USING DIFFERENT RESPONSE-TIME REQUIREMENTS TO SMOOTH TIME-VARYING DEMAND FOR SERVICE

by

Ward Whitt[1]

AT&T Labs

September 9, 1997

Revision: November 20, 1998

Operations Research Letters 24 (1999) 1–10

[1]AT&T Labs, Room A117, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971; email: wow@research.att.com

Abstract

Many service systems have demand that varies significantly by time of day, making it costly to provide sufficient capacity to be able to respond very quickly to each service request. Fortunately, however, different service requests often have very different response-time requirements. Some service requests may need immediate response, while others can tolerate substantial delays. Thus it is often possible to smooth demand by partitioning the service requests into separate priority classes according to their response-time requirements. Classes with more stringent performance requirements are given higher priority for service. Lower capacity may be required if lower-priority-class demand can be met during off-peak periods. We show how the priority classes can be defined and the resulting required fixed capacity can be determined, directly accounting for the time-dependent behavior. For this purpose, we exploit relatively simple analytical models, in particular, $M_t/G/\infty$ and deterministic offered-load models. The analysis also provides an estimate of the capacity savings that can be obtained from partitioning time-varying demand into priority classes.

Keywords: time-varying demand, smoothing time-varying demand, priority queues, nonstationary queues, deterministic fluid models, infinite-server queues, offered-load models

1. Introduction

In many service systems, demand varies significantly by time of day; e.g., see Hall [5]. If a fixed capacity is used to meet all demand, then it is natural to focus on peak demand and use busy-hour engineering or something similar in order to determine the appropriate capacity. We then may do steady-state analysis over the busy hour; see Massey and Whitt [10] and references cited there. Alternatively, if capacity too can be made time-dependent, as with service agents in telephone call centers, then it is natural to do time-dependent staffing, as in Jennings, Massey, Mandelbaum and Whitt [8] and references cited there.

Whether or not capacity can be made time-dependent, having significant variation of demand over time can be costly. If the capacity must be fixed, then a much higher capacity may be needed to meet peak demand than average demand. Even if capacity can be adjusted, it is often quite costly to do so. For example, it may not be easy to rapidly change the number of service agents in telephone call centers, because the service agents may be required to have minimum-length shifts. The expense of adjusting the work force and other aspects of production capacity is well established in the literature; e.g., see Holt, Modigliani, Muth and Simon [6].

Hence, it is natural to consider ways to alter the time-varying demand. A familiar method is time-varying pricing, such as lower telephone rates during off-peak hours. The demand over time can often be smoothed (leveled) by having lower prices during periods of otherwise low-demand; e.g., see Daniel [1].

Instead of altering the time of the service request, an alternative way to smooth demand is to alter the time that service must be provided. The service provider might offer different response times for different prices. Even without pricing, customers may have very different response-time requirements. For example, some calls to police require immediate response, while others do not. Thus, it is natural to partition the demand according to the response-time requirement. If such partitioning is not initially evident, then it may be possible to induce it by an appropriate pricing policy.

Concrete applications are copying and message (e.g., fax) communication services. The service providers might offer immediate service (e.g., within 10 minutes, in part depending on the job length), two-hour service and next-day service, each at successively lower prices. The options are somewhat like regular mail and express mail, but in our context a major motivation for introducing the different response-time guarantees is to smooth out daily demand.

The service provider may well have sufficient capacity to meet daily demand but insufficient capacity to provide immediate service to all service requests. The service provider could just serve everybody on a first-come first-served basis, but greater benefit — both profit to the service provider and customer satisfaction — is likely when demand is partitioned, because there should be significant differences among customer response-time requirements; i.e., it may be possible to provide immediately service when immediate service is really important, once we identify those cases.

We implement our partitioning scheme by assigning priorities to the different classes. Service requests from higher-priority classes are met before service requests from lower-priority classes. We let service be on a first-come first-served (FCFS) basis within each priority class. We think of the priority discipline as being preemptive-resume (i.e., upon arrival of a higher-priority job, the higher-priority job preempts a lower-priority job in service, and later the lower-priority job resumes service where it left off), but we do not dwell on that detail. The appropriate priority scheme may depend on the application.

Of course, there already is a large literature on priority queues, e.g., see Jaiswal [7] and Chapter 3 of Kleinrock [9], no doubt motivated in part by the situations that we consider. However, this literature, like most of the queueing literature more generally, does *not* discuss models with time-dependent arrival rates. The present paper is part of a long-term effort to develop methods for analyzing queueing models with time-dependent arrival rates. (The references give other examples.)

In this paper we directly model the time-dependent offered load. With the time-dependent offered load specified, we indicate how the different service classes can be defined and how a constant capacity can be set. We apply ideas in Jennings et al. [8] and in Duffield and Whitt [2], so we will be brief here. However, there are interesting new ideas. As before, a main idea is to simplify the analysis by focusing on offered load instead of carried load and associated delays and/or lost demand due to blocking. We use a normal approximation to approximate the distribution of the offered load. As in Duffield and Whitt [2], we also use a deterministic approximation, which may be the time-dependent mean offered load, to approximate the delays associated with having insufficient capacity. That is, we use deterministic methods to approximately describe the build up of queues, and the associated delays, when demand exceeds capacity. That part of our approach corresponds to a deterministic fluid approximation and thus follows Newell [11]; also see Hall [5] and Schmidt, Hoeflin and Skoog [14]. A deterministic fluid model tends to be appropriate when the aggregate demand is made up of a relatively large number of small demands, as is likely to occur in a message communication service. A seminal contribution on deterministic fluid approximations was Oliver and Samuel [12]. Segal [15] has recently proposed a mathematical programming approach with a discrete-time deterministic model to determine a good order of service (queue discipline) in each period.

Here is how the rest of this paper is organized: In Section 2 we show how capacity can be set to meet peak demand and daily demand. Both methods use a normal approximation to describe the demand. In Section 3 we consider the case in which the service requests are partitioned into two classes, one with a immediate-service requirement and the other with a daily-service requirement. In Section 4 we show how the analysis in Section 3 should be modified when the two kinds of work are associated with a common service request, and thus come together in a single arrival process.

In Section 5 we show how a deterministic fluid model can be used to treat other classes with intermediate-response-time requirements, in between immediate and daily. Our use of a deterministic fluid model to solve the problem of meeting a response-time constraint is different from our previous normal approximations. For models such as $M_t/M/c$ with a specified capacity c, the time-dependent queue-length distribution could be computed exactly, e.g., see Taaffe and Ong [16], Ong and Taaffe [13] and Jennings et al. [8]. Ong and Taaffe [13] also develop methods for computing the time-dependent waiting-time distribution in this model with the FCFS discipline. That approach might be extended to priority classes, but it has not yet been. So far, there evidently have not yet been any methods developed to compute the time-dependent waiting-time distribution with priorities. Even with the deterministic fluid approximation proposed here, the waiting-time formula associated with any specified capacity in (5.15) is somewhat complicated (but still computationally feasible). Hence we also give simple lower and upper bounds on the maximum waiting time in Section 6. Finally, in Section 7 we state our conclusions.

2. Peak Demand and Daily Demand

Let $A(t)$ denote the number of arrivals (service requests) in the interval $[0, t]$. We assume that the arrival process $\{A(t) : t \geq 0\}$ is a nonhomogeneous Poisson process with (deterministic) arrival rate $\lambda(t)$ at time t. Let S_n be the service requirement of the n^{th} arriving customer. We assume that $\{S_n : n \geq 1\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with a general cumulative distribution function (cdf) G, i.e.,

$P(S_n \leq t) = G(t)$.

There are two natural models for the offered load. The first counts the service requirement the instant it arrives, while the second counts the service requirement spread over time. In the first model, the total demand (offered load) in the interval $[0, t]$ is

$$D(t) = \sum_{i=1}^{A(t)} S_i, \quad t \geq 0 , \tag{2.1}$$

which has mean $\Lambda(t)E[S]$, where S is a generic service time and $\Lambda(t) = \int_0^t \lambda(u)du$. We call the derivative $\lambda(t)E[S]$ the offered load at time t. The first model tends to be appropriate for a single-server model, such as might represent a communication link with fixed available bandwidth (output rate) c serving messages or packets in a first-come-first-served (FCFS) order.

In the second model, the offered load is the number of busy servers in an $M_t/G/\infty$ service system with arrival process $A(t)$ and service times S_n. The idea is that, with ample capacity, the n^{th} service request would start service immediately upon arrival and be in service for a duration S_n after arrival. This second model is appropriate for the situation in which multiple services are being performed concurrently. The capacity c in the actual system then is represented by the number of servers. We shall work with the second model, but our methods apply to both models. We shall also use the first model later.

With the $M_t/G/\infty$ model for offered load, the number of busy servers at each time t has a Poisson distribution with mean

$$m(t) = \int_{-\infty}^{t} G^c(t-u)\lambda(u)du = E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t-S_e)]E[S] , \tag{2.2}$$

where $G^c(t) \equiv 1 - G(t)$ is the complementary cdf and S_e is a random variable with the service-time stationary-excess distribution, i.e.,

$$G_e(t) \equiv P(S_e \leq t) = \frac{1}{E[S]} \int_0^t G^c(u)du ; \tag{2.3}$$

e.g., see Eick, Massey and Whitt [3].

We could also extend the model to allow the service requirement of each customer to be for a random number of servers for the specified holding time, but we do not. With this extension, the total demand at each time has a compound Poisson distribution, i.e., it is distributed as the random (Poisson) sum of the random-server requirements. With or without this generalization, we approximate the offered load distribution at each time by a normal distribution, as in Jennings et al. [8]. In the extension, we need to calculate the variance of the compound Poisson distribution, which is a standard calculation; e.g., see Chapter 12 of Feller [4] and (2.6) below. When the distribution is Poisson, the time-dependent variance equals the mean $m(t)$.

We could also allow non-Poisson (G_t) arrival processes (with time-dependent rates) and time-dependent service-time cdf's, but we do not. See Jennings et al. [8] for ways to treat these extensions.

To select the capacity to meet peak demand, we consider the time of day, say t^*, where the mean offered load $m(t)$ in (2.2) is maximum. Following Section 3 of Jennings et al. [8], we let the number of servers, s, be

$$s = \lceil m(t^*) + 0.5 + z_\alpha \sqrt{m(t^*)} \rceil , \tag{2.4}$$

where α is the target delay probability (probability of having to wait before beginning service), $P(N(0,1) \geq z_\alpha) = \alpha$ with $N(0,1)$ being a standard (mean 0, variance 1) normal random variable, and $\lceil x \rceil$ is the least integer greater than or equal to x. For refinements to this estimate to account for having only finitely many servers, see Section 4 of Jennings et al. [8]. Also see Jennings et al. [8] for comparisons with numerical results and simulations that justify (2.4).

If instead we elected to use the model of offered load in (2.1), then we could still set capacity by (2.4), but using the modified mean $\lambda(t)E[S]$, which turns out to be the pointwise stationary approximation (PSA) for the infinite-server mean; see Eick et al. [3], Massey and Whitt [10] and references therein.

In order to see the potential benefits of smoothing demand over the day, we next consider the capacity required to meet daily demand (any time over the day), and compare that capacity requirement with the greater capacity required to meet peak demand. Suppose that we measure time in minutes and that a day consists of a full 24 hours $= T$ minutes, where $T = 1440$. We also assume that the starting point of the day can be wherever we choose, in order to avoid end effects. Then the total number of arrivals during a day, $A(T)$, has a Poisson distribution with mean

$$\lambda_T = \int_0^T \lambda(t) dt . \tag{2.5}$$

Ignoring end effects, the total demand over a day is the random sum $D(T)$ for $D(t)$ in (2.1). That is, when we consider demand over a full day, we may count the service requirement the instant it arrives. Let a generic service time S have mean m and variance σ^2. Then $D(T)$ has a compound Poisson distribution with mean and variance

$$ED(T) = \lambda_T m \quad \text{and} \quad VarD(T) = \lambda_T(\sigma^2 + m^2) \tag{2.6}$$

for λ_T in (2.5). Like the time-dependent mean $m(t)$ in (2.2), but unlike the stationary demand where $\lambda(t)$ has been replaced by the average demand $\bar{\lambda} \equiv \lambda_T/T$, the variance of the total demand in (2.6) depends on the service-time cdf G beyond its mean m.

Now we consider the number of servers, say s_T, needed to meet all demand over a day. (Again we ignore end effects.) The total capacity over a day can be regarded as Ts_T, the length of the day multiplied by the number of servers. As before, we use a normal approximation. Thus, paralleling (2.4), we let

$$Ts_T \approx ED(T) + z_{\hat{\alpha}}\sqrt{VarD(T)} + 0.5 \ , \tag{2.7}$$

where $\hat{\alpha}$ is the target probability of meeting the daily demand and $P(N(0,1) \geq z_{\hat{\alpha}}) = \hat{\alpha}$ or, more precisely,

$$s_T = \left\lceil \frac{ED(T) + z_{\hat{\alpha}}\sqrt{VarD(T)} + 0.5}{T} \right\rceil \ . \tag{2.8}$$

We might be more demanding in a daily requirement than a peak requirement, so that we could have $\hat{\alpha}$ substantially less than α, but still we may have s_T substantially less than s in (2.4). A rough estimate (lower bound) for s_T is $E[D(T)]/T$. When $m(t^*) >> E[D(T)]/T$, there is potential for substantial gain by replacing a peak-demand constraint with a daily-demand constraint.

3. Partitioning Demand Into Two Classes

We now suppose that the total demand can be partitioned into two independent classes, part of which must be met immediately and part of which must be met on a daily basis. Thus the total Poisson arrival process can be decomposed into two independent Poisson arrival processes, and the total arrival rate $\lambda(t)$ is divided into two components: an immediate-service demand $\lambda_i(t)$ and a daily-service demand $\lambda_d(t)$, i.e.,

$$\lambda(t) = \lambda_i(t) + \lambda_d(t), \quad t \geq 0 \ . \tag{3.1}$$

This partitioning may be achieved by introducing two different prices. These two prices might even be time-dependent. Before the prices are introduced, there may be considerable uncertainty about the resulting partitioning. Indeed, the sum of the rates after introducing the new prices need not equal the rate before the prices are set. We do not address the forecasting issue here and we do not study the effect of price upon demand. We assume that the two demand rates $\lambda_i(t)$ and $\lambda_d(t)$ are known. The service-time distributions may depend on the classes too. Hence we assume that service-time random variables S_i and S_d with associated cdf's G_i and G_d are available.

We can now determine the capacity required to meet peak demand. It is the capacity in (2.4), using an appropriate target α, where $m(t)$ in (2.2) is calculated using $\lambda_i(t)$ and G_i instead of $\lambda(t)$ and G. If the offered load model is (2.1), then instead of (2.4) we would replace $m(t)$ by $\lambda(t)E[S]$. As noted earlier, this is a convenient simple approximation for (2.4) as well.

We call the capacity required to meet immediate demand s_i. Note that the daily demand requirement $\lambda_d(t)$ and S_d play no role in setting s_i. We must also specify the target probability of experiencing delay α for the immediate-delivery class. We should have $\lambda_i(t)$ substantially less than $\lambda(t)$, so that s_i should be substantially less than s, yielding the benefit of the approach.

We now turn to the capacity needed to meet daily requirements. We now consider the total daily demand for both classes, $D_i(T)$ and $D_d(T)$, respectively. We compute each as in (2.1), using the defining variables for that class. Then, assuming that the two kinds of demand are independent, we add the means and variances to get the total mean and variance; i.e., by (2.6), the mean and variance of the total daily demand are

$$ED(T) = \lambda_{iT} m_i + \lambda_{dT} m_d \tag{3.2}$$

and

$$VarD(T) = \lambda_{iT}(\sigma_i^2 + m_i^2) + \lambda_{dT}(\sigma_d^2 + m_d^2) \ . \tag{3.3}$$

We then let the capacity required to meet daily demand be s_T as in (2.8) using an appropriate target delay probability $\hat{\alpha}$ and (3.2) and (3.3). We call this capacity s_d. Finally, we let the overall capacity be the maximum of the two capacities s_i and s_d.

If there is a significant difference between s_i and s_d, then there will be excess capacity for one class. With prices, it is natural to consider price adjustments to make the two capacities s_i and s_d nearly equal. Alternatively, prices could be set so that revenue is maximized subject to the constraints that both kinds of demand are met. For that, though, we need to know how the demand rates depend on the prices. We have provided a framework wherein that can be studied.

4. A Common Arrival Process

In the previous section we considered the case in which the two kinds of service requests arrive in separate independent streams. Now we consider the case in which the two kinds of work are associated with a single service request, arriving together in a single arrival process with rate $\lambda(t)$. The idea now is that part of the required work should be done immediately, while part can be delayed. For example, taking an order from a customer may need to be done immediately, but it may be possible to process the order afterwards. In general, interacting with the customer and providing immediately needed service typically must be done immediately, but writing reports can be delayed.

When the two kinds of work arrive in a common arrival process, there is no change in the analysis of high-priority work. However, there is a change in the analysis of the requirement to meet daily demand. Assuming that $\lambda_i(t) = \lambda_d(t)$, the previously calculated mean in (2.5) and (3.2) is all right, but the variance in (3.3) should be calculated differently.

For the daily demand, it is natural to have a single arrival process with arrival rate $\lambda(t)$. Then the daily demand has mean and variance just as in (2.6), using the total mean λ_T in (2.5), where the mean m and variance σ^2 of the service time apply to both kinds of work, i.e.,

$$m = E(S_i + S_d) \tag{4.1}$$

and

$$\sigma^2 = Var(S_i + S_d) \ , \tag{4.2}$$

where S_i and S_d are the service times associated with the two kinds of work. Now we allow the two components of work to be correlated, so that

$$\sigma^2 = \sigma_i^2 + \sigma_d^2 + 2\,Cov(S_i, S_d) \ , \tag{4.3}$$

where $Cov(S_i, S_d)$ is the covariance. Even if the two components of work are uncorrelated, the variance formula is different from (3.3). Now the mean is $\lambda_T m$ and the variance is

$$Var\, D(T) = \lambda_T(\sigma^2 + m^2) = \lambda_T(\sigma_i^2 + \sigma_d^2 + 2\,Cov(S_i, S_d) + m_i^2 + m_d^2 + 2m_i m_d) \ . \tag{4.4}$$

With the revised variance formula (4.4), we can choose the capacity using (2.7) and (2.8), just as before.

5. Other Demand Classes with Response-Time Limits

We now consider how we can introduce other classes with response-time requirements in between immediate service and daily service. For example, we might want to have a class for which service should be provided within two hours of any request.

For simplicity, we only consider three classes, but more can be treated in essentially the same way. In addition to the immediate and daily demand classes with arrival rates $\lambda_i(t)$ and $\lambda_d(t)$, suppose that we have a class with response-time constraint (e.g., delay less than or equal to r minutes) having arrival-rate function $\lambda_r(t)$. Now, instead of (3.1), we assume that the total arrival rate can be partitioned into three components, i.e.,

$$\lambda(t) = \lambda_i(t) + \lambda_r(t) + \lambda_d(t), \quad t \geq 0 \ . \tag{5.1}$$

Paralleling Section 3, we also have service-time variables S_i, S_r and S_d.

To treat the immediate-service class we do just as in Section 3 (or Section 4, if they arrive in a common arrival process), considering only class i. To treat the response-time requirement, we combine the i and r classes, letting the time-dependent mean $m(t)$ be the sum of the two separate means, each computed as in (2.2). (The sum of two independent Poisson variables is again Poisson.) We indicate how to determine the required capacity s_r below. To treat the daily requirement we combine all three classes, in the same way we combined two classes in Section 3 (or Section 4, if they arrive in a common arrival process). We let the final required capacity be the maximum of the three determined capacities s_i, s_r and s_d.

To determine the new response-time capacity requirement s_r, we consider a single (aggregate) class (containing classes i and r). We suppose that the time-dependent mean offered load $m(t)$ has been computed. We show how to determine if any fixed capacity is adequate to meet the new response-time request. The required capacity to meet the new response time requirement, s_r, is the minimum fixed capacity such that the response-time constraint is met. We now use a different approach. We use a simple deterministic fluid model to estimate the response time (as a function of time), given a specification of capacity and demand. For simplicity, we assume that capacity is a constant c, but the reasoning also applies to time-dependent deterministic capacity.

As in Section 8 of Duffield and Whitt [2] and in previous deterministic fluid model analyses, e.g., in Newell [11] and Hall [5], we simplify the delay analysis by assuming that the demand (offered load) is a deterministic function of time $\delta(t)$. The idea behind this approximation is that the fluctuations over time should be more important than the stochastic fluctuations. In this deterministic model, we assume that work arrives deterministically and continuously at rate $\delta(t)$ and is processed deterministically and continuously at rate c. In reality, both input and output may be random. Then $\delta(t)$ and c are both deterministic approximations.

The first natural value for $\delta(t)$ is simply the time-dependent mean $m(t)$, which we have assumed is available. To be conservative, allowing for stochastic fluctuations, we may alternatively let the deterministic demand be inflated by some standard deviations, i.e.,

$$\delta(t) = m(t) + z_\alpha \sqrt{m(t)} \tag{5.2}$$

for some α, where $P(N(0,1) > z_\alpha) = \alpha$, as in (2.4), but not necessarily the same α.

If we use the alternative model for offered load in (2.1) and (2.6), then (5.2) would be replaced by

$$\delta(t) = ED(t) + z_\alpha \sqrt{Var\, D(t)} \tag{5.3}$$

where

$$ED(t) = \lambda_i(t) m_i + \lambda_d(t) m_d \tag{5.4}$$

and

$$Var\, D(t) = \lambda_i(t)(\sigma_i^2 + m_i^2) + \lambda_d(t)(\sigma_d^2 + m_d^2)\ . \tag{5.5}$$

Similar changes should be made in the setting of Section 4.

We also impose simplifying technical regularity conditions. We assume that the deterministic demand $\delta(t)$ is piecewise smooth, i.e., it is differentiable except for only finitely many jump discontinuities. We also assume that it is right-continuous with left limits. These restrictions evidently impose no practical limitations.

We start by determining the queue length $q(t)$ of work remaining to be processed at time t. We get queues and delays because $\delta(t) > c$ for some t. We only consider a single day. Over a longer period, we are thinking of approximately periodic behavior over successive days. Let the average demand over a day be

$$\bar{\delta} = \frac{1}{T} \int_0^T \delta(t) dt\ . \tag{5.6}$$

We assume that $\bar{\delta} < c$, so that demands do not build up over successive days. We also assume that $\delta(t) > c$ for some t, so that there are delays.

Given the deterministic demand $\delta(t)$ and capacity c, we can determine the (deterministic) time-dependent queue length $q(t)$. Assuming that the day starts with an empty queue, the queue becomes positive for the first time at time

$$t_1 = \min\{t \geq 0 : \delta(t) > c\}\ . \tag{5.7}$$

The queue then remains positive until time

$$t_2 = \min\left\{t > t_1 : \int_{t_1}^{t} (\delta(u) - c)du = 0\right\} . \tag{5.8}$$

When there is a single period of high demand, as in Figure 1, the queue remains positive until the time t_2 at which the area below the curve $\delta(t) - c$ in the subsequent period of low demand (the shaded region in Figure 1) equals the total area above the curve during the period of high demand, which begins at t_1 (the striped region in Figure 1). Indeed,

$$q(t) = 0, \quad 0 \le t \le t_1 \tag{5.9}$$

and

$$q(t) = \int_{t_1}^{t} (\delta(u) - c)du > 0, \quad t_1 < t < t_2 . \tag{5.10}$$

More generally, there may be multiple periods of high demand. (The case of two high-demand periods is depicted in Figure 2.) For $k \ge 0$, recursively define the times by setting $t_0 = 0$,

$$t_{2k+1} = \min\{t \ge t_{2k} : \delta(t) > c\} \tag{5.11}$$

and

$$t_{2k+2} = \min\left\{t \ge t_{2k+1} : \int_{t_{2k+1}}^{t} (\delta(u) - c)du = 0\right\} . \tag{5.12}$$

Then

$$q(t) = 0, \quad t_{2k} < t < t_{2k+1} \tag{5.13}$$

and

$$q(t) = \int_{t_{2k+1}}^{t} (\delta(u) - c)du > 0, \quad t_{2k+1} < t < t_{2k+2} . \tag{5.14}$$

The next step is to determine an associated estimate of delay experienced by a class-r arrival at time t. It is important to note that this estimate should depend on the priority structure, because a class-r arrival must wait, not only for all the work that has accumulated at time t, but also for all higher-priority work that arrives after time t (before the high-priority piece of work can complete service). The remaining work in the system is just $q(t)$. The available capacity at time u after time t is $[c - \delta_i(u)]^+$, where $[x]^+ = \max\{0, x\}$. Here $\delta_i(t)$ represents the total time-dependent demand of all classes having higher priority than r. (With only the three classes i, r and d, this is just $\delta_i(t)$.) Hence, using the fluid model, the waiting time before a class-r arrival at time t can begin service is

$$w(t) = \inf\left\{u \ge 0 : q(t) = \int_{t}^{t+u} [c - \delta_i(v)]^+ dv\right\} . \tag{5.15}$$

Graphically, we can calculate $q(t)$ for any t by plotting $\delta(u)$ for $u \leq t$, and we can calculate $w(t)$ given $q(t)$, by plotting $\delta_i(u)$ for $u \geq t$. In words, $w(t)$ is the time u such that the area under the curve $[c - \delta_i(v)]^+$ over the interval $(t, t+u)$ first equals $q(t)$.

For convenience, it is natural to use discrete-time approximations to compute both $q(t)$ and $w(t)$. Then we would assume that there are demands δ_k at evenly spaced time points t_k; e.g., we might let

$$\delta_k \equiv \int_{(t_{k-1}+t_k)/2}^{(t_k+t_{k+1})/2} \delta(u) du \ . \tag{5.16}$$

The integrals in (5.6), (5.8), (5.10), (5.12), (5.14) and (5.15) would then be replaced by sums.

From (5.15), we can compute the waiting time for a set of representative time points and determine the maximum waiting time, say $w_{\max}$. We say that a response time limit r is met by capacity c if $w_{\max}(c) \leq r$. We let the minimum capacity c such that $w_{\max}(c) \leq r$ be the required capacity s_r. We can find s_r by performing a search over candidate capacities, employing (5.15) in each case. The search is facilitated by the fact that $q(t)$, $w(t)$, and $w_{\max}$, are all decreasing in c; e.g., we can use bisection search.

It is also of interest to see how sensitive the required capacity s_r is to key parameters. We can see the effect of random fluctuations by repeating the analysis after changing the target delay probability α in (5.2) and (5.3). We can also consider changes in the other parameters. By calculating the required capacity s_r as a function of the other parameters, we can see the sensitivity of s_r to the other parameters.

6. Bounds on the Maximum Delay

Since the time-dependent delay in (5.15) is somewhat complicated, it may be useful to have bounds that can serve as convenient simple approximations. A lower bound is obtained by acting as if the delay at time t for class r is just the time required to process the workload at time t, as if the FCFS discipline were in effect for the classes under consideration. Then the maximum delay becomes the maximum queue length divided by c. For this lower bound, the waiting times are maximized when the queue lengths are maximized. Clearly, if the lower bound delay is too large, then so will be the actual delay in (5.15).

The local maxima of the queue length process $\{q(t) : t \geq 0\}$ are those points t at which the demand crosses the capacity from above, i.e., t for which $\delta(t-) \geq c$ while $\delta(t) < c$. We now define these times. With $u_0 = 0$, let u_k be defined recursively for $k \geq 1$ by

$$u_k = \min\{t : u_{k-1} < t \leq T, \quad \delta(u_k-) \geq c, \quad \delta(u_k) < c\} \ , \tag{6.1}$$

where K is the number of such time points, which we assume is finite. Then the maximum queue length, say $q_{\max}$, is the maximum of $q(u_k)$ over k, $1 \leq k \leq K$. An example with two high-demand periods is shown in Figure 2.

We now specify an upper bound on the delay. For it, we act as if the new arrival at time t is a small quantity in a separate lowest-priority class. Then we must wait until the first time after t at which the queue is empty. This upper bound is attained by considering the particles initiating busy periods. The upper-bound waiting time is the sum of the lower-bound waiting time associated with the last particle in the busy period and the length of the busy period. For example, in Figure 2, the lower bound is the maximum of $t_2 - u_1$ and $t_4 - u_2$, while the upper bound is the maximum of $t_2 - t_1$ and $t_4 - t_3$.

These simple bounds enable us to bound the capacities that need to be considered when we seek the minimum required capacity s_r. The bounds themselves may sometimes serve as adequate approximations, e.g., when there is considerable uncertainty about the model parameters. Since both lower and upper bounds are available, their difference provides a bound on the error from using these approximations. It is significant that the upper (lower) bound tends to be correct when the proportion of class-r demand among all class-r or higher demand, $\pi_r(t) \equiv \lambda_r(t)/(\lambda_i(t) + \lambda_r(t))$, tends to be consistently near 0 (1). This is evident because the assumptions underlying the bounds are then satisfied. This limiting behavior suggests an approximation based on both bounds when $\pi_r(t) \approx \pi_r$ for all t (or only those t yielding large congestion), namely,

$$w_{\max} \approx \pi_r w_{\max}^L + (1 - \pi_r) w_{\max}^U \; , \qquad (6.2)$$

where $w_{\max}^L$ and $w_{\max}^U$ are the lower and upper bounds.

7. Conclusions

Our analysis shows how service requests can be partitioned into priority classes according to response-time requirements, and how the capacity needed to meet all requirements can be determined. The highest priority class is for immediate response. We use an $M_t/G/\infty$ model and a normal approximation, with a target probability that a service request will have to be delayed before beginning service, to determine the capacity required for this highest priority class (Section 2).

The lowest priority is for demand to be met over the course of a day (or any other chosen suitably long period). We again use a normal approximation with another target small probability that daily demand will all be met (Section 2). In Section 3 we showed how all service requests can be partitioned into two classes, one for immediate response and one for daily response. The required capacity is then the maximum of the two requirements. In Section 4 we showed how to modify the analysis of daily demand to represent the case in which both types of work come in a common arrival process.

In Section 5 we showed how to treat other classes with the requirement that the response time for each request (before beginning service) be less than some specified value r. For any given response-time limit r, we combine the demand of that class with all higher priority classes to determine a deterministic approximation for the time-dependent queue length $q(t)$. We then use a deterministic fluid model to determine the capacity required to meet this response-time requirement. For the deterministic fluid model, the required capacity is the minimum capacity such that the maximum waiting time over the day is less than or equal to r. To compute the waiting time, we must take account of the priority structure. The final formula is (5.15).

Thus, there is a required capacity generated for each priority class, each involving all higher priority classes. As the priority level decreases, the constraint loosens (because the limit r increases), but also the relevant offered load increases (because the set of service requests that are included grows), so that it is not a priori evident which capacity requirement will be dominant. The final required capacity is the maximum of all the required capacities for the individual priority classes.

The priority classification should be relatively efficient when the individual requirements are not too different. When there are significant differences, then there is excess capacity available for some of the classes at the overall capacity limit. When there are significant differences, it is natural to consider modifications in the priority-class definitions to balance the capacity requirements.

Finally, we can compare alternative priority schemes by comparing their total capacity requirements. We can see how much more capacity is required if all requests are given immediate service (often a big difference). We can see how much less capacity is required if all requests are given only daily service (often a little difference). The analytical models provide a convenient quick rough-cut analysis. The proposed results can be confirmed and refined by computer simulations and, after deployment, by system measurements.

Acknowledgment. I am grateful to my colleague Marc Goldring for helpful comments.

References

[1] J. L. Daniel, Congestion pricing and optimal capacity of large hub airports: a bottleneck model with stochastic queues, *Econometrica* 63 (1993) 327–370.

[2] N. G. Duffield and W. Whitt, Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems* 26 (1997) 69–104.

[3] S. G. Eick, W. A. Massey and W. Whitt, The physics of the $M_t/G/\infty$ queue, *Oper. Res.* 41 (1993) 731–742.

[4] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. I, third edition, Wiley, New York, 1968.

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

[6] C. F. Holt, J. Modigliani, J. Muth and H. Simon, *Planning Production, Inventories and Work Force*, Prentice Hall, Englewood Cliffs, NJ, 1960.

[7] N. K. Jaiswal, Priority Queues, Academic Press, New York, 1968.

[8] O. B. Jennings, A. Mandelbaum, W. A. Massey and W. Whitt, Server staffing to meet time-varying demand, *Management Sci.* 42 (1996) 1383–1394.

[9] L. Kleinrock, *Queueing Systems, vol. II: Computer Applications*, Wiley, New York, 1976.

[10] W. A. Massey and W. Whitt, Peak congestion in multi-server service systems with slowly varying arrival rates, *Queueing Systems* 25 (1997) 157–172.

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

[12] R. M. Oliver and A. H. Samuel, Reducing letter delay in post offices, *Oper. Res.* 10 (1962) 839–892.

[13] K. L. Ong and M. R. Taaffe, Nonstationary queues with interrupted Poisson arrivals and unreliable/repairable servers, *Queueing Systems* 4 (1989) 27–46.

[14] D. C. Schmidt, D. A. Hoeflin and R. A. Skoog, Simple models of complex transient phenomena, *Teletraffic Contributions for the Information Age*, *Proceedings of ITC 15*, V. Ramaswami and P. W. Wirth (eds.), Elsevier, Amsterdam, 1997, pp. 953–964.

[15] M. Segal, Delays in throttled store and forward traffic: a flow model, AT&T Labs, (1997).

[16] M. R. Taaffe and K. L. Ong, Approximating $Ph(t)/M(t)/S/C$ queueing systems, *Ann. Oper. Res.* 8 (1987) 103–116.