REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	11/28/95	Final 1/1/94 - 7/1/95

4. TITLE AND SUBTITLE

Final Report on AFOSR Project:
Stochastic Network Processes

5. FUNDING NUMBERS

F49620-94-1-0034

AFOSR-TR-96

0117

6. AUTHOR(S)

Richard F. Serfozo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205

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

AFOSR/NM
Bldg. 410
Bolling AFB, DC 20332-6446

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

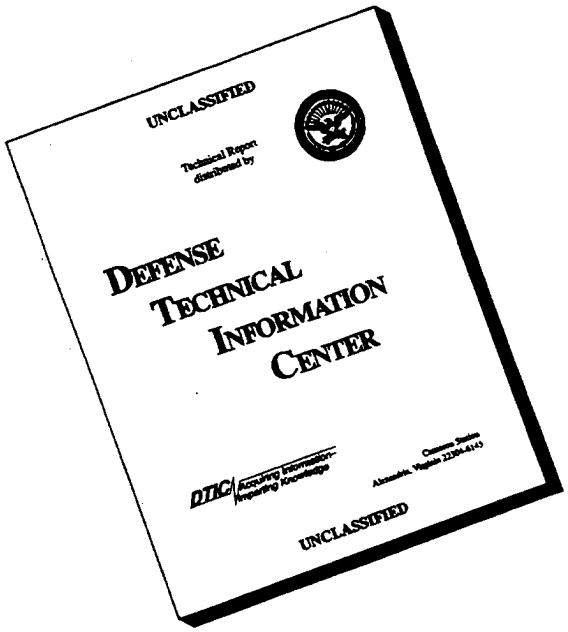
Please see next sheet.

19960320 036

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

This final report summarizes the publications from our research on stochastic network processes that represent the movement of discrete units in networks. Primary examples are the movement of parts and supplies in manufacturing plants and in distribution systems and the movement of data packets and telephone calls in computer and telecommunications networks. The distinguishing feature of our research was the emphasis on the next generation of *intelligent* networks that will be the backbone of our manufacturing and computer systems. In these networks, the processing of units at the nodes and the routing of units typically depend dynamically on the actual network congestion, and units move concurrently (e.g. batch processing). Most of the present theory of stochastic network processes is for unintelligent networks in which the nodes operate independently, the routes of units are independent, and the units move one-at-a-time. A recent focus of our network research was on parallel simulation, which is one of the most promising areas for the use of parallel or distributed processing. We developed stochastic network models for assessing the feasibility and quality of various protocols in parallel simulations and we developed algorithms that can be incorporated as subroutines in certain types of parallel processing simulations, such as queueing networks, in which the system evolution can be represented by recursive equations.

Final Technical Report on AFOSR Project 94-1-0034 Stochastic Network Processes

by
Richard F. Serfozo
Georgia Institute of Technology
November 28, 1995

The general theme of our research has been to develop stochastic network processes for modeling the movement of discrete units in networks. Primary examples are the movement of parts and supplies in manufacturing plants and in distribution systems and the movement of data packets and telephone calls in computer and telecommunications networks. The distinguishing feature of our research is the emphasis on the next generation of *intelligent* networks that will be the backbone of our manufacturing and computer systems. In these networks, the processing of units at the nodes and the routing of units typically depend dynamically on the actual network congestion, and units move concurrently (e.g. batch processing). Most of the present theory of stochastic network processes is for unintelligent networks in which the nodes operate independently, the routes of units are independent, and the units move one-at-a-time. Our goal is to provide an understanding of these more complex intelligent networks by describing their stochastic behavior.

Another focus of our network research is on parallel simulation, which is one of the most promising areas for the use of parallel or distributed processing. One task is to develop stochastic network models for assessing the feasibility and quality of various protocols in parallel simulations. Emphasis is on an optimistic or aggressive *Time Warp protocol* that periodically generates superfluous data that has to be rolled back or discarded. Another task is to develop algorithms that can be incorporated as subroutines in certain types of parallel processing simulations, such as queueing networks, in which the system evolution can be represented by recursive equations.

The following is a summary of the papers we have written for this project during the last two years.

1 Relating the Waiting Times and Queue lengths in Heavy Traffic Systems by W. Szczotka, W. Topolski and the PI. *Stoch. Processes Appl.* **52**, 119-134, 1994.

Diffusion models have been instrumental in analyzing queueing systems in heavy traffic. We show that there is more than the one conventional *reflected Brownian motion* model for such analyses — there are three natural models (including the conventional one) depending on the subtle nature of the heavy traffic. These are established via weak convergence limit theorems. Other major results show that, in heavy traffic, an arriving unit's waiting time before its service begins is approximately equal to the number of units in the system times the average service time (this is not true in light or moderate traffic). This property is useful for obtaining confidence intervals for queue lengths based on waiting time information or vice versa.

2 Campbell's Formula and Applications to Queueing by Schmidt,V. and the PI. *Frontiers in Queueing*, 225-242, 1995.

Campbell's formula for Palm probabilities is a basic tool for deriving properties of stationary queueing systems and stationary processes in general. This study shows that Campbell's formula is essentially Fubini's theorem for random measures — this clarifies many of its previous ad hoc applications and conjectures about its scope. We also give new insights into the versatility of Campbell's formula by establishing several equivalent versions of it. A few of these are known (e.g. the exchange formula and $H = \lambda G$). Also included are applications involving integrals with respect to random product-measures, waiting times in systems, rate conservation laws, sojourn times of processes, travel times in networks, ladder heights in risk processes and virtual delays in queueing systems.

3 Performance Limitations of Parallel Simulations by Chen, L. and the PI. Submitted for publication 1994.

This study shows how the performance of a parallel simulation may be affected by the structure of the system being simulated. We consider a wide class of "linearly synchronous" simulations consisting of asynchronous and synchronous parallel simulations (or other distributed-processing systems), with conservative or optimistic protocols, in which the differences in the virtual times of the logical processes being simulated in real time t are of the order $o(t)$ as t tends to infinity. Using a random time transformation

idea, we show how a simulation's processing rate in real time is related to the throughput rates in virtual time of the system being simulated. This relation is the basis for establishing upper bounds on simulation processing rates. The bounds for the rates are tight and are close to the actual rates as numerical experiments indicate. We use the bounds to determine the maximum number of processors that a simulation can effectively use. The bounds also give insight into efficient assignment of processors to the logical processes in a simulation.

4 Parallel Simulation by Multi-Instruction Longest Path Algorithms by Chen, L. and the PI. Submitted for publication, 1994.

This paper presents several basic algorithms for the parallel simulation of G/G/1 queueing systems and certain networks of such systems. The coverage includes systems subject to manufacturing or communication blocking, or to loss of customers due to capacity constraints. The key idea is that the customer departure times are represented by longest-path distances in directed graphs instead of by the usual recursive equations. This representation leads to algorithms with a high degree of parallelism that can be implemented on parallel computers with single or multiple instruction streams.

5 Little Laws for Waiting Times and Utility Processes by the PI. *Queueing Systems* **17**, 137-181, 1994.

A fundamental law that holds for many service systems is that the average time W that a unit waits in the system is directly proportional to the average queue length L of the system, i.e., $L = \lambda W$, where λ is the average arrival rate of units to the system. This law is known to be universal in the sense that it is true when all three terms exist. Establishing the existence of these limits for a new system, however, may pose a problem and this is where there are still open questions. This study gives a new, comprehensive and simpler aproach for proving Little laws (i.e. establishing the existence of the limits). This includes laws for general utility processes as well as for waiting times. Using this approach, we obtain fundamental Little laws for Markovian, regenerative and stationary systems that apply to large classes of systems; prior to this, there were only ad hoc results in this regard that were unfriendly to non-specialists.

6 Parallel-Processing Times: Extreme Values of Phase-type and Mixed Random Variables by Sungyeol Kang and the PI. Submitted for publication, 1995.

In computer or telecommunications systems, a typical concern is the time it takes for a group of data packets that will eventually constitute one message to be processed by a network of computers. This time is the maximum or *extreme value* of the travel times of the units through the network. A similar concern is the time to complete a job in a manufacturing (or computer) system consisting of many tasks performed in parallel by a network of workstations. An example is that 20 units must complete a PERT network in parallel before they are brought together as one system. A basic problem is to determine the distribution of the time to complete a large number of tasks in parallel. That is, knowing the probability structure of the individual tasks, what is the asymptotic distribution of the maximum of the task times as the number of tasks tends to infinity?

We address this parallel-processing (or extreme-value) problem for the following settings: (a) A task consists of performing a set of randomly selected subtasks in series and the subtask durations have Erlang distributions. (b) The task times are independent, identically distributed phase-type random variables (e.g. completion times of Markovian PERT networks or task-graphs). (c) The tasks are dependent and their distributions are selected by a random environment process such as a Markovian or stationary selection process.

We show that the distributions of the task completion times in these settings are the classical extreme-value distributions, even though the task times are dependent and may involve a series of dependent subtasks. Included are three fundamental results that apply to general extreme-value contexts:
(1) All Phase-type distributions are in the domain of attraction of the Gumbel distribution.
(2) Extreme values of Markovian- or stationary-selected variables are determined by certain states of the selection process whose associated distributions have a "dominant tail".
(3) Extreme values of sums of random variables are determined by parts of the sum that dominate the other parts.

7 Partition-Balanced Markov Processes by Alexopoulos, C., El-Tannir, A. and the PI. Submitted for publication, 1995.

When can the stationary distribution of a Markov process be obtained by pasting together several stationary distributions that represent the process restricted to certain subspaces? This study describes a class of "partition-balanced" Markov process that have this cut-and-paste or divide-and-conquer property. The importance of this property is that the problem of obtaining a stationary distribution on a large space (e.g. for networks) reduces to finding several stationary distributions on smaller subspaces, possibly even by simulations.

The notion of partition-balance is a "macro-reversibility" property resembling the detailed balance property of reversible processes. We present several characterizations of partition-balance and identify subclasses of tree-like, starlike and circular partition-balanced processes. A new "circular birth-death process" is also presented as a vehicle for the analysis. The results are illustrated by a queueing model with controlled service rate, a multi-type service system with blocking and a parallel-processing model. A few comments address partition-balance for non-Markovian processes.

8 Performance of Time Warp Parallel Simulations of Queueing Networks by Chen, L., Das, S., Fujimoto, R. and the PI. Submitted for publication, April 1995.

A new approach for predicting the performance of Time Warp parallel discrete event simulations of queueing networks is presented. Time Warp performance is estimated by constructing and analyzing a second queueing network model that emulates the simulation. Comparisons of predicted performance with measurements from an actual operational system indicate that the model yields accurate perormance predictions.

9 M/M/1 Queueing Decision Processes with Monotone Hysteretic Optimal Polidcies by Kitaev, M. and the PI. Submitted for publication, June 1995.

This study gives further insights into the $M/M/1$ queueing control model developed by Lu and Serfozo (1984). In the model, the rates of the exponential interarrival and service times are controlled and it is established that there exists a monotone optimal policy that increases the service rate and decreases the arrival rate as the queue increases. Also, there are costs of

changing the rates creating a tendency to continue using the rates for longer periods (a retardation or hysteretic effect). This revisit of the model establishes the existence of a monotone hysteretic optimal policy under more general conditions, and it corrects the original proof for the case of average costs.

10 Queueing Networks With String Transitions by Yang, Bingyi. PhD Dissertation, Georgia Institute of Technology, August 1995.

The major contribution of this study is the development of a new Markovian queueing network model whose transitions are determined by a string of information. This string is a vehicle for incorporating a series of activities in the network instantaneously that are involved in the network change. Examples include sending signals or negative customers to several nodes to delete batches of customers, concurrent movements of auxiliary resources, batch arrivals and services, and catastrophes eliminating all customers at a node. Essentially all the existing Markovian queueing network models with tractable equilibrium distributions, including the classical Jackson model and recent models of batch movements and negative customers, are special cases of this new model. The equilibrium distribution and several other properties of the model are derived. Two characteristics of the model are non-standard traffic equations and a new type of partial balance property for the equilibrium distribution.