Optimal allocation of arrivals to a collection of parallel workstations

Madjid Tavana and Jack Rappaport
Management Department, La Salle University, Philadelphia, USA

Received July 1996

Introduction

Increasing global competition and rapid technological changes have forced many firms to take a serious look at their manufacturing function and improve their plant efficiency. Significant improvements can be made by appropriately managing a production facility, especially by reducing the waiting times of the work-in-process. Optimal control and distribution of arrivals has been the subject of research by several authors (Davis, 1977; Ghoneim and Stidham, 1985; Johnson and Stidham, 1980; Stidham, 1978). Most of these studies focus on reducing the amount of waiting time while maintaining throughput in a collection of service facilities, either parallel or serial. Farrell (1976) and Stidham (1985) studied the problem of routeing customers to one of two queues, but without the option of accepting or rejecting an arriving customer. Weber (1978) found that in parallel queues a control strategy that assigns each arrival to the shortest queue minimizes the probability that customers complete service by a given time. Davis (1977) also studied the case of two identical parallel servers with the possibility of rejecting some arriving customers.

Optimal control of arrivals has also been studied in serial queues. Lazar (1983) considered the control of arrivals to a network of queues with the objective of maximizing throughput subject to a response time constraint. Ghoneim and Stidham (1985) studied two queues in series where customers are accepted or rejected with the objective of maximizing the discounted expected benefits over the horizon. Shioyama (1991) demonstrated the optimal control in a system consisting of two stages: the first stage with one server and the second stage with two servers. According to this optimal control, customers are selected to be served in order to minimize expected cost per time unit.

Many researchers have studied optimal allocation of servers in multiserver queues. Dyer and Proll (1977) have proved a conjecture of Rolfe (1971) concerning the allocation of a fixed number of servers to a multiple facility service system each with a different arrival. They show that the expected

The authors wish to thank the editors, the anonymous reviewers, and Dr Prafulla Joglekar, the Lindback Professor of Production and Operations Management at La Salle University, for their invaluable commentary and constructive suggestions.

International Journal of Operations & Production Management, Vol. 17 No. 3, 1997, pp. 305-325.

waiting time formula for *M/M/c* queue is a convex function of *c*. Their result was then used to derive the optimal number of servers minimizing the total expected waiting time in a multiple facility system. Shanthikumar and Yao (1987; 1988) studied the optimal server allocation in a system of multiserver stations with a fixed buffer capacity and an arrival rate that was dependent on the number of jobs in the station.

Hillier and Boling (1966; 1977; 1979) studied the optimal design of unpaced production lines. Their model is based on the classical queueing system with finite queues in series and an infinite queue before the first server. The decision variables are the mean service times allocated at each facility while the objective function is to maximize the mean output rate. Dallery and Stecke (1990), Stecke and Kim (1989), Stecke and Morin (1985), and Stecke and Solberg (1985) considered allocating workloads and machine group sizes in closed queueing networks. Closed queueing networks are useful in representing a wide variety of real manufacturing systems. The problem is to partition *m* servers into *g* groups according to the maximum expected production. The results suggest that the more pooling the better and the more unbalanced configuration the larger the maximum expected production. Moreover, in a system with unequal group sizes, the expected production rate is maximized by assigning a specific unbalanced workload per server to each group.

The queueing system studied here is different from the previous research. Our model does not factor in any cost or reward functions, nor does it require any observation or measurement at the service process (such as routeing a customer to the shorter queue length). In addition, our model considers optimal allocation of arrivals as opposed to the optimal allocation of servers studied by many researchers. In this paper, we present an aggregate optimization model for a collection of *M/M/1* queueing systems. We show that one cannot separately optimize the objective function for each service facility independently since the control of a service facility is affected by and affects the control of every other facility in the system. For example, one can minimize the total average queue size of a collection of *M/M/1* queues aggregated over all queues when the total arrival rate is equal to a constant and the service rate of each facility is given[1]. The problem as it is shown in Figure 1 is to find the optimal allocation of total arrivals among service facilities using the allocation

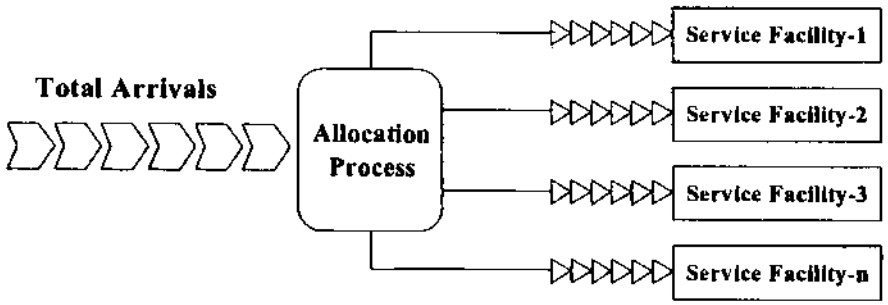

Figure 1.
A graphical representation of the queueing system

process described here. In the second section we present the optimization model followed by a series of sample cases for exponential and general service distributions in sections three and four. Then, in section five we illustrate the effectiveness of the model in a practical application. The final section presents the conclusion and future research directions.

Model

Let Q denote a collection of n *M/M/1* queueing systems, and define the following notations:

- n = number of *M/M/1* queueing systems belonging to *Q;*
- λ_i = arrival rate of the i^{th} system;
- μ_i = service rate of the i^{th} system;
- L_{si} = the average number of customers in the i^{th} system;
- W_{si} = the average time customers spend in the i^{th} system;
- L_{qi} = the average number of customers waiting for service in the i^{th} system;
- W_{qi} = the average time customers wait in line in the i^{th} system.

First we derive the model for L_{si} and W_{si}, considering the following constrained optimization problem:

$$\text{Minimize } \sum_{i=1}^{n} L_{s_i} = \sum_{i=1}^{n} \frac{\left(\frac{\lambda_i}{\mu_i}\right)}{\left(1 - \frac{\lambda_i}{\mu_i}\right)} \tag{1}$$

Subject to:

$$\sum_{i=1}^{n} \lambda i = M$$

$$\text{Where } M \text{ is constant such that } M < \sum_{i=1}^{n} \mu_i \tag{2}$$

λ_is are the decision variables and μs are assumed to be constant. This optimization problem can be solved by Lagrange's method. Assuming that m is the Lagrange multiplier, then the Lagrangean function is defined as:

$$L(\lambda_1, \lambda_2, \ldots \lambda_n, m) = \sum_{i=1}^{n} \frac{\left(\frac{\lambda_i}{\mu_i}\right)}{\left(1 - \frac{\lambda_i}{\mu_i}\right)} - m\left(\sum_{i=1}^{n} \lambda_i - M\right) \tag{3}$$

The solution to the optimization function is found by setting

$$\frac{\partial L}{\partial \lambda_i} = 0 \tag{4}$$

and

$$\frac{\partial L}{m} = 0 \tag{5}$$

This yields the following equation:

$$\frac{\partial L}{\partial \lambda_i} = \frac{\left(1 - \frac{\lambda_i}{\mu_i}\right)\left(\frac{1}{\mu_i}\right) - \left(\frac{\lambda_i}{\mu_i}\right)\left(-\frac{1}{\mu_i}\right)}{\left(1 - \frac{\lambda_i}{\mu_i}\right)^2} - M = 0 \tag{6}$$

The solution to the above equation is:

$$\lambda_i = \mu_i\left[1 - \frac{1}{\sqrt{\mu_i m}}\right] \tag{7}$$

Combining equations (7) and (2) we obtain

$$\sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \frac{\mu_i}{\sqrt{\mu_i m}} = M \tag{8}$$

Algebraically we then derive the following equation:

$$m = \left(\frac{\sum_{i=1}^{n} \sqrt{\mu_i}}{\sum_{i=1}^{n} \mu_i - M}\right)^2 \tag{9}$$

For W_{s_i} the objective function and the optimal solution is the same as L_{s_i} since

$$\frac{\sum_{i=1}^{n} \lambda_i W_{s_i}}{\sum_{i=1}^{n} \lambda_i} = \frac{\sum_{i=1}^{n} L_{s_i}}{M} \tag{10}$$

Next, we derive the model for L_{qi} and W_{qi} considering the following constrained optimization problem:

$$\text{Minimize} \sum_{i=1}^{n} L_{q_i} = \sum_{i=1}^{n} \frac{\left(\frac{\lambda_i}{\mu_i}\right)^2}{\left(1 - \frac{\lambda_i}{\mu_i}\right)} \tag{11}$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i = M \tag{12}$$

Solving equations (4) and (5) as before, we obtain

$$\lambda_i = \mu_i \left[1 - \frac{1}{\sqrt{1 + m\mu_i}}\right] \tag{13}$$

and

$$\sum_{i=1}^{n} \mu_i \left[1 - \frac{1}{\sqrt{1 + m\mu_i}}\right] = M \tag{14}$$

For W_{qi}, the objective function and the optimal solution is the same as L_{qi} since

$$\frac{\sum_{i=1}^{n} \lambda_i W_{q_i}}{\sum_{i=1}^{n} \lambda_i} = \frac{\sum_{i=1}^{n} L_{q_i}}{M} \tag{15}$$

Sample cases for exponential service distribution

In this section, we demonstrate the effectiveness of our model by solving selected sample cases for different values of n (the number of service facilities) and λ_is (the arrival rates). We developed a computer program that uses our model to calculate the optimal values of λ_is for various objective functions including aggregate L_s, W_s, L_q, and W_q. Tables I to V show the resulting performance measures for these queueing systems.

In these cases (n = 1, 2, …, 6), the optimal allocation of arrivals decreases the aggregate L_s, W_s, L_q and W_q, compared to the naive allocation of arrivals.

One factor to note is that the overall improvement in the aggregate waiting is achieved by increasing the utilization of facilities with higher service rates while decreasing the utilization of facilities with lower service rates. Apparently the increase in waiting at the faster service facilities is offset by a larger decrease in waiting of the slower service facilities. This can be illustrated

Naïve		Optimal		Performance measures				Queue's performance improvement (percentage)
λ_1	λ_2	λ_1	λ_2	L_s	W_s	L_q	W_q	
1.80	0.90			18.00	6.67	16.20	6.00	
		1.82	0.88	17.44				3.11
		1.82	0.88		6.46			3.15
		1.82	0.88			15.65		3.40
		1.82	0.88				5.80	3.33
1.60	0.80			8.00	3.33	6.40	2.67	
		1.65	0.75	7.71				3.63
		1.65	0.75		3.21			3.60
		1.65	0.75			6.14		4.06
		1.65	0.75				2.57	3.75
1.40	0.70			4.66	2.22	3.26	1.56	
		1.47	0.63	4.48				3.86
		1.47	0.63		2.13			4.05
		1.47	0.63			3.11		4.60
		1.47	0.63				1.49	4.49
1.20	0.60			3.00	1.67	1.80	1.00	
		1.28	0.52	2.86				4.67
		1.28	0.52		1.59			4.80
		1.28	0.52			1.70		5.56
		1.28	0.52				0.95	5.50

Table I.
Results for the naïve and optimal approaches for sample λ_is, $n = 2$ and $(\mu_1, \mu_2) = (2,1)$

mathematically for the case of two facilities by comparing the marginal change in waiting for the two facilities assuming that $\mu_1 > \mu_2$ and that the initial utilization at both facilities is

$$\rho_0 = \frac{\lambda_1}{\mu_1} = \frac{\lambda_2}{\mu_2}.$$

Then

$$\frac{\partial L_{s_i}}{\partial \lambda_i} = \frac{1}{\mu_i (1 - \rho_i)^2}$$

At the point of equal utilization $\rho_i = \rho_0$, and

$$\frac{\partial L_{s_i}}{\partial \lambda_i} \Bigg|_{\rho_i = \rho_0} = \frac{1}{\mu_i (1 - \rho_0)^2}$$

Naïve			Optimal			Performance measures				Queue's performance improvement (percentage)
λ_1	λ_2	λ_3	λ_1	λ_2	λ_3	L_s	W_s	L_q	W_q	
2.70	1.80	0.90				27.00	5.00	24.30	4.50	
			2.75	1.80	0.85	25.65				5.00
			2.75	1.80	0.85		4.74			5.20
			2.75	1.80	0.85			23.00		5.42
			2.75	1.80	0.85				4.26	5.33
2.40	1.60	0.80				12.00	2.50	9.60	2.00	
			2.50	1.59	0.71	11.33				5.58
			2.50	1.59	0.71		2.36			5.60
			2.49	1.59	0.72			9.00		6.25
			2.49	1.59	0.72				1.88	6.00
2.10	1.40	0.70				6.99	1.67	4.89	1.17	
			2.25	1.39	0.56	6.55				6.30
			2.25	1.39	0.56		1.56			6.59
			2.23	1.38	0.59			4.54		7.16
			2.23	1.38	0.59				1.09	6.84
1.80	1.20	0.60				4.50	1.25	2.70	0.75	
			2.00	1.18	0.42	4.16				7.56
			2.00	1.18	0.42		1.16			7.20
			1.96	1.18	0.46			2.47		8.52
			1.96	1.18	0.46				0.69	8.00

Table II.
Results for the naïve and optimal approaches for sample λ_is, $n = 3$ and $(\mu_1, \mu_2, \mu_3) = (3,2,1)$

Therefore,

$$\frac{\partial L_{s1}}{\partial \lambda_1} < \frac{\partial L_{s_2}}{\partial \lambda_2} \text{ when } \mu_1 > \mu_2.$$

This shows that the marginal change in the average number waiting in the system with respect to λ_i is greater for the slower service facility. Thus reducing the utilization of the slower facility, while increasing the utilization of the faster facility, will result in an overall decrease in aggregate waiting.

Further analysis of sample cases shows that a larger range for the service rates results in greater percentage improvements of our aggregate objectives. For example a two-facility system ($n = 2$) with $\mu_1 = 2$, $\mu_2 = 1$, and $\rho = 0.90$ results in a 3.11 per cent decrease of aggregate L_s as compared with a 7.67 per cent improvement for a six-facility system ($n = 6$) with $\mu_1 = 6$, $\mu_2 = 5$, $\mu_3 = 4$, $\mu_4 = 3$, $\mu_5 = 2$, $\mu_6 = 1$, and $\rho = 0.90$ (see Figure 2).

Table III.
Results for the naïve and optimal approaches for sample $\lambda_i s$, $n = 4$ and $(\mu_1, \mu_2, \mu_3 \mu_4) = (4,3,2,1)$

Naïve				Optimal				Performance measures				Queue's Performance improvement (percentage)
λ_1	λ_2	λ_3	λ_4	λ_1	λ_2	λ_3	λ_4	L_s	W_s	L_q	W_q	
3.60	2.70	1.80	0.90					36.00	4.00	32.40	3.60	
				3.67	2.72	1.77	0.84	33.78				6.17
				3.67	2.72	1.77	0.84		3.75			6.25
				3.67	2.72	1.77	0.84			30.22		6.73
				3.67	2.72	1.77	0.84				3.36	6.67
3.20	2.40	1.60	0.80					16.00	2.00	12.80	1.60	
				3.35	2.44	1.54	0.67	14.89				6.94
				3.35	2.44	1.54	0.67		1.86			7.00
				3.34	2.43	1.54	0.69			11.80		7.81
				3.34	2.43	1.54	0.69				1.48	7.50
2.80	2.10	1.40	0.70					9.32	1.33	6.52	0.93	
				3.02	2.16	1.31	0.51	8.60				7.73
				3.02	2.16	1.31	0.51		1.23			7.52
				3.00	2.14	1.32	0.54			5.94		8.90
				3.00	2.14	1.32	0.54				0.85	8.60
2.40	1.80	1.20	0.60					6.00	1.00	3.60	0.60	
				2.70	1.87	1.08	0.35	5.44				9.33
				2.70	1.87	1.08	0.35		0.91			9.00
				2.64	1.85	1.09	0.42			3.24		10.00
				2.64	1.85	1.09	0.42				0.55	9.17

Naïve						Optimal					Performance measures				Queue's Performance improvement (percentage)
λ_1	λ_2	λ_3	λ_4	λ_5	λ_1	λ_2	λ_3	λ_4	λ_5	L_s	W_s	L_q	W_q	
4.50	3.60	2.70	1.80	0.90						45.00	3.33	40.50	3.00	
					4.60	3.64	2.69	1.75	0.82	41.85				7.00
					4.60	3.64	2.69	1.75	0.82		3.09			7.21
					4.60	3.64	2.69	1.75	0.82			37.42		7.60
					4.60	3.64	2.69	1.75	0.82				2.78	7.33
4.00	3.20	2.40	1.60	0.80						20.00	1.67	16.00	1.33	
					4.20	3.28	2.38	1.50	0.64	18.42				7.90
					4.20	3.28	2.38	1.50	0.64		1.53			8.38
					4.19	3.28	2.38	1.50	0.65			14.57		8.94
					4.19	3.28	2.38	1.50	0.65				1.22	8.65
3.50	2.80	2.10	1.40	0.70						11.65	1.11	8.15	0.78	
					3.80	2.93	2.07	1.24	0.46	10.62				8.84
					3.80	2.93	2.07	1.24	0.46		1.01			9.00
					3.77	2.91	2.06	1.25	0.51			7.34		9.94
					3.77	2.91	2.06	1.25	0.51				0.70	9.62
3.00	2.40	1.80	1.20	0.60						7.50	0.83	4.50	0.50	
					3.40	2.57	1.76	0.99	0.28	6.71				10.53
					3.40	2.57	1.76	0.99	0.28		0.74			10.84
					3.33	2.52	1.75	1.02	0.38			3.99		11.33
					3.33	2.52	1.75	1.02	0.38				0.44	11.00

Table IV.
Results for the naïve and optimal approaches for sample λ_is, $n = 5$ and $(\mu_1, \mu_2, \mu_3\, \mu_4\, \mu_5) = (5,4,3,2,1)$

Table V.
Results for the naïve and optimal approaches for sample λ_is, $n = 6$ and (μ_1, μ_2, μ_3, μ_4, μ_5, μ_6) = (6,5,4,3,2,1)

Naïve						Optimal						Performance measures				Queue's performance improvement (percentage)
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	L_s	W_s	L_q	W_q	
5.40	4.50	3.60	2.70	1.80	0.90							54.00	2.86	48.60	2.57	
						5.52	4.57	3.61	2.66	1.73	0.81	49.86				7.67
						5.52	4.57	3.61	2.66	1.73	0.81		2.63			8.04
						5.52	4.57	3.61	2.66	1.73	0.81			44.59		8.25
						5.52	4.57	3.61	2.66	1.73	0.81				2.36	8.17
4.80	4.00	3.20	2.40	1.60	0.80							24.00	1.43	19.20	1.14	
						5.05	4.13	3.23	2.33	1.45	0.61	21.95				8.54
						5.05	4.13	3.23	2.33	1.45	0.61		1.30			9.09
						5.04	4.12	3.22	2.33	1.46	0.63			17.36		9.58
						5.04	4.12	3.22	2.33	1.46	0.63				1.03	9.21
4.20	3.50	2.80	2.10	1.40	0.70							13.98	0.95	9.78	0.67	
						4.57	3.70	2.84	1.99	1.18	0.42	12.62				9.73
						4.57	3.70	2.84	1.99	1.18	0.42		0.86			9.90
						4.54	3.67	2.82	1.99	1.20	0.48			8.73		10.74
						4.54	3.67	2.82	1.99	1.20	0.48				0.60	10.45
3.60	3.00	2.40	1.80	1.20	0.60							9.00	0.71		0.43	
						4.10	3.27	2.45	1.66	0.90	0.22	7.96		5.40		11.56
						4.10	3.27	2.45	1.66	0.90	0.22		0.63			11.97
						4.01	3.20	2.42	1.66	0.96	0.35			4.74		12.22
						4.01	3.20	2.42	1.66	0.96	0.35				0.38	12.09

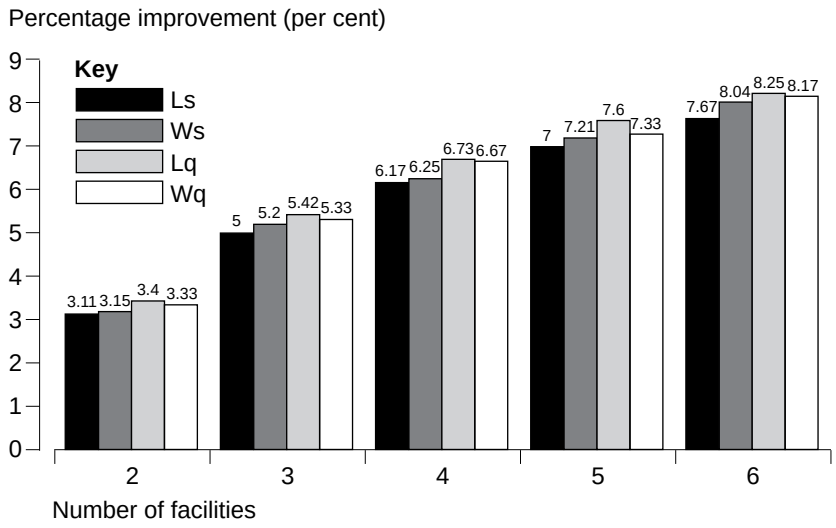

Figure 2.
Queue's performance improvement for sample number of facilities where $\rho = 0.90$

This general result can be seen mathematically by examining a two-facility system, with $\mu_1 > \mu_2$ and $\rho_i = \rho_0$. The net marginal decrease in waiting time with respect to a change in λ_i is equal to

$$\frac{\partial L_{s_2}}{\partial \lambda_2} - \frac{\partial L_{s_1}}{\partial \lambda_1} = \frac{1}{\mu_2(1-\rho_0)^2} - \frac{1}{\mu_1(1-\rho_0)^2} = \frac{1}{(1-\rho_0)^2}\left[\frac{1}{\mu_2} - \frac{1}{\mu_1}\right] \qquad (17)$$

Clearly, a greater difference between $\frac{1}{\mu_2}$ and $\frac{1}{\mu_1}$ results in a greater marginal improvement in aggregate waiting.

Another phenomenon shown in the tables and Figure 3 is that greater percentage improvement and changes in arrival rates result from lower utilized systems.

Using a two-facility system as an example we derive the following mathematically:

Let $\mu_1 > \mu_2$ and initial $\rho_0 = \frac{\lambda_1}{\mu_1} = \frac{\lambda_2}{\mu_2}$.

Now assuming that λ_1 is increased marginally by the amount of ε and λ_2 is decreased by the same amount, λ_1 will be increased relative to λ_2 as long as

$$\frac{\partial L_{s_1}}{\partial \lambda_1} < \frac{\partial L_{s_2}}{\partial \lambda_2},$$

i.e. the marginal increase in the waiting of the first facility resulting from increased λ_1 is offset by a marginal decrease in waiting of the second facility from decreased λ_2. By solving for ε when

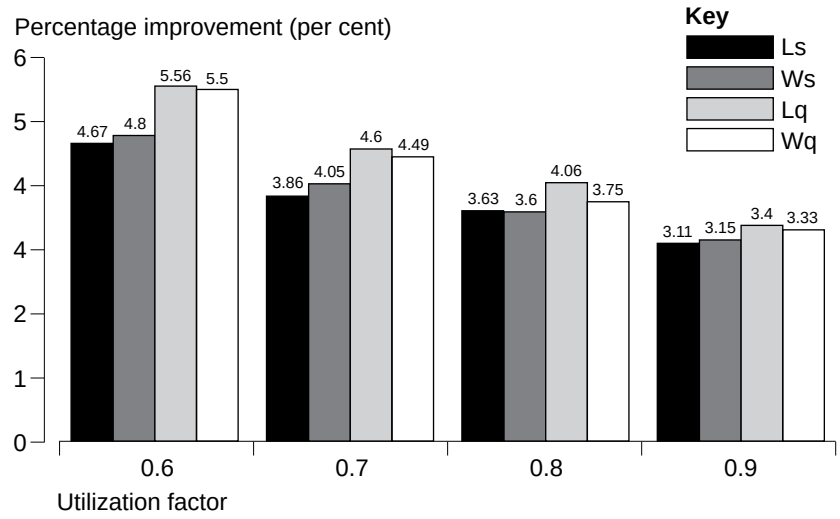

Figure 3.
Queue's performance improvement for sample utilization factors where $n = 2$ and $(\mu_1, \mu_2) = (1,2)$

$$\frac{\partial L_{s_1}}{\partial \lambda_1}\Big|_{\lambda_1 + \varepsilon} = \frac{\partial L_{s_2}}{\partial \lambda_2}\Big|_{\lambda_2 - \varepsilon} \tag{18}$$

we obtain

$$\frac{1}{\mu_1\left(1 - \dfrac{\lambda_1 + \varepsilon}{\mu_1}\right)^2} = \frac{1}{\mu_2\left(1 - \dfrac{\lambda_2 - \varepsilon}{\mu_2}\right)^2}$$

or algebraically we obtain

$$\varepsilon = \left[\frac{(1 - \rho_0)\left[\sqrt{\dfrac{\mu_1}{\mu_2}} - 1\right]}{\left[\dfrac{1}{\mu_2} + \dfrac{1}{\sqrt{\mu_1\mu_2}}\right]}\right] \tag{19}$$

ε represents the largest change in λ_1 and λ_2 that can occur while still providing for marginal improvement in aggregate waiting. Thus ε is greater for lower ρ_0, implying that for lower utilized systems there is a greater amount by which λ_i is changed in order to optimize the aggregate waiting.

Sample cases for general service distribution

Next, we have investigated the results of the model for the general service distributions using the Polleczek-Khintchine formula (Taha, 1982). Let us consider

the same model for the case of other distributions for the service process. The Pollaczek-Khintchine formula gives the value for L_{qi} the single server queueing system for the case of Poisson arrivals and a general service distribution. It is given as follows:

$$L_{q_i} = \frac{\lambda_i^2\left((E_i(t))^2 + \sigma^2_{s_i}\right)}{2(1 - \lambda_i E_i(t))} \tag{20}$$

where

L_{qi} = the average number of customers waiting for service in the i^{th} system;
λ_i = arrival rate of the i^{th} system;
L_{si} = the average number of customers in the i^{th} system;
$E_i(t)$ = mean service time of the i^{th} system;
$\sigma^2_{s_i}$ = variance of the i^{th} system.

This formula can also be written as follows

$$L_{q_i} = \frac{\lambda_i^2 (E_i(t))^2 [1 + k_i]}{2(1 - \lambda_i E_i(t))} \tag{21}$$

where $k_i = \dfrac{\sigma^2_{s_i}}{(E_i(t))^2}$

Note that k_i is the square of the coefficient of variation of the service time distribution. Also note that

$$L_{si} = \lambda_i E_i(t) + L_{qi}$$

where

L_{si} = average number of customers in the i^{th} system.

Generalizing our model to the general service distribution case, our problem is as follows

Minimize $\sum L_{q_i}$

subject to $\sum \lambda_i = M$

as well as

Minimize $\sum L_{s_i}$

subject to $\sum \lambda_i = M$

Note that since the solution to the case of minimizing the sum of ΣW_{qi} is identical to ΣL_{qi} and minimizing ΣW_{si} is identical to minimizing ΣL_{si} they will not be considered here.

Table VI gives the results for the case of $n = 2$, $E_1(t) = 0.5$ and $E_2(t) = 1$. This case corresponds to the case of exponential service time given in Table I where

Variation measures[a]		Naïve				Optimal				Queue's performance improvements (percentage)	
		Arrival rate		performance measures		Arrival rate		Performance measures			
K_1	K_2	λ_1	λ_2	ΣL_{si}	ΣL_{qi}	λ_1	λ_2	ΣL_{si}	ΣL_{qi}	ΣL_{si}	ΣL_{qi}
0.1	0.1	1.80	0.90	10.71	8.91	1.82	0.88	10.40	8.61	2.89	3.36
0.5	0.5	1.80	0.90	13.95	12.15	1.82	0.88	13.53	11.74	3.01	3.37
1.0	1.0	1.80	0.90	18.00	16.20	1.82	0.88	17.44	15.65	3.11	3.39
2.0	2.0	1.80	0.90	26.10	24.30	1.82	0.88	25.27	23.48	3.18	3.37
5.0	5.0	1.80	0.90	50.40	48.60	1.82	0.88	48.75	46.96	3.27	3.37
10.0	10.0	1.80	0.90	90.90	89.10	1.82	0.88	87.89	86.10	3.31	3.36
0.1	10.0	1.80	0.90	50.80	49.00	1.76	0.94	45.40	43.59	10.61	11.04
0.5	5.0	1.80	0.90	32.17	30.37	1.78	0.92	31.34	29.53	2.59	2.78
0.1	5.0	1.80	0.90	30.55	28.75	1.77	0.93	29.04	27.22	4.95	5.33
0.5	2.0	1.80	0.90	30.55	28.75	1.89	0.81	21.04	19.29	31.14	32.91
10.0	0.1	1.80	0.90	20.02	18.22	1.80	0.90	20.02	18.22	0.00	0.00
5.0	0.5	1.80	0.90	50.80	49.00	1.91	0.79	29.23	27.49	42.46	43.90
5.0	0.1	1.80	0.90	32.17	30.37	1.87	0.83	24.00	22.24	25.40	26.78
2.0	0.5	1.80	0.90	20.25	18.22	1.85	0.85	17.55	15.78	12.35	13.41

Note: [a]The results are presented for three sets of cases including: $k_1 = k_2$, $k_1 < k_2$ and $k_1 > k_2$

Table VI.
Results for the naïve and optimal approaches using general queue time for $n = 2$, $E(t_1) = 0.5$, $E(t_2) = 1$, $\rho = 0.90$, and different values for k_1 and k_2

$\mu_1 = 2$ and $\mu_2 = 1$. Notice that the resulting solution is identical to the one for exponential service times when $k_1 = k_2$. The percentage improvement in the objective functions are roughly equivalent to the percentage improvements in the exponential case. However, note that the percentage improvement increases slightly as k_1 and k_2 increase. The effectiveness of the model increases slightly as the coefficient of variation of the service time distribution increases.

The results are different from the exponential service case when there are different k values for each server. Generally speaking the percentage improvement of the model over the naive solution increases as the difference between the k values increases. For our example, the percentage improvement is greatest for cases where there is a large difference in the k values and the higher k value is associated with the larger mean service time and the smaller k is associated with the smaller mean service time. For example in the case of $k_1 = 10$, associated with a mean service time of 1, and $k_2 = 0.1$, associated with a mean service time of 0.5, the percentage improvement is 43.9 per cent when minimizing the sum of the L_{qi}.

We have also simulated a few cases assuming gamma distributions for both the arrival and service rates. We used the same case as above, $n = 2$, $E_1(t) = 1$, $E_2(t) = 2$ and $M = 2.7$. Using a gamma distribution for the interarrival time with various k values ($k = 0.1$, 0.5, 2, 3 and 5) as well as the k values for the service distributions, we obtained results that were the same as above.

Generally speaking, we can say that the model is applicable for the case of the general service distributions, but it is sensitive to the variance and coefficient of variation of these distributions. When the coefficients of variation of the different service distributions are the same, the percentage improvement of the model over the naive solution is the same as in the exponential service case. When there is a large spread in the coefficient of variations, and this spread is applied so that the larger coefficient of variation is applied to the larger mean, then the percentage improvement of the model can be quite substantial as shown in Table VI.

Application problem

Technocraft[2] is the largest manufacturer of engine cylinder blocks in North America. In recent years, the company has experienced a dramatic increase in sales, which prompted management to invest heavily in automation. Figure 4 represents the casting operation layout at Technocraft.

As shown; the casting division is a complex of ten departments located in seven buildings including the pattern and sand shops; as well as the moulding; administrative; fully-automated casting; semi-automated casting; grinding; and quality control, storage and shipping departments. The cylinder block production line begins with a pattern which is made in the pattern shop. Then, a specially mixed sand (made in the sand shop) is moulded around the pattern in the moulding department. When the pattern is removed, the resulting sand

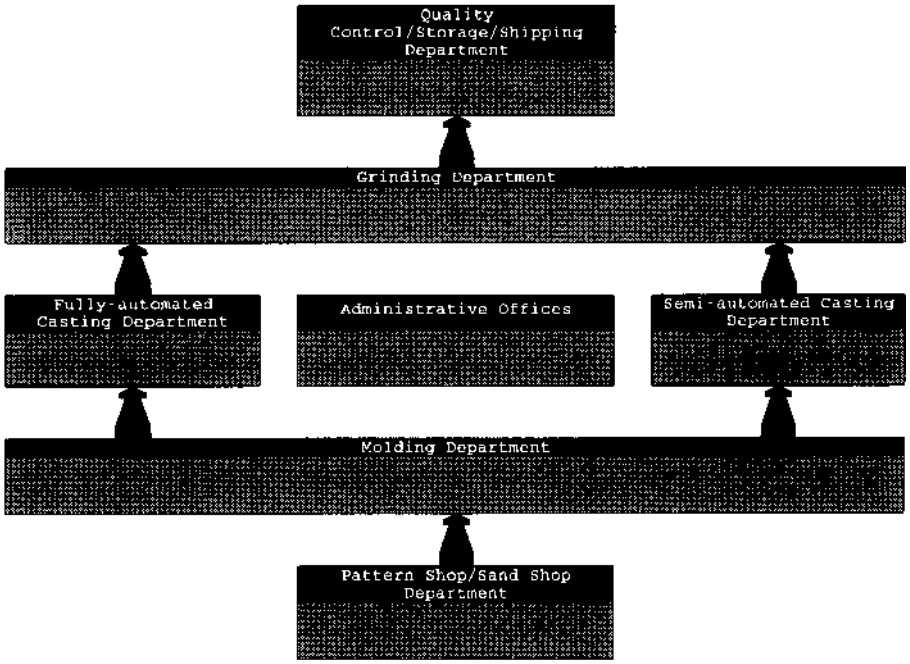

Figure 4.
Technocraft's casting operation layout

moulds form a negative image of the cylinder block. Next, the moulds are distributed among the fully-automated and semi-automated casting departments[3] located in different buildings[4], and separated by the administrative building. During the casting, molten iron is poured into the moulds and allowed to cool. When the iron is solidified, the moulds are moved to the grinding building, where the rough cylinder blocks are subject to grinding. Next, the blocks are inspected for manufacturing defects in the quality control department and if they pass, cylinder blocks are moved to storage to be shipped to the assembly operations. The two casting departments (fully-automated and semi-automated) each had a different service time, since queues formed in each building separately, and the production manager realized that multiple channel queueing models were not applicable to the existing situation. Therefore, he decided to form a dispatching centre within the moulding department and distribute the moulds among the two casting departments according to their service rates. In other words, he asked the dispatching centre to dispatch four moulds (192 per hour) to the fully-automated department for every mould (48 per hour) sent to the semi-automated department[5]. Although this naive distribution seemed logical, casting operations was still experiencing a high waiting time and a large number of cylinder blocks in the queue at a given time.

Using the $M/M/1$ model and hourly data, we learn that the naive allocation of arrivals to the two service facilities would result in the performance measures given in Table VII. As it is shown, the total L_s is 8.00 cylinder blocks (4.00 plus 4.00), the total weighted W_s is 0.0333 hours (0.0208 plus 0.0833), the total L_q is 6.40 cylinder blocks (3.20 plus 3.20), and the total weighted W_q is 0.0267 hours (0.0167 plus 0.0667) for the fully-automated and semi-automated casting departments respectively.

Next we used our model described earlier to find the optimal number of arrivals to be allocated to the two casting departments. Using the Lagrange multiplier, the total L_s, W_s, L_q and W_q are minimized when the allocation of arrivals is changed from the current (naive) 4 : 1 ratio (192 versus 48 cylinder

	Allocation approach						
	Naïve			Optimal			
	Fully-automated casting	Semi-automated casting	Total casting	Fully-automated casting	Semi-automated casting	Total casting	Total casting percentage improvement
λ_i	192	48	240	200	40	240	–
μ_i	240	60	300	240	60	300	–
L_s	4.00	4.00	8.00	5.00	2.00	7.00	12.50
W_s	0.0208	0.0833	0.0333	0.0250	0.0500	0.0292	12.31
L_q	3.20	3.20	6.40	4.09	1.41	5.50	14.06
W_q	0.0167	0.0667	0.0267	0.0208	0.0333	0.0229	14.23

Table VII.
Results from the analytical model

blocks per hour) to an optimal 5 : 1 ratio (200 versus 40 cylinder blocks per hour) for the fully-automated and semi-automated casting departments, as shown in Table VII. With the implementation of the new allocation policy, Technocraft would be able to reduce the total L_s by 12.50 per cent (7.00 versus 8.00 cylinder blocks), the total W_s by 12.31 per cent (0.0292 versus 0.0333 hours), the total L_q by 14.04 per cent (5.50 versus 6.40 cylinder blocks), and the total weighted W_q by 14.23 per cent (0.0229 versus 0.0267 hours).

In an attempt to verify our model, we developed a computer simulation model with the following characteristics:

- Poisson arrival and service process;
- FIFO; and
- no balking or reneging.

This is precisely the system considered analytically. The functions performed by our simulation program are as follows:

- developing the probability distribution of interarrival and service times for the naive and optimal queueing systems; and
- computing various measures of performance for each system.

The program was run using a time unit of one hour for a period of 25 hours (the data for the first two hours of the simulation were dropped to estimate better the performance measures during the steady state period). Our simulation indicated results similar to that of our analytical model. As shown in Table VIII, total L_s will decrease by 11.85 per cent (7.07 versus 8.02 cylinder blocks), the total W_s will decrease by 12.06 per cent (0.0299 versus 0.0340 hours), total L_q will decrease by 12.32 per cent (5.55 versus 6.33 cylinder blocks), and the total weighted W_q will decrease by 12.31 per cent (0.0235 versus 0.0268 hours) for the fully-automated and semi-automated casting departments respectively.

Finally, we employed chi-square tests of goodness of fit to check statistical conformity of the actual probability distributions of arrivals and services to the assumed theoretical Poisson probability distributions. For both approaches

	Allocation approach						
	Naïve			Optimal			
	Fully-automated casting	Semi-automated casting	Total casting	Fully-automated casting	Semi-automated casting	Total casting	Total casting percentage improvement
λ_1	189.80	46.52	236.32	197.34	39.15	236.49	–
μ_1	234.50	60.56	295.06	236.92	57.77	294.69	–
L_s	4.04	3.98	8.02	4.87	2.20	7.07	11.85
W_s	0.02123	0.0856	0.0340	0.0247	0.0562	0.0299	12.06
L_q	3.27	3.06	6.33	4.06	1.49	5.55	12.32
W_q	0.0172	0.0658	0.0206	0.0206	0.0381	0.0235	12.31

Table VIII. Results from the simulation model

		Allocation approach			
		Naïve		Optimal	
		Fully-automated casting	Semi-automated casting	Fully-automated casting	Semi-automated casting
Analytical model	Mean interarrival	0.005210	0.208300	0.005000	0.025000
	Standard deviation interarrival	0.000027	0.0000434	0.000025	0.000625
	Mean service time	0.004170	0.016670	0.004170	0.016670
	Standard deviation service time	0.000017	0.000278	0.000017	0.000278
Simulation model	Mean interarrival	0.005270	0.021500	0.005070	0.025540
	Standard deviation interarrival	0.000028	0.000466	0.000026	0.000599
	Mean service time	0.004260	0.016510	0.004220	0.017310
	Standard deviation service time	0.000019	0.000285	0.000018	0.000347

Table IX.
Interarrival and service time characteristics

(naive and optimal) the test indicated a good fit at a 0.05 significance level. The mean and standard deviation of interarrival and service times for the two systems (naive and optimal) are summarized along with the analytical mean and standard deviation of the interarrival and service times in Table IX.

Conclusion

In this paper we used analytical models and simulation to show that the naive allocation of total arrivals among service facilities in proportion to their service rates does not result in an efficient aggregate queueing system. In addition, we presented an optimization model which uses the Lagrange multiplier in order to find the optimal allocation of arrival rate that minimizes various objectives (L_s, W_s, L_q and W_q) in a group of systems. However, there are several limitations associated with this research.

(1) Although there are limited applications for a collection of $M/M/1$ systems, considering $M/M/c$, we can identify more examples since it would not be necessary to assume different service rates for each service facility for cases where the number of channels of each system are different. For example, an aggregate system with two service facilities may have different staffing capabilities as measured by the number of servers at each facility. In addition applications exist where a general arrival stream must be allocated to facilities with different physical locations, and different service rates. For example, the Internal Revenue Service (IRS) processes forms in various regional offices and tax payers are asked to submit their forms to a previously assigned region. Since the service rates at different regions may depend on the staffing level and the technology used, our model could be applied to find the optimal arrival allocation of tax returns to various regions. Once this optimal allocation

is found, IRS can then assign tax payers to different regions accordingly. Similarly, ships arriving to a general area could be allocated to a variety of ports each with its own service process.

(2) In spite of the fact that the model shows an improvement for all objectives, one could argue that in real-life settings, the facility which incurs the larger utilization and waiting time in the queue and the system might not co-operate.

(3) Another limitation could be unforeseen complication arising at each facility when the utilization goes beyond certain critical ranges. For example, the servers would slow down or machines might break down more when subjected to unusually high utilization.

These kinds of issues must be addressed in any queueing system application, as behavioural and other considerations alter the basic process when control is applied. In heavily loaded systems, it might be useful to study the effect of optimization of the aggregate waiting on the probability that waiting time exceeds given critical values. This could be achieved by imposing the appropriate constraints on waiting, or by using objective functions that measure various stochastic properties of the system such as the probability that aggregate waiting time exceeds critical values. In addition, the issue of fairness must be dealt with and the need for global or aggregate considerations must justify the increase in waiting time for some of the service facilities versus others.

Further variations of the model could be considered by assuming different distributions for arrivals and service. In general, one could study (either analytically or by simulation) the effect that the coefficient of variation in the arrival and service process has on the optimal allocations and the percentage improvements in the objectives. Other interesting cases could arise when considering multi-phase queueing networks whereby arrival streams allocated in the final phase must then be further broken down in successive phases. The objective would be to minimize aggregate waiting time of all service facilities including all phases. Perhaps a dynamic programming formulation could be useful in showing such multi-phase aggregate problems.

Several other objective functions could also be considered. For example, a weighted sum of the average waiting time of each server could be used in situations where it is more important to reduce the waiting time of one server versus another. Another objective function could be to minimize the sum of various percentiles of the waiting time distributions of each server instead of the average waiting time. These examples would involve interesting mathematical problems and could be the subject of future research.

Notes

1. The assumption of a fixed arrival rate is consistent with other models discussed in the literature. For example, Dyer and Proll (1977) and Rolfe (1971) considered the allocation of a fixed number of servers to a multiple facility service system each with a different

arrival. Hillier and Boling (1966; 1977; 1979) considered allocating a fixed amount of service time among a system of finite queues in series. Dallery and Stecke (1990), Stecke and Kim(1989), Stecke and Morin(1985), and Stecke and Solberg (1985) considered allocating workloads and machine group sizes in closed queueing networks. In all of these cases, the models consider the allocation of a fixed amount of resources among various alternatives. The assumption of constant service rate is also consistent with the models discussed earlier which generally assume non-decision variables such as service rate to be constant.

2. The name of the company has been changed to protect the anonymity of this company.
3. The fully-automated casting uses a large articulated robot along with an automated cooling system which allows a service time of 15 seconds per cylinder block compared to a service time of 60 seconds per cylinder block in the semi-automated casting. Two years ago the company decided to fully automate their casting department. However, due to the high cost of automation, a decision was made to complete the project in six years. After one year one casting department was fully automated and the second one was scheduled to go under automation in four years.
4. A comprehensive cost/benefit analysis done by the production engineering group revealed that the cost of changing the layout is so high that it is not feasible to locate the two casting departments in one building.
5. This is due to the fact that since the service time in the fully-automated department is 15 seconds, the service rate would be (3600/15 = 240 per hour) and since service time in the semi-automated department is 60 seconds, the service rate would be (3600/60 = 60 per hour). This 60 : 240 (or 1 : 4) was used as a basis for this distribution. In other words, out of the 240 cylinder blocks arriving into the dispatching centre per hour, 192 were dispatched to the fully-automated department compared to 48 dispatched to the semi-automated department. This dispatching policy resulted in 0.3125 and 1.25 arrival rates for the fully-automated and semi-automated casting departments respectively.

References

Dallery, Y. and Stecke, K. (1990), "On the optimal allocation of servers and workloads in closed queueing networks", *Operations Research*, Vol. 38, pp. 694-703.

Davis, E. (1977), "Optimal control of arrivals to a two-server queueing system with separate queues", PhD dissertation, Graduate Program in Operations Research, North Carolina State University, Raleigh, NC.

Dyer, M. and Proll, L. (1977), "On the validity of marginal analysis for allocating servers in *M/M/c* queues", *Management Science*, Vol. 23, pp. 1019-22.

Farrell, W. (1976), "Optimal switching policies in a nonhomogeneous exponential queueing system", PhD dissertation, Graduate School of Management, University of California, Los Angeles, CA.

Ghoneim, H. and Stidham, S. (1985), "Control of arrivals to two queues in series", *European Journal of Operational Research*, Vol. 21, pp. 399-409.

Hillier, F. and Boling, R. (1966), "The effect of some design factors on the efficiency of production lines with variable operation times", *Journal of Industrial Engineering*. Vol. 17, pp. 657-8.

Hillier, F. and Boling, R. (1977), "Toward characterizing the optimal allocation of work in production line systems with variable operation times", in Roubens, M. (Ed.), *Advances in Operations Research*, North-Holland, Amsterdam, pp. 649-58,

Hillier, F. and Boling, R. (1979) "On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times", *Management Science*, Vol. 25, pp. 721-8.

Johnson, S. and Stidham, S. (1980), "Control of arrivals to a stochastic input-output system", *Advances in Applied Probability*, Vol. 12, pp. 972-99.

Lazar, A. (1983), "Optimal flow control of a class of queueing networks in equilibrium", *IEEE Transactions on Automatic Control*, Vol. 28, pp. 1001-7.

Rolfe, A. (1971), "A note on marginal allocation in multiple-server service systems", *Management Science*, Vol. 17, pp. 656-8.

Shanthikumar, J.G. and Yao, D. (1987), "Optimal server allocation in a system of multi-server stations", *Management Science*, Vol. 33, pp. 1173-80.

Shanthikumar, J. G. and Yao, D. (1988), "On server allocation in multiple center manufacturing systems", *Operations Research*, Vol. 36, pp. 333-42.

Shioyama, T. (1991), "Optimal control of queueing network system with two types of customers", *European Journal of Operational Research*, Vol. 52, pp. 367-72.

Stecke, K. and Kim, I. (1989), "Performance evaluation for systems of pooled machines of unequal sizes: unbalancing versus balancing", *European Journal of Operational Research*, Vol. 24, pp. 22-38.

Stecke, K. and Morin, T. (1985), "The optimality of balancing workloads in certain types of flexible manufacturing systems", *European Journal of Operational Research*, Vol. 20, pp. 68-82.

Stecke, K. and Solberg, J. (1985), "The optimality of unbalancing both workloads and machine group sizes in closed queueing networks of multiserver queues", *Operations Research*, Vol. 33, pp. 882-910.

Stidham, S. (1978), "Socially and individually optimal control of arrivals to a *GI/M/1* queue", *Management Science*, Vol. 24, pp. 1598-610.

Stidham, S. (1985), "Optimal control of admission to a queueing system", *IEEE Transactions on Automatic Control*, Vol. 30.

Taha, H. (1982), *Operations Research*, 3rd ed, Macmillan Publishing Company, New York, NY.

Weber, R. (1978), "On the optimal assignment of customers to parallel servers", *Journal of Applied Probability*, Vol. 15, pp. 406-13.