Resource Pooling in the Presence of Failures: Efficiency versus Risk

Sigrún Andradóttir
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, U.S.A.

Hayriye Ayhan
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, U.S.A.
Phone: 404-894-2308 Fax: 404-894-2301

Douglas G. Down
Department of Computing and Software
McMaster University
Hamilton, Ontario L8S 4L7, Canada

May 2, 2016

Abstract

This paper studies the effects of resource pooling on system performance in the presence of failures. The goal is to understand whether pooling increases efficiency and/or reduces risk. We consider four queueing systems with different degrees of pooling (one has no pooling, one has only queues pooled, one has queues and failures pooled, and one has servers pooled), estimate efficiency via the mean number of customers in each system, and assess risk via the probability that there are many customers in each system. Our results show that when servers are subject to failures, pooling queues is always beneficial, whereas pooling both queues and servers improves efficiency but also increases risk. Thus there is a tradeoff between efficiency and risk in the presence of failures. These conclusions are different from reliable systems where pooling simultaneously improves efficiency and reduces risk and more pooling is better than less pooling (e.g., pooling queues and servers is better than pooling queues only). Thus, insights about resource pooling obtained from studying reliable systems should be used with caution in the presence of failures.

Keywords: unreliable servers, pooling, mean system size, tail asymptotics, stochastic ordering

1 Introduction

This paper considers systems with resources that are subject to failures (in that they experience downtimes during which they cannot serve customers). The objective is to understand the effects of pooling on system performance. More specifically, we consider the following four queueing systems, denoted as $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$, where $\mathcal{P}$ refers to pooling and the subscript indicates what is pooled:

$\mathcal{P}_\emptyset$: s single-server queues in parallel with servers subject to independent failures (no pooling);

$\mathcal{P}_Q$: a single queue with s servers subject to independent failures (the queues are pooled);

$\mathcal{P}_{Q,F}$: a single queue with s servers subject to synchronous failures (the queues and failures are pooled);

$\mathcal{P}_S$: a single-server queue (with the server subject to failures) where the service rate is s times the rate of the servers in the other systems (the servers, and hence the queues and failures, are pooled). This corresponds to the servers working together as a team with no loss of efficiency.

These four systems will be described more precisely in Section 2; the relationship between them is depicted in Figure 1.

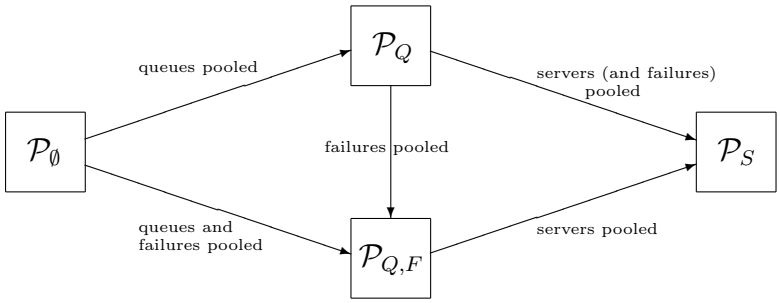

Figure 1: Relationship between systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$.

The performance measures we consider are the mean system size and tail probabilities of the number of customers in the system. These two performance measures capture efficiency (mean system size) and risk (tail probabilities of system size). We capture efficiency with mean system size since this reflects the average behavior of the systems and risk via tail probabilities of system size since this measures undesirable variation (large values, not small) and also provides the entire distribution of system size. Thus, higher efficiency means lower mean system size whereas higher risk means larger tail probabilities. Throughout the paper, we will compute the two performance measures for the four systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$, and compare them to investigate the effects of pooling on system performance. We are specifically interested in identifying the effects that changes made with the intent of increasing efficiency have on risk. In particular, resource pooling is aimed at increasing

operational efficiency, but may expose the servers to the same (or correlated) failures, and hence increase risk.

It is well known that pooling queues improves the performance of reliable systems. For example, an M/M/s queue with arrival rate $s\lambda$ and service rate μ for each server is superior (in the sense of standard stochastic ordering) to s parallel M/M/1 queues, each having an arrival rate of λ and service rate μ, in terms of the number of customers in the system. In fact, Smith and Whitt [19] establish the stronger monotone likelihood-ratio ordering of the number of customers in these two systems. Similarly, it is well known that an M/M/1 queue with arrival rate $s\lambda$ and service rate $s\mu$ is superior to an M/M/s queue with arrival rate $s\lambda$ and service rate μ for each server with respect to the mean number of customers in the system (see for example Wolff [28], page 258). Moreover, when $s = 2$, the number of customers in an M/M/1 queue with arrival rate $s\lambda$ and service rate $s\mu$ is stochastically smaller than the number of customers in the corresponding M/M/s queue (see Wolff [28], pages 257 and 258; a generalization of this result to s servers is provided in Proposition A.1 in the Appendix). We can then conclude that pooling queues and servers improves efficiency and risk (i.e., reduces both the mean and the tail probability of the number of customers) in reliable Markovian systems (however, Scheller-Wolf [18] shows that pooling can increase risk in certain systems with heavy-tailed service times and van Dijk and van der Sluis [21, 22] numerically argue that pooling may not be advantageous when there are two (or more) classes of customers, one class being short jobs, the other long jobs, and investigate how to improve the performance of pooling in this scenario). Note that using basic queueing theory, it is easy to determine that the tail probabilities of system size in these three systems all have a decay rate of $\frac{\lambda}{\mu}$. Interested readers are referred to Benjaafar [2], Calabrese [5], Larson [11], Rothkopf and Rech [17], and Smith and Whitt [19] for additional discussions of the effects of pooling on system performance in reliable systems. Our objective is to investigate if the same comparisons hold when servers are unreliable. To the best of our knowledge, this is the first paper that addresses the performance of pooling when the servers are subject to failures.

Argon and Andradóttir [1], Buzacott [4], Mandelbaum and Reiman [14], and Van Oyen, Gel, and Hopp [23] study pooling in queueing networks. In particular, Argon and Andradóttir [1] provide sufficient conditions for partial pooling of multiple adjacent queueing stations to be beneficial in tandem lines, allowing the service rate of a team of pooled servers to be additive, sub-additive, or super-additive. Buzacott [4] studies two models of pooling stations in a tandem line, namely parallel facilities (where each server completes all tasks in order) and teams (where the total processing time is the maximum duration of the subtasks completed by different team members). He shows that high task processing time variability makes the parallel system attractive compared to the tandem system, but pooling with teams is not superior to the (un-pooled) tandem line unless factors such as motivation improve the performance of team members. Mandelbaum and Reiman [14] show that for a tandem Jackson network, complete pooling always helps, but for Jackson networks with more general routing, complete pooling becomes advantageous only when the service variability is low. Van Oyen, Gel, and Hopp [23] study tandem lines with cross-trained servers. They show that if all servers are identical, then complete resource pooling maximizes throughput along all sample paths. Finally, Borst, Mandelbaum, and Reiman [3] and Wallace and Whitt

[25] discuss the effects of resource pooling in call centers.

Note that none of the studies reviewed in the previous paragraphs consider server failures. We first use results in the literature, probability generating functions, and large deviations techniques to provide closed-form expressions for the mean and tail asymptotics of the number of customers in the systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$. We then compare these systems using closed-form expressions, stochastic ordering techniques, and numerical experiments. We find that some standard ordering results comparing the mean and tail probabilities of the number of customers in pooled and un-pooled systems may be reversed when servers are unreliable. Thus, the benefits of pooling may no longer apply when the servers are subject to failures, and one needs to use caution pooling servers in these systems.

The rest of the paper is organized as follows. In Section 2, we provide a detailed description of the four systems we consider. The steady-state mean numbers of customers for these four systems are given in Section 3. Section 4 focuses on the tail probabilities of the number of customers in the four systems. In Section 5, we compare the systems analytically. Section 6 provides numerical experiments to compare the four systems for different numbers of servers s. In Section 7, we use numerical experiments to investigate if the conclusions of Section 6 hold for systems with reliable servers. Finally, Section 8 concludes the paper, and the Appendix provides proofs of some of the technical results presented in the paper and three new propositions with their proofs.

2 Models

We compare various degrees of pooling under different failure scenarios. For these purposes, we consider queueing systems with $s \geq 2$ servers and with the arrivals following a Poisson process with rate $s\lambda$. Each server has service times that are exponentially distributed with rate μ. The time between failures is exponentially distributed with rate f. Repair times are exponentially distributed with rate r. We assume that servers can fail anytime.

Specifically, we focus on four different systems within the class of systems described above. The first system, $\mathcal{P}_\emptyset$, is s M/M/1 queues in parallel with arrival rate λ to each queue (no pooling). The second system, $\mathcal{P}_Q$, is an M/M/s queue with independent server failures (the queues have been pooled). The third system, $\mathcal{P}_{Q,F}$, is an M/M/s queue with synchronous server failures (the queues and failures have been pooled). Finally, the last system, $\mathcal{P}_S$, is an M/M/1 queue where the service rate is $s\mu$ (the servers have been pooled into a team, implying the pooling of the queues and failures). In all four systems, the proportion of time R (for reliability) that each server is available is the same (i.e., $R = r/(r+f)$). We assume that a failure occurring while a customer is served results in preemption of the customer, and the displaced customer becomes the first in the queue. Since the service times are exponentially distributed, the remaining service time is exponentially distributed with rate μ, like the original service time (because if X is the original service time and Y is an exponential random variable with rate f that is independent of X, then both X and the remaining service time $X - Y$ given that $X > Y$ are exponential with rate μ). Our objective is to compare the mean number of customers and the tail probabilities of the number of customers in these four systems. Note that the stability condition for all four systems is $\lambda < \mu[r/(r+f)] = \mu R$.

The pooling of queues and servers is well known from the study of reliable queueing systems (see, e.g., the literature review in Section 1). The motivation for considering the pooling of failures is that pooling the queues may involve co-locating the servers, and hence subjecting them to the same failures (e.g., power outages).

3 Average Performance (Efficiency)

In this section, we provide expressions for the steady-state mean number of customers in the systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ in Sections 3.1, 3.2, 3.3, and 3.4, respectively.

3.1 Mean Number in System $\mathcal{P}_\emptyset$ – s M/M/1 Queues

Mitrani and Avi-Itzhak [15] derived the probability generating function of the number of customers in the system for an M/M/s queue with independent server failures. In particular, the mean number in the system for an M/M/1 queue with failures is given in equation (29) of [15] and the result below simply follows by multiplying that expression with s:

$$L_\emptyset = \frac{s\lambda((f+r)^2 + f\mu)}{(f+r)(r\mu - \lambda(r+f))}. \tag{1}$$

3.2 Mean Number in System $\mathcal{P}_Q$ – One M/M/s Queue, Independent Failures

As mentioned in Section 3.1, Mitrani and Avi-Itzhak [15] derived the probability generating function of the number of customers in the system for an M/M/s queue with independent server failures. More recently, Pang and Whitt [16] derive many server limit results in M/M/s queues with service interruptions (i.e., their results hold asymptotically as s increases). Even though it is difficult to compute (exactly) the performance measures that we are interested in for systems with general numbers of servers (from the probability generating function), one can obtain the mean number of customers for specific values of s. For example when $s = 2$, we have

$$\begin{aligned} L_Q &= \left(\frac{\mu}{2(f+r)(r\mu - \lambda(r+f))^2}\right)\big[\alpha(\mu(\mu-\lambda)r^2 + \lambda\mu f^2 + \lambda(f+r)^3) \\ &\quad +2\beta(\mu f(2\lambda(f+r) - \mu r) + \lambda(f+r)^3) + \gamma(\mu r(f+r)^2 + \lambda\mu f r)\big], \end{aligned}$$

where

$$\begin{aligned} &\alpha = 2f(r\mu - \lambda(r+f))(\mu + (\lambda+f)z_1)/\big[\mu(f+r)((f+r)(\mu+\lambda+fz_1) + \lambda(1-z_1)(\mu+\lambda))\big], \\ &\beta = (r\mu - \lambda(r+f))[\lambda\mu + \mu r + (fr - \lambda\mu)z_1]/\big[\mu(f+r)((f+r)(\mu+\lambda+fz_1) + \lambda(1-z_1)(\mu+\lambda))\big], \\ &\gamma = [2(\lambda+f)\beta - r\alpha]/\mu, \\ &z_1 = [(2\lambda+\mu+f+r) - \sqrt{(2\lambda+\mu+f+r)^2 - 8\lambda\mu}]/4\lambda. \end{aligned}$$

3.3 Mean Number in System $\mathcal{P}_{Q,F}$ – One M/M/s Queue, Synchronous Failures

There do not appear to be results in the literature for this system. There are results for systems in which the failures (vacations) occur when the system becomes idle (see Chao and Zhao [6], Levy and Yechiali [12], Tian, Li and Cao [20], Vinod [24], and Zhang and Tian [29]), but not for when the failures occur at any point in time. Consequently, our results in this section may be of some independent interest.

Let $p_{k,w}$ ($p_{k,f}$) be the steady-state probability of having k customers in the system and the servers working (failed). The balance equations are

$$\begin{aligned}
(s\lambda + s\mu + f)p_{k,w} &= s\lambda p_{k-1,w} + s\mu p_{k+1,w} + rp_{k,f}, \quad k \geq s, && (2)\\
(s\lambda + k\mu + f)p_{k,w} &= s\lambda p_{k-1,w} + (k+1)\mu p_{k+1,w} + rp_{k,f} \quad 1 \leq k \leq s-1, && (3)\\
(s\lambda + f)p_{0,w} &= \mu p_{1,w} + rp_{0,f}, && (4)\\
(s\lambda + r)p_{k,f} &= s\lambda p_{k-1,f} + fp_{k,w}, \quad k \geq 1, && (5)\\
(s\lambda + r)p_{0,f} &= fp_{0,w}. && (6)
\end{aligned}$$

Define $p_k = p_{k,w} + p_{k,f}$, the steady-state probability that there are k customers in the system. The probability generating function for the number in the system, $P(z)$, may be expressed as $P(z) = P_w(z) + P_f(z)$, where $P_w(z) = \sum_{k=0}^{\infty} z^k p_{k,w}$ and $P_f(z) = \sum_{k=0}^{\infty} z^k p_{k,f}$. Multiplying both sides of (5) with z^k, summing up over all $k \geq 1$, and adding this to (6), we derive

$$P_f(z) = \frac{f}{s\lambda + r - s\lambda z} P_w(z). \tag{7}$$

Multiplying both sides of (2) and (3) with z^k, summing up over all $k \geq 1$, and adding this to (4), we obtain

$$[s\lambda(1-z) + s\mu(1-z^{-1}) + f]P_w(z) = rP_f(z) + p_{0,w}s\mu(1-z^{-1}) + \mu(z-1)\sum_{k=1}^{s-1}(s-k)z^{k-1}p_{k,w}.$$

Combining the last two equations, we have

$$P_w(z) = \frac{(s\lambda + r - s\lambda z)\big[p_{0,w}s\mu(1-z^{-1}) + \mu(z-1)\sum_{k=1}^{s-1}(s-k)z^{k-1}p_{k,w}\big]}{s\lambda(1-z)\big[s\lambda(1-z) + s\mu(1-z^{-1}) + f\big] + rs\big[\lambda(1-z) + \mu(1-z^{-1})\big]}. \tag{8}$$

Note that using equations (3) to (6), one can express $p_{1,w}$, $p_{2,w}$, ..., $p_{s-1,w}$ in terms of $p_{0,w}$. Then using the fact that $P_w(1) = \frac{r}{r+f}$ and equation (8), one can obtain $p_{0,w}$, which together with (7) and (8) will yield $P(z)$.

The previous discussion provides a mechanism for computing $P(z)$, and, hence, $L_{Q,F}$ for specific s. As an example, consider the case with $s = 2$ servers. Then

$$p_{0,w} = \frac{(\mu r - \lambda(r+f))(2\lambda + r)}{\big[2\lambda^2 + \lambda(r+f) + 2\lambda\mu + \mu r\big](r+f)},$$

$$L_{Q,F} = \frac{2\lambda\mu(r^3 + 2f\lambda^2 + rf^2 + 2fr^2 + 2\lambda r^2 + rf\mu + 2f\lambda\mu + 4r\lambda f + 2\lambda f^2)}{\big[2\lambda^2 + \lambda(r+f) + 2\lambda\mu + \mu r\big](r\mu - \lambda(r+f))(r+f)}.$$

3.4 Mean Number in System $\mathcal{P}_S$ – One M/M/1 Queue

The result immediately follows from equation (29) of Mitrani and Avi-Itzhak [15] since we have an M/M/1 queue with server failures, arrival rate λs and service rate μs (see (1)):

$$L_S = \frac{\lambda((f+r)^2 + sf\mu)}{(f+r)(r\mu - \lambda(r+f))}.$$

4 Tail Behavior (Risk)

In this section, we provide exact tail asymptotics for the number of customers in systems $\mathcal{P}_\emptyset$, $\mathcal{P}_S$, and $\mathcal{P}_{Q,F}$ in Sections 4.1, 4.2, and 4.3, respectively. Since obtaining this expression is more challenging for system $\mathcal{P}_Q$ (because the large deviation paths appear to become complex), we will consider the tail probabilities of this system numerically in Section 6. Note that even though the probability generating function for the number in the system is available either in closed form (for systems $\mathcal{P}_\emptyset$ and $\mathcal{P}_S$) or via an algorithm (for systems $\mathcal{P}_Q$ and $\mathcal{P}_{Q,F}$), it is challenging to invert these functions to obtain the exact probabilities.

4.1 Tail Behavior of System $\mathcal{P}_\emptyset$ – s M/M/1 Queues

Examining one queue in isolation, we have that $X_I(t) = (Q_I(t), N_I(t))$ is a Continuous-Time Markov Chain (CTMC), where $Q_I(t)$ is the number of customers in the system and $N_I(t)$ is the status of the server at time t. This system has been studied in Lorek [13], where the techniques in Foley and McDonald [7, 8] are applied to compute the exact tail asymptotics of the steady-state distribution for $X_I(t)$, as well as the corresponding large deviations paths. We are only concerned with the former here. Let $f \sim g$ denote the relationship $\lim_{k\to\infty} f(k)/g(k) = 1$. For $\mathcal{P}_\emptyset$, the system state $X(t) = (Q_1(t), N_1(t), Q_2(t), N_2(t), \ldots, Q_s(t), N_s(t))$ is a CTMC, where $Q_i(t)$ and $N_i(t)$ are independent copies of $Q_I(t)$ and $N_I(t)$, for $i = 1, \ldots, s$ (as each queue operates independently from the others). The proof of the following result is given in the Appendix.

Proposition 4.1 *There exist finite constants c_a, c_b such that*

$$c_a \binom{\ell+s-1}{\ell} \gamma_\emptyset^\ell \le P\{Q_1(t) + Q_2(t) + \cdots + Q_s(t) = \ell\} \le c_b \binom{\ell+s-1}{\ell} \gamma_\emptyset^\ell \tag{9}$$

as $\ell \to \infty$, where

$$\gamma_\emptyset = \frac{2\lambda}{\lambda + r + \mu + f - \sqrt{\alpha_\emptyset}}, \quad \alpha_\emptyset = (\mu - \lambda - f - r)^2 + 4f\mu.$$

4.2 Tail Behavior of System $\mathcal{P}_S$ – One M/M/1 Queue

This system is a direct application of Proposition 2.1 of Lorek [13], where the arrival and service rates are $s\lambda$ and $s\mu$, respectively. So, for finite constants $c_{S,w}$ and $c_{S,f}$,

$$p_{k,w} \sim c_{S,w}\gamma_S^k, \quad p_{k,f} \sim c_{S,f}\gamma_S^k, \tag{10}$$

where

$$\gamma_S = \frac{2s\lambda}{s\lambda + r + s\mu + f - \sqrt{\alpha_S}}, \quad \alpha_S = (s\mu - s\lambda - f - r)^2 + 4sf\mu.$$

Note that the difference between $\gamma_\emptyset$ and γ_S arises from the fact that pooling impacts the arrival and service rates, but not the failure and repair rates (if pooling resulted in failure and repair rates sf and sr, then $\gamma_\emptyset$ and γ_S would be equal).

4.3 Tail Behavior of System $\mathcal{P}_{Q,F}$ – One M/M/s Queue, Synchronous Failures

Let $Q(t)$ be the number of customers in the system at time t. Then $\{Q(t)\}$ is a CTMC and the transition rates for $\mathcal{P}_{Q,F}$ and $\mathcal{P}_S$ agree for all but a finite number of states (those with $Q(t) = 0, 1, \ldots, s-1$). Thus, the decay rates of the tails of the steady-state distribution are identical. In other words, for finite constants $c_{Q,F,w}$ and $c_{Q,F,f}$,

$$p_{k,w} \sim c_{Q,F,w}\gamma_S^k, \quad p_{k,f} \sim c_{Q,F,f}\gamma_S^k. \tag{11}$$

While the decay rate is the same as for system $\mathcal{P}_S$, the constants will in general be different.

5 Comparison

In this section, we compare the means and tail probabilities of the number of customers in the four systems described in Section 2. In what follows, we first summarize the comparison results and then show how these results were obtained by comparing two systems at a time.

5.1 Comparison Summary

We start by addressing the efficiency of the systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$. In particular, we have shown that:

$$\text{If } s = 2, \text{ then } \begin{cases} L_Q, L_S < L_{Q,F} < L_\emptyset, \\ \text{both } L_Q < L_S \text{ and } L_S < L_Q \text{ are possible;} \end{cases} \tag{12}$$

$$\text{If } s > 2, \text{ then } \begin{cases} L_Q, L_S \leq L_\emptyset \text{ (with } L_S < L_\emptyset), \\ L_S \leq L_{Q,F}. \end{cases} \tag{13}$$

In addition, we have:

$$\text{As } \rho \to 1, \begin{cases} L_Q \ll L_S < L_{Q,F} \ll L_\emptyset & \text{if } s = 2, \\ L_S \ll L_\emptyset & \text{if } s > 2, \end{cases} \tag{14}$$

where $\ll$ indicates that the difference between the limits is infinite and $<$ means that the difference is finite. (Note that throughout the paper, $\rho = \frac{\lambda(r+f)}{\mu r} \to 1$ is equivalent to $\lambda \to \frac{\mu r}{r+f}$. Thus, we increase λ as μ, r, and f remain the same). Finally, when it comes to risk, we have shown that:

$$\begin{cases} \gamma_\emptyset < \gamma_S = \gamma_{Q,F}, \\ \mathrm{P}\{N_Q(t) \geq k\} \leq \mathrm{P}\{N_\emptyset(t) \geq k\}, \forall t \geq 0, k \geq 0, \\ \mathrm{P}\{N_S(t) \geq k\} \leq \mathrm{P}\{N_{Q,F}(t) \geq k\}, \forall t \geq 0, k \geq 0. \end{cases} \tag{15}$$

We next prove the comparison results summarized above. In each case, we start with the comparison of the mean number of customers (but in certain cases we will have stronger stochastic ordering results), and then continue with the comparison of the tail asymptotics (except for the cases involving system $\mathcal{P}_Q$). Numerical results are provided in Section 6 to study the tail asymptotics of system $\mathcal{P}_Q$. More specifically, in Sections 5.2, 5.3, and 5.4, we compare the un-pooled system $\mathcal{P}_\emptyset$ with the pooled systems $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$, respectively; in Section 5.5, we compare the two partially pooled systems $\mathcal{P}_Q$ and $\mathcal{P}_{Q,F}$; and in Sections 5.6 and 5.7, we compare the two partially pooled systems $\mathcal{P}_Q$ and $\mathcal{P}_{Q,F}$ with the fully pooled system $\mathcal{P}_S$, respectively.

5.2 Comparison of Systems $\mathcal{P}_\emptyset$ and $\mathcal{P}_Q$

We show that $\mathcal{P}_Q$ (the system with pooled queues) has a better mean system size performance than $\mathcal{P}_\emptyset$ (the un-pooled system). In fact, Proposition 5.1 below proves the stronger result that the number of customers in $\mathcal{P}_\emptyset$ at any time t is stochastically larger than the number of customers in $\mathcal{P}_Q$ at time t. This is due to the fact that $\mathcal{P}_\emptyset$ can have idle servers even when there are customers waiting in line. Moreover, the arrival process can be a general renewal process (i.e., not necessarily Poisson). The proof of Proposition 5.1 is given in the Appendix.

Proposition 5.1 *Consider systems $\mathcal{P}_\emptyset$ and $\mathcal{P}_Q$ with s servers and general renewal arrival processes. Let $N_\emptyset(t)$ and $N_Q(t)$ be the number of customers at time t in these more general versions of systems $\mathcal{P}_\emptyset$ and $\mathcal{P}_Q$, respectively. Then*

$$N_Q(t) \overset{st}{\leq} N_\emptyset(t) \text{ for all } t \geq 0,$$

where st denotes the standard stochastic ordering.

Proposition 5.1 immediately implies that $L_\emptyset \geq L_Q$. For Markovian systems with $s = 2$, we can use the expressions in Sections 3.1 and 3.2 to compute $L_\emptyset - L_Q = \Upsilon/\Delta$, where

$$\begin{aligned}
\Upsilon &= \lambda\big[(1-z_1)\big(2\lambda^2(r+f)^3 + \lambda(2\lambda\mu f^2 + f^4 + 3f^3 r + 3f^2 r^2 + f r^3 + \mu r f^2 + \lambda\mu r f + 2f^3\mu) + \\
&\quad \mu^2(r^2 f + \mu r f)\big) + \lambda\big(7f r^3 + 9f^2 r^2 + 5f^3 r + \mu f^3 + 4\mu r f^2 + 2r^4 + f^4\big) + \\
&\quad \mu(1+z_1)\big(3r f^3 + 3f^2 r^2 + f^4 + f r^3\big) + \mu^2 f^2\big(f + 2r\big) + \\
&\quad f z_1\big(4f^3 r + 6f^2 r^2 + 4f r^3 + f^4 + r^4\big)\big], \\
\Delta &= (f+r)^2(\mu r - \lambda r - \lambda f)[(1-z_1)\lambda(\mu+\lambda) + (r+f)(z_1 f + \lambda + \mu)],
\end{aligned}$$

with z_1 defined in Section 3.2. Note that $\Upsilon > 0$ since $0 < z_1 < 1$ (see Mitrani and Avi-Itzhak [15]) and $\Delta > 0$ since $0 < z_1 < 1$ and the stability condition holds. Furthermore, the above expression verifies that $L_\emptyset - L_Q \to \infty$ as $\rho \to 1$ (note that Δ tends to 0 as ρ tends to 1). This is intuitive because idling the servers (when there are customers waiting) will have higher impact on the mean number of customers in the system as the system load increases and the difference will tend to ∞ (as $\rho \to 1$) as in the case with reliable systems (see [22]).

Proposition 5.1 also implies that $\mathrm{P}\{N_Q(t) \geq k\} \leq \mathrm{P}\{N_\emptyset(t) \geq k\}$ for all $t \geq 0$ and $k \geq 0$, and hence system $\mathcal{P}_Q$ performs better than system $\mathcal{P}_\emptyset$ in terms of both efficiency and risk.

5.3 Comparison of Systems $\mathcal{P}_\emptyset$ and $\mathcal{P}_{Q,F}$

We compute $L_\emptyset - L_{Q,F}$ when $s = 2$ (as we do not have a closed-form expression for $L_{Q,F}$ for general s, see Section 3.3). In this case, we have

$$L_\emptyset - L_{Q,F} = \frac{2\lambda^2\left[r^2 + 2\lambda r + 2rf + 2\lambda f + f\mu + f^2\right]}{[2\lambda^2 + \lambda r + \lambda f + 2\lambda\mu + \mu r][\mu r - \lambda(r+f)]} > 0,$$

which implies that $\mathcal{P}_{Q,F}$, which has correlated server failures, is more efficient than $\mathcal{P}_\emptyset$ where the servers fail independently and have dedicated queues. This again is due to the fact that $\mathcal{P}_\emptyset$ can have idle servers even when there are customers waiting in line. Moreover, as $\rho \to 1$, $L_\emptyset - L_{Q,F}$ approaches ∞ since idling the servers when there are customers waiting is more wasteful as the system load increases.

As far as the comparison of tail asymptotics for these two systems is concerned, we have that $\gamma_\emptyset < \gamma_{Q,F} = \gamma_S$ if (after some algebra):

$$s\sqrt{(\mu - \lambda - f - r)^2 + 4f\mu} < (s-1)(f+r) + \sqrt{(s\mu - s\lambda - f - r)^2 + 4sf\mu}.$$

If we square both sides and rearrange, we see that the above inequality is true if and only if

$$s\lambda f - s\mu r + s\lambda r + sf\mu + 2fr + f^2 + r^2 < (f+r)\sqrt{(s\mu - s\lambda - f - r)^2 + 4sf\mu}.$$

Again squaring both sides and rearranging, we have that $\gamma_\emptyset < \gamma_{Q,F}$ holds if

$$fs^2\mu(\mu r - \lambda r - f\lambda) > 0,$$

which holds under the stability condition. Thus, the pooling of queues and server failures helps to reduce the mean number of customers in the system, but actually has the *opposite effect* on the tail probabilities. Hence, pooling queues and server failures improves efficiency but increases risk in unreliable systems (because the servers are now exposed to the same failures). This suggests that while pooling queues and server failures, one needs to take into the account the tradeoff between risk and efficiency.

Remark 5.1 *Consider a system $\mathcal{P}_F$ that is composed of s M/M/1 queues with synchronized server failures (i.e., only the failures are pooled). Let L_F be the mean number of customers in $\mathcal{P}_F$. Then it is easy to see that $L_F = L_\emptyset$, which is given in equation (1). Moreover, using an argument similar to the one in the proof of Proposition 5.1, one can show that $N_{Q,F}(t) \overset{st}{\leq} N_F(t)$ for all $t \geq 0$, where $N_F(t)$ is the total number of customers in system $\mathcal{P}_F$ at time t, and hence, $\mathrm{P}\{N_F(t) \geq k\} \geq \mathrm{P}\{N_{Q,F}(t) \geq k\}$, for all $k \geq 0$ and $t \geq 0$. Hence, $\mathcal{P}_F$ has the worst performance among the five systems both in terms of the mean and the tail probability of the number of customers in the system.*

5.4 Comparison of Systems $\mathcal{P}_\emptyset$ and $\mathcal{P}_S$

It is easy to see that

$$L_\emptyset - L_S \;=\; \frac{\lambda(s-1)(f+r)}{(\mu r - \lambda r - \lambda f)} > 0.$$

As expected, system $\mathcal{P}_S$, which has a single server whose service rate is s times faster, has lower mean number in the system than system $\mathcal{P}_\emptyset$, where the servers fail independently and have dedicated queues. Moreover, as $\rho \to 1$, $L_\emptyset - L_S$ approaches ∞. This is intuitive since having dedicated servers (which implies servers may idle when there are customers waiting in line) results in higher number of customers in the system as the load increases.

As far as the comparison of the tail asymptotics is concerned, since the decay rate is the same in systems $\mathcal{P}_{Q,F}$ and $\mathcal{P}_S$, the comparison in the previous subsection holds and we have $\gamma_\emptyset < \gamma_S$. Hence, as in Section 5.3, pooling reduces the mean number of customers in the system but not the tail probabilities, and there is again a tradeoff between risk and efficiency.

5.5 Comparison of Systems $\mathcal{P}_Q$ and $\mathcal{P}_{Q,F}$

We prove that for systems with $s = 2$, $\mathcal{P}_Q$ has better performance than system $\mathcal{P}_{Q,F}$ in terms of the mean number of customers. Using the expressions in Sections 3.2 and 3.3, we have

$$L_{Q,F} - L_Q = \frac{\Theta}{\Delta(2\lambda^2 + 2\lambda\mu + \lambda r + \lambda f + \mu r)},$$

where

$$\begin{aligned}
\Theta = {} & \lambda f\big[(\mu r - \lambda(r+f))\big(9z_1\lambda f^2 r + 6z_1 f^2 r^2 + 7z_1\lambda r^2 f + 3\mu r f^2 + 4z_1 f^3 r + 2z_1\lambda^2 r^2 + \\
& 4\lambda^2 r f + 2z_1\lambda r^3 + 3z_1\lambda f^3 + 2z_1\lambda^2 f^2 + \mu f^3 + \mu r^3 + z_1 f^4 + z_1 r^4 + 3\mu r^2 f + 4z_1 r^3 f\big) + \\
& (1-z_1)\big(\mu^4 r^2 + \mu^3 r^3 + 2\mu^4\lambda r + \lambda\mu^3 f r + 4\lambda^4\mu f + 2\lambda^4\mu r + 2\mu^3 r\lambda^2 + 3\mu^3 r^2\lambda + \\
& 2\lambda^3\mu^2 r + 4\lambda^3\mu^2 f + 3\lambda^3\mu f^2 + \mu^2 r^2\lambda^2 + 2\mu^2\lambda^2 f^2\big) + \\
& (\mu-\lambda)(\lambda+\mu)(z_1 r^4 + 2f r^3) + 2r^2 f z_1(\mu^2 r - \lambda^2(f+r)) + \lambda^2\mu f(\mu r - z_1\lambda(r+f)) + \\
& \mu\big(2\mu^2 f r^2 + \lambda\mu f^3 + \lambda^2 f^3 + \mu^2 f^2 r + 3\lambda^3 f^2 + 4\lambda^2\mu f^2 + 2\mu^2 f^2\lambda + 3\lambda^3 r^2 + 8f\mu r^2\lambda + \\
& 11\lambda^2\mu f r + 2\lambda^2 r f^2 + \lambda^2 f r^2 + 6\mu r\lambda f^2 + 9\lambda^3 r f\big) + \lambda r^3\big(3\mu^2 + z_1\mu^2\big) + \\
& \mu z_1\big(3\lambda^3 r^2 + 2f^3\mu\lambda + f^3\mu r + r^2\lambda^3 + 4f\lambda\mu r^2 + \lambda^2 r^2 f + 2r f^3\mu + \lambda^2 r f^2 + 4\lambda\mu f^2 r\big)\big]
\end{aligned}$$

and z_1 and Δ are defined in Sections 3.2 and 5.2, respectively. From the stability condition and the facts that $0 < z_1 < 1$ and $\Delta > 0$, we conclude that $L_{Q,F} - L_Q > 0$. This is due to incurring longer waiting lines since all the servers fail at the same time in the $\mathcal{P}_{Q,F}$ system. Thus, pooling the failures does not improve efficiency. Moreover, using the closed form expression of $L_{Q,F} - L_Q$ above, we can again see that $L_{Q,F} - L_Q$ approaches ∞ as $\rho \to 1$ (recall that $\Delta \to 0$ as $\rho \to 1$, see Section 5.2). This is intuitive because as the load increases, customers accumulate faster in the $\mathcal{P}_{Q,F}$ system since all servers fail at the same time. In fact, we have shown that with fixed R, λ, and μ values, as r and f tend to zero (and hence, failures become infrequent and long), the ratios $\frac{L_\emptyset}{L_{Q,F}}$, $\frac{L_\emptyset}{L_S}$, $\frac{L_{Q,F}}{L_S}$ converge to one, while the ratios $\frac{L_\emptyset}{L_Q}$, $\frac{L_{Q,F}}{L_Q}$, and $\frac{L_S}{L_Q}$ converge to limits that are equal to each other and greater than one (see Proposition A.4).

5.6 Comparison of Systems $\mathcal{P}_Q$ and $\mathcal{P}_S$

For the comparison of systems $\mathcal{P}_Q$ and $\mathcal{P}_S$, we provide a numerical example that demonstrates that one system does not necessarily dominate the other. In other words, the dominant mean

number of customers depends on the system parameters. Figure 2 depicts $L_Q - L_S$ as a function of the arrival rate λ when $s = 2$, $r = 5$, $f = 1$, and $\mu = 4$.

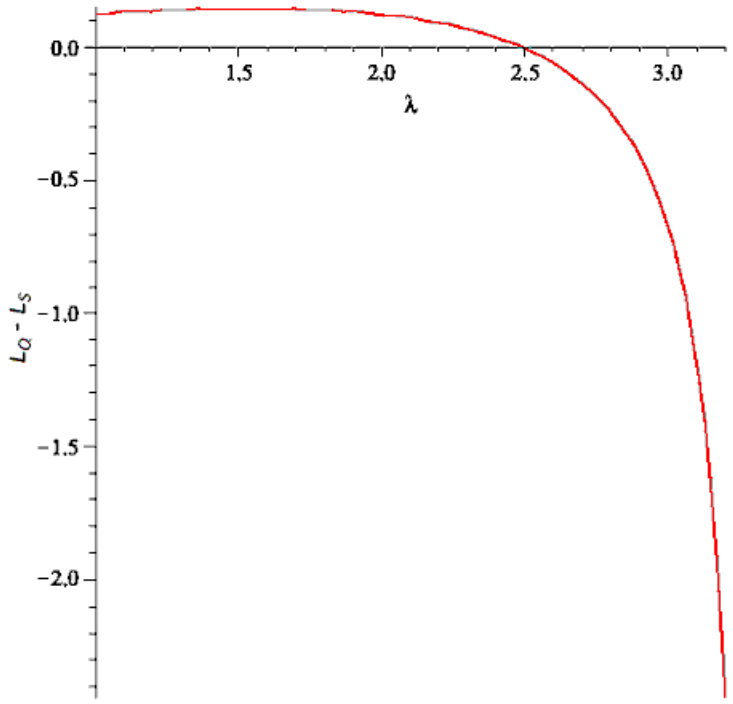

Figure 2: The difference of the mean number of customers in systems $\mathcal{P}_Q$ and $\mathcal{P}_S$ as a function of λ.

When $s = 2$, the expressions in Sections 3.2 and 3.4 yield $L_Q - L_S = \Gamma/\Delta$, where

$$\begin{aligned}\Gamma \;=\; & \lambda\big[(z_1-1)\big(\lambda^2(r+f)^3 + f(2\lambda^2\mu f + f^3\mu + \lambda r^3 + \mu^3 r + \lambda^2\mu r + \lambda^3\mu + 3\mu f r^2 + \\ & 3f\lambda r^2 - r f^2\mu - \mu r^3 - \mu\lambda r^2 + \mu^2 r^2 + 3f^2\lambda r) + \mu\lambda r^3\big) + \\ & (1+z_1)\mu\lambda r f^2 - \lambda\big(3f r^3 + 3f^2 r^2 + f^3 r + \mu f^3 + r^4\big) + \\ & \mu\big(r^4 - \mu f^3 - z_1 f^4 - 2z_1 r f^3 + 2f r^3 - 2\mu r f^2 - 2z_1 f r^2\lambda\big)\big]\end{aligned}$$

and Δ and z_1 are as given in Sections 3.2 and 5.2, respectively. Furthermore, with some algebra one can verify that as $\lambda \to \frac{\mu r}{r+f}$, Γ reduces to

$$\frac{-\mu^2 f[(1-z_1)(rf^3 + f^2r^2 + 4f\mu r^2 + f^2\mu r + 2r^3\mu) + 2r^4]}{(f+r)^2} + \frac{-\mu f^2(\mu f^3 + z_1 f^4 + 5\mu r f^2 + 4z_1 f r^3 + 6z_1 f^2 r^2 + 8\mu r^3 + 4z_1 f^3 r + 10\mu f r^2 + z_1 r^4)}{(f+r)^2}.$$

Thus, $L_Q - L_S$ tends to $-\infty$ as ρ goes to 1, and pooling the servers (and hence the failures) hurts efficiency when ρ is close to one. This is intuitive because customers accumulate quickly in $\mathcal{P}_S$ when the server fails (resulting in a service rate of zero).

Remark 5.2 *Consider a system $\mathcal{P}_{Q,S}$ with pooled queues and servers (but not failures). System $\mathcal{P}_{Q,S}$ performs as an M/M/1 queue, but the head of the line customer is served by the aggregate capacity of all the available servers. Using an argument similar to the proof of Proposition 5.1, one can show that $N_{Q,S}(t) \overset{st}{\leq} N_Q(t)$ for all $t \geq 0$, where $N_{Q,S}(t)$ is the total number of customers in system $\mathcal{P}_{Q,S}$ at time t .*

5.7 Comparison of Systems $\mathcal{P}_{Q,F}$ and $\mathcal{P}_S$

The next proposition shows that the number of customers at any point in time is stochastically less for the system with pooled servers ($\mathcal{P}_S$) than for the system with pooled queues and server failures ($\mathcal{P}_{Q,F}$). Moreover, the result holds for general renewal arrival processes. The proof of Proposition 5.2 is intuitive and is given in the Appendix.

Proposition 5.2 *Consider systems $\mathcal{P}_{Q,F}$ and $\mathcal{P}_S$ with s servers and general renewal arrival processes. Let $N_{Q,F}(t)$ and $N_S(t)$ be the number of customers in these more general versions of systems $\mathcal{P}_{Q,F}$ and $\mathcal{P}_S$, respectively. Then*

$$N_S(t) \overset{st}{\leq} N_{Q,F}(t) \text{ for all } t \geq 0.$$

The above proposition immediately implies that $L_{Q,F} \geq L_S$ which is intuitive since the service rate is always $s\mu$ in the $\mathcal{P}_S$ system (as long as the server is up). Hence, pooling the servers improves efficiency when the failures are already pooled. Moreover, for Markovian systems with $s = 2$ servers, we can quantify the difference in the mean number of customers in these two systems as

$$L_{Q,F} - L_S = \frac{\lambda(r + f + 2\lambda)}{(2\lambda^2 + \lambda r + \lambda f + 2\lambda\mu + \mu r)} > 0. \tag{16}$$

Note that $L_{Q,F} - L_S < 1$. Thus, unlike the differences of the mean number of customers in Sections 5.2, 5.3, 5.4, 5.5, and 5.6, this difference does not tend to infinity as $\rho \to 1$. This is intuitive because in both systems a failure stops the entire service process, and hence, increasing the load has similar effects on both systems.

The tail asymptotics of system size in these two systems have the same decay rate.

6 Numerical Results

In this section, we use numerical experiments to better understand how the systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ compare and to study how our measures of efficiency and risk (mean number of customers and tail asymptotics) depend on s and other system parameters. Our main conclusion from the numerical results is that $\mathcal{P}_Q$ appears to perform the best overall for the systems we considered.

Specifically, we consider systems with $s \in \{2, 3, 5\}$ and focus on six sets of rates as defined below:

- Set 1: $r = 1$, $f = 1$, $\mu = 100$, and $\lambda = 25$;

- Set 2: $r = 10$, $f = 1$, $\mu = 100$, and $\lambda = \frac{500}{11}$;
- Set 3: $r = 10$, $f = 0.1$, $\mu = 100$, and $\lambda = \frac{5000}{101}$;
- Set 4: $r = 1$, $f = 1$, $\mu = 100$, and $\lambda = 45$;
- Set 5: $r = 10$, $f = 1$, $\mu = 100$, and $\lambda = \frac{900}{11}$;
- Set 6: $r = 10$, $f = 0.1$, $\mu = 100$, and $\lambda = \frac{9000}{101}$.

Recall that the combined arrival rate to each system is $s\lambda$, and, thus, the parameters in sets 1, 2, and 3 yield a traffic intensity of $\rho = 0.5$ and the rates in sets 4, 5, and 6 result in systems with traffic intensity of $\rho = 0.9$. On the other hand, the proportion of time that systems with parameter sets 1 and 4 are reliable is $R = 1/2$, as compared to $R = 10/11$ for systems with parameter sets 2 and 5 and $R = 100/101$ for systems with parameter sets 3 and 6. Hence, these six parameter sets consider all combinations of medium and high traffic intensity $\rho \in \{0.5, 0.9\}$ and low, medium, and high server reliability $R \in \{1/2, 10/11, 100/101\}$.

Let N denote the steady-state number of customers in the system. Hence, for all four systems, $\mathrm{E}[N] = L$, and $\mathrm{P}\{N = k\}$ for all $k = 0, 1, \ldots$ denotes the probability mass function of N. Tables 1 through 3 illustrate the steady-state mean number of customers in the system and the tail probabilities $\mathrm{P}\{N > k\}$ for systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ and $k \in \{15, 50, 100, 150, 200\}$ when the rates are chosen from the sets of parameters listed above and $s = 2$, $s = 3$, and $s = 5$, respectively. *The mean number in the system and the tail probabilities were computed exactly* using the probability generating function of N for each system (and were not estimated via simulation). The missing entries in the tables correspond to the cases where a straightforward implementation in Maple did not lead to results (as Maple was unable to compute the derivatives of the probability generating function of N). Furthermore, for parameter sets 3 and 6, we computed the mean number in the system and the tail probabilities for equivalent systems where all rates are multiplied by 10, since Maple was again unable to compute these characteristics for systems with the original rates. For each parameter set in each table, we use bold font to indicate which system yields the best (smallest) value of each performance measure when data for all performance measures are available and the best is unique (up to the indicated precision); the corresponding system(s) are also indicated in bold.

We start by considering the mean number in the system. Tables 1 to 3 are consistent with equations (12) to (14), and *suggest that the results for $s = 2$ also hold for $s > 2$.* More specifically, we know from Section 5 that $L_Q, L_S \leq L_\emptyset$, that $L_S \leq L_{Q,F}$, and that $L_\emptyset - L_S \to \infty$ as $\rho \to 1$. Moreover, for systems with $s = 2$, $L_Q, L_S < L_{Q,F} < L_\emptyset$ and the differences $L_\emptyset - L_{Q,F}$, $L_\emptyset - L_Q$, $L_{Q,F} - L_Q$, and $L_S - L_Q$ tend to ∞ as $\rho \to 1$. The numerical results in Tables 1 to 3 agree with these results, and suggest that $L_Q < L_{Q,F} < L_\emptyset$ and that the differences $L_\emptyset - L_{Q,F}$, $L_\emptyset - L_Q$, $L_{Q,F} - L_Q$, and $L_S - L_Q$ become large as $\rho \to 1$ even when $s > 2$. On the other hand, we know that $L_Q - L_S$ can be negative or positive when $s = 2$ depending on the arrival, service, failure, and repair rates. In our numerical experiments for systems with $s = 2$ and $s = 3$ (we do not have results for system $\mathcal{P}_Q$ when $s = 5$), we have $L_S < L_Q$ for parameter set 3 and L_S is slightly larger than L_Q for parameter set 6. However, L_S is much larger than L_Q for parameter sets 1, 2, 4, and 5. As given in equation (16), when

	System	L	P$\{N > 15\}$	P$\{N > 50\}$	P$\{N > 100\}$	P$\{N > 150\}$	P$\{N > 200\}$
Parameter	$\mathcal{P}_\emptyset$	52.00	0.719531	0.401017	0.153173	0.053871	0.018032
Set 1:	$\boldsymbol{\mathcal{P}_Q}$	**18.12**	**0.306645**	**0.124039**	**0.034043**	**0.009343**	**0.002565**
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	51.20	0.546591	0.345224	0.179066	0.092881	0.048177
$R = 1/2$	$\mathcal{P}_S$	51.00	0.545128	0.344300	0.178587	0.092632	0.048048
Parameter	$\mathcal{P}_\emptyset$	3.65	0.036719	0.000112	0.000000	0.000000	0.000000
Set 2:	$\boldsymbol{\mathcal{P}_Q}$	**1.67**	**0.004758**	**0.000008**			
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	2.97	0.045000	0.001652	0.000015	0.000000	0.000000
$R = 10/11$	$\mathcal{P}_S$	2.65	0.043640	0.001602	0.000014	0.000000	0.000000
Parameter	$\mathcal{P}_\emptyset$	2.20	0.004276	0.000008	0.000000	0.000000	0.000000
Set 3:	$\boldsymbol{\mathcal{P}_Q}$	1.37	**0.000321**	**0.000005**	0.000000	0.000000	0.000000
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	1.53	0.005448	0.000195	0.000002	0.000000	0.000000
$R = 100/101$	$\boldsymbol{\mathcal{P}_S}$	**1.20**	0.005270	0.000188	0.000002	0.000000	0.000000
Parameter	$\mathcal{P}_\emptyset$	468.00	0.982109	0.950218	0.891658	0.823994	0.752295
Set 4:	$\boldsymbol{\mathcal{P}_Q}$	**216.94**	**0.798465**	**0.693832**	**0.570820**	**0.469628**	**0.390398**
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	459.31	0.884064	0.824719	0.746784	0.676214	0.612313
$R = 1/2$	$\mathcal{P}_S$	459.00	0.883515	0.824207	0.746320	0.676214	0.611933
Parameter	$\mathcal{P}_\emptyset$	32.88	0.676240	0.211824	0.031038	0.003959	0.000470
Set 5:	$\boldsymbol{\mathcal{P}_Q}$	**16.20**	**0.349951**				
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	24.33	0.425437	0.154734	0.037708	0.009190	0.002239
$R = 10/11$	$\mathcal{P}_S$	23.88	0.419105	0.152759	0.037227	0.009072	0.002211
Parameter	$\mathcal{P}_\emptyset$	19.76	0.510514	0.049838	0.001424	0.000037	0.000001
Set 6:	$\boldsymbol{\mathcal{P}_Q}$	**10.27**	**0.216603**				
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	11.24	0.227820	0.023596	0.002102	0.000200	0.000019
$R = 100/101$	$\mathcal{P}_S$	10.76	0.218700	0.023107	0.002056	0.000195	0.000019

Table 1: Mean and tail probabilities for number in the system when $s = 2$.

$s = 2$, $L_{Q,F} - L_S < 1$, and the results in Table 2 suggest that this difference is also small when $s = 3$, but we can conclude from Table 3 that $L_{Q,F} - L_S$ *increases as* s *increases*. Finally, we would like to point out that the results of Tables 1 and 2 suggest that *unless the servers are highly reliable, system* $\mathcal{P}_Q$ *in general has significantly smaller mean number of customers than the other three systems*.

We next focus on the tail asymptotics of the number of customers in the system. The results of Section 5 show that the decay rates of the tail probabilities for systems $\mathcal{P}_\emptyset$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ are ordered as $\gamma_\emptyset < \gamma_{Q,F} = \gamma_S$, and that system $\mathcal{P}_Q$ outperforms system $\mathcal{P}_\emptyset$ in terms of tail probabilities (see equation (15)). Tables 1 through 3 illustrate that one does not need to go far in the tail to observe $\gamma_{Q,F} \simeq \gamma_S$. However, the results for parameter set 4 indicate that the ordering between $\gamma_\emptyset$ and $\gamma_{Q,F} = \gamma_S$ may become apparent only for large values of k in P$\{N > k\}$ in certain cases. (This is consistent with the fact that the tail asymptotics for $\mathcal{P}_\emptyset$ include a polynomial term, unlike the tail asymptotics for $\mathcal{P}_{Q,F}$ and $\mathcal{P}_S$, see equations (9), (10), and (11).) For systems with two servers and the rates from parameter set 4, we also computed P$\{N > k\}$ for $k = 650$ and $k = 700$. We had

	System	L	$\mathrm{P}\{N > 15\}$	$\mathrm{P}\{N > 50\}$	$\mathrm{P}\{N > 100\}$	$\mathrm{P}\{N > 150\}$	$\mathrm{P}\{N > 200\}$
Parameter	$\mathcal{P}_\emptyset$	78.00	0.865102	0.596300	0.291977	0.125597	0.049760
Set 1:	$\boldsymbol{\mathcal{P}_Q}$	**11.96**	**0.214479**	**0.079435**	**0.022508**		
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	76.44	0.584804	0.429818	0.276852	0.178325	0.114862
$R = 1/2$	$\mathcal{P}_S$	76.00	0.580000	0.428160	0.275784	0.177637	0.114419
Parameter	$\mathcal{P}_\emptyset$	5.48	0.069653	0.000300	0.000000	0.000000	0.000000
Set 2:	$\boldsymbol{\mathcal{P}_Q}$	**1.87**	**0.002316**				
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	4.17	0.069810	0.007348	0.000295	0.000012	0.000000
$R = 10/11$	$\mathcal{P}_S$	3.48	0.066694	0.007021	0.000282	0.000011	0.000000
Parameter	$\mathcal{P}_\emptyset$	3.29	0.008323	0.000016	0.000000	0.000000	0.000000
Set 3:	$\boldsymbol{\mathcal{P}_Q}$	1.75	**0.000048**				
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	2.03	0.008420	0.000880	0.000035	0.000001	0.000000
$R = 100/101$	$\boldsymbol{\mathcal{P}_S}$	**1.29**	0.008013	0.000838	0.000033	0.000001	0.000000
Parameter	$\mathcal{P}_\emptyset$	702.00	0.997871	0.991300	0.974224	0.948156	0.913965
Set 4:	$\boldsymbol{\mathcal{P}_Q}$	**207.86**	**0.780616**				
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	684.69	0.892684	0.852045	0.797178	0.745845	0.697816
$R = 1/2$	$\mathcal{P}_S$	684.00	0.890000	0.851264	0.796447	0.745161	0.697177
Parameter	$\mathcal{P}_\emptyset$	49.31	0.864000	0.398956	0.085459	0.014359	0.002114
Set 5:	$\boldsymbol{\mathcal{P}_Q}$	**16.21**	**0.344596**				
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	32.32	0.458782	0.220682	0.081305	0.029956	0.011037
$R = 10/11$	$\mathcal{P}_S$	31.31	0.446814	0.216278	0.079683	0.029358	0.010817
Parameter	$\mathcal{P}_\emptyset$	29.65	0.754111	0.132218	0.005671	0.000191	0.000006
Set 6:	$\boldsymbol{\mathcal{P}_Q}$	**10.77**	**0.219253**				
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	12.70	0.243476	0.035323	0.006529	0.001282	0.000252
$R = 100/101$	$\mathcal{P}_S$	11.65	0.223026	0.033942	0.006307	0.001239	0.000243

Table 2: Mean and tail probabilities for number in the system when $s = 3$.

$\mathrm{P}\{N > 650\}$ as 0.245499, 0.250597, and 0.250441, and
$\mathrm{P}\{N > 700\}$ as 0.212541, 0.226916, and 0.226775,

for systems $\mathcal{P}_\emptyset$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$, respectively. Similarly, for systems with three servers and the rates from parameter set 4, we also computed $\mathrm{P}\{N > k\}$ for $k = 900$ and $k = 950$. We had

$\mathrm{P}\{N > 900\}$ as 0.272061, 0.274816, and 0.274564, and
$\mathrm{P}\{N > 950\}$ as 0.241642, 0.257140, and 0.256884,

for systems $\mathcal{P}_\emptyset$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$, respectively. Thus, for this set of rates, $\gamma_\emptyset < \gamma_{Q,F} = \gamma_S$ is observed further in the tail and one needs to consider larger values of k in $\mathrm{P}\{N > k\}$ as s increases. Moreover, the results of Tables 1 and 2 suggest that as in the comparison of L, *the tail probability for the number of customers in the system $\mathcal{P}_Q$ declines faster than it does in the other three systems*. Note that this observation holds even for systems with parameters

	System	L	P$\{N > 15\}$	P$\{N > 50\}$	P$\{N > 100\}$	P$\{N > 150\}$	P$\{N > 200\}$
Parameter	$\mathcal{P}_\emptyset$	130.00	0.972219	0.846337	0.581753	0.340648	0.177490
Set 1:	$\mathcal{P}_Q$						
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	128.35	0.622375	0.516994	0.396634	0.304295	0.233453
$R = 1/2$	$\mathcal{P}_S$	126.00	0.580000	0.510562	0.391700	0.300510	0.230549
Parameter	$\mathcal{P}_\emptyset$	9.13	0.165840	0.001296	0.000001	0.000000	0.000000
Set 2:	$\mathcal{P}_Q$						
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	8.66	0.109407	0.027661	0.003882	0.000545	0.000076
$R = 10/11$	$\mathcal{P}_S$	5.13	0.064235	0.024064	0.003377	0.000474	0.000067
Parameter	$\mathcal{P}_\emptyset$	5.49	0.024310	0.000045	0.000000	0.000000	0.000000
Set 3:	$\mathcal{P}_Q$						
$\rho = 0.5$	$\mathcal{P}_{Q,F}$	5.44	0.013456	0.003360	0.000474	0.000067	0.000009
$R = 100/101$	$\mathcal{P}_S$	1.49	0.011357	0.002878	0.000405	0.000057	0.000008
Parameter	$\mathcal{P}_\emptyset$	1170.00	0.999974	0.999790	0.998937	0.996873	0.992994
Set 4:	$\mathcal{P}_Q$						
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	1137.51	0.901213	0.876249	0.841788	0.808682	0.776878
$R = 1/2$	$\mathcal{P}_S$	1134.00	0.870000	0.891487	0.839417	0.806404	0.774690
Parameter	$\mathcal{P}_\emptyset$	82.19	0.984166	0.738153	0.290812	0.078994	0.017176
Set 5:	$\mathcal{P}_Q$						
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	50.08	0.514850	0.310016	0.165474	0.088329	0.047150
$R = 10/11$	$\mathcal{P}_S$	46.19	0.405847	0.295237	0.157591	0.084122	0.044904
Parameter	$\mathcal{P}_\emptyset$	49.41	0.961504	0.414587	0.039421	0.002216	0.000097
Set 6:	$\mathcal{P}_Q$						
$\rho = 0.9$	$\mathcal{P}_{Q,F}$	17.40	0.314392	0.054232	0.018685	0.006871	0.002528
$R = 100/101$	$\mathcal{P}_S$	13.41	0.227116	0.049039	0.017249	0.006344	0.002334

Table 3: Mean and tail probabilities for number in the system when $s = 5$.

in sets 3 and 6 (i.e., when the proportion of time that each server is up is high).

We conclude this section by discussing how the behavior of systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ depends on the number of servers s, traffic intensity ρ, and server reliability R. Our results indicate that except for system $\mathcal{P}_Q$, the mean number in the system increases as the number of servers s increases. For system $\mathcal{P}_Q$, our results in Tables 1 and 2 show that the mean number in the system can increase or decrease with s. As expected, L is also an increasing function of the traffic intensity ρ. Finally, we observe that the mean number in the system decreases as the proportion of time R that each server is up increases.

On the other hand, tail probabilities of the number of customers in systems $\mathcal{P}_\emptyset$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ decay slower as s increases (the limited results for $\mathcal{P}_Q$ suggest that the tail probabilities can decay faster for larger s in this system). As expected, the decay rate becomes slower as ρ increases. Finally, the decay in the tail probabilities is faster as R increases.

7 Comparison with Reliable Systems

In order to better understand the effects of server failures on the mean number and tail probabilities of system size, we have computed the values in Tables 1 through 3 for corresponding systems with reliable servers. In particular, we consider reliable systems with traffic intensities $\rho = 0.5$ and $\rho = 0.9$. Thus, we have

- Set 7: $\mu = 100$ and $\lambda = 50$;
- Set 8: $\mu = 100$ and $\lambda = 90$.

Tables 4 through 6 provide the mean and the tail probabilities for these two sets of parameters for systems with $s = 2$, $s = 3$, and $s = 5$, respectively. Note that when the servers are reliable, $\mathcal{P}_Q$ and $\mathcal{P}_{Q,F}$ are identical.

	System	L	P$\{N > 15\}$	P$\{N > 50\}$	P$\{N > 100\}$	P$\{N > 150\}$	P$\{N > 200\}$
Parameter Set 7: $\rho = 0.5$	$\mathcal{P}_\emptyset$	2.00	0.000137	0.000000	0.000000	0.000000	0.000000
	$\mathcal{P}_Q$	1.33	0.000020	0.000000	0.000000	0.000000	0.000000
	$\boldsymbol{\mathcal{P}_S}$	**1.00**	**0.000015**	0.000000	0.000000	0.000000	0.000000
Parameter Set 8: $\rho = 0.9$	$\mathcal{P}_\emptyset$	18.00	0.481785	0.028294	0.000265	0.000002	0.000000
	$\mathcal{P}_Q$	9.47	0.195055	0.004883	0.000025	0.000000	0.000000
	$\boldsymbol{\mathcal{P}_S}$	**9.00**	**0.185302**	**0.004638**	**0.000024**	0.000000	0.000000

Table 4: Mean and tail probabilities for number in the system when $s = 2$ and servers are reliable.

	System	L	P$\{N > 15\}$	P$\{N > 50\}$	P$\{N > 100\}$	P$\{N > 150\}$	P$\{N > 200\}$
Parameter Set 7: $\rho = 0.5$	$\mathcal{P}_\emptyset$	3.00	0.000656	0.000000	0.000000	0.000000	0.000000
	$\mathcal{P}_Q$	1.74	0.000029	0.000000	0.000000	0.000000	0.000000
	$\boldsymbol{\mathcal{P}_S}$	**1.00**	**0.000015**	0.000000	0.000000	0.000000	0.000000
Parameter Set 8: $\rho = 0.9$	$\mathcal{P}_\emptyset$	27.00	0.733796	0.089799	0.001497	0.000016	0.000000
	$\mathcal{P}_Q$	10.05	0.207686	0.005199	0.000027	0.000000	0.000000
	$\boldsymbol{\mathcal{P}_S}$	**9.00**	**0.185302**	**0.004638**	**0.000024**	0.000000	0.000000

Table 5: Mean and tail probabilities for number in the system when $s = 3$ and servers are reliable.

Consistent with the literature on queues with reliable servers, Tables 4 to 6 illustrate that $L_S \leq L_Q \leq L_\emptyset$. Thus, $L_\emptyset$ is larger than L_Q and L_S both for reliable and unreliable systems. However, the ordering between L_Q and L_S depends on the system parameters when the servers are subject to failures (see Section 5.6). Moreover, in most of our numerical results in Section 6, we had $L_Q < L_S$ for unreliable systems.

Tables 4 to 6 also show that the tail probabilities decline faster for systems $\mathcal{P}_Q$ and $\mathcal{P}_S$ than for system $\mathcal{P}_\emptyset$ when the servers are reliable. These results are consistent with the

	System	L	P$\{N > 15\}$	P$\{N > 50\}$	P$\{N > 100\}$	P$\{N > 150\}$	P$\{N > 200\}$
Parameter Set 7: $\rho = 0.5$	$\mathcal{P}_\emptyset$	5.00	0.005909	0.000000	0.000000	0.000000	0.000000
	$\mathcal{P}_Q$	4.62	0.000215	0.000000	0.000000	0.000000	0.000000
	$\boldsymbol{\mathcal{P}_S}$	**1.00**	**0.000015**	0.000000	0.000000	0.000000	0.000000
Parameter Set 8: $\rho = 0.9$	$\mathcal{P}_\emptyset$	45.00	0.956826	0.345147	0.016716	0.000365	0.000000
	$\mathcal{P}_Q$	12.91	0.280002	0.007009	0.000036	0.000000	0.000000
	$\boldsymbol{\mathcal{P}_S}$	**9.00**	**0.185302**	**0.004638**	**0.000024**	0.000000	0.000000

Table 6: Mean and tail probabilities for number in the system when $s = 5$ and servers are reliable.

fact that it is beneficial to pool both the queues and the servers in reliable systems, and different from the case with unreliable servers considered in Section 6 where we had the tail probabilities for $\mathcal{P}_Q$ declining much faster than those for $\mathcal{P}_\emptyset$, which in turn decline faster than those of $\mathcal{P}_S$. Hence, in general, it is only beneficial to pool queues in unreliable systems.

We now turn our attention to the dependence of the mean and tail probabilities on s, ρ, and R. In system $\mathcal{P}_\emptyset$ with reliable servers, both the mean and the tail probabilities increase with s. This is consistent with the results for $\mathcal{P}_\emptyset$ with unreliable servers. Similarly, when servers are reliable, the mean number in the system and tail probabilities increase as s increases for system $\mathcal{P}_Q$ (we prove the former result in Proposition A.2 in the Appendix). By contrast, when servers are unreliable, the mean number in the system $\mathcal{P}_Q$ can increase or decrease with s and the tail probabilities of the number in the system decrease (unless the servers are highly reliable) as s increases (whereas the opposite holds for the other systems with unreliable servers). Thus, $\mathcal{P}_Q$ with reliable servers behaves more like $\mathcal{P}_{Q,F}$ with unreliable servers than like $\mathcal{P}_Q$ with unreliable servers. Finally, both the mean and the tail probabilities remain the same for all s in system $\mathcal{P}_S$ with reliable servers. This is reasonable because when the servers are reliable, increasing s in $\mathcal{P}_S$ is equivalent to changing the time scale. This equivalence does not hold for $\mathcal{P}_S$ with unreliable servers (since the failure and repair rates remain unchanged), and increasing s increases the number of customers arriving during down times, which leads to an increase in both the mean and tail probabilities.

As expected, the relationship between the mean (and tail probabilities) of the number in the system and ρ in reliable systems is consistent with the results for unreliable systems. In particular, the mean number in the system increases with ρ, whereas the tail probabilities decay slower for large ρ. Finally, note that the mean and the tail probabilities of the number of customers in systems with rates drawn from parameter sets 7 (8) (depicted in Tables 4 through 6) are lower than those in systems with rates drawn from parameter sets 3 (6) (depicted in Tables 1 through 3). This is consistent with our earlier observation that the mean and the tail probabilities of the number in the system decrease as R increases.

Our numerical results are focused on systems in which failures and repairs occur on a slower scale than arrivals and departures. We conclude this section by considering the mean number of customers in systems with fixed R, λ, and μ values as r and f tend to ∞. This corresponds to having short and frequent breakdowns. As detailed in Proposition A.3, when $s = 2$, we prove that in the limit $L_S < L_Q < L_{Q,F} < L_\emptyset$. This is consistent with the results

of reliable systems. Thus, systems with frequent and short downtimes are similar to reliable systems.

8 Conclusions

In this paper, we investigated the effects of resource pooling on system performance in the presence of failures. The mean and tail probabilities of the number of customers in the system were the performance measures, with the former capturing efficiency and the latter assessing risk. We focused on four systems with differing degrees of pooling, namely $\mathcal{P}_\emptyset$ (s parallel queues with independent server failures), $\mathcal{P}_Q$ (an M/M/s queue with independent server failures – the queues have been pooled), $\mathcal{P}_{Q,F}$ (an M/M/s queue with synchronous server failures – the queues and failures have been pooled), and $\mathcal{P}_S$ (an unreliable single-server queue where the server is s times faster than the servers in the other three systems – the servers, and hence queues and failures, have been pooled). Our objective was to study the tradeoff between efficiency and risk.

Overall, the system $\mathcal{P}_Q$ (with queues pooled) appears to show the best performance, both in terms of efficiency and risk. However, for systems with highly reliable servers, system $\mathcal{P}_S$ can outperform system $\mathcal{P}_Q$ in terms of efficiency. The systems $\mathcal{P}_{Q,F}$ and $\mathcal{P}_S$ in which failures are pooled perform similarly both in terms of efficiency (unless servers are highly reliable) and risk, and outperform the un-pooled system $\mathcal{P}_\emptyset$ in terms of efficiency but not risk. Thus, our results indicate that pooling queues is desirable, but that pooling servers is not necessarily wise in the presence of failures. Moreover, changes aimed at improving efficiency may increase risk and vice versa. By contrast, resource pooling in reliable systems simultaneously improves efficiency and reduces risk, and more pooling is better than less pooling (e.g., pooling queues is good, but pooling queues and servers is better). Thus, one should use caution when applying insights obtained from studying reliable systems to make resource pooling decisions for unreliable systems, and such decisions should address not only efficiency but also risk.

The belief that resource pooling is always beneficial is deep-rooted in our community. For example, the Wikipedia page on pooling [27] starts with the statement that "Pooling is a resource management term that refers to the grouping together of resources (assets, equipment, personnel, effort, etc.) for the purposes of maximizing advantage and/or minimizing risk to the users." Our research shows that pooling may decrease efficiency and increase risk, contrary to conventional wisdom.

Acknowledgments

We thank the Editor and three anonymous referees for their insightful suggestions which we believe have improved the paper. This research was supported by the National Science Foundation under Grant CMMI–0856600. The second author was also supported by the National Science Foundation under Grant CMMI–0969747. The research of the third author was also supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] Argon, N. T., and S. Andradóttir. "Partial Pooling in Tandem Lines with Cooperation and Blocking," *Queueing Systems*, 52:5–30, 2006.

[2] Benjaafar, S. "Performance Bounds for the Effectiveness of Pooling in Multi-processing Systems," *European Journal of Operational Research*, 87:375–388, 1995.

[3] Borst, S.C., A. Mandelbaum, and M.I. Reiman. "Dimensioning Large Call Centers," *Operations Research*, 52:17–34, 2004.

[4] Buzacott, J. A. "Commonalities in Reengineered Business Processes: Models and Issues," *Management Science*, 42:768–782, 1996.

[5] Calabrese, J. B. "Optimal Workload Allocation in Open Networks of Multiserver Queues," *Management Science*, 38:1792–1802, 1992.

[6] Chao, K., and Y. Zhao. "Analysis of Multi-server Queues with Station and Server Vacations," *European Journal of Operational Research*, 110:392-406, 1998.

[7] Foley, R. D., and D. R. McDonald. "Join the Shortest Queue: Stability and Exact Asymptotics," *Annals of Applied Probability*, 11:569-607, 2001.

[8] Foley, R. D., and D. R. McDonald. "Large Deviations of a Modified Jackson Network: Stability and Rough Asymptotics," *Annals of Applied Probability*, 15:519-541, 2005.

[9] Gross, D., and C. M. Harris. *Fundamentals of Queueing Theory*, second edition, John Wiley and Sons, New York, 1985.

[10] Hokstad, P. "Approximations for the M/G/m Queue," *Operations Research*, 26:510–523, 1978.

[11] Larson, R. C. "Perspectives on Queues: Social Justice and the Psychology of Queueing," *Operations Research*, 35:895–905, 1987.

[12] Levy, Y., and U. Yechiali. "An M/M/c Queue with Server's Vacations," *INFOR*, 14:153-163, 1976.

[13] Lorek, P. "The Exact Asymptotic for the Stationary Distribution of Some Unreliable Systems," arXiv:1102.4707, 2011.

[14] Mandelbaum, A., and M. I. Reiman. "On Pooling in Queueing Networks," *Management Science*, 44:971–981, 1998.

[15] Mitrani, I. L., and B. Avi-Itzhak. "A Many-server Queue with Service Interruptions," *Operations Research*, 16:628-638, 1968.

[16] Pang, G., and W. Whitt. "Heavy-Traffic Limits for Many-Server Queues with Service Interruptions," *Queueing Systems*, 61:167-202, 2009.

[17] Rothkopf, M. H., and P. Rech. "Perspective on Queues: Combining Queues Is Not Always Beneficial," *Operations Research*, 35:906–909, 1987.

[18] Scheller-Wolf, A. "Necessary and Sufficient Conditions for Delay Moments in FIFO Multiserver Queues with an Application Comparing s Slow Servers with a Fast One," *Operations Research*, 51:748-758, 2003.

[19] Smith, D. R., and W. Whitt. "Resource Sharing for Efficiency in Traffic Systems," *The Bell System Technical Journal*, 60:39–55, 1981.

[20] Tian, N., Q. Li, and J. Cao. "Conditional Stochastic Decomposition in M/M/c Queue with Server Vacations," *Stochastic Models*, 15:367–377, 1999.

[21] van Dijk, N. M., and E. van der Sluis. "To Pool or not to Pool in Call Centers," *Production and Operations Management*, 17:1–10, 2008.

[22] van Dijk, N. M., and E. van der Sluis. "Pooling is not the Answer," *European Journal of Operational Research*, 197:415–421, 2009.

[23] Van Oyen, M. P., E. G. S. Gel, and W. J. Hopp. "Performance Opportunity of Workforce Agility in Collaborative and Non-collaborative Work Systems," *IIE Transactions*, 33: 761–777, 2001.

[24] Vinod, B. "Exponential Queue with Server Vacations," *Journal of the Operational Research Society*, 37:1007-1014, 1986.

[25] Wallace, R. B., and W. Whitt. "A Staffing Algorithm for Call Centers with Skill-Based Routing," *Manufacturing and Service Operations Management*, 7:276–294, 2005.

[26] Wiens, D. P. "On the Busy Period Distribution of the M/G/2 Queueing System," *Journal of Applied Probability*, 26:858-865, 1989.

[27] Wikipedia. "Pooling (Resource Management)," en.wikipedia.org/wiki/Pooling_(resource_management), accessed on June 29, 2012.

[28] Wolff, R. W. *Stochastic Modeling and the Theory of Queues*, Prentice Hall, New Jersey, 1988.

[29] Zhang, Z. G., and N. Tian. "Analysis on queueing systems with synchronous vacations of partial servers," *Performance Evaluation*, 52:269-282, 2003.

A Appendix

A.1 Proof of Proposition 4.1

The probabilities $p_{k,w}$ and $p_{k,f}$ are as defined in Section 3.3. We apply Proposition 2.1 of Lorek [13] to conclude that for finite constants $c^*_{\emptyset,w}$ and $c^*_{\emptyset,f}$, $p_{k,w} \sim c^*_{\emptyset,w}\gamma^k_\emptyset$ and $p_{k,f} \sim c^*_{\emptyset,f}\gamma^k_\emptyset$

(Lorek [13] discusses how these constants can be computed). Hence, $p_k = p_{k,w} + p_{k,f} \sim c^*_\emptyset \gamma^k_\emptyset$, where $c^*_\emptyset = c^*_{\emptyset,w} + c^*_{\emptyset,f}$. Let $\varepsilon > 0$. Then, there exists a K such that for all $k > K$,

$$\left| \frac{p_k}{c^*_\emptyset \gamma^k_\emptyset} - 1 \right| < \varepsilon. \tag{17}$$

Define

$$A = \min_{k=0,1,\ldots,K} p_k \tag{18}$$

and

$$B = \max_{k=0,1,\ldots,K} \left| \frac{p_k}{c^*_\emptyset \gamma^k_\emptyset} - 1 \right|. \tag{19}$$

We write

$$\lim_{t\to\infty} P\{Q_1(t) + Q_2(t) + \cdots + Q_s(t) = \ell\} = \sum_{k_1+k_2+\cdots+k_s=\ell} \prod_{i:k_i\le K} p_{k_i} \prod_{i:k_i>K} p_{k_i}. \tag{20}$$

We will now provide upper and lower bounds on (20) given that $\ell \ge sK + 1$, so that there is at least one term in the second product. Using (17) and (18), a lower bound is

$$\begin{aligned}
&\sum_{k_1+k_2+\cdots+k_s=\ell} \prod_{i:k_i\le K} A \prod_{i:k_i>K} (1-\varepsilon) c^*_\emptyset \gamma^{k_i}_\emptyset \\
&\quad \ge \sum_{k_1+k_2+\cdots+k_s=\ell} A^{s-1}(1-\varepsilon)^s \min(c^*_\emptyset, (c^*_\emptyset)^s) \gamma_\emptyset^{\ell - \sum_{i:k_i\le K} k_i} \\
&\quad \ge \sum_{k_1+k_2+\cdots+k_s=\ell} A^{s-1}(1-\varepsilon)^s \min(c^*_\emptyset, (c^*_\emptyset)^s) \gamma^\ell_\emptyset \\
&\quad = \binom{\ell+s-1}{\ell} A^{s-1}(1-\varepsilon)^s \min(c^*_\emptyset, (c^*_\emptyset)^s) \gamma^\ell_\emptyset.
\end{aligned}$$

From (17) and (19), when $\ell \ge sK + 1$, (20) has upper bound

$$\begin{aligned}
&\sum_{k_1+k_2+\cdots+k_s=\ell} \prod_{i:k_i\le K} c^*_\emptyset \gamma^{k_i}_\emptyset (B+1) \prod_{i:k_i>K} (1+\varepsilon) c^*_\emptyset \gamma^{k_i}_\emptyset \\
&\quad \le \sum_{k_1+k_2+\cdots+k_s=\ell} \max(1, (c^*_\emptyset)^{s-1})(B+1)^{s-1}(1+\varepsilon)^s \max(1, (c^*_\emptyset)^s) \gamma^\ell_\emptyset \\
&\quad = \binom{\ell+s-1}{\ell} \max(1, (c^*_\emptyset)^{s-1})(B+1)^{s-1}(1+\varepsilon)^s \max(1, (c^*_\emptyset)^s) \gamma^\ell_\emptyset.
\end{aligned}$$

The result now follows. □

A.2 Proof of Proposition 5.1

Let $\{a_n : n \ge 1\}$ denote the sequence of arrival times to both systems, where in $\mathcal{P}_\emptyset$ each arrival is assigned to one of the s queues with equal probability. Define $\{t_n : n \ge 1\}$ as the sequence of event epochs of a Poisson process with intensity $s\mu$. Note that $\{t_n : n \ge 1\}$ is

the sequence of potential service completion times in both systems. Finally, generate the server failure times and repair times so that each server in both systems experiences the same down times (when the server is undergoing repair). For all $t \geq 0$ and stochastic processes $\{N(t)\}$, let $N(t-) = \lim_{s \nearrow t} N(s)$. If all servers are functioning at time t_n and $N_Q(t_n-) \geq s$, then t_n is an actual departure time for $\mathcal{P}_Q$ with probability 1, whereas if $N_\emptyset(t_n-) \geq s$, the probability that t_n is an actual departure time for $\mathcal{P}_\emptyset$ can be less than 1 since some of the s servers may be idle at time t_n in $\mathcal{P}_\emptyset$. Similarly, if $0 < k < s$ servers are down at time t_n and $N_Q(t_n-) \geq s$, then t_n is an actual departure time for $\mathcal{P}_Q$ with probability $\frac{s-k}{s}$, whereas if $N_\emptyset(t_n-) \geq s$, the probability that t_n is an actual departure time for $\mathcal{P}_\emptyset$ can be less than $\frac{s-k}{s}$ since some of the $s-k$ functioning servers may be idle at time t_n in $\mathcal{P}_\emptyset$. If all servers are functioning at time t_n and $N_Q(t_n-) = m < s$, then t_n is an actual departure time for $\mathcal{P}_Q$ with probability $\frac{m}{s}$ whereas if $N_\emptyset(t_n-) = m < s$, the probability that t_n is an actual departure time for $\mathcal{P}_\emptyset$ can be less than $\frac{m}{s}$ since the number of customers being served at time t_n can be less than m in $\mathcal{P}_\emptyset$. Similarly, if $0 < k < s$ servers are down at time t_n and $N_Q(t_n-) = m < s$, then t_n is an actual departure time for $\mathcal{P}_Q$ with probability $\frac{\min\{s-k,m\}}{s}$ whereas if $N_\emptyset(t_n-) = m < s$, the probability that t_n is an actual departure time for $\mathcal{P}_\emptyset$ can be less than $\frac{\min\{s-k,m\}}{s}$ since the number of customers served at time t_n can be less than $\min\{s-k,m\}$ in $\mathcal{P}_\emptyset$. Finally, if all s servers are down at time t_n, then the probability that t_n is a departure time is zero for both systems regardless of the number of customers in the system at time t_n. Since both systems experience the same arrival process $\{a_n : n \geq 1\}$, this immediately yields that $N_Q(t) \overset{st}{\leq} N_\emptyset(t)$ for all $t \geq 0$. □

A.3 Proof of Proposition 5.2

Let $\{a_n : n \geq 1\}$ and $\{t_n : n \geq 1\}$ be defined as in the proof of Proposition 5.1. First assume that both servers are reliable. Then note that if $N_{Q,F}(t_n-) = m \geq 1$, then t_n is an actual departure time for $\mathcal{P}_{Q,F}$ with probability $\frac{\min\{m,s\}}{s}$, whereas if $N_S(t_n-) \geq 1$, then t_n is an actual departure time for $\mathcal{P}_S$ with probability 1. Since both systems experience the same arrival process $\{a_n : n \geq 1\}$, this immediately yields the desired stochastic ordering result for systems with reliable servers. Similarly, one can generate the server failure times and repair times so that both systems experience the same down times. Then having server failures is equivalent to setting the probability of t_n being a departure time equal to 0 (for both systems) if t_n occurs during a server down time, and the result follows. □

A.4 Results for Reliable Systems

The next two propositions are on reliable systems. Consider an M/M/1 queue with arrival rate $s\lambda$ and service rate $s\mu$ and an M/M/s queue with arrival rate $s\lambda$ and service rate μ for each server, where $s \geq 2$. Let $N^{(1)}$ and $N^{(s)}$ be the steady-state number of customers in the single-server and s-server queues, respectively, when $\frac{\lambda}{\mu} < 1$. We have the following result whose proof uses an argument similar to the one in Wolff [28] (pages 257–258).

Proposition A.1 *There exists* $0 < k^* \leq s-1$ *such that* $\mathrm{P}\{N^{(1)} = k\} > \mathrm{P}\{N^{(s)} = k\}$ *for all* $k < k^*$ *and* $\mathrm{P}\{N^{(1)} = k\} < \mathrm{P}\{N^{(s)} = k\}$ *for all* $k \geq k^*$.

Proof Using the corresponding CTMCs, one can immediately show that

$$\begin{aligned}
\frac{\mathrm{P}\{N^{(1)}=k+1\}}{\mathrm{P}\{N^{(1)}=k\}} &= \frac{\lambda}{\mu} \text{ for all } k=0,1,\ldots,\\
\frac{\mathrm{P}\{N^{(s)}=k+1\}}{\mathrm{P}\{N^{(s)}=k\}} &= \frac{s\lambda}{(k+1)\mu} \text{ for all } k=0,1,\ldots,s-2,\\
\frac{\mathrm{P}\{N^{(s)}=k+1\}}{\mathrm{P}\{N^{(s)}=k\}} &= \frac{\lambda}{\mu} \text{ for all } k=s-1,s,\ldots.
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\mathrm{P}\{N^{(1)}=k+1\}}{\mathrm{P}\{N^{(1)}=k\}} &< \frac{\mathrm{P}\{N^{(s)}=k+1\}}{\mathrm{P}\{N^{(s)}=k\}} \text{ for all } k=0,1,\ldots,s-2,\\
\frac{\mathrm{P}\{N^{(1)}=k+1\}}{\mathrm{P}\{N^{(1)}=k\}} &= \frac{\mathrm{P}\{N^{(s)}=k+1\}}{\mathrm{P}\{N^{(s)}=k\}} \text{ for all } k=s-1,s,\ldots.
\end{aligned}$$

The result now follows from the fact that $\sum_{k=0}^{\infty}\mathrm{P}\{N^{(1)}=k\}=\sum_{k=0}^{\infty}\mathrm{P}\{N^{(s)}=k\}=1$. □

We now prove our observation that in a queueing system with Poisson arrivals of rate $s\lambda$ and s reliable servers whose service times are exponentially distributed with rate μ (i.e., system $\mathcal{P}_Q$ with reliable servers), the mean number of customers in the system (L) increases with the number of servers s.

Proposition A.2 *Let L^s denote the steady-state mean number of customers in a reliable $M/M/s$ queue with arrival rate $s\lambda$ and service rate μ for each server, where $\lambda<\mu$. Then $L^s \le L^{s+1}$ for $s=1,2,\ldots$.*

Proof It is well known that

$$L^s = \frac{s\lambda}{\mu}+\frac{(\frac{s\lambda}{\mu})^s\lambda\mu}{s!(\mu-\lambda)^2}\Big(\sum_{i=0}^{s-1}\frac{1}{i!}(\frac{s\lambda}{\mu})^i+\frac{1}{s!}\frac{(\frac{s\lambda}{\mu})^s\mu}{\mu-\lambda}\Big)^{-1}$$

(see for example page 88 of Gross and Harris [9]). This yields

$$L^s = \frac{\lambda[s^2e^{\frac{s\lambda}{\mu}}(\mu-\lambda)^2\Gamma(s,\frac{s\lambda}{\mu})+(\frac{s\lambda}{\mu})^s\mu((s+1)\mu-s\lambda)]}{(\mu-\lambda)\mu[s(\mu-\lambda)e^{\frac{s\lambda}{\mu}}\Gamma(s,\frac{s\lambda}{\mu})+\mu(\frac{s\lambda}{\mu})^s]},$$

where $\Gamma(a,z)=\int_z^{\infty}e^{-t}t^{a-1}dt$ is the incomplete Gamma function. Then some algebra yields

$$\begin{aligned}
&L^{s+1}-L^s\\
&= \frac{\lambda[e^{\frac{\lambda(2s+1)}{\mu}}s(\mu-\lambda)^2\Gamma(s,\frac{s\lambda}{\mu})\Gamma(s+1,\frac{(s+1)\lambda}{\mu})+e^{\frac{s\lambda}{\mu}}s(\mu-\lambda)\lambda(\frac{(s+1)\lambda}{\mu})^s\Gamma(s,\frac{s\lambda}{\mu})]}{\mu[e^{\frac{s\lambda}{\mu}}s(\mu-\lambda)\Gamma(s,\frac{s\lambda}{\mu})+\mu(\frac{s\lambda}{\mu})^s][e^{\frac{(s+1)\lambda}{\mu}}(\mu-\lambda)\Gamma(s+1,\frac{(s+1)\lambda}{\mu})+\lambda(\frac{(s+1)\lambda}{\mu})^s]} \qquad (21)\\
&+\frac{[e^{\frac{s\lambda}{\mu}}s\mu\lambda(\frac{(s+1)\lambda}{\mu})^s\Gamma(s,\frac{s\lambda}{\mu})-e^{\frac{(s+1)\lambda}{\mu}}\mu\lambda(\frac{s\lambda}{\mu})^s\Gamma(s+1,\frac{(s+1)\lambda}{\mu})+\mu\lambda(\frac{s\lambda}{\mu})^s(\frac{(s+1)\lambda}{\mu})^s]}{\mu[e^{\frac{s\lambda}{\mu}}s(\mu-\lambda)\Gamma(s,\frac{s\lambda}{\mu})+\mu(\frac{s\lambda}{\mu})^s][e^{\frac{(s+1)\lambda}{\mu}}(\mu-\lambda)\Gamma(s+1,\frac{(s+1)\lambda}{\mu})+\lambda(\frac{(s+1)\lambda}{\mu})^s]}. \qquad (22)
\end{aligned}$$

Note that the expression in (21) is positive. Furthermore, using the definition of the incomplete Gamma function, with some algebra the numerator of (22) can be written as

$$\lambda\mu s!(\frac{\lambda}{\mu})^s \sum_{k=0}^{s-1} \frac{(s+1)^k s^k}{k!}(\frac{\lambda}{\mu})^k \left[(s+1)^{s-k} - s^{s-k}\right] > 0,$$

which completes the proof. □

A.5 Results for Frequent and Short Downtimes

In this section, we compute the limiting values of the mean number of customers in systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ as both the failure rate f and the repair rate r tend to infinity (at the same rate) and investigate how they compare to each other.

Proposition A.3 *Suppose $s = 2$ and $f = ar$, where $a > 0$ (and, hence, $R = \frac{1}{1+a}$ and $\lambda < \frac{\mu}{1+a}$). Then*

$$\lim_{r\to\infty} L_\emptyset = \frac{2(a+1)\lambda}{\mu - (a+1)\lambda},$$

$$\lim_{r\to\infty} L_Q = \frac{2(a+1)^2\mu\lambda}{(\mu - (a+1)\lambda)((2a+1)\mu + (a+1)\lambda)},$$

$$\lim_{r\to\infty} L_{Q,F} = \frac{2(a+1)\lambda\mu}{(\mu + (a+1)\lambda)(\mu - (a+1)\lambda)},$$

and

$$\lim_{r\to\infty} L_S = \frac{(a+1)\lambda}{(\mu - (a+1)\lambda)}.$$

Thus, we have

$$\lim_{r\to\infty} L_S < \lim_{r\to\infty} L_Q < \lim_{r\to\infty} L_{Q,F} < \lim_{r\to\infty} L_\emptyset.$$

Proof The first part of the proposition follows from taking the limits of the closed form expressions of $L_\emptyset$, L_Q, $L_{Q,F}$, and L_S in Sections 3.1, 3.2, 3.3, and 3.4, respectively. Though the limits for $L_\emptyset$, $L_{Q,F}$, and L_S are straightforward, computing the limit for L_Q is more tedious (due to the square root terms) and requires using the L'Hopital rule. Then with some algebra we have

$$\lim_{r\to\infty} L_\emptyset - \lim_{r\to\infty} L_{Q,F} = \frac{2(a+1)^2\lambda^2}{(\mu + (a+1)\lambda)(\mu - (a+1)\lambda)} > 0,$$

$$\lim_{r\to\infty} L_{Q,F} - \lim_{r\to\infty} L_Q \;=\; \frac{2a(a+1)\lambda\mu}{(\mu+(a+1)\lambda)((2a+1)\mu+(a+1\lambda))} > 0,$$

and

$$\lim_{r\to\infty} L_Q - \lim_{r\to\infty} L_S \;=\; \frac{(a+1)\lambda}{(2a+1)\mu+(a+1)\lambda} > 0.$$

This concludes the proof. □

Note that in the limit $\mathcal{P}_\emptyset$ behaves as two reliable M/M/1 queues with service rate $\frac{\mu}{(a+1)}$. Similarly, in the limit, $\mathcal{P}_S$ behaves as a reliable M/M/1 queue with arrival rate 2λ and service rate $\frac{2\mu}{a+1}$, and $\mathcal{P}_{Q,F}$ behaves as a reliable M/M/2 system with arrival rate 2λ and service rate of each server $\frac{\mu}{a+1}$.

A.6 Results for Infrequent and Long Downtimes

In this section, we compute the limiting values of the mean number of customers in systems $\mathcal{P}_\emptyset$, $\mathcal{P}_Q$, $\mathcal{P}_{Q,F}$, and $\mathcal{P}_S$ as both the failure rate f and the repair rate r tend to 0 (at the same rate) and investigate how they compare to each other.

Proposition A.4 *Suppose $s = 2$ and $f = ar$, where $a > 0$ (and, hence, $R = \frac{1}{1+a}$ and $\lambda < \frac{\mu}{1+a}$). Then*
(i)

$$\lim_{r\to 0} L_\emptyset = \lim_{r\to 0} L_Q = \lim_{r\to 0} L_{Q,F} = \lim_{r\to 0} L_S = \infty,$$

(ii)

$$\lim_{r\to 0} \frac{L_\emptyset}{L_{Q,F}} = \lim_{r\to 0} \frac{L_\emptyset}{L_S} = 1.$$

If $2\lambda \geq \mu$,

$$\lim_{r\to 0} \frac{L_\emptyset}{L_Q} \;=\; \frac{2\lambda(\lambda+\mu)(a+1)}{\lambda(2a\mu+\lambda)+\mu(2\lambda-\mu)} > 1 \tag{23}$$

and if $2\lambda \leq \mu$,

$$\lim_{r\to 0} \frac{L_\emptyset}{L_Q} \;=\; \frac{2(a+1)^2(\lambda+\mu)(\mu-\lambda)}{(a^2\lambda+\mu)(\mu-2\lambda)+(a+1)(\lambda^2+a\mu^2)} > 1, \tag{24}$$

with $\lim_{r\to 0} \frac{L_\emptyset}{L_Q} = \frac{6(a+1)}{4a+1}$ when $2\lambda = \mu$.
(iii)

$$\lim_{r\to 0} \frac{L_{Q,F}}{L_S} = 1,$$

$$\lim_{r\to 0} \frac{L_{Q,F}}{L_Q} = \lim_{r\to 0} \frac{L_S}{L_Q} = \lim_{r\to 0} \frac{L_\emptyset}{L_Q}.$$

Proof Parts (i) and (ii) of the proposition follow from taking the limits of the closed form expressions of $L_\emptyset$, L_Q, $L_{Q,F}$, and L_S in Sections 3.1, 3.2, 3.3, and 3.4, respectively. The expressions in (23) and (24) are greater than 1 since

$$2\lambda(\lambda+\mu)(a+1) - [\lambda(2a\mu+\lambda) + \mu(2\lambda-\mu)] = \lambda^2(2a+1) + \mu^2 > 0$$

and

$$\begin{aligned} & 2(a+1)^2(\lambda+\mu)(\mu-\lambda) - [(a^2\lambda+\mu)(\mu-2\lambda) + (a+1)(\lambda^2+a\mu^2)] \\ = \; & a^2\mu(\mu-\lambda) + a(3\mu^2-5\lambda^2) + \mu(\mu+2\lambda) - 3\lambda^2 > 0 \end{aligned}$$

since in this case $\mu \geq 2\lambda$. Part (iii) follows immediately from part (ii). □