Make It Specialized or Flexible?

Ick-Hyun Nam
College of Business Administration,
Seoul National University

Abstract

Throughput time, the time duration from customer arrival to service completion, is strategically important. In this paper, we consider flexibility as a way to reduce mean throughput time. In case where manufacturing or service systems are modelled as queueing systems, we derive substantial decrease in mean throughput time by incorporating flexibility. We study two types of flexible systems: a parallel flexible system and a serial flexible system. In a case of this paper, the mean throuthput time is reduced to less than a half via flexibility. Thus we can choose flexibility as an alternative for specialization which has been a conventional way to improve production efficiency.

The result in this paper should be helpful to the practitioners(both in manufacturing and service sectors) who try to shorten the response time to customers and gain competitive edge. By making servers more flexible, they can reduce mean throughput time since the flexible servers can help each other and sequencing priority can sometimes be applied appropriately.

1 Introduction

It is emphasized that quick response to customer needs is one of the key success factors. Lead time or throughput time, defined as the time duration from customer arrival to service completion, has been said to have great strategic value. Executives at strategically aggressive companies are altering their measures of performance

from competitive costs and quality to competitive costs, quality, *and* responsiveness. Today's innovation is time-based competition [9].

Many firms have pursued specialization in order to reduce lead time. By having each server focus on a small portion of a job which is simple enough to learn fast, they could get higher production efficiency and speed up response time with some success.

In this paper, we consider flexibility as an alternative way to obtain shorter lead time or quicker response. By increasing each server's capability such that it can process more types of jobs, we show that the mean lead time can be substantially reduced. One source of reduction in mean throughput time when we introduce flexibility is that the servers can help each other when needed. In a flexible system, every server is working on a job when there are waiting jobs. That is, flexibility reduces server idle time and thus jobs are processed more rapidly than in an inflexible system. The other source is that in some situation flexibility allows us to apply an appropriate sequencing priority and thus incurs more reduction in mean throughput time. Due to flexibility we can choose which class of jobs to process first. In a flexible system, by serving first a job requiring less mean service time, we can get more reduction in mean throughput time.

2 Specialization versus Flexibility

For comparison purpose, we consider $M/M/.$ queueing systems for analysis. That is, we consider a manufacturing or a service system modelled as queueing system where both customer interarrival time and service time are exponentially distributed. We will denote λ as the customer arrival rate and μ as the service rate of the system. In this paper, a customer and a job are used interchangeably, and a server can be either a human worker or a machine depending on the situation. Even for more general cases where exponential distribution assumption is violated, our result still applies with revised magnitude.

2.1 Specialization

Since Industrial Revolution, specialization has been emphasized to increase productivity. The tale on pin making is well known. In [1], Adam Smith wrote about an example of the pin-maker emphasizing the effect of the division of labour. The principle was to divide a job into several smaller tasks and to have each worker specialize on a tiny task. That is, by specializing in narrower scope of work, a worker can do his job more efficiently or produce more during the same time period. In this paper, the increase in productivity due to specialization is modelled as $(1+k)\mu$, where $k > 0$. That is, due to learning or experience effect from specialization, we have added service rate of k, which will be called as specialization effect. From the formula for $M/M/1$ queueing system[3], the mean throughput time then becomes

$$W = \frac{1}{(1+k)\mu - \lambda}$$

when we have the specialization effect of k.

2.2 Flexible System

In this paper, we define flexibility as follows: a server is said to be more flexible than another if it can process more scope of jobs. For example, in case worker 1 can process two types of jobs, A and B, while worker 2 can process job A only, then worker 1 is said to be more flexible than worker 2. There are, of course, several definitions of flexibility other than this such as capacity variability, but we stick to the definiton of flexibility in terms of the scope of processing capability of a server.

We will consider two typical kinds of queueing systems in order to show the benefit of flexibility: a parallel queueing system and a serial queueing system. In case there are two classes of customers requiring distinct services, A and B, and we have two servers which can process job A and job B respectively, then we have two independent $M/M/1$ queueing systems. This queueing system will be called a parallel queueing system in this paper.

In case a customer requires a series of jobs, for example A and then B, then we have a serial queueing system as in Figure 2. We will consider how much benefit in

terms of mean throughput time (or equivalently mean queue size in the system from Little's law) we will get when we incorporate flexibility to those two systems.

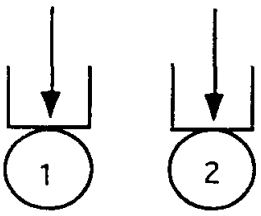

Figure 1. Two M/M/1

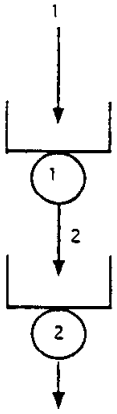

Figure 2. Serial Queueing System

2.2.1 Parallel Flexible System

Suppose we add flexibility to the parallel queueing system as in Figure 1 such that each worker can now process both jobs A and B. We will call this a parallel flexible system. In this parallel flexible queueing system, we now have single queue and two server and thus the system becomes a $M/M/2$ queueing system. In this flexible system, customers arrive and join at a single queue, and have their jobs started when it comes to their turn.

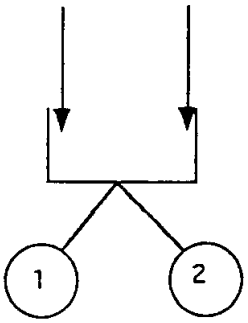

Figure 3. M/M/2 (Parallel Flexible System)

Depending on the traffic intensity of $\rho = \lambda/\mu$, we will take three cases to show the benefit of flexibility. The following table considers the queueing system where the

service rate is 10 and the arrival rate is 2, 5, or 8(three cases).

$\mu = 10$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 6$	$\lambda = 7$	$\lambda = 8$	$\lambda = 9$
Two $M/M/1(\alpha)$	0.13	0.14	0.17	0.20	0.25	0.33	0.50	1.0
$M/M/2(\beta)$	0.10	0.11	0.12	0.13	0.16	0.20	0.28	0.53
α/β	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
$1 + k^*$	1.16	1.21	1.24	1.25	1.24	1.21	1.16	1.09

The first row shows the mean throughput time for each case when we have two parallel distinct queueing systems which are inflexible. The second row gives us the mean throughput time when we add flexibilty and get a parallel flexible queueing system. The third row depicts the improvement ratio which is calculated by dividing the first row by the second. For the parallel flexible system, we note that the ratio becomes $1 + \lambda/\mu$. The last row gives us the k^* such that the queueing system with service rate of $(1 + k^*)\mu$ incurrs the same mean throughput time of corresponding parallel flexible queueing system. That is, k^* is the threshold value of specialization effect which matches the parallel flexibility effect. Thus when the actual specialization effect k is less than k^*, then the benefit of flexibility is greater than that of specialization in the parallel queueing system.

In the table, we can clearly see substantial decrease in mean throughput time by simply adding flexibility. In case 3, for example, we have about 80% decrease in mean throughput time by incorporating flexibility. Then what would be the underlying reason for this benefit? We can think of the source of flexibility effect as follows. In the parallel flexible system, mutual help between servers becomes possible. In the flexible system, no worker is idle as long as there are waiting customers. In the original inflexible system, it is possible that one server is idle while the other has waiting customers. In this case, we suffer server capacity loss. In summary, we have fuller utilization of server capacity in case of flexible system than in inflexible system.

Other than this efficiency due to flexibility, we have another good characteristic in the flexible system. In the parallel flexible system, a customer begins to be served according to his order of arrival. That is, a customer having arrived earlier gets

served earlier than others. The order of arrival is the same as the order of getting service. In the inflexible system, those orders can be reversed. In this sense, customers may experience improved fairness in the parallel flexible system. As an example of parallel flexible system, we can think of a bank where arriving customers join in a single line and get served as soon as one of the several tellers is available. In this case, each customer's starting service is according to his order of arrival, which makes the customers feel fairly treated.

2.2.2 Serial Flexible System

When we incorporate flexibilty into the serial queueing system as in Figure 2, we have a queueing system where two servers can process both jobs A and B. Then we have a serial flexible queueing system depicted as in Figure 4. In this serial flexible queueing system, we have the benefit in reducing mean throughput time due to mutual help between the servers as in the parallel flexible system.

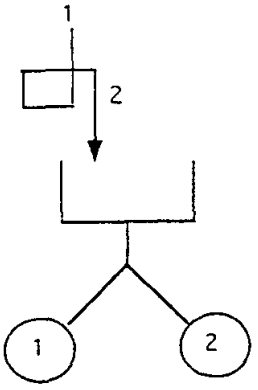

Figure 4. Serial Flexible System

In addition to this, we come to have more discretion which offers more reduction in mean throughput time. That is, unlike the parallel flexible system, we now have the discretion in deciding job orders of processing depending on the type of a job. In job sequencing, we can give priority to job B over job A, that is, we have the servers process job B customers (according to their arrival order) first before job A customers. The idea comes from the $c\mu$ rule which tells us to process first the jobs having shorter mean service time till completion. Since job A has mean service time of 0.2 till completion and job B has 0.1, we give priority to job B over job A. Under

this priority scheme, as soon as a low priority customer(job A) completes its first stage of service, it becomes a high priority customer(job B) and continues to be served till completion. For this reason, applying this sequencing priority to the serial flexible queueing system, we get a $M/E_2/2$ system. Here E denotes Erlang distribution. In this case, the stationary probabilty is very complicated to get[3, 4] and we give simulation results as in the following table. The simulation was run for $10,000$ time units using QSB+.

$\mu = 10$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 6$	$\lambda = 7$	$\lambda = 8$	$\lambda = 9$
Serial $M/M/.(\alpha)$	0.25	0.29	0.33	0.40	0.50	0.67	1.0	2.0
$M/E_2/2(\beta)$	0.20	0.21	0.23	0.25	0.28	0.34	0.46	0.82
α/β	1.21	1.34	1.46	1.60	1.76	1.94	2.16	2.43
$1 + k^*$	1.17	1.24	1.28	1.30	1.31	1.29	1.23	1.14

As we can see in the table, we have even larger reduction ratio than in the parallel flexible queueing system. The extra improvement comes, as mentioned before, from newly assigning sequencing priority. We can thus see that the benefit of flexibilty is substantial, and its magnitude is larger in a serial system than in a parallel system. But we should note that in a parallel flexible system where each type of customer has different mean service time, we can give sequencing priority to the customer class having smaller mean service time and thus get shorter mean throughput time than before the application of sequencing priority.

2.3 Other Considerations

When we incorporate flexibility into a queueing system and operate it, we should consider the setup loss incurring from changing several types of jobs. That is, when a server completes job A and the next job to do is B, he will have to spend some time for switching services. This can be called a setup loss and eats up the benefit of flexiblity. Therefore, when we try to incorporate flexibility, we should also consider the way to shorten switch-over time between jobs. S.M.E.D.(Single Minute Exchange of Die) as in Toyota Production System[7, 8] can be used for this purpose. We can

also think of investment on infrastructure making it possible to have minimal setup time or loss.

We should also consider the investment cost for incorporating flexibility. It is not always possible to get flexibility free of charge. For human servers, we should train them so that they can process multiple jobs and this training may require a fair amount of money. For machine servers, it is usually the case that the more flexible machine is more expensive than others with the same capacity.

3 Conclusion

In the past, mass production based on specialization was thought to be a way of becoming an industry leader. Productivity seemed to be a major factor in competition. When customers did not require variety and demand exceeded supply, firms did not have to worry about the inventory holding cost. They had to find a way to produce more during fixed time period. For this purpose, specialization was chosen and considered to be the best way to improve the production efficiency.

As customers require more variety and supply capacity exceeds demand, firms come to realize that they should produce a variety of products or services according to customer orders. In this era, the lead time becomes more important[9]. In this paper, we considered flexibility as a way to reduce mean throughput time. We defined multi-task personnel or machine as a flexible server, and showed the effectiveness of flexibility in reducing mean throughput time. We considered two types of flexible systems, a parallel flexible system and a serial flexible system, and gave flexibility effects in reducing mean throughput time for three cases respectively. In the serial flexible system, we had even larger efficiency due to sequencing priority than in a parallel flexible system. Readers can refer to [6] in order to understand the flexibility effects for other more general queueing systems.

In the personnel resource management area, the effect of job enlargement or more specifically, job extension [2] has been emphasized. But the focus was mainly on psychological effect such as job satisfaction. Our definition of flexibility in this paper is equivalent to job enlargement in the sense that a server can process increased number of job types. In addition to the psychological effect, we showed the benefit of

flexibility in reducing mean lead time. We should also note that the flexibility idea can be applied not only to manufacturing system but also to service system.

References

[1] Cannan, E.(editor), *An Inquiry into the Nature and Causes of the Wealth of Nations by Adam Smith*, London, Methuen, 1904.

[2] Gordon, J. R., *A Diagnostic Approach To Organizational Behavior*, Allyn and Bacon, Inc., 1987.

[3] Gross, D. and C. M. Harris, *Fundamentals of Queueing Theory*, John Wiley & Sons, 1985.

[4] Hillier, F. S. and O. S. Yu, *Queueing Tables and Graphs*, New York, North Holland, 1981.

[5] Martinich, J. S., *Production and Operations Management*, John Wiley & Sons, Inc., 1997.

[6] Nam, I. H., *Flexibility in Manufacturing:Dynamic Scheduling and Resource Pooling*, Stanford Univ. Ph.D. Dissertation, 1992.

[7] Ohno, T., *Toyota Production System*, Cambridge, MA, Productivity Press, 1988.

[8] Shingo, S., *A Study of the Toyota Production System*, Cambridge, MA, Productivity Press, 1989.

[9] Stalk, G. Jr. and T. M. Hout, *Competing Against Time*, New York, The Free Press, 1990.