Asymptotic Analysis and Performance-based Design of Large Scale Service and Inventory Systems

by

Isilay Talay Degirmenci

Department of Business Administration
Duke University

Date: _____

Approved:

_____
Otis B. Jennings, Advisor

_____
Fernando G. Bernstein

_____
Vidyadhar G. Kulkarni

_____
Paul H. Zipkin

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Business Administration
in the Graduate School of Duke University
2010

Abstract

(Business Administration)

Asymptotic Analysis and Performance-based Design of Large Scale Service and Inventory Systems

by

Isilay Talay Degirmenci

Department of Business Administration
Duke University

Date: _____

Approved:

_____
Otis B. Jennings, Advisor

_____
Fernando G. Bernstein

_____
Vidyadhar G. Kulkarni

_____
Paul H. Zipkin

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Business Administration in the Graduate School of Duke University
2010

Abstract

Many types of services are provided using some equipment or machines, e.g. transportation systems using vehicles. Designs of these systems require capacity decisions, e.g., the number of vehicles. In my dissertation, I use many-server and conventional heavy-traffic limit theory to derive asymptotically optimal capacity decisions, giving the desired level of delay and availability performance with minimum investment. The results provide near-optimal solutions and insights to otherwise analytically intractable problems.

The dissertation will comprise two essays. In the first essay, "Asymptotic Analysis of Delay-based Performance Metrics and Optimal Capacity Decisions for the Machine Repair Problem with Spares," I study the M/M/R machine repair problem with spares. This system can be represented by a closed queuing network. Applications include fleet vehicles' repair and backup capacity, warranty services' staffing and spare items investment decisions. For these types of systems, customer satisfaction is essential; thus, the delays until replacements of broken units are even more important than delays until the repair initiations of the units. Moreover, the service contract may include conditions on not only the fill rate but also the probability of acceptable delay (delay being less than a specified threshold value).

I address these concerns by expressing delays in terms of the broken-machines process; scaling this process by the number of required operating machines (or the number of customers in the system); and using many-server limit theorems (limits

taken as the number of customers goes to infinity) to obtain the limiting expected delay and probability of acceptable delay for both delay until replacement and repair initiation. These results lead to an approximate optimization problem to decide on the repair and backup-capacity investment giving the minimum expected cost rate, subject to a constraint on the acceptable delay probability. Using the characteristics of the scaled broken-machines process, we obtain insights about the relationship between quality of service and capacity decisions. Inspired by the call-center literature, we categorize capacity level choice as efficiency-driven, quality-driven, or quality- and efficiency-driven. Hence, our study extends the conventional call center staffing problem to joint staffing and backup provisioning. Moreover, to our knowledge, the machine-repair problem literature has focused mainly on mean and fill rate measures of performance for steady-state cost analysis. This approach provides complex, nonlinear expressions not possible to solve analytically. The contribution of this essay to the machine-repair literature is the construction of delay-distribution approximations and a near-optimal analytical solution. Among the interesting results, we find that for capacity levels leading to very high utilization of both spares and repair capacity, the limiting distribution of delay until replacement depends on one type of resource only, the repair capacity investment.

In the second essay, "Diffusion Approximations and Near-Optimal Design of a Make-to-Stock Queue with Perishable Goods and Impatient Customers," I study a make-to-stock system with perishable inventory and impatient customers as a two-sided queue with abandonment from both sides. This model describes many consumer goods, where not only spoilage but also theft and damage can occur. We will refer to positive jobs as individual products on the shelf and negative jobs as backlogged customers. In this sense, an arriving negative job provides the service to a waiting positive job, and vice versa. Jobs that must wait in queue before potential matching are subject to abandonment. Under certain assumptions on the magni-

tude of the abandonment rates and the scaled difference between the two arrival rates (products and customers), we suggest approximations to the system dynamics such as average inventory, backorders, and fill rate via conventional heavy traffic limit theory.

We find that the approximate limiting queue length distribution is a normalized weighted average of two truncated normal distributions and then extend our results to analyze make-to-stock queues with/without perishability and limiting inventory space by inducing thresholds on the production (positive) side of the queue. Finally, we develop conjectures for the queue-length distribution for a non-Markovian system with general arrival streams. We take production rate as the decision variable and suggest near-optimal solutions.

To my son, Efe Degirmenci

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

Chapter 1

n	Number of operable units required to work continuously at the demand base.
s_n	Repair center capacity.
m_n	Maximum number of units to be held as spares.
λ	Exponential breakdown rate.
μ	Exponential repair rate.
ρ	$\equiv \lambda/\mu$.
$\bar{s}$	Order n repair capacity parameter to describe s_n.
$\bar{m}$	Order n spares capacity parameter to describe m_n.
$\hat{s}_n$	Order $\sqrt{n}$ repair capacity parameter to describe s_n.
$\hat{m}_n$	Order $\sqrt{n}$ spares capacity parameter to describe m_n.
$\hat{s}$	$\equiv \lim_{n\to\infty} \hat{s}_n$.
$\hat{m}$	$\equiv \lim_{n\to\infty} \hat{m}_n$.
$D_n^{rpl}(s_n, m_n)$	Delay until replacement for capacity choices (s_n, m_n).
$D_n^{rpi}(s_n, m_n)$	Delay until repair initiation of a broken machine for capacity choices (s_n, m_n).
T	Threshold for acceptable delay.
w	Cost rate per server staffed at the repair center.
c	Cost rate for each spare unit at the spares center.

p	Penalty cost rate assessed for a customer waiting for a replacement.
h	Penalty cost rate for broken machine waiting for repair.
$C(s_n, m_n)$	System cost rate for repair and spares capacity level choices s_n, m_n.
Φ	Standard normal distribution function.
ϕ	Standard normal density function.
φ	$\equiv 1 + 1/\rho + \mu T$.

Chapter 2

Q	Queue length for double-sided queue.
n	Index for sequence of double-sided queueing systems defined by scaling assumptions.
p_0	$\equiv P(Q(\infty) = 0)$.
α	Production completion rate.
μ	Customer arrival rate.
γ_1	Inventory perishing rate.
γ_2	Customer abandonment rate.
β	Scaled difference between the production and customer arrival rates.
θ_1	Inventory perishing rate for sequence of systems indexed by n.
θ_2	Customer abandonment rate for sequence of systems indexed by n.
h	Hazard rate of standard normal distribution.
$\hat{h}$	Holding cost rate.
b	Backorder cost rate.
w	Production capacity cost per time unit.
p	Abandonment penalty per customer per time unit.
$C(\alpha)$	Cost function w.r.t production rate chosen.

$\hat{C}(\beta)$	Approximate cost function w.r.t β chosen.
$\hat{T}$	Inventory threshold.
T	Scaling constant for inventory threshold.
μ_P	Interarrival time mean for production completions for the model with general distributions approximation.
μ_C	Interarrival time mean for customer arrivals for the model with general distributions approximation.
$sigma_P^2$	Interarrival time variance for production completions for the model with general distributions approximation.
$sigma_C^2$	Interarrival time variance for customer arrivals for the model with general distributions approximation.

Abbreviations

CLT	Central limit theorem.
ED	Efficiency-driven capacity level preference.
QD	Quality-driven capacity level preference.
QED	Quality- and efficiency-driven capacity level preference.
QED|ED	Quality- and efficiency-driven repair, and efficiency-driven spares capacity level preference (other combinations are also used similarly by expressing repair capacity level preference first).
FIFO	first-in-first-out.
cdf	cumulative distribution function.

Acknowledgements

First, I would like to gratefully acknowledge the enthusiastic supervision of Dr. Otis B. Jennings in all steps of the development of this work. I have learned and achieved more than I ever dreamed of by his precious guiding and help, and I am proud of being one of his students. I am grateful to Prof. Fernando G. Bernstein, Prof. Paul H. Zipkin, and Prof. Vidyadhar G. Kulkarni for taking part in my thesis committee, for critical reading of this dissertation, and for their valuable suggestions and comments. I would like to thank to all my friends, especially Murat Hakan Ozturk and Ugur Celikyurt, for their valuable friendship. And the last but not the least, I would like to thank to my parents for believing in me and trusting me at every step I have taken; and to my husband Oguz and my son Efe for their support and encouragement.

1

Asymptotic Analysis of Delay-based Performance Metrics and Optimal Capacity Decisions for the Machine Repair Problem with Spares

1.1 Introduction

One of the main characteristics of successful systems is when unpredictable interruptions occur, little time is wasted. The secret of this appealing result lies within the system design and management. In this essay, our objective will be to provide insights about how to design a large-scale service system that will lead to less delays after interruptions.

Delays have different implications depending on who is experiencing them. If the service is provided by an operator and equipment pair, e.g. a truck fleet, redundant workforce costs might be incurred if a truck driver (operator) experiences a long delay to receive his truck (equipment) from the repair shop of his fleet. Moreover, a company's image is badly affected if the customers experience unacceptably long delays such as having to wait too much for a new computer from warranty services. Since both types of delay (experienced by operators and customers) will decrease the

profit, we will analyze the relationship between the system design parameters and these delays.

Consider a system serving a demand base requiring needs to be met continuously. This demand might consist of customers who purchased a product with warranty coverage, machines having a common type of component subject to spontaneous breakdowns, or bus routes of a city in need of continuous and smooth transportation. One aspect of delivering exceptional service in these types of systems is when continuous service gets interrupted, delays stay within acceptable bounds. Depending on the sector, the first reflection of this motive on the system design could be on the amount of replacement goods kept by the warranty services, the amount of spare parts kept by manufacturing companies (or maintenance providers), or on the number of spare vehicles in a fleet. Hence, satisfying this type of demand will necessitate carefully made backup capacity investment. Moreover, it is important not to only have enough backup resources but also to have sufficient repair capacity for fixing the broken units so that they can be resent to the demand when needed.

To address the concerns above, we model the system as a machine-repairmen problem with spares where a certain number of identical units have to be kept as working (enough to cover the demand base), and these units break down spontaneously. When this happens, the broken unit (e.g. pc, machine component, city transit bus) is sent to the repair center while the backup center provides an operational unit to the demand as soon as possible. The circulation of units within the system is shown in Figure 1.1. All the spare units at the backup center are identical to the ones at the demand base so all units circulating the system are identical. The servers at the repair center are identical, each server can only work on one broken unit, and each broken unit can only be worked on by one server. Simultaneous presence of idle servers and waiting to be repaired units is not allowed. Our objective is to determine the backup and maintenance capacity levels to keep delays within

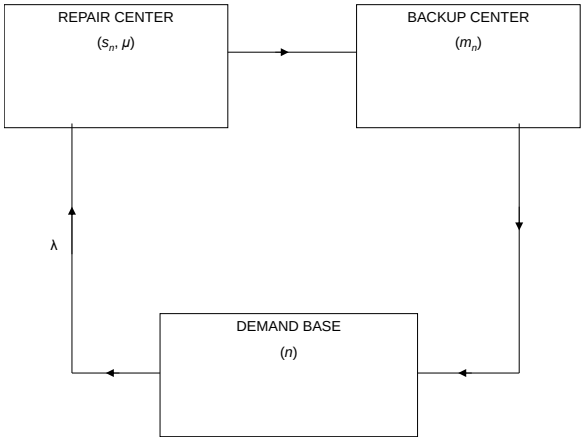

FIGURE 1.1: Circulation of Units Within the System

acceptable bounds and to achieve this with minimum amount of investment. Further details about the model will be given in Section 1.3, but here we will point to some characteristics of the model and the reasons for choosing them in our framework.

First, notice that the size of the demand base is taken as constant (n), and the number of units circulating the system is the sum of the demand size and backup capacity ($n+m_n$). Thus, we are using a closed model. It can be argued that as units start to wear, they will be replaced by new ones before they become unable to be operated in acceptable performance levels, such as retired fleet buses, airplanes, etc. getting replaced by new ones. While a bank customer would like to leave after his transaction is finished, a city needs its transit bus fleet to work for most of the day or a customer values a product when it stays functioning all the time. In this line of work, the total service demand from the system depends on population of a city, market share of a firm, utilization of machines with a common component and other factors that are not subject to frequent major changes. Then, the system capacity will not need to be changed at short intervals, which makes working with a closed model possible.

Second, notice that the units are assumed to be identical. Consider a situation

where you have different types of machines, vehicles, goods, etc. that you need to repair when they break down. Since all of these are different, they might need different tools, spare parts, and skills to fix, which creates extra burden on the maintenance management and adds to the variations in the delays after breakdowns. Moreover, other factors such as training of the operators (for ex, pilot training in airline fleets) may make it difficult to maintain different types of units within the system. Thus, large-scale service systems may opt to have identical operating units, e.g. Southwest and Alaska Airlines.

Finally, notice that we assume the replacement provided by the backup are not necessarily the same unit that has broken down at the demand base. However, the users might be attached to the units and want to retrieve the same unit after repair completion, e.g. taxi drivers might be accustomed to a particular car. When the units are identical, these kinds of attachments will occur much less because running the system more fluidly implies earning more money and hence users are likely to compromise from their personal attachments. In other words, if all the cars are the same and the vehicle they provide after breakdown is the same as the broken one, a taxi driver will prefer to get back to work (if possible, immediately) and earn more money instead of waiting for the repair completion.

To summarize, in this essay we study the machine repair problem with spares and focus on the relationship between the backup and repair capacity investment decisions and two types of delay associated with them: *replacement and repair initiation delays*. Both types of delays occur when a unit breaks down. At that moment, if there is an operational unit waiting at the backup center, it is sent to the demand (or customers) to replace the broken one. The transportation time is assumed to be negligible. Notice that we can deduce whether a delay for replacement would occur by checking the number of broken units (or machines) at the moment of breakdown. Since the model is a closed network model, if there are more broken machines than

the whole backup capacity (or spare machines), it means no operational machines can be kept idle at the backup center and all of them are currently working for customers. Thus, a replacement delay will be experienced.

Similarly, a repair initiation delay will occur if at the time of breakdown there are more broken machines than the whole repair capacity (or repairmen) because all repairmen will be busy working on the previously broken machines. This suggests that all performance measures regarding delays can be obtained by observing the random process of broken machines in the system.

Following the discussion above, although it might seem that obtaining the distribution of the broken machines in the system would be enough to uncover the capacity investment and delay relationship, expressing delays by using the distribution of broken machines usually leads to complex expressions hard to analyze. However, certain scaling approaches can be applied to this process to obtain the limiting values of delay performance measures when the scaling parameters grow large.

Here we scale the process of broken machines by the number of customers (n) and at the same time express the capacity investment decisions as a function of n. Using Markovian assumptions for the time until breakdown and repair, we show that as the number of customers grows large, the process of the ratio of broken machines to the number of customers converges to a deterministic process giving an equilibrium equation. Depending on the preferences of top management about how the firm would place itself on its respective market, this equation establishes the effect of the choice of capacity investment parameters on the broken machines process and hence delays. What would be the firm's level of quality of service? Would they be cost-focused? Would they choose paying more while giving a more satisfactory service? These preferences should be determined for both machine replacement and repair services, and the limiting values of scaled backup and repair capacity levels reflecting all feasible preferences can be explained on a two-dimensional graph (Figure 1.2).

The first performance measure we will establish is the limiting expected delay as n grows large, and a scaled investment cost function will be minimized, giving a set of cost-efficient alternatives for backup and repair capacity investment decisions. Hence, the setting would be to determine the expected delay based on strategic preferences and use this performance measure to obtain the efficient frontier of cost-minimizing capacity level choices.

Besides the optimum capacity investment alternatives that give the minimum cost and an acceptable level of delay, one would like to consider the probability of delay exceeding an arbitrary threshold value. This will also depend on the capacity investment decisions on a more sensitive scaling and the second set of results give the limiting values as n grows large for the distributions of the two types of delay.

Since the distribution of a random quantity possesses a lot more information than its expected value, characterizing it might also require a more structured modeling mechanism. In this study we give a partial characterization of the distribution of delay until replacement and a complete characterization of the delay until repair initiation distribution with graphical representations as well.

We believe this essay contributes to machine repair problem and many server limit theorems literature in two ways. First, as Haque and Armstrong (2007) say, "A factor common to most MIP(machine interference problem) research to date is the tendency to focus on the *mean* value of system performance measures, such as the average number of operational machines. We think it would be of interest to look beyond these averages to also examine some measure of the *variation* in system performance (e.g. standard deviation, 95th percentile, etc.)." By giving analytically tractable approximations for delay distributions we take a first step for another level in performance analysis on machine repair problem. Second, our model has two types of capacity, repairmen and spares, hence we complement staffing decisions in service systems via many-server limit theorems, with the addition of simultaneous

spares capacity decisions. We believe this addition can be useful when services that are provided via the use of some equipment are considered.

In Sections 1.2 we give the literature review. Then, Section 1.3 presents the model and summarizes our method of asymptotic analysis via many-server limit theorems, Section 1.4 provides the analysis for the relationship between the delay until replacement performance and capacity decisions, Section 1.5 gives the same analysis for delay until repair initiation performance, Section 1.6 gives a comparison between the exact or simulation results and the approximations as well as a comparison of systems with exponential and nonexponential distributions, and Section 1.7 gives the conclusion and discusses possible extensions.

1.2 Literature Review

Machine repairmen problem(MRP) or *Machine interference problem* (MIP) is a practical problem with many application areas as manufacturing, telecommunications, coal shipment, aircraft(or other types of fleet) maintenance, warranty services, and so on. The main framework can be summarized as having n machines subject to breakdowns with s servers to repair them. When $n > s$, a machine might not be able to go into repair directly at breakdown if all servers are repairing previously broken machines. Thus, machines may *interfere* with each other.

Summaries of the history of MIP studies can be found in the excellent reviews of Stecke and Aronson (1985), Cho and Parlar (1991), and Haque and Armstrong (2007). As stated in the previous section, additional to the n machines and s_n repairmen, our model has m_n spares and Markovian assumptions for time to failure and time until repair completion. We will now review the studies that have similar structures, approach, and focus in chronological order.

Barlow (1962) studies the same model as ours and obtain the probability distribution of broken machines (given in (1.1)), the distribution of time until all machines

are broken (which they call total failure), and the expected number of total failures until a time t. Then, they obtain these quantities for various modified versions of the main model such as with no spares, generalized breakdown and/or repair distributions, etc. In our study, we focus on the two types of delay described above and obtain computable approximations for means and probability distributions using asymptotic analysis.

Iglehart (1965) study our model; the only difference in their model is they assume arbitrary queue discipline for repair initiation of the machines waiting in the queue at the repair facility whereas we assume FCFS discipline. Here we also employ an additional perspective for delay, the delay until replacement.

They scale the broken machines process by the size of the demand base n, and they show that under a nonrestrictive assumption about the repair and backup capacity choices the scaled broken machines process weakly converges to Ornstein-Uhlenbeck (O-U) process. Then, they give the density of the virtual waiting time, which is the time required to repair all the broken machines in the system at a time t.

Our focus is different in two ways. We do not only study the delay experienced by machines, we also derive the replacement delay. Moreover, we present delay distributions; thus, rather than the congestion of the whole system at a specified time we give information on the waiting experience of a random machine/customer at any time.

Iglehart and Lemoine (1974) again considers the same model as ours and use the same assumption as they did in Iglehart (1965) for backup and repair capacity determination. First, by using classical limit laws on the distribution of broken machines, they give the equilibrium values for this process for some representative cases regarding system parameters. This corresponds to our fluid limit theorem (in Appendix), and we obtain this result for all cases and by using weak convergence instead of classical limit laws. However, they do not prove any limit theorems for

delays; instead they focus on the number of operable machines and obtain equilibrium values for this process. Although the structure of their model would allow them to obtain insights about limiting expected delay, obtaining the delay distribution would require a more detailed model. Their study continues by applying the same analysis on the machine-repairman problem with spares with two types of repairs.

The first studies on MIP considered derivation of some performance measures but did not make further use of these measures. Then, the consequent research made cost analysis using performance metrics such as expected number of broken machines, busy repairmen, etc. Rather than asymptotic analysis, these studies use steady-state analysis directly to construct the optimization problem. The steady-state cost functions are usually not analytically tractable and various search procedures and algorithms have been proposed to obtain the solution. In our cost analysis, using asymptotic methods we obtain an analytically tractable approximate solution (with error size in smaller scale than the square root of the system size) and generate more insights. Unless otherwise mentioned, the following studies focus on M/M/R machine repair problem with spares as cold standbys (spares has a breakdown rate of zero as they wait idle).

Hillard (1976) uses steady-state cost analysis as described above and they develop lemmas to reduce the feasible region their algorithm is searching. Gross et al. (1977) consider a multi-year planning horizon where they allow breakdown and service rates as well as capacity levels to change from year to year. Their cost function includes the expected number of components repaired and they impose a threshold level on the fill-rate (availability of spares). They develop a heuristic algorithm to provide a solution.

Sivazlian and Wang (1989) do steady state cost analysis on M/M/R machine repair problem with warm standbys (spares have a nonzero breakdown rate which is smaller compared to the breakdown rate of operating machines), they consider

machine availability (total fraction of time machines are running) and repairmen utilization among their performance measures. Wang and Sivazlian (1992) repeat the same analysis on a slightly different problem, they assume either cold or hot standbys (spares have the same breakdown rate as the operating machines) and when all repairmen are busy, repairmen switch to a faster repair rate. Wang (1993, 1994, 1995) then extends the analysis by adding the cold and warm mixed spares but with only one repairmen service rate, cold standbys with two failure modes with equal probability of repair, and two types of spares and two repair rates, respectively.

Jain (1997) develops a diffusion approximation to the broken machines process using reflecting boundaries for (m, M) machine repair problem with spares where the system shuts down if there are more than m broken machines. Then they obtain expected number of broken and operating machines, probability of having any spare units waiting as standbys, system availability, and idle repairmen. Their performance measures are focused on system response rather than individual delay experiences and they do not provide any solutions.

Wang and Lee (1998) continue the steady-state cost analysis approach on M/M/R machine-repair problem by assuming cold standbys and multiple failure modes and machines served by one or more repairmen. Jain et al. (2003) study the M/M/R machine repair problem with balking, reneging, spares and two modes of failure, and Singh and Jain (2007) continue on (m, M) machine repair problem with spares and reneging with two repair rates by using transient analysis. They solve sets of linear equations in terms of Laplace transform of the state probabilities. They obtain system availability and mean time to system failure.

Ke and Wang (2007) study the machine repair problem by adding two types of spares, two repair rates, and two vacation policies for repairmen. When a repairmen gets idle, he goes to a vacation, when he comes back he either stays until a machine is assigned to him or continues going vacations until he is assigned to a machine when

he returns. They use matrix-geometric approaches to find the steady state solution, then they derive mean values of several performance metrics, including machine availability and repairmen utilization to construct the expected cost rate. They give numerical examples of solution by direct search algorithm to their cost minimization problem subject to system availability constraint for either of the vacation policies.

Jain et al. (2008) study machine repair problem with warm and cold standbys, permanent and additional repairmen, two failure modes, and balking and reneging. Using matrix-geometric approach they obtain mean values of performance measures and machine availability and expected cost rate to use in cost minimization.

Ke et al. (2009) study the vacation model of Ke and Wang, they assume either cold, warm, or hot standbys, and single, multiple, or hybrid vacations. In hybrid vacations, the repairman waits for a random time and then goes to another vacation after returning from his last one. They produce the expected cost rate through steady state analysis.

As described in the beginning, we use a multi-dimensional many server model approach to address the problem where the offered load to the servers for repair remains roughly equal as we dimension the system. To our knowledge, this approach to the machine repair problem with spares has first been taken by Mandelbaum and Pats (1995, 1998) where they prove the diffusion limit theorem for the number of broken machines in the system but do not study the delay distributions; and then de Véricourt and Jennings (2008) use this approach on machine repair problem without spares to apply on large membership services.

1.3 The Model and Asymptotic Analysis via Many Server Limit Theorems

The system has a total of $n + m_n$ units which are either operable or broken, and the system is required to provide n operable units at all times. However, this may

not always be possible since the units are subject to spontaneous breakdowns. They reside in one of three locations: the demand base, the repair center, and the spare units center.

The demand base is responsible for providing n operable units to work continuously. The repair center has a certain number of identical servers or repairmen, which represents its capacity. To our knowledge of the existing literature on limit theorems, the choice of the capacity level of the repair center can be expressed as a function of the amount of units it will strive to keep working in the demand base; hence, we will denote the repair center capacity by s_n. The last center in the system is the spare units center where undamaged units are kept ready as cold standbys (no breakdowns are assumed to happen to spares during idling time) to replace the units that break down in the demand base, and these units are identical to the ones at the demand base. Based on the same reasoning for the number of servers in the repair center, we denote the choice for the maximum number of units to be held as spares by m_n.

The geographical placement of the three centers in the system is assumed to be such that the transportation times for a unit to circulate between them is negligible and can be omitted in the calculations governing delays.

The mission of each center is as follows: the demand base is supposed to have n units working all the time, the repair center will repair the broken units, and the backup center is supposed to hold up to m_n operable units ready to be sent to the demand base as replacements when breakdowns occur. Thus, if there are more than n operable units in the system, the additional ones will be waiting in the spare center. Thus, a total of $n + m_n$ units will be in the system.

Suppose a unit is working in the demand base and then it breaks down. After the breakdown the system faces two challenges: 1) there is a space open in the demand base and an undamaged unit is required to start working instead of the

broken one as soon as possible, 2) the broken unit has to be going into repair as soon as possible as well. These challenges also correspond to the two types of delay occurring in the system, determining its performance. The time it takes for a spare unit to be provided to the demand base is the first type of delay and the time for repair initiation of a broken unit is the second type of delay. Note that the first type of delay is a performance metric for customer satisfaction, and the second type represents the speed of circulation in the network. Assignment of a repairmen to the broken unit and dispatching of a spare unit to the demand base are done on first-come-first-serve (FCFS) basis. If either or both of these actions cannot be done immediately, two different queues develop; the queue of broken units awaiting repair and the queue of open spaces in the demand base awaiting replacement units.

After a unit breaks down, it goes to the repair center and either starts to be repaired or waits in the queue and then gets repaired. At the end of the repair, the unit is restored to an operational state. If there is an open space in the demand base waiting for a unit to be working, the repaired unit is sent there; otherwise, it is sent to the backup center until its services are needed at the demand base.

Consistent with the earlier claims, it can be seen that the model is closed; the units circulate within the system. Moreover, no additional units are accepted to the system once it starts (it is possible to extract a worn out unit from the system and replace it with a brand new one, as long as the unit is not working at that time and both actions are done simultaneously so that the total number of units in the system will not change).

The source of randomness in the system is in two places; one is the time to failure of a unit once it starts working in the demand base. The second is the time to repair once a repairmen is assigned to a broken unit. We assume the former is exponentially distributed with mean $1/\lambda$, and the latter is also exponentially distributed with mean $1/\mu$. All breakdowns and repairs are assumed to be mutually independent.

We will use the number of broken machines as the key system characteristic for our analysis and denote this stochastic process by $N_n \equiv \{N_n(t), t \geq 0\}$. This process is assumed to be right-continuous with left limits, and the steady state of this process is denoted by N_n^{st}.

The system framework and the aforementioned assumptions implies that N_n^{st} is a birth-and-death process with births corresponding to breakdowns and deaths corresponding to repairs, respectively. Suppose that k represents the number of broken units in the system at the considered moment ($k = 0, 1, 2, \ldots, n + m_n$), x^+ equals x if x is positive and zero otherwise, and $\wedge$ is the minimum operator. Then, since the units break down exponentially and independently from each other, the breakdown rate $\lambda_n(k)$ is λ times the number of units currently operating, which is n or a smaller value if there are more broken units than the spares, thus, $\lambda_n(k) = \lambda(n - [k - m_n]^+)$. When there are more broken down units than the maximum supply of spares ($k > m_n$), it follows that there is a shortfall of $k - m_n$ units at the demand base. The total breakdown rate is proportional to the number of units at the demand base. Similarly, the repair rate $\mu_n(k)$ is μ times the number of repair resources currently utilized, a quantity restricted by the number of broken machines and the number of repairmen; $\mu_n(k) = \mu(k \wedge s_n)$.

Thus, for given n, s_n, and m_n, the probability distribution of the key system characteristic for our analysis, the number of broken machines, is given as follows (suppose that $\rho \equiv \frac{\lambda}{\mu}$; p_0 is the appropriate normalizing constant; and when $s_n = m_n$,

$s_n \wedge m_n = s_n \vee m_n = s_n = m_n$):

$$P(N_n^{st} = k) = p_{n,k}^{st} = \begin{cases} p_0 \frac{n^k}{k!} \rho^k, & k < s_n,\ k < m_n \\ p_0 \frac{n^k}{s_n!} s_n^{s_n - k} \rho^k, & s_n \leq k < m_n \\ p_0 \frac{n^{m_n}}{k!} \frac{n!}{(n-(k-m_n))!} \rho^k, & m_n \leq k < s_n \\ p_0 \frac{n^{m_n}}{s_n!} \frac{n!}{(n-(k-m_n))!} s_n^{s_n - k} \rho^k, & s_n \leq k,\ m_n \leq k. \end{cases} \tag{1.1}$$

Our results are based on the choice of the number of servers and the maximum number of spare units to be kept which correspond to repair and backup capacity, respectively. As stated before, they can both be expressed as functions of n, and therefore we will denote our system design with the pair (s_n, m_n).

The model can be envisioned for systems where a certain number of customers require an equipment to work continuously. For ex, the management of carrier fleets as Fedex and UPS or of a city transit bus fleet will be asked to assign vehicles on a certain number of routes, and in this case the routes will be the customers (denoted by n). On the other hand, maintenance and backup support for warranty services contracts of n sales of cell phones of a particular model or for a job shop where n machine tools of the same type has to be working continuously can also be represented by this model.

As can be seen from (1.1), any direct analysis using the process of broken machines and other derivatives of this process may provide complex expressions hard to solve analytically. To derive analytical solutions, we will consider a sequence of Markovian models as described above, indexed by n ($n \geq 1$), the required number of operating units at the demand base. Besides n, we need to determine the parameters ρ and (s_n, m_n) for each model in the sequence. We assume the same λ and μ (and hence the same ρ) is valid for all models and for each one, (s_n, m_n) is expressed as a function of its respective n value; thus, each sequence will have different capacity choice pairs.

Our approach is to express all the delay performance measures of interest in terms of the steady state distribution of the broken machines process (1.1), and then take the limits of these measures as $n \to \infty$. This way we will obtain asymptotical and analytically tractable approximations to the complex expressions of steady state delay performance measures. The ones we are mainly concerned about are the mean delays until spare unit provision and broken unit repair initiation, and the probability that either type of delay exceeds an arbitrary threshold (the probability of experiencing an unacceptable delay). The following sections demonstrate the relationship between the repair and spares capacity and the explained performance metrics.

1.4 Analysis of the Relationship Between Delay Until Replacement of a Broken Unit and Capacity Decisions

Our aim is to provide insights about how to design a system with acceptable delays, achieved via rationalized capacity level, a process that will aid managerial decisions. We have identified two types of delays that could be of concern regarding company performance. One is the time to provide replacement units to the customers, the other is the time to begin the repair of a damaged unit. The level of backup and repair capacity would be the decisions affecting these two performance measures. Thus, the problem we would like to solve is how to make the best backup and repair capacity level investment decision to give the desired level of delay. In this section, we concentrate mainly on the performance measures related with delay until replacement of a broken unit.

1.4.1 Mean delay until replacement based design

Expressing delay as a function of design parameters

Here we will establish the limiting values of expected delay as n grows large. As discussed in the model section, it can be assumed that the repair and backup capacity

investments can be described in terms of the size of the demand base; and hence, they could be denoted as s_n and m_n, respectively. We assume our family of parameters have the following form:

$$\begin{aligned} s_n &= \bar{s}n + \hat{s}_n\sqrt{n}, \\ m_n &= \bar{m}n + \hat{m}_n\sqrt{n} \end{aligned} \tag{1.2}$$

where $\bar{s}, \hat{s}_n, \bar{m}$, and $\hat{m}_n$ are constants denoting the capacity level investment choices and there exists constants $\hat{s}$ and $\hat{m}$ such that $\hat{s}_n \to \hat{s}$ and $\hat{m}_n \to \hat{m}$ as $n \to \infty$. In other words, $s_n = \bar{s}n + \hat{s}\sqrt{n} + o(\sqrt{n})$ and m_n can alternatively be defined in a similar way.

Observe that $\frac{m_n}{n} \to \bar{m}$ and $\frac{s_n}{n} \to \bar{s}$ as $n \to \infty$. Now consider a possible delay for backup. The machine breaks down, no spares are available, and a customer starts waiting for a functioning machine; however, there are other customers who have started waiting before. When the number of broken machines exceeds the number of spares available (m_n), the customers start forming a queue (managed on a FCFS principle); therefore the customer (including himself) will have to wait for (the number of broken machines at breakdown + 1 - m_n) repairs to be completed. Notice that the repair rate can be arbitrary depending on whether all repairmen are busy during the delay. This may cause the mean delay expressions to be complex; however, we show that asymptotic analysis can help to obtain simple expressions as approximations that provide analytical solutions.

The taxonomy we will employ to categorize the firm's approach in capacity level selection is adapted from the call center literature, where the call center staff level is taken as a function of the call load. The many-server limit theorems applications to the staffing problem in call centers classify the firm preferences into three categories.

The first category is described as efficiency-driven (ED) capacity level determination where high utilization is obtained. For ex, when the size of the demand base

go to infinity, the resources chosen with this preference will be 100% utilized. The second one is quality-driven (QD) determination to provide exceptional service, and contrasting to the previous one, as n goes to infinity the resources will never be 100% utilized. The final one is quality- and efficiency-driven (QED) and inherits the properties of both ED and QD.

The last category was the focus of Halfin and Whitt (1981), while the whole taxonomy was first utilized by Garnett et al. (2002) and has been widely adopted in the literature. Borst et al. (2004) uses this taxonomy to solve an approximate cost minimization problem for staffing in call centers. Our work expands the usage of this taxonomy to the determination of a *pair* of capacity levels: the number of repairmen and maximum number of spare machines. *Our representation will express repairmen capacity level preference first and spare machines level preference second.* Specifically, the repairmen capacity can be ED, QD, or QED. Likewise, the spare machines decision can be ED, QD, or QED as well.

When certain scaling approaches are applied to the birth-and-death process of broken machines, a limiting equilibrium can be obtained, leading to the two dimensional graph of $(\bar{s}, \bar{m})$ pairs (Figure 1.2) partitioned into regions according to the capacity level preference labels they are representing (from Theorem 12 in Appendix). Then, based on the firm's preferences, the limiting value of expected delay can also be expressed as a function of $\bar{s}$ and $\bar{m}$ as below:

Theorem 1. *Let $E[D_n^{rpl}(s_n, m_n)]$ denote the mean steady state delay until replacement. Under (1.2):*

$$\lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)] = \begin{cases} 1/(\bar{s}\mu) - 1/\lambda \quad ED|ED \quad \& \quad QED|ED \\ 1/\mu - (\bar{m}(1+\rho))\,/\,(\mu\rho(1+\bar{m})) \quad QD|ED \\ 0 \quad otherwise. \end{cases} \tag{1.3}$$

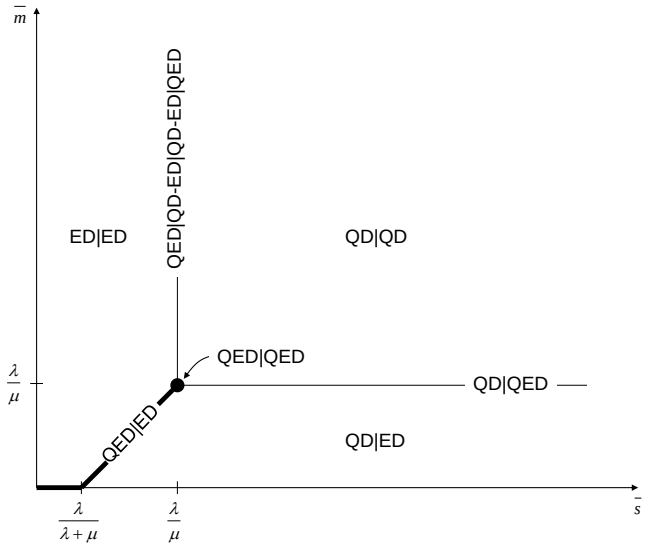

FIGURE 1.2: Pair of capacity level choices representing $(\bar{s}, \bar{m})$ values, cost-effective region in bold

Remark 1. *As can be seen from Theorem 12 in Section A.1.1, the equation* $\bar{s} = 1 + \bar{m} - \frac{\bar{s}}{\rho}$ *holds when capacity level is QED|ED. Hence,* $\lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)]_{QED|ED}$ *can be expressed in two ways: 1)*$1/(\bar{s}\mu) - 1/\lambda$ *(as above), and 2)* $1/\mu - (\bar{m}(1+\rho))$ $/\left(\mu\rho(1+\bar{m})\right)$.

Approximations for Optimal Capacity Decisions Giving the Desired Mean Delay Until Replacement Performance

Let w be the cost rate per server staffed in the repair center, c be the cost rate for each spare unit, and p be the penalty cost rate assessed for each customer waiting for a replacement. Note that $E[D_n^{rpl}(s_n, m_n)]$ is the delay until replacement when a breakdown occurs; hence, $p \cdot E[D_n^{rpl}(s_n, m_n)]$ corresponds to the penalty cost incurred at breakdowns. Then, for a capacity choice of (s_n, m_n), we define the system design cost, $C(s_n, m_n)$, as:

$$C(s_n, m_n) = ws_n + cm_n + npE[D_n^{rpl}(s_n, m_n)]. \tag{1.4}$$

Remark 2. *The delay penalty is scaled-up by the size of the demand base to account for the total inconvenience experienced by the demand. One can envision another formulation that can yield the relative frequency of delay.*

For all n, it is intuitive that $E[D_n^{rpl}(s_n, m_n)]$ is decreasing in both s_n and m_n. Hence, the tradeoff in (1.4) is between the sum of the first two terms and the last term. Thus, from (1.2), we will minimize the following limiting cost function:

$$Min \quad \bar{C}(\bar{s}, \bar{m}) = w\bar{s} + c\bar{m} + p \lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)]. \tag{1.5}$$

Some of the terms in the description of the following solution have strong implications (and they also appear in Figure 1.2). For a machine repairmen problem with no spares and total number of repairmen equal to the number of machines, there will be no interference and hence no positive delay until repair initiation. Since machines will alternate between working and under repair states, with Markovian assumptions, $\lambda/\lambda+\mu$ will be the fraction of time a machine will stay broken. Hence, $\lambda/\lambda+\mu$ will be the fraction of broken machines when the delay until repair initiation is roughly zero. Thus, we may call $\lambda/\lambda+\mu$ as the offered load for a system with no replacements.

Another type of system with ample resources could be one where the breakdown rate is very small compared to the repair rate. In this case, a large system can be envisioned as an open system such as an infinite server queue, and λ/μ will be the expected number of busy servers and thus the offered load.

Theorem 2. *Under the assumption of Theorem 1, the solutions to the minimization problem are summarized below:*

i. If $\sqrt{p/(\mu w)} < \lambda/(\lambda+\mu)$*, then ED|ED is chosen with*

$$\bar{s} = \sqrt{p/(\mu w)}, \bar{m} = 0,$$

ii. if $\lambda/(\lambda+\mu) \leq \sqrt{p/\left[w\mu + (c\mu(\lambda+\mu))/\lambda\right]} \leq \lambda/\mu$, *then QED|ED is chosen with*

$$\bar{s} = \sqrt{p/\left[w\mu + \frac{c\mu(\lambda+\mu)}{\lambda}\right]}, \bar{m} = \sqrt{\left[p(\lambda+\mu)^2\right]/\left[w\mu\lambda^2 + c\mu\lambda(\lambda+\mu)\right]} - 1,$$

iii. if $\lambda/\mu < \sqrt{p/\left[w\mu + (c\mu(\lambda+\mu))/\lambda\right]} < \sqrt{p/(\mu w)}$, *then QED|QED is chosen with*

$$\bar{s} = \bar{m} = \frac{\lambda}{\mu},$$

iv. finally, if $\sqrt{p/\left[w\mu + (c\mu(\lambda+\mu))/\lambda\right]} < \lambda/(\lambda+\mu) < \sqrt{p/(\mu w)}$, *then QED|ED is chosen with*

$$\bar{s} = \lambda/(\lambda+\mu), \bar{m} = 0.$$

No other inequalities are possible.

This theorem shows that to choose the quality level of service and hence the backup and repair capacity levels, only three capacity level preferences can be recommended, ED|ED, QED|ED, and QED|QED. These three preferences constitute the cost efficient region, shown in Figure 1.2 (drawn by Theorem 12 in Appendix).

1.4.2 Distribution of delay until replacement based design

In many real circumstances, the managers will have threshold values for the delays, and they would desire their capacity choices to limit the probability that delays exceed their corresponding thresholds (the probability of unacceptable delay). In the following theorems we will express the limiting delay distribution in terms of capacity choices, and this will guide us to complement the capacity level choices recommended before.

Theorem 3. *Let* $T\epsilon[0,\infty]$. *For* $T > 0$, *a partial characterization of the limiting probability of unacceptable delay until replacement,* $\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > T)$, *is as follows (given (1.2)):*

$$\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > T) = \begin{cases} 1 & for \quad ED|ED, 1-\mu\bar{s}T-\frac{\bar{s}}{\rho} > 0 \\ \Phi\left(-\hat{s}/\bar{s}\right) & for \quad ED|ED, 1-\mu\bar{s}T-\frac{\bar{s}}{\rho} = 0 \\ 0 & for \quad ED|ED, 1-\mu\bar{s}T-\frac{\bar{s}}{\rho} < 0 \\ 0 & for \quad QED|QD-ED|QD-ED|QED \\ 0 & for \quad QED|QED, QD|QED, QD|QD \end{cases} . \tag{1.6}$$

For $T = 0$, the limiting probability of unacceptable delay until replacement, $\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > T)$, is as follows:

$$\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > 0)$$

$$= \begin{cases} 1 & ED|ED, \\ & QED|ED, \\ & QD|ED \\ \\ 0 & QD|QD \\ \\ \{1 & QED|QED, \\ & \hat{m} > \hat{s}, \hat{s} \neq 0 \\ \\ & + \left[\rho \left(1 - \exp\left\{-\left(\hat{s}(\hat{m} - \hat{s})\right)/\rho\right\}\right) \phi\left(\hat{s}/\rho\right)\right] / \left[\hat{s} \ \exp\left\{-\left(\hat{s}(\hat{m} - \hat{s})\right)/\rho\right\} \Phi\left(-\hat{s}/\rho\right)\right] \\ \\ & + \left[\sqrt{\rho}\, \phi\left(\hat{s}/\rho\right) \ \exp\left\{\left(\hat{s}(\hat{m} - \hat{s})\right)/\rho\right\} \ \Phi\left(\hat{s}/\sqrt{\rho}\right)\right] / \left[\phi\left(-\hat{s}/\sqrt{\rho}\right) \ \Phi\left(-\hat{s}/\rho\right)\right]\Big\}^{-1} \\ \\ & \left(1 + \sqrt{2}\, \hat{m}/\sqrt{\pi} + \sqrt{\rho}\right)^{-1} \quad QED|QED, \\ & \qquad\qquad\qquad\qquad\quad \hat{m} > \hat{s}, \hat{s} = 0 \\ \\ & \left(1 + \Phi\left(\hat{m}/\sqrt{\rho}\right) \qquad\qquad QED|QED \right. \\ & \qquad\qquad\qquad\quad \hat{m} < \hat{s} \\ \\ & \left. \cdot \left[\left(\sqrt{1+\rho}\right)^{-1} \ \phi\left(-\hat{m}/\sqrt{\rho}\right) A + \left(\sqrt{\rho}\right)^{-1} \ \phi\left(-\hat{m}/\sqrt{\rho}\right) \ B\right]\right)^{-1} \\ \\ & \left\{1 + \left[\sqrt{\rho}\, \phi\left(\hat{s}/\rho\right) \ \Phi\left(\hat{s}/\sqrt{\rho}\right)\right] / \left[\phi\left(-\hat{s}/\sqrt{\rho}\right) \ \Phi\left(-\hat{s}/\rho\right)\right]\right\}^{-1} \quad \begin{array}{l} QED|QED \\ \hat{m} = \hat{s} \end{array} \\ \\ & \left\{1 + \left[\Phi\left(\hat{m}/\sqrt{\rho}\right) \phi\left(\hat{m}/\sqrt{\rho(1+\rho)}\right)\right] / \qquad QD|QED \right. \\ \\ & \left. \left[\left(\sqrt{(1+\rho)}\right)^{-1} \phi\left(-\hat{m}/\sqrt{\rho}\right) \Phi\left(-\hat{m}/\sqrt{(1+\rho)\rho}\right)\right]\right\}^{-1} \end{cases} \tag{1.7}$$

where

$$A = \left[\Phi\left(\hat{s}\sqrt{1+1/\rho} - \frac{\hat{m}}{\sqrt{1+1/\rho}}\right) - \Phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\right] / \left[\phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\right]$$

$$B = \left[\phi\left(\hat{s}\sqrt{1+1/\rho} - \frac{\hat{m}}{\sqrt{1+1/\rho}}\right)\ \Phi\left(\hat{m} - \left(1+\tfrac{1}{\rho}\right)\ \hat{s}\right)\right] \tag{1.8}$$

$$/\left[\phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\ \phi\left(\left(1+\tfrac{1}{\rho}\right)\hat{s} - \hat{m}\right)\right].$$

Remark 3. *An interesting result is that for ED|ED, the limiting probability of unacceptable delay until replacement only depends on the parameters of repair servers, s_n, defined in (1.2); there is not $\bar{m}$ or $\hat{m}$ included. This implies that for large and highly utilized systems, the repair capacity is the dominant resource in terms of delay until replacement performance.*

Note that for QED|ED and QD|ED, the limiting number of busy repairmen could be smaller than the number of servers, s_n, in order $\sqrt{n}$ sensitivity (see Theorem 12 in Appendix, the fluid limit b is smaller than or equal to $\bar{s}$). In this case, obtaining the limiting probability of unacceptable delay until replacement ($T > 0$) is beyond the scope of this model since the distribution of a random delay until replacement will be hypoexponential and the rates will be determined by the trajectory of the number of busy repairmen during the delay.

For ED|ED above, the equation $1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} = 0$ defines a one-to-one match between T and $\bar{s}$ values. When $\bar{s}$ increases, the corresponding T will be smaller. With smaller threshold, one would expect a higher probability of unacceptable delay ($T > 0$), which is the case with increasing $\bar{s}$ value.

It is not possible to characterize the limiting value for the probability of delay until replacement ($T = 0$) at the region representing the capacity level choices QED|QD-ED|QD-ED|QED at Figure 1.2 ($\bar{s} = \rho < \bar{m}$). When $\hat{s} > 0$, the limit is zero;

however, when $\hat{s} < 0$, it is one. No conclusions can be drawn for $\hat{s} = 0$. One possible explanation for this is as in Section A.3.2 in Appendix, whether there will be a limiting value is determined by the comparison of the fluid limit b with $\bar{m}$, and from Theorem 12 and Table A.4 the limiting initial value of the fluid-scaled process in (A.1), $\bar{\mathbf{b}}(0)$, is equal to the fluid limit b itself; however, QED|ED, ED|QD, and ED|QED all lead to three different (in)equalities between b and $\bar{m}$, leading to three different limiting behaviors of the probability of delay until replacement.

Suppose that the order n parameters in (1.2) are chosen with respect to Theorem 1.2 or via other methods; then, the formulas above for the probability of unacceptable delay until replacement can be used to determine the $(\hat{s}, \hat{m})$ pair giving the desired probability value with minimum cost. This way all the parameters in (1.2) will be determined and an actual capacity level pair (s_n, m_n) can be recommended. It is intuitive that the probability of delay exceeding a threshold is decreasing in the amount of resources used to satisfy the needs of who or what is experiencing the delay; then, if an upper bound for the probability of unacceptable delay has been determined (e.g. in a maintenance contract), Theorem 3 formulas can be used to obtain curves of $(\hat{s}, \hat{m})$ pairs that hit the probability bound to choose the minimum cost pair within (note that using higher values than the curve pairs will only cost more while still satisfying the constraint).

Below we give an example of this for QED|QED, λ is taken as 0.3 and μ is taken as 0.5. We depict nine $(\hat{s}, \hat{m})$ pair curves representing the feasible $(\hat{s}, \hat{m})$ pairs hitting the upper bound (0.1, 0.2, ..., 0.9) for the probability of delay until replacement.

1.5 Analysis of the Relationship Between Delay Until Repair Initiation of a Broken Unit and Capacity Decisions

We will do an analysis for the delay until repair initiation similar to what was done in the previous section. First, we will give the limiting mean delay until repair initiation

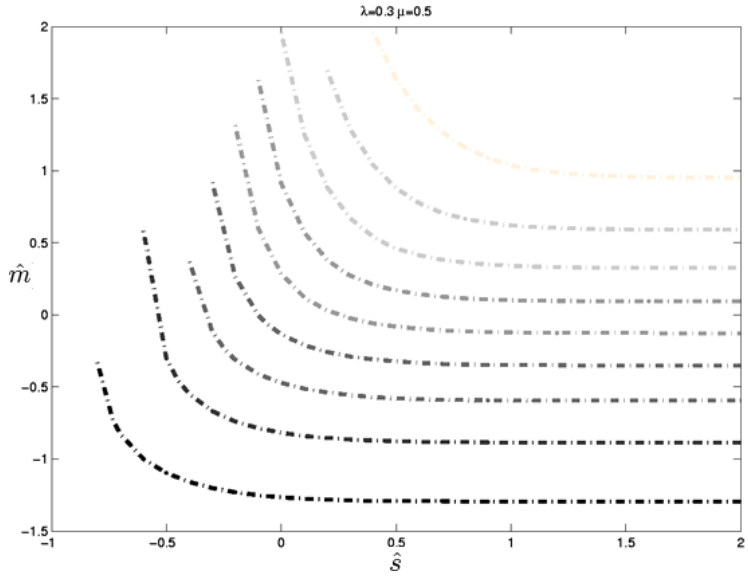

FIGURE 1.3: Curves of capacity level choice QED|QED's order $\sqrt{n}$ parameters, $(\hat{s}, \hat{m})$, hitting the upper bounds 0.1,...,0.9 (from light to dark curves) for delay until replacement

and use this value to minimize an approximate cost function. Then, we will give the limiting probability of unacceptable delay and give another approximation for ED|ED capacity level choices giving the minimum cost rate and at the same time satisfying an asymptotic constraint on the probability of unacceptable delay. We will denote the delay until repair initiation when s_n repair servers and m_n spare units are used as $D_n^{rpi}(s_n, m_n)$.

Theorem 4. *Let $E[D_n^{rpi}(s_n, m_n)]$ denote the mean delay until repair initiation. Under (1.2):*

$$\lim_{n\to\infty} E[D_n^{rpi}(s_n, m_n)] = \begin{cases} \frac{1+\bar{m}}{\bar{s}\mu} - \frac{1}{\lambda} - \frac{1}{\mu} & ED|ED \\ 0 & otherwise. \end{cases} \tag{1.9}$$

where the capacity level preferences are being depicted as in Figure 1.2.

We will use the theorem above to minimize a limiting cost rate similar to (1.5),

which is shown below (h is the penalty cost rate for each broken machine waiting for repair).

$$Min \quad \bar{C}(\bar{s}, \bar{m}) = w\bar{s} + c\bar{m} + h \lim_{n\to\infty} E[D_n^{rpi}(s_n, m_n)]. \tag{1.10}$$

Theorem 5. *Under Theorem 4, the solutions to the minimization problem above are summarized below:*

i. if $\lambda/(\lambda+\mu) \leq c/(w+c+c/\rho) \leq \lambda/\mu$, *then QED|ED (or if* $c/(w+c+c/\rho) = \rho$, *QED|QED) is chosen with*

$$\bar{s} = c/(w+c+c/\rho), \bar{m} = [c(1+1/\rho)]/(w+c+c/\rho) - 1,$$

ii. if $c/(w+c+c/\rho) < \lambda/(\lambda+\mu)$, *and* $\sqrt{h/w\mu} < \lambda/(\lambda+\mu)$, *then ED|ED is chosen with*

$$\bar{s} = \sqrt{h/w\mu}, \bar{m} = 0,$$

iii. if $c/(w+c+c/\rho) < \lambda/(\lambda+\mu)$, *and* $\sqrt{h/w\mu} \geq \lambda/(\lambda+\mu)$, *then ED|ED is chosen with*

$$\bar{s} = \lambda/(\lambda+\mu), \bar{m} = 0,$$

No other inequalities are possible.

Theorem 6. *Let* $T\epsilon[0,\infty]$ *and* $\varphi = 1+1/\rho+\mu T$. *For* $T > 0$, *the limiting probability of unacceptable delay until repair initiation,* $\lim_{n\to\infty} P(D_n^{rpi}(s_n, m_n) > T)$, *is as follows (given (1.2)):*

$$\lim_{n\to\infty} P(D_n^{rpi}(s_n, m_n) > T) =$$

$$\begin{cases} 1 & ED|ED, \\ & \bar{m} - \bar{s} - \frac{\bar{s}}{\rho} + 1 - \mu\bar{s}T > 0 \\ \Phi\left(\left(\hat{m} - \varphi\hat{s}\right)/\sqrt{\bar{s}\left(1/\rho + \mu T\right)}\right) & ED|ED, \\ & \bar{m} - \bar{s} - \frac{\bar{s}}{\rho} + 1 - \mu\bar{s}T = 0 \\ 0 & otherwise. \end{cases} \tag{1.11}$$

For $T = 0$, the limiting probability of unacceptable delay until repair initiation, $\lim_{n\to\infty} P(D_n^{rpi}(s_n, m_n) > T)$, is as follows:

$$\lim_{n\to\infty} P(D_n^{rpi}(s_n, m_n) > 0)$$

$$= \begin{cases} 1 \;\; ED|ED \\ 1 \;\; QED|QD-ED|QD-ED|QED, \hat{s} \le 0 \\ 1 - \left(\frac{\hat{s}}{\hat{s}+\sqrt{\rho}\left[\phi(-\hat{s}/\sqrt{\rho})/\Phi(\hat{s}/\sqrt{\rho})\right]} \right) \;\; QED|QD-ED|QD-ED|QED, \hat{s} > 0 \\ 1 - \left\{ 1 + \left[\sqrt{\rho} \left(1 - \exp\left\{ -\frac{\hat{s}(\hat{m}-\hat{s})}{\rho} \right\} \right) \phi\left(-\frac{\hat{s}}{\sqrt{\rho}} \right) \right] / \left[\hat{s}\, \Phi\left(\frac{\hat{s}}{\sqrt{\rho}} \right) \right] \right. \\ \left. + \left[\phi\left(-\frac{\hat{s}}{\sqrt{\rho}} \right) \exp\left\{ -\frac{\hat{s}(\hat{m}-\hat{s})}{\rho} \right\} \Phi\left(-\frac{\hat{s}}{\rho} \right) \right] \right. \\ \left. / \left[\sqrt{\rho}\, \phi\left(\frac{\hat{s}}{\rho} \right) \Phi\left(\frac{\hat{s}}{\sqrt{\rho}} \right) \right] \right\}^{-1} \;\; QED|QED, \\ \qquad\qquad\qquad\qquad\qquad\qquad \hat{m} \ge \hat{s}, \hat{s} \ne 0 \\ 1 - \left(1 + \sqrt{\frac{2}{\pi}} \;\; \frac{\hat{m}}{\sqrt{\rho}} + \frac{1}{\sqrt{\rho}} \right)^{-1} \qquad\qquad QED|QED, \hat{m} \ge \hat{s}, \hat{s} = 0 \\ \left(1 + \frac{\sqrt{\rho}\, \Phi\left(\hat{m}/\sqrt{\rho}\right)}{\Phi(\hat{m}-\hat{s}-\hat{s}/\rho)} \quad \frac{\phi((1+1/\rho)\hat{s}-\hat{m})\phi\left(\hat{m}/\sqrt{\rho(1+\rho)}\right)}{\phi\left(\hat{s}\sqrt{1+1/\rho}-\hat{m}/\sqrt{1+1/\rho}\right)\phi\left(-\hat{m}/\sqrt{\rho}\right)} \right. \\ \left. + \sqrt{\frac{\rho}{1+\rho}} \quad \frac{\phi\left((1+1/\rho)((1+\rho)\hat{s}-\rho\hat{m})(1+\rho)^{-1}-\hat{m}\right)\phi\left(\hat{m}/\sqrt{\rho(1+\rho)}\right)}{\phi\left(((1+\rho)\hat{s}-\rho\hat{m})(1+\rho)^{-1}\sqrt{1+1/\rho}-\hat{m}/\sqrt{1+1/\rho}\right)\phi\left(-\hat{m}/\sqrt{\rho}\right)} \right. \\ \left. \cdot \frac{\Phi\left(\hat{s}\sqrt{1+1/\rho}-\hat{m}/\sqrt{1+1/\rho}\right)-\Phi\left(\hat{m}/\sqrt{\rho(1+\rho)}\right)}{\Phi(\hat{m}-\hat{s}-\hat{s}/\rho)} \right)^{-1} \;\; QED|QED, \hat{m} < \hat{s} \end{cases} \tag{1.12}$$

$$\lim_{n\to\infty} P(D_n^{rpi}(s_n, m_n) > 0)$$

$$= \begin{cases} \left(1+\sqrt{\frac{\bar{s}}{1+\bar{m}}} \quad \cdot \frac{\phi\left((1+1/\rho)((1+\rho)\hat{s}-\rho\hat{m}))(1+\bar{m})^{-1}-\hat{m}\right)}{\phi\left(((1+\rho)\hat{s}-\rho\hat{m})(1+\bar{m})^{-1}\sqrt{1+1/\rho}-\hat{m}/\sqrt{1+1/\rho}\right)\phi\left(-\hat{m}/\sqrt{\rho}\right)}\right. \\ \left. \cdot\frac{\phi\left(\hat{m}/\sqrt{\rho(1+\rho)}\right)\Phi\left((\hat{s}-\hat{m}\rho/(1+\rho))/\sqrt{(1+\bar{m})\rho/(1+\rho)^2}\right)}{\phi(\sqrt{\rho-\bar{m}}\ \rho^{-1}\ [((1+\rho)\hat{s}-\rho\hat{m})/(1+\bar{m})])\Phi\left((\hat{m}-\hat{s}-\hat{s}/\rho)/\sqrt{\bar{s}/\rho}\right)}\right)^{-1} \quad QED|ED, (\hat{m}=\hat{s}=0)' \\ \left(1+\sqrt{\frac{\bar{s}}{1+\bar{m}}}\right)^{-1} \quad QED|ED, \hat{m}=\hat{s}=0 \\ 0 \qquad otherwise. \end{cases} \tag{1.13}$$

Remark 4. *Note that in (1.11), the region ED|ED, $\bar{m}-\bar{s}-\frac{\bar{s}}{\rho}+1-\mu\bar{s}T=0$ determines a new boundary for ED|ED, we call it QED(T)|ED since for $T=0$ it coincides with the regular QED|ED and depict it in Figure 1.4 (for some $T>0$).*

Remark 5. *In (1.11), from $\bar{m}-\bar{s}-\frac{\bar{s}}{\rho}+1-\mu\bar{s}T=0$, $\Phi\left(\frac{\hat{m}-\varphi\hat{s}}{\sqrt{\bar{s}(1/\rho+\mu T)}}\right)$ can also be represented as $\Phi\left(\frac{\hat{m}-(1+\bar{m})\hat{s}/\bar{s}}{\sqrt{\bar{s}(1+\bar{m}-\bar{s})}}\right)$.*

Now for QED(T)|ED and from (1.2), consider an optimization problem with asymptotic constraints for sufficiently large n,

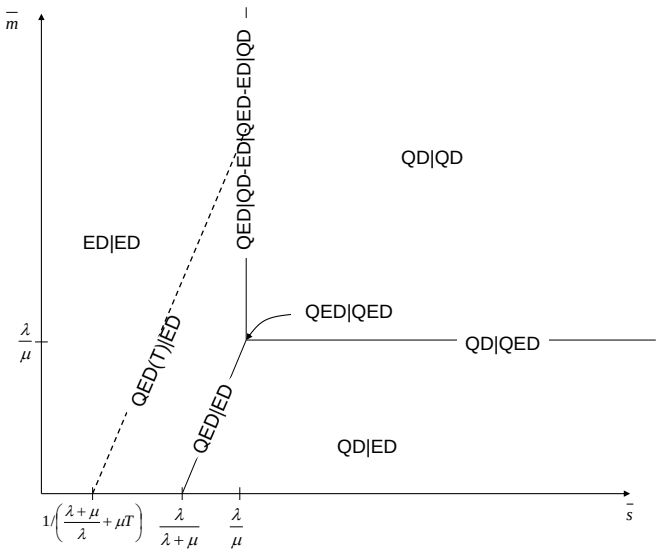

FIGURE 1.4: QED(T)|ED represented by ED|ED, $\bar{m} - \bar{s} - \frac{\bar{s}}{\rho} + 1 - \mu\bar{s}T = 0$, for $T > 0$

$$
\begin{aligned}
&Min \quad C(s_n, m_n) = n[w\bar{s} + c\bar{m}] + \sqrt{n}[w\hat{s} + c\hat{m}] \\
&s.t. \\
&\quad \Phi\left(\frac{\hat{m}-\varphi\hat{s}}{\sqrt{\bar{s}(1/\rho+\mu T)}}\right) \leq \alpha \\
&\quad \bar{m} - \bar{s} - \frac{\bar{s}}{\rho} + 1 - \mu\bar{s}T = 0 \\
&\quad n\bar{m} + \sqrt{n}\hat{m} \geq 0 \\
&\quad n\bar{s} + \sqrt{n}\hat{s} \geq 0 \\
&\quad \bar{s}, \bar{m}, \hat{s}, \hat{m} \, \epsilon \, \mathbb{R}
\end{aligned}
\tag{1.14}
$$

The objective is to minimize the system cost rate for the number of repair servers employed and spare units kept. It is assumed that a threshold value T has been determined exogenously for the acceptable delay. The first constraint represents that the probability of having an unacceptable delay (delay exceeding T) should be

lower than a fixed α value in the limit; the second constraint represents the equality describing the QED(T)|ED in Theorem 6; and the third and fourth constraints represent that number of repair servers and spares cannot be negative.

Since order n parameters are more dominant, one would first want to minimize the first term in the objective function. From the constraints it is desired for $\bar{m}, \bar{s} \geq 0$ and also $\bar{m} - \bar{s} - \frac{\bar{s}}{\rho} + 1 - \mu\bar{s}T = 0$. This gives $w\bar{s} + c\bar{m} = w\bar{s} + c\left(\bar{s} + \bar{s}/\rho + \mu\bar{s}T - 1\right)$. $\bar{s} = c/\left(w + c + c/\rho + \mu cT\right)$ makes this expression zero; however, the corresponding $\bar{m}$ value is $-w/\left(w + c + c/\rho + \mu cT\right) < 0$. Since $\bar{m}$ is increasing in $\bar{s}$ and vice versa (from the equality), we conclude that the optimal order n parameter values are $\bar{s} = 1/\varphi, \bar{m} = 0$.

Applying this in the first constraint gives $\Phi\left(\frac{\hat{m} - \varphi\hat{s}}{\sqrt{1-\varphi^{-1}}}\right) \leq \alpha$, notice that for every $(\hat{s}, \hat{m})$ pair satisfying this inequality, decreasing the $\hat{m}$ value decreases the second term in the cost function in (1.14) while still satisfying the inequality. From the third constraint and the order n optimal values determined above, the smallest feasible value for $\hat{m}$ is zero; and since $\bar{s}$ is positive and dominant, $\hat{s}$ can be negative. From the monotonicity of the standard normal distribution, this brings the solution

$$\hat{s} = -\left(\Phi(\alpha)^{-1}\sqrt{1-\varphi^{-1}}/\varphi\right), \hat{m} = 0.$$

The solution of the optimization problem is

$$\bar{s} = 1/\varphi, \bar{m} = 0, \hat{s} = -\left(\Phi(\alpha)^{-1}\sqrt{1-\varphi^{-1}}/\varphi\right), \hat{m} = 0.$$

Thus, it is best not to use any spares when QED(T)|ED capacity level choice is made. Note that this actually represents the new boundary for ED|ED; hence, in both delay until replacement and delay until repair initiation, when ED|ED is chosen or when very high utilization of both repair and spares capacity is preferred, in terms of delay performance, spares capacity is asymptotically not required.

$E[D_n^{rpl}(s_n, m_n)]$					
n	s_n	m_n	Simulation	Approximation	%Error
24	2	0	3.0058	3	-0.194
60	5	0	2.9975	3	0.0822
108	9	0	3.0203	3	-0.6757
$P(D_n^{rpl}(s_n, m_n) > \lim_{n\to\infty} D_n^{rpl}(s_n, m_n))$					
24	2	0	0.476	0.5	4.809
60	5	0	0.4809	0.5	3.8166
108	9	0	0.497	0.5	0.601

Table 1.1: Comparison of Heavy-traffic Approximations with Simulation Results for Delay until Replacement, ED|ED, $\lambda = 1, \mu = 3, \bar{s} = 1/12, \bar{m} = 0, \hat{s}_n = \hat{m}_n = 0$

$E[D_n^{rpl}(s_n, m_n)]$					
n	s_n	m_n	Simulation	Approximation	%Error
288	84	48	0.1713	0.1429	19.8806
480	140	80	0.1644	0.1429	15.0611
720	210	120	0.1640	0.1429	14.8252

Table 1.2: Comparison of Heavy-traffic Approximations with Simulation Results for Delay until Replacement, QED|ED, $\lambda = 1, \mu = 3, \bar{s} = 7/24, \bar{m} = 4/24, \hat{s} = \hat{m} = 0$

1.6 Numerical Results

In this section, we compare our approximations with the exact values (where obtainable) or simulation results. The capacity level choices we choose to observe are ED|ED, QED|ED, QED|QED since these are the cost-effective alternatives in Section 1.4. When obtaining the data for the probability of unacceptable delay, we used the limiting mean delays we have obtained in Theorems 1 and 4 as thresholds.

In Table 1.1, we see that for small to moderate systems in ED|ED, the fluid level approximations for capacity choices (notice that $\hat{s}_n = \hat{m}_n = 0$) have very small % errors. For these systems it can be more desirable to always use an order n level parameter $\bar{s}_1$ rather than a $(\bar{s}_{n,2}, \hat{s}_{n,2})$ pair so that both alternatives give the same s_n value, and this $\bar{s}_1$ value can be used instead to obtain the approximations.

$E[D_n^{rpi}(s_n, m_n)]$					
n	s_n	m_n	Exact Value	Approximation	%Error
24	8	8	0.1102	0	N/A
48	16	16	0.0765	0	N/A
96	32	32	0.0534	0	N/A
$P(D_n^{rpi}(s_n, m_n) > \lim_{n\to\infty} D_n^{rpi}(s_n, m_n))$					
24	8	8	0.6423	0.634	-1.3124
48	16	16	0.6401	0.634	-0.9558
96	32	32	0.6384	0.634	-0.6903

Table 1.3: Comparison of Heavy-traffic Approximations with Exact Values for Delay until Repair Initiation, QED|QED, $\lambda = 1, \mu = 3, \bar{s} = 1/3, \bar{m} = 1/3, \hat{s} = \hat{m} = 0$

In Table 1.2, we see that the approximations for delay until replacement for QED|ED gives higher % errors with respect to their ED|ED counterparts. We believe the cause of this difference could be the volatility in the number of busy repairmen, that has come as a result of having QED repair capacity level choice; in this case, there will be idle servers in the limit unlike the ED|ED and every delay until replacement will observe a different trajectory of service rates for the service completions occurring during the delay.

It can be observed from Table 1.3 that mean delay values for QED|QED approach the limiting values faster compared to the QED|ED case. The reason for this observation could be because both of the capacity preferences are the same in the QED|QED case, since we only use order n capacity parameters and they are both equal to ρ, this system has the same value for both s_n and m_n. In the following tables, we compare the simulation results between the exponential and nonexponential systems, with breakdown and repair times taken to be other types of distributions. We created two nonexponential cases; for both breakdown and repair distributions, in one case we chose to use an Erlang distribution with the same mean value but one fourth of the variance of the exponential case and in the other a hyper-exponential distribution with the same mean value but 5.5 times the variance of the exponential

$E[D_n^{rpl}(s_n, m_n)]$					
n	s_n	m_n	Sim. Expo.	Sim. Erlang	Sim. Hyper-expo.
24	2	0	3.0058	2.9951	3.0446
60	5	0	2.9975	2.9984	2.9034
108	9	0	3.0203	3.0003	2.9748
$P(D_n^{rpl}(s_n, m_n) > \lim_{n\to\infty} D_n^{rpl}(s_n, m_n))$					
24	2	0	0.476	0.4841	0.4388
60	5	0	0.4809	0.4879	0.4179
108	9	0	0.497	0.4901	0.4561

Table 1.4: Comparison of Simulation Results between the Exponential and Nonexponential Cases for Delay until Replacement, ED|ED, $\lambda = 1, \mu = 3, \bar{s} = 1/12, \bar{m} = 0, \hat{s} = \hat{m} = 0, Erlang \sim (4, 4*\lambda(or\mu)), Hyper{-}expo. \sim (0.8, 4*\lambda(or\mu), 0.2, \lambda(or\mu)/4)$

case.

For ED|ED, there is not a critical difference between the mean delays for exponential and nonexponential cases; however, there is a slight decrease in the probability of unacceptable delay for the hyper-exponential case. We believe the reason for this is that we have used a high probability (0.8) of choosing the repair time as four times the previous rate.

For both QED|ED and QED|QED, we observe that the mean delays are higher for systems with higher variance, but the probability of unacceptable delay is lower for the same threshold value; hence, the delay distribution has a heavier tail when the variance is high and when a delay is experienced it lasts for a longer time compared to a system with lower variance.

1.7 Conclusion and Extensions

In this essay, we study a Markovian machine repair problem with spares where a certain number of identical units have to be kept as working (enough to cover the demand base), and these units break down spontaneously. When this happens, the broken unit (e.g. pc, machine component, city transit bus) is sent to the repair

$E[D_n^{rpl}(s_n, m_n)]$					
n	s_n	m_n	Sim. Expo.	Sim. Erlang	Sim. Hyper-expo.
24	7	4	0.2489	0.2297	0.2807
48	14	8	0.2156	0.2035	0.2374
96	28	16	0.1922	0.1829	0.1969
$P(D_n^{rpl}(s_n, m_n) > \lim_{n\to\infty} D_n^{rpl}(s_n, m_n))$					
24	7	4	0.6646	0.7404	0.5747
48	14	8	0.6779	0.7392	0.5851
96	28	16	0.6761	0.7300	0.562

Table 1.5: Comparison of Simulation Results between the Exponential and Non-exponential Cases for Delay until Replacement, QED|ED, $\lambda = 1, \mu = 3, \bar{s} = 7/24, \bar{m} = 4/24, \hat{s} = \hat{m} = 0$, $Erlang \sim (4, 4 * \lambda(or\ \mu)), Hyper - expo. \sim (0.8, 4 * \lambda(or\ \mu), 0.2, \lambda(or\ \mu)/4)$

$E[D_n^{rpi}(s_n, m_n)]$					
n	s_n	m_n	Exact V. Expo.	Sim. Erlang	Sim. Hyper-expo.
24	8	8	0.1102	0.0792	0.1417
48	16	16	0.0765	0.0582	0.1016
96	32	32	0.0534	0.0398	0.0769
$P(D_n^{rpi}(s_n, m_n) > \lim_{n\to\infty} D_n^{rpi}(s_n, m_n))$					
24	8	8	0.6423	0.6598	0.5918
48	16	16	0.64	0.6701	0.6119
96	32	32	0.6384	0.6594	0.6386

Table 1.6: Comparison of Simulation Results between the Exponential and Non-exponential Cases for Delay until Repair Initiation, QED|QED, $\lambda = 1, \mu = 3, \bar{s} = 1/3, \bar{m} = 1/3, \hat{s} = \hat{m} = 0$, $Erlang \sim (4, 4 * \lambda(or\ \mu)), Hyper - expo. \sim (0.8, 4 * \lambda(or\ \mu), 0.2, \lambda(or\ \mu)/4)$

center while the backup center provides an operational unit to the demand as soon as possible. All units within the system as well as repair servers were assumed to be identical, and the breakdowns were assumed to happen only to the working units and exponentially; repairs were also assumed to be completed exponentially. The breakdowns and repairs were assumed to happen independently.

We demonstrated an asymptotic analysis of the relationship between the delay performance and capacity level choices (the number of repair servers and the spare

units kept) for two types of delays: delay until replacement provision to the demand base, delay until repair initiation of a broken unit. First, we obtained the limiting mean delay for both delays and minimized an approximate cost rate to determine the order n parameters for both types; this demonstrated a *capacity choice frontier*. Second, we gave limiting distributions for both delays and demonstrated their use for either as complements to the first set of results to obtain the order $\sqrt{n}$ parameters or as constraints in alternative approximate optimization problems. Among the interesting results, we find that for highly utilized systems in both repair and spares capacity, delay distribution do not depend on the spares capacity.

Extensions of this work could be to explore on what other complex machine repair, reliability, and closed queuing systems this type of asymptotic analysis can be applied to provide approximations. Moreover, it could be interesting to focus on how to benefit from this approximate analytical performance analysis that establish the relationship between the capacity and delay performance, for modeling competition among maintenance providers according to the amount of resources they are willing to devote and price they are charging; where the buyer of these services (e.g., a transportation firm) is trying to get the best possible delay performance with minimum cost.

2

Diffusion Approximations and Near-Optimal Design of a Make-to-Stock Queue with Perishable Goods and Impatient Customers

2.1 Introduction

In this essay, we approximate and seek the near-optimal design of a make-to-stock queue with perishable goods and impatient customers via modeling the system as a double-sided queueing system subject to abandonment from both sides. For this model, there are two types of exogenous arrivals, one resulting in unit increases in the queue length and the other resulting in unit decreases. We refer to these as positive and negative job arrivals, respectively. The queueing discipline is assumed to be FIFO for both sides. By this we mean that a newly arriving negative (resp. positive) job is matched with the oldest positive (resp. negative) job in queue, if such a job exists. When a match between a positive and a negative job is established, both jobs instantaneously leave the system. In this sense, an arriving negative job provides the service to a waiting positive job, and vice versa. If no such match is immediately available, the arriving job waits in the the queue. Jobs that must wait

in queue before potential matching occurs are subject to abandonment.

There are several potential applications of this model. Our primary motivating example is a production system with perishable inventory and impatient backlogged customers. As our focus will be on this example, we will refer to positive jobs as inventory stock and negative jobs as backlogged customers. A review of inventory models with goods that perish completely after a time delay can be found in Nahmias (1982), whereas Raafat (1991) investigates models of continuously deteriorating inventory. Goyal and Giri (2001) provide recent trends in perishable inventory models. Among the infinite horizon perishable inventory/production models for single items with stationary stochastic demand, this is, to our knowledge, the first study to include customer abandonment.

Our approximation approach falls into the category of conventional heavy traffic limit theory. Specifically, we consider a sequence of diffusion-scaled systems as some parameter – in our case the demand rate μ goes to infinity. Abandonment phenomenon in such models constitutes a significant detour from typical conventional heavy traffic limit theorems and, as such, has a relatively small associated body of literature. Ward and Glynn (2005), Reed and Ward (2008), and Jennings and Reed (2010) all study one-sided queueing models with general arrival, service, and abandonment distributions. The last paper studies an overloaded multi-class system. Theorem 8 of our paper considers the overloaded case, which we refer to as *unbalanced*, as overloadedness can take place on either side of the queue. Unbalancedness is in contrast to a balanced system, which in this work refers to environments in which the product and demand rates differ on a scale smaller than that of the rates themselves.

Our approach is different than the technique employed in the papers above in that we study directly the scalings of steady state distributions, whereas the other works begin by the studying of diffusion-scaled *processes*. Our main results are

first articulated for a Markovian system. Given an original queueing model and its associated steady state distribution, we consider a sequence of scaled steady state queue length distributions. We prove in Theorems 7 and 8 that under certain assumptions on the magnitude of the abandonment rates, the scaled distributions converge to the distribution of a continuous random variable. Slight modifications of the limiting distributions serve as accurate approximations of the distribution of the original system, a fact that we verify in numerical studies in Section 2.4.1.

Further, the limiting distributions allow one to approximate system performance metrics, such as stock-out probabilities, expected queue lengths, lost sales, etc. Given these approximations to system performance, one can formulate a math program involving these approximations and then proceed to optimize. In Section 2.3 we consider an approximate cost minimization problem, where the scaled difference between production rate and customer arrival rate (i.e. the difference between rates of positive and negative jobs) β is the decision variable. The solution from the approximate problem is then interpreted for the original system, yielding an approximately optimal (unscaled) production rate α. We compare this value to the true optimal quantity. Extensive numerical results are presented in 2.4.2. Our conclusion is that our approach is extremely accurate in most reasonable parameter settings. This particular goal of our work, finding the optimal production rate, is comparable but distinguished from Graves (1982), who studies production control rules for queueing systems with customer abandonment.

Heavy traffic limit theory can be an appealing technique for studying large and/or complex systems. One of the benefits of the theory and accompanying analysis is the resulting simplification of otherwise involved formulae, yielding aesthetically pleasing expressions that not only capture the essence of the true solution but also make the system amenable to otherwise elusive insight. Diffusion approximations such as ours strike the delicate balance of preserving the underlying stochasticity of the original

model and doing so in a more tractable form. Another advantage is that increases in the size and complexity of the original models – for instance through the inclusion of multiple customer and inventory classes or generally distributed random variables – can sometimes be accommodated without any additional complexity in the ultimate approximation. Indeed, though proving the result formally is beyond the scope of this work, we dedicate Section 2.6 to conjecturing the analogous approximation under general distributional assumptions for interarrival and abandonment times. The approximation of the queue length distribution is tested using numerical examples.

Alternative asymptotic expressions involving the use of Laplace's method applied to a double-sided queue were employed by Zenios (1999), who proposed a queueing model with reneging as an abstraction of transplant waiting list dynamics. One fundamental distinction between this and our work is that the former assumes abandonment only on the patient side of the queue. Under the assumption that the system is unbalanced, asymptotically, queues form only on the patient side; hence, there's no need for abandonment on the organ side. We address this unbalanced case in Theorem 8. Zenios models the demand and supply of cadaveric organs, and elucidates the relationship between an organ allocation policy and performance metrics such as queue length. Similarly, our model captures an interesting phenomenon emerging within the organ transplant area. A person – call them a would-be *recipient* – requiring an organ transplant may have a loved one – *donor* – who is willing to donate but who is not a viable match for the recipient. All is not lost for this couple. If a similar recipient-donor couple exists such that the organ offered by the donor of the second couple satisfies the needs of the recipient from the first couple and at the same time the needs of the recipient from the second couple matches the organ offered by the donor of the first couple, then these two couples are complimentary. Our queuing model captures the imbalance in the needs of couples looking for matches and the delays until viable matches materialize. In this case, a couple

abandons if an organ is found through other means or transplantation is no longer an option.

There are additional papers to note. Markovian double-sided queues have been studied by Conolly (2002), Ishihara and Utsumi (2005), and Mendoze et al. (2009). The inspiring application area for the first two studies is a taxi stand where passenger arrivals awaiting taxis form one side of the queue, whereas taxis waiting for passengers form the other side. Having abandonments at both sides is natural in this setting.

Conolly (2002) provides transient analysis of the queue length distribution. Ishihara and Utsumi (2005) allow multiple customer classes and investigate queueing dynamics during a busy cycle. Mendoza et al. (2009) impose thresholds on both sides of the queue to assure stability in the absence of abandonment and investigate minimization of the expected total cost. We investigate the role of thresholds in Section 2.5.

Among non-Markovian models, Xiao-ling and Jian (2004) consider the double-sided queue with generalized batch arrivals on one side and Poisson on the other. Similar to Mendoza et al., thresholds are imposed on both sides to ensure stability.

We also see the phenomenon of matching two Poisson streams with each other in finance, specifically in dealership markets, where the market-maker determines the buying and selling prices for stocks. This determines the rates of Poisson arrival streams for buyers and sellers, respectively, and these arrival rates may depend on many factors, e.g. the past stock prices, the mood of the market, etc.

The market-maker starts the market with some inventory of stocks and cash with the objective of keeping a positive amount of both quantities and profiting on the temporary fluctuations of supply and demand. The basic case of having the market maker choosing a single price, hence a single arrival rate for either of buyer and sellers has first been analyzed in Garman (1976) in his seminal paper. Afterwards, many studies have been conducted on market microstructure; Biais et al. (2005) provide

a review. The dealership model has evolved into more complicated forms employing nonstationary behavior for Poisson streams with the focus of exploring the pricing trends; see Horst and Rothe (2008). A major difference important to emphasize is that these models do not include perishability, as this is not a feature of stocks or cash reserve.

The remainder of the essay is organized as follows. In the immediately following section we present the model, framework for diffusion-scaling, and the main asymptotic results for balanced and unbalanced systems. Section 2.3 employs these limiting distributions in estimating various performance measures and obtaining a nearly optimal production rate. Next, numerical studies investigate the performance of the approximation, both in terms of estimating the queue length distribution as well as in terms of finding the optimal production rate. The remainder of the paper considers extensions of the model. Thresholds with and without abandonment are explored in Section 2.5.

Finally, in Section 2.6, we provide a set of equations that govern the virtual waiting time for customers and products under general distributional assumptions on the interarrival and abandonment times. We also provide a conjecture for the steady state of the limiting scaled queue length distribution under these relaxed assumptions. These analogs to the main theorems are validated with simulation results. The paper concludes with closing remarks.

2.2 The model and main results

2.2.1 The model

In this section, we provide the model for the Markovian double-sided queue with unlimited waiting space (or with no thresholds) at both ends, where stability is achieved by having abandonment on both sides. We assume the arrivals generating the unit increases and the arrivals generating the unit decreases in the queue length

occur according to a Poisson process with rate α and μ, respectively; hence, we allow the queue length to be any positive or negative number.

In the previous section, we emphasized that our focus will be on the application of this model to a production system with perishable inventory and impatient customers. We assume that production completions increase the queue length, whereas the customer arrivals decrease it.

Since there are no thresholds on either side, when a product is finished, if there are no customers in the system at that time, it starts waiting to be matched with an arriving customer. If more products get finished before a customer arrives, then a queue of finished products will form on the production side (which corresponds to the positive values of the queue length).

We assume the inventory perishes at an exponential rate γ_1 so all products in the queue will be subject to departure due to perishability as well as due to being matched with an arriving customer, whichever happens first. When a customer arrives, among the products that lasted until that time, the product that arrived at the earliest time will be matched with the customer and they will both leave the queue. Hence, we assume FIFO discipline.

In case there are no products left when the customer arrives, a customer queue may form with similar dynamics where we assume each customer has a patience time which is exponentially distributed with rate γ_2. Thus, a customer will either depart with a product or abandon due to not being matched with one until his/her patience runs out. All production completions, customer arrivals, product perishing, and customer abandonments are assumed to be mutually independent. Figure 2.1 illustrates the model. Note that at any time a queue may form on only one side because the arrivals coming to the opposite side will be matched immediately with the oldest item in the queue.

We denote the queue length process by $Q \equiv \{Q(t), t \geq 0\}$, and it is assumed

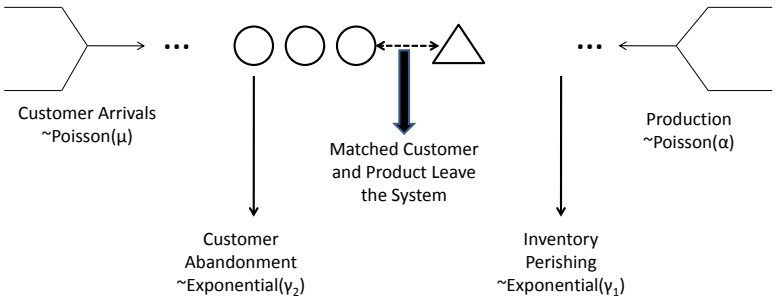

FIGURE 2.1: Markovian Double-sided Queue with No Thresholds

to be right-continuous with left limits. The system framework and aforementioned assumptions imply that Q is a birth-and-death process with births corresponding to production completions or customer abandonments (the latter possible only for transitions made from negative states) and deaths corresponding to customer arrivals or perishing of products (the latter possible only for transitions made from positive states).

Suppose that k represents the value of queue length ($k \in \mathbb{Z}$) at the considered moment. Let $k^+ = max(k, 0)$ and $k^- = max(-k, 0)$. Since the production completions and customer abandonments occur exponentially and independently from each other, the birth rate $\alpha(k)$ is the sum of the production rate α and total abandonment rate, thus, $\alpha(k) = \alpha + k^-\gamma_2$. Similarly, $\mu(k) = \mu + k^+\gamma_1$. For given α, μ, γ_1, and γ_2, the probability distribution of the queue length is given as follows ($p_0 = P(Q(\infty) = 0)$ is the appropriate normalizing constant):

$$P(Q(\infty)=k) \equiv p_k = \begin{cases} p_0\,(\alpha)^k / \prod_{i=1}^{k}(\mu + i\gamma_1) & k > 0 \\ p_0\,(\mu)^{k^-} / \prod_{i=1}^{k^-}(\alpha + i\gamma_2) & k < 0 \\ p_0 = \left[1 + \sum_{k=1}^{\infty}\left(\alpha^k / \prod_{i=1}^{k}(\alpha + i\gamma_1) + \mu^k / \prod_{i=1}^{k}(\mu + i\gamma_2)\right)\right]^{-1}. \end{cases} \tag{2.1}$$

2.2.2 Asymptotic expressions for balanced cases

The ubiquitous assumption is that abandonment rates are relatively small: $max(\gamma_1, \gamma_2) \ll min(\alpha, \mu)$, and γ_1, γ_2 are $o(\mu)$. Here we derive asymptotic expressions for the queue length distribution, $P(Q(\infty) \leq k), (k \in \mathbb{Z})$, for two types of systems: balanced and unbalanced. By a balanced system, we mean that the difference between the production completion rate α and customer arrival rate μ is much smaller compared to the customer arrival rate, i.e. $|\alpha - \mu| \ll min(\alpha, \mu)$.

The asymptotic regime for the balanced system assumes that both rates become infinitely large, with the difference between them being $O\left(\sqrt{\mu}\right)$. Moreover, our numerical analysis shows that the approximation works well (the maximum percentage error between the actual and approximate cumulative distribution function of queue length $< 5\%$), even when: 1) the above assumption is valid, e.g. $\alpha = 150, \mu = 100$; 2) the system is not necessarily large, e.g. $\alpha = 7, \mu = 5$; 3) γ_2 does not satisfy the assumption above, e.g. $\alpha = 80, \mu = 100, \gamma_1 = 1, \gamma_2 = 100$. Out of 132 cases tested, only 5 cases had higher percentage error than 5%, as can be seen from Tables 2.1-2.4.

Towards an approximation we consider a sequence of systems indexed by n such that:

$$\begin{array}{l} production \quad rate = \lambda n + \beta\sqrt{n} \\ demand \quad rate = \lambda n \\ goods \quad abandonment \quad rate = \theta_1 \\ customer \quad abandonment \quad rate = \theta_2. \end{array} \tag{2.2}$$

Note that the difference between the production and demand rates is $O(\sqrt{n})$ whereas the demand rate is $O(n)$. Hence, based on the above definition, we categorize each system in the sequence as balanced. We will denote the double-sided queue length, dependent on n, with $Q^n \equiv \{Q^n(t),\ t \geq 0\}$. Moreover, let $N(\mu, \sigma^2, l, u)$ denote the truncation and renormalization of a normal random variable with mean μ and variance σ^2, where l and u are the lower and upper bounds of the truncation, respectively. Then, the following theorem summarizes the limiting behavior of the queue-length process.

Theorem 7. *Let*

$$\begin{aligned} X^- &\sim N(\beta/\theta_2, \lambda/\theta_2, -\infty, 0) \\ X^+ &\sim N(\beta/\theta_1, \lambda/\theta_1, 0, \infty) \\ a &= \left(1 + \sqrt{\theta_2/\theta_1}\frac{h\left(\beta/\sqrt{\lambda\theta_2}\right)}{h\left(-\beta/\sqrt{\lambda\theta_1}\right)}\right)^{-1}. \end{aligned} \tag{2.3}$$

For the sequence of systems indexed by n with steady state queue length $Q^n(\infty)$, the sequence of out-of-stock probabilities converges:

$$\lim_{n\to\infty} P(Q^n(\infty) \leq 0) = a, \quad as \quad n \to \infty.$$

Moreover, the sequence of queue length distributions converge. That is, for any $b \in \mathbb{R}$ we have

$$\lim_{n\to\infty} P(Q^n(\infty) \leq b\sqrt{n}) = a \cdot P(X^- \leq b) + (1-a) \cdot P(X^+ \leq b). \tag{2.4}$$

Towards approximating the original system we set the parameters for the n^{th} system equal to the functions of the parameters of the real system in the following way:

$$n = \mu/\lambda \tag{2.5a}$$

$$\theta_1 = \gamma_1 \tag{2.5b}$$

$$\theta_2 = \gamma_2 \tag{2.5c}$$

$$\beta = (\alpha - \mu)\sqrt{\lambda/\mu}. \tag{2.5d}$$

Note that λ can be chosen arbitrarily as long as the scale parameter, n, times λ equals μ. In the following section, we will set λ to be 1 for ease of notation.

Hence, for a system with large values of α and μ, based on Theorem 7, we can use the limit shown in (2.4) to approximate the queue length distribution:

$$P(Q(\infty) \leq b) \cong a \cdot P\left(X^- \leq b\sqrt{\lambda/\mu}\right) + (1-a) \cdot P\left(X^+ \leq b\sqrt{\lambda/\mu}\right). \tag{2.6}$$

Because b is either positive or negative, (2.6) simplifies, as either $P\left(X^+ \leq b\sqrt{\lambda/\mu}\right)$ $= 0$ or $P\left(X^- \leq b\sqrt{\lambda/\mu}\right) = 1$.

In the numerical examples in Section 2.4 we observe that the maximum error between the steady state cumulative distribution function of the original system and that of our approximation has been less than 0.05 in all cases, usually it is much smaller.

2.2.3 Asymptotic expressions for unbalanced cases

Now we look at the unbalanced system where the absolute difference between the production and arrival rates, $|\alpha - \mu|$, is $O(\mu)$. Consider a sequence of systems, indexed by n, such that in the n^{th} system we have the following rates:

$$production \quad rate = (\lambda + d)n \tag{2.7a}$$

$$demand \quad rate = \lambda n \tag{2.7b}$$

$$goods \quad abandonment \quad rate = \theta_1 \tag{2.7c}$$

$$customer \quad abandonment \quad rate = \theta_2. \tag{2.7d}$$

Since we consider the unbalanced system, the limiting probability distribution will have all of its mass on the side having the arrival stream with the higher rate. For $d > 0$, it will be on the production side, and for $d < 0$, it will be on the customer side. Since $d = 0$ corresponds to the balanced case with $\beta = 0$, it will not be considered here. Consider an unbalanced system with $\alpha \gg \mu$. The overloaded system (very large values of α and μ) will have the balance equation:

$$\begin{aligned} &\alpha = \mu + Avg.\, Queue\ Length \cdot \gamma_1 \\ &\Rightarrow Avg.\, Queue\ Length = (\alpha - \mu)/\gamma_1 \end{aligned} \tag{2.8}$$

From (2.7) it corresponds to $Avg.\, Queue\ Length = dn/\theta_1$ and from conventional heavy traffic limit theory, it is known that the recurrent states of the limiting queue length distribution will be within $O(\sqrt{n})$ of the centering constant, dn/θ_1. We assume $d \neq 0$.

Theorem 8.

$$\lim_{n\to\infty} P(Q^n(\infty) \leq dn/\theta_1 + b\sqrt{n}) = \Phi\left(b/\sqrt{(\lambda + d)/\theta_1}\right), d > 0 \tag{2.9a}$$

$$\lim_{n\to\infty} P(Q^n(\infty) \leq dn/\theta_2 + b\sqrt{n}) = \Phi\left(b/\sqrt{\lambda/\theta_2}\right), d < 0. \tag{2.9b}$$

The above formula can be used to approximate an unbalanced system in the following way:

$$
\begin{aligned}
P(Q(\infty) \leq b) &\cong \Phi\left(\left(b - \tfrac{\alpha-\mu}{\gamma_1}\right) / \sqrt{\alpha/\gamma_1}\right) \quad d > 0 \\
P(Q(\infty) \leq b) &\cong \Phi\left(\left(b - \tfrac{\alpha-\mu}{\gamma_2}\right) / \sqrt{\mu/\gamma_2}\right) \quad d < 0.
\end{aligned}
\tag{2.10}
$$

2.3 Approximate performance metrics and near-optimal design

We will now provide approximations for the fill rate, average inventory, and average backorder quantities to propose a closed-form cost rate function leading to near-optimal design in the balanced case. Numerical results of the percentage error between the optimal and approximate values will be given in the next section.

In the proof of Theorem 7 we have shown that $P(Q(\infty) \leq 0) \cong a$, where a is as in (2.3). Since this quantity is our proposed approximation for the probability of being out of stock, $(1-a)$ is the approximate fill rate. Without loss of generality, we will assume $\lambda = 1$ for ease of notation. Moreover, from (2.5) and (2.6) for sufficiently large n

$$
\begin{aligned}
&Avg.\ Inv. = E[max(Q(\infty), 0)] \cong (1-a)\,\sqrt{\mu}\,E[X^+ | X^+ > 0] \\
&= (1-a)\sqrt{\mu}\left[\beta/\theta_1 + h\left(-\beta/\sqrt{\theta_1}\right)\sqrt{1/\theta_1}\right]
\end{aligned}
\tag{2.11}
$$

where $X^+ \sim N\left(\beta/\theta_1, 1/\theta_1, 0, \infty\right)$ as in (2.3), and similarly,

$$
\begin{aligned}
&Avg.\ Backorders = E[max(-Q(\infty), 0)] \cong a\,\sqrt{\mu}\,E[-X^- | X^- < 0] \\
&= a\sqrt{\mu}\left[-\beta/\theta_2 + h\left(\beta/\sqrt{\theta_2}\right)\sqrt{1/\theta_2}\right].
\end{aligned}
\tag{2.12}
$$

We now suggest an approximation to the expected cost rate where $\hat{h}$, w, b, and p represent the holding cost, wages, backorder cost, and abandonment penalty per customer per unit time, respectively. Then, with parameters chosen to satisfy (2.5)

and from (2.11) and (2.12), the cost rate C can be expressed and then approximated as follows:

$$
\begin{aligned}
&C(\alpha) = \hat{h}E[max(Q(\infty), 0)] + w\alpha + (b + p\gamma_2)E[max(-Q(\infty), 0)] \\
&[C(\alpha) - w\mu]/\sqrt{\mu} \cong \hat{C}(\beta) = \hat{h}(1-a)\left[\beta/\theta_1 + h\left(-\beta/\sqrt{\theta_1}\right)\sqrt{1/\theta_1}\right] + w\beta \\
&+(b + p\theta_2)a\left[-\beta/\theta_2 + h\left(\beta/\sqrt{\theta_2}\right)\sqrt{1/\theta_2}\right].
\end{aligned}
\tag{2.13}
$$

2.4 Numerical analysis

We will now present our numerical results on the effectiveness of our approximation. The first subsection will be devoted to the effectiveness of the queue length approximation in (2.6). We will focus on the balanced system only.

2.4.1 Tests of the Distributional Results

Tables 2.1-2.4 represent systems of different scales. We first choose a value for the customer arrival rate μ and show the maximum percentage error for different combinations of α, γ_1, and γ_2 choices of the approximate cdf to the actual one. The choice of μ represents the "size" of the system since $|\alpha - \mu| \ll \mu$. We expect our approximation to work better for large values of μ. We observed the steady state and approximate values of $P(Q(\infty) \leq b)$ for $b = -200, -199, \ldots, 200$ and reported the maximum absolute difference between the cdf and approximate cdf values. Both actual and approximate cdf values are very close to zero and one for the b values lower than -200 and higher than 200, respectively, so we do not report on these trivial cases. It can be seen that all percentage maximum absolute errors for the two distributions are less than 5% when $|\alpha - \mu| \ll \mu$ and $\gamma_1, \gamma_2 < 0.1\mu$.

$\mu = 100$	Abandonment Rates,(γ_1, γ_2)						
$\alpha - \mu$	(1,1)	(1,3)	(1,5)	(1,10)	(1,20)	(1,50)	(1,100)
-50	0.66	1.15	1.43	1.77	2.93	5.01	6.88
-40	0.66	1.13	1.35	1.78	2.78	4.20	5.52
-30	0.66	1.07	1.24	1.66	2.30	3.21	4.05
-20	0.66	0.97	1.09	1.25	1.53	2.01	2.50
-10	0.59	0.78	0.88	1.01	1.13	1.12	0.96
0	0.35	0.46	0.51	0.57	0.62	0.65	0.65
10	1.24	1.34	1.39	1.45	1.51	1.55	1.57
20	2.29	2.25	2.26	2.29	2.31	2.34	2.35
30	3.25	3.23	3.22	3.21	3.19	3.17	3.17
40	4.14	4.14	4.14	4.14	4.13	4.13	4.13
50	4.97	4.97	4.97	4.97	4.97	4.97	4.97

Table 2.1: Maximum Absolute Difference in *Percentile* between the Actual and Approximate Cdf Values for $\mu = 100$, $\alpha = 50, 60, \ldots, 150$, $(\gamma_1 = 1, \gamma_2 = 1, 3, 5, 10, 20, 50, 100)$

$\mu = 20$	Abandonment Rates,(γ_1, γ_2)		
$\alpha - \mu$	(1,1)	(1,3)	(1,5)
-8	1.46	2.07	2.39
-6	1.41	1.81	2.04
-4	1.27	1.64	1.62
-2	0.91	1.13	1.34
0	0.73	0.96	1.06
2	1.51	1.78	1.89
4	2.34	2.59	2.70
6	3.35	3.37	3.46
8	4.25	4.13	4.18

Table 2.2: Maximum Absolute Difference in *Percentile* between the Actual and Approximate Cdf Values for $\mu = 20$, $\alpha = 12, 14, \ldots, 28$, $(\gamma_1 = 1, \gamma_2 = 1, 3, 5)$

$\mu = 5$	Abandonment Rates,(γ_1, γ_2)	
$\alpha - \mu$	(1,1)	(1,3)
-3	2.80	4.34
-2	2.46	3.12
-1	1.81	2.36
0	1.34	1.71
1	2.72	3.16
2	4.25	4.45
3	5.94	5.67

Table 2.3: Maximum Absolute Difference in *Percentile* between the Actual and Approximate Cdf Values for $\mu = 5$, $\alpha = 2, 3, \ldots, 8$, $(\gamma_1 = 1, \gamma_2 = 1, 3)$

$\mu = 1$	Abandonment Rates,(γ_1, γ_2)	
$\alpha - \mu$	(0.1,0.1)	(0.1,0.3)
-0.3	2.35	2.93
-0.2	1.92	2.57
-0.1	1.31	1.98
0	1.48	1.86
0.1	2.00	2.31
0.2	2.73	2.62
0.3	3.62	3.21

Table 2.4: Maximum Absolute Difference in *Percentile* between the Actual and Approximate Cdf Values for $\mu = 1$, $\alpha = 0.7, 0.8, \ldots, 1.3$, $(\gamma_1 = 0.1, \gamma_2 = 0.1, 0.3)$

2.4.2 Tests of the Approximation

Here we compare the difference between the optimal production rate α and the recommended rate via approximation. To do this we find the approximate α through finding the β, optimizing (2.13), and calculating the α value it corresponds to via (2.2). Moreover, we also give the percentage error between the actual and approximate objective function values. As can be seen from the Tables 2.5-2.8, although the optimal solution values differ in many cases for systems with small scale, the approximate objective function values are very close to the optimal value. Note that the approximation uses (2.5) for matching the real and approximate system param-

eters. We find the optimal α and the corresponding objective function value. Then, we find the β minimizing $\hat{C}(\beta)$ and use it to derive an estimate for optimal α, $\hat{\alpha}$, via (2.5). The data in Tables 2.5-2.8 is in the format $\frac{C(\hat{\alpha})-C(\hat{\alpha})}{C(\hat{\alpha})}/\frac{\hat{\alpha}-\hat{\alpha}}{\hat{\alpha}}$.

We now demonstrate some characteristics observed on the choice of the cost parameters in the objective function. Notice that for overloaded systems with the customer arrival rate (μ) higher than the production rate (α), from the balance equations, the average queue length will be $((\mu - \alpha)/\gamma_2)$. Hence, for each increase in the production rate you will pay the wage per unit time, w, and have one less customer not being backlogged, which will save b/γ_2. Moreover, you will have one less customer not abandoning the queue and it will save the penalty cost per unit time, p. Thus, it follows that if $w > p + b/\gamma_2$, then there is incentive to produce nothing: $\alpha = 0$. Our numerical examples support this claim. We summarize the observations in Table 2.9 ($\hat{\alpha}$ is the production rate provided by our heuristic and $\alpha*$ is the optimal α).

The p value that makes both sides of the inequality equal to each other is called $p_{critical}$ and we check for higher and lower values of p than $p_{critical}$, where a lower p value means no incentive to produce according to the inequality above. Notice that at some cases although the inequality is not satisfied, there is still no incentive to produce. The reason is this inequality is derived from fluid limit equations for overloaded systems, and it assumes no inventory held when customer arrival rate is higher. However, we observed nonzero expected inventory in the aforementioned cases.

2.5 Thresholds

Here we analyze an extension of our model in Section 2.2 by imposing a threshold on the positive side of the queue, which corresponds to having limited space for inventory

	$w=1$			$w=5$		
$\mu=100$	γ_1,γ_2			γ_1,γ_2		
$(\hat{h},b,p)$	(1,1)	(1,3)	(1,10)	(3,1)	(3,3)	(3,10)
(1, 1, 5)	$\frac{0.00}{0.11}$	$\frac{0.00}{0.14}$	$\frac{0.00}{0.14}$	$\frac{0.00}{0.10}$	$\frac{0.00}{-0.69}$	$\frac{0.01}{-7.60}$
(1, 1, 8)	$\frac{0.00}{-0.01}$	$\frac{0.00}{-0.07}$	$\frac{0.00}{-0.14}$	$\frac{0.00}{0.39}$	$\frac{0.01}{0.82}$	$\frac{0.02}{1.51}$
(1, 3, 5)	$\frac{0.00}{0.03}$	$\frac{0.00}{0.08}$	$\frac{0.00}{0.12}$	$\frac{0.00}{0.37}$	$\frac{0.00}{0.35}$	$\frac{0.00}{-1.74}$
(1, 3, 8)	$\frac{0.00}{-0.08}$	$\frac{0.00}{-0.10}$	$\frac{0.00}{-0.16}$	$\frac{0.01}{0.40}$	$\frac{0.01}{0.81}$	$\frac{0.02}{1.49}$
(1, 10, 5)	$\frac{0.01}{-0.19}$	$\frac{0.00}{-0.07}$	$\frac{0.00}{0.04}$	$\frac{0.00}{0.36}$	$\frac{0.01}{0.82}$	$\frac{0.00}{1.13}$
(1, 10, 8)	$\frac{0.01}{-0.26}$	$\frac{0.01}{-0.21}$	$\frac{0.00}{-0.21}$	$\frac{0.00}{0.29}$	$\frac{0.01}{0.72}$	$\frac{0.02}{1.39}$
(3, 1, 5)	$\frac{0.01}{0.24}$	$\frac{0.01}{0.36}$	$\frac{0.02}{0.47}$	$\frac{0.00}{0.07}$	$\frac{0.00}{-0.85}$	$\frac{0.01}{-8.49}$
(3, 1, 8)	$\frac{0.00}{0.16}$	$\frac{0.01}{0.20}$	$\frac{0.01}{0.22}$	$\frac{0.00}{0.39}$	$\frac{0.01}{0.81}$	$\frac{0.02}{1.55}$
(3, 3, 5)	$\frac{0.01}{0.18}$	$\frac{0.01}{0.32}$	$\frac{0.02}{0.45}$	$\frac{0.00}{0.34}$	$\frac{0.00}{0.27}$	$\frac{0.00}{-2.20}$
(3, 3, 8)	$\frac{0.00}{0.11}$	$\frac{0.00}{0.17}$	$\frac{0.00}{0.20}$	$\frac{0.01}{0.41}$	$\frac{0.01}{0.81}$	$\frac{0.03}{1.54}$
(3, 10, 5)	$\frac{0.00}{0.03}$	$\frac{0.01}{0.20}$	$\frac{0.01}{0.38}$	$\frac{0.01}{0.37}$	$\frac{0.01}{0.81}$	$\frac{0.00}{0.98}$
(3, 10, 8)	$\frac{0.00}{-0.03}$	$\frac{0.00}{0.07}$	$\frac{0.00}{0.15}$	$\frac{0.00}{0.32}$	$\frac{0.01}{0.76}$	$\frac{0.03}{1.46}$

Table 2.5: *Percentage* Error between the Optimal and Approximate Rates (denumerator) and Objective Function (numerator), $\mu = 100$

in a make-to-stock system. Since thresholds provide stability, such systems can also occupy nonperishing inventory. An illustration similar to Figure 2.1 is shown in Figure 2.2. Let the threshold value be $\hat{T}$ and assume the same notation as in Section 2.2. We explore three different models all with thresholds:

(1) with perishing inventory;

	$w=1$			$w=5$		
$\mu=20$	γ_1,γ_2			γ_1,γ_2		
$(\hat{h},b,p)$	(10,1)	(10,3)	(10,10)	(3,1)	(3,3)	(3,10)
(10, 1, 5)	$\frac{0.12}{2.49}$	$\frac{0.28}{5.09}$	$\frac{0.56}{9.06}$	$\frac{0.00}{-0.60}$	$\frac{0.07}{-14.79}$	$\frac{-2.19}{-100.00}$
(10, 1, 8)	$\frac{0.07}{1.82}$	$\frac{0.11}{2.89}$	$\frac{0.11}{3.41}$	$\frac{0.03}{1.78}$	$\frac{0.09}{4.25}$	$\frac{0.26}{10.53}$
(10, 3, 5)	$\frac{0.08}{2.06}$	$\frac{0.24}{4.59}$	$\frac{0.51}{8.59}$	$\frac{0.02}{1.51}$	$\frac{0.00}{-1.04}$	$\frac{0.25}{-100.00}$
(10, 3, 8)	$\frac{0.04}{1.40}$	$\frac{0.08}{2.47}$	$\frac{0.10}{3.14}$	$\frac{0.06}{2.06}$	$\frac{0.11}{4.45}$	$\frac{0.28}{10.42}$
(10, 10, 5)	$\frac{0.01}{0.58}$	$\frac{0.11}{2.89}$	$\frac{0.38}{7.01}$	$\frac{0.07}{2.05}$	$\frac{0.09}{4.25}$	$\frac{0.00}{1.91}$
(10, 10, 8)	$\frac{0.00}{-0.01}$	$\frac{0.02}{1.13}$	$\frac{0.05}{2.21}$	$\frac{0.07}{1.91}$	$\frac{0.16}{4.43}$	$\frac{0.33}{9.94}$

Table 2.6: *Percentage* Error between the Optimal and Approximate Rates (denumerator) and Objective Function (numerator), $\mu = 20$

	$w=1$			$w=5$		
$\mu=5$	γ_1,γ_2			γ_1,γ_2		
$(\hat{h},b,p)$	(3,1)	(3,3)	(3,10)	(1,1)	(1,3)	(1,10)
(1, 1, 5)	$\frac{0.12}{3.47}$	$\frac{0.21}{5.48}$	$\frac{0.34}{7.96}$	$\frac{0.00}{1.63}$	$\frac{0.14}{-100.00}$	$\frac{-5.46}{-100.00}$
(1, 1, 8)	$\frac{0.00}{0.32}$	$\frac{0.00}{-0.65}$	$\frac{0.03}{-1.70}$	$\frac{0.18}{6.34}$	$\frac{0.40}{12.83}$	$\frac{0.89}{25.25}$
(1, 3, 5)	$\frac{0.02}{1.23}$	$\frac{0.11}{3.80}$	$\frac{0.28}{6.90}$	$\frac{0.14}{6.22}$	$\frac{0.04}{10.42}$	$\frac{-10.02}{-100.00}$
(1, 3, 8)	$\frac{0.02}{-1.15}$	$\frac{0.03}{-1.57}$	$\frac{0.04}{-2.15}$	$\frac{0.20}{5.49}$	$\frac{0.42}{11.81}$	$\frac{0.90}{23.78}$
(1, 10, 5)	$\frac{0.19}{-3.60}$	$\frac{0.00}{-0.65}$	$\frac{0.11}{4.03}$	$\frac{0.15}{4.07}$	$\frac{0.40}{12.83}$	$\frac{0.32}{82.49}$
(1, 10, 8)	$\frac{0.40}{-5.02}$	$\frac{0.22}{-4.24}$	$\frac{0.12}{-3.47}$	$\frac{0.11}{3.11}$	$\frac{0.39}{8.94}$	$\frac{0.88}{20.05}$

Table 2.7: *Percentage* Error between the Optimal and Approximate Rates (denumerator) and Objective Function (numerator), $\mu = 5$

	$w=1$		
$\mu=1$	γ_1,γ_2		
$(\hat{h},b,p)$	(1,1)	(1,3)	(1,10)
(3, 1, 5)	$\frac{3.04}{35.71}$	$\frac{7.72}{106.32}$	$\frac{15.61}{371.85}$
(3, 1, 8)	$\frac{1.90}{22.31}$	$\frac{4.03}{43.30}$	$\frac{7.37}{78.65}$
(3, 3, 5)	$\frac{2.25}{26.07}$	$\frac{6.69}{80.85}$	$\frac{14.76}{298.18}$
(3, 3, 8)	$\frac{1.31}{16.87}$	$\frac{3.49}{37.54}$	$\frac{7.03}{73.60}$
(3, 10, 5)	$\frac{0.54}{9.63}$	$\frac{4.03}{43.30}$	$\frac{12.26}{176.73}$
(3, 10, 8)	$\frac{0.22}{5.39}$	$\frac{2.05}{24.79}$	$\frac{5.98}{61.07}$

Table 2.8: *Percentage* Error between the Optimal and Approximate Rates (denumerator) and Objective Function (numerator), $\mu = 1$

(2) with nonperishing inventory, $\alpha > \mu, |\alpha - \mu| \ll \mu$;

(3) with nonperishing inventory, $\alpha < \mu, |\alpha - \mu| \ll \mu$.

Then, the steady state distribution for model (1) is the same as (2.1) but the threshold $\hat{T}$ will be an upper bound for the positive values of the queue length. Thus, the first expression in (2.1) will be valid for only $0 < k < \hat{T}$ and the normalizing constant p_0 will change accordingly. The steady state distribution for models (2) and (3) is as shown below:

$\mu, \gamma_1, \gamma_2, h, b, w$	$p_{critical}$	Tests		$\hat{\alpha}$	$\alpha*$
$(100, 1, 3, 1, 1, 1)$	0.667	p_{low}	0.567	0	0
		p_{high}	0.767	76.62	76.99
$(100, 1, 10, 1, 3, 1)$	0.7	p_{low}	0.6	0	0
		p_{high}	0.8	69.34	69.75
$(100, 1, 3, 1, 1, 5)$	4.667	p_{low}	4.567	0	0
		p_{high}	4.767	68.52	69.68
$(100, 3, 10, 1, 1, 5)$	4.9	p_{low}	4.8	0	0
		p_{high}	5	45.09	48.8
$(100, 10, 10, 1, 10, 5)$	4	p_{low}	3.9	0	0
		p_{high}	4.1	34.54	40.27
$(20, 1, 10, 1, 1, 1)$	0.9	p_{low}	0.8	0	0
		p_{high}	1	6.29	6.85
$(20, 3, 10, 3, 1, 1)$	0.9	p_{low}	0.8	0	0
		p_{high}	1	1.37	3.14
$(20, 10, 10, 3, 3, 5)$	4.7	p_{low}	4.6	0	0
		p_{high}	4.8	0	0
		p_{other}	5.3	2.01	2.97
$(5, 1, 10, 1, 1, 1)$	0.9	p_{low}	0.8	0	0
		p_{high}	1	0	0
		p_{other}	1.5	2.96	2.11

Table 2.9: Optimal and Approximate Production Rates w.r.t Incentive to Produce

$$P(Q(\infty) = k) \equiv p_k = \begin{cases} p_0 \left(\alpha/\mu\right)^k & 0 < k \leq \hat{T} \\ \\ p_0 \left(\mu\right)^{k^-} / \prod_{i=1}^{k^-} (\alpha + i\gamma_2) & k < 0. \end{cases} \tag{2.14}$$

Theorem 9. ***Model 1.*** *With a threshold $T^n = T\sqrt{n}$ on the inventory of goods and perishable inventory (i. e. $\theta_1 > 0$), let*

$$\begin{aligned} X^- &\sim N\left(\beta/\theta_2, \lambda/\theta_2, -\infty, 0\right) \\ X^+ &\sim N\left(\beta/\theta_1, \lambda/\theta_1, 0, T\right) \\ a &= \left(1 + \sqrt{\theta_2/\theta_1} \quad \frac{\Phi\left(T\sqrt{\theta_1/\lambda} - \beta/\sqrt{\lambda\theta_1}\right) - \Phi\left(-\beta/\sqrt{\lambda\theta_1}\right)}{\phi\left(-\beta/\sqrt{\lambda\theta_1}\right)} \cdot h\left(\beta/\sqrt{\lambda\theta_2}\right)\right)^{-1}. \end{aligned} \tag{2.15}$$

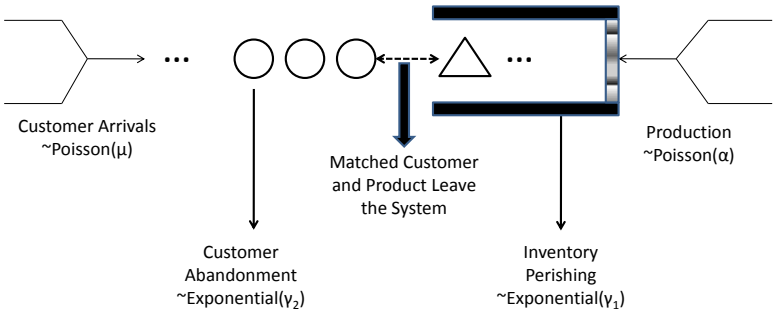

FIGURE 2.2: Markovian Double-sided Queue with Inventory Threshold

For the sequence of systems indexed by n with steady state queue length $Q^n(\infty)$, the sequence of out-of-stock probabilities converges:

$$\lim_{n\to\infty} P(Q^n(\infty) \leq 0) = a, \quad as \quad n \to \infty.$$

Moreover, the sequence of queue length distributions converge. That is, for any $b \in \mathbb{R}$ we have

$$\lim_{n\to\infty} P(Q^n(\infty) \leq b\sqrt{n}) = a \cdot P\left(X^- \leq b\right) + (1-a) \cdot P\left(X^+ \leq b\right). \tag{2.16}$$

Theorem 10. ***Model 2.*** *With a threshold $T^n = T\sqrt{n}$ on the inventory of goods and nonperishing inventory (i. e. $\theta_1 = 0$), and with $\beta > 0$, let*

$$\begin{aligned} X^- &\sim N\left(\beta/\theta_2, \lambda/\theta_2, -\infty, 0\right) \\ (T - X^+) &\sim Exponential\left(-\lambda/\beta, \lambda^2/\beta^2, 0, T\right) \\ a &= \left[1 + \tfrac{\sqrt{\lambda\theta_2}}{\beta}\left[\exp\left\{\beta T/\lambda\right\} - 1\right] h\left(\tfrac{\beta}{\sqrt{\lambda\theta_2}}\right)\right]^{-1}. \end{aligned} \tag{2.17}$$

For the sequence of systems indexed by n with steady state queue length $Q^n(\infty)$, the sequence of out-of-stock probabilities converges:

$$\lim_{n\to\infty} P(Q^n(\infty) \leq 0) = a, \quad as \quad n \to \infty.$$

Moreover, the sequence of queue length distributions converge. That is, for any $b \in \mathbb{R}$ we have

$$\begin{aligned} &\lim_{n\to\infty} P(Q^n(\infty) \leq b\sqrt{n}) \\ &= a \cdot P\left(X^- \leq b\right) + (1-a) \cdot P\left(X^+ \leq b\right). \end{aligned} \tag{2.18}$$

Theorem 11. ***Model 3.*** *With a threshold $T^n = T\sqrt{n}$ on the inventory of goods and nonperishing inventory (i. e. $\theta_1 = 0$), and with $\beta < 0$, let*

$$\begin{aligned} X^- &\sim N\left(\beta/\theta_2, \lambda/\theta_2, -\infty, 0\right) \\ X^+ &\sim Exponential\left(-\lambda/\beta, \lambda^2/\beta^2, 0, T\right) \\ a &= \left[1 - \tfrac{\sqrt{\lambda\theta_2}}{\beta}\left[1 - \exp\left\{\beta T/\lambda\right\}\right] h\left(\tfrac{\beta}{\sqrt{\lambda\theta_2}}\right)\right]^{-1}. \end{aligned} \tag{2.19}$$

For the sequence of systems indexed by n with steady state queue length $Q^n(\infty)$, the sequence of out-of-stock probabilities converges:

$$\lim_{n\to\infty} P(Q^n(\infty) \leq 0) = a, \quad as \quad n \to \infty.$$

Moreover, the sequence of queue length distributions converge. That is, for any $b \in \mathbb{R}$ we have

$$\lim_{n\to\infty} P(Q^n(\infty) \leq b\sqrt{n}) = a \cdot P\left(X^- \leq b\right) + (1-a) \cdot P\left(X^+ \leq b\right). \tag{2.20}$$

2.6 General distributions

In this section we consider our basic model without thresholds as in Figure 2.1. Let the arrival streams for the products and customers be generalized renewal processes but with arbitrary finite variances. Further suppose that abandonment times are generally distributed. We conjecture that the queue length distribution is a normalized weighted average of two truncated normal random variables, just as in the Markovian case.

Let $1/\alpha$ and $1/\mu$ be the product and customer interarrival time means; σ_1^2 and σ_2^2 be the product and customer interarrival time variances, f_1 and f_2 be the density functions of time until product perishing and customer abandonment, respectively. Then, with

$$\begin{aligned} \mu_P &\equiv \tfrac{\alpha-\mu}{\sqrt{\mu} f_1(0)} \\ \mu_C &\equiv \tfrac{\alpha-\mu}{\sqrt{\mu} f_2(0)} \\ \sigma_P^2 &\equiv \left[\sigma_1^2\alpha^3 + \sigma_2^2\mu^3\right] / \left(2\mu f_1(0)\right) \\ \sigma_C^2 &\equiv \left[\sigma_1^2\alpha^3 + \sigma_2^2\mu^3\right] / \left(2\mu f_2(0)\right), \end{aligned} \tag{2.21}$$

we conjecture that (2.6) is still valid where

$$\begin{aligned} X^- &\sim N\left(\mu_C, \sigma_C^2, -\infty, 0\right) \\ X^+ &\sim N\left(\mu_P, \sigma_P^2, 0, \infty\right) \\ a &\equiv \left(1 + \tfrac{\sigma_P}{\sigma_C} h\left(\tfrac{\mu_C}{\sigma_C}\right) / h\left(-\tfrac{\mu_P}{\sigma_P}\right)\right)^{-1}. \end{aligned} \tag{2.22}$$

We have tested our conjecture for eight different systems (ten simulation runs for each), summarized in Table 2.10. Let λ_G represent the reciprocal of the mean for the interarrival time. The arrival streams are combinations of distributions having

Production		Customer Demand		
Dist. Type	C. of Var.	Dist. Type	C. of Var.	Max Dif. in Perc.
Exponential(96)	1	Erl.(2,200)	0.707	1.4622 ± 0.067
Exponential(96)	1	Erl.(5,500)	0.447	1.4605 ± 0.047
Hyp.(0.8,160:0.2,36.923)	1.51	Exp.(100)	1	1.7001 ± 0.062
Hyp.(0.8,500:0.2,22.684)	2.495	Exp.(100)	1	2.1562 ± 0.032
Hyp.(0.8,160:0.2,36.923)	1.51	Erl.(2,200)	0.707	1.6445 ± 0.059
Hyp.(0.8,160:0.2,36.923)	1.51	Erl.(5,500)	0.447	1.5596 ± 0.048
Hyp.(0.8,500:0.2,22.684)	2.495	Erl.(2,200)	0.707	2.17 ± 0.026
Hyp.(0.8,500:0.2,22.684)	2.495	Erl.(5,500)	0.447	2.2042 ± 0.034

Table 2.10: Maximum Absolute Difference in *Percentile* between the Simulated and Approximate Cdf Values from Conjecture for Systems with General Distributions

forms $Exponential(\lambda_G)$,$Erlang\ (k, \lambda_G)$, and $Hyperexponential(p_1, \lambda_G^1, p_2, \lambda_G^2)$.

Note that besides the systems with coefficient of variation equal to 2.495 for production interarrival time, when arrival stream of one side is held at the same distribution, as the coefficient of variation for the other side increases, so does the maximum percentage error between the queue length distribution and our approximation. For the production systems with high coefficient of variation (2.495), we believe having more regular customer arrival times emphasizes the irregularity of production; hence, our conjecture has a higher percentage error as coefficient of variation for customer arrivals decreases.

2.7 Conclusions

In this essay, we studied a make-to-stock system with perishable inventory and impatient customers as a two-sided queue with abandonment from both sides. This model describes many consumer goods, where not only spoilage but also theft and damage can occur. Under certain assumptions on the magnitude of the abandonment rates and the scaled difference between the two arrival rates (products and customers), we suggested approximations to the system dynamics such as average

inventory, backorders, and fill rate via conventional heavy traffic limit theory.

We found that the limiting queue length is well approximated by a random variable best described as a weighted average of two complementary truncated normal random variables, where by "complementary" it is meant that they have non-overlapping support. We extended our results to analyze make-to-stock queues with/without perishability and limiting inventory space by inducing thresholds on the production (positive) side of the queue. Finally, we developed conjectures for the queue-length distribution for a non-Markovian system with general arrival streams and abandonment distributions. We took production rate as the decision variable to suggest near-optimal solutions. It can be interesting to explore different performance metrics to be used for other applications than make-to-stock queues. Moreover, obtaining structural results on the cost approximations that leads to closed form expressions for near-optimal solutions can give more insights.

Appendix A

Proofs of Results in Chapter 1

A.1 Proof of Theorem 1

A.1.1 Preliminary Results - Fluid Limit Theorem

The stochastic process $\{N_n(t), t \geq 0\}$ tracks, as a function of time, the number of broken units in the system, we consider the family of fluid scaled processes $\{\bar{\mathbf{N}}_n, n \geq 1\}$, where $\bar{\mathbf{N}}_n = \{\bar{\mathbf{N}}_n(t), t \geq 0\}$

$$\bar{\mathbf{N}}_n(t) \equiv \tfrac{N_n(t)}{n}, \quad t \geq 0, \tag{A.1}$$

for each $n \geq 0$. We denote the convergence in distribution for sequences of stochastic processes and random variables by $\Rightarrow$, and the term "fluid" is used because the limiting process of (A.1) above is continuous.

The following theorem shows that $\bar{\mathbf{N}}_n$ converges to a deterministic, continuous, monotone process $\bar{\mathbf{b}} = \{\bar{\mathbf{b}}(t), t \geq 0\}$ under fluid scaling. The limiting process $\bar{\mathbf{b}}$ can be used to approximate the transient, fluid-scaled broken-units process, and the limiting value of this process, $b = \lim_{t\to\infty} \bar{\mathbf{b}}(t)$, can be used to approximate the steady state average number of broken units.

Theorem 12. *Let $\bar{\mathbf{N}}_n(t)$ be the scaled number of broken units in the system, defined in (A.1). If $\bar{\mathbf{N}}_n(0) \Rightarrow \bar{\mathbf{b}}(0)$, where $\bar{\mathbf{b}}(0)$ is a deterministic number, then*

$$\bar{\mathbf{N}}_n \Rightarrow \bar{\mathbf{b}}, \quad t \geq 0 \quad as \quad n \to \infty, \tag{A.2}$$

where $\bar{\mathbf{b}}$ obeys the ordinary differential equation (ODE)

$$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - [\bar{\mathbf{b}}(t) - \bar{m}]^+) - \mu(\bar{\mathbf{b}}(t) \wedge \bar{s}), \tag{A.3}$$

and has the steady state values that depend on the decision pair $(\bar{s}, \bar{m})$ (note that repairmen regime is given before the spare one, and $\rho \equiv \frac{\lambda}{\mu}$):

$$b \equiv \lim_{t\to\infty} \bar{\mathbf{b}}(t)$$

$$= \begin{cases} 1 + \bar{m} - \bar{s}/\rho, & ED|ED, \quad \bar{s} < \rho, \quad \bar{m} > \bar{s} + \bar{s}/\rho - 1 \\ \bar{s} = 1 + \bar{m} - \bar{s}/\rho, & QED|ED, \quad \bar{s} < \rho, \quad \bar{m} = \bar{s} + \bar{s}/\rho - 1 \\ \frac{\rho(1+\bar{m})}{1+\rho}, & QD|ED, \quad \bar{s} > \frac{\rho(1+\bar{m})}{1+\rho}, \quad \bar{m} < \rho \\ \bar{\mathbf{b}}(0), & ED|QED, \quad \bar{s} < \bar{\mathbf{b}}(0) = \bar{m}, \quad \bar{s} = \rho, \quad \bar{m} > \rho \\ \rho, & QED|QED, \quad \bar{s} = \rho, \quad \bar{m} = \rho \\ \bar{m}, & QD|QED, \quad \bar{s} > \rho, \quad \bar{m} = \rho \\ \bar{\mathbf{b}}(0), & ED|QD, \quad \bar{s} < \bar{\mathbf{b}}(0) < \bar{m}, \quad \bar{s} = \rho, \quad \bar{m} > \rho \\ \bar{s}, & QED|QD, \quad \bar{s} = \bar{\mathbf{b}}(0) < \bar{m}, \quad \bar{s} = \rho, \quad \bar{m} > \rho \\ \rho, & QD|QD, \quad \bar{s} > \rho, \quad \bar{m} > \rho. \end{cases} \tag{A.4}$$

Remark 6. *Notice that ED|QED, ED|QD, and QED|QD capacity preferences imply to have more machines circulating the system than the repair capacity (and they are all placed on the same region in Figure 1.2 ($\bar{s} = \rho, \bar{m} > \rho$)). We emphasize that these cases are intuitively undesirable and indeed degenerate because the process will be stuck at the initial point $\bar{\mathbf{b}}(0)$.*

Proof of Theorem 12

We follow the framework of Browne and Whitt in [4] which summarizes a useful theorem of Stone in [36]. The fluid scaled process $\bar{\mathbf{N}}_n$ has the state space $\{0, 1/n, 2/n, \ldots, (n-1)/n, 1, (n+1)/n, \ldots, (n+m_n)/n\}$ and drift and diffusion functions

$$\bar{v}_n(x) = \left[\lambda_n(\lfloor nx \rfloor) - \mu_n(\lfloor nx \rfloor)\right]/n = \left[\lambda(n - [\lfloor nx \rfloor - m_n]^+) - \mu(\lfloor nx \rfloor \wedge s_n)\right]/n \tag{A.5}$$

$$\bar{\sigma}_n^2(x) = \left[\lambda_n(\lfloor nx \rfloor) + \mu_n(\lfloor nx \rfloor)\right]/n^2 = \left[\lambda(n - [\lfloor nx \rfloor - m_n]^+) + \mu(\lfloor nx \rfloor \wedge s_n)\right]/n^2, \tag{A.6}$$

respectively, where $\lfloor y \rfloor$ is the largest integer no greater than y. According to the Stone's result, when "appropriate" limits for the sequence of scaled (and centered) drift and diffusion functions are provided, one can obtain a diffusion limit for the scaled (and centered) sequence of processes, the parameters of which are the limiting drift and diffusion functions. Now we take the limit of the sequences in (A.5) and (A.6), and by Stone's result, we obtain the infinitesimal mean and variance of our limiting diffusion:

$$\bar{v}(x) \equiv \lim_{n\to\infty} \bar{v}_n(x) = \lambda(1 - [x - \bar{m}]^+) - \mu(x \wedge \bar{s}) \tag{A.7}$$

and

$$\bar{\sigma}^2(x) \equiv \lim_{n\to\infty} \bar{\sigma}_n^2(x) = 0.$$

Notice, the limiting infinitesimal variance is zero, implying that the limiting diffusion is degenerate. Moreover, given the deterministic initial point, the path of $\bar{\mathbf{b}}$ is fully determined. The drift of the limiting process, (A.7), coincides with the *ODE* in (A.3).

1. $\bar{s} < \bar{m}$	2. $\bar{s} = \bar{m}$	3. $\bar{m} < \bar{s}$
1.1 $\bar{\mathbf{b}}(0) < \bar{s} < \bar{m}$	2.1 $\bar{\mathbf{b}}(0) < \bar{s} = \bar{m}$	3.1 $\bar{\mathbf{b}}(0) < \bar{m} < \bar{s}$
1.2 $\bar{\mathbf{b}}(0) = \bar{s} < \bar{m}$	2.2 $\bar{\mathbf{b}}(0) = \bar{s} = \bar{m}$	3.2 $\bar{\mathbf{b}}(0) = \bar{m} < \bar{s}$
1.3 $\bar{s} < \bar{\mathbf{b}}(0) < \bar{m}$	2.3 $\bar{s} = \bar{m} < \bar{\mathbf{b}}(0)$	3.3 $\bar{m} < \bar{\mathbf{b}}(0) < \bar{s}$
1.4 $\bar{s} < \bar{\mathbf{b}}(0) = \bar{m}$		3.4 $\bar{m} < \bar{\mathbf{b}}(0) = \bar{s}$
1.5 $\bar{s} < \bar{m} < \bar{\mathbf{b}}(0)$		3.5 $\bar{m} < \bar{s} < \bar{\mathbf{b}}(0)$

Table A.1: Possible Cases for the Comparison of the Initial Point of the Fluid Process with Capacity Parameters

	Drift
Trajectory 1	$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda - \mu\bar{\mathbf{b}}(t)$
Trajectory 2	$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda - \mu\bar{s}$
Trajectory 3	$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - \bar{\mathbf{b}}(t) + \bar{m}) - \mu\bar{s}$
Trajectory 4	$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - \bar{\mathbf{b}}(t) + \bar{m}) - \mu\bar{\mathbf{b}}(t)$
	Solution
Trajectory 1	$\bar{\mathbf{b}}(t) = \rho + (\bar{\mathbf{b}}(0) - \rho)e^{-\mu t}$
Trajectory 2	$\bar{\mathbf{b}}(t) = (\lambda - \mu\bar{s})t + \bar{\mathbf{b}}(0)$
Trajectory 3	$\bar{\mathbf{b}}(t) = 1 + \bar{m} - \bar{s}/\rho + \left(\bar{\mathbf{b}}(0) - (1 + \bar{m} - \bar{s}/\rho)\right) e^{-\lambda t}$
Trajectory 4	$\bar{\mathbf{b}}(t) = \frac{\rho(1+\bar{m})}{1+\rho} + \left(\bar{\mathbf{b}}(0) - \frac{\rho(1+\bar{m})}{1+\rho}\right) e^{-(\lambda+\mu)t}$

Table A.2: Possible Drift and Solutions for the Limiting Fluid Process based on the Starting Point

Table A.1 represents all possible cases of values of $\bar{\mathbf{b}}(0)$ relative to $\bar{s}$ and $\bar{m}$. These comparisons dictate the behavior of $\bar{\mathbf{b}}(t)$ after time 0. The comparisons in Table A.1 above might lead to four possible starting solutions for the ODE in (A.3), which are listed in Table A.2 for ease of reference:

Case 1.1: Note that in order to obtain an explicit expression of (A.3) we have to compare $\bar{\mathbf{b}}(0)$ with $\bar{s}$ and $\bar{m}$. Each case number refers to the elements of Table A.1 and hence is matched with one of the trajectories in Table A.2. We have Trajectory 1 here. Since $\lim_{t\to\infty} \bar{\mathbf{b}}(t) = \rho$, if $\bar{\mathbf{b}}(0) < \rho$, $\bar{\mathbf{b}}(t)$ will increase towards ρ; else if $\bar{\mathbf{b}}(0) = \rho$, $\bar{\mathbf{b}}(t)$ will stay at ρ, otherwise $\bar{\mathbf{b}}(t)$ will decrease towards ρ. Having $\bar{\mathbf{b}}(0) < \bar{s} < \bar{m}$ in this case calls for a comparison between ρ and $\bar{s}$.

Case 1.1.1, $\rho < \bar{\mathbf{s}}$**:** The formula for the drift will stay the same until the equilibrium is reached because the monotone drift will not force the process to reach beyond $\bar{s}$; and hence the solution will stay the same. Thus, $b = \lim_{t\to\infty} \bar{\mathbf{b}}(t) = \rho$. Note that we have $b < \bar{s} < \bar{m}$, so this case represents a commodious capacity level choice where both the limiting scaled repairmen and spare machines resources outweigh the limiting scaled value of broken machines in need of them. Hence, the capacity level choices $\rho < \bar{s} < \bar{m}$ represents the QD|QD approach.

Case 1.1.2, $\rho \geq \bar{\mathbf{s}}$**:** Since the $\bar{\mathbf{b}}(t)$ will increase towards $\bar{s}$ in this case, there will be a time t_0 such that $\bar{\mathbf{b}}(t_0) = \bar{s}$ and the analysis to obtain $\lim_{t\to\infty} \bar{\mathbf{b}}(t)$ will follow from Case 1.2 by substituting $\bar{\mathbf{b}}(t_0)$ with $\bar{\mathbf{b}}(0)$ and hence treating time t_0 as the new starting point. This type of connection between the cases will be denoted by $\longrightarrow$; thus, Case 1.1.2 $\longrightarrow$ Case 1.2.

Case 1.2: We have Trajectory 2 here, leading to the following subcases.

Case 1.2.1, $\rho < \bar{\mathbf{s}}$**:** Since the drift is negative, the $\bar{\mathbf{b}}(t)$ process will decrease from the starting point $\bar{s}$; hence Case 1.2.1 $\longrightarrow$ Case 1.1.

Case 1.2.2, $\rho = \bar{\mathbf{s}}$**:** With 0 drift we have $b = \bar{\mathbf{b}}(0) = \bar{s} = \rho < \bar{m}$. Using similar reasoning as in Case 1.1.1, we conclude QED|QD approach is taken here.

Case 1.2.3, $\rho > \bar{\mathbf{s}}$**:** The drift is positive, so Case 1.2.3 $\longrightarrow$ Case 1.3.

The other cases can be analyzed similarly, which is summarized in Table A.4. When capacity level approaches (where possible to finalize) and the regions leading to them (listed as subcases) are matched, one can draw Figure 1.2. □

A.1.2 Preliminary Results - Lemma 1

We will denote e^x with $exp\{x\}$, and $log_e(x)$ with $ln(x)$ below.

Lemma 1. *Let a_1, a_2, b_n^1, and b_n^2 be constants and b_n^1 and b_n^2 depend on n. For any positive real number ϵ, a function $f(n)$ is said to be $o(n^{-\epsilon})$ if $n^\epsilon f(n) \to 0$ as $n \to \infty$.*

Cases	Trj.	Subcases	Fluid Limit	Cap.
1.1	1	1.1.1 $\rho < \bar{s}$	$b = \rho$	QD|QD
		1.1.2 $\rho \geq \bar{s}$	$\longrightarrow$ Case 1.2	
1.2	2	1.2.1 $\rho < \bar{s}$	$\longrightarrow$ Case 1.1	
		1.2.2 $\rho = \bar{s}$	$b = \bar{\mathbf{b}}(0) = \bar{s} = \rho < \bar{m}$	QED|QD
		1.2.3 $\rho > \bar{s}$	$\longrightarrow$ Case 1.3	
1.3	2	1.3.1 $\rho < \bar{s}$	$\longrightarrow$ Case 1.2	
		1.3.2 $\rho = \bar{s}$	$\bar{s} < b = \bar{\mathbf{b}}(0) < \bar{m}$	ED|QD
		1.3.3 $\rho > \bar{s}$	$\longrightarrow$ Case 1.4	
1.4	2	1.4.1 $\rho < \bar{s}$	$\longrightarrow$ Case 1.3	
		1.4.2 $\rho = \bar{s}$	$\bar{s} < b = \bar{\mathbf{b}}(0) = \bar{m}$	ED|QED
		1.4.3 $\rho > \bar{s}$	$\longrightarrow$ Case 1.5	
1.5	3	1.5.1 $\rho < \bar{s}$	$\longrightarrow$ Case 1.4	
		1.5.2 $\rho = \bar{s}$	$\longrightarrow$ Case 1.4	
		1.5.3 $\rho > \bar{s}$	$\bar{s} < \bar{m} < b = 1 + \bar{m} - \bar{s}/\rho$	ED|ED
2.1	1	2.1.1 $\rho < \bar{s}$	$b = \rho$	QD|QD
		2.1.2 $\rho = \bar{s}$	$\longrightarrow$ Case 2.2	
		2.1.3 $\rho > \bar{s}$	$\longrightarrow$ Case 2.2	
2.2	2	2.2.1 $\rho < \bar{s}$	$\longrightarrow$ Case 2.1	
		2.2.2 $\rho = \bar{s}$	$b = \bar{\mathbf{b}}(0) = \bar{s} = \bar{m} = \rho$	QED|QED
		2.2.3 $\rho > \bar{s}$	$\longrightarrow$ Case 2.3	
2.3	3	2.3.1 $\rho < \bar{s}$	$\longrightarrow$ Case 2.2	
		2.3.2 $\rho = \bar{s}$	$\longrightarrow$ Case 2.2	
		2.3.3 $\rho > \bar{s}$	$\bar{m} = \bar{s} < b = 1 + \bar{m} - \bar{s}/\rho$	ED|ED
3.1	1	3.1.1 $\rho < \bar{m}$	$b = \rho$	QD|QD
		3.1.2 $\rho = \bar{m}$	$\longrightarrow$ Case 3.2	
		3.1.3 $\rho > \bar{m}$	$\longrightarrow$ Case 3.2	
3.2	1	3.2.1 $\rho < \bar{m}$	$\longrightarrow$ Case 3.1	
		3.2.2 $\rho = \bar{m}$	$b = \bar{\mathbf{b}}(0) = \bar{m} = \rho < \bar{s}$	QD|QED
		3.2.3 $\rho > \bar{m}$	$\longrightarrow$ Case 3.3	
3.3	4	3.3.1	$(\rho < \bar{m}) \Rightarrow (\frac{\rho(1+\bar{m})}{1+\rho} < \bar{m})$; $\longrightarrow$ Case 3.2	
		3.3.2	$(\rho = \bar{m}) \Rightarrow (\frac{\rho(1+\bar{m})}{1+\rho} = \bar{m})$; $\longrightarrow$ Case 3.2	
		3.3.3 $\bar{m} < \frac{\rho(1+\bar{m})}{1+\rho} < \rho, \bar{s}$	$b = \frac{\rho(1+\bar{m})}{1+\rho}$	QD|ED
		3.3.4 $\bar{m} < \bar{s} = \frac{\rho(1+\bar{m})}{1+\rho} < \rho$	$\longrightarrow$ Case 3.4	
		3.3.5 $\bar{m}, \bar{s} < \frac{\rho(1+\bar{m})}{1+\rho} < \rho$	$\longrightarrow$ Case 3.4	

Table A.3: Fluid Limit Analysis, cases taken from Table A.1

Cases	Trj.	Subcases	Fluid Limit	Cap.
3.4	3	3.4.1 $\rho < \bar{s}$	$\longrightarrow$ Case 3.3	
		3.4.2 $\rho = \bar{s}$	$\longrightarrow$ Case 3.3	
		3.4.3 $\bar{m} < 1 + \bar{m} - \bar{s}/\rho < \bar{s} < \rho$	$\longrightarrow$ Case 3.3	
		3.4.4 $\bar{m} < 1 + \bar{m} - \bar{s}/\rho = \bar{s} < \rho$	$b = 1 + \bar{m} - \bar{s}/\rho$	QED|ED
		3.4.5 $\bar{s} < 1 + \bar{m} - \bar{s}/\rho, \rho$	$\longrightarrow$ Case 3.5	
3.5	3	3.5.1 $\rho < \bar{s}$	$\longrightarrow$ Case 3.4	
		3.5.2 $\rho = \bar{s}$	$\longrightarrow$ Case 3.4	
		3.5.3 $\bar{m} < 1 + \bar{m} - \bar{s}/\rho < \bar{s} < \rho$	$\longrightarrow$ Case 3.4	
		3.5.4 $\bar{m} < 1 + \bar{m} - \bar{s}/\rho = \bar{s} < \rho$	$\longrightarrow$ Case 3.4	
		3.5.5 $\bar{s} < 1 + \bar{m} - \bar{s}/\rho, \rho$	$b = 1 + \bar{m} - \bar{s}/\rho$	ED|ED

Table A.4: Continuation of Table A.4

Then, $\forall \epsilon$ *such that* $0 < \epsilon < \frac{1}{2}$ *we have*

$$ln\left[(a_1 n + b_n^1\sqrt{n})/(a_2 n + b_n^2\sqrt{n})\right] = ln\left(\frac{a_1}{a_2}\right) + o\left(n^{-\epsilon}\right); \tag{A.8}$$

hence, for any $O(n)$ *function* $g(n)$*,*

$$\begin{aligned}&\left[(a_1 n + b_n^1\sqrt{n})/(a_2 n + b_n^2\sqrt{n})\right]^{g(n)} = exp\left\{g(n) ln\left[(a_1 n + b_n^1\sqrt{n})/(a_2 n + b_n^2\sqrt{n})\right]\right\}\\ &= exp\left\{g(n)\left(ln\left(\frac{a_1}{a_2}\right) + o\left(n^{-\epsilon}\right)\right)\right\} = exp\left\{g(n) ln\left(\frac{a_1}{a_2}\right) + o\left(n^{1-\epsilon}\right)\right\}.\end{aligned} \tag{A.9}$$

Proof of Lemma 1

We have

$$\begin{aligned}&(a_1 n + b_n^1\sqrt{n})/(a_2 n + b_n^2\sqrt{n}) = (a_1 a_2 n + a_2 b_n^1\sqrt{n})/(a_2(a_2 n + b_n^2\sqrt{n}))\\ &= (a_1 a_2 n + a_2 b_n^1\sqrt{n} + a_1 b_n^2\sqrt{n} - a_1 b_n^2\sqrt{n})/(a_2(a_2 n + b_n^2\sqrt{n}))\\ &= a_1/a_2 + ((a_2 b_n^1 - a_1 b_n^2)\sqrt{n})/(a_2(a_2 n + b_n^2\sqrt{n})).\end{aligned} \tag{A.10}$$

Using Taylor's approximation for $f(y) = ln(y)$ at the point $y_0 = 1$ gives $ln(y) = (y-1) - \frac{(y-1)^2}{2} + \frac{(y-1)^3}{3} - \frac{(y-1)^4}{4} + \ldots$, and hence

$$\begin{aligned} & ln\left((a_1 n + b_n^1\sqrt{n})/(a_2 n + b_n^2\sqrt{n})\right) \\ & = ln\left\{a_1/a_2 + [(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})]\right\} \\ & = \sum_{k=1}^{\infty}\left\{(a_1/a_2 + [(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})] - 1)^k (-1)^{k-1}\right\}/k. \end{aligned} \tag{A.11}$$

Since

$$\begin{aligned} & (a_1/a_2 + [(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})] - 1)^k \\ & = \sum_{m=0}^{k}\binom{k}{m}(a_1/a_2 + [(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})])^m (-1)^{k-m} \\ & = \sum_{m=1}^{k}\binom{k}{m}\left[(a_1/a_2)^m + \sum_{j=0}^{m-1}\binom{m}{j}(a_1/a_2)^j ([(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})])^{m-j}\right] \\ & \quad \cdot(-1)^{k-m} + (-1)^k \\ & = \sum_{m=0}^{k}\binom{k}{m}(a_1/a_2)^m (-1)^{k-m} \\ & + \sum_{m=1}^{k}\sum_{j=0}^{m-1}\binom{k}{m}\binom{m}{j}(a_1/a_2)^j ([(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})])^{m-j}(-1)^{k-m} \\ & = \left(\frac{a_1}{a_2} - 1\right)^k + o\left(n^{-\epsilon}\right), \end{aligned} \tag{A.12}$$

we have

$$
\begin{aligned}
&ln\left((a_1 n + b_n^1\sqrt{n})/(a_2 n + b_n^2\sqrt{n})\right) \\
&= ln\left(a_1/a_2 + [(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})]\right) \\
&= \sum_{k=1}^{\infty}\left((a_1/a_2 + [(a_2 b_n^1 - a_1 b_n^2)\sqrt{n}]/[a_2(a_2 n + b_n^2\sqrt{n})] - 1)^k (-1)^{k-1}\right)/k \\
&= \sum_{k=1}^{\infty}\left(\left(\left(\tfrac{a_1}{a_2} - 1\right)^k + o\left(n^{-\epsilon}\right)\right)(-1)^{k-1}\right)/k \\
&= \sum_{k=1}^{\infty}\left(\left(\tfrac{a_1}{a_2} - 1\right)^k (-1)^{k-1}\right)/k + o\left(n^{-\epsilon}\right) = ln\left(\tfrac{a_1}{a_2}\right) + o\left(n^{-\epsilon}\right).\square
\end{aligned}
\tag{A.13}
$$

A.1.3 Preliminary Results - Expressing Mean Delay by Conditioning on the Broken Machines Process

Since the system state we follow is the broken machines process we would like to express $E[D_n^{rpl}(s_n, m_n)]$ in terms of this process. Therefore, we compute $E[D_n^{rpl}(s_n, m_n)]$ by conditioning on $\{N_n(t), t \geq 0\}$ at the time of the breakdown. Let τ_b denote the time of the breakdown and note that we assume the system is in steady state. Then,

$$
E[D_n^{rpl}(s_n, m_n)] = \sum_{k=0}^{n+m_n} E[D_n^{rpl}(s_n, m_n) \mid N_n^{st}(\tau_b) = k] P(N_n^{st}(\tau_b) = k). \tag{A.14}
$$

Notice that since we have a closed network, we cannot have $P(N_n^{st}(\tau_b) = k) = P(N_n^{st} = k)$. Instead, let $\{B_n(t), t \geq 0\}$ be the counting process for breakdowns. Then, for all $k = 0, \ldots, n + m_n$,

$$
P(N_n^{st}(\tau_b) = k) = \lim_{t\to\infty} \frac{\int_0^t 1_{(N_n^{st}(s)=k)} dB_n(s)}{B_n(t)} \tag{A.15}
$$

$\{B_n(t), t \geq 0\}$ is a Poisson process with arrival (breakdown) rate proportional to the number of units in use. Let $\{\widehat{M}_n, n \geq 1\}$ and $\{\widetilde{M}_n, n \geq 1\}$ be two indepen-

dent rate 1 Poisson processes, it follows that we have the following distributional equivalences (denoted by $=^d$).

$$\int_0^t 1_{(N_n^{st}(s)=k)} dB_n(s) =^d \widetilde{M}_n \left(\lambda \int_0^t (n - [N_n^{st}(s) - m_n]^+) 1_{(N_n^{st}(s)=k)} ds \right) \tag{A.16}$$

and

$$B_n(t) =^d \widehat{M}_n \left(\lambda \int_0^t (n - [N_n^{st}(s) - m_n]^+) ds \right). \tag{A.17}$$

All states in $\{N_n(t), t \geq 0\}$ are positive recurrent with steady state probabilities shown in (1.1). Thus, every state in $\{0, 1, 2, \ldots n + m_n - 1, n + m_n\}$ can be observed infinitely often, and therefore the birth(breakdown) rate would be positive infinitely often. It follows that the integrals in (A.16) and (A.17) will go to infinity as $t \to \infty$. Using this with $\frac{\widetilde{M}_n(t)}{t} \to 1$ $\left(\frac{\widehat{M}_n(t)}{t} \to 1 \right)$ almost surely as $t \to \infty$ for a rate 1 Poisson process from elementary Markov chain results (see, e.g. Resnick(1998), Section 7.5.1), we can express (A.15) as

$$P(N_n^{st}(\tau_b) = k) = \lim_{t\to\infty} \frac{\widetilde{M}_n \left(\lambda \int_0^t (n - [N_n^{st}(s) - m_n]^+) 1_{(N_n^{st}(s)=k)} ds \right)}{\widehat{M}_n \left(\lambda \int_0^t (n - [N_n^{st}(s) - m_n]^+) ds \right)}$$

$$= \lim_{t\to\infty} \frac{\widetilde{M}_n \left(\int_0^t (n - [N_n^{st}(s) - m_n]^+) 1_{(N_n^{st}(s)=k)} ds \right) \Big/ \left(\int_0^t (n - [N_n^{st}(s) - m_n]^+) 1_{(N_n^{st}(s)=k)} ds \right)}{\widehat{M}_n \left(\int_0^t (n - [N_n^{st}(s) - m_n]^+) ds \right) \Big/ \left(\int_0^t (n - [N_n^{st}(s) - m_n]^+) ds \right)}$$

$$\cdot \lim_{t\to\infty} \frac{\left(\int_0^t (n - [N_n^{st}(s) - m_n]^+) 1_{(N_n^{st}(s)=k)} ds \right) \Big/ t}{\left(\int_0^t (n - [N_n^{st}(s) - m_n]^+) ds \right) \Big/ t} = \frac{E[(n - [N_n^{st} - m_n]^+) 1_{(N_n^{st}=k)}]}{E[(n - [N_n^{st} - m_n]^+)]} = \frac{(n - [k - m_n]^+) p_{n,k}^{st}}{\sum_k (n - [k - m_n]^+) p_{n,k}^{st}} \tag{A.18}$$

where $p_{n,k}$ can be found in (1.1). $\square$

A.1.4 Proof of Formulas for $\lim_{\mathbf{n}\to\infty} \mathbf{E}[\mathbf{D_n^{rpl}}(\mathbf{s_n}, \mathbf{m_n})]$

Proof of Formulas for $\lim_{\mathbf{n}\to\infty} \mathbf{E}[\mathbf{D_n^{rpl}}(\mathbf{s_n}, \mathbf{m_n})]$ *for ED|ED*

Proof of Formulas for $\lim_{\mathbf{n}\to\infty} \mathbf{E}[\mathbf{D_n^{rpl}}(\mathbf{s_n}, \mathbf{m_n})]$ *for ED|ED,* $\mathbf{m_n} \geq \mathbf{s_n}$ First, we give some preliminary results and state some properties used. We recommend checking them as they are referred to in the proof.

Preliminaries-ED|ED-1

a.

$$a_n \equiv n\frac{(n\rho)^k}{k!}, \quad b_n \equiv n\frac{n^k}{s_n!}s_n^{s_n-k}\rho^k, \quad c_n \equiv (n+m_n-k)\frac{n^{m_n}}{s_n!}\frac{n!}{(n+m_n-k)!}s_n^{s_n-k}\rho^k. \tag{A.19}$$

b. Let ϵ be as in Lemma 1 and $\lfloor x \rfloor$ and $\lceil x \rceil$ represent the biggest integer smaller than x and the smallest integer larger than x, respectively. Then, $\frac{1}{2} < 1-\epsilon < 1$. We know that $\bar{s} < \rho$ for ED|ED from Theorem 12. Thus, for sufficiently large n, we have

$$m_n \leq \lfloor n\,(1+\bar{m}-\bar{s}/\rho) - n^{1-\epsilon} \rfloor - 1 \leq \lceil n\,(1+\bar{m}-\bar{s}/\rho) + n^{1-\epsilon} \rceil + 1 \leq n + m_n \tag{A.20}$$

c.

$$A_n \equiv \left[\sum_{k=m_n}^{\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor-1} ((k+1-m_n)/\mu s_n)\, c_n\right] / \left[\sum_{k=0}^{s_n-1} a_n + \sum_{k=s_n}^{m_n-1} b_n + \sum_{k=m_n}^{n+m_n} c_n\right],$$

$$B_n \equiv \left[\sum_{k=\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor}^{\lceil n(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil} ((k+1-m_n)/\mu s_n)\, c_n\right] / \left[\sum_{k=0}^{s_n-1} a_n + \sum_{k=s_n}^{m_n-1} b_n + \sum_{k=m_n}^{n+m_n} c_n\right],$$

$$C_n \equiv \left[\sum_{k=\lceil n(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil+1}^{n+m_n-1} ((k+1-m_n)/\mu s_n)\, c_n\right] / \left[\sum_{k=0}^{s_n-1} a_n + \sum_{k=s_n}^{m_n-1} b_n + \sum_{k=m_n}^{n+m_n} c_n\right]. \tag{A.21}$$

d.

$$A'_n \equiv \left[\sum_{k=m_n}^{\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor-1} c_n\right] / \left[\sum_{k=0}^{s_n-1} a_n + \sum_{k=s_n}^{m_n-1} b_n + \sum_{k=m_n}^{n+m_n} c_n\right],$$

$$B'_n \equiv \left[\sum_{k=\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor}^{\lceil n(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil} c_n\right] / \left[\sum_{k=0}^{s_n-1} a_n + \sum_{k=s_n}^{m_n-1} b_n + \sum_{k=m_n}^{n+m_n} c_n\right], \tag{A.22}$$

$$C'_n \equiv \left[\sum_{k=\lceil n(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil+1}^{n+m_n-1} c_n\right] / \left[\sum_{k=0}^{s_n-1} a_n + \sum_{k=s_n}^{m_n-1} b_n + \sum_{k=m_n}^{n+m_n} c_n\right].$$

e.

Let $X_n \sim Poisson(\frac{s_n}{\rho})$ and *lb* and *ub* denote the lower and upper bound, respectively. Then, from a) above

$$\sum_{k=lb}^{ub} c_n = \sum_{k=lb}^{ub} (n+m_n-k)\frac{n^{m_n}}{s_n!}\frac{n!}{(n+m_n-k)!}s_n^{s_n-k}\rho^k$$

$$= \sum_{k=lb}^{ub} \frac{n^{m_n}}{s_n!}\frac{n!}{(n+m_n-k-1)!}s_n^{s_n-k}\rho^k \left(\frac{s_n}{\rho}\right)^{n+m_n-1}\left(\frac{\rho}{s_n}\right)^{n+m_n-1} exp\{-\tfrac{s_n}{\rho}\} \quad exp\{\tfrac{s_n}{\rho}\}$$

$$= \left(\frac{\rho}{s_n}\right)^{n+m_n-1}\frac{n^{m_n}}{s_n!} \quad n! \quad s_n^{s_n} \quad exp\{\tfrac{s_n}{\rho}\}\sum_{k=lb}^{ub}\frac{1}{(n+m_n-1-k)!}\left(\frac{s_n}{\rho}\right)^{n+m_n-1-k} exp\{-\tfrac{s_n}{\rho}\}$$

$$= \left(\frac{\rho}{s_n}\right)^{n+m_n-1}\frac{n^{m_n}}{s_n!} \quad n! \quad s_n^{s_n} \quad exp\{\tfrac{s_n}{\rho}\}$$

$$\cdot P(n+m_n-1-ub \le X_n \le n+m_n-1-lb) \tag{A.23}$$

Note that (referring to a) above) if $ub = n+m_n$, then $\sum_{k=lb}^{n+m_n} c_n = \sum_{k=lb}^{n+m_n-1} c_n$.

f. From (1.2) and $\bar{s} < \rho$ for ED|ED (Theorem 12),

$$\begin{aligned} i. \quad & \lim_{n\to\infty}\left(n+m_n-1-\lfloor n\left(1+\bar{m}-\bar{s}/\rho\right)-n^{1-\epsilon}\rfloor+1-s_n/\rho\right).\left(\sqrt{s_n/\rho}\right)^{-1} \\ & = \lim_{n\to\infty}\left(n+m_n-1-n\left(1+\bar{m}-\bar{s}/\rho\right)+n^{1-\epsilon}+1-s_n/\rho\right).\left(\sqrt{s_n/\rho}\right)^{-1} \\ & = \lim_{n\to\infty}\left(\hat{m}_n\sqrt{n}+n^{1-\epsilon}-\hat{s}_n\sqrt{n}/\rho\right).\left(\sqrt{(n\bar{s}+\hat{s}_n\sqrt{n})/\rho}\right)^{-1} \to +\infty, \end{aligned} \tag{A.24}$$

$$\begin{aligned} ii. \quad & \left(n-1-s_n/\rho\right).\left(\sqrt{s_n/\rho}\right)^{-1} \\ & = \left(n\left(1-\bar{s}/\rho\right)-1-\hat{s}_n\sqrt{n}/\rho\right).\left(\sqrt{(n\bar{s}+\hat{s}_n\sqrt{n})/\rho}\right)^{-1} \to +\infty, \end{aligned} \tag{A.25}$$

$$
\begin{aligned}
iii. \quad & \lim_{n\to\infty} \left(\lfloor n\left(1+\bar{m}-\bar{s}/\rho\right)-n^{1-\epsilon}\rfloor+1-m_n\right)/\mu s_n \\
&= \lim_{n\to\infty} \lceil \left(n\left(1+\bar{m}+\bar{s}/\rho\right)+n^{1-\epsilon}\rceil+1-m_n\right)/\mu s_n \qquad \text{(A.26)} \\
&= (1-\bar{s}/\rho)/(\mu\bar{s}) = 1/\mu\bar{s}-1/\lambda.
\end{aligned}
$$

Main Part of the Proof

Let $\{W_n(t), t \geq 0\}$ be the stochastic process representing the amount of time the system will provide a backup unit(spare machine) if a breakdown occurs at time t, even if it is impossible for a breakdown to occur (i.e. when $N_n(t) = n + m_n$). We denote the steady state of this process by W_n^{st} and emphasize its dependence on the capacity levels by referring to it as $W_n^{st}(s_n, m_n)$. Let $E[W_n^{st}(s_n, m_n)|N_n^{st}(\tau_b) = k]$ (see Section A.1.3) be the conditional expected delay for backup when a breakdown occurs at the system state k (k does not include the broken unit itself). Then ,

$$
\begin{aligned}
& E[D_n^{rpl}(s_n, m_n)] \\
&= \left(\sum_{k\geq m_n} E[W_n^{st}(s_n, m_n)|N_n^{st}(\tau_b) = k](n - [k - m_n]^+)p_{n,k}^{st} \right) \qquad \text{(A.27)} \\
&\quad / \left(\sum_k (n - [k - m_n]^+)p_{n,k}^{st} \right).
\end{aligned}
$$

Since $m_n \geq s_n$, when a unit breaks down and a replacement cannot be provided instantly (i.e. $N_n^{st} \geq m_n \geq s_n$), it implies all repairmen are busy at the time of the breakdown. Thus, even if we assume no breakdowns occur until the replacement comes (which is $N_n^{st} + 1 - m_n$ repair completions later), the number of working repairmen will fall down to m_n at most. Therefore, all repairmen will stay busy during the delay. Hence, from (1.1), (A.27), Preliminaries-ED|ED-1 a)-c), and the just explained implication, for sufficiently large n we can write

$$E[D_n^{rpl}(s_n, m_n)] = \left(\sum_{k=m_n}^{n+m_n} ((k+1-m_n)/\mu s_n)\, c_n\right) / \left(\sum_{k=0}^{s_n-1} a_n + \sum_{k=s_n}^{m_n-1} b_n + \sum_{k=m_n}^{n+m_n} c_n\right)$$

$$= A_n + B_n + C_n. \tag{A.28}$$

Let $X_n, Y_n \sim Poisson\left(\frac{s_n}{\rho}\right)$ and note that $0 < \epsilon < \frac{1}{2}$ as in Lemma 1. From Preliminaries-ED|ED-1 c)-e),

$$\begin{aligned}
&A_n \leq ((n + m_n - m_n)/\mu s_n)\, A'_n, \\
&A'_n = \left((\rho/s_n)^{n+m_n-1}\, n^{m_n}(s_n!)^{-1} \quad n! \quad s_n^{s_n} \quad exp\{s_n/\rho\}\right) \\
&.P\left(n + m_n - 1 - \lfloor n\left(1 + \bar{m} - \bar{s}/\rho\right) - n^{1-\epsilon}\rfloor + 1 \leq X_n \leq n-1\right) \\
&.\left[\sum_{k=0}^{s_n-1} n\tfrac{(n\rho)^k}{k!} + \sum_{k=s_n}^{m_n-1} n\tfrac{n^k}{s_n!} s_n^{s_n-k}\rho^k + (\rho/s_n)^{n+m_n-1}\, n^{m_n}(s_n!)^{-1}\right. \\
&\left.\cdot n! \quad s_n^{s_n} \quad exp\{s_n/\rho\} P\left(Y_n \leq n-1\right)\right]^{-1}.
\end{aligned} \tag{A.29}$$

Let $\left((\rho/s_n)^{n+m_n-1}\, n^{m_n}(s_n!)^{-1} \quad n! \quad s_n^{s_n} \quad exp\{s_n/\rho\}\right)$ be $factor_1$. Then,

$$\begin{aligned}
&A'_n = P\left(n + m_n - 1 - \lfloor n\left(1 + \bar{m} - \bar{s}/\rho\right) - n^{1-\epsilon}\rfloor + 1 \leq X_n \leq n-1\right) \\
&.\left[\left(\sum_{k=0}^{s_n-1} n\tfrac{(n\rho)^k}{k!}\right)/factor_1 + \left(\sum_{k=s_n}^{m_n-1} n\tfrac{n^k}{s_n!} s_n^{s_n-k}\rho^k\right)/factor_1 + P\left(Y_n \leq n-1\right)\right]^{-1}.
\end{aligned} \tag{A.30}$$

From (1.2), Preliminaries-ED|ED-1 f)i) and ii), and Central Limit Theorem (CLT), for $Z_n, \bar{Z}_n \sim Normal(0,1)$ we have

$$
\begin{aligned}
&P\left(n+m_n-1-\lfloor n\left(1+\bar{m}-\bar{s}/\rho\right)-n^{1-\epsilon}\rfloor+1\leq X_n\leq n-1\right)\\
&\rightarrow P(+\infty<Z_n<+\infty)=0\\
&P\left(Y_n\leq n-1\right)\rightarrow P(\bar{Z}_n<+\infty)=1.
\end{aligned}
\tag{A.31}
$$

From Theorem 12, for ED|ED, for sufficiently large n, we know that $n\rho > s_n - 1$. Hence, for $k = 0, 1, 2, \ldots, s_n - 2$

$$
\left(\tfrac{(n\rho)^{k+1}}{k+1!}\right)/\left(\tfrac{(n\rho)^{k}}{k!}\right)=\tfrac{n\rho}{k+1}\geq\tfrac{n\rho}{s_n-1}>1. \tag{A.32}
$$

Therefore, for $k = 0, 1, 2, \ldots, s_n - 2$, for suff. large n, $\frac{(n\rho)^k}{k!}$ is increasing in k, and

$$
\sum_{k=0}^{s_n-1} n\tfrac{(n\rho)^k}{k!}\leq n\,\tfrac{(n\rho)^{s_n-1}}{(s_n-1)!}s_n. \tag{A.33}
$$

Since,

$$
\begin{aligned}
&n\,\tfrac{(n\rho)^{s_n-1}}{(s_n-1)!}s_n/factor_1=n\quad\tfrac{n^{s_n-1}\rho^{s_n-1}}{(s_n-1)!}\quad s_n\left(\tfrac{s_n}{\rho}\right)^{n+m_n-1}\tfrac{s_n!}{n^{m_n}}\quad\tfrac{exp\{-s_n/\rho\}}{n!\;{s_n}^{s_n}}\\
&=\tfrac{exp\{-s_n/\rho\}}{n^{m_n-s_n}\quad n!}\quad s_n\left(\tfrac{s_n}{\rho}\right)^{n+m_n-s_n},
\end{aligned}
\tag{A.34}
$$

after applying Stirling's approximation, $n!\cong\sqrt{2\pi n}\,n^ne^{-n}$ to $n!$, we have (from (1.2) and Lemma 1 with $0<\epsilon<1/2$)

$$
\begin{aligned}
&\tfrac{exp\{-s_n/\rho\}}{n^{m_n-s_n}n!}s_n\left(\tfrac{s_n}{\rho}\right)^{n+m_n-s_n}\cong\tfrac{1}{\sqrt{2\pi}}\tfrac{s_n}{\sqrt{n}}\left(\tfrac{s_n}{n\rho}\right)^{n+m_n-s_n}exp\{n-\tfrac{s_n}{\rho}\}\\
&=\tfrac{1}{\sqrt{2\pi}}(\bar{s}\sqrt{n}+\hat{s}_n)exp\left\{n-\tfrac{s_n}{\rho}+(n+m_n-s_n)ln\left(\tfrac{\bar{s}}{\rho}\right)+o(n^{1-\epsilon})\right\}.
\end{aligned}
\tag{A.35}
$$

To obtain the limit of the last expression above we will use the following facts:

i) Taylor's approximation for $f(y) = -ln(y)$ at the point $y_0 = 1$ gives $-ln(y) = (1-y) + \frac{(1-y)^2}{2} + \frac{(1-y)^3}{3} + \frac{(1-y)^4}{4} + \ldots$, and we use it for $-ln\left(\frac{\bar{s}}{\rho}\right)$,

ii) $\bar{s} < \rho$ for ED|ED from Theorem 12 and hence $0 < 1 - \frac{\bar{s}}{\rho}$,

iii) $\bar{m} \geq \bar{s}$ (since $m_n \geq s_n$ in this case),

iv) $1 + \bar{m} - \bar{s} \geq 0$ since we assume $n + m_n > s_n$ (otherwise we will have ample repairmen which leads to trivial cases which are out of concern here). Note that at least one of the inequalities in iii) and iv) always have to be strict. Then, we get

$$
\begin{aligned}
&\tfrac{1}{\sqrt{2\pi}}(\bar{s}\sqrt{n} + \hat{s}_n) exp\left\{n - \tfrac{s_n}{\rho} + (n + m_n - s_n) ln\left(\tfrac{\bar{s}}{\rho}\right) + o(n^{1-\epsilon})\right\} \\
&= \tfrac{1}{\sqrt{2\pi}}(\bar{s}\sqrt{n} + \hat{s}_n) \\
&exp\left\{-n\left\{(1 + \bar{m} - \bar{s})\left[\left(1 - \tfrac{\bar{s}}{\rho}\right) + \tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2} + \tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3} + \ldots\right] + \tfrac{\bar{s}}{\rho} - 1\right\}\right\} \\
&.exp\left\{o(n^{1-\epsilon}) - \tfrac{\hat{s}_n\sqrt{n}}{\rho} + (\hat{m}_n - \hat{s}_n)\sqrt{n}\, ln\left(\tfrac{\bar{s}}{\rho}\right)\right\} \\
&= \tfrac{1}{\sqrt{2\pi}}(\bar{s}\sqrt{n} + \hat{s}_n) \\
&/exp\left\{n\left\{(\bar{m} - \bar{s})\left(1 - \tfrac{\bar{s}}{\rho}\right) + (1 + \bar{m} - \bar{s})\left[\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2} + \tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3} + \ldots\right]\right\} + o(n^{1-\epsilon})\right\} \\
&\to \tfrac{+\infty}{+\infty}.
\end{aligned}
\tag{A.36}
$$

L'Hospital's rule gives

$$\lim_{n\to\infty} \tfrac{1}{\sqrt{2\pi}}(\bar{s}\sqrt{n}+\hat{s}_n)$$

$$/\left[exp\left\{n\left\{(\bar{m}-\bar{s})\left(1-\tfrac{\bar{s}}{\rho}\right)+(1+\bar{m}-\bar{s})\left[\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2}+\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3}+\ldots\right]\right\}+o(n^{1-\epsilon})\right\}\right]$$

$$=\lim_{n\to\infty}\tfrac{1}{\sqrt{2\pi}}\bar{s}$$

$$/\left[2\sqrt{n}\left\{(\bar{m}-\bar{s})\left(1-\tfrac{\bar{s}}{\rho}\right)+(1+\bar{m}-\bar{s})\left[\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2}+\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3}+\ldots\right]+o(1)\right\}\right.$$

$$\left..exp\left\{n\left\{(\bar{m}-\bar{s})\left(1-\tfrac{\bar{s}}{\rho}\right)+(1+\bar{m}-\bar{s})\left[\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2}+\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3}+\ldots\right]\right\}+o(n^{1-\epsilon})\right\}\right]$$

$$=0. \tag{A.37}$$

Thus, from (A.33)-(A.37)

$$\begin{aligned}&\lim_{n\to\infty}\left(\sum_{k=0}^{s_n-1} n\tfrac{(n\rho)^k}{k!}\right)/factor_1 \le \lim_{n\to\infty}\left(n\,\tfrac{(n\rho)^{s_n-1}}{(s_n-1)!}s_n\right)/factor_1=0\\ &\Rightarrow \lim_{n\to\infty}\left(\sum_{k=0}^{s_n-1} n\tfrac{(n\rho)^k}{k!}\right)/factor_1=0.\end{aligned} \tag{A.38}$$

Similarly, from Theorem 12, for ED|ED, for sufficiently large n, we know that $n\rho > s_n$. Hence, for $k = s_n, s_n+1, s_n+2, \ldots, m_n-1$

$$\left(\tfrac{(n\rho)^{k+1}}{s_n!{s_n}^{k+1-s_n}}\right)/\left(\tfrac{(n\rho)^k}{s_n!{s_n}^{k-s_n}}\right)=\tfrac{n\rho}{s_n}>1. \tag{A.39}$$

Therefore, for $k = s_n, s_n+1, s_n+2, \ldots, m_n-1$, $\frac{(n\rho)^k}{s_n!{s_n}^{k-s_n}}$ is increasing in k, and

$$\sum_{k=s_n}^{m_n-1} n \quad \tfrac{n^k}{s_n!} \quad s_n^{s_n-k} \quad \rho^k \le n \quad \tfrac{n^{m_n-1}}{s_n!} \quad s_n^{s_n-m_n+1}\rho^{m_n-1}(m_n-s_n). \tag{A.40}$$

Since,

$$\begin{aligned}
&\left(n \quad \tfrac{n^{m_n-1}}{s_n!} \quad s_n^{s_n-m_n+1} \quad \rho^{m_n-1} \quad (m_n-s_n)\right)/factor_1 \\
&= n \quad \tfrac{n^{m_n-1}}{s_n!} \quad s_n^{s_n-m_n+1} \quad \rho^{m_n-1} \quad (m_n-s_n)\left(\tfrac{s_n}{\rho}\right)^{n+m_n-1} \tfrac{s_n!}{n^{m_n}} \quad (n!)^{-1} \quad s_n^{-s_n} \\
&\quad \cdot exp\left\{-\tfrac{s_n}{\rho}\right\} \\
&= (m_n-s_n)\left(\tfrac{s_n}{\rho}\right)^n \quad (n!)^{-1} exp\left\{-\tfrac{s_n}{\rho}\right\},
\end{aligned} \tag{A.41}$$

after applying Stirling's approximation to $n!$, we have

$$(m_n-s_n)\left(\tfrac{s_n}{\rho}\right)^n \quad (n!)^{-1} \quad exp\left\{-\tfrac{s_n}{\rho}\right\} \cong \tfrac{1}{\sqrt{2\pi}} \quad \tfrac{(m_n-s_n)}{\sqrt{n}} \quad \left(\tfrac{s_n}{n\rho}\right)^n \quad exp\left\{n-\tfrac{s_n}{\rho}\right\}. \tag{A.42}$$

Then, from (1.2), Lemma 1, $\bar{s} < \rho$ for ED|ED by Theorem 12 ($0 < \epsilon < \frac{1}{2}$), and the same facts used for (A.35) we get

$$\begin{aligned}
&\tfrac{1}{\sqrt{2\pi}} \tfrac{(m_n-s_n)}{\sqrt{n}} \left(\tfrac{s_n}{n\rho}\right)^n exp\left\{n-\tfrac{s_n}{\rho}\right\} \\
&= \tfrac{1}{\sqrt{2\pi}} \quad ((\bar{m}-\bar{s})\sqrt{n}+\hat{m}_n-\hat{s}_n) \quad exp\left\{n\, ln\left(\tfrac{\bar{s}}{\rho}\right)+o(n^{1-\epsilon})+n-\tfrac{s_n}{\rho}\right\} \\
&= \tfrac{1}{\sqrt{2\pi}} \quad ((\bar{m}-\bar{s})\sqrt{n}+\hat{m}_n-\hat{s}_n) \\
&/ \quad exp\left\{n\left\{\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2}+\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3}+\ldots\right\}+o\left(n^{1-\epsilon}\right)+\tfrac{\hat{s}_n\sqrt{n}}{\rho}\right\} \\
&\to \tfrac{+\infty}{+\infty}.
\end{aligned} \tag{A.43}$$

We apply L'Hospital's rule which gives

$$\lim_{n\to\infty} \tfrac{1}{\sqrt{2\pi}}((\bar{m}-\bar{s})\sqrt{n}+\hat{m}_n-\hat{s}_n)$$

$$/\left[exp\left\{n\left\{\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2}+\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3}+\ldots\right\}+o\left(n^{1-\epsilon}\right)+\tfrac{\hat{s}_n\sqrt{n}}{\rho}\right\}\right]$$

$$= \lim_{n\to\infty} \tfrac{1}{\sqrt{2\pi}}(\bar{m}-\bar{s}) \tag{A.44}$$

$$/\left[2\sqrt{n}\left\{\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2}+\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3}+\ldots+o(1)\right\}\right.$$

$$\left.exp\left\{n\left\{\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^2}{2}+\tfrac{\left(1-\frac{\bar{s}}{\rho}\right)^3}{3}+\ldots\right\}+o\left(n^{1-\epsilon}\right)\right\}\right]=0.$$

Thus, from (A.40)

$$\lim_{n\to\infty}\left(\sum_{k=s_n}^{m_n-1} n\tfrac{n^k}{s_n!}s_n^{s_n-k}\rho^k\right)/factor_1 \leq$$

$$\lim_{n\to\infty}\left(n\tfrac{n^{m_n-1}}{s_n!}s_n^{s_n-m_n+1}\rho^{m_n-1}(m_n-s_n)\right)/factor_1=0 \tag{A.45}$$

$$\Rightarrow \lim_{n\to\infty}\left(\sum_{k=s_n}^{m_n-1} n\tfrac{n^k}{s_n!}s_n^{s_n-k}\rho^k\right)/factor_1=0.$$

Thus, from (A.29)-(A.45)

$$A_n \leq ((n+m_n-m_n)/(\mu s_n))\, A'_n,$$

$$A'_n = P\left(n+m_n-1-\lfloor n\left(1+\bar{m}-\bar{s}/\rho\right)-n^{1-\epsilon}\rfloor+1\leq X_n\leq n-1\right)$$

$$\cdot\left[\left(\sum_{k=0}^{s_n-1} n\tfrac{(n\rho)^k}{k!}\right)/factor_1+\left(\sum_{k=s_n}^{m_n-1} n\tfrac{n^k}{s_n!}s_n^{s_n-k}\rho^k\right)/factor_1+P\left(Y_n\leq n-1\right)\right]$$

$$\to \tfrac{0}{0+0+1}=0$$

$$\Rightarrow \lim_{n\to\infty} A_n \leq \lim_{n\to\infty}\left(\tfrac{n+m_n-m_n}{\mu s_n}\right)A'_n=0.$$

(A.46)

Now we look at B_n. From (A.28) and Preliminaries-ED|ED-1 c)-e) and similar to the analysis made for A_n and A'_n (with $factor_1$ having the same value and $X_n, Y_n \sim Poisson\left(\frac{s_n}{\rho}\right)$),

$$\begin{aligned}
&((\lfloor n\,(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor+1-m_n)/(\mu\,s_n))\,B'_n \le B_n \\
&B_n \le ((\lceil n\,(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil+1-m_n)(\mu\,s_n))\,B'_n \\
&B'_n = P\,(n+m_n-1-\lceil n\,(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil \\
&\le X_n \le n+m_n-1-\lfloor n\,(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor) \\
&\cdot\left[\left(\sum_{k=0}^{s_n-1} n\frac{(n\rho)^k}{k!}\right)/factor_1+\left(\sum_{k=s_n}^{m_n-1} n\frac{n^k}{s_n!}s_n^{s_n-k}\rho^k\right)/factor_1+P\,(Y_n\le n-1)\right]
\end{aligned} \tag{A.47}$$

Similar to Preliminaries-ED|ED-1 f) i) for $Z_n \sim Normal(0,1)$,

$$\begin{aligned}
&P\,(n+m_n-1-\lceil n\,(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil \\
&\le X_n \le n+m_n-1-\lfloor n\,(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor) \\
&\to P(-\infty<Z_n<+\infty)=1
\end{aligned} \tag{A.48}$$

From (A.31)-(A.48) and Preliminaries-ED|ED-1 f) iii),

$$\begin{aligned}
&\lim_{n\to\infty}((\lfloor n\,(1+\bar{m}-\bar{s}/\rho)-n^{1-\epsilon}\rfloor+1-m_n)/(\mu\,s_n))\,B'_n \le \lim_{n\to\infty} B_n \\
&\lim_{n\to\infty} B_n \le \lim_{n\to\infty}((\lceil n\,(1+\bar{m}-\bar{s}/\rho)+n^{1-\epsilon}\rceil+1-m_n)/(\mu\,s_n))\,B'_n \\
&(1/\mu\bar{s}-1/\lambda)\cdot\tfrac{1}{0+0+1} \le \lim_{n\to\infty} B_n \le (1/\mu\bar{s}-1/\lambda)\cdot\tfrac{1}{0+0+1} \\
&\Rightarrow \lim_{n\to\infty} B_n = (1/\mu\bar{s}-1/\lambda)\,.
\end{aligned} \tag{A.49}$$

Now we look at C_n. From Preliminaries-ED|ED-1 c)-e) with $X_n, Y_n \sim Pois.\left(\frac{s_n}{\rho}\right)$ and the same value of $factor_1$,

$$C_n \leq ((n+m_n-m_n)/\mu s_n)\, C'_n,$$

$$C'_n = P\left(X_n \leq n+m_n-1-\lceil n\left(1+\bar{m}-\tfrac{\bar{s}}{\rho}\right)+n^{1-\epsilon}\rceil-1\right)$$

$$\cdot\left[\left(\sum_{k=0}^{s_n-1} n\tfrac{(n\rho)^k}{k!}\right)/factor_1+\left(\sum_{k=s_n}^{m_n-1} n\tfrac{n^k}{s_n!}s_n^{s_n-k}\rho^k\right)/factor_1+P\left(Y_n\leq n-1\right)\right] \tag{A.50}$$

Similarly, from Preliminaries-ED|ED-1 f)i) for $Z_n \sim Normal(0,1)$,

$$P\left(X_n \leq n+m_n-1-\lceil n\left(1+\bar{m}-\tfrac{\bar{s}}{\rho}\right)+n^{1-\epsilon}\rceil-1\right) \to P(Z_n<-\infty)=0 \tag{A.51}$$

Similar to A'_n, we have $\lim\limits_{n\to\infty} C'_n = 0$ and

$$\lim_{n\to\infty} C_n \leq \lim_{n\to\infty}\left(\tfrac{n+m_n-m_n}{\mu s_n}\right) C'_n = 0 \Rightarrow \lim_{n\to\infty} C_n = 0. \tag{A.52}$$

To sum up, for $m_n \geq s_n$ for ED|ED, (A.27)-(A.52) gives

$$\lim_{n\to\infty} E[D_n^{rpl}(s_n,m_n)]$$

$$= \lim_{n\to\infty}\left(\sum_{k\geq m_n}((k+1-m_n)/\mu s_n)\, c_n\right)/\left(\sum_{k=0}^{s_n-1} a_n+\sum_{k=s_n}^{m_n-1} b_n+\sum_{k=m_n}^{n+m_n} c_n\right) \tag{A.53}$$

$$= \lim_{n\to\infty}(A_n+B_n+C_n) = \left(\tfrac{1}{\mu\bar{s}}-\tfrac{1}{\lambda}\right).\square$$

Proof of Formulas for $\lim\limits_{\mathbf{n}\to\infty} \mathbf{E}[\mathbf{D_n^{rpl}}(\mathbf{s_n},\mathbf{m_n})]$ *for ED|ED,* $\mathbf{s_n} > \mathbf{m_n}$ We also have some preliminary results here. Again, we recommend referring to them as they are referred to in the proof.

Preliminaries-ED|ED-2

a.

$$
\begin{aligned}
&d_n \equiv (n+m_n-k)\tfrac{n^{m_n}}{k!}\tfrac{n!}{(n+m_n-k)!}\rho^k,\\
&\tilde{A_n} \equiv \left[\sum_{k=m_n}^{s_n-1} E[W_n^{st}(s_n,m_n)|N_n^{st}(\tau_b)=k]d_n\right] / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right],\\
&\tilde{B_n} \equiv \left[\sum_{k=s_n}^{n+m_n} E[W_n^{st}(s_n,m_n)|N_n^{st}(\tau_b)=k]c_n\right] / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right].
\end{aligned}
\tag{A.54}
$$

b.

$$
\begin{aligned}
\tilde{A_n}' &\equiv \left[\sum_{k=m_n}^{s_n-1} d_n\right] / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right],\\
\tilde{B_n}' &\equiv \left[\sum_{k=s_n}^{n+m_n} c_n\right] / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right].
\end{aligned}
\tag{A.55}
$$

c.

$$
\begin{aligned}
&\sum_{k=m_n}^{s_n-1} d_n = \sum_{k=m_n}^{s_n-1} \tfrac{n^{m_n}}{k!}\tfrac{n!}{(n+m_n-k-1)!}\rho^k\\
&= \tfrac{n^{m_n}\ n!}{(n+m_n-1)!} \quad (1+\rho)^{n+m_n-1} \quad \sum_{k=m_n}^{s_n-1} \tbinom{n+m_n-1}{k}\left(\tfrac{\rho}{1+\rho}\right)^k\left(\tfrac{1}{1+\rho}\right)^{n+m_n-1-k} \qquad \text{(A.56)}\\
&= \tfrac{n^{m_n}\ n!}{(n+m_n-1)!} \quad (1+\rho)^{n+m_n-1} \quad P(m_n \le Y_n \le s_n-1)
\end{aligned}
$$

where $Y_n \sim Binomial\left(n+m_n-1, \tfrac{\rho}{1+\rho}\right)$.

d. From Theorem 12, we know that $\bar{s} < \rho$ for ED|ED, and since $s_n > m_n$ for this case, we have $\bar{m} \le \bar{s} < \rho$. Therefore, from (1.2)

$$\left(m_n - \tfrac{\rho}{1+\rho}(n+m_n-1)\right) / \left(\sqrt{(n+m_n-1)\tfrac{\rho}{(1+\rho)^2}}\right)$$
$$= \{(1+\rho)\bar{m}n + (1+\rho)\hat{m}_n\sqrt{n} - \rho[(1+\bar{m})n + \hat{m}_n\sqrt{n} - 1]\} / \left(\sqrt{(n+m_n-1)\rho}\right)$$
$$\to -\infty. \tag{A.57}$$

Moreover, again from Theorem 12 we have $\bar{s} + \bar{s}/\rho - 1 < \bar{m}$ for ED|ED, hence $\bar{s}\rho + \bar{s} - \rho - \bar{m}\rho < 0$, so

$$\left(s_n - 1 - \tfrac{\rho}{1+\rho}(n+m_n-1)\right) / \left(\sqrt{(n+m_n-1)\tfrac{\rho}{(1+\rho)^2}}\right)$$
$$= [(1+\rho)(\bar{s}n + \hat{s}_n\sqrt{n} - 1) - \rho((1+\bar{m})n + \hat{m}_n\sqrt{n} - 1)] / \left(\sqrt{(n+m_n-1)\rho}\right)$$
$$\to -\infty. \tag{A.58}$$

e.

$$\begin{aligned} &\left\{(1+\bar{m}) \quad ln\left(\tfrac{(1+\bar{m})\rho}{\bar{s}(1+\rho)}\right) + \bar{s} + \tfrac{\bar{s}}{\rho} - (1+\bar{m})\right\} \\ &= \tfrac{\bar{s}\ (1+\rho)}{\rho}\left[\tfrac{(1+\bar{m})\rho}{\bar{s}(1+\rho)} \quad ln\left(\tfrac{(1+\bar{m})\rho}{\bar{s}(1+\rho)}\right) + \left(1 - \tfrac{(1+\bar{m})\rho}{\bar{s}(1+\rho)}\right)\right]. \end{aligned} \tag{A.59}$$

f.

Since, $xlnx - x + 1|_{x=1} = 0$ and $(xlnx - x + 1)' = lnx > 0$ for $x > 1$, $xlnx - x + 1 > 0$ for $x > 1$.

g.

Since $1 + \bar{m} - \bar{s} - \frac{\bar{s}}{\rho} > 0$ for ED|ED from Theorem 12,

$$\left(n + m_n - 1 - s_n - \frac{s_n}{\rho}\right) / \sqrt{\frac{s_n}{\rho}}$$
$$= \left[n\left(1 + \bar{m} - \bar{s} - \frac{\bar{s}}{\rho}\right) + \left(\hat{m}_n - \hat{s}_n - \frac{\hat{s}_n}{\rho}\right)\sqrt{n} - 1\right] / \sqrt{\frac{s_n}{\rho}} \to +\infty. \tag{A.60}$$

h.

Let τ_b be the random time of breakdown of a random unit such that a replacement cannot be provided immediately and a delay will be incurred, and let $W_n^{st,\tau_b}(s_n, m_n)$ denote the delay to provide backup after that breakdown. Then, for an arbitrary $\varepsilon_1 > 0$, let the event $F_n = \{min_{\,t\in(\tau_b,\tau_b+W_n^{st,\tau_b}(s_n,m_n))]} N_n^{st}(t) < s_n - \varepsilon_1.n\}$ and F_n' be its complement. Note that $N_n^{st}(t)$ emphasizes that we are considering the process $\{N_n(t), t \geq 0\}$ after it reaches steady state. Then, we define

$$\tilde{B}_n^1 \equiv \left[\sum_{k=s_n}^{n+m_n} E[W_n^{st,\tau_b}(s_n, m_n) \cdot 1_{F_n} \mid N_n^{st}(\tau_b) = k] \quad c_n\right]$$
$$/ \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right],$$
$$\tilde{B_n}^2 \equiv \left[\sum_{k=s_n}^{n+m_n} E[W_n^{st,\tau_b}(s_n, m_n) \cdot 1_{F_n'} \mid N_n^{st}(\tau_b) = k] \quad c_n\right] \tag{A.61}$$
$$/ \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right].$$

i.

Let $\varepsilon_1, \hat{\varepsilon}_1 > 0$ and $\varepsilon_1 > \hat{\varepsilon}_1$. For $k = s_n, s_n + 1, \ldots, n + m_n$ and for sufficiently large n, $\frac{k}{n} > \bar{s} - \hat{\varepsilon}_1$. Then, for $N_n^{st}(\tau_b) = k$,

$$P\left(min_{\,t\in(\tau_b,\infty)} N_n^{st}(t) < s_n - \varepsilon_1.n \mid N_n^{st}(\tau_b) = k\right)$$

$$= P\left(min_{\,t\in(\tau_b,\infty)}\frac{N_n^{st}(t)}{n} < \frac{s_n}{n} - \varepsilon_1 \;\mid \frac{N_n^{st}(\tau_b)}{n} = \frac{k}{n}\right).$$

From Theorem 12, we know that $\bar{\mathbf{N}}_{\mathbf{n}}(t)$ in (A.1) converges to a deterministic process $\bar{\mathbf{b}}(t)$ if the starting point is deterministic. Moreover, $b = \lim_{t\to\infty}\bar{\mathbf{b}}(\mathbf{t}) > \bar{s}$ for ED|ED. Here we have $\lim_{n\to\infty}\frac{N_n^{st}(\tau_b)}{n} = \lim_{n\to\infty}\frac{k}{n}$ acting as the deterministic starting point (since its value is known when the observation starts), so we have for $k = s_n, s_n+1, \ldots, n+m_n$

$$\begin{aligned}
&P\left(min_{\,t\in(\tau_b,\infty)}\tfrac{N_n^{st}(t)}{n} < \tfrac{s_n}{n} - \varepsilon_1 \;\mid \tfrac{N_n^{st}(\tau_b)}{n} = \tfrac{k}{n}\right)\\
&= P\left(min_{\,t\in(\tau_b,\infty)}\tfrac{N_n^{st}(t)}{n} - \bar{\mathbf{b}}(t) + \bar{\mathbf{b}}(t) < \tfrac{s_n}{n} - \varepsilon_1 \;\mid \tfrac{N_n^{st}(\tau_b)}{n} = \tfrac{k}{n}\right)\\
&\le P\left(min_{\,t\in(\tau_b,\infty)}\bar{\mathbf{b}}(t) - max_{\,t\in(\tau_b,\infty)}|\tfrac{N_n^{st}(t)}{n} - \bar{\mathbf{b}}(t)| < \tfrac{s_n}{n} - \varepsilon_1 \;\mid \tfrac{N_n^{st}(\tau_b)}{n} = \tfrac{k}{n}\right)\\
&= P\left(min_{\,t\in(\tau_b,\infty)}\bar{\mathbf{b}}(t) - \tfrac{s_n}{n} + \varepsilon_1 \;< max_{\,t\in(\tau_b,\infty)}|\tfrac{N_n^{st}(t)}{n} - \bar{\mathbf{b}}(t)| \mid \tfrac{N_n^{st}(\tau_b)}{n} = \tfrac{k}{n}\right)\\
&\to P\left(min_{\,t\in(\tau_b,\infty)}\bar{\mathbf{b}}(t) - \bar{s} + \varepsilon_1 \;< 0 \mid \bar{\mathbf{b}}(\tau_b) = \lim_{n\to\infty}\tfrac{k}{n}\right) = 0\\
&\quad\left(since \lim_{n\to\infty}\tfrac{k}{n} > \bar{s} - \hat{\varepsilon}_1\right)
\end{aligned} \tag{A.62}$$

j.

Since $\bar{s} + \bar{s}/\rho - 1 < \bar{m}$ for ED|ED from Theorem 12, for sufficiently large n we have

$$s_n \le \lfloor n\,(1+\bar{m}-\bar{s}/\rho) - n^{1-\varepsilon}\rfloor - 1 \le \lceil n\,(1+\bar{m}-\bar{s}/\rho) + n^{1-\varepsilon}\rceil + 1 \le n + m_n. \tag{A.63}$$

Main Part of the Proof

From (A.27), (1.1), Preliminaries-ED|ED-2 a), and Preliminaries-ED|ED-1 a), we have

$$
\begin{aligned}
&E[D_n^{rpl}(s_n, m_n)] \\
&= \left[\sum_{k=m_n}^{s_n-1} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k]\, d_n + \sum_{k=s_n}^{n+m_n} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k]\, c_n \right] \\
&\cdot \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n \right]^{-1} = \tilde{A}_n + \tilde{B}_n.
\end{aligned}
\tag{A.64}
$$

Since for a delay to occur, at least m_n units should be broken down and with $s_n > m_n$, at least m_n repairmen should be busy during the whole delay time. Thus, from Preliminaries-ED|ED-2 b) and similar reasoning used for (A.28)

$$
\tilde{A}_n \leq \left((n + m_n - m_n) / \mu m_n \right) \tilde{A}'_n. \tag{A.65}
$$

Let $n^{m_n} n! ((n+m_n-1)!)^{-1} (1+\rho)^{n+m_n-1}$ be $factor_2$. From Preliminaries-ED|ED-2 b) and c),

$$
\begin{aligned}
&\tilde{A}'_n = P(m_n \leq Y_n \leq s_n - 1) \qquad \left(Y_n \sim Binomial\left(n + m_n - 1, \tfrac{\rho}{1+\rho} \right) \right) \\
&\cdot \left[\left(\sum_{k=0}^{m_n-1} a_n \right) / factor_2 + P(m_n \leq Y_n \leq s_n - 1) + \left(\sum_{k=s_n}^{n+m_n} c_n \right) / factor_2 \right]^{-1}
\end{aligned}
\tag{A.66}
$$

From Preliminaries-ED|ED-2 c) and d) and CLT, we have

$$
P(m_n \leq Y_n \leq s_n - 1) \to P(-\infty < Z_n < -\infty) = 0 \tag{A.67}
$$

where $Z_n \sim Normal(0, 1)$.

From:

i) Preliminaries-ED|ED-1 e),

ii) Stirling's approximation applied to $(n+m_n)!$ and $s_n!$,

iii) (1.2) and Lemma 1 ($0<\epsilon<\frac{1}{2}$),

iv) Preliminaries-ED|ED-2 e) and f) with $x \equiv \frac{(1+\bar{m})\rho}{\bar{s}(1+\rho)}$ (note that $\frac{(1+\bar{m})\rho}{\bar{s}(1+\rho)} > 1$ for ED|ED),

v) Preliminaries-ED|ED-2 g), and CLT we have

$$
\begin{aligned}
&\left(\sum_{k=s_n}^{n+m_n} c_n\right)/factor_2 \\
&= \left(\left((\rho/s_n)^{n+m_n-1} n^{m_n}(s_n!)^{-1} \quad n! \quad s_n^{s_n} \quad exp\{s_n/\rho\}\right)/factor_2\right) \\
&\cdot P(X_n \le n+m_n-1-s_n) \\
&= \frac{s_n^{s_n}}{s_n!} \quad \frac{(n+m_n)!}{(n+m_n)} \quad exp\{s_n/\rho\} \quad \left(\frac{\rho}{s_n(1+\rho)}\right)^{n+m_n-1} \quad P(X_n \le n+m_n-1-s_n) \\
&\qquad \left(X_n \sim Poisson\left(\frac{s_n}{\rho}\right)\right) \\
&\cong \sqrt{\frac{n+m_n}{s_n}} \quad \left(\frac{(n+m_n)\rho}{s_n(1+\rho)}\right)^{n+m_n-1} \quad exp\{s_n+s_n/\rho-(n+m_n)\} \\
&\quad \cdot P(X_n \le n+m_n-1-s_n) \\
&= \sqrt{\frac{n+m_n}{s_n}} \quad exp\left\{n\left\{(1+\bar{m}) \quad ln\left(\frac{(1+\bar{m})\rho}{\bar{s}(1+\rho)}\right)+\bar{s}+\frac{\bar{s}}{\rho}-(1+\bar{m})\right\}+o\left(n^{1-\epsilon}\right)\right\} \\
&\quad \cdot P(X_n \le n+m_n-1-s_n) \\
&\to \sqrt{\frac{1+\bar{m}}{\bar{s}}} \cdot +\infty \cdot P(\tilde{Z}_n < +\infty) = +\infty \qquad \left(\tilde{Z}_n \sim Normal(0,1)\right)
\end{aligned}
\tag{A.68}
$$

Since $\left(\sum_{k=0}^{m_n-1} a_n\right)/factor_2 \ge 0$ for all n, from (A.65)-(A.68), $\tilde{A}'_n \to \frac{0}{+\infty} = 0$. From (1.2) and (A.65),

$$\lim_{n\to\infty} \tilde{A}_n \leq \lim_{n\to\infty} (n + m_n - m_n/\mu m_n) \lim_{n\to\infty} \tilde{A}'_n = 0 \Rightarrow \lim_{n\to\infty} \tilde{A}_n = 0. \tag{A.69}$$

Now we look at $\tilde{B}_n$. From Preliminaries-ED|ED-2 a) and h),

$$\tilde{B}_n = \tilde{B}^1_n + \tilde{B}^2_n, \tag{A.70}$$

and for τ_b taken as given in Preliminaries-ED|ED-2 h), let $\bar{F}_n = \{min_{\,t\in(\tau_b,\infty)} N^{st}_n(t) < s_n - \epsilon_1.n\}$ and $\bar{F}'_n$ be its complement. Then,

$$\begin{aligned}
&\tilde{B}^1_n \leq \left[\sum_{k=s_n}^{n+m_n} E[W^{st,\tau_b}_n(s_n, m_n) \cdot 1_{\bar{F}_n} \mid N^{st}_n(\tau_b) = k] \quad c_n\right] \\
&/ \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right] \\
&= \left[\sum_{k=s_n}^{n+m_n} E[W^{st,\tau_b}_n(s_n, m_n) \mid \bar{F}_n, N^{st}_n(\tau_b) = k] \quad P(\bar{F}_n | N^{st}_n(\tau_b) = k) \quad c_n\right] \\
&/ \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right].
\end{aligned} \tag{A.71}$$

For $k = s_n, s_n + 1, \ldots, n + m_n$, the event $\{\bar{F}_n \mid (N^{st}_n(\tau_b) = k)\}$ considers whether the process $\{N_n(t), t \geq 0\}$, after reaching steady state, will ever visit the region $N^{st}_n = \{0, 1, \ldots, \lfloor s_n - \epsilon_1.n \rfloor\}$ given the observation starts at a state in $N^{st}_n = \{s_n, s_n + 1, \ldots, n + m_n\}$ at time τ_b. Note that since N^{st}_n is a birth-and-death process with no absorbing states, $\forall k_1, k_2 \,\epsilon\, \{s_n, \ldots, n + m_n\}$

$$P(\bar{F}_n \mid N^{st}_n(\tau_b) = k_1) = P(\bar{F}_n \mid N^{st}_n(\tau_b) = k_2) \tag{A.72}$$

and we denote this probability with $P^{\tau_b}_{s_n}$. Then, from Preliminaries-ED|ED-2 b) and similar to (A.65) we have

$$\left[\sum_{k=s_n}^{n+m_n} E[W_n^{st,\tau_b}(s_n,m_n) \mid \bar{F}_n, N_n^{st}(\tau_b)=k] \quad P(\bar{F}_n \mid N_n^{st}(\tau_b)=k) \quad c_n\right]$$

$$/\left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right]$$

$$= \left[P_{s_n}^{\tau_b} \quad \sum_{k=s_n}^{n+m_n} E[W_n^{st,\tau_b}(s_n,m_n) \mid \bar{F}_n, N_n^{st}(\tau_b)=k] \quad c_n\right] \tag{A.73}$$

$$/\left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right]$$

$$\leq P_{s_n}^{\tau_b}\left(\tfrac{n}{\mu m_n}\right)\tilde{B}'_n.$$

Thus, combining (A.71) and (A.73) we have

$$\tilde{B}_n^1 \leq P_{s_n}^{\tau_b}\left(\tfrac{n}{\mu m_n}\right)\tilde{B}'_n. \tag{A.74}$$

From Preliminaries-ED|ED-2 b), i), and (1.2), $0 \leq \tilde{B}'_n \leq 1$, $P_{s_n}^{\tau_b} \to 0$, and $n/\mu m_n \to \frac{1}{\mu\bar{m}}$, respectively; thus, we have $\tilde{B}_n^1 \to 0$.

Following Preliminaries-ED|ED-2 h) and similar to (A.71), we have

$$\tilde{B}_n^2 = \left[\sum_{k=s_n}^{n+m_n} E[W_n^{st,\tau_b}(s_n,m_n) \mid F'_n, N_n^{st}(\tau_b)=k] \quad P(F'_n \mid N_n^{st}(\tau_b)=k) \quad c_n\right]$$

$$/\left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right]. \tag{A.75}$$

Let the event F_n and $\bar{F}_n$ be as defined in Preliminaries-ED|ED-2 h) and (A.71), respectively. Then, for any $k = s_n, s_n+1, \ldots, n+m_n$,

$$P(F'_n \mid N_n^{st}(\tau_b)=k) = 1 - P(F_n \mid N_n^{st}(\tau_b)=k), \tag{A.76}$$

and from Preliminaries-ED|ED-2 i),

$$\begin{aligned} &P(F_n \mid N_n^{st}(\tau_b) = k) \leq P(\bar{F}_n \mid N_n^{st}(\tau_b) = k) \to 0 \\ &\Rightarrow P(F_n \mid N_n^{st}(\tau_b) = k) \to 0; \end{aligned} \tag{A.77}$$

and therefore

$$\lim_{n\to\infty} P(F_n' \mid N_n^{st}(\tau_b) = k) = 1. \tag{A.78}$$

Let $\left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right]$ be denoted by $denominator_1$. Thus, for sufficiently large n and $\forall\, \varepsilon_1, \varepsilon_2 > 0$, from (A.75)-(A.78) and Preliminaries-ED|ED-2 h) and i)

$$\begin{aligned} &\left[(1-\varepsilon_2)\sum_{k=s_n}^{n+m_n}\left((k+1-m_n)/\mu s_n\right)c_n\right]/denominator_1 \\ &\leq \tilde{B}_n^2 \leq \left[(1+\varepsilon_2)\sum_{k=s_n}^{n+m_n}\left((k+1-m_n)/\mu(s_n-\epsilon_1.n)\right)c_n\right]/denominator_1. \end{aligned} \tag{A.79}$$

Since $\varepsilon_1, \varepsilon_2 > 0$ is arbitrary, from (A.79)

$$\lim_{n\to\infty} \tilde{B}_n^2 = \lim_{n\to\infty}\left\{\left[\sum_{k=s_n}^{n+m_n}\left((k+1-m_n)/\mu s_n\right)c_n\right]/denominator_1\right\}. \tag{A.80}$$

Now let

$$\lim_{n\to\infty} \tilde{BA}_n^2 \equiv \lim_{n\to\infty} \left[\sum_{k=s_n}^{\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\varepsilon}\rfloor-1} \left((k+1-m_n)/\mu\, s_n\right)\, c_n \right] /denominator_1,$$

$$\lim_{n\to\infty} \tilde{BB}_n^2 \equiv \lim_{n\to\infty} \left[\sum_{k=\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\varepsilon}\rfloor}^{\lceil n(1+\bar{m}-\bar{s}/\rho)+n^{1-\varepsilon}\rceil} \left((k+1-m_n)/\mu\, s_n\right)\, c_n \right] /denominator_1,$$

$$\lim_{n\to\infty} \tilde{BC}_n^2 \equiv \lim_{n\to\infty} \left[\sum_{k=\lceil n(1+\bar{m}-\bar{s}/\rho)+n^{1-\varepsilon}\rceil+1}^{n+m_n} \left((k+1-m_n)/\mu\, s_n\right)\, c_n \right] /denominator_1. \tag{A.81}$$

From (A.80), (A.81), and Preliminaries-ED|ED-2 j), we have

$$\lim_{n\to\infty} \tilde{B}_n^2 = \lim_{n\to\infty} \tilde{BA}_n^2 + \lim_{n\to\infty} \tilde{BB}_n^2 + \lim_{n\to\infty} \tilde{BC}_n^2. \tag{A.82}$$

First we look at the $\tilde{BA}_n^2$. Let $X_n, Y_n \sim Poisson\left(\frac{s_n}{\rho}\right)$ and $factor_1$ as defined in (A.30); then, from Preliminaries-ED|ED-1 e),

$$\tilde{BA}_n^2 \le \left(\tfrac{n+m_n-m_n}{\mu s_n}\right) \left[\sum_{k=s_n}^{\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\varepsilon}\rfloor-1} c_n \right] /denominator_1$$

$$= \left(\tfrac{n}{\mu s_n}\right) P\left(n+m_n-1-\lfloor n\left(1+\bar{m}-\bar{s}/\rho\right)-n^{1-\varepsilon}\rfloor+1 \le X_n \le n+m_n-1-s_n\right)$$

$$\cdot \left[\left(\sum_{k=0}^{m_n-1} n\, \tfrac{(n\rho)^k}{k!} \right) /factor_1 + \left(\sum_{k=m_n}^{s_n-1} \tfrac{n^{m_n}}{k!}\, \tfrac{n!}{(n+m_n-k-1)!}\, \rho^k \right) /factor_1 \right.$$

$$\left. + P(Y_n \le n+m_n-1-s_n) \right]^{-1} \tag{A.83}$$

From Preliminaries-ED|ED-1 f)i) and Preliminaries-ED|ED-2 g), for $Z_n, \bar{Z}_n \sim N(0,1)$

$$P\left(n+m_n-1-\lfloor n\left(1+\bar{m}-\bar{s}/\rho\right)-n^{1-\varepsilon}\rfloor+1\leq X_n\leq n+m_n-1-s_n\right)$$

$$\rightarrow P(+\infty<Z_n<+\infty)=0 \tag{A.84}$$

$$P(Y_n\leq n+m_n-1-s_n)\rightarrow P(\bar{Z}_n<+\infty)=1.$$

Since $s_n>m_n$ here, and $\bar{s}<\rho$ for ED|ED regime from Theorem 12; for sufficiently large n $n\rho>s_n-1>m_n-1$. Hence, for $k=0,1,2,\ldots,m_n-2$

$$\left(\tfrac{(n\rho)^{k+1}}{k+1!}\right)/\left(\tfrac{(n\rho)^k}{k!}\right)=\tfrac{n\rho}{k+1}>\tfrac{n\rho}{m_n-1}>\tfrac{n\rho}{s_n-1}>1. \tag{A.85}$$

Therefore, for $k=0,1,2,\ldots,s_n-2$, for sufficiently large n, $\frac{(n\rho)^k}{k!}$ is increasing in k, and since $s_n>m_n$ in this case,

$$\begin{aligned}
&\left(\sum_{k=0}^{m_n-1} n\tfrac{(n\rho)^k}{k!}\right)/factor_1\\
&=\left(\sum_{k=0}^{m_n-1} n\tfrac{(n\rho)^k}{k!}\right)/\left(\left(\tfrac{\rho}{s_n}\right)^{n+m_n-1}\tfrac{n^{m_n}}{s_n!}\quad n!\quad s_n^{s_n}\quad exp\{s_n/\rho\}\right) \qquad \text{(A.86)}\\
&\leq\left(n\quad\tfrac{(n\rho)^{s_n-1}}{(s_n-1)!}\quad s_n\right)/\left(\left(\tfrac{\rho}{s_n}\right)^{n+m_n-1}\tfrac{n^{m_n}}{s_n!}\quad n!\quad s_n^{s_n}\quad exp\{s_n/\rho\}\right).
\end{aligned}$$

Following (A.34)- (A.36), again we will have

$$\begin{aligned}
&\left(n\,\tfrac{(n\rho)^{s_n-1}}{(s_n-1)!}s_n\right)/\left(\left(\tfrac{\rho}{s_n}\right)^{n+m_n-1}\tfrac{n^{m_n}}{s_n!}\quad n!\quad s_n^{s_n}\quad e^{\frac{s_n}{\rho}}\right)\\
&\cong\tfrac{1}{\sqrt{2\pi}}\quad(\bar{s}\sqrt{n}+\hat{s}_n)\quad exp\{-n\left\{(1+\bar{m}-\bar{s})ln\left(\rho/\bar{s}\right)+\bar{s}/\rho-1\right\}+o(n^{1-\varepsilon})\}.
\end{aligned} \tag{A.87}$$

We know it is assumed $n+m_n\geq s_n$ in general (since otherwise the analysis will be trivial), $\bar{m}>\bar{s}/\rho+\bar{s}-1$ and $\bar{s}<\rho$ for ED|ED from Theorem 12; thus, $1+\bar{m}-\bar{s}\geq 0$, $ln\left(\rho/\bar{s}\right)>0$, and hence we have

$$(1+\bar{m}-\bar{s})ln\,(\rho/\bar{s})+\bar{s}/\rho-1>(\bar{s}/\rho)ln\,(\rho/\bar{s})+\bar{s}/\rho-1=-(\bar{s}/\rho)ln\,(\bar{s}/\rho)+\bar{s}/\rho-1.$$

Note that this is the value of $-xlnx+x-1|_{x=(\bar{s}/\rho)}$ and it is the negative of the function in Preliminaries-ED|ED-2 f). Therefore, we know that $-xlnx-1+x>0$ for $x<1$ and since $\bar{s}/\rho<1$, we conclude that

$$\begin{aligned}&(1+\bar{m}-\bar{s})ln\,(\rho/\bar{s})+\bar{s}/\rho-1>0\\&\Rightarrow \tfrac{1}{\sqrt{2\pi}}\quad(\bar{s}\sqrt{n}+\hat{s}_n)/exp\left\{n\left\{(1+\bar{m}-\bar{s})ln\,(\rho/\bar{s})+\bar{s}/\rho-1\right\}+o(n^{1-\varepsilon})\right\}\rightarrow \tfrac{+\infty}{+\infty}.\end{aligned}\tag{A.88}$$

Then, using L'Hospital's rule similar to (A.37), from (A.86)-(A.88) we conclude that

$$\begin{aligned}&\left(n\quad \tfrac{(n\rho)^{s_n-1}}{(s_n-1)!}\quad s_n\right)/\left(\left(\tfrac{\rho}{s_n}\right)^{n+m_n-1}\tfrac{n^{m_n}}{s_n!}\quad n!\quad s_n^{s_n}\quad exp\{s_n/\rho\}\right)\rightarrow 0\\&\Rightarrow\left(\sum_{k=0}^{m_n-1}n\tfrac{(n\rho)^k}{k!}\right)/factor_1\rightarrow 0.\end{aligned}\tag{A.89}$$

For $k=m_n,m_n+1,\ldots,s_n-1$,

$$n(n-1)(n-2)\ldots(n-(k-m_n))<n.n^{k-m_n}\tag{A.90}$$

and this gives

$$\tfrac{n^{m_n}}{k!}\quad \tfrac{n!}{(n+m_n-k-1)!}\quad \rho^k<n\quad \tfrac{(n\rho)^k}{k!}.\tag{A.91}$$

Then,

$$\left[\sum_{k=m_n}^{s_n-1}\tfrac{n^{m_n}}{k!}\quad \tfrac{n!}{(n+m_n-k-1)!}\quad \rho^k\right]/factor_1\leq\left[\sum_{k=m_n}^{s_n-1}n\quad \tfrac{(n\rho)^k}{k!}\right]/factor_1.\tag{A.92}$$

Similar to (A.86)-(A.89),

$$\begin{aligned}
&\left[\sum_{k=m_n}^{s_n-1} n \quad \tfrac{(n\rho)^k}{k!}\right] / factor_1 \leq \left[n \, \tfrac{(n\rho)^{s_n-1}}{(s_n-1)!} \quad s_n\right] / factor_1 \to 0 \\
&\Rightarrow \left[\sum_{k=m_n}^{s_n-1} \tfrac{n^{m_n}}{k!} \quad \tfrac{n!}{(n+m_n-k-1)!} \quad \rho^k\right] / factor_1 \to 0.
\end{aligned} \tag{A.93}$$

From (A.83)-(A.93) and (1.2),

$$\lim_{n\to\infty} \tilde{BA}_n^2 \leq \lim_{n\to\infty} \left(\tfrac{n}{\mu s_n}\right) \left[\sum_{k=s_n}^{\lfloor n(1+\bar{m}-\bar{s}/\rho)-n^{1-\varepsilon}\rfloor-1} c_n\right] / denominator_1 = 0 \Rightarrow \tilde{BA}_n^2 \to 0. \tag{A.94}$$

Now we look at $\tilde{BB}_n^2$. Let $X_n, Y_n \sim Poisson\,(s_n/\rho)$. From (A.81) and similar to (A.83),

$$\begin{aligned}
&((\lfloor n\,(1+\bar{m}-\bar{s}/\rho) - n^{1-\epsilon}\rfloor + 1 - m_n)\,/\mu\, s_n) \\
&\tilde{BB}_n^{'2} \leq \tilde{BB}_n^2 \leq ((\lceil n\,(1+\bar{m}-\bar{s}/\rho) + n^{1-\epsilon}\rceil + 1 - m_n)\,/\mu\, s_n)\,\tilde{BB}_n^{'2} \\
&\tilde{BB}_n^{'2} = P\,(n + m_n - 1 - \lceil n\,(1+\bar{m}-\bar{s}/\rho) + n^{1-\epsilon}\rceil \leq X_n \\
&\leq n + m_n - 1 - \lfloor n\,(1+\bar{m}-\bar{s}/\rho) - n^{1-\epsilon}\rfloor) \\
&\cdot \left[\left(\sum_{k=0}^{m_n-1} n\tfrac{(n\rho)^k}{k!}\right) / factor_1 + \left(\sum_{k=m_n}^{s_n-1} \tfrac{n^{m_n}}{k!}\tfrac{n!}{(n+m_n-k-1)!}\rho^k\right) / factor_1\right. \\
&\left. + P\,(Y_n \leq n + m_n - 1 - s_n)\right]^{-1}
\end{aligned} \tag{A.95}$$

Since $0 < \epsilon < 1/2$ is arbitrary, from (1.2), CLT, Preliminaries-ED|ED-1 f)i), and the fact that the expressions in the denominator of $\tilde{BB}_n^{'2}$ are the same as in $\tilde{BA}_n^{'2}$, (A.95) gives

$$\lim_{n\to\infty} \tilde{BB}_n^2 = \left(\frac{1-\bar{s}/\rho}{\mu\bar{s}}\right) = \left(\frac{1}{\mu\bar{s}} - \frac{1}{\lambda}\right). \tag{A.96}$$

Now we look at $\tilde{BC}_n^2$. With $X_n, Y_n \sim Poisson\left(\frac{s_n}{\rho}\right)$, from (A.81), and similar to (A.83)

$$\begin{aligned}
&\tilde{BC}_n^2 \leq (n/\mu s_n)\, \tilde{BC}_n^{'2}, \\
&\tilde{BC}_n^{'2} = P\left(X_n \leq n + m_n - 1 - \lceil n\left(1 + \bar{m} - \bar{s}/\rho\right) + n^{1-\epsilon}\rceil - 1\right) \\
&\cdot \left[\left(\sum_{k=0}^{m_n-1} n\, \frac{(n\rho)^k}{k!}\right)/factor_1 + \left(\sum_{k=m_n}^{s_n-1} \frac{n^{m_n}}{k!}\, \frac{n!}{(n+m_n-k-1)!}\, \rho^k\right)/factor_1\right. \\
&\left. + P(Y_n \leq n + m_n - 1 - s_n)\right]^{-1}
\end{aligned} \tag{A.97}$$

From CLT and Preliminaries-ED|ED-1 f)i), for $Z_n \sim Normal(0,1)$

$$P\left(X_n \leq n + m_n - 1 - \lceil n\left(1 + \bar{m} - \bar{s}/\rho\right) + n^{1-\epsilon}\rceil - 1\right) \to P(Z_n < -\infty) = 0 \tag{A.98}$$

The expressions in the denominator of $\tilde{BC}_n^{'2}$ are the same as the ones in the denominator of $\tilde{BA}_n^{'2}$. Hence, $\lim_{n\to\infty} \tilde{BC}_n^2 = 0$.

Thus, for $s_n > m_n$ in ED|ED

$$\begin{aligned}
&\lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)] = \lim_{n\to\infty} (\tilde{A}_n + \tilde{B}_n) = \lim_{n\to\infty} (\tilde{B}_n^1 + \tilde{B}_n^2) = \lim_{n\to\infty} \tilde{B}_n^2 \\
&= \lim_{n\to\infty} \tilde{BB}_n^2 = \left(\frac{1}{\mu\bar{s}} - \frac{1}{\lambda}\right). \square
\end{aligned} \tag{A.99}$$

Proof of Formulas for $\lim_{\mathbf{n}\to\infty} \mathbf{E}[\mathbf{D_n^{rpl}}(\mathbf{s_n}, \mathbf{m_n})]$ *for QD|ED*

Note that we have $\bar{m} < b < \bar{s}$ from Theorem 12; and for sufficiently large n (1.2) gives $s_n > m_n$ as the only case here.

Preliminaries-QD|ED

a.

Let $0 < \epsilon < 1/2$; then, for sufficiently large n, from Theorem 12 we have

$$m_n \leq \lceil \tfrac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rceil \leq \lfloor \tfrac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\epsilon} \rfloor \leq s_n - 1. \tag{A.100}$$

b.

From (1.2) and the inequalities describing QD|ED in Theorem 12,

$$\begin{aligned} &\lim_{n\to\infty} \left[\lfloor \tfrac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor - (n + m_n - 1)\tfrac{\rho}{1+\rho} \right] / \left[\sqrt{(n + m_n - 1)\tfrac{\rho}{(1+\rho)^2}} \right] \\ &= \lim_{n\to\infty} \left[-n^{1-\epsilon} - (\hat{m}_n \sqrt{n} - 1)\tfrac{\rho}{1+\rho} \right] / \left[\sqrt{(n + m_n - 1)\tfrac{\rho}{(1+\rho)^2}} \right] = -\infty, \end{aligned} \tag{A.101}$$

c.

For an $\varepsilon_1 > 0$, let τ_b and $W_n^{st,\tau_b}(s_n, m_n)$ be as in Preliminaries-ED|ED-2 h) and let the event

$$\begin{aligned} F_n = &\left(min_{\, t \,\epsilon\, (\tau_b, \tau_b + W_n^{st,\tau_b}(s_n,m_n)]} \, N_n^{st}(t) < \tfrac{\rho(1+\bar{m})}{1+\rho} n - \epsilon_1 . n - n^{1-\epsilon} \right) \bigcup \\ &\left(max_{\, t \,\epsilon\, (\tau_b, \tau_b + W_n^{st,\tau_b}(s_n,m_n)]} \, N_n^{st}(t) > \tfrac{\rho(1+\bar{m})}{1+\rho} n + \epsilon_1 . n + n^{1-\epsilon} \right) \end{aligned} \tag{A.102}$$

and F_n' be its complement. Finally, let $denominator_1$ be as defined previously. Then,

$$\begin{aligned} \tilde{A}_n^1 &\equiv \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\epsilon} \rfloor} E[W_n^{st,\tau_b}(s_n, m_n).1_{F_n} | N_n^{st}(\tau_b) = k] d_n \right] / denominator_1, \\ \tilde{A}_n^2 &\equiv \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\epsilon} \rfloor} E[W_n^{st,\tau_b}(s_n, m_n).1_{F_n'} | N_n^{st}(\tau_b) = k] d_n \right] / denominator_1. \end{aligned} \tag{A.103}$$

d.

Note that $\varepsilon_1 > 0$ is arbitrary. For $k = \lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rceil, \ldots, \lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\epsilon} \rfloor$, similar to Preliminaries-ED|ED-2 i) and from Theorem 12 we have

$$
\begin{aligned}
& P\left(\left(min_{\,t\,\epsilon\,(\tau_b,\infty)} N_n^{st}(t) < \tfrac{\rho(1+\bar{m})}{1+\rho} n - \varepsilon_1 . n - n^{1-\varepsilon}\right) \bigcup \right. \\
& \left. \left(max_{\,t\,\epsilon\,(\tau_b,\infty)} N_n^{st}(t) > \tfrac{\rho(1+\bar{m})}{1+\rho} n + \varepsilon_1 . n + n^{1-\varepsilon}\right) \mid N_n^{st}(\tau_b) = k\right) \\
& = P\left(\left(min_{\,t\,\epsilon\,(\tau_b,\infty)} \tfrac{N_n^{st}(t)}{n} < \tfrac{\rho(1+\bar{m})}{1+\rho} - \varepsilon_1 - n^{-\varepsilon}\right)\right. \\
& \left. \bigcup \left(max_{\,t\,\epsilon\,(\tau_b,\infty)} \tfrac{N_n^{st}(t)}{n} > \tfrac{\rho(1+\bar{m})}{1+\rho} + \varepsilon_1 + n^{-\varepsilon}\right) \mid \tfrac{N_n^{st}(\tau_b)}{n} = \tfrac{k}{n}\right) \\
& = P\left(\left(min_{\,t\,\epsilon\,(\tau_b,\infty)} \tfrac{N_n^{st}(t)}{n} - \bar{\mathbf{b}}(t) + \bar{\mathbf{b}}(t) < \tfrac{\rho(1+\bar{m})}{1+\rho} - \varepsilon_1 - n^{-\varepsilon}\right) \bigcup \right. \\
& \left. \left(max_{\,t\,\epsilon\,(\tau_b,\infty)} \tfrac{N_n}{n} - \bar{\mathbf{b}}(t) + \bar{\mathbf{b}}(t) > \tfrac{\rho(1+\bar{m})}{1+\rho} + \varepsilon_1 + n^{-\varepsilon}\right) \mid \tfrac{N_n^{st}(\tau_b)}{n} = \tfrac{k}{n}\right) \\
& \leq P\left(\left(min_{\,t\,\epsilon\,(\tau_b,\infty)} \bar{\mathbf{b}}(t) - max_{\,t\,\epsilon\,(\tau_b,\infty)} \left|\tfrac{N_n^{st}(t)}{n} - \bar{\mathbf{b}}(t)\right| < \tfrac{\rho(1+\bar{m})}{1+\rho} - \varepsilon_1 - n^{-\varepsilon}\right)\right. \\
& \left. \bigcup \left(max_{\,t\,\epsilon\,(\tau_b,\infty)} \bar{\mathbf{b}}(t) + max_{\,t\,\epsilon\,(\tau_b,\infty)} \left|\tfrac{N_n^{st}(t)}{n} - \bar{\mathbf{b}}(t)\right| > \tfrac{\rho(1+\bar{m})}{1+\rho} + \varepsilon_1 + n^{-\varepsilon}\right)\right. \\
& \left. \mid \tfrac{N_n^{st}(\tau_b)}{n} = \tfrac{k}{n}\right) \\
& \to P(\left(min_{\,t\,\epsilon\,(\tau_b,\infty)} \bar{\mathbf{b}}(t) < \tfrac{\rho(1+\bar{m})}{1+\rho} - \varepsilon_1\right) \bigcup \left(max_{\,t\,\epsilon\,(\tau_b,\infty)} \bar{\mathbf{b}}(t) > \tfrac{\rho(1+\bar{m})}{1+\rho} + \varepsilon_1\right) \mid \bar{\mathbf{b}}(\tau_b) \\
& = \tfrac{\rho(1+\bar{m})}{1+\rho}) = 0.
\end{aligned}
\tag{A.104}
$$

Main Part of the Proof

Since $s_n > m_n$ is the only possible case, $E[D_n^{rpl}(s_n, m_n)]$ is as defined in (A.64). Let $denominator_1$ be as defined in (A.79) and $0 < \epsilon < 1/2$. Then, from Preliminaries-ED|ED-2 a) and QD|ED a), for sufficiently large n we can write

$$\tilde{A}_n = \Bigg[\sum_{k=m_n}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k] d_n$$

$$+ \sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\epsilon} \rfloor} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k] d_n \tag{A.105}$$

$$+ \sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\epsilon} \rceil}^{s_n - 1} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k] d_n \Bigg] / denominator_1.$$

Since at least m_n units will be broken during the delay, similar to (A.65)

$$\begin{aligned} &\Bigg[\sum_{k=m_n}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k] d_n \Bigg] / denominator_1 \\ &\leq \left(\frac{n}{\mu m_n} \right) \Bigg[\sum_{k=m_n}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor} d_n \Bigg] / denominator_1. \end{aligned} \tag{A.106}$$

Let $factor_2$ be as in (A.66) and $Y_n, \bar{Y}_n \sim Binomial(n + m_n - 1, \frac{\rho}{1+\rho})$. Then, similar to Preliminaries-ED|ED-2 c) and (A.66)

$$\Bigg[\sum_{k=m_n}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor} d_n \Bigg] / denominator_1 = P(m_n \leq Y_n \leq \lfloor \tfrac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor)$$

$$\left[\left(\sum_{k=0}^{m_n - 1} a_n \right) / factor_2 + P(m_n \leq \bar{Y}_n \leq s_n - 1) + \left(\sum_{k=s_n}^{n+m_n} c_n \right) / factor_2 \right]^{-1}. \tag{A.107}$$

Let $Z_n, \bar{Z}_n \sim Normal(0, 1)$. Note that contrary to ED|ED, here we have $\bar{s}\rho + \bar{s} - \rho - \bar{m}\rho > 0$. Then, from CLT, Preliminaries-ED|ED-2 d), and QD|ED b),

$$P(m_n \leq Y_n \leq \lfloor \tfrac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor) \to P(-\infty < Z_n < -\infty) = 0,$$
$$P(m_n \leq \bar{Y}_n \leq s_n - 1) \to P(-\infty < \bar{Z}_n < +\infty) = 1, \tag{A.108}$$

and since $\left(\sum_{k=0}^{m_n-1} a_n + \sum_{k=s_n}^{n+m_n} c_n \right) / factor_2 \geq 0$, from (1.2), (A.106)-(A.108) we have

$$\lim_{n\to\infty} \left[\sum_{k=m_n}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor} d_n \right] / denominator_1 = 0$$
$$\Rightarrow \left[\sum_{k=m_n}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rfloor} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k] d_n \right] / denominator_1 \to 0. \tag{A.109}$$

From Preliminaries-QD|ED c) we have

$$\left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\epsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\epsilon} \rfloor} E[W_n^{st}(s_n, m_n) | N_n^{st}(\tau_b) = k] d_n \right] / denominator_1 = \tilde{A}_n^1 + \tilde{A}_n^2. \tag{A.110}$$

Let

$$\bar{F}_n = \left(min_{\, t \, \epsilon \, (\tau_b, \infty)} \, N_n^{st}(t) < \frac{\rho(1+\bar{m})}{1+\rho} n - \varepsilon_1 . n - n^{1-\varepsilon} \right)$$
$$\bigcup \left(max_{\, t \, \epsilon \, (\tau_b, \infty)} \, N_n^{st}(t) > \frac{\rho(1+\bar{m})}{1+\rho} n + \varepsilon_1 . n + n^{1-\varepsilon} \right)$$

and $\bar{F}_n'$ be its complement. Then, similar to (A.71)

$$\tilde{A}_n^1 \leq \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} E[W_n^{st,\tau_b}(s_n, m_n).1_{\bar{F}_n} | N_n^{st}(\tau_b) = k] d_n \right] / denominator_1,$$

$$= \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} E[W_n^{st,\tau_b}(s_n, m_n) | \bar{F}_n, N_n^{st}(\tau_b) = k] P(\bar{F}_n | N_n^{st}(\tau_b) = k) d_n \right]$$

$/ denominator_1.$

(A.111)

For $k = \lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil, \ldots, \lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor$, the event $\{\bar{F}_n \mid N_n^{st}(\tau_b) = k\}$ considers $\forall \varepsilon_1 > 0$ whether the process $\{N_n(t), t \geq 0\}$, after reaching steady state, will ever visit the region $\{0, 1, \ldots, \lfloor \frac{\rho(1+\bar{m})}{1+\rho} n - \varepsilon_1 . n - n^{1-\varepsilon} \rfloor\} \bigcup \{\lceil \frac{\rho(1+\bar{m})}{1+\rho} n + \varepsilon_1 . n + n^{1-\varepsilon} \rceil, \ldots , n + m_n\}$ if the observation starts at a state in $\{\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil, \ldots, \lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor\}$ at time τ_b. Note that since N_n^{st} is a birth-and-death process with no absorbing states, $\forall k_1, k_2 = \left\{ \lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil, \ldots, \lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor \right\}$, $P(\bar{F}_n | N_n^{st}(\tau_b) = k_1) = P(\bar{F}_n | N_n^{st}(\tau_b) = k_2)$ and we denote this probability with $P^{\tau_b}_{\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}$. Then, from Preliminaries-QD|ED d) we have

$$\tilde{A}_n^1$$

$$\leq \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} E[W_n^{st,\tau_b}(s_n, m_n) \mid \bar{F}_n, N_n^{st}(\tau_b) = k] P(\bar{F}_n \mid N_n^{st}(\tau_b) = k) \quad d_n \right]$$

$$/denominator_1$$

$$= P^{\tau_b}_{\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}$$

$$\cdot \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} E[W_n^{st,\tau_b}(s_n, m_n) \mid \bar{F}_n, N_n^{st}(\tau_b) = k] \quad d_n \right] /denominator_1$$

$$\leq P^{\tau_b}_{\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil} \quad \left(\frac{n}{\mu m_n} \right) \quad \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} d_n \right] /denominator_1 \tag{A.112}$$

From Preliminaries-QD|ED d), $P^{\tau_b}_{\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil} \to 0$; moreover, $\frac{n}{\mu m_n} \to \frac{1}{\mu \bar{m}}$, and

$$0 \leq \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} d_n \right] /denominator_1 \leq 1,$$

then we have $\tilde{A}_n^1 \to 0$. Now we look at $\tilde{A}_n^2$.

Following the same procedure for $\tilde{A}_n^1$, and from Preliminaries-QD|ED c)

$$\tilde{A}_n^2$$

$$= \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} E[W_n^{st,\tau_b}(s_n, m_n) \mid 1_{F'_n}, N_n^{st}(\tau_b) = k] \quad P(F'_n \mid N_n^{st}(\tau_b) = k) \quad d_n \right]$$

$$/denominator_1.$$

(A.113)

For $k = \lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil, \ldots, \lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor$

$$P(F'_n \mid N_n^{st}(\tau_b) = k) = 1 - P(F_n \mid N_n^{st}(\tau_b) = k) \tag{A.114}$$

and from Preliminaries-QD|ED d), as $n \to \infty$

$$\begin{aligned} & P(F_n \mid N_n^{st}(\tau_b) = k) \leq P(\bar{F}_n \mid N_n^{st}(\tau_b) = k) \to 0 \\ & \Rightarrow P(F'_n | N_n^{st}(\tau_b) = k) = 1 - P(F_n \mid N_n^{st}(\tau_b) = k) \to 1. \end{aligned} \tag{A.115}$$

Then, from Preliminaries-QD|ED c) and d), and (A.113)-(A.115) we have for $\varepsilon_1, \varepsilon_2 > 0$

$$\begin{aligned} & (1-\varepsilon_2) \left\{ \left(\lceil \tfrac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil + 1 - m_n \right) / \left[\mu \lfloor \tfrac{\rho(1+\bar{m})}{1+\rho} n + \epsilon_1 . n + n^{1-\varepsilon} \rfloor \right] \right\} \tilde{A}'^2_n \leq \tilde{A}^2_n \\ & \leq (1+\varepsilon_2) \left\{ \left(\lfloor \tfrac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor + 1 - m_n \right) / \left[\mu \lceil \tfrac{\rho(1+\bar{m})}{1+\rho} n - \epsilon_1 . n - n^{1-\varepsilon} \rceil \right] \right\} \tilde{A}'^2_n \\ & \tilde{A}'^2_n = \left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil}^{\lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor} d_n \right] / denominator_1 \end{aligned} \tag{A.116}$$

Similar to Preliminaries-ED|ED-2 c) and (A.107), for $Y_n, \bar{Y}_n \sim Binomial(n + m_n - 1, \frac{\rho}{1+\rho})$,

$$\begin{aligned} & \tilde{A}'^2_n = P(\lceil \tfrac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil \leq Y_n \leq \lfloor \tfrac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor) \\ & \cdot \left[\left(\sum_{k=0}^{m_n - 1} a_n \right) / factor_2 + P(m_n \leq \bar{Y}_n \leq s_n - 1) + \left(\sum_{k=s_n}^{n+m_n} c_n \right) / factor_2 \right]^{-1}. \end{aligned} \tag{A.117}$$

From Preliminaries-QD|ED b), CLT, and (A.108)

$$
\begin{array}{l}
P(\lceil \frac{\rho(1+\bar{m})}{1+\rho} n - n^{1-\varepsilon} \rceil \leq Y_n \leq \lfloor \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rfloor) \to 1 \\
P(m_n \leq \bar{Y}_n \leq s_n - 1) \to 1.
\end{array}
\tag{A.118}
$$

From Theorem 12, for QD|ED regime, for sufficiently large n, we know that $n\rho > m_n - 1$. Hence, similar to (A.32), for $k = 0, 1, 2, \ldots, m_n - 2$, for suff. large n, $(n\rho)^k/k!$ is increasing in k, and

$$
\begin{array}{l}
\left[\sum_{k=0}^{m_n-1} n \frac{(n\rho)^k}{k!} \right] / factor_2 = \left[\sum_{k=0}^{m_n-1} n \frac{(n\rho)^k}{k!} \right] / \left[\frac{n^{m_n} \quad n!}{(n+m_n-1)!} (1+\rho)^{n+m_n-1} \right] \\
\leq \left[n \frac{(n\rho)^{m_n-1}}{(m_n-1)!} m_n \right] / \left[\frac{n^{m_n} \quad n!}{(n+m_n-1)!} (1+\rho)^{n+m_n-1} \right].
\end{array}
\tag{A.119}
$$

Since,

$$
\left[n \frac{(n\rho)^{m_n-1}}{(m_n-1)!} m_n \right] / \left[\frac{n^{m_n} \quad n!}{(n+m_n-1)!} (1+\rho)^{n+m_n-1} \right] = \frac{(1+\rho) \quad m_n}{\rho \quad (n+m_n)} \quad \frac{\rho^{m_n} \quad m_n}{m_n!} \quad \frac{(n+m_n)!}{n! \quad (1+\rho)^{n+m_n}},
\tag{A.120}
$$

and because (1.2) implies $((1+\rho) \quad m_n) / (\rho \quad (n+m_n)) \to \frac{(1+\rho)\bar{m}}{\rho(1+\bar{m})}$ we will focus on finding the limit of $\frac{\rho^{m_n} \quad m_n}{m_n!} \quad \frac{(n+m_n)!}{n! \quad (1+\rho)^{n+m_n}}$.

After applying Stirling's approximation, $n! \cong \sqrt{2\pi n}\, n^n e^{-n}$, we have

$$
\begin{array}{l}
\frac{\rho^{m_n} \quad m_n}{m_n!} \quad \frac{(n+m_n)!}{n! \quad (1+\rho)^{n+m_n}} \\
\cong \frac{\rho^{m_n} \quad m_n}{\sqrt{2\pi} \quad \sqrt{m_n} \quad m_n{}^{m_n} \quad e^{-m_n}} \quad \frac{\sqrt{2\pi} \quad \sqrt{n+m_n} \quad (n+m_n)^{n+m_n} \quad e^{-(n+m_n)}}{\sqrt{2\pi} \quad \sqrt{n} \quad n^n \quad e^{-n} \quad (1+\rho)^{n+m_n}} \\
= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n+m_n}{n}} \sqrt{m_n} \left(\frac{\rho(n+m_n)}{(1+\rho)m_n} \right)^{m_n} \left(\frac{n+m_n}{n(1+\rho)} \right)^n \\
= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n+m_n}{n}} \sqrt{m_n} \quad exp\left\{ m_n \quad ln\left(\frac{\rho(n+m_n)}{(1+\rho)m_n} \right) + n \quad ln\left(\frac{n+m_n}{n(1+\rho)} \right) \right\}
\end{array}
\tag{A.121}
$$

From Lemma 1, for $0 < \varepsilon < 1/2$

$$
\begin{aligned}
&exp\left\{m_n \quad ln\left(\tfrac{\rho(n+m_n)}{(1+\rho)m_n}\right) + n \quad ln\left(\tfrac{n+m_n}{n(1+\rho)}\right)\right\} \\
&= exp\left\{n \quad \left[\bar{m} \quad ln\left(\tfrac{\rho(1+\bar{m})}{(1+\rho)\bar{m}}\right) + ln\left(\tfrac{1+\bar{m}}{1+\rho}\right)\right] + o\left(n^{1-\varepsilon}\right)\right\},
\end{aligned}
\tag{A.122}
$$

and

$$
\begin{aligned}
&\bar{m} \quad ln\left(\tfrac{\rho(1+\bar{m})}{(1+\rho)\bar{m}}\right) + ln\left(\tfrac{1+\bar{m}}{1+\rho}\right) = \bar{m} \quad ln\left(\tfrac{(1+\bar{m})}{(1+\rho)}\right) + \bar{m} \quad ln\left(\tfrac{\rho}{\bar{m}}\right) + ln\left(\tfrac{(1+\bar{m})}{(1+\rho)}\right) \\
&= (1+\bar{m}) \quad ln\left(\tfrac{(1+\bar{m})}{(1+\rho)}\right) + \bar{m} \quad ln\left(\tfrac{\rho}{\bar{m}}\right) \\
&= (1+\rho)\left[\left(\tfrac{1+\bar{m}}{1+\rho}\right) ln\left(\tfrac{(1+\bar{m})}{(1+\rho)}\right) + \tfrac{\bar{m}}{1+\rho} \quad ln\left(\tfrac{\rho}{\bar{m}}\right) - \tfrac{1}{1+\rho} \quad ln\left(\tfrac{\rho}{\bar{m}}\right) + \tfrac{1}{1+\rho} \quad ln\left(\tfrac{\rho}{\bar{m}}\right)\right] \\
&= (1+\rho)\left[\left(\tfrac{1+\bar{m}}{1+\rho}\right) ln\left(\tfrac{(1+\bar{m})}{(1+\rho)} \quad \tfrac{\rho}{\bar{m}}\right) - \tfrac{1}{1+\rho} \quad ln\left(\tfrac{\rho}{\bar{m}}\right)\right] \\
&= \left[(1+\bar{m}) \quad ln\left(\tfrac{(1+\bar{m})}{(1+\rho)} \quad \tfrac{\rho}{\bar{m}}\right) - ln\left(\tfrac{\rho}{\bar{m}}\right)\right] \\
&= \left[-(1+\bar{m}) \quad ln\left(\tfrac{\bar{m}\,(1+\rho)}{(1+\bar{m})\,\rho}\right) - ln\left(\tfrac{\rho}{\bar{m}}\right)\right].
\end{aligned}
\tag{A.123}
$$

Now we apply Taylor's approximation for $f(x) = -ln(x)$ at $x_0 = 1$ (note that $1 - \frac{\bar{m}(1+\rho)}{(1+\bar{m})\rho} = \frac{\rho-\bar{m}}{(1+\bar{m})\rho}$). Then, for $\bar{m} > 0$ and $k \geq 2$, $(1+\bar{m})\left(\frac{\rho-\bar{m}}{(1+\bar{m})\rho}\right)^k < \left(\frac{\rho-\bar{m}}{\rho}\right)^k$, and we have

$$
\begin{aligned}
&\left[-(1+\bar{m}) ln\left(\tfrac{(1+\rho)}{(1+\bar{m})} \quad \tfrac{\bar{m}}{\rho}\right) - ln\left(\tfrac{\rho}{\bar{m}}\right)\right] \\
&= (1+\bar{m})\left[\tfrac{\rho-\bar{m}}{(1+\bar{m})\rho} + \left(\tfrac{\rho-\bar{m}}{(1+\bar{m})\rho}\right)^2 \tfrac{1}{2} + \left(\tfrac{\rho-\bar{m}}{(1+\bar{m})\rho}\right)^3 \tfrac{1}{3} + \ldots\right] \\
&- \left[\tfrac{\rho-\bar{m}}{\rho} + \left(\tfrac{\rho-\bar{m}}{\rho}\right)^2 \tfrac{1}{2} + \left(\tfrac{\rho-\bar{m}}{\rho}\right)^3 \tfrac{1}{3} + \ldots\right] < 0.
\end{aligned}
\tag{A.124}
$$

For $\bar{m} = 0$, we apply L'Hospital's rule to $\bar{m} \quad ln\left(\frac{\rho(1+\bar{m})}{(1+\rho)\bar{m}}\right)$ and

$$\lim_{\bar{m}\to 0} ln\left(\frac{\rho(1+\bar{m})}{(1+\rho)\bar{m}}\right) / (1/\bar{m}) = \lim_{\bar{m}\to 0} 1/\left[1/\bar{m}+1\right] = 0. \tag{A.125}$$

Following (A.122)-(A.125) we have

$$exp\left\{m_n \quad ln\left(\frac{\rho(n+m_n)}{(1+\rho)m_n}\right) + n \quad ln\left(\frac{n+m_n}{n(1+\rho)}\right)\right\} \to exp\{-\infty\} = 0, \tag{A.126}$$

(1.2) and (A.121) with (A.126) gives

$$\begin{aligned} &\tfrac{1}{\sqrt{2\pi}} \quad \sqrt{\tfrac{n+m_n}{n}} \quad \sqrt{m_n} \quad / \quad exp\left\{-\left[m_n \quad ln\left(\tfrac{\rho(n+m_n)}{(1+\rho)m_n}\right) + n \quad ln\left(\tfrac{n+m_n}{n(1+\rho)}\right)\right]\right\} \\ &\to \tfrac{1}{\sqrt{2\pi}}\sqrt{1+\bar{m}} \quad \tfrac{\infty}{\infty}. \end{aligned} \tag{A.127}$$

Applying L'Hospital's rule as in (A.44) gives

$$\begin{aligned} &\tfrac{1}{\sqrt{2\pi}} \quad \sqrt{\tfrac{n+m_n}{n}} \quad \sqrt{m_n} \quad / \quad exp\left\{-\left[m_n \quad ln\left(\tfrac{\rho(n+m_n)}{(1+\rho)m_n}\right) + n \quad ln\left(\tfrac{n+m_n}{n(1+\rho)}\right)\right]\right\} \\ &\to \tfrac{\bar{m}}{\infty} = 0. \end{aligned} \tag{A.128}$$

Then, (A.119)-(A.127) gives

$$\begin{aligned} &\left[\sum_{k=0}^{m_n-1} n\tfrac{(n\rho)^k}{k!}\right] / factor_2 \le \left[n\tfrac{(n\rho)^{m_n-1}}{(m_n-1)!}m_n\right] / \left[\tfrac{n^{m_n}n!}{(n+m_n-1)!}(1+\rho)^{n+m_n-1}\right] \to 0 \\ &\Rightarrow \left[\sum_{k=0}^{m_n-1} n\tfrac{(n\rho)^k}{k!}\right] / factor_2 \to 0. \end{aligned} \tag{A.129}$$

Now we look at $\left[\sum_{k=s_n}^{n+m_n} c_n\right] / factor_2$. Contrary to ED|ED, we have $\frac{(1+\bar{m})\rho}{\bar{s}(1+\rho)} < 1$, and following (A.68) gives $\left[\sum_{k=s_n}^{n+m_n} c_n\right] / factor_2 \to 0$ here.

Following from (A.117) to here gives

$$\tilde{A}_n'^2 =\rightarrow \tfrac{1}{0+1+0} = 1, \tag{A.130}$$

and therefore (since $\varepsilon_1, \varepsilon_2 > 0$ are arbitrary) from (A.116) we get

$$\tilde{A}_n^2 \rightarrow 1/\mu - [\bar{m}(1+\rho)] / [\mu\rho(1+\bar{m})] \tag{A.131}$$

From (A.64), (A.105), and (A.130)

$$\begin{aligned} &\left[\sum_{k=\lceil \frac{\rho(1+\bar{m})}{1+\rho} n + n^{1-\varepsilon} \rceil}^{s_n - 1} E[W_n^{st}(s_n, m_n) \mid N_n^{st}(\tau_b) = k] \quad d_n \right. \\ &\left. + \sum_{k=s_n}^{n+m_n} E[W_n^{st}(s_n, m_n) \mid N_n^{st}(\tau_b) = k] \quad c_n \right] / denominator_1 \\ &\leq \left(\tfrac{n}{\mu m_n}\right) \left(1 - \tilde{A}_n'^2\right) \rightarrow 0, \end{aligned} \tag{A.132}$$

and we have

$$E[D_n^{rpl}(s_n, m_n)] = \tilde{A}_n + \tilde{B}_n \rightarrow 1/\mu - [\bar{m}(1+\rho)] / [\mu\rho(1+\bar{m})] .\square \tag{A.133}$$

Proof of Formulas for $\lim_{\mathbf{n}\to\infty} \mathbf{E}[\mathbf{D_n^{rpl}}(\mathbf{s_n}, \mathbf{m_n})]$ *for QED|ED*

Note that this is the case with $\bar{m} = \bar{s} + \bar{s}/\rho - 1$; when this equality is satisfied, $\lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)]$ formulas for both ED|ED and QD|ED give the same value that could be expressed as either way, and therefore we conclude that $E[D_n^{rpl}(s_n, m_n)] \rightarrow 1/\bar{s}\mu - 1/\lambda \quad = \quad 1/\mu - [\bar{m}(1+\rho)] / [\mu\rho(1+\bar{m})] .\square$

Proof of Formulas for $\lim_{\mathbf{n}\to\infty} \mathbf{E}[\mathbf{D_n^{rpl}}(\mathbf{s_n}, \mathbf{m_n})]$ *for All Other Capacity Level Preferences*

QED|QD and QD|QD: From (A.27) we know

$$E[D_n^{rpl}(s_n, m_n)] \leq \left(\frac{n}{\mu m_n}\right) \left[\sum_{k \geq m_n} (n - [k - m_n]^+) p_{n,k}^{st}\right] / \left[\sum_k (n - [k - m_n]^+) p_{n,k}^{st}\right]. \tag{A.134}$$

Because our system is a closed network, the quantity at the right side is not equal to the steady state probability of broken units being more than the spares. Due to our fluid limit result, we know that the changes in the system state are of $\sqrt{n}$ sensitivity, and it does not matter at the limit when it is possible to consider a corresponding event that the scaled broken units process in (A.1) has a greater value than the limiting scaled value of spares divided by n, $\bar{m}$. In similar fashion to Theorem 9 in [7], it can be shown that

$$\lim_{n\to\infty} \left[\sum_{k \geq m_n} (n - [k - m_n]^+) p_{n,k}^{st}\right] / \left[\sum_k (n - [k - m_n]^+) p_{n,k}^{st}\right] = \lim_{n\to\infty} P(N_n^{st} \geq m_n).$$

Then from Theorem 12 we have $\lim_{n\to\infty} P(N_n^{st} \geq m_n) = \lim_{n\to\infty} P(\frac{N_n^{st}}{n} \geq \frac{m_n}{n}) \to P(b \geq \bar{m}) = 0$, and this implies $\lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)] = 0$.

QD|QED and QED|QED: Let b as described in Theorem 12. Note that in these cases $b = \bar{m}$. As done in QD|ED it could be proven that

$$\left[\sum_{k=\lfloor b.n - n^{1-\varepsilon}\rfloor}^{\lceil b.n + n^{1-\varepsilon}\rceil} (n - [k - m_n]^+) p_{n,k}^{st}\right] / \left[\sum_k (n - [k - m_n]^+) p_{n,k}^{st}\right] \to 1,$$

and therefore we will have

$$\begin{aligned} &\lim_{n\to\infty} \left(\frac{\lfloor b.n - n^{1-\varepsilon}\rfloor + 1 - m_n}{\mu \lfloor b.n + n^{1-\varepsilon}\rfloor}\right) \leq \lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)] \leq \lim_{n\to\infty} \left(\frac{\lfloor b.n + n^{1-\varepsilon}\rfloor + 1 - m_n}{\mu \lfloor b.n - n^{1-\varepsilon}\rfloor}\right) \\ &\Rightarrow \lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)] = 0. \square \end{aligned} \tag{A.135}$$

A.2 Proof of Theorem 2

As can be seen in Theorem 1 $\lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)]$ for ED|ED does not include $\bar{m}$ and the limit for QD|ED does not include $\bar{s}$, so the solution for the cost optimization will have $\bar{m} = 0$ for ED|ED, and from Figure 1.2 it can be seen to reside on two connected line segments; $\bar{m} = 0, \bar{s} \,\epsilon\, (0, \frac{\lambda}{\lambda+\mu})$ and $\bar{m} = \bar{s} + \frac{\bar{s}}{\rho} - 1, \bar{s} \,\epsilon\, (\frac{\lambda}{\lambda+\mu}, \frac{\lambda}{\mu})$. The first line segment is at the bottom of the ED|ED region and the second one is the QED|ED region itself in Figure 1.2, here on we will refer to them as LS 1 and LS 2, respectively.

At LS 1, we have $\lim_{n\to\infty} \frac{C(s_n,m_n)}{n} = w\bar{s} + p\left(\frac{1}{\mu\bar{s}} - \frac{1}{\lambda}\right)$, and after taking the derivatives w.r.t $\bar{s}$

$$\begin{aligned} \left(w\bar{s} + p\left(\tfrac{1}{\mu\bar{s}} - \tfrac{1}{\lambda}\right)\right)' &= w - \tfrac{p}{\mu}\left(\tfrac{1}{(\bar{s})^2}\right) \\ \left(w - \tfrac{p}{\mu}\left(\tfrac{1}{(\bar{s})^2}\right)\right)' &= \tfrac{p}{\mu}\left(\tfrac{2}{(\bar{s})^3}\right) > 0 \end{aligned} \tag{A.136}$$

for $\bar{s} > 0$ since $\bar{s} = 0$ will give $+\infty$ cost it is clearly not the solution. The cost function is convex in this region, and therefore the first-order solution is optimal, which gives $(\bar{s} = \sqrt{p/(w\mu)}, \bar{m} = 0)$. However, this solution is valid only if $\bar{s} = \sqrt{p/(w\mu)} \leq \lambda/(\lambda+\mu)$.

Similarly, for LS 2 from Remark 1 we have $\lim_{n\to\infty} \frac{C(s_n,m_n)}{n} = w\bar{s} + c\left[\bar{s} + \frac{\bar{s}}{\rho} - 1\right] + p\left(\frac{1}{\mu\bar{s}} - \frac{1}{\lambda}\right)$, and after taking the derivatives w.r.t $\bar{s}$ we get

$$\begin{aligned} \left(w\bar{s} + c\left[\bar{s} + \tfrac{\bar{s}}{\rho} - 1\right] + p\left(\tfrac{1}{\mu\bar{s}} - \tfrac{1}{\lambda}\right)\right)' &= w + c + \tfrac{c}{\rho} - \tfrac{p}{\mu}\left(\tfrac{1}{(\bar{s})^2}\right) \\ \left(w + c + \tfrac{c}{\rho} - \tfrac{p}{\mu}\left(\tfrac{1}{(\bar{s})^2}\right)\right)' &= \tfrac{p}{\mu}\left(\tfrac{2}{(\bar{s})^3}\right) > 0. \end{aligned} \tag{A.137}$$

The cost function is convex in this region, and therefore the first-order solution

is optimal, which gives ($\bar{s} = \sqrt{p/\left[\mu(w+c+\frac{c}{\rho})\right]}, \bar{m} = \bar{s} + \frac{\bar{s}}{\rho} - 1$). However, this solution is valid only if $\frac{\lambda}{\lambda+\mu} \leq \bar{s} = \sqrt{p/\left[\mu(w+c+\frac{c}{\rho})\right]} \leq \frac{\lambda}{\mu}$.

Now suppose both conditions obtained for $\bar{s}$ from LS 1 and LS 2 hold. Then this would mean $p/(w\mu) \leq \left(\frac{\lambda}{\lambda+\mu}\right)^2 \leq p/\left[\mu(w+c+\frac{c}{\rho})\right] \Rightarrow \frac{1}{w} \leq 1/\left[w+c+\frac{c}{\rho}\right]$, which is a contradiction. Hence, both conditions cannot simultaneously hold. Thus, there could be four cases:

1) Condition for LS 1 holds only: Then, we have $\sqrt{p/\left[\mu(w+c+\frac{c}{\rho})\right]} < \sqrt{\frac{p}{w\mu}} < \frac{\lambda}{\lambda+\mu}$. From convexity, $\frac{\lambda}{\lambda+\mu}$ is the best solution in LS 2, and since it is also in LS 1 and $\sqrt{\frac{p}{w\mu}}$ is better, $\bar{s} = \sqrt{\frac{p}{w\mu}}, \bar{m} = 0$ is the solution.

2) Condition for LS 2 holds only: $\frac{\lambda}{\lambda+\mu}$ is the best answer for LS 1, and since $\bar{s} = \sqrt{p/\left[\mu(w+c+\frac{c}{\rho})\right]}$ is better, $\bar{s} = \sqrt{p/\left[\mu(w+c+\frac{c}{\rho})\right]}, \bar{m} = \bar{s} + \frac{\bar{s}}{\rho} - 1$ is the solution.

3) Both conditions are not satisfied and $\frac{\lambda}{\mu} < \sqrt{p/\left[\mu(w+c+\frac{c}{\rho})\right]} < \sqrt{\frac{p}{w\mu}}$: $\bar{s} = \frac{\lambda}{\lambda+\mu}$ is the best in LS 1 and $\bar{s} = \frac{\lambda}{\mu}$ is the best in LS 2. Since LS 2 also includes $\bar{s} = \frac{\lambda}{\lambda+\mu}$, $\bar{s} = \frac{\lambda}{\mu}, \bar{m} = \bar{s} + \frac{\bar{s}}{\rho} - 1$ is the solution.

4) Both conditions are not satisfied and $\sqrt{p/\left[\mu(w+c+\frac{c}{\rho})\right]} < \frac{\lambda}{\lambda+\mu} < \sqrt{\frac{p}{w\mu}}$: $\bar{s} = \frac{\lambda}{\lambda+\mu}$ is the best in LS 1 and LS 2 , therefore $\bar{s} = \frac{\lambda}{\lambda+\mu}, \bar{m} = \bar{s} + \frac{\bar{s}}{\rho} - 1$ is the solution. □

A.3 Proof of Theorem 3

A.3.1 Proof of Theorem 3 for $\mathbf{T}\epsilon(\mathbf{0},\infty)$

Proof of Theorem 3 for $\mathbf{T}\epsilon(\mathbf{0},\infty)$*, ED|ED*

Proof of Theorem 3 for $\mathbf{T}\epsilon(\mathbf{0},\infty)$*, ED|ED, and* $\mathbf{m_n \geq s_n}$ From (1.1), (1.2), and similar to (A.27) we can write $P(D_n^{rpl}(s_n, m_n) > T)$ as follows:

$$\begin{aligned} & P(D_n^{rpl}(s_n, m_n) > T) \\ & = \left(\sum_{k \geq m_n} P(W_n^{st}(s_n, m_n) > T | N_n^{st}(\tau_b) = k)(n - [k - m_n]^+) p_{n,k}^{st} \right) \\ & / \left(\sum_k (n - [k - m_n]^+) p_{n,k}^{st} \right) \end{aligned} \tag{A.138}$$

Note that since we have more spare units than our repairmen, whenever a queue for replacement units forms, all repairmen will be busy while the queue is nonempty.

Since a space at the demand base that experiences a breakdown when there are k, $k \geq m_n$, units already broken down has to wait for $k + 1 - m_n$ repairs, from the independence assumptions in the model, the waiting time for repair will be the sum of $k + 1 - m_n$ independent exponentially distributed random variables with mean $1/\mu s_n$. Let a_n, b_n and c_n be as in Preliminaries-ED|ED-1) a), then we have

$$\begin{aligned} & P(D_n^{rpl}(s_n, m_n) > T) = \left[\sum_{k=m_n}^{n+m_n} \sum_{r=0}^{k-m_n} exp\{-\mu s_n T\} \quad \frac{(\mu s_n T)^r}{r!} \quad c_n \right] \\ & / \left[\sum_{k=0}^{s_n - 1} a_n + \sum_{k=s_n}^{m_n - 1} b_n + \sum_{k=m_n}^{n+m_n} c_n \right] \end{aligned} \tag{A.139}$$

Let $factor_1 = \left(\frac{\rho}{s_n}\right)^{n+m_n-1} \frac{n^{m_n}}{s_n!} \quad n! \quad s_n^{s_n} \quad exp\{\frac{s_n}{\rho}\}$ as in (A.30). Then, we can write

$$
\begin{aligned}
&P(D_n^{rpl}(s_n, m_n) > T) \\
&= \lim_{n\to\infty} \left[\sum_{k=m_n}^{n+m_n} \sum_{r=0}^{k-m_n} exp\{-\mu s_n T\} \tfrac{(\mu s_n T)^r}{r!} c_n / factor_1 \right] \\
&/ \left[\sum_{k=0}^{s_n-1} a_n / factor_1 + \sum_{k=s_n}^{m_n-1} b_n / factor_1 + \sum_{k=m_n}^{n+m_n} c_n / factor_1 \right].
\end{aligned}
\tag{A.140}
$$

Note that $\sum_{k=0}^{s_n-1} a_n/factor_1$ is the left side of (A.38) and $\sum_{k=s_n}^{m_n-1} b_n/factor_1$ is the left side of (A.45). Then, from (A.30)-(A.45), we have

$$
\lim_{n\to\infty} \sum_{k=0}^{s_n-1} a_n/factor_1 = \lim_{n\to\infty} \sum_{k=s_n}^{m_n-1} b_n/factor_1 = 0
$$

. Moreover, from Preliminaries-ED|ED-1 e), CLT, and f) *ii*), $\sum_{k=m_n}^{n+m_n} c_n/factor_1 = P(Poisson(s_n/\rho) \le n-1) \to 1$ as $n \to \infty$. From Preliminaries-ED|ED-1 e),

$$
\begin{aligned}
&\sum_{k=m_n}^{n+m_n} \sum_{r=0}^{k-m_n} exp\{-\mu s_n T\} \quad \tfrac{(\mu s_n T)^r}{r!} \quad c_n/factor_1 \\
&= \sum_{k=m_n}^{n+m_n-1} \sum_{r=0}^{k-m_n} exp\{-\mu s_n T\} \quad \tfrac{(\mu s_n T)^r}{r!} \quad \tfrac{1}{(n+m_n-k-1)!} \quad \left(\tfrac{s_n}{\rho}\right)^{n+m_n-k-1} \quad exp\{-\tfrac{s_n}{\rho}\}.
\end{aligned}
\tag{A.141}
$$

Let $k' = n + m_n - 1 - k$; then, $k - m_n = n + m_n - 1 - m_n - k'$, and we have

$$
\begin{aligned}
&\sum_{k=m_n}^{n+m_n-1} \sum_{r=0}^{k-m_n} exp\{-\mu s_n T\} \quad \tfrac{(\mu s_n T)^r}{r!} \quad \tfrac{1}{(n+m_n-k-1)!} \quad \left(\tfrac{s_n}{\rho}\right)^{n+m_n-k-1} \quad exp\{-\tfrac{s_n}{\rho}\} \\
&= \sum_{k'=0}^{n-1} \sum_{r=0}^{n-1-k'} exp\{-\mu s_n T\} \quad \tfrac{(\mu s_n T)^r}{r!} \quad \tfrac{1}{k'!} \quad \left(\tfrac{s_n}{\rho}\right)^{k'} exp\{-\tfrac{s_n}{\rho}\} = P(X_n + Y_n \le n-1)
\end{aligned}
\tag{A.142}
$$

where $X_n \sim Poisson(\mu s_n T)$ and $Y_n \sim Poisson\left(\frac{s_n}{\rho}\right)$. Since sum of two Poisson random variables will have a Poisson distribution with a rate equal to sum of the rates of the two random variables, we continue the last equation as

$$P(X_n + Y_n \leq n-1) = P(H_n \leq n-1) \tag{A.143}$$

where $H_n \sim Poisson(\mu s_n T + \frac{s_n}{\rho})$. We will apply CLT to take the limit as $n \to \infty$ of the last probability above. From (1.2),

$$\begin{aligned}
&\left[n - 1 - \mu s_n T - \tfrac{s_n}{\rho}\right] / \sqrt{\mu s_n T + \tfrac{s_n}{\rho}} \\
&= \left[n\left(1 - \mu\bar{s}T - \tfrac{\bar{s}}{\rho}\right) - \sqrt{n}\left(\mu\hat{s}_n T + \tfrac{\hat{s}_n}{\rho}\right) - 1\right] / \sqrt{n\left(\mu\bar{s}T + \tfrac{\bar{s}}{\rho}\right) + \sqrt{n}\left(\mu\hat{s}_n T + \tfrac{\hat{s}_n}{\rho}\right)} \\
&\to \begin{cases} +\infty & for \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} > 0 \\ -\hat{s}/\bar{s} & for \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} = 0 \\ -\infty & for \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} < 0 \end{cases},
\end{aligned} \tag{A.144}$$

and we have

$$\lim_{n\to\infty} P(X_n + Y_n \leq n-1) = \begin{cases} 1 & for \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} > 0 \\ \Phi\left(-\hat{s}/\bar{s}\right) & for \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} = 0 \\ 0 & for \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} < 0 \end{cases}. \tag{A.145}$$

Starting from (A.138) we conclude

$$\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > T)$$

$$= \lim_{n\to\infty} \left[\sum_{k=m_n}^{n+m_n} \sum_{r=0}^{k-m_n} exp\{-\mu s_n T\} \tfrac{(\mu s_n T)^r}{r!} c_n / factor_1 \right]$$

$$/ \left[\sum_{k=0}^{s_n-1} a_n / factor_1 + \sum_{k=s_n}^{m_n-1} b_n / factor_1 + \sum_{k=m_n}^{n+m_n} c_n / factor_1 \right] \tag{A.146}$$

$$= \lim_{n\to\infty} P(H_n \le n-1)$$

$$= \begin{cases} 1 & for \quad ED|ED, m_n \ge s_n \quad and \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} > 0 \\ \Phi\left(-\hat{s}/\bar{s}\right) & for \quad ED|ED, m_n \ge s_n \quad and \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} = 0 \\ 0 & for \quad ED|ED, m_n \ge s_n \quad and \quad 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} < 0 \end{cases} .\square$$

Proof of Theorem 3 for $\mathbf{T}\epsilon(\mathbf{0}, \infty)$*, ED|ED, and* $\mathbf{s_n \ge m_n}$ Let F_n be defined as in Preliminaries -ED|ED-2 h),and a_n, c_n, d_n, be as in Preliminaries-ED|ED-1) and -2)a).

Then, we define

$$
\begin{aligned}
& P(D_n^{rpl}(s_n, m_n) > T) \\
& \equiv \left[\sum_{k=m_n}^{s_n-1} P(\{W_n^{st}(s_n, m_n) > T\} \cap F_n \mid N_n^{st}(\tau_b) = k) \quad d_n \right] \\
& / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n \right] \\
& + \left[\sum_{k=s_n}^{n+m_n} P(\{W_n^{st}(s_n, m_n) > T\} \cap F_n \mid N_n^{st}(\tau_b) = k) \quad c_n \right] \\
& / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n \right] \\
& + \left[\sum_{k=m_n}^{s_n-1} P(\{W_n^{st}(s_n, m_n) > T\} \cap F_n' \mid N_n^{st}(\tau_b) = k) \quad d_n \right] \\
& / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n \right] \\
& \left[\sum_{k=s_n}^{n+m_n} P(\{W_n^{st}(s_n, m_n) > T\} \cap F_n' \mid N_n^{st}(\tau_b) = k) \quad c_n \right] \\
& / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n \right].
\end{aligned}
\tag{A.147}
$$

Since $P(\{W_n^{st}(s_n, m_n) > T\} \mid F_n, N_n^{st}(\tau_b) = k) \leq 1$, similar to for $\tilde{A}_n$ in (A.65)-(A.69),

$$\left[\sum_{k=m_n}^{s_n-1} P(\{W_n^{st}(s_n,m_n)>T\}\cap F_n \mid N_n^{st}(\tau_b)=k) \quad d_n\right]$$
$$/\left[\sum_{k=0}^{m_n-1} a_n+\sum_{k=m_n}^{s_n-1} d_n+\sum_{k=s_n}^{n+m_n} c_n\right]$$
$$+\left[\sum_{k=m_n}^{s_n-1} P(\{W_n^{st}(s_n,m_n)>T\}\cap F_n' \mid N_n^{st}(\tau_b)=k) \quad d_n\right] \tag{A.148}$$
$$/\left[\sum_{k=0}^{m_n-1} a_n+\sum_{k=m_n}^{s_n-1} d_n+\sum_{k=s_n}^{n+m_n} c_n\right]\to 0;$$

and similar to (A.71)-(A.74) we have

$$\left[\sum_{k=s_n}^{n+m_n} P(\{W_n^{st}(s_n,m_n)>T\}\cap F_n \mid N_n^{st}(\tau_b)=k) \quad c_n\right]$$
$$/\left[\sum_{k=0}^{m_n-1} a_n+\sum_{k=m_n}^{s_n-1} d_n+\sum_{k=s_n}^{n+m_n} c_n\right]\to 0. \tag{A.149}$$

Continuing in the same manner as in (A.75)-(A.80),

$$\lim_{n\to\infty}\left[\sum_{k=s_n}^{n+m_n} P(\{W_n^{st}(s_n,m_n)>T\}\cap F_n' \mid N_n^{st}(\tau_b)=k) \quad c_n\right]$$
$$/\left[\sum_{k=0}^{m_n-1} a_n+\sum_{k=m_n}^{s_n-1} d_n+\sum_{k=s_n}^{n+m_n} c_n\right]$$
$$=\lim_{n\to\infty}\left[\sum_{k=s_n}^{n+m_n}\sum_{r=0}^{k-m_n} exp\{-\mu s_n T\}\tfrac{(\mu s_n T)^r}{r!} \quad c_n\right]/\left[\sum_{k=0}^{m_n-1} a_n+\sum_{k=m_n}^{s_n-1} d_n+\sum_{k=s_n}^{n+m_n} c_n\right]. \tag{A.150}$$

Then, we can write

$$\left[\sum_{k=s_n}^{n+m_n}\sum_{r=0}^{k-m_n} exp\{-\mu s_n T\}\frac{(\mu s_n T)^r}{r!} \quad c_n\right] / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right].$$

$$= \left[\sum_{k=m_n}^{n+m_n}\sum_{r=0}^{k-m_n} exp\{-\mu s_n T\}\frac{(\mu s_n T)^r}{r!} \quad c_n\right] / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right]. \quad \text{(A.151)}$$

$$- \left[\sum_{k=m_n}^{s_n-1}\sum_{r=0}^{k-m_n} exp\{-\mu s_n T\}\frac{(\mu s_n T)^r}{r!} \quad c_n\right] / \left[\sum_{k=0}^{m_n-1} a_n + \sum_{k=m_n}^{s_n-1} d_n + \sum_{k=s_n}^{n+m_n} c_n\right].$$

Let $factor_1$ be as in (A.30). We can write, from (A.83)-(A.93) and (A.141)-(A.142),

$$\begin{aligned}
&\left[\sum_{k=m_n}^{n+m_n}\sum_{r=0}^{k-m_n} exp\{-\mu s_n T\}\frac{(\mu s_n T)^r}{r!} \quad c_n/factor_1\right] \\
&/ \left[\sum_{k=0}^{m_n-1} a_n/factor_1 + \sum_{k=m_n}^{s_n-1} d_n/factor_1 + \sum_{k=s_n}^{n+m_n} c_n/factor_1\right] \qquad \text{(A.152)} \\
&= \begin{cases} 1 & for \quad 1-\mu\bar{s}T-\frac{\bar{s}}{\rho}>0 \\ \Phi\left(-\hat{s}/\bar{s}\right) & for \quad 1-\mu\bar{s}T-\frac{\bar{s}}{\rho}=0 \\ 0 & for \quad 1-\mu\bar{s}T-\frac{\bar{s}}{\rho}<0 \end{cases}.
\end{aligned}$$

Moreover, from Preliminaries-ED|ED-1)e) and (A.83)-(A.93) again

$$\begin{aligned}
&\left[\sum_{k=m_n}^{s_n-1}\sum_{r=0}^{k-m_n} exp\{-\mu s_n T\}\frac{(\mu s_n T)^r}{r!} \quad c_n/factor_1\right] \\
&/ \left[\sum_{k=0}^{m_n-1} a_n/factor_1 + \sum_{k=m_n}^{s_n-1} d_n/factor_1 + \sum_{k=s_n}^{n+m_n} c_n/factor_1\right] \\
&\leq \left[\sum_{k=m_n}^{s_n-1} c_n/factor_1\right] / \left[\sum_{k=0}^{m_n-1} a_n/factor_1 + \sum_{k=m_n}^{s_n-1} d_n/factor_1 + \sum_{k=s_n}^{n+m_n} c_n/factor_1\right] \\
&\to \lim_{n\to\infty} P(n+m_n-s_n < Poisson(s_n/\rho) < n-1); \qquad \text{(A.153)}
\end{aligned}$$

and since

$$\left[n + m_n - s_n - s_n/\rho\right] / \sqrt{s_n/\rho} = \left[n(1 + \bar{m} - \bar{s} - \bar{s}/\rho) + \sqrt{n}(\hat{m} - \hat{s} - \hat{s}/\rho)\right] / \sqrt{s_n/\rho}$$

with inequalities for ED|ED from Theorem 12 and Preliminaries-ED|ED-1 f)ii) gives (using CLT with $Z_n \sim Normal(0,1)$)

$$\lim_{n\to\infty} P(n + m_n - s_n < Poisson(s_n/\rho) < n - 1) = \Phi\left(+\infty < Z_n < +\infty\right) = 0, \tag{A.154}$$

we conclude that

$$P(D_n^{rpl}(s_n, m_n) > T) = \begin{cases} 1 & for \quad ED|ED, s_n > m_n, 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} > 0 \\ \Phi\left(-\hat{s}/\bar{s}\right) & for \quad ED|ED, s_n > m_n, 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} = 0 \\ 0 & for \quad ED|ED, s_n > m_n, 1 - \mu\bar{s}T - \frac{\bar{s}}{\rho} < 0 \end{cases} .\square \tag{A.155}$$

Proof of Theorem 3 for $\mathbf{T}\epsilon(\mathbf{0}, \infty)$*, QED|QD-ED|QD-ED|QED, QED|QED, QD|QED, QD|QD*

Suppose $\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > T) > 0$; from Theorem 1 we know that

$$\lim_{n\to\infty} E[D_n^{rpl}(s_n, m_n)] = 0$$

for these cases. Since it leads to a contradiction,

$$\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > T) = 0$$

here.□

A.3.2 Proof of Theorem 3 for $\mathbf{T} = \mathbf{0}$

Proof of Theorem 3 for $\mathbf{T} = \mathbf{0}$*, ED|ED, QED|ED, and QD|ED*

Note that when $T = 0$, we are actually looking at the probability of delay until replacement, and it considers the event the number of broken units is greater than

the spares at the time of a random breakdown; or in other words, it considers an event that after a breakdown, a space at the demand base will see the system at a state such that the number of broken units is more than the number of spares. Because our system is a closed network, this event is not equal to the steady state probability of broken units being more than the spares. Due to our fluid limit result, we know that the changes in the system state (number of broken units) are of $\sqrt{n}$ sensitivity, and it does not matter at the limit when it is possible to consider a corresponding event that the scaled broken machines process in (A.1) has a greater value than the limiting scaled value of spares divided by n, $\bar{m}$. In similar fashion to Theorem 9 in [7], it can be shown that

$$\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > 0) = \lim_{n\to\infty} P(N_n^{st}(s_n, m_n) \geq m_n) \tag{A.156}$$

From Theorem 12 we have $\lim_{n\to\infty} P(N_n^{st}(s_n, m_n) \geq m_n) = \lim_{n\to\infty} P(N_n^{st}(s_n, m_n)/n \geq m_n/n) \to P(b \geq \bar{m}) = 1$, and this implies $\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > 0) = 1$.

Proof of Theorem 3 for $\mathbf{T} = \mathbf{0}$*, QD|QD*

From Theorem 12

$$\begin{aligned} &\lim_{n\to\infty} P(D_n^{rpl}(s_n, m_n) > 0) = \lim_{n\to\infty} P(N_n^{st}(s_n, m_n) \geq m_n) \\ &= \lim_{n\to\infty} P(N_n^{st}(s_n, m_n)/n \geq m_n/n) \to P(b \geq \bar{m}) = 0. \end{aligned} \tag{A.157}$$

Proof of Theorem 3 for $\mathbf{T} = \mathbf{0}$*, QED|QED*

We separate the proof into four main cases: $\hat{m} > \hat{s}$ and $\hat{s} \neq 0$, $\hat{m} > \hat{s}$ and $\hat{s} = 0$, $\hat{m} < \hat{s}$, and $\hat{m} = \hat{s}$.

Before we study the cases, we will give a useful lemma.

Lemma 2. *Let a_n and c_n be sequences of real numbers with $a_n \to a$ and $c_n \to c$, and b a constant. Then,*

$$e^{-a_n b\sqrt{n}}\left(1+\tfrac{a_n}{\sqrt{n}}\right)^{bn+c_n\sqrt{n}} \to e^{-\frac{a^2 b}{2}+ac} \tag{A.158}$$

Proof of Lemma 2

$$e^{-a_n b\sqrt{n}}\left(1+\tfrac{a_n}{\sqrt{n}}\right)^{bn+c_n\sqrt{n}} = e^{-a_n b\sqrt{n}} e^{(bn+c_n\sqrt{n})ln\left(1+\frac{a_n}{\sqrt{n}}\right)} \tag{A.159}$$

Using Taylor's approximation for $f(x) = ln(1+x)$ at the point $x_0 = 0$ and the definition; a function $f(n)$ is $o(n^x)$ if $\frac{f(n)}{n^x} \to 0$ as $n \to \infty$, we have

$$\begin{aligned} & e^{-a_n b\sqrt{n}} e^{(bn+c_n\sqrt{n})ln\left(1+\frac{a_n}{\sqrt{n}}\right)} = e^{-a_n b\sqrt{n}} e^{(bn+c_n\sqrt{n})\left(\frac{a_n}{\sqrt{n}}-\frac{a_n^2}{2n}+o\left(\frac{1}{n}\right)\right)} \\ & = e^{-a_n b\sqrt{n}} e^{a_n b\sqrt{n}-\frac{a_n^2 b}{2}+o(1)+a_n c_n-\frac{a_n^2 c_n}{2\sqrt{n}}+o\left(\frac{1}{\sqrt{n}}\right)} \to e^{-\frac{a^2 b}{2}+ac}. \quad \square \end{aligned} \tag{A.160}$$

Case 1) $\hat{m} > \hat{s}$ and $\hat{s} \neq 0$: For sufficiently large n, from (1.2) and $\bar{m} = \bar{s} = \rho$ for QED|QED, this leads to $m_n > s_n$.

From (1.1) and similar to above, we have

$$\begin{aligned} & P(D_n^{rpl}(s_n, m_n) > 0) \\ & = \frac{\sum_{k=m_n}^{n+m_n}(n+m_n-k)\frac{n^{m_n}}{s_n!}\frac{n!}{(n+m_n-k)!}s_n{}^{s_n}\left(\frac{\rho}{s_n}\right)^k}{\sum_{k=0}^{s_n-1} n\frac{n^k}{k!}\rho^k + \sum_{k=s_n}^{m_n-1} n\frac{n^k}{s_n!}s_n{}^{s_n}\left(\frac{\rho}{s_n}\right)^k + \sum_{k=m_n}^{n+m_n}(n+m_n-k)\frac{n^{m_n}}{s_n!}\frac{n!}{(n+m_n-k)!}s_n{}^{s_n}\left(\frac{\rho}{s_n}\right)^k} \\ & = \frac{C_n}{A_n+B_n+C_n} = \left(1+\frac{B_n}{C_n}+\frac{A_n}{C_n}\right)^{-1} \end{aligned} \tag{A.161}$$

where we define A_n, B_n, and C_n as

$$A_n \equiv \sum_{k=0}^{s_n-1} n\frac{n^k}{k!}\,\rho^k, B_n \equiv \sum_{k=s_n}^{m_n-1} n\frac{n^k}{s_n!}\,{s_n}^{s_n}\,\left(\frac{\rho}{s_n}\right)^k,$$
$$C_n \equiv \sum_{k=m_n}^{n+m_n}(n+m_n-k)\frac{n^{m_n}}{s_n!}\,\frac{n!}{(n+m_n-k)!}\,{s_n}^{s_n}\,\left(\frac{\rho}{s_n}\right)^k \tag{A.162}$$

Now we look at the values of A_n, B_n, and C_n. Notice that

$$A_n = ne^{-n\rho}e^{n\rho}\sum_{k=0}^{s_n-1}\frac{(n\rho)^k}{k!} = ne^{n\rho}\sum_{k=0}^{s_n-1}e^{-n\rho}\frac{(n\rho)^k}{k!} = ne^{n\rho}P(X_n \le s_n-1) \tag{A.163}$$

where $X_n \sim Poisson(n\rho)$. Moreover, we have

$$B_n = n\frac{{s_n}^{s_n}}{s_n!}\sum_{k=s_n}^{m_n-1}\left(\frac{n\rho}{s_n}\right)^k = n\frac{{s_n}^{s_n}}{s_n!}\left(\sum_{k=0}^{m_n-1}\left(\frac{n\rho}{s_n}\right)^k - \sum_{k=0}^{s_n-1}\left(\frac{n\rho}{s_n}\right)^k\right) \tag{A.164}$$

Multiplying and dividing C_n by ${s_n}^{n+m_n-1}$, ρ^{n+m_n-1}, and $e^{-\frac{s_n}{\rho}}$ and reindexing our summation, we have

$$\begin{aligned}
C_n &= \frac{n^{m_n}}{s_n!}\,n!\,{s_n}^{s_n}\sum_{k=m_n}^{n+m_n}(n+m_n-k)\,\frac{1}{(n+m_n-k)!}\,\left(\frac{s_n}{\rho}\right)^{-k}\\
&= \frac{n^{m_n}}{s_n!}\,n!\,{s_n}^{s_n}\,\frac{\rho^{n+m_n-1}}{{s_n}^{n+m_n-1}}\,e^{\frac{s_n}{\rho}}\sum_{k=m_n}^{n+m_n-1}e^{-\frac{s_n}{\rho}}\left(\frac{s_n}{\rho}\right)^{n+m_n-k-1}\,\frac{1}{(n+m_n-k-1)!}\\
&= \frac{n^{m_n}}{s_n!}\,n!\,{s_n}^{s_n}\left(\frac{\rho}{s_n}\right)^{n+m_n-1}\,e^{\frac{s_n}{\rho}}\sum_{j=0}^{n-1}e^{-\frac{s_n}{\rho}}\left(\frac{s_n}{\rho}\right)^{j}\,\frac{1}{j!}\\
&= \frac{n^{m_n}}{s_n!}\,n!\,{s_n}^{s_n}\left(\frac{\rho}{s_n}\right)^{n+m_n-1}\,e^{\frac{s_n}{\rho}}\,P(Y_n \le n-1),
\end{aligned} \tag{A.165}$$

where $Y_n \sim Poisson\left(\frac{s_n}{\rho}\right)$. By (A.161)–(A.165), we have

$$
\begin{aligned}
&P(D_n^{rpl}(s_n, m_n) > 0) \\
&= \left(1 + \frac{n \frac{{s_n}^{s_n}}{s_n!} \left(\sum_{k=0}^{m_n-1} \left(\frac{n\rho}{s_n}\right)^k - \sum_{k=0}^{s_n-1} \left(\frac{n\rho}{s_n}\right)^k \right)}{\frac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n} \left(\frac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho}} P(Y_n \le n-1)} + \frac{n e^{n\rho} P(X_n \le s_n - 1)}{\frac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n} \left(\frac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho}} P(Y_n \le n-1)} \right)^{-1} \\
&= \left(1 + \left(\frac{s_n}{\rho}\right)^{n+m_n-1} \frac{1}{n^{m_n-1}\, n!\, e^{\frac{s_n}{\rho}}} \left(\sum_{k=0}^{m_n-1} \left(\frac{n\rho}{s_n}\right)^k - \sum_{k=0}^{s_n-1} \left(\frac{n\rho}{s_n}\right)^k \right) \frac{1}{P(Y_n \le n-1)} \right. \\
&\left. + \left(\frac{s_n}{\rho}\right)^{n+m_n-1} \frac{s_n! e^{n\rho - \frac{s_n}{\rho}} P(X_n \le s_n - 1)}{n^{m_n-1}\, n!\, {s_n}^{s_n}\, P(Y_n \le n-1)} \right)^{-1}
\end{aligned}
\tag{A.166}
$$

We divide the expression above into four parts and take the limit as $n \to \infty$ of each part to obtain the limit of the whole expression. These parts are:

$$
\begin{aligned}
&P(X_n \le s_n - 1), \quad P(Y_n \le n-1) \\
&\left(\frac{s_n}{\rho}\right)^{n+m_n-1} \frac{1}{n^{m_n-1}\, n!\, e^{\frac{s_n}{\rho}}} \left(\sum_{k=0}^{m_n-1} \left(\frac{n\rho}{s_n}\right)^k - \sum_{k=0}^{s_n-1} \left(\frac{n\rho}{s_n}\right)^k \right), \\
&\text{and} \quad \frac{s_n!}{n^{m_n-1}} \frac{1}{n!} \frac{1}{{s_n}^{s_n}} \left(\frac{s_n}{\rho}\right)^{n+m_n-1} e^{n\rho - \frac{s_n}{\rho}}
\end{aligned}
\tag{A.167}
$$

From (1.2), CLT and continuity of the standard normal cdf, we have

$$
\begin{aligned}
&P(X_n \le s_n - 1) = P\left(\frac{X_n - n\rho}{\sqrt{n\rho}} \le \frac{s_n - 1 - n\rho}{\sqrt{n\rho}} \right) = P\left(\frac{X_n - n\rho}{\sqrt{n\rho}} \le \frac{s_n - n\rho}{\sqrt{n\rho}} - \frac{1}{\sqrt{n\rho}} \right) \\
&= P\left(\frac{X_n - n\rho}{\sqrt{n\rho}} \le \frac{\left(\frac{s_n}{n} - \rho\right) n}{\sqrt{n}\sqrt{\rho}} - \frac{1}{\sqrt{n\rho}} \right) = P\left(\frac{X_n - n\rho}{\sqrt{n\rho}} \le \frac{1}{\sqrt{\rho}} \left(\frac{s_n}{n} - \rho\right) \sqrt{n} - \frac{1}{\sqrt{\rho}} \frac{1}{\sqrt{n}} \right) \\
&\to \Phi\left(\frac{\hat{s}}{\sqrt{\rho}} \right)
\end{aligned}
\tag{A.168}
$$

$$P(Y_n \le n-1) = P\left(\frac{Y_n - \frac{s_n}{\rho}}{\sqrt{\frac{s_n}{\rho}}} \le \frac{n-1-\frac{s_n}{\rho}}{\sqrt{\frac{s_n}{\rho}}}\right) = P\left(\frac{Y_n - \frac{s_n}{\rho}}{\sqrt{\frac{s_n}{\rho}}} \le \frac{n\rho - s_n}{\rho\sqrt{\frac{s_n}{\rho}}} - \frac{1}{\sqrt{\frac{s_n}{\rho}}}\right)$$

$$= P\left(\frac{Y_n - \frac{s_n}{\rho}}{\sqrt{\frac{s_n}{\rho}}} \le \frac{\left(\rho - \frac{s_n}{n}\right)n}{\sqrt{\rho}\sqrt{s_n}} - \frac{1}{\sqrt{\frac{s_n}{\rho}}}\right) = P\left(\frac{Y_n - \frac{s_n}{\rho}}{\sqrt{\frac{s_n}{\rho}}} \le \frac{\left(\rho - \frac{s_n}{n}\right)\sqrt{n}}{\sqrt{\rho}\sqrt{\frac{s_n}{n}}} - \frac{1}{\sqrt{\frac{s_n}{\rho}}}\right) \to \Phi\left(\frac{-\hat{s}}{\rho}\right). \tag{A.169}$$

To explore the limits of the remaining terms, we will use Stirling's approximation, $n! \cong \sqrt{2\pi n}\, n^n e^{-n}$. Note that by $f(n) \cong g(n)$ we mean $f(n) = g(n)(1 + o(1))$. If $f(n) \cong g(n)$ and $\lim_{n\to\infty} f(n)$ exists, then $g(n)$ has the same limit. Moreover, from (1.2) and Theorem 12, we have $s_n = n\rho + \hat{s}_n\sqrt{n}$ and $m_n = n\rho + \hat{m}_n\sqrt{n}$. Hence, we have

$$\begin{aligned}
&\left(\frac{s_n}{\rho}\right)^{n+m_n-1} \frac{1}{n^{m_n-1}\, n!\, e^{\frac{s_n}{\rho}}} \left(\sum_{k=0}^{m_n-1}\left(\frac{n\rho}{s_n}\right)^k - \sum_{k=0}^{s_n-1}\left(\frac{n\rho}{s_n}\right)^k\right) \cong \\
&\left(\frac{s_n}{\rho}\right)^{n+m_n} \left(\frac{n\rho}{s_n}\right) \frac{1}{n^{m_n}\sqrt{2\pi n}\, n^n\, e^{-n}\, e^{\frac{s_n}{\rho}}} \left(\frac{\left(\frac{n\rho}{s_n}\right)^{m_n} - \left(\frac{n\rho}{s_n}\right)^{s_n}}{\left(\frac{n\rho}{s_n} - 1\right)}\right) \\
&= \left(\frac{s_n}{n\rho}\right)^{n+m_n} \left(\frac{n\rho}{s_n}\right) \frac{1}{\sqrt{2\pi}\sqrt{n}}\, e^{n-\frac{s_n}{\rho}} \left(\frac{\left(\frac{n\rho}{s_n}\right)^{m_n} - \left(\frac{n\rho}{s_n}\right)^{s_n}}{\left(\frac{n\rho}{s_n} - 1\right)}\right) \\
&= \left(\frac{s_n}{n\rho}\right)^{n+m_n} \left(\frac{n\rho}{s_n}\right) \frac{n\rho + \hat{s}_n\sqrt{n}}{\sqrt{2\pi}\,\hat{s}_n\, n}\, e^{-\hat{s}_n\frac{\sqrt{n}}{\rho}} \left(\left(\frac{n\rho}{s_n}\right)^{s_n} - \left(\frac{n\rho}{s_n}\right)^{m_n}\right)
\end{aligned} \tag{A.170}$$

Using Lemma 2,

$$
\begin{aligned}
& e^{-\frac{\hat{s}_n}{\rho}(1+\rho)\sqrt{n}} \left(\frac{s_n}{n\rho}\right)^{n+m_n} = e^{-\frac{\hat{s}_n}{\rho}(1+\rho)\sqrt{n}} \left(1+\frac{\hat{s}_n}{\rho}\frac{1}{\sqrt{n}}\right)^{(1+\rho)n+\hat{m}_n\sqrt{n}} \to e^{-\frac{\hat{s}^2}{\rho^2}\frac{(1+\rho)}{2}+\frac{\hat{s}\hat{m}}{\rho}} \\
& \left(\frac{n\rho}{s_n}\right) = \left(\frac{s_n}{n\rho}\right)^{-1} = \left(1+\frac{\hat{s}_n}{\rho}\frac{1}{\sqrt{n}}\right)^{-1} \to 1^{-1} = 1 \\
& e^{\hat{s}_n\sqrt{n}} \left(\frac{n\rho}{s_n}\right)^{s_n} = e^{\hat{s}_n\sqrt{n}} \left(\frac{s_n}{n\rho}\right)^{-\rho n-\hat{s}_n\sqrt{n}} \to e^{\frac{\hat{s}^2}{\rho^2}\frac{\rho}{2}+\frac{\hat{s}}{\rho}(-\hat{s})} = e^{-\frac{\hat{s}^2}{2\rho}} \\
& e^{\hat{s}_n\sqrt{n}} \left(\frac{n\rho}{s_n}\right)^{m_n} = e^{\hat{s}_n\sqrt{n}} \left(\frac{s_n}{n\rho}\right)^{-\rho n-\hat{m}_n\sqrt{n}} \to e^{\frac{\hat{s}^2}{\rho^2}\frac{\rho}{2}+\frac{\hat{s}}{\rho}(-\hat{m})}
\end{aligned}
\tag{A.171}
$$

and therefore

$$
\begin{aligned}
& \left(\frac{s_n}{n\rho}\right)^{n+m_n} \left(\frac{n\rho}{s_n}\right) \frac{n\rho+\hat{s}_n\sqrt{n}}{\sqrt{2\pi}\,\hat{s}_n\,n} e^{-\hat{s}_n\frac{\sqrt{n}}{\rho}} \left(\left(\frac{n\rho}{s_n}\right)^{s_n} - \left(\frac{n\rho}{s_n}\right)^{m_n}\right) \\
& = e^{-\hat{s}_n\frac{\sqrt{n}}{\rho}} e^{-\hat{s}_n\sqrt{n}} \left(\frac{s_n}{n\rho}\right)^{n+m_n} \left(\frac{n\rho}{s_n}\right) \frac{n\rho+\hat{s}_n\sqrt{n}}{\sqrt{2\pi}\,\hat{s}_n\,n} e^{\hat{s}_n\sqrt{n}} \left(\left(\frac{n\rho}{s_n}\right)^{s_n} - \left(\frac{n\rho}{s_n}\right)^{m_n}\right) \\
& \to e^{-\frac{\hat{s}^2}{\rho^2}\frac{(1+\rho)}{2}+\frac{\hat{s}\hat{m}}{\rho}} \frac{\rho}{\sqrt{2\pi}\hat{s}} \left(e^{-\frac{\hat{s}^2}{2\rho}} - e^{\frac{\hat{s}^2}{2\rho}-\frac{\hat{s}\hat{m}}{\rho}}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\hat{s}^2}{2\rho^2}} \frac{\rho}{\hat{s}} \left(e^{\frac{\hat{s}}{\rho}(\hat{m}-\hat{s})} - 1\right)
\end{aligned}
\tag{A.172}
$$

Hence, we have shown that $\left(\frac{s_n}{\rho}\right)^{n+m_n-1} \frac{1}{n^{m_n-1}\,n!\,e^{\frac{s_n}{\rho}}} \left(\sum_{k=0}^{m_n-1}\left(\frac{n\rho}{s_n}\right)^k - \sum_{k=0}^{s_n-1}\left(\frac{n\rho}{s_n}\right)^k\right)$ in (A.167) converges to $\left[\phi\left(\frac{\hat{s}}{\rho}\right)\rho\left(1-e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\right)\right] / \left[\hat{s}e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\right]$. Finally, for the fourth part of (A.167), we apply Stirling's approximation twice, to yield

$$
\frac{s_n!}{n^{m_n-1}} \frac{1}{n!} \frac{1}{{s_n}^{s_n}} \left(\frac{s_n}{\rho}\right)^{n+m_n-1} e^{n\rho-\frac{s_n}{\rho}} \cong \frac{\sqrt{2\pi s_n}\,{s_n}^{s_n}\,e^{-s_n}}{n^{m_n-1}\sqrt{2\pi n}\,n^n\,e^{-n}} \frac{1}{{s_n}^{s_n}} \left(\frac{s_n}{\rho}\right)^{n+m_n-1} e^{n\rho-\frac{s_n}{\rho}}
\tag{A.173}
$$

Now, from (A.171)

$$\frac{\sqrt{2\pi s_n}\, {s_n}^{s_n}\, e^{-s_n}}{n^{m_n-1}\sqrt{2\pi n}\, n^n\, e^{-n}}\, \frac{1}{{s_n}^{s_n}}\, \left(\frac{s_n}{\rho}\right)^{n+m_n-1}\, e^{n\rho-\frac{s_n}{\rho}} = \sqrt{\rho+\frac{\hat{s}_n}{\sqrt{n}}}\, \left(\frac{s_n}{n\rho}\right)^{n+m_n}\left(\frac{n\rho}{s_n}\right) e^{-\frac{\hat{s}_n\sqrt{n}}{\rho}-\hat{s}_n\sqrt{n}}$$

$$\to \sqrt{\rho}\, e^{-\frac{\hat{s}^2}{2\rho^2}}\, e^{-\frac{\hat{s}^2}{2\rho}}\, e^{\frac{\hat{m}\hat{s}}{\rho}} \tag{A.174}$$

Hence, we have shown that $\frac{{s_n}!}{n^{m_n-1}}\frac{1}{n!}\frac{1}{{s_n}^{s_n}}\left(\frac{s_n}{\rho}\right)^{n+m_n-1} e^{n\rho-\frac{s_n}{\rho}}$ in (A.167) converges to $\left[\sqrt{\rho}\,\phi\left(\frac{\hat{s}}{\rho}\right)\, e^{\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\right] / \left[\phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\right]$.

Therefore, we finally showed that the probability in (A.166) converges to

$$\left(1+\frac{\rho\left(1-\exp\left\{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}\right\}\right)\phi\left(\frac{\hat{s}}{\rho}\right)}{\hat{s}\,\exp\left\{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}\right\}\Phi\left(-\frac{\hat{s}}{\rho}\right)}+\frac{\sqrt{\rho}\,\phi\left(\frac{\hat{s}}{\rho}\right)\,\exp\left\{\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}\right\}\,\Phi\left(\frac{\hat{s}}{\sqrt{\rho}}\right)}{\phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\,\Phi\left(-\frac{\hat{s}}{\rho}\right)}\right)^{-1}$$

as stated in Theorem 3 for the case QED|QED, $\hat{m} > \hat{s}$, and $\hat{s} \neq 0$. Now we proceed to the second case.

Case 2) $\hat{m} > \hat{s}$ and $\hat{s} = 0$: Since $\bar{m} > \bar{s}$ for sufficiently large n, we have the same formula for $P(D_n^{rpl}(s_n, m_n) > 0)$ as Case 1. Hence,

$$\begin{aligned}
A_n &= ne^{n\rho}P(X_n \leq s_n - 1) \\
B_n &= n\frac{{s_n}^{s_n}}{{s_n}!}\left(\sum_{k=0}^{m_n-1}\left(\frac{n\rho}{s_n}\right)^k - \sum_{k=0}^{s_n-1}\left(\frac{n\rho}{s_n}\right)^k\right) = n\frac{{s_n}^{s_n}}{{s_n}!}\left(\frac{n\rho}{s_n}\right)^{s_n}\sum_{k=0}^{m_n-s_n-1}\left(\frac{n\rho}{s_n}\right)^k \\
C_n &= \frac{n^{m_n}}{{s_n}!}\, n!\, {s_n}^{s_n}\, \left(\frac{\rho}{s_n}\right)^{n+m_n-1}\, e^{\frac{s_n}{\rho}}\, P(Y_n \leq n-1),
\end{aligned} \tag{A.175}$$

and $P(D_n^{rpl}(s_n, m_n) > 0) = \left(1+\frac{B_n}{C_n}+\frac{A_n}{C_n}\right)^{-1}$. To find the limit of $P(D_n^{rpl}(s_n, m_n) > 0)$ as $n \to \infty$, we start by taking the limit of B_n/C_n as $n \to \infty$. Using Stirling's approximation with (1.2), we have

$$\frac{B_n}{C_n} = \left(\frac{{s_n}^{s_n}}{s_n!} \left(\frac{n\rho}{s_n}\right)^{s_n} \sum_{k=0}^{{m_n}^{-s_n-1}} \left(\frac{n\rho}{s_n}\right)^k \right)$$

$$/ \left(\frac{{n^{m_n-1}}}{s_n!}\, n!\, {s_n}^{s_n} \left(\frac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho}}\, P(Y_n \le n-1) \right)$$

$$\cong \left(\left(\frac{n\rho}{s_n}\right)^{s_n} \sum_{k=0}^{{m_n}^{-s_n-1}} \left(\frac{n\rho}{s_n}\right)^k \right)$$

$$/ \left(n^{m_n-1}\, \sqrt{2\pi}\, \sqrt{n}\, n^n\, e^{-n} \left(\frac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho}}\, P(Y_n \le n-1) \right)$$

$$= \frac{1}{P(Y_n \le n-1)}\, e^{n-\frac{s_n}{\rho}} \left(\frac{n\rho}{s_n}\right)^{s_n-n-m_n+1} \frac{1}{\sqrt{2\pi}}\, \frac{1}{\sqrt{n}} \sum_{k=0}^{m_n-s_n-1} \left(\frac{n\rho}{s_n}\right)^k$$

$$= \frac{1}{P(Y_n \le n-1)}\; e^{-\frac{\hat{s}_n\sqrt{n}}{\rho}} e^{-\hat{s}\sqrt{n}} e^{\hat{s}\sqrt{n}} \left(\frac{n\rho}{s_n}\right)^{s_n} \left(\frac{s_n}{n\rho}\right)^{n+m_n} e^{-\frac{\hat{s}_n}{\rho}(1+\rho)\sqrt{n}} e^{\frac{\hat{s}_n}{\rho}(1+\rho)\sqrt{n}} \left(\frac{n\rho}{s_n}\right)$$

$$\cdot\, \frac{1}{\sqrt{2\pi}}\, \frac{1}{\sqrt{n}} \sum_{k=0}^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1} \left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^k$$

$$= \frac{1}{P(Y_n \le n-1)}\; e^{\hat{s}_n\sqrt{n}} \left(\frac{n\rho}{s_n}\right)^{s_n} \left(\frac{s_n}{n\rho}\right)^{n+m_n} e^{-\frac{\hat{s}_n}{\rho}(1+\rho)\sqrt{n}} \left(\frac{n\rho}{s_n}\right) \frac{1}{\sqrt{2\pi}}\, \frac{1}{\sqrt{n}}$$

$$\sum_{k=0}^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1} \left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^k \tag{A.176}$$

Since $\hat{s} = 0$ and by (1.2), the Central Limit Theorem (CLT) and continuity of the standard normal cdf, we have

$$[P(Y_n \le n-1)]^{-1} = \left[P\left(\left(Y_n - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}} \le \left(n - \tfrac{s_n}{\rho} - 1\right) / \sqrt{\tfrac{s_n}{\rho}} \right) \right]^{-1}$$

$$= \left[P\left(\left(Y_n - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}} \le \left(-\tfrac{\hat{s}_n\sqrt{n}}{\rho}\right) / \sqrt{n + \tfrac{\hat{s}_n\sqrt{n}}{\rho}} - 1 / \sqrt{n + \tfrac{\hat{s}_n\sqrt{n}}{\rho}} \right) \right]^{-1} \tag{A.177}$$

$$\to [\Phi(0)]^{-1} = 2$$

From $\hat{s}=0$, (A.171) and (A.177)

$$\frac{1}{P(Y_n\leq n-1)}\ e^{\hat{s}_n\sqrt{n}}\left(\frac{n\rho}{s_n}\right)^{s_n}\ \left(\frac{s_n}{n\rho}\right)^{n+m_n}\ e^{-\frac{\hat{s}_n}{\rho}(1+\rho)\sqrt{n}}\ \left(\frac{n\rho}{s_n}\right)\to 2 \tag{A.178}$$

Now we concentrate on $\frac{1}{\sqrt{n}}\sum_{k=0}^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^k$.

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{k=0}^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^k\\&=\frac{1}{\sqrt{n}}\left[1+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^2+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^3+\ldots+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\right]\end{aligned} \tag{A.179}$$

For each value of n, two cases are possible: $\hat{s}_n<0$ and $\hat{s}_n\geq 0$. Hence, for any n, either

$$\begin{aligned}&(\hat{m}_n-\hat{s}_n)\leq\frac{1}{\sqrt{n}}\left[1+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)+\ldots+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\right]\\&\leq(\hat{m}_n-\hat{s}_n)\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\end{aligned} \tag{A.180}$$

or

$$\begin{aligned}&(\hat{m}_n-\hat{s}_n)\geq\frac{1}{\sqrt{n}}\left[1+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)+\ldots+\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\right]\\&\geq(\hat{m}_n-\hat{s}_n)\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\end{aligned} \tag{A.181}$$

In both cases, $\frac{1}{\sqrt{n}}\sum_{k=0}^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^k$ is bounded by $(\hat{m}_n-\hat{s}_n)$ and $(\hat{m}_n-\hat{s}_n)\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}$. Now we will study the limits of both bounds. From $\hat{s}=0$, $(\hat{m}_n-\hat{s}_n)\to\hat{m}$. In addition, from Lemma 2 $\left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1}=$

$\left(1+\frac{\hat{s}_n}{\rho\sqrt{n}}\right)^{-(\hat{m}_n-\hat{s}_n)\sqrt{n}+1} \to 1$. Thus, both bounds approach $\hat{m}$, and we have $\frac{1}{\sqrt{n}} \sum_{k=0}^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1} \left(\frac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^k \to \hat{m}$. The limit of the whole expression is (continuing from (A.176)):

$$\begin{aligned} &\tfrac{1}{P(Y_n\le n-1)}\ e^{\hat{s}_n\sqrt{n}} \left(\tfrac{n\rho}{s_n}\right)^{s_n} \left(\tfrac{s_n}{n\rho}\right)^{n+m_n} e^{-\frac{\hat{s}_n}{\rho}(1+\rho)\sqrt{n}} \left(\tfrac{n\rho}{s_n}\right) \tfrac{1}{\sqrt{2\pi}}\ \tfrac{1}{\sqrt{n}} \\ &\cdot \sum_{k=0}^{(\hat{m}_n-\hat{s}_n)\sqrt{n}-1} \left(\tfrac{n\rho}{n\rho+\hat{s}_n\sqrt{n}}\right)^k \to 2\tfrac{1}{\sqrt{2\pi}}\hat{m} = \sqrt{\tfrac{2}{\pi}}\hat{m} \end{aligned} \tag{A.182}$$

Now we study the limit of $\frac{A_n}{C_n}$. From (A.168), (A.177), $\hat{s}=0$, (A.173), and (A.174), we have

$$\begin{aligned} &\tfrac{A_n}{C_n} = \left(e^{n\rho}P(X_n\le s_n-1)\right) / \left(\tfrac{n^{m_n-1}}{s_n!}\ n!\ {s_n}^{s_n}\ \left(\tfrac{\rho}{s_n}\right)^{n+m_n-1}\ e^{\frac{s_n}{\rho}}\ P(Y_n\le n-1)\right) \\ &= \tfrac{s_n!}{n^{m_n-1}}\ \tfrac{1}{n!}\ \tfrac{1}{{s_n}^{s_n}}\ \left(\tfrac{s_n}{\rho}\right)^{n+m_n-1}\ e^{n\rho-\frac{s_n}{\rho}}\ \tfrac{P(X_n\le s_n-1)}{P(Y_n\le n-1)} \\ &\to \sqrt{\rho}\, e^{-\frac{\hat{s}^2}{2\rho^2}}\ e^{-\frac{\hat{s}^2}{2\rho}}\ e^{\frac{\hat{m}\hat{s}}{\rho}}\ \tfrac{1/2}{1/2} = \sqrt{\rho} \end{aligned} \tag{A.183}$$

Hence, we finally showed that the probability in (A.166) converges to

$$\left(1+\sqrt{\tfrac{2}{\pi}}\hat{m}+\sqrt{\rho}\right)^{-1} \tag{A.184}$$

as stated in Theorem 3 for the case $\hat{m} > \hat{s}$ and $\hat{s}=0$. Now we proceed to the third case.

Case 3) $\hat{m} < \hat{s}$: For sufficiently large n, this leads to $m_n < s_n$.

Define A_n, B_n, and C_n (using (1.1)) as

$$A_n \equiv \sum_{k=0}^{m_n-1} np_{n,k}^{st} = \sum_{k=0}^{m_n-1} n\frac{n^k}{k!}\,\rho^k,$$

$$B_n \equiv \sum_{k=m_n}^{s_n-1} (n+m_n-k)p_{n,k}^{st} = \sum_{k=m_n}^{s_n-1} (n+m_n-k)\frac{n^{m_n}}{k!}\,\frac{n!}{(n+m_n-k))!}\,\rho^k,$$

$$C_n \equiv \sum_{k=s_n}^{n+m_n} (n+m_n-k)p_{n,k}^{st} = \sum_{k=s_n}^{n+m_n} (n+m_n-k)\frac{n^{m_n}}{s_n!}\,\frac{n!}{(n+m_n-k))!}\,{s_n}^{s_n-k}\,\rho^k \tag{A.185}$$

We have

$$P(D_n^{rpl}(s_n,m_n)>0) = \frac{B_n+C_n}{A_n+B_n+C_n} = \left(1+\frac{A_n}{B_n+C_n}\right)^{-1} \tag{A.186}$$

Now we look at the values of A_n, B_n, and C_n. Notice that

$$A_n = ne^{n\rho}e^{-n\rho}\sum_{k=0}^{m_n-1}\frac{(n\rho)^k}{k!} = ne^{n\rho}\sum_{k=0}^{m_n-1}e^{-n\rho}\frac{(n\rho)^k}{k!} = ne^{n\rho}P(X_n\le m_n-1) \tag{A.187}$$

where $X_n \sim Poisson(n\rho)$. Moreover, since $\rho = \frac{\lambda}{\mu}$ we have

$$\begin{aligned}
B_n &= n^{m_n}\,n!\sum_{k=m_n}^{s_n-1}\frac{(n+m_n-k)}{k!\,(n+m_n-k)!}\rho^k = \frac{n^{m_n}\,n!}{(n+m_n-1)!}\sum_{k=m_n}^{s_n-1}\frac{(n+m_n-1)!}{k!\,(n+m_n-k-1)!}\rho^k\,\frac{(1+\rho)^{n+m_n-1}}{(1+\rho)^{n+m_n-1}}\\
&= \frac{n^{m_n}\,n!}{(n+m_n-1)!}\,(1+\rho)^{n+m_n-1}\sum_{k=m_n}^{s_n-1}\binom{n+m_n-1}{k}\rho^k\,\frac{1}{\left(\frac{\lambda+\mu}{\mu}\right)^{n+m_n-1}}\\
&= \frac{n^{m_n}\,n!}{(n+m_n-1)!}\,(1+\rho)^{n+m_n-1}\sum_{k=m_n}^{s_n-1}\binom{n+m_n-1}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n+m_n-1-k}\\
&= \frac{n^{m_n}\,n!}{(n+m_n-1)!}\,(1+\rho)^{n+m_n-1}\,P(m_n\le Y_n\le s_n-1)
\end{aligned} \tag{A.188}$$

where $Y_n \sim Binomial(n+m_n-1, \frac{\lambda}{\lambda+\mu})$.

Multiplying and dividing C_n by ${s_n}^{n+m_n-1}$, ρ^{n+m_n-1}, and $e^{-\frac{s_n}{\rho}}$ and reindexing our summation, we have

$$\begin{aligned}
C_n &= \tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n} \sum_{k=s_n}^{n+m_n} \tfrac{(n+m_n-k)}{(n+m_n-k)!} \left(\tfrac{s_n}{\rho}\right)^{-k} \\
&= \tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n}\, \tfrac{\rho^{n+m_n-1}}{{s_n}^{n+m_n-1}}\, e^{\frac{s_n}{\rho}} \sum_{k=s_n}^{n+m_n-1} e^{-\frac{s_n}{\rho}} \left(\tfrac{s_n}{\rho}\right)^{n+m_n-1-k} \tfrac{1}{(n+m_n-1-k)!} \\
&= \tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n} \left(\tfrac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho}} \sum_{j=0}^{n+m_n-s_n-1} e^{-\frac{s_n}{\rho}} \left(\tfrac{s_n}{\rho}\right)^{j} \tfrac{1}{j!} \\
&= \tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n} \left(\tfrac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho}}\, P(Z_n \le n+m_n-s_n-1),
\end{aligned} \tag{A.189}$$

where $Z_n \sim Poisson\left(\frac{s_n}{\rho}\right)$. By (A.186)–(A.189), we have

$$\begin{aligned}
&P(D_n^{rpl}(s_n, m_n) > 0) = \\
&(1 + ne^{n\rho} P(X_n \le m_n - 1) \\
&/ \Big[n^{m_n}\, n!\, \tfrac{(1+\rho)^{n+m_n-1}}{(n+m_n-1)!}\, P(m_n \le Y_n \le s_n - 1) \\
&+ \tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n} \left(\tfrac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho}}\, P(Z_n \le n+m_n-s_n-1) \Big] \Big)^{-1} \\
&= (1 + P(X_n \le m_n - 1) \\
&\Big\{ n^{m_n-1}\, n!\, \tfrac{(1+\rho)^{n+m_n-1}}{(n+m_n-1)!}\, e^{-n\rho}\, P(m_n \le Y_n \le s_n - 1) \\
&+ \tfrac{n^{m_n-1}}{s_n!}\, n!\, {s_n}^{s_n} \left(\tfrac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho} - n\rho}\, P(Z_n \le n+m_n-s_n-1) \Big\} \Big)^{-1}
\end{aligned} \tag{A.190}$$

To obtain the limit as $n \to \infty$ of the quantity above, we study the limits of the parts:

$$P(X_n \leq m_n - 1), \quad P(m_n \leq Y_n \leq s_n - 1), \quad P(Z_n \leq n + m_n - s_n - 1)$$
$$n^{m_n - 1}\, n!\, \tfrac{(1+\rho)^{n+m_n-1}}{(n+m_n-1)!}\, e^{-n\rho}, \quad \tfrac{n^{m_n-1}}{s_n!}\, n!\, {s_n}^{s_n} \left(\tfrac{\rho}{s_n}\right)^{n+m_n-1} e^{\frac{s_n}{\rho} - n\rho} \tag{A.191}$$

Similar to the previous cases, from (A.4) for QED|QED, (1.2), the CLT and continuity of the standard normal cdf, we have

$$P(X_n \leq m_n - 1) = P\left(\tfrac{X - n\rho}{\sqrt{n\rho}} \leq \tfrac{1}{\sqrt{\rho}} \left(\tfrac{m_n}{n} - \rho\right) \sqrt{n} - \tfrac{1}{\sqrt{n\rho}}\right) \to \Phi\left(\tfrac{\hat{m}}{\sqrt{\rho}}\right) \tag{A.192}$$

$$
\begin{aligned}
&P(m_n \le Y_n \le s_n - 1) \\
&= P\left(\frac{m_n - (n+m_n-1)\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \le \frac{Y_n - (n+m_n-1)\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \le \frac{s_n - 1 - (n+m_n-1)\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \right) \\
&= P\left(\frac{\frac{\mu}{\lambda+\mu}m_n - n\frac{\lambda}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \right. \\
&\left. \le \frac{Y_n - (n+m_n-1)\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \le \frac{s_n - n\rho}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{m_n\frac{\lambda}{\lambda+\mu} - n\rho + n\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{1-\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \right) \\
&= P\left(\frac{\left(\frac{m_n}{n} - \rho\right)\sqrt{n} + \frac{\rho}{\sqrt{n}}}{\sqrt{\left(\frac{n+m_n-1}{n}\right)\rho}} \right. \\
&\left. \le \frac{Y_n - (n+m_n-1)\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \le \frac{\left(\frac{s_n}{n} - \rho\right)\sqrt{n}}{\sqrt{\left(\frac{n+m_n-1}{n}\right)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{m_n\frac{\lambda}{\lambda+\mu} - n\lambda\left(\frac{1}{\mu} - \frac{1}{\lambda+\mu}\right)}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{1-\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \right) \\
&= P\left(\frac{\left(\frac{m_n}{n} - \rho\right)\sqrt{n} + \frac{\rho}{\sqrt{n}}}{\sqrt{\left(1+\frac{m_n-1}{n}\right)\rho}} \right. \\
&\left. \le \frac{Y_n - (n+m_n-1)\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \le \frac{\left(\frac{s_n}{n} - \rho\right)\sqrt{n}}{\sqrt{\left(1+\frac{m_n-1}{n}\right)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{\frac{\lambda}{\lambda+\mu}\left(m_n - n\frac{\lambda}{\mu}\right)}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{1-\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \right) \\
&= P\left(\frac{\left(\frac{m_n}{n} - \rho\right)\sqrt{n} + \frac{\rho}{\sqrt{n}}}{\sqrt{\left(1+\frac{m_n-1}{n}\right)\rho}} \right. \\
&\left. \le \frac{Y_n - (n+m_n-1)\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \le \frac{\left(\frac{s_n}{n} - \rho\right)\sqrt{n}}{\sqrt{\left(1+\frac{m_n-1}{n}\right)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{\sqrt{\frac{\lambda}{\mu}}\left(\frac{m_n}{n} - \rho\right)\sqrt{n}}{\sqrt{\left(1+\frac{m_n-1}{n}\right)}} - \frac{1-\frac{\lambda}{\lambda+\mu}}{\sqrt{(n+m_n-1)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} \right) \\
&\to \Phi\left(\frac{\hat{s}}{\sqrt{(1+\rho)\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}}} - \frac{\sqrt{\frac{\lambda}{\mu}}\hat{m}}{\sqrt{(1+\rho)}} \right) - \Phi\left(\frac{\hat{m}}{\sqrt{(1+\rho)\rho}} \right) \\
&= \Phi\left(\hat{s}\sqrt{1+\frac{1}{\rho}} - \frac{\hat{m}}{\sqrt{1+\frac{1}{\rho}}} \right) - \Phi\left(\frac{\hat{m}}{\sqrt{(1+\rho)\rho}} \right)
\end{aligned}
\tag{A.193}
$$

$$
\begin{aligned}
&P(Z_n \leq n + m_n - s_n - 1) \\
&= P\left(\left(Z_n - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}} \leq \left(n + m_n - s_n - 1 - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}}\right) \\
&= P\left(\left(Z_n - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}} \leq \left(n + m_n - s_n\left(1 + \tfrac{1}{\rho}\right) - 1 - n\rho + n\rho\right) / \sqrt{\tfrac{s_n}{\rho}}\right) \\
&= P\left(\left(Z_n - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}} \leq \left[\left(\tfrac{m_n}{n} - \rho\right) n\right] / \sqrt{\tfrac{s_n}{\rho}} + \left(n + n\rho - 1 - s_n\left(1 + \tfrac{1}{\rho}\right)\right) / \sqrt{\tfrac{s_n}{\rho}}\right) \\
&= P\left(\left(Z_n - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}}\right. \\
&\leq \left[\left(\tfrac{m_n}{n} - \rho\right) \sqrt{n}\right] / \sqrt{\tfrac{s_n}{n\rho}} - \left(1 - n\rho\left(1 + \tfrac{1}{\rho}\right) + s_n\left(1 + \tfrac{1}{\rho}\right)\right) / \sqrt{\tfrac{s_n}{\rho}}\Bigg) \\
&= P\left(\left(Z_n - \tfrac{s_n}{\rho}\right) / \sqrt{\tfrac{s_n}{\rho}}\right. \\
&\leq \left[\left(\tfrac{m_n}{n} - \rho\right) \sqrt{n}\right] / \sqrt{\tfrac{s_n}{n\rho}} - \left(1 + \left(\tfrac{s_n}{n} - \rho\right) \sqrt{n}\left(1 + \tfrac{1}{\rho}\right)\right) / \sqrt{\tfrac{s_n}{n\rho}}\Bigg) \\
&\to \Phi\left(\hat{m} - \hat{s}\left(1 + \tfrac{1}{\rho}\right)\right)
\end{aligned}
\tag{A.194}
$$

As in the previous cases, we use Stirling's approximation, and Lemma 2 to get

$$n^{m_n-1}\, n!\, \tfrac{(1+\rho)^{n+m_n-1}}{(n+m_n-1)!}\, e^{-n\rho} \cong n^{m_n-1} \tfrac{\sqrt{2\pi n}\, n^n\, e^{-n}(n+m_n)}{\sqrt{2\pi(n+m_n)}\,(n+m_n)^{(n+m_n)}e^{-(n+m_n)}}(1+\rho)^{n+m_n-1}\, e^{-n\rho}$$

$$= \left(\sqrt{\tfrac{n+m_n}{n}}\right)^{-1} \left(\tfrac{n+m_n}{n(1+\rho)}\right)^{-(n+m_n)} \left(\tfrac{n+m_n}{n(1+\rho)}\right) e^{m_n-n\rho}$$

$$\to \left(\sqrt{(1+\rho)}\right)^{-1} \exp\left\{-\tfrac{\hat{m}^2}{2(1+\rho)}\right\} = \left(\sqrt{(1+\rho)}\right)^{-1} \exp\left\{-\tfrac{\hat{m}^2}{2\rho} + \tfrac{\hat{m}^2}{2\rho(1+\rho)}\right\}$$

$$= \left(\sqrt{(1+\rho)}\right)^{-1} \phi\left(-\tfrac{\hat{m}}{\sqrt{\rho}}\right) / \phi\left(\tfrac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right) \tag{A.195}$$

Note that the last part of (A.205) is the reciprocal of the last part of (A.167). Hence, from (A.174) we have

$$\tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n}\, \tfrac{\rho}{s_n}^{n+m_n}\, e^{\frac{s_n}{\rho}-n\rho} \to \tfrac{1}{\sqrt{\rho}} \exp\left\{\tfrac{\hat{s}^2(1+\rho)}{2\rho^2} - \tfrac{\hat{s}\hat{m}}{\rho}\right\}$$

$$= \tfrac{1}{\sqrt{\rho}} \left[\phi\left(-\tfrac{\hat{m}}{\sqrt{\rho}}\right) \phi\left(\hat{s}\sqrt{1+\tfrac{1}{\rho}} - \tfrac{\hat{m}}{\sqrt{1+\frac{1}{\rho}}}\right)\right] / \left[\phi\left(\left(1+\tfrac{1}{\rho}\right)\hat{s} - \hat{m}\right) \phi\left(\tfrac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\right] \tag{A.196}$$

Hence, from (A.192)-(A.196) we have

$$
\begin{aligned}
&P(D_n^{rpl}(s_n, m_n) > 0) \\
&= (1 + P(X_n \leq m_n - 1) \\
&\{n^{m_n-1}\, n!\, (1+\rho)^{n+m_n-1}((n+m_n-1)!)^{-1}\, e^{-n\rho}\, P(m_n \leq Y_n \leq s_n - 1) \\
&+n^{m_n-1}(s_n!)^{-1}\, n!\, {s_n}^{s_n}\, \left(\frac{\rho}{s_n}\right)^{n+m_n-1}\, e^{\frac{s_n}{\rho}-n\rho}\, P(Z \leq n + m_n - s_n - 1)\Big\}\Big)^{-1} \\
&\to \left(1 + \Phi\left(\frac{\hat{m}}{\sqrt{\rho}}\right)\right. \\
&\left\{\left(\sqrt{(1+\rho)}\right)^{-1} \frac{\phi\left(-\hat{m}/\sqrt{\rho}\right)}{\phi\left(\hat{m}/\sqrt{\rho(1+\rho)}\right)} \left[\Phi\left(\hat{s}\sqrt{1+\tfrac{1}{\rho}} - \hat{m}/\sqrt{1+\tfrac{1}{\rho}}\right) - \Phi\left(\frac{\hat{m}}{\sqrt{(1+\rho)\rho}}\right)\right]\right. \\
&\left.\left. +\frac{1}{\sqrt{\rho}} \frac{\phi\left(-\hat{m}/\sqrt{\rho}\right)}{\phi((1+1/\rho)\hat{s}-\hat{m})} \frac{\phi\left(\hat{s}\sqrt{1+1/\rho}-\hat{m}/\sqrt{1+1/\rho}\right)}{\phi\left(\hat{m}/\sqrt{\rho(1+\rho)}\right)} \Phi\left(\hat{m} - \hat{s}\left(1+\tfrac{1}{\rho}\right)\right)\right\}\right)^{-1}
\end{aligned}
\tag{A.197}
$$

Case 4) $\hat{m} = \hat{s}$: For any $\epsilon > 0$ and for sufficiently large n, this leads to $s_n - 2\epsilon\sqrt{n} < m_n < s_n + 2\epsilon\sqrt{n}$. It can be shown that the delay is decreasing in the capacity level used. For sufficiently large n,

$$
P(D_n^{rpl}(s_n, s_n + 2\epsilon\sqrt{n}) > 0) < P(D_n^{rpl}(s_n, m_n) > 0) < P(D_n^{rpl}(s_n, s_n - 2\epsilon\sqrt{n}) > 0),
\tag{A.198}
$$

the limiting probability of delay until replacement for this case will be between the limits of the bounds above; so, it will be between the probability of delay until replacement obtained in the case where $\hat{m} > \hat{s}$ and the case where $\hat{m} < \hat{s}$ (remember that these cases correspond to $m_n > s_n$ and $s_n > m_n$, respectively).

Thus, for $\hat{s} \neq 0$ we have

$$
\begin{aligned}
&\left(1+\rho\left(1-e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\right)\phi\left(\frac{\hat{s}}{\rho}\right)/\left[\hat{s}\,e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\Phi\left(-\frac{\hat{s}}{\rho}\right)\right]\right.\\
&\left.+\sqrt{\rho}\,\phi\left(\frac{\hat{s}}{\rho}\right)\,e^{\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\,\Phi\left(\frac{\hat{s}}{\sqrt{\rho}}\right)/\left[\phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\,\Phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\right]\right)^{-1}\\
&<\lim_{n\to\infty}P(D_n^{rpl}(s_n,m_n)>0)\\
&<\left(1+\Phi\left(\frac{\hat{m}}{\sqrt{\rho}}\right)/\left[\frac{1}{\sqrt{1+\rho}}\,\phi\left(-\frac{\hat{m}}{\sqrt{\rho}}\right)A+\frac{1}{\sqrt{\rho}}\,\phi\left(-\frac{\hat{m}}{\sqrt{\rho}}\right)\,B\right]\right)^{-1}
\end{aligned}
\tag{A.199}
$$

where

$$
\begin{aligned}
A&=\left[\Phi\left(\hat{s}\sqrt{1+\tfrac{1}{\rho}}-\frac{\hat{m}}{\sqrt{1+\frac{1}{\rho}}}\right)-\Phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\right]/\phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\\
B&=\left[\phi\left(\hat{s}\sqrt{1+\tfrac{1}{\rho}}-\frac{\hat{m}}{\sqrt{1+\frac{1}{\rho}}}\right)\,\Phi\left(\hat{m}-\left(1+\tfrac{1}{\rho}\right)\,\hat{s}\right)\right]\\
&/\left[\phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\,\phi\left(\left(1+\tfrac{1}{\rho}\right)\hat{s}-\hat{m}\right)\right],
\end{aligned}
\tag{A.200}
$$

and for $\hat{s}=0$ we have

$$
\begin{aligned}
&\left(1+\sqrt{\tfrac{2}{\pi}}\hat{m}+\sqrt{\rho}\right)^{-1}<\lim_{n\to\infty}P(D_n^{rpl}(s_n,m_n)>0)\\
&<\left(1+\Phi\left(\frac{\hat{m}}{\sqrt{\rho}}\right)/\left[\frac{1}{\sqrt{1+\rho}}\,\phi\left(-\frac{\hat{m}}{\sqrt{\rho}}\right)A+\frac{1}{\sqrt{\rho}}\,\phi\left(-\frac{\hat{m}}{\sqrt{\rho}}\right)\,B\right]\right)^{-1}
\end{aligned}
\tag{A.201}
$$

where

$$
\begin{aligned}
A&=\left[\Phi\left(-\frac{\hat{m}}{\sqrt{1+\frac{1}{\rho}}}\right)-\Phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\right]/\phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\\
B&=\left[\phi\left(-\frac{\hat{m}}{\sqrt{1+\frac{1}{\rho}}}\right)\,\Phi\left(\hat{m}\right)\right]/\left[\phi\left(\frac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\,\phi\left(-\hat{m}\right)\right].
\end{aligned}
\tag{A.202}
$$

Letting $\hat{m}\to\hat{s}$ in the inequalities above, for $\hat{s}\neq 0$ we get

$$\left(1+\rho\left(1-e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\right)\phi\left(\frac{\hat{s}}{\rho}\right)/\left[\hat{s}\,e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\Phi\left(-\frac{\hat{s}}{\rho}\right)\right]\right.$$

$$\left.\sqrt{\rho}\,\phi\left(\frac{\hat{s}}{\rho}\right)\,e^{\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\,\Phi\left(\frac{\hat{s}}{\sqrt{\rho}}\right)/\left[\phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\,\Phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\right]\right)^{-1}$$

$$<\lim_{n\to\infty}P(D_n^{rpl}(s_n,m_n)>0) \tag{A.203}$$

$$<\left(1+\rho\left(1-e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\right)\phi\left(\frac{\hat{s}}{\rho}\right)/\left[\hat{s}\,e^{-\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\Phi\left(-\frac{\hat{s}}{\rho}\right)\right]\right.$$

$$\left.+\sqrt{\rho}\,\phi\left(\frac{\hat{s}}{\rho}\right)\,e^{\frac{\hat{s}(\hat{m}-\hat{s})}{\rho}}\,\Phi\left(\frac{\hat{s}}{\sqrt{\rho}}\right)/\left[\phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\,\Phi\left(-\frac{\hat{s}}{\sqrt{\rho}}\right)\right]\right)^{-1}$$

and for $\hat{s}=0$ we get

$$(1+\sqrt{\rho})^{-1}<\lim_{n\to\infty}P(D_n^{rpl}(s_n,m_n)>0)<(1+\sqrt{\rho})^{-1}\,.\square$$

Proof of Theorem 3 for $\mathbf{T}=\mathbf{0}$*, QD|QED*

Since from Theorem 12 and Table A.4 we know that $\bar{m}<\bar{s}$, for sufficiently large n this implies $m_n<s_n$. Let A_n, B_n, and C_n be as in Case 3 above. Then, (A.185)-(A.188) are still valid. Similar to (A.204) and (A.205),

$$P(D_n^{rpl}(s_n,m_n)>0)$$

$$=(1+P(X_n\le m_n-1)$$

$$\left[(1/ne^{n\rho})\,n^{m_n}\,n!\,\tfrac{(1+\rho)^{n+m_n-1}}{(n+m_n-1)!}\,P(m_n\le Y_n\le s_n-1)\right. \tag{A.204}$$

$$\left.+\,(1/ne^{n\rho})\sum_{k=s_n}^{n+m_n}(n+m_n-k)\tfrac{n^{m_n}}{s_n!}\,\tfrac{n!}{(n+m_n-k))!\,s_n{}^{s_n-k}\,\rho^k}\right]\Big)^{-1}$$

To obtain the limit as $n\to\infty$ of the quantity above, we study the limits of the parts:

$$P(X_n \le m_n - 1), \quad P(m_n \le Y_n \le s_n - 1),$$

$$n^{m_n - 1}\, n!\, \tfrac{(1+\rho)^{n+m_n-1}}{(n+m_n-1)!}\, e^{-n\rho}, \quad (1/ne^{n\rho}) \sum_{k=s_n}^{n+m_n} (n + m_n - k) \tfrac{n^{m_n}}{s_n!}\, \tfrac{n!}{(n+m_n-k))!}\, {s_n}^{s_n-k}\, \rho^k. \tag{A.205}$$

From Theorem 12, (A.192) is still valid; since $\bar{s} + \bar{s}/\rho - 1 > \bar{m} \Rightarrow \bar{s} - (1+\bar{m})\frac{\lambda}{\lambda+\mu} > 0$, from (A.193) $P(m_n \le Y_n \le s_n - 1) \to 1 - \Phi\left(\frac{\hat{m}}{\sqrt{(1+\rho)\rho}}\right)$; (A.195) is still valid, and now we look at the limit of the last part.

First, note that

$$\begin{aligned} &(1/ne^{n\rho}) \sum_{k=s_n}^{n+m_n} (n + m_n - k) \tfrac{n^{m_n}}{s_n!}\, \tfrac{n!}{(n+m_n-k))!}\, {s_n}^{s_n-k}\, \rho^k \\ &= \tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n}\ (1/ne^{n\rho}) \sum_{k=s_n}^{n+m_n-1} \tfrac{n!}{(n+m_n-k-1)!} \left(\tfrac{\rho}{s_n}\right)^k, \end{aligned} \tag{A.206}$$

and for $k = s_n, s_n + 1, \ldots, n + m_n - 1$ (and for sufficiently large n)

$$\begin{aligned} &\left(\tfrac{1}{(n+m_n-k)!} \left(\tfrac{\rho}{s_n}\right)^{k+1}\right) / \left(\tfrac{1}{(n+m_n-k-1)!} \left(\tfrac{\rho}{s_n}\right)^k\right) = \tfrac{\rho}{s_n}(n + m_n - 1 - k) \\ &\le \tfrac{\rho}{s_n}(n + m_n - 1 - s_n) \cong (1 + \bar{m} - \bar{s})/(\bar{s}/\rho) < 1. \end{aligned} \tag{A.207}$$

Thus, $\left(\frac{1}{(n+m_n-k-1)!} \left(\frac{\rho}{s_n}\right)^k\right)$ is decreasing for $k = s_n, s_n + 1, \ldots, n + m_n - 1$, and

$$\begin{aligned} &\tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n}\ (1/ne^{n\rho}) \sum_{k=s_n}^{n+m_n-1} \tfrac{n!}{(n+m_n-k-1)!} \left(\tfrac{\rho}{s_n}\right)^k \\ &\le \tfrac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n}\ (1/ne^{n\rho})\, \tfrac{n!}{(n+m_n-s_n-1)!} \left(\tfrac{\rho}{s_n}\right)^{s_n} (n + m_n - s_n) \end{aligned} \tag{A.208}$$

Applying Stirling's approximation gives

$$
\begin{aligned}
&\frac{n^{m_n}}{s_n!}\, n!\, {s_n}^{s_n}\ (1/ne^{n\rho})\, \frac{n!}{(n+m_n-s_n-1)!} \left(\frac{\rho}{s_n}\right)^{s_n} (n+m_n-s_n) \\
&= \frac{n^{m_n}\sqrt{2\pi}n\, n^n\, e^{-n}}{\sqrt{2\pi}\,\sqrt{s_n}s_n^{s_n}e^{-s_n}} s_n^{s_n}\ \frac{1}{n\, e^{n\rho}}\ \left(\frac{\rho}{s_n}\right)^{s_n}\ \frac{(n+m_n-s_n)^2}{(n+m_n-s_n)!} \\
&= n^{n+m_n}\ \frac{1}{\sqrt{s_n}}\ e^{s_n-n}\ \frac{1}{e^{n\rho}}\ \left(\frac{\rho}{s_n}\right)^{s_n}\ \frac{(n+m_n-s_n)^2}{\sqrt{2\pi}\sqrt{n+m_n-s_n}\,(n+m_n-s_n)^{(n+m_n-s_n)}\, e^{-(n+m_n-s_n)}} \\
&= \frac{(n+m_n-s_n)^2}{\sqrt{s_n}\sqrt{n+m_n-s_n}}\quad \frac{1}{\sqrt{2\pi}}\ e^{s_n-n-n\rho+n+m_n-s_n}\ \left(\frac{n}{n+m_n-s_n}\right)^{n+m_n}\ \left(\frac{(n+m_n-s_n)\rho}{s_n}\right)^{s_n}.
\end{aligned}
\tag{A.209}
$$

Similar to (A.35) and from Lemma 1 we continue ($0 < \epsilon < 1/2$)

$$
\begin{aligned}
&= \frac{(n+m_n-s_n)^2}{\sqrt{s_n}\sqrt{n+m_n-s_n}}\quad \frac{1}{\sqrt{2\pi}} \\
&exp\left\{n\left[-(1+\rho)\quad ln(1+\rho-\bar{s})+\bar{s}\quad ln\left(\frac{(1+\rho-\bar{s})\rho}{\bar{s}}\right)\right]+o(n^{1-\epsilon})\right\}.
\end{aligned}
\tag{A.210}
$$

Using Taylor's approximation similar to (A.36),

$$
\begin{aligned}
&-(1+\rho)\quad ln(1+\rho-\bar{s}) = (1+\rho)\left[(\bar{s}-\rho)+\tfrac{(\bar{s}-\rho)^2}{2}+\ldots\right], \\
&\bar{s}\quad ln\left(\frac{(1+\rho-\bar{s})\rho}{\bar{s}}\right) = -\bar{s}\left[\left[\frac{(\bar{s}-\rho)(1+\rho)}{\bar{s}}\right]+\left[\frac{(\bar{s}-\rho)(1+\rho)}{\bar{s}}\right]^2/2+\ldots\right].
\end{aligned}
\tag{A.211}
$$

Since $\bar{m} = \rho$ and we consider the nontrivial cases $n + m_n \geq s_n \Rightarrow 1 + \rho \geq \bar{s}$, observe that for any $k \geq 2$ and for $1 + \rho > \bar{s}$

$$
\begin{aligned}
&(1+\rho)(\bar{s}-\rho) = \bar{s}\left[\frac{(\bar{s}-\rho)(1+\rho)}{\bar{s}}\right], \\
&(1+\rho)(\bar{s}-\rho)^k < \bar{s}\left[\frac{(\bar{s}-\rho)(1+\rho)}{\bar{s}}\right]^k;
\end{aligned}
\tag{A.212}
$$

and after applying L'Hospital's rule as in (A.37) we conclude from (A.208) that

$$\begin{aligned}
&\tfrac{n^{m_n}}{s_n!}\; n!\; {s_n}^{s_n}\; (1/ne^{n\rho})\; \tfrac{n!}{(n+m_n-s_n-1)!} \left(\tfrac{\rho}{s_n}\right)^{s_n} (n+m_n-s_n) \to 0 \\
&\Rightarrow (1/ne^{n\rho}) \sum_{k=s_n}^{n+m_n} (n+m_n-k) \tfrac{n^{m_n}}{s_n!}\; \tfrac{n!}{(n+m_n-k))!}\; {s_n}^{s_n-k}\; \rho^k \to 0.
\end{aligned} \tag{A.213}$$

For $1+\rho = \bar{s}$ we change the last equality in (A.209) in the following way

$$\begin{aligned}
&\tfrac{(n+m_n-s_n)^2}{\sqrt{s_n}\sqrt{n+m_n-s_n}} \quad \tfrac{1}{\sqrt{2\pi}}\; e^{s_n-n-n\rho+n+m_n-s_n}\; \left(\tfrac{n+m_n-s_n}{n}\right)^{-(n+m_n-s_n)}\; \left(\tfrac{n\rho}{s_n}\right)^{s_n} \\
&= \tfrac{(n+m_n-s_n)^2}{\sqrt{s_n}\sqrt{n+m_n-s_n}} \quad \tfrac{1}{\sqrt{2\pi}} \\
&exp\left\{n\left[-(1+\rho-\bar{s}) \quad ln(1+\rho-\bar{s}) + \bar{s} \quad ln\left(\tfrac{\rho}{\bar{s}}\right)\right] + o(n^{1-\epsilon})\right\}
\end{aligned} \tag{A.214}$$

Note that from L'Hospital's rule, $xln(x)|_{x=0} = 0$, and from Theorem 12, $0 < \frac{\rho}{\bar{s}} < 1$, that (A.213) is still valid.

Hence,

$$\begin{aligned}
&P(D_n^{rpl}(s_n, m_n) > 0) \\
&= (1 + P(X_n \le m_n - 1) \\
&/ \left[(1/ne^{n\rho})\, n^{m_n}\; n!\; \tfrac{(1+\rho)^{n+m_n-1}}{(n+m_n-1)!}\; P(m_n \le Y_n \le s_n - 1)\right. \\
&\left. + (1/ne^{n\rho}) \sum_{k=s_n}^{n+m_n} (n+m_n-k) \tfrac{n^{m_n}}{s_n!}\; \tfrac{n!}{(n+m_n-k))!\; {s_n}^{s_n-k}\; \rho^k}\right]\Bigg)^{-1} \\
&\to \left(1 + \left[\Phi\left(\tfrac{\hat{m}}{\sqrt{\rho}}\right)\; \sqrt{(1+\rho)}\; \phi\left(\tfrac{\hat{m}}{\sqrt{\rho(1+\rho)}}\right)\right] / \left[\phi\left(-\tfrac{\hat{m}}{\sqrt{\rho}}\right) \Phi\left(-\tfrac{\hat{m}}{\sqrt{(1+\rho)\rho}}\right)\right]\right)^{-1}. \square
\end{aligned} \tag{A.215}$$

A.4 Proofs for Theorem 4, 5, and 6

The proof for Theorem 4 is omitted because it is very similar to the proof of Theorem 1, because the only difference between these two proofs is that at Theorem 1 instead

of $N_n^{st}+1-m_n$, $N_n^{st}+1-s_n$ repairs have to be completed until the end of a delay until repair initiation since the availability of a repair server is desired in this case.

For Theorem 5, note that from Theorem 4, (1.10) can be written in two different ways: for ED|ED choice in Figure 1.2,

$$Min \quad \bar{C}(\bar{s},\bar{m}) = w\bar{s} + c\bar{m} + p\left(\tfrac{1+\bar{m}}{\bar{s}\mu} - \tfrac{1}{\lambda} - \tfrac{1}{\mu}\right); \tag{A.216}$$

for all the other capacity level choices

$$Min \quad \bar{C}(\bar{s},\bar{m}) = w\bar{s} + c\bar{m}. \tag{A.217}$$

Notice that (A.217) implies, for the capacity level choices other than ED|ED, it is best for a given $\bar{s}$ value to have the minimum $\bar{m}$ value possible and vice versa. Moreover, from (A.216), in this region it is best for a given $\bar{s}$ value to have the minimum $\bar{m}$ value. This ensures that the solution will lie on the two line segments described as LS 1 and LS 2 as in the Proof of Theorem 2: LS 1 is the line segment for the equation $\bar{m}=0$ with end points $(0,0)$ and $(\lambda/(\lambda+\mu),0)$, namely in the region for ED|ED; and LS 2 is the line segment for the equation $\bar{m}=\bar{s}+\bar{s}/\rho-1$ with end points $(\lambda/(\lambda+\mu),0)$ and (ρ,ρ), namely in the region for QED|ED.

First we will consider QED|ED, applying $\bar{m}=\bar{s}+\bar{s}/\rho-1$ to $(A.217)$ we get

$$Min \quad \bar{C}(\bar{s},\bar{m})_{QED|ED} = \bar{s}\,(w+c+c/\rho) - c = 0. \tag{A.218}$$

which gives the unconstrained solution

$$\bar{s} = c/\,(w+c+c/\rho)\,, \bar{m} = [c\,(1+1/\rho)]\,/\,(w+c+c/\rho) - 1.$$

However, the constraint on this solution is that $\lambda/(\lambda+\mu) \le c/\,(w+c+c/\rho) \le \rho$; thus, if the solution above satisfies this, then we will have a zero asymptotic cost rate, which is optimal. Otherwise, one would think either inequalities could be the

case: $c/\left(w+c+c/\rho\right) < \lambda/(\lambda+\mu)$ or $c/\left(w+c+c/\rho\right) > \rho$. The second inequality means negative asymptotic cost rate for QED|ED; however, it leads to $0 > w\lambda + c\lambda$, which is not possible so we concentrate on the first inequality when the unconstrained solution does not lie in QED|ED.

For $c/\left(w+c+c/\rho\right) < \lambda/(\lambda+\mu)$, since it implies an $\bar{s}$ value smaller than equal to $\lambda/(\lambda+\mu)$ will be chosen, we will compare it to the solution for ED|ED. Note that in (A.216), for a specific value of $\bar{s}$, only the smallest value of $\bar{m}$ will give the minimum asymptotic cost rate. For LS 1, $\bar{m}=0$ and (A.216) becomes

$$Min \quad \bar{C}(\bar{s},\bar{m}) = w\bar{s} + p\left(\tfrac{1}{\bar{s}\mu} - \tfrac{1}{\lambda} - \tfrac{1}{\mu}\right). \tag{A.219}$$

The minimizing $\bar{s}$ value for this equation is given in the proof of Theorem 2, and it is $\bar{s} = \sqrt{p/(w\mu)}, \bar{m} = 0$. Thus, if this $\bar{s}$ value is smaller than $\lambda/(\lambda+\mu)$, then it is our overall optimum. If this is not the case, the function in A.219 is convex as demonstrated in the proof of Theorem 2. Thus, $\bar{s} = \lambda/(\lambda+\mu), \bar{m} = 0$ is the point on LS 1 giving the minimum cost rate. Moreover, remember that we came to this case because the unconstrained minimum for QED|ED was on the left side of LS 2; hence, $\bar{s} = \lambda/(\lambda+\mu), \bar{m} = 0$ is the minimizing point for QED|ED as well. This makes it the overall solution.

The proof for Theorem 6 is omitted since it is very similar to the proof of Theorem 3, because the only difference between these two proofs is that at Theorem 6 instead of $N_n^{st}+1-m_n$, $N_n^{st}+1-s_n$ repairs have to be completed until the end of a delay until repair initiation since the availability of a repair server is desired in this case.□

Appendix B

Proofs of Results in Chapter 2

B.1 Proof of Theorem 7

Let $b \leq 0$ and

$$Series_1^n = \sum_{k=1}^{\infty} \left[\left(\lambda n + \beta\sqrt{n}\right)^k / \prod_{i=1}^{k} \left(\lambda n + i\theta_1\right) \right] \tag{B.1a}$$

$$Series_2^n = \sum_{k=1}^{\infty} \left[\left(\lambda n\right)^k / \prod_{i=1}^{k} \left(\lambda n + \beta\sqrt{n} + i\theta_2\right) \right] \tag{B.1b}$$

$$Series_3^n = \sum_{k \leq b\sqrt{n}, k \in \mathbb{Z}} \left[\left(\lambda n\right)^{k^-} / \prod_{i=1}^{k^-} \left(\lambda n + \beta\sqrt{n} + i\theta_2\right) \right]. \tag{B.1c}$$

Then, from (2.1), Q^n has the following distribution:

$$P(Q^n(\infty) \leq b\sqrt{n}) = Series_3^n / \left(1 + Series_1^n + Series_2^n\right). \tag{B.2}$$

We will now obtain the limit of this expression as $n \to \infty$ by investigating each series.

$$
\begin{aligned}
Series_1 &= \sum_{k=1}^{\infty} \left[(\lambda n + \beta\sqrt{n})^k \Big/ \prod_{i=1}^{k} (\lambda n + i\theta_1) \right] \\
&= \sum_{k=1}^{\infty} \left[(\lambda n + \beta\sqrt{n})^k \Big/ \left((\theta_1)^k \prod_{i=1}^{k} \left(\tfrac{\lambda n}{\theta_1} + i \right) \right) \right].
\end{aligned} \tag{B.3}
$$

Note that

$$
\begin{aligned}
&\sum_{k=1}^{\infty} \left[(\lambda n + \beta\sqrt{n})^k \Big/ \left((\theta_1)^k \prod_{i=1}^{k} \left(\lceil \tfrac{\lambda n}{\theta_1} \rceil + i \right) \right) \right] \\
&\leq Series_1 \\
&\leq \sum_{k=1}^{\infty} \left[(\lambda n + \beta\sqrt{n})^k \Big/ \left((\theta_1)^k \prod_{i=1}^{k} \left(\lfloor \tfrac{\lambda n}{\theta_1} \rfloor + i \right) \right) \right].
\end{aligned} \tag{B.4}
$$

We will first consider the limiting behavior of the left side. From Stirling's approximation,

$$
\begin{aligned}
&\sum_{k=1}^{\infty} \left[(\lambda n + \beta\sqrt{n})^k \Big/ \left((\theta_1)^k \prod_{i=1}^{k} \left(\lceil \tfrac{\lambda n}{\theta_1} \rceil + i \right) \right) \right] \\
&= \left(\lceil \tfrac{\lambda n}{\theta_1} \rceil \right)! \sum_{k=1}^{\infty} \left[\left(\tfrac{\lambda n}{\theta_1} + \tfrac{\beta\sqrt{n}}{\theta_1} \right)^{\lceil \frac{\lambda n}{\theta_1} \rceil + k} \Big/ \left(\left(\lceil \tfrac{\lambda n}{\theta_1} \rceil + k \right)! \left(\tfrac{\lambda n}{\theta_1} + \tfrac{\beta\sqrt{n}}{\theta_1} \right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \right) \right] \\
&\equiv \left[\left(\sqrt{2\pi} \sqrt{\lceil \tfrac{\lambda n}{\theta_1} \rceil} \left(\lceil \tfrac{\lambda n}{\theta_1} \rceil \right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \exp\{ - \lceil \tfrac{\lambda n}{\theta_1} \rceil \} \right) \Big/ \left(\left(\tfrac{\lambda n}{\theta_1} + \tfrac{\beta\sqrt{n}}{\theta_1} \right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \right) \right] \\
&\cdot \sum_{k=\lceil \frac{\lambda n}{\theta_1} \rceil + 1}^{\infty} \left(\tfrac{\lambda n}{\theta_1} + \tfrac{\beta\sqrt{n}}{\theta_1} \right)^k / k!.
\end{aligned} \tag{B.5}
$$

Let

$$
\begin{aligned}
A^n &= \left(\left(\tfrac{\lambda n}{\theta_1} + \tfrac{\beta\sqrt{n}}{\theta_1} \right) / \lceil \tfrac{\lambda n}{\theta_1} \rceil \right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \cdot \left(\lceil \tfrac{\lambda n}{\theta_1} \rceil / \tfrac{\lambda n}{\theta_1} \right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \cdot \left(\tfrac{\lambda n}{\theta_1} / \lceil \tfrac{\lambda n}{\theta_1} \rceil \right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \\
B^n &= \exp\{ - \lceil \tfrac{\lambda n}{\theta_1} \rceil \} / \exp\{ - \tfrac{\lambda n}{\theta_1} \},
\end{aligned} \tag{B.6}
$$

then,

$$\begin{aligned}
&\left[\left(\sqrt{2\pi}\,\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}\;\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil\right)^{\lceil\frac{\lambda n}{\theta_1}\rceil}\;\exp\{-\lceil\tfrac{\lambda n}{\theta_1}\rceil\}\right)\Big/\left(\tfrac{\lambda n}{\theta_1}+\tfrac{\beta\sqrt{n}}{\theta_1}\right)^{\lceil\frac{\lambda n}{\theta_1}\rceil}\right]\\
&\cdot\sum_{k=\lceil\frac{\lambda n}{\theta_1}\rceil+1}^{\infty}\left(\tfrac{\lambda n}{\theta_1}+\tfrac{\beta\sqrt{n}}{\theta_1}\right)^{k}/k!\\
&=\left(\sqrt{2\pi}\,\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}\exp\{\tfrac{\beta\sqrt{n}}{\theta_1}\}\;B/A\right)\\
&\cdot\sum_{k=\lceil\frac{\lambda n}{\theta_1}\rceil+1}^{\infty}\exp\{-\left(\tfrac{\lambda n}{\theta_1}+\tfrac{\beta\sqrt{n}}{\theta_1}\right)\}\left(\tfrac{\lambda n}{\theta_1}+\tfrac{\beta\sqrt{n}}{\theta_1}\right)^{k}/k!.
\end{aligned}\tag{B.7}$$

Now let $Y_n \sim Poisson\left(\frac{\lambda n}{\theta_1}+\frac{\beta\sqrt{n}}{\theta_1}\right)$. Then we have

$$\begin{aligned}
&\left(\sqrt{2\pi}\,\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}\exp\{\tfrac{\beta\sqrt{n}}{\theta_1}\}\;B/A\right)\\
&\cdot\sum_{k=\lceil\frac{\lambda n}{\theta_1}\rceil+1}^{\infty}\exp\{-\left(\tfrac{\lambda n}{\theta_1}+\tfrac{\beta\sqrt{n}}{\theta_1}\right)\}\left(\tfrac{\lambda n}{\theta_1}+\tfrac{\beta\sqrt{n}}{\theta_1}\right)^{k}/k!\\
&=\left[\sqrt{2\pi}\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}/\left(1+\tfrac{\beta}{\lambda\sqrt{n}}\right)^{\lceil\frac{\lambda n}{\theta_1}\rceil}\right]\cdot\exp\left\{\tfrac{\beta\sqrt{n}}{\theta_1}\right\}\\
&\cdot\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil/\tfrac{\lambda n}{\theta_1}\right)^{\lceil\frac{\lambda n}{\theta_1}\rceil}\cdot\left[\exp\{-\lceil\tfrac{\lambda n}{\theta_1}\rceil\}/\exp\{-\tfrac{\lambda n}{\theta_1}\}\right]\cdot P\left(Y_n>\lceil\tfrac{\lambda n}{\theta_1}\rceil\right)\\
&=\left[\sqrt{2\pi}\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}\exp\left\{\tfrac{\beta\sqrt{n}}{\theta_1}\right\}/\exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil\ln\left(1+\tfrac{\beta}{\lambda\sqrt{n}}\right)\right\}\right]\\
&\cdot\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil/\tfrac{\lambda n}{\theta_1}\right)^{\lceil\frac{\lambda n}{\theta_1}\rceil}\cdot\left[\exp\{-\lceil\tfrac{\lambda n}{\theta_1}\rceil\}/\exp\{-\tfrac{\lambda n}{\theta_1}\}\right]\cdot P\left(Y_n>\lceil\tfrac{\lambda n}{\theta_1}\rceil\right)
\end{aligned}\tag{B.8}$$

and now we will look at the limiting behavior of the terms in the last expression. x_n is $o(f(n))$ if $\frac{x_n}{f(n)}\to 0$ as $n\to\infty$. Using Taylor's approximation for $f(y)=\ln(1+y)$ at $y_0=0$ ($f(y)=y-y^2/2+y^3/3-\ldots$) gives

$$\exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil \ln\left(1+\tfrac{\beta}{\lambda\sqrt{n}}\right)\right\} = \exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil\left[\tfrac{\beta}{\lambda\sqrt{n}} - \tfrac{\beta^2}{\lambda^2 2n} + o(1/n)\right]\right\}; \tag{B.9}$$

since $\lceil\frac{\lambda n}{\theta_1}\rceil - \frac{\lambda n}{\theta_1} \leq 1$,

$$\begin{aligned}&\lim_{n\to\infty}\exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil\left[\tfrac{\beta}{\lambda\sqrt{n}} - \tfrac{\beta^2}{\lambda^2 2n} + o(1/n)\right]\right\}\\ &= \lim_{n\to\infty}\exp\left\{\left(\tfrac{\lambda n}{\theta_1}\right)\left[\tfrac{\beta}{\lambda\sqrt{n}} - \tfrac{\beta^2}{\lambda^2 2n} + o(1/n)\right]\right\}.\end{aligned} \tag{B.10}$$

Moreover, again using Taylor's approximation for $f(y) = \ln(1+y)$ at $y_0 = 0$

$$\begin{aligned}&\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil/\tfrac{\lambda n}{\theta_1}\right)^{\lceil\frac{\lambda n}{\theta_1}\rceil}\cdot\left[\exp\{-\lceil\tfrac{\lambda n}{\theta_1}\rceil\}/\exp\{-\tfrac{\lambda n}{\theta_1}\}\right]\\ &= \exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil\ln\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil/\tfrac{\lambda n}{\theta_1}\right) - \lceil\tfrac{\lambda n}{\theta_1}\rceil + \tfrac{\lambda n}{\theta_1}\right\}\\ &= \exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil\ln\left(1+\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil - \tfrac{\lambda n}{\theta_1}\right)/\tfrac{\lambda n}{\theta_1}\right) - \lceil\tfrac{\lambda n}{\theta_1}\rceil + \tfrac{\lambda n}{\theta_1}\right\}\\ &= \exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil - \tfrac{\lambda n}{\theta_1}\right)/\tfrac{\lambda n}{\theta_1} + o(1/n) - \lceil\tfrac{\lambda n}{\theta_1}\rceil + \tfrac{\lambda n}{\theta_1}\right\} \to \exp\{0\} = 1.\end{aligned} \tag{B.11}$$

Continuing (B.8), from (B.9)-(B.11), for sufficiently large n,

$$\begin{aligned}&\left[\sqrt{2\pi}\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}\exp\left\{\tfrac{\beta\sqrt{n}}{\theta_1}\right\}/\exp\left\{\lceil\tfrac{\lambda n}{\theta_1}\rceil\ln\left(1+\tfrac{\beta}{\lambda\sqrt{n}}\right)\right\}\right]\\ &\cdot\left(\lceil\tfrac{\lambda n}{\theta_1}\rceil/\tfrac{\lambda n}{\theta_1}\right)^{\lceil\frac{\lambda n}{\theta_1}\rceil}\cdot\left[\exp\{-\lceil\tfrac{\lambda n}{\theta_1}\rceil\}/\exp\{-\tfrac{\lambda n}{\theta_1}\}\right]\cdot P\left(Y_n > \lceil\tfrac{\lambda n}{\theta_1}\rceil\right)\\ &= \left(\sqrt{2\pi}\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}\exp\left\{\tfrac{\beta\sqrt{n}}{\theta_1}\right\}P\left(Y_n > \lceil\tfrac{\lambda n}{\theta_1}\rceil\right)\right)/\exp\left\{\tfrac{\beta\sqrt{n}}{\theta_1} - \tfrac{\beta^2}{2\theta_1\lambda} + o(1)\right\}\\ &= \sqrt{2\pi}\sqrt{\lceil\tfrac{\lambda n}{\theta_1}\rceil}\exp\left\{\tfrac{\beta^2}{2\theta_1\lambda}\right\}P\left(Y_n > \lceil\tfrac{\lambda n}{\theta_1}\rceil\right).\end{aligned} \tag{B.12}$$

Since $Y_n \sim Poisson\left(\frac{\lambda n}{\theta_1} + \frac{\beta\sqrt{n}}{\theta_1}\right)$, from CLT, for sufficiently large n,

$$\sqrt{2\pi}\sqrt{\lceil \tfrac{\lambda n}{\theta_1} \rceil}\exp\left\{\tfrac{\beta^2}{2\theta_1\lambda}\right\} P\left(Y_n > \lceil \tfrac{\lambda n}{\theta_1} \rceil\right) \cong \sqrt{2\pi}\sqrt{\lceil \tfrac{\lambda n}{\theta_1} \rceil}\exp\left\{\tfrac{\beta^2}{2\theta_1\lambda}\right\} \Phi\left(\beta/\sqrt{\lambda\theta_1}\right); \tag{B.13}$$

and from (B.5)-(B.13), with $h(x) = \phi(x)/\Phi(-x)$, the behavior of the left side of (B.4) for large values of n can be expressed as

$$\begin{aligned} &\sum_{k=1}^{\infty}\left[(\lambda n + \beta\sqrt{n})^k \Big/ \left((\theta_1)^k \prod_{i=1}^{k}\left(\lceil \tfrac{\lambda n}{\theta_1} \rceil + i\right)\right)\right] \\ &\cong \sqrt{2\pi}\sqrt{\lceil \tfrac{\lambda n}{\theta_1} \rceil}\exp\left\{\tfrac{\beta^2}{2\theta_1\lambda}\right\} \Phi\left(\beta/\sqrt{\lambda\theta_1}\right) \\ &= \sqrt{\lceil \tfrac{\lambda n}{\theta_1} \rceil}\left\{h\left(-\tfrac{\beta}{\sqrt{\lambda\theta_1}}\right)\right\}^{-1}. \end{aligned} \tag{B.14}$$

Similarly, for sufficiently large n the right side of (B.4) can be expressed as

$$\begin{aligned} &\sum_{k=1}^{\infty}\left[(\lambda n + \beta\sqrt{n})^k \Big/ \left((\theta_1)^k \prod_{i=1}^{k}\left(\lfloor \tfrac{\lambda n}{\theta_1} \rfloor + i\right)\right)\right] \\ &\cong \sqrt{\lfloor \tfrac{\lambda n}{\theta_1} \rfloor}\left\{h\left(-\tfrac{\beta}{\sqrt{\lambda\theta_1}}\right)\right\}^{-1}. \end{aligned} \tag{B.15}$$

The limiting behavior of $Series_2$ and $Series_3$ can be analyzed analogously. Thus, from (B.4)-(B.15), for sufficiently large n,

$$\begin{aligned} Series_1^n &\cong \sqrt{\tfrac{\lambda n}{\theta_1}}\left\{h\left(-\tfrac{\beta}{\sqrt{\lambda\theta_1}}\right)\right\}^{-1} \\ Series_2^n &\cong \sqrt{\tfrac{\lambda n}{\theta_2} + \tfrac{\beta\sqrt{n}}{\theta_2}}\left\{h\left(\tfrac{\beta}{\sqrt{\lambda\theta_2}}\right)\right\}^{-1} \\ Series_3^n &\cong \sqrt{\tfrac{\lambda n}{\theta_2} + \tfrac{\beta\sqrt{n}}{\theta_2}} \quad \tfrac{\Phi\left((b-\beta/\theta_2)/\sqrt{\lambda/\theta_2}\right)}{\phi\left(\beta/\sqrt{\lambda\theta_2}\right)}, \end{aligned} \tag{B.16}$$

which gives, from (B.2), for $b \leq 0$,

$$\begin{aligned}
&\lim_{n\to\infty} P(Q^n(\infty) \le b\sqrt{n}) \\
&= \left(\sqrt{\theta_1}\Phi\left(\left[b-\beta/\theta_2\right]/\sqrt{\lambda/\theta_2}\right)/\phi\left(\beta/\sqrt{\lambda\theta_2}\right)\right) \\
&/\left(\sqrt{\theta_1}\left\{h\left(\beta/\sqrt{\lambda\theta_2}\right)\right\}^{-1}+\sqrt{\theta_2}\left\{h\left(-\beta/\sqrt{\lambda\theta_1}\right)\right\}^{-1}\right).
\end{aligned} \tag{B.17}$$

For $b = 0$, we have $P(Q^n(\infty) \le 0)$. Since our inspiring example is a make-to-stock queue with perishable goods and impatient customers, in this setting this quantity corresponds to out-of-stock probability; which will be denoted by a and shown below:

$$\begin{aligned}
&a \equiv \lim_{n\to\infty} P(Q^n(\infty) \le 0) \\
&= \left(\sqrt{\theta_1}\left\{h\left(\beta/\sqrt{\lambda\theta_2}\right)\right\}^{-1}\right) \\
&/\left(\sqrt{\theta_1}\left\{h\left(\beta/\sqrt{\lambda\theta_2}\right)\right\}^{-1}+\sqrt{\theta_2}\left\{h\left(-\beta/\sqrt{\lambda\theta_1}\right)\right\}^{-1}\right) = \left(1+\sqrt{\theta_2/\theta_1}\frac{h\left(\beta/\sqrt{\lambda\theta_2}\right)}{\left\{h\left(-\beta/\sqrt{\lambda\theta_1}\right)\right\}}\right)^{-1}.
\end{aligned} \tag{B.18}$$

Note that a truncated normal random variable denoted by $N(\mu, \sigma^2, l, u)$ (as explained in Section 2.2) has the distribution function

$$P\left(N(\mu,\sigma^2,l,u) < x\right) = \left[\Phi\left(\tfrac{x-\mu}{\sigma}\right) - \Phi\left(\tfrac{l-\mu}{\sigma}\right)\right] / \left[\Phi\left(\tfrac{u-\mu}{\sigma}\right) - \Phi\left(\tfrac{l-\mu}{\sigma}\right)\right]. \tag{B.19}$$

Thus, for $X^- \sim N\left(\frac{\beta}{\theta_2}, \frac{\lambda}{\theta_2}, -\infty, 0\right)$ and $X^+ \sim N\left(\frac{\beta}{\theta_1}, \frac{\lambda}{\theta_1}, 0, \infty\right)$ we have (remember that $b \le 0$)

$$\begin{aligned}
P\left(X^- < b\right) &= \left[\Phi\left(\tfrac{b-\beta/\theta_2}{\sqrt{\lambda/\theta_2}}\right)\right] / \left[\Phi\left(\tfrac{-\beta/\theta_2}{\sqrt{\lambda/\theta_2}}\right)\right] \\
P\left(X^+ < b\right) &= 0.
\end{aligned} \tag{B.20}$$

It can be seen that

$$\begin{aligned}
&\lim_{n\to\infty} P(Q^n(\infty) \le b\sqrt{n}) \\
&= a \cdot \Phi\left(\left[b - \beta/\theta_2\right] / \sqrt{\lambda/\theta_2}\right) / \Phi\left(-\beta/\sqrt{\lambda\theta_2}\right) \\
&= a \cdot P\left(X^- \le b\right) + (1-a) \cdot P\left(X^+ \le b\right).
\end{aligned} \tag{B.21}$$

Now let $b > 0$, and let

$$Series_4^n = \sum_{1\le k\le b\sqrt{n},\, k\in\mathbb{Z}} \left[\left(\lambda n + \beta\sqrt{n}\right)^k / \prod_{i=1}^{k} \left(\lambda n + i\theta_1\right) \right]. \tag{B.22}$$

Then,

$$P(Q^n(\infty) \le b\sqrt{n}) = P(Q^n(\infty) \le 0) + Series_4 / \left(1 + Series_1 + Series_2\right). \tag{B.23}$$

The limiting behavior of $Series_4$ can be analyzed similar to $Series_1$. The only difference is in the summation, which leads to

$$P\left(\lceil \frac{\lambda n}{\theta_1} \rceil < Y_n < \lceil \frac{\lambda n}{\theta_1} \rceil + b\sqrt{n} \right)$$

in (B.8) and (B.12) instead of $P\left(Y_n > \lceil \frac{\lambda n}{\theta_1} \rceil\right)$. Hence, for sufficiently large n,

$$\begin{aligned}
Series_4^n \cong \sqrt{\tfrac{\lambda n}{\theta_1}} \quad & \left[\Phi\left(\left(b - \beta/\theta_1\right) / \sqrt{\lambda/\theta_1}\right) - \Phi\left(\left(-\beta/\theta_1\right) / \sqrt{\lambda/\theta_1}\right)\right] \\
& / \phi\left(-\beta/\sqrt{\lambda\theta_1}\right),
\end{aligned} \tag{B.24}$$

and we have again

$$\lim_{n\to\infty} P(Q^n(\infty) \le b\sqrt{n})$$

$$= a + \left(\sqrt{\theta_2}\left\{\left[\Phi\left(\left(b-\beta/\theta_1\right)/\sqrt{\lambda/\theta_1}\right) - \Phi\left(\left(-\beta/\theta_1\right)/\sqrt{\lambda/\theta_1}\right)\right]/\phi\left(-\beta/\sqrt{\lambda\theta_1}\right)\right\}\right)$$

$$/\left(\sqrt{\theta_1}\left\{h\left(\beta/\sqrt{\lambda\theta_2}\right)\right\}^{-1} + \sqrt{\theta_2}\left\{h\left(-\beta/\sqrt{\lambda\theta_1}\right)\right\}^{-1}\right). \tag{B.25}$$

Since, for $b > 0$,

$$P\left(X^- < b\right) = 1$$

$$P\left(X^+ < b\right) = \left[\Phi\left(\left(b-\beta/\theta_1\right)/\sqrt{\lambda/\theta_1}\right) - \Phi\left(\left(-\beta/\theta_1\right)/\sqrt{\lambda/\theta_1}\right)\right] / \left[\Phi\left(\frac{\beta/\theta_1}{\sqrt{\lambda/\theta_1}}\right)\right], \tag{B.26}$$

we have again

$$\lim_{n\to\infty} P(Q^n(\infty) \le b\sqrt{n}) = a \cdot P\left(X^- \le b\right) + (1-a) \cdot P\left(X^+ \le b\right). \square \tag{B.27}$$

B.2 Proof of Theorem 8

Let $d > 0$ and for any $b \in \mathbb{R}$, $\left(\frac{dn}{\theta_1} + b\sqrt{n}\right) > 0$ will hold for sufficiently large n.

$$Series_1^n = \sum_{k=1}^{\infty}\left[\left(\lambda n + dn\right)^k / \prod_{i=1}^{k}\left(\lambda n + i\theta_1\right)\right] \tag{B.28a}$$

$$Series_2^n = \sum_{k=1}^{\infty}\left[\left(\lambda n\right)^k / \prod_{i=1}^{k}\left(\lambda n + dn + i\theta_2\right)\right] \tag{B.28b}$$

$$Series_3^n = \sum_{k\le dn/\theta_1 + b\sqrt{n}, k\in\mathbb{Z}}\left[\left(\lambda n + dn\right)^k / \prod_{i=1}^{k}\left(\lambda n + i\theta_1\right)\right]. \tag{B.28c}$$

Then, from (2.1), Q^n has the following distribution:

$$P(Q^n(\infty) \leq dn/\theta_1 + b\sqrt{n}) = (Series_2^n + Series_3^n)/\left(1 + Series_1^n + Series_2^n\right). \quad \text{(B.29)}$$

We will now obtain the limit of this expression as $n \to \infty$ by investigating each series. Similar to the previous section, (B.3)-(B.8) are valid by exchanging $\beta\sqrt{n}$ with dn, which also gives $\ln(1 + d/\lambda)$ instead of $\ln(1 + \beta/\lambda\sqrt{n})$ in (B.8). Now we will look at the limiting behavior of the terms in the expression below which is analogous to (B.8) in the unbalanced case:

$$\begin{aligned} &\left[\sqrt{2\pi}\sqrt{\lceil \tfrac{\lambda n}{\theta_1} \rceil}\exp\left\{\tfrac{dn}{\theta_1}\right\} / \exp\left\{\lceil \tfrac{\lambda n}{\theta_1} \rceil \ln\left(1 + \tfrac{d}{\lambda}\right)\right\}\right] \\ &\cdot \left(\lceil \tfrac{\lambda n}{\theta_1} \rceil / \tfrac{\lambda n}{\theta_1}\right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \cdot \left[\exp\{-\lceil \tfrac{\lambda n}{\theta_1} \rceil\} / \exp\{-\tfrac{\lambda n}{\theta_1}\}\right] \cdot P\left(Y_n > \lceil \tfrac{\lambda n}{\theta_1} \rceil\right). \end{aligned} \quad \text{(B.30)}$$

Note that $\ln(1+x) - x < 0$ for $x > 0$ since its value is 0 at $x = 0$ and its derivative is negative for positive x. Then, as $n \to \infty$

$$\begin{aligned} &\ln(1 + d/\lambda) - d/\lambda < 0 \\ &\Rightarrow -(\lambda n/\theta_1)\ln(1 + d/\lambda) + dn/\theta_1 > 0 \\ &\Rightarrow \exp\left\{\tfrac{dn}{\theta_1}\right\} / \exp\left\{\lceil \tfrac{\lambda n}{\theta_1} \rceil \ln\left(1 + \tfrac{d}{\lambda}\right)\right\} \to \infty. \end{aligned} \quad \text{(B.31)}$$

From (B.11)

$$\left(\lceil \tfrac{\lambda n}{\theta_1} \rceil / \tfrac{\lambda n}{\theta_1}\right)^{\lceil \frac{\lambda n}{\theta_1} \rceil} \cdot \left[\exp\{-\lceil \tfrac{\lambda n}{\theta_1} \rceil\} / \exp\{-\tfrac{\lambda n}{\theta_1}\}\right] \to 1, \quad \text{(B.32)}$$

and from CLT, $P\left(Y_n > \lceil \tfrac{\lambda n}{\theta_1} \rceil\right) \to 1$ (since $d > 0$ and $Y_n \sim Poisson\left(\tfrac{(\lambda+d)n}{\theta_1}\right)$).

Observe that since $d > 0$,

$$
\begin{aligned}
Series_2^n &= \sum_{k=1}^{\infty} \left[(\lambda n)^k / \prod_{i=1}^{k} (\lambda n + dn + i\theta_2) \right] \\
&\leq \sum_{k=1}^{\infty} \left(\lambda n / (\lambda n + dn) \right)^k < \infty.
\end{aligned}
\tag{B.33}
$$

$Series_3^n$ can be analyzed analogously to $Series_1^n$, and this leads to have

$$
P\left(\lceil \frac{\lambda n}{\theta_1} \rceil < Y_n \leq \lceil \frac{\lambda n}{\theta_1} \rceil + dn/\theta_1 + b\sqrt{n} \right)
$$

instead of $P\left(Y_n > \lceil \frac{\lambda n}{\theta_1} \rceil\right)$ in (B.8) and afterwards. Hence, for sufficiently large n,

$$
Series_3^n \cong \sqrt{2\pi}\sqrt{\tfrac{\lambda n}{\theta_1}} \quad \Phi\left(b/\sqrt{(\lambda + d)/\theta_1} \right). \tag{B.34}
$$

From (B.29)-(B.34),

$$
\begin{aligned}
& P\left(Q^n(\infty) < dn/\theta_1 + b\sqrt{n}\right) \\
& \cong \sqrt{2\pi}\sqrt{\tfrac{\lambda n}{\theta_1}}\, \Phi\left(b/\sqrt{(\lambda + d)/\theta_1} \right) \exp\left\{ \tfrac{dn}{\theta_1} \right\} / \exp\left\{ \lceil \tfrac{\lambda n}{\theta_1} \rceil \ln\left(1 + \tfrac{d}{\lambda}\right) \right\} \\
& / \left[1 + \sqrt{2\pi}\sqrt{\tfrac{\lambda n}{\theta_1}} \exp\left\{ \tfrac{dn}{\theta_1} \right\} / \exp\left\{ \lceil \tfrac{\lambda n}{\theta_1} \rceil \ln\left(1 + \tfrac{d}{\lambda}\right) \right\} + Series_2 \right] \\
& \to \Phi\left(b/\sqrt{(\lambda + d)/\theta_1} \right).
\end{aligned}
\tag{B.35}
$$

The case for $d < 0$ is very similar and hence omitted.□

B.3 Proof of Theorems 9-11

Model (1) follows the same way as in the proof of Theorem 7 with $Series_1$ having the upper bound $T\sqrt{n}$. This will lead to have $P\left(\lceil \frac{\lambda n}{\theta_1} \rceil < Y_n \leq \lceil \frac{\lambda n}{\theta_1} \rceil + T\sqrt{n} \right)$ instead of $P\left(\lceil \frac{\lambda n}{\theta_1} \rceil < Y_n \right)$ in (B.8) and (B.12) where $Y_n \sim Poisson\left(\frac{\lambda n}{\theta_1} + \frac{\beta\sqrt{n}}{\theta_1} \right)$. Thus, we

will have $\Phi\left(T\sqrt{\theta_1/\lambda}-\beta/\sqrt{\lambda\theta_1}\right)-\Phi\left(-\beta/\sqrt{\lambda\theta_1}\right)$ instead of $\Phi\left(\beta/\sqrt{\lambda\theta_1}\right)$ in (B.13). Hence, for $b<0$

$$\begin{aligned} Series_1^n &\cong \sqrt{\tfrac{\lambda n}{\theta_1}}\,\tfrac{\Phi\left(T\sqrt{\theta_1/\lambda}-\beta/\sqrt{\lambda\theta_1}\right)-\Phi\left(-\beta/\sqrt{\lambda\theta_1}\right)}{\phi\left(-\beta/\sqrt{\lambda\theta_1}\right)} \\ Series_2^n &\cong \sqrt{\tfrac{\lambda n}{\theta_2}+\tfrac{\beta\sqrt{n}}{\theta_2}}\left\{h\left(\tfrac{\beta}{\sqrt{\lambda\theta_2}}\right)\right\}^{-1} \\ Series_3^n &\cong \sqrt{\tfrac{\lambda n}{\theta_2}+\tfrac{\beta\sqrt{n}}{\theta_2}}\quad \tfrac{\Phi\left((b-\beta/\theta_2)/\sqrt{\lambda/\theta_2}\right)}{\phi\left(\beta/\sqrt{\lambda\theta_2}\right)}, \end{aligned} \tag{B.36}$$

which gives, from (B.2) for $b=0$,

$$a=\left(1+\sqrt{\theta_2/\theta_1}\quad \tfrac{\Phi\left(T\sqrt{\theta_1/\lambda}-\beta/\sqrt{\lambda\theta_1}\right)-\Phi\left(-\beta/\sqrt{\lambda\theta_1}\right)}{\phi\left(-\beta/\sqrt{\lambda\theta_1}\right)}\cdot h\left(\beta/\sqrt{\lambda\theta_2}\right)\right)^{-1} \tag{B.37}$$

and the rest of the proof for Model 1 follows as in Theorem 7. For models (2) and (3) the difference w.r.t the proof of Theorem 7 will be in $Series_1^n$ and $Series_4^n$. We will explore the limiting behavior of $Series_1^n$ and $Series_4^n$ will be similar.

$$\begin{aligned} &\sum_{k=1}^{T\sqrt{n}}\left(\tfrac{\lambda n+\beta\sqrt{n}}{\lambda n}\right)^k=\left(\tfrac{\lambda n+\beta\sqrt{n}}{-\beta\sqrt{n}}\right)\left[1-\left(\tfrac{\lambda n+\beta\sqrt{n}}{\lambda n}\right)^{T\sqrt{n}}\right] \\ &\cong(-1)\left(1+\tfrac{\lambda\sqrt{n}}{\beta}\right)\left[1-\exp\left\{\tfrac{\beta T}{\lambda}\right\}\right]. \end{aligned} \tag{B.38}$$

From (B.2) and (B.36), for $b\le 0$ we have

$$\begin{aligned} &P(Q^n(\infty)\le b\sqrt{n}) \\ &\cong\sqrt{\tfrac{\lambda n}{\theta_2}+\tfrac{\beta\sqrt{n}}{\theta_2}}\quad \tfrac{\Phi\left((b-\beta/\theta_2)/\sqrt{\lambda/\theta_2}\right)}{\phi\left(\beta/\sqrt{\lambda\theta_2}\right)} \\ &/\left(1-\left(1+\tfrac{\lambda\sqrt{n}}{\beta}\right)\left[1-\exp\left\{\tfrac{\beta T}{\lambda}\right\}\right]+\sqrt{\tfrac{\lambda n}{\theta_2}+\tfrac{\beta\sqrt{n}}{\theta_2}}\left\{h\left(\tfrac{\beta}{\sqrt{\lambda\theta_2}}\right)\right\}^{-1}\right), \end{aligned} \tag{B.39}$$

and the rest follows. □

Bibliography

[1] R. E. Barlow. Repairmen problems. In H. Scarf K. J. Arrow, S. Karlin, editor, *Studies in Applied Probability and Management Science*, pages 18–33. Stanford University Press, Stanford, California, 1962.

[2] B. Biais, L. Glosten, and C. Spatt. Market microstructure: a survey of microfoundations, empirical results, and policy implications. *Journal of Financial Markets*, 8(2):217–264, 2005.

[3] S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, Jan/Feb 2004.

[4] S. Browne and W. Whitt. Piecewise-linear diffusion processes. In J. Dshalalow, editor, *Advances in Queueing: Theory, Methods, and Open Problems*, pages 463–480. CRC Press, Boca Raton, FL, 1995.

[5] D. I. Cho and M. Parlar. A survey of maintenance models for multi-unit systems. *European J. of Op. Res.*, 51(1):1–23, March 1991.

[6] B. W. Conolly, P. R. Parthasarathy, and N. Selvaraju. Double-ended queues with impatience. *Computers & Operations Research*, 29(14):2053–2072, December 2002.

[7] F. de Véricourt and O. B. Jennings. Dimensioning large-scale membership services. *Operations Research*, 56(1):173–187, Jan/Feb 2008.

[8] M. B. Garman. Market microstructure. *Journal of Fin. Econ.*, 3:257–275, 1976.

[9] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, Summer 2002.

[10] S. K. Goyal and B. C. Giri. Recent trends in modeling of deteriorating inventory. *European Journal of Operational Research*, 134(1):1–16, March 2001.

[11] S. Graves. The application of queueing theory to continuous perishable inventory systems. *Management Science*, 28(4):400–406, April 1982.

[12] D. Gross, H. D. Kahn, and J. D. Marsh. Queueing models for spares provisioning. *Naval Research Logistics Quarterly*, 24(4):521–536, December 1977.

[13] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, May/Jun 1981.

[14] L. Haque and M. J. Armstrong. A survey of the machine interference problem. *European J. of Op. Res.*, 179(2):469–482, August 2007.

[15] J. E. Hilliard. An approach to cost analysis of maintenance float systems. *AIIE Transactions*, 8(1):128–133, March 1976.

[16] U. Horst and C. Rothe. Queueing, social interactions, and the microstructure of financial markets. *Macroeconomic Dynamics*, 12:211–233, 2008.

[17] D. L. Iglehart. Limiting diffusion approximations for the many server queue and the repairmen problem. *Journal of Applied Probability*, 2(2):429–441, December 1965.

[18] D. L. Iglehart and A. J. Lemoine. Approximations for the repairman problem with two repair facilities, ii: Spares. *Advances in Applied Probability*, 6(1):147–158, March 1974.

[19] S. Ishihara and Y. Utsumi. Time lag dynamics and taxi cycle: taxi service problem. *Sociological Theory and Methods*, 20(2):227–240, July 2005.

[20] M. Jain. An (m, m) machine repair problem with spares and state dependent rates: a diffusion process approach. *Microelectronics Reliability*, 37(6):929–933, June 1997.

[21] M. Jain, G. C. Sharma, , and M. Singh. M/m/r machine interference model with balking, reneging, spares and two modes of failure. *OPSEARCH*, 40(1):24–41, January 2003.

[22] M. Jain, G. C. Sharma, and R. Sharma. Performance modeling of state dpendent system with mixed standbys and two modes of failure. *Applied Mathematical Modelling*, 32(5):712–724, May 2008.

[23] A. J. E. M. Janssen, J. S. H. Van Leeuwaarden, and B. Zwart. Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Adv. Appl. Prob.*, 40(1):122–143, March 2008.

[24] O. B. Jennings and J. E. Reed. An overloaded multiclass fifo queue with abandonments. Submitted to *Op. Res.*, 2010.

[25] J. C. Ke, S. L. Lee, , and C. H. Liou. Machine repair problem in production systems with spares and server vacations. *RAIRO Operations Research*, 43(1):35–54, January 2009.

[26] J. C. Ke and K. H. Wang. Vacation policies for machine repair problem with two type spares. *Applied Mathematical Modelling*, 31(5):880–894, May 2007.

[27] A. Mandelbaum and G. Pats. Stochastic networks. In F. P. Kelly and R. J. Williams, editors, *IMA Volumes in Mathematics*, pages 239–282. Springer, 1995.

[28] A. Mandelbaum and G. Pats. State-dependent stochastic networks. part 1: approximations and applications with continuous diffusion limits. *The Annals of Applied Probability*, 8(2):569–646, May 1998.

[29] G. Mendoza, M. Sedaghat, and K .P. Yoon. Queueing models to balance systems with excess supply. *International Business & Economics Research Journal*, 8(1):91–104, January 2009.

[30] S. Nahmias. Perishable inventory theory: A review. *Operations Research*, 30(4):680–708, Jul/Aug 1982.

[31] F. Raafat. Survey of literature on continuously deteriorating inventory models. *The Journal of Operational Research Society*, 42(1):27–37, January 1991.

[32] J. E. Reed and A. R. Ward. Approximating the gi/gi/1+gi queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33(3):606–644, August 2008.

[33] C. J. Singh and M. Jain. (m, m) machine repair problem with spares and reneging. *Pak. J. Statistics*, 23(1):23–35, January 2007.

[34] B. D. Sivazlian and K. H. Wang. Economic analysis of the m/m/r machine repair problem with warm standbys. *Microelectronics Reliability*, 29(1):25–35, January 1989.

[35] K. E. Stecke and J. E. Aronson. Review of operator/machine interference models. *Intl. J. of Prod. Res.*, 23(1):129–151, Jan/Feb 1985.

[36] C. Stone. Limit theorems for random walks, birth and eath processes and diffusion processes. *Illinois J. Mathematics*, 7(4):638–660, December 1963.

[37] K. H. Wang. Cost analysis of the m/m/r machine-repair problem with mixed standby spares. *Microelectronics Reliability*, 33(9):1293–1301, September 1993.

[38] K. H. Wang. Profit analysis of the machine repair problem with cold standbys and two modes of failure. *Microelectronics Reliability*, 34(10):1635–1642, October 1994.

[39] K. H. Wang. An approach to cost analysis of the machine repair problem with two types of spares and service rates. *Microelectronics Reliability*, 35(11):1433–1436, November 1995.

[40] K. H. Wang and H. C. Lee. Cost analysis of the cold-standby m/m/r machine repair problem with multiple modes of failure. *Microelectronics Reliability*, 38(3):435–441, March 1998.

[41] K. H. Wang and B. D. Sivazlian. Cost analysis of the m/m/r machine repair problem with spares operating under variable service rates. *Microelectronics Reliability*, 32(8):1171–1183, August 1992.

[42] A. R. Ward and P. W. Glynn. A diffusion approximation for a gi/gi/1 queue with balking or reneging. *Queuing Systems*, 50(4):371–400, August 2005.

[43] Y. Xiao-ling and S. Jian. Double-ended queue system with negative customers, 2004. http://scholar.ilib.cn/A-QCode zsdxxb200404004.html.

[44] S. A. Zenios. Modeling the transplant waiting list: A queueing model with reneging. *Queuing Systems*, 31(3-4):239–251, July 1999.

Biography

Isilay, Talay Degirmenci

Birth

Antalya, Turkey on January 19, 1981

Education

- PhD in Business Administration, Duke University, Fuqua School of Business, Durham, NC, USA
- M.S. in Industrial Engineering, Koc University, Istanbul, Turkey
- B.S. in Industrial Engineering, Istanbul Technical University, Turkey

Publications

- Gayon J-P, I. Talay-Degirmenci, F. Karaesmen, and E. L. Ormeci. 2009. Optimal Pricing and Production Policies of a Make-to-Stock System with Fluctuating Demand. *Probability in the Engineering and Informational Sciences*. **23**, 205-230.
- Gayon J-P, I. Talay-Degirmenci, F. Karaesmen, and E. L. Ormeci. 2006. Optimal Policies under Different Pricing Strategies in a Production System with Markov-Modulated Demand. *System Modeling and Optimization - Proceedings of 22nd IFIP TC 7 Conference*, eds. Ceragioli, F., Dontchev, A., Furuta, H., Marti, K., Pandolfi, L., 239-250, Springer.