Submitted to *Management Science*
manuscript MS-00340-2008.R2

Service Interruptions In Large-Scale Service Systems

Guodong Pang, Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699
{gp2224, ww2040}@columbia.edu

Large-scale service systems, where many servers respond to high demand, are appealing because they can provide great economy of scale, producing a high quality of service with high efficiency. Customer waiting times can be short, with a majority of customers served immediately upon arrival, while server utilizations remain close to 100%. However, we show that this confluence of quality and efficiency is not achieved without risk, because there can be severe congestion if the system does not operate as planned. In particular, we show that the large scale makes the system more vulnerable to service interruptions when (i) most customers remain waiting until they can be served, and (ii) when many servers are unable to function during the interruption, as may occur with a system-wide computer failure. Increasing scale leads to higher server utilizations, which in turn leads to longer recovery times from service interruptions and worse performance during such events. We quantify the impact of service interruptions with increasing scale by introducing and analyzing approximating deterministic fluid models. We also show that these fluid models can be obtained from many-server heavy-traffic limits.

Key words: service interruptions; service systems; economy of scale; call centers; many-server queues; deterministic fluid models; heavy-traffic limits; rare-event simulation.

History: This paper was first submitted on November 7, 1922, and has been with the authors for 86 years for 65 revisions.

1. Introduction

A clear trend in call centers and many other service systems is *increasing scale*. Whereas, formerly, a large call center might employ hundreds of agents, today a large call center employs thousands of agents.

1.1. Economy of Scale

As discussed by Gans et al. (2003) and Aksin et al. (2007) in recent surveys of customer contact centers and their modelling, there is a good reason to see increasing scale, because there is a significant economy of scale, strongly supported by queueing theory. Whitt (1992) applied queueing theory to explain, support and quantify the economy of scale in many-server service systems. This economy of scale was further exposed in a cost-benefit framework by Borst et al. (2004). Indeed, from the first study of many-server queueing models by A. K. Erlang a century ago, the advantage of large scale has been recognized and quantified; see Brockmeyer et al. (1948). Simulations and mathematical analysis show that performance tends to improve as the number of servers increases with the utilization per server held fixed. Alternatively, the utilization per server can approach the upper limit 100% as the number of servers increases with various performance measures held fixed. Qualitatively, the advantage of large scale is supported by stochastic comparisons in multi-server queueing models, as in Smith and Whitt (1981). There it is shown that the appropriate overall measures of performance are improved (the level of congestion is decreased) when two separate service systems with common service times are combined.

As we will explain in §3, the advantage of large scale is also supported by stochastic-process limits for many-server queueing models in the *quality-and-efficiency-driven* (QED) *many-server heavy-traffic limiting regime*; e.g., see Halfin and Whitt (1981), Garnett et al. (2002) and Pang et al. (2007). In the QED regime, the scale – as represented by the arrival rate and the number of servers – is allowed to increase without bound, where these two measures of scale increase together appropriately, so that supply matches demand. In that limit, simultaneously, waiting times become negligible (quality), while server utilizations approach 100% (efficiency).

1.2. Problems Posed By Large Scale

The purpose of this paper is to temper the enthusiasm for large scale. Here we emphasize that the quality-and-efficiency gains from large scale are achieved at some risk, because there can be severe congestion if the system does not operate as planned. In particular, here we show that *large scale makes the system more vulnerable to service interruptions* when (i) most customers remain waiting in the system until they can be served, and (ii) many servers are unable to function during the interruption, as may occur with a system-wide computer failure. Moreover, we show how to *quantify* the performance impact of the interruptions in a relatively tractable way, so that the *tradeoff* between efficiency gains and interruption costs associated with increasing scale can be assessed.

We were motivated to consider this case because of our own personal experience at a large Department of Motor Vehicles (DMV) office in New York City, where hundreds of people were waiting to complete various processing tasks. During that visit, there was a system-wide computer failure that lasted for about 90 minutes, which stopped all service, and yet almost all customers waited to complete their business. Customer patience in that setting is understandable, because there was a large invested cost in coming to the DMV office in the first place, and in acquiring a place in line, and the length of the interruption was uncertain. It was natural to hope that the difficulty might be resolved any minute. Because of the large scale, hundreds of customers experienced hours of extra delays. Much more serious consequences could occur in a large hospital, transportation system or food-distribution system. Clearly, customer delays constitute only one component of the cost of service interruptions, but they are an important component.

Our results here are consistent with other recent research exposing difficulties associated with increasing scale. From Whitt (2006a,b) and Bassamboo et al. (2006a,b), we see that large-scale service systems are vulnerable to uncertainty about the arrival and service rates. Whitt (2006a) showed that the sensitivity of the principal performance measures to the arrival rate and the service rate increase with increasing scale. In particular, the arrival-rate and service-rate elasticities of various performance measures grow proportionally to the square root of the number of servers in the QED regime. (For example, if $\mathcal{E}$ is the arrival-rate elasticity of the delay probability, then a 1% increase in the arrival rate tends to produce an $\mathcal{E}\%$ increase in the delay probability.)

We next describe our modelling and analysis approach. At the end of the next section we indicate how the rest of the paper is organized.

2. Modelling

The congestion impact of a service interruption clearly depends on what happens to the customers during the interruption: Arrivals may either continue or stop. The customers already in the system may either remain waiting or they may leave without receiving service.

2.1. The Pure-Delay Model and the Pure-Loss Model

We will consider a range of cases, but we will primarily focus on two extreme cases: First, arrivals may continue and all customers may remain waiting; second, new arrivals may refuse to enter and all customers in the system may leave immediately without receiving service. The first case has delays but no losses, while the second case has losses but no delays. Consistent with intuition, the number of customers affected in the pure-delay case tends to be much greater than in the pure-loss case, because waiting customers not only experience their own delays, but increase the delays of other customers.

We are primarily concerned with the pure-delay case, which requires more careful analysis. In the pure-delay case, the congestion impact of service interruptions increases with increasing scale. As we explain in §3, increasing scale allows the server utilization (or traffic intensity) to be higher, given standard quality-of-service constraints. (That is the much touted economy of scale in this context.) As a consequence the recovery rate after the interruption has ended tends to be slower. Thus, with increasing scale, the recovery time tends to increase and the performance during that event degrades significantly. The bad performance spreads to customers that arrive long after the interruption has ended.

In contrast, in the pure-loss case, the congestion impact is much less. Many customers fail to receive service at all, which naturally may be regarded as a more serious penalty, but there is little impact on other customers that arrive after the interruption has ended. Even though the impact of lost service may be great, relatively few customers will be affected if interruptions are rare. In contrast, with large-scale pure-delay systems, even rare short interruptions can have a dramatic impact on congestion, because they can produce long recovery times.

2.2. A More General Model

But the two extremes discussed above are not the only cases. In practice, service systems tend to operate in between these two extremes, often having customer abandonment after some waiting. Fortunately, abandonment usually tends to make the system behave more like the pure-loss model; see §6.

To provide a basis for further systematic analysis, we also consider a more general model that covers a wide range of intermediate cases, allowing customer abandonment at various rates. When the system is operating normally, customers will be served and abandon from queue at nominal rates. In particular, when there is no interruption, we assume that the system behaves as the Markovian $M/M/n+M$ (Erlang-A or Palm) model with unlimited waiting space, the first-come first-served (FCFS) service discipline, arrival rate λ, individual service rate μ_1 and individual abandonment rate θ_1.

Here is what happens during the service interruptions: First, we assume that arrivals continue arriving at rate λ, even during the interruption. When an interruption occurs, it lasts for a random length of time, the down time D. Throughout that interruption, a random number F of the servers remain functioning, which may range from 0 to n; we think of F as being proportional to n, so that F/n is the random proportion of functioning servers.

Customers in service at functioning servers continue receiving service, but at a new service rate, μ_2 instead of μ_1. That rate μ_2 may be slower than μ_1, reflecting service degradation caused by the interruption, or that rate may be faster, because of a special effort to provide exceptional service during the interruption. Customers in queue continue waiting, but abandon at a new rate, θ_2 instead of θ_1. We would expect to have $\theta_2 > \theta_1$, but we treat the general case.

There are several possible assumptions for the customers that are in service at servers that cease functioning. We assume that these customers remain at these servers, but have high priority (over customers waiting in queue or new arrivals) for newly available functioning servers when they become available, which preserves the FCFS order. These customers at non-functioning servers may also abandon from the system, and do so at a new rate θ_3. We would expect to have $\theta_3 > \theta_1$, but again we treat the general case. We assume that all the service and abandonment times are independent exponential random variables.

The pure-delay model with a system-wide service interruption is obtained as the special case in which there are no functioning servers ($F = 0$) and these abandonment rates - θ_1, θ_2, and θ_3, - are all zero, while the pure-loss model is obtained as the special case in which again $F = 0$ but these abandonment rates are all infinite. We quantify performance, conditional on the pair of random variables (D,F), as a function of the 6-tuple of model parameters $(\lambda,\mu_1,\theta_1,\mu_2,\theta_2,\theta_3)$. However, we especially emphasize the severe performance degradation in the pure-delay case with $F=0$ and $\theta_1=\theta_2=\theta_3=0$.

2.3. Models of the Service Interruptions

There are two different ways to look at these service interruptions. First, we can consider *a single service interruption in isolation*, starting from the system operating normally in steady state. Second, we can look at *a sequence of successive interruptions*, and try to properly account for the cumulative impact of several successive interruptions, possibly with new interruptions occurring before the system has fully recovered from the previous service interruption. We primarily focus on the single-interruption case, because it is much easier to analyze, and because we think that it already captures the main impact, assuming that the service interruptions are relatively rare events.

In both cases, we assume that interruptions occur exogenously. By "exogenous," we mean that they are not caused by events in the queueing system. As stated before, we are thinking of a computer failure or an electronic outage. Thus interruptions are specified a priori. That makes the general model a Markovian many-server queueing model in an exogenous random environment, with the parameter 6-tuple $(\lambda,\mu_1,\theta_1,\mu_2,\theta_2,\theta_3)$ specified above.

To be concrete, for the case of multiple interruptions over time, we consider an alternating-renewal-process environment, but that easily can be generalized if there is good reason to do so (e.g., if interruptions tend to occur in clusters). Here we assume that there is a sequence of independent and identically distributed random vectors $\{(U_k,D_k,F_k):k\geq 1\}$, where U_k is the k^{th} up time, D_k is the k^{th} down time, and F_k is the random number of the n available servers that is functioning throughout the k^{th} down time. The parameter pair (μ_1,θ_1) prevails during each up time, while the parameter triple $(\mu_2,\theta_2,\theta_3)$ prevails during each down time. We assume that the system starts at time $T_0\equiv 0$ at the beginning of the first up interval.

2.4. An Approximating Deterministic Fluid Model

It should be evident that the many-server queueing model in a random environment is quite complicated, even with all the Markovian assumptions that we have made. Thus we resort to approximations. In particular, we propose an approximating deterministic fluid model. Such rough fluid approximations tend to be effective (sufficiently accurate for practical engineering purposes) when the system can be significantly overloaded; e.g., see Newell (1982), Hall (1991), Chen and Yao (1992), Choudhury et al. (1997), Whitt (2002, 2004, 2006a,b,c) and references therein. And that is just what happens with the service interruptions.

The deterministic fluid model is easy to understand. The general idea is to approximate discrete customers by continuous fluid, because we think of there being many servers, a high arrival rate and a large number of customers, so that the discrete nature of individual customers should not be important; the law of large numbers should apply to justify deterministic approximations. We regard the occurrence of interruptions as a random environment, which we leave unchanged. In our approximation, we assume that arrivals, service and abandonment occur deterministically at the specified rates $(\lambda, \mu_1, \theta_1, \mu_2, \theta_2, \theta_3)$, depending on the state of the random environment. Then we obtain a deterministic fluid model in the alternating-renewal-process random environment.

It is significant that the deterministic fluid model, for either an isolated interruption or an alternating-renewal-process random environment, arises as the limit of a sequence of queueing models in a many-server heavy-traffic limit, not only as a direct approximation; see §3. Indeed, the fluid model with interruptions arises in the same many-server limit used to justify the economy of scale. Thus, the problems posed by large scale can be deduced by essentially the same reasoning used to demonstrate the economy of scale, by going further to realistically include service interruptions in the model. The details are given in the e-companion.

2.5. The Rest of the Paper

We start in §3 by reviewing the three many-server heavy-traffic limiting regimes and discussing the mathematical basis for the economy of scale, which leads to increasing server utilizations as scale increases. Then in §4 we consider the pure-delay model in the presence of an isolated service interruption. We show that the performance during an interruption and the following recovery period degrades significantly as the scale increases, when QED scaling is used, as is dictated by the asymptotic analysis discussed in §3. Next in §5 we describe the much milder consequence of a service interruption for the pure-loss model. In §6 we consider the intermediate case with customer abandonment, and show that the system then tends to be more like the pure-loss model, provided the abandonment rates are not too low. In §7 we develop simple approximations for the steady-state performance of the pure-delay fluid model in the alternating-renewal-process random environment, drawing on Kella and Whitt (1992). We see even worse consequences of the interruptions for the pure-delay model with increasing scale when the interruptions can overlap.

We present additional supporting material in the e-companion. In §8 we briefly describe these results. In §EC.1 we analyze the more realistic multiple-interruption model in §2.2. Then in §EC.2 we establish a theorem showing that the approximating deterministic fluid model in a random environment arises as the heavy-traffic limit of properly normalized queueing processes. In §EC.2.2 we contrast our many-server heavy-traffic limit theorem with the earlier single-server analog, established by Kella and Whitt (1990), and extended to networks of queues by Chen and Whitt (1993); see §14.7 of Whitt (2002). There are similarities, but also striking differences. For other work on service interruptions or server vacations, see Zhang and Tian (2003), Altman and Uri (2006) and the references therein.

Finally, in §9 we draw conclusions. We summarize the difficulties uncovered and briefly discuss what might be done to address them. We also discuss remaining research problems.

We establish additional related asymptotic results in another paper, Pang and Whitt (2008). There we establish a many-server heavy-traffic stochastic-process limits for the number in system in the $G/M/n+M$ model, with more general non-Poisson arrival process, in an alternating-renewal-process random environment, with both fluid scaling (dividing by n) and diffusion scaling (dividing by $\sqrt{n}$ after centering). The limit in the QED regime is based on assuming that the down times are asymptotically negligible, of order $O(1/\sqrt{n})$ as $n \to \infty$. Even though these down times are asymptotically negligible, they produce upward jumps in the limit process, showing from another perspective that even short (in fact, asymptotically negligible) service interruptions can have a dramatic impact in the pure-delay system with many servers not functioning during the interruption.

3. The Many-Server Heavy-Traffic Limiting Regimes

In a many-server service system there is the classical *tradeoff* between service quality (captured by low congestion, e.g., delays) and efficiency (low operational cost). Roughly, the *operational cost* for given demand (arrival rate λ) may be taken to be an increasing positive linear function $C_1(n) \equiv a_1 + a_2 n$ of the number of servers, n, while the *congestion cost* may be taken to be the expectation of a function of the steady-state customer delay, which typically is a decreasing positive convex function $C_2(n)$ of n. As a consequence, the overall operational-plus-congestion cost $C(n) \equiv C_1(n) + C_2(n)$ as a function of the staffing level n tends to be a convex function that increases as n becomes either large or small; e.g., see Borst et al. (2004). Hence, there is an overall optimal staffing level $n^* \equiv n^*(\lambda)$ that should be sought.

However, these costs are affected by *scale*, which we can represent simply by either the arrival rate λ or the number of servers n. The commonly accepted way to represent increasing scale (without considering service interruptions) is to consider a sequence of many-server queueing models indexed by the number of servers, n, and let $n \to \infty$. In doing so, it is understood that the associated sequence of arrival rates λ_n should increase as n increases.

To review that framework, we consider a sequence of Markovian $M/M/n+M$ models, as in Garnett et al. (2002). (That is our model without the interruptions.) We assume that the service-time and abandonment-time distributions remain unchanged with n, being exponential with rates μ and θ, respectively. Of course, the increasing number of servers should be in response to increasing demand, as quantified by the customer arrival rate. We assume that the arrival rate in model n is λ_n, where

$$\frac{\lambda_n}{n} \to \bar{\lambda} \quad \text{as} \quad n \to \infty \quad \text{where} \quad 0 < \bar{\lambda} < \infty. \tag{1}$$

The important point is that we need to pay careful attention to the way λ_n grows, being more precise than in (1). It is now well known (Halfin and Whitt 1981, Garnett et al. 2002, Borst et al. 2004, Zeltyn and Mandelbaum 2005) that there are three different limiting regimes in this many-server heavy-traffic setting, depending on what we assume for the traffic intensities $\rho_n \equiv \lambda_n/n\mu$. To define these regimes, we impose a further regularity condition: We assume that

$$(1-\rho_n)\sqrt{n} \to \beta \quad \text{as} \quad n \to \infty \quad \text{where} \quad -\infty \le \beta \le \infty. \tag{2}$$

Without customer abandonment, we obtain the *quality-driven* (QD) regime, the *quality-and-efficiency-driven* (QED) regime (or Halfin-Whitt regime), and the *efficiency-driven* (ED), respectively, when $\beta = +\infty$, $0 < \beta < +\infty$ and $\beta = 0$. With customer abandonment, the QD regime is the same, but the QED and ED regimes change; the QED regime has $-\infty < \beta < +\infty$, while the ED regimes has $\beta = -\infty$.

Without customer abandonment, we need to require that $\rho_n < 1$ for all n to have stability (a proper steady state), which leads to $\beta \ge 0$. However, customer abandonment ($\theta > 0$) keeps the system stable for all $\rho < \infty$.

From a practical perspective, the ED regime with customer abandonment and the QD regime generally can correspond to fixed traffic intensities, with $\rho_n = \rho > 1$ for all n with ED, and $\rho_n = \rho < 1$ for all n with ED. That is achieved by having $\lambda_n = \bar{\lambda} n$ for all n, where $\bar{\lambda} > \mu$ for ED, and $\bar{\lambda} < \mu$ for QD. Henceforth we assume fixed traffic intensities for the ED and QD regimes when we consider them. That makes $\beta = -\infty$ in ED and $\beta = \infty$ in QD.

The importance of the intermediate QED regime is highlighted by the QED-delay-probability theorem: Halfin and Whitt (1981) showed that the steady-state delay probability in the $M/M/n$ model (without customer abandonment) converges to a non-trivial limit, strictly between 0 and 1, in the many-server heavy-traffic setting if and only if the system is in the QED regime. The same result holds for the many-server queue with abandonments; see Garnett et al. (2002) and Zeltyn and Mandelbaum (2005). Thus, everything else being equal (e.g., disregarding the possibility of uncertain arrival rates or service interruptions), the QED-delay-probability theorem implies that a many-server service system should be staffed to be in the QED regime.

We will focus on the more restrictive definition of the QED regime, in which (2) holds with $0 < \beta < \infty$, which applies both with and without abandonment. Finite positive β in (2) implies the celebrated *square-root-staffing* (SRS) *rule*: Starting with any given total arrival rate λ^*, which we let grow, the required number of servers is

$$n \equiv n(\lambda^*) \equiv \frac{\lambda^*}{\mu} + \beta\sqrt{\frac{\lambda^*}{\mu}} + o\,(\lambda^*) \quad \text{as} \quad \lambda^* \to \infty, \tag{3}$$

where $o(t)$ as $t \to \infty$ means a function $h(t)$ such that $h(t)/t \to 0$ as $t \to \infty$; see Proposition 2.1 of Whitt (1992). Fundamentally, the SRS staffing in (3) is a consequence of the central limit theorem.

To appreciate the consequences of being in each of the three regimes, it is helpful to see how the standard steady-state performance measures scale in each regime. The results are summarized in Table 1; see Garnett et al. (2002) and Zeltyn and Mandelbaum (2005). In Table 1, $f(n) \sim g(n)$ as $n \to \infty$ means that $f(n)/g(n) \to 1$ as $n \to \infty$.

regime	Quality-Driven (QD)	Quality&Efficiency-Driven (QED)	Efficiency-Driven (ED)
perf. meas.			
staffing	$n = (\lambda/\mu)(1+\epsilon)$	$n = (\lambda/\mu) + \beta\sqrt{\lambda/\mu}$	$n = (\lambda/\mu)(1-\epsilon)$
waiting prob.	$P(W > 0) \sim \frac{c_1 e^{-c_2 n}}{\sqrt{n}}$	$P(W > 0) \to \alpha \equiv \alpha(\beta, \mu/\theta)$	$P(W = 0) \sim \frac{c_3 e^{-c_4 n}}{\sqrt{n}}$
aban. prob.	$P(Ab \mid W > 0) \sim \frac{c_5}{n}$	$P(Ab \mid W > 0) \sim \frac{\xi}{\sqrt{n}} \equiv \frac{\xi(\beta,\mu/\theta)}{\sqrt{n}}$	$P(Ab) \to \frac{\rho-1}{\rho} > 0$
exp. wait	$E[W \mid W > 0] \sim \frac{c_5}{\theta n}$	$E[W \mid W > 0] \sim \frac{\xi}{\theta\sqrt{n}}$	$E[W \mid \text{served}] \to w^* > 0$

Table 1 How key performance measures scale in the three many-server heavy-traffic limiting regimes, for the model including customer abandonment. The quantities $\alpha, \xi, c_1, \ldots, c_5$ are positive functions of the model parameters (independent of n) with explicit expressions; $\rho \equiv \lambda/n\mu$ is the traffic intensity. The remaining parameter β specifies the quality of service, as in (3).

Note that performance is impressively (even excessively) good in the QD regime for large n. The probability of having to wait before starting service is not only converging to 0 as n increases, but is becoming exponentially small. Moreover, even in that rare event that a customer must wait before starting service, the conditional expected waiting time and the conditional abandonment probability, given that a customer must wait, are both asymptotically negligible, being of order $1/n$.

In fact, the performance can be remarkably good in the QED regime. In the QED regime, the probability of having to wait approaches a constant α, with $0 < \alpha < 1$, which can be made as small as we wish by choosing the parameter β in (2) suitably large. Moreover, just as in the QD regime, the conditional expected waiting time and the conditional abandonment probability, given that a customer must wait, are both asymptotically negligible, being of order $1/\sqrt{n}$. As a consequence, the performance might even be judged to be excessively good in the QED regime.

Depending on the criterion, asymptotically as $n \to \infty$, the optimal staffing with customer abandonment might actually put the system in the ED regime, as studied in Whitt (2004). Table 1 shows that the performance can be good in the ED regime (with customer abandonment), provided that ρ is not much greater than 1 and the abandonment rate is not too low. We can set the staffing level so that ρ is only slightly larger than 1, so that the abandonment rate $(\rho - 1)/\rho$ and the common fixed waiting time w^* are both suitably small. Indeed, even though all customers must wait before starting service in the ED limit, it is possible for only a small proportion to abandon and the waiting times of all served customers to be short.

In summary, the asymptotic behavior in the many-server heavy-traffic regimes provides strong motivation for having increasing scale and operating a large service system in the QED regime or even the ED regime. When we consider interruptions, we will do so with the QED scaling in mind; e.g., see Theorem 1.

4. An Isolated Interruption in the Pure-Delay System

To appreciate the possible problems caused by system-wide service interruptions with large scale, it is revealing to focus on the pure-delay fluid model in the presence of a single isolated interruption. Clearly, we will see essentially the same behavior in the general model when the quantities

$F_k, \theta_1, \theta_2, \theta_3$ are sufficiently small. Here we consider a single interruption and let $F = \theta_1 = \theta_2 = \theta_3 = 0$. We also let $\mu_2 = \mu$. (Since $F = 0$, μ_2 plays no role here.) Since there is no customer abandonment in the pure-delay model, we will also assume that $\lambda < n\mu$ to ensure that the model without interruptions is stable.

When there is no interruption, the total fluid content $X(t)$ evolves according to the nonlinear ODE

$$\dot{X}(t) = \psi_1(X(t)) \equiv \lambda - (X(t) \wedge n)\mu, \quad t \geq 0, \tag{4}$$

where $a \wedge b \equiv \min\{a, b\}$. Since we have assumed that $\lambda < n\mu$, we will have $X(t) < n$ for all $t > 0$ if $X(0) \leq n$. In that case, X will evolve according to the linear ODE

$$\dot{X}(t) = \lambda - \mu X(t), \quad t \geq 0. \tag{5}$$

We solve this linear ODE and others later in this paper by applying the following elementary (well known) lemma:

Lemma 1. *The first-order linear ordinary differential equation* (*ODE*)

$$\dot{x}(t) = a - bx(t), \quad t \geq 0,$$

with initial value $x(0)$, *where* a *and* b *are real numbers, has the unique solution*

$$\left(x(0) - \frac{a}{b}\right) e^{-bt} + \frac{a}{b}, \quad t \geq 0.$$

When $X(0) \leq n$, we can apply Lemma 1 to obtain

$$X(t) = \left(X(0) - \frac{\lambda}{\mu}\right) e^{-\mu t} + \frac{\lambda}{\mu}, \quad t \geq 0, \tag{6}$$

On the other hand, if $X(0) > n$, then we will have $X(\tau) = n$ for some τ and $X(t) < n$ for all $t > \tau$, so that

$$X(t) = \left(n - \frac{\lambda}{\mu}\right) e^{-\mu(t-\tau)} + \frac{\lambda}{\mu}, \quad t \geq \tau, \tag{7}$$

In either case, the steady-state limiting value is λ/μ.

During a service interruption of length D after the system has been in steady state at $\rho n = \lambda/\mu$, $X(t)$ simply increases at rate λ until it reaches the value $X(D) = \rho n + \lambda D$. After the single interruption, $X(t)$ again evolves as the ODE (4). Assuming that ρ is close to 1 and the down time D is not too small, we will have $X(D) > n$. As long as $X(t) > n$, $X(t)$ evolves as the linear ODE

$$\dot{X}(t) = \lambda - n\mu, \quad t \geq 0, \tag{8}$$

so that $X(\bar{\tau}) = n$ for $\bar{\tau} \equiv (X(D) - n)/(n\mu - \lambda)$. After time $\bar{\tau}$, $X(t) \leq n$ and $X(t)$ evolves according to the linear ODE (5), so that

$$X(t) = \left(n - \frac{\lambda}{\mu}\right) e^{-\mu(t-\bar{\tau})} + \frac{\lambda}{\mu}, \quad t > \bar{\tau}, \tag{9}$$

which implies that $X(t)$ gradually approaches its steady-state value $\rho n = \lambda/\mu$, but never quite reaches it in finite time.

In addition to the processes $X(t)$ and $Q(t) \equiv (X(t) - n)^+$, where $(a)^+ \equiv \max\{a, 0\}$, describing the number in system and the queue length, we also want to describe the (virtual) waiting-time process $\{W(t) : t \geq 0\}$; $W(t)$ depicts the time a potential arrival at time t would have to wait *before starting service*. When $X(t) < n$ during an up period U, $W(t) = 0$. At the beginning of a down period of length D, $W(t) = D$. In the middle of that down period, at time t after the down period of length D began, the virtual waiting time is

$$W(t) = (D - t) + Q(t)/\mu n. \tag{10}$$

In order to obtain simple revealing formulas, it is convenient to make further approximations. In order to have a stable system without interruptions, we needed to assume that $\rho \equiv \lambda/n\mu < 1$. However, since we are thinking of large scale, ρ should be close to 1. Hence, we propose simplifying the behavior when $\rho n \leq X(t) \leq n$ by assuming that $X(t)$ decreases linearly throughout the region $X(t) \geq \rho n$. That makes $X(t)$, not only increase linearly during the down period, but also decrease linearly during the following up period, until time $\hat{\tau} \equiv (X(D) - \rho n)/(n\mu - \lambda)$. At time $\hat{\tau}$, $X(t)$ reaches the steady-state value ρn and stays there afterwards. We call that period of linear decrease immediately after the down time the *recovery time* and denote it by the random variable R; R is a deterministic function of the random variable D. (Note that R would be infinite without this last approximation step.) We depict the processes $X(t)$ and $W(t)$ during a typical down time D and the following recovery time in Figure 1.

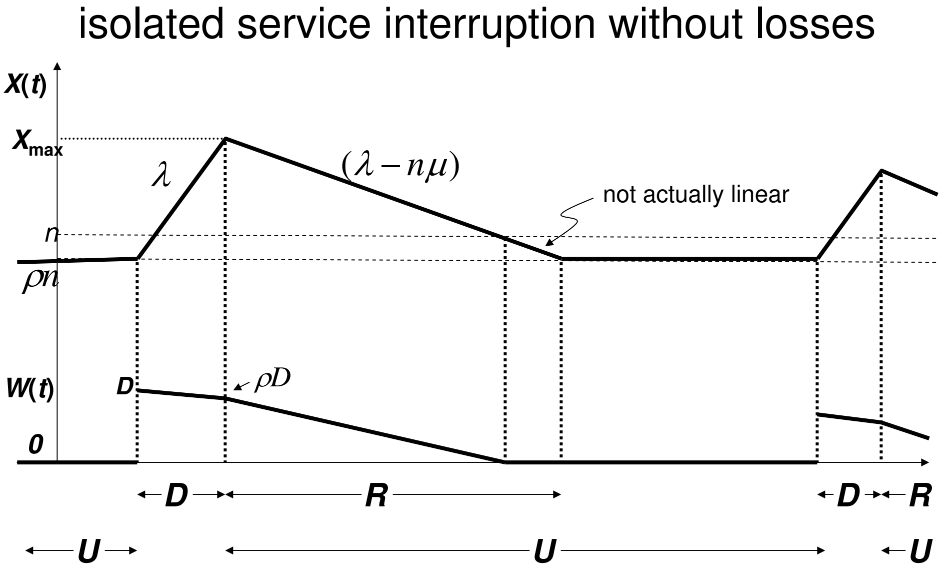

Figure 1 Sample paths of the approximating number in system, X(t), and the associated waiting-time process, W(t), in the pure-delay system experiencing an isolated exogenous service interruption.

Using the same simplification for $Q(t)$ during an interruption, we can approximate the waiting time during an interruption by

$$W(t) \approx D - t + \frac{\lambda t}{\mu n} = D - t + \rho t = D - (1-\rho)t, \tag{11}$$

which decreases linearly to ρD at the end of the down period. During the recovery period, the waiting time is approximately $W(t) \approx Q(t)/\mu n$, which decreases linearly to 0 at the time the queue empties, which is shortly before the end of the recovery period.

Now we want to make another simplifying assumption, again exploiting the fact that the traffic intensity ρ should be close to 1 because of the large scale. In our formulas, we will ignore the region where $\rho n \leq X(t) \leq n$, and act as if the two horizontal dotted lines at n and ρn in Figure 1 coincide. But we leave the rate of linear decrease during the recovery period unchanged. It should be evident that this additional approximation will produce only a minor change in the formulas, but the formulas become more transparent.

performance measure	notation	delay model	loss model
CAPACITY REQUIREMENTS			
maximum queue length	Q_{max}	λD	0
DURATIONS PER INCIDENT			
recovery time	R	$\frac{\rho D}{1-\rho}$	$\approx 4\mu^{-1}$
incident duration	$D+R$	$\frac{D}{1-\rho}$	$D+4\mu^{-1}$
maximum waiting time	W_{max}	D	0
average waiting time	$\bar{W}$	$D/2$	0
TOTAL IMPACT PER INCIDENT			
number of affected arrivals:	N_{tot}		
delayed in the pure-delay model	$N_{tot}^{d} \equiv \lambda(D+R)$	$\frac{\lambda D}{1-\rho}$	0
lost in the pure-loss model	$N_{tot}^{l} \equiv \lambda D$	0	λD
total waiting time	$W_{tot} \equiv N_{tot}\bar{W}$	$\frac{\lambda D^2}{2(1-\rho)}$	0

Table 2 **Approximate performance measures in the pure-delay and pure-loss models associated with an isolated service interruption of random duration D.**

Table 2 summarizes the main performance measures of interest for the pure delay model experiencing an isolated service interruption, exploiting the two approximations above. (Corresponding results for the pure-loss model, discussed in the next section, also appear there.) These performance measures are functions of the down time D and are thus themselves random variables. For example, $\bar{W} \equiv \bar{W}(D)$ represents the average waiting time (before starting service for new arrivals) during the incident, i.e., over the random time interval $D+R$. We make a simplifying assumption here. From Figure 1, it is evident that $\bar{W} = (D(1+\rho-\rho^2))/2 \approx D/2$, assuming that ρ is close to 1.

Particularly significant is the total number of new arrivals delayed, say N_{tot}^{d}, and the total waiting time of all arrivals during the incident (during $D+R$), W_{tot}. Clearly,

$$N^d_{tot} = \lambda(D+R) = \frac{\lambda D}{1-\rho} \quad \text{and} \quad W_{total} = N_{tot}\bar{W} = \frac{\lambda D^2}{1-\rho}. \tag{12}$$

Note that N^d_{tot} counts only new arrivals. It does not count the extra delays experienced by the ρn customers in service at the beginning of the interruption.

The corresponding expected total waiting time per incident (with down time D) is

$$E[W_{total}] = \frac{\lambda(E[D])^2(c_d^2+1)}{1-\rho}, \tag{13}$$

which is directly proportional to the *product* of λ, $1/(1-\rho)$, $(E[D])^2$ and c_d^2+1, where c_d^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of the down time. The SCV c_d^2 shows the importance of the variability of the down time beyond its mean.

At this point, it is insightful to consider the QED many-server heavy-traffic scaling, specified by (2) with $0 < \beta < \infty$. (Indeed, combining Table 2 with the QED scaling is the main innovation in this paper.) As a consequence, $\lambda/\mu n \to 1$ as $n \to \infty$, so that λ is proportional to n while $(1-\rho)^{-1}$ is proportional to $\sqrt{n}$ as $n \to \infty$.

Theorem 1. (QED scaling of the performance measures) *Consider the performance measures for the pure-delay fluid model experiencing one isolated service interruption of random duration D, starting in steady state, as depicted in Table* 2*. If the scale is increased with QED scaling, while the duration of the down time D remaining unchanged, then the performance degrades as follows:*

(*a*) *The expected recovery time for one isolated incident, $E[R]$, and the expected duration of one isolated incident, $E[D+R]$, are inversely proportional to $1-\rho$, and thus are asymptotically proportional to $\sqrt{n}$ as $n \to \infty$.*

(*b*) *The expected number of customers delayed per incident, $E[N^d_{tot}]$, and the expected total waiting time per incident, $E[W_{tot}]$, are proportional to the product of n and $1/(1-\rho)$, and so are proportional to $n^{3/2}$ as $n \to \infty$.*

The conclusions in Theorem 1 remain unchanged if we also consider the customers initially in service when the interruption first occurs. That number ρn is asymptotically negligible compared to $E[N^d_{tot}]$, which is of order $n^{3/2}$.

Theorem 1 also has important implications for multiple interruptions occurring over time according to the random environment process $\{(U_k, D_k) : k \geq 1\}$. We treat multiple interruptions in §7.

There is even a significant consequence if the down times become asymptotically negligible as n increases. Given that the maximum waiting time for any one customer is $W_{max} = D$, the maximum waiting time for any one customer also becomes negligible if D becomes negligible, but the cumulative impact over all customers can remain significant. If the down time D decreases with n more slowly than by $1/\sqrt{n}$, then all those other performance measures will still diverge to infinity as $n \to \infty$. Even if D is of order $1/\sqrt{n}$, there may be a significant impact; see Pang and Whitt (2008) for more on this.

5. The Pure-Loss System

We see very different performance if we assume that all customers in service are lost when a service interruption occurs, and arrivals do not join the system when the system is down. Then the number in system $X(t)$ is always *below* its steady-state value ρn instead of always above. By those assumptions, the pure-loss system completely empties at the beginning of each down time and does not start receiving new arrivals until the down time is over. Thus, the transient behavior after a down period (in each up period after the first down period) is governed by the linear ODE in (5) with initial condition $X(0) = 0$ at the end of the down period, which implies that

$$X(t) = \rho n \left(1 - e^{-\mu t}\right), \quad t \geq 0, \tag{14}$$

where $t = 0$ is the end of the down time, with validity of (14) holding until the next down time.

Figure 2 shows the sample path of the number in system in the pure-loss system, with the same up and down times as for the pure-delay system in Figure 1. The only difference is that customers in service and new arrivals are lost during each down time. After the down time is over, the system needs to ramp up to its steady-state level, but it never quite reaches steady state when the next interruption occurs. Clearly, though, the system regenerates at the beginning of each down time. There can be no overlapping of service interruptions. We give a corresponding figure for the more general model with abandonments in the e-companion.

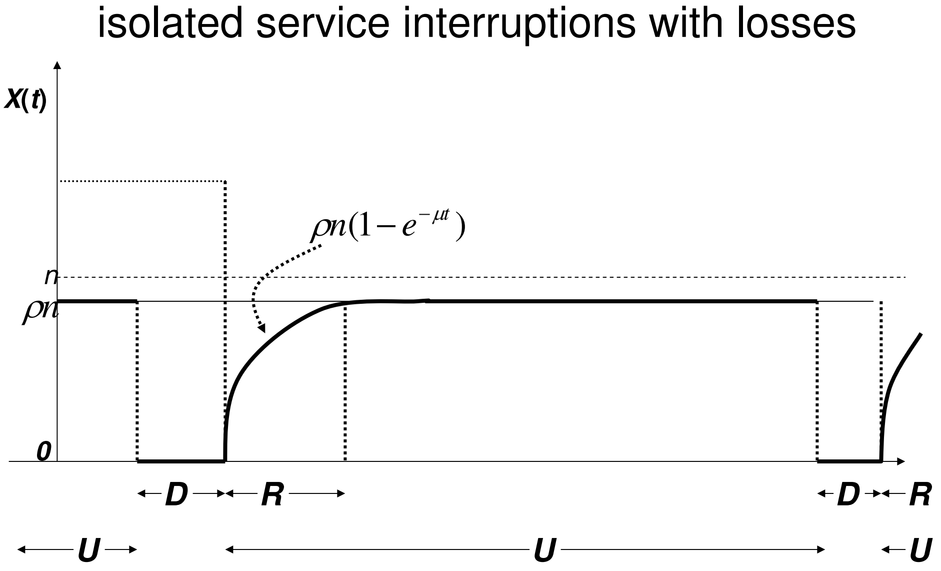

Figure 2 **The number in system, $X(t)$, in the pure-loss fluid model with alternating up and down times distributed as U and D.**

It is significant that *the recovery time in the loss model is much less than in the delay model*, provided that we use good engineering judgment in what we mean by recovery. In the loss model, recovery occurs according to the ODE in (5). From that ODE, we see that the system never completely recovers in finite time, but the practical recovery time (often called the relaxation time) is some multiple of $1/\mu$. In time t/μ, the system has reached a proportion $1-e^{-t}$ of its steady state value. As a rough estimate we take $4/\mu$ as our definition of the recover time. The loss model has reached a proportion $1-e^{-4}=0.982$ of its steady-state value by time $4/\mu$. Clearly this recovery time does not change with increasing scale.

The primary performance issue with this loss model is the lost throughput. For that, we should consider multiple interruptions, i.e., the pure-loss model in the random environment $\{(U_k,D_k):k\geq 1\}$. The long-run proportion of time during which customers are turned away is clearly the long-run proportion of time that the system is down, $\delta=E[D]/(E[D]+E[U])$. However, the proportion of customers lost is greater, because all ρn customers in service at the beginning of the down period are lost as well. The loss probability is

$$P(loss) = \frac{\rho n + \lambda E[D]}{\lambda(E[D+U])} = \delta\left(1 + \frac{1}{\mu E[D]}\right).$$

6. Customer Abandonment

In practice, service systems are rarely pure-delay systems or pure loss systems. Typically, customers will abandon after waiting too long. Clearly, high customer abandonment usually is a sign of poor service, and so is to be avoided. However, even a low level of customer abandonment tends to dramatically shorten the recovery period after a service interruption, making the performance more like in the pure-loss system. Of course, as we observed before, we can obtain a continuum of solutions between the pure-delay model and the pure-loss model by choosing appropriate abandonment parameters. Thus, this conclusion necessarily depends on the abandonment parameters we use.

In this section we will show that typical abandonment rates tend to greatly reduce the impact of the interruptions. Just as in §§4 and 5, we consider a single service interruption in isolation and we assume that all servers cease functioning during the interruption ($F = 0$). Moreover, to highlight the main features with simple formulas, we assume that the many-server system is critically loaded: $\rho \equiv \lambda/\mu n = 1$. As noted by Feldmann et al. (2008), the case $\rho = 1$ ($\beta = 0$ in (2)) should be regarded as a common case with customer abandonments. It also can be regarded as a worst case: If we are able to show that an interruption has limited impact with $\rho = 1$, then it follows that it will have limited impact for all $\rho \leq 1$ (by a sample-path comparison).

The critically loaded assumption implies that the initial steady-state fluid approximation, before the interruption, is $X(0) = n$; i.e., all servers are busy, but there is no queue. Let the interruption begin at time 0. Then the queue length $Q(t) \equiv (X(t) - n)^+$ evolves as the linear ODE $\dot{Q}(t) = \lambda - \theta_2 Q(t)$, $t \geq 0$, with $Q(0) = 0$, just as in the fluid approximation for the queue length in the $M/M/\infty$ queue with arrival rate λ and service rate θ_2 (the abandonment rate from the queue during the interruption). Thus, we obtain the explicit formula

$$Q(t) = \frac{\lambda}{\theta_2}\left(1 - e^{-\theta_2 t}\right), \quad 0 \leq t \leq D, \tag{15}$$

where D is the random down time. From (15), we see that $Q(t)$ increases toward the finite limit λ/θ_2 as $t \to \infty$. Significantly, that limit is independent of the random down time D. Of course, this limiting queue length can be large if θ_2 is small, but otherwise it is controlled. Roughly, the maximum queue length is of the same order $O(n)$ as in the pure-delay model, where it is λD, because $\lambda = \mu n$.

There also are the n customers in service at the start of the interruption, who abandon at rate θ_3. The number of busy servers at time t satisfies the linear ODE $\dot{B}(t) = -\theta_3 B(t)$, $t \geq 0$, so that

$$B(t) = n e^{-\theta_3 t}, \quad 0 \leq t \leq D. \tag{16}$$

In the worst case, $\theta_3 = 0$, so that all these customers remain in the system. Then the number of customers in the system at the end of the interruption is

$$X(D) = n + \frac{\lambda}{\theta_2}\left(1 - e^{-\theta_2 D}\right) \leq n + \frac{\lambda}{\theta_2} = n\left(1 + \frac{\mu}{\theta_2}\right). \tag{17}$$

We are now ready to consider the following recovery period after time D. Here the situation is very different from the pure-delay model. Since $\rho = 1$, the fluid approximation is $X(t) = n + Q(t)$, where $\dot{Q}(t) = -\theta_1 Q(t)$, $t \geq D$, and $Q(D) \leq \lambda/\theta_2$. Thus, $Q(t) = Q(D)e^{-\theta_1(t-D)}$, $t \geq D$. In other words, $Q(t)$ decreases exponentially fast toward 0 at rate θ_1. As a consequence, the duration of the recovery time is approximately $R \equiv R(D) \approx 4/\theta_1$, where the constant 4 is chosen just as in §5 to be within 2% (of the initial queue length) from the target. The main point is that the recovery period is of order O(1) as $n \to \infty$ with critical loading. Indeed, if θ_1 is not too small, then the recovery period is quite short.

The situation is even more favorable if θ_3 is not negligible. A key reference point is $\theta_3 = \mu_1 = \mu_2$. Then the system output is unchanged during the interruption. Then we have abandonment during the interruption at the same rate as service is performed before the interruption. Then we get $X(t) = n$ for all t, independent of the interruption. Then $R = 0$. More generally, we will have $X(t) \geq n$ for all t if $\theta_3 \leq \mu_2 = \mu_1$, but we will have $X(t) \leq n$ for all t if $\theta_3 \geq \mu_2 = \mu_1$.

In summary, we see that the recovery period does not explode as $n \to \infty$ when there is waiting with customer abandonment. The performance problems in §4 stem from the fact that all customers remain waiting until they can be served in the pure-delay system. However, so far we have only considered a single service interruption in isolation. We consider multiple interruptions next.

7. Steady State for the Pure-Delay Model with Multiple Interruptions

We now turn to the more general pure-delay model with multiple interruptions. Now we want to understand the further degradation in performance caused by having new interruptions occur before the system has recovered from the previous interruption.

Fortunately, it is not too difficult to develop a good approximation for the steady-state behavior of the fluid content process $\{X(t) : t \geq 0\}$ for the pure-delay fluid model in the alternating-renewal-process random environment. To obtain a simple approximation, we again make the simplifying assumptions in §4, acting as if the process decreases linearly in the up periods, with a reflecting lower barrier at ρn, where $\rho n \approx n$. (Then the content process coincides with $\rho n + Q(t)$.) For the resulting model with linear decrease during the recovery period, a full analysis has already been done in Kella and Whitt (1992) and the other related references cited there. There it is shown that the full process $X(t)$ and various embedded processes of interest can be analyzed by relating this simplified model to the classical $GI/GI/1$ queue.

If, in addition, we assume that the up times have an exponential distribution, which is very reasonable if we think of the common case in which $U >> D$ and exogenous service interruptions occur according to a Poisson process, then the model is related to the more elementary $M/GI/1$ queue, so that we obtain simple explicit steady-state formulas. In order to understand the impact of overlapping service interruptions, it is helpful to look at these formulas.

In addition to the nominal server utilization $\rho \equiv \lambda/\mu$ before considering the interruptions, we have the *fluid-model traffic intensity* within each up-down cycle,

$$\hat{\rho} \equiv \frac{\lambda E[D]}{(n\mu - \lambda)E[U]} = \frac{\rho E[D]}{(1-\rho)E[U]} = \left(\frac{\rho}{1-\rho}\right)\left(\frac{\delta}{1-\delta}\right), \tag{18}$$

where $\delta \equiv E[D]/(E[D]+E[U])$ is the long-run proportion of down time. Theorem 3 of Kella and Whitt (1992) concludes that steady state exists for this fluid model in a two-state random environment if and only if $\hat{\rho} < 1$. From (18), we see that condition holds if and only if $\delta < 1-\rho$ or, equivalently, $\rho < 1-\delta$. For given δ, this places an upper bound on ρ.

Theorem 2. (QED scaling with repeated interruptions) *For any given long-run proportion of down time $\delta > 0$, the system is unstable if $1-\rho < \delta$, in which case $X(t) \Rightarrow \infty$ as $t \to \infty$. If the scale is increased with QED scaling, with fixed random environment $\{(U_k, D_k, F_k) : k \geq 1\}$, then the system becomes unstable when n is large enough.*

The analysis based on Kella and Whitt (1992) exploits the linear decrease during the recovery period, but recall that in §4 we have exploited an approximation when $X(t)$ falls in the interval $[\rho n, n]$. Since the approximation we use produces a lower bound on $X(t)$, elementary comparison arguments show that the conclusion of Theorem 2 remains valid for the exact fluid model too. Evidently, the conclusion holds for the stochastic queueing model as well for all sufficiently large n, by virtue of the heavy-traffic limit theorem in §EC.2, but there is a limit-interchange argument that we have not yet performed.

We can also do more to describe what happens for lower traffic intensities. To do so, henceforth we assume that $\hat{\rho} < 1$, which guarantees stability. The issue now is to describe the overall steady-state distribution, and calculate associated performance measures. Fortunately, we can perform a more detailed analysis to obtain useful performance measures for the pure-delay model.

From Theorem 3 of Kella and Whitt (1992), we can conclude that

$$X(T_k) - \rho n \Rightarrow Z_e \quad \text{as} \quad k \to \infty, \tag{19}$$

where Z_e has the structure of the steady-state waiting time in the $GI/GI/1$ queue (is a random walk with one reflecting barrier). Since we also assume that U has an exponential distribution, this becomes the $M/GI/1$ queue, so that we obtain the approximations

$$P(Z_e > 0) = \hat{\rho} \quad \text{and} \quad E[Z_e | Z_e > 0] = \frac{\lambda E[D](c_d^2+1)}{2(1-\hat{\rho})}. \tag{20}$$

Thus, in order to justify ignoring overlapping interruptions, we need $\hat{\rho}$ suitably small. As we have stated before, the isolated-interruption assumption only makes sense if δ is very small or, equivalently, if $E[U] >> E[D]$.

The more subtle part of the analysis in Kella and Whitt (1992) is to determine the steady-state distributions of the process $X(t)$ at an arbitrary time t, both unconditioned and conditioned on being in either an up time or a down time. Limits for these require assuming that the distributions of D and U are non-lattice. We conclude this section by stating one such result under that condition:

$$X(t) - \rho n \Rightarrow Z \quad \text{as} \quad t \to \infty, \tag{21}$$

where

$$P(Z > 0) = \delta + (1-\delta)\hat{\rho} = \frac{\delta}{1-\rho} \tag{22}$$

and

$$E[Z|Z > 0] = \frac{\lambda E[D](c_d^2 + 1)}{2(1-\hat{\rho})} > \frac{\lambda E[D](c_d^2 + 1)}{2(1-\rho)} = E[Q(\infty)|Q(\infty) > 0], \tag{23}$$

where we take the formula for $E[Q(\infty)]$ from Table 2. We needed $\delta < 1 - \rho$ in order to have $\hat{\rho} < 1$; that guarantees that $P(Z > 0) < 1$, but $P(Z > 0)$ need not be small. We need δ very small in order to have $P(Z > 0)$ small. That again is achieved if $E[U] >> E[D]$. Formula (23) implies that

$$\frac{E[Z|Z > 0]}{E[Q(\infty)|Q(\infty) > 0]} = \frac{1-\rho}{1-\hat{\rho}} > 1. \tag{24}$$

That ratio approaches 1 as $\delta \downarrow 0$.

8. Generalizing and Formalizing

So far, we have considered only relatively simple models, using a simple direct deterministic fluid approximation. However, we can also generalize and formalize. First, we generalize to the more realistic model in §2.2. Second, we formalize (and further justify) by establishing a many-server heavy-traffic limit. Both steps are carried out in the e-companion; here we discuss the significance.

In EC.1 we consider the more realistic model in §2.2. When we do so, we are no longer able to obtain simple explicit expressions for the relevant performance measures, which are so helpful for providing insight. However, we do give explicit (somewhat complex) expressions for the (deterministic) system content $X(t)$, conditional on a realization of the random environment process $\{(U_k, D_k, F_k) : k \geq 1\}$. That in turn provides the basis for an efficient simulation algorithm to calculate all desired performance measures. It is only necessary to generate an initial segment of the sequence $\{(U_k, D_k, F_k) : k \geq 1\}$ of i.i.d random vectors in $\mathbb{R}^3$ and perform straightforward calculations, using the formulas derived in §EC.1.

From a practical perspective, such more careful analysis might well be very important. The analysis in §§4-6 provides important insight, but it is limited to a single interruption in isolation. We started to consider multiple interruptions over time in §7, but our analysis there was limited to the pure-delay model. In both §4 and §7, we see that there are serious problems with interruptions and increasing scale for the pure-delay model, but many service systems are not pure-delay models. The more detailed analysis with the more general model in 2.2 having both multiple interruptions and customer abandonment may well be needed to better understand specific service systems.

It is important to note that our approach to simulation offers a great advantage compared to direct simulation of the original stochastic model when the interruptions are rare events, which is presumably the common case. (See Chapter VI of Asmussen and Glynn (2007) for background on rare-event simulation.) Our approach replaces the simulation of individual events in the queueing system (arrivals, service completions and abandonments) with deterministic formulas. The stochastic part of our simulation is limited to the interruption process $\{(U_k, D_k, F_k) : k \geq 1\}$, which tends to operate in a much longer time scale. Thus, our approach provides an effective way to perform rare-event simulation in this context, albeit for an approximate model.

Finally, we add theoretical support for our approximating model: the deterministic fluid process in a random environment. In §EC.2 we show that it arises naturally as the many-server heavy-traffic limit for the Markovian queueing systems, which serve as our base system model. The many-server heavy-traffic fluid limit is valid simultaneously in all three regimes: QD, QED and ED. In §EC.2.2 we contrast the many-server-heavy-traffic limit theorem here with the single-server analog established in Kella and Whitt (1990); there are similarities, but also striking differences because of the different scaling. We prove the limit theorem in §EC.2.3.

9. Conclusions

In this paper we have studied the impact of exogenous service interruptions upon multiple-server service systems when the scale (number of servers) increases. The starting point is the widely accepted principle, reviewed in §3, that the quality-efficiency tradeoff supports increasing scale, with $\rho_n \uparrow 1$ as $n \to \infty$. We have used many-server queueing models to study system performance. To obtain tractable results, we have used approximations by deterministic fluid models, and justified them by establishing a many-server heavy-traffic limit; see §EC.2.

We have derived simple descriptions of performance for a single service interruption in isolation when no servers can function during the interruption. We concentrated on two extreme cases - a pure-delay model and a pure-loss model - but we also considered the intermediate case with customer abandonment; see §§4-6. The simple stories are summarized in Figures 1 and 2. *Our main contribution is to look at system performance with service interruptions through the lens of the QED scaling, specified in* (2) *and* (3). That evidently has not been done before.

We showed that the performance degradation increases with scale dramatically with QED scaling for the pure-delay model, but not in the other two cases. The difficulty with system-wide service interruptions and increasing scale seems to be confined primarily to systems where most customers wait until they are served. Clearly, that is not always the case, but it is a common case in some applications. In some cases, as with congestion caused by an accident on a highway, the customers are held captive, having no option to leave. In other cases, such as the DMV example mentioned in §1, there may be a large setup cost in seeking service, which will be wasted and repeated at a later time if the customer elects to leave. Finally, the service may be urgent, as in a hospital emergency room or a technical-support call center, so that the customer does not want to abandon. In those cases, our analysis of interruptions in pure-delay systems is relevant.

The performance problem with service interruptions is serious for pure-delay systems. The quality-efficiency tradeoff provides strong motivation for increasing scale, because as the scale increases, it is possible to meet specified quality-of-service constraints with higher server utilization.

The higher utilizations in turn make the recovery periods longer when there is a service interruption. For any given interruption, the performance degrades as scale increases with QED scaling. Theorems 1 and 2 show the consequences.

However, if customer abandonment is an easy option, then service interruptions may not cause major new problems with increasing scale. Additional analysis is warranted, though, if there are multiple interruptions occurring with some frequency. For such problematic cases, we have provided a basis for doing further analysis with the model introduced in §2.2. The formulas in §EC.1 provide the basis for an efficient simulation algorithm.

9.1. What Can Be Done About It?

Having seen the severe service degradation that can occur from system-wide service interruptions in large service systems where (i) most customers wait until they start service and (ii) management aims to achieve both high quality of service and high efficiency by operating in the QED regime, it is natural to ask: *What can be done about it?*

One possible answer is to avoid large scale, but as discussed in Mitchell (2001), large scale has many advantages, going well beyond what is described by queueing models. Thus, abandoning large scale is unlikely to be acceptable. Thus, we would rephrase the question as: *Given that there will be increasing scale, what can be done to mitigate the performance problems that might arise from service interruptions?*

Our analysis suggests several possible answers:

(i) The first suggestion is to *plan for failures*, by which we simply mean anticipate that any service system is likely to encounter service interruptions, and then prepare for them. We have shown that the performance impact of interruptions can be more significant with increasing scale. Thus the importance may be even greater now and in the future than before.

(ii) The second suggestion is to *operate the service system in the QD regime* instead of the QED regime when the scale gets sufficiently large. Section 4 shows that the performance degradation is significantly reduced if we do not try to extract the last bit of efficiency, and instead keep ρ

bounded away from the critical value 1. Even though we are in the QD regime, we may still have very high server utilizations, e.g., above 95%, so that not much efficiency need be sacrificed.

(iii) The third suggestion is to *take actions to encourage waiting customers to leave and potential new arrivals not to join the system* when an interruption occurs, so that the system will behave more like the loss model than the delay model. Of course, the cost of lost service must be assessed in each application. Sections 4–6 show that the impact upon other customers is much less when customers may leave before receiving service. Customers might be encouraged to not arrive or abandon after entrance by making delay announcements, as studied in Armony et al. (2008) and Ibrahim and Whitt (2008). Indeed, such announcements are already being made in many settings, e.g., the posted warnings about congestion incidents on highways.

(iv) The fourth suggestion is to *take extra measures in order to reduce the frequency of service interruptions*; i.e., reduce δ as defined in §7. We have provided formulas that quantify the benefit, such as (18)-(23). But note that the performance during any one interruption may still be bad.

(v) The fifth suggestion is related to the second, but different: The idea is to *take measures to reduce both the mean and the variance of the down times.* Again, our results can be used to quantify the benefit, e.g., see Table 2 and (13). This suggestion acknowledges that service interruptions are not completely avoidable, and thus suggests acting to reduce their consequence.

(vi) The sixth suggestion amplifies the third in planning to cope with the interruptions when they occur. The idea is to do things differently, so that the predicted bad behavior can be avoided. For example, we suggest *taking measures to have assistance available when a service interruption occurs.* For example, a provision may be made to acquire assistance from other operating service facilities, perhaps using sharing mechanisms as discussed in Perry and Whitt (2008). Since other service facilities are likely to be heavily loaded as well, it would be even better to have access to unused service capacity (e.g., agents in a call center) on short notice, as studied by Bhandari et al. (2008). For example, for call centers, agents at home might be activated upon short notice.

More generally, this means to *seek greater flexibility and adaptability to respond to exceptional circumstances*. We have shown that the performance impact of interruptions may provide strong motivation for such actions.

(vii) The seventh and final suggestion (going beyond our paper), from a broad societal point of view, is to *broadly examine, and possibly redesign, the incentive system* in the service system. The idea is to carefully consider who bears the costs of a service interruption. Given that service interruptions may indeed have serious consequences, system managers are likely to be more motivated to properly manage (prevent, mitigate, etc.) service interruptions if they themselves bear a significant portion of the costs, even if indirectly. The most effective measure may be to ensure that reasonable incentive systems are in place.

9.2. Remaining Problems

Many open problems remain. Perhaps the most compelling direction is to follow up with empirical research. The theory presented here can be viewed as a collection of hypotheses that can be tested empirically. Here are a few candidate hypotheses:

H1: The scale of service systems is increasing over time. And at what rate?

H2: As the scale of service systems increases (given H1), they tend to operate in the QED regime. Or do they, instead, operate in the QD regime, or somewhere in between, perhaps to hedge against risks, reflecting an understanding of the issues we have modelled?

H3: The impact of exogenous service interruptions (as measured by frequency, duration, recovery time, or total waiting time per incident, etc.) is increasing over time. And at what rate?

Much additional work remains to be done along the lines of this paper. For example, it remains to consider service interruptions in more complex service systems, such as multi-class multi-pool call centers with skill-based routing. For these systems, Wallace and Whitt (2005) and Gurvich and Whitt (2007) showed that it suffices to have only minimal cross-training, with each pool only able to serve only two classes (with appropriate network connectivity via chaining); i.e., “a little flexibility goes a long way." (Those results are consistent with earlier work by Jordan and Graves (1995) and others, and more recent work by Bassamboo et al. (2008) and others, in the traditional manufacturing setting.) However, it is evident that service interruptions for some pools can break the connectivity, making it desirable to have additional cross-training. It remains to investigate the impact of service interruptions and model-parameter uncertainty on the benefits of flexibility. The analysis here leads us to anticipate that the need for greater flexibility will make it worthwhile to sacrifice some efficiency in operations.

Acknowledgments. This research was supported by NSF grant DMI-0457095.

References

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: a multi-disciplinary perspective of operations management research. *Production and Operations Management* **16** (6) 665-688.

Altman, E., Y. Uri. 2006. Analysis of customer impatience in queues with server vacations. *Queueing Systems* 52, 261–279.

Armony, M., N. Shimkin, W. Whitt. 2008. The impact of delay announcements upon many-server queues with abandonment. *Oper. Res.*, forthcoming. Available at: http://www.columbia.edu/∼ww2040

Asmussen, S., P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*, Springer, New York.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006a. Design and control of a large call center: asymptotic analysis of an LP-based method. *Operations Research* **54** 419-435.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006b. Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scale fluid limits. *Queueing Systems* **51** 249–285.

Bassamboo, A., R. S. Randhawa, J. A. Van Mieghem. 2008. A little flexibility is all you need: optimality of tailored chaining and pairing. working paper, Northwestern University.

Bhandari, A., A. Scheller-Wolf, M. Harchol-Balter. 2008. An exact and efficient algorithm for the constrained dynamic operatyor staffing problem for call centers. *Management Sci.* **54** (2) 339–353.

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call center. *Oper. Res.* **52** 17–34.

Brockmeyer, E, H. L. Halstrom, A. Jensen. 1948. *The Life and Works of A. K. Erlang*, (eds.), Danish Academy of Technical Sciences, Copenhagen.

Chen, H., W. Whitt. 1993. Diffusion approximations for open queueing networks with service interruptions. *Queueing Systems* **13** 335–359.

Chen, H., D. D. Yao. 1992. A fluid model for system with random disruptions. *Oper. Res.* **40** 239–247.

Choudhury, G. L., A. Mandelbaum, M. I. Reiman, W. Whitt. 1997. Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models* **13** 121–146.

Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* 54 (2) 324–338.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5**(2), 79–141.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 208–227.

Gurvich, I., W. Whitt. 2007. Service-level differentiation in many-server service systems: a solution based on fixed-queue-ratio routing. working paper, Columbia University. Available at: http://www.columbia.edu/∼ww2040

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Hall, R. W. 1991. *Queueing Methods for Services and Manufacturing*, Prentice Hall, Englewood Cliffs, NJ.

Ibrahim, R., W. Whitt. 2008. Real-time delay estimation in overloaded many-server queues with abandonment. Working paper. Available at: http://www.columbia.edu/∼ww2040

Jordan, W. C., S. C. Graves. 1995. Principles on the benefits of manufacturing flexibility. *Managment Science* **41** 577–594.

Kella, O., W. Whitt. 1990. Diffusion approximations for queues with server vacations. *Adv. Appl. Prob.* **22** 706–729.

Kella, O., W. Whitt. 1992. A storage model with a two-stage random environment. *Oper. Res.* **40** 257–262.

Mitchell, I. 2001. Call center consolidation – does it still make sense? *Business Communications Review*, December, 24–28.

Newell, G. F. 1982. *Applications of Queueing Theory*, second edition, Chapman and Hall, London.

Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** 193–267.

Pang, G., W. Whitt. 2008. Heavy-traffic limits for many-server queues with service interruptions. *working paper*. Available at: http://www.columbia.edu/∼ww2040

Perry, O., W. Whitt. 2008. A routing policy for the X call-center model designed to respond to unexpected overloads. working paper, Columbia University. Available at: http://www.columbia.edu/∼ww2040

Smith, D. R., W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* **60** (13) 39–55.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management* **7** 276–294.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Science* **38** 708–723.

Whitt, W. 2002. *Stochastic-Process Limits*, Springer, New York.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50, 1449–1461.

Whitt, W. 2006a. Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Operations Research* **54** 247–260.

Whitt, W. 2006b. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15** (1) 88–102.

Whitt, W. 2006c. Fluid models for multi-server queues with abandonments. *Operations Research* 54, 37–54.

Zhang, Z. G., N. Tian. 2003. Analysis of queueing systems with sunchronous single vacations for some servers. *Queueing Systems* **45** 161–175.

Zeltyn S., A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems* **51** (3/4) 361–402.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

e-Companion

This is an e-companion to *Service Interruptions in Large Scale Service Systems* by Guodong Pang and Ward Whitt. As indicated in §8, we elaborate upon the general model in §2.2 and we establish a many-server heavy-traffic limit, which supports our approximation of the number in system by the deterministic fluid process in a random environment.

We start in §EC.1 by giving an explicit recursive expression for the fluid content $X(t)$, conditional on a realization of the random environment process $\{(U_k, D_k, F_k) : k \geq 1\}$. We next establish the many-server heavy-traffic limit theorem in §EC.2. We first state the result in §EC.2.1. In §EC.2.2 we contrast the many-server heavy-traffic limit with service interruptions with the corresponding single-server heavy-traffic limit with service interruptions established by Kella and Whitt (1990). In §EC.2.3 we prove the theorem.

EC.1. The General Fluid Model in the Random Environment

We now consider the general deterministic fluid model in the alternating-renewal-process random environment $\{(U_k, D_k, F_k) : k \geq 1\}$, as introduced in §§2.2-2.4. The system starts at time $T_0 \equiv 0$ at the beginning of the first up interval U_1. There are n servers. Conditional upon the environment state, the system evolves deterministically according to the parameter 6-tuple $(\lambda, \mu_1, \theta_1, \mu_2, \theta_2, \theta_3)$. In this section we assume that these parameters are all positive and finite (which rules out the pure-delay and pure-loss models). Recall that the customers that were being served by servers that cease functioning remain at those non-functioning servers, but are given high priority for receiving service from newly available functioning servers, ahead of all customers initially in queue and new arrivals.

Our goal is to describe $X(t)$, the total fluid content in the system at time t for all $t \geq 0$. Let $B(t)$ be the fluid content in service being served, let $Q(t)$ be the fluid content waiting in queue, let $Y(t) \equiv B(t) + Q(t)$, and let $I(t)$ be the fluid content at non-functioning servers (which can only be positive during an interruption), all at time t. Clearly, $X(t) = Y(t) + I(t)$. Note that, $Q(t)$ and $I(t)$ decrease when customers either abandon or enter service.

We can calculate the values of the process X conditional upon the sequence $\{(U_k, D_k, F_k) : k \geq 1\}$ by repeated (recursive) application of Lemma 1. We can apply Lemma 1, because the system evolution behaves as a linear ODE in each of finitely many regions. There is a change from one linear ODE to another at boundary levels. All change times have the form

$$\tau_k^{(i)} \equiv \frac{-\log_e (1/(1+\gamma_k^{(i)}))}{\nu^{(i)}} \tag{EC.1}$$

for appropriate constants $\nu^{(i)}$ (a rate) and $\gamma_k^{(i)}$. We now describe the evolution of $X(t)$.

EC.1.1. System Evolution During an Up Interval

We now specify the system evolution (the process $(X(t), Q(t), B(t))$) conditional upon a realization of the random environment process $\{(U_k, D_k, F_k) : k \geq 1\}$). During the k^{th} up interval $[T_{k-1}, T_{k-1} + U_k)$, $k \geq 1$, the total system content $X(t)$ evolves according to the nonlinear ODE

$$\dot{X}(t) = \psi_2(X(t)) \equiv \lambda - \mu_1(X(t) \wedge n) - \theta_1(X(t) - n)^+, \quad T_{k-1} \leq t \leq T_{k-1} + U_k, \tag{EC.2}$$

starting from $X(T_{k-1}-)$. This ODE can be characterized via two linear ODE's in each of two regions: If $X(t) \leq n$, then

$$\dot{B}(t) = \lambda - \mu_1 B(t), \quad T_{k-1} \leq t \leq T_{k-1} + U_k \quad \text{and} \quad X(t) = B(t). \tag{EC.3}$$

If $X(t) > n$, then

$$\dot{Q}(t) = \lambda - \mu_1 n - \theta_1 Q(t), \quad T_{k-1} \leq t \leq T_{k-1} + U_k \quad \text{and} \quad X(t) = n + Q(t). \tag{EC.4}$$

During each up interval, the process $X(t)$ can cross over the level n at most once. It can go down below n from above only if $\lambda < n\mu_1$; it does so at time $T_{k-1} + \tau_k^{(1)}$ for $\tau_k^{(1)}$ in (EC.1); it can go up above n from below only if $\lambda > n\mu_1$; it does so at time $T_{k-1} + \tau_k^{(2)}$ for $\tau_k^{(2)}$ in (EC.1), where $X_{k-1}^u \equiv X(T_{k-1}) = X(T_{k-1}-)$,

$$\nu^{(1)} \equiv \theta_1, \quad \gamma_k^{(1)} \equiv \frac{\theta_1(X_{k-1}^u - n)}{n\mu_1 - \lambda}, \quad \nu^{(2)} \equiv \mu_1 \quad \text{and} \quad \gamma_k^{(2)} \equiv \frac{\mu_1(n - X_{k-1}^u)}{\lambda - n\mu_1}. \tag{EC.5}$$

The full solution is given in Table *EC.1*.

During an Up Time: 6 Cases	formula for $X(T_{k-1}+t)$ for $t \le U_k$ with $X^u_{k-1} \equiv X(T_{k-1})$
1. $\frac{\lambda}{\mu_1} \le n,\ X^u_{k-1} \le n$	$\left(X^u_{k-1} - \frac{\lambda}{\mu_1}\right) e^{-\mu_1 t} + \frac{\lambda}{\mu_1}$
2. $\frac{\lambda}{\mu_1} \le n,\ X^u_{k-1} > n,\ t \le \tau_k^{(1)}$	$\left(X^u_{k-1} + \frac{n\mu_1 - \lambda}{\theta_1}\right) e^{-\theta_1 t} - \left(\frac{n\mu_1 - \lambda}{\theta_1}\right)$
3. $\frac{\lambda}{\mu_1} \le n,\ X^u_{k-1} > n,\ t > \tau_k^{(1)}$	$\left(n - \frac{\lambda}{\mu_1}\right) e^{-\mu_1 (t - \tau_k^{(1)})} + \frac{\lambda}{\mu_1}$
4. $\frac{\lambda}{\mu_1} > n,\ X^u_{k-1} \ge n$	$n + \left(X^u_{k-1} - n - \frac{\lambda - \mu_1 n}{\theta_1}\right) e^{-\mu_1 t} + \left(\frac{\lambda - \mu_1 n}{\theta_1}\right)$
5. $\frac{\lambda}{\mu_1} > n,\ X^u_{k-1} < n,\ t \le \tau_k^{(2)}$	$\left(X^u_{k} - \frac{\lambda}{\mu_1}\right) e^{-\mu_1 t} + \left(\frac{\lambda}{\mu_1}\right)$
6. $\frac{\lambda}{\mu_1} > n,\ X^u_{k-1} < n,\ t > \tau_k^{(2)}$	$n + \left(\frac{\lambda - \mu_1 n}{\theta_1}\right)\left(1 - e^{-\theta_1 (t - \tau_k^{(2)})}\right)$

Table EC.1 The formulas for the time-dependent content process $X(t)$ within the k^{th} up interval starting at $X^u_{k-1} \equiv X(T_{k-1})$ for the six cases that arise.

EC.1.2. System Evolution During a Down Interval

We next consider the k^{th} down interval, beginning at time $T_{k-1} + U_k$. The initial number of customers at non-functioning servers is

$$I(T_{k-1} + U_k) \equiv (X((T_{k-1} + U_k)-) \wedge n) - ((X(T_{k-1} + U_k)-) \wedge F_k).$$

The initial number of customers in queue or functioning servers is $Y(T_{k-1} + U_k) \equiv X((T_{k-1} + U_k)-) - I(T_{k-1} + U_k)$.

There are two main cases here, depending on whether or not the number of customers in service at the beginning of the down interval exceeds the number of functioning servers. The case in which the number exceeds the number of functioning servers divides into two further subcases: (i) only $I(t)$ decreasing, and (ii) after $I(t) = 0$. We consider these three cases in turn:

EC.1.2.1. Case 1. First suppose that $X((T_{k-1} + U_k)-) \le F_k$, so that at the beginning of the down interval there is no queue and all customers that were in service are being served by functioning servers. Then the total system content $X(t)$ evolves according to the nonlinear ODE

$$\dot{X}(t) = \psi_3(X(t)) \equiv \lambda - \mu_2(X(t) \wedge F_k) - \theta_2(X(t) - F_k)^+, \quad T_{k-1} + U_k \le t \le T_k, \tag{EC.6}$$

starting from $X((T_{k-1}+U_k)-)$. If $\lambda > \mu_2 F_k$, then there are two further subcases: The queue may remain empty or it may not; let $T_{k-1}+U_k+\tau_k^{(3)}$ be the time that the queue becomes full if that can happen. The time $\tau_k^{(3)}$ has the form in *(EC.1)* with $X_k^d \equiv X(T_{k-1}+U_k)$,

$$\nu^{(3)} \equiv \mu_2 \quad \text{and} \quad \gamma_k^{(3)} \equiv \frac{\mu_2(F_k - X_k^d))}{\lambda - \mu_2 F_k}. \tag{EC.7}$$

If either $\lambda \le \mu_2 F_k$ or both $\lambda > \mu_2 F_k$ and $t < \tau_k^{(3)}$, then $X(t) = B(t)$, which is governed by the linear ODE

$$\dot{B}(t) = \lambda - \mu_2 B(t), \quad T_{k-1} + U_k \le t \le (T_{k-1} + U_k + \tau_k^{(3)}) \wedge T_k. \tag{EC.8}$$

If $\lambda > \mu_2 F_k$ and $t \ge \tau_k^{(3)}$, then $X(T_{k-1}+U_k+t) = F_k + Q(T_{k-1}+U_k+t) > F_k$, where the queue length $Q(t)$ evolves according to the linear ODE

$$\dot{Q}(t) = \lambda - \mu_2 F_k - \theta_2 Q(t), \quad (T_{k-1} + U_k + \tau_k^{(3)}) \wedge T_k \le t \le T_k. \tag{EC.9}$$

EC.1.2.2. Case 2. Next suppose that $X((T_{k-1}+U_k)-) > F_k$, with $F_k < n$, so that some customers that were being served before the interruption can no longer be served at the beginning of the down interval. Hence, $I(T_{k-1}+U_k) > 0$. Since these customers have priority for entering service, $I(t)$ decreases steadily over time, until it reaches 0 at the time $T_{k-1}+U_k+\tau_k^{(4)}$, according to the linear ODE

$$\dot{I}(t) = -\mu_2 F_k - \theta_3 I(t), \qquad T_{k-1} + U_k \le t \le (T_{k-1} + U_k + \tau_k^{(4)}) \wedge T_k, \tag{EC.10}$$

where $\tau_k^{(4)}$ has the form in (EC.1) with

$$\nu^{(4)} \equiv \theta_3 \quad \text{and} \quad \gamma_k^{(4)} \equiv \frac{\theta_3 I(T_{k-1}+U_k)}{\mu_2 F_k}. \tag{EC.11}$$

During the interval $[T_{k-1}+U_k, T_{k-1}+U_k+\tau_k^{(4)}]$, $B(t)$ remains unchanged, so that the evolution of $Y(t) \equiv X(t) - I(t)$ is determined by the linear ODE for Q:

$$\dot{Q}(t) = \lambda - \theta_2 Q(t). \tag{EC.12}$$

EC.1.2.3. Case 3. It remains to describe what happens after time $T_{k-1}+U_k+\tau_k^{(4)}$ when $X((T_{k-1}+U_k)-) > F_k$. Then the total system content again evolves according to the nonlinear ODE in *(EC.6)*. In this case, we start with $Q(T_{k-1}+U_k+\tau_k^{(4)}) > 0$. If $\lambda \geq \mu_2 F_k$, then the queue will remain positive. If $\lambda < \mu_2 F_k$, then the queue can disappear; that will occur at time $T_{k-1}+U_k+\tau_k^{(4)}+\tau_k^{(5)}$, where $\tau_k^{(5)}$ has the form in (EC.1) with

$$\nu^{(5)} \equiv \theta_2 \quad \text{and} \quad \gamma_k^{(5)} \equiv \frac{\theta_2 (X(T_{k-1}+U_k+\tau_k^{(4)})-n)^+}{\mu_2 F_k - \lambda}. \tag{EC.13}$$

If either $\lambda \geq \mu_2 F_k$ or both $\lambda < \mu_2 F_k$ and $t \leq \tau_k^{(4)}+\tau_k^{(5)}$, then the queue is nonempty and $Q(t)$ evolves according to the linear ODE (EC.9). On the other hand, if $\lambda < \mu_2 F_k$ and $t > \tau_k^{(4)}+\tau_k^{(5)}$, then the queue has become empty, so that $X(t) = B(t)$ evolves according to the linear ODE in *(EC.8)*. The full solution is given in Table *EC.2*.

During a Down Time: 5 Cases	formula for $X(T_{k-1}+U_k+t)$ for $t \leq D_k$ with $X_k^d \equiv X(T_{k-1}+U_k)$
1. For $X_k^d \leq F_k$ and either $\lambda \leq \mu_2 F_k$ or both $\lambda > \mu_2 F_k$ and $t < \tau_k^{(3)}$,	$\left(X_k^d - \frac{\lambda}{\mu_2}\right) e^{-\mu_2 t} + \frac{\lambda}{\mu_2}$ $[Q(t) = I(t) = 0 \text{ here}]$ $[X(T_{k-1}+U_k+\tau_k^{(3)}) = F_k]$
2. For $X_k^d \leq F_k$, $\lambda > \mu_2 F_k$, $t > \tau_k^{(3)}$,	$F_k + \left(\frac{\lambda - \mu_2 F_k}{\theta_2}\right)\left(1 - e^{-\theta_2 (t-\tau_k^{(3)})}\right)$ $[Q(t) > 0,\ I(t) = 0]$
3. For $X_k^d > F_k$, $t \leq \tau_k^{(4)}$, $[I(T_{k-1}+U_k+\tau_k^{(4)}) = 0]$	$F_k + \left((X_k^d - n)^+ - \frac{\lambda}{\theta_2}\right) e^{-\theta_2 t} + \frac{\lambda}{\theta_2}$ $+ \left((X_k^d \wedge n) - (X_k^d \wedge F_k) + \frac{\mu_2 F_k}{\theta_3}\right) e^{-\theta_3 t} - \frac{\mu_2 F_k}{\theta_3}$ [Here $I(t) > 0$ and $X(t) - I(t) > 0$]
4. For $X_k^d > F_k$, $t > \tau_k^{(4)}$, and either $\lambda \geq \mu_2 F_k$ or both $\lambda < \mu_2 F_k$ and $t \leq \tau_k^{(4)}+\tau_k^{(5)}$	$F_k + \left(X(T_{k-1}+U_k+\tau_k^{(4)}) - F_k - \frac{\lambda-\mu_2 F_k}{\theta_2}\right) e^{-\theta_2 (t-\tau_k^{(4)})}$ $+ \left(\frac{\lambda - \mu_2 F_k}{\theta_2}\right)$ $[I(t) = 0$ and $X(t) > F_k]$
5. For $X_k^d > F_k$, $\lambda < \mu_2 F_k$, $t > \tau_k^{(4)}+\tau_k^{(5)}$,	$\left(F_k - \frac{\lambda}{\mu_2}\right) e^{-\mu_2 (t-\tau_k^{(4)}-\tau_k^{(5)})} + \frac{\lambda}{\mu_2}$ $[I(t) = 0$ and $X(t) \leq F_k]$

Table EC.2 The formulas for the time-dependent content process $X(t)$ within the k^{th} down interval starting at $X_k^d \equiv X(T_{k-1}+U_k)$ for the five cases that arise.

EC.1.3. Overall Performance

It is evident that the stochastic processes $\{X_k^u \equiv X(T_k) : k \geq 0\}$ and $\{X_k^d \equiv X(T_k + U_k) : k \geq 1\}$ are discrete-time Markov processes (DTMP's) with stationary transition probabilities for *both* the original Markovian queueing model and the approximating deterministic fluid model, in the same random environment. In both cases, $\{X(t) : t \geq 0\}$ is a regenerative stochastic process. For the fluid model, the transition probabilities of the DTMP given (X_{k-1}^u, U_k, D_k) or $(X_k^d, U_{k+1}, D_{k+1})$ are easily obtained from the distribution of (U_k, D_k, F_k) by combining the results in Tables EC.1 and EC.2.

As indicated in §8, the expressions we have derived in this section can serve as the basis for an efficient simulation algorithm. The long-run behavior of $X(t)$ can be calculated numerically by: (i) generating a long initial segment of the sequence $\{(U_k, D_k, F_k) : k \geq 1\}$, (ii) applying the explicit formulas above, and (iii) computing averages such as

$$\frac{1}{t}\int_0^t 1_{\{X(s)>y\}}\,ds \quad \text{and} \quad \frac{1}{t}\int_0^t 1_{\{X(s)>y, U(s)=1\}}\,ds \tag{EC.14}$$

for large values of t, where $U(t)$ records the state of the environment at time t, being 1 if the system is up and being 0 if the system is down; i.e.,

$$U(t) = 1 \quad \text{if} \quad T_{k-1} \leq t < T_{k-1} + U_k, \quad k \geq 1, \tag{EC.15}$$

and $U(t) = 0$ otherwise.

EC.1.4. The Impact of Different Abandonment Rates

We close this section by giving a figure showing the fluid content in the model with customer abandonment. We assume that $P(F = 0) = 1$. It is to be contrasted with Figures 1 and 2, showing the fluid content in the pure-delay and pure-loss models. Abandonment makes the fluid content lie below the fluid content in the pure-delay model, but above the fluid content in the pure-loss model. Possible intermediate cases are shown in Figure EC.1.

As noted in §6, the fluid content is the same as if there is no interruption at all when $F = 0$ and $\theta_3 = \mu_1$. Then, during the interruption, the fluid content remains constant at the steady-state value

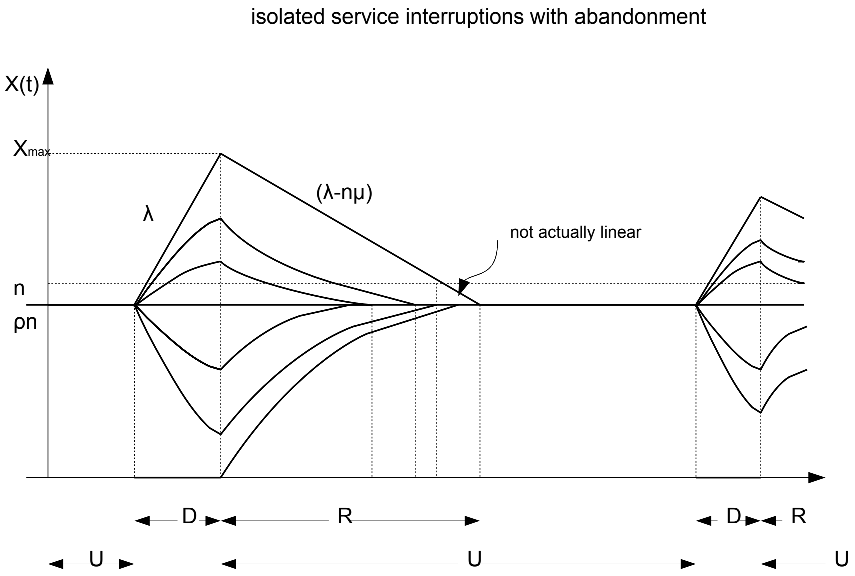

Figure EC.1 Sample paths of the fluid content, X(t), in the system experiencing an isolated exogenous service interruption, when there is customer abandonment at various rates with $F = 0$.

of ρn. During the interruption, customers leave by abandonment instead of by service completion, at precisely the same rate. There is no recovery period at all, because the system content has never changed. That case of constant fluid content also applies with $F > 0$ provided that $\theta_3 = \mu_2 = \mu_1$.

The fluid content curves will be above the horizontal line at ρn when $\theta_3 < \mu_1$ and lie below when $\theta_3 > \mu_1$. The fluid content will rise more above the level n when θ_2 and θ_1 are smaller. If θ_2 and θ_1 are very high, but θ_3 and μ_2 are very small, then the fluid content will tend to be near n during the interruption. The model in §EC.1 can be used to compute what happens in an isolated interruption with general parameters. At the beginning of the down period, the fluid content starts at its steady-state value ρn. Given an interruption with $F = 0$, we start in Case 3 of Table EC.2 and continue with Case 4. Given the length of the down time, we can compute the fluid content at the end of the down period. We then can compute the behavior in the recovery period by applying Table EC.1.

EC.2. The Many-Server Heavy-Traffic Limit

In this section we establish the many-server heavy-traffic limit theorem for the Markovian queueing model in the random environment $\{(U_k, D_k, F_{n,k}) : k \geq 1\}$ model, which supports the approximation in §EC.1 as well as the previous approximations.

EC.2.1. Theorem Statement

Let $X_n(t)$ represent the number of customers in model n at time t and let $\bar{X}_n(t) \equiv n^{-1}X_n(t)$, $t \geq 0$. With this framework, we can establish a fluid limit that is valid in all three many-server heavy-traffic limiting regimes: QD, QED and ED. It can be regarded as a functional weak law of large numbers (FWLLN), for which there will be a functional central limit theorem (FCLT) refinement, which we do not consider. The FCLT extension has centering by the scaled fluid limit $n\bar{X}(t)$ and then division by $\sqrt{n}$. The following is proved in the e-companion. As usual, let $\Rightarrow$ denote convergence in distribution and let $D[0,\infty)$ denote the function space of possible sample paths (real-valued functions on the interval $[0,\infty)$); e.g., see Whitt (2002).

Theorem EC.1. (many-server heavy-traffic fluid limit with unscaled interruptions) *Consider the sequence of Markovian queueing models in a fixed random environment $\{(U_k, D_k) : k \geq 1\}$ in one of the many-server heavy-traffic limiting regimes: QD, QED or ED. Assume that $F_{n,k}/n \Rightarrow \bar{F}_k$ as $n \to \infty$ for each k. If there exists a random variable $\bar{X}(0)$ such that $\bar{X}_n(0) \Rightarrow \bar{X}(0)$ as $n \to \infty$, then the fluid-scaled number-in-system process converges in distribution, i.e.,*

$$\bar{X}_n \Rightarrow \bar{X} \quad in \quad D[0,\infty) \quad as \quad n \to \infty, \tag{EC.16}$$

where the limiting stochastic process $\bar{X} \equiv \{\bar{X}(t) : t \geq 0\}$ has continuous sample paths, evolving as an ODE in the random environment $\{(U_k, D_k, \bar{F}_k) : k \geq 1\}$. The evolution of $n\bar{X}(t)$, conditional on a possible realization of $\{(U_k, D_k, F_k) : k \geq 1\}$ is given in §EC.1*, where $F_k = n\bar{F}_k$.*

EC.2.2. Contrast with Conventional Heavy Traffic

Theorem EC.1 here parallels a conventional heavy-traffic stochastic-process limit for single-server queues with exogenous service interruptions in Kella and Whitt (1990). In the conventional heavy-traffic limits, both the up times and down times were required to grow in the heavy-traffic scaling, whereas here in Theorem EC.1 the up and down times remain fixed. This difference follows naturally from the very different scaling that is used in the two settings.

With only one server (or any fixed finite number of servers), we contract time by the factor $1/(1-\rho)^2$ and we scale space by multiplying by $1-\rho$, where ρ is the traffic intensity, which is increasing toward 1. That means we are describing the behavior of the original system over long time intervals. Consequently, with one server, the up times and down times are required to grow in the limit as $\rho \uparrow 1$, with the up times being of order $O(1/(1-\rho)^2)$ as $\rho \uparrow 1$, while the down times were either of order $O(1/(1-\rho)^2)$ or $O(1/(1-\rho))$. The time scaling for the up times is reasonable to represent infrequent interruptions, which would hopefully reflect the actual situation. The scaling of the down times means that the approximations are only appropriate for relatively long interruptions. Short service interruptions can often be modelled by simply adjusting the distribution of individual service times. In the single-server setting, the performance impact of short service interruptions can usually be captured in models by the increased variance in these individual service times.

In contrast, for the many-server heavy-traffic limits, there is no need to consider large time intervals, because we increase the arrival rate. Thus, we do not scale time at all, but the number of customers waiting still grows, so that we scale space by dividing by n for the fluid limit, as in Theorem EC.1. Since we do not scale time with many servers, the up and down times need not change as $n \to \infty$. Thus, in the many-server heavy-traffic regimes the length of the service interruptions need not be long in order to produce noticeable impact. Indeed, even short service interruptions can have a dramatic impact, as discussed after Theorem 1 in §4. In Pang and Whitt (2008), we obtain a diffusion limit in the QED regime when the down times are assumed to be asymptotically negligible, being of order $O(1/\sqrt{n})$ as $n \to \infty$. Even these asymptotically negligible down times have an impact through jumps up in the limit process.

EC.2.3. Proof of Theorem EC.1

As in Kella and Whitt (1990), it is convenient to do the proof recursively, considering the successive up and down interval in turn, conditioning upon a realization of the random environment $\{(U_k, D_k):$

$k\geq 1\}$. The limit for the final values over one interval becomes the limit for the initial values in the next interval. On each separate interval, we have the transient behavior of a Markovian queue in a stationary deterministic environment. Thus, on each separate interval, we can apply the martingale argument, using random-time-changed rate-1 Poisson processes, as in the review paper by Pang et al. (2007). The proof below actually applies to general non-Poisson arrival processes, provided that they satisfy a FWLLN; see §7.3 of Pang et al. (2007).

First, during the k^{th} up time interval $[T_{k-1}, T_{k-1}+U_k)$, $k\geq 1$, the total system content $X_n(t)$ evolves according to the dynamics,

$$\begin{aligned}X_n(t) &= X_n(T_{k-1}-) + A_n(t) - A_n(T_{k-1}-) - S_{1,k}\Big(\mu_1\int_{T_{k-1}}^{t}(X_n(s)\wedge n)ds\Big)\\ &\quad - L_{1,k}\Big(\theta_1\int_{T_{k-1}}^{t}(X_n(s)-n)^+ds\Big),\end{aligned}\tag{EC.17}$$

for $t\in[T_{k-1}, T_{k-1}+U_k)$, where $X_n(T_{k-1}) = X_n(T_{k-1}-)$, and $\{S_{1,k}(t): t\geq 0\}$ and $\{L_{1,k}(t): t\geq 0\}$ are independent Poisson processes with unit rate.

Next, we consider the k^{th} down time interval $[T_{k-1}+U_k, T_k)$. The initial number of customers at non-functioning servers is $I_n(T_{k-1}+U_k) = (X_n((T_{k-1}+U_k)-)\wedge n) - (X_n((T_{k-1}+U_k)-)\wedge F_{n,k})$, and the initial number of customers in queue or functioning servers is $Y_n(T_{k-1}+U_k) = X_n((T_{k-1}+U_k)-) - I_n(T_{k-1}+U_k)$. We have two cases: $X_n((T_{k-1}+U_k)-)\leq F_{n,k}$ and $X_n((T_{k-1}+U_k)-) > F_{n,k}$.

In the first case when $X_n((T_{k-1}+U_k)-)\leq F_{n,k}$, $I_n(T_{k-1}+U_k)=0$, so the evolution of the total system content $X_n(t)$ is the same as that for many-server queues with $F_{n,k}$ servers, starting from $X_n((T_{k-1}+U_k)-)$, so the process $X_n(t)$ satisfies the dynamics

$$\begin{aligned}X_n(t) &= X_n((T_{k-1}+U_k)-) + A_n(t) - A_n((T_{k-1}+U_k)-) - S_{2,k}\Big(\mu_2\int_{T_{k-1}+U_k}^{t}(X_n(s)\wedge F_{n,k})ds\Big)\\ &\quad - L_{2,k}\Big(\theta_2\int_{T_{k-1}+U_k}^{t}(X_n(s)-F_{n,k})^+ds\Big),\end{aligned}\tag{EC.18}$$

for $t\in[T_{k-1}+U_k, T_k)$, where $X_n(T_{k-1}+U_k) = X_n((T_{k-1}+U_k)-)$, and $\{S_{2,k}(t): t\geq 0\}$ and $\{L_{2,k}(t): t\geq 0\}$ are independent Poisson processes with unit rate, and also independent of $S_{1,k}$ and $L_{1,k}$.

In the second case, when $X_n((T_{k-1}+U_k)-) > F_{n,k}$, the number of customers at the non-functioning servers at the beginning of this down time is positive, i.e., $I_n(T_{k-1}+U_k) > 0$. Since these customers have priority for service, $I_n(t)$ decreases according to the following dynamics

$$\begin{aligned} I_n(t) = {} & I_n(T_{k-1}+U_k) - S_{2,k}\Big(\mu_2 F_{n,k}(t\wedge\tau_{n,k} - (T_{k-1}+U_k))\Big) \\ & - L_{3,k}\Big(\theta_3 \int_{T_{k-1}+U_k}^{t\wedge\tau_{n,k}} I_n(s)ds\Big) \end{aligned} \qquad \text{(EC.19)}$$

for $t \in [T_{k-1}+U_k, \tau_{n,k})$, where $\tau_{n,k} \equiv \inf\{t \in [T_{k-1}+U_k, T_k) : I_n(t) = 0\} \wedge T_k$ with the convention that $\inf\{\emptyset\} = \infty$, and $\{L_{3,k}(t) : t \geq 0\}$ is a Poisson process with unit rate, independent of $S_{2,k}$. The rate-1 Poisson processes $S_{i,k}$'s and $L_{j,k}$'s $(i = 1,2, j = 1,2,3, k = 1,2,3,\ldots)$ are also assumed to be mutually independent. To explain (EC.19), note that these customers that were at non-functioning servers, enter service at functioning servers at rate $\mu_2 F_{n,k}$ as long as there are any such kind of customers in the systems because any customer that completes service by a functioning server will be replaced by customers from the originally interrupted customers; i.e., the rate into service from the customers at the non-functioning servers equals the rate out of service by service completions, until eventually $I(t) = 0$. The remaining system content $Y_n(t) \equiv X_n(t) - I_n(t)$ up to time $\tau_{n,k}$ evolves according to the dynamics

$$\begin{aligned} Y_n(t) = {} & Y_n((T_{k-1}+U_k)-) + A_n(t) - A_n((T_{k-1}+U_k)-) \\ & - L_{2,k}\Big(\theta_2 \int_{T_{k-1}+U_k}^{t} (Y_n(s) - F_k)^+ ds\Big), \end{aligned} \qquad \text{(EC.20)}$$

for $t \in [T_{k-1}+U_k, \tau_{n,k})$, where

$$Y_n(T_{k-1}+U_k) = Y_n((T_{k-1}+U_k)-) \equiv X_n((T_{k-1}+U_k)-) - I_n(T_{k-1}+U_k).$$

We next describe the system evolution after the time $\tau_{n,k}$, at which there are no more customers at non-functioning servers. From time $\tau_{n,k}$ to the end of this down time T_k, the total system content $X_n(t)$ will evolve according to the dynamics

$$X_n(t) = X_n(\tau_{n,k}) + A_n(t) - A_n(\tau_{n,k}-) - S_{2,k}\Big(\mu_2 \int_{\tau_{n,k}}^{t} (X_n(s) \wedge F_{n,k}) ds\Big)$$

$$- L_{2,k}\Big(\theta_2 \int_{\tau_{n,k}}^t (X_n(s) - F_{n,k})^+ ds\Big), \tag{EC.21}$$

for $t \in [\tau_{n,k}, T_k)$, starting from $X_n(\tau_{n,k}) = Y_n(\tau_{n,k})$.

It can be shown that the integral equations in (EC.17)-(EC.21) provide a valid characterization of the corresponding processes, as in Lemma 2.1 of Pang et al. (2007). We can then obtain the corresponding integral equations for associated fluid-scaled processes. For that purpose, define the associated fluid-scaled process $\bar{X}_n \equiv \{\bar{X}_n(t) : t \geq 0\}$ by letting $\bar{X}_n(t) \equiv n^{-1} X_n(t)$, $t \geq 0$, and let the other processes be scaled similarly.

From (EC.17), we directly obtain the integral equation for $\bar{X}_n$ in the k^{th} up time interval $[T_{k-1}, T_{k-1} + U_k)$

$$\begin{aligned} \bar{X}_n(t) &= \bar{X}_n(T_{k-1}-) + \bar{A}_n(t) - \bar{A}_n(T_{k-1}-) - \bar{S}_{n,1,k}(t) - \bar{L}_{n,2,k}(t) \\ &\quad - \mu_1 \int_{T_{k-1}}^t (\bar{X}_n(s) \wedge 1) ds - \theta_1 \int_{T_{k-1}}^t (\bar{X}_n(s) - 1)^+ ds, \end{aligned} \tag{EC.22}$$

for $t \in [T_{k-1}, T_{k-1} + U_k)$, where $\bar{X}_n(T_{k-1}) \equiv \bar{X}_n(T_{k-1}-)$, and

$$\bar{S}_{n,1,k}(t) \equiv \frac{1}{n}\Big(S_{1,k}\Big(\mu_1 \int_{T_{k-1}}^t (X_n(s) \wedge n) ds\Big) - \mu_1 \int_{T_{k-1}}^t (X_n(s) \wedge n) ds\Big), \tag{EC.23}$$

and

$$\bar{L}_{n,1,k}(t) \equiv \frac{1}{n}\Big(L_{1,k}\Big(\theta_1 \int_{T_{k-1}}^t (X_n(s) \wedge n) ds\Big) - \theta_1 \int_{T_{k-1}}^t (X_n(s) \wedge n) ds\Big). \tag{EC.24}$$

The key observation for the martingale argument is that the processes $\hat{S}_{n,1,k}$ and $\hat{L}_{n,1,k}$ are square-integrable martingales, where

$$\hat{S}_{n,1,k} \equiv \sqrt{n} \bar{S}_{n,1,k} \quad \text{and} \quad \hat{L}_{n,1,k} \equiv \sqrt{n} \bar{L}_{n,1,k}, \tag{EC.25}$$

with $\bar{S}_{n,1,k}$ and $\bar{L}_{n,2,k}$ defined in (EC.23) and (EC.24). The associated predictable quadratic variation processes are $\langle \hat{S}_{n,1,k} \rangle \equiv \{\langle \hat{S}_{n,1,k} \rangle(t) : t \in [T_{k-1}, T_{k-1} + U_k)\}$ and $\langle \hat{L}_{n,1,k} \rangle \equiv \{\langle \hat{L}_{n,1,k} \rangle(t) : t \in [T_{k-1}, T_{k-1} + U_k)\}$, where

$$\langle \hat{S}_{n,1,k} \rangle(t) \equiv \mu_1 \int_{T_{k-1}}^t (\bar{X}_n(s) \wedge 1) ds, \quad \text{and} \quad \langle \hat{L}_{n,1,k} \rangle(t) \equiv \theta_1 \int_{T_{k-1}}^t (\bar{X}_n(s) - 1)^+ ds.$$

As usual, we must specify the filtration $\mathbf{F}^u_{n,k} \equiv \{\mathcal{F}^u_{n,k}(t) : t \in [T_{k-1}, T_{k-1} + U_k)\}$. The idea is to condition on the entire arrival process A_n, exogenous service interruption process $\{(U_i, D_i) : i \geq 1\}$, and the functioning servers $\{F_{n,i} : i \geq 1\}$ so that we put the entire sample paths of these stochastic process into the filtration. That is how we achieve conditioning upon the random environment. At the same time, we allow the treatment of general arrival processes.

Thus the filtration is

$$\mathcal{F}^u_{n,k}(t) \equiv \sigma\Big(X_n(T_{k-1}-), S_{1,k}\Big(\mu_1 \int_{T_{k-1}}^t (X_n(s) \wedge n) ds\Big), L_{1,k}\Big(\theta_1 \int_{T_{k-1}}^t (X_n(s) \wedge n) ds\Big)$$
$$\vee\, \sigma\big(\{(U_i, D_i) : i = 1, 2, ...\}, \{F_{n,i} : i \geq 1\}\big) \vee \sigma\Big(A_n(s) : s \geq 0\Big) \vee \mathcal{N}$$

where $\mathcal{N}$ contains all the null sets. This step of the argument follows the proof of Theorem 7.2 in §7.1 of Pang et al. (2007), given that we have conditioned on the entire the exogenous service-interruption process.

Moreover, by Lemmas 3.3, 5.8 and 6.2 of Pang et al. (2007), the sequence of processes $\{(\hat{S}_{n,1,k}, \hat{L}_{n,1,k}) : n \geq 1\}$ defined in equations (EC.25), (EC.23) and (EC.24) is stochastically bounded in the space D^2, i.e., for all $\epsilon > 0$ and $T > T_{k-1} + U_k > 0$, there exists a constant positive K such that

$$P(||(\hat{S}_{n,1,k}, \hat{L}_{n,1,k})||_{T,k} > K) > 1 - \epsilon, \quad \text{for all} \quad n \geq 1,$$

where $\|(x,y)\|_{T,k} \equiv \max\{\|x\|_{T,k}, \|y\|_{T,k}\}$ and $\|x\|_{T,k} \equiv \sup_{t \in [0,T] \cap [T_{k-1}, T_{k-1}+U_k)} \{|x(t)|\}$. Thus we can apply the FWLLN for stochastic bounded sequences of processes taking values in D^2 established in Lemma 5.9 in Pang et al. (2007) to obtain

$$(\bar{S}_{n,1,k}, \bar{L}_{n,1,k}) \Rightarrow (\eta, \eta) \quad \text{in} \quad (D, J_1) \quad \text{as} \quad n \to \infty, \tag{EC.26}$$

where $\eta(t) \equiv 0$ for all $t \geq 0$ and the space D is restricted to the time interval $[T_{k-1}, T_{k-1} + U_k)$.

By the FWLLN in (EC.26) and the assumptions on the arrivals and initial conditions, and the recursive proof procedure, we obtain the joint convergence

$$(\bar{X}_n(T_k-), \bar{A}_n, \bar{S}_{n,1,k}, \bar{L}_{n,1,k}) \Rightarrow (\bar{X}(T_k-), \lambda e, \eta, \eta) \quad \text{in} \quad \mathbb{R} \times \mathrm{D}^3, \quad \text{as} \quad \mathrm{n} \to \infty, \tag{EC.27}$$

where the space D^3 is endowed with the Skorohod J_1 topology and restricted to the time interval $[T_{k-1}, T_{k-1}+U_k)$, and $e(t) \equiv t$ for all $t \geq 0$. Since the limit is a continuous function, the mode of convergence is uniform convergence on bounded subintervals. Given that the total interval $[T_{k-1}, T_{k-1}+U_k]$ is bounded, the mode of convergence is uniform over this interval.

We complete the recursive proof for the time interval $[T_{k-1}, T_{k-1}+U_k)$ by applying the continuous mapping theorem. To do so, consider the deterministic function $x \equiv \psi(b, y)$ determined by the integral representation

$$x(t) = b + y(t) + \int_{T_{k-1}}^{t} h(x(s))ds, \quad t \in [T_{k-1}, T_{k-1}+U_k), \tag{EC.28}$$

where the function $h : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous. By Theorem 4.1 of Pang et al. (2007), the mapping ψ specified by (EC.28) is well defined and continuous when the space D restricted to the time interval $[T_{k-1}, T_{k-1}+U_k)$ is endowed with Skorohod J_1 topology. Moreover, if y is continuous, then so is x.

The integral representation for $\bar{X}_n$ in (EC.22) corresponds to (EC.28) with functions $h(x) = -\mu_1(x \wedge 1) - \theta_1(x-1)^+$ for all $x \in \mathbb{R}$. By conditioning on the random environment $\{(U_k, D_k) : k \geq 1\}$, we can apply continuous mapping theorem with the addition operation and the mapping ψ together with the convergence of the processes $(\bar{X}_n(T_{k-1}-), \bar{A}_n, \bar{S}_{n,1,k}, \bar{L}_{n,1,k})$ in (EC.27) to obtain the weak convergence of the processes $\bar{X}_n$ in (EC.22) to the process $\bar{X}$ in the k^{th} up time interval $[T_{k-1}, T_{k-1}+U_k)$, where the process $\bar{X}$ is given by

$$\bar{X}(t) = \bar{X}(T_{k-1}) + \lambda(t - T_{k-1}) - \mu_1 \int_{T_{k-1}}^{t} (\bar{X}(s) \wedge 1)ds - \theta_1 \int_{T_{k-1}}^{t} (\bar{X}(s) - 1)^+ ds, \tag{EC.29}$$

for $t \in [T_{k-1}, T_{k-1}+U_k)$. Clearly, the integral representation in (EC.29) is equivalent to the ODE in (EC.2) except for the scaling: $X(t) = n\bar{X}(t)$, where X is the content process in §EC.1.

We can now proceed to do essentially the same argument in the down intervals. Thus consider the k^{th} down time interval $[T_{k-1}+U_k, T_k)$ and suppose that $X_n((T_{k-1}+U_k)-) \leq F_{n,k}$, from (EC.18), we obtain the following integral equation for the fluid-scaled process $\bar{X}_n$:

$$\bar{X}_n(t) = \bar{X}_n((T_{k-1}+U_k)-) + \bar{A}_n(t) - \bar{A}_n((T_{k-1}+U_k)-) - \bar{S}_{n,2,k}(t) - \bar{L}_{n,2,k}(t)$$

$$-\mu_2\int_{T_{k-1}+U_k}^t\Big(\bar{X}_n(s)\wedge\frac{F_{n,k}}{n}\Big)ds-\theta_2\int_{T_{k-1}+U_k}^t\Big(\bar{X}_n(s)-\frac{F_{n,k}}{n}\Big)^+ds, \qquad \text{(EC.30)}$$

for $t\in[T_{k-1}+U_k,T_k)$, where $\bar{X}_n(T_{k-1}+U_k)=\bar{X}_n((T_{k-1}+U_k)-)$, and

$$\bar{S}_{n,2,k}(t)=\frac{1}{n}\Big(S_{2,k}\Big(\mu_2\int_{T_{k-1}+U_k}^t(X_n(s)\wedge F_{n,k})ds\Big)-\mu_2\int_{T_{k-1}+U_k}^t(X_n(s)\wedge F_{n,k})\,ds\Big), \quad \text{(EC.31)}$$

and

$$\bar{L}_{n,2,k}(t)=\frac{1}{n}\Big(L_{2,k}\Big(\theta_2\int_{T_{k-1}+U_k}^t(X_n(s)-F_{n,k})^+ds\Big)-\theta_2\int_{T_{k-1}+U_k}^t(X_n(s)-F_{n,k})^+ds\Big), \text{(EC.32)}$$

for $t\in[T_{k-1}+U_k,T_k)$. Recall that we assume that the functioning servers at each interruption grow according to the scale, i.e., $F_{n,k}/n\Rightarrow\bar{F}_k$ as $n\to\infty$.

By an analogous proof to the proof of convergence $\bar{X}_n$ to $\bar{X}$ in the k^{th} up time interval $[T_{k-1},T_{k-1}+U_k)$, we obtain the weak convergence of the processes $\bar{X}_n$ in (EC.30) to the process $\bar{X}$ in k^{th} down time interval $[T_{k-1}+U_k,T_k)$ when $\bar{X}(T_{k-1}+U_k)\le\bar{F}_k$, where the process $\bar{X}$ is defined by

$$\begin{aligned}\bar{X}(t)\,=\,&\bar{X}(T_{k-1}+U_k)+\lambda(t-(T_{k-1}+U_k))-\mu_2\int_{T_{k-1}+U_k}^t\big(\bar{X}(s)\wedge\bar{F}_k\big)ds\\&-\theta_2\int_{T_{k-1}+U_k}^t\big(\bar{X}(s)-\bar{F}_k\big)^+ds\end{aligned} \qquad \text{(EC.33)}$$

for $t\in[T_{k-1}+U_k,T_k)$. The limit in (EC.33) is equivalent to the ODE in (EC.6). When we apply the continuous mapping theorem, we need to consider the deterministic function $x\equiv\phi(b,y,z)$ determined by the integral representation

$$x(t)=b+y(t)+\int_{T_{k-1}+U_k}^t g(x(s),z)\,ds,\quad t\in[T_{k-1}+U_k,T_k), \qquad \text{(EC.34)}$$

where the function $g:\mathbb{R}^2\to\mathbb{R}$ is Lipschitz continuous. It is straightforward to generalize the proof of Theorem 4.1 of Pang et al. (2007) to prove that the mapping ϕ specified by (EC.34) is well defined and continuous when the space D restricted to the time interval $[T_{k-1}+U_k,T_k)$ is endowed with Skorohod J_1 topology. Moreover, if y is continuous, then so is x. The integral representation for $\bar{X}_n$ in (EC.30) corresponds to (EC.34) with functions $g(x,z)=-\mu_2(x\wedge z)-\theta_2(x-z)^+$ for all $x,z\in\mathbb{R}$.

We now consider the remaining case: Again consider the k^{th} down time interval $[T_{k-1}+U_k, T_k)$ but now suppose that $X_n((T_{k-1}+U_k)-) > F_{n,k}$. From (EC.19), (EC.20) and (EC.21), we obtain the following integral equations for the fluid-scaled processes $\bar{I}_n$, $\bar{Y}_n$ and $\bar{X}_n$

$$\begin{aligned}
\bar{I}_n(t) &= \bar{I}_n(T_{k-1}+U_k) - \bar{S}_{n,2,k}(t\wedge\tau_{n,k}) - \bar{L}_{n,3,k}(t\wedge\tau_{n,k}) - \mu_2\frac{F_{n,k}}{n}(t\wedge\tau_{n,k} \\
&\qquad -(T_{k-1}+U_k)) - \theta_3\int_{T_{k-1}+U_k}^{t\wedge\tau_{n,k}}\bar{I}_n(s)ds, \\
\bar{Y}_n(t) &= \bar{Y}_n((T_{k-1}+U_k)-) + \bar{A}_n(t\wedge\tau_{n,k}) - \bar{A}_n((T_{k-1}+U_k)-) \\
&\qquad -\bar{L}_{n,2,k}(t) - \theta_2\int_{T_{k-1}+U_k}^{t\wedge\tau_{n,k}}\Big(\bar{Y}_n(s)-\frac{F_k}{n}\Big)^+ds,
\end{aligned} \tag{EC.35}$$

for $t\in[T_{k-1}+U_k,\tau_{n,k})$, where $\bar{I}_n(T_{k-1}+U_k) = (\bar{X}_n((T_{k-1}+U_k)-)\wedge 1) - (\bar{X}_n((T_{k-1}+U_k)-)\wedge\frac{F_{n,k}}{n})$, $\bar{Y}_n(T_{k-1}+U_k) = \bar{X}_n((T_{k-1}+U_k)-) - \bar{I}_n(T_k+U_k)$, $\bar{S}_{n,2,k}$ and $\bar{L}_{n,2,k}$ are defined in (EC.31) and (EC.32), respectively, and $\bar{L}_{n,3,k}$ is defined by

$$\bar{L}_{n,3,k}(t) = \frac{1}{n}\Big(L_{3,k}\Big(\theta_3\int_{T_{k-1}+U_k}^{t}I_n(s)ds\Big) - \theta_3\int_{T_{k-1}+U_k}^{t}I_n(s)ds\Big)$$

for $t\in[T_{k-1}+U_k,T_k)$, and

$$\begin{aligned}
\bar{X}_n(t) &= \bar{X}_n(\tau_{n,k}) + \bar{A}_n(t) - \bar{A}_n(\tau_{n,k}-) - (\bar{S}_{n,2,k}(t)-\bar{S}_{n,2,k}(\tau_{n,k})) - (\bar{L}_{n,2,k}(t)-\bar{L}_{n,2,k}(\tau_{n,k})) \\
&\quad -\mu_2\int_{\tau_{n,k}}^{t}\Big(\bar{X}_n(s)\wedge\frac{F_{n,k}}{n}\Big)ds - \theta_2\int_{\tau_{n,k}}^{t}\Big(\bar{X}_n(s)-\frac{F_{n,k}}{n}\Big)^+ds,
\end{aligned} \tag{EC.36}$$

for $t\in[\tau_{n,k},T_k)$, where $\bar{X}_n(\tau_{n,k}) = \bar{Y}_n(\tau_{n,k})$. Recall that $\tau_{n,k}$ is the first time at which $I_n(t)=0$. Since the limit $I(t)$ is strictly decreasing and continuous, we have

$$\tau_{n,k}\Rightarrow\tau_k \quad \text{as} \quad n\to\infty, \tag{EC.37}$$

by virtue of the continuous mapping theorem and the first passage function. We have used different notation in §EC.1. We have $\tau_k = T_{k-1}+U_k+\tau_k^{(4)}$ for $\tau_k^{(4)}$ in (EC.11).

We can again apply essential the same argument to establish the weak convergence of the processes $(\bar{I}_n,\bar{Y}_n,\bar{X}_n)$ in (EC.35) and (EC.36) to $(\bar{I},\bar{Y},\bar{X})$ in the k^{th} down time interval $[T_{k-1}+U_k,T_k)$, where the processes $(\bar{I},\bar{Y},\bar{X})$ are defined by

$$\bar{I}(t) = \bar{I}(T_{k-1}+U_k) - \mu_2\bar{F}_k(t\wedge\tau_k-(T_{k-1}+U_k)) - \theta_3\int_{T_{k-1}+U_k}^{t\wedge\tau_k}\bar{I}(s)ds, \tag{EC.38}$$

$$\bar{Y}(t) = \bar{Y}(T_{k-1}+U_k) + \lambda(t-(T_{k-1}+U_k)) - \theta_2 \int_{T_{k-1}+U_k}^{t} (\bar{Y}(s) - \bar{F}_k)^+ ds, \qquad \text{(EC.39)}$$

for $t \in [T_{k-1}+U_k, \tau_k)$, where $\bar{I}(T_{k-1}+U_k) = (\bar{X}(T_{k-1}+U_k) \wedge 1) - (\bar{X}(T_k+U_k) \wedge \bar{F}_k)$, $\bar{Y}(T_k+U_k) = \bar{X}(T_k+U_k) - \bar{I}(T_{k-1}+U_k)$ and

$$\bar{X}(t) = \bar{X}(\tau_k) + \lambda(t-\tau_k) - \mu_2 \int_{\tau_k}^{t} (\bar{X}(s) \wedge \bar{F}_k) ds - \theta_2 \int_{\tau_k}^{t} (\bar{X}(s) - \bar{F}_k)^+ ds, \qquad \text{(EC.40)}$$

for $t \in [\tau_k, T_k)$, where $\bar{X}(\tau_k) = \bar{Y}(\tau_k)$. As noted above, $\tau_k = T_{k-1} + U_k + \tau_k^{(4)}$ for $\tau_k^{(4)}$ in (EC.11), in the case of only $I(t)$ decreasing during a down interval in §EC.1. That completes the proof.

References