AD-A198 821

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

UNITED STATES AIR FORCE RESEARCH INITIATION PROGRAM

CONDUCTED BY UNIVERSAL ENERGY SYSTEMS U.E.S.

Reproduced From
Best Available Copy

1986

TECHNICAL REPORT

VOLUME 2 OF 3

RODNEY C. DARRAH
PROGRAM DIRECTOR, UES

SUSAN K. ESPY
PROGRAM ADMINISTRATOR, UES

LT. COL. CLAUDE CAVENDER
PROGRAM MANAGER, AFOSR

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

DTIC
ELECTE
JAN 16 1988
S D

Best
Available
Copy

SECURITY CLASSIFICATION OF THIS PAGE

(1)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 88-0721
6a. NAME OF PERFORMING ORGANIZATION UNIVERSAL ENERGY SYSTEMS INC. / 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research/XOT
6c. ADDRESS (City, State, and ZIP Code) 4401 Dayton Xenia Rd Dayton OH 45432	7b. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB DC 20332
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR / 8b. OFFICE SYMBOL (If applicable) XOT	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-85-C-0013
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB DC 20332	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102F / PROJECT NO. 3396 / TASK NO. D5 / WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) USAF Research Initiation Program Volume 2	
12. PERSONAL AUTHOR(S) Program Director Rodney C. Darrah	
13a. TYPE OF REPORT Interim	13b. TIME COVERED FROM _____ TO _____
14. DATE OF REPORT (Year, Month, Day) April 1988	15. PAGE COUNT
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

(SEE REVERSE)

DTIC
ELECTE
AUG 16 1988
S D
D

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt. Col. Claude Cavender, Program Manager	22b. TELEPHONE (Include Area Code) (202) 767-4970 / 22c. OFFICE SYMBOL XOT

DD Form 1473, JUN 86 — Previous editions are obsolete. — SECURITY CLASSIFICATION OF THIS PAGE

88 8 12 099

AFOSR-TR- 88-0721

INTRODUCTION

Research Initiation Program - 1985

AFOSR has provided funding for follow-on research efforts for the participants in the Summer Faculty Research Program. Initially this program was conducted by AFOSR and popularly known as the Mini-Grant Program. Since 1983 the program has been conducted by the Summer Faculty Research Program (SFRP) contractor and is now called the Research Initiation Program (RIP). Funding is provided to establish RIP awards to about half the number of participants in the SFRP.

Participants in the 1985 SFRP competed for funding under the 1985 RIP. Participants submitted cost and technical proposals to the contractor by 1 November 1985, following their participation in the 1985 SFRP.

Evaluation of these proposals was made by the contractor. Evaluation criteria consisted of:

1. Technical Excellence of the proposal
2. Continuation of the SFRP effort
3. Cost sharing by the University

The list of proposals selected for award was forwarded to AFOSR for approval of funding. Those approved by AFOSR were funded for research efforts to be completed by 31 December 1986.

The following summarizes the events for the evaluation of proposals and award of funding under the RIP.

A. Rip proposals were submitted to the contractor by 1 November 1985. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share since this is an effort to establish a long term effort between the Air Force and the university.

B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories.

C. Subcontracts were negotiated with the universities. The period of performance of the subcontract was between October 1985 and December 1986.

Copies of the Final Reports are presented in Volumes I through III of the 1985 Research Initiation Program Report. There were a total of 82 RIP awards made under the 1985 program.

Reproduced From
Best Available Copy

AFOSR-TR- 88-0721

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DTIC
This technical report has been reviewed and is
approved for public release IAW AFR 190-12.
Distribution is unlimited.
MATTHEW J. KERPER
Chief, Technical Information Division

Approved for public release;
distribution unlimited.

UNITED STATES AIR FORCE

1986 RESEARCH INITIATION PROGRAM

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

under

USAF Contract Number F49620-85-C-0013

RESEARCH REPORTS

VOLUME II OF III

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington, DC

By

Universal Energy Systems, Inc.

April 1988

Accesion For	
NTIS CRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	

TABLE OF CONTENTS

SECTION	PAGE
INTRODUCTION	i
STATISTICS	ii
PARTICIPANT LABORATORY ASSIGNMENT	vii
RESEARCH REPORTS	xv

INTRODUCTION

Research Initiation Program - 1986

AFOSR has provided funding for follow-on research efforts for the participants in the Summer Faculty Research Program. Initially this program was conducted by AFOSR and popularly known as the Mini-Grant Program. Since 1983 the program has been conducted by the Summer Faculty Research Program (SFRP) contractor and is now called the Research Initiation Program (RIP). Funding is provided to establish RIP awards to about half the number of participants in the SFRP.

Participants in the 1986 SFRP competed for funding under the 1986 RIP. Participants submitted cost and technical proposals to the contractor by 1 November 1986, following their participation in the 1986 SFRP.

Evaluation of these proposals was made by the contractor. Evaluation criteria consisted of:

1. Technical Excellence of the proposal
2. Continuation of the SFRP effort
3. Cost sharing by the University

The list of proposals selected for award was forwarded to AFOSR for approval of funding. Those approved by AFOSR were funded for research efforts to be completed by 31 December 1987.

The following summarizes the events for the evaluation of proposals and award of funding under the RIP.

A. Rip proposals were submitted to the contractor by 1 November 1986. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share since this is an effort to establish a long term effort between the Air Force and the university.

B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories.

C. Subcontracts were negotiated with the universities. The period of performance of the subcontract was between October 1986 and December 1987.

Copies of the Final Reports are presented in Volumes I through III of the 1986 Research Initiation Program Report. There were a total of 98 RIP awards made under the 1986 program.

STATISTICS

Total SFRP Participants	158
Total RIP Proposals submitted by SFRP	134
Total RIP Proposals submitted by GSSSP	7
Total RIP Proposals submitted	141

Total RIP's funded to SFRP	94
Total RIP's funded to GSSSP	4
Total RIP's funded	98

Total RIP's Proposals submitted by HBCU's	14
Total RIP's Proposals funded to HBCU's	9

Laboratory	SFRP Participants	RIP's Submitted	RIP's Funded
AAMRL	9	11 (3 GSSSP)	5 (1 GSSSP)
APL	8	7	6
-AD	11	12 (3 GSSSP)	9 (3 GSSSP)
AEDC	6	6	3
-AL	7	6	5
BRMC	2	2	0
LC	1	1	1
ESMC	0	0	0
ESD	2	1	1
ESC	7	6	5
FDL	13	13 (1 GSSSP)	9
FJSRL	8	8	4
GL	13	9	7
HRL/OT	4	4	2
HRL/LR	4	4	2
HRL/MO	3	2	2
HRL/ID	4	4	3
LMC	2	1	1
-ML	11	8	6
OEHL	4	4	3
RPL	5	4	4
RADC	9	8	7
SAM	17	13	8
WHMC	1	1	1
WL	7	6	4
Total	158	141	98

LIST OF UNIVERSITY THAT PARTICIPATED

Adelphi University	1	Meharry Medical College	2
Alabama A&M University	2	Miami University of Ohio	1
Alabama, University of	5	Miami, University of	1
Alaska, University of	1	Mississippi State University	1
Alfred University	2	Mississippi, University of	1
Auburn University	1	Missouri, University of	2
Boise State University	1	Morehouse College	2
Bradley University	1	Motlow State College	1
Brown University	1	Nebraska, University of	1
Carleton College	1	New Mexico, University of	1
Catholic University of America	1	New Orleans, University of	1
Cedarville College	1	New York University	1
Cincinnati, University of	3	Norfolk State University	1
Colorado, University of	1	North Carolina A&T University	1
Dartmouth College	1	North Carolina, University of	1
Davidson College	1	North Texas State University	1
Dayton, University of	5	Northern Arizona Univeristy	1
Drexel University	1	Northwestern University	1
Duke University	1	Oakwood College	1
Eastern Kentucky University	1	Ohio State University	4
Eastern Montana College	1	Ohio University	3
Edinboro University	1	Oklahoma State University	2
Florida Atlantic University	1	Oklahoma, University of	1
Florida International University	1	Oregon State University	1
Florida State University	2	Pacific University	1
Florida University	3	Paine College	1
Florida, University of	3	Pennsylvania State University	1
Franklin and Marshall College	1	Portland, University of	1
Georgia Institute of Technology	1	Purdue University	2
Georgia, University of	2	Scanton, University of	1
Grambling State University	1	South Carolina, University of	1
Houghton College	1	Southern Illinois University	1
Indiana University	2	Southern Michigan, University of	1
Iowa State University	1	Southern University	1
Iowa, University of	1	Stetson University	1
Jackson State University	4	Stevens Institute of Technology	1
Jefferson State	1	Syracuse, University of	1
Jesm Baromedical Research	1	Tennessee, University of	1
Kansas State Unviersity	1	Texas A&I University	1
Kennesaw University	1	Texas A&M University	2
Lehigh University	1	Texas Southern University	2
Louisiana State University	5	Texas, University of	2
Lowell, University of	2	The Citadel	2
Lyndon State College	1	Toledo, University of	1

(Continued)

LIST OF UNIVERSITY THAT PARTICIPATED

MIT	1	Touglaoo College	1
Maine, University of	1	Trinity University	1
Marquette University	1	Tulsa, University of	2
Mary Washington College	1	U.S. Naval Academy	1
Massachusettes, University of	1	Valparaiso University	1
Vanderbilt University	1	West Virginia University	1
Warren Wilson College	1	Wichita State University	1
Washington State University	2	Wisconsin-Eau Claire	1
Wayne State University	1	Worchester Polytechnic	1
West Florida, University of	1	Wright State University	4
West Georgia College	1	Wyoming, University of	2
		Xavier University	1

PARTICIPANTS LABORATORY ASSIGNMENT

PARTICIPANT LABORATORY ASSIGNMENT (Page 1)

ARMAMENT LABORATORY
(Eglin Air Force Base)

Dr. Prabhat Hajela
University of Flordia
Specialty: Aeronautics & Astronautics

Dr. Boghos D. Sivazlian
The University of Florida
Specialty: Operations Research

Dr. David I. Lawson
Stetson University
Specialty: Mathematics

Mr. Chris Reed (GSRP)
Florida University
Specialty: Aerodynamics

Dr. Barbara Rice
Alabama A&M University
Specialty: Mathematics

Mr. Jim Sirkis (GSRP)
Florida University
Specialty: Engineering Mechanics

Dr. Sally A. Sage
West Georgia College
Specialty: Computer Science

Ms. Jennifer Davidson (GSRP)
Florida University
Specialty: Mathematics

Dr. Meckinley Scott
University of [illegible]
Specialty: Statistics

ARNOLD ENGINEERING DEVELOPMENT CENTER
(Arnold Air Force Systems)

Dr. Glen Johnson
Vanderbilt University
Specialty: Mechanical Eng.

Dr. Arthur A. Mason
The University of Tennessee
Specialty: Physics

ELECTRONIC SYSTEMS DIVISION
(Hanscom Air Force Base)

Dr. Stephan E. Kolitz
University of Massachusetts
Specialty: Operations Research

PARTICIPANT LABORATORY ASSIGNMENT (Page 2)

ENGINEERING AND SERVICES CENTER
(Tyndall Air Force Base)

Dr. Thomas A. Carney
Florida State University
Specialty: Meteorology

Dr. William T. Cooper
Florida State University
Specialty: Chemistry

Dr. Yong S. Kim
The Catholic Univ. of America
Specialty: Civil Engineering

Dr. Cheng Liu
University of North Carolina
Specialty: Civil Engineering

Dr. Roy M. Ventullo
University of Dayton
Specialty: Microbiology

FRANK J. SEILER RESEARCH RESEARCH LABORATORY
(United State Air Force Academy)

Dr. David R. Anderson
University of Colorado
Specialty: Organic Chemistry

Dr. Bernard J. Piersma
Houghton College
Specialty: Physical Chemistry

Dr. William D. Siuru, Jr.
University of Colorado
Specialty: Mechanical Eng.

Dr. Timothy R. Troutt
Washington State University
Specialty: Mechanical Eng.

GEOPHYSICS LABORATORY
(Hanscom Air Force Base)

Dr. John E. Ahlquist
Florida State University
Specialty: Meteorology

Dr. Frank P. Battles
Mass. Maritime Academy
Specialty: Physics

Dr. Wolfgang Christian
Davidson College
Specialty: Physics

Dr. Donald F. Collins
Warren Wilson College
Specialty: Physics

Dr. Patrick T. Gannon, Sr.
Lyndon State College
Specialty: Atmospheric Science

Dr. C. Randal Lishawa
Jefferson State University
Specialty: Physical Chemistry

Dr. Robert M. Nehs
Texas Southern University
Specialty: Mathematics

PARTICIPANT LABORATORY ASSIGNMENT (Page 3)

LOGISTICS COMMAND
(Wright-Patterson Air Force Base)

Dr. Ming-Shing Hung
Kent State University
Specialty: Business Administration
Management Science

LOGISTICS MANAGEMENT CENTER
(Gunter Air Force System)

Dr. Dan B. Rinks
Louisiana State University
Specialty: Quantitative Mgmt. Science

ASTRONAUTICS LABORATORY
(Edwards Air Force Base)

Dr. William M. Grissom
Morehouse College
Specialty: Mechanical Engineering

Dr. Joel R. Klink
Univ. of Wisconsin-Eau Claire
Specialty: Organic Chemistry

Dr. Siavash H. Sohrab
Northwestern University
Specialty: Engineering Physics

Dr. Nicholas E. Takach
University of Tulsa
Specialty: Chemistry

ROME AIR DEVELOPMENT CENTER
(Griffis Air Force Base)

Dr. Donald F. Hanson
University of Mississippi
Specialty: Electrical Engineering

Dr. John M. Jobe
Miami University of Ohio
Specialty: Statistics

Dr. Philipp G. Kornreich
Syracuse University
Specialty: Electrical Engineering

Dr. Mou-Liang Kung
Norfolk State University
Specialty: Mathematics

Dr. Craig G. Prohazka
University of Lowell
Specialty: Electrical Engineering

Dr. Richard S. Quimby
Worcester Polytechnic Institute
Specialty: Physics

Dr. Stephen T. Welstead
University of Alabama
Specialty: Applied Mathematics

PARTICIPANT LABORATORY ASSIGNMENT (Page 4)

WEAPONS LABORATORY
(Kirtland Air Force Base)

Dr. Albert W. Biggs
University of Alabama
Specialty: Electrical Eng.

Dr. Alexandru A. Pelin
Florida International Univ.
Specialty: Computer Science

Dr. Fabian C. Hadipriono
The Ohio State University
Specialty: Engineering, Civil

Dr. Martin A. Shadday, Jr.
University of South Carolina
Specialty: Mechanical Engineering

AERO PROPULSION LABORATORY
(Wright-Patterson Air Force Base)

Dr. Lea D. Chen
The University of Iowa
Specialty: Organic Chemistry

Dr. James C. Ho
Wichita State University
Specialty: Chemistry

Dr. Jacob N. Chung
Washington State University
Specialty: Mechanical Engineering

Dr. Mo Samimy
Ohio State University
Specialty: Mechanical Engineering

Dr. Shirshak K. Dhali
Southern Illinois University
Specialty: Electrical Engineering

Dr. Robert P. Taylor
Mississippi State University
Specialty: Mechanical Engineering

AVIONICS LABORATORY
(Wright-Patterso Air Force Base)

Dr. John Y. Cheung
University of Oklahoma
Specialty: Electrical Engineering

Dr. Dennis W. Whitson
Indiana Univ. of Pennsylvania
Specialty: Physics

Dr. William A. Grosky
Wayne State University
Specialty: Eng. & Applied Science

Dr. George W. Zobrist
University of Missouri-Rolla
Specialty: Electrical Engineering

Dr. Ken Tomiyama
Pennsylvania State University
Specialty: System Science

PARTICIPANT LABORATORY ASSIGNMENT (Page 5)

FLIGHT DYNAMICS LABORATORY
(Wright-Patterson Air Force Base)

Dr. Bor-Chin Chang
Bradley University
Specialty: Electrical Eng.

Dr. George R. Doyle, Jr.
University of Dayton
Specialty: Mechanical Eng.

Dr. Paul S.T. Lee
N.C. A&T State University
Specialty: Quantitative Methods

Dr. V. Dakshina Murty
University of Portland
Specialty: Engineering Mechanics

Dr. Singiresu S. Rao
Purdue University
Specialty: Engineering Design

Dr. Tsun-wai G. Yip
Ohio State University
Specialty: Aeronautics

Dr. Ajmal Yousuff
Drexel University
Specialty: Aeronautics

Dr. Richard W. Young
University or Cincinnati
Specialty: Applied Mechanics

Dr. Peter J. Disimile
University of Cincinnati
Specialty: Fluid Mechanics

MATERIALS LABORATORY
(Wright-Patterson Air Force Base)

Dr. Lokesh R. Dharani
University of Missouri-Rolla
Specialty: Engineering Mechanics

Dr. Gerald R. Graves
Louisiana State University
Specialty: Industrial Engineering

Dr. Gopal M. Mehrotra
Wright State University
Specialty: Metallurgy

Dr. Robert A. Patsiga
Indiana Univ. of Pennsylvania
Specialty: Organic Polymer Chem.

Dr. Nisar Shaikh
University of Nebraska-Lincoln
Specialty: Applied Mathematics

Dr. Stuart R. Stock
Georgia Institute of Technology
Specialty: Metallurgy

PARTICIPANT LABORATORY ASSIGNMENT (Page 6)

HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY
(Wright-Patterson Air Force Base)

Ms. Beverly Girten
Ohio University
Specialty: Physiology

Dr. Albert R. Wellens
University of Miami
Specialty: Experimental Social

Dr. Jacqueline G. Paver
Duke University
Specialty: Biomechanical Eng.

Dr. Robert L. Yolton
Pacific University
Specialty: Psychology, Optometry

Dr. Kuldip S. Rattan
Wright State University
Specialty: Electrical Engineering

HUMAN RESOURCES LABORATORY - LOGISTICS AND HUMAN FACTORS DIVISION
(Wright-Patterson Air Force Base)

Dr. Patricia T. Boggs
Wright State University
Specialty:Decision Science

Dr. Stephen L. Loy
Iowa State University
Specialty: Management Information

HUMAN RESOURCES LABORATORY - OPERATIONS TRAINING DIVISION
(Williams Air Force Base)

Dr. Billy R. Wooten
Brown University
Specialty:Philosophy, Psychology

HUMAN RESOURCES LABORATORY - MANPOWER AND PERSONNEL DIVISION
(Brooks Air Force Base)

Dr. Richard H. Cox
Kansas State University
Specialty:Motor Learning & Control

Dr. Jorge L. Mendoza
Texas A&M University
Specialty: Psychology

USAF OCCUPATIONAL AND ENVIRONMENT HEALTH LABORATORY
(Brooks Air Force Base)

Dr. Clifford C. Houk
Ohio University
Specialty: Inorganic Chemistry

Dr. Ralph J. Rascati
Kennesaw College
Specialty: Biochemistry

Dr. Shirley A. Williams
Jackson State University
Specialty: Physiology

HUMAN RESOURCES LABORATORY - TRAINING SYSTEMS
(Brooks Air Force Base)

Dr. Charles E. Lance
University of Georgia
Specialty: Psychology

Dr. Philip D. Olivier
University of Texas
Specialty: Electrical Engineering

Dr. Doris J. Walker-Dalhouse
Jackson State University
Specialty: Reading Education

SCHOOL OF AEROSPACE MEDICINE
(Brooks Air Force Base)

Dr. Hoffman H. Chen
Grambling State University
Specialty: Mechanical Engineering

Dr. Brenda J. Claiborne
University of Texas
Specialty: Biology

Dr. Vito G. DelVecchio
University of Scranton
Specialty: Biochemistry, Genetics

Dr. Ramesh C. Gupta
University of Maine at Orono
Specialty: Mathematical Statistics

Dr. Frank O. Hadlock
Florida Atlantic University
Specialty: Mathematics

Dr. Parsottam J. Patel
Meharry Medical College
Specialty: Microbiology

Dr. Richard M. Schori
Oregon State University
Specialty: Mathematics

Dr. Shih-sung Wen
Jackson State University
Specialty: Educational Psychology

WILFORD HALL MEDICAL CENTER
(Lackland Air Force Base)

Dr. Donald W. Welch
Texas A&M University
Specialty: Microbiology

RESEARCH REPORTS

MINI-GRANT RESEARCH REPORTS of the

A.F. 1986 RESEARCH INITIATION PROGRAM in vol. 2

include: (to p xviii)

Technical Report Number	Title and Mini-Grant No.	Professor
Volume I		
1	Weather Forecast Evaluation be Decomposition of the Wind Field into Barotropic and Baroclinic Components 760-6MG-041	Dr. Jon Ahlquist
2	An EPR Study of the Role of Catalysts in the Thermal Decomposition of Nitroaromatic Compounds 760-6MG-044	Dr. David Anderson
3	Stellar Scintillometer Based Studies of Optical Turbulence 760-6MG-058	Dr. Frank Battles
4	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-072	Dr. Albert Biggs
5	Basic Research on the Impact of Cognitive Styles on Decision Making 760-6MG-127	Dr. Patricia Boggs
6	A Feasibility Study and Test Application of Uncertainty Estimates to an Atmospheric Dispersion Model with Potential Utility in Air Force Operations 760-6MG-050	Dr. Thomas Carney
7	Design of H Multivariable Optimal Control Systems 760-6MG-013	Dr. Bor-Chin Chang
8	Visualization of Hydrocarbon Jet Diffusion Flames 760-6MG-113	Dr. Lea Chen

9	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-118	Dr. Hoffman Chen
10	Report Not Received In Time. Will Be Provided When Available. 760-6MG-135	Dr. John Cheung
11	Infrared Fluorescense and Photo-acoustic Measurements of NO (v=2) Relazation as a Function of Temp-erature 760-6MG-030	Dr. Wolfgang Christian
12	Heat and Mass Transfer in a Dual-Latent Heat Packed Bed Thermal Storage System 760-6MG-067	Dr. Jacob Chung
13	Long-term Potentiation in Inter-neurons in the Mammalian Brain 760-6MG-101	Dr. Brenda Claiborne
14	The Development of Image Processing Algorithms for AFGL Ultraviolet Camera and Other Imaging Systems 760-6MG-028	Dr. Donald Collins
15	Report Not Received In Time. Will Be Provided When Available. 760-6MG-081	Dr. William Cooper
16	Relationship Between Stages of Motor Learning and Kinesthetic Sensitivity 760-6MG-069	Dr. Richard Cox
17	Received A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-024	Ms. Jennifer Davidson
18	Report Not Received In Time. Will Be Provided When Available. 760-6MG-076	Dr. Vito DelVecchio

19	Investigation of Pulsed Discharges in Nitrogen for Plasma Processing 760-6MG-046	Dr. Shirshak Dhali
20	Modeling of Failure Mechanisms in Ceramic Composites Under Flexural Loading 760-6MG-115	Dr. Lokesh Dharani
21	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-075	Dr. Peter Disimile
22	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-006	Dr. George Doyle
23	Sensitivity of Mesoscale Wind to Variations in Vegetation Canopy Parameters and Surface Properties 760-6MG-100	Dr. Patrick Gannon
24	Effects of Exercise and Dobutamine on Suspension Hypokinesia/ Hypodunamia Deconditioning in Rats 760-6MG-139	Ms. Beverly Girten
25	An Investigation of Computer Communications Using Knowledge-Based Systems 760-6MG-015	Dr. Gerald Graves
Volume II		
26	Droplet Size Distribution Measurement In A Single Element Liquid Rocket Injector 760-6MG-040	Dr. William Grissom
27	A Unified Approach of the Linear Camera Calibration Problem, 760-6MG-070	Dr. William Grosky
28	Survival Anaysis of Radiated Animals for Small Sample Sizes 760-6MG-053	Dr. Ramesh Gupta
29	Report Not Received In Time. Will Be Provided When Available. 760-6MG-054	Dr. Fabian Hadipriono

30	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-073	Dr. Frank Hadlock
31	Studies in Optimum Shape Synthesis for Structures Undergoing Plastic Deformation; 760-6MG-002	Dr. Prabhat Hajela
32	Report Not Received In Time. Will Be Provided When Available. 760-6MG-092	Dr. Donald Hanson
33	Pulsed Power Conductors; 760-6MG-005	Dr. James Ho
34	The Locally Implicit Method for Computational Aerodynamics; 760-6MG-111	Dr. Peter Hoffman
35	Fluorescent Dye Binding Identification of Asbestos on Membrane Filters and in Bulk Materials; 760-6MG-066	Dr. Clifford Houk
36	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Fianl Report. 760-6MG-105	Dr. Ming S. Hung
37	Report Not Received In Time. Will Be Provided When Available. 760-6MG-019	Dr. John Jobe
38	Expert System for Optimal Design. 760-6MG-016	Dr. Glen Johnson
39	Report Not Received In Time. Will Be Provided When Available. 760-6MG-004	Dr. Yong Kim
40	The Synthesis of Some New Energetic Materials; 760-6MG-056	Dr. Joel Klink

41	Report Not Received In Time. Will Be Provided When Available. 760-6MG-094	Dr. Steve Kolitz
42	MBE Grown Al-Cu Alloy Films; 760-6MG-090	Dr. Philipp Kornreich
43	Simulation for Priority Handling Algorithms; 760-6MG-011	Dr. Mou-Liang Kung
44	Received A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-031	Dr. Charles Lance
45	A Neural Network Simulation Generator, Simualtions of Learned Serial Behavior, and a Neural Explanation of Emergent Communication; 760-6MG-001	Dr. David Lawson
46	Data Processing and Statistical Analysis of In-Service Aircraft Transparency Failures 760-6MG-023	Dr. Paul Lee
47	Trajectory Studies of the Bimolecular Reaction of H2Ov/H2O; 760-6MG-107	Dr. C. Lishawa
48	Comparison of Field Rut Depth Measurements and Rutting Predicitions of Asphalt Pavement Under High Tire Pressure and Temperature 760-6MG-009 *NOT PUBLISHABLE AT THIS TIME*	Dr. Cheng Liu
49	Report Not Received In Time. Will Be Provided When Available. 760-6MG-134	Dr. Stephen Loy
50	Received A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-099	Dr. Arthur Mason

13

51	Report Not Received In Time. Will Be Provided When Available. 760-6MG-121	Dr. Gopal Mehrotra
52	Report Not Received In Time. Will Be Provided When Available. 760-6MG-136	Dr. Jorge Mendoza
53	Report Not Received In Time. Will Be Provided When Available. 760-6MG-079	Dr. Dakshina Murty
54	Development of a New Finite Element Grid for Limites Area Weather Models; 760-6MG-120	Dr. Robert Nehs
55	Report Not Received In Time. Will Be Provided When Available. 760-6MG-032	Dr. Philip Olivier
56	Report Not Received In Time. Will Be Provided When Available. 760-6MG-[illegible]	Dr. Parsottam Patel
57	Report Not Received In Time. Will Be Provided When Available. 760-6MG-065	Dr. Robert Patsiga
58	Report Not Received In Time. Will Be Provided When Available. 760-6MG-020	Dr. Jacqueline Paver
59	Automatic Program Generation from Specifications Using Prolong; 760-6MG-117	Dr. Alexandru Pelin
60	Some Novel Aspects of Organic Electrochemistry in Room Temperature Molten Salts 760-6MG-038	Dr. Bernard Piersma
61	Improved Distributed Operating System Communication Protocols; 760-6MG-061	Dr. Craig Prochazka

(cont) → Tunable

62	Tunable Infrared to Visible Light Conversion in Rare Earth and Transition Metal Doped Fluoride Glassess; 760-6MG-042	Dr. Richard Quimby
63	Optimal Structural Modifications to Enhance the Robustness of Actively Controlled Large Flexible Structures; 760-6MG-036	Dr. Singiresu Rao
64	Report Not Received In Time. Will Be Provided When Available. 760-6MG-062	Dr. Ralph Rascati
65	MATRIX-Based Computer Simulation of the Cardiovascular System Under +Gz Stress; 760-6MG-104	Dr. Kuldip Rattan
66	Adaptive Grid Generation Techniques for Transonic Projectile Base Flow Problems. (edc) 760-6MG-034	Mr. Chris Reed
Volume III		
67	Utilization of the Image Algebra 760-6MG-106	Dr. Barbara Rice
68	Simulation Studies of MICAP Allocation Systems for COQ Items 760-6MG-084	Dr. Dan Rinks
69	Computer Modeling of Infrared Signatures 760-6MG-017	Dr. Sally Sage
70	Received A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-059	Dr. Mo Samimy
71	An Intentional Tutor 760-6MG-052	Dr. Richard Schori
72	Report Not Received In Time. Will Be Provided When Available. 760-6MG-025	Dr. Heckinley Scott

73	Report Not Received In Time. Will Be Provided When Available. 760-6MG-089	Dr. Martin Shadday
74	Report Not Received In Time. Will Be Provided When Available. 760-6MG-007	Dr. Nisar Shaikh
75	Report Not Received In Time. Will Be Provided When Available. 760-6MG-142	Mr. Jim Sirkis
76	Two-Dimensional Flight Simulation Model for an Aircraft with a Rapodly Rotating Airfoil 760-6MG-071	Dr. William Siuru
77	Mission Effectiveness Analysis of an Aircraft Attacking Passive Targets 760-6MG-018	Dr. Boghos Sivazlian
78	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-110	Dr. Siavash Sohrab
79	Synchrotron White Beam Topography of Striations and Interface Breakdown in GaAs and of Strain Fields in Si 760-6MG-103	Dr. Stuart Stock
80	Received A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-130	Dr. Nicholas Takach
81	Complete Statistical Classification of Natural Surface Roughness on Gas Turbine Blades 760-6MG-064	Dr. Robert Taylor
82	Evaluation of Atmospheric Effects for Operational Tactical Decision Aid 760-6MG-047	Dr. Ken Tomiyama

83	An Ivestigation Concerning the Formation of a Dynamic Stall Vortex on a Pitching Airfoil 760-6MG-087	Dr. Timothy Troutt
84	Biodegradation of Aqueous Film Forming Foam Components in Laboratory Scale Microcosms 760-6MG-124	Dr. Roy Ventullo
85	Requested A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Fianl Report. 760-6MG-080	Dr. Doris Walker-Dalhouse
86	Received A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Fianl Report. 760-6MG-091	Dr. Donald Welch
87	Effects of Telecommunication Media upon Group Decision Making Processess within a Multi-Team Situation Assessment Task 760-6MG-085	Dr. Albert Wellens
88	Report Not Received In Time. Will Be Provided When Available. 760-6MG-063	Dr. Steve Welstead
89	Can a supervisory Control Simulation System Assess Cognitive Abilities? 760-6MG-049	Dr. Shih-sung Wen
90	Effects on the BICFET of the Fermi Distribution Factor and the Al Mole Fraction 760-6MG-088	Dr. Dennis Whitson
91	The Warehouse Layout Program 760-0MG-038	Dr. Jesse Williams
92	Received A No-Cost Time Extention. To Be Submitted In 1987 Mini-Grant Final Report. 760-6MG-078	Dr. Shirley Williams

93	Report Not Received In Time. Will Be Provided When Available. 760-6MG-051	Dr. Billy Wooten
94	Report Not Received In Time. Will Be Provided When Available. 760-6MG-109	Dr. Tsun-wai Yip
95	Changes in Perceived Workload and Physiological Responses Associated with Monocular Versus Binocular Viewing Conditions 760-6MG-037	Dr. Robert Yolton
96	Finite Element Analysis of Thermo-mechanically Coupled Stress and Temperature Fields 760-6MG-129	Dr. Richard Young
97	Simplification of H∞ Compensators 760-6MG-098	Dr. Ajmal Yousuff
98	Late Appointment. Final Report Will Be Provided When Available. 760-6MG-055	Dr. George Zobrist

1655s

Final Report to Universal Energy Systems
Contract #F49620-85-C-0013/SB5851-0360
Subcontract #S-760-6MG-070

A Unified Approach to the Linear Camera Calibration Problem

William I. Grosky
Computer Science Department
Wayne State University
Detroit, Michigan 48202

1. Introduction

The Numerical Stereo Camera System [Dij84,PoA82] which resides in the Avionics Laboratory of Wright-Patterson Air Force Base utilizes both a passive as well as an active camera to recover 3-D scene information. This is accomplished by solving an overdetermined system of linear equations by the well-known method of least-squares [Gol83]. Specifically, for each camera there are 2 linear equations in the parameters x_W, y_W, and z_W, the world coordinates of a given scene point which is to be determined, where the coefficients are specific functions of x^* and y^*, the known image coordinates of the projection of the given scene point, as well as of the camera geometry (*extrinsic parameters*) and the camera optics (*intrinsic parameters*). The extrinsic parameters give information regarding the camera position and orientation with respect to the world coordinate system, while the intrinsic parameters include the focal length, scale factors to go from units of length to pixels in the image plane, the intersection point of the camera axis with the image plane expressed in pixels, as well as values expressing the different types of possible lens distortions. The term *camera calibration* refers to finding the values of these parameters for a given camera set-up so that the coefficients of x_W, y_W, and z_W in these linear equations can be calculated as functions of x^* and y^*.

There has been much previous work in this area. The techniques to solve this problem range from simple linear equation solving to complex non-linear optimization approaches. The latter methods have been used by [Fai75,Sob74], but are extremely inefficient and must be manually guided. The former methods, most notably used by [Gan84,Str84,Tsa86], while efficient, tend to ignore constraints which the extrinsic and intrinsic parameters must obey. These latter shortcomings become worse when an increasing number of parameters are specified in advance.

In this report, we present unified solutions for many interesting sub-cases of this problem. Most importantly, our solutions satisfy all the necessary constraints as well as being relatively simple to compute.

The organization of this report is as follows. Section 2 derives the linear camera calibration equations. In Section 3, we present our unified solution technique in the context of the, so-called, 1-step method of solution [Tsa86], while Section 4 illustrates various sub-cases of the problem which may be solved using this method. A companion 2-step method [Tsa86] is developed in Section 5. Section 6 presents some experiments we have conducted using our techniques. Finally, in Section 7 we offer our conclusions.

2. The General Linear Camera Calibration Problem

We start with a world coordinate system as shown in Figure 1. We would like to express points in this system with respect to a camera-centered system, x_C, y_C, and z_C, where the camera axis z_C points along the $-z_W$ direction and the (x_C,y_C)-plane is parallel to the image plane. To accomplish this, we first translate the origin of the world system to the focal points of our camera, (x_F,y_F,z_F), and then apply a pan, θ, about the y-axis, a tilt, ϕ, about the x-axis, and finally, a roll, ψ, about the z-axis. See Figure 1. A point $(x_W,y_W,z_W,1)$ expressed in homogeneous coordinates in the world-frame has coordinates $(x_C,y_C,z_C,1) = (x_W,y_W,z_W,1) \times \text{TRAN} \times \text{PAN} \times \text{TILT} \times \text{ROLL}$, expressed in the resulting camera-centered system, where TRANS is the matrix,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -x_F & -y_F & -z_F & 1 \end{bmatrix}$$

PAN is the matrix,

$$\begin{bmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

TILT is the matrix,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi & 0 \\ 0 & \sin\phi & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and ROLL is the matrix,

$$\begin{bmatrix} \cos\psi & -\sin\psi & 0 & 0 \\ \sin\psi & \cos\psi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Defining the extrinsic transformation EXT by TRANS × PAN × TILT × ROLL, we have that EXT is the matrix,

$$\begin{bmatrix} R_{11} & R_{12} & R_{13} & 0 \\ R_{21} & R_{22} & R_{23} & 0 \\ R_{31} & R_{32} & R_{33} & 0 \\ -D_1 & -D_2 & -D_3 & 1 \end{bmatrix}$$

where,

$$R_{11} = \cos\theta\cos\psi + \sin\theta\sin\phi\sin\psi$$

$$R_{12} = -\cos\theta\sin\psi + \sin\theta\sin\phi\cos\psi$$

$$R_{13} = \sin\theta\cos\psi$$

$$R_{21} = \cos\phi\sin\psi$$

$$R_{22} = \cos\phi\cos\psi$$

$$R_{23} = -\sin\phi$$

$$R_{31} = -\sin\theta\cos\psi + \cos\theta\sin\phi\sin\psi$$

$$R_{32} = \sin\theta\sin\psi + \cos\theta\sin\phi\cos\psi$$

$$R_{33} = \cos\theta\cos\phi$$

$$D_1 = x_F R_{11} + y_F R_{21} + z_F R_{31}$$

$$D_2 = x_F R_{12} + y_F R_{22} + z_F R_{32}$$

$$D_3 = x_F R_{13} + y_F R_{23} + z_F R_{33}$$

Using the standard projection equations [DuH73], we get that the screen coordinates (x',y') obey $x' = -Fx_C/z_C$ and $y' = -Fy_C/z_C$, where F is the focal length of the camera system. We now apply a rasterizing transformation to change from length units to pixels. Taking a general linear transformation, we define the pixel coordinates (x*,y*) by

$$\frac{x^* - x_O}{p_x} = a_{11}x' + a_{12}y'$$

and

$$\frac{y^* - y_O}{p_y} = a_{21}x' + a_{22}y'$$

where (x_O,y_O) is the intersection of the image plane with the camera axis expressed in pixels, and px, py are scale factors expressed in units of pixels/unit length.

Expressed in homogeneous coordinates, we thus have that $(x^*,y^*,1) = (x_C,y_C,z_C,1) \times INT$, where INT is the matrix,

$$\begin{bmatrix} p_x F a_{11} & p_y F a_{21} & 0 \\ p_x F a_{12} & p_y F a_{22} & 0 \\ -x_O & -y_O & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus, using $(x^*,y^*,1) = (x_W,y_W,z_W,1) \times EXT \times INT$, we get that $(x^*,y^*,1) = (x_W X_1 + y_W X_2 + z_W X_3 + Z_1, x_W Y_1 + y_W Y_2 + y_W Y_3 + Z_2, D_3 - R_{13}x_W - R_{23}y_W - R_{33}z_W)$, where

$$X_i = p_x F(a_{11}R_{i1} + a_{12}R_{i2}) - x_O R_{i3} \quad \text{for } 1 \le i \le 3$$

$$Y_i = p_y F(a_{21}R_{i1} + a_{22}R_{i2}) - y_O R_{i3} \quad \text{for } 1 \le i \le 3$$

$$Z_1 = x_O D_3 - p_x F(a_{11}D_1 + a_{12}D_2)$$

$$Z_2 = y_O D_3 - p_y F(a_{21}D_1 + a_{22}D_2)$$

which, in turn, implies that,

$$x^* = \frac{x_W X_1 + y_W X_2 + z_W X_3 + Z_1}{D_3 - R_{13}x_W - R_{23}y_W - R_{33}z_W}$$

and

$$y^* = \frac{x_W Y_1 + y_W Y_2 + z_W Y_3 + Z_2}{D_3 - R_{13}x_W - R_{23}y_W - R_{33}z_W}$$

We thus have the linear *camera calibration equations*,

$$x_W(X_1 + x^*R_{13}) + y_W(X_2 + x^*R_{23}) + z_W(X_3 + x^*R_{33}) = x^*D_3 \qquad (1)$$

$$x_W(Y_1 + y^*R_{13}) + y_W(Y_2 + y^*R_{23}) + z_W(Y_3 + y^*R_{33}) = y^*D_3 \qquad (2)$$

Note that we do not include barrel distortion terms in our camera model. This is due to the fact that the images derived from the Numerical Stereo Camera System are rectified to remove all such distortion.

Calibration consists in solving these equations for R_{ij}, $1 \le i,j \le 3$, p_xF, p_yF, x_O, y_O, D_1, D_2, D_3, and a_{ij}, $1 \le i,j \le 2$, so that one can finally find the 3-D coordinates of a scene point corresponding to a known image point, by substituting known values for x^* and y^* in equations 1-2, resulting in 2 linear equations for the unknowns x_W, y_W, and z_W. (Note that by utilizing a second camera, we will have 4 linear equations for these 3 unknowns.) We will solve for these unknowns by substituting known values for x^* and y^*, as well as known values for the corresponding x_W, y_W, and z_W. Note that x_F, y_F, and z_F can be easily calculated from values for the above variables, utilizing the definitions of D_1, D_2, and D_3.

3. A Unified Solution Technique and the 1-Step Method

Notice that equations 1-2 form a homogeneous system in the 12 linearly independent unknowns X_1, X_2, X_3, Y_1, Y_2, Y_3, Z_1, Z_2, R_{13}, R_{23}, R_{33}, and D_3. The fact that these unknowns are linearly independent is important for accuracy, as is mentioned in [Tsa86]. Since we have a homogeneous system, we will put a proper subset of the 12 unknowns on the right-hand side of the equations and solve for the remaining unknowns in terms of the unknowns in the given subset. This is done via the technique of least-squares [Gol83]. This approach, in which we use both equations 1-2 simultaneously, is called the *1-step method* [Tsa86].

Specifically, suppose we have N scene points (x_j,y_j,z_j), $1 \le j \le N$, in the world system, as well as their corresponding image coordinates

$$(x_j^*,y_j^*).$$

Our homogeneous system would then consist of 2N equations. Let $\{\lambda_1,...,\lambda_k\}$, $1 \le k < 12$, be a subset of the 12 unknowns which will be on the left-hand side of our equations, while $\{\rho_1,...,\rho_{12-k}\}$ comprise the remaining unknowns. Our system then becomes of the form $\mathcal{J}\mathcal{G} = \mathcal{K}\mathcal{P}$, where $\mathcal{J}$ is a $2N \times k$ matrix of coefficients, $\mathcal{G}^T = (\lambda_1,...,\lambda_k)$, $\mathcal{K}$ is a $2N \times (12-k)$ matrix of coefficients, and $\mathcal{P}^T = (\rho_1,...,\rho_{12-k})$. We must have $2N \ge k$, and in general, we will have $2N > k$, so that we have an overdetermined system of linear equations to be solved via least-squares. This is done as follows. We have $\mathcal{J}^T\mathcal{J}\mathcal{G} = \mathcal{J}^T\mathcal{K}\mathcal{P}$. Notice that $\mathcal{J}^T\mathcal{J}$ is a square matrix. Thus, $\mathcal{G} = (\mathcal{J}^T\mathcal{J})^{-1}\mathcal{J}^T\mathcal{K}\mathcal{P}$, if $(\mathcal{J}^T\mathcal{J})^{-1}$ exists, which will be the case if the 2N points don't all lie on the same plane. The result of this calculation expresses k unknowns in terms of the remaining 12-k. In solving for our 20 original unknowns, we would then have k equations of the form $\lambda_j = \mathcal{M}\mathcal{P}$, for $\mathcal{M}$ a row vector of coefficients and $1 \le j \le k$. Note that we don't advocate actually computing $\mathcal{M}$ as it is defined above. There are other, more numerically stable techniques, such as the singular value decomposition [Gol83] which may be used. We also have 6 independent constraints on the quantities R_{ij}, $1 \le i,j \le 3$, which express the facts that these are elements of a rotation matrix. These constraints can be written in numerous ways: either that each row is a unit vector and is orthogonal to each of the other 2 rows, or that each column is a unit vector and is orthogonal to each of the other 2 columns, or that some 2 rows are unit vectors, orthogonal to each other, while the remaining row is the cross-product of the 2 given rows, or that some 2 columns are unit vectors, orthogonal to each other, while the remaining column is the cross-product of the 2 given columns.

We thus have k+6 equations, $1 \le k < 11$, in 20 unknowns, which means that not all of the unknowns can be solved for. Let us see how far we can take this solution, however.

For a particular value of k, using at least $\lceil k/2 \rceil$ non-coplanar 3-D points (x_i,y_i,z_i) as well as their known projections

$$(x_j^*,y_j^*),$$

we get, via the method of least-squares, equations of the form

$$p_xF(a_{11}R_{11} + a_{12}R_{12}) - x_OR_{13} = A_1 \quad (3)$$

$$p_xF(a_{11}R_{21} + a_{12}R_{22}) - x_OR_{23} = A_2 \quad (4)$$

$$p_xF(a_{11}R_{31} + a_{12}R_{32}) - x_OR_{33} = A_3 \quad (5)$$

$$p_yF(a_{21}R_{11} + a_{22}R_{12}) - y_OR_{13} = B_1 \quad (6)$$

$$p_yF(a_{21}R_{21} + a_{22}R_{22}) - y_OR_{23} = B_2 \quad (7)$$

$$p_yF(a_{21}R_{31} + a_{22}R_{32}) - y_OR_{33} = B_3 \quad (8)$$

$$R_{13} = C_1 \quad (9)$$

$$R_{23} = C_2 \quad (10)$$

$$R_{33} = C_3 \quad (11)$$

$$x_O D_3 - p_x F(a_{11} D_1 + a_{12} D_2) = E_1 \quad (12)$$

$$y_O D_3 - p_y F(a_{21} D_1 + a_{22} D_2) = E_2 \quad (13)$$

$$D_3 = G \quad (14)$$

where the right-hand side of each equation is a known linear combination of the unknowns in $\mathbf{P} = \{\rho_1,\ldots,\rho_{12-k}\}$. Some of these equations may be identities, however, depending on the nature of $\mathbf{P}$. For example, if $\mathbf{P} = \{D_3, R_{33}\}$, then equations 11 and 14 are not found via least-squares, but are the identities $R_{33} = 0 \bullet D_3 + 1 \bullet R_{33}$ and $D_3 = 1 \bullet D_3 + 0 \bullet R_{33}$, respectively.

Let us define $\mathbf{A} = (A_1, A_2, A_3)$, $\mathbf{B} = (B_1, B_2, B_3)$, $\mathbf{C} = (C_1, C_2, C_3)$, $\mathbf{E} = (E_1, E_2)$, $\Omega_1 = (R_{11}, R_{21}, R_{31})$, $\Omega_2 = (R_{12}, R_{22}, R_{32})$, $\Omega_3 = (R_{31}, R_{32}, R_{33})$, $\mathbf{T}_1 = (a_{11}, a_{12}, 0)$, and $\mathbf{T}_2 = (a_{21}, a_{22}, 0)$. Then equations 3-11 may be expressed as the 3 vector equations,

$$p_x F(a_{11}\Omega_1 + a_{12}\Omega_2) - x_O \Omega_3 = \mathbf{A} \quad (15)$$

$$p_y F(a_{21}\Omega_1 + a_{22}\Omega_2) - y_O \Omega_3 = \mathbf{B} \quad (16)$$

$$\Omega_3 = \mathbf{C} \quad (17)$$

The 6 necessary constraints can now be written as

$$\Omega_1 \bullet \Omega_1 = 1 \quad (18)$$

$$\Omega_2 \bullet \Omega_2 = 1 \quad (19)$$

$$\Omega_3 \bullet \Omega_3 = 1 \quad (20)$$

$$\Omega_1 \bullet \Omega_2 = 0 \quad (21)$$

$$\Omega_1 \bullet \Omega_3 = 0 \quad (22)$$

$$\Omega_2 \bullet \Omega_3 = 0 \quad (23)$$

Using the above equations, we can easily derive that,

$$x_O = -\mathbf{A} \bullet \mathbf{C} \quad (24)$$

$$y_O = -\mathbf{A} \bullet \mathbf{C} \quad (25)$$

Equations 15 and 16 then become 2 simultaneous linear equations for the 2 unknowns Ω_1 and Ω_2. Solving, we get that,

$$p_x F = \frac{|A - (A \bullet C)|}{|T_1|} \tag{26}$$

$$p_y F = \frac{|B - (B \bullet C)|}{|T_2|} \tag{27}$$

$$\Omega_1 = \frac{a_{22}\,|T_1| \dfrac{A - (A \bullet C)C}{|A - (A \bullet C)C|} - a_{12}\,|T_2| \dfrac{B - (B \bullet C)C}{|B - (B \bullet C)C|}}{|T_1 \times T_2|} \tag{28}$$

$$\Omega_2 = \frac{a_{11}\,|T_2| \dfrac{B - (B \bullet C)C}{|B - (B \bullet C)C|} - a_{21}\,|T_1| \dfrac{A - (A \bullet C)C}{|A - (A \bullet C)C|}}{|T_1 \times T_2|} \tag{29}$$

$$\Omega_3 = C \tag{30}$$

Assuming that,

$$C \bullet C = 1 \tag{31}$$

one can verify that these solutions satisfy all constraints 18-23.

Using equations 12 and 13, we also get that,

$$D_1 = \frac{a_{12}\,|T_2| \dfrac{(B \bullet C)G + E_2}{|B - (B \bullet C)C|} - a_{22}\,|T_1| \dfrac{(A \bullet C)G + E_1}{|A - (A \bullet C)C|}}{|T_1 \times T_2|} \tag{32}$$

$$D_2 = \frac{a_{21}\,|T_1| \dfrac{(A \bullet C)G + E_1}{|A - (A \bullet C)C|} - a_{11}\,|T_2| \dfrac{(B \bullet C)G + E_2}{|B - (B \bullet C)|}}{|T_1 \times T_2|} \tag{33}$$

Finally, from equations 15 and 16, we have the following equation which relates T_1 and T_2,

$$\frac{T_1 \bullet T_2}{|T_1|\,|T_2|} = \frac{A \bullet B - (A \bullet C)(B \bullet C)}{|A - (A \bullet C)C|\,|B - (B \bullet C)C|} \tag{34}$$

Taking cross-products of equations 15-17 does not produce any new equations, when one

uses the identity $(\Sigma_1 \times \Sigma_2) \bullet (\Sigma_3 \times \Sigma_4) = (\Sigma_1 \bullet \Sigma_3)(\Sigma_2 \bullet \Sigma_4) - (\Sigma_1 \bullet \Sigma_4)(\Sigma_2 \bullet \Sigma_3)$, along with equation 31.

The above equations do not, as yet, provide numerical solutions for our parameters, as $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are linear combinations of 12-k unknowns. We will next show how to use these equations to solve for these unknowns for various sub-cases of our problem.

4. Some Illustrative Solved Sub-Cases for the 1-Step Method

As the general problem has k+6 equations, $1 \le k \le 11$, in 20 unknowns, by putting $k = 11$, we still cannot completely solve the resulting system. We must specify at least 3 of the 20 unknowns. A particularly interesting case is when we take $a_{11} = 1$, $a_{12} = 0$, $a_{21} = \sin\tau$, and $a_{22} = \cos\tau$. This is the case when the ordinate image-plane axis is not perpendicular to the abscissa image-plane axis, but is offset in the clockwise direction by an angle t. See Figure 2. For this case, set $\mathbf{P} = \{D_3\}$. Thus, $\mathbf{A} = \alpha D_3$, $\mathbf{B} = \beta D_3$, $\mathbf{C} = \gamma D_3$, $\mathbf{E} = \varepsilon D_3$ and $\mathbf{G} = 1 \bullet D_3$, where α, β, γ, and ε are numerically known 3-dimensional vectors. Using equation 31, we can then solve for D_3, after which, back-substituting this value of D_3 in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{E}$, we can solve for τ using equation 34, and x_O, y_O, p_xF, p_yF, Ω_1, Ω_2, Ω_3, D_1, and D_2 using equations 24-33. See [Gan84] for a different interpretation of this situation.

Assume now, for the moment, that we know $\mathbf{T}_1$ and $\mathbf{T}_2$. We then have 16 unknowns. Taking k= 10 thus gives us 16 equations in 16 unknowns. We let $\mathbf{P} = \{D_3, R_{33}\}$. Thus, $\mathbf{A} = \alpha D_3 + \alpha^* R_{33}$, $\mathbf{B} = \beta D_3 + \beta^* R_{33}$, $\mathbf{C} = \gamma D_3 + \gamma^* R_{33}$, $\mathbf{E} = \varepsilon D_3 + \varepsilon^* R_{33}$, and $\mathbf{G} = 1 \bullet D_3 + 0 \bullet R_{33}$, where α, β, γ, ε, α^*, β^*, γ^*, and ε^* are numerically known 3-dimensional vectors having $\gamma_3 = 0$ and $\gamma^*_3 = 1$. We then use equations 31 and 34 to solve for D_3 and R_{33}, followed by back-substituting, as before, to get our final results. When $\mathbf{T}_1 \bullet \mathbf{T}_2 = 0$, which corresponds to perpendicular image-plane axes, solving equations 31 and 34 simultaneously leads one to solve a 4th-degree polynomial equation in the variable $(D_3)^2$. Note that $D_3 > 0$, which narrows down the choices.

The above cases are solved utilizing known corresponding scene and image points such that not all the scene points lie on the same plane. If all the scene points are co-planar, the least-squares based technique previously discussed breaks down, as the matrix for which we need the inverse becomes singular. We can, however, proceed as follows. Assume, without loss of generality, that the plane in which all the scene points lie is $z_W = 0$. Equations 1-2 would then form a homogeneous system in 9 linearly independent unknowns. The unknowns $p_xF(a_{11}R_{31} + a_{12}R_{32}) - x_OR_{33}$, $p_yF(a_{21}R_{31} + a_{22}R_{32}) - y_OR_{33}$, and R_{33} would not appear. We thus would have k+6 equations, $1 \le k \le 8$, in 20 unknowns, an even worse situation than before. It seems that even if we know $\mathbf{T}_1$ and $\mathbf{T}_2$, we must still known at least 2 more unknowns.

As an example, suppose we know x_O and y_O, in addition to $\mathbf{T}_1$ and $\mathbf{T}_2$. Taking $k = 8$, we let $\mathbf{P} = \{D_3\}$. Thus, $\mathbf{A} = \alpha D_3$, $\mathbf{B} = \beta D_3$, $\mathbf{C} = \gamma D_3$, $\mathbf{E} = \varepsilon D_3$, and $\mathbf{G} = 1 \bullet D_3$, where α_3, β_3, γ_3, as well as D_3 are to be determined. Using equations 24, 25, 31, and 34, we have 4 equations for these 4 unknowns. If $\mathbf{T}_1 \bullet \mathbf{T}_2 = 0$, we must solve a cubic equation in the variable $(D_3)^2$.

Cases where other pairs of parameters are known are similarly solved. To solve cases where we know 3 parameters, we let $k = 7$ and $\mathbf{P} = \{D_3, R_{33}\}$.

Table 1 enumerates all presently solved cases when we solve both equations 1 and 2 simultaneously. It is possible to specify more parameters, but this causes computational difficulties in that the degrees of the resulting polynomial equations increase. We have let $\mathbf{P}$ have at most 2 elements. Note that as more information is known, co-planar points allow calibration with less complexity.

5. The 2-Step Method

The 1-step method solved equations 1-2 simultaneously. Now, we show how to solve them sequentially, substituting the solution of the first equation into the second, or vice versa. This method will, in general, allow simpler solutions, but at the price of lower accuracy.

We may either solve equation 1, followed by equation 2, or vice versa. In the former situation, we would have equations 12, 14, 15, and 17 to solve, followed by equations 13 and 16, while in the latter situation, we would have equations 13, 14, 16, and 17 to solve, followed by equations 12 and 15.

In the former case, we would first have solutions to equations 15 and 17 consisting of,

$$x_O = -\mathbf{A} \bullet \mathbf{C} \tag{35}$$

$$p_x F = \frac{|\mathbf{A} - (\mathbf{A} \bullet \mathbf{C})\mathbf{C}|}{|T_1|} \tag{36}$$

$$\Omega_1 = \frac{a_{11}}{|T_1|} \frac{\mathbf{A} - (\mathbf{A} \bullet \mathbf{C})\mathbf{C}}{|\mathbf{A} - (\mathbf{A} \bullet \mathbf{C})\mathbf{C}|} + \frac{a_{12}}{|T_1|} \frac{\mathbf{A} \times \mathbf{C}}{|\mathbf{A} \times \mathbf{C}|} \tag{37}$$

$$\Omega_2 = \frac{a_{12}}{|T_1|} \frac{\mathbf{A} - (\mathbf{A} \bullet \mathbf{C})\mathbf{C}}{|\mathbf{A} - (\mathbf{A} \bullet \mathbf{C})\mathbf{C}|} - \frac{a_{11}}{|T_1|} \frac{\mathbf{A} \times \mathbf{C}}{|\mathbf{A} \times \mathbf{C}|} \tag{38}$$

$$\Omega_3 = \mathbf{C} \tag{39}$$

while in the latter case, we would first have solutions to equations 16 and 17 consisting of,

$$y_O = -\mathbf{B} \bullet \mathbf{C} \tag{40}$$

$$p_y F = \frac{|\mathbf{B} - (\mathbf{B} \bullet \mathbf{C})\mathbf{C}|}{|T_2|} \tag{41}$$

$$\Omega_1 = \frac{a_{21}}{|T_2|}\frac{B-(B\bullet C)C}{|B-(B\bullet C)C|} + \frac{a_{22}}{|T_2|}\frac{B\times C}{|B\times C|} \qquad (42)$$

$$\Omega_2 = \frac{a_{22}}{|T_2|}\frac{B-(B\bullet C)C}{|B-(B\bullet C)C|} - \frac{a_{21}}{|T_2|}\frac{B\times C}{|B\times C|} \qquad (43)$$

$$\Omega_3 = C \qquad (44)$$

Assuming that equation 31 is satisfied, one can verify that both approaches produce solutions satisfying all constraints 18-23. Note, however, that each solution is based on only a single image coordinate.

We now illustrate this technique in the non-coplanar situation, when all we know is T_1 and T_2. Recall that in the 1-step approach, we needed to have $|\mathcal{P}| = 2$. Specifically, we put $\mathcal{P} = \{D_3, R_{33}\}$. In this case, however, we put $\mathcal{P} = \{D_3\}$, resulting in a simpler solution. Thus, $C = \gamma D_3$. Solving equation 1, we would use equation 31 to solve for D_3, then back-substitute this value in equations 35-39 to get numerical values for x_O, p_xF, Ω_1, Ω_2, and Ω_3. Note that in solving equation 1, we would also have an equation in D_1 and D_2. Now, using these values, equation 2 takes the form,

$$\theta_1 y_O + \theta_2 p_y F - p_y F(a_{21}D_1 + a_{22}D_2) = \theta_3 \qquad (45)$$

where θ_1, θ_2, and θ_3 are in terms of the previously found unknowns as well as the image coordinate y^*. We may thus use the same set of points to solve equation 45 via least-squares and y_O, p_yF, and another linear combination of D_1 and D_2. Using these 2 linear equations in D_1 and D_2, we can solve for them individually. We thus have solved numerically for every unknown.

We may turn this process around and start by solving equation 2. Here, we would solve for y_O, p_yF, Ω_1, Ω_2, and Ω_3 first, then using equation 1, we would solve for x_O and p_xF.

Table 2 enumerates all presently solved cases for the 2-step method. Note the contrast with Table 1.

6. Some Experiments Utilizing our Techniques

Using data supplied by the Systems Avionics Division of Wright-Patterson Air Force Base, the only cases we could test were the first 2 cases of Table 1 and the last non co-planar case of Table 2. This is due to the fact that all of the intrinsic and extrinsic parameters were unknown. The given data consisted of 20 points of $(x^*, y^*, x_W, y_W, z_W)$ values and 27 points of $(x^*, y^*, u, v, x_W, y_W, z_W)$ values for the standard calibration object. The former 20 points were to be used for the calibration of the passive camera, while the latter 27 points were to be used for the calibration of the active camera. See Tables 3 and 4, respectively, for a listing of this data.

To see how sensitive the various calibration techniques were to the quantity of points used, we conducted experiments using N points, for $9 \le N \le 20$, for the passive camera. For the active

camera calibration, we used all 27 points in each experiment. Note that the active camera calibration results were only used in solving for the 3-D scene coordinates of given image points and seeing how accurate they were. Solving for the intrinsic and extrinsic parameters were done just from the passive camera calibration.

For Case 1 of Table 1, the results are exhibited in Tables 5a-5f. To determine the accuracy of the results, we used equations 1-2 to solve for x* and y*, respectively. Utilizing the given 20 passive calibration points, we then back-substituted the values of X_1, X_2, X_3, Y_1, Y_2, Y_3, Z_1, Z_2, D_3, R_{13}, R_{23}, R_{33}, x_W, y_W, z_W into these equations to calculate the corresponding values for x* and y*, and compared them to the given values. We also utilized the active camera calibration and solved for the x_W, y_W, z_W values of the given 27 active calibration points, comparing them with their true values. We note that the former results for N = 20 compare precisely with the results achieved by the Numerical Stereo Camera group for this data. This is not surprising due to the fact that even though they don't solve for the intrinsic and extrinsic parameters, the least-squares solution technique is virtually identical. In Tables 5a-5b notice the relative stability of values for the different values of N, except for the values of x_O and y_O. These values seem to be particularly sensitive to the number of points chosen, and hence, to noise.

In Table 5c, pay particular attention to the value of sin τ. This demonstrates that the image axes are not quite perpendicular. Since they are supposed to be perpendicular, we also implemented Case 2 of Table 1 and let $\tau = 0$. This technique is much more complicated than the previous one. Since $\tau = 0$, we had to solve a 4th degree equation for $(D_3)^2$. Even though $D_3 > 0$, we still had up to 4 values of D_3 to consider. We eliminated those values of D_3 which resulted in large errors from back-substituting and comparing the true values of (x*,y*) with the computed values of (x*,y*). These results are shown in Tables 6a-6f. Note here that each value of N gives 2 sets of values for the intrinsic and extrinsic parameters. These 2 sets of values show up for each value of N. Also notice that the set of values exhibiting the smaller errors agrees with that found in Case 1. The errors in Case 2 are very compatible with those found in Case 1: sometimes better, sometimes worse. At the end of this Section, we will present an error analysis of these and other cases which indicates that the Case 2 formulation is less sensitive to error than the Case 1 formulation.

We also tried the 2-step method, assuming that sin $\tau = 0$. This method was a disappointment. Recall that for sin $\tau = 0$, when equation 1 is used alone, the values of x_O, p_xF, D_1, D_3, Ω_1, Ω_2, and Ω_3 are first found, while when equation 2 is used alone, the values of y_O, p_yF, D_2, D_3, Ω_1, Ω_2, and Ω_3 are first found. Since only half as many equations are being used, we initially experimented with this approach using $18 \leq N \leq 20$ points. The values found should have compared favorably with those found for Cases 1 and 2 for N = 9,10. In general, the results were poor. Some values, such as D_3 and p_xF, when equation 1 was used, and D_3 and p_yF, when equation 2 was used, were compatible. Values for x_O in the first case had the wrong sign, however.

Due to these poor results, we increased the range of N up to 47, using all the passive and active points for our calculations. These results for N = 47 are presented here,

Using Equation 1 First

$D_1 = -0.5555678748818383E+01$
$D_2 = 0.9251353575559167E+03$
$D_3 = 0.7829201547124445E+02$

Ω_1 = (0.9979894712507741E+00,
-0.1593003863172282E-01,
-0.6134532697599915E-01)

Ω_2 = (0.1525602072600925E-01,
0.9998181729286259E+00,
-0.1144005739810112E-01)

Ω_3 = (0.6151641329115350E-01,
0.1048117125401979E-01,
0.9980510387474861E+00

$\Omega_1 \bullet \Omega_1$ = 0.1000000000000000E+01

$\Omega_2 \bullet \Omega_2$ = 0.1000000000000000E+01

$\Omega_3 \bullet \Omega_3$ = 0.1000000000000000E+01

$\Omega_1 \bullet \Omega_2$ = -0.1897353801849633E-17

$\Omega_1 \bullet \Omega_3$ = 0.8673617379884035E-18

$\Omega_2 \bullet \Omega_3$ = 0.8673617379884035E-18

x_O = 0.3025599321903800E+02

y_O = -0.1229382964871888E+02

p_xF = 0.3168290708891964E+04

p_yF = -0.1046190042855303E+01

Using Equation 2 First

D_1 = -0.5422030328494327E+02

D_2 = 0.9763671347686911E+01

D_3 = 0.7496470496470639E+02

Ω_1 = (0.9992794859461745E+00,
-0.1370303014013869E-01,
-0.3539400983397718E-01)

Ω_2 = (0.1313878586013341E-01,
0.9997836468007413E+00,
-0.1612550451713603E-01)

Ω_3 = (0.3560732050113826E-01,
0.1564885154856630E-01,
0.9992433298110827E+00

$\Omega_1 \bullet \Omega_1$ = 0.1000000000000000E+01

$\Omega_2 \bullet \Omega_2$ = 0.1000000000000000E+01

$\Omega_3 \bullet \Omega_3$ = 0.1000000000000000E+01

$\Omega_1 \bullet \Omega_2$ =0.7047314121155779E-18

$\Omega_1 \bullet \Omega_3$ = 0.1734723475976807E-17

$\Omega_2 \bullet \Omega_3 = -0.2602085213965211E-16$

$x_O = -0.2463733760501398E+01$
$y_O = 0.5598660453700060E+03$

$p_xF = 0.3304259147950624E+01$
$p_yF = 0.3693358715326379E+04$

We traced down the cause of these results to the solution of the initial least-squares problem. In Case 1, for N = 20, the solutions are,

$X_1/D_3 = 0.4035373463251096E+02$
$X_2/D_3 = -0.6994555310289222E+00$
$X_3/D_3 = -0.2959387694990810E+01$
$Y_1/D_3 = 0.2909991071391078E+00$
$Y_2/D_3 = 0.4912705828771770E+02$
$Y_3/D_3 = -0.8187148116158984E+01$
$R_{13}/D_3 = 0.7439247022530407E-03$
$R_{23}/D_3 = 0.2408085032521297E-03$
$R_{33}/D_3 = 0.1295972726774826E-01$
$Z_1/D_3 = 0.2551636491546525E+03$
$Z_2/D_3 = 0.7867845428491240E+02$

while in the present case, using equation 1 first for N = 47, the solutions are,

$X_1/D_3 = 0.4036247515320486E+02$
$X_2/D_3 = -0.6487010371347563E+00$
$X_3/D_3 = -0.2868196119374038E+01$
$R_{13}/D_3 = 0.7857303573152694E-03$
$R_{23}/D_3 = 0.1338727990451258E-03$
$R_{33}/D_3 = 0.1274780107192484E-01$
$Z_1/D_3 = 0.2550810313218670E+03$

while if we use equation 2 first, the solutions are,

$Y_1/D_3 = 0.3813917438025908E+00$
$Y_2/D_3 = 0.4914043730920537E+02$
$Y_3/D_3 = -0.8257214970198492E+01$
$R_{13}/D_3 = 0.4749878028320434E-03$
$R_{23}/D_3 = 0.2087495916369421E-03$
$R_{33}/D_3 = 0.1332951727458305E-01$
$Z_2/D_3 = 0.7882979394289219E+02$

Notice the compatibility except for the values of R_{13}/D_3, R_{23}/D_3, and R_{33}/D_3. When we present our error analysis, we will see that this least-square problem is more sensitive to errors than the other 2 cases.

Lastly, we compared our techniques with those of [Gan84]. His results are shown in Table

7a-7c. Notice the difference in the dot products of $\Omega_1 \bullet \Omega_2$ and $\Omega_2 \bullet \Omega_3$. His values for D_1, D_3, p_xF, p_yF, x_O, y_O, Ω_1, and Ω_3 are similar to ours. For D_2 and Ω_2, we have

$$\sin\tau\ \Omega_1^{(\text{Our Case 1})} + \cos\tau\ \Omega_2^{(\text{Our Case 1})} = \Omega_2^{(\text{Ganapathy})}$$

and,

$$\sin\tau\ D_1^{(\text{Our Case 1})} + \cos\tau\ D_2^{(\text{Our Case 1})} = D_2^{(\text{Ganapathy})}$$

The (x^*,y^*) and (x_W,y_W,z_W) errors of his approach are otherwise identical to ours.

Let us now compare the tolerance of our techniques to noise. We use a result of [Gol83] which states the following:

Theorem 6.1-3.

Suppose x, r, x', and r' satisfy,

$$\|Ax - b\|_2 = \min,$$

$$r = b - Ax,$$

$$\|(A + \delta A)x' - (b + \delta b)\|_2 = \min,$$

$$r' = (b + \delta b) - (A + \delta A)x',$$

where A and δA are $m \times n$ real matrices with $m \geq n$, $b \neq 0$, and δb is an m element real vector. Assume that,

$$\varepsilon = \max\left\{\frac{\|\delta A\|_2}{\|A\|_2}, \frac{\|\delta b\|_2}{\|b\|_2}\right\} < \frac{\sigma_n(A)}{\sigma_1(A)},$$

and that,

$$\sin\theta = \frac{\rho_{\text{least-squares}}}{\|b\|_2} \neq 1,$$

where $\sigma_1(A)$ and $\sigma_n(A)$ are the largest and the smallest singular values of A, respectively and $r_{\text{least-squares}}$ is the 2-norm of $x_{\text{least-squares}}$, the minimal 2-norm solution to the least-squares problem $Ax = b$.

Then,

$$\frac{\|x' - x\|_2}{\|x\|_2} \leq \varepsilon\left\{\frac{2\kappa_2(A)}{\cos\theta} + \tan\theta\ \kappa_2(A)^2\right\} + O(\varepsilon^2),$$

and,

$$\frac{\|r' - r\|_2}{\|b\|_2} \leq \varepsilon\,(1 + 2\kappa_2(A))\,\min(1, m - n) + O(\varepsilon^2),$$

where $\kappa_2(A) = \sigma_1(A)/\sigma_r(A)$, for $\sigma_1(A)$ as above and $\sigma_r(A)$ the r^{th} largest singular value of A, where r = rank(A).

We assume that x* and y* are accurate to within $1/1000^{th}$ of a pixel. The error upper bounds for Case 1 of Table 1 (and also the technique used by the Numerical Stereo Camera Group and by Ganapathy), and Case 2 of Table 1 are exhibited in Tables 8a-8b. Notice that our improved technique is more resistant to noise than the standard technique. The error bounds for the last non co-planar case of Table 2 for N = 47 is as follows,

Using Equation 1 First

Least-squares error = 0.2405420795083952E+04
Residual error = 0.5819750063015070E+01

Using Equation 2 First

Least-squares error = .3112650142297310E+04
Residual error = 0.6732746874590860E+01

Notice that these errors are much worse that the other techniques, as was expected.

7. Conclusions

We have developed mathematically elegant solutions for the general linear camera calibration problem which are relatively easy to compute. These solutions satisfy all necessary constraints and may be used when certain parameters are known. Some of our techniques have been tried with data provided by the Numerical Stereo Camera Group at Wright-Patterson Air Force Base. One of these techniques happens to be less resistant to noise than the standard techniques. Another one of our techniques was disappointing in its results. With more detailed data we more thoroughly examine our other approaches.

REFERENCES

[Dijk84] J.T. Dijak, *Precise Three-Dimensional Calibration of Numerical Stereo Camera Systems for Fixed and Rotatable Scenes*, AFWAL Technical Report TR-84-1105, Wright-Patterson Air Force Base, 1984

[Duh73] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, Inc., New York, 1973

[Fai75] W. Faig, 'Calibration of Close-Range Photogrammetry Systems: Mathematical Formulation,' *Photogrammetric Engineering and Remote Sensing*, Volume 41 (1975), pp. 1479-1486

[Gan84] S. Ganapathy, 'Decomposition of Transformation Matrices for Robot Vision,' *Proceedings of the IEEE International Conference on Robotics and Automation*,' Atlanta, Georgia, March 1984, pp. 130-139

[Gol83] G.H. Golub, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1983

[PoA82] J.L. Posdamer and M.D. Altschuler, 'Surface Measurement by Space-Encoded Projected Beam Systems,' *Computer Graphics and Image Processing*, Volume 18 (1982), pp. 1-17

[Sob74] I. Sobel, 'On Calibrating Computer Controlled Cameras for Perceiving 3-D Scenes,' *Artificial Intelligence*, Volume 5 (1974), pp. 185-198

[Str84] T.M. Strat, 'Recovering the Camera Parameters from a Transformation Matrix,' *Proceedings of the Image Understanding Workshop*, New Orleans, Louisiana, October 1984, pp. 254-271

[Tsa86] R. Tsai, 'An Efficient and Accurate Camera Calibration Technique for 3-D Machine Vision,' *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, Florida, June 1986, pp. 364-374

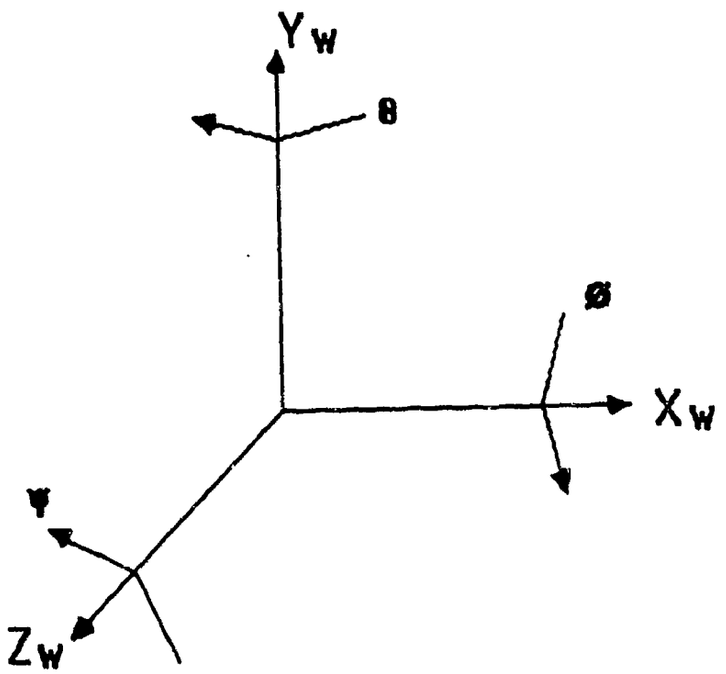


Figure 1 - World Coordinate Frame

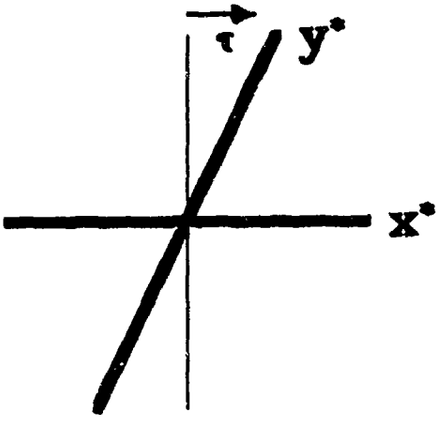


Figure 2 - Skew Angle

Table 1: Solved Cases for the 1-Step Method

Unknowns					Known Affine Transformation T_1, T_2 (K) or Unknown Skew Angle τ (U)	Non Co-Planar (N) or Co-Planar (P)
Scaled Focal Lengths		Displacements		Aspect Ratio		
p_x	p_y	x_O	y_O	p_y/p_x		
×	×	×	×	×	U	N
×	×	×	×	×	K	N
			×		K	P
		×			K	P
		×	×		K	P
	×			×	K	P
	×		×	×	K	P
	×	×		×	K	P
×				×	K	P
×			×	×	K	P
×		×		×	K	P
×	×				K	P
×	×			×	K	P
×	×		×		K	P
×	×	×			K	P

Table 2: Solved Cases for the 2-Step Method with Known Affine Transformation

Unknowns					
Scaled Focal Lengths		Displacements		Aspect Ratio	Non Co-Planar (N) or Co-Planar (P)
p_x	p_y	x_O	y_O	p_y/p_x	
			×		N
		×			N
		×	×		N
	×			×	N
	×	×		×	N
	×	×	×	×	N
×				×	N
×			×	×	N
×		×	×	×	N
×	×			×	N
×	×		×		N
×	×	×			N
×	×	×		×	N
×	×	×	×		N
×	×	×	×	×	N
			×		P
		×			P
	×			×	P
	×		×	×	P
	×	×		×	P
	×	×	×	×	P
×				×	P
×			×	×	P
×		×		×	P
×		×	×	×	P
×	×			×	P
×	×		×	×	P
×	×	×		×	P

Table 3 - Passive Camera Data

x^*	y^*	x_W	y_W	z_W
150.632	80.803	-2.5	0.5	3.0
97.458	145.261	-4.0	1.0	-3.0
149.290	208.382	-2.5	3.0	3.0
97.204	239.473	-4.0	3.0	-3.0
147.415	336.548	-2.5	5.5	3.0
95.390	334.558	-4.0	5.0	-3.0
252.988	337.916	0.0	5.5	3.0
93.701	428.709	-4.0	7.0	-3.0
358.680	340.171	2.5	5.5	3.0
171.781	430.308	-2.0	7.0	-3.0
359.739	210.775	2.5	3.0	3.0
249.319	430.779	0.0	7.0	-3.0
361.369	83.183	2.5	5.0	3.0
327.836	431.901	2.0	7.0	-3.0
255.629	81.922	0.0	5.0	3.0
406.361	433.175	4.0	7.0	-3.0
253.782	209.060	0.0	3.0	3.0
408.177	338.498	4.0	5.0	-3.0
409.838	243.217	4.0	3.0	-3.0
409.900	148.277	4.0	1.0	-3.0

Table 4 - Active Camera Data

x*	y*	u	v	x_w	y_w	z_w
146.14	79.99	73.0	73.0	-2.6069	0.4851	3.0
153.99	207.46	76.0	76.0	-2.3825	2.9751	3.0
147.27	344.90	76.0	76.0	-2.5010	5.6634	3.0
257.47	338.17	96.0	96.0	0.1046	5.4948	3.0
361.43	337.86	114.0	114.0	2.5646	5.4542	3.0
356.90	212.20	112.0	112.0	2.4261	3.0117	3.0
253.27	83.13	93.0	93.0	-0.0644	0.5228	3.0
258.52	206.52	95.0	95.0	0.0907	2.9220	3.0
100.36	152.74	25.0	25.0	-3.9232	1.1577	-3.0
96.48	240.81	25.0	25.0	-3.9927	3.0225	-3.0
93.32	336.84	25.0	25.0	-4.0409	5.0558	-3.0
94.25	420.71	25.0	25.0	-3.9886	6.8306	-3.0
168.59	428.64	40.0	40.0	-2.0841	6.9759	-3.0
253.66	427.11	57.0	57.0	0.0916	6.9183	-3.0
324.15	429.29	70.0	70.0	1.8956	6.9429	-3.0
401.58	429.04	85.0	85.0	3.8764	6.9144	-3.0
253.95	69.34	93.0	93.0	-0.0523	0.2531	3.0
184.24	175.82	81.0	81.0	-1.6743	2.3499	3.0
345.11	201.00	110.0	110.0	2.1435	2.7955	3.0
184.44	275.65	82.0	82.0	-1.6421	4.2972	3.0
321.22	328.46	107.0	107.0	1.6106	5.2847	3.0
158.26	328.81	78.0	78.0	-2.2457	5.3452	3.0
95.65	249.92	25.0	25.0	-4.0107	3.2154	-3.0
101.32	410.34	27.0	27.0	-3.8112	6.6089	-3.0
271.16	391.07	60.0	60.0	0.5280	6.1541	-3.0
387.57	415.34	82.0	82.0	3.5135	6.6301	-3.0
254.46	403.96	57.0	57.0	0.1048	6.4305	-3.0

Table 5a - Intrinsic and Some Extrinsic Parameter Values for Case 1 of Table 1

N	(D_1,D_2,D_3)	(x_O,y_O)	(p_xF,p_yF)
9	(-0.1831585086141255E+01, 0.9595764272024499E+01, 0.7628266917922032E+02)	(0.1811729635180627E+03, 0.5518129888844342E+03)	(0.3090833820861565E+04, 0.3753658262863506E+04)
10	(-0.4464733295194864E+01, 0.9128993394105320E+01, 0.7626684208471991E+02)	(0.7456805801897123E+02, 0.5289013943898248E+03)	(0.3084615320909994E+04, 0.3759059993627350E+04)
11	(-0.7350145450985826E+01, 0.6737235364842700E+01, 0.7547727515821302E+02)	(-0.4104864770754730E+02, 0.4112183481442827E+03)	(0.3034429676230927E+04, 0.3728405543471993E+04)
12	(-0.3676193483904246E+01, 0.6584592974392324E+01, 0.7317454254905257E+02)	(0.1064471597602365E+03, 0.4040610643783840E+03)	(0.2946392929647991E+04, 0.3608228298802134E+04)
13	(-0.2877743233659872E+01, 0.7319033445412277E+01, 0.7340521554601636E+02)	(0.1386822644559356E+03, 0.4401817804977582E+03)	(0.2959971193544175E+04, 0.3617656619134762E+04)
14	(-0.2354735253100525E+01, 0.7145066533756804E+01, 0.7361646307644597E+02)	(0.1597971145181847E+03, 0.4315535734507186E+03)	(0.2970488686863565E+04, 0.3627389064966739E+04)
15	(-0.2589635553503279E+01, 0.6904524305274253E+01, 0.7354863886422616E+02)	(0.1502576650965677E+03, 0.4197118837738590E+03)	(0.2966788498943270E+04, 0.3624123125984675E+04)
16	(-0.3314983053413572E+01, 0.6731830093489205E+01, 0.7425879287285311E+02)	(0.1209932890665242E+03, 0.4112946613107180E+03)	(0.2996495249740732E+04, 0.3659844274120903E+04)
17	(-0.3461786566080526E+01, 0.6001223928892339E+01, 0.7391637504320398E+02)	(0.1150767433432369E+03, 0.3753281576166718E+03)	(0.2980600156247040E+04, 0.3642727540758584E+04)
18	(-0.4510819083218 99E+01, 0.8053690676261673E+01, 0.7575528100290326E+02)	(0.7267736215615843E+02, 0.4763169345148214E+03)	(0.3060973827797450E+04, 0.3734130312614066E+04)
19	(-0.5277656417366124E+01, 0.1058602962860515E+02, 0.7753731157869425E+02)	(0.4158882006778478E+02, 0.6003682930477084E+03)	(0.3139004768081942E+04, 0.3819540386883299E+04)
20	(-0.5059738370646673E+01, 0.9718993818979980E+01, 0.7702205151528707E+02)	(0.5043220502131920E+02, 0.5579805143577485E+03)	(0.3116531860286946E+04, 0.3795319895638180E+04)

Table 5b - Some More Extrinsic Parameter Values for Case 1 of Table 1

N	Ω_1	Ω_2	Ω_3
9	(0.9997643526730505E+00, -0.1574170675975445E-01, -0.1494783571378392E-01)	(0.1541288414594719E-01, 0.9996421376253810E+00, -0.2186411868936040E-01)	(0.1528666499076834E-01, 0.2162857720814593E-01, 0.9996491997302910E+00)
10	(0.9986794183801165E+00, -0.1662672307678447E-01, -0.4861040406826714E-01)	(0.1528748118058132E-01, 0.9994967747965692E+00, -0.2779370594954201E-01)	(0.4904806033989164E-01, 0.2701387145494248E-01, 0.9984310384427710E+00)
11	(0.9961367323453633E+00, -0.1867558213849872E-01, -0.8580695253935020E-01)	(0.1372201168535697E-01, 0.9982245598889294E+00, -0.5796062827351064E-01)	(0.8673705590812407E-01, 0.5655926684762931E-01, 0.9946244177909818E+00)
12	(0.9991475540306655E+00, -0.1619124339067929E-01, -0.3797379243639044E-01)	(0.1397117583202124E-01, 0.9982178346623185E+00, -0.5801690105428676E-01)	(0.3884548262551291E-01, 0.5743690624969224E-01, 0.9975931186009931E+00)
13	(0.9994977505163901E+00, -0.1604573279131964E-01, -0.2732729719283658E-01)	(0.1471029533073186E-01, 0.9987203225881255E+00, -0.4838723448135124E-01)	(0.2806873570289000E-01, 0.4796093940552233E-01, 0.9984547532899931E+00)
14	(0.9996621354376912E+00, -0.1575302860718448E-01, -0.2067503474861381E-01)	(0.1469179239860740E-01, 0.9986152662352405E+00, -0.5051436704578202E-01)	(0.2144215959905586E-01, 0.5019354671290914E-01, 0.9985093097513451E+00)
15	(0.9995923461817180E+00, -0.1557001120171110E-01, -0.2386613089436146E-01)	(0.1437024610577392E-01, 0.9984541829984322E+00, -0.5369115830183542E-01)	(0.2467057927548591E-01, 0.5332630872160107E-01, 0.9982723412558020E+00)
16	(0.9992959191910402E+00, -0.1629665117143272E-01, -0.3379474883366754E-01)	(0.1438153503577134E-01, 0.9983183266034060E+00, -0.5615772625196074E-01)	(0.3465309997891840E-01, 0.5563216631025542E-01, 0.9978498006881994E+00)
17	(0.9992124618761766E+00, -0.1664943078054993E-01, -0.3601738727815045E-01)	(0.1424647918338198E-01, 0.9977196303986192E+00, -0.6597406269224403E-01)	(0.3703368550695294E-01, 0.6540898464459236E-01, 0.9971710840500418E+00)
18	(0.9986037904837991E+00, -0.1707842060077206E-01, -0.4998797036459701E-01)	(0.1508563302469481E-01, 0.9990868063517848E+00, -0.3997470512756205E-01)	(0.5062502649513323E-01, 0.3916479188728599E-01, 0.9979495106310703E+00)
19	(0.9980807390928245E+00, -0.1676143915097318E-01, -0.5961453186522168E-01)	(0.1625743746938968E-01, 0.9998279662114732E+00, -0.8929373346326677E-02)	(0.5975394529945442E-01, 0.7943056025071702E-02, 0.9981815335309166E+00)
20	(0.9982304889691113E+00, -0.1698622305784682E-01, -0.5698560448672955E-01)	(0.1589844514707787E-01, 0.9996836780006571E+00, -0.1948803173703906E-01)	(0.5729860674042830E-01, 0.1854756494280470E-01, 0.9981847812405769E+00)

Table 5c - Yet More Extrinsic Parameter Values for Case 1 of Table 1

N	sin τ	($\Omega_1 \bullet \Omega_1$, $\Omega_2 \bullet \Omega_2$, $\Omega_3 \bullet \Omega_3$)	($\Omega_1 \bullet \Omega_2$, $\Omega_1 \bullet \Omega_3$, $\Omega_2 \bullet \Omega_3$)
9	-0.2426121126232515E-03	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.1463672932855431E-17, 0.0000000000000000E+00, -0.1387778780781446E-16)
10	0.1330101592354715E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(-0.4878909776184770E-18, -0.2602085213965211E-17, -0.2081668171172169E-16)
11	0.9397371720494471E-03	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.8673617379884035E-18, 0.2775557561562891E-16, -0.1647987302177967E-16)
12	-0.2062566834795333E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.8131516293641283E-18, 0.8673617379884035E-18, -0.2949029909160572E-16)
13	-0.2523031214773828E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(-0.6505213034913027E-18, -0.2602085213965211E-17, -0.3382710778154774E-16)
14	-0.3373657202390372E-02	(0.1000000000000000E+01, 0.9999999999999999E+00, 0.1000000000000000E+01)	(-0.2547875105340935E-17, 0.2602085213965211E-17, -0.1908195823574488E-16)
15	-0.3012013650800166E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.2168404344971009E-18, -0.2602085213965211E-17, -0.3209238430557093E-16)
16	-0.3016452288493705E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.5421010862427522E-19, 0.1734723475976807E-17, -0.2688821387764051E-16)
17	-0.3227210156769185E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(-0.2385244779468110E-17, -0.2602085213965211E-17, -0.4163336342344337E-16)
18	-0.2216315421997831E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.1192622389734055E-17, -0.8673617379884035E-18, -0.6938893903907228E-17)

N	sin τ	$(\Omega_1 \bullet \Omega_1, \Omega_2 \bullet \Omega_2, \Omega_3 \bullet \Omega_3)$	$(\Omega_1 \bullet \Omega_2, \Omega_1 \bullet \Omega_3, \Omega_2 \bullet \Omega_3)$
19	-0.1167698302221235E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.9215718466126788E-18, -0.8673617379884035E-18, -0.1474514954580286E-16)
20	-0.1571754743190445E-02	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.3794707603699266E-18, 0.1734723475976807E-17, -0.1387778780781446E-16)

Table 5d - (x*,y*) Errors for the N Passive Points for Case 1 of Table 1

N	Minimum Absolute (x*,y*) Errors	Maximum Absolute (x*,y*) Errors	Average Absolute (x*,y*) Errors
9	(0.4024382966179019E-02, 0.2520039061903745E-01)	(0.6278764950668325E+00, 0.337361043809551 1E+00)	(0.1996207355516254E+00, 0.1619083382402230E+00)
10	(0.1582805831719725E-01, 0.486409651336572 3E-02)	(0.6423776167331354E+00, 0.3579698859302312E+00)	(0.1980711788774709E+00, 0.1433714251349823E+00)
11	(0.1100879490229545E-01, 0.3546885851960724E-02)	(0.6126943580061024E+00, 0.1779486288334908E+01)	(0.2036751592774664E+00, 0.5425102117542626E+00)
12	(0.6028025186776631E-01, 0.4981808377119989E-01)	(0.6251388819773638E+00, 0.5405317225314548E+00)	(0.1866369986295124E+00, 0.2440980163625828E+00)
13	(0.5626308878590436E-01, 0.6862320663533694E-01)	(0.6294361701831015E+00, 0.5330819869760717E+00)	(0.1968863732461189E+00, 0.2435047243560840E+00)
14	(0.1521096431766011E-01, 0.5570217803203903E-01)	(0.6275501543913080E+00, 0.6479275709758099E+00)	(0.1964842256679358E+00, 0.2340618985205403E+00)
15	(0.1060293880712493E-01, 0.1798497007972699E-01)	(0.6381619431186145E+00, 0.6360967072482140E+00)	(0.1986282863810982E+00, 0.2206215893212807E+00)
16	(0.5823335105219485E-03, 0.1669982758357946E-01)	(0.6438620235829795E+00, 0.6840549054947473E+00)	(0.1970078598763643E+00, 0.2150063496161627E+00)
17	(0.5344401058778203E-02, 0.3883330810575103E-01)	(0.6234822848674817E+00, 0.7494738527229288E+00)	(0.2211059828609882E+00, 0.2310087872370876E+00)
18	(0.1093231859528032E-03, 0.6426312489566044E-02)	(0.6983313208065738E+00, 0.7993173679485176E+00)	(0.2657887676045804E+00, 0.2550204308148262E+00)
19	(0.9696063411951172E-02, 0.1659786396299268E-02)	(0.7093636422777081E+00, 0.8721493669122538E+00)	(0.3337950704543147E+00, 0.2726283485370778E+00)
20	(0.1196666111559352E-01, 0.1104991905939912E-01)	(0.8676260585016280E+00, 0.8548678360842814E+00)	(0.3275083442645064E+00, 0.2517123045060989E+00)

Table 5e - (x*,y*) Errors for the Remaining 20-N Passive Points for Case 1 of Table 1

N	Minimum Absolute (x*,y*) Errors	Maximum Absolute (x*,y*) Errors	Average Absolute (x*,y*) Errors
9	(0.4318829708654448E-01, 0.4888170622348298E-01)	(0.1085567570463240E+01, 0.2083018054252555E+01)	(0.517044065021852E+00, 0.1109180894636488E+01)
10	(0.1465021322504185E+00, 0.2065654248379012E+00)	(0.1153880902996320E+01, 0.152252475460716E+01)	(0.4791986001591418E+00, 0.1006125648363220E+01)
11	(0.3546885851960724E-02, 0.2653134689094792E+00)	(0.1779486288334908E+01, 0.1544381105053390E+01)	((0.5425102117542626E+00, 0.7434521067634922E+00)
12	(0.9180147631957070E-01, 0.1671369401896072E+00)	(0.2762056663705039E+01, 0.6350726550878676E+00)	(0.1134050503362083E+01, 0.3644958724234577E+00)
13	(0.2873838530444388E+00, 0.5854218282263446E-01)	(0.2558850310582841E+01, 0.8515339844744432E+00)	(0.1191316316173437E+01, 0.4053227062641841E+00)
14	(0.3079524907702265E+00, 0.1090904863855258E-01)	(0.2363529864137263E+01, 0.7175003472886894E+00)	(0.1193151806351617E+01, 0.3330876993584605E+00)
15	(0.6185071136478868E+00, 0.4214031062230106E-01)	(0.2410267998111863E+01, 0.7195982281504740E+00)	(0.1381879288108943E+01, 0.3893003510051486E+00)
16	(0.5915103810316076E+00, 0.3300457824401093E-01)	(0.2045950735991312E+01, 0.7132511567935467E+00)	(0.1283208487657354E+01, 0.3777877373876892E+00)
17	(0.1274923978958668E+01, 0.3071707323791273E-01)	(0.2138230193414017E+01, 0.4326879986239760E+00)	(0.1606137274645885E+01, 0.2724427290289479E+00)
18	(0.4258943178853656E+00, 0.3103607417319765E-01)	(0.1360616367263560E+01, 0.1590477041850171E+00)	(0.8932553425744629E+00, 0.9504188917910739E-01)
19	(0.4665989062911535E+00, 0.1260788975383633E+00)	(0.4665989062911535E+00, 0.1260788975383633E+00)	(0.4665989062911535E+00, 0.1260788975383633E+00)

Table 5f - (x_W,y_W,z_W) Errors for the 27 Active Points for Case 1 of Table 1

N	Minimum Absolute (x_W,y_W,z_W) Errors	Maximum Absolute (x_W,y_W,z_W) Errors	Average Absolute (x_W,y_W,z_W) Errors
9	(0.1596321594636851E-03, 0.1604080310115830E-03, 0.2398422677807321E-03)	(0.1062606581433911E-01, 0.2571214424643564E-01, 0.6516362741600346E-01)	(0.4291571159947359E-02, 0.8034829800058927E-02, 0.2906674855763060E-01)
10	(0.2231541851920316E-03, 0.9455040128725223E-04, 0.2095923984274828E-02)	(0.1008291397673400E-01, 0.2954237727096110E-01, 0.6431078675795598E-01)	(0.3834371535277658E-02, 0.8537695880473063E-02, 0.2847846561736576E-01)
11	(0.5874181587421923E-03, 0.6441264379895006E-03, 0.1949101560040800E-02)	(0.1360758453313921E-01, 0.2852816216400478E-01, 0.6661126773007187E-01)	(0.6704762345042889E-02, 0.1062646350382908E-01, 0.2832002480480659E-01)
12	(0.5723247195239534E-03, 0.6902966097921404E-03, 0.2902105606079841E-02)	(0.2659297626608903E-01, 0.1715713850385292E-01, 0.8141745984604043E-01)	(0.8657531770053093E-02, 0.6600714604842277E-02, 0.2961807564423347E-01)
13	(0.2366571214813540E-03, 0.5002917458114542E-03, 0.5838282212655699E-03)	(0.2585714019354279E-01, 0.1562269841858321E-01, 0.7988907500026632E-01)	(0.7630623469984926E-02, 0.5930789683703837E-02, 0.2925086968027759E-01)
14	(0.2620432317836929E-03, 0.3081317045505738E-03, 0.2215389212247842E-03)	(0.2014776891788705E-01, 0.1571298107924668E-01, 0.7398021296735702E-01)	(0.6165461561944539E-02, 0.5231145830618758E-02, 0.2877566459091165E-01)
15	(0.3290856669289077E-03, 0.2319125390604082E-03, 0.1482631337419660E-02)	(0.1959090012186748E-01, 0.1571543461366742E-01, 0.7369696110624724E-01)	(0.6309233325292110E-02, 0.5275784601378159E-02, 0.2872111993586621E-01)
16	(0.4680758495689830E-05, 0.2523777423717455E-03, 0.6599012410193339E-03)	(0.1159424805337950E-01, 0.1577635235951957E-01, 0.6715091777789306E-01)	(0.4989976665353942E-02, 0.5496504866989661E-02, 0.2828475089091295E-01)
17	(0.7413343400246575E-03, 0.1078243188135097E-04, 0.2690921676602231E-02)	(0.1178230036326688E-01, 0.1768484491256439E-01, 0.6822952017616446E-01)	(0.5764685763450729E-02, 0.6056611630930922E-02, 0.2813445719662612E-01)
18	(0.2963435168703565E-03, 0.7281338429052653E-03, 0.7494847249265302E-03)	(0.1057436627895414E-01, 0.1599502822766907E-01, 0.6880433799555385E-01)	(0.5089673248663983E-02, 0.6310818465670224E-02, 0.2800549607477694E-01)

N	Minimum Absolute (x_W,y_W,z_W) Errors	Maximum Absolute (x_W,y_W,z_W) Errors	Average Absolute (x_W,y_W,z_W) Errors
19	(0.5136999971262490E-03, 0.1026301206041857E-03, 0.2830297666758241E-03)	(0.1212875722905228E-01, 0.1417640253716423E-01, 0.6807941201017820E-01)	(0.5580253046169232E-02, 0.6202797737520730E-02, 0.2823619399699645E-01)
20	(0.4726746520118397E-03, 0.2427989610949144E-03, 0.2614928934829841E-03)	(0.1084155950314045E-01, 0.1477536029472559E-01, 0.6846505645678480E-01)	(0.5416685985171796E-02, 0.6257914534587310E-02, 0.2808319601662889E-01)

Table 6a - Intrinsic and Some Extrinsic Parameter Values for Case 2 of Table 1

N	(D_1,D_2,D_3)	(x_O,y_O)	(p_xF,p_yF)
9	(-0.2117785606180813E+01, 0.9711010977197643E+01, 0.7648131848313530E+02)	(0.1695765228271484E+03, 0.5574724121093750E+03)	(0.3099312744140625E+04, 0.3763842285156250E+04)
	(0.9979425807990820E+01, 0.4656437420176663E+01, 0.6541544557453690E+02)	(0.6493806152343750E+03, 0.3090334472656250E+03)	(0.2586960205078125E+04, 0.3190581787109375E+04)
10	(-0.2741018071862934E+01, 0.7876156455837568E+01, 0.7479402506322143E+02)	(0.1442289123535156E+03, 0.4675102539062500E+03)	(0.3021485351562500E+04, 0.3684445312500000E+04)
	(0.1710286369428683E+01, 0.4625431296509269E+01, 0.7037678753742525E+02)	(0.3230498046875000E+03, 0.3075107421875000E+03)	(0.2825020263671875E+04, 0.3456023925781250E+04)
11	(-0.6215886917185746E+01, 0.6130165758853136E+01, 0.7439137341692268E+02)	(0.4447823524475098E+01, 0.3815520019531250E+03)	(0.2991073486328125E+04, 0.3672801513671875E+04)
	(0.1516089436755234E+01, 0.2015060658021544E+01, 0.6525479732049557E+02)	(0.3141831054687500E+03, 0.1789614257812500E+03)	(0.2605170654296875E+04, 0.3198914306640625E+04)
12	(-0.6134541786018752E+01, 0.8710342116435703E+01, 0.7594810228422809E+02)	(0.7445616722106934E+01, 0.5084414062500000E+03)	(0.3061356689453125E+04, 0.3748012207031250E+04)
	(0.1495825077341200E+01, 0.2115690913387500E+01, 0.6597090700883577E+02)	(0.3135544433593750E+03, 0.1838894500732422E+03)	(0.2636434814453125E+04, 0.3234786132812500E+04)
13	(-0.5586095349415532E+01, 0.9233850023650223E+01, 0.7740232712588345E+02)	(0.2945146179199219E+02, 0.5340441894531250E+03)	(0.3124048095703125E+04, 0.3816273681640625E+04)
	(0.4012384416602852E+01, 0.2333108645395753E+01, 0.5939107154588387E+02)	(0.4135627441406250E+03, 0.1948193969726563E+03)	(0.2363134033203125E+04, 0.2907780273437500E+04)

N	(D_1,D_2,D_3)	(x_O,y_O)	(p_xF,p_yF)
14	(-0.5763108797928832E+01, 0.1154343362424234E+02, 0.8009674184605099E+02)	(0.2216581726074219E+02, 0.6469587402343750E+03)	(0.3238117675781250E+04, 0.3944461914062500E+04)
	(0.1730919197080688E+01, 0.1837878503278983E+01, 0.6320904673322807E+02)	(0.3234819335937500E+03, 0.1703465881347656E+03)	(0.2531546875000000E+04, 0.3098187011718750E+04)
15	(-0.5710921923129731E+01, 0.1078593195395788E+02, 0.7923857391279075E+02)	(0.2426226806640625E+02, 0.6099584960937500E+03)	(0.3201484863281250E+04, 0.3904172119140625E+04)
	(0.1553486058086590E+01, 0.1739745534151200E+01, 0.6362914262618872E+02)	(0.3163630371093750E+03, 0.1655535583496094E+03)	(0.2548681396484375E+04, 0.3118000244140625E+04)
16	(-0.5711357836575819E+01, 0.1189867305873222E+02, 0.8044786702915324E+02)	(0.2417761230468750E+02, 0.6642822265625000E+03)	(0.3252593017578125E+04, 0.3961281005859375E+04)
	(-0.1085582931278695E+00, -0.5495675906080479E-01, 0.6322840044200547E+02)	(0.2497030639648438E+03, 0.7768769836425781E+02)	(0.2535206787109375E+04, 0.3090893554687500E+04)
17	(-0.6170784606416882E+01, 0.1147673093030904E+02, 0.8085046038246436E+02)	(0.5729712486267090E+01, 0.6436789550781250E+03)	(0.3266466308593750E+04, 0.3982834960937500E+04)
	(-0.2513364780541767E+00, -0.3342507990982095E+00, 0.6292425282120246E+02)	(0.2438914794921875E+03, 0.6406332397460938E+02)	(0.2522071289062500E+04, 0.3074283691406250E+04)
18	(-0.5916601571588777E+01, 0.1116288279780411E+02, 0.7969841595182842E+02)	(0.1588509178161621E+02, 0.6283056640625000E+03)	(0.3222129150390625E+04, 0.3925753662109375E+04)
	(-0.4012377825239448E+00, -0.6936300388133292E+00, 0.6038329885080842E+02)	(0.2381201324462891E+03, 0.4691032409667969E+02)	(0.2425977050781250E+04, 0.2944851074218750E+04)

N	(D_1,D_2,D_3)	(x_O,y_O)	(p_xF,p_yF)
19	(-0.5778371272688401E+01, 0.1188780039101138E+02, 0.7923806736729463E+02)	(0.2133229064941406E+02, 0.6636633300781250E+03)	(0.3208008789062500E+04, 0.3900721191406250E+04)
	(0.1424489714816296E+00, -0.2422193228875957E+01, 0.5064233217780490E+02	(0.2603981933593750E+03, -0.3527543640136719E+02)	(0.2036758300781250E+04, 0.2446442382812500E+04)
20	(-0.5499204284305338E+01, 0.1160534927416639E+02, 0.8023013952787822E+02)	(0.3265710449218750E+02, 0.6497844238281250E+03)	(0.3246262939453125E+04, 0.3949144287109375E+04)
	(-0.3776929171147232E+00, 0.1286340120948821E+02, 0.1783916034245613E+02)	(0.2397465972900391E+03, 0.6796608886718750E+03)	(0.7372712402343750E+03, 0.8561599121093750E+03)

Table 6b - Some More Extrinsic Parameter Values for Case 2 of Table 1

N	Ω_1	Ω_2	Ω_3
9	(0.9996996899353168E+00, -0.1583888870775079E-01, -0.1869918575058691E-01)	(0.1545345309451233E-01, 0.9996687713424279E+00, -0.2058004835213750E-01)	(0.1901895713984270E-01, 0.2028490096658342E-01, 0.9996133262727687E+00)
	(0.9865684982106127E+00, 0.6394200505487427E-03, 0.1633468379873185E+00)	0.1329601629697633E-01, 0.9963597915278151E+00, -0.8420440473799108E-01)	(-0.1628060634285065E+00, 0.9829686814656230E+00, 0.9829686814656230E+00)
10	(0.9995339761353497E+00, -0.1591237920126076E-01, -0.2605046523986839E-01)	(0.1478333436886592E-01, 0.9989668087902511E+00, -0.4297403821332383E-01)	(0.2670736932003259E-01, 0.4256889854785056E-01, 0.9987365044396978E+00)
	(0.9992655311438010E+00, -0.1019815381532174E-01, 0.3693773039395343E-01)	(0.1332816612093323E-01, 0.9962483053655716E+00, -0.8550532733763181E-01)	(-0.3592712423434641E-01, 0.8593783634107756E-01, 0.9956525147004171E+00)
11	(0.9972659790844294E+00, -0.1823689278414770E-01, -0.7160993438311423E-01)	(0.1354852593888468E-01, 0.9977658864140141E+00, -0.6541921241760812E-01)	(0.7264299281870067E-01, 0.6427014586910672E-01, 0.9952850566266459E+00)
	(0.9992041828066820E+00, -0.7376809078964340E-02, 0.3919928251184508E-01)	(0.1217017733021118E-01, 0.9922712588737913E+00, -0.1234894149179302E+00)	(-0.3798536356785535E-01, 0.1238682021377319E+00, 0.9915713694201662E+00)
12	(0.9974373315654879E+00, -0.1685178357247631E-01, -0.6953263255440180E-01)	(0.1468342163238262E-01, 0.9993933991554882E+00, -0.3157896219961446E-01)	(0.7002261583720448E-01, 0.3047705882898497E-01, 0.9970797270812648E+00)
	(0.9992547263322381E+00, -0.7585724433455896E-02, 0.3784770386038777E-01)	(0.1214408985478145E-01, 0.9924919776412608E+00, -0.1217053630672785E+00)	(-0.3664031910727500E-01, 0.1220742851814368E+00, 0.9918444212239932E+00)
13	(0.9979555357743864E+00, -0.1670232980396576E-01, -0.6169101066101056E-01)	(0.1510665859882281E-01, 0.9995414671064597E+00, -0.2624203499434977E-01)	(0.6210102642658957E-01, 0.2525643905592426E-01, 0.9977502567290966E+00)
	(0.9963610592565432E+00, -0.2983224493589524E-02, 0.8518063070372418E-01)	(0.1334727135867717E-01, 0.9925184510056449E+00, -0.1213629876059136E+00)	(-0.8418129212869815E-01, 0.1220582838780727E+00, 0.9889465535547718E+00)

N	Ω_1	Ω_2	Ω_3
14	(0.9979316698046449E+00, -0.1569605834386695E-01, -0.6233791906519810E-01)	(0.1576302232695928E-01, 0.9998755861540519E+00, 0.5825291503569603E-03)	(0.6232101995340779E-01 -0.1563958297766524E-02, 0.9980549306057306E+00)
	(0.9990955459974864E+00, -0.7516209259694723E-02, 0.4185207959407818E-01)	(0.1280684764409833E-01, 0.9917405341187356E+00, -0.1276193466497375E+00)	(-0.4054719005560599E-01, 0.1280399140278120E+00, 0.9909398093700470E+00)
15	(0.9979378937104435E+00, -0.1570511379792730E-01, -0.6223591967150355E-01)	(0.1524686509730857E-01, 0.9998530918876726E+00, -0.7831203443004984E-02)	(0.6234976665127276E-01, 0.6866151997691999E-02, 0.9980307422896734E+00)
	(0.9992184939974095E+00, -0.7790594959765631E-02, 0.3875187587358743E-01)	(0.1272797275949880E-01, 0.9915829967675622E+00, -0.1288454858770541E+00)	(-0.3742191821623413E-01, 0.1292380251769330E+00, 0.9909072271839498E+00)
16	(0.9979922338696359E+00, -0.1518208200156881E-01, -0.6148988146021803E-01)	(0.1550331071184010E-01, 0.9998685321870065E+00, 0.4750336743157146E-02)	(0.6140967751801155E-01, -0.5694095915848407E-02, 0.9980964025477874E+00)
	(0.9998745541381291E+00, -0.1037493277631142E-01, 0.1196815595505377E-01)	(0.1212548355219310E-01, 0.9875324017569503E+00, -0.1569481701988748E+00)	(-0.1019061507971759E-01, 0.1570736013785782E+00, 0.9875343209804207E+00)
17	(0.9976350987704545E+00, -0.1553528224243200E-01, -0.6695419857569478E-01)	(0.1549719048174159E-01, 0.9998793190628917E+00, -0.1088300280581424E-02)	(0.6696302553229068E-01, 0.4812458902983541E-04, 0.9977554564599412E+00)
	(0.9998983923594802E+00, -0.1057065654927565E-01, 0.9563805572261838E-02)	(0.1197822331381769E-01, 0.9867722628932841E+00, -0.1616694880016382E+00)	(-0.7728345434249662E-02, 0.1617676185291259E+00, 0.9867986168779647E+00)
18	(0.9977489847487767E+00, -0.1584284119804838E-01, -0.6516109127047233E-01)	(0.1561738069030328E-01, 0.9998701678096354E+00, -0.3967990008265559E-02)	(0.6521549549884254E-01, 0.2941412433673671E-02, 0.9978668684948586E+00)
	(0.9999170102274650E+00, -0.1171831204713492E-01, 0.5352926352347676E-02)	(0.1246045118571517E-01, 0.9852300210872610E+00, -0.1707821498419658E+00)	(-0.3272585219073351E-02, 0.1708346765477115E+00, 0.9852942725271586E+00)

N	Ω_1	Ω_2	Ω_3
19	(0.9977886657073789E+00, -0.1594379214003305E-01, -0.6452576290199145E-01)	(0.1632614662892001E-01, 0.9998521232607014E+00, 0.5402642623587430E-02)	(0.6443008243157539E-01, -0.6444152641164033E-02, 0.9979014166612840E+00)
	(0.9998524847538297E+00, -0.1132065114859814E-01, 0.1291710451938378E-01)	(0.1386292000669417E-01, 0.9759000489516721E+00, -0.2177772116292521E+00)	(-0.1014042309181133E-01, 0.2179241549569386E+00, 0.9759130260970099E+00)
20	(0.9980842455383521E+00, -0.1569115821215260E-01, -0.5984669048577151E-01)	(0.1574905435273172E-01, 0.9998758530214807E+00, 0.4958140387847611E-03)	(0.5983148080344695E-01, -0.1437392962119310E-02, 0.9982074572985014E+00)
	(0.9996790870784272E+00, -0.6733794352149931E-02, 0.2442086959273657E-01)	(0.1872461982181236E-01, 0.8457169305445573E+00, -0.5333031614408648E+00)	(0.1706198905670390E-01, -0.5335892890640849E+00, -0.8455716167927548E+00)

Table 6c - Yet More Extrinsic Parameter Values for Case 2 of Table 1

N	$A \cdot B - (A \cdot B)(A \cdot C)$	$(\Omega_1 \bullet \Omega_1, \Omega_2 \bullet \Omega_2, \Omega_3 \bullet \Omega_3)$	$(\Omega_1 \bullet \Omega_2, \Omega_1 \bullet \Omega_3, \Omega_2 \bullet \Omega_3)$
9	-0.8585629984736443E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.7144892316679474E-16, 0.3469446951953614E-17, 0.5204170427930421E-17)
	0.2910383045673370E-10	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.1734723475976807E-17, -0.2775557561562891E-16, -0.2775557561562891E-16)
10	0.7712515071034431E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.7031051088568496E-16, 0.1821459649775647E-16, 0.3122502256758253E-16)
	0.6839400157332420E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.7031051088568496E-16, 0.1821459649775647E-16, 0.3122502256758253E-16)
11	0.8620304470241535E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999998E+00)	(0.7892991815694472E-16, 0.1387778780781446E-16, -0.4163336342344337E-16)
	0.9295035852119327E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999998E+00)	(0.1118896642005041E-15, -0.2862293735361732E-16, 0.0000000000000000E+00)
12	0.1107650859921705E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999997E+00)	(0.9475926987523309E-16, 0.0000000000000000E+00, -0.6678685382510707E-16)
	0.1238731783814728E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999997E+00)	(0.1439820485060750E-15, -0.4683753385137379E-16, -0.1387778780781446E-16)
13	0.1368789526168257E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999998E+00)	(0.1158470021300761E-15, -0.1734723475976807E-17, -0.3469446951953614E-16)
	0.1047737896442413E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(0.1474514954580286E-15, -0.5551115123125783E-16, -0.1387778780781446E-16)

N	$A \cdot B - (A \cdot B)(A \cdot C)$	$(\Omega_1 \cdot \Omega_1, \Omega_2 \cdot \Omega_2, \Omega_3 \cdot \Omega_3)$	$(\Omega_1 \cdot \Omega_2, \Omega_1 \cdot \Omega_3, \Omega_2 \cdot \Omega_3)$
14	0.1722582965157926E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.1349730060790783E-15, -0.2602085213965211E-17, -0.2347297703431117E-16)
	0.1119587977882475E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.1457167719820518E-15, -0.8673617379884035E-17, 0.1387778780781446E-16)
15	0.9904397302307189E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(0.8077306185017008E-16, -0.1734723475976807E-17, -0.1994931997373328E-16)
	0.8521965355612338E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(0.1075528555105620E-15, -0.3989863994746656E-16, 0.0000000000000000E+00)
16	0.1538865035399795E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.1217016938614979E-15, 0.0000000000000000E+00, -0.1994931997373328E-16)
	0.8740244084037840E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.1139496483282265E-15, -0.433680868994201E-17. 0.0000000000000000E+00)
17	0.1317800979450112E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.1012881998326692E-15, 0.0000000000000000E+00, -0.3355605723842636E-16)
	0.7139533408917487E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.9334980705100193E-16, -0.7806255641895632E-17 0.2775557561562891E-16)
18	0.1547959982417524E-08	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.1218643241873707E-15, 0.1387778780781446E-16, -0.1734723475976807E-16)
	0.6775735528208315E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.9730714498057402E-16, -0.3469446951953614E-17, 0.2775557561562891E-16)

N	$A \cdot B - (A \cdot B)(A \cdot C)$	$(\Omega_1 \cdot \Omega_1, \Omega_2 \cdot \Omega_2, \Omega_3 \cdot \Omega_3)$	$(\Omega_1 \cdot \Omega_2, \Omega_1 \cdot \Omega_3, \Omega_2 \cdot \Omega_3)$
19	-0.600266503170132E-10	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999997E+00)	(-0.3523657060577889E-17, 0.0000000000000000E+00, -0.8066464163292153E-16)
	0.2273736754432321E-10	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999998E+00)	(0.7155734338404329E-17, -0.3209238430557093E-16, 0.2775557561562891E-16)
20	-0.1246007741428912E-09	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.9022594954152807E-17, 0.8673617379884035E-18, 0.2905661822261152E-16)
	-0.1455191522836685E-10	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.6938893903907228E-17, 0.1824159649775647E-16, -0.1110223024625157E-15)

Table 6d - (x*,y*) Errors for the N Passive Points for Case 2 of Table 1

N	Minimum Absolute (x*,y*) Errors	Maximum Absolute (x*,y*) Errors	Average Absolute (x*,y*) Errors
9	(0.1038079191232555E-02, 0.3421001602861651E-01)	(0.6286088913339825E+00, 0.3319790318711853E+00)	(0.1988452256251350E+00, 0.1630018856877540E+00)
	(0 4004454118234335E-01, 0.2038063291706749E-02)	(0.5867574788557519E+00, 0.6401350394102678E+00)	(0.2953271707577020E+00, 0.3279116899974094E+00)
10	(0.142606973954002E-01, 0.7479686750585302E-02)	(0.6348179006421155E+00, 0.3943399048628180E+00)	(0.1896503845544348E+00, 0.1558605910091799E+00)
	(0.4266083547422284E-01, 0.128994417434000 8E+00)	(0.6117987299097010E+00, 0.5052717611746402E+00)	(0.2074478320208840E+00, 0.2612802870283321E+00)
11	(0.3168155533040817E-01, 0.1087840377954308E-01)	(0.6036039894909067E+00, 0.4564413754530960E+00)	(0.1969811681689480E+00, 0.1699122476516263E+00)
	(0.2218810710672869E-02, 0.1320762049511117E+00)	(0.5833349970152568E+00, 0.7008256650830980E+00)	(0.2497072331409507E+00, 0.4606403399499086E+00)
12	(0.8703168007107109E-01, 0.1715331685218047E-01)	(0.6647641072003374E+00, 0.5999703232546949E+00)	(0.2284077747038156E+00, 0.2254253012744923E+00)
	(0.1949413329464988E-01, 0.6195382484236234E-02)	(0.5847551863258964E+00, 0.6340899260071318E+00)	(0.2351912468785888E+00, 0.4023850983885922E+00)
13	(0.1125627172183385E-01, 0.6516158945186135E-01)	(0.7009385882840853E+00, 0.6265387821669606E+00)	(0.2279221789615354E+00, 0.2593183807604156E+00)
	(0.3391689553030375E-02, 0.1260581112589421E+00)	(0.9005402035541650E+00, 0.1683592363485577E+01)	(0.4236301763954451E+00, 0.6681585978416672E+00)
14	(0.3910154447012459E-01, 0.4195389170502040E-01)	(0.7799874069063542E+00, 0.9028291653651763E+00)	(0.2627670028284957E+00, 0.3252735197775772E+00)
	(0.1983170181034666E-02, 0.3406228817334522E-01)	(0.9301719801702575E+00, 0.1068638235315674E+01)	(0.4434703240952873E+00, 0.4509735825433382E+00)

N	Minimum Absolute (x*,y*) Errors	Maximum Absolute (x*,y*) Errors	Average Absolute (x*,y*) Errors
15	(0.1400675724039502E-01, 0.7858171525469970E-02)	(0.7770885213396177E+00, 0.7770885213396177E+00)	(0.2370576298806988E+00, 0.2370576298806988E+00)
	(0.2945519241915662E-02, 0.2405663196140040E-01)	(0.8961163564725325E+00, 0.9227335623281689E+00)	(0.4123319837272236E+00, 0.4288677967641613E+00)
16	(0.1179340654493899E-01, 0.2712856343981684E-01)	(0.8163156574246671E+00, 0.9807563005031739E+00)	(0.1450540659222931E+00, 0.3086507519686204E+00)
	(0.3381706009427887E-02, 0.4193585682804724E-01)	(0.1389745237915463E+01, 0.1034409506377720E+01)	(0.5075342889959653E+00, 0.3688913434924501E+00)
17	(0.3153436072683036E-02, 0.1904853643484827E-01)	(0.8139733984387156E+00, 0.1110531112214971E+01)	(0.2727015208439005E+00, 0.3514757222521110E+00)
	(0.3551300893057885E-02, 0.1439539767503106E-02)	(0.1369161340419225E+01, 0.1079599446537085E+01)	(0.5339067634679606E+00, 0.3602941310019461E+00)
18	(0.1968116051635249E-01, 0.3427998813203459E-01)	(0.7672689029205984E+00, 0.1014714193430393E+01)	(0.2696167675028717E+00, 0.3317371256639474E+00)
	(0.3988533157548346E-01, 0.1895472304462 5E-01)	(0.2125394047009365E+01, 0.1498513398216790E+01)	(0.7229810427237129E+00, 0.4492775092449117E+00)
19	(0.4407681433132637E-02, 0.1722447797864746E-01)	(0.7165729248358552E+00, 0.9712149091995457E+00)	(0.3143791743059311E+00, 0.3057247903152630E+00)
	(0.9044332329516536E-01, 0.1860912966649053E+00)	(0.3894934530589808E+01, 0.1836819101311974E+01)	(0.1552322710254208E+01, 0.1071198554251172E+01)
20	(0.1734492731146275E-01, 0.1670379557559798E-01)	(0.7897223208551125E+00, 0.1075778664840243E+01)	(0.3125503966216138E+00, 0.3212171096356846E+00)
	(0.1036257113829357E+00, 0.3534209047049220E+01)	(0.3073858433234295E+02, 0.3044195694314681E+02)	(0.1223980530239721E+02, 0.1228249477206069E+02)

Table 6e - (x*,y*) Errors for the Remaining 20-N Passive Points for Case 2 of Table 1

N	Minimum Absolute (x*,y*) Errors	Maximum Absolute (x*,y*) Errors	Average Absolute (x*,y*) Errors
9	(0.5282556432024421E-01, 0.3340408288357821E-01)	(0.9966655872960787E+00, 0.2020797885577174E+01)	(0.5154694235507281E+00, 0.1106040111612314E+01)
	(0.4220995096550268E+00, 0.1047214809718753E+00)	(0.5807964057028755E+01, 0.5496809172280610E+01)	(0.2232862978218517E+01, 0.1969698969950074E+01)
10	(0.1259569168746566E-01, 0.3099382210794879E+00)	(0.1874081794293772E+01, 0.1800636387014592E+01)	(0.6215014168517147E+00, 0.1018194310364604E+01)
	(0.4266083547422284E-01, 0.1289944174340008E+00)	(0.4047674865793454E+01, 0.2988911624906947E+01)	(0.8593811553287612E+00, 0.6577629494429155E+00)
11	(0.7454671306724592E-01, 0.2492016067715603E+00)	(0.2202029543466551E+01, 0.1317093137124459E+01)	(0.7231878605025119E+00, 0.6863573429537801E+00)
	(0.1572158049338555E+00, 0.6856407149729193E-01)	(0.5980939862138086E+01, 0.2394615386539542E+01)	(0.2594549370095079E+01, 0.6408494729278603E+00)
12	(0.6013709916538801E-01, 0.1054753385244567E-01)	(0.1746247173033282E+01, 0.6894886911502738E+00)	(0.6048102344756168E+00, 0.3845130428928343E+00)
	(0.1397090360529347E+00, 0.3775596414613691E-01)	(0.5563564844721043E+01, 0.2336096571546342E+01)	(0.2631154587826801E+01, 0.6800800081143263E+00)
13	(0.6186478434244691E-02, 0.5655274604518112E-01)	(0.1430799759857393E+01, 0.8257604375312333E+00)	(0.5160479863671483E+00, 0.4299245302574326E+00)
	(0.3585646198427384E+00, 0.3017988962727145E+00)	(0.7324551554460925E+01, 0.3644896000514937E+01)	(0.4059119413844785E+01, 0.1363293799701536E+01)
14	(0.6689807256793756E-01, 0.1252561348318650E+00)	(0.9271112133815791E+00, 0.9254983955169820E+00)	(0.4526017591125034E+00, 0.3612994134860041E+00)
	(0.1546297899111906E+00, 0.3417990243966464E+00)	(0.5091554914738936E+01, 0.2790039445228032E+01)	(0.2803109517727950E+01, 0.9388820588853815E+00)

N	Minimum Absolute (x*,y*) Errors	Maximum Absolute (x*,y*) Errors	Average Absolute (x*,y*) Errors
15	(0.1723840661327358E-01, 0.1268572191589072E+00)	(0.1134018792907682E+01, 0.8787144552759649E+00)	(0.4690016354588863E+00, 0.3861179197654792E+00)
	(0.8141962549360038E+00, 0.3889946809692653E+00)	(0.4999756103855418E+01, 0.2734790502843769E+01)	(0.3250154340730955E+01, 0.1033884451179438E+01)
16	(0.2240778425190229E+00, 0.2901047178972682E+00)	(0.9535933863699597E+00, 0.9180769036191201E+00)	(0.5596209123336999E+00, 0.5388656649735770E+00)
	(0.7264496190217358E+00, 0.2718974611050733E+00,	(0.4351397212936376E+01, 0.2471482293558175E+01)	(0.3040634821862533E+01, 0.9870770148983627E+00)
17	(0.2168888826254829E+00, 0.3895626877553013E+00)	(0.9350045956254007E+00, 0.7662439050780918E+00)	(0.5465981160225131E+00, 0.5211065464114242E+00)
	(0.2773971163822296E+01, 0.5148290121510399E+00)	(0.4464570744278660E+01, 0.2466519948500125E+01)	(0.3885744252140322E+01, 0.1221589760173595E+01)
18	(0.2964938036394074E+00, 0.1832561966586184E+00)	(0.8520482700070033E+00, 0.4560667403683212E+00)	(0.5742710368232053E+00, 0.3196614685134698E+00)
	(0.3794674134965931E+01, 0.1056660310515113E+01)	(0.3864864669648114E+01, 0.3087335524442835E+01)	(0.3829769402307022E+01, 0.2071997917478974E+01)
19	(0.6782588973939028E+00, 0.3378244995046593E+00)	(0.6782588973939028E+00, 0.3378244995046593E+00)	(0.6782588973939028E+00, 0.3378244995046593E+00)
	(0.4281334898344710E+01, 0.4623775320636465E+01)	(0.4281334898344710E+01, 0.4623775320636465E+01)	(0.4281334898344710E+01, 0.4623775320636465E+01)

Table 6f - (x_W,y_W,z_W) Errors for the 27 Active Points for Case 2 of Table 1

N	Minimum Absolute (x_W,y_W,z_W) Errors	Maximum Absolute (x_W,y_W,z_W) Errors	Average Absolute (x_W,y_W,z_W) Errors
9	(0.5228363717302287E-04, 0.2525865629205981E-03, 0.1583443346164559E-02)	(0.1067265051264865E-01, 0.2655860448160285E-01, 0.6508124736993626E-01)	(0.4494689455145682E-02, 0.8162101918646364E-02, 0.2917917735282410E-01)
	(0.6936369465173620E-03, 0.8036503227517588E-03, 0.6451612350689917E-02)	(0.9278913672760258E-01, 0.3224102356851688E-01, 0.1054167826351156E+00)	(0.1453622544691902E-01, 0.1125475595143662E-01, 0.3613491417826466E-01)
10	(0.2346477548038732E-05, 0.5012882341597080E-04, 0.2647193748382293E-02)	(0.1050898205916595E-01, 0.2520480359737820E-01, 0.6503140021477516E-01)	(0.4220990968468458E-02, 0.7512093478030253E-02, 0.2861986445991471E-01)
	(0.4118862442681337E-03, 0.6373705499185256E-04, 0.3404028067222065E-03)	(0.4668859447355556E-01, 0.1179277615998053E-01, 0.8747101743851671E-01)	(0.1008546458759866E-01, 0.6053817546429567E-02, 0.3042241657329165E-01)
11	(0.2636014605761350E-03, 0.5031246710414106E-03, 0.9764845105020292E-04)	(0.1600877723727570E-01, 0.2501859971420184E-01, 0.7107960384743528E-01)	(0.7299459213881468E-02, 0.9878650792618139E-02, 0.2848856238192368E-01)
	(0.4397026888036670E-04, 0.1517584123988946E-04, 0.5240563091975625E-02)	(0.8136709931292296E-01, 0.2077872248783574E-01, 0.1169588201349105E+00)	(0.1681513505830325E-01, 0.7310922860314222E-02, 0.2659075735763622E-01)
12	(0.8249884816171615E-03, 0.1348999371328929E-03, 0.4499238450066390E-03)	(0.1609613725150205E-01, 0.1606953849869752E-01, 0.7200710118836584E-01)	(0.6902180051380193E-02, 0.7574206455910087E-02, 0.2897701380658698E-01)
	(0.1252252521846131E-03, 0.2270983074390553E-03, 0.4657992071165573E-02)	(0.7119110414295693E-01, 0.2022441211675963E-01, 0.1081043924320015E+00)	(0.1505812905607160E-01, 0.6819122300431086E-02, 0.3448942071110557E-01)
13	(0.3294533928477339E-03, 0.1731457039058082E-03, 0.2097448754635245E-02)	(0.1120107871262821E-01, 0.1570857498229361E-01, 0.6625327751645949E-01)	(0.5691570500080073E-02, 0.7384784435987552E-02, 0.2896964102130025E-01)
	(0.3706061140139050E-03, 0.8126824767917817E-03, 0.3606593654903456E-03)	(0.1210028247057522E+00, 0.3088696346821895E-01, 0.1414435805021321E+00)	(0.2174946999916754E-01, 0.1325770293265586E-01, 0.4161682685221580E-01)

Table 6f - (x_W,y_W,z_W) Errors for the 27 Active Points for Case 2 of Table 1

N	Minimum Absolute (x_W,y_W,z_W) Errors	Maximum Absolute (x_W,y_W,z_W) Errors	Average Absolute (x_W,y_W,z_W) Errors
9	(0.5228363717302287E-04, 0.2525865629205981E-03, 0.1583443346164559E-02)	(0.1067265051264865E-01, 0.2655860448160285E-01, 0.6508124736993626E-01)	(0.4494689455145682E-02, 0.8162101918646364E-02, 0.2917917735282410E-01)
	(0.6936369465173620E-03, 0.8036503227517588E-03, 0.6451612350689917E-02)	(0.9278913672760258E-01, 0.3224102356851688E-01, 0.1054167826351156E+00)	(0.1453622544691902E-01, 0.1125475595143662E-01, 0.3613491417826466E-01)
10	(0.2346477548038732E-05, 0.5012882341597080E-04, 0.2647193748382293E-02)	(0.1050898205916595E-01, 0.2520480359737820E-01, 0.6503140021477516E-01)	(0.4220990968468458E-02, 0.7512093478030253E-02, 0.2861986445991471E-01)
	(0.4118862442681337E-03, 0.6373705499185256E-04, 0.3404028067222065E-03)	(0.4668859447355556E-01, 0.1179277615998053E-01, 0.8747101743851671E-01)	(0.1008546458759866E-01, 0.6053817546429567E-02, 0.3042241657329165E-01)
11	(0.2636014605761350E-03, 0.5031246710414106E-03, 0.9764845105020292E-04)	(0.1600877723727570E-01, 0.2501859971420184E-01, 0.7107960384743528E-01)	(0.7299459213881468E-02, 0.9878650792618139E-02, 0.2848856238192368E-01)
	(0.4397026888036670E-04, 0.1517584123988946E-04, 0.5240563091975625E-02)	(0.8136709931292296E-01, 0.2077872248783574E-01, 0.1169588201349105E+00)	(0.1681513505830325E-01, 0.7310922860314222E-02, 0.2659075735763622E-01)
12	(0.8249884816171615E-03, 0.1348999371328929E-03, 0.4499238450066390E-03)	(0.1609613725150205E-01, 0.1606953849869752E-01, 0.7200710118836584E-01)	(0.6902180051380193E-02, 0.7574206455910087E-02, 0.2897701380658698E-01)
	(0.1252252521846131E-03, 0.2270983074390553E-03, 0.4657992071165573E-02)	(0.7119110414295693E-01, 0.2022441211675963E-01, 0.1081043924320015E+00)	(0.1505812905607160E-01, 0.6819122300431086E-02, 0.3448942071110557E-01)
13	(0.3294533928477339E-03, 0.1731457039058082E-03, 0.2097448754635245E-02)	(0.1120107871262821E-01, 0.1570857498229361E-01, 0.6625327751645949E-01)	(0.5691570500080073E-02, 0.7384784435987552E-02, 0.2896964102130025E-01)
	(0.3706061140139050E-03, 0.8126824767917817E-03, 0.3606593654903456E-03)	(0.1210028247057522E+00, 0.3088696346821895E-01, 0.1414435805021321E+00)	(0.2174946999916754E-01, 0.1325770293265586E-01, 0.4161682685221580E-01)

N	Minimum Absolute (x_W,y_W,z_W) Errors	Maximum Absolute (x_W,y_W,z_W) Errors	Average Absolute (x_W,y_W,z_W) Errors
14	(0.7191167663152254E-03, 0.5159123325475523E-03, 0.3288484471064779E-02)	(0.1453412581926594E-01, 0.1667273206356509E-01, 0.6266723184920364E-01)	0.5760915485557416E-02, 0.6630674213552699E-02, 0.2950784713382639E-01)
	(0.3210351399041134E-03, 0.8639457931869376E-03, 0.3305708788392092E-02)	(0.5859768405182542E-01, 0.2385729286837579E-01, 0.9619234988853065E-01)	(0.1435799764294138E-01, 0.8847446588805029E-02, 0.3105438663906428E-01)
15	(0.1533217182039248E-03, 0.9383909404747648E-03, 0.4427422041833662E-02)	(0.1303118506073231E-01, 0.1603554119002107E-01, 0.6376879631831756E-01)	(0.5762562698864951E-02, 0.6560221653082244E-02, 0.2937916962760228E-01)
	(0.7604744248665973E-03, 0.1179835579440791E-02, 0.2986440729707995E-02)	(0.5594711177407707E-01, 0.2281192964365752E-01, 0.9447615101626816E-01)	(0.1385280015431861E-01, 0.8260579577557742E-02, 0.3055142515401643E-01)
16	(0.5330730620998203E-03, 0.8653913757528109E-04, 0.5401539428952606E-02)	(0.1626917704397112E-01, 0.1725583138682950E-01, 0.6284294723628828E-01)	(0.5902800493182579E-02, 0.6373972296784429E-02, 0.2969017998872732E-01)
	(0.4911425285027970E-03, 0.6103844265163694E-03, 0.1751577797408421E-02)	(0.3335097441051471E-01, 0.2482347152071131E-01, 0.8046032551797211E-01)	(0.1328042680098461E-01, 0.8046237346116551E-02, 0.2836235034959828E-01)
17	(0.6421475504003737E-04, 0.3611024789607065E-04, 0.5831040525228648E-02)	(0.1790976227874852E-01, 0.1788864716743710E-01, 0.6244366467310769E-01)	(0.6375180863216485E-02, 0.7227373345693220E-02, 0.2968734025613637E-01)
	(0.9699768972946110E-03, 0.3994479546713114E-03, 0.7295528268567164E-03)	(0.3283736080290045E-01, 0.2553266755488193E-01, 0.8262889507751825E-01)	(0.1393149327321614E-01, 0.8271819321469888E-02, 0.2818034149426740E-01)
18	(0.1509801688654999E-03, 0.2767465122606172E-03, 0.3084724717275833E-02)	(0.1471806314261248E-01, 0.1629752095986747E-01, 0.6370859719327693E-01)	(0.5616003467860299E-02, 0.7062799347949174E-02, 0.2913842381232204E-01)
	(0.1606398662111319E-02, 0.1081606124517798E-03, 0.9847731130185888E-03)	(0.4708181027149183E-01, 0.2751604775472560E-01, 0.9333482617503552E-01)	(0.1614174941344779E-01, 0.9282919194190610E-02, 0.3152895573057644E-01)

N	Minimum Absolute (x_W,y_W,z_W) Errors	Maximum Absolute (x_W,y_W,z_W) Errors	Average Absolute (x_W,y_W,z_W) Errors
19	(0.4152474701646369E-05, 0.7676524643747129E-03, 0.3299738634821026E-04)	(0.1447119556894294E-01, 0.1545364193443399E-01, 0.6574244136237373E-01)	(0.5518375592831470E-02, 0.6710339709454375E-02, 0.2876912206904406E-01)
	(0.6321479742432690E-03, 0.3890225595325703E-02, 0.6744919104955116E-03)	(0.9433104724536734E-01, 0.4825932115428144E-01, 0.1210677435848013E+00)	(0.3064600567645777E-01, 0.2091268912586699E-01, 0.4988088806478419E-01)
20	(0.3348569448307082E-03, 0.4316493451910208E-03, 0.6057567760624494E-03)	(0.1573608847422125E-01, 0.1731899954361937E-01, 0.6383076296747237E-01)	(0.5322120547290273E-02, 0.7273140664527009E-02, 0.2912695885603052E-01)
	(0.3599787329378046E-04, 0.4773479688596538E-02, 0.4152591398051353E-02)	(0.1059297954398797E+01, 0.7707613894433230E+00, 0.1207167525062547E+01)	(0.3226197270038626E+00, 0.2515454151239359E+00, 0.3816839398469778E+00)

Table 7a - Intrinsic and Some Extrinsic Parameter Values for Ganapathy's Method

N	(D_1,D_2,D_3)	(x_O,y_O)	(p_xF,p_yF)
9	(-0.1831585-86141248E+01, 0.9596208354345290E+01, 0.768266917922032E+02)	(0.1811729635180630E+03, 0.5518129888844340E+03)	(0.3090833820861565E+04, 0.3753658262863506E+04)
10	(-0.4464733295194675E+01, 0.9123046769864648E+01, 0.7626684208471991E+02)	(0.7456805801897887E+02, 0.5289013943898259E+03)	(0.3084615320909994E+04, 0.3759059993627350E+04)
11	(-0.5308087859627368E+01, 0.6730325185095466E+01, 0.7547727515821302E+02)	(0.4104864770755466E+02, 0.4112183481442833E+03)	(0.3034429676230926E+04, 0.3728405543471993EE+04)
12	(-0.3676193483904240E+01, 0.6592161363107329E+01, 0.7317454254905257E+02)	(0.1064471597602367E+03, 0.4040610643783836E+03)	(0.2946392929647991E+04, 0.3608228298802135E+04)
13	(-0.2877743233659796E+01, 0.7326270786045625E+01, 0.7340521554601636E+02)	(0.1386822644559387E+03, 0.4401817804977597E+03)	(0.2959971193544175E+04, 0.3617656619134762E+04)
14	(-0.2354735253100529E+01, 0.7152969942175281E+01, 0.7361646307644597E+02)	(0.1597971145181846E+03, 0.4315535734507180E+03)	(0.2970488686863565E+04, 0.3627389064966739E+04)
15	(-0.2589635553503231E+01, 0.6912293033137593E+01, 0.7354863886422616E+02)	(0.1502576650965697E+03, 0.4197118837738602E+03)	(0.2966788498943270E+04, 0.3624123125984675E+04)
16	(-0.3314983053413600E+01, 0.6741798955228780E+01, 0.6425879287285311E+02)	(0.1209943890665230E+03, 0.4112946613107173E+03)	(0.2996495249740732E+04, 0.3659844274120903E+04)
17	(-0.3461786566080479E+01, 0.6012364590547883E+01, 0.7391687504320398E+02)	(0.1150767433432388E+03, 0.3753281576166727E+03)	(0.2980600156247040E+04, 0.3642727540758583E+04)
18	(-0.4510819083218941E+01, 0.8063668294055411E+01, 0.7575528100290326E+02)	(0.7267736215616058E+02, 0.4763169345148212E+03)	(0.3060973827797450E+04, 0.3734130312614066E+04)
19	(-0.5277656417365936E+01, 0.1059218512191298E+02, 0.775373115789425E+02)	(0.4158882006779235E+02, 0.6003682930477088E+03)	(0.3139004768081942E+04, 0.3819540386883299E+04)
20	(-0.5059738370646798E+01, 0.9726934481791718E+01, 0.770220515152870 7E+02)	(0.5043220502131413E+02, 0.5579805143577481E+03)	(0.3116531860286946E+04, 0.3795319895638180E+04)

Table 7b - Some More Extrinsic Parameter Values for Ganapathy's Method

N	Ω_1	Ω_2	Ω_3
9	(0.9997643526730505E+00 -0.1574170675975445E-01, -0.1494783571378382E-01)	(0.1317032875061366E-01, 0.9996459273343273E+00, -0.2186049151989079E-01)	(0.1528666499076834E-01, 0.2162857720814593E-01, 0.9996491997302910E+00)
10	(0.9986794183801167E+00, -0.1662672307678440E-01, -0.4861040406826467E-01)	(0.1661581274213621E-01, 0.9994737754253607E+00, -0.2785833813948826E-01)	(0.4904806033989164E-01, 0.2701387145494248E-01, 0.9984310384427710E+00)
11	(0.9984834262753971E+00, -0.1714535625281632E-01, -0.5889712652028585E-01)	(0.1465811234218936E-01, 0.9982065689810615E+00, -0.5804123866373843E-01)	(0.8673705590812407E-01, 0.5655926684762931E-01, 0.9946244177909818E+00)
12	(0.9991475540306655E+00, -0.1619124339067928E-01, -0.3797379243639036E-01)	(0.1191033750601709E-01, 0.9982491068815458E+00, -0.5793845416205835E-01)	(0.3884548262551291E-01, 0.5743690624969224E-01, 0.9975931186009931E+00)
13	(0.9994977505163902E+00, -0.1604573279131959E-01, -0.2732729719283554E-01)	(0.1218848448644401E-01, 0.9987576276975204E+00, -0.4831813284828990E-01)	(0.2806873570289000E-01, 0.4796093940552233E-01, 0.9984547532899931E+00)
14	(0.9996621354376912E+00, -0.1575302860718448E-01, -0.2067503474861387E-01)	(0.1131919142741340E-01, 0.9986627286362480E+00, -0.5044432909885147E-01)	(0.2144215959905586E-01, 0.5019354671290914E-01, 0.9985093097513451E+00)
15	(0.9995923461817180E+00, -0.1567001120171107E-01, -0.2386613089436081E-01)	(0.1135939512862954E-01, 0.9984968521746922E+00, -0.5361902964006970E-01)	(0.2467057927548591E-01, 0.5332630872160107E-01, 0.9982723412558020E+00)
16	(0.9992959191910402E+00, -0.1629665117143274E-01, -0.3379474883366791E-01)	(0.1136714114461474E-01, 0.9983629428223508E+00, -0.5605553051478357E-01)	(0.3465309997891840E-01, 0.5563216631025542E-01, 0.9978498006881994E+00)
17	(0.9992124618761766E+00, -0.1664943978054988E-01, -0.3601738727814982E-01)	(0.1102173638972769E-01, 0.9977681660584518E+00, -0.6585748345715389E-01)	(0.3703368550693294E-01, 0.6540898464459236E-01, 0.9971710840500418E+00)
18	(0.9986037904837992E+00, -0.1707842060077204E-01, -0.4998797036459631E-01)	(0.1287237499261218E-01, 0.9991222037315357E+00, -0.3986381783885198E-01)	(0.5062502649513323E-01, 0.3916479188728599E-01, 0.9979495106310703E+00)
19	(0.9980807390928246E+00, -0.1676143915097316E-01, -0.5961453186521927E-01)	(0.1509196920120242E-01, 0.9998468568729034E+00, -0.8859755470991207E-02)	(0.5975394529945442E-01, 0.7943056025071702E-02, 0.9981815335309166E+00)
20	(0.9982304889691112E+00, -0.1698622305784685E-01, -0.5698560448673118E-01)	(0.1432945200336865E-01, 0.9997091413607912E+00, -0.1939844027113550E-01)	(0.5729860674042830E-01, 0.1854756494280470E-01, 0.9981847812405769E+00)

Table 7c - Yet More Extrinsic Parameter Values for Ganapathy's Method

N	($\Omega_1 \bullet \Omega_1$, $\Omega_2 \bullet \Omega_2$, $\Omega_3 \bullet \Omega_3$)	($\Omega_1 \bullet \Omega_2$, $\Omega_1 \bullet \Omega_3$, $\Omega_2 \bullet \Omega_3$)
9	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.2426121126232542E-03, 0.9367506770274758E-16, -0.5984795992119984E-16)
10	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(0.1330101592354716E-02, 0.2478052485432869E-14, 0.2654126918244515E-15)
11	(0.1000731987300039E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(0.9397371720494538E-03, 0.2705526381388256E-01, 0.1353084311261910E-15)
12	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.2062566834795333E-02, 0.7979727989493313E-16, -0.1396452398161330E-15)
13	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(-0.2523031214773827E-02, 0.1029558382992235E-14, 0.3747002708109903E-15)
14	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.337365720239037E-02, -0.5030698080332741E-16, -0.1925543958334256E-15)
15	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.3012013650800165E-02, 0.6487865800153259E-15, 0.2983724378680108E-15)
16	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.3016452288493704E-02, -0.3720981855970251E-15, -0.2133709875451473E-15)
17	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.9999999999999999E+00)	(-0.3227210156769186E-02, 0.6305719835175694E-15, 0.2081668171172169E-15)
18	(0.1000000000000000E+01, 0.1000000000000000E+01, 0.1000000000000000E+01)	(-0.2216315421997832E-02, 0.7008282842946301E-15, -0.3729655473350135E-16)

N	($\Omega_1 \bullet \Omega_1$, $\Omega_2 \bullet \Omega_2$, $\Omega_3 \bullet \Omega_3$)	($\Omega_1 \bullet \Omega_2$, $\Omega_1 \bullet \Omega_3$, $\Omega_2 \bullet \Omega_3$)
19	(0.100000000000000E+01, 0.100000000000000E+01, 0.100000000000000E+01)	(-0.1167698302221234E-02, 0.2411265631607762E-14, 0.8760353553682876E-16)
20	(0.9999999999999999E+00, 0.100000000000000E+01, 0.100000000000000E+01)	(-0.1571754743190445E-02, -0.1627170620466245E-14, -0.1196959198423997E-15)

Table 8a - Errors in the Least-Squares Approximation for Case 1 of Table 1

N	Least-Squares Error	Residual Error
9	0.3256137891011078E+01	0.2219380229917350E+00
10	0.1760779203087727E+01	0.1645271290708509E+00
11	0.6867349001990916E+00	0.9827664698470877E-01
12	0.8495070554953050E+00	0.1045941302530123E+00
13	0.3068583929983710E+00	0.6056639318606238E-01
14	0.3422768827397124E+00	0.6401691251834318E-01
15	0.3210289220709981E+00	0.6232715459810040E-01
16	0.3572251985032027E+00	0.6572776505502603E-01
17	0.3898026267980995E+00	0.6599538999703393E-01
18	0.3953391871408966E+00	0.6287523862451821E-01
19	0.3351267129082863E+00	0.5525814830772850E-01
20	0.2273352971421487E+00	0.4492611905999137E-01

Table 8b - Errors in the Least-Squares Approximation for Case 2 of Table 1

N	Least-Squares Error	Residual Error
9	0.2035649800065858E+02	0.2074536845554875E+00
	0.5181288602131380E+03	0.2074536845554875E+00
10	0.1029799735808617E+02	0.1424001207269998E+00
	0.2622380220245377E+03	0.1424001207269998E+00
11	0.3288178965760172E+01	0.7695395008157505E-01
	0.8267112161138230E+02	0.7695395008157507E-01
12	0.3932107918472083E+01	0.7819684518194255E-01
	0.9634121946838493E+02	0.7819684518194259E-01
13	0.1700057428554727E+01	0.4980872688669886E-01
	0.4153868426107 9E+02	0.4980872688669888E-01
14	0.1976076903826657E+01	0.5113885566063096E-01
	0.4884834456434653E+02	0.5113885566063099E-01
15	0.1590895826026644E+01	0.4638285897055979E-01
	0.3914920786481321E+02	0.4638285897055981E-01
16	0.1838059986498788E+01	0.4864636493422057E-01
	0.4589131399723282E+02	0.4864636493422058E-01
17	0.1827400160524136E+01	0.4885077508650895E-01
	0.4535567823944662E+02	0.4835077508650895E-01
18	0.1859264126361809E+01	0.4829718622007676E-01
	0.4740237464671525E+02	0.4829718622007678E-01

N	Least-Squares Error	Residual Error
19	0.1778277570756400E+01	0.4577825252561965E-01
	0.4669709808487448E+02	0.4577825252561966E-01
20	0.1567152302091170E+01	0.4080616142991288E-01
	0.4133051064025834E+02	0.4080616142991288E-01

SURVIVAL ANALYSIS OF RADIATED ANIMALS FOR SMALL SAMPLE SIZES

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

Prepared by	Ramesh C. Gupta, Principal Investigator
Academic rank	Professor of Mathematics
Department and	Department of Mathematics
University	University of Maine, Orono, Maine 04469
Date	December 18, 1987

Contract no. F49620-85-C-0013/SB5851-0360

Acknowledgements

I would like to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsoring this research. My special thanks are due to Dr. Richard A. Albanese who was immensly helpful throughout this investigation in discussions and providing the Air Force survival data.

I would also like to thank Dr. Bryce Hartman and Universal Energy Systems for providing me an opportunity to work on this project.

Finally I would like to thank my wife Dr. Pushpa Gupta and my children for their support and encouragement throughout this investigation.

SURVIVAL ANALYSIS OF RADIATED ANIMALS FOR SMALL SAMPLE SIZES

by

Ramesh C. Gupta
Department of Mathematics
University of Maine
Orono, Maine 04469

ABSTRACT

The relative risk is an important parameter in certain epidemiological studies. It is given by the ratio of the rates of attack of a certain disease between the exposed and the control group. This study deals with the tests of hypothesis regarding the relative risks when the survival data are available in the contingency table form. Small sample uniformly most power unbiased tests are derived to (i) test the hypothesis that the relative risk at a particular time is 1 and (ii) to test the hypothesis that the relative risks at two time points are the same. To illustrate our tests, the data obtained from USAFSAM are analyzed. Finally a discussion of the analysis is provided and some further directions of research are pointed out.

1. Introduction

In the Radiation Sciences Division of the USAFSAM the research on the survival analyses of animals exposed to certain levels of radiation is important. Our methods described in a report submitted to AFOSR, are applicable to such a study when one is interested in comparing the rates of recovery of two sets of animals under different levels of radiation. The results of our study suggest that one can estimate the relative risk with the same precision when the data on fewer number of animals is available. Since a study with fewer number of animals involves considerable amount of savings in the cost, a direction in which it was important to follow this research in the Radiation Sciences was to study the comparison of survival curves of these animals for small sample sizes. Most of the results available in the literature are valid for large samples and in our case the data on a large number of animals is not available.

As described in Yochmowitz, Wood and Salmon (1985), in 1964 the USAFSAM and the National Aeronautics and Space Administration initiated a series of studies on the acute effects of protons on rhesus monkeys. The subjects were exposed to single acute whole body doses of mono-energetic 32-, 55-, 138- or 2300- Mev protons representative of the proton spectrum in space. These authors showed that exposure to protons enhanced chronic mortality significantly compared to that of control animals; that the onset and cause of death as well as the shortening of life expectency of the irradiated animals were influenced by the proton energy, dose and sex of the subjects; and that increased mortality in the irradiated animals over the

controls was due largely to development of fatal neoplasms and endometriosis. The data was analyzed by the χ^2 test to detect the differences in mortalities between irradiated and control. The χ^2 test is valid when large numbers of subjects are available. For details see their paper.

In Yochmowitz et. al.'s study, the data required was the actual time of death for different animals. While at USAFSAM I observed that the real situation is as described below.

The Available Data

The experiment starts with a certain number of animals (say N_1) who are irradiated together with a control group of N_2 animals. So the initial table looks like

	Exposed	Non-exposed	Total
Dead	0	0	0
Alive	N_1	N_2	$N_1 + N_2$
Total	N_1	N_2	$N_1 + N_2$

After a time t_1, the experimenter records the following table in his log book.

	Exposed	Non-exposed	Total
Dead	D_{11}	D_{21}	$D_{11} + D_{21}$
Alive	$N_1 - D_{11}$	$N_2 - D_{21}$	$(N_1+N_2) - (D_{11}+D_{21})$
Total	N_1	N_2	$N_1 + N_2$

At time t_2, $t_2 > t_1$, the experimenter records the following

	Exposed	Non-exposed	Total
Dead	D_{12}	D_{22}	$D_{12} + D_{22}$
Alive	$N_1 - D_{12}$	$N_2 - D_{22}$	$(N_1+N_2) - (D_{12}+D_{22})$
Total	N_1	N_2	$N_1 + N_2$

He continues recording the tables like this at times

$0 = t_0 < t_1 < t_2 \ldots < t_k$. We shall call this as the cumulative data.

The data can also be presented as follows:

Let us denote the $(i + 1)^{th}$ interval by $[t_i, t_{i+1})$.

Let $N_{gi}(g = 1, 2)$ denote the number of patients in group g at time t_i.

Let d_{gi} denote the number of deaths among N_{gi} patients in $[t_i, t_{i+1})$. The initial table looks like

	Exposed	Non-exposed	
Dead	0	0	
Alive	$N_1 = N_{10}$	$N_2 = N_{20}$	
Total	N_1	N_2	$N_1 + N_2$

At the end of time t_1, the data is

	Exposed	Non-exposed	
Dead	$D_{11} = d_{10}$	$D_{21} = d_{20}$	
Alive	$N_1-D_{11}=N_{10}-d_{10}=N_{11}$	$N_2-D_{21}=N_{20}-d_{20}=N_{21}$	
Total	N_{10}	N_{20}	$N_{10} + N_{20}$

At the end of time t_2, the data is

	Exposed	Non-exposed	
Dead	$D_{12}-D_{11}=d_{11}$	$D_{22}-D_{21}=d_{21}$	
Alive	$N_{11} - d_{11}$	$N_{21} - d_{21}$	
Total	N_{11}	N_{21}	$N_{11} + N_{21}$

In general at the end of time t_{i+1}, the data is

	Exposed	Non-exposed	Total
Dead	d_{1i}	d_{2i}	$d_{1i} + d_{2i}$
Alive	$N_{1i} - d_{1i}$	$N_{2i} - d_{2i}$	$(N_{1i}+N_{2i})-(d_{1i}+d_{2i})$
Total	N_{1i}	N_{2i}	$N_{1i} + N_{2i}$

We shall call this as the Truncated data.

This study deals with the small sample analyses of such data sets. More specifically in section 2, we briefly describe some of the large sample methods viz Pearson's conditional χ^2 test, Mantel-Haenzel test and the log rank test to (i) test the hypothesis that the relative risk at a particular time is 1 and (ii) to test the hypothesis that the relative risk at all time points is the same. In section 3 we derive the small sample uniformly most powerful unbiased tests for the hypothesis described above. To illustrate our tests, the data obtained from USAFSAM are analyzed in section 4. Finally in section 5, a discussion of the analysis is provided and some further directions of research are pointed out.

2. Large sample methods

1. Pearson's conditional test

Considering the truncated data, the observed proportion of deaths in group g(g = 1, 2) during the period t_i to t_{i+1} is

$\hat{p}_{gi} = \frac{dg_i}{Ng_i}$. The overall proportion of deaths in two groups is

$$\hat{p}_i = \frac{d_{1i} + d_{2i}}{N_{1i} + N_{2i}} = \frac{N_{1i}\hat{p}_{1i} + N_{2i}\hat{p}_{2i}}{N_{1i} + N_{2i}}$$

Conditionally on N_{gi} the number of deaths, d_{gi}, has a binomial distribution. If the mortality probabilities p_{1i} and p_{2i} are the same under the null hypothesis, then the estimated expected number of deaths in group g is

$$E_{gi} = N_{gi}\hat{p}_i$$

and the expected number of survivors is

$$N_{gi} - E_{gi} = N_{gi}(1 - \hat{p}_i)$$

One can then form the Pearson's Chi square statistic with 1 degree of freedom.

$$\chi_i^2(1) = \frac{1}{1 - \hat{p}_i} \sum_{g=1}^{2} \frac{\Sigma(d_{gi} - E_{gi})^2}{E_{gi}}$$

to test for difference between p_{1i} and p_{2i}.

Conditional on the set of N_{gi}'s ($i = 0, 1, 2, \ldots, k - 1$) and supposing the null hypothesis $H_0: p_{1i} = p_{2i}$ for all i to be valid, it is reasonable to regard the $\chi_i^2(1)$ as mutually independent, so that

$$\chi^2(k) = \sum_{i=0}^{k-1} \chi_i^2(1)$$

is approximately chi-square with k degrees of freedom.

2. Mantel-Haenzel test

Again consider the situation as before. The conditional expected value (under the null hypothesis) is

$$E_{1i} = N_{1i}\hat{p}_i = N_{1i} \frac{d_{1i} + d_{2i}}{N_{1i} + N_{2i}}$$

The conditional variance of d_{1i} is

$$\text{Var}(d_{1i}) = V_{1i} = \frac{N_{1i}N_{2i}}{N_{1i} + N_{2i}-1} \hat{p}_i(1 - \hat{p}_i)$$

Under the null hypothesis

$$\chi^2_{1i} = \frac{(d_{1i} - E_{1i})^2}{V_{1i}}$$

is approximately chi-square distributed with 1 degree of freedom, provided neither N_{1i} or N_{2i} are too small.

In order to test the consistent differences between death probabilities in two groups, we test the hypothesis $H_0: p_{1i} = p_{2i}$ vs. $H_1: p_{1i} > p_{2i}$ (or $H_1: p_{1i} < p_{2i}$) for all i as follows:

$$\text{Let } T_1 = \sum_{i=0}^{k-1} (d_{1i} - E_{1i}) = \sum_{i=0}^{k-1} d_{1i} - \sum_{i=0}^{k-1} E_{1i}.$$

Conditionally on fixed values of N_{gi}'s and the $(d_{1i} + d_{2i})$'s, the variance of T_1 under H_0 is given by

$$\text{Var}(T_1) = \sum_{i=0}^{k-1} V_{1i} = \sum_{i=0}^{k-1} \frac{N_{1i}N_{2i}}{N_{1i} + N_{2i} - 1} \hat{p}_i(1 - \hat{p}_i).$$

Then $X_1^2 = \frac{T_1^2}{V_1}$ would be approximately chi-square distributed with 1 degree of freedom and can be used to test the consistent differences between the relative risks in the two groups. For details see Mantal and Haenzel (1950) and Elandt-Johnson and Johnson (1980).

3. Log rank test

The log rank statistic is given by

$$X'^2 = \frac{\left(d_1 - \sum_{i=0}^{k-1} E_{1i}\right)^2}{\sum_{i=0}^{k} E_{1i}} + \frac{\left(d_2 - \sum_{i=0}^{k-1} E_{2i}\right)^2}{\sum_{i=0}^{k-1} E_{2i}}$$

where d_1 and d_2 are the total number of deaths in the two groups. This can also be written as

$$X'^2 = \frac{\left(d - \sum_{i=0}^{k-1} E_{1i}\right)^2}{(1/d)\left(\sum_{i=0}^{k-1} E_{1i}\right)\left(d - \sum_{i=0}^{k-1} E_{1i}\right)}$$

where $d = d_1 + d_2$ is the total number of deaths.

X'^2 has approximately a chi-square distribution with 1 degree of freedom and can be used to test the consistent differences between the relative risks in the two groups.

Relatively, the log rank test will be more conservative than X_1^2 in establishing the difference between the two groups, see Peto and Peto (1972) and Elandt-Johnson and Johnson (1980).

3. Small sample tests

In this section we derive the small sample uniformly most powerful unbiased tests for testing (a) the hypotheses that the relative risk at a particular time is 1 and (b) the hypothesis that the relative risk at two time points is the same.

(a) Testing the relative risk at one point

Suppose the data at a particular time point can be represented as

	Exposed	Non-exposed
Dead	x	y
Alive	m - x	n - y
Total	m	n

Let p_1 be the probability of death in the exposed group and p_2 be the corresponding probability in the non-exposed group.

Let $\Delta = \frac{p_1 q_2}{p_2 q_1}$ be the odds ratio, where $q_i = 1 - p_i$ and $q_2 = 1 - p_2$. It is well known that, in the case of rare diseases, the odds ratio approximates the relative risk.

The likelihood of the data can be written as

$$L = \binom{m}{x} p_1^x q_1^{m-x} \binom{n}{y} p_2^y q_2^{n-y}$$

$$= \binom{m}{x}\binom{n}{y} q_1^m q_2^n \left(\frac{p_1}{q_1}\right)^x \left(\frac{p_2}{q_2}\right)^y$$

$$= \binom{m}{x}\binom{n}{y} q_1^m q_2^n \exp\left[x \ln(p_1/q_1) + y \ln(p_2/q_2)\right]$$

$$= \binom{m}{x}\binom{n}{y} q_1^m q_2^n \exp\left[x\left\{\ln(p_1/q_1) - \ln(p_2/q_2)\right\} + (x + y)\ln(p_2/q_2)\right]$$

$$= \binom{m}{x}\binom{n}{y} q_1^m q_2^n \exp\left[x \ln \Delta + (x + y)\ln(p_2/q_2)\right].$$

This belongs to the one parameter exponential family, and to test $H_0:\Delta = 1$ against $H_1:\Delta > 1$, a uniformly most powerful unbiased test is given by the rule:

Reject H_0 if $X > c$ where c is a constant depending on $x + y$ and is determined so that the conditional probability of rejection given $x + y = t$ is α (the level of significance), see Lehman (1986, page 145). Now the conditional probability is obtained as follows:

$$P(X = x \mid X + Y = t) = \frac{P(X = x,\ Y = t - x)}{P(X + Y = t)}$$

$$= \frac{\binom{m}{x} p_1^x q_1^{m-x} \binom{n}{t-x} p_2^{t-x} q_2^{n-t+x}}{\sum_x \binom{m}{x} p_1^x q_1^{m-x} \binom{n}{t-x} p_2^{t-x} q_2^{n-t+x}}$$

$$= \frac{\binom{m}{x}\binom{n}{t-x}\Delta^x}{\sum_{i=0}^{t} \binom{m}{i}\binom{n}{t-i}\Delta^i}$$

So under the null hypothesis, the conditional distribution of X is given by

$$P(X = x \mid X + Y = t,\ \Delta = 1) = \frac{\binom{m}{x}\binom{n}{t-x}}{\sum_{i=0}^{t} \binom{m}{i}\binom{n}{t-i}},\quad x = 0,\ 1,\ 2,\ \ldots,\ t.$$

(b) Comparing the relative risks at two time points.

Suppose the data at two time points can be represented as

Time point 1

	Exposed	Non-exposed
Dead	x_1	y_1
Alive	$m_1 - x_1$	$n_1 - y_1$
Total	m_1	n_1

Time point 2

	Exposed	Non-exposed
Dead	x_2	y_2
Alive	$m_2 - x_2$	$n_2 - y_2$
Total	m_2	n_2

Let p_{11} and P_{21} be the probabilities of death among the exposed and non-exposed, respectively, at time point 1. Likewise let p_{21} and p_{22} be the probabilities at time point 2. Let $\Delta_1 = p_{11}q_{21}/p_{21}q_{11}$ and $\Delta_2 = p_{12}q_{22}/p_{22}/q_{12}$ be the odds ratio at the two time points where $q_{11} = 1 - p_{11}$ etc. The likelihood of the whole data for the Truncated case can be written as

$$L = \prod_{k=1}^{2} \binom{m_k}{x_k} p_{1k}^{x_k} q_{1k}^{m_k - x_k} \binom{n_k}{y_k} p_{2k}^{y_k} q_{2k}^{n_k - y_k}$$

$$= \prod_{k=1}^{2} \binom{m_k}{x_k} \binom{n_k}{y_k} q_{1k}^{m_k} q_{2k}^{n_k} (p_{1k}/q_{1k})^{x_k} (p_{2k}/q_{2k})^{y_k}$$

$$= \prod_{k=1}^{2} \binom{m_k}{x_k} \binom{n_k}{y_k} q_{1k}^{m_k} q_{2k}^{n_k} \exp\left[\sum_{k=1}^{2} \left(x_k \ln(p_{1k}/q_{1k}) + y_k \ln(p_{2k}/q_{2k})\right)\right]$$

$$= \prod_{k=1}^{2} \binom{m_k}{x_k}\binom{n_k}{y_k} q_{1k}^{m_k} q_{2k}^{n_k} \exp\Big[x_1\big(\ln(p_{11}/q_{11}) - \ln(p_{21}/q_{21})\big)$$

$$+ x_2\big(\ln(p_{12}/q_{12}) - \ln(p_{22}/q_{22})\big)$$

$$+ x_1 \ln(p_{21}/q_{21}) + x_2 \ln(p_{22}/q_{22})$$

$$+ y_1 \ln(p_{21}/q_{21}) + y_2 \ln(p_{22}/q_{22})\Big].$$

$$= \prod_{k=1}^{2} \binom{m_k}{x_k}\binom{n_k}{y_k} q_{1k}^{m_k} q_{2k}^{n_k} \exp\Big[\sum_{i=1}^{2} x_i \ln \Delta_i + \sum_{i=1}^{2} (x_i + y_i)\ln(p_{2i}/q_{2i})\Big]$$

$$= \prod_{k=1}^{2} \binom{m_k}{x_k}\binom{n_k}{y_k} q_{1k}^{m_k} q_{2k}^{n_k} \exp\Big[x_2 \ln(\Delta_2/\Delta_1) + (x_1 + x_2)\ln \Delta_1$$

$$+ \sum_{i=1}^{2} (x_i + y_i)\ln(p_{2i}/q_{2i})\Big]$$

This belongs to the one parameter exponential family, and to test $H_0: \Delta_2 = \Delta_1$ against $H_1: \Delta_2 > \Delta_1$, a uniformly most powerful unbiased test is given by the rule:

Reject H_0 if $X_2 > c$ where c is a constant depending on $x_1 + x_2$, $x_1 + y_1$ and $x_2 + y_2$ and is determined so that the conditional probability of rejection given $x_1 + x_2 = \omega$, $x_1 + y_1 = t_1$, $x_2 + y_2 = t_2$ is α, see Lehman (1986).

Now the conditional probability is obtained as follows:

For i = 1, 2

$$P\left(X_i = x_i \middle| X_i + Y_i = t_i\right) = \frac{P(X_i = x_i,\ Y_i = t_i - x_i)}{P(X_i + Y_i = t_i)}$$

$$= \frac{\binom{m_i}{x_i}\binom{n_i}{t_i - x_i}\Delta_i^{x_i}}{\sum_{x_i}\binom{m_i}{x_i}\binom{n_i}{t_i - x_i}\Delta_i^{x_i}}$$

Therefore, the conditional distribution of X_1 and X_2 given $X_1 + Y_1 = t_1$, $X_2 + Y_2 = t_2$ is

$$= \frac{\binom{m_1}{x_1}\binom{n_1}{t_1 - x_1}\binom{m_2}{x_2}\binom{n_2}{t_2 - x_2}\Delta_1^{x_1}\Delta_2^{x_2}}{\left[\sum_{x_1=0}^{t_1}\binom{m_1}{x_1}\binom{n_1}{t_1 - x_1}\Delta_1^{x_1}\right]\left[\sum_{x_2=0}^{t_2}\binom{m_2}{x_2}\binom{n_2}{t_2 - x_2}\Delta_2^{x_2}\right]}$$

Under H_0, the conditional distribution of X_1 and X_2 given $X_1 + Y_1 = t_1$, $X_2 + Y_2 = t_2$ is

$$C_{t_1}^{(\Delta)} C_{t_2}^{(\Delta)} \binom{m_1}{x_1}\binom{n_1}{t_1 - x_1}\binom{m_2}{x_2}\binom{n_2}{t_2 - x_2}\Delta^{x_1 + x_2}$$

where

$$C_{t_1}^{(\Delta)} = \frac{1}{\sum_{x_1=0}^{t_1}\binom{m_1}{x_1}\binom{n_1}{t_1 - x_1}\Delta^{x_1}} \quad \text{and} \quad C_{t_2}^{(\Delta)} = \frac{1}{\sum_{x_2=0}^{t_2}\binom{m_2}{x_2}\binom{n_2}{t_2 - x_2}\Delta^{x_2}}.$$

Now conditional distribution of X_2 given $X_1 + X_2 = \omega$, $X_1 + Y_1 = t_1$ and $X_2 + Y_2 = t_2$ is given by

$$P(X_2 = x_2 \mid X_1 + X_2 = \omega,\ X_1 + Y_1 = t_1,\ X_2 + Y_2 = t_2)$$

$$= \frac{P(X_1 = \omega - x_2,\ X_2 = x_2 \mid X_1 + Y_1 = t_1,\ X_2 + Y_2 = t_2)}{P(X_1 + X_2 = \omega \mid X_1 + Y_1 = t_1,\ X_2 + Y_2 = t_2)}$$

$$= \frac{\binom{m_1}{\omega - x_2}\binom{n_1}{t_1 - \omega + x_2}\binom{m_2}{x_2}\binom{n_2}{t_2 - x_2}}{\sum_{x_2 = \mathrm{Max}(\omega - t_1, 0)}^{\mathrm{Min}(\omega, t_2)} \binom{m_1}{\omega - x_2}\binom{n_1}{t_1 - \omega + x_2}\binom{m_2}{x_2}\binom{n_2}{t_2 - x_2}}, \quad x_2 \in (\mathrm{Max}(0, \omega - t_1), \mathrm{Min}(\omega, t_2))$$

4. Analysis of the Air Force Data

To illustrate our methods, we have arranged the data at 3 time points, 50, 100 and 150.

1. Analysis of the cumulative data

The cumulative data looks like

	Time 50		Time 100		Time 150	
	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed
Dead	3	1	13	7	22	11
Alive	39	49	29	43	20	39
Total	42	50	42	50	42	50

(a) Pearson's Conditional Test

Time 50

$$\hat{p}_1 = \frac{3 + 1}{42 + 50} = .0435$$

$$E_{11} = 42(.0435) = 1.83$$

$$E_{21} = 50(.0435) = 2.17$$

$$\chi_1^2(1) = \frac{1}{1 - .0435}\left[\frac{(3 - 1.83)^2}{1.83} + \frac{(1 - 2.17)^2}{2.17}\right] = 1.45$$

The chi-square table value at $\alpha = .05$ is 3.84. So at time 50, the Pearson's conditional test does not detect differences.

Time 100

$$\hat{p}_2 = \frac{13 + 7}{42 + 50} = .217$$

$$E_{12} = 42(.217) = 9.13$$

$$E_{22} = 50(.217) = 10.87$$

$$\chi_2^2(1) = \frac{1}{.783}\left[\frac{(13 - 9.13)^2}{9.13} + \frac{(7 - 10.87)^2}{10.87}\right] = 3.85$$

So at time 100, the test rejects the null hypothesis that the relative risk is unity.

Time 150

$$\hat{p}_3 = \frac{22 + 11}{42 + 50} = .359$$

$$E_{13} = 42(.359) = 15.05$$

$$E_{23} = 50(.359) = 17.95$$

$$\chi_3^2(1) = \frac{1}{.641}\left[\frac{(22 - 15.05)^2}{15.05} + \frac{(11 - 17.95)^2}{17.95}\right] = 9.20$$

So at time 150, the death probabilities are not the same.

(b) Small sample test

Time 50

$$P(X = 4) = \frac{\binom{42}{4}\binom{50}{0}}{\binom{92}{4}} = .04$$

$$P(X = 3) = \frac{\binom{42}{3}\binom{50}{1}}{\binom{92}{4}} = .205$$

So reject H_0 if $X > 3$ at level .04. But the observed value of X is 3. So we are not able to reject H_0.

Time 100

$P(X = 20) = .00000000615$

$P(X = 19) = .0000002675$

$P(X = 18) = .0000519$

$P(X = 17) = .00005978$

$P(X = 16) = .0004593$

$P(X = 15) = .002504$

$P(X = 14) = .01006$

$P(X = 13) = .030529$

$P(X = 12) = .071107$

So $P(X > 12) = .04366 + P(X > 11) > .05$. Reject H_0 if $X > 12$ at level .04. But the observed value of X is 13. So at time 100, the null hypothesis that the relative risk is unity is rejected.

Time 150

$P(X = 33) = 4.13169 \times 10^{-17}$

$P(X = 32) = 7.1229 \times 10^{-15}$

$P(X = 31) = 5.07669 \times 10^{-13}$

$P(X = 30) = 2.09836 \times 10^{-11}$

$P(X = 29) = 5.6898 \times 10^{-10}$

$P(X = 28) = 1.08431 \times 10^{-8}$

$P(X = 27) = 1.518039 \times 10^{-7}$

$P(X = 26) = 1.6102 \times 10^{-6}$

$P(X = 25) = 1.32368 \times 10^{-5}$

$P(X = 24) = 8.57943 \times 10^{-5}$

$P(X = 23) = 4.4432 \times 10^{-4}$

$P(X = 22) = 1.85808 \times 10^{-3}$

$P(X = 21) = 6.3263 \times 10^{-3}$

$P(X = 20) = 1.76518 \times 10^{-2}$

$P(X = 19) = .0405$

So $P(X > 19) = .026$ and $P(X > 18) = .066$. We reject H_0 if $X > 19$ at 2.6% level. But the observed value of $X = 22$. So at time 150, the null hypothesis that the relative risk is unity is rejected.

2. Analysis of the Truncated data

The truncated data looks like

	Time 50		Time 100		Time 150	
	Exposed	Non-exposed	Exposed	Non-exposed	Exposed	Non-exposed
Dead	3	1	10	6	9	4
Alive	39	49	29	43	20	39
Total	42	50	39	49	29	43

(a) Pearson's Conditional Test

At time 50 the analysis is the same as for the cumulative data.

Time 100

$$\hat{p}_2 = \frac{10 + 6}{39 + 49} = .1818$$

$$E_{12} = 39(.1818) = 7.10$$

$$E_{22} = 49(.1818) = 8.91$$

$$\chi^2_2(1) = \frac{1}{.818}\left[\frac{(10 - 7.10)^2}{7.10} + \frac{(6 - 8.91)^2}{8.91}\right] = 1.746$$

So at time 100, the test is not able to detect the differences in the death rates.

Time 150

$$\hat{p}_3 = \frac{9 + 4}{29 + 43} = 1.80$$

$$E_{13} = 29(.180) = 5.24$$

$$E_{23} = 43(.180) = 7.76$$

$$\chi^2_3(1) = \frac{1}{.82}\left[\frac{(9 - 5.24)^2}{5.29} + \frac{(4 - 7.76)^2}{7.76}\right] = 5.51$$

So at time 150, the death probabilities are not the same in the two groups.

(b) Small sample test

At time 50, the analysis is the same as for the cumulative data.

Time 100

P(X = 16) = .0000026048

P(X = 15) = .000085091

P(X = 14) = .0012253

P(X = 13) = .0103367

P(X = 12) = .057234

So P(X > 12) = .011 and P(X > 11) = .068. Therefore we reject H_0 if X > 12 at 1.1% level. But the observed value of X is 10. So at time 100, the test does not detect the differences in the death rates.

Time 150

P(X = 13) = .000000957

P(X = 12) = .0000314709

P(X = 11) = .00044059

P(X = 10) = .003486

P(X = 9) = .01743048

P(X = 8) = .0582676

We notice that P(X > 8) is .021 and P(X > 7) = .079. So we reject H_0 if X > 8 at 2.1% level. But the observed value of X is 9. Thus the test detects differences in the death probabilities at 2.1% level.

(c) Mantel-Haenzel test

Time 50

$$V_{11} = \frac{42 * 50}{91} (.0435)(.6965) = .699$$

Also $E_{11} = 1.83$

$$x_{11}^2 = \frac{(3 - 1.83)^2}{.699} = 1.95$$

At time 50 the test does not detect difference in probabilities of death.

Time 100

$$V_{12} = \frac{39 \times 49}{87} (.1818)(.8182) = 3.26$$

Also $E_{12} = 7.10$

$$x_{12}^2 = \frac{(10 - 7.10)^2}{3.26} = 2.52$$

So the test does not detect difference in probabilities of death.

Time 150

$$V_{13} = \frac{29 \times 43}{71} (.18)(.82) = 2.59$$

Also $E_{13} = 5.24$

$$x_{13}^2 = \frac{(9 - 5.24)^2}{2.59} = 5.56,$$ which is more than the critical value 3.84 at the 5% level. Hence the test shows that the death rates in two groups during this period are different.

(d) Log rank test

In order to test the consistent differences between the relative risks in the two groups, we applied the log rank test as follows:

$E_{11} = 1.83$, $E_{12} = 7.10$, $E_{13} = 5.24$, $d_1 = 22$, $d_2 = 11$

$$x'^2 = \frac{(33 - 14.17)^2}{(1/33)(14.17)(18.83)} = 43.65$$

So we reject the null hypothesis at the 5% level.

3. Comparing the relative risks at two points

Let us apply our small sample test derived in section 3 to test $H_0:\Delta_1 = \Delta_2$ against $H_1:\Delta_1 > \Delta_2$ at times 50 and 100.

With $t_1 = 4$, $t_2 = 16$, $\omega = 13$,

$P(X_2 = 13) = .003$

$P(X_2 = 12) = .06$

So we reject H_0 if $X_2 > 11$ at 6.3% level. But the observed value of X_2 is 10. So the test is unable to detect differences at time 50 and 100.

To compare the relative risk at times 100 and 150, we have $t_1 = 16$, $t_2 = 13$, $\omega = 19$ and

$P(X_2 = 13) = .000043$

$P(X_2 = 12) = .001757$

$P(X_2 = 11) = .0216$

$P(X_2 = 10) = .112133$

So $P(X_2 > 10) = .0233$ and $P(X_2 > 9) = .1354$. Therefore we reject $H_0:\Delta_1 = \Delta_2$ if $X_2 > 10$ at 2.3% level. But the observed value of X_2 is 9. So the test is unable to reject H_0.

Similarly it can be checked that the hypothesis $H_0:\Delta_1 = \Delta_2$ against $\Delta_1 < \Delta_2$ is not rejected at the 5% level.

5. Discussion and further direction of Research

The analyses provided in section 4 and the data used are for illustration purpose. The division of the data into three class intervals [0,50), [50,100) and [100,150) is completely arbitrary. The contingency tables used in the analysis were formed from the data

(enclosed) in which the actual times of death were available. In the absence of actual times of death, the small sample results derived in section 3 are more general and can be employed when the data is available in the contingency table form as described in the Introduction. Even though the small sample and large sample procedures yielded similar results in an analyses of the Air Force data, the small sample procedures are uniformly most power unbiased. There is still a need to perform a power study of both the large and small sample procedures. Also the determination of sample sizes, in the case of small sample tests derived in this report, is an important problem especially in epidemiology. In medical studies, where humans and animals are involved, large samples are in general not available. Therefore, it is necessary for the scientist to design an experiment which will need the smallest sample size for the statistical test to achieve a given power. This sample size study should take into account the cost involved as sometimes it is less expensive to take the data on the control group than on the exposed group.

It should be noted that the small sample procedures derived in this report deal with the odds ratio rather than the relative risk. It is well known that for rare diseases, which is our situation, the odds ratio approximates the relative risk. However, in the general situation it is not clear if the hypotheses $H_0:\Delta_1 = \Delta_2$ and $H_0:R_1 = R_2$ are equivalent. In other words is it necessary that a procedure which supports $H_0:\Delta_1 = \Delta_2$ versus $H_a:\Delta_1 > \Delta_2$ also supports $H_0:R_1 = R_2$ versus $H_a:R_1 > R_2$? For example $p_1 = .05$ and $p_2 = .01$ gives a relative risk = 5 and odds ratio 5.210. Also $p_1 = .005$ and $p_2 = .001$ also gives a relative risk = 5 but odds ratio = 5.020.

For the bias caused in approximating the relative risk by the odds ratio, in the case of high values of incidence, see an interesting discussion by Feinstein (1986).

Thus some of the important problems which need further research are as follows:

(1) To perform a power study of the tests in question.

(2) To address the problem of the determination of sample sizes taking into consideration the costs involved.

(3) To study the relationship between odds ratio and relative risk in detail which can provide a clear insight into the methodology involved in these procedures.

References

Elandt-Johnson, R. C. and Johnson, N. L. (1980). Survival Models and Data Analysis. John Wiley & Sons, New York.

Feinstein, A. R. (1986). The bias caused by high values of incidence for p_1 in the odds ratio assumption that $1 - p_1 \approx 1$. Journal of Chronic Diseases 39(6), 485-487.

Lehman, E. L. (1986). Testing Statistical Hypothesis. John Wiley & Sons, New York.

Mantel, N. and Haenzel, W. (1959). Statistical Aspects of the analysis of data from retrospective studies of disease. Journal of National Cancer Institute 22, 719-748.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). Journal of the Royal Statistical Society, Series A, 135, 185-198.

Yochmowitz, M. G., Wood, D. H. and Salmon, Y. L. (1985). Seventeen-year mortality experience of Proton Radiation in Macaca Mulatta. Radiation Research 102, 14-34.

PRODUCT-LIMIT SURVIVAL ANALYSIS GROUP = E (MIF)

CASE LABEL	CASE NUMBER	TIME	STATUS	CUMULATIVE SURVIVAL	STANDARD ERROR
V34 CO M	76	17.00	DEAD	0.9762	0.0235
K19 DE F	84	25.00	DEAD	0.9524	0.0329
I55 DE F	87	28.00	DEAD	0.9286	0.0377
D50 CO M	66	54.00	DEAD	0.9048	0.0453
U06 CO M	72	55.00	DEAD	0.8810	0.0500
J88 DF M	88	56.00	DEAD	0.8571	0.0540
R90 CO M	67	58.00	DEAD	0.8333	0.0575
J54 DE M	80	60.00	DEAD	0.8095	0.0606
122 CO M	69	62.00	DEAD	0.7857	0.0633
K73 DF F	90	74.00	DEAD	0.7619	0.0657
U37 FF F	99	76.00	DEAD	0.7381	0.0678
U67 FF F	98	89.00	DEAD	0.7143	0.0697
T81 CO F	71	100.00	DEAD	0.6905	0.0713
R51 EF F	94	108.00	DEAD	0.6667	0.0727
V84 CO M	79	116.00	DEAD	0.6429	0.0739
M22 EF M	92	120.00	DEAD	0.6190	0.0749
K34 DE M	85	123.00	DEAD	0.5952	0.0757
U53 CO F	73	128.00	DEAD	0.5714	0.0764
P89 EF F	93	133.00	DEAD	0.5476	0.0768
T73 CO F	70	139.00	DEAD	0.5238	0.0771
V56 CO M	78	146.00	DEAD	0.5000	0.0772
V54 CO M	77	148.00	DEAD	0.4762	0.0771
K43 DF F	87	154.00	DEAD	0.4524	0.0769
T66 EF M	95	156.00	DEAD	0.4286	0.0764
F65 BC F	60	160.00	DEAD	0.4048	0.0757
S61 CO F	68	165.00	DEAD	0.3810	0.0747
U36 FF M	97	180.00	DEAD	0.3571	0.0735
U16 FF M	96	191.00	ALIVE		
V22 CO M	75	198.00	ALIVE		
U78 CO M	74	198.00	ALIVE		
L84 EF M	91	199.00	ALIVE		
K02 DE M	83	201.00	ALIVE		
J89 DE M	82	201.00	ALIVE		
J64 DE M	81	201.00	ALIVE		
K71 DE F	86	201.00	ALIVE		
F71 BC F	61	206.00	ALIVE		
F13 BC F	59	207.00	ALIVE		
E93 BC F	58	207.00	ALIVE		
J04 BC M	63	207.00	ALIVE		
J32 BC M	65	207.00	ALIVE		
J70 BC M	64	207.00	ALIVE		
G86 BC M	62	207.00	ALIVE		

MEAN SURVIVAL TIME = 138.93 LIMITED TO 207.00 S.E. = [illegible]

QUANTILE	ESTIMATE
75TH	76.00
MEDIAN (50TH)	146.00

28-28

Arch I

PRODUCT-LIMIT SURVIVAL ANALYSIS GROUP = Cubicle (MIT)

CASE LABEL			CASE NUMBER	TIME	STATUS	CUMULATIVE SURVIVAL	STANDARD ERROR
48P	JA	M	59	18.00	DEAD	0.9825	0.0174
S01	AA	F	1	59.00	DEAD	0.9649	0.0244
DS2	HA	M	34	62.00	DEAD	0.9474	0.0296
L25	EA	F	15	68.00	DEAD	0.9298	0.0338
L27	DA	F	16	70.00	DEAD	0.9123	0.0375
L75	EA	F	17	72.00	DEAD	0.8947	0.0406
96N	JA	M	57	99.00	DEAD	0.8772	0.0435
50K	JA	M	51	105.00	DEAD	0.8596	0.0460
6V4	JA	M	54	111.00	DEAD	0.8421	0.0483
9N5	JA	F	56	121.00	DEAD	0.8246	0.0504
69R	HA	F	37	142.00	ALIVE		
T78	EA	M	26	142.00	DEAD	0.8066	0.0524
67R	HA	F	36	142.00	ALIVE		
EJ4	JA	M	46	143.00	ALIVE		
DL2	JA	M	45	143.00	ALIVE		
61S	JA	F	55	143.00	ALIVE		
11S	HA	F	35	144.00	ALIVE		
A50	IA	M	38	150.00	ALIVE		
7X7	IA	F	43	150.00	ALIVE		
N16	IA	M	40	150.00	ALIVE		
04N	IA	M	42	150.00	ALIVE		
AW4	IA	M	39	150.00	ALIVE		
8X5	IA	F	44	150.00	ALIVE		
IR4	IA	M	41	150.00	ALIVE		
607	JA	F	53	154.00	DEAD	0.7822	0.0562
V66	FA	M	30	160.00	DEAD	0.7577	0.0595
36P	JA	M	49	163.00	ALIVE		
15E	JA	F	47	163.00	ALIVE		
V58	FA	M	29	165.00	DEAD	0.7316	0.0629
6P5	JA	F	52	173.00	ALIVE		
3V4	JA	M	48	173.00	ALIVE		
W43	FA	F	31	180.00	ALIVE		
600	FA	M	33	192.00	ALIVE		
004	FA	M	32	192.00	ALIVE		
U03	FA	M	27	193.00	ALIVE		
U22	FA	M	28	193.00	ALIVE		
899	EA	F	23	196.00	ALIVE		
R39	EA	F	19	197.00	ALIVE		
Q12	EA	M	18	197.00	ALIVE		
R41	EA	F	20	197.00	ALIVE		
S69	EA	F	22	197.00	ALIVE		
S56	EA	M	21	197.00	ALIVE		
T70	EA	M	25	197.00	ALIVE		
T68	EA	M	24	197.00	ALIVE		
911	AA	F	2	198.00	ALIVE		
W02	CA	M	8	198.00	ALIVE		
U74	CA	M	5	198.00	ALIVE		
P46	CA	M	4	198.00	ALIVE		
U76	CA	M	6	198.00	ALIVE		
V86	CA	M	9	198.00	ALIVE		

BMDP1L Page 3
ALL CONTROLS(M+F) VS 500-650RADS(M+F) PROTONS

CASE LABEL			CASE NUMBER	TIME	STATUS	CUMULATIVE SURVIVAL	STANDARD ERROR
U96	CA	M	7	198.00	ALIVE		
L16	BA	M	13	199.00	ALIVE		
J12	BA	M	11	200.00	ALIVE		
J17	BA	F	12	200.00	ALIVE		
L22	BA	M	14	200.00	ALIVE		
J07	BA	F	10	200.00	ALIVE		
S77	BA	F	3	207.00	ALIVE		

MEAN SURVIVAL TIME = 179.75 LIMITED TO 207.00 S.E. = 7.028

QUANTILE	ESTIMATE
75TH	165.00

FINAL REPORT NUMBER 29
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Fabian Hadipriono
760-6MG-054

FINAL REPORT NUMBER 30
REQUESTED A NO-COST TIME EXTENTSION
TO BE SUBMITTED IN 1987 MINI-GRANT FINAL REPORT
Dr. Frank Hadlock
760-6MG-073

Studies in Optimum Shape Synthesis for Structures Undergoing Plastic Deformation

Prabhat Hajela
Aerospace Engineering, Mechanics, and Engineering Science
University of Florida, Gainesville, Florida 32611

March 1988

Final Report

Submitted to Universal Energy Systems
as the Required Report for USAF Mini Grant Program

ABSTRACT

The present report documents the major findings of a year-long study that focussed on the problem of geometric shape optimization. The thrust of the effort was two-fold. The first was to assess the applicability of nonlinear programming based optimization algorithms in the sizing of structures undergoing large strain-rate, plastic deformations under dynamic impact loads. The use of approximate methods in analysis and optimization were explored in this context. A parallel effort was also directed at evaluating alternate methods of analysis for the optimum shape synthesis problem, and the boundary element method emerged as a viable alternative for the given task. The use of this method in elastic shape design is described in this report.

INTRODUCTION

The emergence of high speed digital computing capabilities have added significantly to the potential afforded by nonlinear programming methods of optimization in preliminary and detailed design. Numerous applications of this approach in the design of elastic structural systems are documented in literature. In typical structural design problems, the member sizes of structural components are changed in accordance with a gradient based search procedure, which upon termination, guarantees at least a local optimum in the prescribed design space. The topology of such structural systems is generally fixed.

The present report outlines two problems in structural design, where the structural domain was allowed to change in the redesign process, and can be classified in the general category of shape design problems. One of these problems was lent further complexities in that the structural response for this problem was characterized by material nonlinearities. In other words,

plastic deformations were included in the optimum synthesis problem. The inclusion of plastic collapse as a design constraint for structural synthesis is not a new idea. In fact, some of the earliest developments in structural optimization focussed on ultimate performance requirements as a design criterion. These efforts were largely restricted to rigid-plastic collapse applied to structures governed by piecewise linear constitutive laws. Other studies have approached this problem as a general nonconvex, nonlinear programming problem, with some effort directed at developing computationally viable approximations for this problem, including reducing the size of the problem by mathematical decomposition.

The need for alternative methods of analysis in the shape design problem was prompted because of the special attention that must be given to the computation of response sensitivities with respect to the shape variable. Both finite element and finite difference methods are inherently dependent on domain discretization. A perturbation in the shape variables would require special monitoring to ensure that the domain mesh does not distort and introduce misleading information about the structural response. Further, as the domain was redefined in the sizing process, the domain grid or element mesh would have to be adaptively redefined, resulting in increased computational costs. The boundary element method is based on a boundary discretization only, and furthermore, provides extremely accurate response information at the boundary. The latter contrasts with the finite element method, where response information close to the boundary is most suspect. The boundary element methodology is therefore considered a worthwhile alternative, and was explored in this effort with simplistic structural configurations.

The results described in this report were obtained in collaborative efforts involving the principal investigator and two graduate students. Mr. M. E. Gunger is a candidate for the MS degree at the University of Florida, and worked primarily on the

structural problem involving dynamic plasticity. The other student, Mr. J. Jih is a candidate for the doctoral degree, and was involved in the program exploring boundary element methods for shape design. Brief introductions to each of these problem is given next, with detailed results obtained in these studies included as appendices to this report.

The Target Impact and Penetration Problem

This work focussed on a target impact-penetration problem, where dynamic plasticity plays a dominant role. A framework for an optimization system was first established, using the hydrocode EPIC-2 as the primary analysis tool, coupled to a feasible usable search direction optimization algorithm through a series of pre- and post-processors. Specific test problems were then attempted to validate the synthesis procedure and consisted of a steel projectile impacting a rigid, impenetrable surface as well as a penetrable concrete slab with a specified velocity. The internal and external shape of the projectile shell was allowed to vary in the redesign with an objective of maximizing the internal volume, with design constraints limiting the transient pressure in the explosive and the plastic strain in the structural casing, to specified allowable levels.

The study was particularly useful in identifying the salient features of the problem. These included the need to provide an adaptive grid generation scheme for modelling the structural domain during deformation and redesign to retain computational viability in addition to accuracy, and appropriate definition of design variables and constraints. An additional feature that was added to the methodology was to use the sensitivity of the optimum design to preassigned problem parameters, to predict new optima without actually going through a process of redesign. For example, if the impact velocity or the angle of impact that was used in the design was perturbed by 10%, the design sensitivity allows one to compute a new optimum objective function and new

optimum design variables without actually going through a complete optimization to obtain these results.

BEM in Shape Design

The boundary element method is an alternative approach to obtaining a solution to a given set of differential equations. The basis of this method is in the transformation of the differential equation into an equivalent set of integral equations, the solution of which needs information at end points. A discrete representation of this problem would only require that the boundary of the domain be discretized. The order of the resulting linear system of equations is considerably less than in the finite element method.

The work performed under the present grant is best described as a preliminary effort in the adaptation of boundary elements for optimal shape design. The boundary element technique was developed for torsion problems and plane stress and plane strain problems in elasticity, using both constant and linear boundary elements. The optimal design problem was posed as a nonlinear programming problem, with the linear algebraic system of equations from the boundary element approach treated as equality constraints. The solution to the analysis and the optimization thus proceeded simultaneously. The thrust of the present effort was directed at examining approximation strategies that would influence the computational requirements of this approach. The use of both nodal coordinates and boundary descriptor functions as design variables was studied in relation to these approximation concepts. The influence of alternate formulation of constraints in the design problem was also examined.

Appendices can be obtained from

Universal Energy Systems, Inc.

FINAL REPORT NUMBER 32
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Donald Hanson
760-6MG-092

Final Report
on
PULSED POWER CONDUCTORS

UES/AFOSR Mini Grant Program (1986-87)

James C. Ho
The Wichita State University

INTRODUCTION

Following a 1985 Summer Faculty Research Program, which identified several types of materials as promising candidates for advanced pulsed power conductors, this work has been carried out to experimentally determine their temperature dependence of electrical resistivity and to further survey other existing or new materials of our special interest.

Central to the requirements of high current, pulsed power devices is that the conductors should have high strength at service temperatures, while maintaining reasonable low electrical resistivity. The low resistivity requirement points to the pure metals, particularly aluminum and copper. The pure metals, however, do not normally have enough mechanical strength. Solid solution strengthening through alloying can help in this respect, but has the detrimental effect of increased electrical resistivity. Another much more appropriate strengthening mechanism involves dispersed particles in the pure metal matrix. The so-called dispersion strengthening is the basis of many new structural materials being developed today as a consequence of the advancement in powder metallurgy. The latter has

greatly enhanced the capability of synthesizing almost any conceivable alloy system under equilibrium or nonequilibrium conditions. Furthermore, powder metallurgy's role in net shape forming simplifies processing steps and help form high strength parts of complicated configurations at much reduced cost.

SAMPLE MATERIALS

Samples of several materials commercially available were obtained for evaluation. Since most materials of interest to us are still in their development stage, sample acquisition is limited.

(1) Cu-Aluminum oxide: The properties of this family of dispersion strengthened copper arise from a fine and uniform dispersion of aluminum oxide particles in the copper matrix. These particles range in size about 30 Å to 120 Å with an interparticle spacing between about 300 Å to 1000 Å. The aluminum oxide particles are hard and thermally stable at high temperatures. They retain their original particle size and interparticle spacing even at temperatures approaching the melting point of copper.

The sample used here was obtained from SCM Chemicals. It is made by powder metallurgy and internal oxidation. The

latter produces the finest dispersoid particles and the most uniform distribution, which are critical to obtaining high strength and resistance to softening at elevated temperatures. The processes involve melting a dilute solid solution alloy of aluminum in copper and atomizing the melt by high pressure gas such as nitrogen. The resulting powder is blended with an oxidant comprising fine copper oxide powder. The blend is heated to a high temperature at which the copper oxide dissociates and the oxygen thus produced diffuses into the particles of solid solution copper-aluminum alloy. As aluminum is a stronger oxide former than copper, the aluminum in the alloy gets preferentially oxidized to aluminum oxide. Any excess oxygen left in the powder, after complete oxidation of the aluminum, is reduced by heating the powder in hydrogen or dissociated ammonia atmosphere. The resulting powders are then fabricated into fully dense shapes by various techniques such as rolling or extrusion.

(2) Cu-Nb: A family of reinforced copper with in-situ formed filaments exhibits an increased strength considerably higher than that predicted by the rule of mixtures.[1,2] These composites are usually formed in-situ using as starting materials two-phase alloys prepared either by means of powder metallurgy or by quenching a liquid solution of two components which are mutually insoluble in the solid phase. The third possibility is the cast or directionally solidified eutectic composites. Provided both phases are ductile, such

two-phase alloys can be mechanically processed to large reduction in cross-sectional area until the in-situ formed filaments are sufficiently small.

The Supercon sample used in this study contains 18 wt% Nb in Cu. As a result of extrusion and wire drawing processes, very fine Nb filaments disperse throughout the copper matrix with strong metallurgical bond between them. The room temperature UTS value reaches above 200 ksi.

(3) Al-Fe-Ce: This alloy can play two different roles in high pulsed power applications. In addition to be used as a high strength conductor with reasonably good electrical conductivity, it can also serve as the matrix material for a cryoconductor containing multifilamentary, high purity aluminum with extremely low resistivity at liquid hydrogen temperatures. The latter has been addressed in two recent articles[3,4] and a pending patent.[5] This work deals with the electrical resistivity at elevated temperatures.

The sample materials are powder-metallurgically synthesized at ALCOA,[6] under the AFML sponsorship. A hot-vacuum-pressed (HVP) sample in its as-received condition was measured. For comparison, another sample represents the dynamically recrystallized condition. While the HVP material has greater strength, the cryoconductors employ the dynamically recrystallized material due to processing

concerns.

(4) Al-SiC: A series of high-performance aluminum composites containing SiC particulates or whiskers become available in recent years. The carbides as reinforcing materials to stiffen and strengthen lightweight, but relatively soft aluminum are hard, refractory ceramics, which also have low densities. The whiskers, about 0.5 μm in diameter and 30 μm long, are of particular interest because they are essentially single crystal fibers. The resulting composites exhibit superior microcreep stability, and can be readily fabricated in conventional metal-forming processes.

Samples used in this study include aluminum alloys containing SiC particulates and whiskers from ARCO Chemical Co. and similar materials containing SiC particulates from DWA Composite Specialties.

ELECTRICAL RESISTIVITY MEASUREMENTS

The sample was placed within a tube furnace with temperature control. Argon gas passed through the tube continuously to minimize sample oxidation at elevated temperatures. A chromel-alumel thermocouple was in direct contact with the sample for temperature readings. The electrical resistivity measurements were based on the

standard four-probe method. No significant thermal hysteresis was noticed as long as the sample was not heated up to close to its melting point.

RESULTS AND DISCUSSION

Table 1 lists the experimental data in terms of electrical resistivity $\rho(T)$ and IACS values for each sample. These results are also displayed in Fig. 1 (Cu base materials) and Fig. 2 (Al base materials) as the temperature dependence of the pct IACS values:

$$\text{pct IACS} = (\rho_{Cu}/\rho)\times 100$$

where $\rho_{Cu} = 1.7241\times10^{-6}$ ohm-cm

= International Annealed Copper Standard.

For comparison, curves representing literature data[7] for pure Cu and Al are also shown in the figures.

For all alloys, the reasonably good electrical conductivity (i.e., pct IACS values close to that of pure Cu or Al) at near room temperatures suggests that the dispersed particles or filaments do not behave as strong scattering centers. This, coupled with the enhanced strength due to dispersion strengthening, make these materials suitable for pulsed power applications.

For Al-Fe-Ce, the dispersed intermetallic compounds based on the alloying elements Fe and Ce do not diffuse or contribute appreciably to electron scattering in the HVP condition. Consequently, the mechanical strength holds up quite well to almost 400°C.[6] At higher temperatures, the difference in resistivity between the HVP and dynamically recrystallized samples becomes somewhat less pronounced as expected.

CONCLUSION

(1) As far as electrical conductivity is concerned, all materials studied here exhibit no serious degradation at elevated temperatures approaching the melting point of the matrix metal, Cu or Al.

(2) The actual IACS values of these materials are quite acceptable, in comparison with pure Cu or Al. Their enhanced mechanical properties are achieved through dispersion strengthening. The dispersoids do not form strong scattering centers for conducting electricity.

(3) There are several other materials which may also be considered as potential candidates for pulsed power conductors. For example, a mechanically alloyed, high strength (UTS of 65 ksi near room temperature), high

conductivity (50% IACS) Al is being produced by Novamet of Wyckoff, NJ. Mechanical alloying, a powder metallurgical process, is the combination of plastic deformation, cold welding, and grinding of powder particles during high energy milling.[8] Carbon, derived from organic process control agents, is incorporated into the processed powders and react with aluminum to form very fine aluminum carbides. These carbides and fine oxide particles, derived from the break up of surface films on the initial powder particles, create a dispersion which stabilizes a submicron grain size and greatly enhance the mechanical strength.

Another potential material is Cu or Al cladded in high strength alloys. Pfizer Composite Metal Products has manufactured a stainless steel cladded aluminum by forming a strong metallurgical bond between the two metals through rolling. The aluminum contributes excellent thermal and electrical conductivity to the composite while stainless steel acts as the strength component.

It is recommended that further testing of all potential materials be made towards final selection for a specified application.

REFERENCES

1. J. Bevk, J. P. Harbison, and J. L. Bell, J. Appl. Phys. 49, 6031 (1979).

2. K. R. Karasek and J. Bevk, Scripta Metall. 13, 259 (1979).

3. J. C. Ho, C. E. Oberly, H. L. Gegel, W. T. O'Hara, J. T. Morgan, Y. V. R. K. Prasad, and W. M. Griffith, Proc. 5th IEEE Pulsed Power Conference, Arlington, VA, 1985, pp. 627-629.

4. J. C. Ho, C. E. Oberly, H. L. Gegel, W. M. Griffith. J. T. Morgan, W. T. O'Hara, and Y. V. R. K. Prasad, Adv. Cryogenic Engineering 32, 437 (1986).

5. C. E. Oberly, J. C. Ho, and H. L. Gegel, Air Force Invention #16952 (1985).

6. W. M. Griffith, R. E. Sanders, Jr., and G. J. Hildeman, in High Strength Powder Metallurgy Aluminum Alloys, edited by M. J. Koczak and G. J. Hildeman, the Metallurgical Society of AIME, 1982, pp. 209-224.

7. Thermophysical Properties of High Temperature Solid Materials, Vol. 1, ed. by Y.S. Touloukian, McMillan Co., NY, (1967).

8. J. S. Benjamin and M. J. Bomford, Metall. Trans. 8A, 1301 (1977).

Table 1. Temperature dependence of electrical resistivity and IACS values of several aluminum and copper base alloys.

Cu-aluminum oxide (SCM Metal Products)

T (°C)	ρ(T) (μΩ-cm)	IACS (pct)
22	2.49	69.2
50	2.74	62.9
98	3.09	55.8
147	3.49	49.4
197	3.98	43.4
295	4.73	36.4
390	5.58	30.9
462	6.30	27.4
579	7.37	23.4
673	8.32	20.7
768	9.39	18.4
866	10.51	16.4
903	11.08	15.6
941	11.60	14.9
967	12.00	14.4
992	12.43	13.9

Cu-Nb (Supercon)

T (°C)	ρ(T) (μΩ-cm)	IACS (pct)
22	2.93	58.8
39	3.05	56.5
74	3.28	52.6
122	3.66	47.1
172	3.96	43.5
246	4.40	39.2
343	5.16	33.4
438	6.04	28.5
532	6.83	25.2
625	7.44	23.2
721	8.17	21.1
817	9.02	19.1
864	9.70	17.8
903	9.99	17.3
916	10.28	16.8
941	10.90	15.8
967	11.78	14.6
992	12.26	14.1

Al-Fe-Ce, VHP (ALCOA)

T (°C)	ρ(T) (μΩ-cm)	IACS (pct)
22	5.58	30.9
40	5.80	29.7
78	6.53	26.4
134	7.48	23.0
209	8.93	19.3
283	10.27	16.8
355	11.77	14.6
426	13.34	12.9
473	14.34	12.0
520	15.51	11.1
567	16.74	10.3

Al-Fe-Ce, dynamically recrystallized (ALCOA)

T (°C)	ρ(T) (μΩ-cm)	IACS (pct)
22	5.58	30.9
50	6.03	28.6
98	6.98	24.7
172	8.48	20.3
246	10.10	17.1
319	11.83	14.6
390	13.62	12.7
450	15.07	11.4
496	16.18	10.7
543	17.52	9.8
590	19.14	9.0
614	19.81	8.7
637	20.65	8.3

2124 Al-SiC particulates (ARCO)

T (^{o}C)	ρ(T) (μΩ-cm)	IACS (pct)
22	7.12	24.2
57	7.69	22.4
105	8.69	19.8
134	9.26	18.6
209	10.82	15.9
283	12.32	14.0
355	14.03	12.3
426	16.02	10.8
468	17.23	10.0
520	18.73	9.2
543	19.79	8.7
567	21.08	8.2
579	21.93	7.9
602	23.71	7.3
616	25.77	6.7

2124 Al-SiC Whiskers (ARCO)

T (oC)	ρ(T) (μΩ-cm)	IACS (pct)
22	6.05	28.5
45	6.41	26.9
88	7.26	23.7
159	8.53	20.2
234	9.98	17.3
307	11.50	15.0
379	13.07	13.2
414	14.16	12.2
451	15.73	11.0
485	16.64	10.4
508	17.56	9.8
532	18.33	9.4
574	20.09	8.6
593	21.60	8.0
614	23.05	7.5

2014 Al-SiC particulates (DWA) - 1st sample

T (^{o}C)	ρ(T) ($\mu\Omega$-cm)	IACS (pct)
22	5.58	30.9
62	6.25	27.6
98	6.98	24.7
134	7.81	22.1
172	8.37	20.6
209	9.21	18.7
246	9.88	17.4
283	10.77	16.0
319	11.55	14.9
355	12.28	14.0
390	13.17	13.1
426	14.28	12.1
461	15.40	11.2
485	16.35	10.5
508	17.24	10.0

2014 Al-SiC particulates (DWA) - 2nd sample

T (^{o}C)	ρ(T) ($\mu\Omega$-cm)	IACS (pct)
25	5.58	30.9
43	5.75	30.0
62	6.08	28.4
98	6.70	25.7
134	7.31	23.6
172	8.20	21.0
209	8.98	19.2
246	9.77	17.6
283	10.8	16.0
319	11.6	14.9
355	12.3	14.0
390	13.2	13.1
426	14.3	12.1
461	15.4	11.2
485	16.3	10.6
508	17.2	10.0

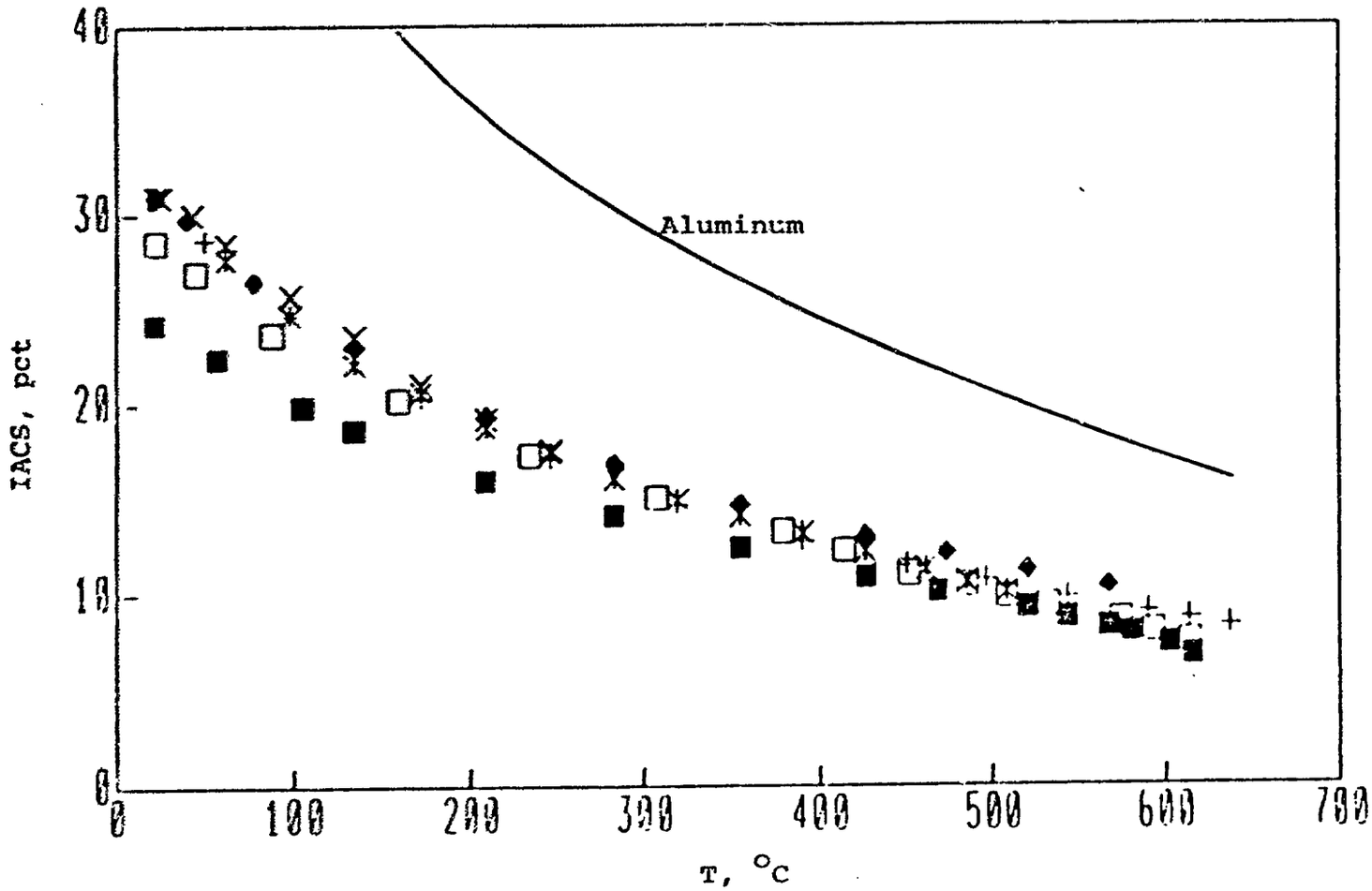


Fig. 2. Temperature dependence of electrical resistivity in terms of IACS values.

◆ : ALCOA Al-Fe-Ce VHP; +: ALCOA Al-Fe-Ce Dynamically Recry.
■ : ARCO Al-SiC Particulates; □: ARCO Al-SiC Whiskers
✳ : DWA Al-SiC particulates, Sample #1
× : DWA Al-SiC particulates, Sample #2

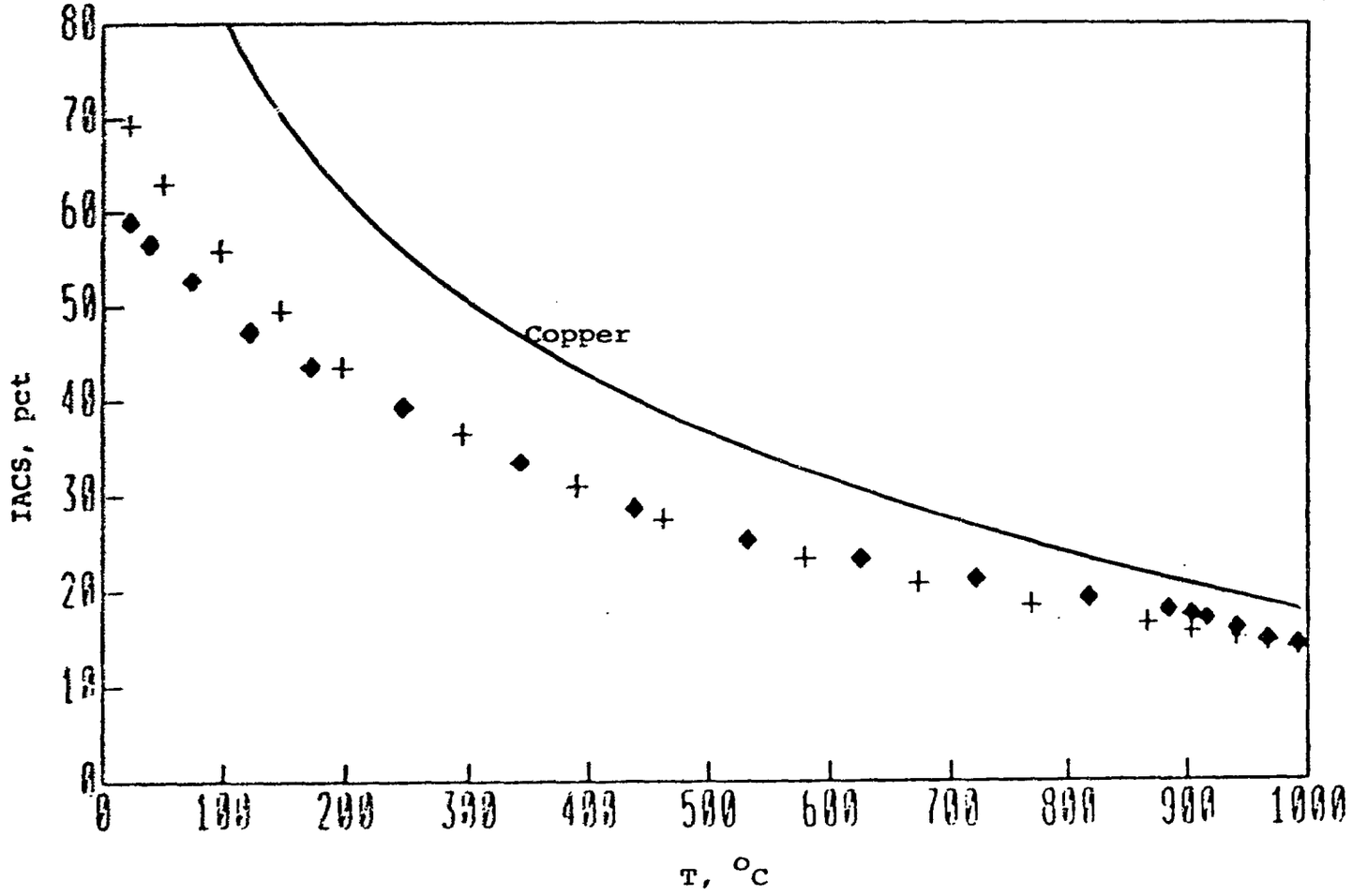


Fig. 1. Temperature dependence of electrical resistivity in terms of IACS values.

◆: SCM Cu-aluminum oxide; +: Supercon Cu-Nb

THE LOCALLY IMPLICIT METHOD FOR COMPUTATIONAL AERODYNAMICS

FINAL REPORT

Peter Hoffman

Computational Mathematics Group

University of Colorado at Denver

1100 Fourteenth Street

Denver, Colorado 80202

submitted to Universal Energy Systems

December 31, 1987

THE LOCALLY IMPLICIT METHOD

FOR COMPUTATIONAL AERODYNAMICS

ABSTRACT

Investigation of the locally implicit method is part of the search for a less expensive method to compute high Reynolds Number flows. Such flows are subject to severe Courant Number restrictions unless implicit methods are used. For general flows, implicit methods use approximate factorization techniques and require structured grids. There is a manpower cost associated with the generation of structured grids, and there will be a future computational cost associated with lack of parallelism in such methods. The locally implicit method, which was studied by an AFOSR-UES investigation during the summer of 1986, appears to overcome these objections. The current research extends this theoretical investigation of one conservation equation in one spatial variable to systems of Navier-Stokes equations in two or three spatial dimensions. The research originally proposed to use a characteristic or splitting approach to extend the method. The report explains the futility of this original approach and then develops the theory based on a new approach.

ACKNOWLEDGMENTS

I would like to thank the United States Air Force, Marshall Kingery, and Universal Energy Systems for this research opportunity.

Furthermore, I appreciate my research relationship with Jim Jacocks, Head, Computational Fluid Dynamics, Propulsion Wind Tunnel Facility, Arnold Engineering Development Center, and with K.C. Reddy, Professor, and Mark Ratcliff, graduate student, University of Tennessee Space Institute. I extend my appreciation to all those associated with Computational Fluid Dynamics at Calspan/AEDC.

Finally, I thank Craig Rasmussen, graduate student, University of Colorado at Denver, who assisted me with this project.

1. Introduction. The locally implicit method was developed by K.C. Reddy and Jim Jacocks of the Computational Fluid Dynamics Section in the Propulsion Wind Tunnel Facility at the Arnold Engineering Development Center. The goal is to reduce the cost presently required to obtain solutions to viscous aerodynamics problems.

If cost were no object, the central difference spatial approximation with implicit Euler time stepping would be a relatively good method; it will be used as a basis for comparison. It is certainly better than explicit time stepping when fine spatial discretizations impose severe Courant Number restrictions. Locally implicit methods avoid the costly linear algebra of implicit solutions by either time-lagging the right point and marching to the right or time-lagging the left point and marching to the left. This is the basic one-point locally implicit method.

An N-point locally implicit method uses N-1 central difference stencils with implicit Euler time stepping. Depending upon the direction of march, either a left or right one-point locally implicit stencil is added to the group so that the resulting system is solvable. It has N equations to solve simultaneously and, hence, is implicit locally.

Because the N-1 fully implicit stencils impose no stability restriction, the stability restriction on the group is imposed by the one-point locally implicit stencil.

2. Objective. The object of this research was to extend the analysis of the one-point locally implicit method. In particular, the following questions were to be investigated:

1. Is there a more effective way to apply the locally implicit method to systems of equations? In particular, can time-accuracy be restored to the method for systems?

2. Is the locally implicit method a good method for a parallel computer, or are there unforeseen problems?

3. Is there a simple multigrid accelerator for this method?

This report reviews the basic locally implicit method, explains the modified equation variant of the method, and compares these with a well-known scheme. Then, the futile attempt to extend the method along the lines of the proposal is discussed. This is followed by a discussion of a successful approach.

3. Basic Method. Consider the one point method for one dimensional linear convection,

$$u_t + cu_x = 0.$$

Marching to the right uses the numerical discretization

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^{n+1}}{2\Delta x} = 0$$

or

$$(I - \frac{c\Delta t}{2\Delta x}E^{-1})u_j^{n+1} = (I - \frac{c\Delta t}{2\Delta x}E)u_j^n, \quad \text{where } u_{j+1} = Eu_j.$$

Fourier transformation gives

$$(1 - \frac{c\Delta t}{2\Delta x}e^{-i\omega h})\hat{u}_j^{n+1} = (1 - \frac{c\Delta t}{2\Delta x}e^{i\omega h})\hat{u}_n, \tag{1}$$

so that the amplification factor $|z| = \left|\frac{\hat{u}^{n+1}}{\hat{u}^n}\right|$ is

$$|z| = \left|\frac{1 - \frac{c\Delta t}{2\Delta x}e^{i\omega h}}{1 - \frac{c\Delta t}{2\Delta x}e^{-i\omega h}}\right| = 1. \tag{2}$$

The method appears to be unconditionally stable according to this modal or von Neumann analysis.

In Eulerian formulations, the convection equation is nonlinear,

$$u_t + uu_x = 0.$$

This equation needs entropy information so viscosity is added

$$u_t + uu_x = \epsilon u_{xx},$$

to eliminate expansion shocks. Numerically, viscosity gives thickness to shocks and permits stable algorithms. Jameson[1] advocates the use of an artificial viscosity which gives a shock thickness proportional to meshwidth; this implies that $\epsilon u_{xx} = \nu_2 \Delta x u_{xx}$. Away from the shock, he switches to a fourth derivative artificial viscosity, $\nu_4 \Delta x^3 u_{xxxx}$, to remove high frequency error components from the numerical solution. Being of third order in Δx, this term cannot contaminate our second order spatial discretization. Both artificial viscosity terms vanish as the mesh is refined,

$$u_t + uu_x - \nu_2 \Delta x u_{xx} + \nu_4 \Delta x^3 u_{xxxx} = 0,$$

and the equation describing the original physics is recovered.

The necessary inclusion of viscosity, real or artificial, couples with convection to destabilize the locally implicit method.

The basic one point locally implicit method has a von Neumann stability criterion of the form

$$C \geq -2 + \alpha V$$

where the signed Courant Number, $C = c\frac{\Delta t}{\Delta x}$, and $V = (\nu_2 + \nu_4)\frac{\Delta t}{\Delta x}$. In aerodynamics, stability problems will occur in this basic method because of the negative characteristic velocities arising in subsonic and transonic calculations.

To see the reasoning in the special case,

$$u_t + cu_x - \nu \Delta x u_{xx} = 0,$$

consider

$$u_j^{n+1} = u_j^n - \frac{C}{2}(u_{j+1}^n - u_{j-1}^{n+1}) + V(u_{j+1}^n + u_{j-1}^{n+1} - 2u_j^{n+1}).$$

Fourier transformation gives, when $h = \Delta x$,

$$\begin{aligned} z &= \frac{1 - \frac{C}{2}e^{i\omega h} + Ve^{i\omega h}}{1 - \frac{C}{2}e^{-i\omega h} + V(2 - e^{-i\omega h})} \\ &= \frac{(1 - \frac{C}{2}\cos\omega h + V\cos\omega h) + i(-\frac{C}{2}\sin\omega h + V\sin\omega h)}{1 - \frac{C}{2}\cos\omega h + V(2 - \cos\omega h)) + i(\frac{C}{2}\sin\omega h + V\sin\omega h)}. \end{aligned}$$

The amplification factor, $|z|$, is not greater than one if and only if

$$(1 - \frac{C}{2}\cos\omega h + V(2 - \cos\omega h))^2 + (\frac{C}{2}\sin\omega h + V\sin\omega h)^2 \geq$$

$$(1 - \frac{C}{2}\cos\omega h + V\cos\omega h)^2 + (-\frac{C}{2}\sin\omega h + V\sin\omega h)^2,$$

which, for $V > 0$, simplifies to

$$C \geq -2 + \alpha V \tag{3}$$

where $\alpha = -2$. Notice that for the highest frequency, $\omega h = \pi$, the denominator of z is

$$1 - \frac{C}{2}e^{-i\pi} + V(2 - e^{-i\pi}) = 0,$$

which vanishes when $C = -2 - 6V$, and unbounded amplification occurs.

A variety of approximations for the viscous term are possible. For $0 \leq \theta \leq 1$,

$$u_j^{n+1} = u_j^n - \frac{C}{2}(u_{j+1}^n + u_{j-1}^{n+1} - 2(\theta u_j^n + (1-\theta)u_j^{n+1}))$$

has the same stability criterion but with

$$\alpha = \begin{cases} \text{positive}, & \text{for } \theta < \frac{1}{2} \\ \text{zero}, & \text{for } \theta = \frac{1}{2} \\ \text{negative}, & \text{for } \theta > \frac{1}{2} \end{cases}.$$

Unfortunately, simple computations with the basic method for, say, $C = +2$, fail miserably. Another explanation is necessary; the notion of group velocity must be introduced. A complete explanation can be found in Vichnevetsky and Bowles[3] or Trefethen[2].

Briefly, the group velocity, $c*$, is function of frequency and is the velocity with which the Fourier component of that frequency is observed to travel. This is not to be confused with the phase velocity, $c*$, which is not observable. Components of differing frequencies and phase velocities interact in the phenomenon called beating to give the observed group velocity.

The group velocity is computed from z as follows,

$$z = Re\, z + i\, Im\, z$$

$$\text{Arg}\, z = \tan^{-1}\left(\frac{Im\, z}{Re\, z}\right)$$

$$\text{phase velocity, } c* = -\frac{\text{Arg}\, z}{\omega \Delta t}$$

$$= -\frac{c\, \text{Arg}\, z}{C\, \omega h},$$

where the Courant Number, $C = \frac{c\Delta t}{h}$, and c is the true characteristic velocity for the partial differential equation. Then,

$$\text{group velocity, } v* = \frac{d}{d\omega}(\omega c*).$$

Using $z = \frac{a^{n+1}}{a^n}$ from equation [1] gives

$$\text{Arg}\, z = \tan^{-1}\left(\frac{-C\sin\omega h + \frac{C^2}{4}\sin 2\omega h}{1 - C\cos\omega h + \frac{C^2}{4}\cos 2\omega h}\right).$$

In the special case, $C = 2$,

$$\text{Arg}\, z = \tan^{-1}\left(\frac{-2\sin\omega h(1-\cos\omega h)}{-2\cos\omega h(1-\cos\omega h)}\right) = \omega h,$$

$$c* = \frac{-c\,\text{Arg}\, z}{C\omega h} = -\frac{c}{2},$$

$$v* = \frac{d}{d\omega}(\omega c*) = -\frac{c}{2}.$$

This indicates that at a Courant Number of 2, components of all frequencies travel in the wrong direction at half their proper speed. This is confirmed numerically. The pulse which should have moved four units to the right, moves instead 2 units to the left.

This is not unusual. Many popular, common schemes exhibit this property at high frequencies. It causes unwanted reflections at boundaries and GKS instability. Artificial viscosity is commonly used to control it.

What is unusual in this case is that all components, both low and high frequencies, travel in the wrong direction for $C = 2$. Artificial viscosity is effective only for controlling high frequencies.

In summary, if viscosity is small, the amplification factor can be unbounded for $C < -2$. The group velocity of the important low frequency components is reversed if $C \geq 2$. Thus, the locally implicit method in its basic form is stable only if $|C| < 2$.

The modified equation is another way of providing the preceding analysis. When the basic locally implicit method is applied to

$$u_t + c u_x - \nu_2 \Delta x\, u_{xx} + \nu_4 \Delta x^3\, u_{xxxx} = 0, \tag{4}$$

the equation actually solved is the modified equation,

$$(1 + \frac{\Delta t}{\Delta x}(-\frac{c}{2} + \nu_2 + 3\nu_4))u_t + c u_x - \nu_2 \Delta x\, u_{xx} + \nu_4 \Delta x_3\, u_{xxxx} = O(\Delta x + \Delta t),$$

where the right hand side vanishes as the mesh is refined for a given Courant Number. The modified equation is derived by replacing each of the finite difference approximations by its Taylor expansion and collecting terms.

If viscosity is negligible, the equation resembles

$$u_t + \frac{c}{1 - \frac{c\Delta t}{2\Delta x}} u_x = 0.$$

Observe that it represents a flow which reverses direction when $\frac{c\Delta t}{2\Delta x} = 1$, that is, when $C = 2$. This agrees with the group velocity analysis.

The viscous portion,

$$u_t - \nu_2 \Delta x\, u_{xx} = 0,$$

is unstable if the time coordinate is reversed. For $\Delta t < 0$, the coefficient $1 + \frac{\Delta t}{\Delta x}(-\frac{c}{2})$ changes sign when $\frac{c|\Delta t|}{2\Delta x} = -1$, or $C = -2$, which coincides with the unbounded amplification predicted by the modal analysis.

4. Modified Equation Method (with alternate marching). If, using the basic locally implicit method, we attempt to solve

$$p u_t + c u_x - \nu_2 \Delta x\, u_{xx} + \nu_4 \Delta x^3\, u_{xxxx} = 0, \tag{5}$$

for some p, we get instead the solution of

$$(p + \frac{\Delta t}{\Delta x}(-\frac{c}{2} + \nu_2 + 3\nu_4))u_t + c u_x - \nu_2 \Delta x\, u_{xx} + \nu_4 \Delta x^3\, u_{xxxx} = O(\Delta x + \Delta t).$$

Wanting the leading coefficient to be one, we solve for p,

$$p = 1 - \frac{\Delta t}{\Delta x}(-\frac{c}{2} + \nu_2 + 3\nu_4),$$

or with general θ,

$$p = 1 - \frac{\Delta t}{\Delta x}(-\frac{c}{2} + (2\theta - 1)(\nu_2 + 3\nu_4)).$$

Discretization of equation [5] with this particular p leads to the solution of the original equation [4] by a method which is automatically stable if only the lower frequencies are considered. Problems at higher frequencies can be handled by a combination of multigrid and artificial viscosity.

The modified equation method has desirable amplification factors for positive Courant Numbers and desirable group velocities for negative Courant Numbers. Therefore, following a forward sweep by a backward sweep

$$u_j^{n+1} = u_j^n - \frac{C}{2}(u_{j+1}^{n+1} - u_{j-1}^n) + V(u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^{n+1})$$

has the effect of providing a superior overall method. Use of $\theta = \frac{1}{2}$ rather than 1 appears to give identical results.

Incidentally, referring back to the basic method, alternate marching failed to be effective at large Courant Numbers because of the unbounded amplification factor for $C \leq -2$. These instabilities are too large to be damped during the alternate sweep.

5. Comparison with a Known Method. The Appendix contains a series of numerically produced tables for various methods. For each method there are two tables: amplification factors and group velocities. The rows of the tables correspond to (signed) Courant Numbers between -5 and $+5$. The columns correspond to frequencies, ωh, between 0 and π in multiples of $\pi/4$.

The amplification factor is the amplification effective in a single time step of the method. The group velocity is the effective numerical speed of propagation of a certain frequency in comparison to the speed for an exact solution.

The known method is central differencing with implicit Euler time stepping. It is, of course, a time accurate method, and the lowest frequency column shows group velocity values of 1.00. It is well known that the higher half of the frequency spectrum has negative group velocities; the values in the right half of the table are negative. The method is stable; the amplification factors are less than or equal to one.

A scheme would be perfect if all the values in both the group velocity and amplification tables were one. No such scheme exists.

Examining the amplification factors for the basic locally implicit method, see values that are greater than one for Courant Numbers less than -2; the scheme is

(von Neumann) unstable for $C < -2$. Examining group velocities, see that the low frequency components travel in the wrong direction for $C > 2$; the scheme is (GKS) unstable for these Courant numbers. For $-2 < C < 2$ the scheme fails to be time accurate; there are no ones in the first column.

For the simple modified equation method, observe that bad group velocities are massed at positive Courant Numbers and that bad amplification factors are at negative Courant Numbers. This motivates alternating the direction of march so that these effects cancel. The first columns show that the method is time accurate.

It is effective. Examining the group velocities, see that few are negative. See that few of the amplification factors are greater than one. A scant amount of artificial viscosity will make this a good method. Again, the first columns are ones, and the method is time accurate.

6. Aerodynamic Systems: Non-conservative Approach. This section chronicles the false starts made during the 1987 research. It was proposed that the matrix of characteristic vectors be used to diagonalize the system and to produce a decoupled system of scalar equations; the modified equation method was to be applied to each of the scalar equations. The proposed approach was found to possess a number of disadvantages: it is non-conservative, it is complicated and difficult to program, and it fails when there are two or more spatial variables.

Consider the system of conservation equations

$$u_t + f_x = 0.$$

This system is written in non-conservation law form as

$$u_t + Au_x = 0,$$

where A is the usual coefficient matrix[4]. Because these systems of equations are hyperbolic, there exists a matrix, Q, of characteristic vectors, that diagonalizes A. That is,

$$Q^{-1}AQ = \Lambda,$$

where Λ is a diagonal matrix with real entries.

The system of conservation laws is replaced by the non- conservative system

$$v_t + \Lambda v_x = 0,$$

where

$$v = Q^{-1}u.$$

This is a decoupled system of scalar equations and the method of the preceding section can be applied to each scalar equation.

This method fails when an additional spatial variable is introduced. Consider

$$u_t + f_x + g_y = 0.$$

This equation is written in non-conservative form as

$$u_t + Au_x + Bu_y = 0.$$

The method fails because the coefficient matrices A and B have different characteristic vectors. There is no longer a matrix Q which can diagonalize A and B simultaneously and decouple the system.

Although the characteristic vector approach was flawed, the research continued by attempting to maintain the notion of characteristic value (characteristic velocity or eigenvalue) which seemed necessary for the application of the modified equation method.

It is a fact from linear algebra[5] that each of the matrices A and B can be written in terms of their characteristic values as

$$A = \sum \lambda_i T_i,$$

where the λ's are the characteristic values and the T's are matrices of rank one. In fact, these matrices are known[4] because

$$T_i = l_i r_i^T,$$

where l_i and r_i are the left and right characteristic vectors of A associated with the characteristic value λ_i. These vectors are orthonormalized so that $l_i^T r_j = \delta_{ij}$.

Knowing that A and B can be written as

$$A = \sum \lambda_i T_i$$

$$B = \sum \mu_j S_j,$$

we again rewrite the original system

$$u_t + f_x + g_y = 0$$

in the form

$$u_t + \sum \lambda_i T_i u_x + \sum \mu_j S_j u_y = 0.$$

Although this system is not decoupled, we explored a corresponding scalar model problem for clues. Consider the scalar equation

$$u_t + cu_x + du_y = 0.$$

In studying this equation, it was found that application of the locally implicit method to

$$pu_t + cu_x + du_y = 0$$

actually solves the modified equation

$$(p - \frac{\Delta t}{2}(\frac{c}{\Delta x} + \frac{d}{\Delta y}))u_t + cu_x + du_y = O(\Delta x + \Delta y + \Delta t).$$

This result is easily extended to systems. Application of the locally implicit method to the system

$$Pu_t + Cu_x + Du_y = 0$$

actually solves the modified system

$$(P - \frac{\Delta t}{2}(\frac{C}{\Delta x} + \frac{D}{\Delta y}))u_t + Cu_x + Du_y = O(\Delta x + \Delta y + \Delta t).$$

It follows that

$$u_t + \sum \lambda_i T_i u_x + \sum \mu_j S_j u_y = 0$$

is solved by applying the locally implicit method to

$$Pu_t + \sum \lambda_i T_i u_x + \sum \mu_j S_j u_y = 0,$$

where P is chosen to be

$$P = I + \frac{\Delta t}{2}(\frac{\sum T_i}{\Delta x} + \frac{\sum S_j}{\Delta y}).$$

In this form, the locally implicit method is non-conservative and extremely cumbersome to understand and to program. Nevertheless, this form is very close to a method which is both simple and, in some sense, conservative.

7. Aerodynamic Systems: Conservative Approach. This section describes the method which evolved from the failures of the characteristic vector approach of the preceding section. The simplicity of the method seems to indicate the maturity of its evolution.

We want to solve

$$u_t + f_x + g_y = 0.$$

Considering the modified equation approach, apply the (right- marching) locally implicit discretization to

$$Pu_t + f_x + g_y = 0,$$

and get:

$$\frac{P(u_{j,k}^{n+1} - u_{j,k}^{n})}{\Delta t} + \frac{f_{j+1,k}^{n} - f_{j-1,k}^{n+1}}{2\Delta x} + \frac{g_{j,k+1}^{n} - g_{j,k-1}^{n+1}}{2\Delta y} = 0.$$

For each of the three finite difference expressions, substitute the corresponding Taylor approximation:

$$\frac{P(u_{j,k}^{n} + u_t \Delta t - u_{j,k}^{n})}{\Delta t},$$

$$\frac{f_{j,k}^{n} + f_x \Delta x - (f_{j,k}^{n} + f_t \Delta t - f_x \Delta x)}{2\Delta x}, \text{or}$$

$$\frac{g_{j,k}^{n} + g_x \Delta x - (g_{j,k}^{n} + g_t \Delta t - g_y \Delta y)}{2\Delta y}$$

to obtain

$$Pu_t + f_x - f_t \frac{\Delta t}{2\Delta x} + g_y - g_t \frac{\Delta t}{2\Delta y} = O(\Delta x + \Delta y + \Delta t).$$

The Chain Rule gives

$$f_t = Au_t$$

$$g_t = Bu_t$$

and therefore the modified equation, the equation actually solved by the locally implicit method, is

$$(P - \frac{\Delta t}{2}(\frac{A}{\Delta x} + \frac{B}{\Delta y}))u_t + f_x + g_y = O(\Delta x + \Delta y + \Delta t).$$

Therefore, the choice

$$P = I + \frac{\Delta t}{2}(\frac{A}{\Delta x} + \frac{B}{\Delta y})$$

gives the solution of the original equation

$$u_t + f_x + g_y = 0.$$

This method applies to time dependent systems in one, two, or three spatial dimensions. As shown earlier, the use of alternate marching directions with this locally implicit method gives numerical results which are comparable to that of an implicit central differencing scheme. In fact, the method was shown to be less dispersive and less dissipative.

In particular, this means the method is time accurate.

Being time accurate, the method is suitable for PNS (parabolized Navier-Stokes) calculations. In PNS calculations, one spatial variable is treated as the time variable. The locally implicit method permits the implicit PNS calculation to be replaced by an inexpensive explicit calculation.

Finally, this method avoids the approximate factorization

$$I - D_x - D_y \approx (I - D_x)(I - D_y)$$

required by the implicit methods in current use.

8. Numerical Example. The numerical solution for the two dimensional shock reflection problem is still forthcoming. The extension of the method as it was outlined in this proposal lead to an impossibly complicated algorithm; my student and I never got the program to produce meaningful results. The newer algorithm that was explained in the last section is expected to be more fruitful.

On another project the same student did produce an interesting parallel programming result. A subroutine of the Conjugate Gradient algorithm, which solves matrix problems, was rewritten in Ada for an eight processor parallel machine. Using seven of the eight processors, the algorithm was speeded up by a factor of 6.96 over its scalar speed. It is especially interesting that this particular algorithm cannot be vectorized to enhance its performance on a Cray.

9. Multigrid and the Locally Implicit Method. Both the multigrid and numerical objectives were subordinate to the theoretical objectives and were compromised. After more consideration, we believe the locally implicit method in its present form is not particularly well suited for a multigrid adaptation. The multigrid strategy is as follows:

1. Vary the grid so that middle frequency components on a fine grid are high frequency components on a coarse grid.

2. Couple this with a scheme that, by whatever means, speeds the time evolution of the high frequency components.

The locally implicit scheme is not well-suited to satisfy the second point; it retards the convective time evolution of the high frequency components. The success of some locally implicit multigrid examples may be due solely to accelerated dissipative time evolution.

10. Conclusions. The research toward a locally implicit method for systems of equations in several dimensions has been successful: a theoretical foundation has been established. This success came only after the expenditure of the majority of the research effort on the futile approach outlined in the proposal.

References

1. Jameson, Antony, "Solution of the Euler Equations for Two Dimensional Transonic Flow by a Multigrid Method," Applied Mathematics and Computation, 13 (1983) 327-355.

2. Trefethen, Lloyd, "Group Velocity in Finite Difference Schemes," SIAM Review, 24 (1982) 113-136.

3. Vichnevetsky,R. and Bowles,J, "Fourier Analysis of Numerical Approximations of Hyperbolic Equations," SIAM 1982.

4. Steger, Joseph L. and Warming, R.F., "Flux Vector Splitting of the Inviscid Gasdynamic Equations with Application to Finite Difference Methods," NASA TM-78605.

5. Golub, G.H. and Van Loan, C.F., "Matrix Computations", Johns Hopkins University Press, Baltimore (1983).

Appendix

Courant Number	Amplification Factors								
	Frequency--->								
5.000	1.00	0.46	0.27	0.21	0.20	0.21	0.27	0.46	1.00
4.500	1.00	0.50	0.30	0.23	0.22	0.23	0.30	0.50	1.00
4.000	1.00	0.55	0.33	0.26	0.24	0.26	0.33	0.55	1.00
3.500	1.00	0.60	0.37	0.30	0.27	0.30	0.37	0.60	1.00
3.000	1.00	0.66	0.43	0.34	0.32	0.34	0.43	0.66	1.00
2.500	1.00	0.72	0.49	0.40	0.37	0.40	0.49	0.72	1.00
2.000	1.00	0.79	0.58	0.48	0.45	0.48	0.58	0.79	1.00
1.500	1.00	0.87	0.69	0.59	0.55	0.59	0.69	0.87	1.00
1.000	1.00	0.93	0.82	0.73	0.71	0.73	0.82	0.93	1.00
0.500	1.00	0.98	0.94	0.91	0.89	0.91	0.94	0.98	1.00
0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
-0.500	1.00	0.98	0.94	0.91	0.89	0.91	0.94	0.98	1.00
-1.000	1.00	0.93	0.82	0.73	0.71	0.73	0.82	0.93	1.00
-1.500	1.00	0.87	0.69	0.59	0.55	0.59	0.69	0.87	1.00
-2.000	1.00	0.79	0.58	0.48	0.45	0.48	0.58	0.79	1.00
-2.500	1.00	0.72	0.49	0.40	0.37	0.40	0.49	0.72	1.00
-3.000	1.00	0.66	0.43	0.34	0.32	0.34	0.43	0.66	1.00
-3.500	1.00	0.60	0.37	0.30	0.27	0.30	0.37	0.60	1.00
-4.000	1.00	0.55	0.33	0.26	0.24	0.26	0.33	0.55	1.00
-4.500	1.00	0.50	0.30	0.23	0.22	0.23	0.30	0.50	1.00

CENTRAL DIFFERENCE WITH IMPLICIT EULER

Courant Number	Ratio of Group to Actual Velocity								
	Frequency--->								
5.000	1.00	0.20	0.05	0.02	-0.00	-0.02	-0.05	-0.20	-1.00
4.500	1.00	0.23	0.06	0.02	-0.00	-0.02	-0.06	-0.23	-1.00
4.000	1.00	0.28	0.08	0.03	-0.00	-0.03	-0.08	-0.28	-1.00
3.500	1.00	0.33	0.10	0.03	-0.00	-0.03	-0.10	-0.33	-1.00
3.000	1.00	0.40	0.13	0.04	-0.00	-0.04	-0.13	-0.40	-1.00
2.500	1.00	0.48	0.17	0.06	-0.00	-0.06	-0.17	-0.48	-1.00
2.000	1.00	0.58	0.24	0.09	-0.00	-0.09	-0.24	-0.58	-1.00
1.500	1.00	0.69	0.33	0.13	-0.00	-0.13	-0.33	-0.69	-1.00
1.000	1.00	0.81	0.47	0.21	-0.00	-0.21	-0.47	-0.81	-1.00
0.500	1.00	0.89	0.63	0.32	-0.00	-0.32	-0.63	-0.89	-1.00
-0.500	1.00	0.89	0.63	0.32	-0.00	-0.32	-0.63	-0.89	-1.00
-1.000	1.00	0.81	0.47	0.21	-0.00	-0.21	-0.47	-0.81	-1.00
-1.500	1.00	0.69	0.33	0.13	-0.00	-0.13	-0.33	-0.69	-1.00
-2.000	1.00	0.58	0.24	0.09	-0.00	-0.09	-0.24	-0.58	-1.00
-2.500	1.00	0.48	0.17	0.06	-0.00	-0.06	-0.17	-0.48	-1.00
-3.000	1.00	0.40	0.13	0.04	-0.00	-0.04	-0.13	-0.40	-1.00
-3.500	1.00	0.33	0.10	0.03	-0.00	-0.03	-0.10	-0.33	-1.00
-4.000	1.00	0.28	0.08	0.03	-0.00	-0.03	-0.08	-0.28	-1.00
-4.500	1.00	0.23	0.06	0.02	-0.00	-0.02	-0.06	-0.23	-1.00

CENTRAL DIFFERENCE WITH IMPLICIT EULER

Courant Number	Amplification Factors								
	Frequency--->								
5.00	1.00	0.99	0.97	0.96	0.95	0.95	0.95	0.95	0.95
4.50	1.00	0.99	0.97	0.96	0.95	0.95	0.95	0.95	0.95
4.00	1.00	0.99	0.97	0.96	0.95	0.95	0.95	0.95	0.95
3.50	1.00	0.98	0.96	0.96	0.95	0.95	0.95	0.95	0.95
3.00	1.00	0.97	0.96	0.96	0.96	0.95	0.95	0.95	0.95
2.50	1.00	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96
2.00	1.00	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
1.50	1.00	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97
1.00	1.00	0.99	0.98	0.98	0.98	0.98	0.97	0.97	0.97
0.50	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.98	0.98
0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.50	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.98	0.97
-1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.97	0.94	0.92
-1.50	1.00	1.00	1.00	1.00	1.00	0.99	0.97	0.92	0.80
-2.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.33
-2.50	1.00	1.00	1.00	1.00	1.00	1.01	1.02	1.09	1.57
-3.00	1.00	1.00	1.00	1.00	1.01	1.02	1.04	1.13	1.29
-3.50	1.00	1.00	1.00	1.01	1.01	1.03	1.06	1.13	1.22
-4.00	1.00	1.00	1.00	1.01	1.02	1.03	1.06	1.13	1.18

BASIC METHOD

Courant Number	Ratio of Group to Actual Velocity								
	Frequency--->								
5.000	-0.67	-0.60	-0.48	-0.40	-0.34	-0.31	-0.30	-0.29	-0.29
4.500	-0.80	-0.70	-0.54	-0.43	-0.37	-0.34	-0.32	-0.31	-0.31
4.000	-1.00	-0.82	-0.60	-0.47	-0.40	-0.36	-0.35	-0.34	-0.33
3.500	-1.33	-1.00	-0.66	-0.50	-0.43	-0.39	-0.38	-0.37	-0.36
3.000	-2.00	-1.20	-0.70	-0.53	-0.46	-0.43	-0.41	-0.40	-0.40
2.500	-4.00	-1.29	-0.68	-0.54	-0.49	-0.46	-0.45	-0.45	-0.44
2.000	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50
1.500	4.00	0.98	-0.09	-0.37	-0.48	-0.53	-0.56	-0.57	-0.57
1.000	2.00	1.30	0.38	-0.14	-0.40	-0.54	-0.62	-0.65	-0.67
0.500	1.33	1.12	0.64	0.15	-0.24	-0.50	-0.68	-0.77	-0.80
-0.500	0.80	0.77	0.68	0.50	0.24	-0.15	-0.64	-1.12	-1.33
-1.000	0.67	0.65	0.62	0.54	0.40	0.14	-0.38	-1.30	-2.00
-1.500	0.57	0.57	0.56	0.53	0.48	0.37	0.09	-0.98	-4.00
-2.000	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
-2.500	0.44	0.45	0.45	0.46	0.49	0.54	0.68	1.29	4.00
-3.000	0.40	0.40	0.41	0.43	0.46	0.53	0.70	1.20	2.00
-3.500	0.36	0.37	0.38	0.39	0.43	0.50	0.66	1.00	1.33
-4.000	0.33	0.34	0.35	0.36	0.40	0.47	0.60	0.82	1.00
-4.500	0.31	0.31	0.32	0.34	0.37	0.43	0.54	0.70	0.80

BASIC METHOD

Courant Number	Amplification Factors								
	Frequency--->								
5.00	1.00	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97
4.50	1.00	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97
4.00	1.00	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97
3.50	1.00	0.99	0.98	0.97	0.97	0.97	0.97	0.97	0.97
3.00	1.00	0.99	0.98	0.97	0.97	0.97	0.97	0.97	0.97
2.50	1.00	0.99	0.98	0.98	0.97	0.97	0.97	0.97	0.97
2.00	1.00	0.99	0.98	0.98	0.98	0.97	0.97	0.97	0.97
1.50	1.00	1.00	0.99	0.98	0.98	0.98	0.98	0.98	0.98
1.00	1.00	1.00	0.99	0.99	0.98	0.98	0.98	0.98	0.98
0.50	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
-0.50	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.97	0.96
-1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
-1.50	1.00	1.00	1.00	1.01	1.02	1.04	1.07	1.11	1.13
-2.00	1.00	1.00	1.01	1.02	1.04	1.06	1.07	1.08	1.08
-2.50	1.00	1.01	1.02	1.03	1.05	1.06	1.06	1.07	1.07
-3.00	1.00	1.01	1.02	1.04	1.05	1.06	1.06	1.06	1.06
-3.50	1.00	1.01	1.03	1.04	1.05	1.05	1.06	1.06	1.06
-4.00	1.00	1.01	1.03	1.04	1.05	1.05	1.05	1.05	1.05

MODIFIED EQUATION METHOD

Courant Number	Ratio of Group to Actual Velocity								
	Frequency--->								
5.000	1.00	0.31	-0.00	-0.10	-0.14	-0.15	-0.16	-0.17	-0.17
4.500	1.00	0.36	0.01	-0.10	-0.14	-0.16	-0.18	-0.18	-0.18
4.000	1.00	0.40	0.03	-0.10	-0.15	-0.18	-0.19	-0.20	-0.20
3.500	1.00	0.46	0.05	-0.10	-0.16	-0.20	-0.21	-0.22	-0.22
3.000	1.00	0.52	0.08	-0.10	-0.18	-0.22	-0.24	-0.25	-0.25
2.500	1.00	0.58	0.13	-0.09	-0.19	-0.24	-0.27	-0.28	-0.29
2.000	1.00	0.65	0.19	-0.07	-0.20	-0.27	-0.31	-0.33	-0.33
1.500	1.00	0.72	0.28	-0.03	-0.21	-0.31	-0.36	-0.39	-0.40
1.000	1.00	0.80	0.39	0.04	-0.20	-0.35	-0.44	-0.49	-0.50
0.500	1.00	0.86	0.54	0.16	-0.15	-0.39	-0.55	-0.64	-0.67
-0.500	1.00	0.97	0.88	0.70	0.40	-0.08	-0.78	-1.59	-2.00
-1.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
-1.500	1.00	1.01	1.04	1.10	1.20	1.36	1.59	1.86	2.00
-2.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
-2.500	1.00	0.97	0.91	0.83	0.77	0.72	0.69	0.67	0.67
-3.000	1.00	0.93	0.80	0.68	0.60	0.55	0.52	0.50	0.50
-3.500	1.00	0.88	0.69	0.56	0.48	0.44	0.42	0.40	0.40
-4.000	1.00	0.82	0.60	0.47	0.40	0.36	0.35	0.34	0.33
-4.500	1.00	0.77	0.52	0.40	0.34	0.31	0.30	0.29	0.29

MODIFIED EQUATION METHOD

Courant Number	Amplification Factors								
	Frequency--->								
5.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01
4.50	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01
4.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01
3.50	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01
3.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.02
2.50	1.00	1.00	1.00	1.00	1.01	1.01	1.02	1.02	1.02
2.00	1.00	1.00	1.00	1.00	1.01	1.02	1.02	1.03	1.03
1.50	1.00	1.00	1.00	1.00	1.00	1.01	1.02	1.04	1.05
1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
0.50	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.98	0.97
0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
-0.50	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.98	0.97
-1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
-1.50	1.00	1.00	1.00	1.00	1.00	1.01	1.02	1.04	1.05
-2.00	1.00	1.00	1.00	1.00	1.01	1.02	1.02	1.03	1.03
-2.50	1.00	1.00	1.00	1.00	1.01	1.01	1.02	1.02	1.02
-3.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.02
-3.50	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01
-4.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01

MODIFIED EQUATION METHOD WITH ALTERNATE MARCHING

Courant Number	Ratio of Group to Actual Velocity								
	Frequency--->								
5.000	1.00	0.51	0.22	0.12	0.08	0.06	0.05	0.04	0.04
4.500	1.00	0.56	0.26	0.15	0.10	0.07	0.06	0.05	0.05
4.000	1.00	0.61	0.31	0.18	0.12	0.09	0.08	0.07	0.07
3.500	1.00	0.67	0.37	0.23	0.16	0.12	0.10	0.09	0.09
3.000	1.00	0.72	0.44	0.29	0.21	0.17	0.14	0.13	0.13
2.500	1.00	0.78	0.52	0.37	0.29	0.24	0.21	0.20	0.19
2.000	1.00	0.82	0.60	0.47	0.40	0.36	0.35	0.34	0.33
1.500	1.00	0.87	0.66	0.53	0.50	0.53	0.61	0.74	0.80
1.000	1.00	0.90	0.69	0.52	0.40	0.33	0.28	0.26	0.38
0.500	1.00	0.92	0.71	0.43	0.12	-0.23	-0.66	-1.11	-1.33
-0.500	1.00	0.92	0.71	0.43	0.12	-0.23	-0.66	-1.11	-1.33
-1.000	1.00	0.90	0.69	0.52	0.40	0.33	0.28	0.26	0.13
-1.500	1.00	0.87	0.66	0.53	0.50	0.53	0.61	0.74	0.80
-2.000	1.00	0.82	0.60	0.47	0.40	0.36	0.35	0.34	0.33
-2.500	1.00	0.78	0.52	0.37	0.29	0.24	0.21	0.20	0.19
-3.000	1.00	0.72	0.44	0.29	0.21	0.17	0.14	0.13	0.13
-3.500	1.00	0.67	0.37	0.23	0.16	0.12	0.10	0.09	0.09
-4.000	1.00	0.61	0.31	0.18	0.12	0.09	0.08	0.07	0.07
-4.500	1.00	0.56	0.26	0.15	0.10	0.07	0.06	0.05	0.05

MODIFIED EQUATION METHOD WITH ALTERNATE MARCHING

1986 USAF-UES MINIGRANT PROGRAM

Sponsored by the
Air Force Office of Scientific Services
Bolling AFB, D.C.

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT

FLUORESCENT DYE BINDING IDENTIFICATION OF ASBESTOS ON MEMBRANE FILTERS AND IN BULK MATERIALS

Prepared by : Cliff Houk, Ph.D.
Academic Rank : Professor
Department and
University : School of Health and Sport Sciences
(Industrial Hygiene/Environmental Health)
Ohio University
Athens, Ohio 45701-2979

Research Location: Athens, Ohio
Date : 3 Feb. 1988

Contract No. : F49620-85-C-0013/SB5851-0360

Introduction:

The fluorescent dye binding properties of several organic dyes for the identification of asbestos in bulk materials and on membrane filters have been studied [1,2] and compared to previously reported results.[3,4,5] Morin and Clayton-Yellow dyes exhibit easily detected fluorescence on bulk samples containing chrysotile asbestos. A preliminary procedure for the specific identification of chrysotile in bulk materials was established.

Fluorescence microscopy data suggested that Morin and Clayton-Yellow bind to other forms of asbestos that may be found on membrane filters. The intensity of the fluorescence produced by other forms of asbestos in bulk materials is too weak to observe with the naked eye but appears detectable with the aid of a photomultiplier tube (PMT) attachment on the fluorescence microscope.

Research Objectives:

(1) To continue the study of the bulk sample analytical procedure to determine its feasibility as a rapid, cost effective, accurate and easy to use field test by Air Force personnel.

(2) To continue the study of fluorescent dye binding identification of asbestos fibers on membrane filters using fluorescent dyes as tags with a combination of epi-fluorescence and phase contrast techniques.

(3) To investigate the kinetics of the binding between asbestos and fluorescent dyes in an effort to quantitate the relationship between observed fluorescence and dye concentration, asbestos concentration and time using spectrofluorometric analysis of aqueous dispersions of "tagged" asbestos.

Goals:

(1) To develop a rapid, accurate, reliable, easy to use and inexpensive field test suitable for the qualitative identification of asbestos in bulk materials by Air Force personnel.

(2) To establish an acceptable procedure for the qualitative identification and counting of asbestos fibers on membrane filters.

(3) To determine the quantitative relationships between dye concentration, asbestos concentration and time and the measured level of fluorescence that may be transferable to membrane filter analyses and answer the questions, "Is it asbestos?" and "How much of it is present?"

Experimental:

(1) Several hundred bulk samples previously screened for asbestos content by PLM will be tested using a procedure initiated at OEHL/SA during the Summer, 1986.

(2) Refinement of the steps in the procedure: sample size, sample treatment, dye concentration, limits of detection will be undertaken. Sampling and analytical protocol will be established for Air Force personnel.

(3) Qualitative identification of asbestos fibers on membrane filters and neat dispersions will be undertaken using a Leitz Dialux-20 microscope equipped to do epi-fluorescence and phase contrast analyses.

(4) Quantitation will be undertaken utilizing existing equipment, spectrofluorimeter and reflectance fluorometers to establish relationships between dye concentration, asbestos concentration, type of asbestos present and time and the observed level of fluorescence.

Results:

Bulk Sample Analyses

A total of 1,014 bulk samples previously identified by PLM analysis were examined under several variations of a scheme reported earlier.[1,7] The chrysotile content of these samples that were positive by PLM analysis ranged from <1% - >75%.

Table I shows the overall results.

TABLE I - TOTAL BULK ANALYSIS RESULTS

No. Samples	No. Correct	% Accuracy	False -	False +
1,014	870	86	70	74

An 86% accuracy may appear less than desirable but it still exceeds that of the current commercial product available today.[4,7] However, when the data are broken out to illustrate the effects of dye concentration and volume of reagent solution used marked improvements in accuracy appear. Tables II-V demonstrate those effects.

Table II represents the results using 1.5 mL total volume of dye and buffer (0.75 mL dye, 0.75 mL buffer) and dye concentration ≥ 0.00001%.

TABLE II - BULK ANALYSIS 1.5 mL Solution, ≥ .00001% dye

No. Samples	No. Correct	% Accuracy	False -	False +
475	472	99	2	1

The two false negative results occurred with a dye concentration of 0.00001% on samples $\leq$ 10% chrysotile.

Steps were taken to investigate reducing the volumes of reagents and concentration of dye utilized in the procedure. Previous efforts had used separate solutions of buffer and dye that were added individually to the bulk sample. Buffer and dye solutions were mixed prior to sample analysis with 0.5 mL of the mixture added directly to the sample. Table III contains the results of that change in procedure.

TABLE III - BULK SAMPLE RESULTS (0.5 mL; 0.00005-0.000075%)

No. Samples	No. Correct	% Accuracy	False -	False +
109	80	73%	3	26

It appeared that the solution volume and dye concentration may be too small so the reagent volume of mixed dye and buffer solutions was increased to 1.0 mL with a final dye concentration of 0.00005%. Table IV contains the results of that change in procedure.

TABLE IV- BULK SAMPLE RESULTS (1.0 mL; 0.00005%)

No. Samples	No. Correct	% Accuracy	False -	False +
360	248	69%	65	47

Data suggested the dye concentration was too low so the dye concentration was increased.

Table V contains the results of that change in procedure.

TABLE V - BULK SAMPLE RESULTS (1.0 mL; 0.00005-0.00015%)

No. Samples	No. Correct	% Accuracy	False -	False +
70	70	100%	0	0

It appears that at least 1.0 mL of reagent solution and a dye concentration of $\geq$ 0.0001% are necessary to approach 100% accuracy.

The eleven step procedure reported earlier has been reduced to four steps:

1) Place 10 mg sample in 1.5 mL, capped, plastic, microcentrifuge test tube.

2) Add 1.0 mL mixed buffer/dye solution, close cap and shake.

3) Expose test tube to UV light against a black background.

4) Observe fluorescence.

Qualitative/Quantitative:

Epi-fluorescent examination of membrane filters was not accomplished because the proper set of filters (exciter, barrier and dichroic) was not available for the Leitz microscope. Filters for another microscope are being ordered to continue this phase of the study.

Preliminary spectrofluorimetric investigation of the quantitative relationships between dye concentration, asbestos concentration, asbestos type with time and intensity were started. Results are inconclusive due to the wavelength limitation of the spectrofluorimeter's PMT. The dye under investigation fluoresces in the red region of the spectrum, >700 nm which is beyond the detection limit of the PMT. An extended range PMT has been ordered. This aspect of the study will be continued upon installation of the necessary PMT.

Conclusion:

The chrysotile bulk sample analytical procedure has been simplified and could be utilized by Air Force personnel following additional field tests outside the laboratory.

Further study is required to identify additional dyes that are specific for other forms of asbestos in bulk materials and to test the

application of the "dye tagging" method to asbestos fibers on membrane filters.

References

1. Houk, C.C.: Fluorescent Dye Binding Analysis for the Identification of Asbestos. Final Report, USAF-UES/SFRP, Aug. 1986.

2. Houk, C.C.: Fluorescent Dye Binding Identification of Asbestos on Membrane Filters and in Bulk Material. AIHC, Montreal, Canada, May 1987.

3. Albright, F.R., D.V. Schumacher, B.J. Feltz and J.A. O'Donnell: A Fluorescent Dye Binding Technique for Detection of Chrysotile Asbestos. Microscope 30:267-280 (1982).

4. Sperduto, B., F. Burrogato, A. Altier and M. Gasperetti: The Asbestos Minerals: Their Identification and Determination. Ann. Inst. Super. Sanita 13:127-136 (1977).

5. Burragato, F. and B. Sperduto: Microscopic Quantitative Determination of Presence of Quartz in Talc by Means of Colouring. Prospects for its use in the Determination of Silicosis Hazards. Securitas 59:939-946 (1974).

6. Baldwin, C.A., H.J. Beaulieu and R.M. Buchan: Application of the [illegible] Asbestos Screening Test in Colorado Schools. Am. Ind. Hyg. Assoc. J. 43:602-604 (1983).

7. Oestenstad, P.K. and V.E. Rose: An Evaluation of the [illegible] Asbestos Screening Test. Am. Ind. Hyg. Assoc. J. 47:245-248 (1986).

FINAL REPORT NUMBER 36
REQUESTED A NO-COST TIME EXTENTSION
TO BE SUBMITTED IN 1987 MINI-GRANT FINAL REPORT
Dr. Ming S. Hung
760-6MG-105

FINAL REPORT NUMBER 37
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. John Jobe
760-6MG-019

FINAL REPORT
Contract No. F49620-85-C-0013/SB5851-0360
Purchase Order No. S-760-6MG-016

"EXPERT SYSTEM FOR OPTIMAL DESIGN"

submitted to

Rodney C. Darrah
Universal Energy Systems
4401 Dayton-Xenia Road
Dayton, OH 45432

November 6, 1987

submitted by

Glen E. Johnson
Mechanical & Materials Engineering
Vanderbilt University
Nashville, TN 37235

This report was published as "A General Approach to Constrained Optimal Design Based on Symbolic Mathematics," C. R. Hammond and G. E. Johnson, in Advances in Design Automation, ed. by S. S. Rao, American Society of Mechanical Engineers, NY, 1987, pp 31-40.

ABSTRACT

This paper presents a new method of optimal design that combines ideas and techniques from both the Method of Optimal Design (MOD) and monotonicity analysis. This new method, while presently limited to constrained monotonic problems, has been automated. The problem is first reduced in size and reformulated using ideas from MOD. The reduced problem is then repeatedly reformulated to develop state and design equations in terms of all possible variable partitions. The complete set of candidate optimal solution points is extracted by monotonicity analysis applied to the trivial constraint sets on the design equations from every possible formulation of the problem. These candidate solution points can then be numerically sorted to determine the optimal solution for a particular numerical case. The new method and its automation are illustrated with two example problems from the literature.

INTRODUCTION

Most non-numerical methods of optimal design have as their basis either the Method of Optimal Design (MOD) developed by R. Johnson (1967, 1978, 1979, 1980), or Monotonicity Analysis introduced by Wilde (1975, 1976) and extended by Papalambros and Wilde (1979, 1980). These methods have been difficult to automate since many of the analysis steps depend on the problem. Zhou and Mayne (1983) developed a program that numerically made the monotonicity decisions and identified active constraints. The user then interactively eliminated the active constraints. Azarm and Papalambros (1984a, 1984b) presented a knowledge based system that uses both local and global information to determine the active set of constraints. Li and Papalambros (1985) developed a production system that incorporates knowledge about the constraints to make deductions about possible active sets.

This paper presents a new method of solution extraction for constrained monotonic optimization problems. This new method is based on the Method of Optimal Design and has been automated using a symbolic math computer program. The method works by developing problem formulations for all of the possible variable partition. Analysis of the design equations in these formulations leads to the complete set of candidate solution points.

Optimization Problem Statement

A general design optimization problem that has been reduced by the Method of Optimal Design can be stated in the form:

Minimize or Maximize

$$f(\underline{x})$$

subject to

$$y_i = g_i(\underline{x}) \quad i = 1,2,\ldots,N_S$$
$$y_{iMIN} <= y_i <= y_{iMAX} \quad i = 1,2,\ldots,N_S$$
$$x_{jMIN} <= x_j <= x_{jMAX} \quad j = 1,2,\ldots,N_D$$

where $y_i = (y_1, y_2,\ldots, y_{NS})$ are the N_S state variables (called "eliminated parameters" in MOD) and $x_j = (x_1, x_2,\ldots, x_{ND})$ are the N_D decision variables (called "related parameters" in MOD). Both the design equation, $f(\underline{x})$, and the state equations, $g_i(\underline{x})$, are in terms of the decision variables.

There are two restrictions on the variables in the method presented in this paper. The first is that the variables are always positive. This is not really a limitation since most engineering variables are always positive or can be transformed so that they are positive. This restriction facilitates the trend or monotonicity analysis and helps in choice of roots when there are multiple roots of an equation. (For example, X^2-1 has roots X=1 and X=-1, This restriction defines x=1 as the only feasible root). The second restriction is that all variables have either both upper and lower limits or no limits. Variables that do not have both upper and lower bounds can be artificially constrained by the designer. (For example a lower limit might be zero, and an upper limit might be some very large positive real number). Variables without any bounds are "free variables" and can be eliminated from the problem.

BASIS FOR THE NEW METHOD OF OPTIMAL DESIGN

A discussion of this new method of solution extraction first requires definition of a few terms:

Vertex:

The point of intersection of N_{DOF} constraints. This point is constraint bound since there are zero degrees of freedom.

Solution Point:

The vertex that is the optimal solution for a specific numerical case.

Candidate Solution Points:
The set of potential solution points. This set is the same for every numerical case. All vertices are not candidate solution points.

Partition:
A division of the set of variables into state and decision variables.

Form of an equation or of the problem:
The variables in an equation or problem are arranged so that the design and state variables are in terms of the decision variables, i.e., the variables are partitioned.

Statement of the New Method of Optimal Design

The set of candidate solution points in constrained monotonic problems can be determined from a trend or monotonicity analysis on all possible forms of the design equation.

Proof Outline for the New Method of Optimal Design

This proof outline holds only for constrained monotonic problems.

1) All candidate solution points occur at vertices since there are no interior points (Wilde, 1975).

2) The constraints on a set of N_{DOF} variables can have many vertices. Of these vertices, only one can be a candidate solution point.

3) The number of degrees of freedom in a problem equals the number of decision variables in that problem: $N_{DOF} = N_D$.

4) There are $A_t = N_V!/(N_D! * N_S!)$ possible unique sets of N_{DOF} variables in a problem (R. Johnson, 1978, 1979, 1980). Thus the variables in a problem can be partitioned into A_t unique sets of state and decision variables and analysis of each partition will lead to the set of candidate solution points.

5) The decision variables in the design equation form the trivial constraint set (i.e., the upper and lower bounds on the decision variables). Ignoring the non-trivial constraints, the trivial constraint set encloses the feasible region. Trend analysis on the design equation identifies the vertex that is the unique optimal solution point in this region. This solution point is one of the set of candidate solution points.

6) Trend analysis of the design equations in all partitions leads to the complete set of candidate solution points.

This new method will first be demonstrated using the pellet production problem.

Example 1: Pellet Production Problem (G. Johnson and Townsend, 1979)

One of the machines involved in the manufacture of plastic pellets is the pelletizer which cuts the strands of plastic into pellets. In this example the pelletizer will be studied to maximize its production rate, Eq. (1), with limits on the diameter and length of the pellets, the number of strands to be cut, the speed of the blades, and the power required to cut the strands, Eq. (2).

Maximize V:

$$V = \left[1 - \frac{0.01\ n}{n_{MAX}} \right] \left(15\ l\ \pi\ \rho\ B\ D^2\ \Omega\ n \right) \frac{KG}{HR} \qquad (1)$$

Subject to:

$$P = 6.6*10^{-6}\ n\ D^3\ \tau\ B\ \Omega \qquad (2)$$

$$P_{MIN} <= P <= P_{MAX}$$

$$D_{MIN} <= D <= D_{MAX}$$

$$l_{MIN} <= l <= l_{MAX}$$

$$\Omega_{MIN} <= \Omega <= \Omega_{MAX}$$

$$n : \text{Integer}, \quad 1..n_{MAX}$$

l is a "truly independent variable" and the optimal value of l is l_{MAX}. It is expedient to solve the problem for fixed n, up to n_{MAX} times. There are two decision variables, D and Ω, so there are two degrees of freedom. The state variable is P.

The problem can be reduced to:

$$\text{Max} \qquad V = A_1 \, A_2 \, D^2 \, \Omega \tag{3}$$

$$\text{Subject to:} \quad P = A_2 \, A_3 \, D^3 \, \Omega \tag{4}$$

where

$$A_1 = \left[1 - \frac{0.01 \, n}{n_{MAX}} \right] (15 \, l_{MAX} \, \pi \, \rho) \tag{5}$$

$$A_2 = B \, n \tag{6}$$

$$A_3 = 6.6 * 10^{-6} \, \tau \tag{7}$$

subject to:

$$P_{MIN} <= P <= P_{MAX}$$

$$D_{MIN} <= D <= D_{MAX}$$

$$\Omega_{MIN} <= \Omega <= \Omega_{MAX}$$

The set of candidate solution points to this problem is:

$$((D_{MAX}, \Omega_{MAX}), (P_{MAX}, \Omega_{MAX}), (P_{MAX}, D_{MIN}))$$

These candidate solution points will now be developed using the new method. In a problem with one state variable and two decision variables, there are

$$N_V!/(N_S! \, N_D!) = 3!/(2! * 1!) = 3$$

different partitions of the variables. The three partitions are shown in Table 1.

Table 1. Partitions of the Pellet Problem

Partition	STATE VARIABLE	DECISION VARIABLE
1	P	D Ω
2	D	P Ω
3	Ω	P D

The problem is presently in the form of Partition #1 - V and P are in terms of D and Ω.

$$V = A_1 A_2 D^2 \Omega \tag{8}$$

$$P = A_2 A_3 D^3 \Omega \tag{9}$$

Trend analysis on Eq. (8) shows that to maximize V, both D and Ω should be set to their maximum values: (D_{MAX}, Ω_{MAX}). This is the first of the candidate solution points in the set.

In the second partition, V and D are in terms of P and Ω. To develop this partition, Eq. (9) is solved for D:

$$D = \left[\frac{P}{A_2 A_3 \Omega}\right]^{1/3} \tag{10}$$

and substituted into Eq. (8) to eliminate D:

$$V = A_1 \left[\frac{P^2 \Omega A_2}{A_3^2}\right]^{1/3} \tag{11}$$

The candidate solution point, (P_{MAX}, Ω_{MAX}) corresponding to this partition, can be determined by inspection of Eq. (11).

The last of the candidate solution points, (P_{MAX}, D_{MIN}), is determined from the third partition. Eq. (9) is solved for Ω:

$$\Omega = \frac{P}{D^3 A_2 A_3} \tag{12}$$

and substituted into Eq. (8) to give:

$$V = \frac{P\ A_1}{D\ A_3} \qquad (13)$$

Thus, the three candidate solution points,

$$\begin{array}{l}((D_{MAX},\ \Omega_{MAX}),\\ (\ P_{MAX},\ \Omega_{MAX}),\\ (\ P_{MAX},\ D_{MIN}))\end{array}$$

have been determined by development and analysis of all of the forms of the design equation. This is summarized in Table 2.

Table 2. Pellet Problem Solutions

Equation	Solution
$V = A_1\ A_2\ \ D^2\ \Omega$	$D_{MAX}\ \Omega_{MAX}$
$V = A_1 \left[\frac{P^2\ \Omega\ A_2}{A_3^{\,2}} \right]^{1/3}$	$P_{MAX}\ \Omega_{MAX}$
$V = \frac{P\ A_1}{D\ A_3}$	$P_{MAX}\ D_{MIN}$

The steps of a numerical analysis of this problem for a specific numerical case could be as follows:

Calculate

$$A_1 = \left[1 - \frac{0.01\ n}{n_{MAX}} \right] (15\ l_{MAX} * \rho) \qquad (14)$$

$$A_2 = B\ n \qquad (15)$$

$$A_3 = 6.6*10^{-6}\ r \qquad (16)$$

From partition 1,

$$P = A_2\ A_3\ \ D_{MAX}{}^3\ \Omega_{MAX} \qquad (17)$$

If $P_{MIN} <= P <= P_{MAX}$ then

Save $V = A_1 A_2 D_{MAX}^2 \Omega_{MAX}$ (18)

as a feasible candidate solution point.

From partition 2,

$$D = \left[\frac{P_{MAX}}{A_2 A_3 \Omega_{MAX}} \right]^{1/3} \quad (19)$$

If $D_{MIN} <= D <= D_{MAX}$ then

Save $V = A_1 \left[\frac{P_{MAX}^2 \Omega_{MAX} A_2}{A_3^2} \right]^{1/3}$ (20)

as a feasible candidate solution point.

From partition 3,

$$\Omega = \frac{P_{MAX}}{D_{MIN}^3 A_2 A_3} \quad (21)$$

If $\Omega_{MIN} <= \Omega <= \Omega_{MAX}$ then

Save $V = \frac{P_{MAX} A_1}{D_{MIN} A_3}$ (22)

as a feasible candidate solution point.

The best of the feasible candidate solution points is selected as the optimal solution. Sometimes, none of the candidate solution points are feasible. This occurs when a feasible region does not exist for the numerical case.

AUTOMATION

The method of analysis presented in this paper has been automated with the symbolic math computer package muMATH 83 on IBM PC and AT micro-computers. muMATH 83 manipulates equations symbolically to solve for variables, simplify equations, differentiate, take absolute values, and perform other needed mathematical tasks. The automation of the method consists of three distinct phases:

1) Conversion of the problem into a reduced form.

2) Development of all partitions.

3) Trend or monotonicity analysis of the design equations.

Each of these will be discussed in detail and demonstrated with Example 2.

Example 2: Torsion Bar Design (R. Johnson, 1967, 1980)

This problem involves the design of a round torsion bar for use as a spring subjected to repeated loadings. It is desirable to minimize the twisting moment, M_t, applied to the machine by the torsion spring for a given amount of energy, PE_s, to be absorbed by the spring. The length, L, and diameter, d, of the bar are limited due to space restrictions. Functional requirements limit the angle of twist, θ, and bar shaft will be designed to prevent fatigue failure with a reasonable factor of safety, N.

Minimize

$$M_t = \frac{2\ PE_s}{\theta} \tag{23}$$

Subject to:

$$\theta = \frac{32\ M_t\ L}{\pi\ d^4\ G} <= \theta_{MAX} \tag{24}$$

$$\tau max = K\ \frac{16\ M_t}{\pi\ d^3} <= \frac{S_e}{(1 + p)\ N} \tag{25}$$

$$L_{MIN} <= L <= L_{MAX}$$

$$d <= d_{MAX}$$

R. Johnson has shown that this problem can be reduced to:

Minimize

$$M_t = \left[\frac{d^4\ PE_s\ *\ G}{16\ L}\right]^{1/2} \tag{26}$$

State Equations:

$$\theta = \left[\frac{64\ PE_s\ L}{d^4\ *\ G}\right]^{1/2} \tag{27}$$

$$\tau max = \left[\frac{16\ K^2\ PE_s\ G}{d^2\ L\ *}\right]^{1/2} \tag{28}$$

Subject to:

$$\theta_{MIN} <= \theta <= \theta_{MAX}$$

$$\tau max_{MIN} <= \tau max <= \tau max_{MAX}$$

$$L_{MIN} <= L <= L_{MAX}$$

$$d_{MIN} <= d <= d_{MAX}$$

Where:

$$\tau max_{MAX} = \frac{S_e}{(1 + p)\ N} \qquad (29)$$

He has also shown that the candidate solution points of this problem are:

Table 3. Torsion Bar Solutions

((d_{MIN}, L_{MAX}),
(θ_{MAX}, any feasible d),
(d_{MIN}, τmax_{MIN}),
(θ_{MAX}, any feasible L),
(L_{MAX}, τmax_{MAX}),
(θ_{MAX}, any feasible τmax)).

Three of the solutions, (θ_{MAX}, any feasible d), (θ_{MAX}, any feasible L), and (θ_{MAX}, any feasible τmax)) are lines rather than points. Along the constraint $\theta = \theta_{MAX}$, the twisting moment, M_t, is dependent only on θ. Thus any feasible point along this constraint can be an optimal solution. This problem can be very difficult to solve using numerical methods since there can be an infinite number of equally optimum solution points.

Conversion of the problem into a reduced form

Analysis of an optimization problem by MOD or the new method requires a reformulation of the problem so that the state and design variables are in terms of the decision variables. In MOD, this reformulation is referred to as the conversion of the Initial Formulation to the Final Formulation.

There are two steps in this reformulation. In the first step, free variables, which do not have limits, are eliminated from the problem (R. Johnson, 1979, 1980). Elimination of the free variables reduces the number of state equations. The second step is the manipulation of the objective function and state equations so that they are in terms of the decision variables.

The interaction of the program with the user as Example 2. is entered into the computer is shown in Figure 1. Once the problem has been entered into the computer, the program displays statistics about the current form of the problem and makes recommendations for the reformulation of the problem. Figure 2. shows the information displayed immediately after entering the problem into the computer. It makes the recommendation that the design variable be eliminated from the state equations (SDE's) using AUTO(). AUTO() eliminates both the free variables from the problem and the design variables from the state equations. This was done in Figure 3. In Figure 3. the program recommends that the equations be manipulated so that θ appears on only one side of the equations. This was done as shown in Figure 4. using MAKESTATE(THETA). After this has been completed, the program determines that the problem is in a final reduced form and recommends solution extraction using DOIT().

Once the problem has been reformulated, then further analysis can reduce the problem size. Some variables may appear only in the design equation and the limit equations. R. C. Johnson (1978, 1980) refers to these variables as "truly independent". No matter how a problem is reformulated, it will always trend in the same direction on these variables. Trend analysis on any design equation will determine the partial solution of the problem for these variables. l in the pellet problem is a "truly independent" variable, it was set equal to l_{MAX}.

Redundant state equations can also be removed from the problem (Wilde, 1975). An example of a redundant state equation is:

$$\text{Volume} = X^3 \tag{30}$$

$$X_{MIN} <= X <= X_{MAX} \tag{31}$$

$$\text{Volume} <= \text{Volume}_{MAX} \tag{32}$$

In this case, the upper limit on the volume is really only an upper limit on X. The true upper limit on X is the smaller of X_{MAX} and $\text{Volume}_{MAX}^{1/3}$. The state equation for the Volume can then be removed from the problem, reducing the number of state equations by one.

Once the problem has been reformulated, then the program identifies and eliminates redundant state equations and analyzes "truly independent" variables. Example 2 does not have either redundant state equations or "truly independent" variables.

Development of all partitions

In step one, the problem was reformulated so that there is a distinct division between state and decision variables. This set of state and decision variables forms one of the partitions that the problem can be reformulated into. In this problem, there are $A_t = N_V!/(N_D!\ N_S!) = 4!/(2!\ 2!) = 6$ possible unique partitions of state and decision variables and six candidate solution points.

An existing partition can be reformulated into another partition by interchanging one of its state variables with a decision variable. This is done by solving the state equation for the decision variable and using this new equation to eliminate the decision variable from the other equations in the formulation. This interchange of variables can be demonstrated by interchanging τ_{max} and d in the torsion bar problem:

Eq. (28) can be solved for d to give:

$$d = \left[\frac{16\ K^2\ PE_s\ G}{\tau_{max}^2\ L\ \pi} \right]^{1/2} \qquad (33)$$

Eq. (33) is then used to eliminate d in Eq. (26) and Eq. (27) resulting in a new partition with state variables d and θ in terms of the decision variables L and τ_{max}:

$$M = \left[\frac{16\ G^3\ K^4\ PE_s^3}{L^3\ \tau_{max}^4\ \pi} \right]^{1/2} \qquad (34)$$

$$\theta = \left[\frac{L^3\ \tau_{max}^4\ \pi}{4\ G^3\ K^4\ PE_s} \right]^{1/2} \qquad (35)$$

$$d = \left[\frac{16\ G\ K^2\ PE_s}{L\ \tau_{max}^2\ \pi} \right]^{1/2} \qquad (36)$$

An orderly procedure for developing all of the partitions based on the interchange of variables has been developed. In this procedure a table of the variable partitions is used to determine partitions that are suitable for modification into new partitions. This is shown in Table 4. for the torsion bar problem.

Table 4. Variable Partitions of the torsion bar problem.

#	STATE VARIABLES	DECISION VARIABLES	LEVEL	PARTITION - INTERCHANGE
1	θ τmax	d L	0	none
2	L τmax	θ d	1	# 1 - θ L
3	θ L	τmax d	1	# 1 - τmax L
4	d τmax	θ L	1	# 1 - θ d
5	θ d	τmax L	1	# 1 - τmax d
6	d L	θ τmax	2	# 2 - τmax d # 3 - θ d # 4 - τmax L # 5 - θ L

The six partitions are grouped in three levels. The partition on level zero is the original partition. The partitions on level one have had one decision variable interchanged with one state variable. The partition on level two has had two decision variables interchanged with two state variables. Each level has one less of the number of original decision variables then the previous level.

New partitions are developed by interchanging a state and a decision variable in an existing partition. Thus, a new partition will have one less of the original decision variables then the partition that it is developed from. It follows that new partitions are developed by modifying a partition from the level below that of the desired partition.

The modifications needed to develop a new partition from an existing partition can be determined by a comparison of the the sets of decision variables from the two partitions. Inspection of the two sets shows that all of the variables but one in a set are common with the variables in the other set. A partition is modified to produce a new partition by interchanging a state with a decision variable. The decision variable to be interchanged is the variable that is not common in the set of decision variables for the partition to be modified. The state variable to be interchanged is the variable that is not common in the set from the desired partition. These two variables are interchanged to produce the desired partition.

Partition #2 was developed by interchanging d and r_{max}. The set of decision variables in partition #1 is d L, and the set in partition #2 is r_{max} L. L is common to both of these sets and r_{max} and d appear in just one of the sets. It follows that r_{max} and d are the variables that should be interchanged. Since r_{max} is not a decision variable in partition #1, it must be a state variable. Thus, the equation for r_{max} in terms of d should be solved for d and used to eliminate d from the other equations to produce the desired partition. The variables to be interchanged to develop a partition are shown in the INTERCHANGE column of Table 4. along with the partition that should be modified.

Partition #6 can be developed from any of the partitions in level one. However, this is not always the case for partitions on level two or higher. All partitions on the next lower level may not be suitable for developing a desired partition since their sets of decision variables may have more then one variable that are not common. Thus, when developing a partition on level two or above, all partitions on the next lower level should be checked until a suitable partition is found.

The development of a new partition consists of the following two steps:

1) Search all partitions on the previous level to find a partition that is suitable for development of the desired partition. A suitable partition is one that exists and whose set of decision variables has only one variable that is not common with the set of decision variables in the desired set.

2) The variables are interchanged in the partition identified in step 1. This is done by solving the state equation for the decision variable identified in step 1. This equation is then used to eliminate the decision variable from the other equations in the partition, producing the new partition.

The automatic development of all of the partitions in Example 2. are shown in Figure 5. The development of these partitions took approximately 60 seconds on an 6 MHz IBM PC/AT.

It is readily apparent that the generation of all of the forms of the problem necessitates that the equations can be solved for all of their variables. For example,

$$Y = X + X \, SIN(X) \qquad (37)$$

cannot be solved for X. Equations such as this are often smooth in the region of interest and can be replaced with approximating functions (R. Johnson, 1980). The approximating function should be of a form that can be easily solved for all of its variables.

In many problems, all of the equations may not be in terms of all of the decision variables. When this occurs some of the equations cannot be solved for all of the decision variables in the problem. Take for example Example 2. If the equation for θ was only in terms of d, then the partition with θ and d as decision variables would not exist since there could not be an equation for τ or L in terms of θ and d. Thus there are five partitions and five candidate solution points since the limits on θ and d can never intersect.

In some cases, global knowledge about the problem can be obtained from the partitions as they are developed. Redundant state equations may not be apparent in some formulations of a problem but appear in others. When a redundancy is found in a new partition, the redundant state variable is removed from the problem, reducing the size of the problem, and any previously developed partitions that have the redundant variable as a state variable are re-used.

Trend or monotonicity analysis of the design equations

The set of candidate solution points is determined from an analysis of the design equations developed in the previous step. Trend or monotonicity analysis

(Wilde, 1975) applied to a design equation will determine the candidate solution point that corresponds to the trivial constraint set on that design equation.

The trend direction of the variables is determined by a first derivative analysis. The first derivative of the design equation is taken for each variable in turn. If the derivative with respect to a given variable is always positive, then the function will always increase if that variable is increased, likewise, the function value will decrease if that variable is decreased. The derivative comparisons are:

|derivative|= derivative -> Always increasing

|derivative|= -derivative -> Always decreasing

derivative = 0 -> Independent

else -> Not monotonic

For example:

$$M_t = \left[\frac{d^4\ PE_s\ \pi\ G}{16\ L}\right]^{1/2} \tag{38}$$

$$\frac{\partial M_t}{\partial d} = \left[\frac{d^2\ PE_s\ \pi\ G}{4\ L}\right]^{1/2} \tag{39}$$

$$\left|\frac{\partial M_t}{\partial d}\right| = \left[\frac{d^2\ PE_s\ \pi\ G}{4\ L}\right]^{1/2} \tag{40}$$

The absolute value of the derivative with respect to d equals the derivative. Thus the function is always increasing with increasing d, all else held constant. Since the minimization of M_t is desired, d should be minimized.

$$M_t = \left[\frac{d^4\ PE_s\ \pi\ G}{16\ L}\right]^{1/2} \tag{41}$$

$$\frac{\partial M_t}{\partial L} = -\left[\frac{d^4\ PE_s\ \pi\ G}{64\ L^3}\right]^{1/2} \tag{42}$$

$$\left|\frac{\partial M_t}{\partial L}\right| = \left[\frac{d^4\ PE_s\ \pi\ G}{64\ L^3}\right]^{1/2} \qquad (43)$$

With respect to L, the function value will decrease as L is increased, all else held constant, since the negative of the derivative equals the derivative. Thus, the candidate solution point corresponding to the trivial constraint set on Eq. (41) is (d_{MIN}, L_{MAX}).

As a second example:

$$M_t = \frac{2\ PE_s}{\theta} \qquad (44)$$

$$\frac{\partial M_t}{\partial \theta} = \frac{-2\ PE_s}{\theta^2} \qquad (45)$$

$$\left|\frac{\partial M_t}{\partial \theta}\right| = \frac{2\ PE_s}{\theta^2} \qquad (46)$$

This shows that the first derivative with respect to θ is always negative. Thus, to minimize M_t, θ should be maximized. Likewise for d:

$$\frac{\partial M_t}{\partial d} = 0 \qquad (47)$$

This candidate solution is independent of d. Any value of d that satisfies all of the limits with $\theta=\theta_{MAX}$ is a candidate optimal point. This candidate solution point turns out actually to be an infinite set of candidate solution points, all with the same value of M_t.

The program automatically makes the first derivative analysis, as shown in Figure 6., for each of the design equations to determine the set of candidate solution points. These points were shown in Table 3. The first derivative analysis took approximately 34 seconds on a 6 MHz IBM PC/AT.

One of the requirements of this new method of optimal design is that all variables must be positive. This requirement was made to facilitate the determination of the absolute values of the derivatives. Some terms such as

$$\left[1 - \frac{0.01\ n}{n_{MAX}} \right] \qquad (48)$$

from the pellet production problem are not clearly positive or negative. However, knowledge about the numerical values can be used to determine the sign of a term. In this case, n cannot be greater then n_{MAX}, so the term is always positive. The program deals with terms such as this by asking the user for more information.

Equations that are not monotonic are handled in the same way as term that are not clearly positive or negative - the program stops and asks for help. The equation may always be monotonic in the area of interest and non-monotonic elsewhere - regionally monotonic (Papalambros and Wilde, 1980). When the equation is monotonic in the area of interest, then this monotonicity can be identified by the designer and the analysis continued.

A more complicated case occurs when the equation is not monotonic in the area of interest. This new method was developed specifically for monotonic problems, however, experience has shown that it can be extended to work with some non-monotonic problems.

CONCLUSIONS AND FUTURE WORK

The new method of optimal design presented in this paper is a powerful method of solution extraction for some classes of constrained monotonic problems. The result of an analysis by this new method is the set of candidate solution points along with the state equations needed to test the feasibility of these points. The optimal solution for a specific numerical case is the best of the feasible candidate solution points. While this method cannot solve all problems, it can often be used to reduce problem size and to reformulate a problem into a more desirable mathematical form for analysis by conventional numerical methods. The method has been successfully used to correctly solve approximately twenty different problems from the literature, including some whose original solution in the literature is incorrect.

The candidate solution points found by this method must be numerically checked to determine which is the optimal solution. The next step in the automation of this method will be to develop an automated method of numerically checking the candidate solution points.

Large problems with many state and decision variables can be difficult and time consuming to solve using this method. For example, if there are eight state and eight decision variables, then there are $16!/(8!\ 8!) = 12870$ different partitions of state and decision variables. Knowledge about the numerical values of the upper and lower limits could be used to eliminate state equations that can never be violated in specific numerical cases (Papalambros and Li, 1985). If four state variables could be eliminated then there would be only $12!/(4!\ 8!) = 495$ different partitions. This reduction of problem size can help make the analysis of large problems feasible, although the solution would be correct only for a subset of the complete set of potential numerical cases.

A previously mentioned area for future work is with non-monotonic problems. Extrema of functions can be determined analytically to determine interior candidate solution points. If the problem is non-monotonic in more then one variable then simultaneous non-linear equations may need to be solved to determine the interior point or points. This has potential for determining the global solution when there are multiple local extrema.

ACKNOWLEDGEMENTS

Research sponsored by the Air Force Office of Scientific Research/AFSC, United States Air Force, under Contract F49620-85-C-0013. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The authors would like to express their thanks to Rodney C. Darrah and Sue Espy of Universal Energy Systems for their support, and to Marshall K. Kingery of Arnold Air Force Station for his support and assistance during this work. We would also like to thank Albert Rich of the Soft Warehouse for his helpful discussions about muMATH.

REFERENCES

Azarm, S., and Papalambros, P., "A Case for a Knowledge-Based Active Set Strategy," Journal of Mechanisms, Transmissions, and Automation in Design, Vol. 106, Mar 1984a, pp. 77-81.

Azarm, S., and Papalambros, P., "An Automated Procedure for Local Monotonicity Analysis," Journal of Mechanisms, Transmissions, and Automation in Design, Vol. 106, Mar 1984b, pp. 82-89.

Johnson, G. E., Strategies for Optimal Design, Doctoral Dissertation, Vanderbilt University, Nashville TN, 1978.

Johnson, G. E., and Townsend, M. A., "Selection and Sensitivity of Pellet Dimensions and Operating Parameters to Maximize Production Rate," Polymer Engineering and Science, Vol. 19, No. 7, May 1979, pp. 500-505.

Johnson, R. C., "Three Dimensional Variation Diagrams for Control of Calculations in Optimum Design," ASME, Journal of Engineering for Industry, Aug 1967, pp. 391-398.

Johnson, R. C., Mechanical Design Synthesis - Creative Design and Optimization, Robert E. Krieger Publishing Company, Malabar, Florida, 1978.

Johnson, R. C., "A Method of Optimum Design," Journal of Mechanical Design, Vol. 101, Oct 1979, pp. 667-673.

Johnson, R. C., Optimum Design of Mechanical Elements, John Wiley and Sons, Inc., NY, NY, 1980.

Papalambros, P. and Li, H. L., "Notes on the Operational Utility of Monotonicity in Optimization," Journal of Mechanisms, Transmissions, and Automation in Design, Vol. 105, June 1983, pp. 174-180.

Papalambros, P. and Li, H. L., "A Production System for Use of Global Optimization Knowledge," Journal of Mechanisms, Transmissions, and Automation in Design, Vol. 107, June 1985, pp. 277-284.

Papalambros, P., and Wilde, D. J., "Global Non-Iterative Design Optimization Using Monotonicity Analysis," Journal of Mechanical Design, Vol. 101, Oct 1979, pp. 645-649.

Papalambros, P. and Wilde, D. J., "Regional Monotonicity in Optimum Design," Journal of Mechanical Design, Vol 102, July 1980, pp. 497-500.

Wilde, D. J., "Monotonicity and Dominance in Optimal Hydraulic Cylinder Design," ASME, Journal Engineering Industry, Vol 97, No. 4, Nov 1975, pp 1390-1394.

Wilde, D. J., " The Monotonicity Table in Optimal Engineering Design," Engineering Optimization, Vol. 2, No. 1, 1976, pp. 29-34.

Zhou, J. and Mayne, R. W., "Interactive Computing in the Application of Monotonicity Analysis," Journal of Mechanisms, Transmissions, and Automation in Design, Vol. 105, June 1983, pp. 181-186.

NOMENCLATURE

MOD - New Method of optimal design:

- A_t = Number of partitions of state and decision variables.
- N_{DOF} = Number of degrees of freedom.
- N_D = Number of decision variables.
- N_S = Number of state variables.
- N_V = Number of variables = $N_D + N_S$.

Example 1. Pelletizer Problem:

- A_1 = Constant defined in Eq. (5).
- A_2 = Constant defined in Eq. (6).
- A_3 = Constant defined in Eq. (7).
- B = Number of blades on the cutting rotor.
- D = Diameter of the strands to be cut into pellets.
- l = Length of an individual pellet.
- n = Number of strands to be cut into pellets.
- n_{MAX} = Maximum number of strands that can be cut.
- P = Power required to cut the pellets.
- V = Volume of pellet production: mass per time.
- Ω = Rotational speed of the cutting rotor.
- ρ = Density of the pellets.
- τ = Shear strength of the plastic strands.

Example 2. Torsion Bar Problem:

- d = Diameter of the torsion spring.
- G = Modulus of rigidity.
- K = Stress concentration factor = 1.6
- L = Length of the torsion spring.
- M_t = Twisting moment.
- N = Factor of safety.
- p = S_e/S_{ty}
- PE_S = Energy to be absorbed by the torsion spring.
- S_e = Fatigue from reversed bending test.
- S_{ty} = Yield strength from a simple tensile test.
- τ_{max} = Maximum shear stress.
- θ = Angle of twist.

Subscripts:

- MAX = Upper bound on a variable.
- MIN = Lower bound on a variable.

Superscripts:

- * = Optimum value.

```
? ENTERPROBLEM();

What is the equation to be optimized ?
MT == 2 PEs/THETA;

What are the other equations ?
Enter 0==0 when done.
# 1    THETA == 32 L MT/(G d^4 PI);
# 2    TAUmax == 16 K MT/(d^3 PI);
# 3    0 == 0;

[PEs, K, TAUmax, MT, PI, d, L, G, THETA] are the
variables in the equations.

What is the variable to be optimized ?  MT;
MINimize or MAXimize ?  MIN;

What are the constrained variables?
Enter DONE when finished.
Constrained variable # 1 ? THETA;
Constrained variable # 2 ? TAUmax;
Constrained variable # 3 ? L;
Constrained variable # 4 ? d;
Constrained variable # 5 ? DONE;

What are the free variables?
Enter DONE when finished.
Free variable # 1 ? DONE;
```

Figure 1. Entering the Torsion Bar Problem

```
0       MT == 2 PEs/THETA

1       THETA == 32 L MT/(G d^4 PI)
2       TAUmax == 16 K MT/(d^3 PI)

In the FF there will be: 2 state equations and 2
decision variables.
Eqn #0  MT          in terms of [THETA]
Eqn #1  THETA       in terms of [MT, L, d]
Eqn #2  TAUmax      in terms of [MT, d]
Limited variables are: [THETA, TAUmax, L, d]
Free variables are [] and should be eliminated.
[TAUmax] are state only.
[L, d] are decision only.
[THETA] are both state and decision.
SDE'S have design vary - Try AUTO();
  or RECOMMEND(); for recommendations.
Problem is not in final formulation form.
@: TRUE
```

Figure 2. Problem as Entered into the Computer

```
? AUTO();

0      MT == 2 PEs/THETA

1      THETA == 64 L PEs/(G THETA d^4 PI)
2      TAUmax == 32 K PEs/(THETA d^3 PI)

In the FF there will be: 2 state equations and 2
decision variables.
Eqn #0  MT          in terms of [THETA]
Eqn #1              in terms of [THETA, L, d]
Eqn #2  TAUmax      in terms of [THETA, d]
Limited variables are: [THETA, TAUmax, L, d]
Free variables are [] and should be eliminated.
[TAUmax] are state only.
[L, d] are decision only.
[THETA] are both state and decision.
SDE's [1] are not solved for problem variables - Try
  SOLVEFOR(eqn number , vary name);
  or MAKESTATE(decision vary name in eqn);
Problem is not in final formulation form.
@: TRUE
```

Figure 3. Elimination of the Design Variable

```
? MAKESTATE(THETA);
#1  THETA == 64 L PEs/(G THETA d^4 PI)
#2  TAUmax == 32 K PEs/(THETA d^3 PI)
Select the equation to become the state equation. 1;

 Solving THETA == 64 L PEs/(G THETA d^4 PI) for THETA
 Root # 1 : THETA == 8 L^(1/2) PEs^(1/2)/(G^(1/2) d^2
PI^(1/2))
 Root # 2 : THETA == -8 L^(1/2) PEs^(1/2)/(G^(1/2) d^2
PI^(1/2))
 Select root # ? 1;

0     MT == G^(1/2) d^2 PI^(1/2) PEs^(1/2)/(4 L^(1/2))

1     THETA == 8 L^(1/2) PEs^(1/2)/(G^(1/2) d^2
PI^(1/2))
2     TAUmax == 4 G^(1/2) K PEs^(1/2)/(L^(1/2) d
PI^(1/2))

In the FF there will be: 2 state equations and 2
decision variables.
Eqn #0  MT          in terms of [L, d]
Eqn #1  THETA       in terms of [L, d]
Eqn #2  TAUmax      in terms of [L, d]
Limited variables are: [THETA, TAUmax, L, d]
Free variables are [] and should be eliminated.
[THETA, TAUmax] are state only.
[L, d] are decision only.
[] are both state and decision.
Problem is in final formulation form.
Use DOIT(); to extract the solutions.
@: TRUE
```

Figure 4. Making θ a State Variable

```
? DOIT();
----------------------------------------

Transforming conversion results into a partition.

----------------------------------------
Partition #1 in terms of: L  d


MT == G^(1/2) d^2 PI^(1/2) PEs^(1/2)/(4 L^(1/2))

TAUmax == 4 G^(1/2) K PEs^(1/2)/(L^(1/2) d PI^(1/2))
THETA == 8 L^(1/2) PEs^(1/2)/(G^(1/2) d^2 PI^(1/2))

MT        is in terms of [L, d]
TAUmax    is in terms of [L, d]
THETA     is in terms of [L, d]
----------------------------------------
Partition #2 in terms of: THETA  d


MT == 2 PEs/THETA

TAUmax == 32 K PEs/(THETA d^3 PI)
L == G THETA^2 d^4 PI/(64 PEs)

MT        is in terms of [THETA]
TAUmax    is in terms of [THETA, d]
L         is in terms of [THETA, d]
--------- ------------------------------
Partition #3 in terms of: TAUmax  d


MT == d^3 PI TAUmax/(16 K)

THETA == 32 K PEs/(d^3 PI TAUmax)
L == 16 G K^2 PEs/(d^2 PI TAUmax^2)

MT        is in terms of [TAUmax, d]
THETA     is in terms of [TAUmax, d]
L         is in terms of [TAUmax, d]
----------------------------------- --------
```

Figure 5. Torsion Bar Partition Generation

```
Partition #4 in terms of: THETA  L


 Solving THETA == 8 L^(1/2) PEs^(1/2)/(G^(1/2) d^2
PI^(1/2)) for d
 Root # 1 : d == 2^(3/2) L^(1/4) PEs^(1/4)/(G^(1/4)
THETA^(1/2) PI^(1/4))
 Root # 2 : d == -2^(3/2) L^(1/4) PEs^(1/4)/(G^(1/4)
THETA^(1/2) PI^(1/4))
 Select root # ? 1;

MT == 2 PEs/THETA

TAUmax == 4 G^(3/4) THETA^(1/2) K PEs^(1/4)/(2^(3/2)
L^(3/4) PI^(1/4))
d == 2^(3/2) L^(1/4) PEs^(1/4)/(G^(1/4) THETA^(1/2)
PI^(1/4))

MT        is in terms of [THETA]
TAUmax    is in terms of [THETA, L]
d         is in terms of [THETA, L]
-----------------------------------------
Partition #5 in terms of: TAUmax  L


MT == 4 G^(3/2) K^2 PEs^(3/2)/(L^(3/2) PI^(1/2)
TAUmax^2)

THETA == L^(3/2) PI^(1/2) TAUmax^2/(2 G^(3/2) K^2
PEs^(1/2))
d == 4 G^(1/2) K PEs^(1/2)/(L^(1/2) PI^(1/2) TAUmax)

MT        is in terms of [TAUmax, L]
THETA     is in terms of [TAUmax, L]
d         is in terms of [TAUmax, L]
-----------------------------------------
Partition #6 in terms of: TAUmax  THETA


 Solving TAUmax == 32 K PEs/(THETA d^3 PI) for d
 Root # 1 : d == 2 4^(1/3) K^(1/3)
PEs^(1/3)/(THETA^(1/3) PI^(1/3) TAUmax^(1/3))
 Root # 2 : d == 2 4^(1/3) #E^(2/3 #I #PI) K^(1/3)
PEs^(1/3)/(THETA^(1/3) PI^(1/3) TAUmax^(1/3))
 Root # 3 : d == 2 4^(1/3) #E^(4/3 #I #PI) K^(1/3)
PEs^(1/3)/(THETA^(1/3) PI^(1/3) TAUmax^(1/3))
 Select root # ? 1;

MT == 2 PEs/THETA
```

```
L == 4^(1/3) G THETA^(2/3) K^(4/3) PEs^(1/3)/(PI^(1/3)
TAUmax^(4/3))
d == 2 4^(1/3) K^(1/3) PEs^(1/3)/(THETA^(1/3) PI^(1/3)
TAUmax^(1/3))

MT          is in terms of [THETA]
L           is in terms of [TAUmax, THETA]
d           is in terms of [TAUmax, THETA]
-----------------------------------------
```

Figure 5. Torsion Bar Partition Generation Continued (to be merged into one figure)

SOLUTION OF

Mt == d^2 (G PI PEs)^(1/2)/(4 L^(1/2))

L max d min

SOLUTION OF

Mt == 2 PEs/THETA

THETA max

SOLUTION OF

Mt == PI d^3 TAUmax/(16 K)

TAUmax min d min

SOLUTION OF

Mt == 2 PEs/THETA

THETA max

SOLUTION OF

Mt == 4 K^2 (G PEs)^(3/2)/(L^(3/2) PI^(1/2) TAUmax^2)

TAUmax max L max

SOLUTION OF

Mt == 2 PEs/THETA

THETA max

Figure 6. Candidate Solution Extraction

FINAL REPORT NUMBER 39
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Yong Kim
760-6MG-004

FINAL REPORT

"The Synthesis of Some New Energetic Materials"

by

Joel R. Klink

University of Wisconsin-Eau Claire

December 31, 1987

in accordance with

Contract No. F49620-85-C-0013/SB5851-0360

Sponsored by: Air Force Office of Scientific Services

Administered by: Universal Energy Services
4401 Dayton-Xenia Road
Dayton, OH 45432

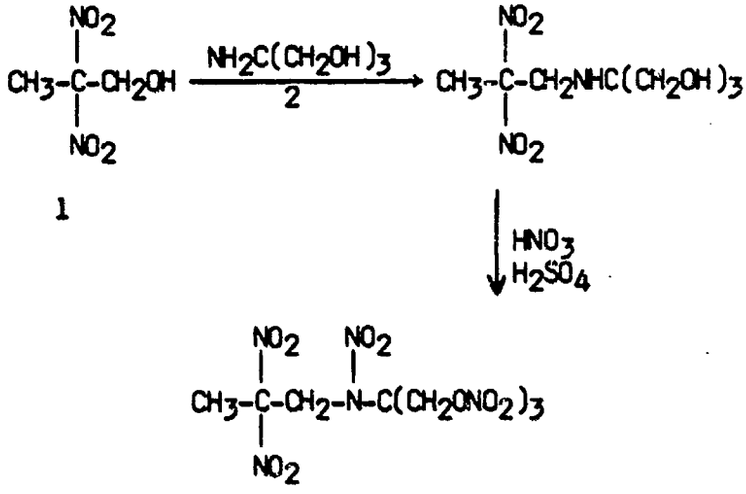
ABSTRACT

The modified Mannich condensation of 2,2-dinitropropanol with tert-alkylamines, $R'R''R'''CNH_2$, where the R groups were varied form three CH_3 groups to three $HOCH_2$ groups was investigated. The condensation is not seriously sterically hindered nor does intromolecular bonding in the hydroxyalkylamines prevent reaction, contrary to an earlier report. Nitration of the Mannich adducts to yield the nitraminoaklyl nitrates was studied. The reaction of 2,2-dinitro-1,3-propanediol with a variety of amines under a range of reaction conditions failed to yield the bis-condensation.

I. Introduction

The investigation undertaken was to attempt to condense 2,2-dinitro-1-propanol, 1, with tris(hydroxymethyl)methylamine, 2, and to fully nitrate the adduct as shown in Scheme 1. It was anticipated that the final pro-

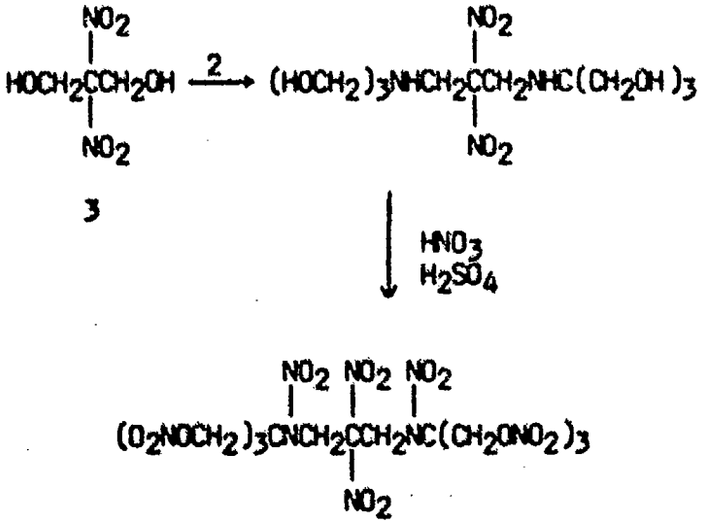
Scheme 1



duct, 2,2-bis(hydroxymethyl)-3,5,5-trinitro-3-aza-1-hexanol trinitrate, would have desirable physical and chemical properties to be utilized as a plasticizer in certain propellant formulations.

A second study sought to find the conditions necessary to accomplish the bis Mannich condensation of 2,2-dinitro-1,3-propanediol, 3, with amines in general and 2 specifically and then to nitrate the adduct as shown in Scheme 2.

Scheme 2



It was expected that the final product of this synthesis would be a potentially useful high melting oxidizer.

II. Results and Discussion

The attractive procedure[1] for the preparation of 1 and 3 in our hands usually afforded little or none of the desired compounds. An unpublished modification of the procedure, which does yield consistent results, was obtained[2] and is detailed in the experimental section.

A. Condensations of 2,2-Dinitropropanol and Nitration of the Adducts.

The modified Mannich condensation of 2,2-dinitroalcohols occurs readily with amines in a variety of solvents.[3] In view of the structural diversity of the amines that have been shown to react and based on the fact that we[4] had shown that 2,2-dinitro-2-fluoroethanol could be condensed with 2 despite a published report to the contrary,[5] we initially attempted to condense 1 with 2 under aqueous conditions. Continuous ether extraction of the reaction mixture failed to yield the desired adduct. Variation of the reaction time, pH, and temperature did not measurably affect the results. Therefore, it was decided to study the reaction of 1 with a series of amines, varying systematically in structure from tert-butylamine to 2, to ascertain if some factor such as a steric effect or intramolecular hydrogen bonding might be hindering the reaction.

As shown in Table 1, 2,2-dinitropropanol reacts readily and in acceptable yield with the hindered amines studied. Furthermore, there appears to be no inhibition of the reaction due to the expected intramolecular hydrogen bonding in the hydroxyalkylamines. We concluded, therefore, that our failure to observe a reaction between 1 and 2 was the result of a poorly designed reaction work-up procedure.

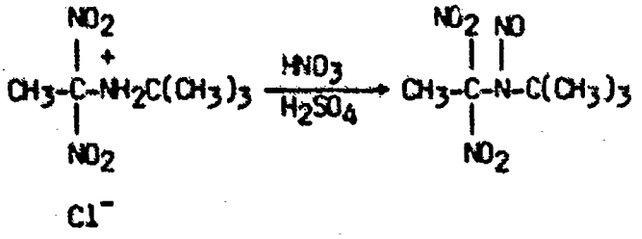
Table I. Mannich Products of 2,2-Dinitropropanol and Amines

Amine, $RR'R''CNH_2$					
R	R'	R''	Solvent	Yield, %[a]	m.p. °C
CH_3	CH_3	CH_3	H_2O	73	144 dec[c]
CH_3	CH_3	$HOCH_2$	CH_2Cl_2	83	133 dec[c]
CH_3	$HOCH_2$	$HOCH_2$	CH_3OH	42[b]	94-95
$HOCH_2$	$HOCH_2$	$HOCH_2$	H_2O	97	128-131

[a]Based on crude product. [b]Recrystallized product. [c]Hydrochloride of adduct.

As detailed in the experimental section, the desired condensation of 1 with 2 has been achieved, based on elemental analysis, but purification of the adduct remains to be accomplished. The initial attempts to recrystallize the crude adduct from acetone/ether yielded unreacted 2, a large amount of an uncharacterized yellow oil and a small amount of relatively pure adduct. We are currently attempting to improve the purification procedure.

Nitration of the Mannich adducts of 2,2-dinitropropanol was carried out in methylene chloride using a mixture of colorless fuming (90%) nitric and fuming (130%) sulfuric acids. This procedure appears, however, to be unsatisfactory. Nitration of the hydrochloride of 2,2-dimethyl-5,5-dinitro-3azahexane, equation 1, yielded the N-nitroso



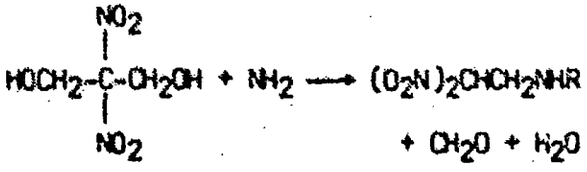
(1)

rather than the N-nitro (nitramine) product as indicated by elemental analysis. With other adducts this nitration procedure has yielded products that also appear to be the N-nitroso derivatives although this remains to be confirmed. It is now obvious that we need to explore other nitration methods to obtain the nitramino nitrate esters.

Since we have been unable so far to obtain the desired adduct of 1 with 2 in pure form, it was decided to nitrate the crude material from the Mannich condensation. Work up of this reaction has yielded a pale yellow oil that is currently being characterized.

B. Condensation of 2,2-Dinitro-1,3-propanediol with Amines

The Mannich condensation of 2,2-dinitro-1,3-propanediol, 3, with amines usually results in condensation with one mole of amine, mono condensation, accompanied by the loss of one mole of formaldehyde[6] (demethylolation), equation 2. Two reports have claimed condensation



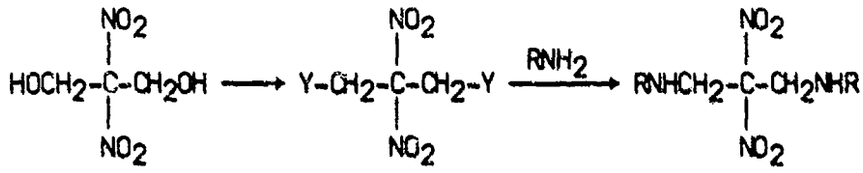
(2)

with two moles of amine, a bis condensation, but no experimental details of Hamel's work[3d] have appeared. The major goal in this portion of the project was to find conditions under which the bis condensation could be achieved. The addition of formaldehyde, which might suppress demethylolation, unfortunately leads to the production of cyclic adducts.[6c-d,7] Thus, this investigation centered on variables such as stiochiometry, solvent type, and amine basicity.

Previous studies on Mannich condensations of 3 were done in aqueous solution. Based on the supposition that demethylolation might be favored by a good ionizing solvent, we attempted to run the reaction in a poor ionizing solvent, a mixture of methylene chloride and acetone (10:1, v:v). In this solvent even with a two-fold excess of isopropyl or isobutylamine, only the mono product could be isolated. The much less basic amine, aniline, also yielded only the mono adduct. The use of methanol as solvent, conducting the reaction at low pH under aqueous conditions and increasing both reaction temperature and time failed to yield the bis adduct.

Based on these results we decided to explore another approach to the synthesis of the bis adduct, namely via a nucleophilic substitution process as shown in Scheme 3.

Scheme 3



Y = Cl, Br, OTos or OSO_2CF_3

Reaction of 3 with thionyl bromide or phosphorus tribromide in pyridine failed to yield the dihalide. The ditriflate was prepared by reaction of 3 with trifluorosulfonic anhydride in pyridine. Since this is a difficult and expensive compound to prepare, it was decided to conduct some preliminary studies on the triflate of 2,2-dinitropropanol as a model system. The requisite triflate has been prepared but reaction of the triflate with amines remains to be investigated.

III. Experimental Section

General Data

Melting points were measured on a Thomas-Hoover capillary tube apparatus and are uncorrected. IR spectra were taken on a Perkin-Elmer 1427 ratio recording spectrometer. ^{1}H and ^{13}C NMR were taken on an IBM NR-80 spectrometer. Chemical shifts are reported in δ values relative to tetramethylsilane as internal standard. Elemental analyses were done at Micro-Tech Laboratories, Inc., Skokie, IL.

2,2-Dinitropropanol, 1

To a continuously stirred mixture of 95.4 g (0.90 mol) of sodium carbonate and sodium nitrite (103.4 g, 1.31 mol) in 450 mL of water at 15°C was added 71.4 g (0.30 mol) of sodium persulfate followed by 19.7 g (0.060 mol) of potassium ferricyanide and, after a few minutes, 22.4 g (0.30 mol) of nitroethane. An exotherm occurred in about 10 minutes, but the reaction temperature was kept below 30°C. After a total of 45 minutes, formalin (37%, 24 g, 0.30 mol) was added followed immediately by sufficient cold 40% phosphoric acid to neutralize the mixture to pH 6-7. The mixture was extracted with ether (3 x 150 mL). The combined ether layers were washed with brine, dried over $MgSO_4$ and the ether roto-evaporated. The residue was vacuum distilled (Kugelrohr) to yield 16.6 g (37%) of translucent solid, bp 110°C, 0.15 mm (lit[1] bp 100-110°C, 0.1 mm).

2,2-Dinitro-1,3-propanediol, 3

To a continuously stirred solution of 95.4 g (0.90 mol) of sodium carbonate in 450 mL of water at 15°C was added successively 103.5 g (1.5 mol)

of sodium nitrite, 71.4 g (0.30 mol) of sodium persulfate, 19.7 g (0.060 mol) of potassium ferricyanide and 18.3 g (0.30 mol) of nitromethane. The ice-bath was removed and stirring continued for 30 minutes before formalin (37%, 72 g, 0.90 mol) was added followed by immediate neutralization to pH 6-7 with cold 40% phosphoric acid. The mixture was extracted with ether (3 x 300 mL) and the combined extracts washed with brine (2 x 200 mL). After drying over $MgSO_4$, evaporation of the solvent left an oil that crystallized upon refrigeration. Recrystallization of the crude material from ethylene dichloride gave 12.9 g (26%) of white needles, mp 137-141°C (lit[1] mp 140-142°C).

General Procedure for the Condensation of 2,2-Dinitropropanol, 1, with Amines

To a magnetically stirred solution of 1 in the appropriate solvent was added an equimolar amount of amine in the same solvent over 5 minutes. The mixture was stirred 1-2 hours at room temperature and worked up by roto-evaporating the solvent to leave a crude residue that was recryatallized or dissolved in anhydrous ethyl ether and converted to the hydrochloride with HCl(g).

2,2-Dimethyl-5,5-dinitro-3-azahexane, 4.

From 3.0 g (0.020 mol) of 1 in 50 mL of methylene chloride and 1.46 g (0.020 mol) of t-butylamine in 15 mL methylene chloride was isolated an orange oil from which 1.5 g (31%) of recrystallized (ether/methanol) hydrochloride, mp 144°C dec, was obtained; ^1H NMR (acetone-D_6) δ 1.40 (s, 9H), 2.45 (s, 3H), 4.18 (s, 2H), 9.5 (bs, 2H).

Anal. Calc. for $C_7H_{16}ClN_3O_4$: C, 34.79; H, 6.67; N, 17.39; Cl, 14.67. Found. C, 34.75, H, 6.60, N, 17.54, Cl, 14.34.

2,2-Dimethyl-5,5-dinitro-3-aza-1-hexanol, 5.

From 3.0 g (0.020 mol) of 1 in 50 mL methylene chloride and 1.78 g (0.020 mol) of 2-amino-2-methyl-1-propanol in 15 mL methylene chloride was obtained 4.3 g (83%) of crude hydrochloride. Recrystallization from methanol/ether gave white needles, mp 136°C dec.

Anal. Calcd for $C_7H_{16}ClN_3O_5$: C, 32.63; H, 6.26; N, 16.31; Cl, 13.76.

Found: C, 35.30; H, 5.86; N, 15.54; Cl, 12.81.

2-Hydroxymethyl-2-methyl-5,5-dinitro-3-aza-1-hexanol, 6.

From 7.5 g (0.050 mol) of 1 in 40 mL methanol and 5.25 g (0.050 mol) of 2-amino-2-methyl-1,3-propanediol in 40 mL of methanol after reluxing for 4 hours was obtained 5.75 g (42%) of recrystallized (ethyl ether) white plates, mp 91-92°C, ^{1}H NMR (acetone-D_6) δ 0.93 (s, 3H), 2.21 (s, 3H), 3.45 (s, 4H), 3.72 (s, 2H), 3.5 (bs, 3H).

Anal. Calcd for $C_7H_{15}ClN_3O_6$: C, 35.44; H, 6.37; N, 17.71.

Found: C, 35.47; H, 6.41; N, 17.70.

2,2-bis(Hydroxymethyl)-5,5-dinitro-3-aza-1-hexanol, 7.

To a stirred solution of 3.00 g (0.020 mol) of 1 in 30 mL water was added 2.42 g of tris(hydroxymethyl)methylamine all at once. The resulting solution was stirred 1.5 h and the water removed at 30°C (rotary evaporation). The yellow viscous oil that remained was refrigerated one day and upon scratching yielded 5.38 g (106%) of yellow solid, mp 56-61°C. Repeated attempts to purify portions of this material by recrystallization have yielded a small amount of pale yellow crystals, mp 129-131°C.

Anal. Calcd. for $C_7H_{15}N_3O_7$: C, 33.20; H, 5.97; N, 16.59.

Found: C, 33.94; H, 6.02; N, 16.32

General Procedure for Nitration of Mannich Adducts.

CAUTION: These nitrated materials are potentially explosive and must be handled with great care utilizing appropriate procedures. To a mixture of the Mannich adduct and methylene chloride at 0-5°C was added dropwise 4 mL of colorless 90% fuming nitric acid per gram of adduct. Next 4 mL of 30% fuming sulfuric acid per gram of adduct was added dropwise maintaining the reaction temperature at 0-5°C. After stirring an additional 0.5 h the ice bath was removed and stirring continued for 3 h. The mixture was poured with stirring into ice, the organic layer separated and the aqueous phase extracted with methylene chloride (3 x 50 mL). The combined extracts were washed twice with 10% aq. sodium bicarbonate, once with brine and dried over $MgSO_4$. The solvent was rotary evaporated to leave the crude nitrate.

Nitration of 4. From 1.29 g of the hydrochloride of 4 was obtained 0.16 g (13%) of recrystallized (acetone) product, mp 115°C dec. ^{1}H NMR, (acetone-D_6). δ 1.69 (s, 9H), 1.85 (s, 3H), 4.77 (s, 2H).

Anal. Calcd for $C_7H_{14}N_4O_5$ (the N-nitroso derivative): C, 35.90; H, 6.03; N, 23.92. Found: C, 35.95; H, 6.11; N, 23.88.

Nitration of 5. Nitration of 1.29 g of the hydrochloride of 5 yielded a pale yellow oil, (0.5 g, 32%) that was passed through a 13 cm silica gel column using methylene chloride as eluent. The oil isolated upon evaporation of the solvent was still impure by ^{1}H NMR studies. Further purification has not been attempted.

Nitration of 6. From 1.19 g of 6 was obtained 1.3 g (64%) of a yellow semisolid that was recrystallized from ether to yield white needles, mp 77-78°C. ^{1}H NMR, (acetone-D_6), δ 1.69 (s, 3H), 2.31 (s, 3H), 5.18 (AB system, J_{AB} 0.24 δ, 4H), 5.23 (s, 2H).

Anal. Calcd. for $C_7H_{12}O_{12}N_6$: C, 22.59; H, 3.25; N, 22.58.

Found: C, 2.62; H, 3.23; N, 21.84.

Nitration of 7. From 1.52 g (0.0060 mol) of crude 7 was obtained 1.86 g (72%) of a pale yellow liquid. Repeated attempts to crystallize this material have failed. Spectroscopic studies are yet to be done.

Condensation Reactions of 2,2-Dinitro-1,3-propanediol,3.

Repeated attempts to condense 3 with amines resulted in the formation of the demethylolated monoadduct. The reaction was carried out in water and in CH_2Cl_2/acetone. Isobutylamine, isopropylamine, and aniline were condensed. In all cases where a solid product was isolated, ^{1}H NMR indicated the presence of the monoadduct based on the ^{1}H NMR triplet at 7.1 ppm. No further purification or characterization of these products has been carried out to date.

2,2-Dinitro-1-propanol Triflate.

To a stirred solution of 1.68 g of trifluoromethanesulfonic anhydride in 50 mL of methylene chloride at 0-3°C was added dropwise under nitrogen a solution of 3.0 g (0.020 mol) of 1 and 1.50 mL of pyridine in 30 mL of methylene chloride. The mixture was stirred overnight after warming to room temperature. The mixture was filtered through silica gel and the solvent evaporated. Vacuum distillation yielded 4.2 g (74%) of a colorless oil. bp 56-65°C, 0.3-0.5 mm. ^{1}H NMR ($CDCl_3$) δ 2.38 (s, 3H), 5.21 (s, 2H).

2,2-Dinitro-1,3-Propanediol Ditriflate.

To a stirred mixture of 1.66 g (0.010 mol) of 3 and 2.3 mL of pyridine in 50 mL of dry methylene chloride at 15°C under nitrogen, was added dropwise a solution of (0.030 mol) of trifluoromethanesulfonic anhydride in 20 mL of methylene chloride keeping the temperature below 20°C. The mixture was stirred 4 h and then filtered through a silica gel column. Roto evaporation of the solvent left yellowish crystals from which 1.54 g (36%) of white needles, mp 52-54°C were obtained by recrystallization. ^{1}H NMR ($CDCl_3$) δ 5.52 (s).

Literature References

1. L. C. Garver, V. Grakauskas, and K. Baum, J. Org. Chem, 1985, 50, 1699.

2. We are indebted to Dr. Lee Garver, Fluorchem, Inc., Azusa, CA, for this procedure.

3. a. K. Baum and W. T. Maurice, J. Org. Chem., 1962, 27, 2231.
   b. M. B. Frankel and K. Klager, J. Am. Chem. Soc., 1957, 79, 2953.
   c. M. B. Frankel and K. Klager, J. Chem. Eng. Data, 1962, 7(3), 412.
   d. E. E. Hamel, Tetrahedron, 1963, 19 (Suppl. 1), 85.

4. Joel R. Klink, FINAL REPORT, 1986 USAF-UES SFRP, "The Synthesis of Fluorodinitroethylnitraminoalkyl Nitrates and Compatibility Studies of GAP-Nitrate and TAET."

5. D. A. Nesterenko, O. M. Savchenko, and L. T. Eremenko, Bull. Acad. Sci, USSR, Div. Chem. Sci., Engl. Trans., 1970, 1039.

6. a. A. H. Feuer, G. B. Bachman, and W. May, J. Am. Chem. Soc., 1954, 76, 5124.

b. D. A. Cichra and H. G. Adolph, J. Org. Chem., 1982, 47, 2474.

c. H. Piotrowska, T. Urbanski, and K. Wejroch-Matacz, Rocz. Chem., 1971 47(7/8), 1267.

d. H. Piotrowska, T. Urbanski, and K. Wejroch-Matacz, Rocz. Chem., 1971, 45(7/8), 2107.

7. D. A. Levins, C. D. Bedford, and S. J. Staats, Propellants, Explos., Pyrotech, 1983, 8(3), 74.

FINAL REPORT NUMBER 41
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Stephen Kolitz
760-6MG-094

MBE GROWN Al Cu ALLOY FILMS

by

Philipp Kornreich
Department of Electrical and Computer Engineering
Syracuse University
Syracuse, NY 13244-1240
(315) 423-4447

FINAL REPORT

Sponsor: AFOSR - Universal Energy Systems

Purchase Order No.: S-760-6MG-090

MBE GROWN Al-Cu FILM

P. Kornreich

ABSTRACT

An MBE machine at RADC-Griffiss A.F.B., Rome, NY was to be used in the study of AlCu alloy films for application as leads in VHSIC's. However, as of the date of this final report, the MBE machine is still not complete. The substrate holder, which was to perform many functions in this machine is awaiting construction. The fact that the MBE machine is actually a modified Auger apparatus requires a complicated substrate holding mechanism. However, we did fabricate AlCu alloy films on Si substrates in the Syracuse University Micro Electronics Laboratory. These films were fabricated in a vacuum system with a residual atmosphere of 20% H_2 and 80% N_2. This greatly reduces oxygen contamination in the film. Oxygen is the major offending impurity in these films. The films are currently waiting for analysis at RADC.

The purpose of this task was to complete the construction of the MBE machine at RADC (RBRE) and to use it to grow ultra pure AlCu alloy films for study as leads in VHSIC's. It should be possible to grow very pure AlCu alloy films in the ultra high vacuum of about 10^{-9} to 10^{-10} torr that exists in the MBE machine.

To this end we first attempted to complete the construction of the MBE machine at RADC. The fabrication of the source holder of the MBE machine was finally completed in the RADC machine shop. We installed the source holder and found that it leaked. We sprayed He gas with a small diameter hose at various locations of the source holder. We used the quadrupole mass spectrometer, in the machine, to detect the He entering through leaks. This allowed us to locate the positions of the leaks. The source holder was disassembled and rewelded.

After the source holder was again installed in the MBE machine a single very small leak was detected. This leak could only be detected when the system was pumped out to a vacuum of about 10^{-8} Torr. The source holder was again disassembled, rewelded, installed in the MBE machine and vacuum tested. This time it worked correctly. The liquid nitrogen deposition shield was installed in the MBE machine. It was leak tested with both liquid nitrogen in its tank and with an empty tank and no leaks were detected.

Fused Quartz Source material containers, that fit in to the source holder, were fabricated by the glass blower in the College of Environmental Science and Forestry at Syracuse University.

A pneumatic source holder shutter actuating mechanism was designed, and fabricated in the RADC machine shop. It was tested and found to function correctly. The source holder with its

pneumatic shutter actuating mechanism and the evaporazation shield are currently installed in the MBE machine.

During the time when the above components were being fabricated we designed a substrate holding mechanism. The substrate holding mechanism has to perform several tasks. It has to function within the constrains of the present system. The mechanism has to hold six substrates. One has to be able to heat each substrate individually to about 600 C. The substrates are loaded unto the substrate holder from an existing load lock. The load lock is located at the opposite side from the deposition source holder in the main vacuum chamber. Thus, one has to be able to rotate the source holder 180. The substrates are loaded unto the source holder through the load lock. The substrate holder is then rotated 180 in order for the substrates to face the deposition sources. However, this is not the only task that the cource holder mechanism has to perform.

An Auger analysis apparatus and sputtering gun are located at 90 with respect to both the sources and the load lock in the main vacuum chamber. The substrates have to be tilted about 30 degrees with respect to the vertical axis of the vacuum system when they face the auger apparatus. Thus, the substrate holder mechanism has to be able to rotate the substrates 90, 180 and tilt the substrates at an angle of 30 about a horizontal axis in the 90 degrees position only. An elliptical grooved cam is used for providing the correct allignment of the substrates in each position. Material for this substrate holder has been on order for some time. It is expected that the material for the substrate holder will arrive shortly.

While the exceeding long process of fabricating of the MBE machine is in progress we deposited AlCu alloy films in the Syracuse Microelectronics Laboratory. Here we have to use a vacuum system

that can only be pumped out to about 10^{-6} Torr. However, we flush the vacuum system three times with a combination of 20% H_2 and 80% N_2 gas. Thus, the residual gas in the system contains 20% H_2 and 80% N_2. The hydrogen tends to reduce the oxygen in the system. This method tends to greatly reduce the oxygen content in the AlCu films. Oxygen is the main offending contaminant in the films. We have used this process in the past in AlCu alloy films. Auger analysis has shown, at that time, that AlCu films prepared by this process are virtually oxygen free.

We also used this method recently to fabricate high purety Al and Cu films. We are currently using these films to calibrate the Auger micro probe at RADC. There are two parts to this calibration. We use the Al and Cu films to calibrate the Auger peak heights for quantative analysis of the CuAl alloy films. We also calibrate sputtering rate of the sputtering gun of Auger micro probe with these films. However, for the calibration of the sputtering gun the film thickness must be known. We measured the film thickness with an interferometer at RADC. We also measured the film thickness with a quartz crystal deposition monitor. Indeed, we had to claibrate the deposition monitor for Cu, Al, and AlCu alloy films by fabricating films and measuring the film thickness with the interferometer. We also have a "formula" which allows us to predict the film thickness from the amount of material that is loaded on the thermal evaporation filament. Of course, all the material has to be evaporated in order for the "formula' to hold. In our metal depositions, indeed, most of the material is evaporated. We found that the film thickness predicted by our "formula" and the ones obtained from interferometer measurements are in good agreement.

We have recently fabricated AlCu alloy films with 4% Cu by volume on Si substrates. We have patterned some of these films into

10mm wide lines. The films are currently awaiting Auger compositional analysis, cross sectional SEM analysis, and electromigration tests at RADC.

Simulation for Priority Handling Algorithms

Final Report

Sponsored by the

Air Force Office of Scientic Services

Bolling AFB, DC

Conducted by the

Universal Energy System, Inc.

Prepared by:	Mou-Liang Kung
Academic Rank:	Associate Professor
Department:	Department of Mathematics and Computer Science
University:	Norfolk State University Norfolk, Virginia 23504
USAF Focus Point:	John Salerno, RADC, Griffiss AFB
Date:	December 15, 1987
Contract No.:	F49620-85-C-0013/SB5851-030
Purchase Order:	No. S-760-6MG-011

Section 1 Activities

According to the approved Project proposal, the following Time-Table was used to implement the activities of the Project on a timely manner. The Time-Table is verified with progress reported in the Comment column.

Period Jan.-May 1987

Activities	Key Personnel	Comments
-Securing secretarial support	PI*	Done (Wanda Morris)
-Securing computer equipment computing time, and office supply	PI	Done (IBM AT system secured, supplies acquired)
-Securing Research Trainee #1	PI	Antonio Ransom was selected and trained
-Attending NTA Conference	PI&RT1*	Student paper presentation on the integrated networks
-Preparing system specification for Preliminary Design	PI	Done (Attached to this report)
-Preliminary Design Review and Visiting RADC/DCLD, Griffiss AFB, NY	PI	Scheduled with John Salerno at RADC on June 4, 1987
-Product Specification and Flow	PI	Included in the Preliminary Design Report
-preparing first period report	PI	Done

* PI = Principal Investigator
* RT1 = Research Trainee #1

Period June-August 1987

-Preliminary Design Review at RADC with Focus Point	PI	June 4 traffic record format obtained
-IN software design defined, Algorithms developed, coding started	PI	June 5
-Securing research trainee #2	PI	Yvette Cherry was selected, trained ,a C menu program written
-Attending Communications Network Management Workshop, The CASE Center, Syracuse, New York	PI	June 30 - July 2
-Presenting the Priority Handling Algorithms at the RADC's Voice/ Data Integrator Design Program Kickoff Meeting with CMC Electronics, Eatontown, N.J.	PI	August 19 - 20 Paper presented, Discussion on the VDI design issues participated
-Traffic Simulator Software received from RADC, software bugs found (It failed to run on PC.)	PI	Received, August 31 Debugging started
-Data Structure (record format) on the traffic files generated from Traffic Simulator changed	PI	August 31, this change resulted in the start of major revision of the integrated node simulation software

Period September - December, 1987

-Paper on Priority Handling Algorithms submitted to Focus Point at RADC	PI	September 14
-Debugging on Traffic Simulator is completed	PI	November 8
-Attending RADC's VDI program meeting at RADC, NY	PI	November 9-10
-Integrated Node simulation coding, debugging & testing	PI	Until Dec. 15, 1987
-Preparing Final Report	PI	December 15, 1987

Section 2 Evaluations

(a) Equipment Procurement

The procurement of office supply of binders, printer ribbon, printer paper were made. A computer workstation is secured to include:

IBM AT with
- -640K RAM and 384K LIM/EMS
- -1.2 M floppy and 720K 3.5 in. floppy drives
- -20 MB harddisk
- -IBM Proprinter
- -MS bus mouse
- -1200 Baud modem
- -Quadram+ EGA card and Multisync monitor,

Computer Desk and chair,

Televideo 970 terminal linked to VAX 11/785 running VMS.

A research account is established. This facility shall be used to import data from the traffic tapes from RADC/DCLD Laboratory and then download to the DOS files to be used on the AT.

A similarly equipped AT workstation in the Computer Science Laboratory at the University is reserved for the Student Research Trainee.

The secured equipment mentioned above seems to satisfy the current need well.

(b) Personnel

(i) Secretarial Assistance: Miss Wanda Morris from the Department of Mathematics and Computer Science at the University is secured for office assistance. Mrs. Jacqueline Clark substituted for Miss Morris soon after Miss Morris left in September.

(ii) Student Research Trainee #1: Mr. Antonio Ransom was selected from the 1987 senior class for his general knowledge in two fields: Electronic Technology and Computer Science. Mr. Ransom (personal vitae attached) worked during the period from Feb 1, 1987 to May 9, 1987. The Principal Investigator has given weekly lectures to Mr. Ransom on the topics included in this Project. Based on the tutored research topic, Mr. Ransom participated in the Student Presentation Competition of the 4th National Technical Association Student Symposium cosponsored by NASA from April 9-11, 1987. Mr. Ransom won the First place honor by presenting the topic of "Integrated Switching Networks" under the supervision of the Principal Investigator. Mr. Ransom was soon employed by AT&T of the research and development division in Chicago, following his graduation in May, 1987.

The effectiveness of this Project in training a student to develop research foundation and interest in the area interest to AFOSS is evaluated as "very satisfactory".

(iii) Student Research Trainee #2: A new student was selected to work for the Summer period from May 18 to July 31 1987. Miss Yvette Cherry, a Computer Science major (personal vitae attached)

was selected based on her programming skillfulness. She was instructed to learn C language and use the Microsoft C compiler. Weekly lectures on this research Project were also scheduled. Her responsibility was to learn as much as possible about the integrated networks and learn C language. Her specific assignment was to write the menus to be used in the simulation software.

(iv) **Principal Investigator**

The Administration duties: The Principal Investigator acquired necessary personnels and equipment as outlined in this project contract. The financial accountability was maintained through periodic examination of this project account at Grants Accounting Division of the Norfolk State University. The Student Research Trainees were taught by the Principal Investigator on the Integrated Network from simple basics. The responsibilities and working schedules for Student Research Trainees were defined. Their work was supervised. The working hours were recorded.

Research responsibilities: The Principal Investigator had a close worling relation with the Focus Point on the research work that is of current interest to RADC. In addition, the Principal Investigator attended the VDI meetings and Network Management workshop as well as being consulted. The Principal Investigator also helped to debug the PC version of the Traffic Generator. Thus successfully helped to port the Traffic Generator from a PDP-11/44 running VENIX to a PC running PC-DOS

The Principal Investigator felt that he has underestimated the completion time of the traffic generator that his IN

software depended on. Furthermore, the IN software development went through a major revision after it was found that the record format of the traffic file generated from the Traffic Generator has been changed in late August. The IN software as it stands now still contains some software bugs. The trial runs on the algorithms are meaningless due to short simulation runs that have been conducted so far. The Principal Investigator felt that the fully working IN software should come along shortly in the future. The Principal Investigator has every intention to see that it is done so and results reported to the Focus Point.

The Priority Handling Algorithms that the Principal Investigator developed are quite promising that they may find immediate application in the VDI Design Program of RADC.

Section 3 Research Results

Preparation

Through the phone conversation with the Project Focus Point at Rome Air Development Center, there was no new acquirement of any network simulation software in the Wide Area Network Laboratory. Thus the Principal Investigator decided to construct an entire simulation software to conduct priority handling algorithm experiment from scratch.

Due to the delay in the University's acquisition of a new computer. An immediate decision was made to start preparation work using an IBM AT.

Student Research Trainees were made to learn the C language , the MS-DOS on the AT and the use of the Microsoft C Compiler. The AT environment is not suitable for running large simulation programs, but the source code written in C shall be easily portable to new machine. Thus the AT and DOS are mainly used for development even though the environment is not as good as UNIX running on a minicomputer.

Work Conducted

1. **Preliminary Design Report**

The Principal Investigator has prepared a Preliminary Design Report in which he outlined the procedures and specifications to the simulation software. The Report was presented to the Focus Point at RADC, Griffiss AFB, NY. The details of the design was discussed and revised. (Please see Appendix B.)

2. Developing Priority Handling Algorithms

The Principal Investigator has outlined 3 experiments to be conducted on the simulation software after its completion. The experiments are testing the performance of the following priority handling algorithms:

- Dynamically adjusting the minimal data packet boundary such that the bandwidth allocation for voice and data is proportional to the "recent" statistics. The adjustment to decrease the voice bandwidth is made "gracefully" to occur only at the completion of some low priority call.

- Similar to item 1 mentioned above, but more drastically terminating lowest priority call without waiting for its completion.

- Data packets are stored in multilevel feedback queues with various "promotion" strategies.

Eventually the algorithms were evolved into the paper which was submitted to RADC. (Please see Appendix C.) The Priority Handling Algorithms developed include four bandwidth allocation methods, and four data queue promotion strategies in queuing and forming SENET frames.

3. VDI Design Program Consultation

The Voice/Data Integrator Design Program initiated by the RADC involves the implementation of Priority Handling Algorithms. Hence the Principal Investigator was invited to attend two meetings that RADC had scheduled with the program contractor: CNC

Electronics. The Principal Investigator was involved in the discussion of technical design sessions in both meetings. The Principal Investigator has a continuing interest in further involvement with the VDI Design Program.

4. The Integrated Node (IN) Simulation Software

The Integrated Node (IN) simulation software that we constructed was to use the traffic file generated by the Traffic Generator software developed at RADC. The record format in the generated traffic file was first designed such that each record be a data packet or a voice call. (Please see Appendix F.) Later it was changed such that each record is a data message with specified number of packets and remainder or a voice call. Thus our software must also be revised to accept and process records of the new format starting at the end of August, 1987. Our IN software was compiled, yet was not debugged, nor tested until November 8, 1987 when the PC version of the Traffic Generator was finally debugged by the Principal Investigator.

Some trial runs have been conducted on the IN software. It seemed that the IN software processed certain type of traffic files well for a short simulation run-time. Bugs occurred when simulation period was set long (beyond 100,000 msec). One major concern has been the queuing of data transaction records, since the IN software was written to run under PC-DOS (V3.2) which can only addrtess 640K bytes (other than ROM) of user's memory space. This number is further lessened by system memory use, and IN software code use. 25-byte buffers are used for each

transaction. This may cause some problem to the few data segments, limited to 64K each, when a large amount of transactions are to be queued.

The VAX8350* with its C compiler has arrived at Norfolk State University but yet to be installed as of December 15, 1987. It is the intention of the Principal Investigator that the IN software shall be ported to the VAX 8350 for continuing testing of the Priority Handling Algorithms. The statistics data are to be gathered, analyzed and reported in the future. The results shall be included in a paper to be prepared for publication in the future.

The source code listing is in Appendix D. The enclosed IBM formatted diskette contains the source code: IN.C, object code: IN.OBJ and executable code: IN.EXE. The softawre was developed on an IBM AT* running PC-DOS 3.20 and Microsoft C Compiler V4.00.

---

* VAX8350 is a trademark of the Digital Equipment Corp.

* IBM AT is a trademark of the IBM Corp.

Section 4 Future Research Plans

The performances of the various priority handling algorithms have yet to be determined. After the IN software is debugged, various testing of the algorithms will be performed. Data then will be analyzed and published.

The next step shall be the implementation of the algorithms in the VDI Design Program if RADC should find it desirable. Finally, the IN simulation plus the Traffic Generator can be expanded to include routing algorithms thus fully simulates an actual integrated node. These fully simulated nodes can then be used to simulate a network to study the interaction of integrated network nodes and network performance.

Appendices can be obtained from

Universal Energy Systems, Inc.

FINAL REPORT NUMBER 44
RECEIVED A NO-COST TIME EXTENTSION
TO BE SUBMITTED IN 1987 MINI-GRANT FINAL REPORT
Dr. Charles Lance
760-6MG-031

1986 USAF-UES Mini Grant Program

Sponsored by the

Air Force Office of Scientific Research

Conducted by the

Universal Energy Systems, Inc.

Final Report

A Neural Network Simulation Generator, Simulations of Learned Serial Behavior, and a Neural Explanation of Emergent Communication

Principal Investigator: David Lawson

Academic Rank: Assistant Professor

Department and University: Mathematics/Computer Science Department, Stetson University, DeLand, Fl 32720

Date: December 31, 1987

Contract No.: F49620-85-C-0013/SB5851-0360

Purchase Order No.: S-760-6MG-001

A Neural Network Simulation Generator,
Simulation of Learned Serial Behavior,
and a Neural Explanation of Emergent Communication

by

David Lawson

Abstract

This monograph has three distinct parts. First, is the development of Brain.Stetson, a laboratory tool which is able to generate neural network simulations. The second part is a series of experiments designed to reveal the neural architecture involved in learned serial behavior. The third is a discussion of insect behavior, and a proposed explanation of the emergence of communication.

Brain.Stetson is an automated program generator which can be used to build a neural network simulation. Its use will result in a substantial reduction in the labor involved in the creation of network simulations.

We use Brain.Stetson in a series of experiments. Each experiment consists of a neural network which is an extension of the network used in the prior experiment. The purpose of the series of experiments is to discover the neural principles and neural mechanisms involved in learned serial behavior. Running a maze is one example of learned serial behavior.

We are guided by an attempt to emulate insect behavior, much of which depends on the ability of the insect to learn a series or sequence of actions based on a sequence of inputs. The nature of learned serial behavior is therefore of interest to us.

We also found that by combining a suggestion by Karl von Frisch with a neural architecture proposed by Stephen Grossberg that we have discovered a neural explanation of emergent communication. The explanation proposes that an accidental use of the neural machinery used to control flight can result in the transfer of neural patterns from one insect to another.

Acknowledgements

I would like to thank the Air Force Office of Scientific Research for sponsoring this research. The research effort is a result of a grant received following my participation in the 1986 Summer Faculty Research Program sponsored by the Air Force Office of Scientific Research.

This research is the result of three distinct groups of people. It grew out of a collaboration with my brother Anton Lawson and my father, C. A. Lawson, and myself. My brother is a professor at Arizona State University and my father was a professor at the University of California, Berkeley. Both of them have long been active researchers in the nature of learning. David Hestenes of Arizona State University introduced us to the work of Stephen Grossberg. The second group is the Air Force Armaments Laboratory at Eglin Air Force Base, especially David Zeigler, with whom I worked so closely, and Dennis Goldstein whose comments have proven so very helpful. The third group are the students I have worked with at Stetson University. Barry Pekin and Melissa Titshaw are the initial architects of the simulation generator. Roy Hale and Brad Williams have been the people who developed the generator into a laboratory tool. Robert Brosofsky and John Carswell are responsible for our next direction, the development of autonomous vehicles.

Table of Contents

Background.

Part 1. The Neural Network Simulation Generator.

Introduction.

Section 1.1. The simulation generator.

Section 1.2. Machine installation.

Section 1.3. Details of machine operation:

the first step - creation of the file *name.txt*, a brain specification.

Section 1.4. Details of machine operation:

the second step - creation of the *name.ups* file and the library routines.

Section 1.5. Details of machine operation:

the third step - running the model.

Section 1.6. Details of machine operation:

the last step - observing the results.

Section 1.7 The architecture of the simulation generator.

Section 1.8. The implementation of the architecture.

Section 1.9. The file structure of the machine.

Part 2. Neural Mechanisms of Serial Behavior.

Introduction.

Section 2.1. The experiments.

Section 2.2. The neural control of sequential behavior: a summary.

Part 3. A Neural Explanation of the Beginning of Communication.

Introduction.

Section 3.1. Insects, intention movement and serial behavior.

Section 3.2. Feedback loops and communication as an emergent phenomenon.

Part 4. Recommendations

References

Figures

Background

Stephen Grossberg has developed a collection of differential equations which we shall refer to as the field equations of the mind. They are generic equations which describe the rate of change of the synaptic strength of a synapse, and the rate of change of the internal voltage of a cell. The assumption is made that the brain can be modeled by a collection of slabs of neurons. A slab could be the retina, the Lateral Geniculate Nuclei (LGN), or the visual cortex. A slab could be some other structure. The cortex is six separate layers, and a model of the cortex could consist of six slabs. A slab for Grossberg is a collection of neurons with a specific function.

figure 1.

A model as a collection of slabs. Each slab is a collection of neurons.

We will give you a brief explanation of why Grossberg has singled out the rate of change in internal voltage, and the rate of change of synaptic strength.

A slab is active when its neurons are firing. Normally neurons will not all fire at once. Pattern registration on a slab (the pattern on a slab) does refer to those neurons which are firing at a specific time. But, it is probably more accurate to say that pattern registration refers to the frequency that each neuron on a slab is firing at a specific time. The rate at which a neuron is firing can be derived from the rate at which the

internal voltage of the neuron is changing. Thus, the acivity of a slab and of the entire brain can be characterized if one can characterize the rate of change of the internal voltage of each neuron.

figure 2. Two neurons.

figure 3. The generic equation for the rate of change of internal voltage.

It has long been conjectured that the rate of release of neurotransmitters will change due to activity of the synaptic knob. Experiments by Eric Kandel and others have verified at this is so. Thus, it can be shown that a pattern of activity which is repeated across a slab will cause the synaptic strength at the site of activity to increase and allow the pattern of activity to be recovered. In other words, long term memory (LTM) seems to be closely related to synaptic strength. Thus, the rate of change of synaptic strength is related to the rate of learning ang forgetting.

figure 4. Grossberg's generic equation for the rate of change of synaptic strength.

Of course the assumption is made that to discover the nature of thought it will be necessary to discover basic principles that govern the

activity of the brain. Neural modelers tend to believe that this means they must discover structural principles. One such example is the ubiquitous On-center, Off-surround (OCOS) archtitecture which can be used for edge detection. OCOS can also be used for gain control. There is also a question of stability. A neural network is a dynamical system, and as such can become hyperactive, inactive, have one or more attractors, etc.

Grossberg has developed a sophisticated structural theory of pattern registration, pattern recognition, and more, in which groups of neurons refered to as slabs have special processing features. Grossberg's theory of adaptive resonance refers to the manner in which the slabs interact with each other.

Brain.Stetson is a tool for the neural modeler. It builds simulations that he can use to test his hypotheses. A neural modeler deals then with connections (the structure of the neural network), and with variables associated with the networks connections and nodes. The variables change over time. The simulations we build have variables of this sort. How they change depends on how the modeler wants them to change. He is able to supply subroutines that will determine how they change. We use variables that we have derived from Grossberg's investigations, and we use subroutines which reflect his differential equations. A modeler can, however, interpret the variables in anyway he wishes (they are just storage locations). He can include routines to implement his own theoretical considerations, and he can define and alter the variables as he wishes.

Part 1. The Neural Network Simulation Generator.

Introduction

The simulation generator, which we refer to as Brain.Stetson, is a general purpose network simulation generator. The experimenter describes the network he wishes to build. He then adds Pascal routines to an internal library which will change the state variables in his network in the manner he wishes for them to change. Brain.Stetson will then generate a program which can be run to simulate network operation.

Once a network has been built it is very easy to experiment with it. One can easily alter or add update routines (routine which change the internal variables of the network). At the same time the modular nature of Brain.Stetson keeps the experimenter from tinkering unnecessarily with his simulation. The initial simulation will remain and can easily be recovered if the experimenter so desires.

We now describe the steps one must take to use Brain.Stetson. Each section to follow will describe a separate step.

Section 1.1. the simulation generator

We have built a simulation generation machine, which we call

...ain.Stetson. The purpose is first: the creation of a neural network, built to specifications presented by a neural modeler, and second: a program capable of running the network created. The machine runs on a VAX 11/750, and uses VAX Pascal, and the DEC Command Language (VAX DCL). The generation machine describes a network. The network consists of slabs of nodes. Each node is thought of as a neuron. All features of the model can be defined by the modeler.

The modeler defines:

1. the connections of the nodes.
2. the update routines.

The update routines determine the transition rules, the state change of a node from time t to time t+1.

Each node of the model has a list of internal variables. These variables are the standard features of the abstract neuron, the synaptic strength, a synaptic decay rate, and the internal voltage of a neuron, and its decay rate. Each neuron can be connected to several other neurons. Each connection (each synaptic knob) has a flag indicating whether the synapse is inhibitory or excitatory.

These internal variables are the variables subject to update. In reality these variables can be ignored and others can be substituted or added. This is possible because it is the update routines that determine the transition rules, and thus the update routines determine whether or not an internal variable is actually used in the model. The routines are written by the user, and are thus under his control.

The simulation machine has been built and refined over the past two years by a team of programmers under the direction of David Lawson. All of the programmers are students at Stetson University. Melissa Titshaw and Barry Pekin are responsible for the initial design of the machine. Roy Hale is responsible for the successful implementation of the machine. Brad Williams is reponsible for the Subroutine Library Facility, the feature which allows modelers to define transition rules of their own design.

Section 1.2. software installation

If you have a VAX you can use Brain.Stetson as a neural modeling tool. Your first step is to load the software into your account. Having done this you must then go through the installation procedure we describe below.

To install the Brain.Stetson software it is necessary to define pathways to various directories, and then compile and link the routines which contain those pathways. This is necessary because Brain.Stetson uses the VAX tree structured pathways of which the username is an integral part. In otherwords, the user must place the pathway to his directory in the appropriate places in the Brain.Stetson software.

figure 1.1

The installation procedure

Section 1.3. details of machine operation:

the first step - creation of the file *name*.txt, a brain specification.

The modeler must first create a text file. This file is used by the brain builder to define the number of nodes in the model, and their connections.

The complete list of the contents of the text file are as follows:

figure 1.2

the contents of a ----.txt file

figure 1.3

an example of a ----.txt file.

figure 1.4

a second example of a ----.txt file

Steps to creation of a ----.txt file.

To create a text file you must follow these simple steps:

First: Log on to your account.

Second: Type GO (and hit the RETURN key)

The RESULT -->

figure 1.5

Third: enter the number 1 (and hit the RETURN key).

The RESULT --->

figure 1.6

Fourth: type in a name (example: Barry) (and hit RETURN) We will use Barry in the rest of our instructions as a generic file name. Where ever Barry appears just substitute the name you typed here.

The RESULT ----> you return to the earlier menu. Enter the number 4 (to quit) and hit RETURN.

The RESULT -----> The file Barry.txt has been created. You can now edit it and enter the proper values for the connections, for the forgetting factors, synaptic strengths, etc.

to do this:

Fifth: type GOTEXT and hit the RETURN key (we will not mention the RETURN key again. We assume everyone is aware of its use) . GOTEXT will place you in the proper subdirectory (the subdirectory which holds the ----.txt file that was created).

Sixth: type set term/width=132

Seventh: type EDIT Barry.txt

Eighth: use the VAX editor of your choice to enter the proper values.

Section 1.4. details of machine operation:

the second step: creation of the name.ups file and the library routines.

Each slab has an associated set of update routines. A slab is meant to be a two dimensional array. As an example, a brain could consist of a retina, an LGN, a visual cortex, and a motor cortex, and each could be a slab. The neurons in each could be subject to different update rules. The retina could take its input from a file, and thus input to a neuron, and the resulting question of whether or not it fired would depend upon the contents of the file. The LGN would take its input from the retina, and feedback from the visual cortex, and have a standard update definition. The visual cortex could also have a standard update routine. The motor cortex could result in motion of the brain, and the retina could then view a different potion of the world. This would require a resulting difference in the file which contains what the retina is looking at. The update routine for the motor slab would have to change the input file.

Standards associated with update routines:

1. they are written in Pascal.
2. each routine is placed in the ----.Library, and has the name *name*.lib.

To view an ---.ups file type GOTEXT. The ---.ups files are in the same

subdirectory as the ---.txt files.

Standards associated with the Ups file.

1. Each slab has its own update routine for voltage update, history update, synaptic update, and voltage reset.

The point: **Each slab is a functional unit with its own set of update routines.** If you feel part of the model should have a different update routine then you must make it a unique slab.

The voltage update is the name of a subroutine that will change the internal voltage of a neuron on the slab in question. The synaptic update is a routine that will change the synaptic strengths of synapses from neurons on the slab. The history update updates an array, a 1 if the neuron fired at time t and a 0 if it didn't. The reset routine will reset a neuron to whatever is specified once it has fired. This is the intent. The modeler can actually do as he wishes.

2. The names of the routines must be listed in a special order. First, list the voltage update routine for each slab. Next, the history update for each routine must be listed. Third, list the synaptic update for each slab, and finally the reset update routines for each slab.

Below is a sample ---.ups file.

figure 1.7

An ---.ups file for a model with two slabs.

To view an update routine type GOLIB. The update routines must all be in the subdirectory in which you find yourself. This is the directory User$disk:[UDD.Neural1.progs.modeltexts] on our system. **Each update routine must end in a .LIB suffix.**

figure 1.8

A standard update procedure written to implement Grossberg's genreic equation for dV/dt.

Section 1.5. details of machine operation:

the third step - running the model.

To run a model requires the following sequence of commands:

Step 1. type GO.

eg. $go

The RESULT --> the menu below will appear.

figure 1.9

The Main Menu

Step 2. Select an option. You will want to select option 2 (eg. type 2), because you have just created a ---.txt file and a ---.ups file. (These must be created before option 2 can be run succesfully). If you have already selected option 2, then you can select either option 2 or 3. If you

select option 2 then a new brain (a new model) will be created. If you select option 3 then the brain you have already created will be run again. If a model is 'run again' it means it will continue to develop. His synaptic strengths will continue to change, his voltages, etc., and they will change from what they were when the model was last run.

In other words:

Option2 - wipes out the old model (if there was one) and a new one created.

Option 3 - develops the present model from its present state.

Section 1.6. details of machine operation:

the last step - observing the results.

The following command sequence will allow one to view the state of the model following an experiment:

Step 1. type go (eg. go to the main menu).

Step 2. select option 3 (run an existing model)

Step 3. follow the directions (pick a model from the the list of models displayed).

Step 4. follow the directions (pick option 1. Examine my brain)

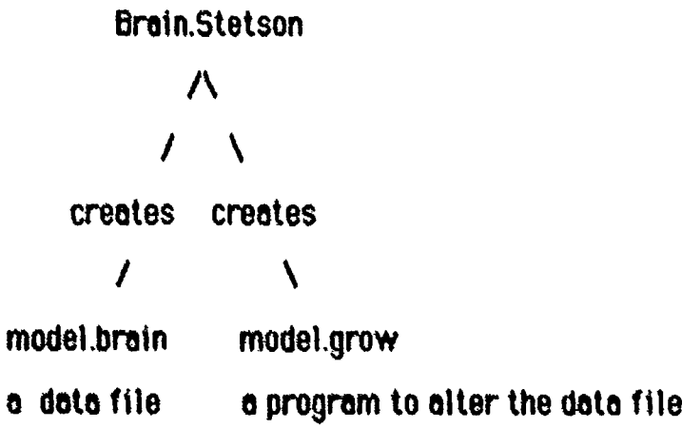
Step 6. follow the directions.

figure 1.10

An example of the output given by the model. This is Slab 2 of Ave2.

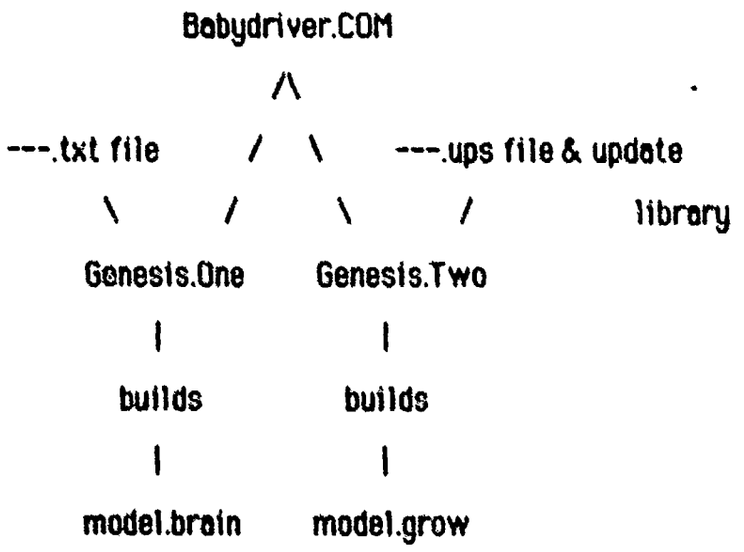
Section 1.7. The architecture of the simulation generator

Brain.Stetson is a simulation generator. As such it has two parts. First it **builds the brain** - a data file. Next, it **builds a program that** updates, or **alters the brain** as time progresses from time step t to time step t+1.



Section 1.8. the implementation of the architecture.

Brain.Stetson was developed on a VAX 11/750. The driver is written in DCL (DEC Command Language), and is a COM file called Babydriver.COM. Babydriver.COM creates two separate PASCAL programs, Genesis.One and Genesis.Two. The two programs are then compiled, linked and run, with Genesis.One creating the brain (the data file) and Genesis.Two creating the program to update the brain.



Because we have designed Brain.Stetson is this fashion we are able to build different size brains, and include different update routines.

Section 1.9. the file structure of the machine.

The VAX has a tree structured or branching file structure and thus Brain.Stetson does as well. We have included in the com file several commands that make it easy to move around within the tree. Once Brain.Stetson is installed you will be able to use the following commands:

$ godoc - to see documentation

$ gotext - to see the ---.txt and ---.ups files

$ gomod - to see the models

$ golib - to see the update routine library

$ gohom - to get back to the root directory

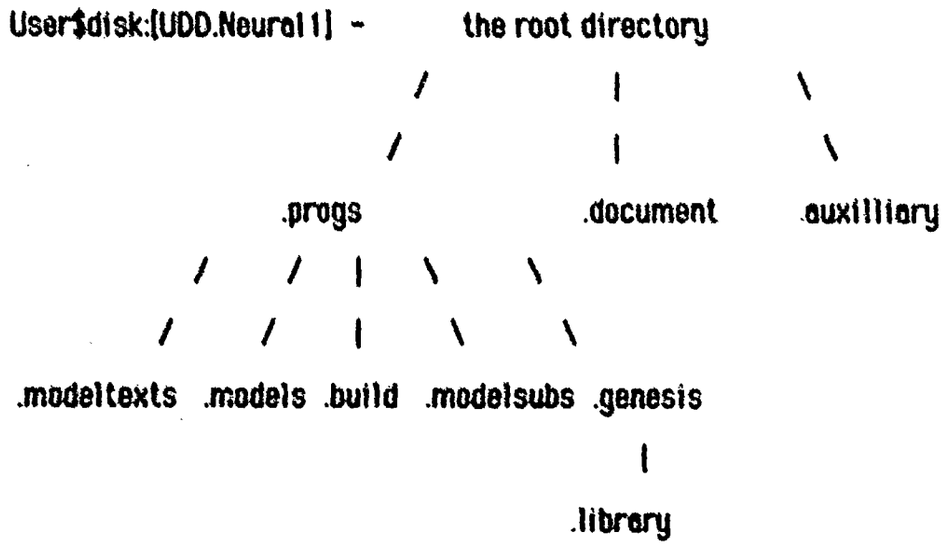
$ gogen - to get to the Genesis directory

$ goprogs - to get to the progs directory

$ gosubs - to get to modelsubs directory

$ gobld - to get to the build directory

This list can be modified of course by changing the login.com file.



Part 2. Neural mechanisms of serial behavior

Our primary concern is and has been with insect navigation, tracking, and target recognition. This encompasses a broad range of behavior which is in part context dependent. Underlying this behavior are neural mechanisms of serial or sequential activity. For this reason we are developing neural models of serial behavior.

Navigation itself has many facets and many subtleties. Different insects have different abilities. In addition, different situations can emphasize different sensory input. Insect sight can be very sophisticated. Bees, for instance, use both landmarks and solar cues for navigation. This involves pattern recognition of some sort, in addition to cues for headings using the position of the sun. Smell is often used for navigation. There is another important consideration. In nature insects do not often perform isolated acts. Instead they perform acts in a continuous series. Each act is dependent on the completion of a preceeding one, not just receipt of the appropriate stimuli. Navigation, tracking and target recognition are infact not even separate activities. They are instead intimately related, and sequential in nature.

As an example, Tinbergen has shown that the hunting behavior of the bee-wolf is sequential. Successful hunting behavior (eg. prey-found) involves first a visual stimulus, the prey must be moving and of the right size; at this stage there is no response to the odor of normal prey even if it is presented close to the wasp. Once the potential prey has been spotted

the wasp flies downwind of it. At this point odor becomes the dominant stimulus. Dummy prey will not be appraoched unless supplied with the correct odor. Finally the wasp seizes the prey, and stings it, evidently employing a tactile sense.

It seems likely that sequential chains are of great importance in insect pattern recognition itself. Such chains form a context in which the identification occurs. In other words, insect shape detection need not be nearly as sophisticated as ours, and yet their identification can be better, because they get can extra cues from context (behavioral chains) as well as from other sensory modalities. For this reason serial or sequential behavior is of great importance, and will form our first topic for neural model experimentation.

Section 2.1 The experiments.

All of the experiments contained within are reproduceable. Our goal is to build a neural model that is able to learn a maze. We have chosen this goal because ants are capable of running mazes, and it would seem to incorporate short term memory, and the transfer from short term to long term memory. It will force us to face questions of inherent interest in such tasks. We are forced to try to determine the neural nature of rehearsal and reward and to discover a neural solution to the credit assignment problem (the question of how one decides exactly what neural

connections are to be rewarded, and by how much).

Experiment 1. Swapper.

The purpose of Swapper is to learn a list. We present Swapper with a list. Typically we would present the same list to Swapper 10 times. We then examine Swapper's neural connections and determine whether of not Swapper was able to associate the elements in the list correctly.

An important note: We will present Swapper with a list of numbers. We could present the numbers to a slab which would have to recognize the number presented. We do not do this. We assume that the number is recognized, and we assume that because of this recognition, a neuron or group of neurons (which recognize this number), then begins to fire.

Swapper's neural architecture.

Swapper is a three slab model. Each Slab has 6 neurons.

Slab 1 is the input slab. The input slab reads a file which contains the list. If the element read is a 1, then neuron 1 of Slab 1 is to fire, if the element read is a 2 then neuron 2 of Slab 1 is to fire, etc. Each neuron in Slab 1 is connected to the corresponding neuron in Slab 2.

Slab 2 is the association slab. Each neuron is Slab 2 is connected to every other neuron in Slab 2. Each neuron is Slab 2 is also connected to the corresponding neuron is Slab 3.

Slab 3 is the output slab. IF neuron n in Slab 2 fires at time t, then neuron n in Slab 3 will fire at time t+1, and n will be written to a file.

Slabs 1 and 3 are really nothing more than a convenient way to

construct our model. Slab 2, the association slab, is the slab of interest.

figure 2.1

Swapper.

Each neuron in Slab 2 is connected to every other neuron in Slab 2. If the list is 1, 2, 3, 4, then at time 1 neuron 1 in Slab 1 will fire. At time 2 neuron 2 on Slab 1 will fire (because Slab 1 is reading the list), and neuron 1 on Slab 2 will fire (because neuron 1 on Slab 1 fired the time step before). At time 3 neuron 3 will fire on Slab 1 and neuron 2 will fire on Slab 2. **Our synaptic update rule is Hebbian**. This means that if neuron **a** fires at time t and neuron **b** fires at time t+1, then the connection strength (the synaptic strength) between neurons **a** and **b** is increased. If **b** does not fire at time t+1, then the synaptic strength connecting **a** to **b** is decreased. Thus, if neuron 1 on Slab 2 fires at time 2 and neuron 2 fires at time 3, then the synaptic strength between the two is increased.

Swapper was able to learn the list. One difficulty did arise. Once Swapper has learned a list he cannot learn another. Instead, when presented with a second list, Swapper quickly becomes overexcited, with all of his neurons firing simultaneously. The reason for this: suppose Swapper has learned the list 1, 2, 3, 4. Suppose that Swapper is then presented with the list 1, 3, 5, 6. This is what happens: neuron 1 fires neuron 2 because the synaptic strength between 1 and 2 is high (since

Swapper learned 1, 2, 3, 4. At the same time 1 is followed by 3. Thus on Slab 2 neurons 2 and 3 fire simultaneously. Since 2 fires, the synaptic strength between 1 and 2 continues to increase.

We would have hoped that Swapper would forget the old list and learn the new one. That does not happen. **Swapper can learn one list, but he cannot forget it.**

Experiment 2. List-Learner.

List-Learner has the same number of slabs that Swapper has, and each slab has the same number of neurons (6). But, List-Learner has additional connections going from Slab 1 to Slab 2.

To construct List-Learner we have taken Swapper and added a **feed-forward on-center,off-surround (OCOS)** structure from Slab 1 to Slab 2. This means that from neuron 1 on Slab 1 there are inhibitory connections to neurons 2, 3, 4, 5, and 6 on Slab 2, as well as the excitatory connection to neuron 1 which is already part of Swapper. This is true for each neuron is Slab 1. This quiets Slab 2 down. **List-Learner can learn a second list.** If a second list is presented to List-Learn, then the feed-forward OCOS inhibits the old list. It does not fire, and the synaptic strengths then die away from disuse.

figure 2.2

List-Learner.

Experiment 3. Avalanche

List-Learner can learn a list 1, 2, 3, 4 and another list 5, 6, 1, and given the first element of either list the correct list can be recovered. This is not true for the nexte example. Given the two lists 1, 2, 3, 4 and 5, 2, 4, 3 it is not possible to correctly recover the element following 5, 2. This is because List-Learner can effectively only look one time step into the past. It will respond the same for 5, 2 and 1, 2. We will use an addaptation of Stephen Grossberg's notion of Avalanche to solve this problem.

figure 2.3

Time-steps side-by-side. This is as if two frames of a movie were appearing on the slab at the same time.

Time t and time t+1 can appear on a slab at the same time. In other words the brain can go backward in time. A method of capturing this notion neurally is to have axons of different length. Suppose one axon, axon 1, is twice the length of another, call it axon 2, and suppose both come from the same neuron N. If neuron N fires at time t, then axon 1 can report that at time t+1, and axon 2 at time t+2. This allows several different things to occur. For one thing, the sequence t, t+1 can be recovered. For another the state of neuron N at times t and t+1 could coexist side-by-side. We use this second option in the construction of Avalanche.

There is another way to say this. Axon 2 is a delayed version of axon 1, and **Avalanche keeps a copy of past events by adding delays.** Because we have given our axons lengths (they can be of length one, two, etc.) we are able to incorporate a delay by altering the length of an axon. There are other ways to do this of course.

figure 2.4

The architecture of Avalanche

Avalanche is List-Learner with an additional slab, a new slab 3 lying between Slab 2 and the Output Slab (the old Slab 3). **Each** neuron in Slab 2 was connected to **each** neuron in the new slab. Infact, each neuron is Slab 2 has 2 axons going to each neuron in the new slab. Of these two axons, one is delayed. In other words, one has length one and one has length two.

Avalanche was a disaster. It was easily overexcited. A slight modification was, however, very successful. That modification is Ava2.

Experiment 4. Ava2

To create Ava2 from Avalanche we merely disconnect most of the neurons going from Slab 2 to the new slab. In Ava2, neuron 1, Slab 2 is connected to only one neuron (neuron 1) of the new slab (the connection was still by two axons). Neuron 2 of Slab 2 is connected to only to one neuron on the new slab (neuron 2), etc.

In other words, to create Ava2 begin with List-Learner, add a new slab between Slab 2 and the Output Slab, and have neuron k of Slab 2

connected to neuron k of the new slab by axons of length 1 and 2. Then have all of the neurons of the new slab connect to each other.

Ava2 now worked as we hoped it would. It did learn a list with changing context. It was possible to recover the list by examining the synaptic strengths of the various connections. Unfortunately it suffered from the same sort flaw as Swapper. If Ava2 was given another context dependent list, it could not forget the first that it learned and relearn the second.

To solve this we could add feedforward OCOS from Slab 2 to the new slab. (eg. this would solve the problem just as it was solved before).

figure 2.5

The architecture of Ava2

We could also solve the problem in another way. Slab 2 (List-Learner) and the new slab (Avalanche) could be just 1 slab. This was conceptually what we meant it to be. To do this make neuron k of the new slab into neuron k of Slab 2. Then Slab 2 becomes the combination of List-Learner and Avalanche. Since List-Learner has OCOS this will give OCOS to Avalanche as well.

Experiment 5. Reward

The architecture of Ava2 insures that the last X time steps can be recovered, where X is the length of the longest neuron. The purpose of Reward is to take advantage of this fact.

For List-Learner to learn a list the list had to be presented to him

several times. We know that we ourselves can occasionally remember things that have occured only once. Infact STM is thought to consist of a feedback loop, with neural impulses spinning on themselves. This is a good explanation of experimental data, and it certainly makes sense. Thus, **Reward contains a feedback loop**. This loop is not always activated. If it were impulses from the past would mingle and interfere with incoming sensory impulses. This may occur. But it may not. In any case, in our first modeling attempt, we choose two options. First, the network could be a feedforward network, just like Ave2. Alternatively, the input slab could be disconnected, and input can come instead from slab 3. This is **feedback. It results in STM and does transfer STM into LTM.** Our first experiments were to have REWARD occur at the end of the data file (EOF). Thus, presenting the model with the list 1, 2, 3, 4, 5, 6 and the EOF (End of File) resulted in the model learning 4, 5, 6.

In other words, Reward allowed the model to spin of the data that caused him (Reward) to be rewarded.

figure 2.6

The architecture of Reward

Reward has a single axon from each neuron k in Slab 3 that connects to neuron k of Slab 1. The voltage update routine for Slab 1 normally takes input data from a file. When it hits EOF it will use the connections from Slab 3 to get input data from Slab 3 (and not from the file). It thus operates as though input from outside the model was inhibited and came instead from slab 3.

Experiment 6. Reward2.

Reward was succesful. But we wanted to extend the ability to reward into the past. In other words, we did not wish to reward only the last three choices, correct moves etc. Once these had been rewarded we wanted them (4, 5, 6 for example) to be able to trigger the reward mechanism themselves. To do this we added a reward neuron and sent axons from each neuron in Slab 3 to the reward neuron. Those axons which fired on the reward neuron had their synaptic strengths increased and were able to fire reward themselves. Once the reward neuron fired we had Slab 1 take its input from Slab 3 instead of from the input file.

figure 2.7

The architecture of Reward2

The result was that 4 and 5 could trigger a reward, causing 3, 4 ,5 to be learned. 3, 4, 5, 6 was then the new chain and reward could be pushed into the past.

Experiment 7. Attention.

We found that Reward2 had a flaw. We weren't able to stop rewarding (we could not stop the STM) and return **attention to input** from outside the model. Thus, we had to add attention, a pseudo reticular formation, to the model. Actually, the attention neuron read from a file. A 1 read by the attention neuron at time t would cause the update routine for the input slab to take input from a data file at time t+1. Reward would then cause

STM to occur, and attention would cause STM to cease and normal operation to continue.

figure 2.8

The architecture of Attention

Mazerunner - The next step.

We have yet to present Attention with a maze. It may be that it is ready to learn a maze. The next step is to add a **Decision Slab**. We expect that a decision slab would be a sinner-take-all slab. A winner-take-all slab is a slab with OCOS, such that exactly one element fires. Then confronted with a choice, turn right, turn left, go straight, or go back, the winner-take-all slab would pick one.

Section 2.2 The neural control of sequential activity: a summary

This monograph began as an investigation into the neural mechanisms which underlay tracking behavior by an insect. Such behavior involves prediction of the path the prey will take. This path may be nothing more than a trajectory, but more sophisticated behavior is a sequence, or

series, of actions. As such, successful prediction requires that the predictor be able to learn a behavioral sequence. Once one embarks on the investigation of tracking behavior it soon becomes clear that a much more general phenomenon is involved. This more general phemomena is the neural mechanisms involved in learning any serial task. Insects do infact have the ability to learn serial tasks. The Beewolf learns the location of a series of nest sights. Ants have been taught to run a maze.

Maze learning can be viewed as a paradigm of serial behavior. We have chosen to do so. Our investigations have resulted in a series of experiments, each embodying a neural principal, and each an extension of the one which preceeds it. Our goal, which have yet to attain, is to build a neural model which is able to learn a maze.

A summary of our experiments:

name	purpose
1. Swapper	- a net which can learn a list.
the problem	- unable to learn a second list.
neural principle	- LTM as Hebbian association via synaptic strength. Synaptic strength given a maximum possible value.
2. List-Learner	- a net which can learn a series of lists.
neural principal	- On-center, Off-surround.
3. Avalanche	- a net which incorporates the past. this net can associate events which occur at different times.

the problem - overexcitation. Incorrectly conceived, the slab added is excitatory only, not OCOS.

neural principal - an adaptation of Stephen Grossberg's avalanche. A delay is added (equivalently axons are given different lengths).

4. Ava2 - just fix Avalanche by getting rid of some neurons.

5. Reward - a reward slab able to inhibit input and induce STM. This can also be interpreted as a neural implementation of rehearsal.

neural principal - STM as cyclic neural activity induced by a feedback loop.

the problem - can't stop reward (can't interrupt STM activity).

6. Attention - Add an attention slab.

neural principal - must have attention (reticular formation activity) to learn (to break out of STM).

the next step:

7. Mazerunner - Can learn a simple maze.

neural principal - winner-take-all decision slab.

Part 3. A neural explanation of the beginning of communication.

We will develop a neural explanation of the rise of communication within the animal world. We propose a neural mechanism which will explain the rise of communication as an emergent phenomenon, an accidental application of the neural machinery used to control flight.

Our development will draw on two diverse sources. We have combined Stephen Grossberg's theoretical model of the development of speech with the behavioral data of the ethologist.

Section 3.1 Insects, intention movements and serial behavior.

We have studied and attempted to emulate insect behavior because insects are behaviorally sophisticated, and yet neurally and behaviorally simpler than man, the mammals, or vertebrates in general. Since insect behavior is simpler, our hope is to discover or develop neural models which will exhibit this behavior.

Our investigation of insect behavior has led along many interesting avenues. The waggle dance, that famous method of bee communication, is perhaps the most startling, but many others exist. Consider, as an example, the intention movement. We begin with an example:

'Birds do something like this: when a bird is ready to take off,

it stretches its neck in the direction of its flight. Such intention movements, as they are called, sometimes influence other animals. In a flock of birds the movements can become infectious and spread until all of the animals are making them.'

(von Frisch, 1962)

An intention action of an animal seems to have no apparent purpose. It does signal the advent of a coming action, but is this of value? It may not be, it may be nothing more than an artifact of the machine, just part of the way it works. But, von Frisch, the man who discovered the waggle dance of the bees, found them of interest. He followed the observation above with another:

'It is possible that among the honeybees the strict pattern of the wagging dance gradually developed out of such intention movements performed by forager bees before they flew off toward their goal.'

(von Frisch, 1962)

We have begun to investigate neural models of serial behavior. Our investigation has led to a question of the neural nature of Short-Term Memory (STM), and of the transfer of STM to Long-Term Memory(LTM). Our models of STM (and independent psychological research) indicate that rehearsal is necessary to transform STM to LTM. This translates within our neural model to something similar to a *dream state* (eg. neural

activity with motor function disconnected).

We all know of instances of dreaming when motor function is not completely disconnected. Sleep walking and talking are good examples. **We conjecture that intention movement is this sort of phenomenon, an analog of sleep walking and talking.**

There are interesting behavioral quirks that can be explained by our conjecture. One such is the following example of insect intention movement. A moth alighting from a flight rocks back and forth rythmically on its feet for a time. The duration of the rocking tends to be related to the length of the flight it has just completed.

We conjecture that this activity is the result of STM activity of the neurons controlling flight, but with motor activity (the crucial portions) disconnected. Some sort of rocking can remain and will do no harm.

Section 3.2 Feedback loops and communication as an emergent phenomenon.

We have conjectured that intention movements are a result of STM activity with motor activity only partially subdued. Assuming that our conjecture is true we can demonstrate that insect neuroanatomy will lead naturally to an emergent model of communication. We begin with a brief description of Grossberg's neural explanation ofspeech (as a motor

activity).

Grossberg's model is a beautiful explanation of the manner in which feedback can be used to finetune a neural mechanism.

figure 3.1

Grossberg's neural model of speech.

A Macrocircuit for the Self-Organization of Recognition and Recall

I have left much of Grossberg's detail out. For a complete explanation I refer you to figure 1 of the article ' Neural Dynamics of Speech and Language Coding: Developmental Programs, Perceptual Grouping, and Competition for Short Term Memory ', by Cohen and Grossberg in Human Neurobiology, 1985.

The intent is that as a child speaks a word or utters a sound the ear receives the sound. The sound waves produced by the vocal track are fed back into the ear. Then, neural connections between the two tracks, the auditory track receiving the sound and the motor track (vocal track) producing the sound can work together. Neural nets or feedback loops connecting the two tracks insure that motor activity can be corrected and finetuned. The motor track produces the activity of course and the auditory track monitors the result. Sophisticated connections between the two must be necessary to allow an organism to finetune motor behavior to fit the environment. Grossberg's neural macrocircuit is a suggested neural architecture that will serve this purpose.

Infact, Grossberg's model is not merely proposed as a method of finetuning motor control. Rather it is a method of using motor activity to help define words, and sentences. He has shown how the recogniton and grouping of sounds at a neural level can be defined by (or become part of) a sensory-motor neural structure. This structure internally develops neural patterns during operation.

We suppose that insects have such neural mechanisms, neural mechanisms that can control, and finetune motor behavior. We mention the following fact:

> 'Sensory input from sensilla on wings, pressure sensitive hairs on head, abdominal nerves and others are fed into multimodal interneurons ... and the insect makes small adjustments to steer in flight.'
>
> (Howard Evans, 1984)

Multimodal neurons are neurons with different functions or different modes. An interneuron can collect data from two or more different senses. There will then be feedback to two or more senses. This sort of architecture can connect two senses. A pattern set up by one sense can develop a pattern within the interneurons which then feed that pattern (in the appropriate form) to another set of senses. It is suspected that it is this sort of phenomena that allows a bee to translate data it receives during the waggle dance about inclination to the vertical into information about angle (or inclination) from the sun. The conjecture is that data from

gravity and sun both connect to the same navigational interneurons. Feedback loops could then translate sensory data from one into sensory data from the other. This would explain why a bee performing a waggle dance is able to perform the dance on a vertical plane in the hive, or (just as easily) on a horizontal plane if exposed to sunlight.

In any case, if a rocking fly were to set up vibrations large enough for another to receive them, and if these vibrations were connected to an internal STM pattern it would be possible to transfer the internal STM pattern from one fly to another.

figure 3.2 Communication from one animal to a second.

In other words, sensory input from receptors on the insect allow it to make small adjustments in flight. These sensory neurons are connected to internal neural patterns which control flight.

- 1. Assume that the neurons which control flight are responsible for intention movements.
- 2. Assume that if these intention movements are received by the sensory neurons, then they will reproduce the internal pattern which controls flight in another animal.
- **3. The result : communication.** The transfer of a neural pattern from the brain of one insect to a similar pattern in the brain of another.

If we are correct, then communication has emerged from feedback loops which control flight, and intention movements which are closely enough related to flight to be able to reproduce the internal neural activity which produced them.

Part 4. Recommendations.

Neither Part 2 nor Part 3 are complete. Extensions are required to both.

For Part 2, the study of serial behavior and STM, two immediate steps remain. First, we must design and test Maze-Runner. Can we build a neural network that can learn a maze? If we are successful, the what additional properties will the network possess? Can it exhibit latent learning for instance? In addition, now close is our model of STM to human STM? If we present our model of STM with one of the standard psychological test, how will it respond? There is one ingredient of explanations of psychological test data on humans that our model does not have. That missing ingredient is a neural implementation of chunking.

Part 3, a theoretical explanation of emergent communication, needs to be tested in some fashion. We have begun to build a robot that can be controlled by a neural network simulation. We hope to build two such robots, and then see if we are able to transfer a neural pattern active in one network to the other, using the method we have described.

REFERENCES

Alloway, T. M. (1972) 'Learning and memory in insects', Annual Review of Entomology, Vol. 17, 43-56.

Cartwright, B. A. and T. S. Collett, (1983) 'Landmark learning in bees,' Journal of Comparative Physiology.A, Vol. 151, 521-543.

Cartwright, B. A., and T. S. Collett, (1979) 'How honey -bees know their distance from a near-by visual landmark,' Journal of Exp. Biol.,Vol. 82, 367-372.

Cartwright, B. A. and T. S. Collett, (1982) 'How honey bees use landmarks to guide their return to a food source,' Nature, Vol.295, Feb. 18, 560-564.

Cohen, Michael and Stephen Grossberg, (1985),"Neural dynamics of speech and language coding: developmental programs, perceptual grouping, and competition for short term memory." Human Neurobiology.

Collett, T. S. and M. F. Land, (1975), 'Visual Spatial memory in a hoverfly,' Journal of Comparative Physiology, Vol. 100, 59-84.

Evans, Howard Ensign (1984) Insect Biology, Addison-Wesley.

Frisch, Karl von.(1962) 'Dialects in the Language of the Bees', Scientific American, Aug. 1962, pp. 79-87.

Frisch, Karl von. (1971) Bees| Their vision, chemical senses, and language, revised edition, 1971, Cornell University Press, Ithaca and London.

Frisch, Karl von. (1967) The Dance Language & Orientation of Bees, The Belknap Press of Harvard University Press, Cambridge, Mass.

Grossberg, Stephen, (1982), ' Why do cells compete? Some examples from

Visual perception,' The UMAP Journal, Vol. III. No. 1.

Grossberg, Stephen, (1983),'Associative and Competitive Principles of Learning and Development,' Competition and Cooperation in Neural Networks, Amari, S. I., and M. Arbib, editors, Springer-Verlag, pp 295-341.

Grossberg, Stephen, (1983), 'A psychological theory of reinforcement, drive, motivation, and attention,' Journal of Theoretical Neurobiology, September, 1983.

Griffin, Donald R. (1976) The Question of Animal Awareness, 1976, The Rockefeller University Press, New York.

Lawson, David, Roy Hale and Brad Williams, (1987),"Brain.Stetson: A tool for the design and test of neural network simulations," Proceedings of the IEEE Southeastcon Conference, Tampa, FL, April,1987.pp.530-534.

Schneirla, T. C. (1953) 'Modifiability in insect behavior', in Insect Psychology, K. D. Roeder (ed.), John Wiley & Sons, Inc.

Tinbergen, N. (1951) The Study of Insects, Oxford University Press.

Wehner, R. and M. V. Srinivasan, (1984) 'The world as the insect sees it,' in Insect Communication, Trevor Lewis (Ed.), Academic Press (London), 29-47.

Wilson, Edward O. (1971) The Insect Societies, The Belknap Press of Harvard University Press, Cambridge and London.

Zeigler, David D. (1986) 'The compound eye: An introduction too the variety of visual capabilities, Goals, and approaches found in the class insecta.' Technical report, AFOSR, Summer Faculty Research Program, 1986.

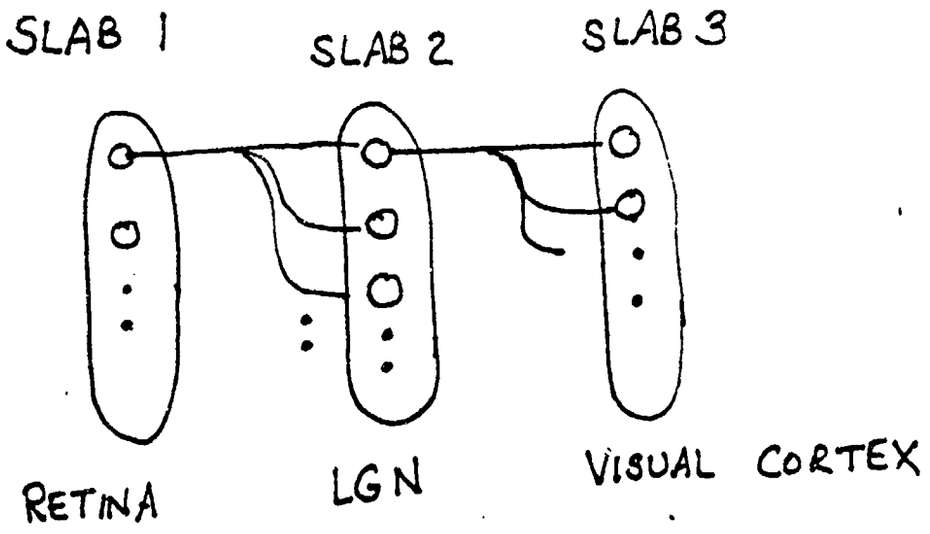
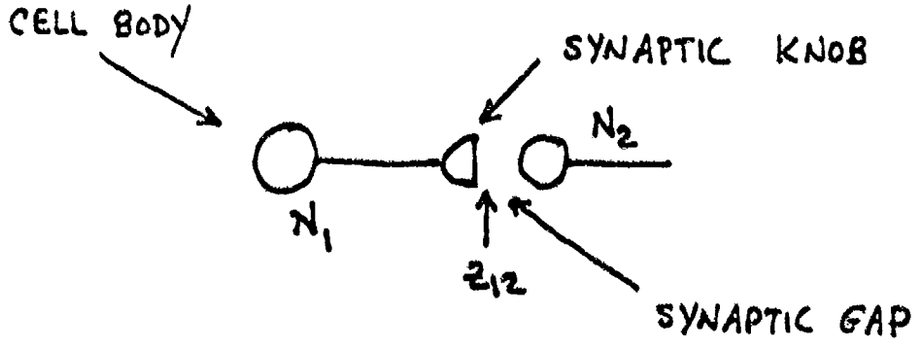


fig 1

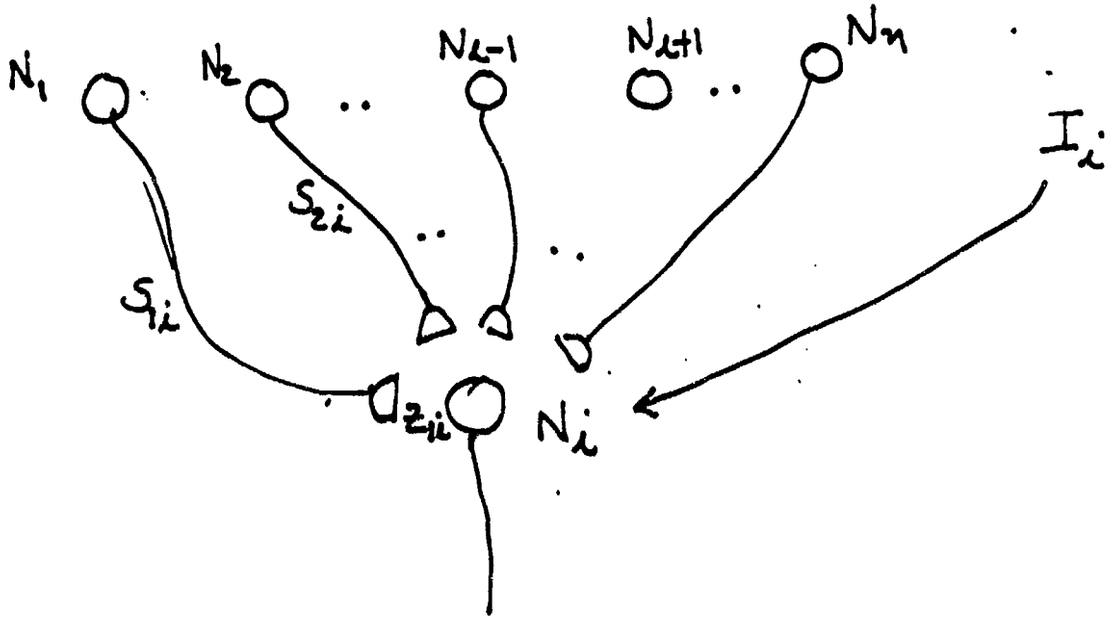


z_{12} = SYNAPTIC STRENGTH FROM NEURON N_1 TO N_2

V_i = INTERNAL VOLTAGE

fig 2

$$\frac{dx_i}{dt} = -A_i X_i + \sum_{k=1}^{n} S_{ki} z_{ki} - \sum_{k=1}^{n} C_{ki} + I_i(t)$$



N_i = NEURON i

X_i = VOLTAGE OF N_i (POTENTIAL DIFF.)

S_{ki} = FREQ of SIGNALS FROM N_k to N_i (S_{ki} CAN = 0 OR 1)

A_i = DECAY RATE OF X_i IN NEURON N_i

C_{ki} = INHIBITORY LINK FROM NEURON N_k to N_i

I_i = INPUT from OUTSIDE of NETWORK $N_1, \ldots N_n$

fig 3

$$\frac{dz_{ij}}{dt} = -B_{ij} z_{ij} + S'_{ij} [x_j]^+$$

z_{ij} = SYNAPTIC STRENGTH from NEURON i to NEURON j

B_{ij} = DECAY RATE (FORGETTING FACTOR)

S'_{ij} = FREQ OF PULSES from N_i to N_j

$$[x_j]^+ = \begin{cases} x_j - T & \text{if } x_j \geq T \quad (T = \text{THRESHOLD}) \\ 0 & \text{OTHERWISE} \end{cases}$$

fig 4

Welcome to Brain.Stetson

This documentation is designed to aid in the installation and use of Brain.Stetson. After following these few simple steps Brain.Stetson will be ready to run on your system.

Program Installation :

Please note that this program depends on the pathname of several directories. All pathnames are relative to the main directory. The current default is [udd.neural1], you will need to replace this with your main directory name. To install this software package on your system please go to all files listed below and substitute your main directory name. To do this you must edit each file and use the SUB command to change all directory pathnames. Please do not forget to re-compile all of the pascall programs and re-link them as follows below.

1) Login.com
2) [.progs.genesis] Babydriver.com
3) [.progs.genesis] Pre_genesis.pas
4) [.progs.genesis] Ganesis.pas
5) [.progs.genesis] Genesisone.pas
6) [.progs.genesis] Ganesisthree.pas
7) [.progs.build] Header.pas
8) [.progs.modelsubs] Fileops.pas
9) [.documents] Dirflow.com
10) [.documents] Dirmove.com

Below is a step by step explanation of the above using the commands to achieve a full installation of Brain.Stetson.

Meanings of Symbols used below
$ --> This is your user prompt
* --> This is the editor prompt
Accnt --> This is the name of your account, i.e. NEURAL1

Note : @LOGIN after step 1 defines all directory movement commands. i.e. GOGEN, GOBLD, etc

Commands for installing Brain.Stetson

```
1)..... $ EDIT LOGIN.COM          <RETURN>
         * SUB/NEURAL1/ACCNT/W     <RETURN>
         * EXIT                    <RETURN>
       $ @LOGIN                    <RETURN>
```

fig 1.1

```
                $ GOGEN                         <RETURN>
2).....         $ EDIT GENBYDRIVER.COM          <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>
3).....         $ EDIT PRE_GENESIS.PAS          <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>
                $ SUBPAS PRE_GENESIS            <RETURN>
4).....         $ EDIT GENESIS.PAS              <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>
                $ SUBPAS GENESIS                <RETURN>
5).....         $ EDIT GENESISONE.PAS           <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>
                $ SUBPAS GENESISONE             <RETURN>
6).....         $ EDIT GENESISTHREE.PAS         <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>
                $ SUBPAS GENESISTHREE           <RETURN>
                $ GOBLD                         <RETURN>
7).....         $ EDIT HEADER                   <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>
                $ SUBPAS HEADER                 <RETURN>
                $ GOSUBS                        <RETURN>
8).....         $ EDIT FILEOPS.PAS              <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>
                $ SUBPAS FILEOPS                <RETURN>
                $ GODOC                         <RETURN>
9).....         $ EDIT DIRFLOW.COM              <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>

10).....        $ EDIT DIRMOVE.COM              <RETURN>
                  * SUB/NEURAL1/ACCNT/W         <RETURN>
                  * EXIT                        <RETURN>

                        Program linking stage
                Note --> This stage may only begin at completion of
                         all above batch jobs.

                $ GOGEN                         <RETURN>
                $ PRELINK                       <RETURN>
                $ GOBLD                         <RETURN>
                $ HLINK                         <RETURN>


                Brain.Stetson is now installed on your computer.
```

FIG 1.1 (CONT.)

his file explains the use of the BRAIN.TXT file. Below is the header created
nen the 'I. build a header file' option is chosen.

```
Text file for brain model "header".

labA  Ai  Aj  Ax  Ay  Synum Length SlabB  Bi  Bj  Bx  By  Thresh  FF     SS    Decay  Excite
```

Each row in the table describes the conections for one synapse of a neuron.
escribing a neuron requires describing the conections for each of its synapses.
ius, describing one neuron may require the use of several rows.

Following are definitions for the columns in the model description table.

```
labA   -      The slab that the firing neuron is on.
:: Aj  -      The dimensions of SlabA.
:: Ay  -      The location of the firing neuron on SlabA.
/num   -      The number of the synapse of the firing neuron. (This number is just used for identificat:
ength  -      The the length of the axon between the firing neuron and the synapse.
              (An axon of length n will have its synapses receive its signals  n time steps after the c
abB    -      The slab that the receiving firing neuron is on.
.: Bj  -      The dimensions of SlabB.
:: By  -      The location of the receiving neuron on SlabA.
:resh  -      The threshold voltage of the firing neuron.
       -      Forgeting factor of the synapse.
.      -      Synaptic strength of the synapse.
·cay   -      The decay rate of the firing neuron.
 cite  -      1 if the synapse is excitatory, 0 if inhibitory.
```


fig 1.2

SLAB 1 → CONNECTED TO SLAB 2

Text file for brain model "eve".

SlabA	Ai	Aj	Ax	Ay	Synum	Length	SlabB	Bi	Bj	Bx	By	Thresh	FF	SS	Decay	Excite
1	1	1	1	1	1	1	2	1	1	1	1	2.00	0.20	1.10	0.50	1
2	1	1	1	1	1	1	2	1	1	1	1	3.00	0.20	1.00	0.50	1

SLAB 2 → CONNECTED TO SLAB 2

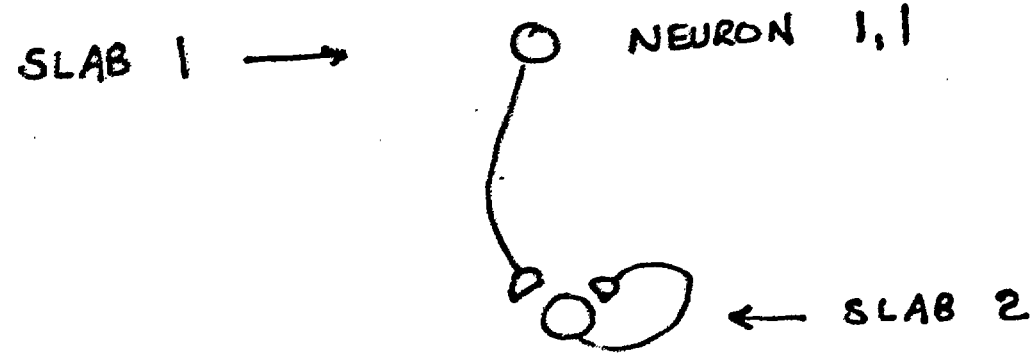


fig 1.3

Text file for brain model "eve".

NEURON 1,1 CONNECTS TO { SLAB 2 NEURONS 1,1 & 1,2

SlabA	Ai	Aj	Ax	Ay	Synum	Length	SlabB	Bi	Bj	Bx	By	Thresh	FF	SS	Decay	Excite
1	1	2	1	1	1	1	2	1	2	1	1	3.00	0.20	1.00	0.50	1
1	1	2	1	1	2	1	2	1	2	1	2	3.00	0.20	1.00	0.50	1
1	1	2	1	2	1	1	2	1	2	1	1	3.00	0.20	1.00	0.50	1
1	1	2	1	2	2	1	2	1	2	1	2	3.00	0.20	1.00	0.50	1

NEURON 1,1 HAS 2 SYNAPSES

LAB 1 IS 1X2

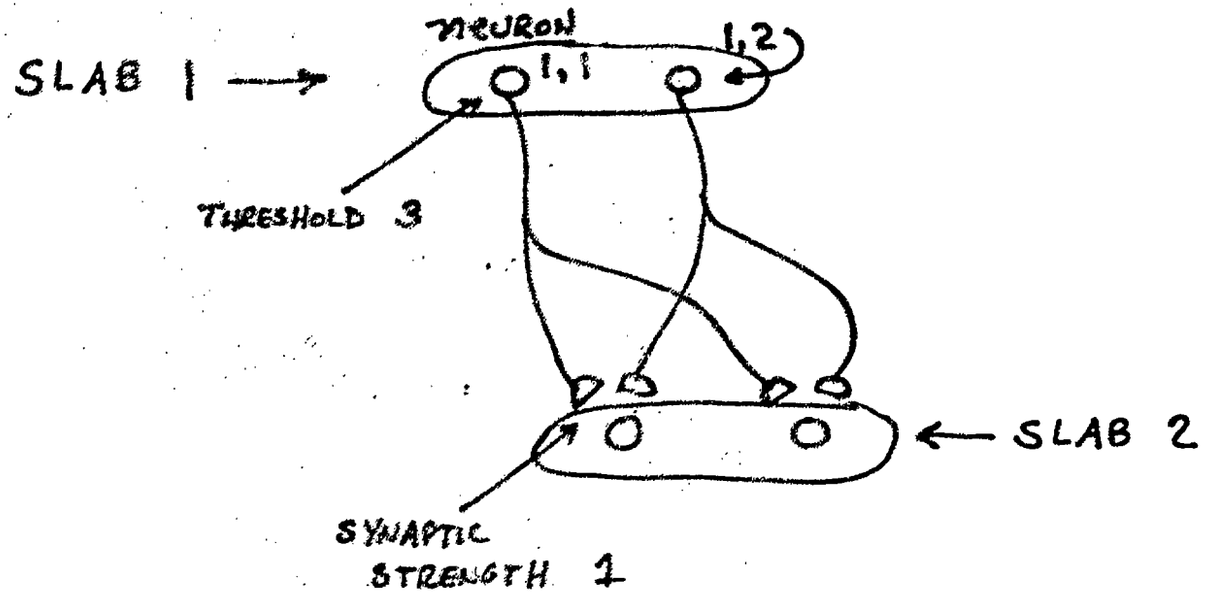


FIG 1.4

```
BRAIN MODELING"

Available Options:"

        1. Build a data file header"
        2. Create a new Model (from a data file)"
        3. Run a pre-existing Model"
        4. Quit"

        Enter Option : "
```

fig 1.5

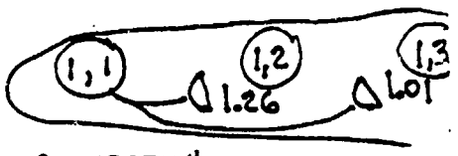
```
Slab #    3 Coordinates : (    1,    1)
Neural Charge    Threshold Charge    Decay Rate
   3.227976          3.800000        0.000000
Synapse Number        SS          FF          Coordinates
     1              0.933954   0.030000        3    1    1
     2              1.261847   0.030000        3    1    2
     3              1.014282   0.030000        3    1    3
     4              0.000000   0.030000        3    1    4
     5              0.000000   0.030000        3    1    5
     6              0.297784   0.030000        3    1    6
     7              1.000000   0.000000        4    1    1
     8              0.450000   0.000000        1    1    1
     9              0.000000   0.000000        0    0    0
History = 0 0 0 0 0 0 1 1 1 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Slab #    3 Coordinates : (    1,    2)
Neural Charge    Threshold Charge    Decay Rate
   3.226415          3.800000        0.000000
Synapse Number        SS          FF          Coordinates
     1              0.603091   0.030000        3    1    1
     2              1.014282   0.030000        3    1    2
     3              1.300973   0.030000        3    1    3
     4              0.000000   0.030000        3    1    4
     5              0.000000   0.030000        3    1    5
     6              0.000000   0.030000        3    1    6
     7              1.000000   0.000000        4    1    2
     8              0.450000   0.000000        1    1    2
     9              0.000000   0.000000        0    0    0
History = 0 0 0 0 0 0 0 1 1 1 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Slab #    3 Coordinates : (    1,    3)
Neural Charge    Threshold Charge    Decay Rate
   0.000000          3.800000        0.000000
Synapse Number        SS          FF          Coordinates
     1              0.336369   0.030000        3    1    1
     2              0.621743   0.030000        3    1    2
     3              2.247218   0.030000        3    1    3
     4              0.000000   0.030000        3    1    4
     5              0.000000   0.030000        3    1    5
     6              0.000000   0.030000        3    1    6
     7              1.000000   0.000000        4    1    3
     8              0.450000   0.000000        1    1    3
     9              0.000000   0.000000        0    0    0
History = 0 0 0 0 0 0 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1
Slab #    3 Coordinates : (    1,    4)
Neural Charge    Threshold Charge    Decay Rate
   1.924474          3.800000        0.000000
Synapse Number        SS          FF          Coordinates
     1              0.288350   0.030000        3    1    1
     2              0.000000   0.030000        3    1    2
     3              0.000000   0.030000        3    1    3
     4              0.497608   0.030000        3    1    4
     5              0.686937   0.030000        3    1    5
     6              0.512998   0.030000        3    1    6
     7              1.000000   0.000000        4    1    4
     8              0.450000   0.000000        1    1    4
     9              0.000000   0.000000        0    0    0
History = 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Slab #    3 Coordinates : (    1,    5)
Neural Charge    Threshold Charge    Decay Rate
   1.924474          3.800000        0.000000
Synapse Number        SS          FF          Coordinates
     1              0.528864   0.030000        3    1    1
     2              0.297784   0.030000        3    1    2
     3              0.000000   0.030000        3    1    3
     4              0.280185   0.030000        3    1    4
     5              0.512598   0.030000        3    1    5
     6              0.708182   0.030000        3    1    6
     7              1.000000   0.000000        4    1    5
     8              0.450000   0.000000        1    1    5
     9              0.000000   0.000000        0    0    0
History = 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Slab #    3 Coordinates : (    1,    6)
Neural Charge    Threshold Charge    Decay Rate
   2.389045          3.800000        0.000000
Synapse Number        SS          FF          Coordinates
     1              0.730005   0.030000        3    1    1
     2              0.545221   0.030000        3    1    2
     3              0.306994   0.030000        3    1    3
     4              0.000000   0.030000        3    1    4
     5              0.268350   0.030000        3    1    5
     6              0.528864   0.030000        3    1    6
     7              1.000000   0.000000        4    1    6
     8              0.450000   0.000000        1    1    6
```

SLAB 3

NEURON (1,1)

SYNAPTIC STRENGTH TO SLAB 3 NEURON (1,2) = 1.2618

NEURON (1,2)



CONNECTION STRENGTHS

NOTE: ONE CAN EXAMINE THIS DISPLAY AND BUILD CONNECTION STRENGTHS

fig 1.6

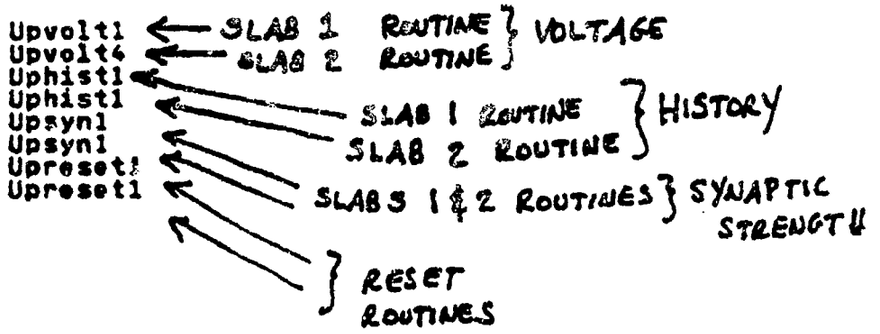


FIG 1.7

```
        BRAIN MODELING"

Available Options:"

        1. Build a data file header"
        2. Create a new Model (from a data file)"
        3. Run a pre-existing Model"
        4. Quit"

        Enter Option : "
```

FIG 1.9

```
{*****************************************************************************}
{* Procedure Voltage1 :  Normal Voltage update based on the internal      =)
{*                       voltage formulas of Grossberg.                   =)
{*****************************************************************************}
{*                                                                          =)
{* ALGORITHM :  1) Create a loop to examine every neuron in the current slab =)
{*              2) Locate the neuron, PostSN - this is the current neuron to =)
{*                 change.                                                  =)
{*              3) calculate the amount of decay in the current voltage of  =)
{*                 PostSN, Delta_V - the change in Voltage.                 =)
{*              4) Check to see if there is any incoming synapses for Postsn.=)
{*              5) If there was any incoming synapses then bring in each one at=)
{*                 a time (PreSN), Check and see if it fired at time t-1, if =)
{*                 so then add the synaptic strength of the PreSN neuron that=)
{*                 connects to the PostSN neuron.  -- note this is only a   =)
{*                 summation of all the incoming synapses to PostSN.        =)
{*              6) Add SS to Delta_V, then add Delta_V to the current Voltage=)
{*                 and loop back to #2 until there are no more Neurons in the=)
{*                 Current slab.                                            =)
{*                                                                          =)
{*  FORMULAS USED :                                                         =)
{*      1) Delta_V = V * Decay * -1   : calculate the amount of decay of the=)
{*         internal voltage, which is a percentage.                         =)
{*      2) Delta_V = Delta_V + SS     : Add the synaptic strengths of all    =)
{*         incoming synapses to the change in volatage. All Excicitory.     =)
{*      3) Delta_V = Delta_V - SS     : Subtract the synaptic strengths of all=)
{*         incoming synapses to the change in volatage. All Inhibitory.     =)
{*      4) V = V + Delta_V + Input                                          =)
{*                                                                          =)
{*****************************************************************************}

[Global] Procedure Voltage1 (Cycle_number,Cur_Cycle:Integer; Slabno:Integer; Var Brnfile:Neurfile; Var Fil

Var
        Loop1, Loop2  : Integer;
        Spot          : Integer;
        Position      : Integer;
        N_Length      : Integer;
        Volt, Delta_V : Real;
        PreSN, PostSN : Neurons;
        Runner        : Insyn_apse;
        Done          : Boolean;

  Begin
  {*****************************************************}
  {** Calculate the Voltage For Each Neuron           **}
  {*****************************************************}
  For Loop1 := 1 to I[SlabNo] do      {** X-corrdinate **}
    For Loop2 := 1 to J[SlabNo] do    {** Y-coordinate **}
      Begin
  {*****************************************************}
  {** Find - locates the proper neuron in the brain   **}
  {** file and reads in the information (PostSN)      **}
  {*****************************************************}
      Postin := Get_Brain(SlabNo,Loop1,Loop2);
      Volt := (PostSN.V);
      Delta_V := (PostSN.V)*(PostSN.Decay)*(-1);
      If PostSN.Insynapse <> 0  {**  If there is any incoming synapses **}
        then begin
```

fig 1.8

```
        Find (Filname2,PostSN.insynapse);    {** Find wh  PostSN connects to **}
        Runner := Filname2^;

        Done := False;
        Repeat
          N_Length := Runner.Length;
{*************************************************}
{** Find - locates the proper neuron in the b ain  **}
{** file and reads in the information (PreSN)      **}
{*************************************************}
          PreSn := Get_Brain(Runner.SlabNo,Runner.Xcoord,Runner.Ycoord);
          Spot := Runne .Synum;
          If Cycle_Number - N_Length <= 0
            Then Position := 50 - ((Cycle_Number - N_Length)*(-1))
            Else Position := Cycle_Numb r - N_Length;
          If Cur_Cycle - N_Length > 0   {** Actual time cycles completed **}
{*************************************************}
{** If both neurons involved (PostSN & PreSN) fired at time **}
{** t then add (or subtract) the synaptic strength from     **}
{** PreSN neur n t  PostSn Delta_V.                         **}
{*************************************************}
```

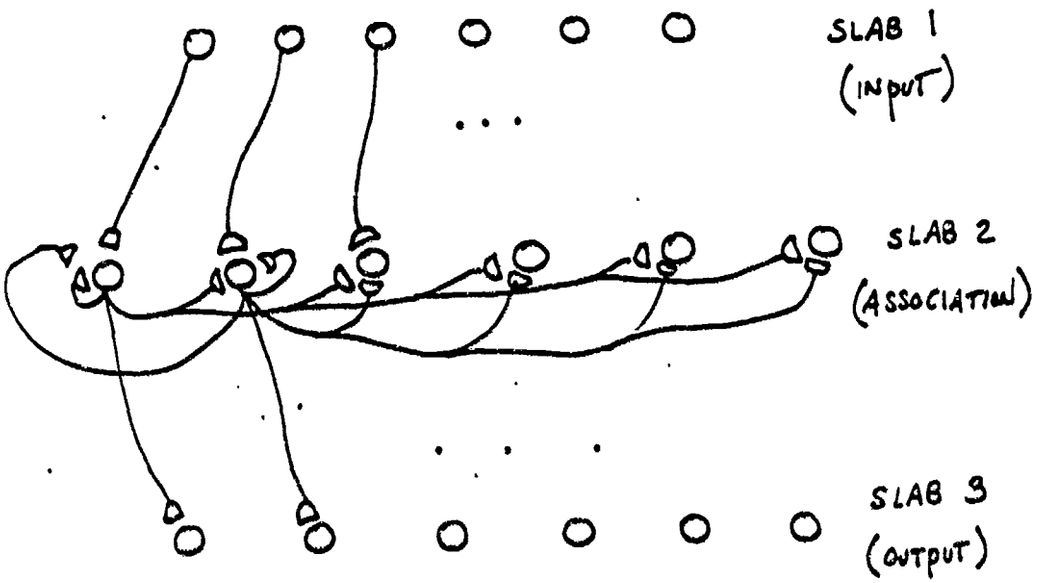
45-58

```
            Then If PreSN.History[Position] > 0    {** Is it firing **}
              Then If PreSN.Axon[Spot].Excite
                Then If PreSN.Axon[Spot].SS   {** Add The Synaptic Strength to the v
                  Then Delta_V := Delta_V + PreSN.Axon[Spot].SS; {** Subtract The Synaptic Strength fro
                  Else Delta_V := Delta_V - PreSN.Axon[Spot].SS;
              If Runner.Next = 0
                Then Done := True
                Else Begin

              Find (Filname2,Runner.next);
              Runner := Filname2^;

            End;
          Until Done;
        End;
{*************************************************}
{** Change the current voltage by Delta_V.      **}
{*************************************************}
        Volt := Volt + Delta_V;
        Volt := Volt + PostSn.S_Input;
        PostSn.V := Volt;
        Put_Brain(PostSn,SlabNo,Loop1,Loop2);
      End;
End;
```

fig 1.8 (cont.)



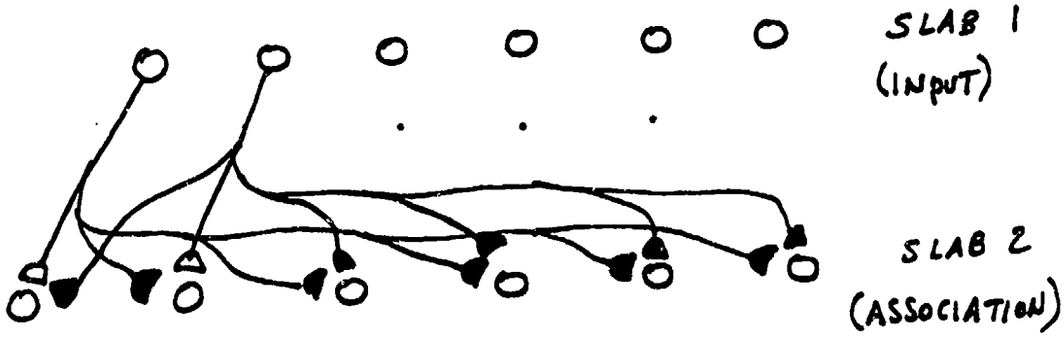
CONNECTIONS SLAB 1 NEURON k → NEURON k SLAB 2

SLAB 2 NEURON k → ALL NEURONS SLAB 2

→ AND TO NEURON k SLAB 3

SLA

fig 2.1



DARKENED SYNAPTIC KNOBS ARE INHIBITORY

THESE ARE THE ADDITIONAL CONNECTIONS

MADE TO fig 2.1

ADDITIONAL CONNECTIONS

SLAB 1 NEURON k → INHIBITORY NEURON 1,.. k-1, k+1, n SLAB 2

fig 2.2

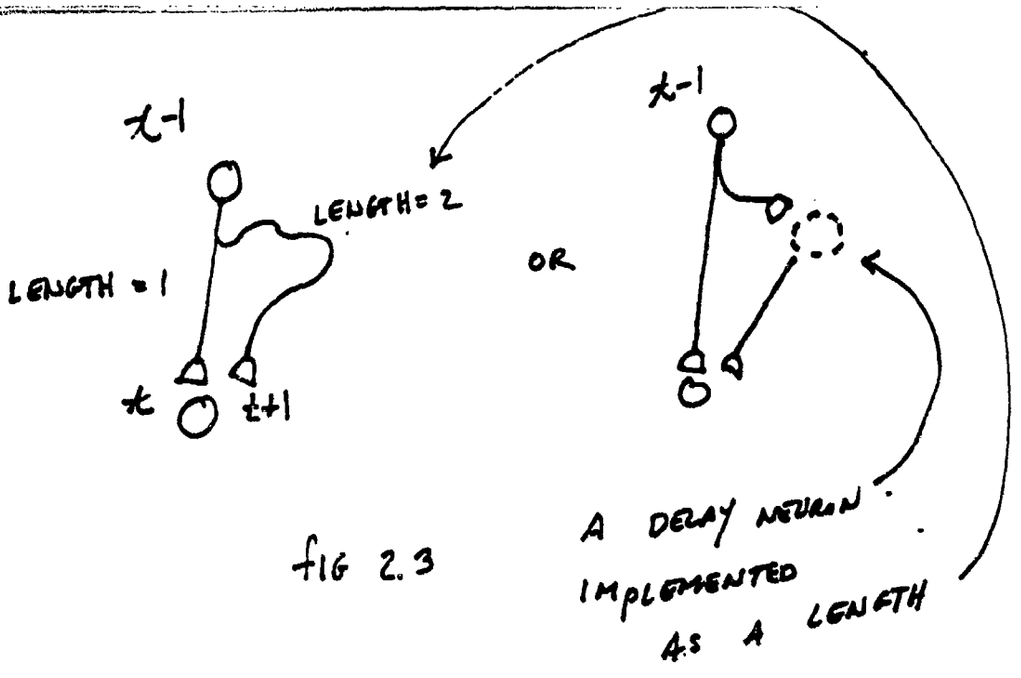
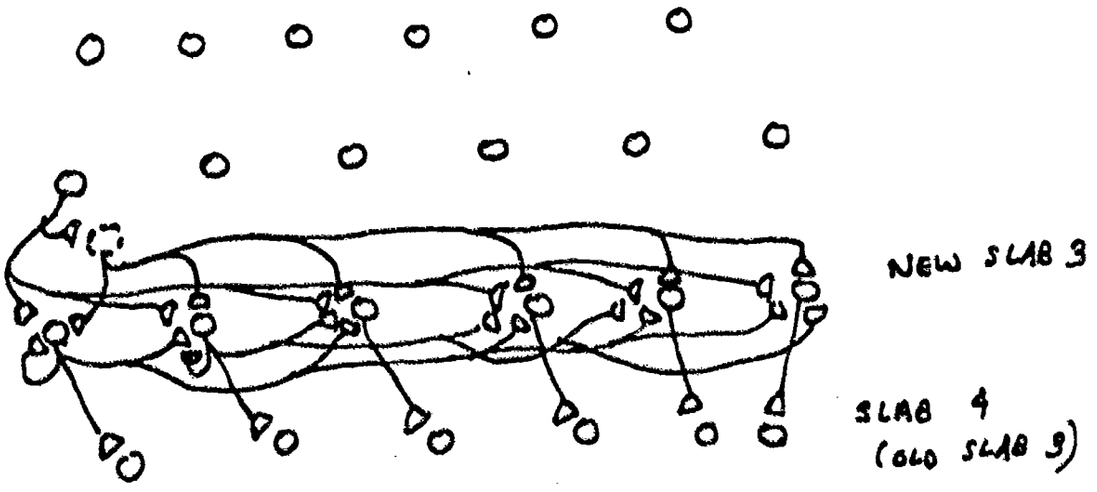


fig 2.3

A DELAY NEURON IMPLEMENTED AS A LENGTH



THE NEW SLAB & NEW CONNECTIONS ADDED TO fig 2.2

fig 2.4

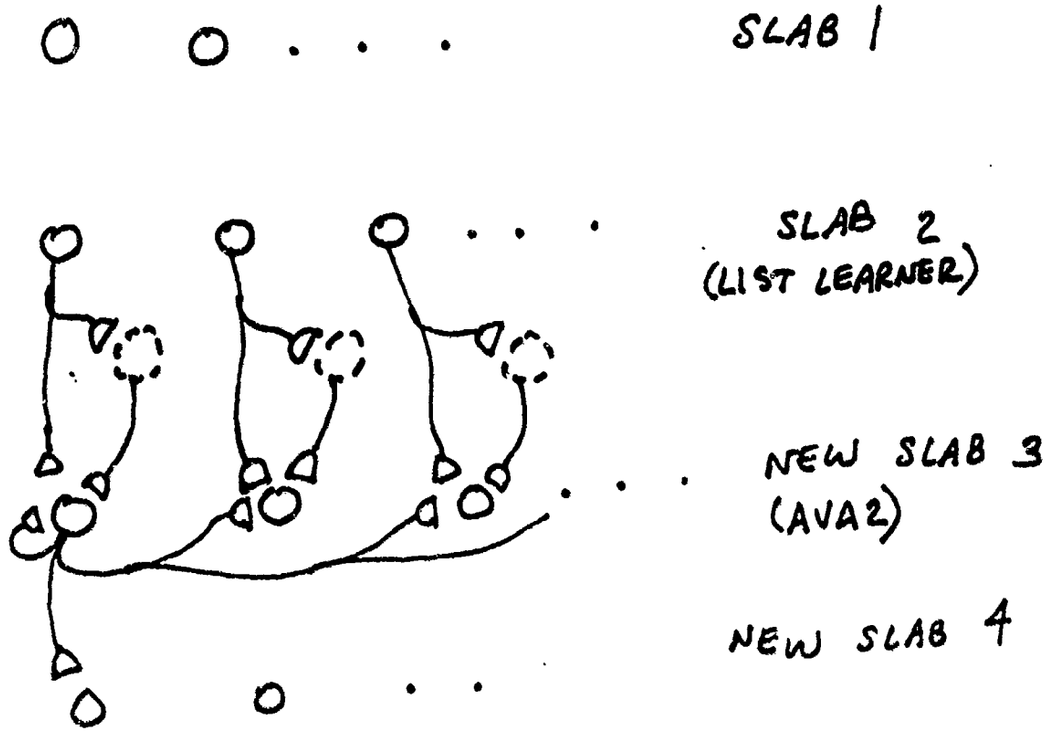
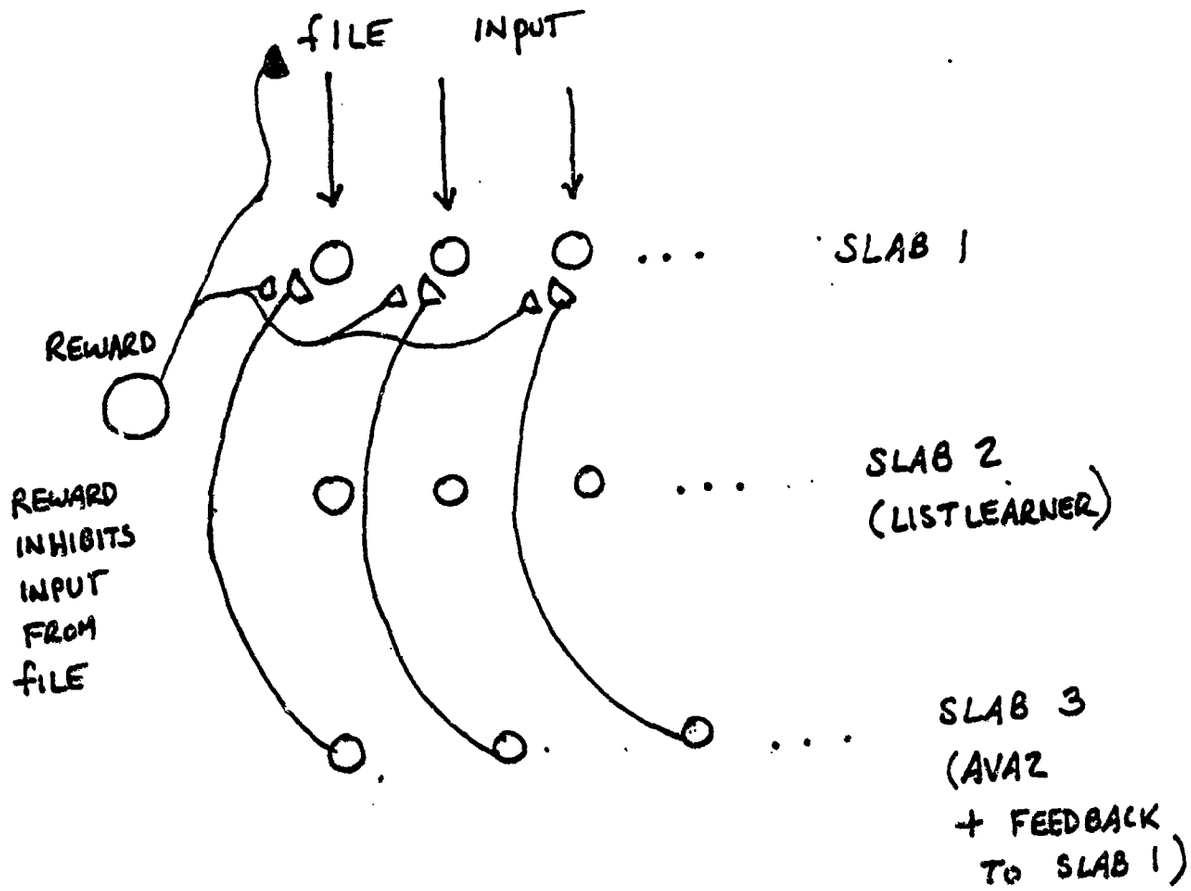


FIG 2.4 IS WRONG ARCHITETURE.

THIS IS THE ADDITIONS TO fig 2.2

AS THEY SHOULD BE MADE

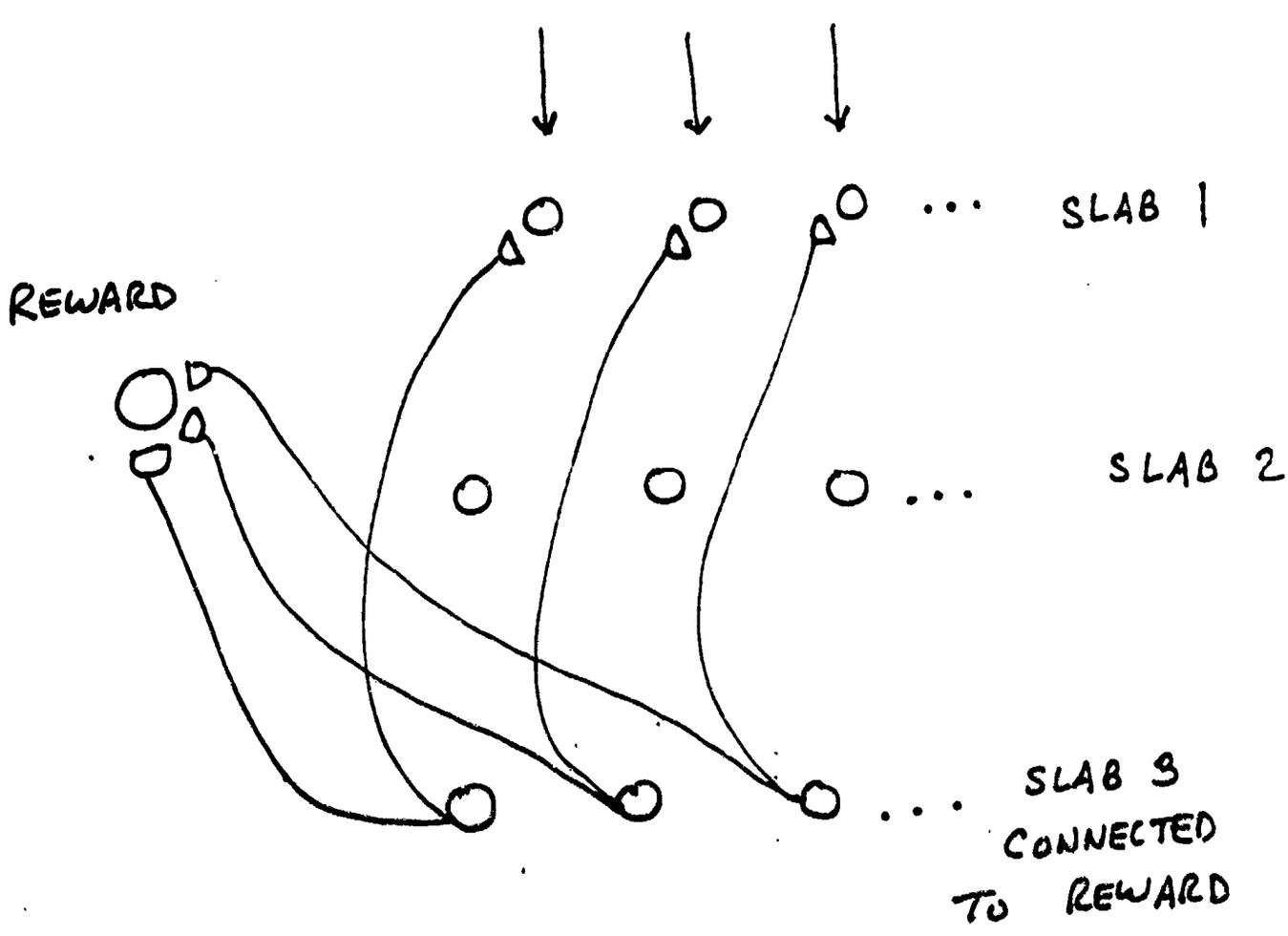
fig 2.5



THE ADDITIONAL CONNECTIONS

TO fig 2.5

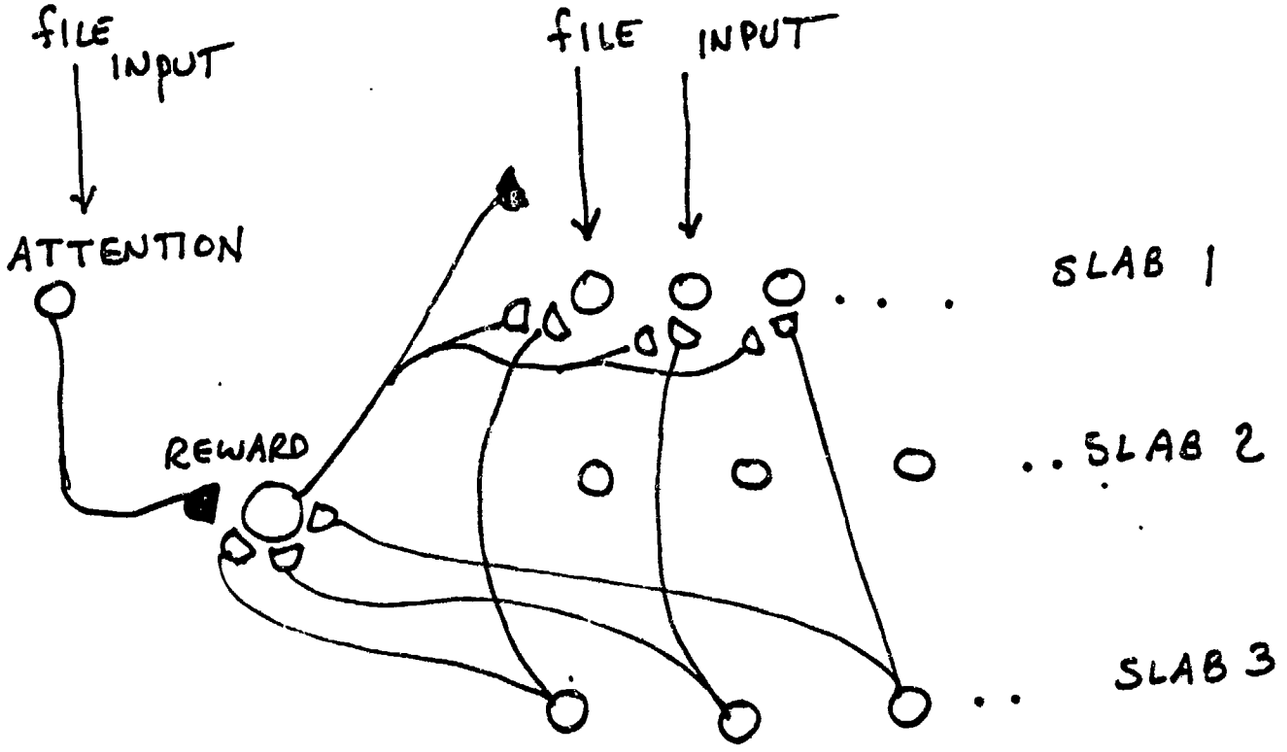
fig 2.6



AXONS FROM SLAB 3 TO REWARD

ADDED

fig 2.7



REWARD
INHIBITS FILE INPUT & ACTIVATES INPUT FROM SLAB 3

ATTENTION IS ADDED TO fig 2.7
AS ARE AXONS FROM SLAB 3 TO REWARD

fig 2.8

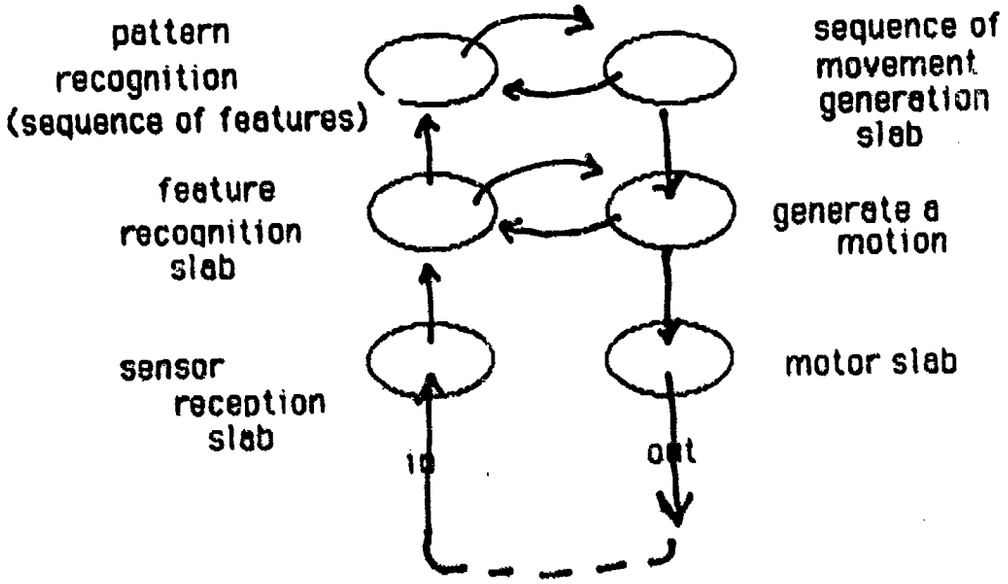
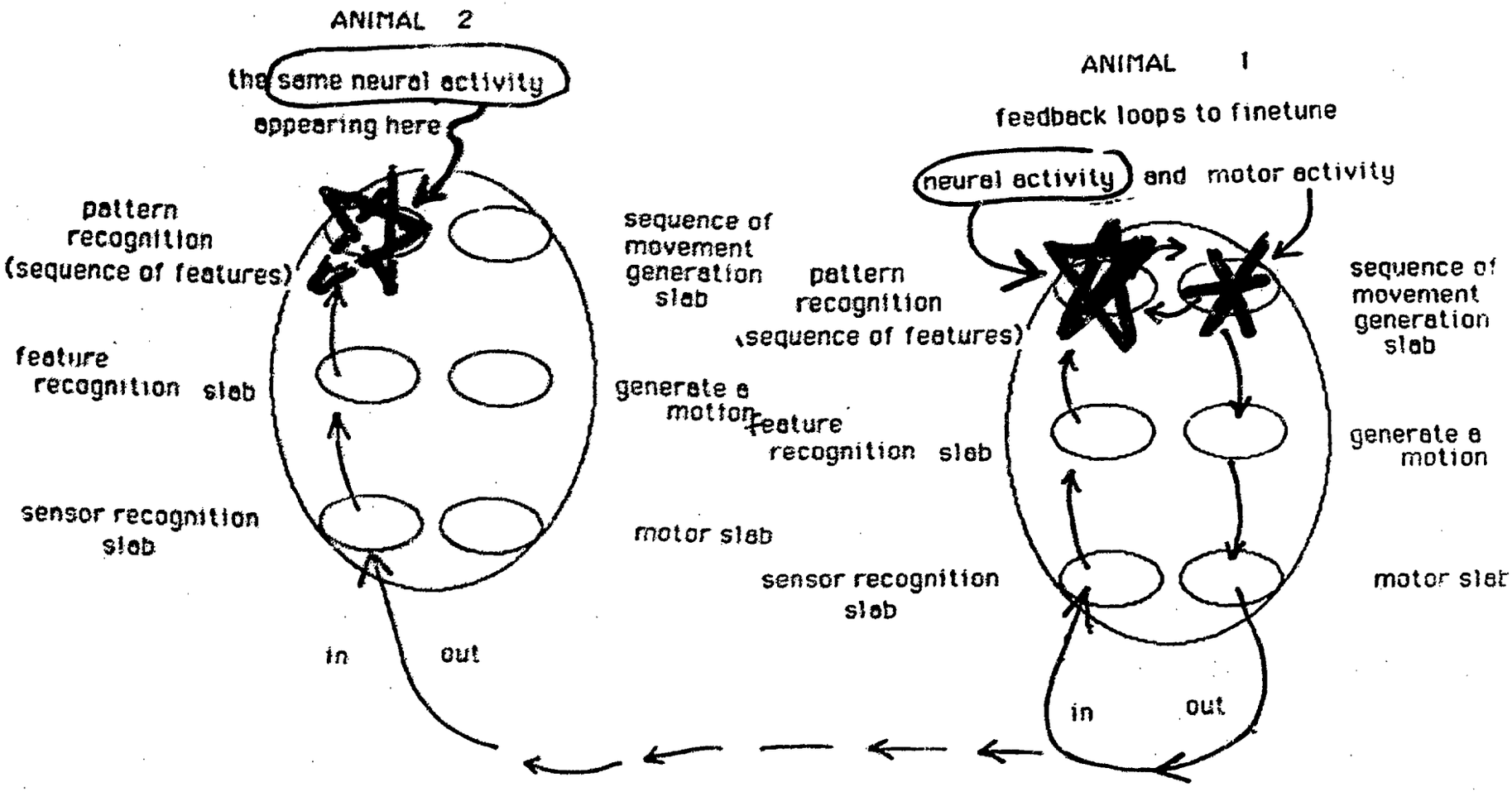

pattern
recognition
(sequence of features)
feature
recognition
slab
sensor
reception
slab
in
out
sequence of
movement
generation
slab
generate a
motion
motor slab

figure 3.1



Each animal has a feedback loop to finetune neural activity. If animal 2's loop captures signals from animal 1, then the neural activity in animal 1 will appear in animal 2.

and VOILA COMMUNICATION ! !

figure 32

DATA PROCESSING AND STATISTICAL ANALYSIS OF IN-SERVICE AIRCRAFT TRANSPARENCY FAILURES*

Paul S. T. Lee, Ph.D.
Associate Professor

SCHOOL OF BUSINESS AND ECONOMICS
NORTH CAROLINA A & T STATE UNIVERSITY

*This project is partially supported by a Mini-Grant from Universal Energy Systems, Inc., and The System Command of the U. S. Air Force. (Grant No.: S-760-6MG-023)

February 1988

ACKNOWLEDGEMENT

This project stems from a follow-up research conducted by the Principle Investigator at the Flight Dynamics Laboratory of the U. S. Air Force in the Summer of 1986. The author acknowledges the financial support of the Universal Energy System, Inc. and the Office of the Scientific Research of the U. S. Air Force in the form of a Mini-Grant Award (Grant No. S-760-6MG-023).

Special thanks go to Mr. Robert McCarty and Mr. Mike Gran, Engineers at the U. S. Air Force, for their technical support throughout the entire duration of the study. I am also indebted to Dr. Japhet Nkonge, Professor of Marketing at the School of Business and Economics for his valuable suggestions on the manuscript preparation.

LIST OF TABLES AND CHARTS

Table 1	Type of Transparency by Vendor
Table 2	Failure Modes by Vendor
Table 3	Failure Modes by Vendor
Table 4	Transparency Failure Modes
Table 5	Vendors of Failed Transparencies by AF Base
Figure 1	Histogram of Transparency Failures (F16)
Figure 2	Histogram of Transparency Failures (F111)

Data Processing and Statistical Analysis of In-Service Aircraft Transparency Failures

I. Introduction

Aircraft transparent enclosures, including windshields and canopies, are high cost items to the U. S. Air Force. In an operational environment, these enclosures are subject to many environmental exposures such as changes in temperature, heavy rainfall, strong pressure, excessive sunlight and abrasion in flight and during cleaning. As a result, many transparent enclosures failed gradually because of the aging process; and others failed abruptly and unexpectedly. Transparency failures are not only costly, but also often impair combat readiness and can cause the loss of aircraft and lives.

The Vehicle Equipment Division of the Air Force Flight Dynamics Laboratory has been seeking the development of the best quality and the most durable aircraft transparencies. In the past several years, it has conducted a major task to collect detailed field in-service data on aircraft transparencies, focusing primarily on those ones taken from F-16 and F-111 aircrafts because of failures of one kind or another.

This report presents the results of the analysis of some of the above data. It is hoped that information generated from these results will be extremely valuable in assessing the quality of structural design of transparencies, the frequency of occurrence of each failure mode, and to determine the possible relationship between these two variables.

II. Objectives

The major objectives of the study are:

(1) To organize the available information obtained from records of the displaced aircraft transparencies into an efficient data base for further study use.

(2) To conduct a statistical analysis using the data base just created relative to transparency failure modes, vendors who supplied the transparencies to the U. S. Air Force and Air Force Bases (AFB) where the aircraft were stationed.

III. Data Source and Data-Base Creation

During the past several years, the Vehicle Equipment Division of the Flight Dynamics Laboratory of the U. S. Air Force has been collecting data at a number of Air Force Bases on aircraft transparencies replaced due to a variety of failures. Engineers at the Vehicle Equipment Division were responsible for the data collection efforts. It is assumed that these data were selected randomly, and no prior discretionary judgement has been injected in the data selection process.

After their collection, these data were sent to the School of Business and Economics at North Carolina A & T State University. They were then evaluated and input into the appropriate microcomputer-based data bases using dBASE III, a data base management computer software.

By the end of August 1987, two data bases, one pertaining to transparencies taken from the F-16 jet fighters and the other to transparencies taken from the F-111 jet fighters, have been created. The F-16 data base consists of 953 records, and the F-111 data consists of 678 records of the displaced aircraft transparencies. After their creation, copies of these data bases were then sent to the Air Force for evaluation to insure their

authenticity and accuracy. Subsequently, errors were corrected.

Table 1 shows a brief profile of the data-base with transparencies from F-16 jet fighters, and Table 2 gives a brief account of the data-base made of transparencies from F-111 aircrafts. As indicated in Table 1, nearly half (45.5%) of the transparencies taken from the F-16 were Texstar's products; about one-third (29.4%) came from Goodyear and the rest from Sieracin. The names of vendors on 48 pieces of transparencies (5.1%) were either missing or not recognizable.

Table 1

TYPE OF TRANSPARENCY BY VENDOR

(F-16)

VENDOR	TOTAL No.	TOTAL Percent	TYPE OF TRANSPARENCY FWD No.	FWD Percent	AFT No.	AFT Percent	N. A. No.	N. A. Percent
Goodyear	280	29.4%	121	25.5%	73	54.0%	86	25.0%
Sieracin	192	20.1	74	15.6	21	15.6	97	28.2
Texstar	433	45.4	237	50.0	39	28.9	157	45.6
N. A.*	48	5.1	42	8.9	2	1.5	4	1.2
Total	953	100%	474	100%	135	100%	344	100%

*N. A. stands for information is "not available" or "not identifiable."

Source: Air Force D-Base, 1987
North Carolina A & T State University

Table 2

FAILURE MODES BY VENDOR

(F-111)

FAILURE MODES	VENDOR				TOTAL	
	Sieracin		PPG		No.	Percent
	No.	Percent	No.	Percent		
Acrylic crazing	13	4.8%	58	19.8%	71	12.6%
Acrylic cracks	35	13.0	29	9.9	64	11.4
Delaminated/ distorted	86	31.9	73	24.9	159	28.2
Scratched	67	24.8	47	16.0	114	20.2
Chipped/pitted	65	24.1	60	20.5	125	22.2
Coating loss	0	0	22	7.5	22	3.9
Others*	4	1.4	4	1.4	8	1.4
Subtotal	270	100.0%	293	100.0%	563	100.0%
N.A.**	55		60		115	
Total	325		353		678	

*Others includes combinations of two or more failure modes.

**N.A. stands for data is not available or missing.

Source: Air Force D-Base, 1987
North Carolina A & T State University

IV. Failure Mode Analysis

Most of the aircraft transparencies, including windshields and canopies, were made of two to three plies of heavy plastics commonly known as acrylics ($C_3H_4O_2$) or polycarbonate materials.

There are many forms of transparency failures. In general, they fall into two categories: Those visible to human eyes and those detectable only by mechanical instruments. Of those visible by human eyes are surface scratches, cracks, crazes, coating loss, poly-delamination or distortion and chips or damages. Those detectable only by mechanical instruments include structural defects, residual stress, cracks or chips at the joints connecting the transparency with the aircrafts.

Transparency failures may result from an aging process or be caused by environmental factors such as wind, rainfall, sunlight (ultra-violet radiation), high speed, and extreme temperature the aircraft would encounter both on the ground or while in-service. Still other failures might have been caused by hail impact, bird strikes, poor maintenance practices and/or low quality of chemicals used in cleaning.

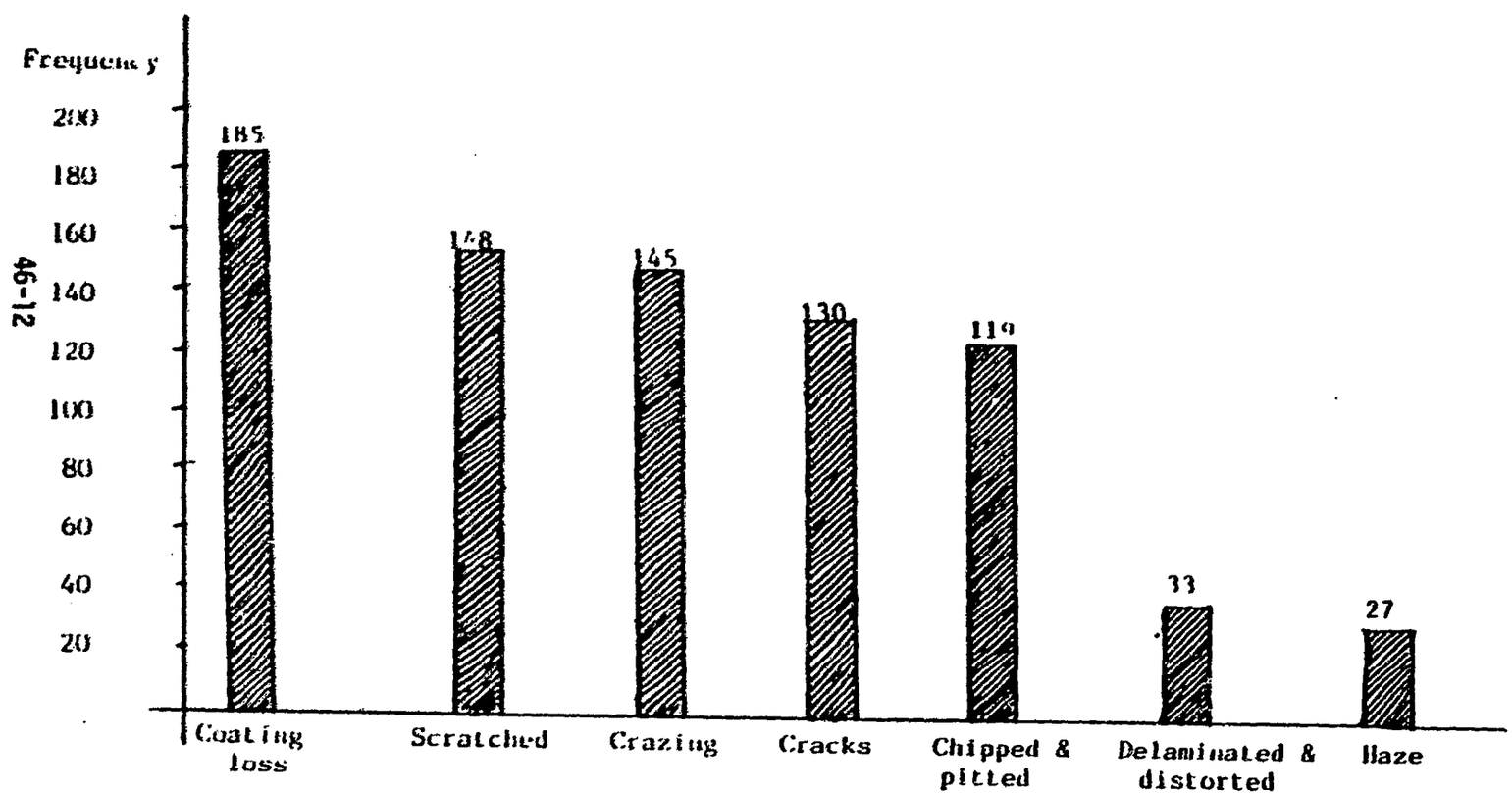
(1). Pareto Analysis

The amount of costs or damages caused by transparency failures varies with different forms of failure, ranging from about $20,000 per windshield replacement to loss of aircraft and/or lives. However, minimizing the frequency or total number of failures is often consistent with reducing the total amount of cost outlays.

In studying business practices, Vilfredo Pareto, a prominent Italian economist, once stated: "In general, 80 percent of the problems could be resolved by using 20 percent of the effort." This 80-20 Pareto principle begins with the examination of the frequency distribution and the histogram of various failure modes. Figure 1 shows the histogram of the transparency failures modes of F-16s, not counting those without failure mode information. It can be seen that coating loss is the predominant failure mode, accounting for nearly one-fourth (23.4%) of transparency failures. Abrasion or scratching is the second most frequent failure, accounting for 18.8% of the total transparency failures. Both coating loss and scratches are mostly caused by poor maintenance practices. Hence, installation of proper maintenance practices may prove to be very cost-effective in reducing total transparency-related expenses.

Figure 1

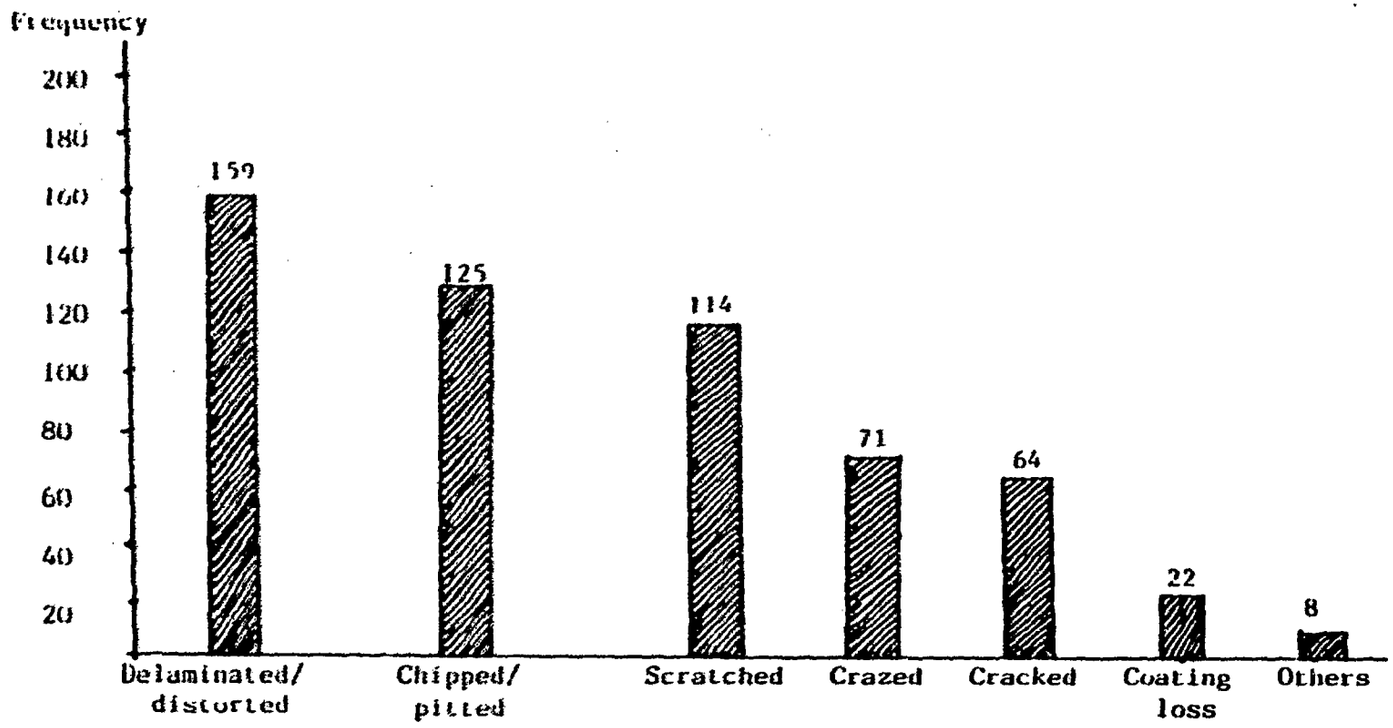
HISTOGRAM OF TRANSPARENCY FAILURES
(F16)



Source: Air Force D-Base, 1987
North Carolina A&T State University

Figure 2

HISTOGRAM OF TRANSPARENCY FAILURES
(F111)



Source: Air Force D-Base, 1987
North Carolina A&T State University

Crazing is the third most frequent transparency failure mode. It is found on nearly one out of every five displaced transparencies. According to recent literature, crazing is defined as fine cracks at or under the surface of a plastic. Transparency crazing may come either from internal structural design or external environmental factors. Studies on the behavior of crazing on acrylics are still in the evolutionary stage, and hence, are beyond the scope of this study.

Cracks rank the fourth in F-16 transparency failures. They are usually found along the edge of the transparency joining with the aircrafts. Hence, their occurrences might have been caused either by poor structural design, by poor quality control at the time of their installation, or by other environmental and human factors, or by a combination of several factors.

Chipping and/or pitting rank the fifth most frequent transparency failure, accounting for about one out of every seven displaced transparencies. Like cracks, chips and pits are mainly caused by hail impact, bird strikes, and poor maintenance practices.

Poly delamination and haze rank the sixth and seventh failure modes respectively. Again, their occurrences could be caused by numerous factors.

(2) Failure Modes by Transparent Vendors

Questions are often being raised as to which vendor produces the best quality transparencies. Which vendor has the most advanced manufacturing technology? And what environmental and human factors affect most strongly the durability of aircraft transparencies? Finding answers to these questions is not an easy task. However, statistical analysis techniques may be used to shed some light toward answering these questions.

Table 3 shows the comparison of failure modes by vendors. It can be seen from the table that variations in failure modes do exist among vendors. For example, coating loss remains to be the most important failure mode for Goodyear. Over one-third (34.6%) of Goodyear's transparency failures were attributed to coating loss. Coating loss to Sieracin, however, is a minor problem; only about one out of every 13 (7.8%) of Sieracin's transparency failures was caused by coating loss. To Texstar, coating loss is a major, but not a predominant, failure mode. About one out of every five failures (20.5%) was due to coating loss.

Table 3

FAILURE MODES BY VENDOR
(F16)

FAILURE MODES	VENDOR						TOTAL	
	Goodyear		Sieracin		Texstar		No.	Percent
	No.	Percent	No.	Percent	No.	Percent		
Acrylic crazing	17	6.7%	42	23.7%	86	21.2%	145	17.4%
Acrylic cracks	23	9.1	57	32.2	43	10.6	123	14.7
Polycarbonate cracks	2	0.1	2	1.1	3	0.7	7	0.8
Delaminated	10	3.9	5	2.8	6	1.5	21	2.5
Distorted	5	2.0	4	2.3	3	0.7	12	1.4
Scratched/ scuffed	37	14.6	19	10.7	29	7.2	85	10.2
Haze	1	0.0	1	0.6	19	4.7	21	2.5
Coating loss	88	34.6	14	7.9	83	20.5	185	22.1
Chipped	30	11.8	19	10.7	43	10.6	92	11.0
Pitted/hail impact	7	2.8	3	1.7	17	4.2	27	3.2
Crazing & cracks	1	0.0	4	2.3	5	1.2	10	1.2
Scratched & coating loss	29	11.4	2	1.1	32	7.9	63	7.6
Others	4	1.6	5	2.8	36	8.9	45	5.4
Total	254	100.0%	177	100.0%	405	100.0%	836	100.0%

$x^2 = 175.000^*$ Degree of freedom = 24
*Very significant at 5% level of significance

Source: Air Force D Base, 1987
North Carolina A & T State University

Cracks remain to be the most frequent failure mode for Sieracin's transparencies. Nearly one out of every three (32.2%) of Sieracin's transparencies failed because of cracks. However, this has not been the case for Goodyear and Texstar. Only 9.1% of Goodyear's transparencies and 10.1% of Texstar's transparencies failed because of cracks.

Acrylic crazing is a predominant failure mode to Texstar, and the second most important problem to Sieracin's products; nearly one out of every four of their transparencies failed because of crazes. On the contrary, crazing is a relatively minor problem to Goodyear; only 6.7% of its transparencies failed because of acrylic crazes.

A Chi-Square, X^2 , test of independence was conducted to test the hypothesis that all the three vendors have the same proportions of various failure modes. The calculated Chi-Square with a value of 175.0 with 24 degrees of freedom is very significant at 5% significant level. The null hypothesis is then rejected. The conclusion is that all three vendors do not have the same proportion of failure modes. In other words, they do not have similar problems as far as transparency failure is concerned.

(3) Transparency Failures by Air Force Bases

The U. S. Air Force has bases around the globe. Many of these bases were exposed to various environmental factors such as extreme temperatures, heavy rainfall, strong dust storms, etc. Undoubtedly, many transparencies failed because of these external environmental factors. Hence, one may wonder whether the distribution of failure modes among various Air Force Bases is the same or not.

Table 4 shows the frequency as well as proportion distribution of failed transparencies among 11 major Air Force Bases where most of the failed transparencies were recorded. It can be seen that acrylic crazing mostly occurred in MacDill AF Base. About three out of every four transparencies crazed were recorded in MacDill. On the other hand, cracks occurred most often in Luke. About one out of every three cracks was recorded at Luke AF Base. Coating loss occurred most often among three bases, namely: Torrejon (19.7%), MacDill (17.9%), and Luke (17.3%). Scratches occurred also most often at MacDill (24.6%) and at Luke (20.6%).

Table 4

TRANSPARENCY FAILURE MODES

AF BASE	Acrylic crazing		Acrylic cracks		Delaminated		Scratched		FAILURE Haze	
	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
Eglin	6	4.2%	2	1.8%	0	0%	4	3.2%	0	0%
G. B.	2	1.4	13	11.4	0	0	6	4.8	0	0
Hahn	0	0	1	0.9	0	0	3	2.4	0	0
Hill	5	3.5	20	17.5	0	0	16	12.7	3	18.8
Kunsan	1	0.7	2	1.8	0	0	1	0.8	1	6.3
Luke	5	3.5	41	36.0	2	10.0	26	20.6	1	6.3
Torrejon	0	0	8	7.0	10	50.0	11	8.7	0	0
MacDill	110	76.9	17	14.9	0	0	31	24.6	7	43.8
Shaw	1	0.7	3	2.6	1	5.0	6	4.8	1	6.3
Misawa	1	0.7	0	0	5	25.0	1	0.8	0	0
Nellis	12	8.4	7	6.1	2	10.0	21	16.7	3	18.8
Total	143	100.0%	114	100.0%	20	100.0%	126	100.0%	16	100.0%

$x2 = 475.8^*$ Degree of freedom = 70

*Very significant at 0.05 level of significance

BY AF BASE

MODES						TOTAL	
Coating Loss		Chipped		Pitted/hail imp.		No.	Percent
No.	Percent	No.	Percent	No.	Percent		
4	2.5%	0	0%	1	4.2%	17	2.5%
10	6.2	10	13.0	0	0	41	6.0
21	13.0	9	11.7	2	8.3	36	5.3
6	3.7	13	16.9	11	45.8	74	10.9
1	0.6	2	2.6	0	0	8	1.2
28	17.3	16	20.8	1	4.2	120	17.6
32	19.7	4	5.2	0	0	65	9.5
29	17.9	9	11.7	7	29.2	210	30.8
16	9.9	1	1.3	0	0	29	4.2
6	3.7	0	0	0	0	13	1.9
9	5.6	13	16.9	2	8.3	69	10.1
162	100.0%	77	100.0%	24	100.0%	682	100.0%

Again, a Chi-Square test of independence was conducted to test the hypothesis that all the major AF Bases have the same failure mode proportion. The large value of calculated Chi-Square value, 475.8 is considered very significant at 5% level of significance, indicating the existence of a significant variation in proportion of transparency failure modes among various AF Bases.

Why did so many failures occur on MacDill AF Base? Was it because most of the failed transparencies were recorded there? Or was it because of some other human and environmental factors? The answers to both questions, nevertheless, might be a positive when one looks at Table 5, which shows the distribution of failed transparencies by vendors as well as by Air Force Bases. It can be seen from Table 5, nearly one-third (28.9%) of all the failed transparencies were recorded at MacDill Air Force Base, even though the proportion of recorded failures by Air Force Bases is different among vendors.

TABLE 5

VENDORS OF FAILED
TRANSPARENCIES BY A F BASE
(F-16)

A F BASE	VENDOR						TOTAL	
	Goodyear		Sieracin		Texstar		No.	Percent
	No.	Percent	No.	Percent	No.	Percent		
Eglin	2	0.8%	4	2.2%	16	4.1%	22	2.6%
G. B.	17	6.4	28	15.4	2	0.5	47	5.6
Hahn	38	14.4	14	7.7	34	8.8	86	10.3
Hill	25	9.5	23	12.6	32	8.2	80	9.6
Kunsan	2	0.8	5	2.8	2	0.5	9	1.1
Luke	43	16.3	31	17.0	58	14.9	132	15.8
Torrejon	49	18.6	0	0	17	4.5	66	7.9
MacDill	31	11.7	43	23.6	167	42.9	241	28.9
Shaw	38	14.4	11	6.1	23	5.9	72	8.6
Misawa	10	3.8	0	0	2	0.5	12	1.4
Nellis	9	3.4	23	12.6	36	9.3	68	8.2
Total	264	100.0%	182	100.0%	389	100.0%	835	100.0%

$x^2 = 227.5$* Degree of freedom = 20

*Very significant at 0.05 level of significance

V. Summary and Conclusions

Aircraft transparent enclosures are high cost items to the U. S. Air Force. Unfortunately, many transparencies failed gradually and others failed abruptly and unexpectedly. Charged with the responsibility to develop high quality and most durable aircraft transparencies, the Vehicle Equipment Division of the Flight Dynamics Laboratory of the U. S. Air Force has been, in the past several years, collecting field in-service data on replaced aircraft transparencies in the hope that valuable information on factors attributable to transparency failures will become available. By controlling some or all of these factors, more durable transparencies might be developed.

During the course of this study, two data bases have been created using data collected by engineers at the Air Force and dBASE III computer software. One consists of 953 records of displaced transparencies taken from F-16 jet fighters; and the other consists of 678 records of displaced transparencies taken from F-111. Data-base creation is one of the objectives that have been met. Through an extensive statistical analysis, coating loss has been found to be a predominant failure mode,

accounting for nearly one-fourth of the displaced transparencies from F-16. Abrasion or scratches is the second most frequent failure mode, accounting for 18.8% of total transparent failures. Both coating loss and scratches, accounting for a combined 42.3% of total transparent failures, are mostly caused by poor maintenance practices. Hence, efforts to install proper maintenance may prove to be very cost-effective in reducing transparency related expenses.

Crazing ranked the third most frequent failure mode. It is found in nearly one out of every five (18.4%) displaced transparencies. And cracks ranked the fourth most frequent failure mode. There could be numerous factors contributing to the formation of crazes and cracks on aircraft transparencies. However, studies on factors contributing to causing aircraft transparencies to craze and to crack have been lacking.

Variations in the frequency of various failure modes do exist among different vendors. For example, coating loss remains to be the most important failure mode for Goodyear. Over one third of Goodyear's transparencies failed were attributable to coating loss. To Sieracin, however, cracks are the utmost important failure mode.

Nearly one out of every three of Sieracin's transparencies failed because of cracks. On the other hand, acrylic crazing is the predominant failure mode to Texstar. Nearly one out of every four of Texstar's transparencies failed because of crazing. Readers are cautioned, however, that this phenomenon may not be interpreted as an indication that one vendor's transparency quality is better than another's, for the simple fact that aircraft transparencies from all vendors were not subjected to the same environmental and human factors. More data are needed to examine the relationship between various human and environmental factors and various failure modes.

REFERENCES

1. Richard B. Chase and Nicholas J. Aquilano: Production and Operations Management - A Life Cycle Approach, Third Edition, Richard D. Irwin, Inc. Homewood, Illinois 1981. p. 490.

2. Kenneth I. Clayton, John F. Milholland and Gregory J. Stenger: Experimental Evaluation of F-16 Polycarbonate Canopy Material, University of Dayton Research Institute, Dayton, Ohio 1981.

3. Kenneth I. Clayton and Blaine S. West: Aircraft Transparency Testing Methodology and Evaluation Criteria, Part I and Part II. University of Dayton Research Institute, Dayton, Ohio 1983.

FINAL REPORT TO UNIVERSAL ENERGY SYSTEMS
(CONTRACT UES/AFOSR F49620-85-C-0013/SB5851-0360)

TRAJECTORY STUDIES OF THE BIMOLECULAR REACTION OF H_2O^+/H_2O

C. Randal Lishawa
Jefferson State Junior College
Birmingham, Alabama 35215

The cross-sections for the bimolecular reaction of H_2O with H_2O^+ over the energy range 0.5 eV to 50 eV have been calculated by classical trajectory methods using the long-range potential of a point ion reacting with a polarizable point dipole. A new program for the IBM-PC was developed to carry out these calculations. These calculations are compared with prior experimental data.

INTRODUCTION

With the refinement of molecular beam techniques in the early 1970's, there was a resurgence of interest in ion-molecule reactions. The reason for this was that experimentally the ion-beam was easily controlled over a wide range of collision energies. Theoretically the long-range potential was easier to describe for ion-molecule reactions than the quantum mechanical description for neutrals. In the 1980's, concern over the fate of the environment has resurfaced an interest in ion-molecule reactions with an emphasis on upper atmosphere chemistry. Such problems as the thinning of the ozone layer, the fate of chlorinated fluorocarbons (CFC's) released into the atmosphere, and the chemical reaction dynamics of the reentry of large payloads from space (such as the Space Shuttle) have been occupying the attention of a segment of the scientific community.

Even though much detailed work has been conducted on quantum mechanical descriptions of chemical reactions,[1-3] some of the older and relatively simple theories still give good descriptions of the reaction cross-sections for ion-molecule reactions.[4] The long-range potential of the ion-permanent dipole has been found to provide an accurate description of charge-exchange reactions as well as other reactive systems.[5-6]

Recent measurements of the reactive cross-section for the H_2O/H_2O^+ system[7] disagree with previous measurements[8] by more than an order of magnitude. One method of examining this difference is to calculate the theoretical cross-sections. The potential energy function describing this

system (and any similar ion-molecule system) is

$$V(R) = -\frac{\alpha q^2}{R^4} - \frac{\mu_D q}{R^2}\cos\theta \qquad (1)$$

where α and μ_D are the polarizabilty and dipole moment of the neutral respectively, q is the charge on the ion and R is the separation of the centers-of-mass. The angle θ is the angle between the line of centers of the reactants and the orientation of the dipole moment. Additional terms involving such quantities as the polarizability and/or the dipole moment of the ion,[9] as well as higher moments[10-12] of the charge distribution, may also be included depending on the availability of the required constants. The difficulty in handling this type of potential energy function lies in the deriving an expression for the geometric term $\cos\theta$. Chesnavich, Su, and Bowers[13] have obtained an expression for the average value of the geometric term from the expression

$$\langle\cos\theta\rangle = \frac{\int\cos(\theta)d\theta}{\int P(\theta)d\theta} \qquad (2)$$

where,

$$P(\theta) \alpha \frac{\sin\theta}{\dot{\theta}} . \qquad (3)$$

Darker and Ridge[14] have approximated the term by

$$P(\theta) \alpha \sin\theta\, e^{\left(\frac{q\mu}{R^2kT}\right)\cos\theta} \qquad (4)$$

which results in the expression

$$\langle\cos\theta(R)\rangle = \pounds\left(\frac{q\mu}{R^2kT}\right), \qquad (5)$$

where $\pounds$ is the Langevin function.[15]

The approach used in this paper is that of Kech,[16] Anderson[17] and Su,[18] and is an alternative to obtaining an explicit functional form for the geometric dependence of the system. In this approach, the initial angle was chosen by Monte Carlo sampling and the equations of motion were integrated. This paper assumes that when R reaches a predefined value (R_0) a reaction will occur. This approach obviously suffers from not being able to predict which of the possible product channels will be populated. It

also does not take into account the internal motions of the reactants, but it has proven to be predictive of several different reaction[19] types.

CALCULATION

A trajectory of the collision pair was calculated by integrating the classical Hamiltonian[20] equations for the system:

$$\frac{\partial H}{\partial x_i} = -\dot{p}_i \qquad \frac{\partial H}{\partial p_i} = \dot{x}_i \tag{6}$$

where the Hamiltonian H is given by

$$H = \frac{p_R^2}{2\mu} + V(R) \tag{7}$$

and x_i ($\dot{x}_i$) is the position (velocity) vector of the i^{th} particle, p_i ($\dot{p}_i$) is the momentum (force) vector for the i^{th} particle, and μ is the reduced mass of the system. A trajectory was counted as reactive if the collision partners reached a separation distance R_0 given by the Langevin criteria[21]

$$R_0 = \left(\frac{\alpha q^2}{2E_T} \right)^{\frac{1}{4}} \tag{8}$$

Although several good trajectory programs were available through the Quantum Chemistry Program Exchange (MERCURY, AB•C, and CTAMYM),[22-24] they were not suited for this study. Therefore, a new program was written to calculate the trajectories. This program was written in Microsoft Fortran V2.0 and run on an IBM-PC microcomputer.

The system of equations generated from Hamilton's formulation, was integrated using the predictor-corrector method of Hamming.[25] Initial seed values required by the Hamming method were generated by fourth order Runga-Kutta employing Kutta's coefficients.[26] The stability of the solution was checked both by reducing the step-size of the integration and by back integration of the trajectory. Calculations began at 16 Å and ended when the trajectory reached R_0 or returned to 16 Å. The impact parameter b and the initial angle of approach θ were chosen by random number (ξ), where b was chosen from $0 \leq \xi \leq 1$ and was weighted as

$$b = b_{max}\xi^{\frac{1}{2}} \tag{9}$$

and θ was allowed to vary from $0 \rightarrow 2\pi$. Other constants used in this study were α = 1.48 Å^3, μ = 1.84 D, and q = 4.80 x 10^{-10} esu.[27]

Two different cases were examined in this study. The first case used equation (1) without any additional constraints on the motion. In the second case the system was constrained to maintain a constant orbital angular momentum (L) by adding an effective potential term of the form[28]

$$V_{eff} = \frac{L^2}{2\mu R^2} = \frac{2E_T{}^2 b^2}{\mu R^2} \tag{10}$$

to equation (1).

The cross-sections calculated were based on the number of trajectories performed and the number of reactions observed.[29] The following equations were used

$$\sigma_r = \pi b_{max}^2 \langle P_r \rangle \tag{11}$$

where b_{max} is the maximum impact parameter sampled. $\langle P_r \rangle$ is given by

$$\langle P_r \rangle = \frac{N_{reactive}}{N} \tag{12}$$

where $N_{reactive}$ is the number of reactive trajectories computed and N is the total number of trajectories sampled. The estimated percentage error is given by

$$\Delta = \left[\frac{(N - N_{reactive})}{N N_{reactive}} \right]^{1/2} \tag{13}$$

For each individual cross-section calculated in this study, 300 trajectories were calculated, with b_{max} equal to 2 Å.

RESULTS

The results of these calculations are summarized in Figure 1. For case 1, where no constraints are imposed on the system (equation 1) we find that the cross-section has a maximum occurring approximately at 2 eV and falls off over the remainder of the energy range.

For case 2 (equation 1 + equation 9), we find that the cross section falls off throughout the range of interest. This cross-section is only about a factor of two above the charge exchange cross-section measured by Lishawa,

Salter, and Murad[7] and nearly an order of magnitude less than the total reaction cross-section reported by Ryan.[8]

Even though this model gives good agreement with the experimental cross-sections of Lishawa et. al.,[7] it does not include detailed examination of the various product channels. Because the model does not incorporate the various product channels, we expect the calculated cross-sections to lie above any individual product channel as we observed in case 2. We also would expect this model to provide only an upper limit to the cross-sections over the collision energy range. To improve the accuracy of this model, work is currently under way to provide data on the cross-sections for the various product channels for this reaction.

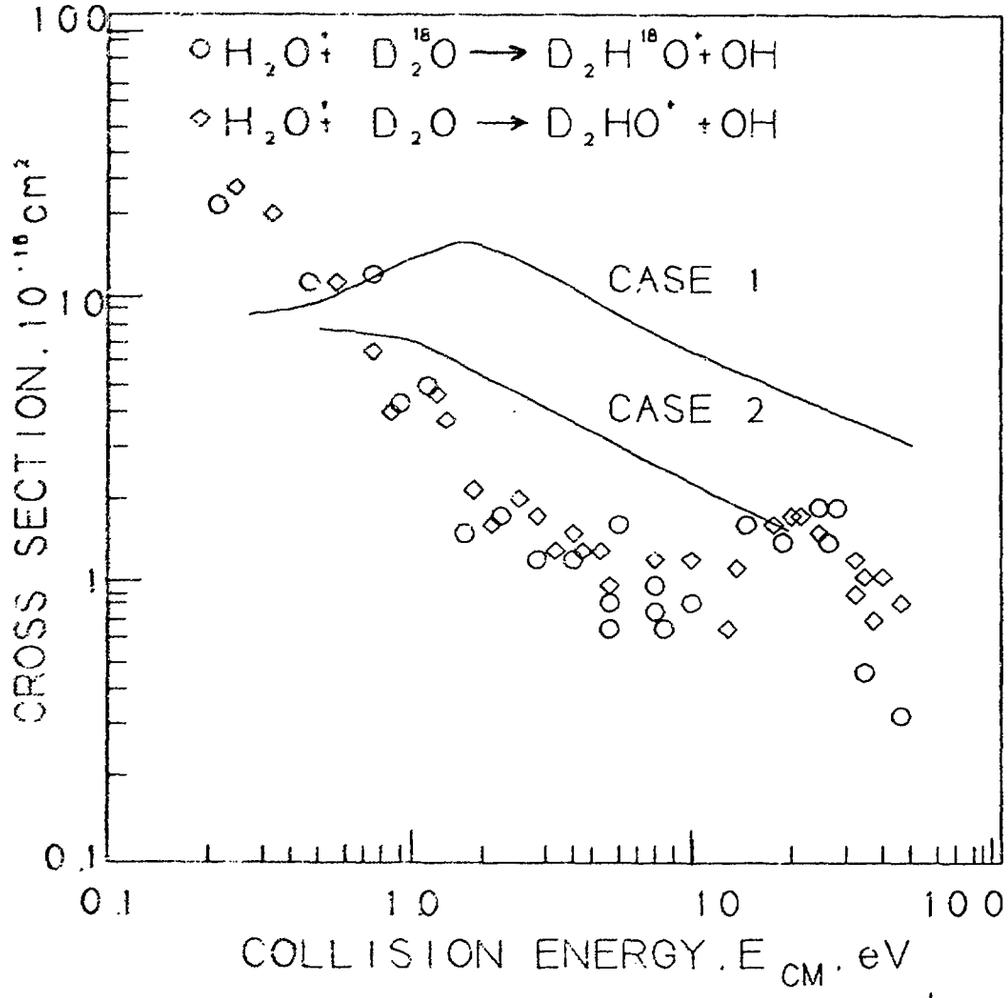


Figure 1. Cross-Sections for the Bimolecular Reaction of H_2O^+/H_2O. The solid lines are theorectical cross-sections as described in the text. The individual symbols are experimental data.

REFERENCES

1. Atom-Molecule Collision Theory: A Guide for the Experimentalist, Bernstein, Richard B. (ed.), Plenum Press: New York, 1979.
2. Theory of Chemical Reaction Dynamics. Vol. I-IV, M. Baer, (ed.) CRC Press Inc.: Boca Raton, FL, 1985.
3. Molecular Reaction Dynamics, Levine, R.D. and Bernstein, R.B., Oxford University Press: New York, 1974.
4. W.J. Chesnavich, T. Su, and M.T. Bowers, "Ion-Dipole Collisions: Recent Theoretical Advances" in Kinetics of Ion-Molecule Reactions, Ausloos, Pierre (ed.), Plenum Press: New York, 1979, pp. 31-53.
5. F.T. Smith, D.L. Heustis, D. Mukherjee, and W.H. Miller, Phys. Rev. Lett. 35, 1073 (1975)
6. L.M. Bass and M.T. Bowers, J. Chem. Phys., 86, 2611 (1987).
7. C.R. Lishawa, R.H. Salter, and E. Murad, in press.
8. Ryan, K.R., J. Chem. Phys., 52, 6009 (1970).
9. S. Hu and T. Su, J. Chem. Phys., 85, 3127 (1986).
10. T. Su, E.C.F. Su, and M.T. Bowers, Int. J. Mass Spectry. Ion Phys., 28, 285 (1978).
11. P.K. Bhowmik and T. Su, J. Chem. Phys., 84, 1432 (1986).
12. L.F. Phillips, Faraday Trans. 2, J. Chem. Soc., 83, 857 (1987).
13. W.J. Chesnavich, T. Su, and M.T. Bowers, J. Chem. Phys., 72, 2641 (1980).
14. R.A. Barker and D.P. Ridge, J. Chem. Phys., 64, 4411 (1976).
15. W.J. Moore, Physical Chemistry, Prentice-Hall, Inc.: Englewood Cliffs, New Jersey, 1972, p. 703.
16. J.C. Keck, J. Chem. Phys., 32, 1035 (1960).
17. J.B. Anderson, J. Chem. Phys., 58, 4684 (1973).
18. T. Su and W.J. Chesnavich, J. Chem. Phys., 76, 5183 (1982).

19. M.T. Bowers and T. Su, "Theory of Ion Polar Molecular Collisions", in Interaction Between Ions and Molecules., P. Ausloos (ed.), Plenum Press: New York, 1975.

20. H. Goldstein, Classical Mechanics, Addison-Wesley: Reading, 1980.

21. P. Langevin, Ann. Chim. Phys., 5, 245 (1905).

22. W.L. Hase, QCPE 453 - MERCURY.

23. S. Chapman, D.L. Bunker, A. Geib, QCPE 273 - A+BC.

24. D.G. Hopper, QCPE 248 - CTAMYM.

25. B. Carnahan, H.A. Luther, and J.O. Wilkes, Applied Numerical Methods, John Wiley & Sons, Inc.: New York, 1969, pp. 391-404.

26. Ibid. pp. 361-380.

27. Op. Cit., W.J. Moore.

28. Op. Cit., R.D. Levine and R.B. Bernstein.

29. L.M. Raff and D.L. Thompson, Theory of Chemical Reaction Dynamics. Vol III, M. Baer (ed.), CRC Press, Inc.: Boca Raton, FL, 1986, pp. 1-108.

FINAL REPORT NUMBER 48
RECEIVED FINAL REPORT
NOT PUBLISHABLE AT THIS TIME
Dr. Cheng Liu
760-6MG-009

FINAL REPORT NUMBER 49
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Stephen Loy
760-6MG-134

FINAL REPORT NUMBER 50
RECEIVED A NO-COST TIME EXTENTSION
TO BE SUBMITTED IN 1987 MINI-GRANT FINAL REPORT
Dr. Arthur Mason
760-6MG-099

FINAL REPORT NUMBER 51
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Gopal Mehrotra
760-6MG-121

FINAL REPORT NUMBER 52
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Jorge Mendoza
760-6MG-136

FINAL REPORT NUMBER 53
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Dakshina Murty
760-6MG-079

1986-1987 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

Development of a New Finite Element Grid for Limited Area Weather Models

Prepared by:	Dr. Robert M. Nehs Principal Investigator
Academic Rank:	Associate Professor
Department:	Department of Mathematics
University:	Texas Southern University
Senior Investigators:	Dr. O. H. Criner Dr. Victor Obot
Graduate Student:	Cornell Wooten

Date:	October 20, 1987
Mini-grant Contract No.	F49620-85-C-0013/SB5851-0360
P.O. No.	S-760-6MG-120

DEVELOPMENT OF A NEW FINITE ELEMENT GRID FOR LIMITED AREA WEATHER MODELS

by Robert M. Nehs

ABSTRACT

During a preliminary study of the use of finite element techniques in limited area atmospheric modeling, a new two dimensional finite element grid was developed. The results of a follow-up investigation of this grid are reported.

Computer models were developed for a variety of problems which could be used with either the new grid or a comparable variable resolution rectangular grid. Experiments were conducted using the Poisson equation with either essential or natural boundary conditions and also the heat equation. Several alternative numerical integration procedures were tested in the program. Further studies were conducted using refinements of both grids. The accuracy of each model generated solution was measured by comparing the results with the analytic solution at each of the grid points of the configuration. The relative efficiencies of the different models were determined by comparing the CPU times required for the numerical computations.

ACKNOWLEDGEMENTS

I would like to acknowledge my indebtedness to Oscar H. Criner, Victor Obot, and Cornell Wooten for their valuable contributions to the work on this project. I wish to express my appreciation for the use of the facilities at Texas Southern University and for the assistance of Etta Walker and Mohammad Ansarizadeh of the Computer Science Department.

Special thanks are due to Don Chisholm and Samuel Yee at the USAF Geophysics Laboratory, the former for the availability of the facilities at the Laboratory during my brief visit and the latter for the several suggestions relating to the project.

Grateful acknowledgement is made of the support from the Air Force Systems Command and the Air Force Office of Scientific Research.

Finally, I wish to thank my family for their patience and encouragement during this project.

1. INTRODUCTION

During the summer of 1986 the author and a graduate student, Cornell Wooten, conducted a preliminary study of the application of the finite element method (FEM) in limited area weather modeling for the Air Force Geophysics Laboratory under the supervision of Samuel Yee. The Canadian limited area finite element model was examined, especially the development of the grid configuration. A key component of this scheme is a two dimensional rectangular grid covering a large region containing a smaller area of forecasting interest. In [4] there is a description of a finite element two dimensional barotropic model which employs a uniform high resolution rectangular grid over the entire domain. In a later paper [5] similar results were achieved for a limited time period using significantly less computer time. This was accomplished through the use of a nonuniform rectangular grid configuration having high resolution over the area of interest and lower resolution outside this area. The best results were produced when the resolution decreased smoothly from the high resolution region outward toward the domain boundary (Figure 1). The savings, both in storage and number of computations, is a result of the reduction in the number of grid points (or nodes). However, there are some areas in the outer region where the resolution is still unnecessarily high.

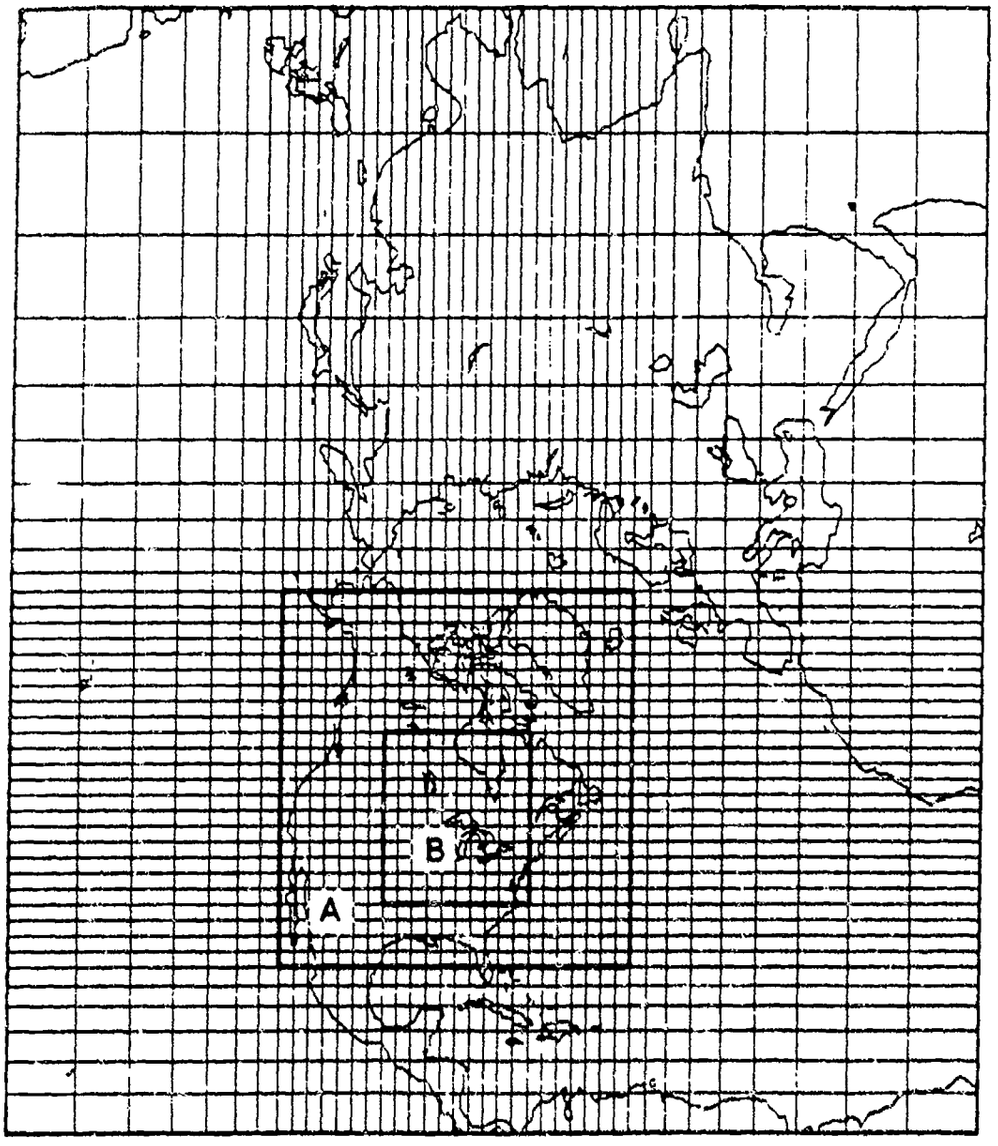


FIGURE 1.

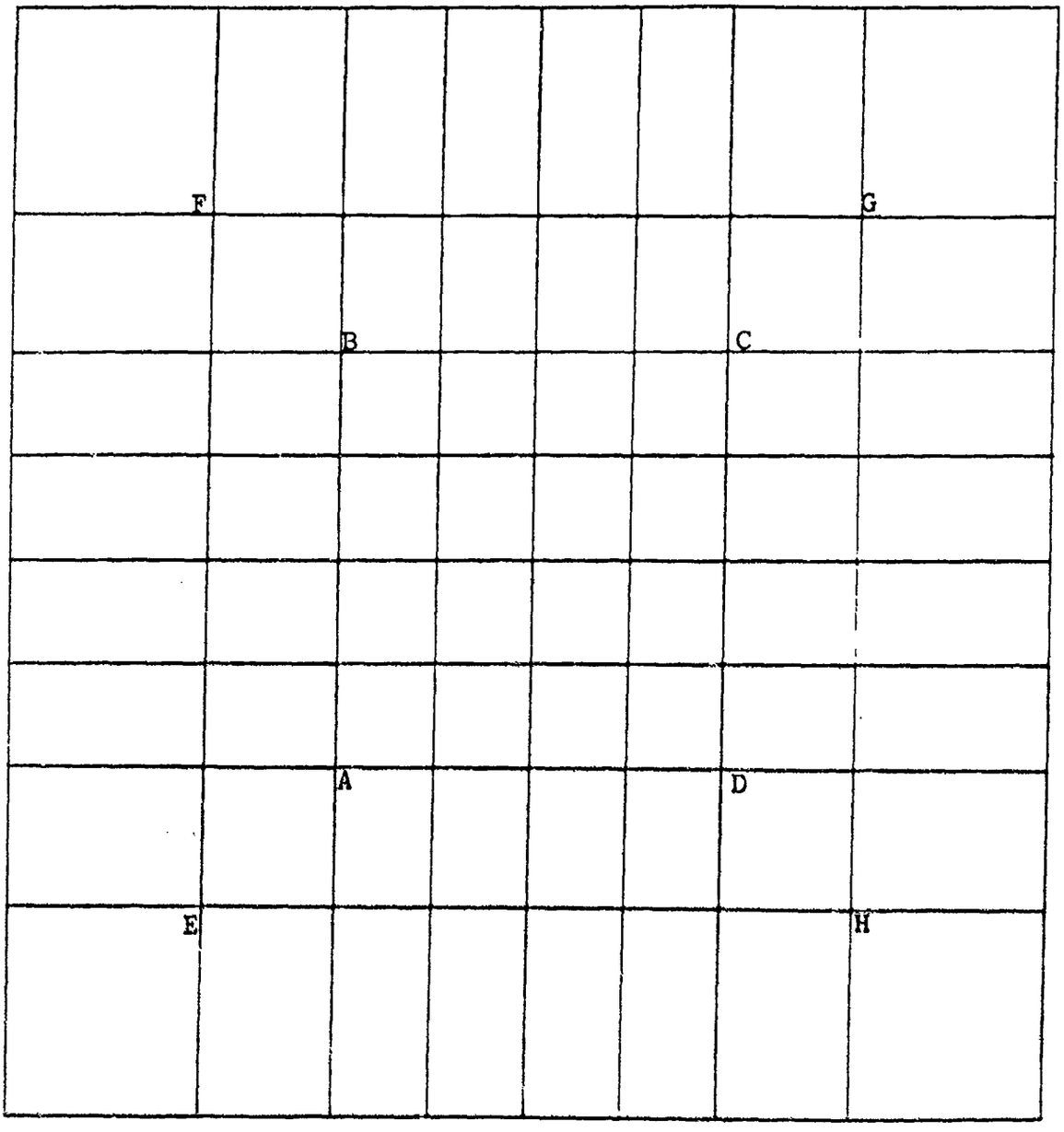


FIGURE 2.

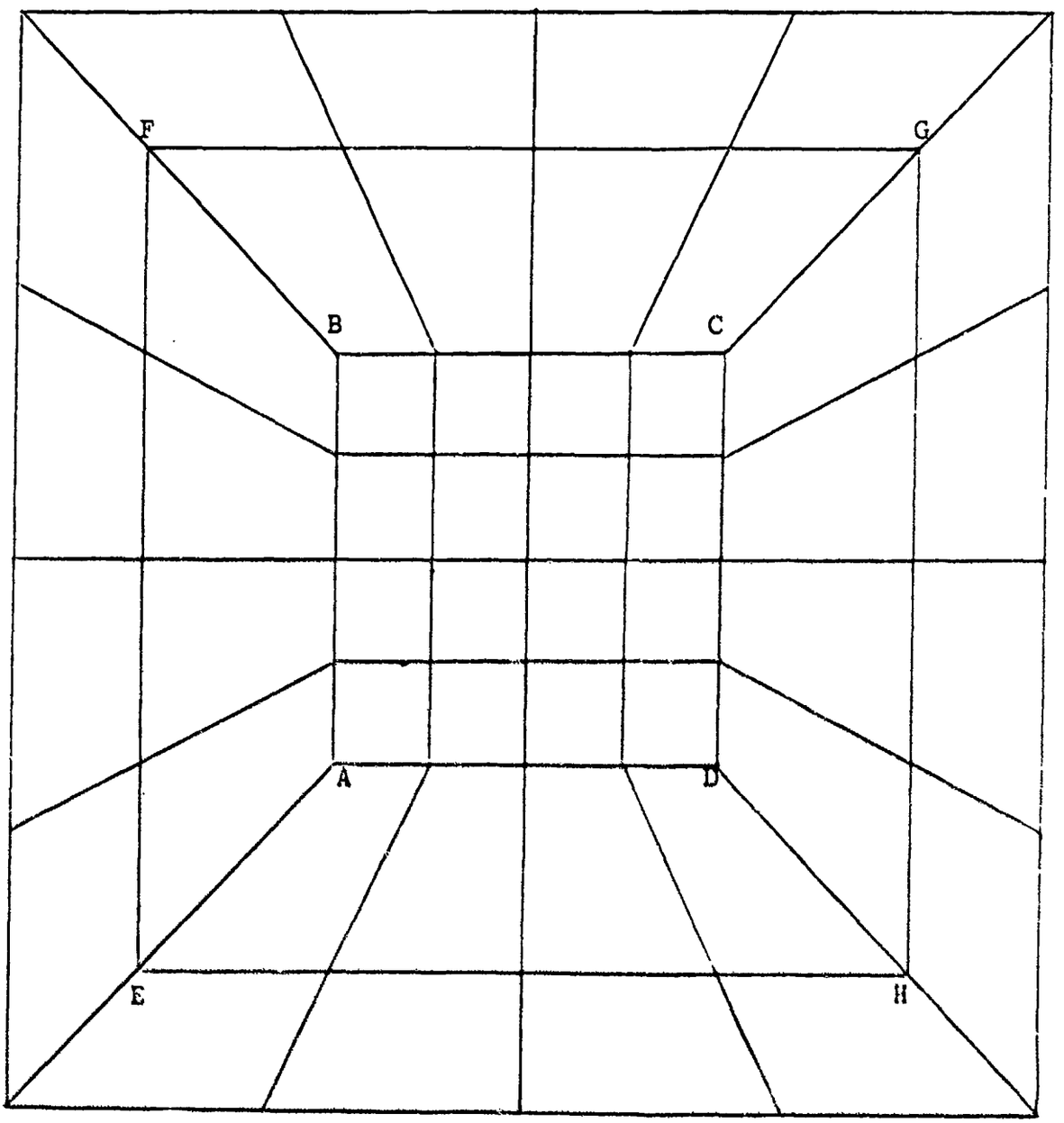


FIGURE 3.

It seemed possible that a more economical mesh could be formed by the use of non-rectangular elements outside the area of interest. Consequently, an alternative configuration, the RECTRAP grid, was developed as part of the 1986 summer project. This grid, which consists of a combination of rectangular and trapezoidal elements, achieves the same uniform high resolution over the forecasting area with fewer total elements. Simplified versions of both the variable resolution rectangular grid and the new mixed element grid are shown in Figures 2 and 3. Square ABCD represents the high resolution area of interest in both configurations. In order to compare these configurations, a finite element program was written for the Poisson problem

$$u_{xx} + u_{yy} = f$$

$$u = g \text{ (boundary condition)}. \qquad (1)$$

It was designed to be used with either of the meshes in Figures 2 and 3. Both grids were tested with this program using $f(x) = \sin(\pi x)$ and $g(x) = 0$ in (1). The corner diagonals of the domain were set at (0,0) and (1,1) and ABCD was chosen to be the centered square A = (.25,.25) and C = (.75,.75). The results were compared with the analytic solution,

$$u = \frac{\sin(\pi x)}{\pi^2}\left(\frac{e^{\pi y}+e^{\pi(1-y)}}{e^{\pi}+1} - 1\right),$$

at the grid points. The new grid produced smaller errors than the rectangular grid at the points inside the high resolution square ABCD and at approximately half of the remaining points.

Investigation of the new mesh was continued under a mini-grant during the spring and summer of 1987. There were two goals of this research project. The first was to develop additional software for finite element modeling and the second was to make further comparisons of the two types of grids. The principal objective was to determine whether the new mixed element grid could achieve nearly as accurate results more efficiently as the comparable rectangular grid.

During the initial stage of the project both grids were tested with several choices of f and g in (1). Results were compared for various placements of the intermediate nodes - ie., the non-boundary grid points outside square ABCD. Alternative numeric integration techniques were tested next to determine the effect on the accuracy and efficiency of the program. A new storage procedure for the FEM system matrix was also tested. Following this the program was modified so that the Poisson equation with a different boundary condition could be studied.

$$u_{xx} + u_{yy} = f$$

$$\frac{\partial u}{\partial n} + gu = h \text{ (B.C.)}$$

At this stage of the project runs were conducted using larger versions of each of the grids. Software was developed that would automatically generate the initial mesh data using minimal input data. Finally, a program was written for a time dependent problem, the heat equation:

$$u_{xx} + u_{yy} = u_t$$

$$u = f \text{ (boundary condition, } t > 0)$$

$$u = g \text{ (initial condition, } t = 0).$$

This program was tested (with each grid) using different time steps.

The accuracy of each run was tested by comparing the results with the analytic solution at each node within the grid. In many cases the root mean squares of the errors and the percent errors were calculated for the nodes over the fine mesh square ABCD and also over the entire grid. The program efficiency was measured using the CPU time required to assemble the system matrix, impose the boundary constraints, and solve the resulting system of equations.

The programs were developed and tested on the AT&T 3B15 at Texas Southern University. After the work was completed Dr. Nehs traveled to Hanscom AFB in Boston to confer with Samuel Yee. The results from the research were discussed along with several ideas for future investigations. Dr. Nehs also presented an informal seminar during which he described briefly the project and some of his results. In addition to this, many of the project programs were run on the VAX 8650 in the Geophysics Laboratory at Hanscom. The results from these tests were essentially the same as those obtained from the 3B15, although the computer times were reduced considerably.

Dr. Robert Nehs was the Principal Investigator for this research project. Dr. Oscar H. Criner and Dr. Victor Obot, two faculty members at Texas Southern University, assisted him as Senior Investigators. Mr. Cornell Wooten served on the project as a Graduate Research Assistant.

2. SUMMARY OF RESEARCH

2.1 Placement of Intermediate Nodes

Each grid in Figures 2 and 3 contains an inner central square ABCD covered by a uniform rectangular mesh that is surrounded by two layers of rectangles or trapezoids. The corners of the domain boundary (the outer square) are fixed at (0,0), (0,1), (1,1), and (1,0) while those of the inner square are A = (.25,.25), B = (.25,.75), C = (.75,.75), and D = (.75,.25). The intermediate square EFGH can be located anywhere between; E = (h,h), F = (h,1-h), G = (1-h,1-h), and H = (1-h,h) where h is the distance from the outer square to the intermediate square, $0 < h < .25$.

The effect that the location of the intermediate square had on the results was studied first. A program was developed that would produce results from the finite element model of Equation (1) for up to ten different values of h and compare each set of results. The error was measured at the grid points, where the error equals

value of analytic solution - approximation from program.

A test run was made with each grid for $f = \sin(\pi x)$ and $g = 0$ using the following values for h:

h = .025, .050, .075, .100, .125, .150, .175, .200, .225.

The results from the rectangular grid are consistent; the

errors are minimized at every node when h = .15. The results from the RECTRAP grid are less conclusive. Within the inner square ABCD the minimum errors are obtained for $.125 \leq h < .200$; the larger values of h produce best results at the most central nodes and the smaller values of h give better approximations at the nodes away from the center of the grid. The root mean square (rms) of the errors over the central grid nodes is minimized when h = .15. Generally, the RECTRAP grid achieves more accurate results over the central grid points while the rectangular grid produces better results over the intermediate nodes.

Similar tests were conducted for various choices of f and g in Equation (1):

PROBLEMS

	f	g
1.	$\sin(\pi x)$	xy
2.	$\sin(\pi x)$	$\sin(x)\cosh(y)$
3.	0	$\sin(\pi x)\sinh(\pi y)$
4.	0	$\sin(\pi x)\cosh(\pi y)$
5.	0	$\cos(\pi x)\sinh(\pi y)$
6.	0	$\cos(\pi x)\cosh(\pi y)$
7.	$-\sin(\pi x)\sin(\pi y)$	0
8.	$-\sin(\pi x)\sin(\pi y)$	$\sin(x)\cosh(y)$
9.	$-\sin(\pi x)\cos(\pi y)$	$\frac{\sin(\pi x)\cos(\pi y)}{2\pi^2}$

In each case the rectangular grid gives optimal or near optimal results at each node within the inner square when h = .15. The results at the intermediate nodes tend to improve when h is decreased. Altogether, the errors are minimized over the entire grid, as measured by the rms, for h between .125 and .175. The results from runs using the RECTRAP grid are not as definite. The value of h which produces the minimum error at a particular node depends on the node and the problem to a greater degree than with the

rectangular grid. Generally, the optimizing values of h for the nodes within ABCD range from .100 to .175 while the rms is minimized or nearly minimized when h = .15. In two extreme cases (Problems 4 and 5 above) the majority of the central nodal errors are minimized for h = .200 and .225 and the rms is minimized when h is .200. The results for the intermediate nodes are even more mixed. In some cases each value of h produces a minimum error for at least one intermediate grid point. For most of the runs the rms of the errors over the entire grid is minimized when h is .125. In Problems 4 and 5 the best value of h is .175 or .200.

The model was run with each of the problems above using h = .15 in both grids. The errors from each of the grids were compared. In most cases the RECTRAP grid give better results at the nodes toward the center and the rectangular grid produces smaller errors at the nodes further away from the center. In three of the problems the rectangular grid produces better results over the entire domain; however, the results from the RECTRAP grid are improved toward the center when h is increased. The results over the intermediate nodes tend to be better in all cases when the rectangular grid is used. When h = .15, the error rms over the entire grid is better for the rectangular grid, in most cases by a factor between 1.5 and 3. The value h = .15 was used with both grids for the remainder of the project.

2.2 Alternative Numerical Integration Formulas

In the FEM program each element in the mesh is mapped onto a standard square e_s with vertices $(\pm 1, \pm 1)$. The integrals necessary to determine the system matrix are calculated over e_s using numerical techniques. In the original program the 9 point (3-by-3) Gaussian quadrature formula was used to approximate these integrals. Software was developed during this project so that the model could be tested using alternative numerical integration schemes; the Gauss 4 point (2-by-2) and the Gauss 16 point (4-by-4) quadrature formulas, the product trapezoidal rule, and the product Simpson rule. The two dimensional trapezoidal rule for e_s is a four point formula based on a bilinear interpolation of the integrand evaluated at the four vertices:

$$f_1 + f_2 + f_3 + f_4.$$

The two dimensional Simpson rule for e_s is a nine point formula using a biquadratic interpolation:

$$\frac{1}{9}(f_1+f_2+f_3+f_4+4(f_5+f_6+f_7+f_8)+16f_9),$$

where f_1, f_2, f_3, f_4 are the values of the integrand f at the vertices, f_5, f_6, f_7, f_8 are the values of f at the midpoints of the sides of e_s, and f_9 is the value at the center. Each of these four integration schemes was tested in the FEM

program using the original problem ($f = \sin(\pi x)$, $g = 0$) and $h = .15$ for both grids. The approximations produced from each of the three Gauss rules and the product Simpson rule are almost the same for the rectangular grid. However, when the product trapezoidal formula is used, the errors are reduced significantly at each of the nodes, especially those within ABCD. At first this seemed surprising since this is the simplest of the five schemes which, together with the 4 point Gauss formula, requires the fewest calculations per application. The phenomenon is probably explained by the facts that the mapping of each rectangular grid element onto e_s is bilinear and the errors are measured at the element vertices.

The five numerical integration procedures were also tested with the RECTRAP grid together with combinations in which a higher order rule was used for the trapezoids in the grid. These consisted of the Gauss n point rule for the rectangles and the Gauss m point rule for the trapezoids, $n < m$, and also the trapezoidal rule and Simpson's rule for the rectangles and the trapezoids respectively. As before, the Gauss formulas, the mixed Gauss formulas, and the product Simpson rule produced similar results. However, unlike the tests with the rectangular grid, the use of the product trapezoidal rule gave the largest errors at each node. (The mappings of trapezoids onto e_s are not

bilinear.) The combination of the trapezoidal rule and Simpson's rule yielded the smallest errors at some, but not all, of the nodes.

2.3 Skyline Storage

The finite element procedure produces a system of linear equations which can be represented by a matrix equation $S \cdot D = C$. Generally S and C are called the system matrix and the system vector respectively. In the original program the system matrix was stored in a packed form determined by the half-bandwidth of the sparse symmetric matrix S. (The half-bandwidth of a matrix equals the maximum number of entries of any one row from the diagonal term to the last nonzero term.) If S is an n x n matrix and m is the half-bandwidth, the packed matrix S^* is an n x m matrix. Any row of S^* contains the first m terms of the corresponding row of S starting with the diagonal term. Thus S^* contains all of the nonzero terms of S on or above the diagonal.

The skyline packing procedure, described in [1], is an alternative storage technique for sparse symmetric matrices. Each column of S from the diagonal term up to the last nonzero term is stored in a vector V^*. A second vector N is needed for bookkeeping purposes, N(k) = the number of terms stored in V^* from column k of S, $1 \leq k \leq n$. Observe that S^*

and V^* are floating point arrays while N is fixed point. Generally the use of the skyline procedure will reduce the storage requirements and the number of calculations necessary to solve the system of equations. Furthermore, it is easier to modify this process when S is non-symmetric.

EXAMPLE

$$S = \begin{bmatrix} 1.0 & 2.0 & 4.0 & 0.0 & 0.0 \\ 2.0 & 3.0 & 0.0 & 0.0 & 0.0 \\ 4.0 & 0.0 & 5.0 & 6.0 & 0.0 \\ 0.0 & 0.0 & 6.0 & 7.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 8.0 \end{bmatrix} \quad S^* = \begin{bmatrix} 1.0 & 2.0 & 4.0 \\ 3.0 & 0.0 & 0.0 \\ 5.0 & 6.0 & 0.0 \\ 7.0 & 0.0 & 0.0 \\ 8.0 & 0.0 & 0.0 \end{bmatrix} \quad V^* = \begin{bmatrix} 1.0 \\ 2.0 \\ 3.0 \\ 4.0 \\ 0.0 \\ 5.0 \\ 6.0 \\ 7.0 \\ 8.0 \end{bmatrix} \quad N = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

This procedure was tested with each grid using modified versions of the original FEM program. The final results are unchanged but the CPU times are improved slightly. (See Section 2.7 for a discussion of the CPU times for the different programs.) The following table compares the storage requirements for each matrix packing technique when the grids in Figures 2 and 3 are used.

GRID	NUMBER OF ARRAY TERMS			
	S	S^*	V^*	N
rectangular	6561	891	801	81
RECTRAP	3249	627	541	57

2.4 Other Types of Boundary Conditions

The boundary of the problem domain, denoted by G, is the outer square. The boundary constraint in Equation (1), u = g on G, is an essential boundary condition; it is imposed directly on the approximating function u at the boundary nodes. Equations (2) and (3) below list other types of boundary conditions (n represents the outward unit normal vector to the boundary).

$$\frac{\partial u}{\partial n} + gu = h \text{ on } G \tag{2}$$

$$u = g \text{ on } G_1, \ \frac{\partial u}{\partial n} = h \text{ on } G_2, \quad G = G_1 \cup G_2 \tag{3}$$

Equation (2) represents a natural boundary condition, the system matrix S and the system vector C are modified to implement this constraint. If $\{\phi_i\}$ is the set of finite element shape functions, $\int_G g\phi_i\phi_j dG$ is added to the i,j-term of S whenever nodes i and j are the endpoints of a boundary segment and $\int_G h\phi_i dG$ is added to the i-term of C when node i is a boundary node. Boundary condition (3) represents a

mixture of the two types of conditions, the first one is imposed on u at the boundary nodes along G_1 and the second one is used to modify those terms of C corresponding to the nodes on G_2.

Using techniques in [1], the finite element program was modified to model boundary value problems of the form

$$u_{xx} + u_{yy} = f$$

$$\frac{\partial u}{\partial n} + gu = h \text{ on } G. \tag{4}$$

It should be noted that the FEM method is applicable to this type of system provided $g > 0$ [3]. The new program was tested on the six problems listed below.

Boundary Value Problems (Equation (4))

1. $f = \sin(\pi x)$, $g = 1$, $h = \begin{cases} 2 - \frac{1}{\pi}\left(\frac{e^{\pi y}+e^{\pi(1-y)}}{e^{\pi}+1}-1\right), & \text{if } x=0,1 \\ 2 + \frac{\sin(\pi x)}{\pi}\cdot\frac{e^{\pi}-1}{e^{\pi}+1}, & \text{if } y=0,1 \end{cases}$

2. $f = 0$, $g = \pi$, $h = \begin{cases} -\pi e^{\pi y}, & \text{if } x=0,1 \\ 0, & \text{if } y=0 \\ \pi e^{\pi}\sin(\pi x), & \text{if } y=1 \end{cases}$

3. $f = 0,\ g = \pi,\ h = \begin{cases} 2\pi - \pi e^{\pi y}, & \text{if } x=0,1 \\ \pi, & \text{if } y=0 \\ 2\pi + 2\pi e^{\pi}\sin(\pi x), & \text{if } y=1 \end{cases}$

4. $f = -(\sin(\pi x)+\sin(\pi y)+2\sin(\pi x)\sin(\pi y)),\ g = \pi,\ h = 0$

5. $f = \dfrac{-\sin(\pi x)\cos(\pi y)}{2\pi^2},\ g = 1,\ h = \begin{cases} \dfrac{20-\cos(\pi y)}{2\pi}, & \text{if } x=0,1 \\ \dfrac{20+\sin(\pi x)}{2\pi}, & \text{if } y=0 \\ \dfrac{20-\sin(\pi x)}{2\pi}, & \text{if } y=1 \end{cases}$

In most of these tests the errors tend to be more uniform over all of the nodes for the rectangular grid. The RECTRAP grid produces better results over most of the inner square nodes in Problems 4 (4/5 of the nodes) and 5 (about 2/3 of the nodes). However, in the first three problems the RECTRAP mesh gives larger errors at all of the grid points, often by factors ranging from 3 to 4.

2.5 Grid Refinements

The next stage of the project consisted of testing the FEM model using larger grids covering the same domain. This involved refining each grid by adding more elements. Grid refinements should improve the accuracy of the results at the expense of increased computer storage and computations. Each grid may be pictured as an inner square ABCD covered by

a uniform rectangular grid containing p nodes per row and p nodes per column. This is surrounded by layers of rectangles or trapezoids separated by intermediate squares (EFGH in Figures 2 and 3). If there are q such squares, then there are q+1 layers. In the original grids, p = 5 and q = 1. Software was developed for testing larger grids of these types. The only data required for the program consists of the values of p and q and the distances h from the outer boundary to each of the intermediate squares. The initial mesh data for each grid is generated by the computer. Tests were conducted with each grid using p = 10 and q = 1, 2, and 3. The first problem, Equation (1) with $f = \sin(\pi x)$ and g = 0, was used for these models.

	p	q	h	
1.	5	1	.15	(original grids)
2.	10	1	.15	
3.	10	2	.12, .20	
4.	10	3	.06, .125, .19	

	Grid	Number of nodes & elements, half-bandwidth		
1.	rectangular	81	64	11
	RECTRAP	57	48	11
2.	rectangular	196	169	16
	RECTRAP	172	153	16
3.	rectangular	256	225	18
	RECTRAP	208	189	18
4.	rectangular	324	289	20
	RECTRAP	244	225	20

Again the errors over the rectangular grids are more uniform. The errors over the fine mesh nodes tend to be smaller for the RECTRAP grid than for the comparable rectangular grid. In fact, many errors at the more central nodes of the original RECTRAP grid (p = 5, q = 1) are no larger than the corresponding errors for the larger rectangular grid (p = 10, q = 2). However, the results for the intermediate nodes are generally more accurate for the rectangular grid than for the RECTRAP grid of the same refinement. The errors decrease with each refinement of the rectangular mesh (by factors of 1/2 from p=5, q=1 to p=10,

q=1; 2/3 from p=10, q=1 to p=10, q=2; and 5/6 from p=10, q=2 to p=10, q=3), the best improvement is observed with the first refinement. On the other hand, the results are not improved to a great extent when the RECTRAP mesh is refined by increasing p from 5 to 10 and keeping q fixed at 1. The errors are decreased when the RECTRAP mesh is refined by fixing p = 10 and increasing q. (In the center the errors are decreased by factors of 1/2, from q=1 to q=2, and 2/3, from q=2 to q=3.)

It should be noted that the choices of h were somewhat arbitrary, no attempt was made to find the optimal placements of the intermediate squares for the grid refinements.

2.6 Time Dependent Problems

During the final phase of the project a FEM model was developed for the initial value-boundary value problem in Equation (5) (the heat equation):

$$\begin{aligned} &u_{xx} + u_{yy} = u_t \\ &u = f \text{ on } G,\ t > 0, \\ &u = g \text{ for } t = 0. \end{aligned} \tag{5}$$

There are several methods for treating such problems; for example, see [1], [2], [6], and [7]. Some of the more

successful modeling techniques employ a combination of a finite element scheme for the spatial x,y-coordinates and a finite difference discretization of the time interval. The model developed for the project uses either of the simple grids (Figures 2 & 3) for the x,y-domain at each time step and a Crank-Nicholson finite difference scheme to move from one time step to the next. A uniform grid was used for the time interval [0,T],

$$0 = t_0 < t_1 < \ldots < t_n = T,$$

$\Delta t = t_j - t_{j-1}$ is the same for each j. The solution is assumed to be of the form

$$u(x,y,t) = \sum_i d_i(t)\phi_i(x,y),$$

where ϕ_i, i=1,2,...,n, are the finite element shape functions and $d_i(t)$ are the unknowns in the problem at each time step. For each time step, $t = t_j$, finite element techniques based on variational principles are used to produce a linear system of equations

$$A \cdot D + B \cdot \frac{dD}{dt} = 0.$$

Here A and B are system matrices containing terms involving the integrals of products of the shape functions or their derivatives. D and dD/dt are system vectors containing the

terms $d_i(t_j)$ $d_i'(t_j)$ respectively. To solve this equation the terms of D are replaced by

$$\frac{d_i(t_j) + d_i(t_{j-1})}{2}$$

and those of dD/dt by

$$\frac{d_i(t_j) - d_i(t_{j-1})}{\Delta t}.$$

The system can be rewritten in the form

$$S \cdot D = C, \text{ where}$$

$$S = \frac{1}{\Delta t} B + \frac{1}{2} A,$$

$$C = (\frac{1}{\Delta t} B - \frac{1}{2} A) \cdot D_1,$$

$$D = (d_i(t_j)), \text{ and } D_1 = (d_i(t_{j-1})), \ i=1,2,\ldots,n.$$

This system can be solved for D whenever D_1 is determined. (Observe D_1 at time step t_j equals D at time step t_{j-1}.) The initial condition, $u = g$, is used to determine D_1 at the first time step $t_0 = 0$. The essential boundary condition, $u = f$, is imposed for each calculation of A and B.

Tests were conducted using two sets of initial-boundary conditions in Equation (5):

1. $f = e^{-\pi^2 t}(\sin(\pi x) + \sin(\pi y)) + 2$

$g = \sin(\pi x) + \sin(\pi y) + 2,$

and

2. $f = 100e^{-\pi^2 t}(\sin(\pi x) + \sin(\pi y)) + 2$

$g = 100(\sin(\pi x) + \sin(\pi y)) + 2.$

Problems 1 and 2 were run using both grids with each of the following time intervals:

	t-interval	Δt	# of time steps
a.	[0,10]	1	10
b.	[0,5]	.1	50
c.	[0,1]	.01	100.

For comparison, the value of $\Delta x = \Delta y$ for the rectangular grid covering ABCD is .125. Since the Crank-Nicholson procedure is unconditionally stable for this problem, the approximate solution will (eventually) converge to the correct values as t increases for any value of Δt. However, smaller time steps will produce more accurate results at each time step. The rms of both the errors and the percent errors over the fine mesh nodes and over the intermediate nodes were printed for each time step.

The following table compares the the percent error rms for the first and last time steps from the tests with Problem 1 above.

grid	Δt	T	rms % errors ABCD nds.	int. nds.
rectangular	1	1	34.2	18.6
		10	5.32	3.92
RECTRAP	1	1	35.1	19.5
		10	5.62	4.22
rectangular	.1	.1	2.06	1.26
		5.0	.0007	.0004
RECTRAP	.1	.1	2.86	1.97
		5.0	.042	.009
rectangular	.01	.01	.0568	.0445
		1.00	.0001	.0001
RECTRAP	.01	.01	.363	.289
		1.00	.0002	.0001

In each run the errors are smaller over the intermediate nodes than the central nodes. There is little difference between the results from the rectangular grid and those from the RECTRAP grid when $\Delta t = 1$; however, the results produced by the rectangular grid are noticeably better when Δt is .1 and .01. In the first two cases the errors decrease with

each time step for both grids while in the last case ($\Delta t = .01$) the errors increase during the first 6 to 9 time steps and then decrease thereafter.

Generally the results from Problem 2 have a similar pattern, although the errors are larger and the convergence is slower. There is one difference, when $\Delta t = .01$ the errors exhibit oscillatory behavior at first before decreasing toward zero. As in the first problem, the rectangular grid produces more accurate results at each time step than the RECTRAP grid when Δt is .1 and .01.

2.7 Program Efficiency

The CPU times required for the calculations in most of the programs were recorded so that comparisons could be made. The times for computing the terms of the system arrays, assembling these, and solving the corresponding system of equations were included while the input-output times and the mesh initialization times were excluded. The times are listed below, those from the AT&T 3B15 first (in minutes:seconds) followed by the VAX 8650 (in seconds):

Programs using the initial grids (Figures 2 & 3)

1. Original problem, Equation (1):

RECTRAP grid:	1:35.27	0.09
rectangular grid:	2:04.78	0.14

2. Alternative integration techniques:

Gauss 4 point:	RECTRAP grid:	0:48.39	0.05
	rectangular grid:	1:03.69	0.09
Gauss 9 point:	RECTRAP grid:	1:35.08	0.09
	rectangular grid:	2:04.48	0.13
Gauss 16 point:	RECTRAP grid:	2:40.67	0.15
	rectangular grid:	3:30.62	0.20
product trapezoid:	RECTRAP grid:	0:40.08	0.05
	rectangular grid:	0:52.04	0.07
product Simpson:	RECTRAP grid:	1:22.41	0.09
	rectangular grid:	1:47.11	0.12

combined techniques; RECTRAP grid only:

Gauss 4 point, 9 point:	1:18.64	0.09
Gauss 4 point, 16 point:	2:02.61	0.13
Gauss 9 point, 16 point:	2:17.22	0.15
pr. trap., pr. Simp.:	1:07.86	0.08

3. Programs using skyline storage, Equation (1):

RECTRAP grid:	1:32.45	0.09
rectangular grid:	1:59.95	0.14

4. Alternative boundary conditions, Equation (4):

RECTRAP grid:	1:39.59	0.10
rectangular grid:	2:12.02	0.14

Programs using larger grids, Equation (1):

5. $p = 10$, $q = 1$

RECTRAP grid:	5:34.74	0.32
rectangular grid:	6:12.13	0.36

6. $p = 10$, $q = 2$

RECTRAP grid:	7:15.42	0.43
rectangular grid:	8.41.93	0.51

7. $p = 10$, $q = 3$

RECTRAP grid:	9:03.86	0.52
rectangular grid:	11:39.93	0.68

Time dependent problems, Equation (4); Figures 2 & 3 grids:

8. $0 \leq T \leq 10$, $\Delta t = 1.0$:

RECTRAP grid:	2:43.29	0.14
rectangular grid:	3:43.57	0.21

9. $0 \leq T \leq 5$, $\Delta t = 0.1$:

RECTRAP grid:	6:24.86	0.35
rectangular grid:	9:05.26	0.52

10. $0 \leq T \leq 1$, $\Delta t = 0.01$:

RECTRAP grid:	11:00.57	0.60
rectangular grid:	15:50.81	0.95

The ratio of the running time using the RECTRAP grid to that using the rectangular grid (trap:rect) is approximately .76 for each run using the simple grids in Figures 2 and 3. (These ratios tend to be smaller for the VAX.) There is a slight decrease of this ratio in the time dependent problems

as the number of time steps increases (trap:rect = .69 for the runs with 100 time steps). The ratio of the number of elements in these grids is .75. The running time ratios for the mesh refinements are also approximately the same as the corresponding element ratios (.90 when p=10, q=1; .83 when p=10, q=2; .78 when p=10, q=3).

3. CONCLUSIONS AND RECOMMENDATIONS

The comparative accuracy of the two grids appears to depend on the problem modeled. The errors are more uniform over the whole domain when the rectangular grid is used in the program. In many runs the RECTRAP mesh tends to give smaller errors than the rectangular mesh at the nodes close to the center and larger errors at the outer nodes. In every case, however, the RECTRAP grid requires less computation time for a given problem. It was observed that the comparative efficiency of the two types of grids can be measured by comparing the total number of elements in each.

The experiments with different placements of the intermediate square EFGH in Figures 2 and 3 indicate there is an optimal location of this square in the rectangular grid that minimizes the errors at nearly every node and that this is nearly the same with each problem tested. No such ideal location is observed for the RECTRAP configuration. Different placements produces optimal results at different nodes and these vary from problem to problem. This may be one advantage of the rectangular grid.

The alternative integration techniques produce little change in the results with the exception of the product trapezoidal rule. When this numerical scheme is used in the models, the errors are noticeably decreased for the

rectangular grid and increased for the RECTRAP grid. It also requires the least amount of CPU time with each grid. This might suggest that the product trapezoidal integration scheme is the best one to use for the rectangular grid, although comparisons of errors at points other than the nodes should be made first.

The skyline matrix storage procedure improves the computational times slightly without affecting the results. Tests using the natural boundary condition in Equation (2) produce similar results, although the rectangular grid tends to out perform the RECTRAP grid at more nodes than in the earlier runs. There is a slight increase in CPU times, too.

When larger grids are used in the FEM model, the errors are decreased and the CPU times are increased. Results from the rectangular grid model seem to improve the most when the grid over the inner square ABCD is refined. The addition of extra intermediate squares to the RECTRAP configuration tends to enhance this model's performance the most.

The experiments involving time dependent problems indicate that the rectangular grid produces better results and faster convergence over both the inner square nodes and the intermediate nodes. This may be explained by the fact that the larger errors which the RECTRAP mesh seems to produce at some of the intermediate nodes tend to propagate

over the rest of the domain at later times steps. Further tests should be conducted to see if smaller time steps are required for the RECTRAP grid in general time dependent problems. If this is the case, the efficiency gained in using the RECTRAP grid for the spatial variables would be lost for dynamic problems because of the increase in the number of time steps.

The results from this project have not demonstrated conclusively that the new RECTRAP grid is a viable alternative to the rectangular grid. Too many questions remain unanswered. However, further investigations can be conducted using the software developed for this work.

RECOMMENDATIONS

1. Errors at points other than the grid nodes should be compared. One suggestion is to develop software that would approximate the L^2-norm of the error.

2. Experiments using different placements of the intermediate square in Figures 2 and 3 should be conducted for the alternative boundary value problems in Equation (4) and the time dependent problems in Equation (5). These would be similar to the studies carried out for Equation (1).

3. The alternative integration techniques should be

investigated further. One question to be answered is whether the product trapezoidal rule is the best integration scheme for the rectangular grid in a wide variety of problems.

4. The effect of different placements of the intermediate squares in the larger refined grids should be studied systematically.

5. Models of additional time dependent problems, such as the wave equation, using both types of grids should be developed. These problems should include ones using alternative types of initial and boundary conditions.

6. Another popular type of two dimensional grid employs triangular elements. Either grid studied during this projects may be converted to such a grid without adding any nodes by subdividing each rectangle and each trapezoid into two triangles. Software should be developed for models using these grids for similar studies.

7. The basic configuration of the two grids in Figures 2 and 3 can be altered to give a better representation of the model grid pictured in Figure 1. The inner square ABCD should be reduced and offset so that it is no longer centered in the domain.

8. A model for a linked system of equations, such as the shallow water equations, should be developed for study.

BIBLIOGRAPHY

1. Akin, J. E., Application and Implementation of Finite Element Methods, New York, Academic Press (1982)

2. Burnett, David S., Finite Element Analysis, Reading, MA, Addison-Wesley Publishing Co. (1987)

3. Reddy, J. N., Applied Functional Analysis and Variational Methods in Engineering, New York, McGraw-Hill Book Co. (1986)

4. Staniforth, A. N. and H. L. Mitchell, "A semi-implicit finite element barotropic model," Mon. Wea. Rev., 105, pp.154-169 (1977)

5. Staniforth, A. N. and H. L. Mitchell, "A variable resolution finite element technique for regional forecasting with the primitive equations," Mon. Wea. Rev., 106, pp.439-447 (1978)

6. Strang, G. and G. J. Fix, An Analysis of the Finite Element Method, New York, Prentice-Hall (1973)

7. Wait, R. and A. R. Mitchell, Finite Element Analysis and Applications, New York, John Wiley and Sons (1985)

FINAL REPORT NUMBER 55
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Philip Olivier
760-6MG-032

FINAL REPORT NUMBER 56
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Parsottam Patel
760-6MG-131

FINAL REPORT NUMBER 57
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Robert Patsiga
760-6NG-065

FINAL REPORT NUMBER 58
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Jacqueline Paver
760-6MG-020

1986-87 MINI GRANT PROGRAM

Sponsored by the

Air Force Office of Scientific Research

Conducted by

Universal Energy Systems, Inc.

Final Report

Automatic Program Generation From Specifications Using Prolog

Prepared by:	Alex Pelin
Academic Rank:	Associate Professor
Department and	School of Computer Science
University:	Florida International University
Research Location:	Weapons Laboratory
	Kirtland AFB, NM.

Date:	October 20, 1987
Contract No.:	F49620-C-0013

ACKNOWLEDGEMENTS

I would like to thank the Air Force Systems Command and the Air Force Office of Scientific Research for the sponsorship of my research.

I conducted most of my research at the Weapons Laboratory during the summer of 1987. There I had the necessary resources to pursue interesting research topics. I would like to thank the chief of the Information Systems Technology Office, Dr. Nazareno Rapignani, for his help in providing resources necessary for my project. I would also like to thank Lt. Col. Ed. Oliver for his help and advice.

I am especially grateful to Captain Robert Millar and to Paul Morrow for helping me on a daily basis. Mr. Morrow wrote a set of programs that are an important part of the project.

I would also like to thank Dr. Roger Clements, Professor Barry McConnell, Diane Wood, Brian Kennedy and John Jordan for their help and friendship.

Finally, I would like to thank my mother, Profira, and my children, Dan and Lydia, for keeping me company during my stay in Albuquerque.

AUTOMATIC PROGRAM GENERATION FROM SPECIFICATIONS USING PROLOG

by

Alex Pelin

ABSTRACT

During the summer of 1986 Michael Slifker and I built a program which generated Prolog programs from input/output specifications. The system was able to generate correct programs for sorting lists. The system used abstract data types as models for data.

This summer I worked on extending the system by adding new abstract data types and developing new heuristics. I developed a generate and test program to e used in building the output program. Paul Morrow wrote an interface that allows the user to enter commands in English. More work needs to be done in this area since the system needs sharper heuristics and more predefined data types.

Considerable attention was given to the problem of validating the input specifications. This problem is difficult, but it is crucial in the program generation process. I wrote a PROLOG program that implemented the Knuth-Bendix completion procedure for a set of equations. I also used the interactive theorem prover ITP to validate data type specifications. Both methods are useful tools, but they require large amounts of internal memory and high run time. Data type validation can be made more efficient by using more heuristics and by employing conditional term rewriting systems.

Progress has been made in the study of the applications of the terms rewriting systems to automatic program generation. More work needs to be done in this area.

1. Introduction

During the summer of 1986 I had an USAF-UES faculty research fellowship at the Weapons Laboratory, Kirtland Air Force Base. I worked together with a graduate student, Michael Slifker, on designing and implementing an expert system which generates Prolog programs from specifications. The system was writen in Prolog and it was able to generate correct programs for simple sorting problems.

By its scope this project falls in the category of program synthesis. Program synthesis is the area of computer science which develops systems that generates programs from specifications. The two main objectives of this area are:

1. To develop formalisms which can describe programming tasks, and
2. To create systems which translate formal specifications into high level language code.

This area is considered part of Artificial Intelligence (AI) since it uses AI techniques to generate software.

Our program synthesizer took the folloving specifications as input:

1. A description of the input data type;
2. A description of the output data type;
3. A description of the input/output relation;
4. A set of tests;
5. A set of transformations.

Figure 1

Based on these specifications, the program generator created a program which took as input an item, i, belonging to the input data type and produced

an item, o, belonging to the output data type. The pair i,o satisfied the input/output relation. The synthesizer used only those transformations and tests specified in the input set.

In one example the user specified the input data type to be list of whole numbers and the output data type to be list of whole numbers sorted in increasing order. The input/output relation was that the output must be obtained from the input list by switching the contents of selected consecutive pairs of list elements. The set of tests consists of checking if a pair of consecutive list elements is in the right order. The synthesizer produced a correct Prolog program for this problem.

The system used the language of abstract data types developed by the ADJ group '[7]) for specifying the input and the output data. In this approach data items are seen as terms which are freely generated by a set of operators ([2], [7]). The semantics of the data type is given by a set of equations (conditional equations) defined on this set of terms.

The transformations are part of the data and are also described by equations. The tests are part of the control structure of the generated program . They are functions on strings. As such they do not preserve equations. The difference between transformations and tests is illustrated by the following example.

The inversion of two consecutive elements in a list can be described by the equation :

$$INV([m|[n|List]]) = [n|[m|List]] \qquad (2)$$

Since INV is part of the data then we must add the equation :

$$INV(Anylist) = Anylist \qquad (3)$$

Equations 2 and 3 yield the equality [m|[n|List]] = [n|[m|List]], where m and n are whole numbers and List is a list of whole numbers. A particular

instance of the last equality is [1,2] = [2,1] . The notation [m,n] is used to denote the list [m|[n|[]]]. The test (predicate) SORTED, which checks if a list of whole numbers is sorted in increasing order, is given by the system of equations shown below.

1. SORTED ([])
2. SORTED ([m])
3. m <= n ---> SORTED ([m|[n|List]]) == SORTED ([n|List])
4. n < m ---> - SORTED ([m|[n|List]])

Figure 4

In figure 4 [] stands for the empty list, [m] is a single element list containing the whole number m, m and n are whole numbers and List is a list of whole numbers. The sign - denotes negation. The symbol == stands for equality of tests. If x == y then this equality of x and y is obtained strictly from the equations that define the data types (without the equations that define transformations) and the equations that define the predicates (tests). In the case of the previous example the relation [1,2] = [2,1] holds but the relation SORTED ([1,2]) == SORTED ([2,1]) does not. This is so since SORTED ([1,2]) is true while SORTED ([2,1]) is false.

During the Fall of 1986 I applied for a minigrant to extend this system in the following directions :

1. Expand the knowledge base of the system to include more data types;
2. Augment the system by adding term rewriting systems;
3. Add an interface that allows the user to describe programs in English;
4. Include more heuristics to handle the set of transformations.

Figure 5

The request for the minigrant was approved and I started working on

the project in January 1987. From May 12 1987 until September 7 1987 , I was in Albuquerque. During that time, I worked full time on the project. I used the Sun 3/160 from the Weapons Laboratory for implementing several programs. The programs were run using the PROLOG interpreter POPLOG, with the exception of the user interface, which was writen by Paul Morrow in Turbo Prolog. The programs were tested with several data sets. Copies of the programs and of the results of those tests can be found in the appendix. I also used the interactive theorem prover ITP to validate data type specifications.

Progress was made in developing a theoretical basis for the concepts employed in the project. Papers [16] and [17] describe these advances. Papers [16] and [17] deal with the AI aspects of the system.

The report is organized as follows. Sections 2-5 deal, respectively, with the directions 1-4 listed in figure 5. Section 6 contains recommendations and section 7 has the bibliography.

2. Adding New Data Types

One objective of the project was to create a library of predefined data types. In the summer of 1986, the data types boolean, whole number and list of whole numbers were created. Last summer I implemented the data type array. In [2] one can find sets of equations that define the data types : integer, rational, queue and set. There is no problem in implementing these types, since they are defined in a structured way.

Last summer the main concern was the validation problem for abstract data types. For example, data type list of whole numbers was defined as having three sorts (types): boolean, natural and list. Each type is defined as a set of strings that must satisfy a set of equations. The sets of strings are generated by the following production rules :

```
natural ---> 0
natural ---> successor(natural)
list    ---> []
list    ---> [natural | list]
boolean ---> True
boolean ---> False
boolean ---> natural <= natural
```

Figure 6

Figure 6 states that : the whole numbers are either 0 or the successor of a whole (natural) number and the lists are either the empty list or are obtained by inserting a natural number in front of a list. The boolean values are the constants True and False and expressions of the type natural <= natural . The semantics of the data type is given by the set of equations shown in figure 7.

1. n <= n = True
2. n <= m = True ---> successor(m) <= n = False
3. n <= m = True ---> n <= successor(m) = True

Figure 7

Figure 7 states that n <= n is true , that n <= m true implies that the relation successor(m) <= n is false and that n <= m true implies that the relation <= holds between n and successor(m). At this point, one may want to check that the relation <= is well defined. This means that for all natural numbers m and n the foloving statements are true :

1. The term n <= m evaluates to either True or False, and
2. The boolean constants True and False are not equal in the system of equations that define the data type.

The above problem is a data validation question.

Data type validation is the problem of deciding if a certain property

follows from the set of axioms that define the data type. As such, it is an automatic theorem proving problem. There are three methods that can be used to validate abstract data type specifications : first order logic, induction and term rewriting systems.

The progress in automating induction has been slow ([4],[6],[14]). The area is crucial to automated theorem proving since many abstract data types are inductively defined. More research needs to be done in this area.

Last summer I used first order logic and term rewriting systems as data type validation tools. The role of term rewriting systems in this project is discussed in the next section.

The interactive theorem prover ITP ([15]) was used to validate first order specifications for data type array. This theorem prover uses the resolution method ([22]) to prove theorems. It can operate either in the batch mode, or in the interactive mode. In the interactive mode the human picks the clauses that are used to generate new clauses. In the batch mode the system makes this decision. The system uses three lists : an axiom list, a set of support list and a demodulator list. This theorem prover works as follows. It takes a clause from the set of support, called the 'given clause' , and uses it to generate new clauses by applying demodulators (simplifications) from the demodulator list, and by unifying it with the clauses in the set of axioms. The new clauses are added to the set of support. Since the system proves theorems by refutation, the user must choose the three lists in such a way that the set of unsatisfiable clauses is not included in the set of axioms. The user must also pick the inference rules that are used (like hyperresolution, UR-Resolution, binary resolution,unit resolution, factoring, paramodulation) to obtain new clauses. As a general rule, the faster inference rules do not guarantee refutation completeness ,i. e. they do not always reach a

contradiction when the set of clauses is unsatisfiable. The main problem with resolution based theorem provers is that they generate an enormous number of useless clauses. In the interactive mode, the user can delete clauses, move clauses from one list to another and choose the 'given' clause.

The PROLOG definition of the data type array is shown on page 1 of the appendix. The translation of these PROLOG clauses into first order clauses is shown on page 2 of the appendix. The data type array(type) is a parameterized data type ([2]). This means that the type of the array is a parameter to the program (equations) that define the data type array. The elements of array(type) are terms of the form pair(x,v), where x is a natural number, called the subscript, and v is an item of sort type.

The program writen on page 2 contains ITP clauses. In ITP variable names start with letters u through z. Names starting with capital letters or with the small letters a-t denote constant predicates or constant functions. The sign '|' stands for disjunction (logical or).

The predicate ARRAY was defined as an expression containing the predicates RELATION, SET, FUNCTIONAL and CONSECUTIVE. ARRAY (x1,x2,x3) means that x3 is an array of type x1 and size x2. The definitions shown on page 2 of the appendix state that the relation ARRAY(x1,x2,x3) holds if and only if the predicates RELATION (x3), FUNCTIONAL (x3), SET(NAT,domain(x3)), SET(x1,codomain(x3)) and CONSECUTIVE(NAT,ZERO,x2,domain(x3)) hold. RELATION(x3) holds if x3 is a set of ordered pairs, SET (y1,y2) is true if all elements in the set y2 are of type y1, and FUNCTIONAL (x3) is true if x3 does not contain two elements, pair(i,v1) and pair(i,v2), with v1=v2. The functions domain and codomain are the domain, respectively the codomain of a relation. CONSECUTIVE(NAT,ZERO,x2,domain(x3)) requires from the set of subscripts (the domain of x3) to contain all natural numbers between ZERO and x2 (the size of the array) . The operation change(A, I, relval(I,A), J, relval(J,A)) switches

the contents of locations I and J in the array A, when I and J are valid subscripts, and leaves the array unchanged when either subscript is out of bound. The function relval(I,A) is the value of the I-th location of the array A.

I tried to prove that change (A, I, relval(I,A), J, relval(J,A)) is an array of the same size as A. I was not successful. The ITP program ran for more than 3 hours and it generated 2,000 clauses but it could not find a contradiction. I repeated the procedure several times by using different partitions of the set of clauses into the three lists (axioms, set of support and demodulators) but the result was the same. I proved, using ITP ,that B is a relation. Even for this proof ITP generated several hundred clauses in the batch mode. The reason for such poor performance is the following. In order to prove that a clause C is a consequence of a set of clauses S, the clause C is negated and the theorem prover is used to derive a contradiction. Since C is a consequence of a subset T of S, the clauses in S - T are useless. At each step ITP chooses the 'given' clause from the set of support . If the 'given clause' is chosen at random, the ratio between the number of clauses that are consequences of the subset T alone and the number of all clauses decreases exponentially in the number of steps. Since many of the clauses generated by the set T alone are also useless in proving C, the complexity of this process is worse than than exponential. For this reason, it is worth to develop algorithms that eliminate useless axioms. The problem is important, since an inexperienced user may keep adding axioms hoping that he will get the desired properties and thinking that redundant equations do no harm. A structured definition of the data type will certainly help. I found ITP to be useful when the user knows how to weed out useless clauses and it was of no use when it was employed in the batch mode. I proved a simple theorem in the axiomatic set

theory calculus in less than 30 steps while ITP, operating in the batch mode, generated more than 600 clauses before reaching a contradiction.

3. Adding Term Rewiting Systems

The purpose of using term rewriting system was two-fold. They can be used for validating data type definitions and they can be employed in the program generation process. Term rewriting systems are sets of rules of the type term1 ---> term2. A term T is simplified by a rule term1 ---> term2, if some instance of term1 in T is replaced by the corresponding instance of term2. In this case the term T called reducible. Otherwise, T is called a normal form. If T' is the result of replacing, in T, the instance of term1 by the corresponding instance of term2 then it is said that T reduces to T'.

Of particular interest are term rewriting systems which are termineting and confluent ([10]). A term rewriting system is terminating if there is no infinite chain of terms T0,T1,T2,...,Tn,... such that T0 reduces to T1, T1 reduces to T2 and so on. A term rewriting system is confluent, if the order in which the rules are applied is irrelevant to the final outcome. If a term rewriting system is both terminating and confluent, then every term has a unique normal form.

Knuth and Bendix ([13]) give an algorithm for finding a confluent and terminating term rewriting system for a set of equations E over a free algebra. The data types used in this project are defined precisely in this manner. If the algorithm finishes with success, then it outputs a set of term rewriting rules R which has the following property: an equation t = t' follows from the set of equations E if and only if the normal forms of t and t', under the system of term rewriting rules R, are identical. The Knuth-Bendix algorithm constructs the set of term rewriting systems by using a predefined ordering on the set of terms.

I implemented the Knuth-Bendix algorithm in PROLOG. It can be found on pages 3-10 of the appendix. The algorithm starts with a set of equations E, a set of (term rewriting) rules R, a maximum number of steps Count and a list of sorts (types) Sort. At each step the algorithm picks an equation from E and attempts transform it into a rule. An equation l = r is transformed into the rewrite rule l ---> r if l is greater than r, and it is transformed into the rule r ---> l if the term r is greater than l. If neither one of these cases applies then the algorithm finishes with failure. The new rule is then simplified by the set of rules in R. If it is not a result of the rules already in R, it is added to R. Otherwise, a new equation is picked and the loop is repeated. The rules in R are checked to see if the new rule simplifies them. The rules that are simplified by the new rule are removed from R and added to the set of equations E. The new rule is matched against the rules that are still left in R in order to form critical pairs. The process of forming critical pairs is similar to the unification process in resolution ([1]). A critical pair is formed when a subterm of the left hand side of a rule l1 ---> r1 can be unified with an instance of the left hand side of another rule l2 ---> r2. For example, the rules (x + y) + z ---> x + (y + z) and x + 0 ---> x form a critical pair since the subterm x + y of the lefthand side of the first rule can be unified with the lefthand side of the second rule. This way, the term (x + 0) + z can be reduced to x + (0 + y) by using the first rule and it can be simplified to x + z by the second rule. The pair x + (0 + z) , x + z is called a critical pair. The critical pairs obtained this way are added to the set of equations E and the loop is repeated. The process finishes with success when the set of equations E becomes empty. It finishes with failure when it cannot transform an equation into a rewrite rule and it stops with a message of 'incomplete' when neither one of these cases occur in the

first Count steps. The algorithm is correct.

The orderings shown on pages 11 and 12 of the appendix were used to run several examples. The first example is shown on pages 13-14. On top,it is shown the initial set of equations. After 30 steps the algorithm printed 'Not enough time' and the contents of the set of equations and of the set of rules at that point. The example shown on page 15 is a case of termination with failure. No ordering was given for the terms in the equation shown at the top of the page. On pages 16,17 and 18 we have three examples of termination with success.

There are some problems with using PROLOG for the Knuth-Bendix algorithm. PROLOG has the advantage that it has a built-in unification algorithm, but it is not fast in doing arithmetic. Since the equations in E must be kept sorted in such a way that the simpler axioms will be eliminated first this creates a problem. Another problem is the built-in backtracking mechanism in PROLOG. Because of it, the algorithm ran out of internal memory on more than one occasion.

4. Adding an User Interface

The system should be used by two classes of users. The first class of users define their own data types, transformations and tests. The second class of users employs the data types, the tests and the transformations that are already in the library. In order to facilitate the use of the system by the second category of users, it has an interface which allovs them to enter commands in English. This program was vriten in Turbo Prolog and it can be found on pages 21-27 of the appendix. A BNF representation of the input grammar can be found on page 19. Examples of input terms and their translations into data type declarations are shown on page 20. The user may enter commands like:

1. A is a real array;
2. B is A reversed;

3. C is A sorted;

4. 10 is the size of A;

5. N is an integer list;

6. R is a heap of W.

Figure 8

The commands shown in figure 8 allow the user to employ the system in the same way that he/she uses a calculator. A more sophisticated interface can be built that will allow the user to receive programs as output. The description of this interface can be found in section 6.

5. Adding New Heuristics

I wrote a PROLOG program ,named loop, which takes as input a term T and a maximum number of steps called Steps. The program loop uses a set of clauses of the form rule(Rule_Number, Condition, Left, Right). They represent conditional rewrite rules Rule_Number : (Condition) ===> Left ---> Right. These conditional rules are more general than the rules described in section 3 because they contain the condition (Condition). A term T reduces to a term T' by using the conditional rule (C) ===> l ---> r if there an instance I of l in T, the condition C is true for the instance I and T' is obtained from T by replacing I by the corresponding instance of r. For example, let (n < m) ===> [m|[n|list]] ---> [n|[m|list]] be the rule and T be the list [4,5,2,3]. The rule can be applied to the subterm [5,2,3] of T. The sublist [5,2,3] is the instance of l = [m|[n|list]] in T. The condition C = (n < m) is true for I since m=5 and n=2. The term T' is obtained from T by replacing the instance I=[5,2,3] in T by the corresponding instance [2,5,3] of the righthand side of the conditional equation. The term T' is the list [4,2,5,3].

The procedure loop also uses a test,called predicate, to describe the goal. The program checks if there is a sequence of conditional rules, called

Transformation, which carries Term into a term Term1 such that predicate(Term1) is true. As the name suggests, loop repeats the procedure loop_body Steps number of times, or until either the success condition or the empty_list condition are met. If the success clause is true then Transformation and Term1 are the output variables. The procedure loop keeps a list of active states sorted according to the predicate less. Each state is of the form pair(List_of _transformations,Term2). The meaning of this representation is that the sequence List_of_transformations carries Term to Term2. The list of active states, called active_list, is sorted according to the ordering on the set of terms. There is also a list of dead terms, called dead_terms_list which is also sorted according to the predicate less. At each step the loop_body picks a state from active_list, called State, to be the expansion state. If the active list is empty then it raises the empty_list flag. Since the list of active states is sorted, it always picks the head of the list. If the expansion state pair(Transformations,Term3) satisfies the goal predicate(Term3) then the flag success is set to true. Otherwise, loop_body generates all children of the expansion state by applying every rule at every address. If the child is not in the dead_terms_list then it is entered in the active list in such a way that the active list is always sorted according to the predicate less.

The program loop can be found on pages 28-32 of the appendix. It was run using the definitions of rule, predicate and less shown on page 32 of the appendix. The two rules are inversion of a pair of consecutive elements in a list and reduction of a duplicate item in the list when the identical items occur next to each other . The goal,described by the clauses named predicate, is to sort the list in increasing order and eliminate all duplicate elements. The program was run with the input list [8,5,3,1,4,5,2,3] two times. In the first run the variable Steps was set to 10 and in the second run it was set to 20. The list of active states was printed when the program exited the loop. In

the first case the program did not finish and it printed the message 'Not enough time'. The list of active states can be found on pages 33-45 of the appendix. In the second case the program finished after 18 steps and it produced the output [1,2,3,4,5,8]. The list of active states for this run can be found on pages 47-70 of the appendix.

Each state is identified by a sequence of transformations (reductions) that brings the original term to the term that characterizes that state. The reductions are displayed by the rule number and the address at which the rule is applied . The list [Head | Tail] is represented in our example by the operation a (Head,Tail). The string 'empty' stands for [].

This program is useful for finding a sequence of transformations that satisfies a given predicate. Since the size of the search space grows faster than exponentially in the size of the input list, it is essential for the system to have good guiding functions (heuristics).

6. Recommendations

Term rewriting systems are going to play an important role both in the the data validation process and in the program generation phase. For data type validation it is important to find terminating and confluent sets of reductions for certain classes of equations, especially for classes of conditional equations. Term rewriting systems are used as an alternative to resolution based theorem provers ([9]). They are also used as an important link between the equational specifications of the data types and their translation into PROLOG programs. The relation between first order logic and PROLOG is complicated ([8]). There has been intense activity in the area of conditional term rewriting systems ([5],[11],[12], [20]).

Equally important is the task of developing heuristics for guiding the search for a solution. The performance of the program described in the previous

section depends on the heuristics employed. It is imperative to develop heuristics for various types of specifications. This project dealt with the case when the output data type was a subset of the input data type. For this problem, the program generator focused on the predicate that defined that subset. Even for this case there is much work to be done. The system can be extended to include the data types : heaps,binary trees, and tries and such transformations as tree traversals, forming search trees and graphs. This may lead to a more general definition of abstract data type than the initial algebra model that was used in the project. The problem of defining a specification language is a difficult one ([8], [21]).

It is also important to have a user interface that allows English commands and generates programs. For example the user may enter the 5 specications from figure 9 as follows:

1. The input sort is real array;
2. The output sort is real array sorted in increasing order;
3. The input/output relation is standard;
4. The transformation is inversion of consecutive pairs;
5. The test is checking any pairs.

Figure 9

From such specifications the user must obtain a sorting program. He does not have to know the data type definitions anymore than the user of the calculator must know number theory. This task is the easiest of the three.

It must also be mentioned that the complexity of program generation is high. Theorem proving, searching state spaces are problems that use large amounts of time and memory. I believe that it would be beneficial to use parallel architectures to solve these problems. Many algorithms, like the Knuth-Bendix completion procedure, can be made to run in parallel.

7. Bibliography

1. Buchberger, B. : Basic Features and Development of the Critical-Pair Completion Procedure, in Rewriting Techniques and Applications, Springer Verlag Lecture Notes in Computer Science, Vol. 202, 1985, pp. 1-45.
2. Ehrig, H. and Mahr, B. : Fundamentals of Algebraic Specification 1, Springer Verlag, 1985.
3. Frenkel, K. : Towards Automating the Software-Development Cycle, CACM, Vol. 28, No. 6, June, 1985, pp. 578-591.
4. Fribourg, L. : A Strong Restriction of the Inductive Completion Procedure, to appear in Lecture Notes in Computer Science, Springer Verlag, 1988.
5. Goebel, R. and Gramlich, B. : TRSPEC User Manual, Universitat Kaiserlautern, 1986.
6. Goebel, R. : Completion of Globally Finite Term Rewriting Systems for Inductive Proofs, SEKI-Report SR-86-06, Universitat Kaiserlautern, June, 1986.
7. Goguen, J. , Thatcher, J. , and Wagner, E. : An Initial Algebra Approach to the Specification, Correctness, and Implementation of Abstract Data Types, in Current Trends in Programmming Methodology, Vol. 4, Edited by R. Yeh, Prentice Hall, 1978, pp. 80-149.
8. Hoare, C. : An Overview of Some Formal Methods for Program Design, Computer, Vol. 20, No. 9, September , 1987, pp. 85-91.
9. Hsiang, J. : Topics in Automated Theorem Proving and Program Generation, Doctoral dissertation, University of Illinois at Urbana-Champain, 1983.
10. Huet, G. : Confluent Reductions : Abstract Properties and Applications

to term Rewriting Systems, JACM, 1980, pp. 767-821.

11. Hussmann, H. : Rapid Prototyping for Algebraic Specifications - RAP-System User's Manual, Universitat Passau, March 1985.
12. Kaplan, S. : Simplifying Conditional Term Rewriting Systems: Unification, Termination and Confluence, L. R. I. Technical Report No. 316, December 1986.
13. Knuth, D. and Bendix, P. : Simple Word Problems in Universal Algebras, in Computational Problems in Abstract Algebra, Edited by J. Leech, Pergamon Press, 1970, pp. 263-270.
14. Lazrek, A. and Lescanne, P. : Proving Inductive Equalities : Algorithms and Implementation, to appear in Lecture Notes in Computer Science, Springer Verlag, 1988.
15. Lusk, E. and Overbeek, R. : The Automated Reasoning System ITP, Argonne National Laboratory Report 84-27, April, 1984.
16. Pelin, A. : Applications of Conditional Term Rewriting Systems to Automatic Program Generation, to appear in Lecture Notes in Computer Science, Springer Verlag, 1988.
17. Pelin, A. and Morrow, P. : Automatic Program Generation from Specifications Using Prolog, to appear in the proceedings of the Third Conference on Artificial Intelligence for Space Applications, November , 1987.
18. Pelin, A. : Computing with Conditional Rules, to be submitted to the Journal of Symbolic Computation in December 1987.
19. Pelin, A. : The Use of Meta-Reductions in Computing Normal Forms, to be submitted to the 15th Colloquium on Automata, Languages and Programming in November 1987.
20. Pelin, A. and Gallier, J. : Building Exact Computation Sequences,

Theoretical Computer Science, October 1987.

21. Roman, G. : A Taxonomy of Current Issues in Requirements Engineering, Computer, Vol. 18, No. 4, April, 1985, pp. 14-22.

22. Wos, L. , Overbeek, R. , Lusk, E. and Boyle, J. : Automated Reasoning: Introduction and Applications, Prentice-Hall, 1984.

Appendices can be obtained from

Universal Energy Systems, Inc.

1986 Mini Grant Program
Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432

Final Report

Some Novel Aspects of Organic Electrochemistry in Room Temperature Molten Salts

Prepared by: Bernard J. Piersma
Academic Rank: Professor of Chemistry
Department and Chemistry Department
University Houghton College

Research Location: Houghton College
Houghton, New York 14744

USAF Researcher: Dr. John Wilkes
The F.J. Seiler Research Laboratory
FJSRL/NC, USAFA
Colorado Springs, CO 80840-6528

Date: December 17, 1987

Some Novel Aspects of Organic Electrochemistry in Room Temperature Molten Salts

by

Bernard J. Piersma

Abstract

Carbonium ions formed by reaction of several alkyl chlorides and an acyl chloride with 1-methyl-3-ethylimidazolium chloride-aluminum chloride melts were detected electrochemically using cyclic voltammetry. 1-Cl-2-methylpropane, 2-Cl-2-methylpropane and butyryl chloride form carbonium ions in neutral (equal mole fractions of MeEtImCl and $AlCl_3$). 2-Cl-propane, 1-Cl-butane and 2-Cl-butane are electrochemically inactive in neutral melt but form carbonium ions when the melt is made acidic (excess $AlCl_3$) and once formed the carbonium ions remain stable when the melt is made neutral. Addition of benzene to melts with carbonium ions produced distinct changes in the CV curves. Detailed studies of butyryl chloride - benzene mixtures as a function of melt acidity demonstrated the complexity of interactions with the melt and the surprising sensitivity of electrochemical behavior to melt composition.

Studies are continuing with the seven stable isomers of chloropentane.

I. Introduction

After receiving a B.A. in chemistry from Colgate University, I studied physical chemistry at St. Lawrence University for two years, receiving an M.S. and did thesis work on chemical kinetics and mechanism studies in nonaqueous solutions. My professional training in electrochemistry was received at the University of Pennsylvania where I received the PhD degree in 1965. My thesis work with Professor J. O'M. Bockris was an investigation of the electrode kinetics and mechanisms of anodic oxidation of hydrocarbons in aqueous solutions primarily using steady-state potentiostatic techniques. Graduate courses in electrode kinetics and the electrical double layer which I took from Prof. Bockris became the basis of his two volume work with A. N. Reddy, Modern Electrochemistry.

In order to learn about transient and perturbation techniques for studying electrochemical process and the structure of the double layer, I worked for about two years after my graduation from Penn. with Sigmund Schuldiner at the U.S. Naval Research Laboratory in Washington as a NAS-NRC Postdoctoral Resident Research Associate. For the last 17 years I have taught at a liberal arts undergraduate college in a sponsored half-time teaching, half-time research position with research

interests in the stability of materials under physiological conditions and the electrochemical processes that occur at cardiac pacemaker electrodes.

For the 1981-82 school year, I was on sabbatical and received a grant from the USAFOSR as a University Resident Research Associate at the Frank J. Seiler Research Laboratory to study electrochemical processes in a recently developed room-temperature molten salt system. One major technical report resulted from the effort, Electrochemical Survey of Selected Cations and Electrode Materials in Dialkylimidazolium Chloroaluminate Melts.(1) To follow up on this work, I received a SFRP appointment for the summer of 1986 at the Frank J. Seiler Research Laboratory and again worked with Dr. John Wilkes. The Mini-Grant received to continue the work initiated during the SFRP appointment requested funding to purchase and install a Vacuum/Atmosphere Company dry box that would provide the inert conditions essential for acceptable studies with the molten salt systems.

II. Objectives of the Research Effort

The specific objectives are listed below as they were developed in the Mini-Grant proposal:

1. To purchase and install a controlled environment system that would maintain the inert atmosphere essential for acceptable work with room temperature molten salts.

2. To continue the electrochemical study of carbonium ion formation in MeEtImCl-$AlCl_3$ melts (Methyl Ethyl Imidazolium Chloride) using cyclic voltammetry with particular emphasis on the role of melt acidity for carbonium ion formation and stability.

3. To continue work in electrochemically detecting intermediates, e.g., "sigma complexes" in the alkylation and acylation mechanisms of Friedel-Crafts type reactions.

4. To explore the scope of organic reactions that can be studied in the MeEtImCl-$AlCl_3$ melts.

III. Purchase and Installation of Dry Box

A Vacuum/Atmosphere Company controlled environment system consisting of the following components was purchased and installed.

Model HE-243-2	Dri-Lab Glove Box
Model Mo40-1-H	Dri-Train Inert Gas Purifier

Model CVP-1	Valves and Plumbing
Model ST-110	Safe Trol
Model AV-1	Auto-Vac
Model RGF-1	Regeneration gas flow control
Model HECS	Support Stand

In addition gloves, filters, glassware for preparation and handling of materials for melt preparation, gas regulators and compressed gases and chemical reagents for melt preparation and for the proposed studies were purchased. The total expenditure for the project was approximately $28,000; with $8,000 being provided as matching funds by Houghton College. An additional purchase for this project was a Houston Model 2000 Omnigraphic X-Y recorder at a cost of approximately $3000, also provided by Houghton College. The target date for having the system fully operational was March 1987. With numerous delays that were not anticipated, the system was not actually in use until about June 1, 1987. Additional significant delays have been encountered in the preparation of melt components of acceptable purity.

The system is presently performing very well and the dry box atmosphere is maintaining a light bulb (with a hole drilled through the glass envelop to expose the filament) for 35-40 days.

IV. Electrochemical Investigation of Carbonium Ions Formed in MeEtImCl-$AlCl_3$ Melts

The Friedel-Crafts reaction using anhydrous aluminum chloride as catalyst has been the most important method for attaching alkyl side chains to aromatic rings.(2) The role of $AlCl_3$ is to abstract the halogen from the alkyl halide generating a carbonium ion, which then acts as an electrophile attacking the aromatic ring. Alternatively a "sigma-complex" may form as an intermediate with the alkyl group being transferred in a single step from the halogen to the aromatic ring. In a recent publication(3) a mechanism involving the formation of a "sigma-complex" as the rate limiting step was proposed for the acylation of benzene in acidic (0.60 and 0.67) MeEtImCl-$AlCl_3$ melts. That work found no acylation reactions in neutral (0.50) melt and the initial reaction rates for substitution increased as the melt acidity increased.

The Lewis acid-base properties of the melt are described by

$$2AlCl_4^- \rightleftharpoons Al_2Cl_7^- + Cl^-$$

with an equilibrium constant on the order of 10^{-17} in the MeEtImCl-$AlCl_3$ melt. In neutral melt (0.50 mole fraction $AlCl_3$) the concentration of $Al_2Cl_7^-$ is negligable, but it increases by addition of $AlCl_3$.

This led to the suggestion that $Al_2Cl_7^-$ was the catalyst responsible for promoting the acylation reaction.(3)

In earlier work Koch, et.al.(4) were successful in electroinitiating Friedel-Crafts transalkylations with hexamethylbenzene in acidic (0.67) ethylpyridinium bromide - $AlCl_3$ melt. Luer and Bartak(5) studied triphenylchloromethane with acidic n-butylpyridinium chloride - aluminum chloride melts. They proposed the following for the formation and cathodic reduction of the triphenyl methyl carbonium ion:

$$Ph_3CCl + Al_2Cl_7^- \rightleftharpoons Ph_3C^+ + 2AlCl_4^-$$

$$Ph_3C^+ + e^- \rightleftharpoons Ph_3C^\cdot$$

$$Ph_3C^\cdot \xrightarrow{\text{slow}} \text{products}$$

The radical dimerizes to produce an oxidizable form of 1-(diphenylmethylene)-4-(triphenylmethyl)-2,5-cyclohexadiene which also isomerizes to [4-(diphenylmethyl) phenyl] triphenylmethane.

In this study we were interested in determing whether carbonium ions could be formed and Fiedel-Crafts type reactions could be carried out in neutral melts. We chose several simple chloroalkanes to look for differences in stability and formation rates of carbonium ions that could be detected electrochemically. Finally, we wanted to further expose the alkylation and acylation mechanisms, particularly in neutral melt, and

attempt to detect "sigma-complexes" or other intermediates electrochemically.

The 1-methyl-3-ethylimidazolium chloroaluminate melts were prepared in the helium-filled Vacuum Atmospheres Corp. glove box maintained at <10 ppm combined water and oxygen as previously described.(6) The organic compounds were dried and then used without further purification. The experimental procedure followed was identical to that reported to UES in the Final Report for my Summer 1986 SFRP appointment. Exactly neutral (0.50 mole fraction of MeEtImCl and of $AlCl_3$) melts were prepared by adjusting the acidity (by addition of solid MeEtImCl or solid $AlCl_3$) until a maximum electrochemical window of about 4.6 volts was obtained. (The electrochemical window, defined as the potential range in which essentially no oxidation or reduction of the melt occurs, was determined at a glassy carbon electrode at a sweep rate of 100 mv/sec.) The studies previously begun during my SFRP appointment were reexamined and extended. The chloroalkanes studied included 2-Cl-propane, 1-Cl-butane, 2-Cl-butane, 1-Cl-2-methylpropane and 2-Cl-2-methylpropane. An acyl chloride, butyryl chloride was added in this study.

Representative cyclic voltammetric curves are illustrated in Figures 1-3 for the alkyl chlorides, Figure 4 for butyryl chloride and Figures 5-9 for mixtures of butyryl chloride and benzene in MeEtImCl-$AlCl_3$ melts. The measurements were made using a P.A.R. Model 173 Potentiostat with a Model 175 Universal Programmer and the CV curves were recorded with the Houston model 2000 Omnigraphic X-Y recorder. In general about 150 mg of organic reactant were added to about 16 g of exactly neutral melt yielding a concentration of about 150 mM in the organic species. After CV responses had been obtained in neutral melt, the solutions were made very slightly acidic by addition of solid $AlCl_3$ and the CV response was again recorded. When the CV behavior indicated no formation of a reducible species, i.e., carbonium ion formation, in a neutral melt, but formation in an acidic melt, the acidic melt was again made neutral by addition of solid MeEtImCl to determine whether carbonium ions formed would remain stable in neutral melt.

A summary of the CV data with additional observations is presented in Tables I and II. The results supported the tenative conclusions made in the 1986 SFRP Final Report. Carbonium ions are formed from 1-Cl-2-methylpropane and 2-Cl-2-methylpropane

in neutral melt but not from 2-Cl-propane, 1-Cl-butane or 2-Cl-butane. Product analyses still need to be completed to account for the additional reduction and oxidation peaks observed. Additions of benzene to a neutral melt containing 2-Cl-2-methylpropane resulted in decrease and finally disappearance of two reduction peaks, suggesting that benzene did react with carbonium ions formed. Comparisons of Figs. 5-9 with Fig 4 for butyryl chloride and butyryl chloride-benzene mixtures for a range of melt acidities shows the complexity and surprising sensitivity of electrochemical behavior in this system. A major reduction peak at about -0.57 V observed in slightly acidic melts in the butyryl chloride - benzene solution splits as the melt acidity is increased and separates as the acidity is further increased. The details of this behavior are summarized in Table II. We are continuing these studies to clarify this very interesting behavior.

V. Recommendations

Because of delays in obtaining the dry box and in getting the system installed, the research accomplished is somewhat behind our projected schedule. We are in a position now to continue the project with our focused effort from January to August 1988. I have been granted a sabbatical for the Spring 1988 semester and anticipate spending the major portion of this time on this project. We have encountered some difficulties in preparing high quality MeEtImCl and expect that conditions for this preparation (i.e., lower relative humidity) will be much more favorable in January and February. We plan to prepare a quantity of MeEtImCl and $AlCl_3$ early in 1988 to supply the program through the next year. The seven isomers of chloropentane will be studied as proposed and accompanied by product analyses.

One additional factor in support of this research project is an award from the Research Corporation to fund this project during the summer of 1988 and 1989. The grant for $16,000 includes support for an undergraduate student to be invovled in the research program.

I am eager to spend essentially full-time on this research project for the next eight months.

LIST OF TABLES

TABLE I Summary of Cyclic Voltammetric Data for Alkly Chlorides and Butyryl Chloride in Neutral and Slightly Acidic Melts

TABLE II Cyclic Voltammetric Data for a Butyryl Chloride - Benzene Mixture Showing Dependence on Melt Composition

ILLUSTRATIONS

For all of the CV diagrams the starting potential is +0.50 volts vs. an Al wire in 0.60 melt. Cathodic sweeps are to the right, anodic to the left, sweep rate is 100 mv/sec.

Figure 1 CV curves for 1-Cl-2-Me-Propane and 2-Cl-2-Me-Propane

a. 0.50 MeEtImCl- $AlCl_3$ melt, sweep limits are +2.40 to -1.90 V

b. 1-Cl-2-2Me-Propane in 0.5 melt, +2.4 to -1.9 V

c. 1-Cl-2-Me-Propane in slightly acidic melt, +2.4 to -1.9V

d. 1-Cl-2-Me-Propane, melt acidity intermediate between that of (b) and (c), +2.4 to -1.9 V

e. 2-Cl-2-Me-Propane, 0.50 melt, +2.45 to -1.30 V

f. 2-Cl-2-Me-Propane, slightly acidic melt, +2.3 to -1.9 V

g. benzene added to solution of (f)

Figure 2 CV curves for 2-Cl-butane and 1-Cl-butane

a. 2-Cl-butane in 0.50 melt, +2.5 to -1.4 V
b. 2-Cl-butane in slightly acidic melt, 2.5 to 1.9 V
c. 2-Cl-butane in neutral melt that had been acidic, 2.5 to -1.6 V
d. 1-Cl-butane in 0.50 melt, 2.4 to -1.9 V
e. 1-Cl-butane in slightly acidic melt, 2.4 to -1.9 V
f. 1-Cl-butane in neutral melt that had been acidic, 2.4 to -1.9 V

Figure 3 CV curves for 2-Cl-Propane and 2-Cl-2-Me-Propane

a. 2-Cl-Propane in 0.50 melt, 2.4 to -1.9 V
b. 2-Cl-Propane in slightly acidic melt, 2.4 to -1.9 V
c. 2-Cl-Propane in neutral melt that had been acidic, 2.4 to -1.9 V
d. benzene added to solution of (c), 1.9 to -1.9 V
e. benzene added to 2-Cl-2-Me-Propane in neutral melt, 2.0 to -1.9 V
f. same as (e) except taken 60 minutes after addition of benzene

Figure 4 CV curves for butyryl chloride

a. 0.5000 melt, 2.5 to 1.9 V

b. butyryl chloride in 0.5000 melt, 2.5 to -1.75 V

c. butyryl chloride in 0.5031 melt, 2.5 to -1.75 V

d. butyryl chloride in 0.5039 melt, 2.5 to -1.8 V

e. butyryl chloride in 0.5049 melt, 2.5 to -1.8 V

Figure 5 CV curves for butyryl chloride with benzene solution, approximately 1.5 X 10^{-3} M in butyryl and approximately 5 X 10^{-3} M in benzene

a. melt comp. is 0.4927, 1.25 to -2.1 V

b. melt comp. is 0.4987, 2.1 to -1.9 V

c. same as (b) showing anodic sweep first

d. melt comp. is 0.5040, 2.1 to -1.9 V

e. melt comp. is 0.5049, 2.1 to -1.9 V

Figure 6 CV curves for butyryl chloride and benzene as in Figure 5

a. melt comp. is 0.5050, 2.1 to -2.1 V

b. melt comp. is 0.5059, 2.1 to -2.1 V

c. melt comp. is 0.5059, effect of continuous sweeping is demonstrated

d. melt comp. is 0.5064, 2.1 to -2.1 V

Figure 7 CV curves for butyryl chloride and benzene as in Figure 5

a. melt comp. is 0.5071, 2.1 to -2.1 V

b. melt comp. is 0.5075, 2.1 to -2.1 V

c. melt comp. is 0.5084, 2.1 to -2.1 V

d. melt comp. is 0.5087, 2.1 to -2.1 V

Figure 8 CV curves for butyryl chloride and benzene as in Figure 5

a. melt comp. is 0.5088, 2.1 to 02.1 V

b. melt comp. is 0.5090, 2.1 to -2.1 V

c. melt comp. is 0.5097, 2.1 to -2.1 V

d. melt comp. is 0.5108, 2.1 to -2.1 V

e. melt comp. is 0.5131, 2.1 to -0.75 V

Figure 9 CV curves for butyryl chloride in slightly acidic melt showing effect of benzene

a. 150 mM butyryl chloride in slightly acidic melt, 2.5 to -1.8 V

b. 300 mM benzene added to solution of (a), 2.05 to -1.8 V, current scale in (b) is double that in (a), i.e., 50 µA/cm in (b) and 25 µA/cm in (a)

REFERENCES

(1) Piersma, B.J. and Wilkes, J.S., FJSRL-TR-82-0004, September, 1982.

(2) Morrison, R.T. and Boyd, R.N., Organic Chemistry, 3rd Ed., Allyn and Bacon, Inc., Boston (1974), P. 378

(3) Boon, J.A., Levisky, J.A., Pflug, J.L. and Wilkes, J.S., Journal of Organic Chemistry, 51, 480 (1986)

(4) Koch, V.R., Miller, L.L. and Osteryoung, R.A., Journal of the American Chemical Society, 98.17, 5277 (1976)

(5) Luer, G.D. and Bartak, D.E., Journal of Organic Chemistry 47, 1238 (1982)

(6) Wilkes, J.S., Levisky, J.A., Wilson, R.A. and Hussey, C.L., Inorganic Chemistry, 21, 1263 (1982)

TABLE I

Solution	Cathodic Peaks			Anodic Peaks			Notes
	$E_{p,c-1}$	$Ep,c-2$	$E_{p,c-3}$	$E_{p,a-1}$	$E_{p,a-2}$	$E_{p,a-3}$	
1. 2-Cl-Propane/ slightly acidic melt	-0.40V	-0.80V	-1.65V	--	--	--	A
2. 2-Cl-Propane/ 0.50 (after being acidic)	--	-0.85	-1.60	--	--	--	B
3. 1-Cl-Butane/ slightly acidic melt	-0.33	-1.58	-1.68	1.23	1.53	--	C
4. 1-Cl-Butane/ 0.50 (after being acidic	-0.35	-1.58	--	1.21	1.58	--	D
5. 2-Cl-Butane/ slightly acidic melt	-0.33	-1.50	--	1.20	1.58	--	E

TABLE I (cont.)

6. 2-Cl-Butane/ 0.50 (after being acidic)	-0.35	-1.10	--	1.20	1.58	--	F
7. 1-Cl-2-Me-Propane/ 0.50 melt	--	-1.0	--	--	--	--	G
8. 1.Cl-2-Me-Propane/ slightly acidic melt	--	-0.80	1.50	--	--	--	H
9. 1-Cl-2-Me-Propane/ (after being acidic)	--	-1.50	--	--	--	--	I
10. 2-Cl-2-Me-Propane/ 0.50 melt	-0.32	-0.99	-1.55	1.23	1.61	1.84	J
11. 2.Cl-2-Me-Propane/ slightly acidic melt	-0.35	-1.00	-1.55	1.11	1.55	--	K

TABLE I (cont.)

12. 2-Cl-2-Me-Propane/ 0.50 (after being acidic)	-0.32	-1.03	-1.35	1.20	1.56	1.75	L
Triphenyl methyl chloride/ 0.52 BPC (Luer & Bartak)	+0.10	--	--	0.17	0.85	2.1	
13. butyryl chloride/ 0.50 melt	--	-0.98	-1.55	--	+1.50	--	M
14. butyryl 0.503 melt	-0.30	-0.70	-1.63	--	--	--	N
15. butyryl chloride/ 0.504-0.505 melt	-0.32	--	-1.63	--	--	2.25	O
16. butyryl chloride/ acidic melt	-0.40	-0.98	-1.30	--	--	2.25	P

Notes:

A: No redox in 0.50 melt. $E_{p,c-1}$ is a shoulder on peak $E_{p,c-2}$.
Al deposition at ~-0.75V and Al stripping at -0.13V

B: No oxidation observed.

C: No redox behavior in 0.50 melt. $E_{p,c-1}$ has a shoulder at -0.20V.
Al deposition at -0.7V and Al stripping at +0.03V.

D: $E_{p,c-1}$ has a shoulder at -0.20V. Anodic peaks are not present without prior reduction.

E: No redox behavior in 0.50 melt. $E_{p,c-1}$ has a shoulder at -0.20V.
Al deposition at -1.35V and Al stripping at 0.00V.

F: Oxidation current anodic to $E_{p,a-2}$ but no defined peaks.

G: Product at $E_{p,c-2}$ is irreversibly absorbed, and with continued cycling, $E_{p,c-2}$ varies from -1.0 to -1.35V. Not anodic peaks observed.

H: Position and height of cathodic peaks are very dependent on melt acidity. No oxidation observed.

I: No oxidation observed.

J: Anodic peaks are not present without prior reduction. After ~24 hours, the anodic peak $E_{p,a-3}$ is not present.

K: Anodic peaks are not present without prior reduction.

L: There is an additional anodic peak before $E_{p,a-1}$ at 0.95V.

M: The anodic current beginning at 1.5 V is only slightly greater than background. The cathodic current at -1.55 V goes off scale and after potential reversal is greater for approximately 200 mv, analogous to metal deposition.

N: Essentially no oxidation observed. The major cathodic peak is at -1.63 V. The small peak at -0.70 V is broad and there are indications of additional small peaks at -1.05 and -1.30 V.

TABLE II

Butyryl Chloride / Benzene / MeEtImCl Melt

Melt Composition	Cathodic Peaks (volts)				Anodic Peaks (volts)			Notes
0.4927	--	-1.00	--	-1.83	0.25	--	--	A
0.4987	--	-1.00	-1.25	--	0.25	1.57	--	B
0.5040	-0.57	--	-1.65	--	--	1.75	1.97	C
0.5050	-0.57	--	-1.65	--	--	1.75	1.97	-
0.5059	-0.55	--	-1.60	-1.97	--	1.73	1.95	D
0.5064	-0.53	-0.58	-1.75	-2.00	-1.55	1.73	--	E
0.5071	-0.50	-0.58	-1.75	-1.95	-1.30	1.72	--	-
0.5075	-0.52	-0.77	-1.75	-1.88	-1.35	1.70	--	F
0.5084	-0.55	-0.80	-1.73	-1.88	--	--	--	G
0.5087	-0.55	-0.76	-1.72	-1.88	--	1.70	--	H
0.5088	-0.50	-0.78	-1.66	-1.87	-0.05	1.40	1.70	I
0.5090	-0.50	-0.80	-1.70	-1.95	-0.05	1.40	1.70	J

TABLE II (Cont.)

0.5097	-0.50	-0.85	-1.82	-1.98	-0.05	1.40	1.73	
0.5108	-0.50	-0.85	-0.97	-1.82	-0.03	1.40	1.71	
0.5131	-0.50	-0.70	--	--	-0.33	0.33	1.55	K

Notes:

A: The major cathodic peak is at -1.83 V. An anodic current begins at about 0.25 V but no peak is observed.

B: Beginning at -1.25 V, the cathodic current rapidly increases and goes off scale. A major anodic peak is observed at 1.57 V.

C: The cathodic peak at -1.65 V is only a shoulder on the increasing cathodic current. Anodic currents are much smaller than cathodic currents.

D: A shoulder is observed on the major cathodic peak at -0.42 V. Anodic currents are small compared to cathodic currents.

E. The cathodic peak formerly appearing at -0.55 V has split giving the two closely spaced peaks.

F: The cathodic peak at -1.75 V appears to have split giving peaks at -1.65 and -1.83 V.

G: A small cathodic peak is observed at -0.23 V. Anodic current is observed but with no peaks observed.

H: Features in the anodic current are becoming evident.

I: The relative heights of the cathodic peaks at -0.5 and -0.8 have shifted.

J: An Al stripping peak is clearly evident at -0.05 V.

K: Al deposition current goes off scale at -0.7 V.

O: The cathodic current increases significantly as melt acidity is increased, particularly the peak at -0.32 V. The peak obtained previously at -0.70 is observed by the higher current although the small peaks at -1.05 and -1.30 are still evident. Very small oxidation currents are observed beginning at 2.25 V.

P: The major cathodic peak is at -0.40 V.

FIGURE 1

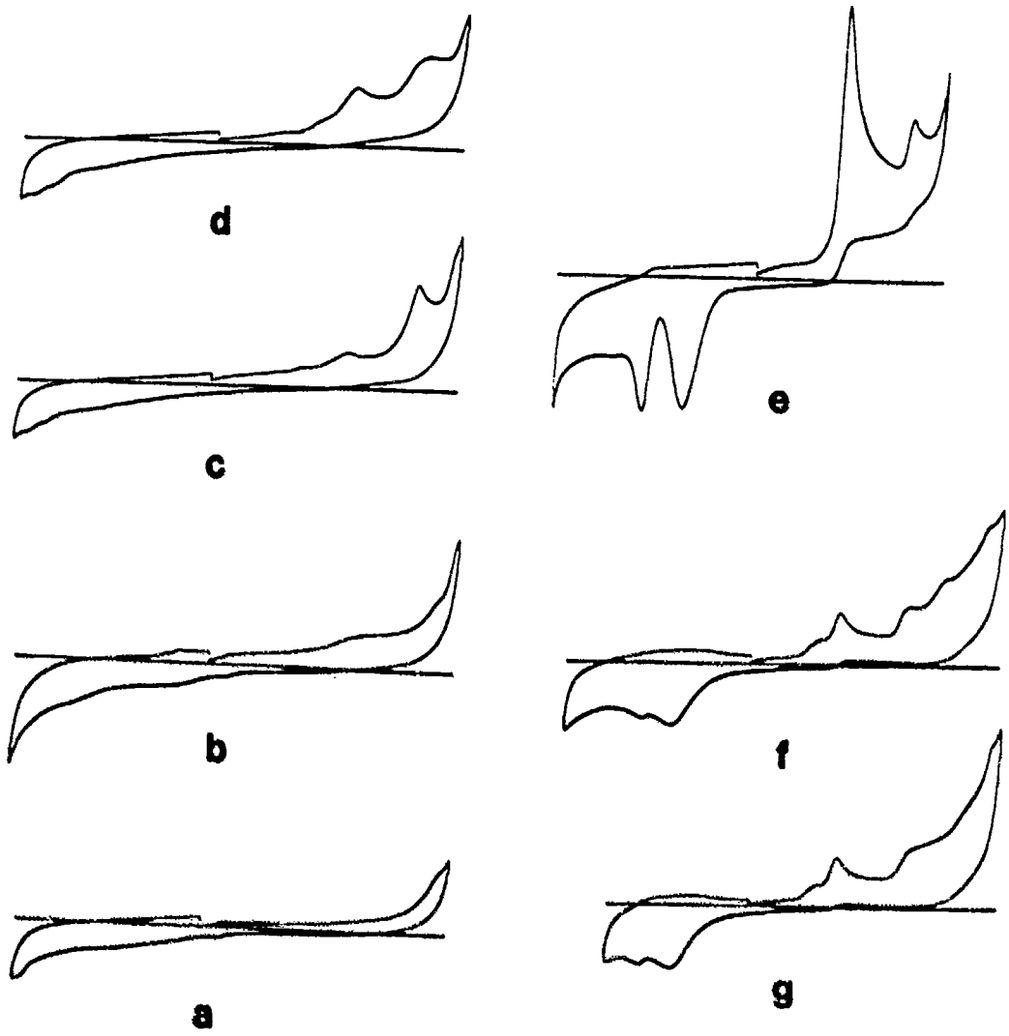


FIGURE 2

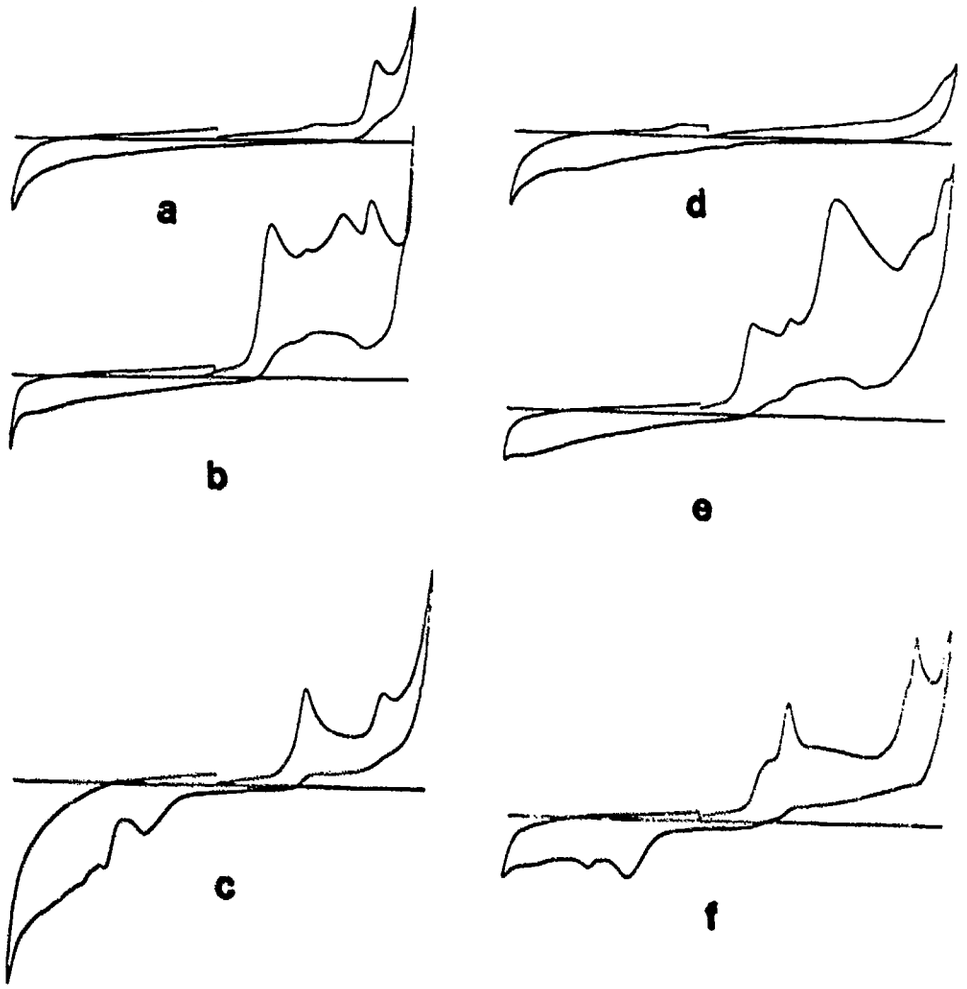


FIGURE 3

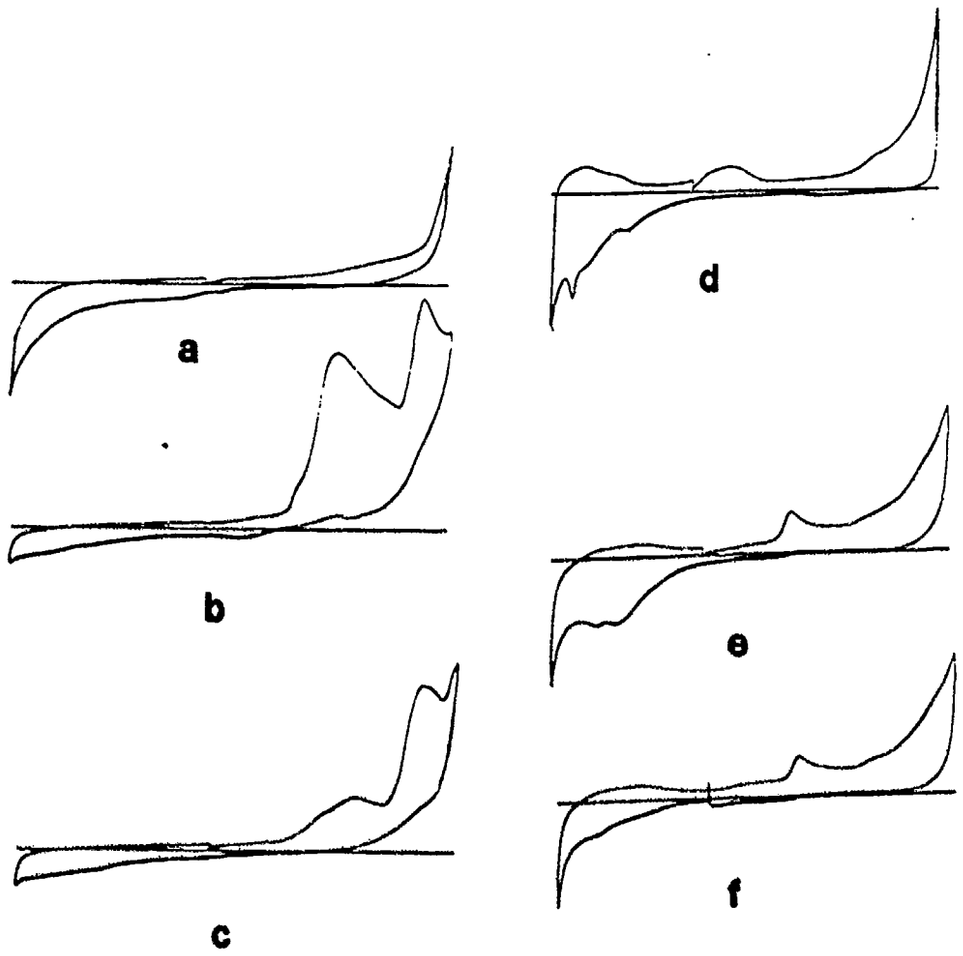


FIGURE 4

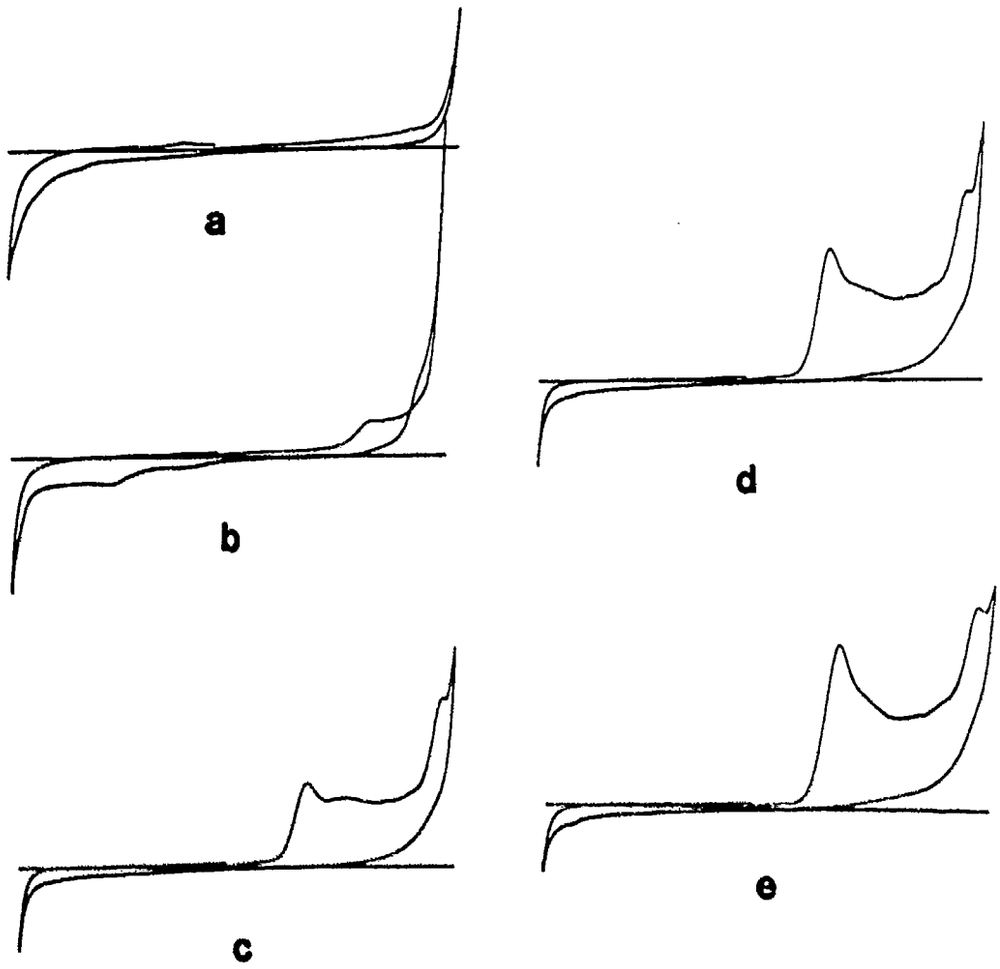


FIGURE 5

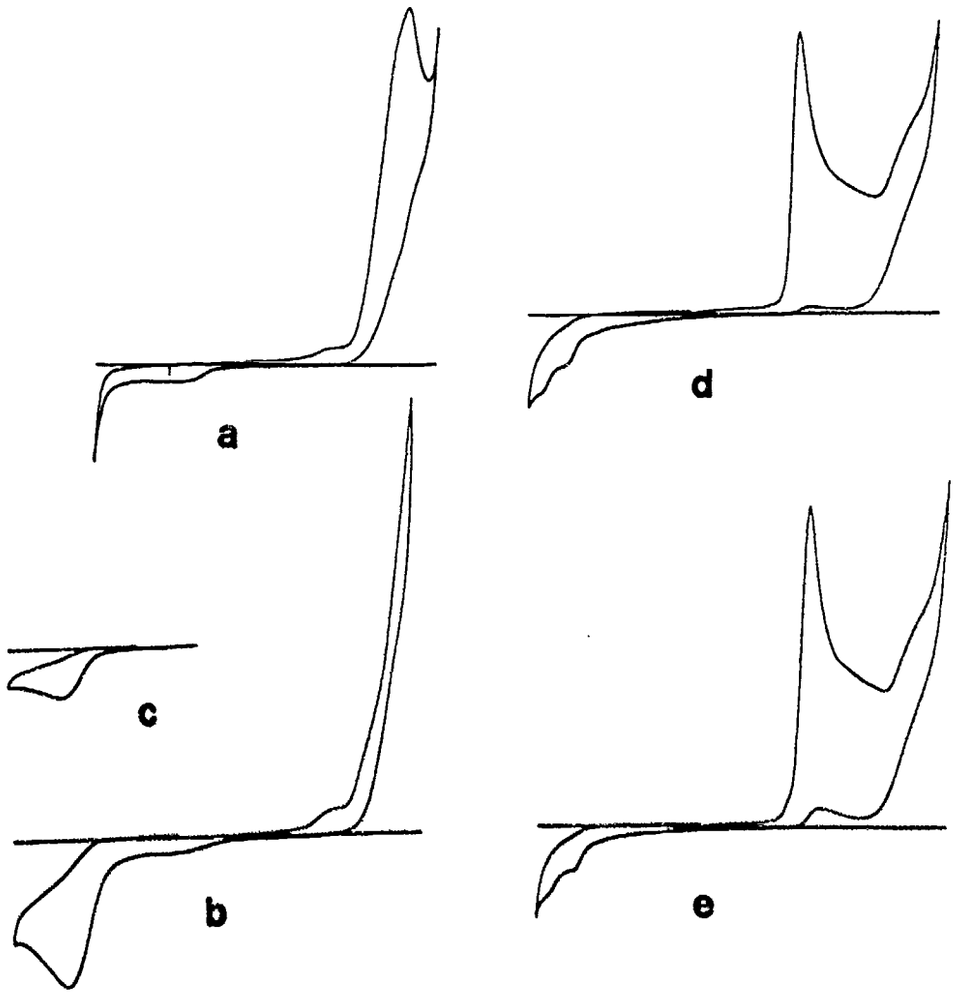


FIGURE 6

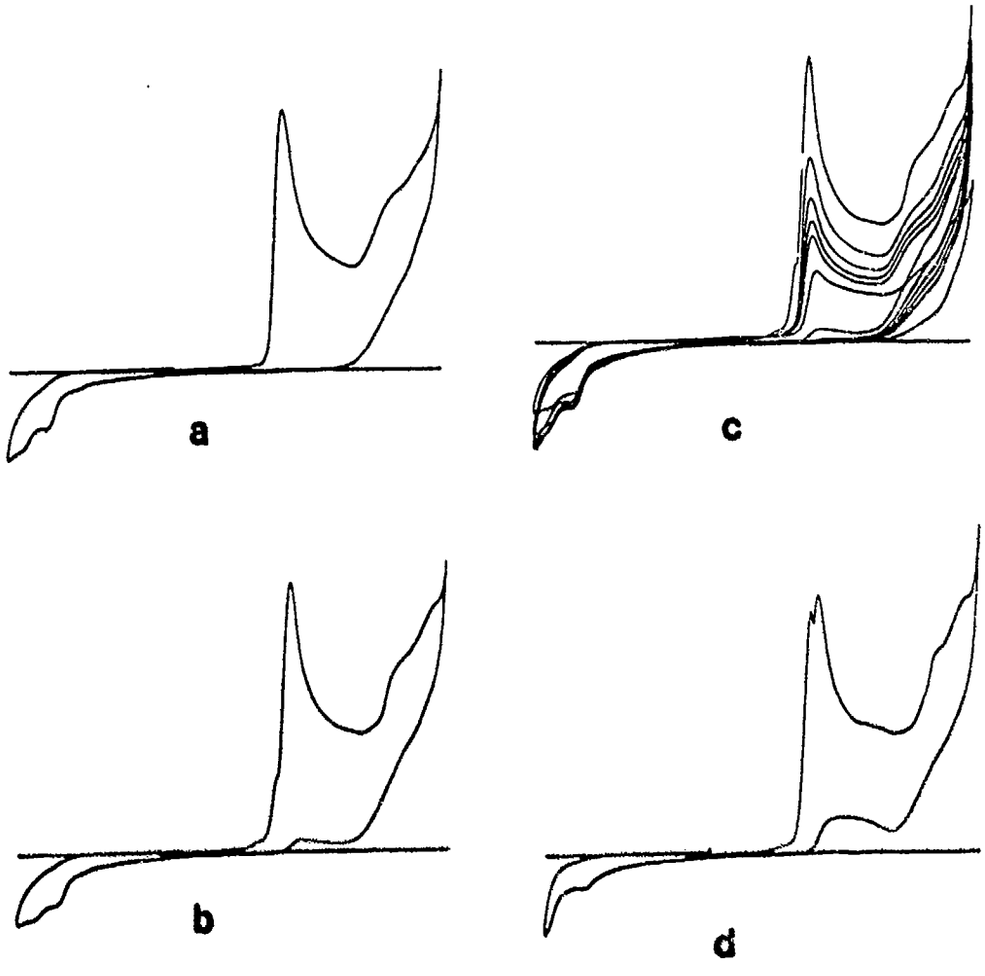


FIGURE 7

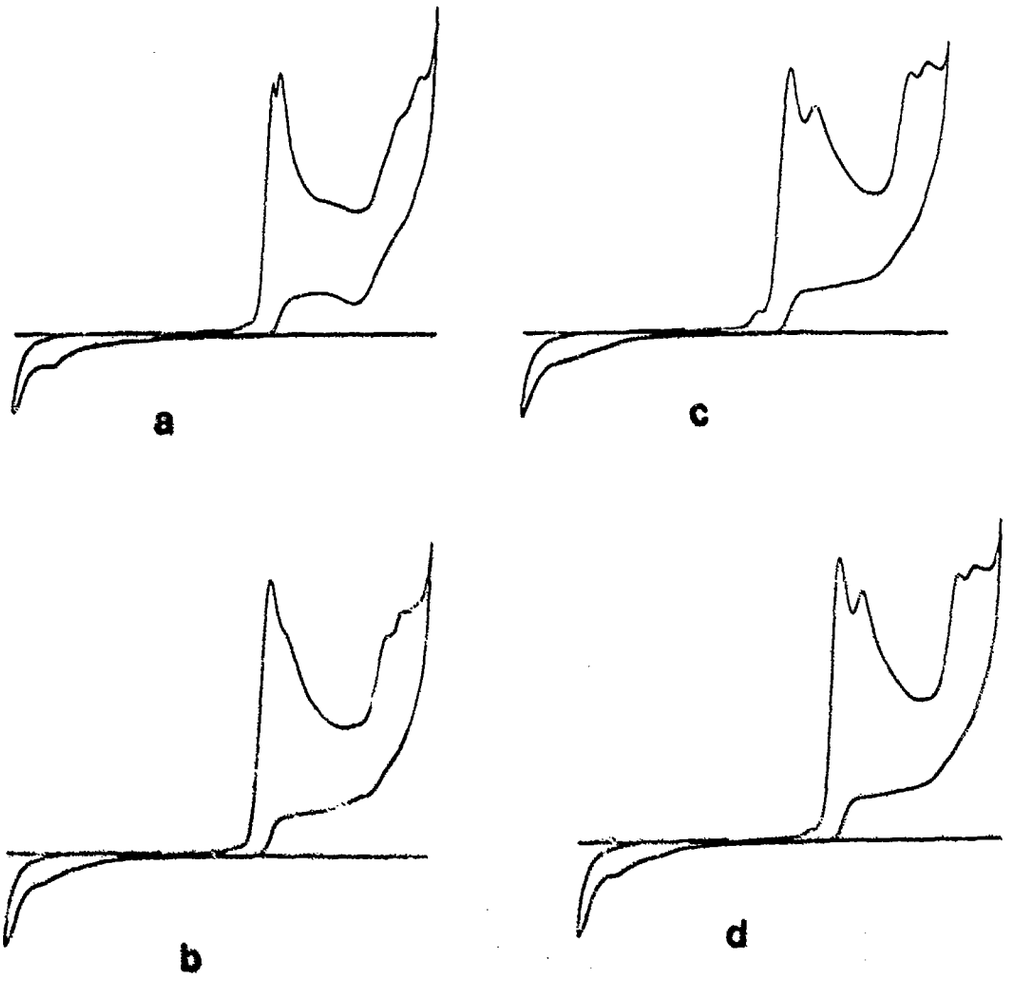


FIGURE 8

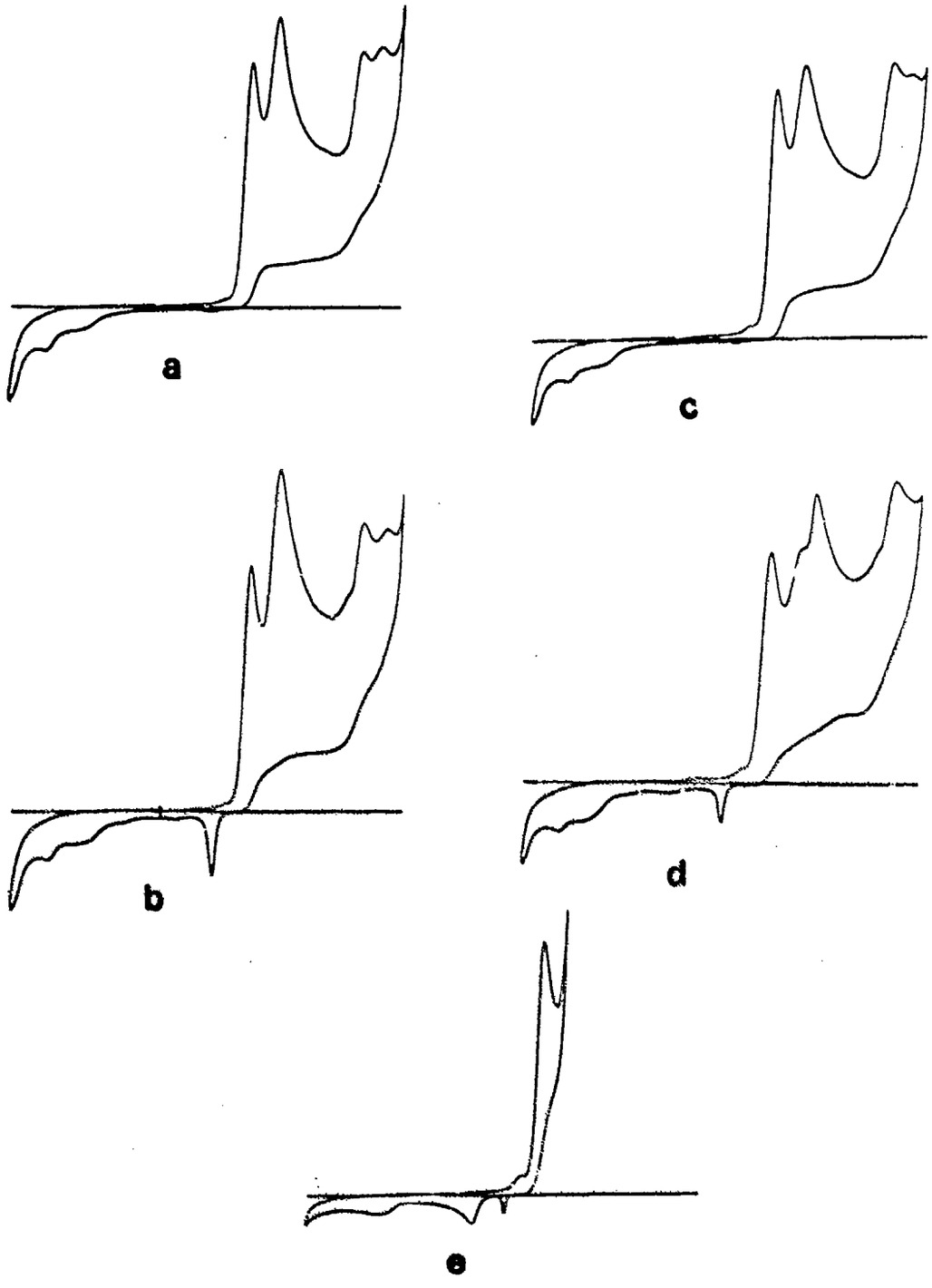
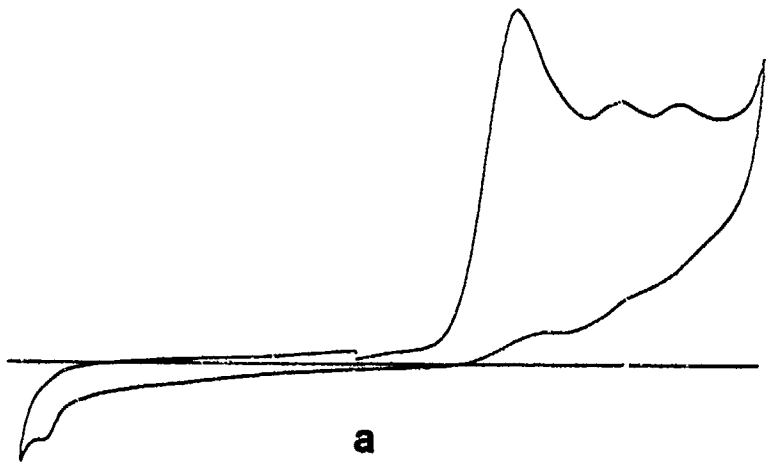
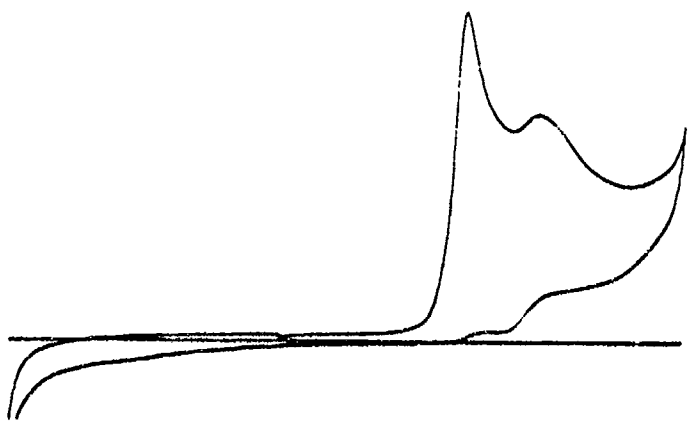


FIGURE 9



a



b

IMPROVED DISTRIBUTED OPERATING SYSTEM COMMUNICATION PROTOCOLS

Final Report for

Contract No. F49620-85-C-0013/SB5851-0360

Purchase Order No. S-760-6MG-061

Technical Report for

Universal Energy Systems

by

Craig G. Prohazka

December 21, 1987

1 INTRODUCTION

This document constitutes the final report for the research conducted by Professor Craig G. Prohazka under the US Air Force Minigrant Program, contract no. F49620-85-C-0013/SB5851-0360, purchase order no. S-760-6MG-061.

The proposal submitted to the Minigrant program specified that our work address the design of protocols for communication between instances of a distributed operating system (DOS) running at different sites. We have investigated four areas of such protocol design.

The first is the design of distributed synchronization problem protocols, presented in section 2. We solved three such problems: the termination detection problem, the mutual exclusion problem, and the distributed bounded buffer producer/consumer problem. The first two of these have earlier been examined by other researchers. Our protocols outperform theirs in several ways, including delay and required number of inter-site messages.

The second area is the design of multiprocessor load sharing protocols. We propose a new load sharing protocol based upon previous researchers' work and some novel thoughts on parallel user process behavior.

The third area is the problem of increasing database access concurrency in a distributed system. It has been claimed (see for example [38]) that nested transactions provide such increased concurrency. In the present report we show that claim to be false. We then propose a simple but effective solution: decreasing data item granularity.

Finally, the fourth area is a continuation of the 1986 Summer Faculty Research Program work. That work determined the general inter-site communication services required by a simple but typical distributed operating system. Such general services include for example send a datagram, broadcast a datagram, establish a virtual circuit, etc. The present report

identifies the specific communication services required by the DOS. Then we show that all the communication services can be layered so that each inter-site DOS communication will find the services it need by e..tering at an appropriate layer and accessing only that and lower layers.

2 DISTRIBUTED SYNCHRONIZATION PROTOCOLS

2.1 An Improved Termination Detection Protocol

2.1.1 Introduction

The termination detection problem is the following. A distributed system has N sites cooperating in some computation. At any point in time each site is either active or passive. It is active if it is executing part of the computation. It is passive if it is idie. An active site may at any time send an activation message destined to any other site. Upon receiving an activation message, a passive site becomes active; an active site is uneffected. An active site may become passive at any time. However a passive site becomes active only after it receives an activation message. Reliable communication is assumed; however we do not assume that messages leaving the same source for the same same destination arrive there in the order in which they left the source. The problem is for any site at any time it chooses to determine if the computation is complete; that is, to determine if all sites are passive and no activation messages are in transit. The site which tests for termination is often called the "distinguished site".

2.1.2 Background

Topor [1] presents a tree-based protocol. The distinguished site is the root vertex of the graph representing the distributed system. A spanning tree is constructed. The termination detection protocol generates waves of signals sent through the tree. First a wave moves from the leaves to the root. If this wave reaches the root without detecting termination, the root site sends another wave of signals to the leaves. It is returned to the root. This sequence of events is repeated until termination is detected.

Sanders [2] proves a necessary and sufficient condition for termination. A local snapshot

of a site is defined to be the value of its state variables at the instant the snapshot is made. The state variables are defined by the termination detection protocol. When the distinguished site wishes to test for termination, it requests a snapshot from each site. The set of local snapshots is called a combined snapshot.

For each snapshot a local time-slice is defined. It is the set of events that occurred at the site before the snapshot was taken. For any combined snapshot, the corresponding combined time-slice is the union of the local time-slices. The necessary and sufficient condition for termination is the following: the combined snapshot indicates that all sites are idle and the associated combined time-slice has the property that no activation message transmission crosses its boundary. This test has the disadvantage that the set of all events occurring at each site must be stored forever.

2.1.3 Improved Protocol

We propose an original token-based protocol for the termination detection problem. Here when the distinguished site wishes to test for termination it initiates a termination detection round. If the result of the round is negative - that is, termination has not yet occurred - then the distinguished site ends the round. If it wishes, it may initiate another round at a later time.

In each round a number of tokens are generated in a distributed fashion and traverse every communication channel in both directions. Note that we assume channels are full duplex. The token format is shown in figure 1.

Figure 1. The token format.

The TAG field identifies the message as a token. The DISTINGUISHED SITE field specifies the distinguished site. The ROUND field specifies the termination detection round in which the token was created. If a site receiving a round i token is active, it sends to the distinguished site a NO-TERMINATION-FOR-ROUND-I message. If any site receives a NO-TERMINATION-FOR-ROUND-I message it generates no more round i tokens. If a site has sent and received a round i token on each of its channels, while being passive the whole time, it transmits an I-BELIEVE-TERMINATION-FOR-ROUND-I message to the distinguished site. If the distinguished site receives an I-BELIEVE-TERMINATION-FOR-ROUND-I message from all other sites, it concludes termination has occurred.

We now formalize the operations performed by the distinguished site and then all non-distinguished sites in two different protocols.

Protocol Performed by the Distinguished Site to Test for Termination for Round i

1. if active, conclude that termination has not occurred and STOP.
2. transmit a round i token to each neighbor.
3. if an activation message destined for any site has arrived, conclude no termination and STOP.
4. if a NO-TERMINATION-FOR-ROUND-I message has been received, conclude no termination and STOP.
5. if a round i token has been received from all neighbors, go to step 6. otherwise, go to step 3.
6. if an I-BELIEVE-TERMINATION-FOR-ROUND-I has been received from all other sites, conclude termination and STOP.

Protocol Performed by a Non-Distinguished Site When it Receives its First Round i Token

1. if active, transmit a NO-TERMINATION-FOR-ROUND-I message to the distinguished site and STOP.
2. if a NO-TERMINATION-FOR-ROUND-I message on its way to the distinguished site has been received, STOP.
3. transmit a round i token to each neighbor.
4. if an activation message destined for any site has arrived, transmit a NO-TERMINATION-FOR-ROUND-I message to the distinguished site and STOP.

5. if a round i token has been received from all neighbors, transmit an I-BELIEVE-TERMINATION-FOR-ROUND-I message to the distinguished site and STOP; otherwise, go to step 3.

In the above protocols, STOP means to stop the execution of the protocol, not to stop the distributed computation whose termination is being tested.

Note that for these protocols, a lost token causes no special problems; that is, it may be retransmitted as any other lost message. Multiple copies of the same same token cause no difficulties at all.

Until now we have assumed that the identity of the distinguished site remains constant from round to round. This is not necessary. In fact, two sites may simultaneously play the role of distinguished site and thereby concurrently test for termination without interfering with each other.

Proof of Protocols

Suppose that the distinguished site concludes no termination for round i. This could occur only if the distinguished site is active during round i or if it received a NO-TERMINATION-FOR-ROUND-I message. In either case, at least one site must have been active when tested. Hence termination had not occurred before the beginning of round i.

Next suppose that the distinguished site mistakenly concludes termination for round i. This occurs only if an activation message destined to some site j_M arrives there after that site has sent an I-BELIEVE-TERMINATION-FOR-ROUND-I message. Figure 2 illustrates site j_M and the path the activation message follows to site j_M.

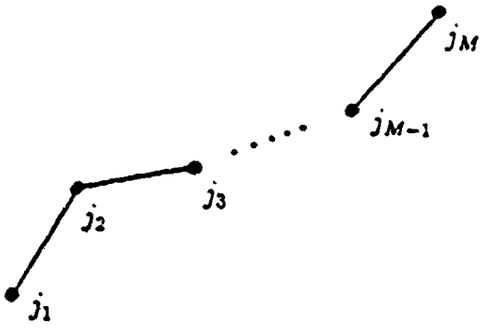


Figure 2. The path followed by an activation message.

Let j_1 be the site closest to site j_M, with distance measured along the path, which transmitted the activation message before it sent its I-BELIEVE-TERMINATION-FOR-ROUND-I message. We may assume without loss of generality that such a site exists by the following argument. Suppose that every site including the activation message's source transmitted the activation message after sending its I-BELIEVE-TERMINATION-FOR-ROUND-I message. Then some earlier activation message must have been in transit to the source of the activation message currently under investigation at the time when the source sent its I-BELIEVE-TERMINATION-FOR-ROUND-I message. In this case we could apply the remainder of this proof to the earlier activation message.

Site j_1 must have transmitted the activation message before it transmitted a round i token to any of its neighbors. However we know that the activation message arrived

at site j_2 after that site sent its I-BELIEVE-TERMINATION-FOR-ROUND-I message. Further, site j_2 could not have sent the latter message unless it had already received a round i token from all its neighbors, including site j_1. This is a contradiction. Hence, the distinguished site cannot mistakenly conclude termination for round i.

QED

2.1.4 Performance

The improved termination detection protocol presented here has the following performance. The number of messages to be sent per round is less than or equal to $N^2 + N - 1$, as explained next. Each site sends a token to each of its neighbors; this requires N^2 messages. Each site except for the distinguished site sends at most one I-BELIEVE-TERMINATION-FOR-ROUND-I message or one NO-TERMINATION-FOR-ROUND-I message; this requires at most $N - 1$ messages.

The delay from the initiation of round i by the distinguished site until that site decides whether termination occurred before round i is at most the time required for a message to follow the longest cycle traversing the distinguished site, assuming that the time required for protocol computation and the time required to forward a message through a site are negligible in comparison with propagation delay.

2.1.5 Comparison

Our termination detection protocol outperforms both Topor and Sanders. The number of messages and the delay required by Topor's protocol seen to be unbounded. Our protocol on the other hand requires at most $N^2 + N - 1$ messages and a time equal to the delay for one maximal cycle traversal. Sanders' approach requires that records of all events at

each site be stored forever. Our protocol requires no such records.

2.2 An Improved Distributed Mutual Exclusion Protocol

2.2.1 Introduction

The distributed mutual exclusion problem is the following. A distributed system has N sites cooperating in some computation. Located at one of these sites is a critical resource, which processes at all sites may need to access. However, at most one site is allowed to access it at a time. Further, there is no access control mechanism provided at the resource's site. Instead the sites must cooperate with each other so as to ensure that the resource is accessed by just one site at a time.

We demand that a distributed mutual exclusion protocol satisfy two additional constraints: absence of deadlock and absence of starvation. Deadlock exists in a distributed system if and only if the following three conditions hold [8].

1. mutual exclusion - at least one resource is non-sharable,
2. no preemption - resources cannot be preempted, and
3. circular wait - there must exist a sequence of sites $s_1, s_2, s_3, ..., s_n$ such that s_i is waiting for a resource held by s_{i+1} for all $i = 1, 2, 3, ..., n-1$ and s_n is waiting for a resource held by s_1.

Starvation occurs when one site must wait indefinitely to access the critical resource even though other sites' requests are being serviced.

2.2.2 Background

Ricart and Agrawala [9] propose a refreshingly readable sequence number based protocol for the distributed mutual exclusion problem. A site wishing to use the critical resource

requests permission from all other sites by broadcasting a REQUEST message with a sequence number. The sequence number functions in a manner similar to a time stamp. When a site receives a REQUEST, it compares this REQUEST with its own unperformed REQUESTS. If it has no such request whose sequence number is smaller than that of the REQUEST just received, it sends a REPLY message back to the requesting site. The REPLY message indicates that the sending site gives the receiving site permission to use the critical resource. The requesting site waits for a REPLY from all other sites. Then it uses the critical resource.

Roberts and Chang [10] present a technique for extrema finding in circular configurations of processes. Chang [11] claims that "a simple modification of this technique to include sequence numbers gives a decentralized mutual exclusion protocol with a better performance than the Ricart-Agrawala requiring only $O(n\log n)$ message passes instead of $O(n^2)$ per critical section." However Chang does not present this modification. Thus we cannot consider Chang's claim as justified.

In a letter in response to Chang, Ricart and Agrawala [see 11] propose the first (of two) modifications of their original protocol. Here all sites are organized into a logical ring. Each REQUEST, including its sequence number, circles completely around the ring, waiting at each site until all local younger requests are serviced. When the REQUEST returns to the initial site, this site accesses the critical resource.

Carvalho and Roucairol [12] present a slight modification of Ricart-Agrawala which under some circumstances allows a site to use the critical resource more than once without sending additional REQUESTs. In their scheme, the permission implied by a REPLY message sent from site i to site j persists until site j sends a REPLY to site i.

In a letter responding to Carvalho and Roucairol, Ricart and Agrawala [see 12] present a second modification of their original protocol. It uses a token to pass the privilege of using the critical resource. A site wishing to use the critical resource broadcasts a REQUEST with sequence number. The site currently holding the token sends it to the REQUESTor if the REQUEST's sequence number is smaller than all his own requests, if any. Then the REQUESTor uses the critical resource.

2.2.3 Improved Protocol

The improved protocol presented here uses a token to control critical resource access. The token format is shown in figure 3.

Figure 3. The token format.

The TAG field identifies the message as a token. The purpose of the REQUEST Y/N field is explained below. The INCARNATION field is used to handle lost and duplicate tokens and is described below.

Basically the token follows a preestablished logical ring. When a site which does not currently need to access the critical resource receives the token, it forwards it to the next site on the ring. However, if it does need to access the critical resource, it sends a

request to the critical resource. The token is piggybacked upon the request. Note that at this point the token has temporarily stopped following the logical ring. Now the purpose of the REQUEST Y/N field can be understood. If this field is set (to Yes), the token is piggybacked upon a request. If it is cleared (to No), the token is not piggybacked; instead it is simply following the logical ring through the critical resource's site. Upon the reception of a token with cleared REQUEST Y/N field, the critical resource's site receives the opportunity to access the critical resource.

Next the request is serviced by the critical resource. When the service is complete, the critical resource sends a reply containing the service result back to the requesting site. Next it forwards the token to the site following the requestor on the logical ring. Now the token has resumed its traversal of the logical ring. When the next site receives the token, it may access the critical resource by repeating the protocol just described.

If the critical resource's site has not seen the token for a preestablished period of time long enough for the token to traverse the entire ring, it creates a new token with an incremented INCARNATION.

Suppose the critical resource's site receives an old token, mistakenly assumed lost. Then it immediately removes it. If the old token was sent to the critical resource's site piggybacked upon a request, then the critical resource's site returns an EXCEPTION message to the requesting site. This EXCEPTION message informs the requesting site that its request was ignored. If the token is not piggybacked, but simply reaches the critical resource's site as it traverses the logical ring, then the critical resource's site just removes the old token.

We now formalize the operations performed by the critical resource's site and then all

other sites in two different protocols.

Protocol Performed by the Critical Resource's Site

1. if no token has been received for T seconds,
   (a) create a new one with an incremented INCARNATION.
   (b) forward the new token to the next site on the logical ring.
2. if a token destined for this site is received and if its INCARNATION is less than the greatest existing INCARNATION,
   (a) remove this token.
   (b) if the token was piggybacked upon a request, send an EXCEPTION message to the requesting site informing it that the request was ignored.
3. if a token with REQUEST Y/N field set (to Yes) destined for this site is received,
   (a) service the request it is piggybacked upon and record the requesting site's identity.
   (b) when the request has been serviced, return the result to the requesting site.
   (c) clear the token's REQUEST Y/N field (to No) and forward it to the site which follows the requesting site on the logical ring.
4. if a token with REQUEST Y/N field cleared (to No) destined for this site is received, determine if this site has a request for the critical resource.
   (a) if so, service the request.
   (b) when the request has been serviced, forward the token to the site which follows the current site (the critical resource's site) on the logical ring.

Protocol Performed by All Other Sites

1. if a token destined for this site is received, determine if the site has a request for the critical resource.

   (a) if so, set the token's REQUEST Y/N field (to Yes) and send the request to the critical resource, piggybacking the token upon it.

   (b) if not, forward the token to the next site on the logical ring.

Proof of Protocols

We must show that the three protocol requirements are satisfied:

1. mutual exclusion
2. absence of deadlock
3. absence of starvation

First is mutual exclusion. The servicing of two or more site's requests for the critical resource may not overlap in time because there is only one token and it is kept at the critical resource while each request is being satisfied.

Second is absence of deadlock. Deadlock cannot exist because the third of the three conditions listed in the introduction to the distributed mutual exclusion problem cannot hold. This is true because there is only one resource (the critical resource) controlled by our protocols.

Last is absence of starvation. Starvation is impossible because the time between the servicing of consecutive requests from a particular site is upper bounded by the time

required for the token to traverse the logical ring once. At most N requests are serviced during that time.

QED

2.2.4 Performance

Our improved distributed mutual exclusion protocol has the following performance. The number of messages to be sent per critical resource access depends upon the number of sites with requests for the critical resource. If every site has one, only two messages are required: one to send the token to the site and one to send the request with the token piggybacked upon it to the critical resource's site. On the other hand, if only one site has requests, then $N+1$ messages are required per access: N messages to forward the token once around the ring and one more to send the request with the token to the critical resource's site.

The delay between consecutive critical resource accesses by any site assuming that every site has a request is at most twice the maximum delay from the critical resource's site to any other site, $2x delay(CRS \Leftrightarrow S)$. This is true under the reasonable assumption that the request service time is smaller than the quantity $2x delay(CRS \Leftrightarrow S)$.

The delay between consecutive critical resource accesses by a particular site depends upon the number of requesting sites. If every site has a request, then the delay is at most twice the maximum delay from the critical resource's site to any other site, multiplied by the number of sites, or $2N x delay(CRS \Leftrightarrow S)$. If only one site has requests, the delay is less than twice the delay from the critical resource's site to the requesting site plus the time required to forward the token completely around the logical ring, in other words $2x delay(CRS \Leftrightarrow S) + delay(RING)$. Again, these calculations assume that the

request service time is less than $2x delay(CRS \Leftarrow S)$.

2.2.5 Comparison

Our distributed mutual exclusion protocol outperforms all previous ones in terms of both number of messages and delay per critical resource access.

Ricart and Agrawala [9] require $2N-1$ messages: $N-1$ to broadcas. a REQUEST, $N-1$ for the REPLYs from all other sites, and 1 to send its request to the critical resource. Our protocol requires only from 2 to $N+1$ messages, as shown above. Their protocol results in a delay of at most $2x delay(CRS \Leftrightarrow S) + 2x delay(S \Leftrightarrow S)$ between accesses by any site, assuming a high request load. Our result, namely $2x delay(CRS \Leftrightarrow S)$, is significantly better. Because access to the critical resource is controlled by sequence numbers, the delay between accesses by a particular site depends upon the order in which requests are generated at different sites. So it is impossible to derive a meaningful bound on this delay.

The performance of Carvalho and Roucairol is in the worst case (which is typically the actual case) the same as Ricart and Agrawala [9].

Ricart and Agrawala in response to Chang requires $N+1$ messages. Again, our protocol requires from 2 to $N+1$. The delay between accesses by any site, assuming that every site has a request is $2x delay(CRS \Leftrightarrow S) + delay(S \Leftrightarrow S)$. Our result is significantly better. Again, the delay between accesses by a particular site cannot be easily bounded because it depends upon request generation order.

Ricart and Agrawala in response to Carvalho and Roucairol requires $N+1$ messages per access. The delay between consecutive accesses by any site is $2x delay(CRS \Leftrightarrow S) + delay(S \Leftrightarrow S)$. In this case as well, the delay between accesses by a particular

site cannot be easily bounded because it depends upon request generation order. Our protocol performs at least as well.

2.3 A Distributed Bounded Buffer Producer/Consumer Problem Protocol

2.3.1 Introduction

The distributed bounded buffer producer/consumer problem is the following. A distributed system has N sites cooperating in some computation. Located at one of these sites is a pool of B buffers, each of which can hold one data item. Processes at other sites may need to access the buffer. In particular, sites from time to time produce data items which must be stored in the buffer pool. Sites also from time to time consume data items after reading them from the buffer pool. As far as the consumer is concerned, all data items are equal; that is, the identity of a data item's producer is ignored. The problem is to ensure, without any access control mechanism at the buffer site, that no producer writes to a full buffer pool and no consumer reads from an empty buffer pool.

We demand that a bounded buffer producer/consumer problem protocol satisfy two additional constraints: absence of deadlock and absence of starvation. These constraints were explained in the introduction to the distributed mutual exclusion problem section.

2.3.2 Background

We have not found any earlier protocol for the distributed bounded buffer producer/consumer problem.

2.3.3 Protocol

We propose an original token-based protocol. The token format is shown in figure 4.

TAG	DESTINATION	REQUEST Y/N	INCARNATION	I

Figure 4. The token format.

The TAG field identifies the message as a token. The purpose of the REQUEST Y/N field is explained below. The INCARNATION field is used to handle lost and duplicate tokens, as described below. The I field is the current value of the number of full buffers in the buffer pool; it equals the number of data items stored in the pool.

Basically the token follows a preestablished logical ring. When a site which does not currently need to access the buffer pool receives the token, it forwards it to the next site on the ring. Now suppose that the site receiving the token has produced X items and so needs to store them in the buffer pool. If $I < B$, it increments I by $min(X, B - I)$ and sends that number of items to the buffer, along with a request to write them to the buffer. The token is piggybacked upon the request. Note that at this point the token has temporarily stopped following the logical ring. Now the purpose of the REQUEST Y/N field can be understood. If this field is set (to Yes), the token is piggybacked upon a request. If it is cleared (to No), the token is not piggybacked; instead it is simply

following the logical ring through the buffer site. Upon the reception of a token with cleared REQUEST Y/N field, the buffer site receives the opportunity to access the buffer pool.

Upon arriving at the buffer site, a token with a REQUEST Y/N field set (to Yes) is immediately forwarded to the next site on the logical ring; that is, the site following the one which produced the X items. If $I = B$ when the producer's site receives the token, it simply forwards it to the next site on the ring.

Now suppose the site receiving the token needs to consume X items. If $I > 0$, it decrements I by $min(X, I)$, sets the REQUEST Y/N field (to Yes) and sends to the buffer site a request to read $min(X, I)$ items. The token is piggybacked upon the request. When the request arrives, the items are sent to the requestor, the token's REQUEST Y/N field is cleared (to No), and the token is dispatched to the next site on the ring, directly from the buffer site. If $I = 0$ when the consumer's site receives the token, it simply forwards it to the next site on the ring.

If the buffer site has not seen the token for a preestablished period of time long enough for the token to traverse the entire ring, it creates a new token with the current value of I and an incremented INCARNATION.

Suppose the buffer site receives an old token, mistakenly assumed lost. Then the buffer site immediately removes it. If the old token was sent to the buffer site piggybacked upon a request, then the buffer site returns an EXCEPTION message to the requesting site. This EXCEPTION message informs the requesting site that its request was ignored. If the token is not piggybacked, but simply reaches the buffer site as it traverses the logical ring, then the buffer site just removes the old token.

We now formalize the operations performed by the buffer site and the all other sites in two different protocols.

Protocol Performed by the Buffer Site

1. if no token has been received for T seconds,
   (a) create a new one with an incremented INCARNATION and the current value of I.
   (b) forward the new token to the next site on the logical ring.
2. if a token destined for this site is received and if its INCARNATION is less than the greatest existing INCARNATION,
   (a) remove this token.
   (b) if the token was piggybacked upon a request, send an EXCEPTION message to the requesting site informing it that the request was ignored.
3. if a token with REQUEST Y/N field set (to Yes) destined for this site is received,
   (a) perform the read or write request it is piggybacked upon.
   (b) clear the token's REQUEST Y/N field (to No).
   (c) forward it to the site which follows the requesting site on the logical ring.
4. if a token with REQUEST Y/N field cleared (to No) destined for this site is received and if this site has produced X items,
   (a) store $min(X, B - I)$ items in the buffer pool.
   (b) increment the token's I field by $min(X, B - I)$.

(c) forward the token to the site which follows the current site (the buffer site) on the logical ring.

5. if a token with REQUEST Y/N field cleared (to No) destined for this site is received and if this site needs to consume X items,

   (a) read $min(X, I)$ items from the buffer pool.

   (b) decrement the token's I field by $min(X, I)$.

   (c) forward the token to the site which follows the current site (the buffer site) on the logical ring.

Protocol Performed by All Other Sites

1. if a token destined for this site is received and if this site has produced X items,

   (a) increment the token's I field by $min(X, B - I)$.

   (b) send $min(X, B - I)$ items to the buffer site with the token piggybacked upon them.

2. if a token destined for this site is received and if this site needs to consume X items,

   (a) decrement the token's I field by $min(X, I)$.

   (b) send a request for $min(X, I)$ items to the buffer site with the token piggybacked upon it.

Proof of Protocols

No producer will write to a full buffer pool and no consumer will read from an empty buffer pool because the I field always contains the number of data items which the buffer pool will contain after the current read or write request, if any, is performed. The REQUEST Y/N field always specifies whether any request is outstanding.

Deadlock is impossible because the third of the three conditions listed in the introduction to the distributed mutual exclusion problem cannot hold. This is true because there is only one resource (the buffer pool) controlled by our protocols.

Last is absence of starvation. Starvation is impossible because the time between the servicing of consecutive requests from a particular site is upper bounded by the time required for the token to traverse the logical ring once. At most N requests are serviced during that time.

QED

2.3.4 Performance

Our distributed bounded buffer producer/consumer protocol has the following performance. The number of messages to be sent per read or write request to the buffer depends upon the number of sites which need to read or write. Note that each read or write request may read or write multiple data items. If every site need to read or write, only two messages are required: one to send the token to the site and one to send the request with the token piggybacked upon it to the buffer site. On the other hand, if only one site has requests, then $N+1$ messages are required per access: N messages to forward the token once around the ring and one more to send the request with the token to the buffer site.

The delay between consecutive buffer accesses by any site assuming that every site has a request is at most twice the maximum delay from the buffer site to any other site, $2x delay(B \Leftrightarrow S)$. This is true under the reasonable assumption that the request service time is smaller than the quantity $2x delay(B \Leftrightarrow S)$.

The delay between consecutive buffer accesses by a particular site depends upon the number of requesting sites. If every site has a request, then the delay is at most twice the maximum delay from the buffer site to any other site, multiplied by the number of sites, or $2N x delay(B \Leftrightarrow S)$. If only one site has requests, the delay is less than twice the delay from the buffer site to the requesting site plus the time required to forward the token completely around the logical ring, in other words $2x delay(B \Leftrightarrow S) + delay(RING)$. Again, these calculations assume that the request service time is less than $2x delay(B \Leftrightarrow S)$.

3 A PROPOSED LOAD SHARING PROTOCOL

3.1 Introduction

Load sharing in a distributed system is the problem of assigning processes to different sites so as to reduce job turnaround time. This can be done by maximizing the utilization of resources while minimizing the communication between sites. Resource utilization maximization tends to distribute processes evenly among the sites. In contrast, minimizing the inter-site communication tends to assign all processes to a single site. So, there exists a conflict between these two goals and a compromise must be made to obtain an optimal load sharing policy.

In this report we present a proposed load sharing protocol based upon previous researchers' work and some novel thoughts on parallel user process behavior. As we will see, a precise definition of the job turnaround time to be minimized is necessary: specifically, our load sharing protocol is intended to minimize the sum of the real times to complete the jobs (parallel application programs) submitted to the system.

3.2 Background

Load sharing has been studied for more than twenty years. Some of the load sharing techniques in use today have been adapted from operations research results. These results concern the utilization of people, equipment, and raw materials. Specifically, if people and equipment are equated with processors and raw materials with computer programs, then these results become immediately applicable. Such an approach to load sharing is called the job-shop approach, because the terminology of manufacturing is applied to load sharing. An early text on this subject is Conway, Maxwell, and Miller [14].

Another approach is the graph theoretic technique [15]. Here a graph is used to express the problem. Each vertex represents a program module. Each edge represents the communication required between the two modules represented by the vertices at either end. Each edge is labelled by a number equal to the time required for inter-site communication should the two modules be assigned to different sites. Then the process-to-site assignment minimizing inter-site communication is found by applying the Ford-Fulkerson algorithm.

Yet another approach is the mathematical programming approach [16,17]. This approach formulates process-to-site assignment as an optimization problem and solves it using mathematical programming methods.

Finally, the heuristic approach provides fast but suboptimal protocols for process-to-site assignment [18].

Next we summarize an extensive literature search of the load sharing problem. Since our goal is the proposal of a load sharing protocol and not a mathematical analysis, we have concentrated upon practical protocols.

First Eager, Lazowska, and Zahorjan [19] present the results of experimental studies of heuristic load sharing protocols. The following are their conclusions.

1. only simple strategies with small amounts of system information are necessary.
2. the cost of process migration is mostly processor time rather than communication time.
3. sender-initiated policies are preferable to receiver-initiated policies at light to moderate loads.

4. receiver-initiated policies are preferable at high system loads but only if the costs of process migration under the two strategies are comparable. That is, receiver-initiated policies may cause the migration of executing processes. Such processes typically have a much larger context than newly created processes. Thus their migration cost is greater.

5. if the cost of process migration under receiver-initiated policies is significantly greater than under sender-initiated policies, then sender-initiated policies are uniformly better.

6. modifying receiver-initiated policies to transfer only newly created processes yields unsatisfactory results.

Leland and Ott [20] present a quasi-heuristic load sharing protocol based upon an experimental study of the behavior of 9.5 million Unix processes created at the Bell Communication Research computer system during a four month period. These are the study's results.

1. processes actually do fall into three groups, as the "folk theorem" claims:

   (a) CPU bound

   (b) IO bound

   (c) normal

2. the overwhelming majority of processes are normal

3. if X is a random variable equal to the amount of CPU time used by an arbitrary process, then given that X is greater than 3 seconds,

$$1 - F(x) = rx^{-c}$$

where r and c are constants and $1.05 < c < 1.25$. Here $F(x)$ is the probability distribution of X.

4. using the same definition of X and assuming that $x > 3$ seconds,

$$E[X - x \mid X > x] \approx k_1 + k_2 x$$

where k_1 and k_2 are constants. In other words, given that a process has used at least 3 seconds of CPU time, the expected value of the remaining CPU time is more or less proportional to the amount of time the process has already used.

Their proposed load sharing protocol included two parts: an initial placement protocol and a process migration protocol. The initial placement protocol chooses the site for a newly created process to begin running at. This protocol attempts to take advantage of the smaller context of newly created processes. The process migration protocol is receiver-initiated. When a processor becomes idle, it broadcasts an auction invitation. Each of the other processors executing at least one process which meets a certain criterion sends a bid containing its load to the idle processor. The latter processor waits for a time for such bids and accepts the bid from the processor with the highest load. Then the winning bidder sends one of the criterion-meeting processes to the idle processor.

Leland and Ott also simulate their protocols. They compare different values of tuneable parameters. In addition, they compare their process migration protocol with a "random" protocol. One result of this study is the conclusion that CPU and IO bound processes benefit from almost any migration policy, but selection heuristics must be carefully chosen

to avoid penalizing the majority of processes while rewarding the CPU and IO bound processes.

Barak and Faradise [21] describe and compare some load sharing protocols implemented in the MOS Multicomputer Operating System. Their conclusions are the following.

1. assigning newly created processes to underloaded machines seems to be an adequate means to achieve load sharing.
2. processes should be migrated to where their IO operations take place.
3. responsibility for the migration of each process may be assigned to the process itself.

Barak and Shiloh [22] present an earlier load sharing protocol implemented in MOS. Their conclusions are the following.

1. a process should remain at its current processor for a certain minimum time before being migrated.
2. a process should be moved to be physically close to the objects it must communicate with.

Wang and Morris [23] propose a performance metric for load sharing protocols called the Q-factor. This metric compares the mean response time for all processes under any given protocol with the mean response time under FCFS. They arrive at the following conclusions.

1. server-initiated protocols usually have a higher Q-factor than source-initiated protocols for the same level of information. This is true since servers are not allowed to be idle while jobs are waiting.

2. the performance of source-initiated protocols degrades as the number of servers becomes large. On the other hand, the performance of server-initiated protocols improves.

3. the performance of server-initiated protocols is less sensitive to service time variability than the performance of source-initiated protocols.

Finally Cabrera [24] presents measurements of process behavior on several Unix installations. Then he analyzes the implication of these measurements on load sharing protocols. His conclusions are the following.

1. for a wide range of lifetimes and systems, at least 40which have a lifetime of greater than T time units have a lifetime greater than $2T$ units.

2. the percentage of processes which do not benefit from remote execution increases more than linearly with increasing CPU power.

3. general purpose load balancing strategies should be based upon a process migration mechanism and driven by the detection of long-lived processes.

4. the soundest processing strategy for short processes is to execute them locally.

5. local schedulers should detect and mark long-lived processes.

6. the scheduler algorithm must be able to differentiate between long-lived processes and processes which use a lot of CPU time.

Next we present our proposed heuristic load sharing protocol based upon the above-mentioned results and some novel thoughts on parallel user process behavior.

3.3 The Process as a Unit of Parallelism

The load sharing protocol we present in this report is based upon the interpretation of user process creation explained below. Even after reading many articles on multiprocessor scheduling ([14] through [37]), we have not discovered any previous expression of this interpretation. Thus we claim that it is novel. Further, as the reader will see, it leads us directly to a simple and satisfying load sharing protocol, presented in section 3.6.

First, we distinguish between a system process and a user process. A system process is one of possibly many processes which implement (distributed) operating system functions - that is, control the system hardware and software resources so as to make their operation both convenient to the user and efficient. On the other had, a user process is one of possibly many processes which cooperate in executing some application program.

Now consider user process creation in a uniprocessor. Why would a user create more than one process to execute a particular application program? The answer is that the user would not, because the fact that only one processor exists implies that only one process could execute at any time in any case. So there is no advantage in creating additional processes.

Next consider user process creation in a multiprocessor. Here multiple processors exist; thus multiple processes may run simultaneously in the execution of the same application program. By comparing the uniprocessor case with the multiprocessor case, we see that a user should create an additional process to help execute an application program only if the following condition holds: there are two parts of the program which will execute faster

on two processors than on one, including the time for interprocessor communication.

The same argument holds even if process creation is decided not by human users, by by software, such as a compiler specially written for a multiprocessor.

Based upon the above argument, in the remainder of the explanation of our proposed load sharing protocol, we assume that multiple user processes exist only to define the parts of an application program which should be executed in parallel upon different processors.

3.4 The Job Graph

We saw in the previous section that multiple user processes serve only to express parallelism in application programs. Specifically, two distinct processes performing different parts of the code of the same application program should exist only if the code executes faster on two processors than on one, including interprocessor communication time. We assume without loss of generality and with only one exception described below that at each process creation, at least two processes are created. The only exception is the creation of the first process to begin the execution of a particular application program. Next we explain why we can assume two processes are created.

We do not allow a single process to be created for the following reason. With the exception mentioned above, a child process is always created by some parent user process. We assume that the parent process continues executing after the child is created; for otherwise, the child process could be considered a continuation of the parent process rather than a distinct process. However, we identify as two different processes the parent process before child creation and the parent process after child creation. The reason is that the parallelism expressed by the parent process before child creation if different from that expressed after child creation. Otherwise, the child process could have been created

at the same time as the parent process was created. Recalling that the function of a process is to express parallelism, it is reasonable to consider the parent process before and after child creation as two distinct processes.

We will represent the creation and termination of processes during the execution of an application program as a graph. An example is shown in figure 5.

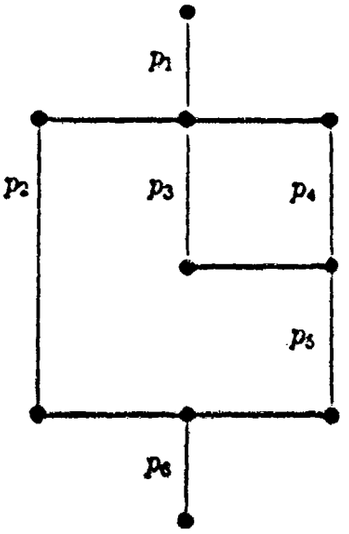


Figure 5. A job graph.

The execution of the application program begins with process p_1, represented by the vertical edge labelled p_1. Process p_1 creates three other processes, which are meant to be executed in parallel, represented by the edges labelled p_2, p_3, and p_4. The horizontal edges connecting p_1, p_2, p_3, and p_4 represent interprocessor communication. As explained in the previous paragraphs, p_1 is considered to be distinct from p_2, p_3, and p_4. Notice that process p_5 is a child of both process p_3 and process p_4. This implies that p_5 needs results from p_3 and p_4. By an argument similar to the one presented above, p_5 is considered distinct from both p_3 and p_4; also p_6 is distinct from p_2 and p_5. Finally process p_6 produces the output of the application program.

The execution of an application program is referred to as a job. In general a job consists of multiple processes organized in a graph as in figure 5. Thus such a graph is called a job graph.

3.5 Local CPU Scheduling

The function of the local CPU scheduler is three-fold. First it must schedule processes present at each site so as to contribute best to minimizing the sum of the real times to complete the jobs. Second it must classify jobs as either normal, CPU bound, or IO bound. This classification is used by the initial placement protocol presented below. Third it must estimate the loads presented to the local site by each of the three classes of jobs. Next we present the proposed local CPU scheduling algorithm. It performs all three functions.

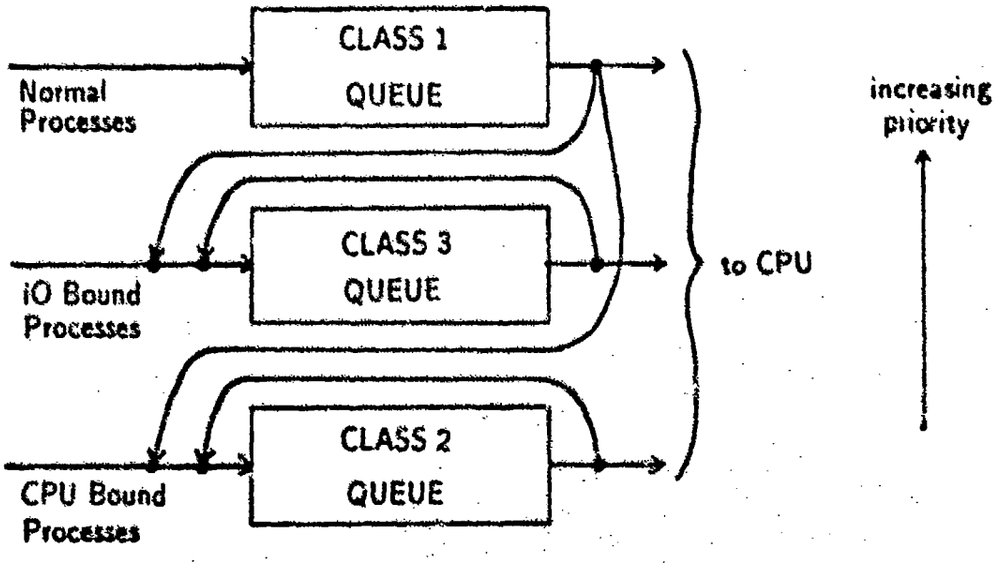


Figure 6. The local CPU scheduler.

As shown in figure 6 each of the processes arriving at a site is placed into one of three queues If the process belongs to a known IO job, it is placed in the class 3 queue. If it belongs to a known CPU bound job, it is placed in the class 2 queue. Otherwise, it is placed in the class 1 queue.

The CPU scheduling algorithm is round robin with a fixed time quantum T, multiple priority queues, and preemption. The CPU is allocated to a process for one time quantum at a time. If the process initiates an IO operation in the middle of a time quantum, the quantum terminates at that point and the next process' time quantum begins. This next process is taken from the class 1 queue, unless it is empty. In that case, it is taken from the class 3 queue, unless it is empty, in which case it is taken from the class 2 queue. Further, if while a class 2 or class 3 process is executing, a class 1 process arrives, the class 2 or 3 process is preempted.

Processes are transferred between the three queues in the following manner. Suppose a process is initially inserted in the class 1 queue. It is given I time quanta in the class 1 queue to complete. If it has not completed by that point, the process is classified as either IO bound or CPU bound. If the percentage of the first I time quanta which it completed with no IO operation causing an early time quantum termination is less than a tuneable parameter E, it is classified as IO bound. Otherwise, it is classified as CPU bound.

The justification for this choice of local CPU scheduling algorithm is the following. Recall that the function of our load sharing protocol is to minimize the sum of the real times required to complete each job. This function is reduced if longer jobs are delayed for

shorter jobs.

Thus we separate normal processes from CPU and IO bound ones, place them in the class 1 queue, and give them highest priority. Actually, we place both normal processes and processes of unknown class in the class 1 queue. This should cause little performance degradation, since the overwhelming majority of processes are normal [20]. But in any case, processes of initially unknown class which turn out to be either CPU or IO bound remain in the class 1 queue at most 1 time quanta.

We also separate IO bound processes from CPU bound processes and give the former higher priority, with the following justification. An executing IO bound process will almost surely not use an entire time quantum; instead it will request an IO operation and block long before the time quantum ends. At the point it does so, the next time quantum begins. Thus the typical CPU burst is of much shorter duration for an IO bound process than for a CPU bound one. Applying the conclusion stated above, namely that longer jobs should be delayed for shorter ones, on a much smaller time scale, we conclude that the CPU scheduling of CPU bound processes should be delayed in favor of IO bound processes.

The CPU scheduler maintains for each process m the total amount of service time it has received, including both CPU time and IO service time. This quantity at real time t is called the age of process m at time t and denoted $a_m(t)$.

The current set of process ages is used to estimate the current load on each processor contributed by each of the three classes, as follows. It has been shown [20,24] that the expected value of the remaining CPU time required by a known CPU bound process is more or less proportional to its current age. Thus we estimate the class 2 load on a

particular site at time t to be

$$\sum_m k_2 a_m(t)$$

where k_2 is some constant of proportionality and m ranges over all class 2 processes at the site in question.

We have seen no estimate of the remaining service time of IO bound processes as was made for CPU bound processes in Leland and Ott [20]. However, it seems very reasonable to extend their argument by estimating the remaining service time of an IO bound process to be proportional to its current age. The main difference then between the behavior of a CPU bound process and an IO bound process is that for the latter almost all the service time is used performing IO requests instead of CPU operations. In summary, we estimate the class 3 load on a particular site at time t to be

$$\sum_m k_3 a_m(t)$$

where k_3 is some constant of proportionality and m ranges over all class 3 processes at the site in question.

Finally, for normal processes any kind of load prediction based upon past behavior seems to be both difficult and pointless. So, mostly for completeness, we propose the following simple load estimate for class1 processes at a particular site.

$$k_1 P$$

where k_1 is some constant of proportionality and P is the number of normal processes at the site in question.

3.6 The Initial Process Placement Protocol

Recall that in section 3.3 we showed that a process should be an expression of parallelism. This leads us to the conclusion that no two processes from the same job should execute at the same site at the same time. So, whenever a set of processes is created, all but one should be moved to other processors. The moving of a process upon its creation is often called initial placement. We use this term throughout the remainder of this report.

The following question arises immediately: at which processor do we initially place a new process? We address this question next. It is well-known (see for example [8]) that a site is utilized most efficiently if its CPU bound load and IO bound load are balanced. In this case IO requests and CPU operations can most frequently be performed in parallel. In a computer system containing only one site these loads are not adjustable; instead, they are determined by the application programs submitted by the users. However, in a computer system with multiple sites CPU bound and IO bound processes can be moved from site to site so as to better balance the load distribution. In particular, we propose to initially place CPU bound and IO bound processes to achieve this goal.

But how do we know whether a process is normal, CPU bound, or IO bound before it starts executing? Of course, we do not. But it is reasonable to assume that the class of a process is determined by the nature of the problem it helps to solve. In other words, problems can be classified as inherently normal, computation intensive, or memory space intensive. Thus all the processes of the same job are expected to be of the same class. Once the first process of a job, for example process p_1 of the job whose graph is shown in figure 5, has been classified, the classes of all the other processes of the same job are assumed known. As we will see below, the misclassification of a process which does not satisfy this assumption has only slight consequences. Specifically, our initial process

placement protocol may misplace it. However, if all the other processes of the same job are well placed, the performance degradation should be small.

Our initial process placement protocol uses the normal, CPU bound, and IO bound load estimates which were presented in section 3.5. They are periodically broadcast by each site, as specified by the protocol. Thus our initial proces. placement protocol, performed by each site, is the following.

1. periodically broadcast the local loads of the three classes 1, 2, and 3.
2. if a job creates process at the local site,
   (a) if the job's class is unknown, assume it is 1 (normal).
   (b) let j denote the job's class.
   (c) choose, from among all sites with no process from the job in question, the site with the smallest class j load.
   (d) place the process at the site chosen in step c above.

3.7 Summary

This part of the report presented first an overview of the load sharing problem in multiprocessor computer systems. This included an extensive literature search and the presentation of the conclusions of those papers which proposed practical, heuristic load sharing solutions. Next we explained that in a multiprocessor system multiple processes executing the same application program exist only to express parallelism. This novel conclusion has an important implication for the design of a load sharing protocol: namely, that at most one process from a particular job should execute at any site at any time. We then defined

the job graph, which was used to display the relationships between different processes from the same job. Next, we specified the local CPU scheduling algorithm, which forms part of our solution. It had three functions: first, to schedule processes on the local CPU; second, to classify jobs as either normal, CPU bound, or IO bound; and third, to estimate the loads presented to the local site by each of the three classes of jobs. Finally, we presented the initial placement protocol, designed to implement load sharing in accordance with all our earlier conclusions.

4 INCREASING DATABASE ACCESS CONCURRENCY IN A DISTRIBUTED SYSTEM

4.1 Introduction

This part of the report addresses the problem of increasing database access concurrency in a distributed system. First we consider nested transactions as a possible solution. It has been claimed (for example, see Walpole, et. al. [38]) that this type of transaction increases database concurrency. In the following sections we show that this claim is false. We do this by first summarizing the classic nested transaction scheme presented in Moss [39]. Then we show that the parallel execution of the subtransactions of a nested transaction at different sites occurs no faster than the serial execution at one site. This is the case because of the large inter-site communication delays present in a distributed system. Next we show that the period of time that accessed data items are locked under a nested transaction is greater than under the serial version of the same transaction. Thus we conclude that nested transactions do not increase database access concurrency. Finally, we suggest that database access concurrency may be effectively increased by decreasing data item granularity.

Although we cannot prove it, we suspect that a similar argument holds for compound transactions, as proposed in Jensen, et. al. [40] and in Jensen, et. al. [41]. That is, we suspect that compound transactions also do not increase database access concurrency over that provided by serial transactions. The reason we cannot prove it is the lack of precise definition of what is meant by compound transactions, and especially what is meant by compensation.

4.2 Nested Transactions

4.2.1 Introduction

We begin by explaining what nested transactions are. First we present a serial transaction, written in terms of a possible set of operations. These operations are intended only to be examples of how thing may be done, not an actual implementation.

```
a_trans=proc(t:tid)
... actions with respect to t ...
end a_trans.
```

Here t is a tid, or transaction ID. When we wish to execute the transaction, we write a simple program such as

```
t:tid:=create_transaction();
a_trans(t);
commit(t);
```

Now suppose we wish to compose the transaction routine a_trans and another transaction routine, say b_trans, into a single transaction. We can do it as follows.

```
t:tid:=create_transaction();
a_trans(t);
b_trans(t);
commit(t);
```

This is fine. But now a_trans and b_trans run serially, as a single serial transaction. No concurrency is possible. However, with nested transactions, it is in principle possible to have concurrent transaction execution. We may execute a_trans and b_trans as a nested transaction in the following way.

```
t:tid:=create_transaction();
t1:tid:=create_transaction();
a_trans(t1);
commit(t1);
t2:tid:=create_transaction();
b_trans(t2);
commit(t2);
commit(t);
```

In the nested transaction both a_trans and b_trans are executed as lower level transactions and and may be executed in parallel at different sites. Here a_trans has tid t1 and b_trans has tid t2. The composition itself is a higher level transaction. It has tid t. The fact that the composition is expressed as a (higher level) transaction guarantees the atomicity of the composition. Thus, if either a_trans or b_trans aborted after the other had finished, the effects of the finished one would have to be undone.

In the previous paragraph, we mentioned that the lower level transactions of a nested transaction may be executed in parallel at different sites. This is true, but actually the following stronger statement can be made. There is no reason to divide a serial transaction into multiple lower level transactions encapsulated in a higher level transaction other than to allow the lower level transactions to execute in parallel on different sites. This is true

by the following argument. A serial transaction cannot commit until all its operations have been performed. But a nested transaction cannot commit until all its lower level transactions have committed; the lower level transactions in turn cannot commit until all their operations have been performed. Similarly, if a serial transaction aborts, all its effects must be undone. If any lower level transaction of a nested transaction aborts, the effects of all its other lower level transactions must be undone. Thus there is no other reason to divide a serial transaction up into a nested transaction other than to allow the lower level transactions to execute in parallel on different sites.

In the remainder of this report, we assume that each lower level transaction of a nested transaction is intended to execute at a particular site, where all the data items it reads or writes are located. In other words, we assume that the transaction programmer (whether a human or an automaton) specially constructed each lower level transaction to read or write data items located at only one site.

The example nested transaction of this section may be illustrated in the following way.

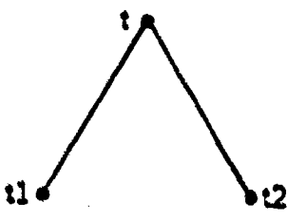


Figure 7. A tree representation of our example nested transaction.

In general, a nested transaction may be represented as a tree. Its depth may of course

be greater than one. Thus we may apply stancard tree terminology to a nested transaction. We may speak of leaf transactions, parent transactions, child transactions, inferior transactions, superior transactions, etc. We may also speak of the root transaction as a top-level transaction and any non-root transaction as a subtransaction.

We may now easily express in tree terminology the following assumption concerning nested transactions. Only leaf transactions read or write data items directly. Higher level transaction dc not; instead, they operate upon data item values which are read and written by leaf transactions. This assumption simplifies the lock inheritance rules, presented in section 4.2.3.

Finally we assume that each site has a transaction manager (TM) which manages all local transactions and can also create at a foreign site a transaction inferior to a local one. The home site of a nested transaction is the site where its top-level transaction is located.

This section has defined the concept of nested transactions and explained their potential advantages. The next section describes in detail the nested transaction facility proposed by Moss. Later we use this facility to investigate the increasing of concurrency in distributed systems.

4.2.3 Locking

Moss proposes the following locking rules, which define the meaning of locks and what is done with an inferior transaction's locks when it commits or aborts.

- a transaction may hold a lock if all other transactions holding a lock on the same data item are superiors of the first transaction.

- when a transaction aborts, all locks are simply discarded. If any superior had held a lock on the same data item, it continues to do so.
- when a transaction commits, all its locks are inherited by its parent, if any.

4.2.4 Transaction Commitment

The transaction commitment rules for nested transactions are the following.

- all of a transaction's children must be committed before the transaction itself can commit.
- all of a transaction's children need not be resolved before the transaction aborts.
- if a transaction aborts, all its inferiors must be aborted.

If a transaction commits, its TM

- sends to its parent's TM a COMMITTED message which informs the parent that it committed and contains a list of the committing transaction's committed inferiors. The parent's TM appends this list to his own list of committed inferiors.
- sends to the TM of each inferior transaction a COMMITTED message. Upon receiving it, the TM then discards all information concerning the inferior transaction. The parent of the committing transaction now has all the pertinent information about the inferior transaction.

Note that subtransaction commitment is provisional. If any superior transaction aborts, then the effects of the subtransaction are undone.

If a subtransaction aborts, its TM

- sends to its parent a TM a message which informs the parent that it aborted.

If a top-level transaction commits, its TM initiates the two-phase commitment protocol. In other words,

- its TM sends to all inferior transactions' TMs a PREPARE message.
- upon receiving a PREPARE message, a TM checks that crashes have not destroyed the inferior's data item modifications. Then it stores both the old and the modified data item values in permanent memory. The TM replies to the top-level TM with a PREPARED message after the data item values have been stored.
- after receiving a PREPARED message from all inferior transactions' TMs, the top-level TM records in permanent memory a notation that the transaction is completing. Then it sends a COMPLETE message to each inferior transaction's TM.
- after receiving a COMPLETE message, an inferior TM discards the old data item values and responds COMPLETED to the top-level TM.

4.2.5 Nested Transactions Do Not Allow Increased Concurrency.

In this section we show that Moss' nested transactions do not fulfill the promise of increased database access concurrency in distributed systems.

We assume that transaction concurrency is controlled by the two-phase locking protocol. In this protocol, a transaction's execution can be divided into two-phases: the first phase during which locks are set and the second phase during which locks are released, but no additional locks are set. In other words, a transaction satisfies the two-phase locking protocol if and only if it sets no lock after it releases the first one.

Nested transactions increase database access concurrency in distributed systems over that provided by serial transactions only if the transactions comprising the nested transaction can execute in parallel at different sites faster than they could execute in serial at one site.

We now show that nested transactions do not accomplish this. We begin by examining the inter-site communication required by an example transaction. The transaction modifies data items A and B assumed to be located at two different sites 1 and 2.

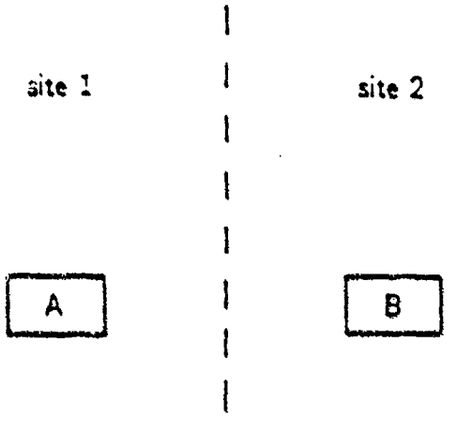


Figure 8. The data items accessed by our example transaction.

First we implement the transaction as a serial transaction and then as a nested transaction. We assume without loss of generality that the serial transaction executes at site 1. The following is the serial transaction.

```
a_trans=proc(t:tid)
read A;
read B;
A:=A+B;
```

```
B:=B-A;
write A;
write B;
end a_trans.
```

This transaction is executed at (site 1) via the following simple program.

```
t:tid:=create_transaction();
a_trans(t);
commit(t);
```

Now let us implement the same modifications of A and B via a nested transaction.

```
b_trans=proc(t2:tid)
read A;
A:=A+B;
B:=B-A;
write A;
end b_trans.

c_trans=proc(t3:tid)
read B;
end c_trans.

d_trans=proc(t4:tid)
write B;
```

```
end d_trans.
```

The nested transaction execution is begun at site 1 via the following program.

```
t1:tid:=create_transaction();
t2:tid:=create_transaction();
b_trans(t2);
commit(t2);
t3:tid:=create_transaction();
c_trans(t3);
commit(t3);
t4:tid:=create_transaction();
d_trans(t4);
commit(t4);
commit(t1);
```

Recall from section 4.2.1 that each subtransaction of a nested transaction is written so as to access data items at only one site and to execute there. Since c_trans and d_trans operate only upon data items at site 2, it is clear that they should be send to site 2. Now let us see if the execution of b_trans at site 1 and the concurrent execution of c_trans and d_trans at site 2 can be completed faster than the serial execution of a_trans at site 1.

Figure 9 shows a time line for messages passed between site 1 and site 2 during the execution of the serial transaction t. Figure 10 shows a time line for the nested transaction t1. We explain figure 9 next. First site 1 locks data item B in accordance with the two-phase locking protocol. Then it reads B. The value of B is returned from site 2 to site

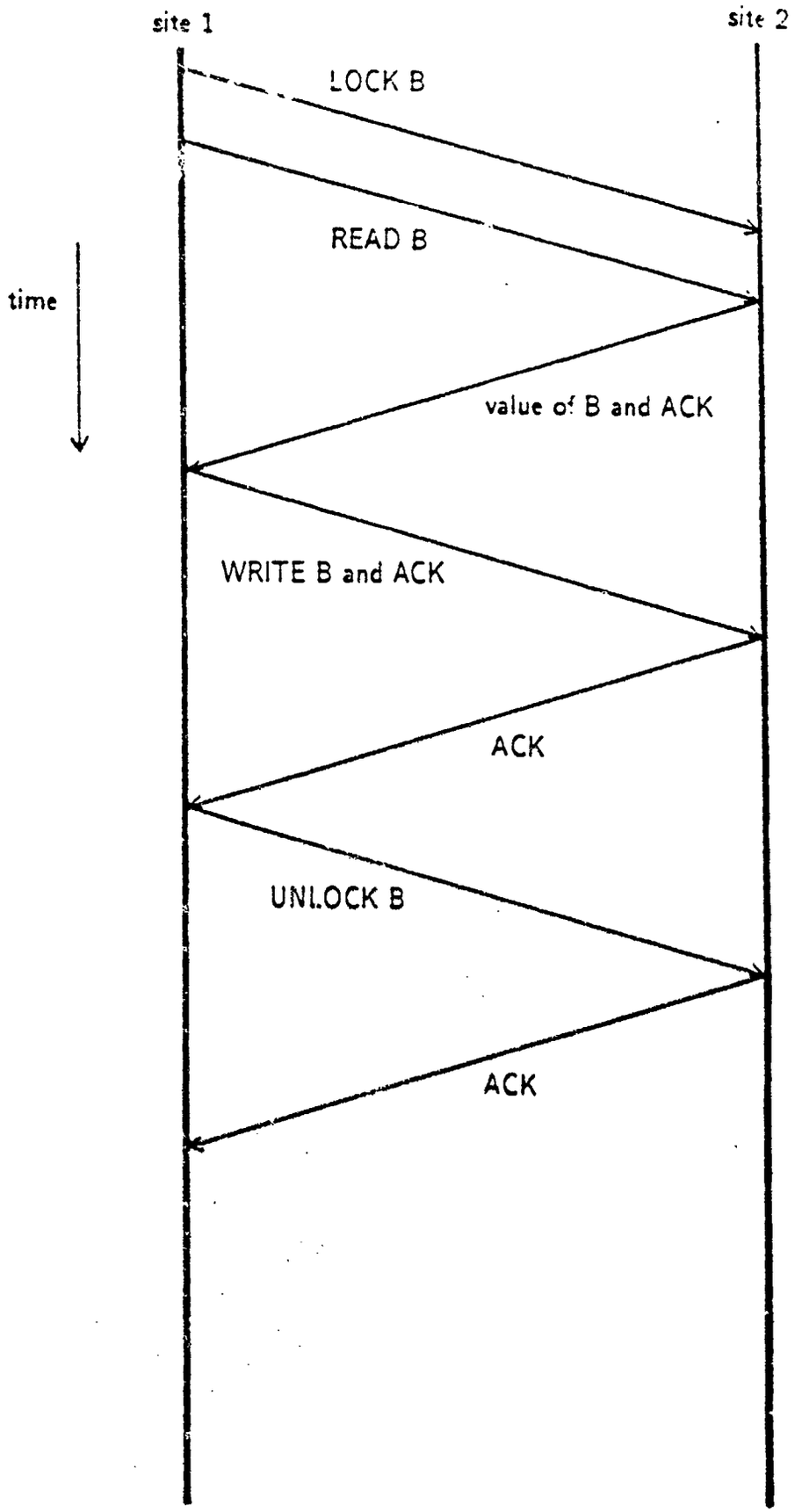


Figure 9. The time line for transaction t.

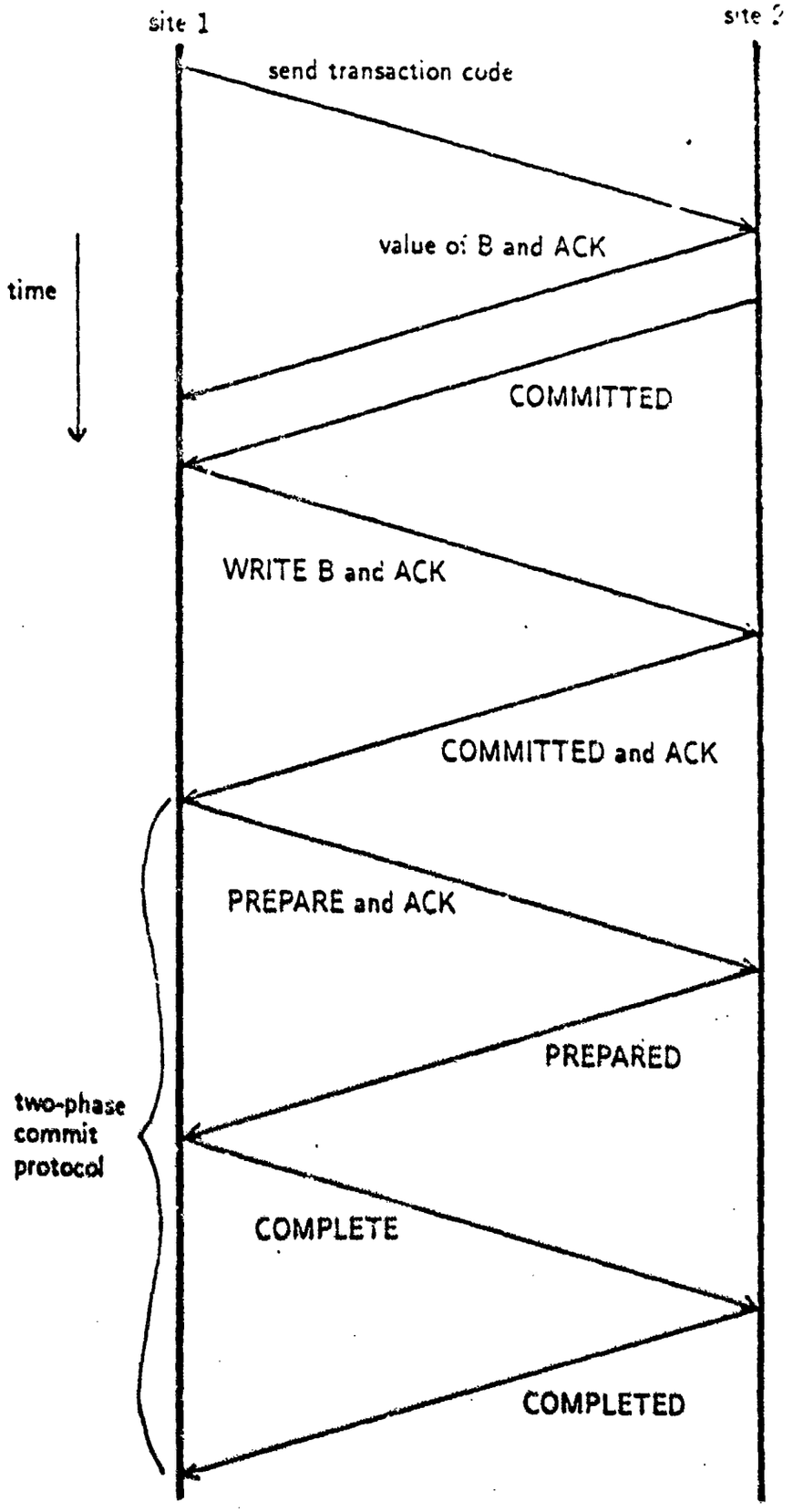


Figure 10. The time line for transaction t1.

1 with an ACK for the LOCK B message. Next site 1 writes the updated value of B to site 2 and piggybacks an ACK for the earlier message which sent B to site 1. Next site 2 ACKs the WRITE B message Then site 1 is sure that B has been updated at site 2 and unlocks B. Finally site 2 ACKs the UNLOCK message. The transaction t is now complete and can be committed.

Next we explain figure 10. First we assume that site 1 recognizes that the inferior transactions t3 and t4 belong at site 2, since they reference only variables located there. Thus, site 1 sends site 2 a message including the code of c_trans and d_trans. Next c_trans is executed at site 2. It reads B and sends the value to site 1. Accompanying that message is an ACK for the transaction code message. Next t3 commits and send a COMMITTED message to site 1, where its parent transaction t1 is located. Then transaction t2 modifies the value of B. The new B value is returned to site 2 along with an ACK for the COMMITTED message. Now transaction t4 writes the new B value at site 2 and returns a COMMITTED message to its parent, namely t1. Finally t1, which is a top-level transaction, starts to commit. It performs the two-phase commit protocol, as shown in figure 10.

Clearly, the nested transaction t1 takes a longer time to execute than the serial transaction t. Thus at least for this example, nested transactions do not increase database access concurrency. But is this true in general? Yes, by the following argument.

We compare the time required to execute a general serial transaction with the time to execute a nested transaction under the assumption that inter-site communication is much greater than computation time. In the serial case, at the beginning of the transaction's execution a LOCK followed by a READ is sent to each site with a data item read and modified by the transaction. The read data items are sent to the site where the transaction

is executing. An ACK for the READ message is piggybacked upon the read data items. It is best to read all needed variables initially, when the transaction begins execution, so that multiple locks and reads to the same site may be combined into a single LOCK message, a single READ message, and a single message containing the read data items. After modifying the data items, the transaction writes them back to their original sites, all of the writes destined for the same site having been combined into a single WRITE message. The transaction then waits for an ACK for each WRITE message. Upon receiving the ACK from a particular site, the transaction UNLOCKS all the data items it has locked at that site. Finally the transaction waits for an ACK for each UNLOCK message and then commits if everything went correctly. A total of six rounds of inter-site communication are required. In each round one message must be sent between the transaction site and each site storing data items modified by the transaction.

In the nested case, at the beginning of the transaction's execution, the code of each inferior transaction is sent to the site where it belongs. It modifies data items there and may even return values of local data items to its home site. If this is necessary, the performance of the nested transaction becomes even worse. Assuming this is not necessary, each inferior transaction sends the home site a COMMITTED message when it finishes; an ACK for the transaction code message is piggybacked upon it. After receiving COMMITTED messages from all inferior transactions, the top-level transaction commits by executing the two-phase commit protocol. This requires a PREPARE message, a PREPARED message, a COMPLETE message, and a COMPLETED message. Again six rounds of inter-site communication are required; each round again requires the transmission of a message between the home site and each site where an inferior transaction is executing.

In conclusion, six rounds of inter-site communication are required for both the serial and the nested versions of the transaction. Assuming that computation time is insignificant with respect to inter-site communication delay, nested transactions offer no advantage over serial transactions.

So far in our examination of the inter-site communication required by nested transactions we have considered only nested transactions whose tree representations have depth one. Is our conclusion true even for nested transactions whose trees have depth greater than one? Yes. As the tree depth increases, more COMMITTED messages must be sent from committing subtransactions to their inferiors. Many of these COMMITTED messages must be send between sites, greatly slowing the execution of the nested transaction.

Our comparison of nested and serial transactions is not complete until we examine the length of time each type of transaction holds its accessed data items locked. We will see that nested transactions actually hold their locks longer than serial transactions and so our earlier result is reinforced. Thus other transactions must wait longer to access the locked data items in the nested case.

As can be seen from an examination of figures 9 and 10, the serial version holds its locks on the data items it accesses at site 2 from the time that the LOCK message arrives until the time that the UNLOCK message arrives. The nested version holds its locks from the time that the subtransaction code reaches site 2 until the time that the COMPLETE message arrives. The latter time interval is greater than the former. Thus, finally we may conclude that nested transactions do indeed not provide increased database access concurrency in a distributed system.

4.3 Increased Database Access Concurrency Via Decreased Data Item Granularity.

The previous sections of this report showed that nested transactions provide no increase in database access concurrency over serial transactions. As mentioned earlier, we suspect that the same is true for compound transactions. Thus, how do we increase database access concurrency? One approach is to decrease data item granularity - that is, to allow smaller sections of memory to be locked. For example, instead of choosing a data item granularity equal to a page, choose a granularity equal to half a page. This approach is both simple and effective, in contrast to the nested transaction approach, which is complex and ineffective.

4.4 Summary

This part of the report addressed the problem of increasing database access concurrency in a distributed system. In short, we showed that nested transactions do not solve the problem because of long inter-site communication delays. We explained that we suspect compound transactions also do not, although the term compound transaction is not well-defined enough to prove this suspicion. Finally we suggested that decreased data item granularity is a simple and effective solution to the problem.

5. PROTOCOLS FOR OUR PROPOSED DISTRIBUTED OPERATING SYSTEM

5.1 Introduction

This section of the report puts the finishing touches on the problem addressed by the 1986 Summer Faculty Research Program (SFRP) work: namely, to determine in detail which services should be supported by protocols for inter-site DOS communication. A slightly modified version of the final report of that work is attached as an appendix. This version was presented at the October 1987 International Federation for Information Processing Conference on Distributed Computing in Amsterdam, The Netherlands. The problem addressed in that report is the identification of the inter-site communication needs of a DOS. The problem was approached by proposing a simple but typical DOS and determining the general communication services it requires - for example, send a datagram, broadcast a datagram, establish a virtual circuit, etc. These terms are explained in the SFRP report. The present report identifies the specific communication services a typical DOS requires, in the following way. First, we define the specific services being considered. Then we specify for each of the inter-site communications required by our proposed DOS which specific services are needed. Finally, we show that the services may be layered so that each inter-site DOS communication will find the services it needs by entering at an appropriate layer and accessing only that and lower layers. We specify the entry layer for each inter-site DOS communication.

5.2 The Specific Communication Services

The final report from our SFRP work identifies the general inter-site communication services required by our proposed typical DOS. From a careful examination of that report, enclosed as an appendix to the present report, one can identify the specific communication

services required. These specific communication services, which are necessary for some communications and unnecessary for others, are

GUARANTEED PACKET ARRIVAL

RESEQUENCING AND DUPLICATE REMOVAL

ERROR CORRECTION CODING

The GUARANTEED PACKET ARRIVAL service ensures that transmitted packets eventually arrive at their destinations with no errors. One implementation of this service is to acknowledge received packets and retransmit unacknowledged packets after a time-out. The RESEQUENCING AND DUPLICATE REMOVAL service reorders and discards packets, if necessary, at their destinations so that one copy of each packet remains and so that the packets are arranged in the same order as that in which they were transmitted. The ERROR CORRECTION CODING services adds redundancy to transmitted packets so that possible errors may be corrected at the destination. This service decreases the delay required to correctly receive a packet which experiences errors on the route to its destination by eliminating the need to wait for the acknowledgement time-out period and the delay caused by retransmitting the erroneous packet. We apply the ERROR CORRECTION CODING service to delay-sensitive communications.

5.3 The Required Specific Inter-Site DOS Communication Services

In this section we list for each of the inter-site communications required by our proposed DOS the necessary general and specific communication services. The required communications are listed in the same order as in the SFRP report and are grouped according to the DOS manager performing the communication. The format of the list is the following. First, the communication is listed, followed by a colon. Next comes the general

communication service it requires. Next comes a list of the three specific communication services and whether each is necessary for the communication being considered. Finally, comes the Entry layer number. This number is explained in section 5.4.

The Specific Inter-Site Communication Services Required by the Security Manager

To increase a classification: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 5

To decrease a classification: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

To migrate a data entity: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

The Specific Inter-Site Communication Services Required by the Resource Manager

To access a foreign resource: send a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 7

To broadcast resource loads: establish virtual circuits to all other sites

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 4

To detect deadlocks: send a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

To elect controllers: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 5

To elect a process to be preempted: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 5

The Specific Inter-Site Communication Services Required by the Entity Manager

To locate the closest replication of a data entity: broadcast a datagram

ERROR CORRECTION CODING necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 7

To read a foreign data entity: send a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 4

To create a foreign data entity: send a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 7

To copy a data entity to a particular foreign site: establish a virtual circuit

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 7

To delete a data entity at a particular foreign site: send a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

To delete all replications of a data entity: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

To create a process at a foreign site: send a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 7

To terminate a process at a foreign site: send a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

The Specific Inter-Site Communication Services Required by the Database Manager

To broadcast a lock for a data entity: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

To broadcast a data item update: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 5

To broadcast an unlock: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 6

To broadcast one datagram for each data entity after system reconnection: broadcast a datagran

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 5

To exchange journals: establish virtual circuits

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 7

To copy all replications of a data entity to a common site: establish virtual circuits

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL necessary

GUARANTEED PACKET ARRIVAL necessary

Entry layer 7

The Specific Inter-Site Communication Services Required by the Fault Manager

To broadcast checkpoints: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 5

To elect a rollback site: broadcast a datagram

ERROR CORRECTION CODING not necessary

RESEQUENCING AND DUPLICATE REMOVAL not necessary

GUARANTEED PACKET ARRIVAL not necessary

Entry layer 5

5.4 A Layering of the Inter-Site DOS Communication Services

The previous section presented the inter-site communication services required by a DOS. The present section presents a layering of these services. This layering has the property that all DOS communications are able to find the communication services they require by entering at an appropriate layer and accessing only that layer and lower layers. The previous section anticipated the results of the present section by listing the entry layer for each inter-site DOS communication.

Figure 11 shows the service layering. There are actually two layering: one provides datagram connections and the other provides virtual circuits. Virtual circuits are not layered upon datagram connections because virtual circuits can be implemented more efficiently directly upon the routing service. To justify this claim we must first explain the function of the ROUTING layer. It determines the channel by which packets leave each site. It sends those packets flowing along a virtual circuit out on the one outgoing channel lying on the virtual circuit. It sends datagram packets out on a channel chosen dynamically according to some routing technique. Thus, the routing service required by datagrams differs fundamentally from that required by virtual circuits. For this reason and others, it is more sensible to have separate datagram and virtual circuit layers than

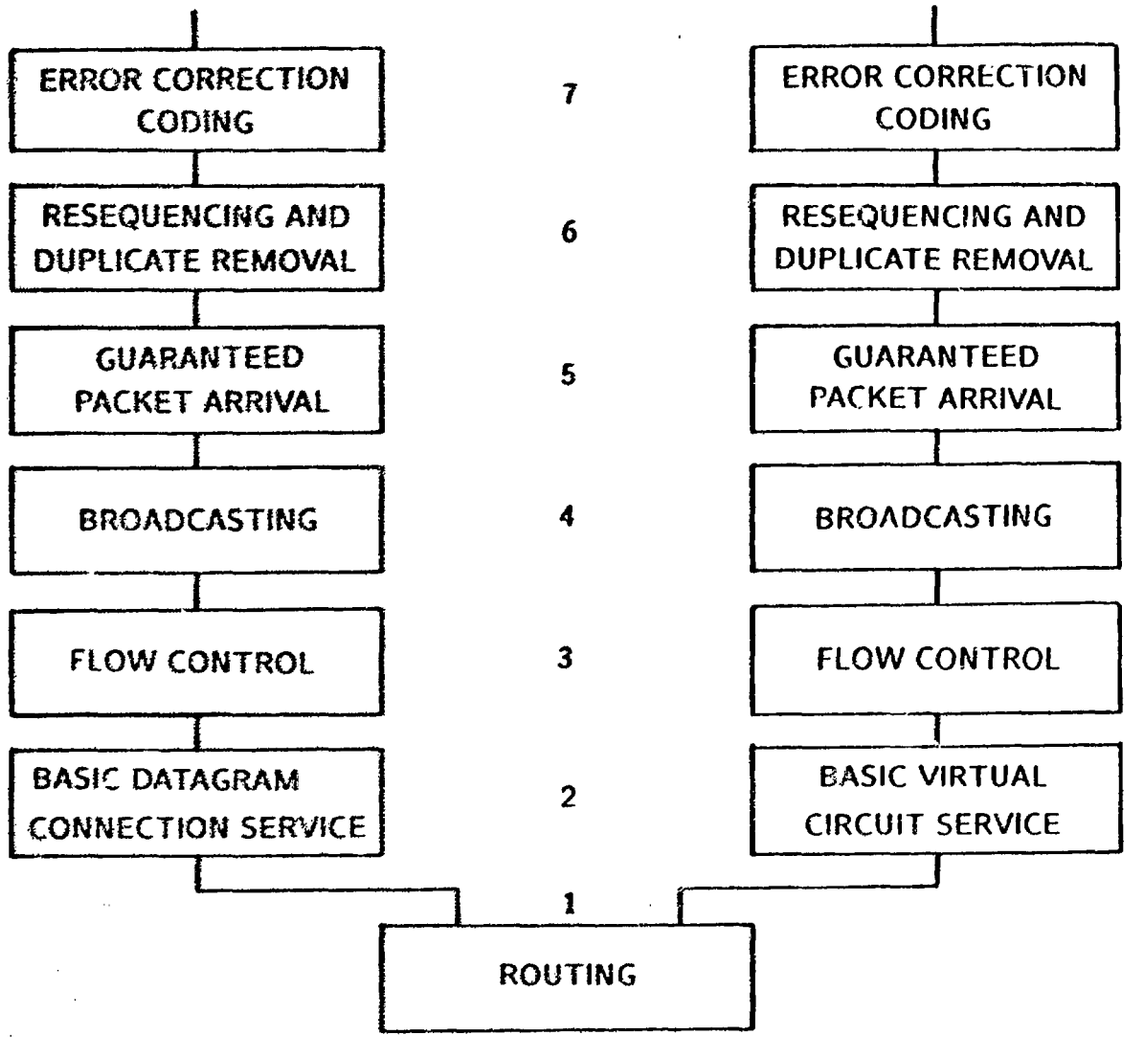


Figure 11. A layering of inter-site communication services.

to layer virtual circuits upon datagram connections.

We next describe the remaining layers. The BASIC DATAGRAM CONNECTION SERVICE layer provides datagram connection establishment and termination as well as packet transmission over a datagram connection. Similarly the BASIC VIRTUAL CIRCUIT SERVICE layer provides virtual circuit establishment and termination as well as packet transmission over a virtual circuit.

The FLOW CONTROL layer provides a means for the destination to govern the amount of data sent by the source. This may be accomplished in the following way. First a sequence number is associated with each packet. This sequence number should be unique over all packets sent over the same datagram connection or virtual circuit. But periodic sequence numbers provide satisfactory performance for large enough periods. Second, a "window" is returned with every acknowledgement indicating the maximum sequence number that the source may send before receiving further permission.

The BROADCASTING layer provides for the transmission of a packet (via either datagram connections or virtual circuits) to all other sites.

The remaining layers provide what we earlier referred to as specific communication services. These services were explained in section 5.2.

5.5 Summary

This part of the report finished the work started in the 1986 SFRP project. In that report, we determined the inter-site communications required by a simple but typical DOS. For each communication we identified the general communication services needed - such as send a datagram, broadcast a datagram, established a virtual circuit, etc. In the present report we identified the specific communication services required, such as error correction

coding, resequencing and duplicate removal, and guaranteed packet arrival. We presented a layering of communication services so that each inter-site DOS communication finds the services it needs by entering at an appropriate layer and accessing only that and lower layers. Finally, we specified the entry layer for each inter-site DOS communication.

6 SUMMARY

This document presented the research conducted by Professor Craig G. Prohazka under the US Air Force Minigrant Program.

The work addressed four problems in the design of protocols for communication between instances of a distributed operating system (DOS) running at different sites. The first was the design of distributed synchronization problem protocols. We solved three such problems: the termination detection problem, the mutual exclusion problem, and the distributed bounded buffer producer/consumer problem. The first two of these have earlier been examined by other researchers. Our protocols outperform theirs in several ways, including delay and required number of inter-site messages. The second area was the design of multiprocessor load sharing protocols. We proposed a new load sharing protocol based upon previous researchers' work and some novel thoughts on parallel user process behavior. Then we showed that, contrary to the claims of other researchers, nested transactions do not provide increased database access concurrency in distributed systems. We then proposed a simple but effective technique to increase this concurrency: decreasing data item granularity. Finally, the fourth area was the completion of the 1986 Summer Faculty Research Program work. That work identified the general inter-site communication services required by a DOS. The present report identified the specific communication services required by the DOS. Then we layered all the communication services so that each inter-site DOS communication will find the services it need by entering at an appropriate layer and accessing only that and lower layers.

7 REFERENCES

1. R. W. Topor, "Termination detection for distributed computations", Information Processing Letters, vol. 18, pp. 33-36, January 1984.

2. B. A. Sanders, "A method for the construction of probe-based termination detection algorithms", International Federation on Information Processing Conference on Distributed Processing, October 1987.

3. N. Francez and M. Rodeh, "Achieving distributed termination without freezing", IEEE Transactions on Software Engineering, vol. 8, no. 3, pp. 287-292, May 1982.

4. B. Szymanski, Y. Shi, and N. Prywes, "Synchronized distributed termination", IEEE Transactions on Software Engineering, vol. 11, no. 10, pp. 1136-1140, October 1985.

5. E. W. Dijkstra and C. S. Scholten, "Termination detection for diffusing computations", Information Processing Letters, vol. 11, no. 1, pp. 1-4, August 1980.

6. T. V. Lakshman and A. K. Agrawala, "Efficient decentralized consensus protocols", IEEE Transactions on Software Engineering, vol. 12, no. 5, pp. 600-607, May 1986.

7. E. W. Dijkstra, W. H. J. Feijen, and A. J. M. van Gasteren, "Derivation of a termination detection algorithm for distributed computations", Information Processing Letters, vol. 16, pp. 217-219, June 1983.

8. J. L. Peterson and A. Silberschatz, Operating System Concepts. Reading, MA: Addison-Wesley, 1985.

9. G. Ricart and A. K. Agrawala, "An optimal algorithm for mutual exclusion in computer networks", Communications of the ACM, vol. 24, no.1, pp. 9-17, January 1981.

10. E. Chang and R. Roberts, "An improved algorithm for decentralized extrema-finding in circular configurations of processes", Communications of the ACM, vol. 22, no. 5, pp. 281-283, May 1979.

11. E. Chang, "On message passing in computer networks", Communications of the ACM, vol. 24, no. 7, July 1981.

12. O. S. F. Carvalho and G. Roucairol, "On mutual exclusion in computer networks", Communications of the ACM, vol. 26, no. 2, pp. 146-148, February 1983.

13. D. S. Hirschberg and J. B. Sinclair, "Decentralized extrema-finding in circular configurations of processors", Communications of the ACM, vol. 23, no. 11, pp. 627-628, November 1980.

14. R. W. Conway, W. L. Maxwell, and L. W. Miller, Theory of Scheduling. Reading, MA: Addison-Wesley, 1967.

15. H. S. Stone, "Multiprocessor scheduling with the aid of network flow algorithms", IEEE Transactions on Software Engineering, vol. 3, no. 1, pp. 85-93, January 1977.

16. W. W. Chu, L. J. Holloway, M. T. Lan, and K. Efe, "Task allocation in distributed data processing", Computer, vol. 13, pp. 57-69, November 1980.

17. P. R. Ma, E. Y. S. Lee, and M. Tsuchiya, "A task allocation model for distributed computing systems", IEEE Transactions on Computers, vol. 31, pp. 41-47, January 1982.

18. K. Efe, "Heuristic models of task assignment scheduling in distributed systems", Computer, vol. 15, pp. 50-56, June 1982.

19. D. L. Eager, E. D. Lazowska, and J. Zahorjan, "A comparison of receiver-initiated and sender-initiated adaptive load sharing", Performance Evaluation Review, vol. 13, no. 2, pp. 1-3, August 1985.

20. W. E. Leland and T. J. Ott, "Load-balancing heuristics and process behavior", Proceedings of Performance '86 and ACM SIGMETRICS 1986 Joint Conference on Computer Performance Modeling, Measurement, and Evaluation, vol. 14, no. 1, pp. 54-69, May 1986.

21. A. Barak and O. G. Paradise, "MOS - A load-balancing UNIX", EUUG Autumn 1986 Conference Proceedings, pp. 273-280, September 1986.

22. A. Barak and A. Shiloh, "A distributed load-balancing policy for a multicomputer", Software - Practice and Experience, vol. 15, no. 9, pp. 901-913, September 1985.

23. Y. T. Wang and R. J. T. Morris, "Load sharing in distributed systems", IEEE Transactions on Computer, vol. 34, no. 3, pp. 204-217, March 1985.

24. L. F. Cabrera, "The influence of workload on load balancing strategies", Proceedings of the Usenix Technical Conference, pp. 446-458, Summer 1986.

25. W. E. Leland and T. J. Ott, "Unix process behavior and load balancing among loosely-coupled computers", International Seminar on Teletraffic Analysis and Computer Performance Evaluation, pp. 191-208, June 1986.

26. F. C. H. Lin and R. M. Keller, "Gradient model: a demand-driven load balancing scheme", 6th International Conference on Distributed Computing Systems Proceedings, pp. 329-336, May 1986.

27. L. M. Ni, C. W. Xu, and T. B. Gendreau, "A distributed drafting algorithm for load balancing", IEEE Transactions on Software Engineering, vol. 11, no. 10, pp. 1153-1161, October 1985.

28. L. M. Ni and K. Hwang, "Optimal load balancing strategies for a multiple processor system", Proceedings of the 1981 International Conference on Parallel Processing, pp. 352-357, 1981.

29. M. J. Gonzales, Jr., "Deterministic processor scheduling", Computer Surveys, vol. 9, no. 3, pp. 173-204, September 1977.

30. A. Hac and T. J. Johnson, "A study of dynamic load balancing in a distributed system", ACM Symposium on Communications Architectures and Protocols, vol. 16, no. 3, pp. 348-356, August 1986.

31. C. E. McDowell and W. F. Appelbe, "Processor scheduling for linearly connected parallel processors", IEEE Transactions on Computer, vol. 35, no. 7, pp. 632-638, July 1986.

32. J. A. Stankovic, "Stability and distributed scheduling algorithms", IEEE Transactions on Software Engineering, vol. 11, no. 10, pp. 1141-1152, October 1985.

33. J. Barhen and E. C. Halbert, "Roses: an efficient scheduler for precedence-constrained tasks on concurrent multiprocessors".

34. C. C. Shen and W. H. Tsai, "A graph matching approach to optimal task assignment in distributed computing systems using a minimax criterion", IEEE Transactions on Computers, vol. 34, no. 3, pp. 197-203, March 1985.

35. C. V. Ramamoorthy, K. M. Chandy, and M. J. Gonzalez, Jr., "Optimal scheduling strategies in a multiprocessor system", IEEE Transactions on Computers, vol. 21, no. 2, pp. 137-146, February 1972.

36. J. A. Stankovic, K. Ramamritham, and S. Cheng, "Evaluation of a flexible task scheduling algorithm for distributed hard real-time systems", IEEE Transactions on Computers, vol. 34, no. 12, pp. 1130-1143, December 1985.

37. A. M. van Tilborg and L. D. Wittie, "Wave scheduling - decentralized scheduling of task forces in multicomputers", IEEE Transactions on Computers, vol. 33, no. 9, pp. 835-844, September 1984.

38. J. Walpole, G. S. Blair, D. Hutchison, and J. R. Nicol, "Transaction mechanisms for distributed programming environments", Software Engineering Journal, pp. 169-177, September 1987.

39. J. E. B. Moss, "Nested transactions: an approach to reliable distributed computing", Ph.D. Dissertation, Technical Report 260, Massachusetts Institute of Technology, Cambridge, MA, April 1981.

40. E. D. Jensen, H. Tokuda, J. W. Wendorf, R. Wendorf, A. M. van Tilborg, "System/Subsystem description of ArchOS (Archons Operating System)", Department of Computer Science, Carnegie-Mellon University, April 1986.

41. E. D. Jensen, H. Tokuda, R. Clark, C. D. Locke, "Functional description of ArchOS (Archons Operating System)", Department of Computer Science, Carnegie-Mellon University, October 1985.

Appendices can be obtained from

Universal Energy Systems, Inc.

FINAL REPORT

Air Force Office of Scientific Research

RESEARCH INITIATION PROGRAM

Conducted by

Universal Energy Systems

Tunable Infrared to Visible Light Conversion in Rare Earth and Transition Metal Doped Fluoride Glasses

UES Project 760
S-760-6MG-042

Contract # F49620-85-C-0013/SSB5851-0360

R. S. Quimby

Department of Physics, Worcester Polytechnic Institute
Worcester, MA 01609
(617) 793-5490

SUMMARY

The overall goal of this project is to explore the possibility of tunable infrared to visible upconversion in a heavy metal fluoride glass (HMFG) co-doped with rare earth and transition metal ions. The transition metal ion provides the "tunable absorption" of infrared light, while the rare earth ion provides the upconversion mechanism via ion-ion interaction. During this grant period, progress has been made on the second step of this process, namely , upconversion via ion-ion interaction between rare earth ions. In particular, the absolute upconversion efficiency for Er^{3+} ions doped in fluorozirconate glass (ZBLA) has been determined. The emphasis of these experiments is to characterize the efficiencies quantitatively, and to determine the fundamental energy transfer parameters which will allow the effect of upconversion in other experimental situations to be determined. The results of these experiments have implications not only for tunable IR to visible upconverters, but also for the operation of rare earth doped solid state lasers. This second application has been a major focus of the research to date (see attached conference abstract). Particular attention has been paid to fiber lasers, and theoretical models of fiber lasers and amplifiers have been developed during this grant period. In the following paragraphs a more detailed discussion of these research results is presented.

ABSOLUTE UPCONVERSION EFFICIENCY

The research to date has focused on upconversion in the system Er:ZBLA glass. There are actually two distinct upconversion processes relevant in this sytem. In the first process, two Er ions in the $^4I_{11/2}$ level interact, yielding an ion in the ground state and one in a higher lying state. Green fluorescence is observed from the higher state ($^4S_{3/2}$). The second upconversion process involves two Er ions initially both in the $^4I_{13/2}$ excited state, interacting to leave one in the ground state and one in the $^4I_{9/2}$ exited state. The first upconversion process has been of primary interest in the experiments to date. Absolute efficiencies were measured using a calibrated integrating shere, with optical filters to block the pump light and pass the green upconverted fluorescence. A tunable dye laser in the range 647-670 nm was used as the pump source, exciting the $^4F_{9/2}$ level. The $^4I_{11/2}$ level is efficiently populated in this way, and upconversion from the $^4I_{11/2}$ level can be determined. Absolute efficiency is defined here as the ratio of emitted power in the green to absorbed power in the red. The absorbed power was determined from the measured incident power and the absorption coefficient measured on a spectrophotometer. With 48 mW of pump light at 652 nm focused to a spot of diameter 50 μm, the absolute upconversion efficiency was measured to be 0.5%. It was

verified that the upconverted fluorescence varied as the square of the pump beam power, as expected for a two step upconversion process.

There are actually two distinct upconversion mechanisms that could be responsible for the observed fluorescence, and it was necessary to distinguish between them before making any conclusions about microscopic energy transfer parameters. The first process is energy transfer upconversion (ETU), which involves energy transfer between two excited Er ions. The second process is excited state absorption (ESA), wherby a single excited Er ion absorbs a second pump photon. In both of these processes the fluorescence intensity is proportional to the square of the pump power. These two processes can be distinguished however, by the way that the upconversion efficiency depends on pump wavelength. In the case of ETU the fluorescence is proportional to the square of the ground state pump cross section, whereas for ESA the fluorescence is proportional to the product of the ground state pump cross section and ESA cross section. By tuning the dye laser over the range 647-670 nm, it was found that the mechanism of upconversion is ETU. This conclusion was further confirmed by observing that the green emission decays with a lifetime of 3 ms, approximately half that of the $^4I_{11/2}$ level. This is expected for ETU since the fluorescence signal is proportional to the square of the population of the $^4I_{11/2}$ level. Lifetime measurements were also made pumping the $^4I_{9/2}$ level at 799 nm. In this case as well, the dominant upconversion process was found to be ETU. Still another way

to distinguish between ESA and ETU is through the concentration dependence of the upconversion efficiency. For ETU the upconversion efficiency should be proportional to the concentration, whereas ESA should result in a concentration-independent efficiency. This has important implications for fiber laser performance, as will be discussed in the following section. Measurements on the concentration dependence of the efficiency are now in progress.

With the mechanism of the observed upconversion firmly established, microscopic energy transfer parameters can be determined from the measured absolute upconversion efficiency. The probability per unit time that an Er ion in the $^4I_{11/2}$ level decays via an upconversion process is given by αN_{ex}, where N_{ex} is the number of Er ions in the excited state $^4I_{11/2}$, and α is the energy transfer parameter. Using the measured absolute upconversion efficiency , the parameter α is calculated to be 2.5×10^{-19} cm^3/s. This parameter can then be used in the rate equations to predict the effect of upconversion in a variety of experimental situations. This will be further discussed in the following section.

EFFECT OF UPCONVERSION ON Er 2.7 μm LASER

Lasing on the $^4I_{11/2}$ - $^4I_{13/2}$ transition in Er^{3+} has been observed in a number of oxide and fluoride crystals. However, lasing has not been acheived in a glass to date, since in oxide based glasses the nonradiative relaxation rate between these levels is very large. Fluoride glasses, on the other hand, are characterized by relatively small nonradiative rates, and lasing between these levels should be efficient. In a collaboration with W. J. Miniscalco at GTE labs, the PI has attempted to achieve lasing in an Er-doped fluoride glass fiber. The doping level was 1% Er and the fluoride glass used was ZBLAN. To date, lasing has not been achieved, due to very high scattering losses in the fiber. In these experiments, however, a very bright green fluorescence was observed, due to upconversion among the Er ions. The question arose as to the effect of this upconversion on laser action at 2.7 μm, since upconversion in this case depletes the population of the upper laser level. The results of the absolute upconversion efficiency measurements can now be used to answer this question.

As an example, consider a 5 μm core single mode ZBLA fiber doped with 1% Er and pumped at 800 nm. With the effects of upconversion included in the rate equations, it is found that the upper laser level population is reduced by approximately 10% when the incident pump power is 5 mW. At

this pump power level, however, saturation of the level populations is already occuring, and it is this saturation rather than upconversion which will be important in limiting the upper laser level population. The effect of upconversion can be reduced even further by lowering the concentration of Er ions, since the upconversion mechanism is now known to be ETU. It is concluded, then, that upconversion from the $^4I_{11/2}$ level will not be a limiting factor in obtaining laser action at 2.7 μm in Er doped ZBLA fiber lasers.

FIBER LASERS AND AMPLIFIERS

The PI has been collaborating with W. J. Miniscalco and L. J. Andrews at GTE labs in the development of fluoride glass fiber lasers, in work related to the goal of this project. We have succeded in demonstrating laser action in a Nd^{3+} doped fluoride glass fiber at 1.3 μm (see attached preprint). This wavelength region has proved difficult in oxide glass hosts due to excited state absorption. Also in collaboration with W. J. Miniscalco at GTE, the PI has developed theoretical models for fiber lasers and amplifiers, which allow calculation of laser thresholds, output vs input, and efficiencies. (see attached conference abstract). Both three and four level systems are treated, and the effects of pump and signal saturation are included. Analytical expressions are derived in the case of the fiber

laser oscillator. For the fiber amplifier it is necessary to numerically integrate the differential equations to obtain the signal gain. These results should prove valuable in optimizing fiber laser and amplifier performance, especially in the case of three level systems where saturation of the level populations is essential for operation of the laser.

FUTURE WORK

The focus so far has been on determining the upconversion efficiency out of the $^4I_{11/2}$ level. It is also important to understand the degree of upconversion out of the $^4I_{13/2}$ level, since this has implications both for laser operation at 2.7 μm and for tunable IR to visible upconversion. This will be the near term goal of future research on this project. With the upconversion energy transfer parameters determined, it will be possible to include upconversion processes in the theoretical models for fiber lasers and amplifiers, to see how upconversion effects the optimum fiber length, for example. The dependence of upconversion on Er concentration will also be studied, to see if there is an optimum doping level for fiber lasers and for the tunable upconverters.

After upconversion in the Er only doped fluoride glass is well understood, the project will move into the second phase. The work here will concentrate on the energy transfer from a transition metal ion such as Ni^{3+} to a rare earth ion such as Er^{3+}. The efficiency of emission of the transition metal ion will be important in the overall process, and initially a Ni only doped fluoride glass sample will be studied, and partially re-crystallized to create a locally crystalline environment for the Ni ion. The fluorescence lifetime will be measured for samples devitrified to various degrees, in an attempt to obtain a long lifetime (hence high efficiency). The project will

conclude with the measurement of absolute upconversion efficiencies in the co-doped system Ni,Er:fluoride glass, along with other transition metal-rare earth combinations.

Appendices I-III can be obtained

from Universal Energy Systems, Inc.

Optimal Structural Modifications to Enhance the Robustness of Actively Controlled Large Flexible Structures

Singiresu S. Rao
Associate Professor

and

Tzong-Shii Pan
Graduate Research Assistant

School of Mechanical Engineering
Purdue University
West Lafayette, IN 47907

submitted to

Universal Energy Systems
Dayton, OH 45432
(Contract No. F49620-85-C-0013/SB5851-0360, Purchase Order No. S-760-6MG-036)

February 1988

1. Introduction

There has been a dramatic increase in the past decade in using active control systems to improve structural performance [1-4]. The major challenge in the field of active control of structures is in the design of control systems for very large space structures. Because of the high cost of lifting mass to orbit, there is a great incentive in making these structures light weight (and therefore flexible). Large space structures are by nature distributed parameter systems with multiple inputs (controls) and a continuum of outputs (displacements). The finite element method is commonly used for the description of these structures. This is a source of parameter errors and truncated (or reduced order) medels in the system. In addition, the structural properties of large space structures cannot be tested before they are put into orbit and hence sizable uncertainties exist in modal parameters.

A great deal of research is currently in progress on developing methods for the simultaneous (integrated) design of the structure and the control system. The weight of the structure was minimized with constraints on the distribution of the eigenvalues and/or damping ratio of the closed loop system by Khot, Eastep and Venkayya [5]. Salama, Hamidi and Demsetz [6] and Miller and Shim [7] considered the simultaneous minimization, in structural and control variables, of the sum of structural weight and the infinite horizon linear regulator quadratic control cost. The structure/control system optimization problem was formulated by Khot, et. al [8] with constraints on the closed loop eigenvalue distribution and the minimum Frobenious norm of the control gains. A unified algorithm for sequential (or simultaneous) design modifications of a closed loop constant gain control system and the flexible structure to be controlled

was presented by Junkins, Bodden and Turner [9]. It can be seen that in all the above works, the consideration of robustness of the control system has been ignored.

The parameter variations introduced by the analysis model, uncertain material properties or optimization may adversely influence the stability and performance characteristics of the control system. The robustness is an extremely important feature of a feedback control design. A robust control design is one which satisfactorily meets the system specifications even in the presence of parameter uncertainties and other modelling errors. Since the system specifications could be in terms of stability and/or performance, two types of robustness, namely, stability robustness and performance robustness, are to be considered in the design stage.

The current published literature on control system robustness addresses either the stability robustness aspect or the performance robustness aspect. Most of the work on the stability robustness (in the control area) was done in the frequency domain using singular value decomposition while much of the useful research on performance robustness was carried out in time domain using sensitivity approaches. Design studies that treated the stability robustness aspect in time domain and studies which combined both stability robustness and performance robustness in the design process have been scarce.

The recent developments in the area of robust multivariable control theory have been summarized by Ridgely and Banda [10]. The stability robustness of linear systems was analyzed in the time domain in [11,12] wherein a bound on the perturbation of an asymptotically stable linear system was obtained to maintain stability using Lyapunov matrix equation solution. In Ref. [13], singular value robustness measures were used to

compare the performance and stability robustness properties of different control design techniques in the presence of residual modal interaction for a flexible spacecraft system. The importance of robustness considerations in the design of flexible space structures was discussed by Hoehne [14].

Gordon and Collins [15] presented a direct design method for solving the problem of robustness to cross-coupling perturbations by treating the feedback gains as design variables. Their method makes use of nonlinear programming techniques along with a time domain pole placement procedure. The time domain stability robustness analysis for time varying perturbation using Lyapunov approach was considered by Yedavalli and Liang [16]. A technique for the improvement of stability robustness by shaping the singular value spectrum using constrained optimization methods was described in Ref. [17].

The following problems are considered in this work:

1. Effect of change in structural parameters on the robustness of space structures.
2. Effect of structural optimization on the robustness of the control system design.
3. Multiobjective optimization of actively controlled structure by treating the structural weight, the stability robustness index, and the performance robustness index as the objectives for minimization.

The dominant closed loop eigenvalue has been used to analyze and test the stability robustness of the system.

2. System Formulation and Basic Equations

The equations of motion of a large space structure with active controls under external forces are given by

$$[M]\ddot{\vec{U}} + [C]\dot{\vec{U}} + [K]\vec{U} = [D]\vec{F} \tag{1}$$

where [M] is the mass matrix, [C] is the damping matrix and [K] is the stiffness matrix. These matrices are of order r which denotes the number of degrees of freedom of the structure. $\vec{U}$ represents the vector of displacements and a dot over a symbol denotes differentiation with respect to time. [D] is a rxp matrix denoting the applied load distribution that relates the control input vector $\vec{F}$ to the coordinate system. The number of components of $\vec{F}$ is assumed to be p. By introducing the coordinate transformation

$$\vec{U} = [\Phi]\,\vec{\eta} \tag{2}$$

where $[\Phi]$ is the rxr modal matrix whose columns are the eigenvectors and $\vec{\eta}$ is the vector of modal coordinates, Eq. (1) can be transformed into a system of uncoupled differential equations as

$$[\bar{M}]\ddot{\vec{\eta}} + [\bar{C}]\dot{\vec{\eta}} + [\bar{K}]\vec{\eta} = [\Phi]^T[D]\vec{F} \tag{3}$$

where

$$[\bar{M}] = [\Phi]^T[M][\Phi] = [I] \tag{4}$$

$$[\bar{C}] = [\Phi]^T[C][\Phi] = \mathrm{diag}[2\xi_i\omega_i] \tag{5}$$

$$[\bar{K}] = [\Phi]^T[K][\Phi] = \mathrm{diag}[\omega_i^2] \tag{6}$$

ξ_i is the damping ratio of ith mode and ω_i is the natural frequency of i^{th} mode. Then Eq. (3) can be converted into a state space representation by using the transformation

$$\vec{x} = (\vec{\eta}, \dot{\vec{\eta}})^T \tag{7}$$

where $\vec{x}$ is the nx1 state variable vector with n=2r. This gives the state input equation

$$\dot{\vec{x}} = [A]\vec{x} + [B]\vec{F} \tag{8}$$

where [A] is the the nxn plant matrix and [B] is the nxp input matrix given by

$$[A] = \begin{bmatrix} [O] & [I] \\ -[\bar{K}] & -[\bar{C}] \end{bmatrix} \tag{9}$$

and

$$[B] = \begin{bmatrix} [O] \\ [\phi]^T \quad [D] \end{bmatrix} \tag{10}$$

The state output equation is given by

$$\vec{y} = [\underline{C}]\vec{x} \tag{11}$$

where $\vec{y}$ is the qx1 output vector and $[\underline{C}]$ is the qxn output matrix. If the sensors and actuators are co-located, then q=p and

$$[\underline{C}] = [B]^T \tag{12}$$

In order to design a controller using a linear quadratic regulator, a performance index J is defined as

$$J = \int_0^\infty (\vec{y}^T [Q]\vec{y} + \vec{F}^T [R]\vec{F})\, dt \tag{13}$$

where [Q] is the state weighting matrix which has to be positive semidefinite and [R] is the control weighting matrix which has to be positive definite.

Equations (8) to (11) are assumed to correspond to the system with nominal design variables. For this system, an optimal state feedback controller is designed by linear quadratic regulator as

$$F = -[R]^{-1}[B]^T[K]\vec{x} \tag{14}$$

$$= -[G]\vec{x}$$

where [K] satisfies the algebraic Riccati equation:

$$0 = [A]^T[K] + [K][A] - [K][B][R]^{-1}[B]^T[K] + [\underline{C}]^T[Q][\underline{C}] \tag{15}$$

This controller minimizes the quadratic performance index J. If there is a bounded uncertainty or disturbance in the design variables, then the plant matrix [A], input matrix [B] and the output matrix [$\underline{C}$] will be changed to [$\tilde{A}$], [$\tilde{B}$] and [$\tilde{C}$] respectively. If the controller designed for the nominal system, given by Eq.(14), is used for this case, it might cause the system to be unstable. This brings the necessity of consideration of the stability robustness.

3. Robustness Analysis and Problem Formulation

3.1 Stability robustness index (β_{sr})

In this work, the optimal control law is used to design the controller of the structure. Under the permissible uncertainty in the design variables, the stability robustness is maximized. For this, a stability robustness index is defined as the change in the dominant eigenvalue of the closed loop system as

$$\beta_{sr} = \frac{\left||\lambda_1| - |\tilde{\lambda}_1|\right|}{|\lambda_1|} \tag{16}$$

where λ_1 is the dominant eigenvalue of the system ([A]-[B][G]), and $\tilde{\lambda}_1$ is the dominant eigenvalue of the system ([$\tilde{A}$]-[$\tilde{B}$][G]). Note that $Re(\lambda_1) < 0$ if and only if the original system is asymptotically stable and $Re(\tilde{\lambda}_1)$ is not guaranteed to be negative. Hence the stability requirement, $Re(\tilde{\lambda}_1) < 0$, is used as a constraint in the optimization

procedure.

3.2 Performance robustness index (β_{pr})

In addition to the stability robustness, it is desirable to retain the performance unchanged when the design variables change. Since the performance cannot remain the same, a performance robustness index is defined as follows. From Eq.(13) ,the steady state solution of the stable system, as $t_f \rightarrow \infty$, gives

$$J = x_0^T [K] x_0 \tag{17}$$

where x_0 denotes the initial state variable vector. For the modified system with $[\tilde{A}]$, $[\tilde{B}]$ and $[\tilde{C}]$

$$\tilde{J} = x_0^T [\tilde{K}] x_0 \tag{18}$$

where $[\tilde{K}]$ satisfies the Lyapunov equation

$$0 = [\tilde{A}]^T [\tilde{K}] + [\tilde{K}][\tilde{A}] + ([G]^T [R][G] + [C]^T [Q][C]) \tag{19}$$

The performance robustness index, β_{pr}, is defined as

$$\beta_{pr} = \left| \frac{J - \tilde{J}}{J} \right| = \frac{x_0^T (K - \tilde{K}) x_0}{x_0^T K x_0} \tag{20}$$†

It has been pointed out in Ref.[19] that the optimal solution of Eq.(20), in general, depends on the initial state x_0. This result is not very useful since the initial state is not always known. In Ref.[19] the effect of x_0 is averaged out by assuming that x_0 is a uniformly distributed random vector whose covariance is given by the identity matrix. It can be shown that the trace of K is proportional to the expected value of J. Hence

† For simplicity, square brackets are not used to denote matrices, hereafter.

the performance robustness index is redefined as

$$\beta_{pr} = \left| \frac{Tr(K) - Tr(\tilde{K})}{Tr(K)} \right| \tag{21}$$

3.3 Guaranteed stability margin

For both nominal and perturbed systems, good dynamic response can be achieved if the real part of every eigenvalue is restricted to be smaller than a specified value, -a, with a>0. The LQ regulator can be modified [21] as follows with the requirement of ([A],[B]) being controllable and ([A],[C]) being observable.

$$\dot{x} = (A + aI)x + Bu \tag{22}$$

$$u = -R^{-1}B^T K_a x$$

$$= -Gx \tag{23}$$

with K_a satisfying the equation

$$0 = (A+aI)^T K_a + K_a(A+aI) - K_a B R^{-1} B^T K_a + C^T Q C \tag{24}$$

Equation (22) can be rewritten as

$$\dot{x} = (A + aI - BG)x \tag{25}$$

whose characteristic equation is given by

$$\det [(A + aI - BG) - \lambda I] = \det [A - BG - \lambda' I] = 0 \tag{26}$$

with $\lambda' = \lambda - a$. Since A is assumed to be stable, then $Re(\lambda') < 0$, and hence $Re(\lambda) < -a$. Thus, the system of Eqs.(8) to (11) with controller (23) provides a guaranteed stability margin.

If the system matrix is made separable, then the guaranteed stability margin technique can be used to move certain eigenvalues to specified values. For a controlled structure, the system equations are given by Eqs.(8) to (11). If the real part of the i^{th}

eigenvalue pair of (A-BG) (G is given in Eq.(14)) is greater than the specified stability margin, then a modified controller gain can be used to satisfy the requirement. In this case, the modified controller can be designed as follows.

$$\tilde{A} = A - \begin{bmatrix} [\bar{I}] & [I] \\ [-\omega_i^2] & [-2\zeta_i\omega_i] + [\bar{I}] \end{bmatrix} \tag{28}$$

with

$$\bar{I} = \begin{cases} 1 \text{ at (i,i)position} \\ 0 \text{ at other positions} \end{cases}$$

$$u = -Gx \tag{29}$$

$$= -R^{-1}B^T\tilde{K}x$$

with $\tilde{K}$ satisfying the algebraic Riccati equation

$$0 = \tilde{A}^T\tilde{K} + \tilde{K}\tilde{A} - \tilde{K}BR^{-1}B^T\tilde{K} + C^TQC \tag{30}$$

3.4 Multiobjective design problem

In this work, the stability robustness, the performance robustness and the total structural weight are considered as the objective functions. The cross section areas of the members are treated as the design variables. The first objective function, the stability robustness index (β_{sr}), describes the relative stability of the system when the design variables change by a specified amount. It is assumed that the controller gains are such that the condition for the stability of the system is satisfied and thus the closed loop system matrix of the perturbed system is still stable. According to this definition, β_{sr}=0 corresponds to highly robust system from the stability point of view. However, β_{sr} will not attain the value zero due to the presence of perturbations in the design variables. The second objective function, the performance robustness index of

the system (β_{pr}), is defined by Eq.(21). Here also, $\beta_{pr}=0$ corresponds to a highly robust system from the performance point of view. The value $\beta_{pr}=0$ will not be attained in practice due to the presence of perturbations in the design variables. The third design objective function, the total structural weight, is given by

$$f_3(x) = \sum_{i=1}^{N} \rho_i l_i A_i \tag{31}$$

with ρ_i, l_i and A_i denoting the density, length and cross-sectional area of the i^{th} member respectively, and N representing the number of members in the truss structure. The following constraints are used during the optimization procedure:

1. Upper and lower bounds on the design variables.
2. Stability requirement, i.e. the requirement of the real parts of the eigenvalues of the closed loop system to be negative.
3. Lower and/or upper bounds on the natural frequencies of vibration of the structure.

In some cases, the closed loop damping ratios of the system may have to be constrained; however, these are not considered in this work.

4. Multiobjective Optimization

As stated in section 3.4, three objective functions are used in the optimization problem. This requires the use of a suitable multiobjective optimization technique for the solution. In this work, the utility function method, the lexicographic and the goal programming methods are used for numerical solution. These methods are briefly described below [20].

4.1 Utility function method

This is a widely used multiobjective optimization method. This method involves the solution of the following problem:

$$\text{Min } U(x) = \sum_{i=1}^{k} w_i f_i(x) \tag{32}$$

subject to

$$g_j(x) \leq 0 \ , \ j=1,2,...,m$$

where w_i is the weight of the i^{th} objective function f_i and $\sum_{i=1}^{k} w_i = 1$. Usually, the scales and units of different objective functions are different. Hence a suitable normalization process has to be used in constructing the objective functions of Eq.(32). A convenient form is to define a new objective function F_i as

$$F_i(x) = \frac{f_i(x) - f_i^*(x^*)}{f_i^*(x^*)} \tag{33}$$

and redefine the objective function of Eq.(32) as

$$\text{Min } U(x) = \sum_{i=1}^{k} w_i F_i(x) \tag{34}$$

4.2 Lexicographic method

In this method, the objectives are ranked in order of importance by the designer. The preferred solution obtained by this method is one which minimizes the objectives starting with the most important one and proceeding according to the order of importance of the objectives. Let the subscripts of the objectives indicate not only the objective function number, but also the priorities of the objectives. Thus $F_1(x)$ and $F_k(x)$ denote the most and the least important objective functions, respectively. Then the first problem is formulated as

Min $F_1(x)$ (35)

subject to

$g_j(x) \leq 0$, j=1,2,...,m

and its solution x_1^* and $F_1^* = F_1(x_1^*)$ are obtained. Then the second problem is formulated as

Min $F_2(x)$ (36)

subject to

$g_j(x) \leq 0$, j=1,2,...,m ,and

$g_{m+1}(x) = F_1(x) - F_1(x_1^*) \leq \epsilon_1$

where ϵ_1 is a small value compared to $F_1(x_1^*)$. This procedure is repeated until all k objectives have been considered. The i^{th} problem is given by

Min $F_i(x)$ (37)

subject to

$g_j(x) \leq 0$, j=1,2,...,m ,and

$g_{m+n}(x) = F_n(x) - F_n(x_n^*) \leq \epsilon_n$, n= 1,2,...,i−1

The solution obtained at the end, i.e. x_k^* is taken as the desired solution x^* of the multiobjective optimization problem.

4.3 Goal programming method

The goal programming method was originally proposed by Charnes and Cooper for linear optimization problem [22]. The method requires goals to be set for each objective that the designer wishes to obtain. A preferred solution is then defined as the one which minimizes the deviations from the set goals. Thus a simple goal programming problem can be defined as

$$\text{Min } F(x) = \left[\sum_{i=1}^{k} \left(\underset{\sim}{F}_i(x) \right)^p \right]^{1/p} , \; p \geq 1 \tag{38}$$

subject to

$$g_j(x) \leq 0 , \; j=1,2,...,m$$

$$\underset{\sim}{F}_j(x) \geq 0 , \; j=1,2,...,k$$

where $\underset{\sim}{F}_j(x) = F_j(x) - F_j(x_j^*)$

5. Computational Procedure

5.1 Analysis

The purpose of the analysis is to study the effects of variations in the parameters of the structure on the robustness of the system. The following procedure is used for this purpose.

1. Start with an initial reference design of the structure and find the corresponding plant matrix A, input matrix B and the output matrix C.
2. Use the LQ regulator design technique to find the optimal control gain G by solving the algebraic Riccati equation.
3. Change the design parameters by known percentage values and find the corresponding A, B and C matrices.
4. Find the stability robustness index δ_{sr} (Eq.(16)) and the performance robustness index δ_{pr} (Eq.(21)).
5. Repeat steps (3) and (4) for different parameter changes. (e.g. nominal design variables, damping ratio, density etc.)

6. Plot a graph between ζ_{sr} or ζ_{pr} and the change in the parameters.

5.2 Design

The purpose of design is to optimize the actively controlled structure by using suitable multiobjective optimization techniques. The procedure is given as follows:

1. From the requirements of stress and deformation, obtain the preliminary design (to be used as the nominal design) of the structure.
2. Construct the plant matrix, input matrix and output matrix.
3. Formulate the multiobjective constrained optimization problem.
4. Minimize the individual objective functions and find the respective minima around the nominal design.
5. Use a suitable multiobjective optimization approach to find a compromise solution.

6. Examples

6.1 Two-Bar Truss

The two-bar truss shown in Fig.1 is selected for its simplicity. A nonstructural mass of 1 unit is attached at node 3. The actuators and the sensors are co-located at node 3 acting in x and y directions. The design variables (cross-sectional areas of the two bars) are restricted to lie between 0.01 and 1.0. The structural damping ratio is considered as 0.01, Young's modulus is assumed to be 10^7, and density is taken to be 4.6. In the performance index, the output weighting matrix Q is assumed to be 1000. I, and the input weighting matrix R is taken to be I, where I is the identity matrix.

The natural frequencies of the closed loop controlled structure are constrained to lie between 20 rad/sec and 40 rad/sec. For a stable open loop system, the corresponding feedback closed loop system must be stable under the optimum control law. But the stability is not guaranteed if there exists disturbances or uncertainties in system parameters. Hence, additional constraints are added on the perturbed closed loop system, namely, that all the eigenvalues of the perturbed closed loop system are restricted to have negative real parts.

6.1.1 Analysis

Figure 2 shows the relationship between the stability robustness index and the design variables, which can be seen to be a smooth convex function. Figure 3 shows the variation of the performance robustness index with the two design variables, it can be seen to be a non-smooth function having several local minima in the design space. Figures 4 to 10 present the variations of stability robustness index, performance robustness index and performance index with changes in the structural damping ratio of the two-bar truss. Figures 4, 6 and 8 show that the stability robustness index drops sharply at a value of the damping ratio of approximately 4%. Figures 5, 7 and 9 indicate that the performance robustness index attains a minimum value at a damping ratio of approximately 2%. Figure 10 shows that the performance index decreases as the damping ratio increases. The effect of the variations in Young's modulus of the material on the stability robustness index, performance robustness index and performance index is shown in Figures 11 to 14. It can be seen from Fig. 11 that the stability robustness index changes very little beyond a value of 5×10^6 of the Young's modulus. The performance robustness index reduces to a minimum value at Young's

modulus (E)= 20×10^6 and then increases for larger values of E. On the other hand, the system performance index increases to a maximum value at E=20×10^6 and then decreases for larger values of E (see Fig.13).

The relationship of the stability robustness index, the performance robustness index and the performance index with the mass density of the material is shown in Fig. 14 to 16. It can be observed that the stability robustness index decreases with an increase in the density of the material. The performance robustness index reduces to a minimum at ρ=1.5 and then increases for higher values of ρ. The performance index attains a maximum value at ρ=1.5 and then decreases monotonically (see Fig.16). Figures 17 to 19 show the variations of the stability robustness index, the performance robustness index and the performance index with a change in the coefficient of the output weighting matrix. An increase in the coefficient of the output weighting matrix implies that the output performance is more important than the control energy. Obviously, it improves the system stability as well as the performance index.

6.1.2 Design

The results of minimization of the individual objective functions are shown in Table 1. The results given by different multiobjective optimization methods are shown in Table 2. The first two columns in Table 2 correspond to formulations #1 and #2 of the utility function method. In formulation #1, w_i are set equal to a fixed value of 1/3 in Eq.(34) while w_i are considered as design variables in formulation #2. The last row of Table 2 gives the values of the global evaluation function which can be used as an index to compare the results of different multiobjective optimization methods. The global evaluation function is defined as

$$F_g(x) = \sum_{i=1}^{3} F_i(x^*) \tag{39}$$

where

$$F_1(x) = \frac{(f_1(x) - 0.009502)}{0.010535},$$

$$F_2(x) = \frac{(f_2(x) - 0.0015899)}{0.0032231},$$

and $$F_3(x) = \frac{(f_3(x) - 23.598)}{43.682}.$$

6.2 Two-Bay Truss

The finite element model of the second example (two-bay truss) is shown in Fig. 20. For this example, nonstructural masses of magnitude 1.29 are attached at nodes 1 to 4. Each node has two degrees of freedom. The actuators and sensors are co-located at nodes 1 to 4 and are assumed to act along the y-direction only. The design variables (cross- sectional areas) are restricted to lie between 0.001 and 0.5. The natural frequencies of the closed loop system are constrained to be larger than 31.62 rad/sec (i.e. $\omega^2 \geq 1000$).

6.2.1 Analysis

This example has ten design variables. Since the display of functional relations in ten dimensional design space is not possible, the variation of the robustness of the system is found by uniformly varying the value of all the ten design variables. The results are shown in Figs.21 to 24. Figure 21 shows the stability robustness index vs the value of the design variables when the permissible change in the design vector is assumed to be -5%. It can be observed that β_{sr} decreases slowly with an increase in the value of the design variables; but it is not a convex function. Hence local minima

are expected in the optimization process , i.e., the optimum solution will depend on the initial guess. Figure 22 shows the variation of the performance robustness index and the performance index with the value of the design variables when the permissible change in the design vector is -5%. In this case, β_{pr} decreases until a value of 0.2 for the design vector and then increases. This relationship is also not a convex function as in the case of Fig.21. On the other hand, the performance index can be seen to decrease monotonically with increase in the value of the design vector. This implies that a stronger structure will induce a smaller displacement and need lesser control energy to obtain good performance. Figures 23 and 24 show the variations of the stability robustness index and the performance robustness index with respect to the design variables when the permissible change in the design vector is assumed as -10%. In these cases also, the trend can be seen to be similar to those observed in Figs. 21 and 22.

6.2.2 Design

The nominal value of the design variables are assumed to be $x_i = 0.1$, $i = 1$ to 10. Table 3 gives the results obtained by optimizing the individual objective functions starting from the nominal design. The results of different multiobjective optimization methods, namely, the utility function method, the Lexicographic method and the goal programming method are compared in Table 4. The last row of Table 4 shows the global evaluation function, F_g, defined as

$$F_g(x) = \sum_{i=1}^{3} F_i(x^*) \qquad (40)$$

where

$$F_1(x) = \frac{(f_1(x) - 0.048694)}{0.000668},$$

$$F_2(x) = \frac{(f_2(x) - 0.008454)}{0.001588},$$

and $$F_3(x) = \frac{(f_3(x) - 0.2275)}{0.02681}.$$

7. Summary and Conclusions

1. The stability robustness index and the performance robustness index considered in this work are highly nonlinear with respect to variations in design variables.

2. The relationships between the stability/performance robustness index and the various system parameters have been determined numerically for the two-bar truss. These results are expected to be useful in choosing a suitable material for a given structure with a specified geometry.

3. The variation of the stability/performance robustness index with changes in design variables has been found to be a non-smooth function. This leads to difficulties in optimisation; one can expect only a local minimum in the neighborhood of the starting design. In general, the local optima are acceptable since the starting design is usually taken as the nominal design which is expected to be a robust design.

4. A major advantage of using nonlinear programming to find the robust control/structural design is that it can be used for large permissible changes in the design variables and/or different constraint specifications.

5. Three multiobjective optimization methods have been used to find the optimal design in both the illustrative examples. For the two-bar truss, the utility

function method with variable coefficients gave the smallest value of the global evaluation function. For the two-bay truss, the goal programming method with p=2 yielded the smallest value for the global evaluation function and the utility function method with variable coefficients gave the second smallest value for the global evaluation function.

Acknowledgement

The valuable suggestions made by Dr. Vipperla B. Venkayya of Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base at various stages of the present research work are gratefully acknowledged.

8. References

1. L.D. Pinson, A.K. Amos and V.B. Venkayya (Eds.), "Modeling, Analysis and Optimization Issues for Large Space Structures," *Proceedings of the NASA-AFOSR Workshop,* Williamsburg, VA, May 13-14, 1982.

2 V.B. Venkayya and V.A. Tischler, "Frequency Control and Effect on the Dynamic Response of Flexible Structures," (84-1044-CP) AIAA Dynamics Specialists Conference, Palm Springs, CA, May 17-18, 1984.

3. D.F. Miller, V.B. Venkayya and V.A. Tischler, "Integration of Structures and Controls--Some Computational Issues," *Proceedings of 24th Conference on Decision and Control,* Ft. Lauderdale, FL, December 1985.

4. M.J. Balas, "Some Trends in Large Space Structure Control Theory: Fondest Hopes; Wildest Dreams," *IEEE Transactions on Automatic Control,* Vol. AC-27, June 1982.

5. N.S. Khot, F.E. Eastep and V.B. Venkayya, "Optimal Structural Modifications to Enhance the Optimal Active Vibration Control of Large Flexible Structures," AIAA/ASME/ASCE/AHS 26th Structures, Structural Dynamics and Materials Conference, Orlando, FL, April 1985.

6. M. Salama, M. Hamidi and L. Demsetz, "Optimization of Controlled Structures," presented at the Jet Propulsion Workshop on Identification and Control of Flexible Space Structures, San Diego, CA, June 4-6, 1984.

7. D.F. Miller and J. Shim, "Combined Structural and Control Optimization for Flexible Systems using Gradient Based Searches," 24th AIAA Aerospace Sciences Meeting, Reno, NV, January 6-8, 1986.

8. N.S. Khot, H. Oz, F.E. Eastep and V.B. Venkayya, "Optimal Structural Designs to Modify the Vibration Control Gain Norm of Flexible Structures," AIAA/ASME/ASCE/AHS 27th Structures, Structural Dynamics and Materials Conference, Paper No. 86-0840-CP, 1986.

9. J.L. Junkins, D.S. Bodden and J.D. Turner, "A Unified Approach to Structure and Control System Design Iterations," Presented at the Fourth International Conference on Applied Numerical Modelling, Tainan, Taiwan, Dec. 27-29, 1984.

10. D.B. Ridgely and S.S. Banda, "Introduction to Robust Multivariable Control," Flight Dynamics Laboratory, Air Force Wright Aeronautical Laboratories, Report No. AFWAL-TR-85-3102, February 1986.

11. R.K. Yedavalli, "Time Domain Robustness Analysis for Linear Regulators," *Proceedings of American Control Conference,* San Diego, June 1984, pp. 975-980.

12. R.K. Yedavalli, S.S. Banda and D.B. Ridgely, "Time Domain Stability Robustness Measures for Linear Regulators," *Journal of Guidance, Control and Dynamics,* Vol. 8, No. 4, 1985, pp. 520-525.

13. R.L. Kosut, H. Salzwedel and A.E. Naeini, "Robust Control of Flexible Spacecraft," *Journal of Guidance and Control,* Vol. 6, March - April 1983, pp. 104-111.

14. V.O. Hoehne, "AFWAL Control Technology Programs," *Proceedings of the Workshop on Identification and Control of Flexible Space Structures,* G. Rodriguez (Ed.), JPL Publication 85-29, Vol. 1, April 1, 1985.

15. V.C. Gordon and D.J. Collins, "Multi-Input Multi-Output Automatic Design Synthesis for Performance and Robustness," AIAA Paper No. 85-1929.

16. R.K. Yedavalli and Z. Liang, "Reduced Conservatism in Time Domain Stability Robustness Bounds by State Transformation: Application to Aircraft Control," AIAA Paper No. 85-1926.

17. V. Mukhopadhyay, "Stability Robustness Improvement Using Constrained Optimization Techniques," AIAA Paper No. 85-1931.

18. S.S. Rao, "Game Theory Approach for Multiobjective Structural Optimization," *Computers and Structures,* 1986, (in press).

19. D.L. Kleinman and M. Athans, "The Design of Suboptimal Linear Time Varying Systems," *IEEE Trans. Automatic Control,* Vol. AC-13, April 1968, pp. 150-158.

20. S.S. Rao, "Design of Vibration Isolation Systems Using Multiobjective Optimization Techniques," ASME paper 84-DET-60.

21. D.H. Jacobson, D.H. Martin, M. Pachter and T. Geveci, "Extensions of Linear Quadratic Control Theory," 1980.

22. A. Charnes and W.W. Cooper, "Goal Programming and Multiple Objective Optimization -- Part I," *European Journal of Operations Research,* Vol. 1, 1977, pp. 39-54.

Table 1.

Single Objective Optimization of Two-Bar Truss.

Permissible design variable change = -5%.

$$\xi = 0.01\,, \qquad \varsigma = 10^3\,, \qquad x(o) = \begin{Bmatrix} 0.1 \\ 0.1 \end{Bmatrix}$$

Minimization of			
Objective	β_{sr}	β_{pr}	Weight, W
C_i i=1,3	1.0 0. 0.	0. 1.0 0.	0. 0. 1.0
x^*	0.14626 0.14626	0.15247 0.13797	0.051301 0.051301
$f_1(x^*)$	0.009502	0.009557	0.020037
$f_2(x^*)$	0.001618	0.0015899	0.004813
$f_3(x^*)$	67.28	66.801	23.598
$f^* = \sum_{i=1}^{3} C_i f_i$	0.009502	0.0015899	23.598

Table 2.

Multiobjective Optimization of Two-Bar Truss

Permissible design variable change = -5%

$$\xi = 0.01\,, \qquad \varsigma = 10^3\,, \qquad x(o) = \begin{Bmatrix} 0.1 \\ 0.1 \end{Bmatrix}$$

	Utility Function Method		Lexicographic Method			Goal Programming Method	
	Const. Coef.	Variable Coef.	Optimization Order			p=1	p=2
			f_1, f_2, f_3	f_2, f_1, f_3	f_2, f_1, f_3		
Optimal Design Variables X^*	X_1 = 0.1309	0.1463	0.15389	0.056296	0.14466	0.051294	0.051807
	X_2 = 0.12709	0.1299	0.13483	0.056303	0.1440	0.051309	0.051807
$f_1(X^*)$	0.010506	0.009950	0.009603	0.018932	0.009604	0.020037	0.019919
$f_2(X^*)$	0.001867	0.001792	0.001669	0.004399	0.001664	0.004809	0.004770
$f_3(X^*)$	59.337	63.549	66.404	25.898	66.391	23.599	23.831
$\sum_{i=1}^{3} F_i(X^*)$	0.999438	0.977291	1.014075	1.819317	1.010398	1.998782	1.980792

Table 3.

Single Objective Optimization of Two-Bay Truss

Permissible design variable change = -5%

$\xi = 0.01$, $\varsigma = 10^3$, $X_i(o) = 0.1$ i=1 to 10

Minimization of			
Objective	β_{sr}	β_{pr}	Weight, W
Optimal Design Variables X_i^* i=1 to 10	0.13816 0.09648 0.13782 0.27661 0.10103 0.27780 0.14537 0.14417 0.14808 0.15119	0.13765 0.09339 0.13777 0.27637 0.09899 0.27725 0.14671 0.14507 0.14998 0.14920	0.11274 0.00100 0.11315 0.33788 0.00100 0.33772 0.11810 0.11884 0.12110 0.12122
$f_1(X^*)$	0.048694	0.048701	0.049354
$f_2(X^*)$	0.010019	0.00981975	0.0084567
$f_3(X^*)$	0.252557	0.252307	0.22730

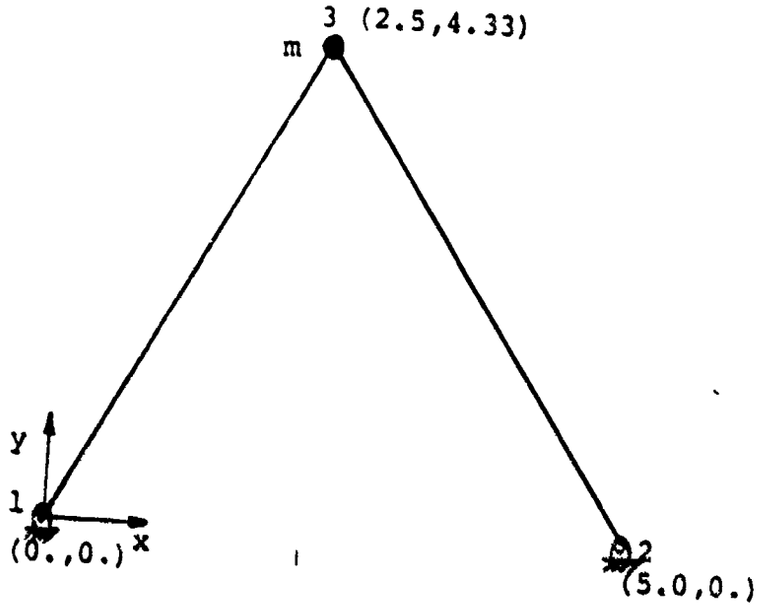
Table 4.

Multiobjective Optimization of Two-Bay Truss

Permissible design variable change = -5%

$\xi = 0.01$, $\varsigma = 10^3$, $X_i(o) = 0.1$ $i = 1,10$

Approach	Utility Function Method		Lexicographic Method		Goal Programming Method
	Const. Coef.	Variable Coef.	Optimization Order f_1, f_2, f_3	Optimization Order f_3, f_1, f_2	p=2
Optimal Design Variables X_i^* i=1,10	0.12247	0.13753	0.13697	0.14197	0.11389
	0.01779	0.09207	0.09588	0.03602	0.00100
	0.11406	0.13603	0.15071	0.16162	0.11142
	0.32120	0.27907	0.29918	0.36978	0.00100
	0.01000	0.09880	0.07856	0.00100	0.00100
	0.33931	0.27920	0.29422	0.34327	0.33617
	0.10835	0.14554	0.15018	0.11642	0.11907
	0.12838	0.14278	0.15061	0.12668	0.11901
	0.13045	0.14917	0.10706	0.10488	0.12277
	0.11834	0.14972	0.12839	0.11305	0.12420
$f_1(X^*)$	0.048769	0.048742	0.048788	0.048664	0.049369
$f_2(X^*)$	0.012034	0.00997	0.010307	0.011721	0.008479
$f_3(X^*)$	0.22903	0.25196	0.24841	0.24438	0.22736
$\sum_{i=1}^{3} F_i(X^*)$	2.431212	1.946222	2.094988	2.694381	0.999241



m = 1.0
E = 10*10**6
Density = 4.6
Nominal area =0.1

Fig. 1 Two-bar truss

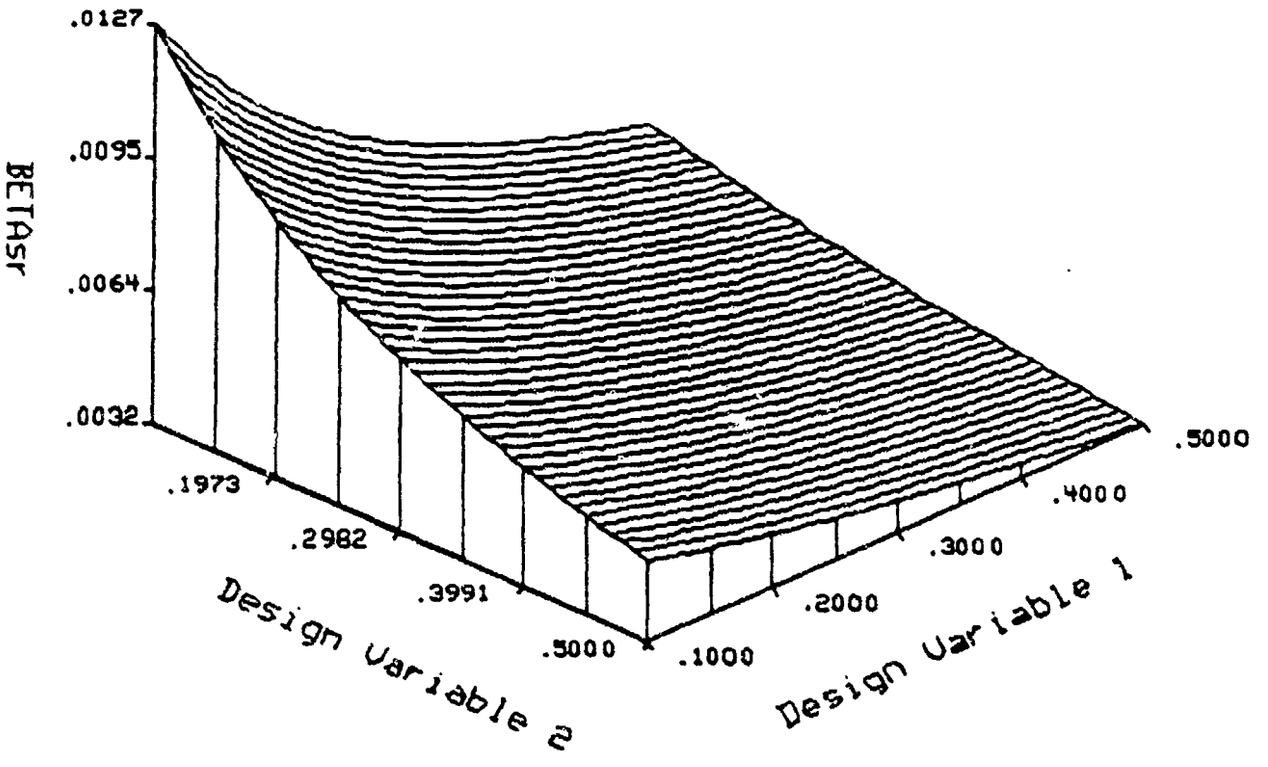


Fig. 2 Stability robustness index vs. design variables

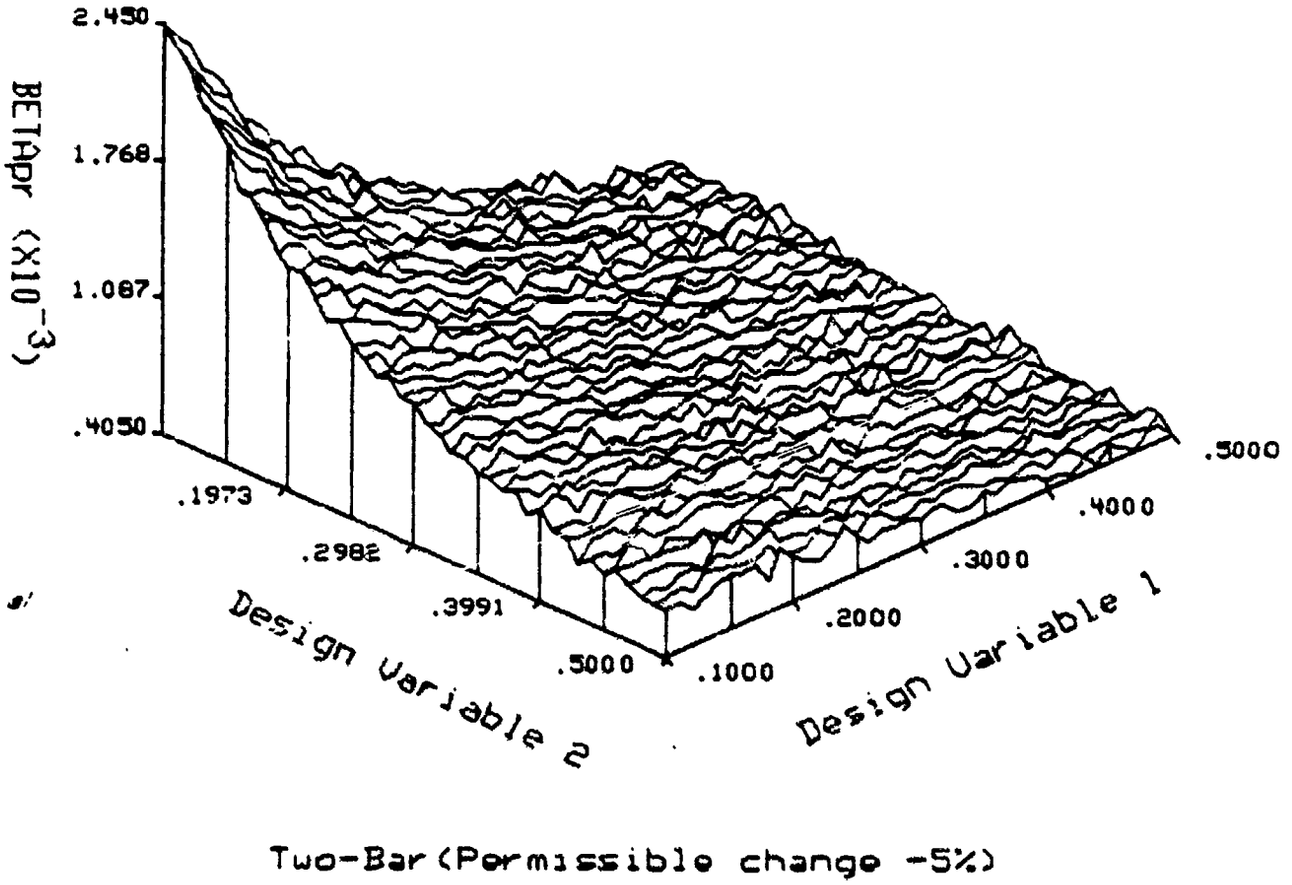


Fig. 3 Performance robustness index vs. design variables

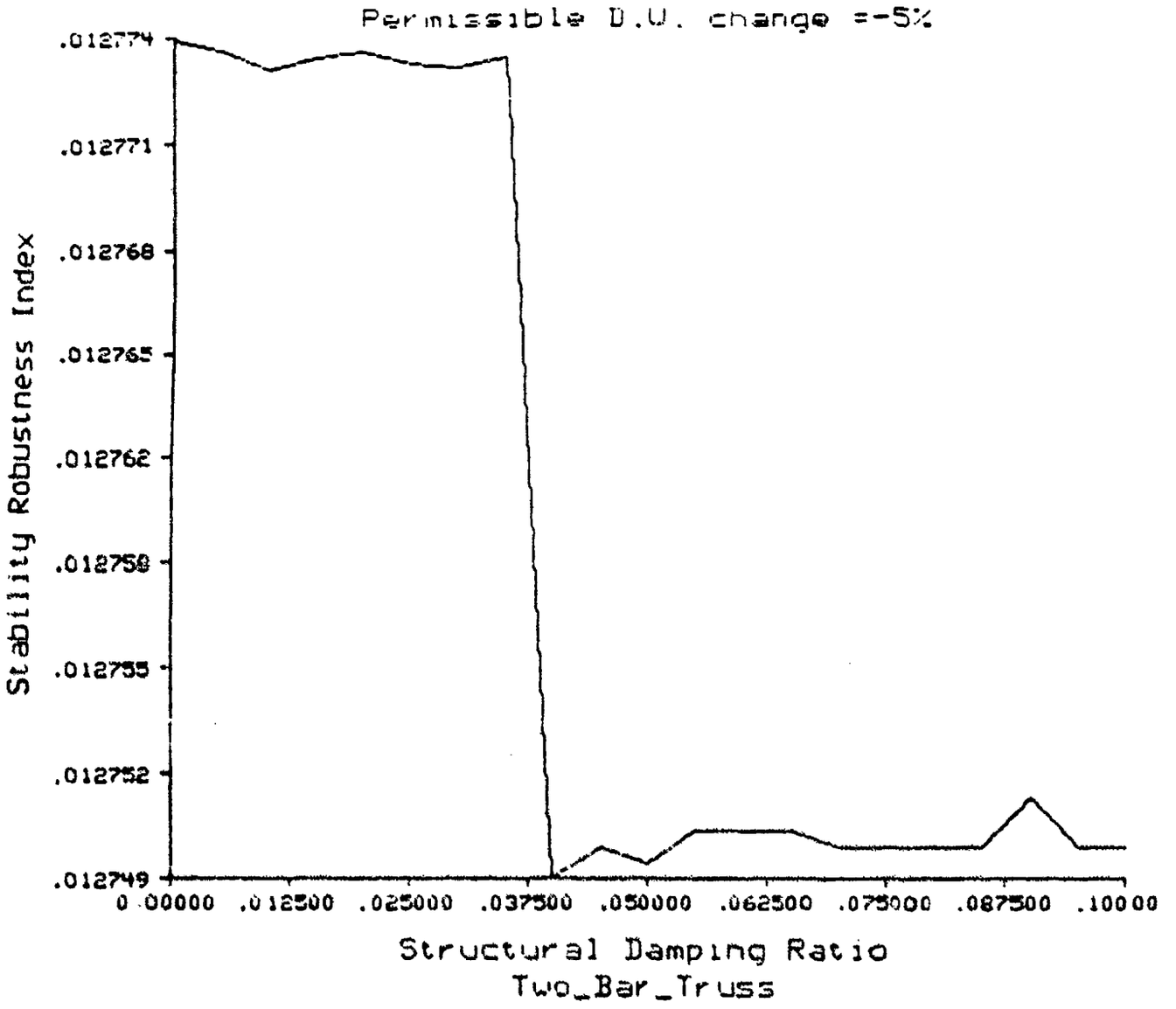


Fig. 4 Variation of λ_u with structural damping ratio (Design variable change = -5%)

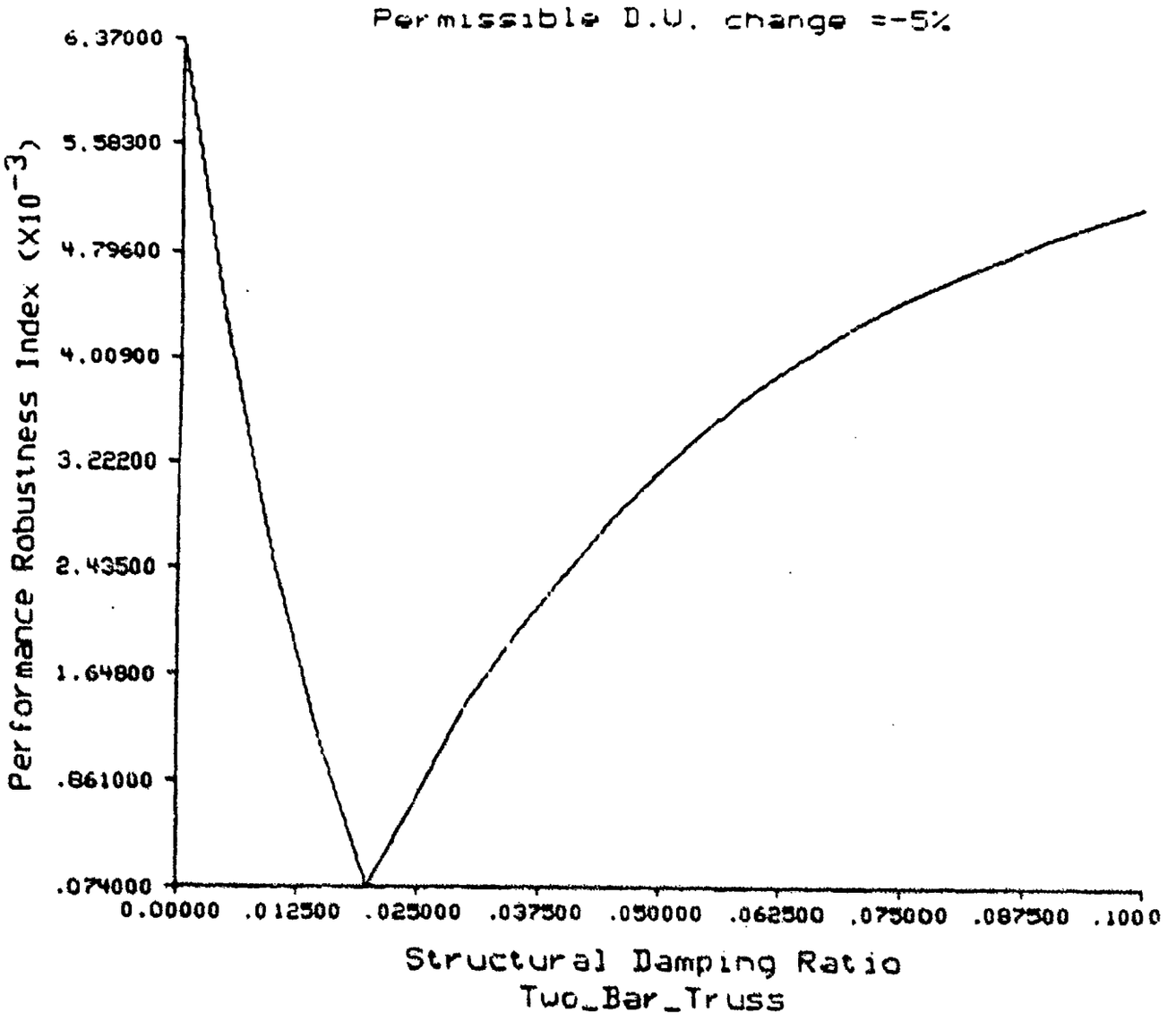


Fig. 5 Variation of γ_{pr} with structural damping ratio (Design variable change = -5%)

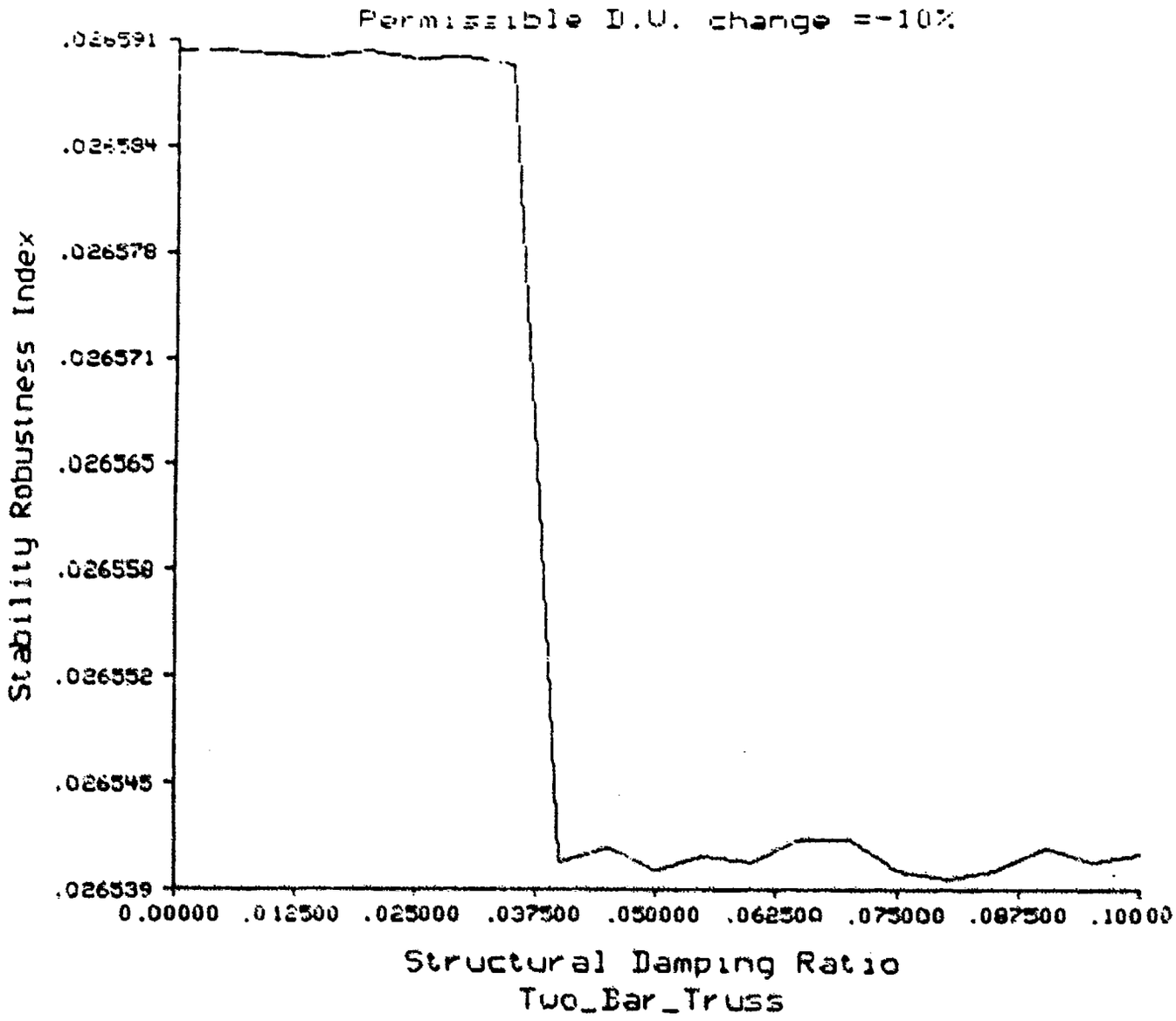


Fig. 6 Variation of λ_{ll} with structural damping ratio (Design variable change = -10%)

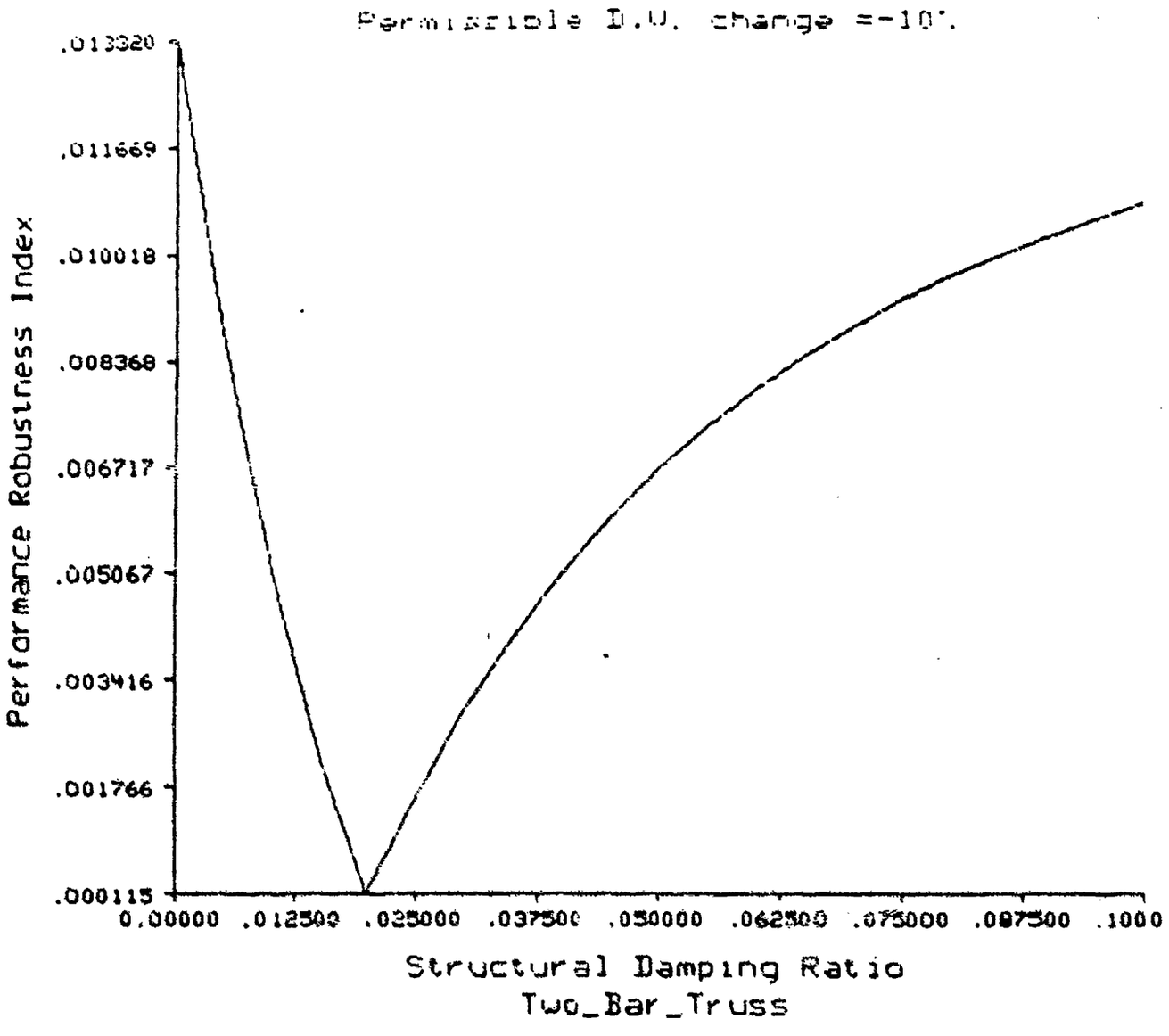


Fig. 7 Variation of λ_p with structural damping ratio (Design variable change = -10%)

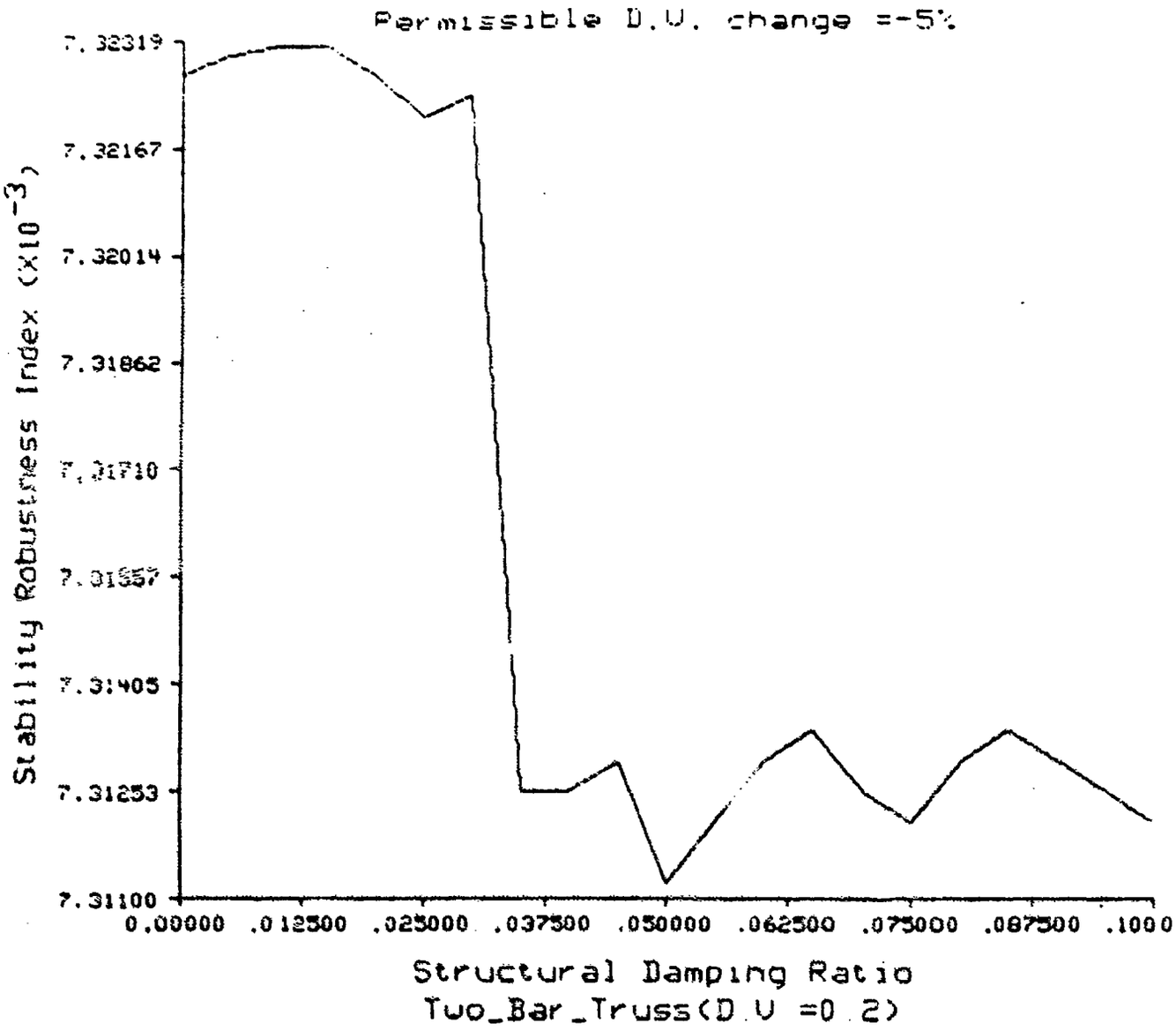


Fig. 8 Variation of ϵ_s with structural damping ratio (Design variable change = -5%, D.V. = 0.2)

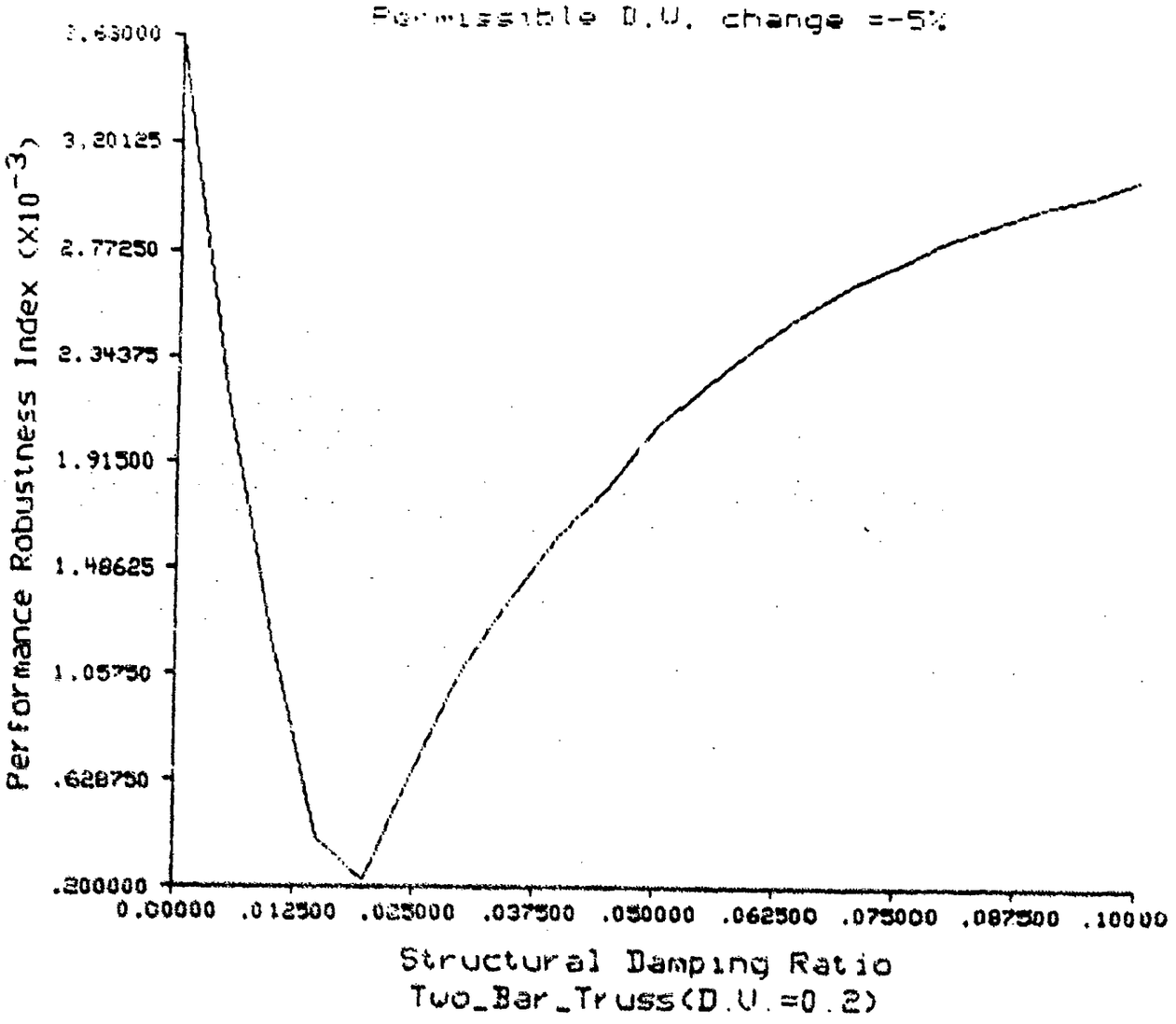


Fig. 9 Variation of ϕ_{p_i} with structural damping ratio (Design variable change = -5%, D.V. = 0.2)

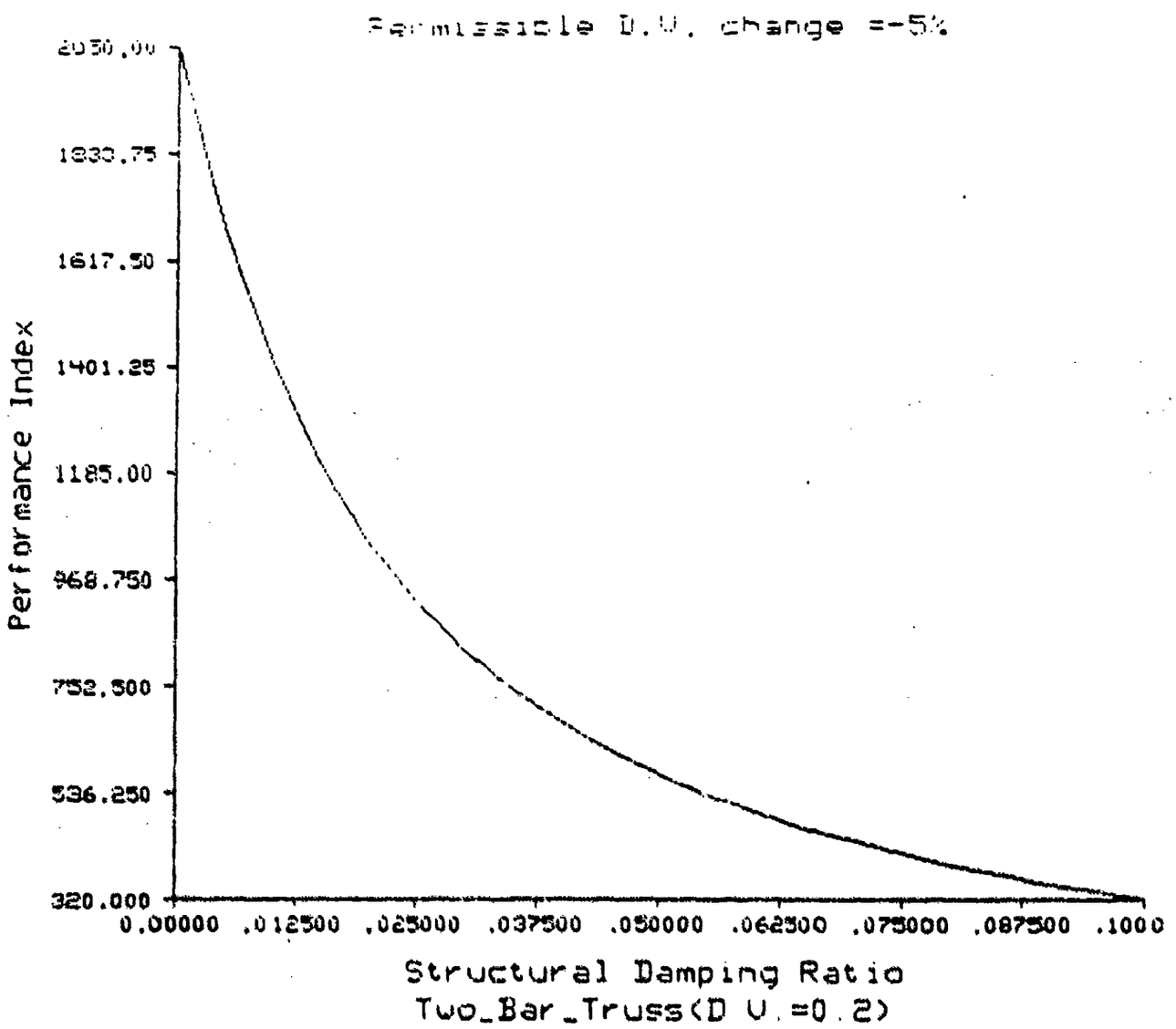


Fig. 10 Variation of performance index with structural damping ratio
(Design variable change = -5%, D.V. = 0.2)

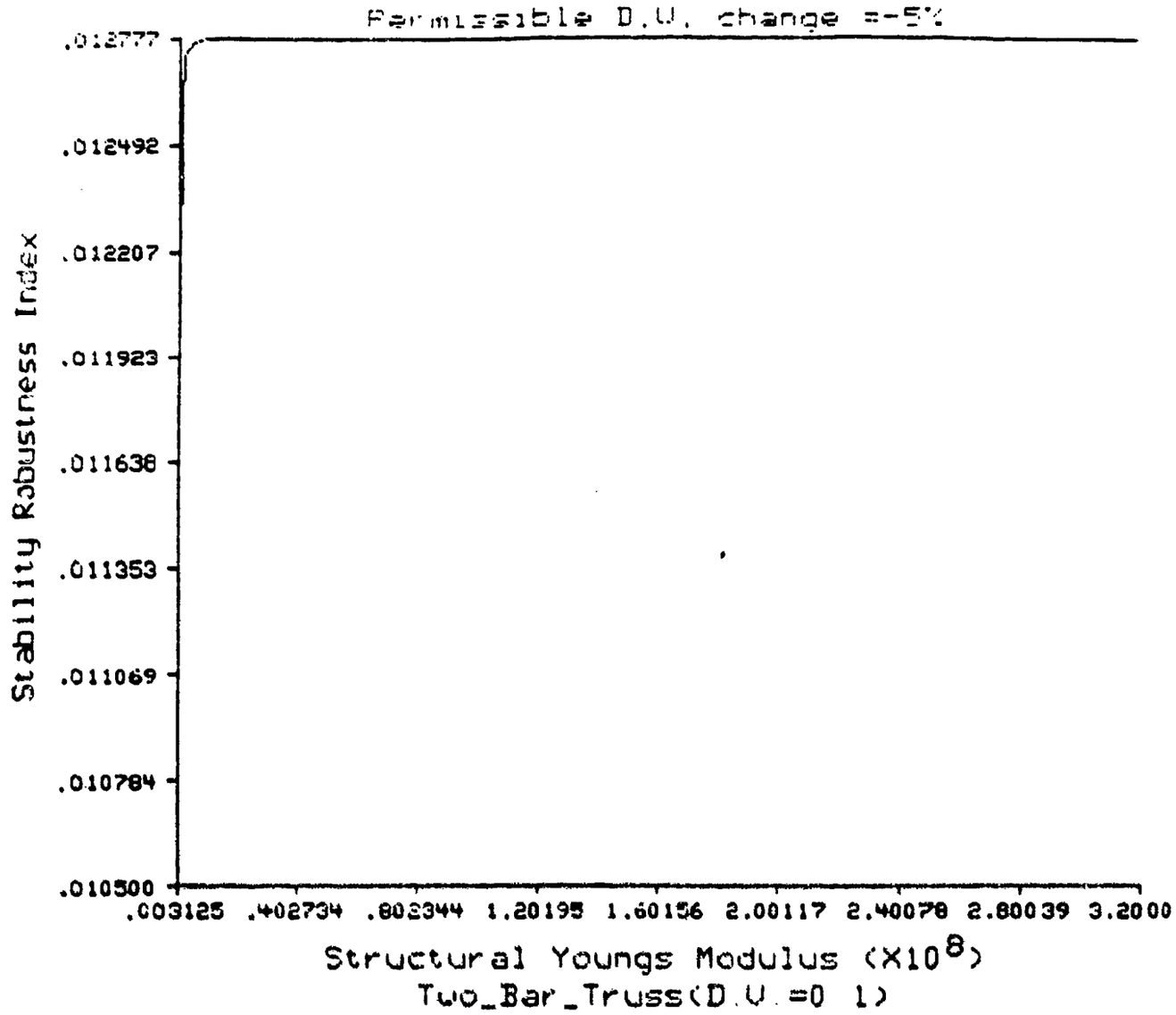


Fig. 11 Variation of λ_u with Young's modulus

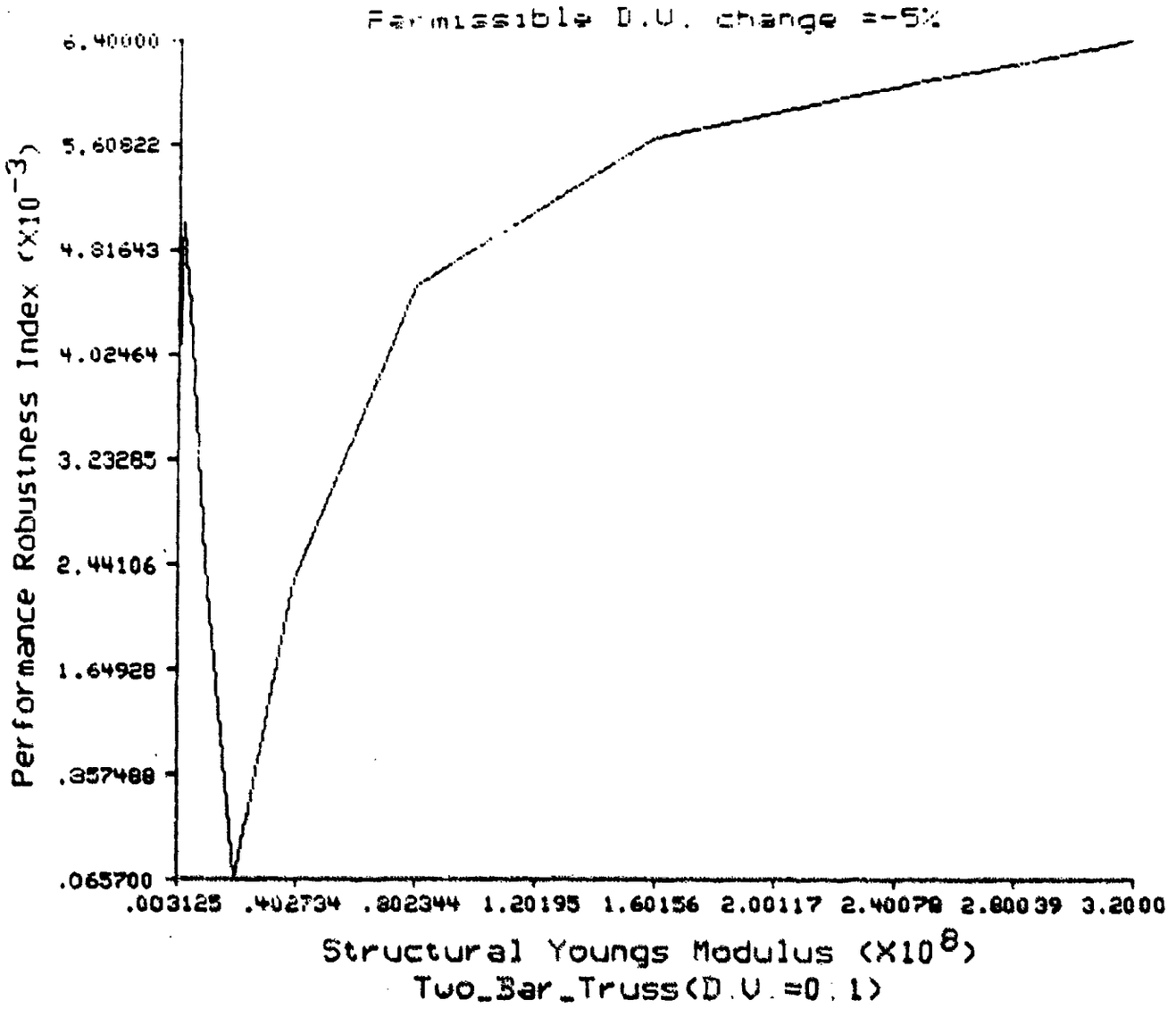


Fig. 12 Variation of ϕ_μ with Young's modulus

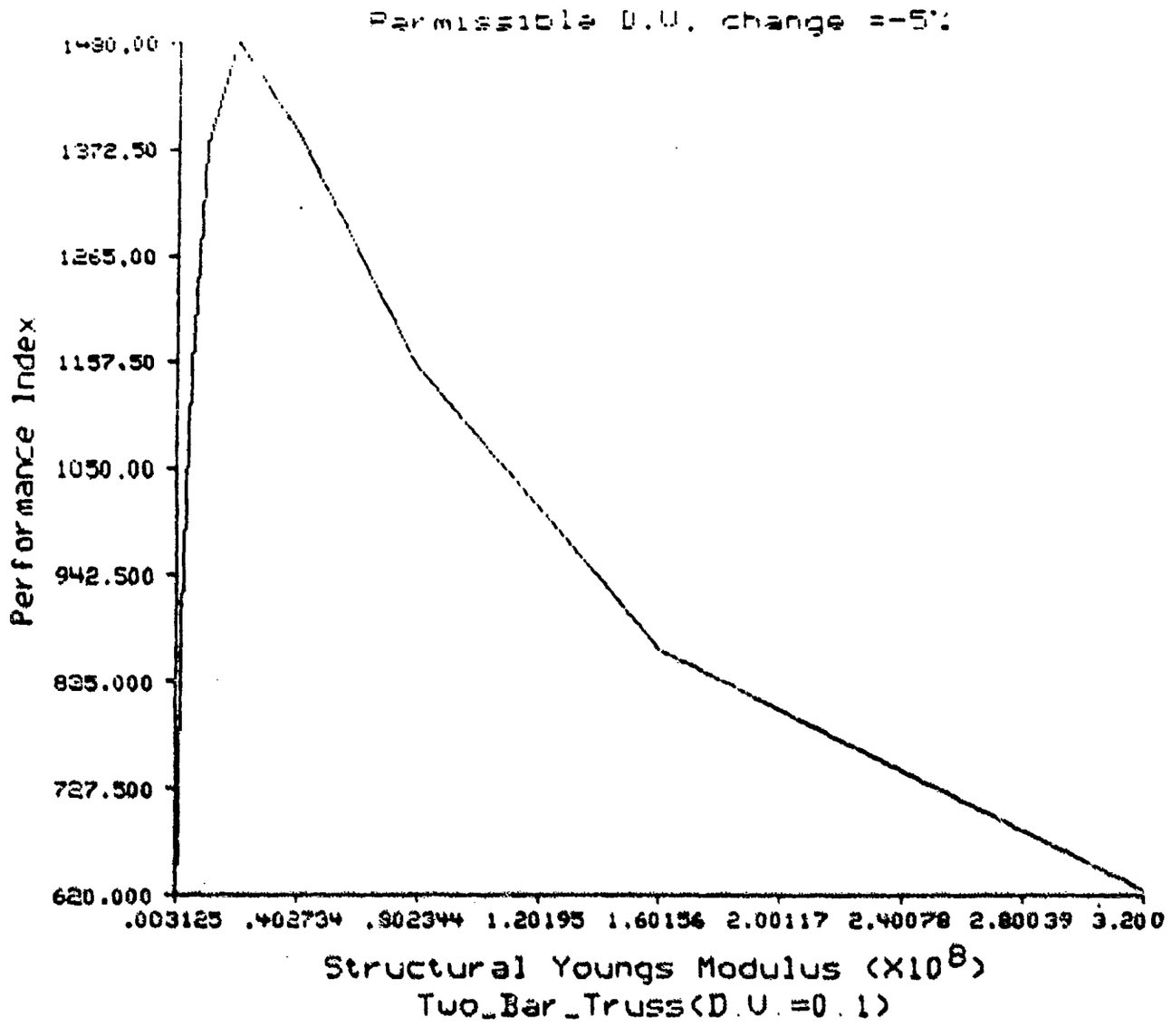


Fig. 13 Variation of performance index with Young's modulus

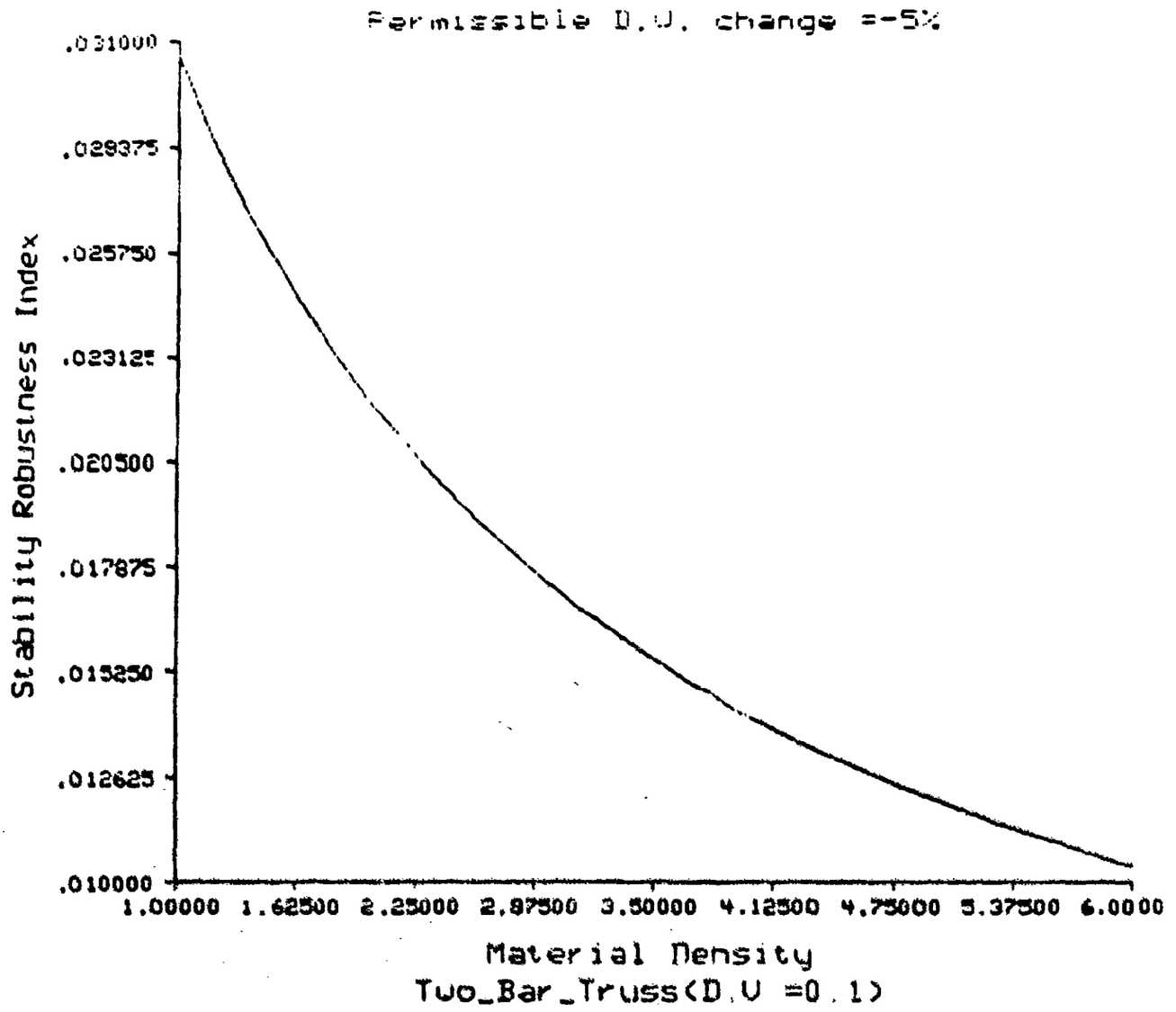


Fig. 14 Variation of [illegible] with density

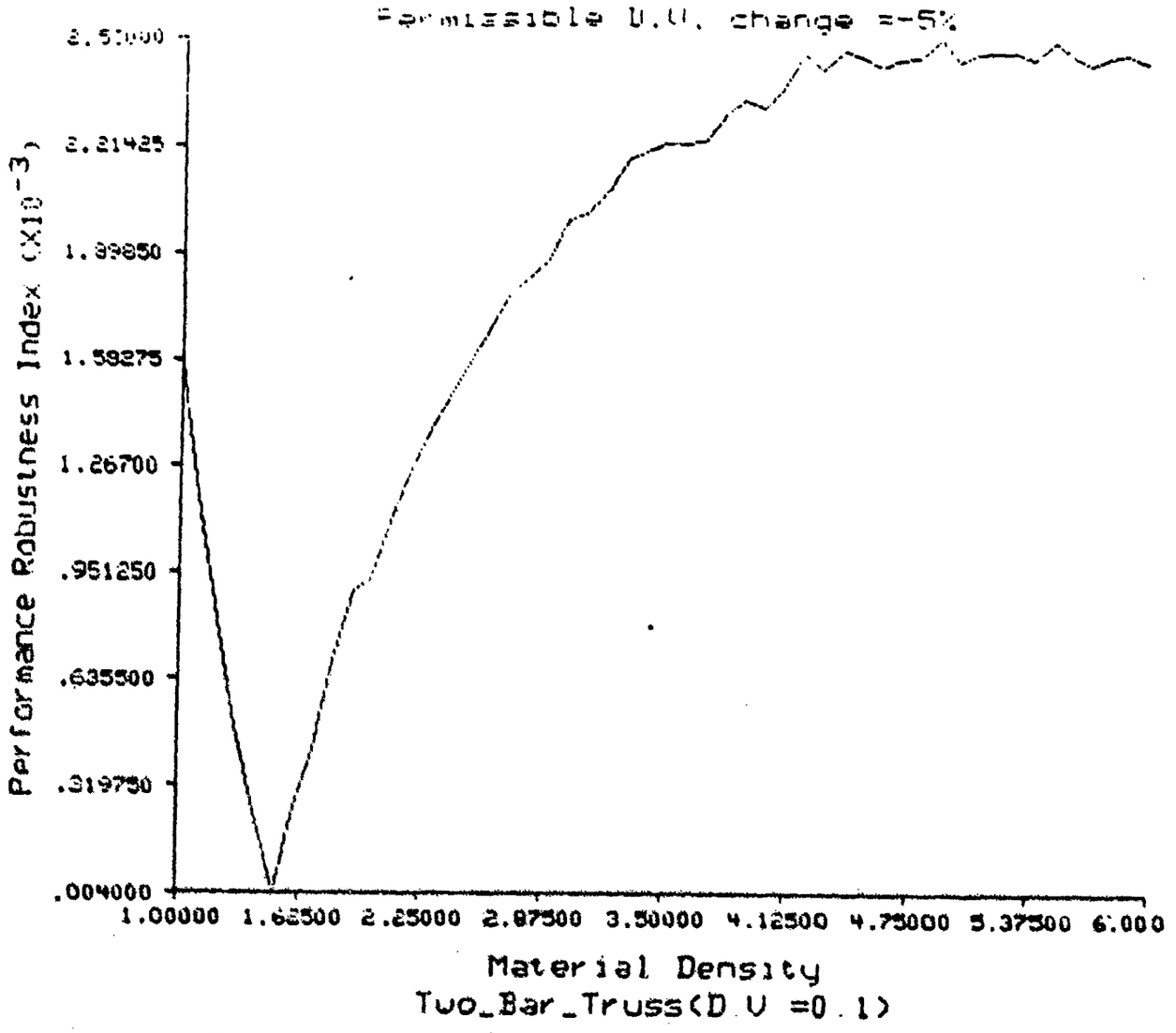


Fig. 15 Variation of β_w with density

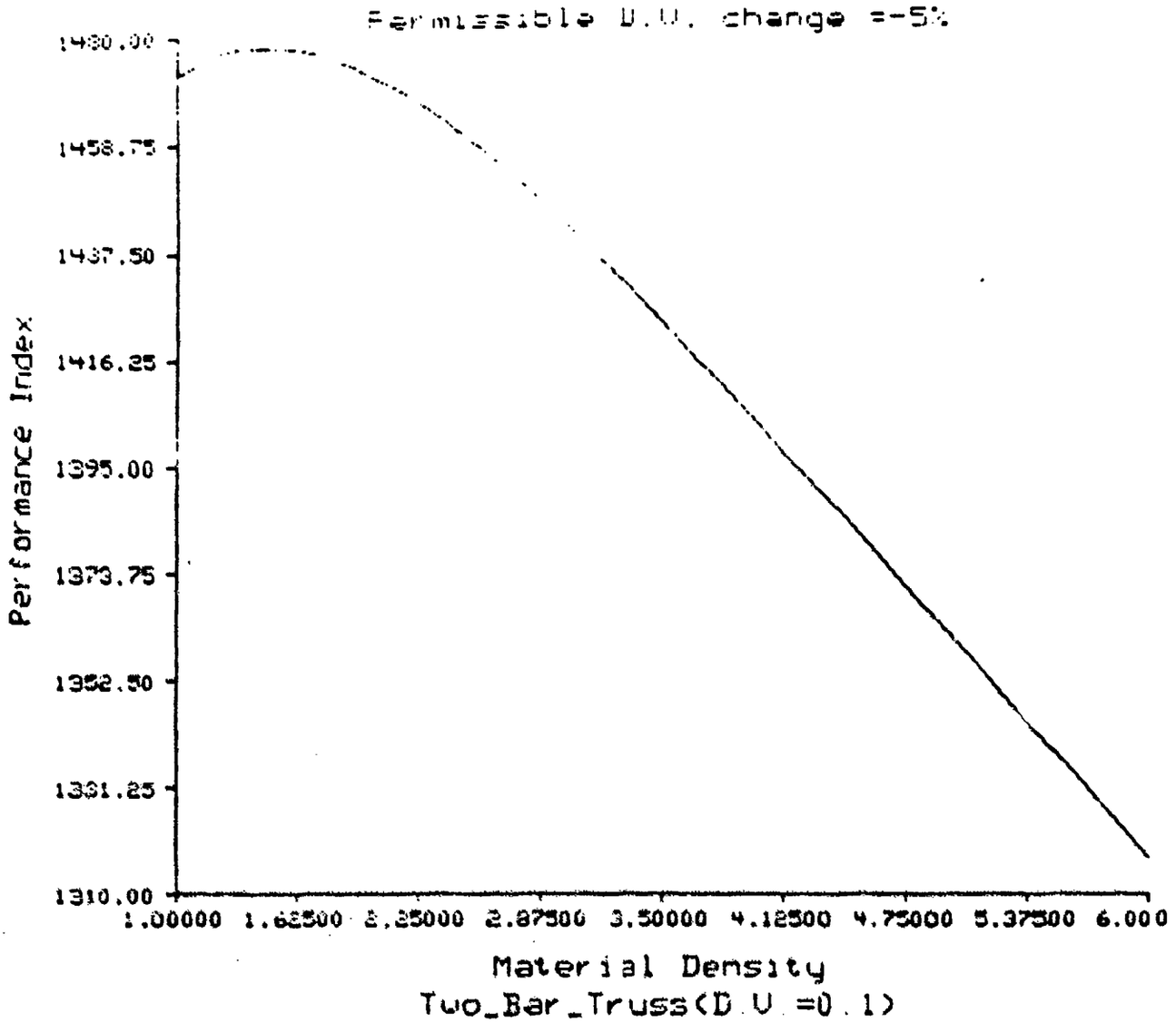


Fig. 16 Variation of performance index with density

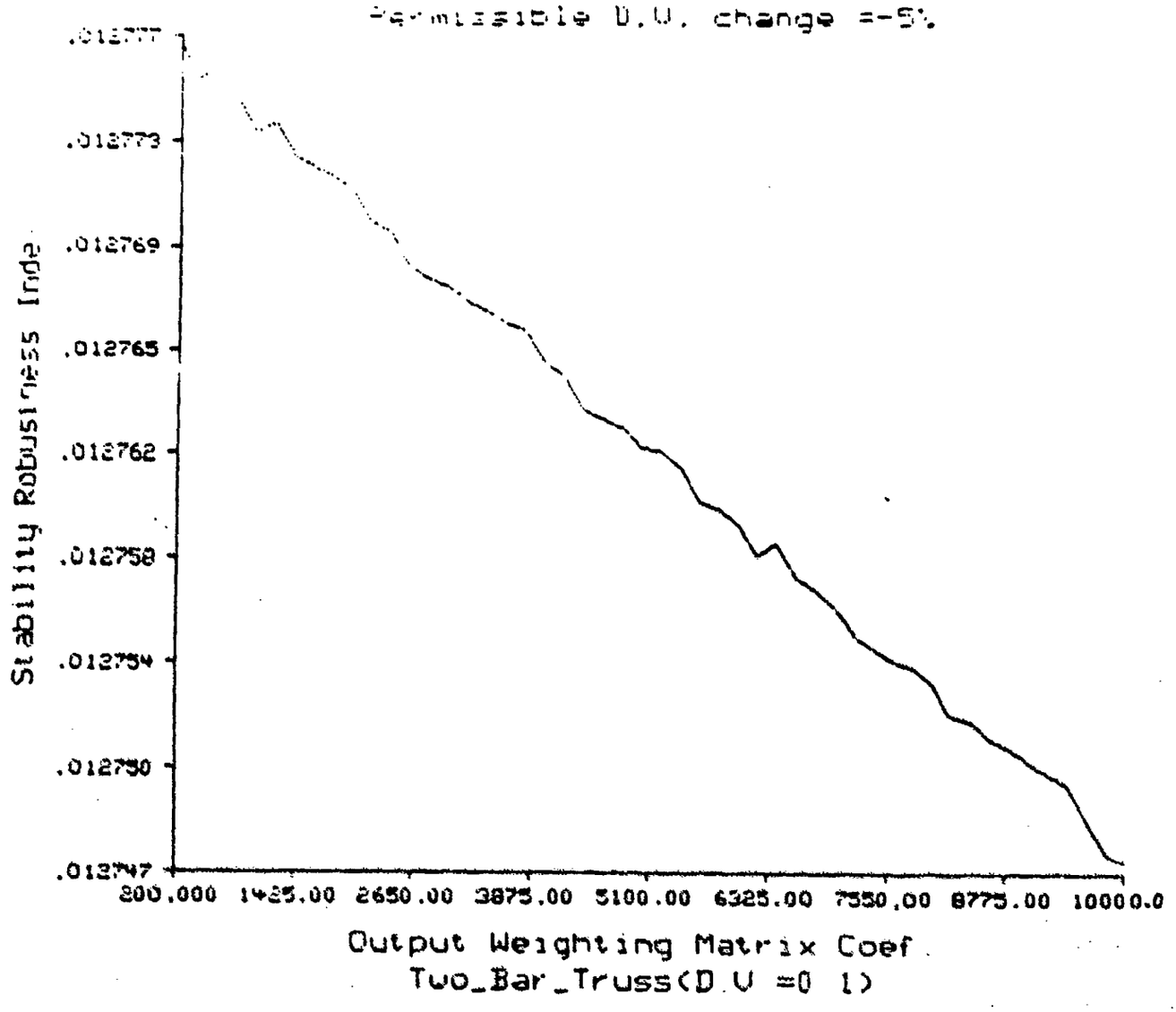


Fig. 17 γ_w versus output weighting matrix coefficient

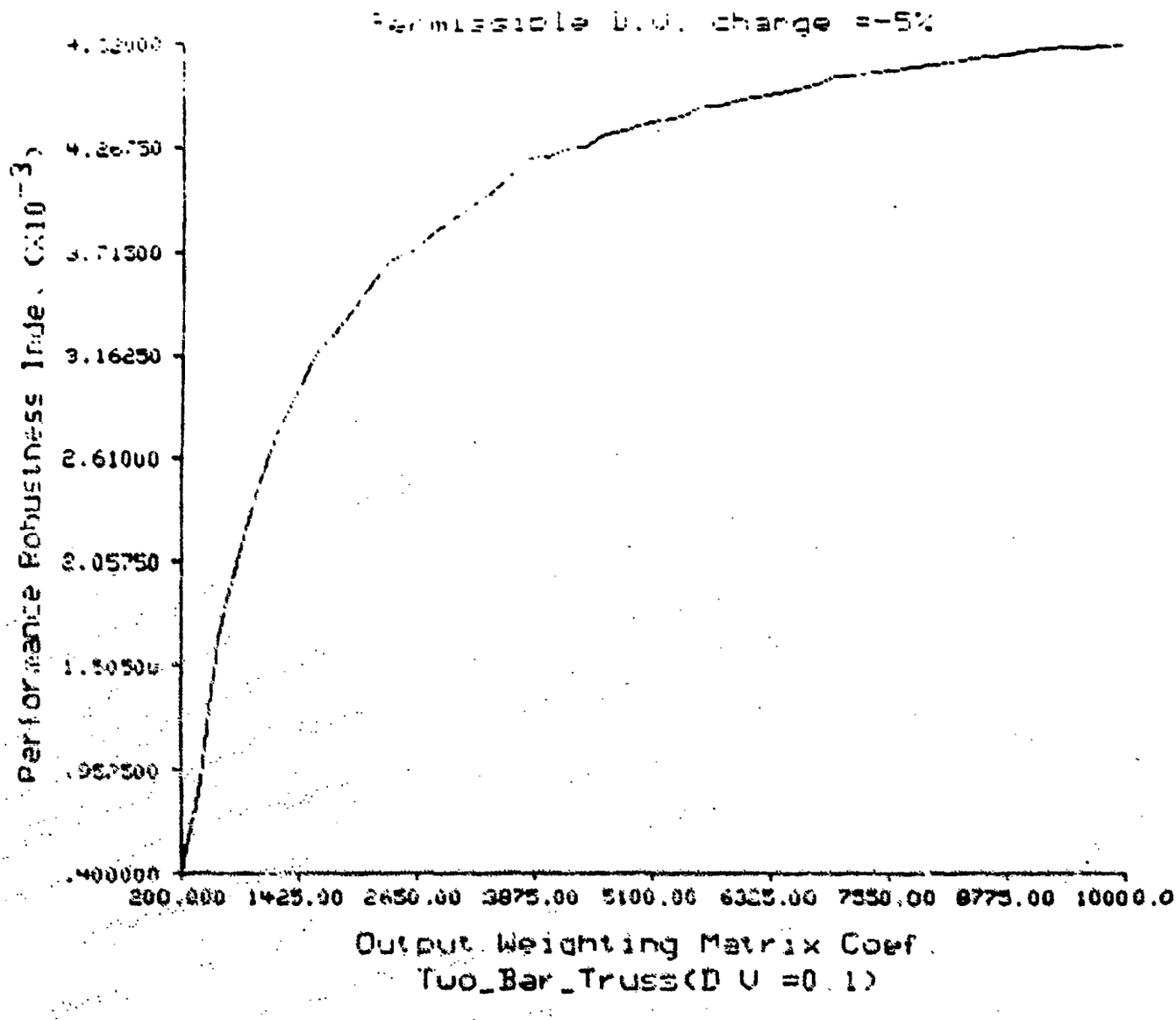


Fig. 18 γ_{pr} versus output weighting matrix coefficient

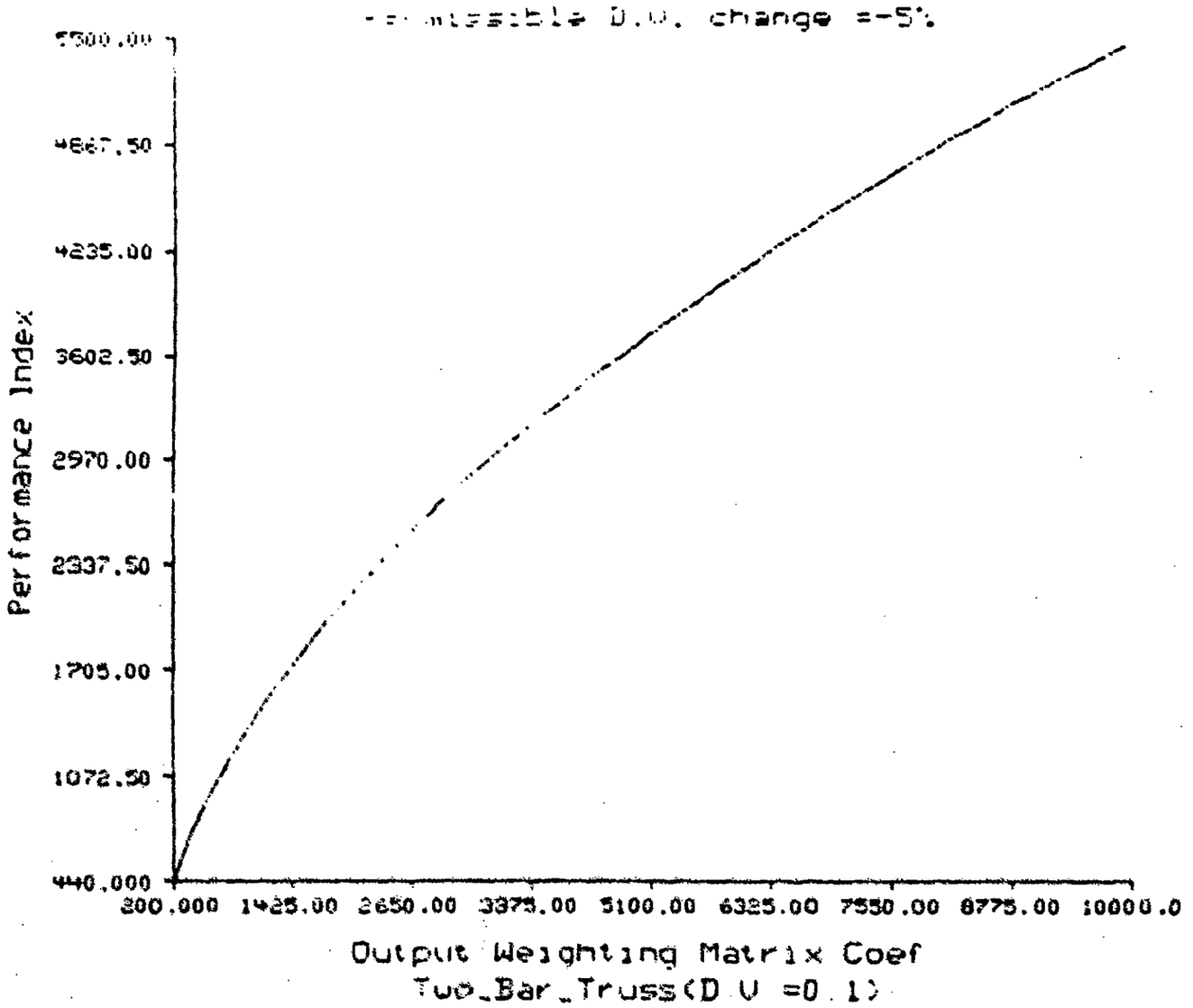
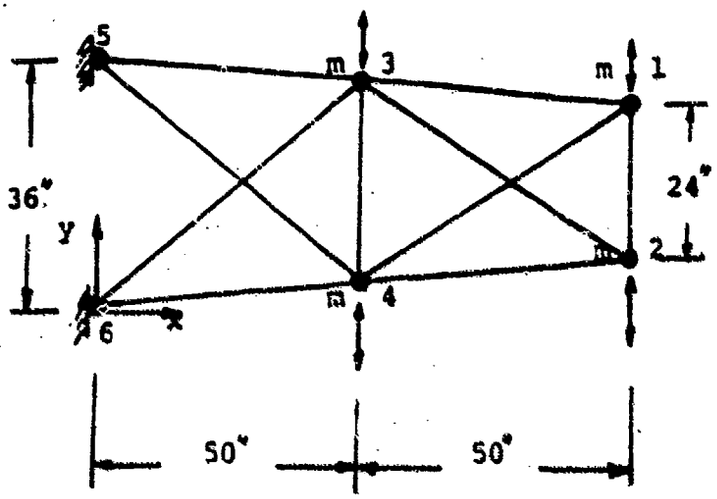


Fig. 19 Performance index versus output weighting matrix coefficient



m = 1.29
E = 10*10**6
Density = 0.1/32.2
Nominal areas = 0.1

Fig. 20 Two-bay truss

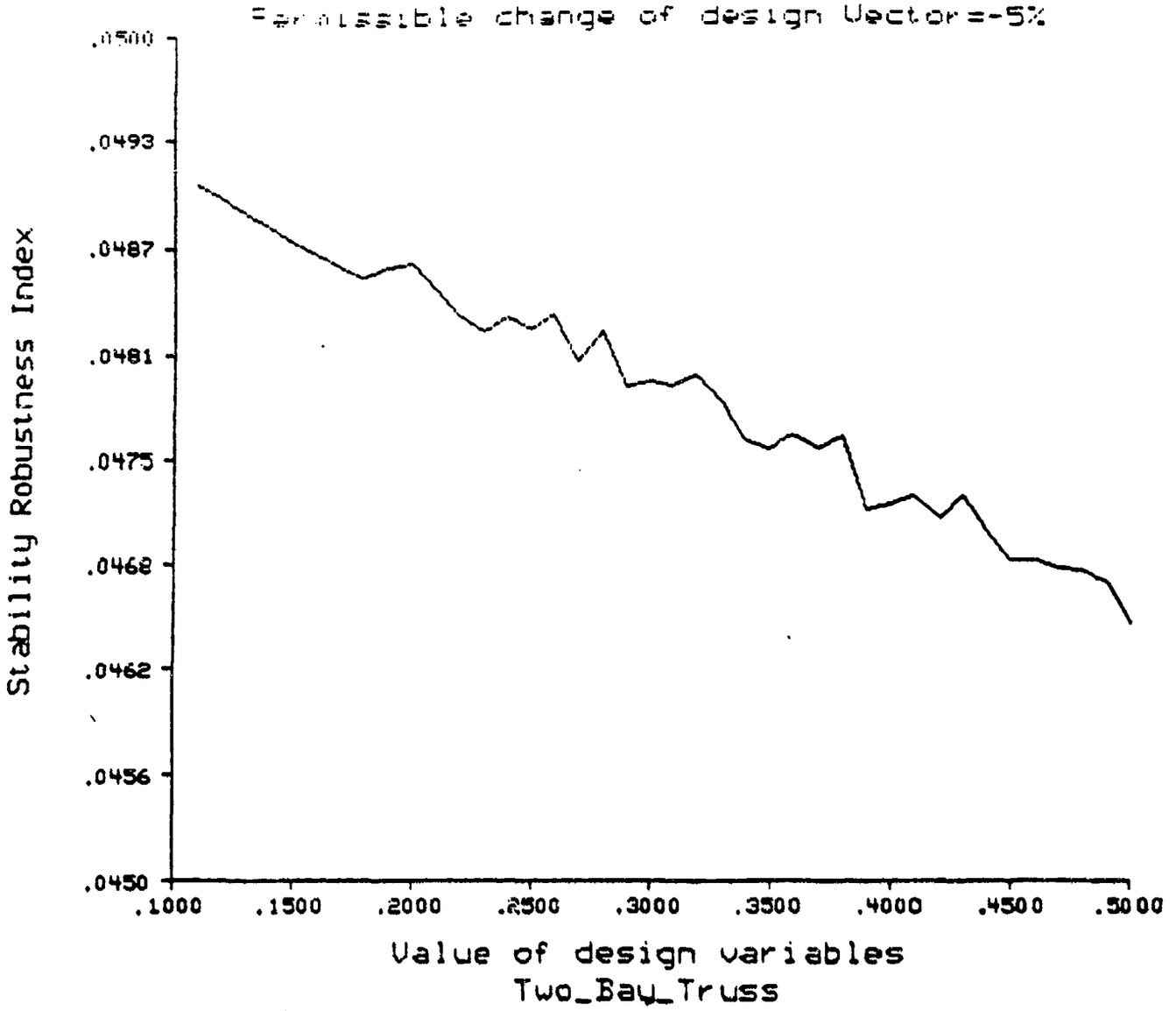


Fig. 21 Stability robustness index vs. value of design variables

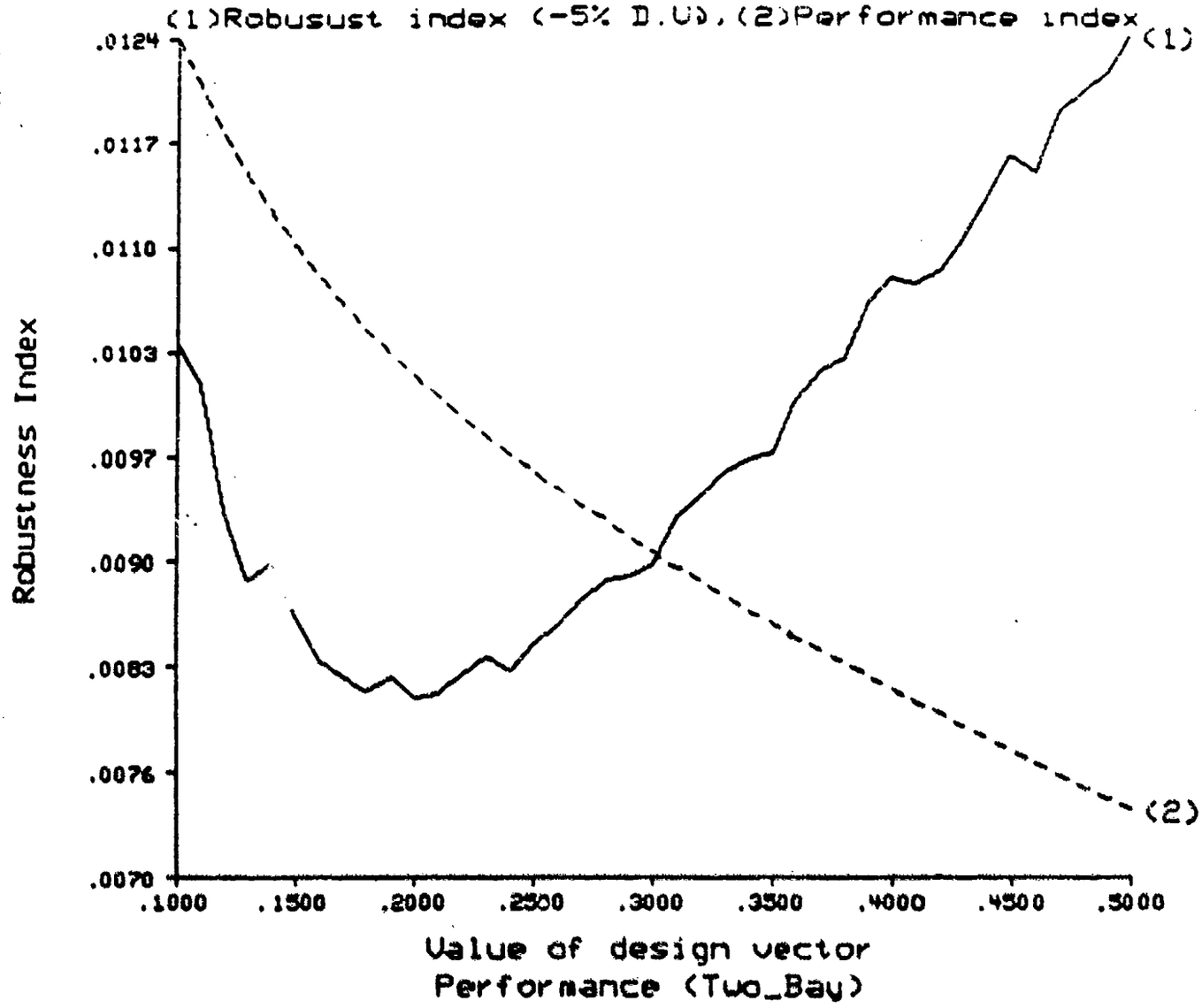


Fig. 22 Robustness index vs. value of design vector

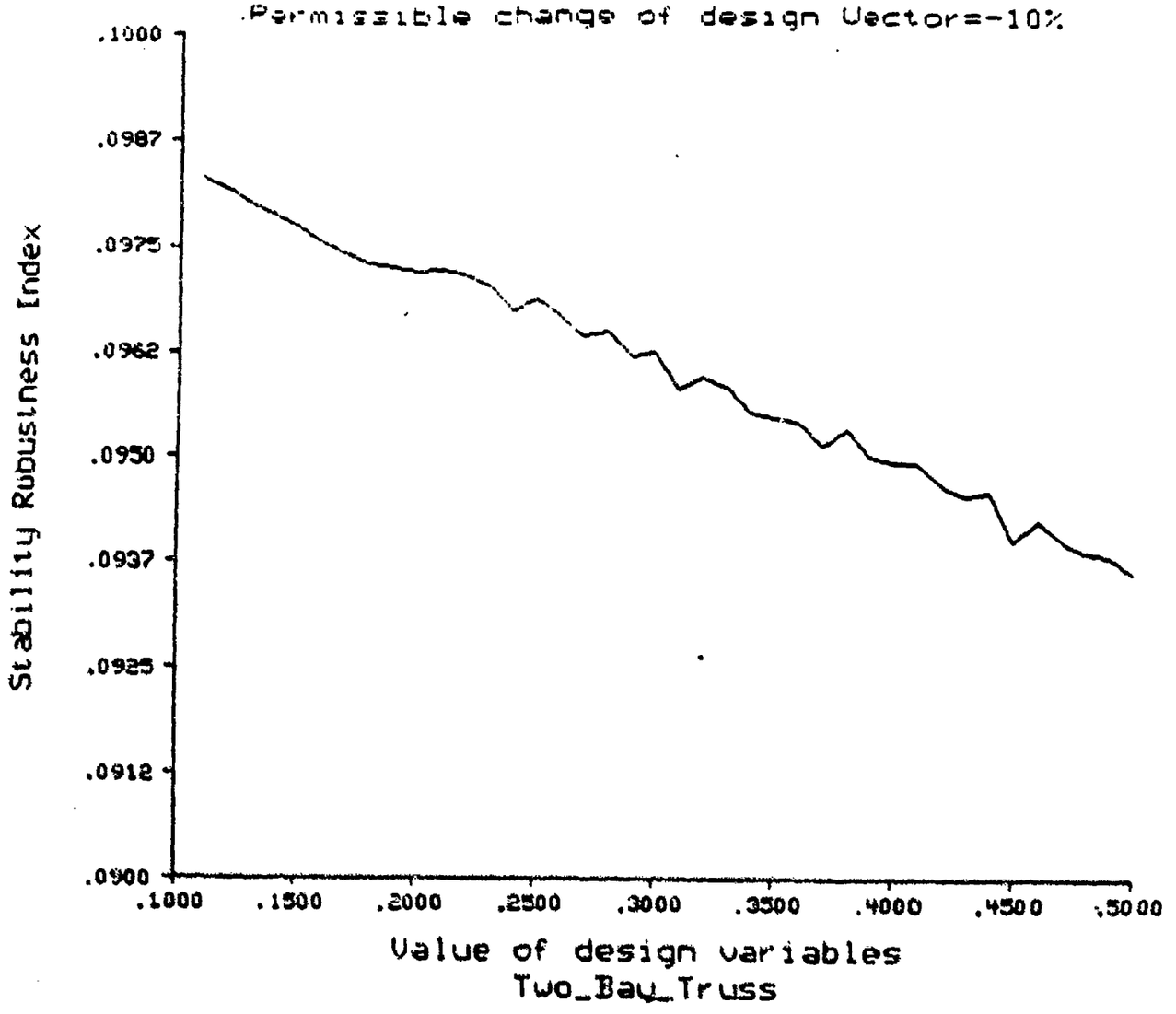


Fig. 23 γ_s versus value of design variables
(Change in design vector = -10%)

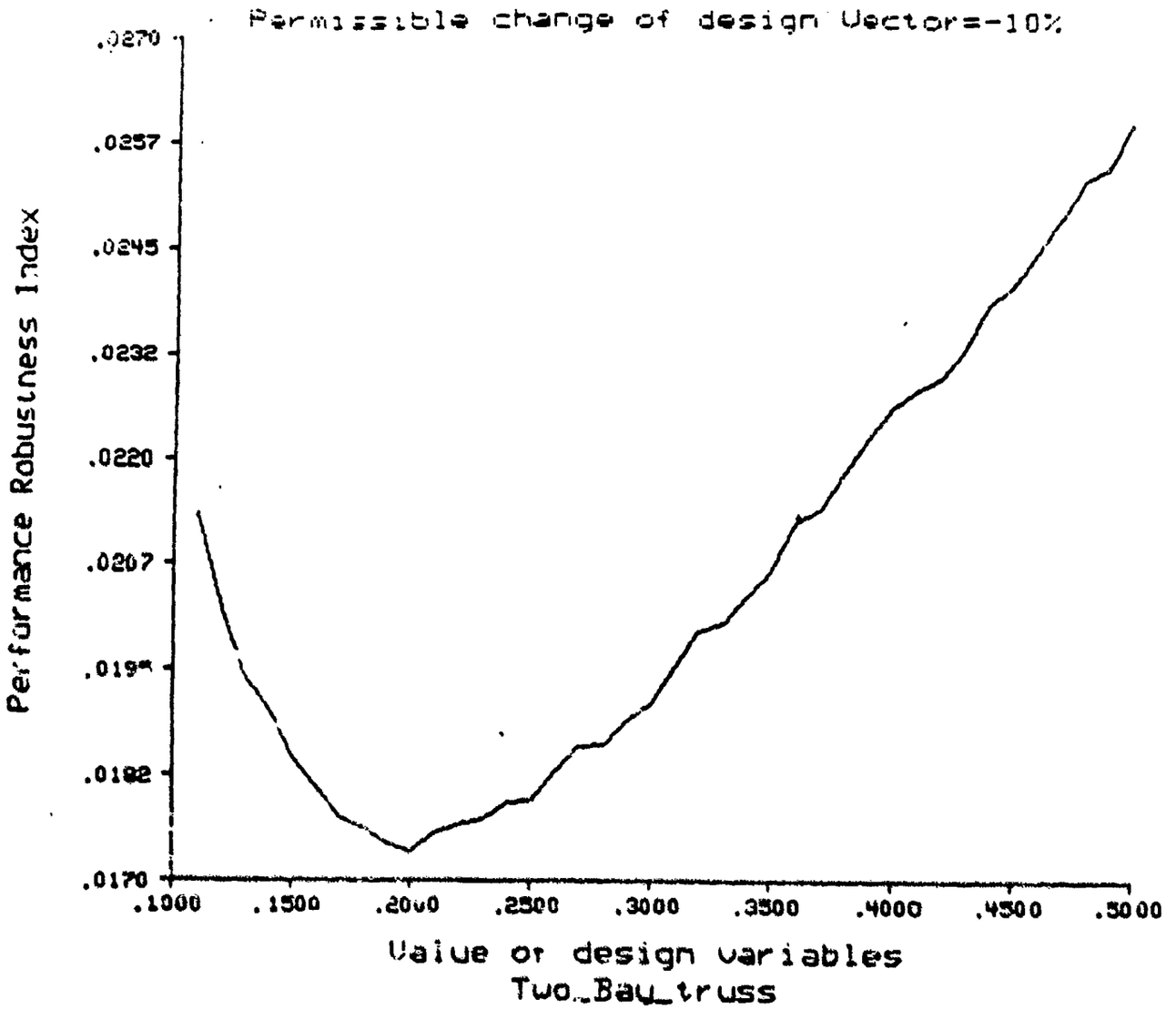


Fig. 24 I_p versus value of design variables (Change in design vector = -10%)

FINAL REPORT NUMBER 64
REPORT NOT RECEIVED IN TIME
WILL BE PROVIDED WHEN AVAILABLE
Dr. Ralph Rascati
760-6MG-062

MINI-GRANT

FINAL REPORT

MATRIXx-BASED COMPUTER SIMULATION OF THE CARDIOVASCULAR SYSTEM UNDER +Gz STRESS

Kuldip S. Rattan
Department of Electrical Systems Engineering
Wright State University
Dayton, OH 45435

Contract No. F49620-85-C-0013/SB5851-0360
Subcontract No. S-760-6mg-104
Title: State Variable Model of the Cardiovascular System under +Gz Stress

MATRIXx-BASED COMPUTER SIMULATION OF THE CARDIOVASCULAR SYSTEM UNDER +Gz STRESS

by

Kuldip S.Rattan
Department of Electrical Systems Engineering
Wright State University
Dayton, OH 45435

ABSTRACT

Acceleration forces in the +Gz direction (head to foot) causes pooling of the blood in the abdomen and legs and reduces the venous return to the heart and upper body. To study this phenomena, a computer simulation of the cardiovascular model under +Gz stress was implemented using HYPER_BUILD of the computer-aided design package MATRIXx. This model incorporates arterial and venous systems, heart, baroreceptors control of the heart rate and venous tones, and inputs for acceleration force and externally applied pressures. Anti-G suit valves, straining maneuvers such as L1 and M1, and seat back angle were added to the model to study their compensatory effects in increasing the +Gz tolerance. Results for each protective mechanism were obtained from the simulation and were compared with those given in the literature.

I. INTRODUCTION

Present-day high performance fighters are designed to generate and sustain acceleration forces at magnitudes and durations that exceed the tolerance limits of their human operators [4]. Acceleration forces in the +Gz direction reduces blood pressure at the eye level causing the pilot to experience peripheral light loss (PLL). Further decrease in the eye level pressure results in central light loss (CLL) and finally complete black-out and loss of consciousness (LOC). In recent years G-induced loss of consciousness (LOC) has been implicated in several aircraft mishaps. Pilots of aircrafts such as F-15 and F-16 , that have very high sustained Gz and high onset rates capabilities , are more likely to be involved in these types of mishaps [4].

The pilot's +Gz tolerance can be increased by applying external pressures to the legs and abdomen using an anti-G suit, straining and tensing the muscels (L-1 and M-1 maneuvers), and increasing the seat back angle. An increased tolerance of up to +2 G may be realized using either one of these techniques [1,2,7,10].

Most of the studies concerned with increasing +Gz tolerance have been done on the Dynamic Environment Simulator (DES) by using human or animal subjects. Centrifugal studies are expensive to run and are difficult to control and reproduce [2]. In addition, the hemodynamic variables needed to objectively assess the +Gz tolerance are hard to measure. Some studies have been done using modeling techniques. The models are used to predict the hemodynamic variables during acceleration stress [2]. Not much work has been done to study the improvement in the +Gz tolerance with protective mechanisms using computer simulation. The objective of this study is to develop a computer simulation of a cardiovascular model in order to study improvement in +Gz tolerance provided by the anti-G suit, increase in the seat back angle, and straining maneuvers such as L-1 and M-1. The model developed by Jaron and Chu [6] was used for this simulation since it incorperates both heart rate and venous tone reflex compensation, and pressures/flows at different parts of the circulatory system can be obtained from the multi-element arterial tree. The model was simulated using HYPER_BUILD of the computer-aided design package, MATRIXx. MATRIXx provides an interactive, menu-driven environment for building, modifying and editing complex computer simulation models. An Anti-G suit, seat back angle, and straining maneuvers such as L-1 and M-1 were added to the model to study their compensatory effects. Results for each protective mechanism were obtained from the simulation and were compared with those given in the literature.

II. THE CARDIOVASCULAR MODEL

The cardiovascular model chosen for this study was the one developed by Jaron and Chu [6]. It consists of a variable compliance left ventricular model driving a multi-element arterial tree. Each element in the arterial tree is affected by the distributed loading of the +Gz stress. The arterial tree is then connected to a lumped systemic venous model which is in turn connected to a lumped pulmonary model. The output of the lumped pulmonary

model feeds back to the heart to complete the circulatory cycle. The pressure at the arterial tree element corresponding to the carotid sinus is monitered continuously by the baroreceptors to drive the control mechanisms. The control mechanisms change heart rate and venous tone (of the lumped systemic model) with the aim of returning the carotid pressure toward normal. Figure 1 shows a block diagram of this model and figure 2 shows the multi-element arterial tree [2].

III. MATRIXx IMPLEMENTATION

The cardiovascular model discussed in the previous section was implemented using the HYPER_BUILD option of the computer-aided design package MATRIXx [8]. MATRIXx is a powerful, programmable, matrix calculator with excellent graphical capabilities which can be used to solve complex, large-scale matrix problems.

HYPER_BUILD provides an interactive, menu-driven graphical environment for building, modifying and editing computer simulation of complex models. Simulating cardiovascular system performance under both nominal and strained environment can be accomplished easily with HYPER_BUILD. HYPER_BUILD also provides modularity in design which makes testing, modifying, and interfacing parts of the model a simple task.

HYPER_BUILD basic building unit is the block and it has a large library of different types of blocks. A very important type of block is the SUPER-BLOCK. This block can have up to six other blocks interfaced in any way. The most important advantage of a SUPER-BLOCK is that it can contain other SUPER-BLOCKS. This permits the nesting of SUPER-BLOCKS as seen in Figure 3. This nesting property makes it possible to implement complex systems which contain any number of blocks.

Figure 4 shows the equivalent circuit of the "A" element. The "A" element is modeled by two nested Super_Blocks. The state-space equations used to implement this segment can be written in the matrix form as

$$\dot{\underline{x}} = A\underline{x} + B\underline{u}$$
$$Y = C\underline{x} + D\underline{u} \qquad (1)$$

where

$$\underline{x} = \begin{bmatrix} FA_n & VA_n \end{bmatrix}^T$$

$$\underline{u} = \begin{bmatrix} PA_{n-1} & PC_n & FA_{n+1} \end{bmatrix}^T$$

$$Y = \left[PA_n \right]$$

$$A = \begin{bmatrix} -\dfrac{R_{1n} + R_{2n}}{L_n} & -\dfrac{1}{C_n L_n} \\ 1 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} \dfrac{1}{L_n} & \dfrac{1}{L_n} & \dfrac{R_{2n}}{L_n} \\ 0 & 0 & -1 \end{bmatrix}$$

$$C = \begin{bmatrix} R_{2n} & \dfrac{1}{C_n} \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & 0 & -R_{2n} \end{bmatrix}$$

and

$$PG_n = (\rho \Delta Z_n \cos\theta_n) \, Gz \qquad (2)$$

= Hydrostatic effect of Gz stress on the n^{th} arterial segment.

θ_n = The angle between the direction of Gz and the orientation of the n^{th} arterial segment.

PA_n = Blood pressure in the n^{th} arterial segment.

FA_n = Blood flow into the n^{th} arterial segment.

VA_n = Volume of blood in the n^{th} arterial segment.

Gz = Acceleration force in the Z direction.

The HYPER_BUILD Super-Blocks used to implement the element "A2" of the arterial tree are shown in figures 5a and 5b. The Super-Block SEGA2, which has as inputs PA_1, FA_3 and Gz, calculates the elements of matrices A and B of equation (1). These are then sent as inputs to the nested Super-Block STATA2.

STATA2 forms the differential equations FA_2 and VA_2 given by equations (1) in the block EQNA2 which are then solved using the integrator FLOW and limited integrator VOLUME. The limited integrator is used because physiologically a segment volume cannot have a negative value. The outputs of STATA2 are PA_2, FA_2, and VA_2.

The "A2" element is then interfaced with the rest of the multi-element arterial tree. Figure 6 shows how SEGA2 is connected to the ascending aorta in Super-Block AAORTA. Figure 7 shows how the AAORTA is interfaced with the rest of the arterial tree in Super-Block ARTERI. The arterial tree is then interfaced with the heart model and the lumped systemic and pulmonary model to form the circulatory system (Figure 8). Figure 9 shows the connection of the circulatory system with the Super-Block GRAVTY which generates the +Gz profile. The Super-Block GSTRES has 100 outputs. Due to memory considerations, not all outputs could be monitored at the same time. For this reason GSTRES is nested in the Super-Block OUTPUT (figure 10). Super-Block OUTPUT can be programmed to monitor up to 11 outputs of the Super-Block GSTRES. The total number of SUPER-BLOCKS needed to implement the cardiovascular model are 72. Variable step Kutta-Merson method was used as the integration algorithm for the simulation which took approximately 72 minutes of CPU time to run a 40 second +Gz profile with simulation step size of 0.0025 second.

IV. UNPROTECTED SYSTEM RESULTS

The outputs of the model under +1 Gz stress were obtained and found to check out with the known physiological values. Figure 11 shows the unprotected system response under +4 Gz gravity profile with an onset rate of 1 G/Sec.. It can be seen from this figure that the pilot will experience both PLL and CLL because the peak systolic carotid pressure at the eye level drops down to 6 mm Hg. However, due to the reflex compensation, the eye level carotid pressure increases to 45 mm Hg and hence his central vision is completely restored as well as most of his peripheral vision.

Figure 12 shows how the peak systolic pressure at the eye level under a +3 Gz gravity profile varies with time for different onset rates. This graph is obtained by fitting the output of the model with exponential curves. It is clear that the minimum peak systolic level tend to have similar time constants. This is due to the fact that at high onset rates, the gravity profile becomes approximately a step function. The cardiovascular system responds to such an input with a similar time constant.

It is also clear from figure 12 that the minimum peak systolic pressure at eye level occurs about 3.5-4.0 seconds after the start of the +Gz profile. Since the venous tone control has a delay of about 5 seconds, it can be concluded that the venous tone control does not play any role in determining the minimum peak systolic pressure and thus does not affect the +Gz tolerance of the pilots (which is determined from the minimum peak systolic pressure at eye level).

V. METHODS OF INCREASING +Gz TOLERANCE

The pilot's +Gz tolerance can be increased by applying external pressures to the legs and abdomen using an anti-G suit, straining and tensing the muscels (L-1 and M-1 maneuvers), and increasing the seat back angle. An increased tolerance of up to +2 G may be realized using either one of these techniques [1,2,7,10].

5.1 ANTI-G SUIT

The function of the Anti-G suit is to increase the pilot's +Gz tolerance by applying pressure to the abdomen and legs. This is especially needed during the first few seconds of the gravity profile when the venous tone control is inactive. The protective G suit garment consists of airtight bladders which are filled with pressurized air delivered from a G sensitive mechanical valve [9].

Figure 13 shows the model of the Anti-G suit. It consists of a transport delay and a first-order lag. The transport delay is due to the connecting air feed line from the valve to the suit and is of the order of 5 msec. [9]. The bladder size and garment tightness govern the lag term which has a time constant of about one second [9]. Since the transport delay time is negligible compared to the suit time constant, the G suit model can be simplified by the first-order lag and can be written as

$$\frac{Pcuff(S)}{Pvalve(S)} = \frac{1}{S + 1} \qquad (3)$$

where Pcuff is the suit pressure in psi and Pvalve is the output pressure of valve in psi.

5.1.1 HIGH-FLOW VALVE

The High-Flow (HF) valve [2] starts operating once the gravity profile exceeds 2.0 G. Once triggered, the valve's output pressure is proportional to the gravity profile. The HF valve can be modeled by a piece-wise linear function

$$P_{HF} = 0 \qquad \text{for } Gz \leq 2$$
$$P_{HF} = 1.8\ (Gz-2) = 1.8\ Gz_m \qquad \text{for } Gz \geq 2 \qquad (4)$$

where P_{HF} is the output pressure of High-Flow valve in psi and Gz_m is the modified acceleration force in the Z-direction. Thus, the HF valve represents an open-loop proportional type of control mechanism.

5.1.2 BANG-BANG SERVO VALVE

A new valve under development is the Bang-Bang Servo (BBS) valve [11]. The BBS valve is a modified version of the High-Flow valve. The modification consists of a solenoid mounted on top of the HF valve. When the Gz profile has a G level $\leq$ 2 G or an onset rate $\leq$ 2 G/sec., the valve responds in the High-Flow mode. If the Gz profile has a Gz level $\geq$ 2 G and an onset rate $\geq$ 2 G/sec., the solenoid is activated causing the valve to operate at its maximum rate. The valve continues to operate in this mode for 1.5 seconds following the triggering of the solenoid. After that, the valve returns to the High-Flow mode of operation. This results in the suit pressure reaching a peak value of about 11 psi. Figure 14 shows the pressure profiles of the BBS valve due to gravity profiles having 3 G/sec. onset rate. It is clear that the valve represents an open-loop proportional and derivative type of control mechanism.

The motivation behind designing the BBS valve is to make the Anti-G suit filling schedule sensitive not only to the gravity profile level but also to its onset rate. This gives the valve a predictive edge and thus improves the performance of the Anti-G suit. It can be seen from figure 14 that the BBS has a very short rise time. However, at low sustained G, if the triggering criteria is met, the BBS valve results in a very large overshoot which could be uncomfortable and even painful to the pilots. Also, the valve uses an analog differentiator to detect the onset rate of the G profile. The performance of such a differentiator is greatly affected by noise and may be unreliable in practical situations.

5.1.3 PROPOSED CLOSED-LOOP SERVO ANTI-G VALVE

To overcome some of the design problems associated with the BBS valve, a preliminary closed-loop controller design is proposed (figure 15). The parameter that is monitered in the feedback is the suit pressure. The compensators are designed to satisfy the following specifications

1. Steady-state value of the suit presuure to unit step Gz_m input is 1.8 psi.
2. Short rise time.

The closed-loop transfer function of the compensated system shown in figure 15 can be written as

$$\frac{Pcuff(S)}{Gz_m(S)} = \frac{K\,(\,K_{P1}\,S + K_{I1}\,)}{S^2 + (\,K\,K_{P2} + 1\,)\,S + K\,K_{I2}} \qquad (5)$$

From the steady-state suit pressure requirement

$$K_{I1} = 1.8\, K_{I2} \qquad (6)$$

Selecting the damping ratio $\zeta = 0.707$ and the natural frequency $w_n = 3$ rad./sec., the compensators coefficients that satisfy equations (5) and (6) are : $K = 9$, $K_{P1} = 1.11$, $K_{I1} = 1.8$, $K_{P2} = 0.362$, $K_{I2} = 1$ and the overall transfer function is given by

$$\frac{Pcuff(S)}{Gz_m(S)} = \frac{10\,S + 16.2}{S^2 + 4.242\,S + 9} \qquad (7)$$

Figure 14 shows the pressure profiles of the closed-loop servo (CLS) valve in comparison with the HF and BBS valves under Gz profile having 3 G/sec. onset rate. It can be seen from this figure that the CLS rise time is simillar to the BBS. However, the CLS valve does not suffer from the excessive overshoot at low sustained +Gz. Also the controllers used in the design are of the proportional and integral type (PI) and thus will overcome the noise problems of the BBS.

5.1.4 SIMULATION RESULTS WITH ANTI-G SUIT

To incorporate the effects of the Anti-G suit in the model, the femoral and abdominal elements of the multi-element arterial tree were modified by adding a pressure source equivalent to the suit pressure. Model results were obtained with and without the Anti-G suit bladder dynamics. To measure the +Gz tolerance in an objective manner, the minimum peak systolic eye level carotid pressures are plotted in figures 16 - 18. The measures taken for tolerance are

PLL: Starts when the systolic blood pressure at eye level falls below 50 mm Hg.

CLL: occurs when the systolic blood pressure at eye level falls below 20 mm Hg.

It can be seen from figure 16 that the standard Anti-G suit model (with no bladder dynamics) improves the CLL tolerance by 1.5 G.. Although this compares favorably with experimental findings [5], the result is somewhat unexpected since the Anti-G suit model used does not take into account the time lag due to suit filling. This could be due to a flaw in the modeling of the femeral and hepatic "B" elements of the multi-element arterial tree. Another reason may be due to the fact that the venous part of the circulatory system is represented by a lumped model. Thus accurate blood distribution in the veins under acceleration stress or external applied pressure is not possible. For the Anti-G suit model with bladder dynamics, it can be seen from figure 17 that for an onset rate of 1 G/sec., the HF and BBS valves

improve the CLL tolerance by about 0.8 G while the CLS valve improves the CLL tolerance by about 1.1 G. The HF valve has been reported to increase CLL tolerance by 1.5 - 2.0 G [11] . This inaccuracy should be expected due to the reasons mentioned above.

For an onset rate of 3 G/sec. (figure 18), the BBS valve increases the CLL tolerance by about 0.5 G above the HF valve. This result does follow experimental findings reported by Van Patten et al [11]. The CLS valve performance in this case is slightly inferior to the BBS valve. However, since the CLS valve performance is sensitive to the onset rate of the G profile, at onset rates > 3 G/sec., it is expected that the CLS valve performance will be closer to, if not better than, the BBS valve performance.

5.2 SEAT BACK ANGLE

It is well known from the literature that increasing the seat back angle increases tolerance to +Gz acceleration. There are two types of reclining seats [2]: **Tilt-Back Seat** (head is lowered) and the PALE (Pelvis and Legs elevating) seat (Figure 19). Both seats have been found to increase the +Gz tolerance. This section investigates the effects of the seat back angle in improving the +Gz tolerance and compares the results with those found in the literature.

Two factors that cause the reduction of the carotid pressure at the eye level are the length of the vertical hydrostatic column between the heart and the brain and the length of the vertical hydrostatic column (referenced to the heart) that causes blood pooling in the lower extremities. Increasing the reclination of the pilot's seat shortens these distances. However, the vertical hydrostatic column distance from the eye to the aortic arch increases as the seat back angle increases from 0° to 30°. this distance reduces below the control value only when the seat back angle becomes larger than 30° (Table 1 of Chu [2]). In this case, the pumping of the blood in the heart-brain column increases and hence the preservation of the cerebral blood flow is maintained. In addition, tilting the seat back reduces the blood pooling tendency in the lower part of the body and hence the venous return is increased. The carotid eye level pressure increases significantly at higher tilt angles and thus significantly increasing the tolerance to +Gz acceleration.

Two seat back angles were used for simulation in this study : 52° and 67° from the vertical. These angles represent actual seats made with angles of 45° and 60° in an aircraft with an assumed angle of attack of 7°. These reclinations were compared with the conventional seat back angle of 17° from the vertical (including the angle of attack). Lisher and Glaister [7] have reported that a 52° and a 67° seat back angle increases tolerance by 0.3 G and 1.4 G respectively, over the 17° conventional seat back angle.

5.2.1 MATRIXx IMPLEMENTATION AND RESULTS OF SEAT BACK ANGLE

The "A" segments of the arterial tree were modified to include the seat back angle. This was done simply by modifying the hydrostatic pressure given in equation (2).

$$PG_n = \rho\, G_z\, \Delta Z_n \cos(\theta_n + \theta_s) \qquad (8)$$

where θ_s is the seat back angle with respect to the vertical. Figures 20-23 show the simulation results due to +4 Gz gravity profile having an onset rate of 1 G/sec using a standard valve and seat back angles of 0°, 17°, 52°, and 67°. It can be seen from these figures that the responses for 17°, 52°, and 67° seat back angles have shapes similar to the one for 0° seat back angle. It can also be seen from these figures that the drop in the steady-state systolic pressure at eye level decreases with higher seat back angles. This is expected since the hydrostatic pressure column, the primarily reason for the steady-state systolic drop, decreases at higher seat back angles. In addition, figures 20 and 21 show that the minimum systolic peak pressure decreases as the seat back angle increases from 0° to 17°. This result agrees favorably with the fact that seat back angle has negative effect in improving +Gz tolerance for angles between 0° to 30°.

Figure 24 shows that the 52° and 67° seat back angles improves the CLL tolerance by about 2.1 G and 4.2 G respectively over the conventional seat back angle for onset rate of 1G/Sec.. The maximum protection that the seat back is expected to provide is at an angle of 90°. In this case the head, heart, and legs lie in the same horizontal plane (prone position) [2]. Figure 25 shows the protected response for +5 Gz with an onset rate of 3 G/sec. and for a seat back angle of 90°. It clearly shows that pilot maintains an adequate vision since the minimum peak systolic carotid pressure is larger than the PLL threshold by 50 mm Hg and hence this posture is an effective way to increase the +Gz tolerance. However, this position is impractical for combat maneuvers.

5.3 STRAINING MANEUVERS

Straining and tensing the musceles is another effective method in raising the CLL threshold. These maneuvers act similar to an Anti-G suit in providing pressures to the abdomen and legs. The only difference is that these straining maneuvers provide an internal instead of external pressure.

5.3.1 M-1 MANEUVER

M-1 maneuver is commonly refered as the "grunt" maneuver since it approximates the physical effort required to lift a heavy weight. This

maneuver basically consists of pulling the head down between the shoulders, slowly and forcefully exhaling through a partially closed glottis, and simultaneously tensing all skeletal muscles [9]. Pulling the head downward shortens the vertical head-heart distance, exhaling forcefully increases the intrathoracic pressure and tensing the abdominal and peripheral muscles raises the diaphragm and compresses the capacitance vessels [9]. To have an optimal effect, the manuever must be repeated once every 4-6 seconds with an active period for 3-5 seconds. This corresponds to holding the lung pressure high for 3-5 seconds ("grunt" phase). The active period is followed by a rapid exhalation and inhalation period for 1 second (figure 26). During the exhalation phase, the intrathoracic pressure drops down to about 50% of its maximum. Since this leads to a drop in the eye level arterial pressure, the pilots are trained to exhale and inhale as fast as they can. Rogers [9] has reported that when the M-1 manuever is properly executed, the intrathoracic pressure increases from 50 to 100 mm Hg. This causes the arterial blood pressure at eye level to increase which results in an increase in the +Gz tolerance of at least 1.5 G. Burton et al [1] have also reported that this maneuver increases +Gz tolerance of up to 2.4 G.

A study in 1985 by Cote, et al. [3] indicated that larger inspiratory volumes enable the generation of larger intrathoracic pressures. Since the subject expels air during the "grunt" phase of the M-1 maneuver, his lung volume decreases. As a result, the intrathoracic pressure falls down to about 95% of its maximum (figure 26) causing the eye level carotid pressure to slightly decrease and hence the +Gz tolerance is decreased. Even though this study was done in 1 G environment, it shows the disadvantage of grunting while exhaling in the M-1 maneuver. In addition, this maneuver distracts the pilot from doing his job properly because of the noise level and laryngeal irritation created by the forced exhalation with a partially closed glottis [10].

5.3.2 L-1 MANUEVER

The L-1 maneuver is similar to the M-1 maneuver except the exhalation phase is done against a completely closed glottis (valsalva maneuver) while tensing all skeletal muscles [10]. Figure 27 shows phases of the L-1 maneuver where the relative intrathoracic pressure magnitude is kept constant during the active phase of L-1 maneuver. A 1972 study by Shubrooks and Leverett [10] reported that L-1 maneuver, if done correctly in conjunction with use of the anti-G suit, result in an equivalent increase in +Gz tolerance of at least 1.5 G.

The increase of intrathoracic pressure, as a result of the straining maneuvers, raises the blood pressure at eye level. Therefore, maximizing intrathoracic pressure during the straining maneuver should help provide the optimum G protection. An experiment [3] was done for eight subjects performing the L-1 maneuver in a 1 G environment to investigate the intrathoracic pressure relationship with inspiratory volumes. The averaged peak intrathoracic pressure was found to be 108 mm Hg. No work has been done to investigate the intrathoracic pressure magnitude under higher +Gz acceleration. It is expected that under +Gz stress, the maximum intrathoracic pressure would decrease since the inspiratory volume decreases as +G

increases. The previous fact is not yet verified, therefore the maximum intrathoracic pressure was kept constant (independent of +Gz acceleration) in this study.

5.3.3 MATRIXx IMPLEMENTATION AND RESULTS OF STRAINING MANEUVERS

Straining maneuvers, including muscle tensing, were incorporated into the cardiovascular model as equivalent external pressure sources applied between the compliance of the segment of interest and the ground. Figure 28a shows the modified equivalent circuit of "A" element. P_l represents the pressure generated by the lungs which effect the great vessels of the thorax, the great veins lacking valves, and the heart chambers (A2-A8 and A17 segments). P_{ab} represents the intraabdominal pressure (A9-A12 segments). The equations that describe P_l and P_{ab} are given by

$$P_l = PR \cdot P_{lmax} \tag{9}$$

$$P_{ab} = PR \cdot P_{abmax} \tag{10}$$

where

$P_{lmax} = P_{abmax} = 108$ mm Hg

PR = Relative magnitude of P_{lmax} or P_{abmax} for M-1 or L-1 profile shown in figures 26 and 27.

This effective lung or abdomen pressure was added to the state-space equation of "A" element by modifying the input vector, $\underline{u}$ and the matrices B and D of equation (1) as shown below.

$$\underline{u} = \begin{bmatrix} PA_{n-1} & PG_n & FA_{n+1} & PR \end{bmatrix}^T$$

$$B = \begin{bmatrix} \frac{1}{L_n} & \frac{1}{L_n} & \frac{R_{2n}}{L_n} & \frac{P_{max}}{L_n} \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & 0 & -R_{2n} & P_{max} \end{bmatrix}$$

Where, P_{max} is either P_{lmax} or P_{abmax}.

Figure 28b shows the modified equivalent circuit of "B" elements of Hepatic, Renal, and Femoral segments as part of the muscular tensing. The equation which describes the muscular tensing is

$$P_m = PR \cdot P_{mmax} \qquad (11)$$

where

P_{mmax} = 70 mm Hg (for the Hepatic and Renal segments)
= 100 mm Hg (for the Femoral segment)

Figure 29 shows the Super-Block M1/L1 which generates the M-1 or the L-1 maneuver profile connected to the circulatory system. Since PLL occurs at about +3 Gz stress for a subject wearing a standard anti-G suit (figure 16), both L-1 an M-1 maneuvers were started when the +Gz profile reached +3 Gz. CPU time for the simulation under M-1 or L-1 maneuver did not increase significantly since only one Super-Block was added to the cardiovascular system.

Figure 30 shows the protected simulation results for L-1 maneuver due to +4 Gz stress with an onset rate of 1G/sec.. It can be seen from this figure that exhalation-inhalation phase between repeated L-1 maneuvers causes a sudden fall in arterial pressure at eye level. This is expected since a quick forcefull exhalation (as part of L-1 maneuver) causes a sudden drop in the intrathoracic pressure. Figure 30 also shows that the pilot initially experiances PLL after exhaling forcefully during the L-1 maneuver. However, his peripheral vision is restored due to the active phase of L-1 maneuver (holding the breath) and the venous tone control.

Figure 31 shows the protected simulation results for M-1 maneuver due to +4 Gz with an onset rate of 1G/Sec.. It can be seen from figures 30 and 31 that L-1 and M-1 maneuvers have similar effect on the eye level carotid pressure. However, since the active ("grunt") phase of M-1 maneuver causes the intrathoracic pressure to drop slightly from its maximum, the eye level carotid pressure also decreases slightly (about 2-5 mm Hg) as compared to L-1 maneuver.

Figure 32 shows that L-1 and M-1 maneuvers, if performed in conjunction with an anti-G suit, increases the +Gz tolerance of the unprotected model by 1.4 G and 1.6 G respectively. These results compare favorably with the experimental findings [10]. Even though both L-1 and M-1 maneuvers provide almost an equal +Gz protection, L-1 maneuver is preferred over the M-1 maneuver becuase of the distraction and laryngeal irritation caused by the M-1 maneuver.

VI. SUMMARY AND CONCLUSIONS

A computer simulation of the cardiovascular system under +Gz stress is carried out in this report. This was done to study improvement in +Gz tolerance provided by an anti-G suit, increase in seat back angle, and straining maneuvers such as L-1 and M-1. The model developed by Jaron and Chu was used for this simulation. An anti-G suit, seat back angle, and straining maneuvers were added to the model to study their compensatory effects. HYPER_BUILD of the computer-aided design package MATRIXx was used to simulate this model.

It was found that the minimum peak systolic pressure at eye level occurs about 3.5-4.0 seconds after the start of the +Gz profile. Since the venous tone control has a delay of about 5 seconds, it was concluded that the venous tone control does not play any role in determining the minimum peak systolic pressure and thus does not effect the +Gz tolerance of the pilots. It was also found that High Flow and Bang-Bang Servo valves improves the Central Light Loss tolerance by about 0.8 G for an onset rate of 1G/sec.. The BBS valve increases the CLL tolerance by about 0.5 G above HF valve for an onset rate of 3G/sec.. These results compares favorably with experimental findings.

It was also found that increasing the seat back angle from 0° to 30° has a negative effect in improving +Gz tolerance. However, increasing the seat back angle further increases CLL tolerance by 2.1 G and 4.2 G for 52° and 67° seat back angles, respectively. A seat back angle of 90° provides an optimum +Gz protection. M-1 and L-1 maneuvers when done in conjunction with standard anti-G suit resulted in an improvement of 1.4 G and 1.6 G, respectively. These results agrees favorably with those found in the literature. L-1 maneuver is prefered over the M-1 maneuver because of the distracion caused by M-1 maneuver.

ACKNOWLEDGEMENTS

The author hereby expresses his gratitude to the Air Force Systems Command, the Air Force Office of Scientific Services Bolling, AFB, DC and the Universal Energy Systems for providing him the opportunity to perform this research under contract no. F49620-85-C-0013.

The author would like to thank Dr. Daniel W. Repperger of AARML/BBS for his help during the course of this research. He would also like to express his thanks to Mr. J. W. Frazier and Sgt. L. Tripp of AARML/BBS for helpful discussions. Mr. Khalid Barazanji of Wright State University was a Graduate Research Assistant for this project. He would like to thank him for his help during the course of this work and for typing this final report.

REFERENCES

1. Burton, R.R., S.D. Leverett, and E.D. Michaelson, "Man at High Sustained +Gz Acceleration: a review," Aerospace Med., 45(10):1115-1136, 1974.

2. Chu, C., "A Mathematical Model of the Cardiovascular System Under +Gz Stress," Ph.D. Dissertation, Drexel University, 1984.

3. Cote R, L. Tripp, T. Jennings, A. Karl, C. Goodyear, and R. Wiley, "Effect of Inspiratory Volume on Intrathoracic Pressure Generated by An L-1 Maneuver," Aviat. Space Environ.. Med., 57:1035-8.

4. Frazier, J.W., R.E. Van Patten, T. Jennings, C.D. Goodyear, D. Ratino and W. Albery, "Evaluation of an Electro-Pneumatic Anti-G Valve in a High Onset Rate Acceleraton Environment," AAMRL-TR-85-051, 1985.

5. Howard, P., "The Physiology of Positive Acceleration," Chapter 23 in: A Textbook of Aviation Physiology, J.A. Gillies, ed., New York, Pergamon Press, 1965.

6. Jaron, Moore and Chu, "A Cardiovascular Model for Studying Impairment of Cerebral Function During +Gz Stress," Aviat. Space Environ. Med., 55(1) 24-31, 1984.

7. Lisher, B.J. and D.H. Glaister, "The Effect of Acceleration and Seat Back Angle on Performance of a Reaction Time Task," Flying Personnel Research Committee, March 1978.

8. "MATRIXx User's Guide," Integrated Systems, Inc., 1986.

9. Rogers, D.B., "A Model for the Energetic Cost of Acceleration Stress Protection in the Human," AMRL-TR-79-58, 1979.

10. Shubrooks, S.J., Jr., and S.D. Leverett, Jr., "Effect of the Valsalva Maneuver on Tolerance to +Gz Acceleration," Appl. Physiol., 34(4):460-466, 1973.

11. Van Patten, R.E., T.J. Jennings, W.B. Albery, J.W. Frazier, C. Goodyear, B. Gruesbeck and D. McCollor, "Development of a Bang-Bang Servo Anti-G Valve for High Performance Fighter Aircraft: Final Report," AFAMRL-TR-85-024, 1

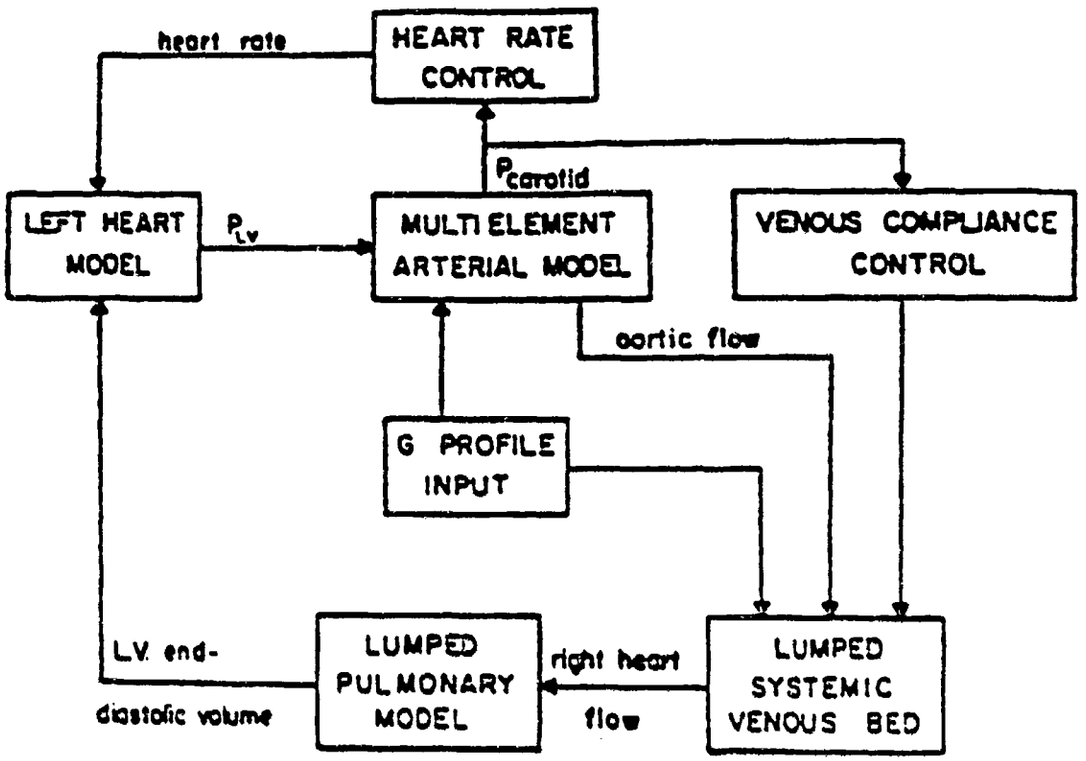


Figure 1. Block Diagram of the Cardiovascular System Model (Chu [2])

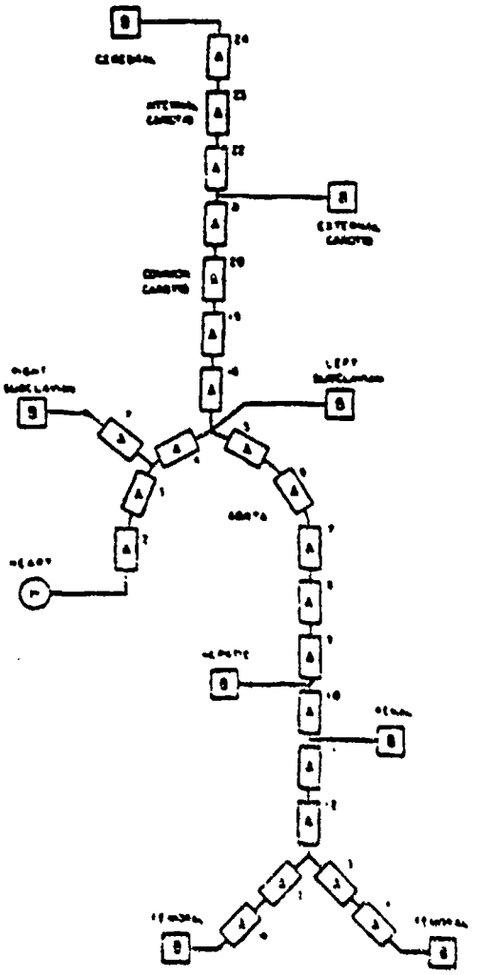


Figure 2. Multi-element Arterial Tree

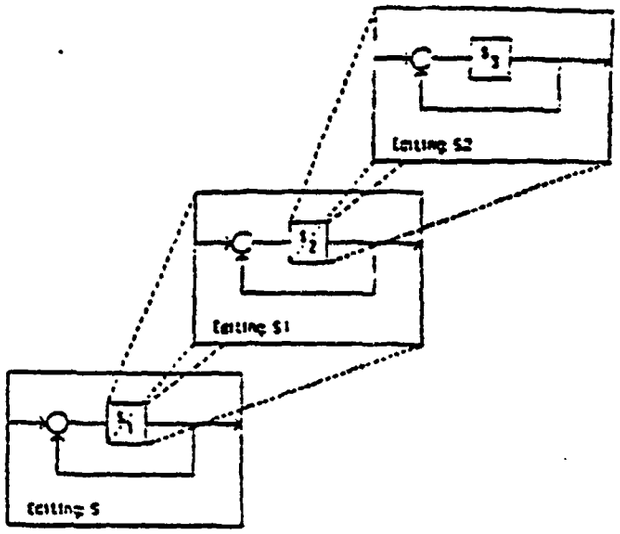


Figure 3. Nesting of Super-Blocks

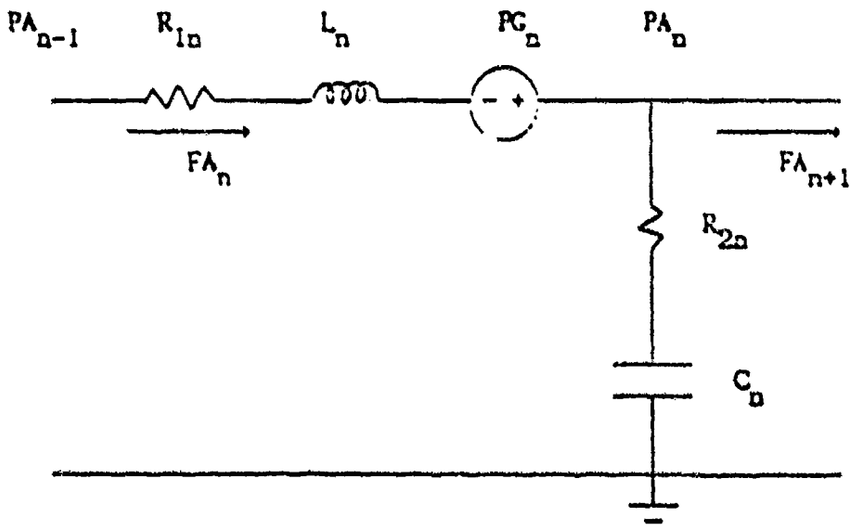


Figure 4 Equivalent Circuit of "A" Element

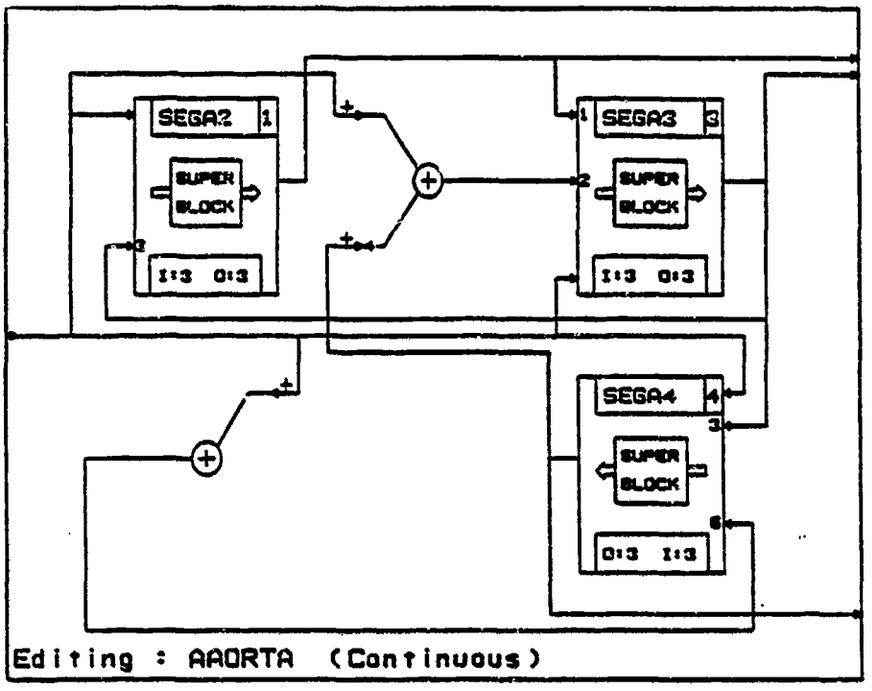


Figure 6. Ascending Aorta

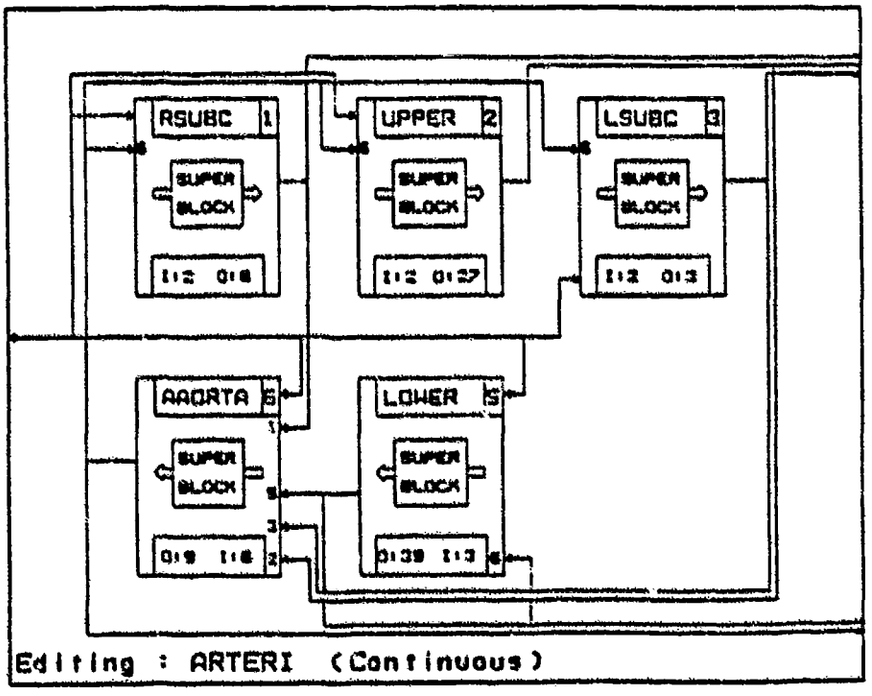


Figure 7. Arterial Tree

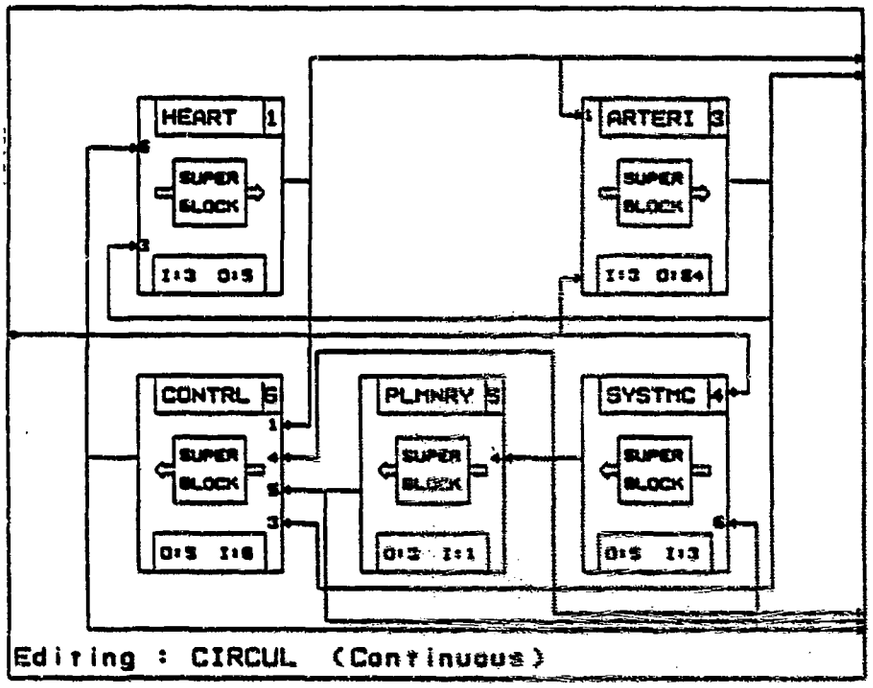


Figure 8. Circulatory System

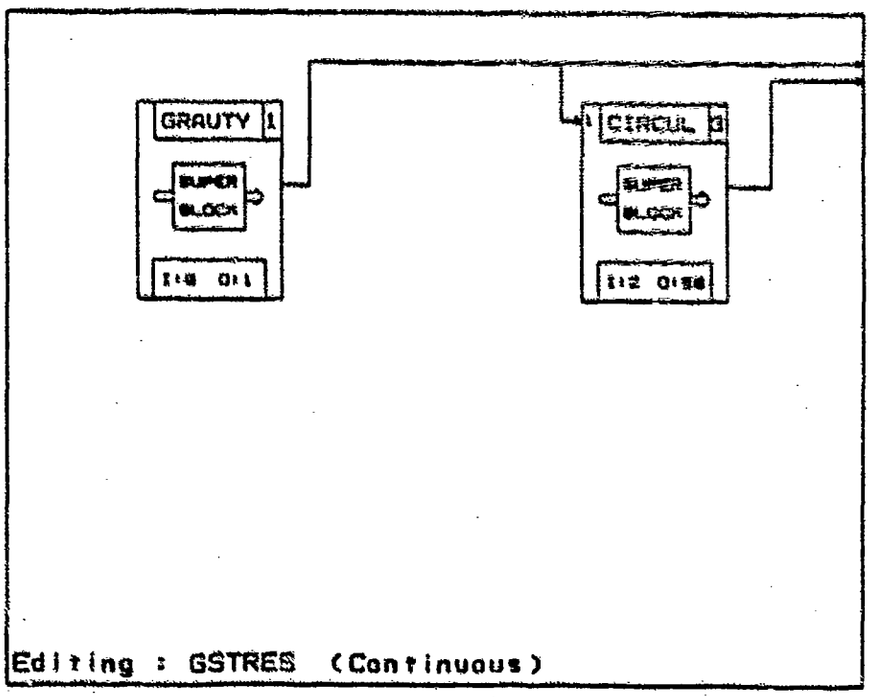


Figure 9. Circulatory System Under +Gz Stress

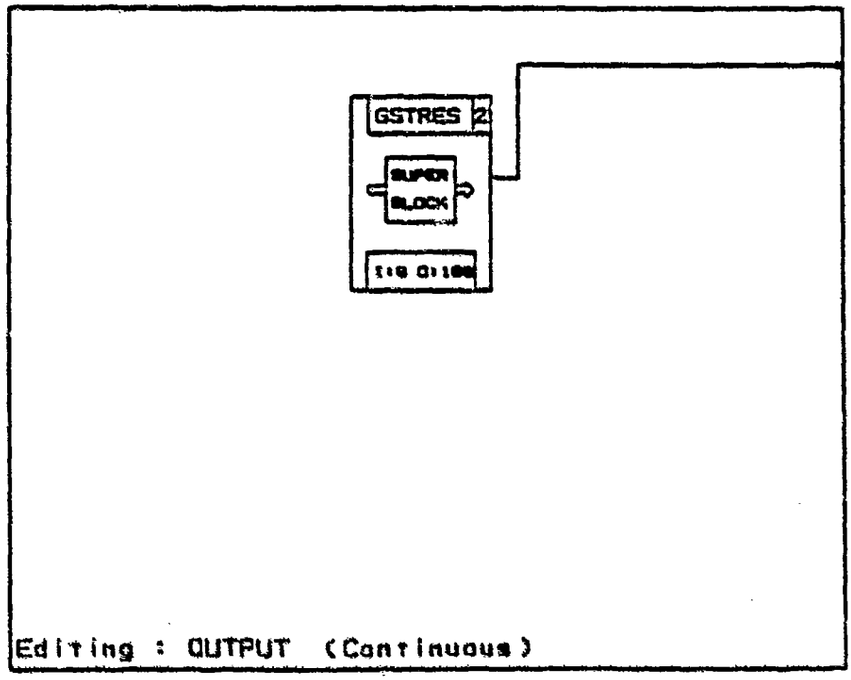


Figure 10 Output of Model

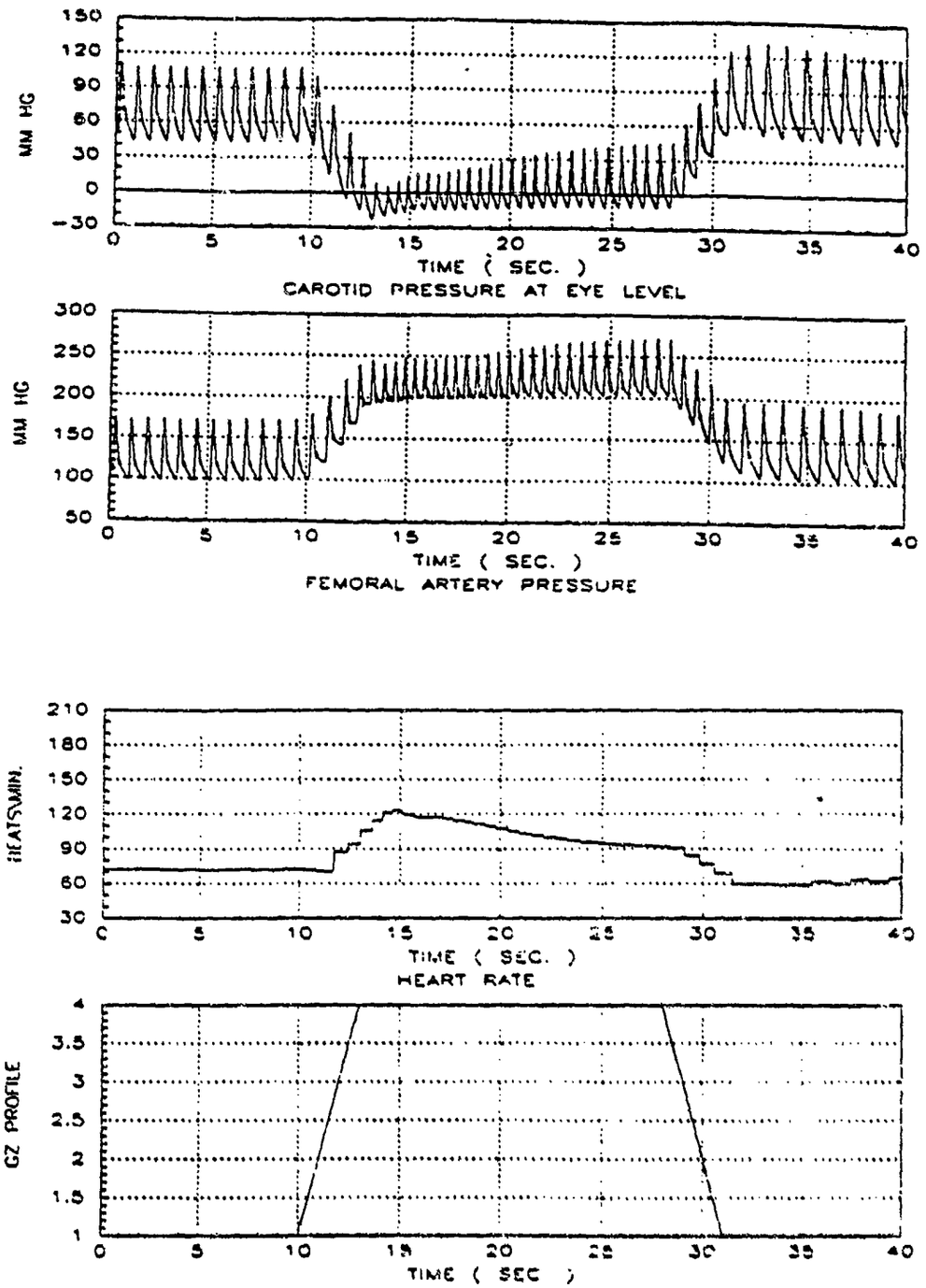


Figure 11. Simulation Results Under A +4 Gz Profile with an Onset Rate of 1G/sec.

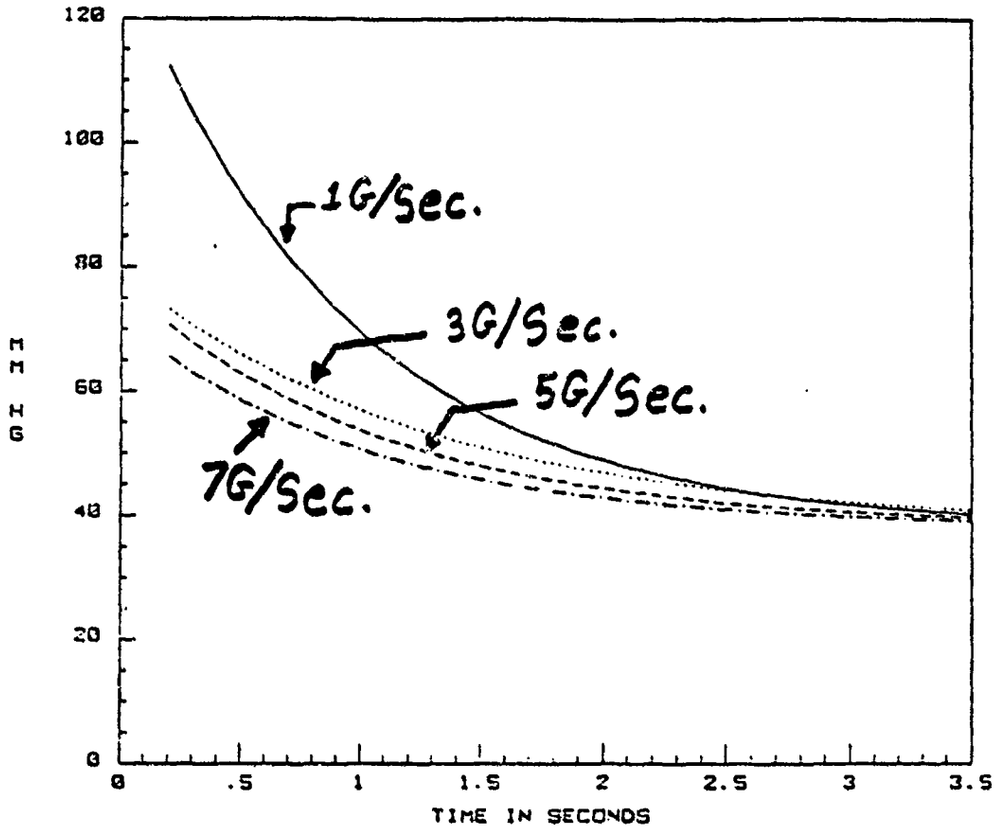


Figure 12. Time Variation of Systolic Pressure at Eye Level Under a +3 Gz Profile with Different Onset Rates

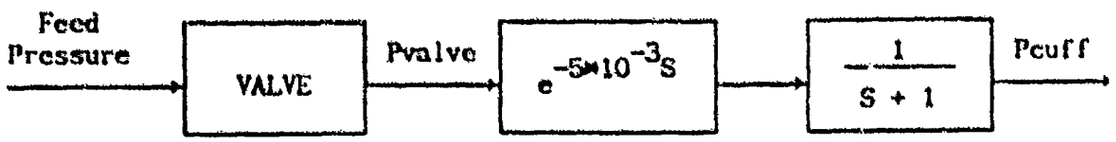


Figure 13. Anti-G Suit Model (Rogers [9])

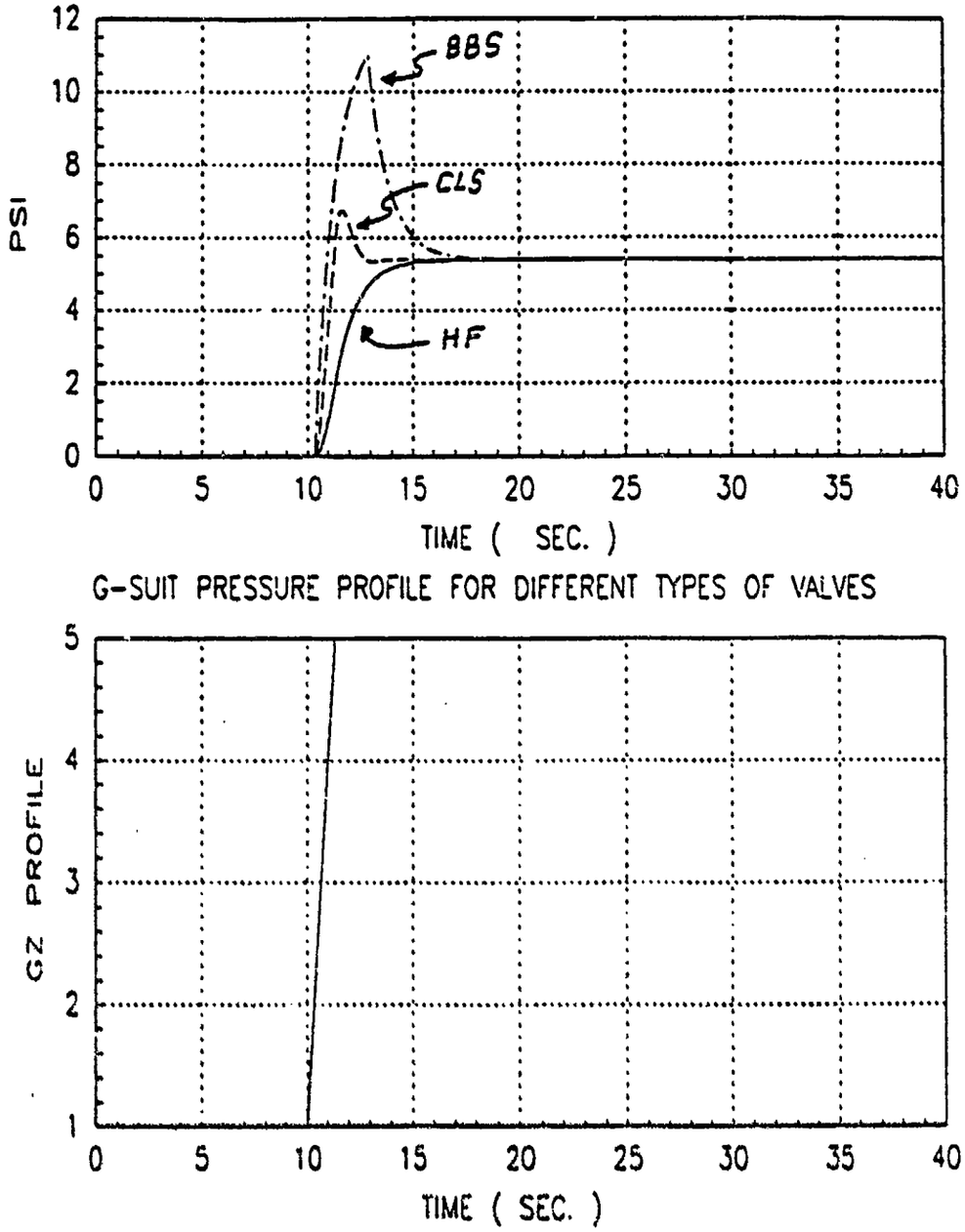


Figure 14. Suit Pressure Profiles for a +5 Gz Stress with an Onset Rate of 3G/sec.

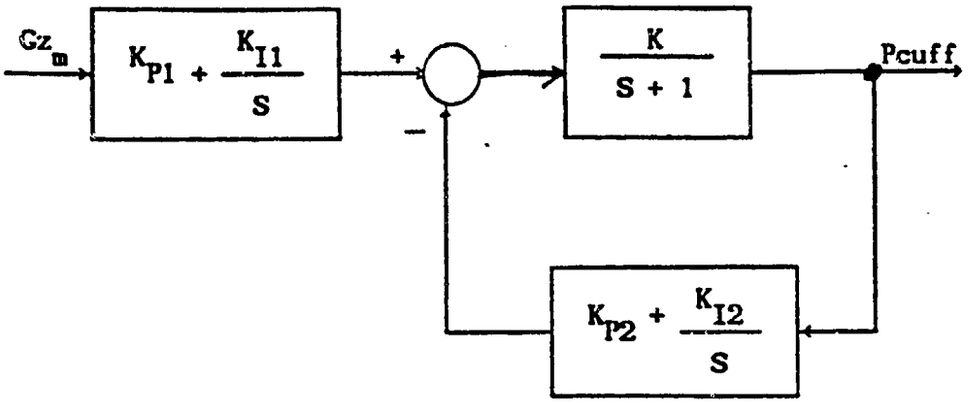


Figure 15. Closed Loop Configuration

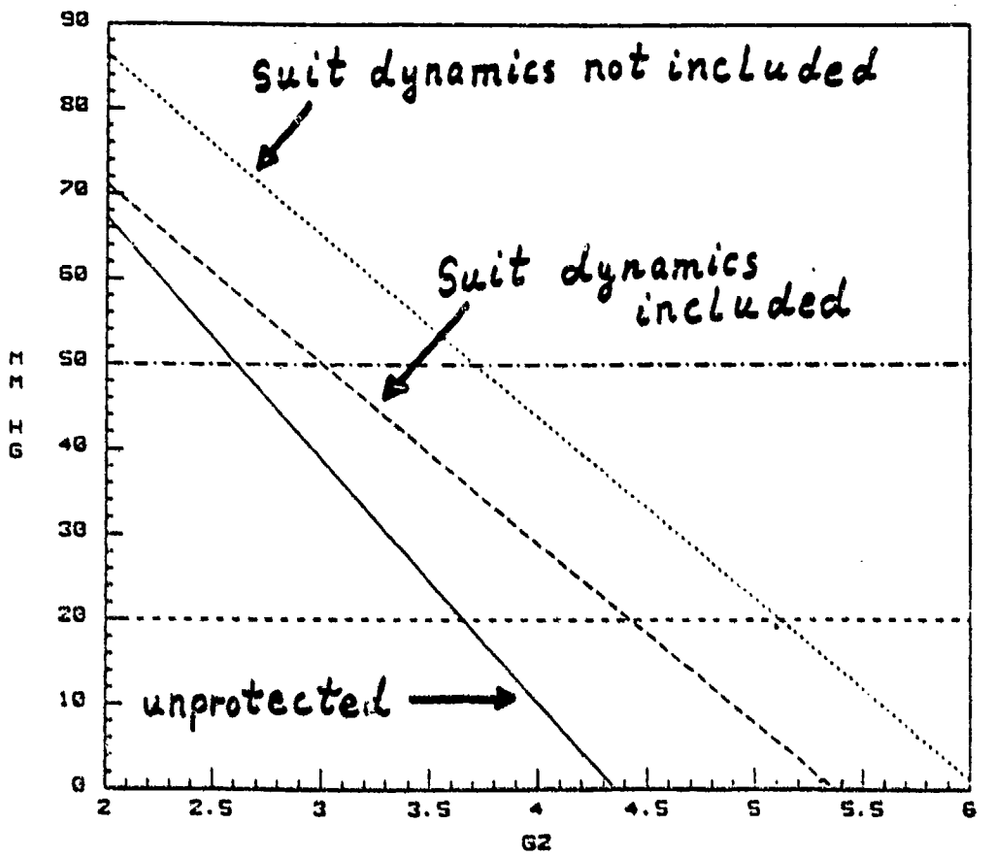


Figure 16. Minimum Systolic Eye Level Carotid Pressure for +Gz Profiles with an Onset Rate of 1G/sec. (Standard Valve)

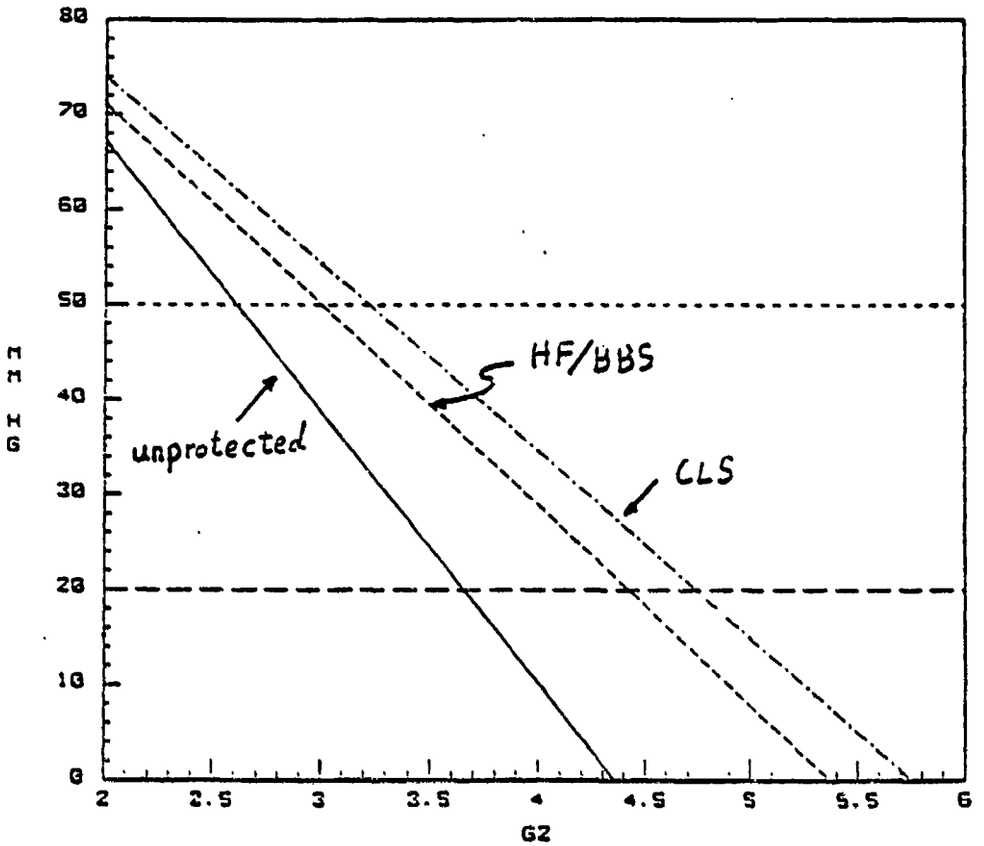


Figure 17. Minimum Systolic Eye Level Carotid Pressure for Different Types of Valves with an Onset Rate of 1G/Sec.

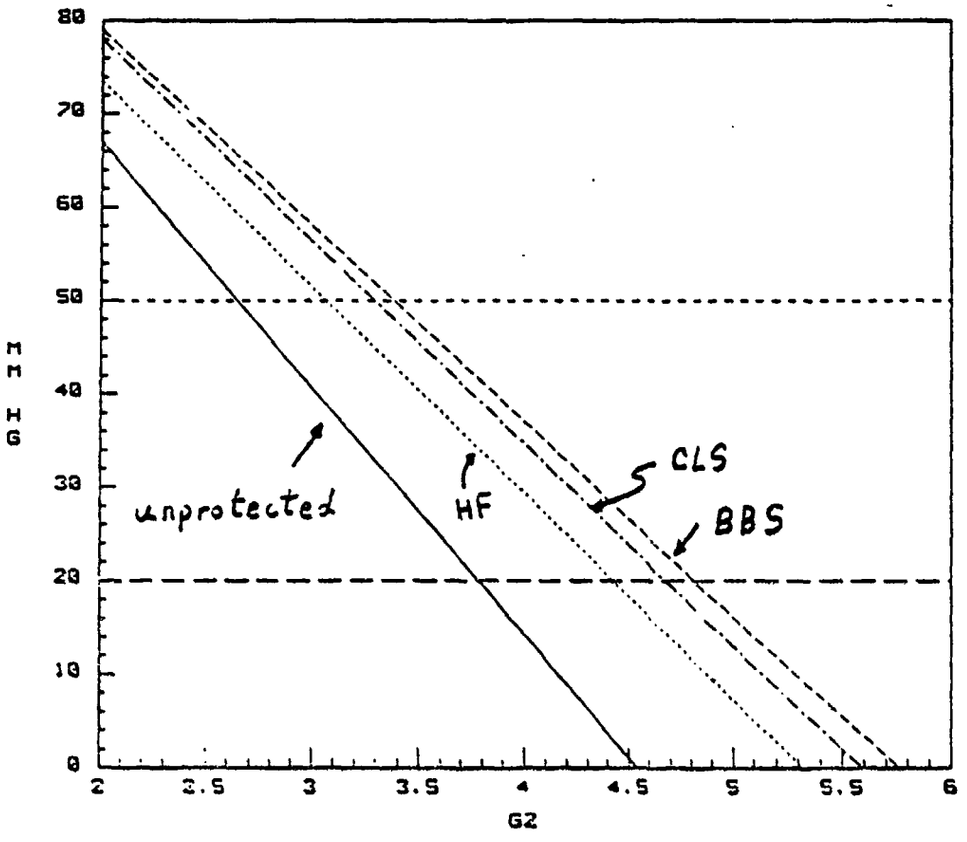


Figure 18. Minimum Systolic Eye Level Carotid Pressure for different Types of Valves with an Onset Rate of 3G/sec.

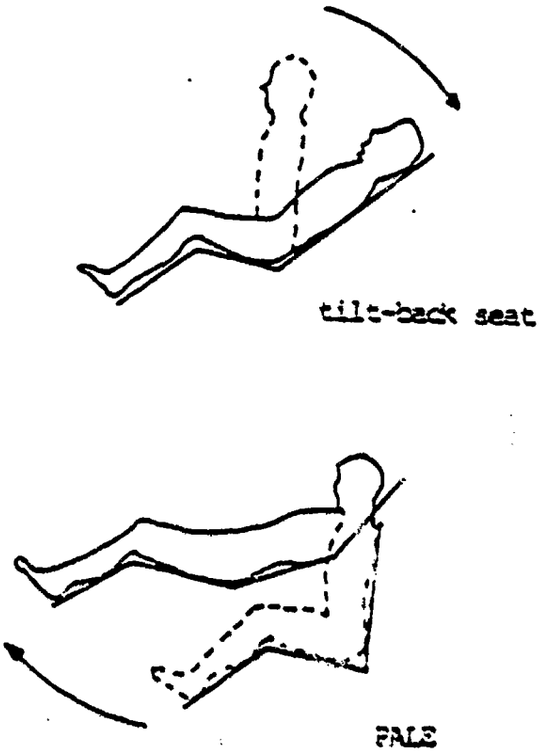


Figure 19. Tilt Back Seat and PALE seat (Chu [2])

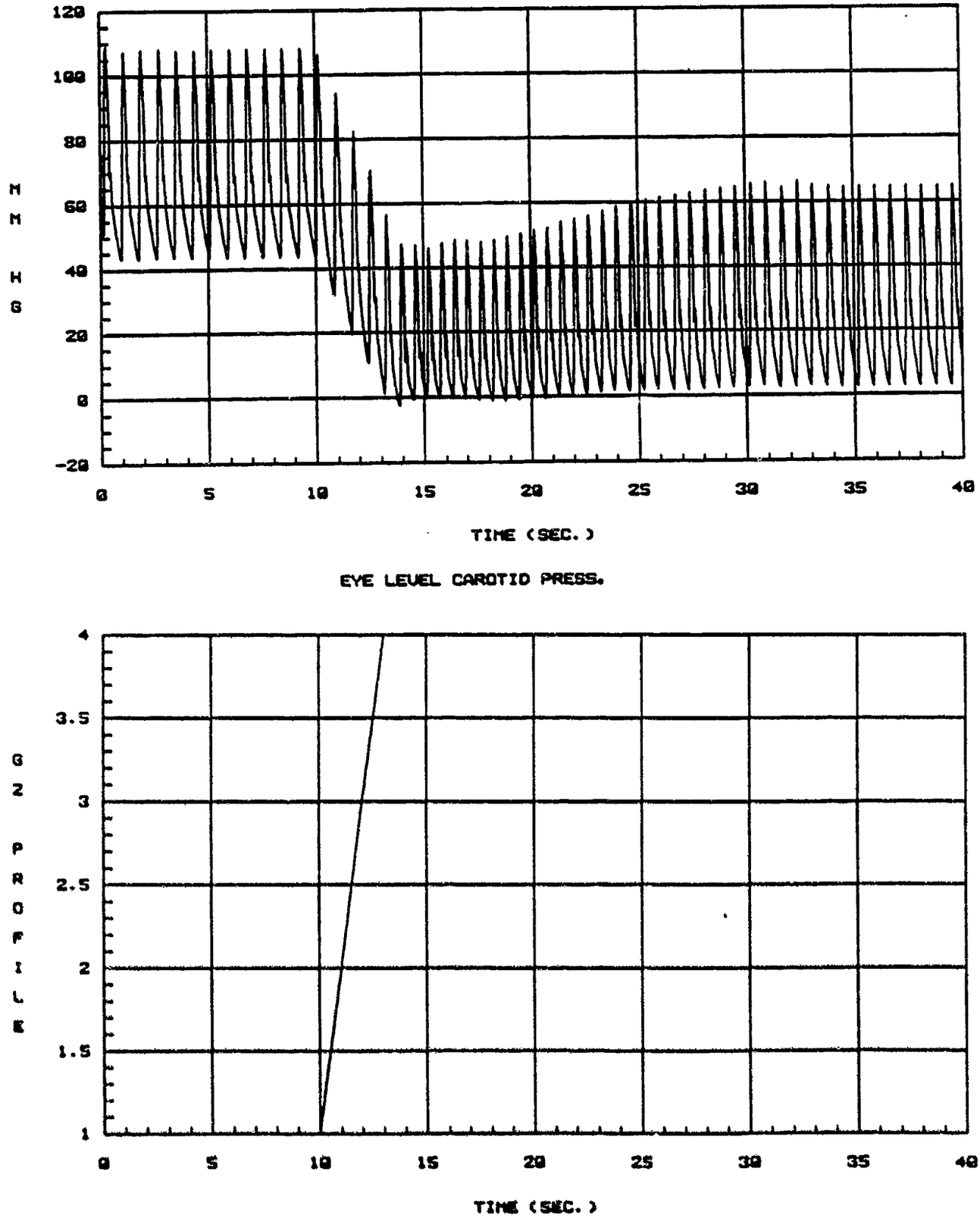


Figure 20. Simulation Result for +3 Gz stress with an Onset Rate of 1G/sec. and Seat Back Angle of 0°

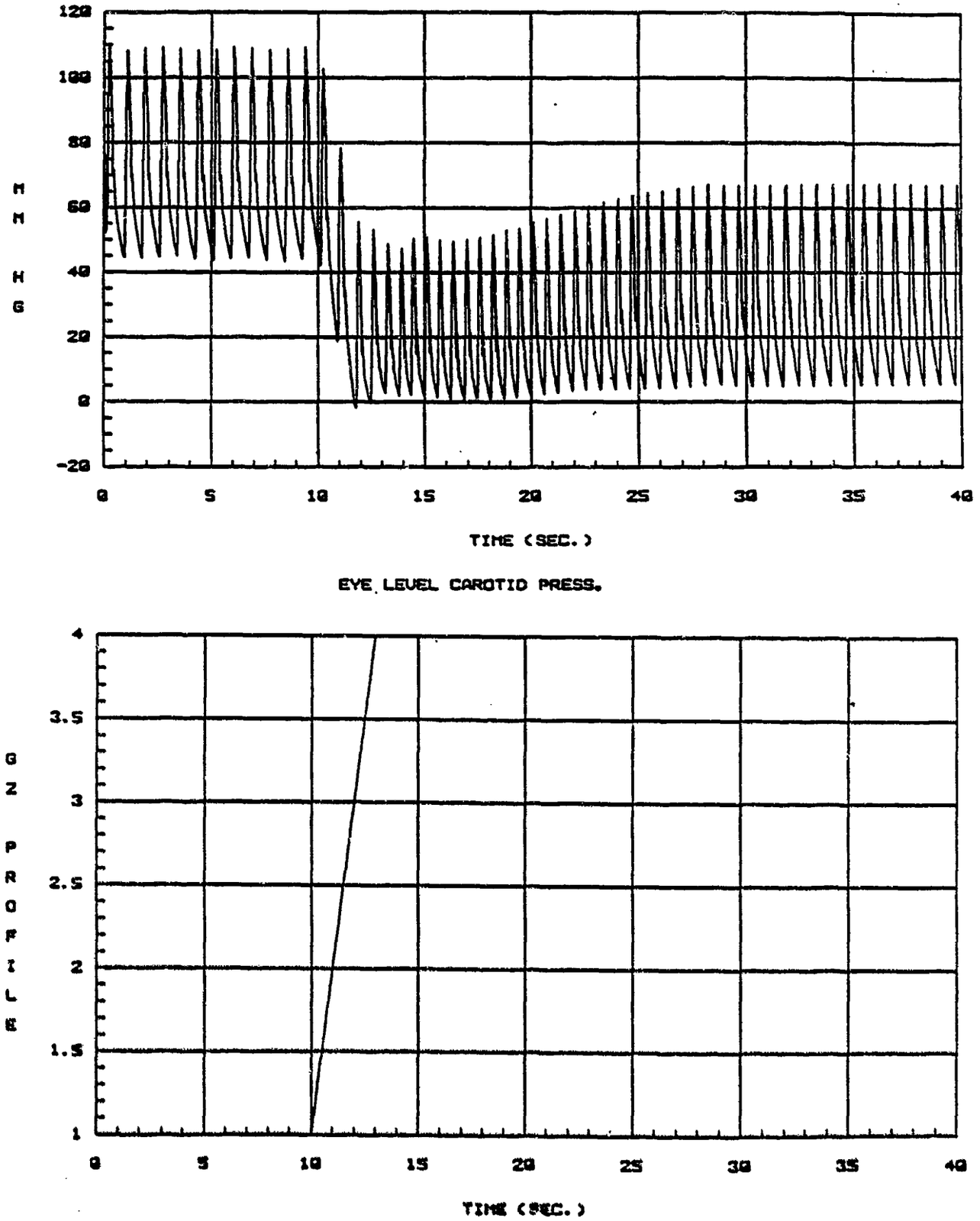


Figure 21. Simulation Result for +3 Gz stress with an Onset Rate of 1G/sec. and Seat Back Angle of 17°

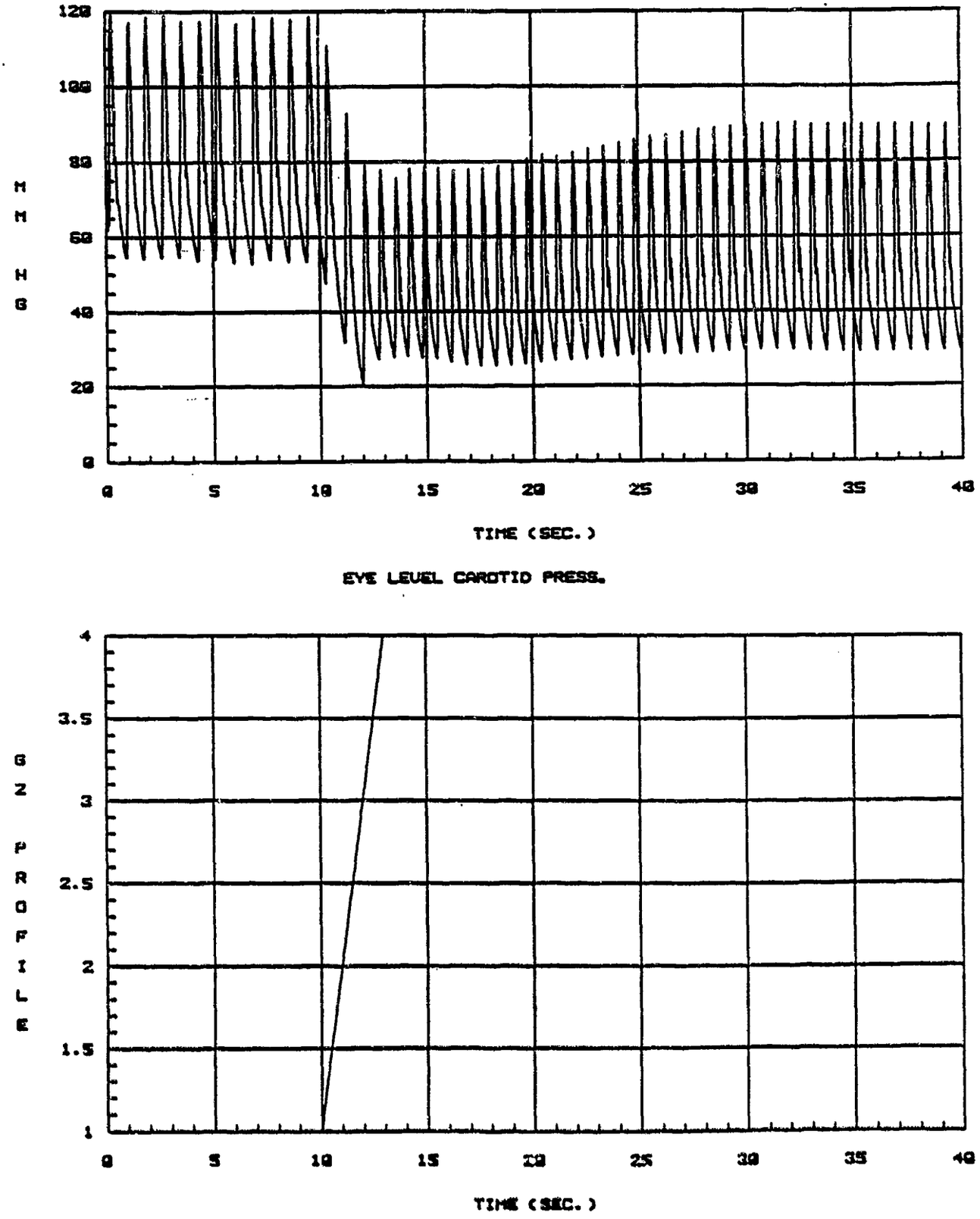


Figure 22. Simulation Result for +3 Gz stress with an Onset Rate of 1G/sec. and Seat Back Angle of 52°

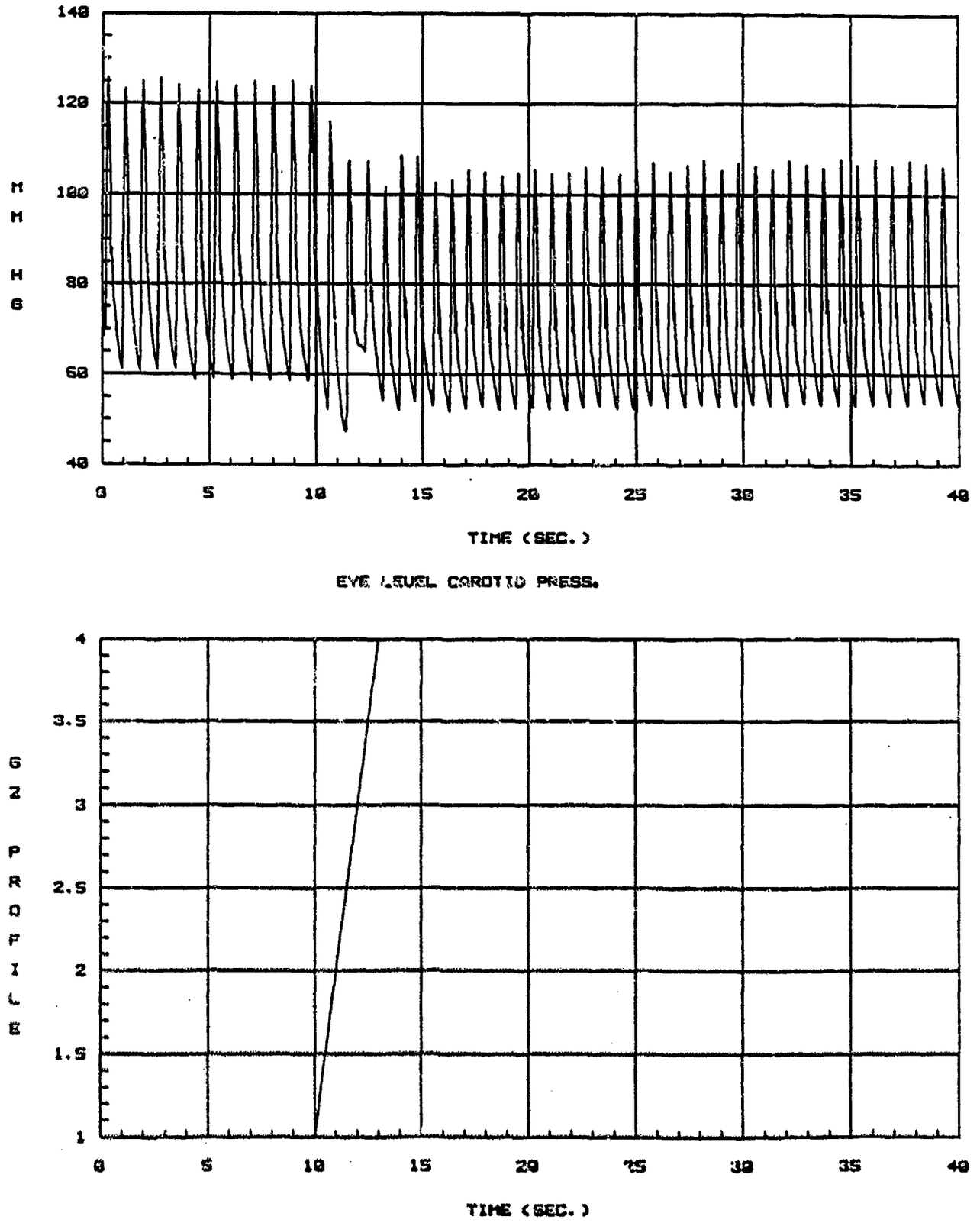


Figure 23. Simulation Result for +3 Gz stress with an Onset Rate of 1G/sec. and Seat Back Angle of 67°

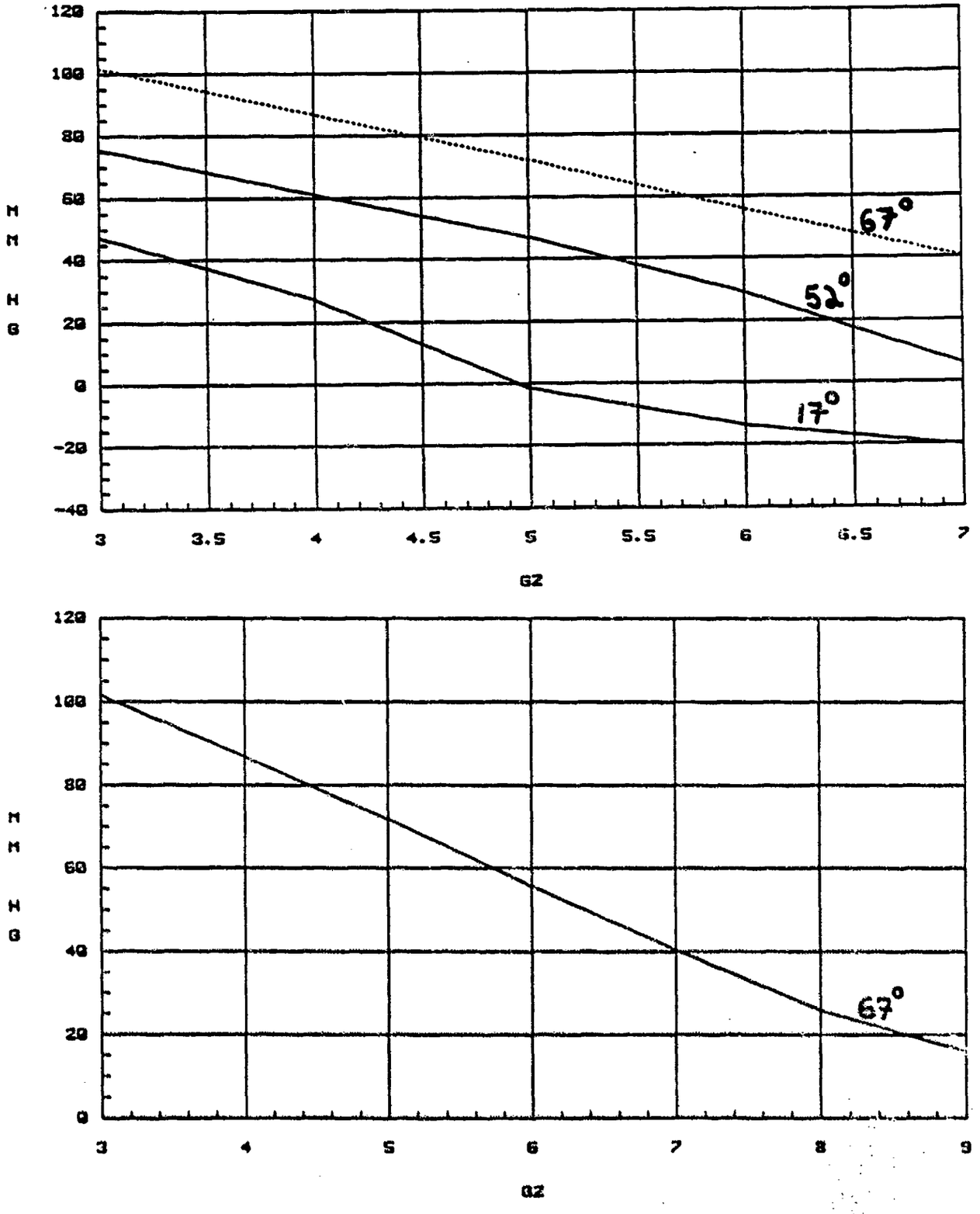


Figure 24. Minimum Systolic Eye Level Carotid Pressure for Different Seat Back Angle (Onset Rate of 1G/sec.)

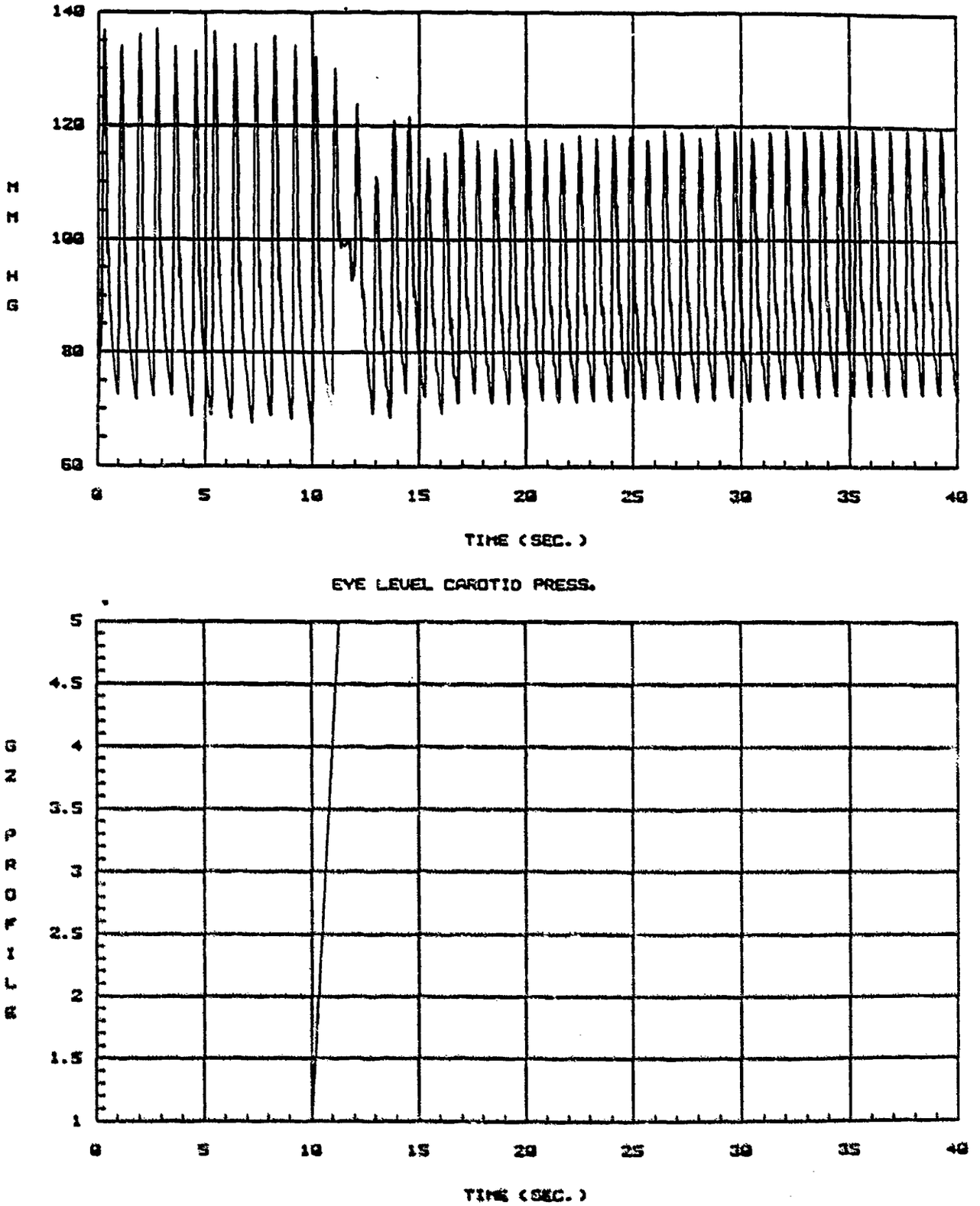


Figure 25. Simulation Results for +5 Gz stress with an Onset Rate of 3G/sec. and Seat Back Angle of 90°

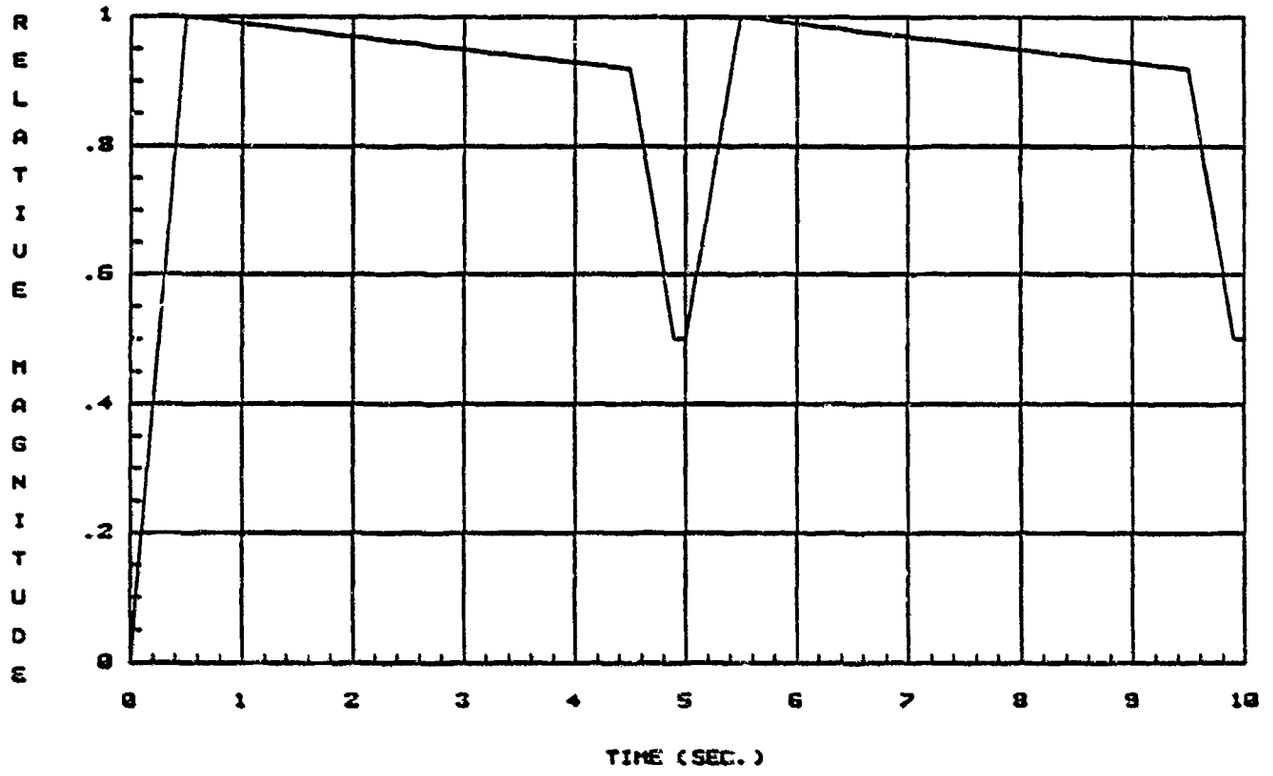


Figure 26. M-1 Maneuver Profile

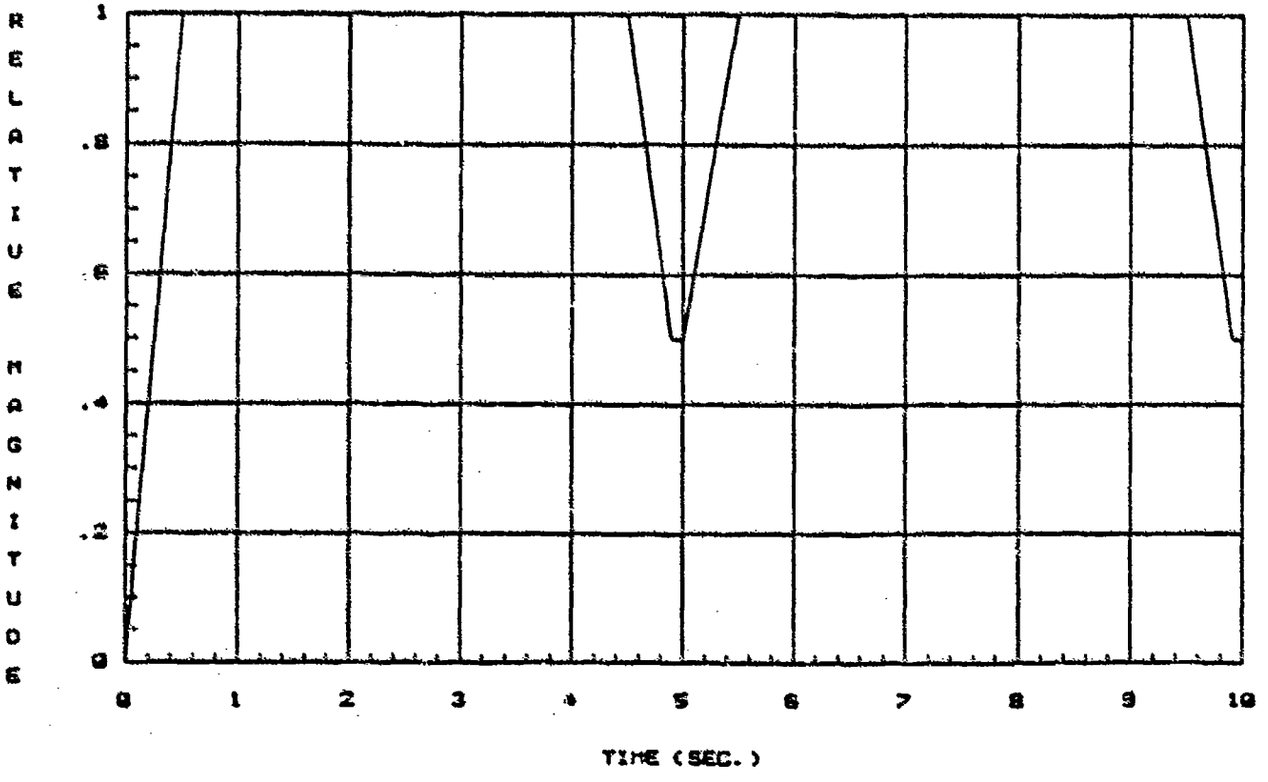


Figure 27. L-1 Maneuver Profile

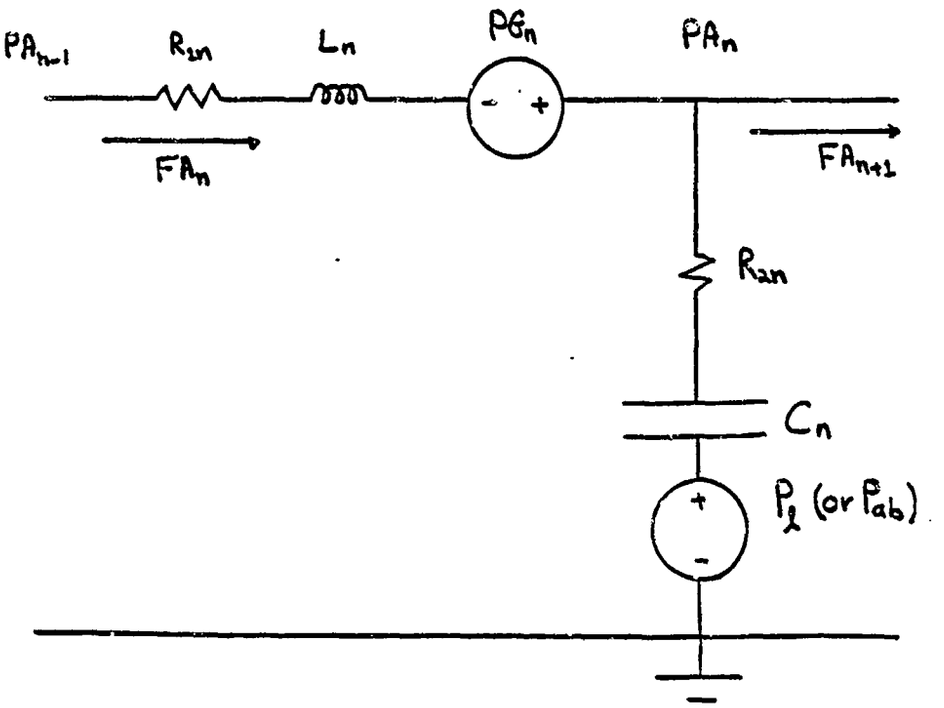


Figure 28a. Modified Equivalent Circuit of "A" Element

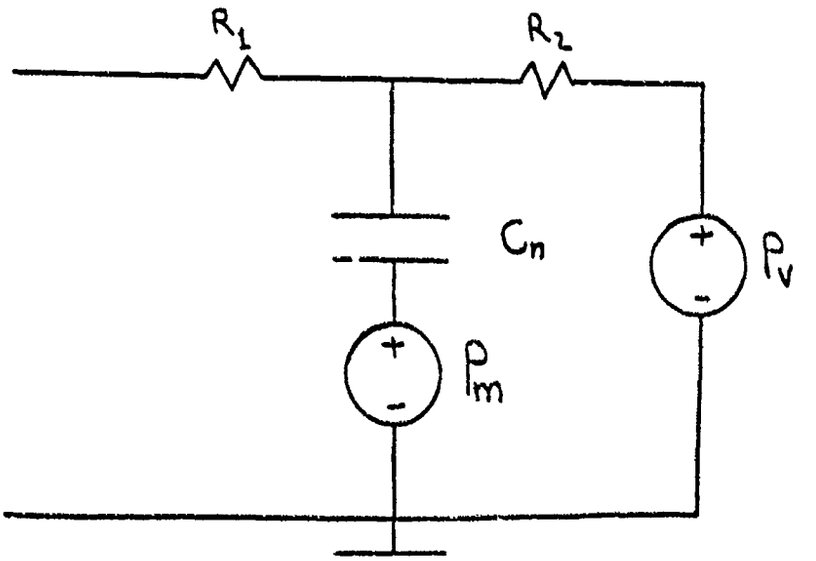


Figure 28b. Modified Equivalent Circuit of "B" Element

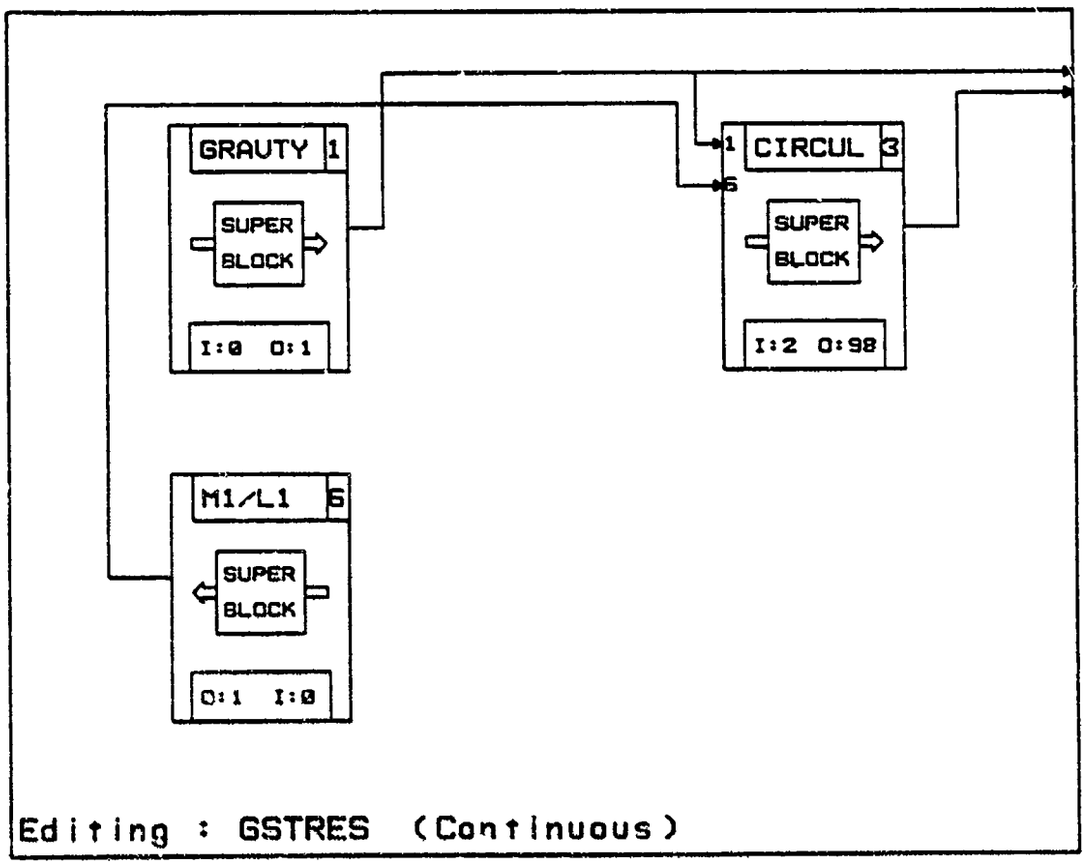


Figure 29. Circulatory System Under +Gz Stress and M-1/L-1 maneuvers

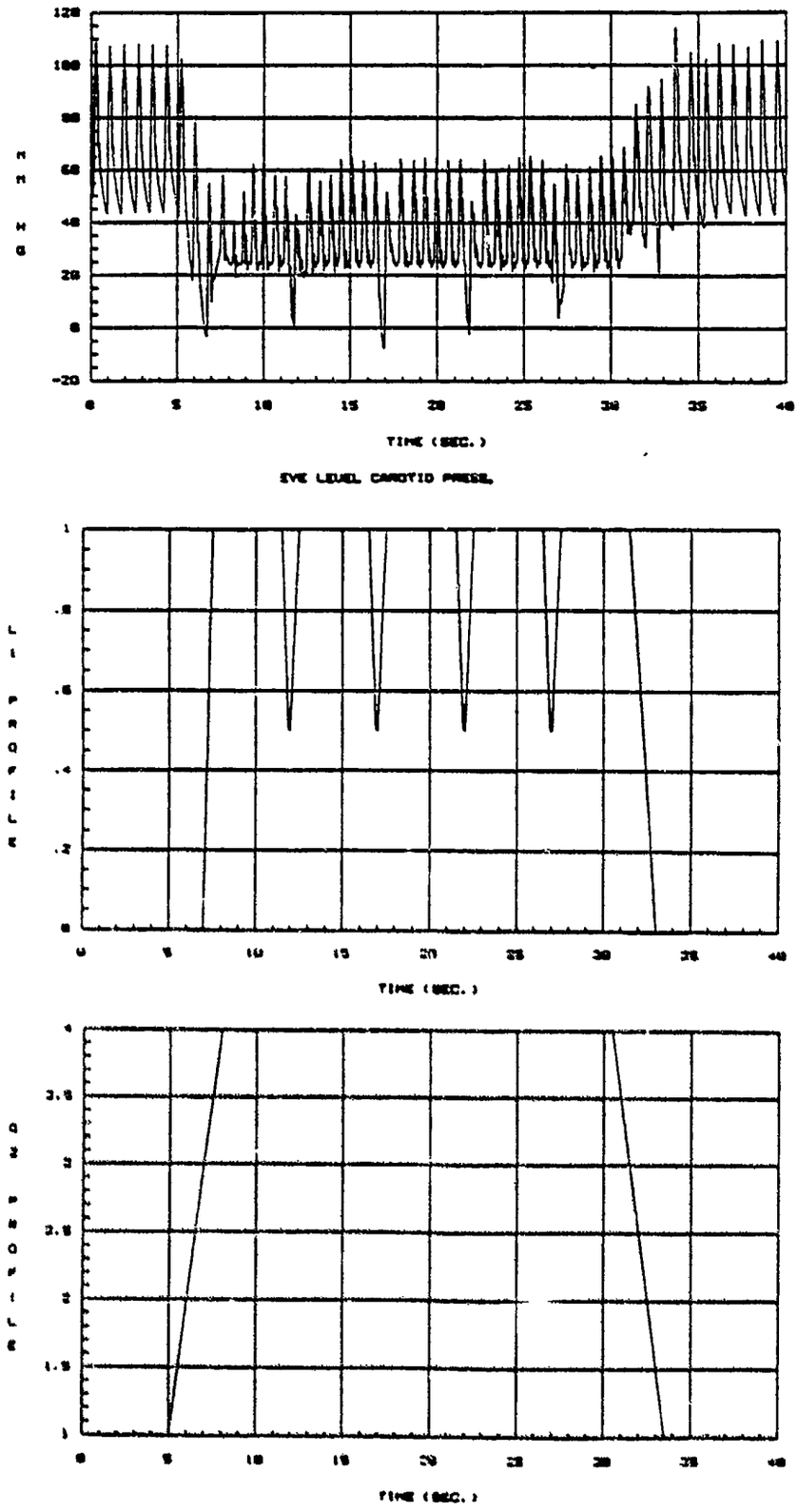


Figure 30. Simulation Results Under +4 Gz Stress and L-1 Maneuver with an Onset Rate of 1G/sec.

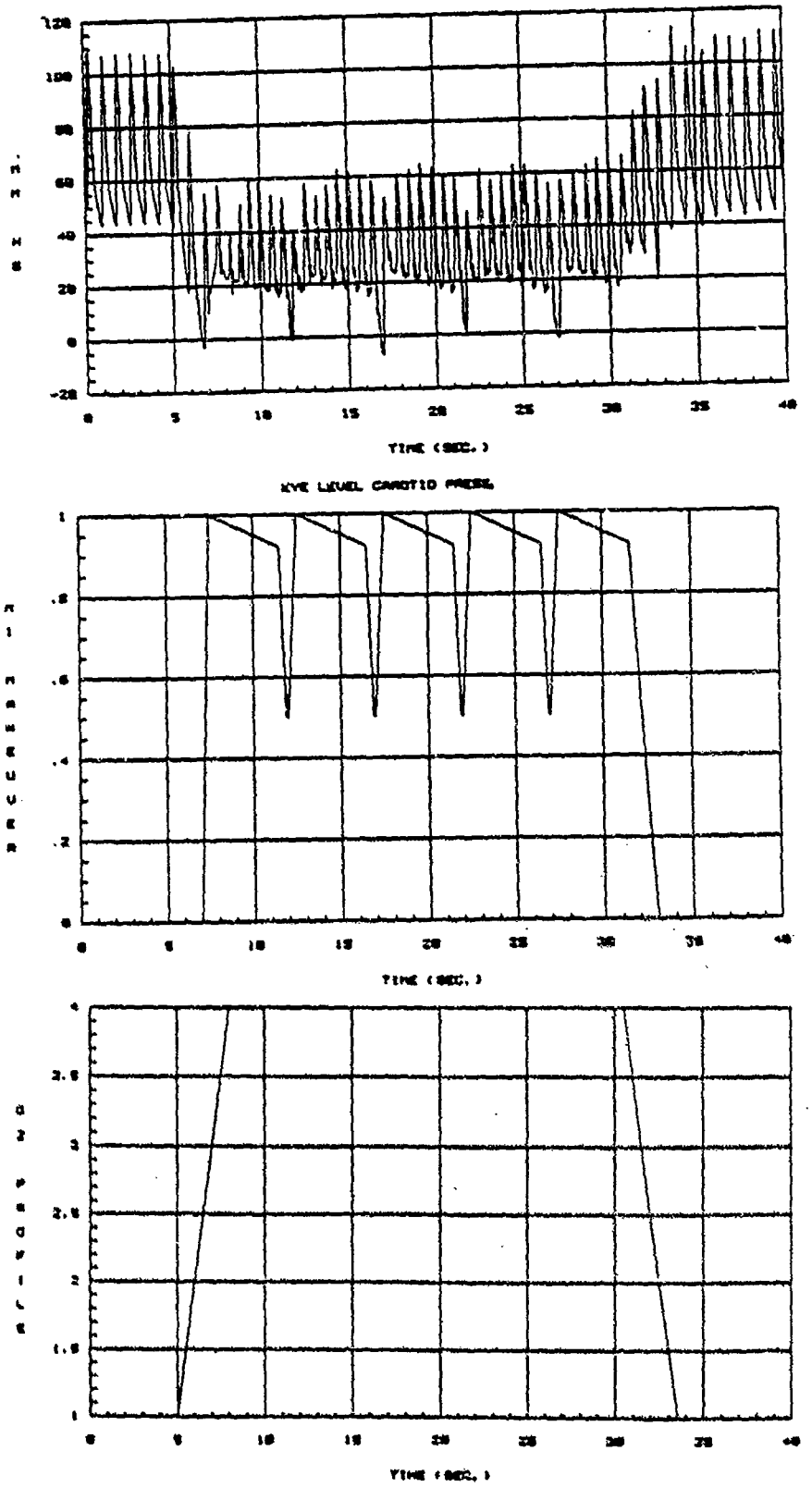


Figure 31. Simulation Results Under +4 Gz Stress and M-1 Maneuver with an Onset Rate of 1G/Sec.

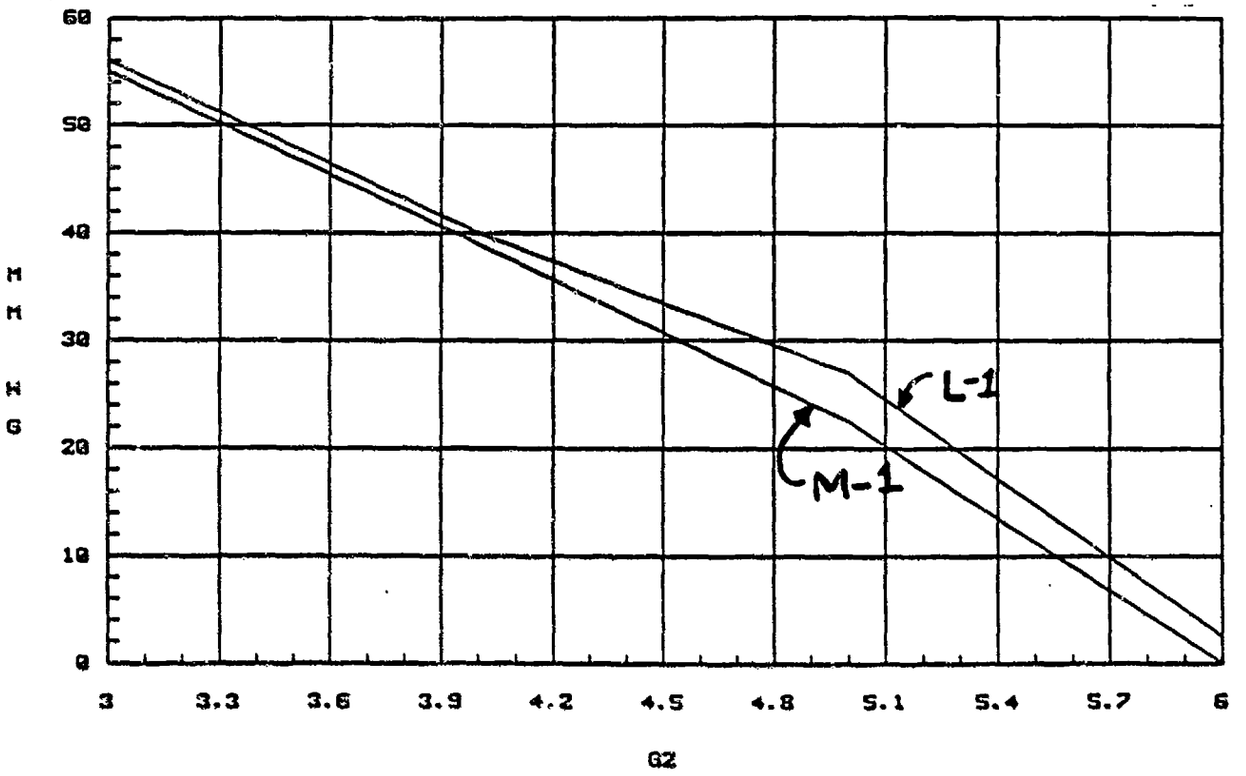


Figure 32. Minimum Systolic Eye Level Carotid Pressure for L-1 and M-1 Straining Maneuvers with an Onset Rate of 1G/Sec.

1986 USAF-UES MINI-GRANT PROGRAM

FINAL TECHNICAL REPORT

Adaptive Grid Generation Techniques for Transonic Projectile Base Flow Problems

Chen-Chi Hsu, Principal Investigator
Christopher Reed, Co-Principal Investigator
Department of Engineering Sciences
University of Florida
Gainesville, Florida 32611

SUMMARY

A two-dimensional adaptive grid generation technique for the numerical simulation of transonic turbulent projectile aerodynamics has been successfully developed. The technique is mainly based on a variational principle which allows for nearly independent control of grid adaptation along individual curvilinear coordinates. Accordingly, an adaptive grid generation code has been developed and coupled to an axisymmetric thin-layer Navier-Stokes code. Numerical experiments conducted for transonic turbulent flows past a projectile model with sting (no base flow) has resulted in a paper (Attachment I) to be presented at the Eighth International Conference on Computing Methods in Applied Sciences and Engineering to be held in Versailles, France, December 14-18, 1987. It was found that the self-adaptive gridding at every time step cannot give steady state solution in the shock-boundary layer interaction regions. This fact could be attributed to the sensitivity of the Beam and Warming algorithm to grid resolution. Hence, the thin-layer Navier-Stokes code has been modified to provide an option of using robust TVD scheme for the numerical simulation. Numerical experiments conducted for assessing the Beam and Warming algorithm, a TVD scheme and a diagonalized TVD scheme developed has resulted in a paper (Attachment II) to be presented at AIAA 26th Aerospace Sciences Meeting to be held on January 11-14, 1988. As reported in the Attachment III, a paper to be given at

AIAA 26th Aerospace Sciences Meeting, the coupling of the adaptive gridding to the TVD scheme has been tested successfully on transonic flows past the projectile model with sting.

The self-adaptive computational method developed has also been tested successfully for projectile base flow problems. Preliminary results obtained for transonic turbulent flow of Mach number 0.96 past a projectile have been reported in Co-Principal Investigator C. W. Reed's Ph.D. dissertation. A copy of the dissertation (107 pages) has been sent to Dr. Lijewski of Eglin Air Force Base. Further assessment of the computational method is being conducted for transonic turbulent flows past a real projectile at zero angle of attack. The results obtained to date show that the self-adaptive computational method is indeed very accurate and robust. The surface pressure computed is in excellent agreement with measured data; unfortunately, there is no base pressure measurement available for assessing the computed base pressure distribution. It is expected that a technical paper will be resulted from the current investigation.

Attachment I

A paper to be presented at the Eighth International Conference on Computing Methods in Applied Sciences and Engineering to be held in Versailles, France, December 14-18, 1987.

An Adaptive Grid Generation Technique for Viscous Transonic Flow Problems

Christopher W. Reed and Chen-Chi Hsu
Department of Engineering Sciences
University of Florida, Gainesville, Florida

ABSTRACT

An adaptive grid generation procedure is developed for viscous flow problems. The equations governing the adaptation are based on a variational statement resulting in a set of elliptic governing equations in which adaptation can occur independently in each coordinate direction. The method allows for explicit control of adaptation and orthogonality while grid smoothness is implicit in the elliptic equations. The adaptive grid generation equations provide a predictable and reliable response to the control functions and are capable of providing the extremely refined mesh in the boundary layer regions. The grid generation equations are coupled with a thin layer Navier-Stokes code to solve a transonic axisymmetric projectile problem. Results obtained for Mach number 0.96 indicate that the adaptive grid can provide a good grid network for viscous transonic flow problems, however it was found that the grid motion and interpolation used in the technique could have an effect on the solution.

INTRODUCTION

The solution adaptive grid generation technique has become an important area in computational fluid dynamics since it has been shown to provide good grid networks for the complex flow fields occurring in transonic and supersonic flows [1,2,3]. The use of boundary fitted curvilinear coordinate systems with transformed governing equations leads naturally to the concept of solution adapted grids. As practical limits are placed on the grid resolution by constraints of computer storage and CPU time, the coordinate spacing in the physical domain is varied to increase resolution in only the large gradient regions. For simple flow problems when the position of important gradients is known, good adapted grids can be obtained with conventional techniques. However, in more complex problems the position and orientation of these important regions are not known a priori and then the development of a good adapted grid is difficult. A solution adaptive grid generation addresses this problem by continuously updating the grid during the solution process such that the important physical gradients are sufficiently resolved as they develop. The purpose of adaptive grid generation, thus, is to increase solution accuracy by reducing the truncation error due to finite difference approximations of the transformed governing equations. Analysis of the truncation error terms [4] as well as experience has shown that enhancement of two other grid characteristics, smoothness and orthogonality will also

reduce the truncation error. Thus good adaptive grid generation should include the optimization of these characteristics as well.

The general approach of most adaptive grid generation schemes is based on minimization techniques. A measure of each desired grid characteristic is defined, and the grid is obtained by minimizing the integral of these measures over the domain. In many instances adaptation is important in only one coordinate direction. Dwyer et al. [5], for example has used an 'equidistribution law' to control spacing in one coordinate direction of a two dimensional combustion problem. One dimensional adaptation can be extended to higher dimensions by successive adaptation in each coordinate direction. In one such approach by Nakahashi and Deiwert [2], a spring analogy is used to include orthogonality. Tension springs are assumed to connect each point along a coordinate and torsional springs are assumed to connect intersecting coordinates. The tension spring constants are synonymous with the control function and the torsional spring constants are determined to prevent skewness. Nakahashi and Deiwert solved a transonic viscous airfoil problem in which the grid was adapted to the density gradient in the streamwise direction to resolve shocks and was adapted to the velocity gradient normal to the airfoil surface to resolve the boundary layer. The one dimensional approach to multidimensional grids has the advantage of efficiency and the independence of control functions in each direction. However, this approach can

lead to problems in maintaining smoothness [6].

Dulikravich and Kennon [7] have used the spring analogy also, but extend it to a multidimensional approach by removing the constraint that grid points move along coordinate lines. This approach can eliminate problems with smoothness but the grid must then be obtained using iterative solution algorithms, which can be time consuming if the initial guess for the grid is far from the solution. To increase efficiency they have used an optimization procedure to solve for the grid. However, since the method is based on the discrete nature of the grid the spring analogy equations are not elliptic and the resulting grid is not guaranteed to maintain a one to one mapping between coordinate systems. Brackbill and Saltzman [1] have developed an adaptive grid generation scheme based on a variational approach. A functional is defined to measure each grid characteristic and the minimization of these functionals results in a set of partial differential equations that govern the grid spacing. With the proper choice of parameters the equations are elliptic which helps to maintain a one to one mapping and results in a smooth grid. They solved an inviscid supersonic flow problem by adapting the grid cell size to a function of the pressure gradient to capture the shocks. As the solution of elliptic equations requires iterative algorithms, this approach appears costly. However, since the grid is updated continuously, only a few, and possibly just one, iteration is required to solve the

equations each time the grid is updated.

The adaptive grid generation scheme presented here is similar in approach to that of Brackbill and Saltzman but is developed for applications to viscous transonic flow problems. The solution of these problems usually contain shock patterns aligned with one coordinate direction and boundary and shear layers parallel to a streamwise coordinate. Also the grid spacing can vary by orders of magnitude along different coordinate directions and it is therefore necessary to use different control functions in each direction. An independent functional is defined to control adaptation in each coordinate direction such that the resulting equations are elliptic. These equations are used to adapt the grid for an axisymmetric transonic projectile flow problem which is solved using an implicit factorized algorithm for the thin layer Navier-Stokes equations.

ADAPTIVE GRID GENERATION

The generation of a grid network can be viewed as the development of a boundary fitted curvilinear coordinate system in which the grid points are defined by the coordinate intersections. In two dimensions the mapping between the curvilinear and Cartesian coordinates is expressed as

$$\xi = \xi\ (x,y)$$
$$\eta = \eta\ (x,y) \qquad (1)$$

and the problem of grid generation becomes one of defining this mapping. To obtain an adaptive grid generation method using variational techniques, a measure of each grid property is defined and its integral is taken over the physical domain. This procedure results in the total functional I_T

$$I_T = \int \frac{\nabla\xi \cdot \nabla\xi}{P}\, dxdy + \int \frac{\nabla\eta \cdot \nabla\eta}{Q}\, dxdy + \lambda \int (\nabla\xi \cdot \nabla\eta)^2\, dxdy \qquad (2)$$

The curvilinear coordinates that minimize this functional represent the grid with the desired properties. The first integrand in Eq. 2 is a measure of adaptation of the grid spacing in the ξ direction to the control function P. When P is small the quantity $\Delta\xi$ must also be small in order to minimize the integral; a small value of $\Delta\xi$ corresponds to large grid spacing. Consequently, when P is large, the spacing will be small. The second functional represents adaptation in the η coordinate direction to the control function Q. The third functional is a measured of orthogonality defined such that an orthogonal grid will minimize the integral. The parameter λ weighs the relative importance of orthogonality to adaptation. A value of 0.5 was used for λ in all calculations. An equivalent set of partial differential equations can be obtained by applying the Euler-Lagrange equations to I_T. These equations are

$$\xi_{xx} + \xi_{yy} + \frac{\lambda}{J^{-1}} (\eta_x{}^2\xi_{xx} + 2\eta_x\eta_y\xi_{xy} + \eta_y{}^2\xi_{yy})$$

$$+ \frac{\lambda}{J^{-1}} ((2\xi_x\eta_x+\xi_y\eta_y)\eta_{xx} + (\xi_x\eta_y+\xi_y\eta_x)\eta_{xy} + (\xi_x\eta_x+2\xi_y\eta_y)\eta_{yy}) = \frac{\nabla\xi\cdot\nabla P}{P} \qquad (3)$$

$$\eta_{xx} + \eta_{yy} + \frac{\lambda}{J^{-1}} (\xi_x{}^2\eta_{xx} + 2\xi_x\xi_y\eta_{xy} + \xi_y{}^2\eta_{yy})$$

$$+ \frac{\lambda}{J^{-1}} ((2\xi_x\eta_x+\xi_y\eta_y)\xi_{xx} + (\xi_x\eta_y+\xi_y\eta_x)\eta_{xy} + (\xi_x\eta_x+2\xi_y\eta_y)\xi_{yy}) = \frac{\nabla\xi\cdot\nabla Q}{Q}$$

where $$J^{-1} = \xi_x\eta_y - \xi_y\eta_x \qquad (4)$$

They will remain elliptic as long as the terms arising from the orthogonality functional are not too large. To obtain a numerical grid in the physical domain the equations are inverted to make x and y the dependent variables. The complete equations are available in reference [8]. Currently, the equations are solved numerically using a Newton Raphson point iterative procedure. The point iterative scheme was chosen specifically over ADI methods since it will be more efficient when only one or two iterations are required for convergence. This will be the case when the grid is adapted continuously during the solution procedure of the compressible flow equations.

As the grid points interior to the domain move to satisfy Eq. (3), it is necessary to move the points along the boundary in

a consistent manner. For this purpose a one dimensional equation analogous to the two dimensional adaption can be derived in a similar way. The resulting equation is

$$s_{\xi\xi} + s_{\xi} P_{\xi}/P = 0 \quad \text{or} \quad s_{\eta\eta} + s_{\eta} Q_{\eta}/Q = 0 \tag{5}$$

where s is the arclength along a boundary coordinate. The analytic solution is Ps_{ξ}=const. which is an 'equidistribution law'. Here, however, Eq. (5) is solved iteratively with Eq. (3) so that the boundary points will remain consistent with the interior points during the solution procedure.

As pointed out earlier, the grids used for viscous transonic flow problems contain highly refined grid spacing in the direction normal to the surface in order to resolve the viscous sublayer. This spacing results in grid cells with large aspect ratios, up to 10^5 near solid boundaries. The solution to elliptic equations becomes inefficient for such large aspect ratios since the motion in one direction will be severely limited by the extremely small spacing in the other. To improve the efficiency of the above method a temporary 'reduced' grid is formed by removing many of the points in the finely clustered boundary layer regions (e.g., 18 points of the 40 used in the normal direction are removed in this study), leaving only enough points to define the coordinate and produce a grid with spacing of equal orders of magnitude in both directions. Equations (3)

and (5) are solved for the reduced grid and then the points previously removed are reinserted along each normal coordinate using the same one dimensional adaptation equation used along the boundaries. It is necessary, however, to modify the control function in the region where points were removed since just a few points will represent the spacing required by those removed. Let Δs_i be the spacing between the i+1 and ith grid points along a normal coordinate line. By approximating Eq. (5) with second order central differences the following relationship can be obtained

$$(\Delta s)_{i+1} = (\Delta s)_i \, C_i$$

$$C_i = \frac{Q_i - 0.5(Q_{i+1} - Q_{i-1})}{Q_i + 0.5(Q_{i+1} - Q_{i-1})} \tag{6}$$

If the i+1 point is removed, the relationship for the spacing between the remaining adjacent points becomes

$$(\Delta s)_{i+2} = (\Delta s)_i \, C_i^*$$

$$C_i^* = \frac{C_i \, C_{i+1}}{1+C_i} \tag{7}$$

where C_{i+1} is eliminated and C_i replaced with C_i^*. This procedure is repeated for each point removed and then the modified control function Q^* can be decoded from C^* by starting at a boundary and solving backwards for each value of Q^*. By

using this approach the same grid spacing that would have occurred without using the reduced grid scheme is obtained.

CONTROL FUNCTIONS

The control functions should, in general, be chosen so that the resulting grid reduces the truncation error of the transformed governing equations. In practice, the relationship between the grid and the truncation error is not known explicitly and the choice of control functions is guided by intuition and experience. A general form of the control function considered is

$$P_i = 1 + \gamma f_i \tag{8}$$

where γ is a parameter and f is some derivative of a flow variable scaled to range between zero and one. By evaluating the constant in the equidistribution law and using Eq. (8) for the control function a general expression for the spacing Δs along a coordinate line can be obtained in discrete form

$$(\Delta s)_i = \frac{S[1/(1 + \gamma f_i)]}{\Sigma[1/(1 + \gamma f_j)]} \tag{9}$$

where S is the total arclength along the coordinate. By writing Eq. (9) for the maximum and minimum spacing and forming their ratio, the expression

$$\frac{(\Delta s)_{max}}{(\Delta s)_{min}} = \gamma + 1 \tag{10}$$

is obtained which shows that γ is related to the ratio of the maximum and minimum spacing. Following the work of Nakahashi and Deiwert [2] we have introduced another parameter α to give more control over the grid point spacing. The general control function considered is

$$P_i = 1 + \gamma f_i^{\alpha} \tag{11}$$

The parameter α can be determined automatically by prescribing the minimum grid point spacing along a coordinate line. An implicit equation for α results in

$$(\Delta s)_{min} \Sigma\left(\frac{1}{1 + \gamma f_i^{\alpha}}\right) - \frac{S}{1 + \gamma} = 0 \tag{12}$$

which can be solved for α using a root finding technique. Thus by prescribing the parameter γ and the minimum spacing, control over the minimum and maximum grid point spacing is obtained through the control function. It should be noted that in practice the prescribed values will not be obtained exactly by the adaptive grid method developed here since the equations containing the control functions are two dimensional everywhere except on the boundaries. Also, the influence of orthogonality will alter the spacing in some regions. However, results have

indicated that this approach yields reliable, predictable control over the grid point spacing. The actual choices for f in Eq. (11) are discussed in the results section.

THIN-LAYER NAVIER-STOKES CODE

The governing equations for transonic projectile flow are the compressible Navier-Stokes equations. These equations are solved with a code developed by Neitubicz et al. [9] for efficient calculations of axisymmetric flow problems. The thin layer approximation is employed and turbulence effects are included through the eddy viscosity model of Baldwin and Lomax [10]. the implicit factorized scheme in delta form, developed by Beam and Warming [11], is used to solve the equations. It is second order in space and first order in time. As the scheme is nondissipative, artificial damping is added to make the scheme more stable. A second order dissipation term is added to the implicit side to retain the block tridiagonal matrix and a fourth order dissipation is added to the explicit side.

Steady state solutions are obtained as the time asymptotic solution of the unsteady equations, thus the scheme is time marching. Adaptive gridding can be added to this scheme simply by updating the grid each time step. The general algorithm is to advance the solution to the governing equations one time step, calculate the control functions based on the current solution and

then update the grid network. The solution on the new grid is then interpolated from the solution on the old grid. For two dimensional grids, a linear interpolation based on three points is used. Each cell of the previous grid is divided into two triangles. Once a point on the new grid is located within a triangle of the old grid, the value of the flow variables at the new point are interpolated from the values at the triangle vertices. The use of interpolation has the advantage that the solution can be kept accurate in time when the grid is adapted after a larger number of time steps rather than every time step.

RESULTS

In order to investigate the proposed adaptive grid technique it has been used to solve a transonic projectile flow problem at zero angle of attack. The projectile is a 6 caliber secant-ogive cylinder boattail (SOCBT) configuration shown in Figure 1 with flow conditions of Mach number 0.96 and Reynolds number of 760000, a case for which experimental data is available [12]. An initial grid configuration is shown in Figure 2 which contains 90 points in the streamwise direction and 40 points in the direction normal to the projectile surface. The boattail is extended downstream and a sting is attached to eliminate the base flow region. The solution contains an expansion wave at the ogive cylinder and cylinder boattail junctures and two shocks, one on the cylinder and one on the boattail. Another important physical

feature of the flow is the boundary layer along the projectile surface. Consequently, the choices for the control functions were the pressure gradient in the streamwise direction and the velocity gradient in the direction normal to the surface. It was found in numerical experiments, however, that the expansions required smaller spacing than the shocks for adequate resolution whereas the pressure gradient put the smallest spacing in the shocks. The dominating feature of the expansion wave is the large curvature of the pressure and therefore the second derivative of the pressure in the streamwise direction was used. This choice also refined the mesh in the vicinity of the shocks since the curvature of the pressure increased at the top and bottom of the shock. In the normal direction the grid point distribution resulting from the velocity gradient was not smooth and led to a rapid stretching of the spacing into the outer flow region. The control function was changed to the exponential of the velocity gradient which resulted in a grid point distribution similar to an exponential clustering function which is known to provide a good distribution.

The first case run on the 90 by 40 grid network had a minimum spacing of 0.04 in the streamwise direction and 0.00002 in the normal direction. The grid network was adapted every time step for 1400 time steps of 0.1 at which point the solution converged. The calculated pressure coefficient is shown in Figure 3 and an expanded view of the adapted grid near the

projectile is shown in Figure 4. The solution compares well with the experimental data except that the minimum pressure at the cylinder-boattail juncture does not agree with the experimental data and the sharp pressure rise appears upstream of the experimentally predicted location. However, a comparison of the pressure contour plot of Figure 5 to the adapted grid network of Figure 4 shows clearly the adaptation of the grid network to the pressure distribution. The extension of the shocks into the flow field is reflected in the grid network as well as the smearing of the shock in the boundary layer region. The smallest grid point spacing along the surface which occurs at the surface junctures, reflects the choice of the pressure curvature over the gradient. The grid point distribution in the normal direction indicates the adaptation of the grid network to the large velocity gradient in the boundary layer.

In order to increase solution accuracy over the boattail, the minimum spacing in the streamwise direction was reduced to 0.025 and the number of points in the normal direction was increased to 50. The solution obtained on this 90 by 50 grid network is shown in Figures 6 and 7. As indicated in Figure 6, the second expansion reaches the experimentally predicted point, however, the two sharp pressure rises do not obtain a steady position but appear to oscillate along the projectile surface. The series of plots range form 1200 to 1800 time steps and differ by 200 time steps. The pressure rise over the cylinder is moving

upstream and that over the boattail is moving downstream. This solution was continued another 2000 time steps but the propagation of the pressure rises did not cease, but rather continued in a cyclic fashion. This result could be caused by error introduced by the interpolation of the solution after each adaptation. Another possible source is the dissipation terms used in the thin-layer Navier-Stokes code. In these items the solution is considered to be a function of the computational domain and the solution's distribution in the physical domain is not considered. As the grid network is adapted, the grid point correspondence between the computational domain and the physical domain change, thus changing the value of the dissipation terms.

In order to investigate this phenomenon, two more cases were obtained on the 90 by 50 grid network. First, the solution obtained at the 1800th time step (curve 4 in Figure 6) was continued without adapting the grid network any further. The solution shown in Figure 7 converged readily, but the sharp pressure rise over the cylinder has moved down stream of the experimentally predicted position. This poor correspondence occurs because the adaptation was stopped when the pressure rise was in the upstream position, and thus leaves poor resolution in the adjacent downstream region. The steady convergence does, however, indicate that the coupling of the adaptive grid generation scheme can affect the solution. It should also be noted this result occurred in the vicinity of the shock boundary

layer regions, which is a sensitive structure. In another case, the solution at 1800 time steps was continued, but the grid network was adapted every 50 time steps for the first 200 time steps and then remained fixed until the solution converged. The calculated solution, shown in Figure 9, shows that the adaptation has improved the solution, in that the calculated pressure rise on the cylinder surface agrees well with the experimental prediction.

CONCLUDING REMARKS

The results calculated so far indicate that the adaptive grid generation equations are capable of providing reliable and predictable response to the control functions. The grids clearly show the shock and expansion features to which they were adapted and the extremely small spacing required to resolve the boundary layer was obtained. The choice of the control functions is not obvious, however, and must be guided by experience. The primary problem encountered is in the coupling of the adaptive grid generation procedure with the thin-layer Navier-Stokes code. It was found that when the grid is adapted every time step the grid motion can affect the solution near sensitive structures such as shocks and it is this process that must be investigated further. However, when the adaptive grid generation technique was modified to adapt the grid only a few times as the solution converged, the scheme worked well, indicating that the scheme is sufficient at

least for steady flow problems.

ACKNOWLEDGEMENT

This work was partially supported by the AFOSR minigrant program and the calculations were performed at the Pittsburgh Supercomputing Center through a grant from NSF.

REFERENCES

[1] Brackbill, J. U., coordinate System Control: Adaptive Meshes, in: J. F. Thompson, Ed., Numerical Grid Generation. (North-Holland 1982) pp. 277-294.

[2] Nakahashi, K. and Deiwert, G. S., A Self-Adaptive Grid Method with Application to Airfoil Flow, AIAA paper 85-1525.

[3] Hsu, C. C. and Tu, C. G., An Adaptive Grid Generation Technique Based on Variational Principles for Transonic Projectile Aerodynamics Computation, Int. J. Numerical Methods in Fluids, Vol. 7, (1987), 567-579.

[4] Thompson, J. F., Numerical Grid Generation, Elsevier Science Publishing Co., New York, 1984.

[5] Dwyer, H. A., Smooke, M. D. and Kee, R. J., Adaptive Gridding for Finite Difference Solutions to Heat and Mass Transfer Problems, in: J. F. Thompson, ed., Numerical Grid Generation, (North-Holland 1982) pp. 339-356.

[6] Thompson, J. F., A Survey of Dynamically-Adaptive Grids in the Numerical Solution of Partial Differential Equations, AIAA-84-1606, AIAA Fluid and Plasma Dynamics Conference, Snowmass, Colorado, 1984.

[7] Kennon, S. R. and Dulikaravich, G. S., Generation of Computational Grids Using Optimization, AIAA Journal, (1986) 1069-1073.

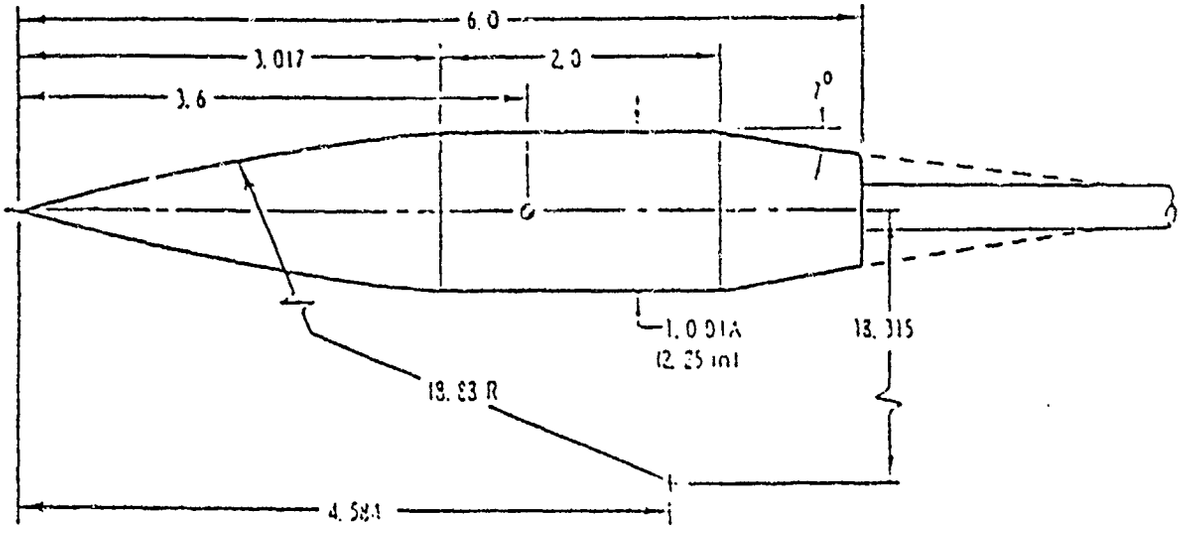
[8] Reed, C. W., Adaptive Grid Generation for Viscous Flow Problems, Final Report, 1986 USAF-UES Graduate Student Summer Support Program, August 1984.

[9] Nietubicz, C. J., Pulliam, T. H. and Steger, J. L., Numerical Solution of the Azimuthal-Invariant Thin-Layer Navier-Stokes Equations. AIAA paper 79-0010, AIAA 17th Aerospace Sciences Meeting, New Orleans, La., January 1979.

[10] Baldwin, B. S. and Lomax, H., Thin-Layer Approximation and Algebraic Model for Separated Turbulent Flows, Paper 78-257, AIAA 16th Aerospace Sciences Meeting, January 1978.

[11] Warming, R. F. and Beam, R., On the Construction and application of Implicit Factored Schemes for Conservation Laws, SIAM-AMS Proceedings, Vol. 2, (1978) 85-99.

[12] Kayser, L. D. and Whiton, F., Surface Pressure Measurements on a Boattailed Projectile Shape at Transonic Speeds, ARBRL-MR-03161, U. S. Army Ballistic Research Laboratory, March 1982.



ALL DIMENSIONS IN CALIBERS

Figure 1. Projectile configuration

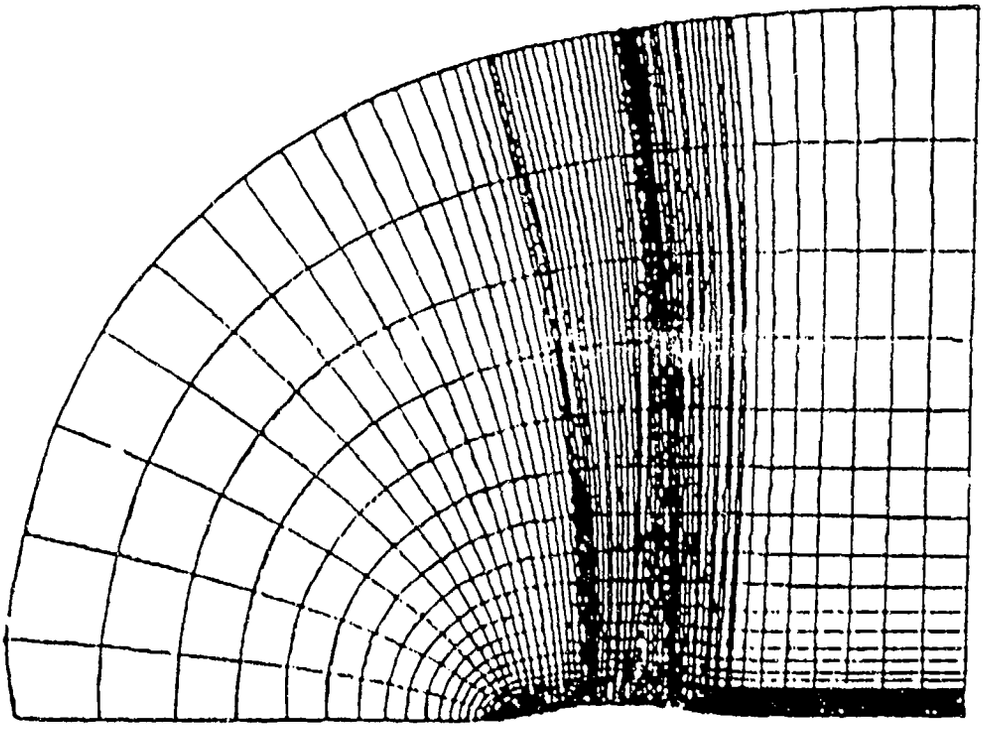


Figure 2. Initial grid configuration

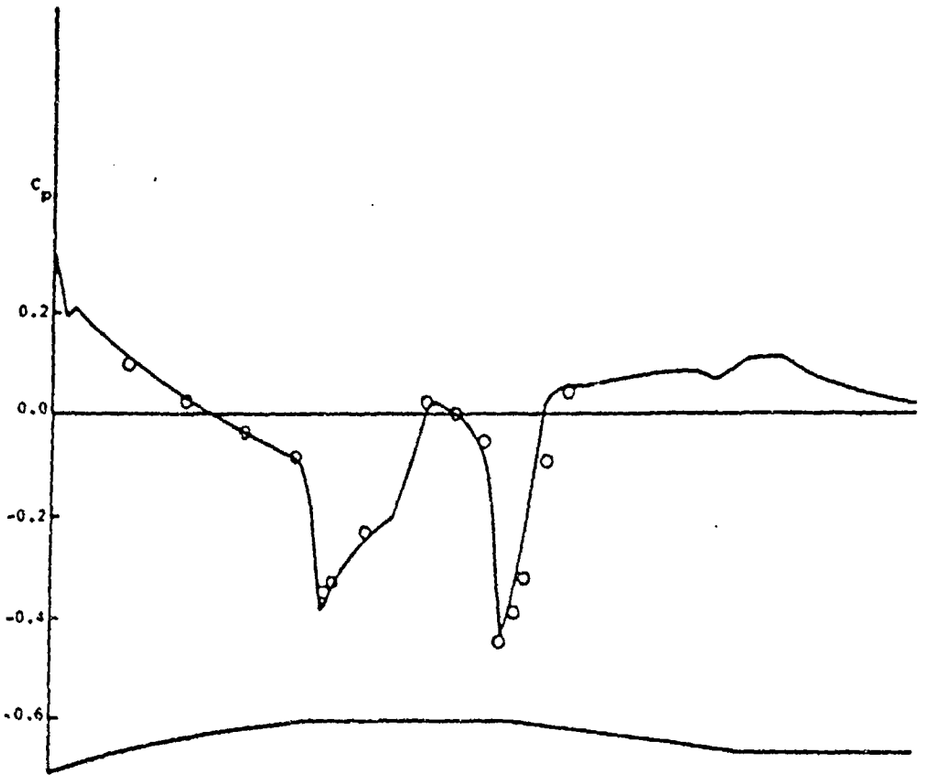


Figure 3. Calculated pressure coefficient on 90 by 40 grid

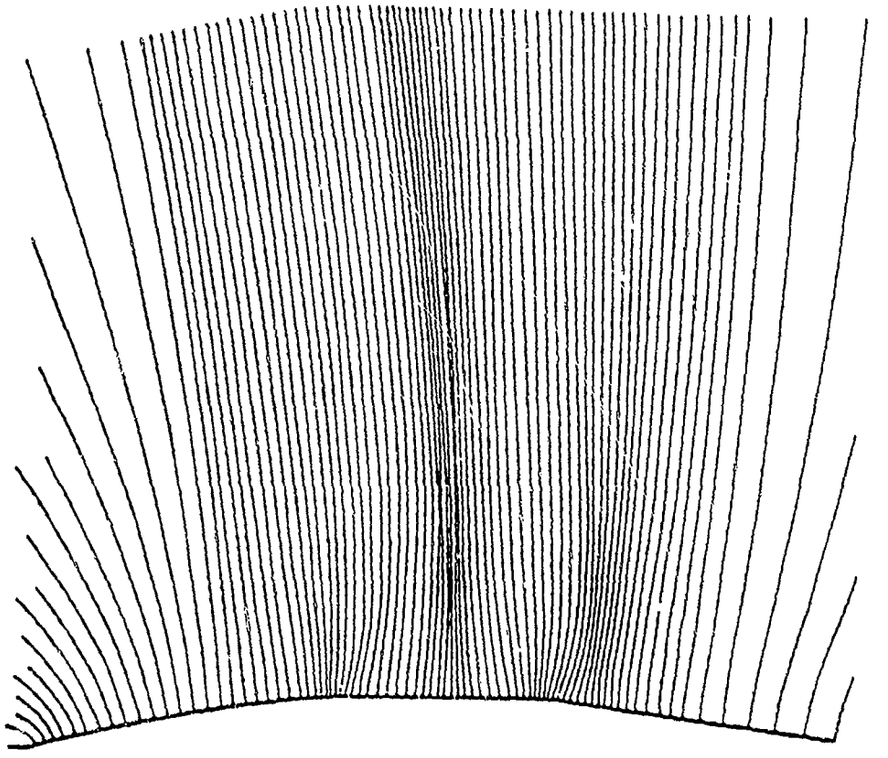


Figure 4. Final adapted grid network (90 x 40)

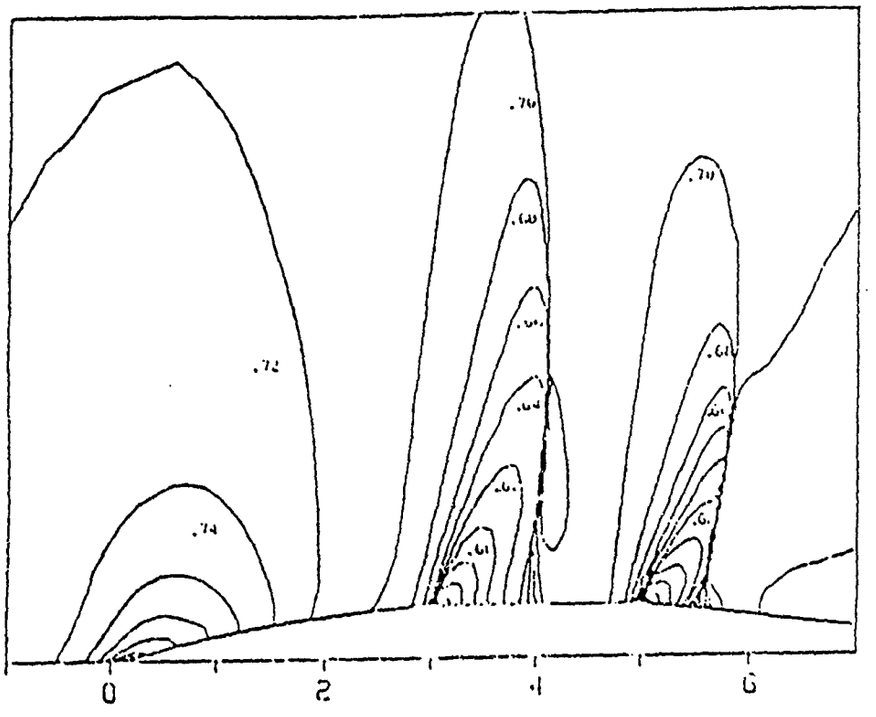


Figure 5 Pressure contours for converged solution

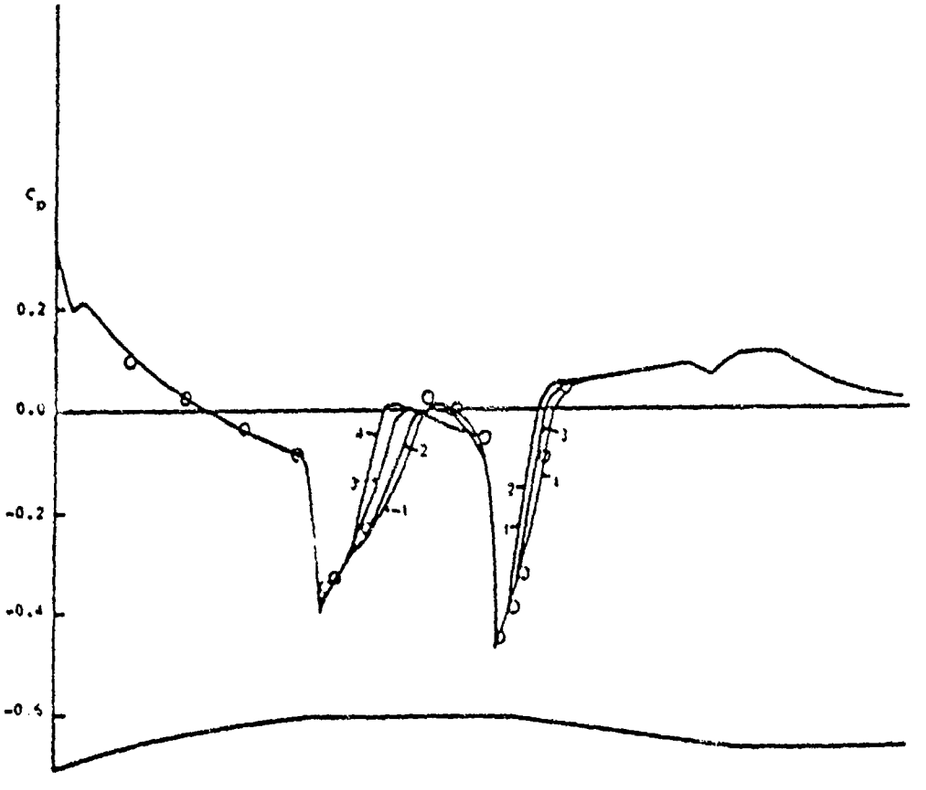


Figure 6. Calculated pressure coefficients on 90 x 50 grid (plotted every 200 time steps)

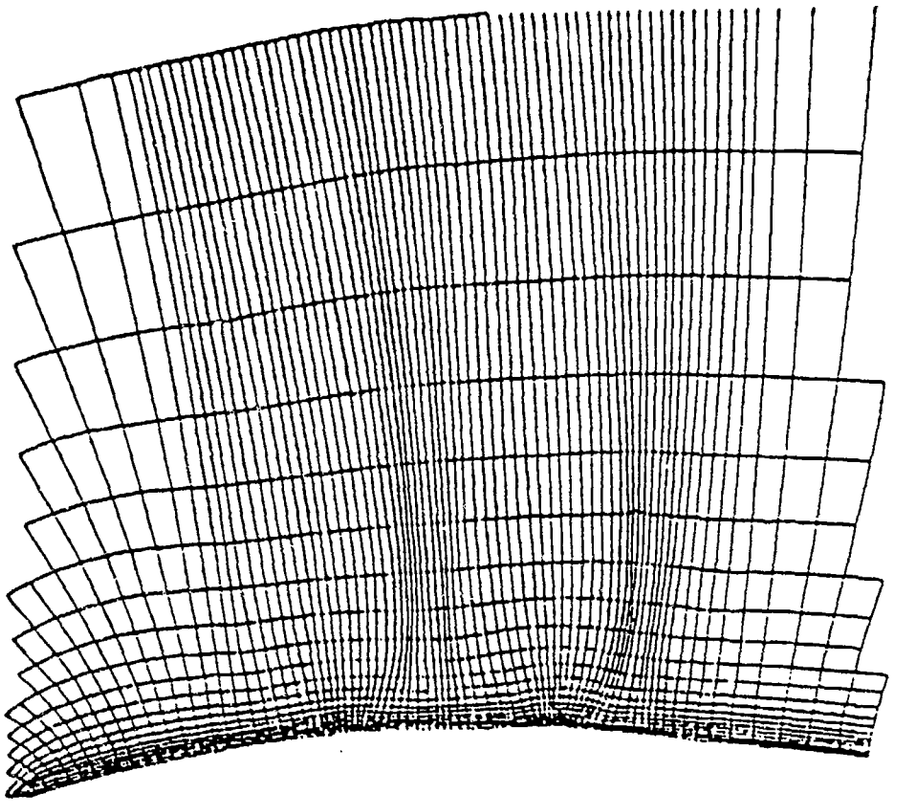


Figure 7. Adapted grid network (90 x 50) at 1800 time steps

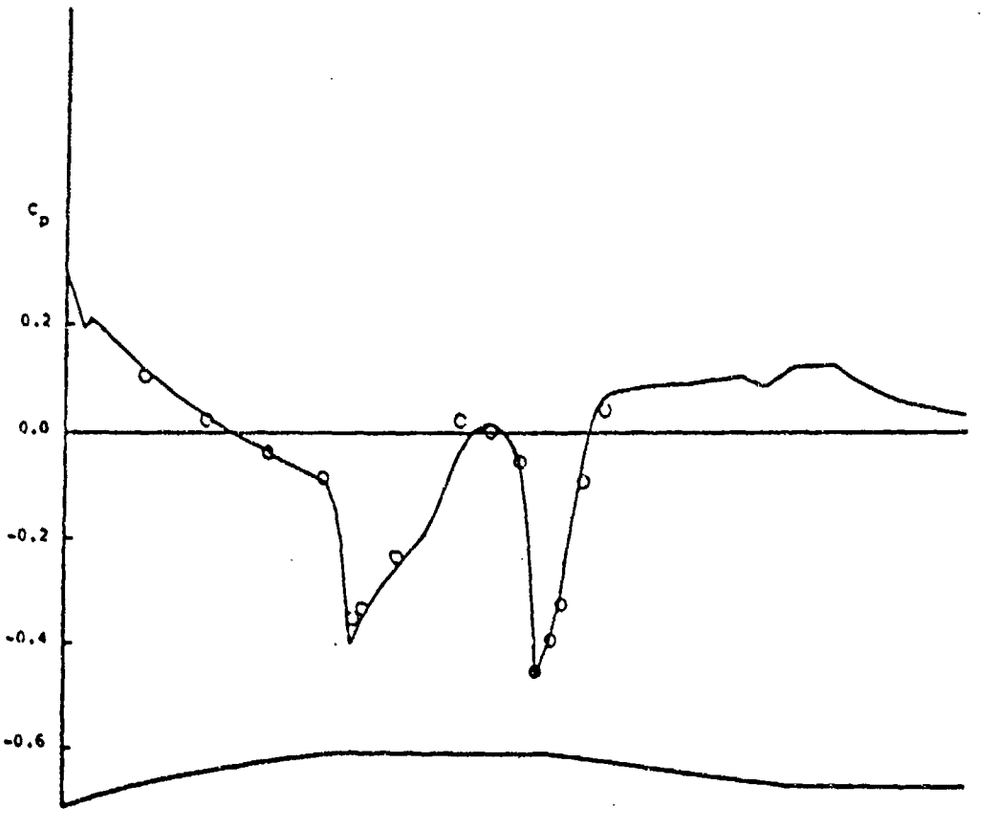


Figure 8. Calculated pressure coefficient (grid no longer adapted)

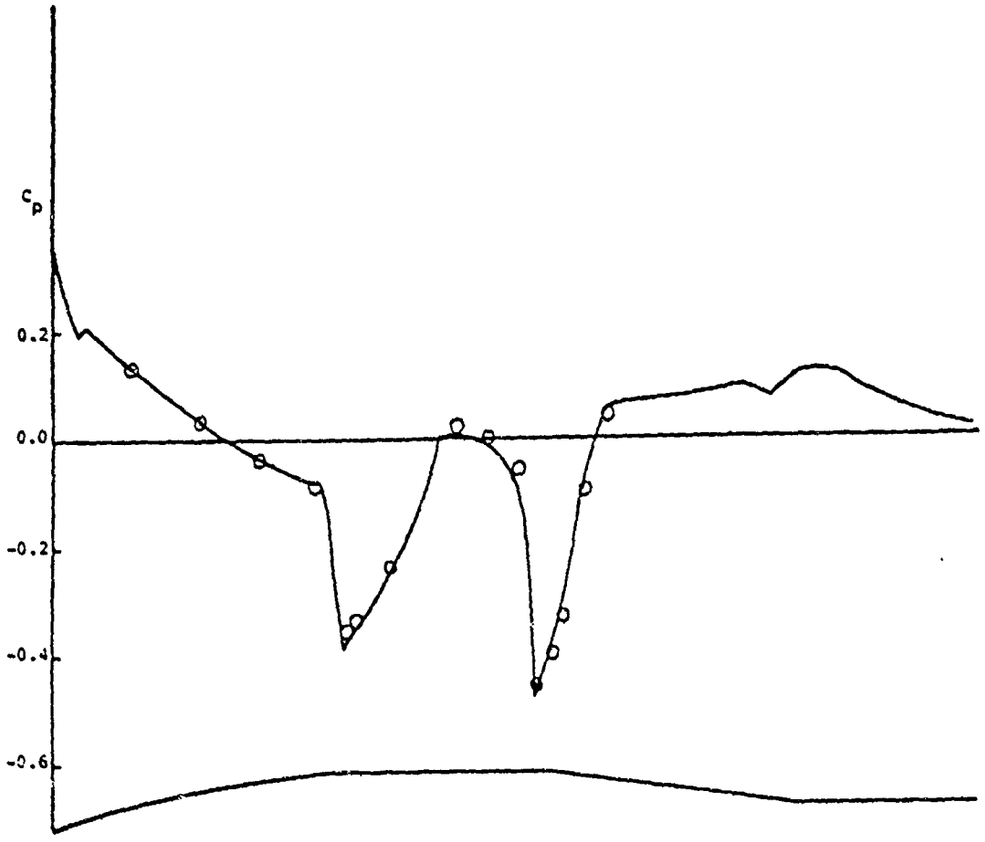


Figure 9. Calculated pressure coefficient (grid adapted every 50 time steps)

Attachment II

A paper to be presented at AIAA 26th Aerospace Sciences Meeting to be held in Reno, Nevada, January 12-14, 1988.

A DIAGONALIZED TVD SCHEME FOR TURBULENT TRANSONIC PROJECTILE AERODYNAMICS COMPUTATION

Nae-Haur Shiau* and Chen-Chi Hsu**.
Department of Engineering Sciences
University of Florida, Gainesville, Florida

Abstract

A diagonalized TVD scheme is proposed and tested on a steady turbulent transonic flow of M_∞=0.96 past a secant-ogive-cylinder-boattail projectile with sting at zero angle of attack. An axisymmetric thin-layer Navier-Stokes code, which is based on the implicit Beam and Warming scheme, obtained from the U.S. Army Ballistic Research Laboratory has been modified for additional options of selecting a diagonalized Beam and Warming scheme, a TVD scheme or the diagonalized TVD scheme. Numerical results computed for the flow problem with different hyperbolic grids have shown that the diagonalized TVD scheme is most effective among the three schemes investigated; it can give acceptably accurate results with rather coarse grid and can save about 35% to 50% of the CPU time over the original Beam and Warming scheme.

Introduction

A 3-D Navier-Stokes code based on unsteady thin-layer Navier-Stokes equations for ideal gas in a transformed boundary-fitted space was developed in 1978 by Pulliam and Steger at NASA Ames for high speed compressible flow problems.[1] In this code the governing equations are approximated by the factorized implicit scheme of Beam and Warming[2] with second order implicit and fourth order explicit artificial dissipation terms added for controlling numerical stability of the solution algorithm. The resulting system of equations to be solved is a block tridiagonal system. The turbulence closure model implemented in the code is an algebraic eddy viscosity model of Baldwin and Lomax.[3] The unsteady Navier-Stokes code has an option for solving inviscid flow problems while a steady solution is resulted from a converged solution of the unsteady flow problem. Since the development of the code, advancements have been made to improve overall computational efficiency of the code. Some of these advancements are an introduction of varying time step to accelerate the convergence rate for steady solution, a reduction of the block tridiagonal matrix inversion work by using simpler matrices, and combinations of dissipation operators to prevent severe oscillations near shocks. The improvements in efficiency, accuracy, and convergence for the solution algorithm of the code have recently been documented and reported by Pulliam and Steger.[4]

Recent publications of Yee and Harten[5] and Yee[6,7] have shown that an implicit total variation diminishing (TVD) scheme can be a very effective numerical scheme for solving steady two-dimensional high speed aerodynamic problems. The implicit TVD scheme in essence is similar to that of Beam and Warming scheme except that it employs more sophisticated dissipation terms which can be switched from a second order accurate scheme to a first order scheme near the shock to provide nonoscillatory and accurate solutions. Hence, the resulting system of equations to be solved is still a block tridiagonal system. Experience shows that an inversion process of the block tridiagonal matrix is rather expensive for a complex aerodynamic problem.

In this study a diagonalized TVD scheme is proposed and investigated for effective computation of steady turbulent transonic projectile aerodynamics. An axisymmetric version of the 1978 3-D Navier-Stokes code obtained from the U.S. Ballistic Research Laboratory is employed in this study. This axisymmetric code has been modified to provide different options for selecting the original Beam and Warming scheme, a diagonalized Beam and Warming scheme, a TVD scheme, or a diagonalized TVD scheme. A turbulent transonic flow of Mach number equal to 0.96 past a secant-ogive-cylinder-boattail (SOCBT) projectile with sting at zero angle of attack is considered for assessing the effectiveness of the four finite-difference schemes.

Governing Equations

For an axisymmetric flow problem, the transformed unsteady thin-layer Navier-Stokes equations for ideal gas can be written in strong conservation law form as[8]

$$\partial_\tau q + \partial_\xi E + \partial_\zeta G = Re^{-1}\partial_\zeta S - H \qquad (1)$$

where (τ,ξ,ζ) is the general curvilinear coordinate system. The unknown vector q, flux vectors E,G,S, and the source vector H are

---

* Graduate student, Aerospace Engineering Program.
** Professor, Department of Engineering Sciences.
Released to AIAA to publish in all forms

$$q=\frac{1}{J}\begin{Bmatrix}\rho\\ \rho u\\ \rho v\\ \rho w\\ e\end{Bmatrix},\qquad E=\frac{1}{J}\begin{Bmatrix}\rho U\\ \rho uU+\zeta_x p\\ \rho vU\\ \rho wU+\zeta_z p\\ (e+p)U-\zeta_t p\end{Bmatrix},$$

$$C=\frac{1}{J}\begin{Bmatrix}\rho W\\ \rho uW+\zeta_x p\\ \rho vW\\ \rho wW+\zeta_z p\\ (e+p)W-\zeta_t p\end{Bmatrix},\qquad S=\frac{1}{J}\begin{Bmatrix}0\\ m_1u_\zeta+m_2\zeta_x\\ m_1v_\zeta\\ m_1w_\zeta+m_2\zeta_z\\ m_1m_3+m_2(\zeta_x u+\zeta_z w)\end{Bmatrix},$$

$$H=\frac{1}{J}\begin{Bmatrix}0\\ 0\\ \rho V[R_\zeta(U-\zeta_t)+R_\zeta(W-\zeta_t)]\\ -\rho VR(V-\eta_t)-p/R\\ 0\end{Bmatrix},$$

in which m_1, m_2, m_3 and the contravariant velocity components U, V and W are given by

$$m_1 = \mu(\zeta_x^2 + \zeta_z^2)$$

$$m_2 = \mu(\zeta_x u_\zeta + \zeta_z w_\zeta)/3$$

$$m_3 = (u^2 + v^2 + w^2)_\zeta/2 + Pr^{-1}(\gamma-1)^{-1}(c^2)_\zeta$$

$$U = \zeta_t + \zeta_x u + \zeta_z w$$

$$V = \eta_t + \eta_y v$$

$$W = \zeta_t + \zeta_x u + \zeta_z w$$

while certain relationships between the computational coordinates (ξ,η,ζ) Cartesian coordinates (x,y,z) and cylindrical coordinates (R,θ,x) can be obtained from the definition of the coordinate systems shown in Figure 1. For turbulent flow problems, the eddy viscosity is approximated by Baldwin-Lomax eddy viscosity model.

Finite-Difference Schemes

An application of a noniterative approximate factorization implicit method of Beam and Warming to the governing Navier-Stokes equations, Eq (1), gives the finite-difference equations

$$\hat{L}_\xi\ \hat{L}_\zeta\ \Delta q^n = R^n \qquad (2)$$

with

$$\hat{L}_\xi = I + \Delta t\ \delta_\xi A^n$$

$$\hat{L}_\zeta = I + \Delta t\ \delta_\zeta C^n - \Delta t\ Re^{-1}\delta_\zeta J^{-1}M^n J$$

$$R^n = -\Delta t\ (\delta_\xi E + \delta_\zeta G + H - Re^{-1}\delta_\zeta S)^n$$

in which A, C and M are the Jacobian matrices defined as

$$A = \frac{\partial E}{\partial q}, \qquad C = \frac{\partial G}{\partial q}, \qquad M = \frac{\partial S}{\partial q}$$

The solution process for Δq^n in Eq.(2) can be carried out in two steps:

$$\hat{L}_\zeta\ q^* = R^n \quad \text{and} \quad \hat{L}_\zeta\ \Delta q^n = q^* \qquad (3)$$

Since both operators $\hat{L}_\xi$ and $\hat{L}_\zeta$ are block tridiagonal matrices, it can be rather expensive to compute Δq^n for a complex flow problem. The number of operations required for solving Eq.(3) could be reduced considerably by Thomas algorithm if both L_ξ and L_ζ were scalar tridiagonal matrices.

For steady flow problems, equation (2) for Δq^n can be considered as an iterative scheme for finding Δq^n; consequently, a different iterative scheme of similar form

$$L_\xi\ L_\zeta\ \Delta q^n = R^n \qquad (4)$$

can also be employed to obtain Δq^n for Eq.(2). This implies that the operators $\hat{L}_\xi$ and $\hat{L}_\zeta$ of Eq.(2) can be modified to solving steady flow problems. A diagonal form of the implicit scheme was proposed by Pulliam and Chaussee of NASA Ames in 1981.[9] Since flux Jacobian matrices A and C have real eigenvalues and a complete set of eigenvectors, they can be decomposed as

$$A = T_\xi \Lambda_\xi T_\xi^{-1} \quad \text{and} \quad C = T_\zeta \Lambda_\zeta T_\zeta^{-1} \qquad (5)$$

Here T_ξ and T_ζ are eigenvector matrices of A and C, respectively, while the diagonal matrices Λ_ξ and Λ_ζ are the corresponding eigenvalue matrices. Use of the decomposition, Eq.(5), and the relations

$$T_\xi\ T_\xi^{-1} = I \quad \text{and} \quad T_\zeta\ T_\zeta^{-1} = I\ ,$$

the operators in Eq.(2) can be written as

$$\begin{aligned}\hat{L}_\xi &= T_\xi[I+\Delta t\ T_\xi^{-1}\delta_\xi(T_\xi\Lambda_\xi)]T_\xi^{-1}\\ \hat{L}_\zeta &= T_\zeta[I+\Delta t\ T_\zeta^{-1}\delta_\zeta(T_\zeta\Lambda_\zeta)\\ &\quad -\Delta t\ Re^{-1}T_\zeta^{-1}\delta_\zeta(J^{-1}MJ)T_\zeta]T_\zeta^{-1}\end{aligned} \qquad (6)$$

One observes that if T^{-1} is put inside the difference operator δ then the second term in the brackets is reduced to a scalar tridiagonal matrix from a block tridiagonal matrix; however, the last term of $\hat{L}_\zeta$ remains the same in structure. Hence, if one selects the modified operators for Eq.(4) as

$$\begin{aligned}L_\xi &= T_\xi(\ I + \Delta t\ \delta_\xi\Lambda_\xi\ )T_\xi^{-1}\\ L_\zeta &= T_\zeta(\ I + \Delta t\ \delta_\zeta\Lambda_\zeta\ )T_\zeta^{-1}\end{aligned} \qquad (7)$$

then the solution process for finding Δq^n involves only tridiagonal systems of equations which can be effectively solved by Thomas algorithm.

For a complex aerodynamic problem, a direct application of Eq.(2) or Eq.(4) often runs into the convergence problem. Hence, implicit dissipation operators D_I and explicit dissipation term D_E must be added to control the nonlinear numerical instability. In fact, a proper choice of the dissipation terms will have a great impact on the efficiency and accuracy of a solution algorithm. In the original axisymmetric Navier-Stokes code, second order implicit dissipation operators

$$D_{I\xi} = -\epsilon_I \Delta t\, J^{-1} \nabla_\xi \Delta_\xi J$$
$$D_{I\zeta} = -\epsilon_I \Delta t\, J^{-1} \nabla_\zeta \Delta_\zeta J \qquad (8)$$

are added to the operators $\hat{L}_\xi$ and $\hat{L}_\zeta$, respectively, while an explicit fourth order dissipation term

$$D_E = -\epsilon_E \Delta t\, J^{-1}[(\nabla_\xi \Delta_\xi)^2 + (\nabla_\zeta \Delta_\zeta)^2] J q^n \qquad (9)$$

is added to R^n on the right hand side of Eq.(2).

The TVD scheme investigated by Yee[6,7] is in essence the same as the Beam and Warming scheme except that the implicit dissipatron operators and the explicit dissipation term are more complex in nature than those of Eqs. (8) and (9). Following the work of Yee, the following dissipation terms have been implemented into the axisymmetric Navier-Stokes code

$$D_{I\xi} = -0.5\ \Delta t\ (\Omega_{j+1/2,\ell} - \Omega_{j-1/2,\ell}) \qquad (10)$$
$$D_{I\zeta} = -0.5\ \Delta t\ (\Omega_{j,\ell+1/2} - \Omega_{j,\ell-1/2})$$

$$D_E = -0.5\ \Delta t\ (T_{j+1/2}\phi_{j+1/2} - T_{j-1/2}\phi_{j-1/2} + T_{\ell+1/2}\phi_{\ell+1/2} - T_{\ell-1/2}\phi_{\ell-1/2}) \qquad (11)$$

In equation (10), the operators Ω are defined as

$$\Omega_{j+1/2,\ell} = |\mathrm{diag}|-\max\Psi(\lambda_\xi^k)||_{j+1/2}\Delta_{j+1/2}$$
$$\Omega_{j,\ell+1/2} = |\mathrm{diag}|-\max\Psi(\lambda_\zeta^k)||_{\ell+1/2}\Delta_{\ell+1/2}$$

in which $\Psi(z)$ is an entropy correction function given by

$$\Psi(z) = \begin{cases} |z| & , \quad |z| \geq \epsilon = 0.25 \\ (\epsilon^2 + z^2)/(2\epsilon) & , \quad |z| < \epsilon = 0.25 \end{cases}$$

and λ_ξ^k, λ_ζ^k are the kth eigenvalue of the Jacobian matrices A and C, respectively. In equation (11), $T_{j+1/2}$, $T_{\ell+1/2}$ are, respectively, the eigenvector matrices T_ξ, T_ζ evaluated at $q_{j+1/2,\ell}$ and at $q_{j,\ell+1/2}$. The $q_{j+1/2,\ell}$ and $q_{j,\ell+1/2}$ are the Roe's average of ($q_{j,\ell}$ and $q_{j+1,\ell}$) and ($q_{j,\ell}$ and $q_{j,\ell+1}$), respectively.[10] The kth element of the vector $\phi_{j+1/2}$ is

$$\phi^k_{j+1/2} = \Psi(\lambda^k_{j+1/2})(g^k_j + g^k_{j+1} - 2\alpha^k_{j+1/2})$$
$$g^k_j = S\ \max[0, \min(|\alpha^k_{j+1/2}|, S\alpha_{j-1/2})] \qquad (12)$$
$$S = \mathrm{sign}(\alpha^k_{j+1/2})$$

$$\alpha_{j+1/2} = T^{-1}_{j+1/2}\ \frac{q_{j+1,\ell} - q_{j,\ell}}{0.5\times(J_{j+1,\ell} + J_{j,\ell})}$$

Similar expressions for elements of $\phi_{\ell+1/2}$ in the ζ direction are used. The form of g in Eq.(12) devised by Harten[11] represents the switching mechanism which can change the dissipation term of Eq.(11) from 2nd order into 1st order at points of extrema. These smart dissipation terms of Eq. (11) provide automatic feedback controlling the amount of numerical smoothing without introducing spurious oscillations near discontinuities. A detailed explanation of these sophisticated dissipation terms can be found in Ref. 5,6 and 7.

The original axisymmetric Navier-Stokes code also has been modified for the diagonalized scheme Eq.(4) with operators given by Eq.(7). A nonlinear artificial dissipation model proposed by Jameson et al.[12] has been employed for the diagonalized scheme by Pulliam.[13] This model is also implemented into the Navier-Stokes code for the transonic projectile aerodynamics computation. The results of our numerical experiments show that the scheme with constant time step is not stable; however, if variable time step is used then the solution algorithm is rather effective. For a diagonalized TVD scheme, we propose that the implicit dissipation operators $D_{I\xi}$ and $D_{I\zeta}$ given by Eq.(10) be added to the modified differential operators L_ξ and L_ζ, respectively and that the explicit dissipation term given by Eq.(11) be added to the right hand side of Eq.(4). Since the implicit dissipation operators are tridiagonal matrices, the resulting systems of equations to be solved for Δq^n remain as the tridiagonal systems. Hence, the Thomas algorithm is programmed in the code for finding Δq^n.

Results and Discussion

In order to assess the accuracy and efficiency of the proposed diagonalized TVD scheme, a steady turbulent transonic flow of M_∞=0.96 past a secant-ogive-cylinder-boattail projectile with sting at zero angle of attack is considered in this study. Surface pressure measurements are available[14] for assessing the numerical results. In the computation, the boattail part of the projectile model was further extended to meet the sting in order to avoid the difficulty of simulating the base flow region. The grid network chosen for the Navier-Stokes computation is a modified adaptive hyperbolic grid; a 90×60 hyperbolic grid used is shown in Figure 2. Five different grids have been selected to investigate

the effectiveness of TVD schemes; accordingly, each case is also solved by the Beam and Warming scheme with the same sequence of time steps used in TVD schemes. These computer codes, Beam and Warming, TVD, diagonalized TVD algorithm, have the same basic structures. The difference between them is described as follows: the computer program of Beam and Warming algorithm is fully vectorized; the TVD algorithm is vectorized except for the computation of artificial dissipations; the diagonalized TVD algorithm is vectorized except for the computation of artificial dissiaptions and tridiagonal solver-Thomas algorithm. In this study a converged steady solution is assumed when the residual becomes less than 10^{-4}.

It is known that the boundary grid points distribution 1-25-58-82-90 for secant-ogive, cylinder, extension boattail and sting of a 90*60 grid, shown in Figure 2, is a very good grid for the flow problem.[15] The surface pressure distribution computed with this grid by the Beam and Warming scheme is presented in Figure 3; indeed, it is in excellent agreement with the measured data. The computed Mach contours also presented in Figure 4, which indicates two shock locations, one in the middle of cylinder part, another in the middle of boattail part, were predicted and agree well with the shadowgraph presented in Ref. 14. The convergence process of the Beam and Warming scheme is given in Figure 5. From the residual plot, it is clearly shown that the solutions converged in an oscillatory manner. In fact, the surface pressure distribution has been plotted at every 100 iterations during the convergence process; the oscillation of solutions has also been observed especially near the shock/ boundary layer interaction regions. The flow problem with the same 90*60 grid is also solved by the TVD scheme and the diagonalized TVD scheme. As one would have expected, the converged solutions obtained from both TVD schemes are nearly exactly the same, since the same explicit dissipation term is used in both schemes. The distribution of surface pressures, computed and given in Figure 3, indicates that the TVD schemes can give very accurate solution for the flow problem. A corresponding Mach contour plot is presented in Figure 6; the sharpness of the two normal shock waves are clearly exhibited. Moreover, the results show that both TVD schemes have almost the same rate of convergence even though the implicit operators are different; and also, both solutions converged monotonically compared with the Beam and Warming scheme presented. This is one of characteristics of TVD schemes. Figure 5 shows that the TVD schemes converge much faster than the Beam and Warming scheme; it takes 3480 time steps for the Beam and Warming scheme, yet only 1620 for the TVD scheme to reach the converged solution.

The ratio of the CPU time on a Cray X-MP/48 required by the Beam and Warming scheme, the TVD scheme and the diagonalized TVD scheme to complete one-time step computation is, respectively, 1.00, 1.76 and 1.47. Hence, the ratio of the total CPU time required to obtain a converged solution is 1.00, 0.82 and 0.68, respectively, for the Beam and Warming scheme, the TVD scheme and the diagonalized TVD scheme. The same flow case has also been solved on a Harris-800 scalar computer. Ratios of the CPU time on a Harris-800 required for the Beam and Warming scheme, the TVD scheme and the diagonalized TVD scheme to complete one-time step computation is 1.00, 1.24 and 0.74, respectively. Ratios for the total CPU time required to obtain a converged solution is 1.00, 0.58 and 0.34 for the Beam and Warming scheme, the TVD scheme and the diagonalized TVD scheme, respectively.

The flow problem is also solved by the Beam and Warming scheme and the diagonalized TVD scheme with 90*50, 90*40 grids for investigating the effectiveness of grid resolution in the normal direction. The surface pressure distributions computed with a 90*40 grid is shown in Figure 7. It is clearly shown that the accuracy of the Beam and Warming scheme is rather sensitive to the grid resolution in the normal direction. The corresponding Mach contour plots computed and presented in Figrues 8 and 9, shows that the Beam and Warming scheme didn't catch the shock in the middle of the cylinder part and contour plot is quite different from the case solved on 90*60 grid shown in Figure 4. Next, two grids 80*60 and 70*60 with boundary grid point distributions 1-20-48-72-80 and 1-20-40-62-70, respectively, was considered for investigation on the effectiveness of grid resolution in the streamwise direction. The surface pressure distribution, presented in Figure 10, shows that 70 boundary grid points in streamwise direction is sufficient for both schemes and computed results also agree very well with experimental data. Moreover, Figure 11 indicates that the convergence process of the Beam and Warming scheme can be very sensitive to the grid resolution in the streamwise direction for the transonic flow problem.

The numerical results obtained in this study clearly show that the proposed diagonalized TVD scheme is an extremely effective solution algorithm for steady turbulent transonic projectile aerodynamics. It can give acceptably accurate solutions with rather coarse grids and can save about 60% to 80% of CPU time over the Beam and Warming scheme on scalar computers, or save about 35% to 50% of CPU time on a vector computer for a partial vectorized diagonalized TVD computer code.

Acknowledgement

The authors wish to thank Dr. H.C. Yee of the NASA Ames Research Center for providing them a copy of her 2-D TVD code. This work was partially supported by a 1986 USAF-UES mini-grant and the calculation was performed by a Harris-800 scalar computer system at the University of Florida and by a Cray X-MP/48 at National Center for Supercomputing Applications (NCSA), University of Illinois. The Cray CPU time was provided by NCSA through a supercomputer service grant.

Reference

1. Pulliam, T.H. and Steger, J.L., "Implicit Finite-Difference Simulations of Three-Dimensional Compressible Flow," AIAA Journal, Vol. 18, Feb. 1980, pp. 159-167.
2. Beam, R.M. and Warming, R.F., "An Implicit Finite-Difference Algorithm for Hyperbolic Systems in Conservation-Law Form," Journal of Computational Physics, Vol. 22, Sep. 1976, pp. 87-110.
3. Baldwin, B.S. and Lomax, H., "Thin Layer Approximation and Algebraic Model for Separated Turbulent Flows," AIAA Paper 78-257, AIAA 16th Aerospace Sciences Meeting, Huntsville, Ala., Jan. 1978.
4. Pulliam, T.H. and Steger, J.L., "Recent Improvements in Efficiency, Accuracy, and Convergence for Implicit Approximate Factorization Algorithms," AIAA Paper 85-0360, AIAA 23rd Aerospace Sciences Meeting, Reno, Nev., Jan. 1985.
5. Yee, H.C. and Harten, A., "Implicit TVD schemes for hyperbolic Conservation Laws in Curvilinear Coordinates," AIAA Paper 85-1513, AIAA 7th Computational Fluid Dynamics Conference, July 1985.
6. Yee, H.C., "On Symmetric and Upwind TVD Schemes," Proceedings of the 6th GAMM Conference on Numerical Methods in Fluid Mechanics, Sep. 1985.
7. Yee, H.C., "Linearized Form of Implicit TVD Schemes for the Multidimensional Euler and Navier-Stokes Equations," International Journal on Computers and Mathematics with Applications, Dec. 1985.
8. Nietubicz, C.J., Pulliam, T.H. and Steger, J.L., "Numerical Solution of the Azimuthal-Invariant Thin-Layer Navier-Stokes Equations," AIAA Paper 79-0010, AIAA 17th Aerospace Sciences Meeting, New Orleans, La., Jan. 1979.
9. Pulliam, T.H. and Chaussee, D.S., "A Diagonal Form of an Implicit Approximate-Factorization Algorithm," Journal of Computational Physics, Vol. 39, Feb. 1981, pp. 347-363.
10. Roe, P.L., "Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes," Journal of Computational Physics, Vol. 43, 1981, pp. 357-372.
11. Harten, A., "High Resolution Schemes for Hyperbolic Conservation Laws," Journal of Computational Physics, Vol. 49, 1983, pp. 357-393.
12. Jameson, A., Schmidt, W. and Turkel, E., "Numerical Solutions of the Euler Equations by Finite Volume Methods Using Runge-Kutta Time-Stepping Schemes," AIAA Paper 81-1259 AIAA 14th Fluid and Plasma Dynamics Conference, Palo Alto, 1981.
13. Pulliam, T.H., "Artificial Dissipation Models for the Euler Equations," AIAA Paper 85-0438, AIAA 23rd Aerospace Sciences Meeting, Reno, Nev., Jan. 1985.
14. Kayser, L.D. and Whiton, P., "Surface Pressure Measurements on a Boattailed Projectile Shape at Transonic Speeds," ARBRL-MR-03161, U.S. Army Ballistic Research Laboratory, March 1982.
15. Hsu, C.C. and Shiau, N.H., "Numerical Simulation of Turbulent Transonic Projectile Aerodynamics," AIAA Paper 87-1237, AIAA 19th Fluid Dynamics, Plasma Dynamics and Lasers Conference, Honolulu, Hawaii, June, 1987.

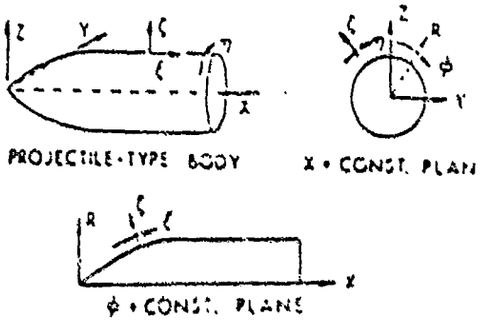


Figure 1. Axisymmetric body and coordinate systems.

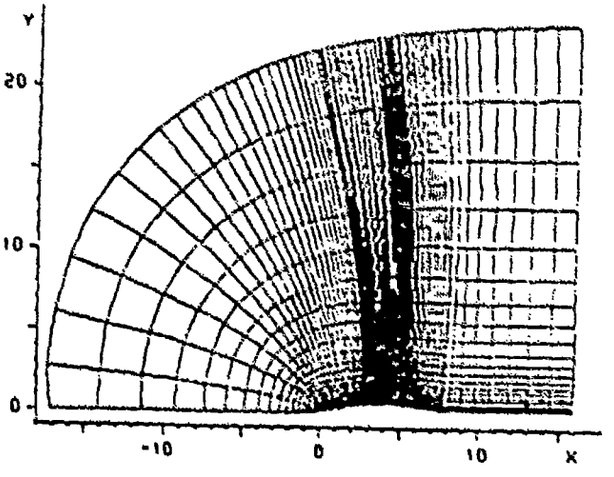


Figure 2. 90×60 hyperbolic grid system.

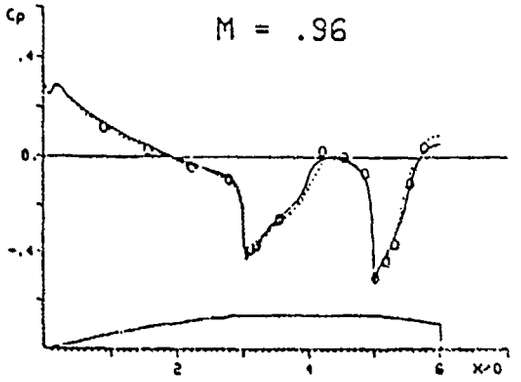


Figure 3. Surface pressure coefficient computed with a 90×60 grid.
O : measured data,
—— diagonalized TVD and TVD schemes,
---- Beam and Warming scheme.

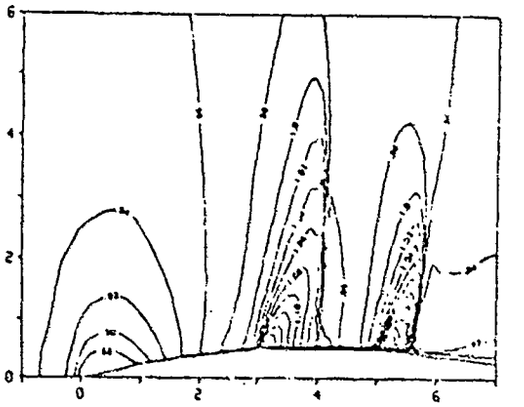


Figure 6. Mach contours resulted from diagonalized TVD scheme with 90×60 grid.

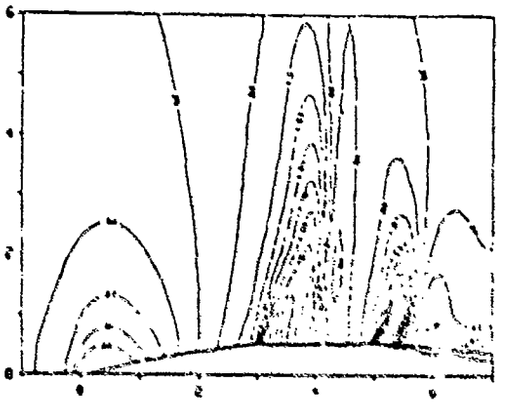


Figure 4. Mach contours resulted from Beam and Warming scheme with 90×60 grid.

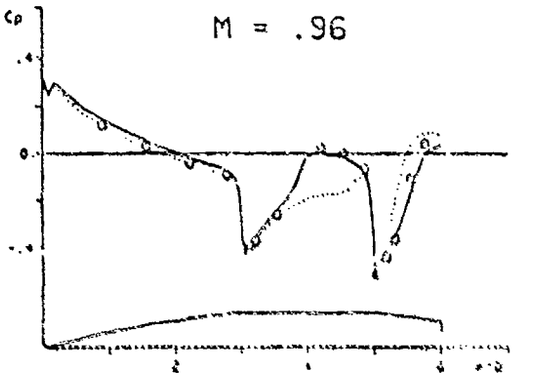


Figure 7. Surface pressure coefficient computed with a 90×40 grid
O : measured data,
—— diagonalized TVD scheme.
---- Beam and Warming scheme.

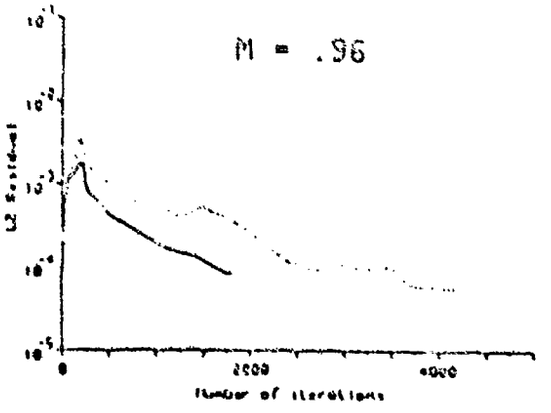


Figure 5. Rate of convergence with 90×60 grid.
—— diagonalized TVD and TVD schemes,
---- Beam and Warming scheme.

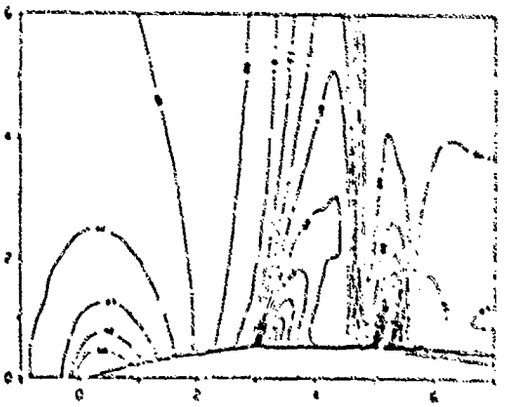


Figure 8. Mach contours resulted from Beam and Warming scheme with 90×40 grid.

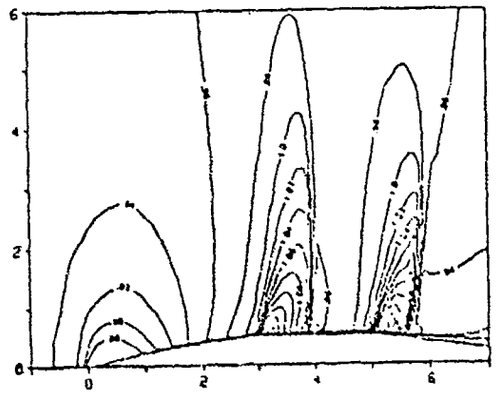


Figure 9. Mach contours resulted from diagonalized TVD scheme with 90×40 grid.

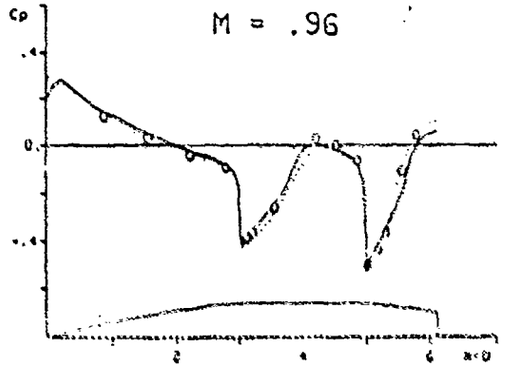


Figure 10. Surface pressure coefficient computed with a 70×60 grid
○ : measured data,
—— diagonalized TVD scheme.
---- Beam and Warming scheme.

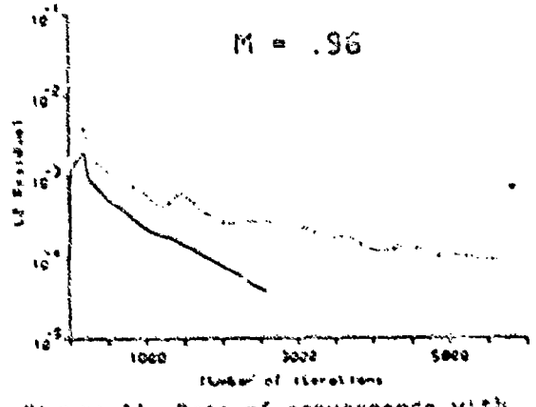


Figure 11. Rate of convergence with 70×60 grid.
—— diagonalized TVD scheme.
--- Beam and Warming scheme.

Attachment III

A paper to be presented at AIAA 26th Aerospace Sciences Meeting, Reno, Nevada, January 11-14, 1988.

AN ADAPTIVE GRID GENERATION TECHNIQUE FOR VISCOUS TRANSONIC FLOW PROBLEMS

C. W. Reed*
Systea Dynamics, Inc., Gainesville, Flordia
C. C. Hsu** and N. H. Shiau***
Department of Engineering Sciences
University of Florida, Gainesville, Florida

Abstract

An adaptive grid generation procedure is developed for viscous flow problems. The equations governing the adaptation are derived using a variational statement resulting in a set of elliptic equations in which adaptation can occur independently in each coordinate direction. The equations allow for explicit control of adaptation and orthogonality while smoothness is inherent in the elliptic equations. They retain a simple relationship between the control functions and the grid point spacing, the minimum and maximum grid point spacing may be specified and the method is capable of providing the extremely refined mesh in the boundary layer regions. The adaptive grid generation technique has been used with a TVD scheme to solve a transonic projectile flow problem. The results indicate that the adaptive grid generation procedure can reliably provide good adaptive grid networks provided proper choices are made for the control functions.

Introduction

The study of self-adaptive grid generation techniques has become an important area of computational fluid dynamics since it has been shown to provide good grid networks for the complex flow fields occuring in transonic and supersonic flows[1,2,3]. The use of boundary fitted curvilinear coordinate systems with transformed governing equations leads naturally to the concept of solution adaptive grid networks. Uniform mesh refinement throughout the entire flow domain is prohibited by existing computer storage and CPU time restrictions, and it is general practice, therefore, to vary the grid point spacing in the physical domain to increase resolution in only those regions in which the solution is changing rapidly. When the position of important solution gradients is known, good adaptive grids can be obtained with conventional techniques. However, if this information is not available a priori, the development of a proper adaptive grid network is difficult. Solution adaptive grid generation addresses this problem by continuously updating the grid network as the solution evolves such that the important physical gradients are sufficiently resolved as they develop.

The characteristics of a good grid network are considered here to include adaptation, orthogonality and smoothness. This conclusion is based both on experience and an analysis of the truncation error terms of finite difference approximations to derivatives transformed onto the computational plane[4]. Thus for the present application of viscous transonic flow over a projectile, a good adaptive grid generation scheme should provide optimization of these three characteristics.

An approach of many proposed adaptive grid generation schemes is based on minimization techniques. A measure of each desired grid characteristics is defined, and the governing equations for the grid are obtained by minimizing the integral of these measures over the domain. Dwyer et al.[5], for instance, adapted the grid point spacing along one family of coordinate lines in a two dimensional problem using an equidistribution law. In another one-dimensional application, Gnoffo[6] used a tension spring analogy in which the adapted grid point spacing along the family of coordinate lines resulted from a minimization of the spring system's potential energy. The one-dimensional approach has been extended to two dimensions by Nakahashi and Deiwert[7] by successively adapting the grid point spacing along each family of coordinate lines. They added torsional springs into the system to control orthogonality. A primary advantage to the one-dimensional approach is that the grid can be adapted independently in each coordinate direction. However, it can be difficult to maintain smoothness in such approaches and highly skewed or overlapped grids may result.

Saltzman and Brackbill[8] have developed an adaptive grid generation scheme in two dimensions using a variational approach. A set of elliptic partial differential equations is obtained by minimizing three functionals which measure the desired grid characteristics of smoothness, orthogonality and adaptation over the domain. They successfully applied this technique in the solution of a supersonic inviscid flow over a step by adapting the grid cell size to a function of the pressure gradient. A primary advantage of this approach is the use of elliptic partial differential equations. The elliptic operator, which results directly from a measure of smoothness, inherently provides a smooth grid network and helps to prevent grid overlapping, a property especially advantageous in three dimensions and for complex geometries.

This method, however, may not be suitable for viscous flow applications. The solution to the

---

* Research Engineer, System Dynamics, Inc.
** Professor, Department of Engineering Sciences
*** Graduate student, Department of Engineering Sciences

Released to AIAA to publish in all forms

viscous transonic projectile flow considered here usually contains shock structures aligned with one coordinate direction and boundary shear layers parallel to a streamwise coordinate. The grid point spacing required for adequate resolution of these structures varies by orders of magnitude and it is therefore desirable to adapt the grid network independently in each direction. The adaptive grid generation procedure proposed here, which is based on a variational approach, has been developed with the goal of providing adaptation independently in each coordinate direction within the framework of a set of elliptic partial differential equations. It is believed that the elliptic equations will reliably provide smoothly varying grids and help prevent grid distortion, skewness and overlapping as the adaptation proceeds.

The application of the variational approach developed here required modifications to both the governing equation for adaptive grid generation and the iterative technique used to solve the equations. These modifications include local scaling of the equations, additional source terms to remove the effects of curved boundaries and the temporary removal of some grid points during the iterative solution of the grid generation equations. The grid generation method has been used with a TVD scheme to solve a viscous transonic axisymmetric projectile flow problem at zero angle of attack. Due to the axial symmetry, the flow domain can be described by two independent spatial coordinates, and thus the adaptive grid generation equations are developed in two dimensions.

Adaptive Grid Generation Equations

A Variational Approach

In the formulation of the grid generation equations, the two curvilinear coordinates are designated as ξ and η and run in the streamwise direction and the direction normal to the projectile surface respectively. The functionals Ip and Iq used to measure adaptation in the ξ and η directions are

$$I_P = \int \frac{\nabla\xi \cdot \nabla\xi}{P} dxdy \qquad I_Q = \int \frac{\nabla\eta \cdot \nabla\eta}{Q} dxdy \qquad (1)$$

Note that the control functions P and Q must be positive definite and that they have an inverse relationship to the grid resolution. That is, in order to minimize the functionals, the grid point spacing must be large when P or Q is small and consequently the spacing will be small when the control functions are large. This dependence must be considered when forming the functional relationship of the control functions to the flow solution. A third functional Io, used to measure orthogonality, is

$$I_O = \int (\nabla\xi \cdot \nabla\eta)^2 \, dxdy \qquad (2)$$

These three functional are combined to form a total functional I, and a parameter λ is included to weight the perceived importance of orthogonality to adaptation:

$$I_T = \int \left\{ \frac{\nabla\xi \cdot \nabla\xi}{P} + \frac{\nabla\eta \cdot \nabla\eta}{Q} + \lambda(\nabla\xi \cdot \nabla\eta)^2 \right\} dxdy \qquad (3)$$

A set of partial differential equations can be obtained, the solution of which minimizes the total functional, by applying the Euler-Lagrange equations. The resulting set of partial differential equations is

$$\frac{1}{P}(\xi_{xx} + \xi_{yy}) + \frac{\nabla P \cdot \nabla\xi}{P^2} + \lambda(\eta_x^2\xi_{xx} + 2\eta_x\eta_y\xi_{xy} + \eta_y^2\xi_{yy} + (2\xi_x\eta_x + \xi_y\eta_y)\eta_{xx} + (\eta_x\xi_y + \eta_y\xi_x)\eta_{xy} + (\xi_x\eta_x + 2\xi_y\eta_y)\eta_{yy}) = 0$$

$$\frac{1}{Q}(\eta_{xx} + \eta_{yy}) + \frac{\nabla Q \cdot \nabla\eta}{Q^2} + \lambda((2\xi_x\eta_x + \xi_y\eta_y)\xi_{xx} + (\eta_x\xi_y + \eta_y\xi_x)\xi_{xy} + (\xi_x\eta_x + 2\xi_y\eta_y)\xi_{xy} + \xi_x^2\eta_{xx} + 2\xi_x\xi_y\eta_{xy} + \xi_y^2\eta_{yy}) = 0 \qquad (4)$$

It should be noted that there was no explicit functional designated to measure smoothness. In this formulation, the smoothness of the grid point lines along coordinate lines can be controlled in the definition of the control functions. Any variation of the grid point spacing between adjacent coordinate lines will be smoothed by the elliptic operator in eqs. (4). It is also apparent that a large portion of the terms in eqs. (4) result from the functional measuring orthogonality. For purposes of efficiency, therefore, it would be beneficial to consider other means of obtaining orthogonality. One such approach is to seek another functional which leads to fewer terms, or possibly delete the orthogonality functional completely and incorporate another technique to obtain orthogonality such as that of Sorenson[1].

Local Scaling

An ordering analysis of eqs. (4) shows that the first and second terms are of order (C/PL^2) and (C/QL^2) respectively where C is a computational length scale and L is a physical length scale. The third term is of order (C^2/L^2). It is therefore useful to scale each term so that the parameter λ will accurately reflect the desired weighting. Due to the use of highly stretched grids in the current applications, the terms in eqs. (4) may vary by orders of magnitude and one global scale may not be sufficient to properly scale each term throughout the entire domain. It was, therefore, found beneficial to use local scales. The local scales used here were incorporated by replacing λ in the first equation by

$$\lambda = \frac{\lambda' J}{P} \qquad (5)$$

and in the second equation by

$$\lambda = \frac{\lambda' J}{Q} \qquad (6)$$

where J is the Jacobian

$$J = x_\xi y_\eta - x_\eta y_\xi$$

and λ' is a constant that weights orthogonality relative to the other terms. The use of local scaling violates the original variational principal, but it has been shown that the use of local scales will yield more reliable control over the grid network[4]. The adaptive grid generation equations become with the local scales

$$\xi_{xx} + \xi_{yy} + \frac{P_\xi}{P}\lambda\left(\alpha_1^2(x_{\xi\xi}-2\alpha_4 x_{\xi\eta}+\alpha_2^2 x_{\eta\eta}) + (2\xi_x x_\xi+(y_\xi y_\eta)x_{\eta\eta} + (\alpha_4\xi_y+\alpha_4\xi_x)x_{\xi\eta}+((\xi_x x_\eta+2(y_\eta y_\xi)\alpha_2 y_{\eta\eta})\right)=0 \tag{7}$$

$$\eta_{xx} + \eta_{yy} + \frac{Q_\eta}{Q}\lambda\left((2\xi_x x_\eta+(y_\eta y_\xi)(x_{\xi\eta}+(\alpha_4\xi_y+\alpha_4\xi_x)(x_{\xi\eta} + ((\xi_x x_\eta+2(y_\eta y_\xi)(x_{\xi\eta} + (\xi_x^2 x_{\eta\eta}+2(x_\xi y_\eta x_{\xi\eta} + \xi_y^2 y_{\eta\eta})\right)=0$$

In order to solve these equations numerically and obtain the grid network it is neccessary to invert these equations so that x and y become the dependent variables. The adaptive grid generation equations become in the computational domain

$$a_1 x_{\xi\xi} + a_2 x_{\xi\eta} + a_3 x_{\eta\eta} + b_1 y_{\xi\xi} + b_2 y_{\xi\eta} + b_3 y_{\eta\eta} = s_1 \tag{8}$$

$$c_1 x_{\xi\xi} + c_2 x_{\xi\eta} + c_3 y_{\eta\eta} + d_1 y_{\xi\xi} + d_2 y_{\xi\eta} + d_3 y_{\eta\eta} = s_2$$

$$\begin{aligned}
a_1 &= -y_\eta(\alpha+\lambda'\beta^2) - 2\lambda'y_\xi\alpha\beta \\
a_2 &= 2y_\eta(\beta+\lambda'\beta\gamma) + \lambda'y_\xi(\alpha\gamma+\beta^2) \\
a_3 &= -y_\eta(\gamma+\lambda'\gamma^2) - 2\lambda'y_\xi\gamma\beta \\
b_1 &= x_\eta(\alpha+\lambda'\beta^2) + 2\lambda'x_\xi\alpha\beta \\
b_2 &= -2x_\eta(\beta+\lambda'\beta\gamma) - \lambda'x_\xi(\alpha\gamma+\beta^2) \\
b_3 &= x_\eta(\gamma+\lambda'\gamma^2) + 2\lambda'x_\xi\gamma\beta \\
c_1 &= y_\xi(\alpha+\lambda'\alpha^2) + 2\lambda'y_\eta\alpha\beta \\
c_2 &= -2y_\xi(\beta+\lambda'\alpha\beta) - \lambda'y_\eta(\alpha\gamma+\beta^2) \\
c_3 &= y_\xi(\gamma+\lambda'\beta^2) + 2\lambda'y_\eta\gamma\beta \\
d_1 &= -x_\xi(\alpha+\lambda'\alpha^2) - 2\lambda'x_\eta\alpha\beta \\
d_2 &= 2x_\xi(\beta+\lambda'\alpha\beta) + \lambda'x_\eta(\alpha\gamma+\beta^2) \\
d_3 &= -x_\xi(\gamma+\lambda'\beta^2) - 2\lambda'x_\eta\gamma\beta
\end{aligned} \tag{9}$$

$$\begin{aligned}
\alpha &= x_\eta^2 + y_\eta^2 \\
\gamma &= x_\xi^2 + y_\xi^2 \\
\beta &= x_\xi x_\eta - y_\xi y_\eta \\
\lambda' &= \lambda/J^2
\end{aligned} \tag{10}$$

$$s_1 = \frac{P_\xi J\alpha}{P} - \frac{P_\eta J\beta}{P} \tag{11}$$

$$s_2 = \frac{Q_\xi J\beta}{Q} - \frac{Q_\eta J\gamma}{Q}$$

Note that the derivative of P in the η direction appears in the source term S_1 for the equation governing adaptation in the ξ direction. This occurs because the contravariant base vector $\nabla\xi$ indicates the spacing in the direction normal to lines of constant η rather than along ξ coordinate lines. If the grid is orthogonal, these two directions are the same and the term containing P_η vanishes since β is then zero. It has been found in the current study that dropping the term containing P_η, and likewise the term containing Q_ξ in the second source term, has a negligable effect on the resulting grid and consequently these terms are eliminated from the equations.

Boundary Adaptation

It is important to adapt the grid points along the boundaries in a consistant manner to those in the interior. For this purpose a one-dimensional analog to the two-dimensional functional for adaption is used

$$I_s = \int \frac{\xi_s}{P} ds \tag{12}$$

where s measures arc length along the bounding ξ coordinate lines. After applying the Euler-Lagrange equation and inverting to make s the dependent variable, the equation for adaptation along the bounding curvalinear coordinate lines is

$$s_{\xi\xi} + s_\xi P_\xi/P = 0 \tag{13}$$

Similarly, for the bounding η coordinate lines the governing equation for grid point spacing is

$$s_{\eta\eta} + s_\eta Q_\eta/Q = 0 \tag{14}$$

where s now measures the arc length along the η coordinate lines.

Curvature Effects

Another modification to the adaptive grid generation equations found to be useful is the elimination of the effects of curved boundaries. It is well known that in using elliptic partial differential equations, curved boundaries will cause either an attraction or repulsion of coordinate lines. In the current applications, this effect will either work with or against the source terms that control the grid point spacing, either increasing or decreasing the grid spacing dictated by the control functions. In order to eliminate this undesirable result, the follwing 'curvature terms' have been derived

$$\begin{aligned}
K_1 &= \frac{y_\eta x_{\eta\eta} - x_\eta y_{\eta\eta}}{\sqrt{\gamma}\ \alpha^{3/2}} \\
K_2 &= \frac{y_\xi x_{\xi\xi} - x_\xi y_{\xi\xi}}{\sqrt{\alpha}\ \gamma^{3/2}}
\end{aligned} \tag{15}$$

and are added to the equations in the source terms. The source terms now become

$$s_1 = \frac{\alpha J P_\xi}{P} - K_1 J^2 \tag{16}$$

$$s_2 = \frac{\gamma J Q_\eta}{Q} - K_2 J^2$$

The derivation of these curvature terms can be found in reference 8. A more general derivation in three dimension has been done by Thompson[9].

Note that the curvature terms contain second derivatives, but are evaluated here locally during each iteration as part of the source terms.

Iterative Solution Method

The elliptic partial differential equations governing grid adaptation are approximated using second order central difference expressions for the derivative terms which results in a set of coupled nonlinear algebraic equations for the x and y position of each grid point. These difference equations are currently solved using a Newton-Raphson iterative method. However it was found that the convergence of the iterative scheme was extremely slow due to the high aspect ratio cells predominant in the boundary layer region of the flow domain and it was therefore neccessary to modify the solution procedure. The source of the poor convergence of the iterative scheme can be seen by considering a grid cell as shown in figure 1 which, for simplicity, is rectangular and aligned with the cartesian coordinates. It can be shown[8] that the movement of a grid point Δx and Δy for one iteration can be related to the cell aspect ratio AR as

$$\Delta x_{i,j} = \frac{P_\xi}{2P}\left(\frac{1}{1+AR}\right)$$

$$\Delta y_{i,j} = \frac{Q_\eta}{2Q}\left(\frac{AR}{1+AR}\right) \qquad (17)$$

In the grid networks used for viscous transonic flow problems, the cell aspect ratio near the projectile surface can be as large as 10^5. Thus the movement of a grid point along the surface (the x direction in figure 1) during the solution procedure will be extremely inhibited by the small spacing in the y direction, and thus results in an extremely inefficient solution process.

An efficient procedure to circumvent this problem has been developed in reference 8. Basically, it consists of temporarily removing many of the ξ coordinate lines near the projectile surface so the the resulting grid contains cells with aspect ratios closer to order 1. Only enough points remain to adequately represent the η coordinate lines. An example is given in figure 2 in which the dashed lines indicate the ξ coordinate lines that are temporarily removed. The solution to the grid with the points removed is obtained using the Newton-Raphson iterative method and then the removed points are reinserted along the η coordinate lines using the equation for one-dimensional adaptation, eq. (16) to govern their relative spacing.

It is necessary, for a successful implementation of this procedure, to modify the control function Q to account for the spacing required of the grid points that were temporarily removed. A method for properly modifying the control function Q can be obtained based on the one-dimensional equations for adaptation. However, since the two-dimensional equations essentially adapt independently in each coordinate direction (and they default to the one-dimensional equations on a rectilinear grid), this method has been found to be suitable as well for use with the two-dimensional adaptive grid generation equations.

The procedure to obtain the modified control function Q*, which is derived in detail in reference 8, is to first form the function C at each point along a η coordinate line

$$C_j = \frac{Q_j + \left(\frac{Q_{j+1}-Q_{j-1}}{2}\right)}{Q_j - \left(\frac{Q_{j+1}-Q_{j-1}}{2}\right)} \qquad (18)$$

For each jth grid point removed, the function C* at the adjacent j-1 position is calculated as

$$\frac{C_j C_{j-1}}{1+C_j} = C^*_{j-1} \qquad (19)$$

and then the modified control function at the j-1 position is

$$Q^*_{j-1} = \left(\frac{1+C^*_{j-1}}{1-C^*_{j-1}}\right)(Q_{j+1}-Q_{j-2}) \qquad (20)$$

Thus as each grid point is removed along the η coordinate line the modified control function at a remaining adjacent point is updated. In regions of the grid in which no grid points are removed, the modified control function Q* is equal to the original control function Q. Note that the modified control function is used only when solving the two-dimensional grid generation equations with the reduced number of grid points. When inserting the removed points along the η coordinate lines, the original control function Q must be used.

This modified solution procedure has been implemented to eliminate the poor convergence due to the large aspect ratio cells predominate in the boundary layer regions. The method essentially results in the same grid network that would have been obtained by directly applying an iterative scheme after a large number of iterations. Furthermore, the computational time required to remove points, modify the control function and reinsert the points is more than compensated by the reduction of calculations in the iterative solution process since many of the grid points are removed (almost 50% in the current applications).

The Control Functions

The control functions provide the link between the flow solution and the grid adaptation and thus indicate the regions in the physical domain for which increased resolution is required. The general form of the control function used here is the same as that used by Nakahashi and Deiwert[9]

$$W = 1 + \gamma h^\sigma \qquad (21)$$

where W indicates either P or Q, γ and σ are parameters that determine the minimum and maximum spacing along a coordinate line and h is a derivative of a flow variable scaled to range between 0 and 1. For instance, if h were chosen as the pressure gradient, the control function would be large when the pressure gradient is large and, as a consequence of the relationship

between the control function and the grid point spacing defined in eqs. (1), the grid point spacing would become small.

The two parameters γ and σ can be used to specify the minimum and maximum grid point spacing along each coordinate line. Following the work of Nakahashi and Diewert, this is done in the following manner. The ratio of the minimum and maximum grid point spacing along any one coordinate line can be shown to be related to γ as

$$\frac{(\Delta s_j)_{max}}{(\Delta s_j)_{min}} = 1 + \gamma \qquad (22)$$

and thus this ratio can be specified by specifying γ. The influence of γ can be determined by first integrating either eq. (13) or eq. (14) once, and then using a discrete representation of the integral. For the ith point the grid point spacing Δs_i is

$$\Delta s_i = \frac{S_t \frac{1}{1+\sigma h_i}}{\sum_j^N \frac{1}{1+\sigma h_j}} \qquad (23)$$

where S_t is the total length of the coordinate line and N is the number of grid points along the line. Evaluation of this integral for the minimum spacing (i.e. $h = 1$) results in the following expression

$$(\Delta s_j)_{min} \sum_j (\frac{1}{1+\sigma h_j^2}) - \frac{S_t}{1+\sigma} = 0 \qquad (24)$$

Thus, by prescribing both γ and the minimum grid point spacing, σ can be determined from eq. (24) using a root finding technique. This procedure is currently being used to specify the minimum and maximum grid point spacing along each coordinate line. Again, the method is developed using the equation for one-dimensional adaptation, but is employed in the two-dimensional adaptive grid generation procedure. In the current applications, the specified spacings have been obtained in the two-dimensional grid networks to within 5% of the specified values. The variation from the prescribed values can also be attributed to the effects of minimizing the orthogonality functional.

Solution-Adaptive Computational Method

The adaptive grid generation technique can be incorporated naturally into the solution process of the flow solver since the flow solver algorithm is time-marching. After a specified number of time steps, the grid network is adapted using control functions based on the current values of the flow solution. Thus the time-marching algorithm of the TVD code used here proceeds as usual with intermittent calls to the adaptive grid generation algorithm to update the grid network. In this approach, however, it is necessary to account for the grid point motion due to the grid adaptation. In the current applications to steady flow problems, the grid network was updated every 200 timesteps. After each grid adaptation, the values of the flow variables are interpolated from the previous grid network to the adapted grid network using a linear interpolation in two dimensions.

The TVD scheme developed by Yee[10] for the multidimensional Navier-Stokes equations is used here as the flow solver. It was obtained by modifying an existing thin layer Navier-Stokes code based on the Beam and Warming scheme[11] that was developed by Steger and Pulliam[12] and later modified by Nietubicz[13] for the efficient calculation of axisymmetric flow solutions. The TVD code employs the thin layer approximation and uses the eddy viscosity turbulence model of Baldwin and Lomax[14].

The computational technique then proceeds by starting the calculations of the TVD scheme on an initial grid network. The initial grid network used here was obtained using the technique of Steger et al.[15]. After each interval of 200 time steps, the grid network is updated using the adaptive grid generation method which is comprised of calculating the control functions, removing grid points in high resolution areas and modifying the control functions, solving the two-dimensional grid generation equations, reinserting the grid points and finally interpolating the flow variables onto the adapted grid network.

This process is continued until the solution of the flow variables converges to a steady-state solution. Consequently the grid network will also cease to change since its configuration is linked to the flow solution through the control functions. For the cases considered here, the solution converged between 1500 and 2000 time steps of 0.1. In each case, the calculations were continued to 2000 time steps.

Results

The self-adaptive computational technique just described has been applied to calculate the axisymmetric flow over a 4 caliber secant-ogive cylinder boattail projectile which is shown in figure 3. The Reynolds number is 76,000 and three Mach numbers 0.91, 0.96 and 1.1 are considered, and are cases for which some experimental data is available for comparison[16].

The initial grid network is shown in figure 4a. Figure 4b shows the grid network near the projectile surface. The boattail is extended downstream and meets a horizontal sting which is used to eliminate the base region from the flow domain. The grid network, which contains 70 points in the streamwise direction and 50 points in the direction normal to the projectile surface, extends out from the projectile four times the projectile length and extends upstream three times the projectile length. Note that most of the points in the normal direction are clustered near the projectile surface. Points along the streamwise coordinates are clustered near the ogive cylinder and cylinder boattail junctures. This grid is typical of those used in viscous transonic projectile flow calculations.

The solution is expected to contain shocks and expansions normal to the projectile surface and a boundary shear layer parallel to the projectile surface. These structures require increased grid resolution and thus guide the choices for the function h in eqs. (21) for the control functions. In the streamwise direction, the grid is adapted to the second derivative of the pressure distribution. Originally, the first

derivative was used, but it was found that too many points clustered in the shock regions at the expense of clustering in the expansion regions. Experience has shown the the pressure expansions require more resolution than the shocks and the control function based on the second derivative of the pressure yielded such a distribution. The specified minimum and maximum grid point spacing for use in the adaptive grid generation algorithm are 0.02 and .25 respectively. In the normal direction, the control function is based on the exponential of the velocity gradient. This function typically resulted in a grid point distribution similar to the exponential clustering function which is known to yield good results. The specified minimum spacing is 2×10^{-4} and the specified maximum spacing is 3. In all cases, the parameter λ' which weights grid orthogonality is 0.5.

The first flow case considered is Mach 0.96. Figure 5 shows a comparison of the computed pressure coefficient to experimental data and the pressure contour plot of figure 6 reveals the structure of the pressure field in the flow domain. The adapted grid network corresponding to this converged solution is shown in figure 7. A comparison of the contour plot and the adapted grid network clearly indicates the expected adaptation to the pressure field. Grid point clustering is evident in both shock regions and in the pressure expansion at the two junctures and it appears that the resolution is greatest in the expansion regions. Note that the clustering in the shocks near the surface is not as strong as that farther from the projectile surface. This is due both to the competition for clustering with the expansion regions and the smearing of the pressure jump in the boundary layer. Adaptation of the grid network to the velocity gradient is also clearly evident as most of the points along the η coordinate lines are clustered near the projectile surface. These results clearly indicate that the adaptive grid generation equations can essentially adapt independently in each coordinate direction.

Two more cases were considered, Mach 0.91 and 1.1, using the same parameters to control the grid adaptation. Figures 8a and 8b show the computed pressure coefficient and corresponding adapted grid network for Mach 0.91 and figures 9a and 9b show the same for Mach 1.1. At Mach 0.91 the two shocks appear further upstream, directly behind the pressure expansions. The pressure field is again clearly indicated in the adaptation of the grid network. At Mach 1.1, no shocks occur over the projectile surface, a result again reflected by the grid adaptation. The adaptive grid generation scheme is thus quite reliable in providing good adapted grid networks provided proper choices are made for the control tions.

Comments

The adaptive grid generation technique, derived from a variational approach, has been shown to adapt independently in each coordinate direction and provide the extremely refined mesh necessary to accurately resolve the boundary layer regions. Modifications to the grid generation technique, including local scaling of the equations, the elimination of the effects of curved boundaries, and use of the modified control function have been found helpful in providing an efficient, reliable adaptive grid generation technique. In applications to viscous transonic projectile problems, the adaptive grid generation technique has been shown, with the proper choice of control functions, to provide good adapted grid networks. The characteristics of the adaptive grid generation equations demonstrated here should be useful in adapting grid networks in complex geometries and in three dimensions.

Aknowledgements

The research presented here was sponsored by a USAF-UES grant. The computations were completed on a Cray XMP computer at the Pittsburg Supercomputing Center. The computer time was furnished by NSF and the Pittsburg Supercomputer Center.

References

1. Saltzman, J. and Brackbill, J.U., "Applications and Generalization of Variational Methods for Generating Adaptive Meshes," Numerical Grid Generation, Ed. J.F. Thompson, North-Holland, pp. 865-884, 1981.

2. Hsu, C.C. and Tu, C.G., "An Adaptive Grid Generation Technique Based on Variational Principals for Transonic Aerodynamic Calculation," To Appear in Int. J. Numerical Methods in Fluids.

3. Nakahashi, K. and Deivert, G.S., "A Self -Adaptive Grid Method with Applications to Airfoil Flow," AIAA Paper 85-1525.

4. Mastin, C.W., "Error Induced By Coordinate Systems," Numerical Grid Generation, Ed. J.F Thompson, North-Holland, p. 41, 1981.

5. Dywer, H.A., Smooke, M.D. and Kee, R.J., "Adaptive Gridding for Finite Difference Solutions to Heat and Mass Transfer Problems," Numerical Grid Generation, Ed. J.F. Thompson, North-Holland, p. 339, 1981.

6. Gnoffo, P.A. "A Vectorized Finite Volume, Adaptive-Grid Algorithm for Navier-Stokes Calculation," Numerical Grid Generation, Ed. J.F. Thompson, North-Holland, p. 819, 1981.

7. Sorenson, R.L., "A Computer Program to generate Two-Dimensional Grids About Airfoils and Other Shapes by the Use of Poisson's Equations," NASA TM 81198, 1980.

8. Reed, C.W. "A Self-Adaptive Computational [illegible] for Transonic Turbulent Projectile Aerodynamics" Ph.D. Dissertation, University of Florida, 1987.

9. Thompson, J.F. "Program EAGLE Numerical grid Generation System User's Manual," AFATL-TR -87-15, Vol. III, 1987.

10. Yee, H.C., "Linearized form of Implicit TVD Schemes for the Multidimensional Euler and Navier-Stokes Equations," Comp. & Math. with Appls., Vol. 12a, Nos. 4/5, pp. 413-432, 1986.

11. Warming, R.F. and Beam, R., "On the Conservation and Application of Implicit Factored Schemes for Conservation Laws," SIAM-AMS Proceedings, Vol. 2, pp. 85-99, 1978.

12. Pulliam, T.H., and Steger, J.L., "Implicit Finite-Difference Simulations of Three -Dimensional Compressible Flow," AIAA Journal, Vol. 18, No. 2, pp. 167-169, 1980.

13. Nietubicz, C.J., Pulliam, T.H. and Steger, J.L., "Numerical Solutions of the Azimuthal Invariant Thin layer Navier-Stokes Equations," AIAA Paper 79-0010.

14. Baldwin, B.S. and Lomax, H., "Thin Layer Approximation and Algebraic Model for Separated Turbulent Flows," AIAA Paper 78 -257.

15. Nietubicz, C.J., Heavey, K.R. and Steger, J.L., "Grid Generation Techniques for Projectile Configurations," Proc. 1982 Army Numerical Analysis and Computers Conference, ARO Rept. 82-3, February 1982.

16. Weyer, L.D. and Whiten, F. "Surface Pressure Measurements in a Boattailed Projectile Shape at Transonic Speeds," ARBRL-MR-03161, U.S. Ar / Ballistic Research Laboratory, March, 1982.

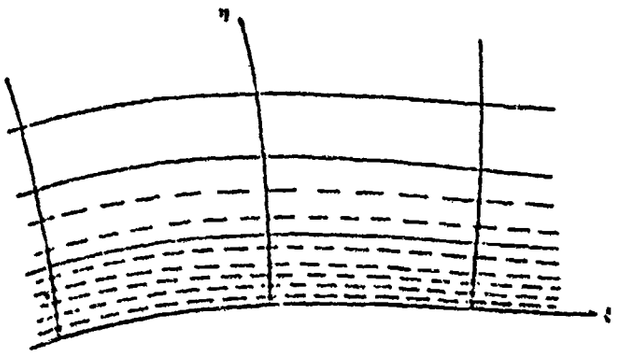


Figure 2. Temporary removal of grid points

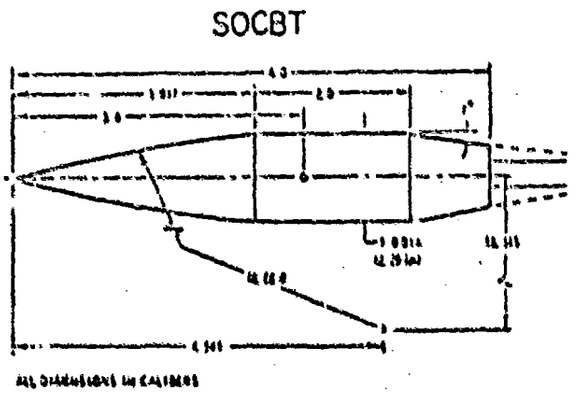


Figure 3. SOCBT projectile

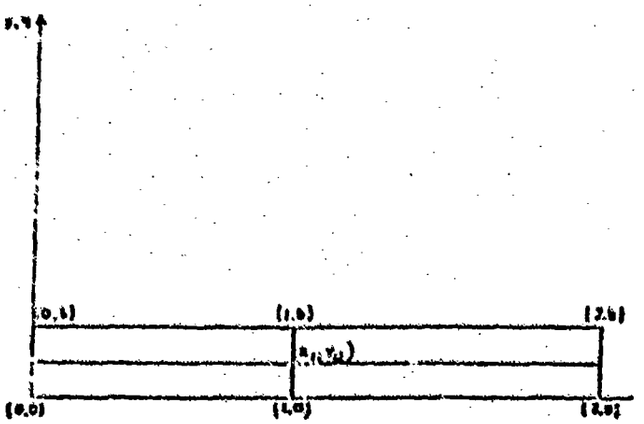


Figure 1. High aspect ratio cell

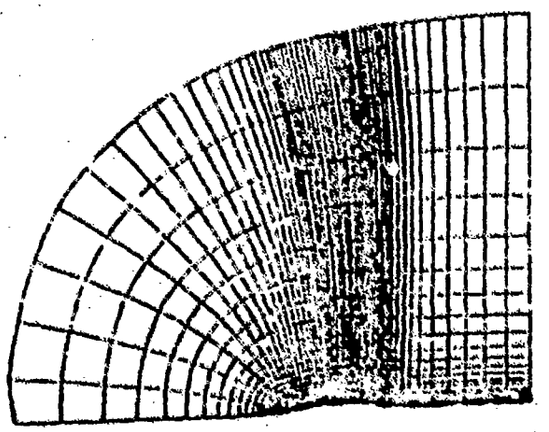


Figure 4a. Initial grid network

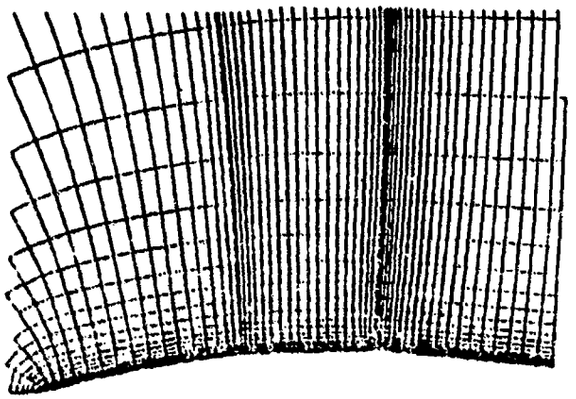


Figure 4b. Initial grid network near the projectile

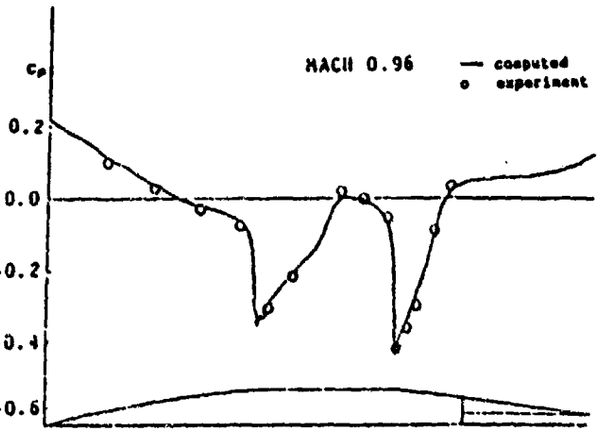


Figure 5. Computed surface pressure coefficient

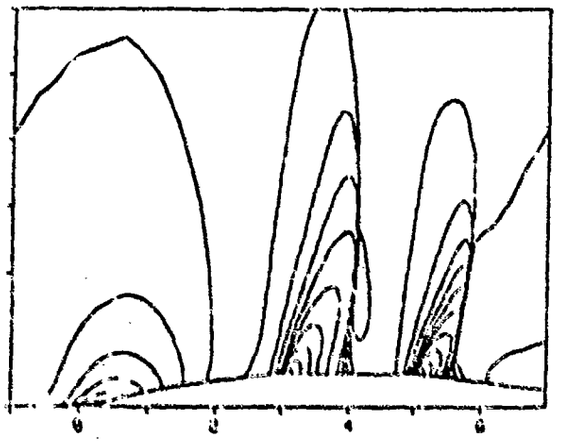


Figure 6. Pressure contours for Mach 0.96

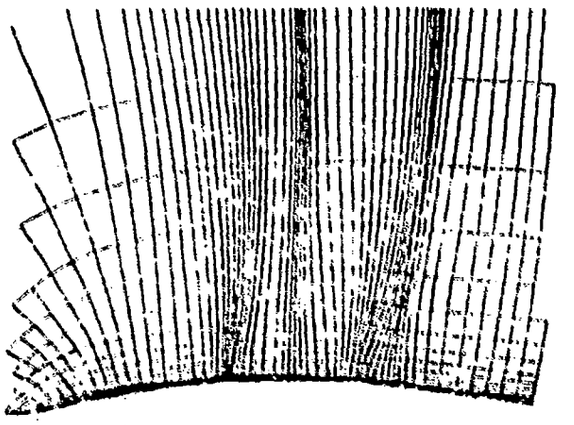


Figure 7. Adapted grid network for Mach 0.96

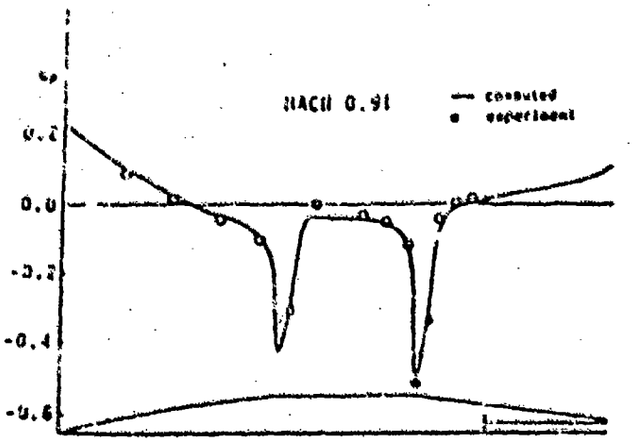


Figure 8a. Computed surface pressure coefficient

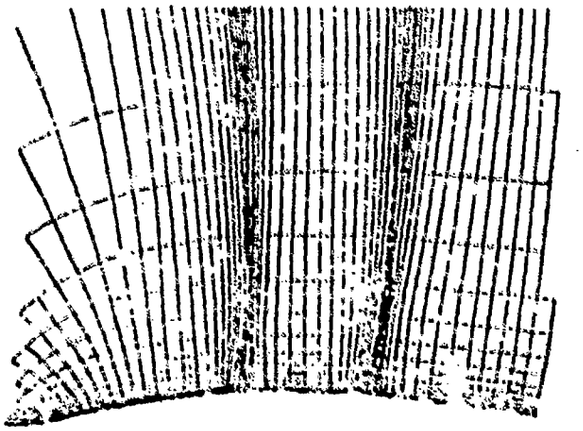


Figure 8b. Adapted grid network for Mach 0.91

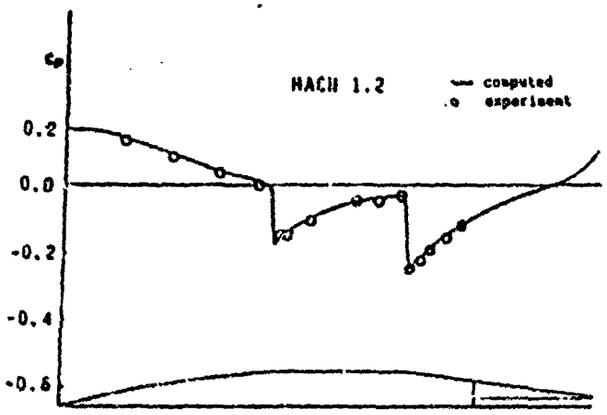


Figure 9a. Computed surface pressure coefficient

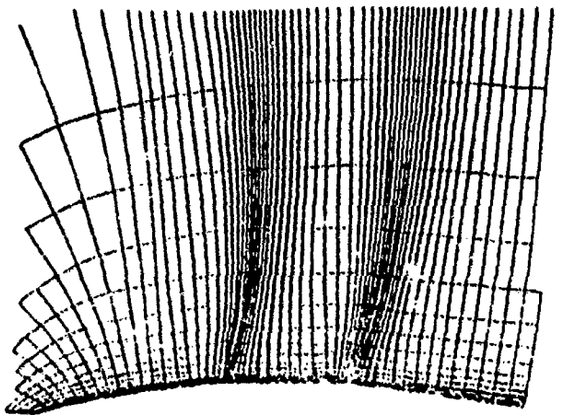


Figure 9b. Adapted grid network for Mach 1.2