LOAN DOCUMENT

PHOTOGRAPH THIS SHEET

AD-A261 991

DTIC ACCESSION NUMBER

LEVEL

1

INVENTORY

AFOSR-TR-93-0113

DOCUMENT IDENTIFICATION

Dec 92

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

DISTRIBUTION STATEMENT

HANDLE WITH CARE

ACCESSION FOR

NTIS	GRA&I	☑
DTIC	TRAC	☐
UNANNOUNCED		☐

JUSTIFICATION

BY

DISTRIBUTION/

AVAILABILITY CODES

DISTRIBUTION	AVAILABILITY AND/OR SPECIAL
A-1	

DISTRIBUTION STAMP

DTIC
ELECTE
MAR 1 0 1993
S C D

DATE ACCESSIONED

INSPECTED 1

DATE RETURNED

93 3 10 006

~~93 3 4 075~~

DATE RECEIVED IN DTIC

93-05124

REGISTERED OR CERTIFIED NUMBER

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-FDAC

DTIC FORM 70A JUN 90

DOCUMENT PROCESSING SHEET

PREVIOUS EDITIONS MAY BE USED UNTIL STOCK IS EXHAUSTED.

LOAN DOCUMENT

UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1992

SUMMER FACULTY RESEARCH PROGRAM (SFRP) REPORTS

VOLUME 4

ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 UPLANDER WAY

CULVER CITY, CA 90230-6608

SUBMITTED TO:

LT. COL. CLAUDE CAVENDER
PROGRAM MANAGER

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

BOLLING AIR FORCE BASE

WASHINGTON, D.C.

DECEMBER 1992

REPORT DOCUMENTATION PAGE

AGENCY USE ONLY	REPORT DATE	REPORT TYPE AND DATES COVERED
	28 Dec 92	Annual 1 Sep 91 - 31 Aug 92

TITLE AND SUBTITLE

V4

1992 Summer Faculty Research Program (SFRP)
Volumes 1 - 16

F49620-90-C-0076

AUTHOR(S)

Mr Gary Moore

PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Research & Development Laboratoreis (RDL)
5800 Uplander Way
Culver City CA 90230-6600

SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NI
110 Duncan Ave., Suite B115
Bldg 410
Bolling AFB DC 20332-0001
Lt Col Claude Cavender

SUPPLEMENTARY NOTES

DISTRIBUTION AVAILABILITY STATEMENT

UNLIMITED

ABSTRACT (Maximum 200 words)

The purpose of this program is to develop the basis for cintinuing research of interest to the Air Force at the institution of the faculty member; to stiumlate continuing relations among faculty members and professional peers in the Air Force to enhance the research interests and capabilities of scientific and engineering educators; and to provide follow-on funding for research of particular promise that was started at an Air Force laboratory under the Summer Faculty Research Program.

During the summer of 1992 185 university faculty conducted research at Air Force laboratories for a period of 10 weeks. Each participant provided a report of their research, and these reports are consolidated into this annual report.

SUBJECT TERMS

SECURITY CLASSIFICATION OF REPORT	SECURITY CLASSIFICATION OF THIS PAGE	SECURITY CLASSIFICATION OF ABSTRACT	
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Standard Form 298 (Rev. 2-89)

UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1992

SUMMER FACULTY RESEARCH PROGRAM (SFRP) REPORTS

VOLUME 4

ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES
5800 Uplander Way
Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Lt. Col. Claude Cavender

Program Manager, RDL
Billy Kelley

Program Administrator, RDL
Gwendolyn Smith

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling Air Force Base
Washington, D.C.
December 1992

PREFACE

This volume is part of a 16-volume set that summarizes the research accomplishments of faculty, graduate student, and high school participants in the 1992 Air Force Office of Scientific Research (AFOSR) Summer Research Program. The current volume, Volume 4 of 16, presents the final research reports of faculty (SFRP) participants at Rome Laboratory.

Reports presented herein are arranged alphabetically by author and are numbered consecutively -- e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

Research reports in the 16-volume set are organized as follows:

VOLUME	TITLE
1	Program Management Report
2	Summer Faculty Research Program Reports: Armstrong Laboratory
3	Summer Faculty Research Program Reports: Phillips Laboratory
4	Summer Faculty Research Program Reports: Rome Laboratory
5A	Summer Faculty Research Program Reports: Wright Laboratory (part one)
5B	Summer Faculty Research Program Reports: Wright Laboratory (part two)
6	Summer Faculty Research Program Reports: Arnold Engineering Development Center; Civil Engineering Laboratory; Frank J. Seiler Research Laboratory; Wilford Hall Medical Center
7	Graduate Student Research Program Reports: Armstrong Laboratory
8	Graduate Student Research Program Reports: Phillips Laboratory
9	Graduate Student Research Program Reports: Rome Laboratory
10	Graduate Student Research Program Reports: Wright Laboratory
11	Graduate Student Research Program Reports: Arnold Engineering Development Center; Civil Engineering Laboratory; Frank J. Seiler Research Laboratory; Wilford Hall Medical Center
12	High School Apprenticeship Program Reports: Armstrong Laboratory
13	High School Apprenticeship Program Reports: Phillips Laboratory
14	High School Apprenticeship Program Reports: Rome Laboratory
15	High School Apprenticeship Program Reports: Wright Laboratory
16	High School Apprenticeship Program Reports: Arnold Engineering Development Center; Civil Engineering Laboratory

1992 FACULTY RESEARCH REPORTS

Rome Laboratory

Report Number	Report Title	Author
1	Toward the Development of a Generalized Method and Code for Analyzing Infinite Arrays of Antennas Printed on Both Sides of Protruding Dielectric Substrates	Dr. Jean-Pierre R. Bayard
2	Statistical Comparison of Several Automatic Target Recognition (ATR) Systems	Dr. Pinyuen Chen
3	Photonics Technology Development at Rome Laboratory	Dr. Richard L. Fork
4	Issues in Adaptive Fault Management for Surveillance C^3I Systems	Dr. Rex E. Gantenbein
5	Atomistic Simulation of Grains in Submicron Aluminum Interconnects	Dr. Surendra K. Gupta
6	Wideband ATM Networks with Adaptive Routine for the Dynamic Theater Environment	Dr. Robert R. Henry
7	Multipath Channel Equalization for Spread Spectrum Communication System	Dr. H. K. Hwang
8	An Investigation of the Benchmark Evaluation Tool	Dr. Khosrow Kaikhah
9	Measurement of Thermophysical Properties of Semiconductors at High Temperature	Dr. Joseph B. Milstein
10	Photonic Delay Line for High-Frequency Radar Systems	Dr. Evelyn H. Monsay
11	User-Based Requirements for Large-Scale Distributed Information Management Systems: Representation for System Designers	Dr. Michael S. Nilan
12	Flux Creep in a Y-Ba-Cu-O Film Characterized by a C-Axis Microstructure Imbedded with A-Axis Oriented Grains	Dr. John L. Orehotsky
13	(Report not received)	
14	Toward Implementation of a Certification Framework for Reusable Software Modules	Dr. Allen S. Parrish
15	Data Association Problems in Multisensor Data Fusion and Multitarget Tracking	Dr. Aubrey B. Poore
16	Thermal Characterization of In-Situ Synthesis for LEC/MLEK Growth of InP Single Crystals	Dr. Vishwanath Prasad
17	(Report not received)	
18	FDTD Analysis of a Novel Anechoic Chamber Absorbing Boundary Condition for EM Scattering Simulation	Dr. Carey M. Rappaport

Rome Laboratory (cont'd)

Report Number	Report Title	Author
19	X-Band T/R Module Conducted Interference Simulation and Measurement	Dr. John P. Rohrbaugh
20	Monte Carlo Validation of a Theoretical Model for the Generation of Non-Gaussian Radar Clutter	Dr. Jorge Luis Romeu
21	Hierarchical and Integrated Modeling and Simulation	Dr. Robert G. Sargent
22	(Report not received)	
23	Metamodel Applications using TERSM	Dr. Jeffrey D. Tew
24	Fractal Image Compression Techniques	Dr. Guttalu R. Viswanath
25	Prominence in Spontaneous Speech: Annotation and Applications	Dr. Colin W. Wightman
26	The Case for Like-Sensor Pre-Detection Fusion	Dr. Peter Willett

TOWARD THE DEVELOPMENT OF A GENERALIZED METHOD AND CODE FOR ANALYZING INFINITE ARRAYS OF ANTENNAS PRINTED ON BOTH SIDES OF PROTRUDING DIELECTRIC SUBSTRATES

Jean-Pierre R. Bayard, Ph.D.
Associate Professor
Department of Electrical & Electronic Engineering

California State University, Sacramento
6000 J Street
Sacramento, CA 95819-6019

Final Report for:
Summer Research Program
Hanscom Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

September 1992

TOWARD THE DEVELOPMENT OF A GENERALIZED METHOD AND CODE FOR ANALYZING INFINITE ARRAYS OF ANTENNAS PRINTED ON BOTH SIDES OF PROTRUDING DIELECTRIC SUBSTRATES

Jean-Pierre R. Bayard
Associate Professor
Department of Electrical & Electronic Engineering
California State University, Sacramento

Abstract

A method for analyzing infinite arrays of antennas printed on both sides of protruding substrates and covered with a dielectric radome is described. By using the equivalence principle, the array unit cell is decomposed into homogeneous regions where the fields are expressed as Floquet summations, and an inhomogeneous cavity region where the fields can be found by a combination of the method of moments and modal analysis. The approach is rigorous in the sense that the combined effects of the radiating element and feed geometry printed on both sides of a protruding substrate are accounted for. It is general, capable of modeling any antenna elements with currents that are perpendicular and parallel to the ground plane. In addition, both the radiating and scattering/receiving modes of operation are treated. The method is used to calculate the active element impedance of an infinite array of dipoles transmission line - coupled to microstrip feeds. Examples of numerical results are presented for various scan conditions and the effects of a near-field dielectric radome are demonstrated.

TOWARD THE DEVELOPMENT OF A GENERALIZED METHOD AND CODE FOR ANALYZING INFINITE ARRAYS OF ANTENNAS PRINTED ON BOTH SIDES OF PROTRUDING DIELECTRIC SUBSTRATES

Jean-Pierre R. Bayard

INTRODUCTION

The need for antennas that can be easily integrated with MMIC modules has drawn a lot of attention to radiating elements that are printed on finite-height dielectric substrates. The geometry which usually is comprised of a dielectric wall protruding a finite height from a ground plane offers the advantage of removing the feed and its spurious effects from the radiating half space. Phased arrays of these elements have been investigated via the infinite array model by many workers [1-6]. In [1-4], the authors have presented analyses and results which ignored either the presence of the dielectric permittivity [1,2], or the radiating effects of the feedlines [3,4]. Both approximations are consequential given that surface wave/blindness phenomena caused by the dielectric substrate or by the feedlines can not be readily identified by these methods.

In [5] and [6], analysis and results for infinite phased arrays of dipoles with and without coplanar feedlines were presented. The method based on the equivalence principle and the method of moments accurately models the presence of the dielectric as well as any electric currents on one side of the protruding substrate. The present effort seeks to extend this method as to allow the treatment of electric currents on both sides of the protruding substrate, the presence of a near-field dielectric radome layer, and the solution to the scattering/receiving problem. This extension requires that the following steps be added to the formulation: 1) Determining Green's functions for current elements parallel and perpendicular to the ground plane on both sides of the substrate, 2) modifying the free-space fields to accommodate a dielectric layer, 3) enforcing the boundary conditions on the additional conductors and at the boundaries of the radome, and 4) modifying the source vector to include a plane wave illumination. The result is a numerical approach which is of extensive applicability, and can be used to model feed arrangements that are more practical than, for example, the center-fed delta-gap generator model used in [5] and [6]. The applicability is, however, limited since in its present form the numerical procedure is not suited for handling tapered currents unless the computer-intensive step approximation model is used.

This approach is outlined in the next section and the reader is referred to [5] for additional details of the analysis. Then the method is applied to a dipole element with straight arms fed by coplanar transmission lines electromagnetically coupled to a microstrip line. Input impedance values are calculated for various scan angles and dielectric parameters for the array with and without the radome cover. Some previously published data are duplicated for verification purposes.

SUMMARY OF THE ANALYSIS

The schematic of the infinite array of dipoles covered with a dielectric radome layer is shown in Figure 1a, and the dipole parameters are defined in Figure 1b. The part of the element that is closer to the ground plane serves as a ground for the microstrip feeder line (dotted line) printed on the other side of the protruding substrate. This part is followed by the coplanar transmission lines typically with a narrow spacing, and connected to the printed dipole arms.

The first step of the analysis employs the equivalence principle at the planes z=d and z=d+l (see Figure 2). This results in three distinct regions coupled via the equivalent electric and magnetic current sources. Region I ($0 \le z \le d$) is an inhomogeneously-filled region with the electric conductor currents present and equivalent sources at z=d. Region II ($d \le z \le d+l$) is an homogeneous dielectric region with equivalent sources on its boundaries. Region III ($z \ge d$) is a free-space region with equivalent sources at z=d. The contributions of all equivalent electric current sources are then removed by inserting perfect conductors at z=d and z=d+l. A schematic of the resulting unit cell model is shown in Figure 2. The remaining part of the formulation deals with finding the fields in each of these 3 regions and matching the boundary conditions to solve for the following unknowns: 1) The feedlines and dipole currents, and 2) the magnetic current sources/responses introduced in the equivalent problem.

REGION I

The fields in region I are generated by the y and z-directed electric currents existing on both sides of the substrate, and by the equivalent magnetic current under the metallized aperture at z=d. We solve for the contributions of the electric and magnetic currents individually, and by superposition, the total response is the sum of the responses due to each type of current. The fields produced by the electric currents are found via a Green's function approach, whereas those produced by the equivalent magnetic current are expressed using modal analysis.

First consider the electric current problem. In order to find the Green's functions for the electric currents in the cavity region, consider a y or z-directed infinitesimal dipole printed at $(0,y_o,z_o)$ or at (t,y_o,z_o). Using the TM_x and TE_x mode expansions, the magnetic and electric vector potentials in the dielectric region are given respectively by

$$\vec{A}^{\,r} = \hat{x} \sum_{n=-\infty}^{+\infty} \sum_{m=1}^{\infty} \sin\left(\frac{m\pi z}{d}\right) e^{-jV_n y} \{A_{mn}^{r} e^{-j\beta_{mn} x} + B_{mn}^{r} e^{j\beta_{mn} x}\},$$

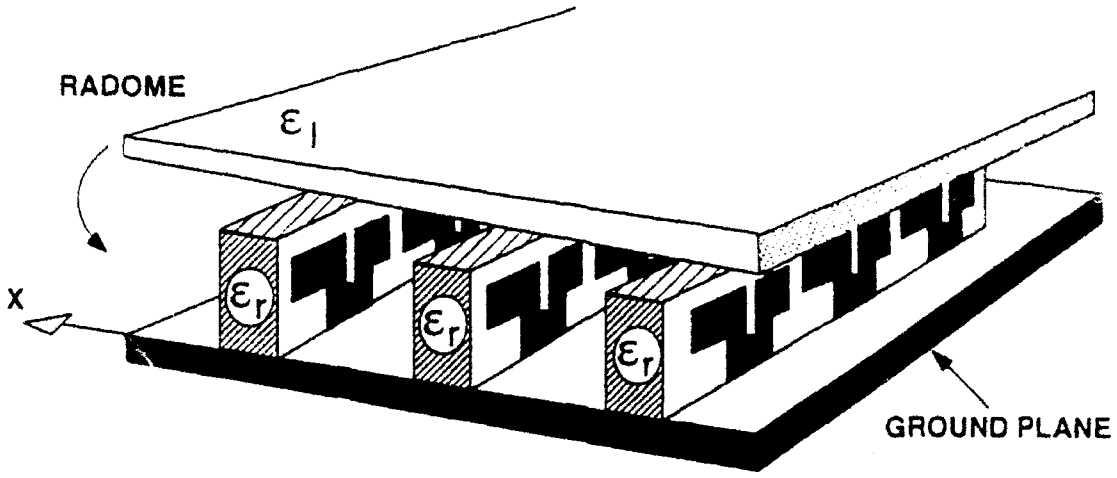
$$\vec{F}^{\,r} = \hat{x} \sum_{n=-\infty}^{+\infty} \sum_{m=0}^{\infty} \cos\left(\frac{m\pi z}{d}\right) e^{-jV_n y} \{C_{mn}^{r} e^{-j\beta_{mn} x} + D_{mn}^{r} e^{j\beta_{mn} x}\}$$

$$\text{where} \quad \beta_{mn}^2 = \epsilon_r k_o^2 - V_n^2 - \left(\frac{m\pi}{d}\right)^2, \qquad (1)$$

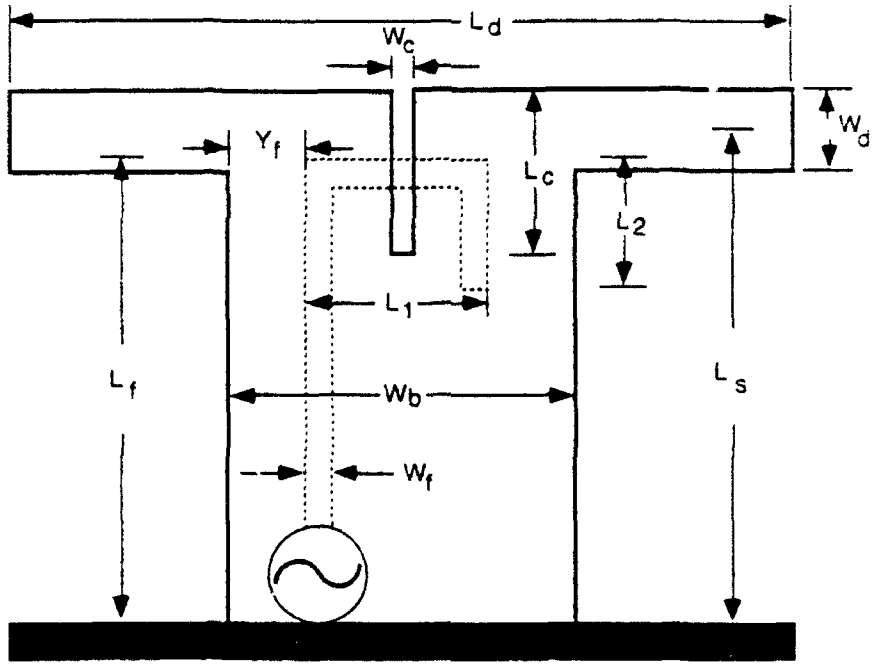
$$V_n = k_o \sin\theta_o \sin\phi_o + \frac{n2\pi}{b},$$

$$k_o = \omega\sqrt{\mu_o \epsilon_o},$$

and θ_o and ϕ_o are the array beam steering angles. In the free-space region, the potential functions can be written straightforwardly from equation (1) by considering modal coefficients, say, A^{o}_{mn}, B^{o}_{mn}, C^{o}_{mn}, D^{o}_{mn},



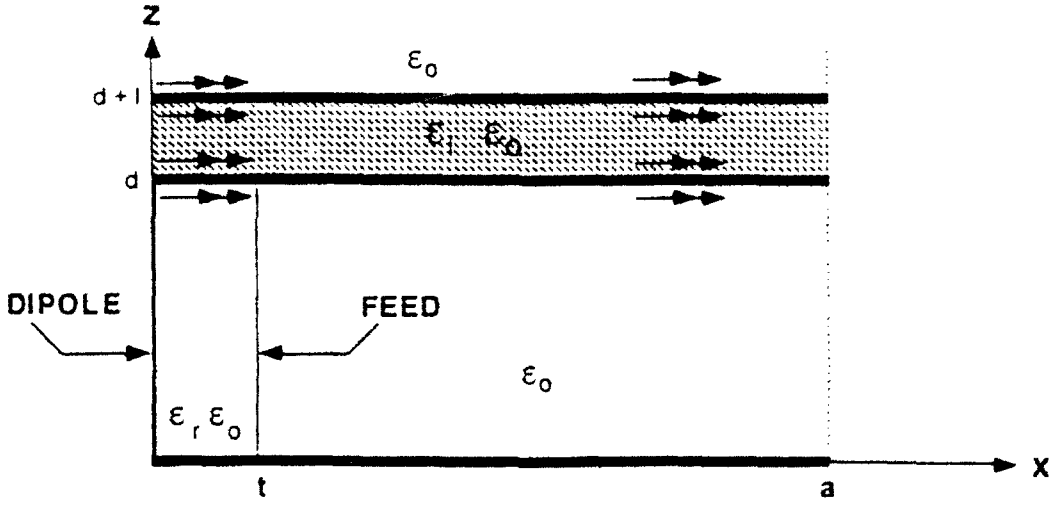
(a) Array geometry



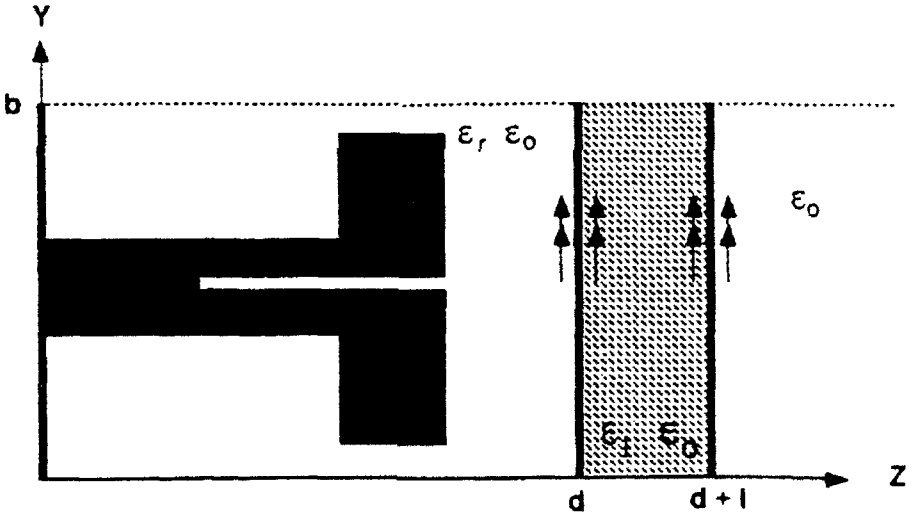
GROUND PLANE

(b) Dipole element

Figure 1. The geometry of the array of dipoles fed by coplanar transmission lines coupled to microstrip feedlines.



(a) z - x plane



(b) y - z plane

Figure 2. The equivalent unit cell with magnetic sources at equivalent planes.

and by letting

$$\beta_{mn}^2=\alpha_{mn}^2=k_o^2-V_n^2-\left(\frac{m\pi}{d}\right)^2. \qquad (2)$$

The Green's function solution is found when the unknown coefficients are determined. This is accomplished by enforcing boundary conditions on the tangential electric and magnetic field components at x=0 and x=t, as well as the conditions of periodicity in x on the solution. These conditions are represented by the following equations which form an 8x8 matrix:

Continuity of E_y at x=t

$$\begin{aligned} &A_{mn}^{x}\left[-\frac{\beta_{mn}V_n}{j\omega\epsilon_o\epsilon_x}e^{-j\beta_{mn}t}\right]+B_{mn}^{x}\left[\frac{\beta_{mn}V_n}{j\omega\epsilon_o\epsilon_x}e^{j\beta_{mn}t}\right]+C_{mn}^{x}\left[\frac{m\pi}{d}e^{-j\beta_{mn}t}\right]+D_{mn}^{x}\left[\frac{m\pi}{d}e^{j\beta_{mn}t}\right] \\ &A_{mn}^{o}\left[\frac{\alpha_{mn}V_n}{j\omega\epsilon_o}e^{-j\alpha_{mn}t}\right]+B_{mn}^{o}\left[-\frac{\alpha_{mn}V_n}{j\omega\epsilon_o}e^{j\alpha_{mn}t}\right]+C_{mn}^{o}\left[-\frac{m\pi}{d}e^{-j\alpha_{mn}t}\right]+D_{mn}^{o}\left[-\frac{m\pi}{d}e^{j\alpha_{mn}t}\right]=0; \end{aligned} \qquad (3)$$

jump discontinuity on H_z and periodicity conditions at x=0

$$\begin{aligned} &A_{mn}^{x}[-jV_n]+B_{mn}^{x}[-jV_n]+C_{mn}^{x}\left[-\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\mu}\right]+D_{mn}^{x}\left[\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\mu}\right]+A_{mn}^{o}\left[jV_ne^{-j\alpha_{mn}a}e^{jU_oa}\right]+ \\ &B_{mn}^{o}\left[jV_ne^{j\alpha_{mn}a}e^{jU_oa}\right]+C_{mn}^{o}\left[\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\mu}e^{-j\alpha_{mn}a}e^{jU_oa}\right]+D_{mn}^{o}\left[-\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\mu}e^{j\alpha_{mn}a}e^{jU_oa}\right]=R^{yo}; \end{aligned} \qquad (4)$$

continuity of E_y and periodicity conditions at x=0

$$\begin{aligned} &A_{mn}^{x}\left[-\frac{\beta_{mn}V_n}{j\omega\epsilon_o\epsilon_x}\right]+B_{mn}^{x}\left[\frac{\beta_{mn}V_n}{j\omega\epsilon_o\epsilon_x}\right]+C_{mn}^{x}\left[\frac{m\pi}{d}\right]+D_{mn}^{x}\left[\frac{m\pi}{d}\right]+A_{mn}^{o}\left[\frac{\alpha_{mn}V_n}{j\omega\epsilon_o}e^{-j\alpha_{mn}a}e^{jU_oa}\right]+ \\ &B_{mn}^{o}\left[-\frac{\alpha_{mn}V_n}{j\omega\epsilon_o}e^{j\alpha_{mn}a}e^{jU_oa}\right]+C_{mn}^{o}\left[-\frac{m\pi}{d}e^{-j\alpha_{mn}a}e^{jU_oa}\right]+D_{mn}^{o}\left[-\frac{m\pi}{d}e^{j\alpha_{mn}a}e^{jU_oa}\right]=0; \end{aligned} \qquad (5)$$

continuity on E_z and periodicity conditions at x=0

$$\begin{aligned} &A_{mn}^{x}\left[-\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\epsilon_o\epsilon_x}\right]+B_{mn}^{x}\left[\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\epsilon_o\epsilon_x}\right]+C_{mn}^{x}[-jV_n]+D_{mn}^{x}[-jV_n]+A_{mn}^{o}\left[\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\epsilon_o}e^{-j\alpha_{mn}a}e^{jU_oa}\right]+ \\ &B_{mn}^{o}\left[-\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\epsilon_o}e^{j\alpha_{mn}a}e^{jU_oa}\right]+C_{mn}^{o}\left[jV_ne^{-j\alpha_{mn}a}E^{jU_oa}\right]+D_{mn}^{o}\left[jV_ne^{j\alpha_{mn}a}e^{jU_oa}\right]=0; \end{aligned} \qquad (6)$$

continuity on E_z at x=t

$$\begin{aligned} &A_{mn}^{x}\left[-\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\epsilon_o\epsilon_x}e^{-j\beta_{mn}t}\right]+B_{mn}^{x}\left[\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\epsilon_o\epsilon_x}e^{j\beta_{mn}t}\right]+C_{mn}^{x}\left[-jV_ne^{-j\beta_{mn}t}\right]+D_{mn}^{x}\left[-jV_ne^{j\beta_{mn}t}\right] \\ &A_{mn}^{o}\left[\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\epsilon_o}e^{-j\alpha_{mn}t}\right]+B_{mn}^{o}\left[-\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\epsilon_o}e^{j\alpha_{mn}t}\right]+C_{mn}^{o}\left[jV_ne^{-j\alpha_{mn}t}\right]+D_{mn}^{o}\left[jV_ne^{j\alpha_{mn}t}\right]=0; \end{aligned} \qquad (7)$$

jump discontinuity on H_y and periodicity conditions at x=0

$$A_{mn}^{e}\left[\frac{m\pi}{d}\right]+B_{mn}^{e}\left[\frac{m\pi}{d}\right]+C_{mn}^{e}\left[-\frac{V_n\beta_{mn}}{j\omega\mu}d\right]+D_{mn}^{e}\left[\frac{V_n\beta_{mn}}{j\omega\mu}\right]+A_{mn}^{o}\left[-\frac{m\pi}{d}e^{-j\alpha_{mn}a}e^{jV_n a}\right]+$$
$$B_{mn}^{o}\left[-\frac{m\pi}{d}e^{j\alpha_{mn}a}e^{jV_n a}\right]+C_{mn}^{o}\left[\frac{V_n\alpha_{mn}}{j\omega\mu}e^{-j\alpha_{mn}a}e^{jV_n a}\right]+D_{mn}^{o}\left[-\frac{V_n\alpha_{mn}}{j\omega\mu}e^{j\alpha_{mn}a}e^{jV_n a}\right]=R^{zo}; \quad (8)$$

jump discontinuity on H_z at x=t

$$A_{mn}^{e}\left[jV_n e^{-j\beta_{mn}t}\right]+B_{mn}^{e}\left[jV_n e^{j\beta_{mn}t}\right]+C_{mn}^{e}\left[\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\mu}e^{-j\beta_{mn}t}\right]+D_{mn}^{e}\left[-\frac{m\pi}{d}\frac{\beta_{mn}}{\omega\mu}e^{j\beta_{mn}t}\right]+$$
$$A_{mn}^{o}\left[-jV_n e^{-j\alpha_{mn}t}\right]+B_{mn}^{o}\left[-jV_n e^{j\alpha_{mn}t}\right]+C_{mn}^{o}\left[-\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\mu}e^{-j\alpha_{mn}t}\right]+D_{mn}^{o}\left[\frac{m\pi}{d}\frac{\alpha_{mn}}{\omega\mu}e^{j\alpha_{mn}t}\right]=R^{yt}; \quad (9)$$

jump discontinuity on H_y at x=t

$$A_{mn}^{e}\left[-\frac{m\pi}{d}e^{-j\beta_{mn}t}\right]+B_{mn}^{e}\left[-\frac{m\pi}{d}e^{j\beta_{mn}t}\right]+C_{mn}^{e}\left[\frac{V_n\beta_{mn}}{j\omega\mu}de^{-j\beta_{mn}t}\right]+D_{mn}^{e}\left[\frac{V_n\beta_{mn}}{j\omega\mu}e^{j\beta_{mn}t}\right]+$$
$$A_{mn}^{o}\left[\frac{m\pi}{d}e^{-j\alpha_{mn}t}\right]+B_{mn}^{o}\left[\frac{m\pi}{d}e^{j\alpha_{mn}t}\right]+C_{mn}^{o}\left[-\frac{V_n\alpha_{mn}}{j\omega\mu}e^{-j\alpha_{mn}t}\right]+D_{mn}^{o}\left[\frac{V_n\alpha_{mn}}{j\omega\mu}e^{j\alpha_{mn}t}\right]=R^{zt}; \quad (10)$$

where $U_o=\sin\theta_o\cos\phi_o$; $R^{yo}=(2/db)\exp(jV_n y_o)\sin(m\pi z_o/d)$, when the impulsive source is a y-directed current at x=0, and 0 otherwise; $R^{zo}=(i_s/db)\exp(jV_n y_o)\cos(m\pi z_o/d)$, ($i_s=1$ for m=0, and $i_s=2$ otherwise) for a z-directed impulsive source at x=0, and 0 otherwise; $R^{yt}=(2/db)\exp(jV_n y_o)\sin(m\pi z_o/d)$, for a y-directed impulsive source at x=t, and 0 otherwise; $R^{zt}=(i_s/db)\exp(jV_n y_o)\cos(m\pi z_o/d)$, when the impulsive source is a z-directed current at x=t, and 0 otherwise. It should be noted that for z-directed sources and m=0 (TE_x modes only) the matrix described above becomes a 4x4 matrix with rows 2, 4, 5 and 7, and columns 1, 2, 5 and 6 eliminated, whereas for y-directed sources and m=0 the solution is zero. Upon solution to equations (3) to (10), any vector component of the Green's function for any of the four impulsive electric current elements can easily be obtained. As usual, the fields produced by the dipole and feed currents can then be found by using convolution.

The formulation of the magnetic current problem follows steps identical to those described in [5]. We begin by expressing the fields in the dielectric region using the following TM_x and TE_x scalar potentials:

$$\phi^{r}=\sum_{q=1}^{\infty}\sum_{n=-\infty}^{\infty}\sin(k_{zM}z)\,e^{-jV_n y}\{\tilde{A}_{qn}^{r}e^{-jk_{rM}x}+\tilde{B}_{qn}^{r}e^{jk_{rM}x}\},$$
$$\psi^{r}=\sum_{q=1}^{\infty}\sum_{n=-\infty}^{\infty}\cos(k_{zE}z)\,e^{-jV_n y}\{\tilde{C}_{qn}^{r}e^{-jk_{rE}x}+\tilde{D}_{qn}^{r}e^{jk_{rE}x}\}, \quad (11)$$

where k_{zM} and k_{zE} are (q,n) propagation constants that will be later found, and

$$\begin{bmatrix}k_{rM}^2\\k_{rE}^2\end{bmatrix}=\epsilon_r k_o^2-V_n^2-\begin{bmatrix}k_{zM}^2\\k_{zE}^2\end{bmatrix}. \quad (12)$$

Similar TM_x and TE_x scalar potentials can be easily written for the free-space region by using other modal

coefficients and letting k_{rM} and k_{rE} (with $\epsilon_r=1$) equal k_{oM} and k_{oE}. At this time, the formulation contains 10 unknowns, the two propagation constants and the eight mode coefficients. Enforcing continuity of the tangential electric and magnetic field components at x=0 and x=t yields two 4x4 homogeneous matrix equations, one for the TM modes, one for the TE modes (see [5] for these equations). The values of k_{zM} and k_{zE} are then found numerically when the determinant of these matrices go to zero, allowing for a non-trivial solution. In addition, for each matrix, the eigenvector associated with a propagation constant ($[f_i]$ for TM modes and $[h_i]$ for TE modes) establishes a relationship between the coefficients of that mode. In doing so, the formulation of the above potentials becomes

$$\begin{aligned}
\phi^r &= \sum_q \sum_n \sin(k_{zM} z)\, e^{-jV_n y} A_{qn} \{f_1 e^{-jk_{rM}x} + f_2 e^{jk_{rM}x}\}, \\
\psi^r &= \sum_q \sum_n \cos(k_{zE} z)\, e^{-jV_n y} B_{qn} \{h_1 e^{-jk_{rE}x} + h_2 e^{jk_{rE}x}\}, \\
\phi^o &= \sum_q \sum_n \sin(k_{zM} z)\, e^{-jV_n y} A_{qn} \{f_3 e^{-jk_{oM}x} + f_4 e^{jk_{oM}x}\}, \\
\psi^o &= \sum_q \sum_n \cos(k_{zE} z)\, e^{-jV_n y} B_{qn} \{h_3 e^{-jk_{oE}x} + h_4 e^{jk_{oE}x}\},
\end{aligned} \tag{13}$$

thus is reduced to 2 unknown coefficients with the only remaining boundary conditions being those at the plane of the source (z=d). As in [5], no attempt is made to find expressions for the magnetic currents themselves. The 2 remaining modal coefficients join the dipole and feed currents as part of the unknowns that will later be determined.

REGION II

The fields in region II are expressed with propagating Floquet modes in x and y and standing wave for the z variation. Their TE_x and TE_y scalar potentials are given by:

$$\begin{aligned}
\psi_x^I &= \sum_{p=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} e^{-jV_n y} e^{-jU_p x} \{T_{pn}^{+} e^{-jk_z(z-d)} + T_{pn}^{-} e^{jk_z(z-d-l)}\} \\
\psi_y^I &= \sum_{p=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} e^{-jV_n y} e^{-jU_p x} \{L_{pn}^{+} e^{-jk_z(z-d)} + L_{pn}^{-} e^{jk_z(z-d-l)}\}
\end{aligned} \tag{14}$$

$$\text{where } U_p = U_o + \frac{p2\pi}{a}, \text{ and}$$

$$k_z^2 = \epsilon_1 k_o^2 - V_n^2 - U_p^2 .$$

REGION III

The formulation for Region III is similar to that in region II except for the z variation which in the present case is represented by a propagating/evanescent function. The TE_x and TE_y potentials are written as:

$$\psi_x^a = \sum_{p=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} M_{pn} e^{-jV_n y} e^{-jU_p x} e^{-jk_z(z-d-l)}$$
$$\psi_y^a = \sum_{p=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} N_{pn} e^{-jV_n y} e^{-jU_p x} e^{-jk_z(z-d-l)} \qquad (15)$$
$$\text{where } k_z^2 = k_o^2 - V_n^2 - U_p^2, \text{ and}$$
$$\text{Imaginary}\{k_z\} \le 0.$$

NUMERICAL IMPLEMENTATION OF THE METHOD

At this point, one is faced with the determination of the electric currents J(y,z), A_{qn}, B_{qn}, T^+_{pn}, T^-_{pn}, L^+_{pn}, L^-_{pn}, M_{pn} and N_{pn} by numerically implementing the electromagnetic boundary conditions at z=d, z=d+l and on the surface S of the electric conductors. The conditions are fulfilled by a Galerkin moment method procedure for the requirements on S, and by using the conjugate of the Floquet functions for testing the boundary condition equations at z=d and z=d+l. Note that all infinite summations must be truncated when the solution, i.e., the active impedance value, has converged. For the conditions on S, we require that:

$$E_z(\vec{J}) + E_z(\vec{M}(z=d-\delta)) = -E_z^i$$
$$E_y(\vec{J}) + E_y(\vec{M}(z=d-\delta)) = -E_y^i \qquad (16)$$

where δ is used here as a limiting quantity to denote the magnetic current M(x,y) under the metallized aperture z=d, and E^i is an impressed electric field produced by an applied voltage source in the radiation problem. For the conditions at z=d, we require that

$$E_x(\vec{M}(z=d-\delta)) - E_x(\vec{M}(z=d+\delta)) = 0$$
$$E_y(\vec{M}(z=d-\delta)) - E_y(\vec{M}(z=d+\delta)) = 0$$
$$H_x(\vec{J}) + H_x(\vec{M}(z=d-\delta)) - H_x(\vec{M}(z=d+\delta)) = 0$$
$$H_y(\vec{J}) + H_y(\vec{M}(z=d-\delta)) - H_y(\vec{M}(z=d+\delta)) = 0 \qquad (17)$$

where as at z= d+l, we have

$$E_x(\vec{M}(z=d+l-\delta)) - E_x(\vec{M}(z=d+l+\delta)) = 0$$
$$E_y(\vec{M}(z=d+l-\delta)) - E_y(\vec{M}(z=d+l+\delta)) = 0$$
$$H_x(\vec{M}(z=d+l-\delta)) - H_x(\vec{M}(z=d+l+\delta)) = 2H_x^{inc}$$
$$H_y(\vec{M}(z=d+l-\delta)) - H_y(\vec{M}(z=+l+\delta)) = 2H_y^{inc} \qquad (18)$$

H^{inc} represents an incident plane wave of the form

$$\vec{H}^{inc} = \vec{H}_o e^{-jk_o \sin\theta_o \cos\phi_o} e^{-jk_o \sin\theta_o \sin\phi_o} e^{jk_o \cos\theta_o} \qquad (19)$$

illuminating the face of the array in the scattering/receiving problem. The integral equations resulting from equation (16) are to be solved by the method of moments. We begin by approximating the surface

current on either side of the substrate by finite sums of overlapping piecewise sinusoidal functions:

$$\vec{J}(y,z) \sim \hat{y}\sum_{i=1}^{M_y} c_i J_i^y(y) + \hat{z}\sum_{j=1}^{M_z} d_j J_j^z(z)$$
$$\sim \hat{y}\sum_{i=1}^{M_y} c_i \frac{\sin k_o(h_i^y - |y-y_i|)}{W_i^y \sin k_o h_i^y} + \hat{z}\sum_{j=1}^{M_z} d_j \frac{\sin k_o(h_j^z - |z-z_j|)}{W_j^z \sin k_o h_j^z} \quad (20)$$
$$\text{for } y_i - h_i^y \le y \le y_i + h_i^y,\ y_i = y_{i-1} + h_i^y,\ z_j - h_j^z \le z \le z_j + h_j^z,\ z_j = z_{j-1} + h_j^z,$$

where M_y and M_z are the number of y and z-directed current modes of length 2h, W_i is the width of the i-th mode, (y_i, z_i) and (y_j, z_j) are the coordinates of the center of the i-th and j-th modes, and c_i and d_j are the expansion coefficients to be found. It should be noted that this formulation permits the selection of individual mode width, length and location, and as a result, can model localized currents that are rapidly varying. This nonuniform segmentation is implemented in a computer code by writing/reading the mode specifications in data files. Using the Galerkin's method, equation (16) can be written in matrix form as:

$$\begin{aligned} [Z^{zy}]\,[c] + [Z^{zz}]\,[d] + [Z^{zM}]\,[A_{qn}] + [Z^{zE}]\,[B_{qn}] &= [V^z] \\ [Z^{yy}]\,[c] + [Z^{yz}]\,[d] + [Z^{yM}]\,[A_{qn}] + [Z^{yE}]\,[B_{qn}] &= [V^y] \end{aligned} \quad (21)$$

where V^z_i and/or $V^y_i = -1/W_i$ when the i-th testing mode corresponds to the source location, and 0 otherwise. The elements of the matrices $[Z^{yy}]$, $[Z^{zz}]$, $[Z^{yz}]$ and $[Z^{zy}]$ are expressed in terms of the inner products of the testing functions J_i^y and J_i^z and the integral operators representing the fields produced by the electric currents in the dielectric part of the cavity region. For example, the elements of $[Z^{yz}]$ defining the y-directed electric field produced by the j-th, z-directed current mode and tested with J_i^y are given by:

$$Z_{i,j}^{yz} = \int_{y_i-h_i}^{y_i+h_i}\int_{z_i-W_i/2}^{z_i+W_i/2} J_i^y(y) \int_{y_j-W_j/2}^{y_j+W_j/2}\int_{\tau}^{z_j+h_j} J_j^z(z_o) \sum_m \sum_n \sin\left(\frac{m\pi z}{d}\right)$$
$$\cdot e^{-jV_n y}\left[-\left(\frac{V_n}{\omega \epsilon_o \epsilon_r}\right)\{A_{mn}^x(-j\beta_{mn})\, e^{-j\beta_{mn}x} + B_{mn}^x(j\beta_{mn})\, e^{j\beta_{mn}x}\}\right. \quad (22)$$
$$\left. + \left(\frac{m\pi}{d}\right)\{C_{mn}^x e^{-j\beta_{mn}x} + D_{mn}^x e^{j\beta_{mn}x}\}\right] dz_o dy_o dz dy$$
$$\text{for } i = 1, 2, \ldots, M_y, \text{ and } j = 1, 2, \ldots, M_z,$$

where x takes on the value of 0 or t depending on the location of the i-th testing mode. The lower limit τ on the z_o integration equals 0 for z-directed modes with their center at z=0, and $z_j - h_j$ otherwise. By so doing, the nonzero currents which exist at the ground plane-feed junction can be treated. We should point out that the integrals present in equation (22) and throughout the numerical method are all evaluated analytically. Equation (22) is written in the present form so that elements of $[Z^{yy}]$, $[Z^{zz}]$ and $[Z^{zy}]$ (not shown here) can be easily derived from it. In the scattering/receiving problem, a delta-gap load can be considered by simply adding its value to a selected self impedance value. The reader should also realize that the Green's function coefficients contain the y_o, z_o functional dependence. The elements of

$[Z^{zM}]$, $[Z^{zE}]$, $[Z^{yM}]$ and $[Z^{yE}]$ represent the fields produced by a particular (q,n) magnetic current mode and tested with the piecewise sinusoidal functions described above. For example, $[Z^{zM}]$ can be found as

$$Z_{i,j}^{zM} = \frac{k_{zM}}{j\omega\epsilon_o\epsilon_r} \int_{y_i-W_i/2}^{y_i+W_i/2} \int_{\tau}^{z_i+h_i} J_i^z(z) \cos(k_{zM}z)\, e^{-jV_n y} \cdot \{f_1 e^{-jk_{zM}x} + f_2 e^{jk_{zM}x}\}\, dz\, dy \qquad (23)$$

where i=1,2, ..., M_z, and j is the index representing the ordering of the unknown modal coefficients A_{qn}.

Matrix equations for equations (17) and (18) can be obtained as well by testing the field quantities with Floquet functions

$$e^{jU_l x}\, e^{jV_k y} \qquad (24)$$

over the equivalence apertures at z=d and z=d+l. For the sake of brevity, expressions for the resulting matrices are not shown here, but most of them can be found in the work presented in [5].

A point of concern relates to the storage and CPU time requirements necessary for implementing the numerical method on a computer. In fact, one might ask why not enforce the conditions at z=d+l analytically rather than numerically. With the exception of the matrices describing the conditions on S, most of the matrices are highly sparse. More specifically, those which come from equation (18) are diagonal matrices. The overall solution to the system of equations is obtained by using a matrix solver which requires only the storage of the nonzero elements ([7] modified by the author to handle complex matrices). The author elected to enforce the conditions at z=d+l numerically because the present method and resulting code are easily extendable to a multilayer radome configuration without any significant increase in storage or CPU time requirements. The storage and CPU requirements are heavily dominated by those needed for satisfying equation (16). As in [5], using the (m,n) Green's function contributions/coefficients in as many matrix elements as possible helps reduce the CPU time requirements.

NUMERICAL RESULTS

In this section, we illustrate the method numerically by computing the element active impedance and scattering coefficient for many scan angles in the E (ϕ_o=90°) and H-plane (ϕ_o=0°) of the array, and various dielectric substrate and radome parameters. In our first example, scattering calculations aimed at providing partial verification of the method and code are presented. The reflection coefficient from an infinite array of unloaded constant-width-slot antenna (CWSA) elements is calculated and compared to waveguide measurement values presented in [8] (see Figure 3). The unit cell dimensions are a=4.755 cm, b=4.43 cm and d=10 cm, and the substrate relative permittivity is set to unity. The CWSA element is made of two rectangular plates whose width (y-dimension) =2.12 cm and height (z-dimension) = 2.95 cm, and separated by 0.19 cm. The current on each plate is approximated with a total of 10 current

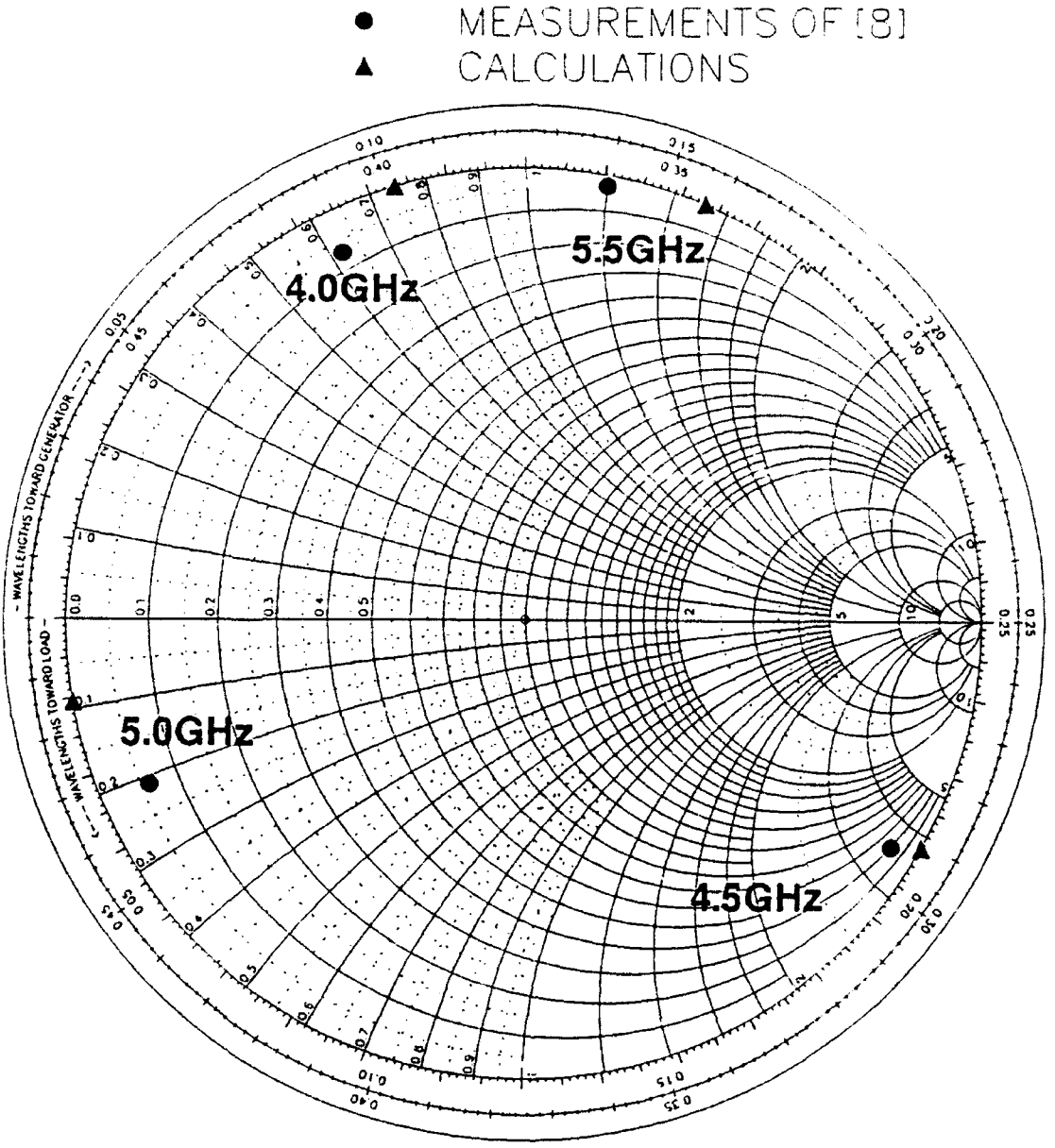


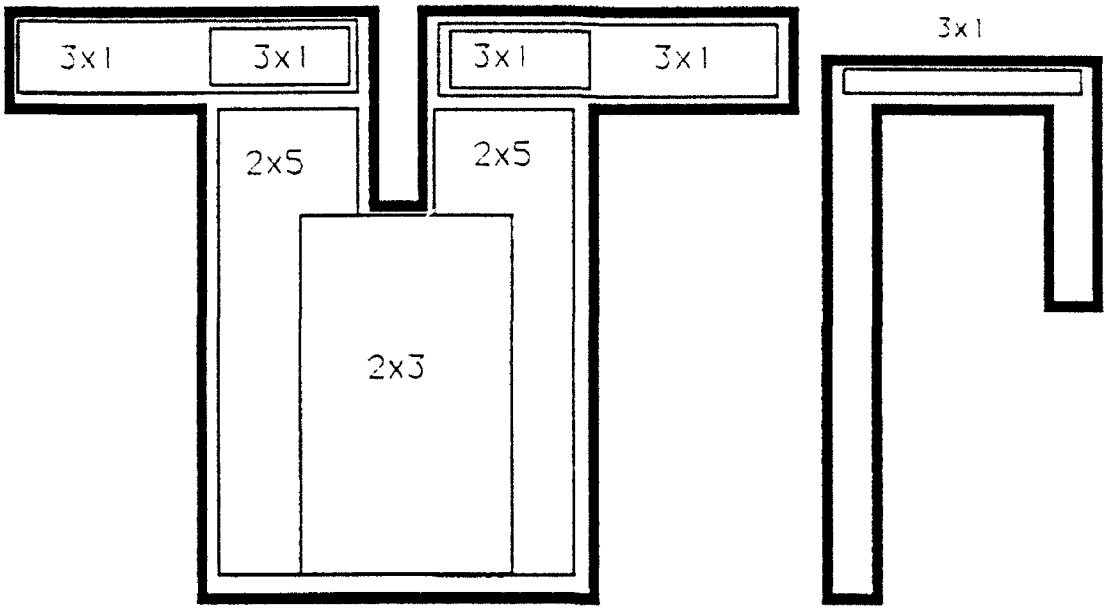
Figure 3. Comparison of waveguide simulator scattering measurements [8] and calculations for an unloaded CWSA array in free-space.

modes J^y and 12 current modes J^z. As in [8], since the equivalence plane is located at z=10 cm, only 18 magnetic modes are used for the cavity region as well as for the free-space region. Fairly good agreement exists between our calculations and the measurements of [8]. There is a slight discrepancy in the angle values (the magnitude of the calculated reflection coefficient is expected to always be unity), but it is certainly comparable to that exhibited in [8] between theory and experiment.

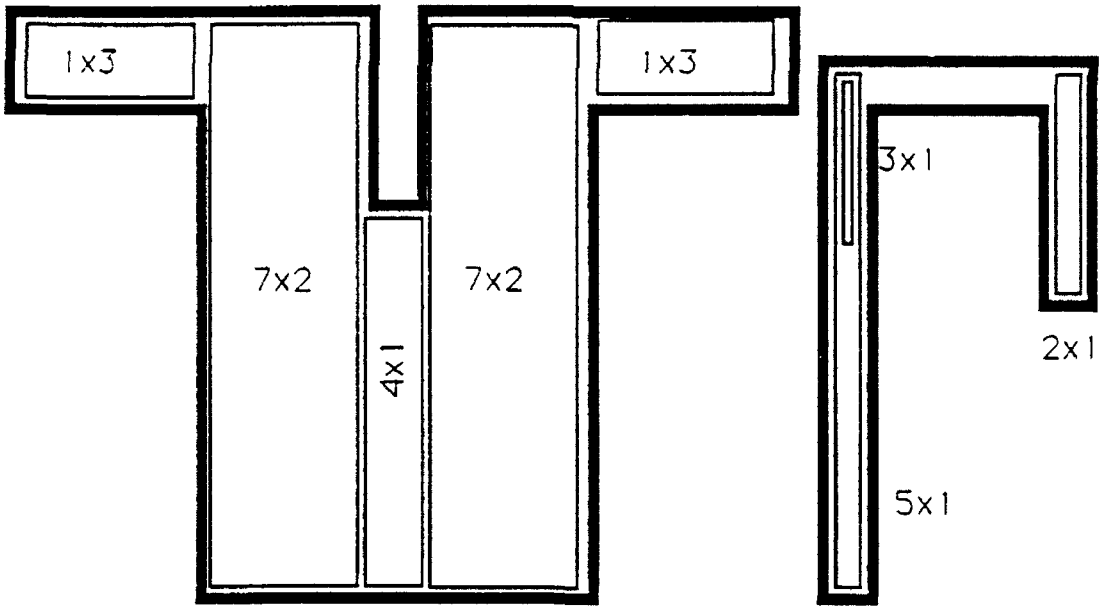
Next, the dipole element shown in Figure 1b is considered. The unit cell dimensions are a=b=0.5λ and d=0.3λ for a frequency of 300 MHz. The element dimensions in wavelength are:

L_d	0.38
W_d	0.05
W_c	0.02
W_s	0.20
L_f	0.255
W_f	0.010
y_f	0.0675
L_c	0.100
L_1	0.065
L_3	0.25
L_2	0.045

In all the cases presented, the calculated input impedance of the element is referenced to the feedline point at the dipole's apex. This can be accomplished in the code by simply selecting the center mode (among an odd number of modes) of the y-directed microstrip feedline segment as the excited mode. For an ideal transmission line, this is equivalent to doing a reference plane extension by the electrical length of the microstrip line from the back wall to the dipole's apex. The reader should be made aware that the coplanar stub length and width, the dipole length, and the microstrip line dimensions and position are parameters all of which affect a great deal the active impedance value. The above dimensions were selected somewhat arbitrarily, however, matching and bandwidth design criteria could be readily implemented on the microstrip line. Another issue of importance relates to the mode lay-out used to model the surface currents on the printed conductors. The printed geometry (feed and antenna) is divided into rectangular regions where uniform current modes are positioned (see Figure 4). The mode density is typically higher in regions located at or near the conductor junctions. For example, consider



(a) y-directed mode lay-out



(b) z-directed mode lay-out

Figure 4. The rectangular regions used to lay-out uniform piecewise sinusoidal modes on the printed conductors.

the left dipole arm in Figure 4a. The y-directed current is approximated with 3 larger modes and 3 smaller modes at the dipole-feed junction. All six modes have a width equal to W_d. A total of 41 y-directed modes and 48 z-directed modes are used in the method. We should point out that this discussion is included not to imply uniqueness of the lay-out, but rather for the sake of completeness if the reader wish to reproduce the results.

Figure 5 shows the active impedance versus scan for the array with and without the radome cover ($\epsilon_r = 2$) and for $\epsilon_r = 2.2$ and $t = 0.01\lambda$. These curves show the presence of a null in the active resistance value near 23 degree scan. This null, evidence of a blindness, can be attributed to radiation from the feed structure (microstrip and/or coplanar stub). No significant change in the blindness location is observed by adding the thin and low-permittivity radome cover. In Figure 6, the substrate relative permittivity is increased to 3. The results appear to also show a null between 20 and 30 degree scan, however more data points need to be calculated in that region.

CONCLUDING REMARKS

A method for analyzing infinite arrays of straight-arm dipoles fed by coplanar stub coupled to microstrip feed has been presented. Preliminary results show the presence of a feed-induced blindness which limits considerably the scan performance of the array. Future work will focus on experimentally verifying the method and code and on calculating more cases with various dielectric substrate and radome parameters.

DELIVRABLES

The following items were delivered to the laboratory:

a) The moment method code (executable version of LOOPC.FOR).

b) Codes for finding the propagation constants and corresponding eigenvectors for the TM and TE modes (executable versions of RT1.FOR, RT2.FOR, EMODE1.FOR, EMODE2.FOR).

c) A code for laying out modes on rectangular regions, and creating data files to be used in LOOPC.FOR (MODE_LAY.FOR).

A member of the antenna group was also trained to understand the method and use the code.

ACKNOWLEDGEMENT

The author wishes to thank Dr. Michael E. Cooley for his helpful discussions and for providing the measured data. Thanks also to Professor Daniel H. Schaubert and Dr. Boris Tomasic for their suggestions.

REFERENCES

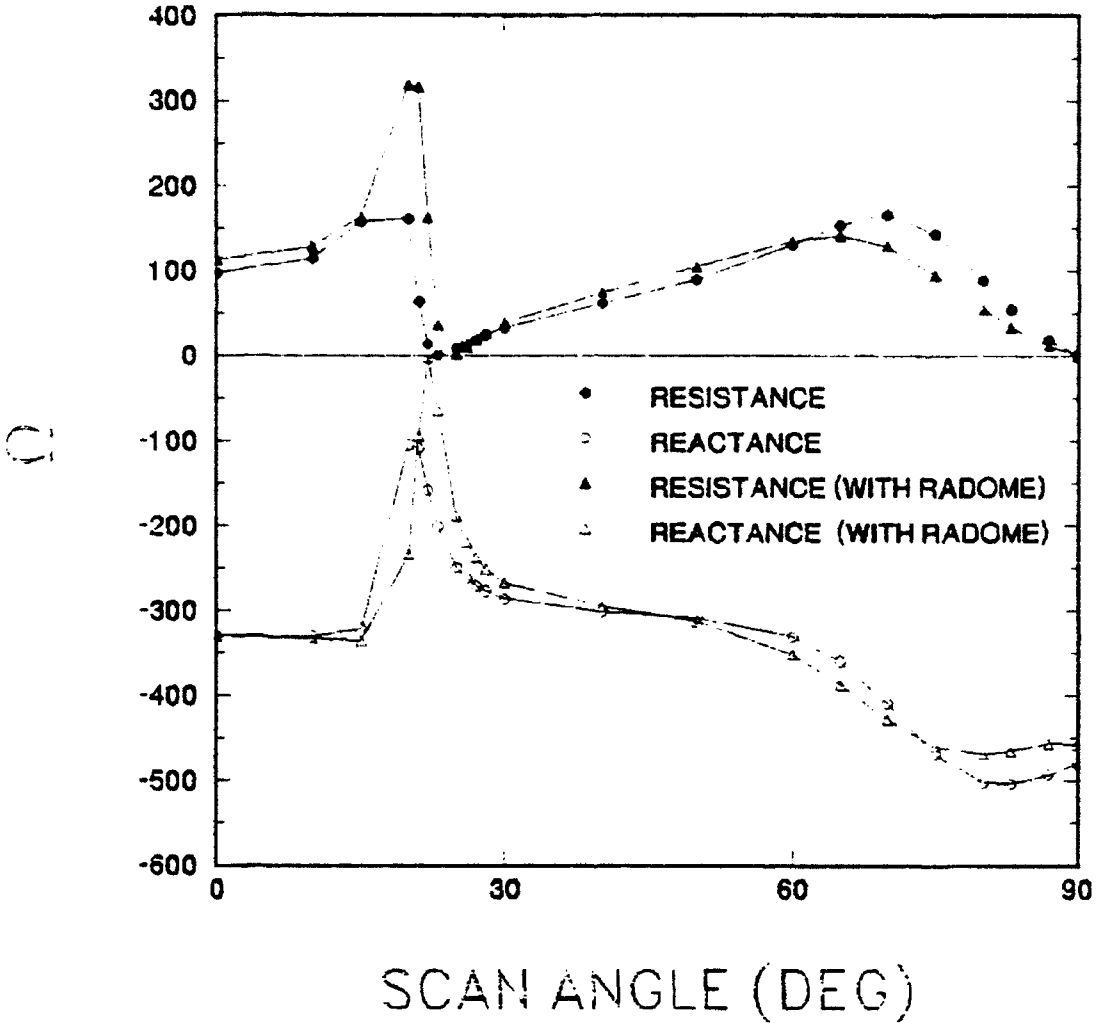


Figure 5. Element impedance versus E-plane scan for $t=0.01\lambda$, $\epsilon_r = 2.2$ and $\epsilon_i = 2$.

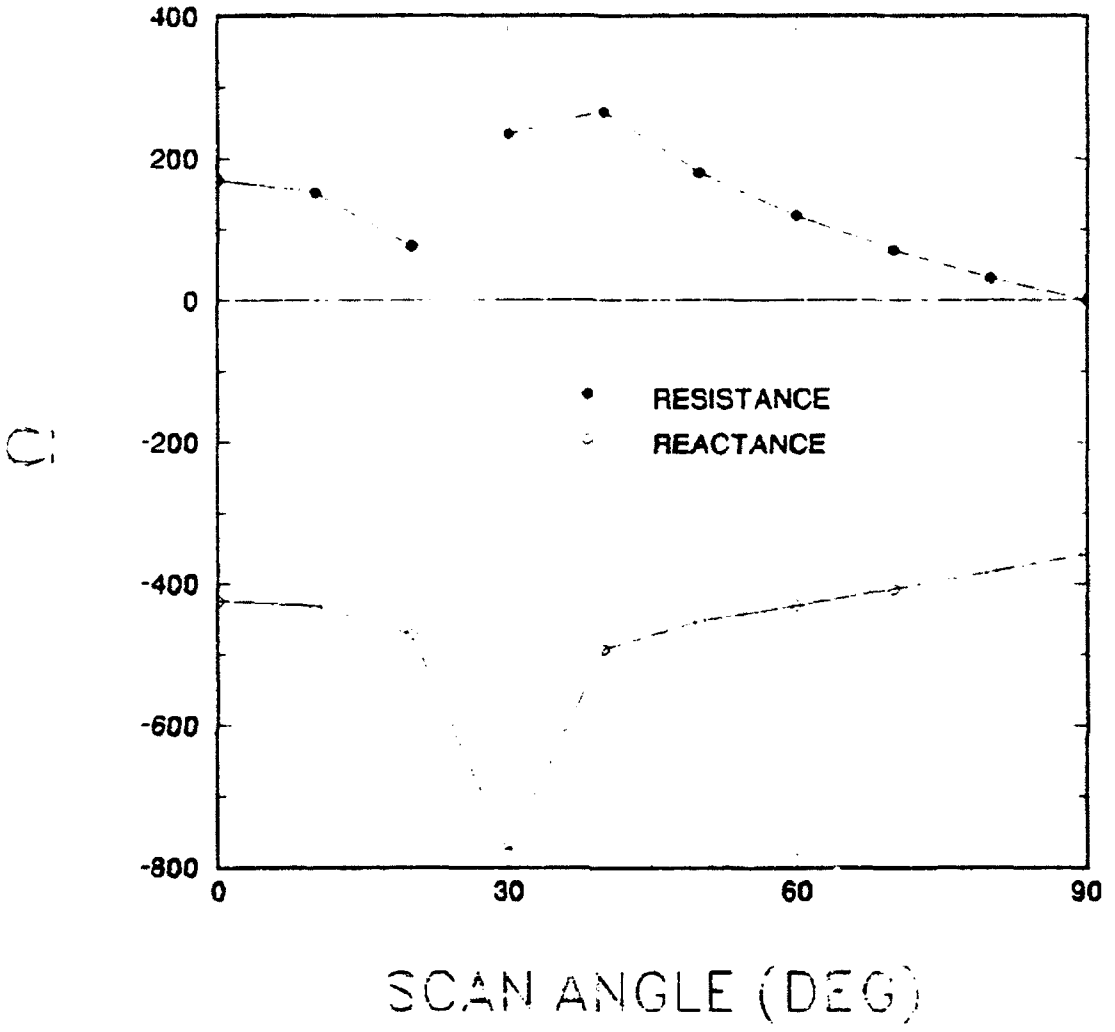


Figure 6. Element impedance versus E-plane scan for t=0.01λ, ϵ_r = 3 and ϵ_l = 1.

[1] Stark, L., "Radiation Impedance of a Dipole in an Infinite Planar Phased Array," Radio Science, Vol. 1 (New Series), No. 3, pp. 361-377, March 1966.

[2] Schuman, H. K., D. R. Pflug, and L. D. Thompson, "Infinite Planar Arrays of Arbitrarily Bent Thin Wire Radiators," IEEE Transactions on Antennas and Propagation, Vol. AP-32, No. 4, pp. 364-377, April 1984.

[3] Chu, R-S., and K-M. Lee, "Radiation Impedance of a Dipole Printed on Periodic Dielectric Slabs Protruding Over a Ground Plane in an Infinite Phased Array," IEEE Transactions on Antennas and Propagation, Vol. AP-35, No. 1, pp. 13-25, January 1987.

[4] Chu, R-S., "Analysis of an Infinite Phased Array of Dipole Elements with RAM Coating on Ground Plane and Covered with Layered Radome," IEEE Transactions on Antennas and Propagation, Vol. AP-39, No. 2, pp. 164-176, February 1991.

[5] Bayard, J-P. R., M. E. Cooley and D. H. Schaubert, "Analysis of Infinite Arrays of Printed Dipoles on Dielectric Sheets Perpendicular to a Ground Plane," IEEE Transactions on Antennas and Propagation, Vol. AP-39, No. 12, pp. 1722-1732, December 1991.

[6] Bayard, J-P. R., M. E. Cooley and D. H. Schaubert, "A General Method for Treating Infinite Arrays of Antennas Printed on Protruding Dielectric Substrates," IEEE/AP-S International Symposium Digest, pp. 600-603, 1991.

[7] Sherman, A. H., "NSPIV, a Fortran Subroutine for Sparse Gaussian Elimination with Partial Pivoting," Collected Algorithm from ACM, Algorithm 533, 1978.

[8] Cooley, M. E., D. H. Schaubert, N. E. Buris, and E. A. Urbanik, "Radiation and Scattering Analysis of Infinite Arrays of Endfire Slot Antennas with a Ground Plane, "IEEE Transactions on Antennas and Propagation, Vol. AP-39, No. 11, pp. 1615-1625, November 1991.

STATISTICAL COMPARISON OF SEVERAL AUTOMATIC TARGET RECOGNITION (ATR) SYSTEMS

Pinyuen Chen
Associate Professor
Department of Mathematics

Syracuse University
Syracuse, New York 13244-1150

Final Report for:
Summer Research Program
Rome Laboratory/IRRE

Sponsored by:
Air Force Office of Scientific Research
Griffiss Air Force Base, New York 13441-5700

September 1992

STATISTICAL COMPARISON OF SEVERAL AUTOMATIC TARGET RECOGNITION (ATR) SYSTEMS

PINYUEN CHEN

Associate Professor
Department of Mathematics
Syracuse University

ABSTRACT

A selection procedure is proposed to select the best Automatic Target Recognition system among several candidate systems based on two performance measures, the probability of detection and the probability of false alarm. Two classical statistical approaches, the indifference zone approach and the subset selection approach, are integrated to form a meaningful approach to meet the evaluation requirements. The proposed procedure allows us to select the best system if there is a significant improvement of the best over the second best, or the procedure will select a subset of systems that contains the best when the difference between the top two systems is not significant. Four examples and a table are given at the end of this report to implement the procedure.

STATISTICAL COMPARISON OF SEVERAL AUTOMATIC TARGET RECOGNITION (ATR) SYSTEMS

PINYUEN CHEN

1. INTRODUCTION

This report proposes a statistical methodology for comparing several Automatic Target Recognition (ATR) systems (algorithm suites). An ATR system is usually composed of algorithms which are arranged in a particular order by the system designer and are fine tuned to meet some specific requirements of the individual program. These algorithms include neural network paradigms, conventional preprocessors and transformations for image processing, and expert systems and artificial intelligence approaches in parallel processing to meet Rome Laboratory's functional requirements for an image exploitation system. While each ATR system may achieve its performance optimality in the class of the systems that are composed of these algorithms, the superiority of an ATR system to other competing ATR systems is often unknown to the designer. In this report, we introduce the statistical ranking and selection theory which will allow us to compare the performances of $k(> 1)$ ATR systems and, among several candidate systems, either (1) to select the best system (i.e. the one with the highest performance rank) when it is significantly better than the others, or (2) to select a subset from all the candidates that contains good systems when there is not a system that outruns the others by a statistically significant margin. The following assumptions are needed for the proposed comparing procedure being applicable to the ATR systems: (1) The same database (or databases) will be used to test all the competing systems. (2) All the systems should use the same standard definitions for the measures of the probability of correct detection, the probability of false alarm, and other relevant measures that were given in "Target Recognizer Definitions and Performance Measures" ATRWG 86-001, provided by ATRWG[1] Evaluation Committee. (3) All the competing systems provided by the designers should give reasonable probabilities of correct detection (usually higher than .7) and probabilities of false alarm (usually lower than .3) for some standard databases suggested or provided by the evaluator.

There have been efforts in evaluating digital imagery exploitation algorithms by the Image Exploitation Branch (IRRE) at Rome Laboratory. The report [1] : Exploitation Evaluation Test Bed—RADC-TR-88-241 (EETB) provided the designers an evaluation methodology for each stage of the ATR process and to help a designer to compare his newly developed algorithm against those already in the EETB algorithm data base and to determine utility

[1] ATRWG is a joint government-industry group that serves as a forum for issues and actions relating to automatic target recognition

in the overall ATR process. Here in this report, we emphasize the comparison of several new ATR systems proposed by the designers and to enable an evaluator to select the best design. The report [2]: Automated Imagery Exploitation Evaluation—RADC-TR-90-292 (AIEE) gave the description of a step-by-step procedure for conducting automated Imagery Exploitation evaluation experiments. In AIEE, statistical response surface methodology was used to fit the data by polynomials and statistical hypothesis testing was used to test the significance of the independent variables. The performance measure (the probability of correct detection) was then predicted from the fitted model. In this report, assuming that the evaluation of each individual system has been done by its designer, we propose a statistical selection procedure to compare the final overall performances of several ATR systems and select the best system with predetermined levels of confidence. In statistical selection problems, one of the two approaches, namely, the indifference zone approach and the subset selection approach, is usually adopted, depending on the application involved (see [3] and [4]). The indifference zone approach, which selects the best population (or system in our application) and guarantees with a specific confidence level whenever the best is better than the second best by a predetermined margin, emphasizes the design aspect of the problem. On the other hand, the subset selection approach, which selects a subset that contains the best population (or system in our application) with a specific confidence level, emphasizes data analysis. Here in our report, we study a composite formulation which integrates the two above mentioned approaches to compare and select among several ATR systems. The idea of integrating the two approaches to form a new formulation was first studied in [5] from a theoretical point of view. It is a meaningful formulation for the purpose of evaluating ATR systems since the Rome Laboratory, as an evaluator, would choose the best system if there is one, or screen out the bad ones and do further investigation on the remaining systems based on other conditions if there is no absolute best in the current situation. The other conditions that the evaluator may consider are the economical considerations, the true target backgrounds, and some prior knowledge about the distribution of the targets.

The report is organized in the following manner. Section 2 reviews the terminologies in automatic target recognition and their definitions, and the terminologies in statistics that are relevant to the current study. Section 3 formally proposes the main procedure and states the theorem in supporting the procedure. In Section 4, we derive the formulas, explain the computing algorithms, and give four examples to illustrate how to use the FORTRAN program and the table. The appendix provides the proofs of the main theorems in Section 3.

2. TERMINOLOGIES AND THEIR DEFINITIONS

One of the main performance objectives of an ATR system is to detect the target in a test image that it has been trained to learn, and not detect non-target areas of the test image as targets. Each ATR system under evaluation is treated as if it was a "black box". An input image is presented and the system generates a response that is compared with the known ground truth. The following definitions, which were extracted from [7] establish a base on which the set of performance probabilities, the major criteria of our statistical comparison of the ATR systems, can be defined formally.

Image: A two-dimensional array, $I(m,n)$, of pixels available as digital quantities from the output of a system. The (i,j)th pixel of an image is denoted as W_{ij}.
Region: A collection of pixels of an image.
Reference Pixel ($C(R)$): A pixel that represents a region. The reference pixel for a region R is denoted as $C(R)$. Some possible types of reference pixel are:
* Central Pixel: The center of gravity of the region.
* Area Median Pixel: The pixel which has half of the pixels in the region to its left and half of the pixels in the region above it.

Ground Truth (G_T): The reference information available regarding the data collection or the data generation process. This information is generally of two types:
* Scenario Information (environment, weather, range to targets, target locations, target aspects, etc.)
* ATR System Information (sensor location, sensor orientation, sensor characteristics, etc.)

Target Region (T): A set of image pixels $T = \{W_{ij}\}$ to which a target label has been assigned by an ATR system.
Detection: The correct association of a target region with a ground truth target region based on an appropriate detection criterion. For each image, only one detection per ground truth target region is allowed. The total number (over all images) of ground truth targets is denoted by N_{GT}. The total number (over all images) of ground truth targets that are detected is denoted by N_D.
Detection Criterion: Rules used to determine whether an ATR system's target region output corresponds to a ground truth target region. Some criteria are listed below.
* Region Intersection Criterion:
  Card$(T \cap G_T) > c$ for any $G_T \longrightarrow$ detection where Card(X) denotes the number of elements in X.
* Reference Distance Criterion:
  $|C(T), C(G_T)| < c$ for any $G_T \longrightarrow$ detection where $|\ \ |$ denotes the norm between the pixels.
* Ground Truth Reference Pixel Criterion:
  $C(G_T) \in T$ for any $G_T \longrightarrow$ detection
* Target Region Reference Pixel Criterion:
  $C(T) \in G_T$ for any $G_T \longrightarrow$ detection

False Alarm: If a target region does not correspond to any ground truth target region based on the chosen detection criterion, then a false alarm has been generated by the system.

Now we are ready to define the performance probabilities.
True Probability of Detection: The unknown true probability of the system labeling a ground target region as a target region.
True Probability of False Alarm: The unknown true probability of the system labeling as a target region a region which is not a ground truth target region.

For a given system, the above two probabilities correspond to the probabilities of two types of error in statistical hypothesis testing problems. If we consider the two probabilities as two functions of the procedure parameters (system thresholds in our application), then they will both be monotone in the same directions (see the report [6]: Neural Network Investigation Final Report—RL/IRRE, Oct. 1991). That is, if the probability of detection

increases as the threshold increases, then the probability of false alarm will also increase. This is a common phenomenon in statistical testing problem. Our primary interest in this report is to compare the candidate systems with respect to their performance measure which can be defined as a meaningful function $f(p_D, p_{FA})$ of the two probabilities $p_D = P(D) =$ probability of detection and $p_{FA} = P(FA) =$ probability of false alarm. Possible choices for $f(p_D, p_{FA})$ are $a_1 p_D - a_2 p_{FA}$ $(a_1, a_2 > 0)$, $b p_D / p_{FA}$ $(b > 0)$, and $c[p_D/(1-p_D)]/[p_{FA}/(1-p_{FA})]$ $(c > 0)$. Although the general theory holds for all the three functions mentioned above with minor modifications, we will restrict ourselves only on $f(p_D, p_{FA}) = p_D - p_{FA}$ and we will denote it by μ which represents the mean performance of the system. Let $\pi_1, \pi_2, \ldots, \pi_k$ be k competing systems. The observed performance measure X_i for each system π_i on an image is defined by

(2.1)
$$X_i = N_{CD}/N_{GT} - N_{FA/No}$$
number of correct target detections/number of ground truth targets - number of false alarms/number of false alarm opportunities

and is assumed to follow a normal distribution with population mean μ_i $(i = 1, 2, \ldots, k)$ with common known variance σ^2. The normal distribution assumption is reasonable since the random variable X is the difference of two binomial variables, each with a fairly large size n. (In report [6], page 25, there were 16 polygons in each of the 424 images.) AIEE (report [2]) also assumed the normal noise for the dependent variable, the probability of a correct detection, in the response surface analysis. The assumption about the common variance is a more restricted one, and needs to be generalized in future studies. Here for the purpose of data analysis, we will use the pooled sample variance (see the definition in EXAMPLE (1) below) as an estimate of the common variance and denote it by σ^2. For the purpose of designing an experiment, we will use a conservative upper bound (see EXAMPLE (4) below) as an estimate of σ^2. Let the ordered values of the k population means be denoted by

(2.2)
$$\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k-1]} < \mu_{[k]}$$

where we assume that $\mu_{[k-1]}$ is strictly less than $\mu_{[k]}$ in order that the best system(with the largest performance mean) should be well-defined. If we let δ denote the difference $\mu_{[k]} - \mu_{[k-1]}$ then by (2.2), $\delta > 0$. The set of parameter vector $\underline{\mu}$ for which $\delta \geq \delta^*$ (where $\delta^* > 0$ is specified) will be called the preference zone PZ; the complementary set of $\underline{\mu}$ for which $\delta < \delta^*$ will be called the indifference zone. We define our goal in two parts, according

to whether the true parameter $\underline{\mu}$ is in the PZ or the IZ, as follows

$$\text{(2.3)} \qquad \text{GOAL} \begin{cases} \text{For } \underline{\mu} \in PZ, \text{ we want to select the best system with confidence level at least } P_1^*. \\ \text{For } \underline{\mu} \in \text{IZ, we want to select a subset containing the best system with a confidence level at least } p_2^* \end{cases}$$

where both p_1^* and p_2^* (as well as δ^*) are specified.

3. The Proposed Procedure

We are considering all the systems being tested on the same image databases, so that there will be a common number of observations n from each of the k populations. Our procedure depends only on the sufficient statistics X_i for the unknown mean performance measure μ_i $(i = 1, 2, \ldots, k)$ and the ordered values of the X_i are denoted by

$$\text{(3.1)} \qquad X_{[1]} \leq X_{[2]} \leq \cdots \leq X_{[k]}$$

Since we are selecting the best system, the two observed measures $X_{[k]}$ and $X_{[k-1]}$ play a special role in our procedure R. Let $c > 0$ and $d > 0$ be constants to be determined (along with n) as a function of δ_1^*, p_1^*, and p_2^* for any given values of k. Assume that c, d, and n are already determined.

Procedure R : If $X_{[k]} - X_{[k-1]} > c$, then we select the system that gives rise to the largest sample mean $X_{[k]}$.

If If $X_{[k]} - X_{[k-1]} < c$, then we select a subset consisting of all those systems π_i with $X_i > X_{[k-1]} - d$.

A correct decision in the PZ (denoted by $CD_1|PZ$) is defined to be the decision of selecting the best and only the best system. A correct decision in the IZ (denoted by $CD_2|IZ$) is defined to be the selection of a subset containing the best system. Let $P(CD_1|PZ)$ and $P(CD_2|IZ)$ denote the probabilities of correct decision under the preference zone and the indifference zone respectively. To implement our proposed procedure, we need to find the procedure parameters c and d so that the lower bounds (or the confidence levels) P_1^* for $P(CD_1|PZ)$ and P_2^* for $P(CD_2|IZ)$ are met. We state in the next two theorems the important results about where the minimum of $P(CD_1)$ occurs in the PZ (denoting it as the least favorable configuration LFC) and where the minimum of $P(CD_2)$ occurs in the IZ (denoting it as the worst configuration WC). The proofs of these results are in the appendix of the report. The LFC and the WC are used in the next section to compute the procedure parameters c and d that are needed to satisfy the probability requirements P_1^* and P_2^*. The tabulated c and d values in the table for each n guarantee the (P_1^*, P_2^*) requirements for the worst situation, and thus they guarantee the (P_1^*, P_2^*) requirements for any situation.

Theorem 3.1. *Under procedure R the LFC for $P(CD_1)$ in the PZ is given by the following configuration:*

$$\mu_{[1]} = \mu_{[2]} = \cdots = \mu_{[k-1]} = \mu_{[k]} - \delta^*. \tag{3.2}$$

Theorem 3.2. *Under procedure R the WC for $P(CD_2)$ in the IZ is given by the following configuration:*

$$\mu_{[1]} = \mu_{[2]} = \cdots = \mu_{[k]} \tag{3.3}$$

4. Formulas, Table, and Examples

Let $X_{(i)}$ denote the sample mean performance measure associated with $\mu_{[i]}$ $(i = 1, 2, \ldots, k)$. Then the probability of a correct decision in the PZ, $P(CD_1|PZ)$ can be written as

$$\begin{aligned} P(CD_1|PZ) &= P(X_{(k)} > X_{(i)} + c, \qquad i = 1, 2, \ldots, k) \\ &= P\left(\frac{X_{(k)} - \mu_{[k]}}{\sigma} > \frac{X_{(i)} - \mu_{[i]}}{\sigma} + \frac{c + \mu_{[i]} - \mu_{[k]}}{\sigma}, \qquad i = 1, 2, \ldots, k\right) \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi\left(z - \frac{c + \mu_{[i]} - \mu_{[k]}}{\sigma}\right) d\Phi(z) \end{aligned} \tag{4.1}$$

Under the LFC given in (3.2), the above expression can be simplified to

$$P(CD_1|LFC) = \int_{-\infty}^{\infty} \Phi^{k-1}\left(z + \frac{\delta^* - c}{\sigma}\right) d\Phi(z) \tag{4.2}$$

The probability of a correct decision in the IZ, $P(CD_2|IZ)$ can be written as

(4.3)

$$\begin{aligned} P(CD_2|IZ) &= P(X_{(k)} > X_{(i)}, \qquad i = 1, 2, \ldots, k) \\ &\quad + \sum_{i=2}^{k-1} P(X_{(i)} > X_{(k)} > X_{(j)}, \qquad j = 1, 2, \ldots, k-1,\ j \neq i;\ X_{(i)} < X_{(k)} + c) \\ &\quad + \sum_{\substack{i=1 \\ i \neq j}}^{k-1} \sum_{j=1}^{k-1} P\left(X_{[k]} = X_{(i)}, X_{[k-1]} = X_{(j)}, X_{(k)} > X_{(j)} - d, X_{(i)} - X_{(j)} < c\right) \\ &= T_1 + T_2 + T_3. \end{aligned}$$

T_1 in (4.3) can be written as

$$T_1 = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi\left(z + \frac{\mu_{[k]} - \mu_{[i]}}{\sigma}\right) d\Phi(z)$$

which can be simplified to the following form under the WC:

$$T_1(WC) = \int_{-\infty}^{\infty} \Phi^{k-1}(z) d\Phi(z) = \frac{1}{k}. \tag{4.4}$$

T_2 in (4.3) can be written as

$$T_2 = \sum_{i=2}^{k-1} \int_{-\infty}^{\infty} \left[\prod_{\substack{j=1 \\ j \neq i}}^{k-1} \Phi\left(z + \frac{\mu_{[k]} - \mu_{[j]}}{\sigma} \right) \right] \left[\Phi\left(z + \frac{\mu_{[k]} - \mu_{[i]} + c}{\sigma} \right) - \Phi\left(z + \frac{\mu_{[k]} - \mu_{[i]}}{\sigma/\sqrt{n}} \right) \right] d\Phi(z)$$

which can be simplified to the following form under the WC:

$$T_2(WC) = (k-1) \int \infty_{-\infty} \Phi^{k-2}(z) \left[\Phi\left(z + \frac{c}{\sigma} \right) - \Phi(z) \right] d\Phi(z) \tag{4.5}$$

T_3 in (4.3) can be written as

$$T_3 = \sum_{\substack{i=1 \\ i \neq j}}^{k-1} \sum_{j=1}^{k-1} \int_{-\infty}^{\infty} \left[\prod_{\substack{m=1 \\ m \neq j \\ m \neq i}}^{k-1} \Phi\left(z + \frac{\mu_{[j]} - \mu_{[m]}}{\sigma} \right) \right] \left[\Phi\left(z + \frac{\mu_{[j]} - \mu_{[i]} + c}{\sigma} \right) - \Phi(z) \right] \cdot$$

$$\left[\Phi(z) - \Phi\left(z + \frac{\mu_{[j]} - \mu_{[k]} - d}{\sigma} \right) \right] d\Phi(z)$$

which can be simplified to the following form under the WC:

$$T_3(WC) = (k-1)(k-2) \int_{-\infty}^{\infty} \Phi^{k-3}(z) \left[\Phi\left(z + \frac{c}{\sigma} \right) - \Phi(z) \right] \left[\Phi(z) - \Phi\left(z - \frac{d}{\sigma} \right) \right] d\Phi(z) \tag{4.6}$$

In Table I, we present the procedure parameters (τ_2, τ_3) where $\tau_2 = c/\sigma$, $\tau_3 = d/\sigma$, for $k = 2, 3, 4, 5$; $\tau_1 = \delta^*/\sigma = 3.0,\ 4.0,\ 5.0$; $P_1^* = .75,\ .90,\ .95$; $P_2^* = .75,\ .90,\ .95$. Here (τ_2, τ_3) is solved from the two integral equations:

$$P(CD_1|LFC) = P_1^* \tag{4.7}$$

and

$$P(CD_2|WC) = P_2^*. \tag{4.8}$$

A FORTRAN program atr.f was written on a SUN 3/60 station in the Image Exploitation Branch (IRRE) at Rome Laboratory to compute $P(CD_1|LFC)$ and $P(CD_2|WC)$ from the formulas (4.2)-(4.8). The program atr.f also finds the solutions for the procedure parameters

τ_2 and τ_3 such that (4.7) and (4.8) are satisfied simultaneously. It is clear that τ_2 can be solved from (4.2) alone. It is also clear from (4.4)-(4.6) that $P(CD_2|WC)$ is an increasing function of τ_3. For all the entries (τ_2, τ_3) in Table I, we solve for τ_2 from (4.7) first and then increase τ_3 gradually and check if (4.8) is satisfied. Approximation formulas (27) in page 55 of [9] was used to evaluate the normal distribution function involved in our calculation. Trapezoidal rule (see for example [10]) was used to evaluate the integrals in the formulas. Examples at the end of this section illustrate how to find the procedure parameters from the program for given k, τ_1, P_1^* and P_2^*.

It should be noticed that for fixed τ_1, $P(CD_2|WC)$ converges to an upper bound as τ_3 goes to infinity. When the given P_2^* value exceeds the upper bound, there does not exist a solution for τ_3 in (4.8). If this is the case, the evaluator should request the designers to test their systems on a larger sample size n (that is, more images) to obtain larger τ_1 and solve for (τ_2, τ_3) under the new n. In our table, a * sign is marked when (τ_2, τ_3) does not exist for the given τ_1. For example, when $k = 4$, $\tau_1 = 3$, $P_1^* = P_2^* = .90$, (τ_2, τ_3) does not exist.

However, if τ_1 is increased to 4, we find that $(\tau_2, \tau_3) = (1.54, 2.46)$ from the table which will satisfy our probability requirements. This property will be studied further in Example (4).

For the special case $k = 2$, the $P(CS_2|WC)$ depends only on τ_1 and τ_2, but not on τ_3. Thus we have used the smallest possible value, namely 0, for all the τ_3 entries in the table. The procedure parameters (τ_2, τ_3) stay the same under the same P_1^* for this case.

It is clear from (4.2), (4.5), and (4.6) that $P(CD_1|LFC)$ and $P(CD_2|WC)$ both increase as σ decreases. Thus the table values of (τ_2, τ_3) also guarantee the P_1^* and P_2^* requirements when σ is replaced by an upper bound. In our application to automatic target recognition, the upper bound of σ can be estimated by the method described in Example (1).

EXAMPLE (1): This and Example (2) show how the program atr.f can be used to select the best system or to select a subset containing the best system. Suppose that 3 competing systems π_1, π_2 and π_3 are tested with the same images used in evaluating the NNIIES system (see report [6]) and they give the following P(FA)'s and P(D)'s:

	System	System	System
$P(D)$	230/285	247/285	234/285
$P(FA)$	81/6499	102/6499	181/6499

The evaluator would like to make sure, with probability at least .90, that he selects the best system if the best performance measure is at least $\delta^* = .1$ higher than the second best, or he will select a subset containing the best, with probability of at least .90. Thus we have $k = 3$, $P_1^* = .90$, and $P_2^* = .90$. The common value σ can be estimated as follows:

Let V_i and W_i denote the sample probability of detection and the sample probability of false alarm respectively for system π_i $(i = 1, 2, 3)$. Our selection statistics X_i is defined as $X_i = V_i - W_i$. Since $285V$ (= the number of detection) and $6499W$ (= the number of false alarms) follow the binomial distributions with parameters $(285, P(D))$, and $(6499, P(FA))$ respectively (see page 25, [6]). The selection statistics $V_i - W_i$ follows a distribution which can be approximated by a normal curve with mean $P(D) - P(FA)$ and variance $\sigma^2 = \mathrm{Var}(V) + \mathrm{Var}(W)$ which can be estimated from the sample by the minimum

variance unbiased estimator $V_i(1-V_i)/n + W_i(1-W_i)/m$ where n and m are the total number of targets and the total number of false alarm opportunities respectively, which are both known to the evaluator. The three sample variances for the common σ^2 are .000547893, .000407835, and .000519692. The unbiased estimate for σ^2 is the average .0004918. Thus $\tau_1 = \delta^*/\sigma = 4.51$. Now we have all the parameters ($k = 3$, $P_1^* = .90$, $P_2^* = .90$, $\tau_1 = 4.51$) that we need to run the FORTRAN program. Fix the values of these parameters at lines 3, 4, 5, and 7 in the program atr.f and compile it by 'f77 atr.f' and run it by 'a.out' on SUN 3/60. It took about 3 minutes to get the result $\tau_2 = 2.27$ and $\tau_3 = 1.31$. Now as defined right below (4.6), $c = (2.27)(.0222) = .0504$, $d = (1.31)(.0222) = .0291$. The sample performance measures for the three systems are respectively .7946, .8510, and .7932. Since $.8510 - .7946 = .0564 > c = .0504$, thus we select system π_2 as the best with confidence level at 90%.

EXAMPLE (2): If the observed performance measures in the above examples are changed to the following numbers:

	System	System	System
$P(D)$	240/285	247/285	234/285
$P(FA)$	81/6499	102/6499	181/6499

Then the unbiased estimate for σ^2 changes to .0004653 and τ_1 changes to 4.64. We rerun the program atr.f and obtain $\tau_2 = 2.40$, $\tau_3 = 1.25$. Hence $c = (2.40)(.0216) = .0518$, $d = (1.25)(.0216) = .0270$. The performance measures now are .8297, .8510, and .7932. Since $.8510 - .8297 = .0213 < c = .0518$, and $.8297 - .7932 = .0365 > d = .0270$, thus we select a subset of two systems π_1 and π_2, and claim that this subset includes the best system with a confidence level at 90%.

EXAMPLE (3): If the required distance δ^* between the best and the second best in Example (1) is reduced to .05, then $\tau_1 = \delta^*/\sigma$ becomes 2.26. Now we rerun the program with the new τ_1 value. The result of the program on the screen is "unable to find tau3—tau1 too small for P2star" The proposed procedure is not able to differentiate the systems and rank among them. We should either lower the confidence level P_2^* or increase the sample size (i.e. the values of 285 and 6499) to make τ_1 larger. The latter one (i.e. to increase the sample size) makes the problem an experimental design problem that we will discuss in Example (4) below. The comment "unable to find tau3—tau1 too small for P_2^*" also means that τ_2 can be found for the given P_1^* (remember that $P(CD_1|LFC)$ depends on τ_1, and τ_2, but not τ_3). Since $P(CD_2|WC)$ depends on τ_1, τ_2, and τ_3, we may need to reduce both P_1^* and P_2^* to find τ_3. Suppose that P_1^* and P_2^* are set to be .75 and .70. Then $(k, P_1^*, P_2^*, \tau_1) = (3, .75, .70, 2.26)$ gives us $(\tau_2, \tau_3) = (.8200, 2.056)$. Now for Example (1), $.8510 - .7946 = .0564 > c = \tau_2 \cdot \sigma = (.8200)(.0222) = .0182$. Thus we select π_2 as the best with confidence level at 75%.

If the δ^* value stays at .05, but P_1^* is increased to .95, when we rerun the program, we will see the response on the screen "unable to find tau2—tau1 too small for P1star." Then we should either lower the P_1^* value or increase the τ_1 value by increasing the sample size n. The latter one will be illustrated in the next example.

EXAMPLE (4): This example shows how Table I can be helpful to design the selection experiment. If the evaluator wishes to select the best system among 4 competing systems

with confidence level .90 whenever the distance δ^* between the true performance measures of the best and the second best is at least .05, or otherwise to select a subset that contains the best system also with confidence level .90. Based on the past experience, the evaluator may require that the competing systems produce at least .90 for the $P(D)$ and at most .01 for the $P(FA)$ (see for example, Figure 3.3 NNIIES Operating Curve in [6]). Then a conservative upper bound for σ^2 can be estimated as follows:

$$\begin{aligned}
\sigma^2 = \mathrm{Var}(V) + \mathrm{Var}(W) &< (.90)(.10)/n + (.01)(.99)/m \\
&= .09/n + .0099/m \qquad \text{where } n \text{ is the number of targets and } m \text{ is the total number of false alarm opportunities.} \\
&< .0999/n \qquad \text{where we assume that } n << m.
\end{aligned}$$

From Table I, the procedure parameter (τ_2, τ_3) exists when τ_1 is fixed at 4. Thus as long as $n > (4)^2(.0999)/(.05)^2 = 639.36$ we get $\tau_1 = \frac{\delta^*}{\sigma/\sqrt{n}} > 4$. Hence we can guarantee that our selection procedure with procedure parameter $(\tau_2, \tau_3) = (1.54, 2.46)$ will achieve the goal.

Table 1

The procedure parameters (τ_2, τ_3) for given
k, τ_1, P_1^*, and P_2^*

k=2

$\tau_1 = 3$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(2.04, 0.)	(1.18, 0.)	*
.90	(2.04, 0.)	*	*
.95	*	*	*

$\tau_1 = 4$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(3.04, 0.)	(2.18, 0.)	(1.67, 0.)
.90	(3.04, 0.)	(2.18, 0.)	*
.95	(3.04, 0.)	*	*

$\tau_1 = 5$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(4.04, 0.)	(3.18, 0.)	(2.67, 0.)
.90	(4.04, 0.)	(3.18, 0.)	(2.67, 0.)
.95	(4.04, 0.)	(3.18, 0.)	(2.67, 0.)

$k = 3$

$\tau_1 = 3$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(1.56, .67)	*	*
.90	(1.56, 3.64)	*	*
.95	*	*	*

$\tau_1 = 4$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(2.56, .36)	(1.76, .55)	(1.28, .98)
.90	(2.56, 1.21)	(1.76, 1.86)	*
.95	(2.56, 1.77)	*	*

$\tau_1 = 5$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(3.56, .32)	(2.76, .35)	(2.28, .40)
.90	(3.56, 1.10)	(2.76, 1.17)	(2.28, 1.30)
.95	(3.56, 1.58)	(2.76, 1.69)	(2.28, 2.20)

Table 1 (continued)

$k = 4$

$\tau_1 = 3$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(1.31, 1.28)	*	*
.90	*	*	*
.95	*	*	*

$\tau_1 = 4$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(2.31, .86)	(1.54, 1.09)	(1.08, 1.66)
.90	(2.31, 1.66)	(1.54, 2.46)	*
.95	(2.31, 2.20)	*	*

$\tau_1 = 5$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(3.31, .32)	(2.54, .84)	(2.08, .90)
.90	(3.31, 1.56)	(2.54, 1.61)	(2.08, 1.74)
.95	(3.31, 2.00)	(2.54, 2.10)	(2.08, 2.39)

$k = 5$

$\tau_1 = 3$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(1.15, 1.65)	*	*
.90	*	*	*
.95	*	*	*

$\tau_1 = 4$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(2.15, 1.14)	(1.39, 1.39)	(.94, 2.20)
.90	(2.15, 1.91)	(1.39, 3.01)	*
.95	(2.15, 2.44)	*	*

$\tau_1 = 5$

$P_2^* \backslash P_1^*$	.75	.90	.95
.75	(3.15, 1.09)	(2.39, 1.10)	(1.94, 1.17)
.90	(3.15, 1.81)	(2.39, 1.86)	(1.94, 1.99)
.95	(3.15, 2.24)	(2.39, 2.34)	(1.94, 2.64)

APPENDIX

We need a lemma of Mahamunulu [8] to prove the theorems 3.1 and 3.2. For $\theta \in \Theta$, let $G_n(x|\theta) = F_\theta(x)$ be a stochastically increasing (SI) family of distribution functions on the real line and let $X_1, X_2, \ldots, X_k$ be independent random variables where the distribution function of X_i is $G_n(x_i|\theta_i)$ $(i = 1, 2, \ldots, k)$.

Lemma (Mahamunulu). *For each i $(i = 1, 2, \ldots, k)$ if $\psi = \psi(X_1, X_2, \ldots, X_k)$ is a non-decreasing function of X_i when all X_j for $j \neq i$ are held fixed, then $E[\psi(X_1, X_2, \ldots, X_k)]$ is a non-decreasing function of θ_i.*

Theorem 3.1. *Under procedure R the LFC for $P(CD_1)$ in the PZ is the configuration as given in (3.2).*

Proof. We start with an arbitrary configuration $\underset{\sim}{\mu}$ in the PZ

$$\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k]} \text{ with } \mu_{[k]} - \mu_{[k-1]} \geq \delta^*. \tag{A.1}$$

Letting $X_{(i)}$ denote the sample mean associated with $\mu_{[k]}$, we have

$$\begin{aligned} P(CD_1|PZ) &= P\{\text{selecting the population associated with } \mu_{[k]}\} \\ &= P\{X_{(k)} > \max_{1 \leq i \leq k-1} X_{(i)} + c\}. \end{aligned} \tag{A.2}$$

Define the function $\psi = \psi(y_1, y_2, \ldots, y_k)$ by

$$\psi = \begin{cases} 1 & \text{if } X_{(k)} > \max_{1 \leq i \leq k-1} X_{(i)} + c \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

Then by (A.1) it follows that $P(CD_1|PZ) = E\{\psi(X_{(1)}, X_{(2)}, \ldots, X_{(k)})\}$. Moreover it is clear from the definition of ϕ that $\psi(y_1, y_2, \ldots, y_k)$ is non-increasing in y_i for $i = 1, 2, \ldots, k-1$ when $y_1 (j \neq i)$ is held fixed and it is non-decreasing in y_k when all the y_i for $i = 1, 2, \ldots, k-1$ are held fixed. Since the $X_{(i)}$ are from an SI family we can use the above lemma to conclude that the $P(CD_1|PZ)$

(1) is non-increasing in $\mu_{[i]}$ for $i = 1, 2, \ldots, k-1$, and

(2) is non-decreasing in $\mu_{[k]}$ for $j = k$. It follows that the LFC for $P(CD_1|PZ)$ is given by the configuration (3.2). This completes the proof of the theorem.

Theorem 3.2. *Under procedure R, the WC for the $P(CD_2|IZ)$ is the configuration as given in (3.3).*

Proof. We start with an arbitrary configuration μ in the IZ:

$$\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k]} \text{ with } \mu_{[k]} - \mu_{[k-1]} < \delta^* \tag{A.4}$$

Let $M_i = \max\{y_\alpha \quad (\alpha = 1, 2, \ldots, k-1; \alpha \neq i)\}$ and $M_0 = \max\{y_\alpha (\alpha = 1, 2, \ldots, k-1)\}$. We define the function $\psi = \psi(y_1, y_2, \ldots, y_k)$ by

$$\psi = \begin{cases} 1 & \text{if } y_k \geq M_0 \\ 1 & \text{if } M_i \leq y_k \leq y_i \leq y + c \text{ for } i = 1, 2, \ldots, k-1 \\ 1 & \text{if } M_i - d \leq y_k \leq M_i \leq y_i \leq M_i + c \text{ for } i = 1, 2, \ldots, k-1 \\ 0 & \text{otherwise} \end{cases} \tag{A.5}$$

If we replace the y_i in (A.5) by $X_(i)$, then by the definitions of CD_2 and the function ψ we have $P(CD_2|IZ) = E\{\psi(X_{(1)}, X_{(2)}, \ldots, X_{(k)})\}$. It can be easily checked that $\psi = \psi(X_{(1)}, X_{(2)}, \ldots, X_{(k)})$ is non-decreasing in $X_{(i)}$ when all $X_{(j)}$ for $j \neq i$ are held fixed $(i = 1, 2, \ldots, k)$ and also that it is non-decreasing in $X_{(k)}$ when all $X_{(j)}$ for $j \neq k$ are held fixed. Again by the lemma, we conclude that $P(CD_2|IZ)$

(1) is non-increasing in $\mu_{[i]}$ for $i = 1, 2, \ldots, k-1$, and

(2) is non-decreasing in $\mu_{[k]}$. It follows that for $t = 1$ the WC for $P(CD_2|IZ)$ occurs when the parameters are all equal as given in (3.3). This completes the proof of the theorem.

REFERENCES

[1] Sanders, A. R. Exploitation Evaluation Test Bed. RADC-TR-88- 241.

[2] Shelton, C. and Sadjadi, F. Automated Imagery Exploitation Evaluation. RADC-Tr-90-292.

[3] Gibbons, J. D., Olkin, I., and Sobel, M. (1977). Selecting and Ordering Populations–A new Statistical Methodology. Wiley, New York.

[4] Gupta, S. S. and Panchapakesan, S. (1979). Multiple Decision Procedures, Wiley, New York.

[5] Chen, P. and Sobel, M. (1987). An Integrated Formulation For Selecting The t Best of K Normal Populations. Communications in Statistics, Theory and Methods, 16, 1, 121-146.

[6] Neural Network Investigation Final Report (1991), Image Exploitation Branch, Rome Laboratory.

[7] ATRWG, Target Recognizer Definition And Performance Measures, ATRWG No. 86-001.

[8] Mahamunulu, D. M. (1967) Some Fixed Sample Ranking And Selection Problems. Annals of Mathematical Statistics, 38, 1079- 1091.

[9] Johnson, N. I. and Kotz, S. (1970) Continuous Univariate Distributions-1, Houghton Mifflin Company, Boston.

[10] Dahlquist, G. and Björck, øa (1974) Numerical Methods, Prentice-Hall, Inc., New Jersey.

PHOTONICS TECHNOLOGY DEVELOPMENT AT ROME LABORATORY

Richard L. Fork
Professor
Leonard Lyon
Sean McCaul
Students
Physics Department

Rensselaer Polytechnic Institute
Troy, NY 12180

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

September 1992

Photonics Technology Development at Rome Laboratory

Richard L. Fork
Professor
Leonard Lyon
Sean McCaul
Students
Department of Physics
Rensselaer Polytechnic Institute

Abstract

Six projects relevant to photonics, high transmission rates, and optical processing were developed: the achieving of lasing action using forsterite crystals; designing, ordering components for, and commencing construction of a figure eight optical fiber laser; the measurement of stimulated brillouin backscattering in optical fiber; the alignment of an external cavity semiconductor laser; setting up a demonstration unit for video signals over optical fibers; construction of a soldering station for bonding leads for a semiconductor laser; and the development of educational modules on fusion splicing, calculating loss through a fusion splice, and optimizing coupling laser emission into a bare fiber.

PHOTONICS TECHNOLOGY DEVELOPMENT AT ROME LABORATORY

Richard L. Fork
Leonard Lyon
Sean McCaul

INTRODUCTION

The summer faculty research participant and two students engaged in six projects related to photonics, high transmission rates, and optical processing at Rome Laboratory. These projects were: the achieving of lasing action using forsterite crystals; designing, ordering components for, and commencing construction of a figure eight optical fiber laser; the measurement of stimulated brillouin backscattering in optical fiber; the alignment of an external cavity semiconductor laser; setting up a demonstration unit for video signals over optical fibers; construction of a soldering station for bonding leads for a semiconductor laser; and the development of educational modules on fusion splicing, calculating loss through a fusion splice, and optimizing coupling laser emission into a bare optical fiber. Each of these projects will be discussed in the order listed above.

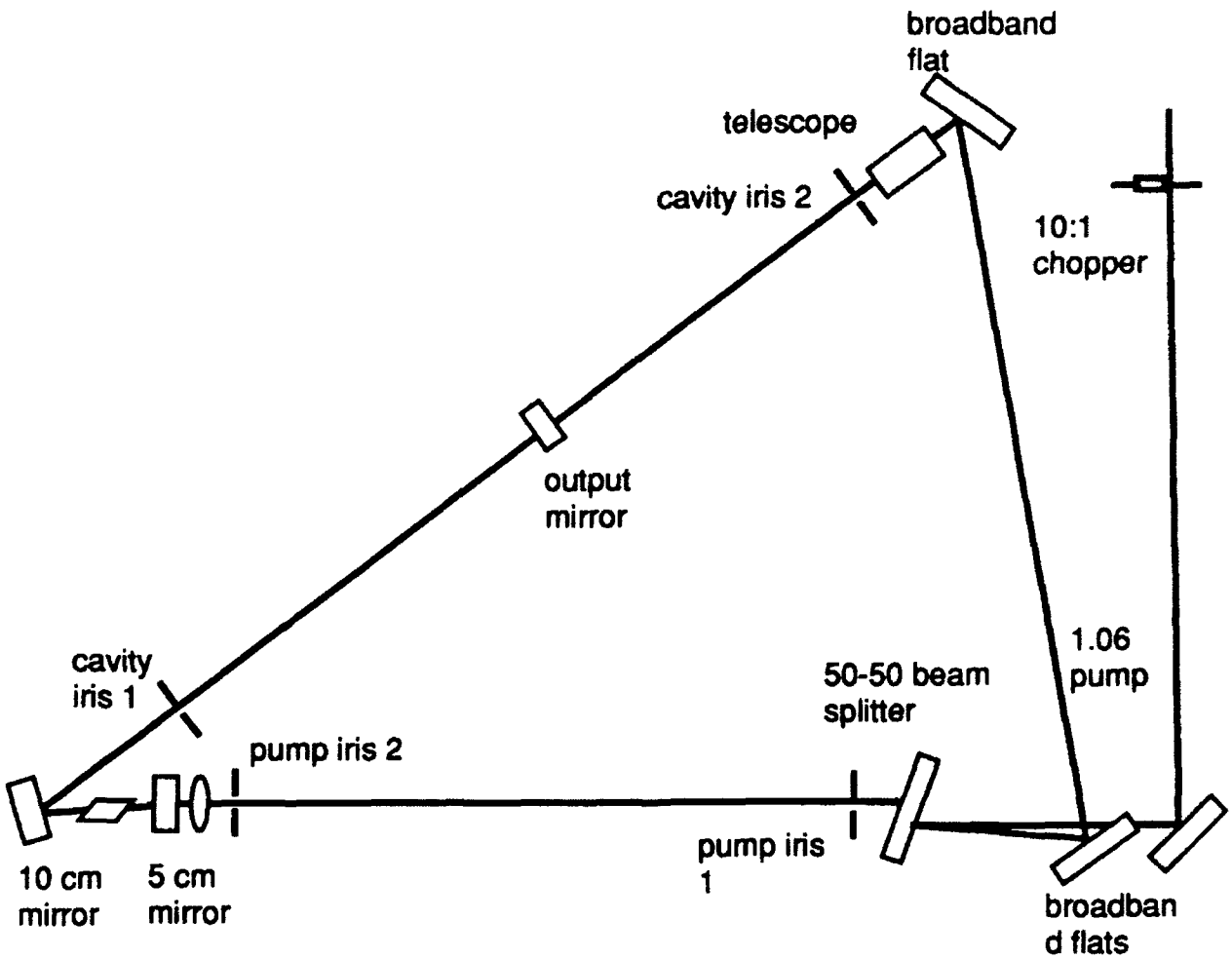
ACHIEVING LASING ACTION USING FORSTERITE CRYSTALS

For a fosterite laser to exhibit lasing action, it is necessary to take certain key steps. The two basic problems are: (1) making the resonator axis for the 1.3 micron radiation and the 1.06 micron pump axis coincident, and (2) obtaining a diagnostic that will guide one from a nearly aligned condition to the above threshold condition. See the figure on page 3-5.

The problem in making the resonator and the pump axis coincident is that the 5 cm mirror through which the pump enters will typically have a "wedged" section where the pump enters. The curved front face and the flat rear face will be parallel only at one location and in general the beam will not be passing through that location. Also, this wedge will change with the position where the pump beam is incident on the mirror. The diagnostic for aligning the pump axis and the cavity axis should provide some way of resolving this problem.

The He-Ne laser emission at 0.633 microns and the1.3 micron emission are useful only for coarse alignment since they will be refracted differently as compared with the 1.06 micron pump on passing through the wedged 5 cm mirror. The lineup strategy that is most reliable is one based on the 1.06 micron pump light. It happens

Lineup of Forsterite Laser



that the 1.3 mirrors are both sufficiently transmissive and sufficiently reflective that one can use them both in transmission and reflection with the 1.06 micron light.

This makes it possible to use the strategy illustrated in the diagram on page 3-5. It is helpful to also have the collimated single mode 1.3 micron light to set up the initial resonator, since the length of the resonator can be properly set and the high reflectivity of the curved mirrors at 1.3 provides an easily observed diagnostic beam.

The basic idea is to create a closed ring structure for the 1.06 light that includes the pump axis both *internally and externally* to the 1.3 micron cavity. One can satisfy this condition by requiring that the 1.06 micron light complete a closed loop around the ring structure. Once the closed loop is established, the path internal to the laser resonator defines the axis for the oscillator and the path outside defines the correct path for the pump light. Only the 1.06 micron light will have this property for any position on the 5 cm mirror.

The spatial positioning of the forsterite crystal within the resonator can be done coarsely by using the infrared viewing card; however, the focused spot is diffuse on the card, so it is difficult to evaluate the quality of the alignment. For some future system it would be much easier if the forsterite crystal was mounted on an adjustable platform. One could then use micrometer adjusts to find

the mid position in all three dimensions.

The strategy used to place the crystal on axis and to establish a uniform set of positions for the laser elements was to establish a plane parallel to the plane of the laser table (the breadboard on top of the optical table). This was done by starting with no refractive elements in the system and requiring the unfocused He-Ne beam to pass through the crystal and to also be parallel to the laser table surface. Some of the He-Ne beam leaks around the crystal, so it is useful to have a block of some kind that prevents the leakage. It is then easy to adjust the He-Ne beam so that it is centered in the crystal by requiring that the transmitted beam symmetrically fill the aperture presented by the crystal. (The He-Ne beam has expanded enough at this point that it slightly overfills the crystal. This makes it easy to find the centered condition.) One then adjusts all four irises to have their apertures at the same height as the forsterite crystal center and consequently to lie in a common plane that is parallel to the laser breadboard and that also passes through the center of the crystal. One must work back and forth with the iris position and He-Ne beam positioning by the two input mirrors until one positions the irises and the beam and the crystal all in a common plane. Three or four iterations are usually sufficient.

The 1.06 light is divided into two beams by a 50-50 beam splitter at the entrance to the laser table. The beam splitter is

oriented near normal incidence so that it shifts the beam position only by a negligible amount. The two irises that define the pump beam entrance path (pump irises 1 and 2) are then used to define the input path for the 1.06 micron beam. One can check with an ir viewer that the beam is centered on the crystal by requiring the transmitted pattern symmetrically fill the crystal. The 5 cm mirror and the focusing lens for the pump are then introduced. It is preferable to introduce these one at a time, requiring in each case that they be approximately centered on the 1.06 beam. There is no vertical adjust on the 5 cm mirror, but it is designed to be centered close to the height of the crystal center. The combination of the lens and the 5 cm mirror should be adjusted to collimate the 1.06 beam and to direct the beam along the axis previously defined by the He-Ne.

The spacing of the 5 cm mirror and the 10 cm mirror is rather critical. This spacing needs to be correct within a distance less than about 1 mm. This constraint arises from the stability condition on these cavities. The amount by which this distance is less than 1 mm depends on how well the astigmatism is compensated. The worse the astigmatism, the smaller the distance. If the astigmatism is severe enough, there may be no condition that will yield lasing. In general, the formed position is at a spacing slightly greater than the confocal condition. This distance can be very small so it is important to try to

closely approximate the correct condition.

As a means of determining the astigmatism has been minimized, the optimum angle of incidence at the first spherical mirror was calculated. The value agreed with previously recommended information for a 1 cm crystal, assuming an index of 1.6 for the crystal. The angle was set close to the calculated angle and the beam quality was found to be very good for one pass through the crystal.

If one examines, however, the return beam reflected from the 5 cm mirror, it shows significant residual uncorrected astigmatism. Since lasing was achieved and the lasing mode was found to be of good quality, it was decided not to pursue this problem further. This observation should, however, be borne in mind in future designs.

The dependence of the astigmatism on the position of the crystal between the two curved mirrors was also explored. The degree of astigmatism changes slightly with crystal position, but probably not enough to be important. The crystal was set at the midpoint between the mirrors as measured by a ruler using the 12.2 mm correction previously determined for the offset of the mirror from the front face. This gives an astigmatism roughly intermediate among the various patterns obtained on translating the crystal and is probably as good a position as any. It is predicted the ability to achieve lasing would not be especially sensitive to the crystal position.

The 1.06 light that passes through the lens and the 5 cm mirror should be reflected by the 10 cm mirror at an angle of 34 degrees, as is the laser oscillator beam. The other two irises (cavity iris 1 and cavity iris 2) are positioned along this beam, and the 10 cm mirror is used to send the beam at the height of the two irises. This then places the pump beam, the cavity axis, and the crystal center all in a common plane parallel to the laser table. This is not absolutely necessary, but makes the alignment procedure much easier.

To set the curved mirror spacing and alignment, the 1.3 micron laser was located as shown in the figure on page 3-5, and collimated by the telescope. The 1.3 micron light was brought through the two cavity irises and reflected by 10 cm mirror. The return beam could be monitored by slightly offsetting it from the incident beam or by observing the beam picked off by a beamsplitter. One adjusts the cavity spacing to cause the return beam to be slightly converging. Thus ensures approximately the correct cavity spacing. This could also be accomplished with the 1.06 beam; however, for a first attempt the 1.3 beam was preferred since one knows the path through the crystal is the same as for the lasing action.

The output mirror is introduced and oriented so that the 1.06 beam is reflected back on itself through the cavity iris 1. The detector is located after the output mirror and the output displayed

on the oscilloscope. The input beam is chopped by a 10:1 chopper to reduce heating. The input lens is adjusted slightly to maximize the detected signal. The 5 cm mirror is also adjusted slightly. At this point, small oscillations on the leading edge of the super luminescence are typically observed using a detector. This appears to be a type of "Q-switching" of the super luminescence that occurs if the 5 cm mirror is aligned to send the luminescence back on itself. One can tell if one has this condition, since the oscillations on the leading edge of the emission are quite sensitive to orientation of the mirror. The oscillations are small, so it is desirable to expand the oscilloscope display to exhibit them as clearly as possible.

One can then adjust the 10 cm mirror slightly and this should bring the resonator over threshold. The oscillations on the leading edge of the super luminescence provide a guide. Larger oscillations tend to mean better cavity alignment.

DESIGN AND CONSTRUCTION OF FIGURE EIGHT LASER

In recent years, erbium doped fiber lasers have attracted a great deal of interest. These lasers have several advantages that make them superior for a number of applications, such as communications, computing and optical probing.

The development of a figure eight laser began with a discussion

of the optimum setup. A decision was made to develop a fiber eight laser similar to the one constructed by Dr. Hercules Avramopoulos (member Technical Staff, AT&T Bell Laboratories). The major difference in the setup was due to the difference in application. Both setups include two fiber loops. The major difference is a plan to add, in the Rome Laboratory figure eight laser, a Fabry Perot interferometer.

During the summer, the following parts were ordered for the construction of the figure eight fiber laser: 50-50 coupler; 10-90 coupler; 30-70 coupler; and a 980 1550 nanometer WOM (wavelength sensitive). The latter will allow power to be taken from the Ti:sapphire laser at the laboratory, which operates at 980 nm, and pump it through the fiber. Two beam expanders, an interference filter, two polarization controllers, and fiber isolators (which controls direction of pulse propagation), were also ordered. By the end of the summer, all the equipment had arrived. The laser was constructed and current work is addressing the task of making it operational.

MEASUREMENT OF STIMULATED BRILLOUIN BACKSCATTERING IN OPTICAL FIBER

Polarization preserving fiber was excited by 500 milliwatts YAG laser power. The input power was varied using a free space

attenuator. The amount of transmission and reflection at different input powers was measured. Data was correlated. It was possible to determine a threshold value for stimulated brillouin backscattering. At the threshold value the scattering exhibited a nonlinear pump dependence.

The main problems in this project were the high threshold value for the stimulated brillouin scattering, and difficulty in coupling power into the fiber to see the effects. Considerable effort was required to eliminate a number of potential sources of error; however, useful data was successfully obtained.

ALIGNMENT OF EXTERNAL CAVITY SEMICONDUCTOR LASER

The alignment of the external cavity semiconductor laser was accomplished in the following manner: An external cavity laser was constructed. A semiconductor chip was hooked up to a laser diode driver. A pair of twenty power objectives were used to focus and recollimate the light coming from the chip. Two mirrors were placed on each side of the objectives, to form the laser cavity.

After aligning the system, the laser was actively mode locked, and the pulses measured to be approximately 70 ps. In addition, the back focal mirror was replaced with a diffraction grating to tune the laser. This was accomplished to some extent, providing 400 nm at tunable bandwidth.

SETTING UP A DEMONSTRATION UNIT FOR VIDEO SIGNALS OVER OPTICAL FIBERS

The setting up of a demonstration unit for video signals over optical fibers was begun. A two channel network, with optical fiber was planned. The network consisted of a optical switch that switch back and forth between two channels. A problem encountered was that the signal coming from a video camera is multimode, and single mode fiber was being used in the experiment. Another problem was that it was not possible to get enough power through the system. Because of all the optical components, about a 23 dB loss was experienced. Additionally, a high frequency electromagnetic pulse from another source on base interfered with the data. It was believed to be radar emanating from another base location.

CONSTRUCTION OF A SOLDERING STATION FOR BONDING LEADS FOR A SEMICONDUCTOR LASER

Construction of a semiconductor chip mounting/soldering station was undertaken. This entailed construction of a pair of vacuum tweezers that could pick up a semiconductor chip, and placing it in position. The station itself worked well, but it was not possible to spread the solder thin enough on the chuck. This caused

the solder to creep up the side of the chip when the system started to cool. This process inevitably shorted out the chip, and consequently the laser would not function.

The problem was determined to lie not in the design of the station or the tweezers but rather it was found that the brass itself was too rough. Precision made pieces of brass will be needed that will be smoother on top.

DEVELOPMENT OF EDUCATIONAL MODULES

The development of three educational modules for upper level undergraduate and first year graduate students studying the field of optics was begun. Preliminary text was written for the following modules based on experiences of students at Rome Laboratory during the summer of 1992: "Fusion Splicing," will provide students with a working knowledge of an Alcoa Fujikura Fusion splicer and simple splicing techniques. "Calculating Loss through a Fusion Splicer," discusses basic techniques of calculating optical path loss and use of power, power meters, pigtailed lasers and fusion splicers. "Optimizing a Laser into a Bare Fiber," describes how to couple a free space beam into an optical fiber. The plan is to take these short descriptions of important laboratory techniques, expand them with

more text, and add video, graphical and possibly computer generated animations to assist students who would like to learn the techniques discussed.

CONCLUSION

In conclusion, six projects were undertaken during the summer of 1992 at the Photonics Laboratory at Griffiss Air Force Base. Both the faculty and student participants found the summer program to be an extremely valuable experience. Much was learned to further photonic technology development at Rome Laboratory, and collaboration between the participants is expected to continue.
more text, and add video, graphical and possibly computer generated animations to assist students who would like to learn the techniques discussed.

ISSUES IN ADAPTIVE FAULT MANAGEMENT FOR SURVIVABLE C^3I SYSTEMS

Rex E. Gantenbein
Associate Professor
Department of Computer Science

University of Wyoming
P.O. Box 3682
Laramie, Wyoming 82071-3682 USA
Internet: rex@corral.uwyo.edu

Final Report for:
AFOSR Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

September 1992

ISSUES IN ADAPTIVE FAULT MANAGEMENT FOR SURVIVABLE C^3I SYSTEMS

Rex E. Gantenbein
Associate Professor
Department of Computer Science
University of Wyoming

ABSTRACT

Most strategies for fault management are effective for a narrow range of fault classes. In survivable distributed systems, a wide range of operating environments may be encountered that require different strategies to be used at different times. This report discusses how adaptive fault management can be used to select the most appropriate methods for assuring survivability under conditions that can suddenly and drastically change.

A model for adaptive fault management is presented that outlines the major research issues: managing adaptivity, modeling the strategies, and evaluating the system behavior in a dynamic environment. A taxonomy for fault management mechanisms that establishes three basic classes of strategies is defined as a basis for modeling these mechanisms. A generalized metric by which the effectiveness of a fault management strategy can be measured against system requirements -- the objective function -- is also defined. This metric can be used to determine how well a survivable system meets its requirements in the current operating environment. The research being carried out into this approach at Rome Laboratory and at Wyoming is described. Finally, some suggestions for future research are given, along with a schema that relates these topics to the current work.

ISSUES IN ADAPTIVE FAULT MANAGEMENT FOR SURVIVABLE C^3I SYSTEMS

Rex E. Gantenbein

INTRODUCTION

A complex system that can justifiably be trusted to deliver its required services when needed must be designed to meet and handle a wide variety of potential problems in the course of its operation. Unfortunately, most strategies intended to provide dependability consider only a few very specific problems. The assumptions that are made about the environment (often to make the proposed solution tractable or efficient) can limit the effectiveness of a method. An effective strategy in one case may be ineffective and even harmful in another.

The result of this optimization for a particular operating environment is that rapid or drastic change, as is sometimes seen in a military command, control, communication, and intelligence (C^3I) system under field-of-combat conditions, can cause catastrophic failure in the system. On the other hand, a system that is designed for a combat environment may not be effective or efficient in "normal" (noncombat) usage. Even the mission of a C^3I system may change over time (from strategic planning to theater operation to tactical battle management, for example), and strategies that are appropriate for one mission may not be for another.

One way in which a system can remain effective in a rapidly changing environment is to be *adaptive*, that is, able to change fault management mechanisms or modify the parameters of a mechanism to respond to changes in the system's behavior, environment, or requirements. Static fault management uses a single

mechanism to respond to errors in the system; adaptive fault management allows the system to bring into play any of a number of alternatives, depending on the nature and severity of the error(s) encountered and the environment in which the error(s) occurred. Under an adaptive paradigm, shown in Figure 1, the system behavior indicated by the system's internal state is continually evaluated in the context of its external state (i.e., the operating environment) and its requirements. When an error (a variation from the required behavior) is detected, the system determines whether static or adaptive recovery is most appropriate and reponds to the error accordingly.

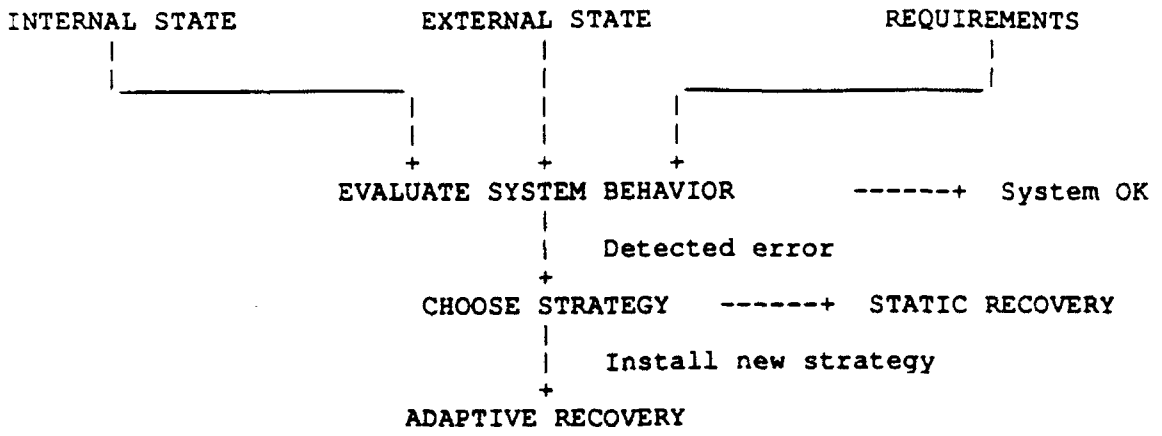


Figure 1. Adaptive vs. static recovery in a system.

A MODEL FOR ADAPTIVE FAULT MANAGEMENT

Our interest in adaptive fault management is motivated by its potential for increasing the *survivability* of C^3I systems. Survivable systems "must continue to perform adequately in the face of various levels of adversity" [Neumann92], as can be encountered in field deployment. Previous work has shown that adaptive fault

management may be able to achieve greater resilience to failures under changing environments [Armstrong91]. Certainly the ability to evaluate system behavior in the context of both the environment and the requirements -- and to change strategies if the existing strategy is not effective -- should decrease a system's susceptibility to changes in the environment.

In the adaptive fault management paradigm, this evaluation and selection of strategies is done dynamically, based on information about the system state. As shown in Figure 2, adaptive fault management utilizes an adaptation data base that contains information about the mechanisms and strategies that may be employed in the computing system. This information consists of a rule base that supports decisions about the appropriateness of each available strategy and characterizations of the mechanisms that can be used to implement them.

At the same time, information about the current external state (including operation mode and mission) and the current internal state (including error rate, fault type, resources available, resource utilization, and workload) can be collected to help the system understand its current operating environment. Using the rule base to relate this state information to the requirements for the system, the *adaptive behavior manager* selects the fault management strategy that can best maintain the required behavior in the current environment. This strategy is used in scheduling of system services and in determining the parameters for the execution of the fault management mechanisms.

This model highlights the major research issues in adaptive fault management. If a system can evaluate the current environment and match a fault management stra-

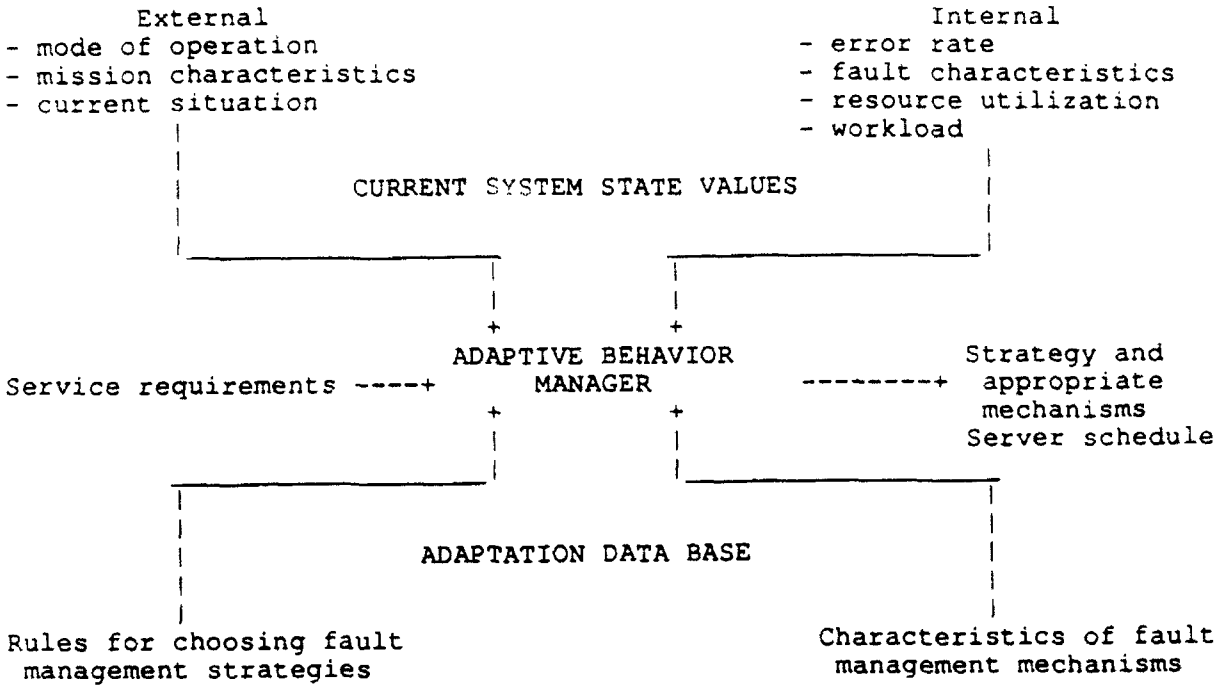


Figure 2. A model for adaptive fault management.

tegy to that environment, then the system is more likely to survive drastic environmental changes. For this approach to succeed, however, we must be able to characterize the environment and the fault management strategies in a manner that can be efficiently instrumented and computed at run time. We must characterize potentially useful strategies so that their effectiveness in a given environment can be evaluated. We must also find a way to represent the relationship between the system behavior and the requirements so that we can determine when a failure occurs or when a change in strategy is appropriate. We address these issues in the following sections of this report.

CHARACTERIZATIONS OF FAULT MANAGEMENT STRATEGIES

Various fault management strategies can be used to resist failures when faults are

present in a C^3I system. The adaptive fault management paradigm suggests that adaptation to a new strategy is advisable whenever the current environment no longer conforms to the assumptions of the current strategy. However, we need to be able to define the assumptions about the environment that are part of each strategy used in the system.

We have begun this study by defining three general classes of fault management strategies [Gantenbein92]: *optimistic*, for environments in which the expected errors can be explicitly detected and handled without long-term effects on other requirements of the system; *pessimistic*, for environments in which the expected number, type, or frequency of errors is sufficient to require the allocation of system resources to fault management in order to avoid catastrophic failure; and *ultrapessimistic*, for environments in which the errors that have occurred require the relaxation or abandonment of requirements in order to preserve the critical services of the system.

In normal operation, a well-designed and carefully implemented system should experience a relatively low error rate. When errors do occur, the recovery time is short enough that system services can be suspended while an error-free state is recreated without significantly hampering the *progress*, or timely (as defined by the requirements) delivery of the expected services, in the system. Under these conditions, an optimistic fault management strategy devotes most of the system resources to providing services and uses information about the application and the requirements to explicitly detect errors.

In C^3I applications, the environment can undergo changes (both internally and externally) that affect the system's ability to make progress under an optimistic stra-

tegy. In such circumstances, pessimistic fault management strategies must be employed that devote system resources not only to error detection but to error avoidance as well, so that progress can still be made in the new, adverse environment. However, the resource overhead for a pessimistic strategy is significant, and it may not be the case that additional resources are available when the environment is no longer supportive of an optimistic strategy.

Furthermore, even if a pessimistic strategy can be effectively applied, there is no guarantee that the environment will not get worse. Under extremely harsh conditions, such as the loss of nodes or communication links resulting from sabotage or hostile action, a system that was making progress under a pessimistic strategy may no longer be able to do so. If the system is to survive, a third class of strategy, which we call ultrapessimistic, is required.

Most well-known fault management techniques fall into one of these three strategy classes, as shown in the taxonomy of Figure 3. This taxonomy allows us to broadly define strategies as well as show how they apply to the three basic resource classes in a distributed system [Kohler81]: processing (generating or transforming information), data (storing information), and communication (transmitting information). We classify individual techniques as optimistic, pessimistic, or ultrapessimistic, depending on the amount of *a priori* information and the resources needed to detect and respond to errors.

RELATING THE ENVIRONMENT TO THE REQUIREMENTS

The identification of these three categories of fault management strategies allows us to consider which techniques may be most effective when the surrounding environ-

	Classification Characteristics	Processing	Data Storage	Communication
1	**Optimistic:** devotes resources primarily to computation; needs *a priori* information for error detection; recovers state by recomputation	Standby versions: Cold hardware sparing, recovery blocks	transactions, checkpointing	retransmission
2	**Pessimistic:** devotes resources largely to error detection and response; needs little *a priori* information for either task; recovers state by masking	Multiple versions: N-modular redundancy, N-version programming, hybrids	replicated data, error-correcting codes	error-correcting codes, multicasting, redundant media
3	**Ultrapessimistic:** devotes resources to critical services; relaxes system requirements; requires *a priori* information for both error detection and response; recovers state by approximation	cold sparing without checkpoints	compensating transactions	network partitioning

Figure 3. A taxonomy for fault management strategies.

ment rapidly and drastically changes. However, it is also necessary that the system designer be able to specify the acceptable system behavior in a given environment, so we can evaluate the behavior to see whether the current fault management strategy is still effective. To do this, we must define an efficient, quantifiable metric that measures changes in the environment and their effect on the server's ability to deliver its service as required. This metric, the *objective function*, is a partial specification of a service that uses derived or observed attributes representing the current environment to compute a set of values that can be used to evaluate the service delivery in that environment [Gantenbein92].

The purpose of the objective function is to capture in a computable definition the expectations for the service provided by a server and the effect of the environment on

delivery of that service. The function must be able to express all of the *objectives* that may affect the server's dependability, including such things as:

- functionality, the extent of the delivery or accessibility of the expected service;
- performance, a measure of the timeliness of the delivery of the service;
- mutual consistency, the agreement among a set of servers on the contents of a resource or the attributes of a cooperatively provided service;
- internal consistency, the degree of correctness within a single server with respect to its view of a resource or service;
- precision, the latitude that will be accepted in the server's delivery of its expected service;
- security, a measure of the server's resilience to maliciously (as opposed to inadvertently) introduced faults; and
- safety, the requirements that deal with the risks associated with delivery of the service.

The values produced by the objective function represent a measure of how close the delivery of the service is to its requirements for each of the defined objectives. This measure is used by the adaptive behavior manager to decide when a new strategy is needed (i.e., when the current strategy is no longer acceptable) and which strategy to apply next (i.e., one that will give acceptable behavior according to the objective function). We can compute a set of objective function values, obtaining the function parameters from the internal and external states of the system (i.e., from the current environment). If these values match their expected values, as defined in the system design, then the server *meets its objectives* under the current environment.

Furthermore, this model allows us to define violations of the objectives and gives us a method of detecting when they occur. If the computed value of an objective does not match its expected value, then we say that an *anomaly* exists in that objective, indicating that the requirement has not been met. Most errors in a service can be modeled as anomalies of the objective function for that service. The advantage of our model is that we can consider a range of values for any objective and thus allow a tolerance in meeting it. This tolerance allows different strategies to be associated with different (possibly overlapping) ranges of objective function values.

Several examples of specifying these objectives in a computable function exist in the literature. Many of these are aimed at static or predictive measurement of a given objective, but it may be the case that they can be used to evaluate a system at run time as well. Obviously, more study is needed before good objective functions can be defined for any or all of the above objectives.

EXPERIMENTS IN SURVIVABLE DISTRIBUTED SYSTEMS

A number of examples exist of systems in which multiple fault management strategies are used to increase the system's resilience to a broad class of anomalies [Gantenbein92]. Our current research, as well as projects being carried out under the auspices of Rome Laboratory, is exploring ways in which this work can be extended into a framework that will increase our confidence in the claim that adaptivity can enhance the survivability of distributed C^3I systems. In this section of the report, we describe some of the background and plans for these experiments.

In-house projects

Survivable distributed C^3I systems are an important Air Force research topic. For example, a recent triservice project centered around a demonstration of cooperation and resource sharing in a C^3I system distributed over several geographically dispersed sites [Gadbois90]. The project showed the feasibility of using distributed computing to support a survivable application. This kind of work could be applied to other similar systems, such as the Joint Surveillance System Distributed Tracker [Leckie90].

The above work is being extended by the Distributed Systems Group of the Computer Systems (C3AB) branch of Rome Laboratory. This group is developing an adaptive, survivable C^3I system, with an demonstration to be given in early 1994. The primary consideration in this experiment is the development of a distributed behavior manager that can evaluate system behavior and determine the ability of the available fault management strategies to respond to variations in the behavior from the requirements [Craig92].

Other considerations in this experiment include modularity of the design (so that the adaptive and survivable components of the system are separated from the mechanisms themselves, making the technology portable to other applications) and scalability (so that increased performance can be achieved at the hardware level without affecting the system design). A related project is looking at the development of metrics and instrumentation tools for the static evaluation of the dependability of these systems. A mechanism for technology transfer of evaluation techniques from the Systems Reliability division of Rome Laboratory is being investigated to see if the techniques used in electronics design can apply to computer systems as well.

Several different testbeds have been used by Air Force researchers as underlying environments for survivability and adaptability experiments. The four most common choices have been ISIS [Birman85], Mach [Rashid89], Cronus [Schantz86], and Alpha [Jensen90]. At this time, Cronus and ISIS figure most prominently in the research plans of Rome Laboratory due to their support of distribution and dependability, coupled with their availability to government researchers. The current projects expect to make use of both testbeds; future considerations include combining the two into a single integrated environment (CrISIS) for dependable distributed system development.

Related research at the University of Wyoming

A pilot study funded by the AFOSR Research Initiation Program, is currently under way at Wyoming investigating some of the related research issues in adaptive, survivable distributed systems. This study, which will be completed by December 1992, has two major foci:

- We are examining a large number of fault management mechanisms to categorize them into the three strategy classes discussed earlier. We will determine which of these categories are populated with well-defined techniques and which are not. This research will provide an understanding of the support for adaptive fault management that is currently available.
- We are investigating distributed testbeds with respect to their support for adaptivity and survivability. We have established a set of criteria for these testbeds and are investigating the mechanisms present in them on that basis. We plan to identify the tools that each testbed under consideration provides for development

and operation of adaptive systems.

These studies should not only lay the foundation for additional research (as described in a later section), but should also provide insight into the mechanisms available for fault management in the demonstration systems. Preliminary results have already been transferred to Rome Laboratory personnel, and copies of reports prepared at the end of the project will be made available as well. Support is being sought for the continued involvement of Wyoming researchers in this project, with the hope of continuing to support the goals of Rome Laboratory in investigating adaptivity and survivability.

Contracted projects

Recent contracts have been established with GE Aerospace and SRI for research into adaptive survivable systems. The GE research, which began in July 1992, will produce a realistic demonstration system based on previously developed C^3I systems and work from a previous contract for investigating adaptive fault tolerance [Armstrong91]. Their expertise in this area should provide this project with the background needed to complete their demonstration system within 18 months. Their main contribution will be the development of an intelligent behavior manager that is able to choose from among a large number of strategies using rule-based decision algorithms, and the mechanisms by which this choice can be efficiently installed in a running system.

The contract to SRI involves longer-term goals, including consideration of the real-time and resource management issues. Their work will lead to a better understanding of the system management problems related to adaptivity. GE has, in its

proposal, offered to organize, in conjunction with Rome Laboratory, a 1993 workshop on trends in adaptive fault management that should bring together ideas from all these projects, as well as from outside researchers with interest in the area.

FUTURE WORK

We have presented a model for adaptive fault management based around a taxonomy of fault management strategies into optimistic, pessimistic, and ultrapessimistic categories. While the proposed taxonomy encompasses many of the well-known fault management strategies for both software and hardware, most strategies fall into the optimistic or pessimistic categories. Very few strategies exist that utilize the relaxation of one or more objectives to avoid catastrophic failure. It is hoped that strategies exist that can be utilized under such conditions to enhance the survivability of C^3I systems.

Clearly, much more research is needed to make adaptivity a useful tool in the design and operation of survivable systems. First and foremost, the taxonomy categories must be formalized and more completely populated with known strategies to define how well we understand the various categories. It appears, based on preliminary work, that many techniques for optimistic and pessimistic fault management exist for processing. Data and communication have a large number of optimistic strategies, but fewer pessimistic. None of these areas have been extensively studied with respect to ultrapessimistic strategies.

Practical issues also need to be considered. The major problem with the use of objective functions is their definition in a complete, yet operationally tractable manner. We need to be able to define objective functions for system components in such a way that they can be evaluated efficiently. We must also consider how and

when to make the transition between strategies, which will involve evaluating the tradeoffs between various component objectives. This transition must not be limited to an optimistic-to-pessimistic-to-ultrapessimistic paradigm, since we must allow for repair of a failed component (and the associated transition in the opposite direction from degradation) as well as a catastrophic loss of resources under an optimistic strategy, which would necessitate a direct transition from an optimistic to an ultrapessimistic strategy. It is not yet clear how much of this can be done other than on an application-specific basis, although general principles should be definable.

Other considerations include how to choose the appropriate strategy from among those available. We need to find an efficient rule base and evaluation method for deciding among strategies in every system designed in this way. The rules for making such decisions will be complex and a general, or at least automatable, adaptive behavior manager will be complex as well. Furthermore, to avoid a single failure point in the manager, it will be necessary to reliably distribute this service (and others) over the system nodes and consider the effects of coincident errors on its stability.

A useful property of this taxonomy is that it can be applied to a problem to define the strategy classes that are best for a given set of circumstances. A design approach for survivable systems can be envisioned that uses these classifications, first, to guide how the system can be constructed to support adaptation among the strategies and, second, to define the classes of strategies that best fit all the environments that the system is expected to encounter (even if the likelihood of encountering any of them is small). It also appears that this approach can be used for subsystem design

and implementation as well. The primary questions here are how to integrate subsystems back into the larger system and how to control adaptivity among multiple concurrent applications in the same system. The relationships among processing, data, and communication are not likely to be orthogonal, so the specification and management of adaptivity at the lower levels will need to be considered carefully to allow the interaction among the subsystems to be managed.

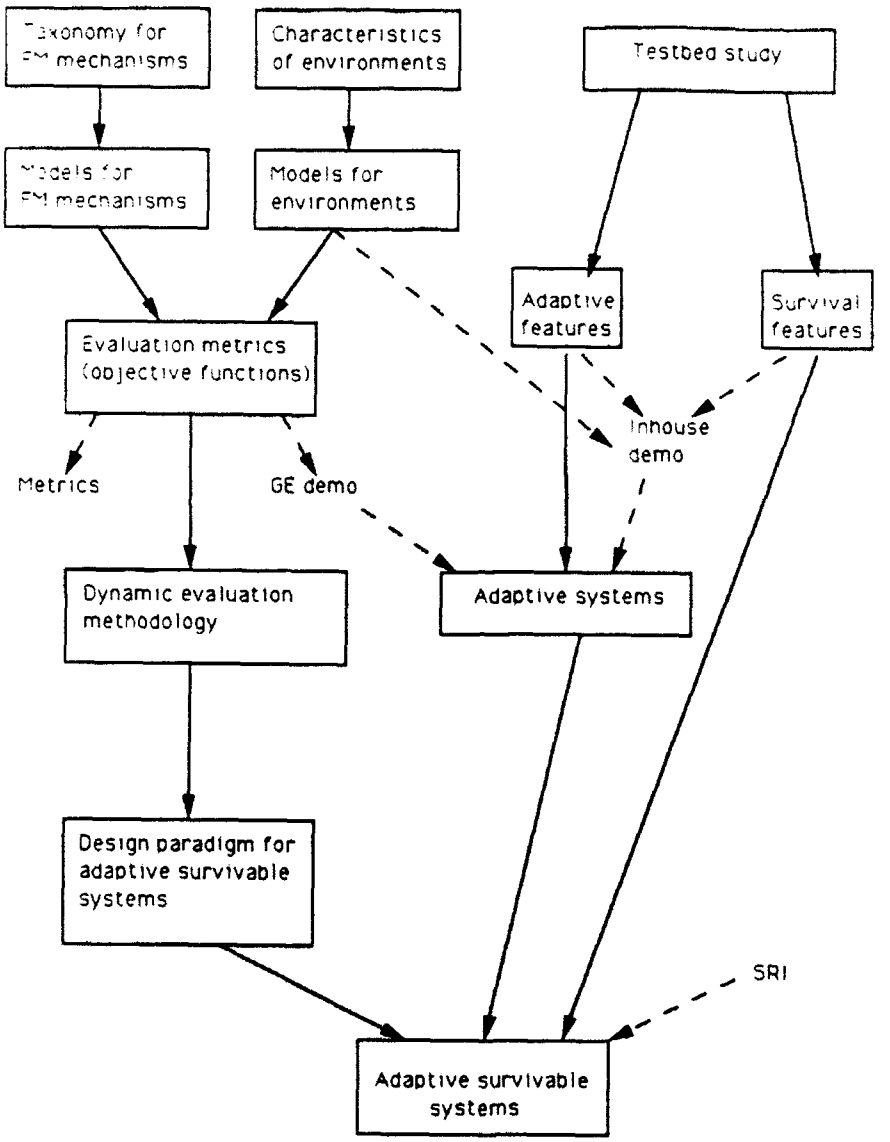


Figure 4. A research schema for adaptive, survivable systems.

The research projects described in this report, in combination with the topics sug-

gested above for future consideration, form a coherent thrust of research into adaptive, survivable systems. The relationships among these topics can be described by the research schema shown in Figure 4. In the short term, such research will help to build a theoretical framework for the application of adaptive fault management to survivable systems. In the long term, however, we see this research as leading towards a generalized methodology for the the dynamic redefinition of system behavior with respect to an operating environment that may change over time. Not only fault management but many other system strategies (for example, task scheduling or resource management) become less effective when the assumptions inherent to them are violated. Since computer-driven systems are likely to be used in a wider variety of application environments in the future, adaptivity is likely to make such systems not only more dependable but more flexible as well.

ACKNOWLEDGEMENTS

The author wishes to thank Waleed Smari and Dr. Gary Craig of Syracuse University and Mary Denz, Jerry Dussault, and Tom Lawrence of Rome Laboratory for their discussions of the ideas in this report. The views and opinions contained in this paper are those of the author and should not be construed as an official Department of Defense position, policy, or decision.

REFERENCES

[Armstrong91]
L.T. Armstrong and T.F. Lawrence, Adaptive Fault Tolerance, *Proc. 1991 Systems Design Synthesis Technology Workshop* (1991), Naval Surface Warfare Center.

[Birman85]
K.P. Birman, Replication and Fault-Tolerance in the ISIS System, *Proc. Tenth ACM Symp. on Operating Systems Principles, Operating Systems Review* 19,5 (1985), 79-86.

[Craig92]
G.L. Craig and V.T. Combs, Resource Management: Support for Survivable and Adaptable C3 Applications, *Proc. 1992 Mohawk Valley Command, Control, Communications, and Intelligence (C3I) Conference* (1992), 295-299.

[Gadbois90]
M.J. Gadbois and A.M. Newton, Triservice Distributed Technology Experiment, *Proc. 1990 Symp. on Command and Control Research* (1990), 150-157.

[Gantenbein92]
R.E. Gantenbein, T.F. Lawrence, and S.Y. Shin, Adaptive Fault Management in Survivable Distributed Systems (in preparation 1992).

[Jensen90]
E.D. Jensen and J.D. Northcutt, Alpha: A Non-Proprietary OS for Large, Complex, Distributed Real-Time Systems, *Proc. Second IEEE Workshop on Experimental Distributed Systems* (1990), IEEE Computer Society Press, 35-41.

[Kohler81]
W.H. Kohler, A Survey of Techniques for Synchronization and Recovery in Decentralized Computer Systems, *Computing Surveys* 13,2 (June 1981), 149-183.

[Leckie90]
R. Leckie and T.P. Humiston, *Joint Surveillance System Distributed Tracker* RADC-TR-90-288, Rome Laboratory (November 1990).

[Neumann92]
P.G. Neumann, Inside Risks: Survivable Systems, *Comm. ACM* 35,5 (May 1992), 130.

[Rashid89]
R. Rashid et al., Mach: A Foundation for System Software, *Proc. 1989 Workshop on Operating Systems for Mission-Critical Computing* (1989), P1-P5.

[Schantz86]
R.E. Schantz, R.H. Thomas, and G. Bono, The Architecture of the Cronus Distributed Operating System, *Proc. Sixth Int. Conference on Distributed Computing Systems* (1986), IEEE Computer Society Press, 250-259.

Atomistic Simulation of Grains in Submicron Aluminum Interconnects

Surendra K. Gupta
Associate Professor
Department of Mechanical Engineering
Rochester Institute of Technology
Rochester, NY 14623

Final Report for:
AFOSR Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

August, 1992

Atomistic Simulation of Grains in Submicron Aluminum Interconnects

Surendra K. Gupta
Associate Professor
Department of Mechanical Engineering
Rochester Institute of Technology

Abstract

A two-dimensional model to simulate motion of atoms in a submicron width aluminum line with bamboo grain structure is developed. The model is implemented on computer platforms supporting the VMS operating system. Real-time graphics is incorporated in the software using the UIS Graphics Library. The computational cell consists of two grains, one in the center and the other split in two halves. One-half is located at the left edge and the other half is located on the right edge so that the edges are properly aligned to create periodic boundary conditions. Atoms in the top and bottom row of subcells are fixed, and the interior atoms interact with one another following the Lennard-Jones potential energy function.

Simulations are performed for a cell consisting of approximately 2000 atoms with a time step of 0.01 pico seconds for 5000 iterations at T = 375°, 400°, 425°, 450°, 475° and 500°K. Total energy in each case is found to remain constant within 0.1%. Overall diffusion coefficients computed during each simulation are plotted against 1/T, and the activation energy for diffusion is found to be 1.292 kcal/mol-K.

Atomistic Simulation of Grains
in Submicron Aluminum Interconnects

Surendra K. Gupta

Introduction

Electromigration (EM), the current-driven transport of the material of a conductor subjected to a high electric current density, has recently attracted intense study because it has been found to limit the further miniaturization of integrated circuits. The metallization in integrated circuits, typically Aluminum and its alloys, carries current densities of the order of 10^{10} Am^{-2}, and under these conditions, metal ions migrate in the direction of electron flow. This can lead to the formation of voids at points in a stripe from which materials is depleted, or hillocks and whiskers at points where material accumulates. The metallization finally fails when voids sever a stripe or hillocks or whiskers cause short-circuits between stripes.

Researchers in the Reliability Physics Branch (ERDR) of Rome Laboratory have been investigating the electromigration phenomenon on several dimensional scales — from subatomic to macroscopic — both experimentally and theoretically. These investigations are divided into four aspects: (*i*) Electronic Structure Studies, (*ii*) Atomic Structure Studies, (*iii*) Void Growth and Microstructure Studies, and (*iv*) Joule Heating and Energy Transfer Studies. These four interrelated studies hope to provide a fundamental understanding of the electromigration phenomenon.

It has been recognized that there is a strong correlation between microstructure and electromigration lifetime in fine metal stripes, especially when the linewidths and thickness of interconnects are reduced to the submicron range thus becoming comparable to the grain

size. For a 1μm thick film, there are only one or two grains across a 1μm wide metal line. With only few grains across the metal line, the contribution of each grain boundary to the overall mass transport becomes more important. Each divergent site in the grain structure can fail without requiring a statistical linkage of damage originating from several divergent sites, as would be the case for a wide metal line with many grains across. This shortens the metal line lifetime and increases the randomness of the failure statistics.

In submicron width aluminum (Al) lines, the grain size is constrained by both the line thickness and width resulting in a columnar grain structure. When the film thickness and linewidth becomes smaller than the grain size, a bamboo grain structure develops. The lifetime of Al lines with such a structure is substantially longer than those of polycrystalline lines since most of the grain boundaries are perpendicular to the direction of current flow. Thus, grain structure and grain boundaries play an important role in electromigration resistance.

This research project represents the first phase of a detailed study of bamboo grain structures in submicron *width* Al *interconnects* by molecular dynamics (MD) simulations. Computer-based atomistic simulation is needed to gain better insight into the electromigration phenomenon because current experimental techniques do not reliably characterize very small voids (< 200 A°) that develop in such interconnects. The goal in this 10-week project was to develop software to perform two-dimensional atomistic simulations with real-time graphics. Initially, the software was developed for a MS-DOS based microcomputer. But due to internal compiler errors and limitations of graphics library, the software was rewritten for VAX/VMS based workstations and minicomputers using UIS Graphics Library.

Theoretical Principles & Numerical Procedures

The structure of grain boundaries has been widely investigated since grain boundaries play a major role in the control of such important phenomena as low and high temperature fracture, creep and corrosion, and electromigration. Unlike other defects such as dislocations, cavities and cracks, a grain boundary does not possess long range elastic stress and strain fields, and

its properties and effects on material behavior are determined by its atomic structure. Thus, continuum theories, which have been very useful in studies of a number of lattice defects, have only a very limited application, and atomistic studies of grain boundaries are essential. These studies have become possible with the advance of high-speed computers. Computers permit the minimization of an assembly of large number of atoms assuming an interatomic interaction.

Interatomic Interactions

The precursor of any atomistic calculation is the knowledge of interaction between the atoms. In the case of simple metals, central forces between the atom pairs have been assumed almost exclusively in atomistic studies, and successfully used in various defect studies. Lennard-Jones (LJ) potential has been found suitable for face centered cubic (fcc) metals with a convenient cut-off of the potential at third or between third and fourth nearest neighbors. LJ potential is a two-body potential energy function of the form:

$$v(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where ϵ is the bond energy at equilibrium ($\approx 6.3 \times 10^{-20}$ J for Al at 0°K), $\sigma = d_0/2^{1/6}$ with d_0 representing the equilibrium interatomic separation (for Al, $d_0 = 2.857 \times 10^{-10}$m), and r is atomic pair separation.

The individual pair forces $\vec{f}_{ij}$ between atom i and atom j given by

$$\vec{f}_{ij} = -\frac{\mathrm{d}v(r)}{\mathrm{d}r}\frac{\vec{r}}{r}$$

are accumulated to give the total force on each atom. In this evaluation, the interaction between a pair of atoms is ignored if their separation r is greater than a chosen cutoff radius r_c. To avoid long range corrections compensating the interactions for $r \geq r_c$, a shifted form of the LJ potential can also be used.

Equations of Motion

In a MD simulation, the initial velocities of all the atoms must be specified. It is usual to choose random velocities, with magnitudes conforming to the required temperature, corrected so that there is no overall momentum. In this project, the velocity components are chosen randomly from a Gaussian distribution conforming to the prescribed temperature and corrected to obtain zero momentum.

The most widely used method of integrating the equations of motion is that initially adopted by Verlet and attributed to Stormer. The basic Verlet scheme has several deficiencies, and modifications have been proposed to tackle these deficiencies. A Verlet-equivalent algorithm which stores positions, velocities, and accelerations all at the same time t recently proposed by Swope, Andersen, Berens and Wilson is implemented. The algorithm minimizes round-off error, and takes the form:

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t \;\; \vec{v}(t) + \frac{1}{2}\delta t^2 \;\; \vec{a}(t)$$

$$\vec{v}(t+\delta t) = \vec{v}(t) + \frac{1}{2}\delta t \;\; [\vec{a}(t) + \vec{a}(t+\delta t)]$$

where $\vec{v}$ and $\vec{a}$ are atomic velocity and acceleration respectively, and δt is the time step. In this form, the method resembles a three-value predictor-corrector algorithm where the position corrector coefficient is zero. The algorithm only requires storage of $\vec{r}$, $\vec{v}$ and $\vec{a}$ for each atom at a given instant t. It involves two stages with a force evaluation in between. First, the new positions at time $t + \delta t$ are calculated, and the velocities at mid-step are computed using

$$\vec{v}(t+\frac{1}{2}\delta t) = \vec{v}(t) + \frac{1}{2}\delta t \;\; \vec{a}(t)$$

Then, the forces (and therefore accelerations) at time $t + \delta t$ are computed, and the velocity move is completed by

$$\vec{v}(t+\delta t) = \vec{v}(t+\frac{1}{2}\delta t) + \frac{1}{2}\delta t \;\; \vec{a}(t+\delta t)$$

At this point, the kinetic energy at $t+\delta t$ becomes available while the potential energy at this time will have already been evaluated in the force loop.

Neighbor List

When the number of atoms is large, the logical testing of every pair in the system (to determine if $r \leq r_c$ for interaction) is inefficient. To keep track of neighbors, a cell index method is developed. The simulation box representing the system is divided into subcells. The length of edge of each subcell is chosen to be equal to or greater than the cut-off radius. The atoms contained in each subcell are indexed in a linked list at each time step. Thus, for the two-dimensional system implemented, if N_c is the approximate number of atoms in each subcell and N is the total number of atoms in the computational cell then only $10NN_c$ pairs are examined for interaction computations at each time step. This contrasts with N^2 examinations per time step with the brute force approach.

Boundary Conditions and Initial Configuration

The atoms in top and bottom row of subcells are held fixed. This assumption is based on the premise that atoms at the edges of stripe are constrained by the passivating layer. By choosing atoms in an entire row of subcells, each interior movable atom is able to interact with a full complement of atoms within a circular zone of the cutoff radius (see figure 1). The left and right edges are periodic, i.e., the atom in the rightmost column of subcells "see" on their right the left most column of subcells and atoms in the leftmost column of see on their left the rightmost column of subcells. In choosing such periodic boundary conditions, the left and the right edges of the computational cell must be carefully aligned in the initial configuration.

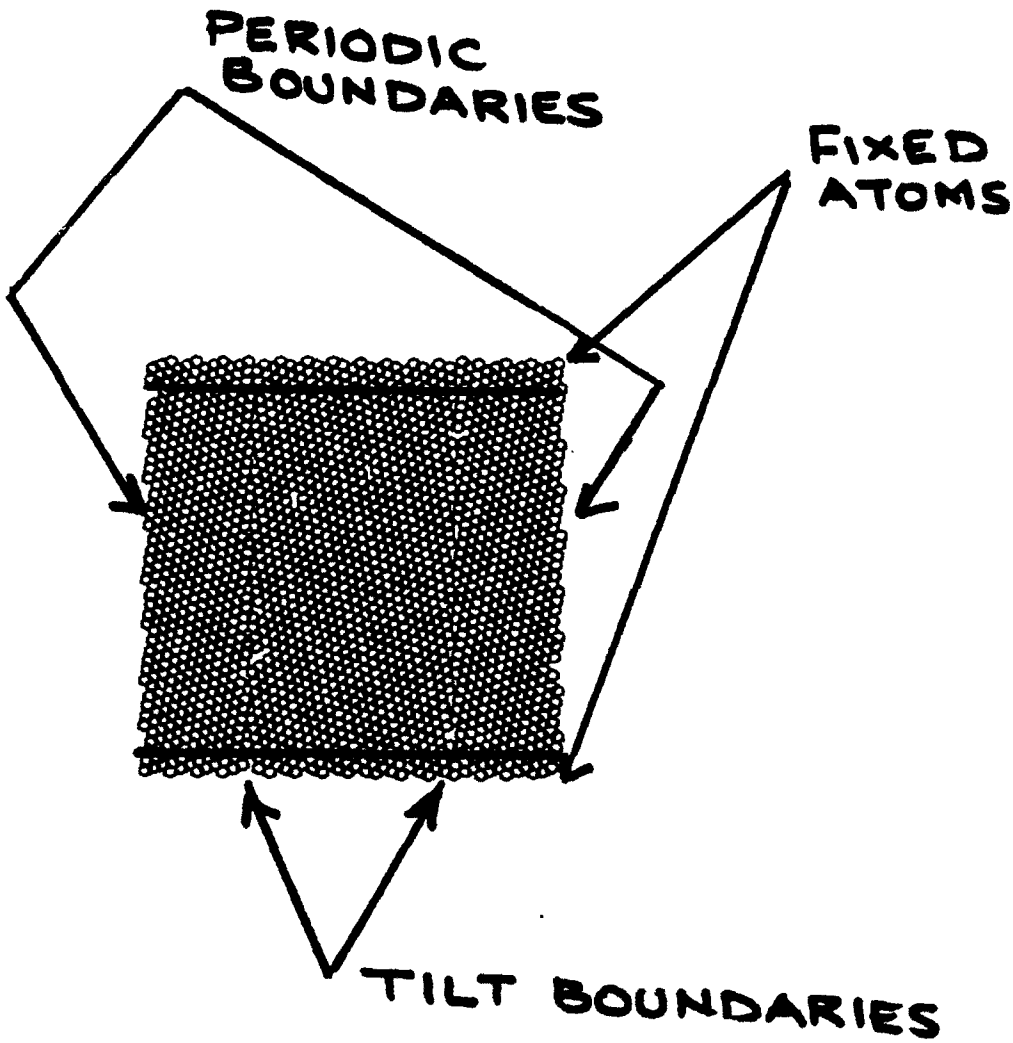


Figure 1: Boundary conditions on the computational cell

For the initial configuration, a coincidence lattice site is specified at (0.25, -0.5) coordinates. Atoms in the first crystal are placed in a closed packed array (representing the {111} plane of a fcc crystal) at a user-specified angle of orientation and split in two halves. The left half is located between $x = 0.25$ and $x = 0.5$, and the right half is located between $x = -0.5$ and $x = -0.25$. This assures that the left and right edges of the computation cell are indeed periodic. Atoms in the second crystal are placed at a different user-specified angle of orientation between $x = -0.25$ and $x = 0.25$. Figure 2 illustrates the initial configuration resulting in 20° symmetric tilt boundaries at $x = -0.25$ and $x = 0.25$ in a computational cell consisting of approximately 2000 atoms.

Diffusion Coefficient

In addition to monitoring potential energy, kinetic energy and total energy of the system at each time step in the simulation, transport coefficients may also be calculated from equilibrium correlation functions by observing Einstein relations. The diffusion coefficient D is given (in two dimensions) by

$$D = \frac{1}{2}\int_0^\infty dt \quad < \vec{v}_i(t) \cdot \vec{v}_i(0) >$$

where $\vec{v}_i(t)$ is the center-of-mass velocity of i^{th} atom at time t. The corresponding Einstein relation, valid for long times, is

$$D = \frac{1}{4t} \quad < |\vec{r}_i(t) - \vec{r}_i(0)|^2 >$$

where $\vec{r}_i(t)$ is the position of i^{th} atom at time t. In this implementation, the averages are computed for each of the movable atoms in the computational cell, the results are added together and divided by the number of movable atoms, to improve statistical accuracy. Precaution is taken to not switch from one periodic image to another by computing net displacement before the image rule is applied.

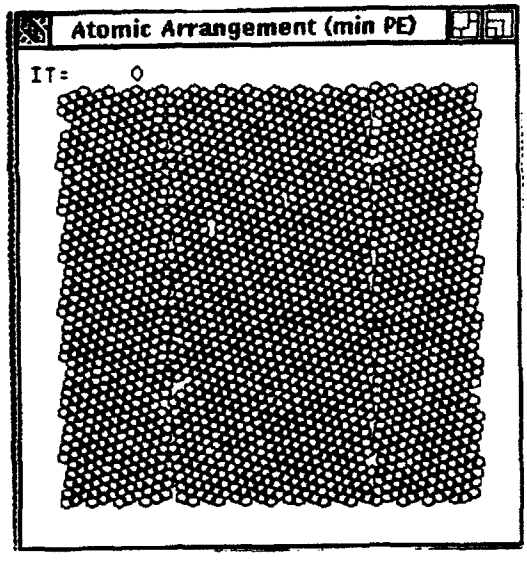


Figure 2: Initial configuration for 20° symmetric tilt boundaries

Program Structure

The program is divided into six files: **ALLCOM.FOR**, **BIX.FOR**, **IO.FOR**, **INI.FOR**, **GRAPH.FOR**, and **PROMPT.FOR**. The code is self-documenting, modular, and written in structured Fortran using VAX-extensions such as structures, records and do-while-loops.

ALLCOM.FOR file contains globally common variables and structure declarations. It serves as an *include* file for most of the modules. The parameter statement sets the limit on maximum number of atoms that can belong to the computational cell, and identifies the input/output devices. Currently, N is limited to 10,000 atoms, but by modifying the parameter statement, a lower or higher limit can be chosen.

BIX.FOR contains the main program, and modules that implement the computation of Lennard-Jones potential, and the Verlet-velocity algorithm.

IO.FOR contains modules that perform input and output functions. Input is permitted interactively from console as well as from an input file. If the latter is chosen, to initialize a simulation an input file of the type **.INP** is needed. To restart a simulation, an input file of the type **.INI**, **.MIN** or **.FIN** is needed. **INI** refers to an initial configuration that is saved at the beginning of a simulation, **MIN** represents the minimum potential energy configuration discovered during a simulation, and **FIN** refers to the configuration at the end of a simulation.

INI.FOR contains modules that set up the initial configuration. Among the inputs required are: the approximate number of atoms to be configured, angles of orientation of the two crystals, time step, number of iterations to be performed, and input/output filenames when applicable.

GRAPH.FOR contains graphic routines that display the current atomic arrangement and that of the minimum potential energy configuration discovered by that time step. Other windows display energy versus time plots, diffusion coefficient versus time plot, and the average coordination number and average bond length profile. The routines use the UIS

Graphics Library.

PROMPT.FOR contains user prompts for input/output functions.

Results and Concluding Remarks

The software has been successfully implemented on a variety of DEC workstations and minicomputers operating under the VAX/VMS operating system. Selected simulations involving 600, 2000 and 9000 atoms have been run successfully.

Results from simulations consisting of approximately 2000 atoms are reported here. Grain boundaries were 20° symmetric tilt type, and the computational cell has 2% sites vacant. Time step used was 10 *femto* seconds, and each simulation was run for 5000 time steps. A typical plot of diffusion coefficient versus time at $T = 450$°K is shown in figure 3.

Simulations were run at $T = 500°$, 475°, 450°, 425°, 400° and 375°K. A straight line fit to $\ln(D)$ versus $1/T$ data using the Least Squares method yields an activation energy of 1.292 kcal/mol-K for diffusion. Since no record is maintained whether the contribution is from movement within the grain or the grain boundary, the activation energy obtained should lie between that for lattice diffusion and for grain boundary diffusion.

The program was run on a timesharing VAX 6000-620 (70 SPEC throughput rating), a standalone MicroVax 3100 (2 SPEC) and a standalone Vaxstation 2000 (1 SPEC) for comparison with identical input data set. VAX 6000 was found to be about 25 times faster than the MicroVax which in turn was approximately 5 time faster than the Vaxstation.

There are several ways in which the program can be enhanced to study the electromigration phenomenon. Naturally, for realistic simulations, the program needs to be modified to model 3-dimensional grains. A proposal is being submitted to AFOSR under its Research Initiation Program to accomplish this task. Driving force gradients which cause electromigration also

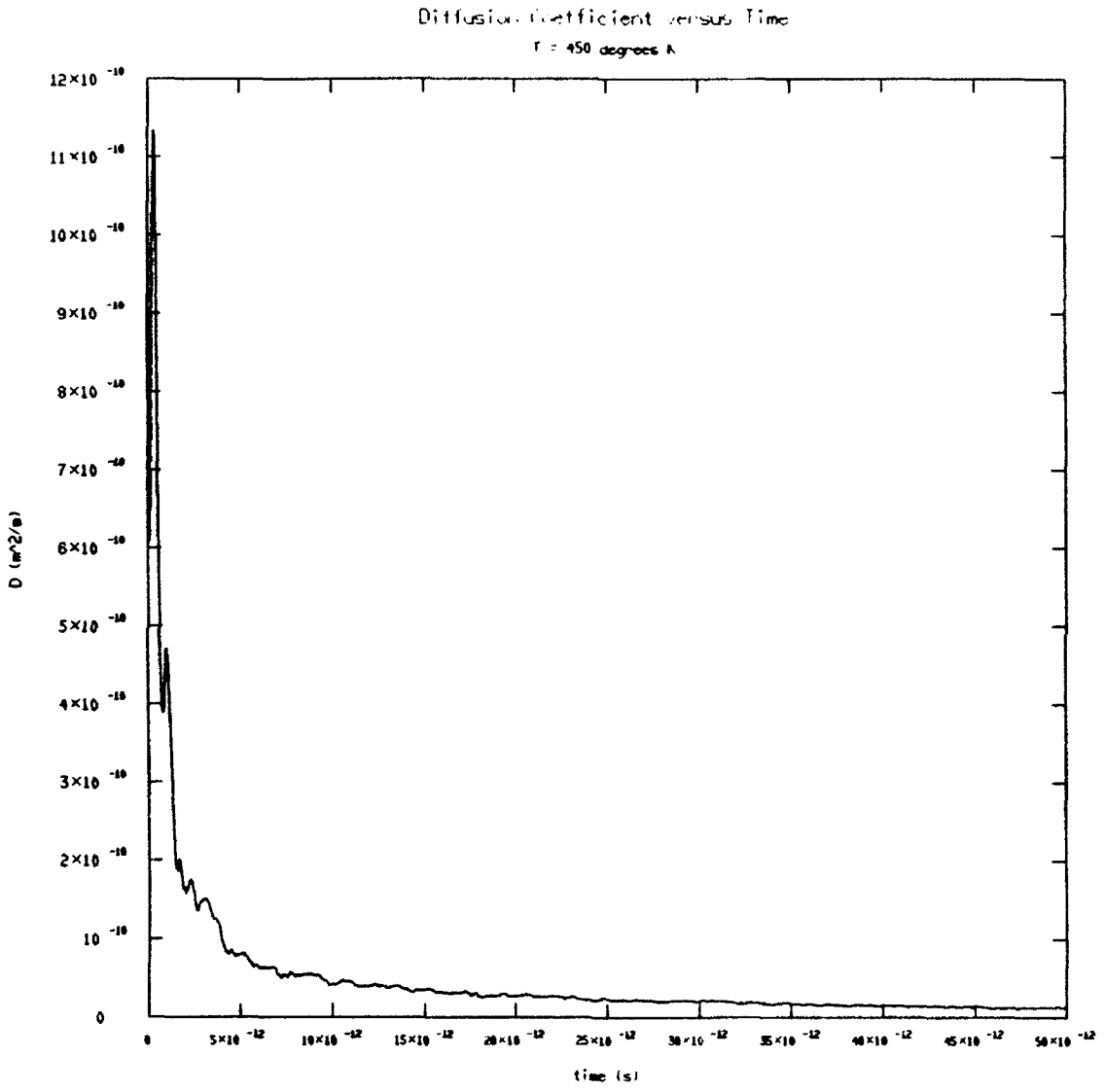


Figure 3: Diffusion Coefficient versus time at $T = 450°K$

need to be incorporated in the model. These gradients are difficult to define at the atomistic level. Efforts are underway in collaboration with Rome Laboratory researchers to investigate fundamental issues involved in defining such gradients at the atomistic level.

WIDEBAND ATM NETWORKS WITH ADAPTIVE ROUTING FOR THE DYNAMIC THEATER ENVIRONMENT

Robert R. Henry
Professor
Department of Electrical & Computer Engineering

University of Southwestern Louisiana
P.O. Box 43890
Lafayette, LA 70504-3890

Final Report for:
AFOSR Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research

August 1992

WIDEBAND ATM NETWORKS WITH ADAPTIVE ROUTING FOR THE DYNAMIC THEATER ENVIRONMENT

Robert R. Henry
Professor
Department of Electrical & Computer Engineering
University of Southwestern Louisiana

Abstract

Traditional tactical and theater military communication networks are characterized by relatively low bandwidth links. The environment is dynamic in the sense that the links are subject to jamming and the nodes to destruction by the enemy. Modern and future military equipment and tactics require the use of wideband links to exchange bandwidth-intensive information such as video and images. However, current and proposed wideband networks such as ATM have been designed for peacetime, ie, well-behaved operation. The research described herein proposes and evaluates three ways in which wideband ATM networks may be adapted for operation in the dynamic theater environment.

WIDEBAND ATM NETWORKS WITH ADAPTIVE ROUTING FOR THE DYNAMIC THEATER ENVIRONMENT

Robert R. Henry

INTRODUCTION

The summer research efforts of the author may be divided into three related and overlapping phases. Phase one studied the various communication research projects being conducted by the Communication Networks Branch of the Telecommunications Division of Rome Laboratory. Phase two was concerned with determining the interrelationships between the projects and identifying promising research areas that are significant to Rome Labs and to this researcher. Phase three began the initial research, and proposes future research plans. This report documents the three phases of the research.

CURRENT PROJECTS

The Evaluation and Development of Multimedia Networks Under Dynamic Stress (EDMUNDS) project [1] is an ongoing effort to develop Secure Tactical Internet Protocols (STIP) that respond well to a stressed tactical environment. Such an environment is characterized by communication links subject to jamming and nodes subject to destruction by the enemy. The assumption is that the links have relatively low bandwidth, and that there is sufficient time to perform sophisticated processing to determine optimal datagram routing.

The Secure Survivable Communication Network (SSCN) project [2] was initiated in an effort to utilize the emerging Broadband Integrated Services Digital Network (B-ISDN) for military applications. This network relies on the Asynchronous Transfer

Mode (ATM) [6] to provide rapid multiplexing and routing of data over wide bandwidth links. Due to high data rates there is relatively little time to determine the optimal route of each packet (cell). Thus connection oriented services have been selected to provide routing that is fixed for the call duration. This is in direct contrast with the STIP routing protocol.

Two other projects, the Media Resource Controller (MRC) [3] and the Multimedia Communication Capability (M2C2), [4] are similar to the EDMUNDS project in the sense that the routing is adaptive and the links are low bandwidth. However, these projects have actually been implemented and demonstrated in the laboratory. Another laboratory oriented project is the Advanced Multi-Media Information Distribution System (AMIDS) [5] in which a distributed Tactical Air Control Center (TACC) is interconnected by fixed wideband links.

Figure 1 graphically compares the network size and link capacity of the above mentioned networks. The ATM and STIP networks cover a wide geographical area, while the AMIDS and MRC span a smaller area corresponding to the tactical environment. ATM and AMIDS have high capacity links in contrast with the lower capacity MRC and STIP links.

STIP NODE ARCHITECTURE

The basic node architecture is shown in Figure 2. As can be seen a packet on any one of the N incoming links is sorted and stored in one of the D destination queues. Each of the L outgoing links has a link scheduler which accepts packets from a tap on each destination queue. Only packets above the tap in the queue are sent to the link scheduler. Figure 2 shows the scheduler and taps for link *1* only, with all other links having a similar structure.

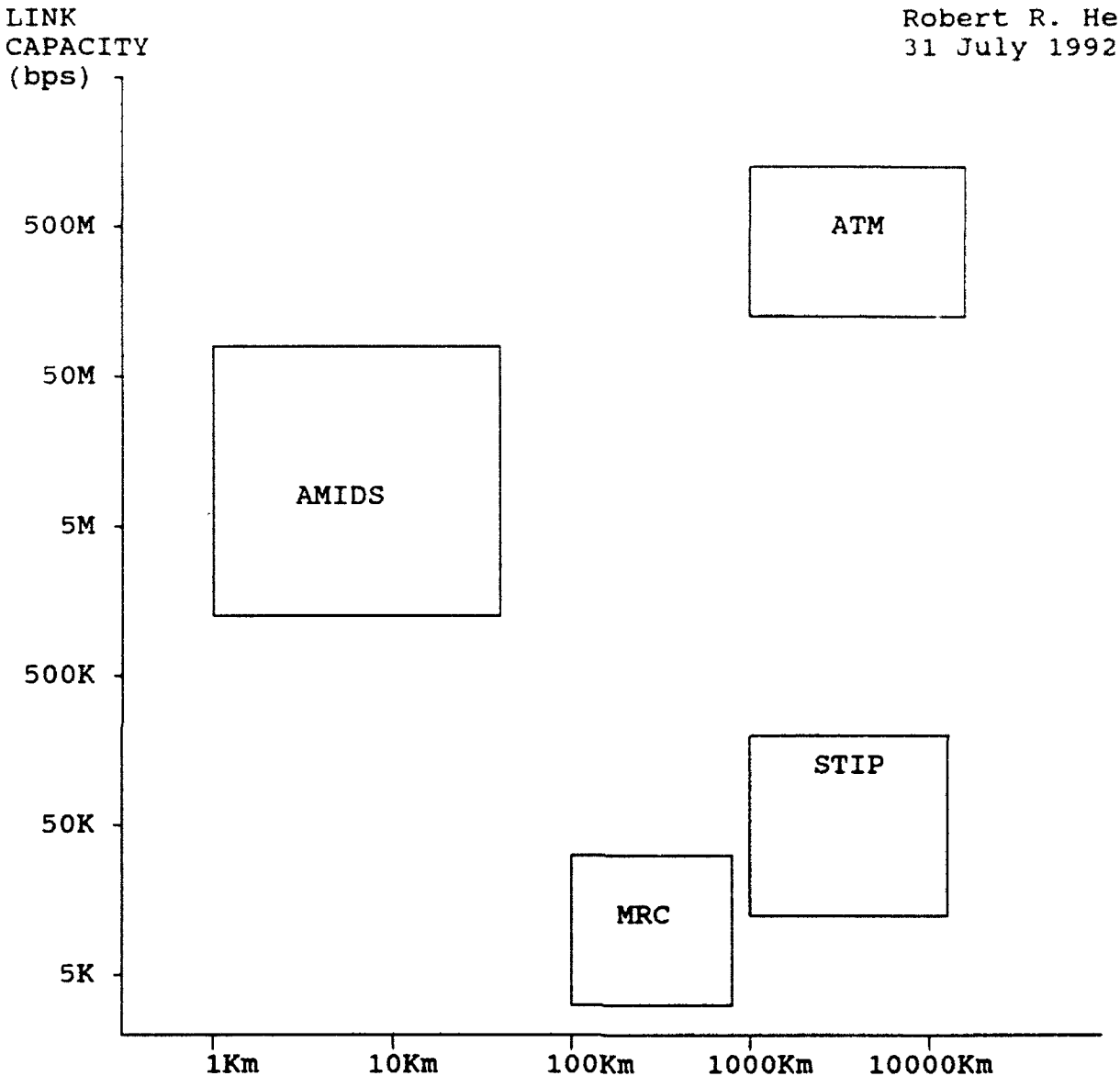


Figure 1. Switching Network Overview.

Link scheduling is accomplished by periodically repeating the following steps:

1- An estimate of the status of each outgoing link is made based on the delay of acknowledgement packets.

2- The delays $D(l,d)$ to each destination d via each link l is determined by exchanging information with neighbor nodes.

3- Flows $f(l,d)$ and queue taps $\theta(l,d)$ are computed for each link-destination pair as follows:

a. A packet entering the queue has the same end-to-end (ete) delay no matter which link is chosen. That is, the differential delay is

$$\tau(d) = q(l,d)/f(l,d) + D'(l,d) \quad \text{for all } l, \text{ and}$$

b. minimize the total delay

$$J = \sum_{d} \tau(d).$$

c. Determination of queue thresholds:

$$\theta(l_{k+1},d) = \theta(l_k,d) + \sum_{j=1}^{k} f(l_j,d)[D(l_{k+1},d) - D(l_k,d)]$$

The effect of the STIP protocol is to distribute packets on all viable links rather than choosing the "best" link. Packets waiting for a "long" time near the bottom on the queue are sent on the link with the smallest delay, while packets that arrived recently near the top of the queue are routed to a link with a longer delay. Thus it is quite likely that packets for the same destination are simultaneously being sent over different links. This is in contrast to the traditional "least-cost" routing paradigm which sends one packet over the "best" link.

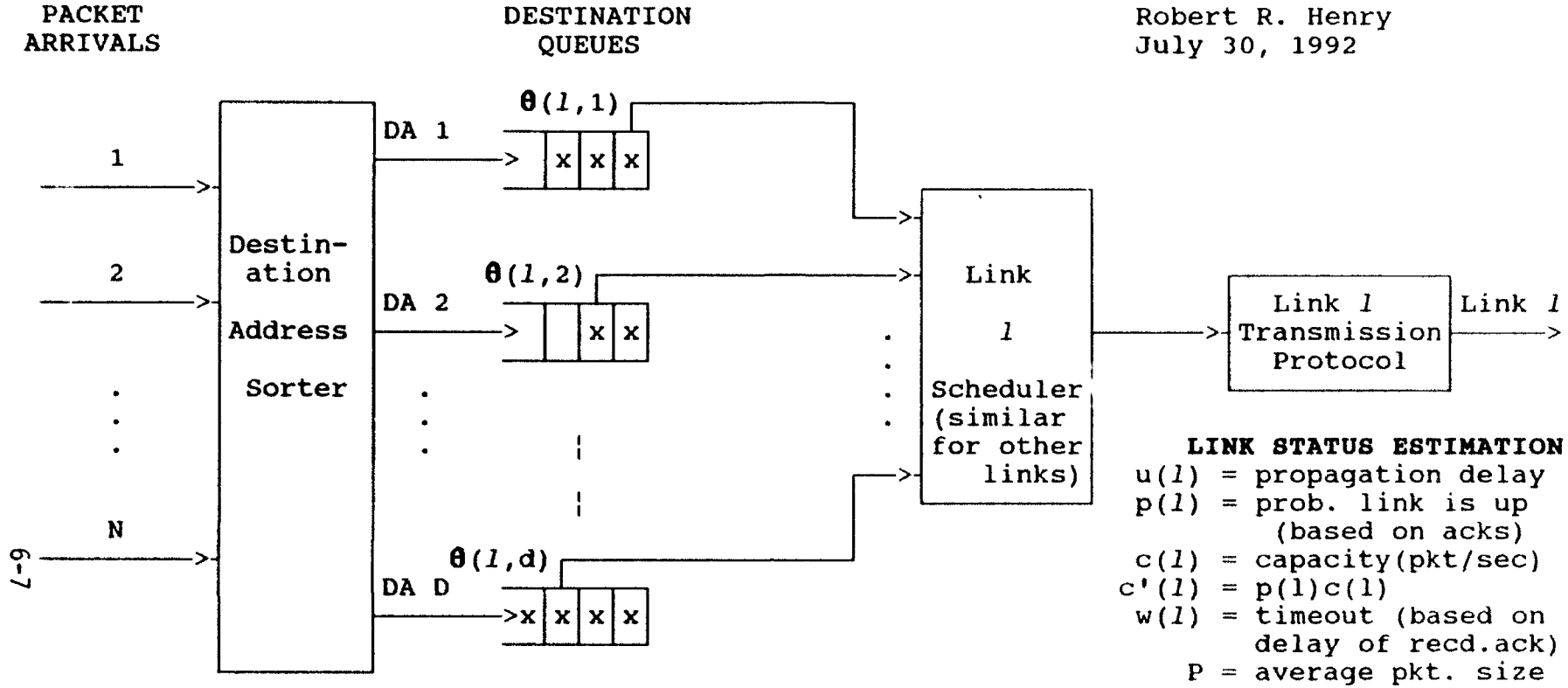
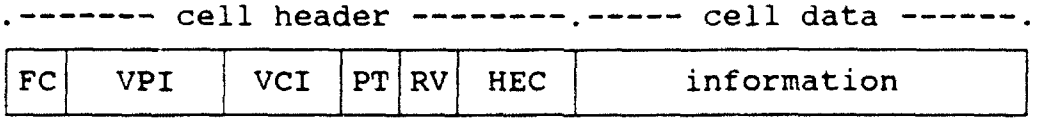


Figure 2. STIP Node Architecture.

ATM NODE ARCHITECTURE

The ATM switch has N input FIFO buffers, each associated with an incoming link, and L output FIFO buffers for the outgoing links as shown in Figure 3. All input and output data is in the form of 53 octet ATM cells [6] with a 5 octet header as shown below.



The ATM switch routes the cell based on the Virtual Path Identifier (VPI) field in the header. An Internal Routing Lookup Table maps each active VPI number to one of the output buffers (links). Figure 3 illustrates how the switch appends an internal node tag to each cell. The node tag causes the cell to take an appropriate path through the switching fabric to the desired output link. Thus the lookup table contents and the VPI field define a virtual circuit route through the switch for all incoming cells.

The control processor has the important function of updating the routing table. It does so by accepting requests to set up a Virtual Channel (VC) from another switch via the control input buffer. In turn the controller can send requests to other switches via the controller output buffer. The physical path through which cells belonging to a session flow is determined and fixed for the duration of the session through a call set-up procedure. The control processor generates updates at a relatively low rate compared to the 150 Mbps rate at which individual cells are routed through the switch.

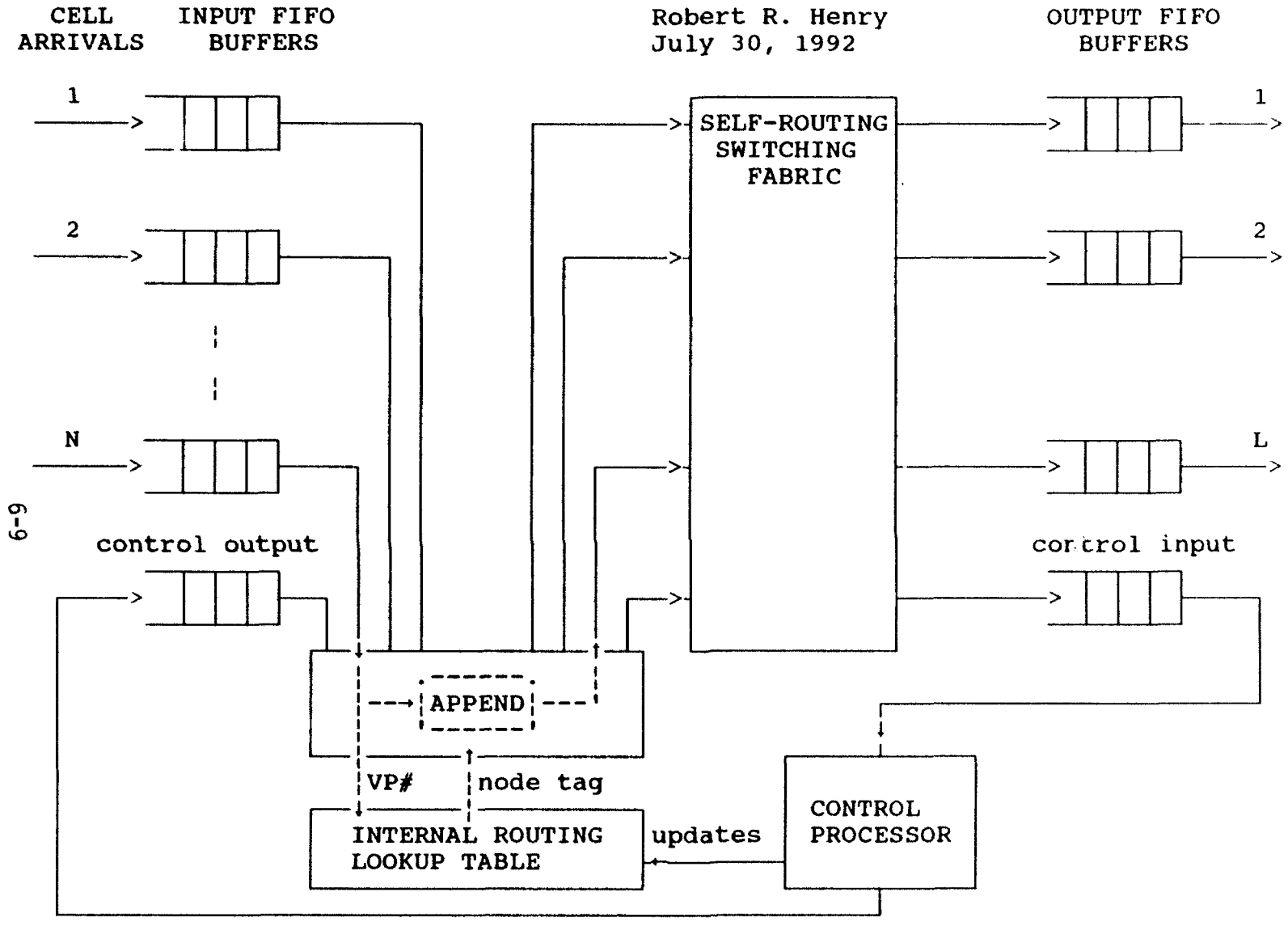


Figure 3. ATM Node i Switch Architecture.

A WIDEBAND DYNAMIC NETWORK

Each of the networks shown in Figure 1 may be classified into one of two groups based on link capacity and the ability to adapt to a stressed environment. The classification is as follows:

ATM/AMIDS: high capacity, low adaptability

STIP/MRC/M2C2: low capacity, high adaptability.

Interconnection between these two dissimilar groups requires careful consideration so that performance does not revert to the least common denominator (ie low capacity and low adaptability). Figure 4 illustrates the networks in these two groups in more detail. As can be seen ATM represents the highest in link capacity, while STIP represents the ultimate in adaptability to a stressed environment.

An underlying concept of the global network is that the National Command Authority (NCA) will be able to communicate directly with personnel in the tactical environment. In order that graphics and video information be delivered in a timely manner, high capacity links are essential. ATM is a candidate for the wide-area part of this network since it is wideband and is being developed commercially. However, since ATM is designed for peacetime (well-behaved) use, STIP-like protocols must be incorporated to provide for survivability in the tactical area. This report documents research by the author that formulates methods which combine the advantages of both the ATM and STIP protocols into a network suitable for the military environment. This is the "RESEARCH" area indicated in Figure 4.

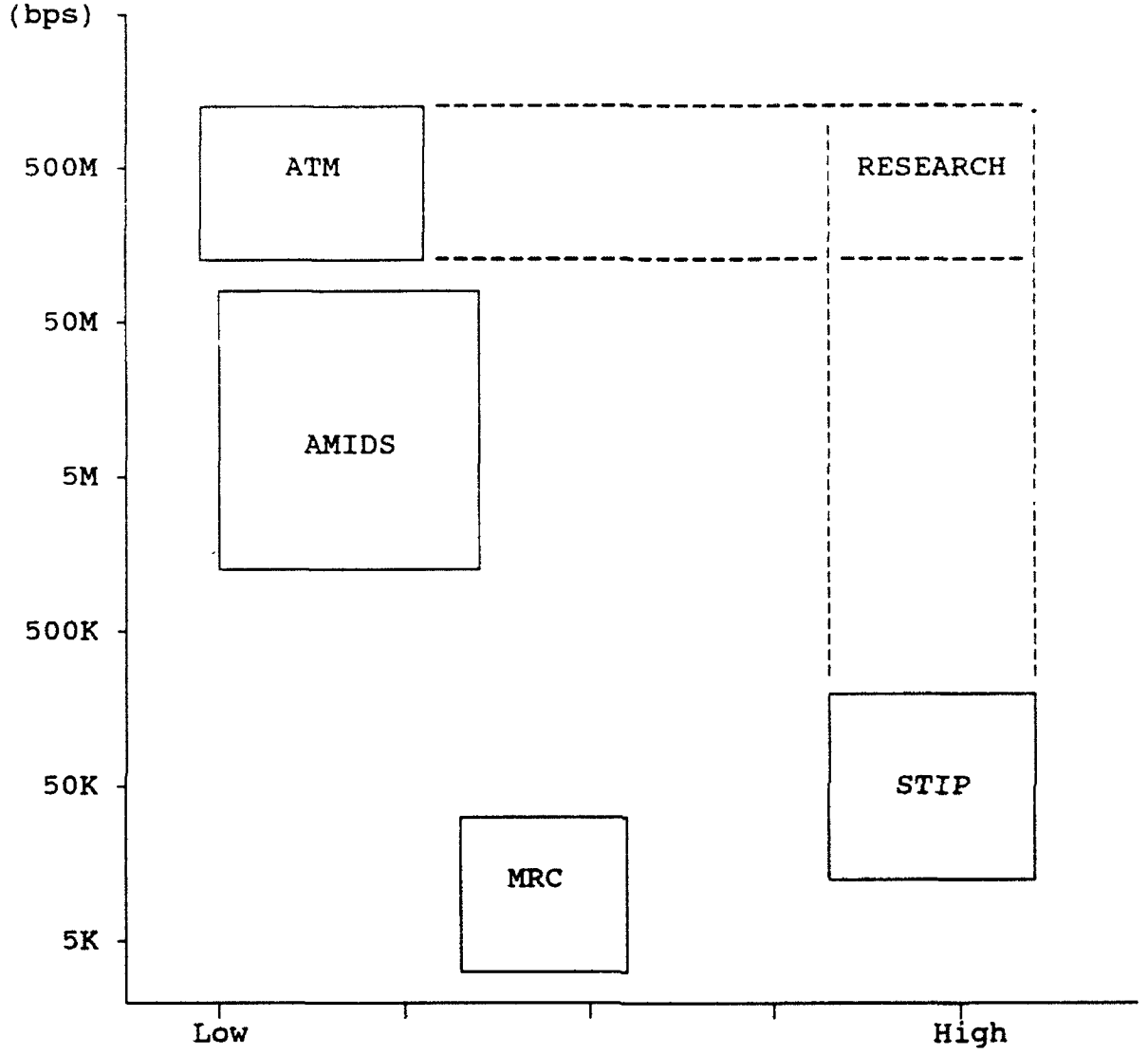
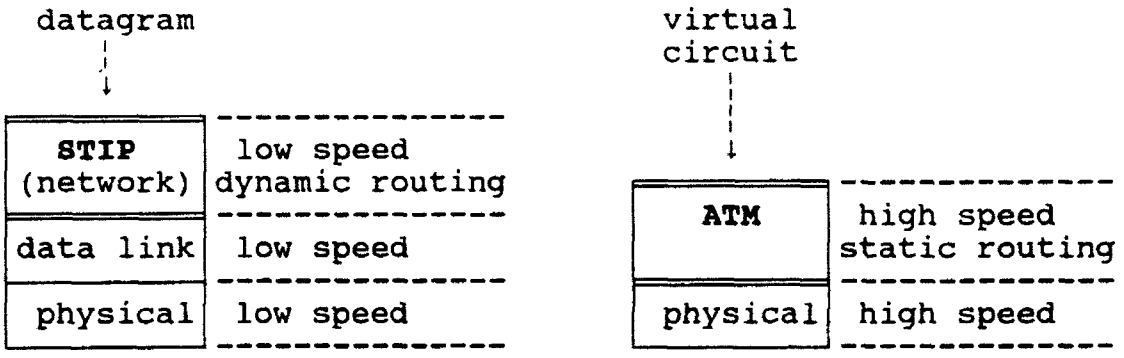


Figure 4. Network Protocol Dynamics.

PROTOCOL STACK COMPARISON

A comparison of the protocol stacks for both the ATM and STIP protocols is shown below, along with the layer performance characteristics, and the service provided to the layer above. The basic dichotomy is apparent - STIP provides datagram service which is low speed with dynamic routing; while ATM provides virtual circuit service at high speed with static routing. In addition, ATM routing is provided at a lower layer than is the more traditional STIP network layer routing.



The challenge then is to provide datagram access to the high speed ATM switch and incorporate the dynamic routing paradigm of the STIP protocol.

DATAGRAM ACCESS TO THE ATM NETWORK

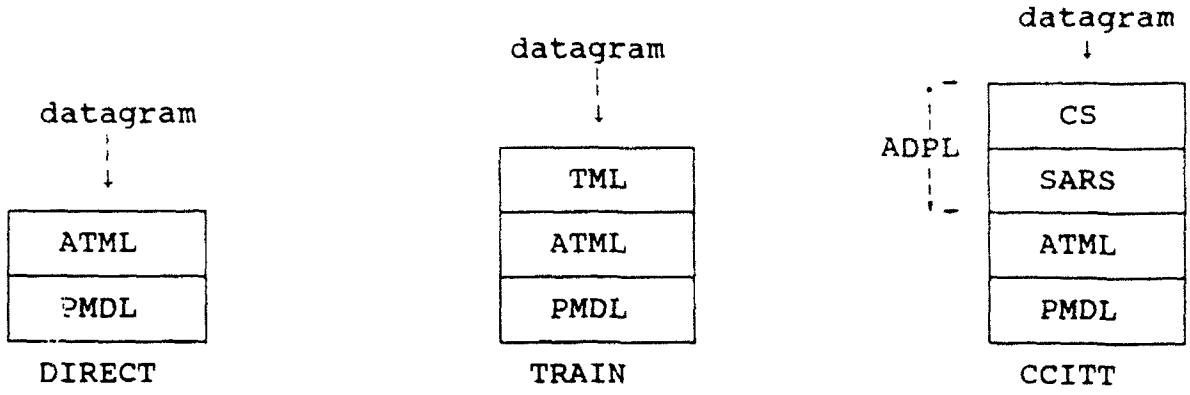
Efforts are currently underway by the CCITT [7] to provide datagram service via an Adaptation Layer (ADPL) that resides on top of the ATM Layer (ATML). This method embeds a 2-octet header and a 2-octet trailer in each cell data field to provide destination addressing, segmentation and reassembly, and error checking. The exact procedures have yet to be determined, but there is currently no effort to incorporate dynamic routing on a datagram basis. The protocol stack for this network is shown in Figure 5a. The corresponding sequence of cells generated by each

datagram is given in Figure 5b.

A literature search reveals current research that focuses on building a connectionless network "on top of" the ATM network using the services defined by the Adaptation Layer. Rahnema [8] proposes the use of this layer to establish permanent VC connections between pairs of connectionless networks. This method does not provide datagram service within the ATM network, but merely point-to-point connections between existing datagram networks. A centralized connectionless server which receives datagrams from remote nodes is proposed by DePrycker [9]. Such a network has a "star" configuration with a disadvantage that the hub is a single-point failure and a prone to congestion. A distributed version in which connectionless servers are placed at each ATM node is suggested by Landegem and Peschi [10]. Each server uses the adaptation layer to exchange datagrams with other nodes. However the authors do not give details of how routing may be accomplished.

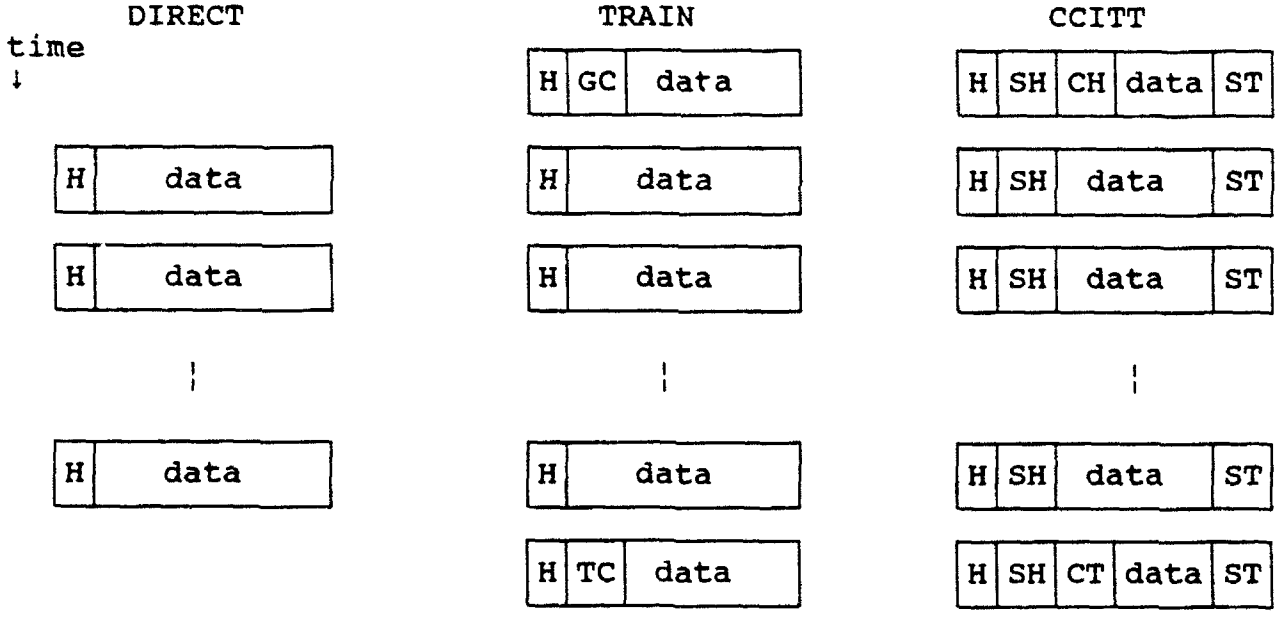
In addition to the CCITT method, two other methods for datagram access proposed by the author are illustrated in Figure 5. The "Direct" method inputs datagram octets directly to the ATM layer. This requires that a subset of VCI numbers by reserved for use as a datagram address field. The ATM switch would route these "datagram" cells separately from the standard VC cells.

The "Train" method uses a Train Mode Layer (TML) to interface between the datagram service and the ATM layer. For each datagram the TML generates a "train" of cells beginning with a Guide Cell, followed by Data Cells, and ending with a Trail Cell. The Guide Cell contains the routing information and leads the following "train" of cells through the network. The Trail Cell then deactivates the track established by the Guide Cell, thereby releasing switch resources for other datagram transmission.



ATML:ATM Layer TML:Train Mode Layer ADPL:Adaptation Layer
PMDL:Physical Media Dependent Layer CS:Convergence Sublayer
SARS:Segmentation & Reassembly Sublayer

Figure 5a. Protocol Stack for the Proposed ATM Datagram Access Networks.



H : ATM cell Header, 5 octets
SH : Segmentation & Reassembly Sublayer Header, 2 octets
ST : Segmentation & Reassembly Sublayer Trailer, 2 octets
CH : Convergence Sublayer Header, 4 octets
CT : Convergence Sublayer Trailer, 4 octets
GC : Guide Cell Header, 4 octets
TC : Trail Cell Header, 4 octets

Figure 5b. Datagram Cell Transmission Sequence for Proposed Networks.

ANALYSIS OF UTILIZATION

The relative performance of each of the three datagram access methods is determined in this section. The metric used is utilization of the physical layer bits in the absence of noise. This gives an upper bound on performance and is useful in evaluating the relative merits of the methods. Let

```
 D =     number of octets in the datagram
HF = 5 = number of octets in the ATM cell header field
DF = 48= number of octets in the ATM cell data field
SH = 2 = number of octets in the SARS header field
ST = 2 = number of octets in the SARS trailer field
CH = 4 = number of octets in the CS header field
CT = 4 = number of octets in the CS trailer field
GC = 4 = number of octets in the Guide Cell header field
TC = 4 = number of octets in the Trail Cell header field.
```

The number of cells required to "package" D datagram octets for DIRECT access is

$$Nc1 = \text{ceiling}[D/DF]. \quad (1)$$

From Figure 5b it can be seen that the utilization then becomes

$$U1 = 100*D/[Nc1*(HF+DF)]. \quad (2)$$

Utilization for the TRAIN access method is determined by including the GC and TC overhead octets in Nc, and yields

$$U2 = 100*D/[(Nc2)*(HF+DF)], \quad (3)$$

where

$$Nc2 = \text{ceiling}[(D+GC+TC)/DF]. \quad (4)$$

In a similar way utilization for the CCITT access method is determined by including the SH and ST overhead octets in Nc, and gives

$$U3 = 100*D/[Nc3*(HF+DF)], \quad (5)$$

where

$$Nc3 = \text{ceiling}[(D+CH+CT)/(DF-SH-ST)]. \quad (6)$$

A plot of U1, U2, and U3 as a function of packet size D is given in Figure 6. The periodic nature of the plot is due to the varying number of octets of data that fit into the last cell of the sequence as D increases. The general conclusion is that a packet size of 200 octets or more is required to yield efficiencies in the 70% to 90% range.

For very large packet sizes the Direct and Train methods perform equally well and approach an efficiency of

$$100*DF/(DF+HF) = 100*48/(48+5) = 90.6\ \%. \qquad (7)$$

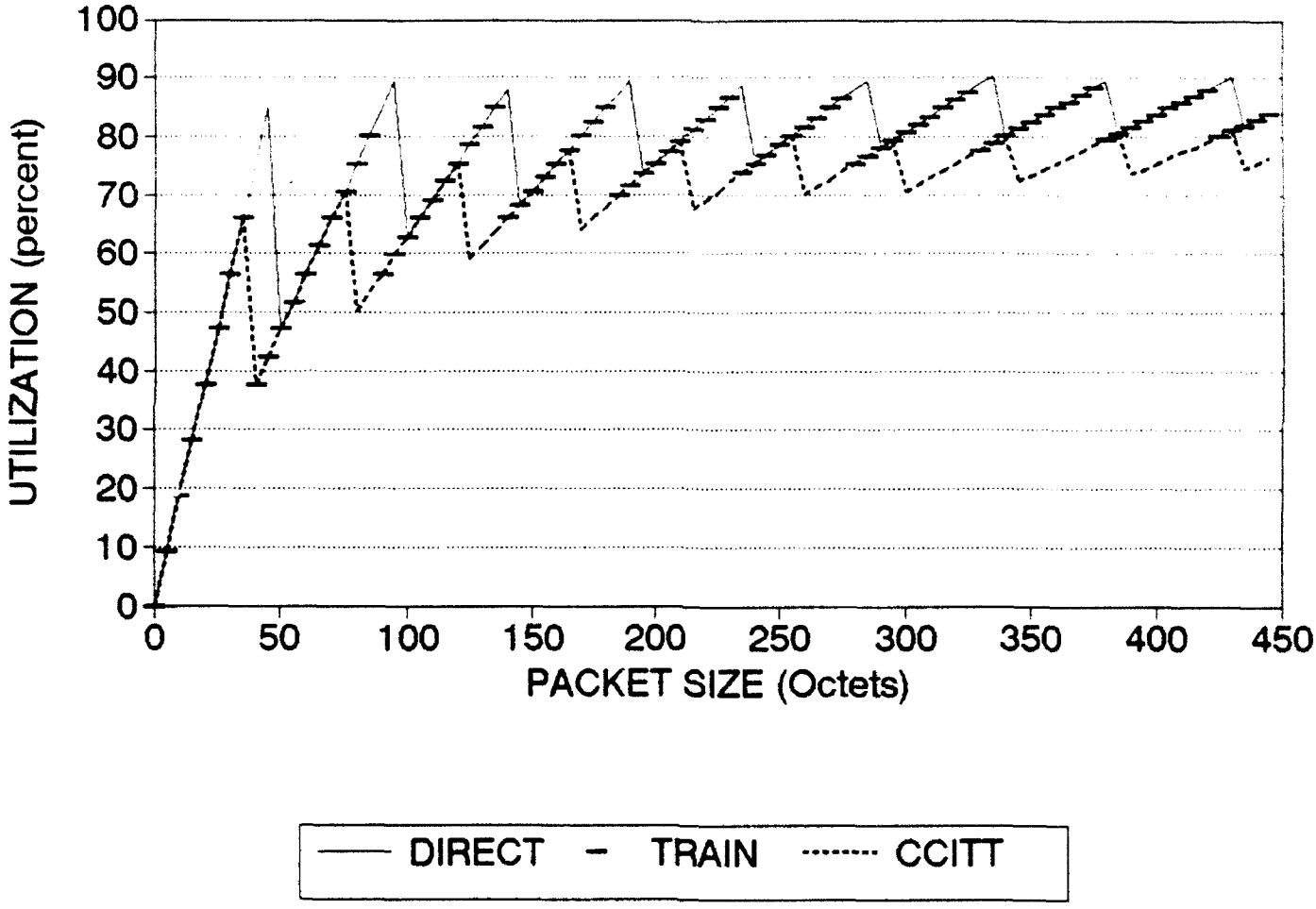
The CCITT method performance is inferior approaching an efficiency of

$$100*DF/(DF+HF+SH+ST) = 100*48/(48+5+2+2) = 84.2\ \%. \qquad (8)$$

DYNAMIC ROUTING IN THE ATM WIDEBAND NETWORK.

To effect dynamic routing at each node, the basic ATM switch architecture and control processor shown in Figure 3 must be enhanced. The processor will be required to determine the status of each outgoing link by monitoring parameters such as the number of cells queued in the output buffers, etc. This information is then exchanged with immediate neighboring switches, and a STIP-like paradigm is used to determine the best links over which to send datagrams to each destination. The routing table is then updated to reflect these changes.

Figure 6.
Datagram Access to ATM Networks



The control processor must identify the destination address and the sequence of incoming cells corresponding to datagrams. Three such possible sequences are shown in Figure 5b, with the destination address being contained in the first cell of the sequence. This requires that the switch "look" into the data field for the CCITT and TRAIN method, which is equivalent to Adaptation Layer and Train Mode Layer processing. The DIRECT method requires information contained in the cell header, corresponding to ATM Layer processing.

CONCLUSION

This research proposes three ways in which a wideband ATM network can be enhanced to operate in the dynamic theater environment. The cell transmission sequence to provide datagram access to the ATM network for each the three methods, DIRECT, TRAIN, and CCITT is defined. A performance analysis reveals that a packet size of 200 octets or more yield ATM utilization in the range of 70% to 90% for all three methods. Modifications to ATM switch architecture and control to incorporate dynamic routing is shown to be feasible.

The research described herein establishes a foundation for future work. A number of issues need further study including the effect of link errors and jamming on performance, and determination optimum packet size. In addition, the details of the dynamic routing algorithm within the ATM network need to be determined and the performance analyzed.

REFERENCES

[1] "Evaluation and Development of Multimedia Networks in Dynamic Stress," Semiannual Project Report, SRI contract with Rome Laboratory, April 5-9, 1992.

[2] "Secure Survivable Communications Network," User Meeting Report, GTE Government Systems contract with Rome Laboratory, July 15-16,1992.

[3] M.T. Rafter and W.C. Walker, "A Distributed Multiple Media Network Architecture," MILCOM '89 Conference Record, October 15-18, 1989, pp.1-6,

[4] "An Overview of the Multimedia Communications Capability Program," Rome Laboratory Telecommunications Division internal report.

[5] D.E. Krzysiak, "Advanced Multi-Media Information Distribution System," MILCOM '90 Conference Record, 1990, pp.35.7.1-35.7.3.

[6] Draft CCITT Recommendation I.150, "B_ISDN ATM Functional Characteristics," 1990.

[7] Draft CCITT Recommendation I.362, "B_ISDN ATM Adaptation Layer Functional Description," 1990.

[8] M. Rahnema, "Frame Relaying and the Fast Packet Switching Concepts and Issues," IEEE Network Magazine, July 1991, pp.18-23

[9] M. De Prycker, "ATM Switching on Demand," IEEE Network Magazine, March 1992, pp.25-28.

[10] T.V. Landegem and R. Peschi, " Managing a Connectionless Virtual Overlay Network On Top of an ATM Network," ICC '91 Conference Record, June, 1991, pp.31.5.1-31.5.5.

United States Air Force
Faculty Summer Research Program

Final Report

Multipath Channel Equalization for Spread Spectrum Communication System

Presented By

Dr. H. K. Hwang
Department of Electrical and Computer Engineering
California State Polytechnic University
Pomona, California 91768-4065

August 28 1992

Abstract

Spread spectrum communication systems use a much wider transmission bandwidth than that is required for information transmission. The advantage of using the excessive bandwidth is that the system becomes less sensitive to noise, jammer, and intersymbol interference (ISI). To further aid the spread spectrum in reducing the ISI, an adaptive equalizer is used. The idea of using adaptive equalizer to suppress the ISI is similar to reference 1, which using the adaptive filter to reject the narrow band interference.

1. Introduction

The spread spectrum system considered in this report is binary signaling, direct sequence spread spectrum system that operates in a multipath environment. More sophisticated signaling systems can be generalized from this discussion. The block diagram of this system is shown in Figure 1.1.

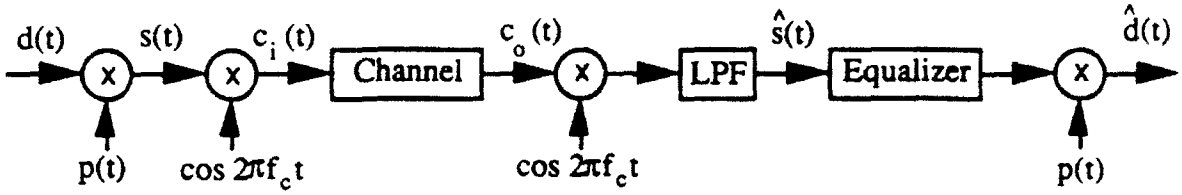


Figure 1.1 Block Diagram of Direct Sequence Spread Spectrum System

In Figure 1.1. d(t), p(t), s(t) are the data signal, PN sequence and spread spectrum waveform. $c_i(t)$ and $c_o(t)$ are the channel input (modulated) and the channel output waveforms. $\hat{s}(t)$ and $\hat{d}(t)$ are the estimated spread spectrum waveform and the estimated data. The relationships between different signals are:

$$s(t) = d(t)p(t) \tag{1.1}$$

$$c_i(t) = s(t) \cos \omega_c t \tag{1.2}$$

Assume the channel consists of a direct path and a multipath with gain a and delay t_d. The block diagram representation of the channel is shown in Figure 1.2.

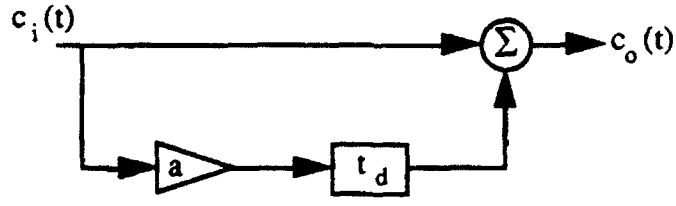


Figure 1.2 Multipath channel model

The channel output waveform $c_o(t)$ is

$$c_o(t) = c_i(t) + ac_i(t-t_d)$$

$$= s(t) \cos \omega_c t + as(t-t_d) \cos \omega_c (t-t_d)$$

$$= s(t) \cos \omega_c t + as(t-t_d) \cos (\omega_c t-\theta) \qquad (1.3)$$

where $\theta = \omega_c t_d$ is the phase term due to multipath delay.

The coherent demodulation process consists of a coherent carrier wave multiplier, then follow by a low pass filter (LPF). The demodulated signal $\hat{s}(t)$ is

$$\hat{s}(t) = [c_o(t) \cos \omega_c t]_{LP}$$

$$= [\, s(t) \cos^2 \omega_c t + as(t-t_d) \cos(\omega_c t-\theta) \cos \omega_c t \,]_{LP}$$

$$= \{\frac{1}{2} s(t)(1 + \cos 2\omega_c t) + \frac{1}{2} as(t-t_d)[\cos\theta + \cos(2\omega_c t-\theta)\,]\}_{LP}$$

Assume the LPF has a gain of 2 in the baseband region, then signal $\hat{s}(t)$ is

$$\hat{s}(t) = s(t) + a\cos\theta\, s(t-t_d) \qquad (1.4)$$

Equation (1.4) shows that there is an interference term $a\cos\theta\, s(t-t_d)$ in the estimated spreaded signal. The equivalent baseband channel model is exactly the same as that in the passband except the gain in the multipath is modified by an additional multiplication factor $\cos\theta$. The signal to interference ratio at the demodulator output is

$$(S/I)_{dem} = \frac{1}{a^2\cos^2\theta} \qquad (1.5)$$

If the interference term can be suppressed at the demodulator output, it is certainly a very attractive feature to enhance the system performance. An adaptive equalizer can be used to undertake this task.

Assume the multipath delay time is longer than a chip time, and N is the number of chip per bit interval. Consider binary signaling case, at a given time t, signal component s(t) is either 1 or -1. The amplitude of the interference term, $a\cos\theta\, s(t-t_d)$ is either $a\cos\theta$ or $-a\cos\theta$. Without adaptive equalizer, the total average energy

per bit interval at the despreading circuit output can be obtained by the following equation.

$$E\{(\sum_{n=1}^{N}[s(n)+a\cos\theta s(n-n_d)]p(n))(\sum_{m=1}^{N}[s(m)+a\cos\theta s(m-n_d)]p(m))\}$$

$$=E\{(\sum_{n=1}^{N}[d(n)+a\cos\theta d(n-n_d)p(n-n_d)p(n)])(\sum_{m=1}^{N}[d(m)+a\cos\theta d(m-n_d)p(m-n_d)p(m)])\}$$

$$=E\{[\sum_{n=1}^{N}d(n)][\sum_{m=1}^{N}d(m)]\} +$$

$$2E\{[\sum_{n=1}^{N}d(n)][\sum_{m=1}^{N}a\cos\theta d(m-n_d)p(m-n_d)p(m)]\} +$$

$$a^2\cos^2\theta E\{[\sum_{n=1}^{N}d(n-n_d)p(n-n_d)p(n)][\sum_{m=1}^{N}d(m-n_d)p(m-n_d)p(m)]\} \quad (1.6)$$

The first term in Equation (1.6) is the bit energy due to signal component only. Since $d(n) = d(m) = \pm 1$ during a bit interval, this term is

$$E\{[\sum_{n=1}^{N}d(n)][\sum_{m=1}^{N}d(m)]\} = N^2 \quad (1.7)$$

The second term is the cross term due to coupling between signal and interference. The third term is due to interference only. Depending on the different scenario, contributions due to the second and third terms are different.

(1) If the period of PN sequence is much longer than the bit interval, then $d(n-n_d)p(n-n_d)p(n)$ at a given bit interval can be modeled as a random variable with equal probability of taking value 1 or -1.

$$d(n-n_d)p(n-n_d)p(n) = e^{j\phi_n} \quad (1.8)$$

where ϕ_n is a random variable with probability density function $p_{\phi_n}(\theta)$.

$$p_{\phi_n}(\theta) = \frac{1}{2}[\delta(\theta) + \delta(\theta-\pi)] \tag{1.9}$$

Since the period of PN sequence is much longer than a bit interval, frequency despreading circuit only performs partial correlation[2,3]. The second term in equation (1.6) becomes

$$2E\{[\sum_{n=1}^{N} d(n)][\sum_{m=1}^{N} a\cos\theta d(m-n_d)p(m-n_d)p(m)]\}$$

$$= \pm 2Na\cos\theta\, E[\sum_{n=1}^{N} e^{j\phi_n}] = 0 \tag{1.10}$$

The third term in Equation (1.6) becomes

$$a^2\cos^2\theta\, E[\sum_{n=1}^{N} e^{j\phi_n} \sum_{m=1}^{N} e^{j\phi_m}] = a^2\cos^2\theta \sum_{n=1}^{N}\sum_{m=1}^{N} E[e^{j(\phi_n+\phi_m)}]$$

Note that $E[e^{j(\phi_n+\phi_m)}] = \begin{cases} 1 & n = m \\ 0 & n \neq m \end{cases}$

then, $a^2\cos^2\theta \sum_{n=1}^{N}\sum_{m=1}^{N} E[e^{j(\phi_n+\phi_m)}] = a^2\cos^2\theta E[\sum_{n=1}^{N} 1]$

$$= Na^2\cos^2\theta \tag{1.11}$$

Hence, the signal to interference ratio at the output of despreading circuit is

$$(S/I)_{des} = \frac{N}{a^2\cos^2\theta} \tag{1.12}$$

The processing gain PG_D is defined as the ratio of $(S/I)_{des}$ and $(S/I)_{dem}$, and it is

$$PG_D = N \tag{1.13}$$

(2)If each bit is modulated by the identical PN sequence, then depending on the relationship between multipath delay time and bit interval, the effect of interference can be quite different.

(2a) If the multipath delay time equals the integer multiple of bit interval, then the timing relationship of signal bit frame and interference bit frame is shown in Figure 1.3.

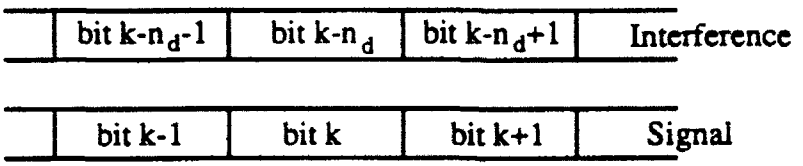


Figure 1.3 Signal and interference bit frame relationship

Since the delay n_d equal integer multiple of the PN period, then,

$$p(n-n_d) = p(n), \text{ and } p(n-n_d)p(n) = 1. \tag{1.14}$$

The second term of Equation (1.6) becomes

$$2E\{[\sum_{n=1}^{N} d(n)][\sum_{m=1}^{N} a\cos\theta d(m-n_d)p(m-n_d)p(m)]\}$$

$$= 2a\cos\theta E[\sum_{n=1}^{N} d(n) \sum_{m=1}^{N} d(m-n_d)] \tag{1.15}$$

Within a bit interval, all d(n) and d(m-n_d) are either 1 or -1 with equal probability. Also, since d(n) and d(m-n_d) are independent, Equation (1.15) is zero.
The third term of Equation (1.6) becomes

$$a^2\cos^2\theta E[\sum_{n=1}^{N} d(n-n_d) \sum_{m=1}^{N} d(m-n_d)] = N^2a^2\cos^2\theta \tag{1.16}$$

Hence, the signal to interference ratio at the output of the despreading circuit is

$$(S/I)_{des} = \frac{1}{a^2\cos^2\theta} \tag{1.17}$$

So, the processing gain for this case is

$$PG_D = 1 \tag{1.18}$$

Hence, under this situation, the despreading circuit do not provide any processing gain. Equalizer is the sole device that suppresses the multipath interference.

(2b) If the multipath delay time is not equal to the integer multiple of bit interval, then integrate the product of p(n) and $p(n-n_d)$ over a bit interval (period of PN) is

$$\sum_{n=1}^{N} p(n-n_d)p(n) = -1 \tag{1.19}$$

The mean value of the product of p(n) and $p(n-n_d)$ integrate over a fraction of bit interval τ (partial correlation) is

$$E[\sum_{\tau} p(n-n_d)p(n)] = 0 \tag{1.20}$$

The timing relationship of signal bit frame and interference bit frame is shown in Figure 1.4, where $\tau_1 + \tau_2 = N$.

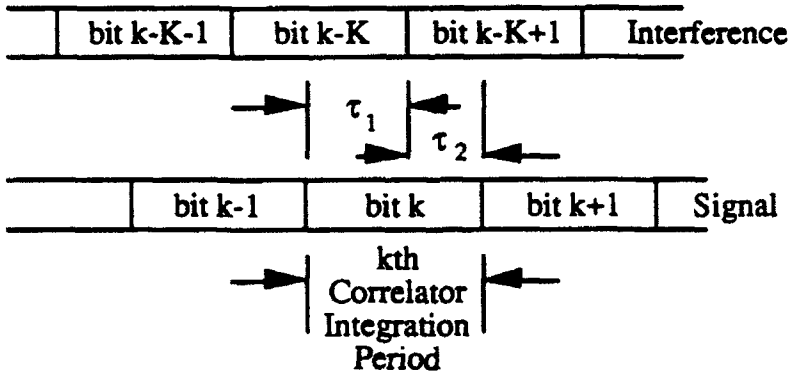


Figure 1.4 Signal and interference bit frame relationship

During the kth integration period, kth signal bit frame overlaps with two interference bit frames. There are three different relations among interference bit k-K, k-K+1 and signal bit k, and they are:

(i) Interference bits at k-K, k-K+1 and signal bit k have the same sign. The probability of this event is 1/4. Under this condition, the second term of Equation (1.6) becomes

$$2E\{[\sum_{n=1}^{N} d(n)][\sum_{m=1}^{N} a\cos\theta d(m-n_d)p(m-n_d)p(m)]\}$$

$$= 2Na\cos\theta E[\sum_{m=1}^{N} p(m-n_d)p(m)]$$

$$= -2Na\cos\theta \qquad (1.21)$$

The third term of Equation (1.6) becomes

$$a^2\cos^2\theta E\{[\sum_{n=1}^{N} d(n-n_d)p(n-n_d)p(n)][\sum_{m=1}^{N} d(m-n_d)p(m-n_d)p(m)]\}$$

$$= a^2\cos^2\theta E\{[\sum_{n=1}^{N} p(n-n_d)p(n)][\sum_{m=1}^{N} p(m-n_d)p(m)]\}$$

$$= a^2\cos^2\theta \qquad (1.22)$$

(ii) Interference bits k-K and k-K+1 have the same sign and they are opposite to the sign of signal bit k. The probability of this event is 1/4. Under this condition, the second term of Equation (1.6) becomes

$$2E\{[\sum_{n=1}^{N} d(n)][\sum_{m=1}^{N} a\cos\theta d(m-n_d)p(m-n_d)p(m)]\}$$

$$= -2Na\cos\theta E[\sum_{m=1}^{N} p(m-n_d)p(m)]$$

$$= 2Na\cos\theta \qquad (1.23)$$

The third term of Equation (1.6) becomes

$$a^2\cos^2\theta E\{[\sum_{n=1}^{N} d(n-n_d)p(n-n_d)p(n)][\sum_{m=1}^{N} d(m-n_d)p(m-n_d)p(m)]\}$$

$$= a^2\cos^2\theta E\{[\sum_{n=1}^{N} p(n-n_d)p(n)][\sum_{m=1}^{N} p(m-n_d)p(m)]\}$$

$$= a^2\cos^2\theta \qquad (1.24)$$

(iii) Interference bits k-K and k-K+1 have different sign. The probability of this event is 1/2. Under this condition, the second term of Equation (1.6) becomes

$$2E\{[\sum_{n=1}^{N} d(n)][\sum_{m=1}^{N} a\cos\theta d(m-n_d)p(m-n_d)p(m)]\}$$

$$= \pm 2Na\cos\theta E[\pm\sum_{\tau_1} p(m-n_d)p(m) \pm \sum_{\tau_2} p(m-n_d)p(m)] \qquad (1.25)$$

Equation (1.25) contains the mean value of two partial correlations, hence, it is zero.
The third term of Equation (1.6) essentially becomes

$$a^2\cos^2\theta E\{[\sum_{n=1}^{N} d(n-n_d)p(n-n_d)p(n)][\sum_{m=1}^{N} d(m-n_d)p(m-n_d)p(m)]\}$$

$$= a^2\cos^2\theta E\{[\sum_{\tau_1} d(n-n_d)p(n-n_d)p(n)+\sum_{\tau_2} d(n-n_d)p(n-n_d)p(n)]$$

$$[\sum_{\tau_1} d(m-n_d)p(m-n_d)p(m)+\sum_{\tau_2} d(m-n_d)p(m-n_d)p(m)]\}$$

$$= a^2\cos^2\theta E\{[\sum_{\tau_1} d(n-n_d)p(n-n_d)p(n)][\sum_{\tau_1} d(m-n_d)p(m-n_d)p(m)] +$$

$$[\sum_{\tau_2} d(n-n_d)p(n-n_d)p(n)][\sum_{\tau_2} d(m-n_d)p(m-n_d)p(m)]\}$$

$$= a^2\cos^2\theta[\tau_1 + \tau_2] = Na^2\cos^2\theta \qquad (1.26)$$

Combine the case of (i), (ii) and (iii), the second term of Equation (1.6) is zero and the third term of Equation (1.6) is

$$\frac{1}{2}(N+1)a^2\cos^2\theta \approx \frac{1}{2}Na^2\cos^2\theta, \qquad \text{for } N >> 1 \qquad (1.27)$$

Hence, the signal to interference ratio at the output of the despreading circuit is

$$(S/I)_{des} = \frac{2N}{a^2\cos^2\theta} \tag{1.28}$$

So, the processing gain for this case is

$$PG_D = 2N \tag{1.29}$$

In summary, the signal processing gain due to the frequency despreading circuit is

$$PG_D = \begin{cases} N & \text{for case (1)} \\ 1 & \text{for case (2a)} \\ 2N & \text{for case (2b)} \end{cases} \tag{1.30}$$

If the signal processing gain is expressed in dB, then the signal processing gain of the system PG_S is the sum of signal processing gains due to equalizer and despreading circuit.

$$PG_S = PG_E + PG_D \tag{1.31}$$

2. Least Mean Square Filtering

A least mean square (LMS) filter is a filter that minimizes the mean square error at the filter output. Most popular LMS filter structure is finite impulse response (FIR) filter because of its structure simplicity and mathematical tractability. Although the theoretical derivation of the LMS filter can be found in some advanced signal processing text[4-6], a summary of the LMS filter is present here as a quick reference.

The structure of FIR adaptive filter is shown in Figure 2.1.

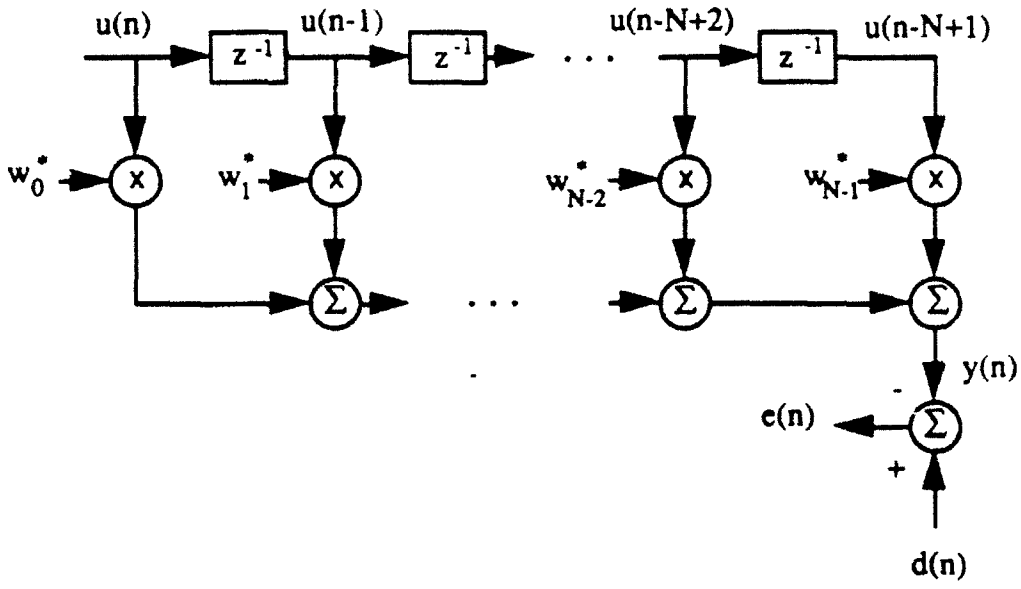


Figure 2.1 FIR adaptive filter

where u(n), y(n), d(n) and e(n) are the filter input, filter output, desired signal, and the error signal respectively. w_0^*, w_1^*, $\cdots$ w_{N-1}^* are the filter coefficients. The error signal e(n) = d(n) - y(n) is used to update the filter coefficients. The input u(n) and desired signal d(n) are assumed to have zero mean.

Define vectors **u**(n) and **w** as:

$$\mathbf{u}(n) = [u(n)\ u(n-1)\ \cdots\ u(n-N+1)]^T \qquad (2.1)$$

$$\mathbf{w} = [w_0\ w_1\ \cdots\ w_{N-1}]^T \qquad (2.2)$$

then the filter output y(n) is

$$y(n) = \mathbf{w}^H \mathbf{u}(n) \tag{2.3}$$

where $\mathbf{w}^H$ is the complex conjugate transpose of $\mathbf{w}$.

The error signal e(n) is

$$e(n) = d(n) - y(n)$$

$$= d(n) - \mathbf{w}^H \mathbf{u}(n) \tag{2.4}$$

The mean square error (MSE) ε is

$$\varepsilon = E[|e(n)|^2]$$

$$= E[e(n)e^*(n)]$$

$$= E\{[d(n) - \mathbf{w}^H \mathbf{u}(n)][d^*(n) - \mathbf{u}^H(n)\mathbf{w}]\}$$

$$= E[|d(n)|^2] - \mathbf{w}^H E[\mathbf{u}(n)d^*(n)] - E[d(n)\mathbf{u}^H(n)]\mathbf{w} +$$

$$\mathbf{w}^H E[\mathbf{u}(n)\mathbf{u}^H(n)]\mathbf{w}$$

$$= \sigma_d^2 - \mathbf{w}^H \mathbf{p} - \mathbf{p}^H \mathbf{w} + \mathbf{w}^H \mathbf{R}\, \mathbf{w} \tag{2.5}$$

where σ_d^2 is the variance of the desired sequence.

$\mathbf{p} = E[\mathbf{u}(n)d^*(n)]$ is the cross correlation vector between $\mathbf{u}(n)$ and $d(n)$.

$\mathbf{R} = E[\mathbf{u}(n)\mathbf{u}^H(n)]$ is the autocorrelation matrix of $u(n)$.

Equation (2.5) shows that the MSE ε is a quadratic function of filter coefficient vector $\mathbf{w}$. There is a unique minimum of ε. To find the unique minimum, simply take the gradient of ε with respect to coefficient vector $\mathbf{w}$ and set it to be a zero vector. From Equation (2.5), the gradient of MSE ε is

$$\nabla\varepsilon = -2\mathbf{p} + 2\mathbf{R}\,\mathbf{w} \tag{2.6}$$

If this gradient vector is zero, the least mean square error state is reached. The filter coefficient associate with this state is the optimal filter coefficient vector $\mathbf{w}_o$.

$$\mathbf{R}\mathbf{w}_o = \mathbf{p} \tag{2.7}$$

Equation (2.7) defined the relationship among $\mathbf{w}_o$, $\mathbf{R}$ and $\mathbf{p}$, is called the Wiener-Hopf equation.

The minimum MSE, ε_{min} is

$$\varepsilon_{min} = \sigma_d^2 - \mathbf{w}_o^H\mathbf{p} - \mathbf{p}^H\mathbf{w}_o + \mathbf{w}_o^H\mathbf{R}\mathbf{w}_o$$

$$= \sigma_d^2 - \mathbf{p}^H\mathbf{w}_o \tag{2.8}$$

Equation (2.7) shows that the optimal filter coefficients can be obtained if the correlation matrix $\mathbf{R}$ and cross correlation vector $\mathbf{p}$ are known. However, the information $\mathbf{R}$ and $\mathbf{p}$ are not available to the receiver system. To find the optimal filter coefficient vector $\mathbf{w}_o$, some other approach is needed. One popular approach is the steepest descent method.

Since the MSE is a quadratic function and has a unique minimum, this minimum can be reached by starting from a random state $\mathbf{w}(0)$, usually $\mathbf{w}(0) = \mathbf{0}$, and then follow the negative gradient of the MSE at that state, and then repeat this process recursively,

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{1}{2}\, a[-\nabla\varepsilon(n)]$$

$$= \mathbf{w}(n) + a[\mathbf{p} - \mathbf{R}\mathbf{w}(n)] \tag{2.9}$$

then, the optimal state $\mathbf{w}_o$ can eventually be reached. Using the definition of $\mathbf{p}$ and $\mathbf{R}$, in Equation (2.9), then the updating equation becomes

$$\mathbf{w}(n+1) = \mathbf{w}(n) + a\{E[\mathbf{u}(n)d^*(n)] - E[\mathbf{u}(n)\mathbf{u}^H(n)]\mathbf{w}(n)\} \tag{2.10}$$

In Equation (2.10), if the expectation is replaced by a single realization, then the computation is referred to as the stochastic gradient estimation method. The coefficient updating equation is

$$\mathbf{w}(n+1) = \mathbf{w}(n) + a\mathbf{u}(n)[d^*(n) - \mathbf{u}^H(n)\mathbf{w}(n)]$$

$$= \mathbf{w}(n) + a\mathbf{u}(n)[d^*(n) - y^*(n)]$$

$$= \mathbf{w}(n) + a\mathbf{u}(n)e^{*}(n) \tag{2.11}$$

If this updating process is performed recursively at every symbol interval, then effectively the mean value computation is carried over time. In component form, this computation algorithm can be represented by the following diagram.

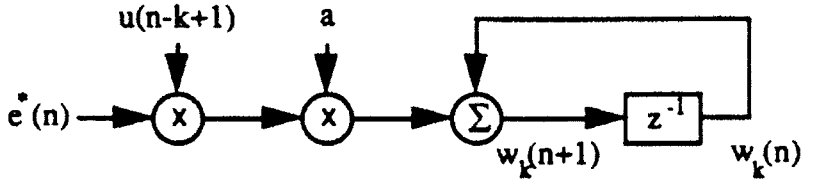


Figure 2.2 LMS filter coefficient updating algorithm

Note that the computation algorithm shown in Figure 2.2 contains a feedback loop. In order to guarantee the convergence of the filter coefficients, the adaptation constant, a, has to confined to a proper region.

Defined the coefficient error vector $\mathbf{v}(n)$ as

$$\mathbf{v}(n) = \mathbf{w}(n) - \mathbf{w}_o \tag{2.12}$$

then, the filter coefficient updating equation (2.9) can be modified as

$$\begin{aligned}\mathbf{v}(n+1) &= \mathbf{w}(n+1) - \mathbf{w}_o \\ &= \mathbf{w}(n) - \mathbf{w}_o + a[\mathbf{p} - \mathbf{R}(\mathbf{w}(n) - \mathbf{w}_o)] - a\mathbf{R}\mathbf{w}_o \\ &= \mathbf{v}(n) - a\mathbf{R}\mathbf{v}(n) \\ &= (\mathbf{I} - a\mathbf{R})\mathbf{v}(n)\end{aligned} \tag{2.13}$$

Using the similarity transformation[7], autocorrelation matrix $\mathbf{R}$ can be transformed into $\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^{H}$ where Λ is a diagonal matrix with diagonal elements λ_i, $i = 1, \cdots N$ being the eigenvalues of $\mathbf{R}$, and columns of matrix $\mathbf{Q}$ are the corresponding eigenvectors.

Since matrix $\mathbf{R}$ is conjugate symmetry, all its eigenvalues are real and positive, and its eigenvector matrix $\mathbf{Q}$ is Hermitian, i. e., $\mathbf{Q}^H = \mathbf{Q}^{-1}$.

Define $\mu(n) = \mathbf{Q}^H v(n)$, Equation (2.13) can be rewritten as

$$\begin{aligned}\mu(n+1) &= \mathbf{Q}^H v(n+1) \\ &= \mathbf{Q}^H(\mathbf{I} - a\mathbf{R})v(n) \\ &= \mathbf{Q}^H\mathbf{Q}(\mathbf{I} - a\Lambda)\mathbf{Q}^H v(n) \\ &= (\mathbf{I} - a\Lambda)\mu(n) \end{aligned} \tag{2.14}$$

Equation (2.14) is the canonical form of coefficient updating equation. If the filter coefficient vector is going to reach the optimal state, $w \rightarrow w_o$, vector $\mu(n)$ has to converge to a zero vector. The component form of Equation (2.14) is

$$\begin{aligned}\mu_k(n+1) &= (1-a\lambda_k)\mu_k(n) \\ &= (1-a\lambda_k)^{n+1}\mu_k(0) \quad k = 1, \cdots N \end{aligned} \tag{2.15}$$

Equation (2.15) states that as long as $|1-a\lambda_k| < 1$, for all k, eventually, filter coefficient vector will converge to its optimal state w_o. So, the condition for convergence is $|1-a\lambda_k| < 1$, for all k. Since all the eigenvalues are real and positive, this condition is

$$a < \frac{2}{\lambda_{max}} \tag{2.16}$$

where λ_{max} is the maximum eigenvector.

However, the information about eigenvalues is not available to the receiver, Equation (2.16) can not use directly.

A conservative estimation of the adaptation constant, a, can be obtained as follow:

Since all eigenvalues are real and positive then

$$\sum_{k=1}^{N} \lambda_k > \lambda_{max}$$

$$\sum_{k=1}^{N} \lambda_k = \text{trace}(\mathbf{R}) = Nr(0) = NP_{av} \tag{2.17}$$

where N is the number of filter taps, r(0) is the autocorrelation of the input sequence of zero lag, or the average input power P_{av}. Using the result of Equation (2.17) in Equation (2.16), the range of adaptation constant is

$$a < \frac{2}{NP_{av}} \tag{2.18}$$

The input signal power of the equalizer can be measured, hence, the adaptation constant, a, can be obtained from Equation (2.18). The updating equation (2.11) is based on stochastic gradient estimation. The adaptation constant used in the adaptive filter is usually much smaller than $\frac{2}{NP_{av}}$.

In summary, the LMS filter coefficient vector updating computation process is:

1. Start the process with an arbitrary initial guess of the filter coefficient vector. Usually an all zero vector is chosen as the filter initial state. $\mathbf{w}(0) = \mathbf{0}$.
2. Using the initial or the present guess, compute the gradient vector of the error performance surface evaluated at the present filter state.
3. Compute the next guess of filter coefficients by making a change of the present value in the direction opposite to the gradient vector (Equation (2.11)).
4. Go back to step 2 and repeat the process.

3. Computer Simulation

Assume the multipath channel impulse response is

$$h_c(n) = \delta(n) + .5\delta(n-1) - .2\delta(n-2) \qquad (3.1)$$

This non-ideal impulse response creates intersymbol interference (ISI). The average ISI introduced by this non-ideal channel is $10 \log(.5^2 + (-.2)^2) = -5.38$ dB.

Assume a FIR equalizer of 11 taps with tap weights $w_0, w_1, \ldots w_{10}$ is used to minimize the ISI. The overall impulse response of the system (channel and equalizer) $h_s(n)$ is the convolution of the channel impulse response $h_c(n)$ and the equalizer tap weights.

$$h_s(n) = \sum_{k=0}^{10} w_k h_c(n-k) \quad n = 0, 1, \ldots 12 \qquad (3.2)$$

The desired system impulse response should be a delayed impulse function such as $h_d(n) = \delta(n-n_d)$, where n_d is the system delay. Equation (3.2) contains 13 simultaneous linear equations. They can be expressed in a compact matrix form as:

$$\mathbf{A}\mathbf{w} = \mathbf{b} \qquad (3.3)$$

where $\mathbf{A} =$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdot & \cdot & 0 \\ .5 & 1 & 0 & 0 & \cdot & \cdot & 0 \\ -.2 & .5 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & -.2 & .5 & 1 & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdot & \cdot & -.2 \end{bmatrix}$$

$$\mathbf{w} = [w_0\ w_1\ w_2 \ldots w_{10}]^T$$

$$\mathbf{b} = [0\ 0\ \ldots 0\ \underset{\uparrow}{1}\ 0\ \ldots 0]^T$$

n_dth sample

Design an equalizer that minimizes the ISI is simply solving equation (3.3) for filter coefficients. Unfortunately, Equation (3.3) contains 13 linear equations with only 11 unknowns. i.e., it is an over specified problem. In general, there is no solution. However, the optimal approximation that minimizes the square error can be obtained from the following equation[7].

$$\mathbf{w}_o = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \qquad (3.4)$$

where $\mathbf{w}_o$ is the optimal solution in the sense of minimum square error. If the components of $\mathbf{w}_o$ are used as the equalizer coefficients, the system impulse response is given as:

$$\mathbf{h}_s = [h_s(0)\ h_s(1)\ \ldots\ h_s(12)]^T = \mathbf{A}\mathbf{w}_o \qquad (3.5)$$

Since $\mathbf{w}_o$ is the least square approximation to the solution of Equation (3.3), the residual ISI after equalization, I_{res}, is

$$I_{res} = [h_s(n_d) - 1]^2 + \sum_{k=0}^{12,} h_s^2(k) \qquad (3.6)$$

where $\sum_{k=0}^{12,}$ is the summation over all k for k not equal n_d.

To estimate the optimal delay n_d, the residual ISI computation of Equation (3.6) has to be carried out for different system delay n_d. The one with the least amount of residual ISI is the optimal system delay. For this example, the optimal system delay is $n_d = 0$. Figure 3.1 shows the system (channel and equalizer) impulse response for $n_d = 0, 1, 2$ respectively. For each case, the system impulse response is approximately an impulse function with the peak at the system delay sample. As the system delay n_d increases, the tail amplitude of the impulse response also increases, thus resulting in a larger amount of residual ISI.

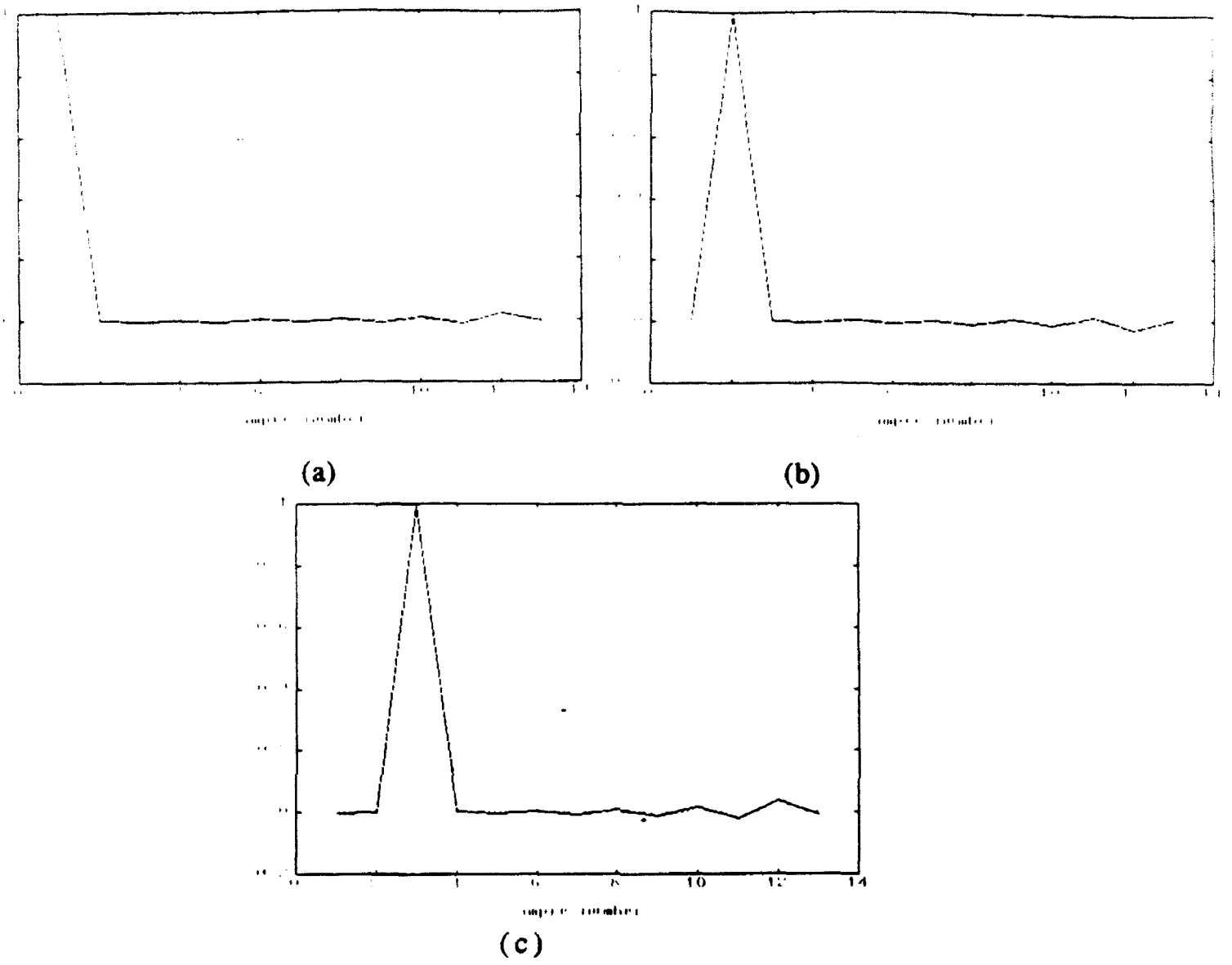


Figure 3.1 System impulse response for $n_d = 0$ (a), 1 (b), and 2 (c)

For the optimal system delay $n_d = 0$, the residual ISI is $I_{res} = 8.5387 \times 10^{-4} = -30.69$ dB. Compared to the unequalized ISI of -5.38 dB, an 25.31 dB ISI suppression is achieved by using an 11 tap equalizer. Even when the timing delay estimation is off by 2 symbol interval, the 11 tap equalizer can still provide 20.64 dB ISI suppression. A longer tap delay line equalizer can provide an even more ISI suppression. Table 3.1 shows for $n_d = 0$, the residual ISI is decreased as the length of equalizer (N) increase. The numbers listed in Table 3.1 represent the maximum available ISI suppression for a given tap length. The actual performance of an equalizer will be worse than the number shown in Table 3.1 due to the particular

computation algorithm and the finite register size used in the tap delay line and equalizer coefficients.

Table 3.1 Residual ISI as a function of equalizer size

N	11	12	13	14	15
I_{res} (dB)	-30.69	-33.05	-35.40	-37.76	-40.12

The frequency responses of the multipath channel, the equalizer, and the system (channel and equalizer) are shown in Figures 3.2 to 3.4. Figure 3.2 shows the multipath channel effectively amplifying the low frequency components and attenuating the high frequency components. Figure 3.3 shows the frequency response of equalizer as it tries to compensate for the non-uniform channel frequency response. Figure 3.4. display the system frequency response; note the uniformity of the overall response as the equalizer tries to compensate the channel effects.

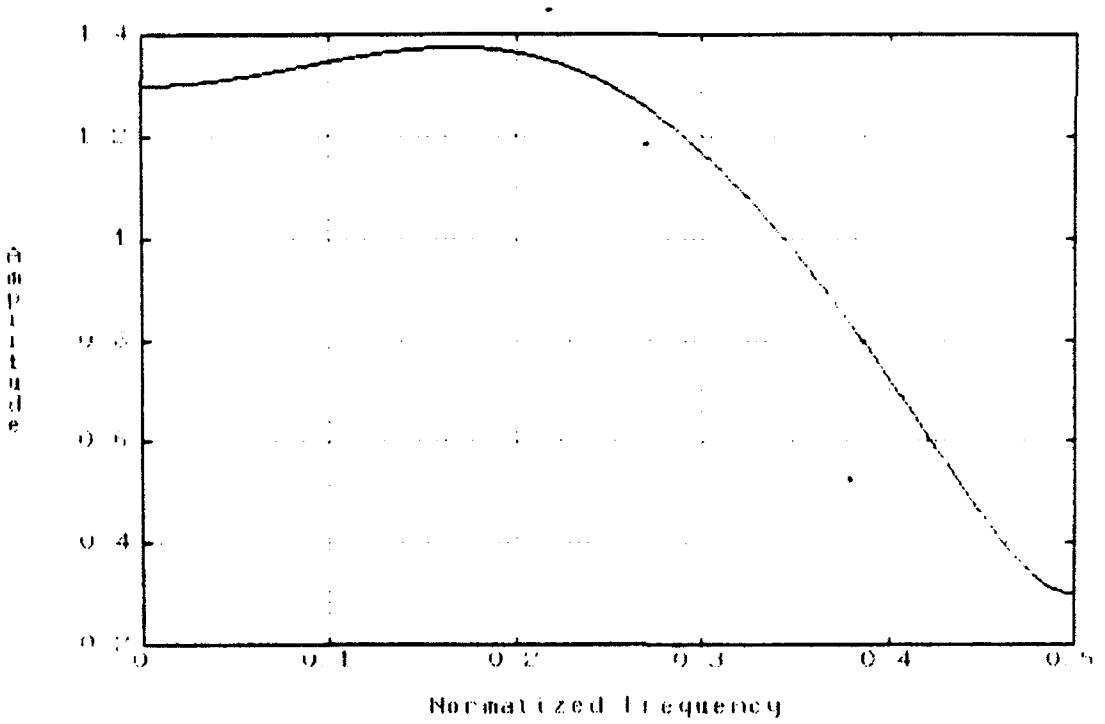


Figure 3.2 Multipath channel amplitude frequency response

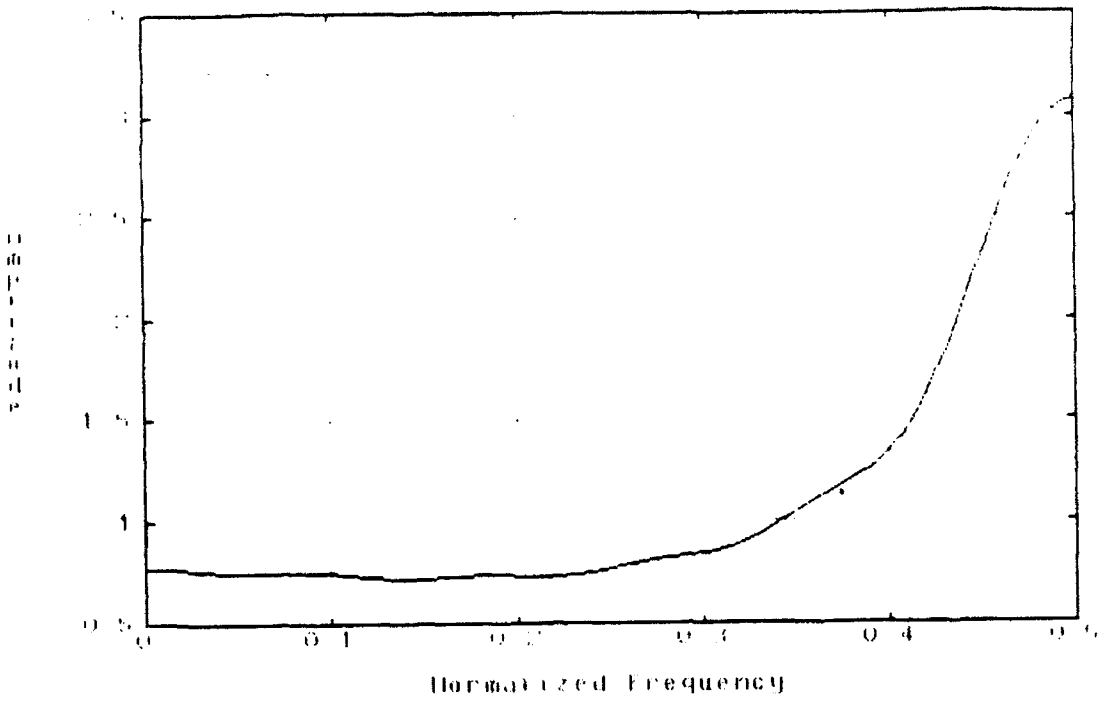


Figure 3.3 Equalizer amplitude frequency response

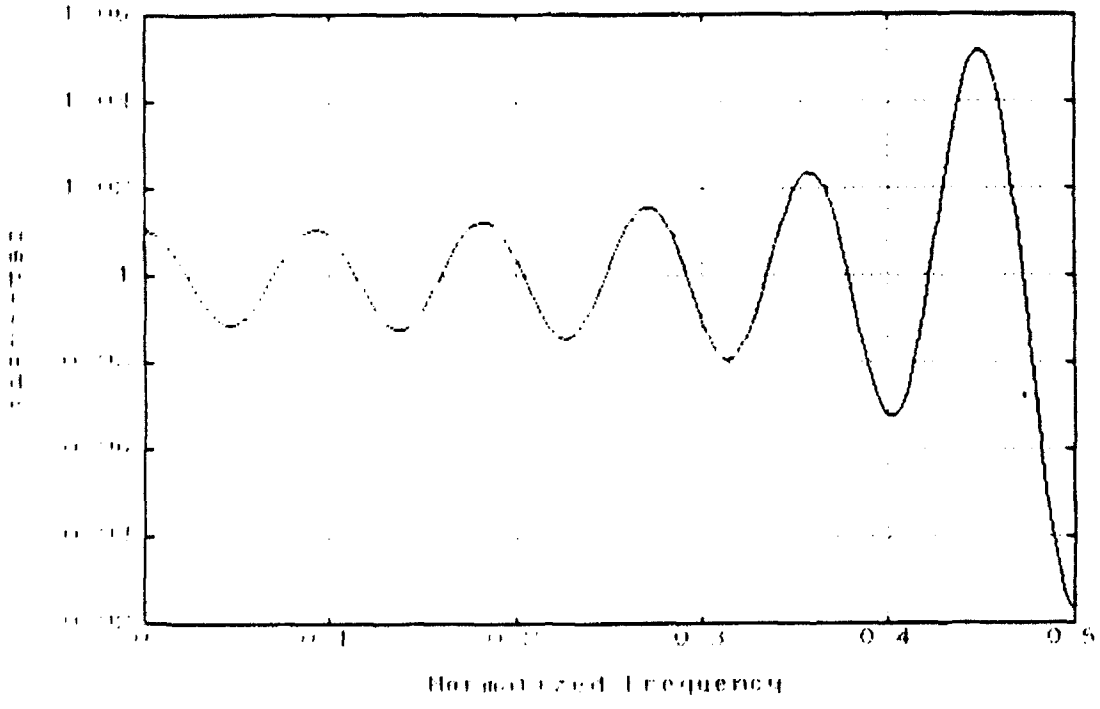


Figure 3.4 System amplitude frequency response

Equation (3.4) states that the optimal equalizer tap coefficients can be obtained by performing a sequence of matrix multiplications and inversion. However, it should be noted that these operations are applicable only if the following conditions are met.

1.The channel impulse response is known.

2.The channel characteristics are stationary.

In a real operating environment, the channel impulse response is usually unknown and it may be time varying. Adaptive filter algorithms do not assume any channel characteristics. For a suitable adaptation constant, it will track time varying channel behavior. Hence, these algorithms can be used to compute the equalizer tap coefficients.

The block diagram shown in Figure 3.5 is used to model a communication system. Note that frequency spread and despread, modulation and demodulation functions are inversion operations, for convenience, they are not included in this block diagram.

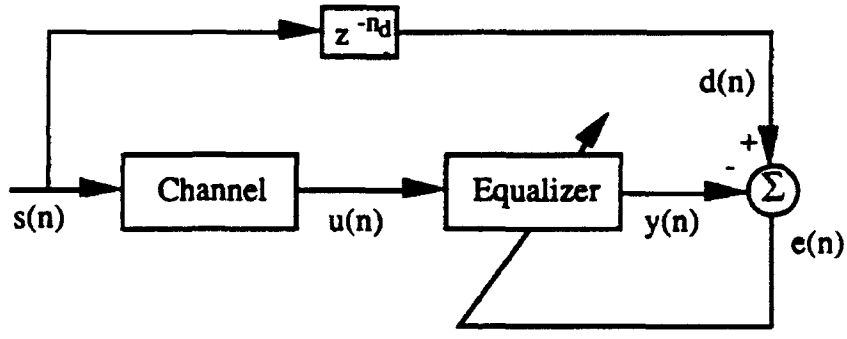


Figure 3.5 Block diagram of communication system

In this block diagram, the signal s(n) is generated from a uniform [0,1] random number generator. When the random number is between 0 and 0.5, assign s(n) = -1, otherwise s(n) =1. The desired signal d(n) is the signal s(n) delayed by the system delay time n_d. The error signal e(n), which is the difference between the desired signal d(n) and the equalizer output y(n), is used to update the equalizer tap coefficients.

In this simulation, the equalizer tap coefficients are assumed to be zero initially. The steepest descent tap updating equation is given by Equation (3.7).

$$w(n+1) = w(n) + au(n)e(n) \qquad (3.7)$$

Since the signal s(n) is generated by a random number generator, sequence s(n) is uncorrelated. The channel impulse response is h(n) = [1, .5, -.2]. The correlation matrix **R** for the equalizer input sequence u(n) is quintdiagonal. That is, the only nonzero elements of **R** are on the main diagonal and four diagonals directly above and below it.

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & r(2) & 0 & \cdot & \cdot & 0 \\ r(1) & r(0) & r(1) & r(2) & \cdot & \cdot & 0 \\ r(2) & r(1) & r(0) & r(1) & \cdot & \cdot & 0 \\ 0 & r(2) & r(1) & r(0) & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdot & \cdot & r(0) \end{bmatrix} \qquad (3.8)$$

where $r(0) = 1^2 + .5^2 + (-.2)^2 = 1.29$
$r(1) = 1x.5 + .5x(-.2) = .4$
$r(2) = 1x(-.2) = -.2$

The eigenvalues of matrix **R** are: {0.1655, 0.3772, 0.6846, 1.0308, 1.3562, 1.6101, 1.7085, 1.7716, 1.7734, 1.8518, 1.8603}.

To guarantee the equalizer coefficient converge to optimal state, the adaptation constant, a, has to satisfy Equation (3.9).

$$a < \frac{2}{\lambda_{max}} = \frac{2}{1.8603} \qquad (3.9)$$

The choice of an adaptation constant is a trade off between the convergence time and the misadjustment due to gradient noise. The adaptation constant used in the equalizer is usually much smaller than the number specified by Equation (3.9).

The convergence time for the equalizer will be shorter with a larger adaptation constant. But, as a result, it processes a larger amount of misadjustment at steady state.

Learning curve is the plot of the square error $e^2(n)$ as a function of sample time n. Figures 3.6 and 3.7 are the learning curves for the adaptation constant a equal .1 and .02, respectively.

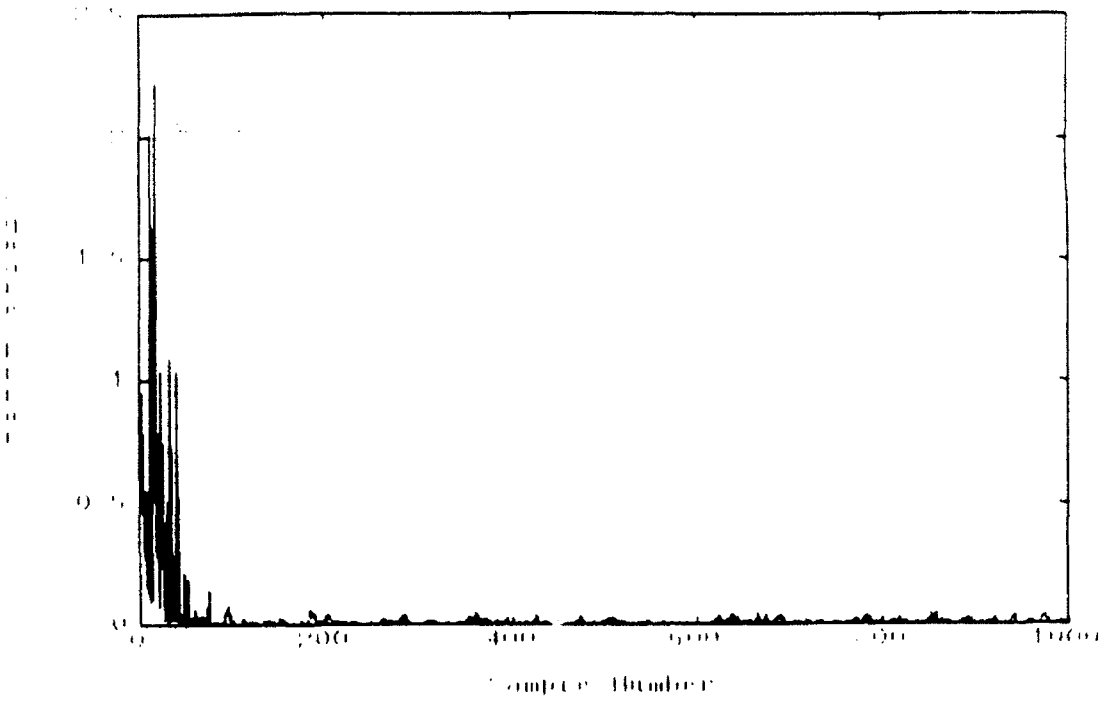


Figure 3.6 Learning curve for a = .1

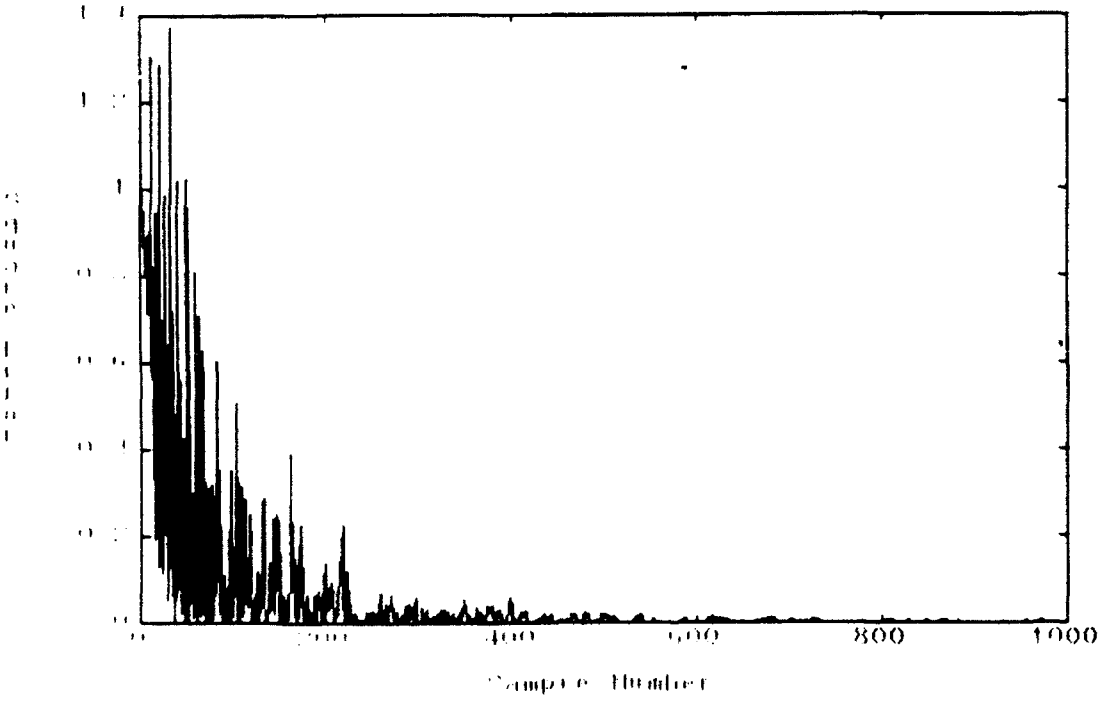


Figure 3.7 Learning curve for a = .02

Figures 3.6 and 3.7 show that the error of equalizer with higher adaptation constant converges faster than that with smaller adaptation constant. But when they converge, the equalizer with smaller adaptation constant has smaller residual error.

Optimal equalizer tap coefficients derive from Equation (3.4) and averaged steady state equalizer coefficients with adaptation constant a = .02, over 500 sample interval (from sample 1000 to 1500), from the steepest descent algorithm are listed in Table 3.2. In Table 3.2, the first and third columns are the optimal filter coefficients for delay n_d = 0 and 2 respectively. The second and fourth columns in Table 3.2 are the averaged steady state equalizer coefficients with delay n_d = 0 and 2 respectively. The numerical values in columns 1,2 and columns 3,4 are very close, which indicates that the steepest descent adaptive filter algorithm indeed converges the equalizer coefficients to the optimal state.

Table 3.2 Optimal equalizer tap coefficients and averaged equalizer tap coefficients derived from steepest descent algorithm.

	n_d= 0		n_d= 2	
	Optimal	Averaged	Optimal	Averaged
w_0	.9991	.9970	-.0015	-.0045
w_1	-.4985	-.4948	.0027	.0080
w_2	.4476	.4435	.9958	.9900
w_3	-.3216	-.3169	-.4941	-.4875
w_4	.2478	.2439	.4419	.4368
w_5	-.1849	-.1805	-.3140	-.3080
w_6	.1376	.1336	.2379	.2324
w_7	-.1001	-.0978	-.1719	-.1693
w_8	.0702	.0696	.1208	.1209
w_9	-.0450	-.0460	-.0774	-.0798
w_{10}	.0248	.0244	.0427	.0424

Once the equalizer reaches the steady state, the numerical value of its tap coefficients fluctuates around the optimal values. Variances of each filter coefficient are computed and listed in Table 3.3. Table 3.3 shows that the numerical values of the variance of filter coefficient increase as the adaptation constant increases. Optimal adaptation constant is a system design trade off among converging time, amount of misadjustment, and capability to track the channel time varying conditions.

Table 3.3 Variance of Equalizer Coefficients

	$n_d = 0$		$n_d = 2$	
	a = .1	a = .02	a = .1	a = .02
w_0	.2237e-3	.6679e-5	.6636e-3	.1685e-4
w_1	.1890e-3	.9872e-5	.5592e-3	.2396e-4
w_2	.1688e-3	.1732e-4	.4983e-3	.4457e-4
w_3	.1758e-3	.1555e-4	.5201e-3	.3607e-4
w_4	.2044e-3	.1987e-4	.6050e-3	.4685e-4
w_5	.1667e-3	.2021e-4	.4933e-3	.4762e-4
w_6	.1723e-3	.1057e-4	.5100e-3	.2341e-4
w_7	.1738e-3	.1103e-4	.5145e-3	.2514e-4
w_8	.1980e-3	.1525e-4	.5863e-3	.3829e-4
w_9	.1553e-3	.1017e-4	.4598e-3	.2651e-4
w_{10}	.1577e-3	.5526e-5	.4668e-3	.1541e-4

In order to achieve fast convergent time and small coefficient variances, "Gear shift" mode is used. Gear shift mode starts the coefficients updating with larger adaptation constant to ensure fast converging time, then shift to smaller adaptation constant to reduce the steady state fluctuation.

4. Equalizer Training Sequence

In the simulation exercises, the desired response is derived from a transmitted signal with appropriate delay. Thus, the error information generated in Figure 3.5 is the true error. Adaptive equalizer converges to the optimal state only if the true error signal is used in the coefficient updating. In the communication receiver system, the desired information is not available. It is thus necessary to use a set of proper training sequences to set up the receiver system. For binary antipodal signaling systems, signal space contains only two points A(+1) and B(-1). A possible set of training sequences is shown in Figure 4.1[8].

Figure 4.1 Training sequences

Each training sequence and its objective are listed as follows:

(a) The P1 sequence is a. constant DC (point A or B) in the baseband, or the carrier frequency in the passband. The main purpose of this sequence is to properly set up the automatic gain control (AGC) circuit of the receiver. If the initial received signal level is too low, then the AGC uses a higher gain factor. Otherwise, a lower gain factor is used. P1 sequence can be detected by observing a constant signal $\hat{s}(t)$ over an extended period of time.

(b) The P2 sequence is an alternate signal between two fixed signal points. For antipodal signaling, the P2 sequence is ABABAB $\cdots$ over an even number of symbol interval. The receiver system usually has a low pass filter (LPF) to filter out the twice carrier components due to demodulation processing. The alternating sequence at the receiver LPF output is a pure tone with period equals twice the symbol period. Hence, it is referred to as the P2 sequence. The purpose of the P2 sequence is to quickly acquire a proper timing sample for the equalizer processor to do further signal processing. Before the proper timing sample can be selected, the receiver first has to detect the presence of P2 sequence.

(b-1) P2 Sequence Detection

Assume that the phase lock loop (PLL) properly synchronizes the transmitter and receiver carrier phase, then during the P2 period, the demodulator output waveform is a pure sinusoid with

frequency f_2 equal half the symbol rate. If the symbol interval T and sampling period T_s are related by $T = kT_s$, then the received waveform during the P2 period is a pure tone with frequency equal to $\frac{1}{2k}$ of sampling frequency. A narrowband band pass filter (BPF) with passband centered at frequency $\frac{1}{2k} f_s$ can be used to detect the presence of P2 sequence. If the energy at the output of this BPF exceeds certain threshold, then the receiver declares that the P2 sequence is detected.

Another possible P2 sequence detection algorithm uses the property that since P2 sequence is a sinusoid with frequency equal half the symbol rate. Thus the sum of two samples spaced one symbol interval apart is zero. If the sum of two samples spaced one symbol interval is smaller than some threshold Δ for an extended period of time (NTh samples), then the receiver declares that the P2 sequence is detected. Using this algorithm, the BPF processing can be eliminated. Flow diagram of this algorithm is shown in Figure 4.2. Notation used in Figure 4.2 are:

n	sample index
k	number of samples per symbol interval
$\hat{s}(n)$	demodulated sample
CNT	P2 sequence counter
NTh	threshold count

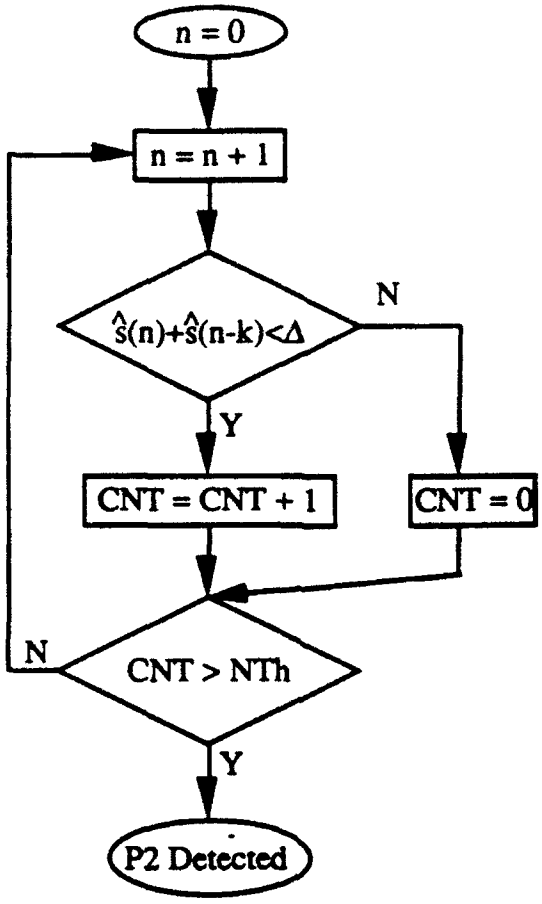


Figure 4.2 Flow diagram of P2 detection algorithm

(b-2) Timing Acquisition

Since P2 sequence is a pure tone, Observe this tone over a period interval (two symbol intervals), the sample that has the maximum absolute value is the proper timing sample. Flow diagram of timing acquisition is shown in Figure 4.3. Notation used in Figure 4.3 are:

n	sample index
$\hat{s}(n)$	demodulated sample
Ts	timing sample
k	number of samples per symbol interval

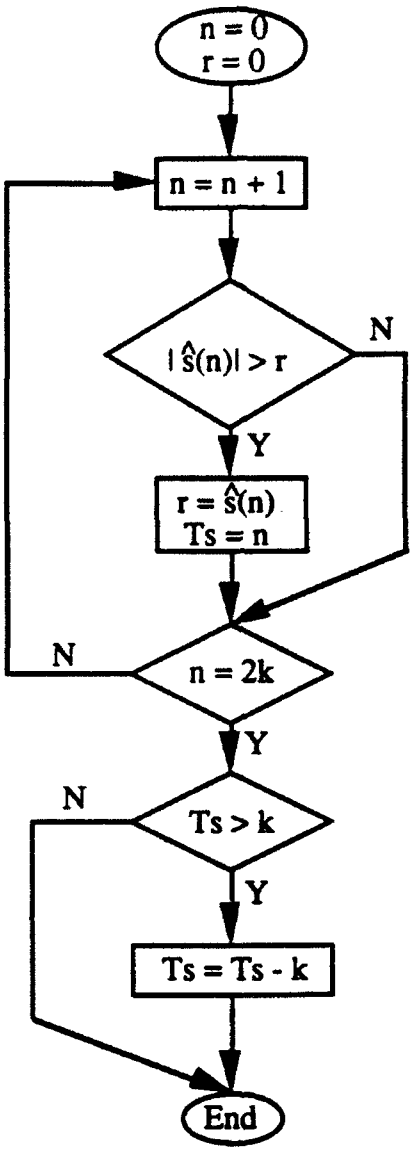


Figure 4.3 Timing acquisition flow diagram

The decision processor uses the timing sample to decide whether the transmitted symbol is A or B. Whenever this sample is greater than zero, then it decides that the received symbol is A. Otherwise, the received symbol is B.

Timing acquisition is a "coarse" time adjustment that allows the receiver to quickly derive the proper timing sample. After this sample is obtained, timing recovery loop takes over to perform the "fine tuning" of timing adjustment.

(c) The PN sequence is a pseudo random sequence generated by a known PN sequence generator. This PN sequence is not used to spread the signal spectrum. Rather, it is used to generate the error

information in the receiver so that the equalizer coefficient updating is based on the "true error" signal.

For example, in a binary antipodal signaling system, PN sequence is a random sequence between point A and B with B being the starting point. Since point B is also the end point of P2 sequence, whenever the receiver decision processor detects a repeat signal point, that point is the starting point of PN sequence. Upon detecting the PN sequence, the local receiver generates the identical PN sequence. Locally generated PN sequence can be used as the "desired signal". The difference between this desired signal d(n) and the equalizer output y(n) is the true error information. Use this error information to update the equalizer tap coefficients guarantee the tap coefficient vector will converge to the optimal state w_o. Figure 4.4 shows the equalizer coefficient updating block diagram during the PN sequence. Error used to update equalizer coefficient vector is the difference between the local PN generator output and the equalizer output.

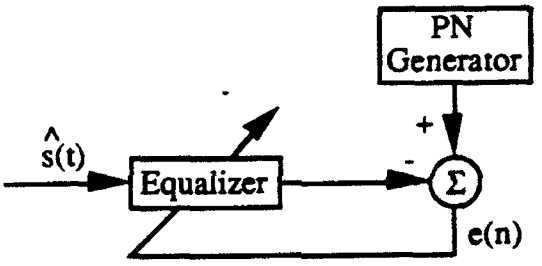


Figure 4.4 Equalizer coefficient updating during PN period

Since PN is a known sequence of fixed length, the receiver can easily estimate the beginning point of the random data. The length of PN sequence should be long enough so that at the end of PN period, the error e(n) is converged to a reasonable small value.

Once the receiver enters the data mode, the error signal is small enough such that the equalizer output signal point is within the decision boundary of the ideal signal point. Error information used to update the equalizer coefficient vector is derived from the difference between the decision processor output and equalizer output. For example, if binary antipodal signaling is used, decision processor output is A whenever equalizer output is greater than zero. Otherwise, decision processor output is B. Figure 4.5 show the block diagram of equalizer coefficient updating during the random data period.

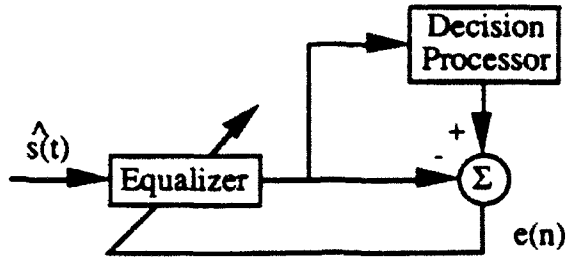


Figure 4.5 Equalizer coefficient updating during random data period

Figure 4.6 shows the receiver turn on sequence when an adaptive equalizer is used to compensate the multipath channel distortion.

Immediate after the receiver is turned on, the P1 sequence detector is enabled. Once the P1 sequence is detected, the receiver enters the phase ϕA. During this phase, receiver measures the amplitude of the received signal $\hat{s}(t)$. If this amplitude is smaller than the desired amplitude, then the AGC uses a higher gain factor. Otherwise, a lower gain factor is used. Once the gain factor is properly set, P2 sequence detector is turned on.

When P2 sequence is detected, the receiver performs timing acquisition to pick up the proper timing sample. The subsequent receiver signals are processed at the symbol rate. After the timing sample is defined, the receiver turn on the PN sequence detector.

The input to the PN sequence detector is the timing sample. At every symbol interval, the decision processor decides whether signal point A or B is received. The beginning of PN sequence is the first time the decision processor makes a repeat decision. At this point, the receiver generates an identical PN sequence, and enable the equalizer. The difference between the PN generator output and equalizer output is the error information. This error information is used to update the equalizer coefficients. If the equalizer operates in gear shift mode, phase ϕC may be further divided into several sub-phases. Starting with a relatively high adaptation constant to ensure fast convergence, the adaptation constant is then successively reduced to a smaller value at each subsequent sub-phase to reduce the misadjustment.

When PN sequence is detected, the receiver starts the PN counter. Every symbol interval, the PN counter is increased by one. When the PN counter count equals the length of the PN sequence (LPN), the receiver declares that it enters the random data mode.

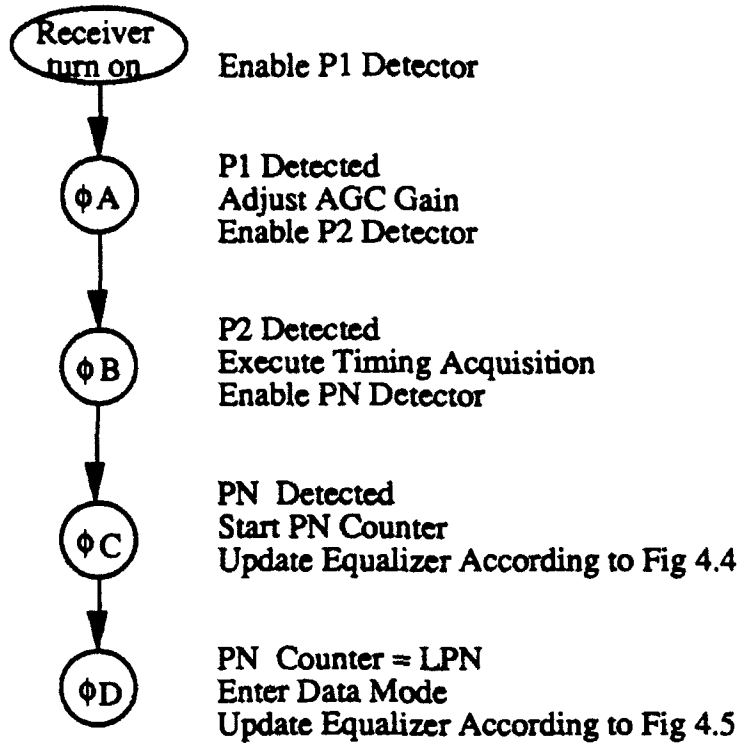


Figure 4.6 Receiver turn on sequence

The turn on sequence indicated in Figure 4.6 only consider the operation of adaptive equalizer. If the other servo loops (phase lock loop, timing recovery loop, etc) are also considered, the turn on sequence has to substantively modified. Also, in order to improve the performance of adaptive equalizer, fractional spaced equalizer[9-10] is considered by several researchers. Detailed system simulations by including all the servo loops and various equalization techniques are proposed as the possible subjects for further research work.

References

1. L. Milstein and R. Iltis
   "Signal Processing for Interference Rejection in Spread Spectrum Communications"
   IEEE ASSP Magazine, April, 1986
2. A. Papoulis
   "Probability, Random Variables and Stochastic Processes"
   p. 345
   McGraw-Hill, 1991
3. G. Cooper and C. McGillem
   "Modern Communications and Spread Spectrum"
   p. 294
   McGraw Hill, 1986
4. B. Widrow and D. Stern
   "Adaptive Signal Processing"
   Prentice Hall, 1985
5. S. Haykin
   "Adaptive Filter Theory"
   Prentice Hall 1991
6. J. Lim and A. Oppenheim
   "Advanced Topics in Signal Processing"
   Prentice Hall 1988
7. G. Strange
   "Linear Algebra and its Applications"
   Harcourt Brace Jovanovich, 1988
8. CCITT Recommendations V27 and V29
9. R. Giltin and S. Weinstein
   "Fractionally-Spaced Equalization: An improved digital traversal equalizer"
   Bell System Technical Journal Vol 60, p. 275-296, 1981
10. F. Ling
    "On the Convergence Speed of Fractionally-Spaced Equalizer Using Intersymbol Interpolation"
    Proc. Intl. Conf. Comm. Tech, 1987

AN INVESTIGATION OF THE BENCHMARK EVALUATION TOOL

Khosrow Kaikhah
Chair
Department of Computer Science

Huston-Tillotson College
1820 East 8th Street
Austin, Texas 78702-2793

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

August 1992

AN INVESTIGATION OF THE BENCHMARK EVALAUATION TOOL

K. Kaikhah
Chair, Department of Computer Science
Huston-Tillotson College

ABSTRACT

Recently, natural language processing has received tremendous support and popularity. As a consequence, the number of natural language processing systems has dramatically increased and the need for a systematic evaluation procedure of such systems seems inevitable. Until recently, there has not been a universal evaluation procedure for evaluating all types of NLP systems. Evaluations of such systems are usually conducted during the implementation phase and, in most cases, do not involve a comprehensive plan or independent evaluators. Developers of NLP systems can benefit from an unbiased evaluation procedure which measures their efforts and the power of their systems. At the same time, the consumers of NLP systems can benefit greatly from an evaluation tool which assists with the selection of the appropriate system for their needs.

The Calspan Corporation has proposed and implemented the Benchmark Evaluation Tool for evaluating all natural language processing systems, regardless of type or application. The study was sponsored by the Rome Laboratory and was concluded in May of 1992. The Benchmark Evaluation Tool is designed to be domain independent. Therefore, it concentrates on the linguistic issues rather than on the application domain. This feature is unique, in that, the tool is sensitive to each individual linguistic capability and not to each individual application. It is composed of twelve independent sections which are designed to progressively test different linguistic features of NLP systems. The Benchmark Evaluation Tool also includes definitions and explanations for each section as well as a five-choice scoring strategy to measure the responses.

Our objective is to investigate the effectiveness of the Benchmark Evaluation Tool by applying the tool to a natural language processing system. This particular system is composed of two major parts: a domain-independent part which has general knowledge of syntactic rules, and a domain-specific part which provides the necessary semantic and pragmatic knowledge for a specified domain. The application domain accompanying the NLP system for testing purposes is an interface to a relational database of *air travel planning information.*

AN INVESTIGATION OF THE BENCHMARK EVALAUATION TOOL

K. Kaikhah

1. Introduction

Although natural language processing has been on the minds of researchers from the early days of the inception of digital computers, it has never enjoyed such a tremendous popularity and support as it has received over the past two decades. As the number of natural language processing systems has increased, so has the need for a systematic evaluation procedure for testing NLP systems. Both producers and consumers of NLP systems can benefit from a well defined evaluation procedure. It can help the producers with conducting an unbiased evaluation of their systems, and can help the consumers with choosing the appropriate system for their needs. The evaluation procedure should not be defined for a particular system, but rather as a blueprint for testing the linguistics features of NLP systems.

Until recently, evaluation procedures have been implemented and administrated by the developers of NLP systems. As a result, evaluations tend to be biased and follow known success patterns. These patterns may not be deliberate, but nevertheless it is the result of being so involved with the development. Therefore, a number of NLP researchers and consumers have expressed their needs and desires for an unbiased and independent evaluation procedure. One should keep in mind that a universal evaluation procedure which can be applied to all NLP systems may be too ambitious. However, a foundation for evaluating systems can be laid out to guide the producers, consumers, as well as the independent evaluators through the evaluation.

The Benchmark Investigation/Identification program sponsored by the Rome Laboratory developed an evaluation tool and application procedure for evaluating natural language processing systems. The duration of the project was eighteen months; it was completed in May of 1992. It produced an evaluation procedure consisting of twelve sections. Each section is designed to test a different linguistics capability of NLP systems and provides brief explanations and definitions of the linguistic feature being tested, patterns that define the structure of the test sentence, example sentences, and criteria against which to evaluate the behavior of the NLP system. Each test sentence is then scored according to the level of system's comprehension. It can range from success (S) to Partial success (P) to No output (N). For more details, see [1].

Most applications of NLP systems involve interactive human-computer interfaces which include: a) Data Base Management Systems, b) Command and Control Systems, c) Decision-Aiding Systems, d) Engineering Design Systems, and e) Diagnostic Systems. The natural language processing system which is used for this investigation is equipped with an interface to a relational database. The system can respond to questions about ground transportation, fares, and flights for the cities of Atlanta, Boston, Baltimore, Denver, Dallas, Fort Worth, Pittsburgh, Philadelphia, Oakland, San Francisco, and Washington D.C. The NLP system analyzes the English sentences with three independent modules *syntactic, semantic, and pragmatic* in order to transform the sentences into application calls. Twenty four different switches control the behavior of the NLP system. By setting the appropriate switches, the system can be prompted to learn unknown grammatical structures and words. This process, however, requires a knowledgeable linguistic trainer, since the NLP system expects meaningful linguistic feedback during training. The parse tree as well as the semantic, and pragmatic analysis of sentences can also be examined, if so desired, by setting the appropriate switches.

The goal of this investigation has been to determine the feasibility and usefulness of a universal evaluation procedure, namely the Benchmark Evaluation Tool. The Benchmark Evaluation Tool is designed to be applicable to all types of NLP systems, therefore, it can be considered to be a universal evaluation tool. We have applied the Benchmark Evaluation Tool to an NLP system and the comprehensive results are included in section 4. The following sections briefly describe the Benchmark Evaluation Tool and the NLP system, respectively.

2. The Benchmark Evaluation Tool

In May of 1992, a Rome Laboratory sponsored project, The Benchmark Investigation/Identification Program was completed by Calspan Advanced Technology Center and their subcontractor, Language Systems Incorporated. The goal of the project was to develop a standard evaluation tool which is domain-independent and which can be applied to all NLP systems, regardless of their types, and without any need for modifying or porting the NLP system to a test domain. For more details, see [1].

There are several areas in which NLP systems can be evaluated. They include: a) linguistic competence, b) end user issues such as reliability and likability, c) system development issues such as maintainability and portability, and d) intelligent behavior issues such as learning and cooperative dialogue. The Benchmark Evaluation Tool has focused on

linguistic competence of NLP systems including lexical, syntactic, semantic, and discourse capabilities. It consists of twelve sections, each of which tests a different feature of the NLP systems. They are: I) Basic Sentences, II) Interrogative Sentences, III) Noun Phrases, IV) Adverbials, V) Verbs and Verb Phrases, VI) Quantifiers, VII) Comparatives, VIII) Connectives, IX) Embedded Sentences, X) Reference, XI) Ellipsis, and XII) Semantics of Events.

The Evaluation Tool is designed for people with non-linguistics backgrounds, therefore it provides instructions and explanatory materials for each section. These materials are provided to assist the evaluators with the creation or tailoring of test sentences and do not include a set of predefined natural language test sentences. The testing is conducted in a progressive manner from elementary sentence types to more complex sentence types. This strategy allows the evaluator to concentrate on a single linguistic feature in each test sentence. If the NLP system fails on a certain linguistic feature, the evaluator is advised not to include the feature in subsequent test sentences. The scoring is done according to the following criteria [1]:

- Success (S): The system successfully met the evaluation criteria stated for the particular test item.
- Correct (C): The system did not successfully meet the evaluation criteria, but produced acceptable/correct output.
- Partially Correct (P): The system did not successfully meet the evaluation criteria, and only produced partially acceptable/correct output.
- Failure (F): The system did not successfully meet the evaluation criteria and produced no correct output.
- No Output (N): The system produced no output.

In short, the Benchmark Evaluation Tool is a procedure that a) produces profiles of NLP systems which are descriptive, hierarchically organized, quantitative, and objective, b) is usable across domains and applications, c) is usable across the different types of NLP systems, and d) is unbiased with respect to linguistic theories and does not require an evaluator who is a trained linguist. In fact, the Evaluation Tool is unique in two features [1]:

- The profiling facility
- Its usability and applicability across domains and applications

3. The NLP System

The NLP system, used in our investigation, consists of two major components: a domain-independent (core) component, and a domain-specific (application) component. The domain-independent routines which include the procedural components for syntactic, semantic, and pragmatic analysis as well as large portions of the grammar and lexicon do not change during porting. However the domain-specific routines which include specialized lexicon, semantic rules, knowledge base, and application-specific routines must be re-implemented to accommodate a new application.

Three distinct modules *syntactic, semantic, and pragmatic* analyze and process the input sentences independently. The syntactic module requires knowledge of lexicon and grammar rules; the semantic module requires the services of semantics rules; and the pragmatic module requires domain knowledge. The NLP system contains twenty four switches which control its behavior. The behavior of each module can be seen by setting the appropriate switches. In certain modes, however, the switches only control what may appear on the screen and not the processing that is going on in the background.

Input to the system can be from external files or keyboard. The output can take several forms depending on the configuration of the switches. The syntactic analysis of the sentences can produce two types of output: a detailed surface structure parse tree, and an operator-argument representation called the Intermediate Syntactic Representation (ISR). ISR is the simplified version of the parse tree with a single canonical form for a number of various surface structures and a lot less detail that is not required by the semantic analyzer. The semantic and pragmatic modules use the ISR as input and produce the Integrated Discourse Representation (IDR). IDRs are application-neutral representations of the meaning of the sentences in the current discourse containing situations described in the input sentences, the entities referred to, and the way the entities participate in the situations.

4. Applying the Benchmark Evaluation Tool to the NLP system

The application domain which accompanied the NLP system for this investigation is an interface to a relational database of air travel planning information. Test sentences were scored according to the syntactic, semantic, and pragmatic processing and comprehension of each sentence. Generally, the sentences which failed to produce an IDR (i.e. failed the analysis of the

core components of the NLP system) received an N, those which produced an IDR but not an application call (i.e. passed the analysis of the core components of the NLP system but produced an incomplete IDR) received an F, sentences which produced an IDR and a partial application call, received either a P or a C depending on the completeness of the application call, and finally, sentences which produced an IDR and a complete application call received an S.

Approximately 88% of the sentence patterns suggested by the Benchmark Evaluation Tool were applicable to the application domain, in other words, a meaningful sentence using the available vocabulary and grammar could be made within the application domain. Some of the sentences seemed extremely unusual (III-2.2.1, VII-1.1, VII-1.2, VII-1.3.1, VII-1.3.2, VII-3.1, VII-6.1, VII6.2), but nevertheless they are all grammatically correct and follow the patterns suggested by the Benchmark Evaluation Tool. The following is a subset of the test sentences and their scores. The complete set of test sentences and scores as well as their output are available upon request from the author or Rome Laboratory.

I Basic Sentences

1. Basic Sentence Types
    - 1.1 Declarative Sentences
        - The Boston to Atlanta flight left Denver. S
    - 1.2 Imperative Sentences
        - List the flights to Denver. P
    - 1.3 Interrogative Sentences
        - Is the Boston to Atlanta flight leaving Denver? F
2. Simple Determiners
    - 2.1 The Indefinite Article
        - The Washington to Denver flight landed in Dallas. F
    - 2.2 The Definite Article
        - What is the fare for the Boston to Denver flight? S
3. Simple Noun Phrase
    - 3.1 Count Nouns
        - Show the Boston flight. P

3.2 Proper Nouns (Not Applicable)*

3.3 Mass Nouns

List all information about Boston to Atlanta flight. S

4. Simple Verb Phrases

4.1 Copular Verb Phrases

Which flight landed in Atlanta? F

4.2 Verb Phrases Involving the Auxiliary Verb DO

Did the Atlanta to Boston flight land? F

4.3 Transitivity

Does the Boston to Atlanta flight serve dinner? S

4.4 Voice (Active & Passive) (Not Applicable)*

II Interrogative Sentences

1. What-questions

1.1 What as a pronoun

What does the Denver flight serve? P

1.2 What as a determiner

What destination does the Atlanta to Denver flight have? F

2. Who-questions (Not Applicable)*

3. Where-questions

Where does the Boston to Denver flight originate? N

4. When-questions

When did the Boston to Atlanta flight arrive? N

5. Which-questions

Which flight left for Atlanta? F

6. How-questions

How short is the Boston to Atlanta flight? N

7. Wh-word in Prepositional Phrase

From which airport did the Boston to Denver flight originate? F

8. Yes/No-questions

Are Denver to Atlanta flights expensive? F

III Noun Phrases

1. Prepositional Phrase as Postmodifier in a Noun Phrase

Does the Atlanta to Denver flight have a stop in Boston? S

---

* A meaningful sentence could not be constructed with the available vocabulary that would satisfy the suggested grammatical pattern recommended by the Benchmark Evaluation Tool.

2. The Noun Head

2.1 Nouns (Covered in Section I)

2.2 Nominals

2.2.1 Adjective Nominal

What is the cheapest to Denver? F

2.2.2 Passive Participles as Nominal (Not Applicable)

2.2.3 Progressive Participle as Nominal (Not Applicable)

3. Determinatives in More Detail

3.1 Predeterminers (Covered in Section IV)

3.2 Central Determiners

3.2.1 Articles (Covered in Section I)

3.2.2 Relative Determiners

The Denver to Atlanta flight which is cheap arrived in Boston. S

3.2.3 Wh-Determiners

3.2.3.1 Indefinite Reference ("Whatever")

("Whatever" not in the lexicon)

3.2.3.2 Definite Reference ("Whichever")

List whichever flight is leaving Denver. P

3.3 Postdeterminers (Covered in Section VI)

3.4 Determinative Combination (Covered in Section VI)

4. Premodification

4.1 Central Premodifiers

4.1.1 Simple Adjectives

What are the expensive flights? P

4.1.2 Superlative Adjectives

The most expensive flight is to Boston. P

4.2 Precentral Premodifiers

What is the very best Boston to Atlanta flight? F

4.3 Postcentral Premodifiers

4.3.1 Passive Participle as Premodifier (Not Applicable)

4.3.2 Progressive Participle as Premodifier (Not Applicable)

4.4 Noun-Noun Phrases or Nominal Compounds

4.4.1 At-Time Nominal Compound Type

Does the Boston to Atlanta flight have a night departure? F

4.4.2 At-Location Nominal Compound Type

The Boston to Atlanta flight originates from Logan airport. F

4.4.3 Purpose-Concerns Nominal Compound Type (Not Applicable)

4.4.4 BELONGS-To Nominal Compound Type (Not Applicable)

4.4.5 PART-OF Nominal Compound Type

Did the Atlanta flight leave Denver for Boston? S

4.4.6 PRODUCES Nominal Compound Type (Not Applicable)

4.4.7 EXECUTED-BY Nominal Compound Type (Not Applicable)

4.4.8 MADE-OF Nominal Compound Type (Not Applicable)

4.4.9 Produced-BY Nominal Compound Type (Not Applicable)

4.4.10 Purpose-BENEFITS Nominal Compound Type (Not Applicable)

4.4.11 Purpose-IS Nominal Compound Type (Not Applicable)

4.4.12 HAS-TYPE Nominal Compound Type (Not Applicable)

4.4.13 USES Nominal Compound Type (Not Applicable)

4.4.14 HAS-PART Nominal Compound Type (Not Applicable)

4.5 Premodifier Combinations

List the cheap short Boston to Denver flights. S

5. Genitives and Alternative forms

5.1 Genitives

List the Atlanta to Boston flight's arrival time. N

5.2 OF as an alternative to the Genitive

List the departure time of the Boston to Atlanta flight. S

5.3 WITH as an alternative to the Genitive

List the Boston to Atlanta flight with the cheapest fare. S

5.4 Combinations of Genitives and Alternatives

5.4.1 Genitive and OF-Phrase

What are the stops of the Atlanta to Boston's flight? F

5.4.2 Combination of OF-Phrase and the WITH-Phrase

List the origin of the Boston to Atlanta flight with a stop in Denver. F

6. Postmodification

6.1 Relative Clauses

6.1.1 Relative Pronoun as Subject

List the flights that stop in Atlanta. P

6.1.2 Possessive Relation Pronoun ("WHOSE") as a Determiner

The flight whose destination is Denver left Boston. N

6.1.3 Relative Pronoun as Object

Is the Boston to Denver flight the flight that stopped in Atlanta? F

6.1.4 Deletion of Relative Pronoun Object (Not Applicable)

6.1.5 Relative Pronoun Prepositional Objects (Not Applicable)

6.1.6 Relative Pronoun as Adverbial

List the cities in which the Boston to Atlanta flight stops. N

6.1.7 Multiple Relative Clauses Within the Same Noun Phrase

The flight that stopped in Atlanta whose destination is Denver landed in Boston. N

6.2 Reduced Relation Clauses (Not Applicable)

IV ADVERBIALS

1. Adverbs and Adverbials

1.1 Adverbials in Initial Position

In Atlanta the flight from Boston landed. F

1.2 Adverbials in Pre-Verb Position

Which flights from Boston to Denver have a stop in Washington. S

1.3 Adverbials in Final Position

List the flight which arrived in Boston today. F

V VERBS AND VERB PHRASES

1. Temporal Aspect of Verb Phrases

1.1 Stative Verbs (Not Applicable)

1.2 Stance (Not Applicable)

1.3 Dynamic Verbs

Which flight left for Denver from Atlanta? S

1.4 Tense/Aspect

1.4.1 Simple Tense

List the flights which leave Boston in 5 minutes. P

1.4.2 Perfect

Which flights have landed in Denver after 10 a.m. F

1.4.3 Progressive

List the flights which were flying to Denver at 10 a.m. P

1.4.4 Perfect Progressive

Which flights have been landing in Denver after 12 p.m.? F

VI QUANTIFIERS

1. Functional Position of Quantifiers

1.1 Quantifiers in Determinative Positions.

When does the flight from Denver arrive? F

1.2 Quantifiers as Noun Phrase Head

List all of the flights leaving Denver for Atlanta. S

2. Semantics of Quantifiers

2.1 Universal Quantifiers

Both of the flights to Denver were canceled. N

2.2 Existential Assertive Quantifiers

Some of the flights to Denver were canceled. N

2.3 Existential Non-Assertive Quantifiers

What time did either of the flights to Denver arrive? P

2.4 Negative Quantifiers

("None" not in the lexicon)

("Neither" not in the lexicon)

2.5 Numerical Quantifiers

List the second flight which arrived in Boston from Denver. S

3. Indefinite Quantifier Pronouns

3.1 Universal

Tell me everything about the Boston to Atlanta flight. F

3.2 Existential Assertion

Tell me something that flights to Denver serve. N

3.3 Existential Non-assertive

Tell me anything that is served on the flights. N

3.4 Negative

Nothing was served on the flight from Boston to Denver. F

4. Existential

4.1 Declaratives

There are two flights from Boston to Denver. C

4.2 In Relative Clauses

List the Boston flights that there are. P

4.3 What Questions

What flights are there to Atlanta from Boston? S

4.4 Yes-No Questions

Is there a flight to Boston? P

4.5 Tag Questions

There is a flight from Boston to Atlanta, Isn't there? N

5. Universal Adjectives

("Entire" not in the lexicon)

VII COMPARATIVES

1. Comparative Adjectives

1.1 Comparison to a Higher Degree

List the flights which are more expensive than the Boston to Atlanta flight is expensive.** N

1.2 Comparison to a Lower Degree

The flights which are less expensive than the Boston to Atlanta flight is expensive are shorter.** N

1.3 Comparisons to the Same Degree

1.3.1 In Assertive Contexts

List the flights which are as cheap as the Boston to Atlanta flight is cheap.** F

1.3.2 In Non-Assertive Contexts

What airlines are not as cheap as the American airlines is?** N

1.4 Comparison to a Constant

List the flights which are more expensive than $200. F

1.5 Comparative Phrases Used as Pre-Modifiers within a Noun Phrase

List the flights which are longer than 3 hours. F

2. Superlatives

2.1 Superlatives to a High Degree

What is the cheapest flight from Denver to Atlanta which stops in Boston? S

2.2 Superlatives to a Low Degree

What is the least expensive fare from Boston to Atlanta? S

3. Comparative Adverbial Phrase

3.1 Comparison to a High Degree

Which flight arrived earlier than the Boston to Atlanta flight arrived? N

3.2 Comparison to a Lower Degree

Which airline discounts fares less often than the American airlines discounts first class fares? N

3.3 Comparison to the Same Degree

3.3.1 In Assertive Context

Which airlines discount fares as often as the American airlines discounts first class fares? N

3.3.2 In Non-Assertive Contexts

Which airlines discount fares not as often as the American airlines discounts first class fares? N

3.4 Comparisons to a Constant

** These constructs may seem unusual but they are grammatically correct and follow the suggested patterns recommended by the Benchmark Evaluation Tool.

Which flight stops more often than three times on each trip? N

4. Superlative Adverbs

4.1 Superlative Adverbs to a High Degree

Which flight stops in Boston most often? F

4.2 Superlative Adverbs to a Low Degree

Which flight stops in Boston least often? F

5. THAN as a Connective Between Mixed Clauses (Not Applicable)

6. Comparative Noun Phrases in Object Position

6.1 Comparative Noun Phrases Using Comparative Quantifiers

List the flights which had more stops in Boston than the Denver to Washington flight had. N

6.2 Comparative Noun Phrases Using Comparative Adjectives

Which flight stops more than the Washington to Atlanta flight stops? F

7. Comparing Objects of VPs with Extra Adjuncts

List the airlines which discounted more fares last week than the American airlines discounted first class fares this week. N

8. Implicit Comparatives

List the flights which are better than the Boston to Atlanta flight. F

9. Multipliers and Fractions in Comparisons

9.1 Multipliers

Which airline has twice as many planes as the American airlines? N

9.2 Fractions

Which flight takes one half as long as the Denver to Atlanta flight? N

VIII CONNECTIONS

1. Coordinators

1.1 AND

1.1.1 Sentential Connectives

List the flights to Atlanta and list the flights to Denver. N

1.1.2 Relative Clauses Conjoined by AND

What are the flights which leave Denver and which stop in Washington? N

1.1.3 Noun Phrase Constructions with the Conjunction AND

List the shortest and fastest flights from Denver to Boston. N

1.1.4 The Conjunction AND in a Premodification of a Noun Phrase

What are the cheapest and the fastest flights from Boston to Denver? F

1.1.5 Verb Phrases with the Conjunction AND

1.1.5.1 BE-Verb Conjoined by AND

List the flights from Denver to Boston which are cheap and are fast. F

1.1.5.2 DO-Verb Constructions with the Conjunction AND (Not Applicable)

1.1.5.3 Full Verb Constructions with the Conjunction AND

The Boston to Atlanta flight arrived and landed. S

1.1.5.4 Mixed Verb Phrases in Constructions with AND

List the flights which are cheap and stop in Atlanta. P

1.1.6 Conjunction of Adverbials

Which flights arrived late to Atlanta yesterday and today? F

1.1.7 Conjunction within Prepositional Phrases

List the flights to Boston and to Atlanta. N

IX EMBEDDED SENTENCES

1. THAT-Clauses

Is it correct that American airlines discounted the first class fares? N

2. WH-Interrogative Clauses

Tell me what is the cheapest flight from Denver to Boston? S

3. Yes-No Interrogative clauses

Tell me if the Boston flight has arrived. P

4. Infinitive Clauses Introduced by To (Not Applicable)

5. -ING Clauses (Not Applicable)

X REFERENCE

1. Specific Reference

1.1 Anaphoric Reference

1.1.1 Pronominal Anaphora

Did the Dallas to Boston flight arrive?

When did it leave Dallas? N

1.1.2 Nominal Anaphora

List the flights to Boston.

Do these flights serve dinner? P

1.1.3 Anaphora with SO and AS

The Boston flight left Denver and so did the Atlanta flight. N

1.1.4 Intra- and Inter-Sentential Anaphora

Which flight stops in Dallas?

Is it an economy flight? P

1.2 Cataphoric References

Before it landed in Boston, did the Atlanta to Boston flight have a stop in Dallas? P

1.3 General Knowledge or the Larger Situation

What is a flight? F

XI ELLIPSIS

1. Elliptical Noun Phrases

1.1 Ellipsis of head only, no Postmodifiers Present

What is the cheapest flight?

What is the fastest? F

1.2 Ellipsis of Premodifier(s) and head, no Postmodifiers Present

What is the cheapest flight to Washington?

What is the fastest? F

1.3 Ellipsis of Post-modifier(s) only

List the arriving flights to Boston.

List the departing. N

1.4 Ellipsis of head and Post-modifier(s)

What is the fastest flight to Boston from Denver?

What is the cheapest? F

1.5 Ellipsis of Premodifier(s) and head and Postmodifier(s)

What is the cheapest Boston flight from Denver?

What is the fastest? F

1.6 Ellipsis of head with Postmodifier(s) retained.

What is the cheapest Boston flight from Denver?

What is the fastest from Denver? F

2. Elliptical Clauses (Not Applicable)

XII SEMANTICS OF EVENTS

1. Main Concept of Events

1.1 Verbs of Motion

Did the Atlanta flight leave Boston? S

1.2 Verbs of Mental Phenomenon (Not Applicable)

1.3 Verbs of Communication

Tell me which flight left Boston for Atlanta? S

1.4 Verbs of Happenings (Not Applicable)

2. Agency

2.1 Agent

The American airlines discounted the first class fares. N

2.2 Co-Agent (Not Applicable)

3. Experiencer (Not Applicable)

4. Beneficiary (Not Applicable)

5. Patient (Not Applicable)

6. Theme

6.1 Theme in Object Position

List the flights which leave Boston for Atlanta. S

6.2 Theme in Subject Position

The flight to Denver was canceled in Atlanta. N

6.3 Co-Theme

Which airline discounts fares to Atlanta? N

7. Source

What was the origin of the flight that just landed in Boston? F

8. Goal

List the flights which are leaving for Boston. P

9. Instrument (Not Applicable)

10. Participant Combination (Not Applicable)

11. Spatial Information of Events

11.1 Position

What flight landed in Washington? F

11.2 Path (Not Applicable)

11.3 Distance

How many miles is the flight form Denver to Washington? F

12. Temporal Information of Events

12.1 Position

List the flights which left for Boston from Denver last week. N

12.2 Duration

List the flights which have landed since last night. N

12.3 Frequency

How often does a flight leave Boston for Denver? N

12.4 Relationship

List the flights which landed in Boston before 5 a.m. F

13. Participant Combination with Time and Location

In Boston on March 26 the flight from Denver landed. F

14. Manner (Not Applicable)

15. Means (Not Applicable)

16. Respect (Not Applicable)

17. Contingency (Not Applicable)

18. Degree (Not Applicable)

19. Modality

List the flights which possibly leave Boston for Denver before 5 a.m. S

20. Evidentials (Not Applicable)

In order to achieve a numeric average score for each section, all scores were transformed, according to the following, and averaged for each individual section.

	Points
Success (S)	10
Correct (C)	8
Partially Correct (P)	5
Failure (F)	2
No Output (N)	0

The comprehensive results of all twelve sections are shown in Figure 1. Full circles represent the average success percentages of all applicable sentences and the empty circles represent the average success percentages of all applicable as well as non-applicable sentences combined. The NLP system performed well on basic sentences, verbs and verb phrases, and semantics of events; however, comparatives and ellipsis sections were among its worst performances. The NLP system is apparently not designed for complicated database comparisons (VII-1.1, VII-3.1), and incomplete sentences (XI-1.1, XI-1.2, XI-1.3).

The overall success rate of the NLP system is 32.4% for applicable sentences and 28.3% for applicable and non-applicable sentences combined. The low score is contributed to the fact that the application domain, which accompanied the NLP system, is extremely narrow. In other words, the application domain has a very limited lexicon and knowledge of grammatical patterns. Therefore, the input sentences must be confined within a restricted boundary. This limitation is expected from all NLP systems which operate within a well-defined domain. The evaluation tool, on the other hand, is designed to test the linguistic flexibility of the NLP systems. For instance, in order to form a complete application call, the NLP system requires the source and destination of each flight to be explicitly stated in each sentence. As expected, the test sentences which satisfied those requirements performed well. Consider the following two sentences: a) Which flight left for Atlanta (II-5), and b) Which flight left for Denver from Atlanta (V-1.3). Although the NLP system can analyze sentence (a) and can produce an IDR, it cannot make a complete application call and hence fails to respond. On the other hand, it has no problem responding to sentence (b). Whenever possible, we have included a source and a destination for each flight. In most of the failure cases (i.e. F scores), the NLP system was able to create an

IDR but the application module, which is responsible for making the application calls, was not able to interpret it due to a lack of information.

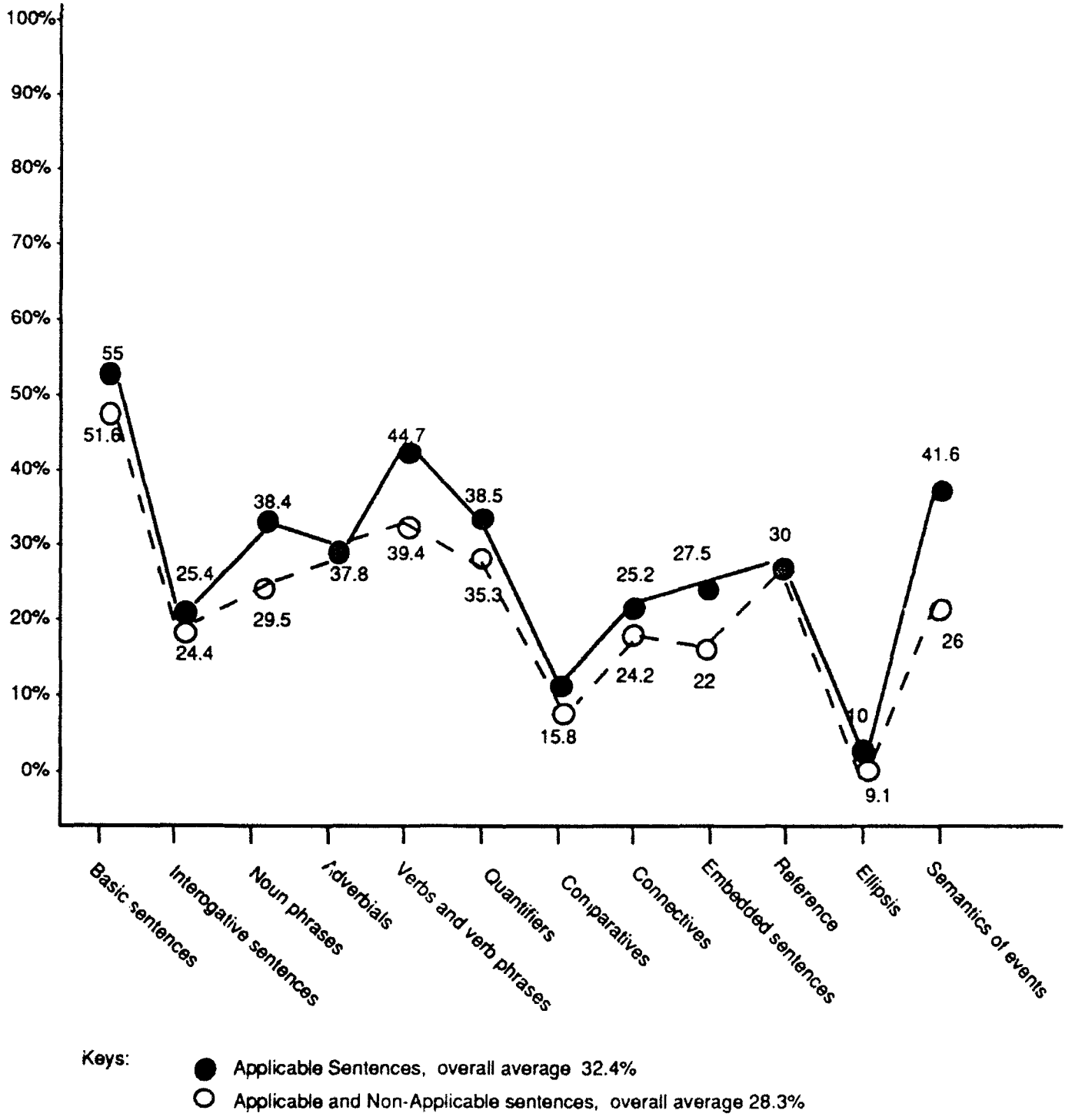


Figure 1: Average Success Percentages

5 Conclusions

Although the idea of testing the sensitivity of individual linguistic capabilities of NLP systems rather than the sensitivity of the systems to individual applications is extremely attractive, it has nevertheless proved to be an ambitious task. Most NLP systems are designed for well-defined domains and applications. Therefore, a general purpose evaluation tool may not be suitable for all types of NLP systems. This was evident from our investigation. The NLP system, in our investigation, has an extremely narrow application domain which responds fairly well to sentences that satisfy its requirements, however the sentences that do not, fail to be analyzed completely. Each type of NLP system possesses certain attributes that are unique. Each type has strengths and weaknesses which are directly associated to the goals and objectives of the system. Therefore, the evaluation procedure should be more sensitive to the type of the NLP system being evaluated. For instance, if the NLP system is a *Data Base Management System*, the evaluation tool must place more emphasis on the interrogative and basic sentences rather than on quantifiers and ellipsis. Since after all, the system is not designed to respond to ellipsis or quantifiers.

Some of the grammar patterns suggested by the Benchmark Evaluation Tool are not used in everyday conversation, however they are perfectly correct. For instance, *'List the flights which are more expensive than the Boston to Atlanta flight is expensive'* (VII-1.1) is grammatically correct, but the second part of the sentence *'flight is expensive'* is normally omitted and is implied by the first part. This may cause some confusion among some of the evaluators. In addition, scoring may also pose some confusion, since the the boundaries between suggested scores are not well-defined and are subjective. Hence, two independent evaluators may score a single NLP system completely different.

Although not all suggested sentence patterns were applicable, nevertheless they helped with defining the boundaries of the NLP system. In many instances such as VII-3.1, VII-3.2, VII-9.1, and VII-9.2 the clash between the wide scope of the evaluation tool and the narrow application domain of the NLP system was clearly evident. In defense of the NLP system, it must be noted that no NLP system can successfully satisfy all the rigorous requirements of the Benchmark Evaluation Tool. The Benchmark Evaluation Tool proved to be extremely helpful in providing guidance and structure for evaluating the NLP system, therefore it should be used as a guide to select the appropriate testing procedures for individual types of systems, rather than as a general purpose evaluation procedure that can be applied to all NLP systems.

Strengths of the Benchmark Evaluation Tool:

- Comprehensive
- Contains detailed explanations
- Independent of NLP systems and their application domain
- Defines the boundaries of NLP systems

Weaknesses of the Benchmark Evaluation Tool:

- Time consuming
- Scope of the evaluation is too wide
- Some suggested patterns seem unusual and are not used in everyday conversation
- Scoring is not well defined

In conclusion, the evaluation process proved to be extremely time consuming. It is conceivable that in the near future the evaluation process could be fully automated. However, in order for an automated evaluator to be successful, the evaluation should be performed in a narrower space with well-defined boundaries. Therefore, there should be several different automated evaluators each specialized for a different type of NLP system. Each automated evaluator would have syntactic, semantic, and pragmatic knowledge of only one type of NLP system and would generate appropriate test sentences. The set of automated evaluators would form a complete collection of tools for evaluating all types of NLP systems. The Benchmark Evaluation Tool will be extremely instrumental in developing the automated evaluators.

6. References

[1]. Benchmark Investigation/Identification Program Volume I; Final Report, Calspan Advanced Technology Center, P.O. Box 400, Buffalo, NY 14225, May 1992.

MEASUREMENT OF THERMOPHYSICAL PROPERTIES OF SEMICONDUCTORS AT HIGH TEMPERATURE

Joseph B. Milstein
Associate Professor
Paul J. Chandonnet
Graduate Student
Department of Electrical Engineering

University of Massachusetts Lowell
One University Avenue
Lowell, MA 01854

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Balling Air Force Base, Washington, D.C.

September 1992

MEASUREMENT OF THERMOPHYSICAL PROPERTIES OF SEMICONDUCTORS AT HIGH TEMPERATURE

Joseph B. Milstein
Associate Professor
Paul J. Chandonnet
Graduate Student
Department of Electrical Engineering
University of Massachusetts Lowell

Abstract

InP is potentially a superior material for use in high-speed devices, microwave devices, optoelectronic devices, and radiation hard, highly efficient solar cells. The production of high quality single crystal substrates is required to attain the technological promise that such materials hold out. The thermophysical properties of these materials in the molten state have an appreciable bearing on the crystal growth process. These properties include the density, thermal conductivity, thermal diffusivity, kinematic viscosity, and emissivity of the liquid. These properties have not been measured for InP, because one requires a pressure of 27 atmospheres of phosphorous above the melt at a melting temperature of 1062 °C (1335 K) in order to maintain stoichiometry.

An Arthur D. Little Model HPCZ High Pressure Furnace, capable of operation to 1500 psi, and an associated 50 KW, 450 KHz Lepel Radio Frequency Generator were used in this work. A cylindrical thermal cell was designed and constructed which permitted the controlled melting of InP under a conventional B_2O_3 encapsulant which was contained in a flat bottomed quartz crucible. Temperature mesurements were made using type K thermocouples, and an Omega Engineering OM-900 computer interfacing module which included a model 992 CPU module. Custom circuitry was designed and constructed to provide electrical power to a secondary molybdenum filament heater under computer control, and with provision for monitoring and logging the voltage and current in a secondary heater. Software to control the experimental procedure and to record data was also written.

The experimental setup was used for measuring properties of copper (a standard material), boric oxide, and indium phosphide.

MEASUREMENT OF THERMOPHYSICAL PROPERTIES OF SEMICONDUCTORS AT HIGH TEMPERATURE

Joseph B. Milstein
Paul J. Chandonnet

INTRODUCTION

Semiconductors, especially III-V materials such as InP and GaAs, are especially attractive materials because they offer the possibility of producing a wide range of improved electronic devices as compared to the work horse semiconductor material, silicon. In particular, InP is potentially a superior material for use in high-speed devices, microwave devices, optoelectronic devices, and radiation hard, highly efficient solar cells.

The production of high quality single crystal substrates is required to attain the technological promise that such materials hold out. The crystal growth technology of these materials has advanced to a considerable extent in recent years [1], but still lags significantly behind that of silicon. There are several reasons why this is true.

First, silicon crystal growth spans about forty years of experience in commercial production. Second, the compound semiconductors present problems relating to such matters as control of stoichiometry. These problems manifest themselves in the necessity to use encapsulants about the melt and/or ambients at high pressures. Specifically for InP, one requires a pressure of 27 atmospheres of phosphorous above the melt at a melting temperature of 1062 °C (1335 K) in order to maintain stoichiometry. Third, the thermophysical properties of these materials in the molten state have an appreciable bearing on the crystal growth process [2]. These properties include the density, thermal conductivity, thermal diffusivity, kinematic viscosity, and emissivity of the liquid. Some of these properties have been measured for only a few compound semiconductors [3], but not for InP. Estimates of the values of some of these properties are presented by Jordan [4].

Accordingly, it was decided that an attempt to determine some of these values by actual measurement would be a useful contribution to our understanding of the crystal growth issues that require solutions if these materials are to become widely used.

METHODOLOGY

An Arthur D. Little Model HPCZ High Pressure Furnace, capable of operation to 1500 psi, and an associated 50 KW Lepel Radio Frequency Generator which operated at 450 KHz were available for use on this project. A cylindrical thermal cell was designed and constructed which permitted the controlled melting of modest amounts (approximately 150 grams) of InP under a conventional B_2O_3 encapsulant which was contained in a flat bottomed quartz crucible. Type K thermocouples were inserted into the thermal cell from the top and bottom along the cell axis. Both thermocouples were monitored using an Omega Engineering OM-900 computer interfacing module, which included a model 992 CPU module, a model 911 General Purpose I/O module, and a model 931 Thermocouple module. Datacould be recorded at a rate of one temperature reading every 50 milliseconds.

Custom circuitry was designed and constructed to provide electrical power to a secondary molybdenum filament heater under computer control, and with provision for monitoring an logging the voltage and current in the secondary heater. This circuit was calibrated both with a series of known resistor loads and with direct meter measurements of current and voltage. Software to control the experimental procedure and to record data was also written. An incandescent light bulb was installed inside the chamber to provide illumination during experimental setup, so as to insure that all of the components were properly assembled prior to the start of a run.

High pressure purified argon gas was provided from cylinders, and the internal pressure of the chamber was monitored by an electronic pressure gauge capable of one psi resolution. The ambient in the system was controlled by performing multiple

fill cycles each of which consisted of pumping down to a residual pressure of less than 100 millitorr with a mechanical pump, followed by backfilling with argon.

The experimental setup was used for measuring properties of copper (a standard material), boric oxide, and indium phosphide.

A graphite susceptor of 4 inch height by 2.5 inch OD and 2.1 inch ID was situated in a 10 turn rf coil. An alumina refractory heat shield and a quartz tube separated the susceptor from the rf coil. The susceptor had a 2.1 inch internal diameter by 0.4 inch recess at its base, which was filled with a fiberfrax felt thermal insulation disc.

A type K thermocouple was axially introduced from the bottom of the furnace and passes through a machineable boron nitride support plate, the fiberfrax disc, and a hole in the bottom of the susceptor, which was a 0.180 inch thick graphite segment.

For measurements on the copper standard and pure boric oxide, a 1 inch tall by 2.080 inch diameter graphite disc was placed in the graphite susceptor simply to place the charge at a known, constant level with regard to the secondary heater.

Boric oxide was contained in a 0.625 inch tall by 1.67 inch diameter quartz crucible which sat on the graphite disc.

To measure indium phosphide, a quartz crucible of 1.625 inch height by 1.67 inch outside diameter was used. By filling this larger crucible with one inch of InP material and then adding the boric oxide encapsulant, the top of the boric oxide stood in the same location the furnace as that of the smaller crucible placed on the graphite disc.

Measurements were made using the bottom thermocouple, and a top thermocouple which was carried by the seed rod and could be raised and lowered along the central axis of the measurement cell.

Two pyrolytic boron nitride crucibles were modified for use as covers to define the volume of the thermal cell and to isolate the specimen under test and its crucible from the rf heater. Since pyrolytic boron nitride has a large thermal

conductivity along the crucible wall and a much smaller conductivity perpendicular to the crucible wall, the crucible tended to make the imposed thermal field due to the primary rf heat source more uniform. One crucible was cut to a height of one inch, and the second was 2 inches tall. Both crucibles had a diameter of approximately two inches.

Secondary heater coils of 0.012 inch diameter molybdenum wire were constructed by wrapping approximately 7 feet of the wire around a mandrel of 0.33 inch diameter. A coil was mounted to the inside of the boron nitride crucible using molybdenum wire as a means of attachment. These coils had a nominal resistance at room temperature of approximately 2.4 ohms, and were used as the secondary heat source for the thermal measurements. It was observed that a coil could generally survive approximately ten experiments before it became too embrittled as a consequence of being heated. Coils could handle approximately 200 watts of power. A readily demountable connection was provided by using a commercial polarized plug and socket set, with the plug attached to the coil and the socket attached to power leads entering a Conax high pressure feedthrough in the chamber wall. The top of the thermal cell was insulated with one inch of fiberfrax board and was covered with a graphite disk which could behave as a susceptor at the cell top.

RESULTS

The uniformity of the temperature in the system was checked by performing profiling measurements, in which the temperature was measured at specific axial positions under constant power input. The system was observed to have a small negative gradient, i.e., slightly cooler as the thermocouple was raised.

The system was first calibrated by using a copper disk as a sample. The disk measured 1.65 inches in diameter by 0.525 inches thick, and was held in a quartz crucible. The bottom thermocouple touched the lower surface of the crucible and the top thermocouple touched the top surface of the copper. Three gas fill cycles were performed, and the system was heated to a

temperature of approximately 750 °C and allowed to equilibrate under constant rf power. After equilibrium was established, a thermal pulse was applied to the top surface of the copper by energizing the secondary molybdenum heater at measured power for a measured duration.

This thermal pulse can be approximated as a step function. The thermal response of the copper was measured at both the top and bottom surfaces. The change in temperature at the top surface could be modelled quite precisely using a function of temperature to the fourth power, which is consistent with the transfer of energy to the block from the secondary heater by optical radiation according to the Stefan-Boltzmann relation. The response of the bottom thermocouple could be modelled by a commonly used one dimensional approximation to the heat flow through an object subjected to a sudden change of temperature at one boundary [5]. While this sudden change is only somewhat accurate, requiring some tens of seconds for the new surface equilibrium temperature to be established, the model appears to fit the data with reasonable precision.

The model describes the system temperature as a function of both position along the unique axis (x), and time (t), i.e., $T(x,t)$. We take $x = 0$ as the surface upon which energy flux impinges. If the material properties do not vary significantly with temperature, the applicable differential equation is

$$\partial^2 T/\partial x^2 = (1/\alpha)\ \partial T/\partial \tau$$

with boundary and initial conditions

$$T(x,0) = T_{init}$$

$$T(0,\tau) = T_{applied}, \text{ and}$$

$$q_o/A = -\kappa\ \partial T/\partial x \quad \text{at } x = 0 \text{ for } \tau > 0.$$

The solution of this differential equation is then

$$T - T_{init} = \{2q_o(\alpha\tau/\Pi)^{0.5}/kA\}\exp(-x^2/4\alpha\tau) - q_o x/kA\{1-\text{erf}(x/2(\alpha\tau)^{0.5})\} \qquad \text{(Equation A)}$$

where q_o is applied power, A is area, k is thermal conductivity, α is thermal diffusivity, and $x > 0$ is distance into the specimen under test. Equation A may be used to obtain a fit to

experimental data quite conveniently using a spreadsheet.

If one models a multilayer system as one in which q_o is constant, but the other parameters are specific to each material as x increases through the composite specimen, one can derive values for the parameters k and α by adding one new material at a time. Since $\alpha = k/\rho c$ where ρ is the density and c is the specific heat, one can estimate the specific heat if one has an estimate of the density. Furthermore, by recording the time evolution of both the temperature on the surface where energy is supplied and the one to which energy is transported through the system, one gets a measure of time constants of the furnace configuration used in the test.

Beginning with copper and quartz, for which all of these parameters are very well known, one can calibrate the value of the thermal flux q_o/A for the system at a given emissivity of the specimen. One then can measure the boric oxide - quartz system and derive values for the relevant thermophysical parameters for boric oxide. Finally one can study the InP - boric oxide - quartz system and deduce values for InP itself.

We have carried out such measurements, with some success. Unfortunately, molten InP is quite corrosive and we were obliged to employ a thermocouple protection tube made of quartz having a wall thickness of 0.7 millimeters to try to protect the top thermocouple from being destroyed. In each experiment the thin walled protection tube failed, which permitted molten InP to attack the thermocouple. For InP we have only obtained consistent data from the bottom thermocouple. We can use this data to draw some limited conclusions.

Calibration runs using the thermal cell with only the quartz crucible as a specimen were performed at pressures of 41 and 496 psi. It was possible to introduce gradients of the order of 5 to 10 degrees across the quartz. It was evident that the temperature measured by the top thermocouple displays an increasing noise component with pressure, which may be attributed to increased convective flow as the pressure is increased. The

noise was essentially zero at 41 psi and increased to about a 2 to 3 degree fluctuation at 496 psi. These results are shown in Fig 1a and 1b.

Calibration measurements using the copper standard were conducted at a baseline temperature of approximately 750 °C at 30 psi, and at approximately 660 °C at 300 psi. Fig 2 shows the temperature excursion for the top thermocouple, as well as a calculated fit to the data for a temperature of approximately 680 °C at 36 psi based on the assumption that energy transfer occurs principally by radiative methods. The fitted curve is in excellent agreement with the experimental data.

Fig 3 shows the thermal history at the lower thermocouple for a power application of approximately 100 watts. Handbook values for the thermal conductivities of copper and fused quartz were used in the calculations based on Equation A above. One concludes that the amount of power entering the front surface of the copper block represents approximately 41% (or that the emissivity was 0.41) for the copper specimen. This is quite reasonable as the emissivity for polished copper is tabulated as 0.03 and that for heavily oxidized copper is 0.75 to 0.8. Our specimen was slightly oxidized, which seems to be in reasonable agreement with the emissivity calculated.

Measurements taken on pure boric oxide were conducted at base temperatures of approximately 960 °C at 45 psi and at approximately 1000 °C at 496 psi. Figs 4 and 5 show results from these experiments. The thermal conductivity (1.63 watts/meter-Kelvin) [6], the viscosity [7], and the specific heat (1.83 J/gram-K) of boric oxide are reported in the literature, measured at 1 atmosphere pressure and 1000 °C. One can use these values to compare to those we obtain at low pressure. Our values at high pressure are the ones that really are important for the growth of InP, and have apparently not been measured before. The data were again analyzed using Equation A.

For both the copper and the boric oxide measurements, a series of thermal pulses were applied to the system and the

responses recorded. Fig 6 summarizes the response of the system to a series of such thermal pulses of various magnitudes. For both materials, the actual system response is essentially a linear function of the magnitude of the applied thermal excitation. Such behavior is a significant confirmation that the system is behaving as a one dimensional thermal system. If the system were two or three-dimensional, it would be extremely fortuitous that the response at the location chosen would be precisely linear for the several thermal pulses introduced.

Fig 7 shows the behavior of the lower thermocouple for measurements made on the InP - boric oxide - quartz system, operated at approximately 1070 °C and 580 psi. Six curves are shown, for thermal pulses of approximately 20 to 200 watts. The response of the lower thermocouple is again very nearly linearly related to the magnitude of the thermal pulse. This system was also examined using Equation A, but we have no measured data for the thermal excursion at the top thermocouple. The precise value of thermal energy entering the system is therefore open to some uncertainty, although we know how much power was supplied by the secondary heater in each case.

CONCLUSIONS

Based on our analysis, we obtain the following results.

The measured thermal conductivity of boric oxide appears to be about 1.5 watts/meter-Kelvin.

Further experimentation using the existing system with a heavier wall protection tube is expected to be successful, which will lead to a measured value for the thermal conductivity of InP.

REFERENCES

1. D.F. Bliss, R.M. Hilton, S. Bachowski and J.A. Adamski, J. Electronic Materials 20, 967 (1991).

2. A.S. Jordan, Journal of Crystal Growth 49, 631 (1980).

3. V. M. Glazov, S.N. Chizhevskaya, and S.B. Evgen'ev, Russian Journal of Physical Chemistry 43, 201 (1969).

4. A.S. Jordan, Journal of Crystal Growth 71, 551 (1985).
5. J.P. Holman, "Heat Transfer", (McGraw-Hill Book Co., New York; 1986), Chapter 3.
6. G.K. Creffield and A.J. Wickens, J. Chem Eng. Data 20, 223 (1975).
7. R.A. Eppler, J. Amer. Ceram. Soc. 49, 679 (1966).

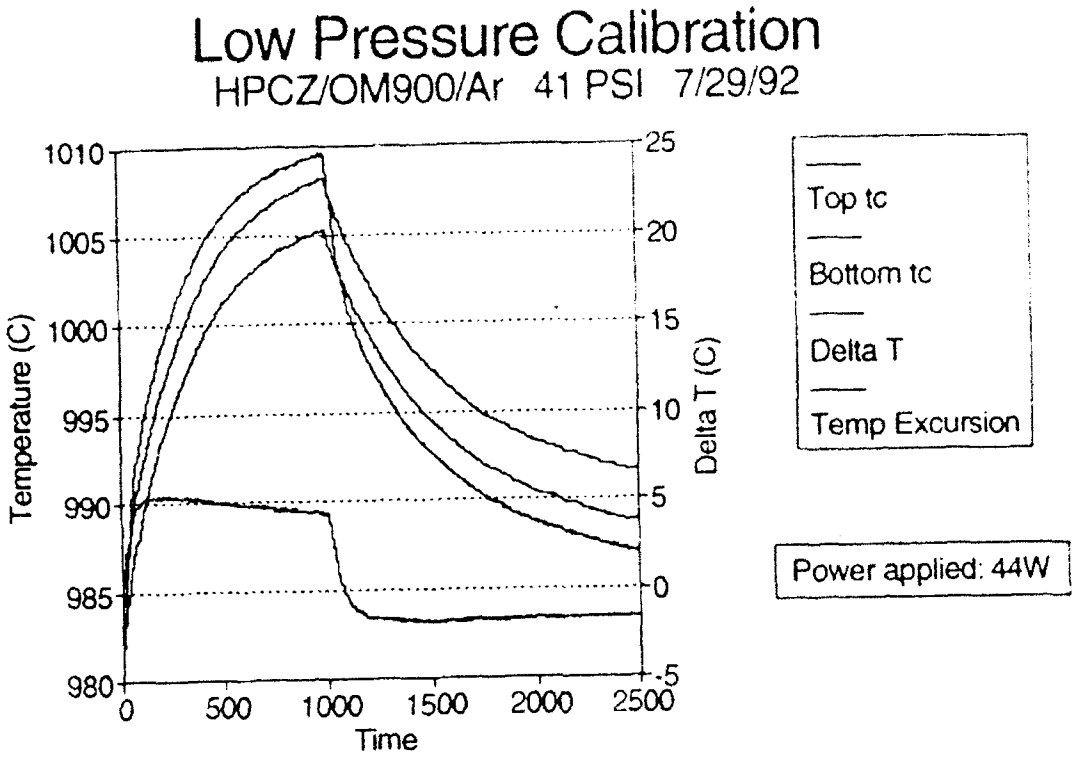


Figure 1a

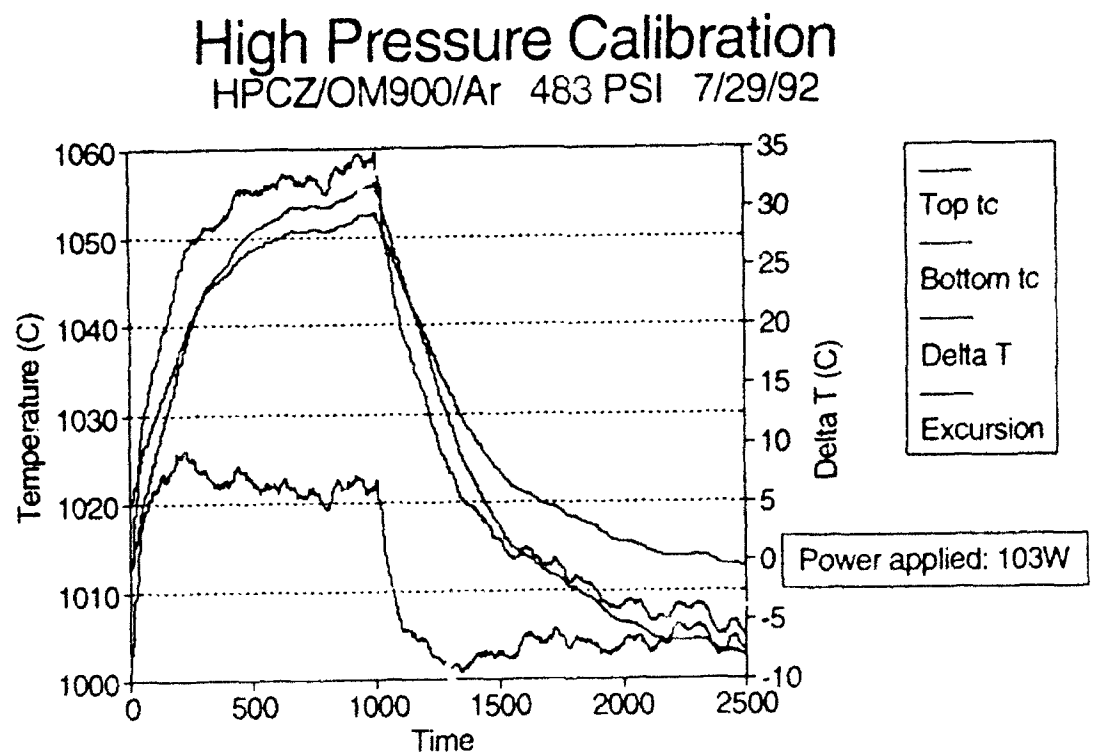


Figure 1b

Heating a Copper Block

HPCZ/OM900 7/13/92

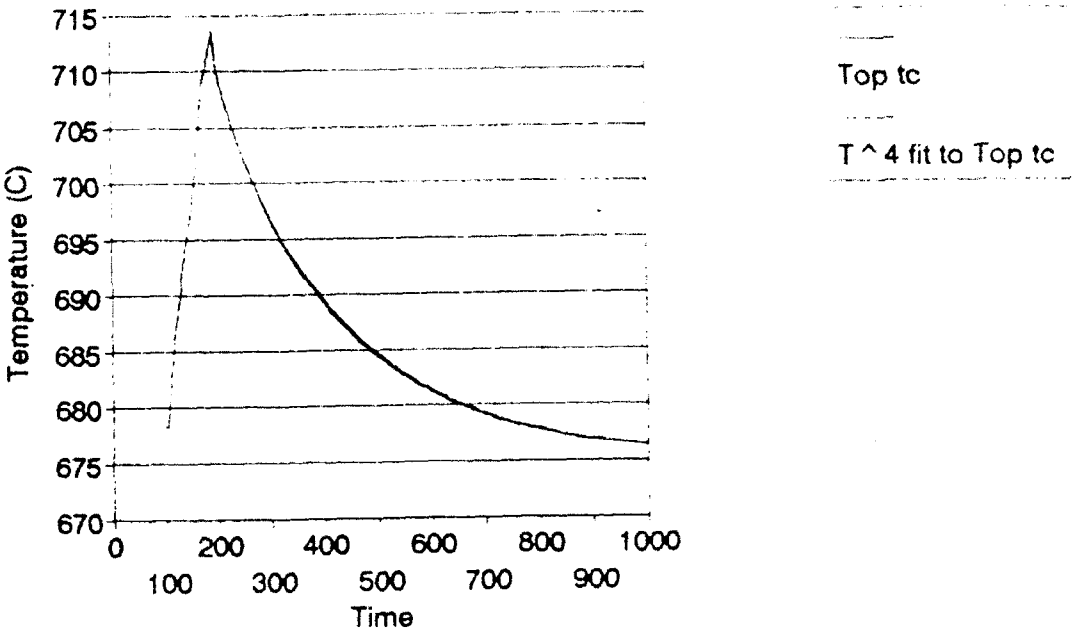


Figure 2

Heating a Copper Block

HPCZ/OM900 7/13/92

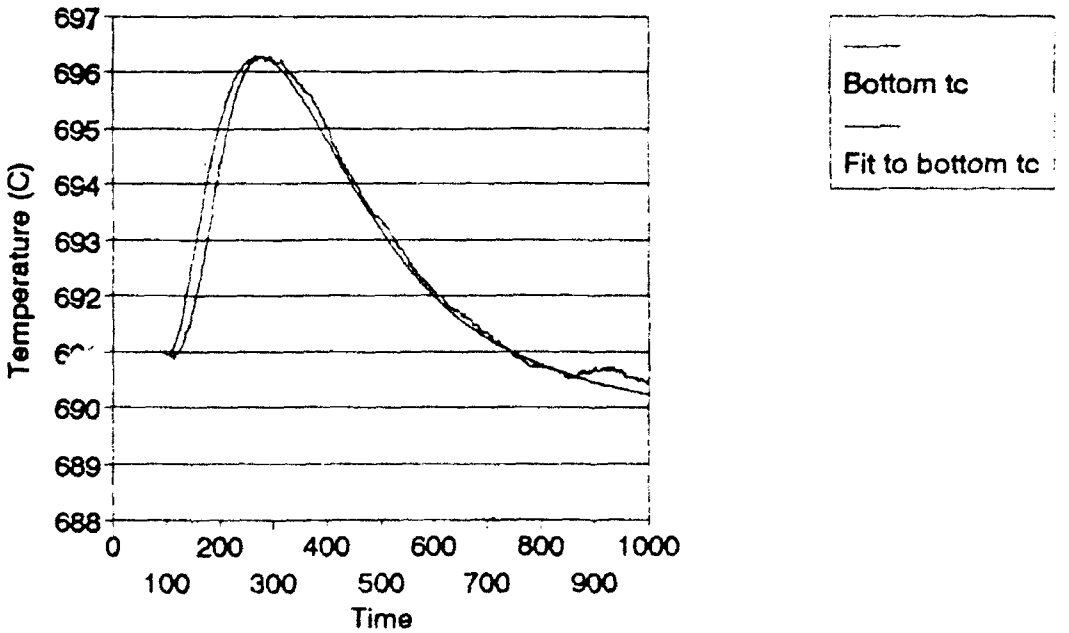


Figure 3

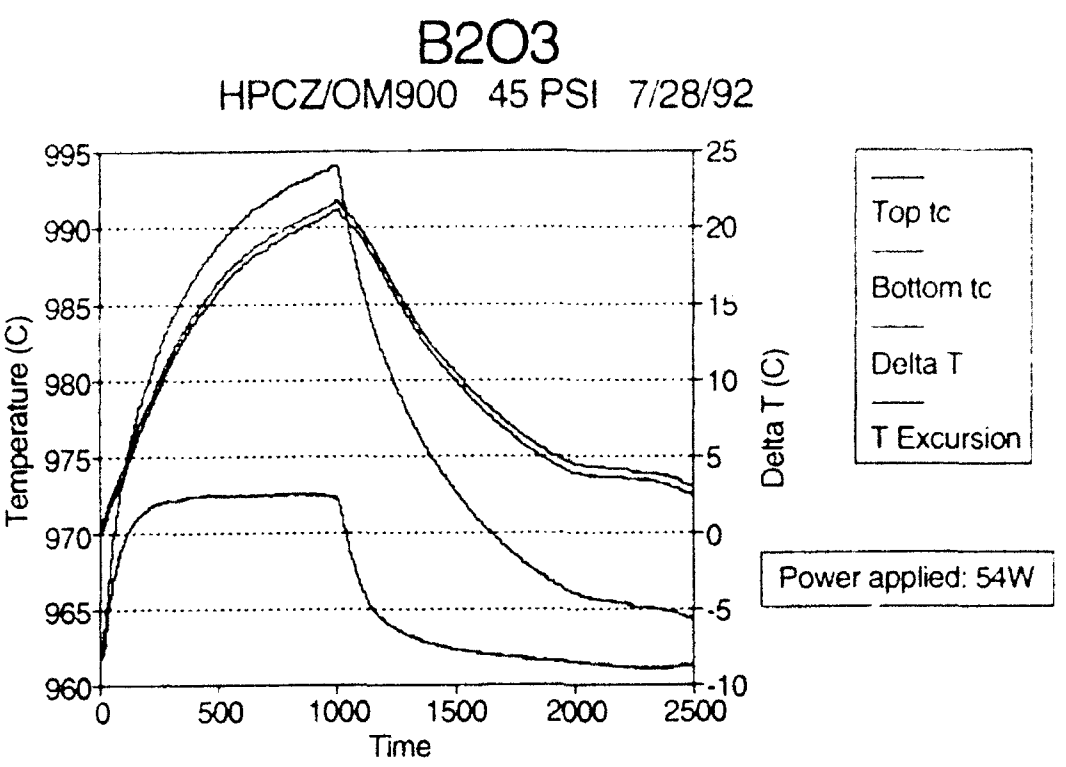


Figure 4

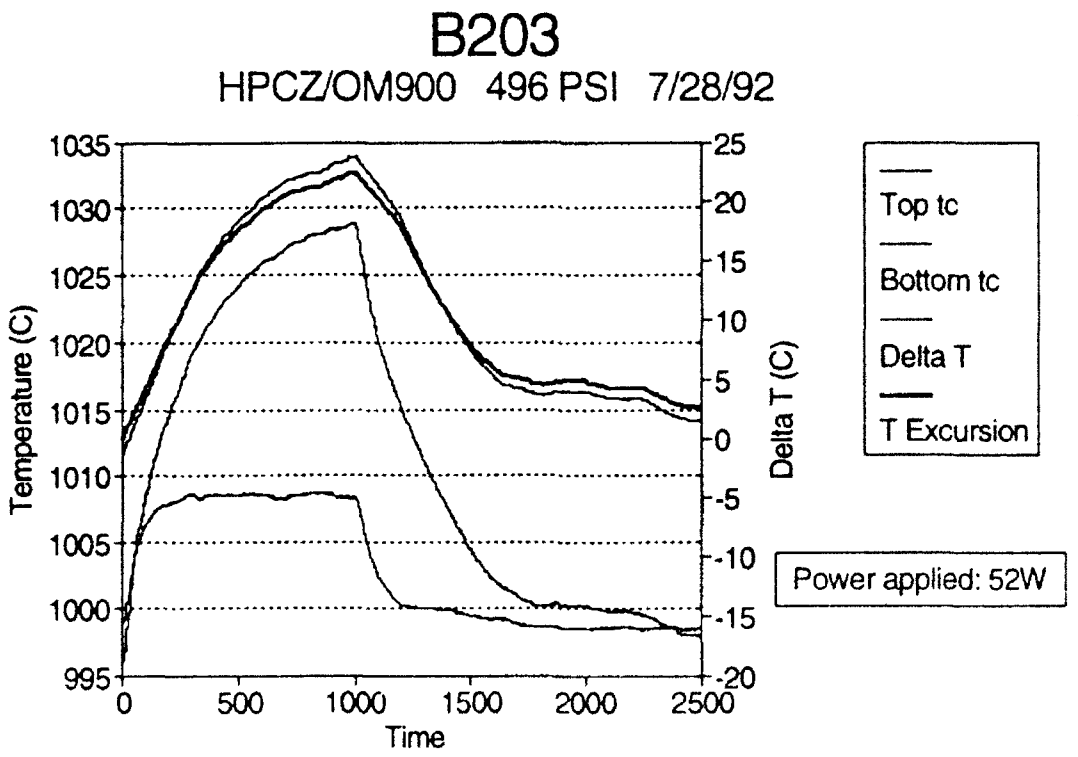


Figure 5

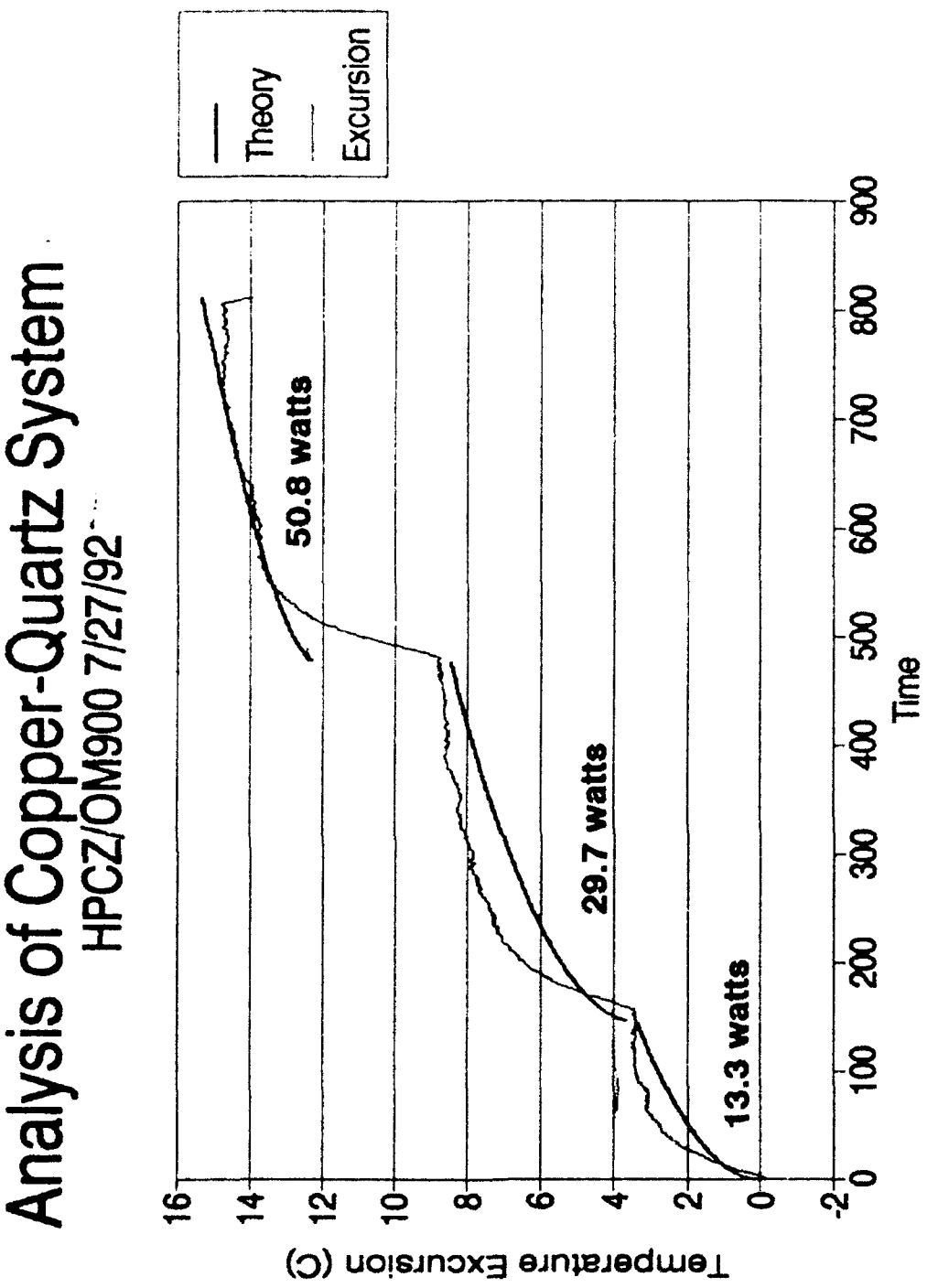

Analysis of Copper-Quartz System
HPCZ/OM900 7/27/92
Theory
Excursion
50.8 watts
29.7 watts
13.3 watts
Temperature Excursion (C)
Time
16
14
12
10
8
6
4
2
0
-2
0
100
200
300
400
500
600
700
800
900

Figure 6

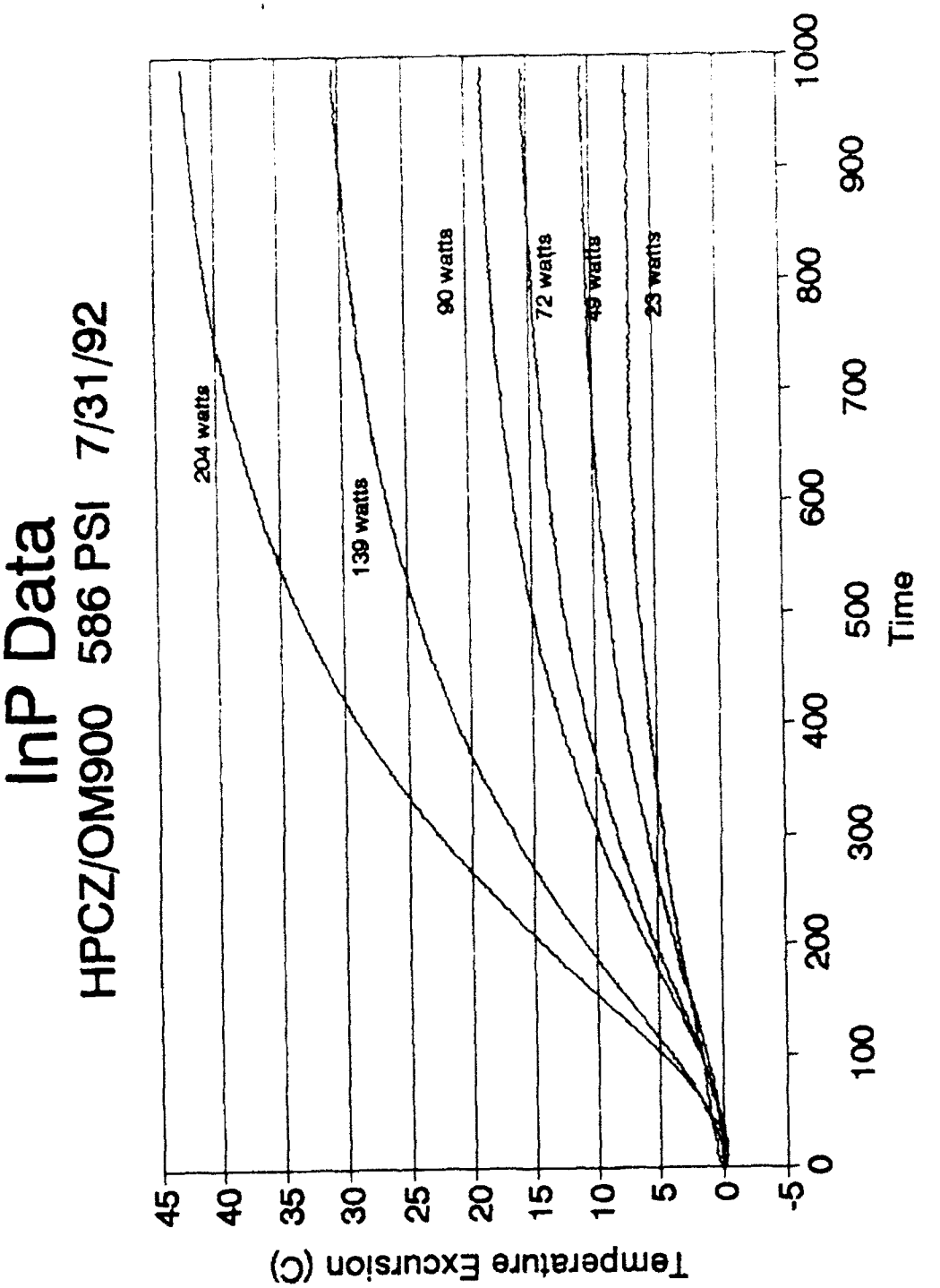

InP Data
HPCZ/OM900 586 PSI 7/31/92
204 watts
139 watts
90 watts
72 watts
49 watts
23 watts
Temperature Excursion (C)
Time
45
40
35
30
25
20
15
10
5
0
-5
0
100
200
300
400
500
600
700
800
900
1000

Figure 7

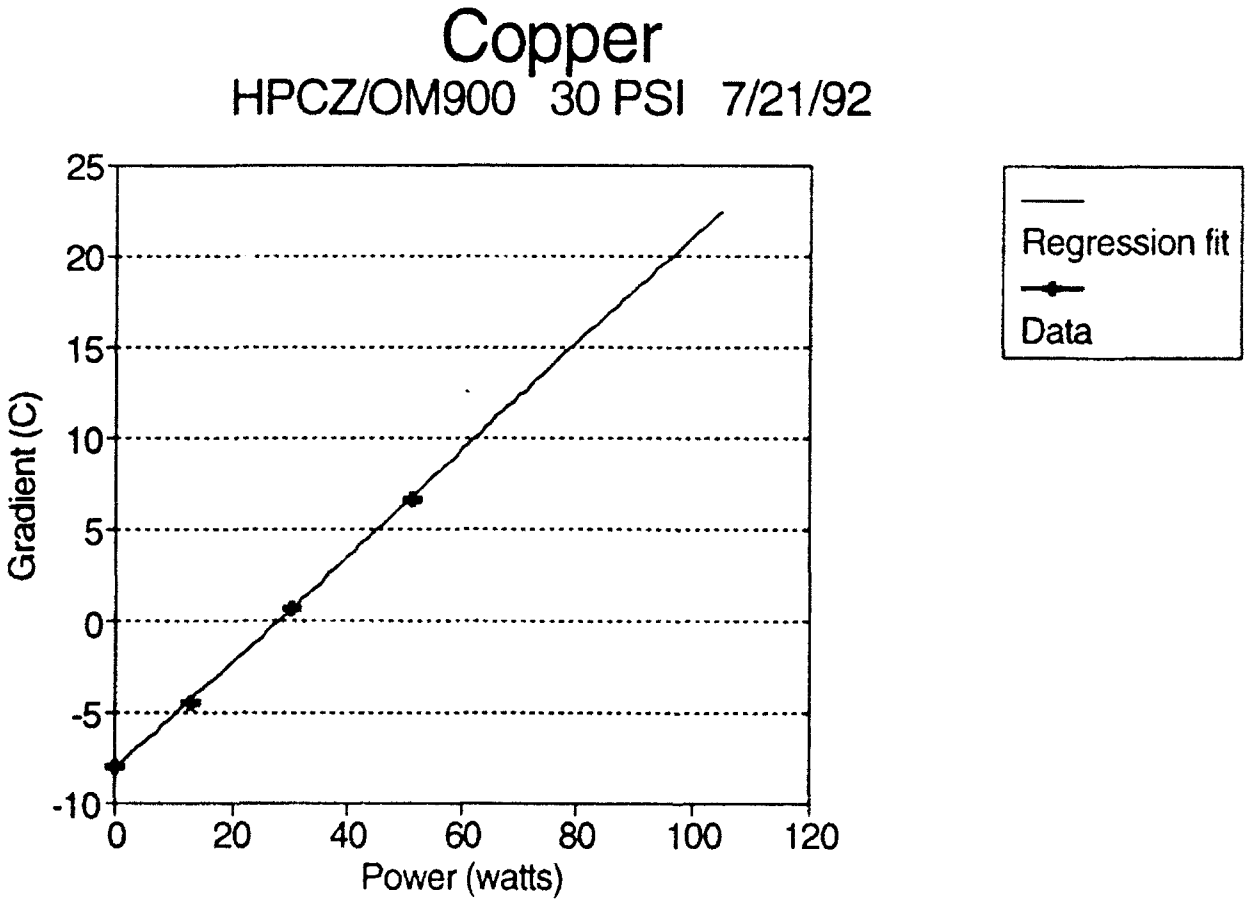


Figure 8. Observed Gradient vs. applied power

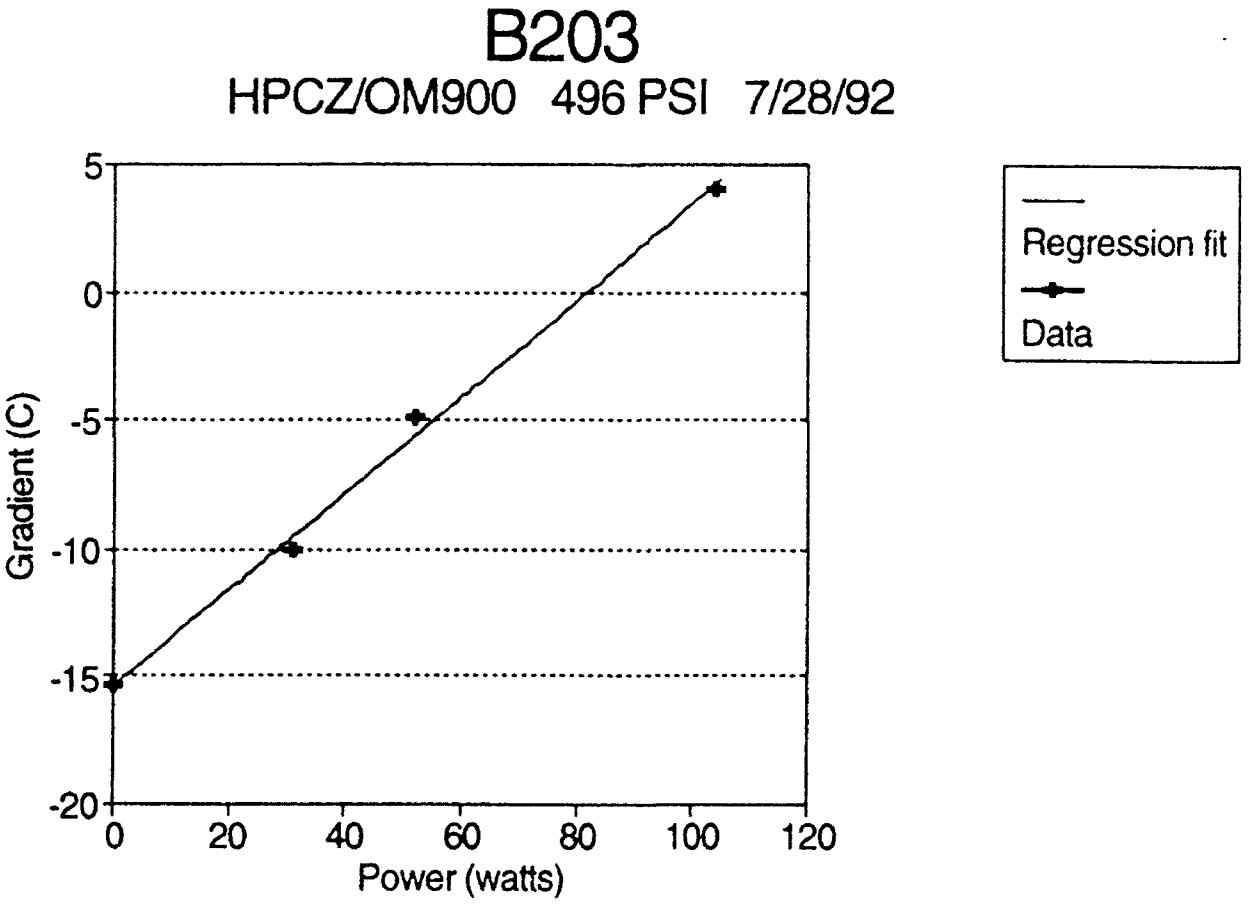


Figure 9. Observed gradient vs. applied power

PHOTONIC DELAY LINE FOR HIGH-FREQUENCY RADAR SYSTEMS

Evelyn H. Monsay
Associate Professor
Department of Physics

Le Moyne College
1419 Salt Springs Road
Syracuse, New York 13214

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Griffiss Air Force Base, Rome, New York

August 1992

PHOTONIC DELAY LINE FOR HIGH-FREQUENCY RADAR SYSTEMS

Evelyn H. Monsay
Associate Professor
Department of Physics
Le Moyne College

Abstract

Good pointing accuracy and sidelobe-level noise control in wideband phased array radar antennas depend on the use of true time delays between the microwave elements. A photonic true time delay line was introduced by Toughlian and Zmuda for systems operating in the hundreds of megahertz frequency range (VHF). In the experimental part of this study, their concept is extended to operation in the radar C band through the use of a dichromatic solid state laser. In addition, a system analysis is initiated which relates system-level concepts such as antenna directivity and noise suppression to device-level concepts such as laser jitter and drift.

PHOTONIC DELAY LINE FOR HIGH-FREQUENCY RADAR SYSTEMS

Evelyn H. Monsay

I. INTRODUCTION

Superior performance of wideband phased-array radar antennas requires true time delay phase shifts between elements. However, building a radar system with conventional true time delay phase shifters is costly and the result is too heavy for many airborne or satellite applications. Previous photonic methods of generating true time delay lines have relied on varying lengths of optical fiber for each element [1]. However, such methods are cumbersome and difficult to implement due in part to large optical losses.

More recently, Toughlian and Zmuda [2-5] developed an RF true time delay phase shifter based on a heterodyne optical system. As indicated schematically in Figure 1, a rotating and translating mirror is used to locate the Bragg-offset beam of the heterodyne system at a specific point in the Bragg cell. At that particular point, a particular phase offset will be imposed on the optical beam relative to the starting point of the acoustic signal in the Bragg cell. Hence, for different points selected along the acoustic wave in the Bragg cell by the motion of the mirror, a different phase will be obtained. Likewise, as the RF signal frequency input to the Bragg cell changes, the optical phase of the heterodyne system will change, resulting in a linear dependence in the RF C-Band of phase on frequency - with a different slope at each mirror-directed location - as required for a true time delay system. For multiple array elements, an array of rotating, translating mirrors can be used to select the appropriate phases.

In Toughlian's and Zmuda's original work, the frequency offset of the Bragg cell determined the frequency band in which a phased array antenna built around the true time delay line would operate. In particular, the initial experimental system operated between 55 - 80 MHz (VHF regime). The authors suggested [4] that high-frequency wideband phased arrays could be produced with true time delay capability if two nominally identical phase-locked lasers operating at slightly different frequencies could be used to establish a baseline frequency shift to a higher band. A compact realization of this function was found in a single, dichromatic solid state laser produced by Amoco laser company (Model ALC 1064-20DS). Using a prototype of this laser, the current study verified operation of a true time delay optical phase shifter for phased array antennas in the RF C-band.

In Section II, the basic operation of the photonic true time delay phase shifter is discussed in more detail. In Section III, some results of an

analysis of radar phased array systems are presented. In particular, the phenomenon of beam "squint", induced when a constant phase shift between elements of an array is used instead of true time delay phase shifts, is considered and illustrated through examples. Some relationships between system-level beamforming specifications and device-level parameters are discussed. In Section IV, an experimental characterization of the Amoco dichromatic laser is presented. A description of the optical breadboard used in the feasibility demonstration, along with results of the experiment, are presented in Section V. Finally, a discussion of these results and recommendations for future experimental and theoretical work are presented in Section VI.

II. PHOTONIC DELAY LINE THEORY

As indicated in the Introduction, Figure 1 presents the architecture of the basic photonic true time delay delay line. The optical configuration is a standard heterodyne system with laser frequencies of ω_S and ω_R and nominally equal amplitudes A, and Bragg cell RF frequency offset ω_M, resulting in optical beams

$$E_{SIGNAL}=A\exp\left[j\left(\left(\omega_S+\omega_M\right)t+\phi\right)\right] \tag{1}$$

$$E_{REFERENCE}=A\exp\left(j\omega_R t\right) \tag{2}$$

where ϕ is an optical phase shift, covering both desired and undesired, ie, noise-induced, possibilities. The heterodyned signal, or coherent addition of the two optical beams, is given by the optical intensity, ie, the absolute value squared of the sum of the two complex beam amplitudes. This operation results in a constant dc background illumination plus the system output signal with frequency dependence

$$I(t)\sim|A|^2\cos\left[\left(\omega_S-\omega_R+\omega_M\right)t+\phi\right] \tag{3}$$

In Equation (3), the self-heterodyne frequency of the laser output, $\Delta\omega=(\omega_S-\omega_R)$, appears, modulated by the RF frequency ω_M. Hence, the system operates in a band defined by lower frequency $\Delta\omega$ and upper frequency $(\Delta\omega+\omega_{M(max)})$. Note that any optical phase shift imposed on either beam, here indicated on the signal beam, comes through directly in the final detected optical signal.

Figure 2 indicates the selective action of the system optics within the

Bragg cell, where a specific phase shift can be "picked off" from the acoustic wave in the cell. For a specific frequency ω_M, the rotating/translating mirror will establish a location, and so a phase value, along the acoustic wave relative to its launch point in the cell. As, ω_M changes, the phase of the acoustic wave at that fixed point in the Bragg cell will also change, such that a linear dependence of optical phase on RF frequency - the hallmark of the true time delay phase shift - is established. Hence, for a wideband RF signal with upper and lower frequencies ω_U and ω_L, respectively, a true time delay phase shift for each frequency component of the RF signal will be available with varying magnitude dependent on which particular point within the Bragg cell is chosen by one (or more, for multiple antenna elements) rotating mirror(s).

III. PHASED ARRAY PERFORMANCE CONSIDERATIONS

A phased array antenna directs its transmit (and/or receive) beam via element-to-element phase shifts which match those of a plane wave front tilted at the desired angle to the plane of the array. (See Figure 3). The array will be designed with a certain center frequency in mind (f_0), and elements will be spaced at distances d_x and d_y in the plane of the array, in the x- and y-directions, respectively.. Then, the array factor $A(\theta,\phi)$ is given by

$$A(\theta,\varphi)=\sum_{m}^{N_x}\sum_{n}^{N_y}\frac{I_{mn}}{I_{00}}\exp\left(j\left[md_x(ku-k_0u_0)+nd_y(kv-k_0v_0)\right]\right) \qquad (4)$$

where uniform element patterns have been assumed [6]. The relative amplitudes of the element excitations, I_{mn}/I_{00}, also will be assumed uniform and equal to one. The observer is located at position (R_0,θ_0,ϕ_0), and the array is scanned to angular position (θ,ϕ). Wavevectors are defined by $k = 2\pi/\lambda$, where λ is the wavelength associated with the signal frequency f, and $k_0 = 2\pi/\lambda_0$, where λ_0 is the wavelength associated with the design frequency f_0. Variables u (u_0) and v (v_0) are defined by $u = \sin\theta\cos\phi$ and $v = \sin\theta\sin\phi$, with the subscript "0" inserted appropriately as needed. Figure 4 shows the beampattern obtained for $\phi = \phi_0 = 0$, and for scan angle $\theta = 30°$, when $f = f_0 = 5$ GHz, for a 36-element square array.

The phenomenon known as beam squint occurs when a scanned phased array is driven at a frequency other than the design frequency with no alteration in the phase shift given to the various elements, ie, for constant phase shift. The effect of beam squint is seen as a shift in the beampattern maximum, creating a beam pointing error, when a frequency of f = 6 GHz is the signal

frequency (Figure 5a), or when f = 4 GHz (Figure 5b). Note also a rise in the sidelobe level for the frequency shifted array which would result in a noisier radar system. However, as shown in Figures 5c and 5d, a true time delay phase offset given to the various array elements - and realized by replacing k_0 with k in Equation (1) - brings the beam maximum back into alignment. Obviously, for transmission or reception of a wideband signal, true time delay phase shifts are required to ensure that the various frequency components of the signal all refer to the same target.

An important goal of this ongoing study is to relate the system-level specifications of importance to the radar user - such as directivity, pointing accuracy and sidelobe noise suppression - to the device-level characteristics of the photonic delay line. Since the photonic delay line is based on a heterodyne optical system, much laser noise suppression will be built in. Any laser instability common to the two frequency-offset beams of a dichromatic laser will be suppressed by the coherent interference of the two beams. Also, the acousto-optic modulator (Bragg cell) can be assumed completely stable. However, trouble can arise in several areas:

1) Frequency jitter in the laser, ie, short-term frequency and mode hopping, will alter directly the value of the true time delay phase shift in an unpredictable manner, introducing pointing errors, etc. into the radar antenna system.
2) Mode hopping in the laser will also create an effective pointing error of the laser beam in the Bragg cell, causing the "pick-off" point along the acoustic wave to move, again resulting in a phase error in the radar antenna system.
3) Any lack of coherence between the two frequencies of the dichromatic laser will result directly in an error in the value of the phase shift for the radar antenna elements.

Quantification of these effects is underway.

IV. DICHROMATIC LASER CHARACTERIZATION

The Amoco prototype dichromatic laser (Model ALC 1064-20DS) is a Nd:YAG solid state laser operating at 1064 microns, but with two mutually orthogonal phase-coupled beams separated in frequency by approximately 4.27 GHz [7]. In order to characterize the laser, the output was directed onto a polarizer, where the two orthogonal polarizations could mix, then fed into a multimode fiber and detected. Figure 6 shows the self-heterodyne signal generated at the nominal interference frequency. Figures 7a and 7b give an indication of the stability of the laser, presenting measurements of short-term frequency

jitter and long-term frequency drift, respectively. The short-term jitter, or width at half-maximum of the heterodyne peak averaged over 16 samples, is about 500kHz. The long-term drift was obtained over a period of 1.5 hours, with the spectrum analyzer set to hold the maximum value in each frequency bin over that time span. The result was a drift of approximately 2 MHz. The self-heterodyne frequency and the relative strength of the two laser frequencies were adjustable, as the two laser frequencies could be moved relative to the gain curve of the laser. It was clear from rotation of the polarizer that one frequency beam was horizontal, ie, parallel to the tabletop, the other was vertical, and that the measured laser output power of approximately 37 milliwatts was about equally split between the beams.

V. EXPERIMENTAL SYSTEM AND RESULTS

A single-element photonic true time delay phase shifter was set up in the lab. Output from the dichromatic laser was split by a polarizing beamsplitter so that the appropriately polarized beam would be directed to the Bragg cell ("signal" beam). The other beam ("reference" beam) was shunted to two mirrors which directed it to a recombining polarizing beamsplitter.

Following the beamsplitter, the signal beam was directed through a cylindrical lens so as to focus the light in the vertical, leaving it unaltered in the horizontal direction. In this configuration, the beam will be optimally utilized by the Bragg cell, all optical power being contained within the height of the acoustic waveform within the cell. Since a band of RF frequencies, from f_L to f_U, will be imposed on the Bragg cell, the outgoing first-order diffracted beam will be spread by an angle

$$\Delta\alpha = \lambda_0 \Delta f / v_s \qquad (5)$$

where $\Delta f = f_U - f_L$, and v_s is the speed of sound in the Bragg cell material, in addition to any initial divergence of the beam from the laser.

The Bragg cell used was a custom-designed indium phosphide (InP) acousto-optic deflector by Brimrose, centered at wavelength 1.3 microns, and with a maximum diffraction efficiency of 15 percent at that wavelength. Since the optical wavelength in use in this experiment was 1.06 microns, it was expected that the diffraction efficiency would be somewhat less; a value of 14.3 percent was measured.

After the signal beam exited the Bragg cell, the zeroth-order (undiffracted) beam was stopped, and the first-order diffracted beam, now

modulated by ω_M, was directed into a spherical lens, chosen so as to stop the angular spread of the beam at that point. The signal beam then fell on a single rotatable mirror, on a Photonic Control piezoelectrically-controlled mirror mount capable of 630 microradian movement. An electronic driver controlled the motion of this mirror, as needed, for mapping out locations along the acoustic wave in the Bragg cell.

The signal and reference beams were recombined with a second polarizing beamsplitter and the output light directed into a multimode fiber for subsequent detection by an Antel Gigahertz Fiber Optic Test System. Since optical power was not very abundant, no polarizer was used to mix the orthogonal signal and reference beams. Instead, the polarization mixing property of the several-foot length of multimode fiber was relied on for the heterodyning of the two signals. Typically, 1.5 to 1.95 milliwatts of optical power were coupled out of the multimode fiber into the photodetector.

Due to a slight mismatch between the center wavelength (1.1 microns) of the polarizing beamsplitters and the nominal laser wavelength, some optical power from each polarization was coupled into each beam (signal and reference). Although the Bragg cell effectively stopped almost all the S-polarization light from the signal beam, the reference was found to carry a good deal of the P-polarization beam, resulting in a sizable self-heterodyne signal out of the Antel photodiode. Figure 8 is the output of the photodiode, showing the large self-heterodyne signal at 4.15 GHz (more typically, this signal was at 4.27 GHz), and the system heterodyne signal at 5.45 GHz, reflecting the 1.3 GHz RF signal driving the Bragg cell. Of course, the self-heterodyne signal is phase coherent with the system heterodyne signal, a fact which was used to great advantage in completing the experimental measurement of phase versus frequency for the system.

In previous measurements of the phase versus frequency characteristic of experimental photonic true time delay lines, a network analyzer was used, with the stimulus being the driving RF signal to the Bragg cell [2]. This method worked well, since the network analyzer works from the known initial phase of the stimulus signal to determine the phase of the response received back - ie, in previous experiments, the output of the photodiode at the Bragg offset frequency. Hence, the stimulus signal was always the same frequency as the response signal in previous characterizations. However, such is not the case for the present system, in which the frequency at which the Bragg cell is driven (the stimulus from the network analyzer) is not the same as the system heterodyne response, which is offset by the dichromatic laser's frequency difference between beams. For a stimulus signal of 1.3 GHz, the network analyzer would have to add a phase coherent frequency offset of about 4.27 GHz

in order to characterize the response at 5.57 GHz, which is illustrated by the system transfer function measured for a stimulus frequency range of 1.29 GHz to 1.31 GHz (Figure 9). Currently available network analyzers do not have such a capability. In order to characterize the phase versus frequency response of the frequency offset photonic true time delay line, the leakage-induced self-heterodyne signal in the photodiode output was used. The self-heterodyne signal, being phase coherent with the system heterodyne signal, could be RF mixed with the system heterodyne signal, downshifting the system heterodyne signal to the same frequency as the stimulus driving the Bragg cell, as indicated in Figure 10. Then, the network analyzer could be used to characterize the system's phase as a function of frequency.

Implementing the frequency offset described above and in Figure 10 required a couple of stages of power amplification, due to the low optical power levels actually coupled out of the Bragg cell and, ultimately, out of the optical fiber. Old, tube-type amplifiers were used, adding some noise to the system, mostly in the forms of frequency spikes and drift.

The results of the characterization of phase versus frequency for the photonic delay line are given in Figure 11 for two mirror positions, taken at extremes of the mirror rotation (differing by about 2°). Two linear phase versus frequency functions are mapped out, with very different slopes for the two different mirror positions - and, hence, pick-off points along the acoustic waveform in the Bragg cell - verifying that true time delay phase shifts can indeed be extracted from the photonic delay line, in this case, for C-band wideband signals.

VI. CONCLUSIONS AND RECOMMENDATIONS

This experiment proves that a photonic system can generate true time delay phase shifts for phased array antennas, providing the capability to build lightweight, potentially lower cost, high-frequency (C-band), wideband radar systems. Some potential sources of error in the phase shifts have been identified; quantification of these effects is currently underway.

In addition to the need to quantify phase errors and cost factors for these photonic delay lines, the potential for creating a frequency-agile phased array antenna system must also be explored. The current system can operate only within a single radar band, fixed by the frequency difference of the two beams out of the dichromatic laser. However, recently developed lasers [8] have a variable frequency difference between orthogonally-polarized beams, controllable in real time, which can provide for shifting between radar bands, while still providing the wideband capability of a true time delay

beamformer. Construction and testing of such a variable-band system will be undertaken in the near future.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Henry Zmuda, Capt. Edward Toughlian, Dr. David Sumberg, and Lt. Michael Caccuitto for their insights and assistance during the course of this work. She would also like to thank James Cusack, William Kaveney, and the AFOSR Summer Faculty Research Program for the excellent opportunities they have provided.

REFERENCES

[1] Soref, R.A., "Application of Integrated Optics and Fiber Optics to Phased-Array Antennas", RADC-TR-84-176, In-House Report, August 1984.

[2] Toughlian, E.N. and Zmuda, H., "A Photonic Variable RF Delay Line for Phased Array Antennas", J. Lightwave Technol., vol. 8, no. 12, pp.1824-1828, 1990.

[3] Toughlian, E.N., Zmuda, H., and Kornreich, P., "A Deformable Mirror-Based Optical Beamforming System for Phased Array Antennas", IEEE Photon. Technol. Lett., vol. 2, no. 6, pp. 444-446, 1990.

[4] Zmuda, H. and Toughlian, E.N., "Variable Photonic Delay Line for Phased Array Antennas and RF/Microwave Signal Processing", RL-TR-91-120, Final Technical Report, June 1991.

[5] Zmuda, H. and Toughlian, E.N., "Adaptive Microwave Signal Processing: A Photonic Solution, Microwave Journal, vol. 35, no. 2, 1992.

[6] von Aulock, W.H., "Properties of Phased Arrays", Proc. IRE, vol. 48, pp. 1715-1727, October 1960.

[7] Leilabady, P.A. and Sipes, D.L., "All Optical Self-Heterodyned Remote Antenna Schemes With Greatly Improved Performance Characteristics", PSAA-91, The Second Annual DARPA/Rome Laboratory Symposium on Photonics Systems for Antenna Applications, December 1991.

[8] Schulz, P.A. and Henion, S.R., "Frequency-modulated Nd:YAG laser", Opt. Lett., vol.16, no. 8, pp. 578-580, 1991.

FIGURES

Figure 1: Photonic true time delay phase shifter - optical layout.

Figure 2: Delay as a function of position of the signal beam along the acoustic waveform in the Bragg cell.

Figure 3: Incidence or projection of a plane wave relative to an array of elements.

Figure 4: Beampattern produced at design frequency $f=f_0=5$ GHz, steered to 30°.

Figure 5: Beampatterns produced by an array of 36 elements designed for frequency $f_0=5$ GHz and steered to 30° with, (a) f=6 GHz, constant phase shift; (b) f=4 GHz, constant phase shift; (c) f=6 GHz, true time delay phase shift; (d) f=4 GHz, true time delay phase shift.

Figure 6: Self-heterodyne signal from Amoco dichromatic laser at $\Delta f=4.28$ GHz.

Figure 7: Characterizations of prototype Amoco dichromatic laser: (a) short-term jitter of about 500 kHz from 16-sample average; (b) long-term drift of about 2 MHz from 1.5 hour sample, using maximum hold function of spectrum analyzer.

Figure 8: Output of Antel photodiode showing "leakage" self-heterodyne peak at 4.15 GHz and system heterodyne peak at 5.45 GHz, phase-locked signals which were then RF mixed.

Figure 9: Transfer function of optical system for stimulus (Bragg cell RF driver) range 1.29 GHz to 1.31 GHz. (Response shown is system heterodyne signal).

Figure 10: Schematic for RF mixing of photodetector output in order to derive nominal 1.3 GHz response signal for network analyzer.

Figure 11: Phase versus frequency of photonic true time delay line for two orientations of the rotating mirror differing by approximately 2°.

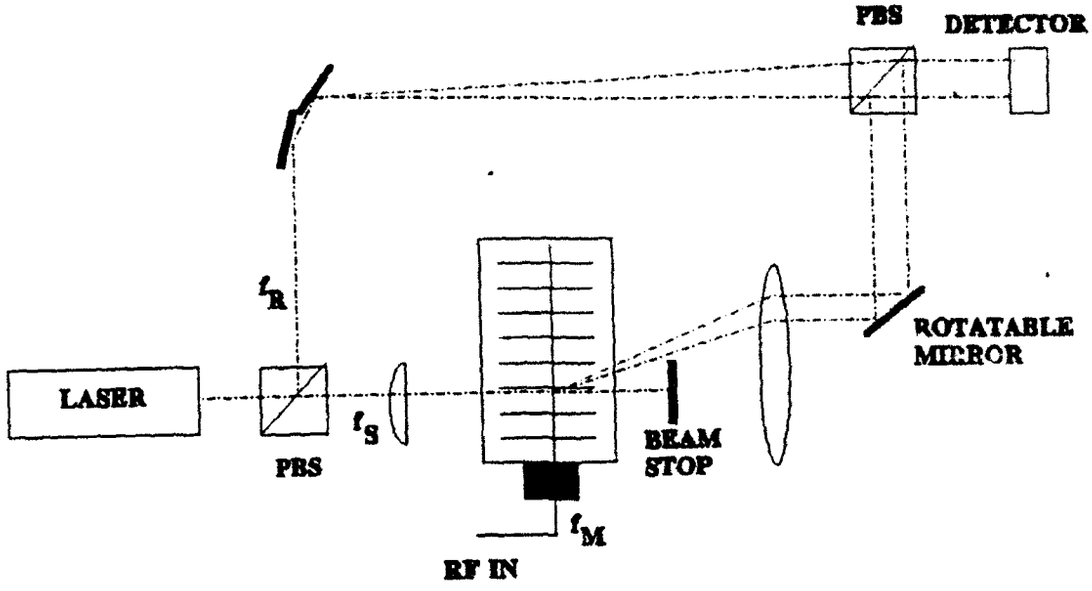


Figure 1

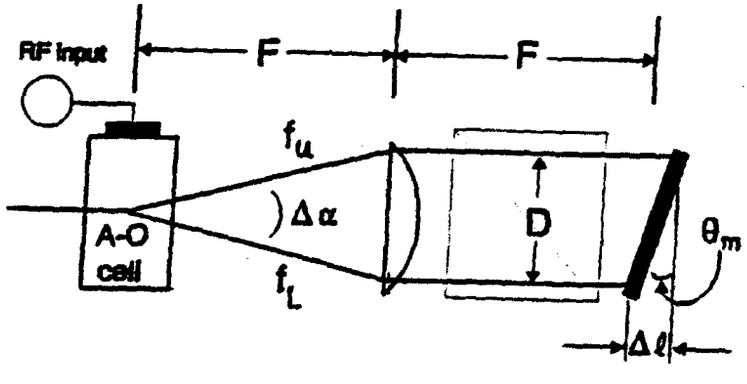


Figure 2

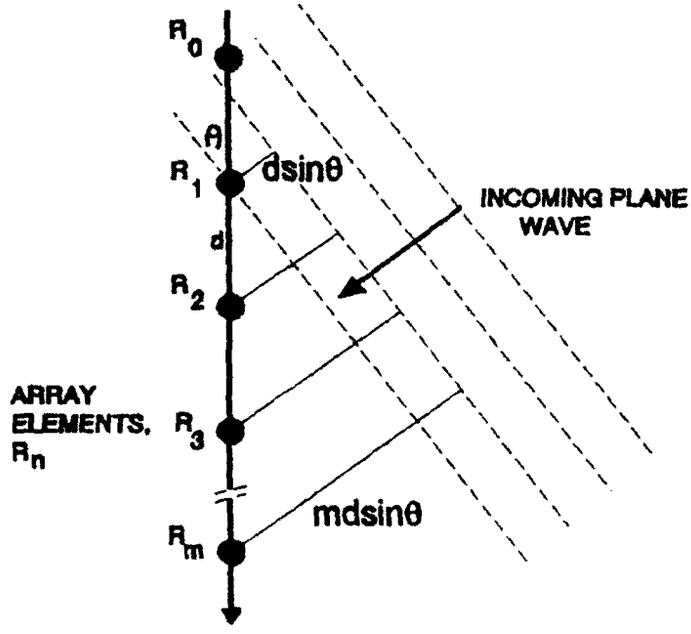


Figure 3

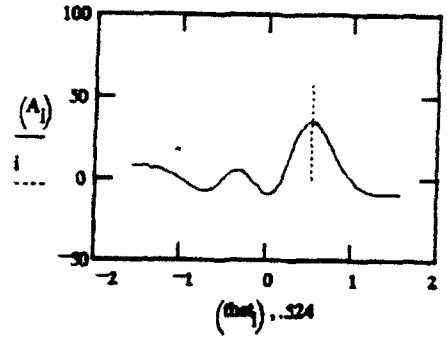


Figure 4

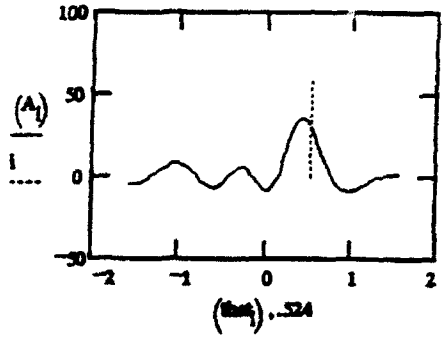


Figure 5a

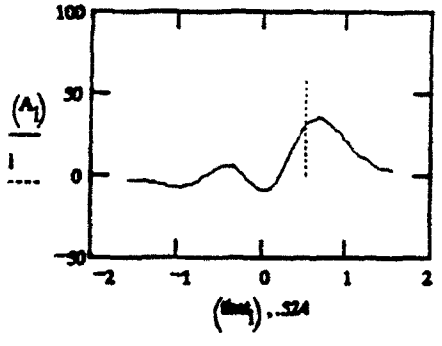


Figure 5b

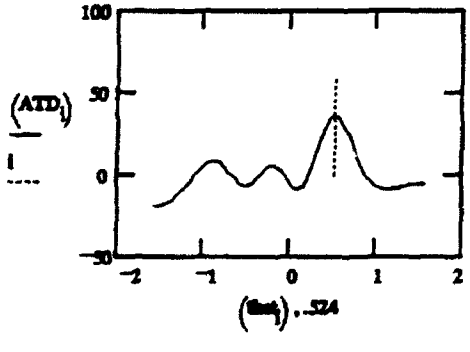


Figure 5c

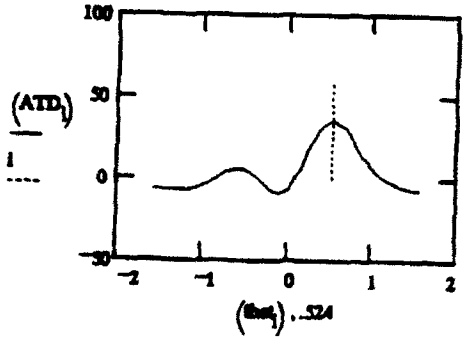


Figure 5d

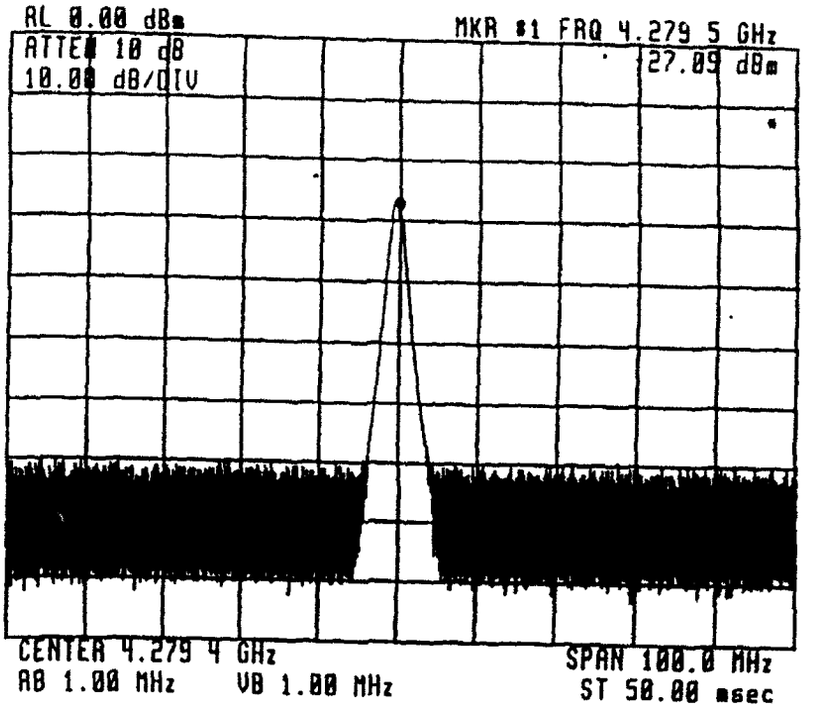

RL 0.00 dBm
MKR #1 FRQ 4.279 5 GHz
ATTEN 10 dB
10.00 dB/DIV
27.09 dBm
CENTER 4.279 4 GHz
SPAN 100.0 MHz
RB 1.00 MHz
VB 1.00 MHz
ST 50.00 msec

Figure 6

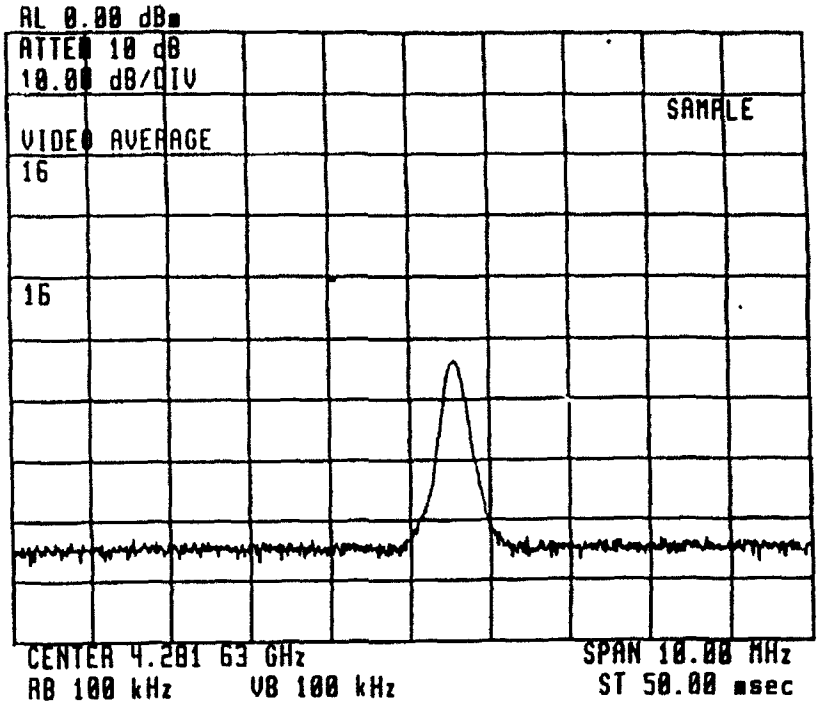


Figure 7a

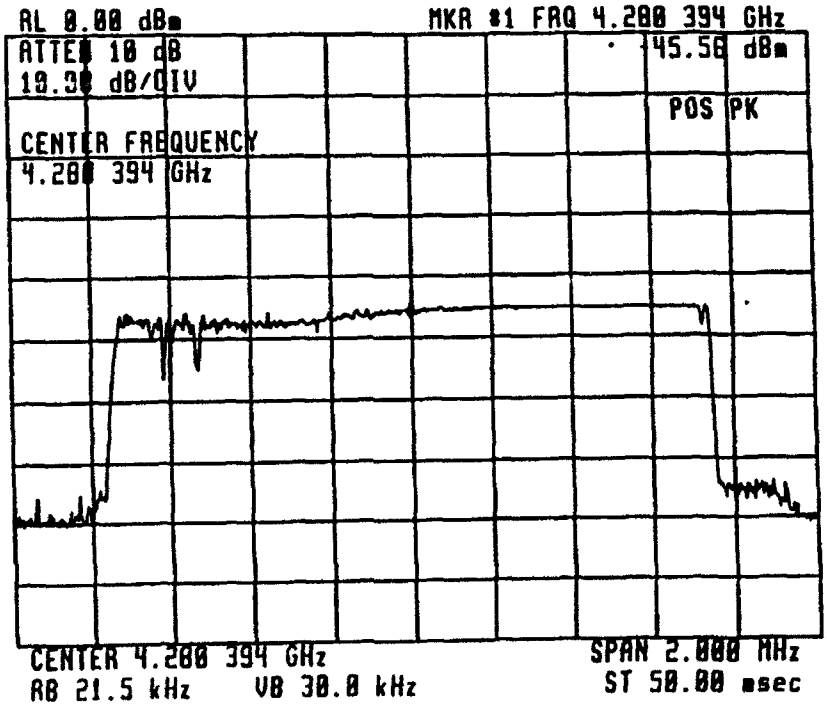


Figure 7b

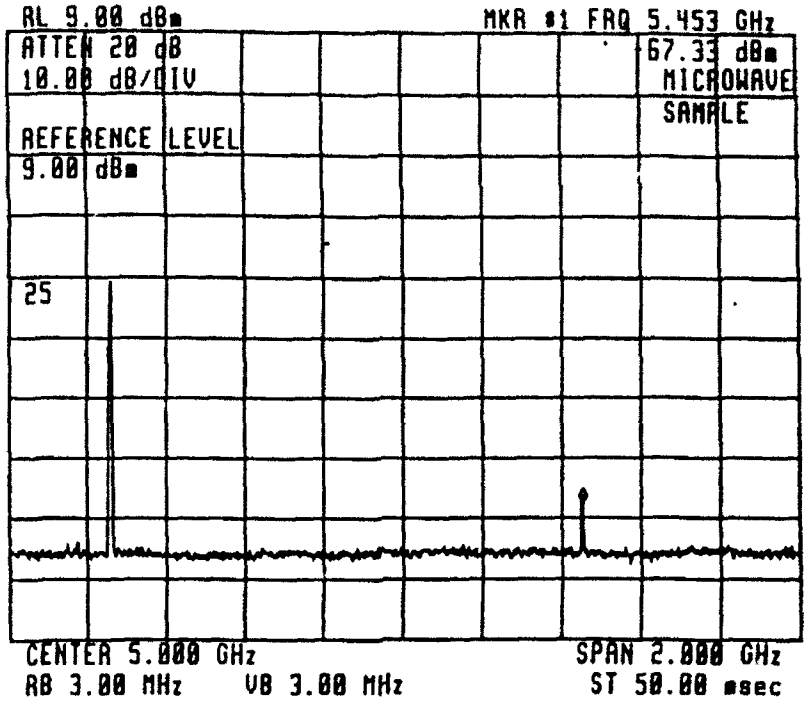


Figure 8

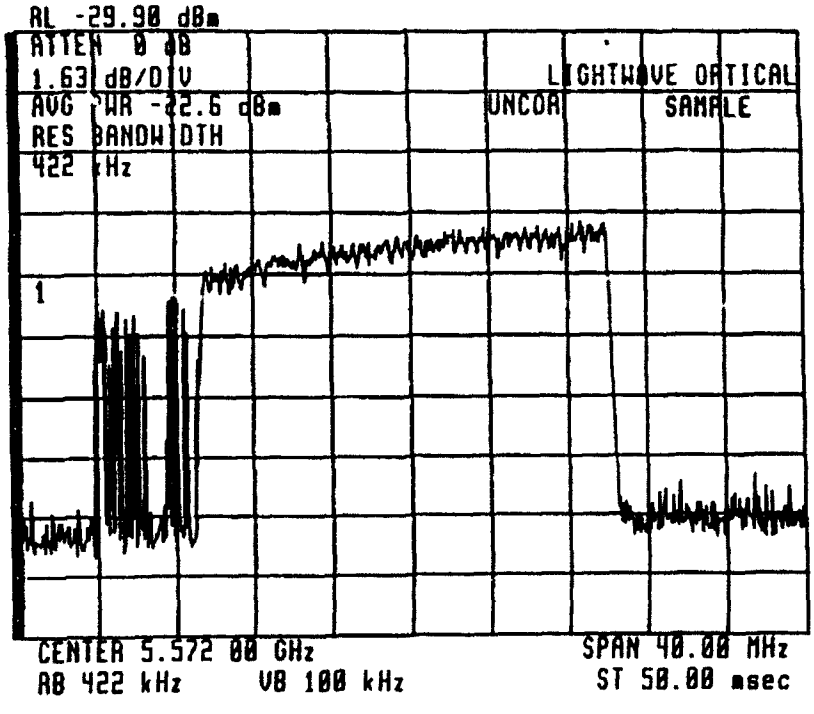


Figure 9

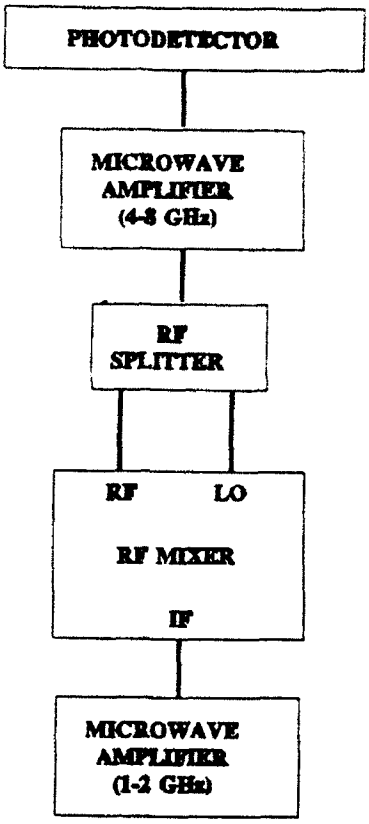


Figure 10

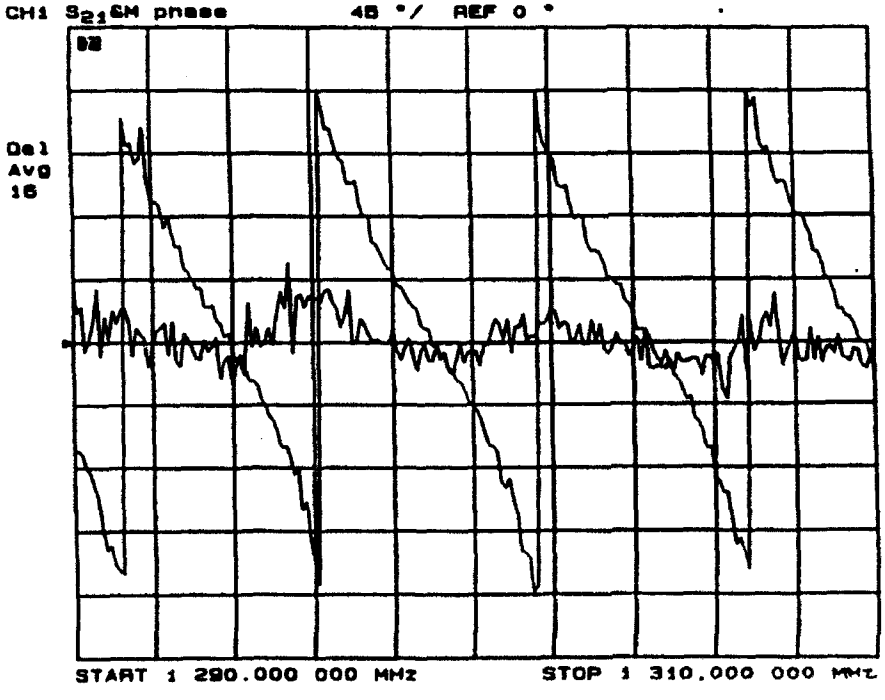


Figure 11

User-Based Requirements for Large-Scale Distributed Information Management Systems: Representation for System Designers

Michael S. Nilan
Associate Professor
School of Information Studies

and

R. David Lankes
Doctoral Student
School of Information Studies

Syracuse University
4-206 Center for Science & Technology
Syracuse, New York 13244-4100

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D. C.

August 1992

User-Based Requirements for Large-Scale Distributed Information Management Systems: Representation for System Designers

Michael S. Nilan
Associate Professor
School of Information Studies
Syracuse University

Abstract

This report describes our efforts this summer to generate a method for translating our user-based information system requirements into a representation form that would be readily interpretable by system designers and system analysts. Typically, the kind of requirements specification that we produce from our user requirements analyses are text-based descriptions of problem solving processes as perceived by a group of users. In the past, these text-based descriptions have proven to be difficult for system designers and analysts to interpret. Since our long term research agenda is oriented towards large-scale information management systems which are more complex than traditional applications, we felt that we needed some systematic and easily interpretable format that we could use for communicating our user requirements. Using a combination of hypertext-like representations and a 3-dimensional virtual reality display, we have been able to create a representation system that not only provides for effective interpretation of our user requirements on the part of system designers and analysts, but the virtual reality graphic display configuration also allows us to represent system components (e.g., display devices, databases, network linkages, etc.) in the same graphic environment. We believe that this combination of hypertext descriptions and virtual reality graphic displays will facilitate the accurate representation of user perspectives in large-scale information resource management systems.

User-Based Requirements for Large-Scale Distributed Information Management Systems: Representation for System Designers

Michael S. Nilan and R. David Lankes

INTRODUCTION

According to the Computer Science and Telecommunications Board of the National Research Council (June 1992), the two most central, important development directions for computing in the United States for the next decade are increased processing capability and improved usability. Of the six "core sub-fields" of computer science and engineering that the Computer Science and Telecommunications Board describe (page 4), three of these fields are clearly pursuing the increased processing capability direction: multiple processors, data communications and networking, and reliability. The other three core sub-fields (software engineering, information storage and management, and user interfaces) would seem to bear the responsibility for pursuing improved usability.

The current trend in computing in general and in information systems in particular is towards distributed environments. This is clear due to the increase in use of Unix as a general, multi-platform operating system; it is clear in the spread of graphic user interfaces (GUI) and windowing systems; it is clear in the rapid spread of local and wide area networks as well as national and international networks; and it is clear in the private sector trend towards "outsourcing" of various computing and information services. One other central phenomenon that has received less attention is how organizations can deal with the exponential growth in the production and use of data/information as they pursue their objectives in this distributed environment. Usability is the essence of coping with this phenomenon.

Our research is directed at enhancing usability for complicated, large-scale information management systems has made significant progress in the user needs assessment or user requirements domain (see Nilan, 1991, 1992a, 1992b, 1992c; Nilan & Hert, 1992; Nilan & Rosenbaum, 1992; Hert & Nilan, 1991; Newby, Nilan & Duvall, 1991). We have a set of coherent needs assessment methods that allow us to generate a dynamic model of the users' perspective of a problem or problem domain that we can use for initial user requirements and we believe that this model can be used later on in the software development cycle for user and system performance evaluation. We also have a systems analysis tool that provides a near real-time feedback loop between system developers/analysts and end-users (Nilan & Travica, 1992). This feedback loop

can be used from the rapid prototyping stage through to the maintenance phase of the software life cycle to provide systems analysts, designers and managers with detailed descriptions of the problems and comments that users have for specific aspects of the system. The analytic procedures associated with the feedback loop provide specific metrics for making system updating decisions (i.e., changing either the data, the code or the peripheral hardware) and for system evaluation (e.g., how well the user is being served by existing database links to the interface). Our user requirements methods are used at the beginning of a system design process and our feedback loop is used from the rapid prototyping stage on. The resulting system (i.e., software, hardware and data) is oriented towards how users perceive the problem the system is designed to solve with several benefits over more conventional design processes. Some of these benefits include: drastically reduced training overhead for new or improved system features; increased effectiveness of human-computer interaction; and usability metrics for system and user evaluation, *particularly* in a distributed operating environment.

What has been missing in our efforts to date however, has been a reliable mechanism to "translate" the initial user problem models (i.e., the user requirements) into a form that was readily interpretable by system designers when we were interested in dealing with a large-scale (and therefore, complicated) information system. What we had been turning over to system designers/developers was a text-based linguistic description of how the users aggregately view the problem that the system designers were trying to develop a system to solve. They had difficulties in understanding how our user problem model could be "translated" into system operation specifications (i.e., user requirements). This difficulty was particularly troublesome when we were addressing a realistically scaled information system (as opposed to a small-scale application) because of the complexity of the model. Therefore, our project for the AFOSR this summer at Rome Laboratories was to develop a representation procedure that would allow system designers to more readily interpret our system performance requirements. Because we are interested in working with large-scale, multimedia information management systems, the additional dimensionality of virtual reality (VR) technology was seen as a potentially powerful way to represent the complexity and interdependencies in our user model.

We used two research teams, one at Syracuse University composed of six doctoral students and one at Rome Labs composed of the principal investigator and one doctoral student. The team at Syracuse University was involved primarily in developing a representation methodology and the team at Rome Labs was concerned with learning the technical capabilities

of the Rome Labs virtual reality graphics system and then creating a three-dimensional demonstration of the representation process on the Rome Lab system. The result of our efforts is reported here and resides as a graphic demonstration on RL/C3AB's virtual reality system. This project represents a significant step in our development of a set of coherent system design methodologies for accurately and effectively representing end-user perspectives in the system design process. We now have a set of three methods for insuring increased usability of large-scale information management systems design (i.e., knowledge acquisition, knowledge representation, and user feedback loop). It is important to note that our efforts this summer have been more successful than we had envisioned. In addition to developing a method to represent user requirements to system developers, the graphic representation also allows us to model various system components, capabilities and requirements within the same display environment. By using virtual reality technology for this graphic representation, system designers and analysts will be able to get a much more wholistic view of the complete system architecture (i.e., hardware, software, data and network linkages) required for a particular application.

THE PROJECT DATABASE

In order to develop our user model representation methods, we first needed a data set that met certain criteria. Among these criteria were that the data set had to be of a realistically sized set of user problems due to our interest in large-scale information management systems; the data had to be composed of dynamic user specifications of problem solving activities, data/information needs associated with those activities; and both of these had to be in user language terms. Although we had empirically created a number of data sets that met the last two of these criteria (e.g., Nilan, 1992b), we had only one data set that met the first criterion. Because we are the only researchers in the U.S. who are involved in this type of model creation, we knew of no existing data sets that met our criteria. Therefore, we used the one data set that met all three criteria.*

This data set was collected for a "Campus-Wide Information System" (CWIS) for Syracuse University. CWISs are currently being implemented all around the world using existing university networks, Internet and Bitnet. These systems typically are simply data that university

* We would like to acknowledge the support of the Nason Foundation, Syracuse University's School of Information Studies, Syracuse University's Computing Services and the students of the IST 720 course in the "Behaviors of Information Users" Fall Semester, 1991 for their generous support in generating the data set for this project.

administrators have sitting around in some database that they make available to students electronically. The data set we collected for Syracuse University was radically different. While we will not attempt a complete description of the data here due to space limitations, we will describe some of its more salient features (the reader is referred to Nilan, 1992a for a conceptual description of this kind of data).

The data was collected in two phases. The first phase was to determine the range of different types of problems students had and involved a total of 28 focus groups of between three and six Syracuse University students who shared some demographic feature (e.g., one focus group was all foreign students, another was all minorities, another was composed of disabled students, etc.). These students were asked to describe problems that they had encountered for which Syracuse University was seen as a source of help or information in solving it. Students were then probed by asking them to think back to when they were making a decision about which university to attend, to when they decided upon Syracuse, to their first semester, to their most recent experiences. Descriptions of all the problems students described were collected. These descriptions were then clustered topically into seven "problem domains" which included: academics, health care, social life, transportation, health care, finances, and library and computer systems. Taken together, all of the problems that students had mentioned in phase one could be classified into one of these problem domains.

In phase two, seven teams of interviewers (with an average of six members per team) completed an average of 43 interviews with students per problem domain (total number of usable interviews was 299) where the respondents were chosen for maximum coverage (i.e., each team had a "quota" of foreign, minority, new, continuing, handicapped, etc.). These interviews, the essence of our knowledge acquisition methods, asked respondents to describe a recent problem in the specified problem domain. This description was in the form of four types of data. First, respondents were asked to describe their problem as a series of activities (things they said or did or thought, things that others said or did or thought, or things that just happened). Next, respondents were asked to describe the questions (things they wanted to find out about, unconfuse, or were just curious about) that they had at each activity in their problem description. It was emphasized that these questions did not need to be asked out loud nor did they need to be answered. Taken together, the activities (arranged on three by five cards horizontally) and the questions (arranged on three by five cards vertically under their respective activity) represent a dynamic "action by cognition matrix" of the respondent's problem. Respondents then rated each

question for its importance in the overall problem solving process and the eight highest rated questions were examined in more detail. The third type of data focused on these eight questions. Respondents were asked how an answer to their question would help them solve the problem, whether they got an answer to their question, if so when they got that answer, whether or not the answer was complete and why so, how satisfied they were with the answer and why so, what sources they considered and why. Finally, respondents filled out a demographic questionnaire that is more typical of traditional knowledge acquisition procedures (e.g., age, experience with computers, major, etc.).

From the data for each problem domain (see Nilan, 1992c for a detailed description of the analytic procedures for this type of data) one to four aggregated problem models were constructed. These models focus on the agreement among respondents reporting on similar problems in the nature of their problem solving activities and are represented in the respondents' own language. The 299 interviews produced a total of twenty-eight individual problem models which represented the fullest range of descriptions of the problems that Syracuse University students have living and studying there. These problem models were all interrelated in that the problem context was the same for all of them (i.e., being a student at Syracuse). This data set met all the criteria set forth for our large-scale information management system requirements: a large number of related user problems (in this data set, we had a total of 28 user problem models that all dealt with study and life at Syracuse University), dynamic models of those problems, with the models in user language.

The task for this project was to see if we could come up with a way to represent the set of user problem models in a way that would effectively and efficiently communicate to a system designer. This involved a series of sub-tasks that relate both to this communication or translation problem but also relate to the larger system issue of usability. For example, because of the large volume of information that is available, users' ability to navigate through databases is currently dependent upon the degree to which the database designer effectively set up the database *and* the extent to which the user knows the database. We are interested in database representations that are based upon the problem being solved rather than relying on a system perspective on database design or upon the user's ability to learn the database. Therefore, the specific representation we want to provide to system designers has some very specific features that address usability. In order to discuss these, the next section will use an object oriented programming metaphor to describe the various aspects of the user problem models into a form that is readily interpretable by system

designers.

THE USERS' ROLE IN SYSTEM DESIGN

The utility of object oriented programming has remained an elusive accomplishment for the last several decades due, to a great extent, with the so-called "granularity" problem, that is, what defines a basic unit or object. Currently, while this problem is less broad than in the past, the granularity problem remains a serious obstacle to creating object oriented operating and programming environments. Our approach to empirically generating user problem models offers one potential solution to the granularity problem that also helps us to communicate with system designers. We call this fortunate juxtaposition a functional approach to object definition. Our argument follows directly from our methodological approach.

Once we have created a user model of a particular problem, we have, in essence, the following products:

- functional objects represented by the aggregated activities or steps in the problem solving process;
- data objects represented by the question and answer sets associated with each activity or step; and
- a set of "dependencies" or relationships among the various functional and data objects.

The functional objects represent specific activities that users undertake at a particular point in the problem solving process. Each functional object is "dependent" upon one or more data objects at that specific point in time. Further, these data objects may have specific relationships with other data objects which are also necessary at that specific point in time (i.e., a problem space or an information space dependency). Each functional and data object has objects that necessarily occur before and after it (i.e., time dependency) that are unique to their specific points in the problem solving process. In other words, when people are solving a problem, they have a series of unique steps that are followed, one after the other. When we look at a large number of people solving the same problem, we find that many of these steps are functionally equivalent *and* that the steps occur in roughly the same temporal order (see Nilan, 1992a, 1992b and Nilan & Rosenbaum, 1992 for discussions of these properties of user models). When we aggregate the information needs of these people and "partition" them (i.e., divide them up according to when in the problem solving process they are most salient), we find that the range of data concerns at any one point in the process is quite manageable (i.e., relatively few questions).

One of the particularly useful aspects of this kind of object approach to representing the

data from our user problem models is that we can create a problem structure that is "recognizable" by novice problem solvers and experts alike. This is the case for two reasons. First, people can see at a glance the whole problem solving process as aggregated across a large number of people. As with language, they understand the process because they have already had experience with similar processes and the language that is used to label the process model comes directly from users rather than from system or content experts - it is not technical language. We refer to this feature as the "navigability" of the model. Its effectiveness can be illustrated if we describe our model as a set of questions (or index terms) used to provide access to our database (or the answers to users' questions). When we look at the whole set of questions, there may be hundreds or thousands of them (in this database, there are over 2,500 questions). For a user to find the specific question s/he has at a particular point in using the system can be a prodigious task! How can you reasonably alphabetize the list? Do you index the list? If so, according to what criteria? Topic? There may still be hundreds of questions in a realistically scaled database that have the same topic. What if the user has a different term to refer to a particular information need? Note that these are common tactics in more traditionally designed systems, manuals and help systems. Whereas with our approach, the user only has to "navigate" to the point in the process where s/he is and look at a list of questions that occur at that point in time (the most in our database is ten). Finding the specific question of interest is trivial!

Therefore, the conceptual approach we took in this project was to try to translate our set of user problem models into a set of objects (functional and data) and specify the dependencies (spatial and temporal) among the various objects. Further, we wanted to be able to present various system objects in the same representation (e.g., display devices, distributed databases, distributed processing capabilities, etc.) so that the user's role in system design could be represented in the same "design space."

METHOD

Employing techniques similar to those we used to aggregate user descriptions into problem models (see Nilan, 1992b), we examined the models for similarities, overlaps and outright redundancy. We also looked for opportunities where the whole set of models taken together would eliminate the need for two models which could be represented together. For example, the three problem models of going to the health center with an injury or illness, getting immunizations, and getting health insurance, while seemingly disparate, nevertheless when

represented together in the whole set of problem models, users had no problem whatsoever in going immediately to the right problem model. In this manner, we were able to reduce the number of problem models from the original 28 down to 18 models.

The next task was to look at each individual model and examine it for its coherence and suitability as a navigation aid. This involved looking at the individual steps or activities in each model and trying to decide if they were necessary for providing access to information. For example, some steps just did not make sense; in one model, the first step was "decide whether or not I needed a job." While several of our respondents had mentioned this as their first step in describing their problem solving process to the interviewers, it doesn't make sense for a user to come to an information system to ask whether or not s/he needs a job! The user will already know whether or not s/he needs a job. Therefore, because we could assume that this step did not need representation in the system, we eliminated it. Another type of adjustment we made to the original models was to combine or eliminate those steps for which respondents had no questions (or data objects). If the step could be combined with the proceeding or following step and still maintain the navigable coherence of the model, this was done to simplify the complexity of each individual model. In some cases, a step would be described to the interviewer as a rhetorical device rather than as a specific point in the problem solving process (these steps *never* had questions associated with them) and so these were also combined or eliminated.

The next step, while not strictly necessary, was taken because it allowed us to simplify the potential representation of the set of problem models to the user at the human-computer interface level. This involved clustering those problem models that seemed to go together by virtue of having similar situational or contextual orientations. For example, the problems of getting admission to a school, registering for a course, declaring a study major, graduating, etc. were put in a cluster labeled "academics" while problems that had less to do with study per se but had to do with life surrounding the university environment were put in a cluster labeled "living." The last cluster was more of a miscellaneous collection of problems that students would all have at some time in their academic careers but were less pervasive. We called this cluster "coping."

RESULTS

The three resulting clusters and their associated problem models have been represented as a three tiered hierarchy that can be used in two significant ways: first as a guide to designing the intellectual structure of the human-computer interface and second, as a basic structure for this

project of providing the system designer with a coherent set of user requirements. Each of these will be discussed in turn below. We believe that the isomorphism between these two functions is a major source of validity and coherence for the approach taken here.

To illustrate how these clusters were configured, Figure 1 shows the "academics" cluster with eight problem models listed under it. These were: admissions, registration declare major, transfer, holds (e.g., a student cannot register for classes until his/her tuition bill is paid), graduation, computers and library. The problem model for admission (the top line in Figure 1) had four steps: take the required exam(s), contact the school, check to see if everything is ok, and complete the (admissions) application. Linked (in a hypertext sense) to each step are a series of questions aggregated across all respondents that we interviewed. These questions (not shown in Figure 1) are also linked (in a hypertext sense) to specific data points in databases maintained by administrative units at Syracuse University that have responsibility for those answers. To navigate this system from the user's point of view, the user would select "academic" as the problem domain that s/he wanted to find out about, would select "admissions" as the specific problem s/he wanted to solve, and then the step that s/he was at in the problem solving process in order to gain access to the specific information needed. This scenario as illustrated by Figure 1 is simply a logical structure that would serve as a starting point for the design of a user interface *and* would serve as a "map" for the subsequent system design linking the interface structure to the requirements for specific database data points.

We next developed a categorization scheme for the various system components and capabilities to include with our problem models in the graphic domain. For example, individual problem clusters are represented as spheres, databases as cones, etc. This will allow system designers/analysts to readily interpret the various system features that are necessary to deliver the functionality of this particular information system. A problem domain "space" was created (albeit arbitrarily) to contain the various user problem clusters (spheres) and the system components and capabilities. (Note this later notion of "capabilities" allows us to specify a certain functionality even if the technology for delivering that functionality is not yet available for implementation. One example might be specifying a parallel processor for use in load balancing across our system network. While parallel machines do exist, their use for load balancing is not developed yet).

Figure 2 presents a generalized example of how this graphic representation might look.

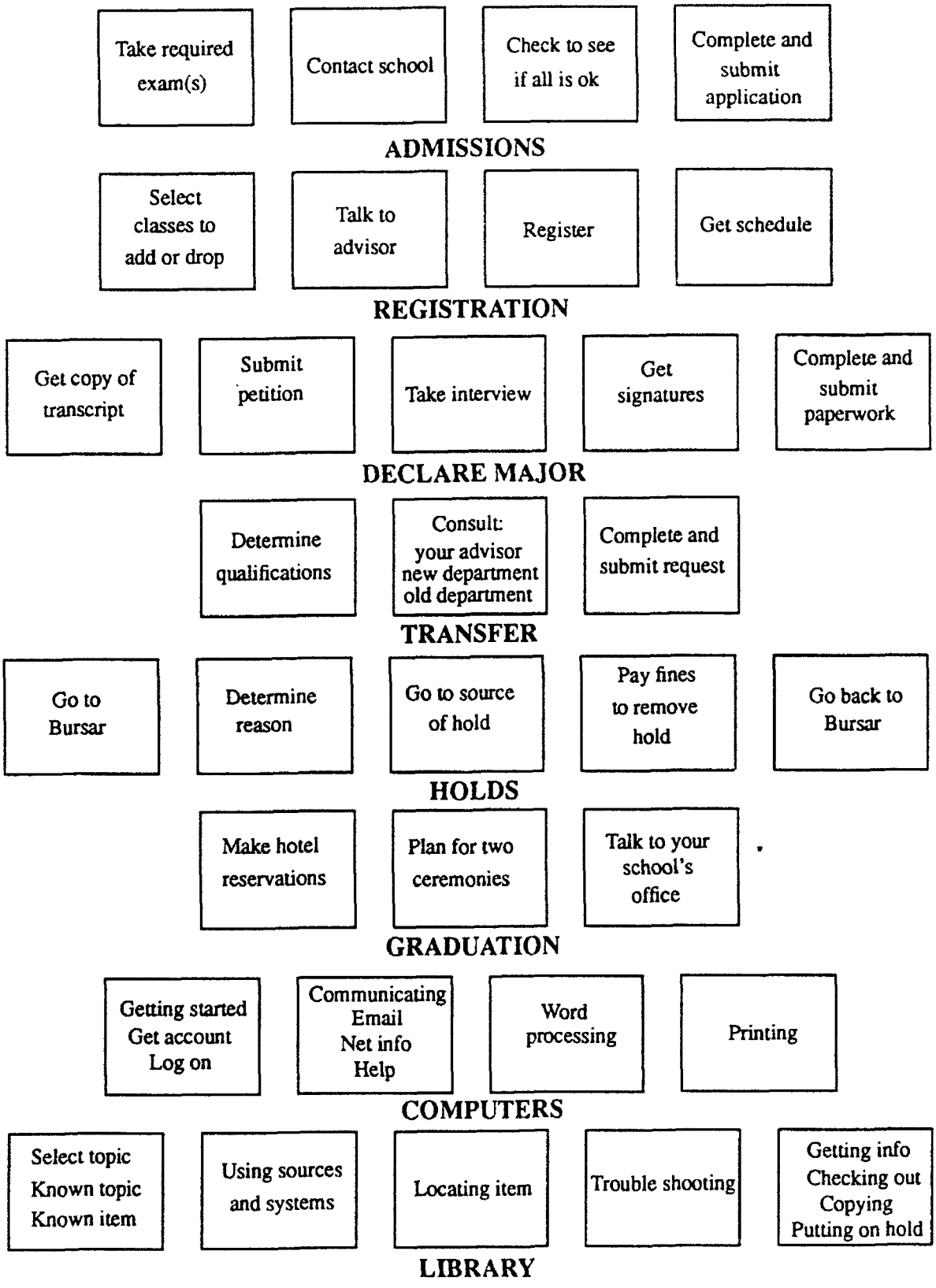


Figure 1. The Academics cluster and its associated user problem models.

Figure 2. Example of a generalized graphic space for user problem model objects and system component objects showing the larger "cognitive" or system space, a window describing the location and nature of linkages among objects, and a "high resolution" window showing the specific user problem that the system designer is looking at.

on an actual screen (it looks different on Rome Labs' Silicone Graphics VR system). This screen has four features worth mentioning here:

- a "cognitive space" window which represents the system space which would include the various user problems (here represented as spheres), and system components (here represented as rectangular prisms);
- movement capability within the space (the four arrows in the center of the screen);
- a descriptive list of the specific linkages between where the user is in the system and specific databases, display devices, processing capabilities, etc. (here presented in the window along the right-hand side of the screen); and
- a "high resolution scheme" which shows the specific user problem model the system analyst is focusing on (note that this model is virtually the same as the "transfer" problem in Figure 1).

In the VR system, we will also have a "meta-" view window that will present the whole system space to the system designer/analyst so that s/he always has an overview of where in the space s/he is at any particular point in time.

The way this graphic presentation is used is as follows: The system designer/analyst could move through the system from the users' perspective by following the hypertext representation (Figure 1) and the graphic representation at the same time. This would be done by moving systematically through each problem model in each cluster until you had exhausted each one. As the system designer/analyst is moving through the user view, s/he can specify the needed system components, databases, processing and display capabilities that can or should be used to insure that the user is getting what is needed in the most effective and efficient manner possible. As new system requirements are articulated, the system designer/analyst can insert them into the space at a convenient place and establish the necessary linkages.

We see this graphic display as more of a design tool (analogous to a CAD or computer assisted design tool) where the space provides an environment for merging user and system views of a particular information resource management application. It is a notation system that will allow for coherency and completeness checks before the system is actually built.

CONCLUSIONS

Our system is in its first iteration and is virtually untested with actual system designers or

analysts. We need to examine how these system users will perceive the utility of both the hypertext representation scheme and of the graphic design tool before we can draw specific conclusions about this particular iteration. Further, there are certain capabilities that are absent from the current iteration. For example, we have no facile editing capability that is operational yet. This editing capability would allow the system designer/analyst to insert system components and capabilities as well as make necessary linkages among system and user components. It would also be useful to have "bench test" capability where the system designer/analyst could simulate the operation of system linkages (or even the whole system) before anything is actually built. Although these capabilities are beyond our goals for this summer, they present some interesting possibilities for continued development of our notion of a knowledge representation scheme for system design.

References

Hert, C. A. and Nilan, M. S., (1991). "User-Based Information Retrieval System Interface Evaluation: An Examination of an On-line Public Access Catalog." **Proceedings for the 54th Annual Meeting of the American Society for Information Science**, **28**, pp. 170-177.

National Research Council, (1992). **Computing the Future: A Broader Agenda for Computer Science and Engineering**. Washington, D.C.: National Academy Press.

Newby, G. B., Nilan, M. S. and Duvall, L. M., (1991). "Towards a Reassessment of Individual Differences for Information Systems: The Power of User-Based Situational Predictors." **Proceedings for the 54th Annual Meeting of the American Society for Information Science**, **28**, pp. 73-81.

Nilan, M.S., (1992a). "Cognitive Space: Using Virtual Reality for Large Information Resource Management Problems." **Journal of Communication**, 42:**4**, in press.

Nilan, M. S., (1992b). "User-Based Information/Communication Systems: Using Sense-Making to Create a User Model for a Desktop Publishing Help and Training System." In B. Dervin (Ed.), **Methodology Between the Cracks: Theory and Method for the Qualitative and Quantitative Study of Human Communication**. Norwood, NJ: Ablex Publishing Company, in press.

Nilan, M.S., (1992c). "Whose Reality? The VR Technology Agenda in Medicine." **Proceedings of the First Annual Medicine Meets Virtual Reality Conference**, **1**,

pp. 183-191.

Nilan, M. S., (1991). Application-Based Utility Evaluation: A Discussion Leading to Assessing the Role of End-Users in the Exploitation of Virtual Reality Technology. Report to Richard Slavinski, Director of Virtual Reality Research, Rome Lab, September, 1991.Also presented to the Air Force Office of Scientific Research, October, 1991.

Nilan, M.S. and Hert, C.A., (1992). "Incorporating the User in System Evaluation and Design." Proceedings of the National Online Meeting, pp. 217-234.

Nilan, M. S. and Rosenbaum, H., (1992). "An Epistemology for Sense-Making Research."In B. Dervin (Ed.), Methodology Between the Cracks: Theory and Method for the Qualitative and Quantitative Study of Human Communication. Norwood, NJ: Ablex Publishing Company, in press.

Nilan, M.S. and Travica, B., (1992). "An On-Line Feedback Loop for End-User and System Analyst Communications for Distributed Systems." Paper submitted to INTERCHI '93.

FLUX CREEP IN A Y-Ba-Cu-O FILM CHARACTERIZED BY A c-AXIS MICROSTRUCTURE IMBEDDED WITH a-AXIS ORIENTED GRAINS

John L. Orehotsky
Professor
Departments of Physics and
Mechanical/Materials Engineering

Wilkes University
River Street
Wilkes-Barre, PA 18766

Final Report for AFOSR Summer
Research Program, Rome Laboratory
Hanscom Air Force Base

Sponsored by Research & Development
Laboratories, Culver City, California

August, 1992

FLUX CREEP IN A Y-Ba-Cu-O FILM CHARACTERIZED BY A c-AXIS MICROSTRUCTURE IMBEDDED WITH a-AXIS GRAINS

John L. Orehotsky
Professor
Departments of Physics and
Mechanical/Materials Engineering
Wilkes University

Abstract

Flux creep in an epitaxial grown, YBCO high temperature superconducting oxide thin film was examined by magnetization measurements in the plane of the film as a function of time at temperatures between 5 and 30°K and in fields between 0.5 and 2.0 Telsa. In this particular film, the magnetization relaxation when plotted on a logarithmic time scale generally exhibited a two-stage kinetic response which was most evident at the lower temperatures and fields. In a field of 0.5T, the experimentally apparent activation energy characterizing the first stage creep response was found to increase with temperature, approaching a plateau level of 33meV at 30°K. At 15°K, the activation energy was found to be relatively field insensitive. These response features for the first stage behavior are in modest agreement with theoretical concepts. Because of insufficient data at long time values, no attempt was made to characterize the second stage behavior. The experimentally observed two-stage kinetic response feature in this film is suggested to be a direct consequence of the particular microstructural details of this film.

Introduction

Before the full potential of the high temperature superconducting oxides can be realized in device applications, the critical current density problem must be resolved. In single crystals and highly oriented, polycrystalline thin films where weak-link grain boundary limitations are minimized, the magnitude of the critical current density at any particular field and temperature appears to be governed by the flux creep effect which is expected to be a structure sensitive phenomena. This suggests that by immobilizing the flux lines on a strong pinning defect structure, exceptionally large and relatively field independent critical current density values will be obtainable. Unfortunately, the theory of flux creep is not precisely established, and experimental investigations often present an assortment of results that sometimes raise more questions than provide answers.

The theory of flux creep originated with an attempt (1) to explain the magnetic relaxation effect typically observed in the low temperature, type II intermetallic superconductors. Inherent in this treatment is the concept of the critical state (2) and the concept that flux pinning defects exist in the material so that in the presence of a magnetic field and these pinning defects, flux lines entering the sample establish a gradient (∇B) and then move collectively in bundles down that gradient by a thermally activated process involving a jump frequency (γ) from pinning site-to-pinning site given by:

$$\gamma = \gamma_o \exp - \frac{U}{kT} \qquad (1)$$

where U is the effective pinning barrier energy and γ_o is the attempt frequency. This effective activation energy is equated to (3) a classical potential well of energy height U_o representing flux pinning in zero field minus the energy associated with a Lorentz-type force (4) experienced by the flux lines in the presence of the gradient where this force is taken to be directly proportional to the gradient. The effective activation energy is then linearly related to the ∇B gradient and can be expressed in general terms as:

$$U = U_o [1 - \frac{\nabla B}{\nabla B_o}]^n \qquad (2)$$

where a value of n=1 specifically refers to the linear gradient assumption, where ∇B_o is the maximum gradient that the flux bundle can experience and still remain pinned at 0°K, and where U_o is expected to decrease slightly with temperature for $T \ll T_c$ and rapidly as T approaches T_c.

Using this linear ∇B driving force relationship for the effective barrier energy, it can be shown that the irreversible magnetization (M_{irr}) associated with flux creep at a given field and temperature when $U_o >> kT$ is logarithmically time dependent.

$$M_{irr}(t) = M_{irr}(0) [1 - \frac{KT}{U_o} \ln(\frac{t}{t_o})] \qquad (3)$$

where t_o is inversely proportional to the attempt frequency. As a consequence, experimental data are typically expressed and analyzed

in the form

$$\frac{1}{M}\frac{dM}{d\ln t} = \frac{-kT}{U_e} \qquad (4)$$

where the experimentally accessible, apparent activation energy U_e equals U_o if the linear gradient relation is physically functional. Using this same linear gradient relationship, Monte Carlo simulations (5-6) show that the time dependence of M_{rem} is linear at very short time values, logarithmic at intermediate time values and exponential at very long time values.

The proposed linear gradient relationship based on the critical state assumption is suggested to be physically unrealistic, and a non-linear gradient expression characterized by n>1 is believed (7-8) to be more representative of reality. Indeed, the linear gradient relation corresponds to a highly questionable "V" shaped pinning potential well, and when a variety of shapes, sizes and distributions of wells are examined (8), the activation energy (Eq 2) could be best approximated by n values in the range $3/2<n<2$ instead of the linear n=1 value.

More recent treatments of U as a function of ∇B (or J through $4\pi J/c = \nabla B$) predict (8-11) a variety of forms other than Equation 2. Thus, the magnetization time decay kinetics can be derived from any particular U-∇B expression (12) and will not necessarily be directly proportional to logarithmic time as predicted in the Anderson-Kim model at low temperatures and fields. Models based

upon collective pinning concepts (10, 13-15) predict a power law logarithmic time dependence for the magnetization at temperatures well below T_c of the form:

$$M(t) = M(o)\left[1 + \frac{KT}{U_o}\ln\left(\frac{t}{t_o}\right)\right]^{1/n} \qquad (5)$$

with values of n both less than and greater than one. This form implicitly suggests a nonlinear U-∇B relationship. The complexity of the theoretical problem becomes even more apparent when it is recognized that a variety of possible pinning defects exist, and these defects represent a distribution of pinning potentials.

Experimentally, the simple n=1 logarithmic time dependence is frequently observed (8, 16-18) in single crystal, polycrystals, and thin film YBCO samples at low temperatures and fields, particularly at longer time values where initial transient effects are no longer operative. At higher fields and temperatures, a simple logarithmic time data fit is not readily apparent. There also exist experimental magnetization data on a single crystal sample of YBCO where the power law form is obeyed (19) over many decades in logarithmic time, with n values that are temperature dependent and range between 0.8 and 1.4.

However, even when a simple logarithmic time dependence is apparent in the magnetization data, the extracted experimental activation energy was found to increase and not, as expected, decrease with temperature. Also, the activation energies extracted from magnetization measurements are typically in the meV energy

range as compared with flux creep activation energies extracted from transport measurements that are typically in the eV range and do systematically decrease with temperature (see reference (8) and the references therein).

To help understand this large discrepancy in reported U_e energy values between two different types of measurements and the apparent anomalous temperature dependence of U_e as measured by magnetic relaxation, the suggestion that the activation energy is probably not linear in ∇B was analyzed by Xu etal (8) in considerable detail. The results showed that the experimentally determined activation energy U_e extracted from magnetization decay measurements is indeed smaller than the actual pinning potential U_o, perhaps by a considerable amount, and that U_e will increase with field and also with temperature, but by progressively smaller amounts (concave down) which did not agree with typical concave up experimental results seen by them and others (6) monitoring flux creep kinetics. In the final analysis, they concluded that either the theories of flux creep are inadequate or that their c-axis oriented granular sample is not appropriate for the task of examining these theories. The suggestion that the granular sample is sufficiently flawed to adequately test the theories should be considered in more detail, particularly since the suggestion, when viewed in the light of typical granular critical current response characteristics, appears to have some merit.

Recent critical current density measurements (20) as a function of field on epitaxial grown thin films characterized by

various microstructures of a-axis grains imbedded in a c-axis-oriented film matrix, displayed critical current values between 10^5 and $10^6 A/cm^2$ that where relatively field insensitive when the field was applied in the plane of the film and, depending on the details of the microstructure, were either field sensitive or insensitive when the field was applied perpendicular to the plane of the film. It would appear that the boundaries between the a-axis and c-axis grains function as effective pinning sites, and that samples with this type of microstructure characterized by a dominant pinning site when the field is applied parallel to the film would provide additional information on the flux creep effect and insight into the prevailing theories of flux creep particularly since some of these theories (12, 16, 21) address the experimental situation where the field is applied in the film plane.

Experimental Procedure

A high quality YBCO thin film was made by off-axis sputter deposition on a 100 oriented $LaAlO_3$ substrate. An X-ray diffraction tracing of the film revealed both the <h00> and <00l> family of diffraction peaks indicating that the film was textured with grains having a- and c-axis orientations perpendicular to the plane of the film. Scanning electron microscopy displayed a basket-weave structure of the a- and c-axis oriented grains.

A Quantum Design SQUID magnetometer was used to measure the magnetic properties of the sample where the sample was mounted so that the magnetic field was in the plane of the film since this arrangement in the critical current results (20) displayed the

maximum flux pinning effect. This mounting arrangement requires considerable care since any angular misalignment caused distorted, asymmetric SQUID wave forms, unsuitable for magnetic measurements. How to achieve the desired magnetic field for the creep measurements is a delicate question since the magnetization decay kinetics particularly at short time values, is dependent on the ramp used to obtain the desire field (6). Xu, etal (8) exercised great care in their measurements and provided a detailed description of their procedures. Except for some minor variations, their experimental prescriptions were followed exactly. To avoid field overshoot and its influence on the measured magnetism of the sample, they indicated the sequence of field steps used to achieve a specific field at which the magnetic relaxation would be measured, but did not indicate the time interval, if any, at a field between steps. A two and one-half minute hold time at each step-field was used in this investigation to allow for an apparent fast flux creep transient effect at each step-field. Also, prior to magnetization measurements at any specific field, the sample was brought to its normal state at 100°K and then zero field cooled to the desired temperature before stepping up to the desired field rather than employing the approach of monitoring the magnetization decay kinetics at various pause fields while tracing the hysteresis loop for increasing and decreasing fields. Otherwise, the fields and temperatures used in this investigation were contained in the range of fields and temperatures used in their investigation where the critical state was suggested to exist. In this investigation

the temperature dependance of flux creep was investigated at 0.5T at temperatures between 5 and 30K. The field dependance of flux creep was monitored at a temperature of 15K for fields between 0.5 and 2.0T.

Each measured magnetization value was the average of 3 scans on the SQUID. The run time (t) associated with each measured magnetization value is defined as the difference between mid-time (t_m) of the 3 scan measurement sequence and the start time (t_s) of the run when the persistent mode switch turns on and the desired final field was obtained after the step sequence.

$$t = t_m - t_s \tag{6}$$

Great care must be exercised in measuring these time values since any data point fit to a logarithmic dependance at the short time end of the scale is crucially dependent on the accuracy of these time measurements. The entire creep run was typically confined to a three hour time span.

Finally, a hysteresis loop was measured to maximum fields of ±2.0T (Fig. 1).

Results

The results for the magnetization behavior at various temperatures for a field of 0.5T as a function of logarithmic time between 5 minutes and 3 hours are shown in Figure 2 where an apparent and unexpected two stage response feature is seen at all temperatures and where the first stage systematically blends into

and becomes increasing less distinguishable from the second stage with increasing temperature. This same data when presented on a normalized magnetization basis where the normalization factor is taken to be the intercept on the magnetization axis at logarithmic time equals zero, is presented in Figure 3. From the slopes of these plots, an apparent pinning energy can be obtained easily as a function of temperature certainly for the first-stage behavior. The apparent pinning energy from the first stage is presented in Figure 4 where the energy is seen to increase concave downward with temperature. Finally, the field dependence of the apparent pinning energy at 15K was determined by evaluating the slopes of the normalized magnetization - ln t response features for fields of .5, 1.0, and 1.5T as shown in Figure 5. First-stage behavior for all three fields have the same slope, showing that the pinning energy is relatively field insensitive.

Discussion

The two-stage kinetic response feature of the magnetization as a function of logarithmic time in the time range between 5 minutes and 3 hours was totally unexpected since it apparently has never been observed before in this time range for the YBCO system. There are numerous possible explanations for this apparent effect of which two are most likely. The first explanation involves the microstructure of this particular film and its orientation with respect to the applied field. One stage could correspond to the kinetics of flux lines entering the a-axis oriented grains and the other stage represents the kinetics of flux penetration into the c-

axis grains. Anisotropic flux penetration is certainly possible in the high anisotropic YBCO crystal structure particularly if the Cu-O planes represent an intrinsic pinning site. The difficulty with this explanation is that the penetration should occur concurrently in both the a- and c-axis grains and not sequentially as implied by the experimental data.

The second possibility is that the first stage would be associated with the kinetics in a partial critical state where the flux front has yet to penetrate to the center of the sample, and the second stage represents the kinetics of flux penetration in the full critical state. This situation of partial and full penetration have been examined in some detail (21) where the derived magnetization decay kinetics displays very definite two-stage response features. Depending upon the temperature/activation energy ratio, both stages could appear to be logarithmic in time. Monte Carlo simulations (6) show that the magnetization kinetics of the partial critical state is not logarithmic in time, but appears to be at the longer time values where the flux front is approaching the center of the film. The attractive feature about this suggestion is that the stages are predicted to occur sequentially as implied in the experimental data. The unattractive features are that the hysteresis behavior would indicate that the sample should be fully penetrated at the experimental creep field value of 0.5T and that the time value characterizing the break between the stages should decrease with increasing temperature which is not readily apparent in the data.

Both suggestions have some merit and the correct interpretation awaits the appropriately definitive experiments.

The activation energy for the first kinetic stage behavior as extracted from the normalized slope characteristics is in the meV energy range and does increase with increasing temperature. Both these features are consistent with previous results obtained from magnetization relaxation measurements (6,8). Unlike previous experimental results, its increasing temperature dependence was concave down as predicted (8). Before this apparently gratifying result can be accepted, it must be recognized that it was obtained from normalized slope characteristics using the total magnetization at logarithmic time equals zero (M (ln t = 0)) instead of M_{irr} (0) as the normalization factor. If M_{irr} (0) is taken to be the difference between M (ln t = 0) and the reversible magnetization M_r where M_r is the mid point value of M on the hysteresis loop (8, 22) shown in Figure 1, M (ln t = 0) is then an adequate representation of M_{irr} (0), and the concave down feature of the experimentally extracted activation energy as a function of temperature is essentially correct.

While the temperature dependence of the experimental activation energy is in agreement with expectations, the field dependence is not. The first stage decay kinetics appear to be field insensitive in disagreement with an expected dependency (8).

Because of a lack of sufficient long time data points, no attempt was made to characterize the second kinetic stage. The straight lines placed on this stage in the various figures are only

done as a means of delineating the two-stage behavior, and should not be regarded as expressing a logarithmic time dependence.

Conclusions

The increasing concave down temperature dependence of the experimentally measured flux creep activation energy for this particular YBCO film with its apparent strong pinning microstructure, provides a measure for experimental support for theoretical concepts of flux creep based on a non-linear gradient functional form for the pinning activation energy where this form is associated with physically realistic shapes for the potential well. The two-stage kinetic response feature evident in the data may also be a result of the characterizing microstructural features of this film.

References

1. P.W. Anderson, Phys. Rev. Lett. 9, 309 (1962)
2. C.P. Bean, Phys. Rev. Lett. 8, 250 (1962)
3. P.W. Anderson and Y.B. Kim, Rev. Mod. Phys. 36, 39 (1964)
4. J. Friedel, P.G. DeGennes, and J. Matricon, Appl. Phys. Lett. 2, 119, (1963)
5. C.W. Hagen, R. Griessen and E. Salomons, Physica C157, 199 (1989)
6. R. Griessen, J.G. Lensink, T.A.M. Schroder and B. Dam, Cryogenics 30, 563 (1990)
7. M.R. Beasley, R. Labush and W.W. Webb, Phys. Rev., 136, A335 (1964)

8. Y. Xu, M. Suenaga, A.R. Moodenbaugh and D.O. Welch, Phys. Rev. B, 10882 (1989)

9. E. Zeldov, N.M. Amer, G. Koren, A. Gupta, R.J. Gambino and M.W. McElfresh, Phys. Rev. Lett. 62, 3093 (1989)

10. M.V. Feigel'man, V.B. Geshkenbein, A.I. Larkin and V.M. Vinokur, Phys. Rev. Lett. 63, 2303, (1989)

11. A. Barone, A.I. Larkin and Y.N. Ovchinnokov, J. Supercond. 3, 155, (1990)

12. D.O. Welsh, Supercond. Sci. Technol. 5, S109, (1992)

13. M.P.A. Fisher Phys. Rev Lett. 62, 1415, (1989)

14. A.P. Malozemoff and M.P.A. Fisher, Phys Rev. B 42, 6784, (1990)

15. D.S. Fisher, M.P.A. Fisher and D.A. Huse Phys. Rev. B 43, 130 (1991)

16. R. Greissen, C.F.J. Flipse, C.W. Hagen, J. Lensink, B. Dam and G.M. Stollman, J. Less-Common Met. 151, 39 (1989)

17. M. Tominen, A.M. Goldman, and M.L. Mecartney, Phys. Rev B 37, 548 (1988)

18. Y. Yeshurun and A.P., Malozemoff, Phys. Rev. Lett. 60, 2202, (1988)

19. J.R. Thompson, Yang Ren Sun and F. Holtzberg, Phys. Rev. B, 44, 458 1991

20. H. Fuke, H. Yoshino, M. Yamazaki, T.D. Thanh, J. Nakamura and D. Ando, Appl. Phys. Lett. 60, 2686 (1992)

21. C.J. Van der Beek, G.J. Nieuwenhuys, and P.H. Kes, Physica C, 185-189, 2241 (1991)

22. M., Konczykowski, A.P. Malozemoff and F. Holtzberg, Physica C, 185-189, 2203 (1991)

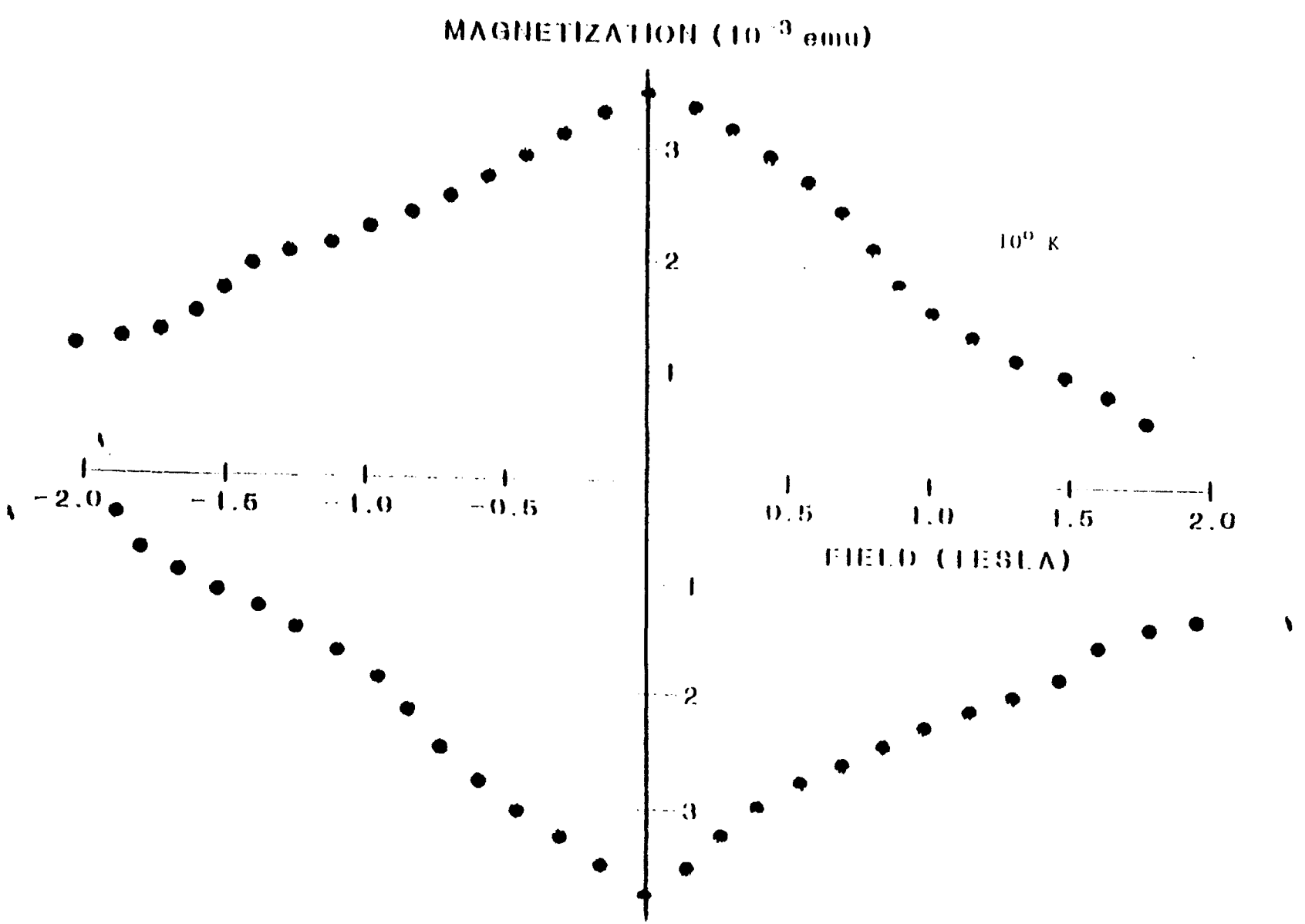


Fig 1 - Magnetic hysteresis at 10K

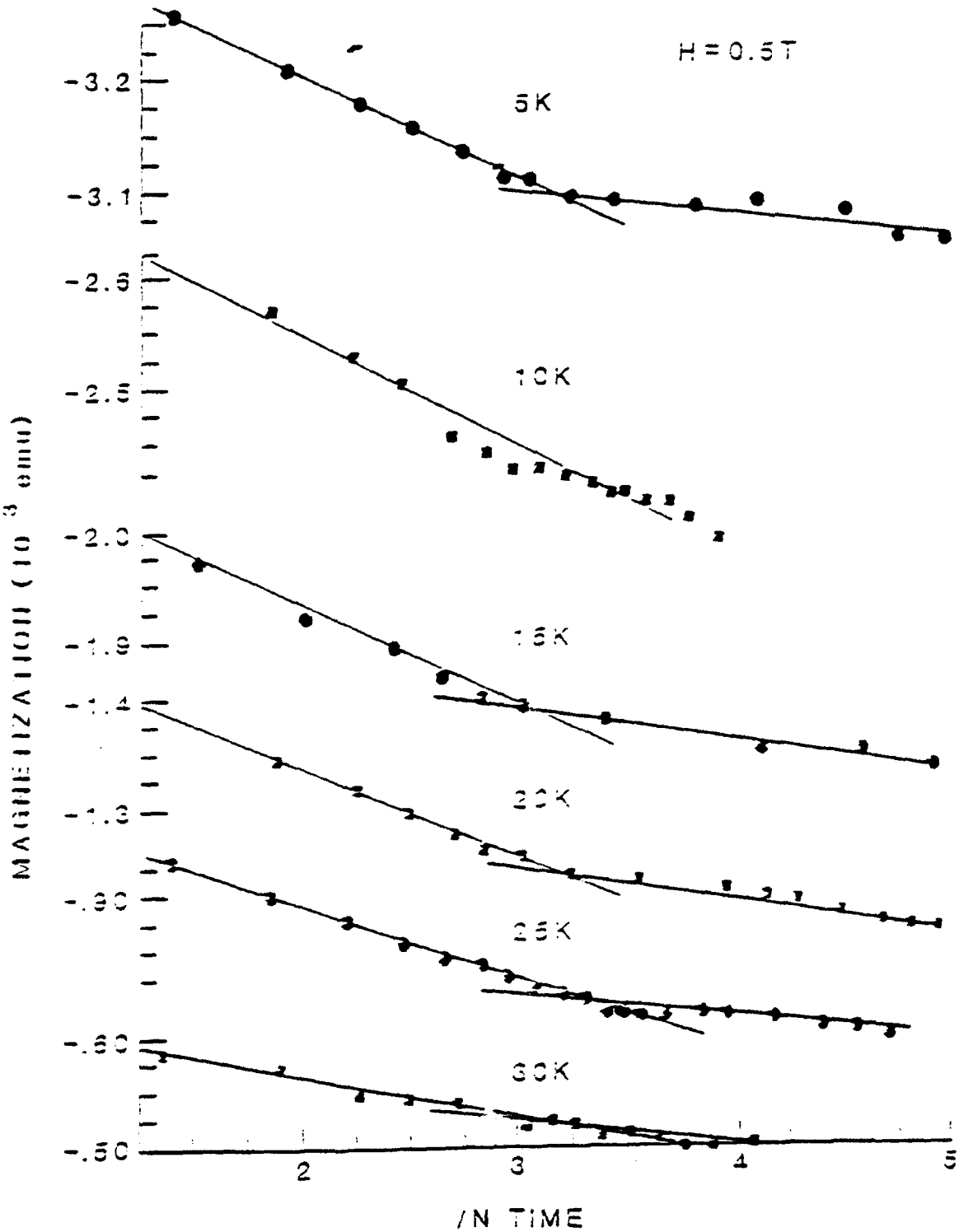


Fig 2 - Magnetization as a function of logarithmic time in minutes at selected temperatures and a field of .5T

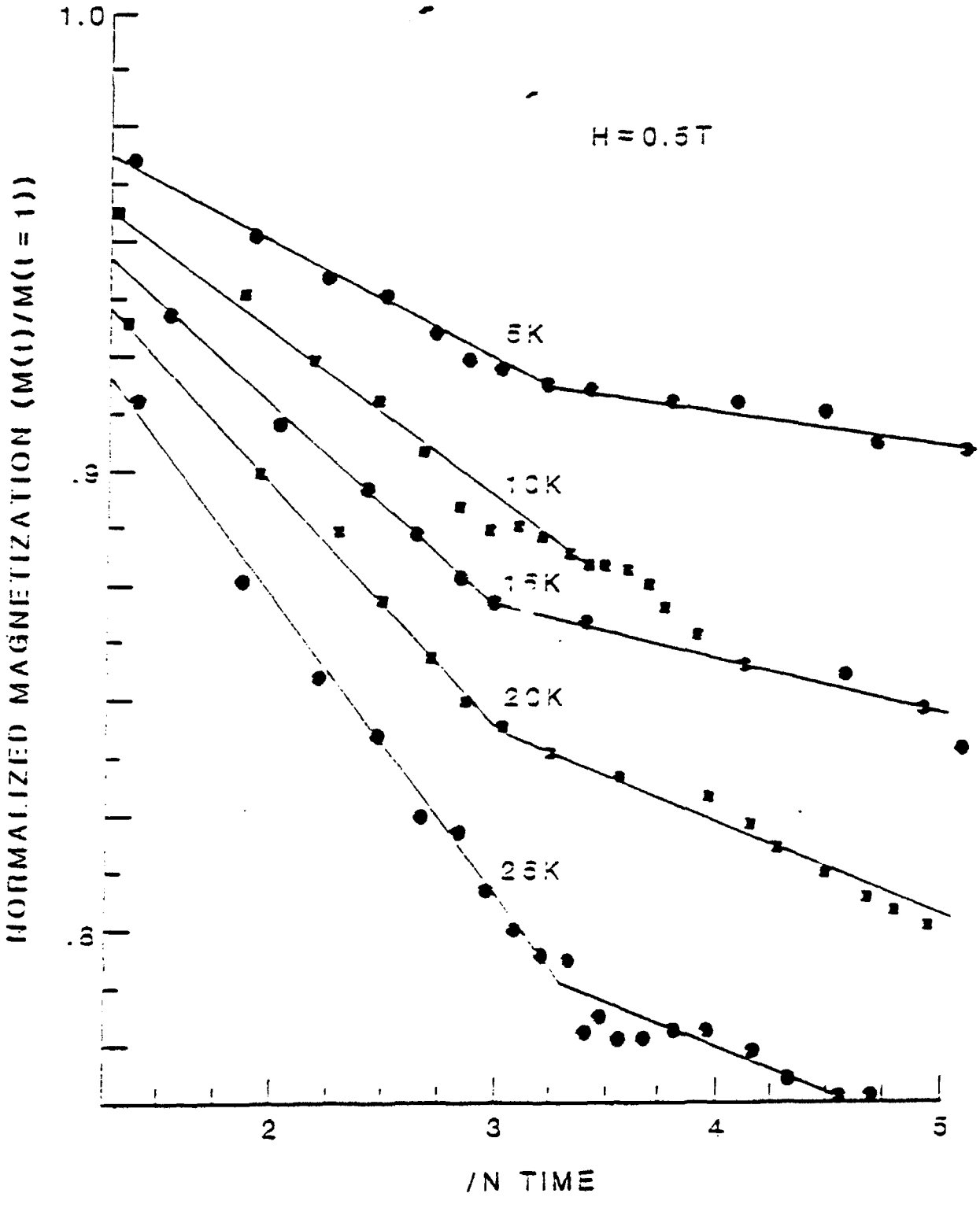


Fig 3 - Normalized magnetization as a function of logarithmic time in minutes at selected temperatures and at a field of 0.5T

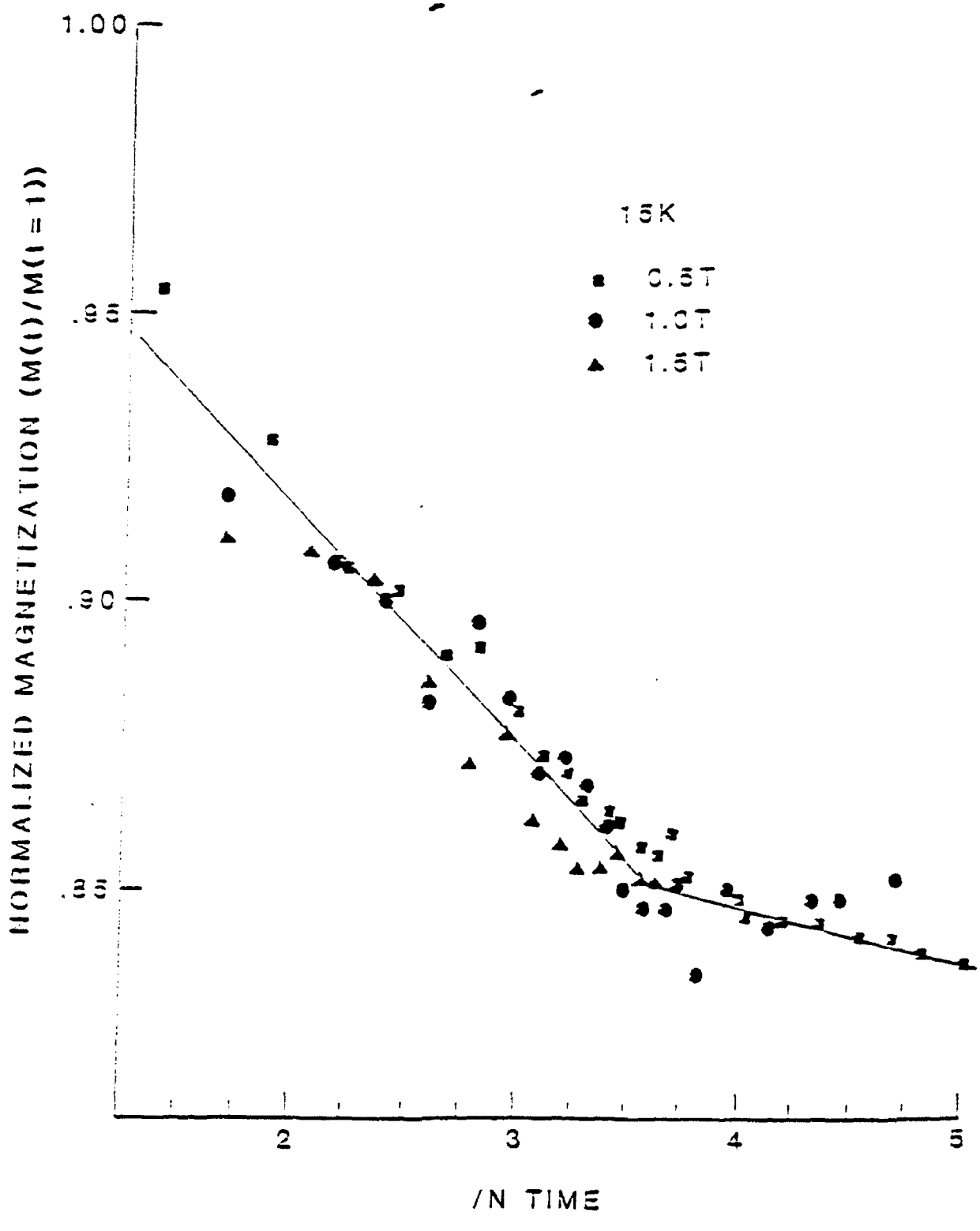


Fig 4 - Normalized magnetization as a function of logarithmic time in minutes at selected fields and at a temperature of 5K

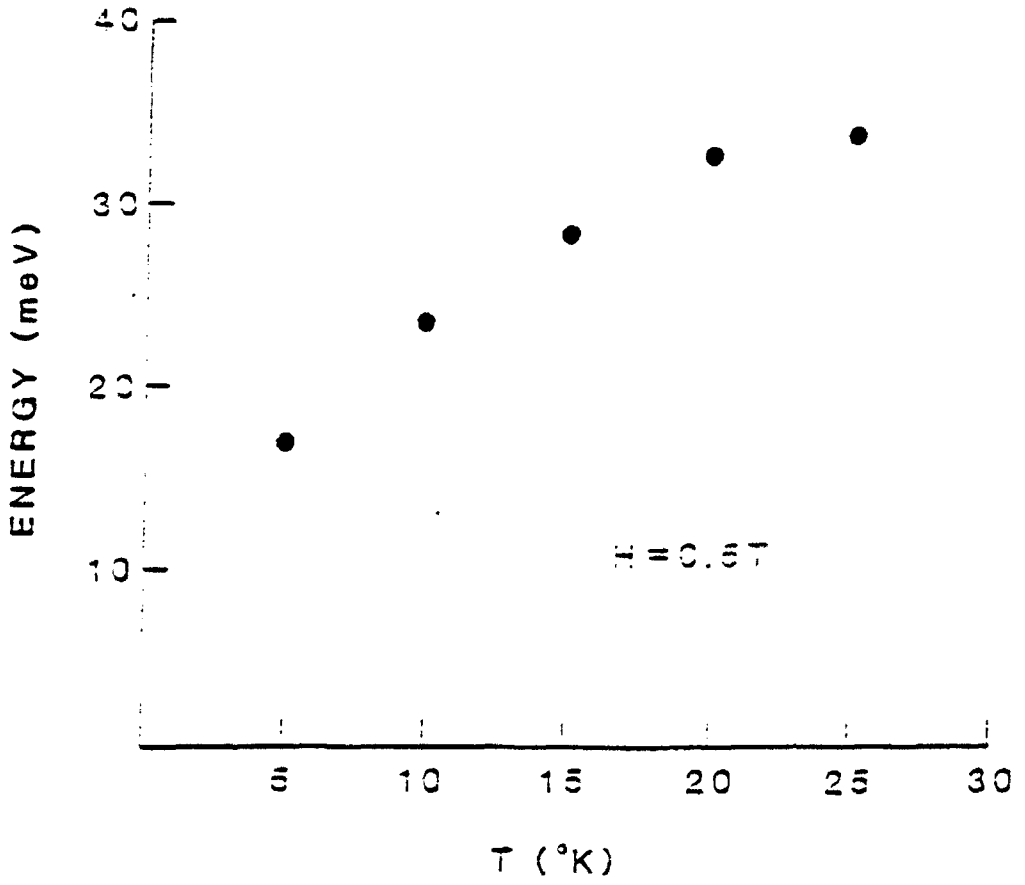


Fig 5 - Temperature dependence of the flux pinning activation energy

THIS PAGE INTENTIONALLY LEFT BLANK

Toward Implementation of a Certification Framework for Reusable Software Modules

Allen S. Parrish
Assistant Professor
Department of Computer Science

The University of Alabama
Tuscaloosa, AL 35487-0290

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Griffiss Air Force Base, Rome NY

July 1992

Toward Implementation of a Certification Framework for Reusable Software Modules

Allen S. Parrish
Assistant Professor
Department of Computer Science
The University of Alabama

Abstract

At Rome Laboratory, a "strawman" framework has been proposed for certifying reusable software. This framework consists of static analysis, including the use of metrics and formal verification, and dynamic analysis, including various testing strategies such as branch coverage and mutation testing. In this paper, we consider the application of this framework to a class of software modules called *passive modules*, which constitute a fundamental part of the vast majority of software reuse. Our emphasis is in two areas. First, we consider an expansion of the framework to include some new dynamic testing strategies, some of which are very important in achieving higher levels of certification of certain passive modules in the context of reuse. We then consider an implementation of the expanded framework, and we discuss some of the requirements of the toolset needed to support this framework.

Toward Implementation of a Certification Framework for Reusable Software Modules

Allen S. Parrish

1. Introduction

The reuse of a variety of software-related assets has been discussed as desirable for a number of years [Hooper 89, Krueger 89, McIlroy 68]. Perhaps the most well-defined form of reuse is that of actual software code. [Hooper 89] classifies reusable code into two categories: *passive* and *dynamic*. Passive reusable code can be used as building blocks in constructing complete programs. That is, passive code does not represent complete programs themselves that can be invoked by a user, but rather represents "blocks" of code that can be combined in the construction of complete programs, and reused from one program to the next. Mathematics and I/O libraries represent examples of passive reusable code.

Dynamic reusable code, on the other hand, represents complete programs that are used to generate other applications (or pieces of applications), based on parameters that are provided as input by the user. Thus, what is reused is not the code itself, but certain common techniques are applied by a single piece of code to produce a family of similar applications. The compiler-compiler YACC, which takes a grammar as input and produces a parser for programs written in that grammar, is an example of dynamic reusable code; the same fundamental approach is reused to generate a variety of parsers that differ only in the grammar they interpret. Because of the relatively high level of abstraction, dynamic code reuse is applicable only in a small set of theoretically well-defined domains [Krueger 89].

Rome Laboratory has developed a "strawman" certification framework for reusable software. It consists of five levels of increasing confidence, depending on the type and amount of certification performed. The basic idea behind this framework is to attach these certification levels to various software units within a reuse library, to give the reuser an idea of the degree of scrutiny that a unit has undergone.

The focus of this paper will be to consider the application of this framework to passive code reuse, which seems to constitute the majority of current reuse practice. There are two

fundamental approaches to passive code reuse: *module-level* reuse and *system-level* reuse. With module-level reuse, a single encapsulated unit is reused, while with system-level reuse, a group of separate modules that are interconnected in some way is reused. Since modules are present with either type of reuse, certification at the module level is important with either type of reuse as well. Thus, to lay an initial foundation for future work in this area, we wish to focus this paper on module-level certification of passive modules. Such an emphasis is directly relevant to current reuse practice within DOD reuse libraries recently examined; virtually all of the software code assets in the RAPID library and more than half of the code assets in the ASSET library are passive modules. Because of the background and interest of the author, much of the paper is devoted to testing related issues, as opposed to static analysis forms of certification.

The work reported in this paper falls into two categories. We first extend the certification framework by considering some new certification strategies that seem to be appropriate for certain types of passive modules. Such strategies are particularly important in obtaining higher levels of confidence in the context of reuse. Once we have extended the framework, we then discuss the kinds of tools that are needed to adequately implement the extended framework.

2. A Model of Passive Modules

In this section, we wish to more formally define the above notion of a "passive module." We will use Ada as a model for discussion, although the ideas here are applicable to languages with similar modularization constructs. Using Ada terminology, we allow two basic categories of passive modules: *subprograms* and *packages*. We note that other types of modules could be allowed (e.g., tasks), but we intentionally wish to keep the model as simple as possible.

Also fundamental to this work is an understanding of some of the types of relationships that can exist between these modules. Three important relationships are *contains*, *is a client of*, and *invokes*. We say that module A *contains* module B if the body of B is textually located somewhere within the body of A. The implication here is that a (re)use of A also involves a (re)use of B. We assume that (a) a subprogram *may or may not* contain (other) nested subprograms and (b) a package *must* contain one or more subprograms. We note that Ada permits a package to contain zero subprograms. However, such packages are uninteresting for (dynamic) testing purposes because they have no (dynamic) functionality to test. Also, Ada

permits packages to contain other packages. Such "nested packages" complicate our model unnecessarily, since the nested packages can be independently tested using the techniques to be discussed here. Thus, we do not explicitly consider this possibility, although it is still within the scope of the results and observations here.

We say module A *is a client of* module B if module A needs B in order to function properly. In Ada, this connection is formalized using a "with" clause: we can say that an Ada module A is a client of B if A (or some module containing A) contains a "with B" clause. Finally, we say that module A *invokes* B if it contains a statement invoking B (either through a procedure or function call). We assume this type of relationship can only exist between two subprograms. We note that B must be within A's visibility. This could be because A and B are contained within the same module, or it could be because A (or some module containing A) is a client of B (or some module containing B). As was the case with the types of modules above, we could expand the model to include other kinds of relationships (e.g., inheritance, as found in C++ and Ada 9x), but again we wish to keep the model as simple as possible, still allowing us to focus on fundamental testing issues.

These relationships allow us to fully depict the structure of typical subprograms and packages. Figure 1 illustrates the structure of a typical subprogram.

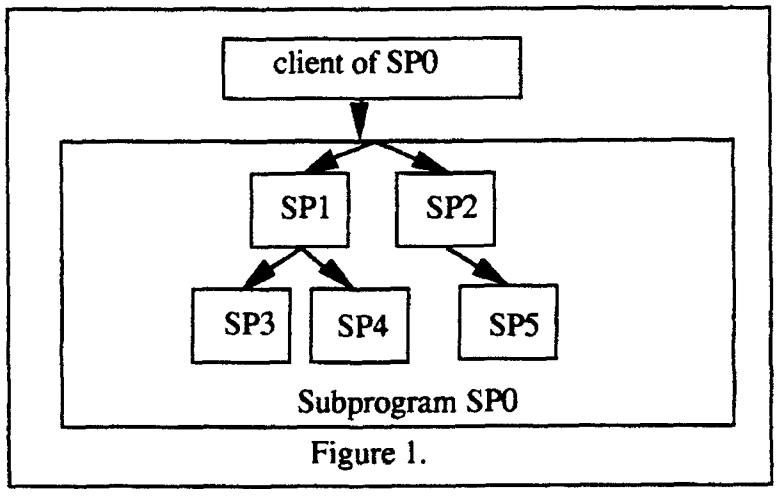


Figure 1.

Figure 1 illustrates subprogram SP0. In this example, SP0 contains other subprograms SP1-SP5; thus, the boxes for these subprograms are inside the box for SP0. However, SP0 also invokes SP1 and SP2 through procedure or function calls (hence the edges from the top of

SP0's box to SP1 and SP2), which in turn invoke the remaining subprograms contained inside SP0. Note that we have not provided any information concerning the structure of the nested subprograms; they also could have their own complex internal structures consisting of other subprograms. The topmost box, representing a "client of SP0," illustrates that external modules that become clients of SP0 can themselves invoke SP0.

In similar fashion, Figure 2 illustrates a typical package P0. In this example, package P0 contains four subprograms, SP1-SP4, that can be invoked by an outside client. (Of course, SP1-SP4 may each have a complex (hidden) structure, such as subprogram SP0 in Figure 1.) Also, subprograms may issue calls to other (independent) subprograms in the package, such as the call illustrated from SP1 to SP2.

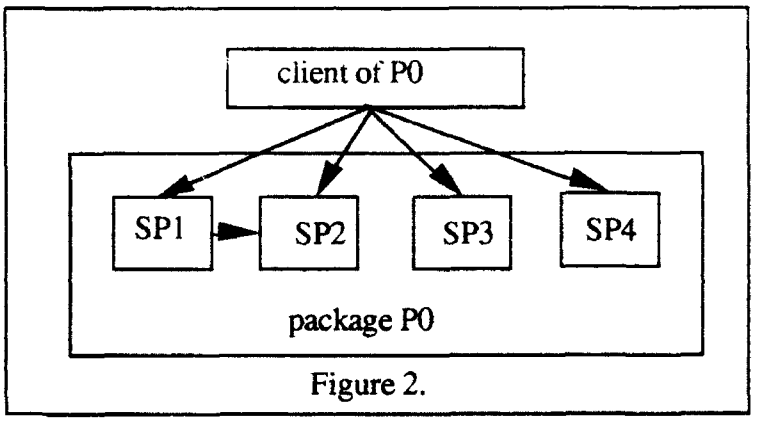


Figure 2.

Finally, we allow modules to have certain additional properties based on the capabilities that Ada provides for subprograms and packages. These properties include *generic parameters* and the exportation of *private types*. Either subprograms or packages may have generic parameters. Only packages may export private types.

With this model of passive modules in mind, we now proceed to address the objectives outlined in the introduction. Section 3 is devoted to extending the certification framework to include new certification strategies, particularly in the area of testing. Section 4 presents a top-level view of a model environment that implements the framework, and develops an idealization of how the environment might be incorporated into actual reuse practice. In keeping with the main emphasis on testing and dynamic analysis, Section 5 then focuses on the dynamic analysis toolset within the certification environment.

3. Extending the Certification Framework

The strawman certification framework developed at Rome Laboratory consists of five discrete levels, as follows:

Level 1: Functional (Black-Box) Testing
Level 2: Branch Testing
Level 3: Quality Metrics Ratings
Level 4: Mutation Testing
Level 5: Formal Verification and/or Performance Evaluation

Higher level numbers are associated with higher levels of certification. Thus, a Level 1 component has undergone very little scrutiny, while a Level 5 component has undergone substantial scrutiny. Certain elements of the framework were intentionally left vague, e.g., whether a Level *n* component must undergo certification techniques at Levels 1 to *n*-1.

There are three deficiencies in this framework that should be immediately addressed:

- There are no intermediate testing techniques between branch testing and mutation testing. Because of the expense of mutation testing, it is therefore very difficult to go beyond branch testing in terms of the level of coverage. Testing techniques that require coverage of combinations of branches should be included.

- There are no certification techniques motivated directly by issues specific to reuse. This means that it may be difficult to distinguish the framework from certification frameworks for conventional software. For this framework to have an impact as a framework for certification for *reuse*, there should exist certification techniques within the framework that are unique to *reusable* software components.

- The one-dimensionality of the framework makes it very rigid and inflexible. Using this framework, there is only one combination of techniques to achieve a particular level.

The remainder of this section addresses these issues. To address the first issue, we introduce *data flow* testing, which provides a straightforward mechanism to require testing of

combinations of branches. To address the second issue concerning the inclusion of reuse-specific techniques, we introduce a new family of testing strategies, which we call *sequence* testing. Finally, to address the third issue concerning the organization of the framework, we suggest a new underlying basis for developing the framework.

3.1 Data Flow Testing

The basic idea behind data flow testing is that paths containing definitions of variables to values, followed by uses of those definitions are executed. In this way, errors associated with a "misuse" of a variable can be detected. A variety of testing techniques based on this idea have been proposed [Rapps 84].

For our purposes here, we consider three techniques which appear to have the most promise: *definition coverage*, *use coverage*, and *DU path coverage*. We use the informal definitions for these techniques given in [Zweben 92]. To understand these techniques, we assume that a subprogram can be represented as a flowgraph, where the nodes and edges are defined as before. We then say that a variable is *defined* at a node if the node consists of a statement that assigns a value to the variable (e.g., through an input statement or the appearance of the variable on the left hand side of an assignment statement). A variable is *used* at a node if the value of the variable is referenced in the node (e.g., through an output statement, or in an expression appearing in a predicate or the right hand side of an assignment statement). A test set achieves *definition coverage* if, for every definition in the graph, it causes some subpath to be traversed from the node containing the definition to a node containing a use of the definition. That subpath must be definition-clear with respect to the variable, in the sense that it cannot redefine the variable that was defined in the initial "definition node." Also, if the node that represents a use of the definition has an out-degree greater than one, all edges emanating from that node must be traversed. That is, "predicate uses" are only considered traversed when each condition for the predicate has been exercised.

A test set achieves *use coverage* if, for every definition in the flowgraph, some definition-clear subpath from the node containing the definition to a node containing a use of that definition is traversed, for every such node containing a use of the definition. Finally, a test set satisfies *DU path coverage* if, for every definition in the flowgraph, *every* definition-clear

subpath from the node containing the definition to a node containing a use of the definition is traversed, for every such node representing a use of the definition. If there are cycles in the flowgraph, the number of such subpaths may be infinite. The standard approach to resolving this issue in the literature is to restrict the DU path coverage criterion to only require the coverage of those subpaths that contain no internal cycles (i.e., the subpaths may be such that the first and last nodes are the same, but no node in between can be the same as any other node in the subpath) [Frankl 88, Zweben 92]. However, this means that if the only path from a definition to use contains a cycle, it is not necessary to cover that definition-use combination at all, which weakens the criterion substantially. An obvious modification [Zweben 92] is to only insist that subpaths from definition to use contain minimum length internal cycles (i.e., if there are cycles, only cover the shortest one).

With these definitions of the criteria, there are some obvious relationships that exist. In particular, the data flow criteria can be ordered in terms of increasing strength: definition coverage ⇒ use coverage ⇒ DU path coverage. That is, if a test set satisfies a stronger criterion (like DU path coverage) it also satisfies any weaker one (like definition or use coverage). Moreover, use coverage (and also DU path coverage) is stronger than branch coverage. However, branch coverage and definition coverage are incomparable; neither is stronger than the other. This means that in terms of providing an intermediate level of testing between branch and mutation testing, use coverage and DU path coverage are perhaps the more likely candidates.

3.2 Sequence Testing

Sequence testing is a technique that is only applicable to packages. Consider the following example of a *stack* package:

```
generic
   type item is private
package stacks is

type stack is limited private;
procedure push(s: IN OUT stack; e: IN item);
procedure pop(s: IN OUT stack; e: OUT item);
procedure clear(s: OUT stack);

private...
end stacks;
```

This stack package has been defined with three operations: *push*, *pop* and *clear*. Sequence testing demands that various sequences of operations be tested (e.g., *push; push; pop; clear; pop; push; clear*). Since clients may interact with the package by invoking various sequences of operations, the purpose of this type of testing is to examine a (hopefully) representative subset of these sequences. Thus, this testing strategy attempts to predict scenarios under which a package might be reused. In this way, this is indeed a testing strategy that is motivated by concerns specific to reuse.

This leads to an important question: How can appropriate sequences be determined? One answer to this question, proposed in [Zweben 92], is to model potential client behavior with respect to a package as a flowgraph. In particular, the indivudal subprograms within the package (in this case, *push*, *pop* and *clear*) can be viewed as nodes in the graph. A directed edge (i.e., a "branch") from subprogram A to subprogram B represents the fact that the sequence A;B is possible in some client program. Note that this very much parallels a common way of viewing code inside an individual subprogram, where nodes might represent statements in the subprogram, and an edge between two statements represents the possibility that control might flow between the statements. Also, note that following a conditional statement (i.e., an *if* statement or a *loop* statement), there are multiple edges: one representing the case where the condition is true, and one representing the case where the condition is false. The same is true with our graph model for packages: for a boolean function exported by a package, there are two edges directed to every subprogram which can legitimately follow the function: one edge represents the case where the function returns "true," and the other edge represents the case where the function returns "false."

The criterion "branch coverage" (or "edge" coverage) requires that the sequence (or sequences) generated for a package contain every edge in the package flowgraph. Often, this will mean that every pair of operations must appear somewhere in the sequence(s), including pairs consisting of repeated invocations of the same operation, since there is normally an edge from every operation to itself. In addition, when boolean functions are involved, some pairs containing boolean functions as the first element of the pair will appear twice: once when the boolean function is true, and once when the boolean function is false. However, it is possible for the functional specifications for a package to disallow clients from invoking certain sequences of subprograms. For example, the designer of a stack package may (although not

necessarily) wish to disallow the sequence *clear;pop*, since this results in a "pop" applied to an empty stack. In this case, there is no edge between *clear* and *pop*, and it is therefore not necessary to cover this sequence with branch coverage.

Data flow analysis can also form a basis for sequence testing. The variables that form the basis for our analysis of data flow are the subprogram parameters. We say that a subprogram contains a *definition* of a parameter if the parameter is an OUT parameter. A subprogram contains a *use* of a parameter if the parameter is an IN parameter, and a parameter *of the same type* has been defined by one of the package's subprograms (from which there is a path in the package flowgraph to the subprogram containing the IN parameter). In this way, the use is of a previously occurring definition, in the same sense that uses are of previous definitions in conventional (sub)programs.

Using the above notion of a flowgraph and the definitions above of "definition" and "use," the data flow techniques defined above for simple subprograms can be defined analogously for packages. Thus, definition coverage at the package level requires covering all subprograms with OUT parameters in sequence with at least one subprogram of an IN parameter of the same type (without any intervening definitions, i.e., without being separated by another subprogram containing the same type as an OUT parameter). Use coverage at the package level requires covering all subprograms with OUT parameters in sequence with all subprograms containing IN parameters of the same type. Since many IN parameters are also OUT (i.e., IN OUT), and no intervening definitions are allowed, then it may be necessary to repeat the initial subprogram with the OUT parameter several times to cover all of these combinations. Finally, DU path coverage requires covering all subprograms with OUT parameters and every (definition-clear) path to every subprogram with an IN parameter of the same type. Note that with a boolean function, to cover the use corresponding to an IN parameter means that the boolean function must be executed when it is both true and false. This may not be possible in combination with some definitions; if so, the objective is to cover as many uses in combination with as many definitions as possible.

As a small example, consider the Ada stack package defined above with operations *push*, *pop* and *clear*. We assume that there are no restrictions imposed by the functional specification on the sequences of operations that clients may invoke within this package. Thus, the following sequence satisfies branch coverage: *clear; push; push; pop; pop; clear; clear; pop; push; clear.*

It is easy to verify that this sequence includes every pair of operations in the package. Note that since an operation is permitted to follow itself, these pairs also include repeated invocations of every operation.

To define the data flow based sequences, we must consider definitions and uses for each operation, as follows.

Operation	Definitions	Uses
Push	stack	stack, item
Pop	stack, item	stack
Clear	stack	

This leads to the following sequences of operations to achieve definition, use, and DU path coverage (although these are not the only choices):

Criterion	Sequence
Definition coverage	*Clear; Push; Pop; Push*
Use coverage	*Clear; Push; Push; Clear; Pop; Pop; Push; Pop*
DU path coverage	Same as use coverage

Some comments are in order regarding the above. First, with a package containing a subprogram that contains only IN parameters (such as a boolean function), definition coverage may not require testing that subprogram at all. This is the case when there is some other subprogram that could serve as the "use" for any corresponding definitions. Second, we note that satisfying use coverage does not ensure satisfying branch coverage. The above is an example of such a case where several branches (that were covered by branch coverage) are uncovered. Thus, use coverage should not be viewed as an augmentation of branch coverage (as is the case with conventional subprograms), but rather as a way to perhaps focus in on those branches that are perhaps important. Lastly, since every operation in this example contained a definition of "stack," there was only one *definition-clear* path from every definition to use. Thus, DU path coverage and use coverage are identical. These two criteria are different only when there are several subprograms without definitions (OUT parameters), and therefore, several different sequences that may be used to get from definitions to uses.

3.3 A New Strawman Framework

Both data flow testing and sequence testing should become part of the strawman certification framework. However, because of the one-dimensionality of the existing framework, it is difficult to augment the framework with new certification approaches; there simply doesn't seem to be any place to put new strategies. Generally speaking, the framework should be modified to become more flexible. We suggest an approach which divides the certification techniques into three components: *functional analysis, source code quality analysis* and *performance analysis.* A "score" is awarded based on the types of certification performed in each of these areas; the three overall scores are then combined using some type of relative weighting scheme to arrive at the overall certification level for the component. Figure 3 illustrates this process.

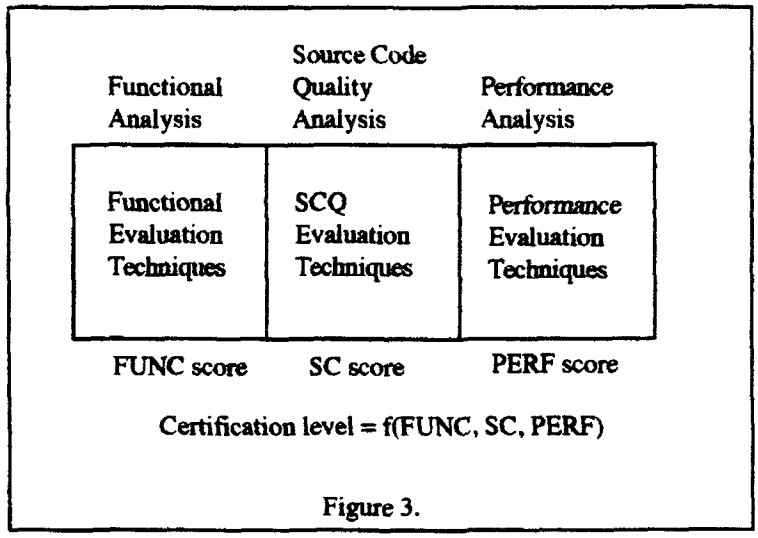


Figure 3.

Roughly speaking, performance analysis techniques could involve various types of analytical (e.g., algorithm analysis) and empirical measurements (e.g., benchmarks). Source code quality metrics could involve a number of things, including SLOC's, cohesion and coupling metrics, cyclomatic complexity, modularity measurements, etc. Functional analysis techniques could involve various testing strategies, along with formal verification. The idea behind functional analysis is an examination of the functionality of the code, to attempt to establish whether or not the code contains defects.

Given the focus of the remainder of the paper on testing, we consider a more detailed model of the functional analysis process. This model is depicted in Figure 4 below.

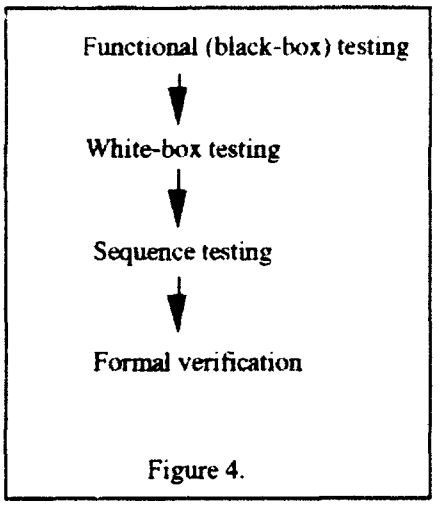


Figure 4.

This model basically says that functional testing (i.e., testing using the requirements to guide the selection of test cases) provides an initial cut at this process. White-box techniques cause the code to be exercised normally beyond functional testing, and therefore provide the next level of confidence in the functional analysis process. We make no attempt to distinguish between various forms of white-box testing (branch testing, mutation testing and data flow testing), although some distinction eventually needs to be made. Sequence testing is a higher-order form of testing, and should only be performed for a package after the individual subprograms within the package have individually undergone testing. Finally, formal verification constitutes a capstone activity that can provide the highest level of confidence for critical applications.

This progression constitutes a high-level process model for functional analysis, by roughly defining the order in which various techniques should be applied. From this model, the "score" for functional analysis should be determined, with increasing scores as one proceeds down this chain. Note that this is the same approach that was used for the original strawman certification framework. However, with the new framework, the *overall* model is now multi-dimensional, with separate "submodels" for performance analysis and source code quality analysis. Consequently, a given overall level can be achieved by more than one combination of activities from these three categories.

4. An Environment Model

Ideally, the certification framework refined in Section 3 should be implemented. In this section, we consider the overall role and scope of an automated environment that implements this framework. We envision two possible objectives for such an environment:

- Certification of new components as they are added to the library.
- Certification of existing components in a reuse library that might be of dubious or unknown quality.

The first objective can be viewed as the primary objective of a certification environment. However, the state of the practice has been such that components have already been placed in reuse libraries with little or no attempt at certification. The second objective allows this problem to be fixed.

In incorporating the certification environment into practice, it can therefore be viewed as a bidirectional gate for component flow into and out of the reuse library. Figure 5 reflects this role (where edges represent component flow).

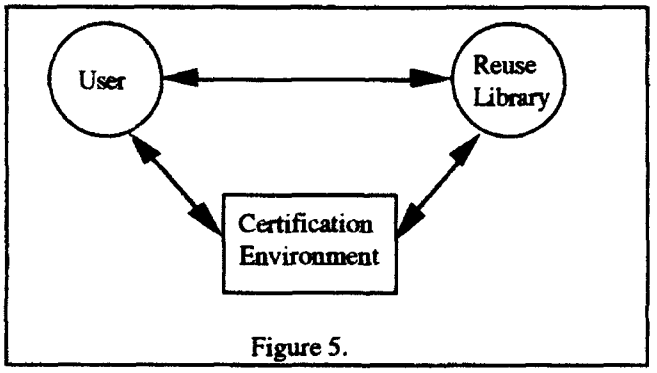


Figure 5.

To make the certification environment a viable, marketable product, it should be configurable to various user environments and reuse libraries. The environment can then stand alone as a product of interest to a variety of potential sponsors.

The basic architecture of the certification environment is organized into three parts:

- Workbench
- Static Analysis Toolset
- Dynamic Analysis Toolset

The workbench consists of those items which are needed for certification activities to take place. A possible candidate workbench of items is provided in Figure 6 below.

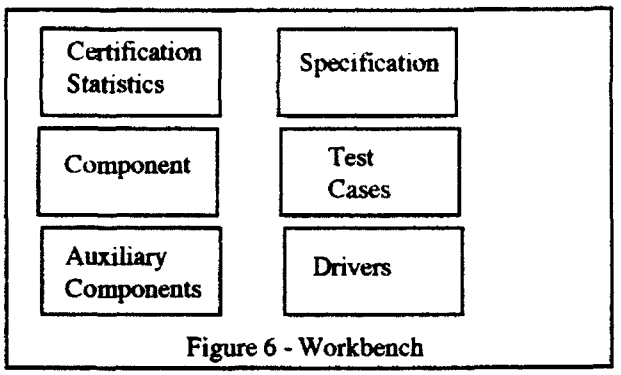


Figure 6 - Workbench

In the above diagram, *certification statistics* refers to an archive of previously performed certification activity. The box labeled *auxiliary components* refers to components that the main component depends on (e.g., via *with* clauses).

The items present on the workbench may be provided by the user, or may be obtained from the reuse library. Ideally, all of these items should be stored and linked with the component in the reuse library; however, this should not be a prerequisite for using the environment. Failure to store items such as test cases, drivers, certification statistics, etc., in the reuse library simply means that this information will not be available to a particular reuser.

The workbench items should be managed in such a fashion that the user is provided information as to what is currently present. Thus, the user should be able to obtain information regarding existing certification statistics, previously run test cases, etc., through a nice interface. The user should then be provided options based on the items present. For example,

if there are previously run test cases, the user should be given the opportunity to perform regression testing using these test cases, and compare the test results to previous test results. Of course, as the user performs new certification activities, the items on the workbench are augmented as appropriate.

The static and dynamic analysis toolsets are composed of tools needed to conduct these types of analysis. The static analysis toolset contains those statically-based metrics which have been chosen for the certification framework. Because the source code quality analysis will probably be based entirely on metrics that are applied statically, such metrics will constitute a large portion of this toolset. These metrics could include things like cohesion and coupling, cyclomatic complexity and modularity. The static analysis toolset could also include techniques to support other static activities, such as mathematical (functional) verification.

The dynamic analysis toolset will contain dynamic coverage metrics (e.g., branch coverage) needed to conduct functional analysis, as well as several auxiliary tools needed to actually conduct dynamic analysis (e.g., a tool to automatically generate drivers for passive modules which cannot themselves be executed). Because of our focus on dynamic analysis, we consider the dynamic analysis toolset in detail in Section 5.

5. Dynamic Analysis Toolset

One possible dynamic analysis toolset can be depicted as follows:

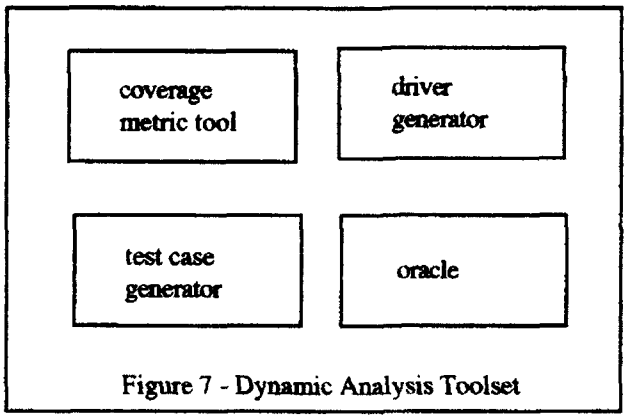


Figure 7 - Dynamic Analysis Toolset

These tools are at various stages in the R&D cycle. We assume that *coverage metric tools*

(e.g., to measure branch coverage) are generally available comercially. *Oracles*, which determine the correctness of a program on a particular test case, require the existence of formal specifications. Tools which do assertion checking, which are available commercially, satisfy this function to a limited extent. A full-blown oracle will probably have to await the introduction of formal specifications into reuse practice. The other two tools are *test case generators* and *driver generators*. Test case generators generate test cases directed toward satisfying certain coverage criteria. Driver generators generate drivers to execute subprograms that cannot be executed independently. Much of the R&D work in both of these areas has centered on programs with numeric inputs and outputs, which is extremely limiting in the face of the robust Ada type system. Our belief is that while manual development of test cases is perhaps tolerable, manual development of drivers involves labor-intensive *programming* effort that could introduce errors into the testing process. Consequently, the driver generator may be the more important of the two tools to develop first. In the remainder of this section, we therefore consider the some requirements and design issues for a driver generator to accomodate the types of issues that are important for Ada components.

A full-blown driver generator must satisfy several requirements. First, the drivers generated by such a tool should be able to handle both new test cases and regression testing over test cases that have been stored in the reuse library and pulled out to the certification environment workbench. As stated above these drivers should be able to handle complex I/O, including user-defined types that are possibly private. Also, the driver generator should permit new drivers to be built on-the-fly which instantiate a generic component in different ways.

In addition, the driver generator should be capable of generating drivers at two levels of testing. First, it should be possible to generate drivers to execute individual subprograms. Also, because Ada packages are important to reuse, the driver generator should also have the capacity to deal with these modules as well. This could simply involve generating drivers to execute individual subprograms within a package, which is conceptually no different than executing an individual subprogram in a module by itself. In addition, however, the driver generator should be capable of generating drivers to issue calls to *sequences* of subprograms within a package, in order to support the sequence testing activity introduced in Section 3.2 as part of the certification framework.

6. Conclusion and Future Work

In this paper, we have proposed high-level, strawman versions of both a certification framework and environment for reusable software components. We envision the primary goal for this research as obtaining a minimal framework and environment prototype that can later be expanded. The basic static and dynamic analysis toolsets to support such a prototype are probably commercially available. However, some additional work is needed, as follows:

- Completion of a rudimentary certification framework which includes standard certification techniques;

- Completion of the basic environment interface and architecture;

- Completion of the driver generator (proposed university research effort), which we have argued is the most fundamental missing link from commercially available dynamic analysis tools;

- New reuse-oriented certification tools are needed to augment the static and dynamic analysis toolsets to give the environment more of a "reuse-based" identity.

One important area of future work not considered here is the extension of the ideas to reusable "systems" of components (i.e., collections of packages and subprograms that work together in some way). These types of unit-oriented certification approaches would certainly be useful for the indivudal modules within such systems. However, it is likely the case that additional certification techniques and tools are needed at the system level. In addition, work is needed to consider certification issues for other types of individual modules such as concurrent units.

References

[ASSET 92] "ASSET Library Repository Catalog," SAIC Corp, April 1992.

[Brown 89] Brown, D. "The Development of a Program Analysis Environment for Ada," Auburn University Technical Report, Department of Computer Science and Engineering, 1989.

[GRC 88] Begley, A., et al. "Ada Test and Verification System (ATVS): Software Requirements Specification," General Research Corp., 1988.

[GRC 92] AdaQuest version 1.1 product flyer, General Research Corporation.

[Hooper 89] Hooper, J. and R. Chester. "Software Reuse Guidelines," Technical Report ASQB-GI-90-015, U.S. Army Institute for Research in Management Information, Communications and Computer Sciences (AIRMICS).

[Krueger 89] Krueger, C. "Models of Reuse in Software Engineering," Technical Report CMU-CS-89-188, School of Computer Science, Carnegie-Mellon University.

[McIlroy 68] McIlroy, M.D. "Mass Produced Software Components," in *Software Engineering: Report on a Conference by the NATO Science Committee*, P. Naur and B. Randell eds., pp. 138-150.

[Parrish 92] Parrish, A., D. Cordes and E. Kortright, "A Framework for Testing Data-Oriented Software Modules," unpublished manuscript.

[Softech 91] "Catalog of Reusable Software Components," RAPID Center Library, 1991.

[Weide 91] Weide, B., S. Zweben and W. Ogden, "Reusable Software Components," in *Advances in Computers*, vol. 33, pp. 1-65.

[Zweben 92] Zweben, S.H., Heym, W. and J. Kimmich, "Systematic Testing of Data Abstractions Based on Software Specifications," to appear in the *Journal of Software Testing, Verification and Reliability*, 1992.

DATA ASSOCIATION PROBLEMS IN MULTISENSOR DATA FUSION AND MULTITARGET TRACKING

by

Aubrey B. Poore
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Boling Air Force Base, Washington, D.C.

September, 1992

DATA ASSOCIATION PROBLEMS IN MULTISENSOR DATA FUSION AND MULTITARGET TRACKING

Aubrey B. Poore
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523

ABSTRACT

The ever-increasing demand in surveillance is to produce highly accurate target and track identification and estimation in real-time, even for dense target scenarios and in regions of high track contention. The use of multiple sensors, through more varied information, has the potential to greatly enhance target identification and state estimation. For multitarget tracking, the processing of multiple scans all at once yields high track identification. However, to achieve this accurate state estimation and track identification, one must solve an NP-hard data association problem in real-time. The only known method for solving this central problem in both multisensor data fusion and multitarget tracking optimally is branch and bound. Herein lies the limitation and difficulty. Data association algorithms based on branch and bound, e.g., multiple hypothesis tracking (MHT), are inevitably faulty in dense scenarios and in regions of high track contention. These failures are not graceful. With the emergence of fast (real-time), near-optimal, and Lagrangian relaxation based algorithms for multidimensional assignment problems, the opportunity now exists to greatly enlarge the number of complex problems solvable in real-time. Thus, in this work a general class of data association problems is formulated as a multidimensional assignment problem. Since the most popular futuristic method for solving difficult data association problems is MHT, the equivalence between solving data association problems by MHT and by multidimensional assignment problems is established. Track initiation and track maintenance using an N-scan moving window are then used as illustrations. MHT is also permeates multisensor data fusion, and thus two classes of simple problems are formulated as multidimensional assignment problems - collocated sensors and noncollocated sensors.

1. Introduction

The ever-increasing demand in surveillance is to produce highly accurate target and track identification and estimation in real-time, even for dense target scenarios and in regions of high track contention. The use of multiple sensors, through varied information, has the potential to greatly enhance target identification and state estimation. For multitarget tracking, the processing of multiple scans all at once yields high track identification. However, one must now solve an NP-hard (NP stands for non deterministic polynomial) data association problem in real-time. The only know method to solve these combinatorial optimization problems optimally is branch and bound, which is essentially the method used in multiple hypothesis tracking (MHT). (Depending on the pruning logic, current techniques in MHT range from explicit enumeration to implicit enumeration using what is essentially a branch and bound method.) Herein lies the fundamental limitation. In dense scenarios and high track contention, the time required to solve the problem optimally can grow exponentially with the size of the problem. This failure is not graceful, i.e., the method is not robust with respect to real-time needs. The emergence of fast (real-time) near optimal algorithms for multidimensional assignment problems is intended to address this problem. Although the Lagrangian relaxation based algorithms are discussed briefly in Section 5 and extensively in our previous work [23-28], the primary objective in this work is to formulate a general class of data association problems for multisensor data fusion and multitarget tracking as multidimensional assignment problems.

Before outlining this paper, it is important to keep in mind the large number of additional issues that one must face in designing algorithms for a tracking system. The surveillance mode can range fro.n wide area with moderate to benign background clutter to tactical with a few targets in high clutter. Targets may range in size from large to small; target density, from high to low; a target may be stealthy or non stealthy; targets may initiate or terminate within the surveillance region; and, the target kinematics can range from simple to highly maneuverable. False reports may arise from a number of sources, so that the clutter environment must be defined. Generally, clutter is characterized as discrete or diffuse (homogeneous) and then further subdivided as stationary or non stationary. One must deal with buildings, mountains, forest, rain, birds, and moving vehicles on the ground. For multiple sensors, one must consider the geometric configuration (e.g., collocated vs. non collocated), passive versus active, and registration errors. Finally, one must also contend with asynchronous measurements, platform motion, and error bias. Thus the data association problem is one of many issues, but it is considered to be the central problem.

The outline of the paper is as follows. The starting point is a generalization of Morefield's work [20] followed by a refinement of this generalization. This development in Section 2 shows precisely the conditions under which multidimensional assignment problems are applicable to data association. Utilizing the results of this section, data association problems solved by MHT in tracking are formulated as assignment problems in Section 3. This development includes track initiation and track maintenance (continuation) using an N-scan sliding window. (The corresponding algorithms have been implemented and extensively tested by

Poore and Rijavec [23-28].) MHT is also used extensively in multisensor data fusion. For example, Blackman [6,7] believes it to be the method of the future and demonstrates its potential use in multisensor fusion, and argues very effectively that it should be an attractive method for dim moving targets. MHT is also at the core of the work of Chong, Mori, and Chang [8] on distributed multisensor multitarget tracking. Although we do not examine each of these problem areas, we do present two example problems in Section 4: non collocated sensors investigated by Deb, Pattipati, and Bar-Shalom [9] and collocated sensors on multiple platforms.

§2. Assignment Formulation of Some General Data Association Problems

Let $Z(k)$ denote a data set of M_k actual reports plus a dummy variable used to denote a missing report, and let Z^N denote the cumulative data set of N such sets defined by

$$Z(k) = \{z_i^k\}_{i=0}^{M_k} \quad \text{and} \quad Z^N = \{Z(1), \ldots, Z(N)\}, \tag{2.1}$$

respectively. Here, the index $i_k = 0$ is reserved for the missing report, so that z_0^k has no physical dimensions. One might think of the cumulative data set Z^N as a matrix with the k^{th} column being the data set $Z(k)$. As such this represents one partitioning of the data. In the course of formulating the data association problem, each data set $Z(k)$ is reordered and the cumulative data set Z^N repartitioned along the rows of this matrix to some benefit. The objective in this section is to formulate a reasonably general class of multidimensional assignment problems for multisensor data fusion and multitarget target tracking, starting from Morefield's work [20]. Variations on the formulation are discussed at the end of this section.

Before beginning the development, some observations about the data sets $Z(k)$ are in order. In multisensor data fusion and multitarget tracking the data sets $Z(k)$ may represent different classes of objects. For example, in centralized fusion [7,31] the objects may all be measurements that represent targets or false alarms. In sensor level fusion the objects may all be tracks. For track initiation in multitarget tracking the objects are observations that must be partitioned into tracks and false alarms. In track maintenance one data set will be tracks and remaining data sets will be observations which are assigned to the existing tracks, false reports, or initiating tracks.

A generalization of Morefield's definition [20] to include missing reports is that a partition $\bar{\gamma}$ of the cumulative data Z^N and the collection $\bar{\Gamma}$ of all such partitions are defined by

$$\bar{\gamma} = \{\bar{\gamma}_1, \ldots, \bar{\gamma}_{n(\bar{\gamma})}\} \tag{2.2a}$$

$$\bar{\gamma}_i \cap \bar{\gamma}_j \subset \{z_0^1, \ldots z_0^N\} \text{ for } i \neq j \tag{2.2b}$$

$$\bar{\gamma}_j \neq \{z_0^1, \ldots, z_0^N\} \text{ for any } j = 1, \ldots, n(\bar{\gamma}) \tag{2.2c}$$

$$Z^N = \{z_0^1, \ldots, z_0^N\} \cup \left[\cup_{j=1}^{n(\bar{\gamma})} \bar{\gamma}_j\right] \tag{2.2d}$$

$$\gamma_i \cap Z(k) \neq \emptyset \text{ for any } j \text{ and } k \tag{2.2e}$$

$$\bar{\Gamma} = \{\bar{\gamma} \mid \bar{\gamma} \text{ satisfies } (2.2a) - (2.2e)\} \tag{2.2f}$$

Note that (2.1b) implies $\bar{\gamma}_i \cap \bar{\gamma}_j$ is either empty or contains one or more missed reports in common when $i \neq j$. Property (2.1c) implies the exclusion of $\{z_0^1, \ldots, z_0^N\}$ from each partition. This is more or less for notational convenience, but one can rationalize this choice by observing the implication - an object is assumed to be detected in at least one of the data sets $Z(k)$. For target tracking or sensor fusion one then chooses the partition $\bar{\gamma} \in \bar{\Gamma}$ which maximizes the posteriori probability $P(\gamma|Z^N)$, i.e., by finding the maximizing partition $\bar{\gamma}$ of the problem

$$\text{Maximize} \left\{ \frac{P(\bar{\gamma}|Z^N)}{P(\bar{\gamma}^0|Z^N)} \mid \bar{\gamma} \in \bar{\Gamma} \right\} \tag{2.3}$$

wherein the partition $\bar{\gamma}^0$ is a reference partition, e.g., the partition consisting of all false reports. (The use of this normalizing constant $P(\bar{\gamma}^0|Z^N)$ does not change the maximizing partition, but does have advantages discussed at the end of this section.) Let $\Upsilon(\bar{\gamma})$ denote the event that $\{\bar{\gamma}\}$ is true. Now $P(\bar{\gamma}|Z^N)$ is developed using Bayesian estimation or as a special case, one frequently assumes no a priori knowledge of the $P(\Upsilon(\bar{\gamma}))$ in which case $P(\Upsilon(\bar{\gamma})) = P(\Upsilon(\bar{\gamma}^0))$ for all partitions $\bar{\gamma} \in \bar{\Gamma}$. One may continue this formulation, add the independence assumptions similar to equations (2.10) and (2.11), and convert this problem (2.3) to a set packing problem as in the work of Morefield [20]; however, the goal here will be to refine the definition of a partition (2.2) in a way that is amenable to the assignment problem.

A *track of reports* $Z_{i_1 \cdots i_N}$ is defined by

$$Z_{i_1 \cdots i_N} = \{z_{i_1}^1, \ldots, z_{i_k}^N\}, \tag{2.4}$$

wherein exactly one report $z_{i_k}^k$ is included from the data set $Z(k)$ for each $k = 1, \ldots, N$, and will be used in the sequel. A *feasible partition* γ of the data set Z^N is defined by

$$\gamma = \{\gamma_1, \ldots, \gamma_{n(\gamma)}\} \tag{2.5a}$$

$$\gamma_i \cap \gamma_j \subset Z_{0 \cdots 0} \equiv \{z_0^1, \ldots, z_0^N\} \text{ for } i \neq j \tag{2.5b}$$

$$Z^N = \{z_0^1, \ldots, z_0^N\} \cup \left[\cup_{j=1}^{n(\gamma)} \gamma_j \right] \tag{2.5c}$$

$$\gamma_j \neq \{z_0^1, \ldots, z_0^N\} \text{ for any } j = 1, \ldots, n(\gamma) \tag{2.5d}$$

$$\text{Each } \gamma_j = Z_{i_1 \cdots i_N} \text{ for exactly one N-tuple } (i_1, \ldots, i_N) \tag{2.5e}$$

$$\Gamma = \{\gamma \mid \gamma \text{ satisfies (2.5}a\text{)} - \text{(2.5}e\text{)}\} \tag{2.5f}$$

A false report $z_{i_k}^k$ *is included in this definition and is denoted by* $Z_{0 \cdots 0 i_k 0 \cdots 0}$. Definitions (2.4) and (2.5) allow one to identify the elements $\gamma_i \in \gamma$ with N-tuples and redefine a feasible partition γ as follows

$$\gamma = \{Z_{i_1 i_2 \cdots i_N} \mid Z_{i_1 i_2 \cdots i_N} \text{ satisfies (2.6}b\text{)} - \text{(2.6}c\text{) for } i_k = 0, \ldots, M_k; k = 1, \ldots, N\} \tag{2.6a}$$

$$Z_{i_1 \cdots i_N} \cap Z_{j_1 \cdots j_N} \subset Z_{00 \cdots 0} \text{ unless } (i_1, \ldots, i_N) = (j_1, \ldots, j_N) \tag{2.6b}$$

$$Z^N = Z_{0 \cdots 0} \cup \left[\cup_{(i_1, i_2, \ldots, i_N) \in \gamma} Z_{i_1 \cdots i_N} \right] \tag{2.6c}$$

$$Z_{0 \cdots 0} \notin \gamma, \tag{2.6d}$$

$$\Gamma = \{\gamma \mid \gamma \text{ satisfies (2.6}a\text{)} - \text{(2.6}d\text{)}\} \tag{2.6e}$$

where the abbreviated notation $(i_1, i_2, \ldots, i_N) \in \gamma$ *means* $Z_{i_1 i_2 \cdots i_N} \in \gamma$. Note that (2.6b) and (2.6c) are valid if and only if each actual report $z^k_{i_k}$, i.e., $i_k \geq 1$ for any $k = 1, \ldots, N$, belongs to exactly one track of reports $Z_{i_1 \cdots i_N} \in \gamma$. The use of the 0-1 variable

$$z_{i_1 \cdots i_N} = \begin{cases} 1, & \text{if } Z_{i_1 \cdots i_N} \in \gamma \ ; \\ 0, & \text{otherwise;} \end{cases} \tag{2.7}$$

yields an equivalent characterization of a feasible partition (2.6) as a solution of the equations

$$\sum_{(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_N) = (0, \ldots, 0)}^{(M_1, \ldots, M_{k-1}, M_{k+1}, \ldots, M_N)} z_{i_1 \cdots i_N} = 1 \text{ for } i_k = 1, \cdots, M_k \text{ and } k = 1, \cdots, N. \tag{2.8}$$

The optimization problem (2.3) is now expressed as

$$\text{Maximize} \left\{ \frac{P(\gamma | Z^N)}{P(\gamma^0 | Z^N)} \mid \gamma \in \Gamma \right\} \tag{2.9}$$

where the definition of the feasible partition is given by (2.6) (or (2.5)) or equivalently by (2.7) and (2.8). The reference partition γ^0 of all false reports is

$$\gamma^0 = \{ Z_{0 \cdots i_k \cdots 0} \mid i_k = 1, \ldots, M_k ; k = 1, \ldots, N \}.$$

Recalling that $P(\gamma | Z^N) = p(Z^N | \Upsilon(\gamma)) P(\Upsilon(\gamma)) / P(Z^N)$, we now introduce the following key independence assumptions that allow one to convert (2.9) to a multidimensional assignment problem:

$$p(Z^N | \Upsilon(\gamma)) = \prod_{(i_1, \cdots, i_N) \in \gamma} p(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) \tag{2.10a}$$

$$P(\Upsilon(\gamma)) = G(Z^N) \prod_{(i_1, \cdots, i_N) \in \gamma} g(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) \tag{2.10b}$$

where $G(Z^N)$ may on the data set Z^N but is independent of any particular partition. Thus

$$P(\gamma | Z^N) = \frac{G(Z^N)}{P(Z^N)} \prod_{(i_1, \cdots, i_N) \in \gamma} p(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) g(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) \tag{2.10c}$$

Next, *if* $Z_{i_1 i_2 \cdots i_N}$ occurs in two distinct feasible partitions, say γ and $\omega \in \Gamma$, then assume

$$p(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) g(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) = p(Z_{i_1 \cdots i_N} | \Upsilon(\omega)) g(Z_{i_1 \cdots i_N} | \Upsilon(\omega)) \tag{2.11}$$

With these two assumptions, one can now formulate the assignment problem in three equivalent ways.

For the *first derivation*, observe that

$$\frac{P(\gamma | Z^N)}{P(\gamma^0 | Z^N)} \equiv L_\gamma \equiv \prod_{(i_1, \ldots, i_N) \in \gamma} L_{i_1 \cdots i_N} \tag{2.12a}$$

where

$$L_{i_1 \cdots i_N} = \frac{p(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) g(Z_{i_1 \cdots i_N} | \Upsilon(\gamma))}{\prod_{k=1, i_k \neq 0}^{N} p(Z_{0 \cdots 0 i_k 0 \cdots 0} | \Upsilon(\gamma^0)) g(Z_{0 \cdots 0 i_k 0 \cdots 0} | \Upsilon(\gamma^0))}. \tag{2.12b}$$

Here the index i_k in the denominator corresponds to the k^{th} index of $Z_{i_1\cdots i_N}$ in the numerator. Next define

$$c_{i_1\cdots i_N} = -\ln L_{i_1\cdots i_N}, \tag{2.13}$$

so that

$$-\ln\left[\frac{P(\gamma|Z^N)}{P(\gamma^0|Z^N)}\right] = \sum_{(i_1,\ldots,i_N)\in\gamma} c_{i_1\cdots i_N}. \tag{2.14}$$

Thus in view of the of the characterization of a feasible partition as a solution of the equations (2.7) and (2.8) and the independence assumptions (2.10) and (2.11), problem (2.9) is equivalent to the following N-dimensional assignment problem:

$$\begin{aligned}
&\text{Minimize} && \sum_{i_1=0}^{M_1}\cdots\sum_{i_N=0}^{M_N} c_{i_1\cdots i_N} z_{i_1\cdots i_N} \\
&\text{Subject To:} && \sum_{i_2=0}^{M_2}\cdots\sum_{i_N=0}^{M_N} z_{i_1\cdots i_N} = 1, \quad i_1 = 1,\ldots,M_1, \\
& && \sum_{i_1=0}^{M_1}\cdots\sum_{i_{k-1}=0}^{M_{k-1}}\sum_{i_{k+1}=0}^{M_{k+1}}\cdots\sum_{i_N=0}^{M_N} z_{i_1\cdots i_N} = 1, \\
& && \qquad \text{for} \quad i_k = 1,\ldots,M_k \text{ and } k = 2,\ldots,N-1, \\
& && \sum_{i_1=0}^{M_1}\cdots\sum_{i_{N-1}=0}^{M_{N-1}} z_{i_1\cdots i_N} = 1, \quad i_N = 1,\ldots,M_N \\
& && z_{i_1\cdots i_N} \in \{0,1\} \text{ for all } i_1,\ldots,i_N.
\end{aligned} \tag{2.15}$$

Note that in this formulation $c_{0\cdots 0 i_k 0\cdots 0} = 0$ for all $i_k = 0,1,\ldots,M_k$ and $k = 1,\ldots,N$ due to the likelihood ratio in (2.12b). Although $c_{0\cdots 0} = 0$ by assumption, it is worth observing that $c_{0\cdots 0} > 0$ (< 0) implies $z_{0\cdots 0} = 0$ (1, respectively), i.e., other than being a 0-1 variable, $z_{0\cdots 0}$ does not enter the constraint equations and is determined directly.

The *second derivation* begins with the following decompositions:

$$-\ln P(\gamma|Z^N) = \sum_{(i_1,\ldots,i_N)\in\gamma} \hat{c}_{i_1\cdots i_N} - \ln(G(Z^N)/P(Z^N)) \tag{2.16a}$$

where

$$\hat{c}_{i_1\cdots i_N} = -\ln\left[p(Z_{i_1\cdots i_N}|\Upsilon(\gamma))g(Z_{i_1\cdots i_N}|\Upsilon(\gamma))\right] \tag{2.16b}$$

and

$$-\ln P(\gamma^0|Z^N) = \sum_{k=1}^{N}\sum_{i_k=0}^{M_k} b_{0\cdots 0 i_k 0\cdots 0} - \ln(G(Z^N)/P(Z^N)) \tag{2.17a}$$

where

$$b_{0\cdots 0 i_k 0\cdots 0} = -\ln\left[p(Z_{0\cdots 0 i_k 0\cdots 0}|\Upsilon(\gamma^0))g(Z_{0\cdots 0 i_k 0\cdots 0})|\Upsilon(\gamma^0))\right]. \tag{2.17b}$$

(By an earlier convention, $b_{0\cdots0} = c_{0\cdots0} = 0$.) Thus

$$
\begin{aligned}
-\ln\left[\frac{P(\gamma|Z^N)}{P(\gamma^0|Z^N)}\right] &= \sum_{(i_1,\ldots,i_N)\in\gamma} \hat{c}_{i_1\cdots i_N} - \sum_{k=1}^{N}\sum_{i_k=0}^{M_k} b_{0\cdots0i_k0\cdots0} \\
&= \sum_{(i_1,\ldots,i_N)\in\gamma} \left(\hat{c}_{i_1\cdots i_N} - \sum_{k=1}^{N} b_{0\cdots0i_k0\cdots0}\right)
\end{aligned}
\tag{2.18}
$$

where the last equality is valid because each nonzero index i_k arises in exactly one $(i_1,\ldots,i_N) \in \gamma$. Now with the identification

$$
c_{i_1\cdots i_N} = \hat{c}_{i_1\cdots i_N} - \sum_{k=1}^{N} b_{0\cdots0i_k0\cdots0}, \tag{2.19}
$$

the formulation of the multidimensional assignment problem (2.15) follows exactly as in that of the first derivation.

The final derivation is based on the following invariance property of the multidimensional assignment problem (2.15).

Invariance Property. *Let $N > 1$, $M_k > 0$ for $k = 1,\ldots,N$, and assume $\hat{c}_{0\cdots0} = 0$. Then the minimizing solution of the following multidimensional assignment problem is independent of any choice of $b_{0\cdots0i_k0\cdots0}$ for $i_k = 1,\ldots,M_k$ and $k = 1,\ldots,N$ provided $b_{0\cdots0} = 0$.*

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{i_1=0}^{M_1}\cdots\sum_{i_N=0}^{M_N}\left(\hat{c}_{i_1\cdots i_N} - \sum_{k=1}^{N} b_{0\cdots0i_k0\cdots0}\right) z_{i_1\cdots i_N} \\
\text{Subject To:} \quad & \sum_{i_2=0}^{M_2}\cdots\sum_{i_N=0}^{M_N} z_{i_1\cdots i_N} = 1, \quad i_1 = 1,\ldots,M_1, \\
& \sum_{i_1=0}^{M_1}\cdots\sum_{i_{k-1}=0}^{M_{k-1}}\sum_{i_{k+1}=0}^{M_{k+1}}\cdots\sum_{i_N=0}^{M_N} z_{i_1\cdots i_N} = 1, \\
& \qquad \text{for} \quad i_k = 1,\ldots,M_k \text{ and } k = 2,\ldots,N-1, \\
& \sum_{i_1=0}^{M_1}\cdots\sum_{i_{N-1}=0}^{M_{N-1}} z_{i_1\cdots i_N} = 1, \quad i_N = 1,\ldots,M_N \\
& z_{i_1\cdots i_N} \in \{0,1\} \text{ for all } i_1,\ldots,i_N.
\end{aligned}
\tag{2.20}
$$

Proof: Let A(z) and B(z) denote the objective function in (2.20) with b removed and b present, respectively. Let $\hat{z}$ and $\hat{y}$ both be feasible solutions of the constraints in (2.20). It suffices to show that $A(\hat{z}) \leq A(\hat{y})$ if

and only if $B(\hat{z}) \leq B(\hat{y})$. To see this, observe

$$
\begin{aligned}
B(\hat{z}) &= \sum_{i_1=0}^{M_1} \cdots \sum_{i_N=0}^{M_N} \Big(c_{i_1 \cdots i_N} - \sum_{k=1}^{N} b_{0 \cdots 0 i_k 0 \cdots 0} \Big) z_{i_1 \cdots i_N} \\
&= \sum_{i_1=0}^{M_1} \cdots \sum_{i_N=0}^{M_N} \hat{c}_{i_1 \cdots i_N} z_{i_1 \cdots i_N} - \sum_{k=1}^{N} \sum_{i_1=0}^{M_1} \cdots \sum_{i_N=0}^{M_N} b_{0 \cdots 0 i_k 0 \cdots 0} z_{i_1 \cdots i_N} \\
&= A(\hat{z}) - \sum_{k=1}^{N} \sum_{i_1=0}^{M_1} \cdots \sum_{i_k=1}^{M_k} \cdots \sum_{i_N=0}^{M_N} b_{0 \cdots 0 i_k 0 \cdots 0} z_{i_1 \cdots i_N} \\
&= A(\hat{z}) - \sum_{k=1}^{N} \sum_{i_k=1}^{M_k} \Big(\sum_{i_1=0}^{M_1} \cdots \sum_{i_{k-1}=1}^{M_{k-1}} \sum_{i_{k+1}=1}^{M_{k+1}} \cdots \sum_{i_N=0}^{M_N} b_{0 \cdots 0 i_k 0 \cdots 0} z_{i_1 \cdots i_N} \Big) \\
&= A(\hat{z}) - \sum_{k=1}^{N} \sum_{i_k=0}^{M_k} b_{0 \cdots 0 i_k 0 \cdots 0}.
\end{aligned}
$$

Similarly, $B(\hat{y}) = A(\hat{y}) - \sum_{k=1}^{N} \sum_{i_k=0}^{M_k} b_{0 \cdots 0 i_k 0 \cdots 0}$. Thus, $B(\hat{z}) - B(\hat{y}) = A(\hat{z}) - A(\hat{y})$, which implies the result. Q.E.D.

The *third derivation* begins with the expression (2.16a) for $P(\gamma|Z^N)$. Since $-\ln(G(Z^N)/P(Z^N))$ in (2.16a) is, by assumption, independent of any partition, the multidimensional assignment problem is posed just as in (2.15) except with c replaced by $\hat{c}$. The above Invariance Property is then used to zero out all cost coefficients of the form $\hat{c}_{0 \cdots i_k \cdots 0}$ by setting $b_{0 \cdots i_k \cdots 0} = \hat{c}_{0 \cdots i_k \cdots 0}$. Examples of the use of this third derivation can be found in the book of Blackman [6].

Several final remarks are in order. The first definition of a partition (2.2) implies that each actual report belongs to exactly one subset $\bar{\gamma}_j$. One can modify this to allow many assignments of one, some, or all the actual reports. The assignment problem (2.15) is changed accordingly. For example, if $z_{i_k}^k$ is to be assigned no more than, exactly, or more than $n_{i_k}^k$ times, then the " = 1" in the constraint (2.15) is changed to " $\leq, =, \geq n_{i_k}^k$," respectively. In making these changes, one must pay careful attention to the independence assumptions (2.10) and (2.11). A key reason for the normalization in (2.9) is as follows. If for a particular combination of reports, $c_{i_1 \cdots i_N} > 0$, then one can preassign (before solving (2.15)) the 0-1 variable $z_{i_1 \cdots i_N} = 0$, since the assignment of the reports in the corresponding $Z_{i_1 \cdots i_N}$ as false reports leads to a lower cost. (This does not mean that each of the actual observations in $Z_{i_1 \cdots i_N}$ is declared to be a false alarm.) Finally, the key independence assumptions (2.10) and (2.11) need not be valid in many potential applications, and this may lead to the exciting possibility of nonlinearity in the formulation!

3. Assignment Formulation of MHT for Multitarget Tracking

Methods for mutitarget tracking generally fall into two categories: sequential and deferred logic. Methods for the former include nearest neighbor, one-to-one or few-to-one assignments, and all-to-one assignments as in the joint probabilistic data association (JPDA) [3]. For track maintenance, the nearest neighbor is valid in the absence of clutter when there is no track contention, i.e., when there is no chance of misassociation.

Problems involving one-to-one or few-to-one assignments are generally formulated as (two dimensional) assignment or multi-assignment problems for which there are some excellent optimal algorithms [4,5]. This methodology is real-time but can result in a large number of partial and incorrect assignments, particularly in dense or high contention scenarios, and thus incorrect track identification. The fundamental difficulty is that decisions, once made, are irrevocable, so that there is no mechanism to correct misassociations. The use of all observations to update a track moderates the misassociation problem and has been highly successful for tracking a few targets in dense clutter [3].

At the other extreme is batch processing in which all observations (from all time) are processed together, but this is computationally too intensive for real-time applications. Between batch and sequential processing lies the deferred logic methods in which several scans of information are considered all at once in data association decisions. The principle method here is called multiple hypothesis tracking (MHT) in which one builds a tree of possibilities, assigns a likelihood based on Bayesian estimation, develops an intricate pruning logic, and then solves the data association problem by methods that range from explicit enumeration to implicit enumeration using what is essentially a branch and bound method. The use of explicit enumeration or branch and bound to solve this NP-hard problem in real-time is inevitably faulty for problems involving dense scenarios or high track contention. Since the time can grow exponentially in the size of the problem, the failure is not graceful, i.e., the method is not robust.

Thus the objective in this section is to demonstrate that data association problems posed in multiply hypothesis tracking (MHT) can be formulated as multidimensional assignment problems, to which new, fast, and near-optimal algorithms are applicable [23-29]. For this purpose we convert the data association problem posed by Reid [30] and modified by Kurien [19] to include maneuvering targets to a multidimensional assignment problem. Two modifications of these formulas yield better performance in practice and discuss them at the end of the first subsection. The track initiation problem and a sliding window implementation for track maintenance, as developed in our previous work [23-29], are presented in the third and fourth subsections, respectively.

§3.1 Target Dynamics. Except in a maneuver, each target is assumed to satisfy a state-space system modeled by

$$\begin{aligned} x(k+1) &= F_k(x(k)) + G_k(x(k))w(k) \\ z(k) &= H_k(x(k)) + v(k) \end{aligned} \tag{3.1}$$

Here $x(k)$ is a vector of n state variable, w is a white noise sequence of normal random variables with zero mean and covariance matrix $Q(k)$, $z(k)$ represents the measurement at time k associated with this particular target, and v is a white noise sequence of normal random variables with zero mean and covariance $R(k)$, $H_k(x(k))$ relates the state to the measurements.

§3.2 Likelihood Calculations. Perhaps the most difficult part of writing down the scoring function (2.13) is the notation involved in the computations. Thus the various expressions are defined here for easier

reference:

(3.2)

P_χ^k	Probability of termination on scan k
P_d^k	Probability of detection on scan k
P_m^k	Probability of a maneuver on scan k
z_i^k	measurement i from scan k
δ^k	number of measurements on scan k originating from previously established tracks
ν^k	number of new targets detected on scan k
f^k	number of false alarms on scan k
M_k	total number of measurements on scan k, i.e., $M_k = \delta^k + \nu^k + f^k$
τ^k	number of targets that were extended from scan $k-1$ to scan k
χ^k	number of terminated (and nondetected) targets on scan k
m^k	number of maneuvering (and detected) targets on scan k

Any maneuvering target is assumed to be detected. These numbers give rise to the following facts: of the τ^k targets that enter scan k from scan $k-1$, χ^k terminate and are not observed on scan k, δ^k tracks are detected, and there are $\tau^k - \chi^k - \delta^k$ missed detections. Of the δ^k detected and continuing targets, m^k perform a maneuver, and $\delta^k - m^k$ continue with the existing dynamic model. The total number of targets on scan k that continue into scan $k+1$ is $\tau_{k+1} = \tau^k - \chi^k + \nu^k$. The following indicator functions are also needed:

(3.3)

$$f_i^k = \begin{cases} 1, & \text{if } z_i^k \text{ is a false alarm;} \\ 0, & \text{otherwise;} \end{cases}$$

$$\nu_i^k = \begin{cases} 1, & \text{if } z_i^k \text{ is a new target;} \\ 0, & \text{otherwise;} \end{cases}$$

$$m_i^k = \begin{cases} 1, & \text{if } z_i^k \text{ originated from a maneuvering target;} \\ 0, & \text{otherwise;} \end{cases}$$

$$\delta_i^k = \begin{cases} 1, & \text{if } z_i^k \text{ belongs to an existing track;} \\ 0, & \text{otherwise;} \end{cases}$$

$$\Delta_{ij} = \begin{cases} 1, & \text{i = j;} \\ 0, & \text{otherwise.} \end{cases}$$

The customary notation here [30] is to define $\Omega^k = \{\Omega_j^k \mid j = 0, 1, \ldots, I_k\}$ to be the set of hypotheses about the feasible partitions of the cumulative set of measurements Z^k into tracks and false alarms. In the notation of the previous section, each $\Omega_j^k = \Upsilon(\gamma^k)$ for some $\gamma^k \in \Gamma^k$ where the dependence of the γ on the number of scans k has been explicitly included as a superscript.

Let $\Omega_{\omega(l)}^{k-1}$ denote that specific hypothesis in Ω^{k-1} that produces Ω_l^k. Let $\psi_l(k)$ denote the hypothesis that indicates the specific status of all targets postulated by $\Omega_{\omega(l)}^{k-1}$ at the scan time k and the specific origin of all reports received at scan time k. Thus

(3.4)

$$\Omega_l^k = \psi_l(k) \cup \Omega_{\omega(l)}^{k-1}.$$

Using Bayes' rule, one can write

$$
\begin{aligned}
P(\Omega_l^k|Z^k) &= P(Z^k,\Omega_l^k)\left[\frac{1}{P(Z^k)}\right] = P(Z(k),Z^{k-1},\psi_l(k),\Omega_{\omega(l)}^{k-1})\left[\frac{1}{P(Z^k)}\right] \\
&= p(Z(k)|\Omega_l^k,Z^{k-1})P(\psi_l(k)|\Omega_{\omega(l)}^{k-1},Z^{k-1})P(\Omega_{\omega(l)}^{k-1}|Z^{k-1})\left[\frac{P(Z^{k-1})}{P(Z^k)}\right]
\end{aligned} \tag{3.5}
$$

The first term in this product is the likelihood of the measurements $Z(k)$ given the association hypothesis. (Each measurement is assigned to a previous track, which can either continue or maneuver, to a new source, or to a false alarm.) Thus

$$
\begin{aligned}
p(Z(k)|\Omega_l^k,Z^{k-1}) = \prod_{i=1}^{M_k}\Big\{ &\left[p_t^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{\delta_i^k(1-m_i^k)}\left[p_m^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{\delta_i^k m_i^k} \\
&\left[p_\nu^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{\nu_i^k}\left[p_f^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{f_i^k}\Big\}.
\end{aligned} \tag{3.6}
$$

Next, one must develop an expression for $P(\psi_l(k)|\Omega_{\omega(l)}^{k-1},Z^{k-1})$. Such a computation can be found in the work of Kurien [19] modeled after the work of Reid [30], and the result is

$$
\begin{aligned}
P(\psi_l(k)|\Omega_{\omega(l)}^{k-1},Z^{k-1}) = &\left\{\frac{\nu^k!f^k!}{M_k!}\mu_f^k(f^k)\mu_\nu^k(\nu^k)\right\}\left\{(P_\chi^k)^{\chi^k}\right\} \\
&\left\{\left[(1-P_\chi^k)(1-P_d^k)(1-P_m^k)\right]^{\tau^k-\delta^k-\chi^k}\right\} \\
&\left\{\left[(1-P_\chi^k)P_d^kP_m^k\right]^{m^k}\right\}\left\{\left[(1-P_\chi^k)P_d^k(1-P_m^k)\right]^{\delta^k-m^k}\right\}
\end{aligned} \tag{3.7}
$$

The combination of (3.5) - (3.7) yields

$$
\begin{aligned}
P(\Omega_l^k|Z^k) = &P(\Omega_{\omega(l)}^{k-1}|Z^{k-1})\left\{\frac{P(Z^{k-1})}{P(Z^k)}\right\}\left\{\frac{\nu^k!f^k!}{M_k!}\mu_f^k(f^k)\mu_\nu^k(\nu^k)\right\} \\
&\left\{(P_\chi^k)^{\chi^k}\left[(1-P_\chi^k)(1-P_d^k)(1-P_m^k)\right]^{\tau^k-\delta^k-\chi^k}\right\} \\
&\prod_{i=1}^{M_k}\Big\{\left[(1-P_\chi^k)P_d^k(1-P_m^k)p_t^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{\delta_i^k(1-m_i^k)} \\
&\left[(1-P_\chi^k)P_m^kP_d^kp_m^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{\delta_i^km_i^k}\left[p_\nu^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{\nu_i^k}\left[p_f^k(z_i^k|\Omega_l^k,Z^{k-1})\right]^{f_i^k}\Big\}
\end{aligned} \tag{3.8}
$$

We normalize this expression by dividing by $P(\Omega_l^0|Z^k)$ where $\Omega_0^k = \{\text{All False Alarms}\}$.

$$
\begin{aligned}
\frac{P(\Omega_l^k|Z^k)}{P(\Omega_0^k|Z^k)} = &\frac{P(\Omega_{\omega(l)}^{k-1}|Z^{k-1})}{P(\Omega_0^k|Z^{k-1})}\left\{\frac{\nu^k!f^k!}{M_k!}\frac{\mu_f^k(f^k)\mu_\nu^k(\nu^k)}{\mu_f^k(M_k)\mu_\nu^k(0)}\right\} \\
&\left\{(P_\chi^k)^{\chi^k}\left[(1-P_\chi^k)(1-P_d^k)(1-P_m^k)\right]^{\tau^k-\delta^k-\chi^k}\right\} \\
&\prod_{i=1}^{M_k}\Big\{\left[(1-P_\chi^k)P_d^k(1-P_m^k)\frac{p_t^k(z_i^k|\Omega_l^k,Z^{k-1})}{p_f^k(z_i^k|\Omega_l^k,Z^{k-1})}\right]^{\delta_i^k(1-m_i^k)} \\
&\left[(1-P_\chi^k)P_m^kP_d^k\frac{p_m^k(z_i^k|\Omega_l^k,Z^{k-1})}{p_f^k(z_i^k|\Omega_l^k,Z^{k-1})}\right]^{\delta_i^km_i^k}\left[\frac{p_\nu^k(z_i^k|\Omega_l^k,Z^{k-1})}{p_f^k(z_i^k|\Omega_l^k,Z^{k-1})}\right]^{\nu_i^k}\Big\}
\end{aligned} \tag{3.9}
$$

If, in addition, $\mu_f^k(f^k) = \exp(-\lambda_f^k)\frac{(\lambda_f^k)^{f^k}}{f^k!}$ and $\mu_\nu^k(\nu^k) = \exp(-\lambda_\nu^k)\frac{(\lambda_\nu^k)^{\nu^k}}{\nu^k!}$ are Poisson probability mass functions, then

$$\frac{P(\Omega_l^k|Z^k)}{P(\Omega_0^k|Z^k)} = \frac{P(\Omega_{\omega(l)}^{k-1}|Z^{k-1})}{P(\Omega_0^{k-1}|Z^{k-1})}\left\{(P_\chi^k)^{\chi^k}\left[(1-P_\chi^k)(1-P_d^k)(1-P_m^k)\right]^{\tau^k-\delta^k-\chi^k}\right\}$$

$$\prod_{i=1}^{M_k}\left\{\left[\frac{(1-P_\chi^k)P_d^k(1-P_m^k)p_t^k(z_i^k|\Omega_l^k,Z^{k-1})}{\lambda_f^k p_f^k(z_i^k|\Omega_l^k,Z^{k-1})}\right]^{\delta_i^k(1-m_i^k)}\right. \tag{3.10}$$

$$\left.\left[\frac{(1-P_\chi^k)P_m^k P_d^k p_m^k(z_i^k|\Omega_l^k,Z^{k-1})}{\lambda_f^k p_f^k(z_i^k|\Omega_l^k,Z^{k-1})}\right]^{\delta_i^k m_i^k}\left[\frac{\lambda_\nu^k p_\nu^k(z_i^k|\Omega_l^k,Z^{k-1})}{\lambda_f^k p_f^k(z_i^k|\Omega_l^k,Z^{k-1})}\right]^{\nu_i^k}\right\}$$

(The Poisson assumption allows one to satisfy (2.10) and (2.11).) For a track $Z_{i_1\cdots i_N}$ define

$$P_\phi^k = \begin{cases} P_\chi^k, & \text{if track } Z_{i_1\cdots i_N} \text{ terminates at scan } k; \\ (1-P_\chi^k)(1-P_d^k)(1-P_m^k), & \text{if track } Z_{i_1\cdots i_N} \text{ has a missed detection on scan } k; \\ 1, & \text{otherwise.} \end{cases} \tag{3.11}$$

The formula for $\frac{P(\Omega_l^k|Z^k)}{P(\Omega_0^k|Z^k)}$ has been developed recursively for $k \geq 1$; however, we now consider N scans of measurements as a single entity. Assuming at least two nonzero indices, each track of measurements $Z_{i_1\cdots i_N}$ in a feasible partition $\gamma \in \Gamma$ is scored as

$$L_{i_1 i_2\ldots i_N} \equiv L(Z_{i_1 i_2\ldots i_N}) \equiv L(z_{i_1}^1 \ldots z_{i_N}^N)$$

$$= \prod_{k=1}^{N}\left\{P_\phi^k\right\}^{\Delta_{0i_k}}\left\{\left[\frac{(1-P_\chi^k)P_d^k(1-P_m^k)p_t^k(z_{i_k}^k|\Omega_l^k,Z^{k-1})}{\lambda_f^k p_f^k(z_{i_k}^k|\Omega_l^k,Z^{k-1})}\right]^{\delta_{i_k}^k(1-m_{i_k}^k)}\right. \tag{3.12a}$$

$$\left.\left[\frac{(1-P_\chi^k)P_m^k P_d^k p_m^k(z_{i_k}^k|\Omega_l^k,Z^{k-1})}{\lambda_f^k p_f^k(z_{i_k}^k|\Omega_l^k,Z^{k-1})}\right]^{\delta_{i_k}^k m_{i_k}^k}\left[\frac{\lambda_\nu^k p_\nu^k(z_{i_k}^k|\Omega_l^k,Z^{k-1})}{\lambda_f^k p_f^k(z_{i_k}^k|\Omega_l^k,Z^{k-1})}\right]^{\nu_{i_k}^k}\right\}^{(1-\Delta_{0i_k})}$$

while

$$L_{0\cdots 0 i_k 0\cdots 0} \equiv 1 \text{ provided } Z_{0\cdots 0 i_k 0\cdots 0} \in \gamma. \tag{3.12b}$$

Each track $Z_{i_1\cdots i_N}$ carries with it such information as the time at which the track initiates, the time at which the target maneuvers or terminates. For example, if this particular target initiates on scan $r > 1$ and terminates on scan $s < N$, then $i_k = 0$ and $\Delta_{0i_k} = 1$ for $k = 1, \ldots, r-1$ and for $k = s, \ldots, N$, so that

$$\{P_\phi^k\}^{\Delta_{0i_k}} = \begin{cases} 1 & \text{if } k = 1, \ldots, r-1 \text{ and } k = s+1, \ldots, N; \\ P_\chi^k & \text{if } k = s. \end{cases}$$

This derivation assumes, as in the work of Kurien [19], that undetected targets do not maneuver and that targets that terminate do so immediately after being detected for the last time. These assumptions are easily removed. Let $N_1 \leq N$ be the last scan on which the target was detected, and define P_T as

$$P_T = \begin{cases} 1 & \text{if } N_1 = N, \\ \prod_{k=N_1+1}^{N}(1-P_\chi^k)(1-P_d^k) + \sum_{k=N_1+1}^{N} P_\chi^k \prod_{j=N_1+1}^{k-1}[(1-P_\chi^j)(1-P_d^j)] & \text{otherwise.} \end{cases}$$

P_T is the probability associated with the trailing missed detections, signifying that target has either terminated or is still present but undetected. The term P_ϕ^k can now be redefined to apply only to missed detections where the hypothesized track could not have terminated, i.e., where target was detected on a later scan.

$$P_\phi^k = \begin{cases} (1 - P_\nu^k)(1 - P_d^k), & \text{if track } Z_{i_1 \cdots i_N} \text{ has a missed detection on scan } k; \\ 1, & \text{otherwise.} \end{cases}$$

The likelihood ratio given in (3.12a) is now redefined to be

$$\begin{aligned} L_{i_1 i_2 \cdots i_N} &\equiv L(Z_{i_1 i_2 \cdots i_N}) \equiv L(z_{i_1}^1 \cdots z_{i_N}^N) \\ &= P_T \prod_{k=1}^{N_1} \left\{ P_\phi^k \right\}^{\Delta_{o_{i_k}}} \left\{ \left[\frac{(1 - P_\nu^k) P_d^k (1 - P_m^k) p_t^k (z_{i_k}^k | \Omega_t^k, Z^{k-1})}{\lambda_f^k p_f^k (z_{i_k}^k | \Omega_t^k, Z^{k-1})} \right]^{\delta_{i_k}^k (1 - m_{i_k}^k)} \right. \\ &\quad \left. \left[\frac{(1 - P_\nu^k) P_m^k P_d^k p_m^k (z_{i_k}^k | \Omega_t^k, Z^{k-1})}{\lambda_f^k p_f^k (z_{i_k}^k | \Omega_t^k, Z^{k-1})} \right]^{\delta_{i_k}^k m_{i_k}^k} \left[\frac{\lambda_\nu^k p_\nu^k (z_{i_k}^k | \Omega_t^k, Z^{k-1})}{\lambda_f^k p_f^k (z_{i_k}^k | \Omega_t^k, Z^{k-1})} \right]^{\nu_{i_k}^k} \right\}^{(1 - \Delta_{o_{i_k}})} \end{aligned} \tag{3.13}$$

§3.3 Track initiation. If one specializes the development in §3.2 to the situation in which each object in the data set $Z(k)$ is a measurement or observation for $k = 1, \ldots, N$, then the problem of partitioning the observations into tracks and false alarms can be posed as the multidimensional assignment problem (2.15) where

$$c_{i_1 \cdots i_N} = -\ln L_{i_1 \cdots i_N} \tag{3.14}$$

and $L_{i_1 \cdots i_N}$ is defined by equation (3.12).

§3.4. Track maintenance using a sliding window. Suppose now that the observations on P ***previous*** scans (of observations) have been partitioned into tracks and false alarms and that K ***new*** scans of observations are to be added. One approach to solving the resulting data association problem is formulate the problem as a track initiation problem with $P+K$ scans. The approach adopted here, however, is to treat the track extension problem within the framework of a window sliding over the observation sets. First assume that the scans of observations are partitioned into three components: D ***discarded*** scans of observations, R ***retained*** scans of observations from the P ***previously*** processed scans, and K ***new*** scans of observations. Thus the number of scans in the sliding window is $N = R + K$ while the number of ***discarded*** scans is $D = P - R$.

Let M_0 denote the number of confirmed tracks previously constructed from the discarded and retained regions that are present at the start of the tracking window. (Tracks terminated in the discarded region are not included in M_0.) Suppose the i_0^{th} such track is denoted by T_{i_0} for $i_0 = 1, \ldots, M_0$. For $i_0 > 0$, the $(N+1)$-tuple $\{T_{i_0}, z_{i_1}^1, \ldots, z_{i_N}^N\}$ will denote a track T_{i_0} plus a set of observations or measurements $\{z_{i_1}^1, \ldots, z_{i_N}^N\}$, actual or dummy, that are feasible with the track T_{i_0}. The $(N+1)$-tuple $\{T_0, z_{i_1}^1, \ldots, z_{i_N}^N\}$ will denote a track that initiates in the sliding window.

Let $\Omega^N(\gamma)$ denote the hypothesis about a partition $\gamma \in \Gamma$ being true, but now conditioned on the truth of the M_0 tracks entering the N-scan window. (Thus the assignments prior to this sliding window are fixed.) The likelihood function is now defined by $L_\gamma = \prod_{\{T_{i_0}, z_{i_1}, \ldots, z_{i_N}\} \in \gamma} L_{i_0 i_1 \cdots i_N}$, where $L_{i_0 i_1 \cdots i_N} = L_{T_{i_0}} L_{i_1 \cdots i_N}$,

$L_{T_{i_0}}$ is the composite likelihood from the discarded scans just prior to the first scan in the window for $i_0 > 0$. $L_{T_0} = 1$, and $L_{i_1 \cdots i_N}$ is defined as in (3.12) for the N-scan window. ($L_{T_0} = 1$ is used for any tracks that initiate in the sliding window.) Thus the track extension problem can be formulated as *Maximize* $\{L_\gamma \mid \gamma \in \Gamma\}$. With the same convention as in Section 3, a feasible partition is one in which every nonzero index on every coordinate appears in exactly one $(N+1)$-tuple in γ. Thus define a zero-one variable $z_{i_0 i_1 \cdots i_N}$ for a track of measurements $\{T_{i_0}, z_{i_1}^1, \ldots, z_{i_N}^N\}$ by

$$z_{i_0 i_1 \cdots i_N} = \begin{cases} 1 & \text{if } \{T_{i_0}, z_{i_1}^1, \ldots, z_{i_N}^N\} \text{ is assigned as a unit,} \\ 0 & \text{otherwise.} \end{cases}$$

and the corresponding cost for the assignment of the sequence $\{T_{i_0}, z_{i_1}, \ldots, z_{i_N}\}$ to a track by $c_{i_0 i_1 \cdots i_N} = -\ln L_{i_0 i_1 \cdots i_N}$. The data association problem of partitioning measurements into true and false tracks, i.e., tracks and false alarms, for *track maintenance* can now be posed as the following multi-dimensional assignment problem:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{i_0=0}^{M_0} \cdots \sum_{i_N=0}^{M_N} c_{i_0 \cdots i_N} z_{i_0 \cdots i_N} \\
\text{Subj. To} \quad & \sum_{i_1=0}^{M_1} \cdots \sum_{i_N=0}^{M_N} z_{i_0 \cdots i_N} = 1, \quad i_0 = 1, \ldots, M_0, \\
& \sum_{i_0=0}^{M_0} \sum_{i_2=0}^{M_2} \cdots \sum_{i_N=0}^{M_N} z_{i_0 i_1 \cdots i_N} = 1, \quad i_1 = 1, \ldots, M_1, \\
& \sum_{i_0=0}^{M_0} \cdots \sum_{i_{k-1}=0}^{M_{k-1}} \sum_{i_{k+1}=0}^{M_{k+1}} \cdots \sum_{i_N=0}^{M_N} z_{i_0 \cdots i_N} = 1, \\
& \qquad \text{for} \quad i_k = 1, \ldots, M_N \text{ and } k = 2, \ldots, N-1, \\
& \sum_{i_0=0}^{M_0} \cdots \sum_{i_{N-1}=0}^{M_{N-1}} z_{i_0 \cdots i_N} = 1, \quad i_N = 1, \ldots, M_N, \\
& z_{i_0 \cdots i_N} \in \{0, 1\} \text{ for all } i_0, \ldots, i_N.
\end{aligned} \tag{3.15}$$

4. Multisensor Fusion

The use of multiple sensors, through more varied information, has the potential to greatly enhance target identification and state estimation. Again, the central problem is that of data association, i.e., that of determining which observations emanate from common targets or sources and which observations are false reports. When one considers techniques for solving the corresponding data association problems, we note once again that MHT is a method that permeates the field [1,2]. This section considers but two examples, but before proceeding, a brief review of some of the issues involved in the design of multisensor fusion algorithms is in order.

Two important decisions involve sensor location, i.e., distributed or collocated, and the level of data association, i.e., central level, sensor level, or hybrids of these two. Sensors spatially distributed enhances geographical diversity and survivability. What's more, the combination of active and passive sensors can

now use the geometry of separation to identify targets. For example, several passive sensors can be used to great benefit; however, the ghosting problem now appears [7]. Other difficulties include the communication complexity and registration errors. In collocated sensor placement, i.e., in the same place or on the same platform, sensor diversity is generally chosen to provide complementary information. For example, a two dimensional radar provides accurate range and moderately accurate azimuth measurements, whereas an infrared sensor provides highly accurate azimuth. The combination can yield highly accurate range and azimuth. Communication complexities and registration errors can be greatly reduced for collocated sensors

The choices of the level of association range from sensor- to central-level tracking with hybrids in between. In the former case, each sensor forms tracks from its own observations and then the tracks from N sensors are fused in a central location. (One may think of N-dimensional assignments here as assigning tracks from the first sensor to tracks from the sensor and so forth to tracks from the N^{th} sensor.) Once the matching is complete, one then combines the tracks with appropriate modification in the statistics. One of the problems here is that the usual error independence assumption is not valid and this introduces additional complexity. Another problem is that the combined track estimates tend to be worse than in central-level fusion of measurements only. Arguments in favor of this method are reduced communications costs and higher survivability since the sensors maintain their own tracks. At the other extreme is centralized fusion in which sensors send measurements to a central processing unit where they are combined to give superior position measurements. From the point of view of track estimation, this method seems to be superior. The difficulties are data association problem, communication costs between the sensor and central processing unit, and the loss of the tracking capability if the central processing unit becomes inoperative.

In the next two subsections examples of noncollocated and collocated sensors are discussed. These examples assume synchronous measurements. This may introduce more error into the observations than that of the sensors themselves and is a topic that must be addressed in future work.

§4.1 Noncollocated Sensors. In this first subsection we closely follow the work of Deb, Pattipati, and Bar-Shalom [9] except that we consider N noncollocated sensors, i.e., all sensors are spatially separated from one another or are on different platforms. The location of the N sensors are assumed known with locations $\{p_k\}_{k=1}^{N}$. The locations and number of the targets are assumed to have unknown locations $\{p^t\}_{t=1}^{M}$. For a three dimensional scenario the notation $p_k = (x_k, y_k, z_k)$ and $p^t = (x^t, y^t, z^t)$ will be used in the sequel. Each of the sensors may be one of three types: a 3D radar, a 2D radar, or 2D passive sensor. The passive sensor measures the azimuth angle and elevation angle of each potential target t, i.e. $z_{i_k}^k = [\theta_{kt}, \phi_{kt}]$; the 2D radar measures azimuth and range, i.e., $z_{i_k}^k = [r_{kt}, \theta_{kt}]$; and, a 3D radar measures all three, i.e., $z_{i_k}^k = [r_{kt}, \theta_{kt}, \phi_{kt}]$. The k^{th} sensor makes M_k actual measurements and we add a dummy variable for a missed detection and denote the corresponding data set by $Z(k) = \{z_{i_k}^k\}_{i_k=0}^{M_k}$ where z_0^k is the dummy variable. The measurements are made synchronously with the following statistical properties

$$z_{i_k}^k = \begin{cases} H(p_k, p_t) + v_{i_k}^k, & \text{if } z_{i_k}^k \text{ is from a true target;} \\ w_{i_k}^k, & \text{if } z_{i_k}^k \text{ is a spurious measurement;} \end{cases} \tag{4.1}$$

where here $H(p_k, p_t)$ is the true observable, $v_{i_k}^k \sim N(0, \Sigma_k)$, and the density of the spurious measurement is assumed uniform and is given by

$$P_{w_{i_k}^k}(w) = \frac{1}{\Psi_k}$$

where Ψ_k is the field of view. Finally, let P_d^k denote the probability of detection of the k^{th} sensor.

Assuming equal priors, the likelihood ratio $L_{i_1 \cdots i_N}$ in (2.12) simplifies to

$$L_{i_1 \cdots i_N} = \frac{p(Z_{i_1 \cdots i_N} | \Upsilon(\gamma))}{\prod_{k=1, i_k \neq 0}^{N} p(Z_{0 \cdots 0 i_k 0 \cdots 0} | \Upsilon(\gamma^0))} \tag{4.2a}$$

Let p^t denote the target position giving rise to the measurements $z_{i_1}^1, \ldots, z_{i_N}^N$. Due to assumption (2.11), we use the abbreviated notation $p(Z_{i_1 \cdots i_N} | \Upsilon(\gamma)) = p(Z_{i_1 \cdots i_N} | p^t)$. Then (4.2a) can be written as [19]

$$L_{i_1 \cdots i_n}(p_t) = \prod_{k=1}^{N} \frac{\left[P_d^k p(z_{i_k}^k | p_t)\right]^{1-\Delta_{0i_k}} \left[1 - P_d^k\right]^{\Delta_{0i_k}}}{\prod_{k=1, i_k \neq 0}^{N} \Psi_k^{-1}} = \prod_{k=1}^{N} \left[P_d^k \Psi_k p(z_{i_k}^k | p^t)\right]^{1-\Delta_{0i_k}} \left[1 - P_d^k\right]^{\Delta_{0i_k}} \tag{4.2b}$$

Then the problem of associating the measurements from all sensors is precisely the multidimensional assignment problem formulated in (2.15) where $c_{i_1 \cdots i_N} = -\ln L_{i_1 \cdots i_n}(p_t)$. Since p^t is unknown, it is replaced by its maximum likelihood estimate

$$\hat{p}^t = \text{Arg Min} L_{i_1 \cdots i_n}(p^t) \tag{4.3a}$$

This $\hat{p}^t$ is, of course, the solution of the statistically weighted nonlinear least squares problem

$$\text{Minimize} \sum_{k=1}^{N} (1 - \Delta_{0i_k}) \left[z_{i_k}^k - H(p_k, p^t)\right]^T \Sigma_k^{-1} \left[z_{i_k}^k - H(p_k, p^t)\right] \tag{4.3b}$$

As an example, $H(p_k, p^t)$ for a 3D-radar is given by

$$H(p_k, p^t) = \begin{bmatrix} \sqrt{\Delta x_k^2 + \Delta y_k^2 + \Delta z_k^2} \\ \arctan\left[\frac{\Delta y_k}{\Delta x_k}\right] \\ \frac{\sqrt{\Delta x_k^2 + \Delta y_k^2}}{\Delta z_k} \end{bmatrix} \tag{4.4}$$

where $(\Delta x_k, \Delta y_k, \Delta z_k) = (x^t - x_k, y^t - y_k, z^t - z_k)$ and (x^t, y^t, z^t) is determined in the course of solving the nonlinear least squares problem (4.3b).

§4.2 Multiple platforms: An example. Consider N spatially separated platforms such that on each platform one as a 2D radar measuring range and azimuth and a passive sensor measuring azimuth and elevation. The location $p_k = (x_k, y_k, z_k)$ of each platform is known: however, the group of M targets (M unknown) and unknown locations $\{p^t\}_{t=1}^{M}$ are observed by the sensors on the various platforms. For a three dimensional scenario the notation $p_k = (x_k, y_k, z_k)$ and $p^t = (x^t, y^t, z^t)$ will be used in the sequel. The passive sensor measures the azimuth angle and elevation angle of each potential target t, i.e. $z_{i_k}^k = [\theta_{kt}, \varphi_{kt}]$; the 2D radar measures azimuth and range, i.e., $z_{i_k}^k = [r_{kt}, \theta_{kt}]$. The k^{th} sensor makes M_k actual measurements and

we add a dummy variable for a missed detection and denote the corresponding data set $Z(k) = \{z_{i_k}^k\}_{i_k=0}^{M_k}$ where z_0^k is the dummy variable. The statistical properties of these measurements are defined by as in (4.1), and the problem is formulated as in (4.2). A minor difference occurs in the computation of p^t. To be precise, let sensors $2k-1$ and $2k$ denote the 2D radar and passive sensors on platform k. Then

This p^t is, of course, the solution of the statistically weighted nonlinear least squares problem

$$\text{Minimize} \sum_{k=1}^{2N} (1 - \Delta_{0i_k}) \left[z_{i_k}^k - H(p_k, p^t)\right]^T \Sigma_k^{-1} \left[z_{i_k}^k - H(p_k, p^t)\right] \tag{4.5}$$

As an example, $H(p_{2k-1}, p^t)$ for a 2D-radar is given by

$$H(p_{2k-1}, p^t) = \begin{bmatrix} \sqrt{\Delta x_k^2 + \Delta y_k^2 + \Delta z_k^2} \\ \arctan\left[\frac{\Delta y_k}{\Delta x_k}\right] \end{bmatrix} \tag{4.6a}$$

and that for the passive sensor by

$$H(p_{2k}, p^t) = \begin{bmatrix} \arctan\left[\frac{\Delta y_k}{\Delta x_k}\right] \\ \frac{\sqrt{\Delta x_k^2 + \Delta y_k^2}}{\Delta z_k} \end{bmatrix} \tag{4.6b}$$

where $(\Delta x_k, \Delta y_k, \Delta z_k) = (x^t - x_k, y^t - y_k, z^t - z_k)$ and (x^t, y^t, z^t) is determined in the course of solving the nonlinear least squares problem (4.3b).

5. Algorithms

Lagrangian relaxation originally gained prominence as a method for obtaining tight bounds for a branch and bound algorithm in Held and Karp's highly successful work on the traveling salesman problem [16,17]. Overviews of this methodology can be found in the works of Geoffrion [14], Fisher [10], Shapiro [31], the book by Nemhauser and Wolsey [21], and the references therein. The particular Lagrangian relaxation scheme used in our [23-29] work is motivated by the relaxation scheme of Frieze and Yadegar [11,12] for three dimensional assignment problems and incorporates the conjugate subgradient algorithms of Wolfe [32,33] for nonsmooth optimization. We also use an adaptation of the reverse auction algorithm of Bertsekas, Castanon, and Tsaknakis [4,5] for the two dimensional assignment problems. Older versions of this algorithm has been previously published [23-28], and the newer versions are the subject of a forthcoming paper [29].

The use of parallel architectures has not been discussed in this work; however, two observations are worthy. Ignoring the zero index and assuming $M_k = M$ for each $k = 1, \ldots, N$ in (3.15), note that one must compute the M^N, which for fixed N is polynomial in the size of the problem M. Contrast this with the number of feasible partitions of the constraints, $(M!)^{N-1}$. The computation of the coefficients requires the solution of M^N nonlinear least squares for multisensor data fusion and M^N filters for multitarget tracking. Each of these is highly independent of one another and ideally suited to parallel computation. On a serial machine, our experience shows that this requires at least one or two orders of magnitude more time than solving the data association problem.

6. References

[1] Y. Bar-Shalom, ed., *Multitarget-Multisensor Tracking: Advanced Applications*, Artech House, Dedham, MA., 1990.

[2] Y. Bar-Shalom, ed., *Multitarget-Multisensor Tracking: Applications and Advances*, Artech House, Dedham, MA., 1992.

[3] Y. Bar-Shalom and T.E. Fortmann, *Tracking and Data Association*, Academic Press, Boston, 1988.

[4] D. P. Bertsekas, *Linear Network Optimization: Algorithms and Codes*, The MIT Press, Cambridge, Mass, 1991.

[5] D. P. Bertsekas, D. A. Castanon, and H. Tsaknakis, "Reverse auction and the solution of inequality constrained assignment problems," preprint, March, 1991.

[6] S. S. Blackman, *Multiple Target Tracking with Radar Applications*, Artech House, Dedham, MA., 1986.

[7] S. S. Blackman, "Association and fusion of multiple sensor data", in [2]

[8] C.-Y. Chong, S. Mori, and K.-C. Chang, "Distributed multitarget multisensor tracking," in [2].

[9] S. Deb, K. R. Pattipati, and Y. Bar-Shalom, "A multisensor-multitarget data association algorithm for heterogeneous systems," preprint, 1992.

[10] M. L. Fisher, "The Lagrangian relaxation method for solving integer programming problem," *Management Science*, Vol. 27, No. 1, 1981, pp. 1-18.

[11] A. M. Frieze, "A bilinear programming formulation of the 3-dimensional assignment problem," *Mathematical Programming*, 7 (1974), pp. 376-379.

[12] A. M. Frieze and J. Yadegar, "An algorithm for solving 3-dimensional assignment problems with application to scheduling a teaching practice," *Journal of the Operational Research Society*, 32 (1981), pp. 989-995.

[13] M. R. Garvey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman & Co., CA, 1979.

[14] A. M. Geoffrion, "Lagrangean relaxation for integer programming," in M. L. Balinski, ed., *Mathematical Programming Study 2: Approaches to Integer Programming*, North Holland Publishing Company, Amsterdam, 1974.

[15] D. L. Hall, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Boston, 1992.

[16] M. Held and R.M.Karp, "The traveling salesman problem and minimal spanning trees," *Operations Research*, 18 (1970), pp. 1138 - 1162.

[17] M. Held and R. M. Karp, "The traveling-salesman problem and minimum spanning trees: Part II," *Mathematical Programming*, 1 (1971), pp. 6-25.

[18] S.A. Hovanessian, *Introduction to Sensor Systems*, Artech House, Dedham, MA., 1988.

[19] T. Kurien, "Issues in the designing of practical multitarget tracking algorithms," in [2].

[20] C. L. Morefield, "Application of 0-1 integer programming to multitarget tracking problems," *IEEE*

Transactions on Automatic Control, Vol. AC-22, No. 3, June 1977, pp. 302-312.

[21] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, John Wiley and Sons, New York, 1988.

[22] K. R. Pattipati, D. Somnath, Y. Bar-Shalom, and R. B. Washburn, "Passive multisensor data association using a new relaxation algorithm," in Bar-Shalom, Y., ed., *Multitarget-Multisensor Tracking: Advanced Applications*, Artech House, Dedham, MA., 1991.

[23] A. B. Poore and N. Rijavec, "A Lagrangian relaxation algorithm for multidimensional assignment problems arising from multitarget tracking," to appear in *SIAM Journal on Optimization*, 1993.

[24] A. B. Poore and N. Rijavec, "A New Class of Methods for Solving Data Association Problems Arising from Multi-target Tracking, *Proceedings of the 1991 American Automatic Control Conference*, Boston, MA, vol 3, 2303-2304.

[25] A. B. Poore and N. Rijavec, "Multitarget Tracking and Multidimensional Assignment Problems, in O. E. Drummond, ed., *Proceedings of the 1991 SPIE Conference on Signal and Data Processing of Small Targets* 1991, vol. 1481, 1991, pp 345 - 356.

[26] A. B. Poore and N. Rijavec, "The Data Association Problem in Multitarget Tracking and Multidimensional Assignment Problems, in G. Frenkel and B. Fridling, eds., the *Proceedings of the SDI Panels on Tracking*, Institute for Defense Analyses, Issue No. 2/1991, p 3-29 to 3-51.

[27] A. B. Poore and N. Rijavec, "Multitarget Tracking, Multidimensional Assignment Problems, and Lagrangian Relaxation," in G. Frenkel and B. Fridling, eds., *Proceedings of the SDI Panels on Tracking*, Institute for Defense Analyses, Issue No. 2/1991, pp 3-51 to 3-74.

[28] A. B. Poore, N. Rijavec, and T. Barker, "Data association for track initiation and extension using multiscan windows, to appear in the *Proceedings of the 1992 SPIE Conference on Small Targets*.

[29] A. B. Poore and N. Rijavec, "Multidimensional assignment problems and Lagrangian relaxation," in preparation.

[30] D. B. Reid, "An algorithm for tracking multiple targets," IEEE Transactions on Automatic Control," Vol. AC-24, No. 6, December 1979, pp. 843-854.

[31] J. F. Shapiro, "A Survey of Lagrangian Techniques for Discrete Optimization," *Annals of Discrete Mathematics*, 5 (1979), pp. 113-138.

[32] E. Waltz and J. Llinas, *Multisensor Data Fusion*, Artech House, Boston, 1990.

[33] P. Wolfe, "A method of conjugate subgradients for minimizing nondifferentiable functions," *Mathematical Programming Study*, 3 (1975), pp. 147-173.

[34] P. Wolfe, " Finding the nearest point in a polytope," *Mathematical Programming*, 11 (1976), pp. 128-149.

THERMAL CHARACTERIZATION OF IN-SITU SYNTHESIS FOR LEC/MLEK GROWTH OF InP SINGLE CRYSTALS

Vishwanath Prasad
Associate Professor
Department of Mechanical Engineering
Columbia University, New York, NY 10027

Final Report for:
Summer Research Program
Rome Laboratory, Hanscom AFB, MA 01731

Sponsored by:
Air Force Office of Scientific Research
Boiling Air Force Base, Washington, D.C.

September 1992

THERMAL CHARACTERIZATION OF IN-SITU SYNTHESIS FOR LEC/MLEK GROWTH OF InP SINGLE CRYSTALS

Vishwanath Prasad
Associate Professor
Department of Mechanical Engineering
Columbia University, New York

Abstract

For one-step, in-situ synthesis of phosphorus vapor and indium melt, and liquid-encapsulated Czochralski growth of InP crystals to succeed and produce single crystals of uniform quality and at lesser cost, it is important to understand the mechanics of heat transfer and gas flow in a high pressure crystal growth (HPCG) system. A series of experiments performed recently in order to characterize the thermal coupling between the melt and the phosphorus injector and to develop an understanding of the buoyancy-induced flow in an HPCG furnace is reported here. It is observed that although the flow in a high pressure puller is turbulent and oscillatory, radiation dominates the heat transfer. The thermal response of the system is therefore quite stable and predictable. An almost linear relationship exists between the power input and the melt temperature. The correlation between the temperatures at various locations of phosphorus injector and the melt is very interesting. The heat of reaction also affects the melt temperature. The phase change phenomenon at the bottom of the phosphorus injector is oscillatory in nature. Theoretical estimates of the strength of gas convection and radiation loss by the melt surface are also presented.

THERMAL CHARACTERIZATION OF IN-SITU SYNTHESIS FOR LEC/MLEK GROWTH OF InP SINGLE CRYSTALS

Vishwanath Prasad

INTRODUCTION

In order to grow single crystals of III-V compounds such as gallium arsenide, indium phosphide and so on, a high pressure crystal growth (HPCG) system is used. The mechanism of energy transport in such pullers is very different from that in the low pressure Czochralski furnaces for silicon crystals. While in a low pressure system, the energy is transferred to the melt and from the melt and crystal, primarily by conduction and radiation, gas convection plays an important role together with other modes of heat transfer in determining the thermal field within an HPCG puller. The buoyancy-driven convective flow in a high pressure chamber is turbulent and oscillatory, and the recirculating gas flow pattern is a strong function of the geometric configuration, temperatures of the r.f. coil, susceptor, melt, crystal and puller walls, and the gas properties. In addition, the convective heat transfer in an HPCG system is strongly coupled with the conduction in various components of this system and radiation exchange between the surfaces which can see each other.

The heat transfer mechanism is further complicated by the process employed to grow the single crystal. Motivated by the high cost of indium phosphide crystals grown by a two-step method - the synthesis of polycrystal by the horizontal Bridgeman method and single crystal growth by the liquid-encapsulated Czochralski (LEC) method, a one-step operation has been proposed and invesigated by several researchers [1-11]. It is basically an LEC method for pulling a single crystal from an "in-situ" synthesized melt of indium phosphide. Recently, magnetically stabilized liquid-encapsulated Czochralski (MLEC) and Kyropoulos (MLEK) growth of InP single crystals have also been reported [12,13]. Unlike the CZ growth of gallium arsenide crystals where gallium and arsenic are covered with boric oxide (B_2O_3) for synthesis at about 800°C and a little above atmospheric pressure, the phosphorus cannot be covered with boric oxide because it sublimes at a very low temperature (about 416°C for red P) and the melting temperature for InP is 1063°C at 27.5 atm. Although a scheme of using liquid phosphorus encapsulated indium for direct synthesis has been proposed recently [9-11], most of the "in-situ" synthesis and growth processes use phosphorus vapor injection into the indium melt at a temperature higher than the melting point of InP.

It is possible to generate phosphorus vapor for injection inside the HPCG system by electrically heating a quartz ampoule containing the solid P, and also to control its temperature by placing a thermocouple on the ampoule [1,2]. However, most of the investigators have preferred to take advantage of the existing thermal gradient in the puller to raise the phosphorus temperature [3-13]. In particular, the heat radiated by the melt can be absorbed by the solid phosphorus for melting and/or sublimation. The convective heat transfer can also help. The temperature of the phosphorus can therefore be controlled by changing the vertical location of the injector. In this scheme, the ampoule is placed directly above the melt and the vapor is transferred to indium melt by a quartz tube whose length is another key variable in the "in situ" synthesis. The heat transfer phenomena during the in-situ synthesis is therefore highly transient. The movement of P-injector changes the temperature field significantly by modifying the radiation exchange between various surfaces and by changing the convective gas flow rate and its structure. The energy generated into the melt due to the reaction of In and P, and the volume increase (density of In = 7.31 and density of InP =4.787) also affect the heat transfer and gas velocity. The temperature and flow fields also vary depending on whether the radial location of P-injector is changed or not. Mostly, the P-injector is moved to and away from the central axis of the puller for synthesis and growth, respectively, while in other arrangements, either the seed rod passes through the injector thereby eliminating its radial movement or phosphorus is vaporized in a horizontal boat away from the growth chamber [7].

For one-step, in-situ synthesis and growth process to succeed and produce InP crystals of uniform quality and at lesser cost, it is important to understand the mechanics of heat transfer and gas flow in the HPCG furnace. This paper reports a series of experiments performed recently in order to characterize the thermal coupling between the melt and the phosphorus ampoule, and to develop an understanding of the convective flow phenomena associated with the MLEK growth of InP single crystals. These experiments were conducted in an HPCG system in which the seed rod passes through the phosphorus injector which can be moved up or down along the central axis independent of the seed rod. It is believed that such a system is less complex than the arrangement in which the injector has to be moved aside before the seed rod can be pushed down for crystal growth. Our measurements demonstrate interesting correlation between the power input and the melt temperature, coupling between the injector and melt temperatures, variation between the melt and crucible bottom temperatures and the effect of heat of reaction. Patterns of gas flow within the HPCG furnace were observed visually and are shown schematically. The temperature measurement at the injector bottom reveals interesting oscillatory phase change phenomena.

SYSTEM OVERVIEW

In order to analyze the heat transfer mechanism and buoyancy-induced gas flow structure in an HPCG puller, it is necessary to consider the growth system in sufficient detail. Although there may exist variations from one puller for in-situ synthesis and growth of InP crystals to the other, major components and process steps are expected to remain the same. Also, we hope that an analysis of convection and radiation mechanism in our puller will help in understanding the transport phenomena in other HPCG systems. With this in mind, a description of the HPCG furnace (Fig. 1) used for ongoing research on InP crystal growth at Rome Laboratory is presented in the following paragraphs.

A 96 mm i.d. quartz crucible of 2 mm thick wall, placed in a 113 mm o.d. graphite susceptor is used for the present experiments. The susceptor is insulated from the r.f. coil by "Fiberfrax Lo-Con Felt" insulation ($k \cong 0.075$ W/m.K at 300°C, 0.13 W/m.K at 600°C and 0.26 W/m.K at 1000°C) placed in an annular space formed by two 2 mm thick quartz cylinders of 114 mm and 134 mm i.d. The gap betwe e inner cylinder and the susceptor wall is about 0.5 mm which can allow some gas flow from underneath the susceptor to the upper portion unless the insulation-filled quartz annulus provides a leakproof joint at the bottom. An opaque (sand blasted) glass plate of 280 mm diameter separates the bottom portion of the crystal growth chamber from its upper region. However, the gas can flow downward through a 2 mm gap between this plate and cold puller wall. (It should be kept in mind that the gaps are generally unavoidable from practical considerations.) The plate is placed on top of the quartz cylinders, 305 mm from the bottom surface of the puller, while the mean height of the upper tube of the r.f. coil is 286 mm.

The graphite susceptor with crucible can be moved up or down, and can be rotated by a shaft. The shaft holds a boron nitride (BN) hollow cylinder which, in turn, supports the susceptor. An insulation to reduce the heat loss from susceptor bottom can be provided by packing the BN cylinder with Lo-Con felt. The shaft to rotate the crucible is hollow from inside so that a tungsten sheathed W26Re-W.5Re thermocouple probe can pass through it and measure temperature at the crucible bottom. The T/c probe rotates with the shaft and hence requires a slip-ring arrangement for temperature reading.

A 75 mm o.d., 100 mm long quartz ampoule is used for phosphorus storage and its vapor injection into the indium melt. This injector has an off-center tube of 12 mm i.d. for the vapor outflow The same tube in an inverted position allows the annular space of the ampoule to be filled (about two-third) with (crushed) red phosphorous. The injector can be hung from

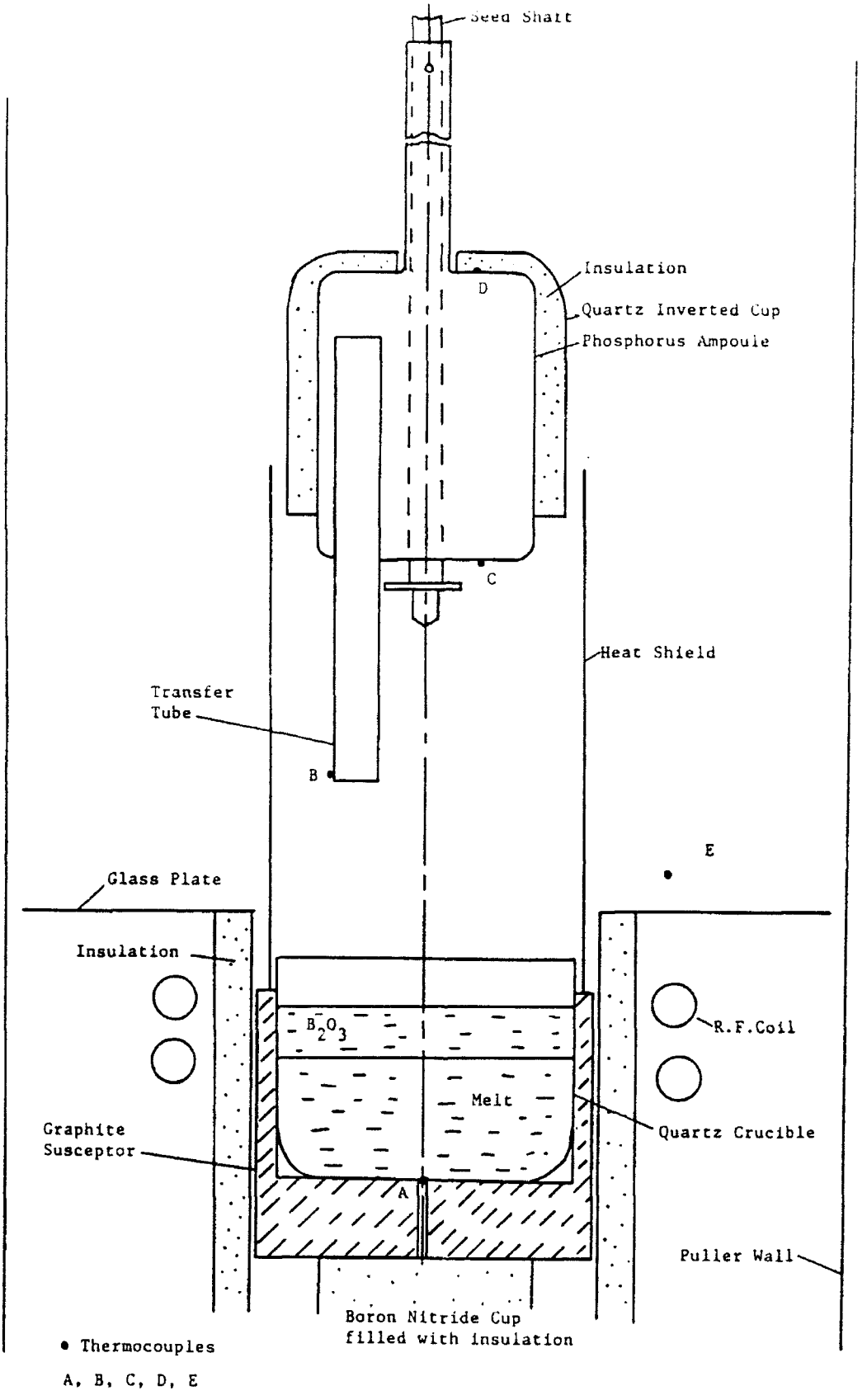


Fig.1 Schematic of the high pressure crystal puller used for on-going research on in-situ synthesis and growth of InP single crystals at Rome Laboratory.

the top with the help of a 12 mm i.d. hollow tube through which a cooled seed shaft can pass. The injector can be moved up or down independent of the seed shaft. The gap (about 2 mm) between the hollow, support tube and the seed shaft can allow the cold gases to flow downward along the seed shaft. To insulate the phosphorus injector from convection losses to cold gases and radiative heat transfer to chamber walls, 8 mm Lo-Con felt insulation is placed between the injector and an inverted quartz cup (Fig. 1). The length of the quartz cup and the insulation is slightly smaller than that of the P-injector in order not to obstruct the view of the boron oxide encapsulated melt surface.

In most of the experiments reported here, a quartz shield of 103 mm i.d. is used to isolate the strong convective flows in regions close to the cold wall from the buoyant flow directly above the melt. Although the heat shield cannot completely eliminate the buoyancy-induced recirculating flows - gas flows downward along the seed shaft and upward along the quartz shield - it can reduce the gas velocity significantly in the space between the melt surface and phosphorus injector. Note that a 4 mm gap exists between the outer surface of the insulating quartz cup and the heat shield even when the upper end of the shield tube is covered by the injector. The shield tube rests on the graphite susceptor and therefore can rotate whenever the crucible is rotated.

In this set-up, the temperature can be measured at three locations other than the crucible bottom - at injector tube tip, at the bottom (outside) surface of the ampoule, and between the insulation and upper surface of the injector (Fig. 1). The chromel-alumel thermocouple attached to the transfer tube is placed in a closed-end, thin quartz cylinder to protect it from the phosphorus reaction and contaminating the indium/InP melt at high temperatures when the tube is dipped into the melt for synthesis. Temperatures can be read by using digital thermometers or can be recorded by strip chart recorders.

There are two viewing ports in this HPCG puller, one is used for video monitoring and photography while the other one helps in visual observation. Since the system is run at a high temperature (above 1063°C) and a high pressure (500-600 psi) at which the density of nitrogen gas is about six to fifteen times larger (as a function of local temperature in the chamber) than at the atmospheric condition, it is possible to see the direction of flow whenever the gas flows over a step. (Note that the encapsulant boric oxide surface is much below the top edge of the crucible wall.) Light reflections also seem to help in this visualization. Although only a small portion of the growth chamber is visible, it is possible to draw flow patterns qualitatively from the observed gas flow behavior.

Whenever an in-situ synthesis is performed, the experiment is started with about 800

gm charge of indium (21 mm high when melted) and 160 gm of B_2O_3 (7 mm thick encapsulating layer) in the crucible, and approximately 260gm phosphorus in the ampoule. After a successful synthesis, the InP melt level rises to about 34 mm high. For present analysis, the information on height of the melt surface and vertical location of the injector are quite important. Several runs were performed with and without the synthesis and crystal growth to study the thermal characteristics of the system.

ORDER-OF-MAGNITUDE ANALYSIS

In this section, an estimate of the strength of gas convection and radiative heat transfer from the melt shall be presented. The information obtained here will help in analyzing the experimental data and visual observation.

Gas Convection. The strength of buoyancy-induced convection flow and heat transfer is determined by the Prandtl number, Pr ($= \mu C_p/k$) and the Rayleigh number,

$$Ra = \rho^2 g \beta C_p L^3 \Delta T/\mu k \qquad (1)$$

where L is the characteristic length of the system and ΔT is the temperature difference producing the buoyancy force. For a gas like nitrogen, μ, C_p, k and Pr remain almost independent of pressure, but vary with temperature. Since nitrogen behaves like an ideal gas between 1-10 MPa pressure and above 300K, the coefficient of isobaric expansion, β can be taken as $1/T_{mean}$. The temperature in the growth chamber varies from about 1400K at the melt surface to about 300K on the puller walls. The temperature difference, ΔT can therefore be taken as 1000K considering 100K drop in temperature across B_2O_3 layer. However, the gas temperature in most part of the puller is much below 1400K. We therefore assume a mean temperature of 600K to estimate the gas properties. Selection of an appropriate length scale is not easy. The inner height of the HPCG system is 850 mm while its diameter is 284 mm. The distance between the B_2O_3 encapsulant surface and the top wall may vary from 550 mm to 650 mm. For L = 600 mm, the gas Rayleigh number is estimated to be 1.1×10^{12}.

The buoyant flow conditions in the HPCG system do not fit into the definitions of classical free convection problems of Benard convection and differentially-heated vertical cavities [14]. The hot B_2O_3 surface and the cold injector bottom provide the so-called Benard conditions whereas the heated quartz shield in the upper portion and the hot vertical insulation layer in the bottom part of the puller together with the cold vertical wall represent a differentially heated vertical cavity configuration. Many other heated and cold surfaces and

gaps between various quartz cylinders further complicate the situation. The Benard convection flow is known to start oscillating at Ra $> 10^5$ leading to random oscillations and turbulent behavior at higher Rayleigh numbers. The vertical cavity flow becomes turbulent at Ra $\cong 10^9$. From theoretical considerations, the gas flow in the HPCG puller can therefore be expected to be turbulent. As is well-known, any step change in the surface over which the flow is taking place, a sudden change in the flow path and passage restrictions further add to the turbulent behavior. Flow turbulence was clearly observed during our experiments. Indeed, the oscillating nature of the gas flow could also be visualized by watching the movement of stray, thin clouds of phosphorus vapor, which appear above the melt time-to-time depending on the rate of reaction and supersaturation of indium melt by the vapor.

Radiation Exchange. Though the gas convection plays an important role in providing desirable (or undesirable!) conditions for the crystal growth, radiation dominates the heat transfer from the melt. For example, if the melt surface (at 1400K) can be treated as a blackbody, it will loose 1.57 kW by radiation to a cold ambient at 300K while the encapsulant B_2O_3 surface at 1300K can dissipate a maximum of 0.5 kW by convection to an infinite ambient at 600K. (An average heat transfer coefficient, h $\cong$ 100 W/m^2.K obtained from a correlation for turbulent free convection from a heated surface Nu = 0.15 Ra$^{1/3}$ has been used to calculate the convection loss [15].) In reality, the radiation loss will decrease if a realistic value for emissivity, ε of In or InP as the case may be, an appropriate value for absorptivity α, of B_2O_3, and the presence of phosphorus injector and heat shield are considered. However, the estimated convective heat transfer will reduce much more if it is treated as an enclosure (formed by injector, heat shield and crucible) problem with temperature at which heat is rejected, much higher than 600K. The presence of a crystal may also modify the above numbers.

It must be mentioned here that the overall convection and radiation heat transfer in a crystal puller is much greater than that calculated above because of the heat dissipation by outside surfaces of the susceptor and also by the r.f. coil. However, this cannot be expected to affect the radiation loss from the melt surface much. On the other hand, the hot gases flowing upward along the outside surface of the quartz shield may reduce convection loss from the encapsulant surface by raising the wall temperature. It should not, however, be perceived that the gas convection directly above the melt is weak. In fact, 100-500W convection loss represents 1.4 to 7 W/cm^2 heat flux which must be transferred by the buoyancy-induced gas flow. A survey of electronic cooling literature can easily demonstrate how large is this heat

flux for free convection dissipation of energy to a gas [16]. It is fortuous that the radiation dominates the heat loss from the melt. The heat absorbed by the phosphorus and its temperature can therefore be easily controlled by changing the vertical location of the injector and the time it spends at selected location(s) since the temperature response to radiation heat transfer is almost instantaneous.

RESULTS AND DISCUSSION

Correlation Between Power Input and Melt Temperature. In order to effectively control a crystal growth furnace it is important to know how does the melt temperature change with a variation in power input and what is the response time. The first set of experimental runs with B_2O_3 encapsulated indium melt in the crucible was conducted to obtain this crucial information. A T/c probe attached to the seed shaft was used to measure the melt temperature at the crucible bottom. While performing this experiment, the quartz ampoule for P-injection and the heat shield were not present in the growth chamber. Figure 2 presents the measured melt temperatures at various power inputs and shows that a linear relationship exists between the two quantities. The melt temperature increases by about 26°C for an increase of 0.2 kV in the power supply, and requires about 12 minutes to stabilize. This is very interesting behavior since the physical phenomena of convection and radiation which govern the heat transfer are highly non-linear.

Effect of Magnetic field. This experiment also confirmed the previous observation [12] that the melt temperature decreases by 4-7°C when a magnetic field of 2000g is applied. This effect of magnetic field was observed by turning the magnet on and off each time the power level was changed and the melt temperature attained a stable temperature. Interestingly, when the magnet was shut off, the melt temperature did not return to the original value, but rather increased by 7-10°C. This increase in temperature was caused by an automatic, small increase in input voltage every time the magnet was turned off. Except that both the magnet and r.f. coil are supplied energy (in parallel) from a single source, there was no direct connection between these two electrical devices. The reason for this increase in supply voltage whenever the magnet is shut off is not clear at this time.

Gas Convection. Another interesting behavior observed during this experiment was a sharp drop in temperature across the indium melt and B_2O_3 layers, 123°C and 186°C, respectively.

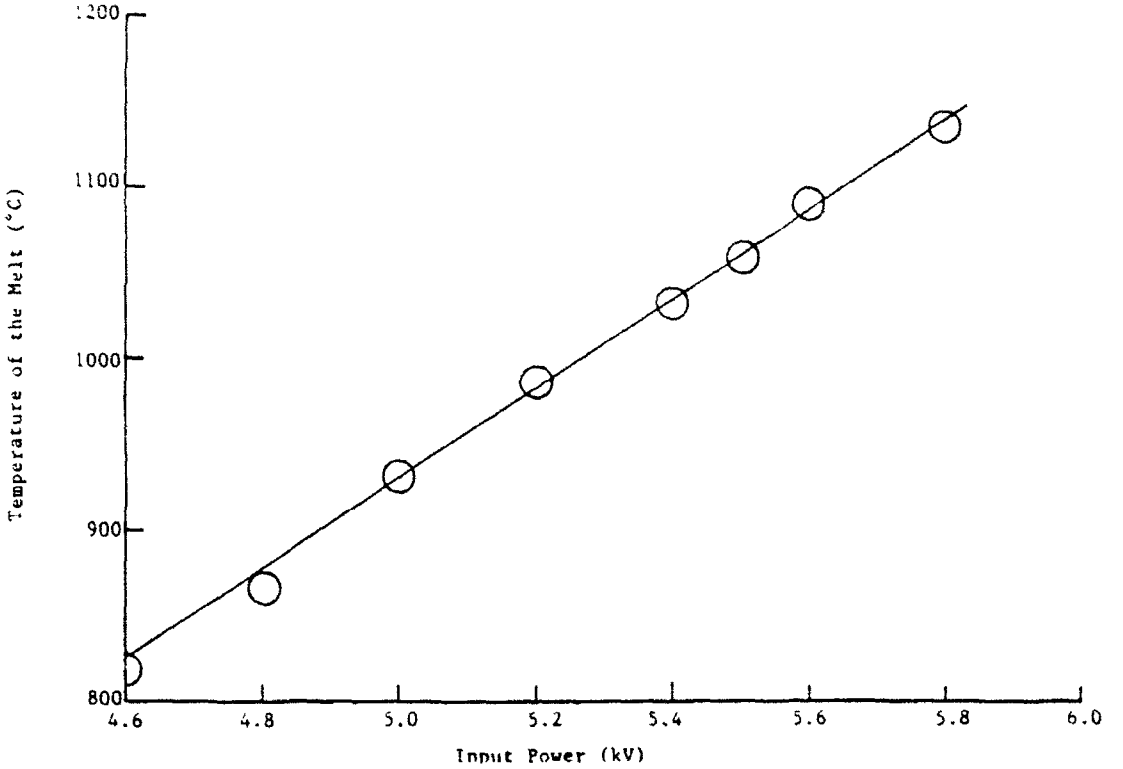


Fig. 2 Variation in Melt Temperature with Power Input.

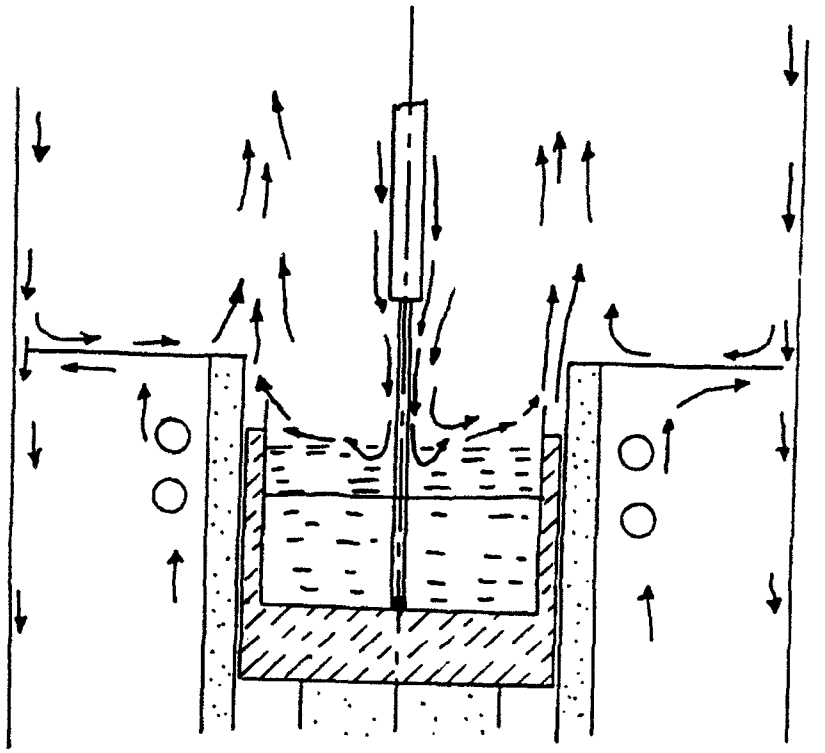


Fig. 3 Recirculatory gas flow pattern in the absence of phosphorus injector and heat shield.

Although within the encapsulant layer, a temperature drop in excess of 100°C has been reported by previous investigators [12], a temperature gradient as large as this one in the melt is unacceptable from crystal growth considerations. Indeed, the large temperature drop observed in this experiment was caused by strong convective currents of N_2 gas. A visual observation of the gas flow indicated that in the absence of P-injector and heat shield, the gas could flow much more strongly and loose more energy on the puller walls. Two primary recirculatory flows were clearly visible (Fig. 3). The downward flow along the seed shaft and T/c probe was so strong that it could create a dip around the T/c probe in the boric oxide layer. This resulted in an extremely high rate of heat transfer, and consequently, a large vertical temperature gradient in this region.

Thermal Response of Phosphorus Injector. The second set of experiment was conducted in order to examine the temperature response of the phosphorus injector with respect to its location from the melt surface. For this run the quartz ampoule was filled with crushed zirconia (SiO_2) to about 2/3rd of its volume and the crucible contained the B_2O_3 encapsulated indium phosphide. Two thermocouples, one at the injector tube tip and the other at its bottom surface, were placed to measure the temperature at two different locations. As shown in Fig. 4, these temperatures change significantly with the injector's distance from the melt surface. Indeed, the response of the thermocouples are very fast which confirm our theory that the radiation plays a dominant role in heat loss from the melt surface even in an HPCG puller.

An interesting feature of Fig. 4 is the drop in bottom temperature after remaining constant for a while even though the power input, the injector location and the temperature at the tip of the transfer tube remain constant. This behavior can be explained only by considering the heat transfer through SiO_2 layers in the injector. When the injector is moved down, its bottom gains more energy because of the increase in radiation view factor between the melt surface and the injector. Since thermal diffusion is an inherently slow process and the thermal diffusivity of SiO_2 is very low, $\alpha \cong 1.4 \times 10^{-6}$ m^2/s, very little of this energy is lost by conduction through zirconia right away. As a result, the temperature first builds up and a balance between the radiation and convection gain and the conduction loss is maintained for a significant amount of time as shown by the isothermal condition in Fig. 4. However, the temperature starts to drop off as the conduction through SiO_2 layers increases. It is expected (although not shown in Fig. 4) that the bottom surface will attain a constant value if the injector is kept stationary at a location for sufficient amount of time.

A series of experiments was then conducted with the injector filled with solid phospho-

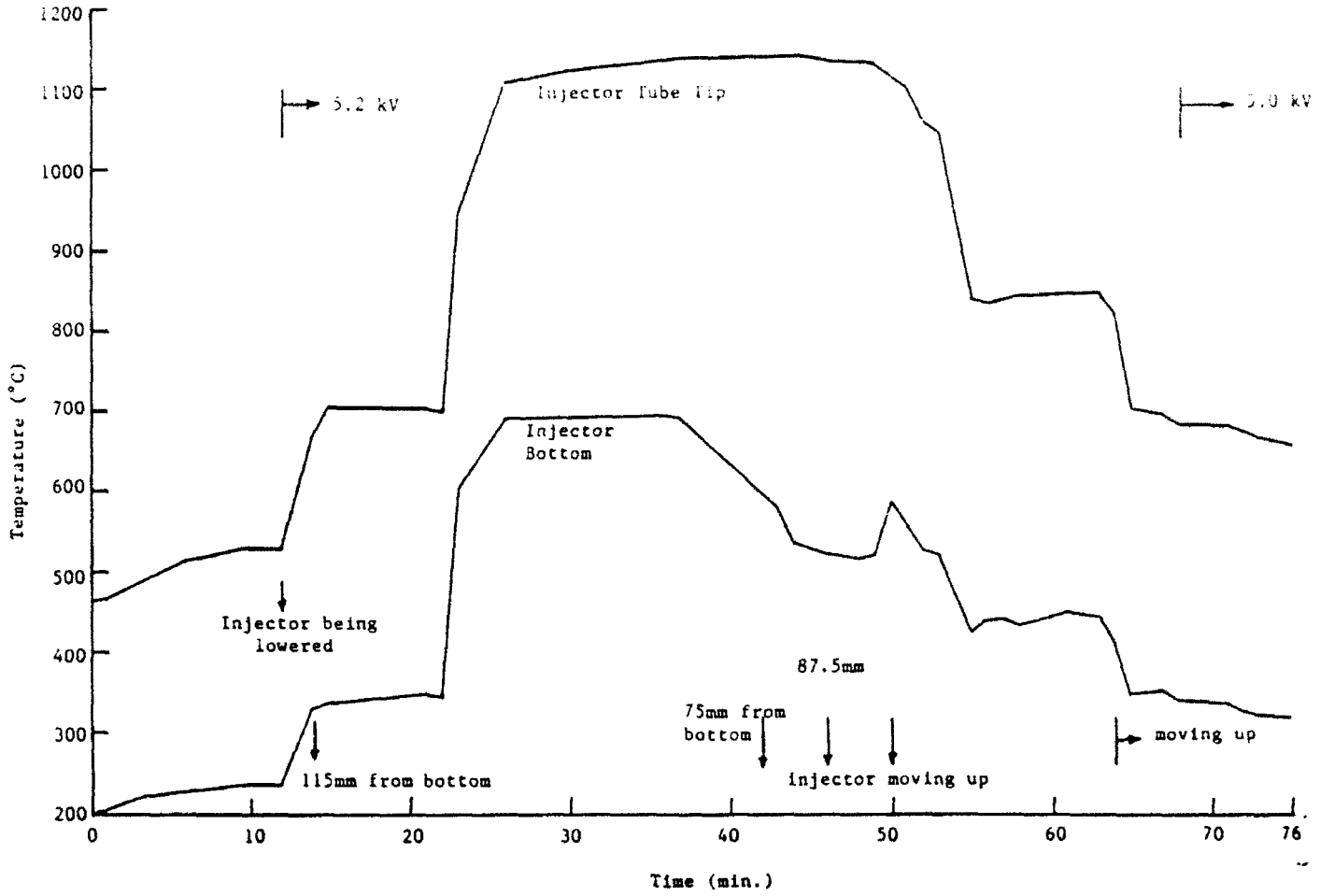


Fig. 4 Temperatures at injector bottom and the tip of the transfer tube when the injector is filled with zirconia (SiO_2).

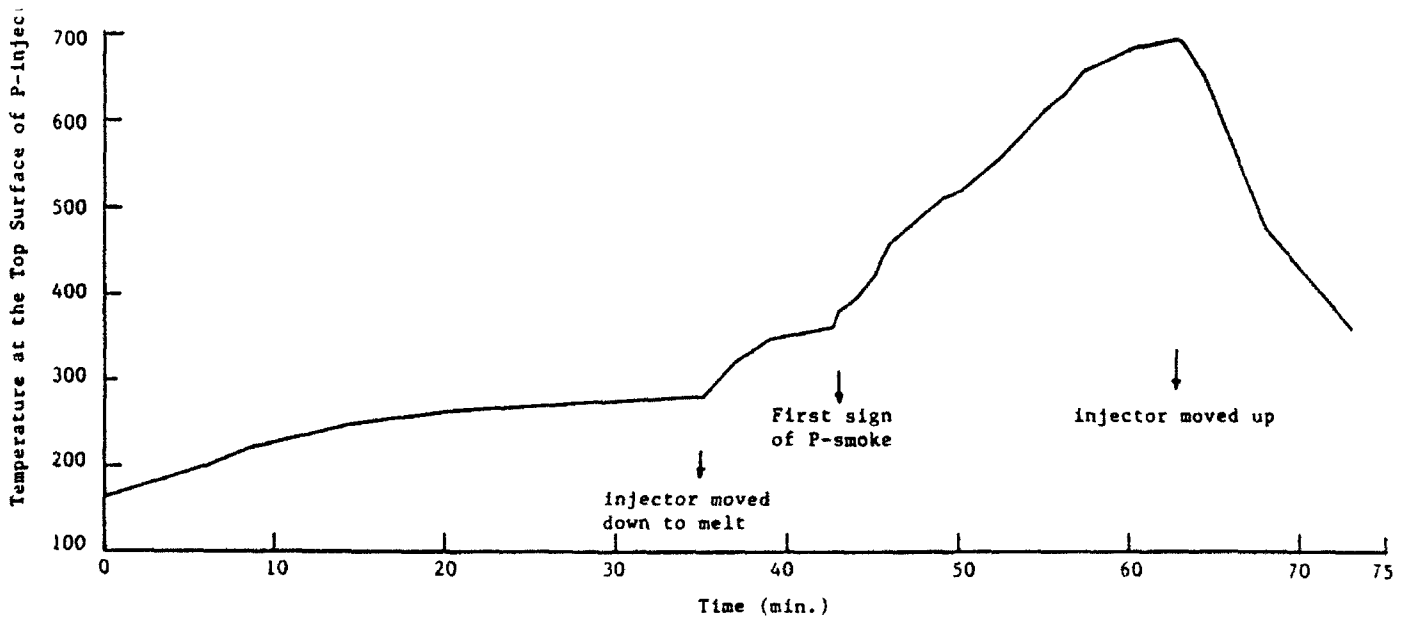


Fig. 5 Temperature at the top of the phosphorus injector during a successful synthesis run. (Crystal growth was partially successful.)

rus. Not all of these runs were successful in both synthesis of phosphorus vapor and indium melt, and the growth of InP crystals. However, they provided valuable information on thermal characteristics of the system. Figure 5 shows the response of the thermocouple at the injector top surface during a successful injection. The temperature at this location is seen to build gradually as the injector is moved down to a new location. This trend is quite expected since there is no direct thermal interaction between the melt surface and the injector top. Energy gained at the bottom must pass through the phosphorus layers where it may be used to raise the phosphorus temperature or phase-change before it reaches the top surface. The decrease in top surface temperature is also gradual when the injector is moved up. This trend is further confirmed by Fig. 6 which presents the temperature readings at three locations on phosphorus injector, the crucible bottom and of the recirculating nitrogen gas at a location outside the heat shield. Figure 6 also demonstrates that the injector bottom and top surfaces attain constant but different temperatures once the injector is moved up after injection and kept stationary at one location.

Oscillatory Nature of Injector Bottom Temperature. An interesting aspect of Fig. 6 is the oscillatory nature of injector bottom temperature when the injector is moved down and the phosphorus vapor is generated. As shown in Figs. 7 and 8 this oscillation is independent of whether the injection is successful or not. The primary reason for this oscillation seems to be the phase change of solid phosphorus at the injector bottom. When most of the energy is used in sublimation or melting of solid P, the temperature does not increase much. However, as soon as a layer of vapor is formed at the bottom, the temperature starts increasing because of the low rate of heat transfer through the vapor layer. Although the localized pressure build up is very high when the vapor is being generated and can sustain the load of the solid phosphorus, this phenomenon is highly unstable. Once a vapor layer is formed at the bottom surface, the heat transfer to solid phosphorus decreases substantially since the vapor is almost opaque and cannot allow much of the radiation to pass through it. The rate of vapor generation then decreases significantly and so does the local pressure at the bottom. As a result, the solid phosphorus drops to the bottom and the temperature starts decreasing. This alternate formation of vapor layer and its disappearance causes the temperature to fluctuate at the bottom surface. Indeed, on certain occasions, we could visually observe the fall of solid P to the bottom. It should also be noted that this phenomenon of vapor layer formation is very similar to that observed in the case of film boiling [15]. However, the film boiling experiment has generally been performed with a liquid for which the vapor is transparent, and hence, the vapor layer

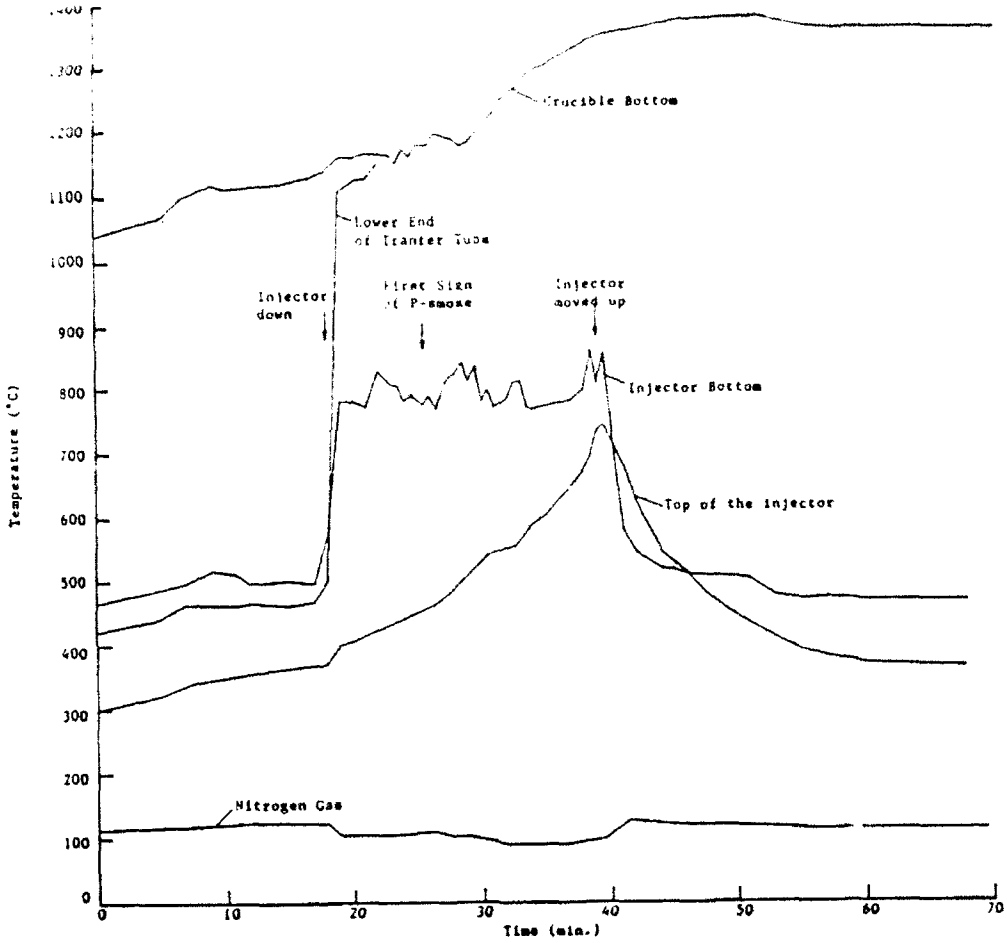


Fig. 6 Temperatures at the crucible bottom, the lower end of the transfer tube, the injector top and bottom, and nitrogen gas (at 69 mm above the glass plate and 48 mm away from the heat shield) during a successful synthesis run.

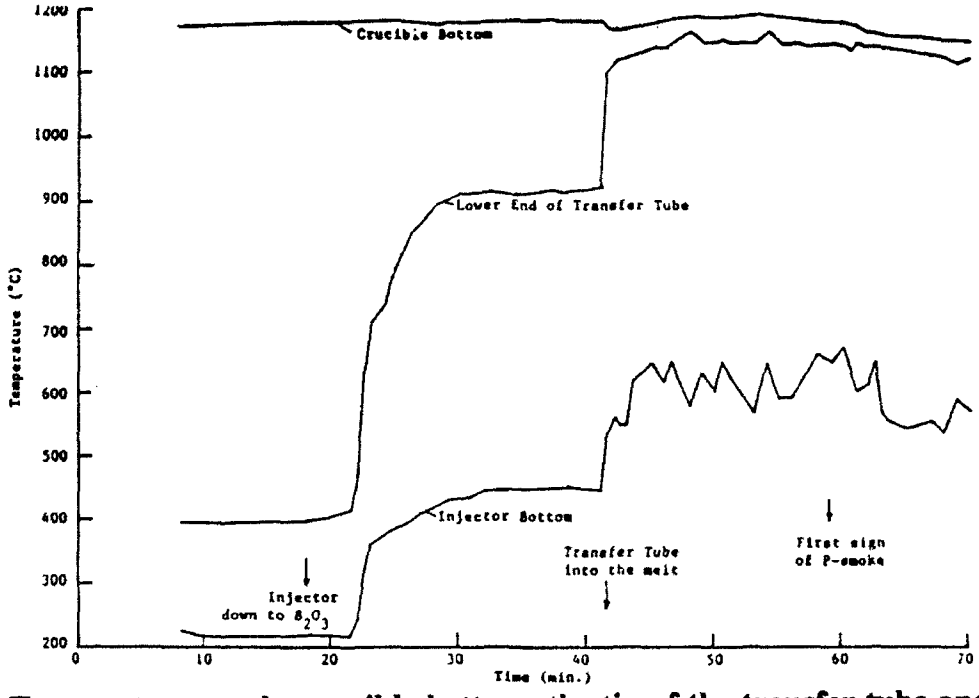


Fig. 7 Temperatures at the crucible bottom, the tip of the transfer tube and the injector bottom. This run was stopped when the transfer tube got broken.

can be sustained by a continuous liquid-vapor phase change by radiation heat transfer to the liquid at its bottom surface.

Melt Temperature. Figures 6, 7 and 8 also present the melt temperature measured by the thermocouple at the injector tube tip and the temperature at the crucible bottom which is also a measure of the melt temperature. The crucible bottom temperature is expected to be higher than the melt temperature except when the reaction is taking place. The exothermic reaction between the phosphorus vapor and indium melt generates heat thereby raising the melt temperature for the period of reaction or a little longer. Figure 7 clearly shows that in the absence of reaction, the thermocouple at the crucible bottom reads a higher temperature than at the injector tube tip. During this experiment, the injector was first lowered to a location close to the B_2O_3 encapsulant surface (barely touching). In this position, we could see the nitrogen gas bubbling at the B_2O_3 surface. Note that the injector contains a large volume (30-40%) of nitrogen which expands when the injector is moved down. The bubbling of N_2 was allowed to continue for about 23 minutes before the injector tube was dipped into the melt. Bubbles of nitrogen leaving the melt surface were still visible for a long time and only after about 16 minutes the first bubble of P-vapor could be seen leaving the melt. Soon after that the injector tube broke at its upper end and the vapor started leaking. The run was then abruptly stopped.

Figure 7 demonstrates some other interesting behavior. It is observed that when the injector tube is dipped into the melt, the crucible bottom temperature decreases slightly because the nitrogen gas bubbles which are at a lower temperature than the melt, gain energy thereby reducing the temperature of the melt. However, this is a very short time phenomenon and the temperature starts increasing after a few minutes, to be discussed later. Although this injection was not successful, the effect of localized build up of temperature due to on and off reaction (some P-vapor coming out with N_2 gas bubbles) is well demonstrated by the peaks of the melt temperature plot (Fig. 7). An increase in melt temperature due to the reaction is also exhibited by Fig. 8 for the case when the injector tube was dipped into the melt only after the first bubble of P-vapor came out of the tube. Unfortunately, the thermocouple on the injector tube could not sustain the corrosive effect of phosphorus reaction, and started giving erroneous readings whenever the reaction took place (Figs. 6 and 8).

It is believed that an increase in melt temperature is caused by

1. a reduction in radiation loss to the puller walls when the injector is moved down, caused by the reduction in radiation view factor between the melt surface and the puller walls,

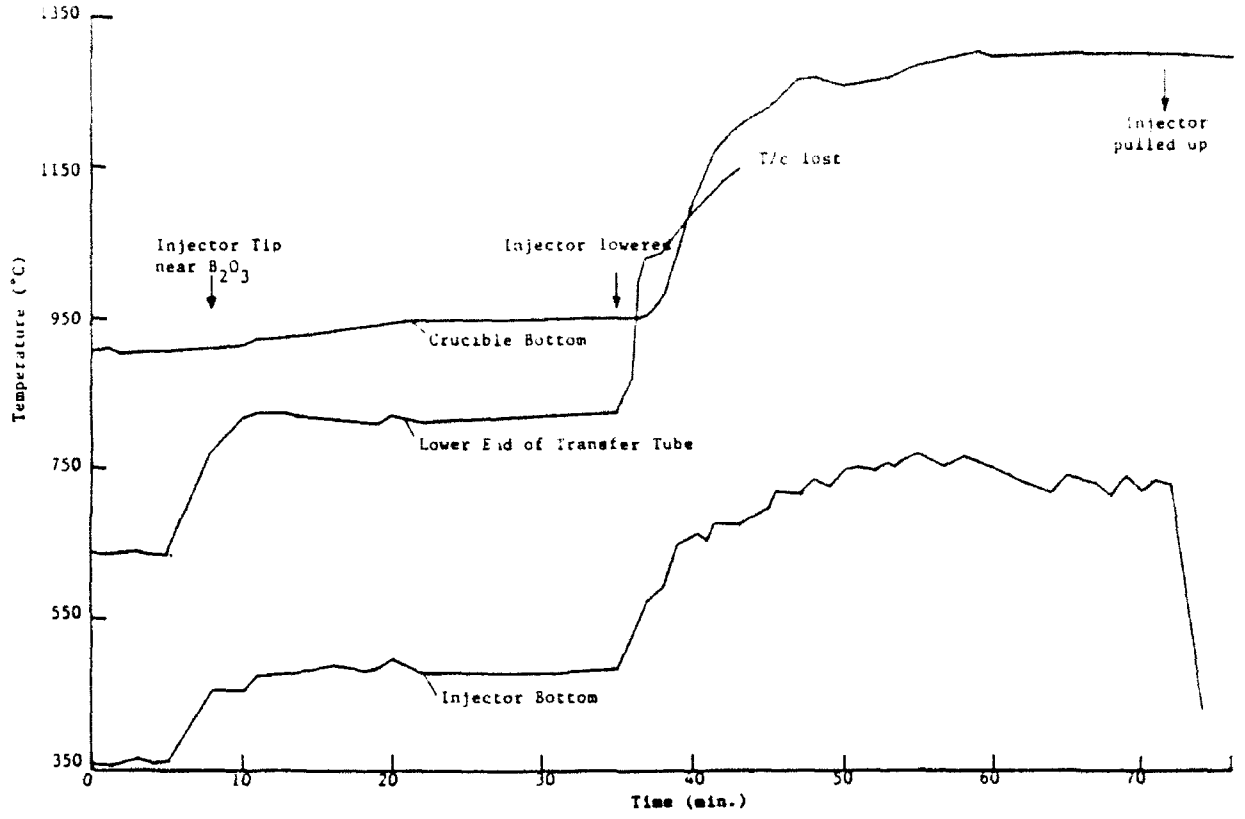


Fig. 8 Temperatures at the crucible bottom, the tip of the transfer tube and the injector bottom. This run was abandoned due to a leak in the transfer tube.

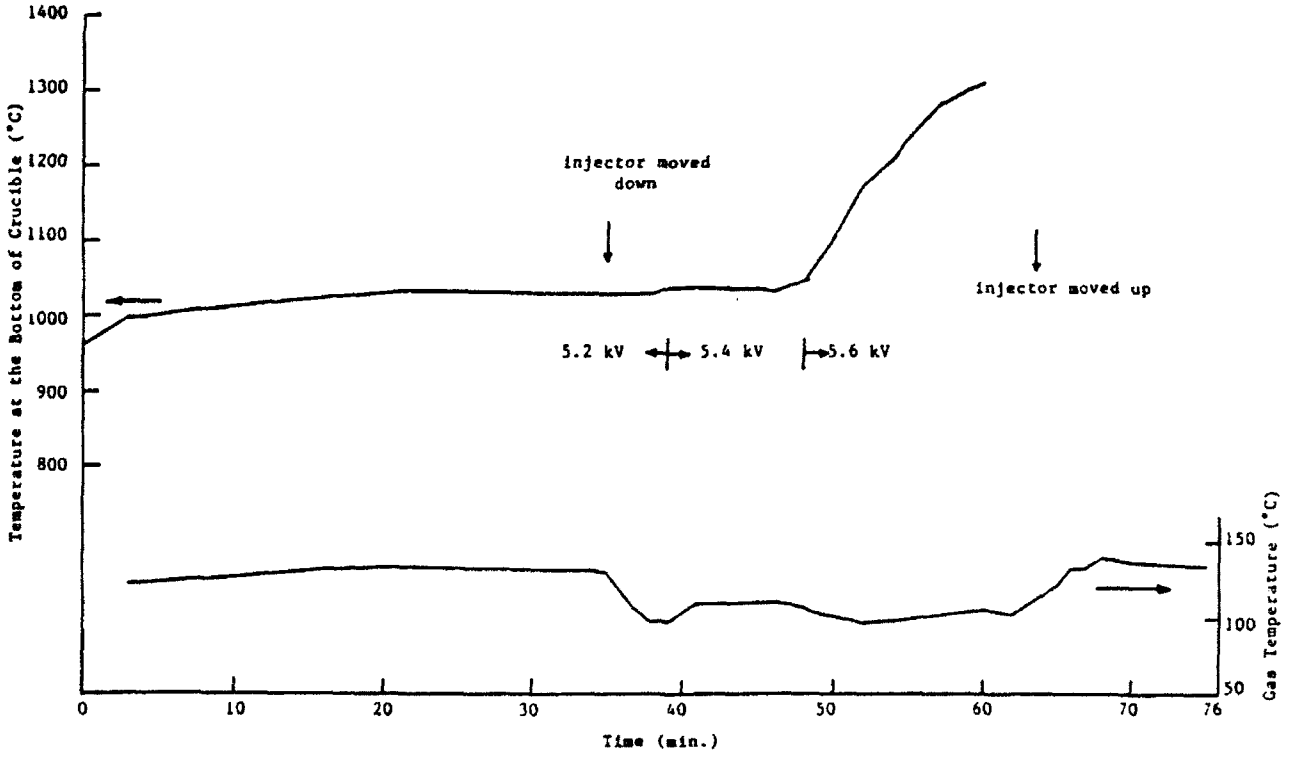


Fig. 9 Temperatures at the crucible bottom and of the N_2 gas (at 69 mm above the glass plate and 48 mm away from the heat shield) during a successful synthesis run.

2. a decrease in radiation heat transfer to solid phosphorus and puller walls whenever the non-transparent phosphorus vapor is present in the vicinity of the melt (This may be caused by the leak(s) in injector tube or vapor leaving the melt due to local super-saturation.), and

3. the exothermic reaction between the P-vapor and indium melt.

In Fig. 6 the temperature at the crucible bottom continues to increase even after the injector has been moved out of the melt, because of the presence of the P-vapor covering the melt surface. Since the phosphorus clouds never cleared during this experiment, this temperature settled at a much higher value than that before the injection. Figure 8 exhibits a similar behavior. Figure 9, on the other hand, shows a temperature build-up due to the heat of reaction, a reduction in radiation loss caused by the presence of P-vapor and an increase in power.

Gas Temperature. Figure 6 and 9 present the gas temperature at a location outside of the heat shield (476 mm below the top or 69 mm above the glass plate and 38 mm away from the puller wall or 48 mm away from the heat shield). As is evident from these figures the convection outside the melt zone is also affected by the location of the injector. When the injector bottom is near the upper end of the heat shield (2-3 mm inside) a large amount of hot gases can leave the melt zone and heat the gas external to the heat shield. (To provide a recirculation, the downward flowing gas enters the melt zone through a small gap around the cold seed shaft.) When the injector is moved down, the rate at which the hot gas leaves the melt zone decreases, and hence, the temperature outside the heat shield. The gas temperature returns to its original value (Fig. 6) when the injector is moved up and the melt comes back to its initial temperature. On the other hand, the gas temperature may be slightly higher if the melt temperature is higher than that before the injection. This shows that in spite of the convective flows being oscillatory, the system is quite stable and its behavior is predictable.

CONCLUDING REMARKS

In order to develop a successful process for one-step, in-situ synthesis of phosphorus vapor and indium melt and growth of InP crystals of large sizes, uniform properties and at lesser cost, it is extremely important to understand the mechanism of heat transfer and gas flow in a high pressure crystal growth system. Although limited to mostly synthesis conditions, the experiments reported here demonstrate interesting relationships between the

temperatures at various locations of the phosphorus injector and the melt. The phase change phenomenon for phosphorus in the bottom portion of the injector is oscillatory in nature. The gas flow is turbulent and oscillatory, and there exists two primary recirculations in this HPCG furnace. From crystal growth considerations, the convective flow should be weakened as much as possible. Fortuously, the overall heat transfer is dominated by radiation, and hence, the the thermal response of this system is quite stable and predictable. Theoretical estimates of gas convection and radiation loss from the melt surface agree qualitatively with the experimental observations.

Even though these experiments provide a much better understanding of the HPCG puller for one-step, in-situ synthesis and growth of InP crystals, they are by no means complete. Similar experiments are needed to characterize the heat transfer and gas flow under the growth conditions. A simultaneous theoretical/numerical and experimental research program will further help in determining the optimized conditions for both the synthesis and the growth, and in suggesting the modifications in the existing system in order to achieve these conditions.

ACKNOWLEDGEMENTS

The author wishes to thank J. Larkin, D. F. Bliss, J. A. Adamski and R. H. Hilton for making his summer visit to Rome Laboratory enjoyable and intellectually exciting. Acknowledgements are also due to J. Milstein of University of Massachusetts, Lowell for helpful discussions and to AFOSR for the financial support.

REFERENCES

1. J. P. Farges, *J. Crystal Growth*, 59, 665-668 (1982).
2. J. P. Farges, *2nd NATO Workshop on Materials Aspects of InP*, 1, 9.1-6 (1983).
3. G. A. Antypass, *2nd NATO Workshop on Material Aspects of InP*, 1, 8.1-9 (1983).
4. S. B. Hyder and C. J. Holloway, Jr., *J. Electron, Mater.*, 12, 575-585 (1983).
5. Y. Sasaki, J. Nakagawa and K. Kurata, *46th Aut. Meet., Jpn. Soc. Appl. Phys.* (1985).
6. S. B. Hyder and G. A. Antypass, *3rd NATO Workshop on Material Aspects of Indium Phosphide*, 1 (1986).
7. D. J. Dowling, R. A. Brunton, D. A. E. Crouch, A. J. Thompson and J. E. Wardill, *J. Crystal Growth*, 87, 37-41 (1988).

8. D. F. Bliss, R. M. Hilton and J. A. Adamski, *4th Int. Conf. on Indium Phosphide and Related Materials*, 262-265 (1992).

9. T. Inada, T. Fujii, M. Eguchi and T. Fukuda, *J. Crystal Growth*, 82, 561-565 (1987).

10. T. Inada, T. Fujii, M. Eguchi and T. Fukuda, *Appl. Phys. Lett.*, 50, 86-88 (1987).

11. K. Kohiro, K. Kainosho, H. Shimakura, T. Fukui and O. Oda, *4th Int. Conf. on Indium Phosphide and Related Materials*, B.5, 35-38 (1992).

12. S. Bachowski, D. F. Bliss, B. Ahern, R. M. Hilton, J. Adamski and D. J. Carlson, *2nd Int. Conf. on Indium Phosphide and Related Materials*, 30-34 (1990).

13. D. F. Bliss, R. M. Hilton and J. A. Adamski, *J. Crystal Growth* (submitted).

14. B. Gebhart, Y. Jaluria, R. L. Mahajan and B. Samakia, *Buoyancy-Induced Flows and Transport*, Hemisphere (1988).

15. J. P. Holman, *Heat Transfer*, McGraw Hill, 7th Edition (1992).

16. A. D. Kraus, and A. Bar-Cohen, *Thermal Analysis and Control of Electronic Equipment*, Hemisphere (1983).

THIS PAGE INTENTIONALLY LEFT BLANK

FDTD ANALYSIS OF A NOVEL ANECHOIC CHAMBER ABSORBING BOUNDARY CONDITION FOR EM SCATTERING SIMULATION

Carey M. Rappaport
Assistant Professor
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115

Final Report for:
Summer Research Program
Rome Laboratories
Hanscom Air Force Base

September 1992

ABSTRACT

During the Summer of 1992, work was performed at Rome Laboratories at Hanscom Air Force Base to develop and test a novel absorbing boundary condition (ABC), used in the finite difference time domain (FDTD) method for simulating electromagnetic wave propagation. This new type of lattice termination algorithm is based on anechoic chamber absorber foam geometry, with specially simulated electric and magnetic conductivity, and electric permittivity and magnetic permeability chosen to prevent reflections and simulate infinite, open free space. The advantage of this ABC over currently used ones is that it prevents reflections from much wider incident angles. Since incident waves need not be normal to this boundary for absorption, the boundary can be placed much closer than previously possible. This new type of ABC can be used to absorb scattered waves in a local sense, so the electrical size of the scattering object does not affect the distance from it to the ABC. Since even a modest decrease in the required separation distance yields a huge saving in the number of required matrix elements for three dimensional geometries, this novel ABC may greatly improve the general applicability of computational electromagnetics.

INTRODUCTION

In finite difference electromagnetic scattering simulation, boundary conditions must be supplied along the surfaces of the scattering objects, as well as into the radiation region. To maintain a finite computation domain, this radiation condition requires zero reflection (total absorption) at the truncated outer domain boundary.

Minimizing the amount of computational space between the scatterer and the mesh termination has long been a difficulty in numerical electromagnetics. Since the equivalence principle allows the computation of the electromagnetic field everywhere given the current on a closed surface, all that is required to solve scattering problems is to find the induced current on the

scattering object. Finding the field distribution in the region surrounding the object is unnecessary except to give the correct current on the object. It is not possible in general, however, to ignore this field, since it is not known what form it will take. For example, merely forcing the scattered field to be zero on a exterior boundary does not prevent reflections there. This is because to do so may necessitate imposing some non-zero, out-of-phase scattered field to cancel the original incident field. This canceling scattered field would represent a computational artifact.

Several types of absorbing boundary conditions (ABC's), with a variety of outer boundary geometries have been described in the literature [1-5]. There are different advantages to the various shapes of this boundary. While a circular boundary is computationally simplest, it is difficult to approximate with rectangular cells, and for long, thin scatterers large amounts of empty, uninteresting space must be included within the computation domain. A rectangular boundary may pose problems at its corners and edges.

The basic principle of each of these ABC's is to use a pseudo-differential annihilation operator on the outer elements of the grid which sets the reflected amplitude of the field components normal to the outer boundary to zero. The standard Dirichlet and Neumann boundary conditions are easy to specify by requiring the field or its normal derivative be zero at the boundary, but these each produce 100% reflection. Ensuring a matched condition with no normal reflections is much more difficult, depending strongly on the boundary geometry and material characteristics. It is an important limitation with these ABC methods that only the normal component to the boundary can be matched. The condition is specified at only the outer element, so it is compact, but inflexible. Reflections for almost normal waves are small, but non-zero, rising to appreciable values after tens of degrees.

More general ABC's which cancel waves incident from angles other than normally incident to the boundary have been proposed [3-5]. These apply approximate solutions to the wave equation at the radiation boundary, with annihilation for multiple discrete angles. Unfortunately,

for each additional angle of annihilation, the order of the differential operator increases. The number of elements in the vicinity of the boundary which must be included in the higher order difference operation thus increases. Although a wide range of incident angles can be absorbed with these ABC's, the resulting complexity at the boundary may become prohibitive.

The ABC's for scattering problems must be placed far enough away from the scatterer for all scattered rays to appear to be normally incident on the absorbing boundary. For large scattering objects, illuminated with microwave frequency radar pulses, an effective grid termination might have to be hundreds or thousands of lattice points away, making the simulation problem unreasonably big and slow to solve.

ANECHOIC CHAMBER-BASED ABC

The idea behind the anechoic chamber ABC was first described in a paper by this author appearing in the *Journal of Electromagnetic Waves and Applications* [6]. This previous work analyzed the anechoic chamber ABC using the Finite Difference Frequency Domain method of computing scattered field. The Finite Difference Time Domain (FDTD) formulation is more useful when considering a finite wave pulse with a continuous spectrum of frequency components, incident from a single angle [7,8].

The fundamental principle of the new ABC is that used with the carbon-loaded absorber foam pyramids lining the interior walls of anechoic antenna test chambers. The steeply slanted lossy material faces absorb some of the incident energy and tend to redirect any reflected waves into other pyramids for additional absorption. The net effect of the wall of pyramids is to absorb all incident waves. And since incident waves from all directions will be absorbed, the lossy pyramids work very well at preventing wall reflections in antenna test chambers. In a two dimensional analysis, which was considered in this project, the pyramids become triangular saw teeth.

With computer modeling, the material characteristics of the absorbing layer need not be those of real dielectric compounds. Instead, the material can be modeled as having both electric and magnetic conductivity σ, σ_M, and with values of permittivity ϵ_1 and permeability μ_1 selected for a perfect match at the second bounce of the normally incident wave. If the vertex angles of these triangles are equilateral, the angle the reflected ray makes with the second triangle face is 90° (0° local incidence angle). This geometry is shown in cross section in Figure 1.

As long as $\mu_0/\epsilon_0 = \mu_1/\epsilon_1 = \sigma_M/\sigma$ (which corresponds to maintaining constant frequency domain wave impedance across the boundary), there will be no reflections for normal incidence to a triangle face.

This requirement is difficult to attain with real materials, but quite easy to specify by computer. It is important to make the conductivities large enough so that the wave quickly decays as it propagates into the absorber medium, but not so large that the decaying field is inadequately sampled on the mesh.

The reflected power is plotted as a function of the angle of incidence of the ray with the boundary α in Figure 2. This is a geometric optics, frequency domain prediction of the reflection characteristics of this boundary. For this test case, $\epsilon_1 = \epsilon_0(1 - j0.01)$ and $\mu_1 = \mu_0(1 - j0.01)$. The discontinuities are the result of calculating the worst general ray for each of the cases. The highest reflected power from this boundary in the entire 180° angular range occurs at $-30°$, with a value of 2.8×10^{-6}, or about -56 dB. The discontinuity occurs at the angle where the ray bouncing off the first triangle edge just misses intersecting the second triangle edge.

This ray analysis completely neglects the effects of diffraction from the vertices of the triangles. The frequency domain analysis [6] showed that this periodic diffraction is a strong effect, especially when the vertices were separated by an integer number of wavelengths.

FDTD SIMULATION

Two dimensional FDTD simulation of this equilateral triangle saw-tooth absorbing layer is examined using a modulated uniform Gaussian pulse plane wave, given by:

$$\overline{E}(x) = \hat{z} E_0 e^{-(\frac{x - x_0 - ct}{W})^2}$$

for normally incident illumination, and

$$\overline{E}(x) = \hat{z} E_0 e^{-(\frac{(x - x_0)\cos\theta + y \sin\theta - ct}{W})^2}$$

for waves incident at angle θ. The magnetic field components, which follow directly from Faraday's law, have the exact same space and time dependence, but are orthogonal to $\overline{E}$ and are reduced in magnitude by the material impedance $\eta = 377\Omega$. This equation is discretized by using $x \rightarrow i\Delta, y \rightarrow j\Delta,\ t \rightarrow n\Delta t,\ W = 20\Delta$. The Gaussian is used since it smoothly approximates a short, causal square pulse in time and space. The computational domain is chosen as a rectangular grid of points $(i, j),\ 0 < i \leq i_{max},\ 0 < j \leq j_{max}$.

Care must be taken at the edges of the computational domain to simulate the incident plane wave's infinite extent. If the wave propagates parallel to the left and right edge boundaries, one good way of simulating infinite extent is to impose a one-dimensional FDTD wave solution on each boundary. This is because the one-dimensional wave only propagates along the boundary, rather than in to or out from it, while the incident two-dimensional wave only propagates in that same direction. Once the incident wave interacts with a scatterer or ABC, or if it propagates obliquely to the side edges, the one-dimensional FDTD is insufficient, and another method—such the Engquist-Majda or even the currently proposed anechoic sawtooth ABC—must be used. The back of the lattice at $i = i_{max}$, is terminated with the usual Engquist-Majda ABC to prevent any waves from returning and confusing the pulse pattern in the lossy medium.

Another important consideration with oblique incidence is to ensure that the entire wavefront propagates across the domain of interest. For example, if the pulse propagates from the southeast toward the northwest, it is essential that the bulk of the peak intensity starts at analysis time $t = 0$ near the northwest corner, past the southwest/northeast diagonal. Otherwise, none of the expected energy would enter the domain from outside the eastern and southern edges.

The calculations are performed with normalized electrical parameters, μ and ϵ, with the only requirements being that $\sqrt{\mu/\epsilon} = \eta = |E|/|H|$, and normalization $\sqrt{\mu\epsilon} = c = 1$ for simplicity. Since for this time-domain analysis, the incident wave is baseband, with no frequency modulation, the time scale is arbitrary, and there is no frequency scale. The only physical scale is the relative lengths of the sawtooth edge and the half-width, W of the Gaussian pulse. The conductivity is specified in terms of the dielectric relaxation time $\tau = \epsilon/\sigma$. Thus, for a wave to decay by a factor of e in 20 time steps, $\sigma = \epsilon/20$. Also, the Courant condition, which specifies the relative sizes of time and space steps in the FDTD algorithm, $c\Delta t/\Delta z$ is chosen to be 0.5, implying that the wave advances $n/2$ steps in space in an interval of n time steps.

COMPUTATIONAL RESULTS

The first FDTD anechoic sawtooth ABC computation is for a normally incident wave, approaching the free-spaced-matched ABC situated parallel to the wavefront at either two-thirds or one-half of the forward direction domain, $i = 2i_{max}/3$ or $i_{max}/2$. To ensure that the one-dimensional wave equation at the lattice edges does not produce spurious reflections, the boundary from free space to the lossy ABC medium is specified as planar instead of sawtooth at the edges (and for a few grid cells in toward the center). Since the lossy ABC medium is matched to free space for normal incidence, their are no reflections at the two edges of the lattice. At other incidence angles there will be significant reflections from this planar region.

Figure 3 depicts a series of "snap-shots" in time for a lattice with width $j_{max} = 200$ by depth $i_{max} = 150$ grid points. As the Gaussian pulse propagates into the page, it just begins to intersect the boundary at the top left at 80 time steps. At 140 time steps, top right, the pulse has entered the lossy medium occupying the rear one-third of the lattice, attenuated on the left and right edges to about one-half its original amplitude, and begun interacting with the sawtooth structure (ABC proper) in the center of the boundary. The ten vertices of the sawteeth are discernible by slight dips in the field intensity across the wavefront. Another 60 time steps later, middle left, the pulse has entirely penetrated the boundary, its intensity is now down to 12 to 15% of the original amplitude, the sawtooth structure effect is clearly visible, and the first evidence of reflection from the sawtooth ABC, with a negative amplitude pulse returning from the boundary, is apparent. Note that there is never any reflection at the perfectly matched left and right edges. At time $250\Delta t$, middle right, the pulse is clearly being absorbed by the Engquist-Majda ABC at the deepest row of the lattice, with amplitude now at about 5%. With maximum scale enhancement, the reflected wave is the dominant feature in the final view, at 300 time steps, bottom, with intensity of -2% of the original signal. It is also seen that this reflected pulse is roughly the same width as the incident Gaussian, but appears to be more tightly bounded from left to right—indicating constructive interference effects from the ends of the sawtooth. For an infinitely wide ABC the reflected wave would be smaller, since there would be none of this edge diffraction interference.

For non-normal incidence, the pulse is assumed to originate from the right-front (southeast), and propagate to the left-back. The sawtooth ABC is specified at $i = i_{max}/2$ and also on the left at $j = j_{max}/3$. This lossy medium geometry is shown in Figure 4.

For waves incident at 45°, Figure 5 shows the progression of snap-shots for time steps: 0, 100, 200, 300, 400, and $600\Delta t$. The symmetry of interaction from the forward and left boundaries is clear. There is a small amount of constructive reflective interference near the point where the two arrays of sawteeth intersect. This is due to small specular reflection off each boundary. However, since these reflections leave at 45°, they head directly for the adjacent sawtooth

array, at the same (negative) incidence angle, and will be absorbed on their second pass. The reflected wave amplitude is about 5% of the incident peak.

Figure 6 is similar to the preceding figure, with 60°, incidence angle, and Figure 7 shows the 75°, incidence case, both with the same time steps as Figure 5, but with proportionally greater spatial grid points in the forward direction to accomadate the entire wavefront. For both these reflection examples, the reflected, or scattered field amplitude is not much greater than it is for normal incidence, with a maximum of about 5%.

Figure 8 shows a comparison of the novel anechoic sawtooth ABC a), with the common Engquist-Majda ABC b) at $t = 210\Delta t$ along the lattice line, $i = 2i_{max}/3$. The latter begins to break down for larger incidence angles. As the incidence angle approaches grazing, almost none of the wave is absorbed. Unlike with the three previous figures, Figure 8 uses the standard Engquist-Majda ABC along the left edge (for both plots) to emphasize the effects due to just the novel ABC on the back boundary. Since the incident angle for the left edge is 30°, the standard ABC is marginally acceptable, but still produces errors. The large amplitude wave on the left in both plots is the remainder of the incident pulse interaction with this left edge Engquist-Majda ABC. The wide positive bulge on the right half of the b) plot is the most important feature of this figure, indicating a significant reflected field, which is in large measure absent in the novel ABC view a). Although the Enqguist-Majda is better for near-normal incidence—as seen from the minimal reflected field at the left boundary, where the incident angle is 30°—at large angles the novel ABC is superior.

Figure 9 shows another test geometry, where the lossy medium with the sawtooth array is on a semi-circle. For any plane-wave pulse, the sawteeth are illuminated with waves from all possible angles. Once again, as with the normal incidence case of Figure 3, there are planar boundaries at each side of the lattice. The pulse begins to interact with the boundary in the first plot, top left, of Figure 10, then is transmitted into the lossy regions on the left and right sides of the semi-circle, and also propagates into the center of the semi circular area in

the top right plot, at $t = 200\Delta t$. The third snap shot, at 240 time steps shows the beginning of negative amplitude field diffracted from the corners of the semi-circle. These coalesce at the fourth view, 80 time steps later. A second reflection becomes apparent at 360 time steps with a positive focused peak, which then propagates towards the original source a times 440, 520, and $600\Delta t$. Although this reflected wave appears to be a large, dominant effect, its amplitude is only about 8% of the original signal, corresponding to 160 times less power than the incident pulse. It may be possible to further reduce this reflected signal by altering the triangle side length with respect to the incident pulse width or semi-circle radius. Continuing research will explore these improvements.

CONCLUSIONS

An improved absorbing boundary condition based on pyramidal anechoic chamber absorber foam prevents reflections from a wide range of incidence angles, and hence could be positioned very close to electromagnetic scatterers. The reduction of unimportant computational space would lead to great savings of computer memory and CPU time, perhaps even allowing the calculation of heretofore "unsolvably big" scattering problems.

Time domain analysis indicates some slight reflected amplitude from the ABC at all angles, but that there is only a small increase in the reflected field for large angles of incidence. The novel ABC is superior to the Engquist-Majda ABC for large incidence angles. Varying the number of grid points per triangle side has less effect on the absorption characteristics of the ABC than varying the pulse width, which in turn matters less than the relative conductivity of the lossy medium. As the σ/ϵ ratio increases, the decay rate in the medium increases, but the reflection coefficient at oblique angles increases also. Future research will be devoted to determining the optimum choices of conductivity and triangle side dimension.

REFERENCES

1. Bayliss, A., Gunzburger, M., and Turkell, E., "Boundary Conditions for the Numerical Solution of Elliptic Equations in Exterior Regions," *SIAM Journal of Applied Mathematics*, vol. 42, 1982, pp. 430-451.

2. Engquist, B., and Majda, A., "Absorbing Boundary Conditions for the Numerical Simulation of Waves," *Mathematical Computation*, vol. 31, 1977, pp. 629-651.

3. Lee, C., Shin, R., Kong, J., and McCartin, B.J., "Absorbing Boundary Conditions on Circular and Elliptic Boundaries," *PIERS Symposium Proceedings*, July 1989, pp. 317-318.

4. Lindman, E., "Free-Space Boundary Conditions for the Time Dependent Wave Equation," *Journal of Computational Physics*, vol. 18, 1975, pp. 66-78.

5. Higdon, R., "Numerical Absorbing Boundary Conditions for the Wave Equation," *Math. Comp.*, vol. 49, 1987, pp.65-90.

6. Rappaport, C. and Bahrmasel, L., "An Absorbing Boundary Condition Based on Anechoic Absorber for EM Scattering Computation," In review for publication in *Journal of Electromagnetic Waves and Applications*.

7. Yee, K.S., "Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media", *IEEE Transactions on Antennas and Propagation*, vol. AP-14, 1966, pp. 302-307.

8. Taflove, A., Umashankar, K., "The Finite-Difference Time-Domain (FDTD) Method for Electromagnetic Scattering and Interaction Problems," *Journal of Electromagnetic Waves and Applications*, vol. 1, 1987, pp. 243-267.

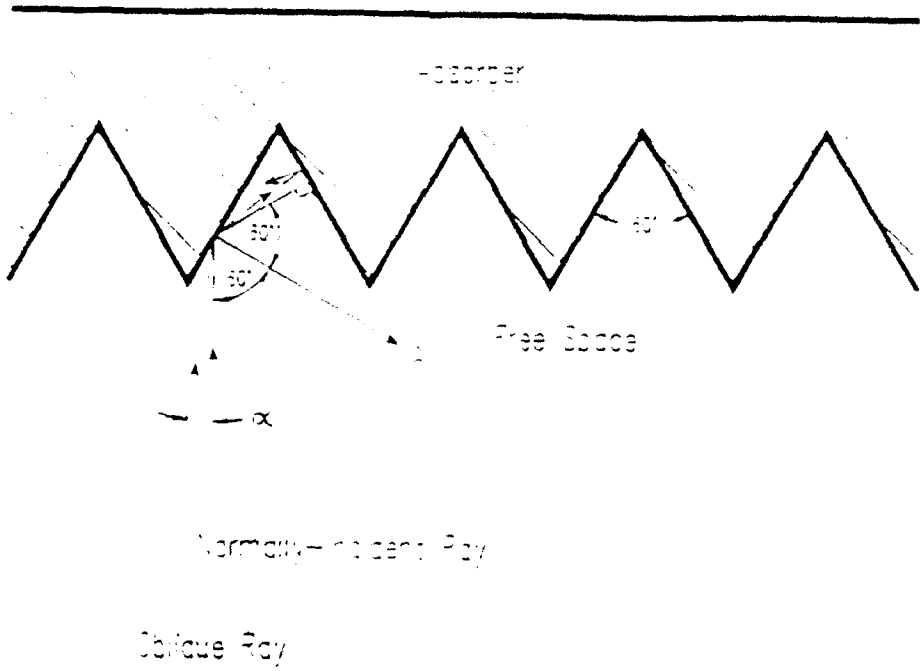


Figure 1: Ray paths for waves normally incident and incident at angle α on equilateral triangular absorbing boundary, showing multiple bounces into medium.

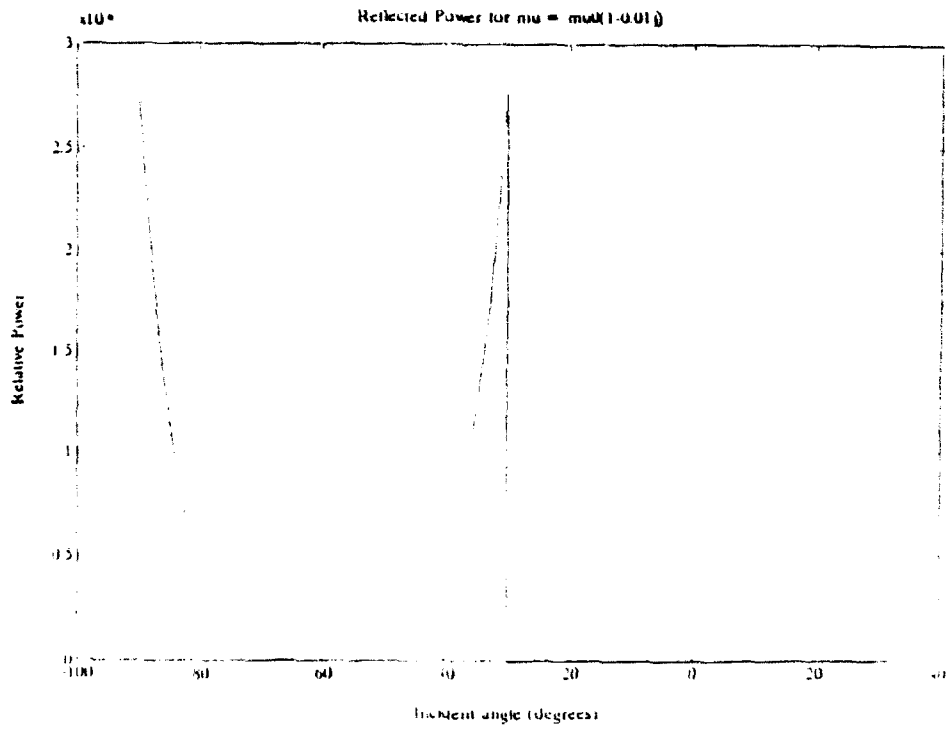


Figure 2: Reflected power ratio for individual rays as a function of incident angle to boundary.

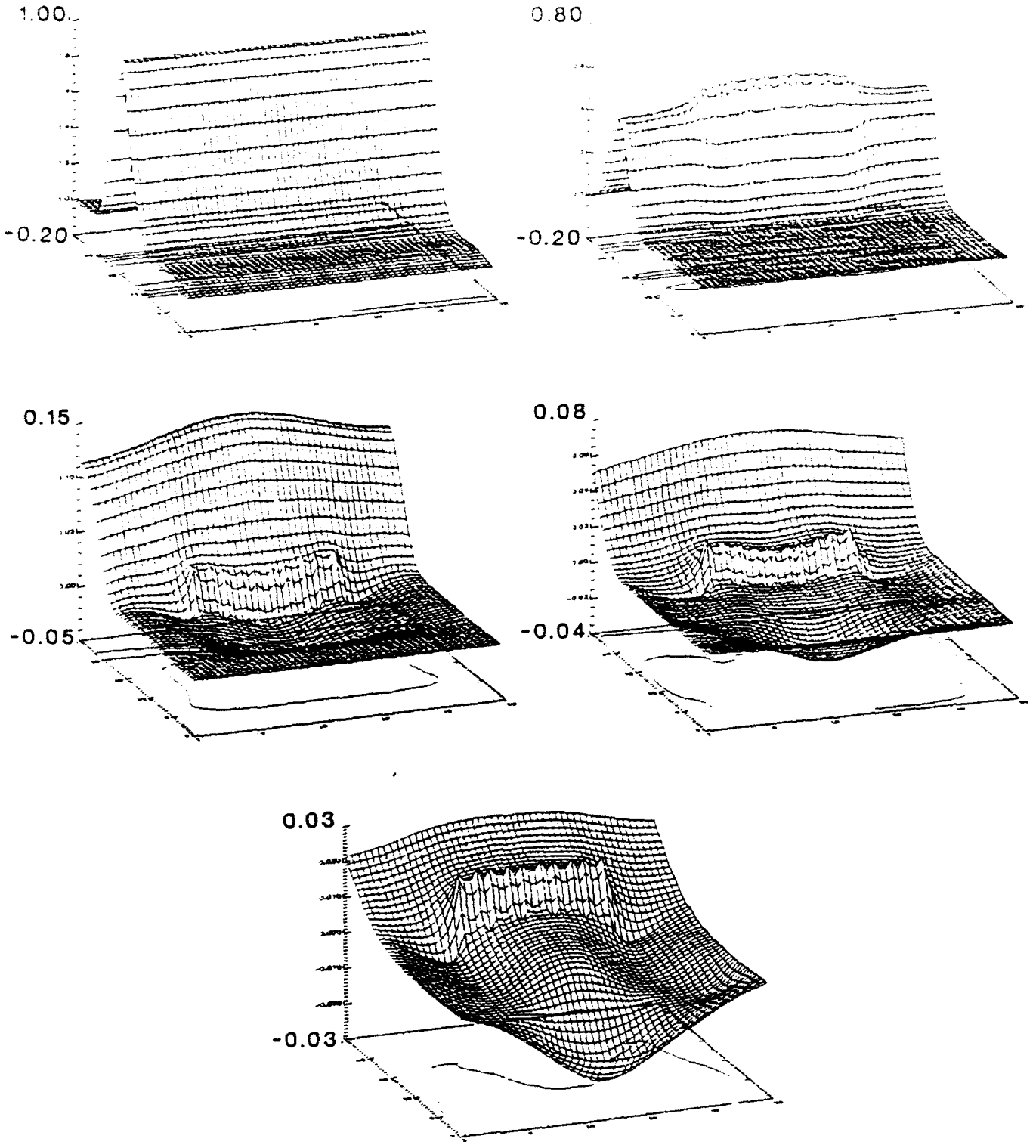


Figure 3: Normally incident Gaussian pulse approaching and interacting with the anechoic sawtooth ABC at $i = 2i_{max}/3$.

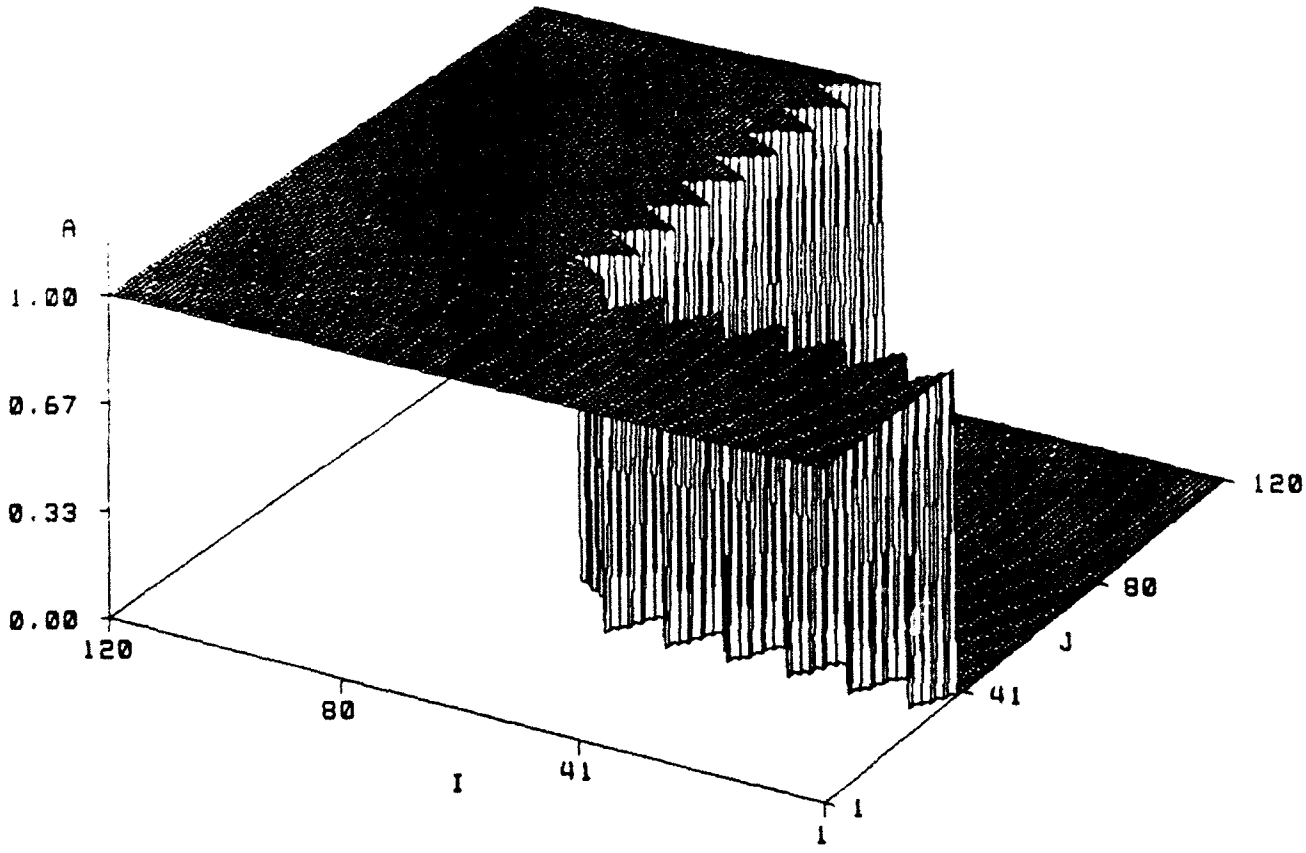


Figure 4: Geometry of anechoic sawtooth ABCs on forward and left boundaries.

Raised area represents matched dissipative medium.

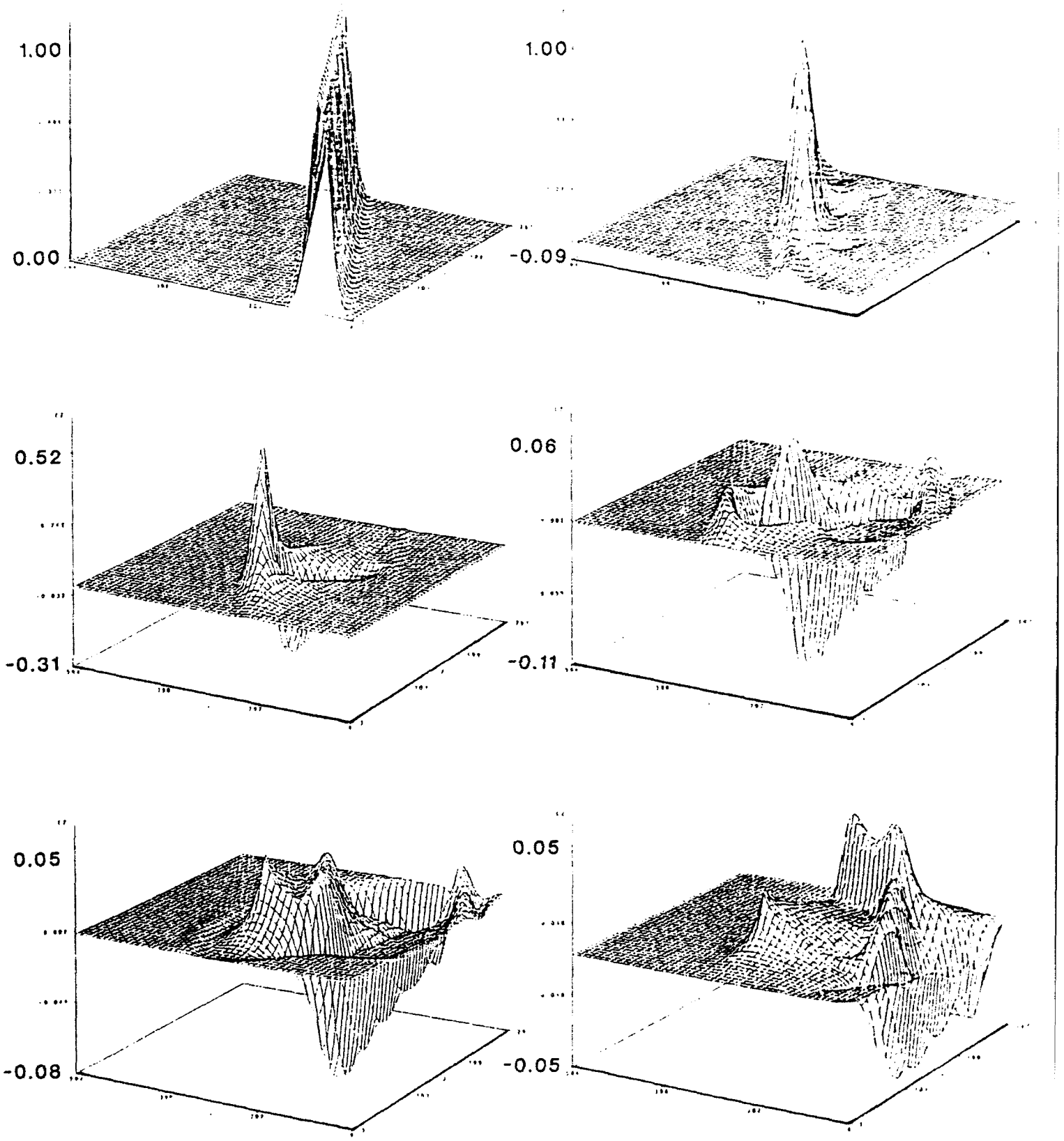


Figure 5: Gaussian pulse incident at 45° on the sawtooth ABC.

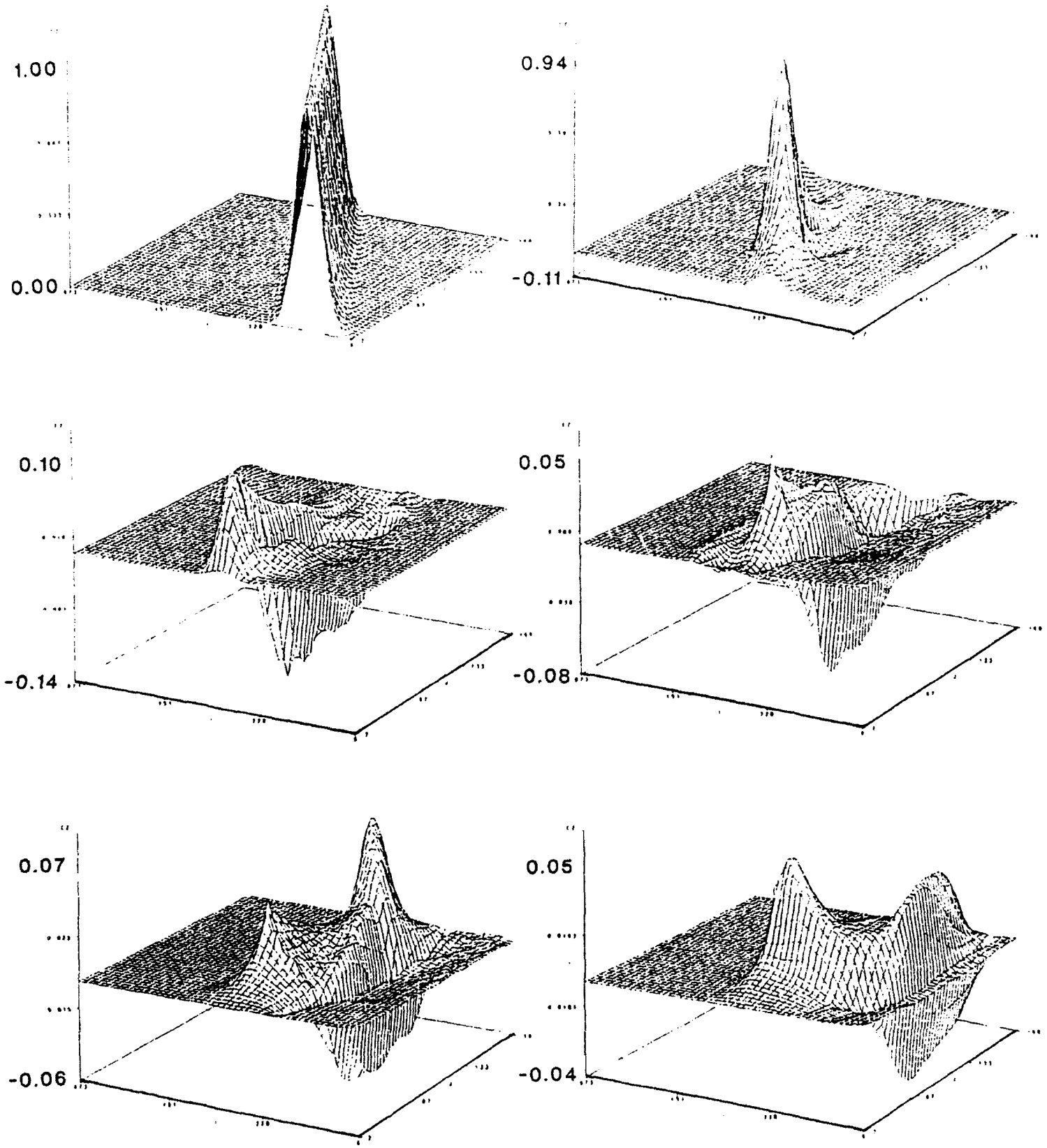


Figure 6: Gaussian pulse incident at 60° on the sawtooth ABC.

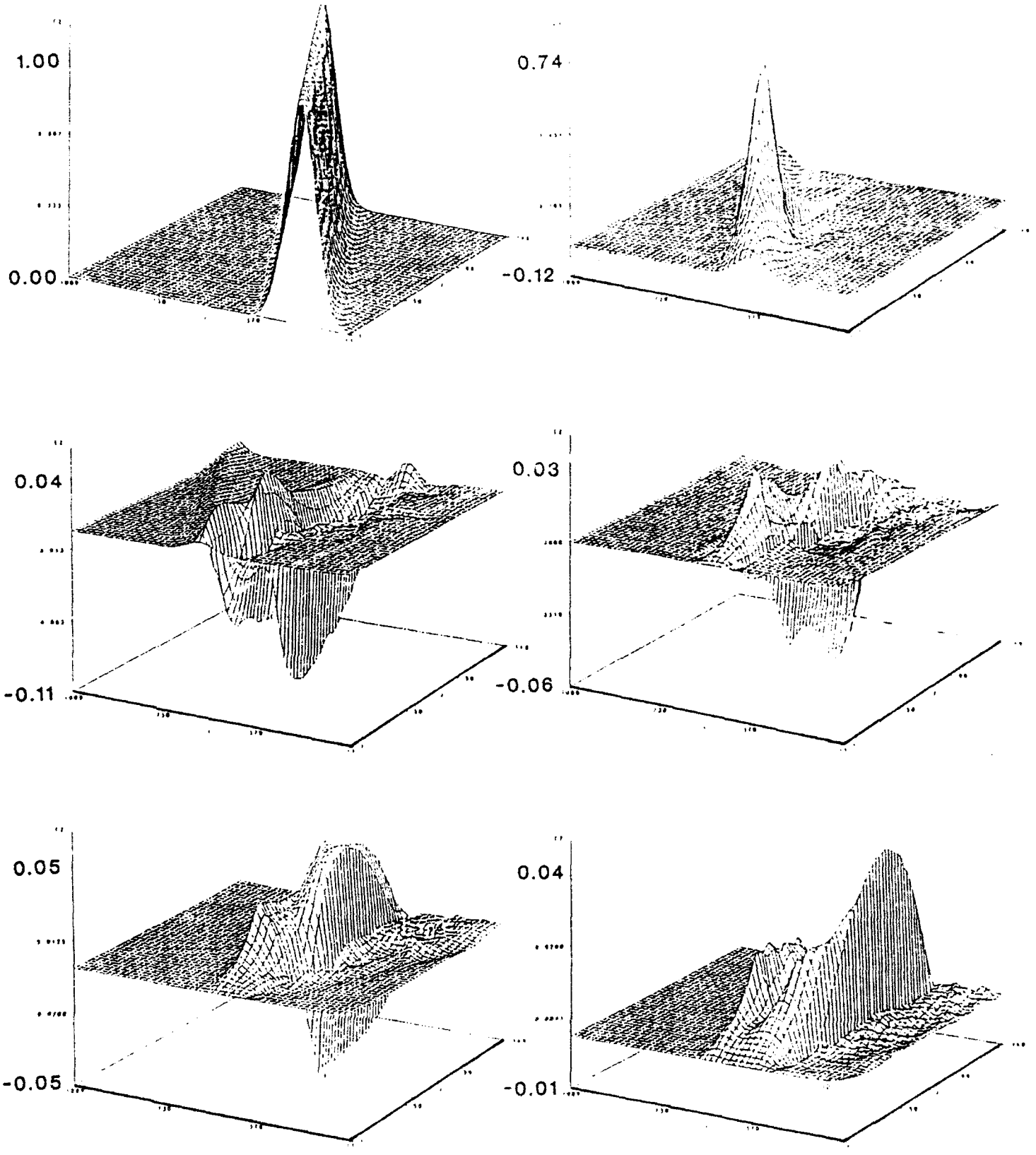
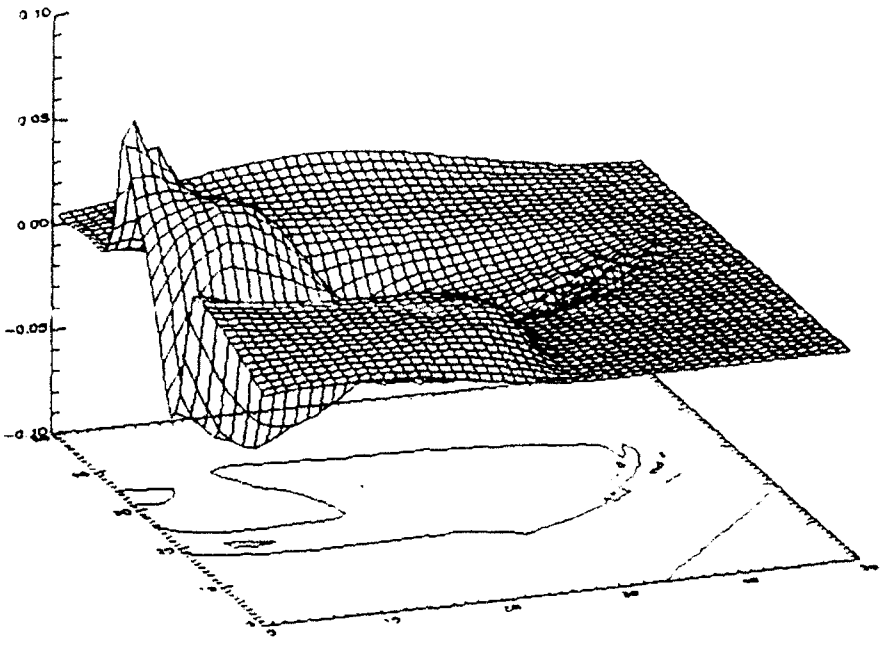
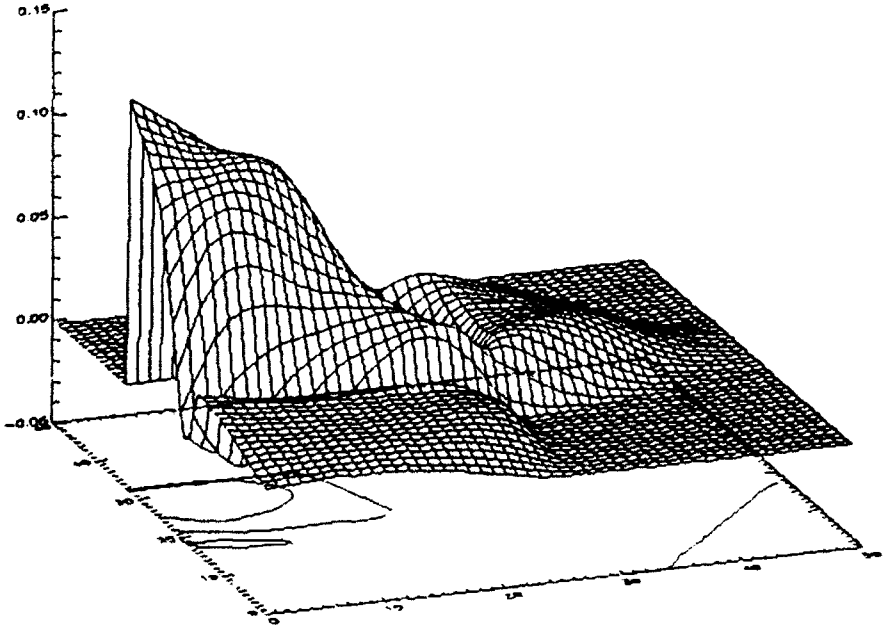


Figure 7: Gaussian pulse incident at 75° on the sawtooth ABC.



a)



b)

Figure 8: Gaussian pulse incident at 60° on a) the novel ABC and b) the Engquist-Majda ABC at 210 Δt.

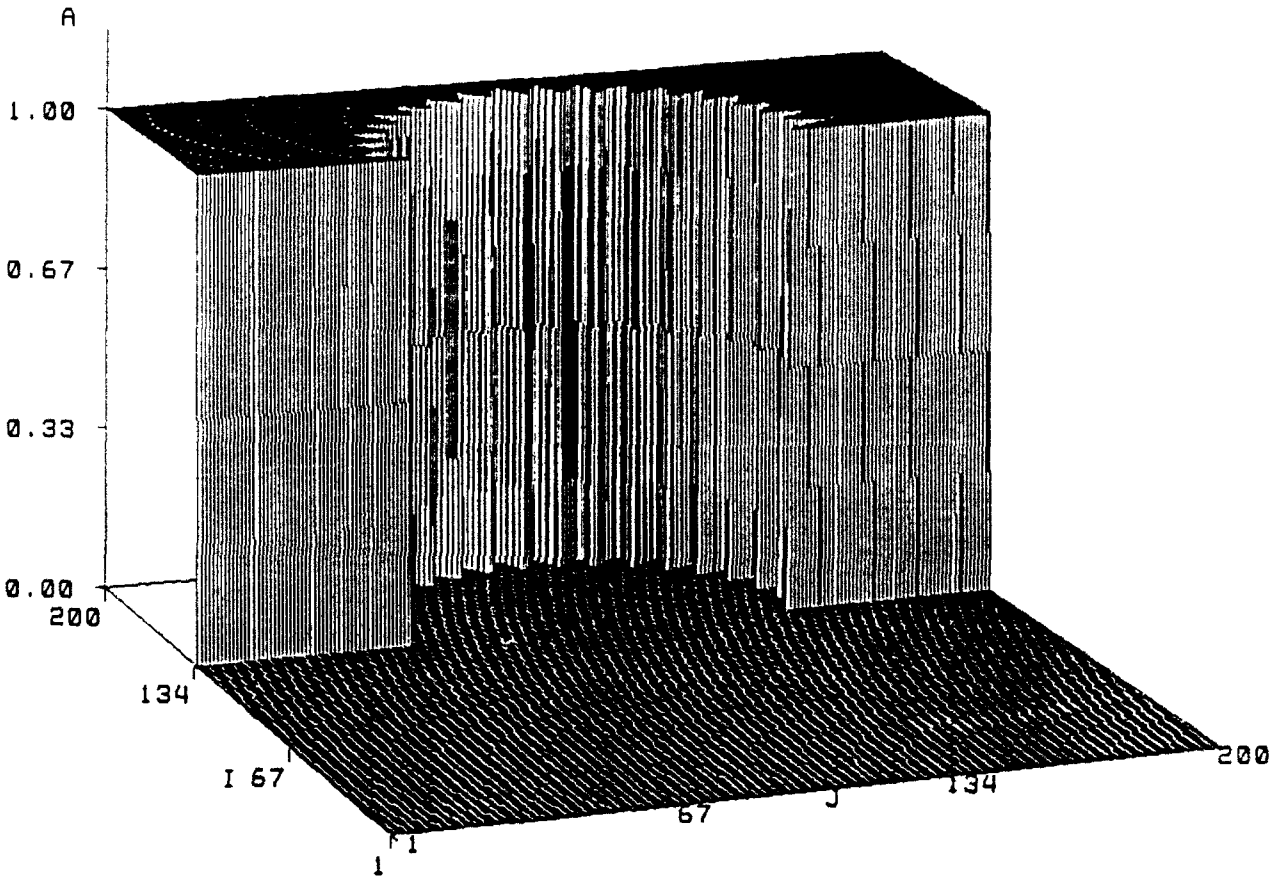


Figure 9: Geometry of semicircular sawtooth ABC boundary.

Raised area represents matched dissipative medium.

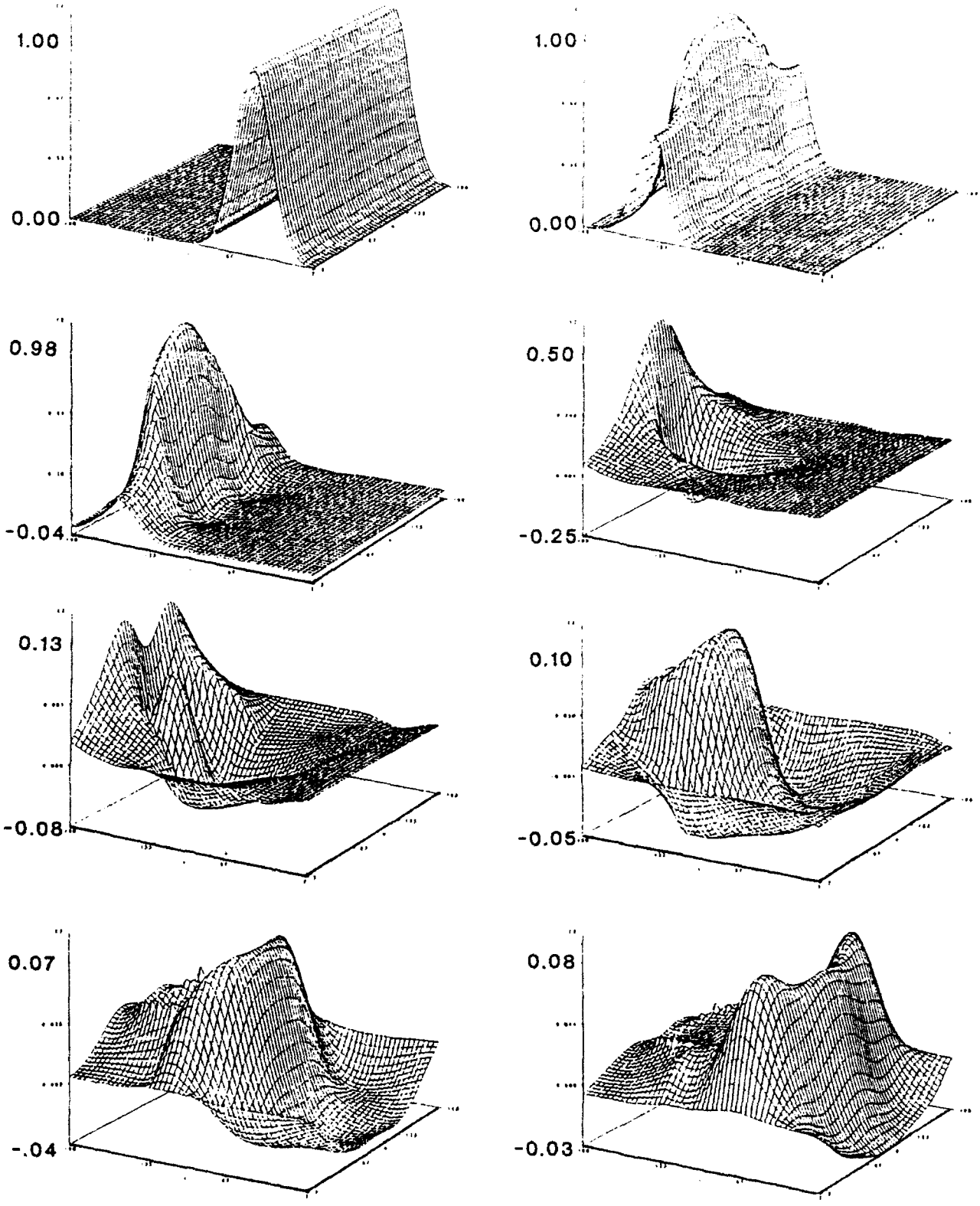


Figure 10: Gaussian pulse normally incident on a semicircular sawtooth ABC array.

X-BAND T/R MODULE
CONDUCTED INTERFERENCE SIMULATION AND MEASUREMENT
Final Report

John P. Rohrbaugh
Senior Research Engineer

and

Randall H. Pursley
Graduate Research Assistant

Georgia Institute of Technology
Georgia Tech Research Institute
Atlanta, GA 30332

Final Report for:

Summer Research Program
Rome Laboratory

Sponsored by:

Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

June 1992

X-BAND T/R MODULE
CONDUCTED INTERFERENCE SIMULATION AND MEASUREMENT
Final Report

John P. Rohrbaugh
Senior Research Engineer
and
Randall H. Pursley
Graduate Research Assistant
Georgia Institute of Technology
Georgia Tech Research Institute
Atlanta, GA 30332

Abstract

Conducted electromagnetic (EM) interference measurements and analyses were performed on X-band Transmit/Receive (T/R) modules built by Raytheon and Texas Instruments. The T/R module's Clock, Mode[1], +5 and -7 volt dc supply input lines and the Output Built-in-Test and Evaluation (OBITE) line were evaluated. The Clock and Mode differential input pins are connected within the T/R module to a CMOS gate array through DS8820/7820 differential line receivers. EM interference effects were simulated using PSPICE, and verified through measurements, to determine if the model of a DS8820/7820 provides accurate EM interference simulation results at very high frequencies. The objectives of performing measurements and simulations on the DS8820 were to demonstrate that interference effects can accurately be determined on simpler devices and models prior to developing more complex and costly products, such as T/R modules.

Limited simulations were also performed on the OBITE driver IC (54ALS03 NAND gate) and Power Condition Monitoring (PCM) circuits that are connected to the +5 and -7 volt dc supply lines. The PCM circuits are used to monitor over-voltage conditions on the +5 supply and over-temperature on the transmit power amplifier and to disable receive and transmit modes in the event of over-voltage or over-temperature conditions occur. All interference effects, with the exception of receiver low noise amplifier (LNA) gain compression, could be simulated. Effects that were duplicated during simulation included Mode words not received properly by the T/R module, and the T/R module receiver LNA cycling off and on with the application and removal of interference to the OBITE, +5 and -7 volt supply lines.

Damage effects that were observed while performing the interference mesurements could not be simulated. Two T/R modules and two DS7820 IC's were damaged over the course of this effort.

---

[1] The Data lines were not tested on this effort. The Mode lines are used to send commands to the T/R module to place the module in transmit, receive, etc., mode-of-operation, hence the nomenclature Mode. The Data lines are used to place the T/R module in a particular state-of-operation once the mode-of-operation has been selected. The default state-of-operation was selected, and thus the reason for not testing the Data lines.

X-BAND T/R MODULE
CONDUCTED INTERFERENCE SIMULATION AND MEASUREMENT
Final Report

John P. Rohrbaugh and Randall H. Pursley

INTRODUCTION

The primary objectives of this effort were to determine the conducted electromagnetic (EM) interference characteristics of X-band Transmit/Receive (T/R) modules built on a previous Rome Laboratory funded effort by Raytheon and Texas Instruments [1] and to determine if EM effects could be evaluated using simpler devices and models prior to the development of more complex and costly products, such as T/R modules. Figure 1 presents a block diagram of the X-band T/R module. Interference signals from 1 MHz to 3 GHz were combined with functional test signals and applied to selected T/R module input pins. The Clock, Mode, and +5 and -7 volt dc supply input lines and the Output Built-in-Test and Evaluation (OBITE) line were evaluated. The +10 volt supply and Data pins were not evaluated since the T/R modules were only tested in the receive mode-of-operation and not in the transmit mode. Phase was not varied during the measurements. The Clock and Mode (and Data) differential input lines connect to a CMOS gate array through DS8820/7820 differential line receivers. The OBITE line is connected to a CMOS gate array through one quarter of a 54ALS03 NAND gate IC. The +5 and -7 volt dc supplies are used to power all electronics within the T/R module with the exception of the transmitter output power amplifier. The +10 volt supply is used to power only the transmitter output power amplifier.

Since the T/R module digital input pins are connected to DS8820/7820 differential receiver IC's, additional stand-alone measurements were performed on this IC. A Tester Interface Unit (TIU) built on a previous Georgia Tech conducted interference measurement program [2] was used to facilitate the performance of the DS8820/7820 IC and T/R module measurements. Interference simulations were also performed to determine if a model could be used to simulate interference effects at very high frequencies.

After completing the T/R module measurements and the simulations and measurements on the DS7820 IC's, additional simulations were initiated, but not completed, for the OBITE output NAND gate and the Power Condition Monitoring (PCM) circuits, using estimates of transistor parameters. The PCM circuits are used to monitor over-voltage conditions on the +5 volt supply and over-temperature on the transmit power amplifier and to disable the receive and transmit modes in the event over-voltage or over-temperature conditions occur. Simulations of the LM103 3-volt regulators used to regulate voltage to the receiver LNA were not possible because transistor and diode parameters could not be obtained prior to the conclusion of this effort. It is felt that interference occuring in the LM103 3-volt regulators would explain the gain enhancement and gain

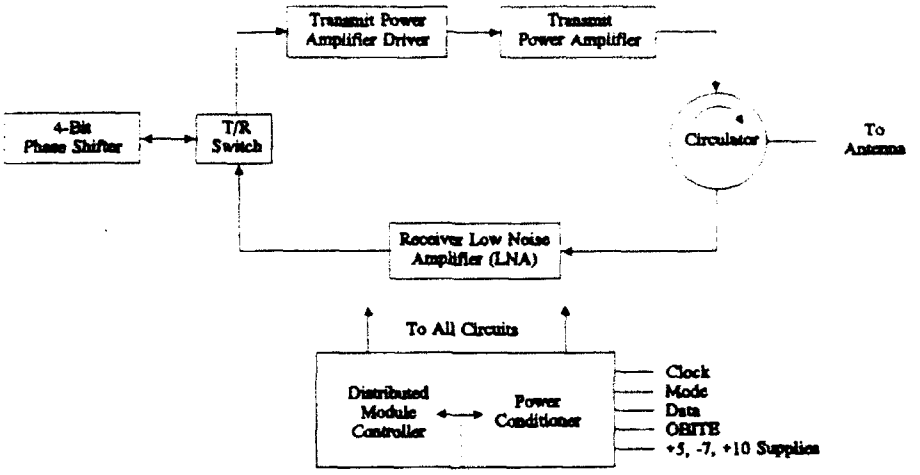


Figure 1. T/R Module Block Diagram.

compression effects that were observed during the T/R module measurements. Simulations of interference coupled to the OBITE pin indicate that a "sneak circuit" may exist that could allow interference to cause all outputs of the NAND gate to go to a logic high state which turns on both the receiver LNA and transmitter power amplifier. This could cause catastrophic failure of the T/R module. Measurements indicate that the OBITE pin is also the most susceptible T/R module pin for interference frequencies greater than 100 MHz.

MEASUREMENT PROCEDURE

Block diagrams of the T/R module conducted interference test setups are given in Figure 2. The interference and functional signals were added using various signal combiners. The signal combiner used depended upon the interference signal frequency, the functional signal frequency, and the input impedance of the pin-under-test. Table 1 summarizes the characteristics of the signal combiners used and their pin applicability. Additional information on the design and construction of the signal combiners can be found in Reference 3.

When testing the T/R modules, functional signals were generated by a T/R Module Signal Controller (MSC) built by Texas Instruments [4, 5]. The MSC is controlled by an HP9836 computer over the IEEE-488

Table 1. Signal Combiner Networks.				
Combiner	Interference Frequency Range	Functional Frequency Range	Maximum Output Power	Pin Applicability
Op Amp	dc - 160 MHz	dc - 160 MHz	+ 14 dBm	Input
250 MHz Dual Quadrature Hybrid	100 - 400 MHz	dc - 50 MHz	50 Watts	Input-Output-Power
750 MHz Dual Quadrature Hybrid	500 - 1000 MHz	dc - 150 MHz	50 Watts	Input-Output-Power
Broadband Dual Quadrature Hybrid	1 - 11 GHz	dc - 750 MHz	50 Watts	Input-Output-Power

Table 2. Interconnect Cable Pin-outs.		
T/R Module Back-Plane	T/R Module 5x3 Rectangular Connector (A Top Left, O Bottom Right	MS3102E22-14S Circular Connector
1 : +10V	B	V
2 : GND	C	C
3 : -7V	A	F
4 : +5V	D	K
5 : Spare	NOT USED	
6 : Spare	NOT USED	
7 : Beam B	L	H
8 : Beam A	K	G
9 : Mode-	J	J
10 : Mode+	O	S
11 : Clock-	G	L
12 : Clock+	F	M
13 : Data-	I	N
14 : Data+	H	P
15 : OBITE	E	R

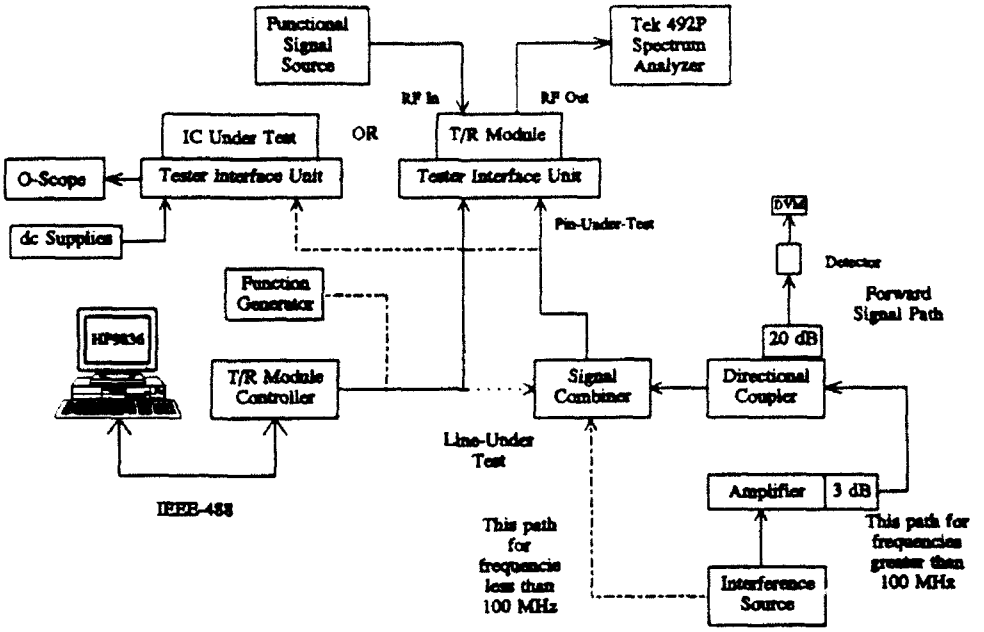


Figure 2. General T/R Module and DS7820 Conducted Interference Test Setups.

data bus. The control code used on this effort was adapted from a previously developed control program [6]. The MSC was also used to route the dc supply voltages to the T/R module-under-test. The T/R module was mounted in a Module Test Fixture (MTF) (constructed specifically for use on this effort by the Rome Laboratory Experimental Fabrication Shop) that in-turn was mounted on the Tester Interface Unit (TIU). The TIU provides 32 controlled impedance paths to device pins-under-test up to frequencies of 12 GHz. The TIU is described in detail in Reference 2. Table 2 summarizes the MSC-to-TIU interface cable connections.

The T/R module conducted interference measurements were semi-automated and controlled by an HP9836 which is used to display instructions to the operator. Commands are sent from the HP9836 through the MSC to turn the T/R module off or to to switch it to the receive mode-of-operation. The interference data was recorded manually and then plotted using GraphTools software[7]. The HP9836 control software is listed in Appendix A.[2]

[2] Appendices are contained in a separate volume that can be obtained from the authors or from Mr. Michael Seifert, Rome Laboratory/ERPT, 525 Brooks Rd., Griffiss AFB, NY 13441-4505.

The goals of the DS8820/7820 IC conducted interference measurements were to determine if the DS8820/7820 changes logic states at interference voltage and power levels that are comparable to those that caused interference in the T/R module. Simulated logic signals were provided using a function generator. The output pin (pin 6) on one-half of the DS8820/7820 was monitored using an oscilloscope. The DS8820/7820 output signal was routed to the oscilloscope through the TIU and a 1 meter coaxial cable. The DS8820/7820 has limited capabilities to drive capacitive loads so the simulated logic signal was limited to 200 kHz using this test configuration.

For interference frequencies greater than 100 MHz (where op amp signal combiners are not used) interference power incident at the T/R module or IC pin-under-test was calculated using forward signal power measured at the direction coupler sample port and then corrected by subtracting measured losses through the directional coupler, the signal combiner, and interface cables. The signal attenuation through the directional coupler sample port was then added to this result to complete the incident interference power calculation.

When using the op amp combiner for interference frequencies less than 100 MHz, incident interference power was calculated using the generator output power that was read directly from the generator output power meter. The generator output power meter reading was then corrected to account for errors in the meter reading using measured calibration results. Incident interference power was then computed by subtracting losses through the op amp combiner and interface cables. Active voltage probes were not available that operated to 100 MHz so the actual voltage levels at the output of the op amp combiner could not be determined directly. Connecting a coaxial cable at the output of the op amp combiner creates a transmission line stub that reduces the power at the combiner output at certain frequencies. Additional problems were encountered using the op amp combiners. It is felt that the op amp combiner output signal level was compressed with the addition of the functional test signal and that the interference power level never actually exceeded more than 15 dBm, although calculations indicate greater power levels were incident on the device pin-under-test.

DS8820/7820 MEASUREMENT RESULTS

Conducted interference measurements were performed on three pins of a DS7820 differential line receiver IC. Measurements were performed on the inverting input (pin 1), the non-inverting input (pin 3), and the +5 volt supply input (pin 14). The +5 volt supply pin was evaluated with the functional signal applied to either pin 1 or 3. The functional signal was applied to one input pin while the other pin was grounded. The interference signal was added to the functional signal or to the +5 volt supply voltage. Combiners were not available to add interference to two differential signals simultaneously and, therefore, measurements of the common-mode rejection capabilities of the DS7820 were not possible.

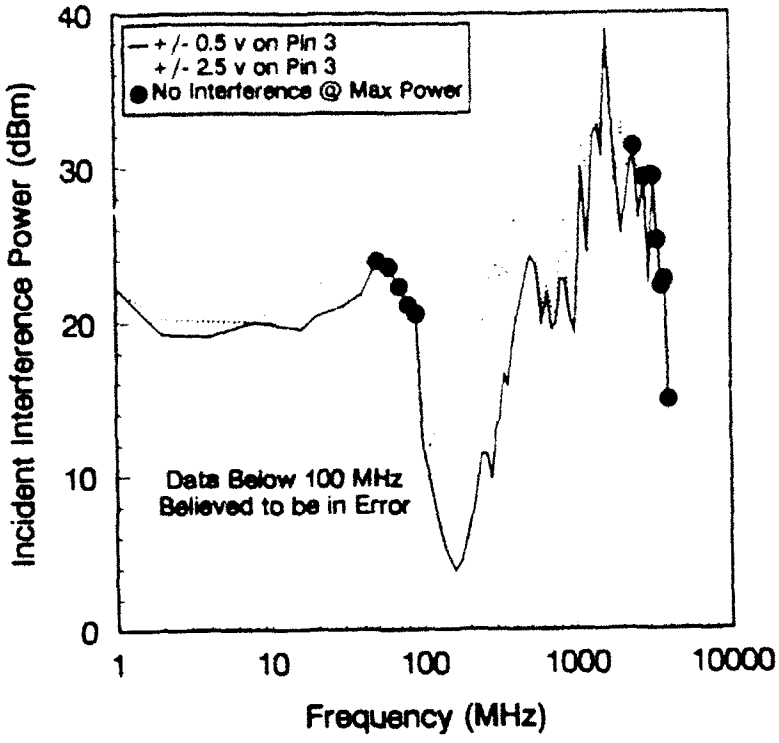


Figure 3. Interference Thresholds of DS8820/7820 +5 Volt Supply Pin.

The DS7820 IC measurements were performed over five interference frequency bands. The measured data is presented in Appendix B in order of increasing interference frequency. Plots of the interference thresholds on the + 5 volt supply with a 200 kHz pulse train applied to pin 3 of the DS8820/7820 are presented in Figure 3. The results for the pulse train applied to pin 1 and with the interference applied to either pin 1 or 3 are given in Appendix B. Solid circles indicate that no interference occurred at the maximum available power, with the maximum available power indicated by the power level shown at the solid circle locations. As explained previously, it is felt that the op amp combiner output was compressed for frequencies less than 100 MHz and that the actual output power was less than that presented in Figure 3.

An incorrect amplifier was used during the 2 to 4 GHz interference measurements and the output power was not sufficient to cause interference above 3 GHz.

The basic interference effects that were observed include (1) logic low pulled high (greater than 1 volt), (2) logic high pulled low (less than 2 volts), (3) output pulse rise and fall times increased, and (4) output pulse width narrowed or increased. At many frequencies, the output pulse width would begin to narrow (or increase)

as if the output were going to be pulled low (or high); however, the output pulse would instead begin to increase (or decrease) after narrowing (increasing), for example, and be pulled high (low) rather than low (high).

The first IC under evaluation was damaged while testing the inverting input (pin 1) over the 100 to 400 MHz frequency range. Measurements were completed at the higher frequencies using a second IC and measurements over the 100 to 400 MHz range repeated. The data obtained using the second IC was nearly identical to that obtained using the first IC and the data for the two IC's is not differentiated. The second IC was damaged over the same frequency range, at the same functional signal level (±2.5 volts), and for the same pin under test (pin 1). The effect observed in both instances was a reduction in operating bandwidth and reduced output amplitude (approximately 2.5 volts rather than 4.5 volts).

T/R MODULE MEASUREMENT RESULTS

Figure 4 illustrates the T/R module interference measurement results over the frequency range of 100 MHz to 2 GHz. (Measurements below 100 MHz were not performed due to problems encountered using the op amp signal combiners on the DS8820/7820 IC measurements.)

The most susceptible pin was the OBITE pin. The primary effect observed while testing the OBITE pin was that the T/R module cycled off and then on again once the interference signal was removed (referred to as "gain collapses" in the Appendix C tables). Three effects were observed on the Clock and Mode lines tied to the DS8820/7820 IC's. The dominant effect at interference frequencies less than approximately 1 GHz was that the T/R module would not go into receive mode or would not turn off when commands were sent over either the Clock or Mode lines. At frequencies greater than approximately 1 GHz, interference either did not occur, gain was compressed, or the module would cycle off and then come back on once interference was removed. The dominant effects observed while testing the +5 and -7 volt supplies were gain compression and the T/R module cycling off and then on again once the interference signal was removed.

Two T/R modules were damaged while performing interference tests. The first (module serial number 135) was damaged while testing the Mode+ input pin at 700 MHz. Measurements were repeated over the frequency range of 500 MHz to 1 GHz using a second module (serial number 144). Module 144 was not susceptible over the frequency range of 500 to 700 MHz at the maximum output power of the interference signal generator (approximately 43 dBm out of the generator and 37 to 39 dBm incident at the T/R module OBITE pin after subtracting losses through the 3 dB attenuator at the generator output, directional coupler, combiner, and interface cables) while the damaged module was susceptible at a level approximately 15 dB below the maximum interference generator output capabilities. This indicates that the first T/R module interference

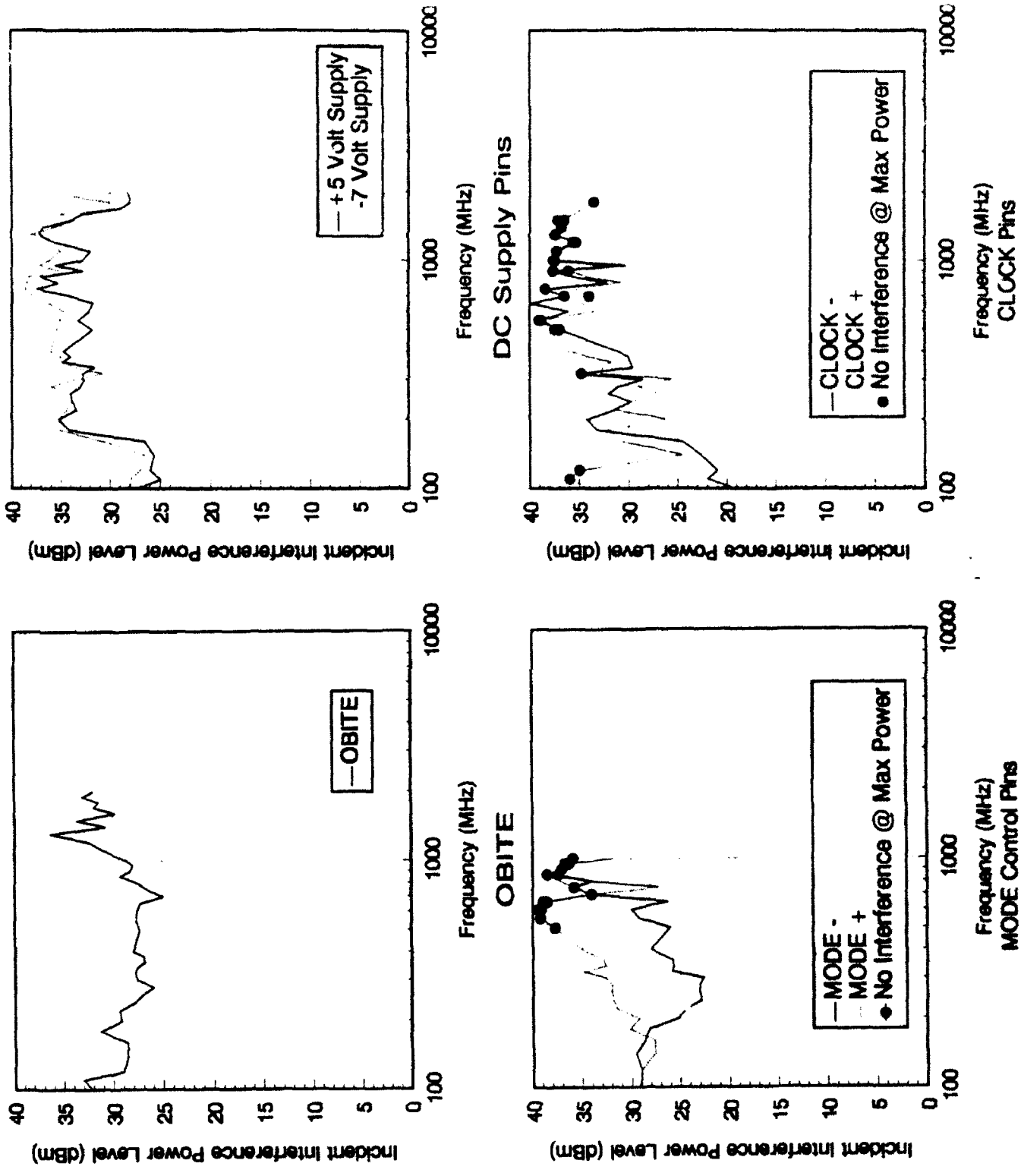


Figure 4. T/R Module Conducted Interference Thresholds.

thresholds gradually degraded prior to catastrophic failure. This effect was consistent with those observed on the DS8820/7820, in that the DS8820/7820 frequency response and output amplitude levels gradually dropped until rated performance specifications were no longer met. Module 144 was damaged while testing the Clock- input pin at 1.5 GHz. No interference was observed over the frequency range of 1 to 1.5 GHz except that the receiver LNA gain gradually dropped by 2 dB without recovering after the interference signal was removed. The gain drop prior to failure of module 144 and the 1 dB gain compression interference effects that were observed could be due to interference occurring within the LM103 3-volt regulators used to power the receiver LNA (and transmitter power amplifier driver). Simulations and measurements of this hypothesis should be performed.

SIMULATION PROCEDURE

Simulations of conducted interference effects were performed using a model of a DS8820/7820 dual differential line receiver IC, a model of a 54ALS03 quad two-input open collector NAND gate, and a model of the T/R module Power Condition Monitor circuit. These simulations were performed over a frequency range of dc to 1 GHz using PSPICE [8]. Simulated interference signals were combined with functional signals in a manner similar to those used in the conducted interference measurements. The techniques described in Reference 9 were used in the performance of the simulations. The objective was to determine if simulated interference voltage levels and effects are comparable for both the DS8820/7820 and the T/R module. The objectives of the 54ALS03 and Power Condition Monitor circuit simulations were to identify other interference effects that alter the operation of the T/R module.

The DS8820/7820 IC is used as the front end of the six differential inputs of the T/R module (Clock+, Clock-, Data+, Data-, Mode+, and Mode-). Failure of this IC due to interference effects would prevent the T/R module from receiving the correct digital commands. The DS8820/7820 simulation model, shown in Figure 5, was developed from information obtained from National Semiconductor. The schematic for this model was found in the National Semiconductor Interface databook [10]. This schematic provided all the necessary resistor values, but National Semiconductor had to be consulted to obtain parameters for the transistors [11]. The transistors in this model were derived from the 2N3932 transistor. The following parameters were changed from the values given in the 2N3932 PSPICE model.

Ideal Maximum Forward Beta	Bf = 90
Ideal Maximum Reverse Beta	Br = 2.5
Zero-Bias Base Resistance	Rb = 21
Collector Ohmic Resistance	Rc = 26

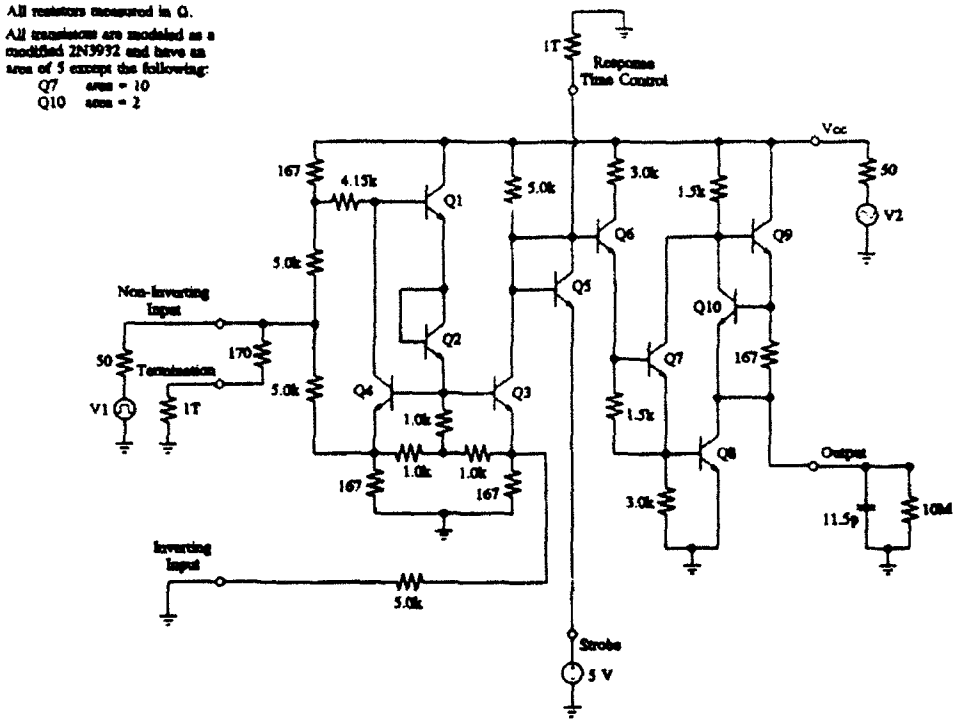


Figure 5. Schematic of the DS8820/7820 Differential Receiver.

Emitter Ohmic Resistance	Re = 1.31
Forward Early Voltage	Vaf = 90
Base-to-Emitter Area Ratio	area = 5

The only exceptions to these changes are the transistors Q7 and Q10. Q7 has a base-to-emitter area ratio of 10 and Q10 has a base-to-emitter area ratio of 2.

As shown in Figure 5, some additional components are added for simulation purposes. V1 is the functional signal (a 500 kHz square wave) and V2 is the interference signal (a CW signal). Both sources have a 50 Ω series resistor to simulate a 50 Ω source impedance. The Termination pin and the Response Time Control pin are tied to 1 Tera-Ω resistors to provide a dc path to ground for simulation purposes. A 5 volt dc source is connected to the Strobe pin to simulate a logic high signal.

Limited simulations were performed on the OBITE line interface IC (54ALS03 NAND gate) and the Power Condition Monitor circuits (see Figures 6 and 7). A PSPICE macro-model of the LN139 quad comparator was used while the 54ALS03 model was obtained from Reference 12. Transistor parameters were

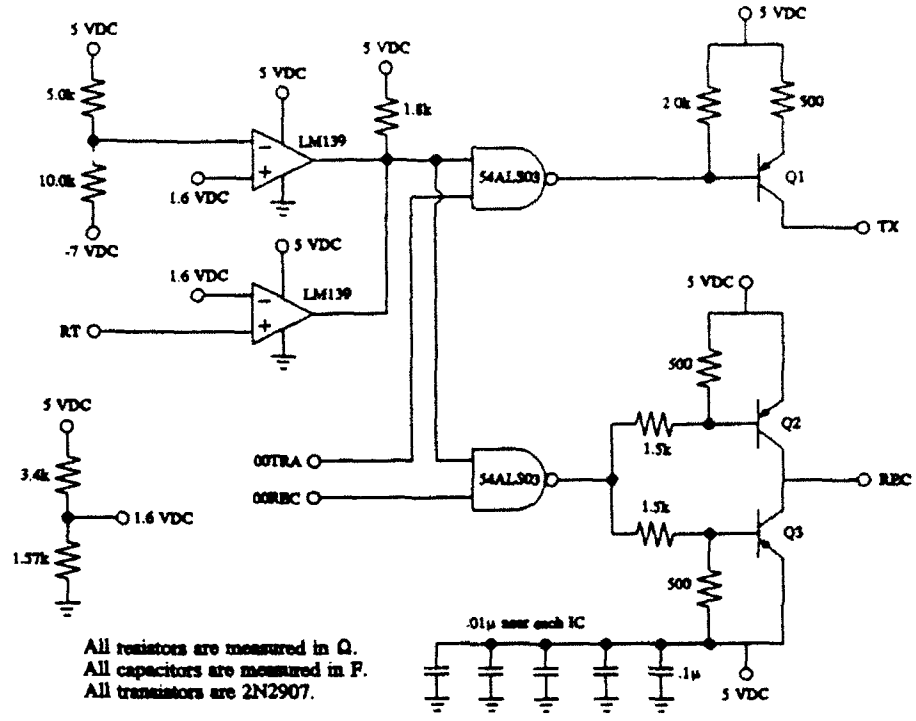


Figure 6. Schematic of the T/R Module Power Condition Monitor Circuit.

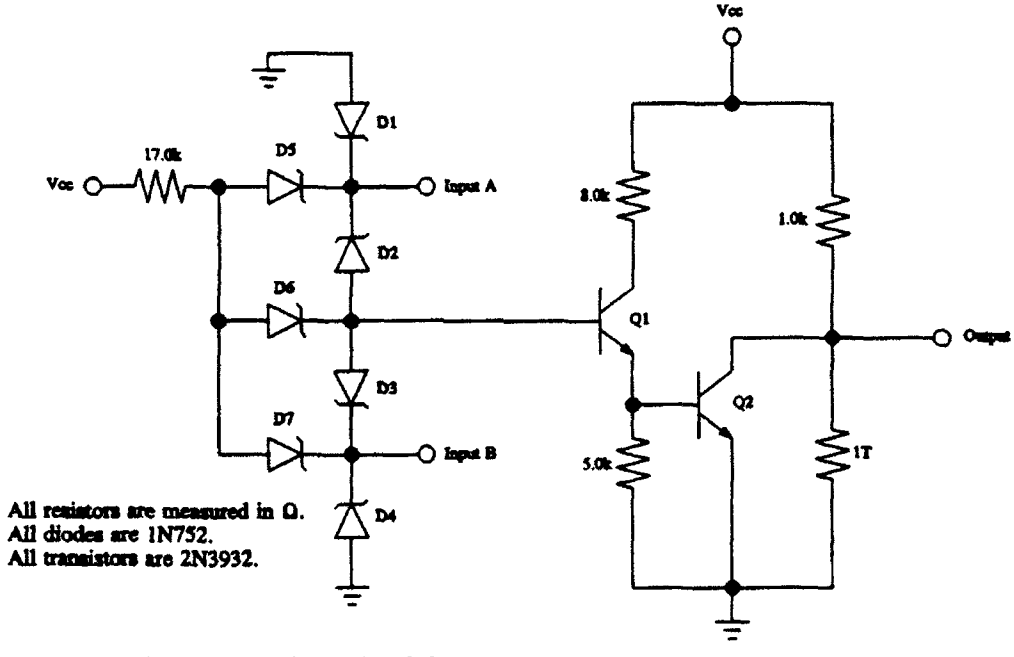


Figure 7. Schematic of the 54ALS03 Open-Collector Nand Gate.

estimated for the 54ASL03 since they were not provided in Reference 12. The Power Condition Monitor circuits are used to monitor the +5 volt supply for over voltage conditions and transmitter power amplifier temperature and to turn the T/R module off in the event of over-voltage or over-temperature. Model parameters for the LM103 3-volt regulator were not obtained prior to the conclusion of this effort and therefore simulation of the receiver LNA regulator circuits could not be performed.

SIMULATION RESULTS

Simulations of the DS8820/7820 dual differential line receiver IC interference effects and thresholds correlate within a few dB of measured results to 1 GHz. The only simulations that were performed involved placing the interference signal on the +5 volt supply signal. All of the measured interference effects were observed with the exception that logic low was pulled high instead of the logic high being pulled low at frequencies greater than 500 MHz. Figure 8 shows simulated and measured interference effects at the DS8820/7820 output pin with the interference level set to the same amplitude in both the measurements and simulations. At low interference frequencies (less than a few hundred MHz), the simulation results agree almost exactly with the measured results. At high frequencies, the interference thresholds are within 3 dB, but the effects differ. For example, simulations show logic low is pulled high at 1 GHz, but measurements show logic high is pulled low at this frequency. The differences observed at high frequencies could be due to the presence of the second half of the DS8820/7820 which is not taken into account during simulations or the presence of other parasitic linear devices, such as parasitic resistors, capacitors and inductors, or non-linear parasitic devices, such as parasitic substrate diodes, that are not accounted for during simulations.

Simulations of interference signals coupled to the OBITE line were performed by tying two of the NAND gate circuits shown previously in Figure 8 together at only the +5 and ground connections. This would simulate two NAND gates in a single package. An interference signal was then applied to one NAND gate output while square waves were applied to all four inputs. The simulations confirmed that interference applied to the first NAND gate output can cause the other NAND gate output logic high output level to be pulled low. This would cause the NAND gate that enables the receiver LNA to cycle off when it is supposed to be on and then cycle back on once the interference signal is removed, as was observed during the T/R module OBITE pin measurements. It was not possible to perform additional simulations to determine if increased interference levels could cause the logic low levels to be pulled high which could cause the transmitter and receiver amplifiers to both be enabled. This condition could lead to catastrophic failure of the T/R module and should be investigated further since this may be a critical "sneak circuit" mode of failure.

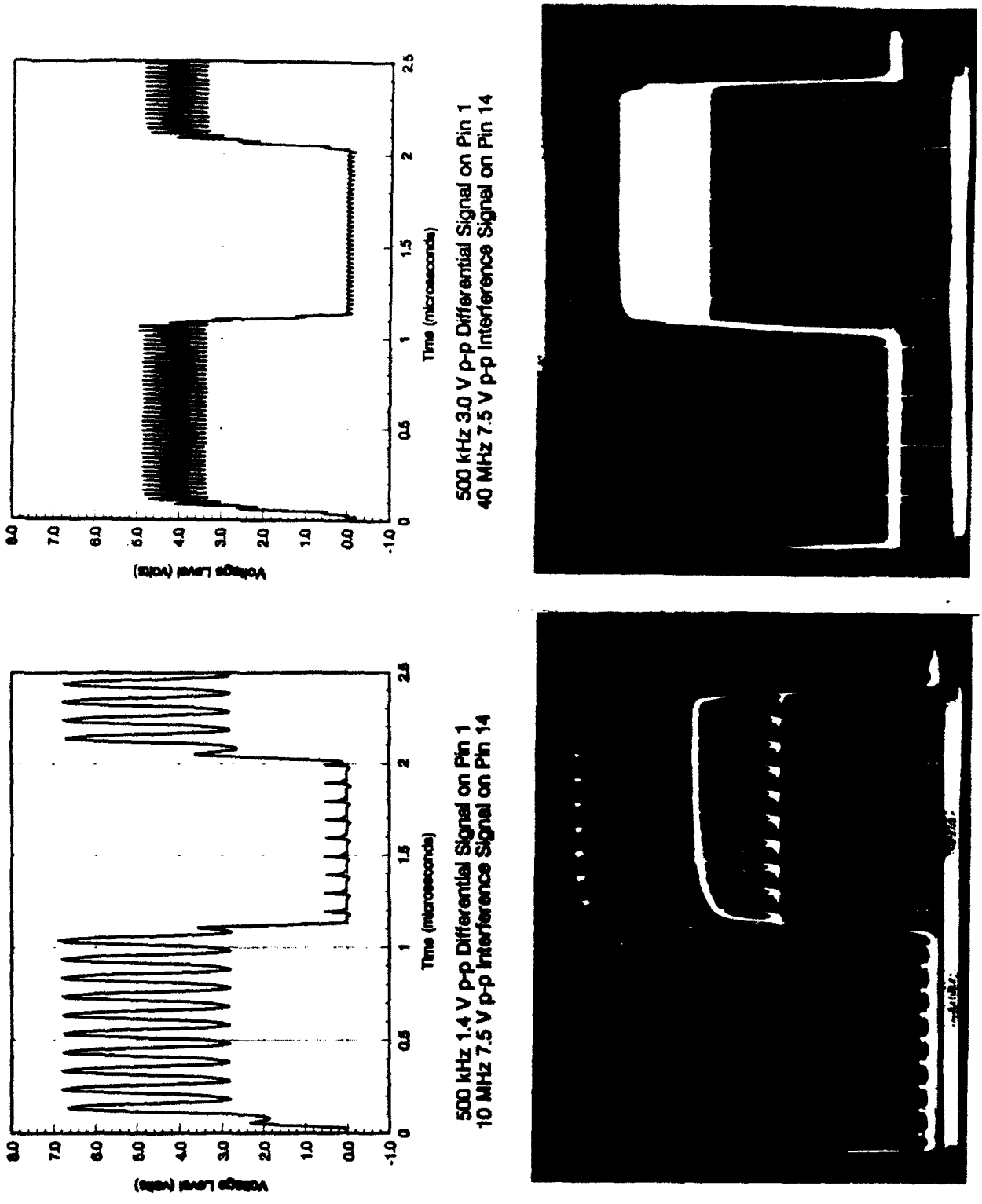


Figure 8. Comparison of DS8820/7820 Measured and Simulated Interference Effects on +5 Volt Supply Pin.

Simulations of interference coupled to the +5 volt supply pin of the complete Power Condition Monitor circuit show that the output transistors could be forced to change states, placing the transmitter on instead of the receiver and vise-versa, or even both on or off (both on would cause catastrophic failure of the T/R module). The power level required to cause interference was higher than that determined through measurements because transistor parameters are not correct.

Additional simulations should be conducted to determine if the transmitter power amplifier gate voltage (-7) could be disabled due to interference whild the drain voltage (+10) is applied. This condition would cause catastrophic failure of the T/R module.

T/R MODULE VULNERABILITY ASSESSMENT

The results of this effort indicate that the Raytheon and Texas Instruments X-band T/R modules are susceptible to EM interference but not whether a vulnerability condition could exist. To perform a vulnerability assessment, an estimate was made of the coupled signal levels due to external EM threats. Radiated field levels from Reference 13 were used to determine if the T/R modules could be vulnerable to upset or damage due to EM field exposure. It was assumed that each pin was connected to a tuned dipole for analysis purposes. Techniques from Reference 14 were used to calculate coupled signal levels given the threat source profile in Figure 9 (from Reference 13) and for separation distances of 1 and 10 km. Given the power density profile in Figure 9, the power received at an IC pin can be calculated using the following equation,

$$P_r = \frac{G_r P_D \lambda^2}{2\pi}$$

or in decibels as

$$P_r = G_r + P_D + 20\log\lambda - 7.98$$

where,

P_r	=	recieved power in dBm
G_r	=	receive antenna gain, which for a tuned dipole is 2.14 dBi
P_D	=	incident power density in dBm per square centimeter
λ	=	wavelength in centimeters

Figure 10 overlays the OBITE susceptibility profile over the calculations of coupled signal levels. Coupling to the OBITE pin could lead to T/R module vulnerability over the entire frequency range of 100 MHz

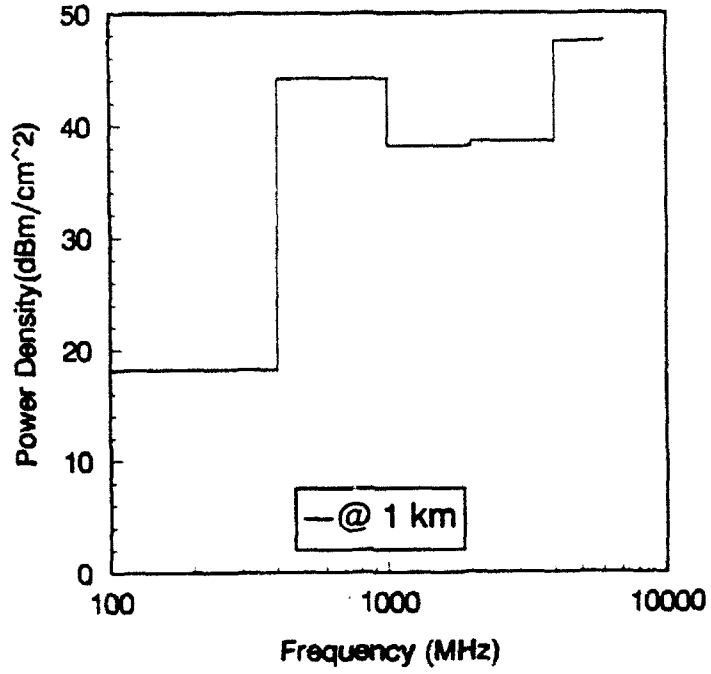


Figure 9. EM Power Density at a Distance of 1 km from Threat Sources.

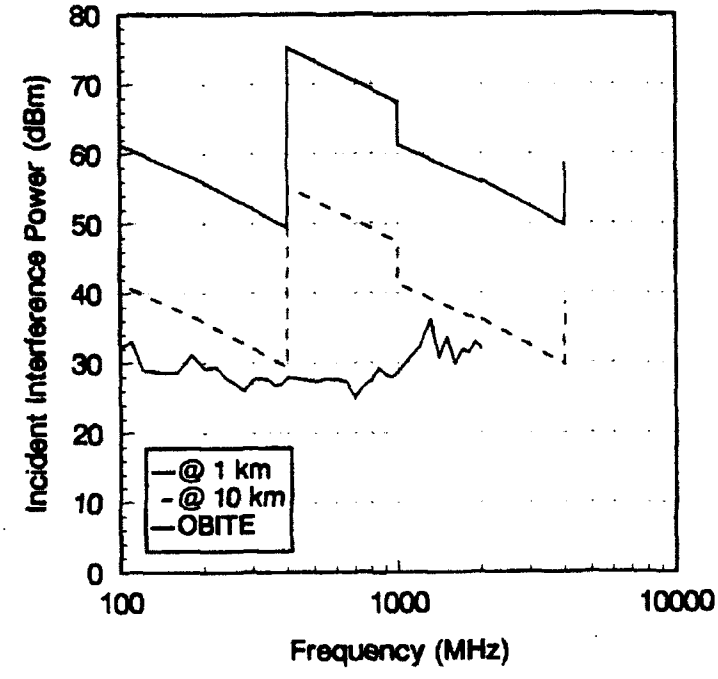


Figure 10. OBITE Pin Susceptibility Thresholds and Calculated Threat Levels.

to 4 GHz for threat sources located at either 1 or 10 km. Additionai filtering, limiting and shielding of up to 60 dB would be required to prevent vulnerability over the frequency range of 100 MHz to 4 GHz.[3]

CONCLUSIONS AND RECOMMENDATIONS

Simulations of the DS8820/7820 interference effects and the resulting susceptibility levels were in close agreement with measured effects and susceptibility thresholds. Accurate models of the OBITE buffer 54ALS03 NAND gate, the Power Condition Monitor circuitry and the ± 3 volt regulator circuitry were not developed during this limited effort. The models that were developed for the 54ALS03 NAND gate and the Power Condition Monitor circuitry were sufficient to verify that most of the measured T/R module interference effects could be correlated with simulated results by modeling relatively simple front-end circuitry of the modules.

A possible "sneak circuit" was identified through simulations of the 54ALS03 NAND gate IC that is shared between the OBITE buffer and the Power Condition Monitor circuit. Interference coupled to the OBITE output (the most susceptible pin based on T/R module measurement results) could cause all outputs of the 54ALS03 to go to either a logic high or low state. If all outputs go to a logic high then both the transmitter and receiver amplifiers would be enabled, resulting in catastrophic failure of the T/R module. Based on this analysis it is recommended that future T/R modules be designed such that input/output buffer IC's and circuitry contained on a common substrate are not shared with internal control circuitry.

To improve the EM reliability of the X-band T/R module, all input and output pins should be limited and filtered at the package I/O terminals. The filter and limiter circuitry should be placed as close to the T/R module as possible.

In order to adequately address EM reliability issues early in the design and test phases of T/R module development, design guidelines and test procedures should be prepared as part of a T/R module procurement package. Test fixtures and equipment should be incorporated into the Rome Laboratory T/R module reliability measurement facility in order to facilitate EM reliability testing.

---

[3] The threat profile given in Figure 9 does not considered classified sources, lightning, electro-static discharge, electronic warfare transmitters, nuclear electromagnetic pulses or high power microwave weapons. Consideration of the above sources could dramatically increase the vulnerability levels and required filtering, limiting and shielding.

Based on the results of this effort, the following additional activities are recommended:

1. Continue simulations of T/R module circuitry to determine the benefits, limitations, and procedures required to accurately simulate EM interference effects, including damage effects.
2. Continue measurements on T/R modules to determine the effects of interference modulation.
3. Design, simulate and construct T/R module protection circuitry.
4. Design and construct improved low frequency combiner networks (interference frequency less than 100 MHz).
5. Develop test procedures, circuitry, fixtures, and control software for use within the Rome Laboratory T/R module reliability measurement facility so that EM reliability measurements can be performed in addition to other environmental tests.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Mr. Mike Little, Rome Laboratory/OCTP for providing T/R modules, documentation, a controller, and controller software for use during testing. The authors would also like to thank Mr. Dennis Walker, of the Rome Laboratory Experimental Fabrication Shop for machining the Module Test Fixture and Mr. Syed Huq of National Semiconductor for his perseverence in obtaining transistor parameters for the DS8820/7820. Special thanks to Mr. John Cleary and Mike Seifert of Rome Laboratory/ERPT for defining and supporting this Summer Research Program effort.

REFERENCES

[1] "Module Validation Program, Final Technical Report," by Raytheon Co., Equipment Division, Wayland, MA, 01778 and Texas Instruments Inc., Defense Systems and Electronics Group, Dallas, TX, 75265, for Rome Air Development Center, Griffiss AFB, NY, 13441, under Contract No. F30602-86-C-0024, 5 February 1990.

[2] J.K. Daher, J.C. Santamaria, R.M. Herkert, and J.M. Goodroe, BIT Technology-VLSI Circuits/Assemblies, Final Technical Report, RADC-TR-90-310, December 1990.

[3] J.K. Daher, J.P. Rohrbaugh, and J.G. Hotchkiss, EMI Test Methodology for High Speed, High Density IC's, Final Technical Report, RADC-TR-86-223, February 1987.

[4] Shiu-Kai Chin, et. al., Module Signal Controller, Final Technical Report, RADC-TR-86-150, September 1986.

[5] Module Controller Documentation. No other reference related information available. Only partial document provided.

[6] MV_TRE computer program, by Ken Bradley, 26 October 1987.

[7] GraphTools, 3-D Visions, Redondo Beach, CA 90277.

[8] PSPICE, MicroSim Corp., Irvine, CA 92718.

[9] J.P. Rohrbaugh and B.R. Farris, MMIC Conducted Interference Test Methodology, Final Technical Report, Volume I, RADC-TR-90-312, Vol I (of two), December 1990.

[10] National Semiconductor Interface Data Book, 1983, National Semiconductor, Santa Clara, CA 95051.

[11] Phone conversations with Mr. Syed Huq, National Semiconductor, Interface and Peripherals Group, Interface Applications, 2900 Semiconductor Drive, Santa Clara, CA, 95052-8090, (408) 721-4874.

[12] TTL Data Book, Texas Instruments, Dallas, Texas.

[13] H. W. Denny, J. K. Daher, and J. P. Rohrbaugh, "Integrated Circuit Technology Assessment (ICTA) Project," Final Technical Report: Definition Phase, for AFEWC/SAX, San Antonio, Texas 78243, under Contract No. F41621-83-C-5015, prepared by the Georgia Tech Research Insititute on GTRI Project A-3657, April 1984.

[14] J. P. Rohrbaugh, B. R. Farris, and R. C. Alford, " EM Performance Monitor," for Rome Laboratory, Griffiss AFB, NY 13441-5700, under Contract F-30602-89-C-0144, on GTRI Project A-8426, October 1990.

MONTE CARLO VALIDATION OF A THEORETICAL MODEL FOR THE GENERATION OF NON GAUSSIAN RADAR CLUTTER.

JORGE LUIS ROMEU
ASSOCIATE PROFESSOR
DEPARTMENT OF MATHEMATICS

STATE UNIVERSITY OF NEW YORK
COLLEGE AT CORTLAND
CORTLAND, NY 13045

FINAL REPORT FOR:
SUMMER RESEARCH PROGRAM
ROME LABORATORY

SPONSORED BY:
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
BOLLING AIR FORCE BASE, WASHINGTON, D.C.

SEPTEMBER 1992

Typeset by AMS-TEX

MONTE CARLO VALIDATION OF A THEORETICAL MODEL FOR THE GENERATION OF NON GAUSSIAN RADAR CLUTTER

JORGE LUIS ROMEU
ASSOCIATE PROFESSOR
DEPARTMENT OF MATHEMATICS
SUNY CORTLAND
RESEARCH FELLOW OF THE
CASE CENTER OF SYRACUSE UNIVERSITY

Abstract

Generation of multivariate Non Gaussian random variates is of importance in radar clutter studies. For when the analytical evaluation of a radar clutter distribution is difficult or impossible, it is through computer simulation studies that this problem is attacked and solved.

A new statistical method, based on SIRP's (Spherically Invariant Random Processes) has been proposed. This theoretical method allows, both, fit testing of a multivariate Non Gaussian process, and the computer generation of these processes. The theoretical method in question works by *decomposing the Non Gaussian process* into the product of two sub processes. One these processes is univariate and drives the distribution of the Non Gaussian process. The other subprocess is multivariate Gaussian.

In theorey, the new method appears sound and correct. However, in practice we deal with limited data. In such cases, results may not always reflect the theoretical properties with the required accuracy. For example, it may not be possible to recognize, from the sample, which is the univariate process that drives the multivariate Non Gaussian one. Or the lack of knowledge about the true covariance matrix of the Non Gaussian process, which is then estimated from the data, may substantially degrade the method's accuracy.

The present report describes a Monte Carlo validation experiment, designed to evaluate this theoretical method. Using this approach, we generate two specific SIRP Processes. Then, we perform goodness-of-fit tests on several variables theoretically derived from the Processes and on the Processes themselves. We use different sample sizes and number of p-variates. We also work with the covariance matrix from which we generated the Process and the covariance matrix estimated from the data. Results are compared and statistical tests are discussed.

MONTE CARLO VALIDATION OF A THEORETICAL MODEL FOR THE GENERATION OF NON GAUSSIAN RADAR CLUTTER.

Jorge Luis Romeu

INTRODUCTION.

The theory of SIRP *Spherically Invariant Random Processes*, has been presented in Rangaswamy, Weiner and Ozturk (1992). It has also been extensively and carefully discussed in the forthcomming Rome Labs document referred in this paper as the Kaman Report (1992).

Succinctly, an SIRP X is defined via the **Representation Theorem**, as the product $X = S*Z$ of two independent random processes S, Z. The first one, S, is univariate and ***drives*** the SIRP process X: i.e. if S varies, so does X.

The second process, independent of S, is the Multivariate Gaussian process $Z \sim MVN_N(0, M)$, which remains the same no matter what is X. We can ***standardize*** S so $\mathcal{E}(S^2) = 1$; then $\Sigma_X = M$.

The conditional density function of $X|S$ is:

$$f_{X|S}(x|s) = (2\pi)^{-N/2}|M|^{-1/2}s^{-N}exp(\frac{-p}{2s^2})$$

From here, the unconditional density function becomes:

$$f_X(x) = (2\pi)^{-N/2}|M|^{-1/2}h_N(p)$$

$$where \quad h_N(p) = \int_0^\infty s^{-N}exp(\frac{-p}{2s^2})f_S(s)ds$$

Here $p = X'\Sigma^{-1}X$, is the ***quadratic form*** of the process X. Function p will play a pivotal role in the SIRP Theory. For, $h_N(p)$ will provide the density function of such random variable p, via:

$$f_P(p) = \frac{1}{2^{N/2}\Gamma(N/2)}p^{N/2-1}\,h_N(p)$$

From the above relations we can verify how, for a given SIRP X, everything is known once we have h_N, the $f_S(.)$ density function of S and the covariance matrix Σ of X.

However, as with any other theoretical model, the SIRP must be validated empirically before proceeding on to its widespread use. There are several reasons for this.

First, the SIRP theory states that X is obtained by the product of S, Z. Consequently, the resulting quadratic form p has a distribution, dependent on S. The marginal distributions of the multivariate X follow the same family of SIRP distributions as X. And the resulting covariance matrix Σ will be obtained, through S, M.

It is however, necessary to verify that such assertions are met in practice. And moreover that they are met in such a way that it has a practical application (i.e. that they can be recognized from the data).

Then, the SIRP model assumes that the covariance matrix Σ, of X **is known**. This seldom occurs in practice, except in the case of the simulation of radar clutter. Such simulations constitute an important application of the SIRP model, allowing the study of some types of radar clutter with difficult or impossible analytical solutions.

It is necessary to verify that the resulting random variates, say S, p can actually be identified under different experimental settings (e.g. different sample sizes, variate correlations ρ, or number of variates N in the process).

A third reason for model validation is to conduct performance studies of interest about the model. The SIRP theory requires knowledge of key elements which are seldom known in practice. This is the case, for example, of the covariance matrix of the SIRP process $X = S * Z$. In practice, Σ is substituted by its estimate Σ^*. It is necessary to study any efficiency loss by such a substitution. And it is necessary to study the sample size requirements, N_r, and the number of variates N, for which such estimation breaks down.

For all these reasons, a Monte Carlo Validation Study is required for the SIRP model described above. However, there are serious problems when attempting such a study. First, validating this model requires testing **both** outputs: the SIRP process X and the quadratic form p. Second, there are no tests for the multivariate SIRP distributions of interest, that we can apply to X. If there were, then the main interest in SIRP theory, the need for the indirect testing of the multivariate process X via its univariate quadratic function p, would disappear.

To circumvent this problem, we approach the validation process through a *two-phase scheme*, taking advantage of the SIRP properties. In the first phase, we implement a multivariate Gaussian Process $X = S * Z$, obtained with S constant, which is a special case of an SIRP. There are several, well investigated, multivariate normality goodness of fit tests, which can be used on process X. And we can simultaneously test for the Chi Square distribution of the quadratic form p.

In the second phase we analyze a **univariate** SIRP, where $N \equiv 1$. We find cases of the *univariate K-Distribution*, analytically simple enough to obtain a closed form for its Cumulative Distribution Function. We need such closed forms to handle the simulation program. We generate these univariate SIRP's following the established model. Next, we test **both** the univariate X and the univariate p.

There is no multivariate test available for the multivariate K, which is one of our multivariate distributions of interest. Hence it is not possible to test the SIRP process directly. In addition, the resulting distribution of the quadratic function p becomes so involved (including Incomplete Bessel Funcions) that it is not possible to obtain a closed form for its CDF. And testing the fit for p becomes convoluted, in a Monte Carlo study.

For these reasons, we will not attempt to analyze the general case. However, our approach does provide validation for the SIRP's. Since it does perform a comprehensive study of the problem, conmensurate with its realistic constraints.

PHASE I: MULTIVARIATE GAUSSIAN SIRP.

By letting S be a unit valued constant, $X = Z$ is an SIRP. By sampling the real multivariate Gaussian $Z \sim MVN_N(0, I_N)$ with covariance identity, we obtain the quadratic function $p = \sum_{i=1}^{N} X_i^2$. When the covariance Σ is known, $p \sim$ *exponential*(N), i.e. the quadratic form is distributed as an exponential random variable.

We apply a battery of four multivariate normality goodness of fit tests to Z and two Kolmogorov Smirnov fit tests to the quadratic form p. One, when Σ is known (denoted P KN/KS in the tables) and another when we estimate it from the data (P ES/KS in the tables). Finally, we apply a second, Chi Square, goodness of fit test to p, for double checking its exponential distribution (P K/CHI in the tables).

Two of our four multivariate normality tests (Ozturk and Romeu, 1990) were recently developed and have good power properties when sample sizes are small (CHOLESKI and SIGMA in the tables). The other two multivariate normality tests (M-SKEW and M-KURT), Mardia's Skewness and Kurtosis, (Mardia (1970)), were studied for small samples in Romeu (1992a). All four of these tests are scale-location invariant. In addition, Romeu (1990) provided empirical critical values when $n < 200$, which improve their efficiency with small samples.

A series of fortran programs were developed with a REXX system program to drive them in a simulation system. In an IBM 3090, using the IMSL random variate generators, this system was run. We simulated correlated (H_1) and uncorrelated (H_0) Gaussian, sample sizes of 50, 100, 200, number of variates 2, 4, 8 and 1000 replications. And we tested the (*i*) generation of multivariate white Gaussian, the quadratic form p with (*ii*) covariance known and (*iii*) covariance estimated from the data.

In Table 1 we show the results for uncorrelated Gaussian (H_0) with two variates and samples of 200 data points. The seven tests applied to the data (four for multivariate normality and three for the univariate quadratic form p) are reasonably close to their expected values (i.e. they are close to their test nominal significance levels $\alpha = 0.1, 0.05, 0.01$). We also verify that for large sample size (i.e. 200 data points), the performance of the quadratic form for p that uses the *sample covariance* is acceptable.

One caveat is due, reagarding our use of the Kolmogorov Smirnov test. It is known that this goodness of fit test is conservative when the parameters are estimated from the data. We have used the approach in Goel (1982), suggested by Allen (1978), and adjusted its significance level. We used four times the nominal level α, to test for that level (i.e. we test at $\alpha = 0.4$ and report at $\alpha = 0.1$, and so on). Observe how, in the *large sample case*, this approach works well.

TABLE 1.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 200 VARIATES: 2

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.09100	0.03800	0.01300
SIGMA	0.07700	0.03900	0.01300
M-SKEW	0.08400	0.04300	0.01500
M-KURT	0.10000	0.05300	0.01000
P -KN/KS	0.09800	0.04500	0.00400
P -ES/KS	0.10600	0.03100	0.00100
P -K/CHI	0.09500	0.05000	0.00900

TABLE 2.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 100 P-VARIATES: 2

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.09800	0.05100	0.00800
SIGMA	0.09200	0.05100	0.00700
M-SKEW	0.08200	0.04500	0.01000
M-KURT	0.09600	0.05500	0.00900
P -KN/KS	0.07100	0.02800	0.00800
P -ES/KS	0.16200	0.04300	0.00000
P -K/CHI	0.08500	0.04300	0.00600

TABLE 3.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 50 P-VARIATES: 2

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.10100	0.04500	0.00700
SIGMA	0.10300	0.04900	0.00600
M-SKEW	0.08200	0.04100	0.00900
M-KURT	0.07200	0.03200	0.00300
P -KN/KS	0.08500	0.04400	0.00700
P -ES/KS	0.19200	0.06000	0.00300
P -K/CHI	0.10600	0.05300	0.01000

In Table 2 we report the same results as above, now for samples of size 100. We can see how most tests are still close to their nominal significance levels. There is one exception: the test for p using Σ^* the covariance estimated from the data (P ES/KS in the tables). Its empirical significance level has gone up to 0.162. The reason for this poor result is due to a *loss of efficiency in the estimation of* Σ *as the sample size decreases.*

Results for Table 3 are obtained for samples of size 50. As the sample size decreases, the test for p, when the covariance matrix Σ is estimated from the data, continues deteriorating. The empirical significance level for this test has now gone up to 0.192.

In Table 4 we present similar results for four variates and sample sizes of 200 datapoints. Observe, for this large sample, that tests are well within their expected values. The test for p, when Σ is estimated, deteriorates faster in the case of four variates ($N = 4$).

In Table 5 we observe the same type of results, now for 100 datapoints. We see that the only test which has deteriorated is that for the quadratic form p, when Σ is estimated from the data. In Table 6 we show similar results, now for samples of 50 datapoints. We observe how the test for p, when Σ is estimated from the data, deteriorates even more with the reduction of the sample.

In Table 7 we show results for eight variates and sample sizes of 200 data points. Observe here that all tests are close to their nominal significance level, except the one for p, when Σ is estimated from the data. In general, as the number of variates N in the multivariate Gaussian increases, the performance of the test requires a larger sample. Hence, where 200 data points was excellent for the bivariate Gaussian, it is no longer so, for the Gaussian with $N = 8$ variates. Of course, it is even worse in the case where the covariance matrix Σ is unknown and estimated from the data.

In Table 8, for a sample size of 100 data points, the performance of the test for p, when Σ is estimated from the data, continues its deterioration. And in Table 9, for $N = 8$ variates and 50 data points, the performance of p, when the covariance matrix Σ is estimated, is even worse (0.199). All other tests are within reasonable bound of their respective nominal levels α.

The above performed tests are not joint tests. Therefore, if one of them, *isolated*, departs from its expected value, is not necessarily indicative of statistical problems. In the long run we will expect, by chance, that some of these tests will fail.

We also explored the problem under the alternative hypothesis H_1, i.e. when the distribution of the SIRP X is not $MVN_N(0, I_N)$. We simulated the SIRV $X \sim MVN_N(0, \Sigma)$, $\Sigma \neq I_N$. In particular, we simulated a multivariate normal with covariance matrix equal to its correlation matrix, with all non diagonal entries $\rho_{ij} = 0.5$, $i \neq j$. In this case, we could assess (i) the effect of miss specifying the covariance matrix, and of (ii) estimating Σ directly from the data and (iii) the power of the SIRP model to identify alternative distributions as such.

In Table 10 we show the results when simulating X from a bivariate correlated Gaussian, with samples of size 200. We can see how poor the agreement is between empirical and nominal significance

TABLE 4.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 200 P-VARIATES: 4

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.10200	0.04900	0.01000
SIGMA	0.10300	0.04500	0.01400
M-SKEW	0.11000	0.06300	0.01200
M-KURT	0.10500	0.05800	0.01100
P -KN/KS	0.07400	0.04100	0.00900
P -ES/KS	0.11400	0.02600	0.00200
P -K/CHI	0.09600	0.04200	0.01000

TABLE 5.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 100 P-VARIATES: 4

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.08900	0.04300	0.00700
SIGMA	0.09500	0.04300	0.01100
M-SKEW	0.09500	0.05290	0.02200
M-KURT	0.11200	0.05600	0.01700
P -KN/KS	0.09000	0.05100	0.00700
P -ES/KS	0.14400	0.02800	0.00300
P -K/CHI	0.08600	0.04000	0.01000

TABLE 6.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 50 P-VARIATES: 4

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.09000	0.04400	0.00400
SIGMA	0.08900	0.04400	0.00600
M-SKEW	0.09600	0.05500	0.00600
M-KURT	0.05900	0.04200	0.00500
P -KN/KS	0.09500	0.03800	0.00900
P -ES/KS	0.16000	0.05400	0.00300
P -K/CHI	0.11200	0.05300	0.01000

TABLE 7.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 200 P-VARIATES: 8

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.10900	0.05200	0.01300
SIGMA	0.10600	0.06500	0.01500
M-SKEW	0.09300	0.04900	0.01100
M-KURT	0.10700	0.04900	0.01100
P -KN/KS	0.11000	0.04900	0.01000
P -ES/KS	0.12500	0.03500	0.00200
P -K/CHI	0.09900	0.04600	0.01200

TABLE 8.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 100 P-VARIATES: 8

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.10000	0.06100	0.01500
SIGMA	0.11300	0.05500	0.01100
M-SKEW	0.10800	0.05100	0.01900
M-KURT	0.10100	0.05300	0.01100
P -KN/KS	0.08300	0.04400	0.00700
P -ES/KS	0.13000	0.03300	0.00100
P -K/CHI	0.09800	0.05800	0.01200

TABLE 9.

TOTAL REJECTIONS FOR N= 1000 TOTAL CASES.

SAMPLE SIZE: 50 P-VARIATES: 8 SEED: 1742315143

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.11600	0.06600	0.01100
SIGMA	0.09600	0.05100	0.01100
M-SKEW	0.08900	0.04500	0.01100
M-KURT	0.08000	0.04000	0.01100
P -KN/KS	0.07600	0.03500	0.00800
P -ES/KS	0.19900	0.05300	0.00100
P -K/CHI	0.08200	0.03700	0.00600

levels of the fit test, for the quadratic form p that assumes the SIRP has $\Sigma = I_N$. The same occurs in Table 11 and 12, for samples of size 100, 50, respectively. Finally, in Table 13 we show the same problem for Gaussians with eight variables and sample sizes 200.

Such disparity between empirical and nominal significance levels can be interpreted in two ways. First, it is a sign that the quadratic function p can actually discriminate between H_0, H_1 with high efficiency. It also warns the user of the dangers in miss specifying a covariance matrix Σ. The first interpretation can be readily used in simulation studies, to assess, say, minimal sample sizes. The second interpretation, estimation of Σ from data, should be used when performing the field analyses.

Phase I of the validation shows how, for the special case of the Gaussian SIRP, (*i*) the theoretical model holds, (*ii*) we can correctly estimate the fit of the quadratic form p, for samples down to size 50 and number of variates up to 8, when the covariance matrix is correctly specified. And (*iii*) the quadratic form p obtained by estimating the covariance matrix from the data, approximates reasonably well for large samples (say of size 200 and above) but not accurately enough for medium 100 or small 50. Finally, (*iv*) the SIRP model can effectively discriminate, through its quadratic function p, an incorrectly specified (alternative) model.

Another caveat is necessary here. It is well known that the distribution of p when the SIRP process is Gaussian and the covariance matrix is estimated from the data is not Chi Square but *Beta*. However, we have used the distribution postulated by the SIRP model, with a specific objective in mind. In other SIRP processes of interest, say in that of the K Distribution, the distribution of the quadratic form p, when the covariance matrix is unknown but estimated from the data, is not available. Only the SIRP theoretical distribution is available for testing. We are intentionally investigating the efficiency loss, when the substitution of Σ^* by Σ is performed.

PHASE II: UNIVARIATE SIRP's.

The univariate SIRP is just a particular case of the general SIRP for $N = 1$. Hence, all properties of the theoretical model should hold, as with $N > 1$. We still define $X = S * Z$. Only now $Z \sim N(0, 1)$, is a standard normal random variable and X is also univariate. We still investigate the problem of the K Distribution, through a special case: the univariate Laplace. This distribution is easily invertible and hence suitable for a Monte Carlo study. A random variable X is Laplace distributed if:

$$f_X(x) = \frac{1}{2\lambda} exp(-\frac{|x-\mu|}{\lambda}), \quad \lambda > 0$$

To obtain a Laplace univariate SIRP X, let the random variable $w \sim exp(1)$, i.e. exponential with mean unit. Making the transformation $y = \sqrt{2w}$ we obtain a Rayleigh distributed random variable y, with $\mathcal{E}(y^2) = 2$ and density function:

$$f_Y(y) = y\, exp(-\frac{y^2}{2}); \quad y > 0$$

TABLE 10.

PERCENT REJECTIONS FOR N= 2000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.09750	0.04800	0.00850
SIGMA	0.10300	0.05350	0.01350
M-SKEW	0.10400	0.05150	0.00650
M-KURT	0.08600	0.04800	0.01250
P-KNOWN	0.67950	0.61400	0.49500

Bivariate Correlated; Sample Size 200.

TABLE 11.

PERCENT REJECTIONS FOR N= 1000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.10100	0.04900	0.00800
SIGMA	0.10200	0.05000	0.00700
M-SKEW	0.09100	0.04800	0.01200
M-KURT	0.09600	0.04900	0.01200
P-KNOWN	0.63200	0.58200	0.49900

Bivariate Correlated; Sample Size 100.

TABLE 12.

PERCENT REJECTIONS FOR N= 2000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.10250	0.05000	0.00650
SIGMA	0.10250	0.04650	0.00950
M-SKEW	0.09900	0.04650	0.01050
M-KURT	0.10250	0.05000	0.00800
P-KNOWN	0.67450	0.63950	0.56750

Bivariate Correlated; Sample Size 50.

TABLE 13.

PERCENT REJECTIONS FOR N= 1000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
CHOLESKI	0.10900	0.05500	0.01300
SIGMA	0.08500	0.04700	0.00700
M-SKEW	1.00000	1.00000	1.00000
M-KURT	0.13600	0.07900	0.02500
P -KN/KS	0.64800	0.64800	0.63000
P -ES/KS	0.07000	0.02400	0.00100
P -K/CHI	1.00000	1.00000	0.99300

Number of Variates: 8; Sample Size 200.

TABLE 14.

PERCENT REJECTIONS FOR N= 5000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.09780	0.04740	0.00980
X=S*Z/KS	0.09340	0.04480	0.01000
S RY/CHI	0.10420	0.05140	0.01160

Univariate SIRP (Ho); Sample Size 200.

TABLE 15.

PERCENT REJECTIONS FOR N= 10000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.09250	0.04420	0.00880
X=S*Z/KS	0.09670	0.04580	0.00910
S RY/CHI	0.10380	0.05060	0.00870

Univariate SIRP (Ho); Sample Size 100.

However, the resulting covariance matrix of the SIRP $X = y * Z$ is, by definition :

$$\Sigma_X = \mathcal{E}(y * Z)(y * Z)' = \mathcal{E}(y^2)\Sigma_Z \neq \Sigma_Z$$

Hence, such a Rayleigh distributed y is not convenient since $\Sigma_X \neq \Sigma_Z$. We seek an SIRP X with the same covariance matrix as Z. Therefore, we transform our original random variable y to one with a unit expectation by redefining:

$$s = \frac{y}{\sqrt{2}} = \sqrt{w}$$

The resulting random variable s, has now expectation $\mathcal{E}(s) = 1$, as desired yielding an SIRP process $X = S * Z$, with covariance matrix $M = \Sigma_X$. The density of the transformed variable s is now:

$$f_S(s) = 2s\,exp(-s^2)$$

To obtain the distribution of the quadratic form $p = x'x = x^2$, following the SIRP model in the Kaman report, we substitute $f_S(.)$ in $h_N(.)$ for $N = 1$:

$$h_N(p) = \int_0^\infty s^{-N} exp(\frac{-p}{2s^2}) f_S(s) ds = \sqrt{\pi} exp(-\sqrt{2p})$$

From the SIRP theory, the distribution of the quadratic form p is then:

$$f_P(p) = \frac{1}{\sqrt{2\pi}\sqrt{p}} h_1(p) = \frac{1}{\sqrt{2p}} exp(-\sqrt{2p})$$

This is still not a good distribution for testing goodness of fit in a Monte Carlo study. It is more convenient to find an equivalent, well known variable with an easily invertible distribution. We perform the transformation $t = \sqrt{2p}$ and the obtain the random variable $t \sim exp(1)$, exponentially distributed with mean unit, easily invertible for CDF evaluation.

We then test that the quadratic form p is distributed following the SIRP model above derived (H_0), by testing that the distribution of the transformed variable $t = \sqrt{2p}$ is exponential with unit mean.

Hence, $x = s * z$, with s, z and $h_N(p)$ as above, is just a (univariate) SIRP. Following the theory developed in the Kaman Report, we obtain the distribution of the resulting SIRP X as:

$$f_X(x) = \sqrt{2\pi}|\Sigma|^{-1/2} h_N(p) = \frac{1}{\sqrt{2}}\, exp\,(-\sqrt{2}|x|)$$

and we recognize it as a Laplace Distribution with $\lambda = \frac{1}{\sqrt{2}}$.

However, the distribution of this resulting SIRP process X is not convenient for performing a goodness of fit test in a Monte Carlo study. We seek an equivalent random variable, with a better suited distribution, that we can easily invert in our simulation.

Making the transformation $u = \sqrt{2}|x|$, we find how the resulting random variable u follows a univariate exponential with mean unit.

We sampled for $n = 25, 50, 100, 200$ from the univariate $x = s * z$ discussed above. The variables x, s, p were tested for their fits and results computed to obtain empirical significance levels, as in Phase One, for $\alpha = 0.1, 0.05, 0.01$ (Tables 14 through 17). As before, we assess the model by the closeness between the theoretical α and its corresponding empirical significance level α^*.

In Table 14 we show the results for 5000 replications of batches of large sample sizes (e.g. 200). Observe that very close agreement is obtained between α and α^* for significance levels $0.1, 0.05, 0.01$. Agreement is obtained in all three tests: for the SIRP process $X = S * Z$, for the quadratic form $p = X^2$ and for the Rayleigh distributed variable S, which drives the SIRP.

Table 15 shows similar results, now for 10,000 replications and samples of size 100. We can still observe a very close agreement between α and α^*, for significance levels $0.1, 0.05, 0.01$.

Tables 16 and 17 have the same type of results for 10,000 and 20,000 replications and samples of sizes $50, 25$, respectively. We observe, as would be expected, some deterioration of the efficiency in the goodness of fit tests, as sample sizes decrease.

We also analyzed alternatives to the null hypothesis, to assess the efficiency of an SIRP to reject a false hypothesis. Previously, we used the SIRP $X = s * z$, Laplace with $\lambda = \frac{1}{\sqrt{2}}$ (H_0). We now generate the SIRP using $X = y * z$ instead (H_1), obtaining a related SIRP, but with different parameter. In what follows, we investigate the sample size requirements to differentiate one SIRP from the other. In Tables 17 through 20 we show our analysis results, for 5000 replications and for $n = 100, 50, 25, 10$, respectively.

In Table 18 we can see, for batches of size (100), that the SIRP X will be correctly identified as not distributed Laplace with parameter $\lambda = \frac{1}{\sqrt{2}}$. The goodness of fit tests applied to X, achieves an empirical significance level α^*, several times larger than the nominal α. But that of p is two times higher than the one for X.

In Table 19 we show similar results, for sample sizes of 50. It is still plain that the SIRP X is correctly rejected with that sample size. The empirical level α^* for the fit of p is still twice as large as that of the fit test for X. In Table 20 we show similar results for sample of size 25. It is still possible to detect the different SIRP, with such reduced sample size. And the performance of p is still much better than that of X.

It is not until the sample size decreases to 10, in Table 21, that it is possible to confound these two closely related (but different) SIRP processes. This table shows, for 5,000 replications, how α is close to α^* when testing the goodness of fit of X. But the SIRP test through the quadratic function p is still rejecing the null hypothesis. **This is a strong and positive result in favor of the SIRP model.**

From Phase II, we conclude that (i) it is equivalent to test for the fit of a univariate SIRP X,

TABLE 16.

PERCENT REJECTIONS FOR N= 10000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.08760	0.04030	0.00570
X=S*Z/KS	0.08590	0.04030	0.00730
S RY/CHI	0.10230	0.04620	0.01070

Univariate SIRP (Ho); Sample Size 50.

TABLE 17.

PERCENT REJECTIONS FOR N= 20000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.08305	0.03755	0.00745
X=S*Z/KS	0.08235	0.03865	0.00635
S RY/CHI	0.09065	0.04210	0.00780

Univariate SIRP (Ho); Sample Size 25.

TABLE 18.

PERCENT REJECTIONS FOR N= 5000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.84360	0.75240	0.62980
X=S*Z/KS	0.46060	0.29020	0.09480
S RY/CHI	0.99960	0.99900	0.99500

Univariate SIRP (H_1); Sample Size 100.

TABLE 19.

PERCENT REJECTIONS FOR N= 5000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.59200	0.46200	0.23520
X=S*Z/KS	0.26780	0.15640	0.04320
S RY/CHI	0.95900	0.93060	0.83820

Univariate SIRP (H_1); Sample Size 50.

TABLE 20.

PERCENT REJECTIONS FOR N= 5000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.34860	0.23040	0.09160
X=S*Z/KS	0.18400	0.10220	0.02540
S RY/CHI	0.74500	0.61500	0.40400

Univariate SIRP (H_1); Sample Size 25.

TABLE 21.

PERCENT REJECTIONS FOR N= 5000 TOTAL CASES.

METHOD:	ALPHA=0.10	ALPHA=0.05	ALPHA=0.01
P=X'X/KS	0.19600	0.11560	0.03280
X=S*Z/KS	0.10080	0.06940	0.01640
S RY/CHI	0.00000	0.00000	0.00000

Univariate SIRP (H_1); Sample Size 10.

directly on X or through its quadratic function p, derived following the SIRP theoretical model. Also (*ii*) that it is possible to perform this equivalent test with excellent results for samples as small as 50 data points. Finally, that (*iii*) when the postulated model is not true, even when it is closely related as in the above case, it is possible to detect and reject such a false hypothesis through the quadratic function p, even more efficiently, than through the fit of the SIRP process X itself.

CONCLUSIONS.

The Monte Carlo study above described has demonstrated, in $N > 2$ dimensions, for the special case of a multivariate Gaussian SIRP, that the model works as postulated. We have also shown that the quadratic form p can actually be obtained and identified for samples down to size 50 and eight variables, when the covariance Σ is known. We also shown that this quadratic form p can also approximate its theoretical distribution when sample sizes are large (say 200 or more). Finally, we have shown that the quadratic function p can correctly discriminate different SIRP models, with high power, when the covariance matrix is known.

We have shown, for $N = 1$, for the general case of SIRP, that the theoretical model works as postulated and that we can use the quadratic function p to test that the SIRP model is the one specified, with samples as small as 50 data points.

Moreover, we have shown how it is possible to discriminate an erroneously postulated SIRP model, on the basis of the univariate quadratic function p. And we have shown how such test based on p is more powerful than the test based on the complete (multivariate) SIRP process X.

This last result is of importance in radar studies. For, if the distribution of certain types of radar input can be predetermined, a test using p can effectively be implemented to recognize these patterns.

Future research on this direction is recommended.

ACKNOWLEDGEMENTS.

The author greatfully acknowledges the cooperation and fruitful interaction with Drs. Don Weiner and Muraly Rangaswamy, of Syracuse University and Jim Michels, of Rome Labs.

REFERENCES

Allen, A. O., *Probability, Statistics and Queueing Theory*, Academic Press, 1978.

Anderson, T. W., *An Introduction to Multivariate Analysis*, Wiley, 1984.

Bratley, P; Fox, B. and L. Schrage, *A Guide To Simulation*, Springer-Verlag, 1983.

Cambanis, S.; Huang, S. and G. Simons, *On the Theory of Elliptically Contoured Distributions*, J. Multivar. Annal. 11 (1981), 368–385.

Chmielewski, M. A., *A Re-Appraisal of Tests for Normality*, Comm. Stat. - Theor. Meth. a10(20) (1981), 2005–2014.

Dudewicz, E. J. and T. G. Ralley, *The Handbook of Random Number Generation and Testing With TESTRAND Computer Code*, American Sciences Press Inc., 1981.

Gnanadesikan, R., *Methods of Statistical Data Analysis of Multivariate Observations*, Wiley, 1977.

Goel, A. L., *Software Rehability Modelling and Estimation Techniques*, RADC-TR-82-263, Griffiss AFB, Rome, NY, 1982.

Johnson, M. E., *Multivariate Statistical Simulation*, Wiley, 1987.

Johnson, N. L. and S. Kotz, *Distributions in Statistics: Continous, Univariate Distributions*, Wiley, 1970.
Johnson, M. E., Chiang, W. and J. S. Ramberg, *Generation of Continuous Multivariate Distributions for Statistical Applications*, Amer. Jour. Math. Manag. Sci. 4 (1984), 225–248.
Johnson, R. A. and D. E. Wichern, *Applied Multivariate Analysis*, Prentice Hall, 1982.
Kaman Sciences, *Contract Report Number F30602-89-C-0082, Task 10*, Rome Labs Technical Report (to appear), 1992.
Kendall, M. G. and A. Stuart, *The Advanced Theory of Statistics*, (Vols. I, II and III), Charles Griffin and Co., London, 1966.
Koziol, J. A., *A Class of Invariant Procedures for Assessing Multivariate Normality*, Biometrika 69 (1982), 423–427.
Koziol, J. A., *On Assessing Multivariate Normality*, JRRS-B 45 (1983), 358–361.
Koziol, J. A., *Assessing Multivariate Normality: A Compendium*, Comm. Stat. 15 (1986), 2763–2783.
Loh W., *Testing Multivariate Normality by Simulation*, Jour. Statist. Comput. Simul. 26 (1986), 243–252.
Mardia, K. V., Kent, J. T. and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979.
Mardia K. V., *Measures of Multivariate Skewness and Kurtosis With Applications*, Biometrika 57 (1970), 519–530.
Ozturk, A. and J. L. Romeu, *A New Graphical Test for Multivariate Normality*, Comm. in Statist. (Simula.) 21(1) (1992).
Press, W. H.; Flannery, B. P.; Teukolsky, S. A. and W. T. Vetterling, *Numerical Recipes: the Art of Scientific Computing*, Cambridge University Press, 1986.
Rangaswamy, M.; Weiner, D. and A. Ozturk, *Computer Generation of Correlated Non Gaussian Clutter for Radar Signal Detection*, IEEE Trans. Aerosp. Electr. Sys. (1992 (to appear)).
Rangaswamy, M.; Weiner, D. and A. Ozturk, *Simulation of Correlated Non Gaussian Interference for Radar Signal Detection*, Proceedings of the 25th Annual ASILOMAR Conference on Signals, Systems and Computers (1991).
Romeu, J. L., *A Simulation Approach for the Analysis and Forecast of Software Productivity*, Journal of Computers and Industrial Engineering 9(2) (1985).
Romeu, J. L., *Teaching Engineering Statistics With Simulation: A Classroom Experience*, Journal of the Institute of Statisticians 35(4) (1986).
Romeu, J. L., *Another Look at the Comparison of the Non Overlapping Batch Means and Area STS Simulation Output Analyses Procedures*, Actas del ISORBAC-2, San Sebastian, 1988.
Romeu, J. L., *A Small Sample Monte Carlo Study of Four System Reliability Bounds*, Journal of Computers and Industrial Engineering 16(1) (1989).
Romeu, J. L., *Development and Evaluation of a General Procedure for Assessing Multivariate Normality*, CASE Center Technical Report 9022. Syracuse University, NY. 13244, 1990.
Romeu, J. L., *A New Multivariate Normality Goodness of Fit Test With Graphical Applications*, Proceedings of the Computers and Industrial Engineering Conference (1991).
Romeu, J. L., *Small Sample Empirical Critical Values as a Tool for the Comparison of Multivariate Normality Goodness of Fit Tests*, Proceedings of the Conference on the Interface Between Statistics and Computer Science (1992a).
Romeu, J. L., *Validation of Multivariate Monte Carlo Studies*, Proceedings of the International Meeting of Statistics in the Basque Country (IMSIBAC-4) (to appear) (1992b).
Shapiro, S. and A. Gross, *Statistical Modeling Techniques*, Marcel Dekker, 1981.
Tong, Y. L., *The Multivariate Normal Distribution*, Springer-Verlag, 1990.

HIERARCHICAL AND INTEGRATED MODELING AND SIMULATION

Robert G. Sargent
Professor

Simulation Research Group
L.C. Smith College of Engineering and Computer Science
Syracuse University
439 Link Hall
Syracuse, NY 13244

Final Report for:
AFOSR Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

August 1992

HIERARCHICAL AND INTEGRATED MODELING AND SIMULATION

Robert G. Sargent
Professor
Simulation Research Group
L.C. Smith College of Engineering and Computer Science
Syracuse University

Abstract

This short paper discusses the three basic approaches to hierarchical and integrated (discrete event) modeling and simulation: software, metamodeling, and "coupling of models" in a new modeling paradigm. Also the desired requirements in a new modeling paradigm for (discrete event) modeling are given.

HIERARCHICAL AND INTEGRATED MODELING AND SIMULATION

Robert G. Sargent

I. INTRODUCTION

In this paper we discuss hierarchical and integrated modeling and simulation. (We are restricting ourselves to discrete event modeling and simulation). With the increased use and interest in large-scale simulation and in model reuse, hierarchical and integrated modeling and simulation is becoming more important to the Air Force [1,7,10,14], Department of Defense (DoD), and other users. Currently, the technology does not, in general, exist for hierarchical and integrated modeling and simulation. Thus, commercial simulation languages do not provide for this capability. Hierarchical and integrated modeling and simulation is starting to receive attention from researchers.

Figure 1 is a generic example of a three-level hierarchical model. This could represent, for example, a simulation model (M_1) of two aircraft battling each other, where submodels $M_{1,1}$ and $M_{1,2}$ are each aircraft, and the components (e.g. $M_{1,1,1}$ and $M_{1,1,2}$) are aircraft components such as radar, weapon systems, etc. (An objective of such a hierarchical model could be the ability to replace $M_{1,1}$ or $M_{1,2}$ with another to simulate different aircraft and also to replace the aircraft components with other component models, e.g. replace $M_{1,1,1}$, to determine their effectiveness.) Hierarchical modelling is also referred to as vertical integration in DoD. Integrated modeling and simulation is another term used, but this often refers to only the integration of submodels into a model, i.e. only a two level hierarchical model.

We are assuming here that the submodels (e.g. $M_{1,1}$ and $M_{1,2}$) at each level interact with at least one other submodel or model during a simulation. This contrast with the situation where each submodel is run separately and used as input to a higher level model. In this latter way, the execution of each model would take place sequentially and is not the subject of discussion here.

HIERARCHICAL AND INTEGRATED MODELING AND SIMULATION

Robert G. Sargent
Professor
Simulation Research Group
L.C. Smith College of Engineering and Computer Science
Syracuse University

Abstract

This short paper discusses the three basic approaches to hierarchical and integrated (discrete event) modeling and simulation: software, metamodeling, and "coupling of models" in a new modeling paradigm. Also the desired requirements in a new modeling paradigm for (discrete event) modeling are given.

HIERARCHICAL AND INTEGRATED MODELING AND SIMULATION

Robert G. Sargent

I. INTRODUCTION

In this paper we discuss hierarchical and integrated modeling and simulation. (We are restricting ourselves to discrete event modeling and simulation). With the increased use and interest in large-scale simulation and in model reuse, hierarchical and integrated modeling and simulation is becoming more important to the Air Force [1,7,10,14], Department of Defense (DoD), and other users. Currently, the technology does not, in general, exist for hierarchical and integrated modeling and simulation. Thus, commercial simulation languages do not provide for this capability. Hierarchical and integrated modeling and simulation is starting to receive attention from researchers.

Figure 1 is a generic example of a three-level hierarchical model. This could represent, for example, a simulation model (M_1) of two aircraft battling each other, where submodels $M_{1,1}$ and $M_{1,2}$ are each aircraft, and the components (e.g. $M_{1,1,1}$ and $M_{1,1,2}$) are aircraft components such as radar, weapon systems, etc. (An objective of such a hierarchical model could be the ability to replace $M_{1,1}$ or $M_{1,2}$ with another to simulate different aircraft and also to replace the aircraft components with other component models, e.g. replace $M_{1,1,1}$, to determine their effectiveness.) Hierarchical modelling is also referred to as vertical integration in DoD. Integrated modeling and simulation is another term used, but this often refers to only the integration of submodels into a model, i.e. only a two level hierarchical model.

We are assuming here that the submodels (e.g. $M_{1,1}$ and $M_{1,2}$) at each level interact with at least one other submodel or model during a simulation. This contrast with the situation where each submodel is run separately and used as input to a higher level model. In this latter way, the execution of each model would take place sequentially and is not the subject of discussion here.

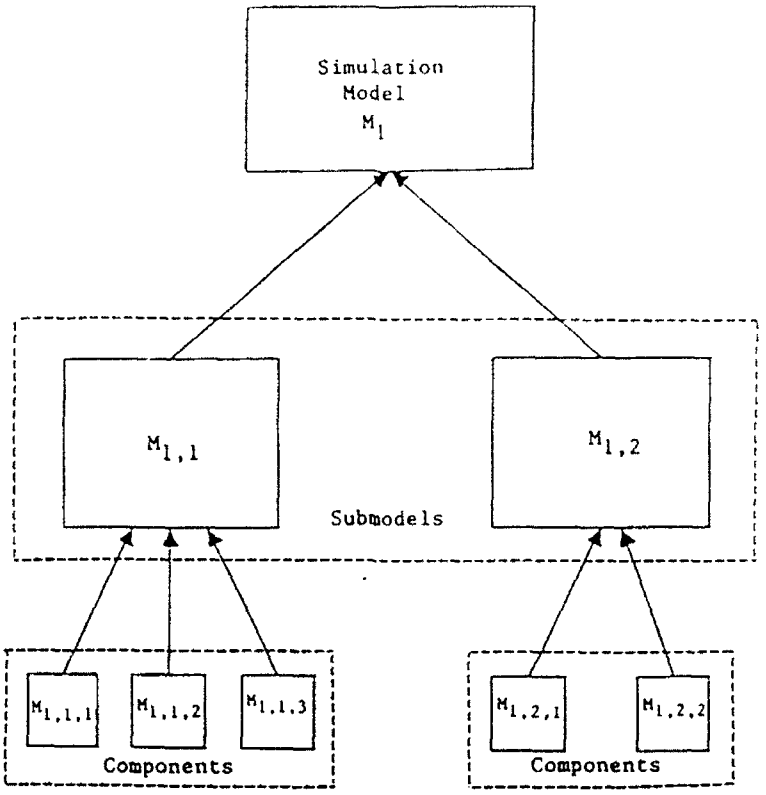


Figure 1. Hierarchical Modeling

We note that most large-scale and DoD simulations are computational intense and thus it is desirable in the future to be able to execute simulations on parallel and distributed computers. Furthermore, we note that it is desirable to be able to easily replace lower level models (e.g. submodels and components in Figure 1) with other models and therefore it is desirable to use the modularity concept [6]. This would also allow for reusable of models. Also, one must be concerned with verification and validation [13].

In the author's view, there are three basic approaches to accomplish hierarchical and integrated models for simulation. The first is through the use of software and this is discussed in Section 2. The second is through the use of metamodels and this way is discussed in Section 3. The third approach is the use of "coupling of models to have closure" through

the development of new model paradigms. Section 4 discusses what is required of a new model paradigm for hierarchical and integrated modeling simulation and briefly mentions one new paradigm under development. The first approach is currently technically feasible but is usually costly. The latter two approaches need research and development.

II. SOFTWARE APPROACH

This author sees three general ways of accomplishing hierarchical and integrated modeling using the software approach. The first way, which is discussed in earlier work by the author [11], is combining (or joining) separate simulation models into "one" simulation model. If the simulation models already exist, then extensive software modification may be required. This joining of simulation models is occasionally done [16].

The second way is to develop a specific software framework (architecture or backbone) for each specific application domain. This way is generally used only for those simulation models where a set of submodels are joined to form a simulation model; i.e. there are only two levels. The "backbone" requires the submodels that are to be connected together (joined) to have a "specific interface" to the backbone. This may require writing an interface to an existing model in order for it to be used. The current development of NCTI (Non-Cooperative Target Identification) Mod I Simulation [9] is an example of this approach.

The third way, is to use message passing and objects. A software framework is needed and models developed for it. This approach requires some overall controlling mechanism and models specifically developed for it. This is currently proposed, e.g. the proposed J-MASS (Joint Modeling and Simulation System) [7] and SAMSON (Simulation and Modeling Supporting Operational Needs) [7] appear to use this way of obtaining hierarchical and integrated simulation models. This will allow model reuse if the models are **appropriately** developed. However, since only a single "event list" will, in general, be used in the controlling mechanism, only limited parallelism will be possible.

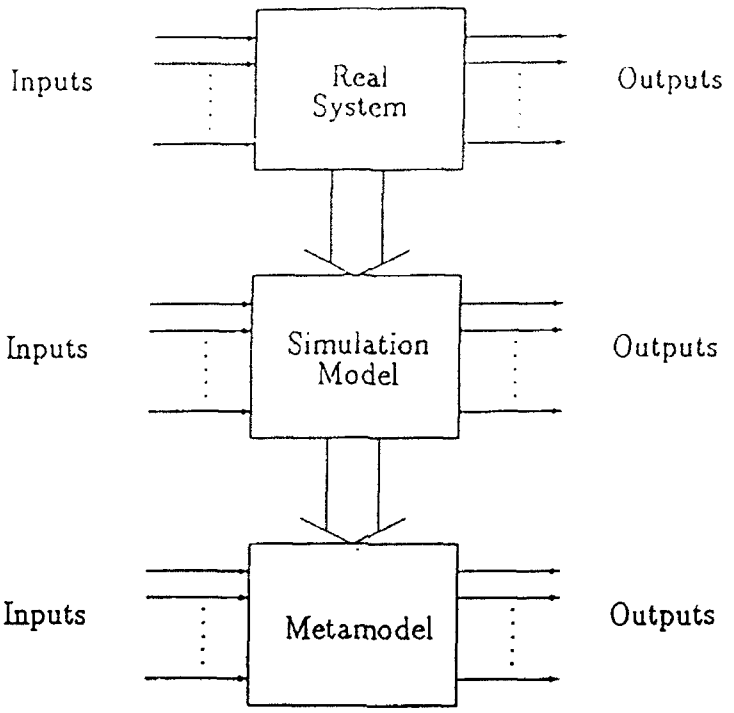


Figure 2. Metamodeling

The software approach is currently feasible. However, it is usually costly to use and models must be specifically developed or modified to use this approach.

III. METAMODEL APPROACH

A metamodel [8] is a model of a model. Figure 2 illustrates the relationships of a real ystem, a simulation model, and a metamodel. We note that a metamodel is an empirical (blackbox) model, often a linear least squares polynomial model, while a simulation model is a causal (mechanism) model. The idea of metamodeling is to replace a simulation model with a simple mathematical equation relating the inputs and the outputs(s) (or response).

In the metamodel approach to hierarchical and integrated modeling, metamodels are used for the lower level models instead of simulation models. This approach requires a metamodel to be developed for each submodel. We note that metamodeling is an evolving field and that there are numerous research issues to be solved [12].

Metamodels are usually expected value models of random (Monte Carlo) simulation models and deterministic models of deterministic simulation models [8,15,20]. This implies that nonrandom values are what are usually the outputs of metamodels. Thus, if the random variability of a simulation model is of interest, then metamodels must have a random component developed. This is not now currently done and is a research topic.

Military simulation models are often what are called terminating simulation models (i.e. nonsteady-state models) and the performance measures (responses) of interest are usually time-dependent. Metamodels have not yet been developed for time-dependent responses (at least as known by this author). If they would be developed for these types of simulation models, the "curse of dimensionality" would generally apply because there are usually many responses of interest and several input scenarios of interest (e.g. if there were 10 responses of interest and 15 input scenarios, there would be 150 responses that would have to be modelled.)

This author believes that the use of the metamodelling approach to hierarchical and integrated modeling will be extremely limited in the near future. This is because there are numerous research issues yet to be solved for "straight-forward" metamodeling, it is currently costly to develop metamodels--in particular if the experimental region (i.e. range of input variables) is large, and metamodels have yet to be developed to be random and to be time-dependent. Thus, considerable research is needed to make metamodeling practical as a general approach to hierarchical and integrated modeling. (However, the use of metamodels will continue to

evolve for their current uses which are very important. See [15] and [20] as an example of metamodels of an air force simulation model.)

IV. NEW PARADIGM APPROACH

Another approach to providing hierarchial and integrated modeling and simulation capability is to develop a new model paradigm that includes this capability. The hierarchical and integrated modeling would, in all likelihood, require that models be "closed under coupling". Zeigler [17,18,19] has developed what is required for closure under coupling and one specific form called DEVS-Scheme.

This author believes that a new paradigm for modeling and simulation should have more than just hierarchical and integrated modeling capability. Based upon my years of experience and further investigation this summer, I believe that a new paradigm should satisfy the following requirements.

* *GENERAL PURPOSE* - the *modelling paradigm* should allow the modeler to model a wide variety of problem types and domains; it should not be primarily for one type of system, e.g. queueing systems or transaction oriented systems.
* *THEORETICAL FOUNDATION* - a theoretical foundation should underlie a model paradigm if modelling is to be moved towards a science.
* HIERARCHICAL CAPABILITY - a modelling paradigm should allow hierarchical modelling so that complex systems can be more easily modelled.
* COMPUTER ARCHITECTURE INDEPENDENCE - the model paradigm should be such that a model can be executed on different computer architectures (e.g. sequential, distributed, or parallel, and be able to take advantage of the architecture that it is executing on) and be transparent to the modeler. This requires that such items as "lookahead" information required for parallel simulation be available from the model itself and not have to be specially added

by the modeler.

* STRUCTURED - a structured approach that guides the user in model development, including hierarchical modelling, should be part of the model paradigm.
* MODEL REUSE - the model paradigm should allow models and submodels to be easily reused and support a model database.
* SEPARATION OF MODEL AND EXPERIMENTAL FRAME - both the model's input and the model's output should be able to be separated from the model itself in the model paradigm.
* GRAPHICAL/VISUAL MODELLING - the model paradigm should allow the capability to have graphical/visual modelling.
* EASE OF MODELLING - the model paradigm should allow a world view(s) of modelling to be used that is easy to model with.
* EASE OF COMMUNICATION - the conceptual model(s) allowed by the model paradigm should be easy to communicate to other parties.
* EASE OF MODEL VALIDATION - the model paradigm should support both conceptual and operational validity.
* ANIMATION - the model paradigm should allow animation to be accomplished without difficulty.
* MODEL DEVELOPMENT ENVIRONMENT - a model development environment can aid in the steps of model development and a model paradigm should allow the use of such an environment.
* EFFICIENT TRANSLATION TO EXECUTABLE FORM - the model paradigm should be capable of allowing efficient model translation to executable code. The paradigm should allow the model to automatically be converted to computer code or allow ease of program verification if it does not.

To accomplish these requirements (which includes closure under coupling), modularity and encapsulation will be required [6]. While the object-oriented approach as provided in object-oriented languages may be sufficient to handle the requirements needed for closure under coupling in

order to achieve hierarchical and integrated modeling, this author believes that the object oriented approach is not sufficient to meet all of the above requirements [6]. Furthermore, this author believes that some type of model representation is needed (where model representation allows for model analysis to be performed on the representation) in order to develop algorithms for the simulation to run on different types of computer architectures such as parallel computers that are transparent to the user.

No "artificial intelligence" capability or "expert systems" were included in the above paradigm requirements. These may be useful but this author does not believe that they are required for the model paradigm he visualizes.

This author (and his graduate students) is in the process of developing a new model paradigm that satisfies the above requirements and also provides some new foundations for discrete event simulation. A new version of the Process-Interaction World View as a modeling approach has been developed along with its theoretic formulation [6]. A model representation based on this new modeling approach called "Control Flow Graphs" has been developed [2,3, and 5]. Algorithms have been developed based on Control Flow Graphs for simulation models to execute on different types of computers that are transparent to the users [4]. The algorithms currently need evaluation and optimization. Research work is now beginning on developing a modeling language, on the details of a structured hierarchical modeling approach (the necessary requirements are satisfied in our new paradigm), and on how this paradigm could be implemented in a (object-oriented) language.

V. SUMMARY AND CONCLUSIONS

We have briefly described the three basic approaches that the author is aware of for providing hierarchical and integrated (discrete event) modeling and simulation. The first approach using software is currently

technically feasible but is usually costly. The second approach using metamodels requires considerable research before becoming feasible as a general approach and may also be costly to use. The third approach based on developing new paradigms requires some research but can be developed in the very near future. This author believes that the latter approach is the way to proceed for the longer term and should be supported and developed. More than one such paradigm should be developed and experimented with in order to learn from them for future improvements. Furthermore, metamodels could be used (with some research and development) within the new paradigm by using metamodels for some of the submodels.

Acknowledgements

The author thanks William Gregory and Alex F. Sisti (of Rome Laboratory) for their assistance during his summer visit at Rome Laboratory.

REFERENCES

[1] Air Force Acquisition Simulation & Modeling Architecture (AFASMA), Overheads of Kickoff Meeting, June 30 - July 1, 1992, Electronic Systems Division, Hanscom AFB, MA.

[2] Cota, B.A. and R.G. Sargent. 1989. Automatic Lookahead Computation for Conservative Distributed Simulation. CASE Center Technical Report 8916, CASE Center, Syracuse University, December 1989.

[3] Cota, B.A. and R.G. Sargent. 1990. A Framework for Automatic Lookahead Computation in conservative distributed simulations. In Distributed Simulation, ed. D. Nicol, the Society for Computer Simulation, 1990.

[4] Cota, B.A. and R.G. Sargent. 1990. Simulation Algorithms for Control Flow Graphs. CASE Center Technical Report 9023, CASE Center Syracuse University, November 1990.

[5] Cota, B.A. and R.G. Sargent. 1990. Control Flow Graphs: A Method of Model Representation for Parallel Discrete Event Simulation. CASE Center Technical Report 9026, CASE Center, December 1990.

[6] Cota, B.A. and R.G. Sargent. 1992. A Modification of the Process Interaction World View, forthcoming in *ACM Transactions on Modeling and Computer Simulation*. (An earlier version was published as A New Version of the Process World View for Simulation Modeling, CASE Center Technical Report 9003, Syracuse University, February 1990.)

[7] J-MASS/SAMSON Briefing Materials, 4-5 February 1992, Office of Aerospace Studies, Kirtland AFB, N.M.

[8] Kleijnen, J.P.C. 1987. Statistical Tools for Simulation Practitioners, Marcel Dekker, 1987.

[9] NCTI Simulation MOD I Build (Rome Laboratory), Overheads for Kickoff Briefing, August 17, 1992, Synectics Corporation.

[10] Report of the Ad Hoc Committee on Modeling and Simulation, United States Air Force Scientific Advisory Board, December 1991.

[11] Sargent, R.G. 1986. Joining Existing Simulation Programs,

Proceedings of the 1986 Winter Simulation Conference, J. Wilson, J. Henriksen, and S. Roberts, eds., pp. 512-516.

[12] Sargent, R.G. 1991. Research Issues in Metamodeling, *Proceedings of the 1991 Winter Simulation*, B.L. Nelson, W.D. Kelton, and G.M. Clark, eds., pp. 37-47.

[13] Sargent, R.G. 1991. Simulation Model Verification and Validation, *Proceedings of the 1991 Winter Simulation*, B.L. Nelson, W.D. Kelton, and G.M. Clark, eds., pp. 888-893.

[14] Sisti, A.F. 1989. A Model Integration Approach to Electronic Combat Effectiveness Evaluation, RADC-TR-89-183.

[15] Tew, J.D., M.A. Zeimer, R.G. Sargent, and A.F. Sisti. 1992. Metamodel Procedures for Air Engagement Simulation Models, forthcoming Rome Laboratory (AFMC) Technical Report.

[16] Two Way Sensor Interface, Interim Technical Report for Rome Laboratory, Contract F30602-89-C-0215, January 1991, by PAR Government Systems Corporation (PGSC Report 90-50).

[17] Zeigler, B.P. 1976. *Theory of Modelling and Simulation*. Wiley, 1976.

[18] Zeigler, B.P. 1984. *Multifaceted Modelling and Discrete Event Simulation*, Academic Press, 1984.

[19] Zeigler, B.P. 1990. *Objective-Oriented Simulation with Hierarchical Modular Models*, Academic Press, 1990.

[20] Zeimer, M.A., J.D. Tew, and R.G. Sargent. 1992. A Metamodel Application using TERSM, Final Report for AFOSR 1992 Summer Research Program, Rome Laboratory.

THIS PAGE INTENTIONALLY LEFT BLANK

METAMODEL APPLICATIONS USING TERSM

Michael A. Zeiner
Graduate Student

and

Dr. Jeffrey D. Tew
Assistant Professor

Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

and

Dr. Robert G. Sargent
Professor

Simulation Research Group
L. C. Smith College of Engineering and Computer Science
Syracuse University
439 Link Hall
Syracuse, New York 13244

Final Report for:
Summer Research Program
Rome Laboratory

Sposored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D. C.

September 1992

METAMODEL APPLICATIONS USING TERSM

Michael A. Zeimer
Graduate Student

and

Dr. Jeffrey D. Tew
Assistant Professor

Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University

and

Dr. Robert G. Sargent
Professor

Simulation Research Group
L. C. Smith College of Engineering and Computer Science
Syracuse University

Abstract

Tactical simulation models are often used to assess vulnerabilities and capabilities of combat systems and doctrines. Due to the complexity of tactical simulation models, it is often difficult to assess the relationship between input factors and the performance of the simulation model. To facilitate this type of assessment, simulation analysts often use the simulation model to emirically construct a *black-box* approximation of the causal and time dependent behavior of the simulation model. This type of approximation is known as a *metamodel* and can be viewed as a summary of the behavior of the simulation model. We demonstrate this technique in the context of an example using TERSM (Tactical Electronic Reconnaissance Simulation Model). The results indicate that metamodeling is applicable to tactical simulation models and that the technique has a wide range of uses.

1. Introduction

Tactical simulation models are often employed by the Department of Defence to assess the capabilities and vulnerabilites of various combat systems and doctrines. These simulation models are usually highly complex and of relatively high dimensionality. That is, the performance of the simulation model is dependent on a large number of parameters or input factors that act and interact in a complex manner. Thus, it is often difficult to assess the relationship of individual input factors to the performance of the simulation model. Recently, a technique known as *metamodeling* has generated interest in the simulation community for its ability to facilitate this type of assessment.

A metamodel is a mathematical approximation of the relationship between a set of input factors and one or more responses. Metamodels are usually estimated empirically via experimentation with a simulation model, and thus, metamodels are models of models. With respect to a given response, a metamodel is *black-box* approximations of the causal (mechanistic) and time dependent behavior of a simulation model. Figure 1 depicts the relationships among the real system, the simulation model, and the metamodel.

In this report, we introduce metamodeling and illustrate its applicability to the analysis of tactical simulations. In Section 2, we summarize the mathematical and statistical concepts and notation of metamodeling. In Section 3, we present an example using TERSM (Tactical Electronic Reconnaissance Simulation Model). In Section 4, we present some conclusions.

2. Metamodels

Metamodels can have various forms, but we restrict our attention to the most commonly used class of models: least squares models. To simplify the discussion, we focus on polynomial and simple transformed response polynomial models of the forms

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

and

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

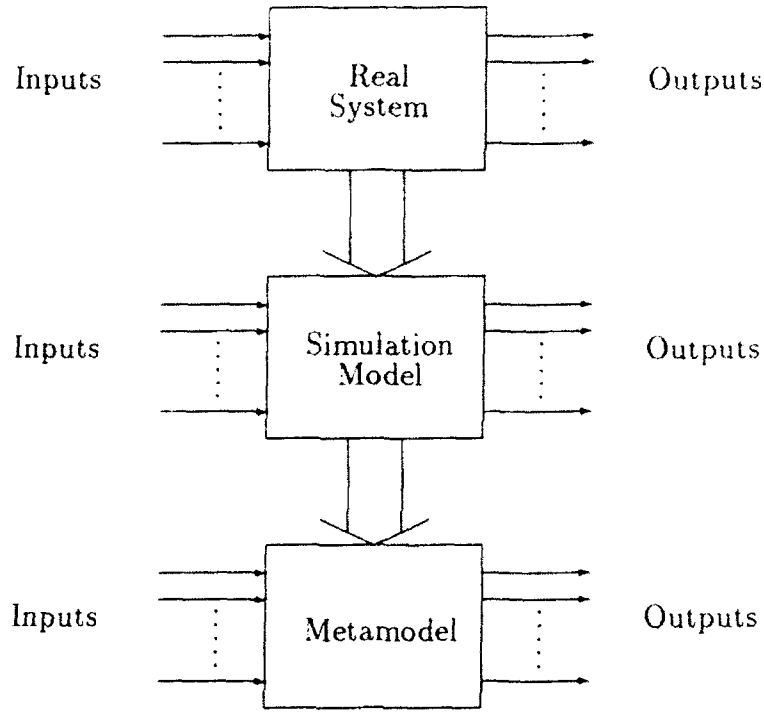


Figure 1: Relationships Among the Real System, Simulation Model, and Metamodel.

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ data matrix containing the levels of the input factors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown metamodel coefficients, $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms, and $\mathbf{y}^*$ is a vector of transformed responses. The transformation on y can be any real function over the range of the untransformed response. Functions such as the square root and natural logarithm are often used to linearize sets of observations in order to obtain simpler and/or better approximations of system behavior.

For example, the relationship between a pair of factors, x_1 and x_2, and a response, y, may have the polynomial form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon \tag{3}$$

or the transformed response polynomial form

$$\sqrt{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon. \tag{4}$$

Both metamodels are said to be linear models since all coefficients have power one.

The type of metamodel to use is often dictated by the purpose of the metamodel and by properties of the system. Metamodels are usually employed for one or more of the following purposes:

1. studying system behavior,
2. predicting responses,
3. sensitivity analysis, or
4. optimization.

Depending on the purpose of the metamodel, the form and the fineness of the approximation may vary greatly. For example, a simple linear approximation is often adequate for studying some elements of system behavior such as the degree to which certain factors affect the response while nonlinear approximations may be more appropriate for prediction.

Some of the important properties of the system that influence the type of model used include:

1. characteristics of the response (discrete or continuous, qualitative or quantitative, random or deterministic, etc.),
2. characteristics of the input factors (discrete or continuous, qualitative or quantitative, random or deterministic, etc.), and
3. dimensions of the experimental region.

In this report, we restrict our attention to systems with quantitative, continuous responses; and quantitative, deterministic input factors. Metamodels for these systems can be estimated using the *method of least squares*. We consider both random and deterministic response cases. Metamodels can also be obtained using more advanced techniques which are beyond the scope of this report. The techniques outlined in the following subsections are applicable to random responses in general, but a subset of the outlined techniques are applicable to the deterministic response case as well. Thus, we will explain all the techniques in terms of the random response case and note exceptions for the deterministic response case.

In Section 2.1, we discuss least squares model estimation. In Section 2.2, we briefly summarize a pair of statistical analysis analysis tools called *analysis of variance* and *statistical inference*. In Section 2.3, we discuss the difference between the concepts of pure error and lack-of-fit in metamodels. In Section 2.4, we briefly discuss some measures and methods for determining the validity of metamodels. In Section 2.5, we introduce and briefly discuss some techniques for efficiently designing experiments. Finally, in Section 2.6, we add some perspective to material in Sections 2.1-2.5 by outlining a general metamodeling process.

2.1. Least Squares Metamodel Estimation

To illustrate the method of least squares, consider a set of observations (y_i, $i = 1, 2, \ldots, n$) and corresponding set of factor levels (x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$) given by

$$\begin{array}{cccccc} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{array}.$$

Suppose we postulate a model given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \varepsilon. \tag{5}$$

In this case the number of parameters, p, is equal to five, corresponding to the number of coefficients in the postulated model. This same model can be written in general vector notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Now suppose that we obtain an estimated model given by

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \tag{6}$$

where $\hat{\mathbf{y}}$ is an $n \times 1$ vector of estimated responses, and $\mathbf{b}$ is a $p \times 1$ vector of estimated model coefficients. The $n \times 1$ vector of deviations of the observations about the fitted model, called the vector of residuals, is given by

$$\mathbf{e} = \hat{\mathbf{y}} - \mathbf{y}. \tag{7}$$

The least squares estimator of $\boldsymbol{\beta}$ is obtained by setting the derivative with respect to $\mathbf{b}$ of the sum of squared residuals equal to zero, such that

$$\frac{\delta}{\delta \mathbf{b}}[\mathbf{e}'\mathbf{e}] = \mathbf{0} \tag{8}$$

$$\frac{\delta}{\delta \mathbf{b}}[(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})] = \mathbf{0} \tag{9}$$

$$-2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{0}. \tag{10}$$

Simplification leads to the least squares estimator

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{11}$$

(see Myers 1990, p.88). Thus, least squares estimates are estimates for which the unweighted sum of squared residuals is minimized.

For example, consider the estimated simple linear regression model given by

$$\hat{y} = b_0 + b_1 x.$$

The ith predicted response, observed response, and residual are denoted by $\hat{y}_i$, y_i, and e_i respectively. These elements are graphically depicted in Figure 2. The given model is a least squares model if and only if the sum of the squared vertical distances from each observation to the fitted model is minimized.

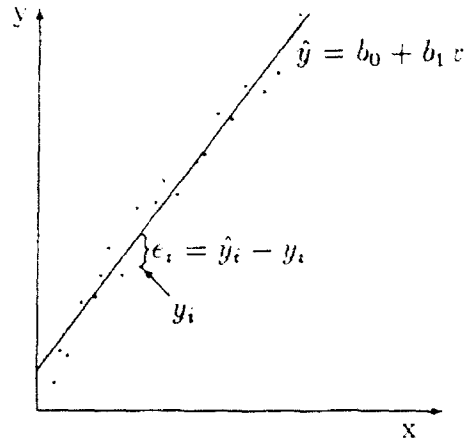


Figure 2: Illustration of Least Squares Principal.

2.2. Analysis of Variance and Statistical Inference

Analysis of variance (ANOVA) and *statistical inference* are statistical methods that are commonly employed to quantify the importance of factors with respect to a given response. They are extremely useful and powerful because they allow the analyst to make statements concerning the statistical significance of various factors. For example, these methods could be used to determine if it is *likely* that x_1x_2 in the metamodel given by equation (5) affects the response. Without ANOVA and statistical inference, randomness and/or lack of fit between the metamodel and the observed data make such conclusions difficult to reach.

The principal drawback of ANOVA and statistical inference is that they require certain assumptions concerning the behavior of $\boldsymbol{\epsilon}$. In particular, the assumption that $\boldsymbol{\epsilon}$ be normally and independently distributed with homogeneous variance is required. This assumption implies that there must be some random *noise* in the response which causes observed responses to be normally distributed with equal dispersion about the estimated model, independent of the location in x-space. These methods can still be employed in violation of the assumptions, but the results are unpredictable, undependable, and should be treated with suspicion especially when the response is

not random (i.e. the deterministic response case).

To illustrate ANOVA and statistical inference, consider the ANOVA table for the metamodel given by equation (5), shown in Table 1. The purpose of ANOVA table construction is to break-down variability in the response and assign portions of the variability to sources of variation based on the observed contribution of each source. Contributions to variability are measured using sums of squares, and the corresponding degrees of freedom represent restriction on the calculation of sums of squares.

Table 1

ANOVA Table for the Metamodel Given by Equation (5)

Source of Variation	Degrees of Freedom	Sum of Squares
Metamodel	5	$\mathbf{y'X(X'X)^{-1}Xy}$
Error	$n-5$	$\mathbf{y'y - y'X(X'X)^{-1}Xy}$
Total	n	$\mathbf{y'y}$

In addition to the basic ANOVA table, it is also possible to subdivide the metamodel sum of squares given in Table 1 in order to account for the variability due to individual model terms. For the metamodel given by equation (5), the variability due to x_1x_2 is given by

$$\mathbf{y'X(X'X)^{-1}Xy - y'X_2(X_2'X_2)^{-1}X_2y}, \tag{12}$$

where $\mathbf{X_2}$ is the data matrix without the fourth column (which corresponds to x_1x_2).

Sums of squares are used to measure the variability in the response because they possess useful distributional properties when our assumptions concerning ϵ hold. We can take advantage of these distributional properties to conduct statistical hypothesis tests (see Myers 1990, p. 95-125). An hypothesis test is a formal means of quantifying the probability that an assertion is incorrect. For example, consider the hypothesis

that $\beta_{12} = 0$, or in other words, that the interaction between x_1 and x_2 is insignificant. In the formal notation of hypothesis testing, this can be stated as

$$H_0 : \beta_{12} = 0 \text{ versus } H_1 : \beta_{12} \neq 0.$$

If the null hypothesis is correct, then under the given assumptions

$$\frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} - \mathbf{y}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2\mathbf{y}}{s^2} = F_{\alpha,1,n-p}, \tag{13}$$

where $F_{\alpha,1,n-p}$ is α point of an F-distribution with 1 and $n-p$ degrees of freedom (see Myers and Milton 1991, p. 116). The value of α for which equation (13) holds is called the *p-value* and is the probability that H_0 is true. Thus, a very low p-value for the given test indicates that it is highly likely that β_{12} explains a significant portion of the variability in the response. Such coefficients are said to be statistically significant.

2.3. Pure Error and Lack Of Fit

The experimental error, $\boldsymbol{\epsilon}$, represents the inability of a metamodel to determine $\mathbf{y}$ exactly. Experimental error is comprised of the effects of all extraneous factors that are not monitored in a study or experiment. The assumption that $\boldsymbol{\epsilon}$ is independently distributed with homogeneous variance depends on the observed factors being uncorrelated with unobserved factors. If this is not the case, then at least some of the error will be location-dependent and is called lack-of-fit (LOF) (see Myers 1990, pp. 117-120).

To illustrate, consider a postulated model given by

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^* \tag{14}$$

and the theoretically correct, or *true*, model given by

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{15}$$

where $\mathbf{X}_1$ is $n \times p$ and corresponds to the p parameters of the postulated model, $\boldsymbol{\beta}_1$ is $p \times 1$ and contains the parameters of the postulated model, $\mathbf{X}_2$ is $n \times (m-p)$ and corresponds to the additional parameters of the true model, $\boldsymbol{\beta}_2$ is $(m-p) \times 1$ and contains the additional parameters of the true model, and $\boldsymbol{\epsilon}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$.

If the true model is fit to the data then the total variability in the response can be written as

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + SS_{PE}, \tag{16}$$

where SS_{PE} represents the variability in the response explained by pure error. If we fit the postulated model, then the total variability in the response can be written as

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} + \mathbf{y}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} + SS_{PE}. \tag{17}$$

If the additional factors in $\mathbf{X}_2$ are functions, such as products, of the factors in $\mathbf{X}_1$, then $\mathbf{y}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$ is LOF variation (see Myers 1990, p. 119). In practice it is common for small amounts of LOF variation to be included in the estimate of error variance. Obviously, the assumption that the error variance is homogeneous is in jeopardy if LOF is included in its estimate since LOF depends on the location in x-space. When experimental error is comprised entirely of LOF, as in the deterministic response case, our assumptions concerning ϵ are questionable.

2.4. Validation

The validity of a metamodel indicates the degree to which the specified purpose of the metamodel can be accomplished (Sargent 1991a). For example, a simple linear approximation may be valid for studying some elements of system behavior but completely invalid as a means of making predictions. Validity is also specific to the experimental region used to develop the metamodel. In other words, the metamodel is expected to be valid for a specific purpose over the experimental region.

Validity can be measured using many available diagnostics. For a complete discussion of diagnostics for the random response case see Myers (1990, Chapter 4). Diagnostics for the deterministic response case are discussed in Kleijnen (1987). In this section, we simply discuss the diagnostics used in the example in Section 3.

One diagnostic that is appropriate for both deterministic and random responses is the squared coefficient of determination, R^2, which is given by

$$R^2 = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}. \tag{18}$$

Note that the numerator is the sum of squares for the metamodel and the denominator is the total sum of squares. Thus, R^2 measures the proportion of the total variability in the response explained by the metamodel. The higher R^2 the better the metamodel fits the given data. While R^2 provides a good, general measure of fit it does not measure the uniformity of fit. In other words, a metamodel with a high R^2 may have some areas of very poor fit as long as there are relatively large areas with very good fits. Also, R^2 only measures the degree to which the estimated metamodel fits the data that is used to estimate the metamodel. Thus, R^2 does not account for fit in areas where there is no data. In addition, for the random responses case, the use of R^2 by itself gravitates the model selection to an overfit model (one which tracks random error). This has the detrimental effect of reducing the prediction capability of the metamodel (see Myers 1990, p. 179-180).

In order to test the validity of a metamodel across the entire experimental region, analysts often advocate a technique known as data splitting (see Myers 1990, 169-170). Data splitting is applied by using some observations to fit a model, and a separate set of observations to measure the validity of the model. This allows the validity of the metamodel to be tested independently of the data used to fit the model. In cases where data is expensive and/or difficult to acquire, this may be impractical. Often data that is not used to fit the metamodel is supplemented with data that is used to fit the metamodel in order to measure validity. This can result in misleading measures of validity.

Another approach to validation involves the use of a diagnostic known as the PRESS statistic (see Myers 1990, pp. 170-178). To calculate the PRESS statistic, a set of PRESS residuals is calculated by mathematically *factoring out* the dependence of each observed residual on the data used to estimate the metamodel. This method eliminates the data splitting problems and is applicable for both the random and deterministic response case. However, the details of the PRESS statistic are beyond the scope of this report.

Two diagnostics that are appropriate for data splitting in the deterministic response case are the maximum absolute error (MAE) and the average absolute relative error (AARE). MAE is simply the absolute value of the largest residual. By basing

model validation on MAE, the model selection is gravitated to a uniform but not necessarily good fit.

AARE is given by

$$AARE = \frac{\sum_{i=1}^{n} |e_i/y_i|}{n}. \tag{19}$$

AARE is similar to R^2 in that it provides a good, general measure of fit, but it has the same drawback as R^2 in that use of AARE does not insure uniformity of fit.

Note that the difficulties with the individual diagnostics can be overcome by using them in combination. This is done in the example in Section 3.

2.5. Design of Experiments

The purpose of experimental design is to obtain *better* estimates and predictive models with fewer observations by carefully constructing $\mathbf{X}$. This is done by preselecting certain values for the k factors in the experiment. Assuming that we are using an unbiased estimator such as a least squares estimator, the quality of experimental designs is usually measured with the variance of prediction. The variance of prediction is the variance of the true population about the fitted model at some arbitrary location in x-space, $\mathbf{x}_0$ and is given by

$$\text{var}[\hat{y}(\mathbf{x}_0)] = \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0. \tag{20}$$

For any set of similarly *scaled* competing designs with the same number of observations, it can be shown that $\text{var}[\hat{y}(\mathbf{x}_0)]$ is minimized when $\mathbf{X}'\mathbf{X} = n\mathbf{I}$ (see Myers 1976, p. 109). In this case, $\mathbf{X}$ is said to be orthogonal. The minimization of the variance of prediction for orthogonal designs is due in part to the fact that orthogonal designs result in models with uncorrelated coefficients. This is not true of any other designs. Thus, orthogonal or near-orthogonal designs should be used whenever possible.

We consider two basic designs in this report: (1) 2^k factorials and (2) central composite designs (CCD). A 2^k factorial experiment consists of k factors each at two levels arranged in all possible factor/level combinations. To obtain an orthogonal design, the levels of the input factors are usually centered and scaled such that high

level of the factor appears as a 1 and the low level of the factor appears as a -1. The centering and scaling transformation is given by

$$x_i = 2\left(\frac{\xi_i - \bar{\xi}}{d_i}\right), \tag{21}$$

where ξ_i is the level of the ith input factor, $\bar{\xi}_i$ is the average of the low and high levels of ξ_i, x_i is the centered and scaled level of the ith input factor, and d_i is the spacing between the low and high levels of ξ_i. While centering and scaling input factors in designed experiments usually results in better metamodels, it may require the analyst to perform some extra work in order to analyze the model. For example, if an analyst needs to predict a response at some point $\boldsymbol{\xi}_0$ using a metamodel for centered and scaled input factors, then he must rescale $\boldsymbol{\xi}_0$ to $\mathbf{x}_0$ using the same centering and scaling formula used in the experiment.

To illustrate, consider a 2^2 factorial experiment replicated r times for the purpose of estimating the regression model given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon. \tag{22}$$

The corresponding design matrix is given by

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_r & \mathbf{1}_r & \mathbf{1}_r & \mathbf{1}_r \\ \mathbf{1}_r & -\mathbf{1}_r & -\mathbf{1}_r & \mathbf{1}_r \\ \mathbf{1}_r & \mathbf{1}_r & -\mathbf{1}_r & -\mathbf{1}_r \\ \mathbf{1}_r & -\mathbf{1}_r & \mathbf{1}_r & -\mathbf{1}_r \end{bmatrix} \tag{23}$$

where r is the number of replications of the experiment and $\mathbf{1}_r$ is an $r \times 1$ column vector of ones.

$$\mathbf{b} = \frac{1}{r}\mathbf{X}'\mathbf{y}. \tag{24}$$

An important advantage of factorial experiments over one-variable-at-a-time experimentation is that we can estimate the interaction effects of the factors on the response. Further details on factorial experiments can be found in Box, Hunter, and Hunter (1978, Chapter 10).

A central composite design (CCD) consists of a 2^k factorial design augmented with $2k + 1$ extra design points to allow the estimation of second order models.

To illustrate, consider a CCD replicated r times for the purpose of estimating the regression model given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon. \tag{25}$$

The corresponding design matrix is given by

$$\mathbf{X} = \begin{bmatrix} 1_r & 1_r & 1_r & 1_r & 1_r & 1_r \\ 1_r & -1_r & -1_r & 1_r & 1_r & 1_r \\ 1_r & 1_r & -1_r & -1_r & 1_r & 1_r \\ 1_r & -1_r & 1_r & -1_r & 1_r & 1_r \\ 1_r & \alpha_r & 0_r & 0_r & \alpha^2{}_r & 0_r \\ 1_r & -\alpha_r & 0_r & 0_r & \alpha^2{}_r & 0_r \\ 1_r & 0_r & \alpha_r & 0_r & 0_r & \alpha^2{}_r \\ 1_r & 0_r & -\alpha_r & 0_r & 0_r & \alpha^2{}_r \end{bmatrix}. \tag{26}$$

Note that when $\alpha = 1$, the CCD design is near-orthogonal and allows for the estimation quadratic curvature. A near orthogonal design for, higher order models can be obtained by layering CCDs. This concept is demonstrated in the example in Section 3. Further details on CCDs can be found in Myers (1976, pp. 127-134).

2.6. The Metamodeling Process

Metamodeling is a developing field so there is no set procedure for constructing a metamodel (see Sargent 1991b). Also, metamodeling is not an algorithm and cannot be generalized for all situations. Often, special information based on the analysts previous experience, theoretical knowledge, or intuition may make it possible to modify or dispense with certain steps in the process. Also note that only a handful of the available analysis techniques have been touched on here. Thus, it is often useful to add steps to the process. The following metamodel process is simply intended to add some perspective to material covered thus far:

1. Determine the purpose of the metamodel (study system behavior, predict responses, etc.).

2. Identify the response.

3. Identify important characteristics of the response (random or deterministic, discrete or continuous, etc.).

4. Identify the input factors that are to be studied with respect to the given response.

5. Identify important charactersitics of the input factors.

6. Specify the experimental region.

7. Select validity measures.

8. Specify the required validity of the model in terms of selected validity measures.

9. Postulate a metamodel based on characteristics of the response and input factors, the dimensions of the experimental region, and the required validity of the metamodel.

10. Select an appropriate experimental design based on the postulated model.

11. Obtain data.

12. Fit the metamodel.

13. Assess the validity of the metamodel.

14. If the model is of desired validity then stop, else postulate a more complex model and repeat steps 9-13.

3. TERSM Example

The Tactical Electronic Reconnaissance Simulation Model (TERSM) was built in 1969 by the Rand Corporation, for the purpose of making comparative performance evaluations of a variety of airborne direction-finding systems. Simulating a reconnaissance mission through a pulsed radar environment, its primary output is a lower bound on the emitter location accuracy attainable by accumulation and processing of bearing measurements. These measures of emitter location accuracy are known as Circular Error Probabilities, of CEPs. In essence, it is the imaginary circle or ellipse around an emitter, of such size that the probability of that emitter's actually falling

in the circle is 50%. Obviously, the smaller the CEP, the more accurate the associated location estimate.

The model was designed to simulate a reconnaissance mission in sufficient detail to assess the influence of variations of system design parameters and input factors on overall system performance. Thus, by altering input factors and parameters, analysts can use TERSM to compare and contrast proposed airborne direction-finding systems and tactics.

To demonstrate the usefulness of metamodeling in this type of assessment and to illustrate the metamodeling process, we elected to perform a system behavior study. In particular, we wanted a reasonably accurate approximation of the relationship between the number of emitters located on the *test* mission within five nautical miles or less CEPs, and four input factors: (1) altitude in feet, (2) velocity in knots, (3) azimuth angle in degrees, and (4) channel capacity in number of channels. The test mission consists of a set of hostile and friendly radar emitters of various types arranged in set locations on a hypothetical battlefield. Historically, the test mission has been the standard scenario used to compare competing systems. We selected the following experimental region: (1) altitude from 5000 to 40000 feet, (2) velocity from 186 to 1150 knots, (3) azimuth angle from 60 to 150 degrees, and (4) and channel capacity from 4 to 30 channels. The selected factors and experimental region were chosen based on previous studies with TERSM. Note that the response is continuous and deterministic and that all of the input factors are continuous and deterministic except for channel capacity, which is discrete and deterministic. To accomplish our stated purpose, we specified the following goals for the validity of the model: (1) R^2 of at least 95%, (2) MAE less than 100, and (3) AARE less than 5%. All 49 observations given in Appendix I were used to validate the model.

To obtain a satisfactory metamodel, we made seven model fitting iterations. The observations used to estimate the metamodels were drawn from the validation test set. Although the stated purpose of the metamodel does not include simplicity, all terms with p-values greater 50% were eliminated sequentially from each model in order to obtain models of manageable size. Usually p-values of around 10-30% are used to eliminate insignificant factors, but since we were dealing with an obvious violation

of assumptions (i.e. a deterministic response), 50% was used for conservatism. All models were estimated for centered and scaled factors levels. The correspondence between the actual input factors used in the experiment and the centered and scaled input factor levels used to fit the metamodels is given in Table 2.

Table 2

Correspondence Between Actual Input Factor Levels and Centered and Scaled Input Factor Levels

Input Factor	Variable	-1	-0.5	0	0.5	1
Altitude	x_1	5000	13750	22500	31250	40000
Velocity	x_2	186	427	668	909	1150
Azimuth	x_3	60	82	105	128	150
Channel Cap.	x_4	4	10	17	24	30

The following is the sequence of postulated and fitted metamodels:

Model (1)

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{23} x_2 x_3 \\ &\quad + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \beta_{134} x_1 x_3 x_4 + \beta_{234} x_2 x_3 x_4 \\ &\quad + \beta_{1234} x_1 x_2 x_3 x_4 + \varepsilon \end{aligned}$$

$$\begin{aligned} \hat{y} &= 224.118 + 85.750 x_2 + 57.750 x_3 + 89.000 x_4 + 27.750 x_1 x_4 + 21.000 x_2 x_3 \\ &\quad + 47.250 x_2 x_4, \end{aligned}$$

Model (2)

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 \\ &\quad + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \beta_{134} x_1 x_3 x_4 \\ &\quad + \beta_{234} x_2 x_3 x_4 + \beta_{1234} x_1 x_2 x_3 x_4 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 + \varepsilon, \end{aligned}$$

$$\begin{aligned} \hat{y} &= 532.633 + 11.944 x_1 + 91.944 x_2 + 61.778 x_3 + 102.333 x_4 + 16.000 x_1 x_3 \\ &\quad + 27.75 x_1 x_4 + 21.000 x_2 x_3 + 47.250 x_2 x_4 + 19.750 x_3 x_4 + 12.250 x_1 x_2 x_3 \\ &\quad + 12.500 x_1 x_3 x_4 - 30.608 x_1^2 - 212.608 x_2^2 - 86.108 x_3^2, \end{aligned}$$

Model (3)

$$\begin{aligned}\ln y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 \\ &\quad + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \beta_{134} x_1 x_3 x_4 \\ &\quad + \beta_{234} x_2 x_3 x_4 + \beta_{1234} x_1 x_2 x_3 x_4 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 + \varepsilon,\end{aligned}$$

$$\begin{aligned}\widehat{\ln y} &= 6.346 - 0.047 x_1 + 0.417 x_2 + 0.306 x_3 + 0.467 x_4 - 0.025 x_1 x_2 \\ &\quad + 0.075 x_1 x_3 + 0.170 x_1 x_4 + 0.098 x_2 x_4 - 0.028 x_3 x_4 \\ &\quad + 0.066 x_1 x_2 x_3 - 0.078 x_2 x_3 x_4 - 0.049 x_1 x_2 x_3 x_4 - 0.133 x_1^2 \\ &\quad - 0.679 x_2^2 - 0.125 x_3^2 - 0.363 x_4^2,\end{aligned}$$

Model (4)

$$\begin{aligned}\sqrt{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{23} x_2 x_3 \\ &\quad + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \beta_{134} x_1 x_3 x_4 + \beta_{234} x_2 x_3 x_4 \\ &\quad + \beta_{1234} x_1 x_2 x_3 x_4 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 + \varepsilon,\end{aligned}$$

$$\begin{aligned}\widehat{\sqrt{y}} &= 23.319 + 2.994 x_2 + 2.042 x_3 + 3.288 x_4 + 0.488 x_1 x_3 + 1.006 x_1 x_4 + 0.407 x_2 x_3 \\ &\quad + 1.155 x_2 x_4 + 0.257 x_3 x_4 - 0.400 x_1 x_2 x_3 - 0.288 x_2 x_3 x_4 - 0.841 x_1^2 \\ &\quad - 5.757 x_2^2 - 0.701 x_3^2 - 2.683 x_4^2,\end{aligned}$$

Model (5)

$$\begin{aligned}y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 \\ &\quad + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \beta_{134} x_1 x_3 x_4 \\ &\quad + \beta_{234} x_2 x_3 x_4 + \beta_{1234} x_1 x_2 x_3 x_4 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 \\ &\quad + \beta_{111} x_1^3 + \beta_{222} x_2^3 + \beta_{333} x_3^3 + \beta_{444} x_4^3 + \beta_{1111} x_1^4 + \beta_{2222} x_2^4 + \beta_{3333} x_3^4 \\ &\quad + \beta_{4444} x_4^4 + \varepsilon,\end{aligned}$$

$$\begin{aligned}\hat{y} &= 549.756 - 32.722 x_1 - 111.537 x_2 - 62.778 x_3 + 180.556 x_4 + 16.441 x_1 x_3 \\ &\quad + 26.647 x_1 x_4 + 20.882 x_2 x_3 + 46.971 x_2 x_4 + 18.676 x_3 x_4 + 12.508 x_1 x_2 x_3 \\ &\quad + 9.785 x_1 x_2 x_4 + 12.892 x_1 x_3 x_4 + 61.972 x_1^2 - 22.962 x_3^2 - 86.491 x_4^2 \\ &\quad + 44.667 x_1^3 + 203.481 x_2^3 - 78.222 x_4^3 - 90.454 x_1^4 - 210.918 x_2^4,\end{aligned}$$

Model (6)

$$\begin{aligned}\ln y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 \\ &\quad + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \beta_{134} x_1 x_3 x_4 \\ &\quad + \beta_{234} x_2 x_3 x_4 + \beta_{1234} x_1 x_2 x_3 x_4 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 \\ &\quad + \beta_{111} x_1^3 + \beta_{222} x_2^3 + \beta_{333} x_3^3 + \beta_{444} x_4^3 + \beta_{1111} x_1^4 + \beta_{2222} x_2^4 + \beta_{3333} x_3^4 \\ &\quad + \beta_{4444} x_4^4 + \varepsilon.\end{aligned}$$

$$\begin{aligned}\widehat{\ln y} &= 6.350 - 0.047 x_1 - 0.308 x_2 + 0.075 x_3 + 0.267 x_4 - 0.022 x_1 x_2 \\ &\quad + 0.073 x_1 x_3 + 0.162 x_1 x_4 + 0.021 x_2 x_3 + 0.099 x_2 x_4 - 0.029 x_3 x_4 \\ &\quad + 0.066 x_1 x_2 x_3 + 0.014 x_1 x_2 x_4 - 0.078 x_2 x_3 x_4 - 0.049 x_1 x_2 x_3 x_4 + 0.260 x_1^2 \\ &\quad - 0.125 x_3^2 - 0.359 x_4^2 + 0.725 x_2^3 + 0.230 x_3^3 + 0.199 x_4^3 - 0.398 x_1^4 - 0.682 x_2^4.\end{aligned}$$

Model (7)

$$\begin{aligned}\sqrt{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 \\ &\quad + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \beta_{134} x_1 x_3 x_4 \\ &\quad + \beta_{234} x_2 x_3 x_4 + \beta_{1234} x_1 x_2 x_3 x_4 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 \\ &\quad + \beta_{111} x_1^3 + \beta_{222} x_2^3 + \beta_{333} x_3^3 + \beta_{444} x_4^3 + \beta_{1111} x_1^4 + \beta_{2222} x_2^4 + \beta_{3333} x_3^4 \\ &\quad + \beta_{4444} x_4^4 + \varepsilon,\end{aligned}$$

$$\begin{aligned}\widehat{\sqrt{y}} &= 23.567 - 0.669 x_1 - 2.842 x_2 + 1.298 x_3 + 3.344 x_4 - 0.491 x_1 x_3 \\ &\quad + 0.963 x_1 x_4 + 0.414 x_2 x_3 + 1.155 x_2 x_4 + 0.231 x_3 x_4 + 0.404 x_1 x_2 x_3 \\ &\quad + 0.198 x_1 x_2 x_4 + 0.201 x_1 x_3 x_4 - 0.285 x_2 x_3 x_4 + 2.037 x_1^2 - 0.788 x_3^2 \\ &\quad - 2.743 x_4^2 + 0.714 x_1^3 + 5.836 x_2^3 + 0.744 x_3^3 - 2.947 x_1^4 - 5.823 x_2^4.\end{aligned}$$

Information on the designs used to estimate these models and the resulting validity measures are given in Table 3.

Table 3					
Summary and Results of the Model Fitting Procedure					
Model	R^2	MAE	AARE	Design	Obs.
1	71.7%	421.6	44.3%	2^k factorial w/ 1 center run	1-17
2	95.5%	200.6	14.2%	CCD w/ $\alpha = 1$	1-25
3	99.2%	256.3	12.3%	CCD w/ $\alpha = 1$	1-25
4	98.2%	225.4	11.4%	CCD w/ $\alpha = 1$	1-25
5	97.4%	120.9	8.3%	2 layer CCD w/ $\alpha_1 = 1$ and $\alpha_2 = 0.5$	1-49
6	99.4%	94.3	4.0%	2 layer CCD w/ $\alpha_1 = 1$ and $\alpha_2 = 0.5$	1-49
7	98.9%	73.51	4.7%	2 layer CCD w/ $\alpha_1 = 1$ and $\alpha_2 = 0.5$	1-49

Note the improvement in the validity as the complexity of the metamodel increases. Both model (6) and (7) satisfied our validity requirement and we stopped here. However, note that the validity test set is the same as the data set used to fit these models. Thus, as discussed in Section 2.4, these models are probably somewhat less valid than the validity measures indicate. If the process was continued, the next step would be to expand the validity test set and reevaluate the validity of the selected models.

Figures 3-6 contain surface and contour plots of models (1), (2), (5), and (7) respectively with CS azimuth angle and CS channel capacity held at zero. The prefix CS indicates the centered and scaled input factor. These graphs allow the estimated relationship between the response, CS altitude, and CS velocity to be observed independently of CS azimuth angle and CS channel capacity. Notice how the increasing complexity of the model manifests itself in increasingly refined shapes and how the *optimum* value of the response changes as the metamodels become more accurate.

Figures 7-11 contain surface and contour plots of model (7) for all possible pairs of CS altitude, CS velocity, CS azimuth angle, and CS channel capacity held at zero. These graphs illustrate the variety of relationships between the response and the input factors that are encapsulated in model (7).

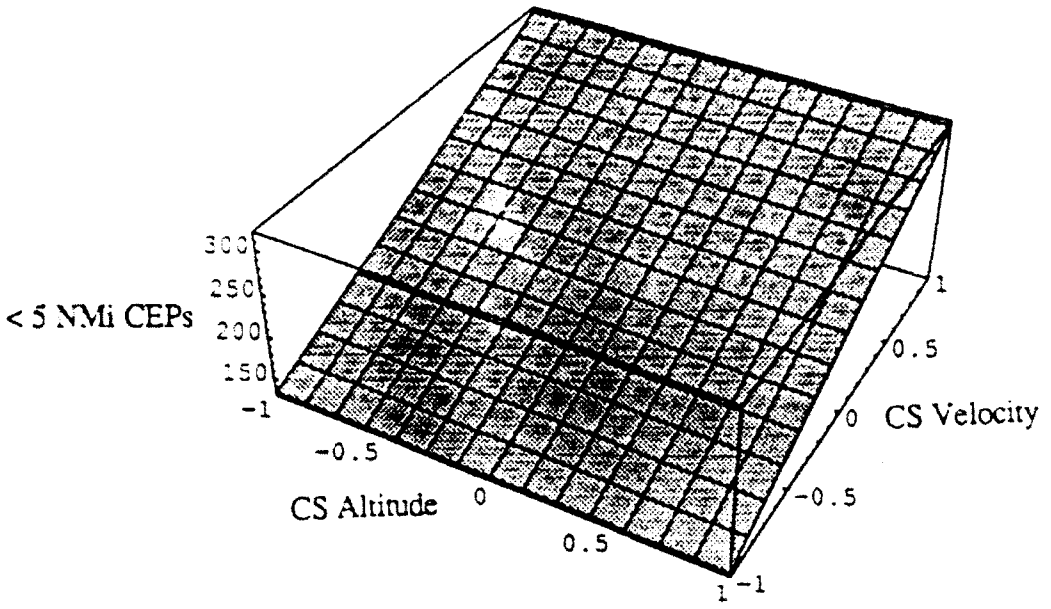


Figure 3a: Surface plot of model (1) with CS azimuth and CS channel capacity equal to zero.

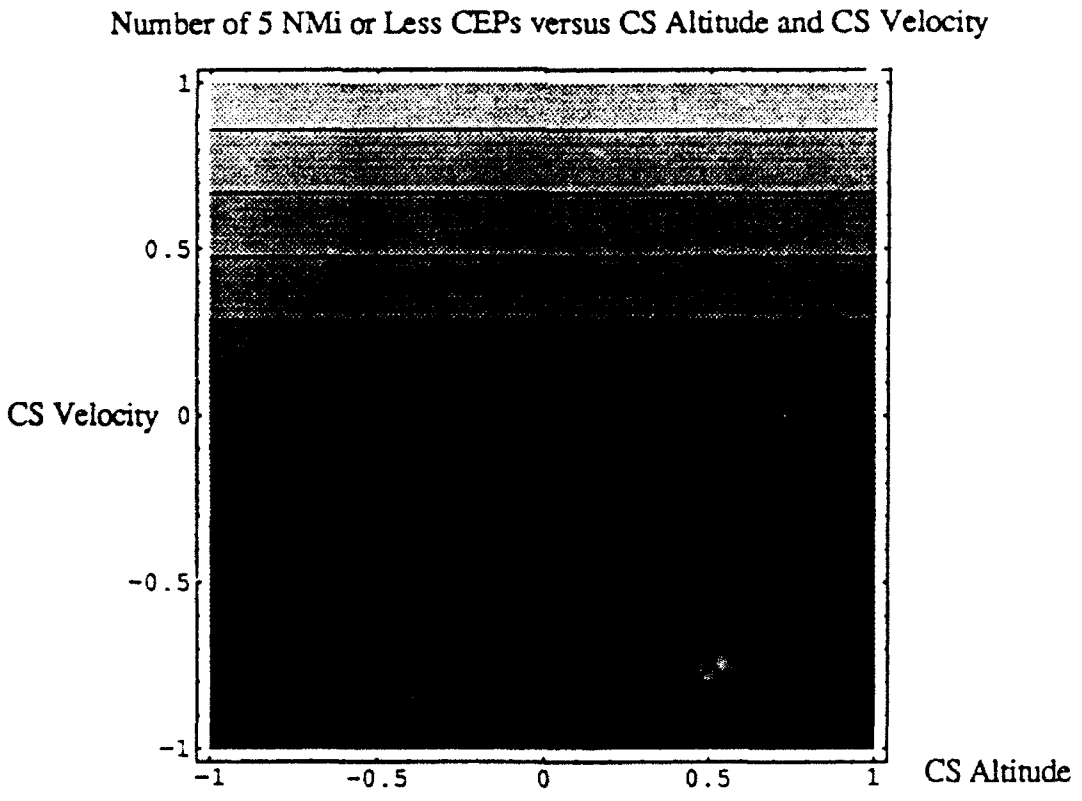


Figure 3b: Contour plot of model (1) with CS azimuth and CS channel capacity equal to zero.

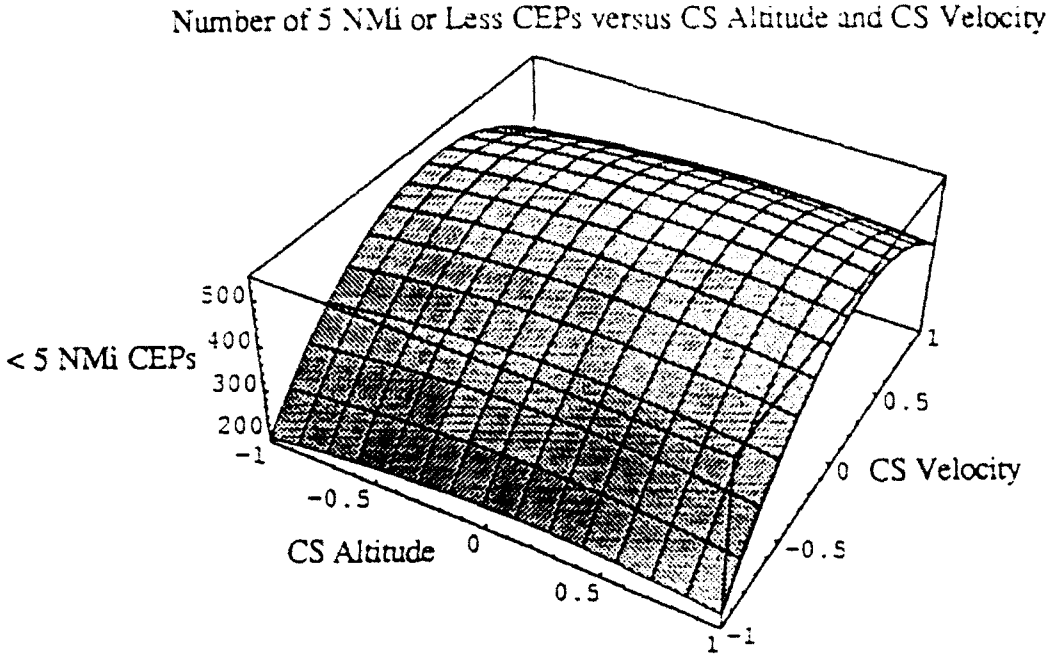


Figure 4a: Surface plot of model (2) with CS azimuth and CS channel capacity equal to zero.

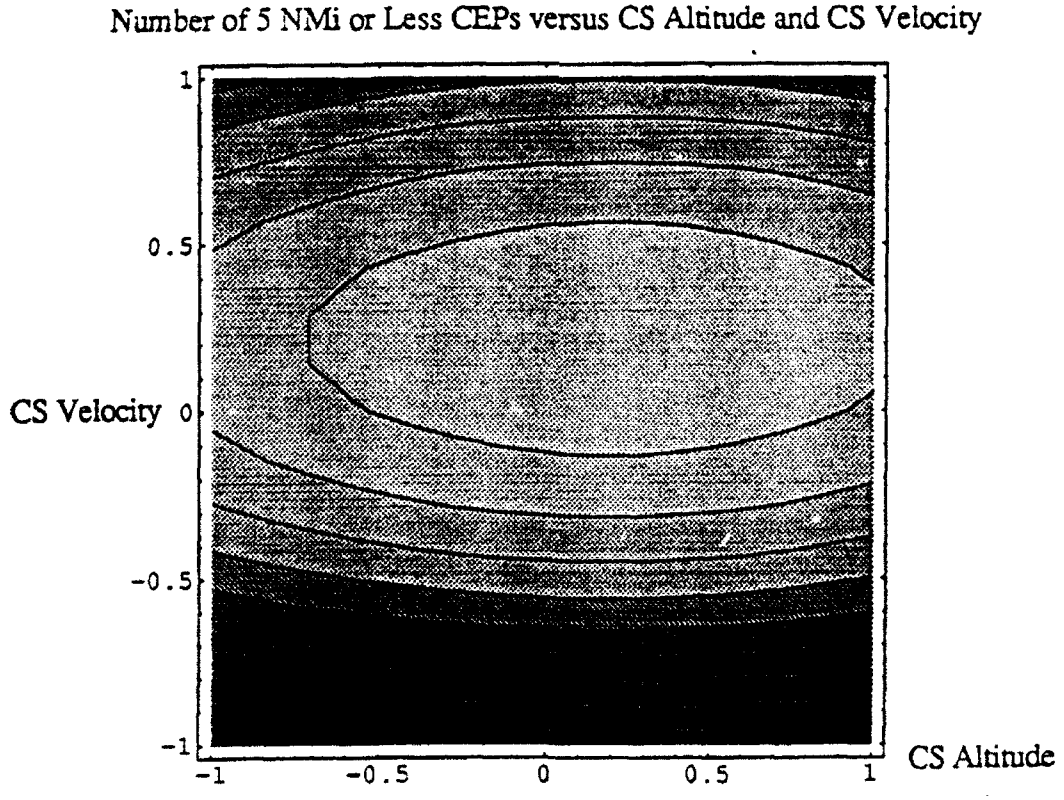


Figure 4b: Contour plot of model (2) with CS azimuth and CS channel capacity equal to zero.

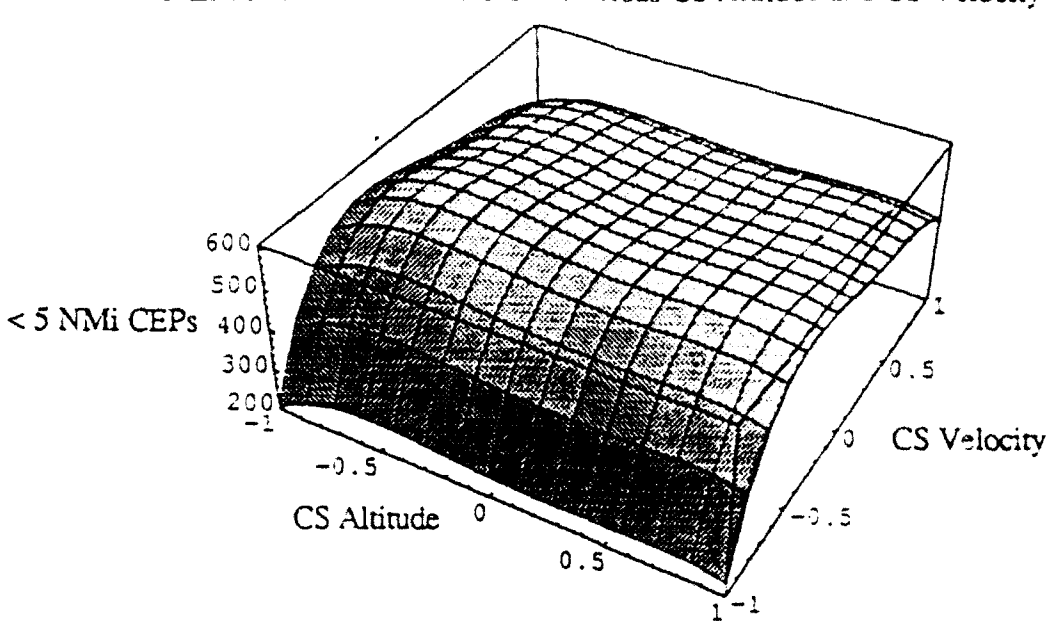


Figure 5a: Surface plot of model (5) with CS azimuth and CS channel capacity equal to zero.

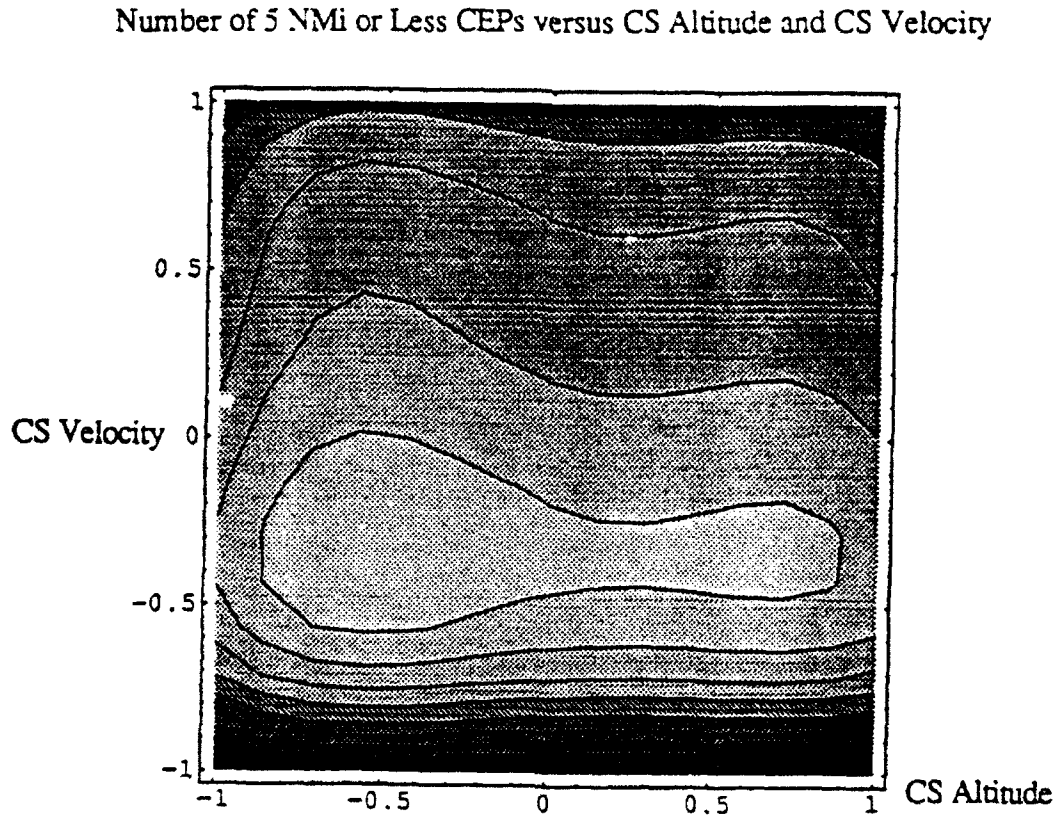


Figure 5b: Contour plot of model (5) with CS azimuth and CS channel capacity equal to zero.

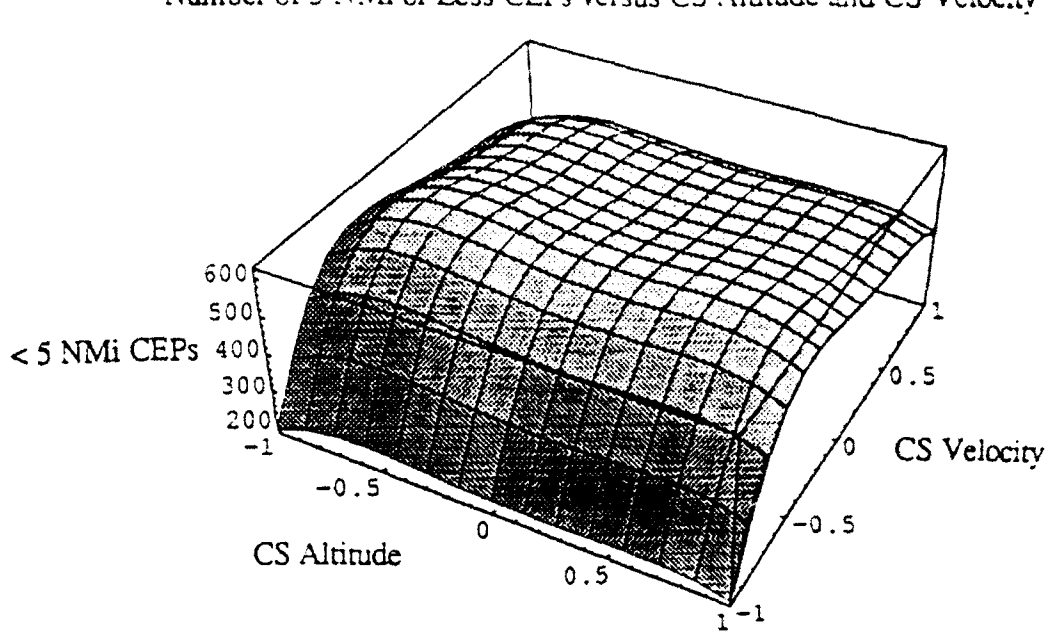


Figure 6a: Surface plot of model (7) with CS velocity and CS channel capacity equal to zero.

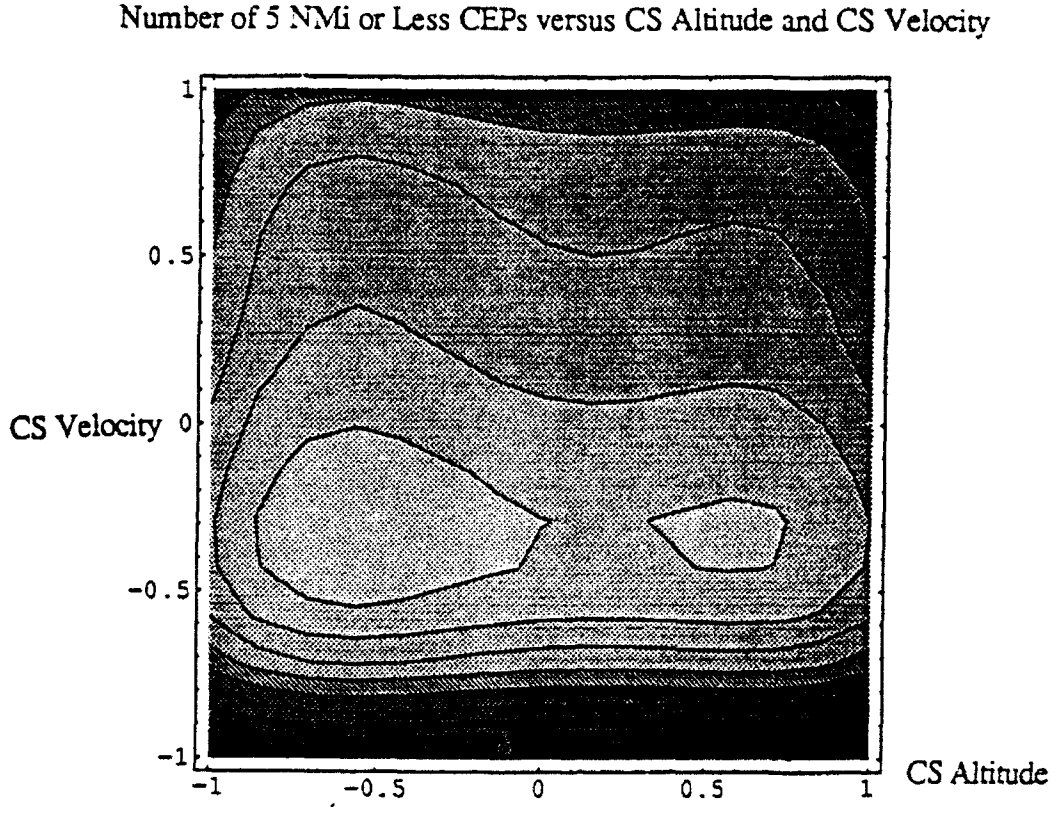


Figure 6b: Contour plot of model (7) with CS velocity and CS channel capacity equal to zero.

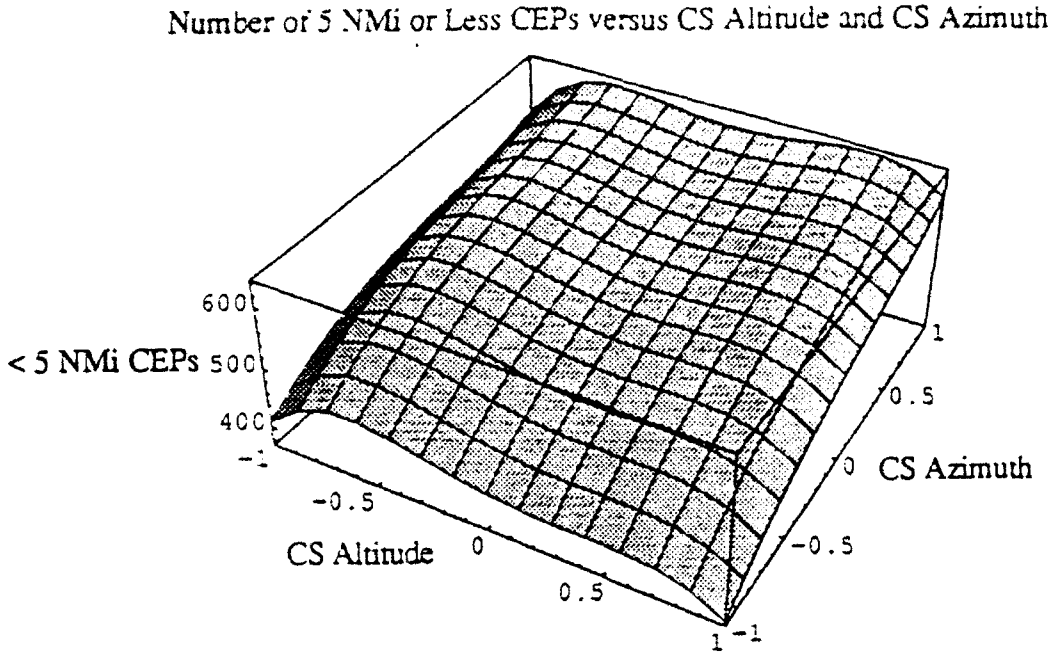


Figure 7a: Surface plot of model (7) with CS velocity and CS channel capacity equal to zero.

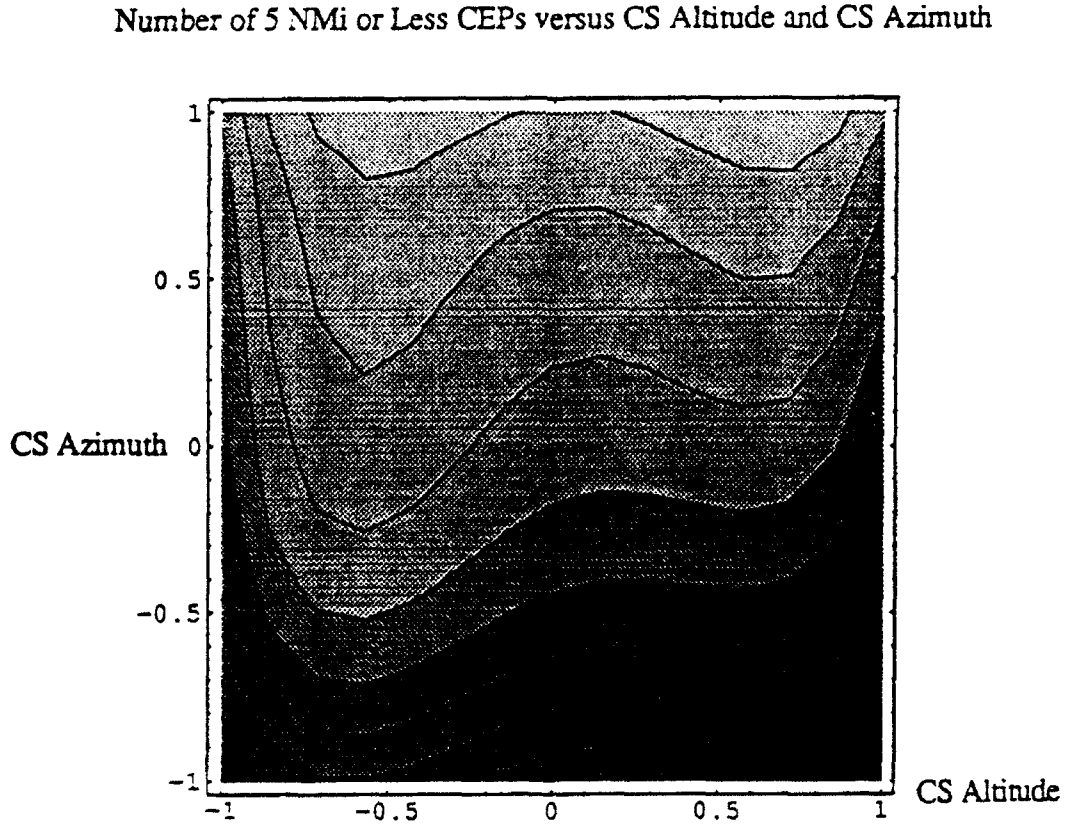


Figure 7b: Contour plot of model (7) with CS velocity and CS channel capacity equal to zero.

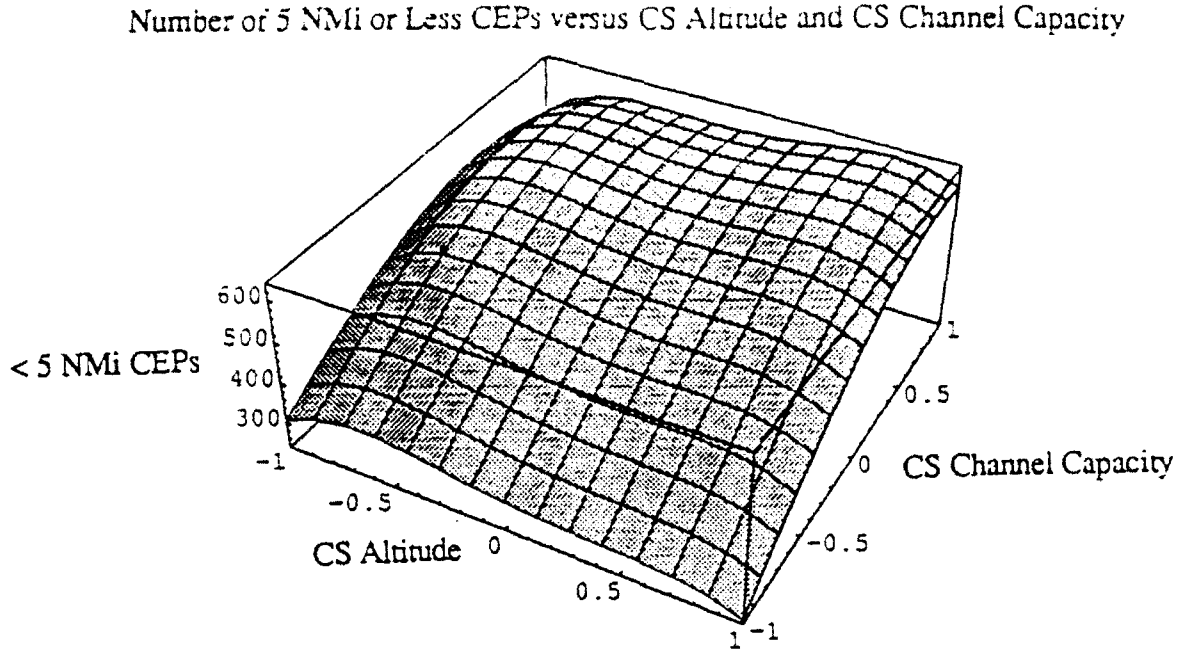


Figure 8a: Surface plot of model (7) with CS velocity and CS azimuth equal to zero.

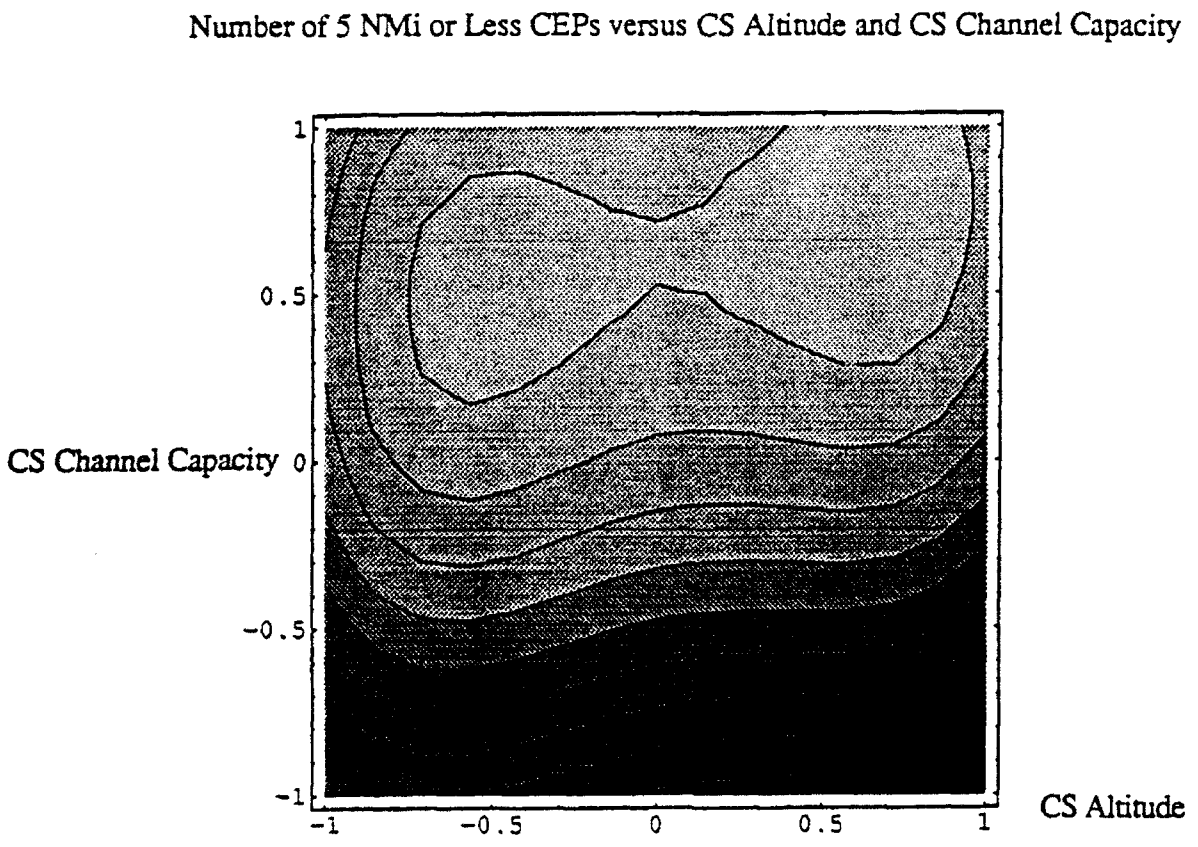


Figure 8b: Contour plot of model (7) with CS velocity and CS azimuth equal to zero.

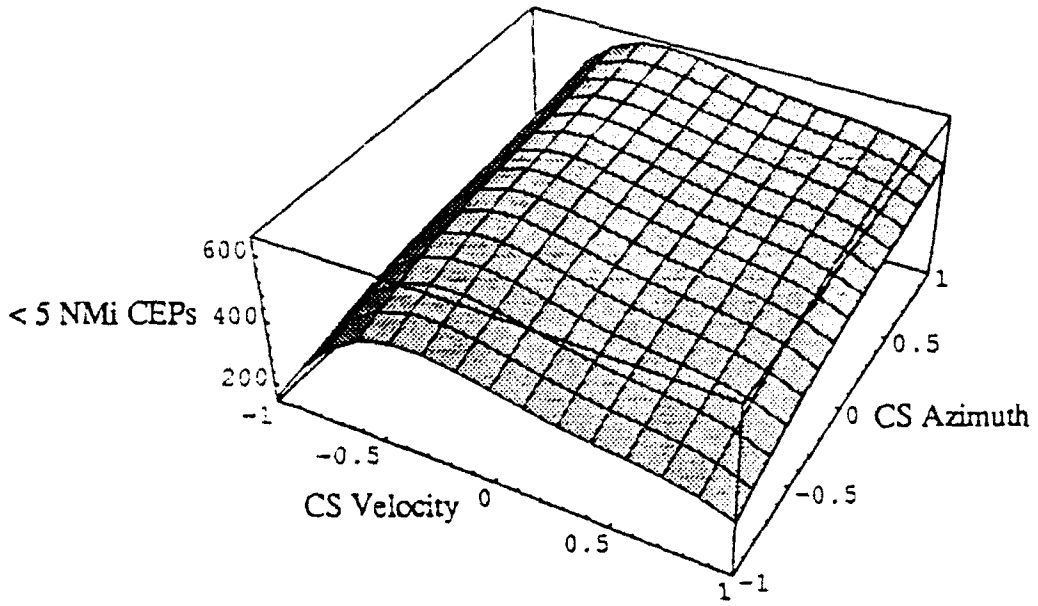


Figure 9a: Surface plot of model (7) with CS altitude and CS channel capacity equal to zero.

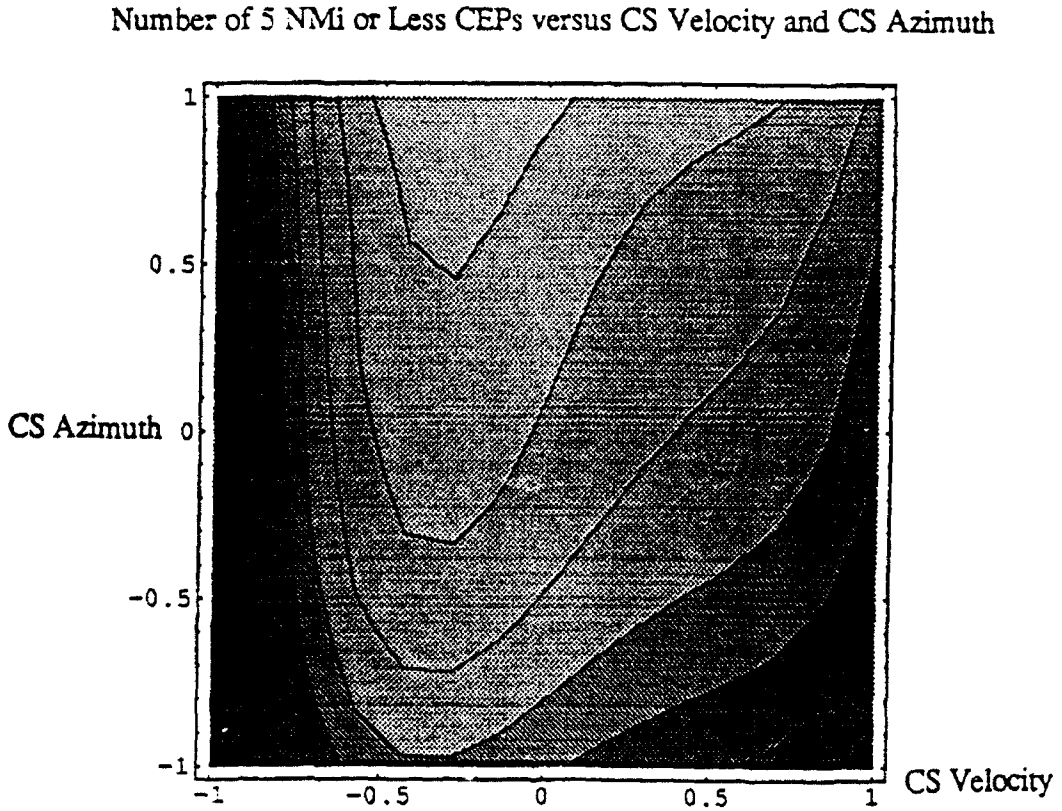


Figure 9b: Contour plot of model (7) with CS altitude and CS channel capacity equal to zero.

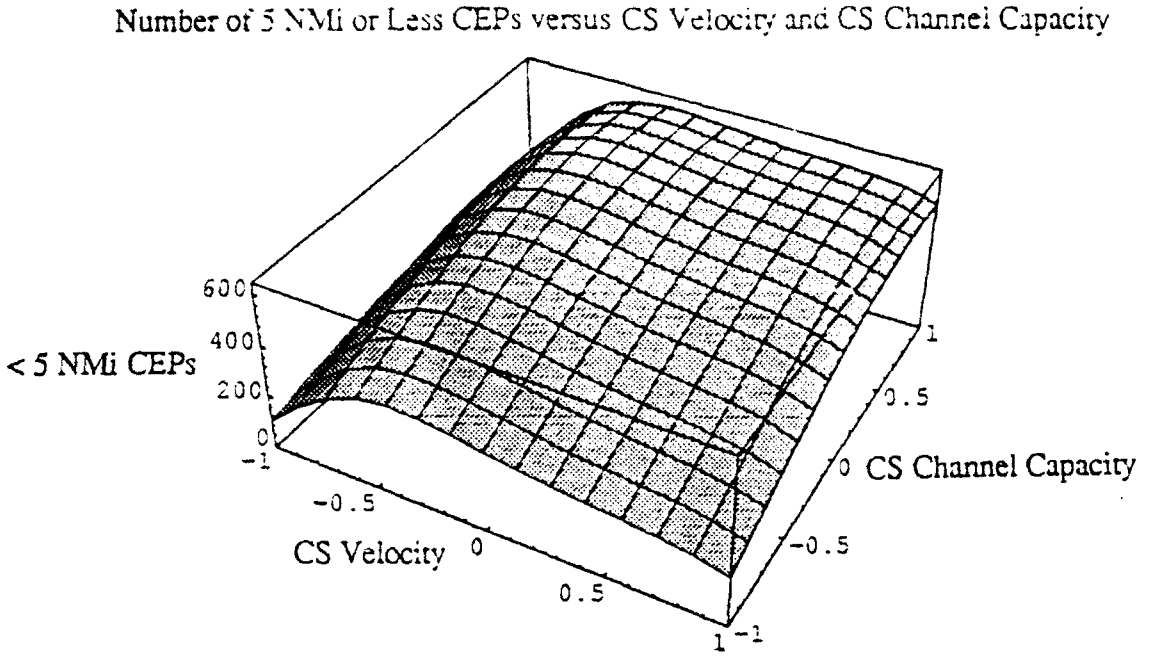


Figure 10a: Surface plot of model (7) with CS altitude and CS azimuth equal to zero.

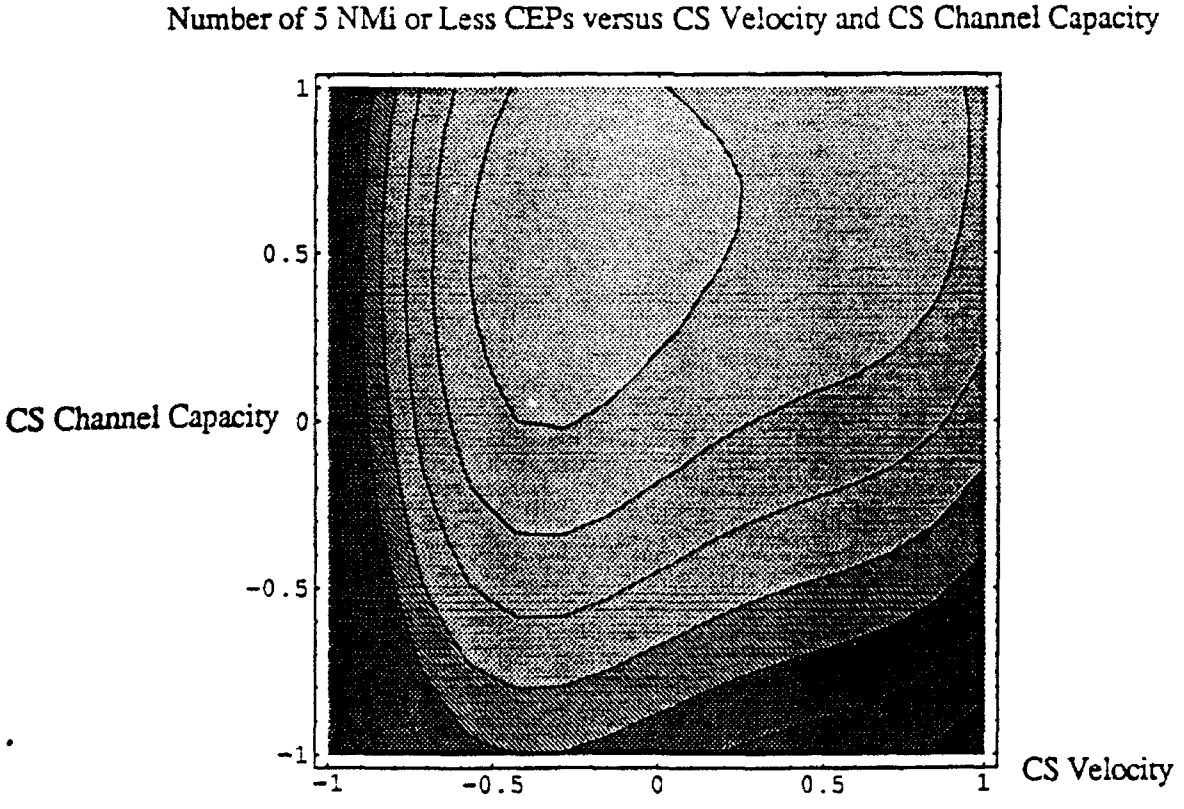


Figure 10b: Contour plot of model (7) with CS altitude and CS azimuth equal to zero.

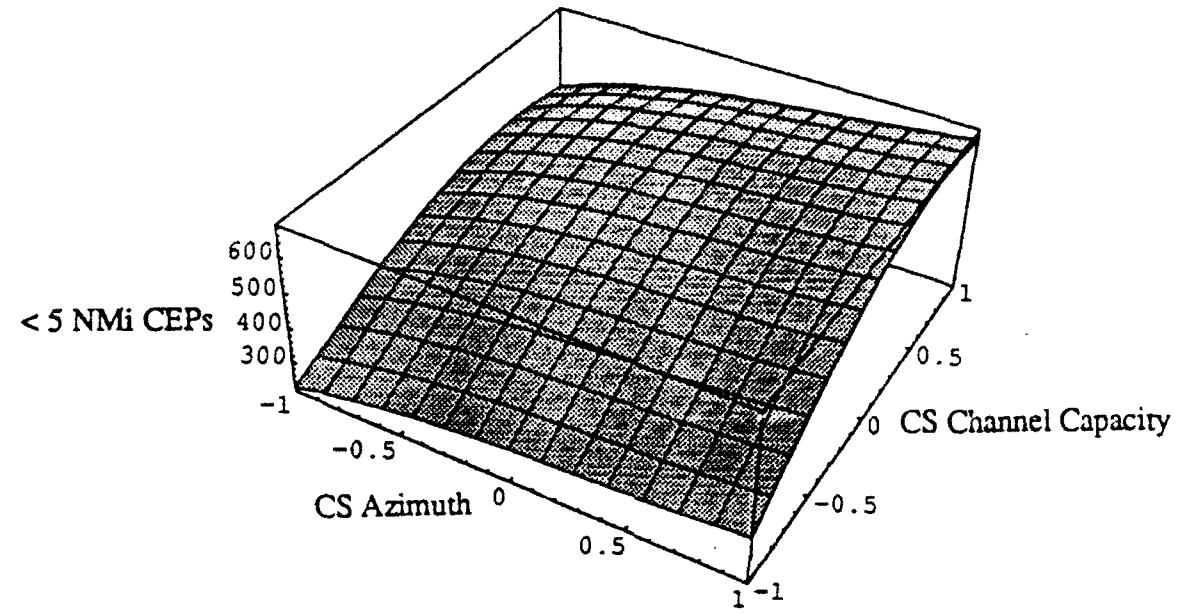


Figure 11a: Surface plot of model (7) with CS altitude and CS velocity equal to zero.

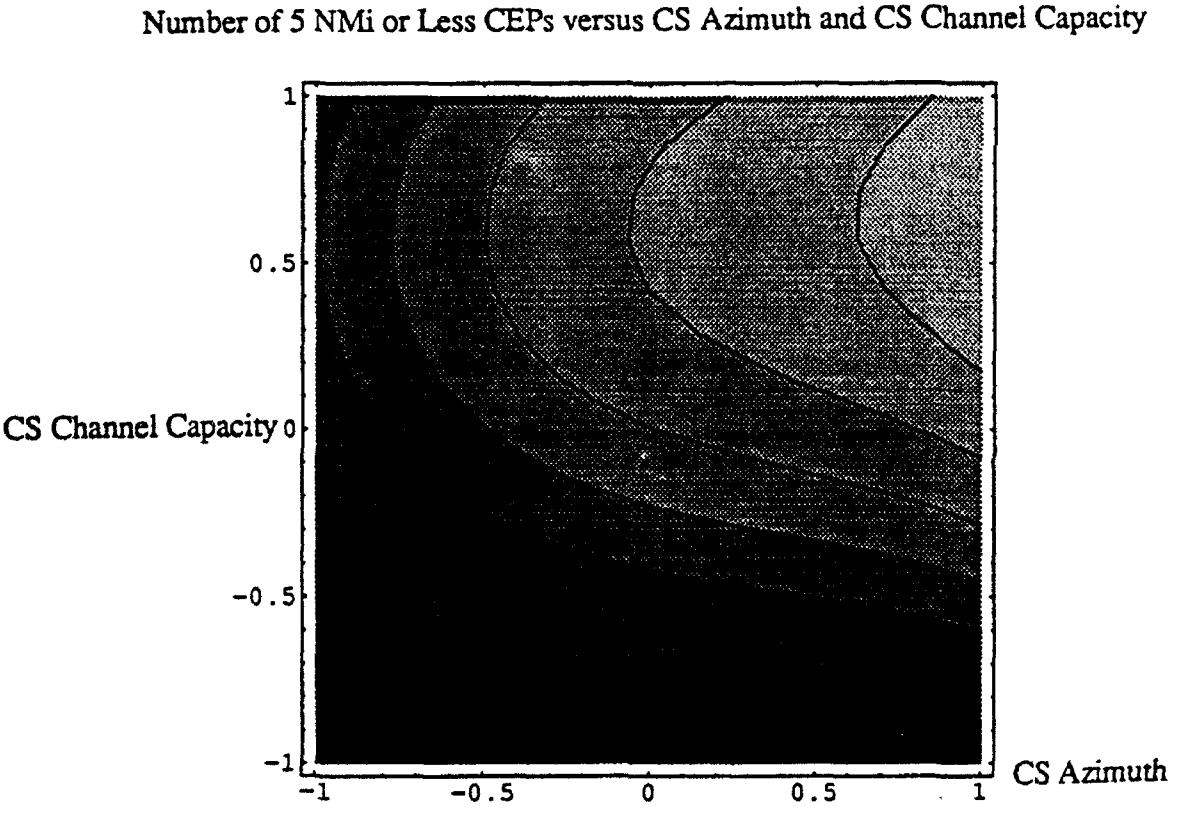


Figure 11b: Contour plot of model (7) with CS altitude and CS velocity equal to zero.

4. Conclusions

In this report, we have introduced and successfully demonstrated a possible application of metamodeling to analysis of tactical simulations. While metamodeling is not appropriate for all simulation analysis problems, it does have a wide range of possible applications. In particular, metamodeling may provide a means of adequately aggregating simulation model behavior in hierarchical modeling schemes (see Sargent (1986) and Sisti (1989 and 1992)).

Appendix I

Table 4					
Data for the TERSM Example					
Run	Altitude (feet)	Velocity (knots)	Azimuth (degrees)	Channel Capacity	< 5 NMi CEPs
1	40000	1150	150	30	615
2	40000	1150	150	4	193
3	40000	1150	60	30	327
4	40000	1150	60	4	53
5	40000	186	150	30	247
6	40000	186	150	4	73
7	40000	186	60	30	111
8	40000	186	60	4	47
9	5000	1150	150	30	436
10	5000	1150	150	4	226
11	5000	1150	60	30	322
12	5000	1150	60	4	138
13	5000	186	150	30	180
14	5000	186	150	4	116
15	5000	186	60	30	98
16	5000	186	60	4	66
17	22500	668	105	17	562
18	5000	668	105	17	439
19	40000	668	105	17	570
20	22500	186	105	17	181
21	22500	1150	105	17	464
22	22500	668	60	17	419
23	22500	668	150	17	607
24	22500	668	105	4	240
25	22500	668	105	30	658
26	31250	909	128	24	621
27	31250	909	128	10	424
28	31250	909	82	24	512
29	31250	909	82	10	347
30	31250	427	128	24	634
31	31250	427	128	10	489
32	31250	427	82	24	570
33	31250	427	82	10	434

Table 4. Continued					
Data for the TERSM Example					
Run	Altitude (feet)	Velocity (knots)	Azimuth (degrees)	Channel Capacity	< 5 NMi CEPs
34	13750	909	128	24	602
35	13750	909	128	10	441
36	13750	909	82	24	560
37	13750	909	82	10	373
38	13750	427	128	24	651
39	13750	427	128	10	526
40	13750	427	82	24	605
41	13750	427	82	10	471
42	13750	668	105	17	580
43	31250	668	105	17	584
44	22500	427	105	17	575
45	22500	909	105	17	529
46	22500	668	82	17	512
47	22500	668	128	17	597
48	22500	668	105	10	441
49	22500	668	105	24	640

References

1. Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. John Wiley, New York.

2. Kleijnen, J. P. C. 1987. *Statistical Tools for Simulation Practitioners*. Marcel Dekker, New York.

3. Myers, R. H. 1976. *Response Surface Methodology*. Edwards Brothers, Ann Arbor, Michigan.

4. Myers, R. H. 1990. *Classical and Modern Regression with Applications*. PWS-KENT, Boston.

5. Myers, R. H. and Milton, J. S. 1991. *A First Course in the Theory of Linear Statistical Models*. PWS-KENT, Boston.

6. Sisti, A. F., 1989. A Model Integration Approach to Electronic Combat Effectiveness Evaluation, RADC-TR-89-183.

7. Sisti, A. F., 1992. Large-Scale Battlefield Simulation Using Multi-Level Model Integration Methodology, RL-TR-92-69.

8. Sargent, R. G. 1986. Joining Existing Simulation Programs. *Proceedings of the 1986 Winter Simulation Conference*, J. Wilson, J Henrikson, and S. Roberts, eds., 512-516.

9. Sargent, R. G. 1991a. Simulation Model Verification and Validation. *Proceedings of the 1991 Winter Simulation Conference*, B. L. Nelson, W. D. Kelton, and G. M. Clark, eds., 34-47.

10. Sargent, R. G. 1991b. Research Issues in Metamodeling. *Proceedings of the 1991 Winter Simulation Conference*, B. L. Nelson, W. D. Kelton, and G. M. Clark, eds., 888-893.

FRACTAL IMAGE COMPRESSION TECHNIQUES

Guttalu R. Viswanath
Professor of Mathematics
Department of Mathematics and Computer Science
South Carolina State University
Orangeburg, S.C. 29117

Patrick K. McCabe
Joseph D. Stooks
RL/IRDD
Griffiss Air Force Base
Rome, N.Y. 13441 - 57000

Final Report for :
Summer Research Program
Rome Laboratory

Sponsored by :
Air Force Office of Scientific Research
Bolling Air Force Base, Washington D.C.

September 1992

FRACTAL IMAGE COMPRESSION TECHNIQUES

Guttalu R. Viswanath
Professor of Mathematics
Department of Mathematics and Computer Science
South Carolina State University

Patrick K. McCabe
Joseph D. Stooks
RL/IRDD
Griffiss Air Force Base

Abstract

Display and analysis of terrain and imagery data has become common in computerized information systems. However, the massive storage and transmission requirements required for terrain and imagery data remain major problem areas. Current Joint Photographic Group(JPEG) algorithms achieve compression ratios no greater than 30 to 1. The development of robust compression algorithms based on fractal mathematics may mitigate the storage and transmission problems and ultimately lead to more cost effective utilization of terrain and imagery data at the unit level. In addition, the compression of data can lead to increased volume of data transmitted in decreased amounts of time. Our investigation identified promising approaches, which are candidates for further development.

FRACTAL IMAGE COMPRESSION TECHNIQUES

Guttalu R. Viswanath
Patrick K. McCabe
Joseph D. Stooks

Introduction

Tactical Air Force users have data communication assets of limited bandwidth. Tactical bandwidths can range from 1.9 Kbps to 56 Kbps. A Battle Target Graphic (BTG) consists of 48 to 50 megabytes of terrain and image data, and 2 to 3 megabytes of annotation. This 50 to 53 megabyte package if shipped electronically over a communications link of 1.9 Kbps would require 50 to 60 hours to complete. Given a communications link of 56 Kbps, approximately 2 hours. A 150 to 1 compression ratio however, would allow the same data to be transmitted in 20 minutes over a 1.9 Kbps link and in less than a minute at 56 Kbps. The need for image data compression becomes apparent.

A wide range of image data compression techniques have been developed over the years [6], [8], [9], [10], and new methods continue to emerge. In the area of image processing, fractals have created a great deal of interest because of the prospect of solving the inverse problem, that of automatically recovering simple rules that describe a computer imagery [1(a, b)], [2], [4], [11].

A Fractal is a geometrical shape that has the following properties: (a) the object is self similar (b) the object has fractional dimension. Fractal curves are associated with many physical and natural phenomenon [5], [7], and have unusual characteristic that they can be defined totally by relatively simple mathematical equations. To put it in a simple way [1(c)], a fractal is a mathematical description(formula) of a picture. Each fractal formula can generate a fractal picture through a repetitive formula known as an Iterated Function System(IFS) [1(a,b)], [3]. Because the fractal formula

describing a picture is much smaller file than the original image file it is easier and more economical to store, handle and transmit images over network, communication lines etc.,. The Fractal Transform Technology [1(c)] by Barnsley automates the process, generating fractal formulas for complex images which otherwise would have taken hours of computing time. This process creates a compressed Fractal Image File(FIF) of a given imagery which can be decompressed for display or transmission.

An article on fractal compression in an issue of the Joint National Intelligence Development Staff(JNIDS) newsletter describes potential applications of this new technology: "The possibility for this technology seems limitless. It will open the door for a variety of a plications, particularly, in the area of multimedia, that are impossible because of today's limited capability to store imagery". Further, the article points out that: "Fractals represent a breakthrough not only for image compression, but also for image understanding. Since the process describes a real image rather than millions of individual points in the image, it opens up all sorts of possibilities for computer vision, image understanding, soft-copy photographic keys, and aids for imagery interpretation".

Methodology

The methodology utilized was straight forward. Investigate current approaches, bound the problem, and develop algorithms to evaluate. Each aspect of the methodology is discussed below.

Current Approaches

Let us take a look at Barnsley's compression algorithm [1(a)]. An image is broken into segments using any of the image processing techniques. We then look up these segments in a library of fractals. The library doesn't contain literal fractals; that would require astronomical amount of storage. Instead, the library contains relatively compact sets of numbers, called iterated function systems(IFS) codes that will reproduce the corresponding fractals. The library's cataloging system is such that images that look alike are close together; nearby codes correspond to nearby fractals which makes it

feasible to set up automated procedure for searching the library to find fractals that approximate a given target image. The Collage theorem guarantees that we can always find a suitable IFS code and gives us a method for doing so. The image components are looked up in the IFS library using Collage theorem and their IFS codes are recorded. Once we looked up all the segments of the image in the library and found the IFS codes, we can throw away the original digitized image and keep only the codes achieving our compression ratio of 10000 to 1 or even better. When the image is to be reconstructed, the IFS codes are input to the random iteration algorithm. The accuracy of the reconstructed image depends on the tolerance setting used during the Collage mapping stage.

Problem Boundaries

As part of the investigation into the applicability of Fractal Image Compression techniques, it was necessary to identify realistic network bandwidth limitations. An experiment was conducted on the MILNET to determine what bandwidth a tactical user could reasonably expect at this time. A SUN 3/80 with 12 Megabytes main memory and 296 Megabytes of available disk space was utilized for this task.

A script file was set up on the SUN to retrieve six files from a remote host. The remote host was the Network Information Center in Menlo Park, California. The files consisted of Request for Comment (RFC) files which are network technical notes. The script was designed to retrieve all six files, every hour, for two weeks. However, the script implementation was not terribly sophisticated. Time required to physically transmit the files was not taken into account and the resulting "clock creep" cost us a measurement every six hours or so. Additionally, a couple days worth of data were lost due to equipment and network problems (thunderstorm induced). Data collected is summarized below; File Name refers to the actual name of the file retrieved, File Size is the physical size of the file in Bytes, Average Transmission Time (seconds) is the average of all 269 retrievals in seconds, and the Average Transmission Time (KBps) is the time required to send 1 KByte.

File Name	File Size (bytes)	Average Transmission Time (seconds)	Average Transmission Time (KBps)
rfc1119.txt	151	0.16	0.92
rfc1018.txt	7,931	7.60	1.02
rfc1136.txt	22,158	16.41	1.32
rfc969.txt	40,894	27.59	1.45
rfc1251.txt	72,721	46.78	1.52
rfc1166.txt	566,778	342.28	1.62

Tactical Air Force users currently can't expect network bandwidths in excess of 1.6K Bytes per second. Straight transmission of a 50 MB Battle Target Graphic would take 9 hours and 40 minutes.

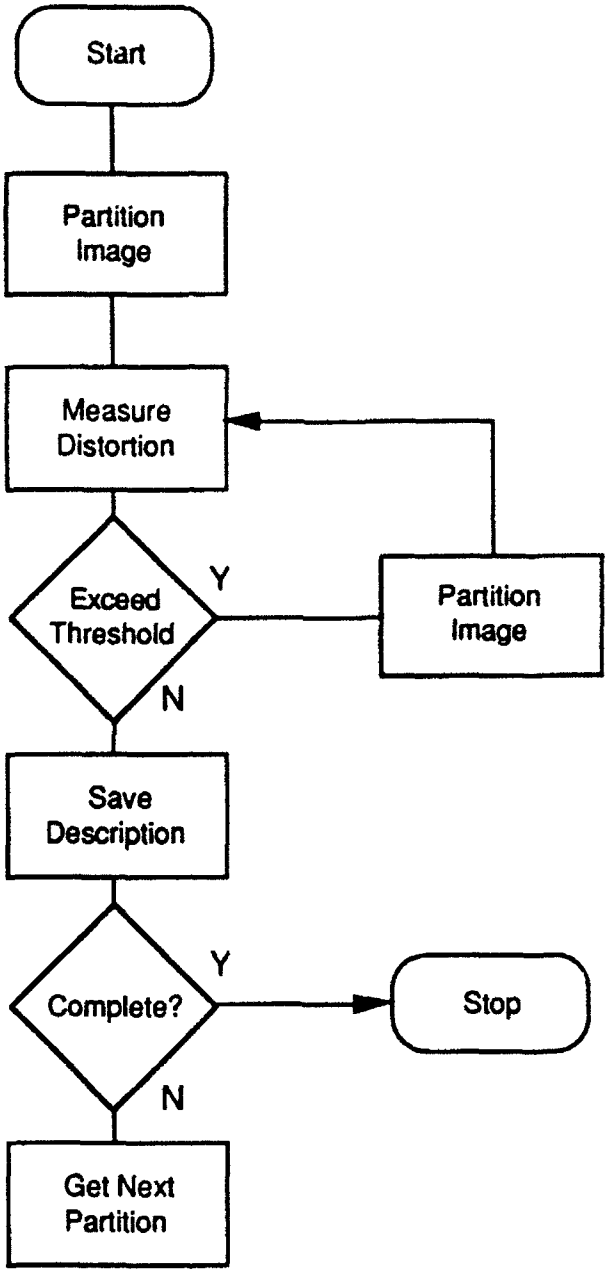
Algorithm Development

Development of our algorithm was influenced heavily by [4]. Our approach is to iteratively partition the image, measure the degree of pixel heterogeneity, generate a value if pixel heterogeneity is less than an arbitrary threshold, or re-partition if the threshold is exceeded (see flowchart 1).

To measure the pixel heterogeneity within a partition, each pixel is compared to all others in the partition. This method is described more precisely as follows:

$$\sum_{i,j=1}^{m,n} \sum_{k,l=1}^{m,n} (p(i,j)-p(k,l))^2$$

Where the partition is a two dimensional array of size m x n. The higher the value, th' greater the diversity of the pixels within a partition.


Start
Partition Image
Measure Distortion
Exceed Threshold
Y
N
Partition Image
Save Description
Complete?
Y
N
Stop
Get Next Partition

Flowchart 1

In conjunction with computing the pixel diversity within the partition, an average value is obtained to use in reconstructing the image. Once partitioning halts, the coordinates of the upper left pixel, the size of the partition, and the derived average pixel value are saved in a stack. The image is reconstructed by popping the coordinates off the stack and generating a square of homogeneous pixels.

Some questions need to be resolved during further evaluation of this particular approach. What is an appropriate threshold value to use when measuring the heterogeneity of pixels in a partition? Also, this approach is compute intensive. It remains to be seen what kind of performance is realized, especially when dealing with color images which require three bytes per pixel instead of one.

Conclusions

Fractal image compression is a lossy technique with respect to imagery. The resultant image is an approximation of the original, not the original image. The applicability of the technique will be a tradeoff of compression required versus accuracy. Imagery may be segmented into sections of interest that are compressed and transmitted using conventional techniques and sections of less importance that are compressed using fractal techniques. Compression factors of thousands to one for useful imagery (certainly from an imagery interpreter perspective) do not appear realizable with this technique. Further investigation is required.

References

[1(a)] Barnsley, Michael F
A better way to compress images,
Byte Magazine, 215 - 223, January 1988.

[1(b)] Barnsley, Michael F
Fractals Everywhere, Academic Press, 1988.

[1(c)] Barnsley, Michael F
Fractal Transform Technology, Iterated Function Systems Inc.,
Norcross, GA 30092.

[2] Culik II, Karel and Dube, Simant
Using fractal geometry for Image Compression, IEEE 1991.

[3] Hodges, Laurie ; Naylor, Bruce ; Demko, Stephen
Construction of fractal objects with Iterated Function Systems, ACM SIGGRAPH, 271 - 278, 1985.

[4] Jacquin, Arnaud , Image Coding Based on a fractal theory of Iterated contractive Image Transformations, IEEE Tran. on Image Processing, Vol. 1. No. 1, 18 - 30, 1992.

[5] Mandelbrot, Benoit
The Fractal Geometry of Nature, W.H. Freeman and Co., San Francisco, CA, 1982.

[6] Netravali, Arun N
Digital Picture - Representation and Compression, Plenum Press, 1988.

[7] Pentland, Alex P
Fractal based description of natural scenes, IEEE Trans. PAMI, 6(6), 661 - 674. (Nov. 1984).

[8] Rabbani, Majid and Jones, Paul W
Digital Image Compression Techniques,
SPIE Optical Engineering Press,
Bellingham, Washington 98227 - 0010, 1991.

[9] Rosenfeld, Azriel and Kak, Avinash C
Digital Picture Processing, Academic Press, 1976.

[10] Trivedi, Mohan M
Selected Papers on Digital Image Processing,
SPIE Optical Engineering Press,
Bellingham, Washington 98227 - 0010, 1990.

[11] Karnin, E and Walach, E
A fractal based approach to Image Compression,
IEEE Trans. on ASSP, 1986.

PROMINENCE IN SPONTANEOUS SPEECH: ANNOTATION AND APPLICATIONS

Colin W. Wightman
Assistant Professor
Department of Electrical Engineering

New Mexico Institute of Mining and Technology
Socorro, NM 87801

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

September 1992

PROMINENCE IN SPONTANEOUS SPEECH: ANNOTATION AND APPLICATIONS

Colin W. Wightman
Assistant Professor
Department of Electrical Engineering
New Mexico Institute of Mining and Technology

Abstract

Prominences, which occur when a speaker emphasizes a word so that it "stands out" to the listener, play a number of important roles in aiding the listener in the interpretation of speech. In particular, prominences are crucial for signaling many of the discourse events which conversants use to coordinate a conversation. In this work, we investigate the ability of untrained human subjects to consistently label prominences in spontaneous speech. Consistent labels will be needed to develop the automated detection algorithms necessary to make use of prominences in automatic speech processing systems. Our results indicate that, even without training, the subjects were able to produce labels in agreement with those done by an expert labeler 83.6% of the time. Moreover, merging the transcriptions done by the test subjects by using a 3-of-8 criterion produced labels in agreement with the expert 88.2% of the time and cut the missed detection rate in half.

PROMINENCE IN SPONTANEOUS SPEECH: ANNOTATION AND APPLICATIONS

Colin W. Wightman

1 Introduction

In spoken English, prominences provide the listener with information related to the semantic content of an utterance and mark significant discourse structures. This observation, however, has had a relatively small influence on the design of current speech understanding systems: prominence remains essentially unused in current speech-based systems. This has occurred for two reasons: (1) the linguistic community has not produced a unified theoretical framework which makes explicit the role of prominence in conveying syntactic, semantic, and discourse-level information, and (2) automatic methods to reliably detect prominences have not been available to developers. Both of these problems have been exacerbated by a lack of speech corpora with consistently labeled prominences. Without such corpora, neither linguistic research, nor algorithmic development can proceed efficiently. In this report, we present a methodology for labeling prominences in spontaneous speech and investigate its potential for the efficient annotation of the large corpora needed to advance the state of the art.

We are concerned with the detection of prominence, which occurs when a speaker emphasizes a word or syllable so that it "stands out" to the listener. Prominences are generally used to focus the listeners attention and mark new or important information. Automatic detection of Prominences may thus aid the interpretation of an utterance by identifying key words and facilitating semantic disambiguation. At a higher level, Prominences play a central role in marking discourse structures which conversants use to identify shifts in topic, corrections, and sub-dialogues. For example, Chen and Withgott [5] have used prominences to automatically select summarizing excerpts from

spontaneous speech. It has also been claimed that words which contain a prominence are more carefully articulated than other words [9]. If this is the case, then automatic detection of prominence could help detect these "islands of reliability" and thus improve recognition performance.

We begin in the next section by reviewing some of the discourse functions served by prominence, and the acoustic correlates which linguists have found to mark prominences. Section 3 then presents a brief discussion of how automated prominence detection might provide significantly enhanced capabilities in several systems. Then, in section 4, we describe a system for annotating prominences in spontaneous speech and present the results of an experiment to determine the ability of human subjects to label prominences consistently. Section 5 then concludes the report by examining the steps needed to realize that potential.

2 Prominence

In conversation, human listeners easily follow changes in speaker, shifts in topic, interruptions, corrections, and asides. Current speech understanding systems, however, do not have such capabilities. In large measure, this is due to the fact that they do not have access to the cues which signal significant discourse events: the speaker's prosody. Most current systems avoid this problem by restricting the user to a small task and constraining the application domain so that any uncertainty about the topic of an utterance could be trivially resolved. In the ATIS task, for example, the user is assumed to be talking only about a single, direct flight, and is limited to a simple question/response interaction. As systems for more complex tasks and with more "natural" speech interfaces are developed, however, the need for discourse analysis becomes critical for providing robust interfaces which the user can engage in a more familiar way.

For an automated system to be able to carry out effective discourse analyses, however, detection

of the prominences which signal many of the significant discourse events will be needed. In the following subsections, we will briefly explore the role of prominence in marking discourse events, and some of the acoustic correlates which influence human listeners perception of prominence.

2.1 Discourse Functions

Prominence is used to mark several different types of information of relevance to discourse analysis: new *vs* old information, confirmations, clarifications, corrections, floor-holding, and interjections[1]. Of these, the most widely investigated is the use of prominence for marking new versus old information (cf. [8]). When a speaker introduces new information, and particularly when doing so changes the focus of the conversation, the first occurrence of the "new" information is highlighted by making it prominent. Subsequent occurrences of the same information are not prominent, however, since, having been introduced, the information becomes "given" in the subsequent discussion. Of course, "given" information is not always unaccented: it may be made prominent to mark other events.

The primary purpose of a conversation is to communicate ideas between the conversants. Consequently, a great deal of conversational speech is devoted to corrections, clarifications, and confirmations, as the conversants work together to ensure that the ideas are correctly transmitted. Prominence plays a central role in this process, marking key words which signpost the dialogue. For example, a speaker might say "I'm going to eat with *Boston* . . . *in* Boston." The first occurrence of *Boston* is emphasized because it is new information. The speaker then realizes their error and repeats the clause emphasizing *in* to mark the repeat as a correction. Notice that, had the speaker corrected their error faster, the prominence would be the only way of detecting this correction: "I'm going to eat with *in* Boston".

[1]The literature contains a vast number of overlapping, and sometimes conflicting, terms. For reasons of simplicity, we will primarily use the taxonomy developed by Allen [1], with extensions as needed.

Prominences are also used to mark clarifications, which provide information needed to interpret a previous statement: "Let's do it. Let's *buy* the house." In this case, the intended action (*buy*) is emphasized to mark it as a clarification (should we buy or rent?). Alternatively, *house* could have been emphasized if the clarification was in answer to the question "should we buy a car or a house?" Note that this pair of examples also demonstrates that interpretation of clarifications often requires knowing the context and current focus of the conversation.

In addition to corrections and clarifications, prominences are used to mark acknowledgments, which signal that the speaker has understood (but not necessarily accepted) the previous statement such as *okay*, *yes*, and *I* see. Similarly, interjections such as *wow!*, *whoa!*, and *never!*, are always prominent.

Finally, speakers often engage in floor-holding when they need time to generate the remainder of their utterance, but do not wish to be interrupted during the break. Floor-holding can take the form of a prominent function word immediately before the break such as in "I'd love to come, *and* ...Alternatively, the speaker may choose to fill the pause with *uhm* ... or *err* ..., which many listeners perceive as prominent.

It should be clear from the above discussion that prominences mark many different kinds of discourse events in several different ways. Indeed, as reported by Chen and Withgott, simply extracting the prominent words from a conversation does not produce an intelligible summary of it [5], but extracting phrases which contain clusters of prominences will. Indeed, while a prominent word is often a crucial content word (such as when introducing new information), many prominent words are function words with very little meaning by themselves. Thus, in terms of discourse, detection of prominences alone is not sufficient. It is, however, necessary both for doing the analyses, and for developing the corpora which will be needed to determine how to do the analyses.

2.2 Acoustic Correlates

In order to detect prominences, we should investigate their acoustic correlates. That is, what changes in the acoustic signal cause a listener to perceive one word as prominent and another not? To answer these questions, we need only consult the extensive linguistics literature on the subject. In essence, there appear to be three principal acoustic correlates: segmental durations, pitch accents, and energy changes. Of these, pitch accents have received the most attention, principally as part of larger phonological theories of intonation (cf. [11, 2, 3, 6]). Pitch accents are distinctive intonational features which cause listeners to perceive the syllable in which they occur as prominent. Although there is considerable disagreement in the literature about whether it is a particular pitch *value* or the pitch *movement* which cues the perception of a pitch accent, there is complete agreement that the intonational contour marks prominences.

In addition to intonational marking, prominences are marked by lengthening of the segmental durations in the prominent syllable. Indeed, Campbell [4] has shown that the segments within a prominent syllable are uniformly lengthened, which makes it possible to distinguish prominence-related lengthening from pre-boundary lengthening which occurs only in the syllable rhyme. Finally, many listeners report amplitude as a correlate of prominence. That is, they think prominent syllables are louder than others. Several perceptual experiments (e.g. [14]), however, have shown that, while amplitude does play a role in the perception of boundaries, that role is small and can be easily overridden by either intonational or durational cues.

3 Utilization of Prominence

Having briefly explored some of the roles prominence plays in normal, human to human, communications, we need to determine if any of those roles are of relevance to communications between

humans and machines. In particular, we need to identify ways in which automated detection of prominence could improve the ability of speech processing systems to fulfill their missions. From this perspective, one can identify three general roles for automatic detection of prominence: (1) labeling prominences in large research corpora, (2) coding prominences for speech compression/conversion, and (3) detecting prominences for discourse analysis.

The most straightforward application of a prominence detection algorithm would be in speech coding. For example, one approach to extremely narrowband coding (*e.g.* 50 baud), is to use a speech recognizer on the transmitting end, send codes to identify the words recognized, and then resynthesize the words at the receiving end. While this method yields perfect intelligibility, it destroys all of the prosodic information in the speech. Automatic detection of prominence would permit each word to be tagged with a one-bit flag indicating whether that word should be resynthesized as prominent or not. This could substantially enhance the utility of the narrowband channel by allowing the users to coordinate their discourse activities in a more normal manner. The simplicity of this application comes from the fact that the system need only be able to detect and synthesize prominences: all of the discourse processing cued by the prominences is done by the users of the system.

A similar application would be to integrate prominence into a voice translation system which translates an utterance spoken in one language to another language. Again, this is done by using a speech recognizer, some translation process, and synthesizing the output. And, as with narrowband coding, adding prominence detection and synthesis would not require understanding all of the ways prominence is used in coordinating a dialog: the users provide that function. What would be required is some modification to the translation process to make sure that the prominence gets resynthesized in the right place. If the speaker stresses the verb of a sentence, for example, the

system's output should stress the verb too, even though it may appear in a completely different place in the sentence.

A much more ambitious application would be a system which actually analyzes the speaker's discourse using the prominences as a cue. For example, automatically detected prominences might be integrated into a gisting system, to help identify and extract the topic of conversation. Such a system would require not only an automatic detection algorithm, it would also require a detailed understanding of how prominence can be used to identify the desired information in a conversation. Such an understanding can only come, however, from further research into the roles of prominence, which brings us to the third application for automated prominence detection: labeling research corpora.

Researching the roles of prominence will require the use of a large (several hours) corpus of speech in which prominences have been labeled. Labeling prominences in such a large corpus, however, is a daunting task. While hand-labeling can be quite accurate, it is also extremely time-consuming: in the test described below, eight person-hours were required to transcribe less than ten minutes of speech. Clearly, labeling of the large corpora required will necessitate the application of an automated labeling algorithm. Even if the labels produced by the algorithm need to be hand corrected in a second pass, this is still likely to be far faster than labeling by hand. Indeed, this is the approach used in the University of Pennsylvania's TREEBANK project [10] which efficiently annotates syntactic bracketing in very large corpora.

4 Transcribing Prominence

In order to study prominence or to develop automatic methods for detecting it, we need to develop a means of consistently transcribing prominences. In this section, we explore the experiences of

other researchers in transcribing prosody, and describe a core transcription system. We then report on an experiment to determine the speed and consistency with which human listeners could label prominences in spontaneous speech.

Conceptually, labeling prominence is very straightforward: the syllables which are perceived as prominent should be marked, leaving the other syllables unmarked. Most listeners, however, claim to hear more than one level of prominence, with some syllables being very prominent, and others being less so, but still emphasized with respect to the remainder. Unfortunately, in two separate studies [12, 5], other researchers have tried to label more than one level of prominence (two levels in [12], and three levels in [5]), and found that, while there was good agreement between labelers about the location of a prominence, there was very poor agreement as to the level of prominence. In both cases, the researchers were forced to merge the multiple levels and simple mark each syllable as *prominent* or *not* prominent.

It is for this reason that the marking of prominence in the core transcription under the TOBI system (a standardized system of prosodic annotation being developed by the speech and linguistics community [13]) is simply a "*" on each prominent syllable. And, in keeping with this standard, we have adopted this notation in the present study. A crucial question remains, however: how consistently can human labelers use even this simple notation. As mentioned above, previous researchers have claimed that labelers are in close agreement on the locations of prominences, but these claims are based on somewhat qualitative obser tions. Neither group has extensively investigated this issue, and no systematic study of the consistency or speed of the labeling has been reported.

4.1 Corpus

We investigated the issue of labeling consistency by conducting an experiment using a corpus of spontaneous speech. The corpus, often referred to as the KING database [7], consists of excerpts from telephone conversations in which the subjects were asked to describe various objects and pictures over the phone. The subjects were recorded both locally, and over the telephone line which makes it possible to evaluate a system on the same speech under both clean and noisy conditions. Roughly fifty speakers took part in the test which was done in ten sessions producing five hundred excerpts of between forty and sixty seconds each.

For our evaluation of prominence labeling, we utilized a subset of this corpus drawn from the first session. We used the excerpts from the first nine speakers, which yielded nine excerpts containing a total of roughly seven minutes of speech.

4.2 Labeling Experiment

For our experiment, we recruited eight labelers who had never previously labeled prosodic information. Transcriptions of each speech utterance were prepared with each word appearing on a single line adjacent to a set of *'s, one "*" for each syllable in the word. These transcriptions were then provided to the eight test subjects who were instructed to circle the "*" corresponding to any syllable which they perceived as prominent. The subjects were told only that a prominent syllable was one which the speaker had emphasized so it "stands out" from the others. Specific acoustic correlates such as lengthening, or pitch accents, were not mentioned.

The test itself was conducted by playing tapes of the target utterances to the panel of labelers. The tape was stopped, and sections repeated until all labelers indicated their readiness to move on. After labeling the ten utterances, the subjects were asked for comments on the labeling process.

	unmarked	prominent
unmarked	20533	2449
prominent	2195	3019

Table 1: Confusions between pairs of labelers.

4.3 Results

We have evaluated the results of our labeling experiment in two ways: (1) in terms of the agreement between the untrained labelers, and (2) in terms of the agreement between the untrained labelers and an expert labeler. In the discussion that follows, agreement or disagreement between labelers was determined by checking to see if they both marked a prominence on *any* syllable in a word. Thus, two labelers who marked prominences on different syllables in the same word were considered to be in agreement. The reason for this is that the test subjects found it very difficult to identify the prominent syllable within a word. More experienced subjects, however, can quite reliably localize the prominent syllable which, in most cases, is just the lexically stressed syllable.

To determine the agreement between untrained labelers, we compared the labels produced by a given labeler pair-wise with the labels produced by each of the other labels. Thus the labels produced by subject 1 were compared with those produced by subject 2, and with those of subject 3, *etc.*, yielding eight sets of comparisons. This was done for each labeler, yielding a total of 72 sets of comparisons. From these comparisons, we then generated the confusion matrix shown in table 1.

Several comments should be made about table 1. First, since none of the label sets can be considered more or less correct than any of the others, the off-diagonal elements in the table represent the same thing: disagreements between labelers. Recognizing this, we can combine them to form three classes: agreement on unmarked words (20533), agreement on prominent syllables (3019), and disagreements (4644). Thus, the labelers were in agreement 83.5% of the time. Considering

expert	*test subjects*	
labels	unmarked	prominent
unmarked	5649	439
prominent	881	1087

Table 2: Confusions between labels produced by untrained labelers and those produced by an experienced expert.

that the labelers were untrained, this is a remarkable degree of consistency.

To evaluate the accuracy of the labels produced by the untrained subjects, we compared them to a set of labels produced by an expert transcriber with several years of experience transcribing prosodic phenomena including prominence. By comparing the expert labels pair-wise with the labels produced by each untrained labeler, we obtained the confusion matrix shown in table 2.

Table 2 differs from table 1 in that we now have a reference set of labels and, consequently, the off-diagonal elements no longer represent the same thing. Instead, they indicate the number of false detections (7.2%) and missed detections (44.7%) of prominences labeled by the expert. Notice that there is a strong bias towards missed detections: the number of prominences labeled by a test subject that were not labeled by the expert is rather small. On the other hand, when the expert labeled a prominence, almost half of the subjects failed to mark it. Thus it appears that the test subjects were much more conservative than the expert, marking only large prominences and missing many of the smaller ones which the expert labeled. Nevertheless, agreement between the untrained labelers and the expert was 83.6%, overall.

While the number of missed detections is disappointing, it may be possible to improve the agreement between the untrained and expert labelers by combining the labels produced by all the untrained labelers into a single transcription. Indeed, this would have to be done eventually, since the goal of a labeling task is to produce one, single transcription. We have investigated combining

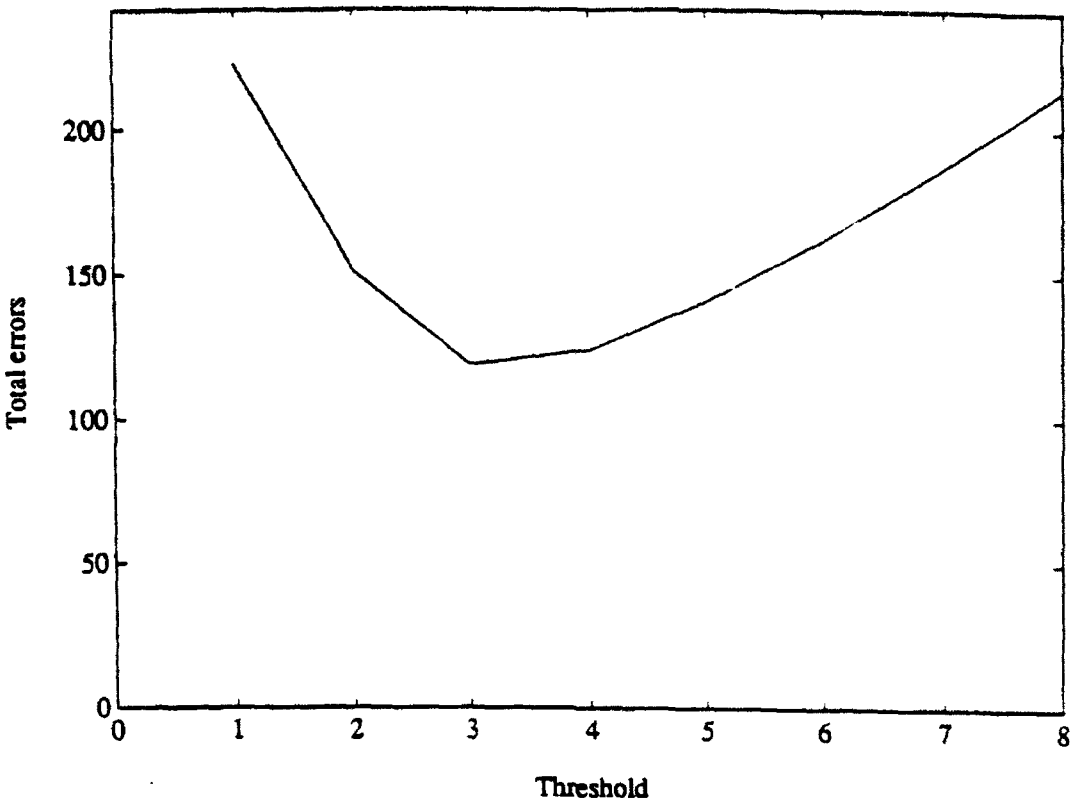


Figure 1: Total disagreements between merged labels and expert labels for different threshold values.

the label sets produced by the untrained labelers by using a "n of m" test. Essentially, a word is marked as prominent in the final transcription only if n, or more, of the m untrained labelers marked it as prominent. In our study, m is equal to eight. To determine the best value for the threshold n, we plotted the total number of errors (false detections and missed detections) for thresholds ranging from one to eight, as shown in figure 1.

From figure 1, it is clear that the best threshold value is 3. Using this value, we produced a single, merged transcription and then generated a confusion matrix between it and the expert labels, as shown in table 3. From the table, it can be seen that the overall agreement has been improved to 88.2%. In particular, merging the transcriptions from the untrained labelers cut the missed detection rate in half (to 23%), while only increasing the false detection rate to 8.1%. Thus, merging the transcriptions of untrained labelers can significantly reduce the number of labeling errors and produce a transcription that is fairly close to one produced by an expert labeler.

expert labels	*merged labels* unmarked	prominent
unmarked	699	62
prominent	57	189

Table 3: Confusions between transcriptions produced by merging the untrained labelers with a threshold of 3, and an expert labeler.

5 Conclusions

In the work reported here, we have laid the groundwork needed to significantly enhance the capability of several speech processing systems through the integration of automatic detection of prominence. We reviewed the roles which prominence plays in discourse, and examined some of the ways which current speech processing systems could make use of them. We also identified the need for a fast, consistent means of labeling prominences.

To determine the ability of untrained human subjects to consistently label prosody, we conducted an experiment and found agreement between the labelers to be 83.5%. However, when we compared these labels with those produced by an expert, we found a missed detection rate of 44.7%. This appeared to be due to conservative labeling on the part of the untrained subjects: they were only marking very large prominences. Combining their individual labels into a single, merged transcription using a 3-of-8 criterion overcame this problem, and cut the missed detection rate in half while raising the agreement with the expert labels to 88.2%.

The merged transcription is still not of sufficient quality that it could be used directly. This could be remedied by hand correcting the merged transcription. Alternatively, it is quite possible that a small amount of training and practice would produce labelers whose merged transcriptions could be used directly. Either way, we have established that the initial, time-consuming process of

marking prominences in spontaneous speech can be done by a panel of relatively inexperienced (and hence low-cost) subjects. This is a valuable result since development of an automatic algorithm will initially require hand labeling a substantial amount of speech for training and testing. The ability to do this quickly, and at low cost, is crucial to the eventual success of any effort to develop an automatic algorithm.

References

[1] J. Allen. "Discourse stucture in the TRAINS project". In *Proc of DARPA Speech and Natural Language Workshop*, San Mateo, CA, February 1991. Morgan Kaufmann.

[2] M. Beckman and J. Pierrehumbert. "Intonational structure in Japanese and English". *Phonology Yearbook 3*, pages 255–309, 1986.

[3] D. Bolinger. "Pitch accent and sentence rhythm". In D. Bolinger, editor, *Forms of English: Accent, Morpheme, Order*. Harvard University Press, Cambridge, MA., 1965.

[4] W. N. Campbell. "Evidence for a syllable-based model of speech timing". In *Proceedings Int. Conf. Spoken Language Processing*, pages 9–12, Kobe, Japan, 1990.

[5] F. Chen and M. Withgott. "The use of emphasis to automatically summarize a spoken discourse". In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing"*, pages I–229, 1992.

[6] A. Cohen and J. 't Hart. "On the anatomy of intonation". *Lingua*, 19:177–179, 1967.

[7] A. Higgens, E. Wrench, L. Bahler, J. Porter, D. Schmoldt, and M. Lipps. "Speaker identification and recognition", Final Report, Contract 88-F744200-000. Technical report, ITT Aerospace/Communications, San Diego, CA, 1991.

[8] J. Hirschberg. "Assigning pitch accent in synthetic speech: The given/new distiction and deaccentability". In *Proc. of the Seventh National Congress*, Boston, 1990. American Association for Artificial Intelligence.

[9] W. Lea. "Prosodic aids to speech recognition". In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice-Hall, Inc., 1980.

[10] M. Marcus and B. Santorini. "Building a very large natural language corpora: The Penn treebank". Submitted manuscript.

[11] J. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, 1980.

[12] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. "The use of prosody in syntactic disambiguation". *J. Acoust. Soc. Am.*, 90:2956–2970, 1991.

[13] K. Silverman, J. Pitrelli, M. Beckman, J. Pierrehumbert, R. Ladd, C. Wightman, M. Ostendorf, and J. Hirschberg. "TOBI: a standard for labeling English prosody". To appear in *Proc. of International Conf. Spoken Language Processing*, Banff, Canada., 1992.

[14] L. Streeter. "Acoustic determinants of phrase boundary perception". *Journal of the Acoustical Society of America*, 64(6):1582–1592, 1978.

THE CASE FOR LIKE-SENSOR PRE-DETECTION FUSION

Peter Willett
Associate Professor
Electrical and Systems Engineering Department

University of Connecticut
U-157, Storrs CT 06269

Final Report for:
Summer Faculty Research Program
Rome Laboratory – OCTM
Griffiss AFB, NY 13441

Sponsored by: Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

September 1992

THE CASE FOR LIKE-SENSOR PRE-DETECTION FUSION

Peter Willett
Associate Professor
Electrical and Systems Engineering Department
University of Connecticut

Abstract

There has been a great deal of theoretical study into decentralized detection networks composed of similar (often identical), independent sensors, and this has produced a number of satisfying theoretical results. At this point it is perhaps worth asking whether or not there is a great deal of point to such study – certainly two sensors can provide twice the illumination of one, but what does this really translate to in terms of performance?

We shall take as our metric the ground area covered with a specified Neyman-Pearson detection performance. To be fair, the comparison will be of a multi-sensor network to a single-sensor system where both have the same aggregate transmitter power. The situations examined are by no means exhaustive but are, we believe, representative.

Is there a case? The answer, as might be expected, is "sometimes". When the statistical situation is well-behaved there is very little benefit to a fused system; however, when the environment is hostile the gains can be significant. We shall see, depending on the situation, gains from co-location, gains from separation, optimal gains from operation at a "fusion range", and sometimes no gains at all.

THE CASE FOR LIKE-SENSOR PRE-DETECTION FUSION

Peter Willett

Introduction

In this report we shall consider the canonical "parallel-topology" decentralized (or distributed) detection network as pictured in Figure 1. With reference to this, there are assumed to be N *sensors*, the i^{th} of which observes data X_i and transmits to a *fusion center* a version of its observation which we label U_i. This latter processor is responsible for the ultimate decision (which we shall assume to be about the presence or absence of a target) obtained via the appropriate weighing of the various sensors' inputs.

There are of course a number of assumptions commonly (but not uniformly) made in studies in decentralized detection, and we list these here numbered 1-3.

Assumption 1 *The sensors are all of the same sort.*

The idea here is that all sensors being considered are, for example, common band radars which do not interfere with one another. This is as opposed to the fusion of different-modality sensors (such as radar and infrared), which is intuitively-appealing; however, much current academic work deals with like-sensor fusion (and particularly with independent and identically distributed (*iid*) sensors, which is perhaps less-defensible practically) and these are what we shall examine.

Assumption 2 *The sensors' observations X_i are independent of one another conditioned on the hypothesis (target present/absent) parameter.*

This is certainly convenient in that a number of theoretical results apply to make optimization and practice more straightforward, such as U_i being a quantized version of X_i [1, 2, 3] and the weighted sum form of the fusion rule [4]. It is also plausible on intuitive grounds, as noise processes observed by different sensors in different locations, probably at different times, are unlikely to be predictable one from another.

Assumption 3 *The sensors are viewing an identical region of space at the same time.*

It would be very unlikely that nonco-located sensors could share a common understanding of space in the sense of a grid of identical and coincident resolution cells. A common grid *is* feasible for co-located sensors, but even here it is clear that a great deal of calibration is required. Also, it is at least unlikely that in all but the most benign of environments sensors will scan the same volume simultaneously. However, it is possible that a practical system incorporating measures dealing with these non-coincidences will share many features with those designed under assumption 3. Most research into decentralized detection begins from here.

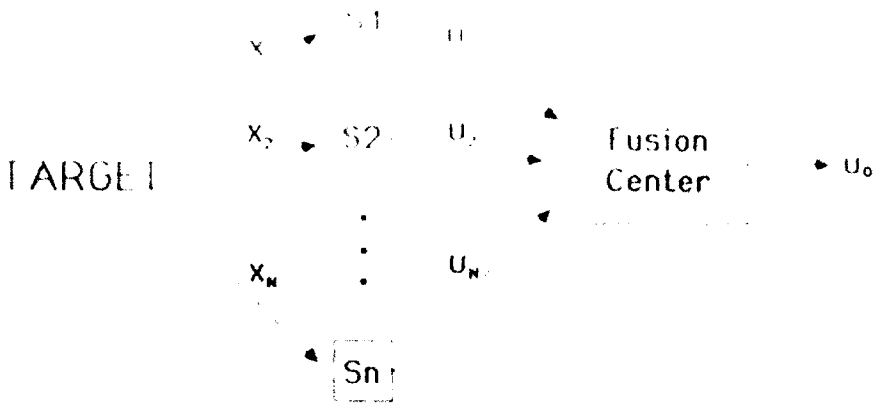


Figure 1: A decentralized detection network with parallel information flow.

Returning to the thrust, is there a case for pre-detection fusion? The obvious response is "yes" – if one has available, for example, two sensors, then combining their data corresponds to somewhat less than but approximately a doubling of received power, and indeed we shall observe attendant gains in performance. However, to say that if one has a resource one should use it is tautological, and in this report our aim is to see whether or not one should deliberately go about building a like-sensor decentralized detection system. To respond to this we must lay some groundwork in the form of further assumptions and the methods used for comparison.

Assumption 4 *The total transmitter power in the system will be held constant, regardless of the number of sensors.*

As mentioned above, to state that fusion of four $100kW$ sensors can outperform one does not lead to any interesting conclusion. However, to be able to say that four $25kW$ sensors are an improvement over a single $100kW$ sensor *is* significant, and we shall wherever possible make this assumption.

Assumption 5 *The comparison will be carried out on the basis of* ground area covered *by a fused-data system having a specified performance.*

For example, we might say that a given single-sensor system can observe targets with a false-alarm probability of 10^{-6} and with a probability of detection of at least 80% within a 7,500 square mile area; if a two-sensor system could likewise cover 15,000 square miles it is a significant improvement, whereas if the figure were 8,000 square miles the superiority could be termed negligible or at least not worth the cost or complication.

Assumption 6 *The signal to noise ratio (SNR) observed by a given sensor varies as the inverse fourth-power of the distance between the sensor and the resolution cell in question.*

The fourth-power law [5] is a well-known feature of the radar equation for free space. An implication is that the area covered varies as the square root of the transmitter power; that is, the area covered by the fusion of two co-located sensors will certainly be no greater than 42% larger than it would be without fusion.

Assumption 7 *We shall assume a Swerling I target model with cell-averaging (CA)-CFAR processing and a homogeneous reference window.*

The model, therefore, is

$$\begin{aligned} \mathbf{H_0}: \quad \Pr(X_i \geq x_i) &= \frac{1}{(1+x_i)^m} \\ \mathbf{H_1}: \quad \Pr(X_i \geq x_i) &= \frac{1}{(1+\frac{x_i}{1+S_i})^m} \end{aligned} \tag{1}$$

where H_0 is the target-absent hypothesis, H_1 is the target-present alternative, X_i is the square-envelope return (a sufficient statistic under this model), and S_i is the SNR as observed by the i^{th} sensor. The parameter m corresponds to the number of reference cells used for noise-power normalization (CFAR), and will usually for our purposes be taken as $m = 8$. One would of course prefer not to make any such assumption; however, without a statistical model there does not appear to be a framework for comparison. The model of assumption 7 has the virtues of being both computationally tractable and of practical interest, but the disadvantage of being perhaps more well-behaved than many real-world situations. At any rate, we shall adopt it in most of what follows, but shall briefly explore the Swerling III/CA-CFAR scenario, a more-unfriendly variable fluctuation model, and heavy-clutter.

Assumption 8 ***Communication from the sensors to the fusion center is binary, and the entire system is optimized under this constraint.***

It should be noted that at the "other extreme" is a system in which the unquantized local likelihood ratios are transmitted – further schemes in which the transmissions are members of M-ary alphabets lie between these two in terms of performance, but are a great deal more complicated to optimize. At any rate, given the independence assumption SNR losses for binary versus unquantized schemes are usually in the range of a few dB; a study based on unquantized likelihood ratio transmissions would show slightly larger ground coverage areas than those we shall present, but would reach similar conclusions.

Positioning of the Sensors

In this section we shall explore the ramifications of "moving" the sensors around spatially. We shall begin with two somewhat trivial cases: that in which sensors are so separated that pre-detection fusion, in the truly collaborative meaning, is not possible; and that in which the sensors are co-located. Both of these provide insight, but are less interesting than the third more general case that sensors are not co-located but do indeed collaborate. It will be shown that as sensors separate their aggregate ground coverage increases, but not, as might be expected, monotonically: there is a "fusion range" at which group decision-making provides significant benefit.

Disjoint Sensors

Consider a group of sensors which are separated by a sufficient distance that they do not collaborate. We may define this formally by denoting as $B_i(\alpha, \beta)$ the ground region covered by sensor i with probability of false alarm α and probability of detection at least β. By "disjoint" we mean that if a point $(x, y) \in B_i(\alpha, \beta)$ with $\beta = \alpha + \epsilon$ and some small ϵ, then $(x, y) \notin B_j(\alpha, \beta)$ for any $j \neq i$. The situation is pictured in Figure 2(a).

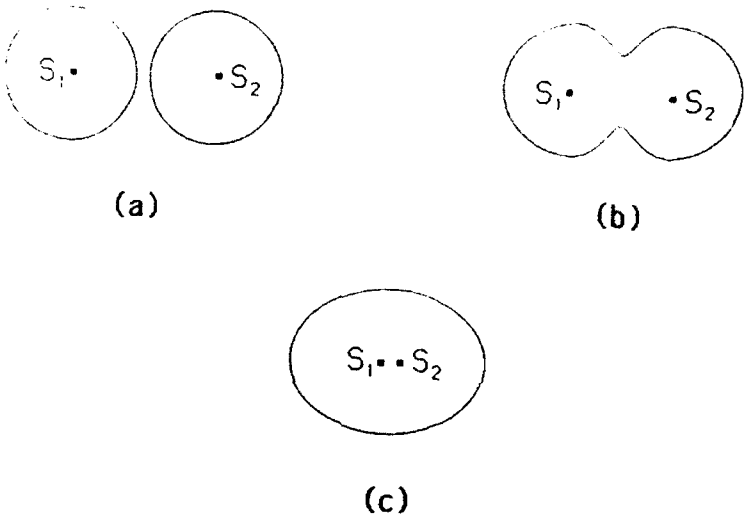


Figure 2: Positioning and qualitative performance of a two-sensor distributed network. (The contours indicate ground covered by the network with a specified performance.) ((a) Disjoint sensors; (b) Collaborating sensors; (c) Almost co-located sensors.)

Returning to our framework, we compare a single sensor with transmitter power P to a group of n disjoint sensors each with transmitter power P/n. Via assumption 6 we have that the maximum range observable with a specified α and β performance is proportional to $P^{1/4}$; the ground area covered is proportional to the square of this maximum range, and hence

$$\text{Area}_{n\ disjoint\ sensors} = \sqrt{n}\,\text{Area}_{single\ sensor} \tag{2}$$

since there are n such sensors.

As intuition suggests, the multiple-sensor system deals more effectively with the inverse fourth-power law than does the single (strong) sensor. One must exercise caution when interpreting this result, of course, since it is tempting to propose a system with 10,000 sensors each with transmitter power $10W$

as being 100 times as good as a single $100kW$ transmitter - in fact it may not be possible to tesselate space in the appropriate manner with a large number of small sensors, and in any case a sensor's "cost" is by no means directly proportional to its transmitter power.

Co-Located Sensors

Let us now turn our attention to the case of the opposite extreme, that in which n sensors each with transmitter power P/n are located at the same site, to be compared to a single sensor with power P. This is as pictured in Figure 2(c).

The simplifying aspect of this scenario is that all sensors operate with the same SNR at all resolution cells – this is truly the *iid* sensor case so often studied. We make two observations: first, that the statistical model satisfies the conditions of [6] that to have all sensors use the identical operating point is at least a *local* (in the sense of optimization) best choice; and second that the performance of a system with $n-j$ sensors each with power P/n decreases with j.

These two facts (which are straightforward enough that we omit the detail) ensure that in an optimal co-located system sensors should indeed use the same operating point, and hence that only "k-out-of-n" fusion rules need be considered. We do not know k *a-priori*; however, proceeding in the manner of [6, 7] we may define

$$g_k(x) = \sum_{l=k}^{n} \binom{n}{l} x^l (1-x)^{n-l} \tag{3}$$

and note that consequently

$$\beta = \max_{k \in \{1,2,\ldots,n\}} \left\{ g_k \left(\beta_{local} \left(g_k^{-1}(\alpha) \right) \right) \right\} \tag{4}$$

where $\beta_{local}(\alpha_{local})$ denotes the local-sensor receiver operating characteristic (ROC). Deriving an optimal system is hence a matter only of computing the maximum of n numbers. An example of this is shown in Figure 3 which shows the regions of false-alarm-rate/SNR space in which the various fusion rules are optimal.

Figure 4 deals more directly with the subject at hand. This is a plot of ground coverage (see assumption 5) versus number of sensors for various probabilities of detection and at a false alarm rate of 10^{-5}. To interpret this plot we note that:

- The "relative ground coverage" of the vertical axis is the actual area covered by the multi-sensor system expressed as a fraction of the single-sensor coverage. That is, a reading of unity on the vertical axis would correspond under our CA-CFAR model to

$$\text{unity area} = \pi \left(\frac{S_1}{\frac{\alpha^{-1/m}-1}{\beta^{-1/m}-1} - 1} \right)^{1/2} \quad [\textit{square distance units}] \tag{5}$$

  where S_1 is the SNR observed by a single-sensor (with the aggregate power) system when the target of interest is one *distance unit* away, and where α and β are the specified false-alarm and detection probabilities, respectively.

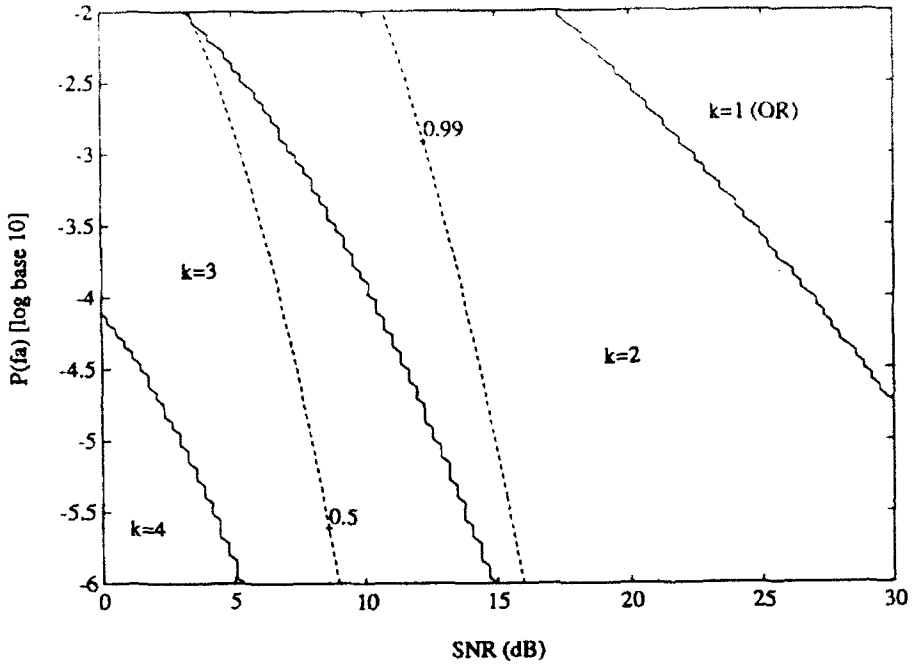


Figure 3: Regions for which various k-out-of-n fusion rules are optimal. (There are six co-located sensors, and the superimposed dashed lines are contours indicating the probability of detection performance. Notice that $k = 5$ and $k = 6$ are optimal for no $P(fa)/SNR$ pair shown on this plot.)

- If a false alarm rate other than 10^{-5} were chosen then the plots would change; however, the qualitative nature would be unaltered.
- There is a fundamental difference between a system with an even number of sensors and an odd – this is manifested in the curves' lack of smoothness.

From Figure 4 we observe an interesting non-uniformity: as the number of co-located sensors increases the ground area covered for low-quality detection (e.g. 50%) *decreases*, while that for high-quality *increases*. This may be a property of the statistical model chosen, but it is intuitive that small and numerous sensors be very effective in a nearby neighborhood, but due to the inverse fourth-power law none can "see" as far as the single large sensor.

More General Sensor Positioning

Consider now the situation of Figure 2(b), in which a more general configuration of sensors provide data to be fused. Such a case is more complicated than either of those previously encountered, since the sensors *do* collaborate, but the differing sensor qualities (i.e. SNRs) must be taken into account. The upshot is that optimization of the network is more involved, and for this we have been forced to appeal to a gradient-based algorithm as described in [8].

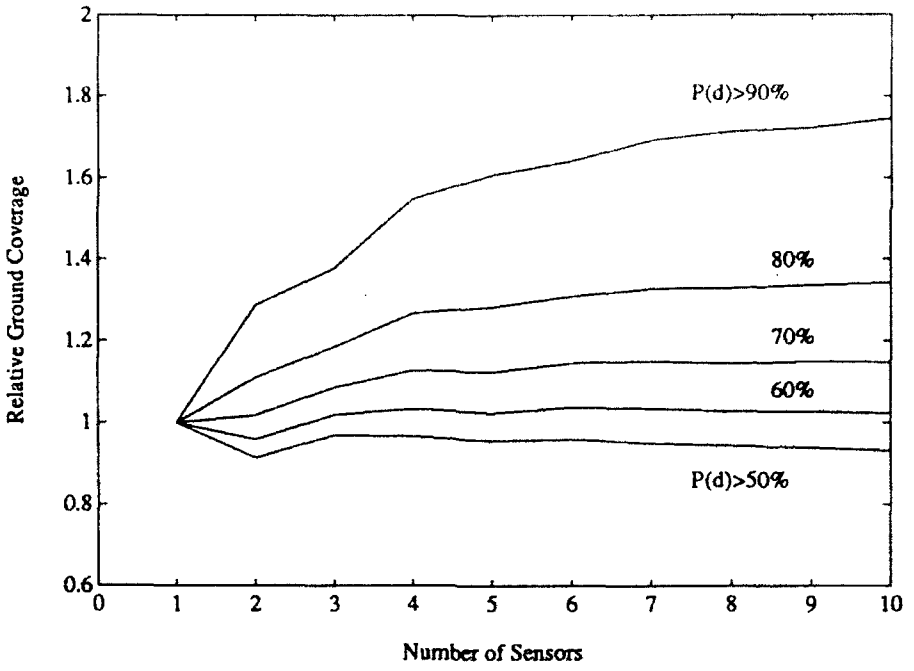


Figure 4: Ground area covered by optimal fused system with various numbers of sensors. (Probability of false alarm 10^{-5}, co-located sensors, equal aggregate power.) Coverage is "relative" to a single-sensor system.

As before we are interested in ground area covered by the network. Our approach is to assume that sensors are spatially "moved" away from some central location at equal rates. Specifically, if we call the distance from each sensor to its nearest neighbor d and take the central point as having coordinates $(0,0)$, then for a two-sensor system the sensors are located at $(-d/2, 0)$ and $(d/2, 0)$ [collinear]; for three sensors at $(-d/2, -\sqrt{3}d/6)$, $(d/2, -\sqrt{3}d/6)$, and $(d/\sqrt{3}, 0)$ [an equilateral triangle]; and for four sensors at $(-d/2, -d/2)$, $(d/2, -d/2)$, $(-d/2, d/2)$, and $(d/2, d/2)$ [a square].

The results are plotted in Figures 5 and 6 in the form of coverage versus d. The difference between the first two figures is that in the former low-quality (probability of detection at least 50% only) coverage is considered and in the latter high-quality coverage (correspondingly 90%), and both are designed for an overall false alarm probability of 10^{-5}. Note that in these figures, and also in those which follow, the results are scaled relative to a single-sensor system. Our notes:

- The "sensor separation" is expressed as a fraction of the maximum single-sensor range, which we shall call the *standard radius*. That is, under our model, with a specified performance (α, β), and given a transmitter/target pair, then the standard radius is equivalent to

$$\text{standard radius} \;=\; \left(\frac{S_1}{\frac{\alpha^{-1/m}-1}{\beta^{-1/m}-1} - 1} \right)^{1/4} \qquad [\textit{distance units}] \tag{6}$$

where S_1 is the SNR observed by a single-sensor system with the target one distance-unit away.

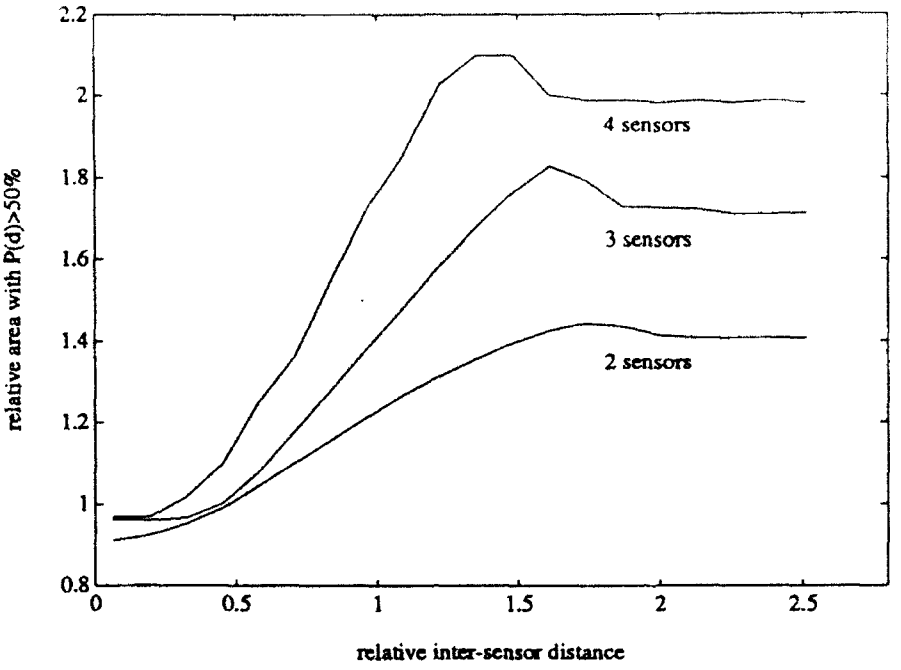


Figure 5: Ground coverage for multi-sensor system with probability of detection greater than 50% and probability of false alarm 10^{-5}. Coverage is "relative" to a single-sensor system, and range is "relative" to the maximum range for a single-sensor system.

- For this scenario when the sensors are more than twice the standard radius apart they are effectively "disjoint". This is manifested by the horizontal curves beyond this point.
- As predicted, for disjoint sensors the coverage varies as the square root of their number; that is, four disjoint sensors cover twice as much ground area as a single sensor with the same aggregate power.
- For a multiple sensor co-located system there is little improvement in the low-quality coverage, but the high-quality coverage is considerably expanded; this is as predicted in Figure 4.
- The increase in coverage is *not monotonic* in separation, particularly for high-quality coverage.

The non-monotonicity is perhaps the most interesting.

The clear peak in coverage is surprising, but referring to Figure 7 the cause is clear. As the sensors move apart the duplication in their initial co-located ground coverage disappears. Before the tesselation advantages of disjointness become dominant, however, there is a range encountered at which the middle of their pattern (i.e. surrounding coordinates $(0,0)$) is insufficiently illuminated by any one of the sensors alone but for which collaborative fusion is possible. For want of a better term we shall say that this happens at the *fusion range*.

To understand this more fully let us turn attention to the optimal fusion rule given in Figure 8. The contours indicate the number of binary n-vectors (there are of course 16 such possible in this case, as

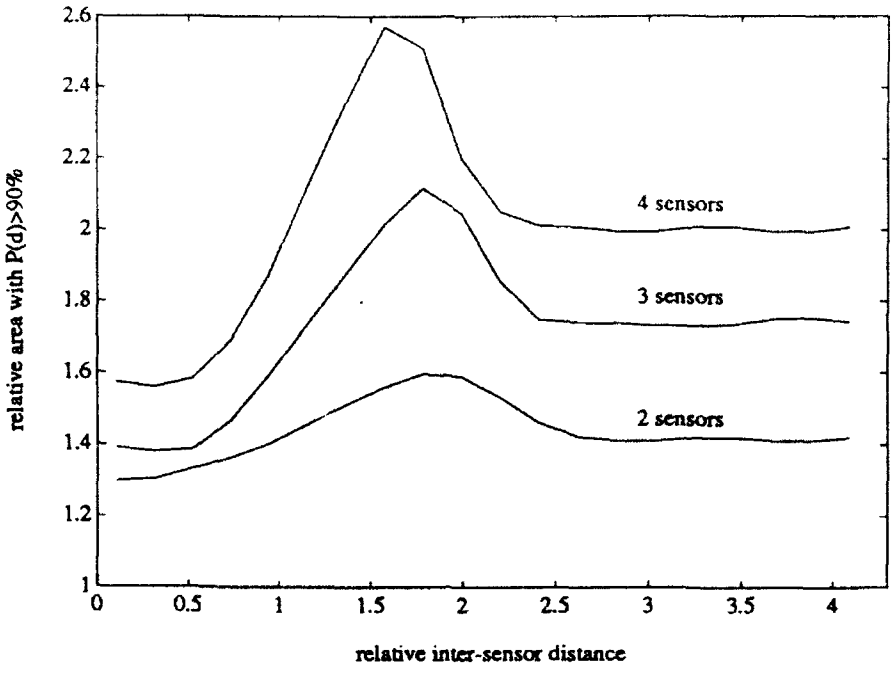


Figure 6: Ground coverage for multi-sensor system with probability of detection greater than 90% and probability of false alarm 10^{-5}. Coverage is "relative" to a single-sensor system, and range is "relative" to the maximum range for a single-sensor system.

$n = 4$) which cause the fusion center to decide for H_1. The reason for the complicated nature of this plot is that in an n-sensor network there are $2^{2^n} - 1$ (65535 in this case) possible fusion rules, and as such presentation is difficult. Certain example points, as indicated by the lettered locations in Figure 8, are illustrated in Table 1. The notation here is Boolean: for example, at location e of Figure 8 the optimal fusion rule should be read "decide H_1 iff $u_3 = 1$ *or* $u_1 = 1$ *and either* $u_2 = 1$ *or* $u_4 = 1$". It should be stressed that Figure 8 is by itself only a partial description, in that contiguous regions with an identical number of of binary 4-vectors comprising the fusion center's decision region for H_1 will not appear different in the plot even if the 4-vectors themselves *are* different. However, there is a symmetry which we have tried to draw out by observing, for example, the rules at locations d and e; similarly, if the location b were close to sensor 3 rather than sensor 1 the rule would be $u_3 + (at\ least\ two\ of\ \{u_1, u_2, u_4\}$.

Other Considerations

Alternative Models

Swerling III

The Swerling I model pre-supposed by equation (1) is applicable to targets composed of many small scatterers, none of which is significantly larger than the others. The Swerling III model is similar, but

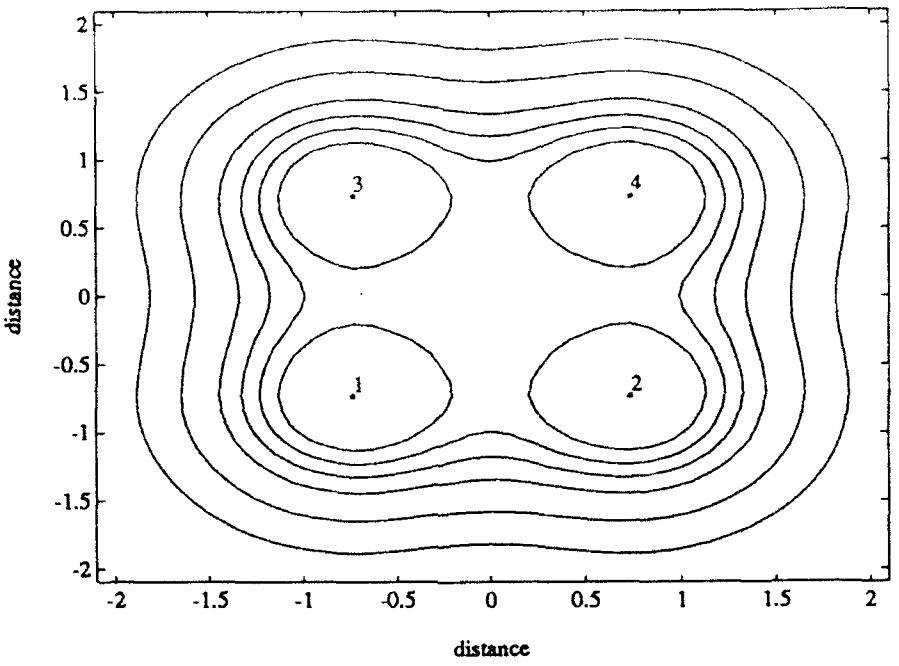


Figure 7: Probability of detection contours for four-sensor system with operation at "fusion range" (i.e. with sensors forming a square of side 1.5 approximately times the standard radius) and a false-alarm rate of 10^{-5}. The contours which surround the sensors enclose points with probability of detection at least 99%, and surrounding contours are those corresponding to probabilities of detection $\{.975, .95, .9., 75, .5\}$.

assumes that one of the scattering centers can be considered dominant. When coupled with a CA-CFAR processor the model becomes [11]:

$$\begin{aligned} \mathbf{H_0}: \quad \Pr(X_i \geq x_i) &\approx \frac{1}{(1+x_i)^m} \\ \mathbf{H_1}: \quad \Pr(X_i \geq x_i) &\approx \frac{1+2(m+1)x_i/(1+S_i)}{(1+\frac{2x_i}{1+S_i})^{(m+1)}} \end{aligned} \tag{7}$$

where as before m is the size of the reference window.

In Figure 9 we duplicate the situation of Figure 6 for the Swerling III case. (Note that both axes are in "relative" units as before, but equations (5) and (6) are modified in an obvious way to reflect the new model.) It is apparent that behavior under the two models is at least qualitatively the same, in that the co-location benefits of fusion are unimpressive, in that there is a "fusion range", and, naturally, in the disjoint-sensor coverage. A difference, however, is that the gains from fusion in the Swerling III case are less impressive.

Variable Aspect

A given complex target, observed from a number of closely-spaced aspects, exhibits radar cross-section fluctuations about a nominal average value, and these fluctuations are widely accepted generally to obey

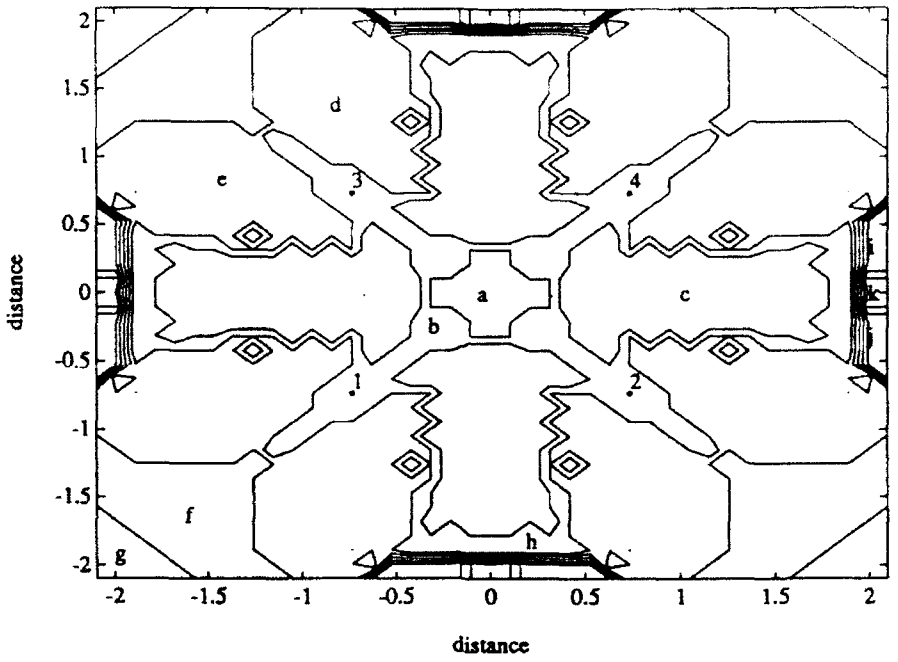


Figure 8: Optimal fusion rules for four-sensor system of previous figure. The asterisks and numbers indicate the sensors' positioning, and the lettered locations refer to the fusion rules as given in Table 1. This figure requires some explanation, for which refer to that table and to the text.

one of the Swerling models [5]. When such a target is illuminated from more than one aspect, however, it seems reasonable to expect not only these fluctuations but also different *nominal* cross-sections. Turning attention again to reference [5], these nominal cross-sections may differ by as much as 20 or $30dB$.

This behavior and its impact on fusion has not previously been discussed, as far as we are aware. The idea is this: given a single-sensor system a target is illuminated from one range of aspects at a time, and hence and hence the *gross* (or nominal) cross-sectional area is simply an SNR parameter, and fine fluctuation give rise to a Swerling model. In a multi-sensor system the various illumination aspects do not, in general, share even nominal cross-sections. For example, in a two-sensor system if one sensor is observing a target from broadside and another from a three-quarter angle (due to the sensors' physical separation), one might expect the former return to be a great deal larger.

There do not appear to be established models for such imbalanced gross cross-sectional areas. We shall thus feel free to use an *ad hoc* bimodal approach, and assume that the hypothesis testing problem is

$$\begin{aligned} \mathbf{H_0}: \quad \Pr(X_i \geq x_i) &= \frac{1}{(1+x_i)^m} \\ \mathbf{H_1}: \quad \Pr(X_i \geq x_i) &= (1-p)\frac{1}{(1+\frac{x_i}{1+S_i^{low}})^m} + p\frac{1}{(1+\frac{x_i}{1+S_i^{high}})^m} \end{aligned} \tag{8}$$

where the average SNR $(1-p)S_i^{low} + pS_i^{high}$ for the i^{th} sensor and a given location does not depend on

Location	Fusion Rule	Cardinality
a	*at least two of* $\{u_1, u_2, u_3, u_4\}$	11
b	$u_1 + (\textit{at least two of } \{u_2, u_3, u_4\})$	12
c	$u_2 + u_4 + u_1.u_3$	13
d	$u_3 + u_4.(u_1 + u_2)$	11
e	$u_3 + u_1.(u_2 + u_4)$	11
f	$u_1 + u_2.u_3$	10
g	$u_1 + u_2.u_3.u_4$	9
h	$u_1 + u_2$	12
i	$u_4.(u_1 + u_2 + u_3)$	7
j	$u_2.(u_1 + u_3 + u_4)$	7
k	$u_2.u_4 + u_1.u_3.(u_2 + u + 4)$	12

Table 1: Fusion rules for selected points of Figure 10. (The third column indicates the number of binary 4-vectors for which the fusion center decides for H_1; also, "+" denotes logical OR and "." denotes logical AND.)

S_i^{high}/S_i^{low}, which we shall vary. It can be shown that the likelihood ratio of sensor i is monotonically increasing in X_i, and hence we may use X_i as a sufficient statistic for generation of U_i, as opposed to resorting to a nonlinear function of X_i.

In Figure 10 we show the results in terms of ground coverage plotted against this ratio, for three sensors, $p = 0.5$, and for various values of $\rho \equiv S_{high}/S_{low}$. (As usual, in a multi-sensor system the power is divided evenly among the sensors. Both axes are in "relative" units expressible via modified versions of equations (5) and (6); however, for each value of ρ the definition of a standard radius and of a unit area of ground coverage is different, corresponding to the model.) The message from the plot is that a single-sensor cannot "see" the target at all when the aspect is unfavorable; in a multi-sensor system the probability is low that the aspect will be unfavorable from *all* viewpoints, and hence the performance is radically improved.

Heavy Clutter

Up to this point we have concerned ourselves with cell-averaging CFAR with a homogeneous background assumed. Often this is not appropriate, in that the background is more accurately epresented by a heavy-tailed distribution such as K or log-normal.

For computational reasons we are not in a position to examine such a case directly; we can, however, to a large extent mimic its behavior by specifying a poor CFAR estimate. To this end we have again used the model of equation (1), but this time with $m = 2$ (only two reference cells used for the CFAR estimate) and yielding an extremely heavy-tailed pair of density functions.

Figure 11 and Table 2 give the fusion rule for a four sensor system in which the sensors form a square of side .75 times the standard radius, with the fused false alarm rate being 10^{-5}. What is notable here is the tendency, as compared to Figure 8, for fewer 4-vectors to produce a target-present report.

A plot of coverage versus separation is shown in Figure 12. It is apparent that for this situation the

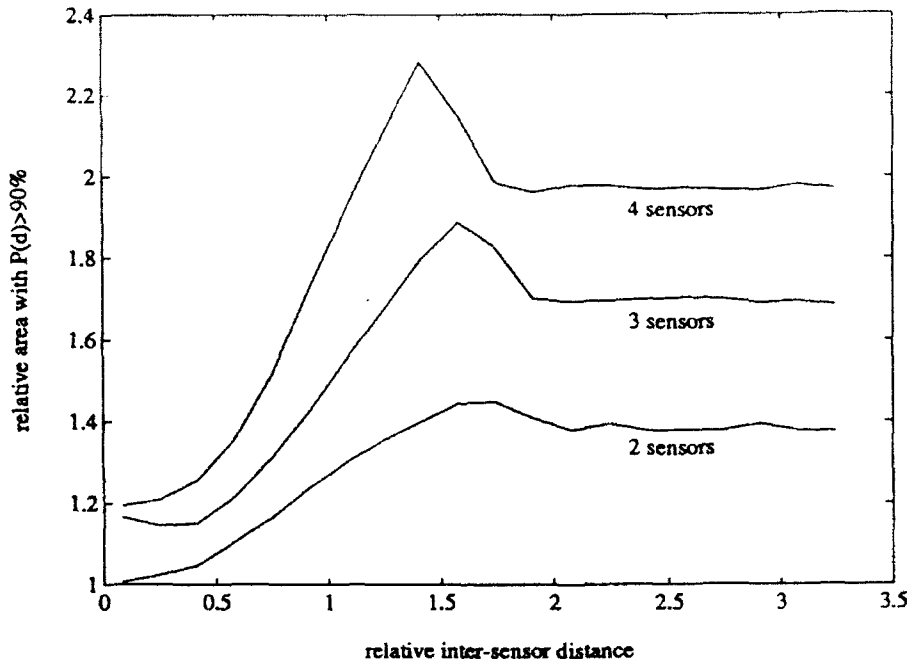


Figure 9: Ground coverage for multi-sensor system with probability of detection greater than 90% and probability of false alarm 10^{-5}, for a Swerling III target. Coverage is "relative" to a single-sensor system, and range is "relative" to the maximum range for a single-sensor system.

gains from like-sensor pre-detection fusion can be very significant. It is also interesting that, although a "fusion range" does indeed exist, it is much less striking than in most previous cases; in fact, the maximum benefit seems to occur for co-located sensors.

Summary

In this report we have tried to explore as impartially as possible the case for like- and independent-sensor fusion. To do this we have taken as our metric the "ground area covered" with a specified (α, β) performance, and in an attempt to avoid such meaningless statements as "two sensors outperform one" we have held the total transmitter power in the system constant, regardless of the number of sensors. Among our findings:

- Optimization of a multi-sensor binary-transmission is an involved process, with a considerable variation in local thresholding and in the fusion rule as a function of the location being tested and the sensor positioning.
- There is a benefit associated with making the sensors "disjoint"; that is, separated by sufficient distance that they do not in any meaningful way collaborate in the fused decision. The gain, subject of course to the aggregate power constraint and as compared to the single-sensor coverage,

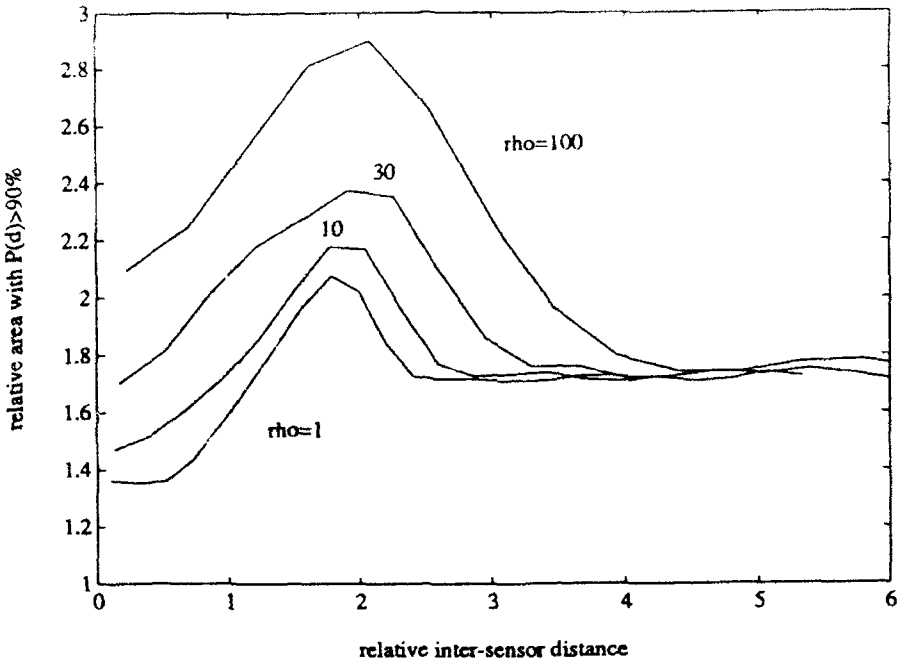


Figure 10: Ground coverage for three-sensor system with probability of detection greater than 90% and probability of false alarm 10^{-5}. The target fluctuates such that at each sensor the SNR is independently either S or ρS, each with probability one-half, and the average SNR is constant. Coverage is "relative" to a single-sensor system, and range is "relative" to the maximum range for a single-sensor system.

is given by the square-root of the number of sensors. The disjoint gain is a benchmark against which improvements of more-complicated systems should be compared.

- When the statistical hypothesis-testing model is unfriendly, then there may be considerable improvement from a co-located multi-sensor system. Conversely, if the model is well-behaved statistically (no heavy-tailed H_0-distributions or severe target fluctuations), then the co-located gain is minimal, and in some cases there is a degradation as compared to the single-sensor system.

- When the sensors are separated from one another by a "fusion range", then by collaboration they are able to detect targets no single sensor could observe were it in isolation. This fusion range produces performance superior to both disjoint and zero-separations, and may be considered fusion at its purest. The peak in coverage is most pronounced for well-behaved models, as in such cases the benefits of co-location are least impressive. The fusion range itself is a function both of the statistical model and of the physical sensor pattern.

- When an unfriendly environment's hostility is dominated by its heavy-tailed H_0-distribution (as, for example, with clutter), there is a tendency for AND-style rules to be optimal, by which we mean that relatively few combinations of sensor outputs result in a target-present decision at the fusion center. Conversely, when hostility is dominated by H_1 variability the tendency is towards

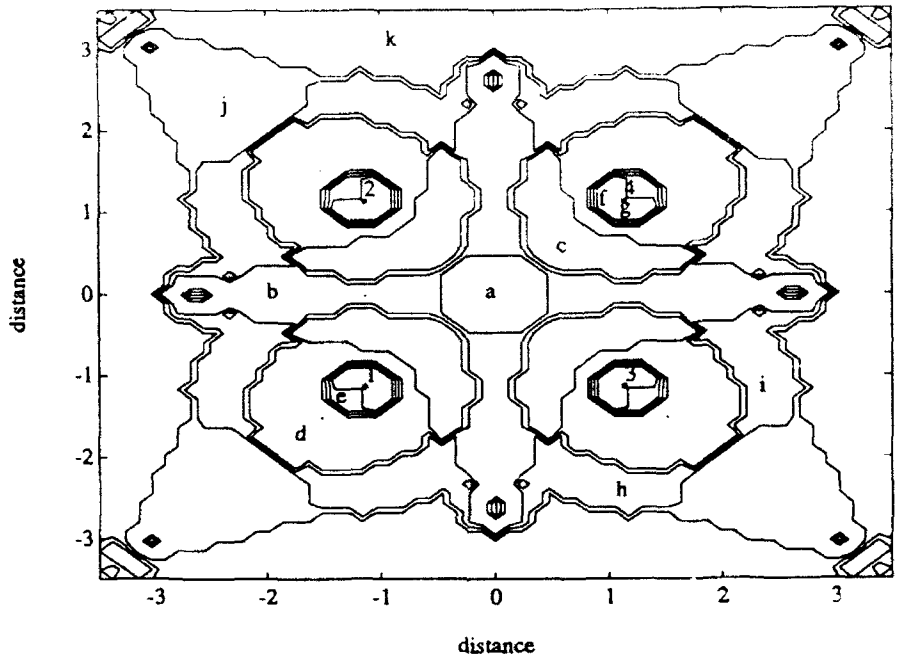


Figure 11: Optimal fusion rules for four-sensor system in heavy clutter. The asterisks and numbers indicate the sensors' positioning, and the lettered locations refer to the fusion rules as given in Table 2.

OR-style rules.

An additional factor is that in a multi-sensor system there will in general be fewer "blind spots" caused by obstacles to propagation, close proximity, and the like. Such factors are difficult to quantify, and beyond this note we make no further comment.

We have explored a number of examples: Swerling I, Swerling III, variable-aspect fluctuation, heavy-clutter, and at a variety of performance levels. While our models are certainly not exhaustive we believe they are at least representative. *Is there a case for like-sensor pre-detection fusion?* Our answer, reduced to a thumbnail: if the statistical environment is hostile there is a *very good case*; conversely, if the environment is well-behaved, then the case is considerably poorer.

In the course of this work a number of points have arisen which we believe worthy of further investigation.

1. Optimal sensor positioning appears to be open. The problem is not one of standard Euclidean distances, and may be made of further interest in that a propitious layout may involve sensors with varying powers.

2. When a given target is viewed from different angles its nominal radar cross-section will vary. In monostatic systems this is not a factor, but if pre-detection fusion is to be further studied some investigation of joint-fluctuation models would appear desirable.

Location	Fusion Rule	Cardinality
a	*at least three of* $\{u_1, u_2, u_3, u_4\}$	5
b	$u_1 u_2 + (u_1 + u_2) u_3 u_4$	6
c	$u_4 (u_1 + u_2 + u_3) + u_1 u_2 u_3$	8
d	$u_1 (u_2 + u_3 + u_4)$	7
e	$u_1 + u_2 u_3$	10
f	$u_4 + u_2 (u_1 + u_3)$	11
g	$u_4 + u_3 (u_1 + u_2)$	11
h	$u_3 (u_1 + u_2 u_4)$	5
i	$u_3 (u_1 u_2 + u_4$	5
j	$u_2 (at\ least\ two\ of\ \{u_1, u_3, u_4\})$	4
k	$u_2 u_4 (u_1 + u_3)$	3

Table 2: Fusion rules for selected points of Figure 15. (The third column indicates the number of binary 4-vectors for which the fusion center decides for H_1.)

3. There are a number of asymptotic results relating the performance of binary-transmission systems to those without quantization, and in these a common result is that there is a $1 - 3dB$ SNR loss from binarization. If there were a similar but more-general non-asymptotic result or worst-case bound, then one could in good conscience turn all attention to the simpler binary networks.

If one were to admit like but non-independent sensors to this study one would probably find even more significant benefits to fusion. As mentioned previously, CFAR processing and joint-fluctuation models may be appropriate candidates for such study. Also as mentioned previously, fusion of unlike sensor reports is probably more intuitively-appealing than that of like sensors. We have not studied these here due to their inherent attributes, such as angle-only or target signature, which go beyond detection performance, but a similar study at the tracking level would prove very interesting.

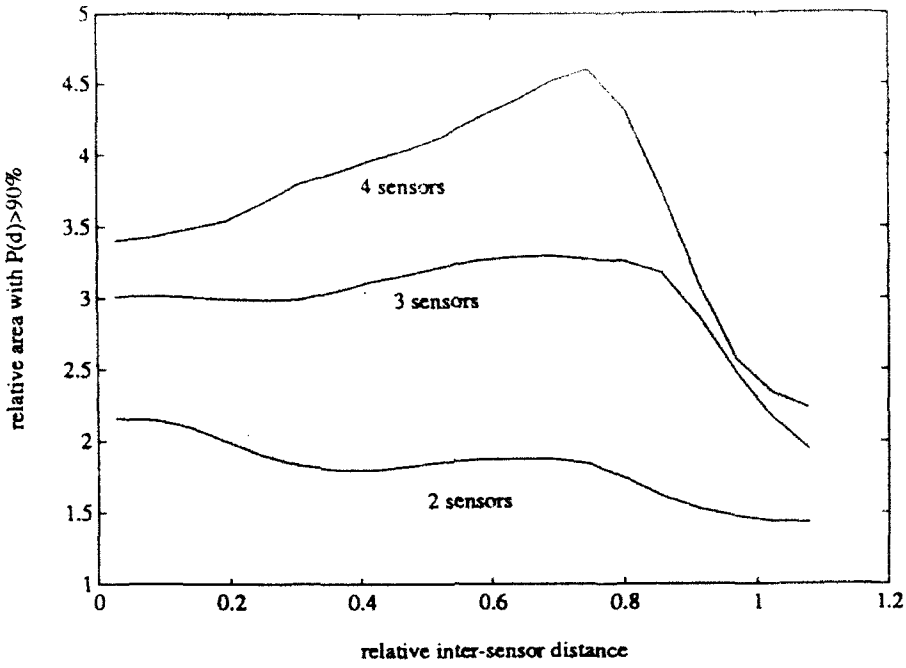


Figure 12: Ground coverage for multi-sensor system with probability of detection greater than 90% and probability of false alarm 10^{-5}, for high-clutter scenario. Coverage is "relative" to a single-sensor system, and range is "relative" to the maximum range for a single-sensor system.

References

[1] D. Warren and P. Willett, "Optimal Decentralized Detection for Conditionally-Independent Sensors", *Proceedings of the American Control Conference*, June 1989.

[2] J. Tsitsiklis, "Decentralized Detection", *Advances in Statistical Signal Processing, Vol 2 [Signal Detection]*, H.V. Poor and J.B. Thomas eds., JAI Press, 1991.

[3] B. Picinbono and P. Duvaut, "Optimum Quantization for Detection", IEEE Transactions on Communications, November 1988.

[4] Z. Chair and P. Varshney, "Optimal Data Fusion in Multiple Sensor Detection", *IEEE Transactions on Aerospace and Electronic Engineering*, January 1986.

[5] M. Skolnik, *Introduction to Radar Systems*, McGraw-Hill, 1980.

[6] P. Willett and D. Warren, "Decentralized Detection – When Are Identical Sensors Identical?", *Proceedings of the 1991 Conference on Information Sciences and Systems*, March 1991.

[7] A. Elias-Fusté, A. Broquetas-Ibars, L. Boutsikaris, J. Abshire, "CFAR Data Fusion Center with Inhomogeneous Receivers", *IEEE Transactions on Aerospace and Electronic Systems*, January 1992.

[8] P. Willett, M. Alford, and V. Vannicola, "The Case for Like-Sensor Pre-Detection Fusion", *submitted to IEEE Transactions on Aerospace and Electronic Engineering*, October 1992.

[9] S. Kassam, *Signal Detection in Non-Gaussian Noise*, Springer-Verlag, 1987.

[10] Benedetto, Biglieri, and Castellani, *Digital Transmission Theory*, McGraw-Hill, 1988.

[11] H. Finn and A. Johnson, "Adaptive Detection Mode with Threshold Control as a Function of Spatially-Sampled Clutter Level Estimates", *RCA Review*, Vol. 29, 1968.

[12] E. Starcewski, Private Communication, Rome Laboratory, Griffiss AFB, NY, August 1992.

[13] P. Willett and D. Warren, "The Suboptimality of Randomized Tests in Decentralized and Quantized Detection Systems", *IEEE Transactions on Information Theory*, March 1992.

[14] D. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, 1989.

[15] P. Willett, M. Alford, and V. Vannicola, "Pre-Detection Fusion with Similar Sensors", *submitted to the 1993 OE/Aerospace Engineering and Sensing Conference*, September 1992.