GENERAL QUEUEING NETWORKS WITH PRIORITIES

Maximum entropy analysis of general queueing network models

with priority preemptive resume or head-of-line and

non-priority based service disciplines

Nasreddine TABET AOUEL

Submitted for the degree

of Doctor of Philosophy

Department of Computing

University of Bradford

(1989)

N. TABET AOUEL

GENERAL QUEUEING NETWORKS WITH PRIORITIES

ABSTACT

Keywords:

Scheduling discipline, priority class, maximum entropy, preemptive resume, head-of-line, generalised exponential, queueing network.

Abstract

Priority based scheduling disciplines are widely used by existing computer operating systems. However, the mathematical analysis and modelling of these systems present great difficulties since priority schedulling is not compatible with exact product form solutions of queueing network models (QNM's). It is therefore, necessary to employ credible approximate techniques for solving QNM's with priority classes.

The principle of maximum entropy (ME) is a method of inference for estimating a probability distribution given prior information in the form of expected values. This principle is applied, based on marginal utilisation, mean queue length and idle state probability constraints, to characterise new product-form approximations for general open and closed QNM's with priority (preemptive-resume, non-preemtive head-of-line) and non-priority (first-come-first-served, processor-sharing, last-come-first-served with or without preemtion) servers. The ME solutions are interpreted in terms of a decomposition of the original network into individual stable G/G/1 queueing stations with assumed renewal arrival processes. These solutions are implemented by making use of the generalised exponential (GE) distributional model to approximate the interarrival-time and service-time distributions in the network. As a consequence the ME queue length distribution of the stable GE/GE/1 priority queue, subject to mean value constraints obtained via classical queueing theory on bulk queues, is used as a 'building block' together with corresponding universal approximate flow formulae for the analysis of general QNM's with priorities. The credibility of the ME method is demonstrated with illustrative numerical examples and favourable comparisons against exact, simulation and other approximate methods are made.

ACKNOWLEDGMENTS

This thesis is dedicated to my parents, whose love and encouragment offered me the opportunity to complete my studies; to my fiancee, Samira, whom I deeply thank for her moral support during the demanding years of the research.

I owe deep gratitude to my research supervisor Dr.D.D. Kouvatsos for his invaluable guidance and inspiration throughout my period of study.

I also wish to extend my thanks to all my friends and relatives, and especially my brother Abdelghani for his great support in Algeria.

Finally I wish to thank the ministry of high education of the Algerian government for the financial support.

DECLARATION

Part of the contribution of this thesis has been presented in conferences and a journal and submitted and/or accepted for publication (see [KOUV,87; 88b; 88c; 88d]).

TABLE OF CONTENTS

GLOSSARY OF NOTATION

$A(.)$: Probability distribution function (PDF) of the interarrival time process.

$A_1(.)$: PDF of the interval of time between the instant the arrival process is switched on and the occurence of the first arrival.

$A^*(.)$, $A_1^*(.)$: are the corresponding L.S.T of $A(.)$ and $A_1(.)$, respectively.

$\tilde{B}$: the bulk size of the arriving jobs in an ordinary GE/G/1 queue.

b_n : $Prob[\tilde{B} = n]$

b_n^i : i-fold convolution of b_n with itself.

$<b_r>$: Mean bulk size of class-r jobs.

β_0: Lagrangian multiplier corresponding to the normalisation constraint.

β_ℓ: Lagrangian multiplier corresponding to the ℓ^{th} constraint.

$\tilde{C}_r$: Completion time of class-r jobs in PR or HOL G/G/1 queue.

$C_r^*(.)$: L.S.T of the the completion time $\tilde{C}_r$.

$C_r^{*(b)}$: L.S.T of the completion time of all members af an arriving bulk in GE/G/1 HOL or PR queue.

C_{ar}^2: Squared coefficient of variation of the interarrival-time of class-r jobs.

C_{sr}^2: Squared coefficient of variation of the service-time of class-r jobs.

$\hat{C}_{sr}^2$: Squared coefficient of variation of the effective service-time of class-r jobs

C_{dr}^2: Squared coefficient of variation of the interdeparture-time of class-r jobs.

$D^*(.)$: L.S.T of the interdeparture-time in ordinary G/G/1 queue.

$f^{(d)}$: Departure formula for the squared coefficient of variation.

$f^{(m)}$: Merging formula.

$f^{(q)}$: Mean queue length formula.

$f^{(s)}$: Splitting formula.

$f_i(\underline{n}_i)$: Unnormalized ME solution of centre-i having $\underline{n}_i=(n_{i1},\ldots,n_{iR})$.

$\tilde{G}_r$: Busy period generated by jobs belonging to classes $\{1,2,\ldots,r\}$.

$G_r^*(.)$: L.S.T of the busy period $\tilde{G}_r$.

$\langle G_r \rangle$: Mean busy period $(\tilde{G}_r)$.

$\hat{G}_r^*(.)$: L.S.T of the remaining busy period $\tilde{G}_r$.

g_r: Lagrangian coefficient corresponding to the utilisation constraint relative to class-r jobs.

g_{ir}: Lagrangian coefficient corresponding to the utilisation constraint relative to class-r jobs at centre-i.

g_{0ir}: Flow-balance correction factor of class-r jobs at centre-i in genereal closed queueing network.

γ_r: Utilisation of the server (G/G/1 priority queue) with respect to jobs belonging to classes s, $s\epsilon\{1,2,\ldots,r\}$.

κ: Tuning parameter of the Hyperexponential (H_2) distributional model

λ_r: Mean arrival rate of class-r job in a multiple-class G/G/1 queue.

$\lambda_r^{(b)} = \lambda_r\sigma_r$: Mean bulk arrival rate of class-r jobs.

$\Lambda_r^{(b)} = \sum_{\ell=1}^{r} \lambda_\ell^{(b)}$: Mean bulk arrival rate of job classes $\{1,2,\ldots,r\}$.

λ_{ir}: Throughput of class-r at centre-i.

λ_{ir}^*: throughput of class-r at centre-i in pseudo-open network.

M: Number of centres in a general QNM.

N_r: Number of jobs of class-r in closed QNM.

$\underline{N}_R$: Population vector in closed QNM.

$\langle n \rangle$: Mean queue length of a single-class G/G/1 queue.

$\langle n_r \rangle$: Mean queue length of class-r jobs in a multiple class G/G/1 queue.

$\langle n_{ir} \rangle$: Mean queue length of class-r jobs at centre-i in a general QNM.

$\langle n_{ir} \rangle [\underline{N}_R]$: Mean queue length of class-r at centre-i in a closed QNM with $\underline{N}_R$ jobs contained in the network.

$\underline{n} = (n_1, n_2, \ldots, n_R)$: Population vector in a G/G/1 queue.

n_{ir}: Number of jobs of class-r at centre-i.

$\underline{n}_i = (n_{i1}, n_{i2}, \ldots, n_{iR})$: Population vector at centre-i.

$\underline{\underline{n}} = (\underline{n}_1, \underline{n}_2, \ldots, \underline{n}_M)$: Population vector in the network.

$\{P_{ir;js}\}$, $\{P_{irj}\}$: Routing frequencies.

$P(\underline{n})$: Joint probability to have $\underline{n}$ jobs in a multiple-class G/G/1 queue.

$P(\underline{S})$: Probability to be in state $\underline{S}$.

$P(\underline{n})$: Joint steady-state probability to have $\underline{n}$ jobs in closed QNM.

$P_i(\underline{n}_i)$: Marginal probability to have $\underline{n}_i$ jobs at centre-i.

$P_{ir}(\underline{n}_{ir})$: Marginal probability to have n_{ir} jobs of class-r at centre-i

.

$P_r^{(a)}$. Probability that an arriver from class-r sees n_r jobs of its own class in the queue.

$P_r^{(d)}$: Probability that a departer from class-r leaves n_r jobs of its own class in the queue.

$Q_r[\,.\,]$: Generating function of the queue length distribution of class-r jobs in a G/G/1 priority queue.

$q_r(.)$: Generating function of the bulk size distribution of class-r jobs in GE/G/1 priority queue.

R: Number of classes in the queueing system.

ρ_r: Utilisation of class-r jobs in a G/G/1 queue

$\rho = \sum_{r=1}^{R} \rho_r$: Overall utilisation in a G/G/1 queue.

ρ_{ir} : Utilisation of class-r jobs at centre-i.

ρ_i : Overall utilisation of centre-i.

ρ^*_{ir}: Utilisation of class-r jobs at centre-i in the pseudo-open network.

ρ^*_i: Overall utilisation of centre-i in the pseudo-open network.

$\hat{\rho}_r$: Utilisation of the virtual server-r.

$\tilde{S}_r$: Service-time of class-r jobs.

$S_r^*(.)$: L.S.T of the service time of class-r jobs

$<S_r>$: Mean service time of class-r jobs.

$\hat{S}_r$: effective service time of class-r jobs.

$\hat{S}^*_r(.)$: L.S.T of the effective service time of class-r jobs.

$S^{*(b)}_r(.)$: L.S.T of service time of all members of an arriving bulk of class-r jobs in GE/G/1 priority queue.

$\sigma_r = 2/(C^2_{ar}+1)$: Parameter of the bulk size distribution of class-r arrival process.

$\tilde{T}_r$: Response time of class-r jobs in G/G/1 queue.

$T^*_r(.)$: L.S.T of reponse time of class-r jobs.

$<Ts_r>$: Mean system response time.

$\tau_r = 2/(C^2_{sr}+1)$: Parameter of the bulk size distribution of class-r service process.

U(t): Unfinished work at time t in G/G/1 queue.

<U> : Mean unfinished work.

$W_r(t)$: Occupation time of the server with respect to class-r jobs at centre-i or the waiting time of class-r jobs in G/G/1 PR or HOL queue.

$W_r^*(.)$: L.S.T of the waiting time of class-r jobs

$\langle W_r \rangle$: Mean waiting time of class-r jobs.

x : Lagrangian coefficient corresponding to the mean queue length constraint in a single-class G/G/1 queue.

x_r : Lagrangian coefficient corresponding to the mean queue length constraint of class-r.

y_r: Lagrangian coefficient corresponding to the idle state probability constraint of class-r.

Z : Normalising constant.

$Z[\underline{N}_R]$: Normalising constant in closed network with $\underline{N}_R$ jobs in the network.

$z[\underline{n}_R, m]$: Normalising constant in closed network with a population vector, $\underline{n}_R$ and m centres.

$z^i[\underline{n}_R, M]$: Auxiliary function, is the normalising constant for closed networks after the removal of centre-i containing $\underline{N}_R - \underline{n}_R$ jobs.

CHAPTER 1

INTRODUCTION

1.1 Overview

Computer systems and communication networks generally consist of set of resources and set of tasks, jobs, or messages, competing for and accessing those resources. In general there are multiple classes of jobs competing for a limited number of resources and inevitably congestion may occur and queues are formed. The quantitative evaluation and performance modelling of such systems is essential for design, development, tuning and configuration purposes. Throughput, utilisation and response time are typical indicators of system performance.

The performance evaluation methods are grouped into three categories, namely measurement or benchmarking, simulation modelling and analytical modelling.

The benchmarking technique involves experimentation on the system under investigation and consequently performance measures can be obtained only when a system has been built, incorporating appropriate instrumentation and is running. This technique is most of the time excessively expensive and although it gives accurate knowledge of the system under specific workloads, it does not provide any insight that would allow generalisation.

The simulation modelling is a statistical description of the behaviour of the real system. It has extremely broad applicability and have been used extensively in system performance evaluation. It can achieve any degree of accuracy required in reflecting the detailed structure of the system. Nevertheless, it becomes much more

costly in terms of computer time as more details are incorporated.

Analytical modelling is a mathematical representation of the system under study. The performance measures of interest are evaluated by solving a set of equations.

A proper model to any system must be validated, projected and finally verified before being used. In the validation phase the real system is identified and parameterised. The appropriate model is then chosen and solved by comparing its outputs to the performance metrics collected from measurements. To assess the performance of the system for eventual modification, for instance upgrading the central processing unit (CPU) or adding more input/output (I/O) devices, the model inputs are modified and projected into the model. Finally, the system workload is adjusted to the new model inputs and comparison between performance measures collected and predicted takes place. Any substantial discrepancy between these quantities results from a misrepresentation of the system characteristics.

Analytic models provide an insight into the key factors affecting the performance of a real or a proposed system and determine performance sensitivity to parameter changes. A robust model must be able to accurately predict the behaviour of an actual system for tuning and capacity planning purposes. Moreover, such models can provide guidance into the overall design of a new system and it can also be useful in the development of more complex computer and communication architectures.

1.2 Queueing network modelling

Queueing network models (QNM's) have become widely accepted as powerful tools for estimating the performance of computer systems and communication networks and optimising their performance. The

performance analysis of QNM's is generally cost effective since it is based on efficient methods of solving mathematical equations. However, in order for these equations to have a tractable solution, certain simplifying assumptions must be made regarding the structure and the behaviour of the QNM. As a result QNM's cannot represent all the details that can be built into simulation models. Nevertheless, QNM's can generally capture most of the important system behaviour in order to make predictions with reasonable accuracy.

1.2.1 Description of QNM

QNM is a collection of queueing stations containing one or more servers arranged in the same disposition as in the real system. Each QNM is characterised mainly by its topology which describes the configuration of the system and the system workload which consists of multiple components, each of which identifies a customer (or job) class. For example, time sharing, batch and transaction processing are three types of workload commonly identified in a multiprogramming environment.

Each customer class defines a specific population which can be open when the number of customers of that class is not limited, or closed if the population size is bounded.

Figure 1.1 shows an example of open QNM consisting of a single class of customers coming from an external source, with mean arrival rate λ and C_a as coefficient of variation[1] of the external interarrival time distribution, P_{ij} is the probability that customer leaving centre-i is directed to centre-j. Moreover, figure 1.2 depicts an example of closed QNM known as the central server model with $1/\mu_i$ and C_{si} as the mean and coefficient of variation of the service time distribution of centre-i, respectively. K terminal users

are sharing the resources. Each user may be either in an operative state (a transaction is waiting or receiving service at the CPU, or at the I/O's devices) or in a think state (the user is in the editing mode). Under heavy traffic, the level of multiprogramming or the number of users in the operative state, N (with $N \leqslant K$) is generally kept fixed due to the finite capacity of the main memory, therefore transactions may have to queue (backlog of jobs) before accessing the main memory. Transactions in central server model alternate sequences of CPU and I/O processing intervals to complete their service demand.

A QNM containing both open and closed classes of customers is said to be mixed. The customer classes are characterised by their workload intensity which consists of either external arrival processes or population sizes depending if the classes are open or closed, respectively and, at every station, by the probability distribution function (PDF) of the service time required.

Each station has an associated queue in which customers may wait prior to receiving service. Thus, a queueing discipline is required to determine the order in which arriving customers receive service.

1: The squared coefficient of variation of a random variable X, Cv_X^2 is defined as i.e.,

$$Cv_X^2 = \frac{Var[X]}{E[X]^2} .$$

Where E[X] and Var[X] are the mean and the variance of X, respectively.

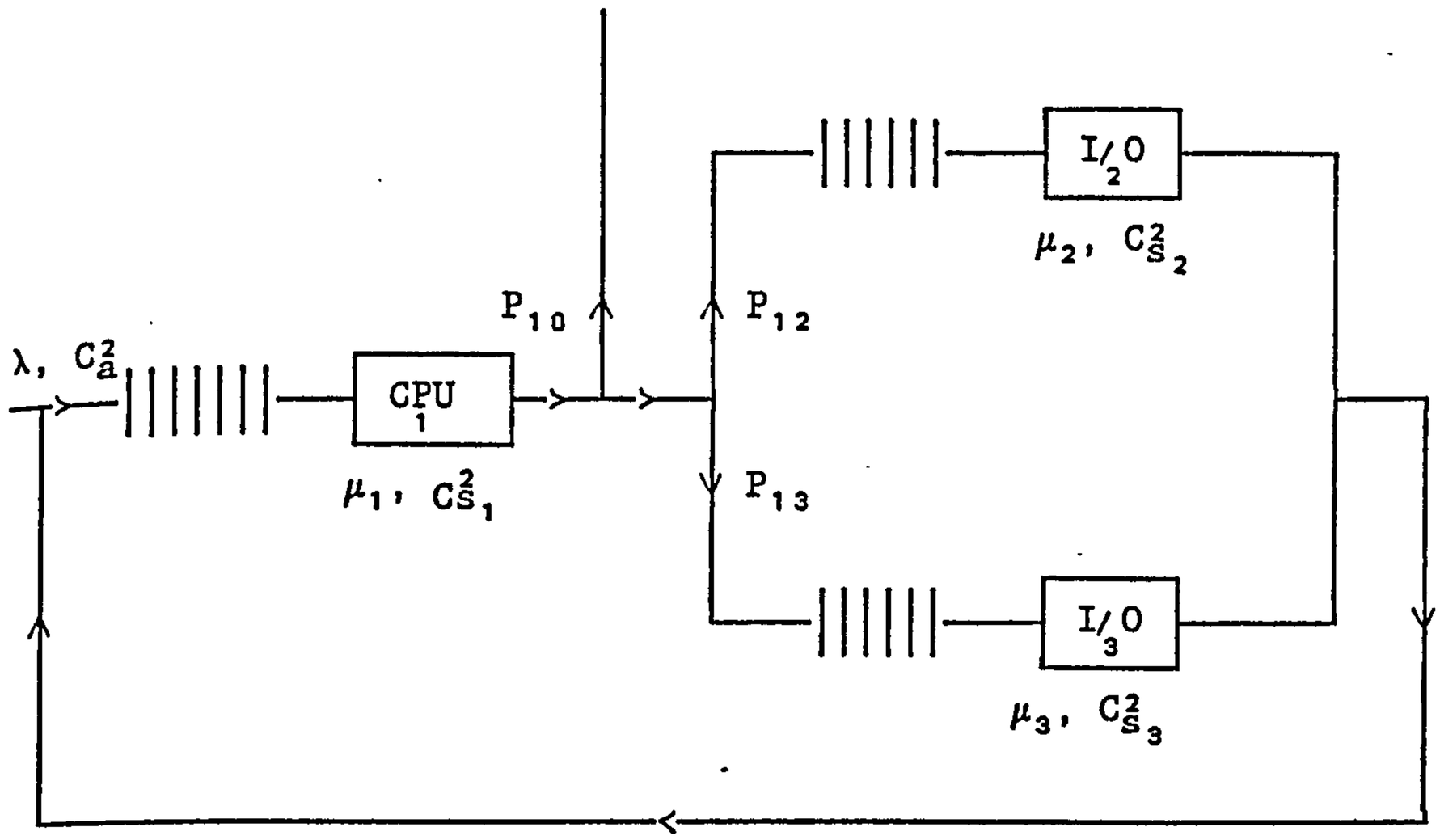

Figure 1.1 General open QNM with single class of customers.

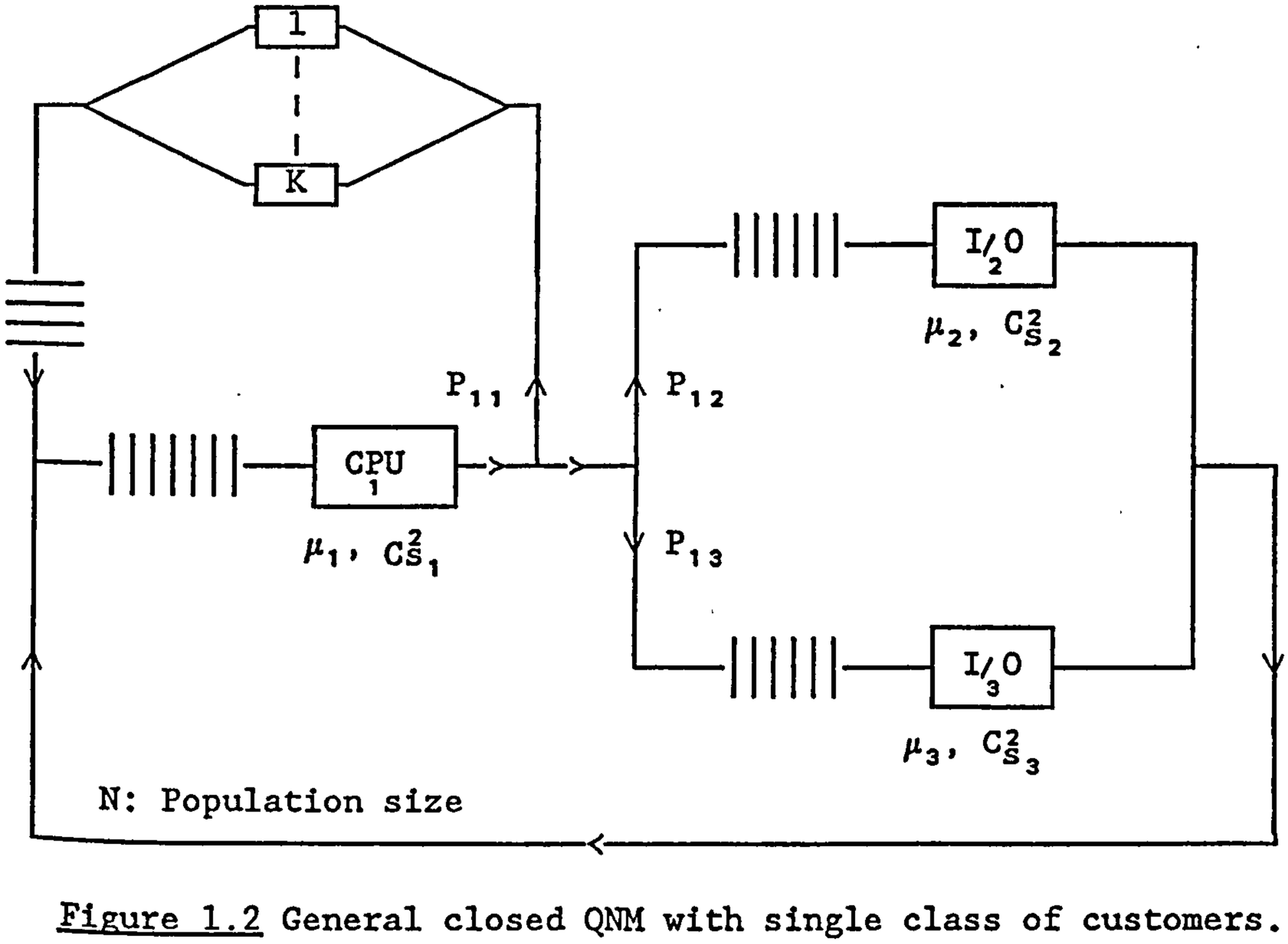

Figure 1.2 General closed QNM with single class of customers.

1.2.2 Queueing disciplines (or scheduling disciplines)

The most popular queueing discipline used is the *first-come-first-served (FCFS)* rule where customers are served in the order of their arrival instant. This discipline has been explored widely for various types of queueing systems involving general interarrival or service time distribution, single or multiple classes, waiting room with finite capacity and finite or infinite population. Several theoretical as well as practical results have been produced in the literature [ALLE,78; KLEI,75,76; GROS,85].

Under *last-come-first-served preemptive resume (LCFS-PR)* rule customers are served upon their arrival at the station and they are taken out of the service by a newly arriving customer. After each completion of service, the preempted customer resumes service from the point where he was interrupted.

Under the round robin (RR) queueing discipline a customer is given continuous service for a maximum interval of time known as a quantum. If the customer's service requirement is not satisfied during the quantum, the customer reenters the queue and waits to receive additional quantum of service, repeating this process until its service requirement is satisfied. Customers in the queue are served in FCFS fashion. Kleinrock [KLEI,64 a] defined the processor sharing (PS) discipline as the limiting case of the round robin when the quantum of time schrinks to zero. Under PS discipline, all customers share equally the capacity of the server on the full time basis.

The *infinite server (IS)* queueing discipline is used when the number of servers of a station is always at least the maximum number of customers at the station. Thus no customer will ever experience a queueing delay waiting for a server. This type of discipline is

typically used to represent user terminals in a time-sharing model.

A queueing discipline in which customers are assigned fixed priorities which determine the order in which they are served is called *a priority queueing discipline*. The next customer to be served is one that has the highest priority among all customers waiting to be served. Two main priority queueing disciplines are used in the existing computer systems or communication networks. If a service, once begun, is not interruptable, the queueing discipline is called priority non-preemptive or head-of-line (HOL) discipline. For example, in some communication network systems, control messages are given transmission priority over user messages; that is, the next message to be transmitted over a line is a control message if there are any messages waiting. Under a HOL discipline, once a message transmission begins it continues to completion without interruption. Control messages are transmitted in the order of their arrival as are user messages. If, on the other hand, an arriving customer interrupts the service of a lower priority customer and begins service or a customer whose service was interrupted resumes service at the point of interruption when there are no higher priority customers to be served, the queueing discipline is called priority preemptive resume (PR) discipline. For example, in some operating systems, system programs are given priority over user programs for execution on the processor. Furthermore, system programs interrupt executing user programs. An interrupted user program later resumes execution at the point of interruption.

Several other scheduling disciplines have been studied in the existing literature, some of them are analysed in [CONW,67].

1.2.3 History

Queueing theory has always played a major role in the elaboration and development of more realistic QNM's. Queueing theory models are considered as very efficient tools for the analysis of computer systems. Unfortunately very little has been done analytically in general queueing networks. This is due to the mathematical complexity involved in the general cases. Most of the queueing theory models are based on *exponential*[1] assumptions (service or interarrival time conforms to an exponential PDF).

One of the first important queueing theory models is the one presented by Scherr [SCHE,67]. He used the "machine repairman model" which had been established years before, to analyse the MIT compatible time shared system (CTSS). Scherr estimated the mean response time with a reasonable accuracy even though the CTSS violated most of his assumptions.

The first exact results in queueing network occured with the work of Jackson [JACK,57] in the analysis of open networks and followed by Gordon and Newell [GORD,67] in their investigation of closed networks. They showed the product form solution of the steady state probabilities in Markovian (or exponential) queueing networks involving single class of customers and FCFS servers. They concluded that each station behaves as it is isolated from the rest of the network despite that the arrival process is not renewal[2].

1: A random variable X conforms to an exponential distribution with parametre $\lambda > 0$ if its PDF is of the form $F(t) = 1-\exp(-\lambda t)$, $t \geq 0$.
2: Random variables which are identical and independent form a renewal process.

Few years later, Basket et al [BASK,75] extended the above results to the case of multiple classes and mixed networks with the following queueing disciplines:

a- FCFS discipline with all the classes having the same exponential service time distribution. The service rate of this station can be load dependent (service rate depends only on the number of customers in the station and not on the total number in the system),

b- PS queueing discipline,

c- LCFS-PR queueing discipline,

d- IS queueing discipline.

Note that under PS, LCFS-PR, IS queueing disciplines, each class of customers may have distinct service time distribution.

These types of QNM's are known as product form (PF) or separable QNM's, their corresponding performance measures such as server utilisation which is the proportion of time the server is busy, the throughput or the mean response time are obtained efficently by fast available computational algorithms.

1.2.4 Solving closed PF-QNM's

Closed PF-QNM's have been proven to be adequate models for wide range of computer systems. They are relatively efficient and easy to implement. Two major algorithms exist for the evaluation of the performance measures.

a- *The convolution algorithm:* this algorithm was developed by Buzen [BUZE,73] for closed networks, with single class of customers, and extended later by Bruel and Balbo [BRUE,80] for various specifications involving multiple classes with or without class switching and load dependent servers. It consists first, of computing

the normalisation constant $Z(\underline{N})$[1] by convoluting arrays according to recursion expressions and then obtaining the performance measures directly from the normalisation constant.

b- *The mean value analysis (MVA) algorithm*: This algorithm was developed by Reiser and Lavenberg [REIS,80], it evaluates the performance metrics directly without explicit computation of the normalisation constant $Z[\underline{N}]$. The MVA algorithm is slightly more computationally efficient and more numerically stable than the corresponding convolution algorithm [BRUE,80]. The MVA also provides a basis for approximation for either large PF-QNM's or nonproduct form (NONPF) networks.

The MVA is based on two simple principles:

a/ Little's formula $L = \lambda T$ which is a general applicable theorem relating the mean queue length L to the throughput λ and the mean response time T.

b/ The "arrival theorem" which states that in stationary PF-QNM, the state distribution that a customer sees upon arrival to a service centre is equal to the steady state distribution of the network with that customer removed [REIS,80].

The primary consequence is that the mean response time of customers of class-r at station-i satisfies the relation

$$T_{ir} = \langle S_{ir} \rangle (1 + A_{ir}) \qquad (1.1)$$

where $\langle S_{ir} \rangle$ is the mean service time of class-r customers at station-i and A_{ir} is the arrival instant queue length at centre-i seen by an arriving class-r customer.

1: $\underline{N} = (N_1, N_2, \ldots, N_R)$ is a vector population with R number of classes.

1.2.5 Solving open PF-QNM's

Open PF-QNM's are solved easily by the standard MVA algorithm where the queue length seen upon arrival at centre-i by a class-r customer, A_{ir}, is equal to the time averaged mean queue length L_i. Therefore using this fact together with Little's law, the mean response time formula (eq 1.1.) becomes:

$$Tir = \frac{<S_{ir}>}{1 - \rho_i} \qquad (1.2)$$

where ρ_i is the overall utilisation of station-i.

1.2.6 Solving NONPF-QNM's

Although the class of PF-QNM's has proven quite useful, there are many important features which, when incorporating into a model, lead to queueing networks violating the PF assumptions, such as:

a- General service time distribution at FCFS servers,

b- blocking or loss due to finite buffer capacity,

c- simultaneous resource possession of two or more resources by a customer,

d- allocation of resources to different customer classes according to priority based disciplines.

NONPF-QNM's are basicaly the most suitable models for computer systems and communication networks [GELE,80]. In principle, the long run (or the equilibrium) distribution of the state process of a network of queues may be obtained from a set of conditions which specify the equilibrium of the state process. These conditions are called the global balance conditions and take the form of a system of

linear equations in the asymptotic state probabilities. Various methods for numerical solution of the global balance equations have been studied. However, the size of the state space of a network state process combinatorially increases with the number of stations, customers, and customer classes. Applying numerical methods to solve the global balance equations is feasible only for networks with relative small sizes. As a consequence, approximation techniques are widely used in the literature.

Nearly all the approximate techniques exhibit one of the three basic approaches:

a/ decomposition:

The original network is decomposed into subnetworks which are solved in isolation by different techniques. Each subnetwork is then replaced by a single composite centre which mimics its behaviour. The aim of this technique is to break down the problem into subproblems that are easier to comprehend and analyse.

An early approach to solve complex models of computer systems is based on the concept of near-decomposability which was used first in econometrics [SIMO,61] and applied to queueing systems by Courtois [COUR,77]. The technique is based on variable aggregation involving a partition of the state space of the system, so that transitions between the resulting groups of states are much weaker than transitions between states of the same group. Each subset is solved in isolation. A macro model is solved for the probability of being in each subset. The probability of being in a state is estimated as the product of the two foregoing probabilties.

Chandy et al [CHAN,75] proposed the so-called flow equivalent service centre decomposition technique which is based on "Norton's theorem" in the analysis of electrical circuits. It involves estimating a service distribution for the composite centre

representing the subnetwork with a common input and a common output. The service time has queue length dependent service rates. The rates are determined by solving, for each population of customers, a network in which the service times of all service centres outside the subnetwork are set to zero, in effect "shorting" out the rest of the network.

An important aspect of decomposition techniques is the estimation of the second moments (or the coefficient of variation) of the job flow distribution when different flows with general distributions are departing, merging and splitting. Problems of this nature are tackled in [TOMA,89] where a comparative study against the existing decomposition techniques is presented.

b/ Modification of the MVA response time:

The standard MVA algorithm becomes computationally prohibitive when solving very large closed PF-QNM's. In this case and for certain types of NONPF-QNM's, approximate methods of solution based on the modification of the mean response formula are widely used..

To reduce time and space complexity for large closed PF-QNM's, Bard [BARD,79] suggested to use approximate and simplistic rather than exact and recursive arrival theorem formula. The most accurate approximation which is still in use today was proposed by Schweitzer [SCHW,79].

For queueing networks presenting NONPF features such as exponential FCFS server with class dependent server, Reiser [REIS,79] proposed to use the following expression for the mean response time:

$$T_{ir} = \langle S_{ir} \rangle + \sum_{t=1}^{R} \langle S_{it} \rangle A_{it} \qquad (1.3)$$

The approximation was extented to handle nonexponential service time at FCFS servers [REIS,79]. Bard [BARD,79] proposed method similar to Schweitzer's for closed QNM's with simultaneous resource possession.

In priority scheduling environment Bryant et al [BRYA,84] and recently Bondi and Chuang [BOND,88] proposed different expressions for the MVA response time formula. The MVA priority approximations are discussed in details in chapter 2.

C/ Iteration:

In this approach a sequence of simplified networks is solved so that, upon convergence, the results obtained closely approximate the solution of the network of interest. If each network in the sequence has PF solution, the overall method is computationally efficient, provided convergence is obtained in reasonable number of iterations. The objective of iterative methods is to determine a fixed point of multidimensional and nonlinear operator "Φ", which is subsequently used to obtain the various performance measures. The equations involved in a fixed point problem are solved by successive substitutions i.e.,

$$x^{n+1} = \Phi(x^n) \qquad (1.4)$$

This approach was first used by Sevcik [SEVC,77a] to represent the effect of PR queueing discipline. Sevcik transformed NONPF-QNM involving priority discipline into PF-QNM by replacing the priority centre by virtual FCFS servers with modified service rates, so that the network can be solved by existing algorithms. Unfortunately their service rates are not known before hand and therefore the solution of such network involves a fixed point problem.

For networks containing only nonexponential FCFS servers, iterative procedures have been used to capture the variability of the flow process (e.g., [GELE,76; SEVC,77b; REIS,74]).

Recently, an alternative method based on the principle of maximum entropy (PME) has received increasing interest. It has been applied successfully in the analysis of QNM's with FCFS servers and general service time [KOUV,83,86a,88a; WALS 84]. The extension of the ME analysis of queueing networks with priority scheduling disciplines is the subject of this thesis.

1.3 Thesis outline

QNM's under priority scheduling disciplines violate the separability conditions and, as a consequence, their solution is obtained efficently only in approximate manner. The techniques adopted are based on the solution of simpler queueing systems that can be used as a basis for the analysis of more general networks. However, very few exact or approximate results have so far been obtained despite the fact that only Markovian queueing systems are used. For example, to the knowledge of the author no closed-form approximation even for the queue length distribution (qld) of a stable M/M/1[1] priority queue has so far been found in the literature.

1: Kendall's notation A/B/C/D/E is used to classify single resource queueing models. In this notation, A describes the arrival process, B specifies the service process, C denotes the number of servers at the station, D specifies the maximum number of customers waiting at the station, and E is the size of the population.
M/M/1 - Poisson (Markov) input, exponential (Markov) service time, 1 server.
M/G/1 - Poisson input, general service time, 1 server.

Note that obtaining exact solution of this qld is computationally prohibitive, particularly as the number of classes increases [JAIS,68; MARK,72; MILL,81]. Thus it is worthwile to search for more efficient techniques for both Markovian and non-Markovian networks.

The PME is considered to be a uniquely correct, self consistent method of inference for estimating a discrete probability distribution based on information in the form of expected values [JAYN,68; SHOR,80]. In this thesis, the PME is used, in conjunction with classical queueing theory to provide an analytic framework for the analysis of QNM's with priorities.

Our main objectives is first to analyse the single server queue under PR and HOL disciplines and then use the results obtained in order to develop more efficient algorithms to solve general QNM's with priorities.

In chapter 2 some definitions and properties of the priority queueing displines are presented together with a review of existing techniques employed to solve QNM's with either PR or HOL discipline.

In chapter 3 the PME is introduced and a summary of some of the useful results obtained by application of the PME to queueing systems is reviewed. The work is mainly based on the properties and the physical interpretation of the generalised exponential (GE) distribution of the form:

$$F(t) = 1 - \frac{2}{C_V^2 + 1} \exp\left\{ - \frac{2}{C_V^2 + 1} \lambda t \right\} \quad (1.4)$$

with $t \geq 0$, and $1/\lambda$, C_V being the mean and the coefficient of variation, respectively.

The implementation of ME solutions requires analytic estimation of expected values used as prior variables. This is achieved by making use of the versatile GE distribution in chapter 4 where exact analytic and closed-form expressions for GE/G/1 priority queue have been derived. These results represent a generalisation of the existing results of the M/G/1 priority PR or HOL queue [JAIS,68].

Chapter 5 determines the ME approximations of a single server queue under either PR or HOL discipline. The PME is applied under two sets of prior information:

a- Normalisation, utilisation and mean queue length.

b- Normalisation, utilisation, mean queue length and the idle state probability.

New analytic results are presented constituing a "building block" for the analysis of more complex queueing network configurations. Some illustrative numerical examples are given at the end of the chapter.

The ME analysis of open and closed networks will be developed in chapter 6 and chapter 7, respectively. The analysis is sanctioned by two major algorithms for the two types of networks. Numerical examples and comparison against exact, simulation and existing approaches will be given at the end of each chapter.

Finally, conclusion and suggestions for future work are contained in chapter 8.

CHAPTER 2

ANALYSIS OF SINGLE QUEUES AND NETWORKS WITH PRIORITIES : REVIEW

The purpose of this chapter is to review some existing results for single station queues and networks with priorities. We first present some generalities refering to priority disciplines and then summarize some basic analyic results of the single server queue under either PR or HOL discpline. These results are used in turn as a basis for the introduction of current techniques for queueing network models with priorities.

2.1. Generalities

When designing a system involving queues, it is very important to predict some of its performance measures under different scheduling disciplines, so that the most appropriate rule can be adopted. The system under study must be seen from different view points with respect to performance measures, e.g., the queue length distribution is of interest from the designer's point of view, the waiting time from the customer's point of view and the utilisation from the server's point of view. Changing the queueing discipline generally affects these performance measures. Quantities such as the "design measure" [JAIS,68,PP.80] which is the ratio of the mean queue length of a given class of customers in isolation over the mean queue length of the same class under a specific service discipline in conjunction with other classes, may be used to determine the effect of the newly introduced rule on the system performance.

For instance, due to the large amount of processing time of batch

jobs at the CPU, interactive jobs, although requiring smaller processing time may be kept waiting for relatively long time if the two types of jobs are served in FCFS fashion, and subsequently deteriorating the performance of the system (increase in the overall mean waiting time). Consequently, the CPU of modern computer systems generally provides preferential treatments to interactive jobs at the expense of batch jobs.

In queueing systems involving priority disciplines, customers are assigned priorities according to either their class membership (*exogeneous priority*) or to any characteristic relating to the state of the system (*endogeneous priority*), e.g., under time-dependent priority [KLEI,64b], the priority depends on the time spent by a customer in the system. Unfortunately, the analysis of queues involving endogeneous priorities is very complex and asaconsequence, very few analytic results have been derived. For more details about these disciplines we refer to [JAIS,68].

In exogeneous priority situations, there are two possible refinements, preemption and nonpreemption. In preemptive cases the customer with the highest priority is allowed to enter service immediatly even if another customer with lower priority is already present in the service. In addition, a decision has to be made whether to continue the preempted customer's service from the point of preemption when resumed (PR service discipline), or to start anew (Priority preemptive repeat). On the other hand, a priority discipline is said to be nonpreemptive (HOL) if there is no interruption of service, and a higher priority customer just goes to the head of the queue to wait his turn.

In the rest of the thesis it is assumed that customers are discriminated according to their class membership and for

convenience, the priority assignement is in inverse order of the class indices. Thus class-1 has the higest priority, class-2 the second highest, ... , class-R is the lowest.In particular, we focus on the investigation of PR and HOL scheduling disciplines which are two of the most appropriate priority disciplines in modelling computer systems and communication networks.

2.2 The single PR or HOL queue

The PR and HOL disciplines have been of great interest for queueing analysts and performance modelling designers during the last three decades. White and Christie [WHIT,58] were the first who studied the PR discipline by examining it in a two class M/M/1 queue. A few years later Miller [MILL,60] investigated similar system with general service time distribution. Heathcote [HEAT,59] considered the time dependent (or transient) distribution of the number of jobs in the M/M/1 priority queue with 2 classes of jobs, and then extended his results to R (R > 2) classes in [HEAT 60]. Avi-itzhak and Naor [AVI-,61] used the "machine breakdown" model in the analysis of the M/G/1 queue with PR, where they represented the service time of lower priority jobs as the operative time of a machine and its waisted time due the presence of higher priority jobs in the system as the down time of the machine.

The HOL discipline was introduced by Cobham [COBH,54] and was studied subsequently by Miller [MILL,60] and Jaiswal [JAIS,62] under different assumptions regarding the service time distribution. It was investigated also by Keilson [KEIL,62] who used the "machine breakdown" model with postponeable interruptions.

The results obtained up to now, are generally restricted to mean value formulae or transforms (Z-transform[1] or Laplace-Stieljes transform[2] (L.S.T)) which are difficult to invert to obtain joint or marginal probability distribution functions. For instance, procedures enabling the computation of the marginal steady state probabilities of an M/M/1 queue are proposed by Marks [MARK,72] and Miller [MILL,81]. However, the procedures are recursive with respect to the population size and therefore become much more complex and computationally expensive for large number of customers in the system or large number of classes. Note that the problem of multiprocessor systems with priorities is examined in [MITR,81;BUZE,83;BOND,84]. In particular, Mitrani and King [MITR,81] have derived the exact solution of a M/M/2 PR queue and have suggested an approximate analysis of M/M/c, with c>2. The solution obtained is expressed in terms of generating function for the joint and marginal queue length distributions that must be evaluated numerically for each set of parameters. On the other hand, Buzen and Bondi [BUZE,83] have proposed a different solution than the one presented by Mitrani and King for the marginal mean response times in M/M/c PR queue (c>2). This solution turns out to be exact when the service-time is class independent. The approximation is later generalised to M/G/c queue [BOND,84].

1:Let N be a discrete random variable with p(n) as PDF. The Z-transform is defined by

$$Q[Z] = \sum_{n=-\infty}^{\infty} p(n)z^n, \qquad |z| < 1$$

where Z is a complex variable.

2: Let X be a continuous random variable with F(t) as PDF. The L.S.T is defined be

$$F^*(\theta) = \int_{t=-\infty}^{+\infty} e^{-\theta t}\, dF(t) \qquad |\theta| < 1$$

where θ is a complex variable.

Before reviewing some of these results, let us first define some fundamental characteristics and properties seen in priority situations.

2.2.1 Basic definitions

Consider the queueing system of Fig.2.1 consisting of a single server with R different types of customers. The interarrival and service time of class-r, r = 1,..,R, jobs conform to arbitrary distributions under either PR or HOL scheduling disciplines. Let $S_r(.)$ be the probability distribution function (PDF) of the service time of class-r customers with mean service rate μ_r and C_{sr}^2 as squared coefficient of variation and $A_r(.)$ is the PDF of the interarrival time of class-r customers with mean arrival rate λ_r and C_{ar}^2 as the squared coefficient of variation.

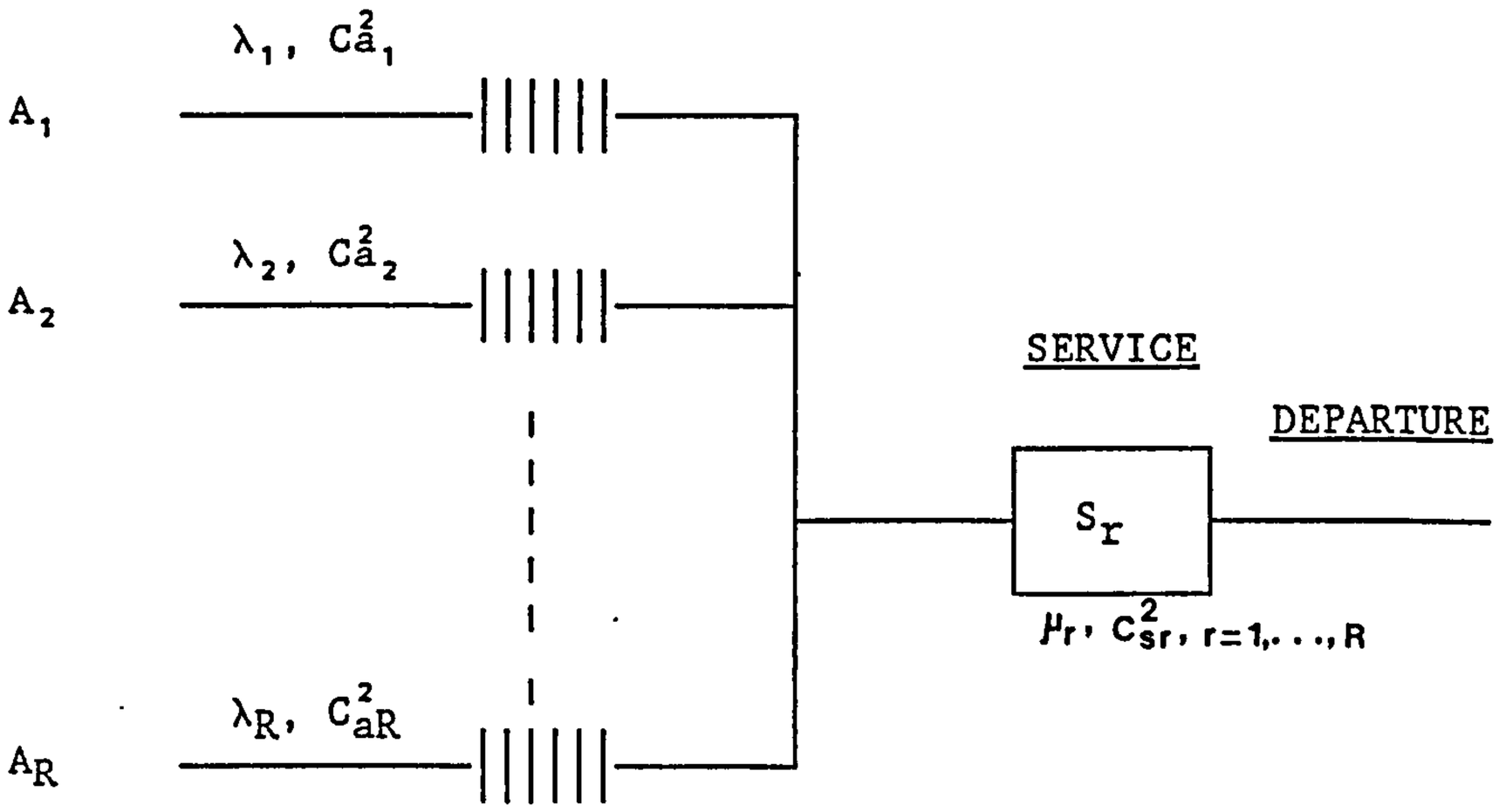

Fig.2.1 Single server queue with PR or HOL

Some other important parameters of a single server queue under PR and HOL disciplines are defined as follows (see Figs.2.2 and 2.3)

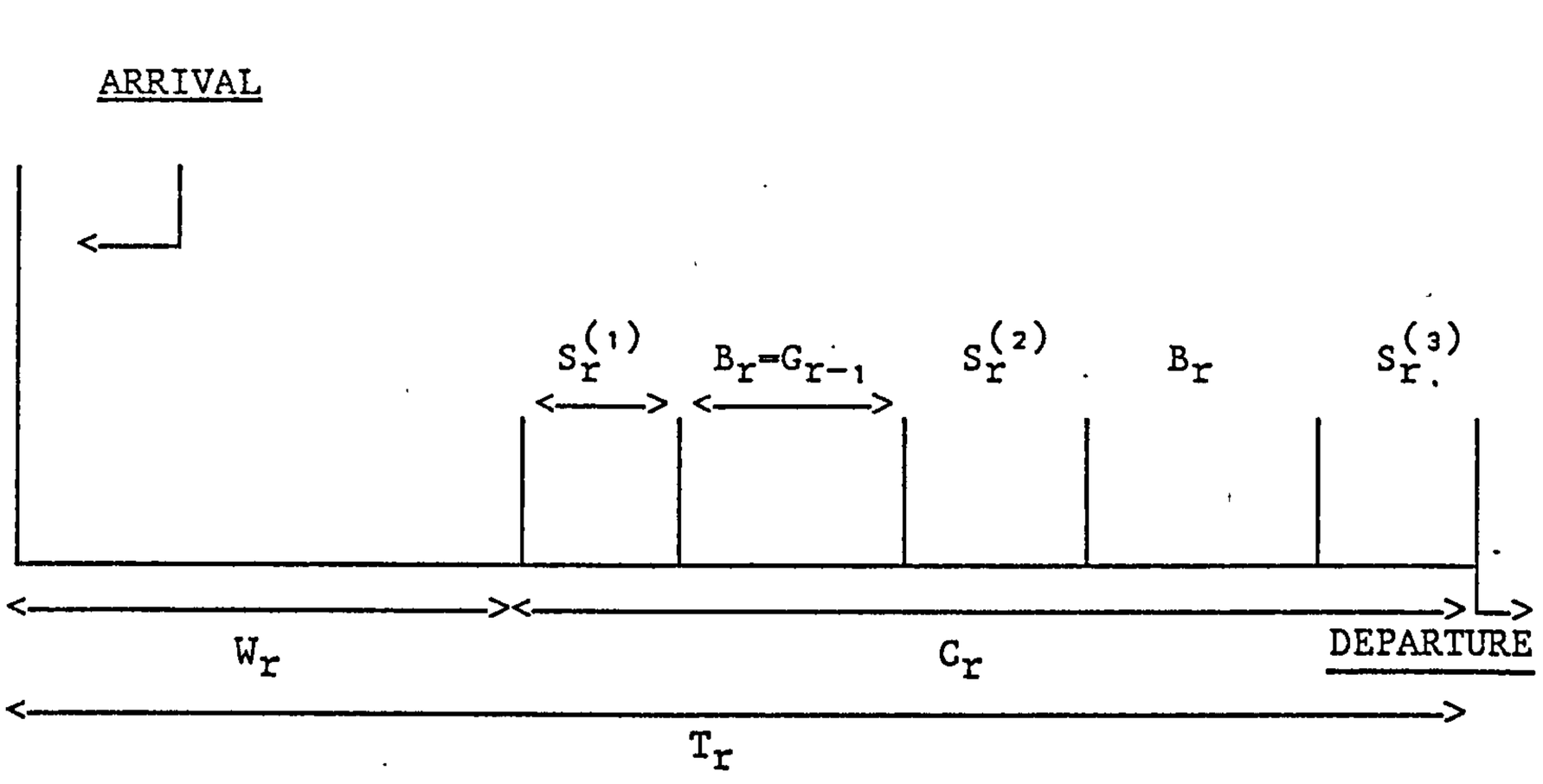

FIG.2.2 Characteristics of class-r job under PR

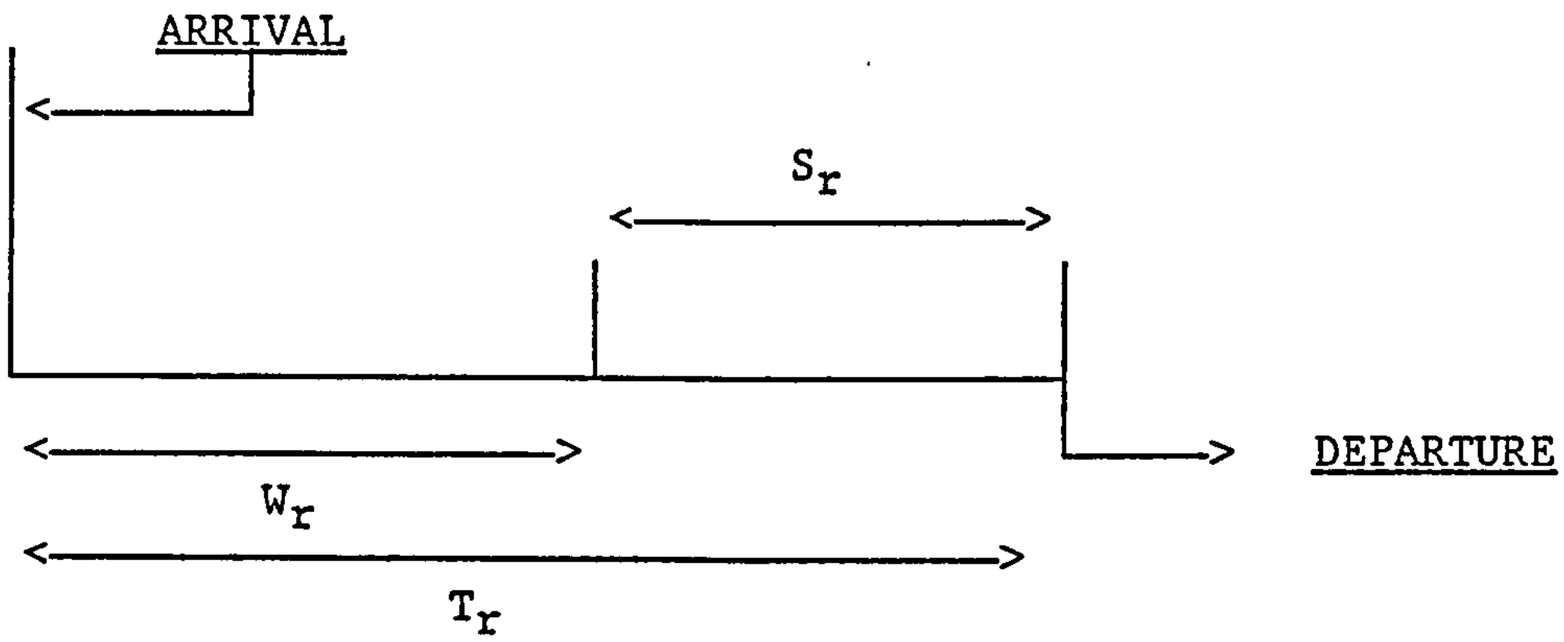

Fig.2.3 Characteristics of class-r job under HOL

W_r : The waiting time of class-r customer or the occupation time of the server with respect class-r customers. It is the time spent by class-r customer in the queue before he receives service for the first time [JAIS,68,pp.20].

$S_r^{(k)}$: The amount of service time received by class-r customer under PR discipline at his k^{th} visit to the server (Fig.2.2)

S_r : The service time of class-r customer.

C_r : The completion time of class-r customer. It is defined as the duration of a period that begins from the instant the service of class-r customer starts and ends at the instant the server becomes free to take the next customer of the same class (if any present) [JAIS,68,pp.56].

T_r : The response time of class-r customer.

G_r : The busy period generated by customers of class 1,...,r. and is defined as the time experienced by the server in serving customers belonging to the classes 1,...,r.

B_r : The breakdown time of class-r customers in PR queue. It is the time waisted by class-r customer after being preempted by a higher priority customer

2.2.2 Properties

A number of important properties are summarised below:

1- PR and HOL are "work-conserving" disciplines since no work (service) is created or destroyed within the system [KLEI,76].

2- Because PR is work-conserving discipline, the following relation is satisfied:

$$S_r = \sum_k S_r^{(k)} \qquad (2.1)$$

Note that, unless the service time is exponentially distributed (memoryless[1]), the relation above is not satisfied for preemptive repeat priority since a preempted customer starts a new service.

1: The exponential distribution is the only continuous distribution possessing the memoryless (or Markovian) property (i.e., if a continuous random variable, X, conforms to an exponential distribution we have :

$$Prob[X-t_0<t \;/\; X>t_0] = Prob[X<t] \quad \text{for } t, t_0 > 0).$$

3- The breakdown time of class-r customer is initiated by the arrival of a higher priority customer and lasts as long as there are higher priority customers in the queue to be served. As a consequence, the breakdown time becomes identical to a busy period generated by higher priority customers.

4- Because PR and HOL are work-conserving disciplines, the server experiences the same busy period under both rules and subsequently class-r customers perceive identical completion time under both rules.

In the next subsections we review briefly some important analytic results which are used as a basis for the approximate and exact analysis of some complex queueing systems with priorities.

2.2.3 M/G/1 priority queue

The stochastic analysis of a single server priority queue has been mainly restricted to pure Markovian arrival process and general service time [JAIS,68]. This is due to tremendous mathematical complications encountered in the general case. Most of the results derived are restricted to mean value formulae or transforms. For instance, the marginal queue length distributions, $\{P_r(n_r)\}$, are only specified by their generating functions, which are almost impossible to invert. A brief description for their derivation follows next.

Consider a queueing system consisting of a single server and R classes of customers arriving in Poisson fashion with rate λ_r (r=1,...,R) and receiving service according to an arbitrary distribution $S_r(.)$. The service discipline is subject to either PR or HOL rule. In particular, under PR discipline, class-r customer is served only when there is no customer belonging to class 1,...,r-1 in the system.

Let us define:

$$\rho_r = \frac{\lambda_r}{\mu_r} \qquad (2.2)$$

ρ_r : The utilisation of the server with respect to class-r customers.

$$\rho = \sum_{r=1}^{R} \rho_r \qquad (2.3)$$

ρ : The overall utilisation of the server.

$$\gamma_r = \sum_{\ell=1}^{r} \rho_\ell \qquad (2.4)$$

γ_r : The utilisation of the server with respect to classes 1,...,r.

For $\rho < 1$ (at equilibrium), the joint steady state queue length distribution exists and the corresponding Z-transform $Q[z_1,z_2,\ldots,z_R]$ is obtained first by considering the system at transient state and taking the result to the limit (time $\longrightarrow\infty$) [JAIS,68]. The expression derived is very complex and is obtained through a lengthy proof. However, the Z-transforms of the marginal steady state queue length distribution have simpler form and are determined by appropriate substitutions, e.g., for class-r we have:

$$Q_r(z_r) = Q[1,1,\ldots,1,z_r,1,\ldots,1] \quad .$$

Where after some manipulations, it is expressed by

$$Q_r(z_r) = (1-\gamma_r)\,\frac{\left[\lambda_r z_r - \Lambda_r + \Lambda_{r-1} G^*_{r-1}(\lambda_r-\lambda_r z_r)\right]}{\lambda_r\left[Z_r - C^*_r(\lambda_r-\lambda_r z_r)\right]}\, C^*_r(\lambda_r-\lambda_r z_r). \qquad (2.5)$$

$$\Lambda_r = \sum_{\ell=1}^{r} \lambda_\ell \qquad (2.6)$$

where $\overset{*}{G}_{r-1}(.)$ and $\overset{*}{C}_r(.)$ are the L.S.T of the busy period of the server with respect to classes $\{1,2,..,r-1\}$ and the class-r completion time, respectively, and can be found [JAIS,68].

The above expression may also be obtained by using simple probabilistic arguments (see appendix A1).

The differentiation of equation (2.5), leads to the following marginal mean queue lengths:

i/ For PR discipline

$$\langle n_r \rangle = \frac{\rho_r}{1-\gamma_{r-1}} + \frac{\sum_{\ell=1}^{r} (\lambda_r/\lambda_\ell)\rho_\ell^2(C_{s_\ell}^2 + 1)}{2(1-\gamma_{r-1})(1-\gamma_r)}, \quad r=1,2...,R \qquad (2.7)$$

ii/ For HOL disciplne

$$\langle n_r \rangle = \rho_r + \frac{\sum_{\ell=1}^{R} (\lambda_r/\lambda_\ell)\ \rho_\ell^2\ (C_{s_\ell}^2 + 1)}{2(1-\gamma_{r-1})(1-\gamma_r)}, \quad r=1,2,...,R \qquad (2.8)$$

Note that the numerator of (eq.2.8) involves the remaining service time of a customer found in service upon the arrival of a class-r customer. It also depends on the parameters of higher priority classes as well as those of lower priority classes, which is expected in HOL situations.

2.3 Queueing networks with priority disciplines

Priority disciplines are very important features to take into account when modelling modern computer systems or communication networks. However, exact solutions of priority QNM's have produced only few analytic results due to the computational expense of solving the global balance equations involved. For example, in Avi-itzhak paper [AVI-,73], the mean system response times are obtained for a homogeneous central server network where all priority classes have identical service times and routing frequencies. Morris [MORR,81] presented the exact analysis of nonhomogeneous two-centre Markovian networks where each centre is under PR or HOL. However, the complexity of the solution of such networks increases with the number of classes and therefore it is of limited practical use. Mitrani [MITR,72] considered special cases of closed queueing networks consisting of two stations serving N customers under either PR or HOL. Each customer in the system is assigned a distinct priority. Exact mean system response times as well as utilisations and throughputs are obtained. However, the exact analysis of QNM's with priorities is generally confined to very small networks with small population sizes and therefore becomes not of practical interest. To this end, approximate methods which are computationally efficient are generally used in performance modelling of computer systems and communication networks. Unfortunately QNM's with priority disciplines violate the conditions of separability [BASK,75; LAZO,84] and therefore cannot be solved directly by existing fast computational algorithms such as convolution and MVA. Special techniques are then required to take into consideration the effect of the discipline and provide solutions to such models. These techniques consist either of solving in isolation the priority centre using decomposition methods

or transforming a nonproduct-form QNM into product-form one using iterations or by modifying the MVA response time formula.

Some known methods are presented below:

2.3.1 Composite centre approximations.

These techniques are based on Norton's theorem (flow equivalent service centre approach) of a decomposition of a network into subnetworks [CHAN,75]. They are exclussively applicable to Markovian queueing networks of central server type. The CPU is subject to priority disciplines (PR or HOL), whereas the set of I/O's are under FCFS rule. The subnetwork which comprises from the I/O units, is solved in isolation by known algorithms and then it is replaced by a single composite centre with a queue-length-dependent-service-rate. The reduced model is finally solved by a global balance technique (c.f. Fig.2.4).

The technique becomes too complex to be of practical value as the number of classes and the population size increases.

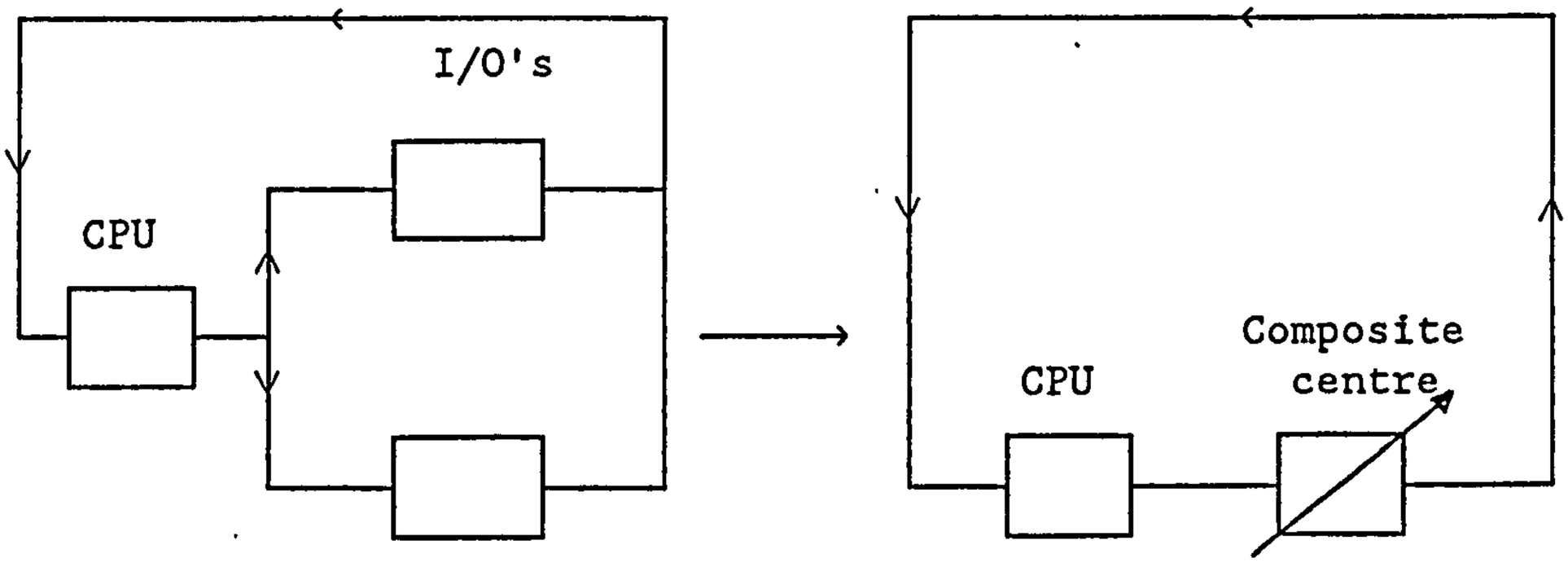

Fig.2.4 Composite centre approximations

For networks with more than two priority classes, Sauer and Chandy [SAUE,75b] analyse each class in turn. They proposed to use a

model with three classes of customers, namely, a designated class, lower composite class and higher composite class which consist of the class under investigation, all lower priority classes and all higher priority classes, respectively. The CPU is assumed to serve customers of the composite classes exponentially in time, with mean service time equals to the weighted sum of the individual mean service times of the classes being coalesced and where the weights are the relative throughputs of the original classes within the composite class. Note that the exponential assumption of the service time of the composite class can be strongly violated, if the classes have very different values of the mean service times.

A variation of this technique is proposed by Chow and Yu [CHOW,83]. It consists of solving the same central server model iteratively by considering two classes at the time. At each iteration a new class is considered together with the composite class of all higher priority classes, the capacity of the I/O's processor (FCFS servers) available to higher priority jobs is reduced because of the possible existence of lower priority jobs. Decomposition as well as global balance techniques are necessary for the solution of such models. The performance measures of interest are given by the last iteration.

In principle, the composite centre approximations can be applied to networks with more than one priority centre. However, the computational cost of the global balance solutions increases rapidly with the number of priority centres.

Note that decomposition technique involving a transformation of priority centres into product-form stations have been proposed by Neuse and Chandy [NEUS,82] to solve QNM's with priorities. However, their approach is iterative with respect the parameters of the modified queues and therefore becomes very costly for large networks.

2.3.2 Shadow CPU based approximations

Consider a queueing network with one or more priority centres and let us assume that the priority feature is the only nonproduct-form characteristic presented by the system in question.

Sevcik [SEVC,77a] proposed to transform the original NONPF queueing network into PF one so that the latter can be solved by current algorthims (convolution, MVA). He suggested to represent the single priority scheduled server (CPU) in the actual system by R virtual (shadow) FCFS service centres in the model, each visited only by a single class. For example, Fig.2.5 illustrates the transformation for 2 classes with CPU service rates μ_{11} and μ_{12}, respectively.

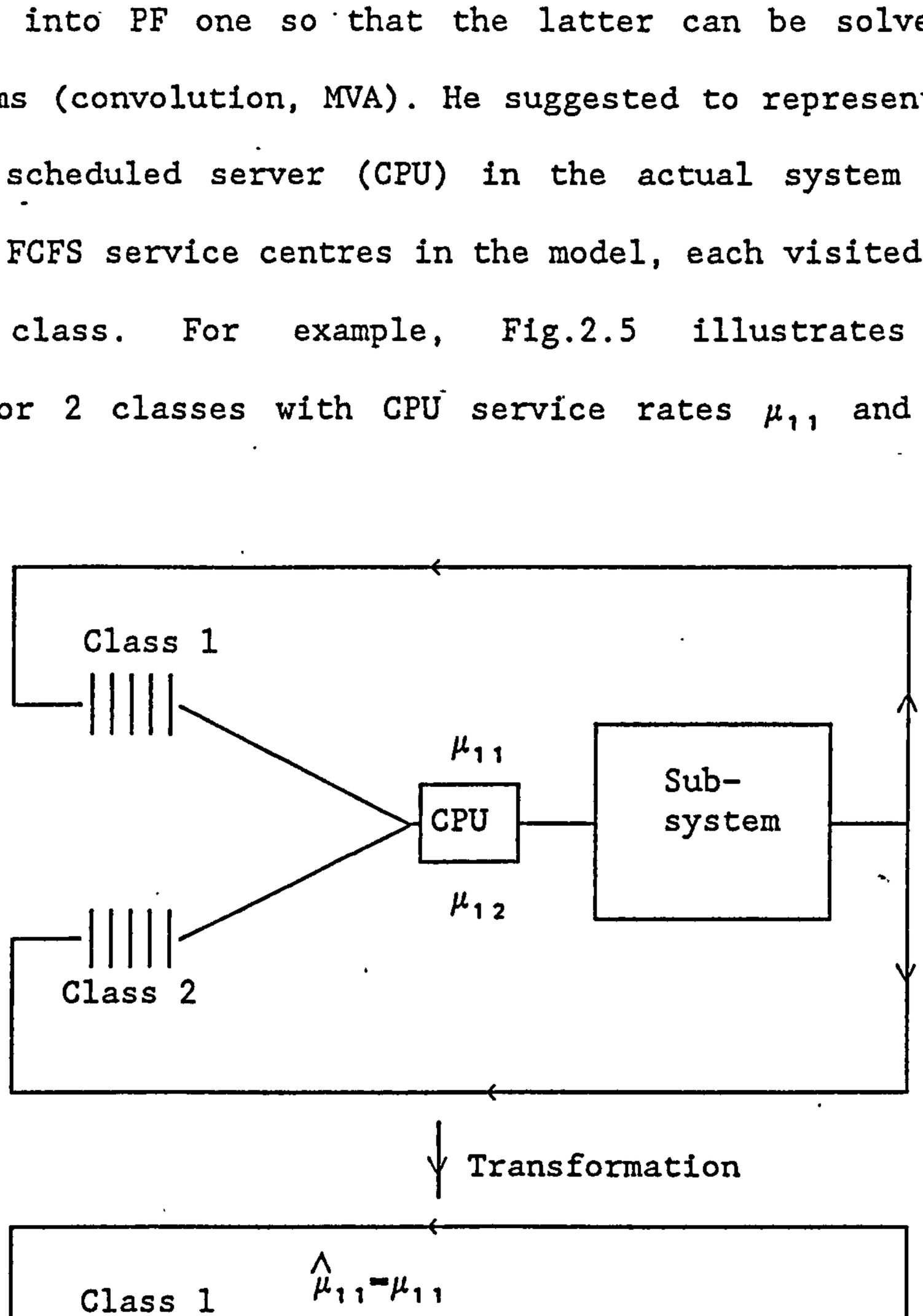

Fig2.5 Shadow CPU transformation

Each "dedicated" shadow centre is assumed to be an exponential server with reduced capacity to account for degradation due to use of the actual server by higher priority jobs. Sevcik's approach (known as the "reduced occupancy approximation" (ROA)) solves networks with PR centres (and not HOL) and the mean effective service time of class-r or the mean service time of the r^{th} virtual server, $1/\hat{\mu}_{ir}$, is taken to be equal to the actual service time of class-r divided by the proportion of the time that the CPU is not serving jobs of higher priority classes.namely,

$$\hat{\mu}_{ir}^{-1} = \frac{\mu_{ir}^{-1}}{1 - \gamma_{ir-1}} \qquad (2.9)$$

Where $\hat{\mu}_{ir}^{-1}$ is the mean effective service time of class-r at centre-i.

Note that the ROA uses exactly the mean completion time of an M/G/1 queue under PR [JAIS,68] as the mean effective service time.

Kaufman [KAUF,84] argued that the mean effective service time of the ROA is structurally flawed. Lower priority jobs do not always perceive their effective service time to be an ordinary completion time (i.e., delay error). In fact, a lower priority arriver which finds the server busy with a higher priority job and no jobs of its own class in the queue, has to wait for the remaining higher priority busy period before commencing its completion time.

Kaufman suggested that the correction factor should be conditioned on the absence of jobs of high-priority classes given that low-priority jobs are present and proposed (via little's law) that the exact mean effective service time to be

$$\hat{\mu}_{ir}^{-1} = \frac{\mu_{ir}^{-1}}{1 - \text{Prob}(m_{ir-1}>0 \;/\; n_{ir>}0)} \qquad (2.10)$$

where

$$m_{ir} = \sum_{\ell=1}^{r} n_{i\ell}$$

Unfortunately , the exact computation of the conditional probability is itself prohibitively expensive. Kaufman suggested an easily computed estimate of this value in terms of the utilisations of the priority centre and demonstrated that its use retains a significant portion of the increased accuracy attained using the exact conditional probability. The approximate mean effective service time used by the modified -ROA (m-ROA) is given by [KAUF,84]

$$\hat{\mu}_{ir}^{-1} = \frac{\gamma_{ir-1}(1-\gamma_{ir-1}) + \delta_{ir}\gamma_{ir}}{\gamma_{ir-1}(1-\gamma_{ir-1})^2 + \delta_{ir}\rho_{ir}} \qquad (2.11)$$

where

$$\delta_{ir} = \sum_{\ell=1}^{r-1} \rho_{i\ell}\ (\mu_{ir}/\mu_{i\ell})$$

The m-ROA has been extended also to nonpreemptive priority discipline, where the exact mean effective service time of class-r jobs is given by

$$\hat{\mu}_{ir}^{-1} = \frac{\mu_{ir}^{-1}}{1 - \text{Prob}(\ m_{ir}^{c} > 0,\ I_s \neq r\ /\ n_{ir}>0\ )} \qquad (2.12)$$

where I_s is the class index of the customer currently in service.

and

$$m_{ir}^{c} = \sum_{\substack{\ell=1 \\ \wedge \\ \ell \neq r}}^{R} n_{i\ell}$$

Unfortunately neither exact nor approximate closed form expression is given for the conditional probability.

Schmitt [SCHM,83,84] uses a load dependent virtual server which provides an exact description of the marginal distributions of the M/M/1 queue with PR or HOL. For example, the mean effective service time of class-r under PR discipline is given by:

$$\hat{\mu}_{ir}^{-1}(n) = \frac{\mu_{ir}^{-1}}{1 - \text{Prob}(m_{ir}>0 \;/\; n_{ir}=n)} \qquad (2.13)$$

However, the conditional probability is a priori unknown. Schmitt suggests to decompose the network into subnetworks and to estimate the probability in question by the global balance technique which may be prohibitively costly for large networks.

2.3.2.1 Shadow CPU algorithm for open networks [SEVC,77a;KAUF,84].

A simple procedure based either on ROA or m-ROA can be used for solving open QNM's with PR servers since the utilisations are obtained easily for external arrival processes. The procedure is summarized in algorithm 2.1.

To this end, we consider an open network with M centres and R classes where some centres are under PR rule and P_{irj} designates the transition probability that class-r job having just finished service at centre-i is directed to centre-j.

The index 0 denotes the outside world (external source).

Begin

STEP 1 {* solve the flow balance equations *}

$$\lambda_{ir} = \sum_{j=0}^{M} \lambda_{jr} P_{jri} \qquad (2.14)$$

STEP 2 {* compute the utilisations *}

$$\rho_{ir} = \frac{\lambda ir}{\mu_{ir}} \quad \text{for } i = 1,\ldots,M \text{ and } r = 1,\ldots,R$$

STEP 3 {* create the R shadow CPU's for each priority centre with $\hat{\mu}_{ir}^{-1}$ as service time of the r^{th} shadow CPU given by (eq.2.9) and (eq.2.11) for ROA and m-ROA, respectively. *}

STEP4 {* Analyse each queueing station separated from the rest of the network and subject to Poisson arrival with rates λ_{ir} obtained in step 1 *}

End.

Algorithm 2.1 ROA and m-ROA for open QNM's

2.3.2.2 Shadow CPU algorithm for closed networks

[SEVC,77a;KAUF,84]

Since the external arrival rate for closed networks is set to zero, the linear system of equations given by (eq.2.14) have infinite number of solutions and therefore the utilisations are a-priori unknown. The problem is solved by the fixed point iteration method with respect to the utilisations. A stepwise presentation of the method is given in algorithm 2.2

Begin

STEP 1 {* Initialisation of the utilisations *}

$\rho_{ir} \longleftarrow 0$ for i = 1,...,M and r=1,...,R

STEP 2 {* Create R shadow CPU's for each priority centre with $\hat{\mu}_{ir}^{-1}$ given by (eq.2.9) and (eq.2.11) for ROA and m-ROA respectively. *}

STEP 3 {* solve iteratively the QNM *}

step 3.1 Use either the standard convolution or the MVA algorithm to compute the statistics

step 3.2 Repeat step 3 until successive estimates of the utilisations are sufficently close.

STEP 4 {* obtain the final performance measures from the last iteration. *}

End.

Algorithm 2.2 ROA and m-ROA for closed QNM's

2.3.3. MVA priority approximations [BRYA,84]

These approaches are based on the standard MVA algorithms, where the mean response time formula is modified to accomodate the priority disciplines.

2.3.3.1 MVA priority approximations for open QNM's

In open QNM's, the solution of the flow balance equations (eq.2.14) is used in (eq.2.7) and (eq.2.8) for the estimation of the mean queue length of each class at PR and HOL server, respectively.

The MVA priority approximation technique is summarized in the algorithm below:

```
Begin
   STEP 1 {* Solve the flow balance equations (eq.2.14) *}
   STEP 2 {* Solve each centre in isolation as an M/M/1 queue *}
             - Use (eq.2.7) for PR centre
             - Use (eq.2.8) for HOL centre
   STEP 3 {* obtain the response times via Little's
             law.*}
End.
```

ALgorithm 2.3 MVA priority approximation for OPEN QNM's

Note that the arrival process to a centre in a queueing network is generally not Poisson. In fact Lower priority jobs experience highly variable interdeparture time process (which is also the interarrival time process to some other centres) due to the variability of higher priority busy period [NAIN,84].

2.3.3.2 MVA approximations for closed QNM's

Using just simple probabilistic arguments, together with the memoryless property of the exponential distribution, Bryant et al [BRYA,84] proposed a modified mean reponse time formula in the context of the standard MVA algorithm for PR and HOL centres.

If the centre-i is subject to PR discipline, the MVA mean response time formula of class-r is given by

$$T_{ir}(\underline{N}_R) = \frac{<S_{ir}> + \sum_{\ell=1}^{r} A_{i\ell}(\underline{N}_R) <S_{i\ell}>}{1 - \sum_{\ell=1}^{r-1} <S_{i\ell}> \lambda_{i\ell}(\underline{N}_R - <n_{i\ell}>\underline{1}_\ell)} \qquad (2.15)$$

Where $\underline{1}_r = (0,\ldots,0,1,0,\ldots,0)$ is the R-dimentional vector with unity at the r^{th} entry and $A_{ir}(\underline{N}_R)$ is given by the arrival instant theorem (averaged mean queue length of class-r jobs at centre-i when the population of the system is $\underline{N}_R - \underline{1}_r$), $<S_{i\ell}>$ and $<n_{i\ell}>$ are the mean service time and the mean queue length of class-ℓ at centre-i, respectively.

If centre i is a HOL station, the mean response time formula is given by

$$T_{ir}(\underline{N}_R) = <S_{ir}> + \frac{\sum_{\ell=1}^{r} A_{i\ell}(\underline{N}_R) <S_{i\ell}> + \sum_{\ell=r+1}^{R} \lambda_{i\ell}(\underline{N}_R - \underline{1}_r) <S_{i\ell}>^2}{1 - \sum_{\ell=1}^{r-1} <S_{i\ell}> \lambda_{i\ell}(\underline{N}_R - <n_{i\ell}>\underline{1}_\ell)} \qquad (2.16)$$

Note that, although the above equations provide the exact mean response time of class-r in the M/M/1 priority queue, the arrival theorem does not hold for priority queueing networks and therefore the formulae used are only approximations. The mean response time formula is kept unchanged for non-priority centres.

Very recently an attempt to capture the effect of the preemptions not only at the priority (PR only) centres, but also at the other stations in the network is made by Bondi and Chuang [BOND,88]. They proposed a new mean response time formulae not only for PR centres but for non-priority ones as well.

Eager and Lipscomb [EAGE,88] extend Bryant et al's approach to

the approximate MVA algorithm [SCHW,79], so that large networks can be solved efficiently without great loss of the accuracy.

The MVA priority based approximations for closed QNM's are depicted in algorithm 2.4.

2.3.4 Discussion

All the present priority approximation techniques, although sometimes highly efficient, they are restricted only to Markovian type networks.

The computational complexity of the global balance technique limits the application of the composite centre approximation to the solution of priority networks with a small number of centres and customers.

The shadow CPU based approximations are iterative methods and although originally implemented on top of the convolution algorithm, they fit naturally on top of the approximate and exact MVA algoritrhms [LAZO,84]. On the other hand, the MVA priority based methods are noniterative and generally more accurate.

However, in contrary to shadow CPU based techniques, the MVA based methods do not capture the behaviour of the null process [ZAHO,87] (i.e., in closed networks, the priority centre should be satured with 100% utilisation if an imaginary low-priority job receives service only at the priority center during its idle time).

Moreover, the ROA is not notably accurate due to several potential sources of error. For example, it fails to capture the "delay error" which stems from the construction of the virtual server though which job of low-priority may commence service immediately on arrival even though high-priority job may be present at the original centre.

```
Begin
   STEP 1 {* Initialise mql of class-r jobs at centre-i *}.
          <n_ir>[0] ⟵ 0 for i=1,...,M and r=1,...,R
STEP 2 {* Compute the statistics for each feasible population
          n = (n_1,...,n_R) *}.
          for n ⟵ 0 to N_R do
          begin
              STEP 2.1
              for r ⟵ 1 to R do
              for i ⟵ 1 to M do
              - obtain the mean response time of class-r at
                  centre-i (use appropriate equations).
          STEP 2.2
          for r ⟵ 1 to R do
          begin
               - evaluate the system throughput of class-r.
               - for i ⟵ 1 to M do
                 begin
                   - Evaluate the throughput of class-r at
                     centre-i for the population n.
                   - obtain the new mean queue length of
                     class-r at centre-i.
                 end;
          end;
   end;
End.
```

Algorithm 2.4 MVA priority approximations for closed QNM's

Although the m-ROA does not suffer too much from the "delay error", it fails to predict accurately the effective service time distribution. For instance the m-ROA assumes exponential effective service time, whereas it is actually highly variable (due to the effect of preemption of high-priority jobs).

Most of the techniques described above suffer from the effect of preemption on the interarrival time variability of low-priority jobs at non-priority centres. Furthermore, they don't take into account the so-called "synchronisation error" [KAUF,84] which arises from the work profile of non-priority servers (i.e., in closed networks, a low-priority job leaving, for example, a PR centre (CPU) will find all high-priority jobs present at non-priority centres (I/O subsystem), a property that violates the arrival theorem and which it is assumed to hold in the MVA based methods).

Note that , although Bondi and Chuang's approximation captures some of 'synchronisation error' and 'the effect of preemption on the interarrival time variability of low-priority jobs', it does not capture the null process behaviour. Furthermore, given that their approach is still based on the exact arrival instant theorem, the technique becomes less accurate mainly when several priority centres are involved.

2.4 Conclusions

The analysis of the single server with priority has been found very difficult to tackle using classical queueing theory. The present literature includes only mean value formulae and transforms which are difficult to invert (i.e., there is no known closed form solution of the steady state probabilities even for the simplest M/M/1 priority queue).

The lack of analytic results in the single server case makes the analysis very complex in the context of priority queueing networks. All existing approaches are based on heuristic approximations. In particular, they experience significant errors when a priority centre has high utilisation mostly attributed to high-priority classes. This may be partly due the fact that these methods do not capture the variability of the interdeparture time and interarrival time per class of both priority and non-priority centres in the network. It is in such cases of high utilisation, that the service centre has the greatest effect on overall performance and that priority scheduling has the largest effect on the service station. Furthermore, not much work has been done so far on QNM's with priorities and general service times. Thus, it is worthwhile to search for new techniques which will improve the accuracy and extend the applicability of the approximate methods.

The Principle of Maximum Entropy (PME) is used in this thesis as a new method to analyse single server queues and QNM's with priorities. In the next chapter, we introduce the PME and review some of its applications in queueing systems.

CHAPTER 3

MAXIMUM ENTROPY ANALYSIS

AND QUEUEING SYSTEMS: A REVIEW

In this chapter we are concerned with the application of the Principle of Maximum Entropy (PME) to queueing systems and the review of some useful relating results.

In section one, a historical background is presented in order to explain the origin of the PME. In section two, we present the ME formalism as a nonlinear programming problem which can be solved by the Lagrange's method of indetermined multipliers, leading to a product-form solution. In section three, we review some applications of the PME to queueing systems. In particular, we show that the PME method when properly applied, permits not only to obtain some exact results which are known from classical queueing theory, but also provides closed-form expressions for the approximate solutions of more general queues. In section four, we introduce the generalised exponential (GE) distribution and see how this distribution is related to the ME solutions. We investigate some of its properties and mainly, we examine its physical interpretations where a rigorous proof for its correspondance to a compound Poisson process is presented. Some useful GE-type formulae are given at the end of the section.

We conclude the chapter by a brief summary on the PME method and the GE distribution.

3.1 BACKGROUND

Consider a system evolving in time or space in accordance with probabilistic laws and may be in any one of a given set of states; the state probabilities are not generally known, but information about the probability distribution may be available in the form of mean values. One should then estimates the state probabilities subject to the mean value constraints.

The Problem of probability assignment has a root in Bernoulli principle (1713) known as 'the principle of insufficient reason' which implies:

' *1. A probability assignment is a state of knowldge.*

2. The outcomes of an event should be considered initially equally probable unless there is evidence to make us to think otherwise. '

For example, consider a system with finite number of states n, the best possible probability assignment P_i to be in state i, i=1,2,...,n, is to set $P_i = 1/n$. In other words, for a system with finite space, the uniform distribution is the least minimally prejudiced distribution in the absence of information.

Jaynes [JAYN,79] extended Bernoulli's principle to the constrained problem, where prior information about the system is available. He used the entropy functional as a measure of the amount of uncertainty, introduced earlier in information theory by Shannon [SHAN,48]. Moreover, he noticed that in the absence of prior information, the entropy attains its maximum when all outcomes of an event are equally probable. He then suggested that one should initially start with a distribution of ME (uniform if it exists), and then 'adjust' this distribution to maximize the entropy subject to what is known.

In information theoretic terms, the ME distribution is

interpreted as the one which is maximally non-commital with regard to missing information and the best supported solution subject to the constraints given. Jaynes [JAYN,57a] justified the use of entropy maximization as follows:

'*the most reasonable assignment for the state probabilities is such that the mathematical uncertainty of the probability distribution is maximized, because if any other assignment were chosen, the amount of mathematical uncertainty of the probability distribution would not reflect adequatly the uncertainty about it.*'

Jaynes' suggestion known also as 'the principle of maximum entropy (PME)', has been shown to be a uniquely correct self-consistent method of inference for estimating probability distributions based on available information given in the form of known (or known to exist) mean values [SHOR,80].

The PME has been applied primarily to statistical mechanics [JAYN,57a,57b], statistics [TRIB,69,chap.6], reliability estimation [TRIB,69,chap.10], queueing theory (e.g., [FERD,70; SHOR,78,82; KOUV,83,86a,88a], and system modelling [BARD,80a,80b] for the analysis of I/O subsystems.

3.2 MAXIMUM ENTROPY FORMALISM

Consider a system Q that has a set of possible discrete states $\{S_0, S_1, \ldots, S_n, \ldots\}$ which may be finite or contable infinite. Let $\tilde{X}$ be the random variable (r.v) describing the state of the system.

With $P_n = \text{Prob}[\tilde{X} = S_n]$.

In addition, there is information available about the system Q in the form of (m+1) constraints,

$$\sum_{S_n \in Q} P_n = 1 \qquad (3.1)$$

$$\sum_{S_n \epsilon Q} f_k(S_n) P_n = \langle f_k \rangle \quad \text{for } k=1,\ldots,m \qquad (3.2)$$

where $\{\langle f_k \rangle\}$ is a set of mean values defined through appropriate functions $\{f_k(S_n);\ S_n \epsilon Q\}$ of system state and equation (3.1) represents the normalisation constraint.

Note that the mean values are estimated either theoretically using analytic expressions or via system measurements.

Because in general the number of possible states is much greater than the number of constraints, there is infinite number of distributions $\{P_n\}$ that satisfy the constraints (3.1) and (3.2). The question is which one to choose?

The PME states of all distributions satisfying the constraints supplied by the given information, the least biased distribution is the one that maximizes the system's entropy function given by the equation below:

$$H(P) = -\sum_{S_n \epsilon Q} P_n \cdot \mathrm{Log}(P_n) \qquad (3.3)$$

The maximization of (3.3) subject to (3.1) and (3.2) is a nonlinear programming problem, which is solved by Lagrange's method of indetermined multipliers leading to the following solution:

$$P_n = \frac{1}{Z} \exp - \left\{ \sum_{k=1}^{m} \beta_k \, f_k(S_n) \right\} \qquad (3.4)$$

where $\{\beta_k\}$ are the Lagrangian multipliers corresponding to constraints (3.2) and Z is known in statistical physics as 'the partition function' which is given by:

$$Z = \exp\{\beta_0\} = \sum_{S_n \epsilon Q} \exp - \left\{ \sum_{k=1}^{m} \beta_k \, f_k(S_n) \right\} \qquad (3.5)$$

Where β_0 is the Lagrangian multiplier that corresponds to the normalisation constraint (3.1).

It can be verified easily that for finite state space and in the absence of constaints (3.2), the ME solution (3.4) reduces to the uniform distribution. This result demonstrates that Bernoulli's principle of insufficient reason is a special case of the PME. In more commun term, the ME solution (3.4) treats all possible alternatives as equally as possible, subject to the information provided.

The maximum entropy (ME) probability distribution (3.4) has many interesting properties which were first explored by Jaynes [JAYN,68,79]. For example, it can be easily shown [TRIB,69,pp.124-125], that the Lagrangian multipliers $\{\beta_k\}$ satisfy the following relations:

$$- \frac{\partial \beta_0}{\partial \beta_k} = \langle f_k \rangle \quad \text{for } k=1,\ldots,m \qquad (3.6)$$

Similarly higher moments of the distribution may also be expressed with respect to the Lagrangian multipliers.

The ME solution (3.4) can also be written in the following simple form:

$$P_n = \frac{1}{Z} \prod_{i=1}^{m} x_i^{f_i(S_n)} \qquad (3.7)$$

where $$x_i = e^{-\beta_i} \qquad (3.8)$$

x_i is defined as the Lagrangian coefficient corresponding to constraint i.

The ME formalism can be applied in the performance analysis of queueing systems since expected value of various distributions of interest are usually known in terms of moments of the interarrival and service time distributions which are generally obtained either analytically or via system measurements.

3.3 APPLICATION OF THE PME TO QUEUEING SYSTEMS.

3.3.1 Single class of customers

The PME has been applied to queueing systems since the early 1970's, Ferdinand [FERD,70] used the principle with only the mean queue length as constraint to derive the steady state probability distribution of M/M/1/N queue by analogy with statistical mechanic. Shore [SHOR,78] investigated the queueing systems M/M/∞//N and M/M/∞, where from an abstract model he determined the ME solution. Few years later, he studied the M/G/1 and G/G/1 queues where higher moments of the service and interarrival time distributions were taken into consideration [SHOR,82]. He based the analysis only on the normalisation and mean queue length constraints to derive a ME solution of geometric type. Moreover, taking the utilisation into account together with the mean queue length and the normalisation as constraints, El-Affendi and Kouvatsos [EL-AF,83] showed that the ME solution of an M/G/1 queue is of modified geometric form.

The two following corollaries expose the two results above:

corollary 3.1 [SHOR,82]

The ME solution of the queue length distribution of a M/G/1 queue given the normalisation constraint

$$\sum_{n=0}^{\infty} P_n = 1 \quad ,$$

and the mean queue length constraint

$$\langle n \rangle = \sum_{n=0}^{\infty} nP_n \quad ,$$

is the geometric distribution given by

$$P_n = (1-x)x^n \quad , \quad n \geq 0 \qquad (3.9)$$

Where x is the Lagrangian coefficient corresponding to the mean queue length constraint and given by

$$x = \frac{\langle n \rangle}{\langle n \rangle + 1} \qquad (3.10)$$

#

Corollary 3.2 [EL-AF,83]

The ME solution of the queue length distribution of a M/G/1 queue given the normalisation, mean queue length and utilisation constraints ($\rho = 1-P_0$) is given by

$$P_n = \begin{cases} 1-\rho & \text{if } n = 0 \\ (1-\rho)gx^n \quad , & \text{for } n > 0 \end{cases} \qquad (3.11)$$

Where g and x are the Lagrangian coefficients corresponding to the utilisation and mean queue length constraints, respectively and

are given by

$$x = \frac{\langle n \rangle - \rho}{\langle n \rangle} \qquad (3.12)$$

$$g = \frac{\rho(1-x)}{x(1-\rho)} \qquad (3.13)$$

#

Note that the mean queue length in M/G/1 queue (Pollaczeck - Khinchin formula) is expressed analytically [KLEI,75,pp.187], and given by

$$\langle n \rangle = \rho + \frac{\rho^2(1 + C_s^2)}{2(1-\rho)} \qquad (3.14)$$

where C_s^2 is the squared coefficient of variation of the service time distribution.

Similar results have been established for G/M/1 in [EL-AF,83] and G/G/1 in [KOUV,88a].

Note that the exact M/M/1 queue length distribution is obtained in both cases by using $C_s^2 = 1$.

The finite capacity queue with general service and interarrival time distributions (G/G/1/N) has been examined by Kouvatsos [KOUV,86a] under the set of four constraints; normalisation, utilisation, mean queue length and the flow balance constraints ($\lambda(1-P_N(N)) = \mu(1-P_N(0))$, where λ and μ are the mean arrival and service rates, respectively, and $P_N(n)$ is the long run probability to have n jobs in the system when the capacity size is limited to N). The Lagrangian coefficients corresponding to the mean queue length and utilisation constraints are estimated by making an assymptotic connection to the infinite capacity (N ——> ∞). In particular, they

are assumed to be invariante to the capacity size.

On the other hand, the ME solution for a stable open network with infinite capacity queues subject to the marginal utilisation and mean queue length constraints, implies a decomposition of the network into individual G/G/1 queues under revised arrival and service process [KOUV,85]. However, for general closed queueing networks, the product-form approximation obtained by entropy maximisation is described in terms of Lagrangian coefficients which in essence are output (unknown) parameters [KOUV,83,86b]. A good estimation of these coefficients via closed-form approximations is necessary in order to establish an efficient implementation via a convolution type algorithm. Towards this goal, The Lagrangian coefficients are approximated by analysing instead a 'pseudo' open network (i.e., a network without a fixed number of jobs and external arrival process) which has a nearly identical topology to that of the original closed network (i.e., both networks have the same number of queues, server characteristics and transition probabilities) satisfying the principles of conservation of flow (expressed by the job flow-balance equations) and the conservation of population (represented by the fixed mean population constraint).

3.3.2 multiple classes of customers

The ME methodology for general networks with multiple classes and FCFS centres, was first proposed by [KOUV,83,85]. A technical correction to the original ME algorithm for closed networks was carried out by Kouvatsos [KOUV,86b] and also by Walstra [WALS,84]. The method proposed is based on the ME solution of a single G/G/1 queue. Therefore, class composition and disaggregation techniques are used to accomodate the multiple-class situation.

Meanwhile, Almond [ALMO,88] used the idea of state partition [SHOR,81] and combinatorics [JAYN,68] to derive a ME distribution for the joint state probabilties of a G/G/1 FCFS queue with multiple classesof jobs. The results obtained are used as a basis for the approximate analysis of general closed queueing networks with FCFS queues, single server and multiple class of jobs by operating directly on the classes.

The analysis is summarized as follows:

Consider a G/G/1 queueing system with R classes of customers arriving arbitrarly from R independent external sources and served according to a general distribution in FCFS fashion.

Let's define S as the state of the system belonging to the state space Q. Each state S designates the number of jobs in the system together with their arrangement in the queue.

Given that the sum of all probabilities must add to one, we must have

$$\sum_{S \epsilon Q} P(S) = 1 \qquad (3.15)$$

It is assumed that the marginal utilisations and mean queue lengths are known to exist and are presented via the following constraints:

-Utilisation constraint

$$\sum_{S \epsilon Q} h_r(S)P(S) = \rho_r, \text{ for } r=1,\ldots,R \qquad (3.16)$$

where $h_r(S) = \begin{cases} 1 & \text{if class-r job is receiving service} \\ 0 & \text{otherwise} \end{cases}$

- Mean queue length constraint

$$\sum_{S \epsilon Q} n_r P(S) = \langle n_r \rangle, \text{ for } r=1,\ldots,R \qquad (3.17)$$

From equation (3.7), the ME solution of the state probabilities is given by:

$$P(S) = \frac{1}{Z} \prod_{r=1}^{R} g_r^{h_r(S)} x_r^{n_r}$$

To derive the ME distribution for the joint queue length distribution; all possible arrangements of the jobs in the queue must be taken into account. After some manipulations, the ME joint queue length distribution is given by the following corollary:

Corollary 3.3

The ME solution of the joint steady state queue length distribution of a stable FCFS G/G/1 queue with R classes of jobs, subject to normalisation, mean queue length and utilisation is given by:

$$P(\underline{n}) = \begin{cases} 1-\rho & \text{for } \underline{n} = \underline{0} \\ (1-\rho) \dfrac{\left[\sum_{r=1}^{R} n_r - 1 \right] !}{\prod_{r=1}^{R} n_r !} \prod_{r=1}^{R} (x_r)^{n_r \sum_{r=1}^{R} n_r g_r} & \text{for } \underline{n} \neq \underline{0} \end{cases} \qquad (3.18)$$

where

$$x_r = \frac{\langle n_r \rangle - \rho_r}{\langle n \rangle} \quad , r=1,2\ldots,R \qquad (3.19)$$

with $\langle n \rangle = \sum_{r=1}^{R} \langle n_r \rangle$, $\rho = \sum_{r=1}^{R} \rho_r$

and

$$g_r = \frac{\rho_r}{\langle n_r \rangle - \rho_r} \quad \frac{\rho}{1-\rho} \quad , r=1,\ldots,R \qquad (3.20)$$

#

The ME distribution (3.18) may also be expressed remarkably in the following recursive form:

$$P(\underline{n}) = \begin{cases} 1-\rho & \text{for } \underline{n} = \underline{0} \\ \\ \sum_{r=1}^{R} x_r \, P(\underline{n} - \underline{1}_r) & \text{for } n_r > 0 \end{cases} \qquad (3.21)$$

Equation (3.21) constitutes the basis of an efficient implementation of the ME solution of a general QNM's with FCFS queues.

The marginal probabilities of class-r jobs are obtained by appropriate summation of the ME joint queue length distribution (3.18) and are given by the following expressions:

$$P_r(n_r) = \begin{cases} 1 - \theta_r & \text{for } n_r = 0 \\ \\ \theta_r(1 - \hat{x}_r)\hat{x}_r^{n_r-1} & \text{for } n_r > o \end{cases} \qquad (3.22)$$

where
$$\theta_r = \frac{\rho \, \langle n_r \rangle}{\langle n_r \rangle + \rho - \rho_r} \quad , r=1,2\ldots,R, \qquad (3.23)$$

and
$$\hat{x}_r = \frac{\langle n_r \rangle - \rho_r}{\langle n_r \rangle + \rho - \rho_r} \quad , r=1,2..,R, \qquad (3.24)$$

Note that the analytic expression of the ME marginal queue length distribution (eq.3.22) is analogous to the ME distribution of an ordinary G/G/1 queue (c.f., [KOUV,88a]).

Although the ME distributions expressed above are initially given only for FCFS servers in [ALMO,88], they are also ME solutions of

G/G/1 queues under LCFS, LCFS without preemption and PS disciplines, since these service disciplines do not discriminate the jobs in the basis of their class membership. As a consequence, the ordering of jobs in the queue is taken into consideration in the same manner as in FCFS case.

Note that because under different disciplines (FCFS, LCFS with or without preemption, PS), different values of mean queue lengths are generated for the same queueing parameters, the ME joint queue length distribution, although given by the same expression, its value differs for various disciplines.

It is also easy to verify that the exact product-form solution [BASK,75] are obtained by appropriate substitutions (i.e., for LCFS or PS with general service time or FCFS with class independent exponential server, the joint steady state queue length distribution is given by:

$$P(\underline{n}) = (1-\rho)\frac{\left[\sum_{r=1}^{R} n_r\right]!}{n_1!\ldots n_R!}\prod_{r=1}^{R}\rho_r^{n_r} \qquad)$$

It is interresting to point out, that whatever the system to be analysed, the ME distribution exhibits a product form of factors which are functions of the Lagrangian multipliers corresponding to the constraints imposed on the distribution. The product-form feature provided by the ME formalism plays a key role in a development and an easy implementation of the ME approximations of general QNM's with FCFS centres [ALMO,88; KOUV,85,86b; WALS 84]. Moreover, as shown in some particular cases of corollaries 3.1, 3.2, and 3.3, the PME is a methodology which provides an alternative way to derive some exact results based on information theoretical approach rather than the

usual stochastic one. The well-known queue length probability distribution of a M/M/1 queue are obtained without supposing a birth-and-death stochastic process.

The above mentioned ME solutions for general QNM's have been implemented computationally by making use of the generalised exponential (GE) distribution in approximating the interarrival and service times. This distribution represents a versatile and robust tool in solving single queue serving as a 'building block' for the analysis of general QNM's. This distribution is reviewed below.

3.4 The generalised exponential (GE) distributional model.

Consider a M/G/1 queue with a single-class of jobs, the ME queue length distribution is of geometric type, when only normalisation and mean queue length constraints are used [SHOR,82]. This distribution is considered as a ME approximation for the M/G/1 queue length distribution, and turns out to be exact when the service conforms to an exponential distribution (G ≡ M). However, if additionally, the utilisation constraint is used, the ME queue length distribution is of modified geometric type [EL-AF,83]. In this case, the following corollary holds:

Corollary 3.4 [EL-AF,83]

The ME solution (3.11) - (3.13) is equivalent to the equilibrium solution of an M/G/1 queueing system with a service time density function of the form

$$f_s(t) = (1-\tau)u_0(t) + \tau^2\mu e^{-\tau\mu t} \quad , \quad t \geq 0 \qquad (3.26)$$

or with probability distribution function

$$F_s(t) = 1 - \tau e^{-\tau\mu t} \quad , \quad t \geq 0 \qquad (3.27)$$

where $$\tau = \frac{2}{1 + C_s^2} \qquad (3.28)$$

and $$u_0(t) = \begin{cases} \infty & t = 0 \\ 0 & t \neq 0 \end{cases}$$

such that $$\int_{-\infty}^{+\infty} u_0(t)\, dt = 1 \quad .$$ #

$u_0(t)$ is known as the unit impulse function [KLEI,75, pp.342] which creates a jump at the origin and subsequently makes the service time a mixed r.v (i.e., continuous variable with a non-zero probability at the origin $F_s(0) = 1 - \tau$).

Note that for $C_s^2 = 1$, we have GE ≡ M.

The L.S.T of the GE distribution with parameters μ and C_s^2 (GE(μ,C_s^2)) is given by

$$F_s^*(\theta) = 1 - \tau + \tau \frac{\tau\mu}{\tau\mu + \theta} \qquad (3.29)$$

where τ is given by (eq. 3.28).

Since the term $\tau\mu/(\tau\mu + \theta)$ represents the L.S.T of the exponential distribution with parameter ($\tau\mu$), the GE distribution may then be considered as a phase-type distribution [KLEI,75,pp.141] with a possibility of a null inter-event (service) time (see Fig.3.1).

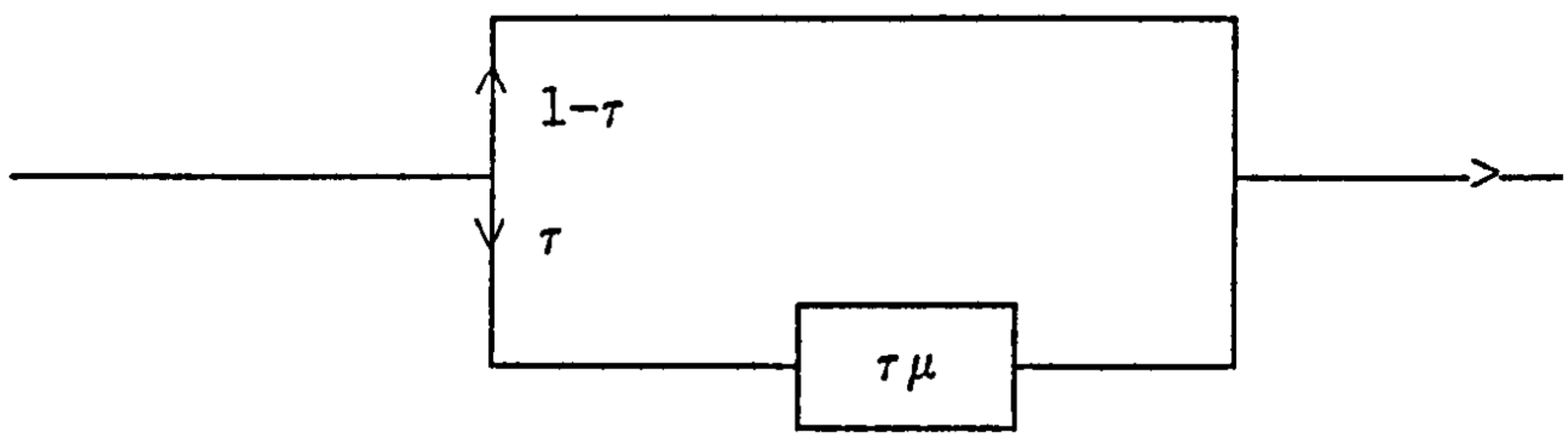

Fig.3.1 GE distributional model

It can be shown by successive differentiation of the L.S.T (3.29), that the moments are given by the following expression:

$$\langle S^n \rangle = \frac{n!}{\mu^n \tau^{n-1}} \qquad (3.30)$$

3.4.1 Properties

Property 3.1 (pseudo-memoryless property)

Given a random variable $\tilde{X}$ (service time) conforming to GE(μ,τ). The remaining service time $\hat{X}$ is distributed exponentially with parameter ($\tau\mu$).

#

Proof

The proof is based on the analytical form of the GE probability distribution function.

Given t, t_0 non-negative reals, clearly we have:

$$\text{Prob}[\ \hat{X} > t\ ] = \text{Prob}[\ \tilde{X} > t+t_0 \ / \ \tilde{X} > t_0\ ]$$

$$= \frac{\text{Prob}[\ \tilde{X} > t+t_0\ ,\ \tilde{X} > t_0\ ]}{\text{Prob}[\ \tilde{X} > t_0\ ]}$$

$$= \frac{\text{Prob}[\ \tilde{X} > t+t_0\ ]}{\text{Prob}[\tilde{X} > t_0\ ]}$$

$$= \frac{\tau e^{-\tau\mu(t+t_0)}}{\tau e^{-\tau\mu t_0}}$$

$$= e^{-\tau\mu t}$$

Taking the complementary probability, leads to

$$\text{Prob}[\ \hat{X} \leqslant t\ ] = 1 - e^{-\tau\mu t} \qquad \text{Q.E.D}$$

property 3.2

Given two independent r.v's X_1 and X_2 conforming to exponential distributions with parameters λ_1 and λ_2, respectively ($\lambda_1 > \lambda_2$). The r.v. X_0 which satisfies the relation $X_2 = X_1 \oplus X_0$, is GE distributed with mean τ/λ_2 and squared coefficient of variation $C^2 = (2-\tau)/\tau$ where $\tau = (\lambda_1-\lambda_2)/\lambda_1$.

Where $\oplus$ is the convolution operator.

#

Proof

Since L.S.T of the convolution of two independent r.v's is the product of L.S.T's, we have:

$$F^*_{X_2}(\theta) = F^*_{X_1}(\theta)\ F^*_{X_0}(\theta)$$

Substituing the L.S.T of the exponential distribution in the above equation, we will have:

$$\frac{\lambda_2}{\lambda_2 + \theta} = \frac{\lambda_1}{\lambda_1 + \theta}\ F^*_{X_0}(\theta)$$

$$\Longleftrightarrow \qquad F^*_{X_0}(\theta) = \frac{\lambda_2(\lambda_1 + \theta)}{\lambda_1(\lambda_2 + \theta)}$$

After some simple calculations, yields

$$F^*_{X_0}(\theta) = 1 - \frac{\lambda_1 - \lambda_2}{\lambda_1} + \frac{\lambda_1 - \lambda_2}{\lambda_1}\ \frac{\lambda_2}{\lambda_2 + \theta}$$

Clearly we recognize the L.S.T of a GE($\alpha\lambda_2, \tau=\frac{\lambda_1-\lambda_2}{\lambda_1}$). Q.E.D

3.4.2 Physical interpretation of GE

As mentioned earlier, the GE distribution is a phase-type distribution consisting of an exponential and null branches and where the selection of the branches is a Bernoulli trial. Therefore, it may also be considered as a hyperexponential-2 (H_2) (Fig.3.2) distribution where one of the stages have zero inter-event time.

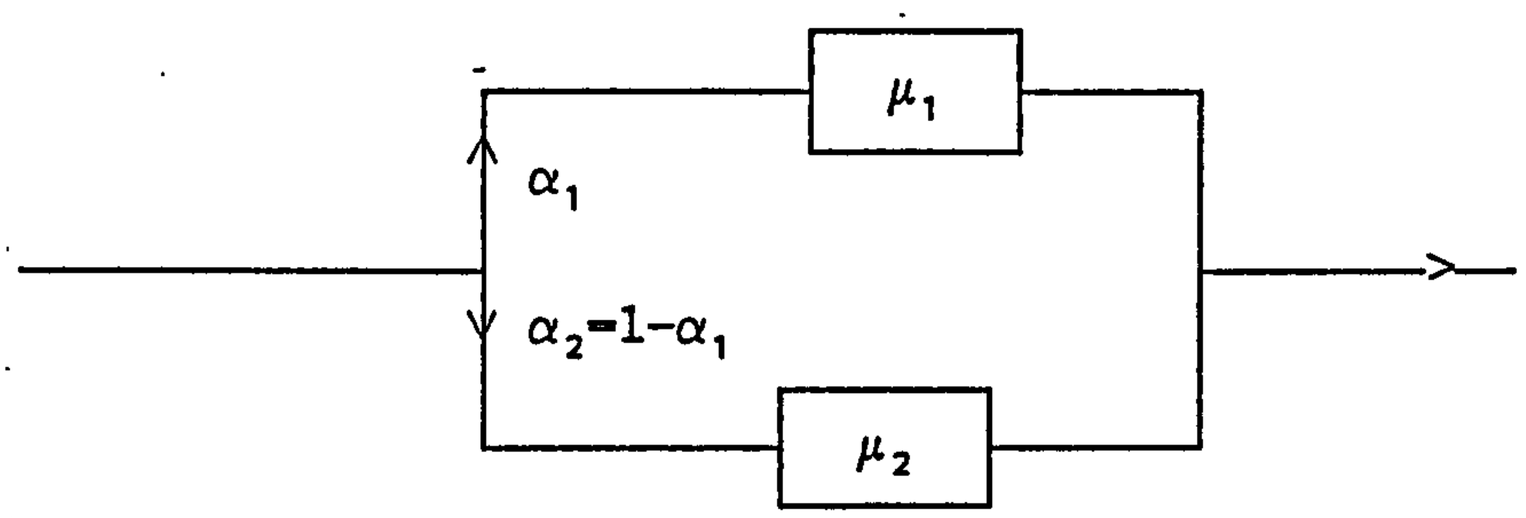

Fig.3.2 H_2 distributional model

On the other hand, the GE server may allow some jobs receiving zero service time (go without receiving service); which means that a 'bulk' of jobs may leave the service at the same time as the one leaving the exponential server.

These two aspects are discussed in the two following subsections.

3.4.2.1 Limiting interpretation of GE [EL-AF,83]

Consider a general distribution with mean μ^{-1} and squared coefficient of variation ($C^2 > 1$). The problem is to determine a two-phase type distribution (H_2) which satisfies the first two moments estimated usually from measurements. In order words, we want

to determine the parameters α_1, α_2, μ_1, μ_2 satisfying the following equations:

$$\frac{\alpha_1}{\mu_1} + \frac{\alpha_2}{\mu_2} = \frac{1}{\mu}$$

$$\frac{2\left\{\frac{\alpha_1}{\mu_1^2} + \frac{\alpha_2}{\mu_2^2}\right\}}{\left\{\frac{\alpha_1}{\mu_1} + \frac{\alpha_2}{\mu_2}\right\}^2} - 1 = C^2$$

The solution of a system of 2 equations and 3 unknowns have generally infinite number of solutions. However, the solutions of the system above form a 'family of solutions' which depends on 'tuning' parameter $\kappa \in]1, +\infty[$ and are given by

$$\alpha_1(\kappa) = \frac{A + B}{C^2 + 1} \qquad (3.31)$$

$$\alpha_2(\kappa) = 1 - \alpha_1(\kappa) \qquad (3.32)$$

$$\mu_1(\kappa) = \kappa\alpha_1(\kappa)\mu \qquad (3.33)$$

$$\mu_2(\kappa) = \frac{\kappa}{\kappa-1}\alpha_2(\kappa)\mu \qquad (3.34)$$

where $$A = \frac{C^2-1}{2} + \frac{2}{\kappa}$$

and
$$B = \frac{1}{2}\left\{ (C^2-1)^2 + \frac{8(C^2-1)}{\kappa} + \frac{8(1-C^2)}{\kappa^2} \right\}^2 .$$

Observing that
$$\lim_{\kappa \to +\infty} \alpha_1(\kappa) = \frac{C^2 - 1}{C^2 + 1}$$

$$\lim_{\kappa \to +\infty} \alpha_2(\kappa) = \frac{2}{C^2 + 1}$$

and $\lim_{\kappa \to +\infty} \mu_1(\kappa) = +\infty$, $\lim_{\kappa \to +\infty} \mu_2(\kappa) = \frac{2}{C^2 + 1}\mu$

Notice that the limiting expressions of $\alpha_2(\kappa)$ and $\mu_2(\kappa)$ when κ goes to infinity, are the parameters of $GE(\mu_2(+\infty), \alpha_2(+\infty))$. Hence, the GE may be interpreted as the limiting case of H_2 when the tuning parameter 'κ' goes to infinity.

Note that for $C^2 < 1$, although the branching probability of GE server is improper ($\tau > 1$), the GE distribution is robust enough to model two-phase type distributions (hypoexponential-2 or h_2) [KOUV,88a; SAUE,75a].

It was shown in [KOUV,88a], that the solution obtained by solving, for example, $H_2/H_2/1$ as $\kappa \to +\infty$, or $h_2/h_2/1$ as $\kappa \to -\infty$ by global balance technique, is identical to GE/GE/1 and corresponds to a ME solution under normalization, utilisation, and mean queue length constraints.

Consequently, the GE distribution is the only phase-type distribution uniquely and completely specified by its first two moments. In an information theoretic context, the exponential (M) is the best supported distribution when only the first moment (mean inter-event time) is available, whereas the GE is the best supported distribution when the first two moments (mean inter-event time, variance) are given.

3.4.2.2 Bulk interpretation of GE

Consider a renewal process where the inter-event time such as the interarrival time to a queueing system conforms to $GE(\lambda, C_a^2)$.

Let N(t) be a r.v. that counts the arrivals occuring during the interval [0,t]. The objective here, is to determine the distribution of the underlying counting process N(t).

Theorem 3.1

The underlying counting process of a GE renewal process with parameters λ and $\sigma = 2/(C_a^2+1)$ is given by a compound Poisson process given by:

$$\text{Prob}[N(t) = n] = \begin{cases} e^{-\sigma\lambda t} & \text{for } n = 0 \\ \sum_{k=1}^{n} \frac{(\sigma\lambda t)^k}{k!} e^{-\sigma\lambda t} \binom{n-1}{k-1} \sigma^k (1-\sigma)^{n-k} & n \geq 1 \end{cases} \quad (3.35)$$

#

The proof is based on renewal theory arguments and can be seen in appendix B (section B1). Note also, that the correspondance between GE distribution and the compound Poisson process has also been mentioned by Whitt [WHIT,82].

Moreover, the bulk size distribution is determined via the following corollary:

Corrolary 3.5

The bulk size $\tilde{B}$ of the compound Poisson process corresponding to $GE(\lambda, C_a^2)$ is geometrically distributed with parameter $\sigma = 2/(C_a^2+1)$

(i.e., $\text{Prob}[\tilde{B} = k] = \sigma(1-\sigma)^{k-1}$)

#

The proof follows from theorem 3.1 and can be seen in appendix B (section B2).

In conclusion on the bulk interpretation of GE, each $GE(\lambda,\sigma=2/(C_a^2+1))$ renewal process can be regarded as a bulk distribution ($M^{[B]}$), where the inter-events between bulks are exponentially distributed with parameter $\sigma\lambda$ and the bulk size conforming to a geometric distribution with parameter σ.

This correspondance to a compound Poisson process has been used extensively to derive very important and useful analytic results. A special application to a GE/G/1 queue is given in the next section.

3.4.3 GE/G/1 queue

The mean queue length in GE/G/1 queue with a single class of jobs is considered as a generalisation of the Pollaczeck-Kinchin formula [KLEI,75,pp.187] relative to queues with Poisson bulk arrivals and is given by the following corollary:

Corrolary 3.6

The mean queue length in GE/G/1 queue with a single class of customers and FCFS scheduling discipline is given by

$$\langle n\rangle = \frac{\rho}{2}\left[\,1+\frac{C_a^2+\rho C_s^2}{1-\rho}\,\right] \qquad (3.36)$$

#

Proof

Using the mean queue length of a $M^{[B]}/G/1$ queue with general bulk size distribution [KLEI,75,pp.235], which is given by

$$\langle n\rangle = \rho + \frac{\rho^2(C_s^2+1) + \frac{\lambda_b}{\mu}\left[\overline{C_b^2}+1-\frac{1}{\langle b\rangle}\right]\langle b\rangle}{1-\rho}$$

where <b> is the mean bulk size, C_b^2 is the squared coefficient of variation of the bulk size distribution, and λ_b is the mean arrival rate of the bulks. Since the interarrival times are GE distributed with parameter (λ, C_a^2) , we do have:

$$<b> = \frac{1}{\sigma} = \frac{C_a^2+1}{2}, \quad C_b^2 = 1-\sigma = \frac{C_a^2-1}{C_a^2+1}$$

Substituing <b> and C_b^2 in the expression of the mean queue length, equation (3.36) follows.

Q E.D

Equation (3.36) is first obtained in [KOUV,83] by using the spectral methods to solve Lindley's integral equation [KLEI,75,pp.275-299]

Furthermore, if a G/G/1 queue is a node in a network of queues, it is of extreme importance to estimate the interdeparture time distribution or at least some of its moments. The interdeparture times in G/G/1 queue are generally not renewals with the exception of the M/M/1 queue [BURK,56]. However, when the interarrival and service time distributions have rational L.S.T, as in the case of GE, the first two moments can be determined easily.

For instance, at equilibrium we have:

$$\lambda_a = \lambda_d$$

where λ_a and λ_d are the mean arrival and departure rates, respectively.

The second moment or equivalently the squared coefficient of variation of the interdeparture time in GE/GE/1 queue is given by the following corollary:

Corollary 3.7 [KOUV,83]

The squared coefficient of variation of the interdepature time

distribution in GE/GE/1 queue is given by

$$C_d^2 = \rho(1-\rho) + (1-\rho)C_a^2 + \rho^2 C_s^2 \qquad (3.37)$$

#

See proof in appendix B (section B3).

For the multiple class case and based on the bulk interpretation of GE as a $M^{[B]}$ process, Georgatsos [GEOR,89] derived the marginal mean queue length of each class in GE/G/1 queue under the FCFS, LCFS-PR, LCFS-NONPR and PS scheduling discplines. The expressions of the mean queue lengths are given by the following corollary:

Corollary 3.8

The marginal mean queue length, $\langle n_r \rangle$, r=1,2,...,R of a stable GE/G/1 queue with R classes of jobs are given by

a/ For FCFS discipline

$$\langle n_r \rangle = \frac{\rho_r}{2}(C_a^2 + 1) + \frac{\sum_{\ell=1}^{R} \frac{\lambda_r}{\lambda_\ell} \rho_\ell^2 (C_{a\ell}^2 + C_{s\ell}^2)}{2(1-\rho)}, \qquad (3.38)$$

b/ For LCFS-PR discipline

$$\langle n_r \rangle = \rho_r \frac{C_{ar}^2 + 1}{2(1-\rho)} \qquad (3.39)$$

c/ For LCFS-NONPR discipline

$$\langle n_r \rangle = \rho_r \frac{C_a^2 + 1 - 2\rho}{2(1-\rho)} + \frac{\sum_{\ell=1}^{R} \frac{\lambda_r}{\lambda_\ell} \rho_\ell^2 (C_{s_\ell}^2 + 1)}{2(1-\rho)} \qquad (3.40)$$

d/ For PS discipline

$$\langle n_r \rangle = \rho_r \left[C_{ar}^2 + \frac{\sum_{\ell=1}^{R} \rho_\ell C_{a\ell}^2}{(1-\rho)} \right] \qquad (3.41)$$

#

3.5 Conclusion

In this chapter we have reviewed the PME and have seen that the application of the PME to queueing systems such as a G/G/1 queue can not only provide some known classical results (i.e., queue length distribution of a M/M/1 queue), but also give analytic and closed-form expressions for the approximate solutions of more general queueing systems (i.e., G/G/1 queue with multiple classes and general QNM's). It has also been mentioned that the ME solution gives the largest and the least biased probabilistic model, treating all the possible states of the system as uniformly as possible, subject to the constraints imposed on the distribution.

Moreover, the ME methods advocates a decomposition of general open networks into individual G/G/1 queues. As a consequence, the ME solution of an isolated G/G/1 queue constitutes the building block for the analysis of QNM's. For general closed networks, the Lagrangian coefficients of the ME solution are evaluated analytically as in the open network case by considering a pseudo open network which has the same configuration as the original closed one, satisfying the principles of conservation of flow and fixed population mean.

In information theoretic context, the exponential distribution has been found to be the best candidate in modelling general

distributions when only the first moment is available, whereas the GE is the best supported distribution when the first and the second moments are taken into consideration.

The GE distribution has been proved to be equivalent to a compound Poisson process or as a limiting phase-type distribution when the tuning parameter 'κ' goes to infinity. This important property of the GE are used to generalize present results (restricted only to pure Poisson process), i.e., queueing system with blocking [XENI,89] decomposition methods of queueing networks [TOMA,89], queueing systems with load dependent routing [GEOR,89] and application to communication networks with window flow control [OTHM,88].

The GE distribution is versatile since it has a pseudo memoryless property which provides a mathematical tractability comparable to that of an exponential distribution. In the next chapter, exact analysis of a GE/G/1 queue under either PR or HOL scheduling discipline is carried out. The results obtained constitute a generalisation to the known M/G/1 formulae [JAIS,68]. The mean value analytic expressions derived are used in turn as prior variables when applying the PME to the above mentioned queues.

CHAPTER 4

The GE/G/1 Priority Queue

4.1 Introduction

This chapter focuses on the PR and HOL GE/G/1 priority queues and carries out exact analysis in order to estabish new analytic approximations for $\langle n_r \rangle$ and $P_r(0)$, r=1,2,...,R, statistics of PR and HOL G/G/1 priority queues, respectively. These statistics as in the case of some Markovian type priority queues (e.g., M/M/1, M/G/1), may be obtained via stochastic arguments at the equilibrium without the prior knowledge of the entire state probability distribution [JAIS,68]. The GE-type analytic expressions for $\langle n_r \rangle$ and $P_r(0)$, r=1,2,...,R, can be used as part of the set of mean value constraints in order to expedite the analytic approximations of the ME solutions of a stable PR and HOL G/G/1 priority queue in the next chapter.

To this end, new exact analytic results for the completion time, busy period, occupation time and response time distributions per priority class are derived for a stable PR and HOL GE/G/1 queue in sections two, three, four and five, respectively. The proofs are generally based on the bulk interpretation of the GE distributional model (c.f. theorem 3.1) having an underlying compound Poisson process (M^B) with geometrically distributed bulk sizes.

In section six, generating functions for the marginal queue length distributions, $Q_r(.)$, r=1,2,...,R, as well as marginal mean queue lengths and idle state probabilities, $P_r(0)$, are derived. A stepwise presentation of the evaluation of $P_r(0)$ for GE service-time, is given at the end of the section.

In section seven, we investigate Kleinrock's conservation law [KLEI,76, pp.117], and suggest new relating equation for a stable and work-conserving (c.f. section 2.2.2) GE/G/1 queue under any non-preemptive scheduling discipline such as HOL rule.

To assess the robustness of the GE-type formulae (mean queue lengths and idle state probabilities), numerical validation examples are displayed in section eight, where comparison against simulations, involving different interarrival-time and service-time distributions per class, is carried out. Furthermore, some useful performance bounds are defined for the marginal mean queue lengths.

Lastly in section 4.9, we conclude this chapter with a brief summary of the results obtained.

4.2 Completion time distribution

The completion time of class-r jobs, r=1,2,...,R, is defined as the time elapsed between the instant a class-r job starts receiving service, until the time the server is alocated to another job of the same class, if any, in the system [JAIS,68,pp.56]. Because PR and HOL are work-conserving disciplines, the following results hold for both of them [JAIS,68,pp.145].

Corollary 4.1

For a stable GE/G/1 queue under either PR or HOL scheduling discipline with R priority classes, the L.S.T of the completion time of class-r, r=1,2,...,R is given by :

$$C_r^*(\theta) = S_r^*[\theta + \Lambda_{r-1}^{(b)}(1-G_{r-1}^*(\theta))] \quad , \qquad (4.1)$$

$$\text{where } \Lambda_{r-1}^{(b)} = \sum_{s=1}^{r-1} \lambda_s \frac{2}{C\hat{a}_s^2+1}$$

and $S_r^*[\,.\,]$, $G_{r-1}^*(.)$ are the L.S.T of the service time of class-r jobs and the L.S.T of the busy period generated by jobs of classes $\{1,2,\ldots,r-1\}$, respectively.

#

Proof

For the sake of simplicity, let's consider a GE/G/1 PR queue. The completion time of class-r jobs starts with the beginning of service of class-r jobs and ends when this job leaves definit e ly the system. In those circumstances, when a class-r job is receiving service, it can only be preempted by higher priority jobs and its service is interrupted during a complete busy period generated by the r-1 higher priority class jobs.

Let $\tilde{S}_r$ and $\tilde{N}_s'$, s=1,...,r-1 be the random variables (r.v.s) describing the service-time of class-r jobs and the number of bulks of higher priority class jobs arriving during $\tilde{S}_r$. Using the delay cycle methodology [KLEI,76,PP.110-113] and Conditioning on the above r.v's., the L.S.T of the completion time of class-r is then given by

$$C_r^*(\theta)\,|_{\tilde{S}_r=t,N_s'=n_s,s=1,\ldots r-1} = e^{-t\theta}[\,G_{r-1}^*(\theta)\,]^n \qquad (4.1a)$$

$$\text{where } n = \sum_{s=1}^{r-1} n_s$$

Given that the interarrival-time distribution of class-s, s=1,...,r-1, jobs conform to $GE(\lambda_s, C_{as}^2)$, class-s jobs arrive in a compound Poisson process with parameter $\lambda_s^{(b)} = 2\lambda_s/(C_{as}^2+1)$ (c.f. theorem 3.1). Subsequently, the r.v. $N'^{(r-1)}=N_1'+N_2'+\ldots N_{r-1}'$ satisfies

the Poisson distribution with parameter, $\Lambda_{r-1}^{(b)}$, given by

$$\Lambda_{r-1}^{(b)} = \sum_{s=1}^{r-1} \lambda_s^{(b)}$$

Applying the law of total L.S.T to equation (4.1a), we obtain:

$$C_r^*(\theta)_{|\tilde{S}_r=t} = e^{-t\theta} \sum_{n=0}^{\infty} e^{-\Lambda_{r-1}^{(b)}t} \frac{(\Lambda_{r-1}^{(b)}t)^n}{n!} [G_{r-1}^*(\theta)]^n$$

$$= e^{-(\theta+\Lambda_{r-1}^{(b)})t} \; e[\Lambda_{r-1}^{(b)} G_{r-1}^*(\theta)t]$$

$$= e^{-[\theta+\Lambda_{r-1}^{(b)}(1-G_{r-1}^*(\theta))]t}$$

Finally, Applying the law of total L.S.T to the equation above, we end up with

$$C_r^*(\theta) = \int_{t=0}^{\infty} e^{-[\theta+\Lambda_{r-1}^{(b)}(1-G_{r-1}^*(\theta))]t} dS_r(t)$$

$$= S_r^*[\theta+\Lambda_{r-1}^{(b)}(1-G_{r-1}^*(\theta))]$$

Q.E.D.

Moreover, it is also important to determine the completion time of all members of an arriving bulk; the quantity that is used in the derivation of the busy period.

If $\tilde{B}_r$ denotes the r.v. describing the bulk size of an arriving bulk of class-r, the conditional L.S.T of the completion time of all members of an arriving bulk is just the product of the L.S.T of the individual completion times and given by:

$$C_r^{*(b)}(\theta)\big|_{\tilde{B}_r=n} = [C_r^*(\theta)]^n$$

Using the Law of total L.S.T, we obtain

$$C_r^{*(b)}(\theta) = \sum_{n=0}^{\infty} \text{Prob}[\tilde{B}_r=n][C_r^*(\theta)]^n$$

$$C_r^{*(b)}(\theta) = q_r[C_r^*(\theta)] ,$$

where $q_r[.]$ is the generating function of the bulk size distribution of class-r.

Since the interarrival-time process is GE distributed, the bulk size conforms to a geometric distribution with parameter $\sigma_r=2/(C\hat{a}_r^2+1)$ (c.f. corollary 3.5). Therefore, using the generating function of the geometric distribution [TRIV,82,pp.579], the L.S.T of the completion time of all members of an arriving bulk is then given by:

$$C_r^{*(b)}(\theta) = \frac{2C_r^*(\theta)}{(C\hat{a}_r^2+1) - (C\hat{a}_r^2-1)C_r^*(\theta)} , \qquad (4.2)$$

where $C_r^*(.)$ is given by (4.1).

4.3 Busy period distribution

The busy period is the time during which the server is busy serving customers in the system without entering the idle state. This entity is of interest from the server's point of view. For a queueing system containing R different classes of customers, with or without priorities, one would like to investigate a busy period process generated by customers belonging to class-s with index $s\epsilon\{1,2,..,r\}$,

for r =1,2,...,R (i.e. this is the sub-busy period during which the server is busy serving jobs of classes 1,2,...,r). The corresponding results are of extreme importance in studying priority queues (e.g., in calculating the idle state probability $P_r(0)$ per class-r, r=1,2..,R).

In the case of a G/G/1 queue, the busy period initiated by job classes {1,2,..,r}, under PR discipline is identical to the one initiated by the same jobs under HOL rule due to the fact that both disciplines are work-conserving [JAIS,68,pp.147]. A special analysis for GE interarrival-time process per class is presented in the two following sub-sections.

4.3.1 GE/G/1 with a single class of customers (r=1)

Let us assume that jobs arrive to the system according to GE distribution with mean arrival rate λ_1 and squared coefficient of variation $C_{a_1}^2$ and are served according to an arbitrary distribution with μ_1 and $C_{s_1}^2$ as mean service rate and squared coefficient of variation, respectively.

Using theorem 3.1, the $GE(\lambda_1, C_{a_1}^2)/G/1$ queue is equivalent to a $M^B/G/1$ where the mean bulk arrival rate is $\lambda_1^{(b)}=2\lambda_1/(C_{a_1}^2+1)$ and the bulk size, $\tilde{B}_1$, is geometrically distributed with mean bulk size $<b_1>=1/\sigma_1=(C_{a_1}^2+1)/2$ (c.f. corollary 3.5).

Let $\tilde{S}_{b_1}$ be the r.v. describing the service time of all members of an arriving bulk.

Conditioning on the bulk size $\tilde{B}_1$, the L.S.T of the r.v. $\tilde{S}_{b_1}$ is :

$$S_1^{*(b)}\Big|_{\tilde{b}_1=n'} = [S_1^*(\theta)]^{n'} \quad ,$$

where $S_1^*(.)$ is the L.S.T of the service time of the individual jobs.

Using the law of total L.S.T as before, the L.S.T of $\tilde{S}_{b_1}$ is then given by:

$$S_1^{*(b)}(\theta) = \sum_{n'=0}^{\infty} [S_1^*(\theta)]^{n'} \text{ Prob}[\tilde{B}_1 = n']$$

$$= q_1[S_1^*(\theta)]$$

where $q_1(.)$ is the generating function of the bulk size distribution. Therefore, by analogy to equation (4.2), the L.S.T of the service time of all members of the arriving bulk is given by:

$$S_1^{*(b)}(\theta) = \frac{2S_1^*(\theta)}{(C\tilde{a}_1^2+1) - (C\tilde{a}_1^2-1)S_1^*(\theta)} \qquad (4.3)$$

From the server point of the view and as far as the busy period is concerned, the system is behaving as an ordinary M/G/1 queue with mean arrival rate $\lambda_1^{(b)}$ and an elongated service time, $\tilde{S}_{b_1}$.

Conditioning on $\tilde{S}_{b_1}$ and the number of bulk of class-1, $\tilde{N}'$, arriving during $\tilde{S}_{b_1}$, the L.S.T of the busy period is then given by:

$$G_1^*(\theta)_{|\tilde{S}_{b_1}=t,\tilde{N}'=n} = e^{-\theta t}[G_1^*(\theta)]^n$$

Since the bulks arrive in Poisson fashion, we will have:

$$G_1^*(\theta)_{|\tilde{S}_{b_1}=t} = \sum_{n=0}^{\infty} e^{-\theta t}[G_1^*(\theta)]^n \, e^{-\lambda_1^{(b)} t} \frac{(\lambda_1^{(b)} t)^n}{n!}$$

$$= e^{-t(\theta+\lambda_1^{(b)}-\lambda_1^{(b)}G_1^*(\theta))}$$

Finally from the law of total L.S.T, we will have:

$$G_1^*(\theta) = S_1^{*(b)}(\theta+\lambda_1^{(b)}-\lambda_1^{(b)}G_1^*(\theta)) \qquad (4.4)$$

which can also be expressed with respect to the GE parameters as follows:

$$G_1^*(\theta) = S_1^{*(b)}\left[\theta + \frac{2\lambda_1}{C\hat{a}_1^2+1}(1-G_1^*(\theta))\right] \qquad (4.5)$$

If the service time conforms to a $GE(\mu_1, C_{s_1}^2)$ distribution, the L.S.T of the busy period is given by the following corollary:

Corollary 4.2

For a stable GE/GE/1 with a single class of jobs, the L.S.T of the busy period is given by the closed-form expression

$$G_1^*(\theta) = \frac{\Phi - \Delta^{\frac{1}{2}}}{2\omega}, \qquad (4.6)$$

where
$$\Phi = \frac{4}{(C\hat{a}_1^2+1)(C\hat{s}_1^2+1)}\left[\mu_1 + \frac{2C\hat{s}_1^2+C\hat{a}_1^2-1}{C\hat{a}_1^2+1}\lambda_1\right]$$

$$+ 2\frac{C\hat{s}_1^2+C\hat{a}_1^2}{(C\hat{a}_1^2+1)(C\hat{s}_1^2+1)}\theta \qquad (4.6a)$$

$$\Delta = \left[\frac{2}{(C\hat{a}_1^2+1)(C\hat{s}_1^2+1)}\right]^2\left\{\left[2(\lambda_1+\mu_1)+(C\hat{a}_1^2+C\hat{s}_1^2)\theta\right]^2-16\lambda_1\mu_1\right\} \qquad (4.6b)$$

$$\omega = 4 \frac{C_{s_1}^2 + C_{a_1}^2}{(C_{a_1}^2+1)^2(C_{s_1}^2+1)} \lambda_1 \qquad (4.6c)$$

#

Proof

Since the service time is GE distributed, we then have (c.f.section 3.4)

$$S_1^*(\theta) = 1-\tau_1 + \tau_1 \frac{\tau_1\mu_1}{\tau_1\mu_1 + \theta} \text{ , with } \tau_1 = \frac{2}{C_{s_1}^2 + 1}$$

The L.S.T of the service time of all members of the bulk (4.3) is subsequently expressed by:

$$S_1^{*(b)}(\theta) = \frac{\sigma_1[\tau_1\mu_1 + (1-\tau_1)\theta]}{\sigma_1\tau_1\mu_1 + (\tau_1+\sigma_1-\sigma_1\tau_1)\theta}$$

where $\sigma_1 = 2/(C_{a_1}^2+1)$.

Substituting the expression above, in equation (4.4), yields

$$G_1^*(\theta) = \frac{\sigma_1[\tau_1\mu_1 + (1-\tau_1)(\theta+\lambda_1\sigma_1-\lambda_1\sigma_1 G_1^*(\theta))]}{\sigma_1\tau_1\mu_1+(\tau_1+\sigma_1-\sigma_1\tau_1)(\theta+\lambda_1\sigma_1-\lambda_1\sigma_1 G_1^*(\theta))}$$

Solving with respect to $G_1^*(\theta)$, we obtain the following quadratic equation:

$$\lambda_1\sigma_1(\tau_1+\sigma_1-\tau_1\sigma_1)[G_1^*(\theta)]^2$$
$$- \{\tau_1\mu_1\sigma_1+(\tau_1+\sigma_1-\tau_1\sigma_1)\theta+2\lambda_1\sigma_1^2+\tau_1\sigma_1\lambda_1(1-2\sigma_1)\}G_1^*(\theta)$$
$$+ \tau_1\sigma_1\mu_1+\sigma_1(1-\tau_1)(\theta+\lambda_1\sigma_1) = 0 \quad ,$$

or in simpler form $\omega[G_1^*(\theta)]^2 - \Phi[G_1^*(\theta)] + \psi = 0$.

The solutions of such equation have generally two real roots,

$$G_1^*(\theta)^{(1)} = \frac{\Phi - \Delta^{\frac{1}{2}}}{2\omega}, \text{ and } G_1^*(\theta)^{(2)} = \frac{\Phi + \Delta^{\frac{1}{2}}}{2\omega},$$

where $\Delta = \Phi^2 - 4\omega\psi$ and is given by equation (4.6b)

However, the solution that we are looking for, must satisfy $G_1^*(0)=1$, this turns out to be $G_1^*(\theta)^{(1)}(.)$. Finally, substituting the expressions of τ_1 and σ_1 in the final solution, equation (4.6) follows.

Q.E.D

Note that the L.S.T of the M/M/1 busy period is obtained by appropriate substitutions of the arrival and service parameters, ($Ca_i^2 = Cs_i^2 = 1$) (c.f. [KLEI,75,pp.215].

4.3.2 GE/G/1 with r (r > 1) classes of jobs.

Since the busy period of a G/G/1 queue with r, r=2,...,R, classes under both PR or HOL discipline is the same, we restrict our analysis when the system in question is under PR rule.

When r classes of jobs are considered, the busy period $\tilde{G}_r$, can be initiated by any job belonging to classes ℓ, ℓ=1,2,...,r, and the corresponding L.S.T is given by the following corollary:

Corollary 4.3

For a stable GE/G/1 PR or HOL queue with r, r=2,...,R, classes of jobs, the L.S.T of the busy period is given by the following recursive form:

$$G_r^*(\theta) = \frac{\lambda_r^{(b)}}{\Lambda_r^{(b)}} G_{r_1}^*(\theta) + \frac{\Lambda_{r-1}^{(b)}}{\Lambda_r^{(b)}} G_{r_2}^*(\theta) \quad , \qquad (4.7)$$

where $G_{r_1}^*(\theta) = C_r^{*(b)}\left[\theta + \frac{2\lambda_r}{C\hat{a}_r^2 + 1}(1-G_{r_1}^*(\theta))\right]$, (4.7a)

and $G_{r_2}^*(\theta) = G_{r-1}^*\left[\theta + \frac{2\lambda_r}{C\hat{a}_r^2 + 1}(1-G_{r_1}^*(\theta))\right]$, (4.7b)

where $C_r^{*(b)}(.)$ is L.S.T of the completion time of all members of an arriving bulk and given by (4.2), $G_{r-1}^*(.)$ is the L.S.T of the busy period with (r-1) classes of jobs and $\Lambda_r^{(b)}$ is given by

$$\Lambda_r^{(b)} = \sum_{s=1}^{r} \frac{2\lambda_s}{C_{\hat{a}s}^2+1}$$

#

Proof

Let us consider first, two classes of customers, (r=2). The busy period is initiated either by low-priority job (class-2), which will constitute our case-1, or by high-priority job (class-1) which will be the case-2.

i/ Case-1: in this case, and with probability $(\lambda_2^{(b)}/\Lambda_2^{(b)})$ the server leaves its idle state when an arrival of bulk of jobs belonging to class-2 occurs. The busy period lasts until there is no class-1 or class-2 jobs in the system. In fact, and as far as class-2 jobs are concerned, the two-class GE/G/1 queue with PR is equivalent to a single-class GE/G/1 queue with service-time replaced by the completion time of class-2 jobs. Therefore, by analogy to equation (4.5), we will have

$$G_{21}^{*}(\theta) = C_2^{*}(b)\left[\theta + \frac{2\lambda_2}{C\tilde{a}_2^2+1}\,(1-G_{21}^{*}(\theta)\right]$$

where $C_2^{*(b)}(.)$ is L.S.T of the elongated completion time of class-2 jobs and given by equation (4.2).

ii/ Case-2: with probability $(\lambda_1^{(b)}/\Lambda_2^{(b)})$, the server enters the busy period when bulk of class-1 jobs arrives in the system to start service. In this case, class-1 jobs go for service during a complete ordinary busy period (with only class-1 jobs involved) before a low-priority job starts receiving service. In other words, case-2 is equivalent to case-1 with the initial delay equal to the busy period of GE/G/1 where only jobs of class-1 are involved.

if $\tilde{Y}_0$ denotes the initial delay in case-2 and $\tilde{N}_2'$ is the number of class-2 bulks arriving during this initial delay, the conditional L.S.T of the busy period in case-2 is

$$G_{22}^{*}(\theta)_{|\tilde{Y}_0=y_0,\tilde{N}_2'=n} = e^{-y_0\theta}[G_{21}^{*}(\theta)]^n$$

Using the fact that class-2 jobs arrive in Poisson compound fashion and applying the law of total L.S.T as in section 4.2, we will end up with the following expression:

$$G_{22}^{*}(\theta) = G_1^{*}\left[\theta + \frac{2\lambda_2}{C\tilde{a}_2^2+1}G_{21}^{*}(\theta)\right]$$

Thus, taking into account both cases, the L.S.T of the busy period of a GE/G/1 queue with 2 classes of jobs under either PR or HOL discipline is given by:

$$G_2^*(\theta) = \frac{\lambda_2^{(b)}}{\Lambda_2^{(b)}} G_{21}^*(\theta) + \frac{\lambda_1^{(b)}}{\Lambda_2^{(b)}} G_{22}^*(\theta)$$

Finally, for r>2, identify that the bulk arrival rate of higher priority class jobs is $\Lambda_{r-1}^{(b)}$ (instead of $\lambda_1^{(b)}$ for the two-class-case) and by induction, equation (4.7) follows.

Q.E.D

4.4 Occupation (waiting) time distribution

The occupation time, $W_r(t)$, of the server at time t, with respect to class-r jobs is defined as the time the server will remain occupied with jobs of equal or higher priority classes, if the arrival process of the r^{th} class units is stoped at time t [JAIS,68,pp.73]. Because jobs are served in FCFS fashion within their class, $W_r(t)$ represents also the time that class-r job will have to wait before receiving its first burst of service, if it arrives at time t. This waiting time obviously will be different when the queueing system in question is under different scheduling disciplines.

4.4.1 PR discipline

The class-r, r=1,2,...,R, occupation time distribution of a GE/G/1 queue under PR discipline is characterised by the following corollary:

Corollary 4.4

For a stable GE/G/1 queue, with R (≥2) classes of jobs under PR scheduling discipline, the L.S.T of the marginal waiting time distribution of class-r jobs, r=1,2,...,R, is given by

$$W_r^*(\theta) = 2(1-\gamma_r)\frac{\theta + \Lambda_{r-1}^{(b)}(1-G_{r-1}^*(\theta))}{\theta\left[C\hat{a}_r^2+1-(C\hat{a}_r^2-1)C_r^*(\theta)\right] - 2\lambda_r(1-C_r^*(\theta))} \quad (4.8)$$

where $C_r^*(.)$ and $G_{r-1}^*(.)$ are given by equations (4.1) and (4.7), respectively.

#

Proof

Let's first consider a GE/G/1 queue under PR discipline with 2 priority classes (R=2). Because of the preemptions, class-2 jobs do not in any way affect the occupation time of the server with respect to class-1 jobs. Therefore, $\tilde{W}_1$ is identical to the waiting time of a job in a single class GE/G/1 queue (i.e., The waiting time in a $M^B/G/1$ queue [KLEI,75,pp.235]) and the correspoding L.S.T is given by

$$W_1^*(\theta) = (1-\rho_1)\frac{\sigma_1\theta(1-S_1^{*(b)}(\theta))}{\left[\theta - \lambda_1^{(b)}(1-S_1^{*(b)}(\theta))\right](1-S_1^*(\theta))}$$

where $S_1^{*(b)}(.)$ is the L.S.T of the service time of all members of the arriving bulk of class-1 jobs.

Substituting the expression of $S_1^{*(b)}(.)$ that is given by equation (4.3) in the equation above, yields

$$W_1^*(\theta) = 2(1-\rho_1)\frac{\theta}{\theta\left[C\hat{a}_1^2+1-(C\hat{a}_1^2-1)S_1^*(\theta))\right]-2\lambda_1(1-S_1^*(\theta))} \quad (4.8a)$$

The occupation time of the server with respect to class-2 jobs, $\tilde{W}_2$, depends on two cases.

i/ Case 1: With probability $(1-\rho_1)$, the busy period is not initiated by high-priority class jobs. Thus, the occupation time $\tilde{W}_2$ is identical to the one perceived in a single-class GE/G/1 queue with the service time replaced by class-2 completion time.

ii/ Case 2: With probability ρ_1, the busy period is initiated by class-1 job, and subsequently the occupation time, $\tilde{W}_2$ consists of the remainig busy period of the server with respect to class-1 jobs and the occupation time encountered in case 1.

Given the two cases above, the L.S.T of the waiting time distribution of class-2 jobs is then given by:

$$W_2^*(\theta) = (1-\rho_1)W_2^{*(1)}(\theta) + \rho_1 \hat{G}_1^*(\theta)W_2^{*(1)}(\theta) \qquad (4.8b)$$

Since in case 1, $\tilde{W}_2$ is identical to the waiting time of a single-class GE/G/1 queue with the service time replaced by the completion time of class-2 jobs , the L.S.T $W_2^{*(1)}(.)$ can be obtained by analogy to equation (4.8a), where $S_1^*(.)$, C_{a_1}, λ_1 and ρ_1 are replaced by $C_2^*(.)$, C_{a_2}, λ_2 and $\rho_2/(1-\rho_1)$, respectively, and is given by

$$W_2^{*(1)}(\theta) = 2\left[1-\frac{\rho_2}{1-\rho_1}\right]\frac{\theta}{\theta\left[C\hat{a}_2^2+1-(C\hat{a}_2^2-1)C_2^*(\theta)\right]-2\lambda_2(1-C_2^*(\theta))} \qquad (4.8c)$$

Note that $\rho_2/(1-\rho_1)$ designates the utilisation of the server with a single class of jobs and the service time taken to be identical to class-2 completion time and $\hat{G}_1(.)$ represents the L.S.T of the remaining busy period of the actual server with respect class-1 jobs.

From renewal theory [COX,65], it is known that the L.S.T of forward reccurent time (residual) of the busy period in a stable G/G/1 queue is given by

$$\hat{G}_1^*(\theta) = \frac{1 - G_1^*(\theta)}{\theta <G_1>} \qquad (4.8d)$$

where $<G_1>$ is the mean busy period of the server with respect class-1 jobs given by

$$<G_1> = \frac{\rho_1}{\lambda_1 \sigma_1 (1-\rho_1)} \qquad (4.8e)$$

Finally, substituting equations (4.8c), (4.8d) and (4.8e) in the equation (4.8b) leads to

$$W_2^*(\theta) = 2(1-\gamma_2) \frac{\theta + \lambda_1 \sigma_1 (1-G_1^*(\theta))}{\theta [C\hat{a}_2^2 + 1 - (C\hat{a}_2^2 - 1) C_2^*(\theta)] - 2\lambda_2 (1 - C_2^*(\theta))} \qquad (4.8f)$$

For R > 2, under steady state, jobs belonging to class-s, s = r+1,...,R, do not have any influence on the waiting time of class-r jobs. Moreover, the first r classes can be split into two groups; the first one consists of all jobs belonging to classes 1,2,...,r-1, and the second group is just class-r jobs. Therefore, by analogy to the two-class case, if the server is serving jobs belonging to the second group, and a job of the first group (high-priority job) arrives to be served, it interrupts the service of the one currently occupying the server. This interruption lasts as long as there is job of the first group in the system. Obviously, the duration of this interruption is equal to the busy period of the server with respect class-s jobs, $s \in \{1,...,r-1\}$. Thus, by analogy to

the two-class case, equation (4.8) is obtained by replacing

$$\text{i/ } \lambda_1^{(b)} = \lambda_1\sigma_1 \text{ by } \Lambda_{r-1}^{(b)} = \sum_{s=1}^{r-1} \lambda_s\sigma_s \quad ,$$

ii/ the duration of the busy period $\tilde{G}_1$ by $\tilde{G}_{r-1}$,

iii/ the duration of the completion time $\tilde{C}_2$ by $\tilde{C}_r$,

iv/ the parameters γ_2, λ_2, Ca_2 by γ_r, λ_r, Ca_r, respectively.

Q.E.D

4.4.2. HOL discipline

The occupation time of the server with respect to class-r jobs (waiting-time of class-r) in GE/G/1 queue with HOL discipline is characterised by the following corollary:

Corollary 4.5

For a stable GE/G/1 queue, with R (>2) classes of jobs under HOL scheduling discipline, the L.S.T of the marginal waiting time distribution of class-r jobs, r=1,2,...,R, is given by

$$W_r^*(\theta) = 2\frac{(1-\rho)\left[\theta+\Lambda_{r-1}^{(b)}(1-\tilde{G}_{r-1}^*(\theta))\right]+\sum_{\ell=r+1}^{R}\lambda_\ell\left[1-S_\ell^*(\theta+\Lambda_{r-1}^{(b)}(1-G_{r-1}^*(\theta))\right]}{\theta\left[C\hat{a}_r^2+1-(C\hat{a}_r^2-1)C_r^*(\theta)\right] - 2\lambda_r(1-C_r^*(\theta))} \qquad (4.9)$$

#

Proof

Let's first consider, a GE/G/1 HOL queue with 2 (R=2) priority classes. Because the completion time is the same under PR or HOL discpline, the occupation time of the server with respect to class-2

jobs under HOL is the same as under PR and is given by equation (4.8f) [JAIS,68,pp134]. However; the occupation time of the server with respect to high-priority jobs is different that the one encountered under PR rule since those jobs cannot take over the service at their arrival instant if low-priority job is found receiving service.

To determine the occupation time of the server with respect to class-1 jobs, we will use certain similarities between some stochastic processes under PR and HOL rules, respectively. In particular, the occupation time of the server with respect to class-1 jobs, $\tilde{W}_1$, is identical to the one encountered under PR discipline if the busy period of the server and the completion time of class-2 jobs are not initiated at the same time. However, if the completion time of class-2 jobs starts with the beginning of the busy period, the occupation time of the server with respect to class-1 jobs will be defined as the sum of the remaining class-2 service time and the occupation time of the server with respect to class-1 under PR discipline. This is because the first class-1 job in the queue starts its service time immediately after the one in service (class-2) leaves the system.

The corresponding L.S.T is then given by

$$W_1^*(\theta) = \left[1 - \frac{\rho_2}{1-\rho_1}\right] W_1^{*(PR)}(\theta) + \frac{\rho_2}{1-\rho_1} \hat{S}_2^*(\theta) W_1^{*(PR)}(\theta) \quad (4.9a)$$

where $W_1^{*(PR)}(.)$ is given by (4.8a) and $\hat{S}_2^*(.)$ is obtained from renewal theory [COX,65] and given by:

$$\hat{S}_2^*(\theta) = \frac{1 - S_2^*(\theta)}{\theta \langle S_2 \rangle} \quad (4.9b)$$

Therefore, substituting (4.8a) and (4.9b) in equation (4.9a) and after simple manipulations yields

$$W_1^*(\theta) = 2\frac{(1-\rho)\theta + \lambda_2(1-S_2^*(\theta))}{\theta\left[C\tilde{a}_1^2+1-(C\tilde{a}_1^2-1)S_1^*(\theta)\right] - 2\lambda_1(1-S_1^*(\theta))} \qquad (4.9c)$$

In the case of R (>2) classes of jobs, the occupation time of the server with respect to class-r jobs, r=1,2,...,R-1 can not only be affected by the remaining service time of a low-priority job (if any in service), but also by the service time of all high-priority jobs which arrive during that remaining service time. However, if the busy period is initiated by high or equal priority job, the occupation time process of the server with respect to class-r jobs is identical to the one encountered under PR discipline.

If $\tilde{Y}_0$ denotes the time waisted by a class-r, r=1,2,...,R job due to the remaining service time of low priority job, if any, found in service (for example belonging to class-ℓ, ℓ>r). The L.S.T of this r.v. can be determined by using the delay cycle methodology [KLEI,76,pp110-113] (c.f. section 4.2) and given by

$$Y_0^*(\theta) = \hat{S}_\ell^*[\theta+\Lambda_{r-1}^{(b)}(1-G_{r-1}^*(\theta))] \qquad (4.9d)$$

where $\hat{S}_\ell^*(.)$ is the L.S.T of the remaining service time of class-s jobs and given by

$$\hat{S}_\ell^*(\psi) = \frac{1 - S_\ell^*(\psi)}{\psi<S_\ell>} \qquad (4.9e)$$

Hence, by analogy to the L.S.T $W_1^*(.)$ (c.f. equation (4.9a)) where the parameters ρ_2, ρ_1, $S_2^*(.)$ and $W_1^{*(PR)}$ are replaced by $(\rho-\gamma_r)$, γ_r, $Y_0^*(.)$ and $W_r^{*(PR)}(.)$,respectively, the following equation is obtained:

$$W_1^*(\theta) = \left[1 - \frac{\sum_{\ell=r+1}^{R} \rho_\ell}{1-\gamma_r}\right] W_r^{*(PR)}(\theta) + \frac{\sum_{\ell=r+1}^{R} \rho_\ell}{1-\gamma_r} Y_0^*(\theta) W_r^{*(PR)}(\theta) \quad (4.9f)$$

Note that the quantity $\rho_\ell/(1-\gamma_r)$ may represent the utilisation of the server of a G/G/1 queue when the service time is replaced by the completion-time of class-ℓ with jobs belonging to classes $\{r+1,\ldots,S-1\}$ removed from the system (i.e., these jobs do not affect the waiting time of class-r jobs).

After substitution of equations (4.8), (4.9d) and (4.9e) in equation (4.9f), equation (4.9) follows.

Q.E.D.

We point out that an analysis, based on the supplementary variable [COX,55] of a $M^B/G/1$ queue with HOL discipline and 2 classes of jobs has been developed in the literature (c.f. [CHAU,83,pp.140-150]). The L.S.T's of the waiting time distibutions of both classes derived by this technique reduces to equation (4.9) when the bulk sizes are geometrically distributed. Furthermore, the M/G/1 waiting time results [JAIS,68] are obtained from (4.9) after appropriate substitutions ($C_{ar}^2=1$, r=1,2,...R).

4.4.3 Partial equilibrium

An important observation should be made here regarding the statistical equilibrium conditions of a GE/G/1 queue with R (≥2) priority classes under either PR or HOL. The steady state waiting time density for the class-r jobs does not impose any restriction on the r+1,...,R classes. Hence, even if the r+1,...,R classes

experience infinite delays, partial equilibrium up to r classes exists provided ($<G_{r-1}> < +\infty$) or ($\lambda_r<C_r> < 1$). The partial equilibrum is inherent in all exogeneous systems because of the state-independent nature of decision-making with regard to the selection of jobs for service.

4.5 Response time distribution

The response time distribution of class-r jobs, $\tilde{T}_r$, depends on the type of the scheduling discipline, but requires only the knowledge of the marginal waiting time and the completion time or service time distributions in the priority situations. The M/G/1 results [JAIS,68] are then generalised via the following corollary to GE interarrival-time distributions.

Corollary 4.6

For a stable GE/G/1 queue with R ($\geq$2) priority classes under either PR or HOL scheduling discipline, the L.S.T, $T_r^*(.)$, of the response time of class-r jobs is given by:

$$T_r^*(\theta) = W_r^*(\theta)C_r^*(\theta) \quad \text{, for PR} \tag{4.10}$$

$$T_r^*(\theta) = W_r^*(\theta)S_r^*(\theta) \quad \text{, for HOL} \tag{4.11}$$

where $S_r^*(.)$ is L.S.T of the service time of class-r jobs, $W_r^*(.)$ is given by (4.8) and (4.9) for PR and HOL, respectively and $C_r^*(.)$ is given by (4.1).

#

Proof

The Proof follows directly from the definitions of the response time in G/G/1 queue under both disciplines. In particular, under PR

discipline, class-r job stays in the service facility during a full waiting time, $\tilde{W}_r$, and the completion time, $\tilde{C}_r$. However, under HOL rule, class-r job is not interrupted when it undergoes service. Therefore, the class-r response time consists of a waiting time, $\tilde{W}_r$ and a complete service time, $\tilde{S}_r$.

Q.E.D.

4.6 Marginal queue length distribution

The marginal queue length distribution of each class, which from the practical point of view are of obvious importance, are determined via the following corollary:

Corollary 4.7

For a stable GE/G/1 queue with R ($\geq$2) priority classes under either PR or HOL discipline, the generating function, $Q_r(.)$, of the marginal class-r queue length distribution, r=1,2...,R is given by:

$$Q_r(z) = T_r^*\left[\frac{2\lambda_r(1-z)}{C\tilde{a}_r^2+1-(C\tilde{a}_r^2-1)z}\right], \quad |z|\leq 1 \qquad (4.12)$$

where $T_r^*(.)$ is the L.S.T of the response time given by (4.10) and (4.11) for PR and HOL, respectively.

#

Proof

Firstly and because a class-r, r=1,2,...,R, arrival or departure in any non-bulk G/G/1 queue with multiple classes will change the state of the system by one position, Cooper results [COOP,81,pp.185-188] hold within each class. Consequently, a class-r arriver to the queue and departer from the same queue will 'see' the same marginal queue length distribution.

$$P_r^{(a)}(n_r) = P_r^{(d)}(n_r) \qquad (4.12a)$$

Systems with batch arrival or/and service processes are generally excluded from the class of those ones that satisfy Cooper's property. However, the above results can be extended to include GE distributions by considering, first a $H_2/G/1$ queue, for which Cooper's results hold and tend the mean service-time of one of the two exponential servers of the H_2 distributional model to zero (c.f. section 3.4.2.1). Cooper's results have also been extended to more complicated systems such as GE/GE/1 or GE/G/c (c.f. [TOMA,89]).

Secondly, using the bulk interpretation of GE, class-r arrivals are occuring in a compound Poisson fashion. Therefore, each member of the bulk (which shares the same view as the rest of the group), will see the same thing as an 'outside' observer. Thus we have

$$P_r(n_r) = P_r^{(a)}(n_r) \qquad (4.12b)$$

using equation (4.12a) together with (4.12b), we obtain the following relation:

$$P_r(n_r) = P_r^{(d)}(n_r) \qquad (4.12c)$$

Given that jobs are served in FCFS fashion within each class, the class-r jobs left behind by a departer from the same class are the jobs that arrived during its staying in the system (response time).

Using the same arguments as in M/G/1 case (c.f section 2.2.3 or appendix A) with $\lambda_r^{(b)}=\lambda_r\sigma_r$ as the mean bulk arrival rate and $q_r(.)$ as the generating function of the bulk size distribution, we obtain

$$Q_r(z)= T_r^*(\lambda_r^{(b)}-\lambda_r^{(b)}q_r(z)) \qquad (4.12d)$$

Using the generating function of the geometric probability distribution function, $q_r(.)$ [TRIV,82,pp.579], equation (4.12) follows.

Q.E.D.

4.6.1 Marginal mean queue lengths

The marginal mean queue lengths, $\langle n_r \rangle$, r=1,2,...,R of PR or HOL GE/G/1 queue which as mean value constraints are of considerable necessity in the application of the PME, are determined via the following theorem.

Theorem 4.1

For a stable GE/G/1 queue with R (≥2) priority classes under either PR or HOL scheduling discipline, the exact marginal mean queue lengths, $\{\langle n_r \rangle\}$, are given by the closed-form expressions:

$$\langle n_r \rangle = \frac{\rho_r}{1-\gamma_{r-1}} + \frac{\rho_r^2(Cs_r^2+1)+\rho_r(1-\gamma_{r-1})(Ca_r^2-1)+\sum_{\ell=1}^{r-1}\frac{\lambda_r}{\lambda_\ell}\rho_\ell^2(Cs_\ell^2+Ca_\ell^2)}{2(1-\gamma_{r-1})(1-\gamma_r)} \quad \text{for PR} \quad (4.13)$$

$$\langle n_r \rangle = \rho_r + \frac{\rho_r(Ca_r^2-1)}{2(1-\gamma_r)} + \frac{\sum_{\ell=1}^{r-1}\frac{\lambda_r}{\lambda_\ell}\rho_\ell^2(Cs_\ell^2+Ca_\ell^2) + \sum_{\ell=r}^{R}\frac{\lambda_r}{\lambda_\ell}\rho_\ell^2(Cs_\ell^2+1)}{2(1-\gamma_{r-1})(1-\gamma_r)} \quad \text{for HOL} \quad (4.14)$$

#

Proof

Using the fact that the mean queue length is the value of the derivative of the generating function, $Q_r(z)$, at the position z=1 [KLEI,75,pp.385], equations (4.13) and (4.14) follow by straight differentiation of the equation (4.12) where the Hopital's rule has

to be used twice to ensure the existence of the derivative at that point.

Q.E.D.

It is very important to point out that, from the two expressions above, we notice that class-r jobs are completely affected by high-priority classes in PR GE/G/1 queue, whereas in HOL case, although completely influenced by high-priority jobs, they are only affected by the service-time of low-priority jobs and not by their arrival process. This is true, since under HOL, class-r job may be delayed only by the remaining service-time of low-priority job (if any in service).

Note also, that the expressions of the mean queue lengths given by Chaudhry and Templeton [CHAU,83] for the two-class $M^B/G/1$ HOL queue reduces to equation (4.14) when the bulk sizes are geometrically distributed. The M/G/1 results [JAIS,68] for both disciplines are obtained by appropriate substitutions.

4.6.2 Marginal idle state probabilities

The marginal idle state probabilities, $P_r(0)$, r=1,2,..., R of a class-r in PR or HOL queue which are of an extreme importance in the maximum entropy analysis of the queueing system under investigation, are established via the next theorem:

Theorem 4.2

For a stable GE/G/1 queue with R (≥2) priority classes under either PR or HOL scheduling disciplines, the exact marginal idle state probabilities, $\{P_r(0)\}$, are given by

$$P_r(0) = (1-\gamma_r)\left[1 + \frac{[\Lambda_{r-1}^{(b)} - \Lambda_{r-1}^{(b)} G_{r-1}^*(\lambda_r\sigma_r))]}{\lambda_r\sigma_r}\right] \quad \text{, for PR} \tag{4.15}$$

$$P_r(0) = \frac{(1-\rho)\left[\Lambda_r^{(b)}\Lambda_{r-1}^{(b)} G_{r-1}^*(\lambda_r\sigma_r))\right] + \sum_{\ell=r+1}^{R} \lambda_\ell\left[1 - S_\ell^*(\Lambda_r^{(b)}\Lambda_{r-1}^{(b)} G_{r-1}^*(\lambda_r\sigma_r))\right]}{\lambda_r\sigma_r G_r^*(\lambda_r\sigma_r)} S_r^*(\lambda_r\sigma_r) \quad \text{for HOL} \tag{4.16}$$

#

Proof

Given that by definition $P_r(0) = Q_r(0)$, the equation above is obtained by evaluating the generating function, $Q_r(z)$, given by (4.12) at the position z=0 with $\sigma_r = 2/(C_{ar}^2+1)$.

Note that in principle, $P_r(0)$ can be obtained for any G-type distribution, however, the analytic expression of $G_{r-1}^*(\lambda_r\sigma_r)$ can be obtained only in some special cases (e.g., M, GE).

Q.E.D.

It is essential to point out that for GE service-time, the various values of $G_{r-1}^*(.)$, r=2,...,R, are calculated recursively by using initially equation (4.6) and (4.7), respectively. However, the time requirements of this procedure grow in a non-linear fashion as the number of classes increases. A Pascal recursive procedure type can be used to compute the marginal idle state probabilities. A stepwise description of the computation of these statistics is presented in algorithm 4.1:

The step 1.2.1 of the algorthm involves the fixed point iteration method, it is therefore essential to have a good 'guess' for the initial value that we are seeking ($G_{s1}^*(\theta)$). One reasonable

initialisation is $G^*_{s,1}(\theta) = 0.5$, since $|G^*_{s,1}(\theta)| \leqslant 1$, for $|\theta| \leqslant 1$ [KLEI,75, PP.383].

Several experiments carried out, have shown that the average number of iterations required for the convergence of step 1.2.1, is about 4.5 (usually between 4 or 5 iterations).

STEP 1:{* Evaluation of $G^*_{r-1}(\lambda_r\sigma_r)$ *}
- If r=1 set $G^*_0(\lambda_1\sigma_1) \longleftarrow 0$;
- If r=2 $G^*_1(\lambda_2\sigma_2) \longleftarrow$ (eq. 4.6); {* Use (eq. 4.6) *}
- If r>2

STEP 1.1:{* Initialisation phase *}
s $\longleftarrow$ r-1; $\theta \longleftarrow \lambda_r\sigma_r$;

STEP 1.2:{* evaluation of $G^*_s(\theta)$ *}
BUSY(s,θ,$G^*_s(\theta)$);

STEP 1.2.1:{* Evaluation of $G^*_{s,1}(\theta)$ *}
{* Solve with respect $G^*_{s,1}(\theta)$ by fixed point iteration *}

STEP 1.2.1.1:{* Initialisation of $G^*_{s,1}(\theta)$ *}
- $G^*_{s,1}(\theta) \longleftarrow 0.5$;

STEP 1.2.1.2:{* Repeat until convergence of $G^*_{s,1}(\theta)$ *}
i/ If s>2 then begin
- $\theta \longleftarrow \lambda_s\sigma_s(1-G^*_{s,1}(\theta)) + \theta$;
- s $\longleftarrow$ s-1;
- BUSY(s,θ,$G^*_s(\theta)$);
end;
else $G^*_{s-1}(\theta) \longleftarrow$ (eq. 4.6);
ii/{*Obtain new value of $G^*_{s,1}(\theta)$ *}
- $G^*_{s,1}(\theta) \longleftarrow$ (eq. 4.7a);

STEP 1.2.2:{* obtain $G^*_{s_2}(\theta)$ from step 1.2.1.2.ii *}

STEP 1.2.3:{* Obtain $G^*_s(\theta)$ *}
$G^*_s(\theta) \longleftarrow$ (eq. 4.7);

STEP 2:{* Evaluation of $P_r(0)$ *}
$P_r(0) \longleftarrow$ (eq. 4.15 if PR) or (eq. 4.16 if HOL);

Algorithm 4.1: Evaluation of $P_r(0)$, r=1,2,...,R in GE/GE/1 queue under PR or HOL

4.7 Conservation law

In priority systems, preferential treatment given to one class of jobs is at the expense of others. This invariance (or conservation according to Kleinrock [KLEI,76]) within priority queueing systems were first studied by Kleinrock [KLEI,65] in the analysis of a M/G/1 non-preemptive queues. The general idea of conservation is that the expected change of a state function is zero over any finite (including infinitisimal) duration of time picked at random in the steady state. For instance, given a stable and specific work-conserving G/G/1 queue (no work or service is created or destroyed within the system) under a non-preemptive discipline, then the following conservation law holds regardless the type of the service discipline [KLEI,76,pp.117].

$$\sum_{r=1}^{R} \rho_r \langle W_r \rangle + \frac{1}{2} \sum_{r=1}^{R} \frac{\rho_r^2}{\lambda_r} (C_{sr}^2+1) = \langle U \rangle \ , \quad (4.17)$$

where $\langle U \rangle$ is the mean overall waiting time or the mean unfinished work [KLEI,75,pp.11].

Given that at equilibrium the mean unfinished work is fixed for any work-conserving discipline, the sum $\sum_{r=1}^{R} \rho_r \langle W_r \rangle$ remains constant for any non-preemptive discipline such as HOL or FCFS.

Moreover, if the interarrival-times conform to GE distribution the above conservation law equation is given by the following corollary:

Corollary 4.8

For a stable GE/G/1 queue with R classes of jobs, under any work-conserving and non-preemptive scheduling discipline (such as HOL), the following relation must be satisfied regardless of that queueing discipline:

$$\sum_{r=1}^{R} \rho_r \langle W_r \rangle + \frac{1}{2} \sum_{r=1}^{R} \frac{\rho_r^2}{\lambda_r} (C_{sr}^2+1) = \frac{1}{2(1-\rho)} \sum_{r=1}^{R} \frac{\rho_r^2}{\lambda_r}(C_{sr}^2+C_{ar}^2) \qquad (4.18)$$

#

Proof

The unfinished work at time t, U(t), of a queueing system is the same under any work-conserving discipline [KLEI,76,pp.113]. Therefore, and for exposition purposes, we will determine the mean unfinished work ,<U>, of a GE/G/1 queue with R classes under the FCFS discipline.

Since class-r, r=1,2,...R, jobs are assumed to arrive according to $GE(\lambda_r, C_{ar}^2)$, thus using the bulk interpretation of GE (c.f. theorem 3.1), the unfinished work, U(t), is identical to the one of single-class M/G/1 with $\Lambda_R^{(b)} = \sum_{r=1}^{R} \lambda_r \sigma_r$ as the mean arrival rate and

$$S_r^{*(m)}(\theta) = \sum_{r=1}^{R} \frac{\lambda_r \sigma_r}{\Lambda_R^{(b)}} S_r^{*(b)}(\theta) \qquad (4.18a)$$

as the L.S.T of the modified (weighted) service-time distribution, where $S_r^{*(b)}(.)$ is given by (4.3) with index '1' replaced by 'r' and $\sigma_r = 2/(C_{ar}^2+1)$.

Note that the unfinished work in an ordinary M/G/1 queue is given by [KLEI,76,pp115]

$$\langle U \rangle = \frac{\Lambda_R^{(b)} \langle S_r^{(m)\,2} \rangle}{2(1-\rho)} \qquad (4.18b)$$

where $\langle S_r^{(m)\,2} \rangle$ is the second moment of the weighted service time $\tilde{S}_r^{(m)}$.

After differentiating twice (4.18a) and evaluating the mean unfinished work, <U> given by (4.18b), equation 4.18 follows after simple calculations.

4.8 Approximations and performance bounds

(See Appendix C, section C1)

The analysis of a G/G/1 priority queue depends on the choice of the general (G-type) priority interarrival-time and service-time distributions and is characterised by measures such as queue length and response time. In this section, we investigate the effect of the distributional form of the interarrival-time and service-time on the mean queue length per class and idle state probability given by ((4.13 or (4.14)) and ((4.15) or (4.16)), respectively. In addition, relative comparisons against simulations on mean queue length and idle state probability approximations with different G-type distributions is carried out. The system under study is a single-server queue under either PR or HOL discipline (fig. 2.1) with various interarrival-time and service-time distributions given known their first two moments.

Tables 4.1–4.4 display the marginal mean queue lengths ((4.13) or (4.14)) and idle state probability ((4.15) and (4.16)), for PR and HOL, respectively, denoted by (GE) and by simulations (SIM) when the interarrival-time (service-time) conforms to a balanced hyperexponential (H_2), (κ=2) for C_{ar}^2, (C_{sr}^2) > 1 or Erlang-2 (E_2)

distribution functions for C_{ar}^2, (C_{sr}^2)=0.5. In particular, for high coefficients of variation, tables (4.1 and 4.2) shows that the GE-type formulae consistently approximate the simulated mean queue length and idle state probabilities with error tolerances [BRYA,84;NEUS,82] generally less than 0.1. On the other hand and for low coefficients of variation, tables (4.3) and (4.4) present the same statistics with simulated interrrival-time (service-time) conforming to E_2 distribution. Note that, although the GE distribution is improper for coefficient of variation less than one, the GE-type formulae are robust enough to approximate the statistics above of priority centre when hypoexponential distributional models are involved [EL-AF,83].

Furthermore, to assess the credibility of the GE/G/1 priority queue analytic approximations, it is practically useful to enhance the above validation by considering more complex queues involving heterogeneous types of interarrival-time and service-time distributions per class including, in addition, deterministic (D) and uniform (U) distributional models. Tables (4.5) and (4.6) support this validation for PR and HOL, respectively, where we notice that the GE-type formulae favourably approximate the simulated quantities.

In practical term, it is simple and much more useful to establish some kinds of bounds for systems involving queues rather than to use sophisticated analysis leading to more computationally expensive techniques. Bounds define the upper (pessimistic for performance measures such as the mean queue lengths) and lower (optimistic) limiting values of the statistics required. They are generally suitable as a first cut modelling technique that can be used to eliminate inadequate alternatives at an early stage of a study. To this end, the marginal mean queue lengths; $\{<n_r>\}$, of a stable G/G/1 queue with 3 classes under either PR or HOL service discipline are depicted in figures (4.1) and (4.2), respectively. Curves are grouped

according to the squared coefficients of variation C^2_{ar} and C^2_{sr} ranging from 0.5 to 18 with fixed λ_r and μ_r, r=1,2,3 (table 4.7). These figures are drawn by applying analytic (ANAL) GE-type formulae and carrying out simulation (SIM) involving E_2, M, and $H_2(\kappa \epsilon [2,50])$ distributions. It can be observed that the GE/G/1 mean queue length (4.13) or (4.14) provides for $<n_r>$, r=1,2,3, pessimistic bounds when C^2_{ar}, $C^2_{sr} > 1$ and optimistic bounds when e.g., C^2_{ar}, $C^2_{sr} = 0.5$. This is typical situation in many experiments, leading to the following conjecture:

Conjecture 4.1:

Consider a stable G/G/1 queue with R (≥2) priority classes under either PR or HOL scheduling disciplines. the analytic measures of the mean queue length, $<n_r>$, r=1,2,...,R and the mean response time, $<T_r>$, of the robust GE/G/1 priority queue, given the first two moments of the interarrival-time and service-time per priority class-r, define pessimistic (optimistic) bounds on the corresponding quantities obtained from simulation models when representing interarrival-time and service-time by hyperexponential (Erlang or hypoexponential) distributions with the same first two moments.

#

The above conjecture is attributed to the fact that the GE distributional model is the limiting case of $H_2(h_2)$ when the tuning parameter κ goes to $+\infty$ ($-\infty$). Note that similar conjecture have been established for a single-class G/G/1 queue by Kouvatsos [KOUV,88a].

4.9 Conclusion

Based on the bulk interpretation of GE, new exact analytic formulae of a stable GE/G/1 queue under PR or HOL discipline are

derived. These expressions constitute a generalisation of present and known results obtained for a pure Poisson arrival process (M/G/1).

We propose, in particular, exact and closed-form expressions for the L.S.T of the busy period of a single-class GE/GE/1 queue which is currently available only for M/M/1 queues. The analysis is extended to more than one class, where we present a recursive formula for the L.S.T of the busy period of GE/G/1 queue with r, (r>1) priority classes. The expression derived is similar to the one proposed by Jaiswal [JAIS,68] in the analysis of a M/G/1 multiple classes queue.

Analogous results are generalised for the completion time, waiting time, response time and queue length distributions. As a consequence, closed-form expressions for the marginal mean queue lengths and idle state probabilities are derived. These results are used as mean value constraints in order to expedite the utility of the ME solution of the G/G/1 priority queue in the next chapter.

Moreover, it is conjectured that the GE-type formulae define some useful performance bounds for the mean queue length, $\langle n_r \rangle$, r=1,2,...,R of a stable G/G/1 priority queue.

CHAPTER 5

ME Analysis of a G/G/1 Priority Queue

5.1 Introduction

A stable G/G/1 priority queue is an important building block in the performance analysis of computer systems and communication networks. In principle, the stochastic analysis of this queue in isolation depends upon the choice of the general (G-type) priority interarrival-time and service-time distributions and is characterised by measures such as queue length and response time.

As mentioned earlier, the analysis of such queue is very difficult to tackle using the classical queueing theory. The corresponding analytic results that are available in the literature are generally restricted to mean values or transforms (L.S.T. and generating functions) which are difficult to invert in order to obtain probability distribution functions, despite the wide use of the Poisson distribution in modelling the arrival process [JAIS,68]. Recursive formulae for the exact queue length distributions of both PR and HOL M/M/1 queues with two classes of jobs have been established by Marks [MARK,73] and Miller [MILL,81]. In particular, Marks solved by induction the priority M/M/1 global balance equations, while Miller applied matrix invariant probability vectors. However, extensions of these formulae to the case of more than two priority classes require more complex algebraic manipulations which are not as yet available.

In this chapter the PME is used to provide a new analytic framework for the approximate analysis of a stable G/G/1 queue with R priority classes, under PR or HOL scheduling disciplines.

The principle is used in section two under two different sets of mean value constraints, {normalisation (norm), marginal utilisations (ρ_r), marginal mean queue lengths ($<n_r>$)} and {norm, ρ_r, $<n_r>$, and marginal idle state probabilities ($P_r(0)$)}, to establish closed-form approximate expressions for the joint queue length distribution of a stable PR or HOL G/G/1 queue with R classes of jobs. New one-step recursions for the ME state probabilities are derived and closed-form approximations for the marginal queue length distributions are established. These results are used in turn as a basis in section three to determine the distribution type of the effective priority service time (a quantity commonly used in the shadow-CPU methods) that corresponds to the ME solutions. As a consequence, new approximate formulae for the mean and coefficient of variation of the effective priority service time are proposed. Moreover, universal flow expressions for the parameters of the departure process from a priority centre are derived at the end of this section. These results are indispensible for the ME analysis of general QNM's containing priority centres.

Numerical validation examples are presented in section four to illustrate the accuracy of the proposed ME solutions in relation to the simulation involving different underlying probability distributions.

Concluding remarks follow in the last section.

5.2 ME solution of a G/G/1 priority queue

Consider a stable G/G/1 queue with R ($R \geq 2$) priority classes of jobs under either PR or HOL service discipline. It is assumed that class-r, $r = 1,2,\ldots,R$, jobs arrive to the system according to an arbitrary distribution with mean arrival rate, λ_r, and squared

coefficient of variation, C_{ar}^2 and they are served by a single general server with mean service rate, μ_r, and squared coefficient of variation, C_{sr}^2 (see figure 2.1).

Let us define the vector state of the system $\underline{S} = (n_1, n_2, \ldots, n_R, u)$, where n_r designates the number of jobs of class-r in the system and u is the variable index indicating the class of the current job in the service (N.B., for an empty system $u = 0$). Note that, in contrast with non-priority based service disciplines, the multidimensional vector, $\underline{S}$, defines without ambiguity the state of the system, since the ordering of jobs in a priority queue is unique for a specific population vector, $\underline{n}$, (i.e., higher priority jobs occupy always the front of the queue, followed by lower priority jobs, whereas the lowest jobs are at the tail of the queue).

Let Q denotes the set of countably infinite states, $\underline{S}$, of the system, and $P(\underline{S})$ be the equilibrium probability that the G/G/1 priority queue is in state $\underline{S}$.

The following analysis assumes that the parameters λ_r, μ_r, C_{ar}^2 and C_{sr}^2, r=1,2,...,R, form a basic set of a prior knowledge and present queue length probability assignment subject to additional prior information.

5.2.1 Case 1 (ME1): Prior information {norm, ρ_r, $\langle n_r \rangle$}

Suppose all that is known about the state probabilities $\{P(\underline{S})\}$ is that the following mean value constraints exist:

i/ The Normalisation (norm),

$$\sum_{\underline{S} \epsilon Q} P(\underline{S}) = 1 \qquad (5.1)$$

ii/ The server utilisation due to class-r, $\rho_r = \lambda_r/\mu_r$, written as:

$$\sum_{\underline{S}\epsilon Q} h_r(\underline{S})\ P(\underline{S}) = \rho_r \qquad (5.2)$$

where $$h_r(\underline{S}) = \begin{cases} 1 \text{ if } u = r \\ 0 \text{ otherwise} \end{cases}$$

iii/ The mean queue length per class r, $\langle n_r \rangle$, $\langle n_r \rangle \geq \rho_r$

$$\sum_{\underline{S}\epsilon Q} n_r P(\underline{S}) = \langle n_r \rangle,\ r=1,2,\ldots,R \qquad (5.3)$$

It is further assumed that the statistics above can be determined via analytic formulae based on stochastic assumptions, although they may also be known numerically via system measurements during finite operational periods [DENN,78].

Since generally, the number of constraints is less than the number of feasible states, there is an infinite number of distributions $\{P(\underline{S})\}$ satisfying the constraints. The problem is which one to choose?

The PME [JAYN,68] states that of all distributions satisfying the constraints supplied by the given information, the form of the minimally biased distribution which should be chosen is the one that maximizes the system's entropy function, H(P), given by

$$H(P) = -\sum_{\underline{S}\epsilon Q} P(\underline{S}) \log\{P(\underline{S})\} \qquad (5.4)$$

The maximization of (5.4) subject to constraints (5.1)-(5.3), can be carried out analytically by using the Lagrange's method of undetermined multipliers leading to the following solutions:

$$P(\underline{S}) = \frac{1}{Z} \prod_{r=1}^{R} g_r^{h_r(\underline{S})} x_r^{n_r} \qquad (5.5)$$

where Z is the normalising constant given by

$$Z = \sum_{\underline{S} \epsilon Q} \prod_{r=1}^{R} g_r^{h_r(\underline{S})} x_r^{n_r} \qquad (5.6)$$

where $g_r = e^{-\beta_1 r}$; $x_r = e^{-\beta_2 r}$, r=1,2,...,R, and $\beta_1 r$, $\beta_2 r$ are the Lagrangian multipliers corresponding to constraints (5.2) and (5.3), respectively.

Note that the joint queue length distribution, $P(\underline{n})$, is obtained by aggregating all the steady state probabilities and is given by

$$P(\underline{n}) = \sum_{u=1}^{R} P(n_1, n_2, \ldots, N_R, u),$$

5.2.1.1 PR discipline

In G/G/1 priority queues with PR scheduling discipline, there is only one ordering of job classes with a job in service , if any, always belonging to the highest priority class present. In this case the index variable, u, of the state of the system is clearly redundant, and the vector state $\underline{S}$ is symbolised directly by the vector $\underline{n} = (n_1, n_2, \ldots, n_R)$.

Using equations (5.5) and (5.6), the ME joint queue length distribution and the Lagrangian coefficients g_r, x_r, r=1,2,...,R, can be obtained analytically via the following theorem:

Theorem 5.1

The joint ME queue length distribution of a stable G/G/1 queue with R (R≥2) priority classes under PR service discipline, subject to constraints (5.1), (5.2) and (5.3) is given by

$$P(\underline{n}) = \begin{cases} \dfrac{1}{Z} = 1-\rho & \text{for } \underline{n} = \underline{0} \\ (1-\rho) g_r \prod_{\ell=r}^{R} x_\ell^{n_\ell} & \text{for } n_1=n_2=\ldots=n_{r-1}=0 \wedge n_r>0 \end{cases} \tag{5.7}$$

$$\text{where } x_r = \frac{\langle n_r \rangle - \rho_r}{\langle n_r \rangle + \gamma_{r-1}} \qquad \text{for } r=1,2,\ldots,R \tag{5.8}$$

$$g_r = \begin{cases} \dfrac{\rho_r}{(1-\rho)} \dfrac{\gamma_r}{(\langle n_r \rangle - \rho_r)} \prod_{\ell=r+1}^{R} \dfrac{\gamma_\ell}{\langle n_\ell \rangle + \gamma_{\ell-1}} & , \ r=1,\ldots,R-1 \\ \dfrac{\rho_R}{(1-\rho)} \dfrac{\rho}{(\langle n_R \rangle - \rho_R)} & , \ r=R \end{cases} \tag{5.9}$$

where ρ_r, ρ, γ_r are given by equations (2.2), (2.3) and (2.4), respectively.

#

The ME solution (5.7) is established directly by using expressions (5.5)-(5.6) for PR discpline. By using constraints (5.1)-(5.3) and queue length distribution (5.7), after some manipulations, expressions (5.8) and (5.9) follows. The full details of the proof can be found in appendix D (section D1).

The marginal queue length distribution of class-r, r=1,2,...,R, jobs are determined via the following corollary:

Corollary 5.1

The marginal ME queue length distribution $P_r(n_r)$ of class-r jobs, r=1,2,...,R, of a stable G/G/1 priority queue under PR service discipline, subject to constraints (5.1)-(5.3) is given by

$$P_r(n_r) = \begin{cases} 1-\rho_r-\gamma_{r-1}x_r \ , \ n_r=0 \\ \\ (\rho_r+\gamma_{r-1}x_r)(1-x_r)x_r^{n_r-1} \ , \ n_r>0, \end{cases} \qquad (5.10)$$

#

Equation (5.10) is obtained by applying the law of total probability to the ME solution (5.7). The detailed proof can be seen in appendix D (section D2).

Note that the marginal queue length distribution of class-1 jobs are identical to the ME solution of a single class G/G/1 queue [KOUV,88a]. This is expected, since under PR discipline, highest priority jobs are not affected by jobs belonging to other classes.

The robust one-step recursions which permit an efficient computation of the ME solution (5.7), are defined via the following corollary.

Corollary 5.2

The joint ME queue length distribution, $P(\underline{n})$, of a stable G/G/1 queue with R (R>2) priority classes under PR sevice discipline, subject to constraints (5.1)-(5.3), satisfies the following one-step recursions.

$$P(\underline{n}) = \begin{cases} g_r x_r P(\underline{0}), \text{ for } \underline{n}=\underline{1}_r \\ \\ \dfrac{g_r x_r}{g_s} P(\underline{n}-\underline{1}_r), \text{ for } \{n_1=\ldots=n_{r-1}=n_{r+1}=\ldots=n_{s-1}=0, n_r=1 \\ \qquad\qquad n_s>0, \ r<s \ \} \qquad (5.11) \\ \\ x_r P(\underline{n}-\underline{1}_r), \text{ for } \{n_1=n_2=\ldots=n_{s-1}=0, \ n_s>0, \ n_r>1, \ r>s \ \} \end{cases}$$

#

The proof is based on the product-form property of the ME solution (5.7) and can be seen in appendix D (section D3).

5.2.1.2 HOL discipline

In the ME analysis of a G/G/1 priority queue with HOL service discipline, the vector state, $\underline{S}$, is identified as $\underline{S}= (n_1,n_2,\ldots,n_R,u)$ with class index 'u' being any element belonging to the integer sub-set $\{1,2,\ldots,R\}$, given that n_u is strictly positive. The vector population $\underline{n}$ is then an aggregate state given by the union of states $\underline{S}$ belonging to the set $Q\underline{n}=\{\underline{S}/\ \underline{S}\epsilon Q \wedge u=1,2,\ldots,R\}$.

The joint ME queue length distribution and the Lagrangian coefficients $\{g_r\}$ and $\{x_r\}$ for a G/G/1 non-preemptive priority (HOL) queue can be obtained analytically via the following theorem:

Theorem 5.2

The joint ME queue length distribution, of a stable G/G/1 queue with R (R≥2) priority classes under HOL service discipline, subject to constraints (5.1), (5.2) and (5.3) is given by

$$P(\underline{n}) = \begin{cases} \dfrac{1}{Z} = 1-\rho & \text{for } \underline{n} = \underline{0} \\ (1-\rho)\prod_{r=1}^{R} x_r^{n_r}\left[\sum_{s=1\wedge n_s>0}^{R} g_s\right], & \text{for } \underline{n}\neq 0 \end{cases} \qquad (5.12)$$

$$\text{where} \quad x_r = \frac{\langle n_r\rangle - \rho_r}{\langle n_r\rangle - \rho_r + \rho} \qquad \text{for } r=1,2,\ldots,R \qquad (5.13)$$

$$g_r = \frac{\rho_r}{(1-\rho)} \frac{\rho}{(<n_r>-\rho_r)} \prod_{\ell=1 \wedge \ell \neq r}^{R} \frac{1}{<n_\ell>+\rho-\rho_r}, \quad r=1,\ldots,R \tag{5.14}$$

#

As in the proof of theorem 5.1, the ME joint queue length distribution given by equation (5.12) is obtained directly by using (5.5) and (5.6) for HOL discipline. The Lagrangian coefficients, $\{g_r\}$ and $\{x_r\}$ are derived by making use of constraints (5.1)-(5.3) and the ME queue length distribution (5.12). More details of the proof are given in appendix D (section D4).

The marginal queue length distribution of class-r jobs is determined via the following corollary:

Corollary 5.3

The marginal ME queue length distribution $P_r(n_r)$ of class r jobs, r=1,2,...,R, of a stable G/G/1 priority queue under HOL service discipline, subject to constraints (5.1)-(5.3) is given by

$$P_r(n_r) = \begin{cases} 1-\rho_r-(\rho-\rho_r)x_r \ , & n_r=0 \\ \\ (\rho_r+(\rho-\rho_r)x_r)(1-x_r)x_r^{n_r-1} \ , & n_r>0, \end{cases} \tag{5.15}$$

#

The proof of corollary 5.3 is based on the law of total probability and can be found in appendix D (section D5).

Note that, as expected, the ME expression of the HOL marginal queue lentgh distribution of class-r, r= 1,2,...,R, depends upon the parameters of higher and lower priority classes.

The one-step recursions of the ME solution (5.12) are given by the following corollary:

Corollary 5.4

The joint ME queue length distribution, $P(\underline{n})$, of a stable G/G/1 queue with R ($R \geq 2$) priority classes under HOL sevice discipline, subject to constraints (5.1)-(5.3), satisfies the following one-step recursions.

$$P(\underline{n}) = \begin{cases} g_r x_r P(\underline{0}), \text{ for } \underline{n}=\underline{1}_r \\ \dfrac{\sum_{\substack{s=1 \wedge n_s \neq 0}}^{R} g_s}{\sum_{\substack{s=1 \wedge n_s \neq 0 \wedge s \neq r}}^{R} g_s} x_r P(\underline{n}-\underline{1}_r), \text{ for } n_r = 1 \\ x_r P(\underline{n}-\underline{1}_r), \text{ for } n_r > 1 \end{cases} \qquad (5.16)$$

#

The above one-step recursions are defined from the ME solution (5.12) and full proof can be seen in appendix D (section D6).

5.2.2 Case 2 (ME2): Prior information {norm, ρ_r, $\langle n_r \rangle$, $P_r(0)$}.

Suppose, in addition to the constraints (5.1) - (5.3), the following mean value constraints are given or known to exist:

iv/ The marginal idle state probabilities, $P_r(0)$, r=1,2,...,R.

$$\sum_{\underline{S} \epsilon Q} V_r(\underline{S}) P(\underline{S}) = \theta_r = 1 - P_r(0), \; r = 1,2,\ldots,R \qquad (5.17)$$

where $V_r(\underline{S}) = \begin{cases} 1 \text{ if } n_r > 0 \\ 0 \text{ otherwise} \end{cases}$

and θ_r is the proportion of time during which at least one job of class-r is present in the queueing system.

Maximizing the entropy function (5.4) subject to constraints (5.1)-(5.3) and (5.17) is carried out using the Lagrange's method of undetermined multipliers, leading to the following ME solution:

$$P(\underline{S}) = \frac{1}{Z}\prod_{r=1}^{R} g_r^{h_r(\underline{S})} x_r^{n_r} y_r^{V_r(\underline{S})} \qquad (5.18)$$

where Z is still the normalising constant and $\{g_r\}$, $\{x_r\}$ and $\{y_r\}$ are the Lagrangian coefficients corresponding to the marginal utilisations, mean queue lengths, and idle state probabilities, respectively.

5.2.2.1 PR discipline

Since under PR discipline, class-1 jobs are not affected by the presence of low-priority jobs, we have $\theta_1=\rho_1$. Therefore, the first idle state probability constraint is clearly redundant, and eventually $y_1=1$. The joint ME distribution is subsequently determined via the following theorem.

Theorem 5.3

The joint ME queue length distribution, of a stable G/G/1 queue with R (R≥2) priority classes under PR service discipline, subject to constraints (5.1), (5.2), (5.3) and (5.17) is given by

$$P(\underline{n}) = \begin{cases} \frac{1}{Z} = 1-\rho & \text{for } \underline{n} = \underline{0} \\ (1-\rho) g_r y_r x_r^{n_r} \prod_{\ell=r+1}^{R} x_\ell^{n_\ell} y_\ell^{V_\ell(\underline{S})} & \text{for } n_1=n_2=\ldots=n_{r-1}=0 \wedge n_r>0 \end{cases} \qquad (5.19)$$

where the Lagrangian coefficients $\{g_r\}$, $\{x_r\}$ and $\{y_r\}$ are given by:

$$x_r = \frac{\langle n_r \rangle - \theta_r}{\langle n_r \rangle} \quad , \quad \text{for } r=1,2,\ldots,R \qquad (5.20)$$

$$g_r = \begin{cases} \dfrac{\rho_1}{(1-\rho)} \dfrac{\rho_1}{(\langle n_1 \rangle - \rho_1)} \prod_{\ell=2}^{R} \dfrac{\gamma_\ell - \theta_\ell}{\gamma_{\ell-1}} & , \text{ for } r=1 \\ \dfrac{\rho_r}{(1-\rho)} \dfrac{\gamma_r - \theta_r}{(\theta_r - \rho_r)} \prod_{\ell=r+1}^{R} \dfrac{\gamma_\ell - \theta_\ell}{\gamma_{\ell-1}} & , \text{ for } r>1 \end{cases} \qquad (5.21)$$

and

$$y_r = \begin{cases} 1 \ , \text{ for } r=1 \\ \dfrac{(\theta_r - \rho_r)}{(\gamma_r - \theta_r)} \dfrac{(1 - x_r)}{x_r} & , \text{ for } r=2,\ldots,R \end{cases} \qquad (5.22)$$

#

The proof follows similar lines to the ones used in theorem 5.1. For more details see appendix D (section D7).

The marginal ME queue length distributions are determined via the following corollary:

Corollary 5.5

The marginal ME queue length distribution $P_r(n_r)$ of class-r jobs, r=1,2,...,R, of a stable G/G/1 priority queue under PR service discipline, subject to constraints (5.1)-(5.3) and (5.17) is given by

$$P_r(n_r) = \begin{cases} 1-\theta_r \ , \ n_r=0 \\ \theta_r(1-x_r)x_r^{n_r-1} \ , \ n_r>0, \end{cases} \qquad (5.23)$$

#

Because the marginal idle state probability of class-r jobs, $P_r(n_r)$, is given as prior information, we then have

$$P_r(0) = 1-\theta_r$$

For $n_r>0$, The proof follows similar steps to the one used in corollary 5.1 (see appendix D, section D2).

The one-step recursions (5.11) (section 5.2.1.1) are generalised via the following corollary:

Corollary 5.6

The joint ME queue length distribution, $P(\underline{n})$, of a stable G/G/1 queue with R ($R\geq 2$) priority classes under PR sevice discipline, subject to constraints (5.1)-(5.3) and (5.17), satisfies the following one-step recursions.

$$P(\underline{n}) = \begin{cases} g_r x_r P(\underline{0}), \text{ for } \underline{n}=\underline{1}_r \\ \dfrac{g_r x_r y_r}{g_s} P(\underline{n}-\underline{1}_r), \text{ for } \{n_1=\ldots=n_{r-1}=n_{r+1}=\ldots=n_{s-1}=0, n_r=1, n_s>0,\ r<s\} \\ x_r P(\underline{n}-\underline{1}_r), \text{ for } \{n_1=n_2=\ldots=n_{s-1}=0,\ n_s>0,\ n_r>1,\ r\geq s\} \\ x_r y_r P(\underline{n}-\underline{1}_r) \text{ for } \{n_1=n_2=\ldots=n_{s-1}=0,\ n_s>0,\ n_r=1,\ r\geq s\} \end{cases} \quad (5.24)$$

#

The one-step recursions are obtained by identifying the ME solution (5.19) as a product-form of factors involving the Lagrangian coefficients $\{x_r\}$, $\{g_r\}$ and $\{y_r\}$. Therefore, following similar lines to the ones used in the proof of corollary 5.2 (see appendix D, section D3), equation (5.24) follows.

5.2.2.2 HOL Discpline

The ME solution for the joint steady state probabilities of a HOL G/G/1 queue, subject to the set of constraints {Norm, ρ_r, $\langle n_r \rangle$, $P_r(0)$} is determined via the following theorem:

Theorem 5.4

The joint ME queue length distribution, of a stable G/G/1 queue with R (R≥2) priority classes under HOL service discipline, subject to constraints (5.1), (5.2), (5.3) and (5.17) is given by

$$P(\underline{n}) = \begin{cases} \dfrac{1}{Z} = 1-\rho & \text{for } \underline{n} = \underline{0} \\ (1-\rho) \prod_{r=1}^{R} x_r^{n_r} y_r^{V_r(\underline{S})} \left[\sum_{s=1 \wedge n_s>0}^{R} g_s \right] & \text{, for } \underline{n} \neq 0 \end{cases} \tag{5.25}$$

where x_r, r=1,2,...,R are also given by equation (5.20), and

$$g_r = \frac{\rho_r}{(1-\rho)} \frac{\rho - \theta_r}{(\theta_r - \rho_r)} \prod_{\ell=1 \wedge \ell \neq r}^{R} \frac{\rho - \theta_\ell}{\rho - \rho_\ell} \quad , \; r=1,\ldots,R \tag{5.26}$$

$$y_r = \frac{\theta_r - \rho_r}{\rho - \theta_r} \frac{1 - x_r}{x_r} \quad , \text{ for } r = 1,2,\ldots,R \tag{5.27}$$

#

As in case 1, the joint steady state vector of a stable G/G/1 queue under HOL discipline is an aggregate state and therefore the probability to be in that state is just the sum of the corresponding individual state probabilities.

$$P(\underline{n}) = \frac{1}{Z} \sum_{r=1 \wedge n_r>0}^{R} g_r \prod_{s=1}^{R} y_s^{V_s(\underline{S})} x_s^{n_s}$$

Following similar steps to the ones used in the proof of theorem 5.3 (see appendix D, section D7), together with the use of the ME distribution given above, the equations of theorem 5.4 follow.

The ME marginal queue length distributions, $\{P_r(n)\}$, are obtained by applying the law of total probability to the ME solution (5.25), leading to the following corollary:

Corollary 5.7

The marginal ME queue length distribution $P_r(n)$ of class r jobs, r=1,2,...,R, of a stable G/G/1 priority queue under HOL service discipline, subject to constraints (5.1)-(5.3) and (5.17) is given by

$$P_r(n_r) = \begin{cases} 1-\theta_r \ , \ n_r=0 \\ \\ \theta_r(1-x_r)x_r^{n_r-1} \ , \ n_r>0 \ . \end{cases} \qquad (5.28)$$

#

Note that, although the ME marginal probabilities of a G/G/1 queue under HOL discipline are given by the same analytic expressions to the ones obtained under PR rule (5.23), their numerical values are different. This is because different mean value constraints are generated for various service disciplines.

Meanwhile, the one-step recursions given by equation (5.16) are generalised following similar steps of the proof of corollary 5.4 (see appendix D, section D6),via the next corollary.

Corollary 5.8

The joint ME queue length distribution, $P(\underline{n})$, of a stable G/G/1 queue with R ($R \geq 2$) priority classes under HOL sevice discipline, subject to constraints (5.1)-(5.3) and (5.17), satisfies the following one-step recursions.

$$P(\underline{n}) = \begin{cases} g_r x_r y_r P(\underline{0}), \text{ for } \underline{n}=\underline{1}_r \\ \dfrac{\sum_{\substack{s=1 \\ s=1 \wedge n_s \neq 0}}^{R} g_s}{\sum_{\substack{s=1 \\ s=1 \wedge n_s \neq 0 \wedge s \neq r}}^{R} g_s} x_r y_r P(\underline{n}-\underline{1}_r), \text{ for } \underline{n} \neq \underline{1}_r \text{ and } n_r = 1 \\ x_r P(\underline{n}-\underline{1}_r), \text{ for } \underline{n} \neq \underline{1}_r \text{ and } n_r > 1 \end{cases} \quad (5.29)$$

#

5.3 On the approximation of the effective service time distribution.

In the previous section, we have established new approximate formulae for the joint and the marginal queue length distributions of a stable G/G/1 priority queue. The ME solutions derived are consistent with some mean values given as prior information.

To use these results in the context of the shadow-CPU based techniques for the analysis of general QNM's with priorities, it is necessary to estimate the effective service-time distributions, $\hat{S}_r$, r=1,2,...,R. In general, these quantities, which are also the service-time perceived by class-r jobs in isolation, depend on the choice of the general (G-type) priority interarrival and service time distributions.

When considering the effective service-time, $\hat{S}_r$, to be the service-time of the r^{th} virtual server, it is very important to stipulate that both the virtual and the original server generate the same marginal probabilities. Moreover, the random variable, $\hat{S}_r$, is generally highly variable ($C^2_{\hat{S}r}>1$) due to the effect of preemptions of high-priority jobs. Therefore, assuming exponential effective service time [SEVC,77a; KAUF,84; SCHM,83,84], may lead to substantial errors in the prediction of the various statistics. Consequently, it is necessary to determine higher moments of the r.v's. $\{\hat{S}_r\}$. To this end, a ME (ME1 for case 1 and ME2 for case 2) approximations of the effective service time, based on the robust GE/G/1 priority queue, are proposed via the following theorem.

Theorem 5.5

The marginal ME (ME1 or ME2)queue length distribution, $P_r(n_r)$, r=1,2,...,R, of a stable GE/G/1 priority queue under either PR or HOL scheduling discipline, is equivalent, to the equilibrium solution of a virtual stable GE/GE/1 queue, r=1,2,...,R, with service time distribution of the GE form

$$\hat{S}_r(t) = (1-\hat{\tau}_r)u_0(t) + \hat{\tau}_r\hat{\mu}_r e^{-\hat{\tau}_r\hat{\mu}_r t} \quad ,t>0, \qquad (5.30)$$

Where $u_0(.)$ is the unit impulse function [KLEI,75,pp.341],

and
$$\hat{\tau}_r = \frac{\sigma_r\delta_r x_r}{x_r - (1-\sigma_r)(1+\hat{\delta}_r x_r)} \qquad (5.31)$$

$$\hat{\mu}_r = \hat{\tau}_r^{-1}\left\{\frac{\lambda_r\sigma_r(1+\hat{\delta}_r x_r - x_r)}{x_r - (1-\sigma_r)(1+\hat{\delta}_r x_r)}\right\} \qquad (5.32)$$

$$\text{with} \qquad \hat{\delta}_r = \frac{(1-\hat{\rho}_r)x_r}{(1-x_r)\hat{\rho}_r} \tag{5.33}$$

$$\text{and} \qquad \hat{\rho}_r = \begin{cases} \rho_r + \gamma_{r-1}x_r \ , \ r=1,2,\ldots,R \ \text{(PR, ME1)} \\ \rho_r + (\rho-\rho_r)x_r, \ r=1,2,\ldots,R \ \text{(HOL, ME1)} \\ \\ \theta_r \ , \ r=1,2,\ldots,R \ \text{(PR or HOL, ME2)} \end{cases} \tag{5.34}$$

Where $\{x_r\}$ are the Lagrangian coefficients corresponding to the mean queue length constraints and are given in case 1 by equation (5.8) and (5.13) for PR and HOL, respectively, or in case 2 by equation (5.20). Moreover, σ_r is the parameter of the r^{th} GE arrival stream, and is given by $\sigma_r=2/(C_{ar}^2+1)$.

#

The proof is based on the bulk interpretation of GE (c.f. theorem 3.1) and is similar to that presented by Kouvatsos [KOUV,88a] for a non-priority GE/G/1 queue with a single class of customers. Full details of the proof can be found in appendix D, section D8.

Moreover, the mean and the squared coefficient of variation of the effective service time of class-r jobs, $\hat{S}_r$, r=1,2,..R, are determined via the following corollary.

Corollary 5.9

The mean and the squared coefficient of variation of the effective service time of class-r jobs, r=1,2...,R, in a stable GE/GE/1 virtual queue satisfy the following relations:

$$\langle\hat{S}_r\rangle = \frac{\hat{\rho}_r}{\lambda_r} \ , \ r=1,2,\ldots,R \tag{5.35}$$

$$\hat{Cs}_r^2 = \frac{2\langle n_r\rangle(1-\hat{\rho}_r) - \hat{\rho}_r(1-\hat{\rho}_r) - \hat{\rho}_r C\hat{a}_r^2}{\hat{\rho}_r^2} \quad (5.36)$$

#

The proof follows directly from theorem 5.5 as follows:

Since the effective service time is of GE type from equations (5.31) and (5.32), we obtain

$$\langle\hat{S}_r\rangle = \frac{1}{\hat{\mu}_r} - \frac{\hat{\delta}_r x_r}{\lambda_r(1 + \hat{\delta}_r x_r - x_r)}$$

Using equation (5.34), together with the expressions of x_r for case 1 or case 2, as appropriate, the equation (5.35) follows.

Note that, since $\hat{\rho}_r$ can also be interpreted as the perceived utilisation of the r^{th} virtual server, it follows from Little's law [LITT,61], $\hat{\rho}_r = \lambda_r\langle\hat{S}_r\rangle$.

Given that $\hat{\tau}_r$ is the branching probability to the exponential effective server of the GE distributional model (c.f. Fig 3.1), we then have

$$\hat{\tau}_r = \frac{2}{\hat{Cs}_r^2 + 1}$$

Solving the equation above with respect to $\hat{C}_{sr}^2$, and subsequently, using equation (5.31), equation (5.36) follows.

Remarks

It is interesting to point out that, although, the results of theorem 5.5 and corollary 5.9 require GE interarrival and arbitrary service time distributions per class to be derived, they may also be used as an approximation for more general processes, provided that

the first two moments are known before hand. As a special application, consider a pure Markovian system ($C_{ar}^2 = C_{sr}^2 = 1$, r=1,2...,R), equation (5.36) presents an approximation for the squared coefficient of variation of the effective service time which is assumed to be equal to one in the context of current shadow-CPU based methods [SEVC,88a; KAUF,84; SCHM,83,84].

Furthermore, when a priority queue is a part of a queueing network, the marginal departure processes generate arrivals to other centres. Therefore, it is necessary in these circumstances to estimate the first two moments of these processes. To this end, the marginal mean departure rate and squared coefficient of variation of the interdeparture time are determined via the following corollary:

Corollary 5.10

The mean departure rate, λ_{dr}, of class r, r=1,2...,R, and the squared coefficient of variation, C_{dr}^2, of the interdeparture time of class-r jobs of a stable GE/GE/1 virtual queue which correspond to the ME solutions (ME1 and ME2) are given by

$$\lambda_{dr} = \lambda_r \tag{5.37}$$

$$C_{dr}^2 = 2\langle n_r\rangle(1-\hat{\rho}_r) + C\hat{a}_r^2(1-2\hat{\rho}_r) \tag{5.38}$$

where $\hat{\rho}_r$ is given by equation (5.34).

#

Equation (5.37) is a consequence of the steady-state condition that is assumed to hold in a long-run.

Meanwhile, equation (5.38) follows directly from theorem 5.5 and corollary 5.9. In fact, since the marginal queue length distributions of a GE/G/1 priority queue correspond to the ones of a single-class

GE/GE/1 with $\hat{\mu}_r$ and $\hat{Cs}_r$ as parameters of the modified service time, it follows from corollary 3.6 (c.f. [KOUV,88a]), that the squared coefficient of variation of the interdeparture time from the r^{th} virtual server is given by

$$C_{dr}^2 = \hat{\rho}_r(1-\hat{\rho}_r) + (1-\hat{\rho}_r)C\hat{a}_r^2 + \hat{\rho}_r^2\hat{Cs}_r^2$$

Using the expression of $\hat{Cs}_r^2$ given by (5.36), equation (5.38) follows after simple calculations.

Note that formula (5.38) does not depend explicitly on the coefficient of variation of the perceived service-time, $\hat{C}_{sr}^2$. Besides, the effect of each service discipline on the marginal departure processes is reflected via the values of the performance measures $<n_r>$ and $\hat{\rho}_r$.

5.4 Numerical results (See Appendix D, section D9)

In this section numerical examples are presented on stable PR and HOL G/G/1 queue in order to demonstrate the credibility of the marginal ME solutions of sections 5.2.1 and 5.2.2, respectively, and also to illustrate how critically the distributional forms of the interarrival and service times per priority class, with known first two moments affect system performance. The ME solutions are based on both GE-type and simulated mean queue length and idle state probability constraints and are validated against simulations using different forms of general distributions given known the first two moments. The simulation results are produced at 95% confidence intervals[1] by making use of the queueing network analysis package QNAP-2 [VERA,84].

1: The tolerance is within 5% of the simulated value.

The ME marginal queue length distributions, denoted by ME1 and ME2 for case 1 and case 2, respectively, of a stable G/G/1 under either PR or HOL service discipline with 4 priority classes are validated in examples (5.1)-(5.10) against simulations (denoted by SIM) involving Erlang-2 (E_2), exponential (M), 'balanced' hyperexponential-2 (H_2), (κ=2) and GE distributions with given first two moments. It can be observed that the accuracy of both ME solutions, under various utilisations and coefficients of variation per class, is consistently comparable to that of simulation models with absolute deviations usually less than 0.1. For class-1 jobs the ME queue length distribution $P_1(n_1)$ for PR discipline which is equivalent to the exact solution of an ordinary single-class GE/GE/1 queue, is fully explored in [KOUV,88a] and therefore omitted in the graphs.

This validation study is further enhanced by considering PR and HOL G/G/1 queues with heteregeneous types of interarrival-time and service-time distributions per class including, in addition, deterministic (D) and uniform (U) models. Clearly, these example queues are strikingly more complex than those considered earlier and, as a consequence, the marginal $\langle n_r \rangle$ and $P_r(0)$, r=1,2,...,R constraints are determined by simulations. To assess the robustness of the resulting hybrid ME1 and ME2 solutions, experiments on various PR and HOL G/G/1 queues with two and three heterogeneous priority classes (examples 5.11-5.14) are carried out in figures (5.11)-(5.14c) and favourable comparisons against simulation results are made.

These experiments indicate that the mean queue lengths, $\{\langle n_r \rangle\}$ and where appropriate, the idle state probabilities, $\{P_r(0)\}$, are sufficient constraints enabling the ME1 and ME2 solutions to predict

the shape of the entire distribution with considerable accuracy. Note that the ME2 approach is generally more robust than that of the ME1, as expected, due to the additional use of prior information via constraints $\{P_r(0)\}$, r=1,2,...R.

5.5 Conclusion

Entropy maximization, subject to two different sets of prior information, drawn from the normalisation, utilisation, mean queue length and idle state probability constraints (case 1 and case 2) is applied to characterise new product or quasi-product form approximations for the joint queue length distributions of both PR and HOL stable G/G/1 queues with R priority classes. Robust 'one-step' recursions are established and two closed-form approximations (ME1 and ME2) for the marginal state probabilities per priority class are derived.

Moreover, these results are used as a basis to analyse a GE/G/1 priority queue under either PR or HOL discipline , and provide new approximations for the mean and squared coefficient of variation of the effective priority service-time distribution in the context of shadow-CPU based methods.

As a consequence of this analysis, a new approximation for the squared coefficient of variation of the interdeparture time process per class is proposed.

Illustrative numerical examples on the marginal queue length distributions are used to demonstrate the credibility of the ME approximations for general single queues with priorities against numerous simulations involving homogeneous and heteregeneous external interarrival-time and service-time distributions per class. The comparative study indicates that the marginal mean queue length and where appropriate, the idle state probabilities per class, are

sufficient constraints enabling the ME solutions to predict the shape of the entire queue length distributions with considerable accuracy.

Finally, The ME solutions of a G/G/1 priority queue derived in this chapter are used in conjunction with classical results (see chapter 4), as a building block for the analysis of general open priority QNM's in the next chapter.

CHAPTER 6

ME ANALYSIS OF GENERAL OPEN QNM's WITH PRIORITIES

In this chapter we investigate open queueing networks at equilibrium with infinite capacities, single servers, multiple job classes, distinct general exogeneous interarrival-time and service-time distributions per class, non-priority (FCFS, LCFS with or without preemptions, PS) or priority (PR or HOL) service disciplines and random routing. In particular, we are interested in a ME solution of such networks primarily because they form the basis for the approximate maximum entropy analysis of closed queueing networks. Such ME approximation suggests a decomposition of an open network into individual multiple class G/G/1 queues at equilibrium with a revised arrival process for each class of jobs.

In this context, the ME solutions of both G/G/1 non-priority and priority queues (c.f. chapters 3 and 4, respectively), are used as building blocks to establish a universal implementation of the ME solution of general open queueing networks with mixed service disciplines and multiple classes.

The analysis is carried out by making use of the robust and versatile GE distribution to model the interarrival-time and service-time of each centre per class. Note that, work on open queueing networks with priorities and general service-time has not yet been reported in the literature.

In section one, we present the ME solution of an open network and suggest an approximation for the marginal queue length distributions given the constraints of utilisations of class-r at centre-i, ρ_{ir}, the marginal queue lengths, $<n_{ir}>$, and when appropriate, the idle

state probabilities, $P_{ir}(0)$.

In section two, we examine the robustness of the universal GE-type flow formulae and present subsequently a stepwise description of a Universal Maximum Entropy (UME) algorithm for the approximate solution of general open queueing networks with mixed service disciplines. The evaluation of system performance measures such as system mean reponse-time per class, in the case where jobs switch class membership as they move from one centre to another, is given in appendix E (section E1).

In section three, we review briefly some existing approximate techniques (i.e., [REIS,74; GELE,76; SEVC,77b; KOUV 85]) which are based on class composition and disaggregation for the solution of general FCFS open QNM's and we suggest an extension to these methods in order to analyse networks where one centre (or more) is subject to PR or HOL scheduling disciplines. More precisely, the notion of the virtual servers (shadow CPU) dedicated to each priority class is adopted by making use of the new GE-type priority formulae given in chapter four.

Numerical validation results, involving ME, simulation and other known approximate methods are included in section four.

Final remarks and comments on the extension of the work follow in section five.

6.1 The ME Approximation

Consider an arbitrary open network at equilibrium with M infinite capacity queues consisting of single servers and subject to abstract service disciplines (i.e., PR, HOL ,FCFS,LCFS with or without preemptions, PS). Jobs of the network belong to R classes and arrive from an external source according to general distributions with mean

$1/\lambda_{0r}$ and squared coefficient of variation C^2_{a0r}, r=1,2...,R, the service-time distributions are characterised by the mean values, $\langle S_{ir} \rangle = 1/\mu_{ir}$ and squared coefficients of variation, C^2_{sir}, i=1,2...M, r=1,2...R. Moreover, jobs may switch class membership as they move from one queue to another.

Let $P_{ir;js}$, i,j= 1,2,...,M, r=1,2,...,R be a transition (1st order Markov chain) matrix describing the routing in the network which is the probability of a class-r job having just completed the service at centre-i joins queue-j in class-s. The imaginary centre 0 represents the outside world where the corresponding class index is redundant (i.e., $P_{0r;js}$ r=1,2,...,R is denoted $P_{0;js}$).

Obviously, the stochastic aspect of the transition matrix $\{P_{ir;js}\}$, requires that for each centre-i and class-r, the following relation must be satisfied:

$$\sum_{j=1}^{M} \sum_{s=1}^{R} P_{ir;js} + P_{ir;0} = 1 \qquad (6.1)$$

The joint state of the network is described by a vector $\underline{\underline{n}}=(\underline{n}_1,\underline{n}_2,\ldots,\underline{n}_M)$, where $\underline{n}_i$ is the state of an individual queue-i (i.e., $\underline{n}_i=(n_{i1},n_{i2},\ldots,n_{iR})$, $n_{ir}\geq 0$, r=1,2,...,R). In this context, $P(\underline{\underline{n}})$ is the equilibrium probability that the system is in state $\underline{\underline{n}}$ and $P_i(\underline{n}_i)$ is the state probability of centre-i and is obtained by applying the law of total probability, namely

$$P_i(\underline{k}) = \sum_{\underline{\underline{n}}=\underline{\underline{0}} \wedge \underline{n}_i=\underline{k}}^{\underline{\underline{\infty}}} P(\underline{\underline{n}}) \qquad (6.2)$$

where $\underline{k}=(k_1,k_2,\ldots,k_R)$ is a constant R-dimensional vector.

Suppose that the following mean value constraints about the state probability $P(\underline{\underline{n}})$ are known to exist:

i/ the normalisation constraint,

$$\sum_{\underline{\underline{n}}=\underline{\underline{0}}}^{\underline{\underline{\infty}}} P(\underline{\underline{n}}) = 1 \qquad (6.3)$$

ii/ the utilisation constraints,

$$\sum_{\underline{\underline{n}}=\underline{\underline{0}}}^{\underline{\underline{\infty}}} h_{ir}(\underline{n}_i)P(\underline{\underline{n}}) = \rho_{ir} \;,\; \begin{matrix} i=1,2,\ldots,M \\ r=1,2,\ldots,R \end{matrix} \qquad (6.4)$$

where
$$h_{ir}(\underline{n}_i) = \begin{cases} 1 \text{ if class-r job is in service} \\ \\ 0 \text{ otherwise} \end{cases}$$

iii/ the mean queue length constraints,

$$\sum_{\underline{\underline{n}}=\underline{\underline{0}}}^{\underline{\underline{\infty}}} n_{ir}P(\underline{\underline{n}}) = \langle n_{ir} \rangle \;,\; \begin{matrix} i=1,2,\ldots,M \\ r=1,2,\ldots,R \end{matrix} \qquad (6.5)$$

The form of the solution, $P(\underline{\underline{n}})$ can be completely specified by maximising the entropy functional of $P(\underline{\underline{n}})$, i.e.,

$$H(P) = -\sum_{\underline{\underline{n}}=\underline{\underline{0}}}^{\underline{\underline{\infty}}} P(\underline{\underline{n}}) \log[\, P(\underline{\underline{n}}) \,] \qquad (6.6)$$

subject to the constraints (6.3)-(6.5), leading to the product-form solution,

$$P(\underline{\underline{n}}) = \frac{1}{Z} \prod_{i=1}^{M} \prod_{r=1}^{R} g_{ir}^{h_{ir}(\underline{n}_i)} x_{ir}^{n_{ir}} \qquad (6.7)$$

where Z is the normalising constant and g_{ir}, x_{ir} are the Lagrangian coefficients corresponding to the utilisation and mean queue length constraints, respectively.

Note that although at this stage the expected values ρ_{ir} and $\langle n_{ir} \rangle$ are not explicitly known in terms of system parameters λ_{ir}, C_{air}, μ_{ir} and C_{sir} they can be incorporated into the ME formalism in order to determine the form of the ME solution $P(\underline{\underline{n}})$.

Notice that the ME expression (6.7) is decomposed into a product-form station-by-station solution, namely

$$P(\underline{\underline{n}}) = \prod_{i=1}^{M} P_i(\underline{n}_i) \qquad (6.8)$$

where $P_i(\underline{n}_i)$ is the ME state probability of centre-i and is given analytically by .

$$P_i(\underline{n}_i) = \begin{cases} (1-\rho_i) \dfrac{(n_i-1)!}{A_i} x_{ir}^{n_{ir}} \left[\sum_{r=1}^{R} g_{ir}^{n_{ir}} \right], & \text{(FCFS,PS,LCFS LCFS-NONPR)} \\ (1-\rho_i)\, g_{ir} \prod_{s=r}^{R} x_{is}^{n_{is}}, \quad (n_{i1}=\ldots=n_{ir-1}=0, n_{ir}>0) & \text{(PR)} \\ (1-\rho_i) \prod_{s=1}^{R} x_{is}^{n_{is}} \left[\sum_{s=1 \wedge n_s \neq 0}^{R} g_{is} \right], & \text{(HOL)} \end{cases} \qquad (6.9)$$

$$\text{where } n_i = \sum_{r=1}^{R} n_{ir} \text{ and } A_i = \prod_{r=1}^{R} n_{ir}!$$

$$\text{and } x_{ir} = \begin{cases} \dfrac{\langle n_{ir}\rangle - \rho_{ir}}{\langle n_i\rangle}, & \text{(FCFS,LCFS,LCFS-NONPR,PS)}, \quad r=1,2..,R \\ \dfrac{\langle n_{ir}\rangle - \rho_{ir}}{\langle n_{ir}\rangle + \gamma_{ir-1}}, & \text{(PR)}, \quad r=1,\ldots,R \\ \dfrac{\langle n_{ir}\rangle - \rho_{ir}}{\langle n_{ir}\rangle + \rho_i - \rho_{ir}}, & \text{(HOL)}, \quad r=1,\ldots,R \end{cases} \tag{6.10}$$

$$\text{where } \langle n_i\rangle = \sum_{r=1}^{R} \langle n_{ir}\rangle, \quad \rho_i = \sum_{r=1}^{R} \rho_{ir}, \quad \gamma_{ir} = \sum_{s=1}^{r} \rho_{is}$$

$$g_{ir} = \begin{cases} \dfrac{\rho_{ir}}{\langle n_{ir}\rangle - \rho_{ir}} \dfrac{\rho_i}{1-\rho_i}, & \text{(FCFS,LCFS,LCFS-NONPR,PS)} \quad r=1,\ldots,R \\ \dfrac{\rho_{ir}}{\langle n_{ir}\rangle - \rho_{ir}} \dfrac{\gamma_{ir}}{1-\rho_i} \prod_{s=r+1}^{R} \dfrac{\gamma_{is}}{\langle n_{is}\rangle + \gamma_{is-1}}, & \text{(PR)}, \quad r=1,\ldots,R \\ \dfrac{\rho_{ir}}{\langle n_{ir}\rangle - \rho_{ir}} \dfrac{\rho_i}{1-\rho_i} \prod_{s=1 \wedge s\neq r}^{R} \dfrac{1}{\langle n_{is}\rangle - \rho_{is} + \rho_i}, & \text{(HOL)} \quad r=1,\ldots,R \end{cases} \tag{6.11}$$

The ME solution of the network, $P(\underline{n})$, at equilibrium, exhibits a decomposition station-by-station solution, implying each centre of the network to be analysed in isolation. To this end, if centre-i is under PR or HOL scheduling discipline, the following marginal idle state probability constraints, $P_{ir}(0)=1-\theta_{ir}$, $r=1,2,\ldots,R$, can also be incorporated into the ME formalism in order to determine the form

of the ME solution $P_i(n_i)$,

$$\sum_{\underline{n}=\underline{0}}^{\infty} V_{ir}(\underline{n}_i)P(\underline{n}) = \theta_{ir} = 1-P_{ir}(0), \quad r=1,\ldots,R \tag{6.12}$$

where $$V_{ir}(\underline{n}_i)= \begin{cases} 1 \text{ if } n_{ir} > 0 \\ \\ 0 \text{ otherwise} \end{cases}$$

with $P_i(\underline{n}_i)$ given by:

$$P_i(\underline{n}_i) = \begin{cases} (1-\rho_i)\, g_{ir} y_{ir} \prod_{s=r}^{R} x_{is}^{n_{is}} y_{is}^{V_{is}(\underline{n}_i)} & (n_{i1}=\ldots=n_{ir-1}=0, n_{ir}>0) \\ & \text{(PR)} \\ (1-\rho_i) \prod_{s=1}^{R} x_{is}^{n_{is}} y_{is}^{V_{is}(\underline{n}_i)} \left[\sum_{s=1 \wedge n_{is}\neq 0}^{R} g_{is} \right] & \text{(HOL)} \end{cases} \tag{6.13}$$

where the Lagrangian coefficients x_{ir}, g_{ir} and y_{ir} are given by:

$$x_{ir} = \frac{<n_{ir}>-\theta_{ir}}{<n_{ir}>} \quad \text{(PR or HOL)} , \quad r=1,\ldots,R \tag{6.14}$$

$$g_{ir} = \begin{cases} \dfrac{\rho_{i1}}{1-\rho_i} \dfrac{\rho_{i1}}{<n_{i1}>-\rho_{i1}} \prod_{s=2}^{R} \dfrac{\gamma_{is} - \theta_{is}}{\gamma_{is-1}} , & r=1, \text{ (PR)} \\ \\ \dfrac{\rho_{ir}}{1-\rho_i} \dfrac{\gamma_{ir}-\theta_{ir}}{\theta_{ir}-\rho_{ir}} \prod_{s=r+1}^{R} \dfrac{\gamma_{is} - \theta_{is}}{\gamma_{is-1}} , & r=2,\ldots,R, \text{ (PR)} \\ \\ \dfrac{\rho_{ir}}{1-\rho_i} \dfrac{\rho_i-\theta_{ir}}{\theta_{ir}-\rho_{ir}} \prod_{s=1 \wedge s\neq r}^{R} \dfrac{\rho_i - \theta_{is}}{\rho_i-\rho_{is}} , & r=1,\ldots,R, \text{ (HOL)} \end{cases} \tag{6.15}$$

$$y_{ir} = \begin{cases} 1, \ r=1, \ (PR) \\ \dfrac{\theta_{ir} - \rho_{ir}}{\gamma_{ir} - \theta_{ir}} \dfrac{1 - x_{ir}}{x_{ir}}, \ r=2,\ldots,R, \ (PR) \\ \dfrac{\theta_{ir} - \rho_{ir}}{\rho_{i} - \theta_{ir}} \dfrac{1 - x_{ir}}{x_{ir}}, \ r=1,\ldots,R, \ (HOL) \end{cases} \quad (6.16)$$

Meanwhile, at the steady state, the ME analysis of centre-i, under FCFS, PS, LCFS or LCFS-NONPR subject to constraints (6.2)-(6.5) and in addition the marginal idle state probability constraints (6.12), has been found very difficult to tackle analytically and as a consequence, no closed-form expressions for the marginal probabilities and the Lagrangian multipliers are derived.

Nevertheless, at equilibrium, the marginal queue length distribution per class for each centre-i, i=1,..,M, given the normalisation, utilisation, mean queue length and when appropriate the idle state probability (if PR or HOL) constraints have modified geometric form, namely

$$P_{ir}(n_{ir}) = \begin{cases} \hat{\rho}_{ir}(1 - \hat{x}_{ir})\hat{x}_{ir}^{n_{ir}-1}, & n_{ir} > 0 \\ & i=1,\ldots,M, \ r=1,\ldots,R \\ 1 - \hat{\rho}_{ir}, & n_{ir} = 0 \end{cases} \quad (6.17)$$

$$\hat{x}_{ir} = \begin{cases} x_{ir} & (PR, \ HOL) \\ & i=1,\ldots,M, \ (6.17a) \\ & r=1,\ldots,R \\ \dfrac{\langle n_{ir} \rangle - \rho_{ir}}{\langle n_{ir} \rangle + \rho_{i} - \rho_{ir}}, & (FCFS, PS, LCFS, LCFS-NONPR) \end{cases}$$

$$\hat{\rho}_{ir} = \begin{cases} \theta_{ir} \ , & \text{(PR, HOL-ME2)} \\ \rho_{ir} + \gamma_{ir-1} x_{ir} \ , & \text{(PR-ME1)} \\ \rho_{ir} + (\rho_i - \rho_{ir}) x_{ir}, & \text{(HOL-ME1), } i=1,\ldots,M \text{ (6.17b)} \\ & \qquad r=1,\ldots,R \\ \dfrac{\rho_i \langle n_{ir} \rangle}{\langle n_{ir} \rangle + \rho_i - \rho_{ir}} \ , & \text{(FCFS,PS,LCFS,LCFS-NONPR)} \end{cases}$$

These ME solutions are used in conjunction with the new GE-type formulae (c.f. sections 3.4.3, 4.6.1 and 4.6.2) as building blocks for the ME approximation of an arbitrary open queueing network in the next section.

6.2 Universal Maximum Entropy (UME) algorithm

The ME approximation (6.8) suggests a decomposition of the network into individual multiple class G/G/1 queues with revised arrival process for each class of jobs. In order to implement the ME solution under an abstract discipline, the flow process in general network should be determined. It is assumed that for each centre-i, i=1,2,..,M, the arriving and departing streams form renewal processes conforming with GE-type underlying interarrival-time and service-time distributions (with known first two moments).It remains now to evaluate the mean rates and squared coefficients of variation of the interarrival and interdeparture processes of class-r, r=1,2,...,R, jobs at queue-i, i-1,...,M.

6.2.1 The interdeparture-time processes

Since in all cases (ME1, ME2 for PR and HOL) and under any service discipline, the steady state ME solution for the marginal queue length distributions of class-r at centre-i, $P_{ir}(n_{ir})$, i=1,2,..,M, r=1,2,...,R are of modified geometric-type (6.17), the mean rate, λ_{dir}, and the squared coefficient of variation, C^2_{dir}, of the interdeparture-time per class can be approximated under an abstract service discipline by considering a virtual FCFS GE/GE/1 queue exclusively dedicated to jobs of class-r with a queue length distribution identical to the marginal ME solution $P_{ir}(n_{ir})$, r=1,2,..., R of the original multiple class GE/G/1 queue (c.f. section 5.3). Thus, by analogy to equations (5.37) and (5.38), we have at equilibrium

$$\lambda_{dir} = \lambda_{ir} \tag{6.18}$$

$$C^2_{dir} = 2\langle n_{ir}\rangle(1-\hat{\rho}_{ir}) + C^2_{air}(1-2\hat{\rho}_{ir}) \tag{6.19}$$

Note that, although the notion of the virtual server is used to estimate the squared coefficient of variation of the interdeparture time of class-r jobs at centre-i, the ME method does not require explicitly the creation of ficticious servers (the configuration of the network is not modified), since the quantity in question, (C^2_{dir}), does not depend directly on the parameters of the 'shadow' servers. Moreover, the effect of each service discipline is reflected via the particular form of the performance measures $\langle n_{ir}\rangle$ and $\hat{\rho}_{ir}$. The statistics $\langle n_{ir}\rangle$ for GE/G/1 queue are given by equations (4.13) and (4.14) for PR and HOL, respectively and by (3.38)-(3.41) for the non-priority disciplines.

6.2.2 The splitting process

For each queue-j, and since jobs may switch class membership, any departing class-s stream denoted by (js)-stream can produce MR potential streams denoted by (js;ir)-stream, directed to queue-i as class-r. Let $\lambda_{djs;ir}$ and $C^2_{djs;ir}$ be the mean rate and squared coefficient for such departing streams. Assuming that the interdeparture process per class from each centre is renewal, the first two moments are determined directly by using the splitting formulae per class [GELE,80; KOUV.,85], namely

$$\lambda_{djs;ir} = \lambda_{js} P_{js;ir} \tag{6.20}$$

$$C^2_{djs;ir} = 1 + P_{js;ir}(C^2_{djs}-1) \tag{6.21}$$

where C^2_{djs} is the squared coefficient of variation of the departing stream of class-s from centre-j which is given by (6.19).

6.2.3 The merging process

The (overall) arrival process of class-r jobs to queue i, i=1,2...,M, is clearly the merging process of MR+1 potential streams of class-r. Assuming the interevent-time distribution of such streams is approximated by GE distributional model with parameters $\lambda_{djs;ir}$ and $C^2_{djs;ir}$, it follows from Kouvatsos [KOUV,85] that the merging stream has also a GE-type interarrival-time distribution with parameters

$$\lambda_{ir} = \lambda_{0;ir} + \sum_{j=1}^{M} \sum_{s=1}^{R} \lambda_{djs;ir} \ , \ i=1,\ldots,M; \ r=1,\ldots,R \tag{6.22}$$

$$C^2_{air} = \left\{ \sum_{j=1}^{M} \sum_{s=1}^{R} \frac{\lambda_{djs;ir}}{\lambda_{ir}} \left(C^2_{djs;ir+1}\right)^{-1} + \frac{\lambda_{0;ir}}{\lambda_{ir}} \left(C^2_{a0;ir+1}\right)^{-1} \right\}^{-1} - 1 \qquad (6.23)$$

6.2.4 UME algorithm

A UME algorithm for general multiple class open network of queues with mixed service disciplines is based on the previous analytic approximations and involves the following steps:

ALGORITHM 6.1: UME algorithm for multiple class general open QNM's with mixed service discipline

INPUT

M,R,

$\{P_{ir;js}\}$:transition probability matrix,

λ_{0r}: external mean arrival rate of class-r jobs, r=1,2...,R,

C^2_{a0r}: squared coefficient of variation of the external interarrival-time of class-r jobs, r=1,...,R.

μ_{ir} : mean service rate of class-r jobs at centre i, i=1,...M, r=1,...R,

C^2_{sir} : squared coefficient of variation of the service time of class-r at centre i, i=1,...,M, r=1,....,R,

DS_i : type of discipline of centre i (FCFS, LCFS, LCFS-NONPR, PS, PR, HOL).

STEP 1 {* Solve the job-flow balance equations for i=1,...,M, r=1,...R, (c.f. (eq. 6.22) *}

STEP 2 {* Compute *}

. $\lambda_{djs;ir}$ (c.f. (eq. 6.20));

. $\rho_{ir} = \lambda_{ir}/\mu_{ir}$, i=1,...,M , r=1,...,R;

.. {* check ρ_i < 1 for i=1,...,M *};

STEP 3 {* Solve the flow equations (6.17)-(6.23) with respect to C^2_{air} i=1,...,M, r=1,...,R *};

STEP 3.1 {* Initialisation of C^2_{air} *}

$C^2_{air} \longleftarrow 1$, for i=1,...,M, r=1,...R;

STEP 3.2 {* Iterate until convergence of C^2_{air} *}

STEP 3.2.1 {* Compute the performance measures $\langle n_{ir} \rangle$, $\hat{\rho}_{ir}$ *}

. $\langle n_{ir} \rangle \longleftarrow f_i(C^2_{ai\ell}, C^2_{si\ell}, \lambda_{i\ell}, \mu_{i\ell}, \ell=1,...,R)$

(c.f (4.13) and (4.14) for PR and HOL respectively and (3.38)-(3.41) for FCFS, LCFS, LCFS-NONPR, PS), i=1,...M, r=1,...R;

. $\hat{\rho}_{ir}$ (c.f. (6.17b)), i=1,...M, r=1,...R;;

STEP 3.2.2 {* Compute C^2_{dir} *}

. C^2_{dir} (c.f.(6.19)) , i=1,...M, r=1,...R;

. $C^2_{djs;ir}$ (c.f.(6.21)), i=1,...M, r=1,...R;

STEP 3.2.3 {* Compute new value of C^2_{air} *}

. C^2_{air} (c.f. (6.23)), i=1,...M, r=1,...R;

STEP 4 {* Evaluation of the performance measures of the network*}

STEP 4.1 {* Obtain the performance measures, $\langle n_{ir} \rangle$ and $\hat{\rho}_{ir}$ from the last iteration *};

STEP 4.2 {* Estimate the ME marginal probabilities *}

- $P_{ir}(n_{ir})$ (c.f. (6.17));

END

Note that algorithm 4.1 may be used in step 3.2.1 if centre-i is under PR or HOL discipline when adapting the ME2 approximation.

Furthermore, step 3 constitutes the major computational cost of the algorithm above. This step is iterative and as a result, the number of operations of the algorithm increases with the number of iterations. The speed of the convergence generally depends upon the initial value of C^2_{air}, i=1,2,...,M and r=1,2,...,R. On the orher hand, the computational cost of step 3 per iteration may be estimated as follows:

If d_{ir} denotes the number of operations required in step 3.2.1 per centre-i and per class-r, the number of operations required by step 3.2 can be evaluated by:

$$O\left[(M+1)^2R^2 + \sum_{i=1}^{M} \sum_{r=1}^{R} d_{ir} \right]$$

where d_{ir} can be estimated as $O[R]$.

However, if centre-i is under priority discipline (PR or HOL) and the ME2 approach is adopted, the computation of the idle state probability, $P_{ir}(0)$, can be very costly for large number of classes with an estimated time complexity derived as $d_{ir} = O[5^{r-1}]$ (c.f. section 4.6.2). It is therefore recommanded to use ME1 approximation for large number of classes. In addition, equations (6.19), (6.21) and (6.23) form a set of MR non-linear independent equations with MR unknowns (i.e., $C^2_{air} = \Phi(C^2_{air})$, i=1,2,...,M, r=1,2,...,R with Φ having non-negative values). Solutions of such systems exist, but it is very difficult to prove analytically the convergence of the fixed point method used in algorithm 6.1. This is due to the complexity of the formulae involved (c.f. $<n_{ir}>$, $P_{ir}(0)$, ρ_{ir}) although experimentally, the ME algorithms have never failed (c.f. [KOUV,85]).

The system mean response time per class when jobs switch class membership as they move from one centre to another, is determined by using the concept of 'equivalence class' [BRUE,80]. The evaluation of this statistic is given in appendix E (section E1).

It is also interesting to point out that the UME algorithm for general open networks with mixed service disciplines provides to the knowledge of the authors, the only tool available to date for calculating analytically an approximate solution for the marginal queue length distribution per class, $P_{ir}(n_{ir})$. In particular, it is the only algorithm for the approximate analysis of general open networks with priorities.

6.3 Maximum Entropy Reduced occupancy approximation (ME-ROA)

In this section, we examine how the present approximations, based on class composition and disaggregation, in solving general open QNM's with FCFS centres, can be adjusted to incorporate the priority disciplines (PR or HOL) by making use of the ME GE-type priority formulae. In particular, we suggest an extension to the previous ME method proposed by Kouvatsos [KOUV,85] in order to solve general open networks with priorities.

First of all, in all methods, the arrival rates of class-r jobs at centre-i are easily obtained by assuming the conservation of job flows (the rate at which jobs arrive at each queue-i is equal to the rate at which they depart from that queue) and are obtained by solving the linear system of equations (6.22). Furthermore, all current approximations for general open networks with FCFS centres, involve a class composition and disaggregation techniques where each centre is analysed in isolation as a G/G/1 queue with a single class of jobs. In addition, the estimation of the squared coefficients of

variation of the flow process requires the aggregation of the classes into a single composite class due to the non-existence of analytic approximation for the coefficient of variation of the interdeparture-time per class, C_{dir}.

To identify the coefficient of variation of the interarrival-time distribution of the composite class at centre-i, C_{ai}, three issues have to be addressed:

i/ the splitting process,

$$C_{dji}^2 = f^{(s)}(P_{ji}, C_{di}^2) \qquad (6.24)$$

ii/ the merging process,

$$C_{ai}^2 = f^{(m)}(\lambda_{ji}, C_{dji}^2, j=0,\ldots,M) \qquad (6.25)$$

iii/ the departure process,

$$C_{di}^2 = f^{(d)}(\rho_i, C_{ai}^2, C_{si}^2) \qquad (6.26)$$

where P_{ji} is the class independent routing frequency that job belonging to the composite class leaving centre-j is directed to centre-i. For each centre-i, i=1,...M, C_{ai}, C_{si}, C_{di} are the aggregate coefficient of variation of the interarrival, service and interdeparture time, respectively.

The analysis of the splitting process is rather straightforward and yields exact result for C_{dji} [GELE,80; KOUV,85] which is given by:

$$C_{dji}^2 = 1 + P_{ji}(C_{dj}^2 - 1) \qquad (6.27)$$

However the analysis of the merging and departure processes depends on the general (G-type) distribution chosen to model the interarrival and service time distributions. As a consequence, the corresponding squared coefficients of variation are estimated in approximate manner, depending on the approach adopted.

Once the first two moments of the flow are estimated, it remains only to determine the statistics of the composite class which of course depends on the parameters and the type of the interarrival and service time distributions. i.e.,

$$\langle n_i \rangle = f^{(q)}(\rho_i, C_{ai}^2, C_{si}^2) \qquad (6.28)$$

Finally, given that jobs are assumed to be served in FCFS fashion at every centre, the statistics per job class are easily obtained by using Little's law [LITT,61] and are given by

$$\langle n_{ir} \rangle = \lambda_{ir}\left[\frac{\langle n_i \rangle - \rho_i}{\lambda_i}\right] + \frac{\rho_{ir}}{\lambda_{ir}} \qquad (6.29)$$

The analytic functions $f^{(m)}(.)$, $f^{(d)}(.)$ and $f^{(q)}(.)$ generally differ from one approximation to another. For example, the 'old' ME approximation for the multiple class open networks [KOUV,85], uses the following expressions:

$$f^{(m)}(\lambda_{ji}, C_{dji}) \longleftarrow -1 + \left\{\sum_{j=0}^{M} \frac{\lambda_{ji}}{\lambda_i}(C_{dji}^2 + 1)^{-1}\right\}^{-1} \qquad (6.30a)$$

$$f^{(d)}(\rho_i, C_{ai}^2, C_{si}^2) \longleftarrow \rho_i(1-\rho_i)+(1-\rho_i)C_{ai}^2 + \rho_i^2 C_{si}^2 \qquad (6.30b)$$

$$f^{(q)}(\rho_i, C_{ai}^2, C_{si}^2) \longleftarrow \frac{\rho_i}{2}\left[1 + \frac{C_{ai}^2 + \rho_i C_{si}^2}{1 - \rho_i}\right] \quad (6.30c)$$

Some other approximations are listed in appendix E (section E2). Their extension to solve priority QNM's, is carried out by appropriate substitutions of the functions $f^{(m)}(.)$, $f^{(d)}(.)$ and $f^{(q)}(.)$.

Since all current approximations for general open networks with FCFS stations, are based on class composition and disaggregation schemes, performance measures of queueing networks with priorities cannot be determined accurately with these techniques. However, to extend these methods to priority situations , it is necessary to create virtual and dedicated FCFS server to each priority class with inflated service time to take into account the degradation of the low-service time due to the presence of high-priority jobs.

In the context of ROA technique [SEVC,77a], the configuration of the original network is modified to capture the effect of higher priority jobs on the system performance so that each priority queue in an open network is replaced by R virtual (shadow) FCFS queues, each one of them exclusively serves jobs of one class-r, r=1,2,...R, according to GE-type effective priority service time distribution with parameters $\hat{\mu}_{ir}$ and $\hat{C}_{sir}^2$ satisfying the relations (5.32) and (5.36), respectively.

By adjusting the values of the transition probabilities $\{P_{ir;js}\}$ and the parameters λ_{or}, C_{aor} of the external interarrival process per class-r, the original open network with R priority classes and M centres is transformed into a non-priority FCFS open network with R classes and M^*, $M^* = m(R-1)+M$ (m is the number of priority centres), queues belonging to two sets, $\hat{Q}$ and Q, associated with virtual and

ordinary centres, respectively.

To this end, the 'old' ME method [KOUV,85] can be applied to produce an aggregate product-form approximation, namely

$$P(\underline{n}) = \prod_{i=1}^{M^*} P_i(n_i) \qquad (6.31)$$

Subject to normalisation and the aggregate constraints of utilisation, ρ_i, and mean queue length, $\langle n_i \rangle$, for each queue-i, where $P(\underline{n})$ is the aggregate state probability of the network with $\underline{n}=(n_1,\ldots,n_{M^*})$, n_i being the number of jobs in queue-i and $P_i(n_i)$ is the marginal aggregate state pro bability of queue-i, i=1,...,M^*. The ME solution (6.31) implies a decomposition of the network into stable FCFS G/G/1 queues {i}, with a single aggregate class of jobs. Thus each G/G/1 queue can be analysed in isolation by determining iteratively the first two moments of the effective priority service-time and the flow (interdeparture-time, splitting and merging) processes via the appropriate GE-type approximation formulae (c.f. (5.35),(5.36)).

A stepwise presentation of the ME-ROA approximation is given next.

Algorithm 6.2: ME-ROA for general open queueing network with priorities.

INPUT

M,R,

$\{P_{ir;js}\}$:transition probability matrix,

λ_{0r}: external mean arrival rate of class-r jobs, r=1,2...,R,

C^2_{a0r}: External squared coefficient of variation of the interarrival-time of class-r jobs, r=1,...,R.

μ_{ir} : mean service rate of class-r jobs at centre-i, i=1,...M, r=1,...R,

C^2_{sir} : squared coefficient of variation of the service time of class-r at centre-i, i=1,...,M, r=1,....,R,

DS_i : type of discipline of centre-i (FCFS, PR HOL).

STEP 1 {* Solve the job-flow balance equations for i=1,...,M, r=1,...R, (c.f. (eq. 6.22) *}

STEP 2 {* Transform the original open network with M FCFS,PR and HOL single server queues into an open network with M^*, M^* <—— (R-1)m + M, FCFS queues *)

STEP 2.1 {* Adjust the values of the new sets of transition probabilities $\{\hat{P}_{ir;js}\}$, i,j=1,...,M^*, r,s=1,...,R, according to the new FCFS network configuration *}

STEP 2.2 {* Adjust the parameters of the virtual servers *}

for i=1,...,$M^*$and r=1,...,R do {step 2.2}

begin

. C^2_{air} <—— 1 ;

if i ϵ $\hat{Q}$ then

begin

. $\hat{\mu}_{ir}$ <—— $1/\langle \hat{S}_{ir} \rangle$; (c.f. (5.35))

. $\hat{C}^2_{sir}$ <—— $(2-\hat{\tau}_{ir})/\hat{\tau}ir$; (c.f. (5.36))

end

else {* if i ϵ Q *}

begin

. $\hat{\mu}_{ir}$ <—— μ_{ir};

. $\hat{C}_{sir} \longleftarrow C_{sir}$;

end

end;

STEP 3 {* obtain the paramaters of the composite class *)

STEP 3.1

$$\text{. } \lambda_0 \longleftarrow \sum_{r=1}^{R} \lambda_{0r};$$

$$\text{. } C_{a0}^2 \longleftarrow \left\{ \lambda_0 / \sum_{r=1}^{R} \lambda_{0r}(C_{a0r}^2 + 1)^{-1} \right\} - 1 \ ;$$

$$\text{for } i=,\ldots,M^* \text{ do } P_{0i} \longleftarrow \sum_{r=1}^{R} \lambda_0/\lambda_{0r} P_{0;ir} \ ;$$

STEP 3.2

for i=1,...,M* do

begin

$$\text{. } \lambda_i \longleftarrow \sum_{r=1}^{R} \lambda_{ir};$$

$$\text{. } \hat{\rho}_{ir} = \lambda_{ir}/\hat{\mu}_{ir} \ , \ r=1 \ ,\ldots,R;$$

$$\text{. } \rho_i \longleftarrow \sum_{r=1}^{R} \hat{\rho}_{ir} \ ;$$

$$\text{. } \mu_i \longleftarrow \lambda_i/\rho_i \ ;$$

$$\text{. } C_{si}^2 \longleftarrow \rho_i^{-1}\mu_i \left[\sum_{r=1}^{R} \hat{\rho}_{ir}(\hat{C}_{sir}^2 + 1)\hat{\mu}_{ir}^{-1} \right] - 1;$$

. $\pi_{ir} \longleftarrow \lambda_{ir}/\lambda_i$, r=1,...,R ;

. $P_{ij} \longleftarrow \sum_{r=1}^{R} \pi_{ir} \sum_{s=1}^{R} \hat{P}_{ir;js}$, j=1,...,M*

end;

STEP 4 {* Solve iteratively with respect to C_{ai}^2, i=1,...,M* *}

{* repeat step 4 until convergence of C_{ai}^2 *}

STEP 4.1 {* evaluate the parameters of the flow processes *}

. $C_{di}^2 \longleftarrow f^{(d)}(\rho_i, C_{ai}^2, C_{si}^2)$; (c.f. (6.30b))

. $C_{dji}^2 \longleftarrow f^{(s)}(P_{ji}, C_{di}^2)$, j=0,...,M*;(c.f. (6.27))

. $C_{ai}^2 \longleftarrow f^{(m)}(\lambda_{ji}, C_{dji}^2)$, j=0,...,M ;(c.f. (6.30a))

STEP 4.2 {* update the values of the parameters $\hat{\mu}_i$ and $\hat{C}_{si}$ for $i \in \hat{Q}$ by using (5.32) and (5.36) as apprpriate and subsequently adjust the new value of $\hat{\rho}_{ir}$, ρ_i, μ_i, C_{si} for $i \in \hat{Q}$ (c.f. step 3.2)*}

STEP 5 {* obtain the marginal statistics *}

STEP 5.1

. $\langle n_i \rangle \longleftarrow f^{(q)}(\rho_i, C_{ai}^2, C_{si}^2)$; (c.f. (6.30c))

. $\langle n_{ir} \rangle \longleftarrow$ (c.f. (6.29));

STEP 5.2 {* Evaluate the Lagrangian coefficients x_{ir}, g_{ir} for $i\epsilon Q$ and x_{ir}, g_{ir} and y_{ir} for $i\epsilon\hat{Q}$ (c.f.(6.10)-(6.11) for ME1 and (6.14)-(6.16) for ME2))*}

STEP 5.3 {* evaluate the marginal queue length probabilities of class-r, r=1,...,R, (c.f. 6.17) *}

END.

Since this algorithm is based on class composition, the computation cost per iteration is estimated as $O[(M^*+1)^2 + \sum_{i=1}^{M^*} d_i]$, where d_i is the time complexity of step 4.2 (that involves the computation of the idle state probabilities) which is comparable in cost to that of algorithm 6.1 (c.f. section 6. 2.4).

6.4 Numerical results and discussions

(See appendix E, section E3)

In this section numerical results are presented to demonstrate the credibility of ME methods (UME, ME-ROA)for GE-type open networks in relation to simulations (SIM) given at 95% confidence intervals and other approximate techniques.

The comparative study focuses on the marginal mean queue lengths, $\langle n_{ir}\rangle$, or the system mean response times per class-r, $\langle Ts_r\rangle$, r=1,...,R, for various open networks of tandem (c.f. Fig 6.1) or cyclic (c.f. Fig. 6.2) configuration with mixed service disciplines.

The ME approximations are obtained by using algorithm 6.1 (UME1 and UME2 for ME1 and ME2 approximation, respectively) or by algorithm 6.2 (ME-ROA).

Table 6.2 displays the system mean response time per class-r, r=1,..,R for various approximations for two Markovian queues in

tandem (M/M/1 ——> ./M/1) with two priority classes (c.f., Table 6.1), $<Ts_r>$ = $<T_{1r}>$ + $<T_{2r}>$ where $<T_{ir}>$, i=1,2, is the class-r mean response time at queue-i, i=1,2, respectively. For exposition purposes the MVA [BRYA,84], the m-ROA [KAUF,84], UME1 and UME2 approximations are presented.

It is observed that for low server utilisations, all the techniques are comparable in accuracy to that of simulation (e.g., experiments 1,8,10,13,14 and 15). This is attributed to the fact that jobs of class-2 are not much influenced by the presence of class-1 jobs . However, for high server utilisations , mostly imposed by high priority class jobs or when μ_{i_2} is much greater than μ_{i_1} the synchronisation error (c.f. section 2.3.4) becomes more influential and the effective priority service-time of class-2 jobs, $<\hat{S}_{i_2}>$, has relatively much higher variability (i.e., $\hat{Cs}_{i_2} > 1$). As a consequence, the m-ROA technique, based on Markovian analysis of separable queueing networks, may produce considerable errors (e.g., experiments 5, 6). Note that the MVA approximation although generally more accurate than the ROA based methods, fails to capture the error due to the high variability of the flow processes, (e.g., experiments 2,5,6, 12). On the other hand, both UME1 and UME2 approaches generate fairly accurate results by reducing the synchronisation error and capturing a great deal of the variability of the interdeparture and interarrival times from and to every centre of the network. Moreover, two (HOL-HOL) queues in tandem with exponential servers are analysed in table 6.4 with the corresponding raw data given in table 6.3. Similarly, it can be seen that the $<Ts_r>$, r=1,2, generated by UME1, UME2 and also those produced by the state dependent-reduced occupancy approximation (sd-ROA) are all very close to the exact solutions supplied by [SCHM,83]. On the other hand, table 6.6 exhibits the $<Ts_r>$, r-1,2, of two (PR-FCFS) queues in tandem having general

interarrival and service time distributions with squared coefficient of variation varying from 1/3 to 30. In this case, none of the present approximations can be used and therefore UME1 and UME2 approximations are the only methods applicable to this type of networks. It is noted that UME2 results are consistently comparable to those of the simulation. The UME1 is somewhat inferior to UME2, particularly as the contribution of high-priority class jobs to the load imposed on the servers increases (e.g., experiments 4, 7, 14). However, It is important to point out that the UME2 approximation requires the exact evaluation of the marginal idle state probability of GE/GE/1 (c.f. algorithm 4.1) and therefore, it can be computationally costly for large number of classes. Consequently, UME1 is recommanded in this case.

The credibility of ME methods is further demonstrated by focussing on two-stage cyclic queues (c.f. fig 7.2) with mixed service disciplines. Tables 6.8a and 6.8b display the marginal mean queue lengths, $\langle n_{ir} \rangle$, r=1,2, with corresponding percentage differences from simulation for centre-1 and centre-2, respectively, in a two-stage Markovian network. It can be observed that the ME2-ROA and UME2 are generally far superior in comparison to ROA [SEVC,77a], m-ROA [KAUF,84], MVA [BRYA,84] methods mainly for high utilisation mostly attributed to high-priority class (e.g., experiments 1, 2, 7, 9 and 10). Note that UME1 and ME1-ROA methods have similar performances to that of ME2-ROA and UME2 and as such has been omitted in the tables. Furthermore, the ROA and m-ROA cannot be applied to networks with HOL centres.

Finally, the two-stage cyclic network with general interarrival and service-time distributions (c.f. Table 6.9) is analysed and the corresponding estimations for the marginal mean queue lengths are depicted in table 6.10a and 6.10b for centre-1 and centre-2

respectively. The service discipline of each station could be either priority (PR or HOL) or non-priority (FCFS, LCFS,LCFSNONPR,PS) based disciplines. It can generally be observed that ME2-ROA and UME2 results are comparable to those obtained by SIM. In particular, the UME2 provides the best overall approximation for GE-type open networks by capturing the influence amongst the various classes of jobs in more direct fashion to that provided by ME2-ROA which uses class composition and disaggregation techniques. Moreover, the computational costs of ME2-ROA and UME2 are comparable in terms of flow iterations when non-priority (FCFS) apply, otherwise the ME2-ROA is clearly less efficient using empirically (ck/R)+k iterations, where c is the number of priority centres (c<M) and k is the corresponding number of iterations for UME2. It is interesting to point out that the ME algorithms produce exact results when appropriate conditions of separability apply [BASK,77]. Note that more numerical results for networks containing FCFS, LCFS, LCFS-NONPR, PS centres can be found in [KOUV,88C; GEOR,89].

6.5 Conclusion

The PME, viewed as inference procedure, is used in this chapter to characterise a new product-form solution for the approximate analysis of arbitrary multiple class open networks of queues at equilibrium with infinite capacities, single servers and mixed service disciplines. The ME approximation implies a decomposition of the network into individual multiple class G/G/1 queues at equilibrium with a revised arrival process for each class of jobs. To this end, the ME solution of a priority G/G/1 queue given in chapter 5 are used to provide approximate solution to more complex type of queueing systems. In particular, the GE distributional model

is used to establish a universal approximation for the flow processes in the network in terms of their first two moments. Moreover, based on class composition and disaggregation, a ME-ROA algorithm for general open networks with mixed (PR, HOL, FCFS) is established . The technique uses the concept of virtual and dedicated server to each priority class (c.f. section 2.3.2). Finally, illustrative numerical examples on marginal mean response time and mean queue lengths at each centre are used to demonstrate the credibility of the ME approximations for general open networks against other known methods relatively to some exact results [SCHM,83] and numerous simulations involving homogeneous and heterogeneous external interarrival-time and service-time distribution per class. The 'new' ME algorithm (UME) seems to capture in a better fashion the influence amongst the various classes in comparison to other approximate methods which are based on class composition and disaggregation (ME-ROA) or the ROA, m-ROA, MVA (applicable only to Markovian networks).

Finally, the GE-type results of open networks obtained by using algorithm 6.1 are used in the next chapter to analyse the corresponding closed networks with priorities satisfying the principles of conservation of flow and conservation of population per class.

CHAPTER 7

General Closed Queueing networks with Priorities

In the previous chapter, approximations of open queueing networks with mixed service disciplines involving priorities are validated. In this chapter the ME analysis is extended to closed QNM's with priorities. The population of each class-r, r=1,2...,R, is bounded and kept fixed (N_r = constant), and external flows are reduced to zero.

In section one, we present the ME product-form approximation of general closed queueing networks with priorities, subject to known mean value constraints, normalisation, utilisations, $\{\rho_{ir}\}$, mean queue lengths, $\{<n_{ir}>\}$ and where appropriate the idle state probabilities $\{P_{ir}(0)\}$ for i=1,...,M, r= 1,...R.

In section two we examine some computational techniques used in the implementation of the ME method. In particular, we will see that the ME method for general closed networks involves two parts:

i/ the fixed population mean solution,

ii/ the ME closed network product-form solution (Convolution).

The fixed population mean solution provides the mean values of the performance measures of a corresponding 'pseudo' open multiple class network having the same transition probabilities and service characteristics as the original closed , but where the mean queue lengths per class (not necessary the number of jobs per class) must satisfy the fixed population mean constraint [KOUV,83], i.e.,

$$\sum_{i=1}^{M} <n_{ir}> = N_r \ , \ r= 1,\ldots,R.$$

This can be carried out by using the UME algorithm 6.1, together with Newton/Raphson iterative method in order to ensure the validity of the relation above. The evaluation of the performance metrics of the pseudo open network enables us to approximate the corresponding Lagrangian coefficients which are convoluted in the second part in order to establish a ME approximation of general closed queueing network with priorities (c.f. section 3.3.1).

In section three, we present an efficient implementation of the ME approximation based on the standard multiple class convolution formulae given in [BRUE,80].

A stepwise presentation of UME algorithm for general closed networks with priorities is presented in section four, followed, in section five by discussions and numerical validation against exact, simulation and current approximations.

Finally, we conclude the chapter by a brief summary of the ME method for the general closed networks.

7.1 ME and closed queueing networks

Consider an arbitrary closed queueing network Q, containing M queueing stations with single general servers and R classes of jobs with priority (PR, HOL) and non-priority (FCFS, PS, LCFS, LCFS-NONPR) scheduling disciplines. For each class of jobs r, r=1,2...,R, there is a fixed number of jobs , N_r, in the network. The service-time distribution of class-r at each centre-i, i=1,2...,M, conforms to an arbitrary distribution with mean service rate, μ_{ir} and squared coefficient of variation, C^2_{sir}, r=1,...,R. Let $\{P_{irj}\}$ be the transition probability matrix describing the routing in the network (i.e., P_{irj} is the probability that a class-r job having just completed service at queue-i joins queue-j). Note that without loss

of generality, we asssume that jobs don't switch class membership as they move from one centre to another. The extension to class switching is straightforward. However, it is required in this case that classes belonging to an equivalence set of class EQ(r) [BRUE,80] have the same priority level (jobs are served in FCFS within each equivalent class) so that the marginal performance measures can be obtained easily after decomposing the equivalent class (see appendix E section E1 for more details about the notion of equivalence class).

As in the ME analysis of open network, let $\underline{\underline{n}}=(\underline{n}_1,\underline{n}_2,\ldots,\underline{n}_M)$ be the joint state of the network where $\underline{n}_i = (n_{i1},n_{i2},\ldots,n_{iR})$ is the marginal state for station-i, and n_{ir}, r=1,...,R is the number of jobs of class-r at the station-i, i=1,...,M, such that

$$\sum_{i=1}^{M} n_{ir} = N_r \ , \ r= 1,\ldots,R.$$

where N_r is a fixed constant belonging to the set of integers.

Let $P(\underline{\underline{n}})$ be the joint probability that the network is at state $\underline{\underline{n}}$ and $P_i(\underline{n}_i)$ is the marginal probability that the station-i is in state $\underline{n}_i$.

Moreover, let $S[\underline{N}_R,M] = \left\{ \underline{\underline{n}} \ / \sum_{i=1}^{M} n_{ir} = N_r, \ r= 1,\ldots R \right\}$

where $\underline{N}_R = (N_1,N_2,\ldots,N_R)$ is a constant population vector.

Given the normalisation, utilisation, $\{\rho_{ir}\}$, mean queue length, $\{<n_{ir}>\}$ and where appropriate the idle state probability, $\{P_{ir}(0)\}$, constraints (c.f. (6.3)-(6.5) and (6.12) with $\underline{\underline{\infty}}$ replaced by $\underline{N}_R$), it follows by analogy to the ME solution of open network, that for every $\underline{\underline{n}}\epsilon[\underline{N}_R,M]$ the joint state probabilities are given by the following

product-form expression:

$$P(\underline{\underline{n}}) = \frac{1}{Z[\underline{N}_R]} \prod_{i=1}^{M} f_i(\underline{n}_i) \qquad (7.1)$$

where $f_i(\underline{n}_i)$ are the unnormalised ME probabilities of an isolated G/G/1 queue-i, under priority (PR, HOL) or non-priority (FCFS, LCFS, LCFS-NONPR, PS) based disciplines, namely

$$f_i(\underline{n}_i) = \begin{cases} \dfrac{(n_i-1)!}{A_i} \prod_{r=1}^{R} x_{ir}^{n_{ir}} \left[\sum_{r=1}^{R} g_{ir}^{n_{ir}} \right], & \text{(FCFS,PS,LCFS LCFS-NONPR)} \\ g_{ir} y_{ir} x_{ir}^{n_{ir}} \prod_{s=r+1}^{R} x_{is}^{n_{is}} y_{is}^{V_{is}(\underline{n}_i)} \; (n_{i_1}=\ldots=n_{ir-1}=0, n_{ir}>0) & \text{(PR)} \\ \prod_{s=1}^{R} x_{is}^{n_{is}} y_{is}^{V_{is}(\underline{n})} \left[\sum_{s=1 \wedge n_{is}\neq 0}^{R} g_{is} \right], \quad \underline{n}_i \neq \underline{0}, & \text{(HOL)} \\ 1, \quad \underline{n}_i = \underline{0}, & \text{(PR,HOL,FCFS,PS,LCFS,LCFS-NONPR)} \end{cases} \qquad (7.2)$$

where $n_i = \sum_{r=1}^{R} n_{ir}$ and $A_i = \prod_{r=1}^{R} n_{ir}!$

where g_{ir}, x_{ir}, y_{ir}, i=1,2,...,M, r=1,2,...,M are the Lagrangian coefficients corresponding to the utilisation, ρ_{ir}, mean queue length, $\langle n_{ir} \rangle$, and idle state probability, $P_{ir}(0)$, constraints, respectively and are given by equations (6.10)-(6.11) for ME1, (6.14)-(6.16) for ME2. $Z[\underline{N}_R]$ is the normalising constant and given by

$$Z[\underline{N}_R] = \sum_{\underline{n} \epsilon S[\underline{N}_R, M]} \prod_{i=1}^{M} f_i(\underline{n}_i) \qquad (7.3)$$

$Z[\underline{N}_R]$ and $f_i(\underline{n}_i)$ are analogous quantities to that used in the standard convolution algorithm [BRUE,80-pp.58] and are obtained through the GE efficient recursive formulae as we will see it in section 7.3.

7.2 Computational techniques

The ME approximation (7.1) cannot be implemented directly since performance measures ρ_{ir}, $\langle n_{ir} \rangle$ and $P_{ir}(0)$ are not known a priori, and subsequently, no closed-form expressions are available for the corresponding Lagrangian multipliers. However, approximate estimates for these coefficient can be obtained by making use of a 'pseudo'[1] open multiple class network, Q^*, with no external arrival or departure processes and almost identical topology (configuration) to that of the original closed network (i.e., both networks have the same number of queues and servers, service-time characteristics and transition probabilities) (c.f. section 3.3.1), satisfying the principles of:

i/ the conservation of flow, expressed by the job flow-balance equations, i.e.,

$$\lambda^*_{ir} = \sum_{j=1}^{M} \lambda^*_{jr} P_{irj} \ , \ i=1,2\ldots,M, \ r=1,2,\ldots,R \qquad (7.4)$$

1: Pseudo open network requires that $\sum_{i=1}^{M} \langle n^*_{ir} \rangle = N_r$, $r= 1,2,\ldots,R$,

whereas for closed network we must have $\sum_{i=1}^{M} n_{ir} = N_r$

where λ^*_{ir} is the throughput of class-r at station-i in the pseudo open network which is proportional to the 'true' throughput of the original closed network, λ_{ir} (i.e., $\lambda_{ir} = \varphi_r \lambda^*_{ir}$, r=1,...,R, $i \epsilon [1,M]$).

ii/ The conservation of population, represented by the fixed population mean constraints:

$$\sum_{i=1}^{M} \langle n^*_{ir} \rangle = N_r, \quad r=1,\ldots,R, \tag{7.5}$$

The performance measures ρ^*_{ir}, $\langle n^*_{ir} \rangle$ and $P^*_{ir}(0)$ of the pseudo open network, Q^*, are determined by assuming that the interarrival, interdeparture of class-r, r=1, ...,R, to and from centre-i, i=1,...,M are renewal processes and conform to GE distribution. These quantities can be determined by algorithm 6.1 with the additional condition given by the fixed population mean constraint (7.5).

Moreover, the job flow-balance linear equations (7.4) are solved within a multiplicative constant per class, φ_r. These constants are determined by applying Newton/Raphson technique to solve the R non-linear equations with R unknowns given by equation (7.5). each set of new values of constant $\{\varphi_r\}$ is validated such that

$$\max_i \left\{ \sum_{r=1}^{R} \varphi_r \lambda^*_{ir} / \mu_{ir} \right\} < 1$$

(otherwise the pseudo open network cannot reach equilibrium).

The solution of equations (7.5) involves the computation of the corresponding Jacobian matrix which requires a good guess of the initial values of $\{\varphi_r\}$. This can be achieved heuristically by estimating the relative loading imposed by class-r on the station-i.

Clearly this parameter depends on the service characteristics, the population vector size, $\underline{N}_R$, and the type of the scheduling discipline adopted.

For queueing networks containing only FCFS servers, Almond [ALMO,88] suggested that the throughput of class-r in the pseudo open network should be normalised with a weighting proportional to the population size of class-r, N_r, e.g.,

$$\lambda^*_{ir} \longleftarrow \frac{N_r \lambda^*_{ir}}{\lambda^*_r} \quad , \; i=1,\ldots,M, \; r=,1,\ldots,R$$

$$\text{where } \lambda^*_r = \sum_{i=1}^{M} \lambda^*_{ir}$$

However, the above initialisation does not generally give convergence in priority situations. This is due to the fact that in priority queueing networks the relative load imposed by class-r on priority station-i, depends not only on the population of class-r but also on those of lower priority classes (since low-priority jobs are kept in the queue at priority centre by high-priority jobs). To overcome this drawback, the following initialisation have been adopted in priority cases:

i/ normalise the throughput

$$\lambda^*_{ir} \longleftarrow \frac{\lambda^*_{ir}}{\sum_{j=1}^{M} \lambda^*_{jr}/\mu_{jr}} \quad , \; i=1,\ldots,M, \; r=1,\ldots,R \quad (7.6a)$$

ii/ estimate the relative load imposed to centre-i by class-r jobs,

$$\lambda^*_{ir} \longleftarrow \lambda^*_{ir} \prod_{s=r}^{R} N_s \ , \ i=1,\ldots,M, \ r=1,\ldots,R \qquad (7.6b)$$

Consequently, $\{\varphi_r\}$ are initialised as follows:

$$\varphi_r = \frac{1 - \varepsilon}{\rho^*_b} \qquad (7.7)$$

where ε is a small value, say 0.01 and ρ^*_b is the overall utilisation of the bottleneck device (i.e.,

$$\rho^*_b = \max_i \left\{ \sum_{r=1}^{R} \lambda^*_{ir}/\mu_{ir} \right\}).$$

Note that whatever the initial values chosen for $\{\lambda^*_{ir}\}$, they must satisfy the job flow balance equations given by (7.4).

Having solved the pseudo open network, the Lagrangian coefficients can then be determined and subsequently the ME solution probabilities (7.1) can be used to approximate the joint steady state of the network. However, unless the network is separable [BASK,75], the throughputs, λ_{ir}, obtained from the truncated ME solution of the pseudo open network do not satisfy the job flow balance equations, (i.e., in general

$$\lambda_{ir} \neq \sum_{j=1}^{M} \lambda_{jr} P_{irj} \ , \ i=1,\ldots,M, \ r=1,\ldots,R$$

To overcome this problem, the ME probabilities are corrected by introducing a new multiplier, g_{0ir}, to adjust the Lagrangian coefficient relative to the utilisation constraints, g_{ir}, i=1,...,M, r=1,...,R. The correcting factors $\{g_{0ir}\}$, are initially set to 1 and are evaluated iteratively until the relation below is satisfied;

$$\frac{\rho^*_{ir}}{\rho_{ir}} = \text{constant for } r=1,\ldots,R \text{ and } \forall\, i\epsilon[1,M].$$

Amongst the heuristic formuale proposed:

- [KOUV,86a]

$$g_{0ir} \longleftarrow \frac{\rho^*_{ir}N_r}{\rho_{ir}\sum_{j=1}^{M}\frac{<n_{jr}>\rho^*_{jr}}{\rho_{jr}}}\, g_{0ir} \qquad (7.8a)$$

- [WALS,84]

$$g_{0ir} \longleftarrow \left[\frac{1-\rho_{ir}}{\rho_{ir}}\right] \frac{\rho^*_{ir}N_r}{\sum_{j=1}^{M}\frac{<n_{jr}>\rho^*_{jr}}{\rho_{jr}} - \rho^*_{ir}N_r}\, g_{0ir} \qquad (7.8b)$$

Both expressions have been thoroughly tested and generally give convergence [ALMO,88; KOUV,86a; WALS,84]. However, the speed of the convergence of each approach varies with the configuration of the network and as a consequence, it is very difficult to know before hand which one is the best. For example, the following expression has been found to be relatively good for certain types of networks (network with high connectivity):

$$g_{0ir} \longleftarrow g_{0ir}\frac{\sum_{j=1}^{M}\rho_{jr}\mu_{jr}P_{jri}}{\mu_{ir}} \qquad (7.8c)$$

The idea behind the above proposition is that the utilisation ρ_{ir}, that we want to determine must satisfy the job flow-balance equations, namely,

$$\rho_{ir}\mu_{ir} = \sum_{j=1}^{M} \rho_{jr}\mu_{jr}P_{jri} \quad , \quad i=1,\ldots,M, \quad r=1,\ldots,R.$$

The choice of the job flow-balance criteria is still an open problem, further suggestions are welcomed and can be easily tested.

7.3 Convolution formulae

The normalisation constant and the performance measures, $\langle n_{ir} \rangle$, ρ_{ir}, λ_{ir}, $i=1,\ldots,M$, $r=1,\ldots,R$ are evaluated by convolution type formulae analogous to the one used in the standard convolution algorithm [BRUE,80].

7.3.1. Computation of the normalising constant

In order to compute the normalising constante $Z[\underline{N}_R]$, we introduce the following definitions:

$\underline{n}_R = (n_1, n_2, \ldots, n_R)$,

$\underline{k}_R = (k_1, k_2, \ldots, k_R)$,

and the auxiliary function $z[\underline{n}_R, m] = \sum_{\underline{n} \in S[\underline{n}_R, m]} \prod_{i=1}^{m} f_i(\underline{n}_i)$

Note that $z[\underline{N}_R, M] = Z[\underline{N}_R]$ (c.f. (eq.7.3)).

The auxiliary function $z[\underline{n}_R, m]$ can also be written in the following form:

$$z[\underline{n}_R,m]=\sum_{\underline{k}_R=\underline{0}}^{\underline{n}_R}\left\{\sum_{\substack{\underline{n}\in S[\underline{n}_R,m]\\ \wedge\ \underline{n}_m=\underline{k}_R}}\prod_{i=1}^{m}f_i(\underline{n}_i)\right\}$$

$$=\sum_{\underline{k}_R=\underline{0}}^{\underline{n}_R}f_m(\underline{k}_R)\left\{\sum_{\underline{n}\in S[\underline{n}_R-\underline{k}_R,m-1]}\prod_{i=1}^{m-1}f_i(n_i)\right\}$$

which leads to $$z[\underline{n}_R,m]=\sum_{\underline{k}_R=\underline{0}}^{\underline{n}_R}f_m(\underline{k}_R)z[\underline{n}_R-\underline{k}_R,m-1] \qquad (7.9)$$

with initial value $$z[\underline{n}_R,1]=f_1(\underline{n}_R) \qquad (7.10)$$

$f_m(\underline{k}_R)$ are given by equations (7.2) and their computation can be carried out recursively using the one-step-recursions property of the ME solution, namely,

- for PR,

$$f_m(\underline{k}R)=\begin{cases}1\ , & \underline{k}_R=\underline{0}\\ g_{mr}x_{mr}y_{mr}f_m(\underline{0})\ , & \underline{k}_R=\underline{1}_r\\ \dfrac{g_{mr}x_{mr}y_{mr}}{g_{ms}}f_m(\underline{k}_R-\underline{1}_r),\ \{k_1=\ldots=k_{r-1}=k_{r+1}=\ldots=k_{s-1}=0,\ k_r=1,\ k_s>0,\ r<s\} & (7.11)\\ x_{mr}f_m(\underline{k}_R-\underline{1}_r),\ \{k_1=k_2=\ldots=k_{s-1}=0,\ k_s>0,\ k_r>1,\ r>s\}\\ x_{mr}y_{mr}f_m(\underline{k}_R-\underline{1}_r),\ \{k_1=k_2=\ldots=k_{s-1}=0,k_s>0,k_r=1,r>s\}\end{cases}$$

- for HOL

$$f_m(\underline{k}_R)= \begin{cases} 1 \ , \ \underline{k}_R = \underline{0} \\ g_{mr}x_{mr}y_{mr}f_m(\underline{0}), \quad \underline{k}_R=\underline{1}_r, \quad r=1,\ldots,R \\ \dfrac{\sum_{\substack{s=1\wedge k_s\neq 0}}^{R} g_{ms}}{\sum_{\substack{s=1\wedge k_s\neq 0\wedge s\neq r}}^{R} g_{ms}} x_{mr}y_{mr}f_m(\underline{k}_R-\underline{1}_r), \ \underline{k}_R\neq\underline{1}_r \text{ and } k_r=1 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad r=1,\ldots,R \\ x_{mr}f_m(\underline{k}_R-\underline{1}_r), \ \underline{k}_R\neq\underline{1}_r \text{ and } k_r>1, \ r=1,\ldots,R \end{cases} \quad (7.12)$$

- for FCFS, LCFS, LCFS-NONPR, PS [ALMO,88]

$$f_m(\underline{k}_R)= \begin{cases} 1 \ , \ \underline{k}_R = \underline{0} \\ g_{mr}x_{mr}f_m(\underline{0}) \ , \ \underline{k}_R = \underline{1}r \quad , \ r=1,\ldots,R \\ \sum_{r=1}^{R} x_{mr}f_m(\underline{k}_R-\underline{1}_r) \qquad \underline{k}_R > \underline{1}_r \ r=1,\ldots,R \end{cases} \quad (7.13)$$

further refinements of expression (7.9) are made if the centre-m is under non-priority based discipline [ALMO,88]. Substituing $f_m(\underline{k}_R)$ in equation (7.9) by its expression (7.13) and after simple manipulations, yields

$$z[\underline{n}_R,m] = z[\underline{n}_R,m-1] + \sum_{r=1}^{R} x_{mr}\left\{z[\underline{n}_R-\underline{1}_r,m-1](g_{mr}-1)+z[\underline{n}_R-\underline{1}_r,m]\right\} \quad (7.14)$$

7.3.2. Computation of the performance measures

The marginal queue length distribution per centre-i, i=1,2,...,M, $P_i(\underline{n}_R,\underline{N}_R)$ are obtained by applying the law of total probability to the ME solution (7.1), namely,

$$P_i(\underline{n}_R,\underline{N}_R) = \sum_{\substack{\underline{n}\in S[\underline{N}_R,M] \\ \underline{n}_i=\underline{n}_R}} P(\underline{n}_1,\underline{n}_2,\ldots,\underline{n}_{i-1},\underline{n}_R,\underline{n}_{i+1},\ldots,\underline{n}_M)$$

$$= \frac{f_i(\underline{n}_R)}{Z[\underline{N}_R]} \sum_{\substack{\underline{n}\in S[\underline{N}_R,M] \\ \underline{n}_i=\underline{n}_R}} \prod_{j=1\wedge j\neq i}^{M} f_j(\underline{n}_j)$$

By defining the following auxiliary function:

$$z^i[\underline{n}_R,m] = \sum_{\substack{\underline{n}\in S[\underline{N}_R,m] \\ \underline{n}_i=\underline{N}_R-\underline{n}_R}} \prod_{j=1\wedge j\neq i}^{m} f_j(\underline{n}_j)$$

$z^i[\underline{n}_R,M]$ is also interpreted as the normalising constant for the network after the removal of centre-i containing $\underline{N}_R-\underline{n}_R$ jobs.

the marginal probabilities $P_i(\underline{n}_R,\underline{N}_R)$ can therefore be written as:

$$P_i(\underline{n}_R,\underline{N}_R) = \frac{f_i(\underline{n}_R)}{Z[\underline{N}_R]}\ z^i[\underline{N}_R-\underline{n}_R,M] \qquad (7.15)$$

where the auxiliary function $z^i[\underline{n}_R,M]$ can be evaluated recursively as follows:

Given that all the probabilities must adapt to one,

$$\sum_{\underline{n}_R=\underline{0}}^{\underline{N}_R} P_i(\underline{n}_R,\underline{N}_R) = 1$$

we have then $$\sum_{\underline{n}_R=\underline{0}}^{\underline{N}_R} \frac{f_i(\underline{n}_R)z^i[\underline{N}_R-\underline{n}_R,M]}{Z[\underline{N}_R]} = 1$$

which gives $$Z[\underline{N}_R] = \sum_{\underline{n}_R=\underline{0}}^{\underline{N}_R} f_i(\underline{n}_R)z^i[\underline{N}_R-\underline{n}_R,M]$$

$$= z^i[\underline{N}_R,M] + \sum_{\underline{n}_R=\underline{0}\wedge\underline{n}_R\neq\underline{0}}^{\underline{N}_R} f_i(\underline{n}_R)z^i[\underline{N}_R-\underline{n}_R,M]$$

where we may have

$$z^i[\underline{N}_R,M] = Z[\underline{N}_R] - \sum_{\underline{n}_R=\underline{0}\wedge\underline{n}_R\neq\underline{0}}^{\underline{N}_R} f_i(\underline{n}_R)z^i[\underline{N}_R-\underline{n}_R,M]$$

or equivalently for population vector $\underline{n}_R$, we have

$$z^i[\underline{n}_R,M] = z[\underline{n}_R,M] - \sum_{\underline{k}_R=\underline{0}\wedge\underline{k}_R\neq\underline{0}}^{\underline{n}_R} f_i(\underline{k}_R)z^i[\underline{n}_R-\underline{k}_R,M] \quad (7.16)$$

Equation (7.16) can also be refined if centre-i is subject to non-priority disciplines and is given by

$$z^i[\underline{n}_R,M] = z[\underline{n}_R,M] - \sum_{r=1}^{R} x_{ir}\left\{z^i[\underline{n}_R-\underline{1}_r,M](g_{ir}-1)+z[\underline{n}_R-\underline{1}_r,M]\right\} \quad (7.17)$$

The mean queue length of class-r at centre-i is given by

$$\langle n_{ir}\rangle = \sum_{\underline{n}_i=\underline{1}_r}^{\underline{N}_R} P_i(\underline{n}_i,\underline{N}_R)n_{ir}$$

$$= \frac{1}{Z[\underline{N}_R]} \sum_{\underline{n}_i=\underline{1}_r}^{\underline{N}_R} f_i(\underline{n}_i)z^i[\underline{N}_R-\underline{n}_i,M]n_{ir} \quad (7.18)$$

if centre-i is under FCFS, LCFS, LCFS-NONPR, PS, the mean queue length is given by

$$\langle n_{ir} \rangle = \frac{1}{Z[\underline{N}_R]} \sum_{s=1}^{R} g_{is} x_{is} \sum_{\underline{n}_i - \underline{1}_s}^{\underline{N}_R} C_i(\underline{n}_i - \underline{1}_s) z^i[\underline{N}_R - \underline{n}_i, M] n_{ir} \quad (7.19)$$

where $C_i(\underline{n}) = \sum_{r=1}^{R} x_{ir} C_i(\underline{n} - \underline{1}_r)$ (7.19a)

and $C_i(\underline{0}) = 1$ (7.19b)

The evaluation of the utilisation ρ_{ir}, depends on the type of the service dsicipline adopted at centre-i. For example, if centre-i is under PR, we have

$$\rho_{ir} = \sum_{\substack{\underline{n}_i - \underline{1}_r \\ \wedge \\ n_{i_1} - n_{i_2} - \ldots - n_{ir-1} = 0}}^{\underline{N}_R} P_i(\underline{n}_i, \underline{N}_R)$$

using equation (7.15), we obtain

$$\rho_{ir} = \frac{1}{Z[\underline{N}_R]} \sum_{\substack{\underline{n}_i - \underline{1}_r \\ \wedge \\ n_{i_1} - n_{i_2} - \ldots - n_{ir-1} = 0}}^{\underline{N}_R} f_i(\underline{n}_i)\, z^i[\underline{N}_R - \underline{n}_i, M] \quad (7.21)$$

Similarly for HOL we obtain

$$\rho_{ir} = \frac{1}{Z[\underline{N}_R]} g_{ir} y_{ir} \sum_{\underline{n}_i - \underline{1}_r}^{\underline{N}_R} x_{ir}^{n_{ir}} \prod_{s=1 \wedge s \neq r}^{R} x_{is}^{n_{is}} y_{is}^{Vis(\underline{n})} z^i[\underline{N}_R - \underline{n}_i, M] \quad (7.21)$$

For FCFS,LCFS,LCFS-NONPR,PS disciplines, the utilisations are evaluated by the following formula [ALMO,88]:

$$\rho_{ir} = \frac{g_{ir}x_{ir}}{Z[\underline{N}_R]} \sum_{\underline{n}_i=\underline{1}_r}^{\underline{N}_R} C_i(\underline{n}_i-\underline{1}_r)z^i[\underline{N}_R-\underline{n}_i,M] \qquad (7.22)$$

where $C_i(\underline{n})$ is given by equations (7.19a-b)

Note in all cases, we have

$$\rho_i = 1 - \frac{z^i[\underline{N}_R,M]}{Z[\underline{N}_R]} \qquad (7.23)$$

The convolution forms of the ME solution has computational time cost of

$$O\left\{ [M'N + 2R(M-M')] \prod_{r=1}^{R} N_r \right\}$$

where M' is the number of priority (PR or HOL) centres and

$$N = \sum_{r=1}^{R} N_r \ .$$

Note that the MVA approximations require $O\left[RM \prod_{r=1}^{R} (N_r+1)\right]$ operations [LAZO,84,pp.140] and therefore is slightly faster but, less accurate than the UME (see numerical results). Furthermore, it is applicable only to Markovian networks (c.f. section 2.3.4).

7.4 UME algorithm for general closed queueing networks with priorities

A solution of general closed queueing network with M centres under either priority (PR and HOL) or non-priority (FCFS, LCFS, LCFS-NONPR and PS) disciplines and R classes of jobs can be approximated by the following algorithm:

Algorithm 7.1

INPUT

- M,R,
- $\underline{N}_R = (N_1, N_2, \ldots, N_R)$, population size vector ,
- for each centre-i
  - type of service discipline,
  - $\underline{\mu}_i = (\mu_{i_1}, \mu_{i_2}, \ldots, \mu_{iR})$, mean service rates
  - $\underline{C^2_{si}} = (C^2_{si_1}, C^2_{si_2}, \ldots, C^2_{siR})$, squared coefficients of variation of the service times,
- $\{P_{irj}\}$, transition probability matrix.

PART A {* Solve the pseudo open network Q^* *}

STEP A.1 {* Solve the job flow-balance equations (7.4) to a multiplicative constant (e.g.,set $\lambda^*_{ir}=1$, r=1,..,R) *}

STEP A.2 {* Estimate the load imposed to centre-i by class-r jobs (e.g., use equations (7.6a-b)) *}

STEP A.3 {* Compute the relative utilisation *}

$$\rho^*_{ir} \longleftarrow \lambda^*_{ir}/\mu_{ir} \ , \ i=1,\ldots,M, \ r=1,\ldots,R$$

STEP A.4 {* obtain the network bottleneck *}

$$\rho_b^* \longleftarrow \max_i \{ \rho_i^* \}$$

STEP A.5 {* Initialisation of a multiplicative constants, φ_r *}

$$\varphi_r \longleftarrow (1-0.01)/\rho_b^* \ , \ r= 1,\ldots,R$$

STEP A.6 {* Solve the pseudo open network as an open network with $\varphi_r\lambda_{ir}^*$ being the arrival rate of class-r at centre-i *}

- {* Iterate step A.6 until convergence of C_{air}^2, i=1,...,M, r=1,...,R *}

STEP A.6.1 {* Obtain the performance measures $\langle n_{ir}\rangle^*, \rho_{ir}^*$, $P_{ir}^*(0)$ using algorithm 6.1 *}

STEP A.6.2 {* Apply Newton/Raphson method to solve the non-linear system of equations (7.5) *}

STEP A.7 {* Evaluate the Lagrangian coefficients, g_{ir}, x_{ir} and y_{ir} as appropriate (c.f. (6.10)-(6.11) for ME1 and (6.14)-(6.16) for ME2) *}

PART B {* Solution of the closed network *}

STEP B.1 {* Initilisation of the job flow-balance correction factors, g_{0ir} *}

$$g_{0ir} \longleftarrow 1 \ , \ i=1,\ldots,M, \ r=1,\ldots,R$$

STEP B.2 {* Iterate until ρ_{ir}^*/ρ_{ir} = constant $\forall i \epsilon [1,M]$ and r =1,2,...,R *}

STEP B.2.1 {* Use convolution formulae (7.9)-(7.23) to obtain the performance metrics, $\langle n_{ir}\rangle$, ρ_{ir} *}

STEP B.2.2 {* Apply job flow-balance correction

(c.f. (7.8a-c)) *}

STEP B.3 {* Obtain the throughputs *}

$$\lambda_{ir} \longleftarrow \mu_{ir}\rho_{ir}$$

END.

7.5 Numerical results and discussions

(See appendix F, section F1)

The performance metrics of the existing and proposed algorithms are compared with QNAP-2 exact or simulation (with 95% confidence intervals) results. The first test-bed network to be investigated is the one analysed by Bondi and Chuang [BOND,88]. It consists of the two-stage cyclic Markovian queueing network (c.f. Fig.7.1) with PR and FCFS centres serving two classes of jobs. The models are grouped into 3 types of networks depending on the utilisation of the PR centre. Each category of networks has a fixed population for high priority class jobs (N_1 = constant) and variable for the low-priority ones (N_2 =1,...,6),(c.f. Table 7.1).

- Type-1 model consists of networks with relatively small utilisation for high-priority class jobs at PR centre ranging from 0.3042 to 0.327 and $\mu_{11} >> \mu_{12}$.

- Type-2 model consists of networks with moderate utilisation for the high-priority class varying from 0.4383 to 0.6167 with $\mu_{11}=\mu_{12}$.

- Type-3 model correponds to networks with relatively high utilisation of the PR centre mostly attributed to high-priority class jobs ($0.5748 < \rho_{11} < 0.8362$) with $\mu_{11} << \mu_{12}$.

The comparative study focusses on the system throughputs (e.g., Tables 7.2a-7.2c) and the mean queue lengths at FCFS centre (e.g. Tables 7.3a-7.3c) of the existing methods ROA [SEVC,77a], m-ROA [KAUF,84], MVA [BRYA,84], the modified MVA (m-MVA) approximation presented by Bondi and Chuang [BOND,88] and the proposed UME2 with the corresponding percentage differences from the exact results given between brackets. It is observed that for low utilisation and $\mu_{11} >> \mu_{12}$, all methods perform well (e.g., Tables 7.2a and 7.3a). However, as the utilisation of the PR centre increases (e.g., Tables 7.2b and 7.3.c), the delay error, the synchronisation error, the failure to account for the effect of preemption on the interarrival-time variability at the nonpreemptive centre and the failure to predict the effective service time accurately (c.f. section 2.3.4) becomes more influential mainly in ROA and m-ROA approches. The MVA which does not use the concept of ficticious servers does not suffer from the delay error and the failure to predict the effective service time accurately and therefore is generally more accurate than the ROA based methods. The m-MVA which captures some of the variablity of the interarrival-time process and reduces the synchronisation error by modifying the mean response time formula in the standard priority MVA algorithm of Bryant-et-al [BRYA,84], is more accurate than the other presently developed algorithm. However, the accuracy of the m-MVA deteriorates as the population of class-2 jobs increases (e.g., Table 7.2b,c). This attributed to the fact that the m-MVA uses the arrival instant theorem which is assumed to hold in priority QNM's. The UME2 being a proper probabilistic approach does not suffer from the synchronisation error and captures in better way the interarrival-time variability. As a consequnce, the UME2 is more accurate than the existing developed approximations. However, the ME

methods still suffer from the fact that the interarrival-time processes are assumed to be renewal and GE distributed.

In tables 7.5a and 7.6c, the test-bed network of table 7.1 with the PR centre replaced by HOL one, is analysed (e.g., Table 7.4). In this case, only the MVA and UME2 can be compared to the exact results. Tables 7.5a-c exhibit the utilisation of the HOL centre and the corresponding standard deviations (since values of utilisations may be very small) from the exact results are given between brackets. Note that the UME2 is still superior in comparison to the MVA mainly for high utilisations of the HOL centre (e.g., Tables 7.5c, 7.6c). Moreover, the two-stage cyclic network with PR and HOL queues and exponential service times, is analysed in table 7.8 with the correponding raw data given in table 7.7. In particular, Table 7.8 displays the marginal mean queue lengths obtained by the MVA and UME2 with the correponding relative errors from the exact results given between brackets. In this case, the UME2 is still more accurate in comparison to MVA with relative error less than 16%. In table 7.10, the two-stage cyclic network (c.f., Fig.7.1) with general (GE-type) service-time distributions is studied with the corresponding characteristics given in table 7.9. Table 7.10 displays the utilisations of the PR centre and the mean queue lengths at FCFS centre for both exact and UME2. The exact GE results are produced by adopting an efficient numerical technique involving the inversion of the state transition matrix [ALMO,88]. Therefore, only networks with relatively small populations (mainly for 3 classes) are considered in the comparative study. It is observed that the UME2 consistently gives reasonnable predictions of the statistics. The UME2 approximation slightly deteriorates for very high variability of the service-time distribution (e.g., experiments 3 and 10). Finally, the central server model with 3 centres and two classes (c.f., Fig.7.2)

is analysed in table 7.12 (raw data given in table 7.11). The service-time distributions conform to GE with mean and coefficient of variation given in table 7.11. Centre-1 is taken to be under PR service discipline, centre-2 under HOL and centre-3 under FCFS. Table 7.12 presents the utilisation of low and high priority classes at PR centre and mean queue lengths of both classes at HOL station obtained by simulation (SIM) (since the networks are too large to be solved exactly) and UME2. It is observed in this case that UME2 still provides a fairly good approximation for the statistics although in experiment 7, the relative error exceeds the 23% because of the very small performance measure value.

7.6 Conclusion

In this chapter the PME for general QNM's is extended to the case of general closed networks with priority (PR or HOL) or non-priority (FCFS, LCFS, LCFS-NONPR, PS) based disciplines. The ME solution is approximated by a truncated ME decomposition solution of a corresponding pseudo open network into individual multiple class G/G/1 queues, which satisfies the principles of population and flow conservation. This solution is implemented by making use of the GE distribution to model the interarrival, service and interdeparture processes (under renewal underlying assumptions). The proposed UME algorthim does not suffer from delay and synchronisation errors or from inconsistent null process behaviour [ZAHO,87] and certainly captures most of the variability of the interarrival and interdeparture processes. The algorithm predicts performance measures for all classes of jobs more accurately than previously developed approximate methods for Markovian networks. Moreover, for general closed networks the UME algorithm is very comparable to the exact or

simulation analysis using the GE distributional model. Improvements of the UME algorithm are clearly important especially to reduce the computational cost of the evaluation of the normalising constant.

CHAPTER 8

CONCLUSION

8.1 Thesis suumary

This research work was motivated by the need of analytic tools for the performance investigation of queueing networks with priorities. These important characteristics are present in most modern computer systems and communication networks but they cannot be represented directly in product-form QNM's [BASK,77]. As a consequence only heuristics are used to incorporate the priority feature in present fast computational algorithms (convolution, MVA). In this thesis, the principle of maximum entropy (PME) has been applied to provide a new analytic framework for the approximate analysis of general open and closed queueing networks involving a mixture of priority (PR, HOL) and non-priority (FCFS, LCFS with or without preemptions, and processor sharing (PS))scheduling disciplines. The PME, subject to closed-form expressions of marginal mean value constraints, provides a more formal rather than a heuristic justification for the decomposition of the network into individual queues. To this end, the ME solution of a single priority G/G/1 queue under either PR or HOL service discipline constitutes a building block for the approximate analysis of open and closed priority queueing networks. The analytic use of the ME solutions is expedicted by carrying out exact analysis based on the bulk interpretation of the generalised exponential (GE) distribution, in order to derive new closed-form expressions for the marginal mean queue length and idle state probability constraints.

In brief the main summary of the thesis is as follows:

i/ In the second chapter, a review of useful results and properties of single queues and networks involving priority disciplines are summarized. In particular, existing techniques which are used for the approximate solution of open and closed priority queueing networks are presented and discussed in detail.

ii/ In the third chapter, the PME is introduced together with the GE distribution. The principle provides a uniquely correct, self-consistent method of inference for estimating a probability distribution given information in the form of expected values. As a result, the GE distribution has been found to be the best supported non-discrete distribution when the first two moments are known. The chapter also presents several properties of the GE which provide at a same time a mathematical tractability comparable to that of an exponential distribution (i.e., pseudo memoryless, bulk interpretation) and variability (C_a^2, $C_s^2 \neq 1$) which establish the GE as a phase-type distribution (i.e., H_2) where one of the two phases has zero interevent-time.

iii) In the fourth chapter, exact analysis on GE/G/1 PR and HOL queue is carried out. New analytic expressions for the corresponding characteristics of these queues are derived. These results which constitute a generalisation of present ones known for a M/G/1 priority queue [JAIS,65] are obtained by making use of the bulk interpretation of GE as a compound Poisson process with geometric distributed bulk sizes and the limiting interpretation of GE as a phase-type distribuion with zero interevent at one of the two phases. Moreover, the M/G/1 conservation law [KLEI,65] is extended to GE/G/1 queues under any work-conserving and non-preemptive discipline. In addition, more useful GE type performance bounds are established for the marginal mean queue lengths.

iv) In the fifth chapter, entropy maximisation, subject to two different sets of prior information, drawn from the normalisation, utilisation, mean queue length and idle state probability constraints (cases 1 and 2), is applied to characterise product-form approximations for the joint queue length distribution of both PR and HOL stable G/G/1 queue with R priority classes. New 'one-step' recursions are established and two closed-form approximations (ME1 and ME2) for the marginal state probabilities per priority class are derived. Moreover, these results are used in the context of the shadow CPU methods as a basis to provide new approximations for the mean and squared coefficient variation of the effective priority service-time distribution per class. As a result, universal formulae for the parameters of the departure process per class are presented. It is interesting to note, that for the first time approximate closed-form expressions for the joint and marginal queue length distributions of a G/G/1 PR or HOL queue are proposed.

v) In the sixth chapter, the PME is used to characterise a new product-form solution for the approximate analysis of arbitrary multiple class open networks of queues with infinite capacities, single servers and mixed service disciplines involving priority (PR and HOL) or non-priority (FCFS, LCFS, LCFS-NON PR, PS) based stations. The new UME algorithm proposed, offers via the GE-type implementation the only analytic tool available in the literature for calculating approximately marginal queue length distributions per class and also analysing bulk-type open queueing networks. Moreover, the GE-type priority approximations are used in conjunction with the shadow CPU technique to extend present methods (based on class compositions and disaggregations for FCFS networks) to analyse queueing networks with priorities.

vi) In the seventh chapter, the GE-type results of open networks are used to analyse a corresponding closed network with similar configuration. In particular, the ME algorithm established previously for the approximate solution of general closed FCFS queueing networks [ALMO,88] is extended to include in addition, PR, HOL, LCFS with or without preemptions and PS disciplines.

The main contributions of the thesis are a) the derivation of new exact closed-form expressions for the GE/G/1 PR or HOL queue. b) The establishment of new closed-form approximations for the joint and marginal queue length distribution of priority PR or HOL queues. c) The development of new ME algorithm for the approximate solution of general open networks with priorities. d) The extension of the UME algorithm (for multiple class closed network with FCFS centres [ALMO,88]) to include in addition, priority (PR and HOL) and non-priority (LCFS with or without preemptions and PS) based disciplines.

8.2 Suggestions for future work

Several issues have been mentioned in this thesis that need further studies. First, proofs of convergence of algorithm 6.1, 6.2 and 7.1 are required to give more credibility to the ME methods. The convergence proofs for iterative algorithms discussed by Agrawal [AGRA,85] may be used as a guide line. Second, the choice of the job flow-balance criteria for the analysis of closed queueing networks is still an important issue to examine in order to minimize the time complexity of the ME approach. Third, the alternative algorithm for the ME approximation of general closed network with a single class of jobs (c.f. [TOMA,89]) where the Lagrangian coefficients are estimated by solving the corresponding open network with external source being

the bottleneck device of the original closed network is worth to extend to multiple classeswith mixed service disciplines.

Throughout our analysis it was assumed that the priority queues are single server stations. The extension of the ME analysis to priority queues with multiple servers is an important step towards the construction of more general and realistic analytic models for modern computer systems with multiple processors. The methodology used in chapter 5 together with the results of Kouvatsos and Almond [KOUV,86c] could act as a basis to the analytic establishment of the ME solution of G/G/c PR or HOL queue. Moreover, the approximate mean queue length of M/G/c PR queue given by Bondi and Buzen [BOND,84] may be generalised to GE/G/c PR queue by using similar arguments to the ones adopted in chapter 4.

Another immediate extension is the investigation of networks with priorities and finite capacity buffers. The work proposed in this thesis in combination with the research presented in [XENI,89] forms a step in this direction.

REFERENCES

[AGRA,85] Agrawal, S.C.,' *Metamodelling, A study of Approximation in Queueing Models*', MIT Press, Herb Schwetman (eds), (1985).

[ALLE,78] Allen, A.O., '*Probability Statistics and Queueing Theory with Computer Science Application*', Academic Press, New-york, (1978).

[ALMO,88] Almond, J.'*Generalised Analytic Queueing Network Models*', Phd thesis, Bradford University, (1988).

[AVI-,61] Avi-itzhak, B . and Naor, P., '*On a Problem of Preemptive priority queueing*', Op. Res., Vol. 9, pp.664-672, (1961)

[AVI-,73] Avi-itzhak, B ,'*Approximate Queueing Models for Multi-programming Computer Systems*', Oper. Res., V.21, No.16, pp.1212-1230,(1973)

[BARD,79] Bard, Y.,'*Some extensions to multipleclass Queueing Network Analysis*', in Performance of Computer systems, eds. Arato,M and Butrimenko, M., Gelenbe, E., North-Holland,pp.51-61, (1979).

[BARD,80a] Bard, Y. '*A Model of Shared DASD and Mulipathing*', Comm. ACM, V. 23, pp.564-572, (1980)

[BARD,80b] Bard, Y.'*Estimation of State Probabilities Using the Maximum Entropy Principle*', IBM j. res. rev. , V. 24, pp. 563-569, (1980).

[BASK,75] Baskett, F., Chandy, K.M., Muntz, R.R., Palacio, F.G., ' *Open, Closed and Mixed Networks of Queues with Different Classes of Customers*',J. ACM, Vol. 22, No. 2, pp.248-260 (1975).

[BOND,84] Bondi, B. and Buzen, J.P.,'*The Response Time of Priority Classes under Preemptive Resume in M/G/m queues*', J. ACM, pp.195-201, (1984).

[BOND,88] Bondi, B. and Chuang, Y.M., '*A New MVA-Based Approximation for Closed Queueing Networks with a Preemptive Priority server*', Perf. Eval., Vol. 8, pp.195-221, (1988).

[BRYA,84] Bryant, R.M., Krezinski, A.E., Lakshimi, M.S., Chandy, K.M, '*The MVA Priority Approximation*, J. ACM trans. on comp. syst., Vol. 2, No. 4, pp.335-359, (1984).

[BRUE,80] Bruel, S.C. and Balbo, G., '*Computational Algorithm for Closed Queueing Networks*', North Holland, (1980).

[BURK,56] Burk, P.J.'*Output of a Queueing system*', Oper. Res., Vol.4, pp.699-704,(1956).

[BUZE,73] Buzen, J.P.,'*Computational Algorithms for Closed Queueing Networks with Exponential Servers*', Communication of the ACM, Vol. 16, No. 9, pp.527-531, (1973).

[BUZE,83] Buzen, J.P. and Bondi, A.B., '*The Response Time of Priority Classes Under Preemptive Resume im M/M/m Queues*', Oper. Res., Vol. 31 (3), pp.456-465, (1983).

[CHAN 75] Chandy,K.M., Herzog, U., Woo,L., '*Approximate Analysis of General Queueing Networks.*', IBM J. RES. Develop., Vol. 19, pp.43-49, (1975).

[CHAU,83] Chaudhry, M.L., Templeton, J.G.C.,'*A First Course in Bulk Queues*', John Willey (eds), New york, (1983).

[CHOW,83] Chow, W.M., Yu, P.S., '*An Approximation Technique for Central Server Queueing Models With a Priority Dispatching rule*', Per.Eval., Vol. 3, North Holland, pp.55-62, (1983).

[COBH,54] Cobham, A., '*Priority Assignement in Waiting Line Problems*' Op.Res., Vol.2, pp.70-76, (1954).

[CONW,67] Conway, R.W., Maxwell, W.L., Miller, L.W., '*Theory of Scheduling*', Addisson Wesley, Reading Mass., (1967).

[COOP,81] Cooper, R.B., '*Introduction to Queueing Theory*', 2nd ed., North Holland, New york, (1981).

[COUR,77] Courtois, P.J.,'*Decomposability: Queueing and Computer Systems Applications.*', Academic Press, (1977).

[COX ,55] Cox, D.R.,'*The Analysis of Non-Markovian Stochastic Process by the Inclusion of Supplementary variables*', Proc. Camb. Phil. Soc., (Math. and Phys. Sci.), Vol. 51, pp.433-441, (1955).

[COX ,62] Cox, D.R., '*Renewal Theory*', Methuen, London, (1962).

[DENN,78] Denning, P.J., Buzen, J.P., '*The Operational Analysis of Queueing Network Models*', Computing Survey, Vol. 10, No. 3 PP.55-62, (1978).

[EAGE,88] Eager, D.L. and Lipscomb, J.N., '*The MVA Priority Approximation*', Perf. Eval., Vol.8, pp.173-193, (1988).

[EL-AF,83] El-Affendi,M.A. and Kouvatsos,D.D.,'*A Maximum Entropy Analysis of The M/G/1 and G/M/1 Queueing Systems at Equilibrium*', ACTA Info., Vol.19, pp.339-355, (1983).

[FELL,68] Feller,w., '*An Introduction to Probability Theory and its Applications.*', John Willey, 3rd edition, Vol. 3, New york, (1968).

[FERD,70] Ferdinand, A.E., '*A Statiscal Mechanical Approach to Systems Analysis*', IBM J. Res. Develop., Vol. 14, PP.539-547, (1970).

[GELE,76] Gelenbe, E. and Pujolle, G., '*The Behaviour of a Single Queue in a General Queueing Network*', Acta Informatica, Vol. 7, pp.123-160, (1976).

[GELE,80] Gelenbe, E. and Mitrani, I.,'*Analysis and Synthesis of Computer Systems*', Academic Press, (1980).

[GEOR,89] Georgatsos, P.H., '*Modelling and Analysis of Computer Communication Networks with Random or Semidynamic Routin*', Forthcoming Ph.d Thesis, Univ. of Bradford, (1989).

[GORD,67] Gordon, W.J., and Newell, G.F., '*Closed Queueing Systems with Exponential Servers*', Oper. Res., Vol. 15, pp.244-265, (1967).

[GROS,85] Gross, D., Harris, C.M., '*Fundamentals of Queueing Theory*', John Willey (2ndeds.), New york, (1985).

[HEAT,59] Heathcote, C.R., '*The Time Dependent Problem for a Queue with Preemptive Priorities*', Op. Res., Vol. 7, pp.670-680, (1959).

[HEAT,60] Heathcote, C.R., '*A Single Queue with Several Preemptive Priority Classes*', Op. Res., Vol. 8, pp.630-638, (1960).

[JACK,57] Jackson, J.R., '*Networks of Waiting Lines*', Oper.Res., Vol.5, pp.518-521, (1957).

[JAIS,62] Jaiswal, N.K., '*Time Dependent Solution of the Head-of-Line Priority Queue*', J. Roy. Stat. Soc., Series B, Vol. 24, pp.91-101, (1962)

[JAIS,68] Jaiswal, N.K., '*Priority Queues*', Academic Press, New york, (1968).

[JAYN,57a] Jaynes, E.T., '*Information Theory and Statistical Mechanics*', Physical Review, Vol. 106, No.4, pp.620-630, (1957).

[JAYN,57b] Jaynes, E.T., '*Information Theory and Statiscal Mechanics II.*' Phys. Rev., Vol. 108, pp. 171-190, (1957).

[JAYN,68] Jaynes, E.T., '*Prior Probabilities*', IEEE Trans. Syst. Sci. Cybern., SSC-4, Vol. 4, pp.227-241, (1968).

[JAYN,79] Jaynes, E.T., '*Where Do We Stand on Maximum Entropy?*', In Maximum Entropy Formalism, R.D. Leven and M.Tribus, Eds., PP. 17-118, MIT Press, Cambridge, Mass, (1979).

[JOHN,79] Johnson, R.W, '*Determining probability distributions by maximum entropy and minimum cross-entropy*', Proc. APL79, ACM 0-8971-005, pp.24-29, (1979).

[KAUF,84] Kaufman, J.S., '*Approximate Method for Networks of Queues with Priorities*', Perf.Eval., Vol. 4, North Holland, pp.183-198, (1984).

[KEIL,62] Keilson, J., '*Queues Subject to service interruption*', Ann. Math. Stat., Vol. 33, pp.1314-1322, (1962).

[KLEI,64a] Kleinrock, L.,' *Analysis of a Time-Shared Processor*', Nav. Res. Log. Quart., Vol. 11,pp.59-73,(1964).

[KLEI,64b] Kleinrock, L.,'*Communication Nets; Stochastic Message Flow and Delay*', Mc Graw-Hill, New york, (1964).

[KLEI,65] Kleinrock, L.,'*A Conservation Law for a Wide Class of Queueing Disciplines*', Nav. Res. Log. Quart., Vol. 12, pp.181-192, (1965).

[KLEI,75] Kleinrock, L.,'*Queueing Systems, Vol. 1: Theory.*', John Willey, New york, (1975).

[KLEI,76] Kleinrock, L.,'*Queueing Systems, Vol. 2: Computer Applications.*', John Willey, New york, (1976)

[KOUV,83] Kouvatsos, D.D.'*Maximum Entropy Methods for General Queueing Networks*', Res. Rep. RCC34, Univ. of Bradford, Bradford, U.K., (1983).

[KOUV,85] Kouvatsos, D.D., '*Maximum Entropy methods for General Queueing Networks*', Modelling Techniques and Tools for Performance Analysis, D.Potier eds., North-Holland, pp.589-608, (1985).

[KOUV,86a] Kouvatsos, D.D., '*Maximum Entropy and the G/G/1/N Queue*', Acta Informatica, Vol.23, pp.545-565, (1986).

[KOUV,86b] Kouvatsos, D.D., '*A universal Maximum Entropy Algorithm for the Analysis of General Closed Networks*', in T.Hasegawa et al eds., IBM and INRIA, North-Holland, pp.113-124, (1986).

[KOUV,86c] Kouvatsos, D.D.,' *two-station cycilc queues with multiple servers of GE-type* ', Papers of Int. Workshop on Computer Perf. Eval. INRIA, Sophia Antiopolis, France, (1986).

[KOUV,87] Kouvatsos, D.D. and Tabet-Aouel, N., '*A Maximum Entropy Priority Approximation for Stable G/G/1 Queue*', Tech. Rep. #DDK/NT-a/1, U. of Bradford, (1987).
(to appear in Acta Informatica)

[KOUV,88a] Kouvatsos,D.D., '*A Maximum Entropy Analysis of the G/G/1 Queue at Equilibrium*', J.Op.Res.Soc., Vol.39, No.2, (1988). PP.183-200.

[KOUV,88b] Kouvatsos, D.D. and Tabet-Aouel, N., '*A Maximum Entropy Method for General Closed Queueing Network with Priority Servers*', Conf. Proc. of 4th UK Computer and Telecommunication Performance Engineering Workshop, Edinburgh, (1988).

[KOUV,88c] Kouvatsos, D.D., Georgatsos, P.H.E and Tabet-Aouel, N., '*A Universal Maximum Entropy Algorithm for General Multiple Class Open Networks with Mixed Service Disciplines*', in Conf.Proc. of 4th Int.Conf. on Modelling Techniques and Tools for Computer Perf. Evaluation, Palma de Mallorca (Spain), (1988).

[KOUV,88d] Kouvatsos, D.D. and Tabet Aouel, N., '*Maximum Entropy Analysis of General Queueing Networks with Priorities*', Paper on Int. Conf. on the Analysis and Control of Large Scale Stochastic Systems, Univ. of North Carolina, Chapel-Hill,USA, May 23-25, (1988)

[LAZO,84] Lazowska,E.D., Zahorjan, J., Graham, G.S. and Sevcik, K.C. '*Quantitative Systems Performance*', Prentice-Hall,(1984).

[LITT,61] Little, J.D.C., '*A Proof of the Queueing Formula L = W* ', OPER. RES., Vol 9, pp.383-387, (1961).

[MARK,73] Marks, B.I.,'*State Probabilities of M/M/1 Priority queues*', Oper. Res., Vol. 21, pp.974-987, (1973).

[MILL,60] Miller, R.G., '*Priority Queues*', Ann. Math. Stat., Vol. 31 pp.86-103, (1960).

[MILL,81] Miller, D.R., '*Computation of Steady-State Probabilities for M/M/1 Priority Queues*', Oper. Res., Vol. 29, pp.945-958 (1981).

[MITR,72] Mitrani, I.,'*A Queueing Model of Priority Multiprogramming*' Tech. Rep., Univ. of Newcastel upon tyne,(1972).

[MITR,81] Mitrani, I. and King, P.J.B.,' *Multiprocessor Systems with Preemptive Priorities*', Perf. Eval., Vol. 1, pp.118-125, (1981).

[MORR,81] Morris, R.J.T.'*Priority Queueing Networks*', Bell System Tech. Jour., Vol. 60, No. 8, pp.1745-1769, (1981).

[NAIN,84] Nain, P.,'*Interdeparture Times from a Queueing Systems with Preemptive Resume Prority*', Perf. Eval.,North-Holand, Vol. 4, pp.93-98, (1984).

[NEUS,82] Neuse, D. and Chandy, K.M., '*HAM: The Heuristic Aggregation Method for Solving General Closed Queueing Network Models of Computer Systems*', Per. Eval. Rev., Vol. 4, pp.195-212, (1982).

[OTHM,88] Othman, A.O.,'*Performance Analysis and Control of Computer Communication Network Models*', Ph.D. Thesis, Dept. of Computing, University of Bradford, U.K., (1988).

[REIS,74] Reiser, M. and Kobayashi, H.'*Accuracy of the Diffusion Approximation for some Queueing Systems*',IBM Jour. Res. of Dev., Vol. 18, pp.110-124, (1974)

[REIS,79] Reiser, M., '*A Queueing Network Analysis of Computer Communication Networks with Window Flow Control*', IEEE Trans. Comm., Vol. Com. 27, pp.1199-1209, (1979).

[REIS,80] Reiser, M., Lavenberg, S.S., '*Mean Value Analysis of Closed Multichain Queueing Networks Models*', J.ACM Vol.22, NO.2, pp.313-322, (1980).

[SAUE,75a] Sauer, C.H., '*Configuration of Computer Systems: An approach using queueing network models*', PhD Thesis, U. of Texas, (1975).

[SAUE,75b] Sauer, C.H., Chandy, K.M., '*Approximate Analysis of Central Server Models*', IBM J. Res. and Develop., Vol.19, No.3, pp.301-313, (1975).

[SCHE,67] Scher, A.L., '*An Analysis of Time-Shared Computer Systems*', MIT Research Monograph, No. 36, MIT Press, (1967).

[SCHM,83] Schmitt, W., '*Approximation Methods for Networks of Queues with Priority*', Proc. of 10th Intern. Tel. Cong., Montreal, (1983).

[SCHM,84] Schmitt, W., '*On Decompositions of Markovian Priority Queues and their Applications to the Analysis of Closed priority Queueing Networks*', in:Perf. 84: Proc. 10th Internat. Symp. Comp. Perf., Gelenbe, E., eds, (North-Holland, Amesterdam), pp.393-407, (1984).

[SCHW,79] Schweitzer, H.D., '*Approximate Analysis of Multiclass Closed Networks of Queues*', Int. Conf. Stochastic Control and Optimization, Amesterdam, Netherlands, (1979).

[SEVC,77a] Sevcik, K.C., '*Priority Scheduling Disciplines in Queueing Network Models of Computer Systems*', Proc. IFIP Congress 77, North Holland Publishing Co., Amesterdam, pp:565-570, (1977).

[SEVC,77b] Sevcik, K.C., Levy, A.I., Tripathi, S.K. and Zahorjan, J.L., '*Improving Approximation of Aggregated Queueing Network Subsytems*', Computer Performance, (North-Holland), in: Chandy, K.M. and Reiser, M. (Eds.), pp.1-22, (1977).

[SHAN,48] Shannon, C.E., '*A mathematical theory of communications*', Bell Syst.Tech.Jour., Vol.27, pp.379-423, pp.622-656, (1948).

[SHOR,78] Shore, J.E., '*Derivation of Equilibrium and Time-Dependent Solutions to M/M/1//N and M/M/1 Queueing Systems using Entropy Maximization*', AFIPS Conf.Proc, Vol.47, pp.483-487, (1978)

[SHOR,80] Shore, J.E.and Johnson, R.W., '*Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy*', IEEE Trans.Inf.Theory, vol. it-26, pp.25-37 (1980).

[SHOR,81] Shore,J.E. and Johnson, R.W., '*Properties of Cross-Entropy Minimization*', IEEE. Trans. on Information Theory, IT-27, pp.472-482, (1981).

[SHOR,82] Shore, J.E., '*Information Theoretic Approximations for M/G/1 and G/G/1 queueing systems*', Acta Informatica, Vol.17, pp.43-61, (1982).

[SIMO,61] Simo, H.A. and Ando, A., '*Aggregation of Variables in Dynamic Systems*', Econometrica, Vol. 29, pp.111-138, (1961).

[TOMA,89] Tomaras, P., '*Decomposition of General Queueing Network Models* ', Ph.d Thesis, Univ. of Bradford, (1989).

[TRIV,82] Trivedi, K.,'*Probability and Statistics with Reliability Queueing and Computer Application*', Prentice-Hall, (1982).

[TRIB,69] Tribus, M., '*Rational Descriptions, Decisions and designs*', Pergamon, New york, (1969).

[VERA,84] Veran, M. and Poitier, M., '*QNAP-2: A Portable Environment for Queueing Network Modelling*', Proc. Int. Conf. on Modelling Techniques and Tools for Performance Analysis, INRIA, (1984).

[WALS,84] Walstra, B.R., '*Iterative Analysis of Networks of Queues*', Ph.D Thesis, Tech. Rep. CSRI-166, Toronto Univ., (1984).

[WHIT,58] White, H. and Christie, L.S., '*Queueing with Preemptive Priorities or with Breakdown*', Op. Res., Vol.6, (1958) pp.79-95.

[WHIT,82] Whitt, W., '*The Marshall and Stoyan Bounds for IMRL/G/1 Queues are tight*', Oper. Res. lett., Vol. 1, pp.209-213, (1982).

[XENI,89] Xenios, N., ' *General Quueing Networks with Blocking*', Ph.D Thesis, Univ. of Bradford, (1989).

[ZAHO,87] Zahorjan, J., Lazowska, E.D. and Sevcik, K.C.,'*The Use of Approximations in Production Performance Evaluation Software*', Proc. Internat. Workshop on Modelling Techniques and Performance Evaluation, Paris, March (1987).

APPENDIX A

A1: Proof of equation 2.5

Given that the marginal arrival processes are Poisson, the marginal queue length distribution which is the distribution seen by an "outside observer" is the same as the distribution seen by an arriving customer. Furthermore, the state of a queueing system under PR or HOL will change by one at each departure or arrival. Therefore a departer and an arriver of the same class will perceive the same marginal distribution [COOP,81, pp.185-188]. Hence the marginal queue length distribution of class-r customers is the same as the one seen by a departing class-r customer. In other words, the following relation is satisfied.

$$\text{Prob}\left[\begin{array}{c} n_r \text{ class-r customers} \\ \text{in the system} \end{array}\right] = \text{Prob}\left[\begin{array}{l} \text{class-r departer} \\ \text{leaves } n_r \text{ class-r} \\ \text{customers behind him} \end{array}\right]$$

Since customers of the same class are served in FCFS order within the class, all class-r customers left in the system by a class-r departer arrive during the response time of the departer, thus:

$$\text{Prob}\left[\begin{array}{c} n_r \text{ class-r customer} \\ \text{in the systme} \end{array}\right] = \text{Prob}\left[\begin{array}{c} n_r \text{ class-r customer} \\ \text{arrives during the} \\ \text{response time of a class-r} \\ \text{departer} \end{array}\right]$$

Conditioning on the response time T_r of a class-r customer , and then using the law of total probability, yields to the following expression of the marginal steady state probabilities:

$$P_r(n_r) = \int_{T_r=0}^{+\infty} \exp(-\lambda_r T_r) \frac{(\lambda_r T_r)}{n_r!} \, dT_r \qquad (A.1.1)$$

After simple manipulations, the Z-transform of the marginal queue length distribution of class-r is given by

$$Q_r(z_r) = T_r^*(\lambda_r - \lambda_r z_r) \qquad (A.1.2)$$

Where $T_r^*(.)$ is the L.S.T of the response time distribution of class-r jobs under either PR or HOL discipline. The corresponding expressions can be seen in [JAIS,68].

Q.E.D.

APPENDIX B

B1: Proof of equation 3.35

First let'sgive the following lemma:

Lemma B.1 [FELL,65,pp.173]

Let a r.v. conforming to a binomial distribution $b(k;n,\sigma)$ such as:

$$b(k;n,\sigma) \triangleq \begin{bmatrix} n \\ k \end{bmatrix}\sigma^k(1-\sigma)^{n-k} \; ; \; B(m;n+1,r) \triangleq \sum_{k=0}^{m} b(k;n,\sigma)$$

the following relation is satisfied:

$$B(m;n+1,\sigma) = B(m;n,\sigma) - \sigma b(m;n,\sigma)$$

proof

The generating function or Z-transform of a binomial distribution with parameters n+1 and σ is:

$$b(z) = (1+\sigma+\sigma z)^{n+1} = (1+\sigma+\sigma z)^n(1+\sigma+\sigma z)$$

This implies that $b(k;n+1,\sigma) = b(k;n,\sigma) \oplus b(k;1,\sigma)$

$$= \sum_{i=0}^{k} b(k-i;n,\sigma)b(i;1,\sigma)$$

$$= b(k;n,\sigma)(1-\sigma) + b(k-1;n,\sigma)\sigma$$

Using the equation above in the expression of $B(m;n+1,\sigma)$, we will have:

$$B(m;n+1,\sigma) = (1-\sigma)\sum_{k=0}^{m} b(k;n,\sigma) + \sigma\sum_{k=1}^{m} b(k-1;n,\sigma)$$

$$= (1-\sigma)\sum_{k=0}^{m} b(k;n,\sigma) + \sigma\sum_{k=0}^{m-1} b(k-1;n,\sigma)$$

$$= \sum_{k=0}^{m} b(k;n,\sigma) - \sigma\sum_{k=0}^{m} b(k;n,\sigma) + \sigma\sum_{k=1}^{m-1} b(k;n,\sigma)$$

Where Finally $B(m;n+1,\sigma) = B(m;n,\sigma) - \sigma b(m;n,\sigma)$.

Q.E.D.

Consider a GE renewal arrival process (i.e., interarrival-times conform to GE distribution with parameter λ and $C_a^2=(2-\sigma)/\sigma)$.

It is known from renewal theory [COX,62], that

$$\text{Prob}[N(t)=n] = A_1(t)\oplus A^{(n-1)}(t) - A_1(t)\oplus A^{(n)}(t) \quad , n\geq 1 \quad \text{(b1.1)}$$

where $A(t)$ and $A_1(t)$ are the probability distribution function of the interarrival time and the time to the first arrival (renewal).

$A^{(n)}(t)$ is the n-fold convolution of A with itself.

$\tilde{A}_1$ is the remaining interarrival time, which due to the pseudo-memoryless property of GE, is exponentially distributed with paramater $\sigma\lambda$.

we have then;

$$A(t) = 1 - \sigma e^{-\sigma\lambda t}$$

and

$$A_1(t) = 1 - e^{-\sigma\lambda t}$$

The right hand side of equation (b1.1) may be expressed with respect L.S.T's as follows:

$$\text{Prob}[N(t)=n] \longleftrightarrow A_1^*(\theta)[A^*(\theta)]^{n-1} - A_1^*(\theta)[A^*(\theta)]^n \quad n\geq 0, t>0 \quad \text{(b1.2)}$$

Let us first evaluate the probability distribution function corresponding to the first term of expression b1.2.

$$A_1^*(\theta)[A^*(\theta)]^{n-1} = \left[\frac{\sigma\lambda}{\sigma\lambda+\theta}\right]\left[1-\sigma+\sigma\frac{\sigma\lambda}{\sigma\lambda+\theta}\right]^{n-1}$$

Using Newton's expansion, we obtain:

$$A_1^*(\theta)[A^*(\theta)]^{n-1} = \left[\frac{\sigma\lambda}{\sigma\lambda + \theta}\right]\sum_{k=0}^{n-1}\binom{n-1}{k}(1-\sigma)^{n-1-k}\sigma^k\left[\frac{\sigma\lambda}{\sigma\lambda + \theta}\right]^k$$

$$= (1-\sigma)^{n-1}\sum_{k=0}^{n-1}\binom{n-1}{k}\left[\frac{\sigma}{1-\sigma}\right]^k\left[\frac{\sigma\lambda}{\sigma\lambda + \theta}\right]^{k+1}$$

Inverting the L.S.T's and integrating, we obtain;

$$A_1(t)\oplus A^{(n-1)}(t) = (1-\sigma)^{n-1}\sum_{k=0}^{n-1}\binom{n-1}{k}\left[\frac{\sigma}{1-\sigma}\right]^k\left\{1 - e^{-\sigma\lambda t}\sum_{i=0}^{k}\frac{(\sigma\lambda t)^i}{i!}\right\}$$

$$= (1-\sigma)^{n-1}\sum_{k=0}^{n-1}\binom{n-1}{k}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$- (1-\sigma)^{n-1}e^{-\sigma\lambda t}\sum_{k=0}^{n-1}\binom{n-1}{k}\left[\frac{\sigma}{1-\sigma}\right]^k\sum_{i=0}^{k}\frac{(\sigma\lambda t)^i}{i!}$$

$$= (1-\sigma)^{n-1}\left[1+\frac{\sigma}{1-\sigma}\right]^{n-1} - (1-\sigma)^{n-1}e^{-\sigma\lambda t}\sum_{i=0}^{n-1}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n-1}\binom{n-1}{k}\left[\frac{\sigma}{1-\sigma}\right]^k$$

Finally, we will have:

$$A_1(t)\oplus A^{(n-1)}(t) = 1 - (1-\sigma)^{n-1}e^{-\sigma\lambda t}\sum_{i=0}^{n-1}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n-1}\binom{n-1}{k}\left[\frac{\sigma}{1-\sigma}\right]^k \text{ for } n\geq 1 \quad (b1.3)$$

Similarly the second term will be given by:

$$A_1(t)\oplus A^{(n)}(t) = 1 - (1-\sigma)^{n}e^{-\sigma\lambda t}\sum_{i=0}^{n}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n}\binom{n}{k}\left[\frac{\sigma}{1-\sigma}\right]^k \text{ for } n\geq 1 \quad (b1.4)$$

Applying (b1.3) and (b1.4) to (b1.1), we will have

$$\text{Prob}[N(t)=n] = (1-\sigma)^n e^{-\sigma\lambda t}\sum_{i=0}^{n}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n}\begin{bmatrix} n \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$- (1-\sigma)^{n-1} e^{-\sigma\lambda t}\sum_{i=0}^{n-1}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n-1}\begin{bmatrix} n-1 \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$= (1-\sigma)^n e^{-\sigma\lambda t}\left[1+\frac{\sigma}{1-\sigma}\right]^n + (1-\sigma)^n e^{-\sigma\lambda t}\sum_{i=1}^{n}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n}\begin{bmatrix} n \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$- (1-\sigma)^{n-1} e^{-\sigma\lambda t}\left[1+\frac{\sigma}{1-\sigma}\right]^{n-1} - (1-\sigma)^{n-1} e^{-\sigma\lambda t}\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n-1}\begin{bmatrix} n-1 \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$-(1-\sigma)^n e^{-\sigma\lambda t}\sum_{i=1}^{n}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n}\begin{bmatrix} n \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k - (1-\sigma)^{n-1} e^{-\sigma\lambda t}\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n-1}\begin{bmatrix} n-1 \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$-(1-\sigma)^n e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^n}{n!}\left[\frac{\sigma}{1-\sigma}\right]^n + (1-\sigma)^n e^{-\sigma\lambda t}\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n}\begin{bmatrix} n \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$- (1-\sigma)^{n-1} e^{-\sigma\lambda t}\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^i}{i!}\sum_{k=i}^{n-1}\begin{bmatrix} n-1 \\ k \end{bmatrix}\left[\frac{\sigma}{1-\sigma}\right]^k$$

$$-\sigma^n e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^n}{n!} + \sum_{k=1}^{n-1}\frac{(\sigma\lambda t)^i}{i!}e^{-\sigma\lambda t}\left\{\sum_{k=i}^{n-1}\begin{bmatrix} n \\ k \end{bmatrix}\sigma^k(1-\sigma)^{n-k} - \sum_{k=i}^{n-1}\begin{bmatrix} n-1 \\ k \end{bmatrix}\sigma^k(1-\sigma)^{n-1-k}\right\}$$

$$-\sigma^n e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^n}{n!} + \sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^i}{i!}e^{-\sigma\lambda t}\left\{1 - \sum_{k=0}^{i-1}\begin{bmatrix} n \\ k \end{bmatrix}\sigma^k(1-\sigma)^{n-k} - 1 + \sum_{k=0}^{i-1}\begin{bmatrix} n-1 \\ k \end{bmatrix}\sigma^k(1-\sigma)^{n-1-k}\right\}$$

$$=\sigma^{n}e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^{n}}{n!}+\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^{i}}{i!}e^{-\sigma\lambda t}\left\{\sum_{k=0}^{i-1}\left[\begin{smallmatrix}n-1\\k\end{smallmatrix}\right]\sigma^{k}(1-\sigma)^{n-1-k}\right.$$

$$\left.-\sum_{k=0}^{i-1}\left[\begin{smallmatrix}n\\k\end{smallmatrix}\right]\sigma^{k}(1-\sigma)^{n-k}\right\} \qquad \text{(b1.5)}$$

Given that $b(k;n,\sigma)\triangleq\left[\begin{smallmatrix}n\\k\end{smallmatrix}\right]\sigma^{k}(1-\sigma)^{n-k}$ and $B(m;n,\sigma)\triangleq\sum_{k=0}^{m}b(k;n,\sigma)$,

where $b(k;n,\sigma)$ is the binomial distribution with parameters n, σ.

Substituing the above quantities in the equation (b1.5), we obtain

$$\text{Prob}[N(t)=n]=\sigma^{n}e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^{n}}{n!}+\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^{i}}{i!}e^{-\sigma\lambda t}\sigma\left\{B(i-1;n-1,\sigma)-B(i-1;n,\sigma)\right\} \qquad \text{(b1.6)}$$

Using lemma B.1 , equation (b1.6) becomes:

$$\text{Prob}[N(t)=n]=\sigma^{n}e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^{n}}{n!}+\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^{i}}{i!}e^{-\sigma\lambda t}\sigma b(i-1;n-1,\sigma)$$

$$=\sigma^{n}e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^{n}}{n!}+\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^{i}}{i!}e^{-\sigma\lambda t}\sigma\left[\begin{smallmatrix}n-1\\i-1\end{smallmatrix}\right]\sigma^{i-1}(1-\sigma)^{n-i}$$

$$=\sigma^{n}e^{-\sigma\lambda t}\frac{(\sigma\lambda t)^{n}}{n!}+\sum_{i=1}^{n-1}\frac{(\sigma\lambda t)^{i}}{i!}e^{-\sigma\lambda t}\left[\begin{smallmatrix}n-1\\i-1\end{smallmatrix}\right]\sigma^{i}(1-\sigma)^{n-i}$$

Finally, the distribution of the renewal counting process N(t) is:

$$\text{Prob}[N(t)=n]=\sum_{i=1}^{n}\frac{(\sigma\lambda t)^{i}}{i!}e^{-\sigma\lambda t}\left[\begin{smallmatrix}n-1\\i-1\end{smallmatrix}\right]\sigma^{i}(1-\sigma)^{n-i}\ ,\ n\geq 1 \qquad \text{(b1.7)}$$

For the special case n = 0, we will have;

$$\text{Prob}[N(t)=n] = 1 - \sum_{n=1}^{\infty} \text{Prob}[N(t) = n]$$

Substituing equation (b1.7) in the equation above, we will have;

$$\text{Prob}[N(t)=0] = 1 - \sum_{n=1}^{\infty} \left\{ \sum_{i=1}^{n} \frac{(\sigma\lambda t)^i}{i!} e^{-\sigma\lambda t} \begin{bmatrix} n-1 \\ i-1 \end{bmatrix} \sigma^i (1-\sigma)^{n-i} \right\}$$

$$= 1 - e^{-\sigma\lambda t} \sum_{i=1}^{\infty} \frac{(\sigma\lambda t)^i}{i!} \sum_{n=i}^{\infty} \begin{bmatrix} n-1 \\ i-1 \end{bmatrix} \sigma^i (1-\sigma)^{n-i}$$

$$= 1 - e^{-\sigma\lambda t} \sum_{i=1}^{\infty} \frac{(\sigma\lambda t)^i}{i!} = 1 - e^{-\sigma\lambda t}(e^{\sigma\lambda t} - 1)$$

Where finally $\text{Prob}[N(t) = n] = e^{-\sigma\lambda t}$ (b1.8)

Q.E.D

B2: Proof of corollary 3.5

Let denote $\text{Prob}[\bar{B} = k] = b_k$, for $k \geq 1$.

It is known that the number of occurences in time t for a compound Poisson process is [GROSS 85, p.285]:

$$\text{Prob}[N(t) = n] = \sum_{i=0}^{n} \frac{(\sigma\lambda t)^i}{i!} e^{-\sigma\lambda t} b_n^{(i)} = e^{-\sigma\lambda t} b_n^{(0)} + \sum_{i=1}^{n} \frac{(\sigma\lambda t)^i}{i!} e^{-\sigma\lambda t} b_n^{(i)}$$

Where $b_n^{(i)}$ is the i-fold convolution of $\{b_n\}$ with itself,

and $$b_n^{(0)} = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \geq 1 \end{cases}$$

Therefore, in GE case, we will have:

$$\text{Prob}[N(t) = n] = \begin{cases} \sum_{i=1}^{n} \frac{(\sigma\lambda t)^i}{i!} e^{-\sigma\lambda t} b_n^{(i)} & , \quad n \geq 1 \\ e^{-\sigma\lambda t} & , \quad n = 0 \end{cases}$$

where,

$$b_n^{(i)} = \begin{bmatrix} n-1 \\ i-1 \end{bmatrix} \sigma^i (1-\sigma)^{n-i}$$

Noticing that the right hand side of the above equation corresponds to a negative binomial distribution with parameter $(n;i,\sigma)$ and due to the fact that the negative binomial is the i-fold convolution of a geometric distribution (σ) with itself, we will have then:

$$b_n = \sigma(1-\sigma)^{n-1} \quad , \text{ for } n \geq 1$$

Q.E.D

B3: Proof of equation 3.37

Consider first a G/G/1 queue with a single class of jobs where we denote $D^*(.)$ as the L.S.T of the interdepature time process and $P_d(n)$ the probability that a departer leaves behind him n jobs in the queue.

Let us consider also a departer leaving the system at a specific time t, the problem is to determine the timc up to the next departer.

With probability $P_d(0)$, the current departer leaves an empty system. Therefore, the time up to the next departer is the sum of the remaining interarrival time and full service time.

On the other hand, with probability $(1-P_d(0))$, the departer in question leaves a non-empty system. In this case, the time up to the next departer is just the duration of the service time.

In terms of L.S.T, this may be formulated as follows:

$$D^*(\theta) = P_d(0)\ \hat{A}^*(\theta)S^*(\theta) + (1-P_d(0))S^*(\theta) \qquad (b3.1)$$

Where $\hat{A}^*(.)$ is the L.S.T of the remaining interarrival time distribution. For GE interarrival process, and because of the pseudo-memoryless property of GE, the remaining interarrival time is exponentially distributed with parameter $\sigma\lambda$. therefore, we have

$$\hat{A}^*(\theta) = \frac{\sigma\lambda}{\sigma\lambda + \theta} \qquad (b3.2)$$

and due to Cooper result [COOP 81,pp.185–188], we have

$P_d(0)$ = Prob[arriver find an empty system].

Using the notion of ordered bulk; in the sense that each member of the arriving bulk is assigned a number which corresponds to the order of occurences. This is true in GE case , since it can be interpreted also as the limiting case of H_2 with negligible interarrival time between members of the bulk.

Thus $P_d(0)$ = Prob[the system is empty]

*Prob[an arriver to be the first member of the bulk]

Because the bulk size is geometrically distributed with parameter σ and because of the memoryless property of the geomtric distribution, we have:

Prob[an arriver to be the first member of the bulk] = σ

Thus $$P_d(0) = \sigma P_0 = \sigma(1-\rho) \qquad \text{(b3.3)}$$

Substituting (b3.2) and (b3.3) in (b3.1) and using the L.S.T of the GE distribution (for the service time), we obtain:

$$D^*(\theta) = \sigma P_0\left[\frac{\sigma\lambda}{\sigma\lambda+\theta}\right]\left[1-\tau+\tau\frac{\tau\mu}{\tau\mu+\theta}\right] + (1-\sigma P_0)\left[1-\tau+\tau\frac{\tau\mu}{\tau\mu+\theta}\right] \qquad \text{(b3.4)}$$

Finally by successive differentiations of equation (b3.4) yields

$$\langle d\rangle = \frac{1}{\lambda} \qquad \text{(b3.5)}$$

which is expected!

The second moment is given by:

$$\langle d^2\rangle = \frac{2}{\tau\mu^2} + \frac{2(1-\rho)}{\sigma\lambda^2} + \frac{2(1-\rho)}{\lambda\mu} \qquad \text{(b3.6)}$$

Using the definition of the squared coefficient of variation together with equations (b3.5) and (b3.6), equation (3.37) follows after some simple manipulations.

Q.E.D

APPENDIX C

C1: Numerical results (chapter 4)

Example 4.1 $H_2/H_2/1$ PR queue (3 Classes)

Table 4.1a: Raw data for PR $H_2/H_2/1$ queue

(Results Table 4.1b)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	4	3	10	9
2	1	7	5	3
3	3	2	15	11

Table 4.1b: Comparison of GE-type $\langle n_r \rangle$ and $P_r(0)$, (GE), in relation to simulations (SIM), r=1,2,3.

	$\langle n_1 \rangle$	$\langle n_2 \rangle$	$\langle n_3 \rangle$	$P_1(0)$	$P_2(0)$	$P_3(0)$
GE	2.4	3.1667	20.50	0.6	0.594	0.254
SIM[1]	2.055	2.250	19.23	0.6	0.589	0.278

1: The tolerance of the confidence intervals is within 5% of the simulated values.

Example 4.2 $H_2/H_2/1$ HOL queue (3 Classes)

Table 4.2a: Raw data for HOL $H_2/H_2/1$ queue

(Results Table 4.2b)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1.5	3	5	5
2	6	7	15	10
3	2.5	2	12.5	2

Table 4.2b: Comparison of GE-type $\langle n_r \rangle$ and $P_r(0)$, (GE), in relation to simulations (SIM), r=1,2,3.

	$\langle n_1 \rangle$	$\langle n_2 \rangle$	$\langle n_3 \rangle$	$P_r(0)$	$P_2(0)$	$P_3(0)$
GE	1.48	16.13	42.089	0.597	0.306	0.144
SIM	1.22	14.43	41.47	0.590	0.307	0.165

Example 4.3 $E_2/E_2/1$ PR queue (3 Classes)

Table 4.3a: Raw data for PR $E_2/E_2/1$ queue

(Results Table 4.3b)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	4	0.5	10	0.5
2	1	0.5	5	0.5
3	3	0.5	15	0.5

Table 4.3b: Comparison of GE-type $\langle n_r \rangle$ and $P_r(0)$, (GE), in relation to simulations (SIM), r=1,2,3.

	$\langle n_1 \rangle$	$\langle n_2 \rangle$	$\langle n_3 \rangle$	$P_1(0)$	$P_2(0)$	$P_3(0)$
GE	0.433	0.416	2.125	0.6	0.627	0.342
SIM	0.494	0.464	2.21	0.599	0.629	0.338

Example 4.4 $E_2/E_2/1$ HOL queue (3 Classes)

Table 4.4a: Raw data for PR $E_2/E_2/1$ queue

(Results Table 4.4b)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1.5	0.5	5	0.5
2	6	0.5	15	0.5
3	2.5	0.5	12.5	0.5

Table 4.4b: Comparison of GE-type $\langle n_r \rangle$ and $P_r(0)$, (GE), in relation to simulations (SIM), r=1,2,3.

	$\langle n_1 \rangle$	$\langle n_2 \rangle$	$\langle n_3 \rangle$	$P_1(0)$	$P_2(0)$	$P_3(0)$
GE	0.357	1.838	4.311	0.6422	0.291	0.255
SIM	0.408	1.949	4.35	0.645	0.280	0.247

Example 4.5 $D,M,H_2/H_2,E_2,U[0,0.4]/1$ PR queue (3 Classes)

Table 4.5a: Raw data for PR $D,M,H_2/H_2,E_2,U[0,0.4]/1$ queue

(Results Table 4:5b)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1.5	0	5	6
2	3	1	15	0.5
3	1	2.5	5	0.333

Table 4.5b: Comparison of GE-type $\langle n_r \rangle$ and $P_r(0)$, (GE), in relation to simulations (SIM), r=1,2,3.

	$\langle n_1 \rangle$	$\langle n_2 \rangle$	$\langle n_3 \rangle$	$P_1(0)$	$P_2(0)$	$P_3(0)$
GE	0.535	1.914	2.344	0.70	0.555	0.487
SIM	0.605	1.946	2.067	0.7	0.589	0.504

Example 4.6 $E_2,D,U[0,1]/H_2,M,E_2/1$ HOL queue (3 Classes)

Table 4.6a: Raw data for HOL $E_2,D,U[0,1]/H_2,M,E_2/1$queue

(Results Table 4.6b)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	2	0.5	10	5
2	4.5	0	15	1
3	1.5	0.333	5	0.5

Table 4.6b: Comparison of GE-type $\langle n_r \rangle$ and $P_r(0)$, (GE),in relation to simulations (SIM), r=1,2;3.

	$\langle n_1 \rangle$	$\langle n_2 \rangle$	$\langle n_3 \rangle$	$P_1(0)$	$P_2(0)$	$P_3(0)$
GE	0.45	1.35	1.45	0.669	0.35	0.393
SIM	0.485	1.506	1.727	0.673	0.342	0.395

Table 4.7: Raw data for PR and HOL G/G/1 queue

(Figs. 4.1-4.2)

Class r	λ_r	C^2_{ar} GE,E_2	M	$H_2(\kappa)$,GE	μ_r	C^2_{sr} GE,E_2	M	$H_2(\kappa)$,GE
1	2	0.5	1	15	20	0.5	1	12
2	1.5	0.5	1	18	30	0.5	1	10
3	1	0.5	1	7	15	0.5	1	5

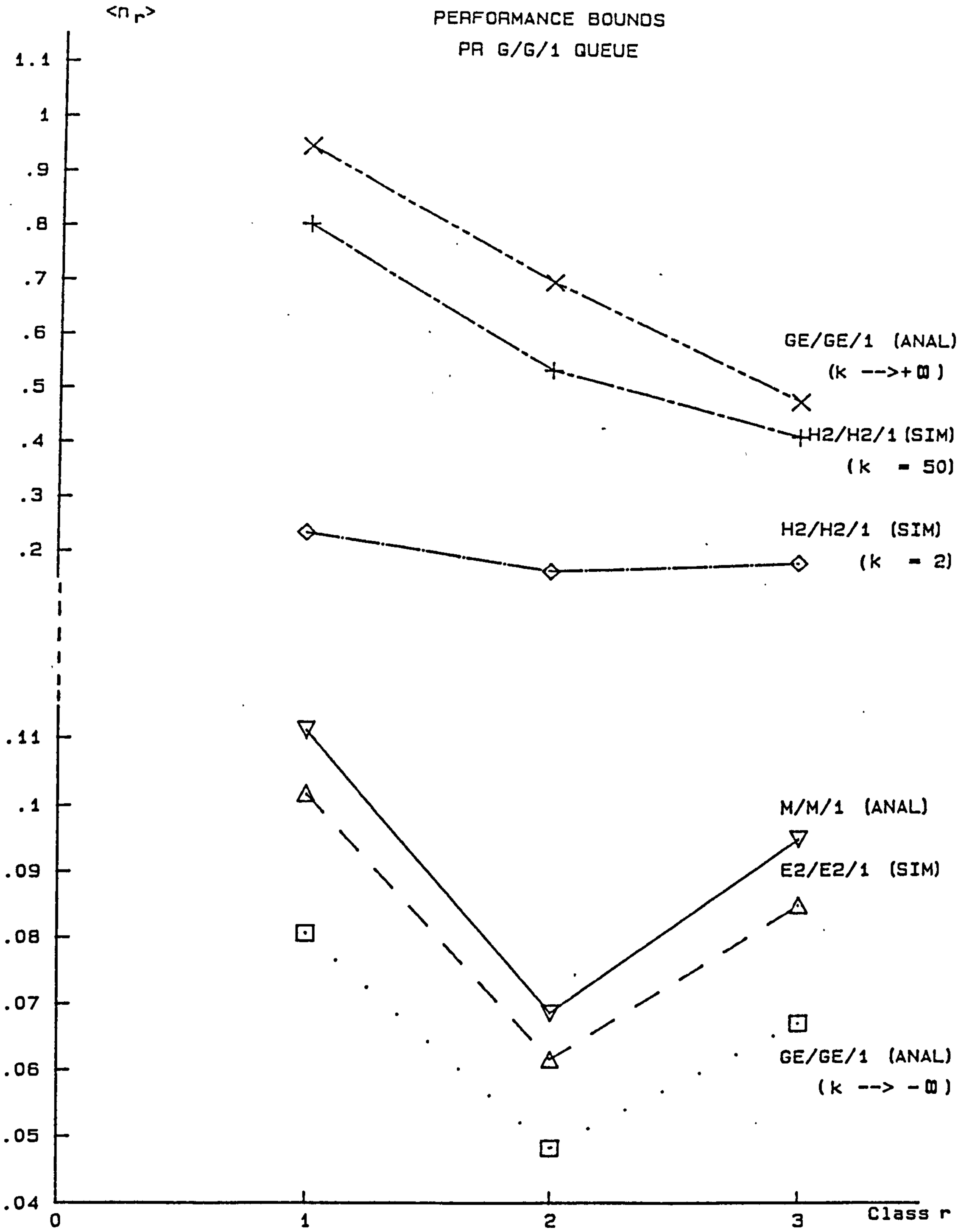

Fig. 4.1. Performance bounds (PR GE/GE/1). $\langle n_r \rangle$ vs class r for various PR G/G/1 simulation (SIM) and analytic (ANAL) solution for the raw data in table 4.7

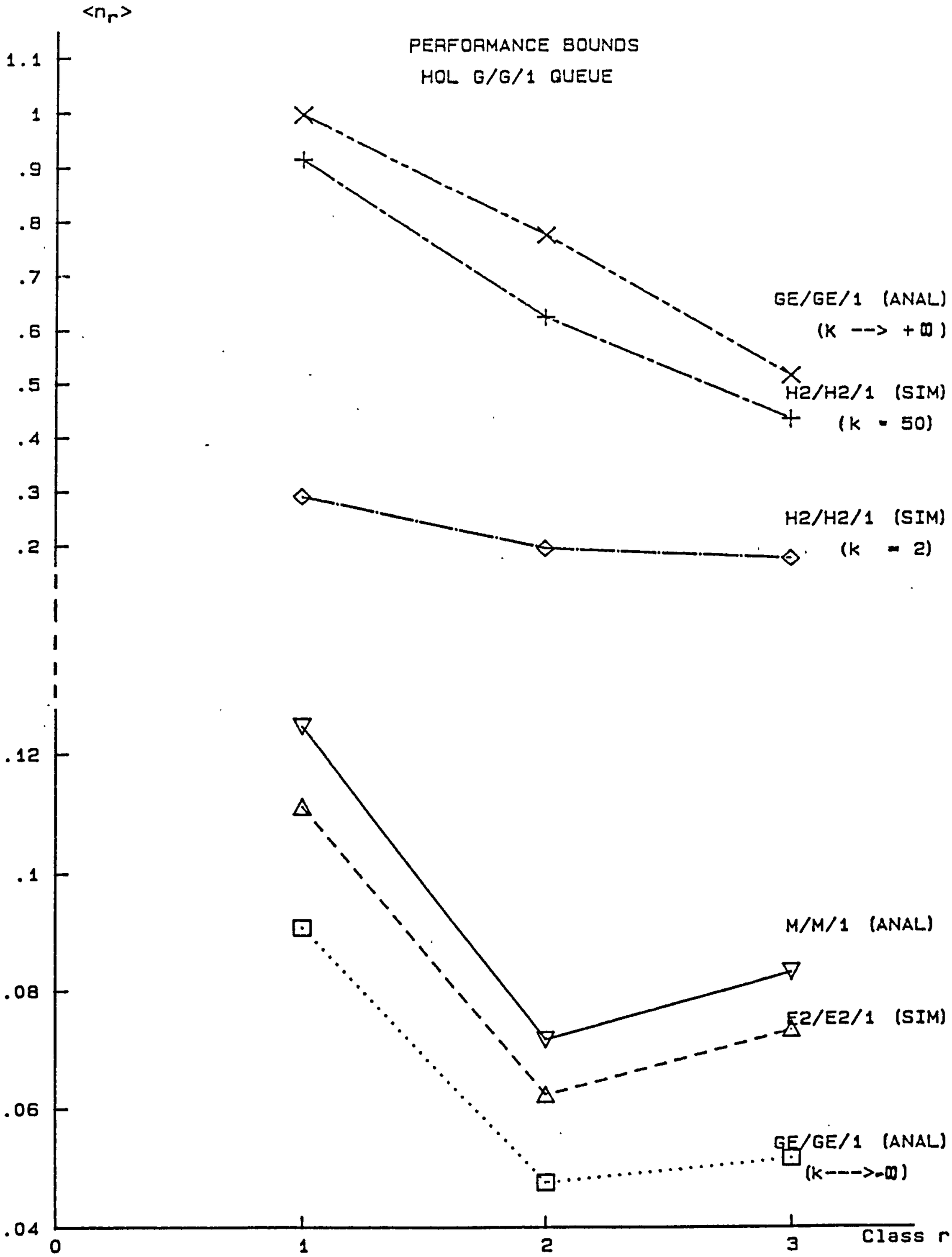

Fig. 4.2 Performance bounds (HOL GE/GE/1). $\langle n_r \rangle$ vs class r

for various HOL G/G/1 simulation (SIM) and analytic (ANAL) solution for the raw data of table 4.7

APPENDIX D

D1: Proof of theorem 5.1

Clearly , for $\underline{n} = \underline{0}$, we do have $P(\underline{0}) = \frac{1}{Z}$ (D1.1)

Let us assume in the following analysis that job of class-r, r=1,2,...,R, is in service. Since the discipline is PR, class-r jobs are the highest priority jobs present in the system. Therefore, the state of the system is clearly identified by $\underline{S}-\underline{n}=(0,\ldots,0,n_r,\ldots,n_R)$.

$$\text{Subsequently } h_\ell(\underline{S}) = \begin{cases} 1 \text{ for } \ell = r \\ \\ 0 \text{ for } \ell \neq r \end{cases}$$

Using the above arguments, the ME joint probability distribution is thus given by

$$P(\underline{n}) = \frac{1}{Z} g_r \prod_{\ell=r}^{R} x_\ell^{n_\ell} \quad ,\underline{n}=(0,\ldots,0,n_r,\ldots,n_R) \quad \text{(D1.2)}$$

The marginal utilisation, ρ_r, r=1,2,..,R, is expressed as follows:

Using the utilisation constraint equation (5.2), we obtain

$$\rho_r = \sum_{n_r=1}^{\infty} \sum_{n_{r+1}=0}^{\infty} \cdots \sum_{n_R=0}^{\infty} P(0,\ldots,0,n_r,\ldots,n_R)$$

$$= \sum_{n_r=1}^{\infty} \sum_{n_{r+1}=0}^{\infty} \cdots \sum_{n_R=0}^{\infty} \frac{1}{Z} g_r \prod_{\ell=r}^{R} x_\ell^{n_\ell}$$

$$= \frac{1}{Z} g_r \sum_{n_r=1}^{\infty} x_r^{n_r} \sum_{n_{r+1}=0}^{\infty} x_{r+1}^{n_{r+1}} \ldots \sum_{n_R=0}^{\infty} x_R^{n_R}$$

Performing all the infinite summations leads to

$$\rho_r = \frac{1}{Z} g_r x_r \prod_{\ell=r}^{R} \frac{1}{1-x_\ell} \qquad \text{(D1.3)}$$

Subsequently, the overall utilisation of the server is given by:

$$\rho = \sum_{r=1}^{R} \rho_r = \frac{1}{Z} \sum_{r=1}^{R} g_r x_r \prod_{\ell=r}^{R} \frac{1}{1-x_\ell} \qquad \text{(D1.4)}$$

Meanwhile, the normalisation constant is determined as follows:

Using the normalisation constraint (eq.5.1) together with (eq.5.5), we will have

$$\sum_{\underline{n}=\underline{0}}^{\infty} \frac{1}{Z} \prod_{r=1}^{R} g_r^{h_r(\underline{S})} x_r^{n_r} = 1$$

which may also be written as follows:

$$\frac{1}{Z} \sum_{n_1=0}^{\infty} g_1^{h_1(\underline{S})} x_1^{n_1} \sum_{n_2=0}^{\infty} g_2^{h_2(\underline{S})} x_2^{n_2} \ldots \sum_{n_R=0}^{\infty} g_R^{h_R(\underline{S})} x_R^{n_R} = 1$$

After performing all the summations, we obtain

$$\frac{1}{Z} + \frac{1}{Z} \sum_{r=1}^{R} g_r x_r \prod_{\ell=r}^{R} \frac{1}{1-x_\ell} = 1$$

Using finally equation (D1.4) leads to

$$Z = \frac{1}{1-\rho} \quad \text{, as expected!} \qquad \text{(D1.5)}$$

Consequently, equation (5.7) is obtained by substitution of (eq.D1.5) and (eq.D1.1) in (eq.D1.2).

Now, let express the Lagrangian coefficients $\{x_r\}$ and $\{g_r\}$ with respect the given mean values $\{\rho_r\}$ and $\{<n_r>\}$.

From the mean queue length constraint (eq.5.3), we have

$$<n_r> = \sum_{\underline{n}=\underline{0}}^{\infty} n_r \, P(\underline{n}) = \sum_{\underline{n}=\underline{0} \wedge n_r=1}^{\infty} n_r \, P(\underline{n})$$

$$= \sum_{n_R=0}^{\infty} \cdots \sum_{n_r=1}^{\infty} \cdots \sum_{n_1=0}^{\infty} n_r \frac{1}{Z} \prod_{s=1}^{R} g_s^{h_s(\underline{S})} x_s^{n_s}$$

$$= \frac{1}{Z} \sum_{n_R=0}^{\infty} g_R^{h_R(\underline{S})} x_R^{n_R} \cdots \sum_{n_r=1}^{\infty} n_r g_r^{h_r(\underline{S})} x_r^{n_r} \cdots \sum_{n_1=0}^{\infty} g_1^{h_1(\underline{S})} x_1^{n_1}$$

$$= \frac{1}{Z} \sum_{n_R=0}^{\infty} g_R^{h_R(\underline{S})} x_R^{n_R} \cdots \sum_{n_r=1}^{\infty} n_r g_r^{h_r(\underline{S})} x_r^{n_r} \cdots \left\{ 1 + \frac{g_1 x_1}{1-x_1} \right\}$$

$$= \frac{1}{Z} \sum_{n_R=0}^{\infty} g_R^{h_R(\underline{S})} x_R^{n_R} \cdots \sum_{n_r=1}^{\infty} n_r g_r^{h_r(\underline{S})} x_r^{n_r} \cdots \sum_{n_2=0}^{\infty} g_2^{h_2(\underline{S})} x_2^{n_2}$$

$$+ \frac{1}{Z} g_1 x_1 \frac{x_r}{1-x_r} \prod_{s=1}^{R} \frac{1}{1-x_s}$$

Performing the same calculations up to the (r-1) sum, we will obtain:

$$\langle n_r \rangle = \frac{1}{Z} \sum_{n_R=0}^{\infty} g_r x_R^{n_R} \cdots \sum_{n_r=1}^{\infty} n_r x_r^{n_r} + \frac{1}{Z} \frac{x_r}{1-x_r} \sum_{\ell=1}^{r-1} g_\ell x_\ell \prod_{s=\ell}^{R} \frac{1}{1-x_s}$$

Proceeding in the same manner in the remaining infinite sums, we end up with the following expression:

$$\langle n_r \rangle = \frac{1}{Z} \frac{1}{1-x_r} g_r x_r \prod_{s=r}^{R} \frac{1}{1-x_s} + \frac{1}{Z} \frac{x_r}{1-x_r} \sum_{\ell=1}^{r-1} g_\ell x_\ell \prod_{s=\ell}^{R} \frac{1}{1-x_s}$$

Using equation (D1.3) in the above equation yields:

$$\langle n_r \rangle = \frac{1}{1-x_r} \rho_r + \frac{x_r}{1-x_r} \sum_{\ell=1}^{r-1} \rho_\ell$$

Finally solving the equation above with respect to x_r, equation (5.8) follows.

The equation (5.9) is therefore obtained by substituting the expression of x_r (5.8) in the equation (D1.3).

Q.E.D

D:2 Proof of corollary 5.1

i/ For $n_r=0$.

$$P_r(0) = \sum_{\underline{n}=\underline{0} \wedge n_r=0}^{\infty} P(\underline{n}) = \frac{1}{Z} + \frac{1}{Z} \sum_{\underline{n}=\underline{1}_s \wedge n_r=0}^{\infty} \prod_{s=1}^{R} g_s^{h_s(\underline{S})} x_s^{n_s}$$

Performing the same type of calculations in the infinite sums as before, we will obtain

$$P_r(0) = \frac{1}{Z} \left\{ 1 + \sum_{s=1 \wedge s \neq r}^{R} g_s x_s \prod_{\ell=s \wedge \ell \neq r}^{R} \frac{1}{1-x_\ell} \right\}$$

Breaking down the above expression into a sum of two terms leads to

$$P_r(0) = \frac{1}{Z} + \frac{1}{Z}\sum_{s=1}^{r-1} g_s x_s \prod_{\ell=s\wedge\ell\neq r}^{R} \frac{1}{1-x_\ell} + \frac{1}{Z}\sum_{s=r+1}^{R} g_s x_s \prod_{\ell=s}^{R} \frac{1}{1-x_\ell}$$

Substituting equations (D1.5) and (D1.3) (given in appendix D1) in the equation above, we will have

$$P_r(0) = 1-\rho + (1-x_r)\sum_{s=1}^{r-1} \rho_s + \sum_{s=r+1}^{R} \rho_s$$

After simple manipulations, we end up with the following expression:

$$P_r(0) = 1-\rho_r-\gamma_{r-1}x_r$$

ii/ For $n_r>0$.

$$P_r(n) = \sum_{\underline{n}=\underline{0}\wedge n_r=n}^{\infty} P(\underline{n}) = \frac{1}{Z} \sum_{\underline{n}=\underline{0}\wedge n_r=n}^{\infty} \prod_{s=1}^{R} g_s^{h_s(\underline{S})} x_s^{n_s}$$

Performing the infinite sums, leads to

$$P_r(n) = \frac{1}{Z} x_r^n \left\{ g_r \prod_{s=r+1}^{R} \frac{1}{1-x_s} + \sum_{\ell=1}^{r-1} g_\ell x_\ell^{n_\ell} \prod_{s=\ell\wedge s\neq r}^{R} \frac{1}{1-x_s} \right\}$$

Using the ME expression of the utilisation, ρ_r, given by equation (D1.3), the equation above is subsequently expressed by

$$P_r(n) = (\rho_r+\gamma_{r-1}x_r)(1-x_r)x^{n-1}, \text{ for } r = 1,2,\ldots,R$$

Q.E.D

D3: Proof of corollary 5.2

The proof is based on the form of the ME solution for the joint queue length distribution.

i/ For $\underline{n} = \underline{1}_r$.

Using ME solution (eq. 5.7), it is clear that we have,

$$P(\underline{0}) = \frac{1}{Z} \text{ , and } P(\underline{1}_r) = \frac{1}{Z} g_r x_r$$

Consequently, we have $P(\underline{1}_r) = g_r x_r P(\underline{0})$.

ii/ For $\underline{n}=\underline{1}_r$, $n_r=1$.

In this case class-r job is assumed to be the highest priority element present in the system. Let us suppose also that 's' is the class index of the job which is going to take over the service, when the class-r job leaves the system. In other words, 's' is the smallest integer greater than r with $n_s>o$.

It is clear then, the successive ME joint probability distributions are given by

$$P(\underline{n}) = \frac{1}{Z} g_r x_r \prod_{\ell=s}^{R} x_\ell^{n_\ell}$$

$$P(\underline{n}-\underline{1}_r) = \frac{1}{Z} g_s \prod_{\ell=s}^{R} x_\ell^{n_\ell}$$

Thus, clearly we have

$$P(\underline{n}) = \frac{g_r x_r}{g_s} P(\underline{n}-\underline{1}_r)$$

iii/ For $\underline{n} \neq \underline{1}_r$, $n_r > 1$.

Given a vector state $\underline{n}=(n_1,n_2,\ldots,n_R)$, and class-s job is receiving service. obviously, we have $n_1=n_2=\ldots=n_{s-1}=0$ and $n_s>0$. Therefore, the ME joint probability is given by

$$P(\underline{n}) = \frac{1}{Z} g_s \prod_{\ell=s}^{R} x_\ell^{n_\ell}$$

For $r\epsilon[s,R]$ with $n_r>1$, the ME solution of the joint probability distribution when one class-r job is removed from the system is given by:

$$P(\underline{n}-\underline{1}_r) = \frac{1}{Z} \left[g_s \prod_{\ell=s \wedge \ell \neq r}^{R} x_\ell^{n_\ell} \right] x_r^{n_r-1}$$

Given the two equations above, we will have the following relation:

$$P(\underline{n}) = x_r P(\underline{n}-\underline{1}_r)$$

Q.E.D.

□4: Proof of theorem 5.2

In contrary to PR discipline, lower priority job may still undergo service while high-priority are arriving to the system. Therefore, using the ME solution (5.5) together with the utilisation constraint (5.2), the ME probability to have $\underline{n}$ jobs with class-r job in service can be clearly given by

$$P(\underline{n}, \text{ class-r job in service}) = \frac{1}{Z} g_r \prod_{s=1}^{R} x_s^{n_s}$$

Thus, using the law of total probability, the joint ME queue length distribution is given by

$$P(\underline{n}) = \frac{1}{Z} \sum_{r=1 \wedge n_r>0}^{R} g_r \prod_{s=1}^{R} x_s^{n_s} \qquad \text{(D4.1)}$$

To determine the Lagrangian coefficients $\{g_r\}$ and $\{x_r\}$, let us first express the class-r utilisation, ρ_r, with respect to $\{g_r\}$ and $\{x_r\}$.

From the utilisation constraint (5.2), we have

$$\rho_r = \sum_{\underline{S} \epsilon Q} h_r(\underline{S}) P(\underline{S}) = \sum_{\underline{n}=\underline{1}_r}^{\infty} P(\underline{n}, \text{ class-r job in service})$$

$$= \frac{1}{Z} g_r \sum_{\underline{n}=\underline{1}_r}^{\infty} \prod_{s=1}^{R} x_s^{n_s}$$

Performing all the infinite summations, yields

$$\rho_r = \frac{1}{Z} g_r x_r \prod_{s=1}^{R} \frac{1}{1-x_s} \qquad \text{(D4.2)}$$

The overall utilisation is then given by

$$\rho = \frac{1}{Z} \sum_{r=1}^{R} g_r x_r \prod_{s=1}^{R} \frac{1}{1-x_s} \qquad \text{(D4.3)}$$

From the normalisation constraint, (eq.5.1), the normalisation constant, Z, can be easily shown to be also given by

$$Z = \frac{1}{1-\rho} \qquad \text{(D4.4)}$$

Therefore, equation (5.12) is obtained by substituting the expression of the normalising constant in (D4.1).

The Lagrangian coefficients $\{x_r\}$ and $\{g_r\}$ are thus evaluated following similar steps to the ones used in appendix D1.

Q.E.D.

D5: Proof of corollary 5.3

i/ For $n_r=0$.

From the law of total probability

$$P_r(0) = \sum_{\underline{n}=\underline{0}\wedge n_r=0}^{\infty} P(\underline{n}) = \frac{1}{Z} \sum_{\underline{n}=\underline{0}\wedge n_r=0}^{\infty} \prod_{s=1}^{R} x_s^{n_s} \left\{ \sum_{\ell=1\wedge n_\ell\neq 0}^{R} g_\ell \right\}$$

Performing the same type of calculations as before, we will have

$$P_r(0) = \frac{1}{Z} \left\{ 1 + \prod_{s=1\wedge s\neq r}^{R} \frac{1}{1-x_s} \sum_{\ell=1\wedge \ell\neq r}^{R} g_\ell x_\ell \right\}$$

Using equations (D4.2) and (D4.3), yields

$$P_r(0) = 1-\rho_r-(\rho-\rho_r)x_r$$

ii/ For $n_r>0$.

Using the law of total probability, the marginal probabilities are given by

$$P_r(n) = \sum_{\underline{n}=\underline{0}\wedge n_r=n}^{\infty} P(\underline{n}) = \frac{1}{Z} \sum_{\underline{n}=\underline{0}\wedge n_r=n}^{\infty} \prod_{s=1}^{R} x_s^{n_s} \left\{ \sum_{\ell=1\wedge n_\ell\neq 0}^{R} g_\ell \right\}$$

$$= \frac{1}{Z} \left\{ \left[\sum_{\ell=1\wedge \ell\neq r}^{R} g_\ell x_\ell + g_r \right] \prod_{s=1\wedge s\neq r}^{R} \frac{1}{1-x_s} \right\} x_r^{n_r}$$

Using then equations (D4.2) and (D4.3), leads to

$$P_r(n) = [\rho_r + (\rho-\rho_r)x_r](1-x_r)x_r^{n-1} \quad , \text{ for } n > 0$$

Q.E.D.

D6: Proof of corollary 5.4

i/For $\underline{n}=\underline{1}_r$.

From the ME solution (5.12), it is clear that

$$P(\underline{1}_r) = \frac{1}{Z} g_r x_r \text{ and } P(\underline{0}) = \frac{1}{Z}$$

Which leads to $P(\underline{1}_r) = g_r x_r P(\underline{0})$

ii/For $\underline{n}\neq\underline{1}_r$ and $n_r=1$.

Given that the ME solution (eq.5.12) can also be written in the following form:

$$P(\underline{n}) = \frac{1}{Z} x_r \prod_{s=1\wedge s\neq r}^{R} x_s^{n_s} \left[\sum_{\substack{\ell=1\wedge\ell\neq r \\ \wedge \\ n_\ell>0}}^{R} g_\ell + g_r \right] ,$$

and since $n_r=1$, the r^{th} entry of the vector $\underline{n}-\underline{1}_r$ is zero, thus expressing $P(\underline{n}-\underline{1}_r)$ as above, The second recursive relation of (eq.5.16) follows.

iii/For $n_r>1$.

$$\text{we have } P(\underline{n}) = \frac{1}{Z}\left[\sum_{\ell=1\wedge n_\ell\neq 0}^{R} g_\ell\right] x_r^{n_r}\prod_{s=1\wedge s\neq r}^{R} x_s^{n_s}$$

$$\text{and} \qquad P(\underline{n}-\underline{1}_r) = \frac{1}{Z}\left[\sum_{\ell=1\wedge n_\ell\neq 0}^{R} g_\ell\right] x_r^{n_r-1}\prod_{s=1\wedge s\neq r}^{R} x_s^{n_s}$$

Clearly the two above equations are related by :

$$P(\underline{n}) = x_r P(\underline{n}-\underline{1}_r)$$

Q.E.D.

D7: Proof of theorem 5.3

Using the ME solution (5.18), and proceeding as in the case 1 (see appendix D1), we obtain the following expressions for $P(\underline{0})$ and ρ_r:

$$P(\underline{0}) = \frac{1}{Z} = 1-\rho \tag{D7.1}$$

$$\rho_r = \frac{1}{Z}\, g_r y_r \frac{x_r}{1-x_r}\prod_{\ell=r+1}^{R}\left[1 + y_\ell\frac{x_\ell}{1-x_\ell}\right] \tag{D7.2}$$

Let us first express $\{\theta_r\}$ with respect to the Lagrangian coefficients $\{x_r\}$, $\{g_r\}$ and $\{y_r\}$.

The idle state probability constraint (eq. 5.17) can be clearly written as

$$\theta_r = \sum_{\underline{n}=\underline{1}_r}^{\infty} P(\underline{n})$$

Using the ME solution (5.18), we will have

$$\theta_r = \frac{1}{Z}\sum_{n_R=0}^{\infty} g_R^{h_R(\underline{S})} y_R^{V_R(\underline{S})} x_R^{n_R} \cdots \sum_{n_r=1}^{\infty} g_r^{h_r(\underline{S})} y_r x_r^{n_r} \cdots \sum_{n_1=0}^{\infty} g_1^{h_1(\underline{S})} y_1^{V_1(\underline{S})} x_1^{n_1}$$

$$= \frac{1}{Z}\sum_{n_R=0}^{\infty} g_R^{h_R(\underline{S})} y_R^{V_R(\underline{S})} x_R^{n_R} \cdots \sum_{n_r=1}^{\infty} g_r^{h_r(\underline{S})} y_r x_r^{n_r} \cdots \left[1+g_1 y_1 \frac{x_1}{1-x_1}\right]$$

$$= \frac{1}{Z}\sum_{n_R=0}^{\infty} g_R^{h_R(\underline{S})} y_R^{V_R(\underline{S})} x_R^{n_R} \cdots \sum_{n_r=1}^{\infty} g_r^{h_r(\underline{S})} y_r x_r^{n_r} \cdots \sum_{n_2=0}^{\infty} g_2^{h_2(\underline{S})} y_2^{V_2(\underline{S})} x_2^{n_2}$$

$$+ \frac{1}{Z} g_1 y_1 \frac{x_1}{1-x_1} \prod_{\ell=2 \wedge \ell \neq r}^{R} \left[1+y_\ell \frac{x_\ell}{1-x_\ell}\right] \frac{y_r x_r}{1-x_r}$$

Thus, recursively we end up with the following expression:

$$\theta_r = \frac{1}{Z} \frac{y_r x_r}{1-x_r} \left\{ g_r \prod_{\ell=r+1}^{R} \left[1+y_\ell \frac{x_\ell}{1-x_\ell}\right] + \sum_{s=1}^{r-1} g_s y_s \frac{x_s}{1-x_s} \prod_{\substack{\ell=s+1 \\ \wedge \ell \neq r}}^{R} \left[1+y_\ell \frac{x_\ell}{1-x_\ell}\right] \right\}$$

Using the expression (D7.2) of ρ_r, θ_r is eventually expressed by:

$$\theta_r = \rho_r + y_r \frac{x_r}{1-x_r} \left\{ \frac{\sum_{s=1}^{r-1} \rho_s}{1+y_r \frac{x_r}{1-x_r}} \right\} \qquad \text{(D7.3)}$$

Solving the above equation with respect to y_r together with the introduction of γ_r given by equation (2.4), leads to equation (5.22).

The Lagrangian coefficients $\{x_r\}$ are obtained by ,first expressing the marginal mean queue length $\{\langle n_r\rangle\}$ with respect to $\{g_r\}$, $\{x_r\}$ and $\{y_r\}$ in equation (5.3) and then performing all the infinite summations leads to

$$\langle n_r\rangle = \frac{1}{Z} g_r y_r \frac{x_r}{(1-x_r)^2} \prod_{\ell=r+1}^{R} \left[1+y_\ell \frac{x_\ell}{1-x_\ell}\right]$$

$$+ \frac{1}{Z} \sum_{s=1}^{r-1} g_s y_s \frac{x_s}{1-x_s} \prod_{\ell=s+1 \wedge \ell\neq r}^{R} \left[1+y_\ell \frac{x_\ell}{1-x_\ell}\right]$$

Substituting the expression of θ_r (D7.3) in the equation above, yields

$$\langle n_r\rangle = \frac{\theta_r}{1 - x_r}$$

Solving the equation above with respect to x_r, equation (5.20) follows.

The Lagrangian coefficients $\{g_r\}$ are the solutions of the equations (D7.2), with the expression (5.22) of $\{y_r\}$ substituted as appropriate in the final solution.

Q.E.D.

D8: Proof of theorem 5.5

Class-r jobs, r= 1,2,...,R are assumed to arrive from an external source according to GE distribution with λ_r and Ca_r^2 as the corresponding mean arrival rate and squared coefficient of variation, respectively. Thus, using theorem 3.1, the arrival process of class-r jobs is a compound Poisson process with mean bulk arrival rate $\lambda_r^{(b)}$, and mean bulk size <b> given by

$$\lambda_r^{(b)} = \frac{2\lambda_r}{Ca_r^2+1} \quad , \quad <b> = \frac{1}{\sigma_r} = \frac{C\hat{a}_r^2+1}{2}$$

Subsequently, the r^{th} virtual queue $GE(\lambda_r, C_{\hat{a}r}^2)/G/1$ is equ ivalent to an ordinary $M^B(\lambda_r^{(b)}, \sigma_r)/G/1$ queue with $\hat{S}_r$ as the service time distribution that we want to determine.

It is known [KLEI 75,pp.235] that the generating function, $P_r(z)$, of the queue length distribution of a stable $M^B/G/1$ single class queue is given by the generalised Pollaczek-Khinchin transform of the form

$$P_r(z) = \hat{S}_r^*(\lambda_r^{(b)}-\lambda_r^{(b)}q_r(z))\frac{(1-\hat{\rho}_r)(1-z)}{\hat{S}_r^*(\lambda_r^{(b)}-\lambda_r^{(b)}q_r(z)) - z} \qquad |z|<1 \quad (D8.1)$$

Where $\hat{S}_r^*(.)$ is the L.S.T of the effective service-time, $\hat{S}_r$, $q_r(.)$ is the Z-transform of the bulk size , and $\hat{\rho}_r=1-P_r(0)$ is the 'perceived' utilisation of the r^{th} virtual, r=1,2,...,R.

Since the bulk size is geomitrically distributed with mean bulk size $(1/\sigma)$, the Z-transform, $q_r(.)$ is given by [FELL,68]

$$q_r(z) = \frac{\sigma_r z}{1 - (1-\sigma_r)z} \qquad r=1,2,...,R, \ |z|<1 \quad (D8.2)$$

On the other hand, the (ME1) and (ME2) marginal queue length distributions (5.10), (5.15), and (5.23) are all of modified geometric-type of the form

$$P_r(n_r) = \begin{cases} 1-\hat{\rho}_r \\ \\ \hat{\rho}_r(1-x_r)x_r^{n_r-1} \end{cases} \quad (D8.3)$$

$$\text{Thus,}\quad P_r^{(ME)}(z) = \sum_{n_r=0}^{\infty} P_r(n_r)z^n = 1-\rho_r+\sum_{n_r=1}^{\infty}\rho_r(1-x_r)x_r^{n_r-1}z^{n_r} \qquad |z|<1$$

Given that $x_r<1$, the generating function above can simply be expressed by

$$P_r^{(ME)}(z) = (1-\hat{\rho}_r)\left[1+\hat{\delta}_r\frac{x_r z}{1-x_r z}\right]$$

Because the ME marginal probabilities must coincide with the steady state probabilities of an ordinary single-class GE/G/1 queue with service-time, $\hat{S}_r$, we must have then $P_r^{(ME)}(z) = P_r(z)$.

Using the above corresponding expressions of $P_r^{(ME)}(.)$ and $P_r(.)$ and solving with repect to $\hat{S}_r^*(.)$, we obtain

$$\hat{S}_r^*(\lambda_r^{(b)} - \lambda_r^{(b)}q_r(z)) = \frac{1-x_r(1-\hat{\delta}_r)}{1+\hat{\delta}_r x_r - x_r z} \qquad \text{(D8.4)}$$

Setting $\Psi = \lambda_r^{(b)} - \lambda_r^{(b)}q_r(z)$, together with the use of (D8.2) and then solving equation (D8.4) with repect to z, yields

$$z = \frac{\lambda_r^{(b)} - \Psi}{\lambda_r^{(b)} - \Psi(1-\sigma_r)}$$

Substituting the above expression of z in equation (D8.4), this latter becomes

$$\hat{S}_r^*(\Psi) = \frac{\dfrac{\lambda_r^{(b)}(1+\hat{\delta}_r x_r - x_r)}{x_r-(1-\sigma_r)(1+\hat{\delta}_r x_r)} + \dfrac{x_r(1-\hat{\delta}_r)-(1-\sigma_r)}{x_r-(1-\sigma_r)(1+\hat{\delta}_r x_r)}}{\dfrac{\lambda_r^{(b)}[1+\hat{\delta}_r x_r - x_r]}{x_r-(1-\sigma_r)(1+\hat{\delta}_r x_r)} + \Psi}$$

where after simple manipulations, we obtain the following known L.S.T of the effective service time of class-r jobs ,r=1,2...R.

$$\hat{S}_r^*(\Psi) = 1-\hat{\tau}_r + \hat{\tau}_r \frac{\hat{\tau}_r\hat{\mu}_r}{\hat{\tau}_r\hat{\mu}_r + \Psi} \qquad \text{(D8.5)}$$

Where $\hat{\tau}_r$ and $\hat{\mu}_r$, are given by (eq. 5.31) and (eq.5.32),respectively.

Finally inverting (D8.5), equation (5.30) follows.

Q.E.D.

D9: Numerical results (chapter 5)

Example 5.1 M/M/1 PR queue (4 Classes)

Table 5.1: Raw data for PR M/M/1 queue

(Figs. 5.1a - 5.1c)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1	1	4	1
2	1	1	10	1
3	0.5	1	2	1
4	0.6	1	6	1

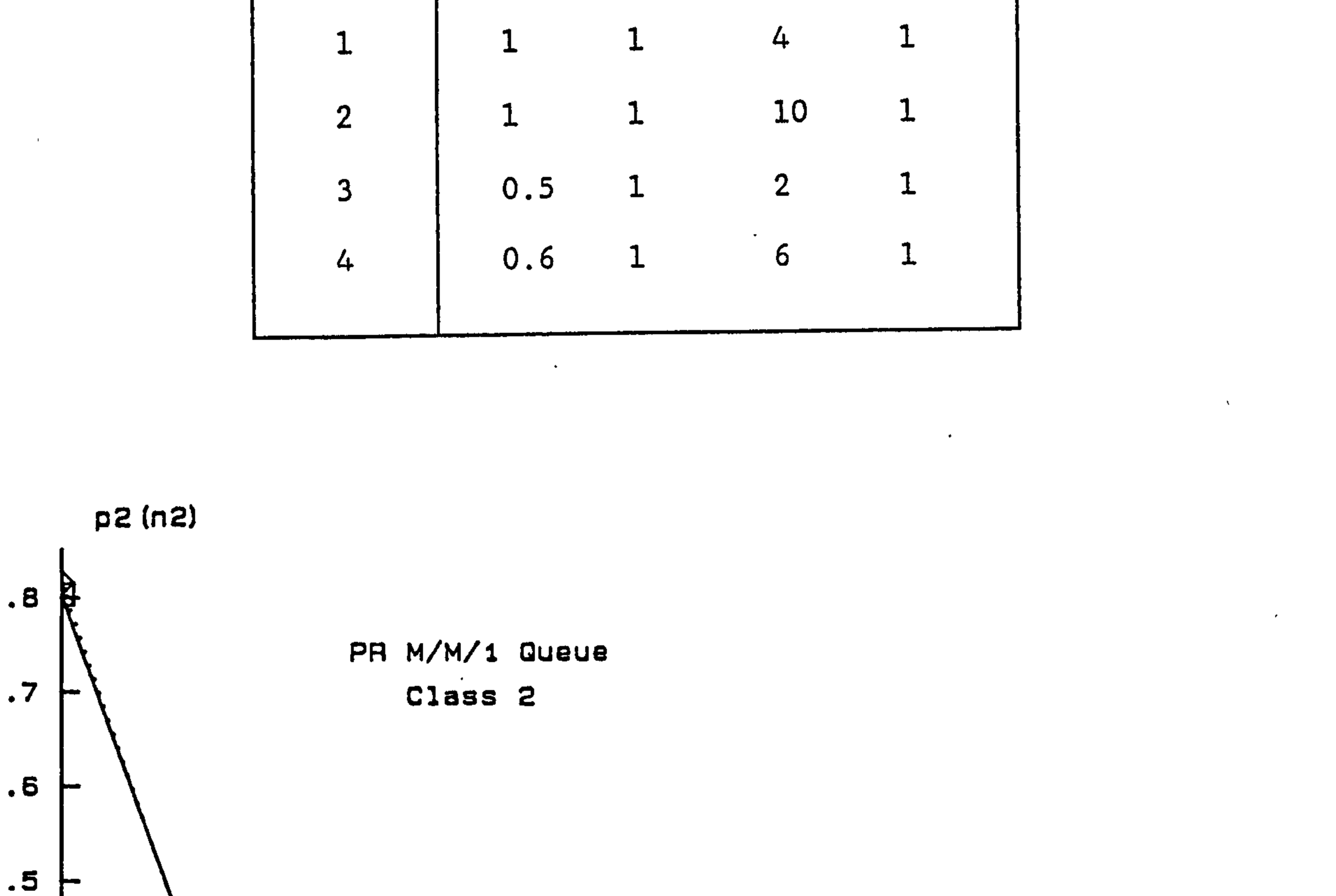

Fig. 5.1a. PR M/M/1 P2(n2) vs n2 Class 2, Table 5.1)

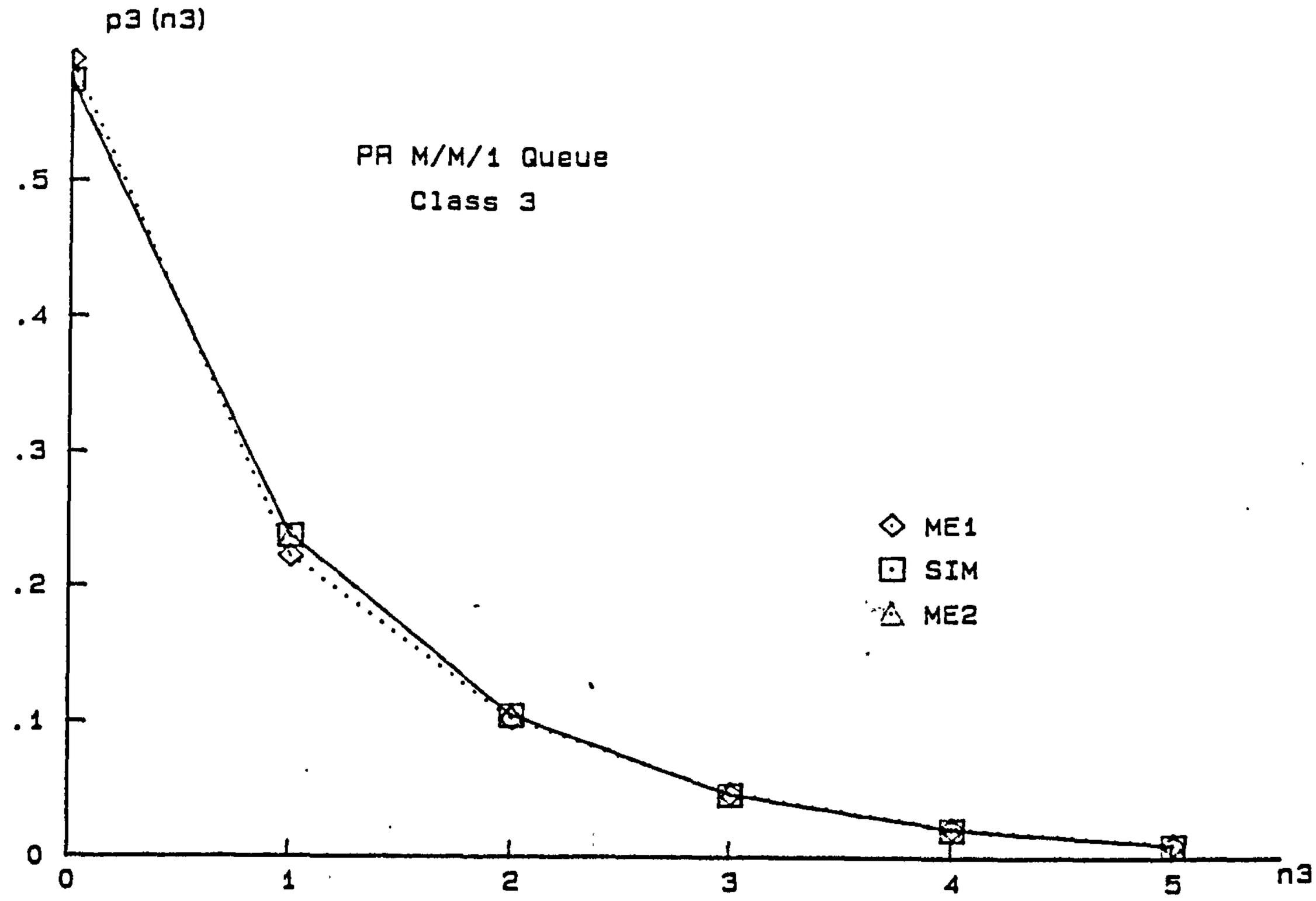

Fig. 5.1b. PR M/M/1 P3(n3) vs n3 (Class 3, Table 5.1)

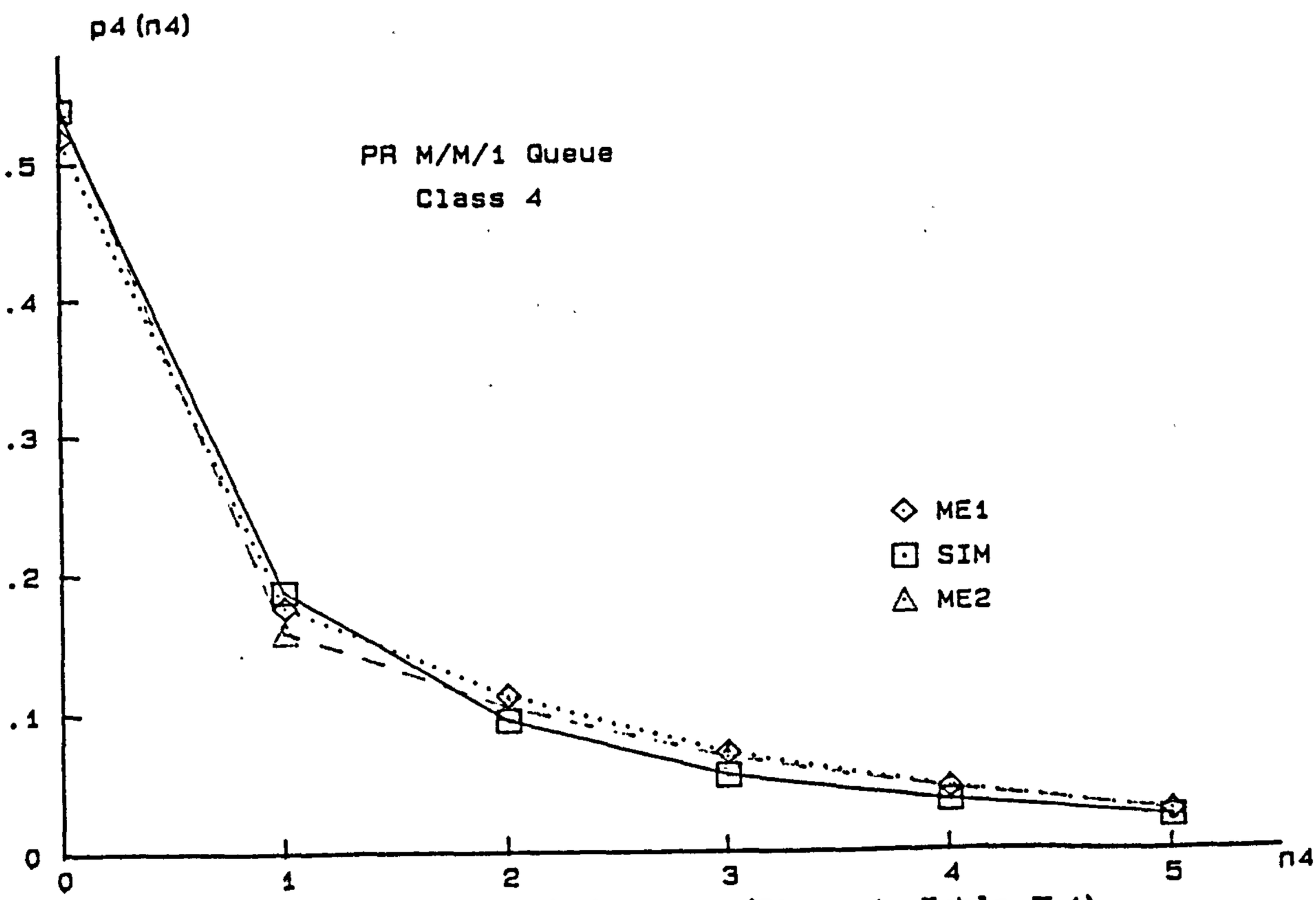

Fig. 5.1c. PR M/M/1 P4(n4) vs n4 (Class 4, Table 5.1)

Example 5.2 $E_2/E_2/1$ PR queue (4 Classes)

Table 5.2: Raw data for PR $E_2/E_2/1$ queue

(Figs. 5.2a - 5.2c)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	0.45	0.5	0.9	0.5
2	0.5	0.5	5	0.5
3	0.25	0.5	5	0.5
4	1.4	0.5	7	0.5

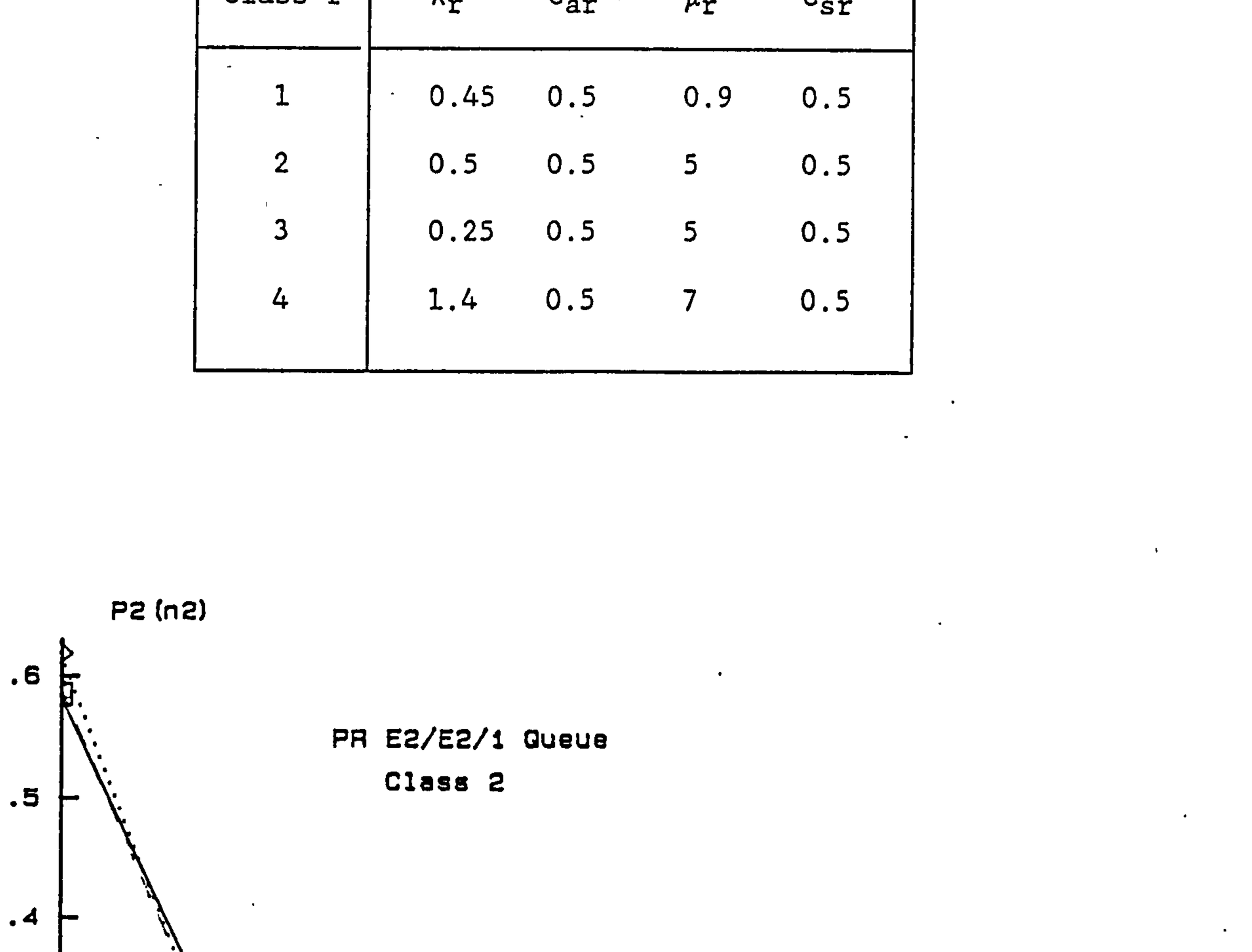

Fig. 5.2a. PR E2/E2/1 P2(n2) vs n2 (Class 2, Table 5.2)

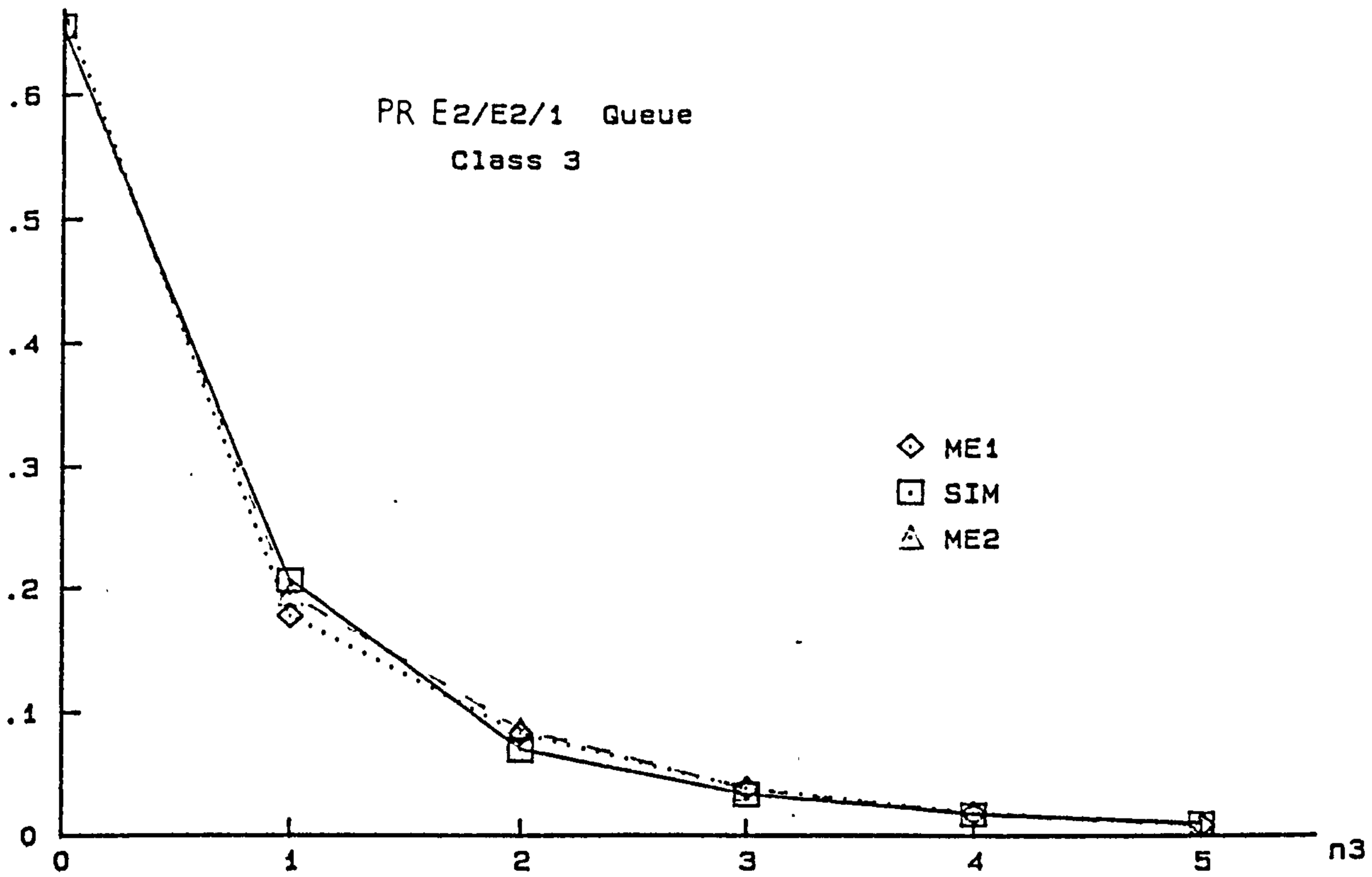

Fig. 5.2b. PR E2/E2/1 P3(n3) vs n3 (Class 3, Table 5.2)

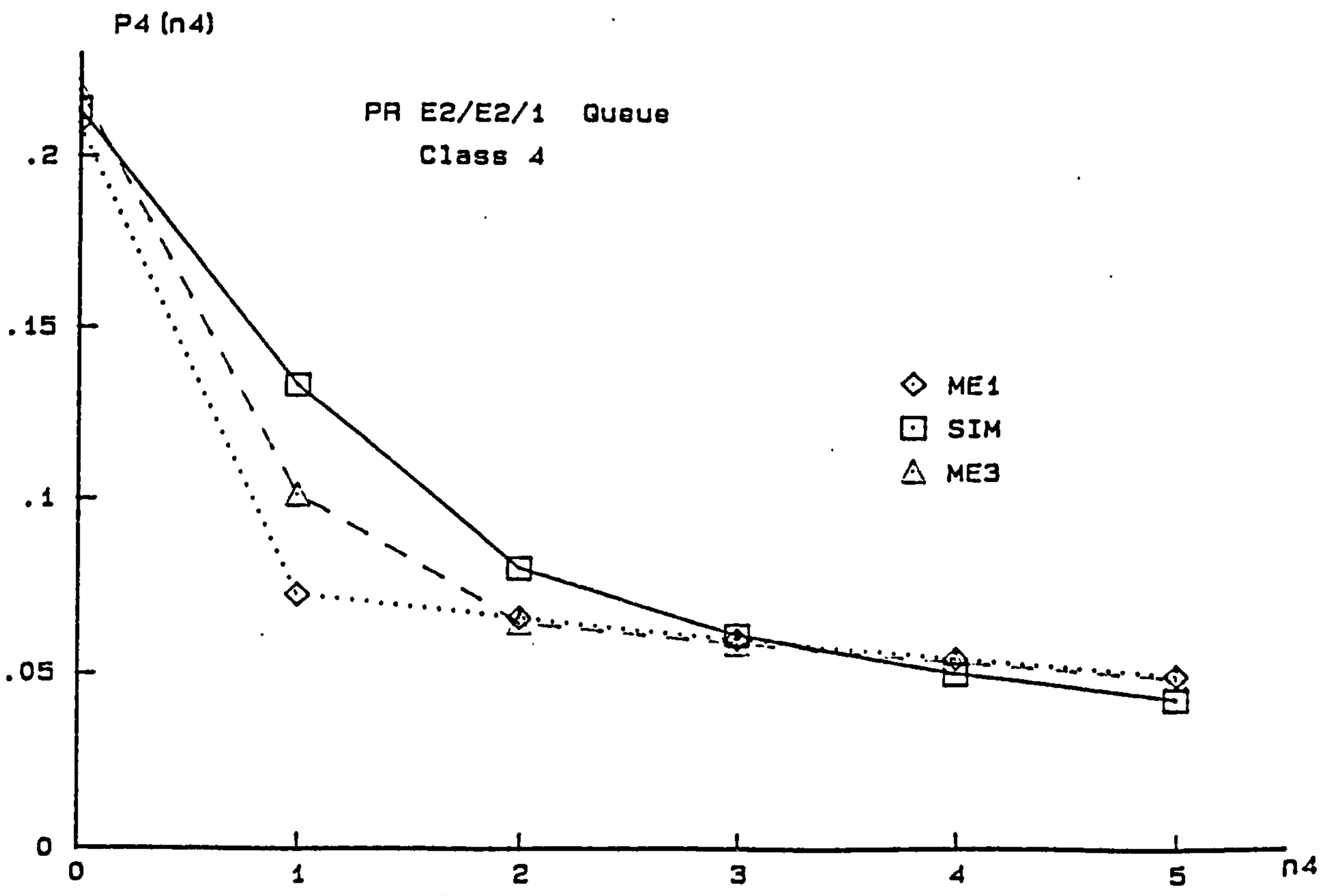

Fig. 5.2c. PR E2/E2/1 P4(n4) vs n4 (Class 4, Table 5.2)

Example 5.3 $H_2/H_2/1$ PR queue (4 Classes)

Table 5.3: Raw data for PR $H_2/H_2/1$ queue

(Figs. 5.3a - 5.3c)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1	5	2	3
2	3	7	30	5
3	1	2	10	3.5
4	6	6	30	4

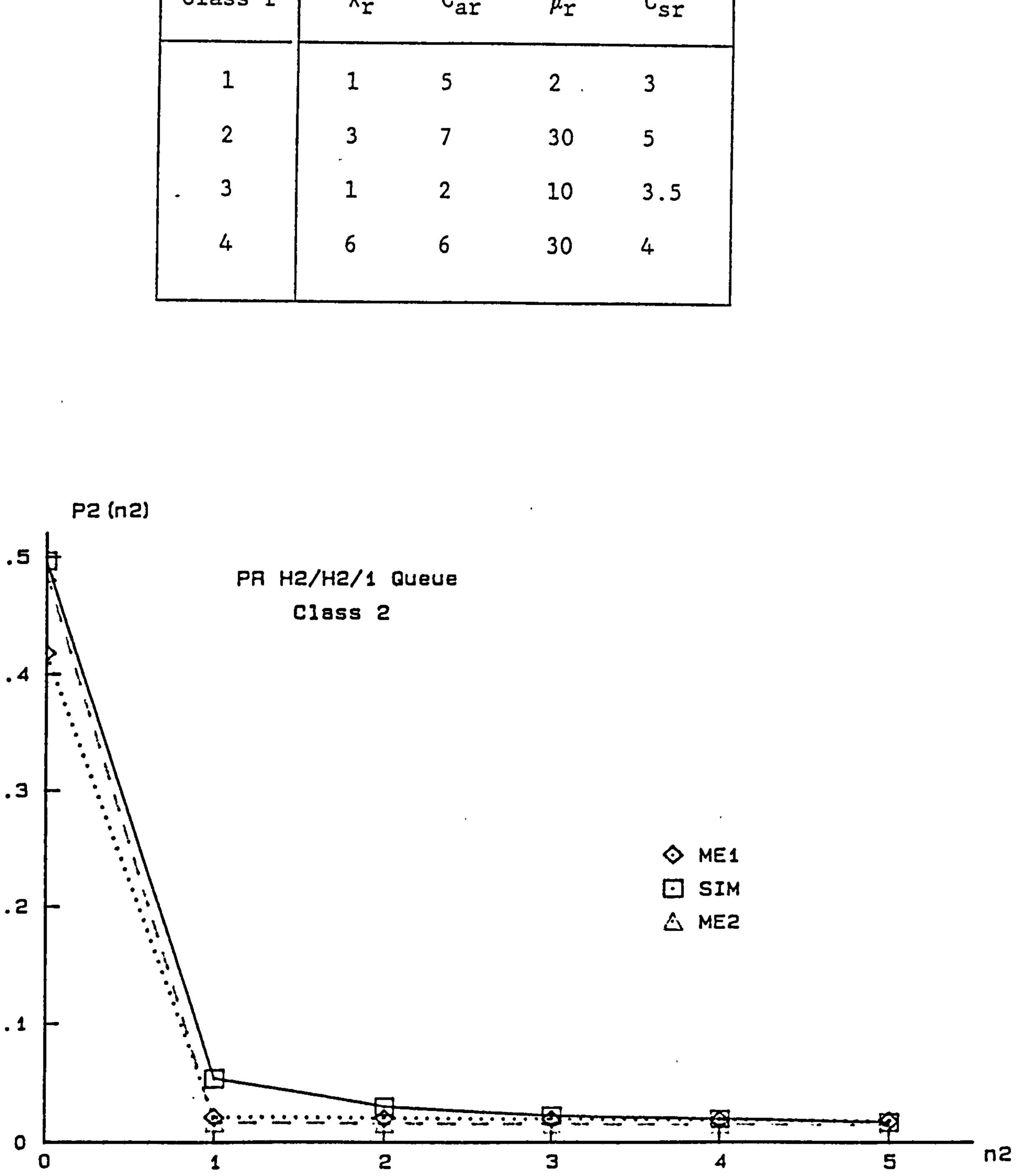

Fig. 5.3a. PR H2/H2/1 P2(n2) vs n2 (Class 2, Table 5.3)

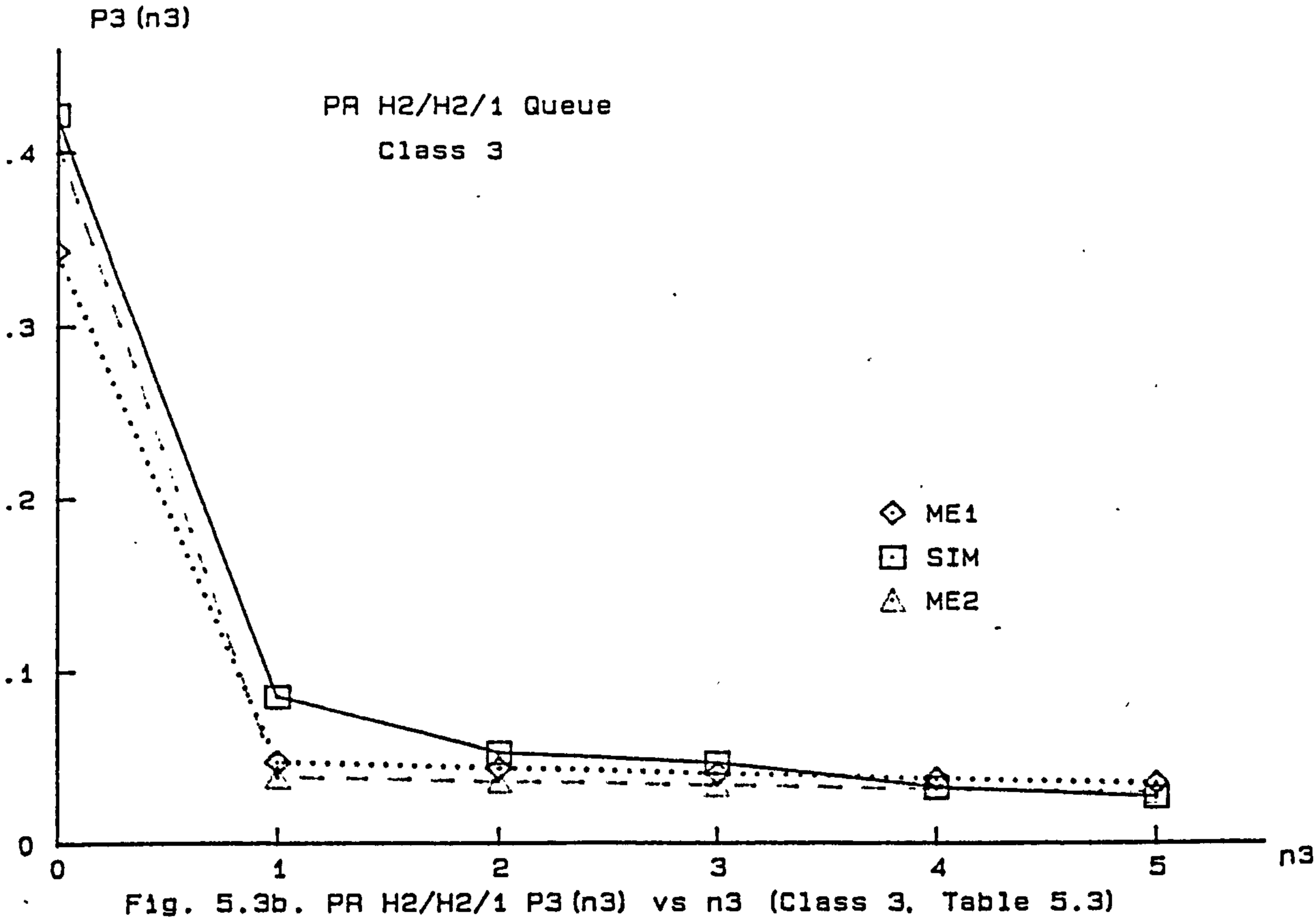

Fig. 5.3b. PR H2/H2/1 P3(n3) vs n3 (Class 3, Table 5.3)

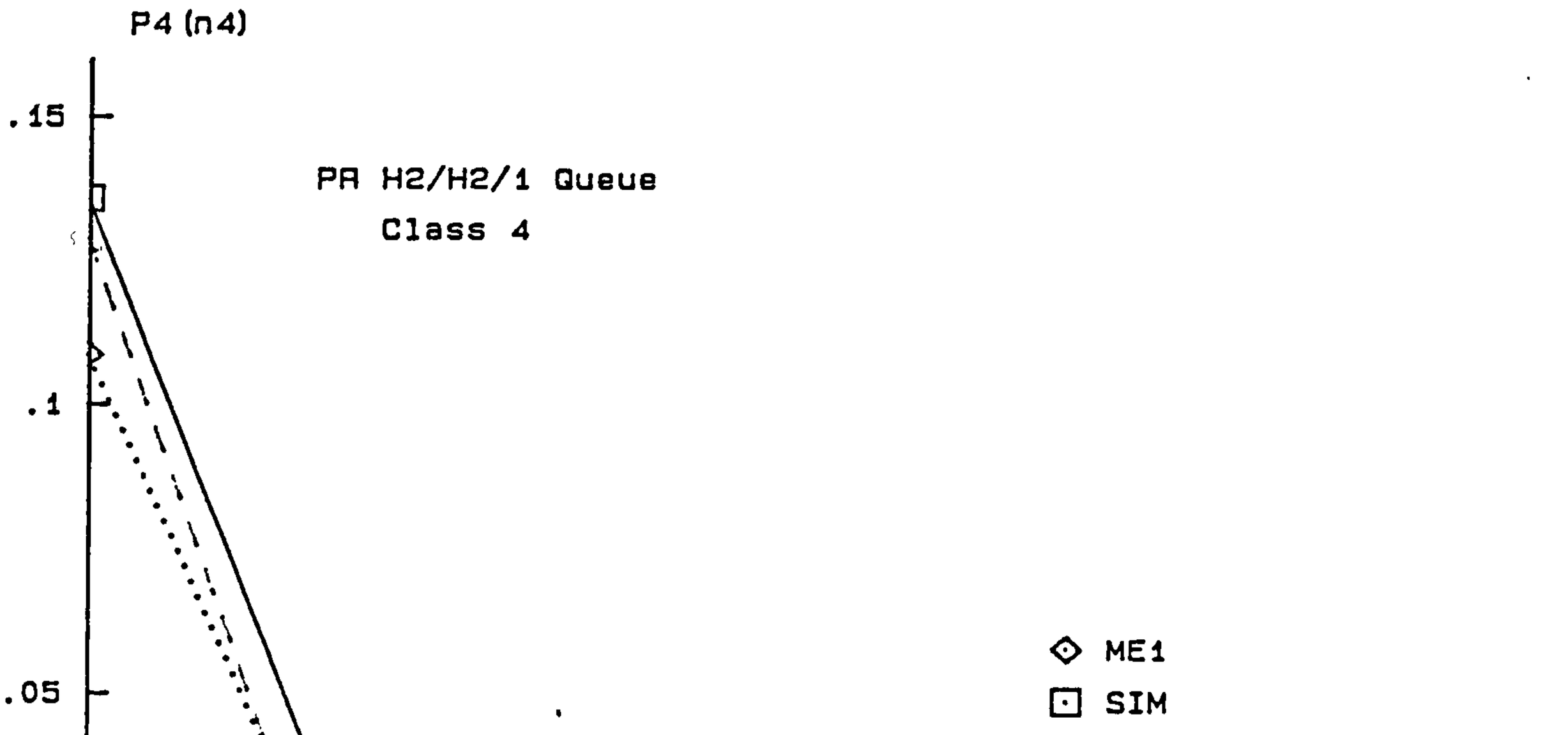

Fig. 5.3c. PR H2/H2/1 P4(n4) vs n4 (Class 4, Table 5.3)

Example 5.4 GE/GE/1 PR queue (4 Classes)

Table 5.4: Raw data for PR GE/GE/1 queue

(Figs. 5.4a - 5.4c)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1	2	4	3
2	1	4	10	2
3	0.5	5	2	2.5
4	0.6	6	6	4

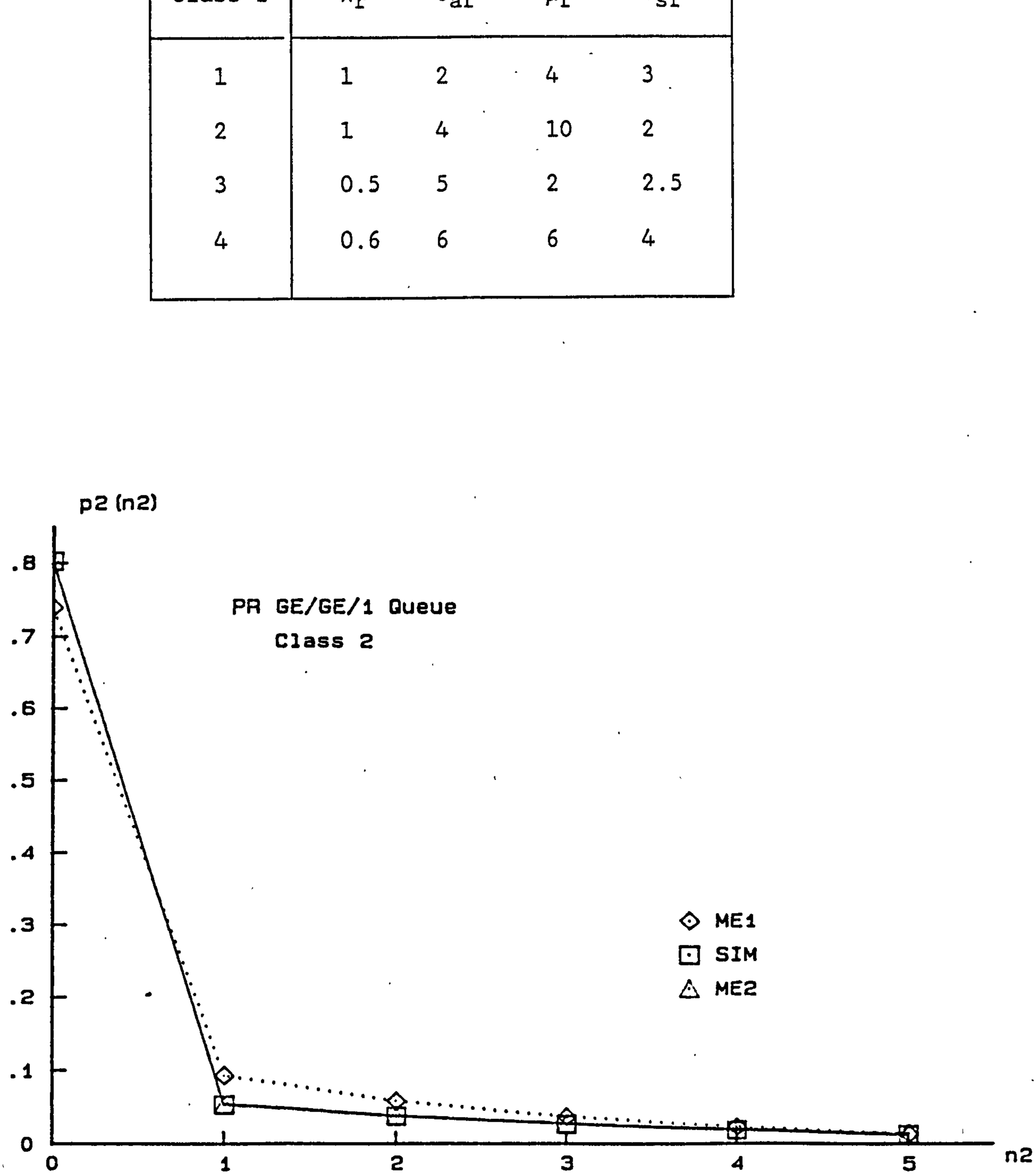

Fig. 5.4a. PR GE/GE/1 P2(n2) vs n2 (class 2, Table 5.4)

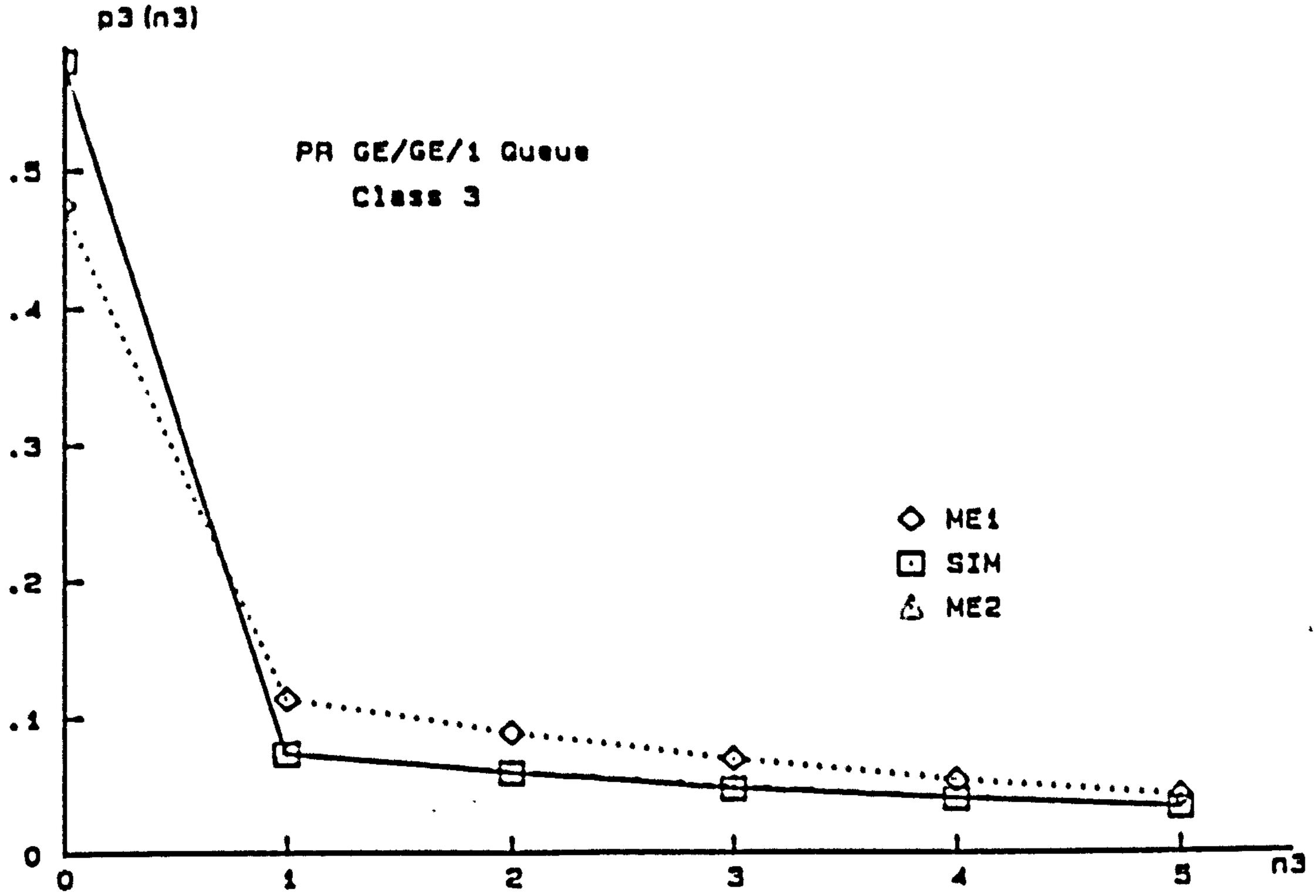

Fig. 5.4b. PR GE/GE/1 P3(n3) vs n3 (class 3, Table 5.4)

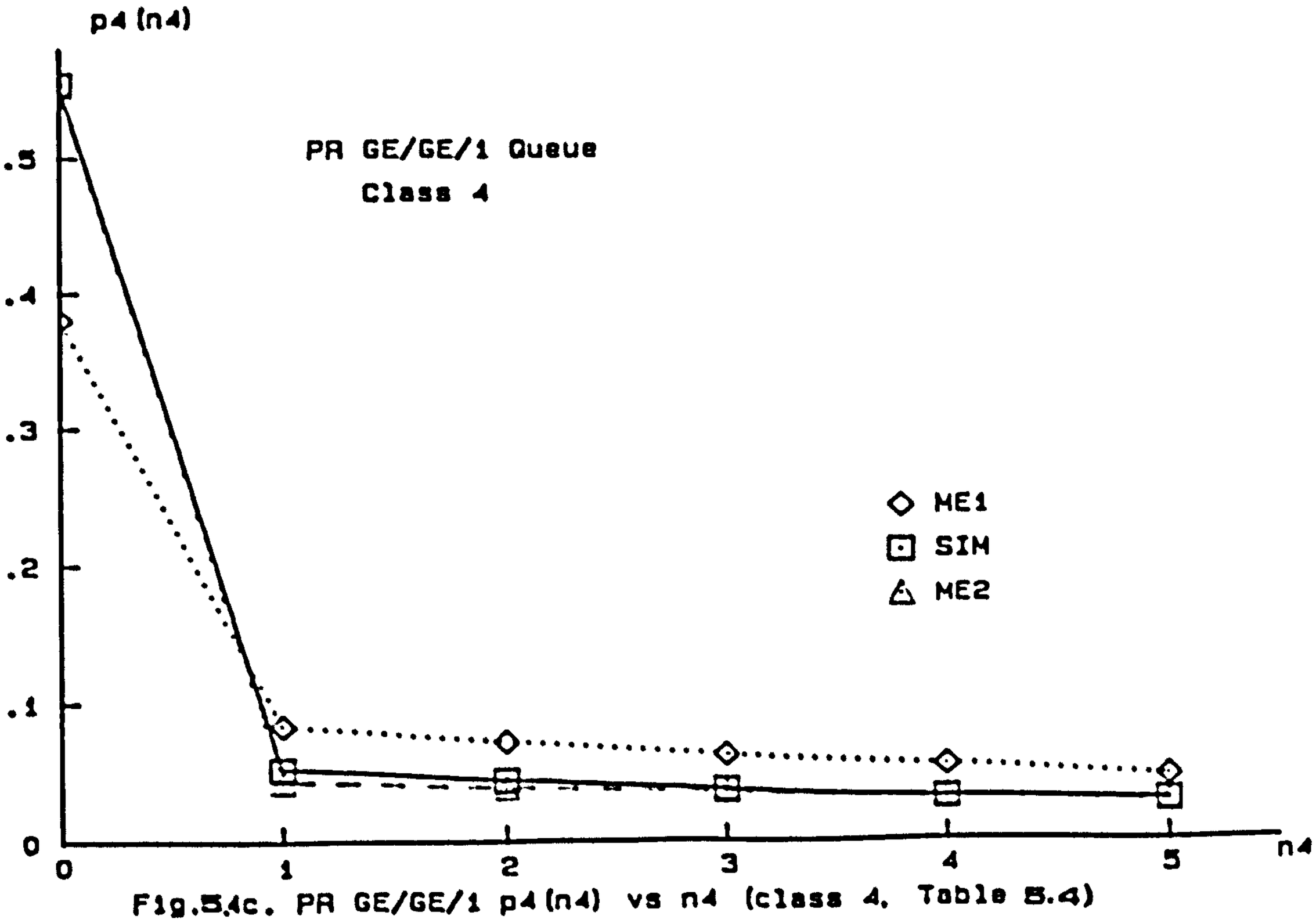

Fig.5.4c. PR GE/GE/1 p4(n4) vs n4 (class 4, Table 5.4)

Example 5.5 E_2/GE/1 PR queue (4 Classes)

Table 5.5: Raw data for PR E_2/GE/1 queue

(Figs. 5.5a - 5.5c)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	2	0.5	5	18
2	4	0.5	20	11
3	1.8	0.5	18	3
4	1	0.5	5	3

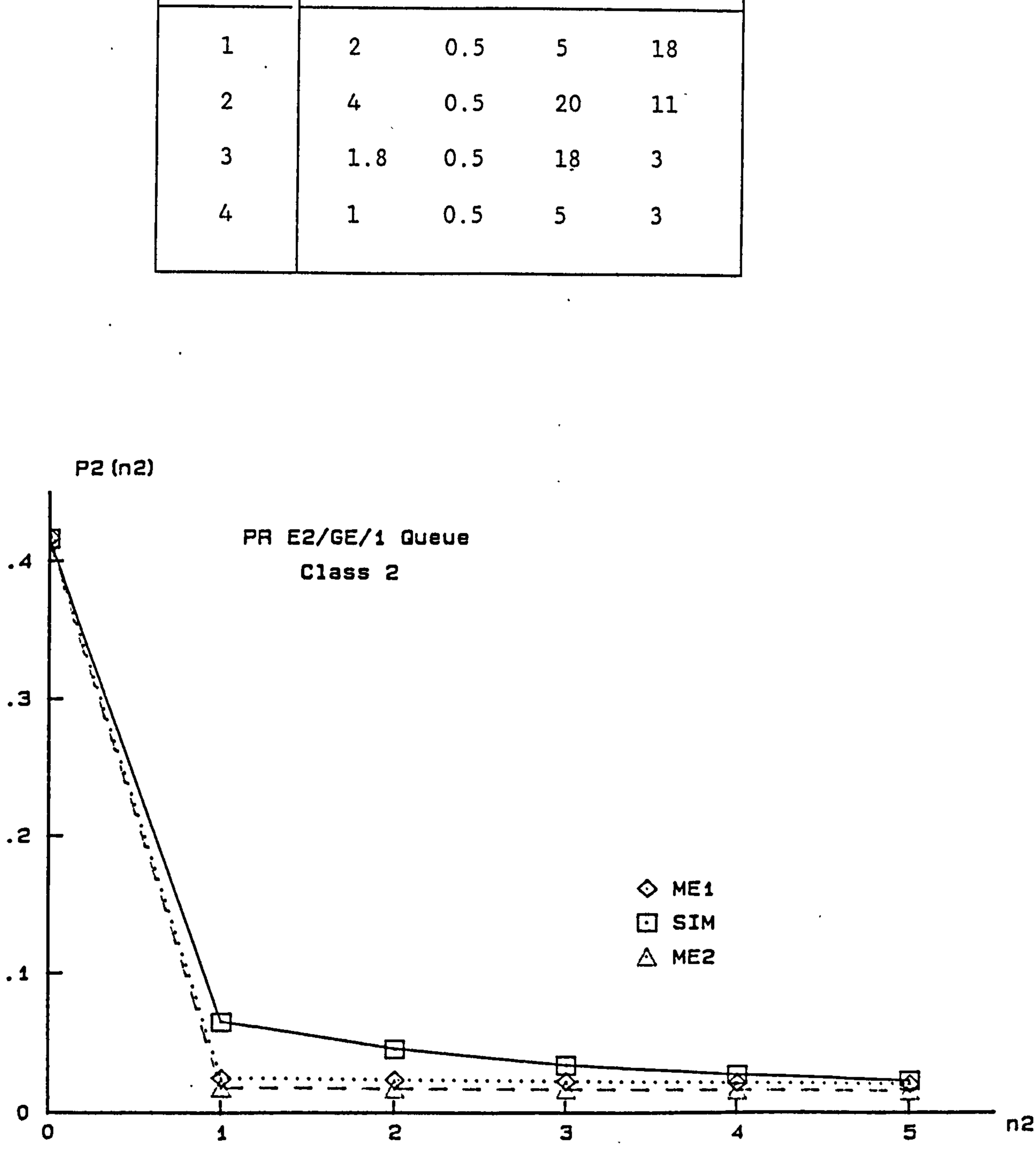

Fig. 5.5a. PR E2/GE/1 P2(n2) vs n2 (Class 2, Table 5.5)

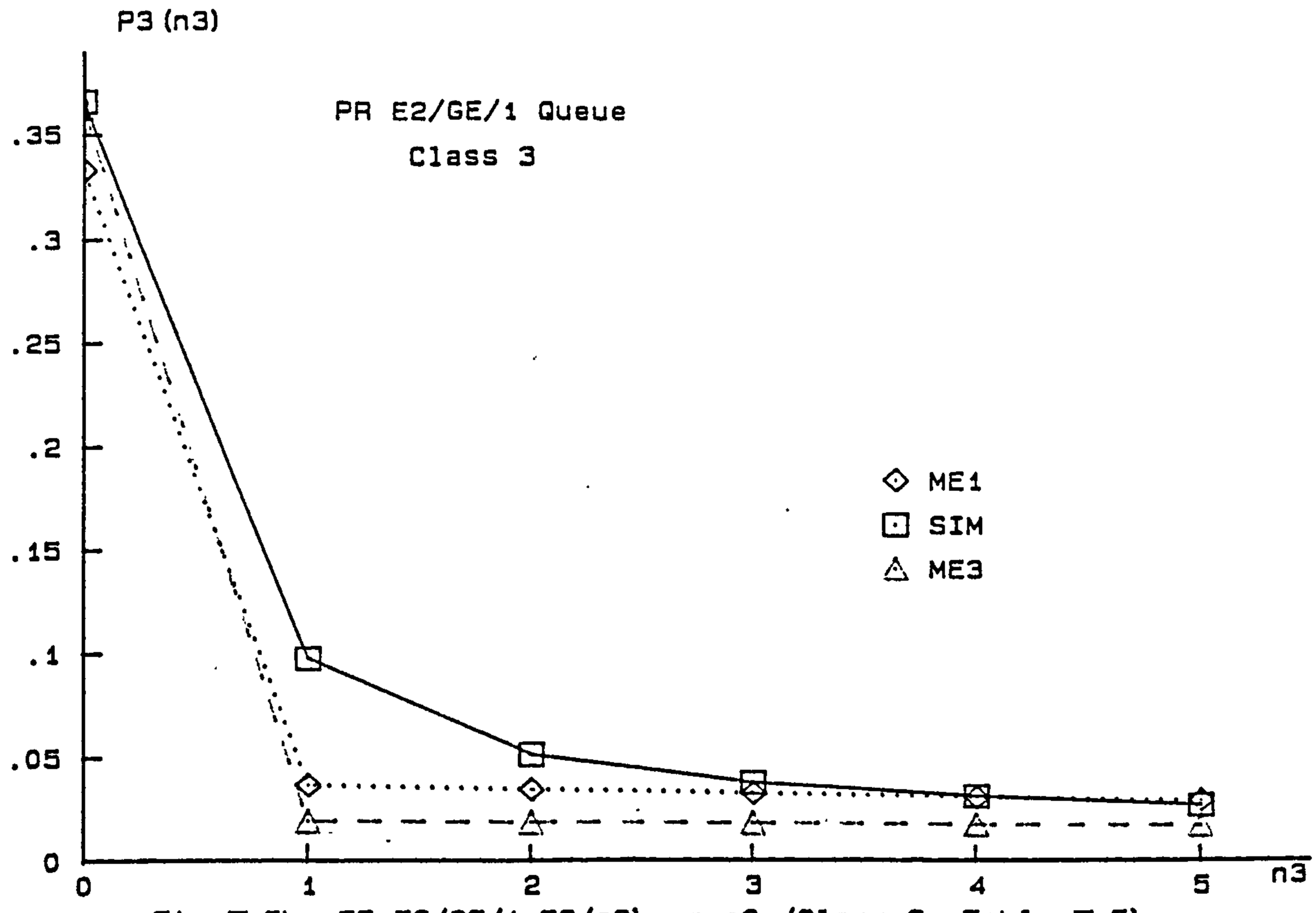

Fig.5.5b. PR E2/GE/1 P3(n3) vs n3 (Class 3, Table 5.5)

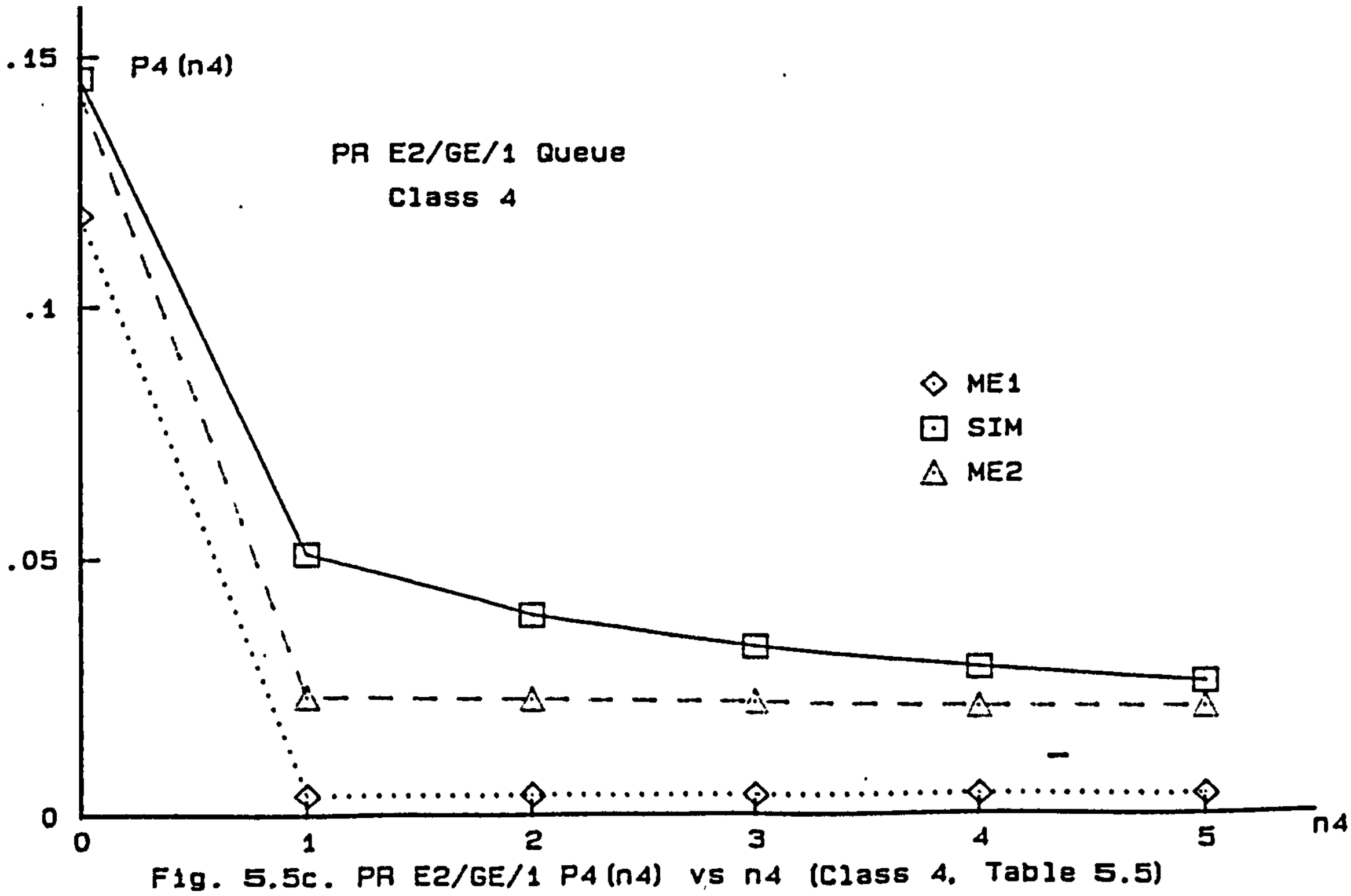

Fig. 5.5c. PR E2/GE/1 P4(n4) vs n4 (Class 4, Table 5.5)

Example 5.6 M/M/1 HOL queue (4 Classes)

Table 5.6: Raw data for HOL M/M/1 queue

(Figs. 5.6a - 5.6d)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	2	1	20	1
2	1.5	1	30	1
3	1	1	15	1
4	1	1	2	1

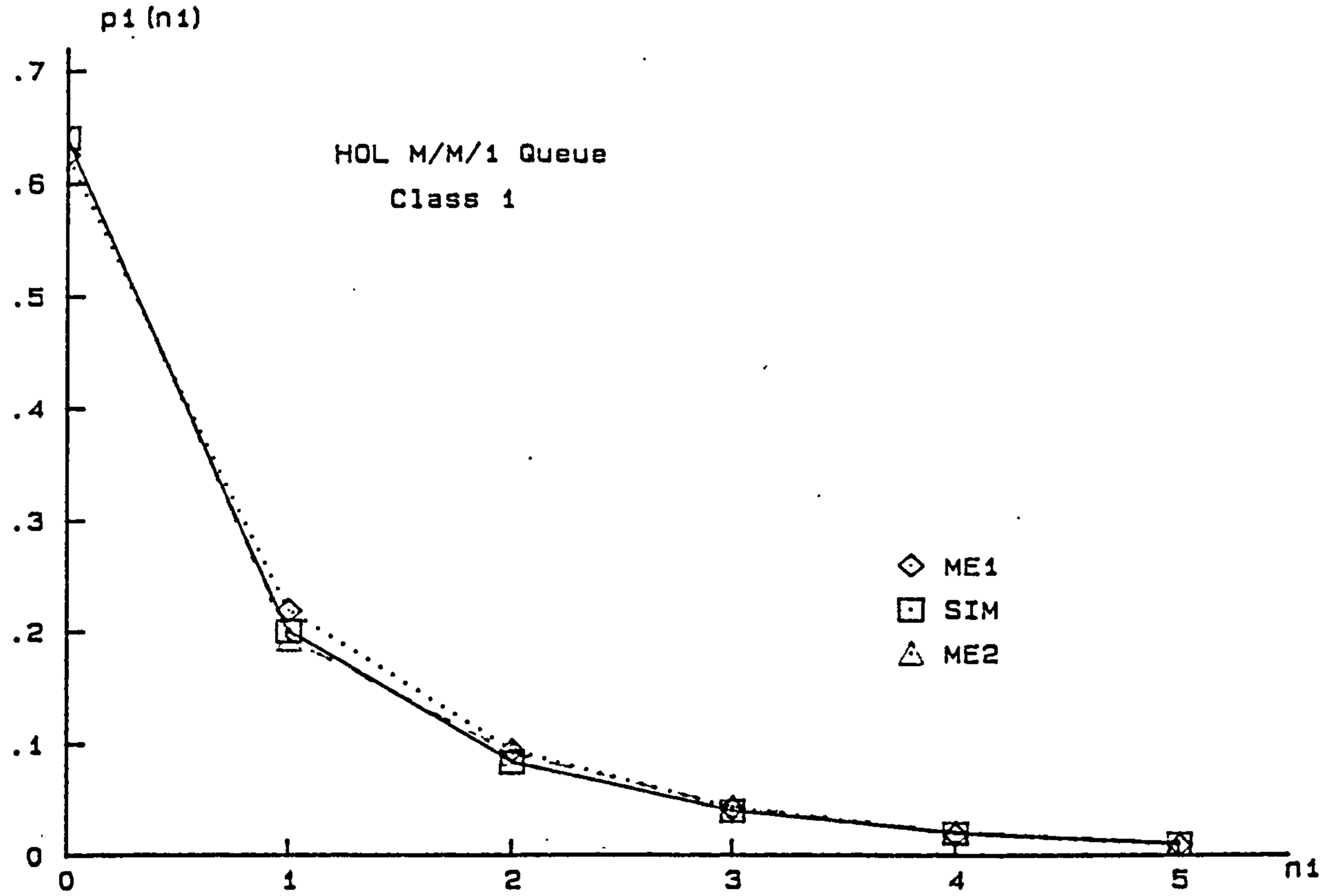

Fig.5.6a. HOL M/M1 p1(n1) vs n1 (class 1, Table 5.6)

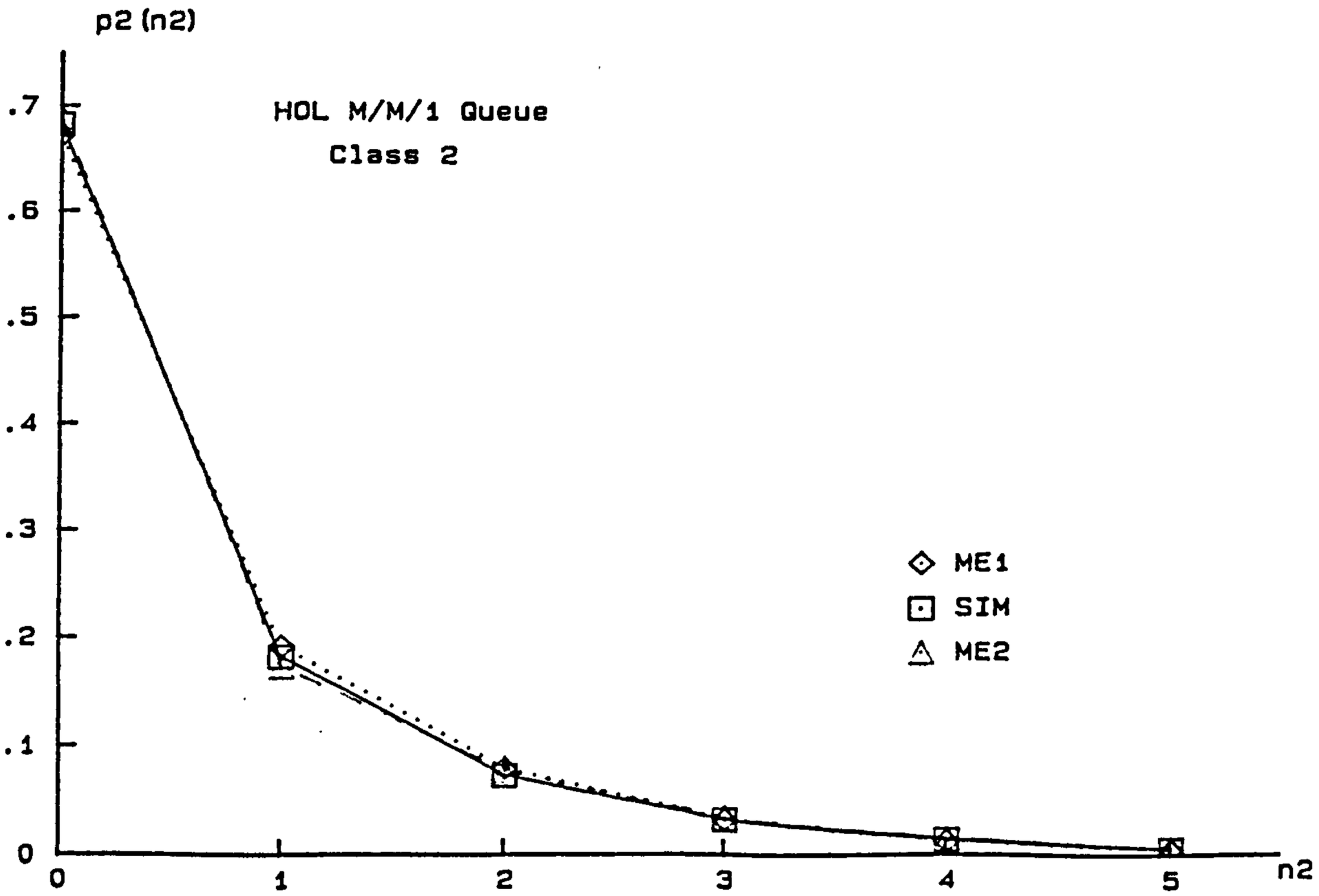

Fig. 5.6b. HOL M/M/1 P2(n2) vs n2 (class 2, Table 5.6)

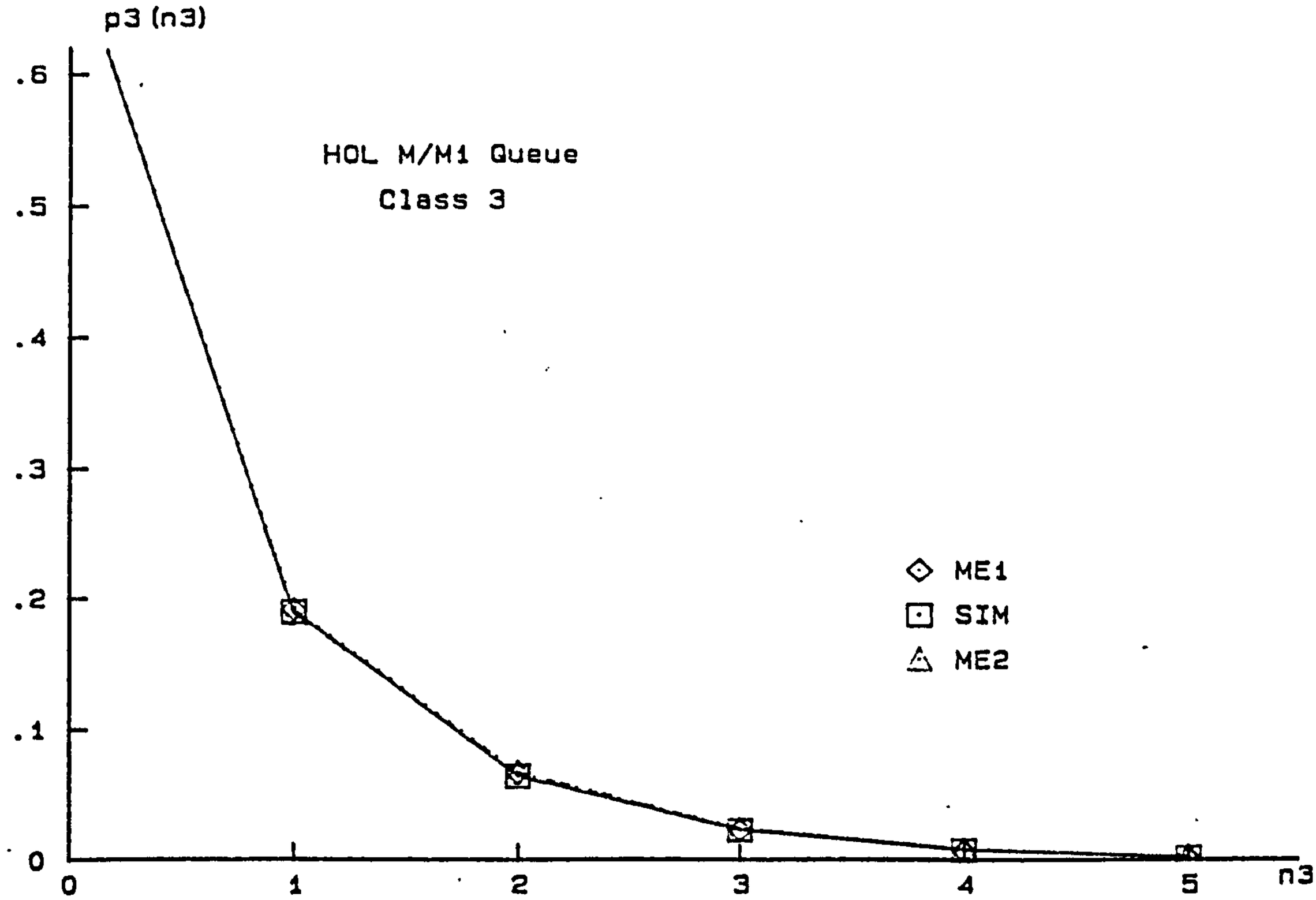

Fig. 5.6c. HOL M/M/1 P3(n3) vs n3 (Class 3, Table 5.6)

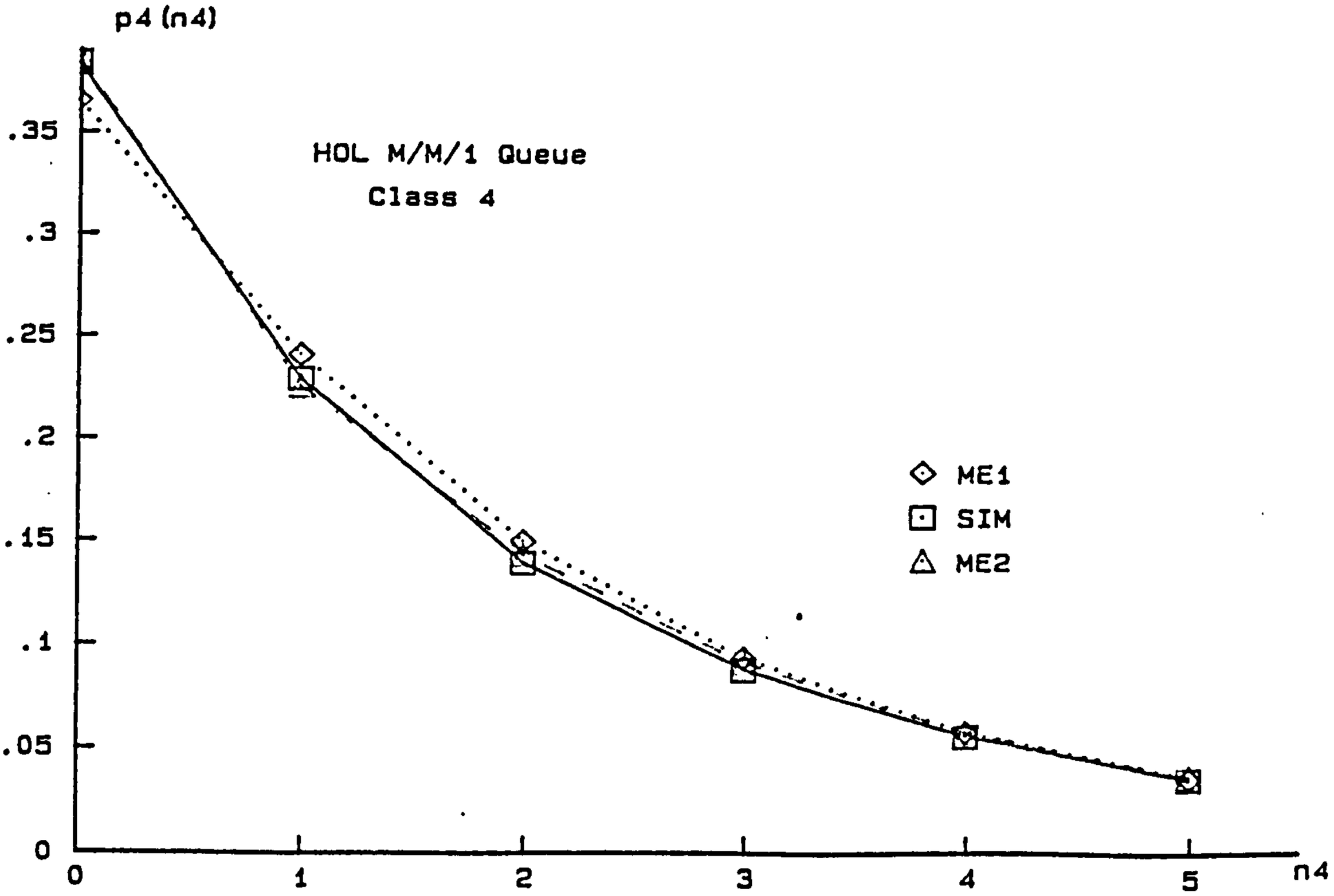

Fig. 5.6d. HOL M/M/1 P4(n4) vs n4 (Class 4, Table 5.6)

Example 5.7 $E_2/E_2/1$ HOL queue (4 Classes)

Table 5.7: Raw data for HOL $E_2/E_2/1$ queue

(Figs. 5.7a - 5.7d)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	3	0.5	4	0.5
2	1	0.5	10	0.5
3	0.05	0.5	2	0.5
4	0.6	0.5	12	0.5

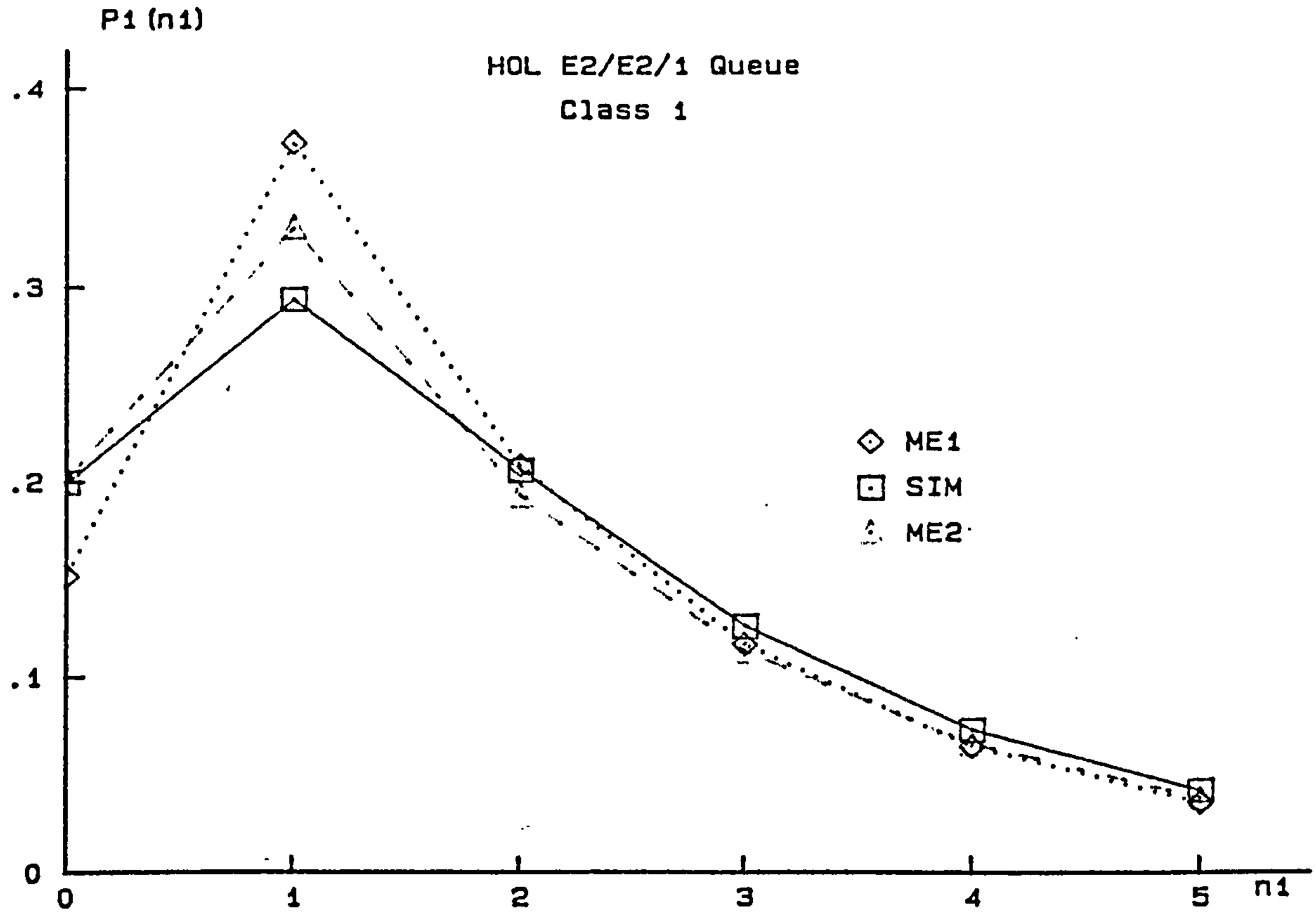

Fig. 5.7a HOL E2/E2/1 P1(n1) vs n1 (class 1, table 5.7)

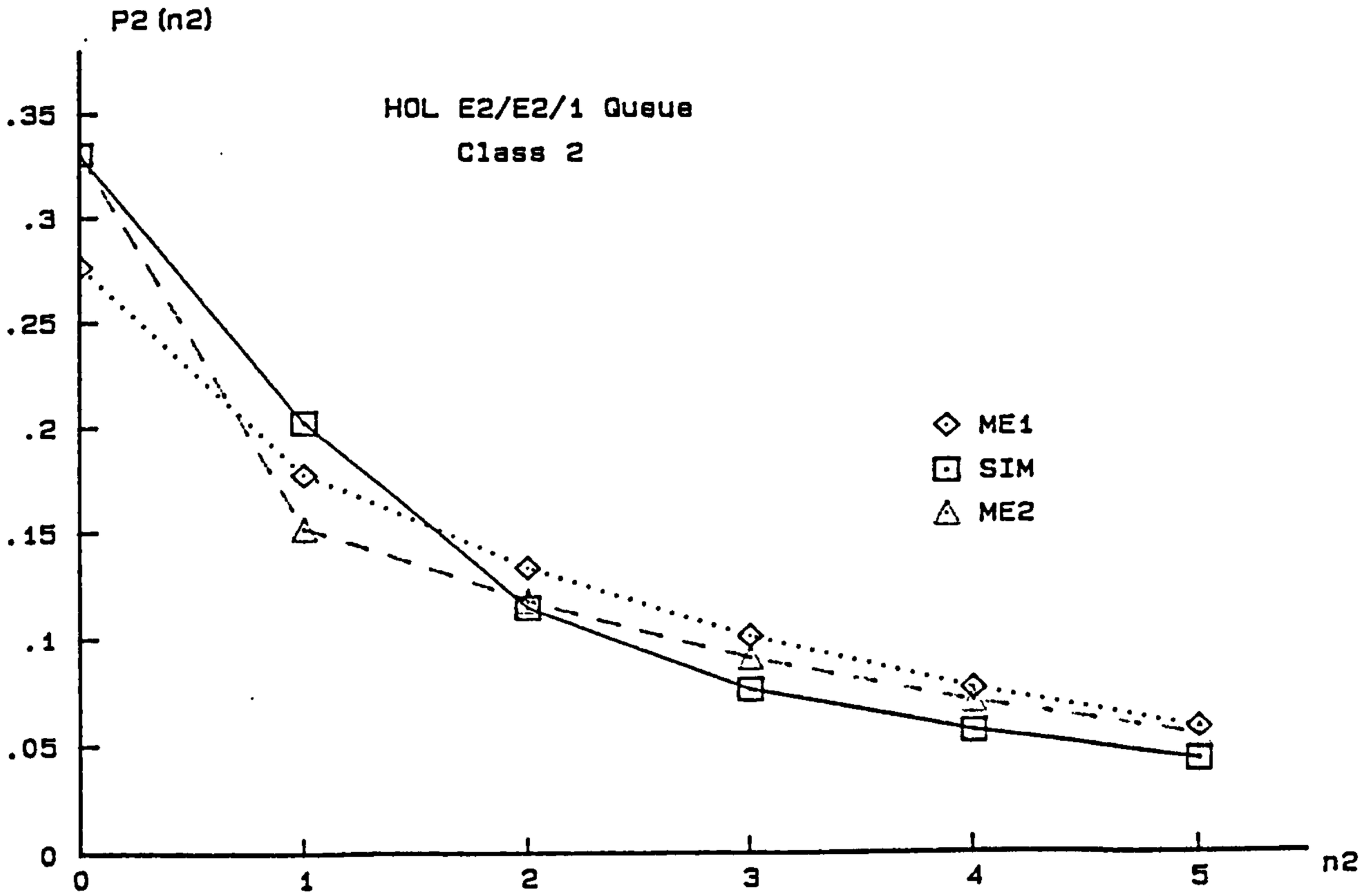

Fig. 5.7b. HOL E2/E2/1 P2(n2) vs n2 (class 2, Table 5.7)

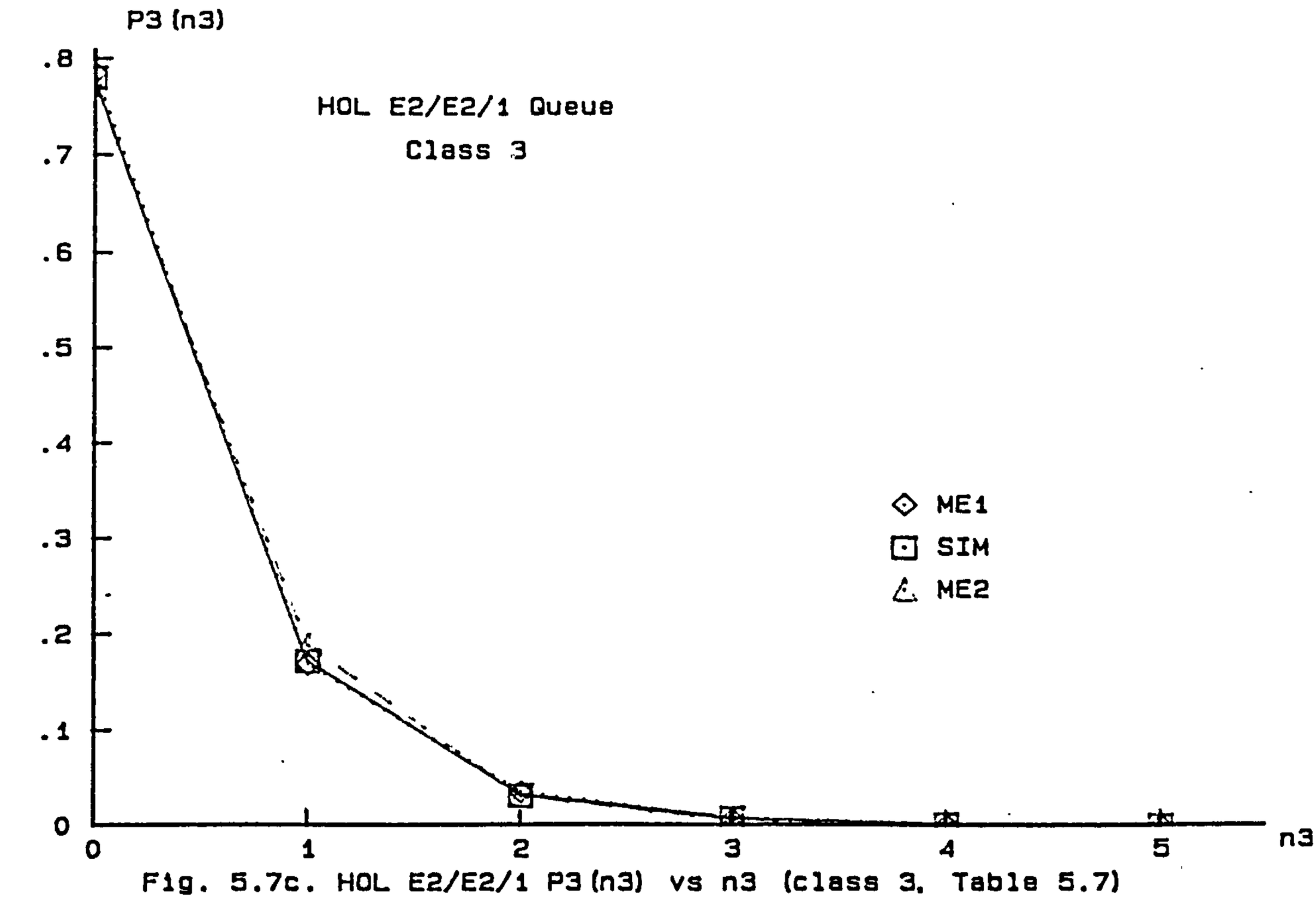

Fig. 5.7c. HOL E2/E2/1 P3(n3) vs n3 (class 3, Table 5.7)

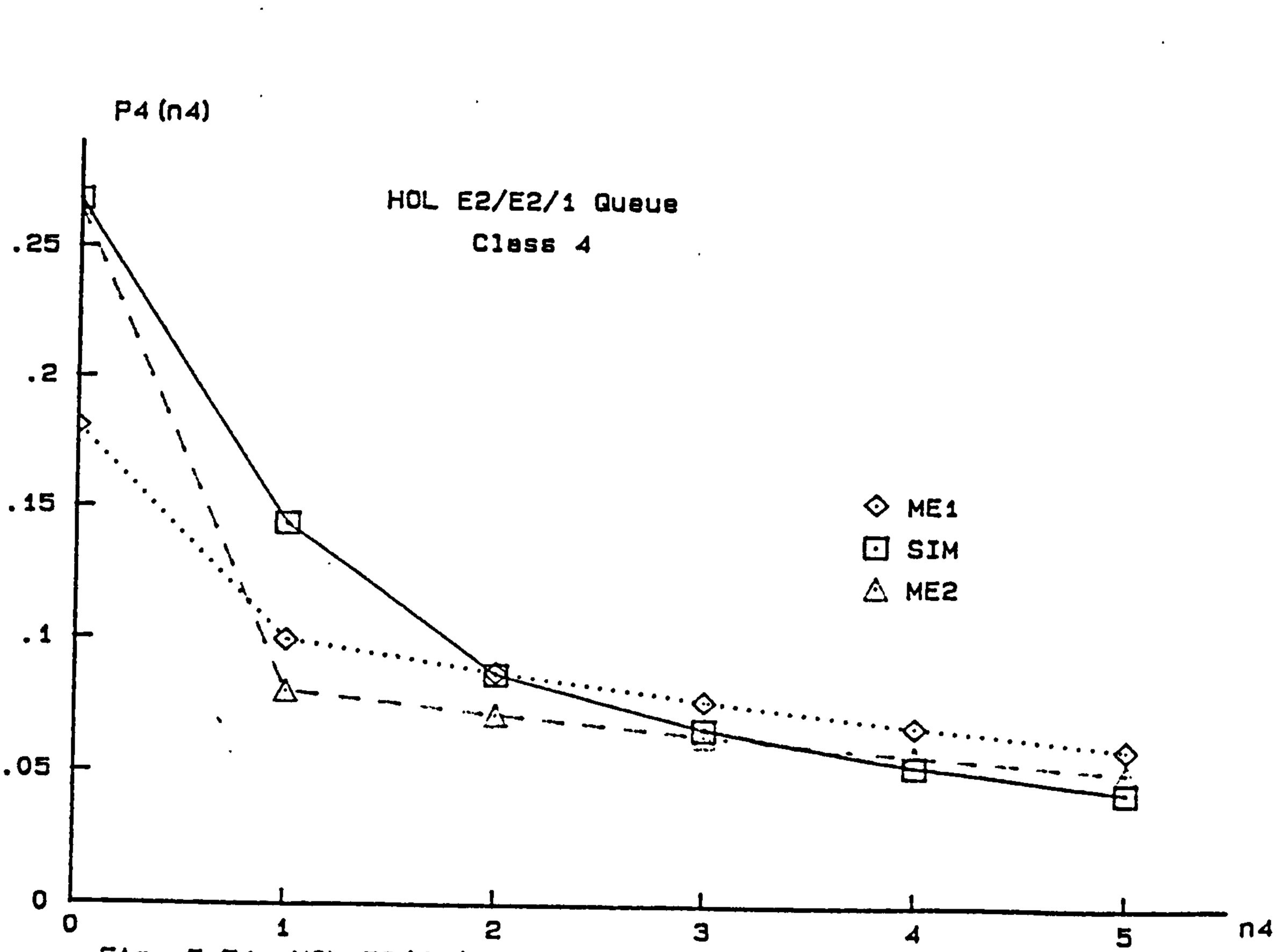

Fig. 5.7d. HOL E2/E2/1 P4(n4) vs n4 (Class 4, Table 5.7)

Example 5.8 $H_2/H_2/1$ HOL queue (4 Classes)

Table 5.8: Raw data for HOL $H_2/H_2/1$ queue

(Figs. 5.8a - 5.8d)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1.5	3	15	6
2	1	2	5	1.5
3	2	5	10	3
4	0.7	4	7	2

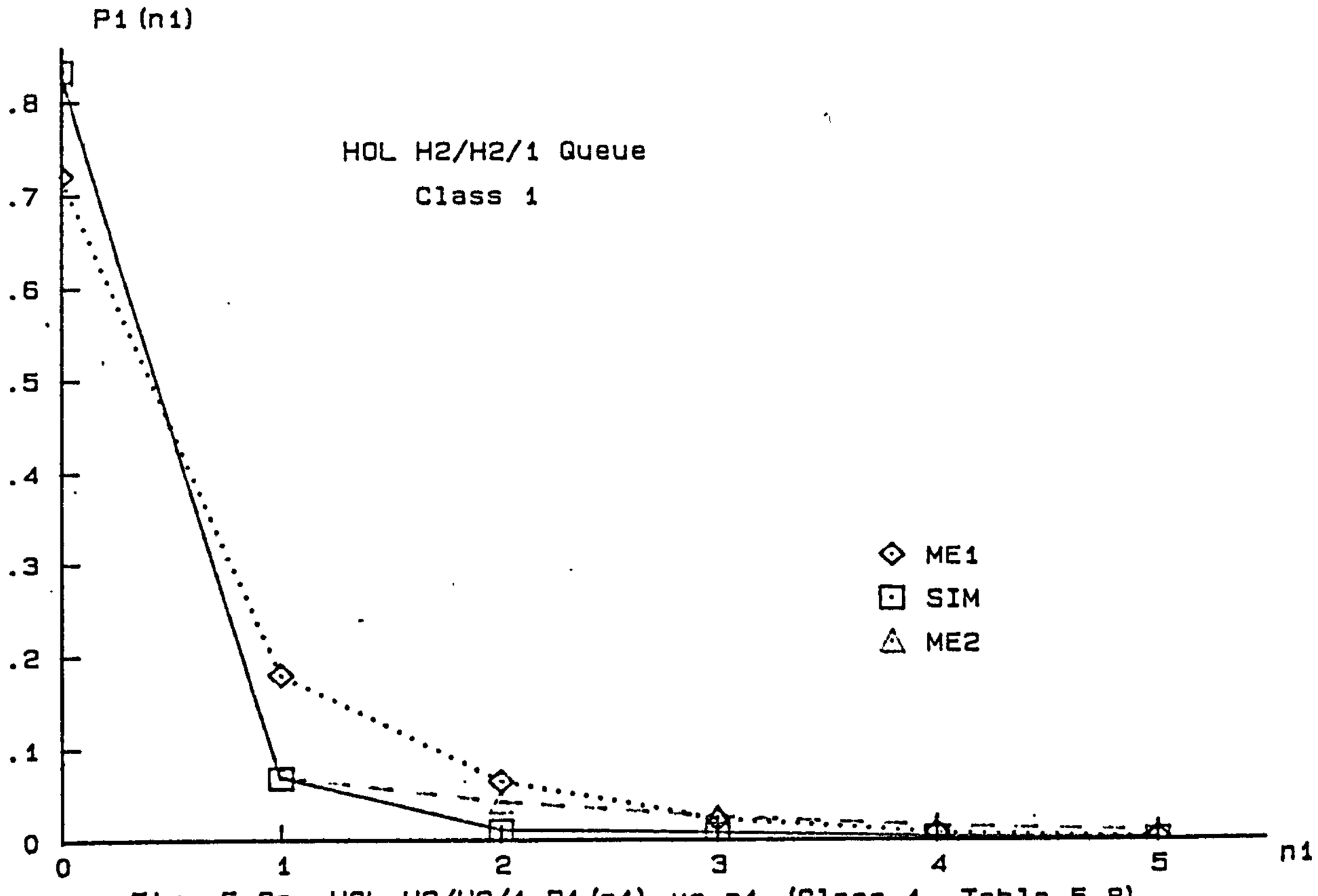

Fig. 5.8a. HOL H2/H2/1 P1(n1) vs n1 (Class 1, Table 5.8)

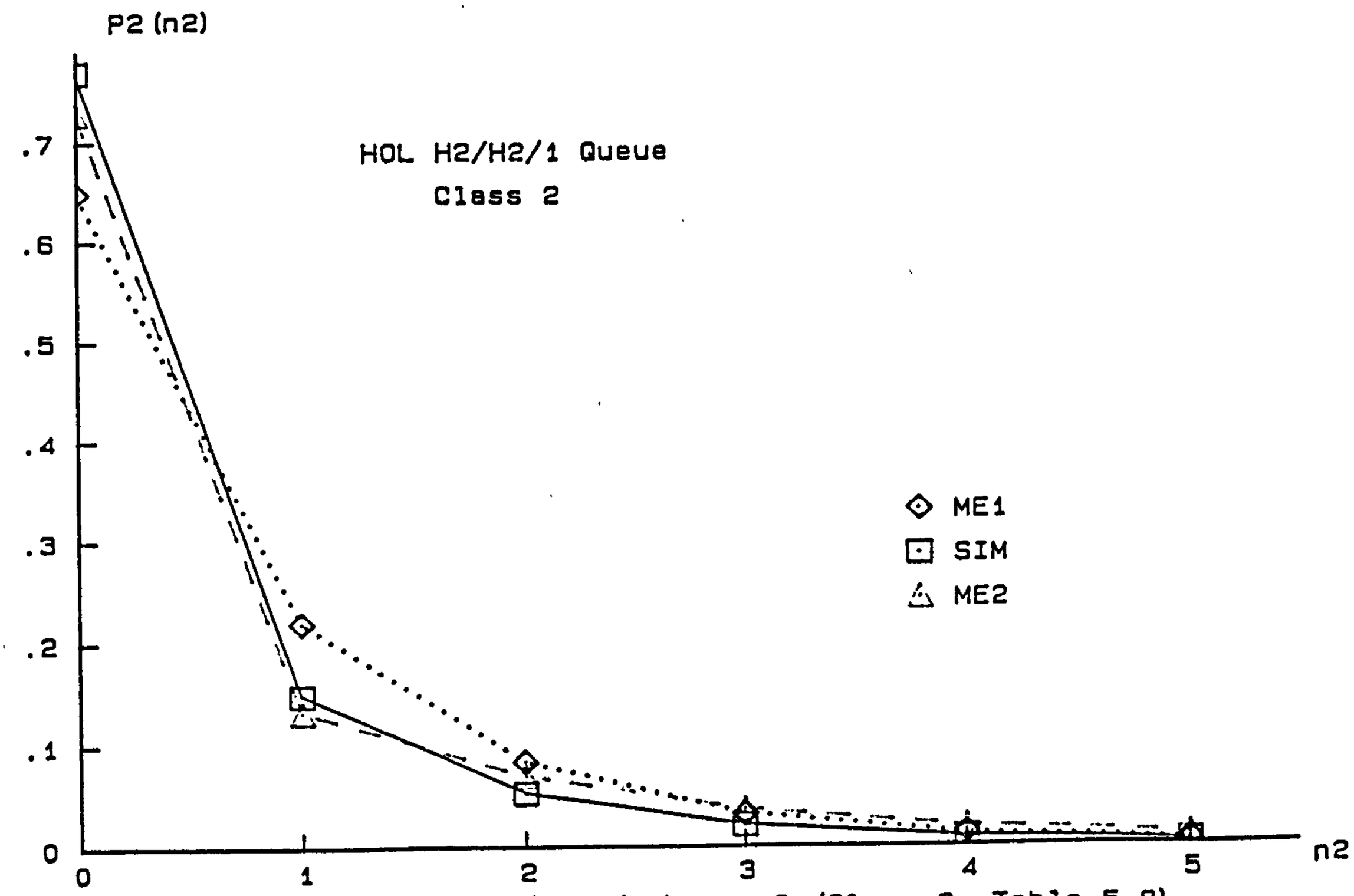

Fig. 5.8b. HOL H2/H2/1 P2(n2) vs n2 (Class 2, Table 5.8)

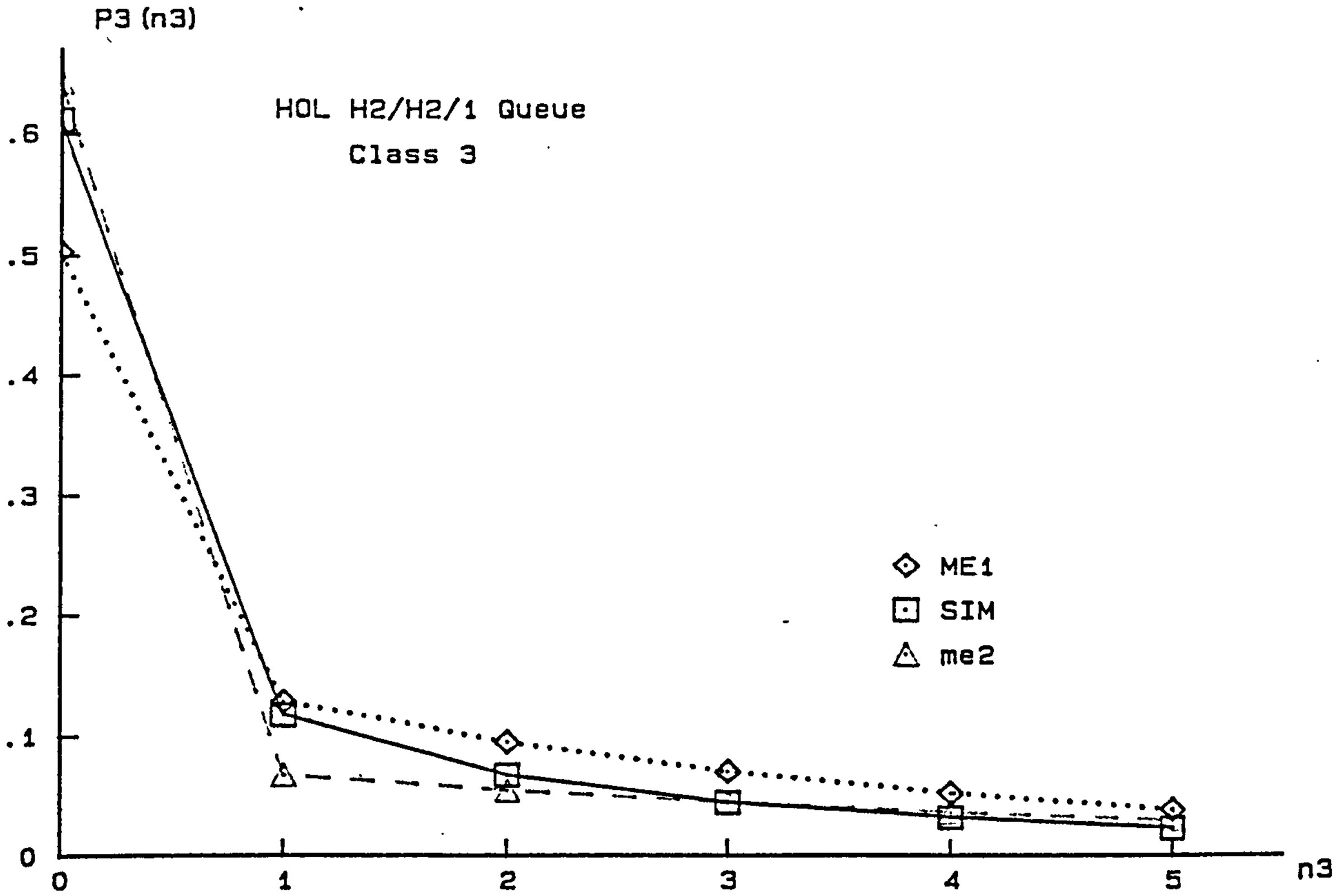

Fig. 5.8c. HOL H2/H2/1 P3(n3) vs n3 (Class 3, Table 5.8)

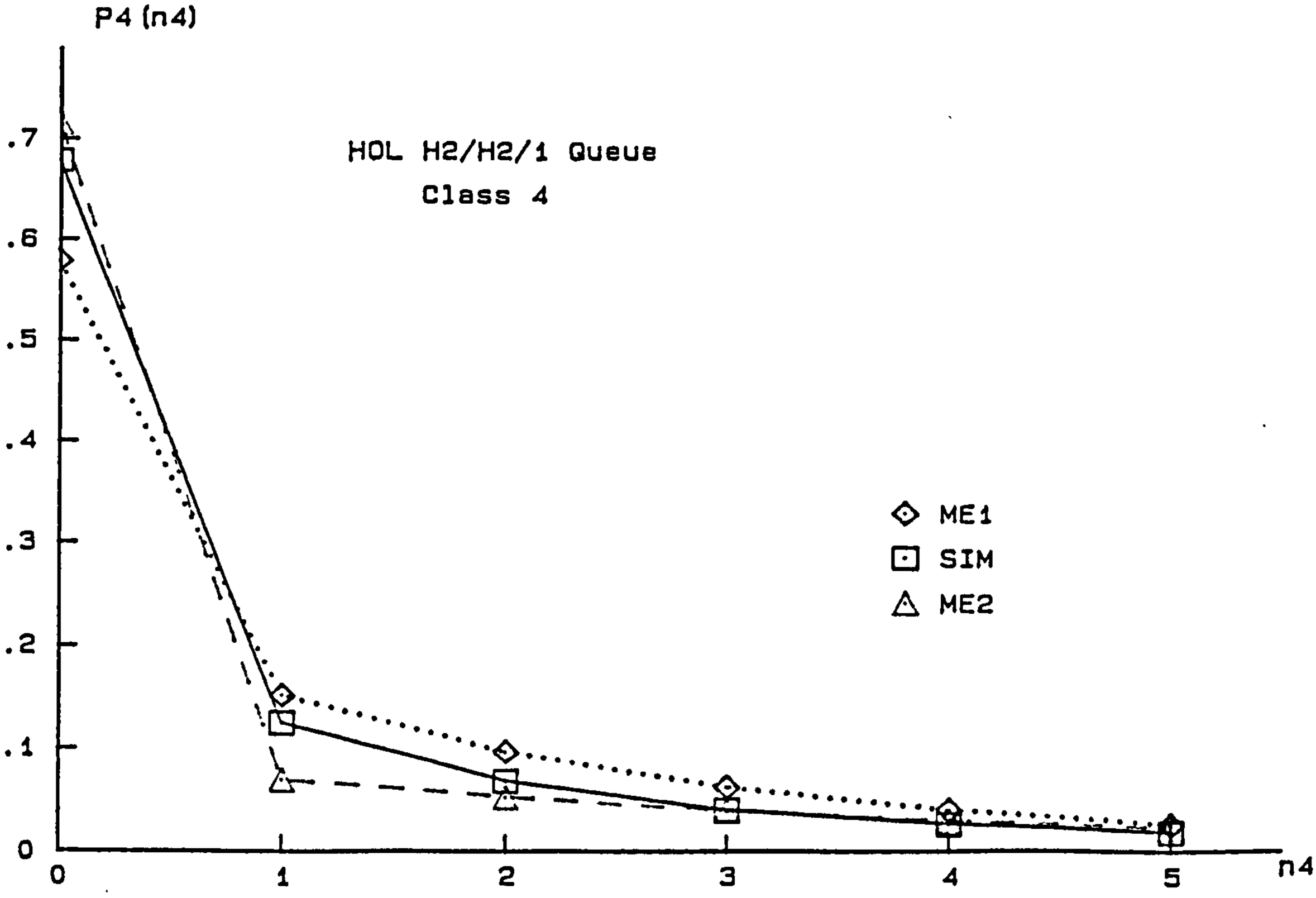

Fig. 5.8d. HOL H2/H2/1 P4(n4) vs n4 (Class 4, Table 5.8)

Example 5.9 GE/GE/1 HOL queue (4 Classes)

Table 5.9: Raw data for HOL GE/GE/1 queue

(Figs. 5.9a - 5.9d)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	2	13	5	18
2	4	9	20	11
3	1.8	5	18	3
4	1	9	5	9

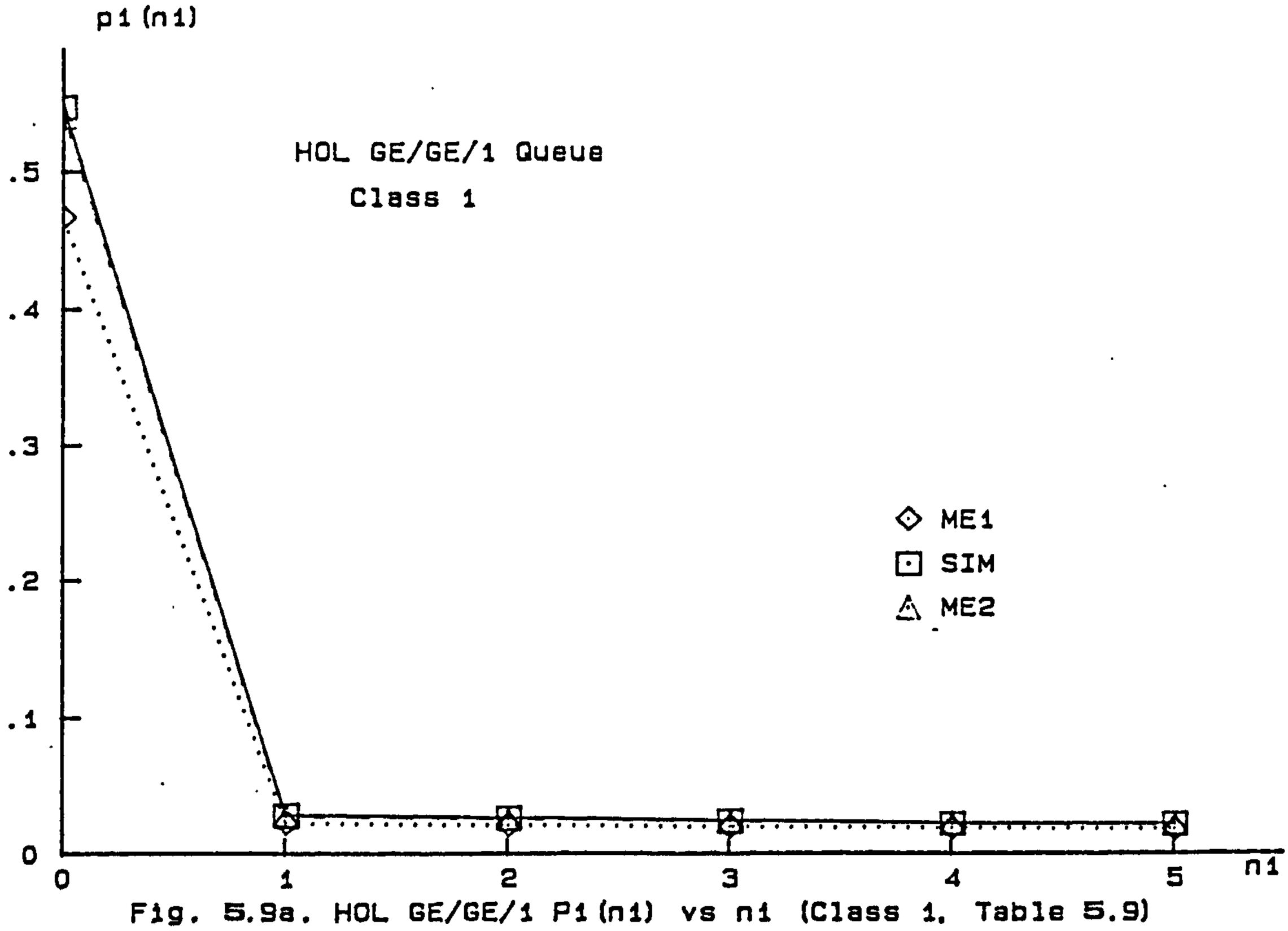

Fig. 5.9a. HOL GE/GE/1 P1(n1) vs n1 (Class 1, Table 5.9)

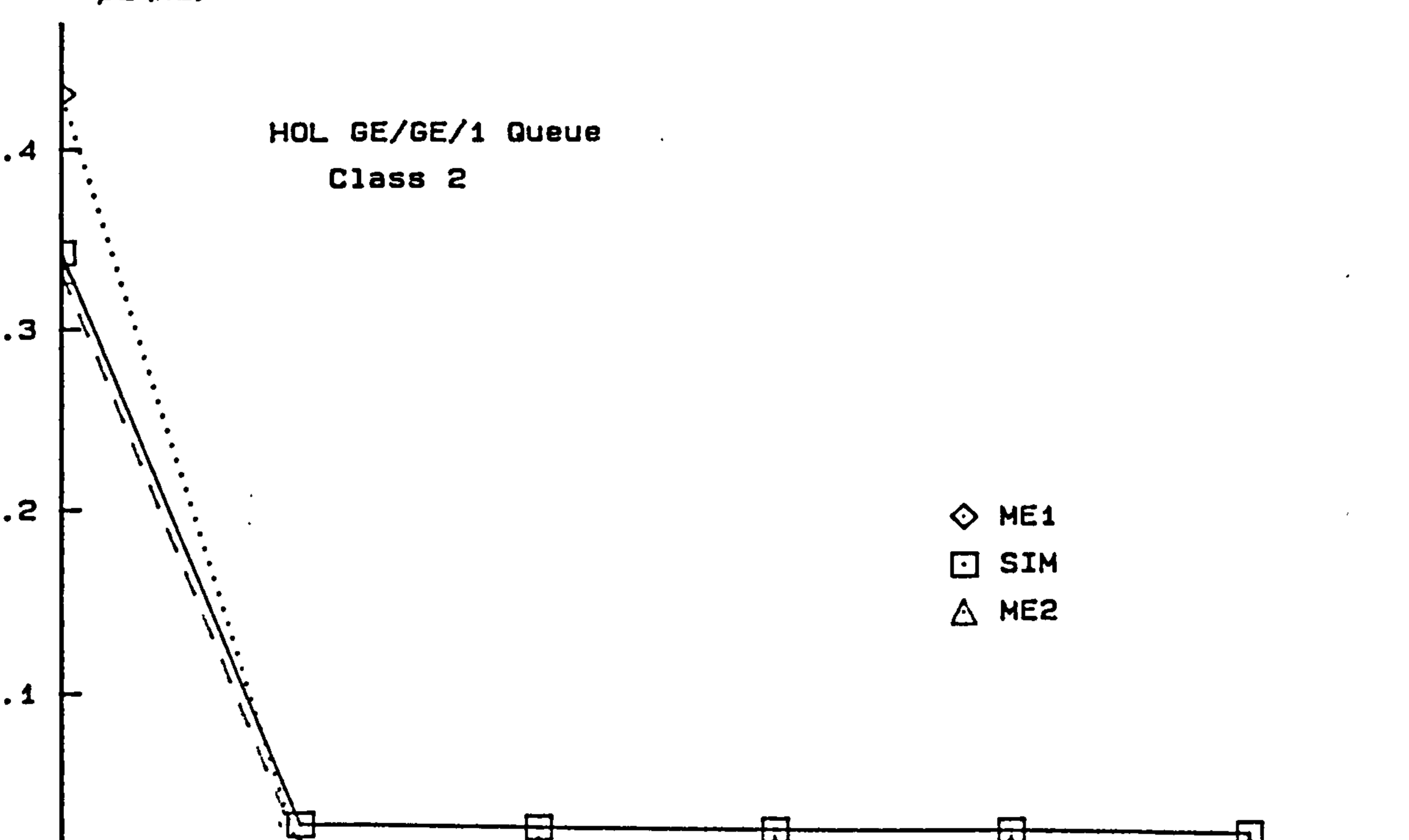

Fig. 5.9b. HOL GE/GE/1 P2(n2) vs n2 (Class 2, Table 5.9)

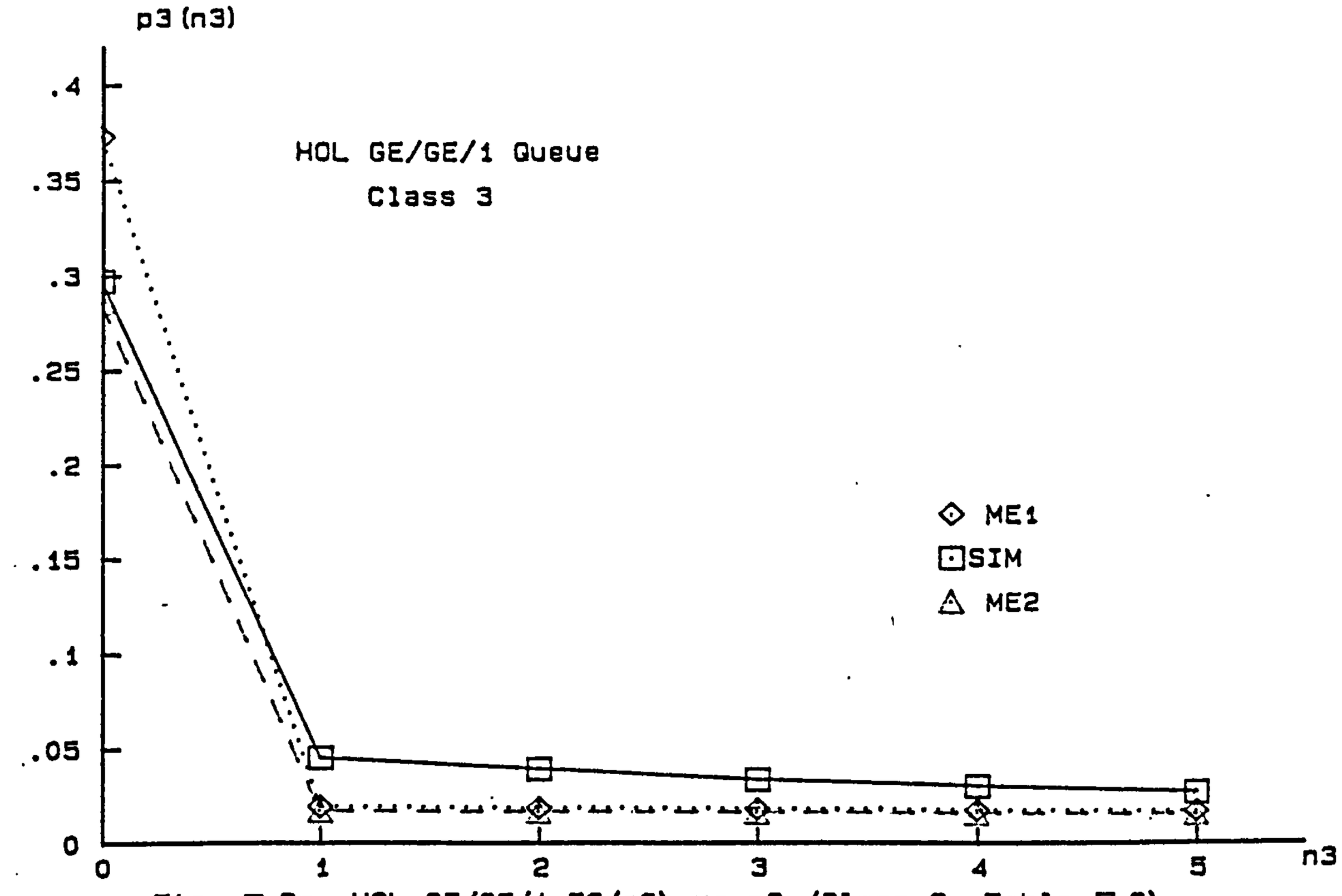

Fig. 5.9c. HOL GE/GE/1 P3(n3) vs n3 (Class 3, Table 5.9)

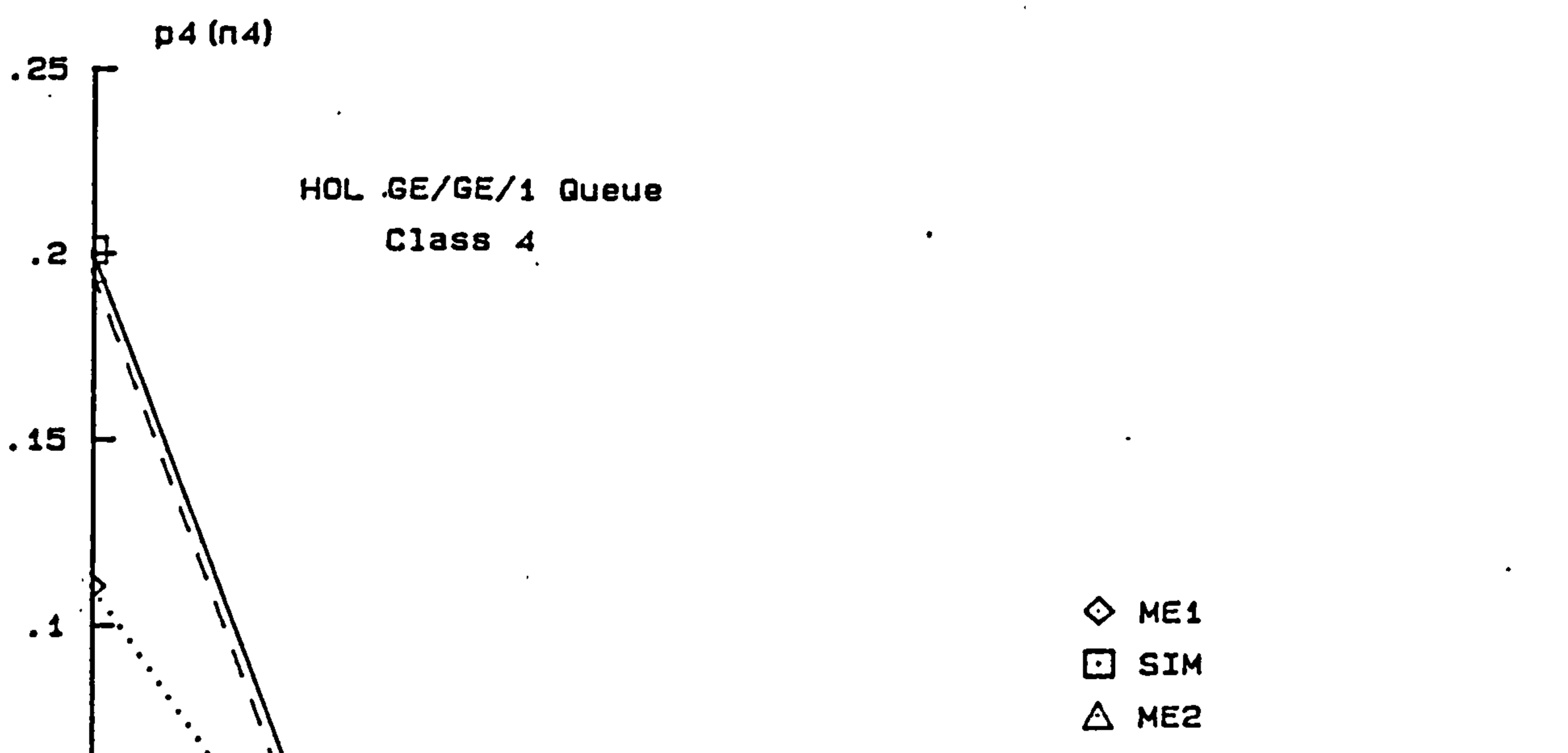

Fig. 5.9d. HOL GE/GE/1 P4(n4) vs n4 (Class 4, Table 5.9)

Example 5.10 $E_2/GE/1$ HOL queue (4 Classes)

Table 5.10: Raw data for HOL $E_2/GE/1$ queue

(Figs. 5.10a - 5.10d)

Class r	λ_r	C_{ar}^2	μ_r	C_{sr}^2
1	1	0.5	5	4
2	1.5	0.5	5	6
3	2	0.5	20	7
4	2	0.5	10	4

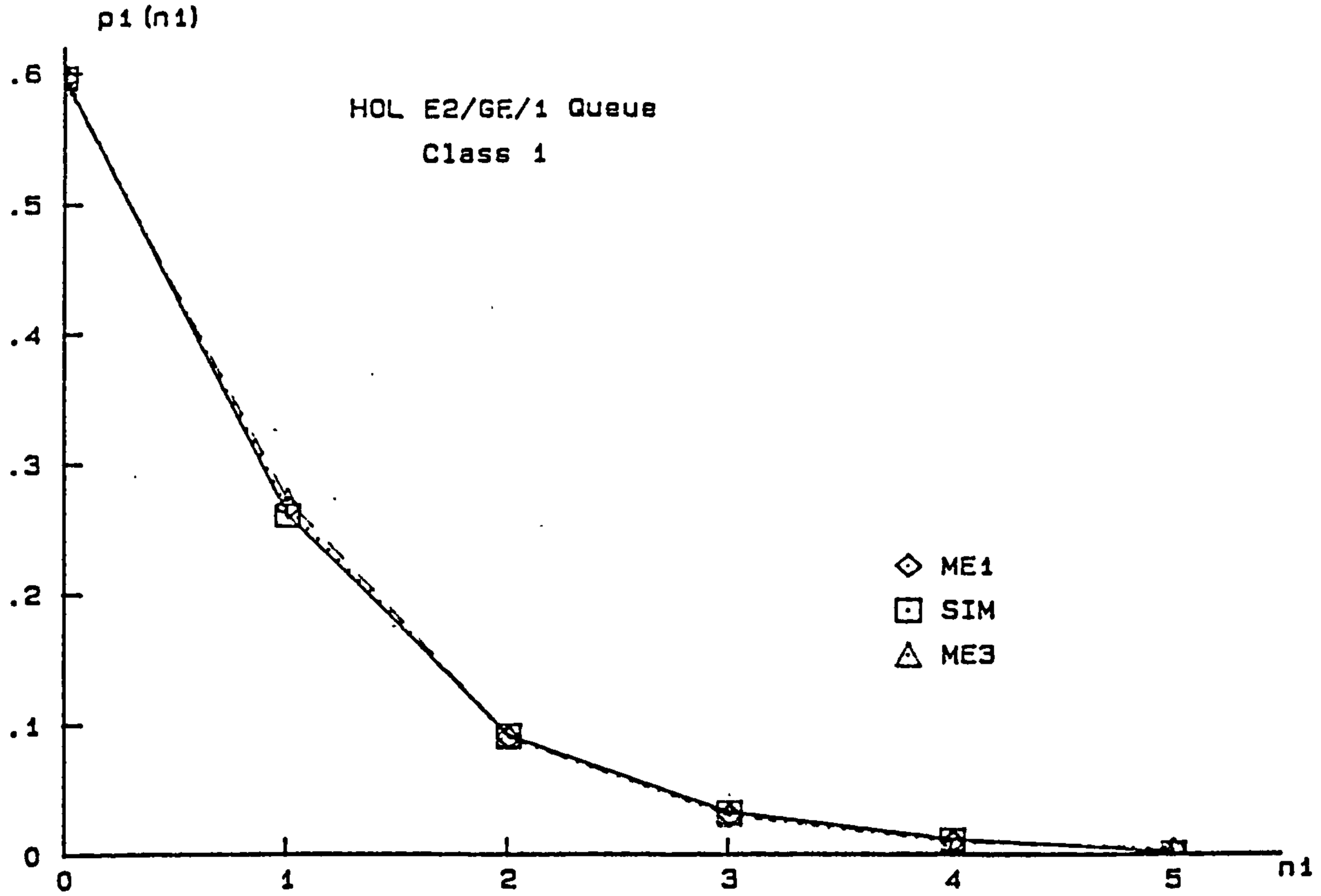

Fig. 5.10a. HOL E2/GE/1 P1(n1) vs n1 (Class 1, Table 5.10)

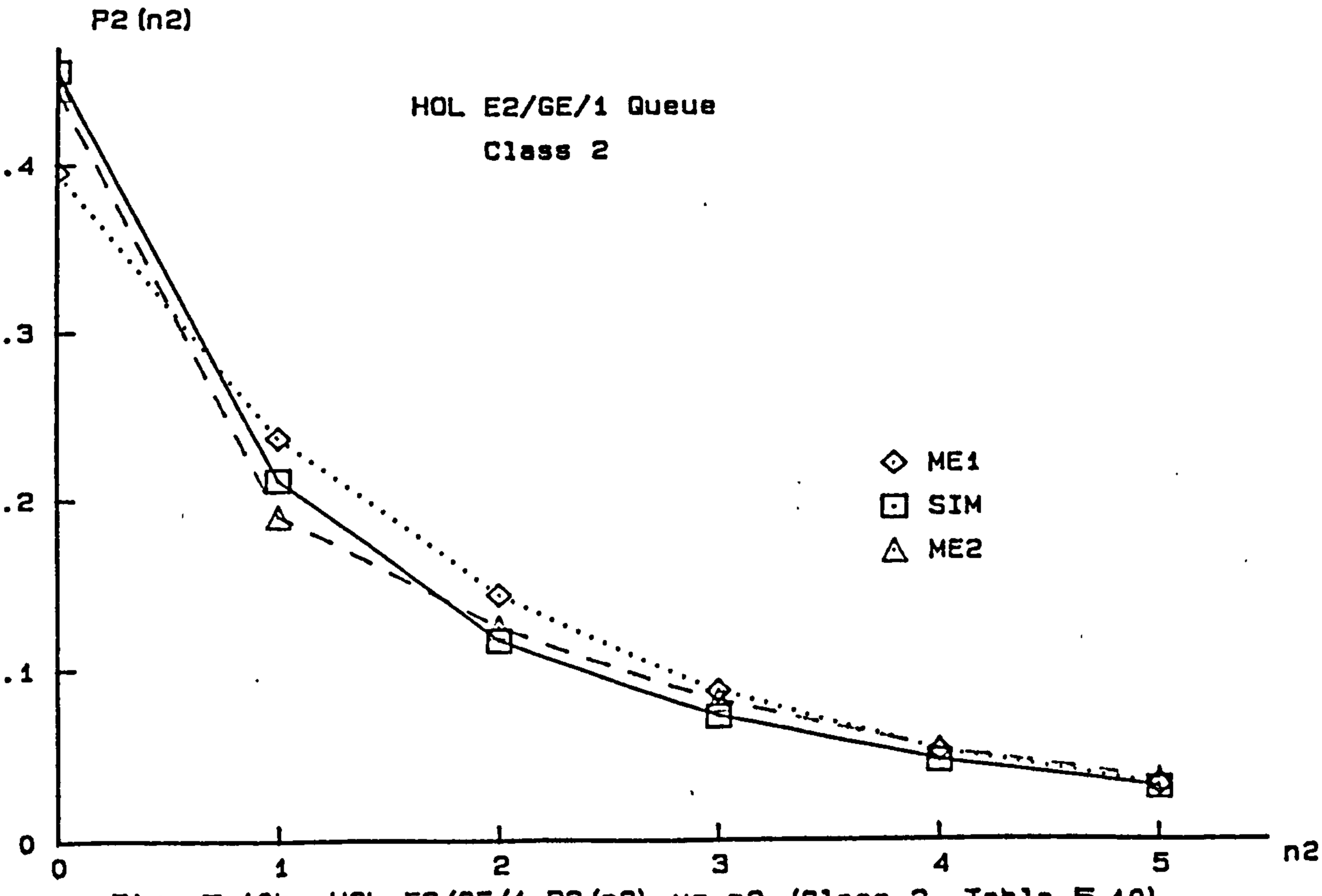

Fig. 5.10b. HOL E2/GE/1 P2(n2) vs n2 (Class 2, Table 5.10)

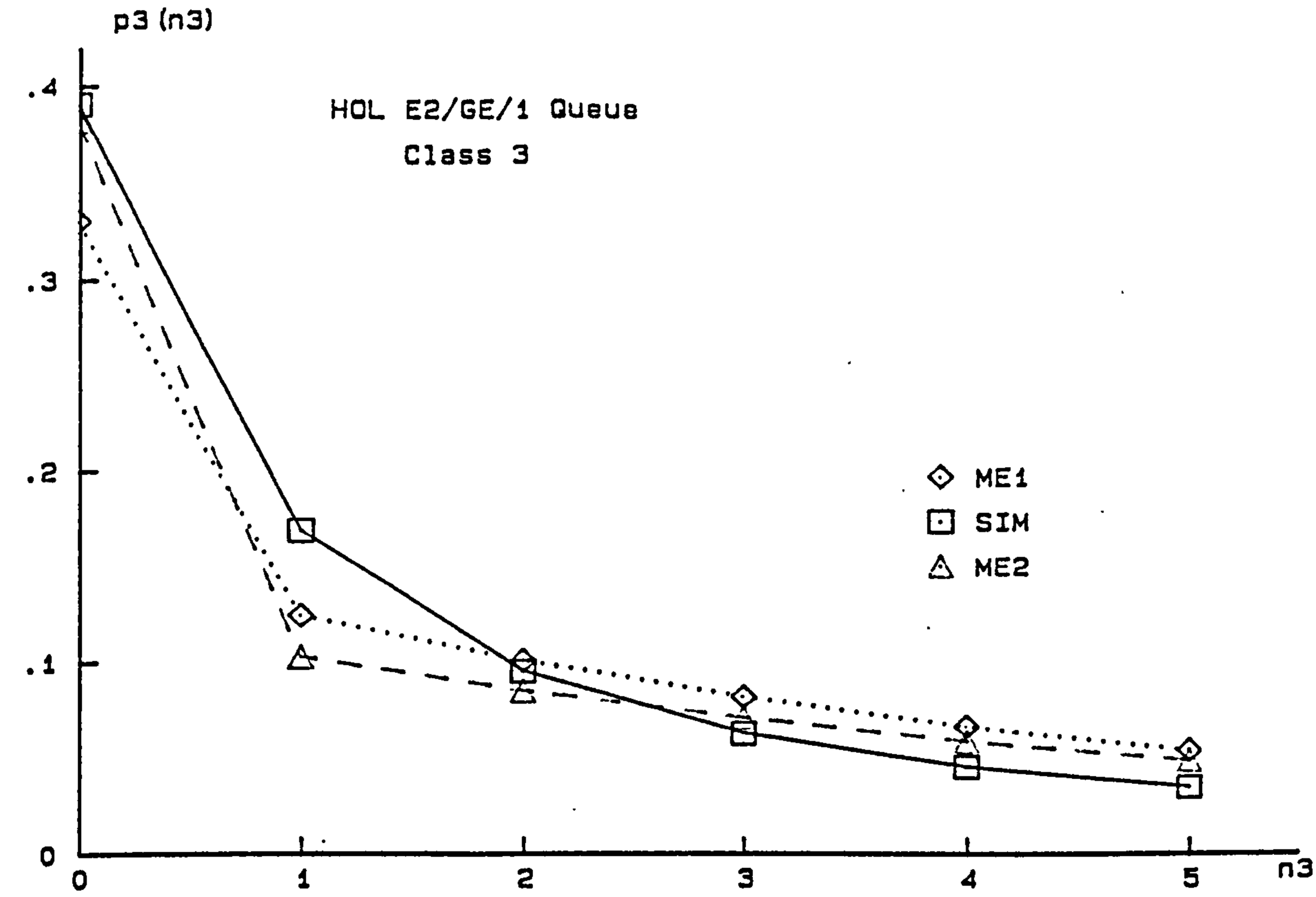

Fig. 5.10c HOL E2/GE/1 P3(n3) vs n3 (class 3, Table 5.10)

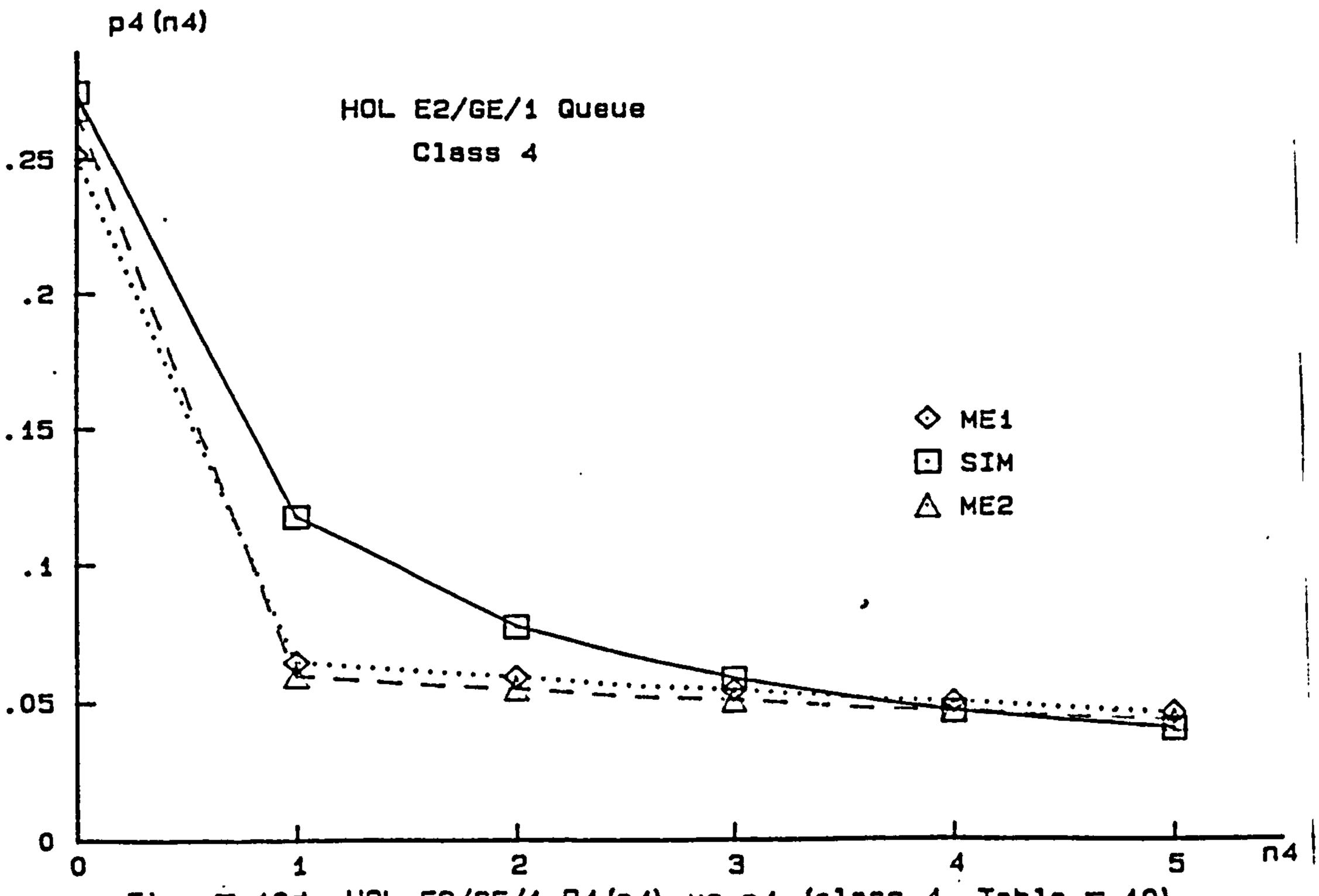

Fig. 5.10d. HOL E2/GE/1 P4(n4) vs n4 (class 4, Table 5.10)

Example 5.11 D,M/GE,U/1 PR queue (2 Classes)

Table 5.11: Raw data for PR D,M/GE,U/1 queue

(Fig. 5.11)

Class r	Distributions		λ_r	C_{ar}^2	μ_r	C_r^2	Simulation Constraints	
	Inter-arrival time	service-time					$\langle n_r \rangle$	$P_r(0)$
1	D	GE	2	0	13.33	7	0.192	0.85
2	M	U[0,2]	1/3	1	1	1/3	0.587	0.6006

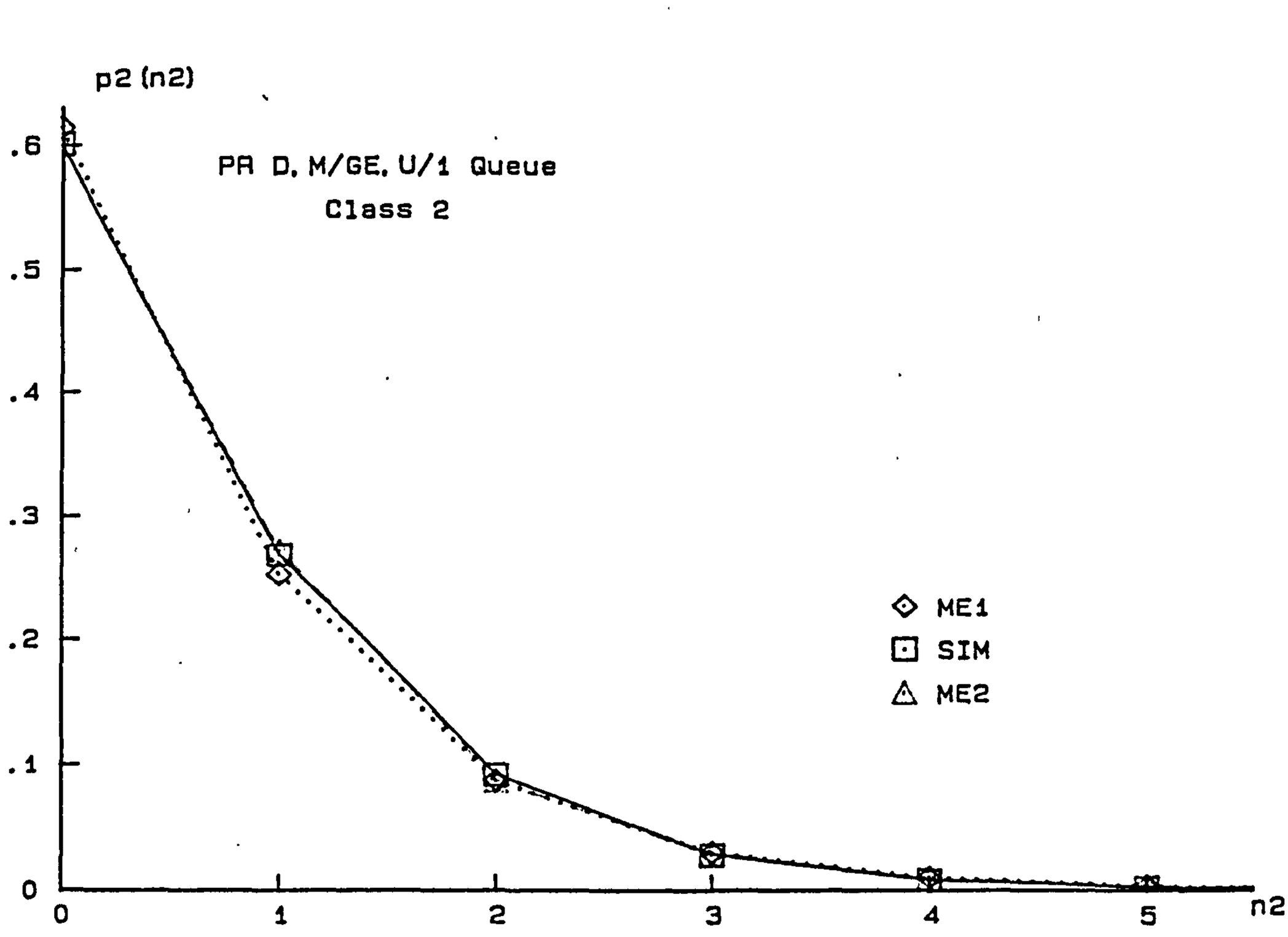

Fig. 5.11. PR D, M/GE, U/1 P2 (n2) vs n2 (class 2, Table 5.11)

Example 5.12 D,U,H_2/H_2,E_2,M/1 PR queue (3 Classes)

Table 5.12: Raw data for PR D,U,H_2/H_2,E_2,M/1 queue

(Figs. 5.12a-b)

	Distributions						Simulation Constraints	
Class r	Inter-arrival time	service-time	λ_r	C_{ar}^2	μ_r	C_r^2	$\langle n_r \rangle$	$P_r(0)$
1	D	H_2	3	0	1	4	0.587	0.7
2	U[0,0.667]	E_2	3	1/3	15	0.5	0.891	0.57
3	H_2	M	2.4	3	12	1	1.86	0.543

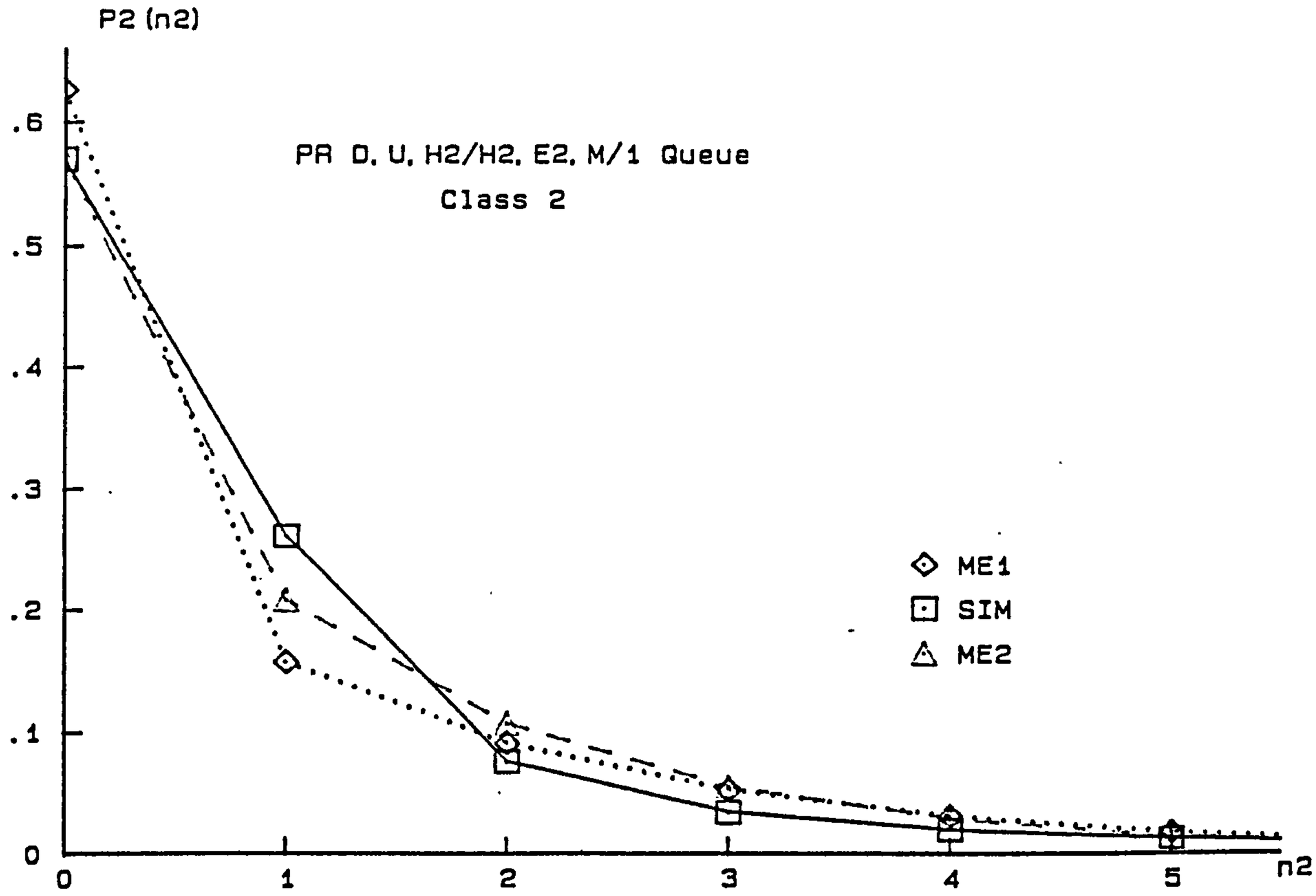

Fig. 5.12a. PR D, U, H2/H2, E2, M/1 P2(n2) vs n2 (Class 2, Table 5.12)

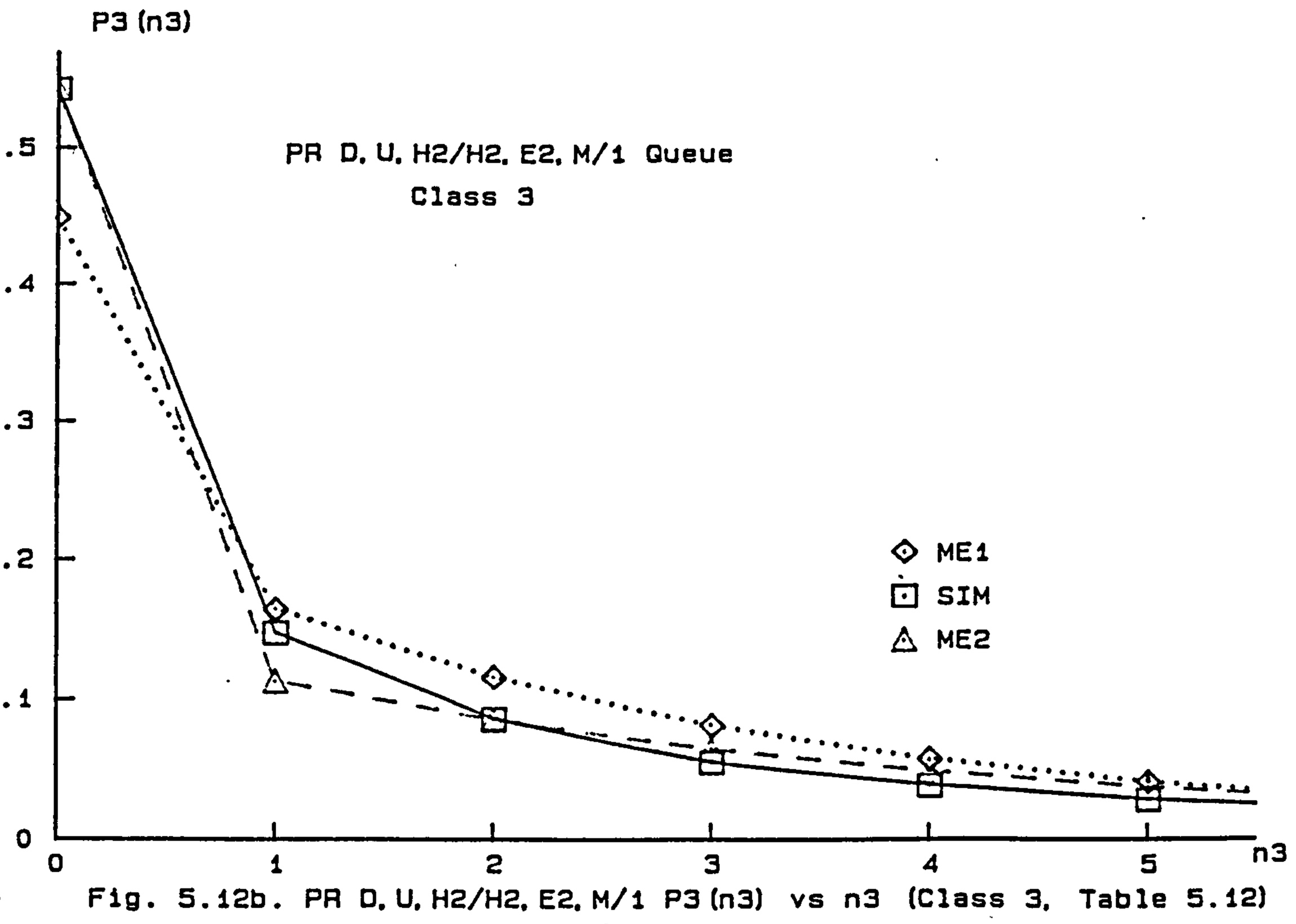

Fig. 5.12b. PR D, U, H2/H2, E2, M/1 P3(n3) vs n3 (Class 3, Table 5.12)

Example 5.13 $U,H_2/M,E_2/1$ HOL queue (2 Classes)

Table 5.13: Raw data for HOL $U,H_2/M,E_2/1$ queue

(Figs. 5.13a-b)

Class r	Distributions: Inter-arrival time	Distributions: service-time	λ_r	C_{ar}^2	μ_r	C_r^2	Simulation Constraints: $\langle n_r \rangle$	Simulation Constraints: $P_r(0)$
1	U[0,10]	M	0.2	1/3	0.5	1	0.76	0.471
2	H_2	E_2	0.2	25	0.5	0.5	14.76	0.335

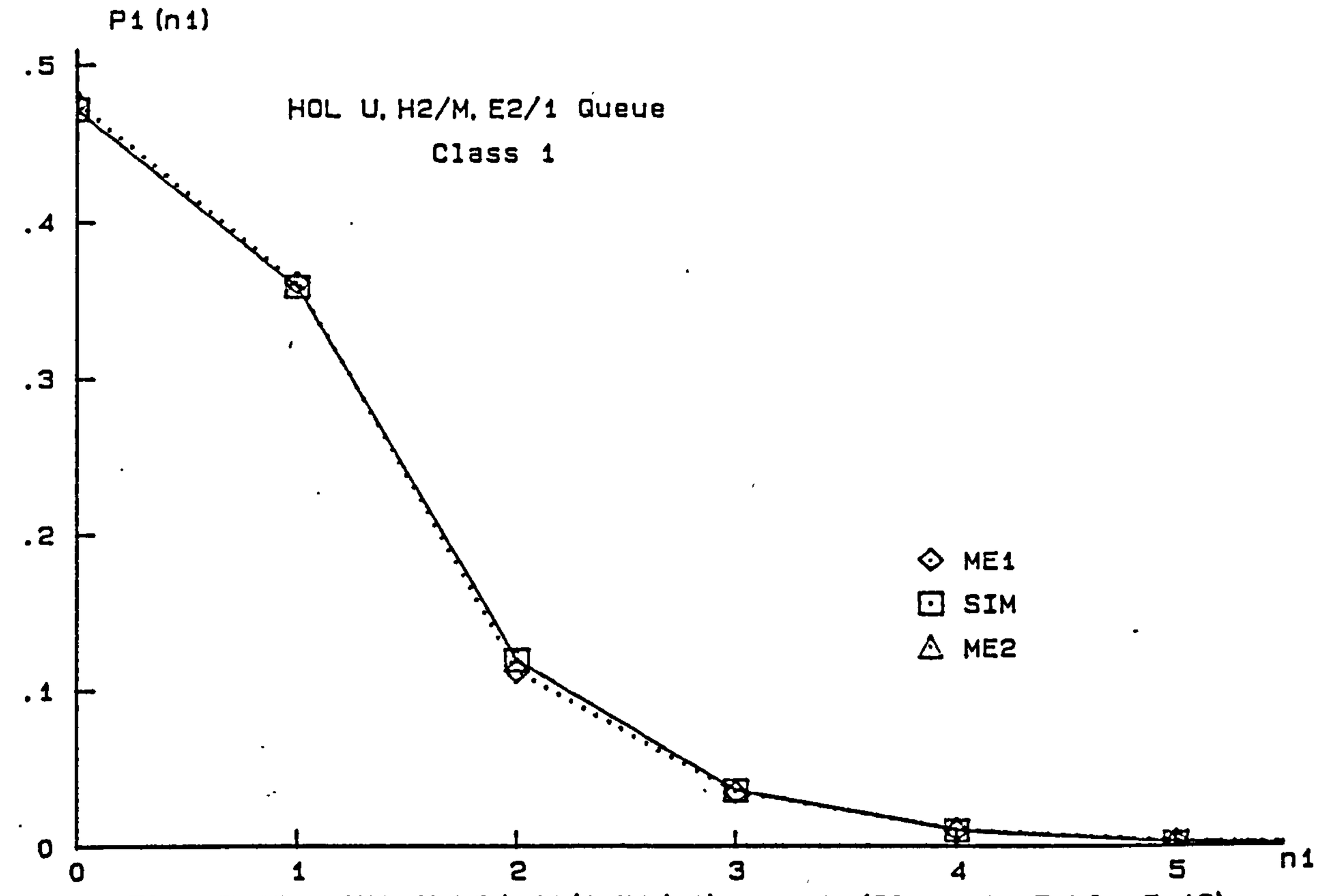

Fig. 5.13a. HOL U, H2/M,E2/1 P1(n1) vs n1 (Class 1, Table 5.13)

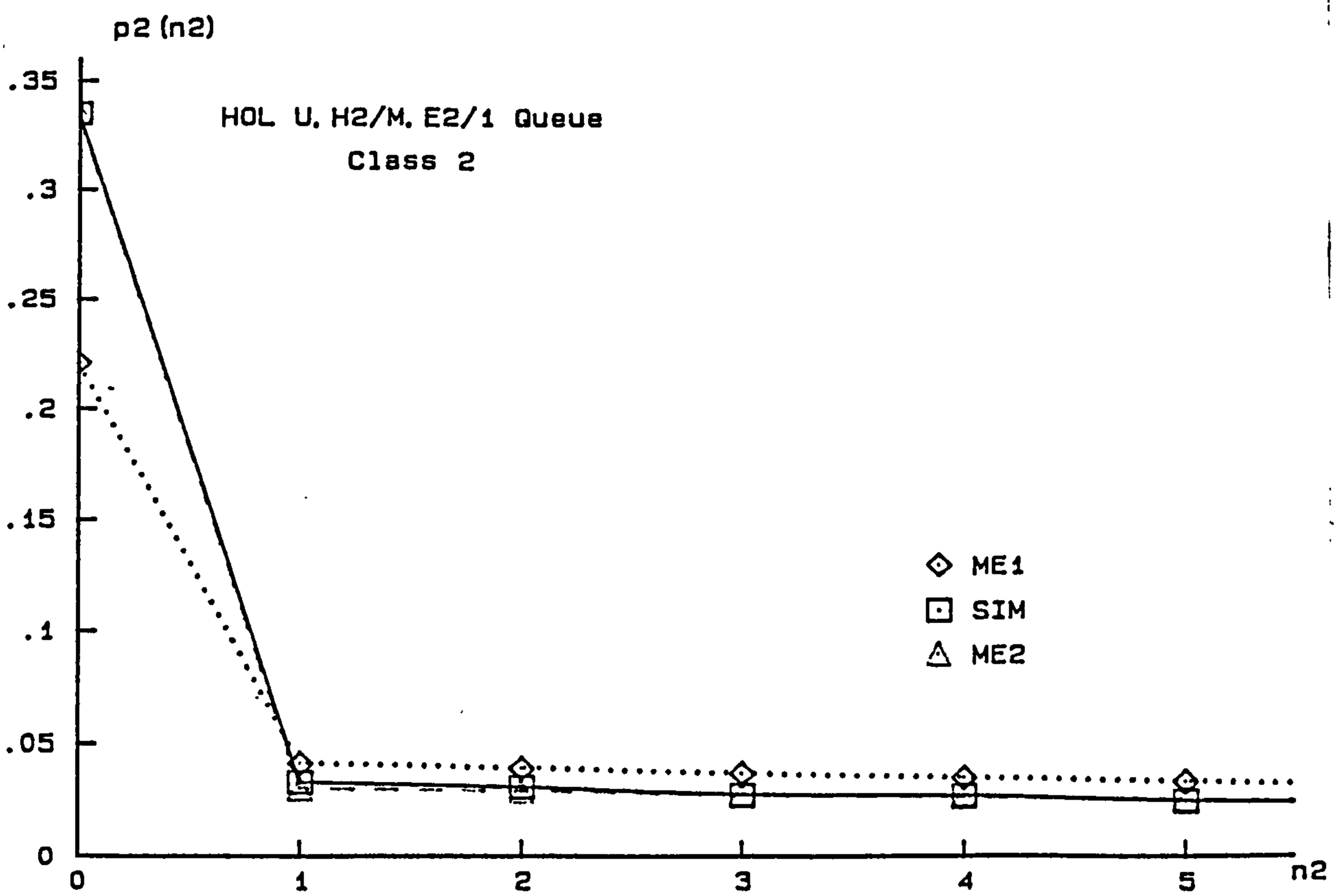

Fig. 5.13b. HOL U, H2/M, E2/1 P2(n2) vs n2 (Class 2, Table 5.13)

Example 5.14 $E_2,M,H_2/GE,M,E_2/1$ HOL queue (3 Classes)

Table 5.11: Raw data for HOL $E_2,M,H_2/GE,M,E_2/1$ queue

(Figs. 5.14a-c)

Class r	Distributions: Inter-arrival time	Distributions: service-time	λ_r	C_{ar}^2	μ_r	C_r^2	Simulation Constraints: $\langle n_r \rangle$	Simulation Constraints: $P_r(0)$
1	E_2	GE	3	0.5	10	4	0.698	0.587
2	M	M	6	1	15	1	3.421	0.283
3	H_2	E_2	2.6	6	12	0.5	12.32	0.258

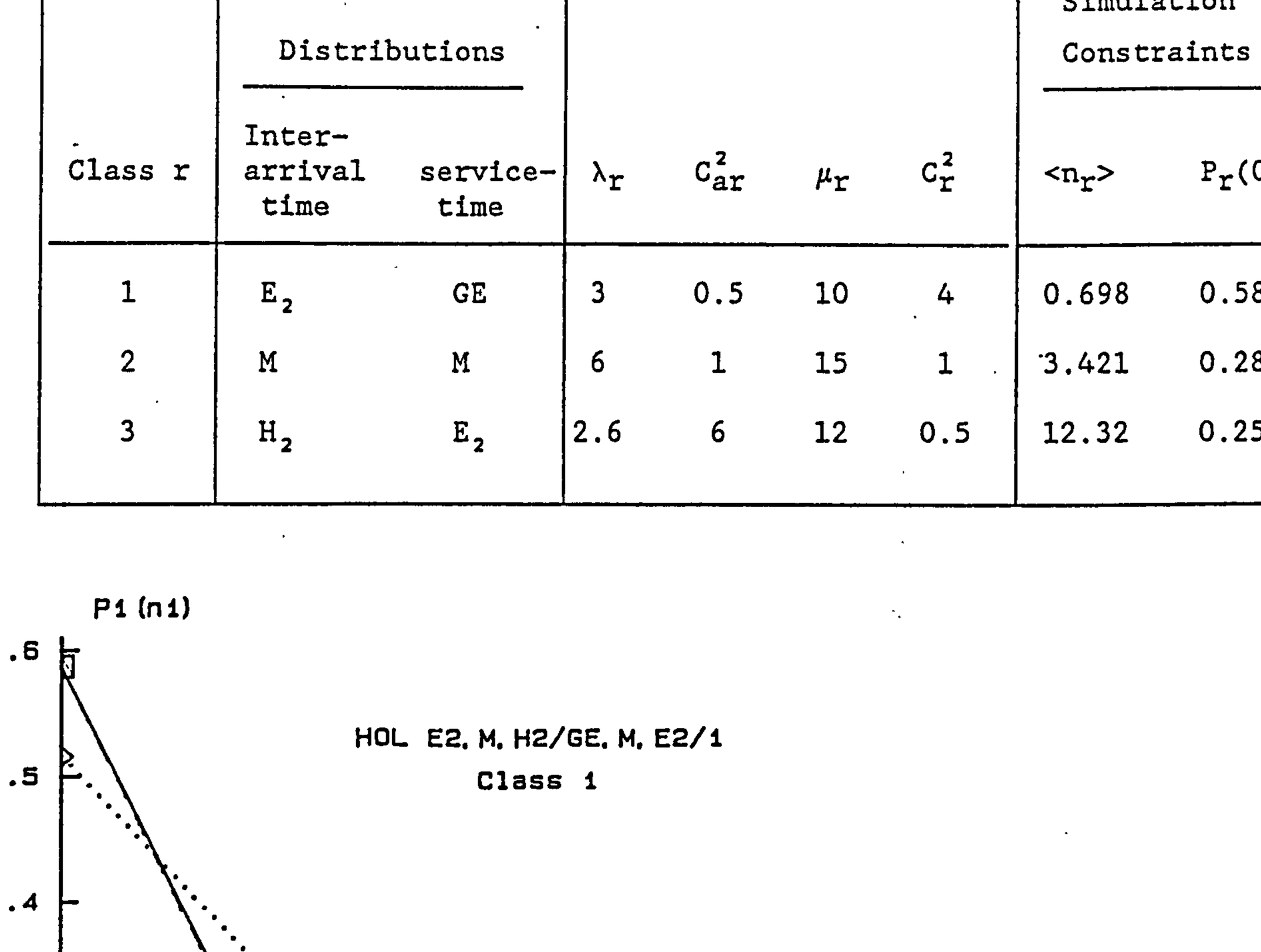

Fig. 5.14a. HOL E2, M, H2/GE, M, E2/1 P1(n1) vs n1 (Class 1, Table 5.14)

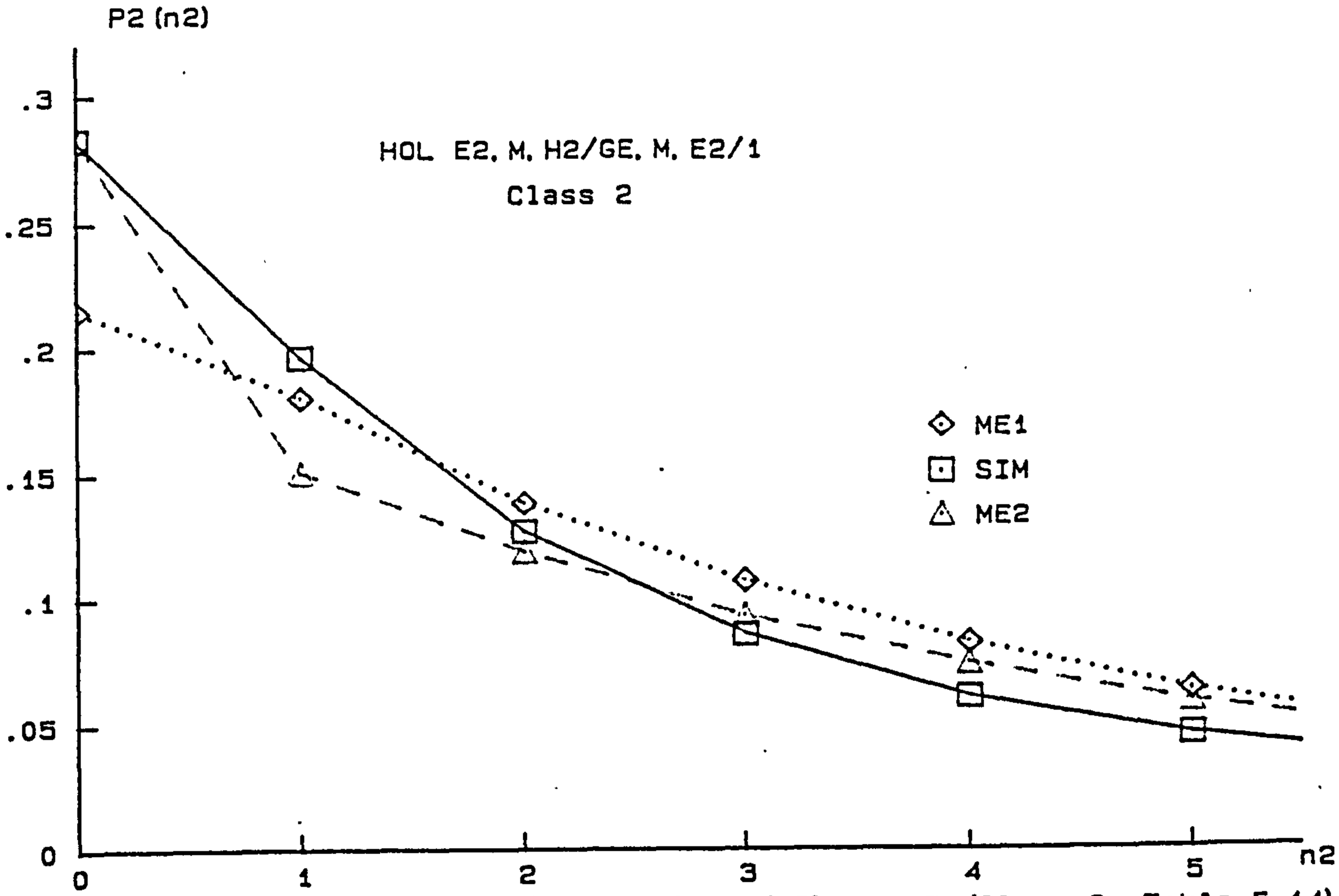

Fig. 5.14b. HOL E2. M. H2/GE. M. E2/1 P2 (n2) vs n2 (Class 2. Table 5.14)

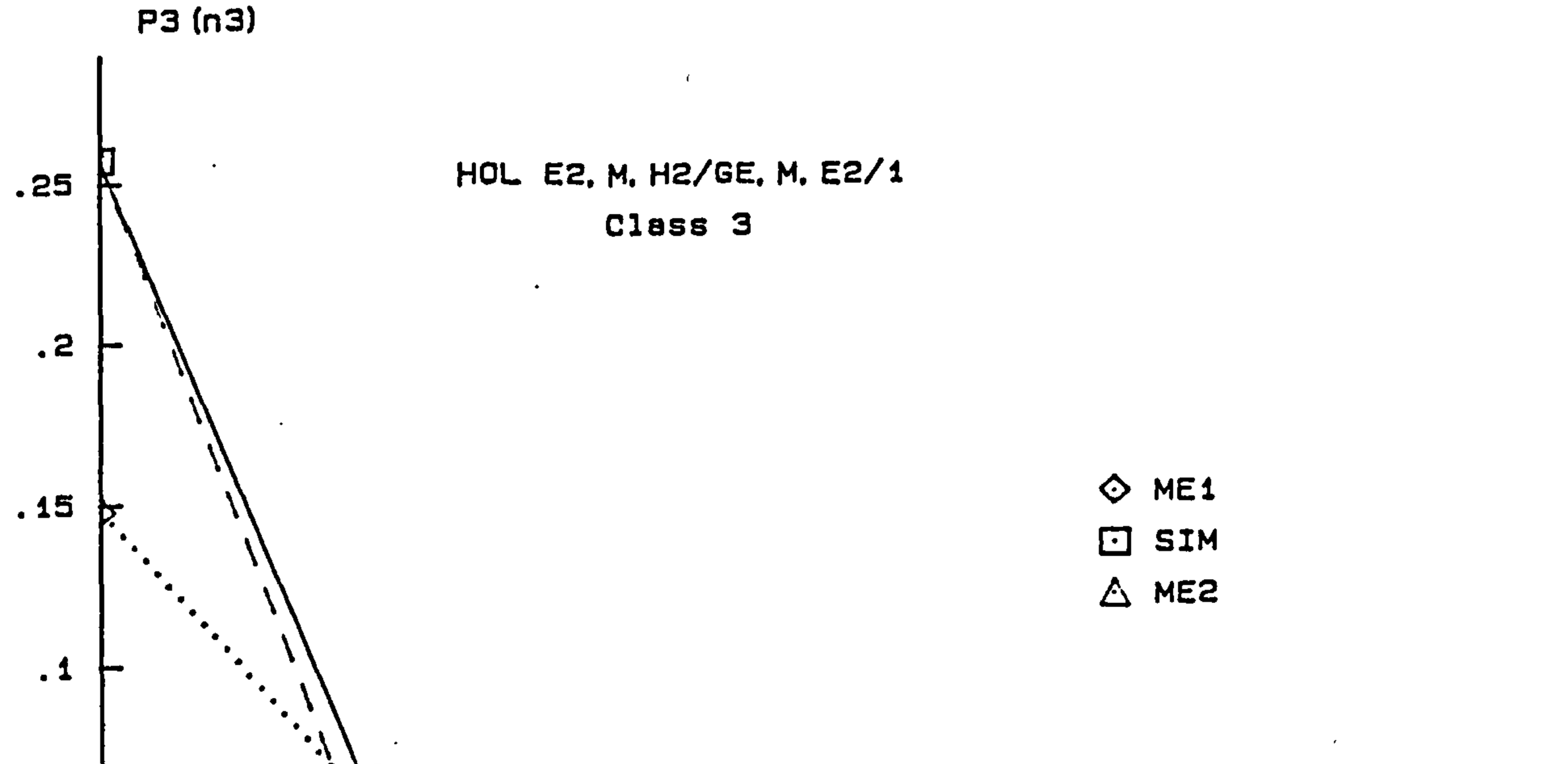

Fig. 5.14c. HOL E2. M. H2/GE. M. E2/1 P3 (n3) vs n3 (Class 3. Table 5.14)

APPENDIX E

E.1 Evaluation of the system mean response-time

The mean queue lengths, $\langle n_{ir} \rangle$ and the mean response time, $\langle T_{ir} \rangle$, per class-r and at each centre-i (per visit) are obtained from the last iteration of the algorithm 6.1. However, when jobs switch class membership as they move from one centre to another, the evaluation of the system mean response time is not straightforward. Nevertheless, the problem is overcome by using the concept of 'equivalence class' introduced earlier by Bruel and Balbo [BRUE,80].

The 'equivalence class-r set' denoted by EQ(r) is defined to be the set of all the classes which communicate with class-r (a class-s communicates with class-r if stage (i,r) can be reached in zero or more transitions from stage (j,s), for any i,j ϵ [1,M] and r,s ϵ [1,R]).

Having done so for all classes r, r=1,...R,, we unify all the sets which have non-empty intersection. The final sets obtained are mutually disjoint and form a composite class called 'equivalence class' denoted by $E_1, E_2, \ldots, E_e$, where obviously $e \leqslant R$. The classes that belong to a class E_r communicate with each other and no with any other class belonging to a different set $E_s, s \neq r$.

In particular if jobs don't switch class membership we have,

$E_r = \{r\}$ for r=1,...,R.

Defining, $\lambda_{i\ell} = \sum_{r \epsilon E_\ell} \lambda_{ir}$ for ℓ=1,...,e as the mean arrival rate of the equivalence class E_ℓ to centre-i, and $v_{i\ell} = \lambda_{i\ell}/\lambda_{0\ell}$, ℓ=1,...,e as the visit ratio (number of visits per unit of time) of class E_ℓ at centre-i. Moreover, a job belonging to equivalence class E_ℓ and is originally of class-r arrives to a centre-i with probability,

$$\alpha_{ir} = \begin{cases} 0 \text{ if } r \notin E_\ell \\ \\ \lambda_{ir}/\lambda_{i\ell} \text{ if } r \epsilon E_\ell \end{cases}$$

and clearly, the visit ratio of class-r at centre-i is given by $v_{ir} = \alpha_{ir} v_{i\ell}$ or equivalently $v_{ir} = \lambda_{ir}/\lambda_{0\ell}$.

Therefore, the mean time that a class-r job spends in centre-i, $\langle Te_{ir} \rangle$, is clearly given by $\langle Te_{ir} \rangle = v_{ir} \langle T_{ir} \rangle = \langle n_{ir} \rangle / \lambda_{0\ell}$, where $r \epsilon E_\ell$.

Finally the system mean system response-time per class-r, $\langle Ts_r \rangle$, is given by,

$$\langle Ts_r \rangle = \sum_{i=1}^{M} \langle n_{ir} \rangle / \lambda_{0\ell}$$

E.2 Approximation methods for general open queueing network with FCFS CENTRES.

B/ Gelenbe and Pujolle approximation [GELE,76]

$$f^{(m)}(\lambda_{ji}, C^2_{dji}) \longleftarrow \lambda_i^{-1} \sum_{j=0}^{M} \lambda_{ji} C^2_{dji}$$

$$f^{(d)}(\rho_i, C^2_{ai}, C^2_{si}) \longleftarrow \rho_i(1-\rho_i) + (1-\rho_i) C_{ai}{}^2 + \rho_i C^2_{si}{}^2$$

$$f^{(q)}(\rho_i, C^2_{ai}, C^2_{si}) \longleftarrow \rho_i \left[1 + \frac{\rho_i (C^2_{ai} + C^2_{si})}{2(1 - \rho_i)} \right]$$

C/ Sevcik et al approximation [SEVC,77b]

$$f^{(m)}(\lambda_{ji}, C^2_{dji}) \longleftarrow 1 + \sum_{j=0}^{M} \frac{\lambda^2_{ji}}{\lambda^2_i} (C^2_{dji}-1)$$

$$f^{(d)}(\rho_i, C^2_{ai}, C^2_{si}) \longleftarrow (1-\rho^2_i)C^2_{ai} + \rho^2_i C^2_{si}$$

$$f^{(q)}(\rho_i, C^2_{ai}, C^2_{si}) \longleftarrow \rho_i \left[1 + \frac{\rho_i(C^2_{ai}+C^2_{si})}{2(1-\rho_i)} \right]$$

D/ Reiser and Kobayashi approximation [REIS,74]

$$f^{(m)}(\lambda_{ji}, C^2_{dji}) \longleftarrow \lambda_i^{-1} \sum_{j=0}^{M} \lambda_{ji} C^2_{dji}$$

$$f^{(d)}(\rho_i, C^2_{ai}, C^2_{si}) \longleftarrow C^2_{si}$$

$$f^{(q)}(\rho_i, C^2_{ai}, C^2_{si}) \longleftarrow \rho_i \left[1-\exp\left\{ \frac{-2(1-\rho_i)}{\rho_i C^2_{ai}+C^2_{si}} \right\} \right]^{-1}$$

E3: Numerical results (chapter6)

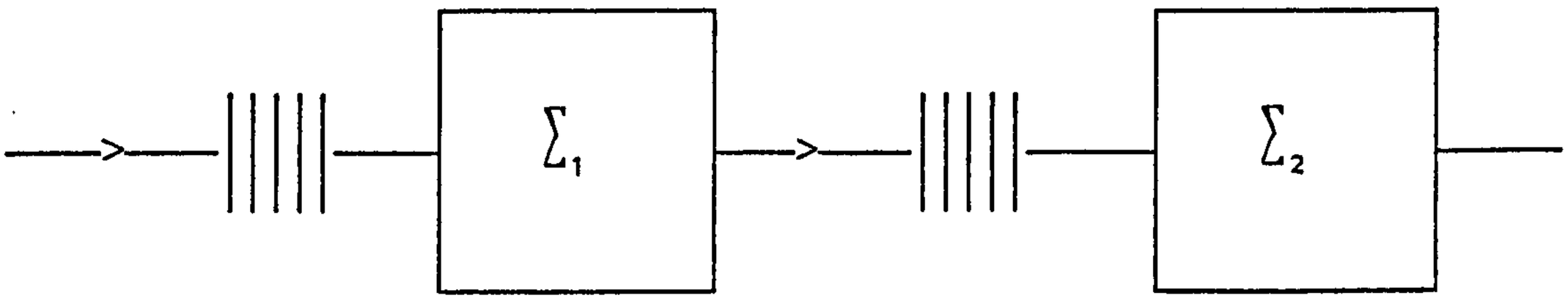

Fig. 6.1 Network of two queues in tandem.

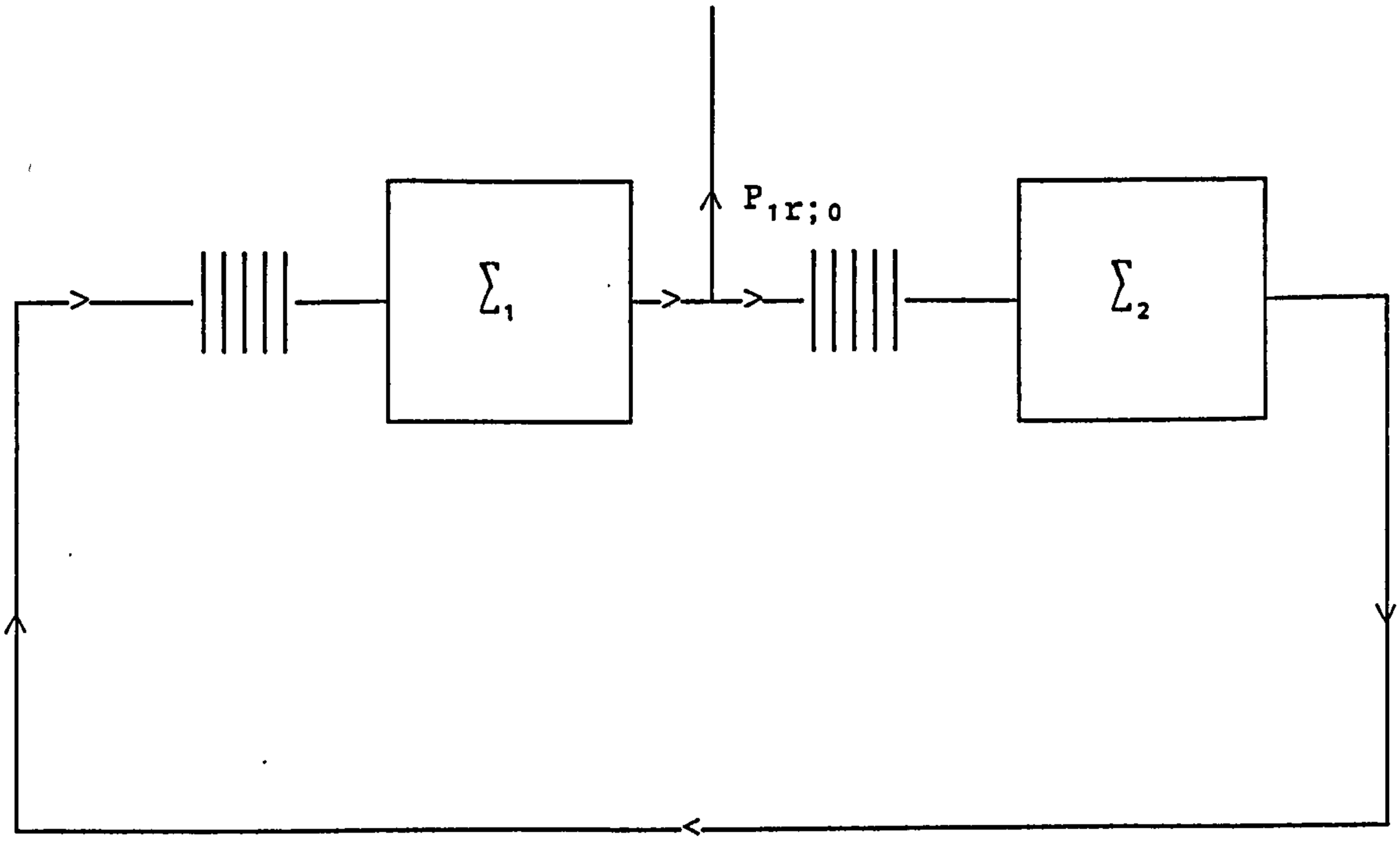

Fig. 6.2 Network with two queues in cycle.

Table 6.1: Raw data for the tandem Markovian network (c.f. Fig. 6.1) with 2 classes (PR-FCFS), M/M/1(PR) ——> ./M/1)(FCFS) (Results-Table 6.2)

Experiment No.	Class r	Raw data					
		λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}
1	1	0.4	1	2	1	2	1
	2	0.4	1	1	1	1	1
2	1	1	1	10	1	2	1
	2	8	1	10	1	24	1
3	1	5	1	10	1	15	1
	2	4	1	10	1	10	1
4	1	3	1	6	1	15	1
	2	2	1	10	1	12	1
5	1	2	1	4	1	10	1
	2	20	1	80	1	40	1
6	1	0.5	1	1	1	1.5	1
	2	20	1	60	1	60	1
7	1	0.8	1	1	1	1	1
	2	0.1	1	1	1	1	1
8	1	0.3	1	2	1	2	1
	2	0.4	1	1	1	1	1

Table 6.1 continued

Experiment No.	Class r	Raw data					
		λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}
9	1	6	1	10	1	12	1
	2	2	1	20	1	10	1
10	1	0.1	1	1	1	3	1
	2	0.05	1	1	1	1	1
11	1	7	1	10	1	14	1
	2	2	1	10	1	8	1
12	1	6	1	12	1	24	1
	2	6	1	120	1	12	1
13	1	0.1	1	1	1	1	1
	2	0.4	1	1	1	1	1
14	1	0.1	1	1	1	1	1
	2	0.4	1	2	1	2	1
15	1	0.1	1	2	1	2	1
	2	0.4	1	1	1	1	1

Table 6.2: Comparison of system mean response $\{<Ts_r>\}$ per priority class-r, r=1,2 for two (PR-FCFS) queues in tandem with exponential servers (Data- Table 6.1).

Exp. No.	Class r	$<Ts_r> = <T_{1r}> + <T_{2r}>$ SIM	MVA	%Dif.	m-ROA	%Dif	UME1	%DIF.	UME2.	%DIF
1	1	2.174	1.875	13.7	1.875	13.7	2.417	-11.2	2.389	-9.9
	2	5.208	5.31	-1.95	5.25	-0.8	5.147	1.17	5.09	2.26
2	1	2.135	3.11	-45.6	3.11	-45.6	2.202	3.13	2.198	-2.95
	2	2.749	1.36	50.5	1.31	52.3	2.747	0.07	2.742	0.25
3	1	0.472	0.45	4.66	0.45	4.66	0.557	-18.6	0.547	-15.88
	2	2.431	2.37	2.51	1.93	20.6	2.427	0.16	2.412	0.78
4	1	0.433	0.438	-1.15	0.438	-1.1	0.449	-3.69	0.449	-3.69
	2	1.075	1.02	5.11	0.793	26.2	1.043	2.97	1.042	3.07
5	1	0.835	0.833	0.24	0.833	0.24	0.912	-9.22	0.915	-9.58
	2	1.591	1.13	28.97	0.22	86.17	1.509	5.15	1.514	4.84
6	1	3.4	4.0	17.64	4.0	17.64	3.682	-8.28	3.683	-8.32
	2	7.924	6.15	22.38	0.296	96.2	7.461	5.84	7.466	5.78
7	1	14.055	15.0	-6.72	15.0	-6.72	15.34	-9.14	15.62	-11.13
	2	57.94	60.0	-3.55	46.67	19.4	60.69	-4.74	61.69	-6.47
8	1	1.955	1.69	13.55	1.69	13.55	2.171	-11.0	2.151	-10.02
	2	4.488	4.64	-3.38	4.60	-2.49	4.532	-0.9	4.49	-0.04

Table 6.2 continued

$$\langle Ts_r \rangle = \langle T_{1r} \rangle + \langle T_{2r} \rangle$$

Exp. No.	Class r	SIM	MVA	%Dif.	m-ROA	%Dif	UME1	%DIF.	UME2.	%DIF
9	1	0.534	0.527	1.31	0.527	1.31	0.553	-3.5	0.555	-3.93
	2	1.11	1.0	9.9	0.83	25.2	1.0	9.9	1.014	8.64
10	1	1.51	1.47	2.64	1.47	2.64	1.51	0.0	1.511	-0.06
	2	2.42	2.398	0.9	2.391	1.19	2.37	2.06	2.374	1.9
11	1	0.66	0.62	6.06	0.62	6.06	0.72	-9.09	0.74	-12.12
	2	3.84	3.83	0.26	2.9	24.48	3.83	0.26	3.86	-0.52
12	1	0.45	0.33	26.66	0.33	26.66	0.47	-4.44	0.485	-7.77
	2	0.66	0.537	18.63	0.44	33.33	0.58	12.12	0.599	9.24
13	1	2.82	3.11	-10.3	3.11	-10.3	3.14	-11.34	3.12	-10.63
	2	4.27	4.22	1.17	4.15	2.81	4.3	-0.7	4.25	0.46
14	1	2.33	2.54	-9.01	2.54	-9.01	2.40	-3.0	2.39	-2.57
	2	1.76	1.58	10.22	1.54	12.5	1.68	4.54	1.67	5.11
15	1	1.65	1.43	13.33	1.43	13.33	1.8	-9.09	1.8	-9.09
	2	3.63	3.68	-1.37	3.67	-1.1	3.65	-0.55	3.64	-0.27

Table 6.3: Raw data for the tandem Markovian network (c.f. Fig. 6.1) with 2 classes (HOL-HOL), M/M/1(HOL) ——> ./M/1(HOL) (Results-Table 6.4)

Experiment No.	Class r	Raw data					
		λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}
1	1	0.1	1	1	1	1	1
	2	0.4	1	1	1	1	1
2	1	0.1	1	1	1	1	1
	2	0.4	1	2	1	2	1
3	1	0.2	1	1	1	1	1
	2	0.4	1	2	1	2	1
4	1	0.1	1	2	1	2	1
	2	0.4	1	1	1	1	1
5	1	0.2	1	2	1	2	1
	2	0.4	1	1	1	1	1
6	1	0.3	1	2	1	2	1
	2	0.4	1	1	1	1	1
7	1	0.4	1	2	1	2	1
	2	0.4	1	1	1	1	1

Table 6.4*: Comparison of system mean response time $\{Ts_r\}$ per priority class r,1,2, for the two queues in tandem , (c.f. Fig.6.1) M/M/1 (R=2,HOL) —>./M/1 (R=2,HOL), (Data - Table 6.3).

Exp. No.	Class r	$\langle Ts_r \rangle = \langle T_{1r} \rangle + \langle T_{2r} \rangle$ EXACT	MVA	%Dif.	sd-ROA	%Dif	UME1	%DIF.	UME2.	%DIF
1	1	3.085	3.11	-0.84	3.11	-0.84	3.11	-0.87	3.109	-0.81
	2	4.212	4.22	-0.19	4.209	0.14	4.364	-3.6	4.247	-0.83
2	1	2.408	2.44	-1.33	2.44	-1.33	2.44	-1.33	2.44	-1.33
	2	1.68	1.635	2.68	1.635	2.68	1.657	1.37	1.643	2.20
3	1	2.709	2.75	-1.51	2.75	-1.51	2.75	-1.51	2.746	-1.64
	2	2.349	2.25	4.2	2.246	4.4	2.3	2.08	2.277	3.06
4	1	1.97	1.894	3.8	1.895	3.81	1.895	3.81	1.895	3.81
	2	3.618	3.626	-0.2	3.624	-0.17	3.703	2.35	3.631	-0.36
5	1	2.073	2.0	3.52	1.998	3.61	2.0	3.52	2.0	3.52
	2	3.98	4.0	-0.5	3.992	-0.3	4.103	-3.09	4.011	-0.78
6	1	2.187	2.117	3.2	2.118	3.16	2.118	3.16	2.117	3.2
	2	4.442	4.48	-0.85	4.464	-0.5	4.627	-1.48	4.508	-1.48
7	1	2.315	2.25	2.8	2.248	2.89	2.25	2.8	2.246	2.98
	2	5.041	5.125	-1.66	5.076	-0.7	5.328	-5.69	5.171	-2.58

* Note that the exact and the sd-ROA results are taken from [SCHM,83].

Table 6.5: Raw data for the tandem general network (c.f. Fig. 6.1) with 2 classes (PR-FCFS), G/G/1(PR) ——> ./G/1(FCFS) (Results - Table 6.6)

Experiment No.	Class r	Raw data λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}
1	1	1	18	4	25	4	1
	2	1	30	10	14	2	1
2	1	1	2	10	2	5	2
	2	3	9	10	9	5	2
3	1	0.1	9	1	2	3	3
	2	0.05	2	1	3	1	2
4	1	3	2	6	1	15	1
	2	2	5	10	2	4	1
5	1	2	2	10	5	4	2
	2	3	5	6	2	10	2
6	1	1	2	4	3	5	3
	2	1	4	10	2	2	1
7	1	3	10	6	5	15	1
	2	2	15	10	2	4	1
8	1	0.2	5	0.4	3	1	2
	2	2	2	8	2	4	3

Table 6.5 continued

Experiment No.	Class r	Raw data λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}
9	1	0.45	3	0.9	5	1	2
	2	0.5	7	5	2	1	3
10	1	0.27	5	0.9	2	1	3
	2	0.5	2	5	3	1	4
11	1	3	0.5	6	3	15	7
	2	4	0.5	10	5	4	3
12	1	4	0.5	12	3	8	6
	2	2	0.5	5	1	6	9
13	1	8	0.5	10	0.5	32	0.5
	2	1	0.5	10	0.5	2	0.5
14	1	2.5	0.5	5	0.5	7.5	0.5
	2	1.5	0.5	5	0.5	3	0.5
15	1	5	1	12.5	1/3	7.5	0.5
	2	4	1	10	1/3	16	0.5

Table 6.6: Comparison of system mean response time $\{<Ts_r>\}$ per priority class-r, r=1,2 for two (PR-FCFS) queues in tandem with general servers G/G/1(PR)—>./G/1(FCFS) (data - Table 6.5)

Exp. No.	Class r	$<Ts_r> = <T_{1r}> + <T_{2r}>$ SIM	UME1	%DIF.	UME2.	%DIF
1	1	8.177	8.337	-1.95	8.879	-8.50
	2	11.695	9.66	17.4	11.29	3.46
2	1	3.712	3.261	12.15	3.603	2.95
	2	5.34	4.593	13.98	5.048	5.47
3	1	6.259	7.32	-16.95	7.322	-16.98
	2	3.981	4.172	-4.79	4.197	-5.42
4	1	1.576	1.204	23.6	1.616	-2.53
	2	3.813	2.86	24.99	3.519	7.71
5	1	2.221	2.169	2.34	2.218	0.13
	2	4.158	3.886	6.54	3.941	5.22
6	1	3.127	2.856	8.66	3.103	0.77
	2	4.536	3.697	18.49	4.092	9.78
7	1	5.04	3.56	29.36	5.116	-1.5
	2	12.494	8.63	30.92	11.12	10.99
8	1	29.34	30.67	-4.51	30.82	-5.04
	2	62.39	57.0	8.64	57.24	8.25

Table 6.6 continued

Exp. No.	Class r	$\langle Ts_r \rangle = \langle T_{1r} \rangle + \langle T_{2r} \rangle$ SIM	UME1	%DIF.	UME2.	%DIF
9	1	74.8	71.24	4.76	86.81	-16.05
	2	90.76	78.74	13.24	95.87	-5.6
10	1	16.8	18.93	-12.93	19.1	-13.69
	2	17.27	16.73	3.15	16.98	1.7
11	1	1.683	1.78	-5.76	1.77	-5.16
	2	3.31	3.18	3.92	3.17	4.22
12	1	2.995	3.243	-8.28	3.21	-7.17
	2	3.851	4.008	-4.07	3.95	-2.57
13	1	1.329	1.286	3.23	1.254	5.6
	2	4.514	4.221	6,49	4.173	7.55
14	1	1.143	1.338	-17.06	1.199	-4.89
	2	2.262	2.44	-7.86	2.255	0.37
15	1	1.114	1.134	-1.79	1.114	0.0
	2	1.581	1.524	3.60	1.498	5.25

•:Note that GE is used for C_{ar} or C_{sr} >1, Erlang-2 for C_a^2, C_s^2 = 0.5, uniform for C_a^2, C_s^2 = 1/3 and deterministic for C_a^2, C_s^2 = 0.

Table 6.7: Raw data for the Cyclic Markovian network (c.f. Fig. 6.2) with 2 or 3 classes and different service disciplines.
(Results - Table 6.8a-b)

Exp. No.	Service Discipline $\sum_1$	$\sum_2$	Class r	Raw data λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}	$P_{1r;0}$
1	PR	FCFS	1	1	1	6.25	1	5	1	0.2
			2	1	1	25	1	15	1	0.4
2	PR	PR	1	1	1	6.25	1	5	1	0.2
			2	1	1	25	1	15	1	0.4
3	PR	HOL	1	3	1	25	1	23.33	1	0.3
			2	2	1	12.5	1	6	1	0.4
4	HOL	FCFS	1	1	1	6.25	1	5	1	0.2
			2	1	1	25	1	15	1	0.4
5	HOL	FCFS	1	3	1	50	1	25	1	0.2
			2	4	1	30	1	20	1	0.4
6	HOL	HOL	1	3	1	50	1	25	1	0.2
			2	4	1	30	1	20	1	0.4

Table 6.7 continued

Exp. No.	Service Discipline $\sum_1$	$\sum_2$	Class r	Raw data λ_{0r}	$C^2_{a_0 r}$	μ_{1r}	$C^2_{s_1 r}$	μ_{2r}	$C^2_{s_2 r}$	$P_{1r;0}$
7	PR	HOL	1	0.4	1	4	1	5	1	0.2
			2	0.15	1	1.875	1	0.9	1	0.4
			3	0.1	1	2	1	1	1	0.5
8	PR	FCFS	1	1	1	12.5	1	7.5	1	0.4
			2	1.5	1	15	1	5	1	0.5
			3	0.9	1	22.5	1	9	1	0.2
9	FCFS	HOL	1	3	1	12	1	10	1	0.5
			2	5	1	41.667	1	25	1	0.4
			3	1	1	33.333	1	23.333	1	0.3
10	PR	FCFS	1	3	1	12	1	10	1	0.5
			2	5	1	41.667	1	25	1	0.4
			3	1	1	33.333	1	23.333	1	0.3

Table 6.8a: Comparison of mean queue lengths for centre-$_1$ and percentage difference from simulation per class-r, R = 2 or 3, for a cyclic queueing network (Fig. 6.2) with exponential servers (Data - Table 6.7).

Exp. No.	Class r	$\langle n_{1r} \rangle$ SIM	ROA	m-ROA	MVA	ME2-ROA	UME2
1	1	4.525	3.999 (11.62%)	3.999 (11.62%)	3.999 (11.62%)	4.29 (5.19%)	3.98 (-12.04)
	2	18.89	0.999 (94.71%)	6.333 (66.47%)	16.99 (10.05%)	18.18 (3.75%)	17.24 (8.73%)
2	1	3.951	3.999 (-1.2%)	3.999 (-1.2%)	3.999 (-1.2%)	4.0 (-1.24%)	3.999 (-1.2%)
	2	20.55	0.999 (95.14%)	6.333 (69.18%)	16.999 (17.32%)	17.71 (13.82%)	17.71 (13.82%)
3	1	0.732	———	———	0.667 (8.88%)	0.653 (10.79%)	0.653 (10.79%)
	2	2.972	———	———	2.666 (10.29%)	2.748 (7.53%)	2.748 (7.53%)
4	1	4.617	———	———	4.09 (11.41%)	4.38 (5.13%)	4.07 (11.84%)
	2	18.96	———	———	16.59 (12.5%)	17.738 (6.44%)	16.837 (11.19%)
5	1	0.672	———	———	0.666 (0.9%)	0.668 (0.6%)	0.673 (-0.14%)
	2	0.988	———	———	1.0 (-1.2%)	0.999 (-1.11%)	1.002 (-1.4%)
6	1	0.64	———	———	0.666 (-4.06%)	0.634 (0.9%)	0.634 (0.9%)
	2	1.062	———	———	1.0 (5.83%)	1.079 (-1.6%)	1.079 (-1.6%)

Table 6.8a. continued

Exp. No.	Class r	$\langle n_{1,r} \rangle$ SIM	ROA	m-ROA	MVA	ME2-ROA	UME2
7	1	1.134	———	———	1.0 (11.81%)	1.071 (5.55%)	1.071 (5.51%)
	2	1.346	———	———	0.979 (27.26%)	1.01 (24.9%)	1.01 (24.9%)
	3	1.922	———	———	1.272 (33.78%)	1.333 (30.6%)	1.532 (20.25%)
8	1	0.256	0.25 (2.34%)	0.25 (2.34%)	0.25 (2.34%)	0.253 (1.17%)	0.253 (1.17%)
	2	0.438	0.333 (23.97%)	0.410 (6.39%)	0.433 (1.14%)	0.436 (0.45%)	0.429 (2.05%)
	3	1.102	0.5 (54.62%)	0.854 (22.50%)	1.05 (4.71%)	1.067 (3.17%)	1.083 (1.17%)
9	1	3.404	———	———	5.0 (-46.8%)	3.717 (-9.19%)	3.655 (-7.37%)
	2	6.246	———	———	3.0 (51.96%)	6.999 (-12.05%)	6.931 (-10.96%)
	3	1.762	———	———	1.0 (43.24%)	1.88 (-6.69%)	1.874 (-6.35%)
10	1	1.024	1.0 (2.34%)	1.0 (2.34%)	1.0 (2.34%)	1.027 (-0.29%)	0.994 (2.93%)
	2	5.878	1.5 (74.48%)	3.189 (45.74%)	6.708 (-14.1%)	6.96 (-18.4%)	6.936 (-17.99%)
	3	8.148	1.0 (87.77%)	5.03 (38.26%)	9.14 (-12.2%)	9.45 (-15.97%)	9.746 (-19.59%)

Table 6.8b: Comparison of mean queue lengths for centre-2 per class-r, R = 2 or 3 , and percentage difference from simulation for a cyclic queueing network (Fig. 6.2) with exponential servers (Data - Table 6.7).

Exp. No.	Class r	$\langle n_{2r} \rangle$ SIM	ROA	m-ROA	MVA	ME2-ROA	UME2
1	1	7.924	7.999 (-0.9%)	7.999 (-0.9%)	7.999 (-0.9%)	8.56 (-8.02%)	7.85 (0.9%)
	2	4.079	0.999 (75.51%)	0.999 (75.51%)	0.999 (75.51%)	2.97 (27.18%)	2.89 (29.15%)
2	1	3.908	3.999 (-2.3%)	3.999 (-2.3%)	3.999 (-2.3%)	4.0 (-2.35%)	3.999 (-2.3%)
	2	17.21	0.999 (94.19%)	5.799 (66.3%)	12.999 (24.46%)	14.56 (15.39%)	14.56 (15.39%)
3	1	1.007	———	———	1.262 (-25.3%)	1.258 (-24.9%)	1.258 (-24.9%)
	2	2.73	———	———	2.561 (6.19%)	2.669 (2.23%)	2.669 (2.23%)
4	1	7.608	———	———	7.999 (-5.14%)	8.299 (-9.08%)	7.50 (1.42%)
	2	3.887	———	———	0.999 (74.3%)	2.888 (25.7%)	2.789 (28.2%)
5	1	2.255	———	———	2.18 (3.32%)	2.34 (-3.77%)	2.31 (-2.44%)
	2	1.255	———	———	1.36 (-8.4%)	1.23 (2.0%)	1.24 (1.19%)
6	1	1.259	———	———	1.269 (-0.79%)	1.225 (2.7%)	1.225 (2.7%)
	2	2.717	———	———	2.093 (22.96%)	2.120 (21.97%)	2.121 (21.93%)

Table 6.8b. continued

Exp. No.	Class r	$\langle n_{2r} \rangle$ SIM	ROA	m-ROA	MVA	ME2-ROA	UME2
7	1	1.135	———	———	1.359 (-19.7%)	1.373 (-20.9%)	1.373 (-20.9%)
	2	0.613	———	———	0.589 (3.9%)	0.609 (-0.6%)	0.609 (-0.6%)
	3	0.568	———	———	0.411 (27.64%)	0.444 (21.83%)	0.484 (14.78%)
8	1	1.871	2.0 (-6.89%)	2.0 (-6.89%)	2.0 (-6.89%)	2.29 (-22.3%)	2.26 (-20.79%)
	2	2.062	3.0 (-45.5%)	3.0 (-45.5%)	3.0 (-45.5%)	2.39 (-15.9%)	2.362 (-14.5%)
	3	4.963	4.0 (19.4%)	4.0 (19.4%)	4.0 (19.4%)	5.43 (-9.40%)	5.399 (-8.8%)
9	1	0.481	———	———	0.498 (-3.53%)	0.517 (7.48%)	0.489 (-1.66%)
	2	1.619	———	———	1.539 (4.94%)	1.662 (-2.65%)	1.665 (-2.84%)
	3	0.936	———	———	1.0 (-6.8%)	1.065 (-13.78%)	1.062 (-13.35%)
10	1	0.719	1.0 (-39.1%)	1.0 (-39.1%)	1.0 (-39.1%)	0.860 (-19.75%)	0.816 (-13.49%)
	2	1.816	1.0 (44.93%)	1.0 (44.93%)	1.0 (44.93%)	1.702 (6.27%)	1.919 (-5.67%)
	3	0.727	0.333 (54.15%)	0.333 (54.15%)	0.333 (51.15%)	0.519 (28.16%)	0.649 (10.72%)

Table 6.9: Raw data for the Cyclic General network (c.f. Fig. 6.2) with 2 or 3 classes and different service disciplines.

(Results - Table 6.10a-b)

No.	Service Discipline $\sum_1$ $\sum_2$	Class r	Raw data λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}	$P_{1r;0}$
1	PR HOL	1	3	2	25	4	23.33	3	0.3
		2	2	6	12.5	5	6	5	0.4
2	FCFS PR	1	2	24	10	25	4	13	0.5
		2	1	12	4	17	2.5	17	0.5
3	L^*-N-P HOL	1	3	0.5	50	4	17.5	1	0.3
		2	2	0.5	12.5	3	6	1	0.5
4	HOL LCFS	1	1	2	6.667	3	3.333	4	0.5
		2	1	4	6.667	2	3.333	3	0.5
5	FCFS HOL	1	0.4	3	4	3	5	2	0.2
		2	0.15	5	1.875	3	0.9	2	0.4
		3	0.1	1	2	5	1	4	0.5
6	PR HOL	1	0.4	3	4	9	2	5	0.4
		2	0.15	5	1.875	3	0.75	2	0.5
		3	0.1	1	2	5	1.6	3	0.2

Table 6.9 continued

No.	Service Discipline $\sum_1$	$\sum_2$	Class r	Raw data λ_{0r}	C^2_{a0r}	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}	$P_{1r;0}$
7	HOL	FCFS	1	2	15	10	10	6	9	0.4
			2	1.5	6	15	2	7.5	14	0.5
			3	1	3	25	7	40	18	0.5
8	PR	HOL	1	0.4	0.5	4	0.5	2	0.5	0.4
			2	0.15	0.5	1.875	0.5	0.75	0.5	0.5
			3	0.1	0.5	2	0.5	1.6	0.5	0.2
9	PR	L*-N-P	1	1	0.5	5	0	10	2	0.5
			2	2	0.5	8.333	0	1.666	3	0.8
			3	1	0.5	10	0	5	4	0.5
10	PS	HOL	1	1.5	1	12.5	0	11.25	0.5	0.4
			2	1.2	1	20.0	0	14	0.5	0.5
			3	2.5	1	62.5	0	50	0.5	0.2

*: L-N-P is the LCFSNONPR service discipline.

Table 6.10a: Comparison of mean queue lengths for centre-1 and percentage difference from simulation per class-r, R = 2 or 3, for a cyclic queueing network (Fig. 6.2), with general service times (Data - Table 6.9).

Exp. No.	Class r	$\langle n_{1r} \rangle$ SIM	ME2-ROA	UME2
1	1	1.435	1.318 (8.15%)	1.307 (8.91%)
	2	11.64	8.134 (30.12%)	9.814 (15.68%)
2	1	138.4	135.2 (2.3%)	160.3 (-15.88%)
	2	80.34	68.11 (15.2%)	82.163 (-2.26%)
3	1	1.13	———	1.453 (-28.38%)
	2	0.693	———	0.75 (-8.37%)
4	1	0.793	———	0.853 (-7.56%)
	2	1.773	———	1.933 (-9.02%)
5	1	7.084	7.42 (-4.74%)	8.30 (-17.16%)
	2	1.931	1.475 (23.61%)	1.72 (10.92%)
	3	0.866	0.776 (10.39%)	0.859 (0.81%)

Table 6.10a. continued

Exp. No.	Class r	$\langle n_1 r \rangle$ SIM	ME2-ROA	UME2
6	1	0.728	0.783 (-7.55%)	0.834 (-14.56%)
	2	0.716	0.596 (16.75%)	0.767 (-15.77%)
	3	2.713	3.217 (-18.57%)	3.141 (-15.77%)
7	1	6.162	5.338 (13.37%)	6.537 (-6.08%)
	2	12.32	8.926 (27.5%)	10.62 (13.79%)
	3	68.64	73.45 (-6.89%)	88.257 (-28.52%)
8	1	0.291	0.275 (5.49%)	0.2588 (11.34%)
	2	0.265	0.271 (-2.14%)	0.235 (11.32%)
	3	0.753	1.002 (33.06%)	0.902 (-19.65%)
9	1	0.574	———	0.498 (13.24%)
	2	1.452	———	1.041 (28.3%)
	3		———	4.326 (26.2%)
10	1	0.75	———	0.806 (-7.46%)
	2	0.282	———	0.323 (-14.53%)
	3	0.562	———	0.563 (-0.3%)

Table 6.10b: Comparison of mean queue lengths for centre-2 and percentage difference from simulation per class-r, R = 2 or 3, for a cyclic queueing network (Fig. 6.2), with general service times (Data - Table 6.9).

Exp. No.	Class r	$\langle n_{2r} \rangle$ SIM	ME2-ROA	UME2
1	1	2.65	3.197 (-20.64%)	3.194 (-20.52%)
	2	8.971	8.264 (7.80%)	8.627 (3.83%)
2	1	10.58	9.73 (3.56%)	10.36 (-2.6%)
	2	83.24	72.89 (12.43%)	76.58 (8.0%)
3	1	1.306	———	1.501 (-14.93%)
	2	1.473	———	1.40 (4.95%)
4	1	1.013	———	0.828 (18.26%)
	2	0.928	———	0.946 (-1.93%)
5	1	2.219	2.418 (-8.9%)	2.443 (-10.09%)
	2	1.239	0.921 (25.66%)	0.988 (20.25%)
	3	0.847	0.696 (17.92%)	0.723 (14.74%)

Table 6.10b. continued

Exp. No.	Class r	$\langle n_1 r \rangle$ SIM	ME2-ROA	UME2
6	1	1.249	1.428 (-14.33%)	1.457 (-16.65%)
	2	0.897	0.759 (15.38%)	0.851 (5.11%)
	3	4.211	4.950 (-17.54%)	4.885 (-16.00%)
7	1	11.27	11.478 (-1.84%)	13.03 (-15.6%)
	2	6.359	5.58 (12.22%)	6.415 (-0,88%)
	3	15.32	14.37 (6.2%)	16.857 (-9.98%)
8	1	0.596	0.693 (-15.8%)	0.633 (-6.12%)
	2	0.362	0.371 (-2.48%)	0.351 (3.03%)
	3	1.308	1.624 (-24.15%)	1.529 (-16.89%)
9	1	0.996	———	1.273 (-27.8%)
	2	0.836	———	0.877 (-4.81%)
	3	1.468	———	1.435 (2.24%)
10	1	0.263	———	0.26 (1.15%)
	2	0.13	———	0.13 (0.0%)
	3	0.57	———	0.792 (-28.03%)

APPENDIX F

F1: Numerical results (chapter 7)

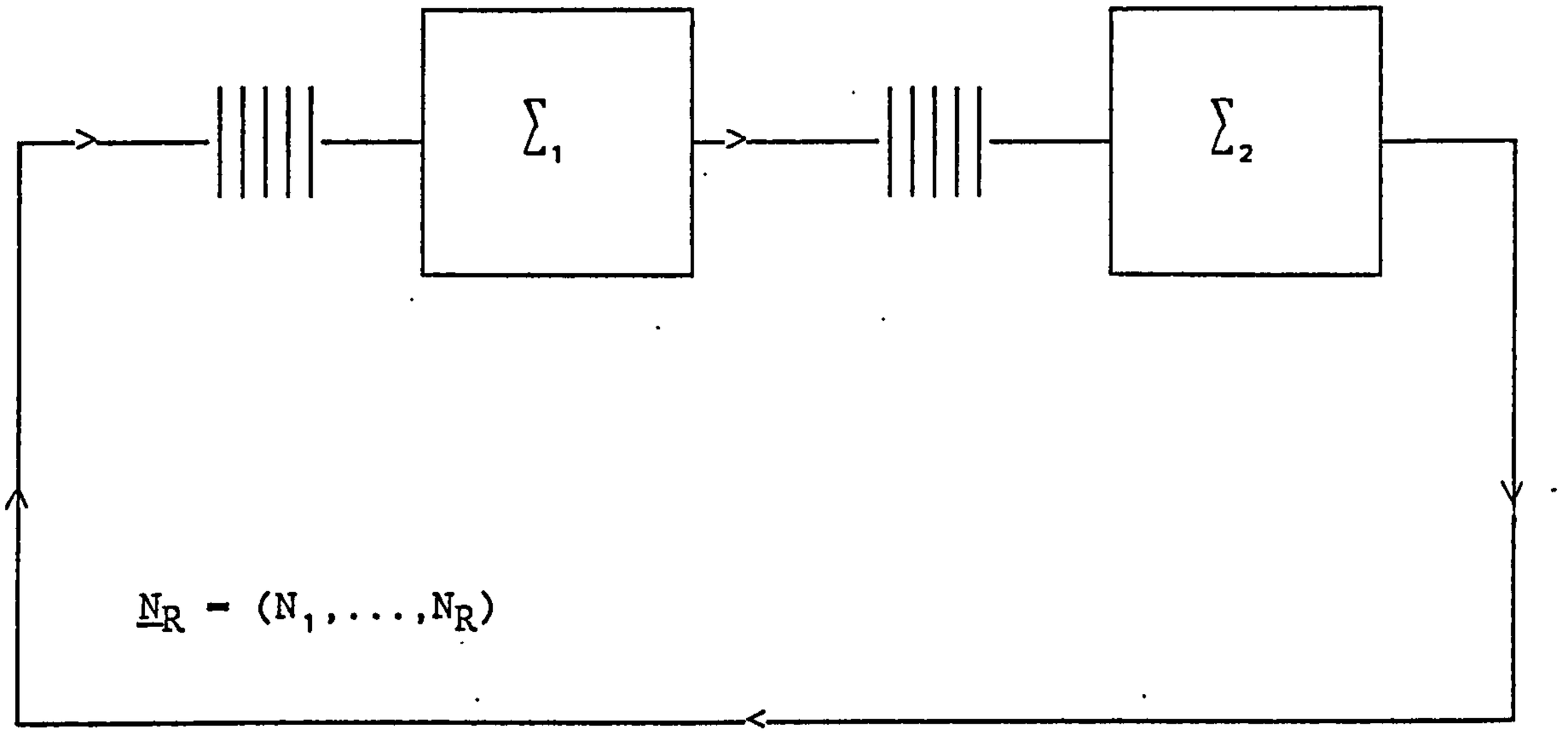

Fig. 7.1 Two stage cyclic closed queueing network.

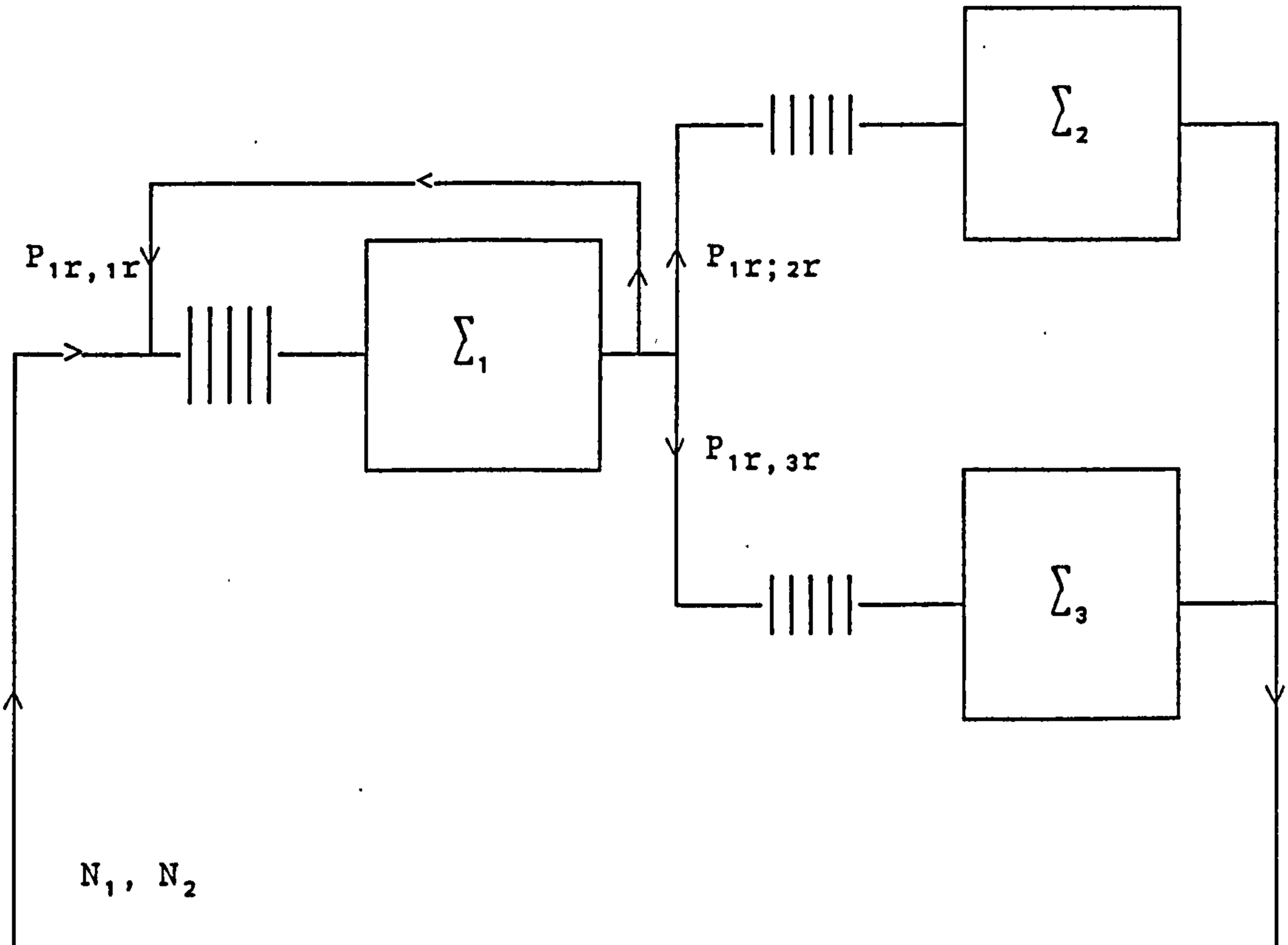

Fig. 7.2 General central server model .

Table 7.1 Raw data for two stage-cyclic Markovian network (PR ——> FCFS), with R = 2, N_1 = 2 and N_2 variable (Fig. 7.1) , (results - Table 7.2a-7.3c).

Model No.	class r	$\sum_1$ (PR)		$\sum_2$ (FCFS)	
		μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}
1	1	50	1	20	1
	2	5	1	20	1
2	1	10	1	10	1
	2	10	1	10	1
3	1	5	1	10	1
	2	50	1	10	1

Table 7.2a: System throughputs and relative errors for model 1. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6. (Data table 7.1).

N_2	Class r	ROA	m-ROA	MVA	m-MVA	EXACT	UME2
1	1	16.028 (-1.9%)	16.046 (-1.85%)	15.981 (-2.25%)	15.941 (-2.5%)	16.35	16.33 (-0.12%)
	2	2.359 (11.67%	2.337 (10.6%)	2.367 (12.04%)	2.122 (0.43%)	2.113	2.193 (3.78%)
2	1	15.494 (-0.99%	15.502 (-0.94%)	15.188 (-2.94%)	15.118 (-3.39%)	15.65	15.66 (0.06%)
	2	3.145 (6.16%)	3.135 (5.81%)	3.187 (7.58%)	2.995 (1.10%)	2.963	3.033 (2.36%)
3	1	15.33 (-0.19%	15.332 (-0.17%)	14.886 (-3.08%)	14.798 (-3.65%)	15.36	15.414 (0.35%)
	2	3.385 (2.72%)	3.381 (2.63%)	3.473 (3.473%)	3.333 (1.17%)	3.295	3.332 (1.45%)
4	1	15.286 (0.17%)	15.287 (0.17%)	14.777 (-3.16%)	14.681 (-3.79%)	15.26	15.329 (0.45%)
	2	3.451 (0.99%)	3.450 (0.97%)	3.57 (4.47%)	3.456 (1.16%)	3.417	3.343 (-0.53%)
5	1	15.276 0.37%)	15.276 (0.37%)	14.74 (-3.15%)	14.64 (-3.8%)	15.22	15.27 (0.33%)
	2	3.467 (0.24%)	3.467 (0.24%)	3.603 (4.16%)	3.50 (1.18%)	3.459	3.463 (0.4%)
6	1	15.273 (0.42%)	15.273 (0.42%)	14.726 (-3.17%)	14.626 (-3.83%)	15.21	15.263 0.35%)
	2	3.471 (0.04%)	3.471 (0.04%)	3.615 (4.11%)	3.515 (1.23%)	3.473	3.471 (-0.05%)

Table 7.2b: System throughputs and relative errors for model 2. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6. (Data table 7.1).

N_2	Class r	ROA	m-ROA	MVA	m-MVA	EXACT	UME2
1	1	5.917 (-4.04%	6.061 (-1.7%)	5.827 (-5.5%)	5.833 (-5.11%)	6.167	6.223 (0.9%)
	2	2.247 (68.6%)	1.815 (36.15%)	1.808 (35.62%)	1.428 (7.17%)	1.333	1.389 (4.2%)
2	1	5.346 (-6.3%)	5.504 (-3.5%)	5.140 (-9.89%)	5.188 (-9.05%)	5.705	5.793 (1.54%)
	2	3.775 (64.48%	3.0113 (31.21%)	3.044 (32.63%)	2.493 (8.63%)	2.295	2.436 (6.14%)
3	1	4.905 (-7.47%	5.054 (-4.67%)	4.583 (-13.55%)	4.669 (-12.04%)	5.302	5.407 (1.98%)
	2	4.141 (36.62%	3.847 (26.92%)	3.940 (30.0%)	3.329 (8.63%)	3.031	3.23 (6.56%)
4	1	4.548 (-8.16%	4.684 (-5.42%)	4.128 (-16.65%)	4.241 (-14.37%)	4.953	5.072 (2.42%)
	2	4.703 (29.99%	4.4622 (23.33%)	4.623 (27.78%)	4.005 (10.71%)	3.618	3.847 (6.32%)
5	1	4.251 (-8.55%	4.374 (-5.91%)	3.752 (-19.29%)	3.881 (-16.5%)	4.649	4.786 (2.94%)
	2	5.142 (25.38%	4.941 (20.47%)	5.162 (25.88%)	5.142 (11.29%)	4.101	4.338 (5.77%)
6	1	3.997 (-8.78%	4.109 (-6.24%)	3.436 (-21.59%)	3.575 (-18.42%)	4.383	4.541 (3.60%)
	2	5.498 (22.01%	5.327 (18.22%)	5.60 (24.28%)	5.033 (11.7%)	4.506	4.735 (5.08%)

Table 7.2c: System throughputs and relative errors for model 3. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6. (Data table 7.1).

N_2	Class r	ROA	m-ROA	MVA	m-MVA	EXACT	UME2
1	1	3.90 (-6.14%	4.169 (-0.32%)	4.067 (-2.12%)	4.013 (-3.43%)	4.156	4.237 (1.9%)
	2	4.119 (319.2%	1.345 (36.96%)	1.398 (42.29%)	1.378 (40.25%)	0.982	0.787 (-19.8%)
2	1	3.391 (-14.4%	3.857 (-2.67%)	3.678 (-7.21%)	3.713 (-6.33%)	3.964	4.097 (3.35%)
	2	6.008 (206.7%	3.388 (72.95%)	2.966 (51.42%)	2.652 (25.37%)	1.959	1.564 (-20.16%)
3	1	2.929 (-21.3%	3.436 (-7.73%)	3.222 (-13.48%)	3.379 (-9.27%)	3.725	3.925 (5.36%)
	2	6.935 (137.6%	5.212 (78.63%)	4.455 (52.67%)	3.879 (32.95%)	2.918	2.362 (-19.05%)
4	1	2.558 (-25.9%	2.977 (-13.8%)	2.799 (-18.96%)	3.025 (-12.42%)	3.454	3.74 (8.28%)
	2	7.414 (93.02%	6.491 (68.98%)	5.777 (50.41%)	5.034 (31.97%)	3.841	3.189 (-16.97%)
5	1	2.268 (-28.3%	2.564 (-18.98%)	2.447 (-22.66%)	2.675 (-15.45%)	3.165	3.555 (12.32%)
	2	7.726 (64.1%)	7.261 (54.23%)	6.815 (44.75%)	6.0434 (28.36%)	4.708	4.007 (-14.89%)
6	1	2.037 (-29.1%	2.235 (-22.28%)	2.167 (-24.57%)	2.358 (-17.92%	2.874	3.363 (17.01%)
	2	7.961 (44.86%	7.713 (40.35%)	7.519 (36.81%)	6.837 (24.4%)	5.496	4.776 (-13.10%)

Table 7.3a: Mean queue lengths at FCFS centre and relative errors for model 1. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6. (Data table 7.1).

N_2	Class r	ROA	m-ROA	MVA	m-MVA	EXACT	UME2
1	1	1.604 (1.02%)	1.603 (0.98%)	1.600 (0.78%)	1.601 (0.85%)	1.588	1.589 (0.06%)
	2	0.305 (-3.62%	0.302 (-4.5%)	0.300 (-5.17%)	0.318 (0.44%)	0.316	0.302 (-4.4%)
2	1	1.614 (0.39%)	1.614 (0.39%)	1.624 (1.0%)	1.625 (1.11%)	1.608	1.609 (0.06%)
	2	0.458 (-7.89%	0.456 (-8.22%)	0.462 (-7.05%)	0.497 (0.0%)	0.497	0.477 (-4.02%)
3	1	1.617 (0.08%)	1.617 (0.08%)	1.633 (1.04%)	1.635 (1.18%)	1.616	1.616 (0.0%)
	2	0.520 (-11.6%	0.519 (-11.76%)	0.536 (-8.99%)	0.583 (-1.05%)	0.589	0.56 (-4.92%)
4	1	1.618 (-0.08%	1.618 (-0.08%)	1.636 (1.05%)	1.638 (1.19%)	1.619	1.618 (-0.08%)
	2	0.541 (-14.1%	0.541 (-14.1%)	0.565 (-10.26%)	0.619 (-1.76%)	0.630	0.593 (-6.23%)
5	1	1.618 (-0.08%	1.618 (-0.08%)	1.637 (1.05%)	1.639 (1.2%)	1.62	1.619 (-0.06%)
	2	0.548 (-15.3%	0.548 (-15.36%)	0.576 (-10.90%)	0.633 (-2.16%)	0.647	0.610 (-5.71%)
6	1	1.618 (-0.08%	1.618 (-0.08%)	1.637 (1.05%)	1.639 (1.2%)	1.62	1.619 (-0.06%)
	2	0.549 (-15.9%	0.549 (-15.9%)	0.581 (-11.13%)	0.638 (-2.3%)	0.653	0.610 (-6.58%)

Table 7.3b: Mean queue lengths at FCFS centre and relative errors for model 2. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6. (Data table 7.1).

N_2	Class r	ROA	m-ROA	MVA	m-MVA	EXACT	UME2
1	1	1.166 (6.06%)	1.137 (3.39%)	1.184 (7.64%)	1.178 (7.09%)	1.10	1.098 (-0.18%)
	2	0.449 (12.37%	0.363 (-9.25%)	0.361 (-9.28%)	0.428 (7.15%)	0.40	0.381 (-4.75%)
2	1	1.261 (6.55%)	1.231 (4.01%)	1.314 (11.03%)	1.301 (9.94%)	1.184	1.180 (-0.3%)
	2	0.898 (10.08%	0.782 (-4.11%)	0.775 (-5.03%)	0.854 (4.75%)	0.816	0.782 (-4.16%)
3	1	1.334 (6.52%)	1.307 (4.35%)	1.410 (12.58%)	1.393 (11.2%)	1.253	1.251 (-0.15%)
	2	1.338 (7.36%)	1.208 (-3.10%)	1.217 (-2.37%)	1.283 (2.91%)	1.247	1.204 (-6.15%)
4	1	1.392 (6.24%)	1.369 (4.42%)	1.484 (13.19%)	1.464 (11.68%)	1.311	1.311 (0.0%)
	2	1.772 (4.93%)	1.632 (-3.35%)	1.677 (-0.68%)	1.715 (1.58%)	1.689	1.646 (-2.54%)
5	1	1.44 (5.89%)	1.419 (4.36%)	1.541 (13.33%)	1.52 (11.81%)	1.36	1.362 (0.14%)
	2	2.199 (2.77%)	2.053 (-4.07%)	2.148 (0.39%)	2.152 (0.57%)	2.14	2.102 (-1.77%)
6	1	1.479 (5.55%)	1.461 (4.23%)	1.587 (13.23%)	1.566 (11.74%)	1.402	1.404 (0.14%)
	2	2.620 (0.85%)	2.469 (-4.95%)	2.626 (1.10%)	2.593 (-0.17%)	2.598	2.569 (-1.11%)

Table 7.3c: Mean queue lengths for FCFS centre and relative errors for model 3. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6. (Data table 7.1).

N_2	Class r	ROA	m-ROA	MVA	m-MVA	EXACT	UME2
1	1	0.789 (23.55%	0.645 (1.03%)	0.762 (19.24%)	0.739 (15.77%)	0.639	0.615 (-3.75%)
	2	0.625 (112.1%	0.204 (-30.69%)	0.219 (-25.47%)	0.413 (40.23%)	0.294	0.257 (-12.58%)
2	1	1.01 (39.12%	0.792 (9.03%)	0.956 (31.65%)	0.90 (23.89%)	0.726	0.676 (-6.88%)
	2	1.518 (114.2%	0.712 (0.49%)	0.587 (-17.06%)	0.905 (27.72%)	0.708	0.61 (13.41%)
3	1	1.194 (44.43%	0.977 (18.26%)	1.131 (36.88%)	1.051 (27.2%)	0.826	0.75 (-10.46%)
	2	2.539 (101.7%	1.515 (20.38%)	1.133 (-9.96%)	1.515 (20.33%)	1.259	1.057 (-16.04%)
4	1	1.327 (42.1%)	1.163 (24.57%)	1.274 (36.39%)	1.19 (27.38%)	0.934	0.834 (-10.92%)
	2	3.582 (83.55%	2.568 (31.62%)	1.886 (-3.31%)	2.273 (16.51%)	1.951	1.619 (-17.01%)
5	1	1.424 (36.42%	1.317 (26.18%)	1.384 (32.57%)	1.311 (25.55%)	1.044	0.919 (-6.40%)
	2	4.614 (65.87%	3.737 (34.33%)	2.835 (1.92%)	3.186 (14.55%)	2.782	2.302 (-17.25%)
6	1	1.496 (30.14%	1.429 (24.34%)	1.467 (27.6%)	1.41 (22.67%)	1.15	1.009 (-5.06%)
	2	5.638 (50.95%	4.91 (31.47%)	3.924 (5.08%)	4.23 (13.23%)	3.735	3.126 (-16.3%)

Table 7.4:Raw data for two stage-cyclic Markovian network (HOL——> FCFS), with R = 2, N_1 = 2 and N_2 variable (Fig. 7.1) , (results - Table 7.5a-7.6c).

Model No.	class r	$\sum_1$ (HOL)		$\sum_2$ (FCFS)	
		μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}
1	1	50	1	20	1
	2	5	1	20	1
2	1	10	1	10	1
	2	10	1	10	1
3	1	5	1	10	1
	2	50	1	10	1

Table 7.5a: Utilisations at HOL centre with standard deviation from exact results for model 1.Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6, (Data table 7.4).

N_2	Class r	MVA	EXACT	UME2
1	1	0.1603 (0.012)	0.1477	0.1795 (0.0318)
	2	0.5934 (0.085)	0.6792	0.5974 (0.081)
2	1	0.1336 (0.003)	0.1304	0.1628 (0.032)
	2	0.862 (0.049)	0.8124	0.7642 (0.048)
3	1	0.1296 (0.003)	0.126	0.1615 (0.035)
	2	0.8988 (0.043)	0.8558	0.813 (0.042)
4	1	0.1282 (0.003)	0.1248	0.162 (0.037)
	2	0.9086 (0.039)	0.869	0.8296 (0.039)
5	1	0.1276 (0.003)	0.1246	0.1656 (0.041)
	2	0.9124 (0.038)	0.8736	0.8348 (0.038)
6	1	0.1272 (0.003)	0.1245	0.1631 (0.038)
	2	0.9168 (0.042)	0.875	0.836 (0.039)

Table 7.5b: Utilisations at HOL centre with standard deviation from exact results for model 2.Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6, (Data table 7.4).

N_2	Class r	MVA	EXACT	UME2
1	1	0.5217 (0.028)	0.545	0.5646 (0.019)
	2	0.222 (0.017)	0.2051	0.192 (0.013)
2	1	0.4575 (0.0235)	0.481	0.5081 (0.027)
	2	0.3592 (0.04)	0.319	0.305 (0.014)
3	1	0.4084 (0.03)	0.4384	0.4661 (0.0277)
	2	0.4539 (0.059)	0.3949	0.3862 (0.008)
4	1	0.3699 (0.036)	0.4066	0.4323 (0.025)
	2	0.5236 (0.073)	0.4506	0.4476 (0.003)
5	1	0.3384 (0.042)	0.381	0.4043 (0.023)
	2	0.5772 (0.083)	0.494	0.496 (0.002)
6	1	0.312 (0.021)	0.3595	0.3808 (0.021)
	2	0.62 (0.09)	0.5294	0.535 (0.005)

Table 7.5c: Utilisations at HOL centre with standard deviation from exact results for model 3. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6, (Data table 7.4).

N_2	Class r	MVA	EXACT	UME2
1	1	0.7992 (0.025)	0.825	0.8484 (0.023)
	2	0.026 (0.003)	0.0227	0.0159 (0.006)
2	1	0.7428 (0.037)	0.7806	0.8154 (0.034)
	2	0.0602 (0.015)	0.0445	0.0298 (0.014)
3	1	0.6534 (0.074)	0.7276	0.7772 (0.049)
	2	0.0935 (0.028)	0.0652	0.0469 (0.018)
4	1	0.5652 (0.104)	0.6698	0.7346 (0.064)
	2	0.1224 (0.038)	0.0843	0.0664 (0.018)
5	1	0.4918 (0.118)	0.6102	0.679 (0.068)
	2	0.1427 (0.041)	0.1017	0.0909 (0.011)
6	1	0.4338 (0.117)	0.5516	0.606 (0.0544)
	2	0.1548 (0.038)	0.1169	0.1157 (0.001)

Table 7.6a: Mean queue lengths at FCFS centre and percentage difference from exact results for model 1. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6, (Data table 7.4).

N_2	Class r	MVA	EXACT	UME2
1	1	0.619 (-19.3%)	0.812	0.906 (11.57%)
	2	0.377 (42.26%)	0.265	0.289 (9.05%)
2	1	0.531 (-28.24%)	0.7405	0.769 (3.91%)
	2	0.431 (11.65%)	0.386	0.398 (3.10%)
3	1	0.523 (-26.75%)	0.7146	0.74 (3.64%)
	2	0.441 (0.45%)	0.439	0.452 (2.96%)
4	1	0.521 (-26.24%)	0.705	0.729 (3.40%)
	2	0.447 (-2.82%)	0.46	0.471 (2.39%)
5	1	0.521 (-25.89%)	0.703	0.723 (2.84%)
	2	0.45 (-4.05%)	0.469	0.477 (1.70%)
6	1	0.521 (-25.78%)	0.702	0.72 (2.56%)
	2	0.452 (-4.23%)	0.472	0.478 (1.27%)

Table 7.6b: Mean queue lengths at FCFS centre and percentage difference from exact results for model 2. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6, (Data table 7.4).

N_2	Class r	MVA	EXACT	UME2
1	1	1.044 (0.57%)	1.038	1.026 (-1.15%)
	2	0.445 (-3.47%)	0.4617	0.441 (4.33%)
2	1	1.126 (3.88%)	1.086	1.07 (-1.47%)
	2	0.894 (-2.08%)	0.913	0.866 (-5.14%)
3	1	1.12 (-1.06%)	1.132	1.118 (-1.23%)
	2	1.371 (0.29%)	1.367	1.304 (-4.6%)
4	1	1.266 (7.74%)	1.1755	1.164 (-0.93%)
	2	1.87 (2.52%)	1.824	1.76 (-3.50%)
5	1	1.323 (8.88%)	1.2151	1.207 (-0.65%)
	2	2.387 (4.46%)	2.285	2.221 (-2.8%)
6	1	1.373 (9.75%)	1.2514	1.246 (-0.4%)
	2	2.92 (6.22%)	2.749	2.696 (-1.92%)

Table 7.6c: Mean queue lengths at FCFS centre and percentage difference from exact results for model 3. Population of class-1 jobs is fixed at 2 and the population of class-2 jobs is varied from 1 to 6, (Data table 7.4).

N_2	Class r	MVA	EXACT	UME2
1	1	0.775 (19.96%)	0.646	0.608 (-5.88%)
	2	0.205 (-35.53%)	0.3185	0.275 (-13.52%)
2	1	0.936 (26.31%)	0.741	0.671 (-9.44%)
	2	0.596 (-22.49%)	0.769	0.637 (-17.16%)
3	1	1.108 (30.50%)	0.849	0.749 (-11.77%)
	2	1.185 (-13.50%)	1.364	1.095 (-19.72%)
4	1	1.256 (30.69%)	0.961	0.843 (-12.27%)
	2	2.015 (-4.18%)	2.103	1.688 (-19.73%)
5	1	1.371 (27.89%)	1.072	0.968 (-9.7%)
	2	3.047 (2.48%)	2.973	2.497 (-16.01%)
6	1	1.458 (23.77%)	1.1789	1.12 (-4.9%)
	2	4.194 (6.15%)	3.951	3.55 (-10.15%)

Table 7.7: Raw data for the two stage cyclic Markovian queueing network (HOL —> PR) with 2 priority classes (Fig.7.1), (results - Table 7.8)

Exp. No.	Class r	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}	N_r
1	1	1/3	1	1/3	1	4
	2	1/3	1	1/3	1	1
2	1	1/3	1	1/3	1	4
	2	1/3	1	1/3	1	4
3	1	1	1	1	1	3
	2	1	1	1	1	5
4	1	1	1	1	1	1
	2	0.1	1	0.1	1	10
5	1	1	1	1	1	3
	2	100	1	100	1	5
6	1	0.2	1	1	1	1
	2	0.02	1	0.02	1	10

Table 7.8: Mean queue lengths at PR centre and percentage difference from exact results for the two stage cyclic Markovian network (HOL —> PR), (Data - table 7.7).

Exp. No.	Class r	MVA	EXACT	UME2
1	1	1.8355 (-5.92%)	1.951	1.9136 (-1.91%)
	2	0.557 (1.51%)	0.5487	0.5734 (4.5%)
2	1	1.7114 (-8.48%)	1.87	1.8235 (-2.48%)
	2	1.5426 (-27.57%)	2.13	2.2328 (4.82%)
3	1	1.1958 (-11.81%)	1.356	1.324 (-2.35%)
	2	3.2271 (21.78%)	2.644	2.744 (-3.78%)
4	1	0.0901 (-39.28%)	0.1484	0.1717 (15.70%)
	2	5.7554 (6.54%)	5.402	5.7554 (-3.88%)
5	1	1.499 (-0.8%)	1.5	1.499 (-0.8%)
	2	3.0133 (20.34%)	2.504	2.504 (0.0%)
6	1	0.0194 (-38.31%)	0.0315	0.0361 (14.6%)
	2	4.9668 (15.88%)	4.286	3.893 (-9.16%)

Table 7.9: Raw data for the two stage cyclic General queueing network (PR ——> FCFS) with 2 or 3 priority classes (Fig. 7.1), (results - Table 7.10)

Exp. No.	Class r	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}	N_r
1	1	1	2	2	5	2
	2	2	3	2	4	5
2	1	0.5	2	2	5	2
	2	2	3	2	4	5
3	1	0.5	25	2	30	2
	2	2	15	2	14	5
4	1	1	5	3	7	5
	2	3	6	1	3	2
5	1	2	3	1	7	3
	2	3	7	5	2	3
6	1	4	3	5	7	3
	2	1	7	2	2	3

Table 7.9 continued

Exp. No.	Class r	μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}	N_r
7	1	1	1	1	1	2
	2	2	1	1	1	2
	3	2	1	1	1	3
8	1	2	4	3	9	1
	2	1	6	2	1	1
	3	1	7	1	5	2
9	1	2	5	5	7	1
	2	3	3	1	9	1
	3	2	2	3	3	1
10	1	2	15	3	27	1
	2	1	35	2	51	1
	3	1	24	1	46	2

Table 7.10: Mean queue lengths at FCFS centre with percentage difference and utilisations at PR centre with the corresponding standard deviations from exact results for a two stage cyclic network , (PR —> FCFS), (data - Table 7.9).

EXP. No.	Class r	EXACT, ρ_{1r}	UME 2, ρ_{1r}	EXACT, $\langle n_{2r} \rangle$	UME 2, $\langle n_{2r} \rangle$
1	1	0.66	0.68 (0.02)	0.82	0.81 (-1.22%)
	2	0.19	0.205 (0.015)	1.33	1.32 (-0.75%)
2	1	0.84	0.86 (2.38%)	0.43	0.388 (-9.76%)
	2	0.096	0.062 (0.034)	0.653	0.672 (2.91%)
3	1	0.76	0.796 (0.036)	0.49	0.434 (-11.42%)
	2	0.0688	0.0364 (0.0324)	0.96	0.974 (1.45%)
4	1	0.847	0.873 (0.025)	1.25	1.09 (-12.8%)
	2	0.031	0.013 (0.018)	0.39	0.389 (-0.8%)
5	1	0.364	0.358 (0.006)	2.13	2.189 (2.77%)
	2	0.11	0.16 (0.05)	1.835	1.90 (3.54%)
6	1	0.584	0.584 (0.0)	1.51	1.54 (1.98%)
	2	0.284 (0.03)	0.314	0.61	0.67 (9.83%)

Table 7.10 continued

EXP. No.	Class r	EXACT, ρ_{2r}	UME 2, ρ_{2r}	EXACT, $\langle n_{2r} \rangle$	UME 2, $\langle n_{2r} \rangle$
	1	0.361	0.333 (0.028)	1.531	1.59 (3.85%)
7	2	0.137	0.15 (0.013)	1.487	1.56 (4.9%)
	3	0.157	0.177 (0.02)	2.033	2.12 (4.28%)
	1	0.48	0.464 (0.016)	0.52	0.535 (2.88%)
8	2	0.175	0.228 (0.053)	0.3705	0.337 (-9.04%)
	3	0.148	0.158 (0.01)	0.4902	0..518 (5.67%)
	1	0.526	0.534 (0.008)	0.473	0.465 (-1.69%)
9	2	0.082	0.077 (0.005)	0.41	0.422 (2.92%)
	3	0.065	0.088 (0.023)	0.352	0.336 (-4.54%)
	1	0.499	0.4958 (0.0032)	0.50	0.504 (0.8%)
10	2	0.156	0.182 (0.026)	0.346	0.334 (-3.46%)
	3	0.092	0.099 (0.007)	0.5206	0.59 (13.33%)

Table 7.11: Raw data for general (GE) central server network (c.f Fig. 7.2), with mixed service disciplines. (results - Table 7.12).

Exp. No.	class r	$\sum_1$ (PR)		$\sum_2$ (HOL)		$\sum_3$ (FCFS)		$P_{1r,2}$	$P_{1r,3}$
		μ_{1r}	C^2_{s1r}	μ_{2r}	C^2_{s2r}	μ_{3r}	C^2_{s3r}		
1	1	1	3	2	2	3	4	0.3	0.3
	2	3	2	4	1	3	4	0.3	0.3
2	1	1	15	2	25	3	46	0.3	0.3
	2	3	16	4	32	3	37	0.3	0.3
3	1	1	3	1	3	1	3	0.5	0.4
	2	10	3	10	3	10	3	0.5	0.4
4	1	1	3	1	3	1	3	0.5	0.4
	2	1	3	100	3	50	3	0.5	0.4
5	1	1	3	1	3	1	3	0.5	0.4
	2	100	3	100	3	100	3	0.5	0.4
6	1	5	15	3	11	4	2	0.4	0.3
	2	2	3	2	9	4	4	0.4	0.3
7	1	5	0.5	3	0.5	4	0.5	0.4	0.3
	2	2	0.5	2	0.5	4	0.5	0.4	0.3

Table 7.12: Utilisations of PR centre with the corresponding standard deviations and mean queue lengths at HOL centre with the corresponding percentage differences from simulations. (Data - Table 7.11).

EXP. No.	Class r	N_r	SIM ρ_{1r}	UME 2 ρ_{1r}	SIM $\langle n_{2r} \rangle$	UME 2 $\langle n_{2r} \rangle$
1	1	3	0.9613	0.985 (0.023)	0.2114	0.191 (-9.17%)
	2	3	0.0319	0.0128	0.032	0.027 (-15.62%)
2	1	3	0.8569	0.862 (0.005)	0.3141	0.299 (-4.81%)
	2	3	0.0452	0.0369 (0.008)	0.2176	0.20 (-8.08%)
3	1	1	0.5182	0.508 (0.0095)	0.2655	0.259 (-2.22%)
	2	4	0.148	0.166 (0.02)	0.97	0.967 (-1.88%)
4	1	1	0.5251	0.525 (0.0004)	0.2658	0.262 (-1.42%)
	2	4	0.4352	0.443 (0.0078)	0.2617	0.247 (-5.61%)
5	1	1	0.5234	0.5239 (0.0005)	0.2636	0.262 (-0.006%)
	2	4	0.0195	0.0205 (0.001)	1.045	1.04 (-0.47%)
6	1	4	0.8076	0.797 (0.01)	0.7631	0.799 (4.7%)
	2	1	0.0694	0.072 (0.0026)	0.1451	0.145 (-0.06%)
7	1	4	0.993	0.994 (0.001)	0.4266	0.48 (12.51%)
	2	1	0.0055	0.0045 (0.001)	0.0039	0.003 (-23.07%)