The Best Order for Queues in Series

Ward Whitt

Management Science, Vol. 31, No. 4 (Apr., 1985), 475-487.

MANAGEMENT SCIENCE
Vol. 31, No. 4, April 1985
Printed in U.S.A.

THE BEST ORDER FOR QUEUES IN SERIES*

WARD WHITT
AT & T Bell Laboratories, Holmdel, New Jersey 07733

An important design problem for queueing systems is to determine the best order for two or more service stations in series. For given external arrival process and given service-time distributions, the object is to determine the order of the stations (to be used by all customers) that minimizes the expected equilibrium sojourn time per customer. Unfortunately, very little is known about this problem because exact analysis is extremely difficult. This paper applies approximation methods recently developed for networks of queues to obtain approximation formulas and useful heuristic design principles.
(TANDEM QUEUES; APPROXIMATIONS; QUEUEING NETWORKS; QUEUEING SYSTEM DESIGN)

1. Introduction and Summary

There are many important design problems for queueing systems about which very little is known because the models tend to be intractable. However, important insights and useful practical guidelines often can be obtained by using approximations. The purpose of this paper is to show how approximation methods can be applied to one important design problem: determining the best order for two or more single-server stations in series. For given external arrival process and given service-time distributions, the object is to determine the order of the stations that minimizes the long-run average sojourn time (time in the system) per customer.

This design problem can arise, for example, in setting up a production line. In some cases the total job is divided into separate tasks that can be performed in any order. The object then is to determine in advance a specific order for the tasks to be followed by all jobs.

In our model, each station has a single server, an unlimited waiting room and the FIFO (first-in first-out) discipline. Each customer (or job) is served once at each station, with the order of the stations being the same for all customers. The stations are mutually independent and independent of the arrival process; i.e., the sequences of service times at the different stations and the sequence of interarrival times are mutually independent. The service times at station i, denoted by S_{ik}, $k \geq 1$, are i.i.d. (independent and identically distributed) with finite mean $\tau_i = ES_{i1}$ and finite squared coefficient of variation c_{si}^2 (variance divided by the square of the mean). The external arrival process is a renewal process, perhaps as a result of approximation, so that the interarrival times, denoted by T_k, $k \geq 1$, are also i.i.d. The interarrival times have finite mean $\lambda^{-1} = ET$, and finite squared coefficient of variation c_a^2. We assume that $\rho_i \equiv \lambda\tau_i < 1$ for each i, so that the system is stable and equilibrium exists. Our object is to arrange the stations to minimize the expected (equilibrium) sojourn time or, equivalently, the expected sum of delays (waiting time excluding service time) at all the stations. We use the squared coefficients of variation c_a^2 and c_{si}^2 to approximately characterize the variability of the general interarrival-time and service-time distributions.

There is of course one familiar special case. If the arrival process is a Poisson process and all the service time distributions are exponential, then the equilibrium departure processes are Poisson processes, the sojourn times (waiting time plus service

*Accepted by John P. Lehoczky; received June 3, 1983. This paper has been with the author 1 month for 1 revision.

0025-1909/85/3104/0475$01.25

time) at the different stations are independent, and the order of the stations does not affect the total sojourn-time distribution; see Burke (1956), Reich (1957), Kelly (1979), and Bermaud (1981). In fact, Weber (1979) has established the remarkable result that the final departure process is independent of the order of the stations for an arbitrary arrival process, given exponential service-time distributions at all stations. As a consequence, in this situation the distributions of the sojourn times in the entire network are independent of the order of the stations. A similar, somewhat more apparent, result had previously been established for stations with deterministic service times by Friedman (1965). As Weber (1979) notes, this property is not maintained when both exponential and deterministic servers are present.

When we drop the special exponential or deterministic assumptions, the order does matter, but unfortunately the general case is very difficult to analyze exactly, primarily because the departure processes are not renewal processes. Important theoretical results for the nonexponential case have been obtained by Tembe and Wolff (1974), and Pinedo (1982a, b), but again under very restrictive conditions, in particular, when the service times at some stations are constant or when the service times at all the stations are nonoverlapping, i.e., ordered with probability one: $P(S_{11} < S_{21}) = 1$.

Unfortunately, there seems to be nothing in the literature that suggests what to do in more general cases. Moreover, none of the results in the literature are quantitative; they do not indicate how much the order matters. In this paper we begin to address this problem. We apply approximation methods for networks of queues in Whitt (1982a, 1983a, 1984d) to obtain heuristic design principles. We provide a method, albeit approximate, to determine what to do and how much it matters. Our procedure also yields conjectures about possible theorems. Moreover, our procedure yields candidate solutions that can be tested more carefully by simulation. If there are n stations in series, then there are $n!$ possible arrangements. For example, if $n = 6$, then $n! = 720$. We can easily calculate the approximate expected sojourn time for all 720 alternatives, but we certainly would not want to simulate all 720 alternatives.

Of course, since the results here are based on approximations, we should have some solid evidence that the approximations are reasonably accurate. We do not make comparisons with simulations here, but extensive comparisons have been made and are reported in Whitt (1983b, 1984d). Based on those papers, we regard a difference in the total expected waiting time for two orderings as significant (compared with typical approximation errors) if it is greater than 10%. The examples in this paper indicate that the order often is significant.

It is important to understand how our approximation procedure works and when it tends to work well, because it does not work well in all circumstances. It is based on approximating the arrival process to each queue by a renewal process partially characterized by the first two moments of the renewal interval. Then the queues are analyzed separately as GI/G/1 queues partially characterized by the first two moments of the interarrival-time and service-time distributions. Even for actual GI/G/1 queues, the approximations are not good for all parameter values and all distributions. In particular, the reliability of the approximations deteriorates in the presence of unusually high variability, especially in the arrival process. Moreover, even for fairly reasonable parameter values, the approximations can perform poorly with unusual distributions, such as two-point distributions; see Whitt (1984a). On the other hand, the approximations usually perform reasonably well for typical distributions; see Klincewicz and Whitt (1984) and Whitt (1983b, 1984a, b).

We aim to treat the more standard probability distributions such as E_k (Erlang), M (exponential) and H_2 (hyperexponential, a mixture of two exponentials), and for H_2 we avoid the extremes, e.g., we have balanced or nearly balanced means (Whitt 1982a, §3). To illustrate, here is the kind of example we want to treat.

EXAMPLE 1. Consider two stations in series with a Poisson arrival process having rate $\lambda = 1$. Let station one have an E_2 service-time distribution with $\tau_1 = 0.9$ and $c_{s1}^2 = 0.5$ and let station two have an H_2 distribution with $\tau_2 = 0.8$ and $c_{s2}^2 = 8.0$ (and balanced means). The literature seems to provide no guidance in this case, but our approximate analysis immediately indicates that it is much better to have station one first. The approximate expected total waiting times with orderings (1, 2) and (2, 1) are 18.4 and 38.6, respectively. The difference is clearly significant, allowing for approximation error.

Since each arrival process is the departure process from the previous queue, we exploit renewal-process approximations for departure processes; see Whitt (1984d) and references there. As in Whitt (1983a), we use the stationary-interval method exclusively, which attempts to capture the stationary distribution of one interdeparture interval without taking into account the dependence among successive intervals. This approximation usually performs well, but it can perform poorly under certain heavy traffic conditions, when the asymptotic method becomes appropriate, but this situation apparently does not often arise in practice. The approximation also tends to perform poorly when several consecutive stations have deterministic or nearly deterministic service-time distributions. Then our approximate decoupling of the stations tends to be unjustified. This phenomenon is well illustrated by considering several consecutive stations with identical deterministic service times; our approximation fails to capture the pipelining effect. It is easy to see that the actual arrival process to all deterministic stations after the first is just a translated version of the departure process from the first station, whereas our approximation has the variability parameters of the successive arrival processes converge to 0 geometrically fast. In fact, in this special case the actual convergence is in one step and the limit is typically not evenly spaced deterministic arrivals. To capture the pipelining effect, we propose a modified approximation procedure; see §4. We first reduce the system by regarding two or more consecutive stations with deterministic service times as a single station for the purpose of calculating expected waiting times (excluding service times). The same procedure is usually appropriate for nonoverlapping service-time distributions too.

The rest of this paper is organized as follows. §2 presents the approximation formulas that enable us to calculate easily the approximate expected sojourn time for n stations in series. §3 discusses the case of two stations in series. §4 describes the refinements for several consecutive queues with deterministic service-time distributions. §5 presents some general heuristic design principles. §6 discusses related theoretical results. §7 considers various special cases for more than two stations in series, including the case of equal service rates and cases of bottleneck stations. Finally, §8 discusses a simple algorithm based on pairwise comparisons of adjacent stations.

2. The Basic Approximation Formulas

In this section we describe a simple procedure for calculating the approximate expected total delay for n stations in series. We begin by approximating the departure processes by renewal processes partially characterized by the first two moments of the renewal interval; see Whitt (1982a, 1983a, 1984d). We recursively apply an approximation for the departure process for a GI/G/1 queue (with a renewal arrival process).

The mean of the renewal interval in the approximating renewal process is just the mean of the interarrival time, so that the departure rate equals the arrival rate. The squared coefficient of variation of the renewal interval in the approximating renewal process is c_d^2, defined by

$$c_d^2 = \rho^2 c_s^2 + (1 - \rho^2) c_a^2, \tag{1}$$

where subscripts indexing the station have been omitted.

Iteratively applying (1), we obtain the following recursive formula for c_{dn}^2, the variability parameter of the departure process from n queues in series:

$$c_{dn}^2 = \rho_n^2 c_{sn}^2 + (1 - \rho_n^2) c_{d,n-1}^2 . \tag{2}$$

The closed-form version of (1) and (2) is

$$c_{dn}^2 = z_{n,1} c_a^2 + \sum_{k=1}^{n} z_{n,k+1} \rho_k^2 c_{sk}^2 , \qquad \text{where} \tag{3}$$

$$z_{n,k} = \prod_{j=k}^{j=n} (1 - \rho_j^2), \quad 1 \leqslant k \leqslant n, \quad \text{and} \quad z_{n,n+1} = 1. \tag{4}$$

As a consequence of (1), c_{dn}^2 is a convex combination of c_a^2 and c_{sj}^2, $1 \leqslant j \leqslant n$. The weight on c_{sj}^2 is increasing in ρ_j and decreasing in ρ_k for $k \neq j$.

Next we approximate the expected equilibrium waiting time (delay excluding service time) in a GI/G/1 queue by

$$EW = \tau\rho(c_a^2 + c_s^2)/2(1 - \rho). \tag{5}$$

This is the MFR (monotone failure rate) approximation discussed and evaluated in Whitt (1982b) and references there, which is a slight simplification of the approximation for EW used in Whitt (1983a). In Whitt (1983a) a refinement of (5) due to Kraemer and Langenbach-Belz (1976) is used when $c_a^2 < 1$, which always reduces (5). In Whitt (1983a), formula (5) is multiplied by $g(\rho, c_a^2, c_s^2)$ when $c_a^2 < 1$, where

$$g(\rho, c_a^2, c_s^2) = \exp\left[\frac{-2(1 - \rho)}{3\rho} \frac{(1 - c_a^2)^2}{(c_a^2 + c_s^2)} \right]. \tag{6}$$

The factor g tends to be significantly less than 1 when ρ, c_a^2 and c_s^2 are relatively small. For example, it is easy to check that (6) yields a significant reduction and is accurate for the D/M/1 queue. However, for simplicity we often use (5). Improved approximations can be expected from (6).

The standard heuristic algorithm, then, is to calculate the approximate value of $EW_1 + \cdots + EW_n$ for each of the $n!$ permutations and choose the smallest value. This can be done with formulas (1)–(6) or via the QNA program described in Whitt (1983a). However, using (5), we can also describe the solution in several special cases in order to obtain simple heuristic design principles.

3. Two Stations in Series

In this section we consider the special case of two stations. Let $T = W_1 + W_2$ be the total waiting time and let $T(1,2)$ and $T(2,1)$ represent the total waiting time as a function of the two possible orderings. From §2, using (5) without (6), we have the approximations

$$ET(1,2) = \frac{\tau_1\rho_1(c_a^2 + c_{s1}^2)}{2(1 - \rho_1)} + \frac{\tau_2\rho_2(\rho_1^2 c_{s1}^2 + (1 - \rho_1^2)c_a^2 + c_{s2}^2)}{2(1 - \rho_2)} \quad \text{and} \tag{7}$$

$$ET(1,2) - ET(2,1) = \frac{\tau_2\rho_2\rho_1^2(c_{s1}^2 - c_a^2)}{2(1 - \rho_2)} - \frac{\tau_1\rho_1\rho_2^2(c_{s2}^2 - c_a^2)}{2(1 - \rho_1)} . \tag{8}$$

As a consequence, for two stations in series we obtain the following simple heuristic

for ordering the stations. For each station i, calculate the quantity δ_i defined by

$$\delta_i = (1 - \rho_i)(c_{si}^2 - c_a^2) \tag{9}$$

and order the stations so that $\delta_1 \leqslant \delta_2$.

For the case $n = 2$, we can see how much the order matters via the approximation formula (8). For the special case in which $\tau_1 = \tau_2 = \tau$, so that $\rho_1 = \rho_2 = \rho$, the approximate difference is

$$ET(1,2) - ET(2,1) = \frac{\rho^4(c_{s1}^2 - c_{s2}^2)}{2\lambda(1-\rho)} . \tag{10}$$

Moreover, when $c_{s1}^2 \leqslant c_{s2}^2$ in this special case,

$$\frac{ET(2,1) - ET(1,2)}{ET(1,2)} = \frac{\rho^2(c_{s2}^2 - c_{s1}^2)}{\left((1+\rho^2)c_{s1}^2 + (2-\rho^2)c_a^2 + c_{s2}^2\right)} \leqslant \rho^2, \tag{11}$$

so that the maximum impact of the order is $100\rho^2\%$. Formula (11) shows that the order should matter less as c_a^2 increases and ρ decreases. On the other hand, the relative difference in (11) actually approaches ρ^2 as $c_a^2 \to 0$ and $c_{s1}^2 \to 0$, so that the order can be significant.

Consistent with the theoretical results in Friedman (1965) and Weber (1979), unequal c_{sj}^2 evidently matters more than unequal τ_j. When $\rho_1 \leqslant \rho_2$ and $c_{s1}^2 = c_{s2}^2 = c_s^2$, it is convenient to look at the normalized difference dividing by the waiting time in a single station having the larger traffic intensity. Then

$$\frac{2\lambda(1-\rho_2)(ET(1,2) - ET(2,1))}{\rho_2^2(c_a^2 + c_s^2)} = \frac{\rho_1^2(c_s^2 - c_a^2)(\rho_2 - \rho_1)}{(c_s^2 + c_a^2)(1-\rho_1)} \leqslant \rho_1^2 . \tag{12}$$

The normalized difference in (12) is small if either ρ_1 is small or ρ_1 is near ρ_2.

4. Pipelining: Deterministic Service Times

As indicated in §1, when several consecutive stations have deterministic service-time distributions, our approximation in §2 does not adequately represent what is happening. To capture the pipelining effect, we suggest a modification of the algorithm. We change the system before calculating the total expected delay. Any time several consecutive stations with deterministic (or nearly deterministic) service-time distributions appear in series, we replace them by the single station among them having the largest service time. It is also natural to apply this method with nonoverlapping service-time distributions; then the single station would be the one with the largest service times plus its own variability parameter. We use the reduced network to calculate the approximate total expected delay, but we use the original expected service times to calculate the expected total sojourn time.

EXAMPLE 2. Consider six stations in series with a Poisson arrival process having rate $\lambda = 1$. Let there be two M stations and four D (deterministic) stations. Let the mean service times be identical: $\tau_1 = \tau_2 = \cdots = \tau_6 = 0.8$. Consider four possible orderings: (D,D,D,D,M,M), (M,D,D,D,D,M), (M,M,D,D,D,D) and (D,M,D,M,D,D). The first three have maximal pipelining, but the fourth does not. These reduce to (D,M,M), (M,D,M), (M,M,D) and (D,M,D,M,D), respectively. The total expected service time at all stations is $6 \times 0.8 = 4.8$. The approximate expected total delays (excluding service times) before and after the reduction are displayed in

TABLE 1

The Approximate Expected Total Delay (Excluding Service Times) in Equilibrium for the Six Queues in Series in Example 2: Illustrating the Pipelining Refinement

The Arrangement	Six-Node Model by (5)	Six-Node Model by (6) using Whitt (1983a)	Reduced Model by (5)	Reduced Model by (6) using Whitt (1983a)
1. *DDDDMM* (Reduction *DMM*)	7.93	7.45	6.89	6.81
2. *MDDDDM* (Reduction *MDM*)	8.38	7.92	7.18	7.11
3. *MMDDDD* (Reduction *MMD*)	9.65	9.34	8.00	8.00
4. *DMDMDD* (Reduction *DMDMD*)	9.49	9.23	8.83	8.65

Table 1. These are computed both by the simple formula (5) and the refinement (6). It is evident that the reduction has a significant impact. Before the reduction, the orderings (M, M, D, D, D, D) and (D, M, D, M, D, D) seem to perform about equally well, but after the reduction, we clearly see the advantage of the pipelining effect. Note that the reduced formula for the (M, M, D, D, D, D) system is exact because the stationary departure process from an M/M/1 queue is Poisson. By each method, (D, D, D, D, M, M) is the preferred order; this is consistent with heuristic design principle P1 in the next section. We regard the advantage over (M, D, D, D, D, M) as marginal, however. Finally, the refinement in (6) does not seem critical for determining which order is best and how much it matters.

5. Heuristic Design Principles

The method we propose is based on a simple formula for the approximate expected total delay for any alternative, so that we have an algorithm for evaluating each alternative. However, if possible, we also want to extract heuristic design principles from the formula. For example, the results of Tembe and Wolff (1974) suggest three plausible heuristic design principles: (i) the stations should be ordered according to the variability of the service-time distributions, with the least variable distributions appearing first, e.g., so that $c_{s1}^2 \leqslant c_{s2}^2 \leqslant \cdots \leqslant c_{sn}^2$ (Theorem 2); (ii) the station should be ordered according to the mean service times, with the largest appearing first (Theorems 1 and 3); and (iii) as a combined principle, if the stations can be ordered so that $c_{s1}^2 \leqslant c_{s2}^2 \leqslant \cdots \leqslant c_{sn}^2$ and $\tau_1 \geqslant \tau_2 \geqslant \cdots \geqslant \tau_n$ both hold, then this order is desirable.

Our analysis indicates that these candidate heuristic design principles are inadequate; we introduce what we believe are better ones. (For the record, these first candidates were not put forth by Tembe and Wolff 1974.) We give evidence in support of heuristic design principle (i) in the case of equal mean service time. There is a simple intuitive explanation. Greater variability in the service times not only tends to increase the waiting time at that station, but it also tends to increase the variability of the departure process, which in turn tends to increase the waiting times at subsequent stations. Hence, it is desirable to have the less variable service times first. (There are limitations to this reasoning, however, because Theorem 1 of Pinedo (1982a) provides a counterexample. The counterexample only applies under very restricted conditions though.)

We argue that heuristic design principle (ii) above is inadequate even if the variability parameters c_{si}^2 or the entire service-time distributions are the same at each station. For the case of two stations, we propose the simple heuristic (9) for ordering

the stations. Note that (9) implies that the order should not matter whenever $c_a^2 = c_{s1}^2 = c_{s2}^2$, i.e., when the variability of all the service times and the external interarrival times are identical, as occurs in the one case we can solve analytically (Poisson arrival process and exponential service times), but (9) does not capture the full implications of Weber's (1979) result. However, when c_a^2 is near $c_{s1}^2 = c_{s2}^2$, our approximation suggests that the order will not matter much. From (9) we conclude that the order should be good if $c_{s1}^2 \leqslant c_a^2 \leqslant c_{s2}^2$. This case was illustrated in Example 1.

If $\tau_1 = \tau_2$, then $\rho_1 = \rho_2$ and $\delta_1 \leqslant \delta_2$ is achieved via (9) by having $c_{s1}^2 \leqslant c_{s2}^2$, which supports the first heuristic design principle above, under the condition of equal mean service times. When $\tau_1 = \tau_2$, it is natural to conjecture that the ordering (1,2) with $c_{s1}^2 \leqslant c_{s2}^2$ is in fact optimal if in addition the service times satisfy a stronger convex stochastic ordering; see Stoyan (1983), Whitt (1984c) and references there. Pinedo (1982a) indicates that we also need to control the variability of the arrival process. In particular, for n stations in series we make the following conjecture.

Conjecture 1. If $Ef(T_{11}) \leqslant Ef(M)$ and $Ef(S_{11}) \leqslant Ef(S_{21}) \leqslant \cdots \leqslant Ef(S_{n1})$ for all convex real-valued functions f, where M is an exponential random variable with mean λ^{-1}, then the optimal order is $(1, 2, \ldots, n)$.

We remark that the condition in the conjecture implies that $\tau_1 = \tau_2 = \cdots = \tau_n$, $c_{s1}^2 \leqslant c_{s2}^2 \leqslant \cdots \leqslant c_{sn}^2$ and $c_a^2 \leqslant 1$. For $n \geqslant 2$, this ordering is optimal by our heuristic; see §7.

If $c_{s1}^2 = c_{s2}^2 = c_s^2$ and $c_s^2 > c_a^2$, then $\delta_1 \leqslant \delta_2$ via (9) by having $\tau_1 \geqslant \tau_2$; however, if $c_a^2 > c_s^2$, then we need $\tau_1 \leqslant \tau_2$. This second case shows that, if (9) is reasonable, then candidate heuristic design principle (ii) is inadequate. For two stations, it is necessary to know the sign of $(c_{si}^2 - c_a^2)$. Moreover, the given order may not be good if $c_{s1}^2 < c_{s2}^2$ and $\rho_1 > \rho_2$, which shows that the combined heuristic design principle (iii) is also inadequate.

Our analysis in §7 produces several refined heuristic design principles for n stations in series:

(P1) If $\tau_1 = \tau_2 = \cdots = \tau_n$, then having $c_{s1}^2 \leqslant c_{s2}^2 \leqslant \cdots \leqslant c_{sn}^2$ is desirable.

(P2) If $c_a^2 \leqslant c_{s1}^2 = c_{s2}^2 = \cdots = c_{sn}^2$, then $\tau_1 \geqslant \tau_2 \geqslant \cdots \geqslant \tau_n$ is desirable.

(P3) If $c_a^2 \geqslant c_{s1}^2 = c_{s2}^2 = \cdots = c_{sn}^2$, then $\tau_1 \leqslant \tau_2 \leqslant \cdots \leqslant \tau_n$ is desirable.

Our analysis also suggests the refinement:

(P4) If $c_a^2 \leqslant c_{s1}^2 \leqslant c_{s2}^2 \leqslant \cdots \leqslant c_{sn}^2$ and $\tau_1 \geqslant \tau_2 \geqslant \cdots \geqslant \tau_n$, then the order is desirable.

Experimental evidence indicates that (P1) operates with much greater force than (P2) or (P3). The order tends to matter much more when the variability parameters differ with common means than when the means differ with constant variability parameters. This is to be expected from Weber (1979).

An important addition to the previous heuristic design principles is the consideration of the variability of the arrival process. It is easy to see that all these refined heuristics are optimal given approximation formula (9) for $n = 2$. For all n, we establish that P1, P2 and P3 are optimal given our approximation formulas and we conjecture that P4 is optimal as well (§7). These heuristic design principles suggest conjectures similar to Conjecture 1. For example, P4 suggests the modification of Conjecture 1 obtained by adding $Ef(T_1) \leqslant Ef(S_{11})$ and requiring that f be nonincreasing and convex. A conjecture related to P2 might have the services times S_{11} destributed as $(c_{sk}/c_{s1})S_{11}$ for each $k \geqslant 1$, a typical case being the exponential distribution. The idea, of course, for theorems is to have more control over the distribution than is provided by the parameters c_a^2 and c_{si}^2 when making comparisons.

Our analysis suggests that for more than two stations simple heuristics, e.g., involving only pairwise comparisons, will not work consistently (§7.4). It seems

necessary to calculate the expected total delay for every permutation and compare the results. The approximation formulas provide a simple way to do this.

6. Related Literature

It is of course important to relate our heuristics to the theoretical results which have been established. It is significant that the conclusions sometimes disagree. With such a conflict, theorems usually fare better than heuristics, but we believe that the theoretical results do not rule out the heuristics here because the theorems apply to very special situations. The deterministic and nonoverlapping service-time distributions in Pinedo (1982a) and Tembe and Wolff (1974) tend to keep the entire system tightly coupled. For our approximations, we usually need more randomness to have appropriate mixing or smoothing.

For example, Theorem 2 of Tembe and Wolff (1974) implies that the optimal order is (1,2) if $c_{s1}^2 = 0$. This is inconsistent with (9) if $c_{s2}^2 < c_a^2(\rho_1 - \rho_2)/(1 - \rho_2)$. To have inconsistency, we thus must have $\rho_1 > \rho_2$. If ρ_1 and ρ_2 are close, c_{s2}^2 will be near 0. On the other hand, if ρ_2 is small, then the order does not matter much.

The theoretical results certainly indicate limitations of our heuristics (in anticipated directions). On the other hand, our heuristics suggest that different heuristics suggested by the theorems may not be appropriate in our setting. For example, our analysis suggests that general rules of thumb involving bowl-shaped distributions of means and variances proposed by Pinedo (p. 323 of 1982a and p. 160 of 1982b) are not appropriate in our setting. (Again for the record, Pinedo did not suggest using his heuristics with an external arrival process.) To be specific, we consider another example.

EXAMPLE 3. Consider five stations in series with a Poisson arrival process having arrival rate $\lambda = 1$. Let there be two H_2 stations, two M stations and one D station. In particular, let $\tau_1 = \tau_2 = \cdots = \tau_5 = 0.8$ and let $c_{s1}^2 = c_{s2}^2 = 8.0$, $c_{s3}^2 = c_{s4}^2 = 1.0$ and $c_{s5}^2 = 0.0$. Our heuristic design principle P1 suggests the order (5,4,3,2,1), whereas Pinedo (1982a, b) suggests (1,3,5,4,2) if we apply the bowl heuristic to our model with an external arrival process. The approximate expected total equilibrium delay for (5,4,3,2,1) and (1,3,5,4,2) computed by Whitt (1983a) are 42.6 and 46.7, respectively. The difference of 9.6 percent is perhaps not conclusive, but certainly (1,3,5,4,2) should not be strongly preferred. Similarly, if the mean service times are changed to $\tau_1 = \tau_2 = 0.5$, $\tau_3 = \tau_4 = 0.8$ and $\tau_5 = 0.9$, our heuristic design principle P4 suggests again (5,4,3,2,1), whereas Pinedo's bowl heuristic suggests again (1,3,5,4,2). (P4 does not apply exactly because $c_a^2 = 1 > 0 = c_{s5}^2$.) The respective approximate total expected delays computed by Whitt (1983a) are 13.8 and 19.3. It is likely that the 39.9 percent difference in favor of P4 is significant. Finally, suppose that again $\tau_1 = \tau_2 = 0.5$, $\tau_3 = \tau_4 = 0.8$ and $\tau_5 = 0.9$, but $c_{si}^2 = 2$ for all i. Then our heuristic design principle P2 suggests again (5,4,3,2,1), whereas Pinedo's bowl heuristic suggests again (1,3,5,4,2). In this case the respective expected equilibrium delays are 23.8 and 23.0. Here the results favor Pinedo's bowl heuristic, but the difference of 3.5 percent cannot be regarded as significant. This last case is consistent with Weber (1979), which suggests that the order should not matter much when the variability parameters are all identical.

Some of the relevant theoretical results are for the related scheduling model in which all customers are initially in the first queue (Pinedo 1982a). It is clear that an external arrival process makes a difference (Example 3 is a demonstration), so that we cannot immediately transfer conclusions from the scheduling model where there is no external arrival process, but the scheduling results do apply in two ways. First, the scheduling model with m customers initially in the system can be regarded as a stationary batch arrival process with low intensity. Moreover, it is possible to regard this arrival process as a renewal process. For example, the interarrival-time would be a two-point distribu-

tion, attaching some probability p to 0 and probability $1 - p$ to some constant K. The number of customers in each batch is then geometrically distributed, but this presents no problem because the scheduling results extend to random numbers of customers initially in the system. However, the arrival rate must be low, so that each finds an empty batch system. It should be clear that this arrival process is quite unusual, so that we would not expect our approximation to perform well for it. For example, the interarrival-time distribution is one of the extremal distributions considered in Whitt (1984a).

We can also apply the scheduling results by regarding the external arrival process as the departure process from an additional station inserted before the others. The service-time distribution at this station is the original interarrival-time distribution. Since the arrival process is renewal, these service times are also independent and identically distributed. However, when we consider possible permutations of the stations, this extra first station is not free to move.

For the scheduling model, another important result has been obtained independently by Dattatreya (1978) and Muth (1979). They showed that the distributions of the sojourn times in the system are unchanged if the order of all the stations is reversed. However, it does not appear that this property can be usefully applied in our setting. It suggests that we could look for appropriate symmetry after the external arrival process has been incorporated by adding a new first station and the interdeparture times in the final departure process are represented by service times in an additional station at the end. The reversal property could be exploited in our situation whenever the departure process from the network happens to be distributed approximately the same as the arrival process for any orderings of stations in between. However, this condition implies that the order of the stations does not matter for our criterion of expected sojourn time.

7. Special Cases for More Than Two Stations

In this section we consider n stations in series, basing all our conclusions on approximation formulas (1)–(9).

7.1. *Equal Mean Service Times*: *Heuristic Design Principle* P1

Suppose that $\tau_1 = \tau_2 = \cdots = \tau_n = \tau$, so that $\rho_1 = \rho_2 = \cdots = \rho_n = \rho$. Then

$$ET = \sum_{i=1}^{n} EW_i = \frac{\tau\rho}{2(1-\rho)} \left\{ c_a^2 \sum_{i=0}^{n-1} (1-\rho^2)^i + \sum_{i=1}^{n} c_{si}^2 \left[1 + \rho^2 \sum_{k=0}^{n-i-1} (1-\rho^2)^k \right] \right\}. \quad (13)$$

From (13) we see that the desired order is independent of c_a^2 in this case. Moreover, since the coefficients of c_{si}^2 are decreasing in i, by the rearrangement theorem, p. 261 of Hardy, Littlewood, and Polya (1967), (13) is minimized by having $c_{s1}^2 \leqslant c_{s2}^2 \leqslant \cdots \leqslant c_{sn}^2$. It is also significant that the reverse order is the worst arrangement, by the same argument.

Note that the weights attached to c_{si}^2 in (13) vary from 2 to $1 + \rho^2$. Hence, it is easy to obtain a rough estimate of how much the order matters. A lower bound on the final sum in (13) is $(1 - \rho^2)(c_{s,\max}^2 - c_{s,\min}^2)$ where $c_{s,\max}^2$ and $c_{s,\min}^2$ are, respectively, the maximum and minimum c_{si}^2. This bound is obtained by considering a simple switch between the first and last stations. A crude upper bound is the lower bound multiplied by $n/2$.

7.2. *Heuristic Design Principles* P2 *and* P3

Next suppose that $c_{s1}^2 = c_{s2}^2 = \cdots = c_{sn}^2 = c_s^2$ and consider any two successive stations, say j and $j + 1$, with c_{aj}^2 the variability parameter of the arrival process to station

j. Note that

$$c_{d,j+1}^2 = (1-\rho_j^2)(1-\rho_{j+1}^2)c_{aj}^2 + \left(1-\left[(1-\rho_j^2)(1-\rho_{j+1}^2)\right]\right)c_s^2, \tag{14}$$

so that $c_{d,j+1}^2$ is independent of the order of stations j and $j+1$. Hence, we can obtain the claimed optimal order according to heuristic design principles P2 and P3 by considering a sequence of pairwise switches using (9), disregarding all other stations. For this reduction, we also use the fact that $c_{aj}^2 \leqslant c_s^2$ for all j if and only if $c_a^2 \leqslant c_s^2$.

7.3. *A More General Pairwise Switch*

Now suppose that we are considering the order of the first two stations of n stations in series in which $c_a^2 \leqslant c_{s1}^2 \leqslant c_{s2}^2 \leqslant c_{sk}^2$ and $\tau_1 \geqslant \tau_2 \geqslant \tau_k$ for $k > 2$. We show that the order $(1,2)$ is better than the order $(2,1)$. Switching to $(2,1)$ increases the total expected waiting time by an amount described by (8), which is bounded below by

$$ET(2,1) - ET(1,2) \geqslant \frac{\rho_1^2\rho_2^2(c_{s2}^2 - c_{s1}^2)}{2\lambda(1-\rho_2)}. \tag{15}$$

On the other hand, it is easy to see that switching from $(1,2)$ to $(2,1)$ lowers c_{d2}^2 by $\rho_1^2\rho_2^2(c_{s2}^2 - c_{s1}^2)$. Here there is a tradeoff in going from $(1,2)$ to $(2,1)$: increase the total expected waiting time at the first two stations and decrease the total expected waiting time at all subsequent stations. However, since $\tau_k \leqslant \tau_2$ and $c_{sk}^2 \geqslant c_{s2}^2$ for all k, the decrease caused in the total expected waiting times at all subsequent stations by a decrease in c_{d2}^2 of γ is

$$\gamma \sum_{k=3}^{n} \frac{\rho_k^2\left[\prod_{j=1}^{k-1}(1-\rho_j^2)\right]}{2\lambda(1-\rho_k)} \leqslant \frac{\gamma}{2\lambda(1-\rho_2)} \sum_{k=3}^{n} \rho_k^2\left[\prod_{j=1}^{k-1}(1-\rho_j^2)\right] \leqslant \frac{\gamma}{2\lambda(1-\rho_2)}. \tag{16}$$

Here $\gamma = \rho_1^2\rho_2^2(c_{s2}^2 - c_{s1}^2)$, so that the decrease in total expected waiting time at all stations after the second is bounded above by the bound in (15). Hence, $(1,2)$ is better than $(2,1)$ for any arrangement of any number of additional stations having $\tau_k \leqslant \tau_2$ and $c_{sk}^2 \geqslant c_{s2}^2$ for $k > 2$. A similar analysis applies if $c_a^2 \geqslant c_{s1}^2 \geqslant c_{s2}^2 \geqslant c_{sk}^2$ and $\tau_1 \leqslant \tau_2 \leqslant \tau_k$ for all $k > 2$. This comparison is a possible tool for proving P4 by considering only pairwise switches, but note that we need c_a^2 appropriately related to c_{s1}^2 and c_{s2}^2 for the two stations under consideration. The more general principle P4 remains a conjecture.

7.4. *A Bottleneck Station*

Next suppose that one station is a bottleneck in that it has a much higher traffic intensity than any other station. If this traffic intensity is sufficiently high, then the total expected delay is dominated by the expected delay at this station and the object is to minimize the c^2 of the arrival process to the bottleneck station. By (5), all stations with $c_{si}^2 < c_a^2$ should appear before the bottleneck station and all stations with $c_{si}^2 > c_a^2$ should appear afterwards. If there are k stations before the bottleneck with $c_{si}^2 < c_a^2$, then to determine their order the object is to minimize

$$c_{dk}^2 = c_a^2 \prod_{j=1}^{k}(1-\rho_j^2) + \sum_{i=1}^{k} \rho_i^2 c_{si}^2 \prod_{j=i+1}^{j=k}(1-\rho_j^2). \tag{17}$$

From (17) we see that the value of c_a^2 does not affect the order. To find the desired order of the k stations before the bottleneck, we can solve (17) for each of the $k!$ permutations and choose the smallest one. When the mean service times of these k stations are identical, we can apply the rearrangment theorem again to conclude that we should have $c_{s1}^2 \geqslant c_{s2}^2 \geqslant \cdots \geqslant c_{sk}^2$, which is the reverse of P1.

TABLE 2
The Service-Time Parameters for Seven Queues in Series Discussed in Example 4

Node	1	2	3	4	5	6	7
τ_j	0.5	0.5	0.5	0.5	0.5	0.5	0.9
c_{sj}^2	1	2	3	5	6	7	4

EXAMPLE 4. We illustrate two special cases by considering seven stations in series with arrival-process parameters $\lambda = 1$ and $c_a^2 = 4$ and service-time parameters in Table 2. By the discussion about the bottleneck station above, the suggested order of the seven stations is (3, 2, 1, 7, 4, 5, 6). By the same reasoning, the worst order should not be the reverse of the best order, i.e., (6, 5, 4, 7, 1, 2, 3), but instead (4, 5, 6, 7, 3, 2, 1). The approximate sum of the expected waiting times by (1)–(6) is 39.1 for (3, 2, 1, 7, 4, 5, 6), 48.4 for (6, 5, 4, 7, 1, 2, 3) and 49.6 for (4, 5, 6, 7, 3, 2, 1). The difference is primarily due to station 7; the expected waiting time at station 7 is 27.3 for (3, 2, 1, 7, 4, 5, 6), 36.6 for (6, 5, 4, 7, 1, 2, 3) and 37.5 for (4, 5, 6, 7, 2, 1). However, if station 7 is removed, then all service rates are equal and the suggested order by principle P1 is (1, 2, 3, 4, 5, 6). The sum of the expected waiting times by (1)–(6) is 11.2 for (1, 2, 3, 4, 5, 6) and 12.7 for (6, 5, 4, 3, 2, 1). This example indicates that myopic comparisons will not suffice. The relative position of stations 1 and 2 depends on the absence or presence of station 7.

7.5. *Two Bottleneck Stations*

To examine additional complexity, now consider the case in which there are two bottleneck stations with equal high traffic intensities. By the discussion above, these two should be ordered so that the one with less variable service time appears first, i.e., so that $c_{s1}^2 \leqslant c_{s2}^2$ for these two. Since the bottleneck stations have very high traffic intensity, the departure process variability parameter is nearly equal to the service time variability parameter for these stations. Hence, all those other stations with $c_{si}^2 > \max\{c_a^2, c_{s1}^2\}$ belong after the second bottleneck station. If $c_a^2 > c_{s1}^2$, then all stations with $c_a^2 > c_{si}^2 > c_{s1}^2$ belong before the first bottleneck station. On the other hand, if $c_a^2 < c_{s1}^2$, then all stations with $c_a^2 < c_{si}^2 < c_{s1}^2$ belong between the two bottleneck stations. It remains to determine where the stations with $c_{si}^2 < \min\{c_a^2, c_{s1}^2\}$ belong before the second bottleneck station.

EXAMPLE 5. To illustrate the two-bottleneck case, consider four stations in series with $\lambda = 1$, $c_a^2 = 4$ and the service-time parameters in Table 3. From the discussion above, station 3 should come before station 4 and stations 1 and 2 should appear before station 4. If stations 1 and 2 appear together, then station 2 should appear first. Hence, we have four permutations to consider: (1, 3, 2, 4), (2, 3, 1, 4), (2, 1, 3, 4) and (3, 2, 1, 4). These are evaluated in Table 4 together with the permutation (4, 3, 2, 1), which we anticipate will be relatively poor. It turns out that (3, 2, 1, 4) is slightly better than the other main candidates, but the differences do not seem significant in view of typical approximation errors. However, all four main candidates are significantly better than the fifth permutation (4, 3, 2, 1).

TABLE 3
The Service-Time Parameters for Four Queues in Series Discussed in Example 5

Node	1	2	3	4
ρ_j	0.5	0.5	0.9	0.9
c_{sj}^2	1	3	7	12

TABLE 4

The Congestion Measures for 5 of the 24 Possible Permutations in Example 5

	Candidate Permutation				
	(1, 3, 2, 4)	(2, 3, 1, 4)	(2, 1, 3, 4)	(3, 2, 1, 4)	(4, 3, 2, 1)
c_{d1}^2	3.25	3.75	3.75	6.43	10.48
c_{d2}^2	6.29	6.38	3.06	5.57	7.66
c_{d3}^2	5.47	5.04	6.25	4.42	6.49
EW_1	1.25	1.75	1.19	1.64	1.88
EW_2	2.32	1.85	1.75	2.36	2.67
EW_3	41.51	43.54	40.74	44.55	70.79
EW_4	70.75	69.01	73.91	66.50	64.80
ET	115.83	116.15	117.59	115.05	140.14

Note. The mean waiting times EW_j are indexed as originally specified, but the variability parameters c_{dj}^2 are indexed according to the order of the station in that permutation.

8. Simple Heuristics

The bottleneck examples in §7 indicate that very simple heuristics cannot be expected to perform well in all cases. In general, it seems desirable to calculate the expected total delay for each of the $n!$ permutations as described in §2.

A simple heuristic that might work reasonably well if there are no bottlenecks is to calculate δ_i in (9) for each station i and then order the stations so that $\delta_1 \leqslant \delta_2 \leqslant \cdots \leqslant \delta_n$. This can be done easily without computer assistance. Although this ordering is optimal by our heuristics for $n = 2$, it is of course not for $n > 2$ because $c_{aj}^2 \neq c_a^2$ for $j \geqslant 2$. Nevertheless, this ordering can be a reasonable candidate. Refinements could be made afterwards by computing c_{di}^2 for each station in this permutation and looking for pairwise switches of adjacent stations to obtain improvement. When considering the stations k and $k+1$ for possible switching, use (9) again but replace c_a^2 with $c_{d,k-1}^2$. The idea is to make one or more pairwise switch, then recompute c_{di}^2 for each station in the new permutation, and look for pairwise switches again. The bottleneck examples suggest, however, that looking for successive local improvements by considering only two stations in isolation is not always effective. This simple heuristic seems promising if the traffic intensities are not relatively high and do not vary too much from station to station.[1]

[1] I am grateful to my colleague Anne Seery for writing and running the QNA program (Whitt 1983a). I am also grateful to Michael Pinedo and Ronald Wolff for bringing references Pinedo (1982a) and Weber (1979) to my attention.

References

BREMAUD, P., *Point Processes and Queues*, Springer-Verlag, New York, 1981.

BURKE, P. J., "The Output of a Queueing System," *Oper. Res.*, 4 (1956), 699–704.

DATTATREYA, E., "Tandem Queueing System with Blocking," Ph.D Dissertation, Dept. of Industrial Eng. and Oper. Res., University of California, Berkeley, 1978.

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

HARDY, G. H., J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, 2nd Ed., University Press, Cambridge, England, 1967.

KELLY, F. P., *Reversibility and Stochastic Networks*, John Wiley and Sons, New York, 1979.

KLINCEWICZ, J. G. AND W. WHITT, "On Approximations for Queues. II. Shape Constraints," *AT&T Bell Lab. Tech. J.*, 63 (1984), 139–161.

KRAEMER, W. AND M. LANGENBACH-BELZ, "Approximate Formulae for the Delay in the Queueing System GI/G/1," Eighth Internat. Teletraffic Cong., Melbourne, 235-1-8, 1976.

MUTH, E., "The Reversibility Property of Production Lines," *Management Sci.*, 25 (1979), 152–159.

PINEDO, M., "On the Optimal Order of Stations in Tandem Queues," in *Applied Probability—Computer Science: The Interface*, Vol. II, R. L. Disney and T. J. Ott (Eds.), Birkhäuser, Boston, 1982a, 307–325.

———, "Minimizing the Expected Makespan in Stochastic Flow Shops," *Oper. Res.*, 30 (1982b), 148–162.

REICH, E., "Waiting Times When Queues Are in Tandem," *Ann. Math. Statist.*, 28 (1957), 768–773.

STOYAN, D., *Comparison Methods for Queues and Other Stochastic Models*, John Wiley and Sons, New York, 1983. (English translation and revision edited by D. J. Daley of *Qualitative Eigenshaften und Abschatzungen Stochasticher Modelle*, 1977.)

TEMBE, S. V. AND R. W. WOLFF, "The Optimal Order of Service in Tandem Queues," *Oper. Res.*, 24 (1974), 824–832.

WEBER, R. R., "The Interchangeability of Tandem ·/M/1 Queues in Series," *J. Appl. Probab.*, 16 (1979), 690–695.

WHITT, W., "Approximating a Point Process by a Renewal Process. I. Two Basic Methods," *Oper. Res.*, 30 (1982a), 125–147.

———, "The Marshall and Stoyan Bounds for IMRL/G/1 Queues Are Tight," *Oper. Res. Letters*, 1 (1982b), 209–213.

———, "The Queueing Network Analyzer," *Bell System Tech. J.*, 62 (1983a), 2779–2815.

———, "Performance of the Queueing Network Analyzer," *Bell System Tech. J.*, 62 (1983b), 2817–2843.

———, "On Approximations for Queues. I. Extremal Distributions," *AT&T Bell Lab. Tech. J.*, 63 (1984a), 115–138.

———, "On Approximations for Queues. III. Mixtures of Exponential Distributions," *AT&T Bell Lab. Tech. J.*, 63 (1984b), 163–175.

———, "Minimizing Delays in the GI/G/1 Queue," *Oper. Res.*, 32 (1984c), 41–51.

———, "Approximations for Departure Processes and Queues in Series," *Naval Res. Logist. Quart.*, 31 (1984d), 499–521.