Tijdschrift voor Economie en Management
Vol. XXXIX, 4, 1994

Queueing Theory and Operations Management

by M. LAMBRECHT* and N. VANDAELE*

I. INTRODUCTION

Waiting is an intimate dimension of our daily lives. Everyone has experienced waiting in line at the supermarket, the bank and any number of other places. We constantly observe traffic, hospital or court congestion, customers or machines are waiting and we experience waiting times for almost every service offered. These waiting-line situations are also called queueing problems. The common characteristic is that a number of physical entities (the arrivals) are attempting to receive service from limited facilities (the servers) and as a consequence the arrivals must sometimes wait in line for their turn to be served. Numerous applications are described and the mathematics of queueing has advanced tremendously over the last 40 years.

The objective of this paper is to focus on operations management applications of queueing theory. The first textbook on the subject: "Queues, Inventories and Maintenance" was written in 1958 by Morse. A tremendous number of queueing problems occur in production and inventory management. Think of the design of facility layouts, staffing decisions, maintenance problems, the physical capacity problem, lead time estimation and lot sizing decisions to mention only a few. Over the last decade Just-In-Time (JIT), Time Based Competition and the Fast Cycle Time strategies gave rise to a renewed interest in queueing. Indeed, a Fast Cycle Time strategy is basically dealing with time, with reduced waiting times and an emphasis on a fast Time-to-

* Department of Applied Economic Sciences, Operations Management Group,K.U. Leuven.
This research was supported by NFWO/FKFO Belgium project 2.0053.93.

Market. It is amazing to realize that with a little understanding of how queues behave, the solution to many operations management problems becomes clear if not obvious.

The paper is organized as follows. We select three major problem areas in operations management: the inventory-capacity trade-off, the impact of uncertainty (disruptions, variability) and capacity utilization on lead time and the impact of lot-sizing on lead times.

We show how insights from queueing theory may be helpful to better manage these issues. It is tempting to treat the subject mathematically, but we opt in this article for a more qualitative approach. The enthusiastic reader however may not underestimate the mathematical intricacies involved.

II. INSIGHTS FROM QUEUEING THEORY

A. *The Capacity-Inventory Trade-Off*

In order to better understand the capacity-inventory trade-off, it is important to understand the nature of the Just-In-Time (JIT) revolution.

The JIT Revolution can be summarized as follows (Zangwill (1992)): "The old viewpoint: Increase inventory, hold a lot in stock, and then you are ready for anything. The new viewpoint: Reduce inventory, cut the production lead time and you can respond fast to anything. These are two opposing views about being responsive to the customer". In the first case companies satisfy customer orders from stock, which is an immediate response. In a JIT environment companies satisfy customer demands with a certain time delay, which of course is kept as small as possible. We view the company more as a queueing system instead of an inventory system. Behind this new viewpoint focussing on a fast response, there is a synergetic chain of manufacturing changes that goes several layers deep. A successful implementation depends on the ability to eliminate all forms of waste, continuous improvement, employee involvement, disciplined implementation, supplier participation, reorganization of the production floor, modular designs, cell layouts, process control and total quality creation. The objective is to improve productivity. Moreover, through this fast response to specific customer needs, it is hoped that it results in an enhanced market power. The improved productivity and the stronger market position are supposed to be the basis for a sustainable competitive advantage.

The question now is: how can we guarantee a fast response without the protection of inventory as JIT asks us to do?

In order to answer this question, let's turn to a basic insight from queueing theory. It is well known that companies that try to operate with tight capacity are forced to carry substantial inventories to protect against unexpected surges in demand and other contingencies (Zipkin (1991)). High levels of capacity utilization cause increased congestion, longer lead times and higher inventories due to uncertainty. So if a company wants to reduce lead times or lower the inventories then it is advisable to have excess capacity. That's the inventory-capacity trade-off. We quote from Zipkin (1991), "Indeed, companies often find that JIT means buying more and better equipment - a serious commitment of capital resources".

In today's manufacturing environment companies are stressing due-date performance, time (cycle time, response time, time-to-market) and reduced inventory levels as primary measures of shop performance. In order to achieve this, companies seek to add capacity cushions in an attempt to become more responsive to customer demands (instead of inventory buffers). This of course is contrary to the traditional performance measure of resource efficiency (high levels of machine utilization). The core problem is the evaluation of the benefits associated with lower inventories versus the lower efficiencies associated with excess capacity. The question is whether a company is better off by replacing inventory by capacity,.. or by keeping the machine assets tight and accepting more inventory.

In order to have some empirical evidence of this phenomenon we analyzed the inventory position and the capital investments in the Belgian metal working industry in the period 1977-1991. Over this period the inventory position measured as work-in-process and finished product inventory relative to value added dropped from 50% to 31%. The investments in material fixed assets relative to value added increased from 32 % to 42 %. Another interesting observation is the following. In the period 1977-1991 total sales in the Belgian industry increased roughly by 300 % (including inflation and taking 1977 as the reference year). Over the same period depreciation charges increased by 420 %.

The decrease in inventory is of course not only attributable to the capacity expansion. A period of economic growth e.g. is always associated with a period of inventory depletion. It is also known that in-

vestments in automation and flexible equipment are larger than the required investments for conventional machinery.

The positive side of the coin is that the increased capital intensity positively contribute to the employee's productivity and that reductions in inventory also help to improve worker productivity. How can reducing inventories improve productivity? One possible mechanism is the dynamic learning process, inventory reduction helps to achieve a higher learning rate through a clearer exposition and easier identification of problems. (Kim (1993)).

There is however also a major drawback associated with the above-mentioned redistribution phenomenon. The question is what happens to companies that heavily invested in plant and equipment and that are confronted with a period of economic recession? The drop in demand, the entrance of many new competitors, and the heavy investment boom created in many industries huge overcapacities, prices dropped, profits disappeared. We again experience a period of intensified price competition (cost cutting programs).

Don't forget that one of the premises of the JIT, Time-based philosophy was the prospect of achieving competitive advantage, higher margins (premiums) and more attractive profits. Now it turns out that excess capacity is an element of rigidity, a source of additional riskiness that may result in more variability of performance.

Is there a solution to this problem? Let us therefore go back to queueing theory. There we learn that variability and uncertainty are the key parameters. The more uncertainty, the more damaging high levels of machine utilizations are on inventories and lead time. We expect in other words lower levels of capacity utilization (more excess capacity) in job-shop manufacturing (e.g. machine building) compared to the more standardized manufacturing environments (e.g. consumer electronics). The argument is that the greater the uncertainty (e.g. in the receipt of customer orders), the higher the negative impact of increased congestion on inventories and lead times. We indeed observe a 10 % point difference in average utilization between the industrial product sector (72 %) and the consumer product sector (82 %) of the Belgian metalworking industry (period 1981-1992).

Every effort to reduce the level of variability (process control, zero defects, better supplier relationships, better forecasting,...) will automatically have a positive impact on inventories. The process of continuous improvement is one of the only ways out to escape from the inventory-capacity conflict, which is basically a conflict between flexi-

bility (responsiveness) and efficiency. Every inventory reduction program should be backed up by efforts of continuous improvement and better capacity management.

So the key to the solution is fighting disruptions caused by process instability and all sorts of unreliabilities. Disruptions lead to unnecessary high capital costs. Fighting disruptions is a learning process offering a clear target for human resources management. Ultimately 'people' implement strategies. Participative management combined with self-directed teams emphasizing joint problem solving and team work, total productive maintenance based on responsibility at the source are all means to achieve the objective.

In the next paragraph we analyze in greater detail the impact of uncertainty and capacity utilization on lead times.

B. *Impact of Disruptions and Capacity Utilization on Lead Time*

The fast cycle strategy and the associated crusade for lower inventories are based on the best known relationship of queueing theory: Little's Law. For simplicity, assume a single server queueing model with an arrival and processing rate of λ and μ customers per time unit. Under steady state conditions, Little's Law combines the two most important operations management performance measures into one formula: the average number of customers in the system, E(L) (equivalent to the average inventory) and the average time units spend in the system, E(W) (equivalent to average lead time).

Little's Law: $E(L) = \lambda.E(W)$

Little's Law which is quite general and applies to any queue discipline specifies how inventory and time in the system are linked. A system containing a lot of inventory inevitably results in long lead times or, conversely reduce inventory and respond fast. The lead time, defined as the total elapsed time from order arrival until the order is finished and the customer is served, consists of two important parts: the waiting time and the processing time. The latter is mostly a fairly stable component of lead time. The average waiting time however is highly sensitive to system conditions such as the level of uncertainty and the capacity utilization. Utilization (ρ) is defined as the ratio of input rate to processing rate:

$$\rho = \lambda/\mu$$

Based on this, one can quantify quite easily the impact of uncertainty and utilization on average lead time.

In general one can state that higher utilizations and/or higher levels of uncertainty cause longer waiting times and consequently longer lead time and higher levels of inventory. This in turn induces strategies to improve performance. One possibility e.g. is to consider capacity expansion (see paragraph A) another is to reduce the uncertainty in the system by eliminating all disruptions. This can be accomplished by automation, a better trained work force, standardisation of processes, more design efforts, improved maintenance practices, quality improvements or in general all efforts related to continuous improvement (Kaizen). A careful analysis of the Japanese Production System immediately reveals that it is based on a combination of both abovementioned strategies.

An exact quantification of the impact for more general situations (multiple machines at a workcenter, multiple part types, different routings, lot sizes, rework and other feedback loops, interdependencies,...) requires of course an in depth mathematical treatment. Most steady state relationships for queueing networks are these days made available in commercial software packages. These software packages capture the main dynamics of the production system in a set of mathematical equations, which are next solved so that the system performance can be obtained with amazingly little computer time. This analytical approach can be viewed as a valuable alternative to the more traditional simulation approach. The analytic approach brings the mathematics of queueing in reach of management who can use it as a dialogue tool to evaluate various strategic options. Examples of commercially available software packages are MPX (Network Dynamics) and QNA (Queueing Network Analyzer, Whitt (1983)).

To fit comfortably as a manufacturing model, a queueing model still exhibits a serious disadvantage. For many queueing systems, means are known but little else. In other words it is possible to express queueing behavior by means of averages (average inventory, average lead time,...). Knowledge of the average lead time alone is insufficient, what is also needed is the variance. Indeed, the variability of the lead time determines to a large extend the probability of meeting quoted or promised lead times. In an inventory driven system demand and production uncertainties are protected by safety stocks, in a fast cycle time

approach the protection is based on a safety time. The safety time can be quantified by means of a multiplier. The question is by what factor do we have to multiply the average lead time so that a quoted lead time is met X % of the time. Traditional inventory theory is mainly concerned with fixing order quantities and safety stocks. The new approach is concerned with quoting reliable lead times and consequently requires a safety time protection. In many cases the issue is not to quote a lead time but to satisfy a market imposed lead time.

It is clear that more safety time will be needed the larger the variability of the lead time. Moreover, the level of capacity utilization is also very important. Higher levels of utilization cause higher lead time variances and service levels will deteriorate. The congestion phenomenon (utilization and uncertainty) is again the key to any lead time reduction program. See Lambrecht, Shaoxiang and Vandaele (1994) for a more detailed discussion.

C. *Lot Sizing and Lead Times*

Another key variable that impacts the lead time is the lot sizing decision. The lot sizing decision is probably the most intensively researched issue in operations management. The traditional approach focuses on balancing ordering costs and inventory holding costs. Since the advent of time-based strategies attention was turned to analyzing the impact of lot sizing on lead time. Traditionally the lead time was held constant, the objective now is to replace the deterministically assumed lead time by a stochastic lead time as a function of the lot size, uncertainty, capacity utilization and other parameters. The determination of this stochastic lead time is based on queueing theory and has been analyzed by Banker et al. (1988), Williams (1984), Zipkin (1986) and Wein ((1990), (1992)).

Amazingly enough this relationship has been misinterpreted by many researchers and practicioneers. The reasoning goes as follows: large lot sizes will lengthen the lead time and small lot sizes will automatically result in short lead times. This is wrong. Queueing theory will keep us on the right path. The rationale goes as follows: for a given setup time, some portion of the available time at a production facility will be spend on performing setups. Total setup time depends of course on the lot size. A small lot size results in a larger proportion of setup time and the capacity utilization of the production facility will increase. So, by manupilating the lot size the capacity utilization can

be changed, and we know from the previous sections that utilization impacts the lead time.

At this point it can be shown that two phenomena are present in the lot sizing decision: a batching effect and a congestion (saturation) effect. A large batch will cause a long lead time (batching effect), but on the other hand very small batches will increase the capacity utilization (the setup time portion), congestion starts and consequently lead times will go up again. Both phenomena result in a convex relationship between lot size and average lead time. The conclusion is that both large and small lot sizes cause long average lead times. Analogous to the previous section it can be shown that the variance of the lead time is also a convex function of the lot size. Consequently, customer service will deteriorate both for very small or large lot sizes. It is interesting to note that exactly the same conclusion is reached in the traditional cost based approach, balancing holding costs and setup costs. In the queueing approach, we balance the batching and the congestion effect. Both approaches will however not result in the same optimal lot sizes.

The full benefits of reduced batch sizes can only be obtained by reducing the level of uncertainty (disruptions), by maintaining a reasonable level of excess capacity or by reducing setup-times. The very popular setup-time reduction programs perfectly fit in this approach, it is an excellent way to realize continuous flow production, short lead times and high service levels. For more details see Karmarkar (1987).

One of the recent developments in computer communication systems such as computer networks opened new perspectives for lot sizing models. A common mode of operation for computer networks is e.g. polling. A polling model is a queueing model composed of a set of queues and a single server who visits the queues in a predetermined order. The data transfer from the terminals to the computer is controlled via a polling scheme in which the computer "polls" the terminals, requesting the data, one terminal at a time (Westrate (1992)). In such a situation it is important to know how long the computer serves the same terminal. The analogy with a lot sizing problem is obvious.

III. CONCLUSION

Most manufacturing operations are stochastic because of uncertainty in the timing of customer orders or the receipt of purchased mate-

rial and because of variability in the processing and set-up times caused by various disruptions. All this increases congestion and consequently inflates lead times and creates excess inventories. In a time-based production environment that's exactly what we want to avoid. So the basic question is how to handle congestion, how to take advantage of the trade-offs between various performance measures such as work-in-process, lead-times and investment in capacity. Insights from queueing theory are of great help here.

A first strategy is to install some capacity in excess of expected demand. Indeed, capacity can be used to buffer the system against unexpected events (instead of the standard inventory buffers). This strategy is somewhat contrary to the traditional performance measure of resource efficiency. That's probably the reason why many companies are reluctant to have large amounts of standby capacity, after all, a large part of the Belgian industrial sector is highly focused on scale intensive activities (VEV, 1994). Instead of focusing on excess capacity it may be advisable to concentrate on a flexible use of the existing capacity (flexible working time schemes). This in turn offers a new incentive for increasing the use of flexible labor, both in terms of the number of people employed (numerical flexibility) and in terms of the mobility of employees to undertake a range of tasks (functional flexibility).

A second strategy is to focus on uncertainty and variability reducing programs. Indeed, the most damaging factor in the pursue of a fast cycle strategy is the existence of all sorts of disruptions. Disruptions lead to congestion, it lowers the speed and it leads to high capital costs and inefficiencies all over. Process stability and reliability are obtained by quality and maintenance improving programs, by better designs and most importantly by installing a problem-solving attitude of all those involved in manufacturing. This is probably best obtained by focussing on small group activities in which learning and knowledge accumulation can result in an enhanced human competence and organizational commitment.

REFERENCES

Banker, R., Datar, S. and Kekre, S., 1988, Relevant Costs, Congestion and Stochasticity in Production Environments, *Journal of Accounting and Economics*, 10, 171-197.

Karmarkar, U., 1987, Lot Sizes, Lead Times and In-Process Inventories, *Management Science*, 33(3), 409-423.

Kim, T., 1993, Reducing Inventory and Improving Productivity: Evidence from the PIMS Data, (Working paper University of California, San Diego).

Lambrecht, M., Shaoxiang Chen and Vandaele, N., 1994, A Lot Sizing Model with Queueing Delays: The Issue of Safety Time, (Onderzoeksrapport 9402, Departement Toegepaste Economische Wetenschappen, K.U. Leuven).
Meyer, C., 1993, Fast Cycle Time, (The Free Press).
Morse, P., 1958, Queues, Inventories and Maintenance (John Wiley).
Vlaams Ekonomisch Verbond, 1994, Op zoek naar Groei: Het Strategisch Plan voor Vlaanderen, (Uitgeverij Pelckmans).
Wein, L., 1990, Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network with Controllable Inputs, *Operations Research*, 38(6), 1065-1078.
Wein, L., 1992, Dynamic Scheduling of a Multiclass Make-to-Stock Queue, *Operations Research*, 40(4), 724-735.
Westrate, J., 1992, Analysis and Optimization of Polling Models, (Doctoral Dissertation, Katholieke Universiteit Brabant).
Whitt, W., 1983, The Queueing Network Analyzer, *The Bell System Technical Journal* 62, 2779-2815.
Williams, T., 1984, Special Products and Uncertainty in Production/Inventory Systems, *European Journal of Operational Research* 15, 46-54.
Zangwill, W., 1992, The Limits of Japanese Production Theory, *Interfaces* 22, 14-25.
Zipkin, P., 1986, Models for Design and Control of Stochastic, Multi-item Batch Production Systems, *Operations Research* 34(1), 91-104.
Zipkin, P., 1991, Does Manufacturing need a JIT-Revolution?, *Harvard Business Review*, Jan-Feb., 40-50.