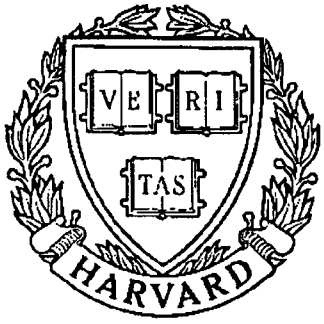
THESIS REPORT
Ph.D.

SYSTEMS
RESEARCH
CENTER



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Performance Evaluation and Optimization of Parallel Systems with Synchronization

by L. Gun
Advisor: A. Makowski

Ph.D. 89-3
Formerly TR 89-64

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1989		**00-00-1989 to 00-00-1989**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Performance Evaluation and Optimization of Parallel Systems with Synchronization	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,The Graduate School,2123 Lee Building,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**117**	

PERFORMANCE EVALUATION AND OPTIMIZATION OF PARALLEL SYSTEMS WITH SYNCHRONIZATION

by

Levent Gün

Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1989

Advisory Committee:

Associate Professor Armand Makowski, Chairman/Advisor
Associate Professor Prakash Narayan
Associate Professor Lawrance Bodin
Assistant Professor Adrianos Papamarcou
Doctor Alain Jean-Marie

ABSTRACT

Title of Dissertation: Performance evaluation and optimization of parallel systems with synchronization

Levent Gün, Doctor of Philosophy, 1989

Dissertation directed by: Armand M. Makowski
Associate Professor
Electrical Engineering Department

This thesis considers synchronization issues such as resequencing and fork/join in parallel architectures. The discussion is carried out in the context of K parallel single server queues with general servers where jobs are subject to resequencing. Both performance evaluation and optimal routing problems are addressed for such systems.

In the first part, Poisson arrivals are assumed to be randomly allocated to different queues according to a Bernoulli switch. The distributions of the various delays in the system are obtained by sample path arguments. The problem of choosing the switching probabilities that minimize the average end-to-end delay is considered. In addition to obtaining exact results in some cases, simple but accurate approximations are provided when the service time distributions are exponential. The simple form of these approximations is then utilized to solve the optimization problem in the case when the service parameters are unknown, and a simple stochastic approximation algorithm is proposed. When the servers are all identical, several useful asymptotic results are obtained as K increases to infinity. Various stochastic monotonicity and convexity results are also provided for this parallel system.

In the second part, the dynamic optimization of the same model is investigated under more general assumptions for the arrival process. The resequencing problem is combined with a fork/join problem, where the incoming packets are broken into smaller subpackets for processing at different queues. The problem of finding the optimal allocation policy that minimizes the average discounted and the long-run average costs is formulated as a Markov Decision problem, where the cost-per-stage is taken as the end-to-end delay of each packet. In both cases, the optimal policy is identified as the one that drives the workload in each queue to a balanced configuration as quickly as feasible.

TO MY BELOVED WIFE HEDİYE

for her incessant love and support

ACKNOWLEDGMENTS

First, I wish to express my deep-felt gratitude to my advisor Prof. Armand M. Makowski for his ceaseless guidance and inspiration during my graduate years, and for his perseverance during the many iterations of this dissertation. I owe him very much for his generous financial assistance and providing me the freedom in selecting research topics.

My special thanks are due to Dr. Alain Jean-Marie for his encouragement and support during the past two years. Chapters II and IV of this thesis are the outcome of many elated hours of joint work with him. He also carefully read the other chapters and improved them by many valuable comments. I also thank to Professors Prakash Narayan, Lawrance Bodin and Adrianos Papamarcou for serving in my advisory committee and carefully reading this dissertation.

My thanks are also due to the financial supporters of this research: National Science Foundation, through NSFD Grant ECS-83-51836, Office of Naval Research, through ONR Grant N00014-84-K-0614, and AT&T Bell Laboratories. This work is conducted in the stimulating atmosphere of the Electrical Engineering Department and the Systems Research Center of the University of Maryland.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

a.s.	Almost surely.
DS	Disordering System.
FCFS	First Come First Served.
Geo	Geometric distribution.
i.i.d.	Independent and identically distributed.
$\mathbb{N}$	The set of non-negative integers.
$P(*)$	Steady state probability that a job has zero resequencing time.
$\mathcal{P}_E$	Projection operator onto a set E.
$\mathbb{R}$	The set of real numbers.
$\mathbb{R}_+$	The set of non-negative real numbers.
RV	Random variable.
$[x]^+$	$\max\{0, x\}, \quad x \in \mathbb{R}$.
$\overline{x}$	$1 - x, \quad 0 \leq x \leq 1$.
$\mathcal{D}$	$\{p \in [0,1]^K : \sum_{k=1}^K p_k = 1, \ \lambda p_k < \mu_k, \ 1 \leq k \leq K\}$
$\mathcal{U}$	$\{u \in [0,1]^K : \sum_{k=1}^K u_k = 1\}$
$\mathcal{U}'$	$\{u \in \mathbb{R}^K : \sum_{k=1}^K u_k = 1\}$
δ_{ij}	Kronecker symbol defined by $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

CHAPTER I

INTRODUCTION AND SUMMARY

Recent advances in computer and communications technology have led to the proliferation of complex parallel and distributed system architectures. On one hand, these new systems offer many advantages over the conventional systems such as resource sharing, reliability and fault tolerance. On the other hand, the parallel and distributed nature of these systems pose fundamental problems related to the interactions between different parts of the system. For instance, the ordering of the packets is of great importance in almost every application where data is transmitted over several communication links. This is also the case in distributed databases or in computing systems where computations have to be executed in a prespecified sequence. In a parallel system different portions (tasks) of a job can follow different paths and overtake one another due to the random nature of delays in various parts of the system. Therefore, some of these tasks have to wait at the destination for those tasks that have entered the system earlier. This type of synchronization delay is called the *resequencing delay* and may crucially affect the performance of the system. It is therefore essential to understand the effects of various system parameters on this synchronization primitive.

The main concern of this thesis is to provide an analytical basis for a better understanding of the *resequencing constraint*, i.e., the requirement that *jobs* have to leave the system in their order of arrival (see (1.1.1)). This synchronization constraint is a basic problem of interest in many parallel and distributed computer, communication and manufacturing systems. Its analysis cannot be handled by the theory of product-form queueing networks, and its study therefore provides new theoretical and experimental challenges.

I.1. MOTIVATION

In order to provide concrete examples, applications from areas as diverse as packet switching, distributed databases and parallel processing are briefly considered next. The reader is referred to the recent survey paper of Baccelli and

Makowski [BaM] and to the M.S. Thesis of Varma [Var.a] for additional examples where resequencing and other forms of synchronization constraints naturally occur.

(i) In packet switching networks such as IBM's System Network Architecture (SNA) [GMN] and the French PTT's public network TRANSPAC [Dan], messages are formatted into packets and are transmitted over an interconnected network. In order to increase the network utilization, packets belonging to the same message are routed over different paths to reach their destination, and these packets may arrive out of sequence at the destination node. Reasons for this include (i) the random nature of the delays in different parts of the network due to different link speeds, random path lengths and varying message sizes, and (ii) the retransmission of erroneous packets such as in Selective-Repeat Automatic Repeat Request protocols [Sch]. In order to achieve message integrity, many communication protocols such as SNA require First-In First-Out delivery. Consequently, the out of sequence packets are stored at the receiver and await the arrival of the packets that have been transmitted before them.

(ii) In distributed databases, different storage sites may contain portions or complete replications of a piece of data. For reasons of reliability, centralized control is not allowed in these systems, and concurrency control mechanisms have been developed to preserve consistency [Ell, LeL]. Due to random communication delays, update requests such as read, write or delete originated from different access sites may arrive out-of-order to each storage site. If the original order of the write and delete requests are not preserved in processing these updates, data in distinct storage sites will no longer be replicas of each other. Therefore, to ensure consistency, the update requests must be resequenced at each storage site before processing.

(iii) In parallel processing machines, different parts of a program are executed on different processors. However, data often needs to be exchanged between the processors, i.e., after completing the execution of a part of a program a

processor may have to wait for the outcome of some other process in order to continue executing the rest of the program. Consequently, the processor may have to wait idle until the necessary synchronization is achieved.

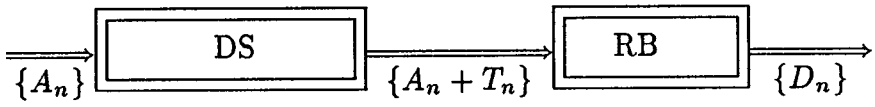
In all three examples, the underlying generic structure is the one given in Figure I.1, i.e., a distributed (disordering) system (DS) followed by a resequencing buffer (RB). In this figure, the $\mathbb{R}_+$-valued sequences $\{A_n,\ n = 0,1,\ldots\}$, $\{T_n,\ n = 0,1,\ldots\}$, $\{D_n,\ n = 0,1,\ldots\}$ and $\{S_n,\ n = 0,1,\ldots\}$ have the interpretation that for all $n = 0,1,\ldots$,

A_n: Arrival epoch of the n^{th} job into the DS, with $A_0 = 0$;

T_n: Delay of the n^{th} job in the DS;

D_n: Departure epoch of the n^{th} job from the RB;

S_n: The system time (or end-to-end delay) of the n^{th} job, i.e., $S_n = D_n - A_n$.



Arriving Jobs Out of Sequence Jobs Resequenced Jobs

Figure I.1.

The Generic Structure

The n^{th} job arrives into the DS at time A_n and experiences a delay of T_n in the system. However, due to the distributed nature of the system, the departure times $\{A_n + T_n,\ n = 0,1,\ldots\}$ of jobs are not in the same order as their arrival times $\{A_n,\ n = 0,1,\ldots\}$. The n^{th} job, upon leaving the DS, enters the RB to await all the jobs that have entered the system earlier. Only after all the jobs which have arrived to the system before it leave the DS, does the n^{th} job leave the RB, i.e., D_n is defined by the condition

$$D_n = \max_{j:\ A_j \le A_n} \{A_j + T_j\}\ , \qquad n = 0,1,\ldots\ . \tag{1.1.1}$$

Throughout this thesis, condition (1.1.1) will be referred to as the *resequencing constraint.* Under this constraint, the system time sequence $\{S_n,\ n = 0, 1, \ldots\}$ satisfies the recursion [BGP]

$$S_n = \max\{T_n\ ,\ S_{n-1} - A_n + A_{n-1}\}\ , \qquad n = 1, 2, \ldots \tag{1.1.2}$$

with $S_0 = T_0$, and the resequencing delay R_n of the n^{th} job is thus

$$R_n = S_n - T_n, \qquad n = 0, 1, \ldots\ . \tag{1.1.3}$$

I.2. LITERATURE SURVEY

This section provides a brief account of the literature on resequencing systems. The recent tutorial paper of Baccelli and Makowski [BaM] contains a more detailed discussion of some of the references given here. Other forms of synchronization constraints are also discussed in [BaM], where some recent developments in the literature are summarized. In all the models, unless otherwise specified, the arrival process is Poisson.

I.2.1. Infinite Server Models

The earliest paper on resequencing systems is due to Kamoun, Kleinrock and Muntz [KKM] who studied the effects of resequencing in an $M/M/\infty$ queue, i.e., when the DS is composed of an infinite number of identical exponential servers. They obtained the steady state statistics of the system time S_n and of the number of jobs in the RB, as well as the bulk size distribution of the output process from the RB. The results of [KKM] were extended to the $M/GI/\infty$ case by Harrus and Plateau [HaP] by means of a similar analysis. Subsequently, Baccelli, Gelenbe and Plateau [BGP] considered the situation where the RB is followed by a single server queue with a general service time distribution. The DS is again implemented by the $M/GI/\infty$ queue. These authors gave recursive formulas of the type (1.1.2) for the end-to-end delay (including the delay in the single server queue), which are then used to derive integral equations for its distribution. The analysis is carried out via factorization methods when the delay distributions in the DS are exponential,

i.e., when DS is the $M/M/\infty$ system. Recently, Gelenbe and Stafylopatis [GeS] considered the model of [HaP] under three *partial* ordering/resequencing disciplines which reflect the random association or locality in the precedence constraints.

I.2.2. Finite Server Models

In the infinite server models of Section I.2.1, the disordering delays $\{T_n,\ n = 0,1,\ldots\}$ form an *independent and identically distributed (i.i.d.)* sequence of random variables (RVs) which is also *independent* from the arrival time sequence $\{A_n,\ n = 0,1,\ldots\}$. When the number of servers is finite, these independence properties are no longer available, and the corresponding results are few in the literature and limited to very special cases. All models assume Poisson arrivals and exponential service time distributions. Two classes of models are considered for the DS, namely (i) models where there is a common buffer attended by parallel servers, (ii) models with parallel single server queues.

(i) Common Buffer Models: The models in this class all lead to a Markovian analysis. However, the lack of independence causes a rapid explosion in the size of the state space with the number of servers. The $M/M/K$ model with identical servers is studied by Bharat-Kumar and Kermani [BKK] who derived an expression for the mean resequencing delay. Agrawal and Ramaswamy [AgR] also considered the model in [BKK] but focused on the distributional aspects of the resequencing delay. Yum and Ngai [YuN] considered the more general $M/M/K/B$ model with heterogeneous servers when the common buffer size B is finite. In the case of heterogeneous servers, it is commonly assumed in the literature that when more than one server is available, jobs are scheduled to the fastest available server. Yum and Ngai obtained the resequencing delay distribution through a numerical algorithm. However, due to the high dimensionality of the state-space, they reported their method to be limited to $K \leq 5$. Lien [Lie] considered the resequencing delay due to the $M/M/2$ queue and obtained the average resequencing delay by a simple yet clever argument. Later on, Varma [Var.a] showed that the Markovian state representation of Lien naturally leads to a matrix-geometric representation for

the steady state buffer occupation probabilities when the $M/M/2$ queue has a finite buffer. However, this approach does not extend to the case $K > 2$. The $M/M/2/B$ model is also considered in the Ph.D. Thesis of Iliadis [Ili] where the distribution of the resequencing delay is obtained under various threshold type scheduling policies (see also [IL.a] and [IL.b]).

(ii) Parallel Buffer Models: The work on this class of models is very recent. Jean-Marie [JeM.b] considered the case of two $M/M/1$ queues in parallel with identical servers when two different Poisson streams of incoming jobs are routed randomly to the queues by two Bernoulli switches. He obtained the distribution of the resequencing delay by a sample path argument. Iliadis and Lien [IL.c] considered two parallel heterogeneous $M/M/1$ queues with two types of traffic, namely (i) jobs allocated to the queues by a Bernoulli switch which are subject to resequencing, and (ii) interfering local traffic to each queue which are not subject to resequencing and leave the system as soon as they are serviced. All arrival processes are assumed to be Poisson and mutually independent. A recursive method is proposed for obtaining the average resequencing time for the jobs allocated by the Bernoulli switch. Their solution method extends to the situation where there are several arrival processes to the Bernoulli switch from various sources when resequencing operates class by class, i.e., when jobs from a given source need to wait only for the jobs that arrive to the DS from that source.

I.2.3. Structural Results

Bounding methodologies based on convex and strong stochastic ordering arguments are used in the M.S. Thesis of Varma [Var.a]. He provides monotonicity results for DSs in (i) and (ii) above under more general assumptions on the service and arrival distributions. Multistage DSs in tandem are also considered in [Var.a] where it is shown that for infinite server DSs resequencing the jobs only after the last stage yields a lower end-to-end delay than when jobs are resequenced after every DS. When the DSs are $GI/GI/K$ queues and the jobs are resequenced after every stage, Varma also showed that a decrease in service times in one of the stages

implies a decrease in the end-to-end delay.

I.2.4. Optimization of Resequencing Systems

Results on the optimal design of resequencing systems are again very few and recent. The static optimization problem of choosing the optimum switching probabilities in the multiple queue model of Section I.2.2(ii) is considered by Jean-Marie [JeM.b] when $K = 2$ with identical exponential servers. Two Poisson input processes are randomly allocated to the queues by two Bernoulli switches when the resequencing operates class by class. Mean queueing and resequencing times are obtained, and the optimal switching probabilities are characterized through the unique solution of a fifth order polynomial.

Varma [Var.b] considered the dynamic allocation of jobs to servers in the $M/M/2$ queue with heterogeneous servers so as to minimize the system time. Using dynamic programming arguments he showed that the faster server should always be kept busy, while the optimal policy that assigns jobs to the slower server is of threshold type in the number of jobs in the common buffer, and is independent of the number of jobs in the RB.

In addition to the papers annotated above, the effect of resequencing on more general systems are considered in [Bac] and [JeM.a]. While [JeM.a] studies the effect of the resequencing delay in Omega networks, [Bac] considers synchronization issues in distributed databases and provides computable bounds for various delays.

I.3. SUMMARY OF THE THESIS

In this thesis we consider the performance evaluation and the optimal routing of jobs in the resequencing system of the type given in Figure I.1. Attention is given here to DSs composed of K parallel single server queues with infinite capacity buffers. The material in Chapters II and IV is joint work with Dr. A. Jean-Marie and is also included in the manuscripts [GJM] and [JMG], respectively.

In Chapter II, jobs arrive according to a Poisson process and are randomly allocated to the queues by a Bernoulli switch. Each queue is attended by a single server with a *general* service time distribution. The service times at different

queues are assumed independent with possibly *different* distributions, i.e., the DS is composed of K parallel $M/GI/1$ queues. We first express the resequencing and system times in terms of the waiting and response times in each one of the parallel $M/GI/1$ queues. This is done by means of a sample path argument which focuses directly on the system time rather than on the resequencing delay. The distributions and first moments of the resequencing and system times are then obtained using these expressions. The problem of choosing the switching probabilities in order to minimize the average end-to-end delay is then addressed for the following two special cases: (i) when the service time distributions are identical, and (ii) when $K = 2$ and the service times are exponential. In the first case, it is shown that the equal load allocation is optimal, and that at this optimum configuration the average system time decreases with K. The second case differs from the model in [JeM] mentioned in Section I.2.2(ii) in that now there is only one arrival process and the servers have different rates. The optimal routing probability is again characterized by the unique root of a fifth order polynomial for certain values of the system parameters, whereas it is best to route all the traffic to the faster server for other values. The equations that define these regions in the parameter space are identified.

The results obtained in the special case (ii) lead to an asymptotically exact *approximation* for K (≥ 2) heterogeneous exponential servers. This approximation has a very simple form and provides insights into the variation of the optimal switching probabilities with the system parameters. Furthermore, for wide range of system parameters it yields a relative error of at most 1% for the mean system time.

The computation of the optimal routing probabilities for the same system without resequencing is also recalled and improved. The solutions to both problems are compared throughout Chapter II. Strong stochastic convexity (see Definition A.II.1 of Appendix II) of the stationary disordering delay in the switching probability vector is also established. In the homogeneous case, equal load allocation is shown to *stochastically* minimize the disordering delay. Furthermore, at this

optimum configuration, stochastic monotonicity and integer convexity properties in K are established for the disordering delay.

Chapter III considers the model of Chapter II when the service times are exponentially distributed, but when both the service and the arrival rates are *unknown*. The simple form of the approximations given in Chapter II is used in a *stochastic approximation* algorithm for computing the *approximate* optimal switching probability vector. The algorithm starts with an initial switching vector and estimates the system parameters by making idle time measurements. The switching vector is then updated using these measurements. The algorithm proceeds by taking new measurements when the system is driven with this new switching vector, and then updates the switching vector again.

Chapter IV studies the *asymptotic* behavior of various delays in the model of Chapter II when the service time distributions are identical and the load is equally (optimally) allocated to the queues. Asymptotic expressions for the distributions of the resequencing and system times are provided as K increases to infinity. Two cases are considered depending on whether the arrival rate into the system is held constant or grows linearly with K.

In the first case, it is proved that the statistics of the system tend to those of the $M/GI/\infty$ system with resequencing. While asymptotic stochastic monotonicity and convexity results (in the sense defined in Chapter IV) are stated for the system time RV, the resequencing delay exhibits different structural characteristics depending on the arrival rate into the system. For example, when the service times are exponential it is (asymptotically) stochastically increasing and concave in K for smaller loads on the system, while for larger loads it is (asymptotically) stochastically decreasing and convex.

The situation is different in the second case. Indeed the expected system and resequencing times both grow as $\log K$ while the average response time is constant for all K. Therefore, although the response time of a job in the queue dominates the resequencing time in the first case, the resequencing delay dominates the response time and has a major impact on the system time in the second case.

In Chapter V, it is assumed that the jobs arrive according to a renewal sequence. The workloads of the jobs are *i.i.d.* and independent from the interarrival times. Upon arrival, instead of the Bernoulli allocation as in the previous chapters, each job is broken into smaller tasks for processing at different queues. These tasks are later assembled in the RB, and a job leaves the system only after all of its tasks are serviced *and* all the other jobs that arrived into the system before it are also assembled and have left the system. The dynamic optimization problem of allocating each job into the parallel queues is considered when the processing *rates* in different queues are identical and constant, i.e., the service time of a task is proportional to its workload and tasks with equal workloads require equal amounts of service time at different queues. The average finite horizon, long-run average and average discounted costs are *all* shown to be minimized by the *same* allocation policy when the cost-per-stage is taken to be the system time of a job. At each arrival epoch, given the workload in each queue and the workload of the arriving job, this optimal policy allocates the workload of the job to the queues so as to derive the workload in each queue into a balanced position as fast as feasible. A simple algorithm that computes this optimal allocation vector is also provided. Under this optimal policy, the steady state system time of a job is obtained as the response time in a $GI/GI/1$ queue.

The case where the jobs are not allowed to be broken into smaller tasks is also briefly considered when $K = 2$. Optimality of joining the queue with the smallest workload is established in this case.

Notation

The following notation and definitions are used throughout this thesis: The (resp. positive) real line is denoted by (resp. $\mathbb{R}_+$) $\mathbb{R}$. The k^{th} component of a column vector x in $\mathbb{R}^K$ is denoted by x_k, while the k^{th} component of the vector $[x]^+$ is defined as $[x_k]^+ := \max\{0, x_k\}$, $1 \leq k \leq K$. The transpose of x is denoted by x^T. The column vector of ones with appropriate dimensions is denoted by e. The notation $\mathcal{P}_E(x)$ denotes the projection of x onto a set $E \subset \mathbb{R}^K$. The distribution function of a RV X will be denoted by the same letter, e.g., $X(x) =$

$P(X \leq x)$. Whenever it is necessary to explicitly indicate the dependence of $X(x)$ on a parameter θ, the notation $X(\theta, x)$ and $X(\theta)$ will be substituted for $X(x)$ and X, respectively. The distinction between the distribution function $X(x)$ and the RV $X(\theta)$ will always be clear from the context. Finally, $\overline{x} = 1 - x$ for every x in $[0, 1]$.

CHAPTER II

RESEQUENCING IN PARALLEL QUEUES WITH BERNOULLI LOADING

II.1. INTRODUCTION

A system with K parallel queues each with an infinite capacity buffer is considered when the jobs are subject to *resequencing.* Jobs arrive according to a Poisson process and are allocated to the queues by a Bernoulli switch. The service times at different queues are assumed independent with possibly *different* distributions. After completing service, a job moves to the RB and awaits the service completion of the jobs which have arrived into the system before it. The model is made precise in Section 2.

In Section 3, the distributions of the resequencing and system times are computed in terms of the distribution functions of the waiting and response times in the parallel $M/GI/1$ queues. Expressions for their first moments are then easily derived.

The problem of choosing the allocation probabilities to minimize the average system time is addressed in Section 4 in two special cases, namely an arbitrary number of identical servers and two queues with exponential servers. Section 5 contains simple approximations to the optimal switching vector when the K parallel servers are exponential.

The solution to the problem *without* the resequencing constraint is briefly recalled [BuC] in order to provide a comparison to the resequencing problem. An algorithm to compute the optimal probabilities is available for the general case in [BuC]. Simplifications to this algorithm are provided in Appendix I.

II.2. THE MODEL

The model consists of K queues in parallel where each queue k, $1 \leq k \leq K$, has an *infinite* capacity buffer and is attended by a *single* server whose service time distribution $B_k(\cdot)$ has finite mean $1/\mu_k$ and variance σ_k^2. The service times are assumed mutually independent. Jobs arrive into the system according to a *Poisson* process with parameter λ, and join the k^{th} queue with probability p_k, $1 \leq k \leq K$,

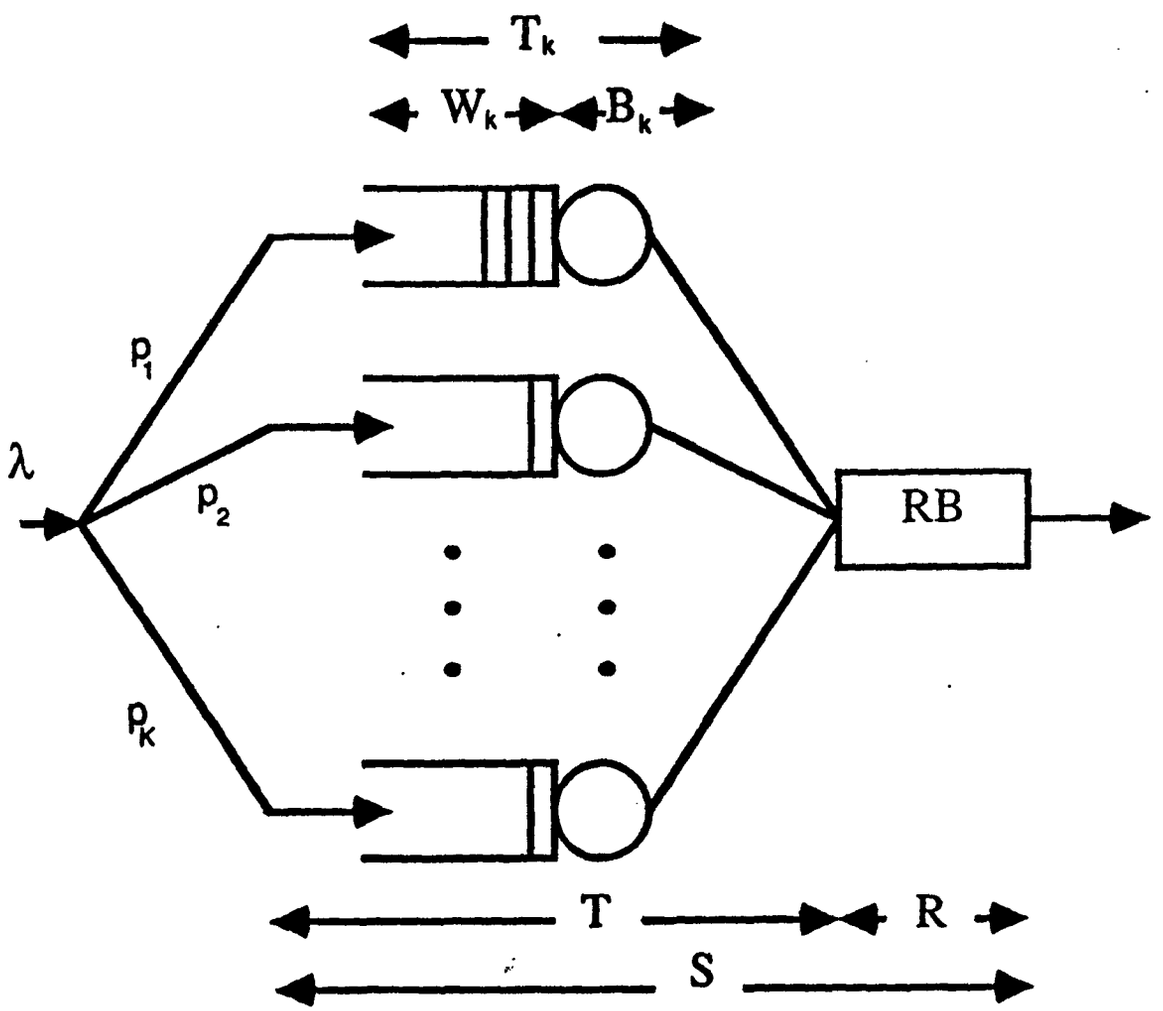


Figure II.1.

The Model

where $\sum_{k=1}^{K} p_k = 1$. The service discipline at each queue is *first come first served* (FCFS), and upon service completion, a job joins the RB to await for the service completion of *all* the jobs that have entered the system before it.

Throughout this chapter, the system is considered in statistical equilibrium. For each queue k, $1 \leq k \leq K$, set

B_k: the *service time* of a job in server k;

W_k: the *waiting time* of a job in queue k;

T_k: the *sojourn time* of a job in queue k (i.e., $T_k = W_k + B_k$),

and for the system, define

T: the *response time* of a job (i.e., the time needed to complete its service);

R: the *resequencing time* of a job (i.e., the time it spends in the RB);

S: the *system time* of a job (i.e., $S = R + T$).

Since the thinning of a Poisson process by a Bernoulli process results in a set of *independent* Poisson processes, the system behaves like a collection of K independent $M/GI/1$ systems with arrival rates $\lambda p_1, \ldots, \lambda p_K$, and with a global ordering/resequencing mechanism. In particular, the RVs W_k and T_k are independent of the RVs W_l and T_l for $1 \leq k \neq l \leq K$. Since the arrival process to each queue is Poisson, the RV W_k is the *workload* of queue k at a random instant [Kle.a], i.e, W_k is the *virtual waiting time* in queue k.

The presence of resequencing does not affect the stability condition of the system [BGP]. Therefore, the system is stable if and only if each queue is stable, which reads

$$0 \leq p_k < \frac{\mu_k}{\lambda}\,, \qquad 1 \leq k \leq K\,. \tag{2.2.1}$$

The utilization of queue k is denoted by $\rho_k = \lambda p_k / \mu_k$, while the system *capacity* is $\mu = \sum_{k=1}^{K} \mu_k$. Note that the stability condition (2.2.1) implies $\lambda < \mu$. Conversely, if this condition is satisfied, then the convex subset $\mathcal{D}$ of $\mathbb{R}^K$ defined by

$$\mathcal{D} \;=\; \{p \in [0,1]^K : \; \sum_{k=1}^{K} p_k = 1 \;\text{ and }\; \lambda p_k < \mu_k,\; 1 \leq k \leq K\}$$

is nonempty.

II.3. RESPONSE AND SYSTEM TIME ANALYSIS

In this section, expressions for the RVs T, R and S are provided in terms of the RVs T_k and W_k, $1 \leq k \leq K$. These expressions then lead to the computation of their distribution functions and first moments.

On an underlying common probability space, define a RV $U : \Omega \to \{1, ..., K\}$ such that $P(U = k) = p_k$, $1 \leq k \leq K$. The RV U is assumed independent of all the RVs introduced so far.

Theorem II.3.1. *The RVs T, R and S are given respectively by*

$$T = T_U \ , \tag{2.3.1a}$$

$$R = [W^* - T_U]^+ \tag{2.3.1b}$$

and

$$S = \max\{W^*, T_U\} \ , \tag{2.3.1c}$$

where

$$W^* := \max\{W_k;\ 1 \leq k \leq K,\ k \neq U\} \ . \tag{2.3.1d}$$

Although a transient version of (2.3.1) also holds for each job n, $n = 0, 1, \ldots$, only the steady state equalities are given here in order to keep the notation to a minimum.

Proof. The expression for T is plain and expresses the Bernoulli allocation of the jobs.

In order to compute R, assume that a "tagged job" C arrives to the Bernouilli switch at time $t = 0$. Let C_k be the last job in the FCFS queue k at time $t = 0^+$ so that C is C_U. Then, with the notation introduced in Section 2, the job $C_k, k \neq U$, completes service at time W_k, so that W^* is the time when the last of the C_k's, $k \neq U$, completes service. The tagged job C completes service at time T_U. If

$T_U < W^*$, then at least one of the C_k, $1 \leq k \neq U \leq K$, has not yet completed service and C has to wait until W^*, i.e., $W^* - T_U$ is his resequencing time. On the other hand, if $T_U \geq W^*$, then C experiences no resequencing delay. Therefore, $R = [W^* - T_U]^+$.

The expression for S now follows immediately.

□

The following lemma presents an equivalent form of Takács' integro-differential equation for the $M/GI/1$ system [Kle.a, p. 230] and will prove useful in computing the distribution of S. The term λp_k is the arrival rate into the k^{th} queue.

Lemma II.3.2. *The equality*

$$T_k(x) = W_k(x) - \frac{1}{\lambda p_k}\frac{dW_k}{dx}(x), \qquad 1 \leq k \leq K \tag{2.3.2}$$

holds for every $x \geq 0$ where $W_k(x)$ is differentiable.

The central result of this section is the following theorem.

Theorem II.3.3. *The distribution functions $T(\cdot)$, $R(\cdot)$ and $S(\cdot)$ are given by*

$$T(x) = \sum_{k=1}^{K} p_k T_k(x) , \tag{2.3.3a}$$

$$R(x) = \sum_{k=1}^{K} p_k \int_0^\infty \prod_{\substack{l=1 \\ l\neq k}}^{K} W_l(x+t) dT_k(t) \tag{2.3.3b}$$

and

$$S(x) = \prod_{k=1}^{K} W_k(x) - \frac{1}{\lambda}\frac{d}{dx}\left(\prod_{k=1}^{K} W_k(x)\right) \tag{2.3.3c}$$

for every $x \geq 0$ where $W_k(x)$, $1 \leq k \leq K$, are all differentiable. Their means are given by

$$ET = \sum_{k=1}^{K} p_k \frac{2 - \rho_k(1 - \sigma_k^2\mu_k^2)}{2(\mu_k - \lambda p_k)} , \tag{2.3.4a}$$

$$ER = ES - ET \tag{2.3.4b}$$

and

$$ES = \int_0^\infty (1 - \prod_{k=1}^{K} W_k(x))dx + \frac{1}{\lambda}(1 - \prod_{k=1}^{K}(1-\rho_k)) . \tag{2.3.4c}$$

Proof. Conditioning (2.3.1*a*) on the value of U gives (2.3.3*a*), from which (2.3.4a) follows by a simple application of the Pollaczek-Khinchin mean value formula [Kle.a, p. 190].

Equation (2.3.3*b*) follows from (2.3.1*b*) by conditioning on U, and noting that

$$P([A]^+ \le x) = P(A \le x) , \qquad x \ge 0 ,$$

and (2.3.4*b*) obvious since $S = T + R$.

To obtain the distribution of S, note that the RVs in the right-hand side in (2.3.1c) are conditionally mutually independent given U. The computation, which makes use of Lemma II.3.2, proceeds as follows: For all $x \ge 0$ where the functions W_k, $1 \le k \le K$, are all differentiable, it is plain that

$$\begin{aligned} S(x) &= \sum_{k=1}^{K} p_k T_k(x) \prod_{\substack{l=1 \\ l \ne k}}^{K} W_l(x) \\ &= \sum_{k=1}^{K} p_k (W_k(x) - \frac{1}{\lambda p_k}\frac{dW_k}{dx}(x)) \prod_{\substack{l=1 \\ l \ne k}}^{K} W_l(x) \\ &= \prod_{k=1}^{K} W_k(x) - \frac{1}{\lambda}\sum_{k=1}^{K} \frac{dW_k}{dx}(x) \prod_{\substack{l=1 \\ l \ne k}}^{K} W_l(x) , \end{aligned}$$

and this rewrites as (2.3.3*c*). Equation (2.3.4*c*) then follows by routine integration since $ES = \int_0^\infty (1 - S(x))dx$ and $W_k(0) = 1 - \rho_k$, $1 \le k \le K$.

□

Comparison of the formulas for $T(x)$ and $S(x)$, or for ET and ES, quickly reveals how the resequencing requirement complicates the analysis. In particular,

whereas ET is expressed in terms of the first two moments of the service times, the distributions of the waiting times are needed to compute ES. This point is made more apparent when considering the expressions of ET and ES when the service times are all exponentially distributed.

Remark II.3.1. Assume the servers to all have exponential service distributions. In that case

$$T_k(x) = 1 - e^{-(\mu_k - \lambda p_k)x} \quad \text{and} \quad W_k(x) = 1 - \frac{\lambda p_k}{\mu_k} e^{-(\mu_k - \lambda p_k)x}, \quad 1 \le k \le K,$$

for all $x \ge 0$, and equations (2.3.4a, c) become

$$ET = \sum_{k=1}^{K} \frac{p_k}{\mu_k - \lambda p_k} \tag{2.3.5a}$$

$$ES = \int_0^\infty \left(1 - \prod_{k=1}^{K} \left(1 - \frac{\lambda p_k}{\mu_k} e^{-(\mu_k - \lambda p_k)x}\right)\right) dx + \frac{1}{\lambda}\left[1 - \prod_{k=1}^{K}(1 - \frac{\lambda p_k}{\mu_k})\right] \tag{2.3.5b}$$

$$= \sum_{k=1}^{K}(-1)^{k+1} \sum_{I \in \mathcal{I}_k} \prod_{i \in I} \left(\frac{\lambda p_i}{\mu_i}\right) \frac{1}{\sum_{i \in I}(\mu_i - \lambda p_i)} + \frac{1}{\lambda}\left[1 - \prod_{k=1}^{K}(1 - \frac{\lambda p_k}{\mu_k})\right], \tag{2.3.5c}$$

where

$$\mathcal{I}_k := \{I \subset \{1, \ldots, K\} : |I| = k\}.$$

Formula (2.3.5c) is better suited for computations for values of K less than approximately 15. For larger values of K, numerical integration of (2.3.5b) appears to be more economical and accurate.

Remark II.3.2. Comparison of the formulas (2.3.3b) and (2.3.3c) shows that the system time has more pleasing properties than the resequencing delay. This appears to be a general phenomenon in the study of resequencing systems (see [BGP] and [Var.a]). For instance, in contrast to (2.3.3c), it was not possible to

eliminate the T_k's in (2.3.3*b*) by using Lemma II.3.2. Similarly, the easiest way to obtain ER is through equation (2.3.4*b*).

Remark II.3.3. The probability of being a *star job*, i.e., of not being resequenced, is

$$P(*) = R(0) = \sum_{k=1}^{K} p_k \int_0^\infty \prod_{\substack{l=1 \\ l \neq k}}^{K} W_l(t) dT_k(t),$$

and has no simple expression in general although special cases are given in Chapter IV.

Remark II.3.4. If

$$W := \max\{W_1, \ldots, W_K\} ,$$

then (2.3.3*c*) and (2.3.4*c*) can be rewritten as

$$S(x) = W(x) - \frac{1}{\lambda} \frac{dW}{dx}(x) , \quad x \geq 0 ,$$

whenever W is differentiable, and

$$ES = EW + \frac{1}{\lambda} P(\text{system not empty}) .$$

Thus, the equation for $S(\cdot)$ is similar to (2.3.2) and seems to suggest that the system admits an equivalent $M/GI/1$ representation, in the sense that $S = W + B$ for some RV B *independent* of W. Unfortunately, such a representation does not seem possible.

Remark II.3.5. The expressions for T, R and S depend only on the waiting and response time distributions of the parallel $M/GI/1$ queueing systems, which are well known at least by their Laplace transforms. If the distributions $B_k(\cdot)$ are of *phase type (PH-type)*, then so are the distributions $W_k(\cdot)$ and $T_k(\cdot)$ (see Appendix II). Since the maximum of independent PH-type distributions is again of PH-type [Neu.b], the distributions of T, R and S will also be of PH-type, in view of Theorem II.3.1. Although, the dimensionality problem in obtaining the maximum of PH-type RVs limits such computations to small values of K, Theorem

II.3.1 may serve to guide approximations in obtaining the performance measures of interest for larger values of K.

II.4. OPTIMIZATION OF THE SYSTEM

The problem of finding the optimal vector of switching probabilities that minimizes the expected response and system times is studied in subsections II.4.1 and II.4.2, respectively. Comparison of both optimization problems provides a better comprehension of the effects of resequencing.

II.4.1. Minimizing the Response Time

The nonlinear program

$$\min_{(p_1,\ldots,p_K)\in\mathcal{D}} ET(p_1,\ldots,p_K)$$

has already been addressed in various forms in the literature, e.g., the Capacity Assignment problem in [Kle.b] or [Kel, Ch.4]. In the present context, it has been solved by Buzen and Chen [BuC]. Nevertheless, it is briefly discussed here in order to compare its solution to the corresponding minimization problem for ES.

The following theorem states that the RV T is stochastically convex with respect to the load allocation vector in the strong *st* sense defined in Appendix II, thus providing a rationale to the intuitive arguments given for the algorithm in [BuC].

Theorem II.4.1. *For all $x \geq 0$, the mapping $p \mapsto T(p,x)$ is (strictly) concave on $\mathcal{D}$, i.e., $\{T(p),\ p \in \mathcal{D}\} \in SCX(st)$.*

Proof. It suffices to show that for every $x \geq 0$, the Hessian matrix of the mapping $p \mapsto T(p,x) = \sum_{k=1}^{K} p_k T_k(p_k,x)$ is negative definite on $\{p \in [0,1]^K;\ \lambda p_k < \mu_k,\ 1 \leq k \leq K\} \supset \mathcal{D}$. Note that this Hessian is diagonal. Furthermore, the diagonal elements are strictly negative since by Theorem A.II.4 each one of the mappings $p_k \mapsto T_k(p_k,x),\ 1 \leq k \leq K$, is concave and decreasing for all $x \geq 0$, and the result thus follows.

□

The following corollary is now immediate by Definition A.II.1.

Corollary II.4.2. *If $\lambda < \mu$, the constrained optimization problem* $\min_{p \in \mathcal{D}} ET(p)$ *has a unique solution.*

Denote this optimum by $p^* = (p_1^*, \ldots, p_K^*)$, and let y be the unique root of the nonincreasing function

$$f(x) = 1 \; - \; \sum_{k=1}^{K} \frac{\mu_k}{\lambda} \left[1 - \sqrt{\frac{1 + \sigma_k^2 \mu_k^2}{2\mu_k x - 1 + \sigma_k^2 \mu_k^2}} \right],$$

defined for $x \geq \max_{1 \leq k \leq K} [(1 - \sigma_k^2 \mu_k^2)/2\mu_k]^+$. It is an easy exercise to show that if $y \geq 1/\mu_k$ for all $1 \leq k \leq K$, then a straightforward Lagrange analysis shows that p^* is given by

$$p_k^* \; = \; \frac{\mu_k}{\lambda} \left[1 - \sqrt{\frac{1 + \sigma_k^2 \mu_k^2}{2\mu_k y - 1 + \sigma_k^2 \mu_k^2}} \right], \qquad 1 \leq k \leq K, \tag{2.4.1}$$

and that it lies in the set $\mathcal{D}$.

On the other hand, if $1/\mu_k < y$, then p_k^* given in (2.4.1) is negative, and by Theorem II.4.1 the minimum is located on the boundary of $\mathcal{D}$, i.e., at least one of the p_k's is 0. The dimension of the problem is thus reduced by at least one and this leads to the algorithm given in Appendix I to compute the vector p^*. Several monotonicity results that yield considerable computational savings in this algorithm are also presented in Appendix I.

Remark II.4.1. When the service time distributions are exponential, $\sigma_k \mu_k = 1$ and (2.4.1) reduces to

$$\lambda p_k^* = \mu_k \; - \; \sqrt{\mu_k} \, \frac{(\mu - \lambda)}{\sum_{i=1}^{K} \sqrt{\mu_i}}, \qquad 1 \leq k \leq K \, . \tag{2.4.2}$$

This dramatically simplifies the computation of p^* as indicated in Appendix I.

Remark II.4.2. Another simplification occurs when all the servers are identical (but otherwise general), in which case

$$p_k^* = \frac{1}{K}, \qquad 1 \leq k \leq K \ ,$$

so that sharing the load equally among the servers is optimal.

The following theorem generalizes this last remark.

Theorem II.4.3. *When the service time distributions are all identical, the probability vector* $p^* = (1/K, \ldots, 1/K)$ *stochastically minimizes the RV* $T(p)$*, i.e.,* $T(p^*) \leq_{st} T(p)$ *for every* p *in* $\mathcal{D}$.

Proof. Since the service distributions are all identical, the distribution functions $T_k(\cdot)$ differ only by the input rate λp_k, i.e., $T_k(p_k, x) = T_1(p_k, x)$, $1 \leq k \leq K$. Therefore, for all $x \geq 0$,

$$T(p,x) = \sum_{k=1}^{K} p_k T_1(p_k, x) \leq T_1(\sum_{k=1}^{K} p_k^2, x) \leq T_1(\sum_{k=1}^{K} \frac{1}{K^2}, x) = T(p^*, x) \ .$$

The first inequality follows from the concavity of the mapping $p \mapsto T_1(p,x)$ for every $x \geq 0$, while the second inequality is a consequence of the fact that

$$\text{argmin}\{\sum_{k=1}^{K} p_k^2;\ p \in \mathcal{D}\} = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$$

since the $\mathbb{R}_+ \mapsto \mathbb{R}_+$ mapping $p \to T_1(p,x)$ is monotone decreasing by Lemma A.II.4.

□

II.4.2. Minimizing the System Time

The complexity of the expression (2.3.4c) makes the solution to the problem

$$\min_{(p_1, \ldots, p_K) \in \mathcal{D}} ES(p_1, \ldots, p_K)$$

much more difficult. Even in the exponential case, the problem is too complicated to be solved "by hand" except for the case $K = 2$. In this case, a solution is presented in the form of the root of a 5^{th} order polynomial. However, the numerical results obtained from the solution of this special case lead to the idea of a simple asymptotic approximation which generalizes to the case $K > 2$. This is discussed in Section II.5.

Another case where the optimization problem can be solved is when all the servers are identical, but otherwise general. For this configuration, it is shown that equal routing probabilities still achieve the minimal average system time.

II.4.2a. The Case of Two Exponential Servers

In the rest of this chapter, when $K = 2$, $\mu_1 \geq \mu_2$ is assumed without loss of generality, and the notation $p = p_1$ and $\overline{p} = 1 - p_1 = p_2$ is used. The set $\mathcal{D}$ can then be parametrized by the scalar p so that the statement "$p \in \mathcal{D}$" is now equivalent to "$(p, \overline{p}) \in \mathcal{D}$".

With this notation, the average system time given in (2.3.5c) takes the form

$$ES(p) = \frac{1}{\mu_1 - \lambda p} + \frac{1}{\mu_2 - \lambda \overline{p}} - \frac{p}{\mu_2} - \frac{\overline{p}}{\mu_1} - \frac{\lambda p \overline{p}(\mu_1 + \mu_2)}{\mu_1 \mu_2 (\mu_1 + \mu_2 - \lambda)} \,. \tag{2.4.3}$$

Differentiating (2.4.3) leads after some simplifications to the relation

$$\frac{d}{dp} ES(p) = \frac{N(p)}{D(p)}$$

with

$$D(p) \;=\; \mu_1 \mu_2 (\mu_1 - \lambda p)^2 (\mu_2 - \lambda \overline{p})^2 (\mu_1 + \mu_2 - \lambda)$$

and

$$N(p) \;=\; (\mu_1 - \lambda p)^2 (\mu_2 - \lambda \overline{p})^2 \, a(p) + \lambda \mu_1 \mu_2 (\mu_1 + \mu_2 - \lambda)^2 \, b(p) \;,$$

where the *linear increasing* functions $a(p)$ and $b(p)$ are given respectively by

$$a(p) \;=\; (\mu_1 + \mu_2 - \lambda)(\mu_2 - \mu_1) - \lambda(1 - 2p)(\mu_1 + \mu_2)$$

and

$$b(p) \;=\; \mu_2 - \mu_1 - \lambda(1-2p) \quad .$$

Note that $D(p) > 0$ under the stability assumptions (2.2.1).

The scalars a_0 and b_0 given by

$$a_0 \;=\; b_0 - \frac{\mu_1 - \mu_2}{2(\mu_1 + \mu_2)} \quad \text{and} \quad b_0 \;=\; \frac{1}{2} + \frac{\mu_1 - \mu_2}{2\lambda}$$

are the unique roots of the equations $a(p) = 0$ and $b(p) = 0$, respectively. Since $\mu_1 \geq \mu_2$ and $\lambda < \mu_1 + \mu_2$, it is easy to check that $\frac{1}{2} \leq a_0 \leq b_0 < \mu_1/\lambda$.

Lemma II.4.4. *The fifth order numerator polynomial $N(p)$ has a unique root r_0 in the interval $\mathcal{I} = (1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda})$. Furthermore this root lies in the interval* $[\max\{a_0, 1 - \frac{\mu_2}{\lambda}\}, b_0]$.

Proof. First, the following observations are made:

(i) For all $p < a_0$, $a(p) < 0$ and $b(p) < 0$, so that $N(p) < 0$, while for all $p > b_0$, $N(p) > 0$ by the same arguments, with $N(p) = 0$ if and only if $p = a_0 = b_0$.

(ii) For $p \leq b_0$, $(1 - \frac{\mu_2}{\lambda}, b_0] \subset \mathcal{I}$, since $b_0 < \mu_1/\lambda$.

(iii) The derivative $N'(p)$ is given by

$$N'(p) = -2\lambda(\mu_1 - \lambda p)(\mu_2 - \lambda\overline{p})a(p)b(p) + a'(p)(\mu_1 - \lambda p)^2(\mu_2 - \lambda\overline{p})^2 + b'(p)\lambda\mu_1\mu_2(\mu_1 + \mu_2 - \lambda)^2 \quad .$$

Two cases have to be studied:

If $1 - \frac{\mu_2}{\lambda} < a_0$, then $[a_0, b_0] \subset \mathcal{I}$ from (ii). According to (iii), $N'(p) > 0$ for every p in (a_0, b_0), since there $a(p)b(p) < 0$ and the last two terms of $N'(p)$ are always positive. The result then follows from (i) since $N(p)$ is monotone increasing in $[a_0, b_0]$.

On the other hand, if $1 - \frac{\mu_2}{\lambda} \geq a_0$, then $N'(p) > 0$ for every p in $[1 - \frac{\mu_2}{\lambda}, b_0]$ by a similar reasoning. Therefore, since $N(1 - \frac{\mu_2}{\lambda}) < 0$ and $N(p) > 0$ for $p > b_0$ by invoking (i), it is plain that $N(p)$ has a unique root r_0 in $(1 - \frac{\mu_2}{\lambda}, b_0] \subset \mathcal{I}$.

The result thus follows by combining these two cases.

□

Corollary II.4.5. *If $\mu > \lambda$, the optimization problem*

$$\min_{(p_1,p_2)\in\mathcal{D}} ES(p_1,p_2)$$

has a unique solution $(p_1^\dagger, p_2^\dagger)$, where

$$p_1^\dagger \;=\; \min\{1, r_0\} \quad \text{and} \quad p_2^\dagger \;=\; 1 - p_1^\dagger \;.$$

Proof. If $r_0 > 1$, then $N(p) < 0$ for all p in $\mathcal{D}$. Therefore, $ES(p)$ decreases monotonically in $\mathcal{D}$, and is minimal when $p = 1$.

On the other hand, if $r_0 \leq 1$, then $ES(p)$ is minimal at $p = r_0$.

□

Remark II.4.3. When $K = 2$ with exponential servers, the average resequencing time takes the form

$$ER(p) = \frac{\overline{p}}{\mu_1 - \lambda p} + \frac{p}{\mu_2 - \lambda\overline{p}} - \frac{p}{\mu_2} - \frac{\overline{p}}{\mu_1} - \frac{\lambda p\overline{p}(\mu_1+\mu_2)}{\mu_1\mu_2(\mu_1+\mu_2-\lambda)} \;.$$

Although the form of $ER(p)$ is very similar to that of $ES(p)$ given in (2.4.3), the optimization problem for $ER(p)$ is quite different. For instance, when $\mu_1 = \mu_2$, elementary arguments show that $p = 1/2$ is a local minimum (resp. maximum) for ER when $\rho > \sqrt{2}-1$ (resp. $\rho \leq \sqrt{2}-1$). Furthermore, for $\rho \geq 1/2$, $p = 1/2$ is the *only* solution of the equation $d\,ER(p)/dp = 0$, whence it is the global minimum. On the other hand, it is plain that for $\rho < 1/2$, $p = 1$ in $\mathcal{D}$ is the global minimum, since $0 = ER(1) < ER(1/2)$. Therefore,

$$\arg\min_{p\in\mathcal{D}} ER(p) = \begin{cases} 1, & \text{if } \rho < 1/2, \\ 1/2, & \text{if } \rho \geq 1/2 \;. \end{cases}$$

Remark II.4.4. When $\mu_1 = \mu_2$, $a_0 = b_0 = 1/2 = p_1^\dagger$ so that $ET(p)$ and $ES(p)$ are both minimized when the load is equally allocated between the servers. This was not evident *a priori*, since $ER(p)$ may be maximum in this configuration, e.g., as would be the case when $\rho \leq \sqrt{2} - 1$ (see also [JeM.b] and [IL.c]).

II.4.2b. Optimal Routing in the Homogeneous Case

In this subsection, the system is studied when all the servers have identically distributed service times. It has been noted in Remark II.4.1 that making all routing probabilities equal minimize the average response time ET for the homogeneous system. The following theorem states that this result still holds for ES and generalizes the observation made in Remark II.4.4. This shows that the response time dominates the resequencing time.

Theorem II.4.6. *When the service time distributions are all identical, the probability vector* $p^* = (1/K, \ldots, 1/K)$ *minimizes the function* $ES(p)$.

Proof. As noted in the proof of Theorem II.4.3, the distribution functions $W_k(\cdot)$ differ only by the input rate λp_k, so that $W_k(p_k, \cdot) = W_1(p_k, \cdot)$, $1 \leq k \leq K$. Equation (2.3.4c) can be rewritten as

$$ES(p) = \int_0^\infty (1 - \prod_{k=1}^{K} W_1(p_k, x))dx + \frac{1}{\lambda}(1 - \prod_{k=1}^{K}(1 - \rho_0 p_k)) . \qquad (2.4.4)$$

It is shown now that both terms in the right-hand side of (2.4.4) are minimum at p^*, so that their sum is therefore minimal at this point.

Strict convexity of the function $x \mapsto \log(1 - \rho_0 x)$ and Lemma A.III.1 of Appendix III imply that the function $\log(\prod_{k=1}^{K}(1 - \rho_0 p_k))$, whence the product $\prod_{k=1}^{K}(1 - \rho_0 p_k)$, is maximum at p^*. The second term is therefore minimum at this point.

On the other hand, it is easy to show by Lemma A.II.4 of Appendix II that for every $x \geq 0$, the positive function $p \mapsto \log W_1(p, x)$ is strictly concave. Therefore, by the same argument, the point p^* maximizes the product $\prod_{k=1}^{K} W_1(p_k, x)$ for all $x \geq 0$, and minimizes the integral in the first term.

□

Remark II.4.5. Since $p^* = (\frac{1}{K}, \ldots, \frac{1}{K})$ maximizes the product $\prod_{k=1}^{K} W_k(p_k, x)$ for all $x \geq 0$, the RV $\max\{W_k;\ 1 \leq k \leq K\}$ is stochastically *smallest* at this point, i.e., when the RVs W_k, $1 \leq k \leq K$, are *i.i.d.* This observation and (2.3.3c) suggest that a stronger (stochastic) optimality result as in Theorem II.4.3 also holds for the RV $S(p)$. This is currently under investigation.

II.4.3. Examples and Comparisons: K = 2

This section aims at a better understanding of the variation of the optimum switching probabilities with the system parameters. For the case $K = 2$ with exponential servers, the notation $p^*(\lambda, \mu_1, \mu_2)$ and $p^\dagger(\lambda, \mu_1, \mu_2)$ is used for p^* and $p^\dagger$ given in (2.4.2) and Corollary II.4.5, respectively, in order to explicitly indicate their dependence on the system parameters. Figure II.1 displays the sets

$$\Gamma_p = \{(\mu_1, \mu_2) \mid \lambda p < \mu_1,\ \lambda(1-p) < \mu_2 \text{ and } p_1^*(\lambda, \mu_1, \mu_2) = p\ \}$$

and

$$C_p = \{(\mu_1, \mu_2) \mid \lambda p < \mu_1,\ \lambda(1-p) < \mu_2 \text{ and } p_1^\dagger(\lambda, \mu_1, \mu_2) = p\ \}\ ,$$

for $0 \leq p \leq 1$ in the (μ_1, μ_2) plane when $\lambda = 1$.

Only the sets Γ_p and C_p, for p ranging from 0.5 to 1 with increments of 0.05, are drawn in Figures II.1a and II.1b, respectively. The sets for values of p smaller than 0.5 follow from symmetry by interchanging μ_1 and μ_2. The dashed curves are the lines $\mu_1 + \mu_2 = \lambda$ (the stability limit), $\mu_1 = \mu_2$ ($\Gamma_{0.5}$ and $C_{0.5}$), $\mu_1 = \mu_2 - 2\lambda$ (the asymptote for the boundary curve of Γ_1) and $\mu_1 = \mu_2 - \lambda$ (the asymptote for the boundary curve of C_1).

An important observation in Figures II.1a and II.1b is that the sets Γ_p and C_p, $0 < p < 1$, are smooth curves. These curves start from the point $(\lambda p, \lambda(1-p))$, since the set $\mathcal{D}$ shrinks to the point $(p_1, p_2) = (\mu_1/\mu, \mu_2/\mu)$ as μ tends to λ (high utilization). This provides a trivial approximation for a heavily loaded system. The points located under the lowest curve in Figures II.1a and II.1b are points

of the sets Γ_1 and C_1, respectively, and correspond to systems where the slower server is not used at all, i.e., $p_1^* = 1$ or $p_1^\dagger = 1$.

Figure II.2 (resp. II.3) provide the same comparison for a DS where an $M/M/1$ queue is in parallel with an $M/D/1$ (resp. $M/Geo/1$) queue. The abbreviation "*Geo*" denotes a geometric distribution, defined by $P(B = k\Delta) = q^k(1-q)$, where Δ is a time constant and q a fixed probability. For this distribution $\mu = (1-q)/\Delta q$ and $\mu^2\sigma^2 = (1+\Delta\mu)$. The parameter Δ is taken to be $1/\lambda$ in Figure II.3. These families of distributions are chosen for their simple relationships between σ and μ.

The sets Γ_p and C_p, for p ranging from 0 to 1 with increments of 0.1 are drawn in Figures II.2 and II.3. As in Figure II.1, points located under the lowest curve are points of Γ_1 (or C_1) and correspond to systems for which the second server is not used. By symmetry, the curve on the top is the boundary of the region Γ_0 (or C_0), and all points located above it correspond to systems where the first server is not used at the optimum.

The points of Γ_p are obtained by solving (2.4.1*b*) numerically. The curves C_p are obtained as follows: Since the first server has exponentially distributed service times, $W_1(\cdot)$ is as given in Remark II.3.1. Therefore, after some simplifications and using the notation of Section II.4.2a, ES reduces to

$$ES(p) = ET_2(p) + \frac{p}{\mu_1} - \frac{p}{\mu_2} - \frac{\lambda p\overline{p}}{\mu_1\mu_2} + \frac{\lambda p}{\mu_1}\frac{W_2^*(\mu_1 - \lambda p)}{\mu_1 - \lambda p},$$

where $W_2^*(\cdot)$ is the Laplace transform of $W_2(\cdot)$. The curves C_p are obtained by solving the equation $d\ ES(p)/dp = 0$ numerically.

The comparison of the curves Γ_p and C_p in Figures II.1-II.3 reveals that $p_1^* \leq p_1^\dagger$ for all values of the system parameters, i.e., the presence of resequencing always decreases the optimal utilization of the slower server.

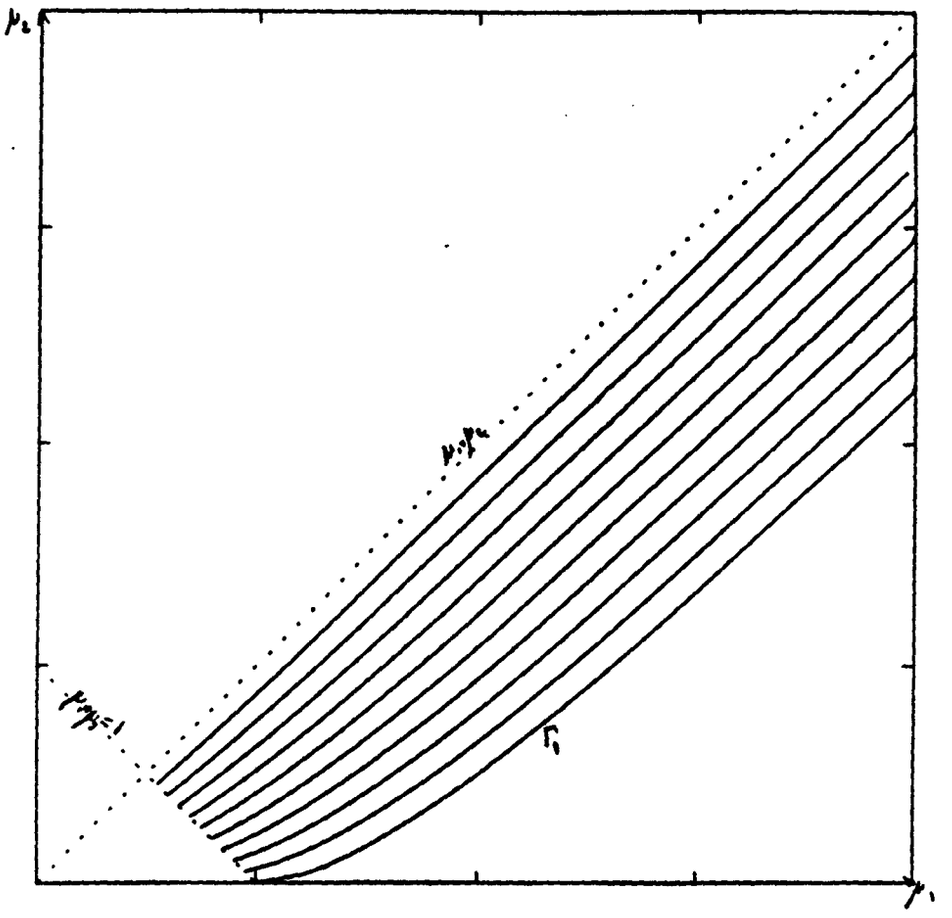


Figure II.1a

The curves Γ_p for $p = 0.5, 0.55, \ldots, 0.95, 1.0$

$M/M/1$ in parallel with $M/M/1$

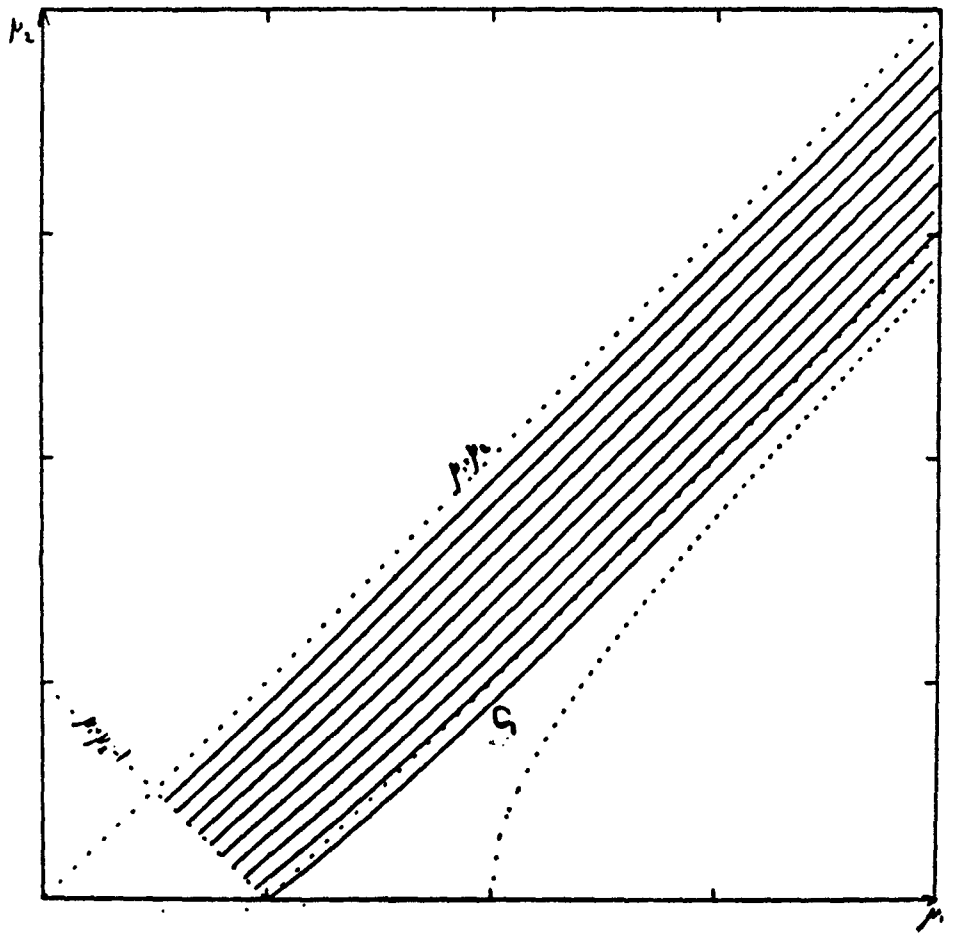


Figure II.1b

The curves C_p for $p = 0.5, 0.55, \ldots, 0.95, 1.0$

$M/M/1$ in parallel with $M/M/1$

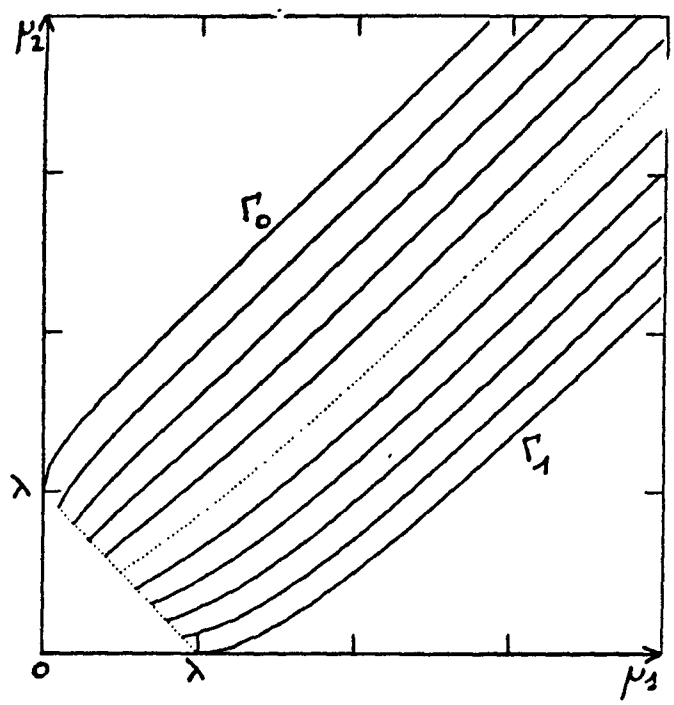


Figure II.2a

The curves Γ_p for $p = 0, 0.1, \ldots, 0.9, 1.0$

$M/M/1$ in parallel with $M/D/1$

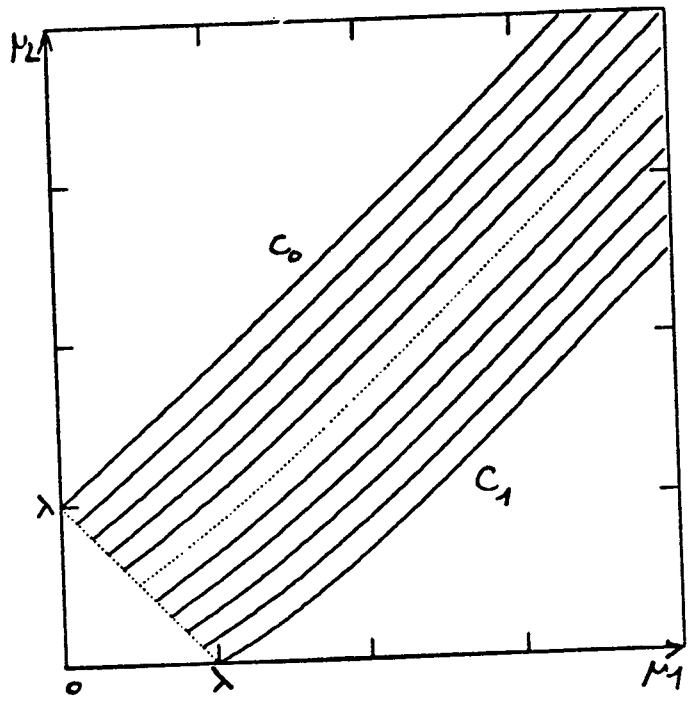


Figure II.2b

The curves C_p for $p = 0, 0.1, \ldots, 0.9, 1.0$

$M/M/1$ in parallel with $M/D/1$

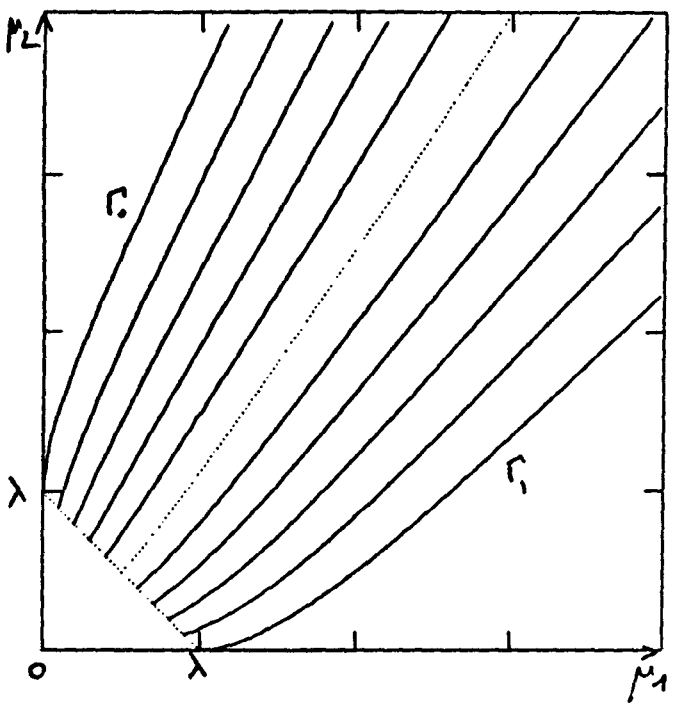


Figure II.3a

The curves Γ_p for $p = 0, 0.1, \ldots, 0.9, 1.0$

M/M/1* in parallel with *M/Geo/1

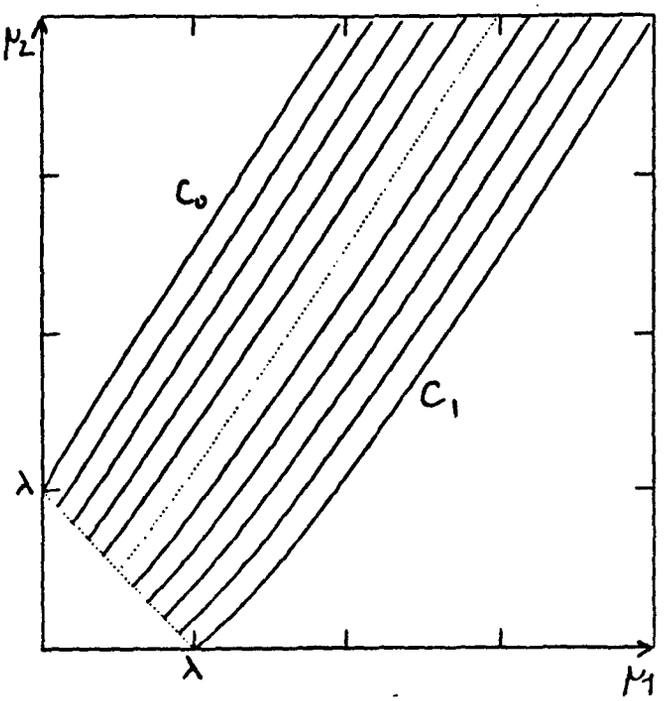


Figure II.3b

The curves C_p for $p = 0, 0.1, \ldots, 0.9, 1.0$

$M/M/1$ in parallel with $M/Geo/1$

II.5. APPROXIMATIONS FOR EXPONENTIAL SERVERS

This section develops approximations for the optimal allocation probabilities $p^\dagger$ and p^* when all the service times are exponentially distributed. The approximations are motivated from Figures II.1a and II.1b, where the level curves Γ_p and C_p, $0 \leq p \leq 1$, have asymptotes that are parallel to the line $\mu_1 = \mu_2$. More importantly, both the curves C_p and their asymptotes remain close to one another for all values of the system parameters. Therefore, the idea behind the approximations is to replace the curves C_p by their asymptotes.

The approximations for the case $K = 2$ are obtained as a natural extension to the discussion in Section II.4.2 and are illustrated separately as they provide a better understanding of the key ideas. Extensions to the case of $K \geq 2$ parallel servers are then provided through the asymptotic expansion of the formula (2.3.5c). Although a simple algorithm is available in Appendix I for computing the vector p^*, a similar approximation is also derived in order to see the effect of resequencing more clearly. The accuracy of these approximations is validated for a wide range of system parameters.

II.5.1. The Case $K = 2$

In this subsection, the notation and definitions given in Section II.4.2a are again adopted. The asymptotes to the curves C_p in Figure I.1a can be obtained directly by noting that

$$a(p) = (\mu_1 + \mu_2)b(p) - \lambda(\mu_2 - \mu_1)$$

and rewritting the equation defining C_p as

$$\frac{N(p)}{\mu_1^5} = 0 = b(p)\left[(1+\frac{\mu_2}{\mu_1})(1-p\frac{\lambda}{\mu_1})^2(\frac{\mu_2}{\mu_1}-\overline{p}\frac{\lambda}{\mu_1})^2 + \frac{\lambda}{\mu_1}\frac{\mu_2}{\mu_1}(1+\frac{\mu_2}{\mu_1}-\frac{\lambda}{\mu_1})^2\right]$$

$$-\lambda(\frac{\mu_2}{\mu_1}-1)(1-p\frac{\lambda}{\mu_1})^2(\frac{\mu_2}{\mu_1}-\overline{p}\frac{\lambda}{\mu_1})^2 \ .$$

Letting μ_1 go to $+\infty$ with μ_2/μ_1 going to some constant α, it follows that $\alpha = 1$, so that the line

$$b(p) = \mu_2 - \mu_1 - \lambda(1-2p) = 0$$

is the asymptote to the curve C_p. As observed from Figure II.1a, the asymptote $b(p) = 0$ is parallel to the $\mu_1 = \mu_2$ line and passes through the point $(\lambda p, \lambda(1-p))$. Since the curve C_p also passes through this point, C_p approaches $b(p)$ under both heavy and light loads, explaining why the approximations are good for all values of the system parameters as indicated in Table II.1.

A similar argument shows that the line $\mu_2 - \mu_1 - 2\lambda(1-2p) = 0$ is the asymptote to the curve Γ_p. The approximations $\hat{p}_1^\dagger$ and $\hat{p}_1^*$ to $p_1^\dagger$ and p_1^* are therefore given by

$$\hat{p}_1^\dagger = \begin{cases} \frac{1}{2} + \frac{\mu_1 - \mu_2}{2\lambda} & \text{if } \mu_1 - \mu_2 < \lambda, \\ 1 & \text{otherwise,} \end{cases} \tag{2.5.1a}$$

and

$$\hat{p}_1^* = \begin{cases} \frac{1}{2} + \frac{\mu_1 - \mu_2}{4\lambda} & \text{if } \mu_1 - \mu_2 < 2\lambda, \\ 1 & \text{otherwise.} \end{cases} \tag{2.5.1b}$$

Remark II.5.1. When $p_1 = \hat{p}_1^\dagger$, both queues are stable, so that this approximation may be used for all values of μ_1 and μ_2 satisfying the condition $\lambda < \mu$. On the other hand, the approximation $\hat{p}_1^*$ is valid only if $2\lambda < \mu_1 + 3\mu_2$.

Remark II.5.2. Note that $\hat{p}_1^\dagger = b_0$, whence by Lemma II.4.4 this is always *greater* than $p_1^\dagger$. On the other hand, $\hat{p}_1^*$ can be shown to be *smaller* than p_1^*.

Remark II.5.3. The approximations are exact when $\mu_1 = \mu_2$ and get better as the service rates get closer to each other.

In order to quantify the accuracy of the approximations, numerical examples are collected in Table II.1 by setting $\lambda = 1$. $ET(p^\dagger)$ is also displayed to see the effect of resequencing on the performance measure.

(μ_1, μ_2)	$p_1^\dagger$	$\hat{p}_1^\dagger$	$\%\epsilon$	$ES(p^\dagger)$	$ES(\hat{p}^\dagger)$	$\%\epsilon$	$ET(p^\dagger)$
(1,0.25)	0.8693	0.875	0.66	10.15	10.19	0.34	7.77
(1,0.50)	0.7425	0.750	1.00	5.12	5.13	0.15	3.95
(1,0.75)	0.6196	0.625	0.87	3.39	3.40	0.35	2.66
(1,1)	0.5000	0.500	0.00	2.50	2.50	0.00	2.00
(2.5,1.5)	0.9532	1.000	4.91	0.665	0.667	0.31	0.648
(2.5,1.75)	0.8405	0.875	4.10	0.647	0.648	0.14	0.607
(2.5,2)	0.7273	0.750	3.12	0.619	0.620	0.05	0.568
(2.5,2.25)	0.6138	0.625	1.83	0.586	0.586	0.01	0.532
(2.5,2.5)	0.5000	0.500	0.00	0.550	0.550	0.00	0.500
(10,9.00)	0.9787	1.000	2.18	0.1111	0.1111	6.0e-3	0.1110
(10,9.25)	0.8592	0.875	1.84	0.1108	0.1108	3.0e-3	0.1095
(10,9.50)	0.7396	0.750	1.41	0.1102	0.1102	1.0e-3	0.1081
(10,9.75)	0.6198	0.625	0.83	0.1092	0.1092	3.0e-4	0.1067
(10,10.0)	0.5000	0.500	0.00	0.1079	0.1079	0.0e-0	0.1053

Table II.1.

Asymptotic Approximations for K = 2

The approximation $\hat{p}_1^\dagger$ yields a relative error of less than 5%. Moreover, the relative errors in ES are much smaller than 1%. For p_1^*, this approximation is not as good for small values of μ_1 and μ_2, for which the curves Γ_p and their asymptote are far apart. However, the curves in Figures II.1-3 indicate that such approximations for p_1^* get better as the service time distributions become less variable. Of course, the easily computable closed form expression given by equation (2.4.2) can always be used. Comparison of $ES(p^\dagger)$ and $ET(p^\dagger)$ shows that resequencing degrades the system performance considerably for $\rho > 0.5$.

II.5.2. The Case K $\geq$ 2

For λ fixed, the optimization problems

$$\min_{p\in\mathcal{D}} ET(p) \quad \text{and} \quad \min_{p\in\mathcal{D}} ES(p)$$

are considered in this section when each μ_k goes to infinity.

It is an easy exercise [Lue] to show that the optimal probability vector $p^*(\mu_1,\ldots,\mu_K)$ (similarly $p^\dagger(\mu_1,\ldots,\mu_K)$) satisfies the Lagrange equations

$$\frac{\partial}{\partial p_k} ET(p) - \gamma^{(1)} - \gamma_k^{(2)} = 0\ , \qquad 1 \le k \le K$$

and

$$\sum_{k=1}^{K} \gamma_k^{(2)} p_k = 0$$

where the constant $\gamma_k^{(2)} \ge 0,\ 1 \le k \le K$.

Let $I^* := \{k :\ p_k^* > 0\}$. Then, $\gamma_k^{(2)} = 0$ for every k in I^*, so that p^* satisfies the equations

$$\frac{\partial}{\partial p_k} ET(p) = \frac{\partial}{\partial p_l} ET(p) \tag{2.5.2}$$

for every k and l in I^*.

Using (2.3.5), the partial derivatives in (2.5.2) are given by

$$\frac{\partial}{\partial p_k} ET(p) \ = \ \frac{\mu_k}{(\mu_k - \lambda p_k)^2}\ , \qquad 1 \le k \le K\ . \tag{2.5.3}$$

For $p^*(\mu_1,\ldots,\mu_K) = p$, equations in (2.5.2) define surfaces in the $(\mu_1,\ldots,\mu_K)$-space. As in the previous subsection, the idea is to approximate these surfaces by their asymptotes to provide asymptotically exact approximations to p^*. For the purpose of obtaining these approximations, let

$$\mu_k = \alpha_k \mu_1 + \delta_k + O(\frac{1}{\mu_1}),$$

for some constants α_k and δ_k with $0 \le \alpha_k \le 1,\ 1 \le k \le K$, and $\alpha_1 = 1$ and $\delta_1 = 0$. Also assume, without any loss of generality, that $\mu_1 \ge \mu_2 \ge \ldots \ge \mu_K$,

so that $p_1^* > 0$, since $\mu_k \geq \mu_l$ implies $p_k^* \geq p_l^*$ by (2.4.2). The equations for the asymptotes can now be obtained as the first order term in μ_1 in the asymptotic expansion of (2.5.2) as $\mu_1 \to \infty$. Using (2.5.3) in (2.5.2), it is readily seen that

$$\lambda(p_k^* - p_l^*) \;=\; \frac{1}{2}\mu_1(\alpha_k - \alpha_l) \;+\; \frac{1}{2}(\delta_k - \delta_l) \;+\; o(1)\,, \qquad k, l \text{ in } I^*\,.$$

Summing this equation over all l in I^* leads to

$$\lambda p_k^* = \frac{\lambda}{|I^*|} + \frac{1}{2}\delta_k - \frac{\sum_{l\in I^*}\delta_l}{2|I^*|} + \frac{1}{2}\mu_1\left(\alpha_k - \frac{\sum_{l\in I^*}\alpha_l}{|I^*|}\right) + o(1)\,, \qquad k \text{ in } I^*\,,$$

after further rearrangements, where $|I^*|$ is the cardinality of I^*. It is clear from this equation that unless $\alpha_k = \sum_{l\in I^*}\alpha_l/|I^*|$, p_k^* cannot be in $\mathcal{D}$ as μ_1 goes to infinity. Therefore, the p_k^*'s are all in $\mathcal{D}$ if and only if all the α_k's for k in I^* are equal to each other. Since $\alpha_1 = 1$, $\alpha_k = 1$ for all k in I^*.

When $\alpha_k = 1$, $\delta_k = \mu_k - \mu_1 + O(1/\mu_1)$ and an easy analysis then leads to

$$\lambda p_k^* \;=\; \frac{\mu_k}{2} \;+\; \frac{2\lambda - \mu_{I^*}}{2|I^*|} \;+\; o(1)\,, \qquad k \text{ in } I^*\,, \tag{2.5.4}$$

where $\mu_{I^*} = \sum_{k\in I^*}\mu_k$. Therefore, the following asymptotic approximation is proposed for p^*

$$\hat{p}^* = \begin{cases} \frac{\mu_k}{2\lambda} + \frac{2\lambda - \mu_{\hat{I}^*}}{2\lambda|\hat{I}^*|} & k \in \hat{I}^* \\ 0 & k \notin \hat{I}^*\,, \end{cases} \tag{2.5.5}$$

where $\hat{I}^* := \{k :\ \hat{p}_k^* > 0\}$.

Since the set $\hat{I}^*$ is in turn determined by $\hat{p}^*$, the approximation (2.5.5) defines an implicit equation for $\hat{p}^*$. However, $\hat{p}^*$ can be obtained through the following simple algorithm with $c = 2$. Algoritm II.1 makes use of the relations $\hat{p}_1^* \geq \ldots \geq \hat{p}_K^*$ by (2.5.5) under the assumed order on μ_k, $1 \leq k \leq K$.

Algorithm II.1.

(i) Set $l \leftarrow K$

(ii) Compute

$$p_l = \frac{\mu_l}{c\lambda} + \frac{1}{l}(1 - \sum_{i=1}^{l} \frac{\mu_i}{c\lambda})$$

(iii) If (a) $p_l > 0$: Stop, and set

$$\hat{p}_k^* = \begin{cases} \frac{\mu_k}{c\lambda} + \frac{1}{l}(1 - \sum_{i=1}^{l} \frac{\mu_i}{c\lambda}) & k = 1, \ldots, l \\ 0 & k = l+1, \ldots, K \ . \end{cases}$$

(b) $p_l \leq 0$: Set $\hat{p}_l^* = 0$ and $l - 1 \rightarrow l$, and go to **(ii)**.

Note that the inequality

$$\mu_k > \frac{\mu_{I^*} - 2\lambda}{|I^*|} \ , \tag{2.5.6}$$

necessarily holds for all k in I^*. As in the case $K = 2$, (2.5.6) does not guarantee stability of the system and the approximation (2.5.5) for p^* may not be used if the additional condition

$$\lambda < \frac{\mu_{I^*}}{2} + \frac{|I^*|}{2} \min_{k \in I^*} \{\mu_k\}$$

is not satisfied.

For the resequencing problem,

$$\frac{\partial}{\partial p_k} ES(p) = \frac{1}{\mu_k} \prod_{\substack{j=1 \\ j \neq i}}^{K} (1 - \frac{\lambda p_j}{\mu_j}) \tag{2.5.7a}$$

$$+ \sum_{l=1}^{K} (-1)^{l+1} \sum_{\substack{I \in \mathcal{I}_l \\ i \in I}} \left(\prod_{\substack{j \in I \\ j \neq i}} \frac{\lambda p_j}{\mu_j} \right) \frac{\lambda}{\mu_i} \frac{\lambda p_i + \sum_{j \in I} (\mu_j - \lambda p_j)}{(\sum_{j \in I} \mu_j - \lambda p_j)^2}$$

$$= \frac{1}{\mu_k} - \lambda \sum_{i \neq k} \frac{p_i}{\mu_k \mu_i} + o(\frac{1}{\mu_1^2}) \ . \tag{2.5.7b}$$

Therefore, $p^\dagger$ satisfies the relations

$$\frac{1}{\mu_k} - \lambda \sum_{i \neq k} \frac{p_i^\dagger}{\mu_k \mu_i} + o(\frac{1}{\mu_1^2}) \;=\; \frac{1}{\mu_l} - \lambda \sum_{i \neq l} \frac{p_i^\dagger}{\mu_l \mu_i} + o(\frac{1}{\mu_1^2}) \;,$$

which rewrites as

$$\lambda(p_k^\dagger - p_l^\dagger) \;=\; \mu_k - \mu_l + \lambda \sum_{i \neq k,l} \frac{p_i^\dagger}{\mu_i}(\mu_k - \mu_l) + o(1) \;,$$

for every k and l in $I^\dagger := \{k : \; p_k^\dagger > 0\}$.

By an argument similar to the one given above, $p_k^\dagger$ are all in $\mathcal{D}$ only if $\alpha_k = 1$ for all k in $I^\dagger$, and the asymptotic expansion of $p^\dagger$ is then

$$\lambda p_k^\dagger \;=\; \mu_k \;+\; \frac{\lambda - \mu_{I^\dagger}}{|I^\dagger|} \;+\; o(1) \;, \qquad k \text{ in } I^\dagger \;. \tag{2.5.8}$$

The condition for $p_k^\dagger$ to be in $I^\dagger$ now reads as

$$\mu_k \geq \frac{\mu_{I^\dagger} - \lambda}{|I^\dagger|} \;, \qquad 1 \leq k \leq K \;. \tag{2.5.9}$$

This region has the same shape as the one defined by (2.5.6), but is "twice narrower", due to the factor 2 in (2.5.6). This is a direct generalization of what has been obtained for the case $K = 2$. The constraint (2.5.9) is sufficient for the vector $p^\dagger$ given in (2.5.8) to be in $\mathcal{D}$ and the implicit equations defining the approximation $\hat{p}^\dagger$ can be given by

$$\hat{p}^\dagger = \begin{cases} \frac{\mu_k}{\lambda} \;+\; \frac{\lambda - \mu_{\hat{I}^\dagger}}{\lambda |\hat{I}^\dagger|} & k \in \hat{I}^\dagger \\ 0 & k \notin \hat{I}^\dagger \;, \end{cases} \tag{2.5.10}$$

where $\hat{I}^\dagger := \{k : \; \hat{p}^\dagger > 0\}$. The approximation $\hat{p}^\dagger$ can again be obtained by Algorithm II.1 with $c = 1$.

Equations (2.5.6) and (2.5.9) indicate that p_k^* and $p_k^\dagger$ obtained from (2.5.4) and (2.5.8), respectively, are strictly positive only if the corresponding μ_k is "large enough". In other words, $\hat{p}_k^*$ (resp. $\hat{p}_k^\dagger$) is set to zero, if μ_k does not satisfy (2.5.6) (resp. (2.5.9)), and (2.5.4) (resp. (2.5.8)) has to be applied to the reduced set of queues. Intuitively, this procedure reflects the fact that, when the load is sufficiently low, it is more advantageous to process all the jobs on a few faster servers. Comparison of (2.5.6) and (2.5.10) also shows that the resequencing requirement tends to further decrease the number of slower servers used.

The randomly selected numerical examples in Table II.2 indicate the accuracy of this asymptotic approximation for $K = 3$ and $\lambda = 1$. The exact solution is obtained by solving the corresponding nonlinear programming problem numerically. As in the case $K = 2$, the approximation $ES(\hat{p}^\dagger)$ seems to be an *upper bound* to $ES(p^\dagger)$, with a maximum relative error of only about 1%.

(μ_1, μ_2, μ_3)	$p^\dagger$	$\hat{p}^\dagger$	$ES(p^\dagger)$	$ES(\hat{p}^\dagger)$	$\%\epsilon$
(1.0,0.8,0.5)	(0.549,0.360,0.091)	(0.567,0.367,0.067)	3.027	3.039	0.40
(1.0,0.5,0.1)	(0.743,0.257,0.000)	(0.750,0.250,0.000)	5.118	5.125	0.14
(1.0,0.2,0.2)	(0.855,0.073,0.073)	(0.867,0.067,0.067)	10.01	10.10	0.90
(1.0,0.1,0.1)	(0.926,0.037,0.037)	(0.933,0.033,0.033)	19.77	19.99	1.11
(0.7,0.4,0.3)	(0.562,0.267,0.171)	(0.567,0.267,0.167)	11.85	11.86	0.08
(0.7,0.3,0.2)	(0.631,0.233,0.135)	(0.633,0.233,0.133)	24.86	24.89	0.12
(0.7,.25,.15)	(0.666,0.217,0.117)	(0.667,0.217,0.117)	51.70	51.72	0.04
(1.2,1.0,0.3)	(0.594,0.406,0.000)	(0.600,0.400,0.000)	2.032	2.033	0.05
(2.0,1.0,1.0)	(0.938,0.031,0.031)	(1.000,0.000,0.000)	0.992	1.000	0.81
(2.0,2.0,1.0)	(0.500,0.500,0.000)	(0.500,0.500,0.000)	0.750	0.750	0.00

Table II.2.

Asymptotic Approximations for K = 3

CHAPTER III

QUASI-STATIC LOAD ALLOCATION IN PARALLEL QUEUES WITH RESEQUENCING

III.1. INTRODUCTION

The model of Chapter II is again considered when the service times are exponentially distributed with rate μ_k in queue k, $1 \leq k \leq K$. The Poisson arrival rate is still denoted by λ and the notation $b_k = \mu_k/\lambda$, $1 \leq k \leq K$, is used.

In this chapter, the system parameters b_k, $1 \leq k \leq K$, are all assumed *unknown*. The *approximate* optimal probability vector $\hat{p}^\dagger$ of the Bernoulli switch that minimizes the average system delay is given in Section II.5 as the projection $\mathcal{P}_{\mathcal{D}}(p)$ of the vector p given by

$$p = b + \frac{1}{K}(1 - b^T e)\, e \tag{3.1.1}$$

onto the probability simplex $\mathcal{D}$, where $b = (b_1, \ldots, b_K)$. The simple form of (3.1.1) is utilized here in a stochastic approximation (SA) algorithm for computing the vector $\hat{p}^\dagger$.

The next section provides a brief introduction to SAs and a framework for the proposed algorithm. In Section 3, an idle time measurement model is described to estimate the system parameters (see also [Bok]). The proposed SA algorithm is given in Section 4 and several numerical examples are collected in Section 5. All the RVs in this chapter are defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

III.2 BACKGROUND ON STOCHASTIC APPROXIMATIONS

Let h be a function from $\mathbb{R}^K \rightarrow \mathbb{R}^K$, and consider the problem of finding the *unique* vector $p^\star$ such that

$$h(p^\star) = 0 \, .$$

There are various iterative methods for determining $p^\star$, such as the Newton's method or its variants. These methods require the evaluation of the function $h(\cdot)$ and its derivatives. In many situations, unfortunately one can only observe $h(p)$ corrupted with error or noise, i.e., only $h(p) + \xi$ is available for some RV ξ.

Robbins and Monro [RoM] suggested the following SA procedure for the solution of this problem: Given an arbitrary initial point p_0[1] and a decreasing sequence of positive numbers $\{a_n \;\; n = 0, 1, \ldots\}$ such that

$$\sum_{n=0}^{\infty} a_n^2 < \infty \quad \text{and} \quad \sum_{n=0}^{\infty} a_n = \infty \, , \tag{3.2.1}$$

set

$$p_{n+1} = p_n - a_n z_n \, , \tag{3.2.2}$$

where z_n is the outcome of the measurement at the n^{th} iterate given by

$$z_n = h(p_n) + \xi_n \, , \qquad n = 0, 1, \ldots \, , \tag{3.2.3}$$

where $\{\xi_n, \;\; n = 0, 1, \ldots\}$ is a sequence of conditionally independent zero-mean RVs. Robbins and Monro proved that the sequence $\{p_n, \;\; n = 0, 1, \ldots\}$ obtained from this procedure converge to $p^\star$ under the conditions given in (3.2.1). The first condition in (3.2.1) guarantees that the jumps $p_{n+1} - p_n$ are damped to achieve convergence, while the second condition ensures that the magnitude of the jumps does not decrease too rapidly to allow recovery from a poor choice of p_0.

Since [RoM], SA algorithms became increasingly popular in applications due to their ease of implementation and to the availability of a comprehensive theory

[1] In this chapter, the subscript n represents the iteration number, so that the k^{th} component of a vector x_n in $\mathbb{R}^K$ is denoted by $x_{k,n}$, $1 \leq k \leq K$.

concerning their asymptotic behavior. The reader in referred to [KuC], [HaN] and the references therein.

In some cases, such as the one considered here, the iterates p_n have to lie in some constraint set E. A modified version of the basic Robbins-Monro algorithm is given by Kushner and Clark [KuC] when E is closed and bounded: The $(n+1)^{st}$ iterate is obtained by *projecting* the vector $p_n - a_n z_n$ onto the set E, i.e.,

$$p_{n+1} = \mathcal{P}_E(p_n - a_n z_n) \ . \tag{3.2.4}$$

Sufficient conditions for almost sure (a.s.) convergence of this modified procedure to $p^\star$ are also established in [KuC].

III.3. THE MEASUREMENT MODEL

In this section, a simple procedure for estimating b_k, $1 \leq k \leq K$, from idle time measurements in the k^{th} $M/M/1$ queue is described. These estimates provide the aforementioned sequence $\{z_n,\ n = 0,1,\ldots\}$. This measurement technique is used by Kumar [Kum] and Bonomi and Kumar [BoK] for similar quasi-static optimization problems.

Consider a stable $M/M/1$ queue with utilization ρ. Assume that this queue is sampled at the points of a Poisson process with rate ν, independent of the arrival and service processes in the queue. If N_τ is the number of samples in the interval $[0,\tau]$, and η_τ is the number of times the queue is sampled idle in this interval, then the *a.s.* relations

$$\lim_{\tau\to\infty} \frac{\eta_\tau}{N_\tau} = \lim_{\tau\to\infty} \frac{\eta_\tau}{\nu\tau} = 1 - \rho \tag{3.3.1}$$

hold (see Appendix IV).

With the RV y_τ defined by

$$y_\tau := (1 - \frac{\eta_\tau}{\nu\tau})^{-1} \ , \tag{3.3.2}$$

it is plain from (3.3.1) that

$$\lim_{\tau\to\infty} y_\tau = \rho^{-1} \ , \qquad a.s.$$

Therefore, y_τ can be written in the form

$$y_\tau = \rho^{-1} + w_\tau \tag{3.3.3}$$

for some RV w_τ, where

$$w_\tau \to 0 \qquad a.s. \qquad \text{as} \quad \tau \to \infty \ . \tag{3.3.4}$$

For the parallel system, assume that each server is sampled independently by a Poisson process with rate ν. After every measurement interval the central scheduler (Bernoulli switch) receives the number of times each server was found idle by the Poisson sampling process. Let $\eta_{k,n}$, $1 \leq k \leq K$ and $n = 0, 1, \ldots$, denote this number for the k^{th} server in the n^{th} measurement interval of length τ_n. The choice of the deterministic sequence $\{\tau_n,\ n = 0, 1, \ldots\}$ is crucial for the convergence of the proposed SA algorithm (see Remark III.4.3). The queue length at the end of the n^{th} measurement interval is used as the initial queue length for the $n+1^{st}$ interval. The switching probability vector p_n is held constant during the n^{th} measurement interval and is updated upon receipt of the new measurements. The arrival rate into queue k during the n^{th} interval is therefore $\lambda p_{k,n}$. Define

$$y_{k,n} := \frac{p_{k,n}}{1 - \frac{\eta_{k,n}}{\nu \tau_n}} \ , \quad 1 \leq k \leq K \ , \quad n = 0, 1, \ldots \ . \tag{3.3.5}$$

When $\nu\tau_n$ is not chosen to be an integer, the denominator is non-zero and $y_{k,n}$ is always well defined.

In view of (3.3.2) and (3.3.3), the following model for the measurement vector y_n with components $y_{k,n}$, $1 \leq k \leq K$, is proposed:

$$y_{k,n} = b_k + w_{k,n} \ , \quad 1 \leq k \leq K \ , \quad n = 0, 1, \ldots \ , \tag{3.3.6}$$

where the noise sequence $\{w_n,\ n = 0, 1, \ldots\}$ has the property

$$w_n \to 0 \qquad a.s. \qquad \text{as} \quad \tau_n \to \infty \ . \tag{3.3.7}$$

III.4. STOCHASTIC APPROXIMATION ALGORITHM

Let the function $h : \mathbb{R}^K \to \mathbb{R}^K$ be defined by

$$h(p) := p - b - \frac{1}{K}(1 - \sum_{k=1}^{K} b_k)\ e\ , \tag{3.4.1}$$

and define the sequence of vectors $\{z_n,\ n = 0, 1, \ldots\}$ by

$$z_n := p_n - y_n - \frac{1}{K}(1 - \sum_{k=1}^{K} y_{k,n})\ e\ , \qquad n = 0, 1, \ldots\ . \tag{3.4.2}$$

It is plain from (3.3.7) and (3.4.1) that

$$z_n = h(p_n) + \xi_n \qquad\qquad n = 0, 1, \ldots$$

where

$$\xi_n = -w_n + \frac{1}{K}\left(\sum_{k=1}^{K} w_{k,n}\right)\ e\ , \qquad n = 0, 1, \ldots\ . \tag{3.4.3}$$

If $h(p^\star) = 0$, then $\hat{p}^\dagger = \mathcal{P}_{\mathcal{D}}(p^\star)$, i.e., $\hat{p}^\dagger$ is the projection of the root of the continuous function h onto $\mathcal{D}$. However, since the vector b is not known and only y_n can be measured, only $z_n = h(p_n) + \xi_n$ is available. This setup naturally calls for the SA algorithm outlined in Section 2. Since the vector p_n needs to be a probability vector in order to obtain the measurements for the $(n+1)^{st}$ interval, a projection algorithm of the type given in (3.2.4) is needed.

To summarize, the following SA algorithm is proposed where the positive sequence $\{a_n,\ n = 0, 1, \ldots\}$ satisfies the conditions in (3.2.1):

Algorithm III.1.

(i) Start with a probability vector p_0, and set $n = 0$.

(ii) Obtain the idle time measurements $\eta_{k,n}$, $1 \le k \le K$, during an interval of length τ_n.

(iii) Compute

$$y_{k,n} = \frac{p_{k,n}}{1 - \frac{\eta_{k,n}}{\nu\tau_n}}\ , \quad 1 \le k \le K\ .$$

(iv) Compute

$$z_n = p_n - y_n - \frac{1}{K}(1 - \sum_{k=1}^{K} y_{k,n})\ e\ .$$

(v) Compute $p_{n+1} = \mathcal{P}_{\mathcal{U}}(p_n - a_n z_n)$.

(vi) Set $n \leftarrow n+1$, and go to **(ii)**.

Remark III.4.1. Since the set $\mathcal{D}$ is unknown, in step **(v)** the projection is done onto the probability simplex $\mathcal{U} := \{p \in [0,1]^K : \sum_{k=1}^{K} p_k = 1\}$.

Remark III.4.2. Note from (3.4.2) that

$$p_n - a_n z_n = \hat{b}_n + \frac{1}{K}(1 - \sum_{k=1}^{K} \hat{b}_{k,n})\ e$$

where

$$\hat{b}_n = (1 - a_n)p_n + a_n y_n\ .$$

Therefore, in step **(v)** of the Algorithm III.1, p_{n+1} is obtained by replacing b with $\hat{b}_n$ in Algorithm II.1.

Remark III.4.3. A convergence proof for the Algorithm III.1 is currently under investigation. Numerical observations in Section 5 indicate that the algorithm converges within a neighborhood of $\hat{p}^\dagger$ even when $\tau_n = \tau,\ n = 0, 1, \ldots$. The choice of τ is also important for the accuracy and the robustness of the algorithm. Clearly, for large values of τ the measurements and the control algorithm derived from it will be more accurate, but the control will be updated less frequently. The reader is referred to [BoK] for a discussion on choosing τ.

The sequence $\{a_n,\ n = 0, 1, \ldots\}$ is typically taken to be

$$a_n = \frac{a}{n+1}\ , \qquad n = 0, 1, \ldots \tag{3.4.4}$$

for some $a > 0$. In this case, the asymptotic normality of the normalized error term $\sqrt{n}(p_n - \hat{p}^\dagger)$ has been established in the literature under various assumptions. The

reader may find an account of the relevant references in [NeH]. Generally speaking, if $\{a_n,\ n = 0, 1, \ldots\}$ is of the form (3.4.4), then $\lim_{n\to\infty} \sqrt{n}\ (p_n - \hat{p}^\dagger)$ is normally distributed with zero mean and covariance matrix C. The matrix C in general depends on the parameter a and the gradient $\nabla h(p^\star)$, and a is chosen so that C is minimal. The interested reader is referred to [NeH, p. 169] for the general form of the C matrix. For the function h given in (3.4.1), $\nabla h(p^\star)$ is the identity matrix, and the matrix C is given, after some elementary computations, by

$$C = \frac{a^2}{(2a-1)}\, C_\xi\ ,$$

where C_ξ is the covariance matrix of the RV ξ_∞. Therefore, C is minimal when $a = 1$, and the sequence, $\{a_n,\ n = 0, 1, \ldots\}$ is thus taken as

$$a_n = \frac{1}{n+1}\ , \qquad n = 0, 1, \ldots\ ,$$

in the following examples.

III.5. NUMERICAL EXAMPLES

In this section the results of a few experiments obtained from a simulation program that implements the SA Algorithm III.1 with $\tau_n = \tau$, $n = 0, 1, \ldots$, is presented. The numerical examples are picked among the ones given in Table II.2 of Chapter II, i.e., $K = 3$ and $\lambda = 1$. The service rates μ_k, $k = 1, 2, 3$, and the corresponding $\hat{p}^\dagger$ vector (from Table II.2) are given in Table III.1 for each example.

Example	ρ	(μ_1, μ_2, μ_3)	$\hat{p}^\dagger$
III.1	0.25	(2.0,1.0,1.0)	(1.000,0.000,0.000)
III.2	0.63	(1.0,0.5,0.1)	(0.750,0.250,0.000)
III.3	0.91	(0.7,.25,.15)	(0.667,0.217,0.117)

Table III.1

Test Examples for the SA Algorithm

In all three examples $\tau = 1000 * \max\{\mu_k^{-1} : k = 1, 2, 3\}$ while ν is such that $\nu\tau = 1000$ (see [BoK]). In all cases $p_0 = (0.33, 0.33, 0.34)$. The curves marked 1, 2 and 3 in Figures III.1-III.3 present the evolution of the optimal switching probabilities $p_{1,n}$, $p_{2,n}$ and $p_{3,n}$, for the first 10, 25 and 35 iterates, respectively. Although the convergence of the proposed algorithm is not established, the iterates converged to within 5% of the limiting values in a few iterations.

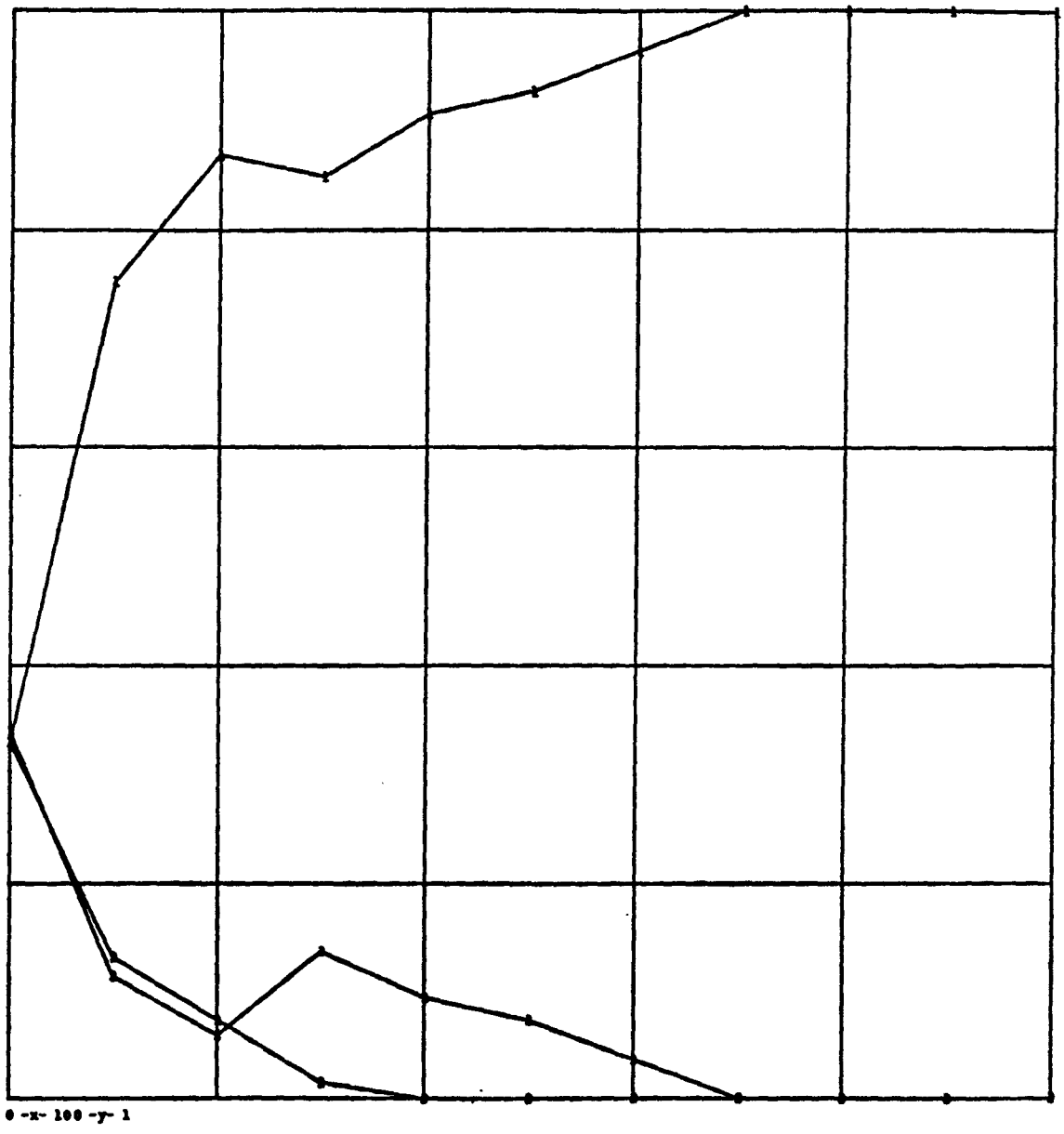


Figure III.1.

p_n, $n = 0, 1, \ldots, 10$ for Example III.1.

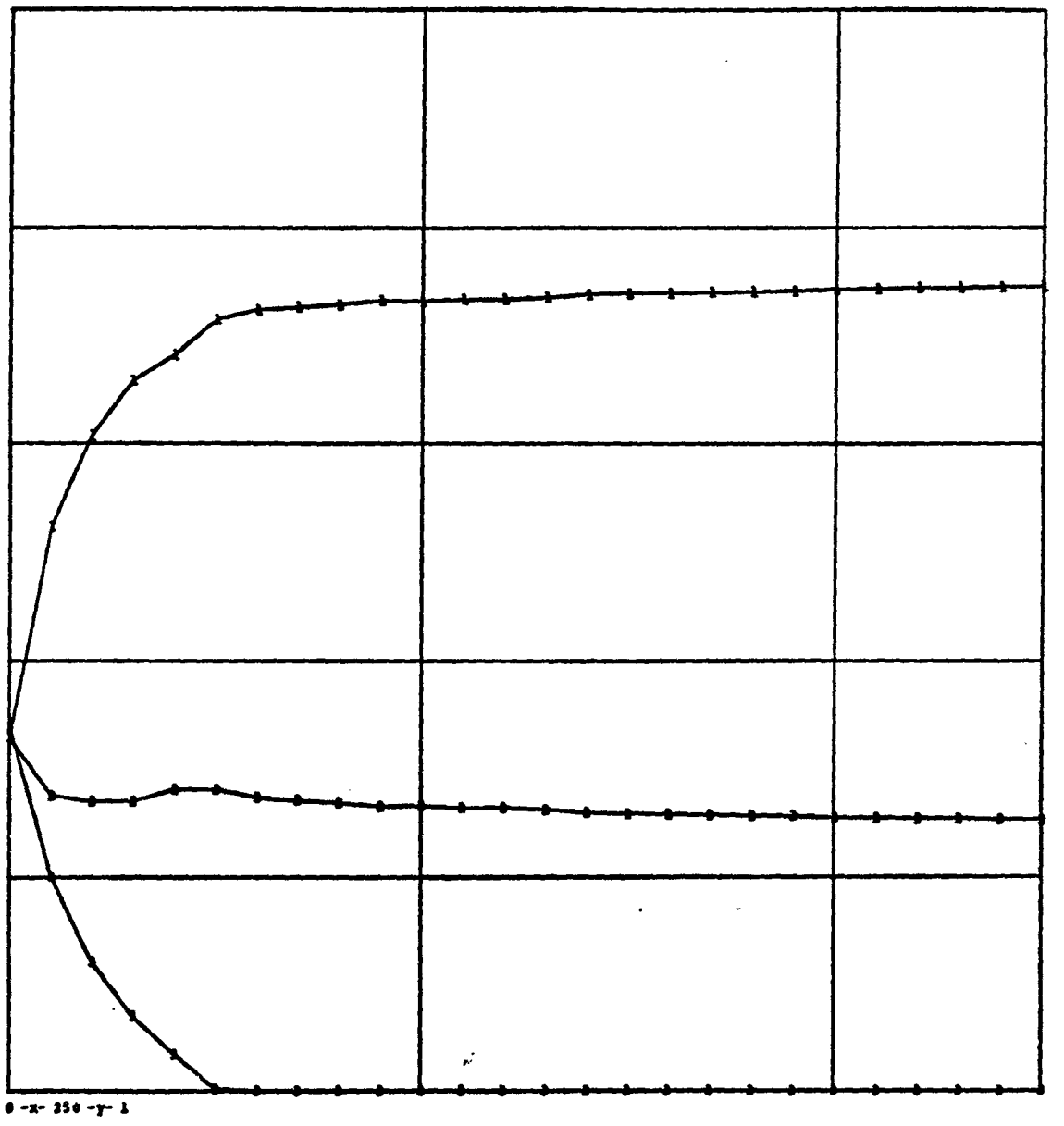


Figure III.2.

p_n, $n = 0, 1, \ldots, 25$ for Example III.2.

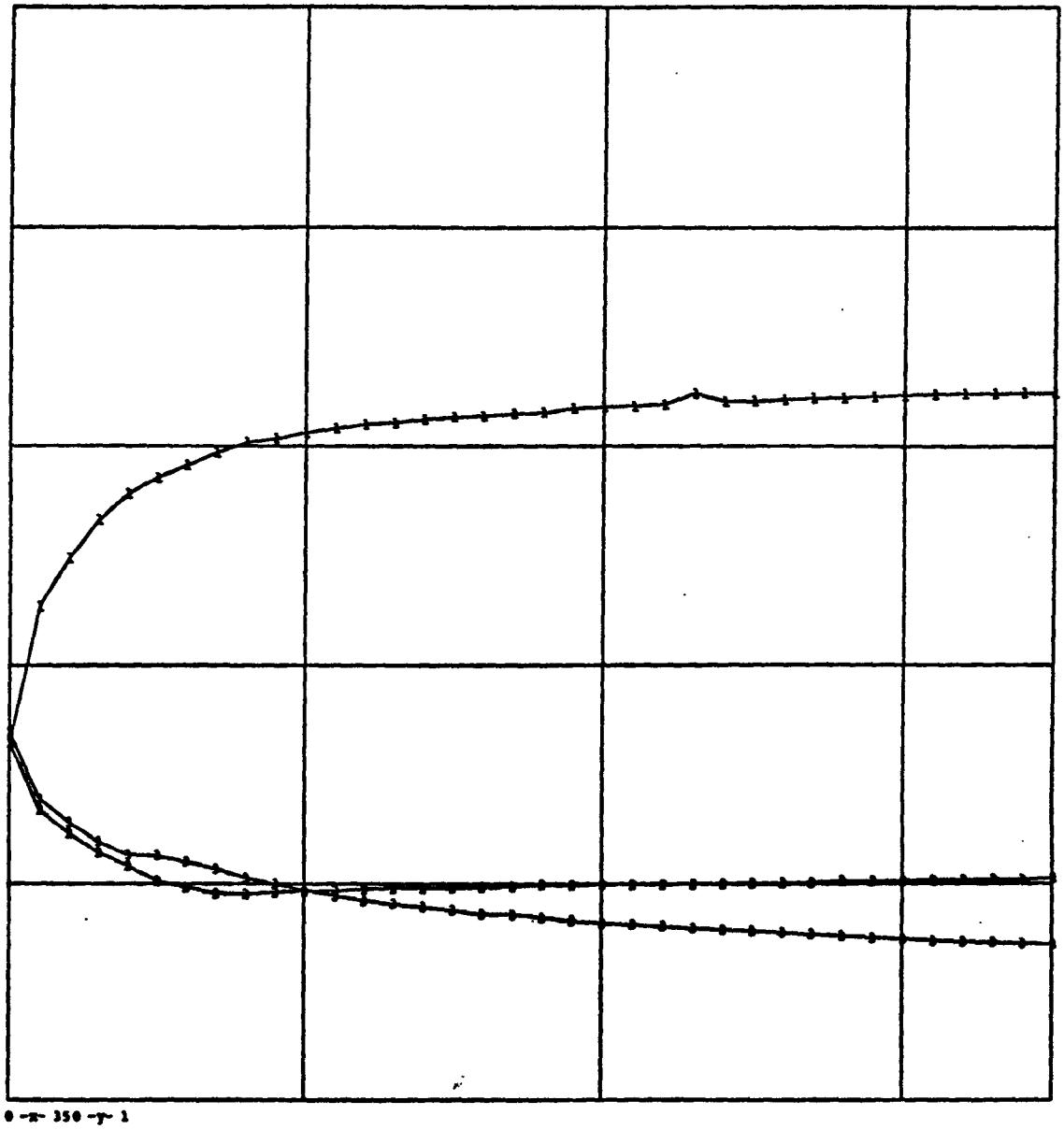


Figure III.3.

p_n, $n = 0, 1, \ldots, 35$ for Example III.3.

CHAPTER IV

ASYMPTOTIC RESULTS FOR PARALLEL QUEUES WITH RESEQUENCING

IV.1. INTRODUCTION

The model of Chapter II is again considered when the load is equally allocated to K parallel queues with *identical* service time distributions. The common service time distribution is denoted by $B(\cdot)$ with finite mean $1/\mu$ so that the system capacity is now $K\mu$. When the load is equally allocated to the queues, the distributions of the waiting and response times are identical for all queues, and are denoted by $W^{(K)}(\cdot)$ and $T^{(K)}(\cdot)$, respectively. Note in this case that the RV $T^{(K)}$ is also the disordering delay. The system time S and the resequencing delay R are also represented by $S^{(K)}$ and $R^{(K)}$, respectively, to indicate their dependence on K.

Attention is given here to the variation of the RVs $T^{(K)}$, $R^{(K)}$ and $S^{(K)}$ with K. In Section 2, it is shown that the RV $T^{(K)}$ is stochastically integer convex and decreasing in K, while $ES^{(K)}$ is decreasing in K when the arrival rate to the system remains constant. Asymptotic expressions in K are also provided for $T^{(K)}(\cdot)$, $R^{(K)}(\cdot)$ and $S^{(K)}(\cdot)$ in Section 2 to establish (i) *convergence* of these distribution functions to the corresponding distributions in the $M/GI/\infty$ system with resequencing and (ii) *asymptotic* monotonicity and integer convexity results for these RVs. It is shown that, while the behavior of $R^{(K)}$ in general depends on the load of the system, $T^{(K)}$ and $S^{(K)}$ always have similar structural characteristics. For instance, $ES^{(K)}$ is also (asymptotically) integer convex and decreasing in K.

In Section 3, the arrival rate to the system is increased linearly with K. A totally different limiting behavior is now observed in this case: $R^{(K)}$ dominates $T^{(K)}$ as both $ER^{(K)}$ and $ES^{(K)}$ grow as $\log K$, while $ET^{(K)}$ remains constant.

IV.2. ASYMPTOTIC RESULTS FOR CONSTANT LOAD

In this section, the limiting behavior of the totally homogeneous system is studied when the arrival rate λ to the system is held *constant*, and the number K of queues is increased. The RVs $W^{(K)}$ and $T^{(K)}$ correspond to the waiting and response time distributions in a $M/GI/1$ system with arrival rate λ/K and service time distribution $B(\cdot)$, respectively. The notation $\rho_0 = \lambda/\mu$ is used throughout this chapter.

The following result follows directly from the convexity result in Appendix II and extends the monotonicity result of [Sto, p. 82].

Corollary IV.2.1. *In the homogeneous system with equal load allocation, the response time $T^{(K)}$ is stochastically integer convex and decreasing in K, i.e. $\{T^{(K)},\ K \in \mathbb{N}\} \in SDCX(st)$.*

Remark IV.2.1. For the system time $S^{(K)}$, only asymptotic results are currently available (see Corollaries IV.2.5 and IV.2.6).

For this totally homogeneous system, equations (2.3.3*b*) and (2.3.3*c*) in Chapter II can be rewritten as

$$R^{(K)}(x) = \int_0^\infty [W^{(K)}(x+t)]^{K-1} dT^K(t) \qquad (4.2.1a)$$

and

$$S^{(K)}(x) = [W^{(K)}(x)]^K - \frac{1}{\lambda}\frac{d}{dx}[W^{(K)}(x)]^K \qquad (4.2.1b)$$

for all $x \geq 0$.

As K goes to infinity, the load in each queue goes to zero, and the distributions of the RVs $W^{(K)}$ and $T^{(K)}$ tend to the unit step function and to $B(\cdot)$, respectively. The expansions of these distributions as $\rho \to 0$ is the subject of the following lemma. For the purpose of the analysis to come, the expansion of the waiting time distribution is given up to the second order term, while a first order expansion suffices for the response time distribution.

Lemma IV.2.2. *Let $W_\rho(\cdot)$ and $T_\rho(\cdot)$ be the distributions of the waiting and response times in a $M/GI/1$ queue with utilization $\rho = \lambda/\mu$, where λ is the arrival rate and $1/\mu$ is the mean of the service time distribution $B(\cdot)$. Then, as $\rho \to 0$,*

$$W_\rho(x) = 1 - \rho(1 - V_1(x)) + \rho^2(V_2(x) - V_1(x)) + o(\rho^2) \quad (4.2.2a)$$

and

$$T_\rho(x) = B(x) - \rho(B(x) - V_3(x)) + o(\rho) \quad (4.2.2b)$$

for all $x \geq 0$, where

$$V_1(x) = \mu \int_0^x (1 - B(t))dt, \qquad V_2(x) = \mu \int_0^x V_1(x-t)(1 - B(t))dt$$

and

$$V_3(x) = \mu \int_0^x B(x-t)(1 - B(t))dt \ . \quad (4.2.2c)$$

Proof. Note that $V_1(\cdot)$ is the distribution function of the *residual* service time V_1, and that, $V_2(\cdot) = V_1 * V_1(\cdot)$ with $*$ denoting the convolution. Equation (4.2.2a) thus easily follows by inverting the first two terms in the expansion of the Pollaczek-Khinchin transform formula [Kle.a, p. 200]

$$W_\rho^*(s) = (1 - \rho) \sum_{k=0}^{\infty} \rho^k [V_1^*(s)]^k \ ,$$

and by rearranging the terms.

Equation (4.2.2b) is plain since $T_\rho(\cdot) = B * W_\rho(\cdot)$ and $V_3(\cdot) = V_1 * B(\cdot)$.

□

Remark IV.2.2. Note that in a $M/GI/1$ queue in statistical equilibrium a job will arrive to an empty system with probability $1 - \rho$, and that there will be exactly one job in the system with probability $\rho + o(\rho)$. This can be proved using the Pollaczek-Khinchin formula for the queue length distribution. Therefore, the

asymptotic expansion of the response time T_ρ can also be obtained by observing that

$$T_\rho = \begin{cases} B & \text{with probability } 1-\rho, \\ V_3 & \text{with probability } \rho + o(\rho)\ . \end{cases}$$

The following asymptotic expansions for the distribution functions and means of the RVs $T^{(K)}$, $R^{(K)}$ and $S^{(K)}$ constitute the main result of this section.

Theorem IV.2.3. *Consider the system of K homogeneous $\cdot/GI/1$ queues in parallel with Poisson arrivals and balanced Bernoulli loading under the resequencing constraint. As K goes to infinity, the asymptotic expansions of the response, resequencing and system time distributions are given by*

$$T^{(K)}(x) = B(x) + \frac{\rho_0}{K}(V_3(x) - B(x)) + o(\frac{1}{K}) \tag{4.2.3a}$$

$$R^{(K)}(x) = \int_0^\infty e^{-\rho_0 \overline{V}_1(x+t)} dB(t) + \frac{1}{K}\int_0^\infty e^{-\rho_0 \overline{V}_1(x+t)}(F(x+t) - \rho_0) dB(t)$$
$$+ \frac{\rho_o}{K}\int_0^\infty e^{-\rho_0 \overline{V}_1(x+t)} dV_3(t) + o(\frac{1}{K}) \tag{4.2.3b}$$

and

$$S^{(K)}(x) = B(x)e^{-\rho_0 \overline{V}_1(x)} + \frac{1}{K}e^{-\rho_0 \overline{V}_1(x)}\left[(F(x) - \rho_0)B(x) + \rho_0 V_3(x)\right] + o(\frac{1}{K}) \tag{4.2.3c}$$

where

$$F(x) = \rho_0^2(V_2(x) - V_1(x)) + \rho_0 \overline{V}_1(x) - \frac{\rho_0^2}{2}\overline{V}_1{}^2(x) \tag{4.2.3d}$$

for all $x \geq 0$. Their means are given by

$$ET^{(K)} = \frac{1}{\mu} + \frac{(1+\sigma^2\mu^2)\rho_0}{2\mu K} + o(\frac{1}{K}) \tag{4.2.4a}$$

$$ER^{(K)} = \int_0^\infty (1 - e^{-\rho_0 \overline{V}_1(x)})B(x)dx + \frac{\rho_0^2}{2K}\left[\frac{e^{-\rho_0} - (1+\sigma^2\mu^2)}{\lambda} + G\right] + o(\frac{1}{K})$$

(4.2.4*b*)

and

$$ES^{(K)} = \frac{1-e^{-\rho_0}}{\lambda} + \int_0^\infty (1 - e^{-\rho_0 \overline{V}_1(x)})dx + \frac{\rho_0^2}{2K}\left[\frac{e^{-\rho_0}}{\lambda} + G\right] + o(\frac{1}{K}) \quad (4.2.4c)$$

where

$$G = \int_0^\infty \left(\overline{V}_1^2(x) + 2(V_1(x) - V_2(x))\right) e^{-\rho_0 \overline{V}_1(x)} dx \ .$$

Proof. Since $T^{(K)}$ denotes the response time in a $M/GI/1$ system with arrival rate λ/K and service time distribution $B(\cdot)$, replacing ρ with ρ_0/K in (4.2.2*b*) gives (4.2.3*a*). Equation (4.2.4*a*) then follows by simple integration.

Equations (4.2.3*b*, *c*) are derived from (4.2.1*a*, *b*) and (4.2.3a), replacing $W^{(K)}(x)$ by its expansion (4.2.2*a*), with λ replaced by λ/K, and noting that as K goes to infinity

$$\left(1 + \frac{u}{K} + \frac{v}{K^2} + o(\frac{1}{K^2})\right)^K = e^u \left(1 + \frac{2v - u^2}{2K} + o(\frac{1}{K})\right) . \quad (4.2.5)$$

Rather than using (4.2.3*c*) to obtain $ES^{(K)}$ by integration, it is much simpler to start from the relation

$$ES^{(K)} = \int_0^\infty \left(1 - [W^{(K)}(x)]^K\right) dx + \frac{1}{\lambda}\left(1 - (1 - \frac{\rho_0}{K})^K\right)$$

derived from (2.4.4), and to replace $W^{(K)}(x)$ by its expansion (4.2.3a). Equation (4.2.4*c*) then follows from (4.2.5) by routine manipulations. Finally, $ER^{(K)}$ is again computed as $ES^{(K)} - ET^{(K)}$.

□

Remark IV.2.3. The terms indicated with the shorthand notation"$o(1/K)$" in (4.2.3) are functions of x. By tedious yet straightforward calculations it can be shown that the integrals involving these functions in Theorem IV.2.3 exist and are still of the order $o(1/K)$, so that the expansions given in the theorem are valid.

Corollary IV.2.4. *For the system of K homogeneous $\cdot/GI/1$ queues in parallel with Poisson arrivals and balanced Bernoulli loading under the resequencing constraint, the limiting distributions of the response, resequencing and system times as K goes to infinity exist and coincide with the corresponding distributions of the $M/GI/\infty$ system with resequencing, namely*

$$T^{\infty}(x) = B(x) \tag{4.2.6a}$$

$$R^{\infty}(x) = \int_0^{\infty} e^{-\rho_0 \overline{V}_1(x+t)} dB(t) \tag{4.2.6b}$$

and

$$S^{\infty}(x) = B(x) e^{-\rho_0 \overline{V}_1(x)} \tag{4.2.6c}$$

for all $x \geq 0$. Their means are given by

$$ET^{\infty} = \frac{1}{\mu} \tag{4.2.7a}$$

$$ER^{\infty} = \int_0^{\infty} (1 - e^{-\rho_0 \overline{V}_1(x)}) B(x) dx = ES^{\infty} - ET^{\infty} \tag{4.2.7b}$$

and

$$ES^{\infty} = \frac{1 - e^{-\rho_0}}{\lambda} + \int_0^{\infty} (1 - e^{-\rho_0 \overline{V}_1(x)}) dx \ . \tag{4.2.7c}$$

Proof. Equations (4.2.6) and (4.2.7) easily follow by letting $K \to \infty$ in (4.2.3) and (4.2.4), respectively. The fact that these distributions coincide with the corresponding quantities in the $M/GI/\infty$ system with resequencing can be verified from the results of [HaP].

□

Remark IV.2.4. In the $M/M/1$ case, the asymptotic expansions $(4.2.2a - b)$ are immediate since

$$T_{\rho}(x) = 1 - e^{-\mu(1-\rho)x} = 1 - e^{-\mu x} - \rho \mu x e^{-\mu x} + o(\rho)$$

and

$$W_\rho(x) = 1 - \rho e^{-\mu(1-\rho)x} \;=\; 1 - \rho e^{-\mu x} - \rho^2 \mu x e^{-\mu x} \;+\; o(\rho)$$

for all $x \geq 0$. Therefore, $(4.2.4a - c)$ simplify as

$$ET^{(K)} \;=\; ET^{\infty} \;+\; \frac{\rho_0}{K\mu} \;+\; o(\frac{1}{K})\,, \tag{4.2.8a}$$

$$ER^{(K)} \;=\; ER^{\infty} \;+\; \frac{1}{2K\mu}(1 - e^{-\rho_0} - 2\rho_0 + 2\rho_0 \int_0^{\rho_0} \frac{1-e^{-t}}{t} dt) \;+\; o(\frac{1}{K}) \tag{4.2.8b}$$

and

$$ES^{(K)} \;=\; ES^{\infty} \;+\; \frac{1}{2K\mu}(1 - e^{-\rho_0} + 2\rho_0 \int_0^{\rho_0} \frac{1-e^{-t}}{t} dt) \;+\; o(\frac{1}{K})\,. \tag{4.2.8c}$$

When the service times are deterministic, $(4.2.4a - c)$ take the form

$$ET^{(K)} \;=\; \frac{1}{\mu} \;+\; \frac{\rho_0}{2K\mu} \;+\; o(\frac{1}{K})\,, \tag{4.2.9a}$$

$$ER^{(K)} \;=\; \frac{\rho_0^2}{6K\mu} \;+\; o(\frac{1}{K}) \tag{4.2.9b}$$

and

$$ES^{(K)} \;=\; \frac{1}{\mu} \;+\; \frac{(3+\rho_0)\rho_0}{6K\mu} \;+\; o(\frac{1}{K})\,. \tag{4.2.9c}$$

By studying the sign of the coefficient functions of $1/K$ in the asymptotic expansions of Theorem IV.2.3, it is now possible to give asymptotic integer convexity and monotonicity results. Additional comments are also made in Remark IV.2.5 without the tedious details in calculations.

Corollary IV.2.5. *In the resequencing system of K homogeneous $\cdot/GI/1$ queues in parallel with Poisson arrivals and balanced Bernoulli loading, the mapping $K \mapsto ES^{(K)}$ is asymptotically integer convex and decreasing.*

Proof. The mean $ES^{(K)}$ is asymptotically integer convex and decreasing if the term in square brackets in (4.2.4c) is strictly positive. By the definitions of the

functions $V_i(\cdot)$, $i = 1, 2$, given in (4.2.2c), it is plain that

$$V_2(x) = \int_0^x V_1(x-t)dV_1(t) \;\leq\; V_1^2(x) \;\leq\; V_1(x)$$

for all $x \geq 0$, and the constant G is thus positive. The result now follows easily since $e^{-\rho_0}/\lambda + G > 0$.

□

Remark IV.2.5. The situation for $R^{(K)}$ is more complex as should be apparent from (4.2.3b). When the service times are *deterministic*, it is plain from (4.2.9b) that $ER^{(K)}$ *decrease* to 0 at least asymptotically. In the exponential case however, the study of the term in parenthesis in (4.2.8b) shows that asymptotically $ER^{(K)}$ increases (resp. decreases) to ER^{∞} for $\rho_0 < \rho_0^*$ (resp. $\rho_0 > \rho_0^*$) with $\rho_0^* \simeq 0.783652$.

Although the integer convexity and monotonicity of $ES^{(K)}$ is given only asymptotically in Corollary IV.2.5, the monotonicity of $ES^{(K)}$ in K follows from Theorem II.4.6 for *all* K.

Corollary IV.2.6. *In the totally homogeneous system, the expected system time* $ES^{(K)}$ *decreases with* K.

Proof. The system with $K-1$ queues can be obtained from the system with K queues by setting $p_k = 1/(K-1)$ for $1 \leq k < K$ and $p_K = 0$. The result therefore follows from Theorem II.4.6 since

$$ES_K \leq ES_K(\frac{1}{K-1}, \ldots, \frac{1}{K-1}, 0) = ES_{K-1} \;,$$

with an obvious meaning to the notation.

□

Asymptotic expansions in (4.2.3) also provide the following weak asymptotic stochastic convexity and monotonicity result for the RV $S^{(K)}$.

Corollary IV.2.7. *If* $S^{(K)}$ *is the end-to-end delay in the resequencing system of* K *homogeneous* $\cdot/GI/1$ *queues in parallel with Poisson arrivals and balanced*

Bernoulli loading, then for all $x \geq 0$, there exists a finite integer $K(x)$ such that the mapping $(K(x), \infty) \mapsto [0,1] : \; K \mapsto P[S^{(K)} > x]$ is integer convex and decreasing.

Note that the constant $K(x)$ depends on x since the term $o(1/K)$ in (4.2.3c) also depend on x as indicated in Remark IV.2.3.

Proof. The result follows if the term in the square brackets in (4.2.3c) is strictly negative for all $x \geq 0$, i.e., if for all $x \geq 0$, $K \mapsto S^{(K)}(x)$ is increasing and concave for $K > K(x)$. By using the definition (4.2.3d) of $F(\cdot)$, this term can be rewritten as

$$(F(x) - \rho_0)\, B(x) + \rho_0 V_3(x) = \rho_0^2 \left(V_2(x) - \frac{V_1^2(x)}{2} - \frac{1}{2} \right) B(x) + \rho_0 \left(V_3(x) - B(x) V_1(x) \right) . \qquad (4.2.10)$$

It was shown in the proof of Corollary IV.2.5 that $V_2(x) \leq V_1^2(x)$. Indeed, it is a simple exercise to show that $V_2(x) < V_1^2(x)$, unless $B(\cdot)$ is the unit step function. Therefore, the first term in (4.2.10) is strictly negative for all $x \geq 0$ since $V_1^2(x) \leq 1$. Similarly,

$$V_3(x) = \int_0^x B(x-t) dV_1(t) \; \leq \; V_1(x) B(x)$$

and the second term in (4.2.10) is also negative for all $x \geq 0$. Therefore, the coefficient function of $1/K$ in (4.2.3c) is strictly negative for all $x \geq 0$, and the result follows.

□

Remark IV.2.6. The stronger asymptotic result, namely, that there exits a finite integer K^* such that the mapping $(K^*, \infty) \mapsto [0,1] : \; K \mapsto P[S^{(K)} > x]$ is integer convex and decreasing for *all* $x \geq 0$, is currently under investigation. Note that this is equivalent to the RVs $\{S^{(K)}, \; K > K^*\}$ being $SDCX(st)$.

To conclude this section, asymptotic expansions of the probability of being a star job are given in some special cases. In the totally homogeneous system, the

formula (4.2.1a) takes the form

$$R^{(K)}(0) = P^{(K)}(*) = \int_0^\infty [W^{(K)}(x)]^{K-1} dT^{(K)}(x).$$

In the exponential case, tedious yet elementary computations show that

$$P^{(K)}(*) = \frac{1}{\rho_0}[1 - (1 - \frac{\rho_0}{K})^K] ,$$

and the asymptotic result

$$P^{(K)}(*) = \frac{1 - e^{-\rho_0}}{\rho_0} + \frac{\rho_0 e^{-\rho_0}}{2K} + o(\frac{1}{K})$$

thus follows from (4.2.5). Note that $P^\infty(*) = (1 - e^{-\rho_0})/\rho_0$ can also be derived from the results of [KKM].

In the deterministic case, the asymptotic result

$$P^{(K)}(*) = 1 - \frac{\rho_0^2}{2K} + o(\frac{1}{K})$$

can be shown to hold. Note that $P^\infty(*) = 1$ as expected.

IV.3. ASYMPTOTICS FOR INCREASING LOAD

In this section, it is assumed that the input rate λ_K into the homogeneous system with K queues is $\lambda_K = K\lambda$ for some fixed λ. Thus, the load of each queue remains fixed at $\rho_0 = \lambda/\mu$ as K varies. Therefore, the common distributions of the *i.i.d.* RVs W_k and T_k, $1 \le k \le K$, are independent of K and are now denoted by $W(\cdot)$ and $T(\cdot)$, respectively. Equation (2.3.4c) now reads as

$$ES^{(K)} = E(\max_{1 \le k \le K} W_k) + \frac{1 - (1 - \rho_0)^K}{K\lambda} . \qquad (4.3.1)$$

The asymptotic method used in [BMS] for the Fork-Join queue applies, and leads to the following theorem.

Theorem IV.3.1. *If the Laplace transform of $B(\cdot)$ is rational, then*

$$ES^{(K)} = \gamma \log K(1 + o(1)) = ER^{(K)} + ET \ , \qquad K \to \infty$$

where the constant γ depends on the distribution $B(\cdot)$.

Proof. Note that the second term in (4.3.1) decreases to zero with rate $1/K$. Under the enforced assumption, the distribution function $W(\cdot)$ also has a rational Laplace transform and the well-known asymptotic result

$$W(x) \ = \ 1 - Ce^{-qx}(1 + o(1)) \ , \qquad x \to \infty \tag{4.3.2}$$

thus holds for some positive constants C and q; see [Bor, p. 129] for a similar proof for the response time distribution. Theorem 7.4 of [BMS] therefore applies to yield

$$E(\max_{1 \le k \le K} W_k) \ = \ \frac{\log K}{q}(1 + o(1)) \ ,$$

and the theorem follows for $ES^{(K)}$ with $\gamma = 1/q$.

The second equality trivially holds. Note that since ET is constant, $ER^{(K)}$ also grow as $\log K$.

□

Remark IV.3.1. This limiting behavior of the resequencing system is similar to the one observed in a Fork-Join system [BMS]. In the corresponding Fork-Join system, a job arriving into the system is *forked* into K tasks and each task is processed in one of the K parallel queues. As soon as *all* the tasks of a job have been serviced, the job is immediately assembled (i.e., tasks are *joined*) and leaves the system. When the service times are *i.i.d.* with a rational Laplace transform and when the arrival process is a renewal process, the moments of the system time in this Fork-Join system have been shown to grow logarithmically in K [BMS].

Remark IV.3.2. When the service times are of PH-type with representation (α, A), the waiting time distribution is again PH-type with representation $(\rho_0 \pi, L)$,

where $\pi = -\mu\alpha A^{-1}$ and $L = A + \rho_0 A^0 \pi$ [Neu.b]. In that case, $-q$ in (4.3.2) is the eigenvalue with largest real part of the matrix L [Neu.b, p. 63].

In the exponential case, $L = \lambda_0 - \mu$, so that $\gamma = 1/(\mu - \lambda_0)$, and

$$ES^{(K)} = \frac{\log K}{\mu - \lambda_0}(1 + o(1)) \ . \tag{4.3.3}$$

Remark IV.3.3. When the load in each queue is held fixed, the limiting behavior of the system with parallel buffers as K goes to infinity differs from that of the $M/GI/\infty$ queue with resequencing, i.e., the limiting behavior of the case when there is a common buffer. To illustrate this, let ES_K^∞ be the average system time when the DS is the $M/M/\infty$ queue with arrival rate λK and service rate μ. Replacing ρ_0 by $K\rho_0$ in (4.2.7c) then yields

$$\begin{aligned} ES_K^\infty &= \int_0^\infty \left(1 - e^{-K\rho_0 e^{-\mu x}}\right) dx \ + \ o(1) \\ &= \frac{1}{\mu} \int_0^{K\rho_0} \frac{1 - e^{-u}}{u} du \ + \ o(1) \\ &= \frac{1}{\mu} \left(\int_0^1 \frac{1 - e^{-u}}{u} du \ + \ \log(K\rho_0) \ - \ \int_1^{K\rho_0} \frac{e^{-u}}{u} du \right) \ + \ o(1) \\ &= \frac{\log K}{\mu}(1 + o(1)) \end{aligned} \tag{4.3.4}$$

where the last equality follows since $\int_1^\infty e^{-u}/u \ du < \infty$ so that

$$\lim_{K \to \infty} \frac{\int_1^{K\rho_0} \frac{e^{-u}}{u} du}{\log \rho_0 K} = 0 \ .$$

Comparison of (4.3.3) and (4.3.4) shows that although both $ES^{(K)}$ and ES_K^∞ grow logarithmically in K, the average end-to-end delay has a smaller growth when there is a common buffer.

Remark IV.3.4. At increasing load, in the exponential case, the probability of being a star job is given by

$$P^{(K)}(*) = \frac{1}{K\rho_0}[1 - (1 - \rho_0)^K] = \frac{1}{K\rho_0} + o(\frac{1}{K}) \ .$$

In conclusion, when the arrival rate into the system is also increased so as to keep a fix load to each queue, the delay due to resequencing grows in $\log K$ while the disordering delay does not change with K. This result indicates that the resequencing delay will have a major impact on the end-to-end delay for highly parallel systems with reasonable load in each queue. This is illustrated in Table IV.1 when the service time distributions are exponential. Table IV.1 presents the ratio $ES^{(K)}/ET$ for various values of K and ρ_0. As apparent from this table, the resequencing delay is much larger than the disordering delay, especially for large values of K and ρ_0.

$\rho_0 \setminus K$	10	30	50	70	ET
0.1	1.40	1.99	2.38	2.67	1.11
0.2	1.96	2.87	3.34	3.66	1.43
0.5	2.34	3.34	3.83	4.15	2.00
0.7	2.62	3.65	4.15	4.48	3.33
0.9	2.83	3.89	4.40	5.24	10.0

Table IV.1.

The ratio ES_K/ET for exponential servers

CHAPTER V

DYNAMIC LOAD ALLOCATION IN PARALLEL QUEUES WITH SYNCHRONIZATION

V.1. INTRODUCTION

In packet switching networks, long messages are broken into several shorter packets which are simultaneously transmitted from source to destination over virtual circuits (parallel channels). Since the transmission time of a message over a channel is proportional to its length, this "pipelining" effect can considerably reduce the transmission time of a message over that of transmitting the message as a single packet. At the destination node the messages are first reassembled and then resequenced.

This chapter considers such a communication system when the parallel channels are *identical.* The DS is again modeled as a system of K parallel single server queues (channels). The assembly and resequencing operations of jobs (messages) are both performed in the RB. The problem of dynamically allocating the workload (message length) of each job to the parallel queues is considered when the interarrival times and workloads of jobs form two mutually independent sequences of *i.i.d.* RVs. Specifically, for each job arriving to the system, the problem of optimally partitioning its workload into $L \leq K$ tasks (packets) for transmission over the parallel queues is studied when the cost-per-stage is taken to be the system time of a job.

The problem is formally defined in Section 2. The dynamic programming methodology is used in Sections 3 and 4 to obtain the optimal allocation policy which minimizes the average discounted cost. It is shown that the optimal policy is Markov stationary and steers the workload in each queue to a balanced position as fast as feasible. In Section 5, the optimal policy for the discounted cost is shown to also minimize the corresponding average finite horizon and the long-run average costs. Section 6 considers the case $K = 2$ and illustrates how the solution

methodology of Sections 3-5 can be used to obtain the optimal scheduling policy when the messages are transmitted as a single packet. Optimality of joining the queue with the smaller workload is established for the discounted, finite horizon and the long-run average costs.

Denote the largest component of any vector x in $\mathbb{R}^K$ by $\|x\|$, i.e., $\|x\| := \max\{x_k;\ 1 \leq k \leq K\}$. A function $f : \mathbb{R}^K \to \mathbb{R}$ is said to be monotone increasing if $x \leq y$ implies $f(x) \leq f(y)$, where the inequality $x \leq y$ is understood componentwise as $x_k \leq y_k,\ 1 \leq k \leq K$.

V.2. THE MODEL AND THE PROBLEM FORMULATION

The model considered here is defined on some probability space $(\Omega, \mathcal{F}, P)$ that carries all the RVs of interest. Jobs arrive to a system of K parallel *identical* queues each with an infinite capacity buffer. The time between the n^{th} and $(n-1)^{st}$ arrivals is denoted by $\tau(n),\ n = 1, 2, \ldots$. The workload of the n^{th} job to arrive to the system is represented by the $\mathbb{R}_+$-valued RV $\sigma(n),\ n = 0, 1, \ldots$, with $\sigma(0) = \Sigma$. It is assumed that the 0^{th} job arrives to the system at time $t = 0$ and finds the initial workload in the system to be the $\mathbb{R}_+^K$-valued RV W, i.e., W_k is the workload of queue k at $t = 0$, not including the workload of the 0^{th} job.

Upon arrival, a job is partitioned into smaller tasks which are allocated to the parallel queues. Let the vector $u(n)$ in $\mathcal{U} := \{u \in [0,1]^K :\ u^T e = 1\}$ represent the allocation of the n^{th} job to the queues, i.e., $\sigma(n)u_k(n)$ is the workload of the k^{th} task of the n^{th} job, $1 \leq k \leq K$. It is assumed that the service rate (channel capacity) in each queue is fixed and thus equal to 1 without loss of generality. Therefore, $\sigma(n)u_k(n)$ is also the service time of the k^{th} task of the n^{th} job. After service completion, the tasks move to the RB to await service completion of the other tasks that belong to the same job. After all the tasks of a job complete their service, the job is reassembled. A reassembled job further awaits in the RB for all the jobs that have arrived to the system earlier to be reassembled.

The following assumptions are made:

(A1) The RVs Σ and W, and the sequences of RVs $\{\tau(n),\ n = 1, 2, \ldots\}$ and

$\{\sigma(n),\ n = 0,1,\ldots\}$ are all *mutually independent.*

(A2) The RVs $\{\tau(n),\ n = 1,2,\ldots\}$ and $\{\sigma(n),\ n = 0,1,\ldots\}$ each form a sequence of *i.i.d.* RVs with common distribution functions $A(\cdot)$ and $B(\cdot)$, respectively, with the properties $EA < \infty$ and $EB < \infty$.

An admissible control policy π is any collection $\pi = \{\pi_n,\ n = 0,1,\ldots\}$ of mappings

$$\pi_n : \mathbb{R}_+ \times \mathbb{R}_+^K \times (\mathcal{U} \times \mathbb{R}_+^K \times \mathbb{R}_+)^n \to \mathcal{U} \ . \qquad n = 0,1,\ldots$$

The collection of all such admissible control policies is denoted in the sequel by Π. For every π in Π, the $\mathbb{R}_+^K$-valued RVs $\{W^\pi(n),\ n = 0,1,\ldots\}$ and the $\mathcal{U}$-valued RVs $\{U^\pi(n),\ n = 0,1,\ldots\}$ are recursively defined by

$$W^\pi(n+1) = [W^\pi(n) + \sigma(n)U^\pi(n) - \tau(n+1)e\ ]^+ \qquad (5.2.1a)$$

and

$$U^\pi(n+1) = \pi_{n+1}\,(\Sigma, W; U^\pi(r), W^\pi(r+1), \sigma(r+1),\ 0 \le r \le n)\ , \qquad (5.2.1b)$$

with initial conditions

$$W^\pi(0) = W \qquad \text{and} \qquad U^\pi(0) = \pi_0(W, \Sigma)\ . \qquad (5.2.1c)$$

For $1 \le k \le K$ and $n = 0,1,\ldots$, the RV $W_k^\pi(n)$ represents the workload of the k^{th} queue at the n^{th} arrival epoch, while the RV $U_k^\pi(n)$ represents the fraction of the n^{th} job allocated to the k^{th} queue, when the policy π is enforced.

If the control policy π is used, then the sojourn time of the k^{th} task of the n^{th} job is $W_k^\pi(n) + U_k^\pi(n)\sigma(n)$, $1 \le k \le K$, and the system time $S^\pi(n)$ of the n^{th} job is thus given by

$$S^\pi(n) = \|W^\pi(n) + \sigma(n)U^\pi(n)\|\ . \qquad (5.2.2)$$

Note that the maximum is taken over *all* queues, even if some of the components of $U^\pi(n)$ are zero, so as to ensure that the jobs leave the system in sequence. Therefore the synchronization delay in $S^\pi(n)$ is due to both the reassembly and resequencing operations.

For every β, $0 < \beta < 1$, the β-discounted cost associated with a policy π in Π is defined by

$$J_\beta(\pi) := E\left[\sum_{n=0}^{\infty} \beta^n S^\pi(n)\right] , \tag{5.2.3}$$

and the minimization problem (P_β) of interest is then

$$(P_\beta): \qquad \textit{Minimize } J_\beta(\pi) \textit{ over } \Pi .$$

Let A and B be two *independent* $\mathbb{R}_+$-valued RVs defined on Ω with probability distribution functions $A(\cdot)$ and $B(\cdot)$, respectively. By simple conditioning, the relation

$$\begin{aligned} &P\{W^\pi(n+1) \le w\ ,\ \sigma(n+1) \le \sigma \mid \mathcal{F}_n^\pi\} \\ &\quad = P\{B \le \sigma\} P\left\{[W^\pi(n) + \sigma(n)U^\pi(n) - Ae]^+ \le w\right\} \end{aligned} \tag{5.2.4}$$

where

$$\mathcal{F}_n^\pi = \sigma\{W, \Sigma\} \vee \sigma\{\sigma(i), W^\pi(i), U^\pi(i-1), \quad i = 1, \ldots n\}$$

is seen to hold under the assumptions (A1) and (A2), for every admissible policy π in Π. Therefore, the joint distribution of the pair $(W^\pi(n+1), \sigma(n+1))$ given $\mathcal{F}_n^\pi$ depends only on $(W^\pi(n), \sigma(n))$ and $U^\pi(n)$. This suggests that the problem (P_β) can be viewed as a Markov decision problem with *augmented* state process $(W^\pi(n), \sigma(n))$. To that effect, for every value β, $0 < \beta < 1$, the *discounted cost-to-go* J_β^π associated with an arbitrary policy π in Π is defined by

$$J_\beta^\pi(w, \sigma) = E\left[\sum_{n=0}^{\infty} \beta^n \|W^\pi(n) + \sigma(n)U^\pi(n)\| \ \Big|\ W = w,\ \Sigma = \sigma\right] \tag{5.2.5}$$

for all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$. The corresponding *value function* V_β is then given by

$$V_\beta(w, \sigma) = \inf_{\pi \in \Pi} J_\beta^\pi(w, \sigma) .$$

V.3. STRUCTURE OF THE VALUE FUNCTION

For any mapping $f : \mathbb{R}_+^K \times \mathbb{R}_+ \mapsto \mathbb{R}_+$, introduce $T_\beta^u f$, u in $\mathcal{U}$, and $T_\beta f$ as the mappings $\mathbb{R}_+^K \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ defined respectively by

$$T_\beta^u f(w,\sigma) := \|w + \sigma u\| + \beta E\left[f\left([w + \sigma u - Ae]^+, B\right)\right] \tag{5.3.1}$$

and

$$T_\beta f(w,\sigma) := \min_{u \in \mathcal{U}} T_\beta^u f(w,\sigma) \tag{5.3.2}$$

for all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$. The expectation in (5.3.1) is taken over the joint distribution of A and B.

Theorem V.3.1. *If the mapping $(w,\sigma) \mapsto f(w,\sigma)$ is monotone increasing, and the mapping $w \mapsto f(w,\sigma)$ is convex for every σ in $\mathbb{R}_+$, then so is $T_\beta f$.*

Proof.

Monotonicity: Let w^1 and w^2 be two vectors in $\mathbb{R}_+^K$, and σ^1 and σ^2 be two scalars in $\mathbb{R}_+$. If $w^1 \leq w^2$ and $\sigma^1 \leq \sigma^2$, then $w^1 + \sigma^1 u \leq w^2 + \sigma^2 u$ for every u in $\mathcal{U}$, and the inequalities

$$\|w^1 + \sigma^1 u\| \leq \|w^2 + \sigma^2 u\| \qquad \text{and} \qquad [w^1 + \sigma^1 u - \tau e]^+ \leq [w^2 + \sigma^2 u - \tau e]^+ \tag{5.3.3}$$

readily follow for every τ in $\mathbb{R}_+$. By the monotonicity of f, the second inequality in (5.3.3) implies

$$f([w^1 + \sigma^1 u - \tau e]^+, \sigma') \leq f([w^2 + \sigma^2 u - \tau e]^+, \sigma') \tag{5.3.4}$$

for every (τ, σ') in $\mathbb{R}_+ \times \mathbb{R}_+$, and the monotonicity of $T_\beta^u f(w,\sigma)$ now follows from the first inequality in (5.3.3), i.e.,

$$T_\beta^u(w^1,\sigma^1) \leq T_\beta^u(w^2,\sigma^2) \tag{5.3.5}$$

holds for every u in $\mathcal{U}$. The monotonicity of $T_\beta f$ is now immediate from (5.3.5).

Convexity: Let α be a scalar in $[0,1]$. First the mapping $(w,u) \mapsto T_\beta^u f(w,\sigma)$ will be shown to be convex for every σ in $\mathbb{R}_+$. Fix σ in $\mathbb{R}_+$, and let u^1 and u^2 (resp. w^1 and w^2) be two vectors in $\mathcal{U}$ (resp. $\mathbb{R}_+^K$). The relations

$$\begin{aligned}\|\sigma(\alpha u^1 + \overline{\alpha} u^2) + \alpha w^1 + \overline{\alpha} w^2\| &= \|\alpha(\sigma u^1 + w^1) + \overline{\alpha}(\sigma u^2 + w^2)\| \\ &\leq \alpha\|\sigma u^1 + w^1\| + \overline{\alpha}\|\sigma u^2 + w^2\| \qquad (5.3.6)\end{aligned}$$

and

$$\begin{aligned}[\alpha w^1 + \overline{\alpha} w^2 + \sigma(\alpha u^1 + \overline{\alpha} u^2) - \tau e]^+ &\stackrel{\cdot}{=} [\alpha(w^1 + \sigma u^1 - \tau e) + \overline{\alpha}(w^2 + \sigma u^2 - \tau e)]^+ \\ &\leq \alpha[w^1 + \sigma u^1 - \tau e]^+ + \overline{\alpha}[w^2 + \sigma u^2 - \tau e]^+ \qquad (5.3.7)\end{aligned}$$

hold for every τ in $\mathbb{R}_+$, owing to the convexity of the mappings $x \mapsto \|x\|$ and $x \mapsto [x]^+$. Therefore,

$$\begin{aligned}&f([\alpha w^1 + \overline{\alpha} w^2 + \sigma(\alpha u^1 + \overline{\alpha} u^2) - \tau e]^+, \sigma') \\ &\leq f(\alpha[w^1 + \sigma u^1 - \tau e]^+ + \overline{\alpha}[w^2 + \sigma u^2 - \tau e]^+, \sigma') \\ &< \alpha f([w^1 + \sigma u^1 - \tau e]^+, \sigma') + \overline{\alpha} f([w^2 + \sigma u^2 - \tau e]^+, \sigma') \qquad (5.3.8)\end{aligned}$$

hold for every (τ, σ') in $\mathbb{R}_+ \times \mathbb{R}_+$. The first inequality in (5.3.8) follows from (5.3.7) and the monotonicity of f, while the second inequality expresses the assumed convexity of f.

The convexity of $T_\beta^u f$ in the variable (w,u) thus follows from (5.3.6) and (5.3.8), i.e., for every σ in $\mathbb{R}_+$,

$$T_\beta^{\alpha u^1 + \overline{\alpha} u^2} f(\alpha w^1 + \overline{\alpha} w^2, \sigma) < \alpha T_\beta^{u^1} f(w^1, \sigma) + \overline{\alpha} T_\beta^{u^2} f(w^2, \sigma)\ , \qquad (5.3.9)$$

and the convexity of $T_\beta f$ in w now follows from Lemma A.III.2 of Appendix III.

□

The key optimality result for problem (P_β) is now discussed.

Theorem V.3.2. *The value function V_β satisfies the Dynamic Programming equation*

$$V_\beta = T_\beta\ V_\beta \qquad (5.3.10)$$

and is monotone increasing, while the mapping $w \mapsto V_\beta(w,\sigma)$ *is convex for every* σ *in* $\mathbb{R}_+$.

Proof. Let $\{V_n,\ n = -1,0,1,\ldots\}$ be a sequence of mappings $\mathbb{R}_+^K \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ defined recursively by

$$V_{n+1} = T_\beta\ V_n\ , \qquad\qquad n = -1,0,1,\ldots$$

where V_{-1} is defined as the zero mapping on $\mathbb{R}_+^K \times \mathbb{R}_+$. Owing to the non-negativity of the cost function for every π in Π, the sequence $\{V_n,\ n = -1,0,1,\ldots\}$ is monotone increasing, and the convergence

$$V_\infty(w,\sigma) := \lim_{n\to\infty}\ V_n(w,\sigma) \tag{5.3.11}$$

thus takes place for all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$. Note from the continuity of the mapping $u \mapsto T_\beta^u V_n(w,\sigma)$ that the level sets $\{u \in \mathcal{U} :\ T_\beta^u\ V_n(w,\sigma) \le \lambda\}$ are compact for every w in $\mathbb{R}_+^K$, σ in $\mathbb{R}_+$ and λ in $\mathbb{R}$. Therefore, it follows from the recursive definition of $\{V_n,\ n = -1,0,1,\ldots\}$ and the Monotone Convergence Theorem that $V_\infty = T_\beta\ V_\infty$. Proposition 13 of [Ber, pp. 264-266] thus implies $V_\beta = V_\infty$ and (5.3.10) is obtained. The second part of the proposition is now an immediate consequence of the fact that monotonicity and convexity are preserved under the limiting operation in (5.3.11).

□

The following properties of the value function is an immediate consequence of Theorem V.3.2.

Corollary V.3.3. *The value function* V_β *exhibits the properties*

$$\min_{(w,\sigma)\in\mathbf{R}_+^K\times\mathbf{R}_+} V_\beta(w,\sigma) = V_\beta(0,0)\ , \tag{5.3.12a}$$

$$\min_{w\in\mathbf{R}_+^K} V_\beta(w,\sigma) = V_\beta(0,\sigma)\ ,\quad \textit{for every } \sigma \textit{ in } \mathbb{R}_+, \tag{5.3.12b}$$

and

$$\min_{\sigma\in\mathbf{R}_+} V_\beta(w,\sigma) = V_\beta(w,0)\ ,\quad \textit{for every } w \textit{ in } \mathbb{R}_+^K\ . \tag{5.3.12c}$$

With a slight abuse of notation, let π^* be any *Markov stationary* policy in Π, induced by the mapping $\pi^* : \mathbb{R}_+^K \times \mathbb{R}_+ \mapsto \mathcal{U}$, which satisfies the relation

$$V_\beta(w,\sigma) = \left(T_\beta^{\pi^*(w,\sigma)}\, V_\beta\right)(w,\sigma) \tag{5.3.13}$$

for all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$. Owing to the compactness of the control set $\mathcal{U}$, there always exists such a policy π^*, and as well known [Ber, Prop.13, p. 264], it is *optimal* for problem (P_β).

The following property of the value function will be crucial in identifying the form of the optimal control.

Lemma V.3.4. *For each k, $1 \le k \le K$, define the mappings $R_k : \mathbb{R}^K \mapsto \mathbb{R}^K$ by*

$$R_k(w) = \begin{cases} (w_{k+1}, \ldots, w_K, w_1, \ldots, w_k) & 1 \le k < K \\ w & k=K\ . \end{cases} \tag{5.3.14}$$

With this notation, the value function V_β has the property

$$V_\beta(w,\sigma) = V_\beta(R_k(w),\sigma)\ , \qquad 1 \le k \le K\ , \tag{5.3.15}$$

for every (w,σ) in $\mathbb{R}_+^K \times \mathbb{R}_+$.

In fact the property (5.3.15) holds for *any* permutation of w. However, the cyclic rotations $R_k(w)$, $1 \le k \le K$, suffice for our purposes.

Proof. For any policy $\pi = \{\pi_n,\ n = 0,1,\ldots\}$, define the policies $R_k\pi$, $1 \le k \le K$, by $R_k\pi = \{R_k(\pi_n),\ n = 0,1,\ldots\}$. Let Π_s be the set of Markov stationary policies in Π. Since the queues are homogeneous and $\|w\| = \|R_k(w)\|$, $1 \le k \le K$, it is plain that the relations

$$J_\beta^\pi(w,\sigma) = J_\beta^{R_k\pi}(R_k(w),\sigma)\ , \quad 1 \le k \le K, \tag{5.3.16}$$

hold for every π in Π_s, w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$. The property (5.3.15) thus follows by noting that $R_k\Pi_s := \{R_k\pi,\ \pi \in \Pi_s\} = \Pi_s$.

□

V.4. THE FORM OF THE OPTIMAL CONTROL

In order to provide an explicit form for the optimal policy π^* given by

$$\pi^*(w,\sigma) = \arg\min_{u\in\mathcal{U}} T_\beta^u \; V_\beta(w,\sigma) \; ,$$

for all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$, consider first the vectors $u^*(w,\sigma)$ which satisfy

$$u^*(w,\sigma) = \arg\min_{u\in\mathcal{U}'} T_\beta^u \; V_\beta(w,\sigma) \tag{5.4.1}$$

where $\mathcal{U}' := \{u \in \mathbb{R}^K : \; u^T e = 1\}$. The following Lemma can be proved by arguments very similar to the ones given in Theorems V.3.1 and V.3.2.

Theorem V.4.1. *For all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$, the $\mathbb{R}_+$-valued function $u \mapsto T^u V_\beta(w,\sigma)$ is convex on the convex set $\mathcal{U}'$.*

The optimal policy $\pi^*(w,\sigma)$ can be obtained from $u^*(w,\sigma)$ by projecting it onto $\mathcal{U}$, owing to the convexity result of Theorem V.4.1. The following Lemma will prove useful in establishing Theorem V.4.3.

Lemma V.4.2. *If the mapping $\psi : \mathbb{R}^K \mapsto \mathbb{R}$ is convex and has the property that $\psi(w) = \psi(R_k(w))$ for all $1 \le k \le K$, then*

$$\min_{\{w\in R^K:\; w^T e=c\}} \psi(w) = \psi(\frac{c}{K}\; e)$$

for every c in $\mathbb{R}$.

Proof. For every vector α in $\mathcal{U}$ and all w in $\mathbb{R}^K$, the inequality

$$\psi\left(\sum_{k=1}^K \alpha_k R_k(w)\right) \le \sum_{k=1}^K \alpha_k \psi(R_k(w)) = \psi(w) \tag{5.4.2}$$

follows from the assumptions enforced on ψ. The result is now obtained by choosing $\alpha = \frac{1}{K}\; e$ in (5.4.2) since $\frac{1}{K}\sum_{k=1}^K R_k(w) = \frac{1}{K}(w^T e)\; e$ for every w in $\mathbb{R}^K$.

□

Theorem V.4.3. *For every σ in $\mathbb{R}_+$ and w in $\mathbb{R}_+^K$, $u^*(w,\sigma)$ in (5.4.1) is given by*

$$u^*(w,\sigma) = \frac{1}{\sigma}\left[\left(\frac{\sigma + w^T e}{K}\right) e - w\right] . \tag{5.4.3}$$

In words, $u^*(w,\sigma)$ steers the workload vector to a balanced configuration, i.e., $u^*(w,\sigma)$ is such that $\sigma u^* + w$ has equal components. Note that this is always possible since $u^*(w,\sigma)$ is in $\mathbb{R}^K$.

Proof. Define the mapping $\psi : \mathbb{R}_+^K \mapsto \mathbb{R}_+$ by

$$\psi(x) := \|x\| + \beta E\left[V_\beta\left([x - A\ e]^+, B\right)\right] . \tag{5.4.4}$$

By Lemma V.3.4, the equalities

$$\begin{aligned}\psi(R_k(x)) &= \|R_k(x)\| + \beta E\left[V_\beta\left([R_k(x) - A\ e]^+, B\right)\right] \\ &= \|x\| + \beta E\left[V_\beta\left(R_k([x - A\ e]^+), B\right)\right] \\ &= \psi(x)\end{aligned}$$

hold for every $1 \le k \le K$. Furthermore, $x \mapsto \psi(x)$ is convex owing to the convexity of the functions $x \mapsto \|x\|$, $[x]^+$ and $V_\beta(x,\cdot)$. Therefore, ψ satisfies the assumptions of Lemma V.4.2.

For every w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$, set $x(u) = w + \sigma u$. Then, equation (5.4.1) can be rewritten as

$$u^*(w,\sigma) = \arg\min_{u \in \mathcal{U}'} \psi\left(x(u)\right) .$$

Since, for every u in $\mathcal{U}'$, $x(u)^T e = \sigma + w^T e$, i.e., $x(u)^T e$ does not depend on u, it follows from Lemma V.4.2 that $u^*(w,\sigma)$ is given by

$$w + \sigma u^*(w,\sigma) = \frac{1}{K}(\sigma + w^T e)\ e\ ,$$

and (5.4.3) thus follows.

□

The following result provides a necessary and sufficient condition for $u^*(w,\sigma)$ given in (5.4.3) to be in $\mathcal{U}$.

Lemma V.4.4. *If*

$$\Delta := \sum_{k=1}^{K}(\|w\| - w_k) \ ,$$

then $u^*(w,\sigma)$ given by (5.4.3) is in the set $\mathcal{U}$ if and only if $\Delta \leq \sigma$.

Proof. Note from (5.4.3) that $u^*(w,\sigma)$ lies in the set $\mathcal{U}$ if and only if

$$0 \leq \frac{(\sigma + w^T\ e)}{K} - \|w\| \tag{5.4.5}$$

since $u^*(w,\sigma)^T e = 1$. It is plain from the definition of Δ that

$$\frac{1}{K}(\sigma + w^T\ e) - \|w\| = \frac{1}{K}(\sigma - \sum_{k=1}^{K}(\|w\| - w_k)) = \frac{\sigma - \Delta}{K}$$

and the result holds.

□

In general, when $u^*(w,\sigma)$ is not in $\mathcal{U}$, the convexity result in Lemma V.4.1. yields the following result.

Corollary V.4.5. *For every σ in $\mathbb{R}_+$ and w in $\mathbb{R}_+^K$, the optimal policy for problem (P_β) is given by*

$$\pi^*(w,\sigma) = \mathcal{P}_{\mathcal{U}}(u^*(w,\sigma)) \tag{5.4.6}$$

where $u^*(w,\sigma)$ is given in (5.4.3).

In words, the optimum allocation strategy π^* steers the system into a balanced configuration as quickly as feasible.

The following corollary is now immediate and states that, if the workloads of all the queues are the same, then the optimum allocation strategy π^* keeps the system in this balanced configuration.

Corollary V.4.6. *For every σ in $\mathbb{R}_+$ and w in $\mathbb{R}_+^K$ such that $w = c\ e$ for some c in $\mathbb{R}_+$, the optimal policy is given by $\pi^*(w,\sigma) = \frac{1}{K}\ e$.*

Finally, for a given w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$, an algorithm is provided for computing the vector $\pi^*(w,\sigma)$. As mentioned before, if $u^*(w,\sigma)$ is not in $\mathcal{U}$, then owing to the convexity of the mapping $\mathcal{U}' \mapsto \mathbb{R}_+ : u \mapsto T_\beta^u V_\beta(w,\sigma)$, $\pi^*(w,\sigma)$ is on the boundary of $\mathcal{U}$, i.e., if $u_l^*(w,\sigma) \leq 0$ for some l, then $\pi_l^*(w,\sigma) = 0$. After setting a component of π^* to zero, the problem reduces to allocating the workload to a reduced set of queues so that the remaining components of π^* should be recalculated from (5.4.3) and (5.4.6). The following result states that if several components of $u^*(w,\sigma)$ are negative, then the corresponding components in $\pi^*(w,\sigma)$ can *all* be set to zero at once, thus facilitating the computations.

Lemma V.4.7. *For every non-empty subset E of $\{1,\ldots,K\}$ with cardinality $|E|$ define the vector $u(E)$ in $\mathbb{R}^{|E|}$ by*

$$u_k(E) := \frac{1}{\sigma}(c(E) - w_k)\ , \quad k \in E \qquad \text{with} \quad c(E) = \frac{\sigma + \sum_{i\in E} w_i}{|E|}\ . \tag{5.4.7}$$

Let l and k be different elements of E. If $u_l(E) < 0$, then $u_k(E) < 0$ implies that $u_k(E \setminus \{l\}) < 0$.

Proof. The result follows from the following routine calculations

$$\begin{aligned} c(E \setminus \{l\}) = \frac{\sigma + \sum_{i\in E} w_i}{|E|-1} - \frac{w_l}{|E|-1} &< \frac{\sigma + \sum_{i\in E} w_i}{|E|-1} - \frac{\sigma + \sum_{i\in E} w_i}{|E|(|E|-1)} \\ &= \frac{\sigma + \sum_{i\in E} w_i}{|E|} < w_k\ , \end{aligned}$$

where the first and second inequalities follow from $c(E) < w_l$ and $c(E) < w_k$, respectively.

□

Lemma V.4.7 leads to the following algorithm

Algorithm V.1.

(i) Set $E \leftarrow \{1,2,\ldots,K\}$.

(ii) Compute the vector $u(E)$ from (5.4.7).

(iii.a) If $u_k(E) \geq 0$ for all k in E, then STOP. The optimum allocation vector is given by

$$\pi_k^*(w,\sigma) = \begin{cases} u_k(E) & k \text{ in } E \\ 0 & k \text{ in } \{1,\ldots,K\} \setminus E \end{cases} .$$

(iii.b) Else, for every k in E such that $u_k(E) < 0$, set $E \leftarrow E \setminus \{k\}$ and go to step **(ii)**.

V.5. THE FINITE HORIZON AND LONG-RUN AVERAGE COSTS

In this section, the finite horizon and the long-run average cost problems are briefly discussed. It is shown that in both cases the optimal policy is the one given for the discounted cost problem.

V.5.1. The Finite Horizon Problem

For any policy π in Π and $n = 0,1,\ldots$, the n-stage total and average expected costs are defined respectively by

$$J_n^\pi(w,\sigma) := E\left[\sum_{i=0}^{n} \|W^\pi(i) + \sigma(i)U^\pi(i)\| \;\middle|\; W = w,\ \Sigma = \sigma\right] \tag{5.5.1}$$

and

$$\widetilde{J}_n^\pi(w,\sigma) := \frac{1}{n+1} J_n^\pi(w,\sigma) \tag{5.5.2}$$

for all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$. The corresponding *value functions* are given respectively by

$$V_n(w,\sigma) = \inf_{\pi\in\Pi} J_n^\pi(w,\sigma)$$

and

$$\widetilde{V}_n(w,\sigma) = \inf_{\pi\in\Pi} \widetilde{J}_n^\pi(w,\sigma) = \frac{1}{n+1} V_n(w,\sigma) .$$

For the total expected cost the dynamic programming equation is given by

$$V_0(w,\sigma) = \min_{u\in\mathcal{U}} \|w + \sigma u\| \tag{5.5.3}$$

$$V_m(w,\sigma) = \min_{u\in\mathcal{U}} \left\{ \|w + \sigma u\| + E\left[V_{m-1}\left([w + \sigma u - Ae]^+, B\right)\right]\right\} .\quad m = 1,\ldots,n$$

Let π_m^* be the vector in $\mathcal{U}$ which minimizes the right hand side of the m^{th} equation in (5.5.3) for $m = 0, 1, \ldots, n$, and let $\pi^* = \{\pi_m^*,\ m = 0, 1, \ldots, n\}$ be the corresponding optimum policy. The following lemma shows that the functions $w \mapsto V_m(w, \sigma),\ m = 0, 1, \ldots, n$ all enjoy the structural properties of the function $V_\beta(w, \sigma)$. Consequently, an argument similar to the one given in Section V.4 shows that $\pi^*(w, \sigma)$ given in (5.4.6) is optimum for the (total expected cost) finite horizon problem as well. To that end, let Rw be any permutation of the vector w in $\mathbb{R}^K$, and let R^{-1} be the inverse of R, i.e., $R^{-1}(Rw) = R(R^{-1}w) = w$.

Lemma V.5.1. *For every σ in $\mathbb{R}_+$, the mappings $w \mapsto V_m(w, \sigma),\ m = 0, 1, \ldots, n$ are convex and have the property $V_m(w, \sigma) = V_m(Rw, \sigma)$.*

The following Lemma will be useful in proving Lemma V.5.1, and follows from Lemma A.III.2 of Appendix III by observing that $R(\mathcal{U}) = \mathcal{U}$.

Lemma V.5.2. *If the mapping $\psi : \mathbb{R}_+^K \times \mathcal{U} \mapsto \mathbb{R}_+$ is jointly convex and has the property $\psi(Rw, Ru) = \psi(w, u)$ for all permutation R, then the mapping $\phi : \mathbb{R}_+^K \mapsto \mathbb{R}_+$ given by*

$$\phi(w) := \min_{u \in \mathcal{U}} \psi(w, u)$$

is also convex and has the property $\phi(w) = \phi(Rw)$.

Proof of Lemma V.5.1. The result follows by induction. The function $(w, u) \mapsto \psi_0(w, u) = \|w + \sigma u\|$ clearly satisfies the assumptions of Lemma V.5.2. The result is therefore true for the function $w \mapsto V_0(w, \sigma)$ for every σ in $\mathbb{R}_+$.

For $m = 1, 2, \ldots$, write

$$V_m(w, \sigma) = \min_{u \in \mathcal{U}} \psi_m(w + \sigma u)$$

where

$$\psi_m(x) = \|x\| + E\left[V_{m-1}\left([x - Ae]^+, B\right)\right] .$$

For every σ in $\mathbb{R}_+$, if the function $V_{m-1}(\cdot, \sigma)$ is convex and has the property $V_{m-1}(R\,\cdot, \sigma) = V_{m-1}(\cdot\ , \sigma)$, then the function $\phi_m(w, u) := \psi_m(w + \sigma u)$ clearly

satisfies the conditions of Lemma V.5.2. The function $w \mapsto V_m(w,\sigma)$ therefore has the desired properties, thus completing the induction argument.

□

Since the optimal policy π^* for the total expected cost is independent of n, it is also optimum for the average finite horizon problem. This can also be seen by writing the corresponding dynamic programming equation for $\widetilde{V}_n$ and repeating the arguments presented above.

V.5.2. The Long-run Average Cost Problem

The long-run average cost for any policy π in Π is given by

$$\widetilde{J}^{\pi}(w,\sigma) = \overline{\lim_{n\to\infty}}\, \widetilde{J}_n^{\pi}(w,\sigma) \tag{5.5.4}$$

for all w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$. Since $\widetilde{J}_n^{\pi}(w,\sigma)$ is minimized by π^* for each n, the long-run average cost is therefore also minimized by π^*. In this section the optimum cost $\widetilde{J}^{\pi^*}(w,\sigma)$ is obtained under the stability condition $EB < K\,EA$.

Let the subset $\mathcal{B}$ of $\mathbb{R}_+^K$ be defined by

$$\mathcal{B} := \{w \in \mathbb{R}_+^K :\ w = c\,e \text{ for some } c \geq 0\}\ .$$

Also define, with a slight abuse of notation, the work process $\{W^{\pi}(t),\ t \geq 0\}$ for every π in Π, where $W_k^{\pi}(t)$ is the workload in queue k at time t under the policy π, $1 \leq k \leq K$. Note that with this definition, if t_n is the n^{th} arrival epoch, then $W^{\pi}(t_n) = W^{\pi}(n)$ for $n = 0, 1, \ldots$.

If the RV T^{π} is defined by

$$T^{\pi} := \inf\{t \geq 0 :\ W^{\pi}(t) \in \mathcal{B}\}\ ,$$

then

$$T^{\pi^*} \leq W \qquad a.s. \tag{5.5.5}$$

Note that T^{π} is the first time the work process reaches a balanced configuration under the policy π in Π. To see (5.5.5), let $T_{\omega}^{\pi^*}$ be a realization of T^{π^*} for $\omega =$

$(w, \sigma(0), \tau(1), \sigma(1), \ldots)$ in Ω. If $\|w\| \leq \tau(1)$, then then the system will be empty at time $\|w\|$, and $T_w^{\pi^*} = \|w\|$. On the other hand, since π^* steers the system to a balanced position, arrivals into the system will only result the system to reach to a balanced position earlier.

Let $\lfloor \cdot \rfloor$ denote the integer part of a real number. The sequence of RVs $\{f_n,\ n = 0, 1, \ldots\}$ defined by

$$f_n := \frac{1}{n+1} \sum_{i=0}^{\lfloor T^{\pi^*} \rfloor} S^{\pi^*}(i) \qquad n = 0, 1, \ldots$$

is uniformly bounded by $\sum_{i=0}^{\lfloor W \rfloor} S^{\pi^*}(i)$ by virtue of (5.5.5) and converge to 0 *a.s.* as $n \to \infty$. It is plain that the RV $\sum_{i=0}^{\lfloor W \rfloor} S^{\pi^*}(i)$ has finite conditional mean given W. Therefore, for some c and σ' in $\mathbb{R}_+$, the relation

$$\lim_{n\to\infty} \widetilde{J}_n^{\pi^*}(w, \sigma) = \lim_{n\to\infty} \widetilde{J}_n^{\pi^*}(c\ e, \sigma') \tag{5.5.6}$$

holds by the Bounded Convergence Theorem [Bil, p. 180] for every w in $\mathbb{R}_+^K$ and σ in $\mathbb{R}_+$.

Under the assumed stability condition $EB < K\ EA$, the right hand side of (5.5.6) converges to the average response time in the $GI/GI/1$ queue with interarrival distribution $A(x)$ and the service time distribution $B(Kx)$ by Corollary V.4.6. The following theorem summarizes this argument.

Theorem V.5.3. *The policy π^* given in (5.4.6) is also optimal for the long-run average problem. Furthermore, if $EB < K\ EA$, then the corresponding long-run average cost is equal to the average response time in the $GI/GI/1$ queue with interarrival time distribution $A(\cdot)$ and the service time distribution $B(K\cdot)$.*

V.6. OPTIMAL SCHEDULING WITH NO PIPELINING: K = 2

This section considers the problem formulated in Section 2 when $K = 2$ and when the control set $\mathcal{U} = \{u \in \{0,1\}^2 :\ u_1 + u_2 = 1\}$, i.e., when the incoming jobs are not allowed to be broken into smaller tasks and are scheduled to the parallel servers in one piece. It is plain that the structural results of Section 3 still hold in this case and that the optimal policy π^* satisfying (5.3.13) is Markov stationary.

First it is shown that the vector $u^*(w,\sigma)$ in $\mathcal{U}' = \{u \in \mathbb{R}^2 : \ u_1 + u_2 = 1\}$ given by (5.4.3) can also be used to characterize the optimal scheduling policy in this case. For every ϵ in $\mathbb{R}_+$, let u_ϵ^1 and u_ϵ^2 be the two vectors in $\mathcal{U}'$ which are at a distance $\sqrt{2}\ \epsilon$ from $u^*(w,\sigma)$, i.e.,

$$u_\epsilon^1 = u^*(w,\sigma) + \epsilon \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{5.6.1a}$$

and

$$u_\epsilon^2 = u^*(w,\sigma) + \epsilon \begin{pmatrix} -1 \\ 1 \end{pmatrix} . \tag{5.6.1b}$$

The following result states that the function $\mathcal{U}' \mapsto \mathbb{R}_+ : u \mapsto T^u V_\beta(w,\sigma)$ is symmetric around $u^*(w,\sigma)$ for every w in $\mathbb{R}_+^2$ and σ in $\mathbb{R}_+$.

Lemma V.6.1. *If the vectors u_ϵ^1 and u_ϵ^2 are as given in (5.6.1), then*

$$T_\beta^{u_\epsilon^1} V_\beta(w,\sigma) = T_\beta^{u_\epsilon^2} V_\beta(w,\sigma)$$

for every ϵ and σ in $\mathbb{R}_+$ and w in $\mathbb{R}_+^2$.

Proof. Since $w + \sigma u^*(w,\sigma) = c\ e$ where $c = (\sigma + w_1 + w_2)/2$, the relations

$$w + \sigma u_\epsilon^1 = c\ e + \sigma\epsilon \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

and

$$w + \sigma u_\epsilon^2 = c\ e + \sigma\epsilon \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

easily follow from (5.6.1). Therefore, the relation

$$R_1(w + \sigma u_\epsilon^1) = R_1(w + \sigma u_\epsilon^2)$$

holds for the permutation R_1 defined in (5.3.14). The result now follows from Lemma V.3.4. and the definition of the function $T_\beta^u V_\beta$ in (5.3.1).

□

It is clear from Lemma V.6.1 that it is optimal to schedule the arriving job to the first (resp. second) queue if $u^*(w,\sigma)$ is closer to $\begin{pmatrix}1\\0\end{pmatrix}$ (resp. $\begin{pmatrix}0\\1\end{pmatrix}$) than $\begin{pmatrix}0\\1\end{pmatrix}$ (resp. $\begin{pmatrix}1\\0\end{pmatrix}$). The following result is therefore obvious given the form of $u^*(w,\sigma)$ in (5.4.3).

Theorem V.6.2. *For every σ in $\mathbb{R}_+$ and w in $\mathbb{R}_+^2$, the optimal policy π^* is given by*

$$\pi^*(w,\sigma) = \begin{cases} \begin{pmatrix}1\\0\end{pmatrix} & \text{if } w_1 \leq w_2 \\ \\ \begin{pmatrix}0\\1\end{pmatrix} & \text{if } w_1 > w_2 \ . \end{cases}$$

In words, it is optimum to join the queue with the smaller workload.

The natural extension of this result to the case of K (≥ 2) queues is currently under investigation.

CONCLUSIONS AND FUTURE RESEARCH

In this thesis, issues of performance evaluation and optimal routing in a system with K parallel queues were studied under the resequencing constraint. In Chapters II-IV, jobs arrive according to a Poisson process and are allocated randomly to the parallel queues by a Bernoulli switch.

In Chapter II, the service time distributions at different queues are assumed to be independent with general and possibly different distributions. The system under the resequencing constraint is compared to the system without the resequencing requirement. It is established that the presence of resequencing severely complicates the analysis and considerably degrades system performance. In contrast to the simple expressions for the response time T, the expressions for the distributions of the resequencing and system delays are very complicated. In particular, the static optimization problem of choosing the Bernoulli switching probabilities to minimize the average system time ES can not be carried out by hand for the general case. On the other hand, the RV T is shown to be stochastically convex in the load allocation vector in the *st* sense defined in Appendix II, and the minimization of the average response time ET is carried out for the general case.

In the presence of resequencing, two special cases are considered for the optimization problem: (i) Identical servers and (ii) Exponential servers. When the service time distributions at different queues are identical, ES is minimized at equal load allocation, and at this optimal configuration, ES is decreasing in K. Stronger results are obtained for the RV T. Indeed, equal load allocation stochastically minimize T, and at this optimum configuration, T is integer convex and decreasing in K in the *st* sense. The weaker result, namely, the asymptotic convexity of ES in K is established in Chapter IV. Monotonicity and convexity results in the *st* sense for the RV S are currently under investigation for an arbitrary K. A weaker asymptotic result in that direction is included in Chapter IV.

For the exponential service time distributions, with even $K = 2$, the optimum switching probability vector does not admit a closed form expression and is char-

acterized by the unique root of a fifth order polynomial. Nevertheless, the study of the optimization problem for the case $K = 2$ lead to simple but very accurate approximations for $K \geq 2$. The asymptotic approximations are not limited to exponential servers and can be applied to any family of distributions with one parameter. For instance, approximations for the optimal probabilities displayed in Figures II.2 and II.3 may be obtained by the same technique. Since the curves C_p displayed in these figures are closer to their asymptotes then the ones drawn in Figure II.1, the asymptotic approximations are believed to be even better for parallel systems with less variable service time distributions.

Both the approximate and the exact (when applicable) expressions for the optimum switching probabilities indicate that the faster servers have to carry more traffic both with and without the resequencing constraint, but that the resequencing requirement tends to further decrease the amount of the traffic switched to the slower servers.

In Chapter III, the simple form of the approximate formula is used in a stochastic approximation algorithm when the system parameters are unknown. The algorithm makes use of some system measurements to update the switching vector. The choice of the lengths of the measurement intervals crucially affects the rate of convergence and the robustness of the algorithm. This is left as a future research topic.

In Chapter IV, the asymptotic behavior of various system delays in K is studied when the load is equally allocated to the homogeneous servers. When the Poisson arrival rate is held fixed, asymptotic expansions are provided for the distribution functions and the first moments of the RVs T, R and S. These asymptotic expressions are used to prove convergence of the system statistics to those of the $M/GI/\infty$ system with resequencing, as K goes to infinity. The asymptotic expansions are also used to establish asymptotic stochastic monotonicity and convexity of the RV S in K, while the RV R asymptotically may increase or decrease depending on the arrival rate. Therefore, despite the different behavior of R, the RVs T and S have (asymptotically) similar structural characteristics.

In the case when the arrival rate into the system is also increased so as to keep a fixed load to each queue, the resequencing delay dominates the queueing delay; while ET remains the same, ER grows to infinity in $\log K$.

In Chapter V, the parallel system is considered under a different set of assumptions on the arrival and service processes. In this chapter, jobs arrive according to a renewal sequence and are broken into smaller tasks for processing at different queues. The servers are assumed identical with fixed capacities. The optimum, dynamic load allocation problem is considered when the workloads of the jobs are *i.i.d.* and independent from the arrival time sequence. The cost associated for allocating each job is chosen to be its end-to-end delay, including the synchronization delays due to the reassembly and resequencing operations. The optimal allocation policy that minimizes both the long run average and the discounted costs is shown to be the one that derives the workloads in the queues to a balanced position as fast as feasible. The same optimization problem when the queues have different processing capacities is currently under investigation.

The solution to the optimization problem considered in Chapter V also sheds light into the characterization of the optimum scheduling policy when the jobs are *not* allowed to be broken into smaller tasks. For the case of two parallel queues, scheduling the incoming jobs to the queue with the smaller workload is shown to be optimal. The optimality of this policy for the case of $K > 2$ parallel queues is also under investigation.

APPENDIX I

COMPUTATION OF THE OPTIMAL PROBABILITY VECTOR p*

In this Appendix an algorithm for computing the optimal probability vector p^* of Section II.4.1 is given. A simplified version of this algorithm is also provided when the service time distributions are exponential.

As noted in Section II.4.1, if the numbers p_k^*, $1 \leq k \leq K$, obtained from (2.4.1) are all non-negative, then they lie in the set $\mathcal{D}$ and constitute the solution to the optimization problem $\min\{ET(p) : p \in \mathcal{D}\}$. On the other hand, if one or more of them are negative, from Theorem II.4.1, the solution is on the boundary of $\mathcal{D}$, i.e., at least one of the p_k^*'s is zero and equation (2.4.1) is applied to the reduced system. The following argument provides a computationally efficient way of computing p^* in this case.

Let the functions f_l, $1 \leq l \leq K$, be defined by

$$f_l(x) = 1 - \sum_{\substack{k=1 \\ k \neq l}}^{K} \frac{\mu_k}{\lambda} \left[1 - \sqrt{\frac{1+\sigma_k^2\mu_k^2}{2\mu_k x - 1 + \sigma_k^2\mu_k^2}} \right]$$

for $x \geq \max_{1 \leq k \leq K}[(1-\sigma_k^2\mu_k^2)/2\mu_k]^+$. The functions f and f_l, $1 \leq l \leq K$, are all decreasing and have unique zeros, denoted by y and y_l, $1 \leq l \leq K$, respectively. Denote p_k^* in equation (2.4.1a) by $p_k^*(y)$. Note that $p_k^*(y)$ and $p_k^*(y_l)$ correspond to the optimal probabilities over the set of indices $E = \{1, 2, \ldots, K\}$ and $E \setminus \{l\}$, respectively. The following lemma leads to Algorithm A.I.1 for computing the optimal probability vector p^*.

Lemma A.I.1. *If $p_l^*(y) < 0$ for some l in E, then $p_k^*(y) < 0$ for some k in $E \setminus \{l\}$ implies $p_k^*(y_l) < 0$.*

Proof. Since the function f_l is decreasing and $f_l(y) = f(y) + p_l^*(y) = p_l^*(y)$, the condition $p_l^*(y) < 0$ implies $y_l < y$. Therefore $p_k^*(y_l) < p_k^*(y) < 0$, since the functions p_k^* are all increasing.

□

Algorithm A.I.1.

(i) Set $E \leftarrow \{1, 2, \ldots, K\}$.

(ii) Compute y as the solution of the equation

$$\sum_{k \in E} \frac{\mu_k}{\lambda} \left[1 - \sqrt{\frac{1 + \sigma_k^2 \mu_k^2}{2\mu_k x - 1 + \sigma_k^2 \mu_k^2}} \right] = 1 .$$

(iii) For all k in E, compute p_k^* as

$$p_k^* = \frac{\mu_k}{\lambda} \left[1 - \sqrt{\frac{1 + \sigma_k^2 \mu_k^2}{2\mu_k y - 1 + \sigma_k^2 \mu_k^2}} \right] .$$

Let $F \subset E$ be such that k is in F if and only if $p_k^* < 0$.

(iv) If $F = \emptyset$, then STOP.

Else, set $E \leftarrow E \setminus F$, $p_k^* \leftarrow 0$ for every k in F, and go to **(ii)**.

Note that the monotonicity of the functions f_l can be used to computational advantage in step **(ii)**. However, this step is still the most computationally intensive step of the algorithm, and must be avoided as much as possible. Lemma A.I.1 allows for the index set E to be reduced to $E \setminus F$ in step **(iv)** at once, instead of reducing E one element at a time by stopping when a negative p_k^* is found. This provides an improvement over the algorithm given in Buzen and Chen [BuC]. Note that $p_k^*(y) > 0$ does not imply $p_k^*(y_l) > 0$, and the algorithm needs to go back to step **(ii)**.

In the exponential case, this algorithm can be improved by using the simplified formula (2.4.2). Assume, without loss of generality, that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$. It is easy to show that if $p_k^* \geq 0$, then $p_{k-1}^* \geq p_k^*$, $1 < k \leq K$. Therefore, if one of the probabilities is positive, so are the ones with a lower index. In particular, if $p_K^* \geq 0$, then all the p_k^*'s will be probabilities. This simplifies step **(iii)** of Algorithm A.I.1 and leads to the following algorithm

Algorithm A.I.2.

(i) Set $n \leftarrow K$

(ii) Compute p_n^*:

$$p_n^* = \frac{\mu_n}{\lambda} - \left(\frac{\sum_{i=1}^n \mu_i}{\lambda} - 1\right)\frac{\sqrt{\mu_n}}{\sum_{i=1}^n \sqrt{\mu_i}}$$

(iii) If $p_n^* < 0$, then set $p_n^* \leftarrow 0$, $n \leftarrow n-1$, and go to **(ii)**.

If $p_n^* \geq 0$, then $p^* = (p_1^*, \ldots, p_n^*, 0, \ldots, 0)$, where

$$p_k^* = \frac{\mu_k}{\lambda} - \left(\frac{\sum_{i=1}^n \mu_i}{\lambda} - 1\right)\frac{\sqrt{\mu_k}}{\sum_{i=1}^n \sqrt{\mu_i}}, \qquad 1 \leq k \leq n.$$

APPENDIX II

STOCHASTIC CONVEXITY RESULTS FOR THE M/GI/1 QUEUE

In this appendix, a notion of strong stochastic convexity [SS.b] is defined and a stochastic convexity result is established for the $M/GI/1$ queue by means of simple arguments. This convexity result states that the stationary waiting and response times in the $M/GI/1$ queue are both stochastically increasing (resp. decreasing) and convex in the arrival (resp. service) rate. This provides an important step in establishing some of the convexity and optimality results in Chapters II and IV.

Let Θ be a convex subset of $\mathbb{R}^K$, and let $\{X(\theta),\ \theta \in \Theta\}$ be a family of $\mathbb{R}_+$-valued RVs.

Definition A.II.1. $\{X(\theta),\ \theta \in \Theta\}$ is *stochastically convex* (resp. convex and increasing/decreasing) on Θ if the $\mathbb{R}^K \to \mathbb{R}$ mapping $\theta \mapsto E\ f(X(\theta))$ is convex (resp. convex and increasing/decreasing) for every *increasing* function $f : \mathbb{R} \mapsto \mathbb{R}$. This is denoted by $\{X(\theta),\ \theta \in \Theta\} \in SCX(st)$ (resp. $SICX(st)/SDCX(st)$).

When θ is $\mathbb{N}$-valued the same definition applies with convexity replaced by integer convexity. It is an easy exercise to show the following result [SS.b].

Lemma A.II.1. *$\{X(\theta),\ \theta \in \Theta\} \in SCX(st)$ (resp. $SICX(st)/SDCX(st)$) if and only if the complementary distribution function $\theta \mapsto P(X(\theta) > x)$ is convex (resp. increasing/decreasing convex) for every $x \geq 0$.*

First, the complementary waiting and response time distributions in the $M/PH/1$ queue are shown to be monotone increasing and convex in the system utilization. Therefore, the waiting and response times in the $M/PH/1$ queue are both stochastically increasing and convex in system utilization by Lemma A.II.1. The results are then extended to the $M/GI/1$ queue using the fact that the class of PH-distributions is *dense* in the space of probability distributions on $[0, \infty)$ [Neu.c].

In this appendix, I denotes the identity matrix with appropriate dimensions, while the $m \times m$ and the $1 \times m$ row vector with zero entries are denoted by O_m

and 0_m, respectively. Finally, the notation $\Re(s)$ denotes the real part of a complex number s.

Waiting and response time distributions for the M/PH/1 queue

Consider an $M/PH/1$ queue under the FCFS discipline where the Poisson arrival process has rate λ and the PH-type service time distribution of order m has an irreducible representation (α, A) with mean $\frac{1}{\mu}$. The matrix A is invertible so that absorption into the state $m+1$ from any initial state is certain [Neu.b, p. 45]. The corresponding $1 \times m$ column vector of absorption probabilities is denoted by A^0, i.e., $Ae = -A^0$. It is assumed that $\alpha_{m+1} = 0$ and that the queueing system is stable, i.e., $\rho := \frac{\lambda}{\mu} < 1$.

Let $W(\cdot)$ be the stationary waiting time distribution. The following well-known result was given in [Neu.a, p. 181].

Theorem A.II.2. *The waiting time distribution $W(\cdot)$ is PH-type and has a representation (γ, L) of order m with*

$$\gamma = \rho\,\pi, \qquad \gamma_{m+1} = 1 - \rho, \tag{A.2.1a}$$

$$L = A + \rho\,A^0\,\pi, \qquad L^0 = (1-\rho)A^0, \tag{A.2.1b}$$

where the probability vector π is uniquely determined by the relations

$$\pi(A + A^0\alpha) = 0_m \quad \textit{and} \quad \pi e = 1.$$

In an $M/GI/1$ queue in statistical equilibrium, the response time T is the sum of two independent RVs, the waiting time W and the service time B. Therefore, in view of Theorem A.II.2 and of closure properties of the PH-distributions [Neu.b, p. 51], it is an easy exercise to see that the stationary response time distribution $T(\cdot)$ is also PH-type of order $2m$ with representation (ξ, Z), where

$$\xi = (\rho\pi, (1-\rho)\alpha) \qquad \text{and} \qquad Z = \begin{pmatrix} L & L^0\alpha \\ O_m & A \end{pmatrix}. \tag{A.2.2}$$

The following theorem shows that a *lower* order PH-representation with only m phases is in fact available for $T(\cdot)$.

Theorem A.II.3. *The response time distribution $T(\cdot)$ is PH-type with representation (α, L).*

Proof. The Laplace-Stieltjes transform $r(s)$ of the PH-distribution (ξ, Z) is given by

$$r(s) = \xi(sI - Z)^{-1}Z^0\,, \qquad \Re(s) \geq 0.$$

By making use of equations (A.2.1) and (A.2.2), direct calculations now yield

$$r(s) = (\rho\pi, (1-\rho)\alpha) \begin{pmatrix} (sI-L)^{-1} & (sI-L)^{-1}L^0\alpha(sI-A)^{-1} \\ O_m & (sI-A)^{-1} \end{pmatrix} \begin{pmatrix} 0_m^T \\ A^0 \end{pmatrix}$$

$$\begin{aligned} &= \rho\pi(sI-L)^{-1}L^0\alpha(sI-A)^{-1}A^0 + (1-\rho)\alpha(sI-A)^{-1}A^0 \\ &= \rho\alpha(sI-A)^{-1}A^0\pi(sI-L)^{-1}L^0 + \alpha(sI-A)^{-1}L^0 \\ &= \alpha\left[\rho(sI-A)^{-1}A^0\pi + (sI-A)^{-1}(sI-L)\right](sI-L)^{-1}L^0 \\ &= \alpha(sI-L)^{-1}L^0\,, \qquad \Re(s) \geq 0, \end{aligned}$$

thus completing the proof.

□

Remark A.II.1. The simple $M/PH/1$ queue provides a building block in the approximate decomposition/aggregation algorithms for analyzing queueing models of real life systems [Gün.b]. In such an iterative algorithmic analysis, low order representations for various distributions in the network are often desirable from a computational standpoint. Theorem A.II.3 serves this purpose and provides a PH-representation for T with only *half* the dimension of (ξ, Z). Furthermore, Theorem A.II.3 shows that the representations of T and W differ *only* in the way they are initialized.

Strong convexity results for the M/GI/1 queue

Theorems A.II.2 and A.II.3 are used to study the variation of the RVs W and T with the system utilization for the $M/PH/1$ queue. The results are then extended to the $M/GI/1$ queue.

The notation $W(\rho)$ and $T(\rho)$ is now used to represent the RVs W and T, respectively, in order to indicate the dependence of their distributions on ρ explicitly. In particular, monotonicity and convexity of the functions $\rho \mapsto \overline{W}(\rho, x) := P\{W(\rho) > x\}$ and $\rho \mapsto \overline{T}(\rho, x) := P\{T(\rho) > x\}$ are established in Theorem A.II.4 for all $x \geq 0$.

Theorem A.II.4. *All the partial derivatives of $\overline{W}(\rho, x)$ and $\overline{T}(\rho, x)$ with respect to ρ exists and are positive for all $x \geq 0$. In particular, the mappings $\rho \mapsto \overline{W}(\rho, x)$ and $\rho \mapsto \overline{T}(\rho, x)$ are both monotone increasing and convex for all $x \geq 0$.*

Equivalently, $\{W(\rho),\ \rho \in [0,1)\}$ and $\{T(\rho),\ \rho \in [0,1)\}$ are both $SICX(st)$.

Proof. By Theorems A.II.2 and A.II.3, the complementary distribution functions $\overline{W}(\rho, x)$ and $\overline{T}(\rho, x)$ are given, respectively, by

$$\overline{W}(\rho, x) = \rho\pi e^{(A+\rho A^0\pi)x}e \qquad \text{and} \qquad \overline{T}(\rho, x) = \alpha e^{(A+\rho A^0\pi)x}e. \qquad (A.2.3)$$

It is a simple exercise to see that if the matrices $\partial^n e^{(A+\rho A^0\pi)x}/\partial\rho^n$ have positive components, then the partial derivatives $\partial^n\overline{W}(\rho, x)/\partial\rho^n$ and $\partial^n\overline{T}(\rho, x)/\partial\rho^n$ are all positive for $n \geq 1$.

Since the matrices A and $A^0\pi$ do not commute, the matrices $\partial^n e^{(A+\rho A^0\pi)}/\partial\rho^n$, $n \geq 1$, do not have a simple closed form expression. Therefore, consider the matrix

$$E(\rho) := e^{\left(x(A+cI)+\rho x A^0\pi\right)} = e^{cx}\ e^{\left(A+\rho A^0\pi\right)x}, \qquad x \geq 0$$

where $c := \max\{-A_{ii} :\ 1 \leq i \leq m\}$. Since both $F := x(A + cI)$ and $G := xA^0\pi$ are positive matrices,

$$E(\rho) = e^{F+\rho G} = \sum_{k=0}^{\infty} \frac{(F + \rho G)^k}{k!}$$

is a polynomial in ρ with positive coefficient matrices. The partial derivatives $\partial^n E(\rho)/\partial\rho^n$ and $\partial^n e^{(A+\rho A^0\pi)x}/\partial\rho^n$ are therefore positive for all $n \geq 1$, and the first part of the theorem thus follows. The monotonicity and convexity of the mappings $\rho \mapsto \overline{W}(\rho, x)$ and $\rho \mapsto \overline{T}(\rho, x)$ are now immediate since $\partial\overline{W}(\rho, x)/\partial\rho$, $\partial^2\overline{W}(\rho, x)/\partial\rho^2$, $\partial\overline{T}(\rho, x)/\partial\rho$ and $\partial^2\overline{T}(\rho, x)/\partial\rho^2$ are all positive for $x \geq 0$.

□

The structural result of Theorem A.II.4 is now extended to the $M/GI/1$ queue.

Corollary A.II.5. *If $W(\rho)$ and $T(\rho)$ are the waiting and response times in an $M/GI/1$ queue with utilization ρ, then $\{W(\rho),\ \rho \in [0,1)\}$ and $\{T(\rho),\ \rho \in [0,1)\}$ are both $SICX(st)$.*

Proof. The result follows from Proposition 8.2.5a of [Sto, pp. 169-170] using the fact that the PH-distributions are dense in the space of probability distributions on $[0, \infty)$.

□

Remark A.II.2. The more general $GI/GI/1$ queue is considered in [SS.a] using a sample path approach and sufficient conditions for *sample path* convexity of the waiting time in the parameter(s) of the service and interarrival times are given (see also [ShY]). However, it does *not* seem possible to establish the stochastic convexity of the waiting and response times in the *arrival rate* from the results in [SS.a] or [ShY]. Moreover, the stochastic convexity considered here is *stronger* than the sample path convexity defined in [SS.a] (see [SS.b]). The sample path approach allowed the consideration of a very general class of problems in [FeG]. However, when specialized to the $M/GI/1$ queue, only the convexity of the *average* waiting time in the arrival rate was established in [FeG].

Remark A.II.3. In [SS.a], it is shown for the $GI/GI/1$ queue that, if the n^{th} service time $S_n(\mu) = X_n/\mu$ where $\{X_n,\ n = 0, 1, \ldots\}$ is a sequence of non-negative *i.i.d* RVs, then the waiting time of each job is stochastically decreasing and convex

in μ in the sample path sense. This result generalizes a result established in [Web] again by sample path arguments, namely that the *average* waiting time for each job is convex in the service rate. For the $M/GI/1$ queue, the *stationary* version of the convexity result of [SS.a] in the stronger $SDCX(st)$ sense easily follows from Corollary A.II.5.

APPENDIX III

This appendix states two technical lemmas. Lemma A.III.1. is used in the proof of Theorem II.4.6, while Lemma A.III.2 is used in proving Theorems V.3.1 and V.5.2.

Lemma A.III.1. *Let I be a convex subset of $\mathbb{R}$ with $\frac{1}{K}$ in I, and let $f : I \mapsto \mathbb{R}$ be a concave function. If $F : I^K \mapsto \mathbb{R}$ is defined by $F(x) = \sum_{k=1}^{K} f(x_k)$, then*

$$\max\{F(x) : \ x \in I^K \quad \text{and} \quad \sum_{k=1}^{K} x_k = 1\} = F(\frac{1}{K}, \ldots, \frac{1}{K}) \ .$$

Proof. The result follows from the following simple argument

$$\begin{aligned} F(x) &= K \sum_{k=1}^{K} \frac{1}{K} f(x_k) \\ &\leq K f(\frac{1}{K} \sum_{k=1}^{K} x_k) \\ &= K f(\frac{1}{K}) \\ &= F(\frac{1}{K}, \ldots, \frac{1}{K}) \ . \end{aligned}$$

□

Lemma A.III.2. *Let $\Omega_i \subseteq \mathbb{R}^{n_i}$, $i = 1, 2$, be two convex sets. If the mapping $\Psi : \Omega_1 \times \Omega_2 \mapsto \mathbb{R}$ is jointly convex, then the mapping $\Phi : \Omega_1 \mapsto \mathbb{R}$ defined by*

$$\Phi(x) = \min_{y \in \Omega_2} \Psi(x, y) \ , \qquad x \text{ in } \Omega_1$$

is also convex.

Proof. The convexity result follows directly from Theorem 5.7 of [Roc, p. 38].

□

APPENDIX IV

A Proof of (3.3.1)

Consider a stable $M/M/1$ queue with utilization ρ, and let $\{T_n,\ n = 1, 2, \ldots\}$ be the Poisson sampling process with rate ν independent of the arrival and the service processes defining the $M/M/1$ queue. Let N_τ be the number of samples in the interval $[0, \tau]$, and let η_τ be the times the server is sampled idle in this interval. If $Q(t)$ is the number of customers in the system at time t, then

$$N_\tau = \sum_{n=1}^{\infty} 1(T_n \leq \tau)$$

and

$$\eta_\tau = \sum_{n=1}^{N_\tau} 1(Q(T_n) = 0)\ ,$$

where $1(\cdot)$ is the indicator function.

Note that $\{Q(T_n),\ n = 0, 1, \ldots\}$ with $T_0 = 0$ is a Markov chain with invariant probability mass $\widetilde{\pi}$, i.e., $\widetilde{\pi}_k = \lim_{n\to\infty} P[Q(T_n) = k]$, $k = 0, 1, \ldots$, and the *a.s.* relation

$$\lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n} 1(Q(T_l) = 0) = \widetilde{\pi}_0$$

thus holds by the Ergodic theorem. Therefore,

$$\lim_{\tau\to\infty} \frac{\eta_\tau}{N_\tau} = \widetilde{\pi}_0 \qquad a.s. \tag{A.4.1}$$

since $\lim_{\tau\to\infty} N_\tau = \infty$. The relation

$$\lim_{\tau\to\infty} \frac{\eta_\tau}{N_\tau} = 1 - \rho \qquad a.s. \tag{A.4.2}$$

now follows from (A.4.1) and the PASTA property [Kle.a].

In order to show the second equation in (3.3.1), divide the interval $[0, \tau]$ into n equal subintervals of length t, i.e., $nt = \tau$. Let η_k be the number of times the

server is sampled idle in the interval $[(k-1)t, kt)$, $k = 1, \ldots, n$, and let N_k be the number of measurements in this interval. Note that $\{N_k,\ k = 1, 2, \ldots\}$ is *i.i.d.* with $E[N_k] = \nu t$, and

$$\eta_\tau = \eta_{nt} = \sum_{k=1}^{n} \eta_k \quad \text{and} \quad N_\tau = N_{nt} = \sum_{k=1}^{n} N_k \ .$$

Therefore,

$$\lim_{n\to\infty} \frac{\eta_{nt}}{N_{nt}} = \lim_{n\to\infty} \frac{\eta_{nt}/n}{\sum_{k=1}^{n} N_k/n} = 1 - \rho \qquad a.s.$$

by (A.4.2), so that the result

$$\lim_{n\to\infty} \frac{\eta_{nt}}{n\nu t} = \lim_{\tau\to\infty} \frac{\eta_\tau}{\tau\nu} = 1 - \rho \tag{A.4.3}$$

holds since

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} N_k = \nu t$$

by the Law of Large Numbers.

REFERENCES

[AgR] Agrawal S. and Ramaswamy R., "Analysis of the resequencing delay for $M/M/m$ systems", *Proc. ACM SIGMETRICS*, pp. 27-35, Alberta, Canada, May 1987.

[Bac] Baccelli F., " A queueing model of timestamp ordering in a distributed system", *Proc. Performance'87*, pp. 413-431, Brussels, Belgium, December 1987.

[BaM] Baccelli F. and Makowski A.M., "Queueing models for systems with synchronization constraints", *Proc. of the IEEE* **77**, *Special Issue on Discrete Event Systems*, pp. 138-161, January 1989.

[BGP] Baccelli F., Gelenbe E. and Plateau B., "An end-to-end approach to the resequencing problem", *J. ACM* **31**, pp. 474-485, July 1984.

[Bil] Billingsley P., *Probability and Measure*, J. Wiley & Sons, New York, NY, 1979.

[BKK] Bharat-Kumar K. and Kermani P., "Analysis of a resequencing problem in communication networks", *Proc. IEEE INFOCOM'83*, pp. 303-310, San Diego, CA, April 1983.

[BMS] Baccelli F., Makowski A.M. and Shwartz A., "The fork-join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds", *Adv. in Appl. Prob.* **21**, September 1989.

[BoK] Bonomi F. and Kumar A., "Adaptive optimal load balancing in a non-homogeneous multiserver system with a central job scheduler", *Proc. of the* 8^{th} *International Conference on Distributed Computer Systems*, pp. 500-507, San Jose, CA, June 1988.

[Bor] Borovkov A.A., *Stochastic Processes in Queueing Theory*, Springer-Verlag, New York-Berlin, 1976.

[BuC] Buzen P. and Chen P.S., "Optimal load balancing in memory hierarchies", *Information Proc.* **74**, pp. 271-275, 1974.

[Dan] Danet A., Despres R., LeRest A., Pichon G. and Ritzenthaler S., "The French public packet switching service: The TRANSPAC network", *Proc. of the* 3^{rd} *Int. Comput. and Comm. Conf.*, pp. 251-260, Toronto, Canada, 1976.

[Ell] Ellis C.A., "Consistency and correctness of duplicate database systems", *Proc. of the* 6^{th} *ACM Symposium on Operating System Principles,* pp. 67-84, West Lafayette, IN, November 1977.

[FeG] Federgruen A. and Groenevelt H., "The impact of the composition of the customer base in general queueing models", *J. Appl. Prob.* **24**, pp. 709-724, 1987.

[GeS] Gelenbe E. and Stafylopatis A., "Delay analysis of resequencing systems with partial ordering", *Proc. Performance'87*, pp. 433-446, Brussels, Belgium, December 1987.

[GJM] Gün L. and Jean-Marie A., "Resequencing in parallel queues with Bernoulli loading", submitted to *Operations Res.*, 1988.

[GMN] Gray J.P. and McNeil T.B., "SNA Multiple system networking", *IBM Systems Journal* **18**, pp. 263-279, 1979.

[Gün.a] Gün, L., "A note on the waiting and response times in the M/PH/1 queue", submitted to *O.R. Letters*, 1989.

[Gün.b] Gün L., *Tandem Queueing Systems Subject to Blocking With Phase Type Servers: Analytic Solutions and Approximations,* M.S. Thesis, Electrical Engineering Department, University of Maryland, College Park, MD, August 1986. Also available as *SRC Technical Report* **87-02**, Systems Research Center, University of Maryland, College Park, MD, 1987.

[HaP] Harrus G. and Plateau B., "Queueing analysis of a re-ordering issue", *IEEE Trans. on Software Engineering* **SE-8**, pp. 113-123, 1982.

[IL.a] Iliadis I. and Lien Y.C., "A generalization of scheduling policies to control resequencing delay", *Proc. IEEE GLOBECOM'87*, pp. 222-226, Tokyo, Japan, November 1987.

[IL.b] Iliadis I. and Lien Y.C., "Resequencing delay for a queueing system with two heterogenous servers under a threshold-type scheduling", *Proc. IEEE INFOCOM'87*, pp. 643-651, San Fransisco, CA, April 1987.

[IL.c] Iliadis I. and Lien Y.C., "Resequencing in distributed systems with multiple classes", *Proc. INFOCOM'88,* pp. 881-888, New Orleans, LA, March 1988.

[Ili] Iliadis I., *Resequencing Control and Analysis in Computer Networks*, Ph.D. Thesis, Columbia University, New York, NY, 1988.

[JeM.a] Jean-Marie A., "Re-routing and resequencing in multistage interconnection networks", *Proc. Int. Conf. on Parallel Proc.*, pp. 453-460, Chicago, IL, August, 1987.

[JeM.b] Jean-Marie A., "Load balancing in a system of two queues with resequencing", *Proc. Performance'87*, pp. 75-88, Brussels, Belgium, December 1987.

[JMG] Jean-Marie A. and Gün L., "Asymptotic results for parallel queues with resequencing", submitted to *J. ACM*, 1989.

[Kel] Kelly F.P., *Reversibility and Stochastic Networks,* J. Wiley & Sons, New York, NY, 1979.

[KKM] Kamoun F., Kleinrock L. and Muntz R., "Queueing analysis of the ordering issue in a distributed database concurrency control mechanism", *Proc. of the 2^{nd} Intl. Conf. on Distributed Computing Systems,* pp. 13-23, Versailles, France, April 1981.

[Kle.a] Kleinrock L., *Queueing Systems, Vol. I: Theory*, J. Wiley & Sons, New York, NY, 1975.

[Kle.b] Kleinrock L., *Queueing Systems, Vol. II: Computer Applications*, J. Wiley & Sons, New York, NY, 1976.

[KuC] Kushner H.J. and Clark D.S., *Stochastic Approximation Methods for Constraint and Unconstrained Systems*, Applied Mathematical Sciences Series **26**, Springer-Verlag, New York, NY, 1978.

[Sch] Schwartz M., *Telecommunication Networks: Protocols, Modeling and Analysis*, Addison-Wesley, Reading, MA, 1987.

[ShY] Shantikumar J.G. and Yao D.D., "Strong stochastic convexity and its applications in parametric optimization of queueing systems", *Proc. of the 27th Conference on Decision and Control*, pp. 657-662, Austin, TX, December 1988.

[SS.a] Shaked M.and Shantikumar J.G., "Stochastic convexity and its applications", *Adv. Appl. Prob.* **20**, pp. 427-446, 1988.

[SS.b] Shaked M. and Shantikumar J.G., "Convexity of a set of stochastically ordered random variables", to appear in *Adv. Appl. Prob.* **22**, March 1990.

[Sto] Stoyan D., *Comparison Methods for Queues and Other Stochastic Models*, Translation D.J. Daley, J. Wiley & Sons, New York, NY, 1983.

[Var.a] Varma S., *Some Problems in Queueing Systems with Resequencing*, SRC *Technical Report* **87-192**, Systems Research Center, University of Maryland, College Park, MD, 1987.

[Var.b] Varma S., "Optimal allocation of customers in a two server queue with resequencing", to appear in *IEEE Trans. on Auto. Control*, 1990.

[Web] Weber R.R., "A note on waiting times in single server queues", *Operations Res.* **31**, pp. 950-951, 1983.

[YuN] Yum T.S.P. and Ngai T.Y., "Resequencing of messages in communication networks", *IEEE Trans. Comm.* **COM-34**, pp. 143-149, 1986.