Fill-Rate Bottlenecks in Production-Inventory Networks

Paul Glasserman • Yashan Wang
Columbia Business School, New York, New York 10027
MIT Sloan School, Cambridge, Massachusetts 02142

The bottleneck in a production-inventory network is commonly taken to be the facility that most limits flow through the network and thus the most highly utilized facility. A further connotation of "bottleneck," however, is the facility that most constrains system-wide performance or the facility at which additional resources would have the greatest impact. Adopting this broader sense of the term, we look for *fill-rate bottlenecks:* facilities in a production-inventory network that most constrain the system-wide fill rate (the proportion of demands filled within a fixed delivery leadtime) or facilities at which either additional production capacity or additional inventory would have the greatest impact on the fill rate.

We consider systems in which various components are produced through a series of stages holding intermediate inventories and are then assembled into finished goods to meet external demands. With each station in the network we associate precise measures of the station's propensity to constrain the fill rate. We call a station with a minimal measure a fill-rate bottleneck and justify this label both theoretically and numerically. Examples show that even the least utilized facility can be a fill-rate bottleneck. Unlike utilization, our bottleneck criteria capture information about process variability.
(*Multistage Assemble-to-Order System*; *Response Time*; *Order Fill Rate*; *Bottleneck*; *Leadtime*; *Inventory*)

1. Introduction and Summary

In the context of production-inventory systems—or any other setting in which jobs or materials flow through a network of resources—a *bottleneck* is generally taken to be the facility or resource that most constrains flow. A focus on bottlenecks influences much of both the theory and practice of operations management, and can be found in the purely practitioner-oriented literature—as in Goldratt (1990), Goldratt and Cox (1985), and Umble and Srikanth (1990)—in highly mathematical models—such as Chen and Mandelbaum (1991) and Harrison and Wein (1990)—and many places in between, including most textbooks on operations. This is hardly surprising, since focusing on bottlenecks is a way of directing attention to the root of a problem and a way of simplifying either a physical process or a mathematical model.

But if the primary notion of "bottleneck" deals narrowly with constraining flow, other connotations are sometimes implicit in its usage: the bottleneck as the facility where additional resources would have the greatest impact, or the bottleneck as the facility that most determines system-wide performance. The perspective associated with Goldratt, for example, focuses on throughput as the main measure of performance. In this case, the most highly utilized facility does determine system-wide performance, and adding capacity at this facility is the *only* way to affect the maximum sustainable throughput.

But these various senses of "bottleneck" need not coincide if we consider other measures of performance.

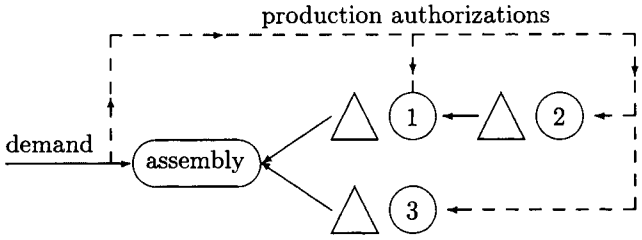
1523-4614/99/0101/0062$05.00

Production-inventory networks in which physical capacity limits and resulting congestion are modeled explicitly have features in common with queueing systems. It is well known in the queueing literature that whereas utilization is determined entirely by *mean* rates of arrival and service, measures like customer delay are sensitive to *variability* in arrival and service times as well. This can be seen explicitly in, for example, formulas for the average waiting time in an $M/G/1$ queue or heavy-traffic approximations to the general single-server queue. This observation is familiar in the literature on models of manufacturing systems; indeed, "the corrupting influence of variability" (as it is called by Hopp and Spearman (1996)) may be considered another central theme of operations management.

Our objective in this article is to put forth measures of a facility's propensity to constrain or dominate performance in a production-inventory network when performance is measured through a service level. More specifically, we work with the *fill rate,* which we take to be the fraction of orders filled within a target delivery leadtime. We analyze system behavior at high fill rates and ask, "At which facility would additional resources have the greatest impact on service? Which facility most constrains or determines the system-wide fill rate?" The measures we propose capture information about both production and demand variability (though perhaps not in an obvious way); consequently, our *fill-rate bottlenecks* need not coincide with the most highly utilized facilities.

A precise specification of the models with which we work will be given in the next section. At this point, we give only a rough description in order to put our comments in context. The essential features of our general setting are illustrated in Figure 1. Circles denote production facilities and triangles denote inventories. Station 1 draws material from the store of items completed at station 2; items completed at stations 1 and 3 are assembled into finished goods. Production and assembly are triggered by the arrival of orders: each order generates a production authorization at every station. Thus, production follows a *one-for-one replenishment* or *base-stock* policy in which each station continues to produce until its inventory of processed items reaches a target (base-stock) level. Production and assembly times are variable; each station produces one

Figure 1 **Circles Denote Production Facilities and Triangles Denote Inventories. Station 1 Uses the Output from Station 2 and the Output from Stations 1 and 3 Are Assembled to Become the Final Product. Orders for the Final Product Trigger Production.**



unit at a time, but the assembly process is uncapacitated in the sense that there is no upper limit on the number of finished goods that can be assembled in parallel.

An order is considered to be filled on time if the time elapsed from the arrival of the order until it is filled does not exceed a fixed *delivery leadtime* x. The fill rate is the long-run fraction of orders filled within this delivery leadtime, and we refer to the complement of the fill rate as the *unfill rate.* Clearly, the fill rate should increase as either the delivery leadtime or the base-stock levels increase. With each station i, we associate a number γ_i measuring that station's propensity to constrain the fill rate as the delivery leadtime increases. Like the usual utilization measure ρ_i, the new measure γ_i is a genuine feature of the station and the order stream in the sense that it can be evaluated by viewing the station in isolation from the rest of the network. We also associate a number α_i with station i that measures its constraining propensity at high levels of inventory. Unlike γ_i, the second measure α_i depends on the system-wide allocation of inventory in addition to purely local characteristics of a station. Unlike ρ_i, both γ_i and α_i capture information about production and demand variability. By comparing values of γ_i or α_i across stations we identify fill-rate bottlenecks. We state a precise result in §2 whose intuitive content is

$$\begin{aligned} &\text{unfill rate} \approx e^{-(\min_i \gamma_i)x} \quad \text{for large } x; \\ &\text{unfill rate} \approx e^{-(\min_i \alpha_i)s} \quad \text{for large } s, \end{aligned} \tag{1}$$

where x, as before, is the delivery leadtime, s is a measure of system-wide inventory, and the minimum in each exponent is taken over all stations in the network.

(We postpone to the next section a discussion of the precise sense in which these approximate relations hold.) Thus, the stations with the smallest values of γ_i and α_i are the ones that most constrain the fill rate and, in this sense, act as bottlenecks.

To illustrate our notions of fill-rate bottlenecks, we present three simple examples.

EXAMPLE 1. This example fits the scheme depicted in Figure 1, except that at this point we omit the intermediate inventories (all base-stock levels are zero) and we take the final assembly operation to be instantaneous. Orders and processing times are variable; the first panel of Table 2 summarizes the distributions used for this example. As indicated in Table 1, the utilization levels at stations 1, 2, and 3 are 95%, 80%, and 90%, respectively. In particular, station 1 has the highest utilization. In contrast, Table 1 shows that station 2 has the smallest γ_i so we predict that the system-wide unfill rate is dominated by station 2. More precisely, if we apply (1) to each station in isolation we arrive at

$$\text{unfill rate for station } i \text{ in isolation} \approx e^{-\gamma_i x}$$

$$\text{for large } x, \quad i = 1, 2, 3.$$

For the system as a whole, (1) predicts an unfill rate approximated by $\exp(-\gamma_2 x)$, since γ_1 and γ_3 are larger than γ_2. If these approximations are valid, then the logarithm of the unfill rate for station 2 in isolation should be roughly the same as the logarithm of the system-wide unfill rate, whereas the logarithms of the unfill rates at stations 1 and 3 should be much smaller, especially for large x. By comparing station-level unfill rates with the system-wide unfill rate we get an indication of which facility most constrains service.

Figure 2 plots the logarithms of the ratios of the three station-level unfill rates to the system-wide unfill rate against increasing values of the delivery leadtime x, corresponding to higher fill rates. (The graph compares values obtained through simulation, not from the approximations.) The striking conclusion from the

Table 1 **Comparison of New and Traditional Bottleneck Measures. The Station with the Smallest γ_i or α_i (Not the Largest ρ_i) Acts as a Bottleneck in Each Example**

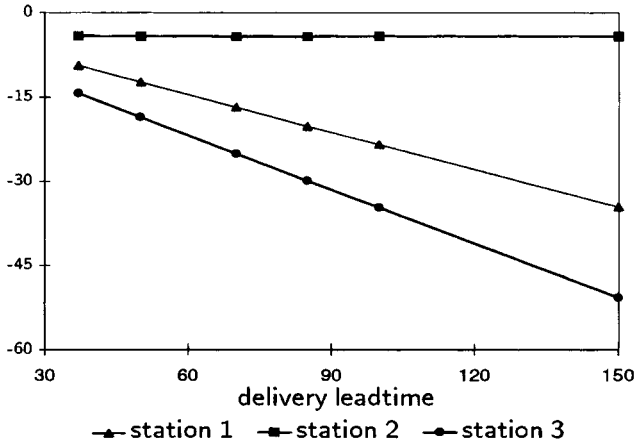
		Station		
		1	2	3
Example 1	ρ_i	0.95	0.80	0.90
	γ_i	0.32	0.10	0.42
Example 2	ρ_i	0.60	0.90	0.75
	γ_i	0.23	0.69	0.55
Example 3	ρ_i	0.60	0.90	0.75
	α_i	0.10	0.31	0.04

Table 2 **Distributions Used in the Examples, the First Panel for Example 1, the Second for Examples 2 and 3**

	Distribution	Mean	C.V.
Order Stream	Erlang	1	$1/\sqrt{3}$
Station 1	Constant	0.95	0
Station 2	Hyperexponential	0.80	2
Station 3	Erlang	0.90	$1/\sqrt{5}$
Order Stream	Erlang	1	$1/\sqrt{3}$
Station 1	Hyperexponential	0.60	2
Station 2	Constant	0.90	0
Station 3	Constant	0.15	0

Note. Order stream refers to times between arrival of customer orders. The distribution associated with each station is the processing time at that station. "C.V." is coefficient of variation, the ratio of the standard deviation to the mean. In Examples 2 and 3, finished goods require a combination of one unit from station 1 and geometrically many from station 3. The mean of the geometric distribution is 5.

Figure 2 **Comparison Between System-Wide and Station-Level Unfill Rates. The Vertical Axis Is the Logarithm of the Ratio of the Station-Level Unfill Rate to the System-Wide Unfill Rate.**



graph is that, as the fill rate increases, stations 1 and 3 become negligible; the system-wide fill rate is determined by that of station 2, *even though this station has the lowest utilization*.

Example 2. For our next example, we make some modifications to the basic system. The distributions are now as in the second panel of Table 2; in addition, each customer order now requires combining one unit completed at station 1 with a variable number of units from station 3. The variability in batch size is represented by a geometric distribution with mean 5. As a result of these changes, the utilization levels at stations 1–3 are now 60%, 90%, and 75%, respectively.

Where would additional production capacity have the greatest impact on service? To address this question we examine the effect of a 0.05 reduction in utilization at each facility, with the coefficients of variation held fixed. The results are summarized in Table 3.

In the base case, with a delivery leadtime of 11 the fill rate is 93.0%. A 5 percentage-point reduction in utilization at station 3 does not change the fill rate; the same reduction in utilization at station 2, which has by far the highest utilization and is thus the throughput bottleneck, lifts the fill rate only slightly, to a level of 93.8%; however, a 5 percentage-point reduction in utilization at station 1, the least utilized, boosts the fill rate to 95.2%. *The same increase in capacity has almost three times as big an impact on service at the least utilized station as at the highest utilized station.* The same conclusion holds at a delivery leadtime of 17. These observations are fully consistent with values in Table 1, showing that station 1 has the smallest γ_i and is therefore the station that most constrains the fill rate.

Example 3. For our last example we introduce inventories and examine increases in fill rate through increases in inventory levels. We use the same distributions as in the base case of Example 2, but we set the delivery leadtime to zero to consider the off-the-shelf fill rate. Each station now keeps inventories of units that have completed processing at that station but have not completed any subsequent stage of production; thus, units in store at station 2 are ready for processing by station 1, and units at stations 1 and 3 are ready to be combined to fill customer orders. New demands trigger production orders just as in the model without inventories. For the initial allocation, we assign the same total inventory to station 3 as to the subsystem consisting of stations 1 and 2, and more inventory at 1 than at 2.

With the initial allocations for stations 1, 2, and 3 set at 25, 5 and 30, respectively, the fill rate is 90.4%. Suppose we can keep 10 additional units of inventory at any one facility. Where should they be added to achieve the greatest improvement in the fill rate? (In comparing inventories at different stations we may be comparing apples and oranges; later we address the issue of making the comparison in common units.) Table 4 compares fill rates under different allocations. As shown there, the fill rate is almost unchanged if the increase is made at station 2 (the one with the highest utilization) or station 1 (the fill-rate bottleneck in Example 2), or shared between them. But the fill rate jumps to 95.4% if the additional inventory is kept at station 3. In this sense, station 3 is a fill-rate bottleneck with respect to increases in inventory levels. This is

Table 3 Comparison of Fill Rate Improvement After a 5 Percentage-Point Decrease in Utilization at Each Station

		fill rate (in %) after a utilization decrease at each station		
delivery leadtime	initial fill rate	station 1	station 2	station 3
11	93.0	95.2	93.8	93.0
17	98.2	99.1	98.5	98.2

Note. Current utilization levels at the three stations are 60%, 90% and 75% respectively.

Table 4 Comparison of Fill Rate Improvement After Inventory Level Increase at Each Station. Utilization Levels at the Three Stations are 60%, 90% and 75%, Respectively

	station inventories			
	1	2	3	fill rate (in %)
initial inventory	25	5	30	90.4
add at 1	35	5	30	90.5
add at 2	25	15	30	90.5
add at 1 and 2	30	10	30	90.5
add at 3	25	5	40	95.4

consistent with (1) and the last row of Table 1, which shows that station 3 has the smallest value of α_i. *The station at which additional inventory has the greatest impact on service need not coincide with the most highly utilized station or even with the station at which additional production capacity would have the greatest impact.*

The examples above consider three different ways the fill rate in the simple system of Figure 1 can become large: through an increase in the delivery leadtime, through an increase in production capacity, and through an increase in inventory. In each case, a facility other than the most highly utilized one displayed bottleneck behavior. We have suggested that these phenomena could have been anticipated from the values of our bottleneck measures in Table 1. A more precise theoretical justification will be developed in the remainder of the paper.

The rest of this paper is organized as follows. Section 2 gives a precise specification of the models we consider and states our central result. As explained there, calculation of γ_i and α_i requires complete distributional information; as a simplification we introduce two-moment approximations to these quantities. These approximations have greater potential for use in practice and also make explicit the role of variability. Section 3 gives additional numerical examples supporting both our basic measures and their approximations. Section 4 records some concluding remarks; the proof of our main result is in an appendix.

2. Main Results

2.1. Model Details and Notation

Figure 3 illustrates the general class of production-inventory networks we consider. Multiple *components* are produced through series of stages. Each stage has a dedicated facility (the circle) and a store of WIP inventory (the triangle). Each production facility processes one unit at a time and operates under an echelon base-stock (or *one-for-one replenishment*) policy, meaning that each demand triggers a production order at every station. (To keep the figure from becoming too complicated we have omitted the path of production authorizations included in Figure 1; they work the same way in this general case.) Initially, a product consists of one unit of each component; we discuss the extension to batch demands in Remark (d), below, after stating our main result. Throughout we assume that the final assembly operation is uncapacitated in the sense that there is no limit (beyond the availability of components) on the number of finished goods that can be assembled in parallel. Assembly times for different orders are i.i.d. and assumed bounded by the delivery leadtime. The *response time* of an order is the time elapsed from the moment it arrives until the moment it is filled.

We use the following notation, often modified by subscripts and superscripts:

A = order interarrival time;
B = unit production interval;
U = assembly time;
R = response time;
s = base-stock level;
x = delivery leadtime.

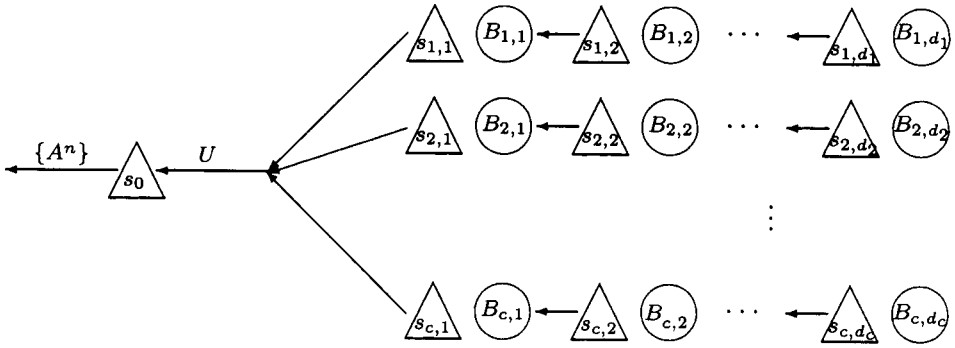
A double subscript i, j refers to stage j of component i; a single subscript j refers to component j, or stage j if there is only one component in the systems; a superscript n refers to the nth order. For example, $B_{i,j}^n$ is the nth production interval at stage j of component i; R_i^n is the time taken to fill the nth order's demand for component i, and R^n is the time taken to fill the nth order completely. An A or B without a superscript refers to a generic version of the random variable. An R without a superscript refers to a steady-state response time. Production intervals and interarrival times are all independent of each other.

For any random variable Y, the symbol ψ_Y denotes the function

$$\psi_Y(\theta) = \log \mathsf{E}[e^{\theta Y}], \tag{2}$$

called the *cumulant generating function* (c.g.f.) (also called the *logarithmic moment generating function*) of Y. The function ψ_Y is convex, and it is differentiable in the interior of its domain (the set of θ at which it is finite). The c.g.f.s of all our input random variables will be finite for some $\theta > 0$, and this implies $\psi_Y'(0) = \mathrm{E}[Y]$ and $\psi_Y''(0) = \mathrm{Var}[Y]$. See Chapter 3 of Kendall (1987) for relevant background; see Glasserman (1997) for several examples of c.g.f.s for distributions commonly used in inventory models.

Figure 3 **Multiple Components Assembled to Product After Serial Production. The Interarrival Times for Product Demand Are $\{A^n\}$. At Stage j, Component i Has a Generic Production Interval $B_{i,j}$ and a Local Base Stock Level $s_{i,j}$. Finished Goods Base-Stock Level Is s_0. Assembly Operation Is Modeled as a Random Delay U. The Arrows Indicate the Direction of Material Flow. The Production Authorization Paths Are Not Shown Explicitly But Would Be Just as in Figure 1.**



2.2. The Fill Rate and Bottleneck Characterizations

Let $s = s_0 + \Sigma_{i,j} s_{i,j}$ be the total inventory level in the system. Let $\bar{s}_{i,j} = s_0 + \Sigma_{k=1}^{j} s_{i,k}$ be the echelon inventory level at station (i, j), and $\pi_{i,j} = \bar{s}_{i,j}/s$ be the proportion of echelon (i, j) base-stock level to the total inventory level. In comparing systems with different total levels of inventory, we keep the fractions $\pi_{i,j}$ fixed. This assumes that the proportion of total inventory held in each facility remains constant, though we could just as easily assume that the proportions of inventory *dollars* (or any similar measure) remain constant, as explained in Remark (c), below. For stability, we assume that $\mathrm{E}[B_{i,j}] < \mathrm{E}[A]$ for all i, j.

Suppose $\gamma_{i,j} > 0$ solves

$$\psi_{(B_{i,j}-A)}(\gamma_{i,j}) = 0 \tag{3}$$

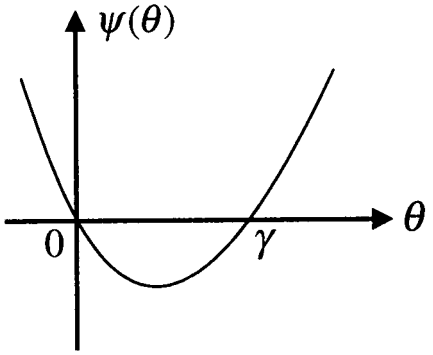
and set $\beta_{i,j} = \psi_{B_{i,j}}(\gamma_{i,j})$, $\alpha_{i,j} = \pi_{i,j}\beta_{i,j}$. The convexity of c.g.f.s implies that a positive solution to (3) is unique if it exists (see Figure 4), and existence holds under virtually all commonly used distributions. Note that the assumption of independence between A and $B_{i,j}$ implies

$$\psi_{(B_{i,j}-A)}(\gamma_{i,j}) = \psi_{B_{i,j}}(\gamma_{i,j}) + \psi_A(-\gamma_{i,j}),$$

which is often used to solve Equation (3).

The fill rate is $P(R \le x)$, the probability that an order is filled within the delivery leadtime. To emphasize its dependence on inventory levels, we frequently write R as $R(s)$. The next result captures the dominant effects of individual stations on the rate of decrease of $P(R(s) > x)$, the unfill rate.

Figure 4 **Graph of ψ and γ. Due to Convexity, $\psi(\theta) = 0$ Has Two Roots, One Is 0, the Other Is $\gamma > 0$.**



THEOREM 1. *Suppose the solutions $\gamma_{i,j}$ all exist. Then*

$$\lim_{x\to\infty} -\frac{1}{x}\log P(R(s) > x) = \gamma, \tag{4}$$

where $\gamma = \min_{i,j} \gamma_{i,j}$; and

$$\lim_{s\to\infty} -\frac{1}{s}\log P(R(s) > x) = \alpha, \tag{5}$$

given that all $\pi_{i,j} > 0$, where $\alpha = \min_{i,j} \alpha_{i,j}$.

REMARKS. (a) We can write the limiting results in Theorem 1 as the following approximations:

$$P(R(s) > x) = e^{-\gamma x + o(x)} \quad \text{as } x \to \infty; \tag{6}$$

$$P(R(s) > x) = e^{-\alpha s + o(s)} \quad \text{as } s \to \infty, \tag{7}$$

where $o(x)$ denotes a quantity for which $o(x)/x \to 0$ as x increases and $o(s)$ is defined similarly. Equation (6) suggests that when the fill rate is high due to slack in the delivery leadtime, it is most constrained by the station with the smallest γ, which we call the *leadtime* (or *capacity*) bottleneck. Equation (7) suggests that when the fill rate is high due to abundant inventories, it is most constrained by the station with the smallest α, which we call the *inventory* bottleneck. (Notice that each $\alpha_{i,j}$ is the product of a term $\beta_{i,j}$ depending solely on the operation of the station (i, j) in isolation, and $\pi_{i,j}$, depending solely on the allocation of inventory.) The distinction between these two types of bottlenecks echoes a distinction made by Umble and Srikanth (1990, p. 83) between two types of resource constraints at a facility.

(b) Theorem 1 is related to results in Glasserman (1997) and Glasserman and Wang (1998), but there are important differences. Glasserman (1997) considers periodic-review serial systems with constant capacities at each stage; as a consequence, the stage with the smallest γ is always the stage with the smallest capacity, so the main issue investigated here is absent in that setting. Glasserman and Wang (1998) consider assemble-to-order systems in which each component undergoes just one stage of production. In that simpler setting, Theorem 1 can be strengthened to give sharper exponential approximations to the unfill rate. For example, in a system with just one component and one production stage, Glasserman and Wang (1998) show that

$$\lim_{x+s\to\infty} e^{\gamma x + \beta s} P(R > x) = C. \tag{8}$$

In the substantially more complex setting considered here, combining stochastic production times, multiple components, and multiple stages per component, it does not seem possible to replace the weaker logarithmic limit of Theorem 1 with this type of exponential asymptotic. For related approximations see Buzacott, Price, and Shanthikumar (1992) and Roundy and Muckstadt (1996).

(c) Through $\pi_{i,j}$ in the definition of $\alpha_{i,j}$, Theorem 1 is formulated in terms of the proportion of inventory units held at each station. Since different components at various stages of production may not be directly comparable, it may be more natural to consider, say, the fraction of inventory *dollars* at each station. This requires only minor modification of the result. Let $c_{i,j}$ be the unit inventory cost at station (i, j) and c_0 the unit cost for finished goods inventory, so the total dollar value of inventory is $s = c_0 s_0 + \Sigma_{i,j}\, c_{i,j}\, s_{i,j}$. Redefine $\pi_{i,j}$ to be $c_{i,j}s_{i,j}/s$, $\pi_0 = c_0 s_0/s$, and assume these fractions remain fixed as s increases. The only modification needed for Theorem 1 is to set $\tilde{\pi}_{i,j} = (\pi_0/c_0) + \Sigma_{l=1}^{j} (\pi_{i,l}/c_{i,l})$. The bottleneck with respect to increases in inventory dollars is the station with the smallest $\tilde{\alpha}_{i,j} = \tilde{\pi}_{i,j}\beta_{i,j}$.

(d) Theorem 1 is formulated with the assumption that each finished product requires exactly one unit of each component. At the expense of further complicating the proof, the result could be generalized to allow random batch sizes for components requiring only a single stage of production. More specifically, we could allow the nth order to require D_i^n units of component i which must then have just one production stage. The D_i^n are independent and identically distributed across orders but may be correlated across components for a fixed order. To account for batch demand, we need to modify the definition of γ in (3) to

$$\psi_{D_i}(\psi_{B_i}(\gamma_i)) + \psi_A(-\gamma_i) = 0.$$

We do not treat batch demand explicitly in this paper except in numerical examples. Glasserman and Wang (1998) show that (8) holds with batch demands, and given this result it becomes possible to extend the proof in the appendix for multiple component systems to accommodate this generalization.

(e) Theorem 1 states two limiting results (4) and (5) for the tail probability of the product response time, one on x and one on s. While it is sometimes possible to determine which regime applies, in other cases, the dominating effect of the stations with the smallest γ or α may not be evident, since x and s are always finite in practice. In that case, we need to take both leadtime and inventory into account and compare $\gamma x + \alpha s$. Stations with the smallest $\gamma x + \alpha s$ are the bottlenecks.

2.3. Two-Moment Approximation

Calculating the parameters γ and β requires knowledge of the distributions of A and B, which may not always

be available. If we have only partial knowledge of the distributions—specifically, the means and variances—we can approximate $\gamma > 0$ through a second-order Taylor series approximation to the ψ_A and ψ_B in (3); i.e., we set

$$\psi_B(\theta) + \psi_A(-\theta) \approx \mathsf{E}[B]\theta + \frac{1}{2}\mathsf{Var}[B]\theta^2 - \mathsf{E}[A]\theta + \frac{1}{2}\mathsf{Var}[A]\theta^2 = 0$$

and solve to get

$$\gamma \approx \frac{2(\mathsf{E}[A] - \mathsf{E}[B])}{\mathsf{Var}[A] + \mathsf{Var}[B]}. \tag{9}$$

Similarly we get the two-moment approximation for β,

$$\beta = \psi_B(\gamma) \approx \mathsf{E}[B]\gamma + \frac{1}{2}\mathsf{Var}[B]\gamma^2, \tag{10}$$

which determines, in turn, an approximation for α. The expression in (9) also arises in heavy-traffic approximations ($1/\gamma$ would be the mean response time in a Brownian approximation for a single-station system), but (10) does not appear to have an obvious counterpart or even an interpretation in the heavy-traffic setting.

The expressions in (9) and (10) show that these quantities reflect variability information, though not in an obvious way. In contrast, utilization depends only on means. Without a two-moment approximation, γ and β capture even more distributional information like skewness. But we also see from (9) that γ (and hence β) is small when $\mathsf{E}[A] - \mathsf{E}[B]$ is close to zero, which is to say utilization is close to 1. Thus, fill-rate bottlenecks will often coincide with the throughput bottleneck, though the examples in §§ 1 and 3 show that this is by no means always the case.

3. Numerical Results

In this section, we use additional numerical examples to illustrate the effectiveness of the proposed new notions of *leadtime* and *inventory* bottlenecks, in comparison with the traditional *throughput* bottleneck. In particular, we give examples where all three types of bottlenecks exist at different stations in the same network. We also discuss the related issue of resource allocation when the fill rate is the main concern. All numerical results were obtained by simulation using a special variance reduction technique (related to the one detailed in Glasserman and Liu 1996) to obtain a high degree of precision in fill rate estimates.

In light of Theorem 1, our bottleneck characterizations are guaranteed to apply at sufficiently high fill rates. The main reason for examining numerical examples is to see if this theoretical property is evident at reasonable parameter values. The examples in this section were chosen to illustrate various effects, including changes in the delivery leadtime, changes in capacities, changes in inventory levels, and changes in the order of stations. They were designed specifically to distinguish between utilization bottlenecks and fill-rate bottlenecks. As discussed in §2.3, the various notions of bottlenecks will often apply to a single station; for purposes of illustration we have generally chosen examples in which this is not the case.

3.1. Leadtime Bottlenecks

To isolate the leadtime effect, in this subsection we consider examples without inventory. We use two approaches to illustrate bottleneck phenomena, paralleling the observation in Examples 1 and 2 of §1. The first approach studies the effect of changes in the delivery leadtime x on the fill rate, and the second approach examines the effect of capacity increases. For the first approach, let R_i be the response time for station i operating in isolation. As x increases, we compare the rate of decrease of the system unfill rate $p \triangleq P(R > x)$ and the station-level unfill rate $p_i \triangleq P(R_i > x)$. We will show that for a leadtime bottleneck station i, p_i decays at the same rate as p while for other stations p_i decays much faster than p. In the second approach we will change γ values through changes in mean production times. We will show that a marginal increase of capacity at the leadtime bottleneck has the greatest effect on the system fill rate.

Consider, again, the three-station serial system in Figure 1. The interarrival time between orders has an Erlang distribution with mean 1 and standard deviation $1/\sqrt{3}$. The three stations have the following mean production intervals: $\mathsf{E}[B_1] = 0.80$, $\mathsf{E}[B_2] = 0.90$ and

$\mathsf{E}[B_3] = 0.95$. Obviously, station 3 is the throughput bottleneck. The production intervals have the following distribution: B_1 is hyperexponential with coefficient of variation $c_{B_1} = 2$; B_2 is Erlang with $c_{B_2} = 1/\sqrt{5}$; B_3 is deterministic. Through equation (3) we can easily calculate $\gamma_1 = 0.10$, $\gamma_2 = 0.42$ and $\gamma_3 = 0.32$. These values predict that station 1, though it has the lowest utilization level, is the leadtime bottleneck. Figure 5 plots the ratio p_i/p, $i = 1, 2, 3$. As x increases, both p_2/p and p_3/p decrease exponentially fast while p_1/p remains nearly constant. It is therefore indeed station 1 that most constrains the system unfill rate p.

For a second perspective (based on changing γ values at each station) we present two classes of examples. The first class refers to the same system as above, and the results are in Table 5. In each row of the table, we first list the current fill rate at the corresponding delivery lead time x. We then calculate the new fill rate after a decrease of 0.05 in each of the $\mathsf{E}[B_i]$ (hence a decrease of 5 percentage points in each ρ_i) with the coefficients of variation held fixed. As shown there, a decrease in utilization at station 1, which has the lowest utilization yet smallest γ, improves service much more than the same decrease at stations 2 or 3.

The second class refers again to a version of the system depicted in Figure 1 for which the results are listed in Tables 6–7. In Table 6, the interarrival time has an Erlang distribution with mean $\mathsf{E}[A] = 1$ and standard deviation $1/\sqrt{3}$; the unit production times have the following distributions, hyperexponential for B_1, deterministic for B_2, and Erlang for B_3 with $c_{B_3} = 1/\sqrt{5}$; each order requires the combination of one unit from station

Figure 5 **Comparison Between System-Wide and Station-Level Unfill Rates of the System in Figure 1. The Vertical Axis Is the Logarithm of the Ratio of the Station-Level Unfill Rate to the System-Wide Unfill Rate.**

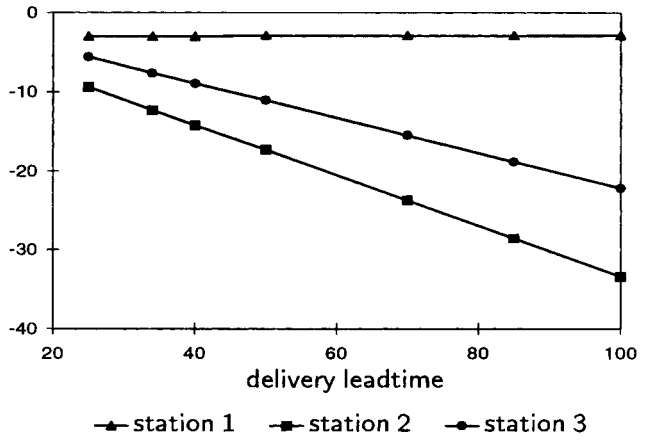


Table 5 **Comparison of Fill Rate Improvement After a 5 Percentage-Point Decrease in Utilization at Each Station in the System of Figure 1. Initial Utilization Levels at the Three Stations are $\rho_1 = 80\%$, $\rho_2 = 90\%$ and $\rho_3 = 95\%$; the Measures of Propensity to Constrain the Fill Rate are $\gamma_1 = 0.10$, $\gamma_2 = 0.42$, and $\gamma_3 = 0.32$**

		fill rate (in %) after a utilization decrease at each station		
delivery leadtime	initial fill rate	station 1	station 2	station 3
25	89.2	93.6	89.4	90.2
34	95.4	98.0	95.7	96.0
40	97.5	99.1	97.6	97.8

Table 6 **Comparison of Fill Rate Improvement After a 5 Percentage-Point Decrease in Utilization at Each Station in a System Depicted in Figure 1. For the Upper Panel, Initial Utilization Levels at the Three Stations are $\rho_1 = 60\%$, $\rho_2 = 90\%$ and $\rho_3 = 80\%$; the Measures of Propensity to Constrain the Fill Rate are $\gamma_2 = 0.69$, $\gamma_3 = 0.67$, and $\gamma_1 = 0.23$ when $c_{B_1} = 2$, $\gamma_1 = 0.10$ when $c_{B_1} = 3$. For the Lower Panel, Initial Utilization Levels at the Three Stations are $\rho_1 = 80\%$, $\rho_2 = 95\%$ and $\rho_3 = 90\%$; the Measures of Propensity to Constrain the Fill Rate are $\gamma_2 = 0.32$, $\gamma_3 = 0.42$, and $\gamma_1 = 0.10$ when $c_{B_1} = 2$, $\gamma_1 = 0.05$ when $c_{B_1} = 3$**

			fill rate (in %) after a utilization decrease at each station		
c_{B_1}	delivery leadtime	initial fill rate	station 1	station 2	station 3
2	15	91.3	93.9	92.5	91.4
	20	97.3	98.5	97.6	97.3
3	18	91.0	94.0	91.6	91.0
	40	99.1	99.6	99.1	99.1
2	25	91.2	95.4	92.5	91.2
	40	97.9	99.3	98.3	97.9
3	50	91.5	95.6	92.0	91.5
	80	97.7	99.2	97.9	97.7

1 and geometrically many from station 3 with mean 5. In each row of the tables, we first list the current fill rate at the corresponding c_{B_1} and x. We then calculate the new fill rate after a decrease of 0.05 in each of the $\mathsf{E}[B_i]$ (hence a decrease of 5 percentage points in each ρ_i) with the coefficients of variation held fixed. As shown there, a decrease in utilization at station 1, which has the lowest utilization yet smallest γ, improves service much more than the same decrease at stations 2 or 3. This is true for both panels in the table; the overall system utilization is higher in the lower panel.

To investigate the role of a station's location in determining whether it is a fill-rate bottleneck, we exchange the distributions and parameters of B_1 and B_2 from our last example. Table 7 shows that the location plays little role since a utilization decrease in station 2 now most improves the service level, as predicted by its γ value.

In all the examples of Tables 5–7, a marginal increase of production capacity at the station with the smallest γ value always improves the fill rate significantly, while the same amount of capacity increase at any other station yields far less, if any, improvement in the fill rate. This is true regardless of the structure of the system, or the overall utilization level, or the relative location of the stations. Interestingly, the leadtime bottlenecks in these examples all have the lowest utilization level in the system.

3.2. Inventory Bottlenecks

In this subsection, we investigate which facility constrains the fill rate when the inventory levels are high. We study the system depicted in Figure 1. Let the interarrivel time A have Erlang distribution with mean $\mathsf{E}[A] = 1$ and $c_A = 1/\sqrt{3}$; let B_1 have hyperexponential distribution with $\mathsf{E}[B_1] = 0.8$ and $c_{B_1} = 2$; let $B_2 = 0.95$ and $B_3 = 0.1$ be constants. A product demand requires one unit of component 1 and D_3 units of component 3, where D_3 has geometric distribution with mean $\mathsf{E}[D_3] = 8.33$. For the initial allocation, we assign the same total inventory to station 3 as to the subsystem consisting of stations 1 and 2, and more inventory at 1 than at 2.

With the initial allocation for stations 1, 2, and 3 set at 70, 5, and 75, respectively, the fill rate is 90.6%. Suppose we can keep 20 additional units of inventory at any one facility. Where should they be added to achieve the greatest improvement in the fill rate? Table 8 compares fill rates under different allocations. As shown there, the fill rate is almost unchanged if the increase is made at station 1 or 2, or shared between them. But the fill rate jumps to 95.0% if the additional inventory is kept at station 3. In this sense, station 3 is a fill-rate bottleneck with respect to increases in inventory levels.

As a comparison, Table 9 shows the impact on the fill rate of the marginal increase in production capacity at each station when stations carry no inventory. As shown there, the service improvement is much greater

Table 7 **Comparison of Fill Rate Improvement After a 5 Percentage-Point Decrease in Utilization at Each Station in the Same Systems as in Table 6, Except that the Distributions of B_1 and B_2 are Exchanged**

			fill rate (in %) after a utilization decrease at each station		
c_{B_2}	delivery leadtime	initial fill rate	station 1	station 2	station 3
2	15	89.8	93.3	94.0	89.8
	20	96.5	97.8	98.4	96.5
3	30	89.6	92.8	93.9	89.6
	50	98.6	99.1	99.5	98.6
2	37	89.5	94.3	95.3	90.4
	50	96.9	98.4	99.6	96.9
3	100	96.3	97.8	98.9	96.3

Table 8 **Comparison of fill rate improvement after inventory level increase at each station. Utilization levels at the three stations are 80%, 95% and 83%, respectively; the α values are $\alpha_1 = 0.05$, $\alpha_2 = 0.14$, and $\alpha_3 = 0.015$.**

	station inventories			
	1	2	3	fill rate (in %)
initial inventory	70	5	75	90.6
add at 1	90	5	75	90.7
add at 2	70	25	75	90.7
add at 1 and 2	80	15	75	90.7
add at 3	70	5	95	95.0

Table 9 **Comparison of fill rate improvement after a 5 percentage-point decrease in utilization at each station. Initial utilization levels at the three stations are $\rho_1 = 80\%$, $\rho_2 = 95\%$ and $\rho_3 = 83\%$; the γ values are $\gamma_1 = 0.10$, $\gamma_2 = 0.32$, and $\gamma_3 = 0.32$.**

		fill rate (in %) after a utilization decrease at each station		
delivery leadtime	initial fill rate	station 1	station 2	station 3
25	90.6	95.1	92.4	90.8
40	97.9	99.3	98.3	97.9

if more capacity is added to station 1, which has the smallest γ, than to station 2 (which has the highest utilization) or station 3 (which is the inventory bottleneck). Again, we see three different types of bottleneck.

4. Conclusions

We have introduced two measures of a station's propensity to constrain the fill rate in a production-inventory network. One measure captures a restraining effect as either the production capacity or the delivery leadtime promised to customers increases; the other captures a similar effect as inventories increase. The smaller the measure, the greater the restraining effect, so a station with the smallest value of either measure is a fill-rate bottleneck—the station that most limits improvements in the fill rate. We have shown both analytically and through numerical examples that the notion of a fill-rate bottleneck can be useful in identifying stations where additional resources would have the greatest impact. Unlike the traditional throughput-oriented, utilization-based bottleneck measure, the new service-oriented measures proposed here are sensitive to process variability.[1]

[1]The authors thank the referees and editors for helpful comments and suggestions. Comments from Michael Harrison helped improve the exposition. This research is supported by NSF grant DMI-9457189.

Appendix: Analysis and Proofs

The main purpose of this appendix is to prove Theorem 1. To that end, we first characterize the steady state response time for a system with only one component but multiple stages, state and prove the results for that system in Proposition 1; then we generalize the results to the multiple component case.

Consider a single component serial system of d stages with unlimited raw material supplying stage d and demands arriving at stage 1. Let $\{t^n; n \geq 1\}$ be the demand arrival epochs, let $\{T_j^n, n \geq 1\}$ be the completion time of the nth production at stage j, with $T_j^0 = 0$. Since the system operates under a base-stock policy with local base-stock level s_j and echelon base-stock level $\bar{s}_j = \Sigma_{k=1}^{j} s_k$, we have for stage d, $T_d^n = T_d^{n-1} \vee t^n + B_d^n$, and for stage j,

$$T_j^n = \begin{cases} T_j^{n-1} \vee t^n + B_j^n, & \text{for } 1 \leq n \leq s_{j+1}, \\ T_j^{n-1} \vee T_{j+1}^{n-s_j+1} \vee t^n + B_j^n, & \text{for } n > s_{j+1}, \end{cases} \tag{11}$$

$j = 1, 2, \ldots, d-1$. (We use $x \vee y$ to denote $\max(x, y)$.) Define

$$Y_j^n = \begin{cases} 0, & \text{for } n \leq s_j, \\ T_j^{n-s_j} - t^n, & \text{for } n > s_j. \end{cases} \tag{12}$$

The response time of the nth order $R^n = (Y_1^n)^+ \triangleq \max(Y_1^n, 0)$.

A key step in analyzing the response time and proving our results is a characterization of the steady state response time, R, in this serial system. Due to the relation between R^n and Y_1^n, it suffices to find the steady-state representation Y_1 of $\{Y_1^n, n \geq 1\}$. To that end, we start with Y_d at stage d and work recursively backward to Y_1. For the first production stage (stage d), we have the following easily verified sample-path result:

$$\begin{aligned} Y_d^n &= T_d^{n-s_d} - t^n \\ &= T_d^{n-s_d} - t^{n-s_d} - (t^n - t^{n-s_d}) \\ &= B_d^{n-s_d} - \sum_{k=n-s_d}^{n-1} A^k + \max_{1 \leq k \leq n-s_d} \sum_{l=k}^{n-s_d-1} (B_d^l - A^l). \end{aligned} \tag{13}$$

For $j < d$ and $n \geq \bar{s}_d$, the release rule (11) and the definition (12) yield

$$\begin{aligned} &Y_j^n + (t^n - t^{n+1-s_j}) \\ &= T_j^{n-s_j} - t^n + t^n - t^{n+1-s_j} \\ &= T_j^{n-1-s_j} \vee T_{j+1}^{n-s_j-s_{j+1}} \vee t^{n-s_j} + B_j^{n-s_j} - t^{n+1-s_j} \\ &= (T_j^{n-1-s_j} - t^{n-1} + t^{n-1} - t^{n-s_j}) \vee (T_{j+1}^{n-s_j-s_{j+1}} - t^{n-s_j}) \vee 0 \\ &\quad + B_j^{n-s_j} - t^{n+1-s_j} + t^{n-s_j} \\ &= (Y_j^{n-1} + (t^{n-1} - t^{n-s_j})) \vee (Y_{j+1}^{n-s_j}) \vee 0 + B_j^{n-s_j} - A^{n-s_j}. \end{aligned}$$

From this recursion, we can express $\{Y_j^n\}$ in terms of $\{Y_{j+1}^n\}$,

$$\begin{aligned} Y_j^n = B_j^{n-s_j} &- \sum_{k=n-s_j}^{n-1} A^k \\ &+ \max_{1 \leq k \leq n-s_j} \sum_{l=k}^{n-s_j-1} (B_j^l - A^l) \\ &\vee \max_{1 \leq k \leq n-s_j} \left(\sum_{l=k}^{n-s_j-1} (B_j^l - A^l) + Y_{j+1}^k \right), \end{aligned} \tag{14}$$

with the convention $\Sigma_{l=k}^{k-1} (B_j^l - A^l) = 0$.

Starting with Equation (13) for stage d, we repeatedly use (14) and obtain

$$
\begin{aligned}
Y_1^n &= B_1^{n-s_1} - \sum_{n-s_1}^{n-1} A^l + \max_{1\le k_1\le n-s_1} \sum_{l=k_1}^{n-s_1-1} (B_1^l - A^l) \\
&\vee \max_{1\le k_1\le n-s_1} \Bigg(\sum_{l=k_1}^{n-s_1-1} (B_1^l - A^l) + B_2^{k_2-s_2} \\
&\quad - \sum_{l=k_1-s_2}^{k_1-1} A^l + \max_{1\le k_2\le k_1-s_2} \sum_{l=k_2}^{k_1-s_2-1} (B_2^l - A^l) \\
&\quad \vee \max_{1\le k_2\le k_1-s_2} \Bigg(\sum_{l=k_2}^{k_1-s_2-1} (B_2^l - A^l) + \cdots + B_d^{k_{d-1}-s_d} \\
&\quad - \sum_{l=k_{d-1}-s_d}^{k_{d-1}-1} A^l + \max_{1\le k_d\le k_{d-1}-s_d} \sum_{l=k_d}^{k_{d-1}-s_d-1} (B_d^l - A^l) \cdots \Bigg) \Bigg) \\
&\triangleq f_n(A^{n-1}, \ldots, A^1; B_1^{n-1}, \ldots, B_1^1; \ldots; B_d^{n-1}, \ldots, B_d^1). \qquad (15)
\end{aligned}
$$

Due to our independence assumptions, we can reverse the order of the sequences $\{A^1, \ldots, A^{n-1}\}$ and $\{B_j^1, \ldots, B_j^{n-1}\}$ in function f_n of (15) without changing the distribution of Y_1^n, i.e.,

$$
\begin{aligned}
Y_1^n &\overset{\mathcal{D}}{=} f_n(A^1, \ldots, A^{n-1}; B_1^1, \ldots, B_1^{n-1}; \ldots; B_d^1, \ldots, B_d^{n-1}) \\
&= B_1^{s_1} - \sum_{l=1}^{s_1} A^l + \max_{s_1\le k_1\le n-1} \sum_{l=s_1+1}^{k_1} (B_1^l - A^l) \\
&\vee \max_{s_1\le k_1\le n-1} \Bigg(\sum_{l=s_1+1}^{k_1} (B_1^l - A^l) + B_2^{k_1+s_2} - \sum_{l=k_1+1}^{k_1+s_2} A^l \\
&\quad + \max_{k_1+s_2\le k_2\le n-1} \sum_{l=k_1+s_2+1}^{k_2} (B_2^l - A^l) \\
&\quad \vee \max_{k_1+s_2\le k_2\le n-1} \Bigg(\sum_{l=k_1+s_2+1}^{k_2} (B_2^l - A^l) + \cdots + \\
&\quad B_d^{k_{d-1}+s_d} - \sum_{l=k_{d-1}+1}^{k_{d-1}+s_d} A^l + \max_{k_{d-1}+s_d\le k_d\le n-1} \\
&\quad \sum_{l=k_{d-1}+s_d+1}^{k_d} (B_d^l - A^l) \cdots \Bigg) \Bigg) \equiv \tilde{Y}_1^n, \qquad (16)
\end{aligned}
$$

where $\overset{\mathcal{D}}{=}$ denotes equal in distribution. Figure 6 gives a graphical representation of $\tilde{Y}_1^n$ as the maximum weight of all the possible paths in the graph, with the weights on the arcs as indicated. A path starts from the lower-left corner, follows the direction of the arcs and may stop anywhere. Note that the maximum length of a path is $n-1$. The values of the k_j in Equation (16) determine on which row the path either stops in column j, or switches from column j to column $j+1$. Note that $\{\tilde{Y}_1^n\}$ is an increasing function of n so it has a limiting distribution. Denote $X_j^l = B_j^l - A^l$ and let $n \to \infty$, we have

$$
\begin{aligned}
\tilde{Y}_1^n &\Rightarrow f_\infty(A^1, A^2, \ldots, B_1^1, B_1^2, \ldots; \ldots; B_d^1, B_d^2, \ldots) \\
&= B_1^{s_1} - \sum_{l=1}^{s_1} A^l + \max_{k_1\ge s_1} \sum_{l=s_1+1}^{k_1} X_1^l \\
&\vee \max_{k_1\ge s_1} \Bigg(\sum_{l=s_1+1}^{k_1} X_1^l + B_2^{k_1+s_2} \\
&\quad - \sum_{l=k_1+1}^{k_1+s_2} A^l + \max_{k_2\ge k_1+s_2} \sum_{l=k_1+s_2+1}^{k_2} X_2^l \\
&\quad \vee \max_{k_2\ge k_1+s_2} \Bigg(\sum_{l=k_1+s_2+1}^{k_2} X_2^l + \cdots + B_d^{k_{d-1}+s_d} \\
&\quad - \sum_{l=1}^{s_d} A^{k_{d-1}+l} + \max_{k_d\ge k_{d-1}+s_d} \sum_{l=k_{d-1}+s_d+1}^{k_d} X_d^l \cdots \Bigg) \Bigg) \equiv Y_1, \qquad (17)
\end{aligned}
$$

where $\Rightarrow$ denotes convergence in distribution. Y_1 has a similar graphical representation as in Figure 6, but the paths may have infinite length. Since $Y_1^n \overset{\mathcal{D}}{=} \tilde{Y}_1^n$, we have $Y_1^n \Rightarrow Y_1$ and the steady state response time $R = (Y_1)^+$.

Proposition 1. *Suppose that in a d-stage serial system $\gamma_1, \gamma_2, \ldots, \gamma_d$ all exist. Then*

$$
\lim_{x\to\infty} -\frac{1}{x} \log P(Y_1 > x) = \min_{1\le j\le d} \gamma_j; \qquad (18)
$$

and if in addition all $\alpha_j > 0$ then

$$
\lim_{s\to\infty} -\frac{1}{s} \log P(Y_1 > x) = \min_{1\le j\le d} \alpha_j. \qquad (19)
$$

Proof. For single stage systems, Glasserman and Wang (1998) showed that

$$
\begin{aligned}
&\lim_{x+s\to\infty} e^{\gamma x+\beta s} P(Y > x) \\
&\quad = \lim_{x+s\to\infty} e^{\gamma x+\beta s} P\Big(B^s - \sum_{l=1}^{s} A^l + \max_{k\ge s} \sum_{l=s+1}^{k} X^l > x\Big) \\
&\quad = C
\end{aligned}
$$

for some constant $C > 0$. Applying this result to station d in isolation gives

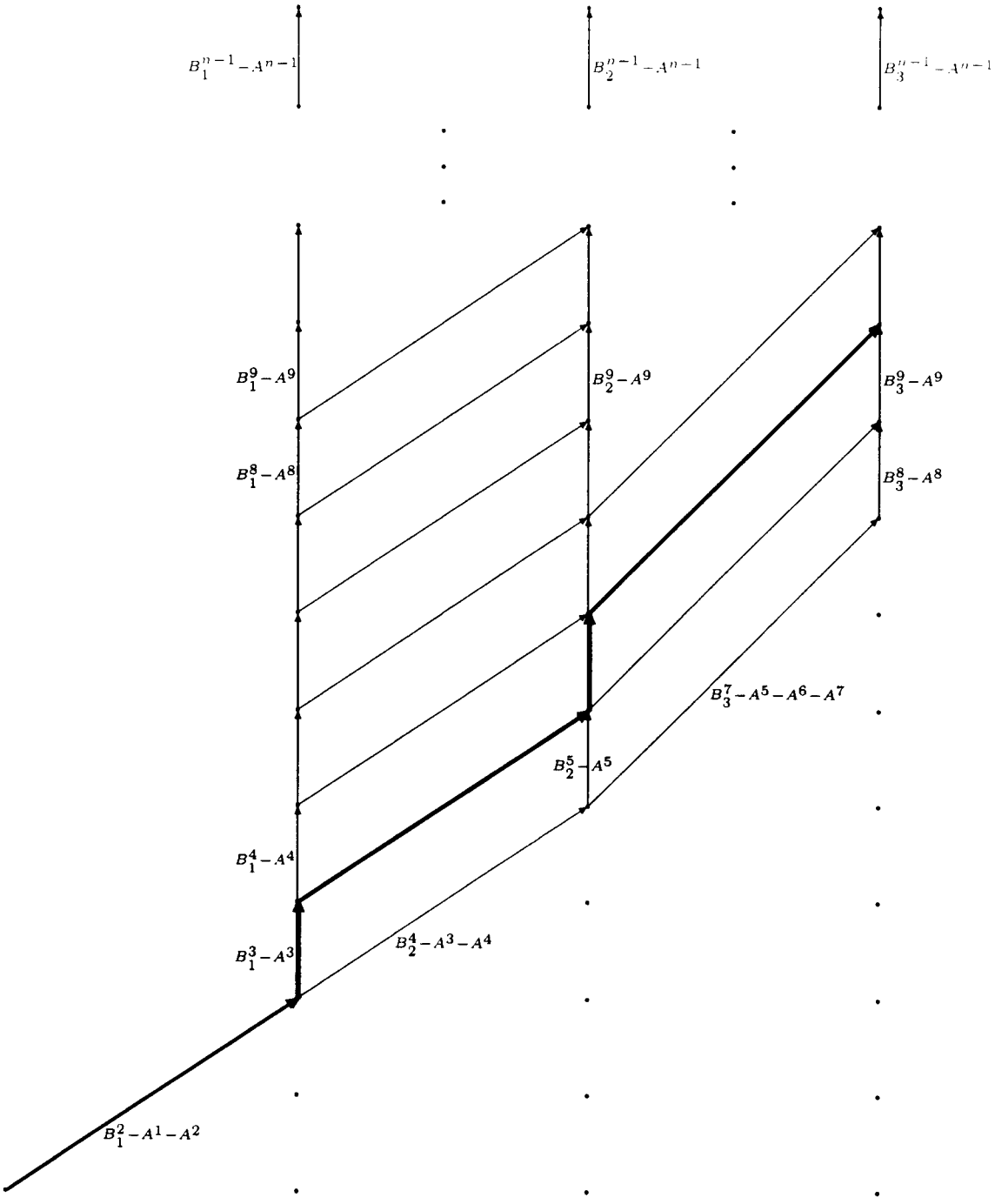
$$
\begin{aligned}
&\lim_{x\to\infty} -\frac{1}{x} \log P(Y_d > x) \\
&\quad = \lim_{x\to\infty} -\frac{1}{x} \log P\Big(B_d^{s_d} - \sum_{l=1}^{s_d} A^l + \max_{k_d\ge s_d} \sum_{l=s_d+1}^{k_d} X_d^l > x\Big) \\
&\quad = \gamma_d, \qquad (20)
\end{aligned}
$$

and

$$
\begin{aligned}
&\lim_{s_d\to\infty} -\frac{1}{s_d} \log P(Y_d > x) \\
&\quad = \lim_{s_d\to\infty} -\frac{1}{s_d} \log P\Big(B_d^{s_d} - \sum_{l=1}^{s_d} A^l + \max_{k_d\ge s_d} \sum_{l=s_d+1}^{k_d} X_d^l > x\Big) \\
&\quad = \beta_d. \qquad (21)
\end{aligned}
$$

For any $1 \le j \le d$, we have the following lower bound on Y_1 (by

Figure 6 **Graphical Representation of $\tilde{Y}^n$ in a Three-Stage Example with $s_1 = 2$, $s_2 = 2$, and $s_3 = 3$. Each Stage Corresponds to a Column, Each Row to a Time Index. The Weight of a Vertical Arc at Column j, Row i is $B_j^{n-i} - A^{n-i}$; The Weight of a Diagonal Arc from Column j, Row i to Column $j+1$ is $B_{j+1}^{n-i+s_{j+1}-1} - A^{n-i} - \ldots - A^{n-i+s_{j+1}-1}$. $\tilde{Y}_1^n$ Takes the Maximum Weight of All the Possible Paths in the Graph. Thick Lines Illustrate One Possible Path.**



selecting a special class of paths which take diagonal arcs from the start all the way up to column j and then stay on column j)

$$Y_1 \geq B_1^{\bar{s}_1} - \sum_{l=1}^{\bar{s}_1} A^l + B_2^{\bar{s}_2} - \sum_{l=\bar{s}_1+1}^{\bar{s}_2} A^l + \cdots + B_j^{\bar{s}_j} - \sum_{l=\bar{s}_{j-1}+1}^{\bar{s}_j} A^l + \max_{k \geq \bar{s}_j} \sum_{l=\bar{s}_j+1}^{k} X_j^l,$$

which immediately yields

$$P(Y_1 > x) \geq P\left(B_1^{\bar{s}_1} + \cdots + B_j^{\bar{s}_j} - \sum_{l=1}^{\bar{s}_j} A^l + \max_{k \geq \bar{s}_j} \sum_{l=\bar{s}_j+1}^{k} X_j^l > x\right)$$

$$> P\left(B_j^{\bar{s}_j} - \sum_{l=1}^{\bar{s}_j} A^l + \max_{k \geq \bar{s}_j} \sum_{l=\bar{s}_j+1}^{k} X_j^l > x\right).$$

By the second inequality of (20) we have

$$\limsup_{x \to \infty} -\frac{1}{x} \log P(Y_1 > x) \leq \gamma_j, \qquad \forall 1 \leq j \leq d,$$

$$\limsup_{x \to \infty} -\frac{1}{x} \log P(Y_1 > x) \leq \min_{1 \leq j \leq d} \gamma_j;$$

and by the second inequality of (21) we have

$$\limsup_{\bar{s}_j \to \infty} -\frac{1}{\bar{s}_j} \log P(Y_1 > x) \leq \beta_j, \qquad \forall 1 \leq j \leq d$$

i.e.,

$$\limsup_{s \to \infty} -\frac{1}{s} \log P(Y_1 > x) \leq \pi_j \beta_j \equiv \alpha_j, \qquad \forall 1 \leq j \leq d,$$

$$\limsup_{s \to \infty} -\frac{1}{s} \log P(Y_1 > x) \leq \min_{1 \leq j \leq d} \alpha_j.$$

So the upper bounds of (18) and (19) have been established.

For the lower bounds, through a crude bound on $P(Y_1 > x)$ we get

$$P(Y_1 > x) \leq \sum_{k_1 \geq s_1} P\left(B_1^{s_1} - \sum_{l=1}^{s_1} A^l + \sum_{l=s_1+1}^{k_1} X_1^l > x\right)$$

$$+ \sum_{k_1 \geq s_1} \sum_{k_2 \geq k_1 + s_2} P\left(B_1^{s_1} - \sum_{l=1}^{s_1} A^l + \sum_{l=s_1+1}^{k_1} X_1^l + B_2^{k_1+s_2} - \sum_{l=k_1+1}^{k_1+s_2} A^l + \sum_{l=k_1+s_2+1}^{k_2} X_2^l > x\right)$$

$$+ \cdots + \sum_{k_1 \geq s_1} \cdots \sum_{k_d \geq k_{d-1} + s_d} P\left(B_1^{s_1} - \sum_{l=1}^{s_1} A^l + \sum_{l=s_1+1}^{k_1} X_1^l + \cdots + B_d^{k_{d-1}+s_d} - \sum_{l=k_{d-1}+1}^{k_{d-1}+s_d} A^l + \sum_{l=k_{d-1}+s_d+1}^{k_d} X_d^l > x\right) \triangleq \Delta_1 + \Delta_2 + \cdots + \Delta_d,$$

where each Δ_j is the sum over $k_1, \ldots, k_j$ of the probability that the path determined by $k_1, \ldots, k_j$ and ending on column j crosses x (cf. Figure 6). It suffices to show that each of the Δ_j satisfies

$$\liminf_{x \to \infty} -\frac{1}{x} \log \Delta_j \geq \min_{1 \leq k \leq d} \gamma_k \quad \text{and}$$

$$\liminf_{s \to \infty} -\frac{1}{s} \log \Delta_j \geq \min_{1 \leq k \leq d} \alpha_k.$$

By Chebychev's Inequality, for $\theta < \min_{1 \leq k \leq j} \gamma_k$

$$\Delta_j \leq \sum_{k_1 \geq s_1} \cdots \sum_{k_j \geq k_{j-1} + s_j} e^{-\theta x} \cdot \mathrm{E}[e^{\theta(B_1 + \ldots + B_j)}] \cdot (\mathrm{E}[e^{-\theta A}])^{\bar{s}_j}$$

$$\cdot e^{(k_1 - s_1)\psi X_1(\theta) + \ldots + (k_j - k_{j-1} - s_j)\psi X_j(\theta)}$$

$$= e^{-\theta x + \bar{s}_j \psi_A(-\theta)} \left(e^{\psi_{B_1}(\theta) + \ldots + \psi_{B_j}(\theta)} \sum_{l_1 \geq 0} (e^{\psi_{X_1}(\theta)})^{l_1} \cdots \sum_{l_j \geq 0} (e^{\psi_{X_j}(\theta)})^{l_j}\right)$$

$$\triangleq e^{-\theta x + \bar{s}_j \psi_A(-\theta)} \cdot M(\theta).$$

where $M(\theta) = (e^{\psi_{B_1}(\theta) + \ldots + \psi_{B_1}(\theta)} \Sigma_{l_1 \geq 0} (e^{\psi_{X_1}(\theta)})^{l_1} \cdots \Sigma_{l_j \geq 0} (e^{\psi_{X_j}(\theta)})^{l_j})$ is finite for all $\theta < \min_{1 \leq k \leq j} \gamma_k$ since $e^{\psi_{X_k}(\theta)} < 1$ for $1 \leq k \leq j$. So

$$\liminf_{x \to \infty} -\frac{1}{x} \log \Delta_j \geq \theta \quad \text{and} \quad \liminf_{\bar{s}_j \to \infty} -\frac{1}{\bar{s}_j} \log \Delta_j \geq -\psi_A(-\theta),$$

for all $\theta < \min_{1 \leq k \leq j} \gamma_k$. Take the supremum over θ, we have

$$\liminf_{x \to \infty} -\frac{1}{x} \log \Delta_j \geq \min_{1 \leq k \leq j} \gamma_k \geq \min_{1 \leq k \leq d} \gamma_k,$$

and

$$\liminf_{s \to \infty} -\frac{1}{s} \log \Delta_j \geq -\pi_j \psi_A\left(-\min_{1 \leq k \leq j} \gamma_k\right)$$

$$= \pi_j \left(\min_{1 \leq k \leq j} \beta_k\right)$$

$$\geq \min_{1 \leq k \leq j} \alpha_k \quad (\text{since } \pi_1 \leq \pi_2 \leq \cdots \leq \pi_j)$$

$$\geq \min_{1 \leq k \leq d} \alpha_k. \quad \square$$

It follows directly from Proposition 1 and $P(R > x) = P((Y_1)^+ > x) = P(Y_1 > x)$ that the response time in a serial system satisfies

$$\lim_{x \to \infty} -\frac{1}{x} \log P(R > x) = \min_{1 \leq k \leq d} \gamma_k, \tag{22}$$

$$\lim_{s \to \infty} -\frac{1}{s} \log P(R > x) = \min_{1 \leq k \leq d} \alpha_k. \tag{23}$$

We now give

Proof of Theorem 1. Let T^n be the completion time of the nth assembly operation; let T_i^n be the finishing time of the nth production of component i. Then

$$T^n = \max_i T_i^{n - s_i, 1} + U^n.$$

The response time of nth order of the product is

$$R^n = (T^{n-s_0} - t^n)^+$$
$$= \left(\max_i T_i^{n-s_{i,1}-s_0} - t^n + U^{n-s_0}\right)^+$$
$$= \left(\max_i Y_i^n + U^{n-s_0}\right)^+$$
$$\Rightarrow \left(\max_i Y_i + U\right)^+ = R,$$

where $(Y_i^n)^+$ is the nth response time of a serial system composed of the stages for component i, but with station $(i, 1)$ carrying $s_{i,1} + s_0$ base stock inventory (cf. definition (12)), and $(Y_i)^+ \equiv R_i$ its steady state version. Equations (22) and (23) specialize to

$$\lim_{x\to\infty} -\frac{1}{x}\log P(R_i > x) = \min_j \gamma_{i,j},$$

and

$$\lim_{s\to\infty} -\frac{1}{s}\log P(R_i > x) = \min_j \alpha_{i,j}.$$

For x with $P(U < x) = 1$, we have

$$P(R > x) = P\left(\max_i Y_i > x - U\right) = P\left(\max_i R_i > x - U\right),$$

which immediately yields the simple bounds

$$\max_i P(R_i > x - U) \le P(R > x) \le \sum_i P(R_i > x - U).$$

From the lower bound, we have

$$\limsup_{x\to\infty} -\frac{1}{x}\log P(R > x)$$
$$\le \min_i \left(\lim_{x\to\infty} -\frac{1}{x}\log P(R_i > x - U)\right) = \min_{i,j} \gamma_{i,j};$$

from the upper bound, we have

$$\liminf_{x\to\infty} -\frac{1}{x}\log P(R > x)$$
$$\ge \min_i \left(\lim_{x\to\infty} -\frac{1}{x}\log P(R_i > x - U)\right) = \min_{i,j} \gamma_{i,j};$$

By combining the two limits we get (4); (5) follows from a similar argument.

References

Asmussen, S. 1987. *Applied Probability and Queues*. Wiley, New York.

Buzacott, J. A., S. M. Price, J. G. Shanthikumar. 1992. Service level in multistage MRP and base stock controlled productions systems. G. Fendel, T. Gulledge, and A. James, eds. *New Directions for Operations Research in Manufacturing*. Springer-Verlag, New York 445–463.

Chen, H., A. Mandelbaum. 1991. Discrete flow networks: Bottleneck analysis and fluid approximation. *Math. Oper. Res.* **16** 408–446.

Glasserman, P. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* **45** 244–257.

——— T. W. Liu. 1996. Rare-event simulation for multistage production-inventory systems. *Management Sci.* **42** 1292–1307.

——— Y. Wang. 1998. Leadtime-inventory trade-offs in assemble-to-order systems. *Oper. Res.* **46** 858–871.

Goldratt, E. 1990. *Theory of Constraints*. North River Press, Inc. Croton-on-Hudson, New York.

——— J. Cox. 1985. *The Goal*. North River Press, Inc., Croton-on-Hudson, New York.

Harrison, J. M., L. Wein. 1990. Scheduling network of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.* **38** 1052–1064.

Hopp, W. J., M. L. Spearman. 1996. *Factory Physics*. Irwin, Chicago, IL.

Roundy, R., J. A. Muckstadt. 1996. Heuristic computation of periodic-review base stock inventory policies. Technical Report, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.

Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York.

Umble, M., M. L. Srikanth. 1990. *Synchronous Manufacturing*. Southwestern Publishing Co., Cincinnati, OH.

Accepted by Michael Harrison; received January 21, 1997. This paper has been with the authors for 16 months for 2 revisions.