ARRANGING QUEUES IN SERIES

S. Suresh
AT&T Engineering Research Center
Murray Hill, NJ 07974

W. Whitt
AT&T Bell Laboratories
Murray Hill, NJ 07974

June 21, 1988

Revision: April 21, 1989

ABSTRACT

For given external arrival process and given service-time distributions, the object is to determine the order of infinite-capacity single-server queues in series that minimizes the long-run average sojourn time per customer. We gain additional insight into this queueing design problem, and congestion in open queueing networks more generally, primarily by performing simulation experiments. We develop a new parametric-decomposition approximation for departure processes, which can be used in general queueing network algorithms as well as in this design problem. For this design problem, we conclude that the key issue is variability: The order tends to matter when the service-time distributions have significantly different variability, and not otherwise. The order also matters less in light traffic, even with a relative difference criterion. Arranging the queues in order of increasing service-time variability, using the squared coefficient of variation as a partial characterization of variability, seems to be an effective simple heuristic. In all cases of two queues in series, the simulation results indicate that the same order is optimal for all combinations of the traffic intensities, suggesting that the light-traffic asymptotics in Greenberg and Wolff (1988) should usually be effective for identifying the best order. Comparisons with simulations for two queues also indicate that the parametric-decomposition approximations provide quite accurate quantitative estimates of the expected sojourn time with each order. However, for more than two queues, the approximations need not properly describe the congestion at a bottleneck queue.

Key Words: queueing networks; tandem queues; departure processes; queueing system design; simulation; variance reduction; common random numbers; approximations; parametric-decomposition approximations.

1. INTRODUCTION

1.1 The Queueing Design Problem

This paper revisits a queueing design problem considered by Tembe and Wolff (1974), Pinedo (1982), Whitt (1985), Greenberg and Wolff (1988) and Wein (1988). The model is an open network of single-server queues in series. Each customer (or job) arrives according to an external arrival process and is served once at each queue, with the order of the queues being the same for all customers. Each queue has unlimited waiting space, the FIFO (first-in first-out) discipline, and i.i.d. (independent and identically distributed) service times that are independent of the other random quantities in the model. The design problem is to determine, for given fixed external arrival process, the order of the queues that minimizes the expected steady state (or long-run average) sojourn time (time in system) per customer. More generally, the object is to determine if the order actually matters, and if it does, which orders are good and which are bad. Even more generally, the object is to determine how queues in series perform.

Given that the external arrival process is a renewal process (i.e., the interarrival times are i.i.d.), which we also assume here, the model is specified by the distribution of the service times at each queue and the distribution of the external interarrival times. This problem is difficult for general distributions (e.g., when they are not all exponential), because exact expressions for the expected steady-state sojourn time typically are unavailable, primarily because the arrival processes to all queues after the first typically are not renewal processes; see Berman and Westcott (1983).

1.2 Parametric-Decomposition Approximations

Whitt (1985) suggested analyzing this design problem *approximately* using parametric-decomposition approximations for queueing networks, as in Whitt (1982, 1983, 1984) and references cited there. With this approximation procedure, each distribution is partially characterized by its first two moments or, equivalently, by its mean and squared coefficient of variation (variance divided by the square of the mean). Then closed-form formulas give an approximate squared coefficient of variation for the arrival process to each queue and an approximate expected steady-state waiting time at each queue; see Section 4. Of course, the expected steady-state sojourn time for queues in series actually depends on the distributions beyond

their first two moments, but experience indicates that fairly good approximations can be obtained given this partial information. Even when the best order can be determined exactly, as in the special cases considered by Tembe and Wolff (1974), the approximations are useful because they provide *quantitative estimates* of the expected sojourn times with each order.

1.3 The Purpose of this Paper

Our goal was to obtain additional insight into this queueing design problem and the parametric-decomposition approximations. We gained additional insight primarily by performing simulation experiments. (None of the previous papers discussing this design problem reported any simulation results.) For the most part, we restrict attention to the special case of only two queues, but we consider a few examples with more queues.

With respect to the queueing design problem, we aim to answer the following questions:

(1) *Can the order of the queues significantly affect performance?*
(Section 3.1)

(2) *If so, when does the order matter?*
(Sections 3.3, 3.4 and 4.5)

(3) *When the order does matter, what orders are good or bad?*
(Section 3.6)

(4) *Are there simple design heuristics that perform reasonably well?*
(Sections 3.2 and 3.6)

(5) *Are there fundamental principles explaining how queues in series perform?*
(Sections 3.5, 5.1 and 5.2)

With respect to the parametric-decomposition approximations, we aim to answer the following questions:

(6) *How accurate are the approximations in Whitt (1983,1985)?*
(Sections 4.1 and 4.3)

(7) *Can these approximations be improved?*

(Section 4.3)

(8) *Can situations be identified where these approximations break down?*

(Section 5.3)

(9) *How should these approximations be regarded, given that recent light-traffic limits in Section 4 of Greenberg and Wolff (1988) contradict design heuristics on p. 481 of Whitt (1985)?*

(Sections 3.2, 5.1 and 6)

(10) *What is the performance of Wein's (1988) new parametric-decomposition approximation, which is derived from the heavy-traffic multi-dimensional diffusion approximation under balanced loading?*

(Section 5.4)

Whitt (1985) previously addressed questions 1-5 and 8, while Whitt (1984), Albin and Kai (1986), Fendick, Saksena and Whitt (1988), and Bitran and Tirupati (1988) addressed questions 7-8. The object now is to say more.

1.4 Organization of this Paper

We begin in Section 2 by describing our simulation experiment to study the design problem for two queues in series. A significant feature is the use of common random numbers in order to obtain greater efficiency (smaller confidence intervals for given computer budget). In Section 3 we present the main conclusions about the design problem, i.e., our answers to questions 1-5 above.

In Section 4 we briefly review the previous parametric-decomposition approximations for queues in series and then propose a new hybrid approximation for departure processes (see (4.4)). In Section 4 we also use the simulations to evaluate the approximations. In Section 5 we review light-traffic and heavy-traffic limits and investigate to what extent this limiting behavior is reflected in systems with typical traffic intensities, e.g., with $0.1 \le \rho_i \le 0.9$ for all i. In Section 5.3 we show how the current implementation of the parametric-decomposition approximation can break down for several queues in series due to the

presence of a bottleneck queue. Finally, in Section 6 we discuss the case in which the service-time distributions have common variability and issues raised by Greenberg and Wolff. Additional tables describing the approximations are in appendices (available from the authors).

2. SIMULATION EXPERIMENTS

To estimate the expected steady-state waiting time at each queue in these models, we used the SIMAN simulation program; see Pegden (1984). In each case, we performed ten independent replications using 30,000 arrivals in each replication and estimated 90% confidence intervals using the *t*-statistic. (The ordinate of the *t*-distribution used was 1.833.) An initial portion of each run (2000 customers) was discarded to allow the system to approach steady state.

2.1 The Experimental Design

A basic problem was to decide what cases to consider. Even when we restrict attention to the special case of two queues in series, which we mostly did, and partially characterize distributions by their first two moments, there are five parameters. Without loss of generality (by choosing the measuring units), we make the external arrival rate 1. Then the mean service time at queue i coincides with the traffic intensity there, which we denote by ρ_i. The parameter five-tuple characterizing the two-queue model is then $(c_a^2, \rho_1, c_{s1}^2, \rho_2, c_{s2}^2)$, where c_a^2 (c_{si}^2) is the squared coefficient of variation of an interarrival time (service time at queue i). Of course, we have yet to specify the distributions to go with the first two moments.

We determined what cases to include in the simulation experiment by first using the approximations to explore the parameter space. We applied calculus with the formulas and a spreadsheet with the numerical calculations to determine when the order should matter and when it should not (see Section 4.5). This preliminary analysis eventually enabled us to see that, at least for the case of two queues, there are only a few basic configurations. We simulated cases in which the approximations predict the order should matter and cases in which the approximations predict the order should not matter.

In particular, we considered various variability parameter triples $(c_a^2, c_{s1}^2, c_{s2}^2)$ for all combinations of the traffic intensities ρ_1 and ρ_2 in a representative range. Assuming that $c_{s1}^2 \le c_{s2}^2$, the three principal cases seem to be:

$$
\begin{aligned}
&(i) \quad ac_a^2 \le c_{s1}^2 \le c_{s2}^2 \\
&(ii)\ c_{s1}^2 \le c_a^2 \le c_{s2}^2 \\
&(iii)\ c_{s1}^2 \le c_{s2}^2 \le c_a^2 \ .
\end{aligned}
\tag{2.1}
$$

The actual cases considered for two queues in series are summarized in Table 1. These cases can be divided into two groups: *different service variability* ($c_{s1}^2 \neq c_{s2}^2$) and *identical service variability* ($c_{s1}^2 = c_{s2}^2$). For $c_{s1}^2 \neq c_{s2}^2$, we considered 5 variability triples (c_a^2, c_{s1}^2, c_{s2}^2), namely, (0.5, 0.5, 2.0), (1.0, 0.5, 8.0), (1.0, 2.0, 4.0), (4.0, 0.5, 1.0) and (4.0, 1.0, 4.0), ordered lexicographically. Note that the three cases in (2.1) are represented by (1.0, 2.0, 4.0), (1.0, 0.5, 8.0) and (4.0, 0.5, 1.0), respectively, while (0.5, 0.5, 2.0) is on the boundary of (i) and (ii) and (4.0, 1.0, 4.0) is on the boundary of (ii) and (iii).

For each queue we considered 4 values of ρ_i: 0.3, 0.6, 0.8 and 0.9. Thus, associated with each variability triple are 32 cases of two queues in series:

$$(4 \text{ values of } \rho_1) \times (4 \text{ values of } \rho_2) \times (2 \text{ orders}) = 32 \text{ cases} .$$

When $c^2 = 0.5$, we used the E_2 distribution (Erlang of order 2, the convolution of two exponentials); when $c^2 = 1.0$, we used the exponential (M) distribution; when $c^2 > 1$, we used the H_2 distribution (hyperexponential, a mixture of two exponentials) with balanced means; i.e., if p_i is the probability of the exponential variable with mean m_i, then we require that $p_1\, m_1 = p_2\, m_2$.

For $c_{s1}^2 = c_{s2}^2$, we used cases analyzed by Greenberg and Wolff (1988) in which the service-time random variables at the two queues, say X_1 and X_2, are related by $X_2 \overset{d}{=} (\rho_2/\rho_1)X_1$, where $\overset{d}{=}$ means equality in distribution; i.e.; the distributions only differ by a scale factor. In particular, we considered the variability triples (1.0, 0.5, 0.5) and (1.0, 4.0, 4.0). For each queue we considered 5 values of ρ_i: 0.1, 0.2, 0.3, 0.6 and 0.9. Thus, for each triple, there are 5 + 4 + 3 + 2 + 1 = 15 cases of two queues in series. (Since the order cannot matter when $\rho_1 = \rho_2$, there are only 10 cases in which the order might matter.) Just as before, we use E_2, M and H_2 distributions when $c^2 = 0.5$, $c^2 = 1.0$ and $c^2 = 4.0$, respectively. However, for the H_2 distributions we considered *both* balanced means and unbalanced means. The case of unbalanced means has one exponential replaced by an atom at zero.

We also conducted experiments with more than two queues in series, which we discuss in Sections 3.6 and 5.3.

2.2 Variance Reduction

We used common random numbers to obtain improved simulation efficiency (smaller confidence intervals for given computer budget), as discussed in Section 2.1 of Bratley, Fox and Schrage (1987). We used the same random numbers to treat cases that differ only by the service times at a queue having different means. For example, E_2 random variables with mean 1 were randomly generated to treat service times with $c_{si}^2 = 0.5$. Then these random variables were multiplied by ρ_i to obtain the random variables at the queue with mean ρ_i in different cases. Similarly, when different orders of queues were considered, the same random variables were used with both orders. With four values of each traffic intensity, we simulated one model with 64 queues as shown in Figure 1 (16 cases $\times$ 2 orders $\times$ 2 queues), instead of 32 independent models of two queues in series. We generated i.i.d. random variables with mean 1 and c_a^2 for the interarrival times, and i.i.d. random variables with mean 1 and c_{si}^2 for the service times at queue *i*. Upon each arrival, we created 32 separate customers (clones) to go through the 32 separate two-queue configurations. The service times at queue *i* with traffic intensity ρ_i were obtained by simply multiplying the given mean-one random variables by ρ_i.

Using common random numbers with both orders was critical for obtaining reliable estimates of the difference between the expected total sojourn times with different orders, given our simulation run lengths. For example, a typical case of two queues in series yielded expected total sojourn times for the two orders of 8.60 and 8.71 with 90% confidence intervals of (± 0.47) and (± 0.46), respectively, while the estimated difference was 0.104 with a 90% confidence interval of (± 0.02). Thus, we can conclude that there is a *statistically significant difference*, although we do *not* regard the 1.2% estimated relative difference as *practically significant* from the perspective of the design problem. (This obviously depends on the context. We give the numbers as well as our interpretation, so readers can judge for themselves.) This method was especially important for detecting statistically significant differences in light traffic; see Tables 7-9.

2.3 Simulation Results for Two Queues

Tables 2-6 contain the simulation results for the five different-service-variability cases in Table 1, while Tables 7-9 contain simulation results for the three identical-service-variability cases. These tables display the estimated expected waiting time at each queue and the total expected waiting time in the two-queue network for each order. Also given is the estimated difference and relative difference of the total expected sojourn times with the two orders. Of course, the difference of the total expected sojourn times coincides with the difference of the total expected waiting times, because the total expected service time is the same for both orders. However, there is a distinction with the relative difference. *The relative difference is for the total expected sojourn times*; i.e., it is the estimated difference divided by the minimum of the estimated total expected sojourn times, including the sum of the expected service times in the denominator, so as not to obtain a meaninglessly inflated number in light traffic.

In Tables 2-9, CSA $= c_a^2$ and CSi $= c_{si}^2$. Order 1 refers to the given triple, while order 2 is the reverse order. The estimated half-width of the 90% confidence interval appears below each simulation estimate; e.g., in Table 2 the entry for queue 1 of order 1 at $\rho_1 = \rho_2 = 0.9$ means that the 90% confidence interval is 3.939 ± 0.189 or $[3.750, 4.128]$.

The statistical precision of the expected waiting time estimates is not great, especially at higher traffic intensities. (Overall, the statistical precision of these estimates is consistent with the approximation formulas in Whitt (1989).) However, the statistical precision of the estimated difference of the expected sojourn times is adequate to meaningfully compare the orders.

3. CONCLUSIONS ABOUT THE DESIGN PROBLEM

In this section we draw some conclusions about the queueing design problem from the simulation results. We start discussing the approximations in Section 4.

3.1 The Order Can Matter

First, there is no doubt that the order can significantly affect the total expected sojourn time. Tables 2, 3, 4 and 6 clearly demonstrate that the order can matter. The order matters most for the variability triple $(1.0, 0.5, 8.0)$. The largest differences occur for $(\rho_1, \rho_2) = (0.9, 0.8)$ and $(0.9, 0.6)$, the first being

Example 1 of Whitt (1985). For $(0.9, 0.8)$, our simulation estimates for the expected total sojourn times for orders $(1, 2)$ and $(2, 1)$ are 19.2 ± 1.2 and 31.8 ± 1.8; the estimated difference is 12.6 ± 1.0 and the estimated relative difference is 60.2%.

3.2 There Are No Switches

For all eight variability triples in Tables 2-9, the simulation results indicate that the same order is optimal for all combinations of the traffic intensities. This consistency suggests that the light-traffic asymptotics in Greenberg and Wolff (1988) should usually be effective for determining the best order for *all* traffic intensities. Similarly, approximations that tend to be more accurate at higher traffic intensities may also be relatively effective in identifying the best order over the full range of traffic intensities.

3.3 The Key Issue is Variability

Consistent with remarks on pp. 479-481 of Whitt (1985), but deserving more emphasis, we conclude that *the key issue is variability*. Tables 7-9 indicate that if the service-time variability is the same or nearly the same at all the queues, then the order should not matter much. For the variability triple $(1.0, 0.5, 0.5)$ in Table 7, we can conclude statistically that the difference is not zero in several cases, but the difference is consistently very small. Only in the case $\rho_1 = \rho_2 = 0.9$, for which we know there is actually no difference at all, is the estimated relative difference greater than 1%. For the variability triples $(1.0, 4.0, 4.0)$ in Tables 8 and 9, the estimated relative difference is always less than 5%, except for a $\rho_1 = \rho_2 = 0.9$ case in which the order does not matter at all. The order clearly matters more in the unbalanced means case than in the balanced means case (average relative difference among the 10 cases where there can be a difference of 2.15% versus 0.36%).

In contrast, Tables 2-6 indicate that if the service-time variability at the queues differs significantly, then the order can matter much more. For $c_{s1}^2 < c_{s2}^2$, the order tends to matter more as c_{s1}^2 decreases and c_{s2}^2 increases. The order also tends to matter more as c_a^2 decreases. For example, the order matters least in Table 5 when c_a^2 is high (4.0) and c_{s2}^2 is 1.0. In contrast, the order matters most in Table 3, where c_a^2 is lower (1.0) and c_{s2}^2 is higher (8.0).

3.4 The Traffic Intensities

As indicated above, when the service-time variability is the same at both queues (Tables 7-9), the order does not matter much for any traffic intensity pair, but in all cases the order tends to matter more, as measured by the relative difference of expected sojourn times, at higher traffic intensities than at lower traffic intensities.

Let $R(\rho_1, \rho_2)$ be the estimated relative difference as a function of ρ_1 and ρ_2. For $c_{s1}^2 < c_{s2}^2$ as in Tables 2-6, $R(\rho_1, \rho_1)$ and $\max\limits_{\rho_2} R(\rho_1, \rho_2)$ are consistently increasing in ρ_1 within the range of traffic intensities considered (with one exception in Table 5). For any given ρ_1, $R(\rho_1, \rho_2)$ appears to be a concave function of ρ_2 with a maximum near but slightly less than ρ_1. The order tends to matter most for $(\rho_1, \rho_2) = (0.9, 0.8)$, with $(0.9, 0.9)$, $(0.8, 0.8)$ and $(0.8, 0.6)$ next.

3.5 The Variability Propagation Principle

There is one fundamental principle that seems remarkably robust in explaining the performance of queues in series:

The Variability Propagation Principle (VPP): Increased variability in the arrival process or the service times of a queue tends to propagate to the departure process from that queue and thus to the arrival process to all subsequent queues.

The VPP is illustrated by the parametric-decomposition approximations for c_{d1}^2, the approximating squared coefficient of variation of the departure process from the first queue in (4.3) and (4.4) below; i.e., c_{d1}^2 is made a convex combination of c_a^2 and c_{s1}^2, so that c_{d1}^2 increases if either c_a^2 or c_{s1}^2 increases. The VPP has implications for performance, because the delays at a queue tend to increase as the variability of the arrival process or the service times increases, as illustrated by approximation formula (4.1) below.

The VPP obviously has implications for many other problems besides this queueing design problem. For example, for production lines that at least roughly behave like our model, the VPP suggests that we should work harder to reduce variability at the front of the line than at the end of the line.

3.6 A Simple Design Heuristic

For our queueing design problem, the VPP suggests the following simple design heuristic.

Simple Design Heuristic: Order the queues so that $c_{s1}^2 \le c_{s2}^2 \le \cdots \le c_{sn}^2$.

This simple design heuristic is an extension of heuristic design principles $P1$ and $P4$ on p. 481 of Whitt (1985). On p. 480, this simple design heuristic was mentioned as a natural simple rule suggested by Tembe and Wolff (1974), but Whitt (1985) suggested that the simple design heuristic should be inadequate. However, our simulation experience indicates that the simple design heuristic performs remarkably well, even better than suggested by the approximations. Indeed, *the simulations indicate that the simple design heuristic is optimal in every case in Tables 2-6.* Apparent advantages of other orderings seem to be primarily due to approximation error. We discuss this further in Section 4.5.

Of course, we know that the simple design heuristic is *not optimal* in general. For the case of non-overlapping service-time distributions, Tembe and Wolff (1974) *proved* that the queue with the larger service times should go first. Since the distribution of the larger service times can have any nonnegative c_{si}^2, the simple design heuristic above can easily be wrong, which is not to say seriously wrong. For example, if in addition the distribution of the smaller service times is deterministic (i.e., the smaller service times are constant), then Theorem 2 of Tembe and Wolff implies that the order does not matter at all.

In Examples 4 and 5 of Whitt (1985) different orderings were proposed for several queues in series with one or two bottlenecks. However, we performed simulation experiments for these examples that indicate that the simple heuristic performs just as well as, in fact, slightly better. In Example 4 with seven queues in series, the simple design heuristic ordering $(1, 2, 3, 7, 4, 5, 6)$ has an estimated expected total sojourn time of 45.47 ± 3.19, whereas the previously proposed ordering of $(3, 2, 1, 7, 4, 5, 6)$ has an estimated expected total sojourn time of 45.57 ± 3.29. (See Whitt (1985) for more details.) The estimated difference (favoring the simple design heuristic) was 0.09 ± 0.15, which of course is not significant statistically or practically.

In Example 5 of Whitt (1985) with four queues in series, the simple design heuristic ordering $(1, 2, 3, 4)$ has an estimated expected total sojourn time of 138 ± 17, whereas the previously proposed

ordering $(3, 2, 1, 4)$ has an estimated expected total sojourn time of 140 ± 18. The estimated difference was 2.0 ± 0.6, so that the difference (again favoring the simple heuristic) is statistically significant. (As before, we consider the relative difference of 1.5% as practically negligible.)

Finally, we remark that the simple design heuristic is *roughly* consistent with the light-traffic limit for two queues when one distribution is exponential in Section 3 of Greenberg and Wolff; i.e., a queue with less (more) variable service-time distribution according to a Laplace transform ordering should go first (second).

3.7 Comparison With the Assembly Line Balancing Model

Although the insights about variability in Sections 3.3, 3.5 and 3.6 should apply to other models, it is important to note that the conclusions may change dramatically when the model is changed. To illustrate this phenomenon, we compare the simple design heuristic above with good design heuristics for the assembly line balancing problem, as discussed by Hillier and Boling (1966, 1979) and others.

The assembly line balancing problem can also be modeled as several queues in series; the object is to allocate tasks (service-time distributions) to the different queues, so as to maximize the throughput (departure rate). A significant feature of the assembly line model is that the queues have limited buffers (finite waiting rooms). In the assembly line problem, there is an unlimited supply of jobs at the front of the line, instead of an external arrival process with a given arrival rate. In our model, the throughput (departure rate) coincides with the arrival rate, which is fixed, whereas in the assembly line model the throughput is a quantity affected by the design, i.e., a quantity less than the processing rate of the slowest server.

As a consequence of these model differences, the mean service times are much more important for the assembly line model than for our model. For the assembly line, something close to balanced means is desirable for the service-time distributions, with the variability of the service times tending to be much less important. Of course, there is the important bowl phenomenon, but the bowl should usually be shallow.

It is probably fair to say that issues about deterministic rates in stochastic models are usually more critical than issues about variability, but in our model the important rate (the arrival rate = the throughput) is fixed, so that the most important issue becomes variability. When the critical issue is variability,

systematic analysis is typically important, because intuition is typically poor.

Note that our model becomes relevant for the assembly line balancing problem as the amount of waiting space in the assembly line queues increases. Indeed, suppose that all queues after the first have unlimited waiting space. Then after the first queue we have a series of infinite-capacity queues with a renewal external arrival process, where the external interarrival-time distribution coincides with the service-time distribution at the first queue. (The first server is assumed to be the slowest.) Then the overall throughput coincides with the service rate at the first queue, and a relevant issue becomes the average sojourn time.

4. THE APPROXIMATIONS

4.1 Expected Waiting Time in a GI/G/1 Queue

A basic approximation for the expected steady-state waiting time before beginning service in a GI/G/1 queue with arrival rate 1, traffic intensity ρ and variability parameters c_a^2 and c_s^2 is

$$E(W) \approx \frac{\rho^2(c_a^2 + c_s^2)}{2\,(1-\rho)}\ ; \tag{4.1}$$

see (44) of Whitt (1983). Approximation (4.1) is the natural heavy-traffic approximation, because it is asymptotically correct as $\rho \to 1$ (the ratio of the two sides approaches 1) and it is exact for M/G/1 for all ρ. A refined approximation developed by Kraemer and Langenbach-Belz (1976) is (4.1) multiplied by an adjustment factor

$$g(\rho, c_a^2, c_s^2) = \begin{cases} \exp\left[-\dfrac{2(1-\rho)}{3\rho}\ \dfrac{(1-c_a^2)^2}{(c_a^2 + c_s^2)}\right], & c_a^2 < 1 \\ \exp\left[-(1-\rho)\ \dfrac{(c_a^2 - 1)}{(c_a^2 + 4\,c_s^2)}\right], & c_a^2 > 1\ . \end{cases} \tag{4.2}$$

In (4.2) the correction factor g is always less than one, reflecting the fact that (4.1) tends to be too high, especially at lower traffic intensities, although (4.1) is not an upper bound. In (45) of Whitt (1983), (6) of Whitt (1985) and here, *we use* (4.2) *only for* $c_a^2 < 1$, *and set* $g = 1$ *for* $c_a^2 \geq 1$.

To measure the accuracy of approximations for the expected waiting time, we use the minimum of the relative error and the absolute error, e.g., the relative error is ($|$approx. $-$ sim. estimate$|$/sim. estimate). Since the arrival rate is 1, the expected waiting time coincides with the expected queue length, so that it seems reasonable to use the absolute error when the simulation estimate is less than one.

With this criterion, the errors in the approximation for the expected steady-state waiting time in a GI/G/1 queue tend to be of order 0.05. To illustrate, in Table 10 we compare the approximations for $E(W)$ in (4.1) and (4.2) with the simulation estimates and exact numerical values from Seelen, Tijms and van Hoorn (1985) for three variability pairs: $(c_a^2, c_{s1}^2) = (0.5, 0.5)$, $(0.5, 2.0)$ and $(4.0, 1.0)$, each for four values of ρ. In these 12 cases, the average error for (4.1), using the exact numerical values where

available, is 0.045. For the same cases, the average error for our adjusted approximation, which incorporates (4.2) when $c_a^2 < 1$, is 0.026.

Of course, there are other candidate approximations for $E(W)$ worth considering, but the major problem is developing approximations for non-renewal arrival processes. For two queues in series, the larger approximation errors tend to occur at the second queue. Thus we often choose to work with (4.1), *without* the correction factor g, because it is easier to work with.

4.2 Approximating Arrival Processes by Renewal Processes

With the parametric-decomposition approximation procedure for open networks of single-server queues, each queue is analyzed approximately as a GI/G/1 queue, so that the expected waiting times are approximated by (4.1), possibly with a refinement such as (4.2), after an approximating arrival variability parameter is determined for each queue. Each queue is analyzed as a GI/G/1 queue by acting as if the interarrival times are i.i.d. with the squared coefficient of variation of an interarrival time equal to the approximating arrival process variability parameter. The mean interarrival time at each queue, which is the reciprocal of the arrival rate, is obtained (usually exactly) by solving the system of traffic rate equations. Of course, for queues in series the arrival rate at each queue is the same as the external arrival rate.

The major difficulty is obtaining a variability parameter to adequately represent complicated non-renewal arrival processes arising in queueing networks. In general, there is more flexibility in the possible approximations than may be apparent, because the approximating arrival variability parameters may depend on the traffic intensities at all the queues. For example, for two queues in series, the approximating variability parameter of the arrival process at the second queue might depend on the traffic intensity of the second queue as well as the traffic intensity of the first queue, even though the arrival process to the second queue is exogeneous to the second queue. Having c_a^2 depend on ρ has been shown to be useful for treating queues with superposition arrival processes; see Albin (1984), Section 4.3 of Whitt (1983) and Albin and Kai.

For two queues in series with order $(1, 2)$ and a renewal arrival process, the critical approximation is the variability parameter of the arrival process at the queue 2, which coincides with the variability parameter

for the departure process from queue 1, say c_{d1}^2. The approximation in Whitt (1983, 1985) is

$$c_{d1}^2 = \rho_1^2 c_{s1}^2 + (1 - \rho_1^2) c_a^2 , \tag{4.3}$$

which is an approximation for the squared coefficient of variation of a stationary interval between departures; see Section 2.2 of Whitt (1984). Note that c_{d1}^2 in (4.3) does *not* depend on ρ_2, even though it could.

4.3 A New Hybrid Approximation for Departure Processes

Using our simulation results and a spreadsheet, we developed a hybrid approximation, in the spirit of Albin, and Albin and Kai, that is a convex combination of the stationary-interval approximation (4.3) and the asymptotic approximation $c_{d1}^2 \approx c_a^2$ (see Whitt (1982) and Sections 1 and 4 of Whitt (1984)), namely,

$$c_{d1}^2 = \rho_2^{10} c_a^2 + (1 - \rho_2^{10})(\rho_1^2 c_{s1}^2 + (1-\rho_1^2) c_a^2) = \rho_1^2(1-\rho_2^{10}) c_{s1}^2 + [1 - \rho_1^2(1-\rho_2^{10})] c_a^2 . \tag{4.4}$$

Note that (4.4) can be applied to general open queueing networks as well as queues in series, e.g., it can be immediately incorporated in the queueing network algorithms in Whitt (1983), Segal and Whitt (1988) and Bitran and Tirupati (1988). We discuss other approximations for c_{d1}^2 in Sections 5.3 and 5.4; a summary of the approximations discussed appears in Table 11.

As in Whitt (1984), we found that natural candidate weighting factors for the convex combination such as ρ_2 or ρ_2^2 instead of ρ_2^{10} in (4.4) do not provide a consistent improvement over (4.3), but that a much higher exponent such as 10 works remarkably well. Of course, for $\rho_2 \leq 0.7$, $\rho_2^{10} \approx 0$ so that (4.4) essentially reduces to (4.3) when $\rho_2 \leq 0.7$. To see the change from (4.3) to (4.4), note that $(0.95)^{10} = 0.599$, $(0.9)^{10} = 0.349$, $(0.8)^{10} = 0.107$, $(0.7)^{10} = 0.028$ and $(0.6)^{10} = 0.006$.

The improvement provided by (4.4) is seen when the second queue has traffic intensity 0.9. For the 52 cases in Tables 2-9 in which the second queue has $\rho = 0.9$, the average error in the expected waiting time at the second queue, which in these cases is always relative error, was reduced from 12.2% to 5.3% by replacing (4.3) with (4.4).

Overall, for the simulation cases in Tables 2-9, the average estimated error in the expected steady-state waiting time at the second queue (using the criterion introduced in Section 4.2) is 0.066 with the simple

approximation (4.1) plus (4.3) and 0.048 with the adjusted approximation, using (4.2) when $c_a^2 < 1$ and (4.4).

Tables 8 and 9 are interesting because they contain quite different service-time distributions with the same parameters. Of course, the mean waiting times are the same at the first M/G/1 queue in Tables 8 and 9, but differences at the second queue reflect effects of the distributions beyond the first two moments. The differences are about the same order as our approximation errors, which suggests that the approximations are reasonable.

4.4 Expected Sojourn Times for Two Queues

Let $T(1, 2)$ and $T(2, 1)$ be the total sojourn time (waiting time plus service time) per customer in equilibrium for two queues with order $(1, 2)$ and $(2, 1)$, respectively. Combining (4.1) and (4.3), we obtain the relatively tractable *simple approximation*

$$E[T(1, 2)] \approx \rho_1 + \frac{\rho_1^2(c_a^2 + c_{s1}^2)}{2(1-\rho_1)} + \rho_2 + \frac{\rho_2^2(\rho_1^2 c_{s1}^2 + (1-\rho_1^2)c_a^2 + c_{s2}^2)}{2(1-\rho_2)} . \tag{4.5}$$

Even though the adjustments discussed above usually improve the quality of the approximation, we use (4.5) in order to obtain a relatively simple expression, which we can analyze with elementary calculus. We sacrifice some accuracy to clearly see the first-order effects. Based on (4.5), we conclude that order $(1, 2)$ is preferred to order $(2, 1)$ if and only if $\delta_1 < \delta_2$, where

$$\delta_i = (1 - \rho_i)(c_{si}^2 - c_a^2) ; \tag{4.6}$$

see (9) of Whitt (1985). Of course, we need to examine (4.5) to see if the predicted difference is truly significant.

To compare the orderings of two queues, we focus on the approximate relative difference $R \equiv R(c_a^2, \rho_1, c_{s1}^2, \rho_2, c_{s2}^2)$, where

$$R = \frac{|E[T(2, 1)] - E[T(1, 2)]|}{\min\{E[T(1, 2)], E[T(2, 1)]\}} . \tag{4.7}$$

Given (4.5) and $E[T(2, 1)] > E[T(1, 2)]$, our simple approximation for R is

$$R \approx N/D \,, \tag{4.8}$$

where

$$\begin{aligned} N &= \rho_1^2\,\rho_2^2\left[(1-\rho_1)c_{s1}^2 - (1-\rho_2)c_{s2}^2 + (\rho_2-\rho_1)c_a^2\right] \\ D &= c_{s2}^2\left[\rho_2^2(1-\rho_1)\right] + c_{s1}^2\left[\rho_1^2(1-\rho_2) + \rho_1^2\,\rho_2^2(1-\rho_1)\right] \\ &\quad + c_a^2\left[\rho_1^2(1-\rho_2) + \rho_2^2(1-\rho_1)(1-\rho_1^2)\right] + 2(1-\rho_1)(1-\rho_2)(\rho_1+\rho_2)\,. \end{aligned} \tag{4.9}$$

Note that (4.6) does *not* always agree with the simple design heuristic in Section 3.6, which we have indicated is correct in all 80 cases in Tables 2-6. However, in only 9 of these 80 cases does the simple approximation dictate the wrong order: $(\rho_1, \rho_2) = (0.6, 0.9)$, $(0.3, 0.9)$, $(0.3, 0.8)$ in Table 4 and $(0.9, 0.8)$, $(0.9, 0.6)$, $(0.9, 0.3)$, $(0.8, 0.6)$, $(0.8, 0.3)$, $(0.6, 0.3)$ in Table 5. For the three cases in Table 4, the simulation estimate of the relative difference in expected sojourn times are –3.9%, –0.4% and –1.0%, while the predicted relative difference with the simple (adjusted) approximations are 0.4% (–0.1%), 0.2% (0.1%) and 0.02% (0.01%), respectively. Thus, for the three cases in Table 4, the approximations correctly predict that the order does not matter.

However, there are two seriously wrong predictions for the variability triple $(4.0, 0.5, 1.0)$ in Table 5. In the six cases here, the simulation estimates of the relative difference are –3.5%, –1.3%, –0.3%, –2.3%, –0.6% and –2.4%, which indicate that the order really does not matter, whereas the simple (adjusted) approximations predict relative differences of 16.1% (4.4%), 17.6% (8.3%), 4.9% (2.8%), 7.9% (5.8%), 3.5% (2.9%), 1.3% (1.3%), respectively. In the first two cases, the simple approximation predicts that the reverse order $(2, 1)$ is significantly better ($\geq 10\%$), whereas actually the order does not matter. However, this approximation error is largely corrected by the refined hybrid approximation (4.4).

As indicated by Greenberg and Wolff (1988), the approximations consistently select the wrong order in Tables 7 and 9, but for the most part the approximations correctly predict that the order actually does not matter. Significant errors only occurs in the case $(\rho_1, \rho_2) = (0.9, 0.6)$ in Tables 8 and 9; then the difference between the simple (adjusted) approximation predicted relative difference and the simulation estimate is 11.0 % (5.0%) and 15.4% (9.3%).

Overall, the approximations predict the relative difference in expected sojourn times for the two orders remarkably well. The average difference between the simple (adjusted) approximation prediction of the relative difference and the simulation estimate in the 125 cases of Tables 2-9 is 4.8% (2.5%). For the adjusted approximation, the maximum difference is 10.1%; the number of cases in which the difference is greater than 8% (5%) is 8 (17). (In evaluating these numbers, recall that there is considerable noise in the simulation estimates.)

4.5 The Perspective of the Separate Queues

It is useful to consider the design problem for two queues *from the perspective of the separate queues.* When we do, we see that there are two cases: one, where the approximations predict that the good order is better for both queues, and the other, where the approximations predict that there are tradeoffs.

The clear choice occurs when the arrival process variability parameter is between the two service-time variability parameters. When $c_{s1}^2 \le c_a^2 \le c_{s2}^2$, the approximations predict that *both* queues are better off *for all combinations of traffic intensities* with the order $(1, 2)$. In (4.6), δ_1 is negative and δ_2 is positive. In fact, this uniform preference remains true for (4.4) and all departure process approximations consistent with

$$\min \{ c_a^2, c_{s1}^2 \} \le c_{d1}^2 \le \max \{ c_a^2, c_{s1}^2 \} . \tag{4.10}$$

Since the conclusion is important, we state it as a proposition, but we omit the elementary proof.

Proposition 1. *Suppose that the approximation for* $E(W)$ *is nondecreasing in* c_a^2 (as in (4.1)) *and the departure process approximation satisfies* (4.10) (*as in* (4.3) *and* (4.4)). *If* $c_{s1}^2 \le c_a^2 \le c_{s2}^2$, *then the arrival process variability parameters and the approximate expected waiting times at the two queues are both smaller for all traffic intensities with order* $(1, 2)$.

Hence, we would predict with more confidence that the simple design heuristic is correct when $c_{s1}^2 \le c_a^2 \le c_{s2}^2$. As indicated in Section 3, this prediction is correct in all simulation cases (Tables 2, 3 and 6). Moreover, in these cases the simulation estimates of the expected waiting times do seem to be less, or essentially the same, at *both* queues with the optimal order. In addition, we would predict that higher moments of the steady-state sojourn time are also less with order $(1, 2)$.

In contrast, if $c_a^2 \le c_{s1}^2 \le c_{s2}^2$, then the approximations predict that both queues want to be first; if $c_{s1}^2 \le c_{s2}^2 \le c_a^2$, then the approximations predict that both queues want to be second. Hence, we expect there to be tradeoffs, so that the order should matter less, which is consistent with our simulation results. When $\rho_1 = \rho_2$, the obvious order is $(1, 2)$, consistent with (4.6) and the simple design heuristic in Section 3.6. When $\rho_1 \neq \rho_2$, (4.6) suggests beginning to favor the queue with the higher traffic intensity. The idea is that the delay at the higher queue should dominate the total sojourn time, so that it is more important to take action to keep the variability of its arrival process as low as possible. Hence, if $c_a^2 \le c_{s1}^2 \le c_{s2}^2$ and $\rho_1 < \rho_2$ or if $c_{s1}^2 \le c_{s2}^2 \le c_a^2$ and $\rho_1 > \rho_2$, then we may want to use the order $(2, 1)$, because it reduces the variability of the arrival process to the queue with the higher traffic intensity. We call this idea the *alleged variability smoothing principle*.

As we explain in Section 5.2 below, the flaw in this reasoning is that in heavy traffic the alleged variability smoothing does not take place. The delay at a second queue with high traffic intensity tends to be nearly the same as if the first queue were not there.

5. ASYMPTOTIC REFERENCE POINTS

Useful insights can be gained from light-traffic and heavy-traffic limits. In this section we review some of these results, discuss their implications, and investigate to what extent the insights are valid for systems with typical traffic intensities, i.e., $0.1 \le \rho_i \le 0.9$.

5.1 Light Traffic

There is a very simple story for single-server queues in light traffic. Assume that customers arrive one at a time (no batch arrivals) in a single-server queue with arrival rate 1 and let $E[W(\rho)]$ be the mean steady-state waiting time before beginning service as a function of the mean service time (traffic intensity) ρ. Of course, in great generality $E[W(\rho)] \to 0$ as $\rho \to 0$. The first important observation is that in great generality

$$\lim_{\rho \to 0} \rho^{-1} E[W(\rho)] = 0 . \tag{5.1}$$

Not only does the mean waiting time become negligible as $\rho \to 0$, but *the mean waiting time relative to*

the mean service time becomes negligible as $\rho \to 0$.

The limit (5.1) is often expressed a different way. In many light-traffic analyses, the mean service time is set equal to 1, so that the arrival rate becomes ρ. Then, instead of (5.1), we have only $E[W(\rho)] \to 0$ as $\rho \to 0$, but the practical consequence is the same: the mean waiting time relative to the mean service time becomes negligible as $\rho \to 0$.

There are several analyses supporting (5.1); e.g., Section 6.8 of Newell (1982), Daley and Rolski (1984, 1988), and Section 4.1 of Fendick and Whitt (1988). This behavior can also be deduced from Wolff (1982) and Greenberg and Wolff, but Greenberg and Wolff conclude (p. 501) that the order can have substantial relative effect on delay in light traffic. Note that approximation (4.1) (which tends to be too high in light traffic) is consistent with (5.1), suggesting that

$$\rho^{-1} E[W(\rho)] = O(\rho) \quad \text{as} \quad \rho \to 0 . \tag{5.2}$$

The only way to support Greenberg and Wolff's interpretation that the order does seriously matter in light traffic is to use the relative difference of the waiting times *excluding the service times*, but we think that it is usually more meaningful to include the service times.

The second important observation concerns the departure process from the queue in light traffic. If the arrival rate is fixed at 1 and $\rho \to 0$, then in great generality the departure process converges with probability 1 to the arrival process; see Proposition 1 of Whitt (1988). Hence, if the arrival process is a renewal process with variability parameter c_a^2, then the departure process is asymptotically the same renewal process with parameter c_a^2. Hence, as $\rho_1 \to 0$, $c_{d1}^2 \approx c_a^2$ is asymptotically correct. Note that approximations (4.3) and (4.4) are consistent with this light-traffic limit.

The overall impact of a queue being in light traffic is that the expected sojourn time in a network containing the queue is asymptotically the same as if the queue were not there. Thus, if one of two queues in series is in light traffic, then the order of the two queues should not matter, even with the relative difference criterion. It is easy to see that approximation (4.8) is consistent with this limit. If either $\rho_1 \to 0$ or $\rho_2 \to 0$, then $R \to 0$.

Moreover, the same conclusion holds if *both* queues are in light traffic. By (5.1), the expected waiting times become asymptotically negligible compared to the expected service times, which of course are the same for the two orders. Again, approximation (4.8) is consistent with this limit; if both $\rho_1 \to 0$ and $\rho_2 \to 0$, then $R \to 0$.

In summary, our analysis supports the conclusion that *the order of queues in series does not matter if all queues, or all but one, are in light traffic*. Of course, it remains to determine what traffic intensities actually constitute light traffic, for which we turn to the simulations.

In fact, it is hard to say what actually constitutes light traffic. One would certainly expect that $\rho_i = 0.3$ is light traffic, but the order still matters when $\rho_1 = 0.9$ and $\rho_2 = 0.3$ in Table 3; indeed the estimated relative difference in expected total sojourn time with the two orders is 19.0%.

The simulations do support the conclusion suggested by the light traffic limits that the order matters *less* as either ρ_1 or ρ_2, or both, get small. The simulations also indicate that for the queue to be negligible, the traffic intensity must be lower as the service-time variability increases. For example, in Table 3, queue 1 with $c_{s1}^2 = 0.5$ is pretty negligible at $\rho = 0.3$, whereas queue 2 with $c_{s2}^2 = 8.0$ is not. In fact, it is only in Table 3 with $c_{s2}^2 = 8.0$, that the relative difference exceeds 10% when $\rho_i = 0.3$ for at least one queue.

5.2 One Queue in Heavy Traffic

When we let $\rho_i \to 1$ for one i, there are two effects to consider: the effect on queue i and the effect on all subsequent queues.

For queue i, approximation (4.1) becomes asymptotically correct (the ratio of the two sides converges to 1) as $\rho_i \to 1$, where c_a^2 is the squared coefficient of variation associated with the external arrival process. For $i = 1$, this just means that approximation (4.1) is good as $\rho \to 1$. For $i \geq 2$, it means that the expected sojourn time at bottleneck queue i is asymptotically the same as if the previous queues were not there; see Iglehart and Whitt (1970), Reiman (1983), and Sections 1 and 4 of Whitt (1984).

As noted in Section 4 of Whitt (1984), approximation (4.3) is *not* consistent with this heavy-traffic limit. In contrast, for $i = 2$, the hybrid approximation (4.4) *is* consistent with this limit. However, for $i > 2$, neither (4.3) nor (4.4) is consistent with this limit. For $i > 2$ and $\rho_i \to 1$, the appropriate

approximate variability parameter for the departure process from queue $(i - 1)$ (to use as the arrival process variability parameter at queue i) is $c_{d(i-1)}^2 \approx c_a^2$. (The service-time variability parameters $c_{s1}^2, \ldots, c_{s(i-1)}^2$ play no role.) Similarly, for general queueing networks, the asymptotic method approximation for the arrival process at queue i becomes appropriate as $\rho_i \to 1$.

The effect of $\rho_i \to 1$ on subsequent queues is fairly clear. Consistent with (4.3) and (4.4), queue i acts like an external source; the departure process approaches a renewal process with the service times as the interdeparture times.

The overall impact of one queue in heavy traffic is now clear: *If one queue is in heavy traffic, then the order will not matter.* As $\rho_i \to 1$, the waiting time at queue i will be the dominant portion of the total sojourn time and the waiting time at queue i will be the same no matter where queue i appears (asymptotically as $\rho_i \to 1$). There is one important exception: If $c_a^2 = c_{s1}^2 = 0 < c_{s2}^2$, then the situation is more complicated; then the first queue is a D/D/1 queue, so that $E(W_1) = 0$ for all ρ_1, and the order should matter as $\rho_1 \to 1$. However, provided that $c_{si}^2 > 0$, the order ultimately does not matter as $\rho_i \to 1$.

As with the light traffic in Section 5.1, it remains to determine at what traffic intensities this heavy-traffic limiting behavior can actually be seen. From Tables 2-9, we certainly cannot conclude that the order does not matter when one queue has traffic intensity 0.9 and the other has a lower traffic intensity. However, the heavy-traffic limit is the basis for the asymptotic method and the refined hybrid approximation (4.4), which produces significant improvement at the second queue when the traffic intensity there is 0.9. The heavy-traffic limit thus provides important insight.

5.3 Approximation Breakdown Due to a Bottleneck Queue

In this section we present simulation results showing that the heavy-traffic limiting behavior with a single bottleneck queue can indeed be realized at typical traffic intensities. These simulation results also show that the parametric-decomposition approximation in Section 4 can perform poorly. For other examples of approximation breakdown, see Section 4 of Whitt (1985), Fendick, Saksena and Whitt, and Bitran and Tirupati.

The model here consists of nine exponential queues in series, with $\rho_i = 0.6$ for $1 \le i \le 8$ and $\rho_9 = 0.9$, so that the last queue is the bottleneck. We let the interarrival times be H_2 with $c_a^2 = 8.0$.

Approximations (4.3) and (4.4) suggest that the appropriate approximating variability parameter for the arrival process to queue i, say c_{ai}^2, should decrease toward 1.0 as i increases, so that queue 9 should behave much like an M/M/1 queue with $E(W_9) \approx 8.1$, whereas the bottleneck heavy-traffic limit suggests, paralleling (4.1), the *bottleneck approximation* at queue i of

$$E(W_i) \approx \frac{\rho_i^2 \, (c_a^2 + c_{si}^2)}{2 \, (1 - \rho_i)} \, , \tag{5.3}$$

which is 36.5 in this case.

In fact, c_{ai}^2 does not quite reach 1 via the approximations, so that the simple approximation for $E(W_9)$ is 8.9 and the adjusted approximation is 9.1. However, the simulation estimate is 30.1 ± 5.1, again using a 90% confidence interval, which is much closer to the bottleneck approximation 36.5.

A similar, but less dramatic result holds for the case of a deterministic arrival process with $c_a^2 = 0.0$. Then the simple and adjusted approximations for $E(W_9)$ are both 8.0, slightly less than the M/M/1 value of 8.1, whereas the bottleneck approximation in (5.3) is 4.05 and the simulation estimate is 5.03 ± 0.22. Again the simulation estimate is much closer to the bottleneck approximation.

These examples show *limitations* in the parametric-decomposition approximations *as currently developed*. We still believe that improved parametric-decomposition approximations can be developed to cover these examples. However, it appears that an appropriate approximating arrival process variability parameter at queue i, say c_{ai}^2, should be a function of $c_a^2, c_{s1}^2, \ldots, c_{s,i-1}^2$ and $\rho_1, \rho_2, \ldots, \rho_{i-1}, \rho_i$. Paralleling (4.10), we are fairly confident that c_{ai}^2 should satisfy the requirement that

$$\min \{c_a^2, c_{s1}^2, \ldots, c_{s,i-1}^2\} \le c_{ai}^2 \le \max \{c_a^2, c_{s1}^2, \ldots, c_{s,i-1}^2\} \, , \tag{5.4}$$

but (4.3) and (4.4) evidently are not always good. However, we expect (4.3) and (4.4) to work well when the bounds in (5.4) are not too far apart.

In order to capture the bottleneck phenomenon, Reiman (1988) proposed a *sequential bottleneck parametric-decomposition approximation* for general queueing networks. The procedure starts by

identifying the queue with the highest traffic intensity (assuming no ties) and applying the bottleneck approximation to it, as in (5.3). This queue, say queue *i*, is then removed from the network and replaced by an external source with a renewal process having the given arrival rate at queue *i* and an arrival variability parameter equal to the service-time variability parameter c_{si}^2. Then the procedure is repeated by identifying the queue with the second highest traffic intensity, and so forth. For two queues in series, this procedure makes $c_{d1}^2 \approx c_a^2$ if $\rho_1 < \rho_2$ and $c_{d1}^2 \approx c_{s1}^2$ if $\rho_2 < \rho_1$; Reiman suggests $c_{d1}^2 \approx \left[c_a^2 + c_{s1}^2\right]/2$ if $\rho_1 = \rho_2$. This procedure has the virtue of capturing the effect of a bottleneck in a general network, so it and variants have promise. However, for only two queues in series, it seems to be dominated by the hybrid approximation in (4.4). For two queues, (4.4) responds more smoothly to changes in the parameters, as it evidently should.

5.4 Two or More Queues in Heavy Traffic

As in Section 5.2, when several queues are in heavy traffic, we can disregard all other queues, but the behavior of the subnetwork of heavily loaded queues in quite complicated, being described (asymptotically) by a multi-dimensional diffusion process; see Iglehart and Whitt, Harrison (1973, 1978), Reiman (1984) and Harrison and Williams (1987). In general, the stationary distribution cannot be readily evaluated, so that we do not know the approximate mean waiting time at the second queue for two queues in series. However, Greenberg (1986, 1987) did apply results from Harrison to show that for two queues with $\rho_1 = \rho_2$ and $c_a^2 = c_{s2}^2$, if the service-time distribution at the first queue is deterministic, then the expected waiting time at the second queue is asymptotically (as $\rho_1 \to 1$) 25% less than it would be if the first queue were not there.

The heavy-traffic behavior of two queues in series when both $\rho_1 \to 1$ and $\rho_2 \to 1$ can be described in two special cases. If $(1 - \rho_2)/(1 - \rho_1) \to \infty$ as $\rho_1 \to 1$ and $\rho_2 \to 1$, then the asymptotic behavior is as if $\rho_1 \to 1$ with $\rho_2 < 1$ fixed, i.e., $c_{d1}^2 \approx c_{s1}^2$ is asymptotically correct. On the other hand, if $(1 - \rho_2)/(1 - \rho_1) \to 0$ as $\rho_1 \to 1$ and $\rho_2 \to 1$, then the asymptotic behavior is as if $\rho_2 \to 1$ with $\rho_1 < 1$ fixed, i.e., $c_{d1}^2 \approx c_a^2$ is asymptotically correct. These different limits reveal a lack of robustness in the behavior of two queues in heavy traffic. This analysis also suggests that numerical evaluation of the

stationary distribution of the two-dimensional diffusion process should prove very useful.

Our design problem has also been studied by Wein (1988) using a heuristic approximation derived from the heavy-traffic diffusion limit, related the case of balanced loading (the traffic intensities of all the queues are high). This approximation makes the approximating arrival variability parameter at each queue after the first equal to the service-time variability parameter at the previous queue, i.e., $c_{d1}^2 \approx c_{s1}^2$ instead of (4.3) or (4.4). For two queues, Wein's approximation is consistent with (4.10) and is asymptotically correct as $\rho_1 \to 1$. Wein's approximation often does not differ much from (4.3), but it does not seem to be an improvement over (4.3). (Note that (4.4) puts more weight on c_a^2 instead of c_{s1}^2.) Even though Wein's approximation is derived from heavy-traffic considerations, it does *not* capture the bottleneck phenomenon; i.e., as explained above, $c_{d1}^2 \approx c_{s1}^2$ is asymptotically correct when $\rho_1 \to 1$ and $\rho_2 \to 1$ only if $(1 - \rho_2)/(1 - \rho_1) \to \infty$. Since Wein's approximation is a *heuristic* based on the heavy-traffic limit, poor performance of Wein's approximation does not imply poor performance based on the full multi-dimensional diffusion limit. We expect that the full diffusion limit *will* tell the proper story.

6. QUEUES WITH COMMON SERVICE-TIME VARIABILITY

Two theoretical results support the conclusion that the order should not matter much when the service-time variability parameters are all nearly identical. First, Friedman (1965) showed that the order does not matter at all when all the service-time distributions are deterministic. Second, Weber (1979) showed that the order does not matter at all when all the service-time distributions are exponential, i.e., the sojourn-time distributions are independent of the order. In both cases, the arrival process and the traffic intensities can be arbitrary. See Lehtonen (1986), Anantharam (1987) and Tsoucas and Walrand (1987) for alternate proofs of Weber's interesting result. As indicated in Whitt (1985), the approximate relative difference in (4.8) when $c_{s1}^2 = c_{s2}^2 = 0$ or when $c_{s1}^2 = c_{s2}^2 = 1$ should be viewed as a *measure of approximation error*.

Light-traffic limits in Section 4 of Greenberg and Wolff prove that the expected sojourn time for two queues in series is *not* independent of the order when the service-time distributions differ only by a scale factor, e.g., if the service-time distributions are both uniform or if they are both E_2. Our simulation results in Tables 7-9 also confirm (but of course do not prove conclusively) that the expected sojourn time is not

independent of the order in these cases, but our simulation results also indicate that the order does not matter much, as one would expect from the results of Friedman and Weber.

In Section 4 of their paper, Greenberg and Wolff show that the optimal light-traffic ordering *contradicts* general ordering heuristics $P2$ and $P3$ on p. 481 of Whitt (1985) concerning the case of queues with identical c_{si}^2, and this conclusion is supported by our simulation results. Indeed, as indicated in Whitt (1985), the approximations themselves do not strongly support heuristics $P2$ and $P3$. The quantitative estimates provided by the approximations mostly suggest (correctly) that the order does not matter in these cases.

As indicated in Section 3.2, the light-traffic asymptotics seem effective for identifying which order is best. However, the approximations also seem effective in providing a quantitative estimate of the expected sojourn time with each other. As indicated in Section 4.4, the average (maximum) error in the predicted relative difference of the expected sojourn times for two queues in series with the adjusted approximation over the 110 cases in Tables 2-9 was 2.5% (10.1%).

REFERENCES

ALBIN, S. L., ‘‘Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues,’’ *Oper. Res.* 32 (1984), 1133-1162.

ALBIN, S. L. and S. KAI, ‘‘Approximation for the Departure Process of a Queue in a Network,’’ *Naval. Res. Logist. Quart.* 33 (1986), 129-143.

ANANTHARAM, V., ‘‘Probabilistic Proof of the Interchangeability of ·/M/1 Queues In Series,’’ *Queueing Systems* 2 (1987), 387-392.

BERMAN, M. and M. WESTCOTT, ‘‘On Queueing Systems with Renewal Departure Processes,’’ *Adv. Appl. Prob.* 15 (1983), 657-673.

BITRAN, G. R. and D. TIRUPATI, ‘‘Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference,’’ *Management Sci.* 34 (1988), 75-100.

BRATLEY, P., B. L. FOX and L. E. SCHRAGE, *A Guide to Simulation*, Second Ed., Springer-Verlag, New York, 1987.

DALEY, D. J. and T. ROLSKI, ‘‘A Light Traffic Approximation for a Single-Server Queue,’’ *Math. Opns. Res.* 9 (1984), 624-628.

DALEY, D. J. and T. ROLSKI, ‘‘Light Traffic Approximations for Queues, II,’’ The University of North Carolina, Chapel Hill, 1988.

FENDICK, K. W., V. R. SAKSENA and W. WHITT, ‘‘Dependence in Packet Queues,’’ *IEEE Trans. Commun.*, 1988, to appear.

FENDICK, K. W. and W. WHITT, ‘‘Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue,’’ *IEEE Proceedings on Dynamics of Discrete Event Systems*, Y. C. Ho (ed.), 1988, to appear.

FRIEDMAN, H. D., ‘‘Reduction Methods for Tandem Queueing Systems,’’ *Oper. Res.* 13 (1965), 121-131.

GREENBERG, B. S., *Queueing Systems with Returning Customers and the Order of Tandem Queues*, Ph.D. dissertation, University of California at Berkeley, 1986.

GREENBERG, B. S., ‘‘On Niu’s Conjecture for Tandem Queues,’’ *Adv. Appl. Prob.* 19 (1987), 751-753.

GREENBERG, B. S. and R. W. WOLFF, ‘‘Optimal Order of Servers for Tandem Queues in Light Traffic,’’ *Management Sci.* 34 (1988), 500-508.

HARRISON, J. M., ‘‘The Heavy Traffic Approximation for Single Server Queues in Series,’’ *J. Appl. Prob.* 10 (1973), 613-629.

HARRISON, J. M., ‘‘The Diffusion Approximation for Tandem Queues in Heavy Traffic,’’ *Adv. Appl. Prob.* 10 (1978), 886-905.

HARRISON, J. M. and R. J. WILLIAMS, ‘‘Brownian Models of Open Queueing Networks with Homogeneous Customer Populations,’’ *Stochastics* 22 (1987), 77-115.

HILLIER, F. S. and R. W. BOLING, ‘‘The Effect of Some Design Factors on the Efficiency of Production Lines with Variable Operations Times,’’ *J. Indust. Engr.* 17 (1966), 651-658.

HILLIER, F. S. and R. W. BOLING, ‘‘On the Optimal Allocation of Work in Symmetrically Unbalanced Production Line Systems with Variable Operations Time,’’ *Management Sci.* 25 (1979), 721-728.

IGLEHART, D. L. and W. WHITT, ‘‘Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks, and Batches,’’ *Adv. Appl. Prob.* 2 (1970), 355-369.

KRAEMER, W. and M. LANGENBACH-BELZ, ‘‘Approximate Formulae for the Delay in the Queueing System GI/G/1. *Eighth Int. Teletraffic Congress*, Melbourne, 1976, 235-1-8.

LEHTONEN, T., ‘‘On the Ordering of Tandem Queues With Exponential Servers,’’ *J. Appl. Prob.* 23 (1986), 115-129.

NEWELL, G. F., *Applications of Queueing Theory*, Second Edition, Chapman and Hall, London, 1982.

PEGDEN, C. D., *Introduction to SIMAN*, Systems Modeling Corporation, State College, PA, 1984.

PINEDO, M., ‘‘On the Optimal Order of Stations in Tandem Queues,’’ *Applied Probability — Computer*

Science: The Interface, Vol. II, R. L. Disney and T. J. Ott (eds.), Birkhäuser, Boston, 307-325, 1982.

REIMAN, M. I., ‘‘Some Diffusion Approximations with State-Space Collapse,’’ *Proc. Int. Seminar on Modeling and Perf. Eval. Methodology*, eds. F. Baccelli and G. Fayolle, Springer-Verlag, Berlin, 209-240, 1983.

REIMAN, M. I., ‘‘Open Queueing Networks in Heavy Traffic,’’ *Math. Oper. Res.* 9 (1984), 441-458.

REIMAN, M. I., ‘‘Asymptotically Exact Decomposition Approximations for Queueing Networks,’’ Talk at the ORSA/TIMS National Meeting, Washington D.C., April 1988.

SEELEN, L. P., H. C. TIJMS and M. H. VAN HOORN, *Tables for Multi-Server Queues*, North-Holland, Amsterdam, 1985.

SEGAL, M. and W. WHITT, ‘‘A Queueing Network Analyzer for Manufacturing,’’ *Proceedings* 12^{th} *Int. Teletraffic Congress*, Torino, Italy, June 1988.

SURESH, S. and W. WHITT, ‘‘Arranging Queues in Series: A Simulation Experiment,’’ submitted to *Management Sci.*, 1989.

SURESH, S. and W. WHITT, ‘‘The Heavy-Traffic Bottleneck Phenomenon in Open Queueing Networks,’’ submitted to *Oper. Res. Letters*, 1989.

TEMBE, S. V. and R. W. WOLFF, ‘‘The Optimal Order of Service in Tandem Queues,’’ *Oper. Res.* 24 (1974), 824-832.

TSOUCAS, P. and J. WALRAND, ‘‘On the Interchangeability and Stochastic Ordering of ·/M/1 Queues in Tandem,’’ *Adv. Appl. Prob.* 19 (1987), 515-520.

WEBER, R. R., ‘‘The Interchangeability of Tandem ·/M/1 Queues in Series,’’ *J. Appl. Prob.* 16 (1979), 690-695.

WEIN, L. M., ‘‘Ordering tandem queues in heavy traffic,’’ Sloan School of Management, MIT, 1988.

WHITT, W., ‘‘Approximating a Point Process by a Renewal Process, I: Two Basic Methods,’’ *Oper. Res.* 39 (1982), 125-147.

WHITT, W., ‘‘The Queueing Network Analyzer,’’ *Bell System Tech. J.* 62 (1983), 2779-2815.

WHITT, W., ‘‘Approximations for Departure Processes and Queues in Series,’’ *Naval Res. Logist. Quart*. 31 (1984), 499-521.

WHITT, W., ‘‘The Best Order for Queues in Series,’’ *Management Sci.* 31 (1985), 475-487.

WHITT, W., ‘‘A Light-Traffic Approximation for Single-Class Departure Processes from Multi-Class Queues,’’ *Management Sci.* 34 (1988), 1333-1346.

WHITT, W., ‘‘Planning Queueing Simulations,’’ *Management Sci.* 35 (1989), to appear.

WOLFF, R. W., ‘‘Tandem Queues with Dependent Service in Light Traffic,’’ *Oper. Res.* 30 (1982), 619-635.

Table for Results	Variability Parameters c_a^2	c_{s1}^2	c_{s2}^2	Traffic Intensities ρ_i
different service variability				
2	0.5	0.5	2.0	
3	1.0	0.5	8.0	
4	1.0	2.0	4.0	0.3, 0.6, 0.8, 0.9
5	4.0	0.5	1.0	$(4 \times 4 = 16$ cases$)$
6	4.0	1.0	4.0	
identical service variability				
7	1.0	0.5	0.5	0.1, 0.2
				0.3, 0.6, 0.9
8	1.0	4.0	4.0	
	(H_2 with balanced means)			$(5 + 4 + 3 + 2 + 1 = 15$ cases$)$
9	1.0	4.0	4.0	
	(H_2 with unbalanced means)			

Table 1. The simulation cases for two queues in series.

Variability Parameters	Traffic Intensity ρ	Approximations		Simulation Estimate	Tabled Numerical Values	Difference	
		(4.1)	(4.2)			(4.1)	(4.2)
$c_a^2 = 0.5$	0.9	4.05	3.98	3.91	3.92	0.033	0.015
	0.8	1.60	1.54	1.48	1.49	0.074	0.034
$c_s^2 = 0.5$	0.6	0.45	0.40	0.38	0.38	0.070	0.020
	0.3	0.064	0.044	0.040	0.039	0.025	0.005
$c_a^2 = 0.5$	0.9	10.13	10.05	9.92		0.021	0.013
	0.8	4.00	3.93	3.90	not	0.026	0.008
$c_s^2 = 2.0$	0.6	1.13	1.08	1.06	available	0.066	0.019
	0.3	0.161	0.138	0.126		0.035	0.012
$c_a^2 = 4.0$	0.9	20.25	19.51	21.16	20.17	0.004	0.033
	0.8	8.00	7.42	8.10	7.86	0.018	0.056
$c_s^2 = 1.0$	0.6	2.25	1.94	2.12	2.06	0.092	0.058
	0.3	0.320	0.247	0.246	0.243	0.077	0.004

Table 10. A comparison of approximations for the expected steady-state waiting time at the first (GI/G/1) queue with simulation estimates and numerical values from Seelen, Tijms and van Hoorn (1985).

Approximation Method	Variability Parameter c_{d1}^2
stationary-interval, (4.3)	$\rho_1^2\, c_{s1}^2 + (1-\rho_1^2)\, c_a^2$
asymptotic method (limit as $\rho_2 \to 1$ or as $\rho_1 \to 0$)	c_a^2
hybrid of stationary-interval and asymptotic, (4.4)	$(1-\rho_2^{10})\rho_1^2\, c_{s1}^2 + \left[1-(1-\rho_2^{10})\rho_1^2\right] c_a^2$
Wein's heuristic based on multi-dimensional diffusion approximation (also limit as $\rho_1 \to 1$)	c_{s1}^2
Reiman's sequential bottleneck	c_a^2 if $\rho_1 < \rho_2$ c_{s1}^2 if $\rho_1 > \rho_2$ $\frac{(c_a^2 + c_{s1}^2)}{2}$ if $\rho_1 = \rho_2$

Table 11. Candidate approximate variability parameters for the departure process from the first queue of two queues in series.

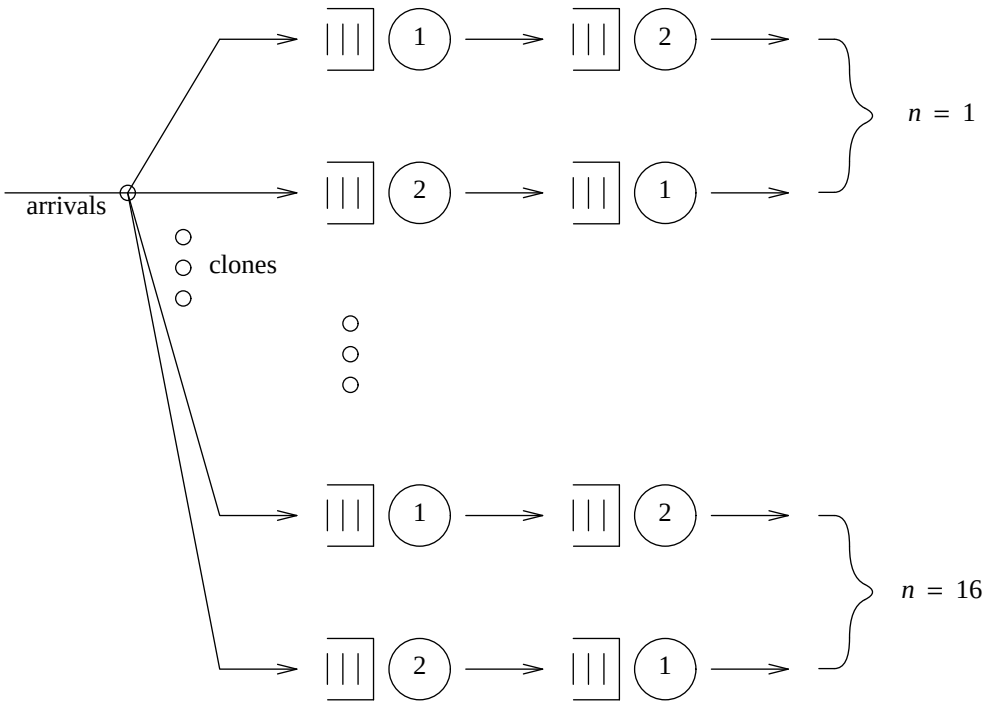

Figure 1. The modified simulation model for greater efficiency: one network with 64 queues (4 values of ρ_1, 4 values of ρ_2, 2 orders, 2 queues).