Call centres with balking and abandonment: from queueing to queueing network models

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
in the
Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan

By

Zhidong Zhang

June 2010

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

The research on call centres has attracted many researchers from different disciplines recently. In this thesis, we focus on call centre modelling, analysis and design. In terms of modelling, traditionally call centres have been modelled as single-node queueing systems. Based on the Semiopen Queueing Network (SOQN) model proposed by Srinivasan et al. [42], we propose and study SOQN models with balking and abandonment (both exponential and general patience time distributions). In addition, we study the corresponding single-node queueing systems and obtain new results. For each model, we study the queue length distribution, waiting time distribution and the related performance measures. To facilitate the computation, we express the performance measures in terms of special functions. In terms of call centre design, we develop a design algorithm to determine the minimal number of CSRs (S) and trunk lines (N) to satisfy a given set of service level constraints.

The explicit expressions for performance measures obtained allow for theoretical analysis of the performance measures. For example we prove monotonicity and convexity properties of performance measures for the $M/M/S/N$ and $M/M/S/N+M$ models. We also study the comparison of different patience time distributions for the $M/M/S/N+G$ model.

We provide numerical examples for each model and discuss numerical results such as monotonicity properties of performance measures. In particular, we illustrate the efficacy of our design algorithm for various models including patient, balking and abandonment models. The impact of model parameters on the design of call centres is also discussed based on the numerical examples. The results are computed using Matlab, where special functions are available.

Acknowledgements

This thesis grew out of a research project provided by my supervisor Prof. Raj Srinivasan. I am sincerely grateful to Prof. Raj Srinivasan for his invaluable advice and patient guidance. This thesis could not have been finished without his constant help and support. I would like to thank the members of my committee, Prof. Mikelis G. Bickis, Prof. Chris Soteros, Prof. William H. Laverty, Prof. Winfried K. Grassmann and my external examiner Prof. Noah F. Gans for reading my thesis and valuable suggestions. Last but not least, I want to thank my family and friends, for their support and encouragement.

To

My father

Dehua Zhang

My wife

Jiezhi Qi

My daughter

Erin Jiaqi Zhang

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

α, Abandonment rate, page 84

λ, Poisson arrival rate, page 5

μ, Exponential service rate, page 5

θ, Exponential service rate for Node 1, page 7

$I_A(k)$, Indicator function, page 48

K, Number of waiting spaces (called buffer), page 4

$M(n)/M/S/N$, $M/M/S/N$ with state-dependent arrival rates, page 6

$M/M/S$, Poisson arrival, exponential service, S servers, infinite trunk lines, page 6

$M/M/S+M$, $M/M/S$ with exponential abandonment, page 6

$M/M/S/N$, Poisson arrival, exponential service, S servers, N trunk lines, page 5

$M/M/S/N+G$, $M/M/S/N$ with general abandonment, page 6

$M/M/S/N+M$, $M/M/S/N$ with exponential abandonment, page 6

$M/M/S/S$, Poisson arrival, exponential service, S servers, S trunk lines, page 6

N, Number of trunk lines, page 4

S, Number of CSRs, page 4

ACD, Automatic Call Distributor, page 2

ANI, Automatic Number Identification, page 2

ASA, Average Speed of Answer, page 10

AWT, Acceptable Waiting Time, page 10

CQN, Closed Queueing Network, page 44

CSRs, Customer Service Representatives, page 2

CTI, Computer-Telephony Integration, page 3

CTMC, Continuous Time Markov Chain, page 15

DNIS, Dialed Number Identification Service, page 2

FCFS, First Come First Served, page 3

IVRU or VRU, Interactive Voice Response Unit, page 2

PABX or PBX, Private Automatic Branch Exchange, page 2

PASTA, Poisson Arrivals See Time Averages, page 15

PSTN, Public Service Telephone Network, page 2

SBR, Skills-Based Routing, page 3

SL, Service Level, page 9

SOQN+G, Semiopen Queueing Network Model with general abandonment, page 13

SOQN+M, Semiopen Queueing Network Model with exponential abandonment, page 13

SOQN, Semiopen Queueing Network Model, page 12

TSF, Telephone Service Factor, page 10

WFM, Workforce Management, page 9

Chapter 1

Introduction

The research on call centres has attracted many researchers from different disciplines. Mandelbaum [32] provides a comprehensive research bibliography with abstracts including disciplines such as Operations Research, Statistics, Psychology, Information Technology, Industrial Engineering etc. The research on call centres is reviewed and extended in the tutorial and survey paper by Gans et al. [20]. We will focus on call centre modelling, analysis and design in this thesis. In this chapter, we first give a definition of call centres and briefly explain the operational process of an inbound call centre. This serves as a background introduction to our study in this thesis. Then we describe two main types of research problems in call centres that we will address. We conclude this chapter with an outline and the main contributions of the thesis.

1.1 Introduction to call centres

1.1.1 What is a call centre?

A *call centre* is a department within a company or a third-party organization that answers incoming telephone calls from customers (often for the purposes of product support), or that makes outgoing telephone calls to customers (for example, tele-marketing). If such a department also responds to letters, faxes, e-mails, and similar written correspondence, it is called a *contact centre*.

A call centre may only handle incoming telephone calls initiated by customers (inbound calls) or only make outgoing telephone calls to customers (outbound calls). There are also call centres who deal with both types of calls. In most contact centres inbound calls still form the majority of contacts with customers. Also inbound calls are more time demanding than other types of contacting manners (like letters or e-mails) when it comes to waiting times or response times. Therefore in this thesis we only focus on *inbound call centres*,

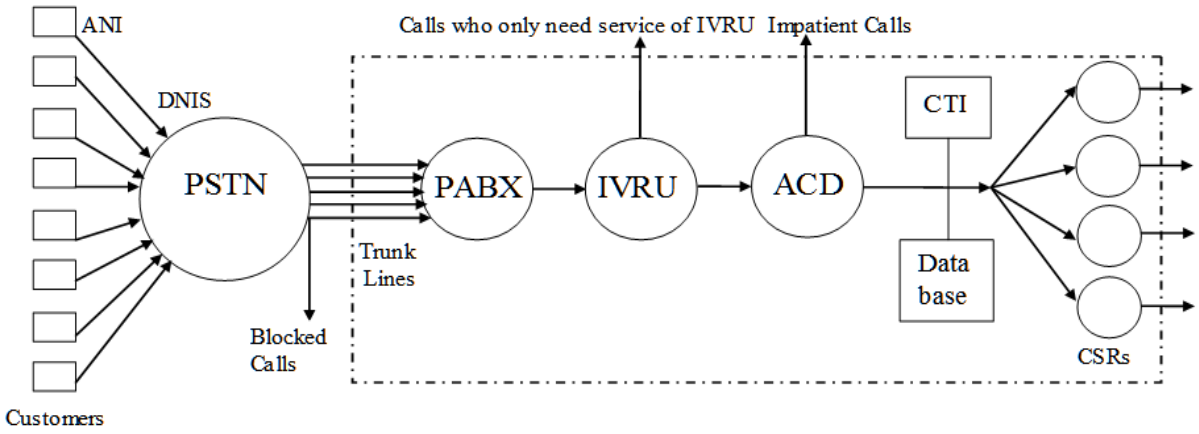

Figure 1.1: Operational process of an inbound call centre

which only involve dealing with incoming telephone calls from customers.

An inbound call centre contains a collection of CSRs (Customer Service Representatives) who provide service through talking to customers over telephones. CSRs are supported by quite elaborate equipment, such as a Private Automatic Branch Exchange (PABX or PBX) , an Interactive Voice Response Unit (IVRU or VRU), an Automatic Call Distributor (ACD), and computers, etc. [42]. For details of components and the operational process of an inbound call centre, see Figure 1.1.

1.1.2 Operational process of an inbound call centre

Almost everyone has the experience of calling a call centre. In the following we will briefly describe components and the operational process of an inbound call centre based on the description in [20]. The basic process is described in Figure 1.1.

When customers want to receive service from a call centre, they dial a special number provided by the call centre. The Public Service Telephone Network (PSTN) company then uses the Automatic Number Identification (ANI) number (the phone number from which the customer dials) and the customer's Dialed Number Identification Service (DNIS) number (the special number being dialed) to connect the customer to the PABX privately-owned by the call centre.

There are telephone lines (often called trunk lines) connecting the PABX to PSTN. If a trunk line is free, the customer seizes it. Otherwise the customer will receive a busy signal and will be rejected. We will say this customer is *blocked*. Once the call is accepted, the

customer will be connected through the PABX to the IVRU. Usually the IVRU provides several options for customers to choose and may also provide some automatic service for customers. After interaction with the IVRU, those customers who complete their service at the IVRU leave the system and release the trunk lines. For some call centres, most customers only need service of the IVRU without requiring the service of CSRs. For instance it is reported in [20] that about 80% of calls in banking call centres are fully self-served using an IVRU.

If the customer requires the service of a CSR, the call will be handed from the IVRU to the ACD. The ACD, a highly sophisticated specialized switch, is designed to route calls to individual CSR based on the specific needs of calls. Skills-Based Routing (SBR) is one example, in which different types of calls are routed to the best available CSR with the appropriate skill according to preprogrammed rules. However, how to get the optimal routing rules or policy of SBR is a complex control problem and we will not address this problem in this thesis. If no appropriate CSRs are available, the customer is informed to wait and join a queue at the ACD. This customer is called *delayed*. The ACD will decide when the customer gets served according to a preprogrammed queueing discipline (usually FCFS, i.e., First Come First Served). While waiting at the ACD, delayed customers may be exposed to music and sometimes are informed of their expected delay. Delayed customers may decide that the service is not worth the wait and may hang up before they are served. In this case they are said to *abandon* or *renege* and they are called *impatient* customers.

Customers who do not abandon will eventually be connected to a CSR. While serving a customer, the CSR works via a PC supported by Computer-Telephony Integration (CTI), which is technology that allows interactions on a telephone and a computer to be integrated or coordinated. CTI will help ACD to route the call, help the CSR to get the caller's information from the database and hence facilitate the service process. After the customer receives the service and leaves, the CSR still needs some wrap-up time to finish the whole service process and then may be available for the next customer. The service time is the sum of talk time and wrap-up time.

Abandoned and blocked customers may try to call again after some time and these calls are referred to as *retrials*. Those who finish talking with a CSR may also need further help and call back again hence they become *return* customers or *feedback* customers. Note that these two types of customers are not shown in Figure 1.1.

1.2 Main research problems in call centres

1.2.1 Call centre modelling and performance analysis problems

In order to analyze a call centre, we need to first give a mathematical model of the given call centre. The more realistic is the model, the better, but it will be more difficult to analyze and solve the model. How to model a call centre is a complex issue and usually the first step of call centre research.

Customers call the call centre independently and in a random way, each with a random required service time. Thus, the call centre is driven by random arriving calls and random service times. Since there is uncertainty in future call arrival times and required service times of arriving calls, it is necessary to use stochastic models. Also from the operational process of an inbound call centre described in Section 1.1.2, a queueing model is a natural choice. Traditionally queueing theory was established from telephone circuits design problems pioneered by A.K. Erlang, a Danish scientist who worked at the Copenhagen Telephone Company [17].

Next, we will give an illustrated example to show how to model a call centre using queueing theory and we will explain some concepts in the context of queueing theory as well.

A natural queueing model

A natural queueing model for a simplified call centre is depicted in Figure 1.2, which is adapted from [30]. The number of waiting spaces (called buffer) at the ACD is K and the number of CSRs is S. Thus there are $N = K + S$ trunk lines altogether at the PABX for this call centre. If a call finds all N trunk lines occupied upon arrival, it will receive a busy signal and is *blocked* from entering the system. Otherwise it is either connected to the system and occupies one of the free trunk lines or just does not want to enter the system and hangs up (this is called to *balk*). If it enters the system and there is at least one free CSR, the call gets service immediately. Otherwise it is *delayed* and has to wait in a queue at the ACD for a CSR to become available. While waiting, calls may become impatient and hang up, or *abandon (renege)* the system before being served and thus release the trunk line. The queueing discipline is usually FCFS. After served by a CSR, the call leaves

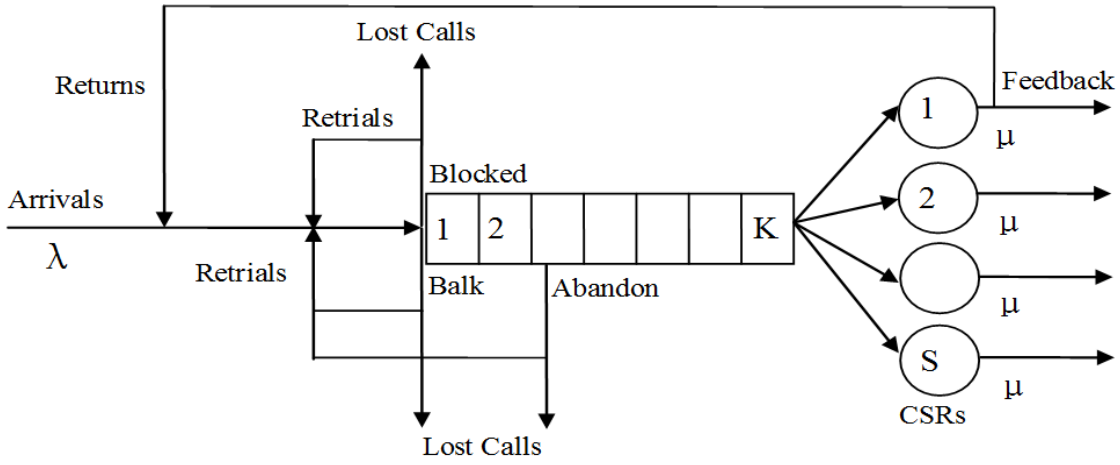

Figure 1.2: A natural queueing model

the system and it releases both the trunk line and the CSR and these resources become available to other arriving calls. Some served calls may choose to call back (*feedback*) and become *returns*. Some of those calls who do not get served (blocked, abandon or balk) may call again and they become *retrials*. The remaining calls become *lost calls*. The above model is also similarly described in [20].

If we assume that the calls arrive according to a Poisson process with rate λ and that the required service times of the calls are i.i.d. exponential with mean $1/\mu$, then this model is a simple $M/M/S/N$ queueing system with features such as balking, abandonment, retrial, and feedback, where we have used a notation similar to Kendall's notation [24]. In this notation, a queueing system is represented by $(.)/(.)/S/N$, where the first position describes the arrival process; the second position describes the service process. Some symbols we will use to represent arrival or service process include M for i.i.d. exponential service or Poisson arrival, D for i.i.d. deterministic and G for i.i.d. general. In the third and fourth position, S is the number of servers (CSRs in our case) and N is the maximum number of calls in the system i.e., in the queue or in service (number of trunk lines in our case). Note that Kendall's notation only includes the first three positions.

Performance analysis of the natural queueing model

Without features such as balking, abandonment, retrial, and feedback, the $M/M/S/N$ queueing system has a closed-form solution for the stationary (steady-state or long-run) queue length (number of calls in the system) distribution and waiting time distribution, from which we can obtain average queue length, average waiting time, probability of blocking etc. This analyzing process of the model is called *performance analysis* and the outputs are called *performance measures*, which reflect the performance of a call centre or a queueing system. We will review the performance analysis of $M/M/S/N$ queueing system and give the definition of the above performance measures in Section 2.4, Chapter 2.

The performance measures are useful information in the design and management of call centres. They can be used to determine the service levels (explained in Section 1.2.2) of call centres. They are also the input of the call centre design problem, which will be elaborated in Section 1.2.2.

However, not all queueing models can be analyzed exactly to obtain performance measures as $M/M/S/N$ model. For example, if we include additional features, the model may become impossible to solve and other techniques have to be used to analyze the model such as numerical methods, approximations, and simulation.

Other queueing models for call centres

The natural queueing model described above only allows homogeneous calls and CSRs. However in practice, call centres may have multi-type calls and multi-skilled CSRs which requires SBR and this is difficult to analyze. The models of call centres can be categorized as SBR and non-SBR types of models. Non-SBR models, as we have seen, can be thought of as base models which do not consider multi-type calls and CSRs. There is an extensive literature on SBR models; for example see Standford and Grassmann [43].

However, in the following we only consider non-SBR models and give a list of common queueing models for call centres.

1. Markovian queueing models ($M/M/S, M/M/S/S, M/M/S/N$)

2. Balking models (state-dependent arrival rates $M(n)/M/S/N$).

3. Abandonment models ($M/M/S+M, M/M/S/N+M$ and $M/M/S/N+G$).

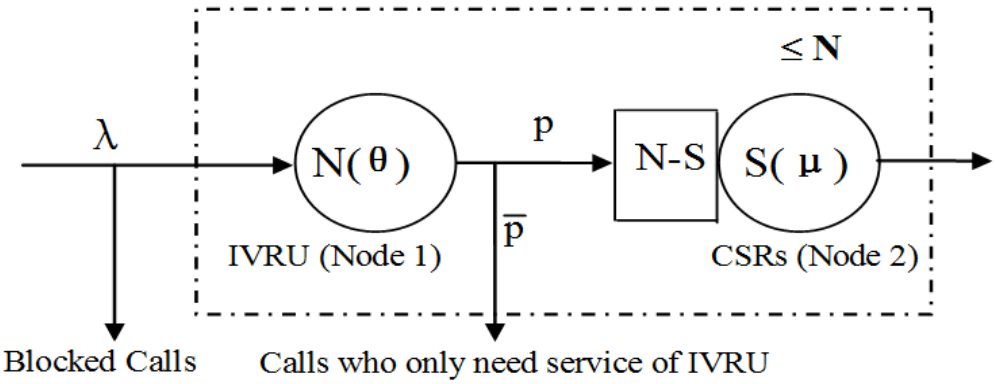

Figure 1.3: Semiopen queueing network model description and parameters

4. Retrial models.

5. Time-dependent queueing model (arriving and serving rate are time-dependent $M(t)/M(t)/S$).

6. Non-Markovian queueing models (general interarrival and service time $G/G/S$).

We will give a detailed review on the first class of queueing models in Chapter 2. The balking and abandonment models will be reviewed in the related chapters later in the thesis. We will not consider class 4, 5 and 6 in this thesis.

The base model in the thesis: Semiopen Queueing Network Model (SOQN)

Queueing models introduced above do not consider the role of IVRU. However, from Figure 1.1, we know that IVRU plays an important role in call centres; the calls get service from IVRU and a large proportion of calls only need self-service with IVRU and then leave the system without requiring the service of CSRs for some call centres. For details about the role of IVRU, see the thesis by Khudyakov [27].

In order to capture the role of the IVRU as well as the CSRs, Srinivasan et al. [42] proposed and analyzed a flow controlled network model. Their model is described in Figure 1.3.

The model is a semiopen queueing network with two nodes in series. The first node (Node 1, representing IVRU) has N servers each with exponential service rate θ. The second node (Node 2, representing CSRs) has S ($\leq N$) servers each with exponential service rate μ and are independent of the IVRU processing times. The maximum number

of calls in the network is N (representing the total number of trunk lines) i.e., if an arriving call finds N calls in the system, it will be blocked. Hence there is no queue at Node 1 and there are at most $N - S$ calls waiting at Node 2. In [42], it is assumed that the number of IVRU servers and the total number of trunk lines are both N, which reflects the fact that IVRU servers are often cheaper so that call centres can assign an IVRU server for each trunk line.

Arriving calls come to the system according to a Poisson process with rate λ. If the call is not blocked, it is immediately processed by the IVRU. Once the call is finished with the IVRU, it leaves the system with probability $\overline{p} = 1 - p$ and releases the trunk line, or it proceeds to Node 2 with probability p and holds the trunk line. If a CSR is free, the call is served, otherwise it waits for a CSR, but in either case, it always holds the trunk line. Once the call is processed by a CSR it releases both the CSR and the trunk line. Note that the trunk line is held by a call from the moment it enters the system until it leaves the system and this is not the case in the usual tandem queueing network. But the semiopen network model with maximum N calls in the system nicely characterizes this property.

Let $\{Q_1(t), Q_2(t)\}$ represent the number of calls at time t at Node 1 and 2 respectively. Note that $Q_1(t) + Q_2(t) \leq N$ for all $t \geq 0$. It is well-known that π_{ij}, the stationary distribution of $\{Q_1(t), Q_2(t)\}$, has a product form solution [14] and in Chapter 3 we will derive π_{ij} using two methods. From this, Srinivasan et al. obtained the stationary distribution π_{ij}, $P(blocking)$ and stationary waiting time distribution of a call conditioned that it joins Node 2 after finishing the IVRU process. We will give a detailed review of this model in Chapter 3.

In this thesis, this SOQN model will be our base model and we will extend this model by incorporating two more features: balking (state-dependent arrival rates) and abandonment (both exponential and general patience time distributions).

1.2.2 Call centre design problem

Queueing design problem has a long history and is one of the main applications of queueing theory. For example, see [23]. In the context of call centres, the queueing design problem becomes call centre design problem. Specifically, the term *call centre design problem* denotes the problem of how to determine the minimal number of CSRs (S) and trunk lines (N) to satisfy a given set of service level constraints (quality). Solving the call centre

design problem is the ultimate goal of modelling and performance analysis of call centres.

There are two key concepts in call centre design problem: service level and operating costs, which will be explained later. The service level reflects the performance of call centres and the operating costs are mainly incurred by the number of CSRs and trunk lines. Typically, there is a trade-off between service level (quality) and operating costs (efficiency). Higher service level (high quality) requires higher costs (low efficiency) and lower costs (high efficiency) cause lower service level (low quality). Quantitatively, this problem is formalized as an optimization problem, which will be elaborated later in this section. In practice, Workforce Management (WFM) software tools try to make this trade-off as optimal as possible [20].

Service level (quality)

Service level (SL) is a complex concept in call centre industry and involves many aspects of service process. Customer satisfaction, CSR effectiveness, etc. are qualitative service levels, which depend on product-related skills of CSRs. We are only concerned with quantitative service levels, which are usually related to performance measures on delay of a call centre. These service levels give constraints to the operation of call centres. The call centre industry uses several forms of service levels in their operation. In Table 1.1, we list some common service level constraints.

Typically call centres allow calls to wait for service if all CSRs are busy. Let random variable W_q denote the stationary waiting time of a call in the queue, assuming the system is stable. The first three service levels in Table 1.1 are all related to W_q. The other three service levels are the probability of blocking (used in finite buffer models), the probability of balking (used in balking models) and the probability of abandonment (used in abandonment models).

Note that service levels are special performance measures of call centre queueing models chosen by call centre manager to give constraints to the operation of call centres. If the call centre queueing model can be analyzed exactly, then the exact expressions of these service levels can be obtained. The expressions of these service levels will involve the model parameters and they are the output of the call centre performance analysis problem.

Service Level Constraints	Explanations
ASA=$E(W_q)<b$	ASA (Average Speed of Answer) or the average waiting time of a call before it gets served is less than b.
TSF=$P(W_q \leq \text{AWT})>b$	TSF (Telephone Service Factor) is the proportion of calls that has to wait shorter than a specified amount of time (called the Acceptable Waiting Time, or AWT). Typically the AWT is 20 seconds and b=80%, hence the famous 80-20 TSF means $P(W_q \leq 20\ sec)>80\%$.
$P(delay)=P(Wq>0)<b$	The proportion of calls that have to wait for service is less than b.
$P(blocking)<b$	The proportion of blocked calls is less than b. It is an important performance measure if we model a call centre as a finite buffer system (insufficient trunk lines).
$P(balking)<b$	The proportion of balking calls is less than b, used in balking models.
$P(Ab)=P(abandonment)<b$	The proportion of calls that abandon before being served is less than b.

Table 1.1: Common service level constraints

Costs (efficiency)

The operating costs of call centres are mainly incurred by the costs of CSRs. Personnel-related operating costs (CSRs salaries) account for about 60 to 70 percent of the total operating costs [20]. Therefore it is a common strategy for call centre managers to assign as small as possible number of CSRs to satisfy the given SL constraints, which makes high efficiency of CSRs.

Another cost is incurred by the cost of trunk lines which call centres may have to lease from the telephone company. This cost is usually smaller than that of CSRs.

There are other costs associated with service levels. For example, waiting or delay cost; a call centre which provides toll-free services has to pay for the time those calls waiting in the queue. Also order-taking businesses can sometimes estimate the opportunity cost of lost sales due to blocking or abandonment. We will not consider these kinds of costs in this thesis.

Formulations of the design problem

There are different formulations of this design problem (the fundamental trade-off between service level and operating costs) according to different economic structures and cost or

revenue analysis.

Traditionally call centres are seen as cost centres. In this case the optimization problem is formalized as

$$\begin{cases} \min \ f(S,N) \\ s.t.\ SL\ constraints, \end{cases}$$

where the objective function $f(S,N)$ is the costs of CSRs and trunk lines and the SL constraints are usually TSF and $P(blocking)$. For example, see [13, 35, 42]. Cleveland and Mayben [16] explained the traditional design method using Erlang B and Erlang C in isolation to solve this problem.

Another economic model considers call centres to be profit centres. There is a revenue for completing a customer service, and the objective is to maximize its profit, defined as revenue minus costs. G.M. Koole and A. Pot [31] studied this problem in a $M/M/S/N+M$ system. By giving the cost of a trunk line, the cost of a CSR, and the profit per handled call, they obtained a function $g^{S,n}$, which is the average long-run expected revenue for S CSRs and n additional trunk lines. They then gave an efficient optimization procedure to find (S,n) which maximizes $g^{S,n}$. However, since this is a model for profit centre, they did not consider the service level constraints.

Borst et al. [9] referred to this problem as "dimensioning" and solved it for the Erlang-C model. They considered two models. One is the so-called *optimization*, where they tried to minimize the sum of CSRs cost and delay cost. There are no service level constraints since they have considered delay cost, which is a kind of cost associated with service levels. Then they argued that in practice, design is rarely determined through optimization in terms of cost, since there is no standard practice for quantifying delay cost. Hence they gave the second model with constraints called *constraint satisfaction*, which can be formalized as,

$$\begin{cases} \min \ S \\ s.t.\ TSF\ constraints. \end{cases} \tag{1.1}$$

This model does not require the specific costs of CSRs and delay.

In summary, the optimal trade-off problem will be formalized into an optimization problem with or without constraints and output performance measures of the call centre performance analysis problem will be involved in the construction of objective function and constraints.

In chapter 7, we will give a design algorithm to find the smallest pair of (S,N) in a

given sense to satisfy a certain set of service level constraints. Since the emphasis of this thesis is on call centre modelling and performance analysis, we will not give the specific costs of S and N and therefore we will not solve the design problem through optimization in terms of costs. Our formulation of the design problem is similar to (1.1) with additional decision variable N and more constraints. See Chapter 7 for details of our method.

1.3 Outline and contributions of the thesis

In Chapter 2, we will give a detailed review of single-node Markovian queueing models of call centres, specifically $M/M/S/S$ model and related Erlang B formula, $M/M/S$ model and related Erlang C formula and the more general $M/M/S/N$ model. We will focus on the computational aspects of the exact performance measures of these well-known models, which is a little different from the standard textbook. Especially for $M/M/S/N$ model, based on the work of [35], we will express the performance measures in terms of Erlang B formula, which facilitates the computation. We will also prove new monotonicity properties of performance measures with respect to N. These monotonicity properties are very important to the call centre design algorithm we have developed in Chapter 7.

In Chapter 3, we will focus on the base model in the thesis: Semiopen Queueing Network Model (SOQN) proposed in [42]. We use two methods to derive the product form solution of the queue length process and the distribution of the total number of calls in the system. We propose an algorithm to compute the blocking probability. In the derivation of the waiting time distribution, we use a new method compared to the one used in the original paper [42].

In Chapter 4, we will study the balking phenomenon of call centres using both single-node Markovian models and the SOQN model. We first give a review of single-node state-dependent balking model $M(n)/M/S/N$. The main work is on the SOQN model with balking, where we prove that the queue length process still has product form solution in equilibrium and we also derive the waiting time distribution. The analysis on SOQN model with balking is new.

Another important feature of call centres is abandonment. In Chapter 5, we will study the exponential abandonment models of call centres and analyze three models, $M/M/S+M$, $M/M/S/N+M$ and SOQN+M which is SOQN model with exponential

abandonment. For single-node models: $M/M/S+M$ and $M/M/S/N+M$, we again focus on the computational aspects by expressing the exact performance measures in terms of special functions and Erlang B formula. For $M/M/S/N+M$ model, we will prove monotonicity and concavity properties of performance measures using a method that simplifies the work in [26]. For SOQN+M model, our work is a generalization and correction of [45].

Chapter 6 deals with the general abandonment models of call centres, which generalize the exponential abandonment of models studied in Chapter 5. We will first give a review of single-node general abandonment models including infinite buffer $M/M/S+G$ and finite buffer $M/M/S/N+G$ model and then study the SOQN+G model, which is SOQN model with general abandonment. For single-node models: $M/M/S+G$ and $M/M/S/N+G$, our work is based on several previous works and is a generalization of them. For example we will express the performance measures of $M/M/S/N+G$ model in terms of new building blocks. We will also study the comparison of different patience time distributions and generalized some previous results to $M/M/S/N+G$ model. The work on SOQN+G model is new.

All the above chapters are about the call centre modelling and performance analysis problems, where for each model we study the queue length distribution, waiting time distribution and related performance measures. We try to express the performance measures in terms of special functions to facilitate the computation. In Chapter 7, we will move to the second problem of call centre research: the call centre design problem, where we develop a design algorithm to determine the minimal number of CSRs (S) and trunk lines (N) to satisfy a given set of service level constraints. The algorithm is based on the monotonicity properties of some performance measures with respect to N. We also give numerical examples in the end. Finally in Chapter 8, summary and future work for further research are given.

Chapter 2

Review of Single-node Markovian Queueing Models of Call Centres

2.1 Introduction

In this chapter, we will give a detailed review of single-node Markovian queueing models of call centres. These models are Markovian, which means that the interarrival and service times are all assumed to have i.i.d. exponential distributions. In Table 2.1, we list some main Markovian queueing models and their performance measures. Note that these models are standard and can be found in most queueing textbooks such as [17]. However, we will focus on the computational aspects of the exact performance measures of these well-known models.

Model Notation	Name of Models	Performance Measures(SL)
$M/M/S/S$	Erlang B blocking/loss model	$P(blocking)=p_S$ (Erlang B formula)
$M/M/S$	Erlang C delay model	$P(delay)=P(W_q>0)$ (Erlang C formula) TSF;ASA
$M/M/S/N$	Blocking and delay model A trade-off between blocking and delay	$P(delay)=P(W_q>0)$;$P(blocking)=p_N$ TSF;ASA

Table 2.1: Some Markovian queueing models

Among these three models, $M/M/S/N$ is the most general model and the other two models are special cases. To use $M/M/S/N$ to model call centres, we assume arriving calls and CSRs are homogeneous so that we do not need to consider SBR. We also assume calls are patient and there is no balking. These two features will be studied in the later chapters. The common features of this class of models are in the following.

1. Arriving calls. There is only one class of arriving calls (homogeneous) who arrive

at the call centre according to a homogeneous Poisson process with rate λ. Arriving calls are patient and there is no balking.

2. CSRs. There is only one class of S homogeneous CSRs. The service times are i.i.d. exponentially distributed with mean μ^{-1}.

3. Routing. Arriving calls are served according to FCFS (first come first served) discipline.

4. Number of waiting places. The number of waiting places is $N-S$ with 0 in $M/M/S/S$ model and ∞ in $M/M/S$ model.

One desirable property of this class of models is that we have closed-form solutions of almost all stationary performance measures, though sometimes complicated and hard to use in terms of computation.

One type of performance measure is related to the number of calls. The Markovian property implies that these models are memoryless so that we can use a single variable (number of calls) to represent the system state in order to get a Continuous Time Markov Chain (CTMC). Let $Q(t)$ be the number of calls in the system (queue length), $Q_q(t)$ be the number of calls waiting in the queue at time t, then $Q(t)$ and $Q_q(t)$ are both birth-death processes. We are mainly concerned with stationary distributions of $Q(t)$ and $Q_q(t)$ if they exist, with corresponding variables denoted by Q and Q_q respectively. If the system is stable, one can obtain the distribution of Q, which is $p_i := P(Q = i)$, $i = 0, 1, ...$ From this, we can obtain performance measures such as $E(Q), E(Q_q)$ (called **Q1**), $E(Q_q|Q_q > 0)$ (called **Q2**). By the PASTA (Poisson Arrivals See Time Averages) property [48], we have $P(delay) = \sum_{i=S}^{\infty} p_i$ $(M/M/S)$ and $P(blocking) = p_S$ $(M/M/S/S)$ or $P(blocking) = p_N$ $(M/M/S/N)$.

Another type of performance measure is related to the stationary waiting time of a call in the queue (denoted by W_q) and in the system (denoted by W). The distributions of W_q and W are harder to obtain than that of Q. From the distribution of W or W_q, performance measures regarding the delay such as $E(W), E(W_q), E(W_q|W_q > 0) = E(W_q)/P(W_q > 0)$ (called **DLYDLY**), $TSF = P(W_q \leq AWT)$ can be calculated.

Little's formula relates the expectation of the number in the system and the delay; for example $E(Q) = \lambda E(W)$. This is a quite general result, applicable to any ergodic queueing models, even non-Markovian queueing models.

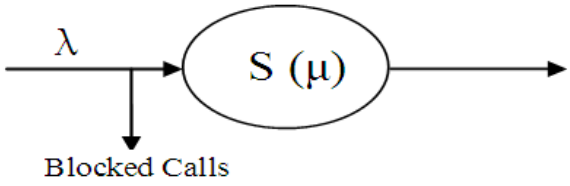

Figure 2.1: $M/M/S/S$ model description and parameters

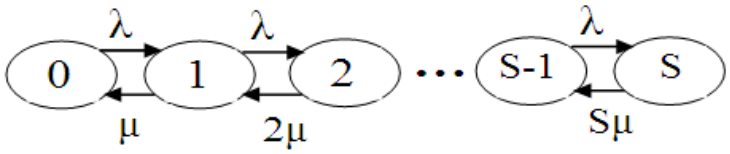

Figure 2.2: $M/M/S/S$ model stationary state transition diagram

2.2 $M/M/S/S$ model and Erlang B formula

In this section, we will give a review of $M/M/S/S$ model, which is the classical model for the analysis of telephone networks with S trunk lines and no buffers. The model description and parameters are shown in Figure 2.1. In this model, $Q(t)$ is a finite birth-death process with birth rate

$$\lambda_i = \begin{cases} \lambda & \text{if } 0 \leq i < S \\ 0 & otherwise \end{cases}$$

and state-dependent death rate $\mu_i = i\mu, i = 0, 1, ...S$. The stationary state transition diagram of $Q(t)$ is shown in Figure 2.2.

Since this is a finite buffer model, we can always obtain the stationary distribution of $Q(t)$ (no stability condition) by solving the global or cut balance equations derived from Figure 2.2. The solution is

$$p_i = \frac{a^i/i!}{\sum_{k=0}^{S} a^k/k!}, \ 0 \leq i \leq S$$

where $a := \lambda/\mu$ is called the *offered load*, a measure of demand made on the system. a is a dimensionless value but is given a unit called *erlangs*, after A.K. Erlang who laid the foundations of queueing theory [17].

Another important process is the number of calls an arriving call sees upon arrival, defined as $Q_a(t)$. The distributions $P(Q_a(t) = i)$ and $P(Q(t) = i)$ are called the arriving customer's distribution and the outside observer's distribution respectively in [17]. The following derivation comes from page 77 of [17]. Let $C(t, t+h)$ be the event that a call arrives in $(t, t+h)$. Note that this event does not imply the call will necessarily enter the system and affect the state of $Q(t)$. Then we have, for $0 \leq i \leq S$,

$$\begin{aligned} P(Q_a(t) = i) &= \lim_{h\to 0} P(Q(t) = i | C(t, t+h)) \\ &= \lim_{h\to 0} \frac{P(Q(t) = i, C(t, t+h))}{P(C(t, t+h))} \\ &= \lim_{h\to 0} \frac{P(C(t, t+h)|Q(t) = i)P(Q(t) = i)}{P(C(t, t+h))} \\ &= \lim_{h\to 0} \frac{(\lambda h + o(h))P(Q(t) = i)}{\lambda h + o(h)} \\ &= P(Q(t) = i), \end{aligned}$$

where we have used the fact that for Poisson process, $P(C(t, t+h)|Q(t) = i) = P(C(t, t+h)) = \lambda h + o(h)$. If we define $\{a_i\}$ as the stationary distribution of $Q_a(t)$, we have $a_i = p_i$, $0 \leq i \leq S$. The above property is called the PASTA property which is true for any queueing systems with Poisson arrival.

$P(blocking)$ is defined as the probability that an arriving call in equilibrium finds S calls in the system (all CSRs are busy), i.e., $P(blocking) = a_S$. By the above PASTA property we have the Erlang B formula

$$P(blocking) = a_S = p_S = B(S, a) = \frac{a^S/S!}{\sum_{k=0}^{S} a^k/k!}. \tag{2.1}$$

Note that this is a truncated Poisson distribution. Also these results are insensitive to the service distribution; it is applicable to $M/G/S/S$ model with mean service time $1/\mu$. See page 83 of [17].

In practice, Erlang B formula (2.1) brings some computation problems; it involves factorials, power functions and the denominator becomes larger and larger when S becomes larger, resulting in overflow errors unless a and S are relatively small. There is an iterative formula used to find $B(S, a)$ efficiently to solve this problem, i.e., for $S \geq 0$,

$$B(S, a) = \frac{aB(S-1, a)}{S + aB(S-1, a)}, \tag{2.2}$$

where $B(0, a) = 1$. The above formula can be used to calculate $B(1, a), B(2, a)$...until $B(S, a)$ for any S and a.

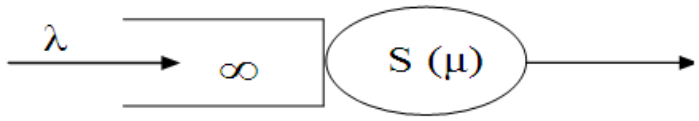

Figure 2.3: $M/M/S$ model description and parameters

Since the model is a loss model, $a' := E(Q_b) = a[1 - B(S, a)]$ is called the *carried load*, where we obtain the last equality by Little's formula applying to the number of busy servers and Q_b is a random variable representing the number of busy servers in equilibrium. The *utilization*

$$v := \frac{a'}{S} = \frac{a[1 - B(S, a)]}{S} = \rho[1 - B(S, a)] < 1$$

is the proportion of time that a server is busy, where $\rho := \frac{a}{S}$ is called the *traffic intensity*. Therefore we have an important lower bound for $B(S, a)$, i.e.,

$$B(S, a) > 1 - \frac{1}{\rho}. \tag{2.3}$$

Another important property of $B(S, a)$ is the monotonicity property of $B(S, a)$ with respect to S. $B(S, a)$ is a decreasing function of S, which is intuitively true and can be proved using the following [35]

$$B(S, a) = \frac{aB(S-1, a)}{S + aB(S-1, a)} > \frac{aB(S-1, a)}{a[1 - B(S-1, a)] + aB(S-1, a)} = B(S-1, a) \tag{2.4}$$

since

$$a[1 - B(S-1, a)] < S - 1 < S.$$

Since this model does not allow calls to wait, there is no waiting time in the queue. Waiting time in the system is just the service time. We do not need to consider the performance measures related to waiting time.

2.3 $M/M/S$ model and Erlang C formula

In this section, we will give a review of $M/M/S$ model, which allows infinite number of buffers. The model description and parameters are shown in Figure 2.3.

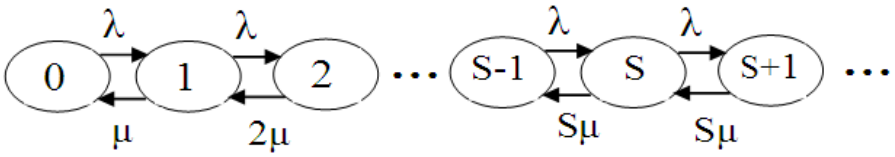

Figure 2.4: $M/M/S$ model stationary state transition diagram

2.3.1 Queue length process

In this model, $Q(t)$ is an infinite birth-death process with birth rate $\lambda_i = \lambda$ and state-dependent death rate

$$\mu_i = \begin{cases} i\mu & \text{if } 0 \le i \le S \\ S\mu & \text{if } i > S \end{cases}.$$

The stationary state transition diagram of $Q(t)$ is shown in Figure 2.4. The stability condition is $\rho = \frac{\lambda}{S\mu} = \frac{a}{S} < 1$. Since the model is not a loss model, the carried load is equal to the offered load, i.e., $a' = a$ so that the utilization $v = \rho$.

Under the stability condition: $0 < \rho < 1$, the stationary distribution can be obtained by solving the global or cut balance equations derived from Figure 2.4. The solution is

$$p_i = \begin{cases} \frac{a^i}{i!} p_0 & \text{if } 0 \le i \le S \\ \frac{a^i}{S! S^{i-S}} p_0 & \text{if } i \ge S \end{cases} \tag{2.5}$$

where

$$p_0 = \left(\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^S}{S!} \frac{1}{1-\rho} \right)^{-1}. \tag{2.6}$$

The PASTA property also holds in this model so that we have $a_i = p_i$ for $i \ge 0$. The Erlang C formula (the probability of delay of an arriving call in equilibrium) is

$$P(delay) = \sum_{i=S}^{\infty} a_i = \sum_{i=S}^{\infty} p_i = C(S,a) = \frac{\frac{a^S}{S!(1-\rho)}}{\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^S}{S!(1-\rho)}} = \frac{a^S}{S!(1-\rho)} p_0. \tag{2.7}$$

Hence

$$p_0 = \frac{C(S,a) S!(1-\rho)}{a^S}, \tag{2.8}$$

which is easier to calculate compared to (2.6) but involves $C(S,a)$. The calculation of $C(S,a)$ is based on that of $B(S,a)$ by the relationship

$$C(S,a) = \frac{B(S,a)}{B(S,a)\rho + 1 - \rho}$$

derived directly from (2.7) and (2.6) or

$$C(S,a) = \frac{\rho}{\rho + (1-\rho)/B(S-1,a)}$$

by additionally using (2.2). The above formula and (2.4) also show that $C(S,a)$ is a decreasing function of S, which is intuitively true. After calculating p_0, p_i can be calculated using (2.5) for any i. This calculation method is better than using (2.5) and (2.6) directly.

Note that $p_S = \frac{a^S}{S!}p_0$, therefore $p_S = C(S,a)(1-\rho)$ by (2.8). Substituting $p_0 = \frac{S!}{a^S}p_S$ into (2.5), we have p_i in terms of p_S, i.e.,

$$p_i = \begin{cases} \frac{S!}{a^{S-i}i!}p_S & \text{if } 0 \le i \le S \\ \rho^{i-S}p_S & \text{if } i \ge S \end{cases}. \tag{2.9}$$

Using our new expression for p_S and the above formula for p_i is even better, especially for calculating p_i, $i \ge S$ since this does not involve factorials.

2.3.2 Waiting time distribution

In order to find $TSF = P(W_q \le AWT)$, we need to find $P(W_q \le t)$: the probability of waiting time in the queue is less than or equal to $t \ge 0$ for a typical call in equilibrium. It is easy to find $P(W_q > t)$ first [24]

$$\begin{aligned} P(W_q>t) &= \sum_{i=S}^{\infty} P(W_q > t \mid i \text{ calls in the system upon arrival}) P(i \text{ calls in the system upon arrival}) \\ &= \sum_{i=S}^{\infty} P(\text{completion time of } i-S+1 \; calls > t) a_i \end{aligned}$$

Because of the PASTA property, $P(i$ calls in the system upon arrival$) = a_i = p_i$, the stationary distribution of i call in the system in equilibrium. Since $i \ge S$, the time between successive completions is exponential with rate $S\mu$. Therefore the completion time of $i - S + 1$ calls Y_i has an Erlang distribution $Er(i-S+1, S\mu)$ and the survival function of Y_i is

$$\begin{aligned} \overline{F}_{Y_i}(t) &= P(Y_i > t) \\ &= P(\text{completion time of } i-S+1 \; calls > t) \\ &= \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!}. \end{aligned}$$

It turns out that, after changing the order of two sums, we have

$$P(W_q > t) = \sum_{i=S}^{\infty} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} p_i = \frac{a^S}{S!(1-\rho)} p_0 e^{-(S\mu-\lambda)t} = C(S,a) e^{-(S\mu-\lambda)t}. \quad (2.10)$$

Also

$$E(W_q) = \int_0^{\infty} P(W_q > t) dt = \frac{C(S,a)}{S\mu - \lambda} \quad (2.11)$$

and $P(W_q = 0) = 1 - C(S,a)$. By Little's formula, we also have the mean number of calls waiting in the queue

$$E(Q_q) = \lambda E(W_q) = \frac{\rho C(S,a)}{1-\rho}.$$

Hence, both $TSF = P(W_q \leq AWT)$, $ASA = E(W_q)$ and $E(Q_q)$ can be expressed by $C(S,a)$, which is very useful in computation.

One simple staffing method is to find the minimum number of S under the SL constraint: $ASA \leq ASA^*$, where ASA^* is a given time, i.e., $S^* = min\{S \mid ASA \leq ASA^*\}$. Since $C(S,a)$ is a decreasing function of S we have ASA is a decreasing function of S, so that S^* is well-defined.

In practice, $M/M/S$ is commonly used in call centres for modelling, performance analysis, and staffing because of its closed-form solutions of almost all interested performance measures. It is usually assumed that the arrival and service rate are piece-wise constant. Then $M/M/S$ is applied to each time interval using the parameters belonging to that interval. Also the stationary performance measures are used since experience shows that this type of stochastic processes converges fast to its stationary state [29].

There are some shortcomings using the $M/M/S$ model. First, it assumes there are infinite number of buffers, hence no blocked calls. If blocking is not a rare event in a call centre, $M/M/S$ is not a good model and we may choose $M/M/S/N$ model. Second, it does not consider abandonment, which is shown in [21] to be a very important issue. The simplest abandon model is $M/M/S + M$.

2.4 $M/M/S/N$ model

In this section, we will give a review of $M/M/S/N$ model. Also based on the work of [35], we will express the performance measures in terms of Erlang B formula, which facilitates the computation. We will also prove new monotonicity properties of some performance measures with respect to N.

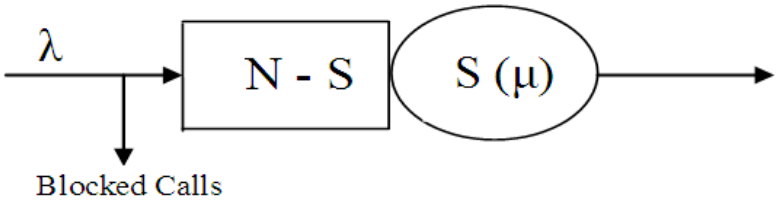

Figure 2.5: $M/M/S/N$ model description and parameters

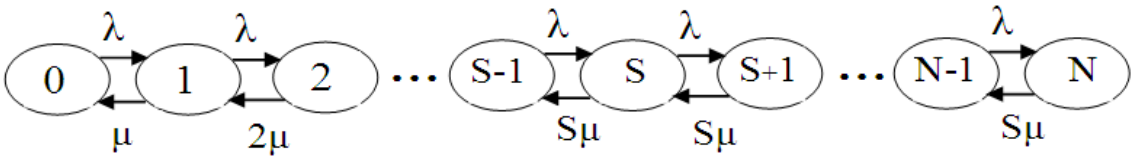

Figure 2.6: $M/M/S/N$ model stationary state transition diagram

$M/M/S/N$ is a more realistic model for call centres, which allows both delay and blocking. Recall that S represents the number of CSRs and N represents the number of trunk lines (the number of buffers plus CSRs). If $N = \infty$, it reduces to $M/M/S$ model and if $N = S$, it reduces to $M/M/S/S$ model. Recall that in Section 1.2.1, we use $M/M/S/N$ as the basis of the natural queueing model of call centres. The model description and parameters are shown in Figure 2.5.

2.4.1 Queue length process

In this model, $Q(t)$ is a finite birth-death process with birth rate

$$\lambda_i = \begin{cases} \lambda & \text{if } 0 \leq i < N \\ 0 & otherwise \end{cases}$$

and state-dependent death rate

$$\mu_i = \begin{cases} i\mu & \text{if } 0 \leq i \leq S \\ S\mu & \text{if } S \leq i \leq N \end{cases}.$$

The stationary state transition diagram of $Q(t)$ is shown in Figure 2.6. Since this is a finite buffer model, the stationary distribution of $Q(t)$ can be obtained by solving the global or

cut balance equations derived from Figure 2.6 (no stability condition). The solution is

$$p_i = \begin{cases} \frac{a^i}{i!} p_0 & \text{if } 0 \le i \le S \\ \frac{a^i}{S! S^{i-S}} p_0 & \text{if } S \le i \le N \end{cases} = p_0 \gamma(i), \tag{2.12}$$

where

$$p_0 = \begin{cases} \left(\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^S}{S!} \frac{1-\rho^{N-S+1}}{1-\rho} \right)^{-1}, & \text{if } \rho \neq 1 \\ \left(\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^S}{S!} (N-S+1) \right)^{-1}, & \text{if } \rho = 1 \end{cases} = \left[\sum_{i=0}^{N} \gamma(i) \right]^{-1}$$

and

$$\gamma(i) := \begin{cases} \frac{a^i}{i!}, & \text{if } 0 \le i \le S \\ \frac{a^i}{S! S^{i-S}}, & \text{if } S \le i \le N \end{cases}.$$

In particular, we have the delay probability: $P(delay) = \sum_{i=S}^{N-1} p_i$, the blocking probability: $P(blocking) = p_N$ and $P(\text{no-delay}) = 1 - P(blocking) - P(delay)$ by the PASTA property.

Since $p_S = \frac{a^S}{S!} p_0$, substituting $p_0 = \frac{S!}{a^S} p_S$ into (2.12), we have p_i in terms of p_S, i.e.,

$$p_i = \begin{cases} \frac{S!}{a^{S-i} i!} p_S & \text{if } 0 \le i \le S \\ \rho^{i-S} p_S & \text{if } S \le i \le N \end{cases} \tag{2.13}$$

where for $\rho \neq 1$,

$$\begin{aligned} p_S &= \left[\sum_{i=0}^{S} \frac{S!}{a^{S-i} i!} + \sum_{i=S+1}^{N} \rho^{i-S} \right]^{-1} \\ &= \left[\frac{1}{B(S,a)} + \frac{\rho(1-\rho^{N-S})}{1-\rho} \right]^{-1} \\ &= \frac{(1-\rho) B(S,a)}{1-\rho+\rho B(S,a)(1-\rho^{N-S})} \end{aligned}$$

and for $\rho = 1$,

$$p_S = \frac{B(S,S)}{1 + (N-S) B(S,S)}. \tag{2.14}$$

Then for $\rho \neq 1$,

$$\begin{aligned} P(blocking) = p_N &= \rho^{N-S} p_S \\ &= \frac{(1-\rho) B(S,a) \rho^{N-S}}{1-\rho+\rho B(S,a)(1-\rho^{N-S})}. \end{aligned} \tag{2.15}$$

Similarly

$$\begin{aligned} P(delay) &= \sum_{i=S}^{N-1} p_i = \sum_{i=S}^{N-1} \rho^{i-S} p_S = \frac{1-\rho^{N-S}}{1-\rho} p_S \\ &= \frac{1-\rho^{N-S}}{1-\rho} \frac{(1-\rho)B(S,a)}{1-\rho+\rho B(S,a)(1-\rho^{N-S})} \\ &= \frac{B(S,a)(1-\rho^{N-S})}{1-\rho+\rho B(S,a)(1-\rho^{N-S})}. \end{aligned} \tag{2.16}$$

Finally

$$\begin{aligned} P(\text{no-delay}) &= 1 - P(blocking) - P(delay) \\ &= \frac{(1-\rho)[1-B(S,a)]}{1-\rho+\rho B(S,a)(1-\rho^{N-S})}. \end{aligned} \tag{2.17}$$

We have expressed the above performance measures in terms of $B(S,a)$ and we can also use (2.2) to get expressions in terms of $B(S-1,a)$

$$P(blocking) = \frac{(1-\rho)B(S-1,a)\rho^{N-S+1}}{1-\rho+\rho B(S-1,a)(1-\rho^{N-S+1})},$$

$$P(delay) = \frac{\rho B(S-1,a)(1-\rho^{N-S})}{1-\rho+\rho B(S-1,a)(1-\rho^{N-S+1})},$$

and

$$P(\text{no-delay}) = \frac{1-\rho}{1-\rho+\rho B(S-1,a)(1-\rho^{N-S+1})}.$$

When $\rho = 1$, we can obtain the corresponding formulas similarly by (2.14). We can also apply L'Hospital's rule in the above formulas (2.15,2.16,2.17). The results are listed in the following.

1. $P(blocking) = \frac{B(S,S)}{1+(N-S)B(S,S)} = \frac{B(S-1,S)}{1+(N-S+1)B(S-1,S)}$.
2. $P(delay) = \frac{(N-S)B(S,S)}{1+(N-S)B(S,S)} = \frac{(N-S)B(S-1,S)}{1+(N-S+1)B(S-1,S)}$.
3. $P(\text{no-delay}) = \frac{1-B(S,S)}{1+(N-S)B(S,S)} = \frac{1}{1+(N-S+1)B(S-1,S)}$.

Remark 2.4.1 *Note that the above formulas (2.15) and (2.17) are also given in [35].*

In the following we will study the special cases of the above performance measures. The following new result is obvious.

Theorem 2.4.1 *When $N = S$, the model reduces to $M/M/S/S$ model and we have $P(blocking) = B(S,a)$, $P(delay) = 0$ and $P(no\text{-}delay) = 1 - B(S,a)$.*

We have the following results when $N \to \infty$.

Theorem 2.4.2 *When $N \to \infty$, we have:*

1. $\lim_{N\to\infty} P(blocking) = \begin{cases} 1 - \frac{1}{\rho}, & if\ \rho \geq 1 \\ 0, & if\ 0 < \rho \leq 1 \end{cases}$.

2. $\lim_{N\to\infty} P(delay) = \begin{cases} \frac{1}{\rho}, & if\ \rho \geq 1 \\ C(S,a), & if\ 0 < \rho < 1 \end{cases}$.

3. $\lim_{N\to\infty} P(no\text{-}delay) = \begin{cases} 0, & if\ \rho \geq 1 \\ 1 - C(S,a), & if\ 0 < \rho < 1 \end{cases}$.

Proof. The first result is also given in [35] and the other two are new. We only give the proof of the second result and others can be proved similarly. For $0 < \rho < 1$, we have

$$P(delay) = \frac{B(S,a)(1-\rho^{N-S})}{1-\rho+\rho B(S,a)(1-\rho^{N-S})}$$

approaches to $\frac{B(S,a)}{1-\rho+\rho B(S,a)} = C(S,a)$ when $N \to \infty$. For $\rho = 1$,

$$P(delay) = \frac{(N-S)B(S,S)}{1+(N-S)B(S,S)} = \frac{B(S,S)}{\frac{1}{N-S}+B(S,S)}$$

approaches to 1 when $N \to \infty$. For $\rho > 1$,

$$\begin{aligned} P(delay) &= \frac{B(S,a)(1-\rho^{N-S})}{1-\rho+\rho B(S,a)(1-\rho^{N-S})} \\ &= \frac{B(S,a)\frac{1-\rho^{N-S}}{\rho^{N-S}}}{\frac{1-\rho}{\rho^{N-S}}+B(S,a)\frac{1-\rho^{N-S}}{\rho^{N-S-1}}} \end{aligned}$$

approaches to $\frac{1}{\rho}$ when $N \to \infty$. ∎

Remark 2.4.2 *When $0 < \rho < 1$, performance measures in this theorem reduce to the corresponding performance measures of $M/M/S$ model.*

Since the model is a loss model, the carried load is, for $\rho \neq 1$,

$$\begin{aligned} a' &= E(Q_b) \\ &= a\,[1 - P(blocking)] \\ &= \frac{a - a\rho + a\rho B(S,a)(1-\rho^{N-S-1})}{1-\rho+\rho B(S,a)(1-\rho^{N-S})} \end{aligned}$$

The utilization

$$v = \frac{a'}{S} = \frac{\rho - \rho^2 + \rho^2 B(S,a)(1-\rho^{N-S-1})}{1-\rho+\rho B(S,a)(1-\rho^{N-S})} < 1$$

is the proportion of time that a server is busy.

For $\rho = 1$,

$$a' = S\frac{1+(N-S-1)B(S,S)}{1+(N-S)B(S,S)}$$

and

$$v = \frac{a'}{S} = \frac{1+(N-S-1)B(S,S)}{1+(N-S)B(S,S)}.$$

$M/M/S/N$ can be thought of as a $M/M/S$ model conditioning that no more N calls are in the system. We can get (2.13) using (2.9) by conditional argument.

2.4.2 Waiting time distribution

We define $\overline{W}_q$ as the stationary waiting time in the queue for all calls (the blocked calls have ∞ waiting time). Therefore $\overline{W}_q$ has a mass at ∞ and $P(\overline{W}_q = \infty) = P(blocking)$. Also $\overline{W}_q$ has a mass at 0 and $P(\overline{W}_q = 0) = P(\text{no-delay})$. This definition also appeared in Stolletz [44]. We have, for $t > 0$,

$$\begin{aligned} P(\overline{W}_q > t) &= P(\overline{W}_q > t, \text{non-blocking}) + P(\overline{W}_q > t, blocking) \\ &= P(\overline{W}_q > t, \text{non-blocking}) + P(blocking). \end{aligned}$$

To find $P(\overline{W}_q > t, \text{non-blocking})$, we use the same idea as in $M/M/S$ model

$$\begin{aligned} &P(\overline{W}_q > t, \text{non-blocking}) \\ &= \sum_{i=S}^{N-1} P(\overline{W}_q > t, \text{non-blocking} \mid i \text{ calls in system upon arrival}) P(i \text{ calls in system upon arrival}) \\ &= \sum_{i=S}^{N-1} P(\text{completion time of } i-S+1 \text{ calls} > t) a_i \\ &= \sum_{i=S}^{N-1} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} p_i. \end{aligned}$$

Typically we are more concerned with the conditional waiting time of a call given that this call is not blocked. Let W_q be this waiting time, which has no mass at ∞, and we have

$$\begin{aligned} P(W_q{>}t) = P(\overline{W}_q > t | \text{non-blocking}) &= \frac{P(\overline{W}_q > t, \text{non-blocking})}{P(\text{non-blocking})} \\ &= \sum_{i=S}^{N-1} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \frac{p_i}{1-p_N} = \sum_{i=S}^{N-1} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} q_i. \end{aligned} \tag{2.18}$$

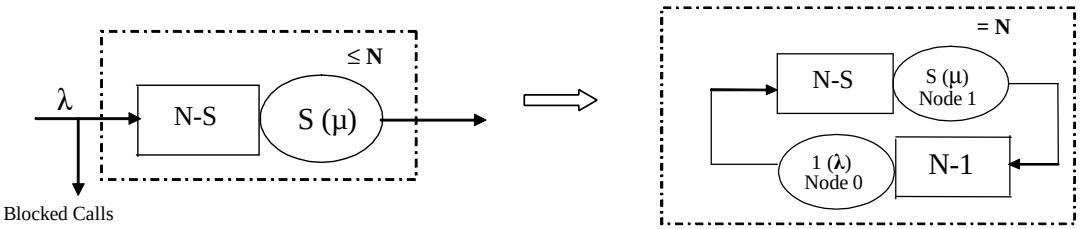

Figure 2.7: From SOQN to CQN for $M/M/S/N$ model

Here $q_i := \frac{p_i}{1-p_N}, 0 \leq i \leq N-1$ is the probability of a call finding i calls in the system given that it is not blocked, which is called the arrival-point probability in [24], page 77. q_i can be derived alternatively by applying the following idea, which was used to solve SOQN model in [14]. We can think of $M/M/S/N$ model as a SOQN model with one node and at most N calls in the system. By introducing a fictitious node (Node 0) which has one server with service rate

$$\mu_0(j) = \begin{cases} \lambda & if \quad j > 0 \\ 0 & if \quad j = 0 \end{cases},$$

we convert the model to an equivalent two-node CQN and they have the same stationary distribution (2.12). See Figure 2.7 for details.

Now we can use the Arrival Theorem of CQN [7] to get an alternative derivation of q_i. The Arrival Theorem shows that for product form CQN, arrivals to Node 1 (in this case) see the same distribution as the stationary distribution of Node 1 of the same CQN with one less call. Therefore we have, by (2.12), for $0 \leq i \leq N-1$

$$\begin{aligned} q_i &= P(\text{arrival finds } i \text{ calls at Node 1}) \\ &= p_i^{(N-1)} = p_0^{(N-1)}\gamma(i) = \frac{\gamma(i)}{\sum_{i=0}^{N-1}\gamma(i)} \\ &= \frac{p_0\gamma(i)}{\sum_{i=0}^{N-1}p_0\gamma(i)} = \frac{p_i}{1-p_N}, \end{aligned}$$

where $p_i^{(N-1)}$ is the stationary distribution of Node 1 having i calls of the same CQN with $N-1$ calls.

Therefore by (2.18) and (2.13) we obtain, for $\rho \neq 1$

$$\begin{aligned} P(W_q{>}t) &= \sum_{i=S}^{N-1} q_i \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \\ &= \frac{p_S}{1-p_N} \sum_{i=S}^{N-1} \rho^{i-S} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \\ &= \frac{p_S}{1-p_N} \sum_{j=0}^{N-S-1} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \sum_{i=j}^{N-S-1} \rho^i \\ &= \frac{p_S e^{-S\mu t}}{1-p_N} \sum_{j=0}^{N-S-1} \frac{(S\mu t)^j}{j!} \frac{\rho^j - \rho^{N-S}}{1-\rho} \end{aligned} \tag{2.19}$$

and for $\rho = 1$,

$$\begin{aligned} P(W_q{>}t) &= \frac{p_S e^{-\lambda t}}{1-p_N} \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j}{j!} (N-S-j) \\ &= \frac{p_S}{1-p_N} \left[e^{-\lambda t}(N-S) \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j}{j!} - e^{-\lambda t} \sum_{j=1}^{N-S-1} \frac{(\lambda t)^j}{j!} j \right] \\ &= \frac{p_S}{1-p_N} \left[(N-S)(1-P(N-S,\lambda t)) - \lambda t (1-P(N-S-1,\lambda t)) \right] \end{aligned} \tag{2.20}$$

where $P(N-S,\lambda t) = \frac{\gamma(N-S,\lambda t)}{\Gamma(N-S)}$ is the regularized Gamma function [2] and we have used the well-known identity,

$$e^{-\lambda t} \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j}{j!} = 1 - P(N-S,\lambda t). \tag{2.21}$$

Note that

$$\gamma(x,y) := \int_0^y t^{x-1} e^{-t} dt, x > 0, y \geq 0$$

is the incomplete Gamma function and

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt, x > 0$$

is the Gamma function.

In addition, the following performance measures can be easily obtained in terms of q_i.

1. $P(\text{no-delay}|\text{non-blocking}) = P(W_q = 0) = \sum_{i=0}^{S-1} q_i$.
2. $P(delay|\text{non-blocking}) = P(W_q > 0) = \sum_{i=S}^{N-1} q_i$.
3. $ASA = E(W_q) = \int_0^\infty P(W_q > t)dt = \int_0^\infty \sum_{i=S}^{N-1} q_i \overline{F}_{Y_i}(t)dt = \sum_{i=S}^{N-1} q_i \int_0^\infty \overline{F}_{Y_i}(t)dt = \frac{1}{S\mu} \sum_{i=S}^{N-1} q_i(i-S+1)$.

Note that the previous results (2.19) and (2.20) are standard and can be found in many queueing textbooks.

Following the same idea as the last section, to facilitate the computation and analysis, we will express the above performance measures in terms of $B(S,a)$. For example, for $\rho \neq 1$,

$$\begin{aligned} P(delay|\text{non-blocking}) = P(W_q > 0) &= \sum_{i=S}^{N-1} q_i \\ &= \sum_{i=S}^{N-1} \frac{p_i}{1-p_N} = \frac{P(delay)}{1-P(blocking)} \\ &= \frac{1-\rho+\rho B(S,a)(1-\rho^{N-S})}{1-\rho+\rho B(S,a)(1-\rho^{N-S-1})} \frac{B(S,a)(1-\rho^{N-S})}{1-\rho+\rho B(S,a)(1-\rho^{N-S})} \\ &= \frac{B(S,a)(1-\rho^{N-S})}{1-\rho+\rho B(S,a)(1-\rho^{N-S-1})}. \end{aligned}$$

Similarly

$$\begin{aligned} P(\text{no-delay}|\text{non-blocking}) = P(W_q = 0) \\ &= 1 - P(delay|\text{non-blocking}) \\ &= \frac{(1-\rho)[1-B(S,a)]}{1-\rho+\rho B(S,a)(1-\rho^{N-S-1})} \end{aligned}$$

and the above two formulas are also given in [35]. Finally

$$\begin{aligned} ASA = E(W_q) \\ &= \frac{1}{S\mu} \sum_{i=S}^{N-1} q_i(i-S+1) \\ &= \frac{p_S}{S\mu(1-p_N)} \sum_{i=0}^{N-S-1} \rho^i(i+1) \\ &= \frac{p_S}{S\mu(1-p_N)} \left(\sum_{i=0}^{N-S-1} \rho^{i+1} \right)' \\ &= \frac{p_S}{S\mu(1-p_N)} \frac{1-(N-S+1)\rho^{N-S}+(N-S)\rho^{N-S+1}}{(1-\rho)^2} \\ &= \frac{1}{S\mu(1-p_N)} \cdot \frac{B(S,a)(1-\rho)}{1-\rho+\rho B(S,a)(1-\rho^{N-S})} \cdot \frac{1-(N-S+1)\rho^{N-S}+(N-S)\rho^{N-S+1}}{(1-\rho)^2} \\ &= \frac{1}{S\mu(1-p_N)} \frac{B(S,a)[1-(N-S+1)\rho^{N-S}+(N-S)\rho^{N-S+1}]}{[1-\rho+\rho B(S,a)(1-\rho^{N-S})](1-\rho)} \\ &= \frac{B(S,a)[1-(N-S+1)\rho^{N-S}+(N-S)\rho^{N-S+1}]}{[1-\rho+\rho B(S,a)(1-\rho^{N-S-1})](1-\rho)S\mu}. \end{aligned} \tag{2.22}$$

When $\rho = 1$, we can obtain the corresponding formulas similarly. We can also apply L'Hospital's rule in the above formulas. The results are listed in the following.

1. $P(W_q{>}t) = \frac{B(S,S)}{1+(N-S-1)B(S,S)} \left[(N-S)(1-P(N-S,\lambda t)) - \lambda t(1-P(N-S-1,\lambda t))\right]$.
2. $P(\text{no-delay}|\text{non-blocking}) = \frac{1-B(S,S)}{1+(N-S-1)B(S,S)}$.
3. $P(delay|\text{non-blocking}) = \frac{(N-S)B(S,S)}{1+(N-S-1)B(S,S)}$.
4. $ASA = E(W_q) = \frac{(N-S)(N-S+1)B(S,S)}{2\lambda[1+(N-S-1)B(S,S)]}$.

The distribution of W_q can also be expressed in terms of $B(S,a)$ with another derivation. Note that this method has been used in [17] page 95 for $M/M/S$ model. We have that $P(W_q{>}t)$ can be factorized as follows,

$$\begin{aligned} P(W_q{>}t) =& P(W_q{>}t|W_q{>}0)P(W_q{>}0) \\ =& P(W_q{>}t|W_q{>}0)P(delay|\text{non-blocking}), \end{aligned}$$

where for $\rho \neq 1$,

$$\begin{aligned} P(W_q{>}t|W_q{>}0) &= P(W_q{>}t|S \leq Q \leq N-1) \\ &= \sum_{i=0}^{N-S-1} P(W_q{>}t|Q = S+i; S \leq Q \leq N-1)P(Q = S+i|S \leq Q \leq N-1) \\ &= \sum_{i=0}^{N-S-1} \left(\sum_{j=0}^{i} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \right) \frac{\rho^i}{1+\rho+...+\rho^{N-S-1}} \\ &= \sum_{j=0}^{N-S-1} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \sum_{i=j}^{N-S-1} \frac{\rho^i}{1+\rho+...+\rho^{N-S-1}} \\ &= \sum_{j=0}^{N-S-1} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \frac{\rho^j - \rho^{N-S}}{1-\rho^{N-S}} \\ &= \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-S\mu t}}{j!} \frac{1-\rho^{N-S-j}}{1-\rho^{N-S}}. \end{aligned}$$

So that we have for $\rho \neq 1$

$$P(W_q{>}t) = P(delay|\text{non-blocking}) \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-S\mu t}}{j!} \frac{1-\rho^{N-S-j}}{1-\rho^{N-S}}, t \geq 0. \tag{2.23}$$

In the above derivation, we use the fact that

$$P(Q = S+i|S \leq Q \leq N-1) = \frac{\rho^i}{1+\rho+...+\rho^{N-S-1}}, \text{ for } i = 0,1,...N-S-1, \tag{2.24}$$

which can be proved using the definition of conditional probability and (2.12). For $\rho = 1$,

$$P(W_q{>}t) = P(delay|\text{non-blocking}) \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-\lambda t}}{j!} \left[1 - \frac{j}{N-S}\right]. \tag{2.25}$$

It can be shown that (2.23), (2.25) and (2.19), (2.20) are equivalent respectively. Note that the above formulas (2.23) and (2.25) are also given in [35] using a similar method.

Also using (2.23),

$$\begin{aligned}
ASA = E(W_q) &= \int_0^\infty P(W_q{>}\, t)dt \\
&= P(delay|\text{non-blocking}) \sum_{j=0}^{N-S-1} \frac{\rho^j - \rho^{N-S}}{1-\rho^{N-S}} \int_0^\infty \frac{(S\mu t)^j e^{-S\mu t}}{j!} dt \\
&= P(delay|\text{non-blocking}) \sum_{j=0}^{N-S-1} \frac{\rho^j - \rho^{N-S}}{(1-\rho^{N-S})S\mu} \\
&= P(delay|\text{non-blocking}) \frac{1-\rho^{N-S}(1+(1-\rho)(N-S))}{(1-\rho)(1-\rho^{N-S})S\mu} \\
&= \frac{B(S,a)(1-\rho^{N-S})}{1-\rho+\rho B(S,a)(1-\rho^{N-S-1})} \frac{1-\rho^{N-S}(1+(1-\rho)(N-S))}{(1-\rho)(1-\rho^{N-S})S\mu} \\
&= \frac{B(S,a)[1-\rho^{N-S}(1+(1-\rho)(N-S))]}{[1-\rho+\rho B(S,a)(1-\rho^{N-S-1})](1-\rho)S\mu}
\end{aligned}$$

which is the same as (2.22).

By Little's formula, we also have the mean number of calls waiting in the queue

$$\begin{aligned}
E(Q_q) &= \lambda E(\overline{W}_q, \text{non-blocking}) \\
&= \lambda E(\overline{W}_q|\text{non-blocking}) P(\text{non-blocking}) \\
&= \lambda E(W_q)\left[1 - P(blocking)\right].
\end{aligned}$$

Therefore we have

$$E(Q_q) = \begin{cases} \frac{\rho B(S,a)[1-(N-S+1)\rho^{N-S}+(N-S)\rho^{N-S+1}]}{(1-\rho)[1-\rho+\rho B(S,a)(1-\rho^{N-S})]}, & \text{if } \rho \neq 1 \\ \frac{(N-S)(N-S+1)B(S,S)}{2[1+(N-S)B(S,S)]}, & \text{if } \rho = 1. \end{cases}$$

In the following we will study the special cases of the above performance measures. The following new results are obvious.

Theorem 2.4.3 *When $N = S$, the model reduces to $M/M/S/S$ model and we have $P(delay|non\text{-}blocking)$=0, $P(no\text{-}delay|non\text{-}blocking) = 1$, $P(W_q{>}t) = 0$, $E(Q_q) = 0$ and $ASA = 0$.*

Theorem 2.4.4 *When $N \to \infty$, we have:*

1. $\lim_{N\to\infty} P(delay|non\text{-}blocking) = \begin{cases} 1, & if\ \rho \geq 1 \\ C(S,a), & if\ 0 < \rho < 1 \end{cases}$.

2. $\lim_{N\to\infty} P(no\text{-}delay|non\text{-}blocking) = \begin{cases} 0, & if\ \rho \geq 1 \\ 1 - C(S,a), & if\ 0 < \rho < 1 \end{cases}$.

3. $\lim_{N\to\infty} P(W_q > t) = \begin{cases} 1, & if\ \rho \geq 1 \\ C(S,a)e^{-(S\mu-\lambda)t}, & if\ 0 < \rho < 1 \end{cases}$.

4. $\lim_{N\to\infty} ASA = \begin{cases} \infty, & if\ \rho \geq 1 \\ \frac{C(S,a)}{S\mu-\lambda}, & if\ 0 < \rho < 1 \end{cases}$.

5. $\lim_{N\to\infty} E(Q_q) = \begin{cases} \infty, & if\ \rho \geq 1 \\ \frac{\rho C(S,a)}{1-\rho}, & if\ 0 < \rho < 1 \end{cases}$.

Remark 2.4.3 *When $0 < \rho < 1$, performance measures in this theorem reduce to the corresponding performance measures of $M/M/S$ model.*

2.4.3 Monotonicity properties

We have studied the special cases of the performance measures when $N \to \infty$ or $N = S$ for $M/M/S/N$ model. However it is also important to study the monotonicity properties of some performance measures with respect to N. We will see in Chapter 7 that these monotonicity properties are very important to the call centre design algorithm we will develop. Some of the following results have been proved in [35] using different methods.

Theorem 2.4.5 *When S and other parameters are fixed, $P(blocking)$ is a strictly decreasing function of N.*

Proof. When $\rho = 1, P(blocking) = \frac{B(S,S)}{1+(N-S)B(S,S)}$. The property is obviously true. For $\rho \neq 1$, we have

$$P_N(blocking) = \frac{(1-\rho)B(S,a)\rho^{N-S}}{1-\rho+\rho B(S,a)(1-\rho^{N-S})}.$$

We only need to prove that for $N \geq S$,

$$\frac{P_{N+1}(blocking)}{P_N(blocking)} = \frac{\rho\left[1-\rho+\rho B(S,a)(1-\rho^{N-S})\right]}{1-\rho+\rho B(S,a)(1-\rho^{N-S+1})} < 1. \tag{2.26}$$

For $\rho > 1$, the above (2.26) is equivalent to

$$\rho\left[1-\rho+\rho B(S,a)(1-\rho^{N-S})\right] > 1-\rho+\rho B(S,a)(1-\rho^{N-S+1})$$

or the well-known inequality (2.3)

$$B(S,a) > 1-\frac{1}{\rho}.$$

For $0 < \rho < 1$, the above (2.26) is equivalent to

$$\rho\left[1-\rho+\rho B(S,a)(1-\rho^{N-S})\right] < 1-\rho+\rho B(S,a)(1-\rho^{N-S+1}),$$

which is again

$$B(S,a) > 1-\frac{1}{\rho}.$$

■

Remark 2.4.4 *This result has been proved in [35] using a similar method.*

Theorem 2.4.6 *When S and other parameters are fixed, $P(nodelay|non\text{-}blocking)$ is a strictly decreasing function of N and $P(nodelay)$ is a strictly decreasing function of N as well.*

Proof. We have

$$P(nodelay|\text{non-blocking})=\begin{cases} \frac{(1-\rho)[1-B(S,a)]}{1-\rho+\rho B(S,a)(1-\rho^{N-S-1})}, & \text{if } \rho \neq 1 \\ \frac{1-B(S,S)}{1+(N-S-1)B(S,S)}, & \text{if } \rho = 1 \end{cases}.$$

Therefore, the theorem holds for $\rho = 1$ and $0 < \rho < 1$. For $\rho > 1$, we have

$$P(nodelay|\text{non-blocking})=\frac{(\rho-1)[1-B(S,a)]}{\rho-1+\rho B(S,a)(\rho^{N-S-1}-1)}.$$

Hence the theorem also holds.

Since

$$P(nodelay)=\begin{cases} \frac{(1-\rho)[1-B(S,a)]}{1-\rho+\rho B(S,a)(1-\rho^{N-S})}, & \text{if } \rho \neq 1 \\ \frac{1-B(S,S)}{1+(N-S)B(S,S)}, & \text{if } \rho = 1 \end{cases},$$

the proof is similar for $P(nodelay)$. ■

Since $P(delay|\text{non-blocking}) = 1 - P(nodelay|\text{non-blocking})$, we have the following corollary, which is also proved in [35] directly.

Corollary 2.4.1 *When S and other parameters are fixed, $P(delay|non\text{-}blocking)$ is a strictly increasing function of N.*

Since

$$P(delay) = [1 - P(blocking)]\, P(delay|\text{non-blocking}),$$

we have, by Theorem 2.4.5 and Corollary 2.4.1, the following new result.

Theorem 2.4.7 *When S and other parameters are fixed, $P(delay)$ is a strictly increasing function of N.*

For the waiting time distribution, we have the following result.

Theorem 2.4.8 *When S and other parameters are fixed, $P(W_q{>}t)$ is a strictly increasing function of N.*

Proof. From (2.23) and (2.25), we have

$$P(W_q{>}t) = \begin{cases} P(delay|\text{non-blocking}) \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-S\mu t}}{j!} \frac{1-\rho^{N-S-j}}{1-\rho^{N-S}}, & \text{if } \rho \neq 1 \\ P(delay|\text{non-blocking}) \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-\lambda t}}{j!} \left[1 - \frac{j}{N-S}\right], & \text{if } \rho = 1 \end{cases}.$$

For $\rho = 1$, since $P(delay|\text{non-blocking})$ is a strictly increasing function of N we only need to prove

$$f(N) := \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-\lambda t}}{j!} \left[1 - \frac{j}{N-S}\right]$$

is a strictly increasing function of N. We have

$$\begin{aligned} f(N+1) &= \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-\lambda t}}{j!} \left[1 - \frac{j}{N+1-S}\right] + \frac{(\lambda t)^{N-S} e^{-\lambda t}}{(N-S)!} \left[1 - \frac{N-S}{N+1-S}\right] \\ &> \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-\lambda t}}{j!} \left[1 - \frac{j}{N-S}\right] = f(N) \end{aligned}$$

since each term in the above is positive and for $0 \leq j \leq N-S-1$,

$$1 - \frac{j}{N+1-S} \geq 1 - \frac{j}{N-S}.$$

For $\rho \neq 1$, since $P(delay|\text{non-blocking})$ is a strictly increasing function of N we only need to prove

$$g(N) := \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-S\mu t}}{j!} \frac{1-\rho^{N-S-j}}{1-\rho^{N-S}}$$

is a strictly increasing function of N. We have

$$g(N+1) = \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-S\mu t}}{j!} \frac{1-\rho^{N+1-S-j}}{1-\rho^{N+1-S}} + \frac{(\lambda t)^{N-S} e^{-S\mu t}}{(N-S)!} \frac{1-\rho}{1-\rho^{N+1-S}}$$
$$> \sum_{j=0}^{N-S-1} \frac{(\lambda t)^j e^{-S\mu t}}{j!} \frac{1-\rho^{N-S-j}}{1-\rho^{N-S}} = g(N)$$

since each term in the above is positive and for $0 \leq j \leq N-S-1$, we can prove that

$$\frac{1-\rho^{N+1-S-j}}{1-\rho^{N+1-S}} \geq \frac{1-\rho^{N-S-j}}{1-\rho^{N-S}}. \tag{2.27}$$

We will prove (2.27) in the following. (2.27) is equivalent to

$$\frac{1-\rho^{N+1-S-j}}{1-\rho^{N+1-S}} - \frac{1-\rho^{N-S-j}}{1-\rho^{N-S}}$$
$$= \frac{(1-\rho^{N-S})(1-\rho^{N+1-S-j}) - (1-\rho^{N+1-S})(1-\rho^{N-S-j})}{(1-\rho^{N+1-S})(1-\rho^{N-S})} \geq 0.$$

We only need to prove

$$(1-\rho^{N-S})(1-\rho^{N+1-S-j}) - (1-\rho^{N+1-S})(1-\rho^{N-S-j}) \geq 0$$

since the denominator is positive for $\rho \neq 1$. The above is equivalent to

$$(\rho^j - 1)(\rho - 1) \geq 0,$$

which is true for $\rho \neq 1$. ∎

Remark 2.4.5 *Note that this theorem agrees with Corollary 2.4.1, since*

$$P(W_q > 0) = P(delay | non\text{-}blocking).$$

Remark 2.4.6 *This result is also proved in [35] using a method involving stochastic ordering.*

Corollary 2.4.2 *When S and other parameters are fixed, $ASA = E(W_q)$ and $E(Q_q)$ are both strictly increasing functions of N.*

2.4.4 Numerical examples

To give some numerical illustrations of the above monotonicity properties for $M/M/S/N$ model, we will consider some numerical examples in the following. Let $\lambda = 8, \mu = 1$ be the

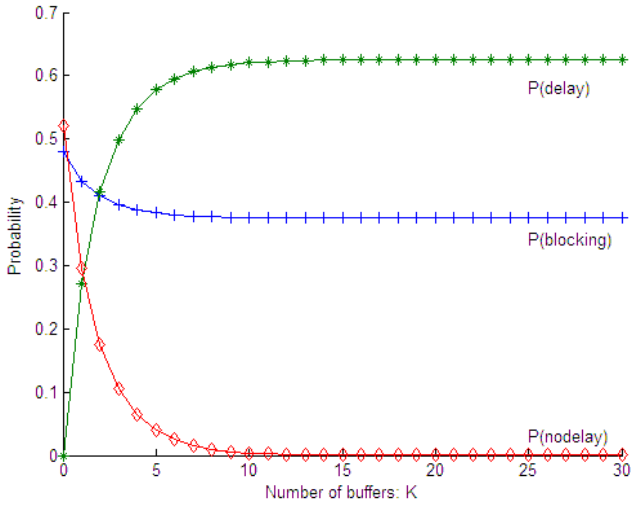

Figure 2.8: $P(blocking), P(nodelay)$ and $P(delay)$ of Example 2.1 for $M/M/S/N$ model

common parameters. We consider three cases: Example 2.1 has $S = 5$ with $\rho = \frac{\lambda}{S\mu} > 1$; Example 2.2 has $S = 9$ with $0 < \rho < 1$ and Example 2.3 has $S = 8$ with $\rho = 1$. Let buffer size $K = N - S$ change from 0 to 30. We compute the results using Matlab.

We first consider three unconditional probabilities: $P(blocking), P(nodelay)$ and $P(delay)$. The results are shown in Figure 2.8 for Example 2.1, in Figure 2.9 for Example 2.2 and in Figure 2.10 for Example 2.3. We find that in all three examples, these probabilities have the monotonicity properties we have proved. For the limiting cases when $N \to \infty$, Theorem 2.4.2 can also be verified. In Example 2.1, $\rho^{-1} = 0.625$ so that $\lim_{N\to\infty} P(blocking) = 1-\rho^{-1} = 0.375$; $\lim_{N\to\infty} P(delay) = \rho^{-1} = 0.625$ and $\lim_{N\to\infty} P(\text{no-delay}) = 0$. In Example 2.2, $0 < \rho < 1$ so that $\lim_{N\to\infty} P(blocking) = 0$; $\lim_{N\to\infty} P(delay) = C(S,a) = 0.653$ and $\lim_{N\to\infty} P(\text{no-delay}) = 1 - C(S,a) = 0.347$. In Example 2.3, $\rho^{-1} = 1$ so that $\lim_{N\to\infty} P(blocking) = 0$; $\lim_{N\to\infty} P(delay) = 1$ and $\lim_{N\to\infty} P(\text{no-delay}) = 0$.

Next we will consider the conditional probabilities: $P(nodelay|\text{non-blocking}), P(W_q{>}t)$ and $P(delay|\text{non-blocking})$, where $t = 0.5$. The results are shown in Figure 2.11 for Example 2.1, in Figure 2.12 for Example 2.2 and in Figure 2.13 for Example 2.3. We find that in all three examples, these probabilities have the monotonicity properties we have proved. For the limiting cases when $N \to \infty$, Theorem 2.4.4 can also be verified. In Example 2.1 and 2.3, $\rho \geq 1$ so that $\lim_{N\to\infty} P(delay|\text{non-blocking}) = 1$; $\lim_{N\to\infty} P(nodelay|\text{non-blocking}) =$

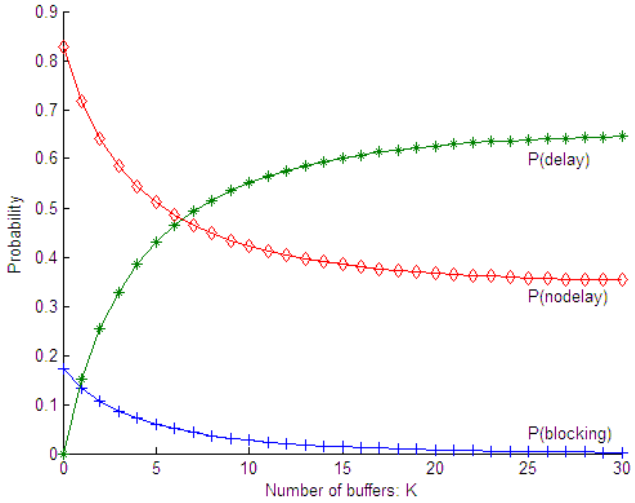

Figure 2.9: $P(blocking), P(nodelay)$ and $P(delay)$ of Example 2.2 for $M/M/S/N$ model

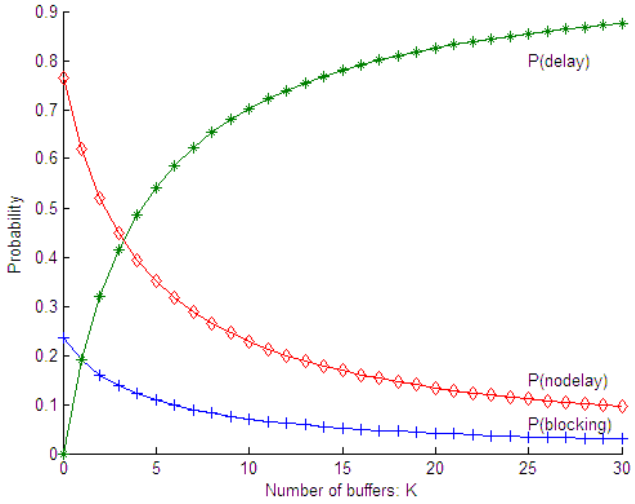

Figure 2.10: $P(blocking), P(nodelay)$ and $P(delay)$ of Example 2.3 for $M/M/S/N$ model

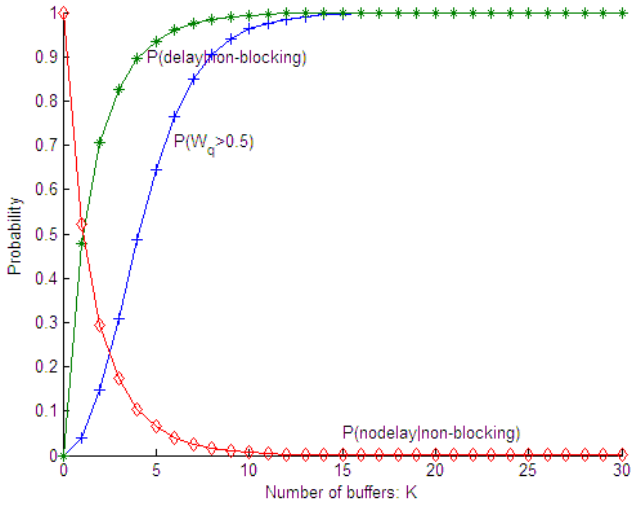

Figure 2.11: $P(nodelay|\text{non-blocking})$, $P(delay|\text{non-blocking})$ and $P(W_q > 0.5)$ of Example 2.1 for $M/M/S/N$ model

0 and $\lim_{N\to\infty} P(W_q > 0.5) = 1$. In Example 2.2, $0 < \rho < 1$ so that

$$\lim_{N\to\infty} P(delay|\text{non-blocking}) = C(S,a) = 0.653;$$

$$\lim_{N\to\infty} P(nodelay|\text{non-blocking}) = 1 - C(S,a) = 0.347$$

and

$$\lim_{N\to\infty} P(W_q > 0.5) = C(S,a)e^{-(S\mu-\lambda)0.5} = 0.396.$$

2.5 Summary

In this chapter, we gave a detailed review of single-node Markovian queueing models of call centres including $M/M/S/S$ model with Erlang B formula, $M/M/S$ model with Erlang C formula and the more general $M/M/S/N$ model. We focused on the computational aspects of the exact performance measures of these well-known models. Especially for $M/M/S/N$ model, we expressed performance measures in terms of Erlang B formula, which facilitates the computation as well as analysis. Based on this analysis, we proved monotonicity properties for performance measures with respect to N. These properties

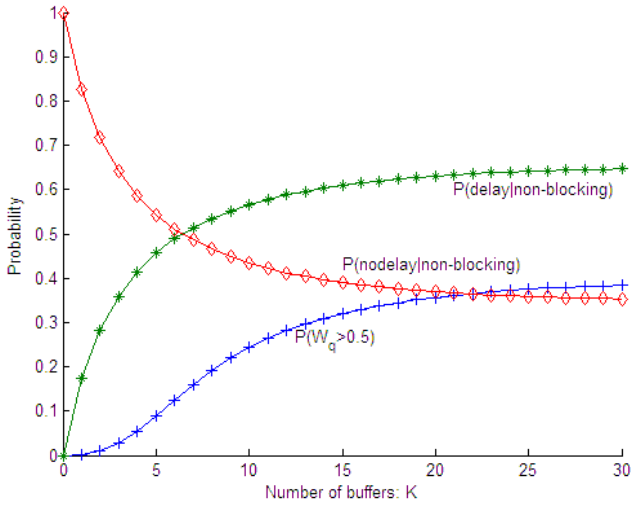

Figure 2.12: $P(nodelay|\text{non-blocking}), P(delay|\text{non-blocking})$ and $P(W_q > 0.5)$ of Example 2.2 for $M/M/S/N$ model

have been verified using numerical examples and are important to the call centre design algorithm we will develop in Chapter 7.

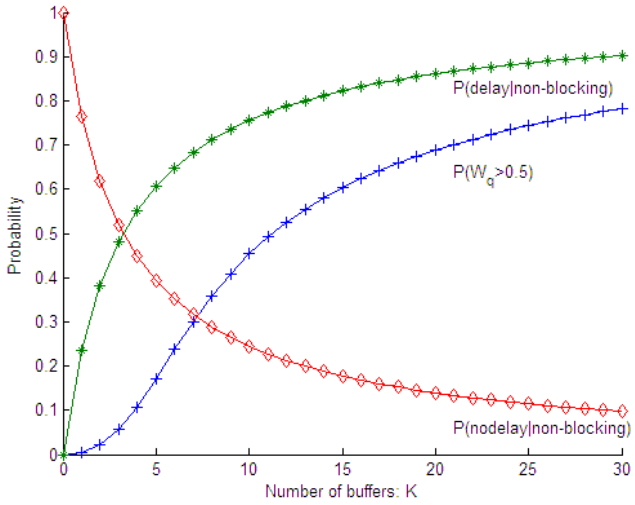

Figure 2.13: $P(nodelay|\text{non-blocking})$, $P(delay|\text{non-blocking})$ and $P(W_q > 0.5)$ of Example 2.3 for $M/M/S/N$ model

Chapter 3

SOQN Model of Call Centres

As we have mentioned in Section 1.2.1, the SOQN model proposed by Srinivasan et al. [42] will be our base model. In this chapter we will review the work of [42] in detail and obtain new results. The model is a natural extension of $M/M/S/N$ model.

3.1 Modelling motivation and model description

The following modelling motivation and model description is adapted from [42]. Recall that an inbound call centre consists of an IVRU to provide some self-service to calls and a large proportion of calls only need self-service with IVRU and leave the system without requiring the service of CSRs for some call centres. For example a customer may call to a bank call centre to check the account balance, transfer money only using IVRU without requiring the service of CSRs. However most traditional call centre models ignore this part and model the call centre as a single-node queue, as we have seen in Chapter 2. Srinivasan et al. [42] first proposed and analyzed a two-node network model to capture the role of the IVRU as well as the CSRs. The thesis by Khudyakov [27] from Technion-Israel Institute of Technology used this model. The Ph.D. research proposal by Yom-Tov [49] also used a similar model. However their focus is on the QED (Quality and Efficiency Driven) approximation on the performance measures of the model.

We will repeat Figure 1.3 as Figure 3.1 here to show the model description. The arriving calls come to the system according to a Poisson process with rate λ. If the arriving call finds less than N calls in the system (the system is not full), it is admitted into the system and it is processed by IVRU immediately. Otherwise it is blocked and leaves the system. There are N IVRU servers and the processing times are assumed to be i.i.d. exponential random variables with rate θ. After finishing the service with an IVRU server, the call may leave the system with probability $\overline{p} = 1 - p$ or it proceeds to request service from a

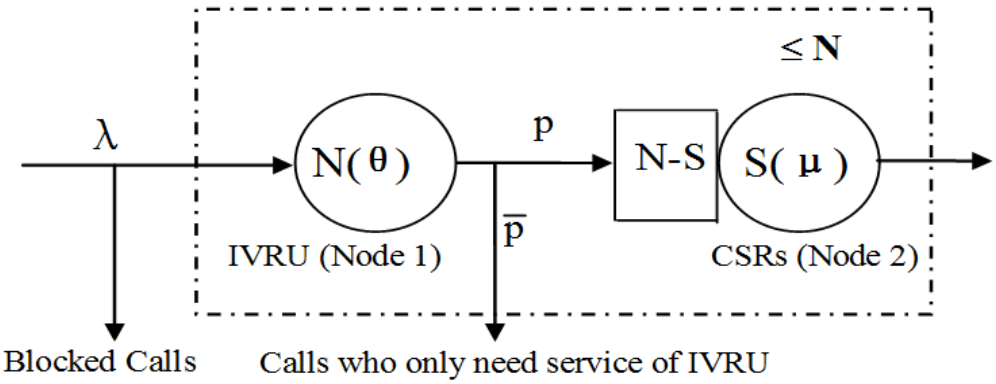

Figure 3.1: Semiopen queueing network model description and parameters

CSR with probability p. If a CSR is free, the call is served, otherwise it waits for a CSR and joins a queue. There are $S(\leq N)$ CSRs and the service times are assumed to be i.i.d. exponential random variables with rate μ and are independent of the arrival process and the IVRU processing times. Once the call is processed by a CSR it releases both the CSR and the trunk line and leaves the system.

Note that this model is a network model with two nodes in series and at most N calls in the system. Node 1 models the IVRU with N servers and no queue since there are at most N calls in the system. It also can be thought of as having infinite servers i.e., a $M/M/\infty$ queue since there are no calls waiting for service at this node. Node 2 models the CSRs with S servers and the queue in front of them. This node can be thought of as a $M/M/S$ queue since there is no blocking at this node.

3.2 Product form solution of the queue length process

Let $Q(t) = \{Q_1(t), Q_2(t)\}$ represent the number of calls at time t at Node 1 and 2 respectively, i.e., the queue length process. Note that $Q_1(t) + Q_2(t) \leq N$ for all $t \geq 0$. From the description of the model, we know that $Q(t)$ is a finite CTMC and has a stationary distribution with corresponding variables denoted by $Q = \{Q_1, Q_2\}$. The state space of Q is $\Omega = \{(i,j)|i+j \leq N, (i,j) \in Z_+^2\}$, where there are i calls at Node 1 and j calls at Node 2. In fact the system can be thought of as a flow controlled Jackson network or SOQN, which has a product form solution for the stationary distribution $\pi_{ij} := P(Q_1 = i, Q_2 = j)$ as stated in [42]. We will show how to obtain the product form solution π_{ij} using two

methods in the following.

3.2.1 Direct method

In this section, we will use the result for SOQN in [14] to obtain the product form solution. The traffic equations for our SOQN model are

$$\begin{cases} \lambda_1 = \lambda \\ \lambda_2 = \lambda_1 p \end{cases}.$$

Hence, the unique solution is $\lambda_1 = \lambda$ and $\lambda_2 = \lambda p$, which give the effective arrival rates to Node 1 and 2. The service rates are state-dependent, i.e. $\mu_1(m) = m\theta, 0 \leq m \leq N$ and

$$\mu_2(m) = \begin{cases} m\mu & if \quad 0 \leq m \leq S \\ S\mu & if \quad S \leq m \leq N \end{cases}.$$

We define

$$M_i(n) = \begin{cases} \prod_{m=1}^{n} \mu_i(m) & if \quad n > 0 \\ 1 & if \quad n = 0 \end{cases}, \ i = 1, 2.$$

Hence, $M_1(n) = n!\theta^n, 0 \leq n \leq N$ and

$$M_2(n) = \begin{cases} n!\mu^n & if \quad 0 \leq n \leq S \\ S^{n-S}S!\mu^n & if \quad S \leq n \leq N \end{cases}.$$

For $i = 1, 2$, we define the mutually independent random variables Y_i with probability mass function as follows

$$P(Y_i = n) = P(Y_i = 0)\frac{\lambda_i^n}{M_i(n)}, n = 0, 1, 2, ...$$

Theorem 2.5 in [14] page 22 states that the stationary distribution of the queue length process is:

For all $(i, j) \in \Omega$,

$$\begin{aligned} \pi_{ij} &= \frac{1}{P(Y_1 + Y_2 \leq N)} P(Y_1 = i) P(Y_2 = j) \\ &= \frac{1}{P(Y_1 + Y_2 \leq N)} P(Y_1 = 0)\frac{\lambda_1^i}{M_1(i)} P(Y_2 = 0)\frac{\lambda_2^j}{M_2(j)} \\ &= \frac{P(Y_1 = 0)P(Y_2 = 0)}{P(Y_1 + Y_2 \leq N)} \frac{\lambda^i}{M_1(i)} \frac{\lambda^j p^j}{M_2(j)} \\ &= \pi_{00} \frac{\lambda^i}{M_1(i)} \frac{\lambda^j p^j}{M_2(j)}, \end{aligned} \tag{3.1}$$

where the normalizing constant is

$$\pi_{00}^{-1} = \frac{P(Y_1 + Y_2 \leq N)}{P(Y_1 = 0)P(Y_2 = 0)} = \frac{\sum\limits_{0 \leq i+j \leq N} P(Y_1 = i)P(Y_2 = j)}{P(Y_1 = 0)P(Y_2 = 0)} \\ = \sum_{0 \leq i+j \leq N} \frac{\lambda^i}{M_1(i)} \frac{\lambda^j p^j}{M_2(j)}. \tag{3.2}$$

3.2.2 Method of CQN

The SOQN model can also be viewed as an open queueing network with two nodes and state-dependent arrival rate. For example in our case, the state-dependent arrival rate is

$$\lambda(i,j) = \begin{cases} \lambda & if \quad i+j < N \\ 0 & otherwise \end{cases}.$$

To find the stationary distribution of the queue length process for this open network with state-dependent arrival rate, we introduce a fictitious node (Node 0) which has one server with service rate

$$\mu_0(m) = \begin{cases} \lambda & if \quad m > 0 \\ 0 & if \quad m = 0 \end{cases}.$$

The state-dependent arrival rate is simply modeled by having only N calls circulating in the network. In this way, our model is converted to an equivalent three-node closed network with the same stationary distribution. See Figure 3.2 for details. Note that this idea was used to solve SOQN model in [14] and we have converted the $M/M/S/N$ model to an equivalent two-node CQN model in Chapter 2 using the same idea. We just need to find the stationary distribution for the CQN in order to find the π_{ij}. We will use the procedure for CQN in [14] to obtain the product form solution.

The traffic equations of the closed network are

$$\begin{cases} v_1 = v_0 \\ v_2 = v_1 p \\ v_0 = v_1 \overline{p} + v_2 \end{cases}. \tag{3.3}$$

By letting $v_1 = v_0 = \lambda$, we can easily get the solution: $v_1 = v_0 = \lambda$ and $v_2 = \lambda p$. These are the effective arrival rates for these three nodes if we see them in isolation. The service rates are $\mu_i(m), i = 0, 1, 2$, as we defined before. Also $M_0(n) = \lambda^n, n > 0$. For $i = 0, 1, 2$, we define the mutually independent random variables Y_i with probability mass function as

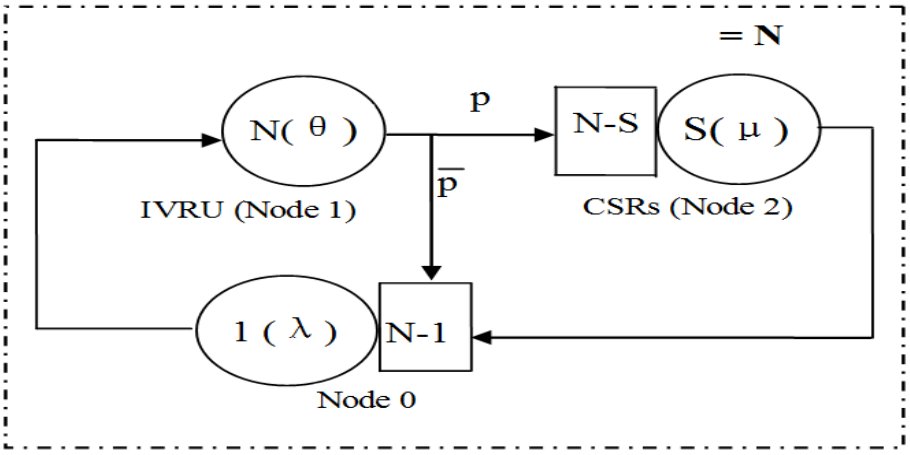

Figure 3.2: The equivalent CQN for the SOQN model

follows

$$P(Y_i = n) = P(Y_i = 0)\frac{v_i^n}{M_i(n)}, n = 0, 1, 2, ...$$

For Node 0 $(M/M/1/N)$, the stationary distribution is

$$P(Y_0 = k) = P(Y_0 = 0)\frac{\lambda^k}{M_0(k)} = P(Y_0 = 0)\frac{\lambda^k}{\lambda^k} = P(Y_0 = 0), \quad 0 \le k \le N$$
$$i.e., \; P(Y_0 = k) = P(Y_0 = 0) = \frac{1}{N+1}, \quad 0 \le k \le N.$$

For Node 1 $(M/M/N/N)$, the stationary distribution is

$$P(Y_1 = i) = P(Y_1 = 0)\frac{\lambda^i}{M_1(i)}, \; 0 \le i \le N.$$

For Node 2 $(M/M/S/N)$, the stationary distribution is

$$P(Y_2 = j) = P(Y_2 = 0)\frac{(\lambda p)^j}{M_2(j)}, \; 0 \le j \le N.$$

Hence by Theorem 2.3 in [14] page 20, we have the stationary distribution of the CQN:

For all $(k, i, j) \in \mathcal{Z}_+^3$ such that $i + j + k = N$,

$$\begin{aligned}\pi_{kij} &= \frac{1}{P(Y_0 + Y_1 + Y_2 = N)}P(Y_0 = k)P(Y_1 = i)P(Y_2 = j) \\ &= \frac{P(Y_0 = 0)P(Y_1 = 0)P(Y_2 = 0)}{P(Y_0 + Y_1 + Y_2 = N)}\frac{\lambda^i}{M_1(i)}\frac{\lambda^j p^j}{M_2(j)} \\ &= \pi_{00}\frac{\lambda^i}{M_1(i)}\frac{\lambda^j p^j}{M_2(j)}\end{aligned}$$

where the normalizing constant is

$$\begin{aligned}
\pi_{00}^{-1} &= \frac{P(Y_0+Y_1+Y_2=N)}{P(Y_0=0)P(Y_1=0)P(Y_2=0)} \\
&= \frac{\sum\limits_{k+i+j=N} P(Y_0=k)P(Y_1=i)P(Y_2=j)}{P(Y_0=0)P(Y_1=0)P(Y_2=0)} \\
&= \frac{P(Y_0=0)\sum\limits_{k+i+j=N} P(Y_1=i)P(Y_2=j)}{P(Y_0=0)P(Y_1=0)P(Y_2=0)} \\
&= \frac{\sum\limits_{0\le i+j\le N} P(Y_1=i)P(Y_2=j)}{P(Y_1=0)P(Y_2=0)} \\
&= \sum_{0\le i+j\le N} \frac{\lambda^i}{M_1(i)}\frac{\lambda^j p^j}{M_2(j)},
\end{aligned}$$

which is the same as (3.2). Note that the above distribution π_{kij} is actually a two dimensional distribution and it is the stationary distribution of the SOQN. If we let $k = N-i-j$, the above can be written equivalently as

$$\begin{aligned}
&\pi_{ij} = \pi_{00}\tfrac{\lambda^i}{M_1(i)}\tfrac{\lambda^j p^j}{M_2(j)} \\
&\forall (i,j) \in \Omega,
\end{aligned}$$

which is the same as (3.1).

Remark 3.2.1 *π_{kij} is actually the solution of the following global balance equations*

$$\begin{aligned}
&\pi_{kij}(\mu_0(k)+\mu_1(i)+\mu_2(j)) \qquad (3.4)\\
&= \pi_{(k+1)(i-1)j}\delta(i)\mu_0(k+1) + \pi_{(k-1)(i+1)j}\delta(k)\mu_1(i+1)\overline{p} \\
&+\pi_{k(i+1)(j-1)}\delta(j)\mu_1(i+1)p + \pi_{(k-1)i(j+1)}\delta(k)\mu_2(j+1) \\
&\forall (k,i,j) \in \mathcal{Z}_+^3 \text{ such that } k+i+j=N \\
&\text{where } \delta(i) = \begin{cases} 1 & if \quad i>0 \\ 0 & if \quad i=0 \end{cases}.
\end{aligned}$$

Remark 3.2.2 *The solution of (3.3) is not unique. For example if we let $v_1 = v_0 = 1$, we obtain the solution: $v_1 = v_0 = 1$ and $v_2 = p$. However we will obtain the same π_{kij}.*

In the following we will express the solution using the notation by Srinivasan et al. [42] and obtain the marginal distribution, which are not given in [42]. Let

$$\beta(j) := \begin{cases} j! & for\ 0 \le j \le S \\ S!S^{j-S} & for\ S \le j \le N \end{cases};$$

$a_1 = \frac{\lambda}{\theta}; a_2 = \frac{\lambda p}{\mu}$, then $M_2(j) = \beta(j)\mu^j$ and

$$\pi_{ij} = \pi_{00}\frac{\lambda^i}{M_1(i)}\frac{\lambda^j p^j}{M_2(j)} = \pi_{00}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}, (i,j) \in \Omega,$$

where

$$\begin{aligned}\pi_{00}^{-1} &= \sum_{0\le i+j\le N}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)} = \sum_{k=0}^{N}\sum_{j=0}^{k}\frac{a_1^{k-j}}{(k-j)!}\frac{a_2^j}{\beta(j)} \\ &= \sum_{k=0}^{S}\sum_{j=0}^{k}\frac{a_1^{k-j}a_2^j}{(k-j)!j!} + \sum_{k=S+1}^{N}\left(\sum_{j=0}^{S}\frac{a_1^{k-j}a_2^j}{(k-j)!j!} + \sum_{j=S+1}^{k}\frac{a_1^{k-j}a_2^j}{(k-j)!S!S^{j-S}}\right).\end{aligned}$$

Applying binomial formula, the following is obtained in [42],

$$\begin{aligned}\pi_{00}^{-1} &= \sum_{k=0}^{S}\frac{(a_1+a_2)^k}{k!} + \sum_{k=S+1}^{N}\frac{(a_1+a_2)^k}{k!} + \sum_{k=S+1}^{N}\sum_{j=S+1}^{k}\left(-\frac{a_1^{k-j}a_2^j}{(k-j)!j!} + \frac{a_1^{k-j}a_2^j}{(k-j)!S!S^{j-S}}\right) \\ &= \sum_{k=0}^{N}\frac{(a_1+a_2)^k}{k!} + \sum_{k=S+1}^{N}\sum_{j=S+1}^{k}\frac{a_1^{k-j}a_2^j}{(k-j)!}\left(\frac{1}{S!S^{j-S}} - \frac{1}{j!}\right).\end{aligned}$$

The marginal distribution for Node 1 is

$$\pi_{i*} := P(Q_1 = i) = \sum_{j=0}^{N-i}\pi_{ij} = \pi_{00}\sum_{j=0}^{N-i}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}, 0 \le i \le N$$

where π_{00}^{-1} has another expression $\pi_{00}^{-1} = \sum_{i=0}^{N}\sum_{j=0}^{N-i}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}$. Hence the mean number of calls at Node 1 is

$$E(Q_1) = \sum_{i=0}^{N} i\pi_{i*} = \pi_{00}\sum_{i=0}^{N} i\sum_{j=0}^{N-i}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}. \tag{3.5}$$

The marginal distribution for Node 2 is

$$\pi_{*j} := P(Q_2 = j) = \sum_{i=0}^{N-j}\pi_{ij} = \pi_{00}\sum_{i=0}^{N-j}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}, 0 \le j \le N$$

where π_{00}^{-1} has the third expression $\pi_{00}^{-1} = \sum_{j=0}^{N}\sum_{i=0}^{N-j}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}$. Hence the mean number of calls at Node 2 is

$$E(Q_2) = \sum_{j=0}^{N} j\pi_{*j} = \pi_{00}\sum_{j=0}^{N} j\sum_{i=0}^{N-j}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}$$

and the mean number of calls waiting in the queue at Node 2 is

$$E(Q_{2q}) = \sum_{j=S+1}^{N}(j-S)\pi_{*j} = \pi_{00}\sum_{j=S+1}^{N}(j-S)\sum_{i=0}^{N-j}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}.$$

3.3 The stationary distribution of the total number of calls in the system

3.3.1 Direct method

The following direct method is given by [42]. The stationary distribution of the total number of calls in the system (denoted by $Q = Q_1 + Q_2$) can be derived by the stationary distribution of the queue length process π_{ij} using the relationship

$$\pi_k := P(Q = k) = \sum_{j=0}^{k} \pi_{(k-j)j}.$$

There are two cases:

1. $0 \leq k \leq S$:

$$\pi_k = \pi_{00} \sum_{j=0}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} = \pi_{00} \frac{(a_1 + a_2)^k}{k!}.$$

2. $S < k \leq N$:

$$\begin{aligned} \pi_k &= \pi_{00} \left(\sum_{j=0}^{S} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!S!S^{j-S}} \right) \\ &= \pi_{00} \left(\frac{(a_1 + a_2)^k}{k!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \left(\frac{1}{S!S^{j-S}} - \frac{1}{j!} \right) \right). \end{aligned}$$

Therefore, the probability π_k that there are exactly $0 \leq k \leq N$ calls in the system is equal to

$$\pi_k = \pi_{00} \left(\frac{(a_1 + a_2)^k}{k!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \left(\frac{1}{S!S^{j-S}} - \frac{1}{j!} \right) I_{(S,\infty)}(k) \right) \tag{3.6}$$

where $I_A(k) = \begin{cases} 1 & if\ k \in A \\ 0 & otherwise \end{cases}$ is the indicator function and $\pi_0 = \pi_{00}$. The mean number of total calls in the system, which is not available in [42], is

$$E(Q) = \sum_{k=0}^{N} k\pi_k = \sum_{k=0}^{N} k \sum_{j=0}^{k} \pi_{(k-j)j} = E(Q_1) + E(Q_2).$$

The blocking probability is given in [42] using the PASTA property,

$$P(blocking) = \pi_N = \sum_{j=0}^{N} \pi_{(N-j)j}.$$

In the following, we will derive the utilization for Node 1 which is not available in [42]. We can write down Little's formula for Node 1

$$E(Q_1) = \lambda \left[1 - P(blocking)\right] \frac{1}{\theta} = a_1 \left[1 - P(blocking)\right],$$

which is easy to verify since, by (3.5),

$$\begin{aligned}\frac{E(Q_1)}{a_1} &= \pi_{00} \sum_{i=0}^{N} i \sum_{j=0}^{N-i} \frac{a_1^{i-1}}{i!} \frac{a_2^j}{\beta(j)} = \pi_{00} \sum_{i=1}^{N} \sum_{j=0}^{N-i} \frac{a_1^{i-1}}{(i-1)!} \frac{a_2^j}{\beta(j)} \\ &= \pi_{00} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1-i} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1-i} \pi_{ij} = 1 - P(blocking).\end{aligned}$$

Hence the carried load for Node 1 is $E(Q_1) = a_1 \left[1 - P(blocking)\right]$. The utilization

$$v_1 = \frac{E(Q_1)}{N} = \frac{a_1 \left[1 - P(blocking)\right]}{N} < 1$$

is the proportion of time that an IVRU server is busy.

3.3.2 Throughput method

In the following, we will use the throughput method in [14] to derive this distribution. By Proposition 2.7 in [14] page 25, the stationary distribution of the total number of calls in SOQN follows the same distribution as the number of calls in a birth-death queue with the following specification: constant arrival rate λ and state-dependent service rates that are equal to the throughput function, $TH(k)$, of the same network operating in a closed fashion with $1 \leq k \leq N$ calls.

According to [14], the corresponding CQN of our SOQN still has two nodes and the effective arrival rates are $w_i = \lambda_i/\lambda$, i.e. $w_1 = 1, w_2 = p$. Define

$$g_1(k) = \frac{w_1^k}{M_1(k)} = \frac{1}{k!\theta^k}, 0 \leq k \leq N,$$

and

$$g_2(k) = \frac{w_2^k}{M_2(k)} = \begin{cases} \frac{p^k}{k!\mu^k} & if \quad 0 \leq k \leq S \\ \frac{p^k}{S^{k-S}S!\mu^k} & if \quad S \leq k \leq N \end{cases}.$$

In [14], the throughput of this CQN with k calls in the system is defined as $TH(k) = \frac{G(2,k-1)}{G(2,k)}, 1 \leq k \leq N$, where

$$\begin{aligned}G(2,k) &= \sum_{i+j=k} g_1(i) g_2(j) = \sum_{j=0}^{k} g_1(k-j) g_2(j) \\ &= \sum_{j=0}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^k}{M_2(j)}. \end{aligned} \tag{3.7}$$

Let $\pi_k, 0 \leq k \leq N$ be the stationary distribution of the total number of calls in the SOQN. By Proposition 2.7 in [14] page 25, π_k is the stationary distribution of a birth-death process with birth rate λ and death rate $TH(k) = \frac{G(2,k-1)}{G(2,k)}$.

Hence, the total number of calls in the SOQN has the following distribution,

$$\begin{aligned}\pi_k &= \pi_0 \frac{\lambda^k}{\prod_{j=1}^{k} TH(j)} = \pi_0 \frac{\lambda^k}{\prod_{j=1}^{k} \frac{G(2,j-1)}{G(2,j)}} = \pi_0 \frac{\lambda^k}{\frac{G(2,0)}{G(2,k)}} \\ &= \pi_0 \frac{\lambda^k}{\frac{1}{G(2,k)}} = \pi_0 \lambda^k G(2,k), 0 \leq k \leq N, \end{aligned} \tag{3.8}$$

where the normalizing constant π_0 is $\left[\sum_{k=0}^{N} \lambda^k G(2,k)\right]^{-1}$. Since $M_2(j)$ has different expression for different interval of j, we have from (3.7) for $0 \leq k \leq S$,

$$\begin{aligned} G(2,k) &= \sum_{j=0}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{M_2(j)} \\ &= \sum_{j=0}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{j!\mu^j} = \frac{1}{k!}(\frac{1}{\theta} + \frac{p}{\mu})^k = \frac{(a_1+a_2)^k}{\lambda^k k!} \end{aligned} \tag{3.9}$$

and for $S < k \leq N$,

$$\begin{aligned} G(2,k) &= \sum_{j=0}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{M_2(j)} \\ &= \sum_{j=0}^{S} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{j!\mu^j} + \sum_{j=S+1}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{S^{j-S}S!\mu^j} \\ &= \sum_{j=0}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{j!\mu^j} - \sum_{j=S+1}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{j!\mu^j} + \sum_{j=S+1}^{k} \frac{1}{(k-j)!\theta^{k-j}} \frac{p^j}{S^{j-S}S!\mu^j} \\ &= \frac{1}{k!}\left(\frac{1}{\theta} + \frac{p}{\mu}\right)^k + \sum_{j=S+1}^{k} \frac{1}{(k-j)!} \left(\frac{1}{\theta}\right)^{k-j} \left(\frac{p}{\mu}\right)^j \left(\frac{1}{S^{j-S}S!} - \frac{1}{j!}\right) \\ &= \frac{(a_1+a_2)^k}{\lambda^k k!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{\lambda^k (k-j)!} \left(\frac{1}{S^{j-S}S!} - \frac{1}{j!}\right) \end{aligned} \tag{3.10}$$

Substituting $G(2,k)$ back to (3.8), we get the same expression as (3.6).

3.3.3 Calculation of the blocking probability

Assuming that the number of CSRs S is fixed, we are interested in the blocking probability while the number of trunk lines N is increased by one starting from S. Let $K = N - S =$

$0, 1, 2...$The blocking probability is denoted by $\beta_K := \pi_{S+K} = \pi_N$. From (3.8), we have

$$\beta_K = \pi_0^{(K)} \lambda^{S+K} G(2, S+K),$$

where $\pi_0^{(K)} = \left[\sum_{i=0}^{S+K} \lambda^i G(2,i) \right]^{-1}$ is the normalizing constant when the buffer size is K.

When $K = 0$,

$$\pi_0^{(0)} = \left[\sum_{i=0}^{S} \lambda^i G(2,i) \right]^{-1} = \left[\sum_{i=0}^{S} \frac{(a_1+a_2)^i}{i!} \right]^{-1}$$

and

$$\beta_0 = \left[\sum_{i=0}^{S} \frac{(a_1+a_2)^i}{i!} \right]^{-1} \lambda^S G(2,S) = \frac{\frac{(a_1+a_2)^S}{S!}}{\sum_{i=0}^{S} \frac{(a_1+a_2)^i}{i!}} = B(S, a_1+a_2)$$

is the Erlang B formula. Note that

$$\begin{aligned} \pi_0^{(K+1)} &= \left[\sum_{i=0}^{S+K+1} \lambda^i G(2,i) \right]^{-1} = \left[\sum_{i=0}^{S+K} \lambda^i G(2,i) + \lambda^{S+K+1} G(2, S+K+1) \right]^{-1} \\ &= \left[\left(\pi_0^{(K)} \right)^{-1} + \lambda^{S+K+1} G(2, S+K+1) \right]^{-1}, K \geq 0. \end{aligned}$$

We have the following algorithm to calculate $\beta_1, \beta_2, \beta_3...$until β_K, where $K > 0$.

Algorithm 3.3.1 *Algorithm to get blocking probabilities $\beta_1, \beta_2, \beta_3...$until β_K for $K > 0$*

1 Initialize: $\beta_0 = B(S, a_1+a_2); \pi_0^{(0)} = \frac{\beta_0}{\frac{(a_1+a_2)^S}{S!}}$;

2 FOR $n = 0$ *to* $K-1$

2.1 Compute increment $= \lambda^{S+n+1} G(2, S+n+1)$

2.2 Compute $\pi_0^{(n+1)} = \left[\left(\pi_0^{(n)} \right)^{-1} + increment \right]^{-1}$

2.3 Compute $\beta_{n+1} = \pi_0^{(n+1)} * increment$

ENDFOR

3.3.4 Other performance measures

In this section, we will derive new results not available in [42]. We can obtain other performance measures from the stationary marginal distribution of Node 2. When we look

at Node 2 only, we have

$$\begin{aligned}
1 &= \pi_{*N} + (1-p)\sum_{j=0}^{N-1}\pi_{*j} + p\sum_{j=0}^{N-1}\pi_{*j} \\
&= \pi_{0N} + (1-p)\sum_{j=0}^{N-1}\left[\sum_{i=0}^{N-1-j}\pi_{ij} + \pi_{(N-j)j}\right] + p\sum_{j=0}^{N-1}\left[\sum_{i=0}^{N-1-j}\pi_{ij} + \pi_{(N-j)j}\right] \\
&= \pi_{0N} + \sum_{j=0}^{N-1}\pi_{(N-j)j} + (1-p)\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij} + p\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij} \\
&= P(blocking) + (1-p)\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij} + p\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}, \qquad (3.11)
\end{aligned}$$

where π_{*j} is the stationary marginal distribution of Node 2. Hence, we have:

1. $P(\text{only self-served by Node 1}) = (1-p)\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
2. $P(\text{no-delay}, entry) = p\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij}$, where the event *entry* means non-blocking and joining Node 2.
3. $P(delay, entry) = p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
4. $P(entry) = p\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.

Now Little's formula for busy CSRs at Node 2 is

$$E(Q_{2b}) = \lambda P(entry)\frac{1}{\mu} = ap\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij},$$

which is easy to verify since

$$\begin{aligned}
ap\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij} &= a_2\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + a_2\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij} \\
&= \sum_{j=0}^{S-1}(j+1)\sum_{i=0}^{N-1-j}\pi_{i(j+1)} + S\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{i(j+1)} \\
&= \sum_{j=1}^{S} j\sum_{i=0}^{N-j}\pi_{ij} + S\sum_{j=S+1}^{N}\sum_{i=0}^{N-j}\pi_{ij} \\
&= E(Q_{2b}),
\end{aligned}$$

where we have used the fact that $\frac{\lambda p}{(j+1)\mu}\pi_{ij} = \pi_{i(j+1)}$ for $j < S$ and $\frac{\lambda p}{S\mu}\pi_{ij} = \pi_{i(j+1)}$ for $j \geq S$. These equations can be verified by the product form solution π_{ij} (3.1). Hence the

carried load for Node 2 is $a' = E(Q_{2b}) = aP(entry)$. The utilization

$$v = \frac{a'}{S} = \rho P(entry) < 1$$

is the proportion of time that a CSR is busy.

3.4 Waiting time distribution and mean waiting time

Another important performance measure is $TSF = P(W_q < AWT)$, which is related to the waiting time distribution. As $M/M/S/N$ model in Chapter 2, we only need to consider those calls given that they are not blocked and join Node 2 (or given entry), since there is no queue at Node 1. Let W_q denote the conditional stationary waiting time of calls given entry, which is the time spent by an entry call in the queue of Node 2 until starting to get service. This definition comes form [42] and they derived the result using $\chi(k, j)$ (refer to Section 3.4.2). We will derive the waiting time distribution using q_j and the Arrival Theorem. In the end we will show that the results obtained using these two methods are equivalent.

3.4.1 Using q_j

In order to find the distribution of W_q, we need to find the probability of finding j calls at Node 2 by an arriving call from Node 1 at arrival instant. As in $M/M/S/N$ model, we use q_j to denote this probability. According to the Arrival Theorem of CQN [7] mentioned in Chapter 2, for $0 \leq j \leq N-1$, we have

$$\begin{aligned} q_j &= P(\text{the call finds } j \text{ calls at Node 2} \mid \text{Non-blocked and join Node 2}) \\ &= P(j \text{ calls at Node 2 for the same network with one less call in equilibrium}) \\ &= \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)}, \end{aligned} \tag{3.12}$$

where $\pi_{ij}^{(N-1)}$ is the stationary distribution of the queue length process of the same network with one less call and has been obtained before.

Once we have the expression of q_j, the following performance measures can be easily obtained similar to the $M/M/S/N$ model.

1. $P(\text{no-delay}|entry) = P(W_q = 0) = \sum_{j=0}^{S-1} q_j = \sum_{j=0}^{S-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)}$.

2. $P(delay|entry) = P(W_q > 0) = \sum_{j=S}^{N-1} q_j = \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)}$.

3. $TSF = P(W_q \le t) = 1 - \sum_{j=S}^{N-1} q_j \sum_{k=0}^{j-S} \frac{(S\mu t)^k e^{-S\mu t}}{k!} = 1 - \sum_{j=S}^{N-1} q_j \overline{F}_{Y_j}(t)$, $t \ge 0$, where $Y_j \sim Er(j - S + 1, S\mu)$.

4. $ASA = E(W_q) = \int_0^\infty P(W_q > t)dt = \int_0^\infty \sum_{j=S}^{N-1} q_j \overline{F}_{Y_j}(t)dt = \sum_{j=S}^{N-1} q_j \int_0^\infty \overline{F}_{Y_j}(t)dt = \frac{1}{S\mu} \sum_{j=S}^{N-1} q_j(j - S + 1)$.

q_j is actually the probability of a call finding j calls at Node 2 given entry. Using (3.11) and conditional argument, it is easy to see that

$$q_j = \frac{p \sum_{i=0}^{N-1-j} \pi_{ij}}{1 - P(blocking) - (1-p) \sum_{j=0}^{N-1} \pi_{*j}} = \frac{\sum_{i=0}^{N-1-j} \pi_{ij}}{\sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}}, 0 \le j < N. \tag{3.13}$$

Comparing (3.12) and (3.13), we have the relationship: $\pi_{ij}^{(N-1)} = \frac{\pi_{ij}}{\sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}}$, which is clearly true since $\pi_{ij}^{(N-1)}$ is π_{ij} under the condition that $0 \le i + j \le N - 1$.

Now Little's formula for calls waiting in the queue at Node 2 is

$$\begin{aligned} E(Q_{2q}) &= \lambda P(entry) E(W_q) \\ &= \lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} E(W_q), \end{aligned}$$

which is easy to verify since

$$\begin{aligned} \lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} E(W_q) &= \lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} \frac{1}{S\mu} \sum_{k=S}^{N-1} q_k(k - S + 1) \\ &= \frac{\lambda p}{S\mu} \sum_{k=S}^{N-1} \sum_{i=0}^{N-1-k} \pi_{ik}(k - S + 1) \\ &= \sum_{k=S}^{N-1} \sum_{i=0}^{N-1-k} \pi_{i(k+1)}(k - S + 1) \\ &= \sum_{k=S+1}^{N} (k - S) \sum_{i=0}^{N-k} \pi_{ik} = E(Q_{2q}), \end{aligned}$$

where we have used the fact that $\frac{\lambda p}{S\mu} \pi_{ik} = \pi_{i(k+1)}$ for $k \ge S$ and (3.13).

3.4.2 Using $\chi(k, j)$

In Srinivasan et al. [42], the same results have been obtained in terms of $\chi(k, j)$, which is defined as: For $0 \le j < k \le N, \chi(k, j)$ is the probability that the system is in state

$(k-j,j)$, when a call (among the $k-j$ calls) is about to leave Node 1 and join Node 2. Using Bayes' theorem,

$$\begin{aligned}
&\chi(k,j) := P(\text{system in state } (k-j,j) \mid \text{ call is about to leave Node 1 and join Node 2}) \\
&= \frac{P\left(\text{call is about to leave Node 1 and join 2} \mid \text{system in state } (k-j,j)\right)\pi_{(k-j)j}}{\sum_{l=0}^{N}\sum_{m=0}^{l} P\left(\text{call is about to leave Node 1 and join 2} \mid \text{system in state } (l-m,m)\right)\pi_{(l-m)m}} \\
&= \frac{(k-j)\theta\pi_{(k-j)j}}{\sum_{l=0}^{N}\sum_{m=0}^{l}(l-m)\theta\pi_{(l-m)m}} \\
&= \frac{(k-j)\pi_{(k-j)j}}{\sum_{l=1}^{N}\sum_{m=0}^{l-1}(l-m)\pi_{(l-m)m}}.
\end{aligned}$$

Then we have:

1. $P(\text{no-delay}|entry) = P(W_q = 0) = \sum_{k=1}^{N}\sum_{j=0}^{k\wedge S-1}\chi(k,j) = \sum_{j=0}^{S-1}\sum_{k=j+1}^{N}\chi(k,j)$.
2. $P(delay|entry) = P(W_q > 0) = \sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)$.
3. $TSF = P(W_q \le t) = 1-\sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)\sum_{l=0}^{j-S}\frac{(S\mu t)^l e^{-S\mu t}}{l!} = 1-\sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)\overline{F}_{Y_j}(t)$, $t \ge 0$, where $Y_j \sim Er(j-S+1, S\mu)$.
4. $ASA = E(W_q) = \int_0^\infty P(W_q > t)dt = \sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)\sum_{l=0}^{j-S}\int_0^\infty \frac{(S\mu t)^l e^{-S\mu t}}{l!}dt = \frac{1}{S\mu}\sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)(j-S+1)$.

We have the following new result to relate q_j with $\chi(k,j)$.

Theorem 3.4.1 *For $0 \le j < N$, we have $q_j = \sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)} = \sum_{k=j+1}^{N}\chi(k,j)$.*

Proof. For $0 \le j < N$, let $i := k-j > 0$, then $k = i+j$. We have

$$\chi(k,j) = \chi(i+j,j) = \frac{i\pi_{ij}}{\sum\limits_{0\le l+m\le N} l\pi_{lm}} = \frac{i\pi_{ij}}{\sum\limits_{1\le l+m\le N} l\pi_{lm}}.$$

Hence

$$\begin{aligned}
\sum_{k=j+1}^{N}\chi(k,j) &= \sum_{i=1}^{N-j}\frac{i\pi_{ij}}{\sum\limits_{1\le l+m\le N} l\pi_{lm}} = \sum_{i=1}^{N-j}\frac{i\pi_{00}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}}{\sum\limits_{1\le l+m\le N} l\pi_{00}\frac{a_1^l}{l!}\frac{a_2^m}{\beta(m)}} \\
&= \sum_{i=1}^{N-j}\frac{i\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}}{\sum\limits_{1\le l+m\le N} l\frac{a_1^l}{l!}\frac{a_2^m}{\beta(m)}} = \sum_{i=1}^{N-j}\frac{a_1\frac{a_1^{i-1}}{(i-1)!}\frac{a_2^j}{\beta(j)}}{a_1\sum\limits_{1\le l+m\le N}\frac{a_1^{l-1}}{(l-1)!}\frac{a_2^m}{\beta(m)}} \\
&= \sum_{i-1=0}^{N-1-j}\frac{\frac{a_1^{i-1}}{(i-1)!}\frac{a_2^j}{\beta(j)}}{\sum\limits_{0\le l-1+m\le N-1}\frac{a_1^{l-1}}{(l-1)!}\frac{a_2^m}{\beta(m)}}.
\end{aligned}$$

Now let $n := i-1$ and $p := l-1$, we have

$$\sum_{k=j+1}^{N} \chi(k,j) = \sum_{n=0}^{N-1-j} \frac{\frac{a_1^n}{n!}\frac{a_2^j}{\beta(j)}}{\sum_{0\le p+m\le N-1} \frac{a_1^p}{p!}\frac{a_2^m}{\beta(m)}} = \sum_{n=0}^{N-1-j} \frac{\frac{a_1^n}{n!}\frac{a_2^j}{\beta(j)}}{\pi_{00}^{(N-1)}} = \sum_{n=0}^{N-1-j} \pi(n,j)^{(N-1)} = q_j$$

where $\pi_{00}^{(N-1)}$ is the normalizing constant for the same network with one less call. ∎

Using this result, we can prove the equivalence of the two sets of the performance measures in terms of q_j and $\chi(k,j)$ respectively. For example

$$\begin{aligned} P(delay|entry) = P(W_q > 0) &= \sum_{j=S}^{N-1} q_j \\ &= \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} = \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \\ &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j). \end{aligned}$$

Other performance measures can be verified similarly.

3.5 Numerical examples

To give some numerical illustrations for the SOQN model, we will consider the following example for $P(blocking)$ and $P(W_q > t)$. This example is given in Srinivasan et al. [42], which deals with a call load of 250 calls per half an hour period. The average talk time is estimated to be 180 seconds and the average IVRU processing time is 0.01 seconds or 100 seconds representing fast and slow IVRU servers respectively. Therefore, our parameters are $\lambda = 250/1800, \mu = 1/180, \theta = 100$ or 0.01. We let $t = 20$ seconds. To illustrate the effect of buffer size, we fix $S = 5$ and let buffer size $K = N - S$ change from 0 to 30. We also consider two cases: $p = 1$ and $p = 0.1$.

For $p = 1$, the results are shown in Figure 3.3 for Example 3.1 ($\theta = 100$) and in Figure 3.4 for Example 3.2 ($\theta = 0.01$) respectively. Since $p = 1$, all calls need the service of CSRs, which makes $P(blocking)$ and $P(W_q > 20)$ pretty high in both examples. In Example 3.1, $\theta = 100$, which means we have fast IVRU servers (processing time is 0.01 seconds). In this case, the model is close to $M/M/S/N$ model as mentioned in [42]. In Example 3.2, $\theta = 0.01$, which means we have slow IVRU servers (processing time is 100 seconds). In

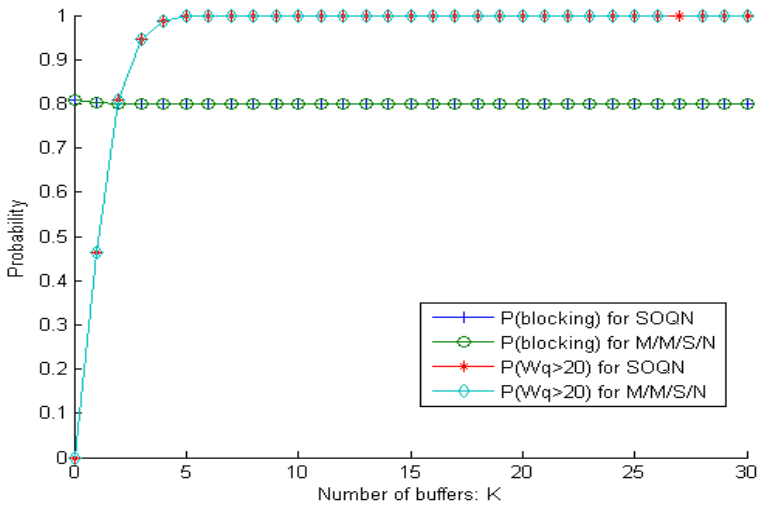

Figure 3.3: $P(blocking)$ and $P(W_q > 20)$ for Example 3.1

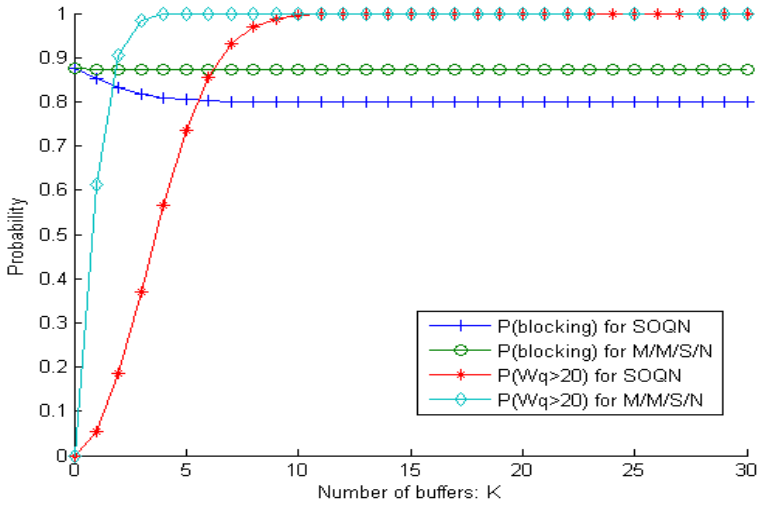

Figure 3.4: $P(blocking)$ and $P(W_q > 20)$ for Example 3.2

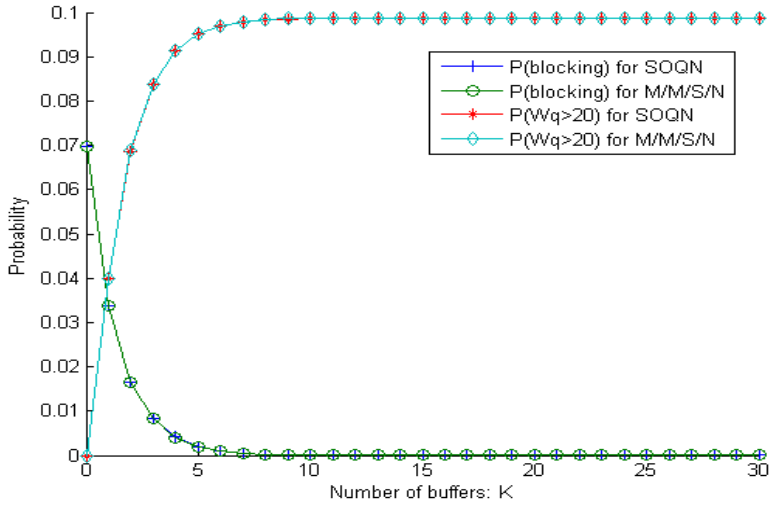

Figure 3.5: $P(blocking)$ and $P(W_q > 20)$ for Example 3.3

this case, for $P(blocking)$, the model can still be approximated by $M/M/S/N$ model as in [42] using new service rate $\widehat{\mu}$, where $\frac{1}{\widehat{\mu}} = \frac{1}{\theta} + \frac{1}{\mu}$. Now since the service time is longer than that in Example 3.1, we have $P(blocking)$ is higher compared to Example 3.1. The above discussion can be verified numerically using the result of $M/M/S/N$ model in Chapter 2 and the approximation is also shown in Figure 3.3 and Figure 3.4. However we find that the approximation in Example 3.2 is not as good as in Example 3.1. The limiting case when K is large can be verified as well. For instance, in Example 3.1, we have $\rho = \frac{\lambda}{S\mu} = 5 > 1$. Therefore we have $\lim_{N\to\infty} P(blocking) = 1 - \frac{1}{\rho} = 0.8$ and $\lim_{N\to\infty} P(W_q > t) = 1$ according to Theorem 2.4.2 and 2.4.4.

For $p = 0.1$, the results are shown in Figure 3.5 for Example 3.3 ($\theta = 100$) and in Figure 3.6 for Example 3.4 ($\theta = 0.01$) respectively. Now since $p = 0.1$, only 10% of calls need the service of CSRs, which makes $P(blocking)$ and $P(W_q > 20)$ pretty low when compared to the case $p = 1$. Similarly as discussed in the case $p = 1$, both examples can be approximated by $M/M/S/N$ model using new arrival rate λp. Again we find that the approximation in Example 3.4 is not as good as in Example 3.3 as shown in Figure 3.5 and Figure 3.6. In Example 3.3, we have $\rho = \frac{\lambda p}{S\mu} = 0.5 < 1$. Therefore we have $\lim_{N\to\infty} P(blocking) = 0$ and $\lim_{N\to\infty} P(W_q > t) = C(S, a)e^{-(S\mu-\lambda)t} = 0.0988$ according to Theorem 2.4.2 and 2.4.4.

We have observed the similar monotonicity properties of $P(blocking)$ and $P(W_q > t)$ as in $M/M/S/N$ model in all the above examples although we are not able to prove them.

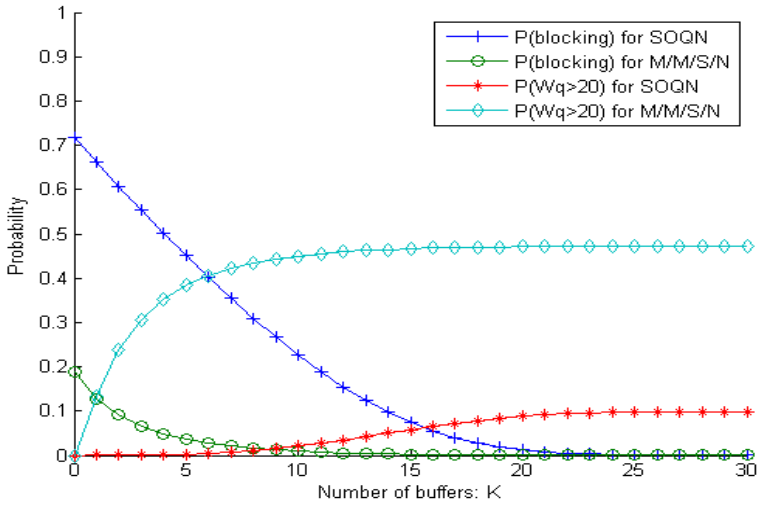

Figure 3.6: $P(blocking)$ and $P(W_q > 20)$ for Example 3.4

Therefore we have the following conjecture. When S and other parameters are fixed, $P(blocking)$ is a strictly decreasing function of K and $P(W_q{>}t)$ is a strictly increasing function of K. This conjecture is intuitively correct and we will use it in Chapter 7 for the call centre design problem.

3.6 Summary

In this chapter, we studied the SOQN model proposed in [42]. We derived the product form solution of the queue length process and the distribution of the total number of calls in the system using two methods. We also proposed an algorithm to compute the blocking probability. In the derivation of the waiting time distribution, we used a new method compared to the one used in the original paper. Finally we provided numerical examples to illustrate the effect of buffer size to performance measures such as $P(blocking)$ and $P(W_q > t)$.

Chapter 4

Balking Models of Call Centres

In Figure 1.2, we find that if an arriving call finds all CSRs are busy and even there are free waiting spaces the call may choose not to enter the system and leave the system upon arrival. This is referred to as balking. Typically a monotonically decreasing function b_i of system size i (called balking function) is employed to include balking to queueing models [24], page 123-124. b_i can be explained as the probability of entering the system if there are i calls in the system upon arrival of a call and the balking probability is $1 - b_i$. Hence, we assume that for a queueing model with S servers,

$$b_i = 1, 0 \leq i < S \text{ and } 0 < b_{i+1} \leq b_i < 1, i \geq S.$$

Now our arriving rate becomes $\lambda_i = \lambda b_i$. Obviously, balking model is equivalent to the state-dependent arrival model with arrival rates λ_i.

Gross and Harris [24], page 94 provided possible balking functions for single server queueing models: $\frac{1}{i+1}, \frac{1}{i^2+1}$, $e^{-\alpha i}, \alpha > 0$, and $e^{-\alpha i/\mu}, \alpha > 0$. Note that the function $e^{-\alpha i/\mu}, \alpha > 0$, depending not only on i but also on the service rate μ, is a more realistic balking function.

In this chapter we will study the balking phenomenon of call centres using both single-node Markovian models and the SOQN model. We first give a review of single-node state-dependent balking model $M(n)/M/S/N$, which is standard and appeared in many textbook. The main work is on the SOQN model with balking, where we prove that the product form solution of the queue length process still holds and we also derive the waiting time distribution.

4.1 Single-node state-dependent balking model $M(n)/M/S/N$

In this section, we will give a review of $M(n)/M/S/N$ model, i.e., the $M/M/S/N$ model with state-dependent balking. Now since the model has finite buffers, the non-balking

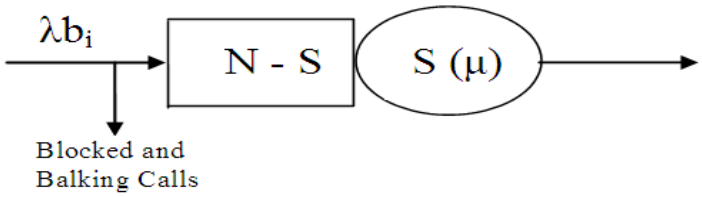

Figure 4.1: $M(n)/M/S/N$ model description and parameters

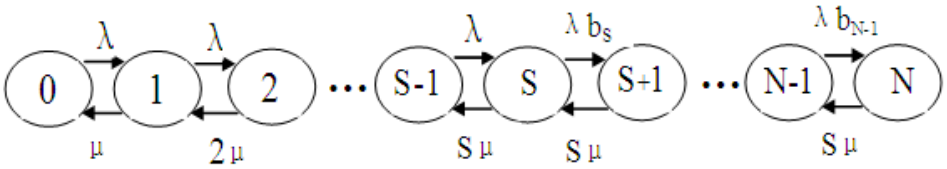

Figure 4.2: $M(n)/M/S/N$ model stationary state transition diagram

probability b_i given i in the system satisfies

$$b_i = 1, 0 \leq i < S \text{ and } 0 < b_{i+1} \leq b_i < 1, S \leq i < N-1. \tag{4.1}$$

All other model description and assumptions are the same as $M/M/S/N$ model. The model description and parameters are shown in Figure 4.1.

4.1.1 Queue length process

In this model, $Q(t)$ is a finite birth-death process with birth rate $\lambda_i = \lambda b_i$ and state-dependent death rate

$$\mu_i = \begin{cases} i\mu & \text{if } 0 \leq i \leq S \\ S\mu & \text{if } S \leq i \leq N \end{cases}.$$

The stationary state transition diagram of $Q(t)$ is shown in Figure 4.2. Since this is a finite buffer model, the stationary distribution of $Q(t)$ can be obtained by solving the global or cut balance equations derived from Figure 4.2 (no stability condition). The solution is

$$p_i = p_0 \gamma(i) \prod_{j=0}^{i-1} b_j = \begin{cases} \frac{a^i}{i!} p_0 & \text{if } 0 \leq i \leq S \\ \frac{a^i}{S! S^{i-S}} \prod_{j=S}^{i-1} b_j p_0 & \text{if } S \leq i \leq N \end{cases}$$

where p_0 is

$$p_0 = \left[\sum_{i=0}^{S} \frac{a^i}{i!} + \frac{a^S}{S!} \sum_{i=S+1}^{N} \rho^{i-S} \prod_{j=S}^{i-1} b_j \right]^{-1} = \left[\sum_{i=0}^{N} \gamma(i) \prod_{j=0}^{i-1} b_j \right]^{-1}$$

and

$$\gamma(i) = \begin{cases} \frac{a^i}{i!} & \text{if } 0 \le i \le S \\ \frac{a^i}{S! S^{i-S}} & \text{if } S \le i \le N \end{cases}$$

as defined in Chapter 2.

The previous results are standard and can be found in many queueing textbooks. However the following idea is new. We can obtain this distribution by using the method of CQN as we have done for $M/M/S/N$ model in Chapter 2 by introducing a fictitious node (Node 0) which has one server with service rate

$$\mu_0(j) = \lambda b_{N-j} \quad \text{for } 0 \le j \le N.$$

It can be proved that the solution of this CQN has product form (see the proof of the similar SOQN case in Section 4.3). Using the same method as used in Chapter 3, we have the stationary distribution of having i calls at Node 1 and j calls at Node 0 is:

For all $(i, j) \in \mathcal{Z}_+^2$ such that $i + j = N$,

$$\pi_{ij} = \pi_{00} \frac{\gamma(i)}{\lambda^i} \frac{1}{\lambda^j \prod_{m=1}^{j} b_{N-m}}.$$

Therefore the stationary distribution of having i calls at Node 1 is

$$\pi_i = \frac{\pi_{00}}{\lambda^N} \gamma(i) \frac{1}{\prod_{m=1}^{N-i} b_{N-m}}$$

since $i + j = N$. The above has the same distribution as p_i since

$$\begin{aligned} \frac{p_i}{\pi_i} &= \frac{p_0 \gamma(i) \prod_{j=0}^{i-1} b_j}{\frac{\pi_{00}}{\lambda^N} \gamma(i) \frac{1}{\prod_{m=1}^{N-i} b_{N-m}}} = \frac{\lambda^N p_0}{\pi_{00}} \prod_{j=0}^{i-1} b_j \prod_{m=1}^{N-i} b_{N-m} \\ &= \frac{\lambda^N p_0}{\pi_{00}} \prod_{j=0}^{N-1} b_j \end{aligned}$$

is a constant, not involving i.

Note that the arrival process is still Poisson process, although the arriving calls enter the system and affect the system state with probability b_i. Therefore from the PASTA property, we have for $0 \le i \le N$, $a_i = p_i$ and the following results.

1. $P(blocking) = p_N$.

2. $P(\text{no-delay}) = \sum_{i=0}^{S-1} a_i = \sum_{i=0}^{S-1} p_i$.

3. $P(delay, entry) = \sum_{i=S}^{N-1} b_i a_i = \sum_{i=S}^{N-1} b_i p_i$.

4. $P(entry) = \sum_{i=0}^{S-1} a_i + \sum_{i=S}^{N-1} b_i a_i = \sum_{i=0}^{N-1} b_i p_i$.

5. $P(balking) = \sum_{i=S}^{N-1} \overline{b}_i a_i = \sum_{i=S}^{N-1} \overline{b}_i p_i$.

4.1.2 Waiting time distribution

Similar as in $M/M/S/N$ model, if we define $\overline{W}_q$ as the stationary waiting time in the queue for all calls (the blocked and balking calls have ∞ waiting time), then $\overline{W}_q$ has a mass at ∞ and $P(\overline{W}_q = \infty) = P(blocking\ and\ balking) = 1 - P(entry)$. We have

$$\begin{aligned} P(\overline{W}_q > t) &= P(\overline{W}_q > t, entry) + P(\overline{W}_q > t, blocking\ and\ balking) \\ &= P(\overline{W}_q > t, entry) + P(blocking\ and\ balking). \end{aligned}$$

To find $P(\overline{W}_q > t, entry)$, we use the same idea as $M/M/S$ model

$$\begin{aligned} &P(\overline{W}_q > t, entry) \\ &\quad = \sum_{i=S}^{N} P(\overline{W}_q > t, entry \mid i \text{ calls in the system upon entry}) P(i \text{ calls in the system upon entry}) \\ &\quad = \sum_{i=S}^{N-1} P(\text{completion time of } i - S + 1 \ calls > t) a_i b_i \\ &\quad = \sum_{i=S}^{N-1} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} p_i b_i. \end{aligned}$$

Usually we are more concerned with the conditional waiting time of a call given entry. Let W_q be this waiting time and we have

$$\begin{aligned} P(W_q > t) = P(\overline{W}_q > t | entry) &= \frac{P(\overline{W}_q > t, entry)}{P(entry)} \\ &= \sum_{i=S}^{N-1} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} \frac{p_i b_i}{\sum_{i=0}^{N-1} b_i p_i} = \sum_{i=S}^{N-1} \sum_{j=0}^{i-S} \frac{(S\mu t)^j e^{-S\mu t}}{j!} q_i. \end{aligned}$$

Here $q_i := \frac{p_i b_i}{\sum_{i=0}^{N-1} b_i p_i}, 0 \le i \le N-1$ is the probability of a call finding i calls in the system given entry, which can be derived alternatively in the following way [17].

Let $E(t, t+h)$ be the event that a call arrives in $(t, t+h)$ and enters the system. Using Bayes' theorem it follows that for $0 \leq i \leq N-1$,

$$\begin{aligned} q_i &= \lim_{t\to\infty}\lim_{h\to 0} P(Q(t)=i|E(t,t+h)) \\ &= \lim_{t\to\infty}\lim_{h\to 0} \frac{P(Q(t)=i, E(t,t+h))}{P(E(t,t+h))} \\ &= \lim_{t\to\infty}\lim_{h\to 0} \frac{P(E(t,t+h)|Q(t)=i)P(Q(t)=i)}{\sum_{i=0}^{N-1} P(E(t,t+h)|Q(t)=i)P(Q(t)=i)} \\ &= \lim_{h\to 0} \frac{(\lambda b_i h + o(h))p_i}{\sum_{i=0}^{N-1}(\lambda b_i h + o(h))p_i} \\ &= \frac{\lambda b_i p_i}{\sum_{i=0}^{N-1}\lambda b_i p_i} = \frac{b_i p_i}{\sum_{i=0}^{N-1} b_i p_i}. \end{aligned}$$

Once we have the expression of q_i, the following performance measures can be easily obtained.

1. $P(\text{no-delay}|entry) = P(W_q = 0) = \sum_{i=0}^{S-1} q_i$.
2. $P(delay|entry) = P(W_q > 0) = \sum_{i=S}^{N-1} q_i$.
3. $ASA = E(W_q) = \int_0^\infty P(W_q > t)dt = \int_0^\infty \sum_{i=S}^{N-1} q_i \overline{F}_{Y_i}(t)dt = \frac{1}{S\mu}\sum_{i=S}^{N-1} q_i(i-S+1), where\ Y_i \sim Er(i-S+1, S\mu)$.

Note that the results in this section are also standard and appeared in many textbooks such as Riordan [39] page 113.

4.1.3 Special cases of b_i

In the literature, there are some studies of this model with special forms of balking function b_i. In an earlier paper [3], the authors studied a $M/M/1$ model with balking and obtained a series of stationary performance measures. The balking function is

$$b_i = \begin{cases} 1 - i/N & \text{if } 0 \leq i \leq N \\ 0 & otherwise \end{cases}.$$

Here N is a measure of call's willingness to enter the system and it is obvious that there are at most N calls in the system. Essentially the authors used the same method as above and the difference is they have an explicit expression for b_i. After some algebraic manipulation they managed to express the performance measures by special functions such as Gamma and Beta functions, which facilitate the computation. This model can also be seen as a

machine interference problem with N machines and the individual failure rate $\sigma = \lambda/N$, since $b_i = 1 - i/N = (N-i)/N$ and $\lambda b_i = (N-i)\sigma$ is the failure rate when i is in repair. In a later paper [4], the authors studied a similar model with a different balking function, for $0 \leq \beta \leq 1$,

$$b_i = \begin{cases} \beta/i & \text{if } i \geq 1 \\ 1 & \text{if } i = 0 \end{cases}$$

where β is a measure of call's willingness to enter the system. Now there is no limit on system capacity ($N \to \infty$). They obtained similar results as [3].

For multiserver case, Reynolds [38] used the balking function

$$b_i = \begin{cases} \frac{1}{i-S+2} & \text{if } i \geq S \\ 1 & \text{if } i < S \end{cases}$$

to analyze an infinite capacity queue. When $S = 1$, the stationary queue length Q has a Poisson distribution with parameter a, just as a $M/M/\infty$ queue [28], page 100. For multiserver and finite capacity case, Abou-El-Ata et al. [1] analyze a $M/M/S/N$ queue with a general balking function

$$b_i = \begin{cases} \frac{\beta(1-(i-S+1)/N)}{(i-S+2)^m} & \text{if } S \leq i < N \\ 1 & \text{if } 0 \leq i < S \end{cases}$$

where β is a measure of a call's willingness to join the queue and m is a non-negative integer.

4.2 Two-node network model with state-dependent balking and state-dependent service

In this section, we will look into two-node network model with state-dependent balking coefficient b_j and state-dependent service rate. Here b_j satisfies conditions (4.1). We will show that this model has a product form solution for the queue length process.

4.2.1 Model description

The model is a tandem queueing network with two nodes in series. Node i has one server with state-dependent exponential service rate $\mu_i(n)$, $i = 1, 2$, where n is the number of calls at Node i. Arriving calls to Node 1 are Poisson with arrival rate λ. After the service

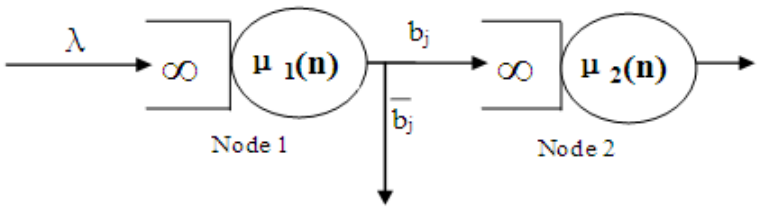

Figure 4.3: Model description and parameters for the two-node network model

with Node 1 is completed, the call leaves the network with probability $\bar{b}_j$ and it joins Node 2 with probability $b_j = 1 - \bar{b}_j$, where j is the number of calls at Node 2 when the call leaves Node 1. After the service with Node 2, it leaves the network. Figure 4.3 gives a picture of the model.

4.2.2 Product form solution of the queue length process

Let $Q(t) = (Q_1(t), Q_2(t))$ be the queue length process of our tandem queueing network, where $Q_i(t)$ is the queue length of Node $i, i = 1, 2$ at time t. From the description of the model, we know that $Q(t)$ is a two dimensional CTMC. We assume that the stationary distribution of $Q(t)$ exists and let $\pi_{ij} = P(Q_1 = i, Q_2 = j)$ be the stationary probability of having i calls at Node 1 and j calls at Node 2. In order to find π_{ij}, we first draw the stationary state transition diagram in Figure 4.4.

From the stationary state transition diagram, we can get the global balance equations

$$
\begin{aligned}
&\pi_{ij}(\lambda + \mu_2(j) + \mu_1(i)) \\
&= \pi_{(i-1)j}\lambda + \pi_{i(j+1)}\mu_2(j+1) + \pi_{(i+1)(j-1)}\mu_1(i+1)b_{j-1} + \pi_{(i+1)j}\mu_1(i+1)\bar{b}_j,\ i \geq 1, j \geq 1 \\
&\pi_{0j}(\lambda + \mu_2(j)) = \pi_{0(j+1)}\mu_2(j+1) + \pi_{1(j-1)}\mu_1(1)b_{j-1} + \pi_{1j}\mu_1(1)\bar{b}_j,\ i = 0, j \geq 1 \\
&\pi_{i0}(\lambda + \mu_1(i)) = \pi_{(i-1)0}\lambda + \pi_{i1}\mu_2(1) + \pi_{(i+1)0}\mu_1(i+1)\bar{b}_0,\ i \geq 1, j = 0 \\
&\pi_{00}\lambda = \pi_{01}\mu_2(1) + \pi_{10}\mu_1(1)\bar{b}_0,\ i = 0, j = 0
\end{aligned}
\tag{4.2}
$$

where the last three equations are boundary cases.

It is typically hard to solve these global balance equations directly. Instead we will solve a set of more strict balance equations: local balance equations. They are also referred to as partial or station balance equations, which are described in [10] as

"At each state the flow out of a station due to the departure of customers is

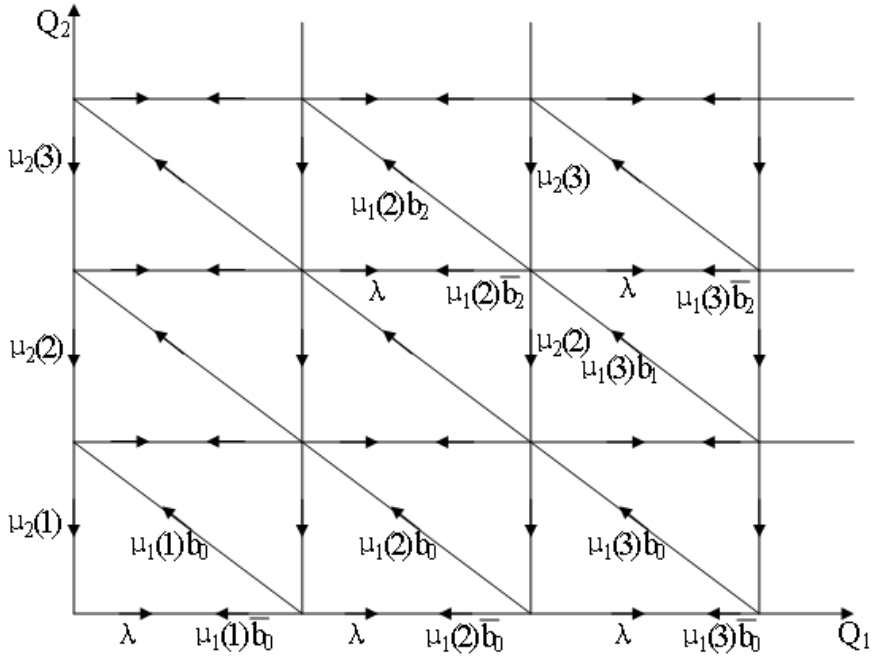

Figure 4.4: Stationary state transition diagram for the two-node network model

balanced by the flow into that same station due to the arrival of customers."

For Node 1 the station balance equations are

$$\pi_{ij}\mu_1(i) = \pi_{(i-1)j}\lambda,\ i \geq 1, j \geq 0. \tag{4.3}$$

For Node 2 the station balance equations are

$$\pi_{ij}\mu_2(j) = \pi_{(i+1)(j-1)}\mu_1(i+1)b_{j-1},\ i \geq 0, j \geq 1. \tag{4.4}$$

For the whole network, the station balance equations are

$$\pi_{ij}\lambda = \pi_{i(j+1)}\mu_2(j+1) + \pi_{(i+1)j}\mu_1(i+1)\bar{b}_j,\ i \geq 0, j \geq 0. \tag{4.5}$$

It is easy to see that if π_{ij} satisfies these station balance equations, it also satisfies the global balance equations (4.2) since the sum of three station balance equations will be the global balance equations. In other words the global balance equations have been decomposed into the sum of three station balance equations, which are easier to deal with. We can think of the station balance equations as the generalization of cut equations of birth-death process to two dimensions.

Theorem 4.2.1 *For the above tandem queueing network model, the stationary queue length distribution π_{ij} has a product form*

$$\pi_{ij} = \prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00}, \quad i \geq 0, j \geq 0$$

where $a_1(n) = \lambda/\mu_1(n)$; $a_2(n) = \lambda/\mu_2(n)$;

$$\pi_{00} = \left[\sum_{i=0}^{\infty} \prod_{n=1}^{i} a_1(n) \sum_{j=0}^{\infty} \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1} \right]^{-1}$$

and the empty product is 1 by convention. The stability conditions are

$$\sum_{i=0}^{\infty} \prod_{n=1}^{i} a_1(n) < \infty \text{ and } \sum_{j=0}^{\infty} \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1} < \infty.$$

Proof. We need to solve station balance equations (4.3) (4.4) (4.5). From (4.4) we have

$$\pi_{(i+1)j} = \frac{\mu_2(j+1)}{\mu_1(i+1)b_j}\pi_{i(j+1)}, \ i \geq 0, j \geq 0.$$

Substituting the above to (4.5), we get

$$\begin{aligned}
\pi_{ij}\lambda &= \pi_{i(j+1)}\mu_2(j+1) + \pi_{(i+1)j}\mu_1(i+1)\bar{b}_j \\
&= \pi_{i(j+1)}\mu_2(j+1) + \frac{\mu_2(j+1)}{\mu_1(i+1)b_j}\pi_{i(j+1)}\mu_1(i+1)\bar{b}_j \\
&= (\mu_2(j+1) + \frac{\mu_2(j+1)\bar{b}_j}{b_j})\pi_{i(j+1)} \\
&= \frac{\mu_2(j+1)}{b_j}\pi_{i(j+1)}, \ i \geq 0, j \geq 0.
\end{aligned}$$

Then we have

$$\pi_{i(j+1)} = a_2(j+1)b_j\pi_{ij} = a_2(j+1)a_2(j)b_jb_{j-1}\pi_{i(j-1)} = ... = \prod_{n=1}^{j+1} a_2(n) \prod_{n=1}^{j+1} b_{n-1}\pi_{i0}, i \geq 0, j \geq 0,$$

i.e.,

$$\pi_{ij} = \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{i0}, i \geq 0, j \geq 0,$$

where the empty product is 1 by convention. From this we have

$$\pi_{0j} = \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00}, \ j \geq 0. \tag{4.6}$$

On the other hand, from (4.3) we have

$$\pi_{ij} = a_1(i)\pi_{(i-1)j} = ... = \prod_{n=1}^{i} a_1(n)\pi_{0j}, \ i \geq 0, j \geq 0. \tag{4.7}$$

Combining the above two formulas (4.6) and (4.7), we have

$$\pi_{ij} = \prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00},\ i \geq 0, j \geq 0.$$

Since $\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\pi_{ij} = 1$, we have

$$\begin{aligned} 1 &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00} \\ &= \sum_{i=0}^{\infty}\prod_{n=1}^{i} a_1(n) \sum_{j=0}^{\infty}\prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00}. \end{aligned}$$

Therefore

$$\pi_{00} = \left[\sum_{i=0}^{\infty}\prod_{n=1}^{i} a_1(n) \sum_{j=0}^{\infty}\prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\right]^{-1}$$

and the stability conditions are:

$$\sum_{i=0}^{\infty}\prod_{n=1}^{i} a_1(n) < \infty \text{ and } \sum_{j=0}^{\infty}\prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1} < \infty.$$

It can be verified that this solution satisfies the boundary cases in (4.2) as well. ∎

Remark 4.2.1 *The solution has a product form: Node 1 is a $M/M(n)/1$ queue with arrival rate λ and state-dependent service rate $\mu_1(n)$. Node 2 is a $M(n)/M(n)/1$ queue with state-dependent arrival rate λb_j and service rate $\mu_2(n)$. This is intuitively true since Node 1 has the Poisson in and Poisson out property in equilibrium.*

Remark 4.2.2 *If the service rates at two nodes are constant, then we have the special case: $\pi_{ij} = (1-a_1)a_1^i a_2^j \prod_{n=1}^{j} b_{n-1}\pi_0, i \geq 0, j \geq 0$ where $a_1 = \lambda/\mu_1$; $a_2 = \lambda/\mu_2$; $\pi_0 = (1 + \sum_{j=1}^{\infty} a_2^j \prod_{n=1}^{j} b_{n-1})^{-1}$ is the probability of 0 customers at Node 2, when Node 2 is thought of as an isolated queue with state-dependent arrival rate λb_j and service rate μ_2. The stability conditions are: $a_1 < 1$ and $\sum_{j=1}^{\infty} a_2^j \prod_{n=1}^{j} b_{n-1} < \infty$. In this case Node 1 is a $M/M/1$ queue and Node 2 is a $M(n)/M/1$ queue.*

Remark 4.2.3 *If $\mu_1(n) = n\mu_1$ and $\mu_2(n) = (S \wedge n)\mu_2$, then we have Node 1 is a $M/M/\infty$ queue with arrival rate λ and service rate μ_1. Node 2 is a $M(n)/M/S$ queue with state-dependent arrival rate λb_j and service rate μ_2. The product form solution is*

$$\pi_{ij} = \ e^{-a_1}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)} \prod_{n=1}^{j} b_{n-1}\pi_0, \quad i \geq 0, j \geq 0$$

where $a_1 = \lambda/\mu_1$; $a_2 = \lambda/\mu_2$;

$$\beta(j) = \begin{cases} j! & for\ 0 \le j \le S \\ S!S^{j-S} & for\ S \le j \le N \end{cases}$$

and

$$\pi_0 = \left[\sum_{j=0}^{\infty} \frac{a_2^j}{\beta(j)} \prod_{n=1}^{j} b_{n-1} \right]^{-1}$$

is the probability of 0 customers at Node 2 when Node 2 is thought of as an isolated queue with state-dependent arrival rate λb_j *and service rate* μ_2*. The stability condition is*

$$\sum_{j=0}^{\infty} \frac{a_2^j}{\beta(j)} \prod_{n=1}^{j} b_{n-1} < \infty.$$

This solution still holds if we allow Node 1 has general service time with mean $1/\mu_1$*, i.e., a* $M/G/\infty$ *queue since* $M/G/\infty$ *has the same stationary distribution as* $M/M/\infty$ *queue and it has the Poisson in and Poisson out property in equilibrium as well.*

Remark 4.2.4 *For different form of* b_j*, we can obtain some special models. For example, in case of constant service rates, if*

$$b_j = \begin{cases} 1 & if \quad 0 \le j < N \\ 0 & if \quad\ \ j \ge N \end{cases},$$

Node 2 becomes $M/M/1/N$*, where* $N-1$ *is the number of buffers. Our result becomes*

$$\pi_{ij} = (1-a_1)a_1^i a_2^j \pi_0,\ i \ge 0, 0 \le j \le N,$$

where $a_1 = \lambda/\mu_1$; $a_2 = \lambda/\mu_2$; $\pi_0 = (\sum_{j=0}^{N} a_2^j)^{-1} = \frac{1-a_2}{1-a_2^{N+1}}$*. Hence*

$$\pi_{ij} = (1-a_1)a_1^i a_2^j \frac{1-a_2}{1-a_2^{N+1}},\ i \ge 0, 0 \le j \le N.$$

4.2.3 Semiopen case

In this case, we have a state-dependent arrival rate to Node 1, which is

$$\lambda(i,j) = \begin{cases} \lambda & if \quad i+j < N \\ 0 & otherwise \end{cases}$$

where i and j are the number of calls at Node 1 and 2 respectively upon arrival. In other words the model is a semiopen network model with state-dependent balking and state-dependent service. The maximum number of calls in the system is N. Now the model is a

finite state model and there always exists π_{ij}, the stationary distribution of $Q(t)$. Similarly we can show that π_{ij} has a product form solution.

Theorem 4.2.2 *For the above semiopen network model, the stationary queue length distribution π_{ij} has a product form*

$$\pi_{ij} = \prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00}, \quad i \geq 0, j \geq 0, i+j \leq N$$

where $a_1(n) = \lambda/\mu_1(n)$; $a_2(n) = \lambda/\mu_2(n)$;

$$\pi_{00} = \left[\sum_{0 \leq i+j \leq N} \prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1} \right]^{-1}$$

and the empty product is 1 by convention.

Proof. Now the global balance equations become

$$\begin{aligned}
&\pi_{ij}(\lambda + \mu_2(j) + \mu_1(i)) = \pi_{(i-1)j}\lambda + \\
&\pi_{i(j+1)}\mu_2(j+1) + \pi_{(i+1)(j-1)}\mu_1(i+1)b_{j-1} + \pi_{(i+1)j}\mu_1(i+1)\bar{b}_j, \ i \geq 1, j \geq 1, i+j \leq N-1; \\
&\pi_{ij}(\mu_2(j) + \mu_1(i)) = \pi_{(i-1)j}\lambda + \pi_{(i+1)(j-1)}\mu_1(i+1)b_{j-1}, \ i \geq 1, j \geq 1, i+j = N; \\
&\pi_{0j}(\lambda + \mu_2(j)) = \pi_{0(j+1)}\mu_2(j+1) + \pi_{1(j-1)}\mu_1(1)b_{j-1} + \pi_{1j}\mu_1(1)\bar{b}_j, \ i = 0, 1 \leq j \leq N-1; \\
&\pi_{0N}\mu_2(N) = \pi_{1(N-1)}\mu_1(1)b_{N-1}, \ i = 0, j = N; \\
&\pi_{i0}(\lambda + \mu_1(i)) = \pi_{(i-1)0}\lambda + \pi_{i1}\mu_2(1) + \pi_{(i+1)0}\mu_1(i+1)\bar{b}_0, \ 1 \leq i \leq N-1, j = 0; \\
&\pi_{N0}\mu_1(N) = \pi_{(N-1)0}\lambda, \ i = N, j = 0; \\
&\pi_{00}\lambda = \pi_{01}\mu_2(1) + \pi_{10}\mu_1(1)\bar{b}_0, \ i = 0, j = 0.
\end{aligned} \tag{4.8}$$

where the last six equations are boundary cases. Basically the equations are the same as (4.2) except that we have more boundary conditions. Again we will try to solve the station balance equations first. For Node 1 the station balance equations are

$$\pi_{ij}\mu_1(i) = \pi_{(i-1)j}\lambda, \ i \geq 1, j \geq 0, i+j \leq N. \tag{4.9}$$

For Node 2 the station balance equations are

$$\pi_{ij}\mu_2(j) = \pi_{(i+1)(j-1)}\mu_1(i+1)b_{j-1}, \ i \geq 0, j \geq 1, i+j \leq N. \tag{4.10}$$

For the whole network, the station balance equations are

$$\pi_{ij}\lambda = \pi_{i(j+1)}\mu_2(j+1) + \pi_{(i+1)j}\mu_1(i+1)\bar{b}_j, \ i \geq 0, j \geq 0, i+j \leq N-1. \tag{4.11}$$

Now the global balance equations have been decomposed into the sum of three station balance equations, which are easier to deal with. Similarly as in the proof of Theorem 4.2.1, we can solve station equations (4.9) (4.10) (4.11) and obtain the solution

$$\pi_{ij} = \prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00},\ i \geq 0, j \geq 0, i+j \leq N.$$

Since $\sum_{0 \leq i+j \leq N} \pi_{ij} = 1$, we have

$$1 = \sum_{0 \leq i+j \leq N} \prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1}\pi_{00}.$$

Therefore

$$\pi_{00} = \left[\sum_{0 \leq i+j \leq N} \prod_{n=1}^{i} a_1(n) \prod_{n=1}^{j} a_2(n) \prod_{n=1}^{j} b_{n-1} \right]^{-1}$$

and we do not need any stability conditions.

It can be verified that this solution satisfies the boundary cases in (4.8) as well. ∎

Remark 4.2.5 *Comparing to the open network case, the semiopen network has the same distribution except the normalizing constant due to the finite state space. We can think of the semiopen network as the open network conditioned that it has no larger than N calls in the system.*

Remark 4.2.6 *The semiopen network model is more useful in modelling since in reality we never have infinite buffers or trunk lines in case of call centres. In fact the semiopen network model with state-dependent balking of call centres studied in the next section is a special case of this model with specific service rates at two nodes.*

4.3 SOQN model with state-dependent balking of call centres

In this section, we will introduce balking to our SOQN model. For the queue length process, the SOQN model with state-dependent balking is just a special case of the model in the last section and also has product form solution for the stationary distribution. We will give an alternative proof as well. Waiting time distribution and mean waiting time will also be studied.

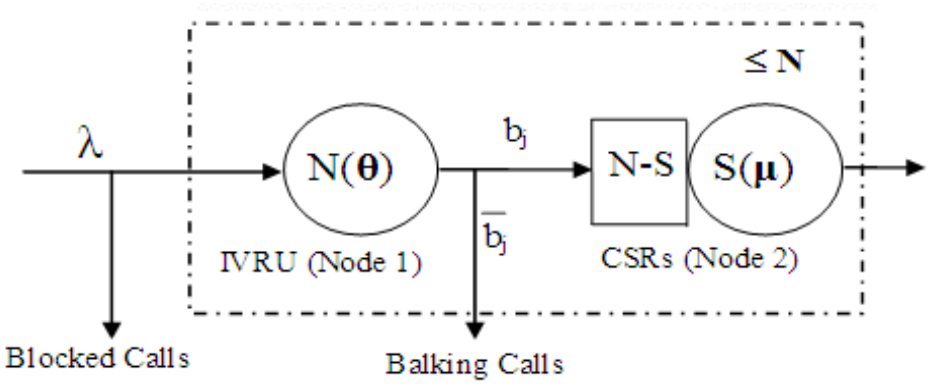

Figure 4.5: SOQN model with state-dependent balking

4.3.1 Model description

The model is a semiopen network with two nodes in series. Node 1 models the IVRU with N servers each with exponential service rate θ and Node 2 models the CSRs with S ($\leq N$) servers each with exponential service rate μ. The maximum number of calls in the network is N, i.e., if an arriving call finds N calls in the system, it will be blocked and rejected entering the system. Hence there is no queue at Node 1 and there are at most $N - S$ calls waiting at Node 2. Arriving calls can enter the network only through Node 1 according to a Poisson process with arrival rate λ. After the service with Node 1 is completed, the call leaves the network with probability $\bar{b}_j$ and it joins Node 2 with probability $b_j = 1 - \bar{b}_j$ where j is the number of calls at Node 2 when the call leaves Node 1. Here b_j satisfies conditions (4.1). If there are free CSRs at Node 2, the call is served by one of S CSRs, otherwise it waits in the queue and leaves the network after the service with Node 2. The original SOQN model in Chapter 3 can be thought of as a constant balking model. Figure 4.5 gives a picture of the model.

4.3.2 Product form solution of the queue length process

From the above model description, we know that it is a special case of the semiopen network model studied in Section 4.2 with service rates $\mu_1(n) = n\theta$ and $\mu_2(n) = (S \wedge n)\mu$. Now we have that Node 1 is similar to a $M/M/\infty$ queue with arrival rate λ and service rate θ.

Node 2 is similar to a $M(n)/M/S$ queue with state-dependent arrival rate λb_j and service rate μ. According to Theorem 4.2.2, the product form solution is

$$\pi_{ij} = \pi_{00} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)} \prod_{n=1}^{j} b_{n-1}, \qquad 0 \le i+j \le N, \tag{4.12}$$

where $a_1 = \lambda/\theta$; $a_2 = \lambda/\mu = a$;

$$\beta(j) = \begin{cases} j! & for\ 0 \le j \le S \\ S!S^{j-S} & for\ S \le j \le N \end{cases}$$

and

$$\pi_{00} = \left[\sum_{0 \le i+j \le N} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)} \prod_{n=1}^{j} b_{n-1} \right]^{-1}.$$

4.3.3 An alternative proof

In the following we will use the similar method in Chapter 3 to prove that for this model, the stationary distribution of queue length process has product form solution. As in Chapter 3, to solve the stationary distribution of queue length process for SOQN, we introduce a fictitious node (Node 0) with service rate

$$\mu_0(m) = \begin{cases} \lambda & if \quad m > 0 \\ 0 & if \quad m = 0 \end{cases}$$

to convert our model to three-node closed network. We then prove that the converted closed network still has product form solution. Our method is first to write down the global balance equations and then guess the solution. Finally we will verify that the solution satisfies the global balance equations. This solution for converted closed network is actually the solution for the original semiopen network.

Let π_{kij} be the stationary probability of having k calls at Node 0, i calls at Node 1, and j calls at Node 2 respectively. The global balance equations for this CQN model is

$$\begin{aligned} &\pi_{kij}(\mu_0(k) + \mu_1(i) + \mu_2(j)) \qquad (4.13) \\ &= \pi_{(k+1)(i-1)j}\delta(i)\mu_0(k+1) + \pi_{(k-1)(i+1)j}\delta(k)\mu_1(i+1)\bar{b}_j \\ &+\pi_{k(i+1)(j-1)}\delta(j)\mu_1(i+1)b_{j-1} + \pi_{(k-1)i(j+1)}\delta(k)\mu_2(j+1) \\ &\forall (k,i,j) \in \mathcal{Z}_+^3 \text{ such that } k+i+j = N, \end{aligned}$$

where $\delta(i) = \begin{cases} 1 & if \quad i > 0 \\ 0 & if \quad i = 0 \end{cases}$ accounts for the boundary cases. Actually the above global balance equations are equivalent to the global balance equations (4.8) for SOQN model.

Now we need to solve the above equations. According to our method, we first assume the solution has product form and then follow the usual way to get the solution. In the end we will verify that the solution satisfies the above global balance equations (4.13).

We can write down the traffic equations of our closed network

$$\begin{cases} v_1 = v_0 \\ v_2 = v_1 b_j \\ v_0 = v_1 \bar{b}_j + v_2 \end{cases}.$$

By letting $v_1 = v_0 = \lambda$, we can easily get the solution: $v_1 = v_0 = \lambda$ and $v_2 = \lambda b_j$. These are the arriving rates for these three nodes if we see them in isolation. For service rates, as in Chapter 3, we have

$$M_i(n) = \begin{cases} \prod_{m=1}^{n} \mu_i(m) & if \quad n > 0 \\ 1 & if \quad n = 0 \end{cases} \quad i = 0, 1, 2. \tag{4.14}$$

If we look at these three nodes in isolation, they are all truncated birth-death process. From the well-known stationary distribution of birth-death process:

$$p_j = \begin{cases} \prod_{i=1}^{j} \frac{\alpha_{i-1}}{\beta_i} p_0 & if \quad j > 0 \\ (1 + \sum_{j=1}^{\infty} \prod_{i=1}^{j} \frac{\alpha_{i-1}}{\beta_i})^{-1} & if \quad j = 0 \end{cases}$$

where α_i $i \geq 0$ and β_i $i \geq 1$ are state-dependent Birth and Death rate respectively, we have: For Node 0, the stationary distribution is

$$P(Y_0 = k) = P(Y_0 = 0)\frac{\lambda^k}{M_0(k)} = P(Y_0 = 0)\frac{\lambda^k}{\lambda^k} = P(Y_0 = 0), \quad 0 \leq k \leq N;$$

For Node 1, the stationary distribution is

$$P(Y_1 = i) = P(Y_1 = 0)\frac{\lambda^i}{M_1(i)}, \quad 0 \leq i \leq N;$$

For Node 2, the stationary distribution is

$$P(Y_2 = j) = P(Y_2 = 0)\frac{\lambda^j}{M_2(j)} \prod_{n=1}^{j} b_{n-1}, \ 0 \leq j \leq N.$$

Therefore, if our model has product form solution, the solution should be

$$\pi_{kij} = \kappa \frac{\lambda^i}{M_1(i)} \frac{\lambda^j}{M_2(j)} \prod_{n=1}^{j} b_{n-1} \tag{4.15}$$

$\forall (k,i,j) \in \mathcal{Z}_+^3$ such that $k+i+j=N$ and κ is the normalizing constant.

Lemma 4.3.1 *Our closed network has the product form solution (4.15).*

Proof. We need to verify that the above solution (4.15) satisfies the global balance equations (4.13). Without loss of generality, we only check for $k \geq 1, i \geq 1$ and $j \geq 1$. The boundary cases can be checked similarly. If we substitute the solution (4.15) to (4.13), the left hand side of (4.13) becomes

$$\begin{aligned} & \pi_{kij} \left[\mu_0(k) + \mu_1(i) + \mu_2(j)\right] \\ &= \left[\lambda + \mu_1(i) + \mu_2(j)\right] \kappa \frac{\lambda^i}{M_1(i)} \frac{\lambda^j}{M_2(j)} \prod_{n=1}^{j} b_{n-1} \\ &= \kappa \left[\frac{\lambda^{i+j+1}}{M_1(i)M_2(j)} + \frac{\lambda^{i+j}}{M_1(i-1)M_2(j)} + \frac{\lambda^{i+j}}{M_1(i)M_2(j-1)} \right] \prod_{n=1}^{j} b_{n-1}. \end{aligned} \tag{4.16}$$

The first term of the right hand side of (4.13) becomes

$$\begin{aligned} & \pi_{(k+1)(i-1)_j} \delta(i) \mu_0(k+1) \\ &= \kappa \frac{\lambda^{i-1}}{M_1(i-1)} \frac{\lambda^j}{M_2(j)} \prod_{n=1}^{j} b_{n-1} \lambda \\ &= \kappa \frac{\lambda^{i+j}}{M_1(i-1)M_2(j)} \prod_{n=1}^{j} b_{n-1}. \end{aligned}$$

The second term of the right hand side of (4.13) becomes

$$\begin{aligned} & \pi_{(k-1)(i+1)_j} \delta(k) \mu_1(i+1) \bar{b}_j \\ &= \kappa \frac{\lambda^{i+1}}{M_1(i+1)} \frac{\lambda^j}{M_2(j)} \prod_{n=1}^{j} b_{n-1} \mu_1(i+1) \bar{b}_j \\ &= \kappa \frac{\lambda^{i+j+1}}{M_1(i)M_2(j)} \prod_{n=1}^{j} b_{n-1}(1-b_j) \\ &= \kappa \frac{\lambda^{i+j+1}}{M_1(i)M_2(j)} \prod_{n=1}^{j} b_{n-1} - \kappa \frac{\lambda^{i+j+1}}{M_1(i)M_2(j)} \prod_{n=0}^{j} b_n. \end{aligned}$$

The third term of the right hand side of (4.13) becomes

$$\begin{aligned}
&\pi_{k(i+1)(j-1)}\delta(j)\mu_1(i+1)b_{j-1} \\
&= \kappa\frac{\lambda^{i+1}}{M_1(i+1)}\frac{\lambda^{j-1}}{M_2(j-1)}\prod_{n=1}^{j-1} b_{n-1}\mu_1(i+1)b_{j-1} \\
&= \kappa\frac{\lambda^{i+j}}{M_1(i)M_2(j-1)}\prod_{n=1}^{j} b_{n-1}.
\end{aligned}$$

The forth term of the right hand side of (4.13) becomes

$$\begin{aligned}
&\pi_{(k-1)i(j+1)}\delta(k)\mu_2(j+1) \\
&= \kappa\frac{\lambda^{i}}{M_1(i)}\frac{\lambda^{j+1}}{M_2(j+1)}\prod_{n=1}^{j+1} b_{n-1}\mu_2(j+1) \\
&= \kappa\frac{\lambda^{i+j+1}}{M_1(i)M_2(j)}\prod_{n=0}^{j} b_{n}.
\end{aligned}$$

Adding the right hand side of the last four formulas, we have

$$\kappa\left[\frac{\lambda^{i+j}}{M_1(i-1)M_2(j)} + \frac{\lambda^{i+j+1}}{M_1(i)M_2(j)} + \frac{\lambda^{i+j}}{M_1(i)M_2(j-1)}\right]\prod_{n=1}^{j} b_{n-1}$$

which equals to the right hand side of (4.16). ∎

Let π_{ij} be the stationary probabilities of having i calls at Node 1 and j calls at Node 2. Since

$$M_1(i) = i!\theta^i, 0 \le i \le N$$

and $M_2(j) = \beta(j)\mu^j$. We can easily get the following result, which agrees with (4.12).

Theorem 4.3.1 *Our semiopen network model with state-dependent balking has product form solution*

$$\pi_{ij} = \pi_{00}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}\prod_{n=1}^{j} b_{n-1}, \qquad 0 \le i+j \le N,$$

where $a_1 = \lambda/\theta$; $a_2 = \lambda/\mu$;

$$\beta(j) = \begin{cases} j! & for\ 0 \le j \le S \\ S!S^{j-S} & for\ S \le j \le N \end{cases}$$

and

$$\pi_{00} = \left[\sum_{0\le i+j\le N}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}\prod_{n=1}^{j} b_{n-1}\right]^{-1}.$$

Remark 4.3.1 *This model is a generalization of the model studied in [42] where* $b_j = p$, $0 \le j < N$.

4.3.4 Blocking probability

As in [42], we can similarly get the stationary probabilities π_k for $0 \leq k \leq N$ that there are exactly k calls in the system

$$\pi_k = \sum_{j=0}^{k} \pi_{(k-j)j}.$$

We will distinguish two cases:

1. $0 \leq k \leq S$:

$$\pi_k = \pi_{00} \sum_{j=0}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} = \pi_{00} \frac{(a_1 + a_2)^k}{k!}.$$

2. $S < k \leq N$:

$$\begin{aligned}
\pi_k &= \pi_{00} \left[\sum_{j=0}^{S} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!S!S^{j-S}} \prod_{n=1}^{j} b_{n-1} \right] \\
&= \pi_{00} \left[\sum_{j=0}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} \prod_{n=1}^{j} b_{n-1} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \prod_{n=1}^{j} b_{n-1} \left(\frac{1}{S!S^{j-S}} - \frac{1}{j!} \right) \right].
\end{aligned}$$

Therefore, the probability π_k that there are exactly $0 \leq k \leq N$ calls in the system is equal to

$$\pi_k = \pi_{00} \left[\sum_{j=0}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} \prod_{n=1}^{j} b_{n-1} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \prod_{n=1}^{j} b_{n-1} \left(\frac{1}{S!S^{j-S}} - \frac{1}{j!} \right) I_{(S,\infty)}(k) \right]$$

where

$$\begin{aligned}
\pi_{00} &= \pi_0 \\
&= \left[\sum_{k=0}^{S} \frac{(a_1 + a_2)^k}{k!} + \sum_{k=S+1}^{N} \left(\sum_{j=0}^{S} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!S!S^{j-S}} \prod_{n=1}^{j} b_{n-1} \right) \right]^{-1} \\
&= \left[\sum_{k=0}^{N} \sum_{j=0}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} \prod_{n=1}^{j} b_{n-1} + \sum_{k=S+1}^{N} \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \prod_{n=1}^{j} b_{n-1} \left(\frac{1}{S!S^{j-S}} - \frac{1}{j!} \right) \right]^{-1}
\end{aligned}$$

and the blocking probability is $P(blocking) = \pi_N$.

4.3.5 Other performance measures

We can get other performance measures from the stationary marginal distribution of Node 2. When we look at Node 2 only, we have

$$
\begin{aligned}
1 &= \pi_{*N} + \sum_{j=0}^{S-1} \pi_{*j} + \sum_{j=S}^{N-1} b_j \pi_{*j} + \sum_{j=S}^{N-1} \overline{b}_j \pi_{*j} \\
&= \pi_{0N} + \sum_{j=0}^{S-1} \left[\sum_{i=0}^{N-1-j} \pi_{ij} + \pi_{(N-j)j} \right] + \sum_{j=S}^{N-1} b_j \left[\sum_{i=0}^{N-1-j} \pi_{ij} + \pi_{(N-j)j} \right] + \sum_{j=S}^{N-1} \overline{b}_j \left[\sum_{i=0}^{N-1-j} \pi_{ij} + \pi_{(N-j)j} \right] \\
&= \pi_{0N} + \sum_{j=0}^{N-1} \pi_{(N-j)j} + \sum_{j=0}^{S-1} \sum_{i=0}^{N-1-j} \pi_{ij} + \sum_{j=S}^{N-1} b_j \sum_{i=0}^{N-1-j} \pi_{ij} + \sum_{j=S}^{N-1} \overline{b}_j \sum_{i=0}^{N-1-j} \pi_{ij} \\
&= P(blocking) + \sum_{j=0}^{S-1} \sum_{i=0}^{N-1-j} \pi_{ij} + \sum_{j=S}^{N-1} b_j \sum_{i=0}^{N-1-j} \pi_{ij} + \sum_{j=S}^{N-1} \overline{b}_j \sum_{i=0}^{N-1-j} \pi_{ij}, \qquad (4.17)
\end{aligned}
$$

where π_{*j} is the stationary marginal distribution of Node 2. Hence, we have:

1. $P(\text{no-delay}, entry) = \sum_{j=0}^{S-1} \sum_{i=0}^{N-1-j} \pi_{ij}$.
2. $P(delay, entry) = \sum_{j=S}^{N-1} b_j \sum_{i=0}^{N-1-j} \pi_{ij}$.
3. $P(balking) = \sum_{j=S}^{N-1} \overline{b}_j \sum_{i=0}^{N-1-j} \pi_{ij}$.

4.3.6 Waiting time distribution and mean waiting time

In order to find the important performance measure: $TSF = P(W_q < AWT)$, we need to have the waiting time distribution. As in SOQN model, we only need to consider those calls given they are not blocked and join Node 2 without balking (or given entry), since there is no queue at Node 1. Let W_q denote the conditional stationary waiting time of calls given entry, which is the time spent by an entry call in the queue of Node 2 until starting to get service.

Using q_j

To find the distribution of W_q, we need first to find q_j, the probability of finding j calls at Node 2 by a joining call from Node 1 at arrival instant. q_j is actually the probability of a call finding j calls at Node 2 given entry. Using (4.17) and conditional argument, it is

easy to see that

$$q_j = \frac{b_j \sum_{i=0}^{N-1-j} \pi_{ij}}{1 - P(blocking) - \sum_{j=S}^{N-1} \overline{b}_j \sum_{i=0}^{N-1-j} \pi_{ij}} \tag{4.18}$$
$$= \frac{b_j \sum_{i=0}^{N-1-j} \pi_{ij}}{\sum_{j=0}^{S-1} \sum_{i=0}^{N-1-j} \pi_{ij} + \sum_{j=S}^{N-1} b_j \sum_{i=0}^{N-1-j} \pi_{ij}}, \ 0 \le j < N.$$

Once we have the expression of q_j, the following performance measures can be easily obtained similar to the SOQN model.

1. $P(\text{no-delay}|entry) = P(W_q = 0) = \sum_{j=0}^{S-1} q_j$.
2. $P(delay|entry) = P(W_q > 0) = \sum_{j=S}^{N-1} q_j$.
3. $TSF = P(W_q \le t) = 1 - \sum_{j=S}^{N-1} q_j \sum_{k=0}^{j-S} \frac{(S\mu t)^k e^{-S\mu t}}{k!} = 1 - \sum_{j=S}^{N-1} q_j \overline{F}_{Y_j}(t)$, $t \ge 0$, where $Y_j \sim Er(j - S + 1, S\mu)$.
4. $ASA = E(W_q) = \int_0^\infty P(W_q > t)dt = \int_0^\infty \sum_{j=S}^{N-1} q_j \overline{F}_{Y_j}(t)dt = \sum_{j=S}^{N-1} q_j \int_0^\infty \overline{F}_{Y_j}(t)dt = \frac{1}{S\mu} \sum_{j=S}^{N-1} q_j (j - S + 1)$.

Using $\chi(k, j)$

An alternative method is to use the method of Srinivasan et al. [42], where they used $\chi(k, j)$, which can be defined here as: For $0 \le j < k \le N, \chi(k, j)$ is the probability that the system is in state $(k - j, j)$, when a call (among the $k - j$ calls) is about to leave Node 1 and join Node 2 without balking. Using Bayes' theorem, we derive

$$\begin{aligned}
&\chi(k,j) := P(\text{system in state } (k-j,j) \mid \text{call is about to leave Node 1 and join Node 2}) \\
&= \frac{P\left(\text{call is about to leave Node 1 and join 2} \mid \text{system in state } (k-j,j)\right) \pi_{(k-j)j}}{\sum_{l=0}^{N} \sum_{m=0}^{l} P\left(\text{call is about to leave Node 1 and join 2} \mid \text{system in state } (l-m,m)\right) \pi_{(l-m)m}} \\
&= \frac{(k-j)\theta b_j \pi_{(k-j)j}}{\sum_{l=0}^{N} \sum_{m=0}^{l} (l-m)\theta b_m \pi_{(l-m)m}} \\
&= \frac{(k-j) b_j \pi_{(k-j)j}}{\sum_{l=1}^{N} \sum_{m=0}^{l-1} (l-m) b_m \pi_{(l-m)m}}.
\end{aligned}$$

Then we have:

1. $P(\text{no-delay}|entry) = P(W_q = 0) = \sum_{k=1}^{N} \sum_{j=0}^{k \wedge S-1} \chi(k, j) = \sum_{j=0}^{S-1} \sum_{k=j+1}^{N} \chi(k, j)$.
2. $P(delay|entry) = P(W_q > 0) = \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k, j)$.

3. $TSF = P(W_q \le t) = 1 - \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \frac{(S\mu t)^l e^{-S\mu t}}{l!} = 1 - \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \overline{F}_{Y_j}(t)$, $t \ge 0$, where $Y_j \sim Er(j - S + 1, S\mu)$.

4. $ASA = E(W_q) = \int_0^\infty P(W_q > t)dt = \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \int_0^\infty \frac{(S\mu t)^l e^{-S\mu t}}{l!} dt = \frac{1}{S\mu} \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j)(j - S + 1)$.

Again we have the following result to relate q_j with $\chi(k,j)$.

Theorem 4.3.2 *For $0 \le j < N$, we have $q_j = \sum_{k=j+1}^{N} \chi(k,j)$.*

Proof. For $0 \le j < N$, let $i := k - j > 0$, then $k = i + j$. We have

$$\chi(k,j) = \chi(i+j,j) = \frac{ib_j\pi_{ij}}{\sum\limits_{0 \le l+m \le N} lb_m\pi_{lm}} = \frac{ib_j\pi_{ij}}{\sum\limits_{1 \le l+m \le N} lb_m\pi_{lm}}.$$

Hence

$$\begin{aligned}
\sum_{k=j+1}^{N} \chi(k,j) &= \sum_{i=1}^{N-j} \frac{ib_j\pi_{ij}}{\sum\limits_{1 \le l+m \le N} lb_m\pi_{lm}} = \sum_{i=1}^{N-j} \frac{ib_j\pi_{00}\frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}\prod\limits_{n=1}^{j} b_{n-1}}{\sum\limits_{1 \le l+m \le N} lb_m\pi_{00}\frac{a_1^l}{l!}\frac{a_2^m}{\beta(m)}\prod\limits_{n=1}^{m} b_{n-1}} \\
&= \sum_{i=1}^{N-j} \frac{a_1 b_j\pi_{00}\frac{a_1^{i-1}}{(i-1)!}\frac{a_2^j}{\beta(j)}\prod\limits_{n=1}^{j} b_{n-1}}{a_1\sum\limits_{1 \le l+m \le N} b_m\pi_{00}\frac{a_1^{l-1}}{(l-1)!}\frac{a_2^m}{\beta(m)}\prod\limits_{n=1}^{m} b_{n-1}} \\
&= \sum_{i-1=0}^{N-1-j} \frac{b_j\pi_{00}\frac{a_1^{i-1}}{(i-1)!}\frac{a_2^j}{\beta(j)}\prod\limits_{n=1}^{j} b_{n-1}}{\sum\limits_{0 \le l-1+m \le N-1} b_m\pi_{00}\frac{a_1^{l-1}}{(l-1)!}\frac{a_2^m}{\beta(m)}\prod\limits_{n=1}^{m} b_{n-1}}.
\end{aligned}$$

Now let $q := i - 1$ and $p := l - 1$, we have

$$\begin{aligned}
\sum_{k=j+1}^{N} \chi(k,j) &= \sum_{q=0}^{N-1-j} \frac{b_j\pi_{00}\frac{a_1^q}{q!}\frac{a_2^j}{\beta(j)}\prod\limits_{n=1}^{j} b_{n-1}}{\sum\limits_{0 \le p+m \le N-1} b_m\pi_{00}\frac{a_1^p}{p!}\frac{a_2^m}{\beta(m)}\prod\limits_{n=1}^{m} b_{n-1}} = \frac{b_j\sum_{q=0}^{N-1-j}\pi_{00}\frac{a_1^q}{q!}\frac{a_2^j}{\beta(j)}\prod\limits_{n=1}^{j} b_{n-1}}{\sum\limits_{0 \le p+m \le N-1} b_m\pi_{00}\frac{a_1^p}{p!}\frac{a_2^m}{\beta(m)}\prod\limits_{n=1}^{m} b_{n-1}} \\
&= \frac{b_j\sum_{q=0}^{N-1-j}\pi_{qj}}{\sum_{m=0}^{S-1}\sum_{p=0}^{N-1-m}\pi_{pm} + \sum_{m=S}^{N-1} b_m\sum_{p=0}^{N-1-m}\pi_{pm}} \\
&= q_j.
\end{aligned}$$

■

Using the relationship $q_j = \sum_{k=j+1}^{N} \chi(k,j)$, we can prove the equivalence of the two sets of the performance measures in terms of q_j and $\chi(k,j)$ respectively. For example

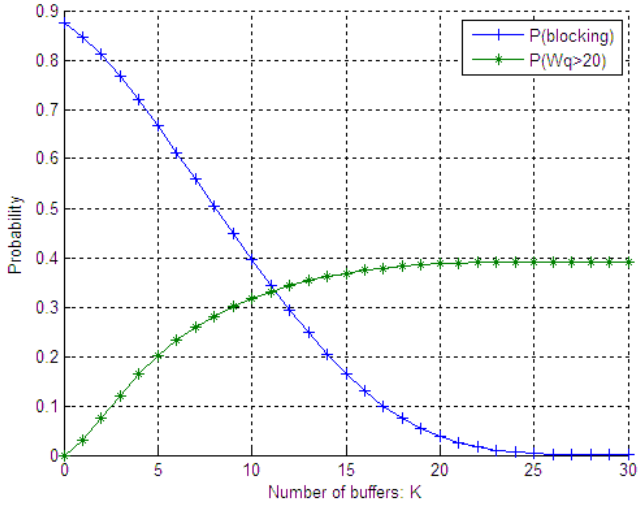

Figure 4.6: $P(blocking)$ and $P(W_q > 20)$ for Example 4.1

$$P(delay|entry) = \sum_{j=S}^{N-1} q_j = \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) = \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j).$$

Other performance measures can be verified similarly.

4.3.7 Numerical examples

To give some numerical illustrations for the SOQN model with state-dependent balking, we will consider the following example for $P(blocking)$ and $P(W_q > t)$. This example is similar to the example discussed in Chapter 3 where the parameters are $\lambda = 250/1800, \mu = 1/180, t = 20$ seconds, $\theta = 0.01$. To illustrate the effect of buffer size, we fix $S = 5$ and let buffer size $K = N - S$ change from 0 to 30. In addition we assume the balking function b_j has the form

$$b_j = \begin{cases} \frac{1}{(j-S+1)^m+1} & \text{if } S \le j < N \\ 1 & \text{if } 0 \le j < S \end{cases},$$

where non-negative m is a measure of a call's willingness to join the queue. We will consider two cases: $m = 10$ representing higher balking probability and $m = 0.5$ representing lower balking probability. The results are shown in Figure 4.6 for Example 4.1 ($m = 10$) and in Figure 4.7 for Example 4.2 ($m = 0.5$) respectively.

Since we have lower balking probability in Example 4.2 compared to that in Example

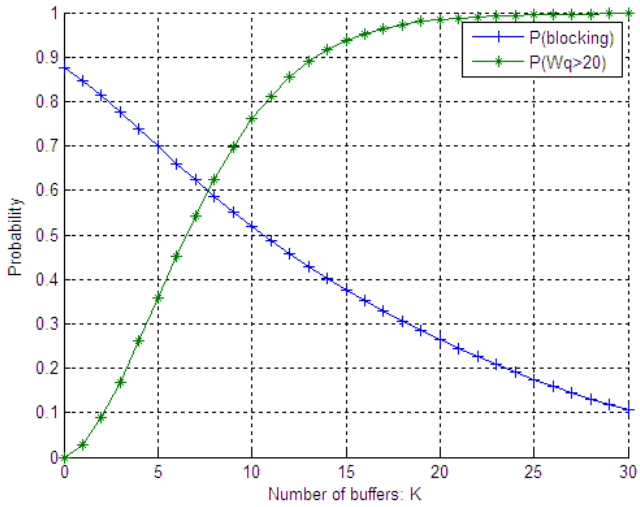

Figure 4.7: $P(blocking)$ and $P(W_q > 20)$ for Example 4.2

4.1, both $P(blocking)$ and $P(W_q > t)$ in Example 4.2 are higher than those in Example 4.1. Also from these two examples, we observed the similar monotonicity properties of $P(blocking)$ and $P(W_q > t)$ as in SOQN model. Again we have the following conjecture. When S and other parameters are fixed, $P(blocking)$ is a strictly decreasing function of K and $P(W_q{>}t)$ is a strictly increasing function of K. This conjecture is intuitively correct and we will use it in Chapter 7 for the call centre design problem.

4.4 Summary

In this chapter, we studied the balking phenomenon of call centres using single-node Markovian models and the SOQN model. We first gave a short review of the single-node state-dependent balking model $M(n)/M/S/N$, which is standard. The main work is on the SOQN model with balking, where we proved that the product form solution of the queue length process still holds. We also derived the waiting time distribution and other performance measures. In the end numerical examples were given to illustrate the effect of balking.

Chapter 5

Exponential Abandonment Models of Call Centres

Another important feature of call centres is abandonment. State-dependent balking can be thought of as the abandonment of *aware* calls who know or are informed of the state of the system upon arrival and hence abandon *before* entering the system. Unlike balking, abandonment (reneging) describes the phenomenon that *unaware* calls leave the system *after* entering the system.

The following is a common approach to include the abandonment into the queueing system. Consider a queueing system which allows for waiting in the buffer if an arriving call finds all servers busy. To incorporate the abandonment, it is assumed that there is a random variable X for each call that quantifies the call's patience. There are two types of abandonment [5]. One is *patience time on waiting*, which means an *unaware* call will leave the system if its waiting time is longer than its patience time X while waiting for the service in the queue. Hence once the call begins its service, it never abandon. The other is *patience time on sojourn*, which means an *unaware* call will leave the system if its sojourn time is longer than its patience time X irrespective of whether or not it is being served. Since in call centres, the first type of abandonment is more realistic, we will only focus on this one.

For different calls, patience times are assumed to be i.i.d. with mean $1/\alpha$, which is natural for the invisible queues occurring in call centres [47], and they are independent of all other model elements as well. If X is infinite, the model reduces to a model without abandonment discussed earlier. We are interested in a call who arrives at the system at stationary state. Let V be this call's *offered waiting time* in the queue (i.e., stationary waiting time of a call with infinite patience [33]). If $V \geq X$, the call will abandon; otherwise the call will get served. Now $W_q := V \wedge X$ is the stationary waiting time in the queue of

this call until it gets served or abandons.

One important concept in abandonment models is (stationary) abandonment rate function. For example, see [3], [6], [24], page 95, and [36]. We introduce the definition of [36],

Definition 5.0.1 *For $t, \varepsilon \in R^+$ and $i \in N$, let*

$\Psi_i(t, \varepsilon) :=$ *the probability that a call misses its deadline(abandon) during* $[t, t+\varepsilon)$, *given there are i calls in the system at time t.*

Define $r_i(t) = \lim_{\varepsilon \to 0} \frac{\Psi_i(t,\epsilon)}{\varepsilon}$, *and assuming at stationary state,* $r_i = \lim_{t \to \infty} r_i(t)$ *is called the stationary abandonment rate function.*

Movaghar [36] gave an explicit expression of r_i for $M/M/S{+}G$ model. For $M/M/S{+}M$ model it reduces to

$$r_i = \begin{cases} 0 & \text{if } 0 \le i \le S \\ (i-S)\alpha & \text{if } i > S \end{cases} . \tag{5.1}$$

In this chapter, we will focus on the exponential abandonment model of call centres, i.e., $X \sim \exp(\alpha)$. We will analyze three models, $M/M/S+M$, $M/M/S/N+M$ and SOQN+M which is SOQN model with exponential abandonment.

5.1 $M/M/S+M$ (Erlang-A model)

$M/M/S+M$ generalizes the $M/M/S$ model by including exponential abandonment, i.e., patience time X is assumed to have exponential distribution with mean $1/\alpha$. Palm first introduced this model and Garnett et al. [21] referred to it as Erlang-A model (A for Abandonment, and for the fact that it interpolates between Erlang-C and Erlang-B [33]). The model description and parameters are shown in Figure 5.1. Note that if $\mu = \alpha$, $M/M/S+M$ model becomes $M/M/\infty$ model; if $\alpha = 0$ (no abandonment) it becomes $M/M/S$ model and if $\alpha = \infty$ it becomes $M/M/S/S$ model.

5.1.1 Queue length process

The following discussion is mainly based on the work of [33]. Since the patience times are assumed to be i.i.d. exponentially distributed, the queue length process $Q(t)$ is an infinite

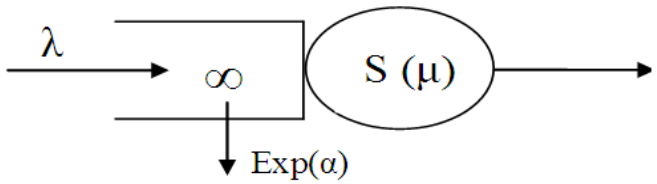

Figure 5.1: $M/M/S+M$ model description and parameters

birth-death process with birth rate $\lambda_i = \lambda$ and state-dependent death rate

$$\mu_i = \begin{cases} i\mu & \text{if } 0 \le i \le S \\ S\mu + (i-S)\alpha & \text{if } i > S \end{cases},$$

where we used the fact that for $M/M/S+M$ model, the abandonment rate is given in (5.1). We can also derive this result directly in the following by Definition 5.0.1.

Theorem 5.1.1 *For $M/M/S+M$ model, we have the stationary abandonment rate function*

$$r_i = \begin{cases} 0 & \text{if } 0 \le i \le S \\ (i-S)\alpha & \text{if } i > S \end{cases}.$$

Proof. By the definition of r_i (Definition 5.0.1), we have $r_i = \lim_{t\to\infty}\lim_{\varepsilon\to 0}\frac{\Psi_i(t,\epsilon)}{\varepsilon}$, where $\Psi_i(t,\varepsilon) = P(\text{a call abandon during } [t,t+\varepsilon)|\ i \text{ calls in the system at time } t)$. For $0 \le i \le S$, $\Psi_i(t,\varepsilon) = 0$, Hence $r_i = 0$. For $i > S$, there are $i-s$ calls in the queue, each with an exponentially distributed patience time X with mean $1/\alpha$. Then $\Psi_i(t,\varepsilon) = (i-s)\alpha\varepsilon + o(\varepsilon)$, which is independent of t because of the memoryless property of exponential distribution. Hence $r_i = (i-S)\alpha$. ∎

Remark 5.1.1 *It is easy to see that for $i > S$, $r_i = (i-S)\alpha$ is the hazard rate of an exponential distribution with mean $\frac{1}{(i-S)\alpha}$, which is the distribution of $Y = min(X_1, ...X_{i-S})$, where X_n, $n = 1, 2, ...i-S$ have exponential distribution with mean $1/\alpha$, representing patience time of calls in the queue.*

The stationary state transition diagram of $Q(t)$ is shown in Figure 5.2. The stationary distribution can be obtained by solving the global or cut balance equations derived from

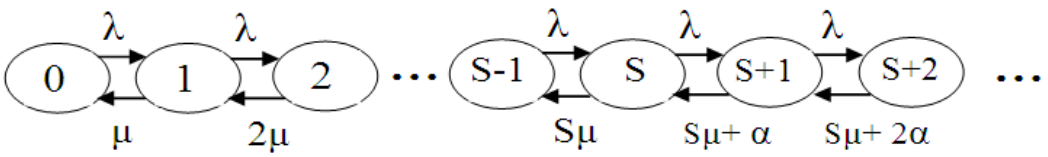

Figure 5.2: $M/M/S+M$ model stationary state transition diagram

Figure 5.2 as in [33]. The solution is

$$p_i = \begin{cases} \frac{a^i}{i!} p_0 & \text{if } 0 \le i \le S \\ \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i}[S\mu+(j-S)\alpha]} \frac{a^S}{S!} p_0 & \text{if } S < i \end{cases}$$

where p_0 is

$$p_0 = \left(\sum_{i=0}^{S} \frac{a^i}{i!} + \frac{a^S}{S!} \sum_{i=S+1}^{\infty} \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i}[S\mu + (j-S)\alpha]} \right)^{-1}.$$

This distribution exists if the infinite sum in p_0 converges. In [33], it is proved that since for $0 < \alpha \le \infty$,

$$\begin{aligned} &\sum_{i=0}^{S} \frac{a^i}{i!} + \frac{a^S}{S!} \sum_{i=S+1}^{\infty} \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i}[S\mu + (j-S)\alpha]} \\ &\le \sum_{i=0}^{\infty} \frac{(\lambda/(\mu \wedge \alpha))^i}{i!} = e^{\lambda/(\mu \wedge \alpha)} < \infty, \end{aligned} \tag{5.2}$$

p_0 always converges, i.e., there always exists the stationary distribution p_i.

In [33], it is also shown that p_i can be expressed in terms of p_S

$$p_i = \begin{cases} \frac{S!}{i! a^{S-i}} p_S & \text{if } 0 \le i \le S \\ \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i}[S\mu+(j-S)\alpha]} p_S & \text{if } S < i \end{cases} \tag{5.3}$$

where p_S is

$$\begin{aligned} p_S = \frac{a^S}{S!} p_0 &= \frac{\frac{a^S}{S!}}{\sum_{i=0}^{S} \frac{a^i}{i!} + \frac{a^S}{S!} \sum_{i=S+1}^{\infty} \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i}[S\mu+(j-S)\alpha]}} \\ &= \frac{B(S,a)}{1 + B(S,a) \sum_{i=S+1}^{\infty} \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i}[S\mu+(j-S)\alpha]}}. \end{aligned}$$

The expressions of p_i have infinite sums that can cause computational problems. In [33], special functions have been used to overcome this problem. They provided expressions

in terms of special functions for some performance measures. In this section we will follow their idea and provide more results.

As in Chapter 2, let

$$\gamma(x,y) := \int_0^y t^{x-1}e^{-t}dt, x > 0, y \geq 0$$

be the incomplete Gamma function and let

$$\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt, x > 0$$

be the Gamma function. Define

$$A(x,y) := M(1,x+1,y) = \sum_{i=0}^{\infty} \frac{y^i}{\prod_{j=1}^{i}(x+j)} = \sum_{i=0}^{\infty} \frac{\Gamma(x+1)y^i}{\Gamma(x+i+1)}, \quad x > 0, y \geq 0$$

where

$$M(a,b,z) = \sum_{i=0}^{\infty} \frac{a(a+1)...(a+i-1)z^i}{b(b+1)...(b+i-1)i!}$$

is the Kummer's hypergeometric function. It is well-known that [2]

$$A(x,y) = \frac{xe^y}{y^x}\gamma(x,y) \tag{5.4}$$

since

$$\gamma(x,y) = e^{-y}y^x \sum_{i=0}^{\infty} \frac{y^i}{\prod_{j=0}^{i}(x+j)}, x > 0, y \geq 0.$$

Now we will introduce the following notations: $C := S\mu/\alpha; \eta := \lambda/\alpha; \rho := \frac{\eta}{C} = \lambda/S\mu; \rho_i := \frac{\eta^i}{\prod_{j=1}^{i}(C+j)} = \frac{\Gamma(C+1)}{\Gamma(C+i+1)}\eta^i$. Therefore

$$A(C,\eta) = \sum_{i=0}^{\infty} \rho_i.$$

Also for simplicity we will replace $B(S,a)$ with B and $A(C,\eta)$ with A. Therefore from (5.3) we have

$$p_i = \begin{cases} \frac{S!}{i!a^{S-i}}p_S & \text{if } 0 \leq i \leq S \\ \rho_{i-S}p_S & \text{if } S < i \end{cases}$$

where

$$p_S = \frac{B}{1+(A-1)B} \tag{5.5}$$

as given in [33].

From the above, performance measures related to p_i can be obtained. For example, by the PASTA property, the probability of delay among all calls (abandon or served) is obtained in [33],

$$P(delay) = P(W_q > 0) = \sum_{i=S}^{\infty} p_i = p_S \sum_{i=S}^{\infty} \rho_{i-S} = \frac{AB}{1+(A-1)B} \tag{5.6}$$

where $W_q = V \wedge X$ is the stationary waiting time in the queue of a call until it gets served or abandons. The above formula corresponds to Erlang C formula in $M/M/S$ model. When $\alpha = 0$, $\rho_i = \rho^i$ and $A = \frac{1}{1-\rho}$ under the condition $0 < \rho < 1$. In this case the above formula reduces to Erlang C formula.

We can also derive three kinds of mean number of calls which are not available in [33].

1. Mean number of busy servers $E(Q_b)$

$$\begin{aligned} E(Q_b) &= \sum_{i=1}^{S} i p_i + S \sum_{i=S+1}^{\infty} p_i \\ &= p_S \left[\sum_{i=1}^{S} i \frac{S!}{i! a^{S-i}} + S \sum_{i=S+1}^{\infty} \rho_{i-S} \right] \\ &= p_S \left[a(\frac{1}{B} - 1) + S(A-1) \right] \\ &= \frac{B}{1+(A-1)B} \frac{a(1-B)+(A-1)BS}{B} \\ &= \frac{a(1-B)+(A-1)BS}{1+(A-1)B}. \end{aligned} \tag{5.7}$$

2. Mean number of calls waiting in the queue $E(Q_q)$

$$\begin{aligned} E(Q_q) &= \sum_{i=S+1}^{\infty} (i-S) p_i = p_S \sum_{i=1}^{\infty} i \rho_i \\ &= p_S \sum_{i=1}^{\infty} i \frac{\Gamma(C+1)\eta^i}{\Gamma(C+i+1)} \\ &= p_S \left[\frac{d}{d\eta} \sum_{i=1}^{\infty} \frac{\Gamma(C+1)\eta^{i+1}}{\Gamma(C+i+1)} - \sum_{i=1}^{\infty} \frac{\Gamma(C+1)\eta^i}{\Gamma(C+i+1)} \right] \\ &= p_S \left[\frac{d}{d\eta} [\eta(A-1)] - (A-1) \right] \\ &= p_S \eta \frac{d}{d\eta} A = p_S \eta \frac{1+A(\rho-1)}{\rho} \\ &= \frac{BC[1+A(\rho-1)]}{1+(A-1)B}, \end{aligned} \tag{5.8}$$

where we have used (5.5) and the fact that

$$\frac{d}{d\eta}A(C,\eta) = \frac{1 + A(C,\eta)(\rho - 1)}{\rho}. \tag{5.9}$$

3. Mean number of calls in the system $E(Q)$

$$E(Q) = E(Q_b) + E(Q_q).$$

5.1.2 Waiting time distribution

Performance measures on waiting time are not available in [33]. In this section, we will study waiting time distribution in detail and express the results in terms of special functions and Erlang B formula. As discussed earlier, $W_q = V \wedge X$ is defined to be the stationary waiting time in the queue of a call until it gets served or abandons, where V is the offered waiting time and X is the patience time. To find $P(W_q > t)$, as in Movaghar [36], we will define an important random variable: *conditional offered waiting time* $V_i :=$ The waiting time of a call with infinite patience in the queue, given it finds i calls in the system upon arrival, for $i \geq 0$. By the total probability law and the PASTA property, we have,

$$P(V > t) = \sum_{i=S}^{\infty} p_i P(V_i > t),\ t \geq 0. \tag{5.10}$$

Hence as long as we have the distribution of V_i, we will get the distribution of V. Furthermore we will have the distribution of W_q

$$P(W_q > t) = P(V \wedge X > t) = P(V > t)P(X > t) = e^{-\alpha t}P(V > t), \tag{5.11}$$

since V and X are independent. The following theorem for the distribution of V_i is adapted from a similar result in Wang [45].

Theorem 5.1.2 *For $0 \leq i < S$, $V_i = 0$. For $i \geq S$, $V_i = \sum_{n=0}^{i-S} \phi_n$, where $\phi_n \sim \exp(S\mu + n\alpha)$ so that for $t \geq 0$ the density function of V_i is*

$$f_{V_i}(t) = \sum_{n=0}^{i-S} A_{n,i-S}(S\mu + n\alpha)e^{-(S\mu+n\alpha)t} \tag{5.12}$$

and the survival function of V_i is

$$P(V_i > t) = \sum_{n=0}^{i-S} A_{n,i-S}e^{-(S\mu+n\alpha)t}, \tag{5.13}$$

where $A_{n,i-S} := \prod_{\substack{k=0 \\ k\neq n}}^{i-S} \frac{C+k}{k-n}$.

Proof. For $0 \leq i < S$, the call gets served immediately so that $V_i = 0$. For $i \geq S$, the call will be in the $i-S+1$ position of the queue upon arrival. There are $i-S$ calls ahead of him and each has an abandonment rate α. Also there are S calls being served and each has a service rate μ. Therefore the time for the call to go from position $i-S+1$ to position $i-S$ is $\exp(S\mu+(i-S)\alpha)$. Generally, the time for the call to go from position $n+1$ to position n is $\exp(S\mu+n\alpha)$, for $0 \leq n \leq i-S$, where position 0 means the time point when the call leaves the queue and starts to get served. Hence the total time of an infinite patience call will wait in the queue given it finds i calls in the system is the sum of $i-S+1$ independent exponential variables with rates $S\mu+n\alpha$, i.e., $V_i = \sum_{n=0}^{i-S} \phi_n, where\ \phi_n \sim \exp(S\mu+n\alpha)$. The Laplace transform of V_i is $f_{V_i}^*(s) = \prod_{n=0}^{i-S} \frac{S\mu+n\alpha}{S\mu+n\alpha+s}$, which can be expressed in partial fraction form as $f_{V_i}^*(s) = \sum_{n=0}^{i-S} \frac{A_{n,i-S}(S\mu+n\alpha)}{S\mu+n\alpha+s}$ where $A_{n,i-S} := \prod_{\substack{k=0 \\ k\neq n}}^{i-S} \frac{S\mu+k\alpha}{(k-n)\alpha}$. See Cox [18] page 17. Therefore we have the density function of V_i as (5.12) and the survival function of V_i as (5.13). Since $P(V_i > 0) = 1$ we have the identity: $\sum_{n=0}^{i-S} A_{n,i-S} = 1$. The distribution of V_i is called hypoexponential. See Ross [40] for the proof of (5.12) using the mathematical induction method. ∎

Remark 5.1.2 *By using the identity:* $\prod_{\substack{k=0 \\ k\neq n}}^{i-S} \frac{1}{(k-n)} = \frac{(-1)^n}{n!(i-S-n)!}$, *we have*

$$A_{n,i-S} = \frac{(-1)^n}{n!(i-S-n)!} \frac{\prod_{k=0}^{i-S}(C+k)}{C+n}. \tag{5.14}$$

Hence we obtain other expressions of (5.12) and (5.13), i.e.,

$$f_{V_i}(t) = \frac{\alpha \prod_{k=0}^{i-S}(C+k)}{(i-S)!}(1-e^{-\alpha t})^{i-S} e^{-S\mu t}\ i \geq S, t \geq 0, \tag{5.15}$$

and

$$P(V_i > t) = \frac{\prod_{k=0}^{i-S}(C+k)}{(i-S)!} \sum_{n=0}^{i-S} \binom{i-S}{n} \frac{(-1)^n}{C+n} e^{-(S\mu+n\alpha)t}\ i \geq S,\ t \geq 0, \tag{5.16}$$

which also appeared in Riordan [39].

The following result is a special case of the result by Baccelli and Hebuterne [5].

Theorem 5.1.3 *The density function of* V *is*

$$f_V(t) = \begin{cases} p_S S\mu e^{\eta} e^{-(S\mu t+\eta e^{-\alpha t})} & if\ t > 0 \\ \sum_{i=0}^{S-1} p_i & if\ t = 0 \end{cases}, \tag{5.17}$$

and the survival function of V is

$$P(V > t) = p_S C e^{\eta} \eta^{-C} \gamma(C, \eta e^{-\alpha t}),\ t \geq 0.$$

Proof. It is obvious that V has a mass of $\sum_{i=0}^{S-1} p_i$ at 0 and for $t > 0$ by using (5.3) and (5.15), we have

$$\begin{aligned} f_V(t) &= \sum_{i=S}^{\infty} p_i f_{V_i}(t) = p_S \sum_{i=S}^{\infty} \frac{\eta^{i-S}}{\prod_{k=1}^{i-S}(C+k)} \frac{\alpha \prod_{k=0}^{i-S}(C+k)}{(i-S)!} (1-e^{-\alpha t})^{i-S} e^{-S\mu t} \\ &= p_S \sum_{i=S}^{\infty} \frac{\eta^{i-S} S\mu}{(i-S)!} (1-e^{-\alpha t})^{i-S} e^{-S\mu t} \\ &= p_S S\mu e^{-S\mu t} \sum_{i=S}^{\infty} \frac{\left[\eta(1-e^{-\alpha t})\right]^{i-S}}{(i-S)!} \\ &= p_S S\mu e^{-S\mu t} e^{\eta(1-e^{-\alpha t})} = p_S S\mu e^{\eta} e^{-(S\mu t + \eta e^{-\alpha t})}. \end{aligned}$$

So that

$$\begin{aligned} P(V > t) &= \int_t^{\infty} f_V(u) du = p_S S\mu e^{\eta} \int_t^{\infty} e^{-(S\mu u + \eta e^{-\alpha u})} du \\ &= p_S S\mu e^{\eta} \int_0^{\eta e^{-\alpha t}} e^{-\left[S\mu(-\frac{1}{\alpha}\ln\frac{z}{\eta}) + z\right]} \frac{1}{z\alpha} dz \\ &= p_S C e^{\eta} \eta^{-C} \int_0^{\eta e^{-\alpha t}} z^{C-1} e^{-z} dz \\ &= p_S C e^{\eta} \eta^{-C} \gamma(C, \eta e^{-\alpha t}), \end{aligned}$$

where we have used the substitution $\eta e^{-\alpha u} = z$. ∎

Remark 5.1.3 *By (5.5) and the relationship between A and $\gamma(C,\eta)$ (5.4), we have*

$$P(V > t) = p_S C e^{\eta} \eta^{-C} \gamma(C, \eta e^{-\alpha t}) = \frac{AB}{1+(A-1)B} \frac{\gamma(C, \eta e^{-\alpha t})}{\gamma(C,\eta)}.$$

Remark 5.1.4 *According to (5.11),*

$$\begin{aligned} P(W_q > t) &= p_S C e^{\eta} \eta^{-C} \gamma(C, \eta e^{-\alpha t}) e^{-\alpha t} \\ &= \frac{AB}{1+(A-1)B} \frac{\gamma(C, \eta e^{-\alpha t})}{\gamma(C,\eta)} e^{-\alpha t} = P(W_q > 0) \frac{\gamma(C, \eta e^{-\alpha t})}{\gamma(C,\eta)} e^{-\alpha t}, \end{aligned} \tag{5.18}$$

where $P(W_q > 0) = P(V > 0) = \frac{AB}{1+(A-1)B}$ is consistent with (5.6).

Remark 5.1.5 *We have $P(V > t|delay) = \frac{P(V>t, delay)}{P(delay)} = \frac{P(V>t)}{P(V>0)} = \frac{\gamma(C,\eta e^{-\alpha t})}{\gamma(C,\eta)}$, which is consistent with Riordan [39] page 111 but they used a different method. Also they have*

$$P(W_q > t|delay) = \frac{e^{-\alpha t}\gamma(C, \eta e^{-\alpha t})}{\gamma(C,\eta)}. \tag{5.19}$$

Since the model is an abandonment model, call centre managers may only consider the waiting time of those calls given they are served. We define "Sr" as the event that a call is served and $P(Sr) = 1 - P(Ab)$ denotes the probability of served calls. In [33] a four-dimensional waiting time performance measure for the abandonment model has been proposed, which is listed in Table 5.1.

$P(W_q \leq AWT, Sr)$	Fraction of well-served
$P(W_q > AWT, Sr)$	Fraction of served, with a potential for improvement
$P(W_q > \varepsilon, Ab)$	Fraction of poorly-served abandoned calls (ε is a very short time)
$P(W_q \leq \varepsilon, Ab)$	Fraction of those whose service-level is undetermined

Table 5.1: Four-dimensional waiting time performance measure

To find the four-dimensional performance measure listed in Table 5.1, we first consider $P(W_q > t, Sr)$. Using the density of V (5.17), we have

$$\begin{aligned}
P(W_q > t, Sr) &= P(V \wedge X > t, V < X) \\
&= P(V > t, V < X) \\
&= \int_t^\infty P(v < X) f_V(v) dv \\
&= p_S S\mu e^\eta \int_t^\infty e^{-\alpha v} e^{-(S\mu v + \eta e^{-\alpha v})} dv \\
&= p_S S\mu e^\eta \int_t^\infty e^{-\left[(S\mu+\alpha)v + \eta e^{-\alpha v}\right]} dv \\
&= p_S S\mu e^\eta \int_0^{\eta e^{-\alpha t}} e^{-\left[(S\mu+\alpha)(-\frac{1}{\alpha}\ln\frac{z}{\eta}) + z\right]} \frac{1}{z\alpha} dz \\
&= p_S C e^\eta \eta^{-C-1} \int_0^{\eta e^{-\alpha t}} z^C e^{-z} dz \\
&= p_S C e^\eta \eta^{-C-1} \gamma(C+1, \eta e^{-\alpha t}).
\end{aligned} \tag{5.20}$$

Using the fact that

$$\gamma(C+1, \eta e^{-\alpha t}) = C\gamma(C, \eta e^{-\alpha t}) - (\eta e^{-\alpha t})^C e^{-\eta e^{-\alpha t}},$$

we have that $P(W_q > t, Sr)$ can be factorized as

$$\begin{aligned} P(W_q > t, Sr) &= p_S C e^{\eta} \eta^{-C-1} \left[C\gamma(C, \eta e^{-\alpha t}) - (\eta e^{-\alpha t})^C e^{-\eta e^{-\alpha t}} \right] \\ &= p_S C e^{\eta} \eta^{-C} \gamma(C, \eta e^{-\alpha t}) e^{-\alpha t} \left[\eta^{-1} e^{\alpha t} C - \eta^{-1} e^{\alpha t} \frac{(\eta e^{-\alpha t})^C e^{-\eta e^{-\alpha t}}}{\gamma(C, \eta e^{-\alpha t})} \right] \\ &= P(W_q > t) \frac{e^{\alpha t}}{\rho} \left[1 - \frac{1}{A(C, \eta e^{-\alpha t})} \right] \end{aligned} \tag{5.21}$$

and $P(W_q > t, Ab)$ can be factorized as

$$\begin{aligned} P(W_q > t, Ab) &= P(W_q > t) - P(W_q > t, Sr) \\ &= P(W_q > t) \left[1 - \frac{e^{\alpha t}}{\rho} \left(1 - \frac{1}{A(C, \eta e^{-\alpha t})} \right) \right]. \end{aligned}$$

From the above, we have

$$\begin{aligned} P(Ab) &= P(W_q > 0, Ab) \\ &= P(W_q > 0)[1 - \frac{1}{\rho} \left(1 - \frac{1}{A} \right)] \\ &= \frac{[1 + A(\rho - 1)]B}{\rho[1 + (A-1)B]}, \end{aligned} \tag{5.22}$$

which is consistent with the result in [33]. Also

$$\begin{aligned} P(W_q > 0, Sr) &= P(W_q > 0) - P(Ab) \\ &= \frac{AB}{1 + (A-1)B} - \frac{[1 + A(\rho - 1)]B}{\rho[1 + (A-1)B]} \\ &= \frac{(A-1)B}{\rho[1 + (A-1)B]}, \end{aligned} \tag{5.23}$$

which is the same as (5.21) when $t = 0$. Therefore from (5.22) and (5.23) we have

$$P(Sr) = 1 - P(Ab) = \frac{\rho(1 - B) + (A-1)B}{\rho[1 + (A-1)B]} \tag{5.24}$$

and

$$\begin{aligned} P(W_q = 0, Sr) &= P(Sr) - P(W_q > 0, Sr) \\ &= \frac{\rho(1 - B) + (A-1)B}{\rho[1 + (A-1)B]} - \frac{(A-1)B}{\rho[1 + (A-1)B]} \\ &= \frac{1 - B}{1 + (A-1)B}, \end{aligned}$$

which can also be derived directly from (5.3) using $P(W_q = 0, Sr) = \sum_{i=0}^{S-1} p_i$. Finally the other two performance measures in Table 5.1 can be easily derived since

$$P(W_q \leq t, Ab) = P(Ab) - P(W_q > t, Ab)$$

and

$$P(W_q \le t, Sr) = P(Sr) - P(W_q > t, Sr).$$

We can also derive some conditional performance measures using the above results.

1. $P(W_q > t|Ab) = \frac{P(W_q>t,Ab)}{P(Ab)} = \frac{P(W_q>t,Ab)}{P(W_q>0,Ab)} = \frac{\eta e^{-\alpha t}\gamma(C,\eta e^{-\alpha t})-\gamma(C+1,\eta e^{-\alpha t})}{\eta\gamma(C,\eta)-\gamma(C+1,\eta)}$.

2. $P(W_q > t|Sr) = \frac{P(W_q>t,Sr)}{P(Sr)}$.

3. $P(W_q > t|delay, Sr) = \frac{P(W_q>t,Sr)}{P(delay,Sr)} = \frac{P(W_q>t,Sr)}{P(W_q>0,Sr)} = \frac{\gamma(C+1,\eta e^{-\alpha t})}{\gamma(C+1,\eta)}$.

4. $P(Ab|W_q > t) = \frac{P(W_q>t,Ab)}{P(W_q>t)} = 1 - \frac{e^{\alpha t}}{\rho}\left[1 - \frac{1}{A(C,\eta e^{-\alpha t})}\right]$.

5. $P(Sr|W_q > t) = \frac{P(W_q>t,Sr)}{P(W_q>t)} = \frac{\gamma(C+1,\eta e^{-\alpha t})}{\eta e^{-\alpha t}\gamma(C,\eta e^{-\alpha t})} = \frac{e^{\alpha t}}{\rho}\left[1 - \frac{1}{A(C,\eta e^{-\alpha t})}\right]$.

Note that in the above, $P(W_q > t|Ab)$ and $P(W_q > t|delay, Sr)$ also appeared in [39].

5.1.3 Mean waiting time

To find the various mean waiting times, we start from the mean delay of the delayed calls (**DLYDLY**). By (5.19), we have

$$\begin{aligned}
E(W_q|delay) &= \int_0^\infty P(W_q > t|delay)dt \\
&= \frac{1}{\gamma(C,\eta)}\int_0^\infty e^{-\alpha t}\gamma(C,\eta e^{-\alpha t})dt \\
&= \frac{1}{\alpha\gamma(C,\eta)}\int_0^\infty e^{-\alpha t}\left(\int_0^{\eta e^{-\alpha t}} z^{C-1}e^{-z}dz\right)d\alpha t \\
&= \frac{1}{\alpha\gamma(C,\eta)}\int_0^\infty e^{-u}\left(\int_0^{\eta e^{-u}} z^{C-1}e^{-z}dz\right)du \\
&= \frac{1}{\alpha\gamma(C,\eta)}\int_0^\eta \left(\int_0^{-\ln z/\eta} e^{-u}du\right)z^{C-1}e^{-z}dz \\
&= \frac{1}{\alpha\gamma(C,\eta)}\left[\gamma(C,\eta) - \frac{1}{\eta}\gamma(C+1,\eta)\right] \\
&= \frac{1}{\alpha}\left[1 - \frac{\gamma(C+1,\eta)}{\eta\gamma(C,\eta)}\right] \qquad (5.25) \\
&= \frac{1}{\alpha}\left[1 - \frac{1}{\rho}\left(1 - \frac{1}{A}\right)\right], \qquad (5.26)
\end{aligned}$$

where (5.25) also appeared in [39]. Hence we have the mean delay for all calls

$$\begin{aligned} E(W_q) &= E(W_q|delay)P(W_q > 0) \\ &= \frac{1}{\alpha}[1 - \frac{1}{\rho}\left(1 - \frac{1}{A}\right)]\frac{AB}{1 + (A-1)B} \\ &= \frac{[1 + A(\rho - 1)]B}{\alpha\rho[1 + (A-1)B]}, \end{aligned} \tag{5.27}$$

which can also be derived directly from (5.18). By comparing with (5.8), we have Little's formula for calls waiting in the queue: $E(Q_q) = \lambda E(W_q)$. Note that (5.26) also appeared in [33], where it was derived directly using Little's formula.

From (5.20), we have

$$\begin{aligned} E(W_q, Sr) &= \int_0^\infty P(W_q > t, Sr)dt \\ &= \int_0^\infty p_S C e^\eta \eta^{-C-1}\gamma(C+1, \eta e^{-\alpha t})dt \\ &= p_S C e^\eta \eta^{-C-1}\int_0^\infty \gamma(C+1, \eta e^{-\alpha t})dt \\ &= p_S C e^\eta \eta^{-C-1}\int_0^\infty \left(\int_0^{\eta e^{-\alpha t}} z^C e^{-z} dz\right) dt \\ &= p_S C e^\eta \eta^{-C-1}\frac{1}{\alpha}\left[\ln(\eta)\gamma(C+1,\eta) - \frac{d}{dC}\gamma(C+1,\eta)\right] \\ &= \frac{p_S}{\alpha} C e^\eta \eta^{-C-1}\gamma(C+1,\eta)\left[\ln(\eta) - \frac{\frac{d}{dC}\gamma(C+1,\eta)}{\gamma(C+1,\eta)}\right] \\ &= \frac{p_S(A-1)}{\alpha\rho}\left[\ln(\eta) - \frac{d}{dC}\ln\gamma(C+1,\eta)\right], \end{aligned}$$

where we have changed the order of integration to get the above result. Then

$$E(W_q, Ab) = E(W_q) - E(W_q, Sr).$$

Hence we have $E(W_q|Sr) = \frac{E(W_q,Sr)}{P(Sr)}$ and $E(W_q|Ab) = \frac{E(W_q,Ab)}{P(Ab)}$.

Since

$$P(W_q > t|delay, Sr) = \frac{\gamma(C+1, \eta e^{-\alpha t})}{\gamma(C+1,\eta)},$$

we have

$$\begin{aligned} E(W_q|delay, Sr) &= \int_0^\infty P(W_q > t|delay, Sr)dt \\ &= \frac{1}{\gamma(C+1,\eta)}\int_0^\infty \gamma(C+1, \eta e^{-\alpha t})dt \\ &= \frac{1}{\alpha}\left[\ln(\eta) - \frac{d}{dC}\ln\gamma(C+1,\eta)\right], \end{aligned}$$

which also appeared in [39]. Therefore

$$\begin{aligned} E(W_q, delay, Sr) &= P(W_q > 0, Sr)E(W_q|delay, Sr) \\ &= \frac{(A-1)B}{\rho[1+(A-1)B]}\frac{1}{\alpha}\left[\ln(\eta) - \frac{d}{dC}\ln\gamma(C+1,\eta)\right] \\ &= \frac{p_S(A-1)}{\alpha\rho}\left[\ln(\eta) - \frac{d}{dC}\ln\gamma(C+1,\eta)\right] \\ &= E(W_q, Sr), \end{aligned}$$

where we have used (5.23) and (5.5).

5.1.4 Probability of abandonment

This section is mainly based on the work of [33]. Since $M/M/S+M$ is an abandonment model, the probability of abandonment is an important performance measure. We already got $P(Ab)$ in (5.22) when considering the four-dimensional performance measure of waiting time. Usually for abandonment model, $P(Ab)$ can be derived in the following way as in [33]. We first condition on the state seen by a call and then sum up all the possibilities. To that end we define $P_i(Ab) = P(\text{the call will abandon} \mid i \text{ calls in the system upon arrival})$. The following result has been given in [33].

$$P_i(Ab) = P(V_i > X) = \begin{cases} 0 & 0 \le i < S \\ \frac{(i-S+1)\alpha}{S\mu+(i-S+1)\alpha} & i \ge S \end{cases} = \frac{r_{i+1}}{S\mu + r_{i+1}}. \tag{5.28}$$

Remark 5.1.6 *The result $P_i(Ab) = \frac{r_{i+1}}{S\mu+r_{i+1}}$ also holds for general abandonment model $M/M/S/N+G$ as proved in Movaghar [36].*

In the same way as we get the distribution of V from V_i, we have, by (5.9),

$$\begin{aligned} P(Ab) = P(V > X) &= \sum_{i=S}^{\infty} p_i P_i(Ab) \\ &= \sum_{i=S}^{\infty} p_i \frac{(i-S+1)\alpha}{S\mu + (i-S+1)\alpha} \\ &= p_S \sum_{i=S}^{\infty} \frac{\eta^{i-S}}{\prod_{k=1}^{i-S}(C+k)} \frac{(i-S+1)\alpha}{S\mu + (i-S+1)\alpha} \\ &= p_S \sum_{j=0}^{\infty} \frac{\eta^{j}}{\prod_{k=1}^{j}(C+k)} \frac{j+1}{C+j+1} \\ &= p_S \frac{d}{d\eta}(A-1) = p_S \frac{1+A(\rho-1)}{\rho} \\ &= \frac{[1+A(\rho-1)]B}{\rho[1+(A-1)B]}, \end{aligned} \tag{5.29}$$

which is the same as (5.22).

From (5.29) and by using the cut balance equation between state i and $i+1$

$$\lambda p_i = [S\mu + (i-S+1)\alpha] p_{i+1}, for\ i \geq S,$$

we have

$$P(Ab) = \sum_{i=S}^{\infty} p_i \frac{(i-S+1)\alpha}{S\mu + (i-S+1)\alpha} = \sum_{i=S}^{\infty} \frac{(i-S+1)\alpha p_{i+1}}{\lambda} = \frac{\alpha \cdot E(Q_q)}{\lambda}.$$

Hence we obtain the following rate balance equation

$$\lambda P(Ab) = \alpha \cdot E(Q_q), \tag{5.30}$$

which shows the stationary balance between the rate that calls abandon the queue and the rate that abandoned calls (i.e., calls who eventually abandon) enter the system. The above can also be proved by referring to (5.8) and (5.22).

Applying Little's formula to Q_b, we get $E(Q_b) = \lambda P(Sr)\frac{1}{\mu}$. Hence

$$P(Sr) = \frac{E(Q_b)}{a} \tag{5.31}$$

and by (5.7), we have the same expression of $P(Sr)$ as (5.24). In general, since $1 = P(Ab) + P(Sr)$, and from (5.30) and (5.31), we have the rate conservative equation in equilibrium

$$\lambda = \alpha \cdot E(Q_q) + \mu E(Q_b),$$

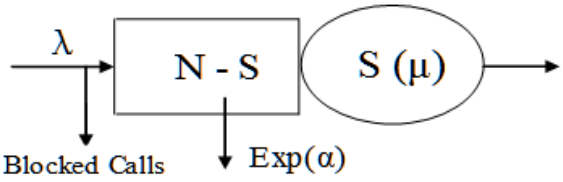

Figure 5.3: $M/M/S/N+M$ model description and parameters

which shows the rate of incoming calls equals to the sum of abandon and service rates in equilibrium.

Comparing (5.22) with (5.27), we have found a remarkable property of Erlang-A model:

$$P(Ab) = \alpha \cdot E(W_q),$$

which can be easily proved by using (5.30) and Little's formula to Q_q, $E(Q_q) = \lambda E(W_q)$. This property is also proved in [33]. In addition they demonstrated this property using real data from a call centre.

As before, $a = \lambda/\mu$ is called the offered load and here since the model is a loss model, $a' = E(Q_b) = aP(Sr)$ is called the carried load. The utilization

$$v = \frac{a'}{S} = \frac{aP(Sr)}{S} = \rho P(Sr) < 1$$

is the proportion of time that a CSR is busy.

5.2 $M/M/S/N+M$ model

$M/M/S/N+M$, a finite version of $M/M/S+M$ model, is more general and allows delay, blocking and abandonment. The model is the same as $M/M/S+M$ except that it has a finite number of buffers $K = N-S$. An arriving call finding all servers busy and all buffers occupied will be blocked. The model description and parameters are shown in Figure 5.3.

5.2.1 Queue length process

As in $M/M/S+M$ model, the patience times are assumed to be i.i.d. exponentially distributed and we have the similar abandonment rate

$$r_i = \begin{cases} 0 & \text{if } 0 \le i \le S \\ (i-S)\alpha & \text{if } S < i \le N \end{cases}. \tag{5.32}$$

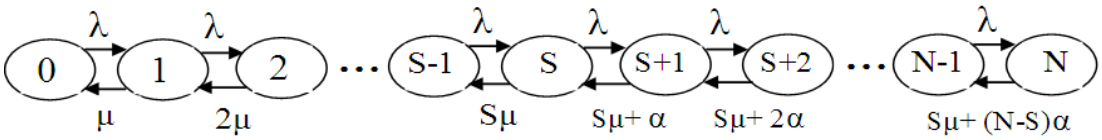

Figure 5.4: $M/M/S/N + M$ model stationary state transition diagram

The queue length process $Q(t)$ is a finite birth-death process with birth rate $\lambda_i = \lambda$ and state-dependent death rate

$$\mu_i = \begin{cases} i\mu & \text{if } 0 \leq i \leq S \\ S\mu + (i-S)\alpha & \text{if } S < i \leq N \end{cases}.$$

The stationary state transition diagram of $Q(t)$ is shown in Figure 5.4. We can conclude from the diagram that if $N = \infty$, it becomes $M/M/S + M$; if $\alpha = 0$ (no abandonment) it becomes $M/M/S/N$; if $\alpha = \infty$ it becomes $M/M/S/S$ and if $\alpha = \mu$ it becomes $M/M/N/N$. Since this is a finite state model, we can always obtain the stationary distribution of $Q(t)$ (no stability condition) by solving the global or cut balance equations derived from Figure 5.4. The solution is

$$p_i = \begin{cases} \frac{a^i}{i!} p_0 & \text{if } 0 \leq i \leq S \\ \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i} [S\mu + (j-S)\alpha]} \frac{a^S}{S!} p_0 & \text{if } S < i \leq N \end{cases}$$

where p_0 is

$$p_0 = \left(\sum_{i=0}^{S} \frac{a^i}{i!} + \frac{a^S}{S!} \sum_{i=S+1}^{N} \frac{\lambda^{i-S}}{\prod_{j=S+1}^{i} [S\mu + (j-S)\alpha]} \right)^{-1}.$$

Performance measures in terms of special functions

The previous results are standard. For example refer to Stolletz [44]. In the following we will define a new function $D(C, \eta, N)$ to express the performance measures. First p_i can also be expressed in terms of p_S

$$p_i = \begin{cases} \frac{S!}{i! a^{S-i}} p_S & \text{if } 0 \leq i \leq S \\ \rho_{i-S} p_S & \text{if } S < i \leq N \end{cases} \tag{5.33}$$

where p_S is

$$p_S = \frac{a^S}{S!}p_0 = \frac{B}{1 + B\sum_{i=S+1}^{N}\rho_{i-S}}$$

and $\rho_i = \frac{\eta^i}{\prod_{j=1}^{i}(C+j)} = \frac{\Gamma(C+1)}{\Gamma(C+i+1)}\eta^i$ as defined before. Since the sum in p_S is finite, we cannot use the special function $A(x, y)$ to express it, as in $M/M/S + M$ model. However we will define

$$\begin{aligned} D(N) := D(C, \eta, N) &= \sum_{i=0}^{N-1}\rho_i = \sum_{i=0}^{N-1}\frac{\Gamma(C+1)}{\Gamma(C+i+1)}\eta^i \\ &= \Gamma(C+1)e^{\eta}\eta^{-C}[P(C,\eta) - P(C+N,\eta)] \\ &= A - \Gamma(C+1)e^{\eta}\eta^{-C}P(C+N,\eta), \end{aligned}$$

where $P(C, \eta) = \frac{\gamma(C,\eta)}{\Gamma(C)}$ is the regularized Gamma function and the above equality can be proved by a property of $P(C, \eta)$,

$$P(C+1,\eta) = P(C,\eta) - \frac{e^{-\eta}\eta^C}{\Gamma(C+1)}.$$

Now p_S can be written in terms of B and $D(N)$

$$p_S = \frac{B}{1 + B[D(N-S+1) - 1]}.$$

In particular, by the PASTA property, we have the blocking probability

$$\begin{aligned} P(blocking) = p_N &= p_S\rho_{N-S} \\ &= \frac{B\rho_{N-S}}{1 + B[D(N-S+1) - 1]} \\ &= \frac{B[D(N-S+1) - D(N-S)]}{1 + B[D(N-S+1) - 1]}, \end{aligned} \tag{5.34}$$

and the probability of delay among all calls (abandoned, served or blocked)

$$\begin{aligned} P(delay) &= \sum_{i=S}^{N-1} p_i = p_S\sum_{i=S}^{N-1}\rho_{i-S} = p_S\sum_{i=0}^{N-S-1}\rho_i \\ &= \frac{BD(N-S)}{1 + B[D(N-S+1) - 1]}. \end{aligned} \tag{5.35}$$

We can also derive three kinds of mean number of calls.

1. Mean number of busy servers $E(Q_b)$

$$\begin{aligned}
E(Q_b) &= \sum_{i=1}^{S} i p_i + S \sum_{i=S+1}^{N} p_i \\
&= p_S \sum_{i=1}^{S} i \frac{S!}{i! a^{S-i}} + S \left(1 - \sum_{i=0}^{S} \frac{S!}{i! a^{S-i}} p_S \right) \\
&= p_S a \left(\frac{1}{B} - 1 \right) + S \left(1 - \frac{1}{B} p_S \right) \\
&= S + p_S a \left(\frac{1}{B} - 1 \right) - \frac{S}{B} p_S \\
&= S + \frac{B}{1 + B \sum_{i=1}^{N-S} \rho_i} \frac{a(1-B) - S}{B} \\
&= S + \frac{a(1-B) - S}{1 + B \sum_{i=1}^{N-S} \rho_i} \\
&= S + \frac{a(1-B) - S}{1 + B[D(N-S+1) - 1]} \\
&= \frac{a(1-B) + SB[D(N-S+1) - 1]}{1 + B[D(N-S+1) - 1]}.
\end{aligned} \tag{5.36}$$

2. Mean number of calls waiting in the queue $E(Q_q)$

$$\begin{aligned}
E(Q_q) &= \sum_{i=S+1}^{N} (i - S) p_i = p_S \sum_{i=1}^{N-S} i \rho_i \\
&= p_S \sum_{i=1}^{N-S} i \frac{\eta^i}{\prod_{j=1}^{i} (C + j)} \\
&= p_S \sum_{i=1}^{N-S} \left[\frac{1}{\prod_{j=1}^{i-1} (C + j)} - \frac{C}{\prod_{j=1}^{i} (C + j)} \right] \eta^i \\
&= p_S \left[\sum_{i=1}^{N-S} \frac{\eta^i}{\prod_{j=1}^{i-1} (C + j)} - \sum_{i=1}^{N-S} \frac{C \eta^i}{\prod_{j=1}^{i} (C + j)} \right] \\
&= \eta p_S \left[\sum_{i=0}^{N-S-1} \rho_i - \frac{1}{\rho} \sum_{i=1}^{N-S} \rho_i \right] \\
&= \frac{\eta B \left\{ D(N-S) - \frac{1}{\rho} [D(N-S+1) - 1] \right\}}{1 + B[D(N-S+1) - 1]} \\
&= \frac{\eta B D(N-S) - CB[D(N-S+1) - 1]}{1 + B[D(N-S+1) - 1]},
\end{aligned} \tag{5.37}$$

which can also be derived using the similar method as in $M/M/S + M$ model by the fact that

$$\frac{d}{d\eta} D(N+1) = D(N) - \frac{D(N+1) - 1}{\rho}.$$

3. Mean number of calls in the system $E(Q)$

$$E(Q) = E(Q_b) + E(Q_q).$$

5.2.2 Probability of abandonment

Since there are some blocking calls and only non-blocking calls can abandon, we have $P(Ab) = P(Ab, \text{non-blocking})$. As in $M/M/S + M$ model we define $P_i(Ab) = P_i(Ab, \text{non-blocking}) = P(\text{the non-blocking call will abandon} \mid i \text{ calls in the system upon arrival})$. Then for $0 \le i < N$, similar as (5.28), we still have

$$P_i(Ab) = \frac{r_{i+1}}{S\mu + r_{i+1}} = \begin{cases} 0 & 0 \le i < S \\ \frac{(i-S+1)\alpha}{S\mu+(i-S+1)\alpha} & S \le i < N \end{cases} \tag{5.38}$$

and

$$P_i(Sr) = P_i(Sr, \text{non-blocking}) = \begin{cases} 1 & 0 \le i < S \\ \frac{S\mu}{S\mu+(i-S+1)\alpha} & S \le i < N \end{cases}.$$

Now

$$\begin{aligned} P(Ab) = P(Ab, \text{non-blocking}) &= \sum_{i=S}^{N-1} P_i(Ab) P(i \text{ calls in the system upon arrival}) \\ &= \sum_{i=S}^{N-1} \frac{(i-S+1)\alpha}{S\mu + (i-S+1)\alpha} a_i = \sum_{i=S}^{N-1} \frac{(i-S+1)\alpha}{S\mu + (i-S+1)\alpha} p_i \end{aligned} \tag{5.39}$$

where we have used the PASTA property, i.e., $a_i = p_i$. Similarly,

$$\begin{aligned} P(Sr) &= P(Sr, \text{non-blocking}) \\ &= \sum_{i=0}^{S-1} p_i + \sum_{i=S}^{N-1} \frac{S\mu}{S\mu + (i-S+1)\alpha} p_i. \end{aligned} \tag{5.40}$$

In the following we will derive expressions in terms of functions B and $D(N)$. By using the cut balance equation between state i and $i+1$

$$\lambda p_i = [S\mu + (i-S+1)\alpha] p_{i+1}, \quad \text{for } S \le i < N,$$

we have, from (5.39),

$$P(Ab) = \sum_{i=S}^{N-1} \frac{(i-S+1)\alpha}{S\mu + (i-S+1)\alpha} p_i = \sum_{i=S}^{N-1} \frac{(i-S+1)\alpha p_{i+1}}{\lambda} = \frac{\alpha \cdot E(Q_q)}{\lambda}.$$

Hence we obtain the same rate balance equation as (5.30)

$$\lambda P(Ab) = \alpha \cdot E(Q_q). \tag{5.41}$$

Now by (5.37) and the above, we have

$$P(Ab) = \frac{E(Q_q)}{\eta} = p_S \left[\sum_{i=0}^{N-S-1} \rho_i - \frac{1}{\rho} \sum_{i=1}^{N-S} \rho_i \right] \tag{5.42}$$

or the expression in terms of $D(N-S+1)$,

$$P(Ab) = \frac{BD(N-S) - \frac{1}{\rho}B[D(N-S+1)-1]}{1+B[D(N-S+1)-1]}.$$

Applying Little's formula to Q_b, we get $E(Q_b) = \lambda P(Sr)\frac{1}{\mu}$, which, combining (5.36), gives us

$$\begin{aligned} P(Sr) = \frac{E(Q_b)}{a} &= \frac{1}{\rho} + \frac{(1-B) - \frac{1}{\rho}}{1 + B\sum_{i=1}^{N-S}\rho_i} \\ &= \frac{1 - B + \frac{1}{\rho}B[D(N-S+1)-1]}{1+B[D(N-S+1)-1]}. \end{aligned} \tag{5.43}$$

In general, since $1 - p_N = P(Ab) + P(Sr)$, and from (5.42) and (5.43), we have the rate conservative equation in equilibrium

$$\lambda(1-p_N) = \alpha \cdot E(Q_q) + \mu E(Q_b),$$

which shows the rate of non-blocking calls equals to the sum of abandon and service rates in equilibrium. Again we have the carried load $a' = E(Q_b) = aP(Sr)$ and the utilization

$$v = \frac{a'}{S} = \frac{aP(Sr)}{S} = \rho P(Sr) < 1,$$

which is the proportion of time that a CSR is busy.

5.2.3 Waiting time distribution

As in $M/M/S/N$ model, we define $\overline{W}_q$ as the stationary waiting time in the queue until abandonment or starting to get service for all calls (the blocked calls have ∞ waiting time), then $\overline{W}_q$ has a mass at ∞ and $P(\overline{W}_q = \infty) = P(blocking)$. Also $\overline{W}_q$ has a mass at 0 and $P(\overline{W}_q = 0) = P(\text{no-delay})$. We have

$$\begin{aligned} P(\overline{W}_q > t) &= P(\overline{W}_q > t, \text{non-blocking}) + P(\overline{W}_q > t, blocking) \\ &= P(\overline{W}_q > t, \text{non-blocking}) + P(blocking). \end{aligned}$$

To find $P(\overline{W}_q > t, \text{non-blocking})$, for $0 \le i < N$, we define W_{qi} as the waiting time of a non-blocking call given that it finds i in the system, i.e.,

$$P(W_{qi} > t) = P(\overline{W}_q > t, \text{non-blocking} \mid i \text{ calls in the system upon arrival}).$$

Then by (5.12), we have for $S \le i < N$,

$$P(W_{qi} > t) = P(V_i \wedge X > t) \tag{5.44}$$
$$= P(V_i > t)P(X > t) = e^{-\alpha t}P(V_i > t) = \sum_{n=0}^{i-S} A_{n,i-S}e^{-[S\mu+(n+1)\alpha]t}$$

and $P(W_{qi} > t) = 0$ for $0 \le i < S$. Therefore,

$$P(\overline{W}_q > t, \text{non-blocking}) = \sum_{i=S}^{N-1} P(W_{qi} > t)P(i \text{ calls in the system upon arrival})$$
$$= \sum_{i=S}^{N-1}\sum_{n=0}^{i-S} A_{n,i-S}e^{-[S\mu+(n+1)\alpha]t}a_i = \sum_{i=S}^{N-1}\sum_{n=0}^{i-S} A_{n,i-S}e^{-[S\mu+(n+1)\alpha]t}p_i,$$

where we have used the PASTA property, i.e., $a_i = p_i$.

Usually we are more concerned with the conditional waiting time of a call until abandonment or starting to get service given that this call is not blocked. Let W_q be this waiting time, which has no mass at ∞. We have

$$P(W_q{>}t) = P(\overline{W}_q > t|\text{non-blocking}) = \frac{P(\overline{W}_q > t, \text{non-blocking})}{P(\text{non-blocking})}$$
$$= \sum_{i=S}^{N-1}\sum_{n=0}^{i-S} A_{n,i-S}e^{-[S\mu+(n+1)\alpha]t}\frac{p_i}{1-p_N} = \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S}e^{-[S\mu+(n+1)\alpha]t}, \tag{5.45}$$

where $q_i := \frac{p_i}{1-p_N}, 0 \le i \le N-1$ is the probability of a call finding i calls in the system given that it is not blocked, which is the arrival-point probability in Chapter 2.

In addition, the following performance measures can be easily obtained in terms of q_i.

1. $P(\text{no-delay}|\text{non-blocking}) = P(W_q = 0) = \sum_{i=0}^{S-1} q_i$.
2. $P(delay|\text{non-blocking}) = P(W_q > 0) = \sum_{i=S}^{N-1} q_i$.

Following the same idea as the last section, to facilitate the computation and analysis, we will express the above performance measures in terms of functions $B(S, a)$ and $D(N)$. We have, by (5.35) and (5.34),

$$P(delay|\text{non-blocking}) = P(W_q > 0) = \sum_{i=S}^{N-1} q_i$$
$$= \sum_{i=S}^{N-1}\frac{p_i}{1-p_N} = \frac{P(delay)}{1-P(blocking)}$$
$$= \frac{BD(N-S)}{1+B[D(N-S)-1]}. \tag{5.46}$$

and

$$P(\text{no-delay}|\text{non-blocking}) = 1 - P(delay|\text{non-blocking}) = \frac{1-B}{1+B[D(N-S)-1]}.$$

To find the waiting time distribution for served calls given non-blocking, i.e., $P(W_q{>}t, Sr)$, we first need to find $P(W_{qi}{>}t, Sr)$. We have for $S \le i < N$, by (5.12),

$$\begin{aligned} P(W_{qi} > t, Sr) &= P(V_i \wedge X > t, V_i < X) \\ &= P(V_i > t, V_i < X) \\ &= \int_t^\infty P(v < X) f_{V_i}(v) dv \\ &= \int_t^\infty e^{-\alpha v} f_{V_i}(v) dv \\ &= \sum_{n=0}^{i-S} A_{n,i-S}(S\mu + n\alpha) \int_t^\infty e^{-\alpha v} e^{-(S\mu+n\alpha)v} dv \\ &= \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}. \end{aligned} \tag{5.47}$$

and $P(W_{qi} > t, Sr) = 0$ for $0 \le i < S$. Then

$$P(W_q{>}t, Sr) = \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} \tag{5.48}$$

and using (5.45),

$$\begin{aligned} P(W_q{>}t, Ab) &= P(W_q{>}t) - P(W_q{>}t, Sr) \\ &= \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}. \end{aligned} \tag{5.49}$$

When $t = 0$ in (5.47), we obtain

$$P(W_{qi} > 0, Sr) = \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha}$$

so that

$$P_i(Ab) = P(W_{qi} > 0, Ab) = 1 - P(W_{qi} > 0, Sr) = \sum_{n=0}^{i-S} A_{n,i-S} \frac{\alpha}{S\mu + (n+1)\alpha}, \tag{5.50}$$

where we have used $\sum_{n=0}^{i-S} A_{n,i-S} = 1$. Comparing (5.50) and (5.38), we have an identity involving $A_{n,i-S}$:

$$\sum_{n=0}^{i-S} \frac{A_{n,i-S}}{S\mu + (n+1)\alpha} = \frac{i-S+1}{S\mu + (i-S+1)\alpha}. \tag{5.51}$$

From (5.50), we also have

$$P(W_q > 0, Ab) = P(Ab|\text{non-blocking}) = \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{\alpha}{S\mu + (n+1)\alpha} \tag{5.52}$$

and

$$P(Ab) = P(Ab|\text{non-blocking})P(\text{non-blocking}) = \sum_{i=S}^{N-1} p_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{\alpha}{S\mu + (n+1)\alpha}, \tag{5.53}$$

which is the same as (5.39) by referring to (5.51), but involving $A_{n,i-S}$.

Similarly when $t = 0$ in (5.48), using (5.51), we obtain,

$$\begin{aligned}
P(W_q>0, Sr) = P(\overline{W}_q > 0, Sr|\text{non-blocking}) &= \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} \\
&= \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \left[1 - \frac{\alpha}{S\mu + (n+1)\alpha}\right] \\
&= \sum_{i=S}^{N-1} q_i \left[1 - \sum_{n=0}^{i-S} A_{n,i-S} \frac{\alpha}{S\mu + (n+1)\alpha}\right] \\
&= \sum_{i=S}^{N-1} q_i \left[1 - \frac{(i-S+1)\alpha}{S\mu + (i-S+1)\alpha}\right] \\
&= \sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu + (i-S+1)\alpha},
\end{aligned}$$

where we have used $\sum_{n=0}^{i-S} A_{n,i-S} = 1$. Therefore

$$\begin{aligned}
P(Sr|\text{non-blocking}) &= P(\overline{W}_q > 0, Sr|\text{non-blocking}) + P(\overline{W}_q = 0, Sr|\text{non-blocking}) \\
&= \sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu + (i-S+1)\alpha} + \sum_{i=0}^{S-1} q_i,
\end{aligned}$$

which can also be obtained from (5.40).

Now we can derive the conditional waiting, which are more useful in practice, using the above results.

1.

$$\begin{aligned}
P(W_q > t|Ab) &= \frac{P(W_q > t, Ab)}{P(Ab|\text{non-blocking})} \\
&= \frac{\sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{1}{S\mu+(n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}}{\sum_{i=S}^{N-1} q_i \frac{i-S+1}{S\mu+(i-S+1)\alpha}}.
\end{aligned}$$

2.

$$P(W_q > t|Sr) = \frac{P(W_q > t, Sr)}{P(Sr|\text{non-blocking})} = \frac{\sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu+n\alpha}{S\mu+(n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}}{\sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu+(i-S+1)\alpha} + \sum_{i=0}^{S-1} q_i}.$$

3.

$$P(W_q > t|Sr, delay) = \frac{P(W_q > t, Sr)}{P(Sr, delay|\text{non-blocking})} = \frac{\sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu+n\alpha}{S\mu+(n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}}{\sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu+(i-S+1)\alpha}}.$$

5.2.4 Mean waiting time

By (5.45), we have the mean waiting time for all calls given entry

$$\begin{aligned} E(W_q) &= \int_0^\infty \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} e^{-[S\mu+(n+1)\alpha]t} dt \\ &= \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} \frac{A_{n,i-S}}{S\mu+(n+1)\alpha}, \end{aligned} \tag{5.54}$$

which involves $A_{n,i-S}$ and makes the computation harder. We can obtain an alternative expression by considering the mean waiting time of the non-blocking call given the call finds i calls upon arrival $E(W_{qi})$, which is 0 for $0 \le i < S$ and for $S \le i < N$,

$$\begin{aligned} E(W_{qi}) &= \int_0^\infty P(W_{qi} > t) dt \\ &= \int_0^\infty P(V_i > t) e^{-\alpha t} dt \\ &= \frac{1}{\alpha} P(V_i > X) = \frac{1}{\alpha} P_i(Ab) \\ &= \frac{i-S+1}{S\mu+(i-S+1)\alpha}. \end{aligned}$$

Hence we have a new expression not involving $A_{n,i-S}$ for $E(W_q)$,

$$E(W_q) = \sum_{i=S}^{N-1} q_i \frac{i-S+1}{S\mu+(i-S+1)\alpha}, \tag{5.55}$$

which can also be derived from (5.54) by using (5.51). Comparing (5.52) with (5.54), we have found a similar property as Erlang-A model:

$$P(Ab|\text{non-blocking}) = \alpha \cdot E(W_q)$$

or

$$P(Ab) = \alpha \cdot E(\overline{W}_q, \text{non-blocking})$$

which again can be easily proved by using (5.41) and Little's formula:

$$E(Q_q) = \lambda E(\overline{W}_q, \text{non-blocking}).$$

By (5.48), we have the mean waiting time for served calls given entry

$$\begin{aligned} E(W_q, Sr) &= \int_0^\infty \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} dt \\ &= \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} \frac{A_{n,i-S}(S\mu + n\alpha)}{[S\mu + (n+1)\alpha]^2}, \end{aligned} \tag{5.56}$$

which involves $A_{n,i-S}$. In the following we will obtain an alternative expression not involving $A_{n,i-S}$ for $E(W_q, Sr)$ using a similar method in [45]. Consider the mean waiting time for served calls given the non-blocking call finds i calls upon arrival $E(W_{qi}, Sr)$, which is 0 for $0 \le i < S$ and for $S \le i < N$,

$$\begin{aligned} &E(W_{qi}, Sr) \\ &= E(V_i, V_i < X) = E\left(\sum_{n=0}^{i-S} \phi_n, V_i < X\right) \\ &= \sum_{n=0}^{i-S} E(\phi_n, V_i < X) = \sum_{n=0}^{i-S} E(\phi_n I_{(V_i,\infty)}(X)) \\ &= \sum_{n=0}^{i-S} \int_0^\infty \cdots \int_0^\infty t_n \left(\int_0^\infty I_{(t_0+t_1+\cdots+t_{i-S},\infty)}(x) f_X(x) dx \right) f_{\phi_0}(t_0) f_{\phi_1}(t_1) \cdots f_{\phi_{i-S}}(t_{i-S}) dt_0 \cdots dt_{i-S} \\ &= \sum_{n=0}^{i-S} \int_0^\infty \cdots \int_0^\infty t_n e^{-\alpha(t_0+t_1+\cdots+t_{i-S})} f_{\phi_0}(t_0) f_{\phi_1}(t_1) \cdots f_{\phi_{i-S}}(t_{i-S}) dt_0 \cdots dt_{i-S} \\ &= \sum_{n=0}^{i-S} \int_0^\infty t_n e^{-\alpha t_n} f_{\phi_n}(t_n) dt_n \prod_{m\neq n, m=0}^{i-S} f^*_{\phi_m}(\alpha) \\ &= \sum_{n=0}^{i-S} \int_0^\infty t_n e^{-\alpha t_n} (S\mu + n\alpha) e^{-(S\mu+n\alpha)t_n} dt_n \prod_{m\neq n, m=0}^{i-S} \frac{S\mu + m\alpha}{S\mu + (m+1)\alpha} \\ &= \sum_{n=0}^{i-S} \frac{S\mu + n\alpha}{[S\mu + (n+1)\alpha]^2} \left[\frac{S\mu}{S\mu + n\alpha} \frac{S\mu + (n+1)\alpha}{S\mu + (i-S+1)\alpha} \right] \\ &= \frac{S\mu}{S\mu + (i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{1}{S\mu + (n+1)\alpha}, \end{aligned} \tag{5.57}$$

where $\phi_n \sim \exp(S\mu + n\alpha)$ with the Laplace transform denoted by $f^*_{\phi_n}$ and $I_A(x) = \begin{cases} 1 & if\ x \in A \\ 0 & otherwise \end{cases}$ is the indicator function. Hence we have

$$E(W_{qi}|Sr) = \frac{E(W_{qi}, Sr)}{P_i(Sr)} = \sum_{n=0}^{i-S} \frac{1}{S\mu + (n+1)\alpha}. \tag{5.58}$$

Now by the total probability law, we have

$$E(W_q, Sr) = \sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu + (i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{1}{S\mu + (n+1)\alpha}, \tag{5.59}$$

which is not involved $A_{n,i-S}$ any more.

By (5.47), we have an alternative expression for $E(W_{qi}, Sr)$,

$$E(W_{qi}, Sr) = \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{[S\mu + (n+1)\alpha]^2}. \tag{5.60}$$

Comparing (5.57) and (5.60), we have another identity involving $A_{n,i-S}$:

$$\sum_{n=0}^{i-S} \frac{A_{n,i-S}(S\mu + n\alpha)}{[S\mu + (n+1)\alpha]^2} = \frac{S\mu}{S\mu + (i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{1}{S\mu + (n+1)\alpha}. \tag{5.61}$$

Whitt [46] used a different method to derive $E(W_{qi}|Sr)$ (5.58) for $S \le i < N$. After the call joins the queue, he will be in the $i - S + 1$ position and will be in the $i - S$ position after a random time of $\phi_{i-S+1} \sim \exp(S\mu + (i - S + 1)\alpha)$, since we are under the condition that this call will not abandon. Generally, the time for the call to go from position $n + 1$ to position n is $\exp(S\mu + (n + 1)\alpha)$, for $0 \le n \le i - S$, where position 0 means the time point when the call leaves the queue and starts to get served. Hence,

$$E(W_{qi}|Sr) = \sum_{n=0}^{i-S} E(\phi_{n+1}) = \sum_{n=0}^{i-S} \frac{1}{S\mu + (n+1)\alpha},$$

which is the same as (5.58). This method can also be used to derive the survival function $P(W_{qi} > t, Sr)$ for $S \le i < N$,

$$\begin{aligned} P(W_{qi} > t|Sr) &= P(\sum_{n=0}^{i-S} \phi_{n+1} > t) \\ &= \sum_{n=0}^{i-S} B_{n,i-S} e^{-[S\mu+(n+1)\alpha]t} \\ &= \frac{S\mu + (i-S+1)\alpha}{S\mu} \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}, \end{aligned}$$

where

$$B_{n,i-S} := \prod_{\substack{k=0 \\ k\neq n}}^{i-S} \frac{S\mu + (k+1)\alpha}{(k-n)\alpha} = \frac{\prod_{k=0}^{i-S}[S\mu + (k+1)\alpha]}{\prod_{\substack{k=0 \\ k\neq n}}^{i-S}[(k-n)\alpha]\,[S\mu + (n+1)\alpha]}$$

$$= \frac{S\mu + (i-S+1)\alpha}{S\mu} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} \tag{5.62}$$

and we have used the argument in Cox [18] page 17 (described in the proof of Theorem 5.1.2) to derive the distribution of the sum of exponential variables with different means. Then

$$P(W_{qi} > t, Sr) = P(W_{qi} > t|Sr)P_i(Sr) = \sum_{n=0}^{i-S} A_{n,i-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t},$$

which is the same as (5.47). However Whitt [46] did not provide the above formula. He only gave the Laplace transform and then used numerical inversion to calculate the survival function for any t.

By (5.49), we have the mean waiting time for abandoned calls given entry,

$$E(W_q, Ab) = \int_0^\infty \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} A_{n,i-S} \frac{\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} dt$$

$$= \sum_{i=S}^{N-1} q_i \sum_{n=0}^{i-S} \frac{A_{n,i-S}\alpha}{[S\mu + (n+1)\alpha]^2}, \tag{5.63}$$

which also involves $A_{n,i-S}$. However we can have an expression without $A_{n,i-S}$ in the following,

$$\begin{aligned} E(W_q, Ab) &= E(W_q) - E(W_q, Sr) \\ &= \sum_{i=S}^{N-1} q_i \frac{i-S+1}{S\mu + (i-S+1)\alpha} - \sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu + (i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{1}{S\mu + (n+1)\alpha} \\ &= \sum_{i=S}^{N-1} q_i \frac{1}{S\mu + (i-S+1)\alpha} \sum_{n=0}^{i-S} \left[1 - \frac{S\mu}{S\mu + (n+1)\alpha}\right] \\ &= \sum_{i=S}^{N-1} q_i \frac{\alpha}{S\mu + (i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{n+1}{S\mu + (n+1)\alpha}, \end{aligned} \tag{5.64}$$

where we have used (5.55) and (5.59).

Now we can derive the conditional mean waiting time, which are more useful in practice, using the above results.

1.

$$E(W_q|Ab) = \frac{E(W_q, Ab)}{P(Ab|\text{non-blocking})} = \frac{\sum_{i=S}^{N-1} q_i \frac{1}{S\mu+(i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{n+1}{S\mu+(n+1)\alpha}}{\sum_{i=S}^{N-1} q_i \frac{i-S+1}{S\mu+(i-S+1)\alpha}}.$$

2.

$$E(W_q|Sr) = \frac{E(W_q, Sr)}{P(Sr|\text{non-blocking})} = \frac{\sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu+(i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{1}{S\mu+(n+1)\alpha}}{\sum_{i=S}^{N-1} q_i \frac{S\mu}{S\mu+(i-S+1)\alpha} + \sum_{i=0}^{S-1} q_i}.$$

3.

$$E(W_q|Sr, delay) = \frac{E(W_q, Sr)}{P(Sr, delay|\text{non-blocking})} = \frac{\sum_{i=S}^{N-1} q_i \frac{1}{S\mu+(i-S+1)\alpha} \sum_{n=0}^{i-S} \frac{1}{S\mu+(n+1)\alpha}}{\sum_{i=S}^{N-1} q_i \frac{1}{S\mu+(i-S+1)\alpha}}.$$

5.2.5 Response time

Whitt [46] considered mean response time for all calls, where the response time $\overline{W}$ is defined as 0 for abandonment and blocked calls. For other calls, the response time $\overline{W}$ is defined as the sojourn time (waiting time plus service time). Then the mean response time $E(\overline{W}) = E(\overline{W}, Sr)$ is derived first by considering the mean response time of the served call given the call finds i calls upon arrival $E(\overline{W}_i, Sr)$, which is $\frac{1}{\mu}$ for $0 \le i < S$ and for $S \le i < N$, by (5.58)

$$\begin{aligned} E(\overline{W}_i, Sr) &= E(\overline{W}_i|Sr)P_i(Sr) \\ &= [E(Y|Sr) + E(W_{qi}|Sr)]P_i(Sr) \\ &= \left[\frac{1}{\mu} + \sum_{n=0}^{i-S} \frac{1}{S\mu+(n+1)\alpha}\right] \frac{S\mu}{S\mu+(i-S+1)\alpha}, \end{aligned}$$

where $Y \sim \exp(\mu)$ is the service time. Therefore, by (5.40) and (5.59)

$$\begin{aligned} E(\overline{W}) &= \frac{1}{\mu}\sum_{i=0}^{S-1} p_i + \sum_{i=S}^{N-1} p_i \left[\frac{1}{\mu} + \sum_{n=0}^{i-S} \frac{1}{S\mu+(n+1)\alpha}\right] \frac{S\mu}{S\mu+(i-S+1)\alpha} \\ &= \frac{1}{\mu}P(Sr) + E(W_q, Sr)[1 - P(blocking)] \end{aligned}$$

and

$$E(\overline{W}|Sr) = \frac{E(\overline{W})}{P(Sr)} = \frac{1}{\mu} + E(W_q|Sr). \tag{5.65}$$

For $P(\overline{W} > t)$, Whitt [46] first gave the Laplace transform of $\overline{W}$: $f^*_{\overline{W}}(s)$ then $P(\overline{W} > t)$ for any t can be calculated by numerically inverting its Laplace transform $(1 - f^*_{\overline{W}}(s))/s$.

In the following, we will use this method to find the analytic expression of $P(\overline{W} > t)$ as we have done for $P(W_{qi} > t|Sr)$ before. For $0 \leq i < S$, $P(\overline{W}_i > t, Sr) = e^{-\mu t}$ and for $S \leq i < N$,

$$\begin{aligned} P(\overline{W}_i > t|Sr) &= P(Y + \sum_{n=1}^{i-S+1} \phi_n > t) \\ &= \sum_{n=0}^{i-S+1} C_{n,i-S+1} e^{-(S\mu+\delta_n)t} \end{aligned}$$

where $C_{n,i-S+1} := \prod_{\substack{k=0 \\ k\neq n}}^{i-S+1} \frac{S\mu+\delta_k}{\delta_k-\delta_n}$, $\delta_n = \begin{cases} (1-S)\mu & for\ n = 0 \\ n\alpha, & for\ 0 < n \leq i-S+1 \end{cases}$ and we have used the argument in Cox [18] page 17 to derive the distribution of the sum of exponential variables with different means. The above can be expressed in terms of $B_{n,i-S} =$

$\prod_{\substack{k=0\\k\neq n}}^{i-S} \frac{S\mu+(k+1)\alpha}{(k-n)\alpha}$ as shown in the following,

$$
\begin{aligned}
P(\overline{W}_i > t|Sr) &= C_{0,i-S+1}e^{-(S\mu+\delta_0)t} + \sum_{n=1}^{i-S+1} C_{n,i-S+1}e^{-(S\mu+\delta_n)t} \\
&= C_{0,i-S+1}e^{-\mu t} + \sum_{n=1}^{i-S+1} \frac{S\mu+\delta_0}{\delta_0-\delta_n} \prod_{\substack{k=1\\k\neq n}}^{i-S+1} \frac{S\mu+\delta_k}{\delta_k-\delta_n} e^{-(S\mu+\delta_n)t} \\
&= C_{0,i-S+1}e^{-\mu t} + \sum_{n=1}^{i-S+1} \frac{\mu}{(1-S)\mu-n\alpha} \prod_{\substack{k=1\\k\neq n}}^{i-S+1} \frac{S\mu+k\alpha}{(k-n)\alpha} e^{-(S\mu+n\alpha)t} \\
&= C_{0,i-S+1}e^{-\mu t} + \sum_{n=1}^{i-S+1} \frac{\mu}{(1-S)\mu-n\alpha} \frac{\prod_{k=1}^{i-S+1}(S\mu+k\alpha)}{\alpha^{i-S}(S\mu+n\alpha)} \prod_{\substack{k=1\\k\neq n}}^{i-S+1} \frac{1}{(k-n)} e^{-(S\mu+n\alpha)t} \\
&= C_{0,i-S+1}e^{-\mu t} + \sum_{n=1}^{i-S+1} \frac{\mu}{(1-S)\mu-n\alpha} \frac{\prod_{k=1}^{i-S+1}(S\mu+k\alpha)}{\alpha^{i-S}(S\mu+n\alpha)} \frac{(-1)^{n-1}}{(n-1)!(i-S+1-n)!} e^{-(S\mu+n\alpha)t} \\
&= C_{0,i-S+1}e^{-\mu t} + \sum_{n=0}^{i-S} \frac{\mu}{(1-S)\mu-(n+1)\alpha} \frac{\prod_{k=0}^{i-S}[S\mu+(k+1)\alpha]}{\alpha^{i-S}[S\mu+(n+1)\alpha]} \frac{(-1)^{n}}{n!(i-S-n)!} e^{-[S\mu+(n+1)\alpha]t} \\
&= C_{0,i-S+1}e^{-\mu t} + \mu\sum_{n=0}^{i-S} \frac{1}{(1-S)\mu-(n+1)\alpha} \prod_{\substack{k=0\\k\neq n}}^{i-S} \frac{S\mu+(k+1)\alpha}{(k-n)\alpha} e^{-[S\mu+(n+1)\alpha]t} \\
&= \prod_{k=1}^{i-S+1} \frac{S\mu+k\alpha}{k\alpha-(1-S)\mu} e^{-\mu t} + \mu\sum_{n=0}^{i-S} \frac{1}{(1-S)\mu-(n+1)\alpha} B_{n,i-S} e^{-[S\mu+(n+1)\alpha]t}.
\end{aligned}
$$

Now by using the relationship between $B_{n,i-S}$ and $A_{n,i-S}$ (5.62), we have for $S \leq i < N$,

$$
\begin{aligned}
&P(\overline{W}_i > t|Sr) \\
&\quad= \prod_{k=1}^{i-S+1} \frac{S\mu+k\alpha}{k\alpha-(1-S)\mu} e^{-\mu t} \\
&\quad+ \frac{S\mu+(i-S+1)\alpha}{S} \sum_{n=0}^{i-S} \frac{S\mu+n\alpha}{[(1-S)\mu-(n+1)\alpha][S\mu+(n+1)\alpha]} A_{n,i-S} e^{-[S\mu+(n+1)\alpha]t}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
P(\overline{W}_i > t, Sr) &= P(\overline{W}_i > t|Sr)P_i(Sr) \\
&= \frac{S\mu}{S\mu+(i-S+1)\alpha} \prod_{k=1}^{i-S+1} \frac{S\mu+k\alpha}{k\alpha-(1-S)\mu} e^{-\mu t} \\
&\quad+ \mu\sum_{n=0}^{i-S} \frac{S\mu+n\alpha}{[(1-S)\mu-(n+1)\alpha][S\mu+(n+1)\alpha]} A_{n,i-S} e^{-[S\mu+(n+1)\alpha]t}.
\end{aligned}
$$

Now we have

$$P(\overline{W} > t) = P(\overline{W} > t, Sr) = \sum_{i=0}^{S-1} p_i e^{-\mu t} + \sum_{i=S}^{N-1} p_i P(\overline{W}_i > t, Sr)$$

$$= \sum_{i=0}^{S-1} p_i e^{-\mu t} + \sum_{i=S}^{N-1} p_i \left(\begin{array}{c} \frac{S\mu}{S\mu+(i-S+1)\alpha} \prod_{k=1}^{i-S+1} \frac{S\mu+k\alpha}{k\alpha-(1-S)\mu} e^{-\mu t} \\ +\mu \sum_{n=0}^{i-S} \frac{S\mu+n\alpha}{[(1-S)\mu-(n+1)\alpha][S\mu+(n+1)\alpha]} A_{n,i-S} e^{-[S\mu+(n+1)\alpha]t} \end{array} \right).$$

5.2.6 A numerical approximation method

Garnett et al. [21] gave a general way to compute the performance measures of $M/M/S/N+M$ model numerically. They found many performance measures that are of interest can be expressed as expectations of simple functions of V and X, where V is the conditional offered waiting time of an infinite patient call given the call is not blocked so that V has no mass at ∞. A representative list is shown in the following table [21].

$f(V,X)$	$E[f(V,X)]$
$I_{(X,\infty)}(V)$	$P(Ab\|\text{non-blocking})$
$I_{(t,\infty)}(V \wedge X)$	$P(W_q > t)$
$I_{(t,\infty)}(V \wedge X) I_{(X,\infty)}(V)$	$P(W_q > t, Ab)$
$(V \wedge X) I_{(X,\infty)}(V)$	$E(W_q, Ab)$
$(V \wedge X) I_{(t,\infty)}(V \wedge X) I_{(X,\infty)}(V)$	$E(W_q, W_q > t, Ab)$
$g(V \wedge X)$	$E(g(W_q))$

Some other important performance measures may be expressed in terms of these performance measures. For example

$$P(Ab|W_q > t) = \frac{P(V \wedge X > t, V > X)}{P(V \wedge X > t)} = \frac{E[I_{(t,\infty)}(V \wedge X) I_{(X,\infty)}(V)]}{E[I_{(t,\infty)}(V \wedge X)]}.$$

To calculate $E[f(V,X)]$, Garnett et al. [21] considered the following decomposition

$$E[f(V,X)] = E[f(V,X) I_{(0,\infty)}(V)] + E[f(V,X) I_{\{0\}}(V)]$$

$$= E[f(V,X) I_{(0,\infty)}(V)] + E[f(0,X)] \sum_{i=0}^{S-1} q_i$$

and argue that for all functions f which seem of interest, $E[f(0,X)]$ evaluates to 0 or 1. Therefore only the first expression needs to be calculated so that the key is to find the density $f_V(t)$ for $t > 0$. We know from Theorem 5.1.3 that for $M/M/S+M$,

$$f_V(t) = \begin{cases} p_S S\mu e^{\eta} e^{-(S\mu t + \eta e^{-\alpha t})} & \text{if } t > 0 \\ \sum_{i=0}^{S-1} p_i & \text{if } t = 0 \end{cases}. \tag{5.66}$$

For $M/M/S/N+M$ model, the density function of V is given in [21] and we summarize the result in the following,

Theorem 5.2.1 *For $M/M/S/N+M$ model, the density function of V is,*

$$f_V(t) = \begin{cases} \frac{p_S S\mu}{1-p_N} e^{\eta} e^{-(S\mu t+\eta e^{-\alpha t})} \left[1 - P(N-S, \eta - \eta e^{-\alpha t})\right] & \text{if } t > 0 \\ \sum_{i=0}^{S-1} q_i & \text{if } t = 0 \end{cases}.$$

Proof. It is obvious that V has a mass of $\sum_{i=0}^{S-1} q_i$ at 0 and for $t > 0$ by using (5.33) and (5.15), we have

$$\begin{aligned} f_V(t) &= \frac{1}{1-p_N} \sum_{i=S}^{N-1} p_i f_{V_i}(t) = \frac{p_S}{1-p_N} \sum_{i=S}^{N-1} \frac{\eta^{i-S}}{\prod_{k=1}^{i-S}(C+k)} \frac{\alpha \prod_{k=0}^{i-S}(C+k)}{(i-S)!} (1-e^{-\alpha t})^{i-S} e^{-S\mu t} \\ &= \frac{p_S}{1-p_N} \sum_{i=S}^{N-1} \frac{\eta^{i-S} S\mu}{(i-S)!} (1-e^{-\alpha t})^{i-S} e^{-S\mu t} \\ &= \frac{p_S S\mu e^{-S\mu t}}{1-p_N} \sum_{j=0}^{N-S-1} \frac{[\eta(1-e^{-\alpha t})]^j}{j!} \qquad (5.67) \\ &= \frac{p_S S\mu}{1-p_N} e^{\eta} e^{-(S\mu t+\eta e^{-\alpha t})} \left[1 - P(N-S, \eta - \eta e^{-\alpha t})\right], \qquad (5.68) \end{aligned}$$

where we have used (2.21). ∎

Remark 5.2.1 *Another expression of $f_V(t)$ for $t > 0$ can be obtained by expanding (5.67)*

$$\begin{aligned} f_V(t) &= \frac{p_S S\mu e^{-S\mu t}}{1-p_N} \sum_{j=0}^{N-S-1} \frac{1}{j!} \left[\eta(1-e^{-\alpha t})\right]^j \\ &= \frac{p_S S\mu e^{-S\mu t}}{1-p_N} \sum_{j=0}^{N-S-1} \frac{\eta^j}{j!} \sum_{k=0}^{j} \binom{j}{k} (-1)^k e^{-\alpha k t}. \qquad (5.69) \end{aligned}$$

Now by using (5.69), (5.67) and (5.68), Garnett et al. [21] listed three methods to evaluate $E[f(V,X)]$.

1. $E[f(V,X)] = \frac{p_S S\mu}{1-p_N} \sum_{j=0}^{N-S-1} \frac{\eta^j}{j!} \sum_{k=0}^{j} \binom{j}{k} (-1)^k \int_0^\infty \left[\int_0^\infty f(v,x) \alpha e^{-\alpha x} dx\right] e^{-S\mu v} e^{-\alpha k v} dv$. The advantage is the integral is easy and the disadvantage is the alternating signs in the sum. This is actually the method we have used before in this chapter; if we expand $A_{n,i-S}$ in the expressions of performance measures obtained before, we will get the same expressions here. However we do obtain some performance measures which do not involve $A_{n,i-S}$ and hence easy to compute such as $P(Ab)$ and $E(W_q, Sr)$ etc.

2. $E[f(V,X)] = \frac{p_S S\mu}{1-p_N}\sum_{j=0}^{N-S-1}\frac{\eta^j}{j!}\int_0^\infty\left[\int_0^\infty f(v,x)\alpha e^{-\alpha x}dx\right]e^{-S\mu v}(1-e^{-\alpha v})^j dv$. This method removes the alternating signs in the sum and has only one sum. However the outer integral usually has to be solved numerically even though the inner integral can be solved analytically.

3. $E[f(V,X)] = \frac{p_S S\mu}{1-p_N}e^{\eta}\int_0^\infty\left[\int_0^\infty f(v,x)\alpha e^{-\alpha x}dx\right]e^{-(S\mu v+\eta e^{-\alpha v})}\left[1-P(N-S,\eta(1-e^{-\alpha v}))\right]dv$. This method has no sum at all and usually the inner integral can be solved analytically and the outer integral has to be solved numerically. For example

$$
\begin{aligned}
&P(Ab|\text{non-blocking})\\
&= \frac{p_S S\mu}{1-p_N}e^{\eta}\int_0^\infty\left[\int_0^\infty I_{(x,\infty)}(v)\alpha e^{-\alpha x}dx\right]e^{-(S\mu v+\eta e^{-\alpha v})}\left[1-P(N-S,\eta(1-e^{-\alpha v}))\right]dv\\
&= \frac{p_S S\mu}{1-p_N}e^{\eta}\int_0^\infty(1-e^{-\alpha v})e^{-(S\mu v+\eta e^{-\alpha v})}\left[1-P(N-S,\eta(1-e^{-\alpha v}))\right]dv\\
&= \sum_{i=S}^{N-1}q_i - \frac{p_S S\mu}{1-p_N}e^{\eta}\int_0^\infty e^{-\alpha v}e^{-(S\mu v+\eta e^{-\alpha v})}\left[1-P(N-S,\eta(1-e^{-\alpha v}))\right]dv,
\end{aligned}
$$

where we have used the fact that

$$\frac{p_S S\mu}{1-p_N}e^{\eta}\int_0^\infty e^{-(S\mu v+\eta e^{-\alpha v})}\left[1-P(N-S,\eta(1-e^{-\alpha v}))\right]dv = \sum_{i=S}^{N-1}q_i$$

since $f_V(t)$ in Theorem 5.2.1 is a density function.

For $M/M/S+M$ (5.68) reduces to (5.66) and we have

$$E[f(V,X)] = p_S S\mu e^{\eta}\int_0^\infty\left[\int_0^\infty f(v,x)\alpha e^{-\alpha x}dx\right]e^{-(S\mu v+\eta e^{-\alpha v})}dv,$$

which is usually integrable and the various performance measures of $M/M/S+M$ model have been derived using this method before in Section 5.1.

5.3 Monotonicity and concavity properties of performance measures for $M/M/S/N+M$ model

In this section, we will study the monotonicity and concavity properties of some performance measures for $M/M/S/N+M$ model. This type of study has a long history for many types of queueing models, for example [25]. Recently Jouini et al. [26] studied the monotonicity and concavity properties of $M/M/S/N+M$ model. Our method here is easier than the one used in [26] and we also have more results. The monotonicity properties are very important to the call centre design algorithm we will develop in Chapter 7.

5.3.1 Monotonicity properties with respect to buffer size K

We will first study the monotonicity properties of some performance measures with respect to the buffer size K, which is $N - S$ as defined in Chapter 1, when other parameters are fixed. Before giving the main results, we will prove a useful lower bound for $B(S, a)$ in the following, which is established by Sobel [41] and also appears in [25]. This result is also obtained in Chapter 2 and here we will use a direct method to prove it.

Lemma 5.3.1 *For Erlang B formula, we have $B(S, a) > 1 - 1/\rho$, where $\rho = \frac{a}{S}$.*

Proof. When $0 < \rho \leq 1$, it is obviously true. When $\rho > 1$, we want to prove

$$\frac{a^S/S!}{\sum_{i=0}^{S} a^i/i!} > 1 - \frac{S}{a}.$$

We have

$$\begin{aligned}
&\left(1 - \frac{S}{a}\right) \sum_{i=0}^{S} a^i/i! \\
&= \left(1 + a + \frac{a^2}{2!} + ... + \frac{a^S}{S!}\right) - \frac{S}{a}\left(1 + a + \frac{a^2}{2!} + ... + \frac{a^S}{S!}\right) \\
&= \frac{a^S}{S!} - \frac{S}{a} + \left(1 + a + \frac{a^2}{2!} + ... + \frac{a^{S-1}}{(S-1)!}\right) - \frac{S}{a}\left(a + \frac{a^2}{2!} + ... + \frac{a^S}{S!}\right) \\
&\leq \frac{a^S}{S!} - \frac{S}{a} + \left(1 + a + \frac{a^2}{2!} + ... + \frac{a^{S-1}}{(S-1)!}\right) - \left(\frac{1}{a}a + \frac{2}{a}\frac{a^2}{2!} + ... + \frac{S}{a}\frac{a^S}{S!}\right) \\
&= \frac{a^S}{S!} - \frac{S}{a} < \frac{a^S}{S!}.
\end{aligned}$$

■

Probability of served calls

We first consider the probability of served calls (5.43)

$$P^{(K)}(Sr) = \frac{1}{\rho} + \frac{1 - B - 1/\rho}{1 + B\sum_{i=1}^{K} \rho_i},$$

where as in the following, we use upper index K to denote that the performance measure is a function of K. We have the following result.

Theorem 5.3.1 *For $M/M/S/N + M$ model, $P^{(K)}(Sr)$ is strictly increasing in the buffer size K. $P^{(K)}(Sr)$ approaches to $P(Sr)$ of $M/M/S+M$ model while K approaches to infinity and it approaches to $P(Sr) = 1 - B$ of $M/M/S/S$ model while K approaches to* 0.

Proof. To show that $P^{(K)}(Sr)$ is strictly increasing in the buffer size K, it suffices to show that for $K \geq 0, P^{(K+1)}(Sr) - P^{(K)}(Sr) > 0$. We have

$$P^{(K+1)}(Sr) - P^{(K)}(Sr) = [(1-B) - 1/\rho] \left[\frac{1}{1 + B\sum_{i=1}^{K+1} \rho_i} - \frac{1}{1 + B\sum_{i=1}^{K} \rho_i} \right] > 0,$$

since $(1-B) - 1/\rho < 0$ by Lemma 5.3.1 and $1 + B\sum_{i=1}^{K+1} \rho_i > 1 + B\sum_{i=1}^{K} \rho_i$.

While K approaches to infinity, $\sum_{i=1}^{\infty} \rho_i = A - 1$. So we have

$$\begin{aligned} P^{(\infty)}(Sr) &= \frac{1}{\rho} + \frac{(1-B) - 1/\rho}{1 + B\sum_{i=1}^{\infty} \rho_i} \\ &= \frac{1}{\rho} + \frac{(1-B) - 1/\rho}{1 + [A-1]B} \\ &= \frac{(1-B) + (A-1)B/\rho}{1 + [A-1]B}, \end{aligned}$$

which is $P(Sr)$ of $M/M/S+M$ model. While K approaches to 0, $\sum_{i=1}^{0} \rho_i = 0$. So we have $P^{(0)}(Sr) = 1 - B$. ∎

Remark 5.3.1 *Since we have $E(Q_b) = aP(Sr)$ (5.43), the above monotonicity result also applies to $E(Q_b)$, the expected number of busy servers in equilibrium, or the carried load.*

Remark 5.3.2 *In [26], the authors gave the same result but their proof is much more complex than ours. The reason is that they express $P^{(K)}(Sr)$ in terms of p_i for $i > S$ (refer to formula (3.2) in [26]). However $P^{(K)}(Sr)$ can be written in terms of p_i for $i \leq S$ only since $E(Q_b) = aP^{(K)}(Sr)$ and $E(Q_b) = \sum_{i=1}^{S} ip_i + S(1 - \sum_{i=0}^{S} p_i)$.*

Probability of blocking

We next consider the probability of blocking (5.34),

$$P^{(K)}(blocking) = \frac{B\rho_K}{1 + B\sum_{i=1}^{K} \rho_i}.$$

We have the following result.

Theorem 5.3.2 *For $M/M/S/N+M$ model, $P^{(K)}(blocking)$ is strictly decreasing in the buffer size K. $P^{(K)}(blocking)$ approaches to 0 while K approaches to infinity and it approaches to $P(blocking) = B$ of $M/M/S/S$ model while K approaches to 0.*

Proof. Since $P^{(K)}(blocking) > 0$, to show that $P^{(K)}(blocking)$ is strictly decreasing in the buffer size K, it suffices to show that for $K \geq 0$, $\frac{P^{(K+1)}(blocking)}{P^{(K)}(blocking)} < 1$. We have

$$\begin{aligned}\frac{P^{(K+1)}(blocking)}{P^{(K)}(blocking)} &= \frac{\rho_{K+1}}{\rho_K}\frac{1+B\sum_{i=1}^{K}\rho_i}{1+B\sum_{i=1}^{K+1}\rho_i} \\ &= \frac{\eta}{C+K+1}\frac{1+B\sum_{i=1}^{K}\rho_i}{1+B\sum_{i=1}^{K}\rho_i+B\rho_{K+1}} \\ &= \frac{\eta}{C+K+1}\frac{1+B\sum_{i=1}^{K}\rho_i}{1+B\sum_{i=1}^{K}\rho_i+B\frac{\eta\rho_K}{C+K+1}} < 1\end{aligned} \tag{5.70}$$

is equivalent to the inequality

$$\eta B\rho_K > [\eta - (C+K+1)]\left(1+B\sum_{i=1}^{K}\rho_i\right). \tag{5.71}$$

We will prove (5.71) in the following for $K \geq 0$ by induction. When $K = 0$, we need to prove $\eta B > \eta - (C+1)$, which is true since $\eta B > \eta - C$ by Lemma 5.3.1. Assuming that (5.71) is true for $K \geq 0$, we will prove

$$\eta B\rho_{K+1} = \frac{\eta^2 B\rho_K}{C+K+1} > [\eta - (C+K+2)]\left(1+B\sum_{i=1}^{K}\rho_i + B\rho_{K+1}\right).$$

To simplify the expressions, we introduce the following notations: $\Delta_1 := C+K+1 > 0$, $\Delta_2 := \eta - (C+K+2)$ with $\Delta_1 + \Delta_2 + 1 = \eta$. Now noting that $\rho_{K+1} = \frac{\eta}{\Delta_1}\rho_K$, the above inequality becomes

$$\eta^2 B\rho_K > \Delta_1\Delta_2 + \Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i + \Delta_2 B\eta\rho_K,$$

which is equivalent to

$$\eta^2 B\rho_K - \Delta_2 B\eta\rho_K = (\Delta_1+1)\eta B\rho_K > \Delta_1\Delta_2 + \Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i.$$

Using the assumption, we have

$$\begin{aligned}(\Delta_1+1)\eta B\rho_K &> (\Delta_1+1)(\Delta_2+1)\left(1+B\sum_{i=1}^{K}\rho_i\right)\\ &= (\Delta_1\Delta_2+\Delta_1+\Delta_2+1)\left(1+B\sum_{i=1}^{K}\rho_i\right)\\ &= (\Delta_1\Delta_2+\eta)\left(1+B\sum_{i=1}^{K}\rho_i\right)\\ &= \Delta_1\Delta_2+\Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i+\eta\left(1+B\sum_{i=1}^{K}\rho_i\right)\\ &> \Delta_1\Delta_2+\Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i,\end{aligned}$$

which is what we want to prove.

Since $\sum_{i=1}^{\infty}\rho_i = A-1 < \infty$, we have $\lim_{K\to\infty}\rho_K = 0$. Hence while K approaches to infinity, $P^{(K)}(blocking)\to 0$. On the other hand, while K approaches to 0, we have $\rho_0 = 1$, $\sum_{i=1}^{0}\rho_i = 0$. Therefore $P^{(0)}(blocking) = B$. ∎

Probability of abandonment

Now we consider the probability of abandonment (5.42),

$$P^{(K)}(Ab) = \frac{B\left(\sum_{i=0}^{K-1}\rho_i - \frac{1}{\rho}\sum_{i=1}^{K}\rho_i\right)}{1+B\sum_{i=1}^{K}\rho_i}.$$

Theorem 5.3.3 *For $M/M/S/N+M$ model, $P^{(K)}(Ab)$ is strictly increasing in the buffer size K. $P^{(K)}(Ab)$ approaches to 0 while K approaches to 0 and it approaches to $P(Ab)$ of $M/M/S+M$ model while K approaches to infinity.*

Proof. To show that $P^{(K)}(Ab)$ is strictly increasing in the buffer size K, it suffices to show that for $K\geq 0, P^{(K+1)}(Ab) - P^{(K)}(Ab) > 0$. We have

$$\begin{aligned}&P^{(K+1)}(Ab) - P^{(K)}(Ab)\\ &= \frac{B\left(\sum_{i=0}^{K}\rho_i - \frac{1}{\rho}\sum_{i=1}^{K+1}\rho_i\right)}{1+B\sum_{i=1}^{K+1}\rho_i} - \frac{B\left(\sum_{i=0}^{K-1}\rho_i - \frac{1}{\rho}\sum_{i=1}^{K}\rho_i\right)}{1+B\sum_{i=1}^{K}\rho_i}\\ &= \frac{B\left[\left(1+B\sum_{i=1}^{K}\rho_i\right)\left(\sum_{i=0}^{K}\rho_i - \frac{1}{\rho}\sum_{i=1}^{K+1}\rho_i\right) - \left(1+B\sum_{i=1}^{K+1}\rho_i\right)\left(\sum_{i=0}^{K-1}\rho_i - \frac{1}{\rho}\sum_{i=1}^{K}\rho_i\right)\right]}{\left(1+B\sum_{i=1}^{K+1}\rho_i\right)\left(1+B\sum_{i=1}^{K}\rho_i\right)}.\end{aligned}$$

Hence we need to prove that for $K \geq 0$,

$$\begin{aligned}
&\left(1+B\sum_{i=1}^{K}\rho_i\right)\left(\sum_{i=0}^{K}\rho_i-\frac{1}{\rho}\sum_{i=1}^{K+1}\rho_i\right)-\left(1+B\sum_{i=1}^{K}\rho_i+B\rho_{K+1}\right)\left(\sum_{i=0}^{K-1}\rho_i-\frac{1}{\rho}\sum_{i=1}^{K}\rho_i\right)\\
&=\left(1+B\sum_{i=1}^{K}\rho_i\right)\left(\rho_K-\frac{1}{\rho}\rho_{K+1}\right)-B\rho_{K+1}\left(\sum_{i=0}^{K-1}\rho_i-\frac{1}{\rho}\sum_{i=1}^{K}\rho_i\right)\\
&=\left(1+B\sum_{i=1}^{K}\rho_i\right)\rho_K-\frac{1}{\rho}\rho_{K+1}-B\rho_{K+1}\sum_{i=0}^{K-1}\rho_i\\
&=\left(1+B\sum_{i=1}^{K}\rho_i\right)\rho_K-\rho_{K+1}\left(\frac{1}{\rho}+B\sum_{i=0}^{K}\rho_i-B\rho_K\right)>0. \qquad (5.72)
\end{aligned}$$

Now dividing both sides of (5.72) by $\rho_K > 0$ and using $\rho_{K+1} = \frac{\eta}{C+K+1}\rho_K$, it can be shown that (5.72) is equivalent to the inequality

$$\eta B\rho_K > (\eta - C - K - 1)\left(1+B\sum_{i=1}^{K}\rho_i\right)+\eta\left(\frac{1}{\rho}+B-1\right). \qquad (5.73)$$

Since $\eta\left(\frac{1}{\rho}+B-1\right) > 0$ by Lemma 5.3.1, the above is a stronger result than (5.71). We will prove this inequality for $K \geq 0$ by induction in the following. When $K = 0$, the inequality is

$$\eta B > (\eta - C - 1) + \eta\left(\frac{1}{\rho}+B-1\right) = \eta B - 1,$$

which is clearly true.

Assuming that (5.73) is true for $K \geq 0$, we will prove

$$\begin{aligned}
\eta B\rho_{K+1} &= \frac{\eta^2 B\rho_K}{\Delta_1}\\
&> (\eta - C - K - 2)\left(1+B\sum_{i=1}^{K}\rho_i+\frac{B\eta\rho_K}{\Delta_1}\right)+\eta\left(\frac{1}{\rho}+B-1\right)\\
&= \Delta_2\left(1+B\sum_{i=1}^{K}\rho_i+\frac{B\eta\rho_K}{\Delta_1}\right)+\eta\left(\frac{1}{\rho}+B-1\right)
\end{aligned}$$

or equivalently,

$$\eta^2 B\rho_K - \Delta_2 B\eta\rho_K = (\Delta_1+1)\eta B\rho_K > \Delta_1\Delta_2 + \Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i + \Delta_1\eta\left(\frac{1}{\rho}+B-1\right).$$

Using the assumption (5.73), we have

$$
\begin{aligned}
(\Delta_1+1)\eta B\rho_K &> (\Delta_1+1)\left[(\Delta_2+1)\left(1+B\sum_{i=1}^{K}\rho_i\right)+\eta\left(\frac{1}{\rho}+B-1\right)\right] \\
&= (\Delta_1\Delta_2+\Delta_1+\Delta_2+1)\left(1+B\sum_{i=1}^{K}\rho_i\right)+(\Delta_1+1)\eta\left(\frac{1}{\rho}+B-1\right) \\
&= (\Delta_1\Delta_2+\eta)\left(1+B\sum_{i=1}^{K}\rho_i\right)+\eta\left(\frac{1}{\rho}+B-1\right)+\Delta_1\eta\left(\frac{1}{\rho}+B-1\right) \\
&= \Delta_1\Delta_2+\Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i+\eta\left(1+B\sum_{i=1}^{K}\rho_i\right)+\eta\left(\frac{1}{\rho}+B-1\right)+\Delta_1\eta\left(\frac{1}{\rho}+B-1\right) \\
&= \Delta_1\Delta_2+\Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i+\eta\left(B\sum_{i=0}^{K}\rho_i+\frac{1}{\rho}\right)+\Delta_1\eta\left(\frac{1}{\rho}+B-1\right) \\
&> \Delta_1\Delta_2+\Delta_1\Delta_2 B\sum_{i=1}^{K}\rho_i+\Delta_1\eta\left(\frac{1}{\rho}+B-1\right),
\end{aligned}
$$

which is what we want to prove.

While K approaches to 0, $\rho_0=1$, $\sum_{i=1}^{0}\rho_i=0$ and since

$$
P^{(K)}(Ab)=\frac{B[1+\sum_{i=1}^{K-1}\rho_i-\frac{1}{\rho}\sum_{i=1}^{K}\rho_i]}{1+B\sum_{i=1}^{K}\rho_i}=\frac{B[1+\sum_{i=1}^{K}\rho_i-\rho_K-\frac{1}{\rho}\sum_{i=1}^{K}\rho_i]}{1+B\sum_{i=1}^{K}\rho_i}
$$

we have $P^{(0)}(Ab)=0$. Since $\sum_{i=1}^{\infty}\rho_i=A-1$, we have that while K approaches to infinity

$$
\begin{aligned}
P^{(\infty)}(Ab) &= \frac{B[1+A-1-\frac{1}{\rho}(A-1)]}{1+B(A-1)} \\
&= \frac{B[1+A(\rho-1)]}{\rho[1+(A-1)B]},
\end{aligned}
$$

which is $P(Ab)$ of $M/M/S+M$ model. ∎

Remark 5.3.3 *Since we have $E(Q_q)=\eta P(Ab)$ (5.42), the above monotonicity result also applies to $E(Q_q)$, the expected number of calls waiting in the queue in equilibrium. Also by Little's formula, $E(Q_q)=\lambda E(\overline{W}_q, \text{non-blocking})$, the above monotonicity result applies to $E(\overline{W}_q, \text{non-blocking})$ as well. In addition by Remark 5.3.1, the above monotonicity result also applies to $E(Q)=E(Q_b)+E(Q_q)$, the expected number of calls in the system in equilibrium.*

The relationship between $P(Ab)$ and $P(Sr)$

In the following, we will study the relationship between $P^{(K)}(Ab)$ and $P^{(K)}(Sr)$. We know that both performance measures increase with buffer size K. For $K=0$, $P^{(0)}(Ab)=0$

and $P^{(0)}(Sr) = B$ so that $P^{(0)}(Sr) > P^{(0)}(Ab)$. What is the relationship between these two probabilities for general K? We have the following result.

Theorem 5.3.4 *For $M/M/S/N+M$ model, $P^{(K)}(Sr) > P^{(K)}(Ab)$ when $1 \le K \le C$ or when $1 \le K \le \lfloor C \rfloor$ since K is an integer, where $\lfloor C \rfloor = \left\lfloor \frac{S\mu}{\alpha} \right\rfloor$ is the floor function i.e., the largest integer not greater than C.*

Proof. We have

$$\begin{aligned}
&P^{(K)}(Sr) - P^{(K)}(Ab) \\
=&\frac{1}{\rho} + \frac{1 - B - 1/\rho}{1 + B\sum_{i=1}^{K}\rho_i} - \frac{B\left(\sum_{i=0}^{K-1}\rho_i - \frac{1}{\rho}\sum_{i=1}^{K}\rho_i\right)}{1 + B\sum_{i=1}^{K}\rho_i} \\
=&\frac{1 + B\sum_{i=1}^{K}\rho_i + \rho(1 - B - 1/\rho) - \rho B\left(\sum_{i=0}^{K-1}\rho_i - \frac{1}{\rho}\sum_{i=1}^{K}\rho_i\right)}{\rho\left(1 + B\sum_{i=1}^{K}\rho_i\right)} \\
=&\frac{B\left(2\sum_{i=1}^{K}\rho_i - \rho\sum_{i=1}^{K-1}\rho_i\right) + \rho(1 - 2B)}{\rho\left(1 + B\sum_{i=1}^{K}\rho_i\right)} \\
=&\frac{B(2-\rho)\sum_{i=1}^{K}\rho_i + \rho B\rho_K + \rho(1-2B)}{\rho\left(1 + B\sum_{i=1}^{K}\rho_i\right)}.
\end{aligned}$$

Hence $P^{(K)}(Sr) > P^{(K)}(Ab)$ if and only if

$$B(2-\rho)\sum_{i=1}^{K}\rho_i + \rho B\rho_K + \rho(1-2B) > 0$$

or equivalently

$$B\left[2\rho + (\rho - 2)\sum_{i=1}^{K-1}\rho_i - 2\rho_K\right] < \rho. \tag{5.74}$$

We will prove (5.74) for $1 \le K \le C$ by induction. When $K = 1 \le C$, (5.74) becomes

$$\begin{aligned}
&B\left[2\rho + (\rho - 2)\sum_{i=1}^{0}\rho_i - 2\rho_1\right] \\
&= 2B(\rho - \rho_1) = \frac{2B\rho}{C+1} < \rho
\end{aligned}$$

since $C \ge 1$ and $B < 1$.

Assuming that (5.74) holds for $1 \leq K \leq C-1$, we have

$$\begin{aligned} & B\left[2\rho + (\rho-2)\sum_{i=1}^{K}\rho_i - 2\rho_{K+1}\right] \\ &= B\left[2\rho + (\rho-2)\sum_{i=1}^{K-1}\rho_i - 2\rho_K\right] + B(\rho\rho_K - 2\rho_{K+1}) \\ &< \rho + B(\rho\rho_K - 2\rho_{K+1}) \end{aligned}$$

by assumption. Now we need to show that

$$\rho\rho_K - 2\rho_{K+1} \leq 0, \text{ when } 1 \leq K \leq C-1. \tag{5.75}$$

We have

$$\begin{aligned} \rho\rho_K - 2\rho_{K+1} &= \rho\rho_K - 2\frac{\eta}{C+K+1}\rho_K \\ &= \rho_K\left(\rho - \frac{2\eta}{C+K+1}\right) \\ &= \rho_K\left(\frac{2\eta}{2C} - \frac{2\eta}{C+K+1}\right) \leq 0 \end{aligned}$$

since $\rho_K > 0$ and $C+K+1 \leq 2C$. ∎

This theorem shows that for all $0 \leq K \leq \lfloor C \rfloor$, $P^{(K)}(Sr) > P^{(K)}(Ab)$ or equivalently

$$B\left[2\rho + (\rho-2)\sum_{i=1}^{K-1}\rho_i - 2\rho_K\right] < \rho.$$

What is the behavior of $P^{(K)}(Sr)$ and$P^{(K)}(Ab)$ for $K > \lfloor C \rfloor$? We first consider the case when $K = \lfloor C \rfloor + 1$. In this case, we have $P^{(\lfloor C \rfloor+1)}(Sr) \leq P^{(\lfloor C \rfloor+1)}(Ab)$ if and only if

$$B\left[2\rho + (\rho-2)\sum_{i=1}^{\lfloor C \rfloor}\rho_i - 2\rho_{\lfloor C \rfloor+1}\right] \geq \rho. \tag{5.76}$$

To consider the general case $K > \lfloor C \rfloor + 1$, we need the following Lemma.

Lemma 5.3.2 *Let $g(K) = 2\rho + (\rho-2)\sum_{i=1}^{K}\rho_i - 2\rho_{K+1}$ then $g(K)$ strictly increases in K when $K > C-2$.*

Proof. For any integer $K > C-2$, we only need to prove $g(K+1) - g(K) > 0$. We have

$$\begin{aligned} & g(K+1) - g(K) \\ &= (\rho-2)\sum_{i=1}^{K+1}\rho_i - 2\rho_{K+2} - (\rho-2)\sum_{i=1}^{K}\rho_i + 2\rho_{K+1} \\ &= \rho\rho_{K+1} - 2\rho_{K+2} > 0 \end{aligned}$$

if and only if $K > C - 2$ by referring to (5.75). ∎

Now for general $K > \lfloor C \rfloor + 1$, we have the following result.

Theorem 5.3.5 *Under the condition (5.76), we have that for $K > \lfloor C \rfloor + 1$, $P^{(K)}(Sr) < P^{(K)}(Ab)$.*

Proof. From the proof of Theorem 5.3.4, we know that $P^{(K)}(Sr) < P^{(K)}(Ab)$ if and only if

$$B\left[2\rho + (\rho - 2)\sum_{i=1}^{K-1}\rho_i - 2\rho_K\right] = Bg(K-1) > \rho.$$

Hence if condition (5.76) holds, i.e., $Bg(\lfloor C \rfloor) \geq \rho$, we have for $K > \lfloor C \rfloor + 1$, by Lemma 5.3.2 and since $K - 1 > \lfloor C \rfloor > C - 2$,

$$Bg(K-1) > Bg(\lfloor C \rfloor) \geq \rho,$$

which means $P^{(K)}(Sr) < P^{(K)}(Ab)$ for $K > \lfloor C \rfloor + 1$. ∎

Waiting time distribution $P(W_q > t)$

For the waiting time distribution, we will first consider the monotonicity property with respect to buffer size K for $P(W_q > 0) = P(delay|\text{non-blocking})$ which is, by (5.46),

$$P^{(K)}(delay|\text{non-blocking}) = \frac{BD(K)}{1 + B[D(K) - 1]}.$$

Theorem 5.3.6 *For $M/M/S/N+M$ model, $P^{(K)}(delay|non\text{-}blocking)$ is strictly increasing in the buffer size K. $P^{(K)}(delay|non\text{-}blocking)$ approaches to $P(delay)$ of $M/M/S+M$ model while K approaches to infinity and it approaches to 0 while K approaches to 0.*

Proof. Since $P^{(K)}(delay|\text{non-blocking}) > 0$, to show that $P^{(K)}(blocking)$ is strictly increasing in the buffer size K, it suffices to show that for $K > 0$,

$$\frac{P^{(K+1)}(delay|\text{non-blocking})}{P^{(K)}(delay|\text{non-blocking})} > 1.$$

We have that

$$\begin{aligned}\frac{P^{(K+1)}(delay|\text{non-blocking})}{P^{(K)}(delay|\text{non-blocking})} &= \frac{BD(K+1)}{1 + B[D(K+1) - 1]} \cdot \frac{1 + B[D(K) - 1]}{BD(K)} \\ &= \frac{D(K+1) + BD(K)D(K+1) - BD(K+1)}{D(K) + BD(K)D(K+1) - BD(K)} \\ &> 1\end{aligned}$$

is equivalent to

$$D(K+1) - BD(K+1) > D(K) - BD(K)$$

or

$$(1-B)D(K+1) > (1-B)D(K)$$

which is clearly true since $B < 1$ and $D(K)$ is an increasing function by its definition.

For the limiting cases, the results are obvious since $D(\infty) = A$, $D(0) = 0$ and from (5.6), we have for $M/M/S + M$ model,

$$P(delay) = \frac{AB}{1 + (A-1)B}.$$

■

In order to prove the monotonicity property with respect to buffer size K for general $P(W_q > t)$, we need the following lemmas. The first lemma gives an alternative expression for $P(W_{qi} > t)$, the waiting time of a non-blocking call given that it finds i in the system. Substituting $A_{n,i-S}$ by (5.14) in (5.44), we have

$$P(W_{qi} > t) = \frac{\prod_{k=0}^{i-S}(C+k)}{(i-S)!} \sum_{n=0}^{i-S} \binom{i-S}{n} \frac{(-1)^n}{C+n} e^{-[S\mu+(n+1)\alpha]t}. \tag{5.77}$$

However the above is the sum of terms with alternating signs and to overcome this problem, we prove the following result.

Lemma 5.3.3 *For $M/M/S/N + M$ model,*

$$P(W_{qi} > t) = \frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{q=0}^{i-S} \frac{\left[\eta(1-e^{-\alpha t})\right]^q}{q!\rho_{q-1}}$$

and when $\alpha \to 0$, the above approaches to $P(W_{qi} > t)$ of $M/M/S/N$ model.

Proof. We will first prove the equality

$$\sum_{q=j}^{m} \frac{\prod_{k=0}^{q-1}(C+k)}{j!(q-j)!} = \binom{m}{j} \frac{\prod_{k=0}^{m}(C+k)}{m!(C+j)} \tag{5.78}$$

for integers $m \geq j \geq 0$ by induction. When $m = j$, the equality is clearly true. Assuming it holds for $m = p > j$, we will prove it still holds for $m = p + 1$. We have

$$\begin{aligned}
\sum_{q=j}^{p+1} \frac{\prod_{k=0}^{q-1}(C+k)}{j!(q-j)!} &= \sum_{q=j}^{p} \frac{\prod_{k=0}^{q-1}(C+k)}{j!(q-j)!} + \frac{\prod_{k=0}^{p}(C+k)}{j!(p+1-j)!} \\
&= \frac{\prod_{k=0}^{p}(C+k)}{j!(p-j)!(C+j)} + \frac{\prod_{k=0}^{p}(C+k)}{j!(p+1-j)!} \\
&= \frac{(p+1-j)\prod_{k=0}^{p}(C+k)}{j!(p+1-j)!(C+j)} + \frac{(C+j)\prod_{k=0}^{p}(C+k)}{j!(p+1-j)!(C+j)} \\
&= \frac{\prod_{k=0}^{p+1}(C+k)}{j!(p+1-j)!(C+j)} = \binom{p+1}{j} \frac{\prod_{k=0}^{p+1}(C+k)}{(p+1)!(C+j)},
\end{aligned}$$

which completes the proof of (5.78). Now by (5.77) and (5.78), we will get another expression for $P(W_{qi} > t)$ which is better in terms of computation,

$$\begin{aligned}
P(W_{qi} > t) &= \frac{\prod_{k=0}^{i-S}(C+k)}{(i-S)!} \sum_{n=0}^{i-S} \binom{i-S}{n} \frac{(-1)^n}{C+n} e^{-[S\mu+(n+1)\alpha]t} \\
&= e^{-(S\mu+\alpha)t} \sum_{n=0}^{i-S} \frac{\prod_{k=0}^{i-S}(C+k)}{(i-S)!} \binom{i-S}{n} \frac{(-1)^n}{C+n} e^{-n\alpha t} \\
&= e^{-(S\mu+\alpha)t} \sum_{n=0}^{i-S} \sum_{q=n}^{i-S} \frac{\prod_{k=0}^{q-1}(C+k)}{n!(q-n)!} (-1)^n e^{-n\alpha t} \\
&= e^{-(S\mu+\alpha)t} \sum_{q=0}^{i-S} \sum_{n=0}^{q} \frac{\prod_{k=0}^{q-1}(C+k)}{n!(q-n)!} (-1)^n e^{-n\alpha t} \\
&= e^{-(S\mu+\alpha)t} \sum_{q=0}^{i-S} \frac{\prod_{k=0}^{q-1}(C+k)(1-e^{-\alpha t})^q}{q!}.
\end{aligned}$$

Since by definition, $\rho_q = \frac{\eta^q}{\prod_{k=1}^{q}(C+k)}$, we have

$$\prod_{k=0}^{q-1}(C+k) = \frac{C\eta^{q-1}}{\rho_{q-1}} = \frac{C\eta^q}{\eta\rho_{q-1}} = \frac{\eta^q}{\rho\rho_{q-1}},$$

where $\rho_{-1} := \rho^{-1}$. Therefore

$$P(W_{qi} > t) = \frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{q=0}^{i-S} \frac{\left[\eta(1-e^{-\alpha t})\right]^q}{q!\rho_{q-1}}.$$

When $\alpha \to 0$, we have $\rho_{q-1} \to \rho^{q-1}$ so that

$$\begin{aligned}\lim_{\alpha\to 0} P(W_{qi} > t) &= e^{-S\mu t} \lim_{\alpha\to 0} \sum_{q=0}^{i-S} \frac{\left[\eta(1-e^{-\alpha t})\right]^q}{q!\rho^q} \\ &= e^{-S\mu t} \sum_{q=0}^{i-S} \frac{\lim_{\alpha\to 0}\left[\frac{S\mu(1-e^{-\alpha t})}{\alpha}\right]^q}{q!} \\ &= e^{-S\mu t} \sum_{q=0}^{i-S} \frac{(S\mu t)^q}{q!},\end{aligned}$$

which is $P(W_{qi} > t)$ of $M/M/S/N$ model studied in Chapter 2. ∎

Remark 5.3.4 *This alternative expression is given in [19] without a proof. Note that in this expression, each term in the sum is positive, which is better in terms of computation and is a fact that we will apply in the proof of the monotonicity property for* $P(W_q > t)$.

The second lemma is an alternative derivation of $P(W_q>t)$ using a different method and this method has been used in Chapter 2 for $M/M/S/N$ model as well.

Lemma 5.3.4 *For* $M/M/S/N + M$ *model,*

$$P(W_q>t) = P(delay|non\text{-}blocking)\frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{j=0}^{N-S-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \left[1 - \frac{D(j)}{D(N-S)}\right]. \tag{5.79}$$

Proof. We have that $P(W_q>t)$ can be factorized as follows,

$$\begin{aligned}P(W_q>t) &= P(W_q>t|W_q>0)P(W_q>0) \\ &= P(W_q>t|W_q>0)P(delay|\text{non-blocking}).\end{aligned}$$

Now

$$
\begin{aligned}
P(W_q{>}t|W_q{>}\,0) &= P(W_q{>}t|S \le Q \le N-1) \\
&= \sum_{i=0}^{N-S-1} P(W_q{>}t|Q = S+i; S \le Q \le N-1) P(Q = S+i|S \le Q \le N-1) \\
&= \sum_{i=0}^{N-S-1} P(W_{q(S+i)} > t) \frac{\rho_i}{1+\rho_1+...+\rho_{N-S-1}} \\
&= \frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{i=0}^{N-S-1} \left(\sum_{j=0}^{i} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \right) \frac{\rho_i}{1+\rho_1+...+\rho_{N-S-1}} \\
&= \frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{j=0}^{N-S-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \sum_{i=j}^{N-S-1} \frac{\rho_i}{1+\rho_1+...+\rho_{N-S-1}} \\
&= \frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{j=0}^{N-S-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \left[1 - \frac{D(j)}{D(N-S)}\right],
\end{aligned}
$$

where we have used Lemma 5.3.3 for $P(W_{q(S+i)} > t)$. Also we have used the fact that

$$
P(Q = S+i|S \le Q \le N-1) = \frac{\rho_i}{1+\rho_1+...+\rho_{N-S-1}}, \text{ for } i = 0, 1, ...N-S-1,
$$

which can be proved similarly as (2.24) in Chapter 2. Therefore,

$$
P(W_q{>}t) = P(delay|\text{non-blocking}) \frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{j=0}^{N-S-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \left[1 - \frac{D(j)}{D(N-S)}\right].
$$

■

Now we can prove the following result on the monotonicity property with respect to K for $P(W_q{>}\,t)$.

Theorem 5.3.7 *For $M/M/S/N+M$ model, $P^{(K)}(W_q{>}t)$ is strictly increasing in the buffer size K. $P^{(K)}(W_q{>}t)$ approaches to $P(W_q{>}t)$ of $M/M/S+M$ model while K approaches to infinity and it approaches to 0 while K approaches to 0.*

Proof. By (5.79), we have

$$
P^{(K)}(W_q{>}t) = P(delay|\text{non-blocking}) \frac{e^{-(S\mu+\alpha)t}}{\rho} \sum_{j=0}^{K-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \left[1 - \frac{D(j)}{D(K)}\right].
$$

Since $P(delay|\text{non-blocking})$ is strictly increasing in the buffer size K by Theorem 5.3.6, we only need to prove

$$
h(K) := \sum_{j=0}^{K-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \left[1 - \frac{D(j)}{D(K)}\right]
$$

is a strictly increasing function of K. We have

$$h(K+1) = \sum_{j=0}^{K-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \left[1 - \frac{D(j)}{D(K+1)}\right] + \frac{\left[\eta(1-e^{-\alpha t})\right]^K}{K!\rho_{K-1}} \left[1 - \frac{D(K)}{D(K+1)}\right]$$
$$> \sum_{j=0}^{K-1} \frac{\left[\eta(1-e^{-\alpha t})\right]^j}{j!\rho_{j-1}} \left[1 - \frac{D(j)}{D(K)}\right] = h(K)$$

since each term in the above is positive and for $0 \le j \le K-1$, we can prove that

$$1 - \frac{D(j)}{D(K+1)} \ge 1 - \frac{D(j)}{D(K)}.$$

The above is true since $D(K+1) > D(K)$.

For the limiting cases, the results are obvious since

$$P^{(K)}(W_q{>}t) = \frac{1}{1-p_{K+S}} \sum_{i=S}^{K+S-1} p_i P(W_{qi} > t).$$

■

5.3.2 Concavity property with respect to buffer size K

We will consider the concavity property of the probability of served calls with respect to buffer size K. We have the following result.

Theorem 5.3.8 *For $M/M/S/N+M$ model, $P^{(K)}(Sr)$ is a strictly concave function in the buffer size K.*

Proof. Since

$$P^{(K)}(Sr) = \frac{1}{\rho} + \frac{1 - B - 1/\rho}{1 + B\sum_{i=1}^{K} \rho_i},$$

we have

$$U_K := P^{(K+1)}(Sr) - P^{(K)}(Sr)$$
$$= (1 - B - 1/\rho) \left(\frac{1}{1 + B\sum_{i=1}^{K+1} \rho_i} - \frac{1}{1 + B\sum_{i=1}^{K} \rho_i} \right).$$

To prove concavity, we will prove that U_K is strictly decreasing, i.e., $\frac{U_{K+1}}{U_K} < 1$ for $K \ge 0$ since $U_K > 0$ by Theorem 5.3.1. Let $f(K) := 1 + B\sum_{i=1}^{K} \rho_i$, then we need to prove

$$\frac{U_{K+1}}{U_K} = \frac{\frac{1}{f(K+2)} - \frac{1}{f(K+1)}}{\frac{1}{f(K+1)} - \frac{1}{f(K)}}$$
$$= \frac{\frac{\rho_{K+2}}{f(K+2)f(K+1)}}{\frac{\rho_{K+1}}{f(K+1)f(K)}} < 1,$$

which is equivalent to $\rho_{K+2}f(K) < \rho_{K+1}f(K+2)$. However from (5.70) in the proof of Theorem 5.3.2, we have $\rho_{K+1}f(K) < \rho_K f(K+1)$ for $K \geq 0$. Hence

$$\rho_{K+1}f(K+2) > \rho_{K+2}f(K+1) > \rho_{K+2}f(K),$$

since $f(K+1) > f(K)$ for $K \geq 0$. ∎

Remark 5.3.5 *Since we have $E(Q_b) = aP(Sr)$ (5.43), the above concavity result also applies to $E(Q_b)$, the expected number of busy servers in equilibrium, or the carried load.*

Remark 5.3.6 *In [26], the authors gave the same result but their proof is much more complex than ours.*

5.3.3 Monotonicity properties with respect to S

Now we will study the monotonicity properties of performance measures with respect to the number of CSRs S while the number of trunk lines $N = S + K$ is fixed. Since the Erlang B formula $B(S, a)$ is involved with S, we will replace it with $B(S)$ here. Also since $K = N - S$, we have the probability of served calls,

$$P^{(S)}(Sr) = \frac{1}{\rho} + \frac{1 - B(S) - 1/\rho}{1 + B(S)\sum_{i=1}^{N-S}\rho_i},$$

Since $B(S)$ is a decreasing function of S, it is easy to see that $P^{(S)}(Sr)$ is increasing with respect to the number of CSRs S while the number of trunk lines N is fixed. However similar results for other performance measures are harder to obtain and we will leave this as a future research.

5.3.4 Numerical examples

To give some numerical illustrations of the above results, we will consider the following two examples. Let $\mu = 1, \alpha = 2$ be the common parameters. Example 5.1 has parameters $\lambda = 8$ and $S = 5$ while Example 5.2 has parameters $\lambda = 40$ and $S = 10$. Let buffer size K change from 0 to 30. We compute the results using Matlab, where special functions such as $A(x, y)$ etc. are available.

We first consider three probabilities: $P(Sr), P(Ab)$ and $P(blocking)$. The results are shown in Figure 5.5 for Example 5.1 and in Figure 5.6 for Example 5.2.

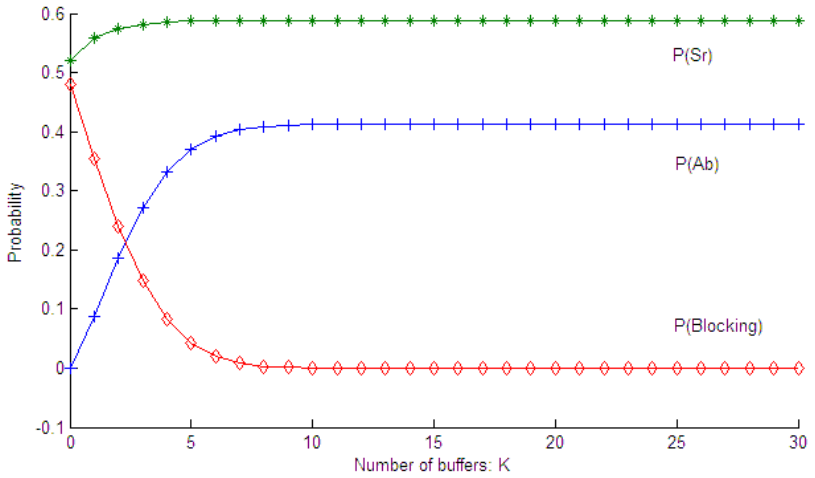

Figure 5.5: $P(Sr), P(Ab)$ and $P(blocking)$ of Example 5.1 for $M/M/S/N+M$ model

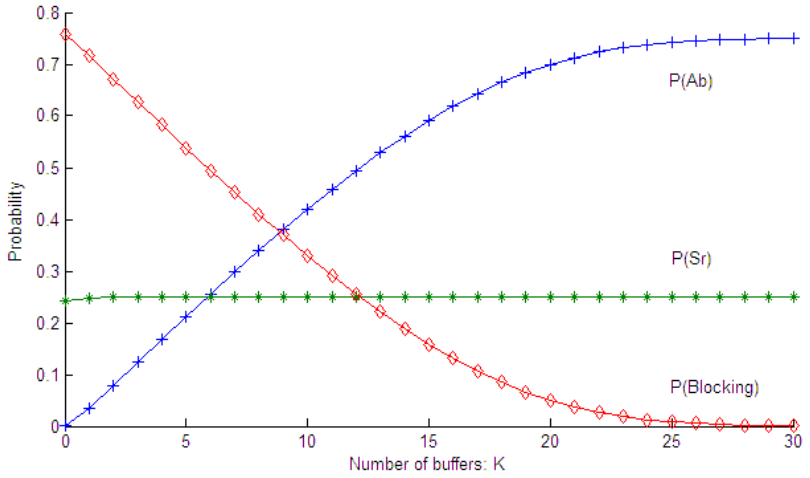

Figure 5.6: $P(Sr), P(Ab)$ and $P(blocking)$ of Example 5.2 for $M/M/S/N+M$ model

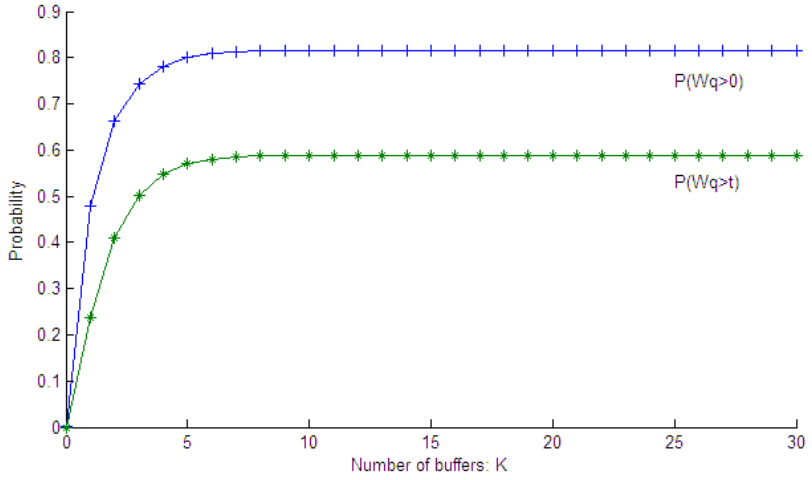

Figure 5.7: $P(Wq > 0)$ and $P(Wq > t)$ of Example 5.1 for $M/M/S/N + M$ model

For both examples, we find that the three probabilities have the monotonicity properties we have proved in the corresponding theorems. For Example 5.2 the condition (5.76) is satisfied. Therefore we have when $K \leq \lfloor C \rfloor = 5$, $P^{(K)}(Sr) > P^{(K)}(Ab)$ and when $K > \lfloor C \rfloor = 5$, $P^{(K)}(Sr) < P^{(K)}(Ab)$, as proved in Theorem 5.3.4 and Theorem 5.3.5. However for Example 5.1 the condition (5.76) is not satisfied. We only have the result that when $K \leq \lfloor C \rfloor = 2$, $P^{(K)}(Sr) > P^{(K)}(Ab)$ according to Theorem 5.3.4. Note that Example 5.1 also appears in [26] for $P(Sr)$ and our numerical results agree with that in [26].

We then consider two probabilities: $P(W_q > 0) = P(delay|\text{non-blocking})$ and $P(W_q > t)$. The results are shown in Figure 5.7 for Example 5.1, where we let $t = 0.1$. Both $P(W_q > 0)$ and $P(W_q > t)$ are increasing in the buffer size K as we have proved. When $K = 0$ both $P(W_q > 0)$ and $P(W_q > t)$ are 0 and when K approaches to infinity, it can be verified that both $P(W_q > 0)$ and $P(W_q > t)$ approach to the corresponding performance measures of $M/M/S + M$ model.

5.4 SOQN model with exponential abandonment of call centres (SOQN+M)

In this section we will study the SOQN model with exponential abandonment of call centres denoted by SOQN+M, which is a generalization of $M/M/S/N+M$ model to the two-node SOQN case. Our work here is a generalization and correction of [45]; we have more results and some derivations are new. We will first give a model description and then derive the main performance measures of this model.

5.4.1 Model description

The model is a semiopen network with two nodes in series. Node 1 models the IVRU with N servers each with exponential service rate θ and Node 2 models the CSRs with S ($\leq N$) servers each with exponential service rate μ. The maximum number of calls in the network is N, i.e., if an arriving call finds N calls in the system, it will be blocked and rejected entering the system. Hence there is no queue at Node 1 and there are at most $N-S$ calls waiting at Node 2. Arriving calls can enter the network only through Node 1 according to a Poisson process with arrival rate λ. After the service with Node 1 is completed, the call leaves the network with probability $\overline{p} = 1-p$ and it joins Node 2 with probability p. If there are free CSRs at Node 2, the call is served by one of S CSRs. Otherwise it waits in the queue to get service.

We model abandonment at Node 2 and assume that upon joining the queue, calls start the patience times which are i.i.d. exponential with rate α. If the virtual waiting time V for the call is longer than its patience time (denoted by X), the call will abandon, leave the system and release the trunk line. Otherwise it gets service with a CSR and releases both the CSR and the trunk line and leaves the system after the service. The arrival, service and abandonment processes are all assumed to be independent. The original SOQN model in Chapter 3 can be thought of as this model with $\alpha = 0$. Figure 5.8 gives a picture of the model. Note that this model has been similarly described in [45].

5.4.2 Product form solution of the queue length process

Let $Q(t) = (Q_1(t), Q_2(t))$ be the queue length process, where $Q_i(t)$ is the queue length (number of calls) of Node $i, i = 1, 2$ at time t. From the description of the model, we find

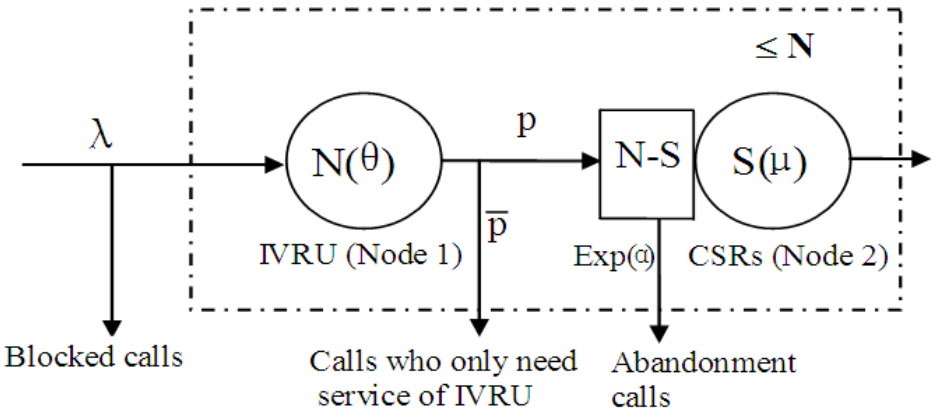

Figure 5.8: SOQN model with exponential abandonment

that $Q(t)$ is a finite two dimensional CTMC with restriction $Q_1(t) + Q_2(t) \leq N$ for all $t \geq 0$. Hence the stationary distribution of $Q(t)$ exists and let $\pi_{ij} = P(Q_1 = i, Q_2 = j)$ be the stationary probability of having i calls at Node 1 and j calls at Node 2 with the state space $\Omega = \{(i,j)|i+j \leq N, (i,j) \in Z_+^2\}$.

For Node 2, as in $M/M/S/N+M$ model, we have the abandonment rate

$$r_j = \begin{cases} 0 & \text{if } 0 \leq j \leq S \\ (j-S)\alpha & \text{if } S < j \leq N \end{cases}$$

since the exponential abandonment. Following the same way as $M/M/S/N+M$ model, we can incorporate the abandonment rate r_j into the service rate for Node 2. Then it is easy to know that this model is a special case of the semiopen network model studied in Section 4.2, Chapter 4 with state dependent service rates $\mu_1(j) = j\theta$,

$$\mu_2(j) = \begin{cases} j\mu & for\ 0 \leq j \leq S \\ S\mu + (j-S)\alpha & for\ S < j \leq N \end{cases}$$

and constant balking. Now we have that Node 1 is similar to a $M/M/\infty$ queue with arrival rate λ and service rate θ. Node 2 is similar to a $M/M/S/N+M$ queue with arrival rate λp, service rate μ and abandonment rate α. Hence we have the following result.

Theorem 5.4.1 *For SOQN+M model, the stationary distribution of the queue length process $Q = \{Q_1, Q_2\}$ has product form solution,*

$$\pi_{ij} = \pi_{00} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}, \quad 0 \leq i+j \leq N \ , \tag{5.80}$$

where $a_1 = \lambda/\theta$; $a_2 = p\lambda/\mu = pa$;

$$\beta(j) := \begin{cases} j! & for\ 0 \le j \le S \\ \frac{S!a^{j-S}}{\rho_{j-S}} & for\ S < j \le N \end{cases};$$

$$\rho_{j-S} = \frac{\eta^{j-S}}{\prod_{k=1}^{j-S}(C+k)};$$

$\eta = \lambda/\alpha$; $C = \frac{S\mu}{\alpha}$ *and* $\pi_{00} = \left[\sum_{0\le i+j\le N} \frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)}\right]^{-1}$ *is the normalizing constant.*

Remark 5.4.1 π_{ij} *is actually the solution of the following global balance equations*

$$\begin{aligned}
&\pi_{ij}(\lambda+\mu_2(j)+\mu_1(i)) = \pi_{(i-1)j}\lambda + \\
&\pi_{i(j+1)}\mu_2(j+1) + \pi_{(i+1)(j-1)}\mu_1(i+1)p + \pi_{(i+1)j}\mu_1(i+1)\overline{p},\ i \ge 1, j \ge 1, i+j \le N-1; \\
&\pi_{ij}(\mu_2(j)+\mu_1(i)) = \pi_{(i-1)j}\lambda + \pi_{(i+1)(j-1)}\mu_1(i+1)p,\ i \ge 1, j \ge 1, i+j = N; \\
&\pi_{0j}(\lambda+\mu_2(j)) = \pi_{0(j+1)}\mu_2(j+1) + \pi_{1(j-1)}\mu_1(1)p + \pi_{1j}\mu_1(1)\overline{p},\ i = 0, 1 \le j \le N-1; \\
&\pi_{0N}\mu_2(N) = \pi_{1(N-1)}\mu_1(1)p,\ i = 0, j = N; \\
&\pi_{i0}(\lambda+\mu_1(i)) = \pi_{(i-1)0}\lambda + \pi_{i1}\mu_2(1) + \pi_{(i+1)0}\mu_1(i+1)\overline{p},\ 1 \le i \le N-1, j = 0; \\
&\pi_{N0}\mu_1(N) = \pi_{(N-1)0}\lambda,\ i = N, j = 0; \\
&\pi_{00}\lambda = \pi_{01}\mu_2(1) + \pi_{10}\mu_1(1)\overline{p},\ i = 0, j = 0.
\end{aligned} \tag{5.81}$$

Remark 5.4.2 *This result has been given in [45] using different notations.*

The explicit expression for π_{00}^{-1} was also derived in [45],

$$\begin{aligned}
\pi_{00}^{-1} &= \sum_{0\le i+j\le N} \frac{a_1^i}{i!}\frac{a_2^j}{\beta(j)} = \sum_{k=0}^{N}\sum_{j=0}^{k} \frac{a_1^{k-j}}{(k-j)!}\frac{a_2^j}{\beta(j)} \\
&= \sum_{k=0}^{S}\sum_{j=0}^{k} \frac{a_1^{k-j}a_2^j}{(k-j)!j!} + \sum_{k=S+1}^{N}\left(\sum_{j=0}^{S} \frac{a_1^{k-j}a_2^j}{(k-j)!j!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j}a_2^j\rho_{j-S}}{(k-j)!S!a^{j-S}}\right).
\end{aligned}$$

Then applying binomial formula yields

$$\begin{aligned}
\pi_{00}^{-1} &= \sum_{k=0}^{S} \frac{(a_1+a_2)^k}{k!} + \sum_{k=S+1}^{N} \frac{(a_1+a_2)^k}{k!} + \sum_{k=S+1}^{N}\sum_{j=S+1}^{k}\left(-\frac{a_1^{k-j}a_2^j}{(k-j)!j!} + \frac{a_1^{k-j}a_2^j\rho_{j-S}}{(k-j)!S!a^{j-S}}\right) \\
&= \sum_{k=0}^{N} \frac{(a_1+a_2)^k}{k!} + \sum_{k=S+1}^{N}\sum_{j=S+1}^{k} \frac{a_1^{k-j}a_2^j}{(k-j)!}\left(\frac{\rho_{j-S}}{S!a^{j-S}} - \frac{1}{j!}\right),
\end{aligned}$$

which will reduce to p_0^{-1} of $M/M/S/N+M$ model if we let $\theta = \infty$ and $p = 1$. In addition when $\alpha = 0$, the above will reduce to π_{00}^{-1} of SOQN model.

The marginal distribution for Node 1 is

$$\pi_{i*} := P(Q_1 = i) = \sum_{j=0}^{N-i} \pi_{ij} = \pi_{00} \sum_{j=0}^{N-i} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}, 0 \leq i \leq N$$

where π_{00}^{-1} has another expression $\pi_{00}^{-1} = \sum_{i=0}^{N} \sum_{j=0}^{N-i} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}$. Hence the mean number of calls at Node 1 is

$$E(Q_1) = \sum_{i=0}^{N} i\pi_{i*} = \pi_{00} \sum_{i=0}^{N} i \sum_{j=0}^{N-i} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}. \tag{5.82}$$

The marginal distribution for Node 2 is

$$\pi_{*j} := P(Q_2 = j) = \sum_{i=0}^{N-j} \pi_{ij} = \pi_{00} \sum_{i=0}^{N-j} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}, 0 \leq j \leq N$$

where π_{00}^{-1} has the third expression $\pi_{00}^{-1} = \sum_{j=0}^{N} \sum_{i=0}^{N-j} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}$. Hence the mean number of calls at Node 2 is

$$E(Q_2) = \sum_{j=0}^{N} j\pi_{*j} = \pi_{00} \sum_{j=0}^{N} j \sum_{i=0}^{N-j} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}$$

and the mean number of calls waiting in the queue at Node 2 is

$$E(Q_{2q}) = \sum_{j=S+1}^{N} (j-S)\pi_{*j} = \pi_{00} \sum_{j=S+1}^{N} (j-S) \sum_{i=0}^{N-j} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}.$$

5.4.3 Blocking probability

In [45], the stationary probabilities π_k for $0 \leq k \leq N$ that there are exactly k calls in the system has been derived similarly as in [42],

$$\pi_k := P(Q = k) = \sum_{j=0}^{k} \pi_{(k-j)j},$$

where $Q = Q_1 + Q_2$. There are two cases:

1. $0 \leq k \leq S$:

$$\pi_k = \pi_{00} \sum_{j=0}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} = \pi_{00} \frac{(a_1 + a_2)^k}{k!}.$$

2. $S < k \leq N$:

$$\begin{aligned} \pi_k &= \pi_{00} \left(\sum_{j=0}^{S} \frac{a_1^{k-j} a_2^j}{(k-j)!j!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j \rho_{j-S}}{(k-j)!S!a^{j-S}} \right) \\ &= \pi_{00} \left(\frac{(a_1 + a_2)^k}{k!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \left(\frac{\rho_{j-S}}{S!a^{j-S}} - \frac{1}{j!} \right) \right). \end{aligned}$$

Therefore, the probability π_k that there are exactly $0 \leq k \leq N$ calls in the system is equal to

$$\pi_k = \pi_{00} \left(\frac{(a_1 + a_2)^k}{k!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \left(\frac{\rho_{j-S}}{S! a^{j-S}} - \frac{1}{j!} \right) I_{(S,\infty)}(k) \right)$$

and the blocking probability is $P(blocking) = \pi_N$. It is obvious that $\pi_0 = \pi_{00}$.

The mean number of total calls in the system is

$$E(Q) = \sum_{k=0}^{N} k\pi_k = \sum_{k=0}^{N} k \sum_{j=0}^{k} \pi_{(k-j)j} = E(Q_1) + E(Q_2).$$

For Node 1, we apply the same Little's formula as in Chapter 3

$$E(Q_1) = a_1 \left[1 - P(blocking)\right]$$

which is also the carried load for Node 1. The utilization

$$v_1 = \frac{E(Q_1)}{N} = \frac{a_1 \left[1 - P(blocking)\right]}{N} < 1$$

is the proportion of time that an IVRU server is busy.

5.4.4 Probability of abandonment and other performance measures

Since abandonment occurs only at Node 2 while calls are waiting, we need to consider the probability of abandonment among those calls who are not blocked and join Node 2 denoted as $P(Ab|entry)$ where the event *entry* means non-blocking and joining Node 2. We use the same idea as before to derive $P(Ab|entry)$; we first condition on the state seen by an entry call and then sum up all the possibilities.

$$\begin{aligned} P(Ab|entry) &= \sum_{j=S}^{N-1} P_j(Ab) P(\text{the call finds } j \text{ calls at Node 2}|entry) \\ &= \sum_{j=S}^{N-1} q_j \frac{(j-S+1)\alpha}{S\mu + (j-S+1)\alpha}, \end{aligned} \tag{5.83}$$

where we have used (5.38) for $P_j(Ab)$. q_j has been defined in Chapter 3 and determined by the Arrival Theorem of CQN [7]: $q_j = \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)}$. Hence we have

$$P(Ab|entry) = \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \frac{(j-S+1)\alpha}{S\mu + (j-S+1)\alpha}.$$

Furthermore, Theorem 3.4.1 in Chapter 3 shows that

$$q_j = \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} = \sum_{k=j+1}^{N} \chi(k,j), \tag{5.84}$$

which also holds here since the proof of that theorem does not involve the specific form of $\beta(j)$. Therefore we have an expression in terms of $\chi(k,j)$

$$\begin{aligned} P(Ab|entry) &= \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \frac{(j-S+1)\alpha}{S\mu + (j-S+1)\alpha} \\ &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \frac{(j-S+1)\alpha}{S\mu + (j-S+1)\alpha}, \end{aligned}$$

which is the same result as obtained in Wang [45].

Now we have

$$\begin{aligned} P(Sr|entry) = 1 - P(Ab|entry) &= \sum_{j=0}^{S-1} q_j + \sum_{j=S}^{N-1} q_j \frac{S\mu}{S\mu + (j-S+1)\alpha} \\ &= \sum_{j=0}^{S-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} + \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \frac{S\mu}{S\mu + (j-S+1)\alpha} \\ &= \sum_{j=0}^{S-1} \sum_{k=j+1}^{N} \chi(k,j) + \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \frac{S\mu}{S\mu + (j-S+1)\alpha} \\ &= \sum_{k=1}^{N} \sum_{j=0}^{k\wedge S-1} \chi(k,j) + \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \frac{S\mu}{S\mu + (j-S+1)\alpha}. \end{aligned}$$

According to Chapter 3, the probability of entry is

$$P(entry) = p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}$$

so that we have

$$\begin{aligned} P(Ab) = P(Ab, entry) &= P(Ab|entry)P(entry) \\ &= \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \frac{(j-S+1)\alpha}{S\mu + (j-S+1)\alpha} \left(p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} \right) \\ &= p \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \frac{\pi_{ij}(j-S+1)\alpha}{S\mu + (j-S+1)\alpha} \end{aligned}$$

by the relationship

$$\pi_{ij}^{(N-1)} = \frac{\pi_{ij}}{\sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}}$$

as proved in Chapter 3. Also

$$\begin{aligned}
P(Sr) &= P(Sr, entry) = P(Sr|entry)P(entry) \\
&= \left(\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)} + \sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)}\frac{S\mu}{S\mu+(j-S+1)\alpha}\right)\left(p\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}\right) \\
&= p\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu}{S\mu+(j-S+1)\alpha}.
\end{aligned}$$

Now Little's formula for busy CSRs at Node 2 is

$$\begin{aligned}
E(Q_{2b}) &= \lambda P(Sr)\frac{1}{\mu} \\
&= ap\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + ap\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu}{S\mu+(j-S+1)\alpha}
\end{aligned}$$

which is easy to verify since

$$\begin{aligned}
&ap\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + ap\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu}{S\mu+(j-S+1)\alpha} \\
&= a_2\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + a_2\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu}{S\mu+(j-S+1)\alpha} \\
&= \sum_{j=0}^{S-1}(j+1)\sum_{i=0}^{N-1-j}\pi_{i(j+1)} + S\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{i(j+1)} \\
&= \sum_{j=1}^{S} j\sum_{i=0}^{N-j}\pi_{ij} + S\sum_{j=S+1}^{N}\sum_{i=0}^{N-j}\pi_{ij} \\
&= E(Q_{2b}),
\end{aligned}$$

where we have used the fact that

$$a_2\pi_{ij} = (j+1)\pi_{i(j+1)} \ for\ 0 \le j < S$$

and

$$a_2\mu\pi_{ij} = [S\mu + (j-S+1)\alpha]\,\pi_{i(j+1)} \ for\ j \ge S \tag{5.85}$$

which can be verified by the product form solution (5.80). Hence the carried load for Node 2 is

$$a' = E(Q_{2b}) = aP(Sr).$$

The utilization

$$v = \frac{a'}{S} = \rho P(Sr) < 1$$

is the proportion of time that a CSR is busy.

Similarly as in Section 3.3.4, Chapter 3, we have the following performance measures.

1. $P(\text{only self-served by Node 1}) = (1-p)\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
2. $P(\text{no-delay}, entry) = p\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
3. $P(delay, entry) = p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
4. $P(\text{no-delay}|entry) = \sum_{j=0}^{S-1} q_j = \sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)} = \sum_{j=0}^{S-1}\sum_{k=j+1}^{N}\chi(k,j) = \sum_{k=1}^{N}\sum_{j=0}^{k\wedge S-1}\chi(k,j)$.
5. $P(delay|entry) = \sum_{j=S}^{N-1} q_j = \sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)} = \sum_{j=S}^{N-1}\sum_{k=j+1}^{N}\chi(k,j) = \sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)$.

5.4.5 Waiting time distribution

To obtain performance measures related to waiting time, such as the four-dimensional performance measure mentioned in $M/M/S+M$ model, we will study the stationary waiting time distribution in the following. As SOQN model in Chapter 3, we only need to consider those calls given they are not blocked and join Node 2 (or given entry), since there is no queue at Node 1. Let W_q denote the conditional stationary waiting time of calls given entry, which is the time spent by an entry call in the queue of Node 2 until abandonment or starting to get service.

We follow the same conditional argument as before and by (5.44) and (5.84), we have

$$\begin{aligned}
P(W_q > t) &= \sum_{j=S}^{N-1} P(W_{qj} > t) P(\text{the call finds } j \text{ calls at Node 2}|entry) \\
&= \sum_{j=S}^{N-1}\sum_{n=0}^{j-S} A_{n,j-S} e^{-[S\mu+(n+1)\alpha]t} q_j \\
&= \sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \sum_{n=0}^{j-S} A_{n,j-S} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{j=S}^{N-1}\sum_{k=j+1}^{N} \chi(k,j) \sum_{n=0}^{j-S} A_{n,j-S} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{k=S+1}^{N}\sum_{j=S}^{k-1} \chi(k,j) \sum_{n=0}^{j-S} A_{n,j-S} e^{-[S\mu+(n+1)\alpha]t},
\end{aligned}$$

which is the same result as obtained in Wang [45].

Similarly by (5.47) and (5.84), we have

$$\begin{aligned}
P(W_q{>}t, Sr) &= \sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} A_{n,j-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \sum_{n=0}^{j-S} A_{n,j-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \sum_{n=0}^{j-S} A_{n,j-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{n=0}^{j-S} A_{n,j-S} \frac{S\mu + n\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}
\end{aligned}$$

and

$$\begin{aligned}
P(W_q{>}t, Ab) &= P(W_q{>}t) - P(W_q{>}t, Sr) \\
&= \sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} A_{n,j-S} \frac{\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \sum_{n=0}^{j-S} A_{n,j-S} \frac{\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \sum_{n=0}^{j-S} A_{n,j-S} \frac{\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t} \\
&= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{n=0}^{j-S} A_{n,j-S} \frac{\alpha}{S\mu + (n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}.
\end{aligned}$$

Now we can derive the conditional waiting, which are more useful in practice, using the above results.

1.

$$\begin{aligned}
P(W_q > t|Ab) &= \frac{P(W_q > t, Ab)}{P(Ab|entry)} \\
&= \frac{\sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} A_{n,j-S} \frac{1}{S\mu+(n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}}{\sum_{j=S}^{N-1} q_j \frac{j-S+1}{S\mu+(j-S+1)\alpha}}.
\end{aligned}$$

2.

$$\begin{aligned}
P(W_q > t|Sr) &= \frac{P(W_q > t, Sr)}{P(Sr|entry)} \\
&= \frac{\sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} A_{n,j-S} \frac{S\mu+n\alpha}{S\mu+(n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}}{\sum_{j=0}^{S-1} q_j + \sum_{j=S}^{N-1} q_j \frac{S\mu}{S\mu+(j-S+1)\alpha}}.
\end{aligned}$$

3.

$$P(W_q > t|Sr, delay) = \frac{P(W_q > t, Sr)}{P(Sr, delay|entry)}$$
$$= \frac{\sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} A_{n,j-S} \frac{S\mu+n\alpha}{S\mu+(n+1)\alpha} e^{-[S\mu+(n+1)\alpha]t}}{\sum_{j=S}^{N-1} q_j \frac{S\mu}{S\mu+(j-S+1)\alpha}}.$$

5.4.6 Mean waiting time

Following the same methods as in $M/M/S/N+M$ model, we can obtain some mean waiting times. The only difference is that here we have

$$q_j = \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} = \sum_{k=j+1}^{N} \chi(k, j).$$

Also since $\pi_{ik}^{(N-1)} = \frac{\pi_{ik}}{\sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}}$, we have

$$q_k = \frac{\sum_{i=0}^{N-1-k} \pi_{ik}}{\sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}}. \tag{5.86}$$

First by (5.54) and (5.55) the mean waiting time for all calls given entry is

$$E(W_q) = \sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} \frac{A_{n,j-S}}{S\mu + (n+1)\alpha}$$
$$= \sum_{j=S}^{N-1} q_j \frac{j-S+1}{S\mu + (j-S+1)\alpha}. \tag{5.87}$$

Comparing (5.83) with (5.87), we have found a similar property as Erlang-A model:

$$P(Ab|entry) = \alpha \cdot E(W_q). \tag{5.88}$$

Now Little's formula for all calls waiting in the queue at Node 2 is

$$E(Q_{2q}) = \lambda P(entry) E(W_q) \tag{5.89}$$
$$= \lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} E(W_q),$$

which is easy to verify since

$$\begin{aligned}\lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} E(W_q) &= \lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} \sum_{k=S}^{N-1} q_k \frac{k-S+1}{S\mu+(k-S+1)\alpha} \\ &= \sum_{k=S}^{N-1} \sum_{i=0}^{N-1-k} \lambda p \pi_{ik} \frac{k-S+1}{S\mu+(k-S+1)\alpha} \\ &= \sum_{k=S}^{N-1} \sum_{i=0}^{N-1-k} \pi_{i(k+1)}(k-S+1) \\ &= \sum_{k=S+1}^{N} (k-S) \sum_{i=0}^{N-k} \pi_{ik} = E(Q_{2q}),\end{aligned}$$

where we have used (5.85) and (5.86). From (5.88) and (5.89), we obtain the similar rate balance equation as (5.41):

$$\lambda P(Ab) = \alpha \cdot E(Q_{2q}).$$

Next by (5.56) and (5.59) the mean waiting time for served calls given entry

$$\begin{aligned}E(W_q, Sr) &= \sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} \frac{A_{n,j-S}(S\mu+n\alpha)}{[S\mu+(n+1)\alpha]^2} \\ &= \sum_{j=S}^{N-1} q_j \frac{S\mu}{S\mu+(j-S+1)\alpha} \sum_{n=0}^{j-S} \frac{1}{S\mu+(n+1)\alpha}.\end{aligned}$$

At last by (5.63) and (5.64) the mean waiting time for abandoned calls given entry

$$\begin{aligned}E(W_q, Ab) &= \sum_{j=S}^{N-1} q_j \sum_{n=0}^{j-S} \frac{A_{n,j-S}\alpha}{[S\mu+(n+1)\alpha]^2} \\ &= \sum_{j=S}^{N-1} q_j \frac{\alpha}{S\mu+(j-S+1)\alpha} \sum_{n=0}^{j-S} \frac{n+1}{S\mu+(n+1)\alpha}.\end{aligned}$$

Now we can derive the conditional mean waiting time, which are more useful in practice, using the above results.

1.

$$\begin{aligned}E(W_q|Ab) &= \frac{E(W_q, Ab)}{P(Ab|entry)} \\ &= \frac{\sum_{j=S}^{N-1} q_j \frac{1}{S\mu+(j-S+1)\alpha} \sum_{n=0}^{j-S} \frac{n+1}{S\mu+(n+1)\alpha}}{\sum_{j=S}^{N-1} q_j \frac{j-S+1}{S\mu+(j-S+1)\alpha}}.\end{aligned}$$

2.

$$E(W_q|Sr) = \frac{E(W_q, Sr)}{P(Sr|entry)} = \frac{\sum_{j=S}^{N-1} q_j \frac{S\mu}{S\mu+(j-S+1)\alpha} \sum_{n=0}^{j-S} \frac{1}{S\mu+(n+1)\alpha}}{\sum_{j=0}^{S-1} q_j + \sum_{j=S}^{N-1} q_j \frac{S\mu}{S\mu+(j-S+1)\alpha}}.$$

3.

$$E(W_q|Sr, delay) = \frac{E(W_q, Sr)}{P(Sr, delay|entry)} = \frac{\sum_{j=S}^{N-1} \frac{q_j}{S\mu+(j-S+1)\alpha} \sum_{n=0}^{j-S} \frac{1}{S\mu+(n+1)\alpha}}{\sum_{j=S}^{N-1} \frac{q_j}{S\mu+(j-S+1)\alpha}}.$$

5.4.7 Numerical examples

To give some numerical illustrations for the SOQN model with exponential abandonment, we will consider the following example for $P(blocking), P(W_q > t), P(Ab)$ and $P(Sr)$. This example is similar to the example discussed in Chapter 3 where the parameters are $\lambda = 250/1800, \mu = 1/180, t = 20$ seconds and $\theta = 0.01$. We also let $p = 0.5$ here. To illustrate the effect of buffer size, we fix $S = 5$ and let buffer size $K = N - S$ change from 0 to 30. In addition we will compare two cases for $\alpha : \alpha = 0.001$ and $\alpha = 0.1$ representing lower and higher abandonment rate respectively.

In Figure 5.9, we compare $P(blocking)$ for different abandonment rate α. It is obvious that $P(blocking)$ is a strictly decreasing function of K and higher abandonment rate makes lower $P(blocking)$. In Figure 5.10, we compare $P(W_q > 20)$ for different abandonment rate α. It is obvious that $P(W_q > 20)$ is a strictly increasing function of K and higher abandonment rate makes much lower $P(W_q > 20)$. Note that we have observed similar monotonicity properties with respect to buffer size K for $P(blocking)$ and $P(W_q > 20)$ as $M/M/S/N + M$ model and we will use this observation in Chapter 7 for the call centre design problem. Also when we compute $P(W_q > 20)$, we have used the expression of $P(W_{qi} > t)$ in Lemma 5.3.3, which makes the computation more stable than using (5.77), especially for very small α.

Next we will consider the unconditional probabilities $P(Ab)$ and $P(Sr)$, shown in Figure 5.11 and Figure 5.12 respectively. It is clear that both $P(Ab)$ and $P(Sr)$ increase when the buffer size K increases as proved in $M/M/S/N + M$ model. Also the higher abandonment rate makes lower $P(Sr)$ and higher $P(Ab)$ consistent with our intuitiveness.

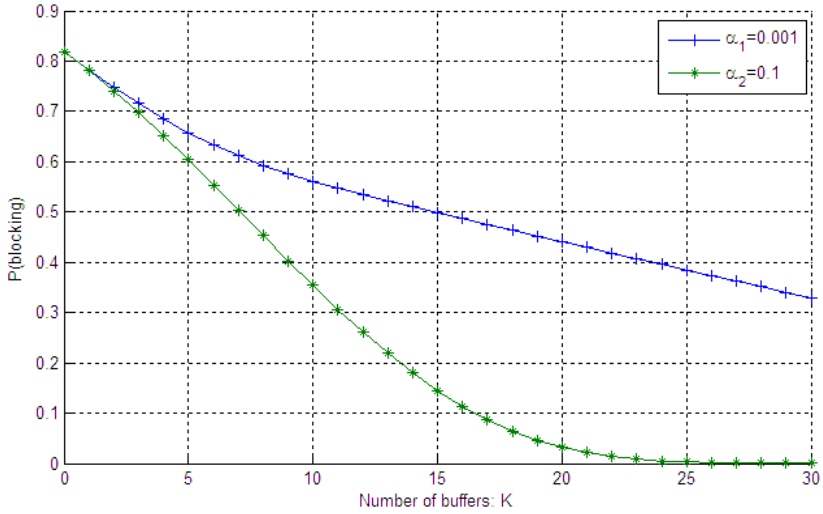

Figure 5.9: $P(blocking)$ for different abandonment rate α

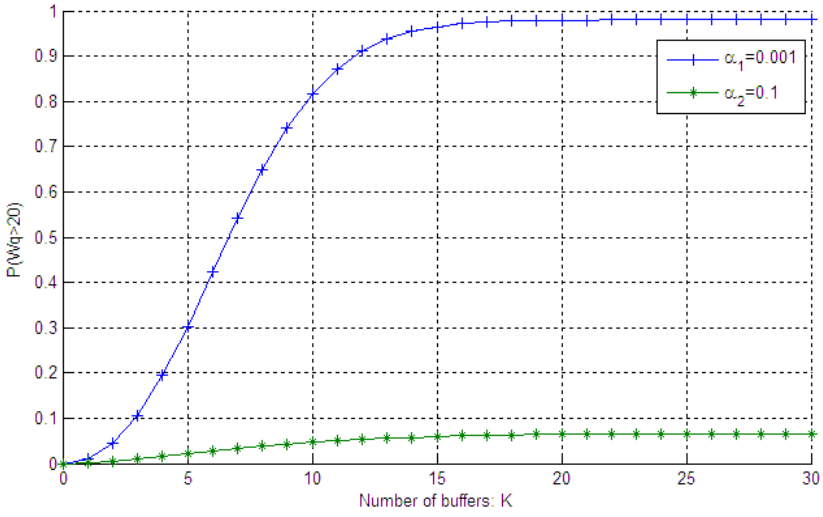

Figure 5.10: $P(W_q > 20)$ for different abandonment rate α

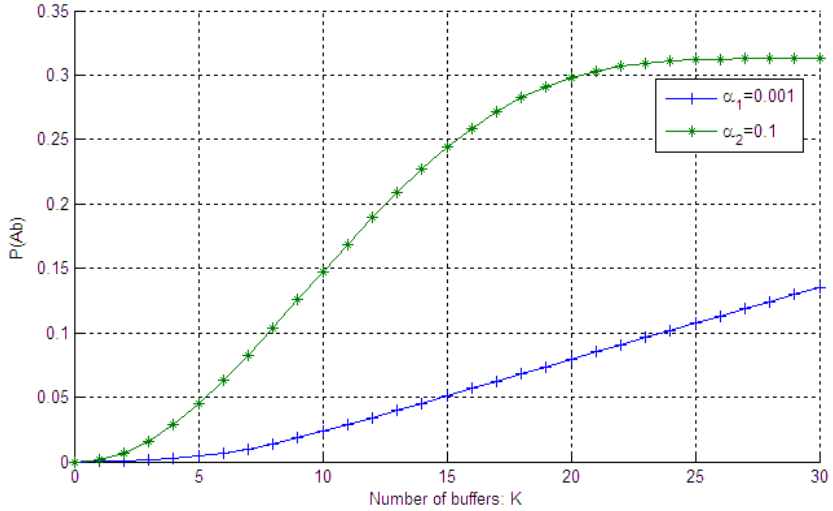

Figure 5.11: $P(Ab)$ for different abandonment rate α

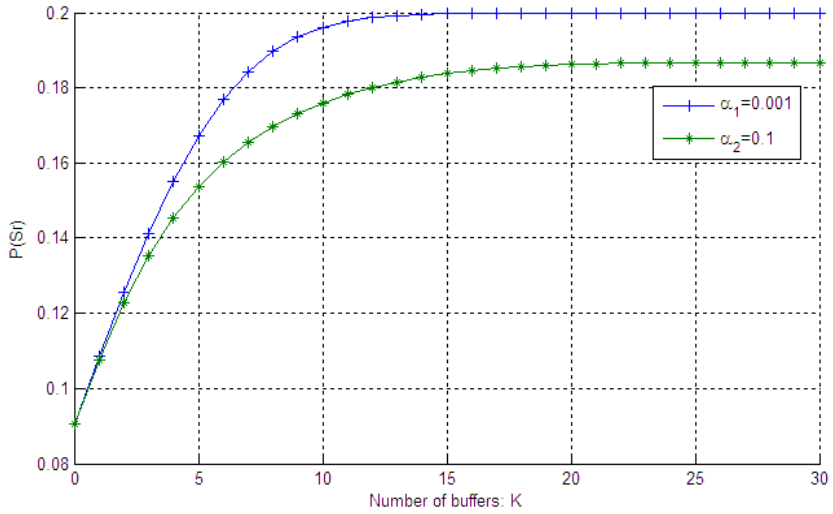

Figure 5.12: $P(Sr)$ for different abandonment rate α

5.5 Summary

Abandonment phenomenon in call centres is very important as shown in the literature. In this chapter, we studied the exponential abandonment model of call centres and analyzed three models, $M/M/S+M$, $M/M/S/N+M$ and SOQN+M. For single-node models, we again focused on the computational aspects by expressing the exact performance measures in terms of special functions and Erlang B formula. The analysis is new and we have provided a unified and comprehensive list of expressions for performance measures. Based on the performance analysis of $M/M/S/N+M$ model, we proved monotonicity and concavity properties with respect to buffer size K for $P(Sr)$ using a method that simplifies the work in [26]. Monotonicity properties for $P(blocking)$ and $P(W_q > t)$ were also proved and these properties are important to the call centre design algorithm in Chapter 7. For SOQN+M model, our work is a generalization and correction of [45]; we used a new approach not only to rederive the formulas for the performance measures correctly, but also introduce new results. In the end we provided numerical examples to illustrate the effect of abandonment rate for SOQN+M model.

Chapter 6

General Abandonment Models of Call Centres

In this chapter we will study the general abandonment models of call centres, which generalize the exponential abandonment models studied in Chapter 5 to the general abandonment. We will only consider the multiserver case in the following. Among the different patience time distributions, deterministic distribution has been given much attention in the earlier literature since it is relatively easier to study and has some real applications; some physical systems will not allow calls to wait more than a fixed time and equivalently this means that calls have a deterministic patience time on waiting if we assume that calls will not abandon by themselves. Barrer[6] studied $M/M/S+D$ model. He obtained the stationary distribution of number of calls in the system p_i by defining and finding the abandonment rate function mentioned in Chapter 5 (Definition 5.0.1). Gnedenko and Kovalenko[22] solved the same model using supplementary variable method and provided a rigorous analysis of the model. Choi and Kim [15] analyzed the $M/M/S+D$ model by finding another simple Markov process using supplementary variable method. Their method can handle priority queues with two classes of customers and impatience in the class of higher priority. Boots and Tijms[8] gave a simple and insightful solution for $P(Ab)$, which is exact for $M/M/S+D$ and $M/G/1+D$ queues and provided an excellent heuristic for the $M/G/S+D$ queue.

Baccelli and Hebuterne [5] studied the more general $M/M/S+G$ queue with general patience time distribution. They obtained the stationary distribution for the actual and virtual offered waiting time process by constructing a Markov process and solving the Kolmogorov equations. They then gave the relevant performance measures. However the stationary distribution of number of calls in the system p_i for $i \geq S$ were not given. Independently Movaghar [36] studied $M(n)/M/S+G$ model and Brandt and Brandt [11] analyzed $M(n)/M(n)/S+G$ model in which arrival and service rates are allowed to depend on the system size n. Hence these are more general model which can include balking as

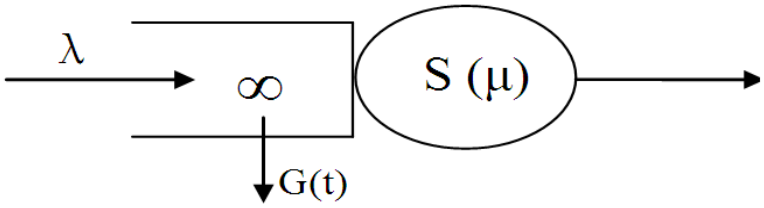

Figure 6.1: $M/M/S+G$ model description and parameters

well as finite buffers.

Basically when the patience time distribution is not exponential, the number of calls in the system by itself is no longer a Markov process. We have to construct an appropriate Markov process by adding other variables. The above method is called supplementary variable method and includes the work of [22], [15], [5] and [11]. These works differ from each other with the choice of supplementary variables. There is another method which is heavily dependent on the concept of *the stationary abandonment rate function* r_i (Definition 5.0.1). They argued that once r_i is found, then the system can be seen as a Markov process and p_i can be obtained using birth-death process. Hence this method is less rigorous than supplementary variable method although they produce the same p_i. The work of [6] and [36] belong to this method.

In this chapter, we will first give a review and new generalization of single-node general abandonment models including infinite buffer $M/M/S+G$ and finite buffer $M/M/S/N+G$ model and then study the SOQN+G model, which is SOQN model with general abandonment.

6.1 $M/M/S+G$ model

This model is a generalization of $M/M/S+M$ model with a general abandonment distribution. The patience times X are assumed to be i.i.d. and generally distributed with mean α^{-1} and distribution function $G(t)$, hence the corresponding survival function $\overline{G}(t) = P(X > t)$. We assume X is positive, i.e., $G(0) = 0$. The model description and parameters are shown in Figure 6.1.

This section is mainly based on the work of Baccelli and Hebuterne [5]. They assumed

that arriving calls are aware of their offered waiting time V (defined in Chapter 5) upon arrival. Thus, if $V > X$, calls abandon immediately and do not join the queue. However, this model coincides with regular $M/M/S+G$ model in terms of abandonment probability and offered waiting time V since the finally abandoned calls in $M/M/S+G$ model will not influence V and hence $P(Ab)$; they can be discarded upon arrival as the model studied here [5].

For single server queue, let w_n be the unfinished work of the system at the n-th customer arriving time and w_n is called the *actual offered waiting time* of the n-th customer. Let $V(t)$ be the *virtual offered waiting time* at time t (i.e., the offered waiting time of a hypothetical infinitely-patient call arriving at time t). Baccelli and Hebuterne gave the stability conditions of these two processes and showed that their stationary distributions coincide, which are actually the distribution of V for $M/G/1+G$ model. This result also holds for multiserver queue $M/M/S+G$.

Baccelli and Hebuterne [5] chose $V(t)$ as the supplementary variable to construct the following Markov process $\{(N(t), V(t)), t \geq 0\}$, where $N(t)$ is the number of busy CSRs and $V(t)$ is the virtual offered waiting time at time t, which is strictly positive when $N(t) = S$ and equals zero otherwise. The state space is $\{0, 1, 2, ...S-1, S\} \times R^+$. Then they considered the following functions

$$\begin{cases} v(x) = \lim\limits_{t\to\infty} \lim\limits_{dx\to 0} \frac{P(N(t)=S, x<V(t)\leq x+dx)}{dx}, & x \geq 0 \\ p_i = \lim\limits_{t\to\infty} P(N(t) = i, V(t) = 0), & 0 \leq i \leq S-1 \end{cases}$$

where $v(x)$ is the density of V for $x \geq 0$.

The Kolmogorov equations for $(N(t), V(t))$ in equilibrium are

$$\begin{cases} \lambda p_0 = \mu p_1 & \\ (\lambda + i\mu)p_i = \lambda p_{i-1} + (i+1)\mu p_{i+1} & 0 < i < S-1 \\ (\lambda + (S-1)\mu)p_{S-1} = \lambda p_{S-2} + v(0) & \\ v(x) = \lambda p_{S-1} \exp(-S\mu x) + \lambda \int_0^x \overline{G}(u) v(u) \exp[-S\mu(x-u)]du, & x > 0. \end{cases}$$

The solution is

$$\begin{cases} p_i = \frac{a^i}{i!} p_0, & 0 \leq i \leq S-1 \\ v(x) = \lambda p_{S-1} \exp\{\lambda \int_0^x \overline{G}(u)du - S\mu x\}, & x \geq 0 \end{cases}$$

and by the normalizing condition $\sum_{i=0}^{S-1} p_i + \int_0^\infty v(x)dx = 1$,

$$p_0 = \left[\sum_{i=0}^{S-2} \frac{a^i}{i!} + \frac{a^{S-1}}{(S-1)!}(1+\lambda J) \right]^{-1}, \tag{6.1}$$

where

$$J = \int_0^\infty \exp\{\lambda \int_0^x \overline{G}(u)du - S\mu x\}dx.$$

The stability condition is $\lambda\overline{G}(\infty) < S\mu$ which can be explained as the CSRs must be able to overcome the traffic consisting of customers with infinite patience to make the system steady.

Since Baccelli and Hebuterne did not provide p_i for $i \geq S$, the mean number of calls waiting in the queue $E(Q_q)$ cannot be obtained directly using p_i although it can be derived using Little's formula and the results for mean waiting time as in Zeltyn [50]. However the mean number of busy servers $E(Q_b)$ can be expressed only in terms of p_i for $0 \leq i \leq S-1$,

$$\begin{aligned} E(Q_b) &= \sum_{i=0}^{S-1} ip_i + S\sum_{i=S}^{\infty} p_i \\ &= ap_0 \sum_{i=0}^{S-2} \frac{a^i}{i!} + S(1 - \sum_{i=0}^{S-2} \frac{a^i}{i!} p_0 - p_{S-1}) \\ &= (a-S)\sum_{i=0}^{S-2} p_i + S(1-p_{S-1}). \end{aligned}$$

Hence the utilization

$$\begin{aligned} v = \frac{a'}{S} &= \frac{E(Q_b)}{S} \\ &= (\rho - 1)\sum_{i=0}^{S-2} p_i + 1 - p_{S-1} \end{aligned}$$

which is the proportion of time that a CSR is busy. By Little's formula for Q_b, we have $E(Q_b) = \lambda P(Sr)\frac{1}{\mu}$. Therefore,

$$\begin{aligned} P(Sr) &= \frac{E(Q_b)}{a} \\ &= (1 - \frac{1}{\rho})\sum_{i=0}^{S-2} p_i + \frac{1-p_{S-1}}{\rho} \end{aligned}$$

and

$$\begin{aligned} P(Ab) &= 1 - P(Sr) \\ &= (1 - \frac{1}{\rho})(1 - \sum_{i=0}^{S-2} p_i) + \frac{p_{S-1}}{\rho} \\ &= (1 - \frac{1}{\rho})(1 - \sum_{i=0}^{S-1} p_i) + p_{S-1} \end{aligned}$$

which is formula (5.9) in [5] and the previous derivation also appeared in [50].

Since the distribution of V is known, the distribution of W_q and its expectation, as well as various conditional versions as studied in Chapter 5, can be obtained. Zeltyn [50] gave a comprehensive list of exact formulas for $M/M/S+G$ performance measures based on the work of Baccelli and Hebuterne [5]. Define $H(x) := \int_0^x \overline{G}(u)du$. They gave the following building blocks,

$$\begin{aligned} J &= \int_0^\infty e^{\lambda H(x) - S\mu x} dx, \\ J_1 &= \int_0^\infty x e^{\lambda H(x) - S\mu x} dx, \\ J_H &= \int_0^\infty H(x) e^{\lambda H(x) - S\mu x} dx, \\ J(t) &= \int_t^\infty e^{\lambda H(x) - S\mu x} dx, \\ J_1(t) &= \int_t^\infty x e^{\lambda H(x) - S\mu x} dx, \\ J_H(t) &= \int_t^\infty H(x) e^{\lambda H(x) - S\mu x} dx, \\ \epsilon &= B(S-1, a)^{-1} = \frac{1-B}{B}\rho. \end{aligned}$$

Then it can be shown that almost all the performance measures can be expressed in terms

of these building blocks. For example

$$\begin{aligned}
\sum_{i=0}^{S-2} p_i &= \frac{\epsilon - 1}{\epsilon + \lambda J}, \\
p_{S-1} &= \frac{1}{\epsilon + \lambda J}, \\
P(delay) = P(W_q > 0) &= P(V > 0) = \frac{\lambda J}{\epsilon + \lambda J}, \\
P(Ab) &= \frac{1 + (\lambda - S\mu)J}{\epsilon + \lambda J}, \\
P(Sr) &= \frac{\epsilon - 1 + S\mu J}{\epsilon + \lambda J}, \\
E(Q_b) &= \frac{a(\epsilon - 1) + S\lambda J}{\epsilon + \lambda J}, \\
v &= \frac{\rho(\epsilon - 1) + \lambda J}{\epsilon + \lambda J}.
\end{aligned}$$

See [50] for the complete list.

For the special case $M/M/S + M$ model, $\overline{G}(u) = e^{-au}$ and $\lambda p_{S-1} = S\mu p_S$ so that

$$\begin{aligned}
v(x) &= \lambda p_{S-1} \exp\{\lambda \int_0^x \overline{G}(u)du - S\mu x\} \\
&= S\mu p_S \exp\{\eta(1 - e^{-\alpha x}) - S\mu x\} \\
&= p_S S\mu e^{\eta} e^{-(S\mu x + \eta e^{-\alpha x})},
\end{aligned}$$

which agrees with the result (5.17) in Chapter 5. Also it can be shown that for $M/M/S+M$ model, using the notation in Chapter 5, we have

$$\begin{aligned}
J &= \frac{e^{\eta}}{\alpha}\eta^{-C}\gamma(C,\eta) = \frac{A}{S\mu}, \\
J(t) &= \frac{A}{S\mu}\frac{\gamma(C,\eta e^{-\alpha t})}{\gamma(C,\eta)} = J\frac{\gamma(C,\eta e^{-\alpha t})}{\gamma(C,\eta)}, \qquad (6.2) \\
J_H &= \frac{1}{\alpha}\left[\frac{A}{S\mu} - \frac{A}{\lambda C}\frac{\gamma(C+1,\eta)}{\gamma(C,\eta)}\right] \\
&= \frac{1}{\alpha}\left[\frac{A}{S\mu} - \frac{A-1}{\lambda}\right] = \frac{1 + (\rho - 1)A}{\lambda\alpha} = \frac{1 + (\lambda - S\mu)J}{\lambda\alpha}, \\
J_H(t) &= \frac{1}{\alpha}\left[\frac{A}{S\mu}\frac{\gamma(C,\eta e^{-\alpha t})}{\gamma(C,\eta)} - \frac{A}{\lambda C}\frac{\gamma(C+1,\eta e^{-\alpha t})}{\gamma(C,\eta)}\right] \qquad (6.3) \\
&= \frac{1}{\alpha}\left[J(t) - \frac{A}{\lambda}\frac{\gamma(C,\eta e^{-\alpha t}) - C^{-1}(\eta e^{-\alpha t})^C e^{-\eta e^{-\alpha t}}}{\gamma(C,\eta)}\right].
\end{aligned}$$

The above special building blocks for $M/M/S + M$ model are also included in Zeltyn [50]. However they did not provide the expressions of J_1 and $J_1(t)$ for $M/M/S + M$ model. We

have that for $M/M/S+G$ model,

$$\begin{aligned}\int_t^\infty J(x)dx &= \int_t^\infty \left[\int_x^\infty e^{\lambda H(u)-S\mu u}du\right]dx \\ &= \int_t^\infty (u-t)e^{\lambda H(u)-S\mu u}du \\ &= J_1(t) - tJ(t)\end{aligned}$$

so that

$$J_1(t) = \int_t^\infty J(x)dx + tJ(t)$$

and

$$J_1 = \int_0^\infty J(x)dx.$$

Hence, for $M/M/S+M$ model, we have

$$\begin{aligned}J_1 &= \frac{A}{S\mu\gamma(C,\eta)}\int_0^\infty \gamma(C,\eta e^{-\alpha x})dx \\ &= \frac{e^\eta}{\alpha\eta^C}\int_0^\infty \gamma(C,\eta e^{-\alpha x})dx\end{aligned}$$

and

$$J_1(t) = \frac{e^\eta}{\alpha\eta^C}\left[\int_t^\infty \gamma(C,\eta e^{-\alpha x})dx + t\gamma(C,\eta e^{-\alpha t})\right].$$

The above special building blocks for $M/M/S+M$ model, together with the formulas of performance measures for $M/M/S+G$ model provided by Zeltyn [50], can be used to derive various performance measures for $M/M/S+M$ model, which will agree with our results obtained in Chapter 5.

6.2 $M/M/S/N+G$ model

The model is a generalization of finite buffer $M/M/S/N+M$ model with a general abandonment distribution. Again the patience times X are assumed to be i.i.d. and generally distributed with mean α^{-1} and distribution function $G(t)$, hence the corresponding survival function $\overline{G}(t) = P(X > t)$. We assume X is positive, i.e., $G(0) = 0$. The model description and parameters are shown in Figure 6.2.

As we have mentioned, Movaghar [36] studied $M(n)/M/S+G$ model using *abandonment rate function* r_i and Brandt and Brandt [11] analyzed $M(n)/M(n)/S+G$ model using supplementary variable method. Note that the input symbol $M(n)$ means the input

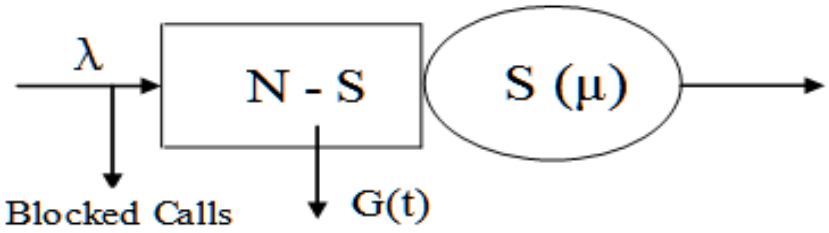

Figure 6.2: $M/M/S/N+G$ model description and parameters

is a general state-dependent Poisson process, i.e., the arrival rate is dependent on the number of calls n in the system. Hence this model can include both finite and infinite buffer models. When $\lambda_n \equiv \lambda > 0$ for $0 \leq n < N$ and $\lambda_n \equiv 0$ for $n \geq N$, we have $M/M/S/N+G$ model. The study in this section on $M/M/S/N+G$ model is mainly based on the work of [36]. However some results are new. For example, following the idea of building blocks of performance measures for $M/M/S+G$ model in Zeltyn [50], we introduce new building blocks for $M/M/S/N+G$ model.

6.2.1 Queue length process

In [36], Movaghar first gave the definition of the offered waiting time U, which is defined to be ∞ for blocked calls. However in the finite buffer model, he actually studied V, the conditional offered waiting time of an infinite patient call given the call is not blocked so that V has no mass at ∞. Movaghar obtained the distribution of V using a different method than Baccelli and Hebuterne [5]. He also obtained the stationary distribution of queue length p_i for all i and other stationary performance measures such as $P(AB)$, and $P(blocking)$ etc.

As we mentioned in Chapter 5 (Definition 5.0.1), the author gave a definition of abandonment rate function r_i, which is first introduced by Barrer [6] for $M/M/S+D$ model. Another important concept is the *conditional offered waiting time* given the non-blocked typical call in equilibrium finds i calls in the system, denoted by V_i for $0 \leq i < N$, which is also defined in Chapter 5 (Section 5.1.2). The first result he obtained is similar to formula (5.38) for $M/M/S/N+M$ model in Chapter 5 and is given below

$$P_i(Ab) = P_i(Ab, \text{non-blocking}) = P(V_i > X) = \frac{r_{i+1}}{S\mu + r_{i+1}}, S \leq i < N. \tag{6.4}$$

By using probabilistic arguments, the author derived the density of V_i

$$f_{V_i}(t) = \begin{cases} 0, & 0 \le i < S \\ \frac{g_{i-S}(t)}{g^*_{i-S}(S\mu)} e^{-S\mu t}, & S \le i < N \end{cases} \tag{6.5}$$

where

$$g_i(t) = \left[\int_0^t \overline{G}(u)du\right]^i = H(t)^i$$

and $g_i^*(s) = \int_0^\infty e^{-st} g_i(t)dt$ is the Laplace transform of $g_i(t)$. Then the explicit form for abandonment rate function can be obtained using (6.4)

$$r_i = \begin{cases} 0, & 0 \le i \le S \\ (i-S)\frac{g^*_{i-S-1}(S\mu)}{g^*_{i-S}(S\mu)} - S\mu, & S < i \le N \end{cases}. \tag{6.6}$$

In case of $M/M/S/N+M$ model, we have

$$g_i(t) = \left[\frac{1-e^{-\alpha t}}{\alpha}\right]^i,$$

and

$$\begin{aligned} g_i^*(S\mu) &= \int_0^\infty e^{-S\mu t} g_i(t)dt = \int_0^\infty e^{-S\mu t}\left[\frac{1-e^{-\alpha t}}{\alpha}\right]^i dt \\ &= \frac{1}{\alpha^i}\int_0^\infty e^{-S\mu t}[(1-e^{-\alpha t}]^i dt \\ &= \frac{1}{\alpha^i}\sum_{k=0}^{i}\binom{i}{k}(-1)^k \int_0^\infty e^{-(S\mu+\alpha k)t}dt \\ &= \frac{1}{\alpha^i}\sum_{k=0}^{i}\binom{i}{k}(-1)^k \frac{1}{S\mu+\alpha k} \\ &= \frac{i!}{\prod_{k=0}^{i}(S\mu+\alpha k)} \end{aligned}$$

where we have used the identity:

$$\sum_{k=0}^{i}\frac{1}{k!(i-k)!}(-1)^k\frac{1}{S\mu+\alpha k} = \frac{\alpha^i}{\prod_{k=0}^{i}(S\mu+\alpha k)}$$

which is equivalent to $\sum_{k=0}^{i} A_{k,i} = 1$ as shown in Chapter 5. Using the above, (6.5) and (6.6) reduce to the corresponding formulas (5.15) and (5.32) respectively in Chapter 5 for $M/M/S+M$ and $M/M/S/N+M$ models.

To obtain the stationary distribution of queue length p_i for all i, Movaghar argued that once r_i is found, p_i will satisfy a set of difference equations similar as the global balance

equations of a birth-death process with abandonment rate r_i as additional death rate. Therefore for $M/M/S+G$ model p_i satisfy the following equations.

$$\begin{cases} 0 = -\lambda p_0 + \mu p_1 \\ 0 = \lambda p_{i-1} - (\lambda + \min(S,i)\mu + r_i)p_i + (\min(S,n+1)\mu + r_{i+1})p_{i+1}, \quad i > 0 \end{cases}.$$

The solution, in view of (6.6), is

$$p_i = \begin{cases} p_0 \frac{a^i}{i!} & \text{if } 0 \le i \le S \\ p_0 \frac{a^{S-1}}{(S-1)!} \frac{\lambda^{i-S+1} g^*_{i-S}(S\mu)}{(i-S)!} & \text{if } i \ge S \end{cases}, \tag{6.7}$$

where

$$\begin{aligned} p_0^{-1} &= \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}}{(S-1)!} \sum_{i=S}^{\infty} \frac{\lambda^{i-S+1} g^*_{i-S}(S\mu)}{(i-S)!} \\ &= \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda}{(S-1)!} \sum_{i=S}^{\infty} g^*_{i-S}(S\mu) \frac{\lambda^{i-S}}{(i-S)!} \\ &= \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda}{(S-1)!} \sum_{i=S}^{\infty} \int_0^{\infty} e^{-S\mu t} H(t)^{i-S} dt \frac{\lambda^{i-S}}{(i-S)!} \\ &= \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda}{(S-1)!} \int_0^{\infty} e^{\lambda H(t) - S\mu t} dt \\ &= \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda J}{(S-1)!} \end{aligned}$$

which is the same as (6.1). Therefore

$$p_S = p_0 \frac{a^S}{S!} = \frac{\frac{a^S}{S!}}{\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda J}{(S-1)!}} = \frac{\rho}{\epsilon + \lambda J}.$$

Note that the stationary distribution of number of calls in the system p_i for $i \ge S$ are unavailable in Baccelli and Hebuterne [5]. The mean number of calls waiting in the queue $E(Q_q)$ (not given in [36]) is

$$\begin{aligned} E(Q_q) &= \sum_{i=S+1}^{\infty} (i-S) p_i = p_0 \frac{a^{S-1}}{(S-1)!} \sum_{i=S+1}^{\infty} (i-S) \frac{\lambda^{i-S+1} g^*_{i-S}(S\mu)}{(i-S)!} \\ &= p_{S-1} \sum_{i=0}^{\infty} \frac{\lambda^{i+2} g^*_{i+1}(S\mu)}{i!} \\ &= \lambda p_{S-1} \int_0^{\infty} \lambda H(t) e^{-S\mu t} \sum_{i=0}^{\infty} \frac{[\lambda H(t)]^i}{i!} dt \\ &= \frac{\lambda^2 J_H}{\epsilon + \lambda J} \end{aligned}$$

which is the same as the result obtained in [50] using Little's formula.

Similarly for the finite buffer $M/M/S/N+G$ model, Movaghar [36] obtained the solution

$$p_i = \begin{cases} p_0 \frac{a^i}{i!} & \text{if } 0 \le i \le S \\ p_0 \frac{a^{S-1}}{(S-1)!} \frac{\lambda^{i-S+1} g^*_{i-S}(S\mu)}{(i-S)!} & \text{if } S \le i \le N \end{cases}, \tag{6.8}$$

where

$$p_0^{-1} = \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda}{(S-1)!} \sum_{i=S}^{N} \frac{\lambda^{i-S} g^*_{i-S}(S\mu)}{(i-S)!}.$$

Performance measures in terms of building blocks

In this section, following the idea of building blocks of performance measures for $M/M/S+G$ model in Zeltyn [50], here for $M/M/S/N+G$ model, we will introduce, for $N \ge S$,

$$\begin{aligned} J^{(N)} &:= \sum_{i=S}^{N} \frac{\lambda^{i-S} g^*_{i-S}(S\mu)}{(i-S)!} \qquad (6.9) \\ &= \int_0^\infty e^{-S\mu t} \sum_{i=0}^{N-S} \frac{[\lambda H(t)]^i}{i!} dt \\ &= \int_0^\infty e^{-S\mu t} e^{\lambda H(t)} [1 - P(N-S+1, \lambda H(t))] dt \\ &= \int_0^\infty e^{\lambda H(t) - S\mu t} dt - \int_0^\infty e^{\lambda H(t) - S\mu t} P(N-S+1, \lambda H(t)) dt \\ &= J - \frac{\int_0^\infty e^{\lambda H(t) - S\mu t} \gamma(N-S+1, \lambda H(t)) dt}{(N-S)!} \\ &= J - \frac{\int_0^\infty e^{\lambda H(t) - S\mu t} \left[\int_0^{\lambda H(t)} x^{N-S} e^{-x} dx \right] dt}{(N-S)!} \\ &= J - \frac{\int_0^\eta x^{N-S} e^{-x} J(H^{-1}(x/\lambda)) dx}{(N-S)!}, \qquad (6.10) \end{aligned}$$

where $H^{-1}(x)$ is the inverse function of $H(x)$ and $H^{-1}(x)$ exists since $H(x)$ is an increasing function. Then we have the following new expressions in terms of $J^{(N)}$,

$$p_0^{-1} = \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda J^{(N)}}{(S-1)!}, \tag{6.11}$$

$$p_{S-1} = p_0 \frac{a^{S-1}}{(S-1)!} = \frac{\frac{a^{S-1}}{(S-1)!}}{\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda J^{(N)}}{(S-1)!}} = \frac{1}{\epsilon + \lambda J^{(N)}} \tag{6.12}$$

and

$$p_S = p_0 \frac{a^S}{S!} = \frac{\frac{a^S}{S!}}{\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda J^{(N)}}{(S-1)!}} = \frac{\rho}{\epsilon + \lambda J^{(N)}}. \tag{6.13}$$

Also, by the PASTA property, we have the probability of delay among all calls (abandoned, served or blocked)

$$P(delay) = \sum_{i=S}^{N-1} p_i = p_0 \frac{a^{S-1}}{(S-1)!} \sum_{i=S}^{N-1} \frac{\lambda^{i-S+1} g_{i-S}^*(S\mu)}{(i-S)!}$$
$$= p_{S-1} \lambda J^{(N-1)} = \frac{\lambda J^{(N-1)}}{\epsilon + \lambda J^{(N)}}, \tag{6.14}$$

the probability that the call get service without delay

$$P(\text{no-delay}) = \sum_{i=0}^{S-1} p_i = \sum_{i=0}^{S-1} p_0 \frac{a^i}{i!} = \frac{\sum_{i=0}^{S-1} \frac{a^i}{i!}}{\sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1} \lambda J^{(N)}}{(S-1)!}} = \frac{\epsilon}{\epsilon + \lambda J^{(N)}} \tag{6.15}$$

and the blocking probability

$$P(blocking) = p_N = 1 - \frac{\epsilon}{\epsilon + \lambda J^{(N)}} - \frac{\lambda J^{(N-1)}}{\epsilon + \lambda J^{(N)}} = \frac{\lambda (J^{(N)} - J^{(N-1)})}{\epsilon + \lambda J^{(N)}}, \tag{6.16}$$

since $P(blocking) = 1 - P(no - delay) - P(delay)$.

We can also express three kinds of mean number of calls in terms of $J^{(N)}$.

1. Mean number of busy servers $E(Q_b)$

$$E(Q_b) = \sum_{i=0}^{S-1} i p_i + S \sum_{i=S}^{N} p_i$$
$$= a p_0 \sum_{i=0}^{S-2} \frac{a^i}{i!} + S(1 - \sum_{i=0}^{S-2} p_0 \frac{a^i}{i!} - p_{S-1})$$
$$= (a - S) \sum_{i=0}^{S-2} p_0 \frac{a^i}{i!} + S(1 - p_{S-1})$$
$$= (a - S)(\sum_{i=0}^{S-1} p_i - p_{S-1}) + S(1 - p_{S-1})$$
$$= \frac{a(\epsilon - 1) + S\lambda J^{(N)}}{\epsilon + \lambda J^{(N)}}, \tag{6.17}$$

   where we have used (6.12) and (6.15).

2. Mean number of calls waiting in the queue $E(Q_q)$

$$E(Q_q) = \sum_{i=S+1}^{N} (i - S) p_i = p_0 \frac{a^{S-1}}{(S-1)!} \sum_{i=S+1}^{N} (i - S) \frac{\lambda^{i-S+1} g_{i-S}^*(S\mu)}{(i-S)!}$$
$$= p_{S-1} \sum_{i=0}^{N-S-1} \frac{\lambda^{i+2} g_{i+1}^*(S\mu)}{i!}$$
$$= \frac{\lambda^2 J_H^{(N-1)}}{\epsilon + \lambda J^{(N)}}, \tag{6.18}$$

where $J_H^{(N)}$ for $N \geq S$ is defined to be

$$\begin{aligned}
J_H^{(N)} &:= \sum_{i=0}^{N-S} \frac{\lambda^i g_{i+1}^*(S\mu)}{i!} \\
&= \int_0^\infty H(t) e^{-S\mu t} \sum_{i=0}^{N-S} \frac{[\lambda H(t)]^i}{i!} dt \\
&= \int_0^\infty H(t) e^{-S\mu t} e^{\lambda H(t)} [1 - P(N-S+1, \lambda H(t))] dt \\
&= \int_0^\infty H(t) e^{\lambda H(t) - S\mu t} dt - \int_0^\infty H(t) e^{\lambda H(t) - S\mu t} P(N-S+1, \lambda H(t)) dt \\
&= J_H - \frac{\int_0^\infty H(t) e^{\lambda H(t) - S\mu t} \gamma(N-S+1, \lambda H(t)) dt}{(N-S)!} \\
&= J_H - \frac{\int_0^\infty H(t) e^{\lambda H(t) - S\mu t} \left[\int_0^{\lambda H(t)} x^{N-S} e^{-x} dx \right] dt}{(N-S)!} \\
&= J_H - \frac{\int_0^\eta x^{N-S} e^{-x} J_H(H^{-1}(x/\lambda)) dx}{(N-S)!}.
\end{aligned} \tag{6.19}$$

3. Mean number of calls in the system $E(Q)$

$$E(Q) = E(Q_b) + E(Q_q).$$

6.2.2 Probability of abandonment

As in $M/M/S/N+M$ model, since there are some blocking calls and only non-blocking calls can abandon, we have $P(Ab) = P(Ab, \text{non-blocking})$. To derive $P(Ab, \text{non-blocking})$ we first condition on the state seen by a non-blocking call and then sum up all the possibilities. Define $P_i(Ab) = P_i(Ab, \text{non-blocking}) = P(\text{the non-blocking call will abandon} \mid i \text{ calls in the system upon arrival})$. Movaghar [36] showed that

$$P_i(Ab) = P(V_i > X) = \frac{r_{i+1}}{S\mu + r_{i+1}} = \begin{cases} 0 & 0 \leq i < S \\ 1 - \frac{S\mu}{i-S+1} \frac{g_{i-S+1}^*(S\mu)}{g_{i-S}^*(S\mu)} & S \leq i < N \end{cases} \tag{6.20}$$

and

$$P_i(Sr) = P_i(Sr, \text{non-blocking}) = \begin{cases} 1 & 0 \leq i < S \\ \frac{S\mu}{i-S+1} \frac{g_{i-S+1}^*(S\mu)}{g_{i-S}^*(S\mu)} & S \leq i < N \end{cases}.$$

Now

$$P(Ab) = P(Ab, \text{non-blocking}) = \sum_{i=S}^{N-1} P_i(Ab) P(i \text{ calls in the system upon arrival})$$

$$= \sum_{i=S}^{N-1} \left[1 - \frac{S\mu}{i-S+1} \frac{g^*_{i-S+1}(S\mu)}{g^*_{i-S}(S\mu)}\right] a_i$$

$$= \sum_{i=S}^{N-1} \left[1 - \frac{S\mu}{i-S+1} \frac{g^*_{i-S+1}(S\mu)}{g^*_{i-S}(S\mu)}\right] p_i \tag{6.21}$$

$$= \sum_{i=S}^{N-1} \frac{r_{i+1}}{S\mu + r_{i+1}} p_i \tag{6.22}$$

where we have used the PASTA property, i.e., $a_i = p_i$. Similarly,

$$P(Sr) = P(Sr, \text{non-blocking})$$

$$= \sum_{i=0}^{S-1} p_i + \sum_{i=S}^{N-1} \frac{S\mu}{i-S+1} \frac{g^*_{i-S+1}(S\mu)}{g^*_{i-S}(S\mu)} p_i. \tag{6.23}$$

Performance measures in terms of building blocks

In the following we will derive expressions of the above performance measures in terms of $J^{(N)}$ using another method. By Little's formula for Q_b, we have $E(Q_b) = \lambda P(Sr)\frac{1}{\mu}$, which, combining (6.17), gives us

$$P(Sr) = \frac{E(Q_b)}{a} = \frac{\epsilon - 1 + S\mu J^{(N)}}{\epsilon + \lambda J^{(N)}}. \tag{6.24}$$

This formula can also be obtained directly from (6.23). Therefore we have

$$P(Ab) = P(Ab, \text{non-blocking}) = 1 - P(blocking) - P(Sr)$$

$$= 1 - \frac{\lambda(J^{(N)} - J^{(N-1)})}{\epsilon + \lambda J^{(N)}} - \frac{\epsilon - 1 + S\mu J^{(N)}}{\epsilon + \lambda J^{(N)}}$$

$$= \frac{1 + \lambda J^{(N-1)} - S\mu J^{(N)}}{\epsilon + \lambda J^{(N)}}. \tag{6.25}$$

Comparing the above with $E(Q_q)$ (6.18), we find that the rate balance equation

$$\lambda P(Ab) = \alpha \cdot E(Q_q)$$

is usually not true unless the patience time has exponential distribution (i.e., $M/M/S/N + M$ model). In this case, later we will show that (refer to (6.45))

$$J_H^{(N-1)} = \frac{1 + \lambda J^{(N-1)} - S\mu J^{(N)}}{\lambda\alpha}$$

so that we have $\lambda P(Ab) = \alpha \cdot E(Q_q)$ holds for $M/M/S/N+M$ model.

In general, for $M/M/S/N+G$ model since $1-P(blocking) = P(Ab)+P(Sr)$, we have the rate conservative equation in equilibrium

$$\begin{aligned}\lambda\left[1-P(blocking)\right] &= \lambda P(Ab) + \lambda P(Sr) \\ &= \lambda P(Ab) + \mu E(Q_b),\end{aligned}$$

which shows the rate of non-blocking calls equals to the sum of total abandonment and service rates in equilibrium. Again we have the carried load

$$a' = E(Q_b) = aP(Sr)$$

and the utilization

$$\begin{aligned}v = \frac{a'}{S} &= \frac{aP(Sr)}{S} = \rho P(Sr) \\ &= \frac{\rho(\epsilon-1)+\lambda J^{(N)}}{\epsilon+\lambda J^{(N)}} < 1,\end{aligned}$$

which is the proportion of time that a CSR is busy.

6.2.3 Waiting time distribution

As in $M/M/S/N+M$ model, for performance measures related to waiting time, we are mainly interested in W_q: the conditional waiting time of a call until abandonment or starting to get service given that this call is not blocked. It is obvious that

$$P(W_q > t) = P(V \wedge X > t) = P(V > t)\overline{G}(t)$$

where V is the conditional offered waiting time of an infinite patient call given the call is not blocked, which is independent of the patience time X. Movaghar [36] used the same method as we have done in $M/M/S/N+M$ model to obtain the distribution of V

$$P(V > t) = \sum_{i=S}^{N-1} q_i P(V_i > t) = \frac{\sum_{i=S}^{N-1} p_i P(V_i > t)}{1-p_N} \tag{6.26}$$

where $q_i = \frac{p_i}{1-p_N}$ is the arrival-point probability mentioned in Chapter 2. Movaghar [36] proved the density of V_i (6.5) so that for $S \le i < N$,

$$\begin{aligned}P(V_i > t) &= \int_t^\infty \frac{g_{i-S}(u)}{g^*_{i-S}(S\mu)} e^{-S\mu u} du \\ &= \frac{1}{g^*_{i-S}(S\mu)} \int_t^\infty H(u)^{i-S} e^{-S\mu u} du.\end{aligned} \tag{6.27}$$

Performance measures in terms of building blocks

In the following we will express performance measures in terms of new building blocks. We have, by (6.8), (6.12), (6.16), (6.26) and (6.27),

$$\begin{aligned} P(V>t) &= \frac{\sum_{i=S}^{N-1} p_i P(V_i>t)}{1-p_N} \\ &= \frac{1}{1-p_N}\sum_{i=S}^{N-1}\frac{p_i}{g^*_{i-S}(S\mu)}\int_t^\infty H(u)^{i-S}e^{-S\mu u}du \\ &= \frac{\lambda p_{S-1}}{1-p_N}\int_t^\infty e^{-S\mu u}\sum_{i=S}^{N-1}\frac{[\lambda H(u)]^{i-S}}{(i-S)!}du \\ &= \frac{\lambda J^{(N-1)}(t)}{\epsilon+\lambda J^{(N-1)}}, \end{aligned} \tag{6.28}$$

where we have defined, for $N \geq S$,

$$\begin{aligned} J^{(N)}(t) &:= \int_t^\infty e^{-S\mu u}\sum_{i=S}^{N}\frac{[\lambda H(u)]^{i-S}}{(i-S)!}du \\ &= \int_t^\infty e^{-S\mu u}e^{\lambda H(u)}(1-P(N-S+1,\lambda H(u))du \\ &= \int_t^\infty e^{\lambda H(u)-S\mu u}du - \int_t^\infty e^{\lambda H(u)-S\mu u}P(N-S+1,\lambda H(u))du \\ &= J(t) - \int_t^\infty e^{\lambda H(u)-S\mu u}P(N-S+1,\lambda H(u))du \qquad (6.29) \\ &= J(t) - \frac{J(t)\int_0^{\lambda H(t)}x^{N-S}e^{-x}dx + \int_{\lambda H(t)}^{\eta}x^{N-S}e^{-x}J(H^{-1}(x/\lambda))dx}{(N-S)!} \\ &= J(t)\left[1-P(N-S+1,\lambda H(t))\right] - \frac{\int_{\lambda H(t)}^{\eta}x^{N-S}e^{-x}J(H^{-1}(x/\lambda))dx}{(N-S)!} \qquad (6.30) \\ &= J(t)\left[1-P(N-S+1,\lambda H(t))\right] - \frac{\int_t^\infty [\lambda H(x)]^{N-S}e^{-\lambda H(x)}J(x)\lambda\overline{G}(x)dx}{(N-S)!}. \end{aligned}$$

Hence

$$P(W_q>t) = P(V>t)\overline{G}(t) = \frac{\lambda\overline{G}(t)J^{(N-1)}(t)}{\epsilon+\lambda J^{(N-1)}} \tag{6.31}$$

and the density of V for $t>0$ is, by (6.29),

$$\begin{aligned} f_V(t) &= -\frac{\lambda}{\epsilon+\lambda J^{(N-1)}}\frac{dJ^{(N-1)}(t)}{dt} \\ &= -\frac{\lambda}{\epsilon+\lambda J^{(N-1)}}\frac{d\left[J(t)-\int_t^\infty e^{\lambda H(u)-S\mu u}P(N-S,\lambda H(u))du\right]}{dt} \\ &= \frac{\lambda}{\epsilon+\lambda J^{(N-1)}}\left[e^{\lambda H(t)-S\mu t} - e^{\lambda H(t)-S\mu t}P(N-S,\lambda H(t))\right] \\ &= \frac{\lambda e^{\lambda H(t)-S\mu t}}{\epsilon+\lambda J^{(N-1)}}\left[1-P(N-S,\lambda H(t))\right]. \end{aligned} \tag{6.32}$$

We now consider the waiting time distribution for abandonment calls given non-blocking, by (6.32) and (6.28),

$$\begin{aligned}
P(W_q{>}t, Ab) &= P(V \wedge X > t, V > X) \\
&= P(X > t, V > X) \\
&= \int_t^\infty P(V > x) dG(x) \\
&= P(V > x)G(x)|_t^\infty - \int_t^\infty G(x) dP(V > x) \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \int_t^\infty G(x) e^{\lambda H(x) - S\mu x} \left[1 - P(N - S, \lambda H(x))\right] dx - P(V > t)G(t) \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \int_t^\infty G(x) e^{\lambda H(x) - S\mu x} \left[1 - P(N - S, \lambda H(x))\right] dx - \frac{\lambda J^{(N-1)}(t)}{\epsilon + \lambda J^{(N-1)}} G(t) \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \left[\int_t^\infty G(x) e^{\lambda H(x) - S\mu x} \left[1 - P(N - S, \lambda H(x))\right] dx - J^{(N-1)}(t)G(t)\right] \\
&= \frac{\lambda \left[J_G^{(N-1)}(t) - J^{(N-1)}(t)G(t)\right]}{\epsilon + \lambda J^{(N-1)}},
\end{aligned} \tag{6.33}$$

where similarly as (6.30) we have defined, for $N \geq S$,

$$\begin{aligned}
J_G^{(N)}(t) &:= \int_t^\infty G(x) e^{-S\mu x} \sum_{i=S}^{N} \frac{[\lambda H(x)]^{i-S}}{(i-S)!} dx \\
&= \int_t^\infty G(x) e^{\lambda H(x) - S\mu x} \left[1 - P(N - S + 1, \lambda H(x))\right] dx \\
&= \int_t^\infty G(x) e^{\lambda H(x) - S\mu x} dx - \int_t^\infty G(x) e^{\lambda H(x) - S\mu x} P(N - S + 1, \lambda H(x)) dx \\
&= J_G(t) \left[1 - P(N - S + 1, \lambda H(t))\right] - \frac{\int_{\lambda H(t)}^{\eta} x^{N-S} e^{-x} J_G(H^{-1}(x/\lambda)) dx}{(N-S)!}
\end{aligned}$$

and

$$\begin{aligned}
J_G(t) &:= \int_t^\infty G(x) e^{\lambda H(x) - S\mu x} dx \\
&= \frac{(\lambda - S\mu) J(t) + e^{\lambda H(t) - S\mu t}}{\lambda}.
\end{aligned}$$

The last equality comes from

$$\int_t^\infty \left[\lambda \overline{G}(x) - S\mu\right] e^{\lambda H(x) - S\mu x} dx = \int_t^\infty de^{\lambda H(x) - S\mu x} = -e^{\lambda H(t) - S\mu t}.$$

Therefore by (6.30), we have

$$
\begin{aligned}
&J_G^{(N)}(t) \\
&= J_G(t)\left[1 - P(N-S+1, \lambda H(t))\right] - \frac{\int_{\lambda H(t)}^{\eta} x^{N-S} e^{-x} J_G(H^{-1}(x/\lambda))dx}{(N-S)!} \\
&= \frac{(\lambda - S\mu)J(t) + e^{\lambda H(t) - S\mu t}}{\lambda}\left[1 - P(N-S+1, \lambda H(t))\right] - \\
&\frac{\int_{\lambda H(t)}^{\eta} x^{N-S} e^{-x}\left[(\lambda - S\mu)J(H^{-1}(x/\lambda)) + e^{x - S\mu H^{-1}(x/\lambda)}\right]dx}{\lambda(N-S)!} \\
&= \frac{(\lambda - S\mu)J(t) + e^{\lambda H(t) - S\mu t}}{\lambda}\left[1 - P(N-S+1, \lambda H(t))\right] - \\
&\frac{(\lambda - S\mu)\int_{\lambda H(t)}^{\eta} x^{N-S} e^{-x} J(H^{-1}(x/\lambda))dx + \int_{\lambda H(t)}^{\eta} x^{N-S} e^{-S\mu H^{-1}(x/\lambda)}dx}{\lambda(N-S)!} \\
&= \frac{(\lambda - S\mu)J^{(N)}(t) + e^{\lambda H(t) - S\mu t}\left[1 - P(N-S+1, \lambda H(t))\right]}{\lambda} - \frac{\int_{\lambda H(t)}^{\eta} x^{N-S} e^{-S\mu H^{-1}(x/\lambda)}dx}{\lambda(N-S)!} \\
&= \frac{(\lambda - S\mu)J^{(N)}(t) + e^{\lambda H(t) - S\mu t}\left[1 - P(N-S+1, \lambda H(t))\right]}{\lambda} - \frac{\int_t^{\infty} \left[\lambda H(x)\right]^{N-S} e^{-S\mu x}\overline{G}(x)dx}{(N-S)!} \\
&= \frac{(\lambda - S\mu)J^{(N)}(t) + e^{\lambda H(t) - S\mu t}\left[1 - P(N-S+1, \lambda H(t))\right]}{\lambda} \\
&- J^{(N)}(t) + J^{(N-1)}(t) + J_G^{(N)}(t) - J_G^{(N-1)}(t)
\end{aligned}
$$

so that $J_G^{(N-1)}(t)$ can be expressed in terms of $J^{(N-1)}(t)$ and $J^{(N)}(t)$:

$$
J_G^{(N-1)}(t) = \frac{\lambda J^{(N-1)}(t) - S\mu J^{(N)}(t) + e^{\lambda H(t) - S\mu t}\left[1 - P(N-S+1, \lambda H(t))\right]}{\lambda} \tag{6.34}
$$

and

$$
J_G^{(N-1)} := J_G^{(N-1)}(0) = \frac{\lambda J^{(N-1)} - S\mu J^{(N)} + 1}{\lambda}.
$$

We now have, from (6.33),

$$
P(W_q > 0, Ab) = \frac{\lambda J_G^{(N-1)}}{\epsilon + \lambda J^{(N-1)}}.
$$

However by definition $P(W_q > 0, Ab) = P(Ab|\text{non-blocking})$. Therefore we have

$$
\begin{aligned}
P(Ab) &= P(Ab|\text{non-blocking})[1 - P(blocking)] \\
&= P(W_q > 0, Ab)\left[1 - P(blocking)\right] \\
&= \frac{\lambda J_G^{(N-1)}}{\epsilon + \lambda J^{(N-1)}} \frac{\epsilon + \lambda J^{(N-1)}}{\epsilon + \lambda J^{(N)}} \\
&= \frac{\lambda J^{(N-1)} - S\mu J^{(N)} + 1}{\epsilon + \lambda J^{(N)}},
\end{aligned}
$$

which is the same as (6.25).

The waiting time distribution for served calls given non-blocking is

$$\begin{aligned} P(W_q{>}t, Sr) &= P(W_q{>}t) - P(W_q{>}t, Ab) \\ &= \frac{\lambda\overline{G}(t)J^{(N-1)}(t)}{\epsilon+\lambda J^{(N-1)}} - \frac{\lambda\left[J_G^{(N-1)}(t) - J^{(N-1)}(t)G(t)\right]}{\epsilon+\lambda J^{(N-1)}} \\ &= \frac{\lambda\left[J^{(N-1)}(t) - J_G^{(N-1)}(t)\right]}{\epsilon+\lambda J^{(N-1)}}. \end{aligned} \tag{6.35}$$

Therefore

$$\begin{aligned} P(W_q{>}0, Sr) =& P(Sr, delay|\text{non-blocking}) \\ &= \frac{\lambda\left[J^{(N-1)} - J_G^{(N-1)}\right]}{\epsilon+\lambda J^{(N-1)}}. \end{aligned}$$

Now we can derive the conditional waiting, which are more useful in practice, using the above results.

1.

$$\begin{aligned} P(W_q > t|Ab) &= \frac{P(W_q > t, Ab)}{P(Ab|\text{non-blocking})} \\ &= \frac{\frac{\lambda\left[J_G^{(N-1)}(t) - J^{(N-1)}(t)G(t)\right]}{\epsilon+\lambda J^{(N-1)}}}{\frac{\lambda J_G^{(N-1)}}{\epsilon+\lambda J^{(N-1)}}} = \frac{J_G^{(N-1)}(t) - J^{(N-1)}(t)G(t)}{J_G^{(N-1)}} \\ &= \frac{\lambda\left[J_G^{(N-1)}(t) - J^{(N-1)}(t)G(t)\right]}{1+\lambda J^{(N-1)} - S\mu J^{(N)}}. \end{aligned}$$

2.

$$\begin{aligned} P(W_q > t|Sr) &= \frac{P(W_q > t, Sr)}{P(Sr|\text{non-blocking})} \\ &= \frac{\frac{\lambda\left[J^{(N-1)}(t) - J_G^{(N-1)}(t)\right]}{\epsilon+\lambda J^{(N-1)}}}{\frac{\epsilon-1+S\mu J^{(N)}}{\epsilon+\lambda J^{(N-1)}}} = \frac{\lambda\left[J^{(N-1)}(t) - J_G^{(N-1)}(t)\right]}{\epsilon-1+S\mu J^{(N)}}. \end{aligned} \tag{6.36}$$

3.

$$\begin{aligned} P(W_q > t|Sr, delay) &= \frac{P(W_q > t, Sr)}{P(Sr, delay|\text{non-blocking})} \\ &= \frac{\frac{\lambda\left[J^{(N-1)}(t) - J_G^{(N-1)}(t)\right]}{\epsilon+\lambda J^{(N-1)}}}{\frac{\lambda\left[J^{(N-1)} - J_G^{(N-1)}\right]}{\epsilon+\lambda J^{(N-1)}}} = \frac{J^{(N-1)}(t) - J_G^{(N-1)}(t)}{J^{(N-1)} - J_G^{(N-1)}}. \end{aligned}$$

6.2.4 Mean waiting time

In this section, we will express various mean waiting time in terms of building blocks. By (6.31), we have the mean waiting time for all calls given entry

$$\begin{aligned}
E(W_q) &= \int_0^\infty \frac{\lambda \overline{G}(t) J^{(N-1)}(t)}{\epsilon + \lambda J^{(N-1)}} dt \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \int_0^\infty \overline{G}(t) J^{(N-1)}(t) dt \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \int_0^\infty J^{(N-1)}(t) dH(t) \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \left[J^{(N-1)}(t) H(t) |_0^\infty - \int_0^\infty H(t) dJ^{(N-1)}(t) \right] \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \int_0^\infty H(t) e^{\lambda H(t) - S\mu t} \left[1 - P(N-S, \lambda H(t)) \right] dt \\
&= \frac{\lambda J_H^{(N-1)}}{\epsilon + \lambda J^{(N-1)}},
\end{aligned} \tag{6.37}$$

which can also be derived using Little's formula,

$$E(Q_q) = \lambda \left[1 - P(blocking) \right] E(W_q)$$

and then use (6.18) and (6.16).

By (6.35) and (6.34), we have the mean waiting time for served calls given entry

$$\begin{aligned}
E(W_q, Sr) &= \int_0^\infty \frac{\lambda \left[J^{(N-1)}(t) - J_G^{(N-1)}(t) \right]}{\epsilon + \lambda J^{(N-1)}} dt \\
&= \frac{\lambda}{\epsilon + \lambda J^{(N-1)}} \int_0^\infty \left[J^{(N-1)}(t) - J_G^{(N-1)}(t) \right] dt \\
&= \frac{1}{\epsilon + \lambda J^{(N-1)}} \int_0^\infty \left[S\mu J^{(N)}(t) - e^{\lambda H(t) - S\mu t} \left[1 - P(N-S+1, \lambda H(t)) \right] \right] dt \\
&= \frac{1}{\epsilon + \lambda J^{(N-1)}} \left[S\mu \int_0^\infty J^{(N)}(t) dt - \int_0^\infty e^{\lambda H(t) - S\mu t} \left[1 - P(N-S+1, \lambda H(t)) \right] dt \right] \\
&= \frac{S\mu J_1^{(N)} - J^{(N)}}{\epsilon + \lambda J^{(N-1)}},
\end{aligned}$$

where we have defined, for $N \geq S$,

$$\begin{aligned} J_1^{(N)} &:= \int_0^\infty J^{(N)}(t)dt \qquad (6.38) \\ &= \int_0^\infty \left[\int_t^\infty e^{\lambda H(u) - S\mu u}(1 - P(N-S+1, \lambda H(u))du \right] dt \\ &= \int_0^\infty u e^{\lambda H(u) - S\mu u} \left[1 - P(N-S+1, \lambda H(u))\right] du \\ &= \int_0^\infty u e^{\lambda H(u) - S\mu u} du - \int_0^\infty u e^{\lambda H(u) - S\mu u} P(N-S+1, \lambda H(u)) du \\ &= J_1 - \frac{\int_0^\eta x^{N-S} e^{-x} J_1(H^{-1}(x/\lambda))dx}{(N-S)!}. \qquad (6.39) \end{aligned}$$

The mean waiting time for abandoned calls given entry is

$$\begin{aligned} E(W_q, Ab) &= E(W_q) - E(W_q, Sr) \\ &= \frac{\lambda J_H^{(N-1)}}{\epsilon + \lambda J^{(N-1)}} - \frac{S\mu J_1^{(N)} - J^{(N)}}{\epsilon + \lambda J^{(N-1)}} \\ &= \frac{\lambda J_H^{(N-1)} - S\mu J_1^{(N)} + J^{(N)}}{\epsilon + \lambda J^{(N-1)}}. \end{aligned}$$

Now we can derive the conditional mean waiting time, which are more useful in practice, using the above results.

1.

$$\begin{aligned} E(W_q|Ab) &= \frac{E(W_q, Ab)}{P(Ab|\text{non-blocking})} \\ &= \frac{\frac{\lambda J_H^{(N-1)} - S\mu J_1^{(N)} + J^{(N)}}{\epsilon + \lambda J^{(N-1)}}}{\frac{\lambda J_G^{(N-1)}}{\epsilon + \lambda J^{(N-1)}}} = \frac{\lambda J_H^{(N-1)} - S\mu J_1^{(N)} + J^{(N)}}{\lambda J_G^{(N-1)}}. \end{aligned}$$

2.

$$\begin{aligned} E(W_q|Sr) &= \frac{E(W_q, Sr)}{P(Sr|\text{non-blocking})} \\ &= \frac{\frac{S\mu J_1^{(N)} - J^{(N)}}{\epsilon + \lambda J^{(N-1)}}}{\frac{\epsilon - 1 + S\mu J^{(N)}}{\epsilon + \lambda J^{(N-1)}}} = \frac{S\mu J_1^{(N)} - J^{(N)}}{\epsilon - 1 + S\mu J^{(N)}}. \end{aligned}$$

3.

$$\begin{aligned} E(W_q|Sr, delay) &= \frac{E(W_q, Sr)}{P(Sr, delay|\text{non-blocking})} \\ &= \frac{\frac{S\mu J_1^{(N)} - J^{(N)}}{\epsilon + \lambda J^{(N-1)}}}{\frac{\lambda \left[J^{(N-1)} - J_G^{(N-1)} \right]}{\epsilon + \lambda J^{(N-1)}}} = \frac{S\mu J_1^{(N)} - J^{(N)}}{\lambda \left[J^{(N-1)} - J_G^{(N-1)} \right]}. \end{aligned}$$

Comparing the results obtained in this section with Section 6.1, it can be shown that when $N \to \infty$, building blocks of $M/M/S/N+G$ model will approach to building blocks of $M/M/S+G$ model. For example, $J^{(N)}$ will approach to J and $J^{(N)}(t)$ will approach to $J(t)$. Therefore performance measures of $M/M/S/N+G$ model will approach to the corresponding performance measures of $M/M/S+G$ model.

6.2.5 $M/M/S/N+M$ model

We have studied $M/M/S/N+M$ model in Chapter 5. In this section we will show that the formulas for general abandonment will reduce to the results obtained before in Chapter 5 for exponential abandonment i.e., $\overline{G}(x) = e^{-\alpha x}$ and hence $H(x) = \alpha^{-1}(1-e^{-\alpha x})$. We will first derive some building blocks for $M/M/S/N+M$ model in the following.

$J^{(N)}$ will reduce to an expression involved with $D(N-S+1)$ which is, as defined in Chapter 5,

$$D(N-S+1) = \sum_{i=0}^{N-S} \rho_i = A - \Gamma(C+1)e^{\eta}\eta^{-C}P(C+N-S+1,\eta). \tag{6.40}$$

We will prove this result by first providing the following lemma.

Lemma 6.2.1 *For any $N \geq S$, we have*

$$\int_0^{\eta} x^{N-S}e^{-x}\gamma(C,\eta-x)dx = \Gamma(C)\Gamma(N-S+1)P(C+N-S+1,\eta).$$

Proof. By changing the order of integration we have for any $N \geq S$,

$$\int_0^{\eta} x^{N-S}e^{-x}\gamma(C,\eta-x)dx = \int_0^{\eta} x^{C-1}e^{-x}\gamma(N-S+1,\eta-x)dx. \tag{6.41}$$

We will prove the lemma by mathematical induction. For $N=S$, we have, by (6.41),

$$\begin{aligned}
\int_0^{\eta} e^{-x}\gamma(C,\eta-x)dx &= \int_0^{\eta} x^{C-1}e^{-x}\gamma(1,\eta-x)dx \\
&= \int_0^{\eta} x^{C-1}e^{-x}\left[1-e^{x-\eta}\right]dx \\
&= \int_0^{\eta} x^{C-1}e^{-x}dx - e^{-\eta}\eta^{C}C^{-1} \\
&= \gamma(C,\eta) - e^{-\eta}\eta^{C}C^{-1} \\
&= \frac{\gamma(C+1,\eta)}{C} = \Gamma(C)P(C+1,\eta).
\end{aligned}$$

Therefore the lemma is proved for $N = S$. Assuming that it is true for $N = K$, we will check for $N = K + 1$. Again by (6.41), we have

$$\begin{aligned}
&\int_0^\eta x^{K+1-S}e^{-x}\gamma(C,\eta-x)dx \\
&= \int_0^\eta x^{C-1}e^{-x}\gamma(K-S+2,\eta-x)dx \\
&= \int_0^\eta x^{C-1}e^{-x}\left[(K-S+1)\gamma(K-S+1,\eta-x)-(\eta-x)^{K-S+1}e^{x-\eta}\right]dx \\
&= (K-S+1)\int_0^\eta x^{C-1}e^{-x}\gamma(K-S+1,\eta-x)dx - \int_0^\eta x^{C-1}e^{-x}(\eta-x)^{K-S+1}e^{x-\eta}dx \\
&= (K-S+1)\int_0^\eta x^{K-S}e^{-x}\gamma(C,\eta-x)dx - e^{-\eta}\int_0^\eta x^{C-1}(\eta-x)^{K-S+1}dx.
\end{aligned}$$

Now using integration by parts continuously, we have

$$\begin{aligned}
\int_0^\eta x^{C-1}(\eta-x)^{K-S+1}dx &= \frac{K-S+1}{C}\int_0^\eta x^{C}(\eta-x)^{K-S}dx \\
&= \frac{K-S+1}{C}\frac{K-S}{C+1}\int_0^\eta x^{C+1}(\eta-x)^{K-S-1}dx \\
&= \frac{K-S+1}{C}\frac{K-S}{C+1}\cdots\frac{1}{C+K-S}\int_0^\eta x^{C+K-S}dx \\
&= \frac{\Gamma(C)\Gamma(K-S+2)}{\Gamma(C+K-S+2)}\eta^{C+K-S+1}. \qquad (6.42)
\end{aligned}$$

Hence using the above and the assumption, we obtain

$$\begin{aligned}
&\int_0^\eta x^{K+1-S}e^{-x}\gamma(C,\eta-x)dx \\
&= (K-S+1)\frac{\Gamma(C)\Gamma(K-S+1)\gamma(C+K-S+1,\eta)}{\Gamma(C+K-S+1)} - \frac{e^{-\eta}\Gamma(C)\Gamma(K-S+2)}{\Gamma(C+K-S+2)}\eta^{C+K-S+1} \\
&= \frac{\Gamma(K-S+2)\Gamma(C)\gamma(C+K-S+1,\eta)(C+K-S+1)-e^{-\eta}\Gamma(C)\Gamma(K-S+2)\eta^{C+K-S+1}}{\Gamma(C+K-S+2)} \\
&= \frac{\Gamma(K-S+2)\Gamma(C)\left[\gamma(C+K-S+1,\eta)(C+K-S+1)-e^{-\eta}\eta^{C+K-S+1}\right]}{\Gamma(C+K-S+2)} \\
&= \frac{\Gamma(C)\Gamma(K-S+2)\gamma(C+K-S+2,\eta)}{\Gamma(C+K-S+2)} \\
&= \Gamma(C)\Gamma(K-S+2)P(C+K-S+2,\eta),
\end{aligned}$$

which completes the proof. ∎

Now we have the following result about $J^{(N)}$ for $M/M/S/N+M$ model.

Theorem 6.2.1 *For $M/M/S/N+M$ model,*

$$J^{(N)} = \frac{D(N-S+1)}{S\mu}. \qquad (6.43)$$

Proof. By definition (6.9), the above lemma, the expression of $J(t)$ (6.2) and the fact that for exponential abandonment $H^{-1}(x/\lambda) = -\alpha^{-1}\ln(1-\frac{x}{\eta})$, we have

$$\begin{aligned} J^{(N)} &= J - \frac{\int_0^\eta x^{N-S}e^{-x}J(H^{-1}(x/\lambda))dx}{(N-S)!} \\ &= J - J\frac{\int_0^\eta x^{N-S}e^{-x}\gamma(C,\eta-x)dx}{(N-S)!\gamma(C,\eta)} \\ &= \frac{A}{S\mu} - \frac{A}{S\mu}\frac{P(C+N-S+1,\eta)\Gamma(C)}{\gamma(C,\eta)} \\ &= \frac{A - A\Gamma(C)\gamma(C,\eta)^{-1}P(C+N-S+1,\eta)}{S\mu} \\ &= \frac{A - Ce^{\eta}\eta^{-C}\Gamma(C)P(C+N-S+1,\eta)}{S\mu} \\ &= \frac{D(N-S+1)}{S\mu}, \end{aligned} \tag{6.44}$$

where we used (6.40). ∎

Next we will consider $J_H^{(N)}$. We have the following result for $M/M/S/N+M$ model.

Theorem 6.2.2 *For $M/M/S/N+M$ model,*

$$J_H^{(N)} = \frac{1+\lambda J^{(N)} - S\mu J^{(N+1)}}{\lambda\alpha} = \frac{1+\rho D(N-S+1) - D(N-S+2)}{\lambda\alpha}. \tag{6.45}$$

Proof. From the definition (6.19),

$$J_H^{(N)} = J_H - \frac{\int_0^\eta x^{N-S}e^{-x}J_H(H^{-1}(x/\lambda))dx}{(N-S)!}.$$

For exponential abandonment we have

$$J_H = \frac{1+(\lambda - S\mu)J}{\lambda\alpha}$$

and by (6.3),

$$\begin{aligned} J_H(H^{-1}(x/\lambda)) &= \frac{1}{\alpha}\left[J\frac{\gamma(C,\eta-x)}{\gamma(C,\eta)} - \frac{A}{\lambda C}\frac{\gamma(C+1,\eta-x)}{\gamma(C,\eta)}\right] \\ &= \frac{1}{\alpha}\left[J\frac{\gamma(C,\eta-x)}{\gamma(C,\eta)} - \frac{A}{\lambda C}\frac{C\gamma(C,\eta-x) - (\eta-x)^C e^{x-\eta}}{\gamma(C,\eta)}\right] \\ &= \frac{1}{\alpha}\left[(J-\frac{A}{\lambda})\frac{\gamma(C,\eta-x)}{\gamma(C,\eta)} + \frac{A}{\lambda C}\frac{(\eta-x)^C e^{x-\eta}}{\gamma(C,\eta)}\right]. \end{aligned}$$

Then by Lemma 6.2.1 and (6.42)

$$
\begin{aligned}
&\frac{\int_0^\eta x^{N-S}e^{-x}J_H(H^{-1}(x/\lambda))dx}{(N-S)!} \\
&= \frac{\int_0^\eta x^{N-S}e^{-x}\left[(J-\frac{A}{\lambda})\frac{\gamma(C,\eta-x)}{\gamma(C,\eta)}+\frac{A}{\lambda C}\frac{(\eta-x)^C e^{x-\eta}}{\gamma(C,\eta)}\right]dx}{\alpha(N-S)!} \\
&= \frac{(\lambda J-A)\int_0^\eta x^{N-S}e^{-x}\gamma(C,\eta-x)dx}{\lambda\alpha(N-S)!\gamma(C,\eta)}+\frac{Ae^{-\eta}\int_0^\eta x^{N-S}(\eta-x)^C dx}{\lambda\alpha C(N-S)!\gamma(C,\eta)} \\
&= \frac{(\lambda J-A)\Gamma(C)P(C+N-S+1,\eta)}{\lambda\alpha\gamma(C,\eta)}+\frac{A\Gamma(C+1)\eta^{C+N-S+1}}{\lambda\alpha C e^{\eta}\gamma(C,\eta)\Gamma(C+N-S+2)} \\
&= \frac{J(\lambda-S\mu)\Gamma(C)P(C+N-S+1,\eta)}{\lambda\alpha\gamma(C,\eta)}+\frac{\Gamma(C+1)\eta^{N-S+1}}{\lambda\alpha\Gamma(C+N-S+2)} \\
&= \frac{J(\lambda-S\mu)\left[S\mu J-D(N-S+1)\right]}{\lambda\alpha S\mu J}+\frac{\rho_{N-S+1}}{\lambda\alpha} \\
&= \frac{J(\lambda-S\mu)-(\rho-1)D(N-S+1)+\rho_{N-S+1}}{\lambda\alpha},
\end{aligned}
$$

where we have used the result derived from (6.44)

$$
\frac{\Gamma(C)P(C+N-S+1,\eta)}{\gamma(C,\eta)}=\frac{S\mu J-D(N-S+1)}{S\mu J}.
$$

Therefore

$$
\begin{aligned}
J_H^{(N)} &= \frac{1+(\lambda-S\mu)J}{\lambda\alpha}-\frac{J(\lambda-S\mu)-(\rho-1)D(N-S+1)+\rho_{N-S+1}}{\lambda\alpha} \\
&= \frac{1+(\rho-1)D(N-S+1)-\rho_{N-S+1}}{\lambda\alpha} \\
&= \frac{1+(\rho-1)D(N-S+1)-D(N-S+2)+D(N-S+1)}{\lambda\alpha} \\
&= \frac{1+\rho D(N-S+1)-D(N-S+2)}{\lambda\alpha} \\
&= \frac{1+\lambda J^{(N)}-S\mu J^{(N+1)}}{\lambda\alpha}.
\end{aligned}
$$

■

Using the above results for $J^{(N)}$ (6.43) and $J_H^{(N)}$ (6.45), we can derive performance measures related to queue length process and the probability of abandonment as shown above in Section 6.2.1 and Section 6.2.2 for $M/M/S/N+M$ model. It can be shown that these formulas (in terms of $\epsilon, J^{(N)}$ and $J_H^{(N)}$) are the same as formulas (in terms of functions B and D) we have obtained in chapter 5.

For performance measures related to waiting time, we find that the key building block

is $J^{(N)}(t)$ and for $M/M/S/N+M$ model by (6.30)

$$J^{(N)}(t) = J(t)\left[1 - P(N-S+1, \eta(1-e^{-\alpha t}))\right] - \frac{\int_{\eta(1-e^{-\alpha t})}^{\eta} x^{N-S} e^{-x} J(H^{-1}(x/\lambda))dx}{(N-S)!}$$
$$= J(t)\left[1 - P(N-S+1, \eta(1-e^{-\alpha t}))\right] - J\frac{\int_{\eta(1-e^{-\alpha t})}^{\eta} x^{N-S} e^{-x} \gamma(C, \eta - x)dx}{(N-S)!\gamma(C,\eta)}$$

and by Lemma 6.2.1

$$\int_{\eta(1-e^{-\alpha t})}^{\eta} x^{N-S} e^{-x} \gamma(C, \eta - x)dx$$
$$= \int_0^{\eta e^{-\alpha t}} \left[y + \eta(1-e^{-\alpha t})\right]^{N-S} e^{-\left[y+\eta(1-e^{-\alpha t})\right]} \gamma(C, \eta - \left[y + \eta(1-e^{-\alpha t})\right])dy$$
$$= e^{-\eta(1-e^{-\alpha t})} \int_0^{\eta e^{-\alpha t}} \left[y + \eta(1-e^{-\alpha t})\right]^{N-S} e^{-y} \gamma(C, \eta e^{-\alpha t} - y)dy$$
$$= e^{-\eta(1-e^{-\alpha t})} \int_0^{\eta e^{-\alpha t}} \sum_{i=0}^{N-S} \frac{(N-S)!}{i!(N-S-i)!} y^{N-S-i} \left[\eta(1-e^{-\alpha t})\right]^i e^{-y} \gamma(C, \eta e^{-\alpha t} - y)dy$$
$$= e^{-\eta(1-e^{-\alpha t})} \sum_{i=0}^{N-S} \frac{(N-S)!\left[\eta(1-e^{-\alpha t})\right]^i}{i!(N-S-i)!} \int_0^{\eta e^{-\alpha t}} y^{N-S-i} e^{-y} \gamma(C, \eta e^{-\alpha t} - y)dy$$
$$= e^{-\eta(1-e^{-\alpha t})}(N-S)!\Gamma(C) \sum_{i=0}^{N-S} \frac{\left[\eta(1-e^{-\alpha t})\right]^i P(C+N-S-i+1, \eta e^{-\alpha t})}{i!}$$

So that

$$J^{(N)}(t) = J(t)\left[1 - P(N-S+1, \eta(1-e^{-\alpha t}))\right]$$
$$- \frac{Je^{-\eta(1-e^{-\alpha t})}(N-S)!\Gamma(C)\sum_{i=0}^{N-S} \frac{\left[\eta(1-e^{-\alpha t})\right]^i P(C+N-S-i+1, \eta e^{-\alpha t})}{i!}}{(N-S)!\gamma(C,\eta)}$$
$$= J(t)\left[1 - P(N-S+1, \eta(1-e^{-\alpha t}))\right]$$
$$- \alpha^{-1}\eta^{-C}\Gamma(C)e^{\eta e^{-\alpha t}} \sum_{i=0}^{N-S} \frac{\left[\eta(1-e^{-\alpha t})\right]^i P(C+N-S-i+1, \eta e^{-\alpha t})}{i!}.$$

Other building blocks such as $J_1^{(N)}$ and $J_G^{(N-1)}(t)$ are all based on $J^{(N)}(t)$ and their expressions are too complex and we will not show them here. Plugging these building blocks to the formulas in Section 6.2.3 and Section 6.2.4, we have performance measures related to the waiting time, which are not given explicitly in Chapter 5 since there we have used the arrival-point probability q_i and $A_{n,i-S}$ to express performance measures.

6.2.6 $M/M/S/N+D$ model

$M/M/S/N+D$ model (deterministic patience time) is another important special case of $M/M/S/N+G$ model. In this case we have that the patience time $X = \alpha^{-1}$ is a constant

and

$$G(t) = \begin{cases} 0, & t < \alpha^{-1} \\ 1, & t \geq \alpha^{-1} \end{cases}.$$

Hence for $t \geq 0$,

$$H(t) = \int_0^t \overline{G}(u)du = \begin{cases} t, & 0 \leq t < \alpha^{-1} \\ \alpha^{-1}, & t \geq \alpha^{-1} \end{cases}$$

and

$$g_i(t) = H(t)^i = \begin{cases} t^i, & 0 \leq t < \alpha^{-1} \\ \alpha^{-i}, & t \geq \alpha^{-1} \end{cases},$$

so that

$$\begin{aligned} g_i^*(S\mu) &= \int_0^\infty e^{-S\mu t} g_i(t)dt = \int_0^{\alpha^{-1}} e^{-S\mu t} g_i(t)dt + \int_{\alpha^{-1}}^\infty e^{-S\mu t} g_i(t)dt \\ &= \int_0^{\alpha^{-1}} e^{-S\mu t} t^i dt + \alpha^{-i} \int_{\alpha^{-1}}^\infty e^{-S\mu t} dt \\ &= \frac{1}{(S\mu)^{i+1}} \gamma(i+1, C) + \frac{\alpha^{-i}}{S\mu} e^{-C} \\ &= \frac{1}{(S\mu)^{i+1}} \left[i\gamma(i, C) - C^i e^{-C} \right] + \frac{\alpha^{-i}}{S\mu} e^{-C} \\ &= \frac{i\gamma(i, C)}{(S\mu)^{i+1}} = \frac{i! P(i, C)}{(S\mu)^{i+1}} \end{aligned}$$

for $i \geq 0$, where we define $P(0, C) = 1$. The distribution of queue length Q is, according to (6.8),

$$p_i = \begin{cases} p_0 \frac{a^i}{i!} & \text{if } 0 \leq i \leq S \\ p_{S-1} \rho^{i-S+1} P(i-S, C) & \text{if } S < i \leq N \end{cases},$$

where

$$p_0^{-1} = \sum_{i=0}^{S} \frac{a^i}{i!} + \frac{a^{S-1}}{(S-1)!} \sum_{i=S+1}^{N} \rho^{i-S+1} P(i-S, C).$$

Comparing with the general formula

$$p_0^{-1} = \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1} \lambda J^{(N)}}{(S-1)!},$$

we have for $M/M/S/N + D$ model

$$J^{(N)} = \frac{1 + \sum_{i=1}^{N-S} \rho^i P(i, C)}{S\mu} = \frac{\sum_{i=0}^{N-S} \rho^i P(i, C)}{S\mu}$$

which is an alternative expression of $J^{(N)}$ obtained in the following. To obtain all the performance measures, we only need to give the expressions of all building blocks. Zeltyn [50] give the expressions of building blocks for $M/M/S+D$ model as follows.

If $\rho \neq 1$ i.e., $\lambda \neq S\mu$

$$
\begin{aligned}
J &= \frac{\rho e^{\eta-C}-1}{\lambda-S\mu}, \\
J_H &= \frac{1}{(\lambda-S\mu)^2} + \frac{[\eta(\rho-1)-1]\,e^{\eta-C}}{(\lambda-S\mu)^2}, \\
J_1 &= \frac{1}{(\lambda-S\mu)^2} + \left[\frac{\eta(\rho-1)-1}{(\lambda-S\mu)^2} + \frac{1}{(S\mu)^2}\right] e^{\eta-C} = J_H + \frac{e^{\eta-C}}{(S\mu)^2}, \\
J(t) &= \begin{cases} \frac{\rho e^{\eta-C}-e^{(\lambda-S\mu)t}}{\lambda-S\mu}, & 0 \le t < \alpha^{-1} \\ \frac{e^{\eta-S\mu t}}{S\mu}, & t \ge \alpha^{-1} \end{cases}, \\
J_H(t) &= \begin{cases} \frac{e^{(\lambda-S\mu)t}-e^{\eta-C}}{(\lambda-S\mu)^2} - \frac{te^{(\lambda-S\mu)t}}{\lambda-S\mu} + \frac{\rho e^{\eta-C}}{\alpha(\lambda-S\mu)}, & 0 \le t < \alpha^{-1} \\ \frac{e^{\eta-S\mu t}}{S\mu}, & t \ge \alpha^{-1} \end{cases}, \\
J_1(t) &= \begin{cases} \frac{e^{(\lambda-S\mu)t}-e^{\eta-C}}{(\lambda-S\mu)^2} - \frac{te^{(\lambda-S\mu)t}}{\lambda-S\mu} + \left[\frac{\rho}{\alpha(\lambda-S\mu)} + \frac{1}{(S\mu)^2}\right] e^{\eta-C}, & 0 \le t < \alpha^{-1} \\ \left[\frac{t}{S\mu} + \frac{1}{(S\mu)^2}\right] e^{\eta-S\mu t}, & t \ge \alpha^{-1} \end{cases}.
\end{aligned}
$$

If $\rho = 1$ i.e., $\lambda = S\mu$

$$
\begin{aligned}
J &= \frac{1}{\alpha} + \frac{1}{\lambda}, \\
J_H &= \frac{1}{2\alpha^2} + \frac{1}{\lambda\alpha}, \\
J_1 &= \frac{1}{2\alpha^2} + \frac{1}{\lambda\alpha} + \frac{1}{\lambda^2}, \\
J(t) &= \begin{cases} \frac{1}{\alpha} + \frac{1}{\lambda} - t, & 0 \le t < \alpha^{-1} \\ \frac{e^{\eta-\lambda t}}{\lambda}, & t \ge \alpha^{-1} \end{cases}, \\
J_H(t) &= \begin{cases} \frac{1-\alpha^2 t^2}{2\alpha^2} + \frac{1}{\lambda\alpha}, & 0 \le t < \alpha^{-1} \\ \frac{e^{\eta-\lambda t}}{\lambda}, & t \ge \alpha^{-1} \end{cases}, \\
J_1(t) &= \begin{cases} \frac{1-\alpha^2 t^2}{2\alpha^2} + \frac{1}{\lambda\alpha} + \frac{1}{\lambda^2}, & 0 \le t < \alpha^{-1} \\ \left[\frac{t}{\lambda} + \frac{1}{\lambda^2}\right] e^{\eta-\lambda t}, & t \ge \alpha^{-1} \end{cases}.
\end{aligned}
$$

To obtain the performance measures for $M/M/S/N+D$ model, we need further to give other building blocks based on the above. Now we have $H^{-1}(t) = t$ for $0 \le t < \alpha^{-1}$.

Hence for $0 \leq x < \eta$,

$$J(H^{-1}(x/\lambda)) = J(x/\lambda) = \begin{cases} \frac{\rho e^{\eta - C} - e^{(1-\rho^{-1})x}}{\lambda - S\mu}, & \rho \neq 1 \\ \frac{1}{\alpha} + \frac{1-x}{\lambda}, & \rho = 1 \end{cases},$$

$$J_H(H^{-1}(x/\lambda)) = J_H(x/\lambda) = \begin{cases} \frac{e^{(1-\rho^{-1})x} - e^{\eta - C}}{(\lambda - S\mu)^2} - \frac{x e^{(1-\rho^{-1})x}}{\lambda^2 - \lambda S\mu} + \frac{\rho e^{\eta - C}}{\alpha(\lambda - S\mu)}, & \rho \neq 1 \\ \frac{1-\eta^{-2}x^2}{2\alpha^2} + \frac{1}{\lambda\alpha}, & \rho = 1 \end{cases},$$

$$J_1(H^{-1}(x/\lambda)) = J_1(x/\lambda) = \begin{cases} \frac{e^{(1-\rho^{-1})x} - e^{\eta - C}}{(\lambda - S\mu)^2} - \frac{x e^{(1-\rho^{-1})x}}{\lambda^2 - \lambda S\mu} + \left[\frac{\rho}{\alpha(\lambda - S\mu)} + \frac{1}{(S\mu)^2}\right] e^{\eta - C}, & \rho \neq 1 \\ \frac{1-\eta^{-2}x^2}{2\alpha^2} + \frac{1}{\lambda\alpha} + \frac{1}{\lambda^2}, & \rho = 1 \end{cases}.$$

By the general formulas of building blocks (6.10), (6.19), (6.39) and (6.30), it can be shown that:

If $\rho \neq 1$ i.e., $\lambda \neq S\mu$,

$$J^{(N)} = \frac{\rho e^{\eta - C}\left[1 - P(N-S, \eta)\right] + \rho^{N-S+1} P(N-S, C) - 1}{\lambda - S\mu},$$

$$J_H^{(N)} = \frac{1}{(\lambda - S\mu)^2} - \frac{\eta^{N-S+1} e^{-C}}{S\mu(\lambda - S\mu)(N-S)!} + \frac{\left[\eta(\rho-1) - 1\right] e^{\eta - C}\left[1 - P(N-S+1, \eta)\right]}{(\lambda - S\mu)^2} + \frac{\left[(N-S+1)(\rho-1) - 1\right] \rho^{N-S+1} P(N-S+1, C)}{(\lambda - S\mu)^2},$$

$$J_1^{(N)} = J_H^{(N)} + \frac{e^{\eta - C}\left[1 - P(N-S+1, \eta)\right]}{(S\mu)^2},$$

$$J^{(N)}(t) = \begin{cases} \frac{\rho e^{\eta - C}[1 - P(N-S+1, \eta)]}{\lambda - S\mu} + \frac{e^{(\lambda - S\mu)t}[P(N-S+1, \lambda t) - 1]}{\lambda - S\mu} & 0 \leq t < \alpha^{-1} \\ + \frac{\rho^{N-S+1}[P(N-S+1, C) - P(N-S+1, S\mu t)]}{\lambda - S\mu}, & \\ \frac{[1 - P(N-S+1, \eta)] e^{\eta - S\mu t}}{S\mu}, & t \geq \alpha^{-1} \end{cases}.$$

If $\rho = 1$ i.e., $\lambda = S\mu$,

$$
\begin{aligned}
J^{(N)} &= \left[\frac{1}{\alpha} + \frac{1}{\lambda} - \frac{\eta^{(N-S+1)}e^{-\eta}}{\lambda(N-S)!}\right] + \left[\frac{N-S}{\lambda} - \frac{1}{\alpha}\right] P(N-S+1,\eta) \\
&= \frac{1+\eta-\eta P(N-S,\eta) + (N-S)P(N-S+1,\eta)}{\lambda}, \\
J_H^{(N)} &= \left[\frac{1}{2\alpha^2} + \frac{1}{\lambda\alpha}\right][1-P(N-S+1,\eta)] + \frac{(N-S+3)(N-S+2)P(N-S+1,\eta)}{2\lambda^2} \\
&\quad - \frac{(N-S+3)\eta^{N-S+1}e^{-\eta} + \eta^{N-S+2}e^{-\eta}}{2\lambda^2(N-S)!} \\
&= \left[\frac{1}{2\alpha^2} + \frac{1}{\lambda\alpha}\right][1-P(N-S+1,\eta)] + \frac{\gamma(N-S+3,\eta)}{2\lambda^2(N-S)!}, \\
J_1^{(N)} &= J_H^{(N)} + \frac{1-P(N-S+1,\eta)}{\lambda^2}, \\
J^{(N)}(t) &= \begin{cases} \left[\frac{1}{\alpha} + \frac{1}{\lambda} - t\right] + \frac{[\lambda t - N + S - 1]P(N-S+1,\lambda t)}{\lambda} & \\ \quad + \left[\frac{N-S}{\lambda} - \frac{1}{\alpha}\right] P(N-S+1,\eta) + \frac{(\lambda t)^{N-S+1}e^{-\lambda t} - \eta^{N-S+1}e^{-\eta}}{\lambda(N-S)!}, & 0 \le t < \alpha^{-1} \\ \frac{[1-P(N-S+1,\eta)]e^{\eta-\lambda t}}{\lambda}, & t \ge \alpha^{-1} \end{cases}.
\end{aligned}
$$

There are some common performance measures of $M/M/S+D$ and $M/M/S/N+D$ models. Referring to the corresponding general formulas of $M/M/S+G$ model and using the results we obtained in this section, we have for $M/M/S+D$ and $M/M/S/N+D$ models:

1. Abandonment rate (use (6.6))

$$
r_i = \begin{cases} 0, & 0 \le i \le S \\ S\mu\frac{P(i-S-1,C)-P(i-S,C)}{P(i-S,C)}, & i > S \end{cases}.
$$

2. The density of V_i (use (6.5))

$$
f_{V_i}(t) = \begin{cases} 0, & 0 \le i < S \\ \frac{(S\mu)^{i-S+1}}{P(i-S,C)(i-S)!}t^{i-S}e^{-S\mu t}, & i \ge S, t < \alpha^{-1} \\ \frac{(S\mu)^{i-S+1}}{P(i-S,C)(i-S)!}\alpha^{S-i}e^{-S\mu t}, & i \ge S, t \ge \alpha^{-1} \end{cases}.
$$

Hence $P(V_i > t) = 0$, for $0 \le i < S$. For $i \ge S$, we have when $t \ge \alpha^{-1}$

$$
\begin{aligned}
P(V_i > t) &= \frac{(S\mu)^{i-S+1}}{P(i-S,C)(i-S)!}\int_t^\infty \alpha^{-(i-S)}e^{-S\mu u}du \\
&= \frac{(S\mu)^{i-S+1}\alpha^{-(i-S)}}{P(i-S,C)(i-S)!}\frac{1}{S\mu}e^{-S\mu t} \\
&= \frac{C^{i-S}}{P(i-S,C)(i-S)!}e^{-S\mu t}
\end{aligned}
$$

and when $t < \alpha^{-1}$

$$
\begin{aligned}
P(V_i > t) &= \frac{(S\mu)^{i-S+1}}{P(i-S,C)(i-S)!}\left(\int_t^{\alpha^{-1}} u^{i-S}e^{-S\mu u}du + \int_{\alpha^{-1}}^{\infty}\alpha^{-(i-S)}e^{-S\mu u}du\right) \\
&= \frac{(S\mu)^{i-S+1}}{P(i-S,C)(i-S)!}\frac{\gamma(i-S+1,C)-\gamma(i-S+1,S\mu t)}{(S\mu)^{i-S+1}} + \frac{C^{i-S}}{P(i-S,C)(i-S)!}e^{-C} \\
&= \frac{\gamma(i-S+1,C)-\gamma(i-S+1,S\mu t)}{P(i-S,C)(i-S)!} + \frac{C^{i-S}}{P(i-S,C)(i-S)!}e^{-C} \\
&= \frac{\gamma(i-S+1,C)-\gamma(i-S+1,S\mu t)+C^{i-S}e^{-C}}{P(i-S,C)(i-S)!} \\
&= \frac{(i-S)\gamma(i-S,C)-\gamma(i-S+1,S\mu t)}{P(i-S,C)(i-S)!} \\
&= 1 - \frac{P(i-S+1,S\mu t)}{P(i-S,C)}.
\end{aligned}
$$

In [36], Movaghar gave the above results for $M/M/S+D$ model in terms of distribution function of Erlang distribution with parameters $(i, S\mu)$ i.e.,

$$F_{Y_i}(x) = \frac{(S\mu)^i}{(i-1)!}\int_0^x e^{-S\mu t}t^{i-1}dt,$$

where $Y_i \sim Er(i, S\mu)$. Actually

$$F_{Y_i}(\alpha^{-1}) = \frac{(S\mu)^i}{(i-1)!}\int_0^{\alpha^{-1}} e^{-S\mu t}t^{i-1}dt = \frac{\gamma(i, S\mu\alpha^{-1})}{(i-1)!} = P(i,C)$$

so that the above results agree with those obtained in [36].

6.2.7 An alternative method

Independently, in [11, 12], Brandt and Brandt considered $M(n)/M(n)/S+G$ queueing system. The first $M(n)$ means the input is a general state-dependent Poisson process, i.e., the arrival rate is dependent on the number of calls n in the system. The second $M(n)$ means the cumulative service rate is generally dependent on the number of calls n in the system. This is a quite general model and some special cases are in the following. If $\lambda_n > 0$ for $0 \le n < N$ and $\lambda_n \equiv 0$ for $n \ge N$, then we have $M(n)/M(n)/S/N+G$ model. If $\mu_n = min(n,S)\mu$ for $n \ge 0$, then we have $M(n)/M/S+G$ model and if additionally $\lambda_n \equiv \lambda > 0$ then we have $M/M/S+G$ model. In this section we will focus on the results of $M/M/S/N+G$ model and show that this alternative method produces the same result as we obtained before.

As we have mentioned before, using the supplementary variable method, the authors construct a complex Markov process by including the residual and original patience times of waiting calls. Specifically the Markov process is

$$(Q(t), X_1(t), ..., X_{L(t)}(t); U_1(t), ..., U_{L(t)}(t)) \tag{6.46}$$

where $Q(t)$ is the queue length at time t; $L(t) = (Q(t)-S)_+$ is the number of calls waiting in the queue at time t; $X_i(t)$ and $U_i(t)$ are the residual and original patience times of waiting calls for the queueing position $i = 1...L(t)$ and $i = 1$ is the first call in the queue, which will be potentially the next call for service. The stationary distribution of this process is

$$\begin{aligned} &P_i(x_1, ...x_l; u_1, ...u_l) \\ &= \lim_{t\to\infty} P(Q(t) = i; X_1(t) \le x_1, ..., X_l(t) \le x_l; U_1(t) \le u_1, ..., U_l(t) \le u_l) \end{aligned}$$

where $l = (i - S)_+$ and the density is

$$p_i(x_1, ...x_l; u_1, ...u_l) = \frac{\partial^{2l}}{\partial x_1 ... \partial x_l \partial u_1 ... \partial u_l} P_i(x_1, ...x_l; u_1, ...u_l).$$

The authors then derived a system of integral equations for the density $p_i(x_1, ...x_l; u_1, ...u_l)$ of the Markov process (6.46). By solving these equations explicitly they obtained $p_i(x_1, ...x_l; u_1, ...u_l)$, the stationary queue length distribution Q, various conditional waiting time distributions and mean waiting times. The results are summarized in the following.

Let

$$F(\xi) := \int_0^{\xi/(S\mu)} \overline{G}(u)du, \xi > 0$$

and the constants

$$F_j := \frac{1}{j!} \int_0^\infty F(\xi)^j e^{-\xi} d\xi.$$

Then the stationary distribution of number of calls in the system is

$$p_i = \begin{cases} gS!\mu^S \frac{a^i}{i!} & \text{if } 0 \le i \le S \\ g\lambda^i F_{i-S} & \text{if } S \le i \le N \end{cases} \tag{6.47}$$

where

$$g^{-1} = S!\mu^S \sum_{i=0}^{S-1} \frac{a^i}{i!} + \lambda^S \sum_{i=0}^{N-S} \lambda^i F_i$$

and hence $p_0 = gS!\mu^S$.

The above results are actually the same as results obtained before (6.8) since we have

$$F(\xi) = \int_0^{\xi/(S\mu)} \overline{G}(u)du = H(\xi/S\mu), \xi > 0$$

and

$$F_i = \frac{1}{i!}\int_0^\infty F(\xi)^i e^{-\xi} d\xi = \frac{1}{i!}\int_0^\infty H(\xi/S\mu)^i e^{-\xi} d\xi = \frac{S\mu}{i!}\int_0^\infty H(t)^i e^{-S\mu t} dt = \frac{S\mu}{i!} g_i^*(S\mu). \tag{6.48}$$

After replacing F_{i-S} in (6.47) with (6.48), we will obtain the same stationary distribution (6.8). This fact confirms the argument of Movaghar [36] that once r_i is found, p_i will satisfy a set of difference equations similar as the global balance equations of a birth-death process with abandonment rate r_i as additional death rate even though the queue length process $Q(t)$ itself is not a birth-death process.

For the waiting time distribution and the mean waiting time of $M/M/S/N+G$ model, the following formulas are obtained in [11, 12], where we also use their notation.

1.

$$\begin{aligned} P(W_q > t|Sr) &= 1 - W_S(t) \\ &= \frac{S\mu p_S}{\lambda P(Sr, \text{non-blocking})} \sum_{j=0}^{N-S-1} \frac{\lambda^{j+1}}{j!} \int_{S\mu t}^\infty F(\xi)^j F'(\xi) e^{-\xi} d\xi. \end{aligned}$$

2.

$$\begin{aligned} P(W_q > t|Ab) &= 1 - W_I(t) \\ &= \frac{S\mu p_S}{\lambda P(Ab, \text{non-blocking})} \sum_{j=0}^{N-S-1} \frac{\lambda^{j+1}}{j!} \int_{S\mu t}^\infty F(\xi)^j \left[F'(S\mu t) - F'(\xi)\right] e^{-\xi} d\xi. \end{aligned}$$

3.

$$\begin{aligned} P(W_q > t) &= P(Sr|\text{non-blocking}) P(W_q > t|Sr) + P(Ab|\text{non-blocking}) P(W_q > t|Ab) \\ &= \frac{S\mu p_S}{\lambda P(\text{non-blocking})} F'(S\mu t) \sum_{j=0}^{N-S-1} \frac{\lambda^{j+1}}{j!} \int_{S\mu t}^\infty F(\xi)^j e^{-\xi} d\xi. \end{aligned}$$

4.

$$E(W_q|Sr) = EW_S = \frac{p_S}{\lambda P(Sr, \text{non-blocking})} \sum_{j=1}^{N-S} \frac{\lambda^j}{j!} \int_0^\infty F(\xi)^j (\xi - 1) e^{-\xi} d\xi.$$

5.

$$E(W_q|Ab) = EW_I = \frac{p_S}{\lambda P(Ab, \text{non-blocking})} \sum_{j=1}^{N-S} \frac{\lambda^j}{j!} \int_0^\infty F(\xi)^j (j + 1 - \xi) e^{-\xi} d\xi.$$

6.

$$E(W_q) = P(Sr|\text{non-blocking})EW_S + P(Ab|\text{non-blocking})EW_I$$
$$= \frac{p_S}{\lambda P(\text{non-blocking})} \sum_{j=1}^{N-S} \frac{\lambda^j}{(j-1)!} \int_0^\infty F(\xi)^j e^{-\xi} d\xi.$$

It can also be shown that the above formulas are the same as we have obtained before. For example since

$$F'(\xi) = \frac{1}{S\mu} H'(\xi/S\mu) = \frac{1}{S\mu} \overline{G}(\xi/S\mu),$$

we have

$$P(W_q > t|Sr) = \frac{S\mu p_S}{\lambda P(Sr, \text{non-blocking})} \sum_{j=0}^{N-S-1} \frac{\lambda^{j+1}}{j!} \int_{S\mu t}^\infty F(\xi)^j F'(\xi) e^{-\xi} d\xi$$
$$= \frac{p_S}{P(Sr, \text{non-blocking})} \sum_{j=0}^{N-S-1} \frac{\lambda^j}{j!} \int_{S\mu t}^\infty [H(\xi/S\mu)]^j \, \overline{G}(\xi/S\mu) e^{-\xi} d\xi,$$

where, by (6.13) and (6.24),

$$\frac{p_S}{P(Sr, \text{non-blocking})} = \frac{p_S}{P(Sr)} = \frac{\frac{\rho}{\epsilon + \lambda J^{(N)}}}{\frac{\epsilon - 1 + S\mu J^{(N)}}{\epsilon + \lambda J^{(N)}}} = \frac{\rho}{\epsilon - 1 + S\mu J^{(N)}}$$

and

$$\sum_{j=0}^{N-S-1} \frac{\lambda^j}{j!} \int_{S\mu t}^\infty [H(\xi/S\mu)]^j \, \overline{G}(\xi/S\mu) e^{-\xi} d\xi$$
$$= S\mu \sum_{j=0}^{N-S-1} \frac{\lambda^j}{j!} \int_t^\infty [H(u)]^j \, \overline{G}(u) e^{-S\mu u} du$$
$$= S\mu \int_t^\infty \overline{G}(u) e^{\lambda H(u) - S\mu u} [1 - P(N-S, \lambda H(u))] du$$
$$= S\mu \left[\int_t^\infty e^{\lambda H(u) - S\mu u} [1 - P(N-S, \lambda H(u))] du - \int_t^\infty G(u) e^{\lambda H(u) - S\mu u} [1 - P(N-S, \lambda H(u))] du \right]$$
$$= S\mu \left[J^{(N-1)}(t) - J_G^{(N-1)}(t) \right].$$

Hence we have

$$P(W_q > t|Sr)$$
$$= \frac{\rho}{\epsilon - 1 + S\mu J^{(N)}} S\mu \left[J^{(N-1)}(t) - J_G^{(N-1)}(t) \right]$$
$$= \frac{\lambda \left[J^{(N-1)}(t) - J_G^{(N-1)}(t) \right]}{\epsilon - 1 + S\mu J^{(N)}}$$

which is (6.36).

In [12] the authors also gave the expression of the abandonment rate function based on results from [11]

$$r_i = \frac{F_{i-S-1}}{F_{i-S}} - S\mu,\ S < i \le N. \tag{6.49}$$

Since $F_i = \frac{S\mu}{i!} g_i^*(S\mu)$, the above is the same as (6.6). For conditional offered waiting time V_i considered in [36], the authors [12] presented a new proof for the density of V_i for $M(n)/M(n)/S + G$ model by using results from [11].

A Markov approximation of $M/M/S/N + G$ model

In [12], a Markov approximation of $M/M/S/N+G$ model is proposed. The approximation model is denoted by $M/M/S/N + M(\beta_i)_{i=1}^{N-S}$ where $M(\beta_i)_{i=1}^{N-S}$ means that with each waiting place, numbered by $i = 1, 2, ..., N - S$ where $i = 1$ is the first call to be served potentially (i.e., if not abandon), there is a position-related exponential abandonment rate β_i. Hence a call waiting on the ith position has an exponential rate β_i to abandon the queue. The calls behind him move up according to FCFS discipline and change their abandonment rate according to their new positions. The total abandonment rate with j calls in the system is

$$r_{j,\beta} = \sum_{i=1}^{j-S} \beta_i, \quad S < j \le N.$$

If all $\beta_i = \alpha$ is a constant, then the model becomes $M/M/S/N + M$ model.

It is obvious that the process $Q(t)$ of this model is a birth-death process with the stationary distribution

$$p_i = \begin{cases} gS!\mu^S \frac{a^i}{i!} & \text{if } 0 \le i \le S \\ \frac{g\lambda^i}{\prod_{j=S+1}^{i} (S\mu + r_{j,\beta})} & \text{if } S \le i \le N \end{cases} \tag{6.50}$$

where

$$g^{-1} = S!\mu^S \sum_{i=0}^{S-1} \frac{a^i}{i!} + \lambda^S \sum_{i=0}^{N-S} \frac{\lambda^i}{\prod_{j=1}^{i} (S\mu + r_{j+S,\beta})}.$$

The key idea of the approximation is to find β_i such that two distributions (6.47) and (6.50) are the same. Comparing the above two distributions, the authors obtained for

$S \leq i \leq N$,

$$F_{i-S} = \frac{1}{\prod_{j=S+1}^{i} (S\mu + r_{j,\beta})}.$$

Solving the above gives

$$r_{i,\beta} = \frac{F_{i-S-1}}{F_{i-S}} - S\mu,\ S < i \leq N,$$

which is actually the same as (6.49). Hence we find that when the stationary distributions are fitted, the abandonment rate function r_i and $r_{i,\beta}$ are also fitted. This fact again confirms the argument of Movaghar [36] that once r_i is found, p_i will satisfy a set of difference equations similar as the global balance equations of a birth-death process with abandonment rate r_i as additional death rate even though the queue length process $Q(t)$ itself is not a birth-death process.

Now since it has been proved in [12] that r_i is a strictly increasing positive function, the method to find β_i in [12] is,

$$\beta_1 = r_{S+1}$$
$$\beta_i = r_{S+i} - r_{S+i-1},\ 1 < i \leq N - S.$$

The fact that p_i and abandonment rate function r_i are fitted for two models if we choose β_i as above also implies the fitting of those performance measures which only involved p_i and r_i, such as $P(Ab), P(blocking), E(Q_q)$ and $E(W_q)$ etc. However this is not true for the various waiting time distributions for two models.

6.2.8 Types of patience time distributions

In $M/M/S/N+G$ model, patience times are assumed to be i.i.d. and generally distributed with distribution function $G(t)$. We have mainly studied exponential and deterministic patience time distributions before for the performance analysis of models. In this section we will consider more types of distributions such as Erlang and uniform distributions. However our focus is not in the performance analysis of each model. Instead we will try to give an order relationship for some performance measures of those models with different patience time distributions assuming that all other model parameters are the same.

In [34], the authors studied the impact of customer's patience on delay and abandonment for $M/M/S+G$ model. They gave an order of delay and abandonment assuming an order relationship between two different patience time distributions. Also they got

the result that the deterministic patience time distribution minimizes $P(AB)$, maximizes $P(W_q > 0)$ and $E(W_q)$. In [37], similar results have been obtained for $M/M/1/N+G$ model with patience time on sojourn, i.e., calls could abandon the system even they are in service. We will generalize these results to $M/M/S/N+G$ model in this section.

For $M/M/S/N+G_A$ and $M/M/S/N+G_B$ models, where we have used the lower index A and B to distinguish two models with different patience time distributions as in the rest of this section, we will assume that all model parameters are the same except that the distributions of the patience time X are different with the same mean α^{-1}. We will assume the following order condition for distribution functions $G_A(t)$ and $G_B(t)$

$$H_A(x) = \int_0^x \overline{G}_A(u)du \geq H_B(x) = \int_0^x \overline{G}_B(u)du, \text{ for } x \geq 0. \tag{6.51}$$

Note that both [34] and [37] used this order condition. We will first prove the following lemma and then give some order relationship for some performance measures under some conditions.

Lemma 6.2.2 *Under the condition (6.51), for $M/M/S/N+G$ model, we have:*

1. $g^*_{A,j}(S\mu) \geq g^*_{B,j}(S\mu)$ *for* $0 \leq j \leq N-S$.
2. $J^{(N)}_A \geq J^{(N)}_B$.

Proof. These are obvious from the definition of $g^*_j(S\mu)$ and $J^{(N)}$, i.e.,

$$g^*_j(S\mu) = \int_0^\infty e^{-S\mu t} H(t)^j dt$$

and

$$J^{(N)} = \sum_{i=S}^{N} \frac{\lambda^{i-S} g^*_{i-S}(S\mu)}{(i-S)!}.$$

■

Theorem 6.2.3 *Under the condition (6.51), for $M/M/S/N+G$ model, we have:*

1. *For the stationary distribution of number of calls in the system,*

$$p_{A,i} \leq p_{B,i}, \ for \ 0 \leq i \leq S.$$

2. *For the probability that the call get service without delay,*

$$P_A(no\text{-}delay) \leq P_B(no\text{-}delay).$$

3. For the mean number of busy servers in the queue,

$$E_A(Q_b) \geq E_B(Q_b).$$

4. For the probability of served calls,

$$P_A(Sr) \geq P_B(Sr).$$

5. For the conditional delay,

$$P_A(delay|non\text{-}blocking) \geq P_B(delay|non\text{-}blocking).$$

Proof. 1. By (6.11),

$$p_0^{-1} = \sum_{i=0}^{S-1} \frac{a^i}{i!} + \frac{a^{S-1}\lambda J^{(N)}}{(S-1)!}.$$

Therefore from Lemma 6.2.2, we have

$$p_{A,0} \leq p_{B,0}.$$

Also by (6.8),

$$p_i = p_0 \frac{a^i}{i!},\ 0 \leq i \leq S.$$

Hence

$$p_{A,i} \leq p_{B,i},\ 0 \leq i \leq S. \tag{6.52}$$

2. P(no-delay) is the probability that the call get service without delay. This can be proved by (6.52) and the fact that

$$P(\text{no-delay}) = \sum_{i=0}^{S-1} p_i = \frac{\epsilon}{\epsilon + \lambda J^{(N)}}.$$

3. The mean number of busy servers $E(Q_b)$ can be written in terms of p_i, $0 \leq i \leq S$ as

$$\begin{aligned} E(Q_b) &= \sum_{i=0}^{S-1} i p_i + S \sum_{i=S}^{N} p_i \\ &= \sum_{i=0}^{S-1} i p_i + S(1 - \sum_{i=0}^{S-1} p_i) \\ &= S - \sum_{i=0}^{S-1} (S-i) p_i. \end{aligned}$$

Therefore by (6.52), we have

$$E_A(Q_b) \geq E_B(Q_b).$$

4. This is obvious since

$$P(Sr) = \frac{E(Q_b)}{a}.$$

5. By (6.31),

$$P(delay|\text{non-blocking}) = P(W_q > 0) = \frac{\lambda J^{(N-1)}}{\epsilon + \lambda J^{(N-1)}}.$$

Hence

$$\begin{aligned}
&P_A(delay|\text{non-blocking}) - P_B(delay|\text{non-blocking}) \\
&= \frac{\lambda J_A^{(N-1)}}{\epsilon + \lambda J_A^{(N-1)}} - \frac{\lambda J_B^{(N-1)}}{\epsilon + \lambda J_B^{(N-1)}} \\
&= \frac{\lambda J_A^{(N-1)} \left[\epsilon + \lambda J_B^{(N-1)}\right] - \lambda J_B^{(N-1)} \left[\epsilon + \lambda J_A^{(N-1)}\right]}{\left[\epsilon + \lambda J_A^{(N-1)}\right] \left[\epsilon + \lambda J_B^{(N-1)}\right]} \\
&= \frac{\lambda\epsilon \left[J_A^{(N-1)} - J_B^{(N-1)}\right]}{\left[\epsilon + \lambda J_A^{(N-1)}\right] \left[\epsilon + \lambda J_B^{(N-1)}\right]} \geq 0
\end{aligned}$$

by Lemma 6.2.2. ∎

Theorem 6.2.4 *Under the condition (6.51) and the condition*

$$\frac{J_A^{(N)}}{J_A^{(N-1)}} \geq \frac{J_B^{(N)}}{J_B^{(N-1)}}, \tag{6.53}$$

for $M/M/S/N + G$ model, we have:

1. $P_A(blocking) \geq P_B(blocking)$.

2. $P_A(Ab) \leq P_B(Ab)$.

3. $P_A(Sr|non\text{-}blocking) \geq P_B(Sr|non\text{-}blocking)$.

Proof. 1. Condition (6.53) is equivalent to

$$J_A^{(N)} J_B^{(N-1)} \geq J_B^{(N)} J_A^{(N-1)}. \tag{6.54}$$

By taking away $J_A^{(N-1)} J_B^{(N-1)}$ from both sides we have

$$J_A^{(N)} J_B^{(N-1)} - J_A^{(N-1)} J_B^{(N-1)} \geq J_B^{(N)} J_A^{(N-1)} - J_A^{(N-1)} J_B^{(N-1)}$$

or

$$\left[J_A^{(N)} - J_A^{(N-1)}\right] J_B^{(N-1)} \geq \left[J_B^{(N)} - J_B^{(N-1)}\right] J_A^{(N-1)}.$$

By Lemma 6.2.2, $0 \leq J_B^{(N-1)} \leq J_A^{(N-1)}$. Therefore from the above

$$J_A^{(N)} - J_A^{(N-1)} \geq J_B^{(N)} - J_B^{(N-1)}. \tag{6.55}$$

On the other hand, by (6.16) we have

$$P(blocking) = \frac{\lambda(J^{(N)} - J^{(N-1)})}{\epsilon + \lambda J^{(N)}}.$$

Hence

$$\begin{aligned}
&P_A(blocking) - P_B(blocking) \\
&= \frac{\lambda(J_A^{(N)} - J_A^{(N-1)})}{\epsilon + \lambda J_A^{(N)}} - \frac{\lambda(J_B^{(N)} - J_B^{(N-1)})}{\epsilon + \lambda J_B^{(N)}} \\
&= \lambda \frac{(J_A^{(N)} - J_A^{(N-1)})(\epsilon + \lambda J_B^{(N)}) - (J_B^{(N)} - J_B^{(N-1)})(\epsilon + \lambda J_A^{(N)})}{(\epsilon + \lambda J_A^{(N)})(\epsilon + \lambda J_B^{(N)})} \\
&= \lambda \frac{\epsilon\left[J_A^{(N)} - J_A^{(N-1)} - (J_B^{(N)} - J_B^{(N-1)})\right] + \lambda\left[J_A^{(N)} J_B^{(N-1)} - J_B^{(N)} J_A^{(N-1)}\right]}{(\epsilon + \lambda J_A^{(N)})(\epsilon + \lambda J_B^{(N)})} \geq 0
\end{aligned}$$

by (6.54) and (6.55).

2. We have

$$P(Ab) = 1 - P(blocking) - P(Sr).$$

Now since

$$P_A(blocking) \geq P_B(blocking),$$

and by Theorem 6.2.3 $P_A(Sr) \geq P_B(Sr)$, we have,

$$P_A(Ab) \leq P_B(Ab)$$

3. Since

$$P(Sr|\text{non-blocking}) = \frac{P(Sr)}{1 - P(blocking)},$$

the result is obvious by the two previous results. ∎

Theorem 6.2.5 *Under the condition (6.51) and the condition*

$$\frac{J_A^{(N)}}{J_A^{(N-1)}} \leq \frac{J_B^{(N)}}{J_B^{(N-1)}},$$

for $M/M/S/N + G$ model, we have

$$P_A(delay) \geq P_B(delay).$$

Proof. By (6.14) we have

$$P(delay)=\frac{\lambda J^{(N-1)}}{\epsilon+\lambda J^{(N)}}.$$

Hence

$$\begin{aligned}
&P_A(delay)-P_B(delay)\\
&=\frac{\lambda J_A^{(N-1)}}{\epsilon+\lambda J_A^{(N)}}-\frac{\lambda J_B^{(N-1)}}{\epsilon+\lambda J_B^{(N)}}\\
&=\frac{\lambda J_A^{(N-1)}\left[\epsilon+\lambda J_B^{(N)}\right]-\lambda J_B^{(N-1)}\left[\epsilon+\lambda J_A^{(N)}\right]}{\left[\epsilon+\lambda J_A^{(N)}\right]\left[\epsilon+\lambda J_B^{(N)}\right]}\\
&=\frac{\lambda\epsilon\left[J_A^{(N-1)}-J_B^{(N-1)}\right]+\lambda^2\left[J_B^{(N)}J_A^{(N-1)}-J_A^{(N)}J_B^{(N-1)}\right]}{\left[\epsilon+\lambda J_A^{(N)}\right]\left[\epsilon+\lambda J_B^{(N)}\right]}\geq 0.
\end{aligned}$$

■

After providing the order relationship of some performance measures of $M/M/S/N+G$ model under some order conditions of patience time distribution in the above, next we will consider the special properties of the deterministic patience time distribution. We have the following Lemma from [34] and [37]. Both papers have this result although their proofs are different. The proof given here is adapted from [37].

Lemma 6.2.3 *For $x\geq 0$,*

$$H_D(x)=\int_0^x \overline{G}_D(u)du\geq H(x)=\int_0^x \overline{G}(u)du,$$

where

$$\overline{G}_D(u)=\begin{cases}1, & u<\alpha^{-1}\\ 0, & u\geq\alpha^{-1}\end{cases}$$

is the survival function of the deterministic distribution $X=\alpha^{-1}$ and $\overline{G}(u)$ is the survival function of any patience time distribution with mean α^{-1}.

Proof. For $0\leq u<\alpha^{-1}$, $\overline{G}_D(u)=1\geq\overline{G}(u)$. Hence for $0\leq x<\alpha^{-1}$, $H_D(x)\geq H(x)$. For $x\geq\alpha^{-1}$

$$\begin{aligned}
H_D(x)&=\int_0^{\alpha^{-1}}\overline{G}_D(u)du+\int_{\alpha^{-1}}^x\overline{G}_D(u)du\\
&=\int_0^{\alpha^{-1}}\overline{G}_D(u)du=\alpha^{-1}\geq H(x)
\end{aligned}$$

since $H(x)$ is a non-decreasing function and $H(\infty) = \alpha^{-1}$. ∎

Therefore by the above Lemma, Lemma 6.2.2 and Theorem 6.2.3 we have the following extremal property of the deterministic patience time among all patience time distributions with mean α^{-1} for $M/M/S/N + G$ model.

Theorem 6.2.6 *For $M/M/S/N + G$ model, we have:*

1. $g^*_{D,j}(S\mu) \geq g^*_j(S\mu)$ *for* $0 \leq j \leq N - S$.
2. $J^{(N)}_D \geq J^{(N)}$.
3. $p_{D,i} \leq p_i$, *for* $0 \leq i \leq S$.
4. $P_D(\textit{no-delay}) \leq P(\textit{no-delay})$.
5. $E_D(Q_b) \geq E(Q_b)$.
6. $P_D(Sr) \geq P(Sr)$.
7. $P_D(\textit{delay}|\textit{non-blocking}) \geq P(\textit{delay}|\textit{non-blocking})$.

We will consider Erlang patience time distribution with mean α^{-1} in the following, i.e.,

$$\overline{G}_{E(k)}(u) = e^{-k\alpha u} \sum_{i=0}^{k-1} \frac{(k\alpha u)^i}{i!} = \frac{(k\alpha)^k}{(k-1)!} \int_u^\infty e^{-k\alpha t} t^{k-1} dt, \ u \geq 0$$

where $\overline{G}_{E(k)}(u)$ is the survival function of Erlang distribution with parameters k and $k\alpha$. In [37], the following result has been proved.

Lemma 6.2.4 *For all $k \geq 1$ and $x \geq 0$,*

$$H_{E(k+1)}(x) = \int_0^x \overline{G}_{E(k+1)}(u)du \geq H_{E(k)}(x) = \int_0^x \overline{G}_{E(k)}(u)du.$$

Therefore, as in the case of deterministic patience time, we have the following theorem.

Theorem 6.2.7 *For $M/M/S/N + E(k)$ model, where $E(k)$ means the patience time has Erlang distribution with parameters k and $k\alpha$, we have for all $k \geq 1$:*

1. $g^*_{E(k+1),j}(S\mu) \geq g^*_{E(k),j}(S\mu)$ *for* $0 \leq j \leq N - S$.
2. $J^{(N)}_{E(k+1)} \geq J^{(N)}_{E(k)}$.

3. $p_{E(k+1),i} \leq p_{E(k),i}$, *for* $0 \leq i \leq S$.

4. $P_{E(k+1)}(\textit{no-delay}) \leq P_{E(k)}(\textit{no-delay})$.

5. $E_{E(k+1)}(Q_b) \geq E_{E(k)}(Q_b)$.

6. $P_{E(k+1)}(Sr) \geq P_{E(k)}(Sr)$.

7. $P_{E(k+1)}(\textit{delay}|\textit{non-blocking}) \geq P_{E(k)}(\textit{delay}|\textit{non-blocking})$.

Remark 6.2.1 *When $k = 1$, we have $M/M/S/N+M$ model. The above theorem shows the extremal property of the exponential patience time distribution among all Erlang patience time distributions with mean α^{-1}.*

Remark 6.2.2 *When $k \to \infty$, we have $M/M/S/N + D$ model. The above theorem shows the extremal property of the deterministic patience time distribution among all Erlang patience time distributions with mean α^{-1}, which is consistent with Theorem 6.2.6.*

At last we will consider the uniform patience time distribution with mean α^{-1} and the support $(0, 2\alpha^{-1})$. In this case

$$\overline{G}_U(u) = \begin{cases} 1 - \frac{u}{2\alpha^{-1}}, & 0 \leq u < 2\alpha^{-1} \\ 0, & u \geq 2\alpha^{-1} \end{cases}.$$

In [37], the following result about the uniform patience time distribution has been proved.

Lemma 6.2.5 *For all $x \geq 0$,*

$$H_U(x) = \int_0^x \overline{G}_U(u)du \geq H_{E(1)}(x) = \int_0^x \overline{G}_{E(1)}(u)du.$$

Therefore, we have the following theorem.

Theorem 6.2.8 *For $M/M/S/N + M$ and $M/M/S/N + U$ models, where U means the patience time has uniform distribution with mean α^{-1} and the support $(0, 2\alpha^{-1})$, we have:*

1. $g^*_{U,j}(S\mu) \geq g^*_{M,j}(S\mu)$ *for* $0 \leq j \leq N - S$.

2. $J_U^{(N)} \geq J_M^{(N)}$.

3. $p_{U,i} \leq p_{M,i}$, *for* $0 \leq i \leq S$.

4. $P_U(\textit{no-delay}) \leq P_M(\textit{no-delay})$.

5. $E_U(Q_b) \geq E_M(Q_b)$.

6. $P_U(Sr) \geq P_M(Sr)$.

7. $P_U(\textit{delay}|\textit{non-blocking}) \geq P_M(\textit{delay}|\textit{non-blocking})$.

6.3 SOQN model with general abandonment of call centres (SOQN+G)

In this section we will study the SOQN model with general abandonment of call centres denoted by SOQN+G, which is a generalization of $M/M/S/N+G$ model to the two-node SOQN case. We will first give a model description and then derive the main performance measures of this model.

6.3.1 Model description

The model is a semiopen network with two nodes in series. Node 1 models the IVRU with N servers each with exponential service rate θ and Node 2 models the CSRs with S ($\leq N$) servers each with exponential service rate μ. The maximum number of calls in the network is N, i.e., if an arriving call finds N calls in the system, it will be blocked and rejected entering the system. Hence there is no queue at Node 1 and there are at most $N-S$ calls waiting at Node 2. Arriving calls can enter the network only through Node 1 according to a Poisson process with arrival rate λ. After the service with Node 1 is completed, the call leaves the network with probability $\overline{p} = 1-p$ and it joins Node 2 with probability p. If there are free CSRs at Node 2, the call is served by one of S CSRs. Otherwise it waits in the queue to get service.

We model abandonment at Node 2 and assume that upon joining the queue, calls start the patience times X which have i.i.d. general distribution with mean α^{-1} (the distribution function is denoted by $G(t)$). If the waiting time for the call is longer than its patience time X, the call will abandon, leave the system and release the trunk line. Otherwise it gets service with a CSR and releases both the CSR and the trunk line and leaves the system after the service. The arrival, service and abandonment processes are all assumed to be independent. The original SOQN model in Chapter 3 can be thought of as this model with

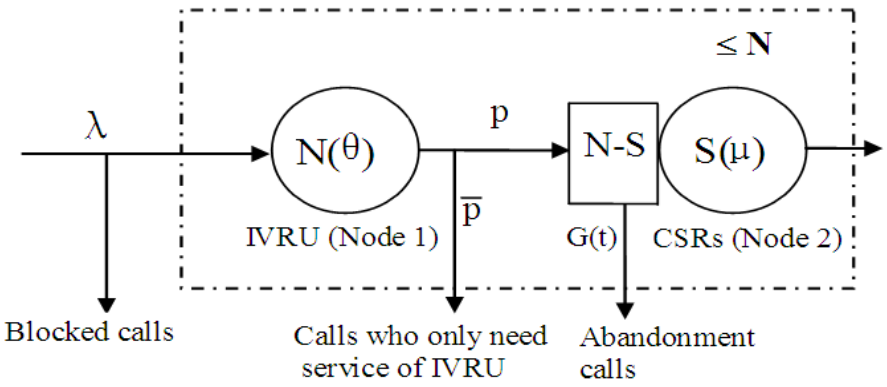

Figure 6.3: SOQN model with general abandonment

$\alpha = 0$ and the SOQN+M model studied in Chapter 5 is a special case of this model with exponential patience time distribution. Figure 6.3 gives a picture of the model.

6.3.2 Product form solution of the queue length process

Let $Q(t) = (Q_1(t), Q_2(t))$ be the queue length process, where $Q_i(t)$ is the queue length (number of calls) of Node $i, i = 1, 2$ at time t. We assume that the stationary distribution of $Q(t)$ exists and let $\pi_{ij} = P(Q_1 = i, Q_2 = j)$ be the stationary probability of having i calls at Node 1 and j calls at Node 2 with the state space $\Omega = \{(i, j)|i+j \leq N, (i, j) \in Z_+^2\}$.

For Node 2, as in $M/M/S/N + G$ model, we have the abandonment rate

$$r_j = \begin{cases} 0 & \text{if } 0 \leq j \leq S \\ \frac{F_{j-S-1}}{F_{j-S}} - S\mu, & \text{if } S < j \leq N \end{cases} \tag{6.56}$$

where $F_j = \frac{S\mu}{j!} g_j^*(S\mu)$ as in (6.48). Following the same way as $M/M/S/N + G$ model in Movaghar [36], we can incorporate the abandonment rate r_j into the service rate for Node 2. We have the argument that π_{ij} will still satisfy the global balance equations (4.8) in Chapter 4 with new service rates even though the queue length process $Q(t)$ itself is not a Markov process. Now it is easy to know that this model is a special case of the semiopen network model studied in Section 4.2, Chapter 4 with state dependent service rates $\mu_1(i) = i\theta$,

$$\mu_2(j) = \begin{cases} j\mu & for\ 0 \leq j \leq S \\ S\mu + r_j & for\ S < j \leq N \end{cases}$$

and constant balking. We have that Node 1 is similar to a $M/M/\infty$ queue with arrival rate λ and service rate θ. Node 2 is similar to a $M/M/S/N+G$ queue with arrival rate λp, service rate μ and general abandonment. Therefore we have the following result.

Theorem 6.3.1 *For SOQN+G model, the stationary distribution of the queue length process $Q = \{Q_1, Q_2\}$ has product form solution,*

$$\pi_{ij} = \pi_{00} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}, \quad 0 \le i+j \le N \ , \tag{6.57}$$

where $a_1 = \lambda/\theta$; $a_2 = p\lambda/\mu = pa$;

$$\beta(j) := \begin{cases} j! & for\ 0 \le j \le S \\ \frac{S! a^{j-S}}{\rho_{j-S}} & for\ S < j \le N \end{cases} ;$$

$$\rho_{j-S} := \frac{\lambda^{j-S}}{\prod_{n=S+1}^{j}(S\mu + r_n)}$$

and $\pi_{00} = \left[\sum_{0 \le i+j \le N} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)} \right]^{-1}$ *is the normalizing constant.*

Remark 6.3.1 *Using the expression of the abandonment rate r_n (6.56), we have*

$$\rho_{j-S} = F_{j-S} \lambda^{j-S} \tag{6.58}$$

or

$$\rho_{j-S} = \frac{S\mu g^*_{j-S}(S\mu)\lambda^{j-S}}{(j-S)!}.$$

In Section 5.3.2, Chapter 5, we already obtained some performance measures related to π_{ij} for SOQN+M model. Those expressions are still valid here and the only difference is that now we have a general expression of ρ_{j-S} valid for any patience time distributions. We repeat those results here.

1. The explicit expression for π_{00}^{-1},

$$\pi_{00}^{-1} = \sum_{k=0}^{N} \frac{(a_1+a_2)^k}{k!} + \sum_{k=S+1}^{N} \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \left(\frac{\rho_{j-S}}{S! a^{j-S}} - \frac{1}{j!} \right),$$

which will reduce to p_0^{-1} of $M/M/S/N+G$ model if we let $\theta = \infty$ and $p = 1$.

2. The marginal distribution for Node 1 and its mean,

$$\pi_{i*} := P(Q_1 = i) = \pi_{00} \sum_{j=0}^{N-i} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}, 0 \le i \le N,$$

$$E(Q_1) = \sum_{i=0}^{N} i\pi_{i*} = \pi_{00} \sum_{i=0}^{N} i \sum_{j=0}^{N-i} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}.$$

3. The marginal distribution for Node 2 and its mean,

$$\pi_{*j} := P(Q_2 = j) = \pi_{00} \sum_{i=0}^{N-j} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}, 0 \le j \le N,$$

$$E(Q_2) = \sum_{j=0}^{N} j\pi_{*j} = \pi_{00} \sum_{j=0}^{N} j \sum_{i=0}^{N-j} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}.$$

4. The mean number of calls waiting in the queue at Node 2,

$$E(Q_{2q}) = \sum_{j=S+1}^{N} (j-S)\pi_{*j} = \pi_{00} \sum_{j=S+1}^{N} (j-S) \sum_{i=0}^{N-j} \frac{a_1^i}{i!} \frac{a_2^j}{\beta(j)}.$$

6.3.3 Blocking probability

In Section 5.3.3, Chapter 5, we already obtained the stationary probabilities π_k for $0 \le k \le N$ that there are exactly k calls in the system and the blocking probability for SOQN+M model. Those expressions are still valid here and we summarize in the following, where random variable $Q = Q_1 + Q_2$ is the number of calls in the system.

1. The probability π_k for $0 \le k \le N$ and the blocking probability,

$$\pi_k = \pi_{00} \left(\frac{(a_1+a_2)^k}{k!} + \sum_{j=S+1}^{k} \frac{a_1^{k-j} a_2^j}{(k-j)!} \left(\frac{\rho_{j-S}}{S! a^{j-S}} - \frac{1}{j!} \right) I_{(S,\infty)}(k) \right),$$

$$P(blocking) = \pi_N.$$

2. The mean number of total calls in the system,

$$E(Q) = \sum_{k=0}^{N} k\pi_k = \sum_{k=0}^{N} k \sum_{j=0}^{k} \pi_{(k-j)j} = E(Q_1) + E(Q_2).$$

For Node 1, we apply the same Little's formula as in Chapter 3

$$E(Q_1) = a_1 \left[1 - P(blocking)\right]$$

which is also the carried load for Node 1. The utilization

$$v_1 = \frac{E(Q_1)}{N} = \frac{a_1 \left[1 - P(blocking)\right]}{N} < 1$$

is the proportion of time that an IVRU server is busy.

6.3.4 Probability of abandonment and other performance measures

As in Section 5.3.4, we only need to consider the probability of abandonment among those calls who are not blocked and join Node 2 denoted as $P(Ab|entry)$ where the event *entry* means non-blocking and joining Node 2. We have

$$\begin{aligned} P(Ab|entry) &= \sum_{j=S}^{N-1} P_j(Ab) P(\text{the call finds } j \text{ calls at Node 2}|entry) \\ &= \sum_{j=S}^{N-1} q_j \frac{r_{j+1}}{S\mu + r_{j+1}}, \end{aligned}$$

where we have used (6.4) for $P_j(Ab)$ and as in Section 5.3.4, $q_j = \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)}$. Hence by (6.56) and (6.58), we have

$$\begin{aligned} P(Ab|entry) &= \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \frac{F_{j-S} - S\mu F_{j-S+1}}{F_{j-S}} \\ &= \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \frac{\rho\rho_{j-S} - \rho_{j-S+1}}{\rho\rho_{j-S}}. \end{aligned}$$

We can also utilize the result in Chapter 3, as in Section 5.3.4,

$$q_j = \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} = \sum_{k=j+1}^{N} \chi(k,j) \tag{6.59}$$

to obtain an expression in terms of $\chi(k,j)$

$$\begin{aligned} P(Ab|entry) &= \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \frac{F_{j-S} - S\mu F_{j-S+1}}{F_{j-S}} \\ &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \frac{F_{j-S} - S\mu F_{j-S+1}}{F_{j-S}}. \end{aligned}$$

Therefore we have

$$\begin{aligned} P(Sr|entry) = 1 - P(Ab|entry) &= \sum_{j=0}^{S-1} q_j + \sum_{j=S}^{N-1} q_j \frac{S\mu F_{j-S+1}}{F_{j-S}} \\ &= \sum_{j=0}^{S-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} + \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \frac{S\mu F_{j-S+1}}{F_{j-S}} \\ &= \sum_{j=0}^{S-1} \sum_{k=j+1}^{N} \chi(k,j) + \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \frac{S\mu F_{j-S+1}}{F_{j-S}} \\ &= \sum_{k=1}^{N} \sum_{j=0}^{k\wedge S-1} \chi(k,j) + \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \frac{S\mu F_{j-S+1}}{F_{j-S}}. \end{aligned}$$

According to Chapter 3, the probability of entry is $P(entry) = p\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$ so that we have

$$\begin{aligned}
P(Ab) &= P(Ab|entry)P(entry) \\
&= \sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)}\frac{F_{j-S}-S\mu F_{j-S+1}}{F_{j-S}}\left(p\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}\right) \\
&= p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}\frac{F_{j-S}-S\mu F_{j-S+1}}{F_{j-S}}
\end{aligned}$$

by the relationship

$$\pi_{ij}^{(N-1)} = \frac{\pi_{ij}}{\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}}$$

as proved in Chapter 3. Also

$$\begin{aligned}
P(Sr) &= P(Sr|entry)P(entry) \\
&= \left(\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)} + \sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)}\frac{S\mu F_{j-S+1}}{F_{j-S}}\right)\left(p\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}\right) \\
&= p\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu F_{j-S+1}}{F_{j-S}}.
\end{aligned}$$

Now Little's formula for busy CSRs at Node 2 is

$$\begin{aligned}
E(Q_{2b}) &= \lambda P(Sr)\frac{1}{\mu} \\
&= ap\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + ap\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu F_{j-S+1}}{F_{j-S}}
\end{aligned}$$

which is easy to verify since

$$\begin{aligned}
&ap\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + ap\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu F_{j-S+1}}{F_{j-S}} \\
&= a_2\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij} + a_2\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\frac{\pi_{ij}S\mu F_{j-S+1}}{F_{j-S}} \\
&= \sum_{j=0}^{S-1}(j+1)\sum_{i=0}^{N-1-j}\pi_{i(j+1)} + S\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{i(j+1)} \\
&= \sum_{j=1}^{S}j\sum_{i=0}^{N-j}\pi_{ij} + S\sum_{j=S+1}^{N}\sum_{i=0}^{N-j}\pi_{ij} \\
&= E(Q_{2b}),
\end{aligned}$$

where we have used the fact that

$$a_2\pi_{ij} = (j+1)\pi_{i(j+1)} \; for \; 0 \le j < S$$

and

$$\lambda p\pi_{ij} = a_2\mu\pi_{ij} = [S\mu + r_{j+1}]\,\pi_{i(j+1)} = \frac{F_{j-S}}{F_{j-S+1}}\pi_{i(j+1)} \; for \; j \ge S \tag{6.60}$$

which can be verified by the product form solution π_{ij} (6.57). Hence the carried load for Node 2 is $a' = E(Q_{2b}) = aP(Sr)$. The utilization

$$v = \frac{a'}{S} = \rho P(Sr) < 1$$

is the proportion of time that a CSR is busy.

Similarly as in Section 3.3.4, Chapter 3, we have the following performance measures.

1. $P(\text{only self-served by Node 1}) = (1-p)\sum_{j=0}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
2. $P(\text{no-delay}, entry) = p\sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
3. $P(delay, entry) = p\sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}$.
4. $P(\text{no-delay}|entry) = \sum_{j=0}^{S-1} q_j = \sum_{j=0}^{S-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)} = \sum_{j=0}^{S-1}\sum_{k=j+1}^{N}\chi(k,j) = \sum_{k=1}^{N}\sum_{j=0}^{k\wedge S-1}\chi(k,j)$.
5. $P(delay|entry) = \sum_{j=S}^{N-1} q_j = \sum_{j=S}^{N-1}\sum_{i=0}^{N-1-j}\pi_{ij}^{(N-1)} = \sum_{j=S}^{N-1}\sum_{k=j+1}^{N}\chi(k,j) = \sum_{k=S+1}^{N}\sum_{j=S}^{k-1}\chi(k,j)$.

6.3.5 Waiting time distribution

For waiting time distribution, as SOQN+M model in Chapter 5, we only need to consider those calls given they are not blocked and join Node 2 (or given entry), since there is no queue at Node 1. Let W_q denote the conditional stationary waiting time of calls given entry, which is the time spent by an entry call in the queue of Node 2 until abandonment or starting to get service. We have

$$P(W_q > t) = P(V \wedge X > t) = P(V > t)\overline{G}(t)$$

where V is the conditional offered waiting time of an infinite patient call given entry, which is independent of the patience time X.

To obtain the distribution of V, we follow the same conditional argument as before. First we have, by (6.27),

$$\begin{aligned} P(V_j > t) &= \frac{\int_t^\infty H(u)^{j-S} e^{-S\mu u} du}{g^*_{j-S}(S\mu)} \\ &= \frac{L_{j-S}(t)}{g^*_{j-S}(S\mu)}, \end{aligned}$$

where we have defined $L_{j-S}(t) := \int_t^\infty H(u)^{j-S} e^{-S\mu u} du$ and $L_{j-S}(0) = g^*_{j-S}(S\mu)$. Then by (6.59), we have

$$\begin{aligned} P(V > t) &= \sum_{j=S}^{N-1} P(V_j > t) P(\text{the call finds } j \text{ calls at Node 2} | entry) \\ &= \sum_{j=S}^{N-1} \frac{L_{j-S}(t)}{g^*_{j-S}(S\mu)} q_j \\ &= \sum_{j=S}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}^{(N-1)} \frac{L_{j-S}(t)}{g^*_{j-S}(S\mu)} \\ &= \sum_{j=S}^{N-1} \sum_{k=j+1}^{N} \chi(k,j) \frac{L_{j-S}(t)}{g^*_{j-S}(S\mu)} \\ &= \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \frac{L_{j-S}(t)}{g^*_{j-S}(S\mu)} \end{aligned} \tag{6.61}$$

and

$$dP(V > t) = -\sum_{j=S}^{N-1} q_j \frac{H(t)^{j-S} e^{-S\mu t}}{g^*_{j-S}(S\mu)} dt. \tag{6.62}$$

Hence

$$P(W_q > t) = \sum_{j=S}^{N-1} q_j \frac{L_{j-S}(t)\overline{G}(t)}{g^*_{j-S}(S\mu)}. \tag{6.63}$$

The waiting time distribution for abandonment calls given entry is

$$\begin{aligned}
P(W_q{>}t, Ab) &= P(V \wedge X > t, V > X) \\
&= P(X > t, V > X) \\
&= \int_t^\infty P(V > x) dG(x) \\
&= P(V > x)G(x)|_t^\infty - \int_t^\infty G(x) dP(V > x) \\
&= \int_t^\infty G(x) \sum_{j=S}^{N-1} q_j \frac{H(x)^{j-S} e^{-S\mu x}}{g^*_{j-S}(S\mu)} dx - P(V > t)G(t) \\
&= \sum_{j=S}^{N-1} q_j \frac{\int_t^\infty G(x)H(x)^{j-S} e^{-S\mu x} dx}{g^*_{j-S}(S\mu)} - \sum_{j=S}^{N-1} q_j \frac{L_{j-S}(t)}{g^*_{j-S}(S\mu)} G(t) \\
&= \sum_{j=S}^{N-1} q_j \frac{\int_t^\infty G(x)H(x)^{j-S} e^{-S\mu x} dx - G(t)L_{j-S}(t)}{g^*_{j-S}(S\mu)} \\
&= \sum_{j=S}^{N-1} q_j \frac{\overline{G}(t)L_{j-S}(t) - \int_t^\infty \overline{G}(x)H(x)^{j-S} e^{-S\mu x} dx}{g^*_{j-S}(S\mu)},
\end{aligned}$$

where we have used (6.62) and (6.61). Then

$$\begin{aligned}
P(W_q{>}t, Sr) &= P(W_q{>}t) - P(W_q{>}t, Ab) \\
&= \sum_{j=S}^{N-1} q_j \frac{\overline{G}(t)L_{j-S}(t)}{g^*_{j-S}(S\mu)} - \sum_{j=S}^{N-1} q_j \frac{\overline{G}(t)L_{j-S}(t) - \int_t^\infty \overline{G}(x)H(x)^{j-S} e^{-S\mu x} dx}{g^*_{j-S}(S\mu)} \\
&= \sum_{j=S}^{N-1} q_j \frac{\int_t^\infty \overline{G}(x)H(x)^{j-S} e^{-S\mu x} dx}{g^*_{j-S}(S\mu)}.
\end{aligned}$$

To facilitate the computation, using integration by parts, it can be shown that

$$\int_t^\infty \overline{G}(x)H(x)^{j-S} e^{-S\mu x} dx = \frac{S\mu L_{j-S+1}(t) - e^{-S\mu t} H(t)^{j-S+1}}{j - S + 1}$$

so that

$$P(W_q{>}t, Sr) = \sum_{j=S}^{N-1} q_j \frac{S\mu L_{j-S+1}(t) - e^{-S\mu t} H(t)^{j-S+1}}{(j - S + 1) g^*_{j-S}(S\mu)} \tag{6.64}$$

and

$$P(W_q{>}t, Ab) = \sum_{j=S}^{N-1} q_j \frac{(j - S + 1)\overline{G}(t)L_{j-S}(t) - S\mu L_{j-S+1}(t) + e^{-S\mu t} H(t)^{j-S+1}}{(j - S + 1) g^*_{j-S}(S\mu)}.$$

Now we can derive the conditional waiting, which are more useful in practice, using the above results.

1.

$$P(W_q > t|Ab) = \frac{P(W_q > t, Ab)}{P(Ab|entry)} = \frac{\sum_{j=S}^{N-1} q_j \frac{(j-S+1)\overline{G}(t)L_{j-S}(t) - S\mu L_{j-S+1}(t) + e^{-S\mu t}H(t)^{j-S+1}}{(j-S+1)g^*_{j-S}(S\mu)}}{\sum_{j=S}^{N-1} q_j \left[1 - \frac{S\mu g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}\right]}.$$

2.

$$P(W_q > t|Sr) = \frac{P(W_q > t, Sr)}{P(Sr|entry)} = \frac{\sum_{j=S}^{N-1} q_j \frac{S\mu L_{j-S+1}(t) - e^{-S\mu t}H(t)^{j-S+1}}{(j-S+1)g^*_{j-S}(S\mu)}}{\sum_{j=0}^{S-1} q_j + \sum_{j=S}^{N-1} q_j \frac{S\mu g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}}.$$

3.

$$P(W_q > t|Sr, delay) = \frac{P(W_q > t, Sr)}{P(Sr, delay|entry)} = \frac{\sum_{j=S}^{N-1} q_j \frac{S\mu L_{j-S+1}(t) - e^{-S\mu t}H(t)^{j-S+1}}{(j-S+1)g^*_{j-S}(S\mu)}}{\sum_{j=S}^{N-1} q_j \frac{S\mu g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}}.$$

6.3.6 Mean waiting time

By (6.63), we have the mean waiting time for all calls given entry

$$\begin{aligned} E(W_q) &= \sum_{j=S}^{N-1} q_j \int_0^\infty \frac{L_{j-S}(t)\overline{G}(t)}{g^*_{j-S}(S\mu)} dt \\ &= \sum_{j=S}^{N-1} \frac{q_j}{g^*_{j-S}(S\mu)} \int_0^\infty L_{j-S}(t) dH(t) \\ &= \sum_{j=S}^{N-1} \frac{q_j g^*_{j-S+1}(S\mu)}{g^*_{j-S}(S\mu)}. \end{aligned}$$

Now Little's formula for all calls waiting in the queue at Node 2 is

$$\begin{aligned} E(Q_{2q}) &= \lambda P(entry) E(W_q) \\ &= \lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} E(W_q), \end{aligned} \tag{6.65}$$

which is easy to verify since

$$\begin{aligned}\lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} E(W_q) &= \lambda p \sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij} \sum_{k=S}^{N-1} q_k \frac{g^*_{k-S+1}(S\mu)}{g^*_{k-S}(S\mu)} \\ &= \sum_{k=S}^{N-1} \sum_{i=0}^{N-1-k} \lambda p \pi_{ik} \frac{g^*_{k-S+1}(S\mu)}{g^*_{k-S}(S\mu)} \\ &= \sum_{k=S}^{N-1} \sum_{i=0}^{N-1-k} \pi_{i(k+1)}(k-S+1) \\ &= \sum_{k=S+1}^{N} (k-S) \sum_{i=0}^{N-k} \pi_{ik} = E(Q_{2q}),\end{aligned}$$

where we have used a similar result as (6.60),

$$\lambda p \pi_{ik} = a_2 \mu \pi_{ik} = [S\mu + r_{k+1}]\, \pi_{i(k+1)} = \frac{(k-S+1) g^*_{k-S}(S\mu)}{g^*_{k-S+1}(S\mu)} \pi_{i(k+1)} \ for\ k \geq S$$

and a result in Chapter 5 (5.86),

$$q_k = \frac{\sum_{i=0}^{N-1-k} \pi_{ik}}{\sum_{j=0}^{N-1} \sum_{i=0}^{N-1-j} \pi_{ij}}.$$

Next by (6.64) the mean waiting time for served calls given entry is

$$\begin{aligned}E(W_q, Sr) &= \sum_{j=S}^{N-1} q_j \frac{S\mu \int_0^\infty L_{j-S+1}(t)dt - \int_0^\infty e^{-S\mu t} H(t)^{j-S+1} dt}{(j-S+1) g^*_{j-S}(S\mu)} \\ &= \sum_{j=S}^{N-1} q_j \frac{S\mu \int_0^\infty L_{j-S+1}(t)dt - g^*_{j-S+1}(S\mu)}{(j-S+1) g^*_{j-S}(S\mu)} \\ &= \sum_{j=S}^{N-1} q_j \frac{S\mu \int_0^\infty u e^{-S\mu u} H(u)^{j-S+1} du - g^*_{j-S+1}(S\mu)}{(j-S+1) g^*_{j-S}(S\mu)}.\end{aligned}$$

At last the mean waiting time for abandoned calls given entry is

$$\begin{aligned}E(W_q, Ab) &= E(W_q) - E(W_q, Sr) \\ &= \sum_{j=S}^{N-1} q_j \frac{g^*_{j-S+1}(S\mu)}{g^*_{j-S}(S\mu)} - \sum_{j=S}^{N-1} q_j \frac{S\mu \int_0^\infty u e^{-S\mu u} H(u)^{j-S+1} du - g^*_{j-S+1}(S\mu)}{(j-S+1) g^*_{j-S}(S\mu)} \\ &= \sum_{j=S}^{N-1} q_j \frac{(j-S+2) g^*_{j-S+1}(S\mu) - S\mu \int_0^\infty u e^{-S\mu u} H(u)^{j-S+1} du}{(j-S+1) g^*_{j-S}(S\mu)}.\end{aligned}$$

Now we can derive the conditional mean waiting time, which are more useful in practice, using the above results.

1.

$$E(W_q|Ab) = \frac{E(W_q, Ab)}{P(Ab|entry)} = \frac{\sum_{j=S}^{N-1} q_j \frac{(j-S+2)g^*_{j-S+1}(S\mu) - S\mu \int_0^\infty ue^{-S\mu u} H(u)^{j-S+1} du}{(j-S+1)g^*_{j-S}(S\mu)}}{\sum_{j=S}^{N-1} q_j \left[1 - \frac{S\mu g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}\right]}.$$

2.

$$E(W_q|Sr) = \frac{E(W_q, Sr)}{P(Sr|entry)} = \frac{\sum_{j=S}^{N-1} q_j \frac{S\mu \int_0^\infty ue^{-S\mu u} H(u)^{j-S+1} du - g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}}{\sum_{j=0}^{S-1} q_j + \sum_{j=S}^{N-1} q_j \frac{S\mu g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}}.$$

3.

$$E(W_q|Sr, delay) = \frac{E(W_q, Sr)}{P(Sr, delay|entry)} = \frac{\sum_{j=S}^{N-1} q_j \frac{S\mu \int_0^\infty ue^{-S\mu u} H(u)^{j-S+1} du - g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}}{\sum_{j=S}^{N-1} q_j \frac{S\mu g^*_{j-S+1}(S\mu)}{(j-S+1)g^*_{j-S}(S\mu)}}.$$

6.3.7 Numerical examples

To give some numerical illustrations for the SOQN model with general abandonment, we will consider the following example for $P(blocking), P(W_q > t), P(Ab)$ and $P(Sr)$. This example is similar to the example discussed in Chapter 3 where the parameters are $\lambda = 250/1800, \mu = 1/180, t = 20$ seconds, $\theta = 0.01$. We also let $p = 0.5$ here. To illustrate the effect of buffer size, we fix $S = 10$ and let buffer size $K = N - S$ change from 0 to 20. In addition we will compare three different patience time distributions (exponential, deterministic and uniform) with the same mean α^{-1} where $\alpha = 0.02$. For the uniform patience time distribution, we assume the support is $(0, 2\alpha^{-1})$ with mean α^{-1}.

In Figure 6.4, we compare $P(blocking)$ for different patience time distributions with the same mean. Among three patience time distributions, exponential patience has the best performance and deterministic patience has the worst. However, we find that $P(blocking)$ is not sensitive to the patience time distribution, especially for smaller K. Also $P(blocking)$ is a strictly decreasing function of K in all cases. In Figure 6.5, we compare $P(Wq > 20)$ for different patience time distributions with the same mean. Again, among three patience time

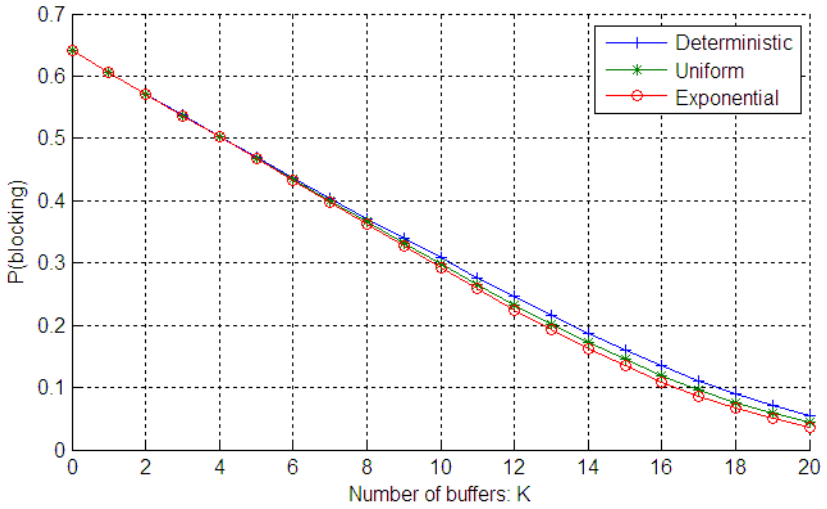

Figure 6.4: $P(blocking)$ for different patience time distributions with the same mean

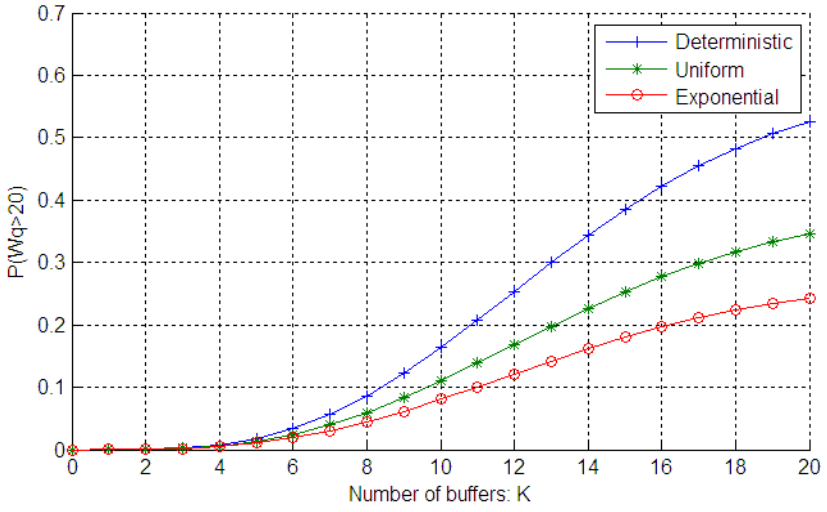

Figure 6.5: $P(Wq > 20)$ for different patience time distributions with the same mean

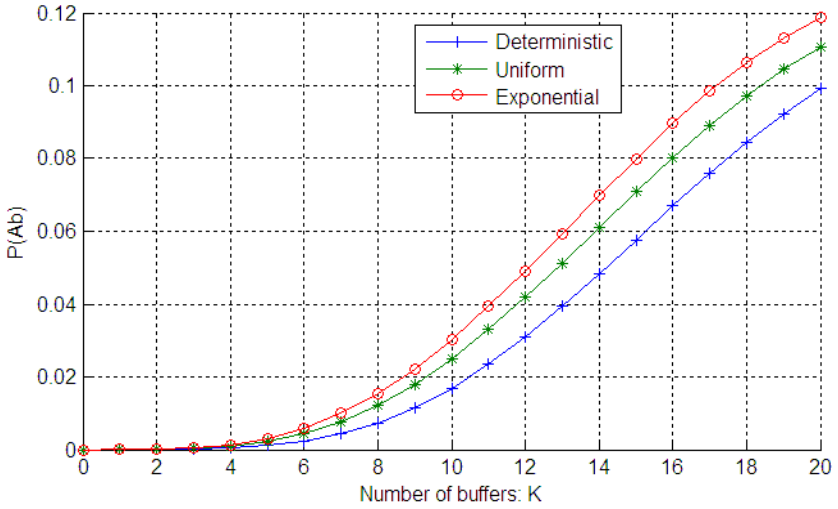

Figure 6.6: $P(Ab)$ for different patience time distributions with the same mean

distributions, exponential patience has the best performance and deterministic patience has the worst. And we find that $P(Wq > 20)$ is sensitive to the patience time distribution, especially for lager K. Also $P(Wq > 20)$ is a strictly increasing function of K in all cases. Note that we have observed similar monotonicity properties with respect to buffer size K for $P(blocking)$ and $P(W_q > 20)$ in $M/M/S/N+M$ model and we will use this observation in Chapter 7 for the call centre design problem.

Next we will consider the unconditional probabilities $P(Ab)$ and $P(Sr)$, shown in Figure 6.6 and Figure 6.7 respectively. It is clear that both $P(Ab)$ and $P(Sr)$ increase when the buffer size K increases as proved in $M/M/S/N+M$ model. For both $P(Ab)$ and $P(Sr)$, among three patience time distributions, deterministic patience has the best performance and exponential patience has the worst. The fact that in terms of $P(Sr)$, deterministic patience has the best performance agrees with Theorem 6.2.6 for $M/M/S/N+G$ model. Also we find that in terms of $P(Sr)$, uniform patience has better performance than exponential patience, which agrees with Theorem 6.2.8 for $M/M/S/N+G$ model. Note that $P(Ab)$ is sensitive to the patience time distribution, especially for lager K while $P(Sr)$ is not sensitive to the patience time distribution, especially for smaller K.

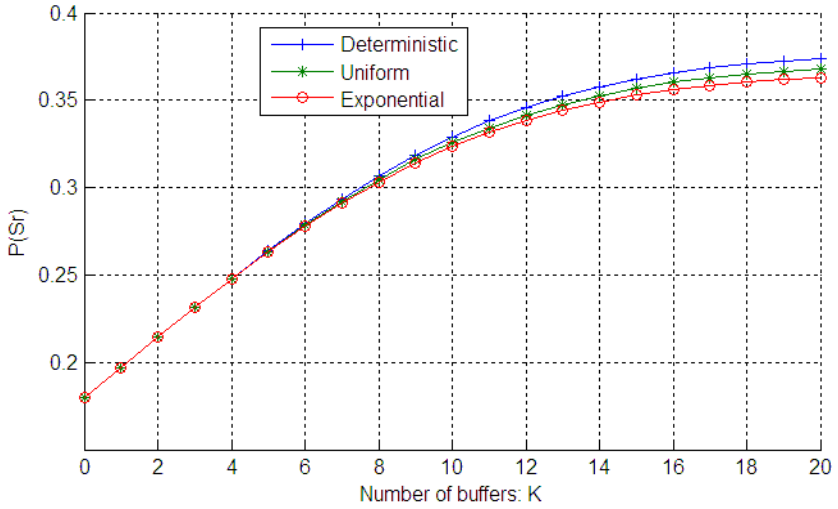

Figure 6.7: $P(Sr)$ for different patience time distributions with the same mean

6.4 Summary

In this chapter, exponential abandonment models were generalized to the general abandonment case. For single-node models, our work is based on several previous works and is a generalization of them. For example we derived a comprehensive list of formulas for performance measures in terms of new building blocks for $M/M/S/N+G$. Two special cases were studied for the building blocks: exponential and deterministic abandonment. Based on the performance analysis of $M/M/S/N+G$ model, we studied the comparison of different patience time distributions including Deterministic, Erlang and Uniform. For network model, we proposed and studied SOQN+G model in detail. In the end, numerical examples were given to illustrate the effect of different patience time distributions for SOQN+G model.

Chapter 7

An Algorithm for Call Centre Design Problem

7.1 Introduction

As discussed in Chapter 1, there are two main research problems in call centres. In previous chapters, we have been focusing on the call centre modelling and performance analysis problem. We proposed several SOQN models of call centres and gave a detailed performance analysis for each model. We also reviewed the corresponding single-node queueing models and obtained new results. Explicit expressions for performance measures can be obtained after performance analysis. Given the parameters of a call centre, these performance measures provide the call centre manager a sense of how the call centre is performing. We can also make some "what-if" experiment using expressions of performance measures, which is often desirable to the call centre manager. Another use of the expressions of performance measures involves theoretical analysis of the performance measures with respect to some model parameters as we have shown in Chapter 5 for monotonicity and convexity properties of some performance measures.

In this chapter we will focus on the call centre design problem. We will develop a design algorithm to determine the minimal number of CSRs (S) and trunk lines (N) to satisfy a given set of service level constraints. The algorithm is based on the monotonicity properties of some performance measures with respect to N. We also provide some numerical examples to illustrate the efficacy of our algorithm.

7.2 Design algorithm

In our design algorithm, for simplicity we will not attach specific costs for S and N. Therefore we will not solve the design problem through optimization in terms of costs. However, we will use the definition of cost ordering among (S, N) pairs defined in [35] and

finding the smallest pair of (S, N) in terms of this cost ordering will be our objective. This definition follows the fact that usually the cost of a trunk line is insignificant compared to the cost of a CSR. The following definition is given in [35].

Definition 7.2.1 *(S_1, N_1) is less than or equal to (S_2, N_2) in terms of cost ordering, denoted as $(S_1, N_1) \leq_C (S_2, N_2)$ if and only if $S_1 < S_2$ or we have $S_1 = S_2$ and $N_1 \leq N_2$.*

Then our design problem can be formalized as

$$\begin{cases} \min\ (S, N) \text{ in terms of cost ordering} \\ s.t.\ SL\ constraints. \end{cases}$$

The SL constraints can be any combination of two performance measures (SL_1 and SL_2) satisfying the following monotonicity properties with respect to S and N.

1. When S and other parameters are fixed, SL_1 is a decreasing function of N.
2. When N and other parameters are fixed, SL_1 is a decreasing function of S.
3. When S and other parameters are fixed, SL_2 is an increasing function of N.
4. When N and other parameters are fixed, SL_2 is a decreasing function of S.

A common combination is $SL_1 = P(blocking)$ and $SL_2 = 1 - TSF = P(W_q > t)$ as in [13, 35, 42] and we will use this combination as well. In this case, the design problem can be formalized as

$$\begin{cases} \min\ (S, N) \text{ in terms of cost ordering} \\ s.t. \begin{cases} SL_1 = P(blocking) < b \\ SL_2 = P(W_q > t) < c \end{cases} \end{cases}. \tag{7.1}$$

Our algorithm works for any model as long as the above monotonicity properties are met for SL_1 and SL_2. For example in [35], it has been proved that these monotonicity properties hold for $SL_1 = P(blocking)$ and $SL_2 = P(W_q > t)$ in $M/M/S/N$ model. Furthermore in Chapter 2 we have proved other monotonicity properties with respect to N in $M/M/S/N$ model. However for other models, it may be difficult to prove them in theory although they are intuitively correct.

For the design problem (7.1), in view of the monotonicity properties of SL_1 and SL_2 mentioned above, we have the feasible region and the optimal solution as shown in Figure 7.1.

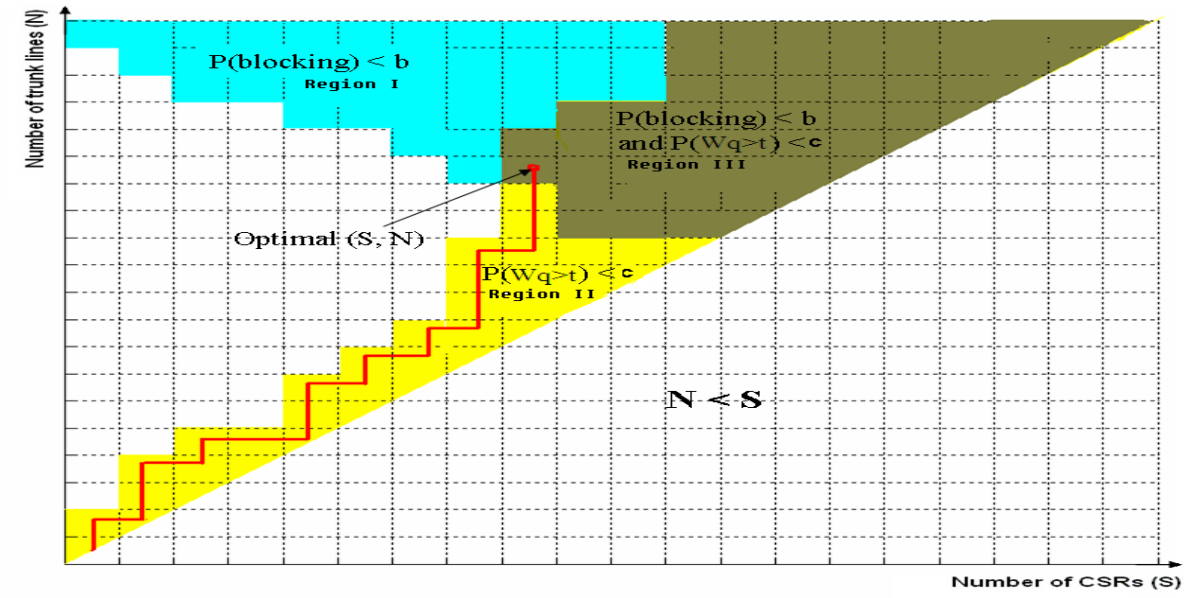

Figure 7.1: The feasible region and the optimal solution of the design problem (7.1)

In Figure 7.1, Region I is where (S, N) satisfies $P(blocking) < b$ and it has this shape because we assume the following monotonicity properties for $P(blocking)$: $P(blocking)$ is a decreasing function of N when S is fixed and a decreasing function of S when N is fixed. Region II is where (S, N) satisfies $P(W_q > t) < c$ and it has this shape because we assume the following monotonicity properties for $P(W_q > t)$: $P(W_q > t)$ is an increasing function of N when S is fixed and a decreasing function of S when N is fixed. Region III is where (S, N) satisfies both $P(blocking) < b$ and $P(W_q > t) < c$ and it is the intersection of Region I and II. According to the definition of cost ordering, the optimal solution is the bottom of the most left column of Region III.

Basically our design algorithm is a searching algorithm, which can be described following the line in Figure 7.1. We start with $N = S = 1$, where $P(W_q > t) = 0$ since no calls can wait. Then we increase N until we reach the smallest N where $P(W_q > t) < c$ is satisfied. This is guaranteed by the monotonicity property of $P(W_q > t)$ with respect to N when S is fixed. At this point we check the constraint $P(blocking)$. If $P(blocking) \geq b$, there is no intersection of Region I and II at the current S and we will increase S by 1 and keep N unchanged. The new (S, N) will satisfy $P(W_q > t) < c$ by the monotonicity property of $P(W_q > t)$ with respect to S when N is fixed. We then repeat the previous procedures until $P(blocking) < b$ for some S. Now we already obtain the right S and we

just need to reduce N until we reach the smallest N where $P(blocking) < b$ is satisfied. This is guaranteed by the monotonicity property of $P(blocking)$ with respect to N when S is fixed. The searching algorithm stops until we obtain the optimal solution or there is no solution when (S, N) reaches the given maximum (S, N). From the above description, we find that the monotonicity property of $P(blocking)$ with respect to S is not important except that it will guarantee that we will obtain an optimal solution faster i.e., the optimal S is smaller as shown in Figure 7.1. The algorithm can be written in the form of pseudocode as follows.

Algorithm 7.2.1 *Searching algorithm to get the optimal* (S, N)

1 Initialize parameters $(S = 1, N = 1, maxS, maxN)$

2 WHILE true

2.1 Compute delay $= P(W_q > t)$

2.2 WHILE delay $< c$ *AND* $N \leq maxN$

$N = N + 1$

Compute delay $= P(W_q > t)$

ENDWHILE

2.3 $N = N - 1$

2.4 Compute block $= P(blocking)$

2.5 IF block $\geq b$ *THEN*

$S = S + 1$

IF $S > maxS$ *THEN OUTPUT no solution; BREAK; ENDIF*

$N = N + 1$

ELSE

OUTPUT S

$N = N - 1$

Compute block $= P(blocking)$

WHILE block $< b$

$N = N - 1$

Compute block $= P(blocking)$

ENDWHILE

OUTPUT $N + 1$

BREAK

ENDIF

ENDWHILE

The above algorithm is similar to the algorithm given in [27]. However we can generalize this algorithm to cope with the case where only the following two monotonicity properties with respect to N are required. Note that we have proved these two monotonicity properties with respect to N for $M/M/S/N+M$ model in Chapter 5.

1. When S and other parameters are fixed, $P(blocking)$ is a decreasing function of N.
2. When S and other parameters are fixed, $P(W_q > t)$ is an increasing function of N.

In this case, when we increase S by 1 and keep N unchanged, we cannot guarantee that the new (S, N) will still satisfy $P(W_q > t) < c$ since we do not have the monotonicity property of $P(W_q > t)$ with respect to S when N is fixed. We can make some adjustment to the algorithm to make sure that after increasing S by 1, the new (S, N) will still be within the region $P(W_q > t) < c$ by reducing N. The generalized algorithm is in the following.

Algorithm 7.2.2 *The generalized searching algorithm to get the optimal* (S, N)

1 Initialize parameters $(S = 1, N = 1, maxS, maxN)$

2 WHILE true

2.1 Compute delay $= P(W_q > t)$

2.2 IF delay $< c$ *THEN*

WHILE delay $< c$ *AND* $N \leq maxN$

$N = N + 1$

Compute delay $= P(W_q > t)$

ENDWHILE

$N = N - 1$

ELSE

WHILE delay $\geq c$

$N = N - 1$

Compute delay $= P(W_q > t)$

ENDWHILE

ENDIF

2.3 Compute block $= P(blocking)$

2.4 IF block $\geq b$ *THEN*

$S = S + 1$

IF $S > maxS$ *THEN OUTPUT no solution; BREAK; ENDIF*

$N = N + 1$

ELSE

OUTPUT S

$N = N - 1$

Compute block $= P(blocking)$

WHILE block $< b$

$N = N - 1$

Compute block $= P(blocking)$

ENDWHILE

OUTPUT $N + 1$

BREAK

ENDIF

ENDWHILE

7.3 Numerical examples

In this section, we will provide some numerical examples to illustrate the efficacy of our design algorithm for different models analyzed in this thesis. We will start with models with patient calls, i.e., $M/M/S/N$ and SOQN models. Then we will consider SOQN model with different balking functions. In the end we will focus on abandonment models, i.e., $M/M/S/N + M$ model, SOQN+M model and SOQN+G model.

7.3.1 $M/M/S/N$ and SOQN models

$M/M/S/N$ and SOQN models are models with patient calls and we have obtained performance measures for these two models such as $P(blocking)$ and $P(W_q > t)$ in Chapter 2 and Chapter 3 before. Now we will apply our algorithm developed in this chapter to solve the design problem, i.e., to determine the minimal number of CSRs (S) and trunk lines (N) in terms of cost ordering given service level constraints on $P(blocking)$ and

$P(W_q > t)$. For SOQN model, we will focus on a call centre example given in Srinivasan et al. [42], which deals with a call load of 250 calls per half an hour period. The average talk time is estimated to be 180 seconds and the average IVRU processing time is 0.01 seconds or 100 seconds representing fast and slow IVRU servers respectively. Therefore, our parameters are $\lambda = 250/1800, \mu = 1/180, \theta = 100$ or 0.01. The service level constraints are $P(blocking) < b = 0.01$ and $P(W_q > t$ seconds $) < c = 0.2$, where $t = 20$. We also consider four cases: $p = 0.1, 0.5, 0.9$ and 1 as in [42]. For $M/M/S/N$ model, in order to get reasonable comparison results with SOQN model, we will add the IVRU processing time to the average talk time to obtain the new average talk time. Therefore we have $\lambda = 250/1800, \mu = 1/(180 + 0.01)$ or $1/(180 + 100)$ with the same service level constraints.

We will also consider the traditional approach to determine the minimal number of CSRs (S) and trunk lines (N) given service level constraints on $P(blocking)$ and $P(W_q > t)$, which is called EBC method in [42]. This method uses the $M/M/S$ model and the $M/M/N/N$ model in isolation and is also described similarly in [35]. We will describe the EBC method in [35] using the following algorithm. Similarly as in $M/M/S/N$ model, we consider two cases: $\mu = 1/(180 + 0.01)$ or $1/(180 + 100)$.

Algorithm 7.3.1 *The traditional EBC method to get the optimal* (S, N)

1 Obtain the adjusted arrival rate: $\lambda^* = \lambda(1 - b)$

2 Minimize S *such that* $P(W_q > t) = C(S, a^*)e^{-(S\mu - \lambda^*)t} < c$, *where* $a^* = \lambda^*/\mu$. *The solution is* S^*. *(Note that we use* $M/M/S$ *model here and* $P(W_q > t)$ *in* $M/M/S$ *model is a decreasing function of* S*).*

3 Obtain the adjusted mean service time: $\frac{1}{\mu^*} = \frac{1}{\mu} + E(W_q)$.

(Note that $E(W_q) = \frac{C(S^*, a^*)}{S^*\mu - \lambda^*}$ *is the mean waiting time in* $M/M/S$ *model)*

4 Minimize N *such that* $B(N, \lambda/\mu^*) < b$. *The solution is* N^*.

We apply the design algorithm to $M/M/S/N$ model and SOQN model with different p using the above example. We also implement the EBC method. For each model, we consider two cases: $\theta = 0.01$ or 100 for SOQN model, which corresponds to $\mu = 1/(180 + 100)$ or $1/(180 + 0.01)$ for $M/M/S/N$ model and the EBC method. The optimal design parameters (S, N) after running the algorithm are shown in Figure 7.2 and Figure 7.3 for $\theta = 0.01$ and $\theta = 100$ respectively.

For both $\theta = 0.01$ and $\theta = 100$, we find that the results in EBC method and $M/M/S/N$

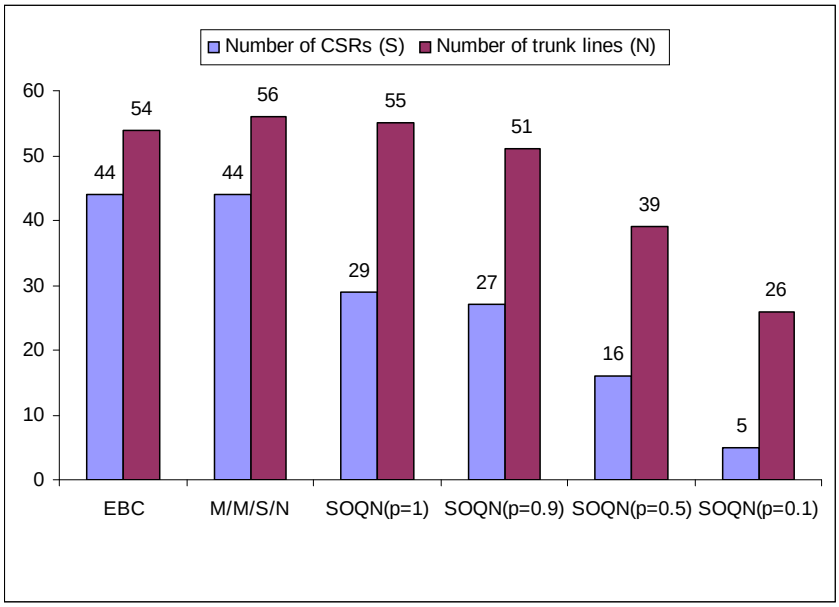

Figure 7.2: The optimal design parameters (S, N) for models with patient calls when $\theta = 0.01$

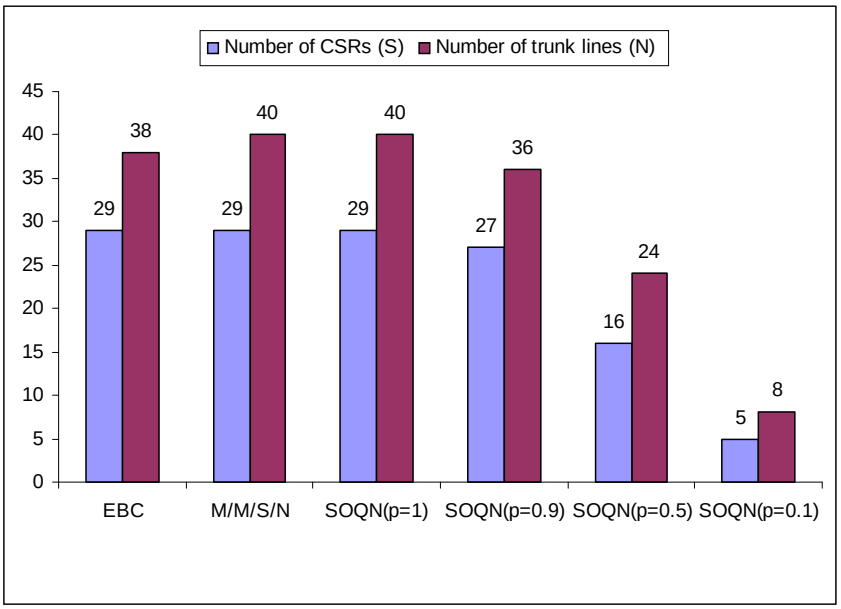

Figure 7.3: The optimal design parameters (S, N) for models with patient calls when $\theta = 100$

model are pretty close, although the EBC method underestimates the number of trunk lines a little bit compared to $M/M/S/N$ model as observed in [35]. However when $\theta = 0.01$, i.e., the IVRU servers are slow, the SOQN model with $p = 1$ requires only 29 CSRs, much smaller than 44, which is produced both by EBC method and $M/M/S/N$ model. When $\theta = 100$, i.e., the IVRU servers are fast, we have the same number of CSRs $S = 29$ for all three models. In this case, we can actually ignore the role of IVRU since the IVRU servers are so fast. For SOQN model, when p is decreased from 1 to 0.1 we find the required (S, N) are decreased as well since less calls require the service of CSRs. Another interesting observation for SOQN model is that we have the same optimal S for all p in Figure 7.2 and Figure 7.3, which shows that the required number of CSRs S is not sensitive to the IVRU processing rate θ. However the IVRU processing rate θ will mainly affect the required number of trunk lines N as shown in these two figures; the optimal N for $\theta = 0.01$ is consistently larger than that for $\theta = 100$ for all p.

In Table 7.1, we list $P(blocking)$ and $P(W_q > 20)$ corresponding to the optimal design parameters (S, N) for all models. We find that the performance of the optimal design

	$\theta = 0.01$		$\theta = 100$	
Models	$P(blocking)$	$P(W_q > 20)$	$P(blocking)$	$P(W_q > 20)$
EBC	0.0120	0.1453	0.0134	0.1418
$M/M/S/N$	0.0092	0.1644	0.0098	0.1630
SOQN($p = 1$)	0.0097	0.1644	0.0098	0.1629
SOQN($p = 0.9$)	0.0096	0.1197	0.0092	0.1187
SOQN($p = 0.5$)	0.0098	0.1452	0.0083	0.1478
SOQN($p = 0.1$)	0.0084	0.0911	0.0082	0.0838

Table 7.1: $P(blocking)$ and $P(W_q > 20)$ corresponding to the optimal design parameters (S, N) for all models

parameters produced by EBC method misses the blocking target ($P(blocking) < 0.01$) for both $\theta = 0.01$ and $\theta = 100$. For example, when $\theta = 0.01$, the optimal design parameters produced by EBC method are $S = 44$ and $N = 54$ which make $P(blocking) = 0.012 > 0.01$. However the optimal design parameters produced by $M/M/S/N$ model are $S = 44$ and $N = 56$ which make $P(blocking) = 0.0092 < 0.01$. Therefore an additional 2 trunk lines will make the performance meet the blocking target without increasing $P(W_q > 20)$ too

much (from 0.1453 to 0.1644, still less than 0.2). The same can be said for $\theta = 100$. For SOQN models, we find that the performance is under both targets.

7.3.2 SOQN model with balking

In this section, we will focus on the design problem for SOQN model with balking. We have obtained performance measures for this model such as $P(blocking)$ and $P(W_q > t)$ in Chapter 4. Now we will apply the design algorithm to the call centre example discussed before, i.e., we have $\lambda = 250/1800, \mu = 1/180, \theta = 100$ or 0.01. The service level constraints are $P(blocking) < b = 0.01$ and $P(W_q > t$ seconds $) < c = 0.2$, where $t = 20$. In addition we assume the balking function b_j has the form,

$$b_j = \begin{cases} \frac{1}{(j-S+1)^m+1} & \text{if } S \leq j < N \\ 1 & \text{if } 0 \leq j < S \end{cases},$$

where non-negative m is a measure of a call's willingness to join the queue. To show the impact of different balking functions to the optimal design parameters, we will let m be $0, \frac{1}{3}, 1, 2, 3, 4$ representing from lower balking probability to higher balking probability. The optimal design parameters (S, N) after running the algorithm are shown in Figure 7.4 and Figure 7.5 for $\theta = 0.01$ and $\theta = 100$ respectively.

For both $\theta = 0.01$ and $\theta = 100$, we find that the optimal design parameters (S, N) become smaller when b_i decreases more quickly i.e., when m is bigger. A similar observation as in SOQN model is that we have the same optimal S for all m in Figure 7.4 and Figure 7.5, which shows that the required number of CSRs S is not sensitive to the IVRU processing rate θ. However the IVRU processing rate θ will mainly affect the required number of trunk lines N as shown in these two figures; the optimal N for $\theta = 0.01$ is consistently larger than that for $\theta = 100$ for all m.

7.3.3 Exponential abandonment models

This section deals with the design problem for exponential abandonment models studied in Chapter 5 and we will focus on $M/M/S/N + M$ model and SOQN+M model. For SOQN+M model, we will apply the design algorithm to the call centre example discussed before, i.e., we have $\lambda = 250/1800, \mu = 1/180, \theta = 100$ or 0.01. We also fix $p = 1$ here. The service level constraints are $P(blocking) < b = 0.01$ and $P(W_q > t$ seconds$) < c = 0.2$,

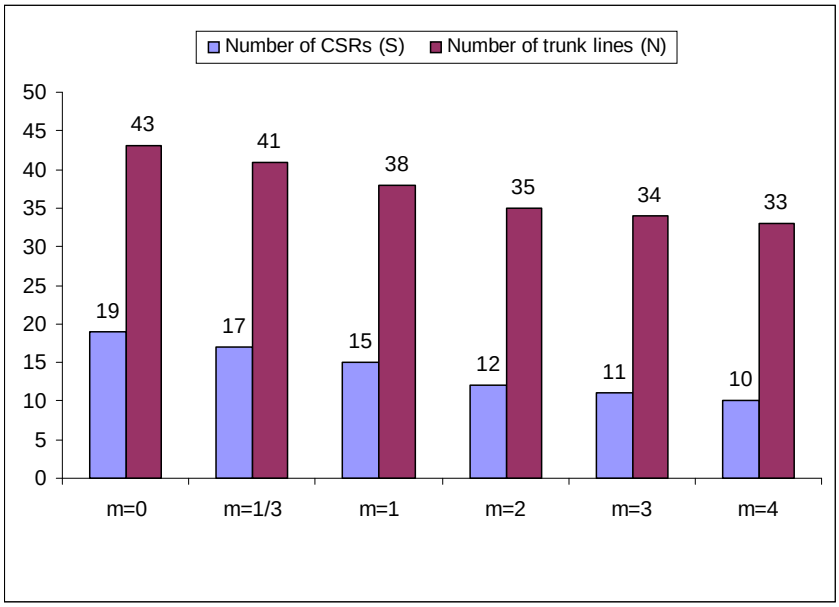

Figure 7.4: The optimal design parameters (S, N) for SOQN models with balking when $\theta = 0.01$

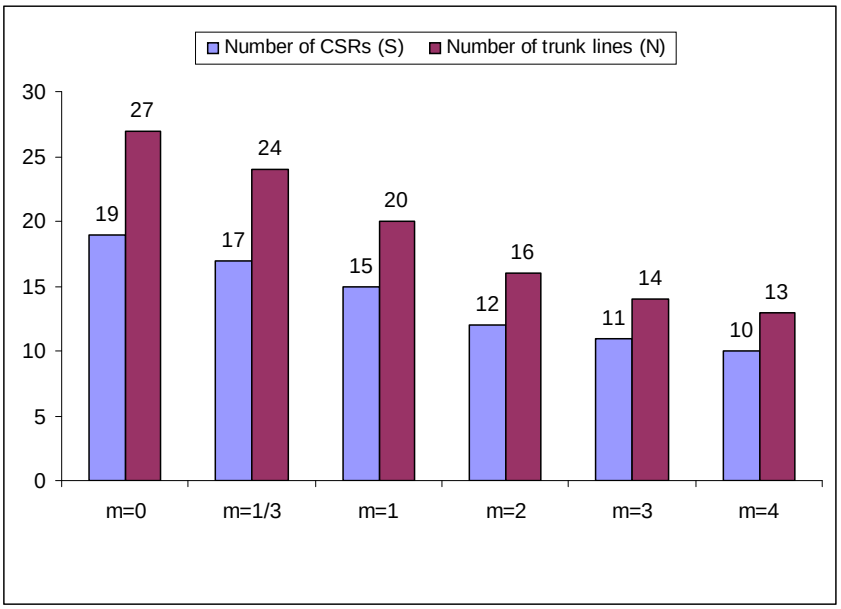

Figure 7.5: The optimal design parameters (S, N) for SOQN models with balking when $\theta = 100$

where $t = 20$. In addition, to show the impact of different abandonment rate to the optimal design parameters we will assume α be $0, 0.01, 0.02, 0.03, 0.04$ representing from lower abandonment rate to higher abandonment rate, where $\alpha = 0$ corresponds to the patient models. Similar as in the patient models, for $M/M/S/N+M$ model, we will add the IVRU processing time to the average talk time to obtain the new average talk time. Therefore we have $\lambda = 250/1800, \mu = 1/(180 + 0.01)$ or $1/(180 + 100)$ with the same service level constraints. The optimal design parameters (S, N) after running the algorithm are shown in Table 7.2 and Table 7.3 for $\theta = 0.01$ and $\theta = 100$ respectively.

	$M/M/S/N + M$		SOQN+M	
Abandonment rate	# of CSRs S	# of trunk lines N	# of CSRs S	# of trunk lines N
$\alpha = 0$	44	56	29	55
$\alpha = 0.01$	38	47	25	50
$\alpha = 0.02$	33	41	21	46
$\alpha = 0.03$	27	34	18	43
$\alpha = 0.04$	22	29	14	39
$\alpha = 0.05$	17	24	11	36

Table 7.2: The optimal design parameters (S, N) for models with exponential abandonment when $\theta = 0.01$

	$M/M/S/N + M$		SOQN+M	
Abandonment rate	# of CSRs S	# of trunk lines N	# of CSRs S	# of trunk lines N
$\alpha = 0$	29	40	29	40
$\alpha = 0.01$	25	34	25	34
$\alpha = 0.02$	21	29	21	29
$\alpha = 0.03$	18	25	18	25
$\alpha = 0.04$	14	21	14	21
$\alpha = 0.05$	11	18	11	18

Table 7.3: The optimal design parameters (S, N) for models with exponential abandonment when $\theta = 100$

For both $\theta = 0.01$ and $\theta = 100$, it is clear that the optimal design parameters (S, N) become smaller when α gets bigger since more calls abandon the system. Also when

$\alpha = 0$, we have the same (S, N) as patient models since there are no abandonment. When $\theta = 0.01$, i.e., the IVRU servers are slow, the SOQN+M model requires less CSRs than those required by $M/M/S/N$ model for all α. However when $\theta = 100$, i.e., the IVRU servers are fast, we have the same optimal design parameters (S, N) for both $M/M/S/N$ model and SOQN+M model. In this case, we can actually ignore the role of IVRU since the IVRU servers are so fast. Again for SOQN+M model we observe the same optimal S for all α in these two tables, which shows that the required number of CSRs S is not sensitive to the IVRU processing rate θ. However the IVRU processing rate θ will mainly affect the required number of trunk lines N; the optimal N for $\theta = 0.01$ is consistently larger than that for $\theta = 100$ for all α.

7.3.4 SOQN model with general abandonment

In this section, we will study the design problem for SOQN model with general abandonment, i.e., SOQN+G model discussed in Chapter 6. We will compare three different patience time distributions (exponential SOQN+M, deterministic SOQN+D and uniform SOQN+U) with the same mean α^{-1} where $\alpha = 0.01$. For SOQN+U model, we assume the support is $(0, 2\alpha^{-1})$ with mean α^{-1}. The other call centre parameters are similar as before, i.e., we have $\lambda = 250/1800, \mu = 1/180, \theta = 100$ or 0.01. We also fix $p = 1$ here. The service level constraints are $P(blocking) < b = 0.01$ and $P(W_q > t \text{ seconds}) < c = 0.2$, where $t = 20$. The optimal design parameters (S, N) after running the algorithm are shown in Table 7.4.

	$\theta = 0.01$		$\theta = 100$	
Models	# of CSRs S	# of trunk lines N	# of CSRs S	# of trunk lines N
SOQN+M	25	50	25	34
SOQN+U	27	52	27	36
SOQN+D	29	55	29	40

Table 7.4: Comparison of the optimal design parameters (S, N) for SOQN models with general abandonment when $\alpha = 0.01$

From Table 7.4, one finds that there is an order relationship among three patience time distributions in terms of the values of the optimal design parameters (S, N) for both $\theta = 0.01$ and $\theta = 100$. The optimal (S, N) required by deterministic patience is bigger than

that required by uniform patience and the optimal (S, N) required by uniform patience is bigger than that required by exponential patience. This fact is consistent with the same order relationship in terms of $P(blocking)$ and $P(W_q > t)$ observed in Chapter 6 (Refer to Figure 6.4 and Figure 6.5). In other words, among three patience time distributions, exponential patience has the best performance and deterministic patience has the worst in terms of $P(blocking)$ and $P(W_q > t)$. Also we have observed in Chapter 6 that in terms of $P(Ab)$, deterministic patience has the best performance and exponential patience has the worst (Refer to Figure 6.6), which implies there are more abandonment in SOQN+M model than other two models. This fact also explains why SOQN+M requires the least (S, N). Again for all three models we observe the same optimal S for both cases, $\theta = 100$ or 0.01, which shows that the required number of CSRs S is not sensitive to the IVRU processing rate θ. However the IVRU processing rate θ will mainly affect the required number of trunk lines N; the optimal N for $\theta = 0.01$ is consistently larger than that for $\theta = 100$ for all three models.

7.4 Summary

Call centre design problem was studied in this chapter. We first formalized the problem into an optimization problem and then described the searching algorithm. Numerical examples were given to illustrate the efficacy of our algorithm for various models including patient, balking and abandonment models.

Chapter 8

Summary and Future Work

In this last chapter, we will provide a summary of the work in this thesis and point out some possible future work.

8.1 Summary

Our work is about call centre modelling, analysis and design. In terms of modelling, traditionally call centres have been modelled as single-node queueing systems. Based on the SOQN model proposed by Srinivasan et al. [42], we have studied the SOQN model with balking and abandonment (both exponential and general patience time distributions). We also studied the corresponding single-node queueing systems and obtained new results. In terms of call centre design, we have developed a design algorithm to determine the minimal number of CSRs (S) and trunk lines (N) to satisfy a given set of service level constraints. Our main contributions are listed in the following.

1. For the single-node Markovian queueing models of call centres, especially for $M/M/S/N$ model, based on the work of [35], we expressed the performance measures in terms of Erlang B formula, which facilitates the computation. We also proved new monotonicity properties of several performance measures with respect to N based on the expressions in terms of Erlang B formula.

2. For SOQN model, we used two methods to derive the product form solution of the queue length process and the distribution of the total number of calls in the system. We proposed an algorithm to compute the blocking probability. In the derivation of the waiting time distribution, we used a new method than the one used in the original paper [42].

3. For SOQN model with balking, we proved that the queue length process still has product form solution in equilibrium and we also derived the waiting time distribution.

4. For single-node exponential abandonment models: $M/M/S+M$ and $M/M/S/N+M$, we focused on the computational aspects by expressing the exact performance measures in terms of special functions and Erlang B formula. The analysis is new and we have provided a unified and comprehensive list of expressions for performance measures. Especially for $M/M/S/N+M$ model, we proved monotonicity and concavity properties for $P(Sr)$ using a method that simplifies the work in [26]. Monotonicity properties for $P(blocking)$ and $P(W_q > t)$ were also proved and these properties are important to the call centre design algorithm in Chapter 7.

5. For SOQN+M model, our work is a generalization and correction of [45]; we used a new approach not only to rederive the formulas for the performance measures correctly, but also introduce new results.

6. For single-node general abandonment models: $M/M/S+G$ and $M/M/S/N+G$, our work is based on several previous works and is a generalization of them. Again we focused on the computational aspects of the performance measures. For example we expressed the performance measures of $M/M/S/N+G$ model in terms of new building blocks. We also studied the comparison of different patience time distributions and generalized some previous results to $M/M/S/N+G$ model.

7. We proposed and studied SOQN+G model in detail.

8. We have developed a design algorithm to determine the minimal number of CSRs (S) and trunk lines (N) to satisfy $P(blocking)$ and $P(W_q > t)$ constraints. We have provided some numerical examples to illustrate the efficacy of our algorithm for some models we have studied. There are some interesting observations based on the numerical examples. For example the required number of CSRs S is not sensitive to the IVRU processing rate θ in all SOQN models (patient, balking and abandonment). And θ will mainly affect the required number of trunk lines N. Another example is for SOQN+G model, there is an order relationship among three patience time distributions (Deterministic, Uniform and Exponential) in terms of the values of the optimal design parameters (S, N). Exponential patience requires the least (S, N) and therefore has the best performance while deterministic patience requires the biggest (S, N) and therefore has the worst performance. This observation is consistent with the theoretical results proved for $M/M/S/N+G$ model.

8.2 Future work

In this section, we will discuss the possible future work.

1. We have used the direct algebraic method to prove monotonicity and convexity properties of several performance measures for $M/M/S/N$ and $M/M/S/N+M$ model. However it seems hard to prove the monotonicity and convexity properties for other models. Other methods could be explored for this purpose.

2. For $M/M/S/N+G$ model, the comparison of different patience time distributions can be studied further and it would be nice if the results can be generalized to SOQN+G model.

3. Retrial queues are characterized by the feature that any arriving call who finds all servers or all waiting positions occupied may repeat its demand after a random amount of time. Retrial is another important feature for call centres. There is a large literature devoted to retrial queues. However only a few papers studied retrial in the context of call centres. A possible reason is that call centres are characterized by a large number of servers, while only retrial queues with several servers have closed-form solution. How to incorporate retrial to the semiopen network model and how to analyze it will be a direction of future work.

Bibliography

[1] Abou El-Ata, M. O., and Hariri, A. M. A. The $M/M/C/N$ queue with balking and reneging. *Computers and Operations Research 19* (1992), 713–716.

[2] Abramowitz, M., and Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1972.

[3] Ancker, C., and Gafarian, A. Some queueing problems with balking and reneging I. *Operations Research 11* (1963), 88–100.

[4] Ancker, C., and Gafarian, A. Some queueing problems with balking and reneging II. *Operations Research 11* (1963), 928–937.

[5] Baccelli, F., and Hebuterne, G. On queues with impatient customers. In *Performance '81* (Amsterdam, The Netherlands, 1981), F. Kylstra, Ed., North-Holland Publ. Cy., pp. 159–179.

[6] Barrer, D. Queueing with impatient customers and ordered service. *Operations Research 5* (1957), 650–656.

[7] Bolch, G., Greiner, S., de Meer, H., and Trivedi, K. S. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. Wiley-Interscience, New York, NY, USA, 1998.

[8] Boots, N. K., and Tijms, H. A multiserver queueing system with impatient customers. *Management Science 45*, 3 (1999), 444–448.

[9] Borst, S., Mandelbaum, A., and Reiman, M. Dimensioning large call centers. *Operations Research 52*, 1 (2004), 17–34.

[10] Boucherie, R. *Product-Form in Queueing Networks*. Ph.D. Thesis, Free University, Amsterdam, The Netherlands, 1992.

[11] Brandt, A., and Brandt, M. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation 35*, 1-2 (1999), 1–18.

[12] Brandt, A., and Brandt, M. Asymptotic results and a markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems 41* (2002), 73–94.

[13] Brandt, A., Brandt, M., Spahl, G., and Weber, D. Modeling and optimization of call distribution systems. In *Proceedings of the 15th International Teletraffic Conference* (1997), V. Ramaswani and P. E. Wirth, Eds., Elsevier Science, pp. 133–144.

[14] Chen, H., and Yao, D. *Fundamentals of queueing networks: performance, asymptotics and optimization*. Springer-Verlag, New York, 2001.

[15] Choi, B. D., Kim, B., and Chung, J. $M/M/1$ queue with impatient customers of higher priority. *Queueing Systems 38*, 1 (2001), 49–66.

[16] Cleveland, B., and Mayben, J. *Call Center Management On Fast Forward*. Call Center Press, Maryland, 1997.

[17] Cooper, R. *Introduction to Queueing Theory*. North-Holland, New York, 1981.

[18] Cox, D. R. *Renewal Theory*. Chapman and Hall, London, 1967.

[19] Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., and Avramidis, A. N. Markov chain models of a telephone call center with call blending. *Computers and Operations Research 34*, 6 (2007), 1616 – 1645. Part Special Issue: Odysseus 2003 Second International Workshop on Freight Transportation Logistics.

[20] Gans, N., Koole, G., and Mandelbaum, A. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management 5* (2003), 79–141.

[21] Garnett, O., Mandelbaum, A., and Reiman, M. Designing a call center with impatient customers. *Manufacturing and Service Operations Management 4*, 3 (2002), 208–227.

[22] Gnedenko, B., and Kovalenko, I. *Introduction to Queueing Theory*. Birkhauser Boston Inc., Cambridge, MA, 1968.

[23] Grassmann, W. K. Finding the right number of servers in real-world queuing-systems. *Interfaces 18*, 2 (Mar-Apr 1988), 94–104.

[24] Gross, D., and Harris, C. M. *Fundamentals of Queueing Theory*, third ed. Wiley, New York, NY, 1998.

[25] Harel, A. Convexity properties of the Erlang loss formula. *Operations Research 38*, 3 (1990), 499–505.

[26] Jouini, O., and Dallery, Y. Monotonicity properties for multiserver queues with reneging and finite waiting lines. *Probab. Eng. Inf. Sci. 21*, 3 (2007), 335–360.

[27] Khudyakov, P. *Designing a Call Center with an IVR (Interactive Voice Response)*. M.Sc. Thesis, Technion, Haifa, Israel, 2006.

[28] Kleinrock, L. *Queueing Systems, Vol. 1.* Wiley, New York, NY, 1975.

[29] Koole, G. Performance analysis and optimization in customer contact centers. In *QEST '04: Proceedings of the The Quantitative Evaluation of Systems, First International Conference* (Washington, DC, USA, 2004), IEEE Computer Society, pp. 2–5.

[30] Koole, G., and Mandelbaum, A. Queueing models of call centers: An introduction. *Annals of Operations Research 113* (2002), 41–59.

[31] Koole, G., and Pot, A. A note on profit maximization and monotonicity for inbound call centers. Tech. rep., Department of Stochastics, Vrije Universiteit Amsterdam, 2006.

[32] Mandelbaum, A. Call centers (centres) research bibliography with abstracts. Tech. rep., Technion, Haifa, Israel, Version 7: May 4, 2006.

[33] Mandelbaum, A., and Zeltyn, S. The Palm/Erlang-A queue, with applications to call centers. Tech. rep., Teaching note to Service Engineering course, 2004.

[34] Mandelbaum, A., and Zeltyn, S. The impact of customers patience on delay and abandonment: some empirically-driven experiments with the $M/M/n+G$ queue. *OR Spectrum 26* (July 2004), 377–411(35).

[35] Massey, W., and Wallace, R. An optimal design of the $M/M/C/K$ queue for call centers. Tech. rep., Department of Operations Research and Financial Engineering ,Princeton University, 2006.

[36] Movaghar, A. On queueing with customer impatience until the beginning of service. *Queueing Systems 29*, 2/4 (1998), 337–350.

[37] Movaghar, A. On queueing with customer impatience until the end of service. *Stochastic Models 22* (Number 1/2006), 149–173(25).

[38] Reynolds, J. F. The stationary solution of a multiserver queuing model with discouragement. *Operations Research 16*, 1 (1968), 64–71.

[39] Riordan, J. *Stochastic service systems*. John Wiley and Sons, Inc, New York and London, 1962.

[40] Ross, S. M. *Introduction to Probability Models, Eighth Edition*. Academic Press, January 2003.

[41] Sobel, M. J. Note–simple inequalities for multiserver queues. *MANAGEMENT SCIENCE 26*, 9 (1980), 951–956.

[42] Srinivasan, R., Talim, J., and Wang, J. Performance analysis of a call center with interactive voice response units. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research 12*, 1 (June 2004), 91–110.

[43] Stanford, D., and Grassmann, W. K. Bilingual server call centers. In *Analysis of Communi-cation Networks: call centers, traffic and performance* (2000), D. McDonald and S. R. E. Turner, Eds., American Mathematical Society, Providence, RI, pp. 31–47.

[44] Stolletz, R. *Performance analysis and optimization of inbound call centers*. No. 528 in Lecture notes in economics and mathematical systems. Springer, Berlin [u.a.], 2003.

[45] Wang, J., and Srinivasan, R. Performance analysis of a call center with interactive voice response units and abandonment. Tech. rep., University of Saskatchewan, 2005.

[46] Whitt, W. Improving service by informing customers about anticipated delays. *Management Science 45*, 2 (1999), 192–207.

[47] Whitt, W. Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Operations Research 54*, 2 (2006), 247–260.

[48] Wolff, R. W. Poisson Arrivals See Time Averages. *OPERATIONS RESEARCH 30*, 2 (1982), 223–231.

[49] Yom-Tov, G. *Queues in Hospitals: Semi-Open Queueing Networks in the QED Regime*. Ph.D. Research Proposal, Technion, Haifa, Israel, 2007.

[50] Zeltyn, S. *Call centers with impatient customers: Exact analysis and many-server asymptotics of the M/M/n+G queue*. Ph.D. Thesis, Technion, Haifa, Israel, 2005.