Division of Research
School of Business Administration
The University of Michigan

March 1992

PRODUCTION RATES FOR UNPACED PRODUCTION LINES WITH SERIAL WORK STATIONS AND PARALLEL SERVICE FACILITIES

Working Paper No. 678

Michael J. Magazine
University of Waterloo
and
Kathryn E. Stecke
The University of Michigan

PRODUCTION RATES FOR UNPACED PRODUCTION LINES WITH SERIAL WORK STATIONS AND PARALLEL SERVICE FACILITIES

Michael J. Magazine
University of Waterloo
Department of Management Sciences
Waterloo, Ontario, Canada

and

Kathryn E. Stecke
The University of Michigan
Graduate School of Business Administration
Ann Arbor, Michigan

March, 1992

ABSTRACT

Unpaced serial production lines with parallel service facilities are examined in order to determine how their output rates may be improved through the manipulation of the various design variables, such as the allocation of facilities to stations, allocation of workload to stations, and placement of buffers between stations. The knowledge gained through many empirical investigations is utilized to make observations and to formulate conjectures about the output maximizing configurations for production lines which have multiple service facilities and finite buffers at each of several work stations in series. Comparisons to recent related results are provided.

1. INTRODUCTION

This paper considers the problem of improving the output rates from unpaced production lines having a fixed process flow and finite buffers by manipulation of the following parameters:

i. the number of work stations;

ii. the number of parallel facilities at each work station;

iii. the amount of buffer storage between work stations; and

iv. the distribution of workload among the stations.

These parameters arise as design variables when the production line is not mechanically paced (e.g., by a conveyor belt) and when the sequence of a customer's or a part's service is fixed (e.g., by precedence rules). Customers or orders arrive at the service system to be processed by any one of the available service facilities at the first work station. They move through the succeeding work stations as their current servicing is completed and one of the next service facilities becomes free. The first station is assumed to have an infinite queue and a supply of customers preceding it, while the last station has an ample number of storage locations succeeding it.

While on its journey through the production line, the customer or part can experience one of two states at any arbitrary time: being serviced or waiting for service. Service facilities, on the other hand, can experience one of three states: busy servicing a customer, waiting for a customer to arrive (starved), or waiting to pass a customer on to the next work station (blocked). Other than the first work station that can never be starved and the last that can never be blocked, all work stations' service facilities experience these three states.

One problem addressed here is to determine the amount of *available service capacity*, or *workload*, that should be allocated to each *station* to maximize output rate. Another problem is to determine the best configuration of facilities among stations to maximize output. Allocating both workloads and facilities to stations is also addressed.

For the purposes of this study, all service mechanisms are assumed to have exponentially distributed service times, which are independent from facility to facility. The observation that the exponential distribution is sufficient in studies such as these was made by Hillier and So [1989],

who examined coefficients of variation ranging from .7 to 2.5. This variability is important to consider in modern production systems to accommodate greater product diversity. The studies provided nearly identical results for these values. In addition, all service facilities are considered to be reliable to the extent that breakdowns are rare and can be excluded from the analysis.

There has been considerable work on improving the efficiency of such production lines, but in most cases, only a single service facility is assumed at each station. For example, Wild and Carnall [1976] consider the effect of buffer storage capacity on the output of a series of work stations, each consisting of a single facility. Altiok and Stidham [1983], Buzacott [1963, 1967, 1971], Yamashina and Okamura [1983], Freeman [1964], and Barter [1962] show how buffers can be used to improve the system performance. Muth [1973] considers the effect of service time variability on system efficiency, while Payne, Slack and Wild [1972] show how a line is effected by changes in the order of service. Hillier and Boling [1967a, 1967b, 1972] realized that unbalancing the workloads in these systems can lead to improvements in system efficiency giving rise to the "bowl phenomenon" that they conjecture is optimal. Generalizations of parameters and improvements in computations have been done by them [1967a, 1972], in addition to Rao [1976], Gershwin [1987], and Magazine and Silver [1978].

Some recent relevant research on allocating facilities to and/or workload among stations in serial production systems has been done by Hillier and So [1989, 1991] and Hillier, Boling and So [1990]. Hillier and So [1989] look at *where* to place a fixed number of extra facilities (over an initial equal allocation to all stations) so as to maximize throughput. Among their conclusions are that the interior stations, especially the center stations, should be given preference over the end stations for receiving an extra facility when workload is balanced. Also, for the case where the total number of facilities is an integer multiple of the number of stations, an equal allocation of servers is optimal. Their study then focuses on allocating the extra facilities after an equal allocation to all stations, and for queue capacities of zero or one.

Hillier et al. [1990] investigate the problem of determining the optimal allocation of buffer storage between stations of a serial production line given an *equal* allocation of workload to the stations. Their conclusions are that it is generally better to give preferential treatment to stations at

or near the center of the line and that when the total *amount* of storage space also is a decision variable, the optimal solution follows a "storage bowl phenomenon" whereby the allocation of buffer space follows an inverted bowl pattern. Because of the discreteness of the buffers, this phenomenon doesn't always hold.

Hillier and So [1991] study the simultaneous optimization of facility and workload allocation. One very interesting result is the "L-phenomenon," whereby throughput is maximized by assigning all extra facilities over one per station to just one of the end stations, while also allocating to this station by far the largest workload per server. They also find that extra facilities add far more throughput per facility than an initial single-facility per station (the "multiple-facility phenomenon").

There is a related literature in some studies on the optimal allocations of facilities to work stations and of workloads to work stations using *closed* networks of multiserver queues. See Dallery and Stecke [1990], Shanthikumar and Yao [1986, 1987, 1988, 1989], So [1989], Stecke [1986], Stecke and Kim [1989], Stecke and Morin [1985], and Stecke and Solberg [1985]. The latter shows the benefits from allocating both unbalanced workloads and unbalanced facilities per station.

The most important contributions of these papers deal with the qualitative statements that can be derived from them. We know that these production lines can be made more efficient by: reducing the variability of service; increasing the amount of storage; or reducing the number of work stations. We even have an understanding of where we can expect to have the greatest marginal improvements when several of these alternatives are available to us. We want to increase this understanding by continuing the work on production line design intuition. Previous work has looked at the tradeoffs involved regarding these parameters mentioned above. We introduce a new parameter, that of dividing the allocated work at a station into several parallel facilities. Such an instance could come about if the work itself can be split, such as in piece work, where the processing time on each item is small relative to the total processing time in the time period; or, where the work capacity can be divided--this could be workers, although our continuity

assumption would make this unlikely unless the number of workers is large. Finally, the workload may consist of several operations, which can then be partitioned.

It would be useful for design purposes to be able to make additional design statements, regarding the *effect of multiple facilities* at each station. Would a single faster server be more efficient than several slower servers in parallel? Where is storage most critical when a particular design configuration is presented to us? Does the bowl phenomenon still lead to optimal-shaped allocations? We hope to help answer these and other questions by examining a variety of situations and making several observations and conjectures. The motivation behind these conjectures will be from computational results.

Some of our results are similar to previous results in the literature, while some results are different. This can be because of the particular models used, the particular scaling of the parameters chosen, and the assumptions made. Following the presentation of our model and results, we compare our model, assumptions and results with those of previous studies, to try to explain the observed similarities and differences.

2. MODEL FORMULATION

Our production line consists of N serial work stations. Each work station has F_i (i=1, 2, ...,N) service facilities in parallel and a finite number of storage locations, S_i (i=1,2,...,N-1), after it as depicted in Figure 1.

Mathematically, we can state the problem as:

Maximize $R(N; F_i; \mu_{ij}; S_i)$

$$\text{s.t.} \quad \sum_{i=1}^{N} \left[\sum_{j=1}^{F_i} \mu_{ij} \right]^{-1} = N,$$

$$\mu_{ij} > 0, \qquad i=1,2,...,N \text{ and } j=1,2,...,F_i$$

where

R = expected *output rate* for the production line

N = number of *work stations*

F_i = number of parallel service *facilities* at work station i (i=1,2,...N)

μ_{ij} = the *mean service rate* for the jth service facility at the ith work station ($j=1,2,...,F_i$; $i=1,2,....,N$)

S_i = the *number of storage spaces* following work station i ($i=1,2,...,N-1$).

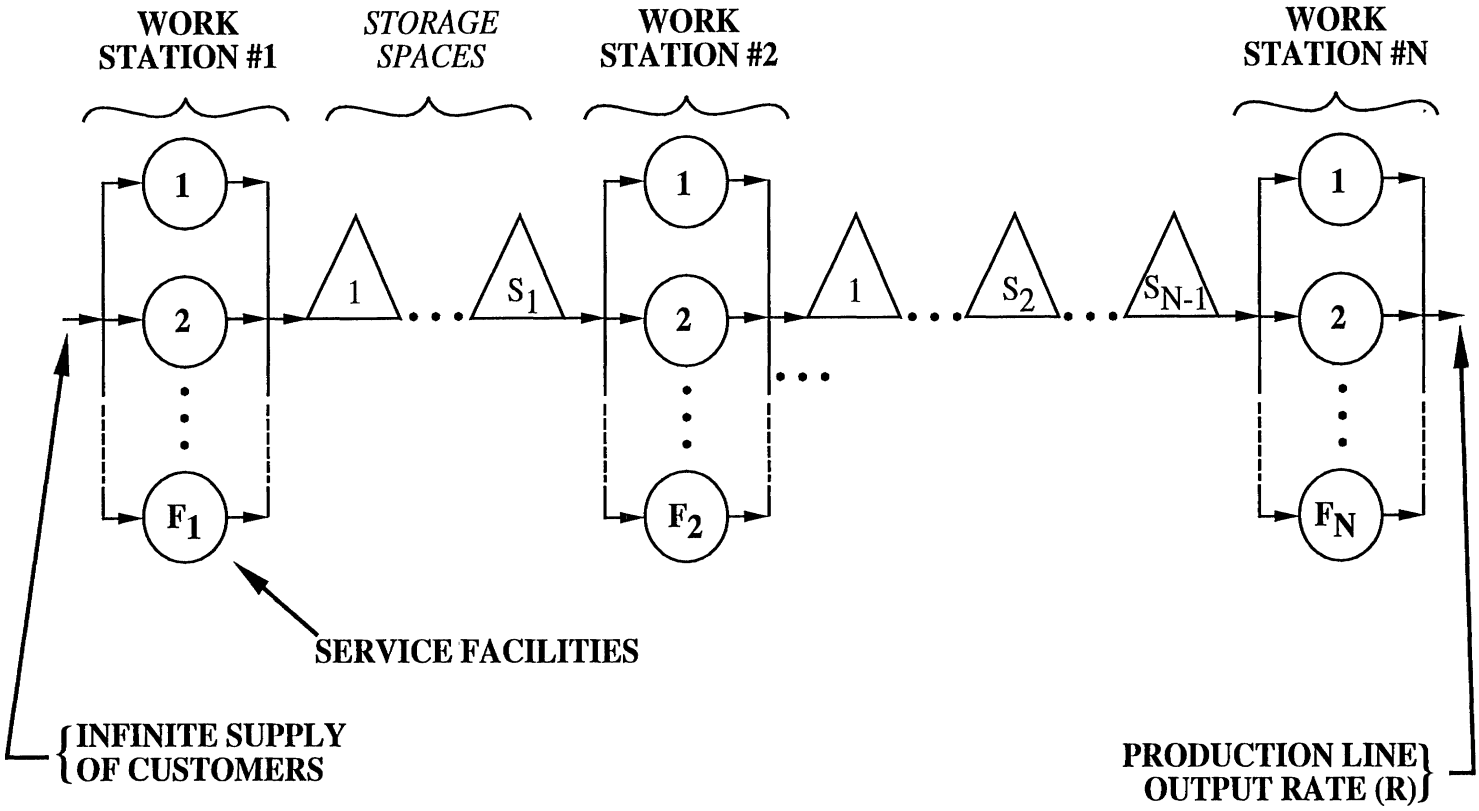

Figure 1. An N-Station Production Line.

The measure of effectiveness considered here is the *expected output rate*, or throughput, of the production line. So that various systems can be compared, the *objective function* is constrained by the *amount of service time capacity available*, i.e.,

$$\sum_{i=1}^{N} \left[\sum_{j=1}^{F_i} \mu_{ij} \right]^{-1} = N. \qquad (1)$$

This is the total *workload* required by the system. One of the design parameters is how to allocate this fixed amount of work or service time capacity to the stations in the production system. This is equivalent to allocating service rates to the work stations, as we will see in the next section. If each server is always busy, then one customer will leave the system each time period, on the average.

We assume that the rate of service at each parallel facility at a particular station is the same, because of the symmetry of these facilities. This is easily verified for two facilities at a given

station. Therefore, when $\mu_{i1} = \mu_{i2} = ...\mu_{iF_i}$ (i=1,2,...,N), then $\sum_{j=1}^{F_i} \mu_{ij} = F_i \mu_{i1} = \mu_i$ and equation (1) becomes: $\sum_{i=1}^{N} [\mu_i]^{-1} = N$.

If $\mu_1 = \mu_2 = ...= \mu_N = 1$, then the production line is said to be "*balanced*". When the service times are *constant* (i.e., the variance $\sigma^2_{ij} = 0$) then balancing maximizes R, which is then equal to 1. The value of R, therefore, can be used to measure the efficiency of the system, relative to the deterministic balanced case.

To find exact solutions for these problems, we would have to work with state space equations whose number explodes as the size of the problems grow. A general recursive formula for the number of states is:

$$SS_N = SS_{N-1} (F_{N-1} + S_{N-1} + F_N + 1) - SS_{N-2} F_{N-1} (F_{N-1} + 1)/2,$$

where SS_n = the number of states when there are n work stations

$SS_1 = 1$ and $SS_0 = 0$.

Table 1 indicates this explosiveness. Because of the computational burden, we investigate serial lines with two and three work stations, up to seventeen facilities, and up to six buffer spaces.

Table 1. State Space Explosion.

Number of Storage Spaces	Number of Work Stations											
	2			3			4			5		
	Number of Service Facilities											
	1	2	3	1	2	3	1	2	3	1	2	3
0	3	5	7	8	22	43	21	95	259	55	409	1555
1	4	6	8	15	33	58	56	180	416	209	981	2980
2	5	7	9	24	46	75	115	301	621	558	1668	4899

Because of this computational burden, theoretical results, for the most part, have not been tractable. We have verified our conjectures (summarized in Section 5) for small problems. The mathematics, however, adds nothing to our understanding of these systems, and will not be repeated here.

3. EFFICIENCY OF TWO SERIAL WORK STATIONS

To understand how different parameters affect the output of serial work stations, let us first analyze a two work station configuration, i.e., N = 2. An obvious advantage of this instance is that we can analytically determine R for several different parameter configurations. In addition, some properties of serial production lines can be verified for N = 2.

The problem is then to:

$$\text{maximize } R(2; F_1, F_2; \mu_1, \mu_2; S)$$
$$\text{s.t.} \quad 1/\mu_1 + 1/\mu_2 = 2$$
$$\mu_1 \text{ and } \mu_2 > 0.$$

Here we can clarify the relationship between the average service rates and the average service times. For a two-station system, the sum of the average service times equals the number of stations, which is the total service capacity:

$$1/\mu_1 + 1/\mu_2 = 2.$$

Solving for the average service rate at station 2,

$$\mu_2 = \frac{\mu_1}{2\mu_1 - 1}.$$

Then the sum of the average service rates is:

$$\mu_1 + \mu_2 = \frac{2\mu_1^2}{2\mu_1 - 1}.$$

This can help to clarify one of the differences in scaling between our study and those of Hillier and So [1989, 1991b]. Although all of these studies examine workload and/or facility allocation, there are different scalings of the output rate. For example, Hillier and So [1989] require that

$$\sum_{i=1}^{N} F_i = F. \tag{2}$$

Hillier et al. [1990] constrain the output according to the number of buffers:

$$\sum_{i=1}^{N-1} S_i = S.$$

Hillier and So [1991b] constrain both the sum of facilities (2) and the sum of workloads (1). One difference is that because of equation (2), the output rate of Hillier and So [1989, 1991b] is usually greater than one, but cannot exceed the minimum number of servers at any station. Such differences in scaling may sometimes cause the occasional differences in results.

As a result of our choices of scaling (see equation (1)), our output rate is always less than or equal to 1 and the capacity of the system remains constant. In Hillier and So [1989], the addition of extra facilities adds additional capacity to the system and makes comparisons difficult. In addition, our results aren't as intuitive. It isn't surprising that additional capacity would increase the throughput. It is, however, less obvious what would happen if we split the capacity allocated to a workstation. Results from classical queueing theory suggest that the two slower parallel servers may not be as efficient as the single faster server. We believe this is an important difference in our results,

The design parameters that we consider are: (1) the addition and placement of facilities; (2) the allocation of the two units of work capacity, i.e., $1/\mu_1$ and $1/\mu_2$; and (3) the effects of changes in the buffer storage S. Through direct comparison of R, we can verify first, that each facility is given equal amounts of the service capability afforded to that station and therefore, specifying μ_i and F_i determines μ_{ij}, $j = 1,2,...F_i$ and secondly, that the "reversibility property holds." This property says that for any design configuration, $(\bar{F}_1, \bar{F}_2, \bar{\mu}_1, \bar{\mu}_2, \bar{S})$ leading to output $\bar{R}$, there is another configuration, where $F_1 = \bar{F}_2$, $F_2 = \bar{F}_1$, $\mu_1 = \bar{\mu}_2$, $\mu_2 = \bar{\mu}_1$, and $S = \bar{S}$ having the same output rate. This result has been proven for single facility systems by Muth [1973] and for somewhat more general systems by Dattatreya [1978] including serial systems of parallel stations when N = 2. However, it has been shown to not hold for serial systems of parallel stations with N > 2 (see Yamazaki et al. [1985] and Hillier and So [1989]). They show that although reversibility does not hold, the differences in throughput for the mirror image lines are very small, 0.0015, for lines with multiserver stations. They conclude that reversibility is close enough that only one of the two configurations needs to be tested. The purpose of each of these facts is to use

them to reduce the number of possible configurations that we must look at in order to feel confident about coming to conclusions. It is sufficient for us to investigate only one of the mirror images.

The mechanism we use will be to hold all but one of the controllable parameters fixed and systematically vary only this one parameter at a time. Values of R are computed in each instance and patterns follow. R is calculated from the steady-state balance equations. After much tedious algebra, we arrive at:

$$R = \mu_1/(2\mu_1 - 1)\left[1 - P \sum_{n=0}^{F_2} [(F_2 - n)\ \{F_2(2\mu_1 - 1)\}^n / F_2 n!]\right],$$

where

$$P = \left\{1 + \sum_{n=1}^{F_2} \left[\{F_2(2\mu_1 - 1)\}^n / n!\right] + \left[F_2(2\mu_1 - 1)\right]^{F_2} / F_2! \sum_{n=1}^{S} (2\mu_1 - 1)^n \right.$$

$$\left. + F_2^{F_2}(2\mu_1 - 1)^{F_2+S} / F_2! \sum_{n=1}^{F_1} \left[\frac{F_1!}{(F_1-n)!}(2\mu_1 - 1)^n / F_1^n\right]\right\}^{-1}.$$

Although it is possible to present lengthy tables showing R for various parameter values, we feel that the information we wish to convey is better done with curves. For readability, these are drawn continuously, by interpolating the results from discrete trials.

3.1 Unbalancing Workloads When $F_1 = F_2$

To begin with, there is an equal number of facilities at each work station and there is no storage between stations 1 and 2. Figure 2a shows the effect of unbalancing workloads for several configurations F_1/F_2, where $F = F_1 + F_2$. As we move away from the balanced case, $\mu_1 = \mu_2 = 1.0$, it is evident that R decreases. In addition, as F increases, both the output rate and the importance of balancing increases.

From Figures 2a and 2b, we note that the detrimental effect of unbalancing on R grows as F increases. Hence, the larger the number of facilities, the larger the penalty for not balancing. These results agree with those of Stecke and Morin [1984] and Stecke and Solberg [1985] using closed queueing networks of multiserver queues.

3.2 Unbalancing Workloads When $F_1 \neq F_2$

The above results do not hold when the number of facilities at each work station is not the same. To illustrate this, let us examine the extreme case where $F_1 = 1$ and $F_2 = F - 1$. (From reversibility results, this is the same as $F_1 = F - 1$ and $F_2 = 1$.) The direction of unbalance is important in these cases. From Figure 3a, we can see that *R increases* as we shift the burden of effort or *increased workload* (decreased service rate) towards the station with fewer facilities. Here, the single-facility station receives the most work (or equivalently, it works at a slower rate). The multi-facility station receives less work. Each facility works at a faster rate. An alternative interpretation is the following. Given a fixed amount of work, output is increased by having each facility of the parallel station working faster than the single facility. This does not agree with the results of Stecke and Solberg [1985] using closed queueing networks of multiserver queues. In that model, it is best to assign a higher workload per facility to the station with more facilities.

Clearly, there is a point below which decreasing μ_1 will also decrease R. We know that this number is less than 1.0, but will withhold discussion of the optimal allocation of workload for now. Once again, as is indicated in Figure 3b, the effects on R of unbalancing are magnified as F increases.

It is also the case that unbalancing in the direction of allocating greater workloads to those stations having fewer facilities increases R for those intermediate cases, where $1 < F_1, F_2 < F - 1$. See Figures 3c and 3d. As we might expect, the closer the configuration is to the balanced case, the less unbalancing is necessary before R starts decreasing. It is also the case that unbalancing too far creates a more dramatic fall-off in R when F is large.

3.3 Arrangement of Service Facilities when Workloads are Balanced

We know for any configuration of facilities, in which direction we should unbalance. Let us assume now that this configuration is a design parameter and under our control. Specifically, we first assume that *workloads are balanced* ($\mu_1 = \mu_2 = 1.0$) and S = 0. We start with $F_1 = F_2 = 1$ and increment, by one facility at a time.

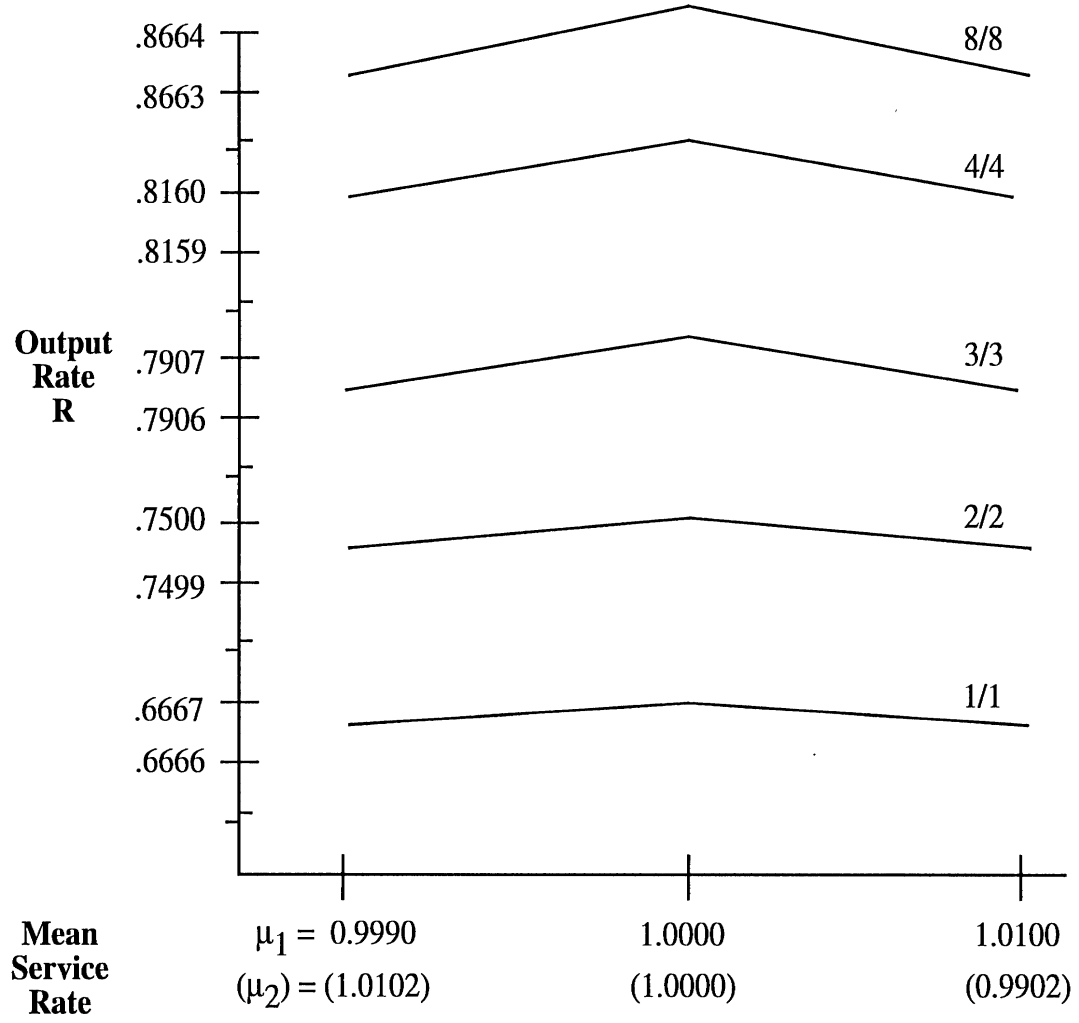

Figure 2a. Output Rates when F_1=F_2, S=0, and Workloads Vary.

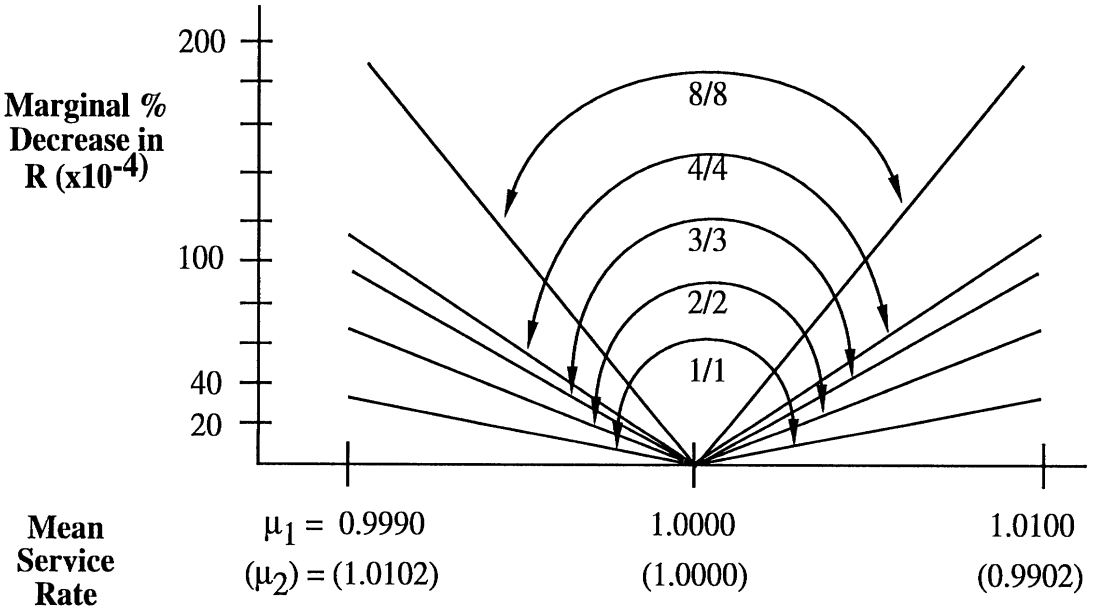

Figure 2b. Marginal % Decrease in R as Workloads Vary.

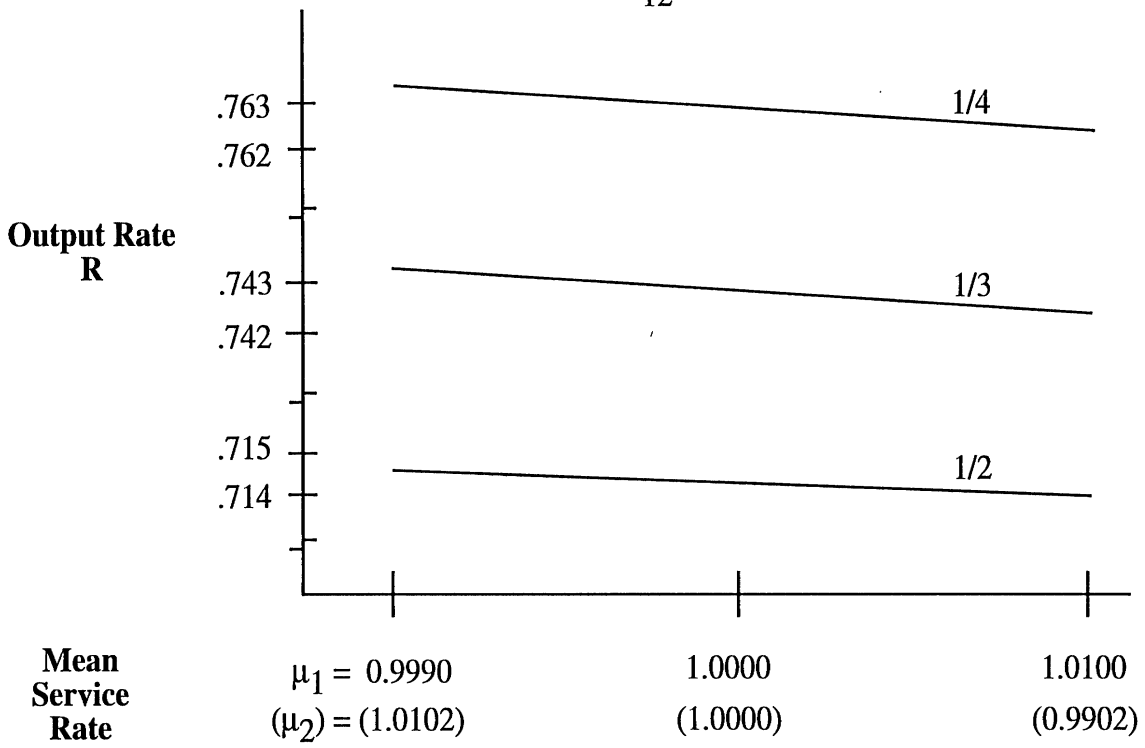

Figure 3a. Output Rates for Configurations 1/F-1 and S=0.

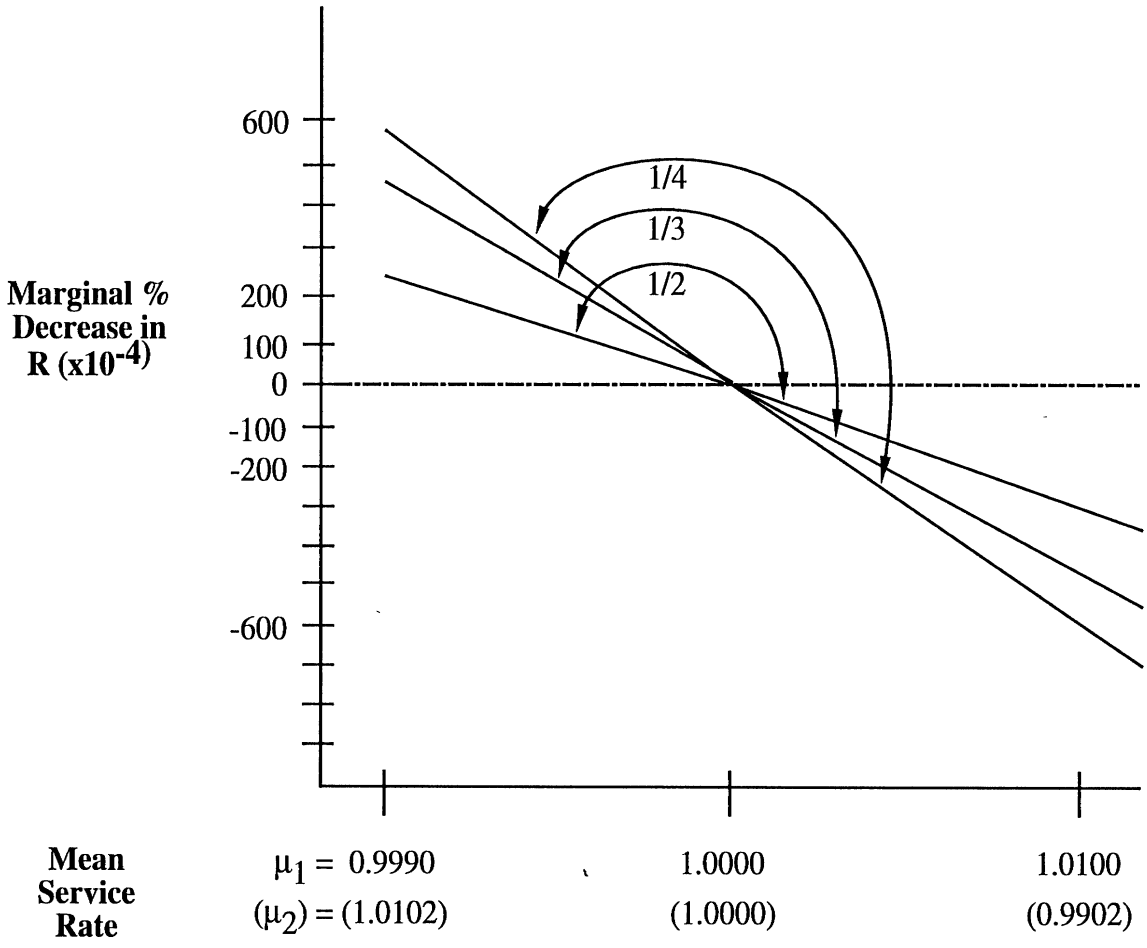

Figure 3b. Marginal % Increase in R for Various Configurations as Workloads Vary.

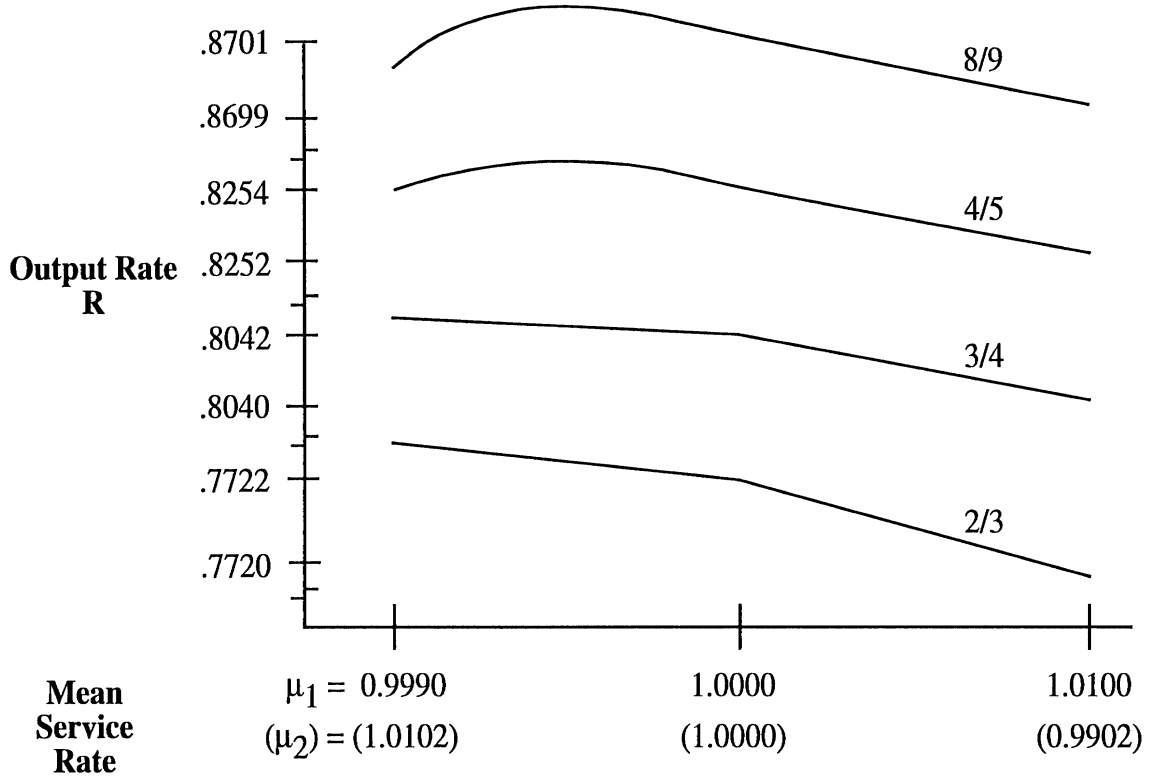

Figure 3c. Output Rates for Configurations where $1 < F_1, F_2 < F\text{-}1$ and $S = 0$.

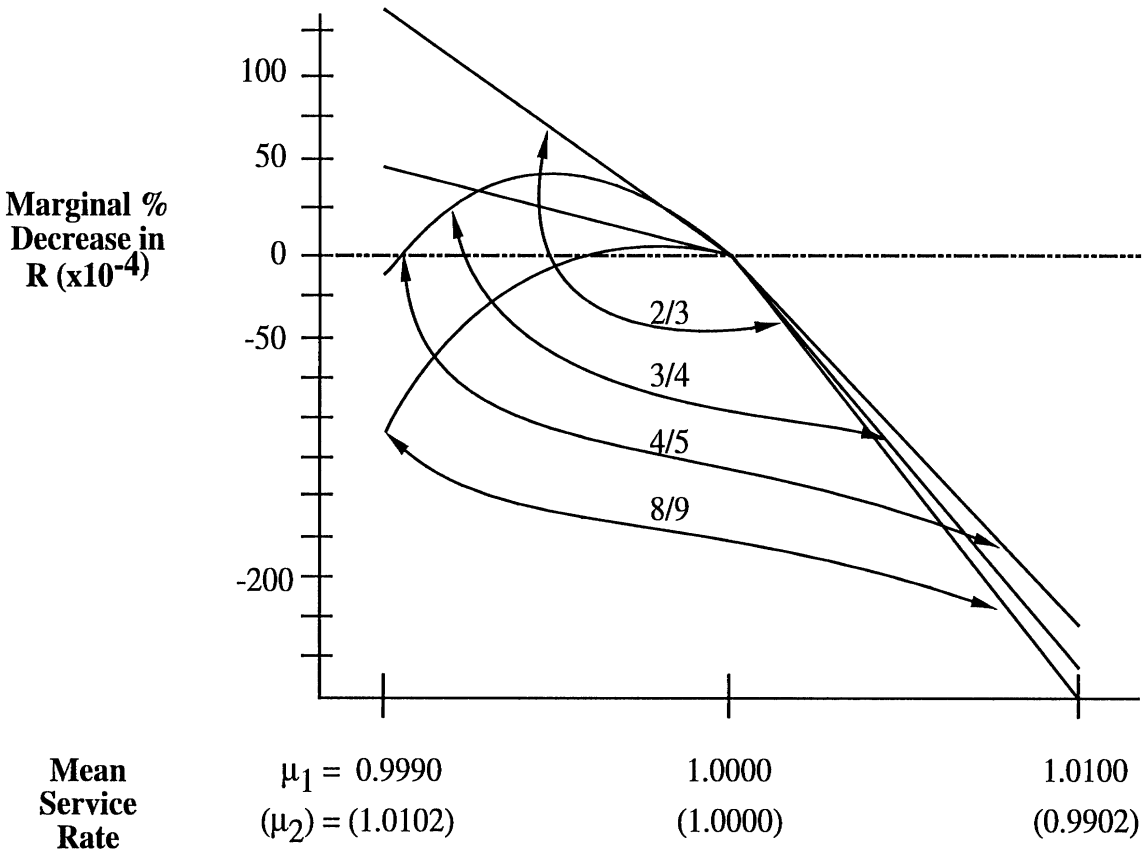

Figure 3d. Marginal % Increase in R for Various Configurations as Workloads Vary.

Figure 4 indicates the effect on R for different values of F_1/F_2. It is clear that the best configuration attempts to equalize F_1 and F_2. Hence, if one facility is to be added to the system, it would be best to place it in the station with the fewest number of facilities. It should also be noted that configurations with larger F need not be better if they aren't carefully allotted. For instance, configuration 3/3 (with an output rate of .79070) is a little better than 6/1 (with output rate of .79056) in spite of the additional facility. These results also agree with those of Stecke and Solberg [1985]. The results here are more directly comparable particularly since in this section, there is the constraint that workloads are balanced. Of course the 6/1 configuration can provide a higher output rate than 3/3 if the workloads are allowed to be unbalanced, as we shall see in Section 3.5.

3.4 Effect of Storage Spaces

The third design variable, and the most significant in terms of its effect on R, is the number of available buffer spaces between the two work stations. Figure 5a illustrates the increase in the output rate for *balanced workloads* and *evenly* (when possible) *split facilities*. Note that the increase in R (by almost 13%) from configuration 1/1 by changing S from 0 to 1 is the same as incrementing the facility configuration to 2/2.

In Figure 5b, the marginal percentage increases in R are shown as S changes from 0 to 1, 2, or 3. The positive effect of adding storage spaces diminishes as F is increased. From Figure 5c, the converse is also seen to be true. That is, for larger S, the positive effect of adding facilities also decreases.

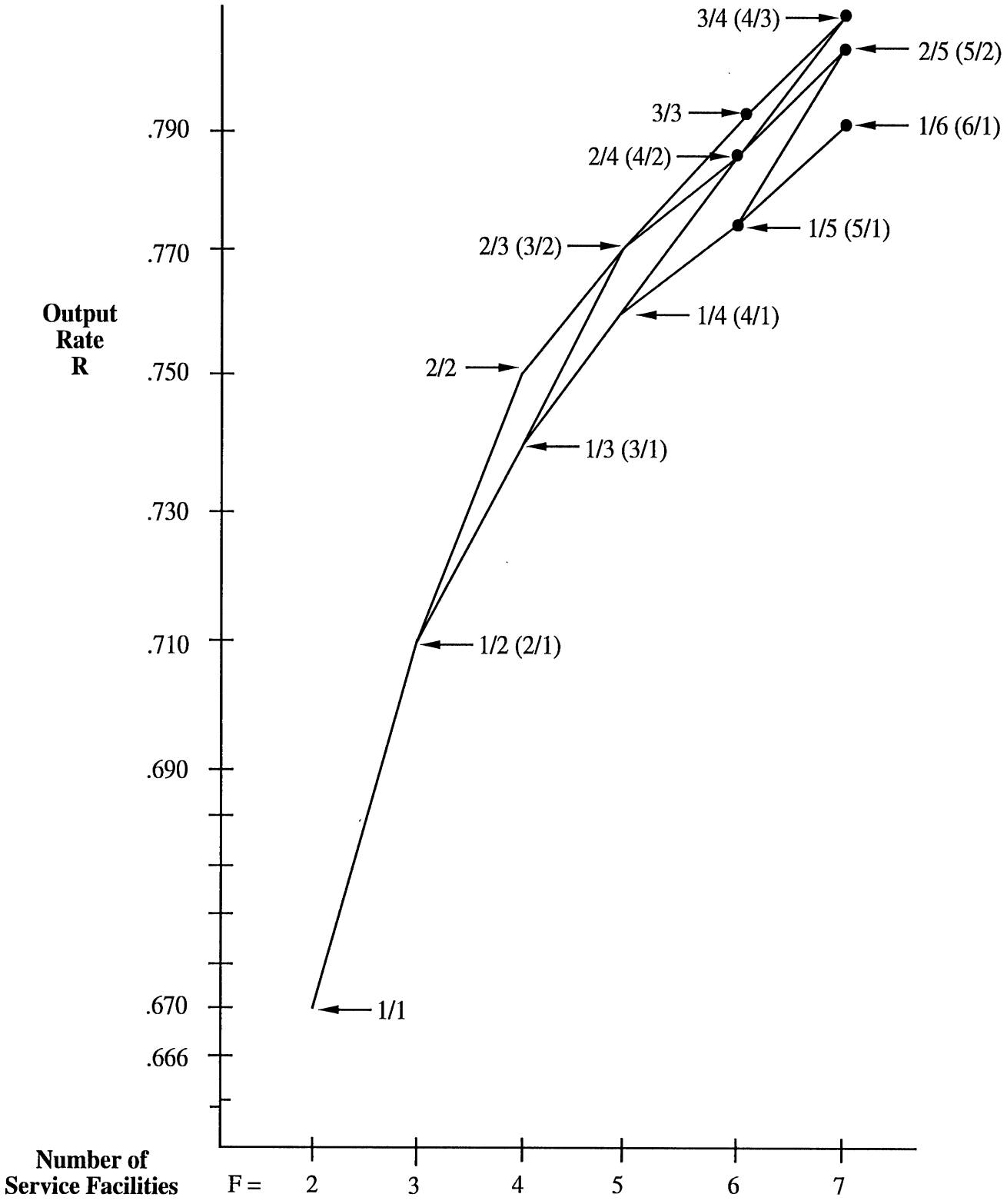

Figure 4. Output Rates for Various Configurations when *Workloads are Balanced* and S = 0.

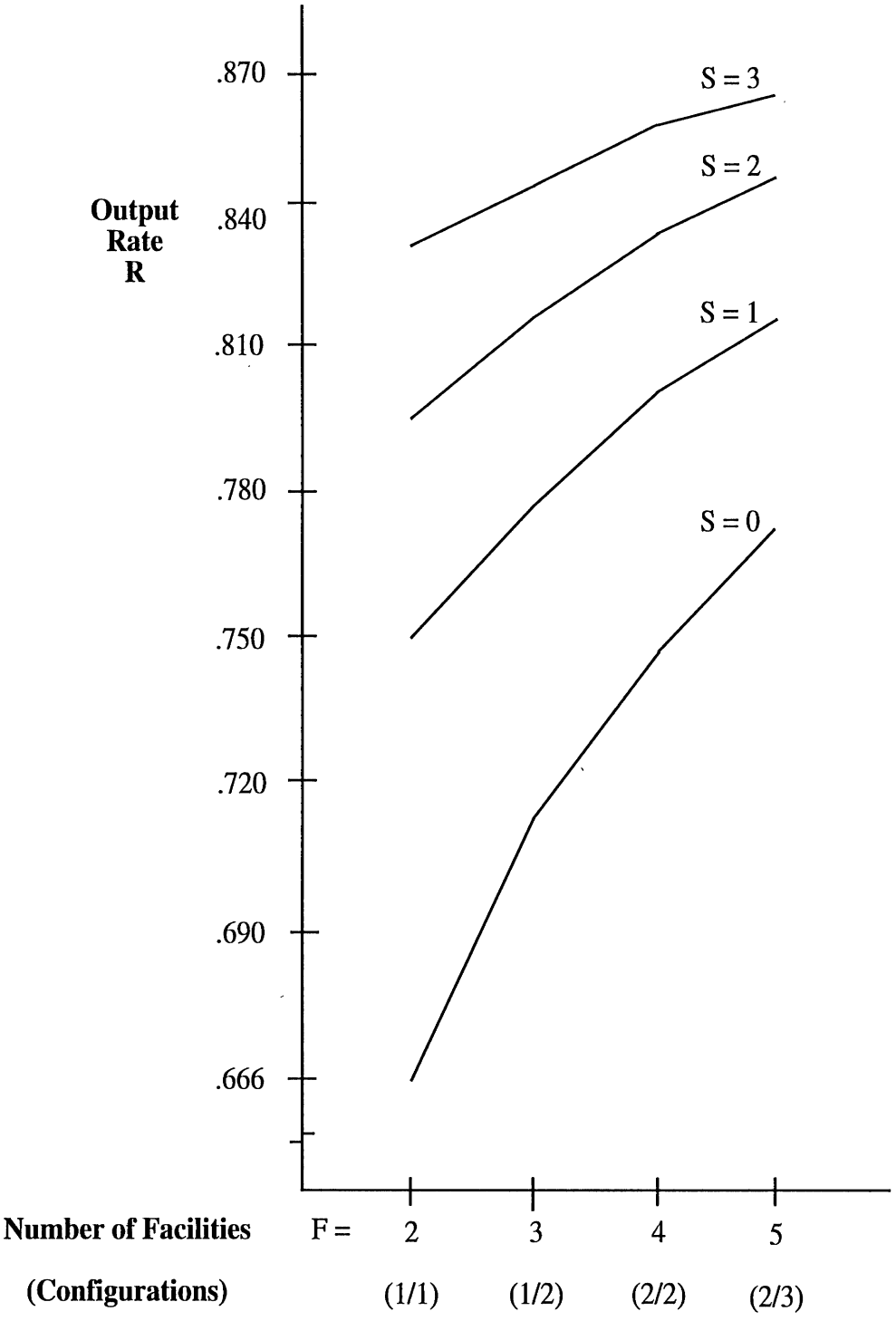

Figure 5a. Output Rates for Evenly Split Configurations Having Balanced Workloads while Varying the Amount of Buffer Space.

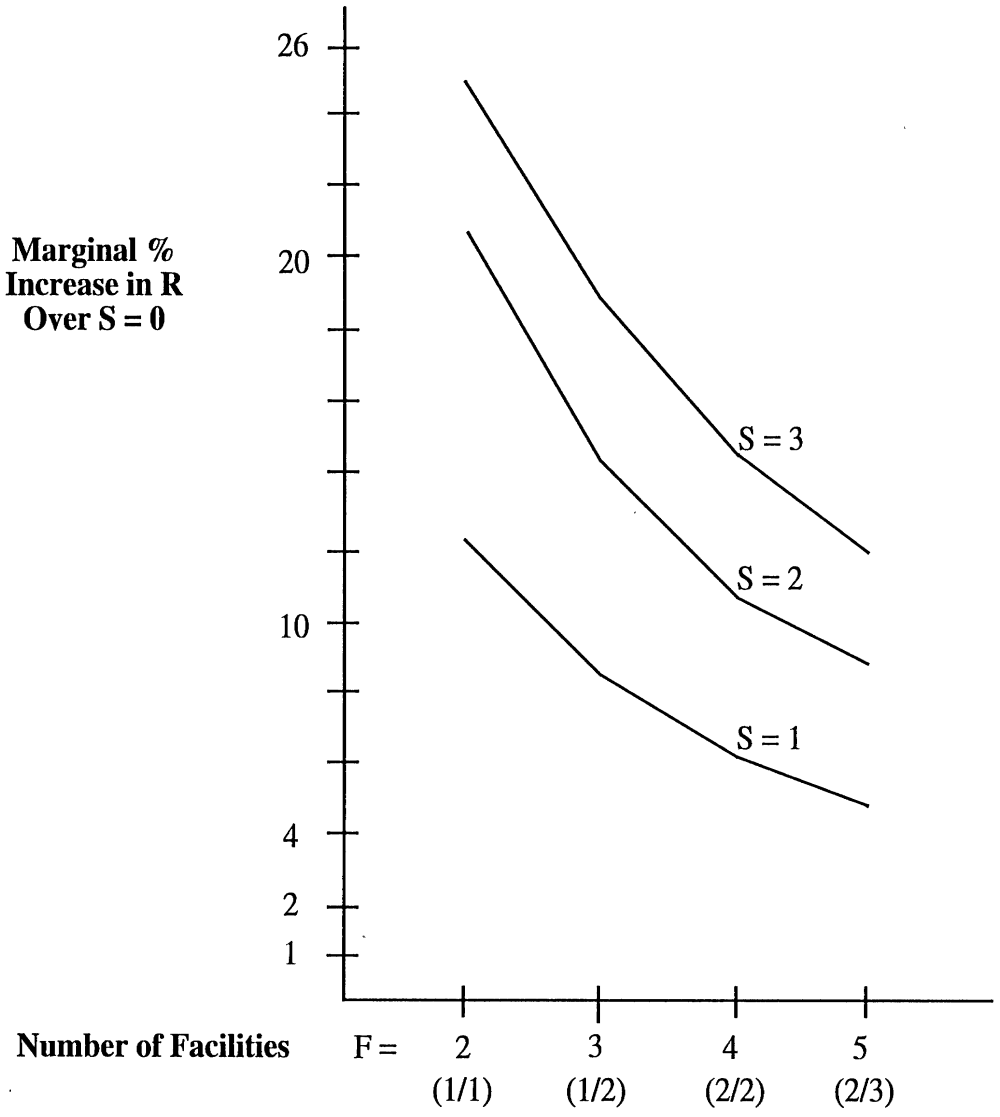

Figure 5b. Marginal % Increase in R by Varying the Number of Buffer Spaces.

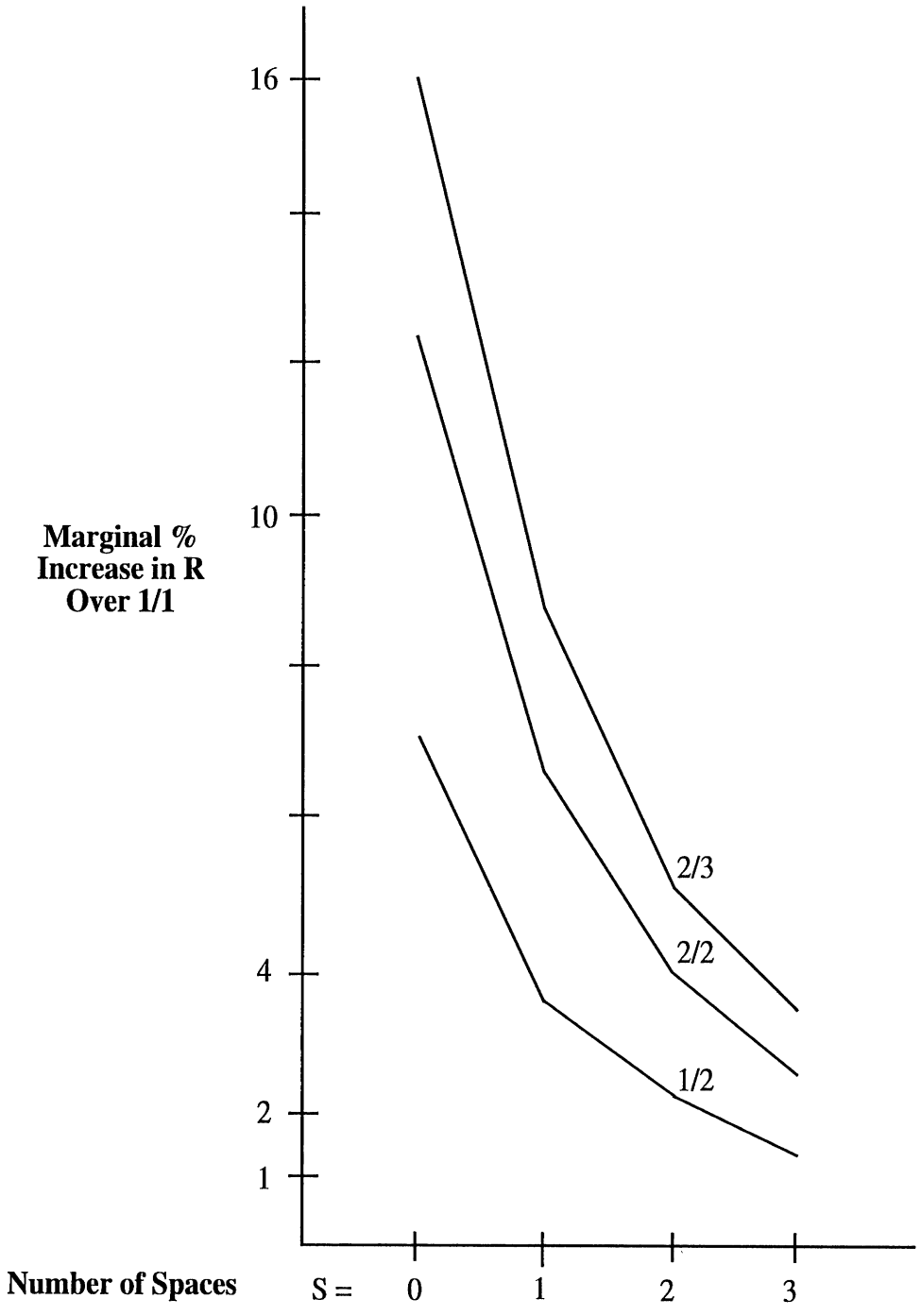

Figure 5c. **Marginal % Increase in R for Various Configurations and while Varying the Number of Buffer Spaces.**

3.5 Determination of "Optimal" Parameter Values

Understanding the effect of changes in the parameters assists us in choosing the best configuration under our control. Obviously, R increases asymptotically to 1.0 as S or F are increased. We would like to determine the best allocation of fixed resources by fine tuning the maximum value of R. We say fine tuning because of the, sometimes, very small improvements in R over the balanced configuration. For example, for the configuration 1/6, the optimal unbalancing of the service rate at station 1 by 4% (from $\mu_1 = 1.0$ to $\mu_1 = .96$) results in a .19% increase in R. Also, some parameters have to be changed in discrete amounts.

Figure 6a shows the locations of the maximum values of R for various fixed configurations of facilities when S = 0. As F increases, not achieving optimal unbalancing has a marginally increasing effect when F is very unevenly split, e.g., 1/F-1. The effect of increasing F, however, is diminishing as the configuration approaches the evenly split case. In addition, the *amount of unbalancing* workloads required to maximize production *increases* with F for the extreme case (1/F-1) and *decreases* as the configuration approaches the evenly split case.

Finally, it should be noted that if unbalancing and distributing facilities is left to our control, both should be as balanced as possible. This is the opposite conclusion to that of Stecke and Solberg [1985], where for a closed network of two connected queues, the maximum expected output is achieved by unbalancing both the configuration and the workload. Reasons for the different conclusions may be the different situations, assumptions, and models. Here, there is no buffer and a fixed route. Stecke and Solberg allow random routing and allow adequate buffer space to hold all waiting customers.

Figure 6b illustrates the effect of increasing S for a fixed F (7 in this case) with various configurations. For each configuration, the amount of unbalancing that is necessary to attain the maximize the output rate for any fixed S decreases as the facilities are more evenly split, but, the fall-off from too much unbalancing is faster in these cases.

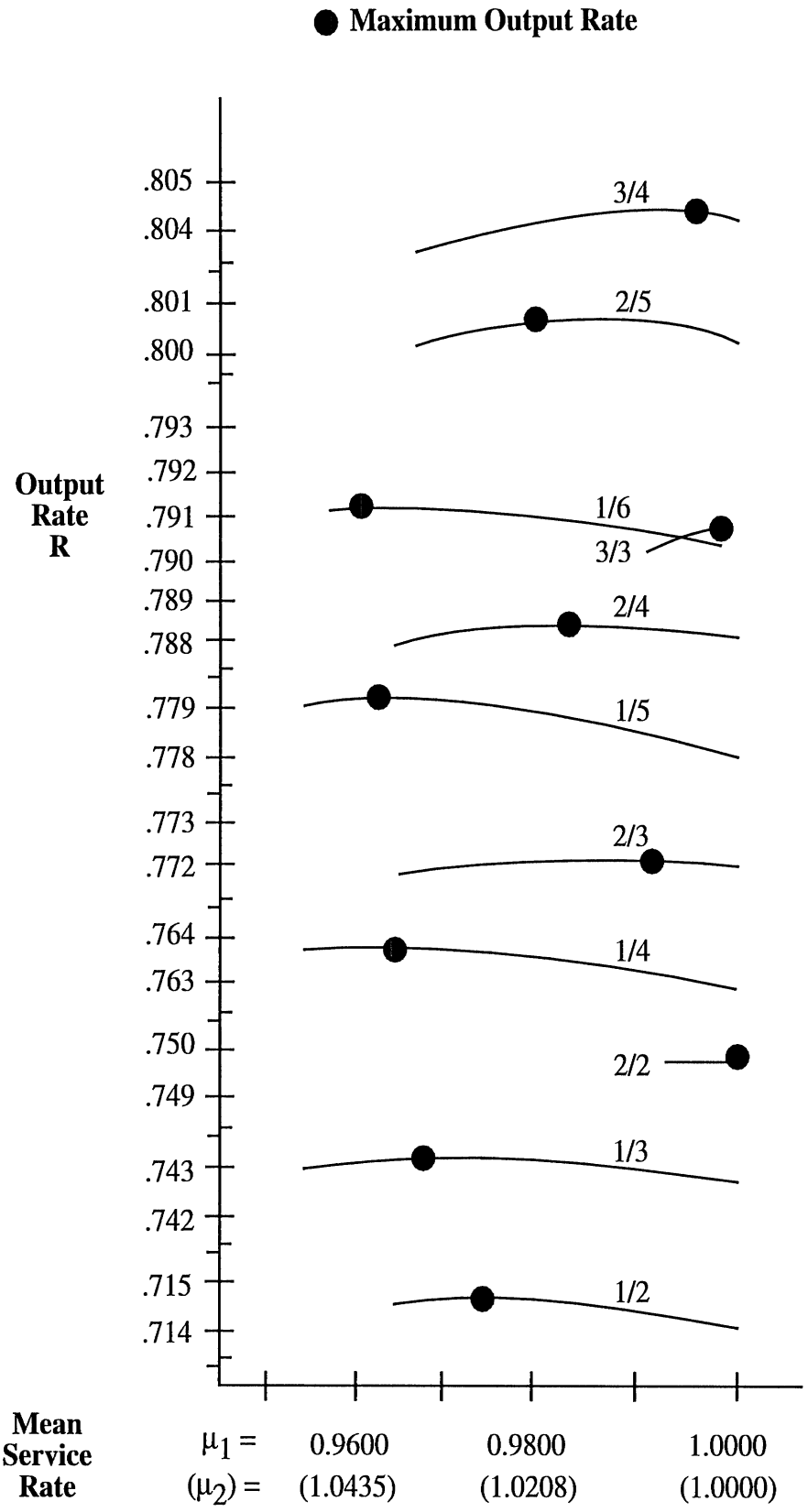

Figure 6a. Maximum Output Rates for Various Configurations and Workloads for S = 0.

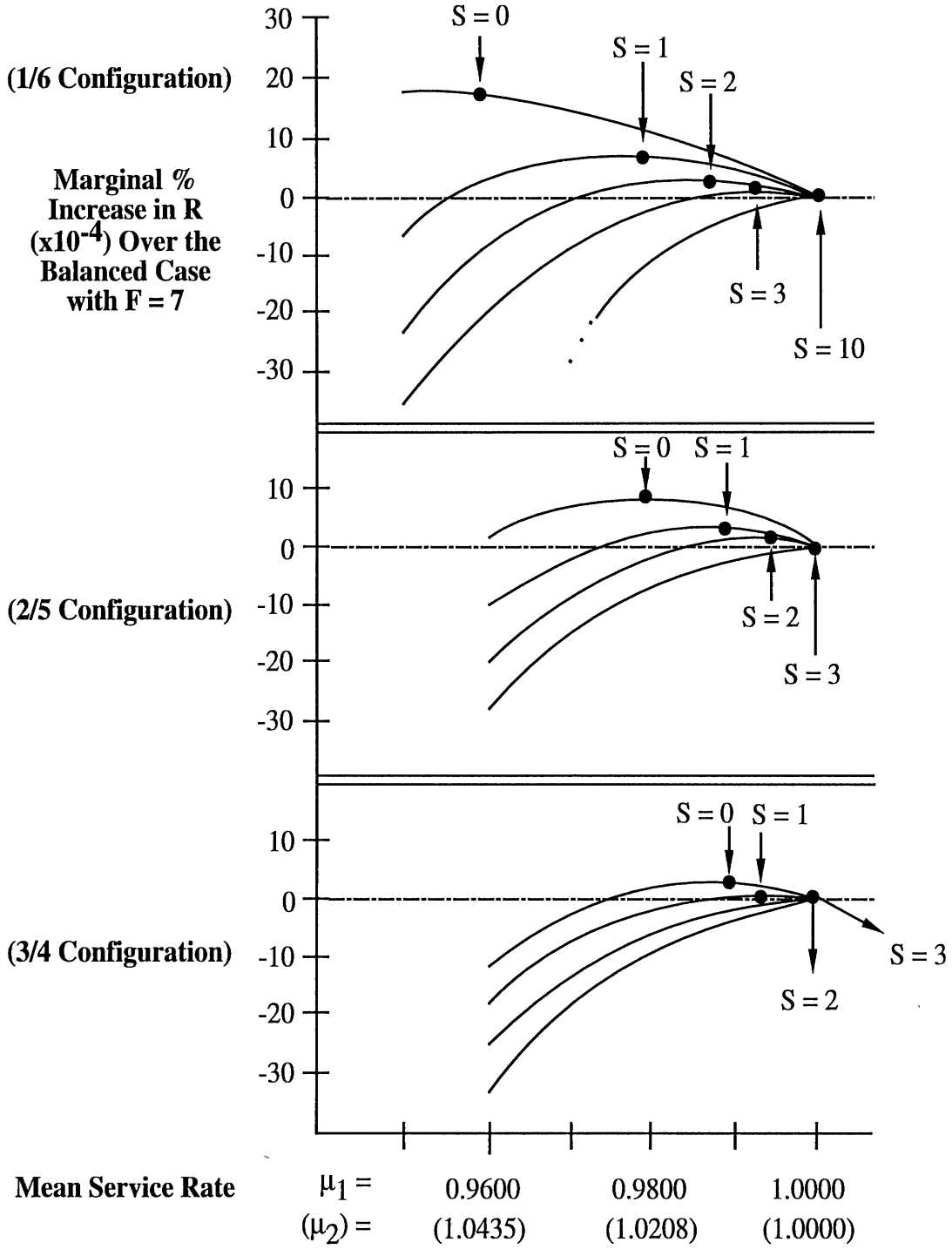

Figure 6b. Marginal % Increase in Production Rate for Various Configurations of Seven Facilities while Varying the Amount of Buffer Space.

4. PARAMETER SENSITIVITY WHEN N = 3

It would be difficult to try to draw general conclusions concerning the best configurations simply from the two work station case. We might expect similar behavior for several parameter changes as the number of work stations increases to three, but certain differences are also anticipated. For instance, when N = 3 we first encounter the "bowl phenomenon" described in Hillier and Boling [1967a, 1967b] for serial systems.

The design parameters we now consider are the addition and placement of facilities, the allocation of total workload among the stations, the increment and placement of storage buffers after the first and second stations, and the increase in the number of stations from two to three. Once again we can assume that the facilities at each work station have the same mean service rates and that reversibility is close enough that we need not explicitly consider the symmetric configurations.

In the subsections that follow, we again, by means of curves, attempt to describe the behavior of R as each parameter is allowed, separately, to vary. From this, we attempt to make statements regarding optimal design configurations when several parameters can vary. We use F and S to denote the total number of facilities and storage spaces available. We also use $F_1/F_2/F_3$ to denote the allocation of the facilities to the different work stations.

4.1 Arrangement of Service Facilities

All combinations of configurations of service facilities from F = 3 to F = 12 were examined to see the effect of different permutations on R. It is assumed initially in this subsection that S = 0 and that the *workload is balanced.* Figure 7 depicts the output rate for many of these configurations. Table 2 provides the best configuration of facilities for each value of F. It is apparent for each F > 3 that the optimal configuration has $F_2 > F_1 = F_3$. As F increases, this bowl phenomenon" effect increases and the *unbalancing of the configuration* in favor of the middle station increases. In other words, as F increases, the middle station needs to be more efficient, so it takes the largest number of the available facilities. As F is increased to F + 1, the optimal allocation of facilities continues to add a facility to the middle station until the unbalancing is too

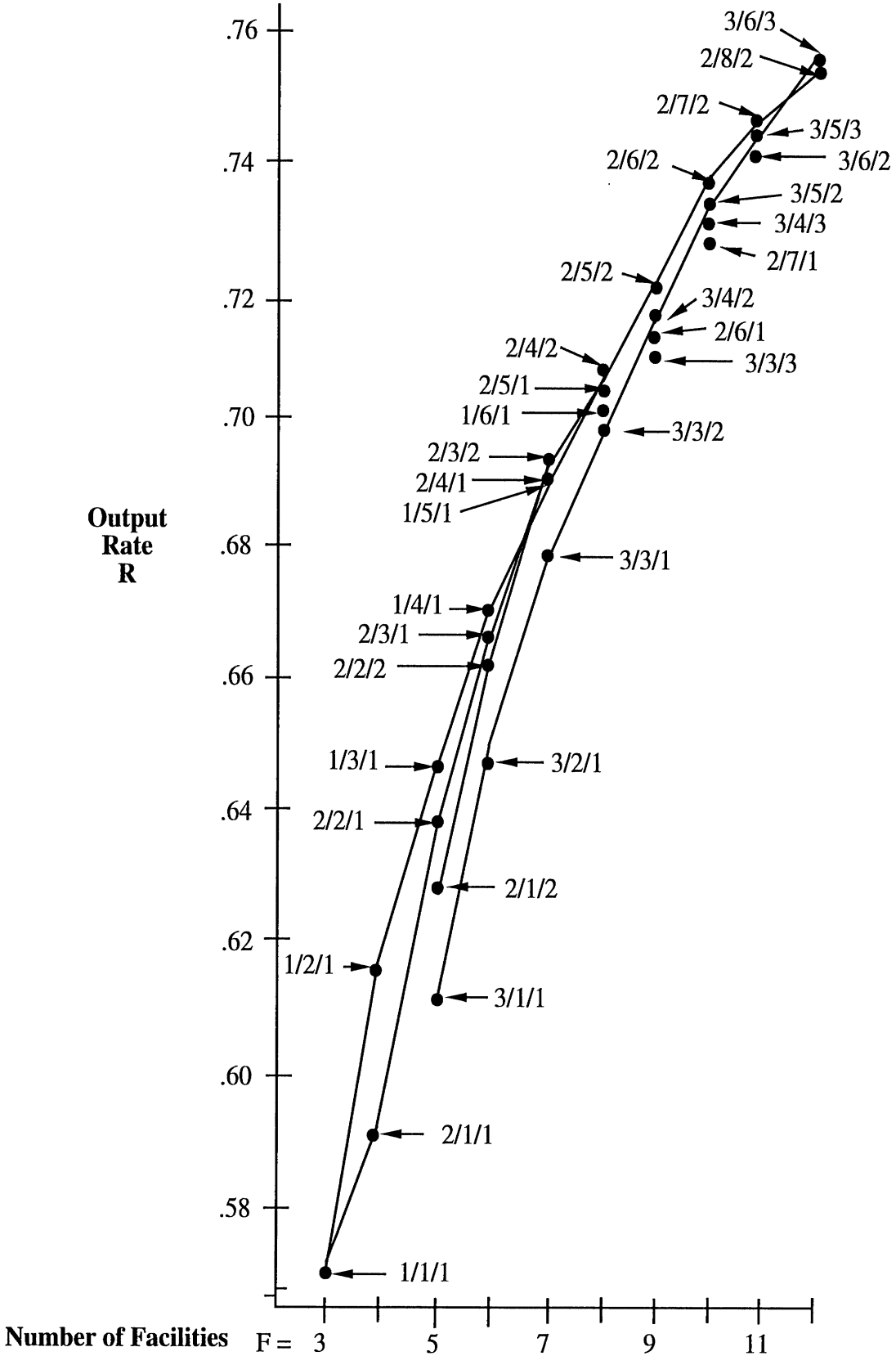

Figure 7. Output Rates for Various F when all $\mu_i = 1$ and $S_1 = S_2 = 0$.

Table 2. Optimal Configurations for Various F When All $\mu_i = 1$ and $S_1 = S_2 = 0$.

F	Optimal Configurations of F		
	F_1	F_2	F_3
4	1	2	1
5	1	3	1
6	1	4	1
7	2	3	2
8	2	4	2
9	2	5	2
10	2	6	2
11	2	7	2
12	3	6	3

severe. At this point, a facility is taken away from the middle station and a facility is added to the first and third stations. In any case, the optimum allocation preserves symmetry. (Note the cases of going from F = 6 to 7 and from 11 to 12.) Also, it is evident that the marginal effect of increasing F decreases as does the penalty for not using the optimal configuration.

These results can also be used to determine the best placement of a new facility when a given configuration (not necessarily optimal) is provided. For instance, if the present design is 2/1/1 and a new facility can be added, we can see that R(2/2/1) > R(2/1/2) > R(3/1/1).

Hillier and So [1989] observe a more subdued bowl phenomenon than occurs here. For example, here the configuration 1/4/1 is better than 2/2/2. See Figure 7. This is not true for Hillier and So. Both are modeling three-station serial production lines with no buffer and balanced workloads. One difference is in the scaling of the output rate. Ours is always less than one. For Hillier and So, R is less than the minimum number of facilities per station for a particular configuration. This scaling *allows* 2/2/2 ($R \leq 2$) to be better than 1/4/1 ($R \leq 1$) here. Further research is required to better understand this. In all cases, there is a bowl phenomenon.

4.2 Unbalancing Workloads

It is not surprising that when the μ_i (i = 1,2,3) are allowed to vary, the symmetric "bowl" configuration is not necessarily optimal. Figure 8 provides optimal values for μ_1 and μ_3 for various configurations. S is still 0 and $\mu_2 = \mu_1\mu_3/(3\mu_1\mu_3-\mu_1-\mu_3)$. When $F_1 = F_3$, we can see that $\mu_1 = \mu_3$. In addition, at optimum, μ_1 and μ_3 are *less than one* for *all* configurations considered. This means that μ_2 is always greater than one.

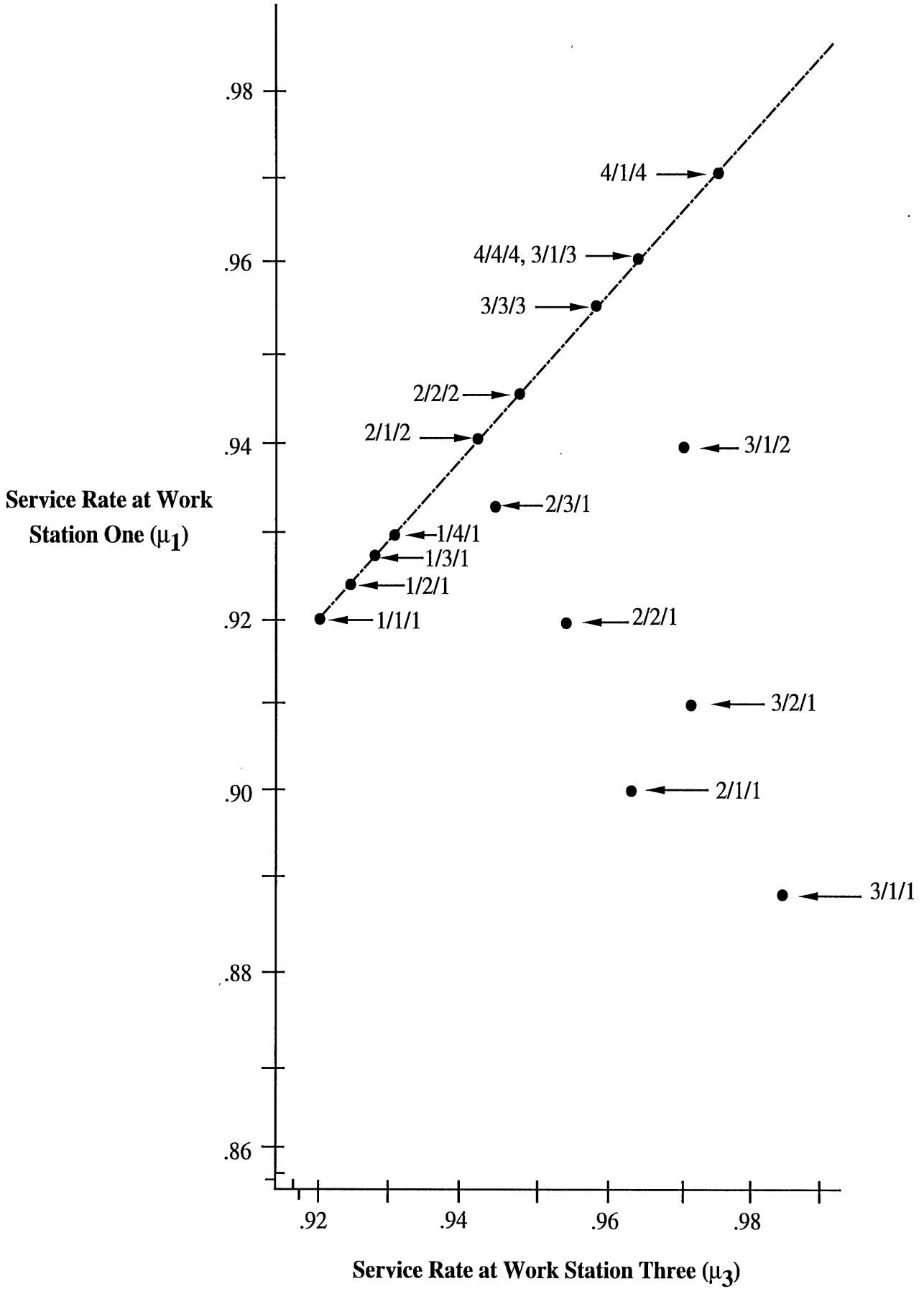

Figure 8. Optimal Workload Allocations for Various Configurations when S = 0.

It certainly appears that, at least for configurations $F_1/F_2/F_3$ that are close to optimal, the bowl phenomenon is once more present. As F increases, the amount of unbalancing of the optimal workloads seems to decrease. Finally, when $F_1 > F_3$ (for any F_2), then $\mu_1 < \mu_3$.

4.3 Allocation of Storage Spaces

Table 3 presents the optimal storage space arrangements for various configurations *with balanced workloads*. When the production line is symmetrical ($F_1 = F_3$), then the maximum value of R is achieved when $S_1 = \lfloor S/2 \rfloor$ ($\lfloor x \rfloor$ is the greatest integer in x) or $S_2 = \lfloor S/2 \rfloor$. When S is odd, both arrangements of S_1, S_2 give the same R. For example, configuration 1/3/1 with S = 3, $S_1 = 2$, and $S_2 = 1$ (or $S_1 = 1$ and $S_2 = 2$) gives an output rate of .74555.

Table 3. Optimal Buffer Space Arrangements for Various Configurations and with Balanced Workloads.

Facility Configurations			Optimal Storage Space Arrangements											
			S = 1		S = 2		S =3		S = 4		S = 5		S = 6	
F_1	F_2	F_3	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
1	1	1	1 or 0	0 1	1	1	2 or 1	1 2	2	2	3 or 2	2 3	3	3
1	2	1	1 or 0	0 1	1	1	2 or 1	1 2	2	2	3 or 2	2 3	3	3
2	1	1	0	1	1	1	1	2	2	2	2	3	3	3
3	1	1	0	1	1	1	1	2						
2	2	1	0	1	1	1	1	2						
2	1	2	1 or 0	0 1	1	1	2 or 1	1 2						
1	3	1	1 or 0	0 1	1	1	2 or 1	1 2						
4	**1**	**1**	0	1	**0**	**2**	1	2						
3	2	1	0	1	1	1	1	2	2	2			3	3
3	1	2	0	1	1	1	1	2	2	2			3	3
2	3	1	0	1	1	1	1	2	2	2			3	3
2	2	2	1 or 0	0 1	1	1	2 or 1	1 2						
1	4	1	1 or 0	0 1	1	1	2 or 1	1 2						

When $F_1 > F_3$, then there is an attempt to provide more storage spaces before the third work station. The integer restrictions on S_1 and S_2 may still keep them equal until the difference between F_1 and F_3 grows sufficiently to make them unequal. This can be seen when S = 2. The two buffer spaces are split equally when 2/1/1 is the configuration, but when it is 4/1/1, the optimal use of the two spaces is $S_1 = 0$ and $S_2 = 2$.

In Table 4 depicting output rates, we can see that the general trend is for R to increase, but at a decreasing rate as S is increased. For example, from configuration 2/3/1 of Table 5 and S = 1, 2, 3, and 4, we have marginal % increases in R of 5.66%, 4.14%, 3.25%, and 2.58%, respectively. This decreasing rate is not monotonic, however, as can be seen from configuration 2/1/2 of Table 5. At times, the more efficient split of S is a dominating factor over the decreasing effect of increments in S. This is especially true when $F_1 = F_3$ and S is incrementing from an odd to an even number, where an even split can be best.

Table 4. Output Rates from the Optimal Buffer Space Arrangements for Various Configurations and with Balanced Workloads.

Facility Configurations			Output Rate from Optimal Storage Space Arrangements						
F_1	F_2	F_3	S = 0	S = 1	S = 2	S = 3	S = 4	S = 5	S = 6
1	1	1	.56410	.61333	.67047	.70032	.73402	.75434	.77671
1	2	1	.61513	.65716	.70217	.72788	.75575	.77347	.79253
2	1	1	.59310	.64695	.68751	.71957	.74541	.76687	.78489
3	1	1	.60982	.66609	.69791	.73136			
2	2	1	.63940	.68315	.71666	.74360			
2	1	2	.62618	.66345	.70638	.73063			
1	3	1	.64746	.68452	.72253	.74555			
4	1	1	.62107	.67915	.71000	.73975			
3	2	1	.65392	.69854	.72567	.75342	.77102		.80447
3	1	2	.64410	.68369	.71753	.74298	.76539		.79920
2	3	1	.66834	.70620	.73544	.75933	.77895		.80973
2	2	2	.66572	.69758	.73230	.75330			
1	4	1	.67101	.70342	.73764	.75888			

Table 5. Marginal % Increase in R for Increments in S with F_i Constant and with Balanced Workloads.

Facility Configurations			Marginal % Increase in R					
F_1	F_2	F_3	S = 1	S = 2	S = 3	S = 4	S = 5	S = 6
1	1	1	8.73	9.32	4.45	4.81	2.77	2.97
1	2	1	6.83	6.85	3.66	3.83	2.34	2.46
2	1	1	9.08	6.27	3.29	4.66	2.88	2.35
3	1	1	9.23	4.78	4.79			
2	2	1	6.84	4.91	3.76			
2	1	2	5.95	6.47	3.43			
1	3	1	5.72	5.55	3.19			
4	1	1	9.35	4.54	4.19			
3	2	1	6.82	3.88	3.82	2.34		
3	1	2	6.15	4.95	3.55	3.02		
2	3	1	5.66	4.14	3.25	2.58		
2	2	2	4.79	4.98	2.87			
1	4	1	4.83	4.86	2.88			

From the selected cases in Table 6, it also appears that R increases more slowly as F increases and S is held constant. For example, for S = 0 and configuration 1/1/1, as a facility is added incrementally to the middle station, the marginal % increase in R is decreasing, from 9.05% to 5.26% to 3.64%, respectively.

Table 6. Marginal % Increase in R for Increments in F with S Constant and with Balanced Workloads.

Facility Configurations			Marginal % Increase in R			
F1	F2	F3	S = 0	S = 1	S = 2	S = 3
1	1	1	-	-	-	-
1	2	1	9.05	7.15	4.73	3.94
1	3	1	5.26	4.16	2.90	2.43
1	4	1	3.64	2.76	2.09	1.79
1	1	1	-	-	-	-
2	1	1	5.14	5.48	2.54	2.75
3	1	1	2.82	2.96	1.51	1.64
4	1	1	1.84	1.96	1.73	1.15

4.4 Unbalancing Workloads as Fine Tuning

There is more potential for improving R by unbalancing workloads when N = 3 rather than 2. In particular, unbalancing is applicable for non-optimal configurations. Consider, for instance, the case of F = 6. The optimal configuration of facilities is 1/4/1 when *balancing* workloads occurs, with an output rate of .67101. If, however, the given configuration were 2/3/1, unbalancing workloads is not only better than the balanced workload case for 2/3/1 (.67283 versus .67101), but it is better than the balanced workload R for the 1/4/1 case. This suggests that, in general, the *optimal allocation* of workload when the facilities are nonsymmetrical would be *unbalanced.*

Unbalancing can't always help poor facility design. For instance, when we distribute the six facilities 2/2/2, no amount of unbalancing workloads will make it dominate the 1/4/1 balanced case. The balanced workload allocation for configuration 2/2/2 gives an output rate of .66572. Hence, there seems to be great potential for unbalancing "near optimal" configurations to further improve the output rate. A possible explanation may stem from the fact that facilities may only be moved discretely and facility changes may overstep the optimal R, whereas the workload can be continuously varied.

4.5 Optimal R for Constrained Variables

Table 7 demonstrates the effects of fixing a parameter on the optimal R for fixed resources. In particular, consider F = 6 and S = 4. The values which are in boldface were fixed in each of these cases.

Table 7. Optimal Output Rates When Some Parameters are Fixed.

Case No.	Configuration								Output Rate (R)
	F_1	μ_1	S_1	F_2	μ_2	S_2	F_3	μ_3	
1	1	0.965	2	4	1.078	2	1	0.965	.78355
2	**2**	0.960	2	**3**	1.072	2	**1**	0.975	.78156
3	1	**1.000**	2	4	**1.000**	2	1	**1.000**	.78084
4	4	0.995	**1**	1	1.061	**3**	1	0.950	.76333

In the first case of Table 7, each parameter was allowed to be free. The other three cases fixed one of these parameters at non-optimal values and permitted the other two parameters to be design variables, chosen to maximize the output rate. It is difficult to draw conclusions from the results of Table 7, but it appears that moderate deviations from the optimal values of these parameters do not always significantly effect R. The effect is greatest when an available resource is small and changes in the associated parameters are discrete. It seems that when F and S are fixed, but large, and the μ_i are nearly balanced, then moderately non-optimal configurations will have output rates which are close to the optimal unconstrained arrangement.

4.6 Increasing N from 2 to 3

There seems to be a significant effect on R when the number of stations is increased from 2 to 3. Table 8 shows a small example, where the main change of interest has been to increase the number of stations (and, therefore, the amount of service capacity given by equation (1)) from 2 to 3.

Table 8. Output Rates as a Function of the Number of Stations.

Configuration					Characteristics			Output Rate (R)	% Change in R
F_1	S_1	F_2	S_2	F_3	N	F	S		
1	0	1	-	-	2	2	0	.66667	-
1	0	2	-	-	2	3	0	.71429	+ 7.1
1	0	1	0	1	3	3	0	.56410	**-15.4**
1	1	1	-	-	2	2	1	.75000	-
1	1	2	-	-	2	3	1	.77778	+ 3.7
1	1	1	0	1	3	3	1	.61333	**-18.2**
1	2	1	-	-	2	2	2	.80000	-
1	2	2	-	-	2	3	2	.81818	+ 2.3
1	1	1	1	1	3	3	2	.67047	**-16.2**
1	3	1	-	-	2	2	3	.83333	-
1	3	2	-	-	2	3	3	.84615	+ 1.5
1	2	1	1	1	3	3	3	.70032	**-16.0**

In each case of Table 8, the optimal allocation of μ is used to calculate the optimal production rate. It is expected that further increases in N would cause a further reduction in R. A possible explanation for this is the increased probabilities of blocking and starving for any

workpiece as N is increased. It is anticipated, however, that this effect would decrease as N is increased from higher values.

If one has a choice of whether to add a facility to an existing station or to form another station, then it is always better to add it to one of the existing stations. Some examples of four facilities with various storage sizes show decreases of 15-18% in the output rate by forming a third station, rather than adding a facility to one of the existing two, which increases R. The table is omitted here for space consideration.

5. OBSERVATIONS FOR GENERAL SYSTEMS

From Sections 3 and 4, it is possible to identify many trends that we would expect to continue if the number of stations were allowed to grow beyond three. We don't reiterate each of these, but only the most salient, by stating the general expectations in the form of conjectures and observations.

Conjecture 1. For two stations, if $F_1 \neq F_2$, then the optimal workload allocation is unbalanced by giving most work to the station with fewer facilities.

Conjecture 2. For two stations, output rate is maximized by balancing the facilities per station and the workload per station as much as possible.

Conjecture 3. For any set of parameters that maximizes R, given fixed resources, the *symmetrical allocation property* holds, i.e.,

$$F_i = F_{N+1-i}, \quad i = 1,2,...,N$$
$$\mu_i = \mu_{N+1-i}, \quad i = 1,2,...,N$$
$$S_j = S_{N-j}, \quad j = 1,2,...,N-1.$$

Conjecture 4. For any set of parameters that maximizes R, the *bowl phenomenon* holds, i.e.,

$$F_1 < F_2 < ... < F_{\lfloor \frac{N}{2} \rfloor}$$

and $$\mu_1 < \mu_2 < ... < \mu_{\lfloor \frac{N}{2} \rfloor}.$$

Conjecture 5. A unit increase in the number of work stations decreases R, even if this is accompanied by unit increases in each of the resources, i.e., $R(N; F; \mu; S) > R(N + 1; F + 1; \mu; S + 1)$.

Conjecture 6. As S or $F \rightarrow \infty$, the optimal distribution of workload approaches a balanced situation and $R \rightarrow 1.0$. In addition, for large S and F, the balanced workload line is approximately equal to the optimal unbalanced distribution of workload.

Conjecture 7. If there is a choice between adding one facility or one storage space to a production line, while keeping the distribution of service constant, the expected output rate is *always increased more* by adding a storage space.

The work of Hillier and Boling [1967a, 1967b] began as an empirical study of single-facility stations with several conjectures, the main one being the bowl phenomenon. We have observed a similar pattern even when the stations may have parallel facilities, as have Hillier and So [1989]. It appears that the general principle is to make the stations towards the center of the line more efficient. This can be done by increasing the service rate per facility for the middle stations (that is, decreasing the workload) or by increasing the number of parallel facilities at these center stations. Also, as the number of storage spaces (or, as the number of facilities) increases, blocking and starving occur less frequently and deviations from balanced systems have lessened effects. This is also experienced in the single facility per station case.

There is an improvement in output rate by adding parallel facilities. This is not obvious because parallel stations are not better in M/M/1 queues. Splitting a station into several facilities is counter to previous queueing theory results. One explanation is that parallel facilities gives us a mechanism for protecting against blocking and starving, by providing additional storage.

The assumption of exponential service times is convenient and not realistic, but has a certain robustness. Hillier and So [1989, 1991] conclude that the approximation of an exponential distribution should suffice whenever the coefficient of variation is between .7 and 2.5. The problems of blocking and starving occur because of the variability of service times from station to station. Exponentially distributed service times lead to this high variability and, hence, we would

expect to need unbalancing. Realistic distributions would likely have lower coefficients of variation. Analyzing extreme cases, such as deterministic and exponential distributions, provides us with information on the sensitivity of such parameters.

Some of the results on the optimality of unbalancing configurations and/or workloads are similar and some different from the results of several queueing network studies. In particular, as opposed to the observations in this paper, the results of Stecke and Solberg [1985] indicate that expected production is maximized by unbalancing both the configuration and the workload allocation when $N > 2$. However, another difference is that there is no bowl phenomena observed in the queueing network studies. Some reasons for these differences are that the queueing networks are modeling different systems. In particular, the closed networks of arbitrarily connected queues are modeling job shop types of systems where workpieces have alternative routes. Also, the queueing network models require an adequate buffer in front of each work station to hold all parts that may need to wait.

The problem of designing or improving production facilities is likely not amenable to general solution procedures until we have a better understanding of the effects of the parameters inherent in these systems. One of the aims of this paper is to increase our understanding of some of these effects.

ACKNOWLEDGEMENT

We'd like to acknowledge Robert Robinson of the University of Waterloo for the computational experiments.

BIBLIOGRAPHY

ALTIOK, TAYFUR and STIDHAM, SHALER, Jr., "The Allocation of Interstage Buffer Capacities in Production Lines," IIE Transactions, Vol. 15, No. 4, pp. 292-299 (December 1983).

BARTER, K., "A Queuing Simulator for Determining Optimum Inventory Levels in a Sequential Process," Journal of Industrial Engineering, Vol. 13, pp. 245-252 (1962).

BUZACOTT, JOHN A., "Prediction of the Efficiency of Production Systems without Internal Storage," International Journal of Production Research, Vol. 6, p. 173 (1963).

BUZACOTT, JOHN A., "Automatic Transfer Lines with Buffer Stocks," Journal of Industrial Engineering, Vol. 5, p. 184 (1967).

BUZACOTT, JOHN A., "The Role of Inventory Banks in Flow Line Production Systems," International Journal of Production Research, Vol. 9, No. 4, pp. 425-436 (1971).

CHOW, W., "Buffer Capacity Analysis for Sequential Production Lines with Variable Process Times," International Journal of Production Research, Vol. 25, No. 8, pp. 1183-1196 (1987).

DALLERY, YVES and STECKE, KATHRYN E., "On the Optimal Allocation of Servers and Workloads in Closed Queueing Networks," Operations Research, Vol. 38, No. 4, pp. 694-703 (July-August 1990).

DATTATREYA, E. S., "Tandem Queuing Systems with Blocking," Ph.D. Dissertation, University of California, Berkeley (1978).

FREEMAN, DAVID R. and JUCKER, T. V., "The Line Balancing Problem," Journal of Industrial Engineering, Vol. 18, pp. 361-364 (1967).

FREEMAN, M. C., "The Effect of Breakdowns and Interstage Storage on Production Line Capacity," Journal of Industrial Engineering, Vol. 15, pp. 194-200 (1964).

GERSHWIN, STANLEY B., "An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking," Operations Research, Vol. 35, No. 3, pp. 291-305 (May-June 1987).

HILLIER, FREDERICK S. and BOLING, RONALD W., "The Effect of Some Design Factors on the Efficiency of Production Lines with Variable Operation Times," Journal of Industrial Engineering, Vol. 17, No. 12, pp. 651-658 (1967a).

HILLIER, FREDERICK S. and BOLING, RONALD W., "Finite Queues in Series with Exponential or Erlang Service Times - A Numerical Approach," Operations Research, Vol. 15, No. 2, pp. 286-303 (March-April 1967b).

HILLIER, FREDERICK S. and BOLING, RONALD W., "Optimal Allocation of Work in Production Line Systems with Variable Operation Times," Technical Report No. 16, Department of Operations Research, Stanford University (1972).

HILLIER, FREDERICK S., BOLING, RONALD W. and SO, KUT C., "Toward Characterizing the Optimal Allocation of Storage Space in Production Line Systems with Variable Operation Times," Technical Report, Department of Operations Research, Stanford University (March 1990).

HILLIER, FREDERICK S. and SO, KUT C., "The Assignment of Extra Servers to Stations in Tandem Queueing Systems with Small or No Buffers," Performance Evaluation, Vol. 10, pp. 219-231 (1989).

HILLIER, FREDERICK S. and SO, KUT C., "The Effect of the Coefficient of Variation of Operation Times on the Allocation of Storage Space in Production Line Systems," IIE Transactions (March 1991a).

HILLIER, FREDERICK S. and SO, KUT C., "On the Simultaneous Optimization of Server and Work Allocations in Production Line Systems with Variable Processing Times," GSM Working Paper #DS91005, Graduate School of Management, University of California, Irvine (February 1991b).

MAGAZINE, MICHAEL J. and SILVER, G. L., "Heuristics for Unbalancing Serial Production Systems," International Journal of Production Research, Vol. 16, No. 6, pp. 169-181 (1978).

MUTH, E. J., "The Production Rate of a Series of Work Stations with Variable Service Times," International Journal of Production Research, Vol. 11, No. 2, pp. 155-169 (1973).

PAYNE, S., SLACK, N., and WILD, R., "A Note on the Operating Characteristics of Balanced and Unbalanced Production Flow Lines," International Journal of Production Research, Vol. 10, No. 1, pp. 93-98 (1972).

RAO, NORI P., "A Generalization of the 'Bowl Phenomenon' in Series Production Systems," International Journal of Production Research, Vol. 14, No. 4, pp. 437-443 (1976).

SHANTHIKUMAR, J. GEORGE and YAO, DAVID D. W., "The Effect of Increasing Service Rates in a Closed Queueing Network," Journal of Applied Probability, Vol. 23, pp. 474-483 (1986).

SHANTHIKUMAR, J. GEORGE and YAO, DAVID D. W., "Optimal Server Allocation in a System of Multiserver Stations," Management Science, Vol. 33, No. 9, pp. 1173-1180 (September 1987).

SHANTHIKUMAR, J. GEORGE and YAO, DAVID D. W., "On Server Allocation in Multiple Center Manufacturing Systems," Operations Research, Vol. 36, No. 2, pp. 333-342 (March-April 1988).

SHANTHIKUMAR, J. GEORGE and YAO, DAVID D. W., "Optimal Buffer Allocation in a Multicell System," International Journal of Flexible Manufacturing Systems, Vol. 1, No. 4, pp. 347-356 (September 1989).

SO, KUT C., "Allocating Buffer Storages in a Flexible Manufacturing System," International Journal of Flexible Manufacturing Systems, Vol. 1, No. 3, pp. 323-337 (June 1989).

STECKE, KATHRYN E., "On the Nonconcavity of Throughput in Certain Closed Queueing Networks," Performance Evaluation, Vol. 6, No. 3, pp. 293-305 (August 1986).

STECKE, KATHRYN E. and KIM, ILYONG, "Performance Evaluation for Systems of Pooled Machines of Unequal Sizes: Unbalancing Versus Balancing," European Journal of Operational Research, Vol. 42, No. 1, pp. 22-38 (September 1989).

STECKE, KATHRYN E. and MORIN, THOMAS L., "The Optimality of Balancing Workloads in Certain Types of Flexible Manufacturing Systems," European Journal of Operational Research, Vol. 20, No. 7, pp. 68-82 (1985).

STECKE, KATHRYN E. and SOLBERG, JAMES J., "The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Queueing Networks of Multiserver Queues," Operations Research, Vol. 33, No. 4, pp. 882-910 (July-August, 1985).

WILD, R., BUXEY, G. M., and SLACK, N. D. C., "Production Flow Line System Design - A Review," American Institute of Industrial Engineers Transactions, Vol. 5, p. 37 (1973).

WILD, R. and CARNALL, C. A., "The Location of Variable Work Stations and the Performance of Flow Lines," International Journal of Production Research, Vol. 14, No. 6, pp. 703-710 (1976).

YAMASHINA, H. and OKAMURA, K., "Analysis of In-Process Buffers for Multi-Stage Transfer Line Systems," International Journal of Production Research, Vol. 21, No. 2, pp. 183-195 (March 1983).

YAMAZAKI, G., KAWASHIMA, T. and SAKASEGAWA, H., "Reversibility of Tandem Blocking Queueing Systems," Management Science, Vol. 31, pp. 78-83 (1985).