Fluctuation Analysis in Stochastic System with Parallel Queues

by

Ahmed I. Merie

Bachelor of Science
Statistics
Mosul University
2001

Master of Science
Statistics
Mosul University
2004

A dissertation
submitted to
Florida Institute of Technology
in partial fulfillment of the requirements
for the dgree of

Doctor of Philosophy
in
Operations Research

Melbourne, FL
July 2018

The undersigned committee hereby recommends that the attaches document be accept as fulfilling in part the requirements for the degree of

Doctor of Philosophy of Operations Research

"Fluctuation Analysis in Stochastic System with Parallel Queues"
a dissertation by Ahmed Merie

Committee:

_____
Jewgeni Dshalalow, Dr. Sci.
Professor, Department of Mathematical Sciences
Dissertation advisor.

_____
Kastro M. Hamed, Ph.D.
Professor and Head of Department of Education and Interdisciplinary Studies

_____
Munevver M. Subasi, Ph.D.
Associate Professor, Department of Mathematical Sciences

_____
Barry L. Webster, Ph.D.
Assistant Professor, Department of Engineering Systems

_____
Ugur Abdulla, Ph.D., Dr. Sci., Dr. rer. nat. habil.
Professor & Head of Department of Mathematical Sciences

Abstract

Fluctuation Analysis in Stochastic System with Parallel Queues

by

Ahmed I. Merie

Principal Advisor

Jewgeni H. Dshalalow, Dr. Sci.

In this dissertation, We analyze a complex queueing system with a single server operating in three different modes and, dependent on circumstances, servicing two different queues simultaneously. There are different switching policies that specify when the server takes one or two queues. Main techniques are based on fluctuation analysis. We study an enhanced hysteretic control system, with primary and secondary queues and random batch service. When the primary queue down-crosses r, the server operates on two parallel lines servicing them asynchronously until the primary line of remaining units is processed or the number of serviced secondary units is at least S which ever comes first. The server waits thereafter is the total quantity of primary units is less than N (In chapter II we assumed $N = 1$). The server capacity of primary units is limited by R, with two options: $r \leq R \leq N$ and $R > N$. Using fluctuation analysis we obtain closed-form distributions of available units during key periods of time and the steady state distribution of the primary units. We illustrate analytical tractability by numerous analytical and computational examples.

Table of Content

List of Keywords

Single-server queue systems

hysteresis

bivariate fluctuations

marked point processes

semi-regenerative process

fluctuation theory

stochastic games.

List of Abbrevations

PDF:	probability distribution function.
pdf:	probability dinsity function.
LST:	Laplace-Stieltjes transform.
r.v.:	random ariable.
a.s.:	almost surely.

List of Figures

Acknowledgment

I am very thankful to my adviser for his fundamental role in Jewgeni Dshalalow my doctoral work. provided me with every bit of guidance, Dr. Dshalalow assistance, and expertise that I needed during my first few semesters; then, when I felt ready to venture into research on my own and branch out into new research areas, he gave me feedback, advice, and encouragement. I am also grateful to the remaining members of my dissrtation committee: Dr. Subasi, Dr. Webster and Dr. Hamed, K.. I am deeply indebted to Dr. White, a good friend ever since, for all his help, advice, and encouragement. I would like to thank my friends in the department of Mathematics at Florida institute of technology for all the great times that we have shared. I am deeply thankful to my family for their love, support, and sacrifices. I am particularly thankful to My mother, and my kids Dalya, Sama and Shuaib. Without them, this Dissertation would never have been written. I dedicate this thesis to the memory of my father Idrees Merie, whose role in my life was, and remains, immense. This last word of acknowledgment I have saved for my dear wife Arwa Alsulaiman, who has been with me all these years and has made them the best years of my life.

Dedication

To my mother and memory of my father

CHAPTER I

INTRODUCTION

1.1. OVERVIEW

In our work, we study classes of stochastic systems equipped with two parallel queues and customers of two priorities. In his primary mode (mode 1), a single server processes batches of first priority customers, under the requirement that at least r customers are present in the queue. Otherwise, the server enters mode 2 turning to a secondary queue, however continuing servicing primary customers (if available), although in a slower pace. The server exits mode 2 whenever one of the two events takes place: the server is done with all remaining first-priority customers or the server processes at least a fixed number of second-priority jobs. There are eventually new arrivals of primary customers in the system during mode 2. If their cumulative number is at least $N(\geq r)$, the server takes a batch of customers between r and R (where R is server's capacity) and upon completing their service, he returns to mode 1. If the number of first-priority customers is lower than N, the server rests until that number is crossed by new arrivals. This period is referred to as mode 3.

The input stream of first-priority jobs is compound, meaning that jobs arrive at random in batches of random sizes and they line up in a buffer waiting for being processed. The server then chooses a batch of a random size between r and R for service or else proceeds to a complex cycle of modes where he most often deals with two queues simultaneously. To make our modeling more realistic, we assume

that the named queues are processed asynchronously. A further challenge comes from figuring out the time when the server is to stop working on the two queues, that is when all primary jobs are fully processed or the number of secondary jobs (that the server also processes in random batches) hits or exceeds some S. Notice that even when the number of secondary jobs crosses S at some point, the server still keeps on working on a primary job whose service may not be interrupted which adds to its analytical complexity.

This and similar situations often arise in computers or servers in which the operating system is programmed in such a way that whenever the number of primary jobs drops and the CPU intensity slows down, the system begins to operate on secondary tasks such as monitoring external threats to the system (such as spams, viruses, and malware), reorganizing hard drive sectors, cleaning the system registry, and backing up most sensitive files, to name a few. Most of these tasks are regarded as a routine maintenance of the system, and the server initiates this routine, while still working on high priority jobs, just in a slower pace. Since those different operating modes are preprogrammed (that accounts to a lion portion of an underlying operating system), due to the unpredictability of underlying processes, it is difficult to foresee all possible outcomes and make the system work optimally. We know that some upscale computers are more reliable, faster, and exhibit a better control. While it takes more efforts to employ better control modes, real challenges require a more advanced approach.

We study three different variants of the above system, all overlapping on mode 2 that we tackle by game-theoretical tools with two hostile players. Unlike a classical

antagonistic game of two players, in our situation, one player has an edge over the other player and as such he is given a slightly preferential treatment. The game-theoretical approach here is intertwined with fluctuation theory developed in Dshalalow [22] that we further embellish. Other tools we apply relate to Markov processes. In our case, due to its complexity, the queueing process is not Markov, but "on occasion" it exhibits Markov property, in fact the strong Markov property. It turns out that successive completions of so-called *service cycles* form a sequence of Markov times thereby making the queueing process *semi-regenerative*. A service cycle can be *simple* if it consists of a single service. Another type of the cycle is *complex* that includes three named modes. These moments are when our queueing process *conditionally regenerates*. That is, the future of the process that enters a particular state upon any such moment replicates itself. Throughout this thesis, we introduce and analyze such service cycles over which we then form an embedded Markov chain that turns out to be homogeneous, aperiodic and under a certain (necessary and sufficient) condition gets recurrent positive. We do it for all three models.

We deal with every segment of the cycles subject to different probabilistic analysis ranging from Laplace-Stieltjes transforms, fluctuation theory, game theory, to Markov processes. In the end, we focus on obtaining the results, such as steady state probability distribution of the named queueing process and various performance measures, in a closed form. We support our claim by numerous computational examples and special cases that validate the results and show their analytical tractability.

Brief Description of the special Model. In chapter II, we study a class of queues with a single server who occasionally works on two queues simultaneously. This is

a routine maintenance of an operating system that should not be confused with a vacation policy (well-known in queueing), because the server is still available in the main operating mode.

The queueing system consists of two queues. One of them is a queue of primary units that arrive in the system according to a marked Poisson process and get organized in an infinite-capacity buffer. Another queue is of units (in an unlimited quantity) that are placed and stored in a secondary buffer.

In the named paper [31], the server begins to operate on two queues simultaneously whenever the primary queue drops below a fixed threshold r. The server routinely takes batches of exactly r primary jobs at a time when it operates in so-called *mode* 1. Once the traffic of incoming jobs slows down, the server still takes primary units for service (if available), but singly and in different intensity. At the same time, the server works on secondary jobs in batches of random quantities. This was defined as *mode* 2. Mode 2 ends whenever a minimum of S secondary jobs is processed or when the server is done with the rest of the primary jobs (not counting new arrivals), whichever of the two events comes first. The work on a primary job is not interrupted even when the server has processed at least S secondary jobs, because the server has to complete this pending job. If upon his exit form mode 2, the primary buffer is not empty, the server will service one of the units. Otherwise, the server is waiting (mode3) for a new arriving batch of primary units before taking just one for service (mode4). This whole period with service was referred as *mode* 3. The end of mode 3 was the end of the service cycle.

The servicing facility includes a single server that in **mode 1** services batches of constant size r of primary customers if available. To avoid triviality, we assume that $r > 1$. Whenever the server has r or more primary customers in the system upon the release of a previous group, a simple *service cycle* keeps running and consists of mere service periods.

When the queue length drops below r, but with at least one primary customer in the queue, the server begins to work on two queues simultaneously. He continues servicing available primary customers one-by-one (according to a different service time distribution) and at the same time, he starts processing units (in random batches) from the secondary buffer. Because service of primary and secondary units are of different nature, server's work on two queues is asynchronous. This servicing mode is referred to as **mode 2a**. So mode 2a follows mode 1 if the queue of primary customers upon the end of a service drops below r, but not exhausted. However, if the queue is exhausted, the server enters **mode 2b** according to which he processes only secondary units dealing just with one queue. In a nutshell, we see two different versions of **mode 2**.

While in mode 2b, where service works on the secondary queue, we assume that there are always jobs available. We set up a limit of S jobs for the server to process before he can move to a next mode. Obviously, because the server processes batches of random sizes, at some point, their total number will rather exceed S than equal S sharp making S a relative threshold. The nearest time when it takes place (the *first passage time*) and the number of processed secondary units (the *first excess over* S) is not a routine task and it will be discussed in forthcoming sections.

The analysis of mode 2a is more complex, because the time when it ends will be stipulated by the following events. If the server is done with servicing all those primary units, whereas he has not finished servicing S secondary units yet, he nevertheless leaves mode 2a. If at some point of time the server is done with secondary units (hitting or exceeding S) while being in a middle of servicing one of the primary units, he does not exit mode 2a until he completes servicing that primary unit. At the same time, the server keeps on processing secondary units even in further excess of S. As soon as he is done with that primary unit, he then exits mode 2a. In this case, he may leave some primary units behind, whereas he has definitely served S or more secondary units, all in the spirit of a relative priority service. The figure below depicts servicing process during mode 2a displaying its two variants we spoke of.

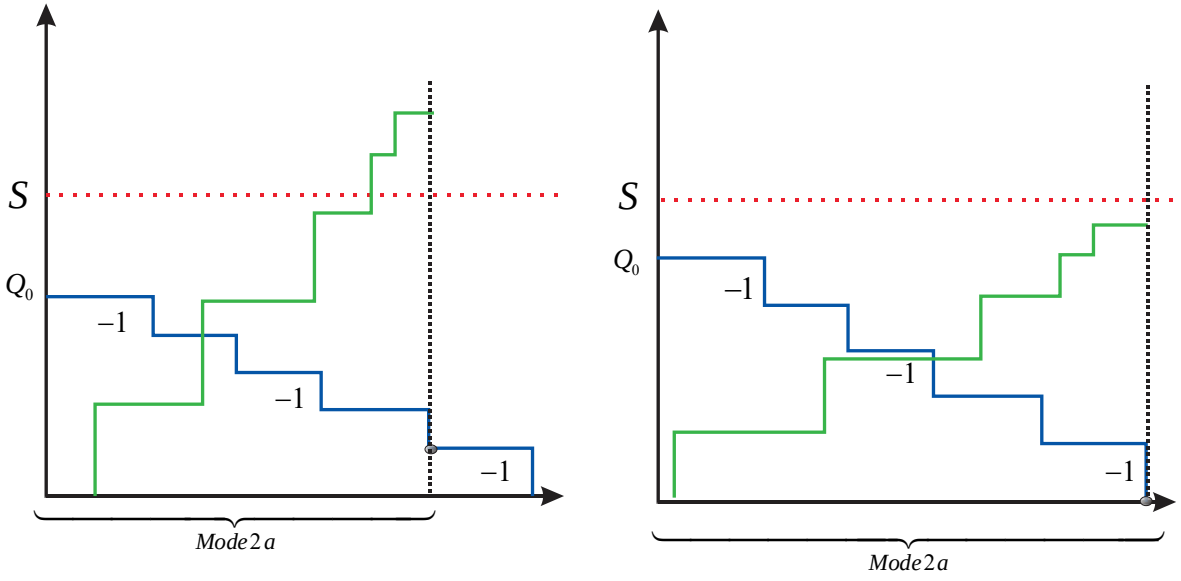


Figure 1-1 mode 2a

In figure 1-1 the green step line shows the number of processed secondary jobs, and the monotone non-decreasing blue step line records primary units and their departure times from the system. In the first figure, we see the server is done with secondary jobs first at some point of time. However, he was busy with a primary unit and he does not exit mode 2a until he is done with that unit and he keeps on working on secondary jobs until that time. The second figure depicts a situation when the server ends servicing primary units while not being done with secondary units. He nevertheless exits mode 2a. (Of course there is a chance that the server is done with primary units and by that time, the number of secondary units he processed that far also S or more.)

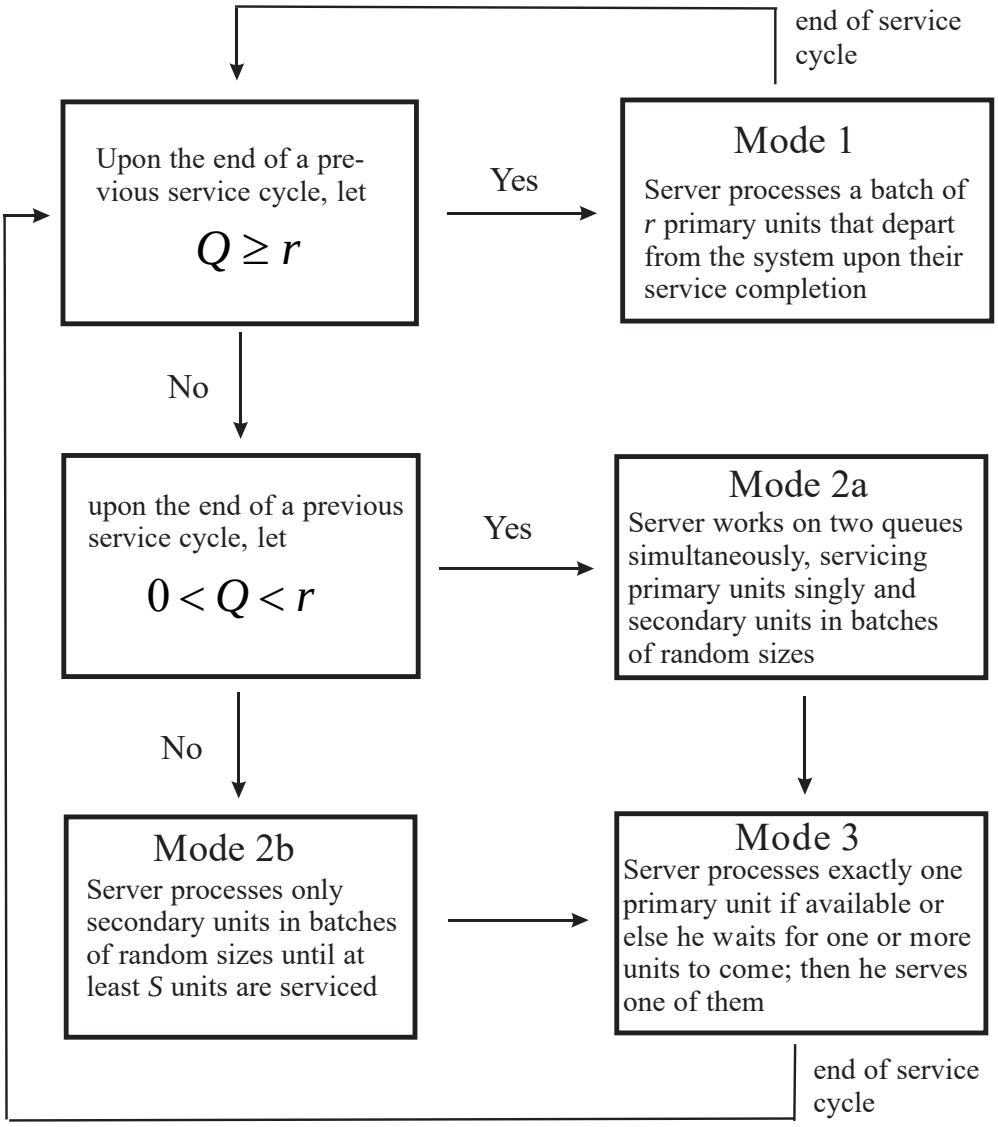
The information needed upon server's exit from mode 2a includes the *first passage time*, the number of primary units so far serviced or left behind, and the number of processed secondary units. The fluctuation analysis of this period will be discussed in a forthcoming section.

It is possible that after mode 2, there are some primary units from those left behind or/and those accumulated from new arrivals. Should that be the case, the server takes just one of them for service. If however, by the end of mode 2, there is no primary unit in the system, the server waits until a first batch of customers arrives, so that he picks out one of them and processes this customer. This closes the entire service cycle. It is convenient to interpret this period as **mode 3** and break it in two phases. During the first phase, the server waits for a first group of customers to arrive (if the primary buffer is empty). Phase 2 of **mode 3** consists of service of a single customer. If upon the end of mode 2, the buffer is not empty, the server picks

out one customer and starts servicing it immediately following mode 2. The second variant of mode 3 also consists of two phases, just the first phase is instantaneous.

Now after a service completion in mode 3, the server ends the entire service cycle passing through modes 2 and 3 followed by a new service cycle. Thus it is subject to the same rule applied to a simple service, that is, if by the end of a cycle, the queue length is r or greater, he resumes mode 1. Otherwise, he returns to modes 2 and 3. We note that not only such a policy is common in operating systems, but it allows to upgrade mode 3 implementing the classic N-Policy where in our present case $N = 1$. In our present model, servicing just one unit in mode 3 gives the system one more chance to raise the queue to r or more units in an attempt to reduce the frequency of passing to mode 2. With $N \geq r$, such an effort will be redundant. [32]

The following chart illustrates the process during one service cycle that starts with $Q(\geq 0)$ primary customers.



Firure 1-2

Such a variety and complexity of modes is paramount for flexibility of modeling computer operating systems and their systematic navigation. On the other hand, they present an analytical challenge and anxiety about tractability of the mathematical outcome. Maneuvering with thresholds offers new avenues for stochastic control that had previously not been explored. Consequently, a closed form (rather than an algorithm) is most desirable.

1.2. METHOLOGY

The main emphasis in chapter II is on the methods rather than modeling of the special model ($N = 1$). They relate to fluctuation theory of multivariate marked point processes with dependent marking and their behavior about critical thresholds. The literature on fluctuations [9, 10, 13, 19, 21, 24, 26, 48, 57] and their applications to queueing [23, 25, 28-30, 56] is very rich. Here the goal of our analysis is obtain analytically tractable results (rather than numerical) that can be better adapted to control problems. Whereas optimizing the control thresholds is beyond the scope of our current discussion, it definitely pertains to our forthcoming work, in which we plan to revisit this model and study an underlying continuous time parameter queueing process.

Related Past Work. **Comparison with Some Similar Systems**. Most articles on server vacations are to some extent precursors of our system, although unlike this work, they do not include specifics on what a server or servers are doing during their occasional maintenance (referred to as *vacations*) [23, 39, 41, 42, 50-52, 60, 63]. The literature on those systems where server's maintenance is closest to ours, to the best of our knowledge, is very narrow [4, 7, 8]. These papers include details on maintenance, on how *secondary jobs* are processed and whether this maintenance is limited by their number or by the time. Speaking of the time limitation, it pertains to the T-policy [40, 53].

Polling queues [11, 15, 47, 58, 61] have some resemblance with our model. It is typically assumed that a single server visits several queues in some order, very

often cyclically. Such systems find applications in computer networks and telecommunications, manufacturing and road traffic management. The term *polling system* goes back to 1957 to model a single repairman who serviced machines in the British cotton industry. However, in our model, not only does the server visits another queue, but in certain modes he also serves two queues simultaneously.

In Alzahrani and Dshalalow [8], the server leaves the main period, equivalent to our mode 1, when the queue is exhausted, and starts servicing a line of secondary jobs in batches of random size. He needs to process at least L units before he considers to resume mode 1. (Since the sizes of batches to be processed are random, the total quantity of secondary units will likely exceed L.) Further on, when the server is ready to resume his work on primary units and the queue by then accumulates to at least N customers, the underlying service cycle ends. Otherwise, the server returns to the secondary units and processes exactly one batch of them and he repeats doing this over and over again until the line of primary hits N or more of them for the first time. This is an N-Policy queueing system with two types of customers, which pertains to our present work. However, except for the element of the N-Policy, this model is more primitive than ours, because there the server never works on two lines simultaneously.

Unlike the paper by Alzahrani and Dshalalow, in Abolnikov et. al., [4] the authors worked on two lines simultaneously. This model, however, differs from ours by the way the multitasking mode operates. In Abolnikov et. at., the server suspends working on the queue (our mode 1), whenever the number of units in the primary queue drops below r (just the same as ours). Then he turns to the secondary jobs

and processes them in random batches until he is done with M or more of them. It resembles (but not identical to) our mode 2b. During the same time, the primary queue replenishes with new arrivals that at some point can raise the total number of jobs to L or more. If this event occurs during server's work on the secondary jobs, he works simultaneously on two queues (of primary and secondary jobs). In contrast, in our present paper, the server enters mode 2a with the primary buffer filled with $0 < i < r$ units and he works on two queues simultaneously, regardless on whether or not the new arrivals fill in the buffer to an L. This mode continues until the server is done with all secondary jobs or with all primary jobs he started off with, whichever of these two events comes first. If he is done with the secondary jobs first, the server keeps on processing the rest of the primary jobs. Now the shortcoming of this model is that during server's multitasking mode or single-tasking mode when he completes his work on the primary queue, there may eventually be new arrivals, but they are not counted and the work on them is assigned to a different mode, with a different service distribution. Not so in our present model, because when entering mode 2, the server does not bother taking care of any new arrivals and if the queue length was positive, he works on two queues simultaneously. This model seems to be better organized.

In Alghamdi and Dshalalow [7], the authors studied a queueing model in which a single server performs maintenance-type vacations working on secondary queues and taking random batches for service. As in Alzahrani and Dshalalow [8], there was no simultaneous work on two queues. However, when the server works on the secondary customers with a total maintenance time limited by a constant T, although he does not interrupt his service until he is done with a current batch.

Upon returning to the primary facility, with no customer waiting, the server rests until a first group of primary units arrives, thereby closing the service cycle. This system refers to as a T-policy system (originated in various articles the seventies), but it is far more complex.

Dshalalow, Kim, and Tadj, in [30], introduced a bilevel hysteretic control discipline that includes two control thresholds r_1 and r_2 when the server leaves mode 1 for vacations and N-Policy by which the server returns to mode 1. Dependent on how many units are left behind, the server leaves on single or multiple vacations. No secondary queue was mentioned. All three thresholds $r_1, r_2,$ and $N,$ are subject to optimal control rendered in the paper. This paper also relates to Abolnikov, Dshalalow, and Treerattrakoon [3]. Other hysteretic control queues were studied in Dshalalow and Dikong [28, 29].

Main Results and Layout of chapter II. In sections 2.1, we formalize the model followed by key results on fluctuation theory pertaining to our system in sections 2.2-2.4. Each of these latter sections deals with modes 2 and 3 separately providing us with explicit formulas of underlying functionals and examples and special cases confirming their closed forms claims. Section 2.5 yields a Kendall's type formula for the stationary distribution of the embedded queueing process.

1.3. BRIEF DESCRIPTION OF THE SYSTEM AND COMPARISON WITH A PREVIOUS MODEL

In chapter III, we continue our work started in [31] (Chapter II) on a queueing system where a server on occasion processes (asynchronously) two different queues. The general model N-policy [32] underwent several significant modifications and embellishments. Firstly, we allow the server to take any number of primary units in one servicing batch from r to R available in the buffer. So that R is the server's capacity, not rigidly r. However, the server does not take less than r units in a batch. This policy alone is known as the r-R-quorum [32].

Now, with the buffer down-crossing r, the server initiates mode 2 as in [32] under similar specifications. However, upon exiting from mode 2, unlike the assumptions in chapter II, it is now required a minimum of N primary units available to have the server resume his work on the primary queue from which at least r are to be taken. Here we consider two different models: (1) with $r \leq R \leq N$ and (2) $r \leq N < R$, that are to be treated separately. If upon server's exit from mode 2, and entering *mode* 3, the primary buffer content is below N, the server must wait until it replenishes to at least N or more units. Because the input is not ordinary, the time of the resumption of service and the number of units at this time are yet to be determined using fluctuation analysis. Furthermore, we have to take into account the initial number of primary units at the beginning of mode 3 which can be greater than or equal to N or less than N. Then the server takes a $\min\{R, Q\}$ units for service to end an underlying service cycle, where Q is the queue length of primary units by the end of mode 3 [32].

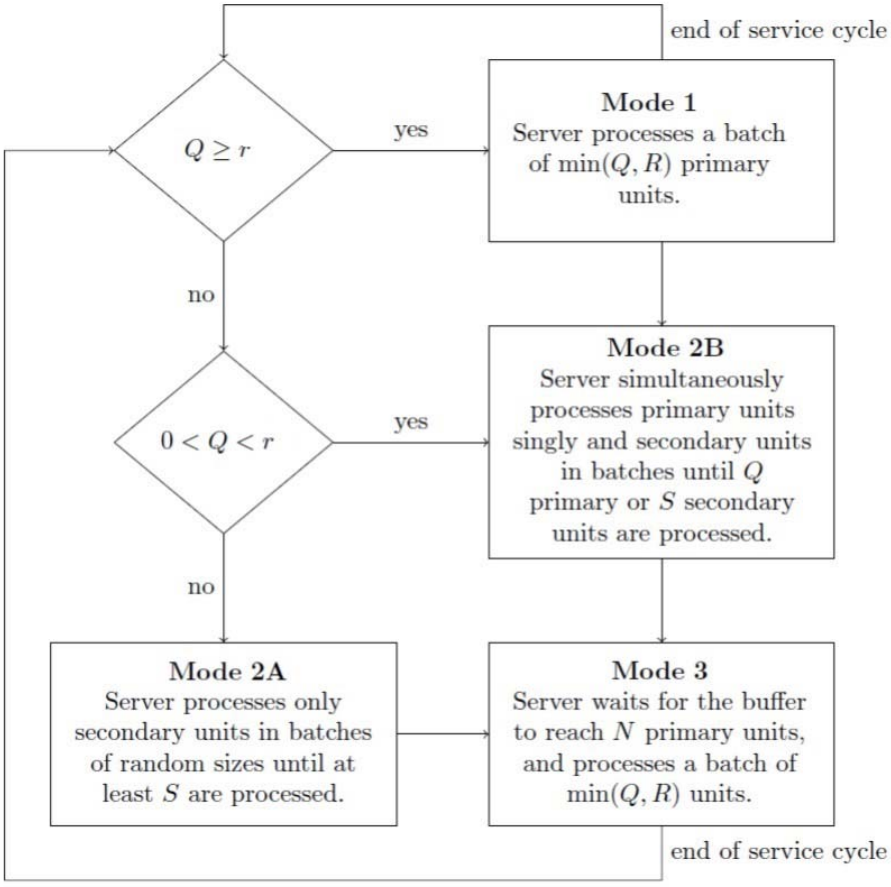


Figure 1-3 The following chart illustrates the process during one service cycle that starts with $Q(\geq 0)$ primary customers.

With the three parameters $r, R, N,$ intertwined as per the above settings, the system is under the *hysteretic control policy*. The different from chapter II is considering that the input is not ordinary and for that reason alone, the tools of fluctuation analysis (previously developed by the first author) become very vital. To be more specific, upon exiting from mode 2 and going after N (or more) primary units requires a more complex chaining of fluctuation functionals. The latter, in turn, requires a further evidence of analytical tractability far beyond what

was rendered in chapter II [31]. To address that, we expand on various examples, special cases, and simulation to support our claim.

Hysteresis is one of the common forms systems with *state dependent parameters*. (See a survey in Dshalalow [22].) In its standard scenario, hysteresis assumes that it takes the queue to down-cross threshold r to alter its conventional mode, and it takes the queue to up-cross another threshold N to return to that mode, under $r \leq N$, cf., Bekker [12], Dikong and Dshalalow [18], Dshalalow [23], Dshalalow and Dikong [28], Dudin and Chakravarthy [33], Ke [43], Loris-Teghem [45], Tadj and Ke [53], and Teghem [59]. Another type of hysteretic control goes by changing the number of active servers in multiserver queues, as in Kim et al. [44]. Besides the above mentioned applications of hysteretic control in computer servers (see also Golubchik [38], Chan et al. [16]), hysteretic control is often implemented in telecommunications (cf. Dudin and Nishimura [34], Semenova [49], Ait-Salaht and Castel-Taleb [6], Gaidamaka et al. [36], and a very comprehensive survey in Vishnevskii and Dudin [62]), some of which are related to Internet telephony for voice and video calls, in private Internet Protocol (IP) telephone systems and instant messaging over IP networks, Pechinkin and Razumchik [46] and Zhennovyi and Zhennovyi [63]. In particular, we can see hysteretic control occurring in applications related to the Session Initiation Protocol (or SIP), which is a communication protocol for signaling and controlling multimedia communication sessions. (See [Abaev et al. [1-2].) Some applications of hysteretic control relate to cellular systems, see Choi and Sohrabi [17]. Further common applications of hysteretic control occur in production-inventory models; see an interesting paper by Boxma et al. [14].

In [31] , we implement fluctuation techniques that are a key engine in our analysis. However, due to the above augmentations, fluctuation analysis in this paper is more complex compared to that in [31]. Next we consider the embedded queueing process upon the ends of *consecutive service cycles*, which are stopping times with respect to the filtration induced by the queueing process. A service cycle can be limited to just a service time of a batch of customers while in mode 1 or it extends by the length of time from the beginning of mode 2 to the end of mode 3, followed by the service time. The service times of primary units are generally distributed, but being of different kinds throughout modes 1, 2 and 3. More on these are in the forthcoming sections [32].

1.4. APPROACH AND OTHER RELEVANT MODELS

As stated, our major innovative tools relate to the theory of fluctuations, with some general work of [13, 19-21, 27, 56] and, in particular, as applied to queueing [4, 5, 18, 23, 28, 30, 31]. The goal of these methods is to obtain closed-form expressions rather than rely on algorithms, because the former are better suited for control problems. Typical fluctuation analysis can be outlined as a probabilistic problem of finding the time and location of a randomly walking particle on a multi-dimensional random grid that tries to escape from a compact set. Because the particle is not continuously moving over the grid, the particle is not just crossing the boundary of that set, but rather jumping off the set, thereby making the problem of finding the first passage time and the first location of the particle outside the set more challenging.

Applied to our system [32], the particle's motion is represented by the server asynchronously processing two queues (two-dimensional case) during mode 2b and then crossing one of the two thresholds assigned for either queue. The escape time and location of the particle (the queue status) is then chained with the new time and location of the particle escaping from a second set. This is when in mode 3, the queue crosses threshold N. Furthermore, in our case, the components of the particle walk are competing against each other.

From the modeling point of view, the system of chapter III relates to systems with vacations in chapter II [31], although very remotely, because in those systems, a server or servers are just vacant during some specified periods of time. There are a few papers like [4, 31], our recent work included, where server's absence is caused by an additional work integrated into the system and a continual work on the primary units at a slower pace, that pertain to our present paper. A complete survey of other related models and models with vacations can be found in [31] where we carried a detailed analysis on the analogy and differences between them.

The chapter III model has also some resemblance with polling queues [14, 32, 58, 61] where it is assumed that a server attains to several queues in some order, often cyclically. Such systems find applications in computer networks and telecommunications, manufacturing and road traffic management. (See a further discussion in [31].)

In chapter III section 3.1, we formalize the model followed by pertinent results from fluctuation theory. Section 3.2, 3.3 and 3.4 deal with modes 2 and 3 separately

providing us with explicit formulas of underlying functionals as well as examples and special numerical cases confirming their closed forms claims. Section 3.5 Section 7 yields a Kendall's-type formula for the stationary distribution of the embedded queueing process treating two analytically distinct models under the assumptions that $r \leq R \leq N$ and $r \leq N < R$. Section 8 deals with the mean value of the stationary service cycle.

CHAPTER 2

FLUCTUATION ANALYSIS IN QUEUES WITH SEVERAL OPERATIONAL MODES AND PRIORITY CUSTOMERS

2.1. FORMAL DESCRIPTION OF THE SYSTEM

Assume that all random variables (r.v.'s) and processes are considered on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$. (A prototype of our system is M/G/1, but our system is far more complex.)

Input. We assume that the input process to the system is a stationary marked Poisson

$$\mathcal{I} = \sum_{k=1}^{\infty} \xi_k \varepsilon_{t_k} \quad (\varepsilon_a \text{ is a Dirac mass at } a) \tag{2.1.1}$$

with position independent marking, the associated support counting measure $\mathcal{S}_I = \sum_{k=1}^{\infty} \varepsilon_{t_k}$ of intensity $E\mathcal{S}_I = \lambda|\cdot|$ ($|\cdot|$ is the Borel-Lebesgue measure, so that $\lambda = E\mathcal{S}_I[0,1]$), the arriving groups of units $\xi_1, \xi_2, \ldots \in [\xi]$ (the P-equivalence class of r.v.'s) with a common pgf (probability generating function) $a(z) = Ez^{\xi}$.

Queueing Process. We denote $Q(t)$ the number of all primary units in the system at time $t \geq 0$. We define $Q(t)$ as a piecewise linear process with right-continuous paths adapted to $(\mathcal{F}_t)$.

Service Cycle. There are three distinct processing modes administered during service cycles. *Service cycles* form a point process

$$\mathcal{C} = \sum_{n=0}^{\infty} \varepsilon_{T_n}, \tag{2.1.2}$$

where $T_0 = 0,$ under the following specifications. For the notational convenience, we formalize the modes during the first service cycle. Suppose that at time $T_0 = 0,$ the number of primary units $Q_0 = Q(T_0) = Q(T_0 +) \geq r,$ where r is assumed to be greater than one. Then the server takes a batch of r units and processes all of them at once during a time σ_1 after which all those r units depart from the system at time $T_1 = T_0 + \sigma_1$. The first cycle then ends at time T_1. We say that the server operates in ***mode* 1**. The r.v. $\sigma_1 \in [\sigma],$ where $[\sigma]$ is an equivalence class of all r.v.'s with a common probability distribution P_σ and the Laplace-Stieltjes transform (LST) $\beta(\theta) = Ee^{-\theta\sigma}$.

Now suppose that at time $T_0,$ the number of primary units Q_0 is less than r. If $Q_0 > 0,$ the server operates simultaneously on two different facilities. He begins processing the queue of Q_0 jobs singly, one-by-one, and at the same time, a secondary line of jobs, in batches of random sizes. We assume that the total number of second-priority jobs is unlimited. This process lasts until one of the two events takes place. All of Q_0 units are serviced or at least S secondary units are serviced, whichever comes first. There is a further condition imposed on service of primary customers. When the server has processed S or more units and still in a middle of servicing a primary unit, he does not leave the mode until he is done with that primary unit. We further assume that the server still keeps on servicing secondary

units until that moment. While it definitely makes the analysis of such system more challenging, this configuration is most practical when programming realistic operating systems. We further notice that because underlying batches of secondary jobs are of random sizes, their total numbers serviced rather exceeds than equals S. We will formalize the game-theoretical aspect of this process during ***mode 2a*** and utilize and further embellish fluctuation analysis to address the complexity of this situation.

Therefore, by the end of mode 2a, the primary buffer includes primary units that could have been left behind from mode 2a and further replenished with new arrivals. This number however, can be zero, and in the case, the server waits for at least one arriving batch and then picking out one for service. A period, from the completion of mode 2a, that can include waiting, until the end of service, is referred to as ***mode* 3**. With a non-empty primary buffer, the server immediately takes one of the available units for service, herewith reducing mode 3 to just a service time. It is convenient to break mode 3 in two phases: waiting phase (1 that can also be instantaneous) and servicing phase (2).

We note that such a service cycle includes a total of two modes: mode 2a and mode 3. T_1 is the end of mode 3 and a service cycle.

Now if the number of primary units Q_0 upon the beginning of the first service cycle is zero, the server processes only secondary jobs and when their number crosses S, the server leaves ***mode* 2b** and moves to mode 3. This version of mode 3 carries the same protocol as that preceded by mode 2a.

In a nutshell, a service cycle can be a simple one that includes only service of r primary units of mode 1, or it can be a compound cycle furnished with modes 2 and 3. By the end of a service cycle, the server moves to a next cycle under the same specifications. We will argue that the completions of service cycles $\{T_n\}$ form a sequence of stopping times relative to the filtration $(\mathcal{F}_t)$.

Game-Theoretic Aspect of Mode 2a. With $0 < Q_0 < r$, the server starts processing the primary units from that queue one-by-one specified by the following marked point process

$$(\mathcal{A}, \tau) = \sum_{k=1}^{\infty} X_k \varepsilon_{\tau_k}, \tag{2.1.3}$$

where $\tau_1, \tau_2, \ldots$ are successive completions of service of primary units during mode 2a. Even though we let $X_k = 1$ for all k's, we prefer to keep them listed in $(\mathcal{A}, \tau)$. The server also works on the secondary queue of jobs processing them according to the marked point process

$$(\mathcal{V}, \mathcal{S}) = \sum_{j=1}^{\infty} v_j \varepsilon_{s_j}, \tag{2.1.4}$$

where $s_1, s_2, \ldots$ are successive completions of service of the secondary units during this mode and $v_1, v_2, \ldots$ are respective random sizes of batches. It is reasonable to assume that the processes $(\mathcal{A}, \tau)$ and $(\mathcal{V}, S)$ are independent.

Note that with the common notion of random measures representing the above point processes, $(\mathcal{A},\tau)(B)$ and $(\mathcal{V},\mathcal{S})(B)$ give the number of primary and secondary jobs performed during a time set B if B is any Borel subset of $\mathbb{R}_+$. Having this in mind, due to the priority discipline imposed on the exit from mode 2a (in favor of primary jobs),

$$Y_1 = (\mathcal{V},\mathcal{S})[0,\tau_1],\ Y_2 = (\mathcal{V},S)(\tau_1,\tau_2],\dots \tag{2.1.5}$$

are the increments of secondary units processed between the epochs $\tau_1,\tau_2,\dots$ of successive service completions introduced in (2.1.3). This forms a new bivariate marked point process

$$(\mathcal{A},\mathcal{B},\tau) = \textstyle\sum_{k=1}^{\infty}(X_k,Y_k)\varepsilon_{\tau_k} \tag{2.1.6}$$

(embedded in $(\mathcal{A},\tau)\otimes(\mathcal{V},\mathcal{S})$) making components $X,Y,$ and τ mutually dependent. Again, Y_k is the number of secondary units processed in interval $(\tau_{k-1},\tau_k]$.

Now we introduce the pair

$$\begin{aligned}\nu_1 &= \inf\{n\geq 1 : A_n = X_1+\dots+X_n = i\},\\ \nu_2 &= \inf\{n\geq 1 : B_n = Y_1+\dots+Y_n \geq S\},\end{aligned} \tag{2.1.7}$$

of the random indices and the index

$$\nu := \nu_1 \wedge \nu_2 \tag{2.1.8}$$

referred to as the *exit index*. The r.v. τ_ν is called the *first passage time* (named so in the theory of fluctuations). Thus, A_ν and B_ν are the quantities of primary and secondary units, respectively, processed by the completion of mode 2a at time τ_ν.

Consequently, upon the end of mode 2a, there are $Q_0 - A_\nu$ primary units that remain in the system from the total of Q_0 units upon the beginning of mode 2a. B_ν is the number of secondary units processed by τ_ν. Of course, the queue of $Q_0 - A_\nu$ primary units (possibly > 0) will be joined by new arrivals during the interval $[0, \tau_\nu]$. This will be yet another random component by the end of mode 2a to worry about.

The process $(\mathcal{A}, \mathcal{B}, \tau)$ of (2.1.6) does not run indefinitely, but it is terminated according to game-theoretic principles, namely, with some two generic players A and B fighting or competing against each other until one of them is defeated at time τ_ν. Namely, we assume that players B and A hit each other upon random times $\tau_1, \tau_2, \ldots,$ exerting respective random damages $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots.$ A and B can sustain Q_0 and S amount of casualties, respectively, after which one of the players is defeated. That is, the game ends at τ_ν with one or both players ruined. In reality, the game is a little more complicated, because player A exerts damages to player B upon times $s_1, s_2, \ldots,$ whereas the casualties $Y_1, Y_2, \ldots$ to player B are those observed upon times $\tau_1, \tau_2, \ldots$ and thus this information is delayed compared to the real time events. The game can be more precisely formalized by the process

$$(\mathcal{A}, \mathcal{B}, \tau)_{\nu} = \sum_{k=1}^{\nu} (X_k, Y_k)\varepsilon_{\tau_k} \tag{2.1.9}$$

terminated at τ_ν, with A_ν and B_ν being predicted cumulative casualties to players A and B, respectively, upon the end of the game at τ_ν.

Mode 2b. Compared to mode 2a, mode 2b is far simpler, but it is still a subject of some formalities of fluctuation theory. In the event mode 2 begins with zero primary units in the system, the server processes only secondary units under the conditions specified in (2.1.4), namely

$$(\mathcal{V}, \mathcal{S}) = \sum_{j=1}^{\infty} v_j \varepsilon_{s_j}.$$

However, it ends as soon as the number of processed secondary units crosses S (that is equal or greater than S at some point s_k). With the exit index

$$\rho = \inf\{n \geq 1 : V_n = v_1 + \ldots + v_n \geq S\},$$

the r.v. s_ρ is the first passage time at which mode 2b ends with V_ρ being the total number of secondary units processed by the end of mode 2b. Thus, the terminated version of (2.1.4) reads

$$(\mathcal{V}, \mathcal{S})_{\rho} = \sum_{j=1}^{\rho} v_j \varepsilon_{s_j}.$$

We will deal with the number of primary units that enter the system during mode 2b later on. Unlike the game-theoretic situation in mode 2a, here the exit takes place in real time.

Mode 3. If the number of primary units in the system by the end of mode 2 (versions a or b) is zero, the server waits for one batch of customers to arrive to begin his service being phase 1 of mode 3. His waiting time is obviously exponential with parameter λ (due to the nature of input). In the event the number of primary units is positive, we let phase 1 of mode 3 last zero time. The situation upon the end of phase 1 is not trivial, because upon that epoch of time, the system inherits an unspecified number of primary units somehow accumulated after mode 2. This is a separate problem that we address in the next section. From the quantity of primary units after phase 1, the server picks out one and processes it some random time Σ. This is a second phase of mode 3 that closes the service cycle.

From the above formalism of the first service cycle, the queueing process $Q(t)$ is semi-regenerative with respect to the sequence $\{T_n\}$ of cycle completions being stopping times relative to the filtration $(\mathcal{F}_t)$. The embedded process $\{Q_n = Q(T_n)\}$ is thus a homogeneous Markov chain to be treated in section 2.5.

2.2. FLUCTUATIONS OF MULTIVARIATE PROCESSES IN THE CONTEXT OF MODE 2A

Pertinent to the forthcoming investigation will be the result [26] by Dshalalow on multivariate marked point processes in a generic form. (See also Dshalalow [21].) Let

$$(\mathcal{A}, \mathcal{B}, \mathcal{P}, \tau) = \sum_{k=0}^{\infty} (X_k, Y_k, P_k)\varepsilon_{\tau_k} \tag{2.2.1}$$

be a marked delayed renewal process with position dependent marking and the mutually dependent marks (X_k, Y_k, P_k) and ε_α being a point mass at $a \in \mathbb{R}$. Here $\sum_{k=0}^{\infty} \varepsilon_{\tau_k}$ is the underlying support point renewal process on the positive real axis. We assume that the tri-variate marks (X_k, Y_k, P_k) form a sequence of conditionally independent vectors valued in $\mathbb{N}_0^3 = \mathbb{N}_0 \times \mathbb{N}_0 \times \mathbb{N}_0$. More specifically, $(X_k, Y_k, P_k, \Delta_k)$ for $k = 0, 1, \ldots,$ are independent $(\Delta_k = \tau_k - \tau_{k-1}, \tau_{-1} = 0,$ valued in $\mathbb{R}_+)$ and for $k = 1, 2, \ldots$ are identically distributed with common joint transforms

$$\gamma_0(u, v, w, \theta) = Eu^{X_0}v^{Y_0}w^{P_0}e^{-\theta\tau_0},\ \gamma(u, v, w, \theta) = Eu^{X_1}v^{Y_1}w^{P_1}e^{-\theta\Delta_1}, \tag{2.2.2}$$
$$|u| \leq 1, |v| \leq 1, |w| \leq 1, \mathrm{Re}\theta \geq 0.$$

Of the four components $(\mathcal{A}, \mathcal{B}, \mathcal{P}, \tau)$, we identify the first two as *active* and the rest as *passive components*. The active components will appreciate upon the times $\tau_0, \tau_1, \tau_2, \ldots$ until one of them crosses its respective specified threshold. These components can be interpreted as casualties to two mutually antagonistic players A and B. The passive components represent a neutral player P who does not take part in the game, but still sustain damages upon the times τ. The values of all other components will be recorded. More formally, let

$$
\begin{aligned}
&A_n = X_0 + \ldots + X_n \text{ (active)} \\
&B_n = Y_0 + \ldots + Y_n \text{ (active)} \\
&\Pi_n = P_0 + \ldots + P_n \text{ (passive)} \\
&n = 0, 1, \ldots
\end{aligned} \tag{2.2.3}
$$

Define,

$$
\nu := \min\Big\{\nu_1 = \inf\{m \geq 0 : A_m \geq M\}, \nu_2 = \inf\{n \geq 0 : B_n \geq N\}\Big\} \tag{2.2.4}
$$

and call it the *exit index*. Then, the r.v.'s $\tau_\nu, A_\nu, B_\nu, \Pi_\nu$ are of interest. τ_ν is referred to as the *first passage time*. One of the two random values A_ν or B_ν will cross fixed specified thresholds R or S, respectively. We will focus the joint transform

$$
\Phi(u, v, w, \theta) = Eu^{A_\nu} v^{B_\nu} w^{\Pi_\nu} e^{-\theta\tau_\nu}, |u| \leq 1, |v| \leq 1, |w| \leq 1, \text{Re}\theta \geq 0. \tag{2.2.5}
$$

In light of Theorem 3.1 below, let us define the following operator $\mathcal{D}$ applied to a function $\varphi : \mathbb{C}^3 \to \mathbb{C}$ analytic at $x = y = 0$ as

$$
\begin{aligned}
&\mathcal{D}_{x,y}^{k,m}\varphi(x, y, z) \\
&= \begin{cases} \lim_{x\to 0, y\to 0} \frac{1}{k!m!} \frac{\partial^{k+m}}{\partial x^k \partial y^m} \Big[\frac{1}{(1-x)(1-y)}\varphi(x, y, z)\Big], & k, m \geq 0 \\ 0, & k < 0 \text{ or } m < 0. \end{cases}
\end{aligned} \tag{2.2.6}
$$

It is called the *$\mathcal{D}$-operator*.

Theorem 2.2.1. (Dshalalow [21, 26].) *The functional Φ satisfies the following formula*:

$$\Phi(u,v,w,\theta) = \gamma_0(u,v,w,\theta)$$

$$- \mathcal{D}_{x,y}^{R-1,\,S-1}\left[\gamma_0(u,v,w,\theta)\Big(1 - \tfrac{\gamma(u,v,w,\theta)-\gamma(ux,vy,w,\theta)}{1-\gamma(ux,vy,w,\theta)}\Big)\right]. \tag{2.2.7}$$

In particular, for the process without a delay, that is, with $X_0 = Y_0 = P_0 = \tau_0 = 0$ a.s., (2.2.7) *reduces to*

$$\Phi(u,v,w,\theta) = 1 - [1-\gamma(u,v,w,\theta)]\mathcal{D}_{x,y}^{R-1,\,S-1}\tfrac{1}{1-\gamma(ux,vy,w,\theta)}. \tag{2.2.8}$$

□

In the context of our model during a period of mode 2a, formula (2.2.8) will be modified as follows. We recall that when entering mode 2a, the primary buffer "inherits" $Q_0 = i$ units $(0 < i < r)$ that will be the initial quantity of the primary units, whereas $\tau_0 = Y_0 = P_0 = 0$ by our setting. However, to apply Theorem 1 we set $X_0 = 0,$ whereas i is identified as the threshold R. In this case, the servicing process during mode 2a is monotone increasing instead of being decreasing. This makes

$$\gamma_0(u,v,w,\theta) = 1.$$

Remark 2.2.1. As per section 2.1, those primary units will be served one-by-one (thereby letting $X_1 = X_2 + \ldots = 1$ a.s.) until all of them are processed or until the

total number of the secondary units served in batches $Y_1, Y_2, \ldots$ hits S, whichever of the two events takes place first. We recall that all new arrivals of primary units that occur during mode 2a are not served during that period. Furthermore, we note the following.

(a) The primary and secondary units in mode 2a are served at different times asynchronously and their service times are differently distributed;

(b) Since the primary units during mode 2a have a higher priority than the secondary units, their service is not interrupted. For example, if at some point, the number of secondary units has reached or even exceeded S for the first time, while a primary unit was being served, the latter service will not be terminated, but sustained until it is completed. Thus, by the time the server is done with that unit, he kept on servicing secondary units in further excess of S. The total number of the secondary units processed by the end of service of that primary unit will then be recorded and assume value B_ν at τ_ν. Not all of the i primary units may be processed by τ_ν. The total number of them having left the system by the time τ_ν is A_ν.

(c) We assume that the service times $\Delta_1, \Delta_2, \ldots$ $(\Delta_k = |(\tau_{k-1}, \tau_k]|)$ of the primary units during mode 2a is general under the Laplace-Stieltjes transform $\gamma(\theta) = Ee^{-\theta\Delta}$, where $\Delta_1, \Delta_2, \ldots$ are independent and identically distributed r.v.'s from an equivalence class $[\Delta]$ (different from $[\sigma]$). Furthermore, $\tau_n = \Delta_0 + \Delta_1 + \ldots + \Delta_n$, with $\Delta_0 = 0$. The Δ's represent one of the two passive components.

(d) The services times $s_1, s_2 - s_1, s_3 - s_2 \ldots$ of the respective batches $v_1, v_2, v_3, \ldots$ of the secondary units are also independent and they are independent of $\Delta_1, \Delta_2, \ldots$ and as per section 2, $s_1, s_2 - s_1, \ldots \in [Exp(\mu)]$. We can say that

$(\mathcal{V}, \mathcal{S}) = \sum_{j=1}^{\infty} v_j \varepsilon_{s_j}$ of (2.1.4) is a marked Poisson process with position independent marking and unit intensity μ. The pgf of the batch size v_1 is $b(z) = Ez^{v_1}$, $|z| \leq 1$.

(e) New primary units enter the system during mode 2a according to the marked Poisson process $\mathcal{I}$ of (2.1) (with position independent marking) of unit intensity λ; the respective batches arrive in sizes of $\xi_1, \xi_2, \ldots$ under the common pgf (probability generating function) $a(u) = Eu^{\xi_1}$. □

In a nutshell, upon the time τ_ν, the number of the primary units will be A_ν and of the secondary units - B_ν. $A_\nu \leq i$, while the number B_ν can be less than S or greater than or equal to S, dependent on what the server manages to complete first. Not only can the value B_ν exceed S, but it can be arbitrarily greater than S possibly exceeding it by the sizes of respective batches of secondary units processed after threshold S was exceeded. Furthermore, the newly arrived primary units are placed in the queue, and they will be processed in the order of their arrivals during mode 3. They will represent another passive component P. Here is how we define $P_1, P_2, \ldots$ $(P_0 = 0)$. We let P_k equal the number of batches of the new primary units that enter the system during the time interval $(\tau_{k-1}, \tau_k]$ of the service duration $\Delta_k = |(\tau_{k-1}, \tau_k]|$ of the kth primary unit (from the group of the i remaining in the system upon the beginning of mode 2a).

Under the conditions specified in remark 3.1, we can show that

$$\gamma(u, v, w, \theta) = u\gamma[\theta + \lambda - \lambda a(w) + \mu - \mu b(v)]. \tag{2.2.9}$$

Indeed,

$$\gamma(u,v,w,\theta) = Eu^{X_1}v^{Y_1}w^{P_1}e^{-\theta\Delta_1} = E\big[E\big[u^1v^{Y_1}w^{P_1}e^{-\theta\Delta_1}|\Delta_1\big]\big]$$

$$= uE\big[e^{-\theta\Delta_1}E\big[v^{Y_1}|\Delta_1\big]E\big[w^{P_1}|\Delta_1\big]\big] = uE\big[e^{-\theta\Delta_1}e^{\lambda[a(w)-1]}e^{\mu[b(v)-1]}\big]$$

$$= u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(v)].$$

Consequently, the status of the system upon the end of mode 2a is determined by the functional

$$\Phi(u,v,w,\theta) = 1-[1-\gamma(u,v,w,\theta)]\mathcal{D}_{x,y}^{i-1,\,S-1}\frac{1}{1-\gamma(ux,vy,w,\theta)}$$

$$= 1-[1-u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(v)]]$$

$$\times\,\mathcal{D}_{x,y}^{i-1,\,S-1}\frac{1}{1-ux\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]} \tag{2.2.10}$$

as per Theorem 1. To simplify formula (2.2.10), we note that from (2.2.6),

$$\mathcal{D}_{x,y}^{k,m} = \mathcal{D}_x^k\circ\mathcal{D}_y^m = \mathcal{D}_y^m\circ\mathcal{D}_x^k, \tag{2.2.11}$$

where

$$\mathcal{D}_x^k\varphi(x,y,z) = \begin{cases} \lim_{x\to 0}\frac{1}{k!}\frac{\partial^k}{\partial x^k}\big[\frac{1}{1-x}\varphi(x,y,z)\big], & k\geq 0 \\ 0, & k<0. \end{cases} \tag{2.2.11a}$$

From (2.2.11-2.2.11a) and Dshalalow [17],

$$\mathcal{D}_x^{i-1}\frac{1}{1-ux\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}=\frac{1-u^i\gamma^i[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}{1-u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}$$

herewith making (2.2.10) reduce to

$$\Phi(u,v,w,\theta)=Eu^{A_\nu}v^{B_\nu}w^{\Pi_\nu}e^{-\theta\tau_\nu}$$

$$=1-[1-u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(v)]]$$

$$\times\mathcal{D}_y^{S-1}\frac{1-u^i\gamma^i[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}{1-u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}. \tag{2.2.12}$$

Because for now we are not interested in the quantity of the secondary units B_ν processed by the end of mode 2a, we can suppress it by letting $v=1$. At this point we can also leave the first passage time out of consideration by letting $\theta=0$. These assumptions have Φ reduce to what we denote by $h(u,w)$,

$$h(u,w):=Eu^{A_\nu}w^{\Pi_\nu}=\Phi(u,1,w,0)$$

$$=1-[1-u\gamma[\lambda-\lambda a(w)+\mu-\mu b(y)]]$$

$$\times\mathcal{D}_y^{S-1}\frac{1-u^i\gamma^i[\lambda-\lambda a(w)+\mu-\mu b(y)]}{1-u\gamma[\lambda-\lambda a(w)+\mu-\mu b(y)]}. \tag{2.2.13}$$

Understandably, $Eu^{A_\nu}w^{\Pi_\nu}$ is not exactly what we are looking for, since it is not the genuine status of the system by the end of Mode 2a. Namely, we do not need just A_ν - the number of processed primary units, but $i-A_\nu,$ i.e., the actual number of

units left unserved from the total of i units at the beginning of mode 2a. In other words, we need the functional

$$\Gamma_0^{(i)}(z) = Ez^{i-A_\nu+\Pi_\nu},$$

because it gives us the total number of the primary units by the end of mode 2a. This can be obtained from (2.2.13) through the following modification

$$\Gamma_0^{(i)}(z) = z^i Ez^{-A_\nu} w^{\Pi_\nu}\big|_{w=\frac{1}{z}} = z^i h\big(\tfrac{1}{z}, z\big)$$

$$= z^i - [z - \gamma(\lambda - \lambda a(z))]\mathcal{D}_y^{S-1} \tfrac{z^i - \gamma^i[\lambda - \lambda a(z) + \mu - \mu b(y)]}{z - \gamma[\lambda - \lambda a(z) + \mu - \mu b(y)]}. \tag{2.2.14}$$

In the sequel, we will also need

$$\Gamma_0^{(i)}(0) = \gamma(\lambda)\mathcal{D}_y^{S-1}\gamma^{i-1}(\lambda + \mu - \mu b(y)). \tag{2.2.14a}$$

Example 2.2.1. To illustrate formula (2.2.14) for $\Gamma_0^{(i)}(z) = Ez^{i-A_\nu+\Pi_\nu}$ consider the following special case. Assume that $i = 2$ and $\tau_1 \in [Exp(\gamma)]$, that is $\gamma(\theta) = \frac{\gamma}{\gamma+\theta}$. Then,

$$\varphi_i(y, z) = \tfrac{z^i - \gamma^i[\lambda - \lambda a(z) + \mu - \mu b(y)]}{z - \gamma[\lambda - \lambda a(z) + \mu - \mu b(y)]} \tag{2.2.15}$$

in (2.2.14) immediately reduces to

$$\varphi_2(y, z) = z + \gamma[\lambda - \lambda a(z) + \mu - \mu b(y)] \tag{2.2.16}$$

alone by setting $i = 2$. Then with $\gamma(\theta) = \frac{\gamma}{\gamma+\theta}$, (2.2.16) turns to

$$\varphi_2(y, z) = z + \frac{\gamma}{\gamma+\lambda-\lambda a(z)+\mu-\mu b(y)}. \tag{2.2.17}$$

Assume further that $b(y) = y^2$, that is the server processes exactly two secondary units at a time leading to

$$\varphi_2(y, z) = z + \frac{\gamma}{\gamma+\lambda-\lambda a(z)+\mu-\mu y^2} = z + \frac{f}{1-ay^2}, \tag{2.2.18}$$

where

$$\begin{aligned} a &= a(y, z) = \frac{\mu}{\gamma+\mu+\lambda-\lambda a(z)} \quad \text{and} \\ f &= f(y, z) = \frac{\gamma}{\gamma+\lambda-\lambda a(z)+\mu}. \end{aligned} \tag{2.2.19}$$

Now, from Dshalalow [17], for an $m \in \mathbb{N}$,

$$\mathcal{D}_y^m \tfrac{1}{1-ay^2} = \begin{cases} \frac{1-a^{\frac{1}{2}(m+2)}}{1-a}, & m \text{ even} \\ \frac{1-a^{\frac{1}{2}(m+1)}}{1-a}, & m \text{ odd} \end{cases} \tag{2.2.20}$$

yielding

$$\mathcal{D}_y^{S-1}\varphi_2(y, z) = z + f\delta(a), \tag{2.2.21}$$

where

$$\delta(a) = \delta(a(y, z)) = \begin{cases} \frac{1-a^{\frac{1}{2}(S+1)}}{1-a}, & S \text{ odd} \\ \frac{1-a^{S/2}}{1-a}, & S \text{ even} \end{cases} \tag{2.2.22}$$

Thus

$$\Gamma_0^{(2)}(z) = z^2 - \Big[z - \tfrac{\gamma}{\gamma+\lambda-\lambda a(z)}\Big](z + f\delta(a))$$

$$= f\delta(a)\tfrac{\gamma}{\gamma+\lambda-\lambda a(z)} + z\Big[\tfrac{\gamma}{\gamma+\lambda-\lambda a(z)} - f\delta(a)\Big]$$

and from (2.2.19),

$$1 - a = \tfrac{\gamma+\lambda-\lambda a(z)}{\gamma+\mu+\lambda-\lambda a(z)} \quad \text{implying that } \tfrac{1}{1-a} = \tfrac{\gamma+\mu+\lambda-\lambda a(z)}{\gamma+\lambda-\lambda a(z)}$$

and

$$f\tfrac{1}{1-a} = \tfrac{\gamma}{\gamma+\lambda-\lambda a(z)+\mu}\tfrac{\gamma+\mu+\lambda-\lambda a(z)}{\gamma+\lambda-\lambda a(z)} = \tfrac{\gamma}{\gamma+\lambda-\lambda a(z)}.$$

Assume that $S = 2M$ is even. Then $\delta(a) = \big(1 - a^M\big)\tfrac{1}{1-a}$ and

$$f\delta(a) = \left[1 - \Big(\tfrac{\mu}{\gamma+\mu+\lambda-\lambda a(z)}\Big)^M\right]\tfrac{\gamma}{\gamma+\lambda-\lambda a(z)}$$

implying that

$$\Gamma_0^{(2)}(z) = \left[1 - \Big(\tfrac{\mu}{\gamma+\mu+\lambda-\lambda a(z)}\Big)^M\right]\Big(\tfrac{\gamma}{\gamma+\lambda-\lambda a(z)}\Big)^2$$

$$+ z\tfrac{\gamma}{\gamma+\lambda-\lambda a(z)}\left[1 - \left[1 - \Big(\tfrac{\mu}{\gamma+\mu+\lambda-\lambda a(z)}\Big)^M\right]\right]$$

finally arriving at

$$\Gamma_0^{(2)}(z) = \frac{\gamma}{\gamma+\lambda-\lambda a(z)}$$

$$\times\left\{\left[1-\left(\frac{\mu}{\gamma+\mu+\lambda-\lambda a(z)}\right)^M\right]\frac{\gamma}{\gamma+\lambda-\lambda a(z)} + z\left(\frac{\mu}{\gamma+\mu+\lambda-\lambda a(z)}\right)^M\right\} \tag{2.2.23}$$

and

$$\Gamma_0^{(2)}(0) = \left(\frac{\gamma}{\gamma+\lambda}\right)^2\left[1-\left(\frac{\mu}{\gamma+\mu+\lambda}\right)^M\right]. \tag{2.2.24}$$

□

2.3. FLUCTUATION ANALYSIS IN THE CONTEXT OF MODE 2B

Because mode 2b starts off when the queue of the primary units drops down to zero upon the end of a service, the formalism of mode 2b is much simpler than that of mode 2a. In this case, the server deals only with a single line of secondary units processing them in batches $v_1, v_2, \ldots$ in accordance with a general renewal process

$$(\mathcal{V}, S) = \sum_{k=1}^{\infty} v_k \varepsilon_{s_k}, \tag{2.3.1}$$

not as in Remark 3.1(d), where we assumed that $(\mathcal{V}, \mathcal{S})$ was Poisson. Now $(\mathcal{V}, S)$ is with position independent marking, where $\sum_{k=1}^{\infty}\varepsilon_{s_k}$ is the underlying support counting measure with s_1 being distributed in accordance with the LST

$$\zeta(\theta) = Ee^{-\theta s_1}, \tag{2.3.2}$$

(that is different from our assumptions in section 2.2), whereas $v_1, v_2, \ldots$ are as in section 2.1 (that is $Ez^{v_1} = b(z)$, however, with no restrictions). Mode 2b lasts until the total number of processed secondary units is at least S. During the period of mode 2b, the primary line may accumulate units waiting for the server to resume its work. However, it may not and the primary buffer until the end of mode 2a can still stay empty. We therefore need to adhere the other component of the primary units to $\mathcal{V}$:

$$(\mathcal{V}, \mathit{\Pi}) = \sum_{k=1}^{\infty} (v_k, \pi_k) \varepsilon_{s_k}. \tag{2.3.3}$$

Component $\mathit{\Pi}$ is obviously passive, while the $\mathcal{V}$ component is active on its threshold S. Let

$$\rho = \inf\{n \geq 0 : V_n = v_1 + \ldots + v_n \geq S\} \tag{2.3.4}$$

denote the exit index from mode 2b. Then, s_ρ is the first passage time (i.e., the end of mode 2b if it started at time 0) and $\mathit{\Pi}_\rho$ is the total number of primary units that enter the system during mode 2b.

The joint functional

$$\Phi(z, w, \theta) = Ez^{\mathit{\Pi}_\rho} w^{V_\rho} e^{-\theta s_\rho} \tag{2.3.5}$$

gives us the status of the system upon the exit from mode 2b. Analogous to section 2.1, we are interested in the marginal transform

$$\Gamma_0^{(0)}(z) = \Phi(z,1,0) = Ez^{\Pi_\rho}. \tag{2.3.6}$$

If $g(z,w,\theta) = Ew^{v_1}z^{\pi_1}e^{-\theta s_1},$ then

$$\Phi(z,w,\theta) = 1 - \{1-g(z,w,\theta)\}\mathcal{D}_y^{S-1}\tfrac{1}{1-g(z,wy,\theta)}, \tag{2.3.7}$$

and in its (2.3.6)-form

$$\Gamma_0^{(0)}(z) = 1 - \{1-g(z,1,0)\}\mathcal{D}_y^{S-1}\tfrac{1}{1-g(z,y,0)}. \tag{2.3.8}$$

Applying similar arguments as in section 2.2 and under the assumption that $(\mathcal{V},\mathcal{S})$ is with the position independent marking we have

$$g(z,w,\theta) = Ew^{v_1}E\big[e^{-\theta s_1}E[z^{\pi_1}|s_1]\big] = b(w)E\big[e^{-\theta s_1}e^{\lambda\theta[a(z)-1]}\big]$$

$$= b(w)\zeta(\theta+\lambda-\lambda a(z)). \tag{2.3.9}$$

Inserting (2.3.9) in (2.3.8), with $w=1$ and $\theta=0,$ we get

$$\Gamma_0^{(0)}(z) = 1 - [1-\zeta(\lambda-\lambda a(z))]\mathcal{D}_y^{S-1}\tfrac{1}{1-b(y)\zeta(\lambda-\lambda a(z))}. \tag{2.3.10}$$

In addition, we also need

$$\Gamma_0^{(0)}(0) = 1 - [1 - \zeta(\lambda)]\mathcal{D}_y^{S-1}\frac{1}{1-b(y)\zeta(\lambda)}. \tag{2.3.10a}$$

Formula (2.3.7) is due to Dshalalow [13]. The operator $\mathcal{D}$ was defined in (2.3.11a). $\Gamma_0^{(0)}(z)$ in (2.3.10) gives the quantity of primary units by the end of mode 2b.

2.4. FLUCTUATION ANALYSIS OF MODE 3

Upon the beginning of mode 3, the primary queue may or may not have customers waiting. If there is at least one customer in the primary buffer, the server immediately resumes his service. Otherwise, he waits for a first batch of customers to arrive, and only then does he resume service. Thus, mode 3 consists of two phases: the first one is waiting and the second one is servicing. The first phase is instantaneous if there is at least one primary unit upon the beginning of mode 3. To run through the first phase of mode 3 not knowing whether or not there are any primary customers by the end of mode 2 (a or b) can be a relatively minor nuisance. Fluctuation analysis is not only a remedy for this predicament, but also for a more complex configuration with the N-Policy (mentioned in section 1.2). A pertinent formula from Dshalalow [19] instructs us how to chain the outcome of mode 2, using it as an initial value expressed by functional $\Gamma_0^{(i)}(z)$, and then predict the outcome of the next phase.

We formalize the problem by introducing the following notation. Denote

$$\eta_i = \begin{cases} \Pi_\rho, & Q_0 = i = 0 \\ i - A_\nu + \Pi_\nu, & 0 < Q_0 = i < r \end{cases} \tag{2.4.1}$$

(the number of primary units at the beginning of mode 3) and

$$\Sigma_i = \begin{cases} \eta_i + \xi_\rho \delta_{0,\eta_i}, & Q_0 = i = 0 \\ \eta_i + \xi_\nu \delta_{0,\eta_i}, & 0 < Q_0 = i < r, \end{cases} \tag{2.4.2}$$

(the number of primary units upon the end of mode 3) where

$\delta_{i,j}$ is the Kronecker delta

$\xi_\nu \in [\xi]$ is the size of the first batch of customers arriving after τ_ν

$\xi_\rho \in [\xi]$ is the size of the first batch of customers arriving after τ_ρ.

Π_ρ is the total number of primary units that enter the system during mode 2b.

Π_ν is the total number of primary units that enter the system during mode 2a.

A_ν is the total number of primary units processed from a total of i units upon beginning of mode 2a.

Thus, Σ_i is the total number of the first priority (i.e., primary) customers in the system upon the end of mode 3. Denote

$$\alpha_i(z) = Ez^{\Sigma_i}.$$

(2.4.3)

Then, given that mode 3 begins with the number of units specified by the functional $\Gamma_0^{(i)}(z)$ ($Q_0 = i = 0$ for mode 2b and $0 < Q_0 = i < r$ for mode 2a), using [19] we have

$$\alpha_i(z) = \Gamma_0^{(i)}(z) - [1 - \Gamma(z)]\mathcal{D}_x^{N-1}\left(\frac{\Gamma_0^{(i)}(xz)}{1-\Gamma(xz)}\right). \tag{2.4.4}$$

The threshold $N\ (\geq 1)$ is what the primary queue needs to cross before the server resumes his work. In our case $N = 1$ and $\Gamma(z)$ reduces to just $a(z)$ being the pgf of ξ. Thus, with $N = 1,$ formula (2.4.4) turns to

$$\begin{aligned}\alpha_i(z) &= \Gamma_0^{(i)}(z) - [1 - \Gamma(z)]\mathcal{D}_x^0\left(\frac{\Gamma_0^{(i)}(xz)}{1-\Gamma(xz)}\right) \\ &= \Gamma_0^{(i)}(z) - [1 - a(z)]\Gamma_0^{(i)}(0), i = 0, 1, \ldots, r-1.\end{aligned} \tag{2.4.5}$$

in accordance with (2.2.11a) and the fact that $\Gamma(0) = a(0) = 0$. For $\Gamma_0^{(i)}(z)$ and $\Gamma_0^{(i)}(0)$ we will apply formulas (3.14-3.14a) and (4.10-4.10a), respectively, for $0 < Q_0 = i < r$ and $Q_0 = i = 0$. Formula (5.5) that gives $\alpha_i(z)$, could be alternatively obtained from probabilistic arguments, but they would be pretty hard to articulate.

2.5. QUEUEING PROCESS

As previously introduced (section 1.2), $Q(t)$ is the number of all primary units in the system at time $t \geq 0,$ defined as a piecewise linear process with right-continuous paths. The sequence $T_0 = 0, T_1, T_2, \ldots$ of successive service cycles completions is a sequence of stopping times relative to the filtration $(\mathcal{F}_t)$, upon which $Q(t)$ conditionally regenerates and thus forms a semi-regenerative process. If $Q_{n-1} = Q(T_{n-1}) \geq r,$ the next service begins immediately and it lasts $\sigma_n \in [\sigma]$ being arbitrarily distributed. Whenever $Q_{n-1} \geq r,$ from T_{n-1} to $T_n,$ the system is

in mode 1. So given $Q_{n-1} \geq r,$ the transitions from T_{n-1} to T_n are very similar to that of the M/G/1-queue, namely,

$$P_i(z) : = E\left[z^{Q_1} | Q_0 = i\right] = E\left[z^{i-r} z^{W_1}\right]$$

$$= z^{i-r} \beta(\lambda - \lambda a(z)), i \geq r, \tag{2.5.1}$$

where $\beta(\theta) : = Ee^{-\theta\sigma}$ is the LST of service time σ and W_1 is the number of customers that enter the system during that service period.

When $Q_{n-1} < r,$ the system enters mode 2 (versions a or b), with the server servicing two queues or one queue as per sections 3-5, followed by mode 3. Within mode 3, the server processes a single primary customer during phase 2 and thereby ending an underlying service cycle. $\{Q_n\}$ is a homogenous Markov chain, specified by the transitions

$$Q_1 = \begin{cases} \Sigma_{Q_0} + W_1 - 1, & 0 \leq Q_0 < r \\ Q_0 + W_1 - r, & Q_0 \geq r \end{cases} \tag{2.5.2}$$

where W_1 is the number of primary units arriving at the system during a service (of one or r customers) in mode 1, and Σ_i was introduced in section 5.

For $Q_0 = i < r,$ we use formula (2.4.5) that gives the number of customers in the system by the end of mode 3. It will be integrated into the conditional expectation like (2.5.1) as follows.

$$P_i(z) = E\left[z^{Q_1}|Q_0 = i\right] = z^{-1}Ez^{\Sigma_i}Ez^{W_1}$$

$$= z^{-1}\alpha_i(z)\beta(\lambda - \lambda a(z)), i < r, \tag{2.5.3}$$

where $\alpha_i(z)$ satisfies formulas (2.4.5), (2.2.14-2.2.14a), and (2.3.10-2.3.10a). From (2.5.3), with the notation $p_{ij} = P\{Q_1 = j|Q_0 = i\}$ we have

$$p_{ij} = \begin{cases} P\{V_1 = j - (i-r)\} = \begin{cases} q_{j-(i-r)}, & j \geq i - r \\ 0, & j < i - r \end{cases}, & i \geq r \\ P\{\Sigma_i - 1 + W_1 = j\} = \sum_{k=0}^{j} \underbrace{P\{W_1 = j-k\}}_{q_{j-k}} \underbrace{P\{\Sigma_i = k+1\}}_{\sigma_{i,k+1}}, & i < r \end{cases} \tag{2.5.4}$$

Hence the transition probability matrix P reads

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} & \cdots\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{r-1,0} & p_{r-1,1} & p_{r-1,2} & p_{r-1,3} & p_{r-1,4} & \cdots\cdots \\ q_0 & q_1 & q_2 & q_3 & q_4 & \cdots\cdots \\ 0 & q_0 & q_1 & q_2 & q_3 & \cdots\cdots \\ 0 & 0 & q_0 & q_1 & q_2 & \cdots\cdots \\ 0 & 0 & 0 & q_0 & q_1 & \cdots\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{2.5.5}$$

Because $q_j > 0$ and $\sigma_{i,k+1} > 0$ in (2.5.4) for all $k = 0, 1, \ldots, j,$ the upper block of this matrix (rows from 0 to $r-1$) is filled with all positive elements, whereas the lower complementary block matrix is a triagonal matrix. It is called (Abolnikov and Dukhovny [3]) a Δ_{r+1}-*matrix*. From (2.5.4-2.5.5), we conclude that the Markov chain $\{Q_n\}$ is irreducible and aperiodic. By Abolnikov-Dukhovny [3], it is recurrent positive if and only if $P_r'(1) < r$ (where $P_i(z)$ is the generating function of the ith row of P). The latter condition reduces to

$$L = ab\lambda < r. \tag{2.5.6}$$

(L is often referred to as the *offered load* although this notion applies to more basic systems.)

The pgf $P(z)$ of the invariant probability measure $\mathbf{p} = (p_0, p_1, \dots)$ of $\{Q_n\}$ (under condition (2.5.6)) satisfies the expression

$$P(z) = \sum_{i=0}^{\infty} p_i P_i(z)$$

that reduces to the Kendall's (Pollaczek-Khintchine)-type formula after a straightforward algebra, with (2.5.1-2.5.3) in mind:

$$P(z) = \beta(\lambda - \lambda a(z)) \frac{\sum_{i=0}^{r-1} p_i \{\alpha_i(z) z^{r-1} - z^i\}}{z^r - \beta(\lambda - \lambda a(z))}. \tag{2.5.7}$$

The above formula for $P(z)$ contains r unknown probabilities $p_0, \dots, p_{r-1}$ which can be computed numerically using the following procedure. First, it can be readily shown that

$$P(z) = \beta(\lambda - \lambda a(z)) \left[z^{r-1} \sum_{i=0}^{r-1} p_i \alpha_i(z) + z^{-r} \underbrace{\sum_{i=r}^{\infty} p_i z^i}_{\mathcal{T}_r(z)} \right] \tag{2.5.8}$$

We use the notation

$$P(z) = \underbrace{\sum_{i=0}^{r-1} p_i z^i}_{H_r(z)} + \underbrace{\sum_{i=r}^{\infty} p_i z^i}_{\mathcal{T}_r(z)}.$$

Then,

$$\mathcal{T}_r(z)[1 - \beta(\lambda - \lambda a(z))z^{-r}] = \beta(\lambda - \lambda a(z))\sum_{i=0}^{r-1} p_i \alpha_i(z) - \sum_{i=0}^{r-1} p_i z^i$$

$$= \sum_{i=0}^{r-1} p_i[\alpha_i(z)\beta(\lambda - \lambda a(z)) - z^i]$$

implying that

$$\mathcal{T}_r(z) = z^r \frac{1}{z^r - \beta(\lambda - \lambda a(z))} \sum_{i=0}^{r-1} p_i \Big[z^{-1} \alpha_i(z) \beta(\lambda - \lambda a(z)) - z^i \Big] \tag{2.5.9}$$

that is,

$$\mathcal{T}_r(z) = z^r \frac{N(z)}{D(z)}, \tag{2.5.10}$$

where

$$N(z) = \sum_{i=0}^{r-1} p_i \Big[z^{-1} \alpha_i(z) \beta(\lambda - \lambda a(z)) - z^i \Big] \tag{2.5.11}$$

and

$$D(z) = z^r - \beta(\lambda - \lambda a(z)). \tag{2.5.12}$$

This is the Kendall's-type formula for the r-tail $\mathcal{T}_r(z)$ of $P(z)$.

Now according to [3], the denominator $D(z)$ has exactly r roots in $\overline{B}(0,1)$ of which root $z_0 = 1$. Furthermore, according to Dukhovny [35], all roots on the boundary $\partial B(0,1)$ are simple. Since $\mathcal{T}_r(z)$ is obviously analytic inside $B(0,1)$ and continuous on the boundary $\partial B(0,1),$ all roots $z_1, \ldots, z_{r-1}$ of the denominator are the roots of the numerator $N(z)$. We deal with $z_0 = 1$ separately from the condition $P(1) = 1$. These are additional r conditions in r unknown probabilities $p_0, \ldots, p_{r-1}$ that fill the pgf $H_r(z) = \sum_{i=0}^{r-1} p_i z^i$. In the procedure below we assume that all roots are simple (which is the most common practical case). Otherwise, if some roots are multiple, we differentiate $N(z)$ accordingly and then substitute those roots.

So we have

$$N(z_j) = \sum_{i=0}^{r-1} p_i \left[z_j^{-1} \alpha_i(z_j) \beta(\lambda - \lambda a(z_j)) - z_j^i \right]$$

$$= \sum_{i=0}^{r-1} p_i a_{ij} = 0, j = 0, \ldots, r-1 \tag{2.5.13}$$

implying that

$$(N(z_0), \ldots, N(z_{r-1})) = (p_0, \ldots, p_{r-1}) A_0 = \mathbf{0}, \tag{2.5.14}$$

where

$$A_0 = \begin{pmatrix} b_{00} & a_{01} & \ldots & a_{0,r-1} \\ b_{10} & a_{11} & \ldots & a_{1,r-1} \\ \vdots & \vdots & \vdots & \vdots \\ b_{r-1,0} & a_{r-1,1} & \ldots & a_{r-1,r-1} \end{pmatrix}. \tag{2.5.15}$$

One can easily see that the column $\begin{pmatrix} b_{00} \\ b_{10} \\ \vdots \\ b_{r-1,0} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, because $z_0 = 1$. Thus matrix A_0 is singular and with $z_0 = 1$ we arrive at identity $0 = 0$. We therefore replace A_0 matrix with $\widetilde{A}_0$ in which the first column is dropped herewith arriving at $r - 1$ equations. The rank of matrix $\widetilde{A}_0$ is obviously $r - 1$ and it is not a square matrix. An additional equation comes from the condition

$$1 = \sum_{i=0}^{r-1} p_i + \mathcal{T}_r(1) = \sum_{i=0}^{r-1} p_i + \lim_{z \to 1-} \frac{N'(z)}{D'(z)}. \tag{2.5.16}$$

So we have

$$N'(1) = \sum_{i=0}^{r-1} p_i[\alpha_i'(1) + L - 1 - i] \tag{2.5.17}$$

and

$$D'(1) = r - L. \tag{2.5.18}$$

We recall that $P_r'(1) = L < r$ is a necessary and sufficient condition for the embedded Markov chain to be recurrent positive. So, $D'(1) > 0$. Thus we have from (2.5.16-2.5.18)

$$1 = \sum_{i=0}^{r-1} p_i + \frac{1}{r-L} \sum_{i=0}^{r-1} p_i[\alpha_i'(1) + L - 1 - i]$$

or in the form

$$1 = \sum_{i=0}^{r-1} p_i \big\{ 1 + \tfrac{1}{r-L}[\alpha_i'(1) + L - 1 - i] \big\}$$

$$\text{implying} \qquad 1 = \tfrac{1}{r-L} \sum_{i=0}^{r-1} p_i [\alpha_i'(1) + r - i - 1] = \sum_{i=0}^{r-1} p_i a_{i0} \tag{2.5.19}$$

or in the form

$$(p_0, \ldots, p_{r-1}) \begin{pmatrix} a_{00} \\ a_{10} \\ \vdots \\ a_{r-1,0} \end{pmatrix} = 1. \tag{2.5.20}$$

Now replacing the first column of matrix A_0 with column $\begin{pmatrix} a_{00} \\ a_{10} \\ \vdots \\ a_{r-1,0} \end{pmatrix}$ we arrive at the following system of linear equations in unknowns $p_0, \ldots, p_{r-1}$:

$$(p_0, \ldots, p_{r-1}) A = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{e} \in \mathbb{R}^r, \tag{2.5.21}$$

where

$$A = (a_{ij} : i, j = 0, \ldots, r-1). \tag{2.5.22}$$

System (2.5.21) has a unique solution guaranteed by the ergodicity condition $L < r$ and the above analyticity arguments. We conclude this discussion by mentioning that there are many numerical methods and software modules that deal with calculating the roots of the equation $z^r - \beta(\lambda - \lambda a(z))$ once the functional $\beta(\lambda - \lambda a(z))$ gets specified.

SUMMARY. In this chapter we analyze a complex queueing system with two queues and one server that occurs in common computer operating systems. Here when the intensity of input calms down, the server begins to operate in a slow pace due to servicing two (instead of one) queues. It is normally a routine maintenance on registry and cleaning (among other things) that real computer servers are programmed to render. The latter should not be confused with a conventional vacation policy, because the server is still available in the main operating mode. The server gets back to the primary queue and dedicates himself entirely to this activity when the intensity of the arriving jobs picks up again. There are provisions in such operating systems to reduce bouncing between operating modes driven by an overly volatile input that we also included in our analysis. A common remedy against this adverse effect would be to administer the N-Policy, that we however tried to avoid in order to soften the complexity of fluctuation analysis we use throughout and to improve tractability of our results, although some light form of the N-Policy is incorporated in one of the operating modes.

We focused on the queueing process embedded over departure epochs of primary units at consecutive ends of service cycles (not exactly in the spirit of classical M/G/1 systems). Among main accomplishments in this work, we implement fluctuation analysis and elements of stochastic antagonistic games that enable us to arrive at closed form functionals. Another novelty was to model a class of real-world operating systems coming as close as possible to their realistic complexities. The experience gained in this modeling lets us further embellish our system by assimilating the N-Policy and come up with a minimal collateral damage to the closed form functionals we deem to target. This is however, our forthcoming work.

Implementation of stochastic control in the underlying system would be yet another major enhancement proceeded from further analysis of a continuous time parameter queuing process. Many more examples and illustrations of the results are to be obtained to include the mean queue length, waiting time, and mean service cycle in the equilibrium, to name a few.

CHAPTER 3

FLUCTUATION ANALYSIS IN PARALLEL QUEUES WITH HYSTERETIC CONTROL

3.1. FORMALISM

Our system is under the following specifications. We denote $Q(t)$ the continuous time parameter queueing process on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$. This process gives the total number of primary units in the system at time $t \geq 0$ and it is adapted to $(\mathcal{F}_t)$. We postulate $Q(t)$ as a piecewise constant process with right-continuous paths. The queueing process is formed of input and servicing processes. The input process is a stationary marked Poisson process

$$\mathfrak{P} = \sum_{k=1}^{\infty} \mathfrak{a}_k \varepsilon_{t_k} \quad (\varepsilon_a \text{ is a unit measure at } a) \tag{3.1.1}$$

with position independent marking, the associated support counting measure $\mathfrak{S} = \sum_{k=1}^{\infty} \varepsilon_{t_k}$ of intensity $E\mathfrak{S} = \lambda|\cdot|$ where $|\cdot|$ is the Borel-Lebesgue measure and $\lambda = E\mathfrak{S}[0,1]$ is the rate. Here the marks $\mathfrak{a}_1, \mathfrak{a}_2, \ldots \in [\mathfrak{a}]$ represent batches of arriving units, with the common pgf $a(z) = Ez^{\mathfrak{a}}$ of a batch size, and $t_1, t_2, \ldots$ are the successive arrivals of those batches of units.

Arriving units line up in the buffer and are processed in the order of their arrivals, except that the arrived batches of units can be allocated within different servicing groups of units. Speaking of servicing groups, server's capacity is R (positive integer), but the server is to take at least $r(\leq R)$ units for service if available. In the basic M/G/1 queue, $T_0, T_1, T_2, \ldots$ are successive moments of customers

departures from the system after they are fully processed. Not so in our system, because the nature of T_n depends upon the mode the system is in, although T_n's are some but not all departure epochs. If at time $T_n,$ the primary buffer count is r or more units, the server takes $\min\{Q(T_n), R\}$ units for processing and by time T_{n+1} this group departs from the system. Thus the period of time between T_n and T_{n+1} refers to **mode 1**.

If at time $T_n,$ there are fewer than r primary units, the server enters one of the two modes. With no primary units, the server begins processing batches from an existing secondary buffer of units that are of random seizes until at least S of them are processed. We assume that the contents of that buffer is unlimited. The service process during this mode (called **mode 2a**) evolves in accordance with the marked point process

$$(\mathcal{V}, \mathcal{S}) = \sum_{j=1}^{\infty} v_j \varepsilon_{s_j}, \tag{3.1.2}$$

where v_j is the size of the jth batch of secondary units and s_j is the time when the service of that batch ends. Without the restriction imposed on the total number of processed secondary units, $(\mathcal{V}, \mathcal{S})$ would continue endlessly. Here is how we implement that restriction. Let

$$\rho : = \inf\{n \geq 1 : V_n = v_1 + \ldots + v_n \geq S\}. \tag{3.1.3}$$

Then we call ρ the *exit index* from mode 2a, and s_ρ is the *first passage time* when the server is done with secondary units. V_ρ is the total number of secondary units

processed during mode 2a (which is S or more). With this modification (3.2) must rightfully read as

$$(\mathcal{V}, \mathcal{S})_\rho = \sum_{j=1}^{\rho} v_j \varepsilon_{s_j}. \tag{3.1.4}$$

We postpone the discussion on further analytical steps regarding marked random measure of (3.1.4) and turn to mode 2b. In another version of mode 2, called **mode 2b**, we assume that the number of primary units, while dropped below r, is positive. In this case, the server works on two parallel queues simultaneously, but asynchronously, processing two buffers of primary and secondary units. Because the number of primary units is perceived relatively low, the server slows down its attention to the primary units processing them singly, one-by-one, and with a different distribution. There are various complexities though. For one, since primary and secondary units are of different nature, their service time distributions are different. Secondly, if the server is done with S or more secondary units, he cannot end mode 2b abruptly, because it is possible that a primary unit will still be in service and the server therefore needs to finish it. During this interval of time, the server will continue servicing secondary units even in further excess of S. If he is done with all primary units left behind after mode 1, but he did not finish servicing at least S secondary units, he nevertheless departs from mode 2b. To continue with this formalism, let us introduce the following marginal process. Let

$$(\mathcal{X}, \tau) = \sum_{k=1}^{\infty} x_k \varepsilon_{\tau_k}, \tag{3.1.5}$$

describe that evolution of processing primary units where $\tau_1, \tau_2, \ldots$ are successive completions of services of primary units during mode 2b, with $x_0 = 0$ and $x_1 = x_2 = \ldots = 1$. This is a marked point process with all marks but x_0 being constant 1. The server also takes on the secondary queue of jobs processing them forming the second marked point process

$$(\mathcal{V}, \mathcal{S}) = \sum_{j=1}^{\infty} v_j \varepsilon_{s_j}$$

already introduced in (3.1.2). It is reasonable to assume that the processes $(\mathcal{X}, \tau)$ and $(\mathcal{V}, S)$ are independent. Merging the two in one reads

$$(\mathcal{X}, \mathcal{Y}, \tau) = \sum_{k=0}^{\infty} (x_k, y_k) \varepsilon_{\tau_k}. \tag{3.1.6}$$

Notice that $(\mathcal{V}, S)$ has changed, because its reading goes after epochs τ_k's, that is $(\mathcal{V}, \mathcal{S})$ observed not upon its original times s_k's. Here is what (3.1.6) exactly means. $(\mathcal{X}, \mathcal{Y}, \tau)$ is a bivariate marked random process observed exclusively upon times τ_k's (where decision makings take place), with exactly 1 primary unit processed upon τ_k and y_k secondary units processed in interval $(\tau_{k-1}, \tau_k]$. This can be zero of them or one or more batches. Let $0 < Q_0 < r$ be the number of primary units left behind at time T_0 upon entering mode 2b. For convenience we operate on the first service cycle lasting from T_0 to T_1.) We introduce the following index

$$\nu := Q_0 \wedge \inf\{n \geq 1 : Y_n = y_1 + \ldots + y_n \geq S\} \tag{3.1.7}$$

referred to as the exit index. The r.v. τ_ν is the first passage time. If X_k is the number of primary units processed singly by time τ_k, then X_k obviously equals k. Thus, X_ν is the random number of primary units processed by the completion of mode 2b at time τ_ν. Consequently, upon the end of mode 2b, there are the rest of $Q_0 - X_\nu$ primary units that remain in the system (from the total of Q_0 units available at the beginning of mode 2b) upon the end of mode 2b. B_ν is the number of secondary units processed by τ_ν. Of course, the queue of $Q_0 - X_\nu$ primary units (possibly > 0) will be joined by new arrivals during the interval $[0, \tau_\nu]$. This will be yet another random component by the end of mode 2b to worry about.

The process $(\mathcal{X}, \mathcal{Y}, \tau)$ of (3.1.6) does not run indefinitely, but it is terminated according to the same guidelines as in mode 2a. Namely as

$$(\mathcal{X}, \mathcal{Y}, \tau)_\nu = \sum_{k=1}^{\nu} (1, y_k)\varepsilon_{\tau_k}. \tag{3.1.8}$$

By leaving mode 2 (a or b), the number of primary units may increase at expense of new arrivals. The next is to implement the "N-Policy" that requires at least N primary units to have the server resume mode 1 service. That is not granted and in this case the server waits and rests for the primary buffer to replenish to at least N units to get started. Alternatively, there can already be that many units available upon the end of mode 2. These two variants are combined in mode 3 that can be instantaneous or it can last a random time period dependent on the contents of primary units at the end of mode 2.

We close this section with the notion of a service cycle that is instrumental to the formation of the embedded queueing process. We start off with assigning to a

service cycle one of the two periods. If after servicing a batch of units of a random size between r and R and their departure from the system at time $T_n,$ the primary buffer $Q_n = Q(T_n)$ is filled with r or more units, the server continues working in mode 1 and he takes a batch of units not in excess of R for service that ends at T_{n+1}. The period of time between T_n and T_{n+1} is called a simple service cycle. If $Q_n < r,$ the server enters a period of modes 2 (versions a or b) and 3, followed by processing a random batch of units by T_{n+1}. The size of that batch is R in the model when $R \leq N$ or with $R > N,$ it is $\min\{Q, R\},$ where without too much formalities, Q is the queue length upon the end of mode 3. To tell a simple cycle from another type, we will refer the latter to as a complex service cycle.

In the forthcoming sections we revisit the formalism of the modes and argue that the queueing process $Q(t)$ is semi-regenerative relative to the sequence $\{T_n\}$ of stopping time with respect to filtration $(\mathcal{F}_t)$ over which we build the embedded Markov process.

3.2. BASIC FLUCTUATION ANALYSIS

To continue with the system we need to use results from Dshalalow [15-16] regarding consecutive exits and switching between the modes. We start off with a simpler variant pertaining to mode 2a and enhance the marked random measure $(\mathcal{V}, \mathcal{S})_\rho = \sum_{j=1}^{\rho} v_j \varepsilon_{s_j}$ of (3.1.4) by one more component that registers the number of primary units entering the system during the period of time when the server is busy processing secondary units.

Denote π_k the total number of primary units that enter the system in interval $(s_{k-1}, s_k]$ that is between two successive departures of batches $k-1$ and k. Then the extended process reads

$$(\mathcal{V}, \mathcal{S}, \mathit{\Pi})_\rho = \sum_{j=1}^{\rho} (v_j, \pi_k)\varepsilon_{s_j}. \tag{3.2.1}$$

For example, $(\mathcal{V}, \mathcal{S}, \mathit{\Pi})_\rho[0, s_\rho]$ gives the number of processed secondary units and newly arrived primary units during the period of time from the beginning to the end of mode 2a. Denote

$$\mathit{\Pi}_n = \pi_1 + \ldots + \pi_n. \tag{3.2.2}$$

Thus $\mathit{\Pi}_\rho$ gives the total number of primary units that arrive in the system during the period of time the system runs mode 2a. This enhancement changed the random measure from position independent to position dependent marking, because the added component π_k depends on $(s_{k-1}, s_k]$. The joint functional

$$\Phi(z, w, \theta) = Ez^{\mathit{\Pi}_\rho} w^{V_\rho} e^{-\theta s_\rho},\ \|z\| \le 1, \|w\| \le 1, \mathrm{Re}\theta \ge 0, \tag{3.2.3}$$

gives the status of the system upon exit from mode 2a, where ρ is the exit index defined in (3.1.3), s_ρ is the first passage time (i.e., the exit time from mode 2a), and $\mathit{\Pi}_\rho$ is the content of the primary buffer upon the exit. Note that of the two random marks, v's and π's, the first is active and the second one is passive. This is because the accumulation of serviced second priority batches stops as soon as the total crosses threshold S regardless of the collection of primary units. However, we are

more interested in the quantity of primary units by then. Theorem 4.1 below is due to Dshalalow [20] adapted to our settings.

Let $[\mathbb{C}^2, \mathbb{C}, \varphi]$ be a function (on $\mathbb{C}^2$ and valued in $\mathbb{C}$). Define the following operator

$$\mathcal{D}_x^k \varphi(x,z) = \begin{cases} \frac{1}{k!} \lim_{x\to 0} \frac{\partial^k}{\partial x^k} \left[\frac{1}{1-x}\varphi(x,z)\right], & k \geq 0 \\ 0, & k < 0 \end{cases} \tag{3.2.4}$$

referred to as the $\mathcal{D}$-operator.

According to Dshalalow [27], $\mathcal{D}$ is a linear operator with fixed points at constant functions in variable x that also truncates series. It is a useful tool in discrete operational calculus.

Theorem 3.2.1 (Dshalalow [20]). *Let*

$$g(z,w,\theta) = Ew^{v_1} z^{\pi_1} e^{-\theta s_1},\ \|z\| \leq 1, \|w\| \leq 1, \mathrm{Re}\theta \geq 0, \tag{3.2.5}$$

Then

$$\Phi(z,w,\theta) = 1 - \{1 - g(z,w,\theta)\}\mathcal{D}_y^{S-1} \frac{1}{1-g(z,wy,\theta)}. \tag{3.2.6}$$

We assume that the support counting measure $\sum_{k=1}^{\infty} \varepsilon_{s_k}$ forms a regular renewal process with the inter-renewal times distributed according to

$$\zeta(\theta) = Ee^{-\theta s_1} \tag{3.2.7}$$

(i.e., a general distribution expressed as a LST). The marks $v_1, v_2, \ldots$ are iid r.v.'s assuming positive integer values distributed according to

$$b(z) = Ez^{v_1} \tag{3.2.8}$$

(no restrictions). Earlier, we assumed that the marginal process $(\mathcal{V}, S) = \sum_{k=1}^{\infty} v_k \varepsilon_{s_k}$ is with position independent marking. Thus, by the assumptions made in the beginning of section 3,

$$g(z, w, \theta) = Ew^{v_1} E\left[e^{-\theta s_1} E[z^{\pi_1} | s_1]\right] = b(w) E\left[e^{-\theta s_1} e^{\lambda\theta[a(z)-1]}\right]$$
$$= b(w)\zeta(\theta + \lambda - \lambda a(z)). \tag{3.2.9}$$

In the forthcoming sections we will be interested in the marginal functional

$$\Gamma_0^{(0)}(z, \theta) = \Phi(z, 1, \theta) = 1 - \{1 - g(z, 1, \theta)\}\mathcal{D}_y^{S-1} \tfrac{1}{1-g(z,y,\theta)}$$
$$= 1 - [1 - \zeta(\theta + \lambda - \lambda a(z))]\mathcal{D}_y^{S-1} \tfrac{1}{1-b(y)\zeta(\theta+\lambda-\lambda a(z))}. \tag{3.2.10}$$

To demonstrate the analytical tractability of formula (3.2.10) just obtained consider the following special case.

Proposition 3.2.2. *Suppose secondary units are being processed in batches of sizes distributed geometrically with parameter* p *(and* $q = 1 - p$*), i.e.*

$$b(y) = \tfrac{py}{1-qy}. \tag{3.2.11}$$

Then the functional $\Gamma_0^{(0)}(z,\theta)$ *reduces to*

$$\Gamma_0^{(0)}(z,\theta) = \zeta(\theta+\lambda-\lambda a(z))[p\zeta(\theta+\lambda-\lambda a(z))+q]^{S-1}. \tag{3.2.12}$$

Proof. For brevity write $\zeta = \zeta(\theta+\lambda-\lambda a(z))$. Then, under assumption (3.2.11), the expression $\psi = \psi(z,y,\theta) = \frac{1}{1-b(y)\zeta(\theta+\lambda-\lambda a(z))}$ reduces to

$$\psi = \tfrac{1-qy}{1-(p\zeta+q)y}. \tag{3.2.13}$$

Then, applying the $\mathcal{D}$-operator to (3.2.13) and using the properties of the $\mathcal{D}$-operator (Dshalalow [27]) we have

$$\mathcal{D}_y^{S-1}\psi = \mathcal{D}_y^{S-1}\Big(\tfrac{1-qy}{1-(p\zeta+q)y}\Big) = p\sum_{j=0}^{S-2}(p\zeta+q)^j + (p\zeta+q)^{S-1}. \tag{3.2.14}$$

Then, by (3.2.10) and (3.2.14),

$$\Gamma_0^{(0)}(z,\theta) = 1-(1-\zeta)\left(p\sum_{j=0}^{S-2}(p\zeta+q)^j + (p\zeta+q)^{S-1}\right) = \zeta(p\zeta+q)^{S-1}. \quad \square$$

Example 3.2.1. Under the conditions of Proposition 3.2.2, consider the marginal functional

$$Ee^{-\theta\tau_\rho} = \Gamma_0^{(0)}(1,\theta) = \zeta(\theta)[p\zeta(\theta)+q]^{S-1}.$$

Then, the mean time of mode 2a is

$$Es_\rho = -\lim_{\theta\to 0}\tfrac{d}{d\theta}Ee^{-\theta s_\rho} = b(q+pS\,),$$

where $b = Es_1$ is the mean batch size of serviced secondary units.

If on the other hand, $\theta = 0$,

$$Ez^{\Pi_\rho} = \Gamma_0^{(0)}(z,0) = \zeta(\lambda - \lambda a(z))[p\zeta(\lambda - \lambda a(z)) + q]^{S-1}.$$

Then, the mean number of the arrivals to the system during mode 2a is

$$E[\Pi_\rho] = \lim_{z\to 1} \frac{d}{dz} Ez^{\Pi_\rho} = -\lambda a'(1)\zeta'(0)[1 + p(S-1)]$$

$$= \lambda ab(q + pS) = \lambda a Es_\rho,$$

where $a = E\mathfrak{a}_1$ the mean batch size of arriving units and $b = Es_1$ is the mean batch size of serviced secondary units. Note that λa is the expected primary batch arrivals per unit time, times the expected number of batch sizes, i.e. the expected number of customers arriving per unit time. It therefore agrees with the intuitive interpretation that $E\Pi_\rho = \lambda a Es_\rho$, the expected number of primary arrivals during mode 2a equals the expected number of arrivals per unit time multiplied by the duration of mode 2a, in agreement with renewal processes, although we must mention that $(\mathcal{V}, \mathcal{S}, \Pi)_\rho$ over the interval $[0, s_\rho]$ does not behave like a stationary and independent increments process, because (Dshalalow [19-21]) all successive subintervals of secondary units' departures are neither independent nor identically distributed.

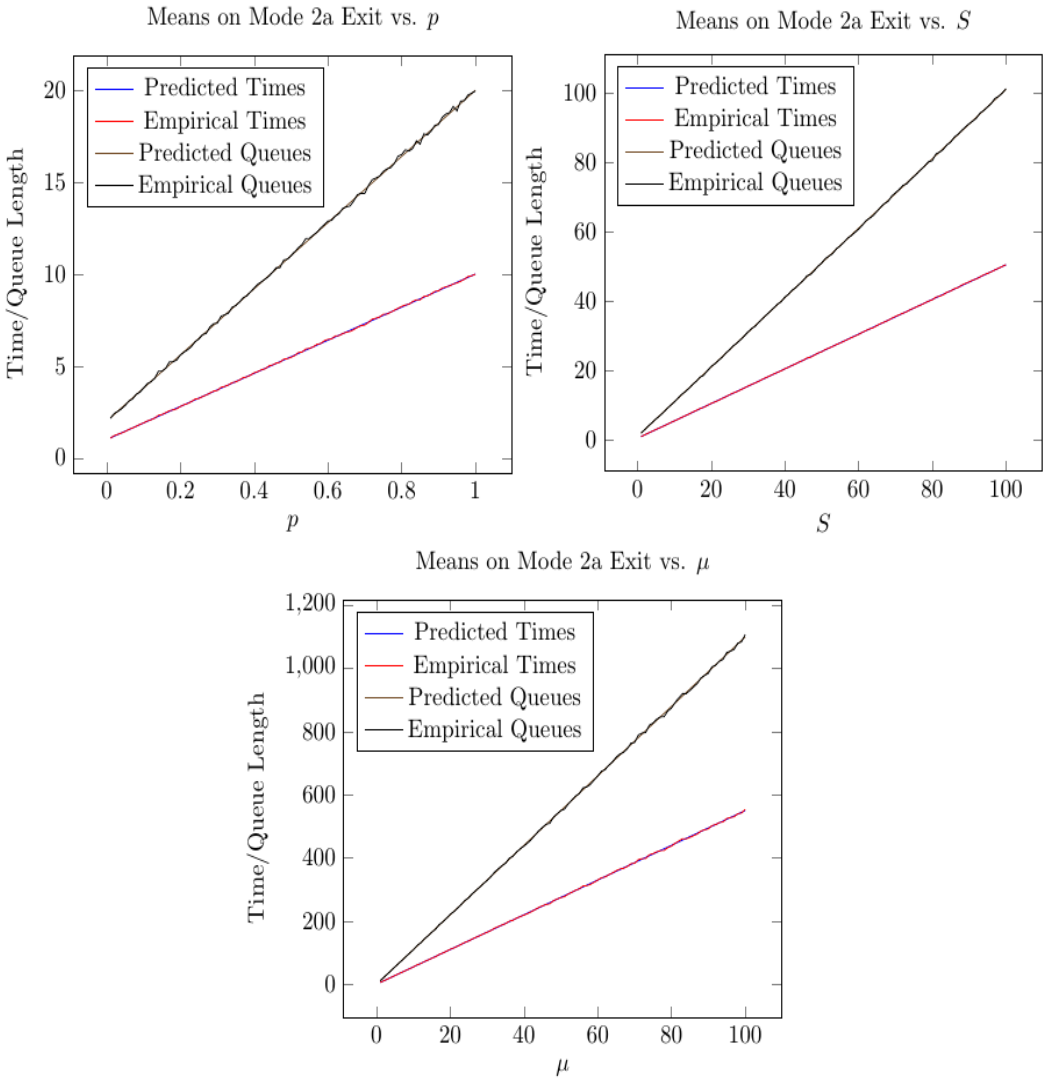


Figure 3.2.1: Mode 2a for parameters $S = 10$, $\lambda = \mu = 1$, $\alpha = p = 0.5$ (with one parameter varying). When varying p (in $\{0.1, 0.2, ..., 0.99\}$), S (in $\{1, 2, ..., 100\}$), and μ (in $\{1, 2, \dots, 100\}$), the mean queue length at the end of the mode and mean time for the mode are plotted as predicted by the formulas above along with the empirical mean measured for 10,000 simulations for each set of parameters. The Predicted values refer to the population values, while the empirical values is the sample simulation. As the figures show, The sampling curve is almost overlapping of the population curve.

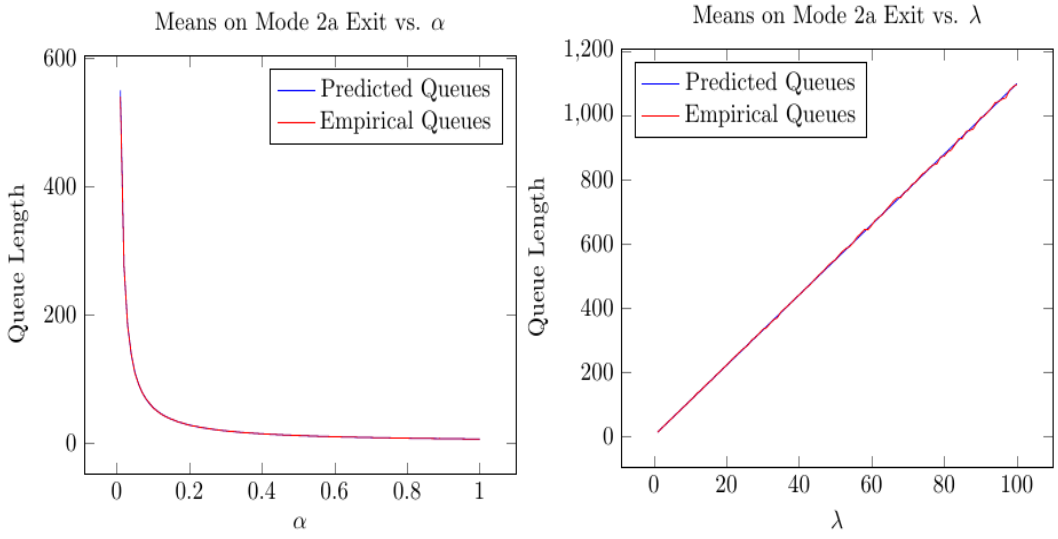


Figure 3.2.2: Mode 2a for parameters $S = 10$, $\lambda = \mu = 1$, $\alpha = p = 0.5$ (with one parameter varying). When varying α (in $\{0.1, 0.2, ..., 0.99\}$) and λ (in $\{1, 2, ..., 100\}$), the mean queue length at the end of the mode is plotted as predicted by the formulas above along with the empirical mean measured for 10,000 simulations for each set of parameters. □

Example 3.2.2. Under the assumptions of Example 3.2.1, suppose the service times of secondary units are independent of batch sizes and are exponentially distributed with parameter μ, that is, $\zeta(\theta) = \frac{\mu}{\theta+\mu}$. Then

$$Ee^{-\theta s_\rho} = \zeta(\theta)[p\zeta(\theta) + q]^{S-1} = \sum_{j=0}^{S-1} \binom{S-1}{j} p^j \left(\frac{\mu}{\theta+\mu}\right)^{j+1} q^{S-1-j}.$$

Then, applying the inverse Laplace transform to $\frac{1}{\theta} Ee^{-\theta s_\rho}$, we find

$$P\{s_\rho \le x\} = \sum_{j=0}^{S-1} \binom{S-1}{j} p^j q^{S-1-j} P(j+1, \mu x),$$

where $P(j+1, \mu x) = \frac{\Gamma(j+1) - \Gamma(j+1, \mu x)}{\Gamma(j+1)}$ is the upper regularized gamma function.

For the queue length,

$$\begin{aligned} Ez^{\Pi_\rho} &= \zeta(\lambda - \lambda a(z))[p\zeta(\lambda - \lambda a(z)) + q]^{S-1} \\ &= \sum_{j=0}^{S-1} \binom{S-1}{j} p^j \left(\frac{\mu}{\mu + \lambda\left(\frac{1-z}{1-\beta z}\right)}\right)^{j+1} q^{S-1-j}. \end{aligned}$$

Abbreviate $c = \frac{\lambda+\beta\mu}{\lambda+\mu}$, then notice that

$$\left(\frac{\mu}{\mu+\lambda\left(\frac{1-z}{1-\beta z}\right)}\right)^{j+1} = \left(\frac{\mu}{\mu+\lambda}\right)^{j+1}\sum_{k=0}^{j+1}\binom{j+1}{k}(-\beta)^k\frac{z^k}{(1-cz)^{j+1}}.$$

The distribution of Π_ρ depends on the following term, by Leibnitz formula,

$$\frac{1}{n!}\lim_{z\to 0}\frac{\partial^n}{\partial z^n}\left(\frac{z^k}{(1-cz)^{j+1}}\right) = \frac{1}{n!}\sum_{l=0}^{n}\binom{n}{l}c^l\frac{(j+l)!}{j!}$$

$$1_{\{k=n-l\}}k! = \frac{k!}{n!}\binom{n}{n-k}c^{n-k}\frac{(j+n-k)!}{j!} = \binom{j+n-k}{j}c^{n-k}.$$

So we have

$$P\{\Pi_\rho = n\} = \sum_{j=0}^{S-1}\binom{S-1}{j}p^j q^{S-1-j}\left(\frac{\mu}{\mu+\lambda}\right)^{j+1}\sum_{k=0}^{j+1}\binom{j+1}{k}\binom{j+n-k}{j}(-\beta)^k\left(\frac{\lambda+\beta\mu}{\lambda+\mu}\right)^{n-k}.$$

These results may be validated with simulation as well, but to avoid redundancy, we include results of some more sophisticated simulation validating the distributions of the time of modes 2b and 3, and distribution of the queue length after passing through these two modes in a later section. □

3.3. FLUCTUATION ANALYSIS OF PROCESSES WITH TWO ACTIVE COMPONENTS AND ITS RAMIFICATIONS

To proceed with mode 2b we extend $(\mathcal{X}, \mathcal{Y}, \tau) = \sum_{k=1}^{\infty}(x_k, y_k)\varepsilon_{\tau_k}$ by adhering yet another, passive, component φ_k that will represent the number of new primary units arriving in the system in the period $(\tau_{k-1}, \tau_k]$ during which exactly one primary unit gets processed. This will make

$$(\mathcal{X}, \mathcal{Y}, \mathcal{P}, \tau) = \sum_{k=1}^{\infty}(x_k, y_k, \varphi_k)\varepsilon_{\tau_k} \tag{3.3.1}$$

a marked random measure with two active $[(\mathcal{X},\mathcal{Y}),$ with increments $(x_k,y_k)]$ and one passive $[\mathcal{P},$ with increments $\varphi_k]$ components, of which $(\mathcal{X},\mathcal{Y})$ are competing against each other under some priority for $\mathcal{A}$. The active components $(\mathcal{X},\mathcal{Y})$ will appreciate upon times $\tau_1,\tau_2,\ldots$ when approaching their thresholds Q_0 for $\mathcal{X}$ and S for $\mathcal{Y}$. The values of the passive component will be just registered upon exit from mode 2b. We recall that $x_1=x_2=\ldots=1$, whereas y_k is the number of secondary units processed in interval $(\tau_{k-1},\tau_k],$ and φ_k is the number of primary units that arrived in the system during the same interval. Notice that the values y_k's differ from v_k's (of section 4), because they are observed during different intervals.

This version of fluctuations is more complex compared with that in section 4 regarding mode 2a. In particular, we need to include an initial condition as per (3.3.1) inherited from the previous phase.

Let $\Delta_k=\tau_k-\tau_{k-1},k=1,2,\ldots.$ Then the random vectors $(x_k,y_k,\varphi_k,\Delta_k)$ are independent and identically distributed for $k=1,2,\ldots,$ with the common join transform

$$\gamma(u,v,w,\theta)=Eu^{x_1}v^{y_1}w^{\varphi_1}e^{-\theta\Delta_1},\ |u|\le 1,|v|\le 1,|w|\le 1,\mathrm{Re}\theta\ge 0. \quad (3.3.2)$$

We assume that the services times $s_1,s_2-s_1,s_3-s_2\ldots$ of the respective batches $v_1,v_2,v_3,\ldots$ of the secondary units are mutually independent and independent from $\Delta_1,\Delta_2,\ldots.$ Furthermore, during server's work on two parallel queues, the service times are different from those specified in mode 2a, now assuming $s_1,s_2-s_1,\ldots\in[\mathrm{Exp}(\mu)]$. We can say that $(\mathcal{V},\mathcal{S})=\sum_{j=1}^{\infty}v_j\varepsilon_{s_j}$ forms a marked Poisson process with position independent marking and unit intensity μ. For

notational convenience, the pgf of the batch sizes is still the same $b(z) = Ez^{\nu_1}$, $|z| \leq 1$.

With the notation

$$\gamma(\theta) = Ee^{-\theta\Delta_1}, \tag{3.3.3}$$

$$\gamma(u,v,w,\theta) = Eu^{x_1}v^{y_1}w^{\varphi_1}e^{-\theta\Delta_1} = uE\big[e^{-\theta\Delta_1}E[v^{y_1}|\Delta_1]E[w^{\varphi_1}|\Delta_1]\big]$$

$$= uE\big[e^{-\theta\Delta_1}e^{\lambda[a(w)-1]}e^{\mu[b(v)-1]}\big] = u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(v)]. \tag{3.3.4}$$

For $n = 1, 2, \ldots$ define the following sums:

$$\begin{aligned} X_n &= x_1 + \ldots + x_n \ \ (= n \text{ in our case}) \\ Y_n &= y_1 + \ldots + y_n \\ P_n &= \varphi_1 + \ldots + \varphi_n. \end{aligned} \tag{3.3.5}$$

We are interested in r.v.'s $X_\nu, Y_\nu,$ and P_ν (ν is the exit index defined in (3.1.7)) representing the number of primary units processed by τ_ν, secondary units processed by τ_ν, and primary units that entered the system by τ_ν, respectively. For this reason we introduce the functional

$$\Psi(u,v,w,\theta) = Eu^{X_\nu}v^{Y_\nu}w^{P_\nu}e^{-\theta\tau_\nu}, |u| \leq 1, |v| \leq 1, |w| \leq 1, \mathrm{Re}\theta \geq 0. \tag{3.3.6}$$

Next we introduce another version of the $\mathcal{D}$-operator applied to a function $\varphi : \mathbb{C}^3 \to \mathbb{C}$ analytic at $x = y = 0$ as

$$\mathcal{D}_{x,y}^{k,m}\varphi(x,y,z) =$$

$$= \begin{cases} \lim_{x\to 0, y\to 0} \frac{1}{k!m!} \frac{\partial^{k+m}}{\partial x^k \partial y^m} \left[\frac{1}{(1-x)(1-y)} \varphi(x,y,z) \right], & k,m \geq 0 \\ 0, & k<0 \text{ or } m<0. \end{cases} \tag{3.3.7}$$

Under the above notation and specifications, the following assertion is due to Dshalalow [21].

Theorem 3.3.1. *The functional Ψ satisfies the following formula*:

$$\Psi(u,v,w,\theta) = 1 - [1-\gamma(u,v,w,\theta)]\mathcal{D}_{x,y}^{Q_0-1,S-1} \frac{1}{1-\gamma(ux,vy,w,\theta)}, \tag{3.3.8}$$

where $\gamma(u,v,w,\theta) = u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(v)]$ *and* Q_0 *is set to equal* $i \in \{1,\dots,r-1\}$. □

To simplify formula (5.8) we note that

$$\mathcal{D}_{x,y}^{k,m} = \mathcal{D}_x^k \circ \mathcal{D}_y^m = \mathcal{D}_y^m \circ \mathcal{D}_x^k, \tag{3.3.9}$$

where the original $\mathcal{D}$-operator was introduced in (3.2.4).

From Dshalalow [20],

$$\mathcal{D}_x^{i-1} \frac{1}{1-\gamma(ux,vy,w,\theta)} = \mathcal{D}_x^{i-1} \frac{1}{1-ux\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}$$

$$= \frac{1-u^i\gamma^i[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}{1-u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}$$

herewith making (5.8) reduce to

$$\Psi(u,v,w,\theta) = Eu^{X_\nu}v^{Y_\nu}w^{P_\nu}e^{-\theta\tau_\nu}$$

$$= 1 - [1 - u\gamma[\theta + \lambda - \lambda a(w) + \mu - \mu b(v)]]$$
$$\times \mathcal{D}_y^{S-1} \tfrac{1-u^i\gamma^i[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}{1-u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(vy)]}. \tag{3.3.10}$$

Because for now we are not interested in the quantity of the secondary units Y_ν processed by the end of mode 2b, we can suppress it by letting $v = 1$ making Ψ reduce to

$$h(u, w, \theta) : = Eu^{X_\nu} w^{P_\nu} e^{-\theta\tau_\nu} = \Psi(u, 1, w, \theta)$$
$$= 1 - [1 - u\gamma[\theta + \lambda - \lambda a(w)]]$$
$$\times \mathcal{D}_y^{S-1} \tfrac{1-u^i\gamma^i[\theta+\lambda-\lambda a(w)+\mu-\mu b(y)]}{1-u\gamma[\theta+\lambda-\lambda a(w)+\mu-\mu b(y)]}. \tag{3.3.11}$$

$Eu^{X_\nu} w^{P_\nu} e^{-\theta\tau_\nu}$ is not exactly what we are looking for, since it does not give us the status of the system by the end of mode 2b, but we need

$$\Gamma_0^{(i)}(z, \theta) = Ez^{i-X_\nu+P_\nu} e^{-\theta\tau_\nu}, \tag{3.3.12}$$

to have the total number of the primary units by the end of mode 2a. This can be obtained from (3.3.11) through the following modification:

$$\Gamma_0^{(i)}(z, \theta) = z^i Ez^{-X_\nu} w^{P_\nu} e^{-\theta\tau_\nu}\big|_{w=\frac{1}{z}} = z^i h\big(\tfrac{1}{z}, z, \theta\big)$$
$$= z^i - [z - \gamma(\theta + \lambda - \lambda a(z))]$$
$$\times \mathcal{D}_y^{S-1} \tfrac{z^i-\gamma^i[\theta+\lambda-\lambda a(z)+\mu-\mu b(y)]}{z-\gamma[\theta+\lambda-\lambda a(z)+\mu-\mu b(y)]}, i = 1, \ldots, r-1. \tag{3.3.13}$$

Example 3.3.1. To illustrate formula (3.3.13) for the functional $\Gamma_0^{(i)}(z,\theta)$ consider the following special case. If $i=1$, clearly

$$\Gamma_0^{(1)}(z,\theta)=\gamma(\theta+\lambda-\lambda a(z)).$$

For the rest of $i=2,\ldots,r-1$, we apply the $\mathcal{D}$-operator and then simplify (3.3.13). As such, suppose that $\Delta_1\in[Exp(\gamma)],$ i.e. $\gamma(\theta)=\frac{\gamma}{\gamma+\theta}$. Then the expression inside the $\mathcal{D}$-operator can be modified as

$$\begin{aligned}\psi_i(y,z)&=\tfrac{z^i-\gamma^i[\theta+\lambda-\lambda a(z)+\mu-\mu b(y)]}{z-\gamma[\theta+\lambda-\lambda a(z)+\mu-\mu b(y)]}\\&=\sum_{j=0}^{i-1}\Big(\tfrac{\gamma}{\gamma+\theta+\lambda-\lambda a(z)+\mu-\mu b(y)}\Big)^j z^{i-1-j}.\end{aligned}\tag{3.3.14}$$

Assume further that $b(y)=\frac{py}{1-qy},$ that is, the server processes batches of secondary units distributed geometrically with parameter p and $q=1-p$. Then,

$$\psi_i(y,z)=\sum_{j=0}^{i-1}\Big(\tfrac{\gamma(1-qy)}{\gamma+\theta+\lambda-\lambda a(z)+\mu-((\gamma+\theta+\lambda-\lambda a(z))q+\mu)y}\Big)^j z^{i-1-j}.$$

With the notation

$$\begin{aligned}&c_q=(\gamma+\theta+\lambda-\lambda a(z))q+\mu\ \ (0<q<1)\ \ \&\\&c_1=\gamma+\theta+\lambda-\lambda a(z)+\mu\end{aligned}\tag{3.3.15}$$

we have

$$\begin{aligned}\psi_i(y,z)&=\sum_{j=0}^{i-1}\Big(\tfrac{\gamma(1-qy)}{c_1-c_qy}\Big)^j z^{i-1-j}\\&=\sum_{j=0}^{i-1}\Big(\tfrac{\gamma}{c_1}\Big)^j z^{i-1-j}\sum_{k=0}^{j}\tbinom{j}{k}(-q)^k y^k\tfrac{1}{\Big(1-\frac{c_q}{c_1}y\Big)^j}.\end{aligned}\tag{3.3.16}$$

Applying the property of the $\mathcal{D}$-operator that $\mathcal{D}_x^k(x^j g(x)) = \mathcal{D}_x^{k-j}(g(x))$ [20] and its linearity,

$$\mathcal{D}_y^{S-1}(\psi_i(y,z)) = \sum_{j=0}^{i-1}\left(\frac{\gamma}{c_1}\right)^j z^{i-1-j} \times \sum_{k=0}^{j}\binom{j}{k}(-q)^k \mathcal{D}_y^{S-1-k}\left(\frac{1}{\left(1-\frac{c_q}{c_1}y\right)^j}\right) \tag{3.3.17}$$

and then using property (vii) of the $\mathcal{D}$-operator from [20] we have

$$\mathcal{D}_y^{S-1}(\varphi_i(y,z)) = \sum_{j=0}^{i-1}\left(\frac{\gamma}{c_1}\right)^j z^{i-1-j} \times \sum_{k=0}^{j}\binom{j}{k}(-q)^k \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\left(\frac{c_q}{c_1}\right)^n. \tag{3.3.18}$$

Finally, from (3.3.13),

$$\Gamma_0^{(i)}(z,\theta) = z^i - \left[z - \frac{\gamma}{\gamma+\theta+\lambda-\lambda a(z)}\right]\sum_{j=0}^{i-1}\left(\frac{\gamma}{c_1}\right)^j z^{i-1-j}\sum_{k=0}^{j}\binom{j}{k}(-q)^k \times \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\left(\frac{c_q}{c_1}\right)^n. \tag{3.3.19}$$

For example, if $i = 2$, the latter expression reduces to

$$\Gamma_0^{(2)}(z,\theta) = z^2 - \left[z - \frac{\gamma}{\gamma+\theta+\lambda-\lambda a(z)}\right] \times \left[z + \left(\frac{\gamma}{c_1}\right)\left(p\frac{1-\left(\frac{c_q}{c_1}\right)^{S-1}}{1-\left(\frac{c_q}{c_1}\right)} + \left(\frac{c_q}{c_1}\right)^{S-1}\right)\right]. \qquad \square$$

Example 3.3.2. In the context of formula (3.3.19) of Example 3.3.1, consider the marginal distribution of the first passage time τ_ν, that is, for $z = 1$,

$$\Gamma_0^{(i)}(1,\theta) = Ee^{-\theta\tau_\nu} = 1 - \sum_{j=0}^{i-1}\gamma^j \sum_{k=0}^{j}\binom{j}{k}(-q)^k \times \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\frac{\theta((\gamma+\theta)q+\mu)^n}{(\gamma+\theta)(\gamma+\theta+\mu)^{n+j}}. \tag{3.3.20}$$

To find the mean of τ_ν, calculate

$$\begin{aligned} E\tau_\nu &= (-1)\frac{\partial}{\partial\theta}\left(\Gamma_0^{(i)}(1,\theta)\right)\Big|_{\theta=0} \\ &= \sum_{j=0}^{i-1}\gamma^j \sum_{k=0}^{j}\binom{j}{k}(-q)^k \\ &\quad \times \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\frac{\partial}{\partial\theta}\left(\frac{\theta((\gamma+\theta)q+\mu)^n}{(\gamma+\theta)(\gamma+\theta+\mu)^{n+j}}\right)\Big|_{\theta=0}. \end{aligned} \tag{3.3.21}$$

Here

$$\frac{\partial}{\partial\theta}\left(\frac{\theta((\gamma+\theta)q+\mu)^n}{(\gamma+\theta)(\gamma+\theta+\mu)^{n+j}}\right)\Big|_{\theta=0} = \frac{(\gamma q+\mu)^n}{\gamma(\gamma+\mu)^{n+j}}$$

implying that

$$E\tau_\nu = \frac{1}{\gamma}\sum_{j=0}^{i-1}\left(\frac{\gamma}{\gamma+\mu}\right)^j \sum_{k=0}^{j}\binom{j}{k}(-q)^k \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\left(\frac{\gamma q+\mu}{\gamma+\mu}\right)^n. \tag{3.3.22}$$

Setting $\gamma = \mu = 1$, $q = \frac{1}{2}$, then we have

$$E\tau_\nu = \sum_{j=0}^{i-1}\left(\frac{1}{2}\right)^j \sum_{k=0}^{j}\binom{j}{k}\left(-\frac{1}{2}\right)^k \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\left(\frac{3}{4}\right)^n. \tag{3.3.23}$$

□

Example 3.3.3. Continuing with Example 3.3.1, with $\theta = 0$, we arrive at the marginal distribution of the total number of primary units at the end of mode 2b.

$$\begin{aligned} \Gamma_0^{(i)}(z,0) = Ez^{i-X_\nu+P_\nu} &= z^i - \left[z - \tfrac{\gamma}{\gamma+\lambda-\lambda a(z)}\right] \\ &\quad \times \sum_{j=0}^{i-1}\left(\tfrac{\gamma}{\gamma+\lambda-\lambda a(z)+\mu}\right)^j z^{i-1-j}\sum_{k=0}^{j}\binom{j}{k}(-q)^k \\ &\quad \times \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\left(\tfrac{(\gamma+\lambda-\lambda a(z))q+\mu}{\gamma+\lambda-\lambda a(z)+\mu}\right)^n. \end{aligned} \tag{3.3.24}$$

Suppose the arriving batches have a geometric distribution with parameter $\alpha \in (0,1)$ to yield $a(z) = \frac{\alpha z}{1-\beta z}$ and have the relevant expressions inside (3.3.23) read

$$\frac{\gamma}{\gamma+\lambda-\lambda a(z)} = \frac{\gamma}{\gamma+\lambda\left(1-\frac{\alpha z}{1-\beta z}\right)} = \frac{\gamma(1-\beta z)}{\gamma+\lambda-(\gamma\beta+\lambda)z},$$

$$\frac{\gamma}{\gamma+\lambda-\lambda a(z)+\mu} = \frac{\gamma}{\gamma+\lambda\left(\frac{1-z}{1-\beta z}\right)+\mu} = \frac{\gamma(1-\beta z)}{\gamma+\mu+\lambda-((\gamma+\mu)\beta+\lambda)z},$$

$$\frac{(\gamma+\lambda-\lambda a(z))q+\mu}{\gamma+\lambda-\lambda a(z)+\mu} = \frac{\left(\gamma+\lambda\left(\frac{1-z}{1-\beta z}\right)\right)q+\mu}{\gamma+\lambda\left(\frac{1-z}{1-\beta z}\right)+\mu} = \frac{(\gamma+\lambda)q+\mu-((\gamma q+\mu)\beta+\lambda q)z}{\gamma+\mu+\lambda-((\gamma+\mu)\beta+\lambda)z}.$$

To find the mean primary queue length at the end of mode 2b, we take the derivative and then the limit as $z \to 1$ of the z-dependent terms,

$$\frac{\partial}{\partial z}\left[z^{i-1-j}\left(z - \tfrac{\gamma(1-\beta z)}{\gamma+\lambda-(\gamma\beta+\lambda)z}\right)\left(\tfrac{\gamma(1-\beta z)}{\gamma+\mu+\lambda-((\gamma+\mu)\beta+\lambda)z}\right)^j\left(\tfrac{(\gamma+\lambda)q+\mu-((\gamma q+\mu)\beta+\lambda q)z}{\gamma+\mu+\lambda-((\gamma+\mu)\beta+\lambda)z}\right)^n\right]\Bigg|_{z=1}$$

altogether reducing to

$$= \left(1 - \tfrac{\lambda}{\alpha\gamma}\right)\left(\tfrac{\gamma}{\gamma+\mu}\right)^j\left(\tfrac{\gamma q+\mu}{\gamma+\mu}\right)^n.$$

Finally,

$$E[i - X_\nu + P_\nu] = i - \Big(1 - \tfrac{\lambda}{\alpha\gamma}\Big)\sum_{j=0}^{i-1}\Big(\tfrac{\gamma}{\gamma+\mu}\Big)^j \sum_{k=0}^{j}\binom{j}{k}(-q)^k$$
$$\sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\Big(\tfrac{\gamma q+\mu}{\gamma+\mu}\Big)^n. \tag{3.3.25}$$

With $\gamma = \lambda = \mu = 1$ and $q = \alpha = \frac{1}{2}$ we have

$$E[i - X_\nu + P_\nu] = i + \sum_{j=0}^{i-1}\big(\tfrac{1}{2}\big)^j$$
$$\times \sum_{k=0}^{j}\binom{j}{k}\big(-\tfrac{1}{2}\big)^k \sum_{n=0}^{S-1-k}\binom{j+n-1}{n}\big(\tfrac{3}{4}\big)^n. \tag{3.3.26}$$

□

The figures below validate the special cases of Examples 3.3.3 and 3.3.2 showing their high accuracy by comparison of our results in (3.3.23) and (3.3.26) with those obtained by simulation.

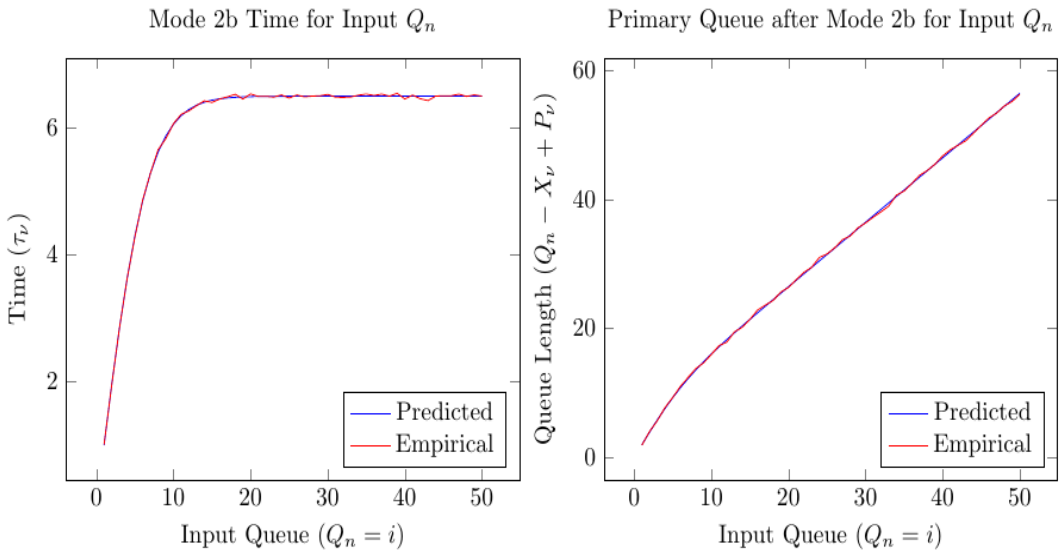


Figure 3.3.1: Predicted and empirical post-mode 2b time and queue length for $(\lambda, \mu, \gamma, p, \alpha, S) = \left(1, 1, 1, \frac{1}{2}, \frac{1}{2}, 10\right)$ for various pre-mode 2b queue lengths.

3.4. ANALYSIS OF MODE 3

Upon the beginning of mode 3, the primary queue may or may not have enough primary units available to resume mode 1 (required at least r units). To reduce the number of returns to mode 2, it is now required an even higher number of primary units. Namely, if there are N or more customers in the primary buffer, the server immediately resumes his service. Otherwise, he waits until the buffer accumulates to N or more units total (including some units that have been waiting or left behind after mode 2b) and only then does he resume service. Thus, mode 3 consists of two period types: first type is instantaneous and the second type is a random period dependent on the number of waiting units and specified by the new first passage time. A relevant formula from Dshalalow [17] instructs us how to chain the outcome of mode 2, using it as an initial value at the beginning of mode 2

expressed by functional $\Gamma_0^{(i)}(z,\theta)$, and then how to predict the outcome of the next phase.

Functional $\Gamma_0^{(i)}(z,\theta)$ specifies the distribution of the number (η_i) of primary units by the end of mode 2 (case a, when $i=0$ and case b when $0<i<r$) jointly with the duration of mode 2. If η_i is that number, then we have from the previous sections

$$\eta_i = \begin{cases} \Pi_\rho, & Q_0 = i = 0 \\ i - X_\nu + P_\nu, & 0 < Q_0 = i < r \end{cases} \tag{3.4.1}$$

with

$$\Gamma_0^{(i)}(z,\theta) = Ez^{\eta_i} e^{-\theta\tau_\nu},\ i = 0,\dots,r-1, \tag{3.4.2}$$

where

Π_ρ is the total number of primary units that enter the system during mode 2a.

P_ν is the total number of primary units that enter the system during mode 2b.

X_ν is the number of primary units processed from a total of i units upon beginning of mode 2b.

Now η_i can be less than N or it can be greater than or equal to N. We combine these two cases in one using fluctuation analysis. Define the *exit index*

$$\xi_i = \min\{n \geq 0 : \Sigma_i = \eta_i + \textstyle\sum_{j=1}^{n} \mathfrak{a}_j \geq N\}, \tag{3.4.3}$$

where $\mathfrak{a}_1, \mathfrak{a}_2, \ldots$ are the sizes of batches of units arriving at respective times $\tau_\nu + \delta_1 + \delta_2 + \ldots$ or $\tau_\rho + \delta_1 + \delta_2 + \ldots$ during mode 3, where δ's are interarrival times ($\delta_j = t_j - t_{j-1}$) of those groups of units, and with Σ_i being the total number of units by the end of mode 3. We set $\sum_{j=1}^{n} = 0$ if $n = 0$. In this case, Σ_i reduces to η_i.

Without loss of generality and for notational convenience, we enumerate those quantities by $1, 2, \ldots,$ although counted from $T = 0$, during mode 3, $\mathfrak{a}$'s and δ's must have different indices. Consequently, mode 3 ends at time

$$\mathcal{T}_i = \begin{cases} \tau_\rho + \delta_1 + \ldots + \delta_{\xi_0}, & i = 0 \\ \tau_\nu + \delta_1 + \ldots + \delta_{\xi_i}, & 0 < i < r \end{cases} \tag{3.4.4}$$

if counted from time zero. In the event $\eta_i \geq N, \mathcal{T}_i = \tau_\rho$ or τ_ν.
So with Σ_i being the number of primary units by the end of mode 3 and $\mathcal{T}_i$ - the time end of mode 3, define the functional

$$\alpha_i(z, \theta) = E z^{\Sigma_i} e^{-\theta \mathcal{T}_i}, i = 0, \ldots, r-1. \tag{3.4.5}$$

Then, given that mode 3 begins with the number of units specified by the functional $\Gamma_0^{(i)}(z)$ ($Q_0 = i = 0$ for mode 2a and $0 < Q_0 = i < r$ for mode 2b), using [20] we have

$$\alpha_i(z, \theta) = \Gamma_0^{(i)}(z, \theta) - \left[1 - E z^{\mathfrak{a}_1} e^{-\theta \delta_1}\right] \mathcal{D}_x^{N-1} \left(\frac{\Gamma_0^{(i)}(xz, \theta)}{1 - E(xz)^{\mathfrak{a}_1} e^{-\theta \delta_1}} \right), \tag{3.4.6}$$

where $Ez^{\mathfrak{a}_1}e^{-\theta\delta_1} = Ez^{\mathfrak{a}_1}Ee^{-\theta\delta_1} = a(z)\frac{\lambda}{\lambda+\theta}$ due to the position independent marking of the Poisson input process. Formula (3.4.6) thus can be rewritten as

$$\alpha_i(z,\theta) = \Gamma_0^{(i)}(z,\theta) - \big[1 - a(z)\tfrac{\lambda}{\lambda+\theta}\big]\mathcal{D}_x^{N-1}\phi_i(xz,\theta), i = 0,\dots,r-1, \tag{3.4.7}$$

where

$$\phi_i(xz,\theta) = \tfrac{\Gamma_0^{(i)}(xz,\theta)}{1-\frac{\lambda}{\lambda+\theta}a(xz)}. \tag{3.4.8}$$

The threshold $N\ (\geq 1)$ is what the primary queue needs to cross before the server resumes his work. In the event $\eta_i \geq N$ a.s., the $\mathcal{D}$-operator will automatically purge the expression in parentheses of (3.4.8) reducing $\alpha_i(z,\theta)$ to $\Gamma_0^{(i)}(z,\theta)$.

Example 3.4.1. We want to calculate $\alpha_i(z,\theta)$ under the assumptions of Example 5.1. From (3.3.19),

$$\begin{aligned}\Gamma_0^{(i)}(xz,\theta) = z^i - \Big[z - \tfrac{\gamma}{\gamma+\theta+\lambda-\lambda a(z)}\Big]\sum_{j=0}^{i-1}\Big(\tfrac{\gamma}{c_1(zx,\theta)}\Big)^j z^{i-1-j}\sum_{k=0}^{j}\tbinom{j}{k}(-q)^k \\ \times \sum_{l=0}^{S-1-k}\tbinom{j+l-1}{n}\Big(\tfrac{c_q(zx,\theta)}{c_1(zx,\theta)}\Big)^l\end{aligned} \tag{3.4.9}$$

where $c_s(zx,\theta) = (\gamma+\theta+\lambda-\lambda a(xz))s + \mu$ $(0 < s \leq 1,$ see (3.3.15)), so the term $\phi_i = \phi_i(xz,\theta)$ of (3.4.8) inside the $\mathcal{D}$-operator in (3.4.7) is

$$\begin{aligned}\phi_i &= \tfrac{\Gamma_0^{(i)}(xz,\theta)}{1-\frac{\lambda}{\lambda+\theta}a(xz)} \\ &= \tfrac{(xz)^i}{1-\frac{\lambda}{\lambda+\theta}a(xz)}\end{aligned} \tag{3.4.10}$$

$$
\begin{aligned}
&- \tfrac{xz}{1-\frac{\lambda}{\lambda+\theta}a(xz)} \sum_{j=0}^{i-1} \left(\tfrac{\gamma}{c_1(zx,\theta)} \right)^j (zx)^{i-1-j} \\
&\times \sum_{k=0}^{j} \binom{j}{k} (-q)^k \sum_{n=0}^{S-1-k} \binom{j+n-1}{n} \left(\tfrac{c_q(zx,\theta)}{c_1(zx,\theta)} \right)^n \qquad (3.4.11) \\
&+ \tfrac{\gamma}{\gamma+\theta+\lambda-\lambda a(xz)} \tfrac{1}{1-\frac{\lambda}{\lambda+\theta}a(xz)} \\
&\times \sum_{j=0}^{i-1} \left(\tfrac{\gamma}{c_1(zx,\theta)} \right)^j (zx)^{i-1-j} \sum_{k=0}^{j} \binom{j}{k} (-q)^k \\
&\times \sum_{l=0}^{S-1-k} \binom{j+l-1}{l} \left(\tfrac{c_q(zx,\theta)}{c_1(zx,\theta)} \right)^l . \qquad (3.4.12)
\end{aligned}
$$

We apply the $\mathcal{D}$-operator separately to (3.4.10), then to (3.4.11), and lastly to (3.4.12). The expression in (3.4.10) can be rewritten as

$$
\begin{aligned}
&\tfrac{(xz)^i}{1-\frac{\lambda}{\lambda+\theta}\frac{\alpha zx}{1-\beta zx}} \qquad (3.4.13) \\
&= \tfrac{z^i x^i (1-\beta zx)}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right) zx} . \qquad (3.4.14)
\end{aligned}
$$

Thus, the $\mathcal{D}$-operator of (3.4.10, 3.4.14) is

$$
\begin{aligned}
&\mathcal{D}_x^{N-1} \left(\tfrac{(xz)^i}{1-\frac{\lambda}{\lambda+\theta}\frac{\alpha zx}{1-\beta zx}} \right) = z^i \mathcal{D}_x^{N-1-i} \left(\tfrac{1-\beta zx}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right) zx} \right) \\
&= z^i \mathcal{D}_x^{N-1-i} \left(\tfrac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right) zx} \right) - \beta z^{i+1} \mathcal{D}_x^{N-1-i-1} \left(\tfrac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right) zx} \right) \\
&= z^i \sum_{j=0}^{N-1-i} \left(\tfrac{\lambda+\beta\theta}{\lambda+\theta} \right)^j z^j - \beta z^{i+1} \sum_{j=0}^{N-1-i-1} \left(\tfrac{\lambda+\beta\theta}{\lambda+\theta} \right)^j z^j . \qquad (3.4.15)
\end{aligned}
$$

Next, we focus on (3.4.11). The $\mathcal{D}$-operator will be interchanged with various finite sums in (3.4.11) by its linearity and will actually need to be applied to the following.

$$
\begin{aligned}
&\tfrac{1}{1-\frac{\lambda}{\lambda+\theta}\frac{\alpha z x}{1-\beta z x}}\,(zx)^{i-j}c_q(zx,\theta)^l\Big(\tfrac{1}{c_1(zx,\theta)}\Big)^{j+l}\\
&=\tfrac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}(zx)^{i-j}\tfrac{1}{(\gamma+\theta+\lambda)^{j+l}}\\
&\quad\times\sum_{m=0}^{l}\tbinom{l}{m}k_2(\theta)^m(zx)^m k_1(\theta)^{l-m}\tfrac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l}}\\
&\quad\times\sum_{n=0}^{j+1}\tbinom{j+1}{n}(-\beta z x)^n, \qquad (3.4.16)
\end{aligned}
$$

where

$$
k_1(\theta)=(\gamma+\theta+\lambda)q+\mu \ \text{ and } \ k_2(\theta)=(\beta(q(\gamma+\theta)\ +\ \mu)+\lambda q). \quad (3.4.17)
$$

Applying the $\mathcal{D}$-operator to the x terms, we find

$$
\begin{aligned}
&\mathcal{D}_x^{N-1}\left(x^{i-j+m+n}\tfrac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}\,\tfrac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l}}\right)\\
&=\mathcal{D}_x^{N-1-i+j-m-n}\left(\tfrac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}\,\tfrac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l}}\right)\\
&=\sum_{r=0}^{N-1-i+j-m-n}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^r z^r\mathcal{D}^{N-1-i+j-m-n-r}\left(\tfrac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l}}\right)\\
&=\sum_{r=0}^{N-1-i+j-m-n}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^r z^r\sum_{s=0}^{N-1-i+j-m-n-r}\tbinom{j+l+s-1}{s}\left(\tfrac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)^s z^s.
\end{aligned}
$$

Lastly, we focus on (3.4.12). The $\mathcal{D}$-operator will be interchanged with various finite series in (3.4.12) by linearity and will actually need to be applied to the following.

$$\frac{\gamma}{\gamma+\theta+\lambda-\lambda a(xz)}(zx)^{i-1-j}\frac{1}{1-\frac{\lambda}{\lambda+\theta}a(xz)}c_q(zx,\theta)^l\left(\frac{1}{c_1(zx,\theta)}\right)^{j+l}$$

$$=\frac{\gamma}{(\gamma+\theta+\lambda)^{j+l+1}}\frac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}(zx)^{i-1-j}$$

$$\times\sum_{m=0}^{l}\binom{l}{m}k_2(\theta)^m(zx)^m k_1(\theta)^{l-m}\frac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l+1}}$$

$$\times\sum_{n=0}^{j+2}\binom{j+2}{n}(-\beta zx)^n,$$

The x-dependent terms in (3.4.12) are

$$x^{i-1-j+m+n}\frac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}\frac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l+1}}.$$

Applying the $\mathcal{D}$-operator,

$$\mathcal{D}^{N-1}\left(x^{i-1-j+m+n}\frac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}\frac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l+1}}\right)$$

$$=\mathcal{D}^{N-i+j-m-n}\left(\frac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}\frac{1}{\left(1-\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)zx\right)^{j+l+1}}\right)$$

$$=\sum_{r=0}^{N-i+j-m-n}\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)^r z^r\sum_{s=0}^{N-i+j-m-n-r}\binom{j+l+s}{s}\left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)^s z^s.$$

Altogether,

$$\alpha_i(z,\theta)=\Gamma_0^{(i)}(z,\theta)-\left[1-a(z)\frac{\lambda}{\lambda+\theta}\right]\mathcal{D}_x^{N-1}(\psi_i)$$

$$= \Gamma_0^{(i)}(z,\theta) - \left[1 - a(z)\tfrac{\lambda}{\lambda+\theta}\right]\Bigg[z^i \sum_{j=0}^{N-1-i}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^j z^j - \beta z^{i+1}\sum_{j=0}^{N-1-i-1}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^j z^j$$

$$- \sum_{j=0}^{i-1}\gamma^j \sum_{k=0}^{j}\binom{j}{k}(-q)^k \sum_{l=0}^{S-1-k}\binom{j+l-1}{l}\tfrac{1}{(\gamma+\theta+\lambda)^{j+l}}\sum_{m=0}^{l}\binom{l}{m}k_2(\theta)^m k_1(\theta)^{l-m}$$

$$\times \sum_{n=0}^{j+1}\binom{j+1}{n}(-\beta z)^n \sum_{r=0}^{N-1-i+j-m-n}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^r$$

$$\times \sum_{s=0}^{N-1-i+j-m-n-r}\binom{j+l+s-1}{s}\left(\tfrac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)^s z^{i-j+m+n+r+s}$$

$$+ \tfrac{\gamma}{\gamma+\theta+\lambda}\sum_{j=0}^{i-1}\gamma^j \sum_{k=0}^{j}\binom{j}{k}(-q)^k \sum_{l=0}^{S-1-k}\binom{j+l-1}{l}\tfrac{1}{(\gamma+\theta+\lambda)^{j+l}}$$

$$\times \sum_{m=0}^{l}\binom{l}{m}k_2(\theta)^m k_1(\theta)^{l-m}\sum_{n=0}^{j+2}\binom{j+2}{n}(-\beta)^n \sum_{r=0}^{N-i+j-m-n}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^r$$

$$\times \sum_{s=0}^{N-i+j-m-n-r}\binom{j+l+s}{s}\left(\tfrac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)^s z^{i-1-j+m+n+r+s}\Bigg]. \tag{3.4.18}$$

The marginal transform of $\mathcal{T}_i$ is

$$Ee^{-\theta\mathcal{T}_i} = \alpha_i(1,\theta)$$

$$= E\left[e^{-\theta\tau_\nu}\right] - \tfrac{\theta}{\lambda+\theta}\Bigg[\sum_{j=0}^{N-1-i}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^j - \beta \sum_{j=0}^{N-i-2}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^j$$

$$- \sum_{j=0}^{i-1}\gamma^j \sum_{k=0}^{j}\binom{j}{k}(-q)^k \sum_{l=0}^{S-1-k}\binom{j+l-1}{l}\tfrac{1}{(\gamma+\theta+\lambda)^{j+l}}\sum_{m=0}^{l}\binom{l}{m}k_2(\theta)^m k_1(\theta)^{l-m}$$

$$\times \sum_{n=0}^{j+1} \binom{j+1}{n} (-\beta)^n \sum_{r=0}^{N-1-i+j-m-n} \left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)^r$$

$$\times \sum_{s=0}^{N-1-i+j-m-n-r} \binom{j+l+s-1}{s} \left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)^s$$

$$+ \frac{\gamma}{\gamma+\theta+\lambda} \sum_{j=0}^{i-1} \gamma^j \sum_{k=0}^{j} \binom{j}{k} (-q)^k \sum_{l=0}^{S-1-k} \binom{j+l-1}{l} \frac{1}{(\gamma+\theta+\lambda)^{j+l}}$$

$$\times \sum_{m=0}^{l} \binom{l}{m} k_2(\theta)^m k_1(\theta)^{l-m} \sum_{n=0}^{j+2} \binom{j+2}{n} (-\beta)^n \sum_{r=0}^{N-i+j-m-n} \left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)^r$$

$$\times \sum_{s=0}^{N-i+j-m-n-r} \binom{j+l+s}{s} \left(\frac{\beta(\gamma+\theta)+\lambda}{\gamma+\theta+\lambda}\right)^s \Bigg]. \tag{3.4.19}$$

Then the mean is

$$E\mathcal{T}_i = E[\tau_\nu] - \frac{1}{\lambda} \Bigg[\alpha(N-i-2)+1$$

$$+ \sum_{j=0}^{i-1} \left(\frac{\gamma}{\gamma+\lambda}\right)^j \sum_{k=0}^{j} \binom{j}{k} (-q)^k \sum_{l=0}^{S-1-k} \binom{j+l-1}{l} \frac{1}{(\gamma+\lambda)^l}$$

$$\times \sum_{m=0}^{l} \binom{l}{m} k_2(0)^m k_1(0)^{l-m} \sum_{n=0}^{j+1} \binom{j+1}{n} (-\beta)^n$$

$$\times \sum_{r=0}^{N-1-i+j-m-n} \sum_{s=0}^{N-1-i+j-m-n-r} \binom{j+l+s-1}{s} \left(\frac{\beta\gamma+\lambda}{\gamma+\lambda}\right)^s$$

$$- \left(\frac{\gamma}{\gamma+\lambda}\right) \sum_{j=0}^{i-1} \left(\frac{\gamma}{\gamma+\lambda}\right)^j \sum_{k=0}^{j} \binom{j}{k} (-q)^k$$

$$\times \sum_{l=0}^{S-1-k} \binom{j+l-1}{l} \frac{1}{(\gamma+\lambda)^l} \sum_{m=0}^{l} \binom{l}{m} k_2(0)^m k_1(0)^{l-m}$$

$$\times \sum_{n=0}^{j+2} \binom{j+2}{n} (-\beta)^n \sum_{r=0}^{N-i+j-m-n} \sum_{s=0}^{N-i+j-m-n-r} \binom{j+l+s}{s} \left(\frac{\beta\gamma+\lambda}{\gamma+\lambda}\right)^s \Bigg]. \tag{3.4.20}$$

The marginal transform of Σ_i is

$$Ez^{\Sigma_i} = \alpha_i(z,0)$$

$$= E\left[z^{i-X_\nu+P_\nu}\right] - [1-a(z)]\left[\sum_{j=0}^{N-1-i} z^{i+j} - \beta \sum_{j=0}^{N-1-i-1} z^{i+1+j}\right.$$

$$- \sum_{j=0}^{i-1}\gamma^j \sum_{k=0}^{j} \binom{j}{k} (-q)^k \sum_{l=0}^{S-1-k} \binom{j+l-1}{l} \frac{1}{(\gamma+\lambda)^{j+l}}$$

$$\times \sum_{m=0}^{l} \binom{l}{m} k_2(0)^m k_1(0)^{l-m} \sum_{n=0}^{j+1} \binom{j+1}{n}$$

$$\times (-\beta z)^n \sum_{r=0}^{N-1-i+j-m-n} \sum_{s=0}^{N-1-i+j-m-n-r} \binom{j+l+s-1}{s} \left(\frac{\beta\gamma+\lambda}{\gamma+\lambda}\right)^s z^{i-j+m+n+r+s}$$

$$+ \frac{\gamma}{\gamma+\lambda} \sum_{j=0}^{i-1}\gamma^j \sum_{k=0}^{j} \binom{j}{k} (-q)^k \sum_{l=0}^{S-1-k} \binom{j+l-1}{l} \frac{1}{(\gamma+\lambda)^{j+l}} \sum_{m=0}^{l} \binom{l}{m} k_2(0)^m k_1(0)^{l-m}$$

$$\times \sum_{n=0}^{j+2} \binom{j+2}{n} (-\beta)^n \sum_{r=0}^{N-i+j-m-n}$$

$$\left.\times \sum_{s=0}^{N-i+j-m-n-r} \binom{j+l+s}{s} \left(\frac{\beta\gamma+\lambda}{\gamma+\lambda}\right)^s z^{i-1-j+m+n+r+s}\right]. \tag{3.4.21}$$

Then the mean of Σ_i is

$$E\Sigma_i = E[i - X_\nu + P_\nu] + a\Bigg[\alpha(N-i-2)+1$$

$$-\sum_{j=0}^{i-1}\gamma^j\sum_{k=0}^{j}\binom{j}{k}(-q)^k\sum_{l=0}^{S-1-k}\binom{j+l-1}{l}\frac{1}{(\gamma+\lambda)^{j+l}}\sum_{m=0}^{l}\binom{l}{m}k_2(0)^m k_1(0)^{l-m}$$

$$\times\sum_{n=0}^{j+1}\binom{j+1}{n}(-\beta)^n\sum_{r=0}^{N-1-i+j-m-n}\sum_{s=0}^{N-1-i+j-m-n-r}\binom{j+l+s-1}{s}\left(\frac{\beta\gamma+\lambda}{\gamma+\lambda}\right)^s$$

$$+\frac{\gamma}{\gamma+\lambda}\sum_{j=0}^{i-1}\gamma^j\sum_{k=0}^{j}\binom{j}{k}(-q)^k\sum_{l=0}^{S-1-k}\binom{j+l-1}{l}\frac{1}{(\gamma+\lambda)^{j+l}}$$

$$\times\sum_{m=0}^{l}\binom{l}{m}k_2(0)^m k_1(0)^{l-m}\sum_{n=0}^{j+2}\binom{j+2}{n}(-\beta)^n$$

$$\times\sum_{r=0}^{N-i+j-m-n}\sum_{s=0}^{N-i+j-m-n-r}\binom{j+l+s}{s}\left(\frac{\beta\gamma+\lambda}{\gamma+\lambda}\right)^s\Bigg]. \tag{3.4.22}$$

□

Example 3.4.2. We revisit Example 3.2.1 with $\Gamma_0^{(0)}(z,\theta)$ in order to calculate $\alpha_0(z,\theta)$, the joint transform of the queue length by the end of mode 3 and the total duration of modes 2a and 3. By formula (3.2.12),

$$\Gamma_0^{(0)}(z,\theta) = \zeta(\theta+\lambda-\lambda a(z))[p\zeta(\theta+\lambda-\lambda a(z))+q]^{S-1}.$$

To continue we make the following assumptions on ζ and $a(z)$. We have previously set service times of secondary units exponential with parameter μ. Further, assume that $a(z) = \frac{\alpha z}{1-\beta z}$, i.e. arrivals come in geometric batches with parameter α (and $\beta = 1-\alpha$). Thus,

$$\zeta(\theta+\lambda-\lambda a(z)) = \frac{\mu}{\mu+\theta+\lambda-\lambda a(z)} = \frac{\mu}{\mu+\theta+\lambda\left(1-\frac{\alpha z}{1-\beta z}\right)}$$

$$= \frac{\mu}{\mu+\theta+\lambda\left(\frac{1-z}{1-\beta z}\right)} = \frac{\mu(1-\beta z)}{(\mu+\theta)(1-\beta z)+\lambda(1-z)}$$

$$= \frac{\mu(1-\beta z)}{\mu+\theta+\lambda-((\mu+\theta)\beta+\lambda)z} = \frac{\mu}{\mu+\theta+\lambda}\frac{1-\beta z}{1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}z}. \tag{3.4.23}$$

Then,

$$\Gamma_0^{(0)}(z,\theta) = \frac{\mu}{\mu+\theta+\lambda}\frac{1-\beta z}{1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}z}\left[p\left(\frac{\mu}{\mu+\theta+\lambda}\frac{1-\beta z}{1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}z}\right)+q\right]^{S-1}. \tag{3.4.7}$$

This allows us to calculate $\alpha_0(z,\theta)$ by formulas (3.4.7-3.4.8). First,

$$\alpha_0(z,\theta) = \Gamma_0^{(0)}(z,\theta) - \left[1-a(z)\frac{\lambda}{\lambda+\theta}\right]\mathcal{D}_x^{N-1}(\phi_0)$$

where

$$\phi_0 = \frac{\Gamma_0^{(0)}(xz,\theta)}{1-\frac{\lambda}{\lambda+\theta}a(xz)} = \frac{\mu}{\mu+\theta+\lambda}\frac{1-\beta zx}{1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx}$$

$$\times\left[\frac{p\mu}{\mu+\theta+\lambda}\frac{1-\beta zx}{1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx}+q\right]^{S-1}\frac{1-\beta zx}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}.$$

Then

$$\left[p\left(\frac{\mu}{\mu+\theta+\lambda}\frac{1-\beta zx}{1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx}\right)+q\right]^{S-1}$$

$$= \sum_{j=0}^{S-1}\binom{S-1}{j}\left(\frac{p\mu}{\mu+\theta+\lambda}\right)^j q^{S-1-j}\left(\frac{1-\beta zx}{1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx}\right)^j.$$

implying that

$$\phi_0 = \frac{\mu}{\mu+\theta+\lambda}\sum_{j=0}^{S-1}\binom{S-1}{j}\left(\frac{p\mu}{\mu+\theta+\lambda}\right)^j q^{S-1-j}$$

$$\times \frac{1}{\left(1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx\right)^{j+1}}\frac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}(1-\beta zx)^{j+2}$$

$$= \frac{\mu}{\mu+\theta+\lambda}\sum_{j=0}^{S-1}\binom{S-1}{j}\left(\frac{p\mu}{\mu+\theta+\lambda}\right)^j q^{S-1-j}$$

$$\times \sum_{k=0}^{j+2}\binom{j+1}{k}(-\beta z)^k x^k \frac{1}{\left(1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx\right)^{j+1}}\frac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}.$$

Applying the operator $\mathcal{D}_x^{N-1}$ to the x terms we get

$$\mathcal{D}_x^{N-1}(\psi_0) = \mathcal{D}_x^{N-1-k}\left(\frac{1}{\left(1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx\right)^{j+1}}\frac{1}{1-\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)zx}\right)$$

$$= \sum_{l=0}^{N-1-k}\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)^l \mathcal{D}^{N-1-k-l}\left(\frac{1}{\left(1-\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}zx\right)^{j+1}}\right)$$

$$= \sum_{l=0}^{N-1-k}\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)^l \sum_{m=0}^{N-1-k-l}\binom{j+m}{m}\left(\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}\right)^m.$$

Altogether,

$$\alpha_0(z,\theta) = \Gamma_0^{(0)}(z,\theta) - \left[1-a(z)\frac{\lambda}{\lambda+\theta}\right]\frac{\mu}{\mu+\theta+\lambda}\sum_{j=0}^{S-1}\binom{S-1}{j}\left(\frac{p\mu}{\mu+\theta+\lambda}\right)^j q^{S-1-j}$$

$$\times \sum_{k=0}^{j+2}\binom{j+2}{k}(-\beta z)^k \sum_{l=0}^{N-1-k}\left(\frac{\lambda+\beta\theta}{\lambda+\theta}\right)^l \sum_{m=0}^{N-1-k-l}\binom{j+m}{m}\left(\frac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}\right)^m.$$

The marginal transform of $\mathcal{T}_0$ is

$$Ee^{-\theta\mathcal{T}_0} = \alpha_0(1,\theta) = E\left[e^{-\theta\tau_\rho}\right] - \tfrac{\theta}{\lambda+\theta}\tfrac{\mu}{\mu+\theta+\lambda}\sum_{j=0}^{S-1}\binom{S-1}{j}\left(\tfrac{p\mu}{\mu+\theta+\lambda}\right)^j q^{S-1-j}$$

$$\times\sum_{k=0}^{j+2}\binom{j+2}{k}(-\beta)^k\sum_{l=0}^{N-1-k}\left(\tfrac{\lambda+\beta\theta}{\lambda+\theta}\right)^l\sum_{m=0}^{N-1-k-l}\binom{j+m}{m}\left(\tfrac{(\mu+\theta)\beta+\lambda}{\mu+\theta+\lambda}\right)^m.$$

Thus,

$$E\mathcal{T}_0 = E[\tau_\rho] + \tfrac{\mu}{\lambda(\lambda+\mu)}\sum_{j=0}^{S-1}\binom{S-1}{j}\left(\tfrac{p\mu}{\lambda+\mu}\right)^j q^{S-1-j}\sum_{k=0}^{j+2}\binom{j+2}{k}$$

$$\times(-\beta)^k(N-1-k)\sum_{m=0}^{N-1-k-l}\binom{j+m}{m}\left(\tfrac{\lambda+\beta\mu}{\lambda+\mu}\right)^m.$$

Then, the marginal transform of Σ_0 is

$$Ez^{\Sigma_0} = \alpha_0(z,0) = E\left[z^{\Pi_\rho}\right] - (1-a(z))\tfrac{\mu}{\mu+\lambda}\sum_{j=0}^{S-1}\binom{S-1}{j}\left(\tfrac{p\mu}{\mu+\lambda}\right)^j q^{S-1-j}$$

$$\times\sum_{k=0}^{j+2}\binom{j+2}{k}(-\beta)^k z^k(N-1-k)\sum_{m=0}^{N-1-k-l}\binom{j+m}{m}\left(\tfrac{\lambda+\beta\mu}{\lambda+\mu}\right)^m$$

implying that the mean of Σ_0 is

$$E\Sigma_0 = E[\Pi_\rho] + a\tfrac{\mu}{\mu+\lambda}\sum_{j=0}^{S-1}\binom{S-1}{j}\left(\tfrac{p\mu}{\mu+\lambda}\right)^j q^{S-1-j}$$

$$\times\sum_{k=0}^{j+2}\binom{j+2}{k}(-\beta)^k(N-1-k)\sum_{m=0}^{N-1-k-l}\binom{j+m}{m}\left(\tfrac{\lambda+\beta\mu}{\lambda+\mu}\right)^m.$$

As expected, we have $E\Sigma_0 = \lambda a E\mathcal{T}_0$. □

3.5. QUEUEING PROCESS

Recall that $Q(t)$ is the number of all primary units in the system at time $t \geq 0$, defined as a piecewise linear process with right-continuous paths. The sequence $T_0 = 0, T_1, T_2, \ldots$ of successive service cycle completions is a sequence of stopping times relative to the filtration $(\mathcal{F}_t)$, upon which $Q(t)$ conditionally regenerates forming a semi-regenerative process. If $Q_{n-1} = Q(T_{n-1}) \geq r$, the next service begins immediately and lasts $\sigma_n \in [\sigma]$ being arbitrarily distributed. Whenever $Q_{n-1} \geq r$, from T_{n-1} to T_n, the system is in mode 1. So given $Q_{n-1} \geq r$, the transitions from T_{n-1} to T_n are very similar to that of the M/G/1-queue, namely,

$$\begin{aligned} P_i(z) :&= E\left[z^{Q_1} | Q_0 = i\right] = E\left[z^{(i-R)^+} z^{W_1}\right] \\ &= z^{(i-R)^+} \beta(\lambda - \lambda a(z)), i \geq r, \end{aligned} \tag{3.5.1}$$

where

$$\beta(\theta) :\ = Ee^{-\theta\sigma} \text{ (with } b = E\sigma) \tag{3.5.2}$$

is the LST of service time σ and W_1 is the number of customers that enter the system during that service period. As far as parameters R and N, mentioned in

section 1, we consider two different versions of the system, under $r \leq R \leq N$ (referred to as *model* 1) and $r \leq N \leq R$ (*model* 2).

Model 1. $\boldsymbol{R \leq N}$. When $Q_{n-1} < r,$ the system enters mode 2 (versions a or b), with the server servicing two queues or one queue as per sections 3-6, followed by mode 3. Under mode 3, the server waits for the primary buffer content to replenish to N or more customers or he skips the wait if that number is available right upon his exit from mode 2. In either case the server processes a batch of R units and ends an underlying service cycle. Due to the above considerations regarding the semi-regenerative nature of $Q(t)$ with respect to the sequence $\{T_n\},$ $\{Q_n\}$ is a homogeneous Markov chain, specified by the transitions

$$Q_1 = \begin{cases} \Sigma_{Q_0} - R + W_1, & 0 \leq Q_0 < r \\ (Q_0 - R)^+ + W_1, & Q_0 \geq r \end{cases} \tag{3.5.3}$$

where W_1 is the number of primary units arriving at the system during service of a batch of customers (from r to R in mode 1 and R after mode 3), and Σ_i was introduced in (3.4.3).

For $Q_0 = i < r,$ we use formulas (3.4.7-3.4.8) that give the number of customers in the system by the end of mode 3. It will be integrated into the conditional expectation like (3.5.1) as follows.

$$\begin{aligned} P_i(z) &= E\left[z^{Q_1} | Q_0 = i\right] = z^{-R} E z^{\Sigma_i} E z^{W_1} \\ &= z^{-R} \alpha_i(z) \beta(\lambda - \lambda a(z)), i < r, \end{aligned} \tag{3.5.4}$$

where $\alpha_i(z) := \alpha_i(z,0)$ is the marginal functional of $\alpha_i(z,\theta)$:

$$\alpha_i(z) = Ez^{\Sigma_i} = \Gamma_0^{(i)}(z,0) - [1-a(z)] \times \mathcal{D}_x^{N-1} \tfrac{\Gamma_0^{(i)}(xz,0)}{1-a(xz)}, i = 0,\dots,r-1, \tag{3.5.5}$$

where

$$\Gamma_0^{(0)}(z,\theta) = 1 - [1-\zeta(\lambda-\lambda a(z))]\mathcal{D}_y^{S-1} \tfrac{1}{1-b(y)\zeta(\lambda-\lambda a(z))}, \tag{3.5.6}$$

$$\Gamma_0^{(i)}(z,0) = z^i - [z-\gamma(\lambda-\lambda a(z))]\mathcal{D}_y^{S-1} \times \tfrac{z^i-\gamma^i[\lambda-\lambda a(z)+\mu-\mu b(y)]}{z-\gamma[\lambda-\lambda a(z)+\mu-\mu b(y)]}, i = 1,\dots,r-1. \tag{3.5.7}$$

are reduced from formulas (3.2.10) and (3.3.13).

From (3.5.3), with the notation $p_{ij} = P\{Q_1 = j | Q_0 = i\}$ we have

$$p_{ij} = \begin{cases} P\{W_1 = j-(i-R)^+\} = \begin{cases} q_{j-(i-R)^+}, & j \ge i-r \\ 0, & j < i-r \end{cases}, & i \ge r \\ P\{\Sigma_i + W_1 = j+R\} = \sum_{k=N}^{j+R} \underbrace{P\{W_1 = j-k+R\}}_{q_{j-k+R}} \underbrace{P\{\Sigma_i = k\}}_{\alpha_k}, & i < r, j \ge N-R \\ 0, & i < r, j < N-R \end{cases} \tag{3.5.8}$$

where $q_k = P\{W_1 = k\} > 0, k = 0,1,\dots,$ is the probability distribution of the number of arriving primary units in the system during a service time of a primary unit that can be obtained from the expansion of $\beta(\lambda-\lambda a(z))$, whereas $\alpha_k > 0, k = N, N+1,\dots, (\alpha_k = 0, k < N)$ is the probability distribution of

arriving primary units during mode 3 and it can be derived from the expansion of the functional $\alpha_i(z)$ of formulas (3.5.7.5-3.5.7).

Hence the transition probability matrix P reads

$$P = \begin{pmatrix} p_{00}=0 & 0 & \dots & p_{0,N-R-1}=0 & p_{0,N-R}>0 & \dots\dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{r-1,0}=0 & 0 & \dots & p_{r-1,N-R-1}=0 & p_{r-1,N-R}>0 & \dots\dots \\ q_0 & q_1 & \dots & q_{N-R-1} & q_{N-R} & \dots\dots \\ 0 & q_0 & \dots & q_{N-R-2} & q_{N-R-1} & \dots\dots \\ 0 & 0 & \dots & q_{N-R-3} & q_{N-R-2} & \dots\dots \\ 0 & 0 & \dots & q_{N-R-4} & q_{N-R-3} & \dots\dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{3.5.9}$$

The left upper corner block of this matrix (rows from 0 to $r-1$ and columns from 0 to $N-R-1$) is a zero matrix. The rest of the elements in the rows 0 to $r-1$ are strictly positive. The Markov chain $\{Q_n\}$ is irreducible. Indeed, for $N \leq 2R$, it is readily seen that P^2 already has all elements $p_{ij}^2 > 0$ for $0 \leq i < r$ and $j = 0, 1, \dots$. In can be readily validated that for $kR < N < (k+1)R$ and $k \geq 2$, P^k has all elements $p_{ij}^k > 0$ for $0 \leq i < r$ and $j = 0, 1, \dots$. Matrix P of (7.9) is referred to (Abolnikov and Dukhovny [1]) as a Δ_{r+1}-*matrix*. From (7.9), we easily deduce that the Markov chain $\{Q_n\}$ is aperiodic. By Abolnikov-Dukhovny [1], it is recurrent positive if and only if $P_r'(1) < r$ (where $P_i(z)$ is the generating function of the ith row of P). The latter condition reduces to

$$L = ab\lambda < r, \tag{3.5.10}$$

where a is the mean batch size of arriving primary units and b is the mean service time of a batch of primary units in mode 1. L is often referred to as the *offered load* although this notion applies to more basic systems.

The pgf $P(z)$ of the invariant probability measure $\mathbf{p} = (p_0, p_1, \dots)$ of $\{Q_n\}$ (under condition (3.5.10)) can be easily found from (3.5.1-3.5.3) and equation $P(z) = \sum_{i=0}^{\infty} p_i P_i(z)$:

$$P(z) = \beta(\lambda - \lambda a(z)) \frac{\sum_{i=0}^{r-1} p_i[\alpha_i(z) - z^i] + \sum_{i=r}^{R-1} p_i(z^R - z^i)}{z^R - \beta(\lambda - \lambda a(z))}. \tag{3.5.11}$$

$P(z)$ can be seen as

$$P(z) = \beta(\lambda - \lambda a(z)) \frac{N(z)}{D(z)}, \tag{3.5.12}$$

where

$$N(z) = \sum_{i=0}^{r-1} p_i[\alpha_i(z) - z^i] + \sum_{i=r}^{R-1} p_i(z^R - z^i) \tag{3.5.13}$$

and

$$D(z) = z^R - \beta(\lambda - \lambda a(z)). \tag{3.5.14}$$

Now according to Abolnikov and Dukhovny [3], the denominator $D(z)$ has exactly R roots in the closed disk $\overline{B}(0, 1)$ of which root $z_0 = 1$. Furthermore, according to Dukhovny [26], all roots on the boundary $\partial B(0, 1)$ are simple. Since $N(z)$ is analytic inside $B(0, 1)$ and continuous on the boundary $\partial B(0, 1)$, all roots

$z_1, \ldots, z_{R-1}$ of the denominator are the roots of the numerator $N(z)$ counting their possible multiplicities. We deal with root $z_0 = 1$ separately from the condition $P(1) = 1$. Altogether, there are R conditions in R unknown probabilities $p_0, \ldots, p_{R-1}$.

Assuming all roots distinct (which is the most likely scenario) and proceeding with the substitution of the root z_j into $N(z)$ we have

$$N(z_j) = \sum_{i=0}^{R-1} p_i a_{ij} = 0, j = 1, \ldots, R-1 \tag{3.5.15}$$

where

$$a_{ij} = \begin{cases} \alpha_i(z_j) - z_j^i, & 0 \le i < r \\ z_j^R - z_j^i, & r \le i < R \end{cases} \tag{3.5.16}$$

If some roots are multiple, then we will differentiate $N(z)$ accordingly and then substitute multiple roots. The process is fairly routine. Now directly from (7.11) under $\lim_{z \to 1} P(z) = 1$ we obtain

$$\sum_{i=0}^{R-1} p_i a_{i0} = 1, \text{where } a_{i0} = \frac{1}{R-L} \cdot \begin{cases} \overline{\alpha}_i - i, & 0 \le i < r \\ R - i, & r \le i < R \end{cases}, \tag{3.5.17}$$

where $\overline{\alpha}_i = E\Sigma_i$. Writing down (7.15-7.18) in the matrix form we have

$$(p_0, \ldots, p_{R-1})A = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathrm{e} \in \mathbb{R}^R, \tag{3.5.18}$$

where

$$A = \begin{pmatrix} a_{00} & a_{01} & \dots & a_{0,\,R-1} \\ a_{10} & a_{11} & \dots & a_{1,R-1} \\ \vdots & \vdots & \vdots & \vdots \\ a_{R-1,0} & a_{R-1,1} & \dots & a_{R-1,R-1} \end{pmatrix} \tag{3.5.19}$$

is a nonsingular matrix of rank R and the vector $(p_0, \dots, p_{R-1})$ can be found from equation (3.5.18-3.5.19) numerically. We conclude this discussion by mentioning that there are many numerical methods and software modules that deal with calculating the roots of the equation $z^R - \beta(\lambda - \lambda a(z))$ when the functional $\beta(\lambda - \lambda a(z))$ becomes specified.

The above results can be summarized in the following theorem.

Theorem 3.5.1. *The queueing process $Q(t)$ in the system with three modes, parallel service discipline, under the r-R-N hysteretic control $(R \leq N)$ is semi-regenerative with respect to the sequence $\{T_n\}$ of successive service cycles over which $\{Q_n = Q(T_n)\}$ is a homogeneous Markov chain with transition probability matrix P of* (3.5.9). *It is irreducible and aperiodic. The invariant probability measure $\mathbf{p} = (p_0, p_1, \dots)$ exists and it is unique if and only if $L = ab\lambda < r$ and under this condition, it satisfies formulas* (3.5.11) *and* (3.5.16-3.5.19). □

Model 2. $R > N$. The semi-regenerative nature of the queueing process is very similar to that of Model 1 and thus we omit any further discussion on this. The difference between the two models lies in the epoch following mode 3 with formula (3.5.3) modified as follows.

$$Q_1 = \begin{cases} (\Sigma_{Q_0} - R)^+ + W_1, & 0 \le Q_0 < r \\ (Q_0 - R)^+ + W_1, & Q_0 \ge r. \end{cases} \tag{3.5.20}$$

For $Q_0 = i < r,$ we have

$$P_i(z) = E\big[z^{Q_1} | Q_0 = i\big] = \delta_i(z) E z^{W_1} = \delta_i(z)\beta(\lambda - \lambda a(z)), i < r, \tag{3.5.21}$$

where from (3.5.20),

$$\delta_i(z) = E z^{(\Sigma_i - R)^+}. \tag{3.5.22}$$

Using property (xi), formula (2.1.11) Dshalalow [27],

$$\delta_i(z) = \mathcal{D}_x^R\big\{\alpha_i(x) + z^{-R}[\alpha_i(z) - \alpha_i(xz)]\big\}, \tag{3.5.23}$$

where $\alpha_i(z)$ satisfies formulas (3.5.5-3.5.7), explicitly as

$$\begin{aligned} \delta_i(z) = \mathcal{D}_x^R \Bigg\{ & \Gamma_0^{(i)}(x) - [1 - a(x)]\mathcal{D}_w^{N-1}\left(\frac{\Gamma_0^{(i)}(xw)}{1 - a(xw)}\right) \\ & + z^{-R}\Bigg[\Gamma_0^{(i)}(z) - [1 - a(z)]\mathcal{D}_w^{N-1}\left(\frac{\Gamma_0^{(i)}(wz)}{1 - a(wz)}\right) \\ & - \Gamma_0^{(i)}(xz) - [1 - a(xz)]\mathcal{D}_w^{N-1}\left(\frac{\Gamma_0^{(i)}(xwz)}{1 - a(xwz)}\right)\Bigg]\Bigg\}, \; i < r. \end{aligned} \tag{3.5.24}$$

Now form (3.5.20),

$$P_i(z) = E\big[z^{Q_1} | Q_0 = i\big] = \beta(\lambda - \lambda a(z)) \cdot \begin{cases} z^{(i-R)^+}, & i \ge r \\ \delta_i(z), & i < r \end{cases} \tag{3.5.25}$$

From (3.5.20), with the notation $p_{ij} = P\{Q_1 = j | Q_0 = i\}$, we have

$$p_{ij} = \begin{cases} P\{W_1 = j - (i-R)^+\} = \begin{cases} q_{j-(i-R)^+}, & j \geq i - r \\ 0, & j < i - r \end{cases}, & i \geq r \\ q_j \sum_{k=N}^{R} \alpha_k^{(i)} + \sum_{k=R+1}^{R+j} \alpha_k^{(i)} q_{j+R-k}, & i < r, j \geq 0 \\ \text{where } \alpha_k^{(i)} = P\{\Sigma_i = k\} > 0, & k \geq N, \end{cases} \tag{3.5.26}$$

where $q_k = P\{W_1 = k\} > 0, k = 0, 1, \ldots,$ are the probability distribution of the number of arriving primary units in the system during a service time of a primary unit that can be obtained from the expansion of $\beta(\lambda - \lambda a(z))$, whereas $\alpha_k^{(i)} > 0, k = N, N+1, \ldots, (\alpha_k^{(i)} = 0, k < N)$ is the probability distribution of arriving primary units during mode 3 and it can be derived from the expansion of the functional $\alpha_i(z)$ of formula (3.4.5). Thus all $p_{ij} > 0,$ for $i < r$ and $j = 0, 1, \ldots$

Hence the transition probability matrix P reads

$$P = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0,\,j} & p_{0,\,j+1} & \cdots\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{r-1,0} & p_{r-1,\,1} & \cdots & p_{r-1,\,j} & p_{r-1,\,j+1} & \cdots\cdots \\ q_0 & q_1 & \cdots & q_j & q_{j+1} & \cdots\cdots \\ 0 & q_0 & \cdots & q_{j-1} & q_j & \cdots\cdots \\ 0 & 0 & \cdots & q_{j-2} & q_{j-1} & \cdots\cdots \\ 0 & 0 & \cdots & q_{j-3} & q_{j-2} & \cdots\cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{3.5.27}$$

The left upper corner block matrix (rows from 0 to $r-1$) is strictly positive. All elements $q_k > 0$. Thus, $\{Q_n\}$ is irreducible. Analogous to Model 1, matrix P is also a Δ_{r+1}-*matrix*. From (3.5.27), it follows that the Markov chain $\{Q_n\}$ is aperiodic. By Abolnikov-Dukhovny [3], it is recurrent positive if and only if

$P_r'(1) < r$ (where $P_i(z)$ is the generating function of the ith row of P). The latter reduces to the same condition as in Model 1,

$$L = ab\lambda < r \tag{3.5.28}$$

as in Model 1, where L the *offered load*.

The pgf $P(z)$ of the invariant probability measure $\mathbf{p} = (p_0, p_1, \dots)$ of $\{Q_n\}$ (under condition (3.5.28)) satisfies the formula

$$P(z) = \beta(\lambda - \lambda a(z)) \frac{\sum_{i=0}^{r-1} p_i[\delta_i(z) z^R - z^i] + \sum_{i=r}^{R-1} p_i(z^R - z^i)}{z^R - \beta(\lambda - \lambda a(z))}. \tag{3.5.29}$$

Finding unknown probabilities $p_i, i = 0, \dots, R-1$ is a matter of the same routine as in Model 1. Let $z_1, \dots, z_{R-1}$ be the roots of the denominator $D(z)$, all except, $z_0 = 1$ in $\overline{B}(0,1)$. Then assuming all roots distinct (which is the most likely scenario) and proceeding with the substitution of the root z_j into $N(z) = \sum_{i=0}^{r-1} p_i[\delta_i(z) z^R - z^i] + \sum_{i=r}^{R-1} p_i(z^R - z^i)$ we have

$$N(z_j) = \sum_{i=0}^{R-1} p_i a_{ij} = 0, j = 1, \dots, R-1, \tag{3.5.30}$$

where

$$a_{ij} = \begin{cases} \delta_i(z_j) z_j^R - z_j^i, & 0 \le i < r \\ z_j^R - z_j^i, & r \le i < R \end{cases} \tag{3.5.31}$$

If some roots are multiple, then we will differentiate $N(z)$ accordingly and then substitute multiple roots. Now directly from (3.5.29), under $\lim_{z\to 1}P(z)=1$, we obtain

$$\sum_{i=0}^{R-1} p_i a_{i0} = 1, \text{where } a_{i0} = \frac{1}{R-L}\cdot \begin{cases} \overline{\delta}_i + R - i, & 0 \le i < r \\ R-i, & r \le i < R \end{cases}, \tag{3.5.32}$$

where $\overline{\delta}_i = E(\Sigma_i - R)^+$. Writing down (3.5.30-3.5.32) in the matrix form we have

$$(p_0, \ldots, p_{R-1})A = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathrm{e} \in \mathbb{R}^R, \tag{3.5.33}$$

where

$$A = \begin{pmatrix} a_{00} & a_{01} & \ldots & a_{0,\,R-1} \\ a_{10} & a_{11} & \ldots & a_{1,R-1} \\ \vdots & \vdots & \vdots & \vdots \\ a_{R-1,0} & a_{R-1,1} & \ldots & a_{R-1,R-1} \end{pmatrix} \tag{3.5.34}$$

is a nonsingular matrix of rank R and the vector $(p_0, \ldots, p_{R-1})$ can be found from equation (3.5.33-3.5.34) numerically.

The variant of Theorem 3.5.1 in the context of Model 2 will read

Theorem 3.5.2. *The queueing process $Q(t)$ in the system with three modes, parallel service discipline, under the r-N-R hysteretic control $(N < R)$ is semi-regenerative with respect to the sequence $\{T_n\}$ of successive service cycles over which $\{Q_n = Q(T_n)\}$ is a homogeneous Markov chain with transition probability*

matrix P of (3.5.27). *It is irreducible and aperiodic. The invariant probability measure* $\mathbf{p} = (p_0, p_1, \dots)$ *exists and it is unique if and only if* $L = ab\lambda < r$ *and under this condition, it satisfies formulas* (3.5.29) *and* (3.5.31-3.5.34). □

3.6. MEAN STATIONARY SERVICE CYCLE

Because service cycles in this system are important time epochs of decision makings over which the queueing process conditionally regenerates, we would like discuss the computational aspect of service cycles $\{T_n\}$ in some form and lay foundation for forthcoming analysis of the continuous time parameter process $Q(t)$. First, we notice that the interval $(T_n, T_{n+1}]$ contains the nth service time and possibly a time when the server processes two queues and an idle period dependent on what state the queue was at T_n. The mean value of the length of this interval in equilibrium is called the *mean stationary service cycle* and it is defined as

$$E\mathcal{C} = \overline{\mathcal{C}} = \ < \mathrm{p}, \mathrm{c} > \ = \sum_{i=0}^{\infty} p_i c_i, \tag{3.6.1}$$

where

$$\mathrm{c} = \{c_i = E[T_1 | Q_0 = i], i = 0, 1, \dots\} \tag{3.6.2}$$

and **p** is the invariant probability measure of P. Below is the following rational of (3.6.1-3.6.2):

$$\begin{aligned} E[T_{n+1} - T_n] &= \sum_{i=0}^{\infty} E[T_{n+1} - T_n | Q_n = i] P\{Q_n = i\} \\ &= \sum_{i=0}^{\infty} E[T_1 | Q_0 = i] P\{Q_n = i\} = \sum_{i=0}^{\infty} c_i P\{Q_n = i\}. \end{aligned}$$

Obviously,

$$c_i = \begin{cases} \overline{T}_i + b, & 0 \le i < r \\ b, & r \le i \end{cases} \tag{3.6.3}$$

where $\overline{T}_i = ET_i$. Therefore,

$$\begin{aligned} E[T_{n+1} - T_n] &= \textstyle\sum_{i=0}^{r-1} P\{Q_n = i\}\overline{T}_i + b\sum_{i=0}^{\infty} P\{Q_n = i\} \\ &= \textstyle\sum_{i=0}^{r-1} P\{Q_n = i\}\overline{T}_i + b. \end{aligned}$$

Passing to the limit in the last equation under the condition that $L = ab\lambda < r$ we have

$$\overline{C} = \lim_{n\to\infty} E[T_{n+1} - T_n] = \textstyle\sum_{i=0}^{r-1} p_i\overline{T}_i + b = \ < \mathrm{p}, \mathrm{c} > . \tag{3.6.4}$$

Now we come to compute $\overline{C}$ explicitly. Without loss of generality we restrict our discussion to Model 1. Model 2 can be rendered very similarly.

From

$$\alpha_i(1,\theta) = Ee^{-\theta\mathcal{T}_i} = \Gamma_0^{(i)}(1,\theta) - \big[1 - \tfrac{\lambda}{\lambda+\theta}\big]\mathcal{D}_x^{N-1}\left[\tfrac{\Gamma_0^{(i)}(x,\theta)}{1-a(x)}\right]$$

we obtain

$$\begin{aligned} \overline{T}_i &= (-1)\alpha_i'(1,0) \\ &= (-1)\tfrac{d}{d\theta}\Gamma_0^{(i)}(1,\theta)|_{\theta=0} - (-1)\big[0 - \lambda(-1)\tfrac{1}{\lambda^2}\big]\mathcal{D}_x^{N-1}\left[\tfrac{\Gamma_0^{(i)}(x,0)}{1-a(x)}\right] \\ &= bg_i + \tfrac{1}{\lambda}\mathcal{D}_x^{N-1}\left[\tfrac{\Gamma_0^{(i)}(x,0)}{1-a(x)}\right],\ i = 0,\dots,r-1. \end{aligned}$$

Now

$$\alpha_i(z) = \alpha_i(z,0) = Ez^{\Sigma_i} = \Gamma_0^{(i)}(z,0) - [1-a(z)]\mathcal{D}_x^{N-1}\left[\frac{\Gamma_0^{(i)}(x,0)}{1-a(x)}\right]$$

implying that

$$\overline{\alpha}_i = \tfrac{d}{dz}\Gamma_0^{(i)}(z,0)|_{z=1} - [0-a]\mathcal{D}_x^{N-1}\tfrac{\Gamma_0^{(i)}(x,0)}{1-a(x)}$$

$$= i - (1-\delta_{0,i})g_i + \lambda a\overline{T}_i,$$

with

$$\lambda a\overline{T}_i = \lambda a b_i g_i + a\mathcal{D}_x^{N-1}\tfrac{\Gamma_0^{(i)}(x,0)}{1-a(x)},$$

and

$$g_i := \begin{cases} \mathcal{D}_y^{S-1}\frac{1}{1-b(y)}, & i=0 \\ \mathcal{D}_y^{S-1}\left[\frac{1-\gamma^i(\mu-\mu b(y))}{1-\gamma(\mu-\mu b(y))}\right], & i=1,\ldots,r-1 \\ i-R, & i=r,\ldots,R-1. \end{cases}$$

So we have

$$\overline{\alpha}_i - i = -(1-\delta_{0,i})g_i + \lambda a\overline{T}_i,\ i=0,\ldots,r-1.$$

From (7.11) and (7.17),

$$R - L = \textstyle\sum_{i=0}^{r-1}(\overline{\alpha}_i - i)p_i + \sum_{i=r}^{R-1}(R-i)p_i$$

implying that

$$R - L = \lambda a\textstyle\sum_{i=0}^{r-1}\overline{T}_i p_i - \sum_{i=1}^{r-1}p_i g_i + \sum_{i=r}^{R-1}(R-i)p_i$$

$$= \lambda a(\overline{C} - b) - \textstyle\sum_{i=1}^{r-1}p_i g_i + \sum_{i=r}^{R-1}(R-i)p_i$$

and thus

$$\lambda a\overline{C} = R + \textstyle\sum_{i=1}^{R-1}p_i g_i.$$ □

REFERENCES

[1] Abaev, P. and Razumchik, Queuing model for SIP server hysteretic overload control with bursty traffic, in *Internet of Things, Smart Spaces, and Next Generation Networking*, 13th International Conference, NEW2AN 2013, and 6th Conference, ruSMART 2013, St. Petersburg, Russia, August 2013, Proceedings, Eds. Balandin, S., Andreev, S., and Koucheryavy, Ye., 383-396, Springer-Verlag, Berlin, 2013.

[2] Abaev, P., Gaidamaka, Yu., and Samouylov, K.E., Hysteretic control technique for overload problem solution in network of SIP servers, *Computing and Informatics*, **33** (2014), 218-236.

[3] Abolnikov, L. and Dukhovny, A., Markov chains with transition delta-matrix: ergodicity conditions, invariant probability measures and applications, *J. Appl. Math. Stoch. Anal.*, **4**:4 (1991), 335-355.

[4] Abolnikov, L., Agarwal, R. V. and Dshalalow, J. H., Random walk analysis of parallel queueing stations. *Mathematical and Computer Modelling*, **47** (2008), 452-468.

[5] Abolnikov, L., Dshalalow, J.H., and Treerattrakoon, A., On dual hybrid queueing systems, *Nonlinear Analysis: Hybrid Systems*, **2**:1 (2008), 96-109.

[6] Ait-Salaht, F. and Castel-Taleb, H., The threshold based queueing system with hysteresis for performance analysis of clouds, 2015 International Conference on Computer, Information and Telecommunication Systems (CITS) in *IEEE Transactions on Computers* (2015), 1-5.

[7] Alghamdi, A. and Dshalalow, J. H. Multiphase fluctuation analysis in a queue with an enhanced maintenance. Continuous time parameter process, *Nonlinear Studies*, **17**:3 (2010), 199-215.

[8] Alzahrani, M.S. and Dshalalow, J. H., Fluctuation analysis in a queue with (L-N)-policy and secondary maintenance. Discrete time parameter process. *Engineering Simulation,* **33**:4 (2011), 15-34.

[9] Andersen, E.S., On the fluctuation of sums of random variables I, *Math. Scand.*, **1** (1953), 263-285.

[10] Andersen, E.S., On the fluctuation of sums of random variables II, *Math. Scand.*, **2** (1954), 195-223.

[11] Avrachenkov, K., Perel, E., and Yechiali, U., Finite-buffer polling systems with threshold-based switching policy, *TOP*, **24** (2016) 541-571.

[12] Bekker, R., Queues with Lévy input and hysteretic control, *Queueing Syst.*, **63** (2009), 281-299.

[13] Bingham, N.H., Random walk and fluctuation theory, in *Handbook of Statistics* (Eds. D.N. Shanbhag and C.R. Rao), Volume **19,** 2001, Elsevier Science, 171-213.

[14] Boxma, O., Löpker, A., and Perry, D., On a make-to-stock production/mountain model with hysteretic control, *Annals of Operations Research*, **241**:1-2 (2016), 53–82.

[15] Cao, J. and Xie, W., Stability of a two-queue cyclic polling system with BMAPs under gated service and state-dependent time-limited service disciplines, *Queueing Syst.*, **85** (2017), 117-147.

[16] Chan, C.W., Armony, M., and Bambos, N., Maximum weight matching with hysteresis in overloaded queues with setups, *Queueing Syst.,* **82** (2016) 315–351.

[17] Choi, S.H. and Sohrabi, K., Analysis of a mobile cellular systems with hand-off priority and hysteresis control, in: INFOCOM 2000 *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings*, IEEE, 2000.

[18] Dikong, E.E. and Dshalalow, J.H., Bulk input queues with hysteretic control, *Queueing Syst.*, **32** (1999), 287-304.

[19] Dshalalow, J.H., On termination time processes, in: *Studies in Applied Probability; Essays in honour of Lajos Takács* (ed.'s J. Galambos and J. Gani), Applied Probability Trust, Sheffield, U.K., 325-336, 1994.

[20] Dshalalow, J.H., Excess level processes in queueing, in: *Advances in Queueing*, ed. by Dshalalow, J.H., CRC Press, Boca Raton, FL, 243-262, 1995.

[21] Dshalalow, J.H., On the level crossing of multi-dimensional delayed renewal processes, *J. Appl. Math. Stoch. Anal.*, **10**:4 (1997), 355-361.

[22] Dshalalow, J.H., Queueing systems with state dependent parameters; in *Frontiers in Queueing*, ed. by J.H. Dshalalow, CRC Press, Boca Raton, FL, 1997, 61-116.

[23] Dshalalow, J.H., Queues with hysteretic control by vacation and post-vacation periods, *Queueing Syst.*, **29** (1998), 231-268.

[24] Dshalalow, J.H., First excess level of vector processes, *J. Appl. Math. Stoch. Anal.*, **7**:3, 457-464, 1994.

[25] Dshalalow, J.H., First excess level analysis of random processes in a class of stochastic servicing systems with global control, *Stoch. Anal. Appl.*, **12**:1, 75-101, 1994.

[26] Dshalalow, J.H., Excess level processes in queueing, in: *Advances in Queueing*, CRC Press, Boca Raton, FL, 243-262, 1995.

[27] Dshalalow, J.H., *Lecture Notes on Stochastic Processes*, Florida Institute of Technology, Melbourne, FL, 2012.

[28] Dshalalow, J.H. and Dikong, E.E., On generalized hysteretic control queues with modulated input and state dependent service, *Stochastic Analysis and Applications*, **17**:6, 937-961, 1999.

[29] Dshalalow, J.H. and Dikong, E.E., Bulk input queues with hysteretic control, *Queueing Systems*, **32**, 287-304, 1999.

[30] Dshalalow, J.H., Kim, S. and Tadj, L., Hybrid queueing systems with hysteretic bilevel control policies, *Nonlinear Analysis*, **65**:11 (2006), 2153-2168.

[31] Dshalalow, J.H. and Merie, A., Fluctuation analysis in queues with several operational modes and priority customers, *TOP* **26**:2 (2018), 309-333.

[32] Dshalalow, J.H., Merie, A. and White, R. Fluctuation analysis in parallel queues with hysteretic control, *submitted.*

[33] Dudin, A. and Chakravarthy, S., Optimal hysteretic control for the BMAP/G/1 system with single and group service modes, *Annals of Operations Research*, **112** (2002), 153–169.

[34] Dudin, A. and Nishimura, S., Optimal hysteretic control for a BMAP/SM/1/N queue with two operation modes, *Math. Problems in Engineering*, **5** (2000), 397-419.

[35] Dukhovny, A., Multiple roots of some equations in queueing theory, *Stochastic Models,* **10**:2 (1994), 519-524.

[36] Dukhovny, A., Multiple roots of some equations in queueing theory, *Stochastic Models,* **10**:2 (1994), 519-524.

[37] Gaidamaka, Y., Pechinkin, A., Razumchik, R., Samouylov, K. and Sopin, K., Analysis of an M/G/1/R queue with batch arrivals and two hysteretic overload control policies, *Int. J. Appl. Math. Comput. Sci.*, **24**:3 (2014), 519–534.

[38] Golubchik, L. and Lui, J.C.S., Bounding of Performance Measures for a Threshold-based Queueing System with Hysteresis, *IEEE Transactions on Computers*, **51**: 4 (2002), 353-372.

[39] Gupta, U. C., and Sikdar, K., The finite-buffer M/G/1 queue with general bulk-service rule and single vacation, *Performance Evaluation*, **57**:2 (2004), 199-219.

[40] Heyman, D., The T-policy for the $M/G/1$ queue, *Man. Sci.*, **23** (1977), 775-778.

[41] Jain, M., Sharma, R., and Sharma, G.C, Multiple vacation policy for $M^X/H_k/1$ queue with un-reliable server, *Journal of Industrial Engineering International*, **9**:36 (2013), 1-11.

[42] Jian, M. and Jian A., Working vacations queueing model with multiple types of server breakdowns, *Applied Mathematical Modelling*, **34** (2010), 1-13.

[43] Ke, J-C., An M/G/1 queue under hysteretic vacation policy with an early startup and un-reliable server, *Math. Meth. Oper. Res.*, **63** (2006), 357–369.

[44] Kim, C., Dudin, A.N., Dudin, S., and Dudina, O., Hysteresis control by the number of active servers in queueing system with priority service, *Performance Evaluation*, **101** (2016), 20-33.

[45] Loris-Teghem, J., Hysteretic contorl of an M/G/1 queueing system with two service time distributions and removable server, in *Point Processes and Queueing Problems*, Colloquia Mathematics Societatis Janos Bolyai, Hungary, **24** (1978), 291-305.

[46] Pechinkin, A.V. and Razumchik, R.V., Stationary characteristics of $M_2/G/1/r$ system with hysteretic policy for arrival rate control, *Journal of Communications Technology and Electronics*, **58**:12 (2013), 1282–1291.

[47] Perel, E. and Yechiali, U., Two-queue systems with switching policy based on the queue which is not being served, *Stochastic Models*, to appear (published online May, 2017).

[48] Redner, S., *A Guide to First-Passage Processes*, Cambridge University Press, Cambridge, 2001.

[49] Semenova, O.V., Optimal hysteresis control for BMAP/SM/1 queue with MAP-input of disasters, *Quality Technology & Quantitative Management*, **4**:3 (2007), 395-405.

[50] Sikdar, K., and Gupta, U. C., Analytic and numerical aspects of batch service queues with single vacation, *Computers & Operations Research*, **32**:4 (2005), 943-966.

[51] Sikdar, K., and Gupta, U. C., On the batch arrival batch service queue with finite buffer under server's vacation: $M^X/G^Y/1/N$ queue, *Computers & Mathematics with Applications*, **56**:11 (2008), 2861-2873.

[52] Solanki, A., Transient Behaviour of batch arrival queue with N-policy and single vacation M^X/G/1/N-policy, *Modeling of Engineering and Technological Problems*, **1146** (2009), 479-487.

[53] Tadj, L. and Ke, J-C., Control policy of a hysteretic bulk queueing system, *Math. Computer Modeling*, **41**:4-5 (2005), 571-579.

[54] Takagi, H., Analysis and application of polling models, in *Performance Evaluation*: *Origins and Directions*, by Harling G., Lindemann, C., and Reiser, Springer-Verlag, Berlin, 2000.

[55] Tadj, L. and Choudhury, G., The M^X/G/1 queue with unreliable server, delayed repairs, and Bernoulli vacation schedule under T-Policy, Applications and Applied Mathematics, **8**:2 (2013), 346-365.

[56] Takács, L., On fluctuations problems in the theory of queues, *Adv. Appl. Probab.*, **8**:3 (1976), 548-583.

[57] Takács, L., On fluctuations of sums of random variables, in Studies in Probability and Ergodic Theory, *Advances in Mathematics; Supplementary Studies*, **2**, ed. by G.-C. Rota, (1978), 45-93.

[58] Takagi, H., Analysis and application of polling models, in *Performance Evaluation*: *Origins and Directions*, by Harling G., Lindemann, C., and Reiser, Springer-Verlag, Berlin, 2000.

[59] Teghem, J., Control of the service process in a queueing system, *European J. of Oper. Res.*, **23**:2 (1986), 141-158.

[60] Tian, N. and Zhang, Z.G., *Vacation Queueing Models*, Springer 2006.

[61] Van der Gaast, J.P, Adan, I.J.B.F., de Koster, R.B.M., The analysis of batch sojourn-times in polling systems, *Queueing Syst.*, **85** (2017), 313-335.

[62] Vishnevskii., V.M. and Dudin, A.N., Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks, *Automation and Remote Control*, **78**:8 (2017), 1361–1403.

[63] Wu, W., Tang, Y., and Yu, M., Analysis of an M/G/1 queue with multiple vacations, N-Policy, unreliable service station and repair facility failures, *Int. Journ, Supply and Oper. Management,* **1**:1 (2014), 1-19.

[64] Zhennovyi, Yu. V. and Zhennovyi, K. Yu., Stationary characteristics of an $M_2^X/M/n$ queue with hysteretic control of the input flow intensity, *Journal of Communications Technology and Electronics*, **59**:6 (2014), 614–621.