Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues

Ward Whitt

AT&T Bell Laboratories, Murray Hill, NJ 07974-0636, USA

Received August 1992; in final form August 1993

Our primary purpose in this paper is to contribute to the design of admission control schemes for multi-class service systems. We are motivated by emerging high-speed networks exploiting asynchronous transfer mode (ATM) technology, but there may be other applications. We develop a simple criterion for feasibility of a set of sources in terms of "effective bandwidths". These effective bandwidths are based on asymptotic decay rates of steady-state distributions in queueing models. We show how to compute asymptotic decay rates of steady-state queue length and workload tail probabilities in general infinite-capacity multi-channel queues. The model has *m* independent heterogeneous servers that are independent of an arrival process which is a superposition of *n* independent general arrival processes. The contribution of each component arrival process to the overall asymptotic decay rates can be determined from the asymptotic decay rates produced by this arrival process alone in a G/D/1 queue (as a function of the arrival rate). Similarly, the contribution of each service process to the overall asymptotic decay rates can be determined from the asymptotic decay rates produced by this service process alone in a D/G/1 queue. These contributions are characterized in terms of single-channel asymptotic decay-rate functions, which can be estimated from data or determined analytically from models. The asymptotic decay-rate functions map potential decay rates of the queue length into associated decay rates of the workload. Combining these relationships for the arrival and service channels determines the asymptotic decay rates themselves. The asymptotic decay-rate functions are the time-average limits of logarithmic moment generating functions. We give analytical formulas for the asymptotic decay-rate functions of a large class of stochastic point processes, including batch Markovian arrival processes. The Markov modulated Poisson process is a special case. Finally, we try to put our work in perspective with the related literature.

1. Introduction and summary

A fundamental problem in the design of multi-service communication networks is *admission control*; see Roberts [50] for a broad overview. Network providers would like to develop a good way in which to decide whether or not to satisfy each successive connection request. Due to the diverse traffic that may be carried by these networks, the different sources can be expected to make very different demands on the network. Consequently, it is natural to seek an appropriate notion of *effective bandwidth* associated with each kind of source, so that n sources with "bandwidths" α_i, $1 \le i \le n$, can be regarded as feasible if and only if $\sum_{i=1}^{n} \alpha_i \le C$, where C is the *capacity*, i.e. the total available "bandwidth". (See section 12 for a discussion of related literature.) The network provider might admit each prospective new connection whenever feasible or sometimes elect to reserve space for "more desirable" future connections, as with *trunk reservation* in circuit-switched networks; see Mitra et al. [42]. We are motivated by potential applications to multi-service communication networks, but obviously other applications are possible.

THE EFFECTIVE BANDWIDTH CONCEPT

In developing an appropriate notion of effective bandwidth, it is natural to go beyond deterministic rates (peak or average) and consider the resulting congestion at network resources. We consider the critical network resource to be a switch, but again, other application are possible. To analyze the congestion at a single switch, it is natural to use a $\sum G_i/D/1$ queueing model, which has one server with deterministic service times, unlimited (as an approximation to large finite) waiting space and an arrival process that is the result of statistical multiplexing, i.e. it is the superposition of independent general stationary arrival processes, each of which may be quite "bursty". This model is difficult to analyze exactly, because there are typically many component arrival processes, but not nearly so many that the superposition process will be nearly Poisson.

Our approach is to focus on the *steady-state* behavior of the queue. If sources come and go relatively slowly compared to the way congestion changes, this steady-state view may be appropriate. However, if sources come and go relatively quickly, it may be much better to base admission decisions on a transient analysis reflecting the present state of congestion in the system. Here, we consider only steady-state analysis, ignoring the current congestion.

We primarily base our analysis of the $\sum G_i/D/1$ queue, and related models with more general service processes, on asymptotic analysis of steady-state tail probabilities. First, however, it seems appropriate to discuss further the general notion of effective bandwidth. In the context of the $\sum G_i/D/1$ queue, we represent the available bandwidth C as a maximum level of some form of congestion. Then we want α_i to be the level of congestion due to source i. However, *a basic property of queueing models is that the level of congestion caused by any one source depends strongly on the other sources present.* This phenomenon is well illustrated by the formula $\rho/(1-\rho)$ for mean number in system in an M/M/1 queue; the effect of increasing ρ (e.g. the derivative of $\rho/(1-\rho)$) increases as ρ increases.

It seems reasonable to hope that α_i can perhaps be regarded, at least approximately, as an *increasing function of the total level of congestion.* It thus seem reasonable to hope that the congestion x associated with n independent sources with *congestion function* α_i, $1 \le i \le n$, would be the unique fixed point of the equation

$$\sum_{i=1}^{n} \alpha_i(x) = x. \tag{1.1}$$

An implication of (1.1) as a general relation is that *any n independent sources with congestion functions α_i should be equivalent to a single source with congestion function $\sum_{i=1}^{n} \alpha_i$*. This property is desirable, but it is not necessarily possible to have it in practice. We will discuss a situation where this property does in fact hold.

In this context, a way to get around the non-constancy of the function α_i is to compute $\alpha_i(C)$ at the prescribed critical congestion level C. It is easy to see that n sources with individual congestion functions α_i are feasible with a total capacity C if and only if

$$\sum_{i=1}^{n} \alpha_i(C) \le C. \tag{1.2}$$

It is important to note that if there is approximately an equality in (1.2) with n given sources, then $\sum_{i=1}^{k} \alpha_i(C)$ may be only a very crude upper bound on the congestion with only the first k sources when $k < n$. Nevertheless, the notion of effective bandwidth based on (1.2) may be very useful.

In this paper, as measures of congestion, we primarily focus on the steady-state tail probabilities of the queue length (number in system) and the workload (virtual waiting time), but it is important to note that the above reasoning applies much more generally. We illustrate this generality in section 11 by discussing effective bandwidths in the context of mean workload instead of tail probabilities. For the multi-service communication networks using the emerging asynchronous transfer mode (ATM) technology, it seems natural to develop performance criteria in terms of tail probabilities because they are natural surrogates for the very small cell blocking probabilities contemplated for ATM networks. However, the most appropriate criterion may be different in other applications.

ASYMPTOTICS FOR STEADY-STATE TAIL PROBABILITIES

In great generality, the tail probabilities of the steady-state queue length and workload are asymptotically geometric and exponential, respectively, and these limits are often good approximations for times that are only moderately large as well as very large; see Tijms [58], Abate et al. [1–3], and references cited in those sources.

In particular, let Q and W denote the steady-state queue length and workload at an arbitrary time. We often have

$$\sigma^{-k} P(Q > k) \to \beta \qquad \text{as } k \to \infty \tag{1.3}$$

and

$$e^{\eta x} P(W > x) \to \alpha \qquad \text{as } x \to \infty, \tag{1.4}$$

where β, σ, α and η are positive constants, independent of k and x. In this context, it is natural to represent the capacity C in terms of the $(100p)$th percentiles (e.g. $p = 1 - 10^{-9}$ is currently of interest), which are approximately given by

$$q_p \equiv \inf\{k : \beta\sigma^k < 1 - p\} \approx \frac{\log(1-p) - (\log\beta)}{\log\sigma} \approx \frac{-\log(1-p)}{1-\sigma} \tag{1.5}$$

for p, σ and β suitably near 1, and

$$w_p = \inf\{x : \alpha e^{-\eta x} < 1 - p\} \approx \frac{-\log(1-p) + \log\alpha}{\eta} \approx \frac{-\log(1-p)}{\eta} \tag{1.6}$$

for p and α suitably near 1, and η suitably near 0. Note that in (1.5) and (1.6), attention is focused on the asymptotic decay rates σ and η, and here we shall restrict attention to these simple parameters.

Given that we will focus on the asymptotic decay rates σ and η in (1.3)–(1.6), it is natural to consider even weaker (more general) limits, namely,

$$\frac{\log P(Q > k)}{k} \to \log\sigma \quad \text{as } k \to \infty \tag{1.7}$$

and

$$\frac{\log P(W > x)}{x} \to -\eta \quad \text{as } x \to \infty. \tag{1.8}$$

Limits of the form (1.7) and (1.8) are naturally considered in the context of large deviations theory; see Bucklew [10], and Dembo and Zeitouni [18]. Limits of the form (1.7) and (1.8) are established in Chang [11], Chang et al. [12], and Glynn and Whitt [26,27].

Our analysis of steady-state tail probabilities is based on the idea that the simple percentile approximations in (1.5) and (1.6) are sufficiently accurate for engineering purposes. It is thus important to note that *such approximations might actually not be sufficiently accurate.* The very small target cell blocking probabilities in the range 10^{-9} that are contemplated for ATM networks suggest that the small-tail asymptotics in (1.3)–(1.6) should be excellent. However, surprisingly, we have developed algorithms for BMAP/GI/1 queues with superposition arrival processes and made numerical comparisons of the approximations with exact values that show that approximations (1.5) and (1.6) can be quite inaccurate for some models with multiple sources; see Choudhury et al. [14,15]. The difficulty is that the asymptotic constants α and β in (1.3) and (1.4) can be very different from 1. This can be explained by the fact that the asymptotic constants α and β in (1.3) and (1.4) tend to be asymptotically exponential in the number of sources when the arrival rate is fixed (achieved by scaling the component arrival processes). The asymptotic constants become small (large) as the number of sources increases when the component arrival processes are more (less) bursty than a Poisson process; see [14]. Here, we assume that approximations (1.5) and (1.6) are in fact satisfactory. In applications of these ideas, this should be checked.

A GENERAL MULTI-CHANNEL MODEL

As a general model for establishing the limits (1.7) and (1.8) and developing an associated concept of effective bandwidths, we propose the multi-channel queue considered in Iglehart and Whitt [31,32]. In this model, there is unlimited waiting room and m heterogeneous servers that operate independently of the arrival process. The arrival process is the superposition of n independent component arrival processes. The number of arrivals in channel i in the time interval $[0, t]$ is $A_i(\lambda_i t)$, $1 \le i \le n$. The parameter λ_i represents the *arrival rate*, so that $A_i \equiv \{A_i(t) : t \ge 0\}$ is understood to be a *rate-1 process*. Customers are assigned to the first available server in order of arrival, with some specified procedure to break ties. (We conjecture that the tie-breaking procedure does not affect the asymptotic decay rates, and we do not focus on it.) The number of service completions by server j during the first t units of time that server j is busy is $S_j(\mu_j t)$, $1 \le j \le m$. The parameter μ_j is the (potential) *service rate* of server j, so that $S_j \equiv \{S_j(t) : t \ge 0\}$ is understood to be a *rate-1 process*. We assume that

$$\lambda \equiv \sum_{i=1}^{n} \lambda_i < \sum_{j=1}^{m} \mu_j = 1, \tag{1.9}$$

so that proper steady-state distributions should exist; we *assume* that they do.

The model we have just described is a conventional queueing model with discrete customers, which might be cells in an ATM network. However, the processes A_i and S_j *need not be integer-valued.* The analysis here applies to general nondecreasing processes (and even more generally), including the fluid models considered by Anick et al. [5], and Elwalid and Mitra [19] and more general ones.

The first main idea is to allow the stochastic point processes A_i and S_j to be very general, just as for the heavy-traffic limits in [31,32]. *We do not directly make any independence, Markov or stationarity assumptions about each process.* We do assume that the $m + n$ processes A_i, $1 \le i \le n$, and S_j, $1 \le j \le n$, are *mutually independent*, but even this is not necessary. If we do not have this independence, we can start with the assumed asymptotic behavior of the "net input" process $\sum_{i=1}^{n} A_i(\lambda_i t) - \sum_{j=1}^{m} S_j(\mu_j t)$ (see section 5). The independence property does play a vital role, however. It allows us to determine the effect of each channel by analyzing it separately.

ASYMPTOTIC DECAY-RATE FUNCTIONS

The second main idea here is that the asymptotic behavior in (1.7) and (1.8) should depend on the limiting behavior of the logarithmic moment generating functions of A_i, $1 \leq i \leq n$, and S_j, $1 \leq j \leq m$. In particular, we define the *single-channel asymptotic decay-rate functions*

$$\psi_{A_i}(\theta) = \lim_{t\to\infty} t^{-1} \log E e^{\theta A_i(t)}, \quad 1 \leq i \leq n, \tag{1.10}$$

and

$$\psi_{S_j}(\theta) = \lim_{t\to\infty} t^{-1} \log E e^{\theta S_j(t)}, \quad 1 \leq j \leq m. \tag{1.11}$$

We assume that these limits exist for all relevant θ (see below). The idea is that the asymptotic decay rate σ in (1.7) is $e^{-\zeta}$, where ζ is the unique positive root of the *queue-length decay-rate equation*

$$\delta(\theta) \equiv \lambda\psi_A(\theta) + \psi_S(-\theta) = 0, \tag{1.12}$$

with

$$\lambda\psi_A(\theta) \equiv \sum_{i=1}^{n} \lambda_i \psi_{A_i}(\theta) \quad \text{and} \quad \psi_S(\theta) \equiv \sum_{j=1}^{m} \mu_j \psi_{S_j}(\theta). \tag{1.13}$$

In other words, the second idea is to apply large deviations theory, as in Bucklew [10] and Dembo and Zeitouni [18]. Supporting theory is developed in Glynn and Whitt [26,27]. Important contribution to both the first two ideas was made by Chang [11].

A significant feature of (1.12) and (1.13) is the *linearity*. The decay rate σ depends on the component processes $A_1, \ldots, A_n$ and $S_1, \ldots, S_m$ only via the composite asymptotic decay-rate functions $\lambda\psi_A$ and ψ_S in (1.13). Hence, the superposition of n i.i.d. arrival processes $\{A_i(\lambda_i t) : t \geq 0\}$, $1 \leq i \leq n$, has the same asymptotic decay-rate function as $\{A_1(n\lambda_1 t) : t \geq 0\}$, the one component arrival process with n times its original rate, i.e. $\sum_{i=1}^{n} \lambda_i \psi_{A_i} = n\lambda_1 \psi_{A_1}$.

The associated asymptotic decay rate η in (1.8) is

$$\eta = \sum_{i=1}^{n} \lambda_i \psi_{A_i}(\zeta) = -\sum_{j=1}^{m} \mu_j \psi_{S_j}(-\zeta), \tag{1.14}$$

where ζ is the unique positive root of (1.12). Alternatively, η can be determined directly as the unique positive root of the *workload decay-rate equation*

$$\lambda\psi_A(-\psi_S^{-1}(-\theta)) = \theta. \tag{1.15}$$

Given η from (1.15) and (1.13), we can find the queue-length asymptotic decay rate by inverting the relations in (1.14); i.e. $\sigma = e^{-\zeta}$, where

$$\zeta = -\psi_S^{-1}(-\eta) = \psi_A^{-1}(\eta/\lambda). \tag{1.16}$$

We hasten to point out that we have *not yet proved* that in full generality the asymptotic behavior of Q and W in (1.7) and (1.8) actually is determined by the decay-rate equations (1.12) and (1.15). However, there is strong supporting evidence, some of which we present here. In particular, this procedure has been *proved to be correct* for the special case in which all component arrival processes A_i and service processes S_j are phase-type (PH) renewal processes in theorem 5 of Neuts [46]. (The representation in [46] is different but equivalent.) One purpose of this paper is to show that the paper [46] has important applications to the effective bandwidth problem. Another purpose is to significantly generalize [46]. We provide supporting theory in [3,26,27]. However, in general, the procedure in (1.12)–(1.16) remains a mathematical conjecture. Nevertheless, we think that this can serve as a useful engineering principle.

EFFECTIVE BANDWIDTH BASED ON ASYMPTOTIC DECAY-RATE FUNCTIONS

We now indicate how the asymptotic decay-rate functions and the decay-rate equations (1.12) and (1.15) can be used to create effective bandwidths. First, we suppose that system capacity is specified in terms of a percentile q_p of the steady-state queue-length distribution as in (1.5). For a given p, this percentile bound determines an upper limit on the asymptotic decay rate σ^* via (1.5). From the perspective of the queue-length decay-rate equation (1.12), $\zeta^* = -\log\sigma^*$. We assume that the service processes are fixed. For *given service processes* S_j with rates μ_j, $1 \le j \le m$, we work with the asymptotic decay-rate function ψ_S in (1.13). For any arrival process A_i with rate λ_i, we let its *effective bandwidth* be $\eta_i^* \equiv \lambda_i \psi_{A_i}(-\log\sigma^*)$. We then say n such arrival processes are *feasible* if and only if

$$\sum_{i=1}^{n} \eta_i^* \equiv \sum_{i=1}^{n} \lambda_i \psi_{A_i}(-\log\sigma^*) \le -\psi_S(-\zeta^*) \equiv -\psi_S(\log\sigma^*) \equiv \eta^*, \qquad (1.17)$$

where η^* is the workload asymptotic decay rate associated with the service process S and the critical queue-length asymptotic decay rate σ^*, as specified in (1.14). In other words, for (1.2) we let the capacity be $C = -\psi_S(\log\sigma^*)$ and let the effective bandwidth be $\alpha_i(C) = \lambda_i\psi_{A_i}(-\log\sigma^*)$. Analysis of (1.12) shows that the actual asymptotic decay rate associated with the n sources A_i is less than or equal to σ^* if and only if (1.17) holds.

Moreover, since $0 < \sigma^* < 1$ and $\psi_S(0) = 0$, we see that the capacity $-\psi_S(\log\sigma^*)$ is positive and increasing in σ^*. In addition, the effective bandwidths η_i^* in (1.16) are positive and increasing in both λ_i and σ^*.

Alternatively, we can focus on the workload asymptotic decay rate η in (1.8). Starting with a capacity constraint expressed in terms of a percentile w_p of the steady-state workloads, we obtain a constraint on η via (1.6). Now a collection of arrival processes is feasible if its workload asymptotic decay rate η satisfies $\eta \ge \eta^*$. Using (1.15), we let the capacity be $\eta^* = -\psi_S(-\zeta^*)$ and we let the effective bandwidths be $\eta_i^* = \lambda_i\psi_{A_i}(-\psi_S^{-1}(-\eta^*)) = \lambda_i\psi_{A_i}(\zeta^*)$, where $e^{-\zeta^*}$ is the queue-length asymptotic decay rate associated with η^* via (1.14) and (1.16). We then obtain the feasibility condition

$$\sum_{i=1}^{n} \lambda_i \psi_{A_i}(\zeta^*) \equiv \sum_{i=1}^{n} \lambda_i \psi_{A_i}(-\psi_S^{-1}(-\eta^*)) \le \eta^* \equiv -\psi_S(-\zeta^*), \tag{1.18}$$

which in fact is identical to (1.17).

It is easy to see why we obtain the same feasibility equation starting with σ^* or η^*, assuming that these are consistent. Given ψ_S and σ^*, η^* is determined by (1.14); i.e. we obtain $\eta^* = -\psi_S(\log \sigma^*)$. Alternatively, given ψ_S and η^*, σ^* is determined by $\sigma^* = \exp(\psi_S^{-1}(-\eta^*))$.

ORGANIZATION OF THE REST OF THIS PAPER

In the remainder of this paper, we explain and (partly) justify the decay-rate equations (1.12)–(1.16), which are the basis for the effective-bandwidth feasibility criterion in (1.17) and (1.18). Here is how the rest of this paper is organized. We begin by discussing an illustrative elementary M/M/1 example in section 2. We then establish basic properties of the asymptotic decay-rate functions in (1.10) and (1.11) in section 3. In section 4, we relate the asymptotic decay-rate functions of a counting process and its inverse partial sum process. In particular, we show that these asymptotic decay-rate functions are themselves, in some sense, inverse functions of each other.

In section 5, we present a simple queueing model for which we can justify the key equations (1.12)–(1.16) above. In section 6, we indicate how to calculate the asymptotic decay-rate functions for a large class of stochastic processes. In section 6, we consider Markov renewal processes and batch Markovian arrival processes (BMAPs) as in Lucantoni [37,38]. These classes include ordinary renewal processes and Markov modulated Poisson processes (MMPPs) as special cases. Also in section 6, we present additional theoretical support for the key equations (1.12)–(1.16).

In section 7, we discuss an alternative way to obtain the asymptotic decay-rate functions from the asymptotic decay rates actually observed in test queues. Also in section 7, we point out that the asymptotic decay-rate functions can be regarded as relations between the two decay rates σ and η in (1.3) and (1.4). In section 8, we discuss heavy-traffic asymptotic expansions for the asymptotic decay rates, which lead to simple approximations for the effective bandwidths. This analysis also establishes important connections to heavy-traffic theory, such as in Iglehart and Whitt [31,32].

In section 9, we show how the asymptotic decay rates can be used to determine how many servers are needed, given a congestion constraint and a fixed arrival process. This is in some sense the dual of the admission control problem. In section 10, we discuss numerical methods that can be used to calculate the asymptotic parameters in (1.3) and (1.4).

In section 11, we show how effective bandwidths can be determined in the framework of (1.2) with a criterion based on the mean steady-state workload instead of percentiles of tail probabilities. Finally, in section 12 we review related literature and describe the history of our involvement with this problem.

2. The M/M/1 example

To see what is going on, it is helpful to consider a simple example. Hence, we consider the $\Sigma_{i=1}^{n} \mathrm{M}_i/\mathrm{M}/1$ queue, which has an arrival process consisting of n independent Poisson processes. Since the superposition of independent Poisson processes is again a Poisson process, this is in fact just the M/M/1 queue. Consider n independent component Poisson arrival processes with rates λ_i, $1 \le i \le n$, where $\Sigma_{i=1}^{n} \lambda_i = \lambda < 1$, and exponential servive times with mean 1, so that (1.9) holds. Note that the M/M/1 queue can be constructed from a potential service process S, which is also Poisson and independent of the arrival process. (The specific construction is discussed in section 5.)

In this case, the processes A_i and S are all Poisson, so that the asymptotic decay-rate functions in (1.10) and (1.11) are easily computed; they are

$$\psi_{A_i}(\theta) = \psi_S(\theta) = \mathrm{e}^{\theta} - 1. \tag{2.1}$$

Note that

$$\lambda \psi_A(\theta) \equiv \sum_{i=1}^{n} \lambda_i \psi_{A_i}(\theta) = \lambda(\mathrm{e}^{\theta} - 1); \tag{2.2}$$

i.e. ψ_A has the same form of ψ_{A_i}, as it should since they all are rate-1 Poisson processes.

Next note that the queue-length decay-rate equation (1.12) here is

$$\lambda(\mathrm{e}^{\theta} - 1) + (\mathrm{e}^{-\theta} - 1) = 0. \tag{2.3}$$

After making the change of variables $z = \mathrm{e}^{\theta}$, we see that (2.3) is easily solved: equation (2.3) has the two roots $\theta = 0$ and $\theta = -\log \lambda$. Since $0 < \lambda < 1$, $\zeta = -\log \lambda$ is the unique positive root of (2.3). Since $\sigma^{-1} = \mathrm{e}^{\zeta}$, $\sigma = \lambda = \rho$, where ρ is the traffic density. Since Q has the geometric distribution $P(Q = k) = (1 - \rho)\rho^k$, $k \ge 0$, the asymptotic decay rate σ is indeed ρ.

From (1.14), we can calculate the associated workload asymptotic decay rate η; we obtain

$$\eta = \lambda(\mathrm{e}^{-\log \lambda} - 1) = -(\mathrm{e}^{\log \lambda} - 1) = 1 - \lambda. \tag{2.4}$$

Alternatively, we can apply (1.15). For this purpose, we need to compute the inverse function ψ_S^{-1}. From (2.1), we see that

$$\psi_S^{-1}(\theta) = \log(1 + \theta) \quad \text{and} \quad -\psi_S^{-1}(-\theta) = -\log(1 - \theta). \tag{2.5}$$

Hence, the workload decay-rate equation in (1.15) becomes

$$\lambda(e^{-\log(1-\theta)} - 1) = \lambda\left(\frac{1}{1-\theta} - 1\right) = \theta. \tag{2.6}$$

Equation (2.6) has the two roots $\theta = 0$ and $\theta = 1 - \lambda$. Since $0 < \lambda < 1$, $\zeta = 1 - \lambda$ is the unique positive root to (2.6). Since $P(W > x) = \rho e^{-(1-\rho)x}$, the asymptotic decay rate is indeed $\eta = 1 - \rho = 1 - \lambda$.

Now consider the effective bandwidths. Suppose that an upper bound σ^* has been specified for the queue-length asymptotic decay rate σ. From (1.17) and (2.1), we see that the capacity is

$$C = -\psi_S(\log \sigma^*) = 1 - \sigma^* \equiv \eta^* \tag{2.7}$$

and the effective bandwidth of source i is

$$\alpha_i(C) = \lambda_i \psi_{A_i}(-\log \sigma^*) = \lambda_i\left(\frac{1}{\sigma^*} - 1\right) = \lambda_i \frac{(1-\sigma^*)}{\sigma^*}. \tag{2.8}$$

From (2.7) and (2.8), we see that the n sources are feasible with capacity $1 - \sigma^*$ in (2.7) if and only if

$$\sum_{i=1}^{n} \lambda_i \leq \sigma^*, \tag{2.9}$$

which of course is easy to verify directly for this simple example.

As indicated above, the approximations based on the limits in (1.3) and (1.4) are *exact* for all x and k in the M/M/1 model; i.e. $\sigma^{-k}P(Q > k) = \beta$ for all k and $e^{\eta x}P(X > x) = \alpha$ for all x, where $\alpha = \beta = \lambda = \rho$, $\sigma = \lambda$ and $\eta = 1 - \lambda$. As a consequence, the effective-bandwidth procedure for admission control tends to work very well for the $\sum_{i=1}^{n} M_i/M/1$ model considered in this section. As shown in [14], poor performance can occur when the sources deviate substantially from Poisson processes.

3. Properties of the asymptotic decay-rate functions

We have seen that the asymptotic decay rates σ and η in (1.3)–(1.8) can be characterized as roots of equations (1.12) and (1.15) involving the single-channel asymptotic decay-rate functions ψ_{A_i} and ψ_{S_j} defined in (1.10) and (1.11). There may be no root, in which case the theory does not apply. (See example 5 of [1].) However, there always is *at most one root*, so that when we succeed in finding a root, the decay rates are unambiguously determined. The uniqueness follows from convexity. Simple algorithms for finding the root when there is one also follow from convexity.

For any random variable X, let $\psi_X(\theta)$ be the logarithm of its moment generating function, i.e.

$$\psi_X(\theta) = \log E e^{\theta X}. \tag{3.1}$$

Obviously, it is possible to have $E e^{\theta X} = +\infty$, in which case $\psi_X(\theta) = +\infty$; we will be concerned with $\psi_X(\theta)$ in the region where it is finite. The *logarithmic moment generating function* ψ_X in (3.1) is also called the *cumulant generating function* of X; see p. 20 of Johnson and Kotz [33]. It plays a key role in large deviations theory; e.g. see p. 26 of Dembo and Zeitouni [18]. Here, and in the large deviations theory more generally, it is important that ψ_X is a convex function. This is an easy consequence of Hölder's inequality; e.g. see p. 47 of Chung [17].

PROPOSITION 1

For any real-valued random variable X, the logarithmic moment generating function $\psi_X(\theta)$ in (3.1) is a convex function of θ in the region where it is finite.

Proof

Choose p such that $0 < p < 1$. Apply Hölder's inequality to conclude that

$$\begin{aligned}\psi_X(p\theta_1 + (1-p)\theta_2) &= \log(E e^{p\theta_1 X} e^{(1-p)\theta_2 X}) \\ &= \log\left(\left(E e^{p\theta_1 X \frac{1}{p}}\right)^p \left(E e^{(1-p)\theta_2 X \frac{1}{1-p}}\right)^{1-p}\right) \\ &\le p \log E e^{\theta_1 X} + (1-p) \log E e^{\theta_2 X} \\ &= p\psi_X(\theta_1) + (1-p)\psi_X(\theta_2). \qquad \square\end{aligned}$$

Now let $Z \equiv \{Z(t) : t \ge 0\}$ be any real-valued stochastic process. As a stochastic process generalization of (3.1), let ψ_Z be the *limiting time-average of the cumulant generating functions*; i.e. let

$$\psi_Z(\theta) \equiv \lim_{t\to\infty} t^{-1} \log E e^{\theta Z(t)}, \tag{3.2}$$

assuming that the limit exists.

We call ψ_Z the *asymptotic decay-rate function* of the stochastic process Z. Let (3.2) also apply to discrete-time processes. Then we can let t run through the positive integers. Alternatively, we can let $Z(t) = Y_{\lfloor t \rfloor}$ for some sequence $\{Y_n\}$, where $\lfloor t \rfloor$ is the greatest integer less than or equal to t.

With (3.2), we are thinking of processes that have stationary increments or asymptotically stationary increments, so that $\psi_Z(\theta)$ is typically nondegenerate for a range of θ of interest. For example, we show that (3.2) is well defined if the process Z has stationary and independent increments. For a discrete-time process, this is just partial sums of i.i.d. random variables. The bounding condition below is then unnecessary.

PROPOSITION 2

If the stochastic process Z has stationary and independent increments and $Ee^{\theta Z(s)} \le M < \infty$, $0 \le s \le 1$, then

$$\psi_Z(\theta) = \psi_{Z(1)}(\theta) \equiv \log Ee^{\theta Z(1)}. \tag{3.3}$$

Proof

Let $\lfloor t \rfloor$ be the greatest integer less than or equal to t. By the stationary and independent increments property,

$$Ee^{\theta Z(t)} = (Ee^{\theta Z(1)})^{\lfloor t \rfloor} Ee^{\theta Z(t - \lfloor t \rfloor)}.$$

Hence,

$$\log Ee^{\theta Z(t)} = \lfloor t \rfloor \log Ee^{\theta Z(1)} + \log Ee^{Z(t - \lfloor t \rfloor)}. \tag{3.4}$$

Since

$$\log Ee^{\theta Z(t - \lfloor t \rfloor)} \le \log M,$$

the desired conclusion (3.3) follows easily from (3.4). □

It is significant that ψ_Z in (3.2) is also a convex function whenever the limit exists, because the limit of a sequence of convex function is convex.

PROPOSITION 3

If the limit in (3.2) exists, then the function ψ_Z is convex.

When considering integer-valued random variables and processes (e.g. counting processes), it is often natural to consider *logarithmic generating functions* $\log Ez^X$, which is also known as the *factorial cumulant generating function*; see p. 21 of Johnson and Kotz [33]. By making a change of variables, we obtain the logarithmic moment-generating function and the desired convexity.

We now continue to establish basic properties of asymptotic decay-rate functions. In particular, we establish a calculus for computing new asymptotic decay-rate functions from given ones. Chang [11] first developed this calculus for upper bounds. This calculus parallels the mapping arguments for constructing new functional limit theorems from given ones in Whitt [60]. We begin with an elementary *superposition* result.

PROPOSITION 4 (SUPERPOSITION)

If $Z_1, \ldots, Z_n$ are independent stochastic processes with well-defined asymptotic decay-rate functions, then the superposition process $Z_1 + \ldots + Z_n$ has a well-defined asymptotic decay-rate function

$$\psi_{Z_1 + \ldots + Z_n} = \psi_{Z_1} + \ldots + \psi_{Z_n}.$$

Proof

By the independence,

$$Ee^{\theta[Z_1(t)+\ldots+Z_n(t)]} = \prod_{i=1}^{n} Ee^{\theta Z_i(t)}.$$

Hence,

$$\log Ee^{\theta[Z_1(t)+\ldots+Z_n(t)]} = \sum_{i=1}^{n} \log Ee^{\theta Z_i(t)}. \qquad \square$$

We now turn to *composition*, i.e. a random time change of one stochastic process by another. Let $(Z_1 \circ Z_2)(t) = Z_1(Z_2(t))$, $t \geq 0$, with the understanding that the values of $Z_2(t)$ are contained in the time domain of Z_1. The following is an extension of results in Chang [11], and Glynn and Whitt [26].

THEOREM 5 (COMPOSITION)

Let Z_1 and Z_2 be independent processes with $\psi_{Z_1}(\theta)$ well defined in a neighborhood of $\hat{\theta}$ and $\psi_{Z_2}(\theta)$ well defined in a neighborhood of $\psi_{Z_1}(\hat{\theta})$. In addition, suppose that $Ee^{\theta_1 M_1(t)} < \infty$ and $Ee^{\theta_2 M_2(t)} < \infty$ for all t and θ_1 in a neighborhood of $\psi_{Z_2}(\hat{\theta})$ and θ_2 in a neighborhood of $\hat{\theta}$, where $M_i(t) = \sup\{Z_i(s) : 0 \leq s \leq t\}$. Then, $\psi_{Z_1 \circ Z_2}(\hat{\theta})$ is well defined and

$$\psi_{Z_1 \circ Z_2}(\hat{\theta}) = (\psi_{Z_2} \circ \psi_{Z_1})(\hat{\theta}) = \psi_{Z_2}(\psi_{Z_1}(\hat{\theta})). \tag{3.5}$$

Proof

First, condition on the value of $Z_2(t)$ to obtain

$$Ee^{\hat{\theta} Z_1(Z_2(t))} = \int Ee^{\hat{\theta} Z_1(x)}\, dP(Z_2(t) \leq x).$$

Now apply (3.2) to Z_1 and Z_2 to deduce that for any ε there is a t_0 such that

$$\begin{aligned} Ee^{\hat{\theta} Z_1(Z_2(t))} &\leq \int e^{x(\psi_{Z_1}(\hat{\theta})+\varepsilon)}\, dP(Z_2(t) \leq x) + Ee^{\hat{\theta} M_1(t_0)} \\ &\leq e^{t\psi_{Z_2}(\psi_{Z_1}(\hat{\theta})+\varepsilon)+\varepsilon} + Ee^{(\psi_{Z_1}(\hat{\theta})+\varepsilon)M_2(t_0)} + Ee^{\hat{\theta} M_1(t_0)} \end{aligned}$$

for all t suitably large. Hence,

$$\overline{\lim_{t\to\infty}}\, t^{-1} \log Ee^{\hat{\theta} Z_1(Z_2(t))} \leq \psi_{Z_2}(\psi_{Z_1}(\hat{\theta}) + \varepsilon) + \varepsilon.$$

Since ε was arbitrary and ψ_2 is continuous at $\psi_1(\hat{\theta})$,

$$\overline{\lim_{t\to\infty}}\, t^{-1} \log Ee^{\theta Z_1(Z_2(t))} \le \psi_2(\psi_{Z_1}(\hat{\theta}))\,.$$

The reasoning in the other direction is essentially the same. □

It is significant that the order of Z_1 and Z_2 is *reversed* in (3.5), i.e. $\psi_{Z_1 \circ Z_2} = \psi_{Z_2} \circ \psi_{Z_1}$. An important special case is when one of the functions is constant. For this purpose, let e be the *identity function*, i.e. $e(t) = t$ for $t \ge 0$. It is elementary that

$$\psi_{\lambda e}(\theta) = \lambda\theta. \tag{3.6}$$

If we think of ψ as a map taking stochastic processes Z with time parameter set $[0, \infty)$ into real-valued functions $\psi_Z(\theta)$ with argument set $(-\infty, \infty)$, formula (3.6) says that ψ maps λe into λe.

The following special case of theorem 5 is elementary.

PROPOSITION 6

Assuming that ψ_Z is well defined at the relevant arguments,

$$\psi_{Z\circ\lambda e}(\theta) = (\psi_{\lambda e} \circ \psi_Z)(\theta) = \lambda\psi_Z(\theta) \tag{3.7}$$

and

$$\psi_{\lambda Z} = \psi_{\lambda e \circ Z} = \psi_Z \circ \psi_{\lambda e} = \psi_Z(\lambda\theta). \tag{3.8}$$

Formula (3.7) states that a deterministic homogeneous change of time scale by λ in Z results in ψ_Z being multiplied by λ; formula (3.8) states that multiplying Z by λ corresponds to a deterministic homogenous change of time scale by λ in ψ_Z.

4. The asymptotic decay-rate functions of inverse processes

A key ingredient of most queueing models is an arrival process. An arrival process can be represented via the counting process $A \equiv \{A(t) : t \ge 0\}$ or its inverse partial sum process $U \equiv \{U_n : n \ge 1\}$. Then $A(t)$ represents the number of arrivals in the interval $[0, t]$, $t \ge 0$, while U_n represents the arrival epoch of customer n, $n \ge 1$. The fundamental inverse relation is

$$U_n \le t \text{ if and only if } A(t) \ge n. \tag{4.1}$$

In this section, we present the relation between the asymptotic decay-rate functions ψ_A and ψ_U of inverse processes A and U. In particular, ψ_A is closely related to the inverse of ψ_U; more precisely, we show that $\psi_A(\theta) = -\psi_U^{-1}(-\theta)$. This inverse relation was first stated in the original version of this paper, and first proved in Glynn and Whitt [27].

It is significant that the process A need not in fact be a counting process. It suffices to assume that A is nonnegative and nondecreasing. Then the inverse process can be defined by

$$U(t) = \inf\{s : A(s) > t\}, \quad t > 0. \tag{4.2}$$

For example, this extension is useful to cover fluid models.

Given that $A(t)$ and U_n are nonnegative for all arguments, the asymptotic decay-rate functions ψ_A and ψ_U must be *nondecreasing* as well as convex (proposition 3). Moreover, $\psi_A(0) = \psi_U(0) = 0$. Let

$$\beta_A^u = \sup\{\theta : \psi_A(\theta) < \infty\} \quad \text{and} \quad \beta_A^l = \sup\{\theta : \psi_A(\theta) = \psi_A(-\infty)\} \tag{4.3}$$

and similarly for β_U^u and β_U^l.

We refer to Glynn and Whitt [27] for a careful statement and proof of the following theorem. The important point is that where the functions are finite,

$$\psi_U(\theta) = -\psi_A^{-1}(-\theta) \quad \text{and} \quad \psi_A(\theta) = -\psi_U^{-1}(-\theta).$$

Possible inverse asymptotic decay-rate functions are displayed in the same graph in fig. 1. The function ψ_U appears in the usual position, while ψ_A increases to the left with its argument increasing downwards.

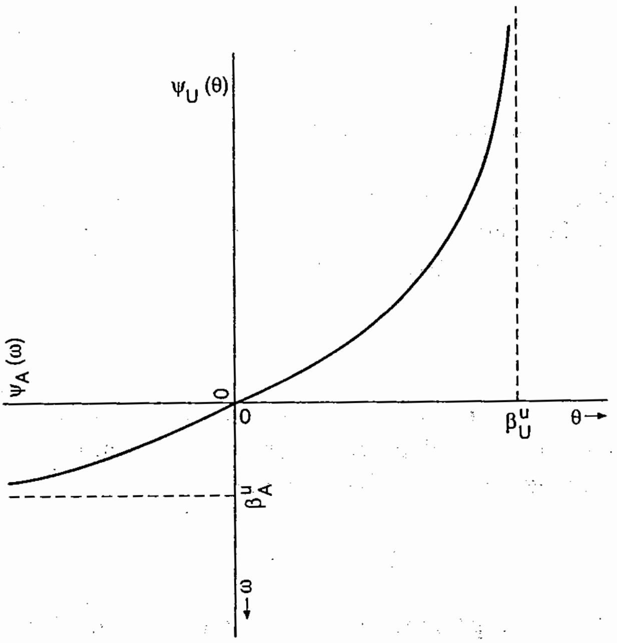

Fig. 1. Possible inverse decay-rate functions ψ_U and $\psi_A = -\psi_U^{-1}(-\cdot)$ with finite asymptotes β_U^u and β_A^u.

THEOREM 7 (GLYNN AND WHITT [27])

Under appropriate regularity conditions [27], the process A satisfies (3.2) with asymptotic decay-rate function ψ_A if and only if U satisfies (3.2) with asymptotic decay-rate function ψ_U, in which case

$$\psi_A(\theta) = \begin{cases} -\beta_U^u, & \theta < \beta_A^l = -\psi_U(\beta_U^u), \\ -\psi_U^{-1}(-\theta), & \beta_A^l \le \theta \le \beta_A^u, \\ +\infty, & \theta > \beta_A^u = -\psi_U(\beta_U^l) = -\psi_U(-\infty), \end{cases}$$

and

$$\psi_U(\theta) = \begin{cases} -\beta_A^U, & \theta < \beta_U^l = -\psi_A(\beta_A^u), \\ -\psi_A^{-1}(-\theta), & \beta_U^l \le \theta \le \beta_U^u, \\ +\infty, & \theta > \beta_U^u = -\psi_A(\beta_A^l) = -\psi_A(-\infty). \end{cases}$$

To illustrate, suppose that $A = \lambda e$ for $\lambda > 0$. Then, by (4.2), $U = \lambda^{-1}e$. By (3.2), $\psi_A = \lambda e$ and $\psi_U = \lambda^{-1}e$. Hence,

$$\psi_U(\theta) = \lambda^{-1}\theta = \psi_A^{-1}(\theta) = -\psi_A^{-1}(-\theta).$$

For a more interesting example, suppose that A is a Poisson counting process with rate λ as in section 2, so that $\psi_A(\theta) = e^{\theta} - 1$. Then U is the partial sum of n i.i.d. exponential random variables with mean λ^{-1}, so that $\psi_U(\theta) = -\log(1-\theta)$, $\theta < 1$. Note that indeed theorem 7 holds with $\beta_A^l = -\infty$, $\psi_A(-\infty) = -1 = -\beta_U^u$ and $\beta_A^u = -\beta_U^l = +\infty$.

These two examples are quite easy, because for them we can easily compute *both* ψ_A and ψ_U. The power of theorem 7 comes when we can easily compute only *one* of ψ_A and ψ_U. We illustrate this for renewal processes.

PROPOSITION 8

Let $A \equiv \{A(t) : t \ge 0\}$ be a renewal counting process associated with i.i.d. interrenewal times distributed as U_1. Then (3.2) is valid for A with asymptotic decay-rate function $\psi_A(\theta) = -\psi_U^{-1}(\theta)$, where

$$\psi_U(\theta) = \log Ee^{\theta U_1}.$$

We now combine proposition 4 and theorem 7 to describe the asymptotic decay rate of the arrival time sequence associated with a superposition arrival process.

PROPOSITION 9

Let $\{U_n : n \geq 1\}$ be the arrival time sequence associated with an arrival counting process $A = A_1 + \ldots + A_m$. If the processes A_i are mutually independent and satisfy (3.2) with asymptotic decay-rate function ψ_{A_i}, $1 \leq i \leq m$, and appropriate regularity conditions holds, then U satisfies (3.2) and its asymptotic decay-rate function is $\psi_U(\theta) = -\psi_A^{-1}(\theta)$, where $\psi_A = \psi_{A_i} + \ldots + \psi_{A_m}$.

We can now combine propositions 8 and 9 to obtain the asymptotic decay-rate function for the arrival time sequence in a superposition of independent renewal processes. If U is the arrival time sequence associated with A, where $A = A_1 + \ldots + A_m$ and A_i are mutually independent renewal processes with interrenewal times distributed as U_1^i, then

$$\begin{aligned} \psi_U(\theta) &= -\psi_A^{-1}(-\theta), \\ \psi_A(\theta) &= \psi_{A_1}(\theta) + \ldots + \psi_{A_m}(\theta), \\ \psi_{A_i}(\theta) &= -\psi_{U^i}^{-1}(-\theta), \quad 1 \leq i \leq m, \\ \psi_{U^i}(\theta) &= \log E e^{\theta U_1^i}, \quad 1 \leq i \leq m. \end{aligned} \tag{4.5}$$

This chain of reasoning closely parallels the approach to heavy-traffic limit theorems for queues with superposition arrival processes in Iglehart and Whitt [31,32].

5. A basic queueing model

In this section, we show how the asymptotic decay-rate functions discussed in sections 2–4 determine the asymptotic decay rates of steady-state queueing distributions. For this purpose, we consider a relatively simple queueing model. In particular, let $I(t)$ represent the total input (of work or customers) to a queue in the time interval $[0, t]$ and let $O(t)$ represent the total *potential* output from the queue in the interval $[0, t]$. In this section, we assume that the workload can be represented in terms of the *net input process*

$$N(t) = I(t) - O(t), \quad t \geq 0. \tag{5.1}$$

In particular, we assume that the workload at time t is simply the familiar *one-sided reflection map* applied to N, i.e.

$$W(t) = \max\{W(0) + N(t),\ N(t) - \inf_{0 \leq s \leq t} N(s)\}, \quad t \geq 0. \tag{5.2}$$

It is well known that if $W(0) = 0$ and the net input process N has stationary increments with $N(t) \to -\infty$ as $t \to \infty$ w.p.1, then $W(t)$ converges in distribution as

$t \to \infty$, a proper limit, say W; see section 6 of Borovkov [9]. We are interested in the asymptotic relation (1.8) for W.

First, however, we point out that it is by no means automatic to have $W(t)$ defined so simply in terms of the basic processes I and O. Formula (5.2) is valid for the queue-length process in the M/M/1 queue as in section 2 (so that theorem 10 below justifies section 2), but rarely more generally. However, this construction is valid quite generally for the workload (or virtual waiting time) process in the infinite-capacity single-server queue. Then $I(t)$ represents the sum of the service times of all arrivals in $[0, t]$, and $O(t) = t$. For the workload process, more general output processes may occur because the server operates in a random environment. Such a random environment may be due to occasional service interruptions, e.g. machine breakdowns. Then $O(t) - O(s)$ is still bounded above by $t - s$, a property we use in theorem 10 below.

For queue-length processes, this model occurs under the "autonomous service" assumption, see p. 235 of Borovkov [9], which means that the server keeps on working even if no work is present. For many queue-length processes, though, this representation serves only as a rough approximation. A *key idea is that the asymptotic decay rates evidently occur as if this model were valid.* (This can be proved in several cases, but remains to be fully proved.)

The following is a positive result in the context of this model. This is a minor generalization of theorem 4 of Glynn and Whitt [26], which in turn follows quite quickly from the corresponding discrete-time result in theorem 1 of [26], and Chang [11]. This discrete-time result generalizes the classical GI/GI/1 result in Smith [52], and on p. 269 of Asmussen [6] by replacing the independence assumption by a condition on the logarithmic moment-generating functions. For the discrete-time result, no upper bound on $O(\delta)$ is required below. Let $\Rightarrow$ denote convergence in distribution.

THEOREM 10 (CHANG [11], AND GLYNN AND WHITT [26])

Suppose that the net input process N has stationary increments and $EI(1) < EO(1) < \infty$. Also, suppose that $W(0) = 0$ and there is a constant M such that $O(\delta) < M$ w.p.1 for all sufficiently small δ. If there exists a function ψ_N and positive constants θ^* and ε^* such that

(i) $t^{-1} \log Ee^{\theta N(t)} \to \psi_N(\theta)$ as $t \to \infty$ for $|\theta - \theta^*| < \varepsilon^*$,

(ii) ψ_N is finite in a neighborhood of θ^* and differentiable at θ^* with $\psi_N(\theta^*) = 0$,

(iii) $t^{-1} \log Ee^{\theta^* N(t)} < \infty$ for all $t > 0$, then $W(t) \Rightarrow W$ as $t \to \infty$ and (1.8) holds, i.e.

$$x^{-1} \log P(W > x) \to -\theta^* \quad \text{as } x \to \infty .$$

For applications, the point of theorem 10 is that *the asymptotic decay rate η of the steady-state queueing variable W is determined by the asymptotic decay-rate* *function ψ_N of the net input process N.* In particular, η is the positive root of the equation $\psi_N(\theta) = 0$. Since $\psi_N(0) = 0$ and ψ_N is convex (proposition 3), there is at most one positive root to this equation. In theorem 10, it is significant that the input process I and the potential output process O *need not be independent*. However, if they are independent, then

$$\psi_N(\theta) = \psi_I(\theta) + \psi_O(-\theta) \tag{5.3}$$

and the key condition (ii) in theorem 10 is of the form (1.12).

We now want to apply theorem 10 to justify, at least heuristically, the queue-length and workload decay-rate equations (1.12) and (1.15), and the auxiliary equations (1.14) and (1.16). Moreover, we want to relate the workload decay-rate equation to a corresponding waiting-time decay-rate equation in the single-server model.

The general idea is that these decay-rate equations can be obtained from the net input process as described above, even though that simple model is often not precisely correct. For the queue-length process in the multi-channel model of section 1, the net input process (obtained by assuming that the servers work continuously) is $\sum_{i=1}^{n} A_i(\lambda_i t) - \sum_{j=1}^{m} S_j \mu_j(t)$. By virtue of the composition property in theorem 5, the asymptotic decay-rate function of this net input process is

$$\lambda\psi_A(\theta) + \psi_S(-\theta) = \sum_{i=1}^{n} \lambda_i \psi_{A_i}(\theta) + \sum_{j=1}^{m} \mu_j \psi_{S_j}, \tag{5.4}$$

just as in (1.12). Thus, assuming that the simple model applies, theorem 10 justifies (1.12). However, such a representation is not valid in general. It is valid, however, in the $\Sigma G_i/M/1$ queue and the discrete-time $\Sigma G_i/D/1$ queue, in which all arrivals occur at integer multiples of the deterministic service time. This last model is a natural candidate for ATM networks, and is in fact used by Chang [11] and Sohraby [53,54].

More generally, though, (1.12) represents a natural conjectured generalization of theorem 10. We conjecture that (1.12) is still valid even though the reflection map representation in section 5 does not hold. In the next section, we present additional support for (1.12).

Turning to the workload, we see that the net input process (again assuming that the servers work continuously) is the sum of all service times assigned to the arrivals in [0, t] minus t (assuming that the sum of the service rates is 1). In other words,

$$N(t) = V_{A(\lambda t)} - t, \quad t \geq 0, \tag{5.5}$$

where V_n is the nth service completion time in the superposition of the service processes $S(t) = \sum_{j=1}^{m} S_j(\mu_j t)$ and $A(\lambda t) \equiv A(\lambda_1 t) + \ldots + A(\lambda_n t)$. Let ψ_V be the asymptotic decay-rate function of the process $V \equiv \{V_n : n \geq 1\}$.

Note that V is the inverse process of S. Hence, we can combine theorems 5 and 7 and proposition 6 to obtain the asymptotic decay-rate functions

$$\psi_V(\theta) = -\psi_S^{-1}(-\theta), \tag{5.6}$$

$$\psi_N(\theta) = \psi_{V\circ A\circ\lambda}(\theta) = \lambda\psi_A(-\psi_S^{-1}(\theta)), \tag{5.7}$$

which is of the same form as (1.15). Hence, if this model is valid, theorem 10 implies (1.15).

Now, doing a change of variables $\theta = -\psi_S(-\omega)$, (5.7) above becomes

$$\lambda\psi_A(\omega) + \psi_S(-\omega) = 0, \tag{5.8}$$

which is just (1.12). Hence, we have the following important result.

THEOREM 11

A root ζ of (1.12) exists if and only if a root η of (1.15) exists. If these roots exist, then (1.14) and (1.16) hold.

Now we consider the waiting-time sequence in a single-server model. The net input process is now the sequence $\{V_n - U_n : n \geq 1\}$, where V_n is the nth service completion time and U_n is the nth arrival time. Let $\psi_V(\theta)$ and $\psi_U(\theta/\lambda)$ be the asymptotic decay-rate functions of $\{V_n\}$ and $\{U_n/\lambda\}$. Then the asymptotic decay-rate function of the net input process (again assuming the server is working continuously) is

$$\psi_V(\theta) + \psi_U(-\theta/\lambda) = -\psi_S^{-1}(-\theta) - \psi_A^{-1}(-\theta)/\lambda). \tag{5.9}$$

The waiting-time decay-rate equation thus can be written as

$$-\psi_S(-\theta) = \psi_A^{-1}(-\theta/\lambda). \tag{5.10}$$

Applying the function $\lambda\psi_A$ to both sides, we see that (5.10) is equivalent to (1.15). In other words, *the asymptotic decay rates of the workload and waiting time in a single-server queue coincide*. For further discussion, see Abate et al [2], and Glynn and Whitt [26].

6. Arrival processes with Markov structure

The definition of the asymptotic decay-rate function ψ_A in (1.10) shows how an arrival process A contributes to the asymptotic decay rates σ and η in (1.3)–(1.8) via the asymptotic behavior of its logarithmic moment-generating function. (We will focus on arrival processes, but what we say also applies to service processes.) For applications, it is also important to know how to *compute* the asymptotic decay-rate function ψ_A in terms of basic model parameters. For renewal processes and superpositions of renewal processes, we have indicated how to do the calculations in propositions 8 and 9. We now consider other arrival processes.

A large class of arrival processes can be represented as *batch Markovian arrival processes* (BMAPs). A BMAP is a convenient representation of the *versatile Markovian point process* in Neuts [43,47] or N process in Ramaswami [49]. The BMAP and the BMAP/GI/1 queue were introduced and investigated by Lucantoni [37]. A tutorial on the BMAP/G/1 queue is in Lucantoni [38]. The BMAP is a generalization of the *Markovian arrival process* (MAP) introduced by Lucantoni et al. [39] allowing batch arrivals. The MAP includes *Markov modulated Poisson processes* (MMPPs), *phase-type* (PH) renewal processes and superpositions of these proceses as special cases. The MMPP models changes of phase without arrivals and arrivals without changes of phase; the MAP models these as well as arrivals and changes of phase occurring simultaneously. Since independent superpositions of MAPs and BMAPs are again processes of the same kind, this is a good framework to study the multi-channel queues introduced in section 1.

In addition to the arrival counting process A, in a BMAP there is an *auxiliary phase process* J, assuming values in $\{1,\dots,m\}$. The process $(A,J)\equiv\{(A(t),J(t)):t\ge 0\}$ is modeled as a continuous-time Markov chain with a specially structured infinitesimal generator matrix $\hat{Q}$. In particular $\hat{Q}$ can be represented in block form as

$$\frac{\hat{Q}}{\lambda}=\begin{pmatrix} D_0 & D_1 & D_2 & D_3 & \cdots \\ & D_0 & D_1 & D_2 & \cdots \\ & & D_0 & D_1 & \cdots \\ & & & D_0 & \cdots \end{pmatrix}, \qquad (6.1)$$

where λ is the overall arrival rate, D_k, $k\ge 0$, are $m\times m$ matrices, D_0 has negative diagonal elements and nonnegative off-diagonal elements, and $D\equiv\sum_{k=0}^{\infty}D_k$ is an irreducible infinitesimal generator. (Our formulas are somewhat different from those in Lucantoni [37] because we factor out in λ in (6.1). A transition from (i,j) to $(i+k,l)$ corresponds to a batch arrival of size k and a transition from auxiliary state i to auxiliary state l. It is possible to have transitions between auxiliary states without arrivals (nondiagonal elements of D_0) and it is possible to have arrivals without changing the auxiliary state (diagonal elements of D_k for $k\ge 1$). An MAP is the special case in which the matrices D_k are matrices of 0's for $k\ge 2$.

Let π be the steady-state probability vector associated with D, i.e. determined by $\pi D=0$ and $\pi e=1$, where e is a vector of 1's and 0 is a vector of 0's (which should be clear from the context). A fundamental role is played by the *BMAP matrix generating function*

$$D(z)\equiv\sum_{k=0}^{\infty}D_k z^k. \qquad (6.2)$$

We assume that $D(z)$ has a radius of convergence $z^*>1$. When D_k is a matrix of 0's for all $k\ge k_0$, as is the case for the ordinary MAP (then $k_0=2$), $z^*=\infty$. Having

$z^* > 1$ implies that $D(z)$ can be regarded as an analytic function of a complex variable z for $|z| < z^*$. The kth derivative $D^{(k)}(z)$ is then finite and analytic for all k and $|z| < z^*$ as well.

Specifying the overall arrival rate λ separately in (6.1) means that

$$\pi\left(\sum_{k=1}^{\infty} kD_k\right)e = \pi D^{(1)}(1)e = 1. \tag{6.3}$$

As in Lucantoni [37], let the marginal conditional distribution of $(A(t), J(t))$ be given by

$$P_{ij}(n,t) = P(A(t) = n,\ J(t) = j \,|\, A(0) = 0,\ J(0) = i) \tag{6.4}$$

and let

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t)z^n \tag{6.5}$$

be the associated *counting process matrix generating function*, which is given explicitly by

$$P^*(z,t) = \mathrm{e}^{D(z)t}, \quad t \geq 0, \tag{6.6}$$

see eq. (8) of ref. [37].

Given any initial vector $\tilde{\pi}$ on the phase space, the counting process $A(t)$ has probability distribution

$$P_{\tilde{\pi}}(N(t) = n) = \sum_{i=1}^{m} \sum_{j=1}^{m} \tilde{\pi}_i P_{ij}(n,t) = \tilde{\pi}P(n,t)e \tag{6.7}$$

and moment generating function

$$E_{\tilde{\pi}} z^{A(t)} = \tilde{\pi}P^*(z_1 t)e = \tilde{\pi}\exp(D(z)t)\,e. \tag{6.8}$$

Hence, the asymptotic decay-rate function ψ_A of the BMAP A with initial phase distribution $\tilde{\pi}$ can be expressed as

$$\psi_A(\theta) = \lim_{t\to\infty} t^{-1} \log \tilde{\pi}\mathrm{e}^{D(\mathrm{e}^{\theta})t}e. \tag{6.9}$$

Theorem 1 of Choudhury and Whitt [16] shows that the asymptotic decay-rate function ψ_A can be characterized in terms of the *Perron–Frobenius eigenvalue* of the matrix $D(z)$, say $pf(D(z))$. This Perron–Frobenius eigenvalue is well defined for $0 < z < z^*$; see chapter 1 and section 3 of Seneta [51], and section 3 of Abate et al. [3].

THEOREM 12 (CHOUDHURY AND WHITT [16])

For any θ with $\theta < \log z^*$ and any initial probability vector $\tilde{\pi}$,

$$\psi_A(\theta) = pf(D(e^{\theta})). \tag{6.10}$$

As indicated at the beginning of this section, many processes are special cases of BMAPs; see Lucantoni [37,38], and Neuts [47] for background. For the special case of an MAP, $D(z) = D_0 + D_1 z$, so that (6.10) becomes

$$\psi_A(\theta) = pf(D_0 + D_1 e^{\theta}). \tag{6.11}$$

An MMPP is a special case of an MAP with $D_0 = M - \Lambda$ and $D_1 = \Lambda$, where M is the infinitesimal generator matrix of the Markovian environment provess and Λ is the associated diagonal matrix of Poisson arrival rates. Hence, for an MMPP characterized by the pair (M, Λ), (6.11) becomes

$$\psi_A(\theta) = pf(M - \Lambda + \Lambda e^{\theta}). \tag{6.12}$$

A PH renewal process with representation (α, T) as in chapter 2 of Neuts [44] is a special case of an MAP with $D_0 = T$ and $D_1 = (-Te)\alpha$. Hence, for a PH renewal process, (6.11) becomes

$$\psi_A(\theta) = pf(T + (-Te)\alpha e^{\theta}). \tag{6.13}$$

So far in this section, we have characterized the asymptotic decay-rate function of a BMAP. The BMAPs are also important because *they provide a specific structure in which we can rigorously justify the assertions in section 1*, in particular, the decay rate equations (1.12) and (1.15), and the auxiliary equations (1.14) and (1.16). Under regularity conditions, the limits (1.3) and (1.4) are shown to hold in BMAP/GI/1 queues and MAP/MSP/1 queues in Abate et al. [3].

The structure of a BMAP that is a superposition of independent BMAPs is easily treated using the Kronecker sum $\oplus$; e.g. see p. 243 of Neuts [46]. The following is theorem 3 of Choudhury and Whitt [16].

THEOREM 13 (CHOUDHURY AND WHITT [16])

Consider n independent BMAPs characterized by pairs $(A_i(t), J_i(t))$ with arrival rates λ_i, $m_i \times m_i$ matrices D_{ik}, $k \geq 0$, $1 \leq i \leq n$, matrix generating functions $D_i(z)$, and asymptotic decay-rate functions ψ_{A_i}. Then the pair $(A(t), J(t))$, where $A(t) = A_1(t) + \ldots + A_n(t)$ and $J(t) = (J_1(t), \ldots, J_n(t))$ determines another BMAP with arrival rate $\lambda = \lambda_1 + \ldots + \lambda_n$, associated $m \times m$ matrices D_k, where $m = \prod_{i=1}^{n} m_i$ and

$$\lambda D_k = \lambda_1 D_{1k} \oplus \ldots \oplus \lambda_n D_{nk}, \quad k \geq 0, \tag{6.14}$$

and matrix generating function

$$D(z) = \left(\frac{\lambda_1}{\lambda}\right) D_1(z) \oplus \ldots \oplus \left(\frac{\lambda_n}{\lambda}\right) D_n(z), \tag{6.15}$$

which has Perron–Frobenius eigenvalue

$$pf(D(z)) = \frac{\lambda_1}{\lambda} pf(D_1(z)) + \ldots + \frac{\lambda_n}{\lambda} pf(D_n(z)) . \qquad (6.16)$$

The single-channel asymptotic decay-rate function ψ_A is thus

$$\lambda \psi_A(\theta) = \lambda_1 \psi_{A_1}(\theta) + \ldots + \lambda_n \psi_{A_n}(\theta) . \qquad (6.17)$$

Proof

First, (6.14) and (6.15) involve the standard construction. Then (6.16) holds because $pf(M_1 \oplus M_2) = pf(M_1) + pf(M_2)$ for matrices for which pf is well defined. Finally, (6.17) then follows from theorem 12. □

The limits (1.3) and (1.4) for BMAP/GI/1 queues established in [3] requires regularity conditions. The weaker limits (1.7) and (1.8) hold for this model in greater generality by virtue of Glynn and Whitt [26] and theorem 1 of Choudhury and Whitt [16]. The limits in Glynn and Whitt [26] show that even the BMAP structure is not needed.

So far, we have not yet said anything about rigorous support for (1.12)–(1.16) for multi-server queues, and indeed much here remains only a conjecture. However, strong positive support for (1.12)–(1.16) exists in the results for GI/PH/m queues with heterogeneous servers established by Takahashi [57], and Neuts and Takahashi [48]. The single-server MAP/MSP/1 results in [3] extend to the multi-server MAP/MSP/m model by the same arguments used for treating PH/PH/m or GI/PH/m models; see p. 206 of Neuts [44].

An alternative general framework for arrival processes instead of BMAPs is *Markov renewal processes* (MRPs) or *semi-Markov processes* (SMPs). These processes are characterized by an *SMP transition matrix* $F(x)$, with $F_{ij}(x)$ being the probability starting with an arrival in state i, the time until the next arrival will be less than or equal to x and the next state will be j. The arrival-time asymptotic decay-rate function $\psi_U(\theta) \equiv -\psi_A^{-1}(-\theta)$ is given by $\log pf(\hat{F}(-\theta))$, where pf is the Perron–Frobenius eigenvalue and $\hat{F}(s)$ is the matrix Laplace–Stieltjes transform of F, i.e.,

$$\hat{F}_{ij}(s) = \int_0^\infty \mathrm{e}^{-sx} \mathrm{d}F_{ij}(x); \qquad (6.18)$$

see p. 237 of Neuts [46] and the appendix of Neuts [47]. Note that an ordinary renewal process is a special case of a one-dimensional MRP; then $pf(\hat{F}(-\theta)) = \hat{F}(-\theta)$.

In fact, as pointed out by Lucantoni and Neuts [40], a (B)MAP can be represented as a special case of a (batch) Markov renewal process. A batch Markov renewal process is specified by a sequence $\{F_k(x) : k \geq 1\}$ of semi-Markov matrices with the $(i,j)$the element of $F_k(x)$ representing the probability that the time until the next batch arrival is less than or equal to x, that the size of the batch is k, and that the phase just after the arrival is j, given that the phase just after the previous batch arrival is i. The BMAP is represented as a batch Markov renewal process via

$$F_k(x) = \int_0^x e^{D_0 u} D_k \, du = (1 - e^{D_0 x})(-D_0^{-1}) D_k. \qquad (6.19)$$

It is instructive to look at MMPPs because they are special cases of *both* MRPs and BMAPs. In the framework of MRPs, an MMPP has SMP transition matrix

$$F(x) = \int_0^x \exp[(M - \Lambda)u]\Lambda \, du, \quad x \geq 0, \qquad (6.20)$$

where M is the infinitesimal generator matrix of the Markovian environment process and Λ is the associated diagonal matrix of Poisson arrival rates. In this case,

$$\hat{F}(s) = (sI - M + \Lambda)^{-1}\Lambda, \qquad (6.21)$$

so that

$$\psi_U(\theta) = \log pf((-\theta I - M + \Lambda)^{-1}\Lambda). \qquad (6.22)$$

We can now independently derive the asymptotic decay-rate function ψ_A for an MMPP given in (6.12) by applying (6.22) and the inverse relation in theorem 7.

PROPOSITION 14

From the MRP formula (6.22) and theorem 7, it follows that for MMPPs (6.12) holds, i.e.,

$$\psi_A(\theta) = pf((e^\theta - 1)\Lambda + M).$$

Proof

Given (6.22), let x be a right (positive, real, unique up to constant multiple) eigenvector of $\hat{F}(-\theta)$ corresponding to $e^{-y} = e^{\psi_U(\theta)}$. Then $\theta = \psi_U^{-1}(-y) = -\psi_A(y)$ and

$$e^{-y}x = (\psi_A(y)I - M + \Lambda)^{-1}\Lambda x,$$

so that

$$\psi_A(y)x = ((e^y - 1)\Lambda + M)x. \qquad (6.23)$$

Finally, we apply section 2.3 of Seneta [51] to confirm that (6.23) implies that $\psi_A(y)$ is the Perron–Frobenius eigenvalue associated with the matrix $(e^y - 1)\Lambda + M$. (Its off-diagonal elements are nonnegative.) □

7. Test queues and relations between decay rates

It is significant that the effect of one arrival or service process on the asymptotic decay rates σ and η in a multi-channel queue depends on this process only via its asymptotic decay-rate function. Thus, the individual arrival and service processes can be analyzed separately. Indeed, an alternative way to define and/or characterize the asymptotic decay-rate functions is to use an *indirect operational approach*, paralleling the definition of the normalized mean workload in Fendick and Whitt [24].

In particular, we can *define* the asymptotic decay-rate function ψ_A of an arrival process A in terms of the asymptotic decay rate that actually prevails in a convenient test queue with arrival process A. A natural test queue for this purpose is a single-server queue with deterministic service times having mean 1. Then we let $\eta(\lambda)$ be the workload asymptotic decay rate in (1.4) that actually prevails in a G/D/1 queue with arrival process $\{A(\lambda t) : t \geq 0\}$. To obtain a full function, we consider this decay rate $\eta(\lambda)$ as a function of the arrival rate λ, where the arrival rate varies over the full range $0 < \lambda < 1$. Since the service times are deterministic with mean 1, the service-time asymptotic decay-rate function is $\psi_S(\theta) = \theta$. Hence, by (1.15), $\eta_i(\lambda)$ is determined by the equation

$$\lambda \psi_A(\eta(\lambda)) = \eta(\lambda), \quad 0 < \lambda < 1. \tag{7.1}$$

Given the function $\eta(\lambda)$, $0 < \lambda < 1$, we can define $\psi_A(\theta)$ as the unique increasing function satisfying (7.1).

Equivalently, if $\sigma(\lambda)$ is the queue-length decay-rate in (1.3), we can apply (1.12) to obtain

$$\lambda \psi_A(-\log \sigma(\lambda)) = -\log \sigma(\lambda), \quad 0 < \lambda < 1. \tag{7.2}$$

Equations (7.1) and (7.2) are in fact equivalent, because for deterministic service time

$$\eta(\lambda) = -\psi_S(-\zeta(\lambda)) = \zeta(\lambda) = -\log \sigma(\lambda) \tag{7.3}$$

by (1.14).

The function $\{\sigma(\lambda) : 0 < \lambda < 1\}$ in (7.2) is the *caudal characteristic curve* in Neuts [46], and the function $\{\eta(\lambda) : 0 < \lambda < 1\}$ in (7.1) and (7.3) is the delay analog. The asymptotic decay-rate function ψ_A is obtained as a function of these curves in a test queue.

Similarly, we can define the asymptotic decay-rate function ψ_S in terms of the asymptotic decay rate $\eta(\lambda)$ that actually prevails in a D/G/1 queue when the service process is S, as a function of the arrival rate λ. Setting $\lambda\psi_A(\theta) = \lambda\theta$ in (1.15), we obtain the equation

$$-\lambda \psi_S^{-1}(-\eta(\lambda)) = \eta(\lambda), \quad 0 < \lambda < 1, \tag{7.4}$$

which is equivalent to

$$\psi_S(-\eta(\lambda)/\lambda) = -\eta(\lambda), \quad 0 < \lambda < 1. \tag{7.5}$$

Starting with (1.12), we also obtain (7.5). The asymptotic decay-rate ψ_S is the unique increasing function satisfying (7.5) for $0 < \lambda < 1$.

Note that eq. (7.1) is similar to, but not identical to, eq. (7.5). Recall that the congestion is in general not the same in a D/G/1 with service process S and a G/D/1 queue with arrival process S and common arrival and service rates.

We have used a D process to create test queues. We could also use other processes for this purpose. For example, if the arrival or service process is Poisson, then the asymptotic decay-rate function is as in (2.1).

From (1.14), we see that the asymptotic decay-rate functions $\lambda\psi_A$ and ψ_S each can be thought of as functions mapping potential queue-length decay rates ζ into workload decay rates η. Similarly, by (1.16) the inverse functions $-\psi_S^{-1}$ and $\psi_A^{-1}(\cdot/\lambda)$ map the other way. Each asymptotic decay-rate function determines a relationship *between* σ and η. Combining arrival and service asymptotic decay-rate functions $\lambda\psi_A$ and ψ_S then determines $\sigma(\lambda)$ and $\eta(\lambda)$ themselves. Thus, we see that *two relations between the queue length and the workload asymptotic decay rates determine the decay rates themselves*. This seems to be closely related to $H = \lambda G$ (the extension of $L = \lambda W$); see [62]. Indeed, it may be a consequence of the distributional version of $L = \lambda W$; see section 8.4 of [62]. From that perspective, we would want to replace the workload by the waiting time, which we should be able to do because they should have the same asymptotic decay rates, as indicated at the end of section 5; also see [2].

8. Heavy-traffic asymptotic expansions

Due to the convexity of the asymptotic decay-rate functions, the decay-rate equations (1.12) and (1.15) are often not difficult to solve. Nevertheless, it is useful to consider approximate solutions obtained by approximating the asymptotic decay-rate functions $\psi_{A_i}(\theta)$ in (1.10) and $\psi_{S_j}(\theta)$ in (1.11) by Taylor series expansions about $\theta = 0$. As shown in Abate et al. [1], Abate and Whitt [4], and Choudhury and Whitt [16], this analysis leads to heavy-traffic approximations for the asymptotic decay rates σ and η in (1.3)–(1.8); i.e. we obtain

$$\eta(\rho) = c_1(1-\rho) + c_2\frac{(1-\rho)^2}{2} + \ldots + c_k\frac{(1-\rho)^k}{k!} + \ldots \tag{8.1}$$

and

$$\sigma(\rho)^{-1} = d_1(1-\rho) + d_2\frac{(1-\rho)^2}{2} + \ldots + d_k\frac{(1-\rho)^k}{k!}, \tag{8.2}$$

where $\rho \equiv \lambda$ is the traffic intensity. For example, for the BMAP/GI/1 queue, the first seven values of the coefficients c_k and d_k in (8.1) and (8.2) are given in Choudhury and Whitt [16]. (These expansions are in fact obtained from the logarithmic generating functions $\log Ez^{A(t)}$, instead of the logarithmic moment generating function $\log Ee^{\theta A(t)}$ used here.)

This analysis is important because it reveals which properties of the processes the asymptotic decay rates σ and η primarily depend upon. In particular, for general stationary point processes, the coefficients c_k and d_k in (8.1) and (8.2) depend on the first $k+1$ asymptotic cumulants of the processes A_i, $1 \le i \le n$ and S_j, $1 \le j \le m$. The first terms in (8.1) and (8.2) coincide with the familiar heavy-traffic approximations, obtained by letting $\rho \to 1$. It is significant that the order of the two limits $x \to \infty$ and $\rho \to 1$ does not matter. In other words, when ρ is not too small, the small-tail asymptotic approximations considered here tend to be close to heavy-traffic approximations. This simple heavy-traffic approximation is

$$\eta = \sigma^{-1} - 1 \approx c_1(1-\rho), \tag{8.3}$$

where

$$c_1 = \frac{2}{\lambda c_A^2 + c_S^2}, \tag{8.4}$$

with

$$\lambda c_A^2 = \sum_{i=1}^{n} \lambda_i c_{A_i}^2, \quad c_S^2 = \sum_{j=1}^{n} \mu_j c_{S_j}^2, \tag{8.5}$$

$$c_{A_i}^2 = \lim_{t\to\infty} \frac{\operatorname{Var} A_i(t)}{\lambda_i t} \quad \text{and} \quad c_{S_j}^2 = \lim_{t\to\infty} \frac{\operatorname{Var} S_j(t)}{\mu_j t}. \tag{8.6}$$

In (8.5), $c_{A_i}^2$ is the *limiting value of the index of dispersion for counts* (IDC) of A_i as $t \to \infty$ and $\lambda_i c_{A_i}^2$ is the *asymptotic variance of* A_i; e.g. see Iglehart and Whitt [31,32], Sriram and Whitt [55], Fendick et al. [22,23], and Fendick and Whitt [24]. These simple one-term approximations have been independently proposed by Sohraby [53,54].

This heavy-traffic analysis also suggests how the congestion might be characterized without asymptotics. Paralleling the use of the IDC in [24], we might characterize an arrival (or service) process A by first k cumulants of $A(t)$ as functions of time.

9. How many servers are needed?

We have focused on the question: How many arrival processes can a queue support, with given service process and given congestion percentile criterion? By essentially the same reasoning, we can instead consider the arrival process as fixed and ask the question: How many servers of different kinds are needed, with given congestion percentile criterion?

For this purpose, suppose that a multi-channel arrival process with asymptotic decay-rate function $\lambda\psi_A$ is given. Also, suppose that an upper bound σ^* has been placed on the queue-length asymptotic decay rate σ. Paralleling (1.17), we conclude that the m service processes $S_1, \ldots, S_m$ with asymptotic decay-rate functions $\mu_1 \psi_{S_1}, \ldots, \mu_m \phi_{S_m}$ are feasible if and only if

$$-\sum_{i=1}^{m} \mu_i \psi_{S_i}(\log \sigma^*) \geq \lambda \psi_A(-\log \sigma^*). \tag{9.1}$$

We could call $\lambda \psi_A(-\log \sigma^*)$ the *overall service requirement* and $\mu_i \psi_{S_i}(\log \sigma^*)$ the *effective service units* provided by server i. The feasibility criterion (9.1) says that the sum of the effective service units provided must be at least the overall service requirement.

10. Numerical methods

The discussion so far has been focused on how the asymptotic decay rates σ and η in (1.3) and (1.4) depend on the model structure (the arrival and service processes) and on the basic model data (e.g. the interrenewal-time distributions in the renewal process case). For the purpose of actually computing the asymptotic parameters in (1.3) and (1.4), however, there are simple procedures that do not rely on these results.

First, given any method for calculating the tail probabilities (including estimation from data), we can estimate the asymptotic parameters by taking logarithms. For example, the parameters η and α in (1.4) are the slope and intercept for the limiting linear form of $\log P(W > x)$. These parameters can be estimated by linear regression.

Second, given an explicit expression for the transform of the steady-state distribution of interest, we can perform elementary asymptotic analysis to determine the asymptotic parameters. For example, suppose that we are given the Laplace–Stieltjes transform $E\mathrm{e}^{-sW}$ of the steady-state workload W. Then the associated Laplace transform of $P(W > x)$ is

$$\hat{W}^c(s) \equiv \int_0^\infty \mathrm{e}^{-sx} P(W > x)\, \mathrm{d}x = \frac{1 - E\mathrm{e}^{-sW}}{s}. \tag{10.1}$$

We can find the asymptotic decay rate η and the asymptotic constant α in (1.4) by finding the right-most singularity of $\hat{W}^c(s)$ and often by exploiting the final-value theorem for Laplace transforms:

$$\alpha \equiv \lim_{x \to \infty} \mathrm{e}^{\eta x} P(W > x) = \lim_{s \to -\eta} (s + \eta) \hat{W}^c(s + \eta), \tag{10.2}$$

for which we *assume* that the limit (1.4) is valid, which means that the singularity is a simple pole; see Abate et al. [3]. This assumption is supported by extensive experience showing that (1.4) indeed is typically valid. (There are exceptions, however; e.g. see example 5 of [1].)

If we can represent $\hat{W}^c(s)$ as $N(s)/D(s)$, where the singularity is a root of $D(s) = 0$, then we can also express α analytically as

$$\alpha = N(-\eta)/D'(-\eta), \tag{10.3}$$

where D' is the derivative of D.

Finally, we mention that Choudhury and Lucantoni [13] have developed methods for numerically calculating the asymptotic parameters in (1.3) and (1.4) based on calculating higher-order moments from moment generating functions.

11. Effective bandwidths for a mean-workload criterion

In this section, we show how the general effective bandwidth procedure based on (1.2) can be applied with the mean steady-state workload. As before, we let the service rate be fixed at 1. Of course, as an approximation, we can regard $EW \approx 1/\eta$ for η in (1.4), so that the mean workload and the reciprocal of the asymptotic decay rate can serve as rough approximations for each other. Moreover, in [1] the product ηEW is proposed as a rough approximation for the asymptotic constant α in (1.4), so that we can use approximations for EW to obtain a full approximation based on (1.4). However, here we consider EW directly.

Heavy-traffic approximations for the mean workload in multi-channel queues follow from Iglehart and Whitt [31,32], and Fendick et al. [22,23], but experience has shown that the quality of these heavy-traffic approximations deteriorates significantly with a large number of arrival or service channels. (Interestingly, a similar phenomenon can occur with the limits (1.3) and (1.4); see Choudhury et al. [14].) Indeed, different asymptotics then come into play, in which the number of channels becomes large as the traffic intensity approaches 1; see Whitt [61]. Hence, in Sriram and Whitt [55], Heffes and Lucantoni [29], and Fendick and Whitt [24], refined approximations for the mean workload in $\Sigma G_i/G/1$ queues were developed, based on indices of dispersion (variance-time curves), and these could be considered here. In the context of common deterministic service times of primary concern here, the index of dispersion for work (IDW) in [24] reduces to the *index of dispersion for counts* (IDC), defined by

$$I_{ci}(t) = \frac{\operatorname{Var} A_i(t)}{E\, A_i(t)}, \quad t \geq 0, \tag{11.1}$$

where Var is the variance. Indeed, with i.i.d. service times that are independent of the arrival process, and that have squared coefficient of variation c_s^2, the IDW is $I_w(t) = I_c(t) + c_s^2$; see eq. (59) of ref. [24].

The mean workload in a $\Sigma G_i/D/1$ queue can be written as

$$EW \approx \frac{c_z^2(\lambda)}{2(1-\lambda)}, \tag{11.2}$$

where $c_z^2(\lambda)$ is called the *normalized mean workload.* As discussed in [24], it reflects the variability contribution of the input to EW.

A main theme of [23] and [24] is that the normalized mean workload $c_z^2(\lambda)$, and thus EW itself, can be approximated reasonably well based on the IDW or, in our case, the IDC, which in turn can be estimated from data or calculated analytically from models.

In contrast to the multi-class M/G/1 queue considered in Kelly [34] and Fendick et al. [22], in general the normalized mean workload $c_z^2(\lambda)$ can vary significantly as a function of λ, which limits the possibilities for effective bandwidths. Moreover, *$c_z^2(\lambda)$ need not be an increasing function of λ.* (A simple example is the $E_2/M/1$ queue in fig. 1 of [24].). Hence, even though we can apply (11.2) to express the constraint $EW \leq C$ as

$$\lambda\left(1 + \frac{c_z^2(\lambda)}{2C}\right) \leq 1 \tag{11.3}$$

as in eq. (3.7) of Kelly [34], we cannot do as much with it, because $c_z^2(\lambda)$ is not necessarily increasing.

However, a simple way to obtain a sufficient condition (rather than a necessary and sufficient condition) for feasibility is to work with the *maximal normalized mean workload,* which we define as

$$\hat{c}_z^2(\lambda) = \sup_{\lambda_0 \leq u \leq \lambda} c_z^2(u), \tag{11.4}$$

where λ_0 is some minimal traffic intensity that we are certain the system will not be operating below. We include λ_0 because there often is a decrease in $c_z^2(\lambda)$ near 0. However, *the important point is the upper limit* λ in (11.4). By considering the supremum up to λ in (11.4), instead of over [0, 1], we succeed in avoiding known common steep increases in $c_z^2(\lambda)$ in the neighborhood of $\lambda = 1$, as shown for model 2 in fig. 2 of ref. [24].

Paralleling eq. (3.7) of Kelly [34], we can obtain a sufficient condition (rather than a necessary and sufficient condition) for feasibility using (1.2) with effective bandwidths

$$\alpha_i(C) = \lambda_i\left(2 + \frac{\hat{c}_z^2(\lambda)}{C}\right) \tag{11.5}$$

and the constraint $\sum \alpha_i \leq 1$.

It is appropriate to point out a difference between our setting and that of Kelly [34]. Here, we are considering the $\sum G_i/D/1$ model with common fixed service times having mean 1. In contrast, in [34] the arrivals bring in their own work requirements. This case is treated by simply changing λ_i above to $\rho_i = \lambda_i \tau_i$ above, where τ_i is the mean service requirement for each class-i customer. This is equivalent to working with IDW instead of the IDC. Note that with this interpretation, (11.5) reduces to eq. (3.7) of ref. [34] in this case of the multiclass M/G/1 model considered there. (That example is also discussed in [22].)

To apply (11.4) and (11.5), we need a formula for $c_z^2(\lambda)$. For this, we can apply the IDC, which itself can be estimated from data or calculated analytically for models. As in eq. (56) of ref. [24], suppose that each IDC is based on scaling giving it rate 1. If the component processes actually have rate λ_i with $\sum\lambda_i = \lambda$, then for superpositions of independent processes,

$$I_c(\lambda t) = \sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda}\right) I_{ci}(\lambda_i t). \tag{11.6}$$

Notice that we have linearity in (11.6), as in (1.2) and (1.12), subject to appropriate *time scaling*.

The fundamental problem in developing the approximation is relating the time t in $I_c(\lambda t)$ to the traffic intensity λ in $c_z^2(\lambda)$. One simple approximation considered in [24], which was motivated by the concept of relaxation time, is

$$c_z^2(\lambda) = I_c(t(\lambda)), \tag{11.7}$$

where

$$t(\lambda) = \frac{\lambda I_w(\infty)}{2(1-\lambda)^2}; \tag{11.8}$$

see eq. (14) of ref. [24]. Formulas (11.6)–(11.8) could be used to compute the mean workload associated with multiple sources. (Other approximations are also discussed in [23] and [24].)

Even though $c_z^2(\lambda)$ need not be increasing in λ, it is easy to show by a sample-path argument that the steady-state mean and, in fact, the entire steady-state distribution increases as additional sources are added. Indeed, to see that the workload at any time (starting empty) increases with the addition of a new source, apply eqs. (21)–(24) of ref. [24] and note that all increments of the net input process necessarily increase.

Thus, we can apply (11.6)–(11.8) with a criterion of mean workload to determine whether or not to admit each successive source. Because of the monotonicity, when sources leave, feasibility will be maintained.

12. History

The emerging high-speed networks exploiting ATM technology have stimulated an enormous literature. So many people have written about topics related to the present paper that it is difficult to properly trace the development of key ideas. Nevertheless, it seems that some history might be helpful, even if it is only one person's view. As indicated at the outset, Roberts [50] gives a broad overview.

My interest in queues with superposition arrival processes goes back to the heavy-traffic studies in Iglehart and Whitt [31,32]. On the surface, the analysis here may seem quite different, but there are many similarities in the two kinds of asymptotic analysis, some of which we pointed out in section 8. In particular, the linearity in (1.13) also occurs in the heavy-traffic formulas in (8.5). Motivated by applications to statistical multiplexing, I worked on this problem further in Sriram and Whitt [55], Fendick and Whitt [24], Fendick et al. [22,23], and Berger and Whitt [8].

My interest in effective bandwidths stems from discussions with Richard Gibbens in 1990. He suggested that it was an interesting idea, but I was skeptical. At that time, I wrote some (largely negative) notes about effective bandwidths, which were evidently the basis for the acknowledgement in Kelly [34]. At that time, I did not appreciate the power of the reasoning using (1.2). My interest in effective bandwidths was rekindled by the papers by Kelly [34], and Gibbens and Hunt [25]. A seminal paper in the whole area was Hui [30]. Another key paper that appeared concurrently with [25] and [34] is Guerin et al. [28].

From these papers, the power of (1.2) and the relevance of asymptotic decay rates and logarithmic moment generating functions are evident, but it is no doubt fair to say that a general theory of asymptotic decay rates for multiple sources was missing. At that time, I realized that a substantial basis for such a theory of asymptotic decay rates already existed in Marcel Neuts' work on asymptotic decay rates [44–48], especially his fundamental paper on the caudal characteristic curve [46]. A contribution here is to point out the importance of Neuts' work for the effective bandwidth problem. Variants of the general decay-rate equation (1.12) were obtained, evidently for the first time, by Neuts [46]. In particular, theorems 4 and 5 there focus on multiple channels, and yield versions of (1.12). The effective bandwidth analysis here in (1.17) follows easily from (1.2) and (1.12).

Independently of effective bandwidths, in 1992 I began studying approximations for steady-state tail probabilities in queues with Joseph Abate and Gagan L. Choudhury. At first, we focused on the GI/GI/1 queue, but we soon considered more general models. With the help of David Lucantoni, we began looking at tail probabilities in the BMAP/GI/1 and GI/PH/1 queues [1–4, 14–16]. A significant component of this work was developing algorithms for computing the steady-state distributions, but we were also interested in developing and evaluating approximations based on asymptotics, as in (1.3) and (1.4). The power of these asymptotic approximations is pointed out by Tijms [58]. There is a substantial history in risk theory, as is pointed out on p. 269 of Asmussen [6]. Recent related work is contained in Van Ommeren [59], Asmussen and Perry [7], and references in these sources. A significant feature of [1], [2], [4] and [16] is the development of effective approximations for the asymptotic parameters in (1.3) and (1.4).

While working on these asymptotics, it occurred to me that what we were learning ought to be relevant for the effective bandwidth problem. In particular, as indicated in section 6, the BMAP/GI/1 queue is a natural model for studying the superposition arrival processes arising in the effective bandwidth problem, because superpositions of independent BMAPs are again BMAPs. Hence, I wrote the first version of this paper. However, beyond pointing out the relevance of known theory about asymptotic decay rates, such as in [46], the first version of this paper was largely speculative, certainly much more so than the present version. This final version incorporates many things I have learned since then. In particular, my completed papers [3], [16], [26] and [27] were significantly motivated by the desire to substantiate conjectures in the first version of this paper. The main ideas in this paper were already in the first version, but theoretical justification has been added.

I had the opportunity to hear Chang talk about his work [11] at Performance'92 in Newport. However, at that time his focus was primarily on bounds (the analog of theorem 10 here, theorem 3.9(ii) of [11], appears only in the 1993 revision). Hence, I did not at first realize the importance of his work for asymptotics. It is significant that he develops the general approach using logarithmic moment generating functions. Chang's minimum envelope rate (MER) is a traffic characterization essentially equivalent to the asymptotic decay-rate functions here. Thus, for the effective bandwidth problem, Chang [11] was evidently the first to develop a general framework like that proposed here. Like the first version of this paper, the first version of [11] did not make contact with general large deviation theory, such as the Gärtner–Ellis theorem. However, the final version of [11] and Chang et al. [12] do. An additional paper is by Kesidis et al. [35].

I learned about the relevance of the Gärtner–Ellis theorem and general large deviations theory from Peter Glynn, who helped me to obtain the substantial supporting theory in [26,27]. As indicated above, [26] overlaps with Chang [11], and Chang et al. [12].

At the same time I was writing the first version of this paper, Anwar Elwalid and Debasis Mitra were writing their effective bandwidth paper [19]. Anwar Elwalid and Debasis Mitra evidently had obtained their effective bandwidth results in [19,20] quite a bit earlier; see the citation to [20] in [21]. However, I did not see [19] until after I completed the first version of this paper, and I did not see [20] until it was issues in 1993. Even though Anwar Elwalid and Debasis Mitra are my colleagues, their work came largely as a surprise to me. I knew that they were thinking about the effective bandwidth problem, but from their previous work I had thought they were primarily focusing on full spectral expansions. Elwalid and Mitra [19] independently arrive at essentially the same solution as in (1.17) for a special class of Markov-modulated fluid models, and provide strong theoretical support. Their essential reasoning in [19] and [20] follows from the seminal paper by Anick et al. [5]. Additional contributions appear in Mitra [41], Elwalid et al. [21], and Stern and Elwalid [56]. Formula (7.24) of ref. [20] is essentially the same as decay-rate equation (1.12) here for the MMPP/MMPP/m model, which is a special case of the MAP/MSP/m model discussed here at the end of section 6. In addition to addressing the effective bandwidth problem, this work is important because it treats the full spectral expansion.

One of the main ideas in my work with Joseph Abate and Gagan Choudhury [1,4,16] is developing heavy-traffic asymptotic expansions for the asymptotic decay rates, as discussed here in section 8. We later learned that Sohraby [53,54] had similar ideas, although he considers only the first term, as in Neuts [46].

More recently, with our algorithm for the BMAP/GI/1 queue, Gagan Choudhury, David Lucantoni and I have been investigating the quality of the effective bandwidth approximation. As indicated in section 1, we found that the quality of the approximation can deteriorate dramatically as the number of sources increases, even when the target tail probability is very small, such as 10^{-9} [14]. In a simulation study, our colleague Kiran Rege reached a similar conclusion (but for larger blocking probabilities, e.g. 10^{-4}). The effective bandwidth approach to admission control may nevertheless be effective, but these numerical results indicate that some caution is needed, because there can be serious difficulties in some parameter regions. The difficulties we discovered with the effective bandwidth approximation are consistent with previous difficulties encountered with queues having superposition arrival processes, i.e. in models of statistical multiplexing; see Sriram and Whitt [55], Fendick and Whitt [24], Fendick et al. [22,23], and Berger and Whitt [8].

As shown in Chang et al. [12], the asymptotic decay rates can be very effective for speeding up simulations to estimate tail probabilities. This shows that the effective bandwidth theory may have important applications beyond those originally intended.

Acknowledgements

I thank my co-authors of recent related papers – Joseph Abate, Arthur Berger, Gagan Choudhury, Peter Glynn and David Lucantoni – for helping to motivate and improve this paper. I thank the authors of the cited unpublished papers for showing their work to me.

References

[1] J. Abate, G.L. Choudhury and W. Whitt, Exponential approximation for tail probabilities in queues, I: waiting times, Oper. Res., to appear.

[2] J. Abate, G.L. Choudhury and W. Whitt, Exponential approximations for tail probabilities in queues, II: sojourn time and workload (1992), submitted.

[3] J. Abate, G.L. Choudhury and W. Whitt, Asymptotics for steady-state tail probabilities in structured Markov queueing models, Stochastic Models 10 (1994), to appear.

[4] J. Abate and W. Whitt, A heavy-traffic expansion for asymptotic decay rates in multi-channel queues, Oper. Res. Lett., to appear.

[5] D. Anick, D. Mitra and M.M. Sondhi, Stochastic theory of a data-handling system with multiple sources, Bell Syst. Tech. J. 61(1982)1871–1894.

[6] S. Asmussen, *Applied Probability and Queues* (Wiley, New York, 1987).

[7] S. Asmussen and D. Perry, On cycle maxima, first passage problems and extreme value theory for queues, Stochastic Models 8(1992)421–458.

[8] A.W. Berger and W. Whitt, The pros and cons of a job buffer in a token-bank rate-control throttle, IEEE Trans. Commun., to appear.

[9] A.A. Borovkov, *Stochastic Processes in Queueing Theory* (Springer, New York, 1976).

[10] J.A. Bucklew, *Large Deviation Techniques in Decision, Simulation and Estimation* (Wiley, New York, 1990).

[11] C.S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, IEEE Trans. Aut. Cont., to appear. (Preliminary version in CDC'92.)

[12] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, Effective bandwidth and fast simulation of ATM in tree networks, IBM T.J. Watson Research Center, Yorktown Heights, NY (1992).

[13] G.L. Choudhury and D.M. Lucantoni, Numerical computation of the moments of a probability distribution from its transform, Oper. Res., to appear.

[14] G.L. Choudhury, D.M. Lucantoni and W. Whitt, Squeezing the most out of ATM (1993), submitted.

[15] G.L. Choudhury, D.M. Lucantoni and W. Whitt, An algorithm for a large class of G/G/1 queues, in preparation.

[16] G.L. Choudhury and W. Whitt, Heavy-traffic asymptotic expansions for the asymptotic decay rate in BMAP/G/1 queues, Stochastic Models, to appear.

[17] K.L. Chung, *A Course in Probability*, 2nd ed. (Academic Press, New York, 1974).

[18] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications* (Jones and Bartlett, Boston, 1993).

[19] A.I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, IEEE Trans. Networks, to appear. (Abbreviated version in *INFOCOM'93*.)

[20] A.I. Elwalid and D. Mitra, Markovian arrival and service communication systems: spectral expansions, separability and Kronecker-product forms, AT&T Bell Laboratories, Murray Hill, NJ (1993).

[21] A.I. Elwalid, D. Mitra and T.E. Stern, Statistical multiplexing of Markov modulated sources: theory and computational algorithms, in: *Teletraffic and Data Traffic in a Period of Change, ITC-13*, eds. A. Jensen and B. Iversen (Elsevier, Amsterdam, 1991) pp. 495–500.

[22] K.W. Fendick, V.R. Saksena and W. Whitt, Dependence in packet queues, IEEE Trans. Commun. 37(1989)1173–1183.

[23] K.W. Fendick, V.R. Saksena and W. Whitt, Investigating dependence in packet queues with the index of dispersion for work, IEEE Trans. Commun. 39(1991)1231–1244.

[24] K.W. Fendick and W. Whitt, Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue, Proc. IEEE 77(1989)171–194.

[25] R.J. Gibbens and P.J. Hunt, Effective bandwidths for the multi-type UAS channel, Queueing Systems 9(1991)17–28.

[26] P.W. Glynn and W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, J. Appl. Prob. 31(1994), to appear.

[27] P.W. Glynn and W. Whitt, Large deviations behavior of counting processes and their inverses (1993), submitted.

[28] R. Guerin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE J. Select. Areas Commun. SAC-9(1991)968–981.

[29] H. Heffes and D.M. Lucantoni, Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, IEEE J. Select. Areas Commun. SAC-4(1986) 856–868.

[30] J.Y. Hui, Resource allocation for broadband networks, IEEE J. Select. Areas Commun. SAC-6(1988)1598–1608.

[31] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, I, Adv. Appl. Prob. 2(1970) 150–177.

[32] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, II: sequences, networks and batches, Adv. Appl. Prob. 2(1970)355–369.

[33] N.L. Johnson and S. Kotz, *Discrete Distributions* (Wiley, New York, 1969).

[34] F.P. Kelly, Effective bandwidths at multi-class queues, Queueing Systems 9(1991)5–16.

[35] G. Kesidis, J. Walrand and C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, Research Report 18588, IBM T.J. Watson Research Center, Yorktown Heights, NY (1992).

[36] J.F.C. Kingman, A convexity property of positive matrices, Quart. J. Math. 12(1961)283–284.

[37] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stochastic Models 7(1991)1–46.

[38] D.M. Lucantoni, The BMAP/G/1 queue: a tutorial, in: *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, eds. L. Danatiello and R. Nelson (Springer, New York, 1993).

[39] D.M. Lucantoni, K.S. Meier-Hellstern and M.F. Neuts, A single-server queue with server vacations and a class on non-renewal arrival processes, Adv. Appl. Prob. 22(1990)676–705.

[40] D.M. Lucantoni and M.F. Neuts, The customer delay in a single server queue with a batch Markovian arrival process, AT&T Bell Laboratories, Holmdel, NJ (1993).

[41] D. Mitra, Stochastic theory of a fluid model of producers and consumers coupled by a buffer, Adv. Appl. Prob. 20(1988)646–676.

[42] D. Mitra, R.J. Gibbens and B.D. Huang, State dependent routing on symmetric loss networks with trunk reservations, I, IEEE Trans. Commun. 40(1992)477–482.

[43] M.F. Neuts, A versatile Markovian point process, J. Appl. Prob. 16(1979)764–779.

[44] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models* (The Johns Hopkins University Press, Baltimore, 1981).

[45] M.F. Neuts, Stationary waiting-time distributions in the GI/PH/1 queue, J. Appl. Prob. 18(1981) 901–912.

[46] M.F. Neuts, The caudal characteristic curve of queues, Adv. Appl. Prob. 18(1986)221–254.

[47] M.F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications* (Marcel Dekker, New York, 1989).

[48] M.F. Neuts and Y. Takahashi, Asymptotic behavior of the stationary distributions in the GI/PH/C queue with heterogeneous servers, Z. Wahrscheinlichkeitsth. 57(1981)441–452.

[49] V. Ramaswami, The N/G/1 queue and its detailed analysis, Adv. Appl. Prob. 12(1980)222–261.

[50] J.W. Roberts, *Performance Evaluation and Design of Multiservice Networks, COST 224 Final Report* (Commission of the European Communities, Luxembourg, 1992).

[51] E. Seneta, *Nonnegative Matrices and Markov Chains*, 2nd ed. (Springer, New York, 1981).

[52] W.L. Smith, On the distribution of queueing times, Proc. Cambridge Phil. Soc. 49(1953)449–461.

[53] K. Sohraby, On the asymptotic behavior of heterogeneous statistical multiplexer with applications, *INFOCOM'92*, Florence, Italy.

[54] K. Sohraby, On the theory of general on–off sources with applications in high-speed networks, *INFOCOM'93*, San Francisco, CA.

[55] K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, IEEE J. Select. Areas Commun. SAC-4(1986)833–846.

[56] T.E. Stern and A.I. Elwalid, Analysis of separable Markov-modulated rate models for information-handling systems, Adv. Appl. Prob. 23(1991)105–139.

[57] Y. Takahashi, Asymptotic exponentiality of the tail of the waiting-time distribution in a PH/PH/c queue, Adv. Appl. Prob. 13(1981)619–630.

[58] H.C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach* (Wiley, New York, 1986).

[59] J.C.W. Van Ommeren, Exponential expansion for the tail of the waiting-time probability in the single-server queue with batch arrivals, Adv. Appl. Prob. 20(1988)880–895.

[60] W. Whitt, Some useful functions for functional limit theorems, Math. Oper. Res. 5(1980)67–85.

[61] W. Whitt, Queues with superposition arrival processes in heavy traffic, Stoch. Proc. Appl. 21(1985) 81–91.

[62] W. Whitt, A review of $L = \lambda W$, Queueing Systems 9(1991)235–268.