Lot-Sizing and Lead-time Performance in a Manufacturing Cell

UDAY S. KARMARKAR
The Graduate School of Management
University of Rochester
Rochester, New York 14627

SHAM KEKRE
Eastman Kodak Company
Rochester, New York 14650

SUNDER KEKRE
The Graduate School of Industrial Administration
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

SUSAN FREEMAN
Eastman Kodak Company
Rochester, New York 14650

Many complex multi-item manufacturing shops have high levels of work-in-process because of queueing delays at machines and consequently long manufacturing lead times. These delays are directly related to lot sizes. Two alternative approaches applied to modeling a manufacturing cell — simulation and a novel analytical lot-sizing model (Q-LOTS) — provide very similar results and validate the analytical model.

Complex, multi-item job shops invariably have high levels of work-in-process and long manufacturing lead times because of the queueing delays at work centers. These problems are well recorded (for example, Burbidge [1975]). Estimates suggest that typically only 10 to 15 percent of shop time for a job is spent in actual processing. However, the causes underlying this phenomenon have not been well understood until recently. In particular, Karmarkar [1983a, 1983b] shows that for closed job shops lot-sizing policy is a major determinant of the extent of queueing delays. Yet, most such job shops fail to take this important consequence into account in establishing lot sizes.

We describe two independent attempts to analyze these phenomena: a simulation model developed by Eastman Kodak's Apparatus Division that examined the behavior of a particular manufacturing cell as the lot-size policy for the cell was changed; and an analytical model [Karmarkar 1983; Karmarkar et al. 1983, 1984]. These two approaches afforded on one hand an interesting opportunity to use the analytical model to confirm the empirically observed results from the simu-

0092-2102/85/1502/0001$01.25
This paper was refereed.

PRODUCTION/SCHEDULING — LOT SIZING (CLOSED SHOP)

lation and to reveal the underlying mechanisms involved. On the other, the simulation provided the means to validate the general model (Q-LOTS) with a specific instance.

Simulation of a Manufacturing Cell

The manufacturing cell in question was organized to improve the production of a group of similar parts that had proved to be troublesome because of long production lead times, high in-process inventories, and difficulties in coordinating assemblies. The 13 parts were grouped on the basis of similar process characteristics, and a production cell, separated from the functionally organized shop, was created to produce them. The cell contains 10 major processing work centers with three other minor preparation and finishing operations. Because three of the major centers have more than one machine, there are, altogether, 15 machines in the cell. The work centers include a manual lathe, an NC lathe, and routine operations (drill, punch), as well as certain proprietary metal forming processes which are quite complex. Part flow through the cell is not uniform and varies across parts with some recirculation or multiple visits for certain parts.

The simulation model was motivated by a need to predict and understand the operating characteristics of the cell. One major task was to devise appropriate lot-sizing policies for the cell; for example, to investigate whether the number of setups should be reduced on bottleneck machines. As it turned out, the reverse was, in a sense, the better policy.

The simulation was written in GPSS, which suited the discrete-event nature of the problem. Once the capacities at work centers are fixed, the simulation is driven by the annual demand for each part and the lot sizes chosen. Lots are released at uniform intervals to the cell, and data is collected about queueing times, total lead time, the number of setups made, and work-in-process inventory. While a detailed scientific validation of the simulation against the cell has not been done for various pragmatic reasons, over a year of experience with the cell and the simulation through a variety of parametric changes has convinced its users that the simulation is an accurate representation of cell behavior for the purposes at hand.

Results from the Simulation

The simulation was used to study capacity and design problems as well as lot sizing. Initially the characteristics of the cell were studied under the lot-sizing policies obtained from conventional EOQ models used by an existing production control system. Then the lot sizes were perturbed to examine the consequences. As a first arbitrary attempt, all lot sizes were cut in half. Surprisingly, this did not result in the catastrophic consequences that EOQ models would have foretold. Instead, lead time and work-in-process (WIP) both showed reductions while productivity did not drop. Further across-the-board cuts worked well up to a point, but then lead time and WIP deteriorated abruptly as queues started to appear at various points in the system.

Without recounting all the details, the set of possible lot sizes (a 13-dimensional vector) was searched, guided chiefly by intuition and trial and error. A significant difficulty in this process was caused by the way in which queues would shift to different machines as lot-size patterns were altered. Roughly speaking, the search rules were

(1) Start with fairly large lot sizes,
(2) Reduce all lot sizes until a queue appears,
(3) Increase lot sizes on those parts which have a high setup time on the machine with a queue,
(4) Keep reducing lot sizes on other parts until a queue appears elsewhere, and so on.

Experimentation over several weeks led to substantial improvements, reducing lead times and work-in-process by a factor of over 50 percent compared to the initial results.

The Analytical Approach

The extant literature on lot-sizing methods by and large does not address the issues of manufacturing lead times and work-in-process. Yet many practitioners and firms have intuitively understood that increasing lot sizes increases production times. Indeed, this is easily demonstrated in the context of a deterministic model [Karmarkar 1983a, 1983b]. Interestingly, this phenomenon is referred to by Magee [1956] in a description of a product cycling problem and by Sugimori et al. [1977] in their discussion of the Toyota Kanban system. It is also mentioned by Sasser et al. [1982] in the Granger Transmission case study. What seems to be less well understood is the countervailing phenomenon that small lot sizes exacerbate the queueing and sequencing delays that occur in complex shops by increasing the load on work centers. While this phenomenon is caused by the higher number of setups, the usual device of using setup costs is an ineffectual and incorrect representation of actual behavior. This is because setup costs are based on a view of capacity as a binding constraint, whereas in job shops, queues effectively discourage high-loading (utilization) levels long before nominal capacity is reached. In essence, the cost of excessively small lot sizes is due to the long lead times and high levels of work-in-process caused by queues.

There have been many studies which model manufacturing facilities as queueing systems [Buzacott 1974, 1980; Koenigsberg and Mamer 1982; Shanthikumar and Buzacott 1981; Solberg 1977; Stecke and Solberg 1981; and Suri 1983]. However, most have not included the effect of lot-sizing policies. One exception is the paper by Zipkin [1983]. Although oriented towards somewhat different concerns, Zipkin has independently developed an approach that is mathematically similar to ours. A paper by Seidmann and Schweitzer [1983] also considers the effect of batch sizes for the special case of flexible manufacturing systems.

The impact of batch size on lead times can be intuitively understood as follows. Consider a machine or resource at which batches queue up, waiting to be processed. For simplicity, the batches are taken to be alike. Each batch requires a setup plus some processing time which, unlike the setup, depends on batch size. If the rate of arrival of work is held constant and batch size is increased, the time that a batch spends on the machine increases linearly; hence the total work that arrives at the queue while a batch is being processed is greater even though the *number* of batches in queue may not change very much. Thus an arriving batch sees more work waiting ahead of it and also requires more time for its own processing. Since the effect of the fixed setups is diminishing, average queueing time and total time in system eventually increase linearly as batch sizes are increased.

Now consider the impact of reducing batch size. Work arrives at the machine at the same rate but because it does so in smaller batches, the amount of time spent on setups increases. Thus, although the real or productive utilization of the machine remains unchanged, the total work load (intensity) increases. At some point this leads to the buildup of large queues which cause queue times to rise sharply even though processing time per batch continues to drop. Clearly, there is a lower limit on batch sizes at the point where the total processing time plus setup time exceeds the time available on the machine.

In the case where the machine is modeled heuristically as an M/M/1 queue processing identical items, it can be shown (appendix) that the average time T spent in the system by a batch is given by

$$T = \frac{(\tau + Q/P)}{1 - (D/P) - (D\tau/Q)} \qquad (1)$$

where D = Total work to be done (units/time)

P = Processing rate at the machine (unit/time)

Q = Batch size

τ = Setup time per batch.

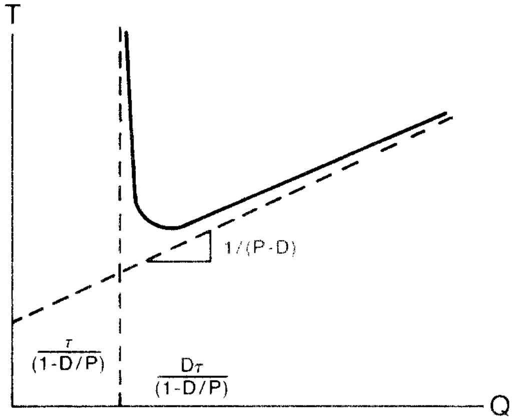

Figure 1: Average waiting times *(T)* as a function of lot size *(Q)*.

The batch size Q cannot be smaller than $D\tau(1-D/P)$. For large Q, approximately $T = Q/(P-D)$ where $1/(P-D)$ is the average number of batches in the system. Figure 1 shows a graph of (T) average waiting times versus Q.

This queueing model was extended to the multi-item case [Karmarkar 1983a; Karmarkar et al. 1983] by modeling the facility as an M/G/1 queue where an exact expression is available for the average time in queue. Next, the case of manufacturing systems with many work centers with several machines at each center was modeled as an open network of M/G/c queues [Karmarkar et al. 1984]. The treatment of this case is necessarily heuristic since no exact analysis exists. At each stage the queueing model was imbedded in an optimization model that determined the best lot sizes for a given objective function. The most general case, which requires the solution of a nonlinear program, has subsequently been coded as a computer program called Q-LOTS. For convenience, we use this term for our analytical approach.

Comparison and Validation

In comparing the two approaches, it is important to remember that

(1) The assumptions underlying the analytical model are quite different from those underlying the simulation mechanism. In particular, the former assumes randomness in arrivals, while the simulation uses a uniform rate of release of work to the cell.

(2) The real cell is a different matter again — the stochastic queueing model may be a better representation of actual behavior since the arrival of batches to the cell is *not* uniform.

The first comparison was qualitative. The behavior observed empirically was explained by the queueing mechanism. In turn, by scaling the lot sizes in the simulation by a constant factor, the characteristics exhibited in Figure 1 were

Part Number	Q-LOTS Lot Size	Part Lead Time (days)	Best Search Lot Sizes	Lead Times (days)
1	168	9.35	270	13.41
2	112	11.01	168	15.31
3	84	8.50	50	9.63
4	158	4.71	94	4.57
5	179	3.98	90	3.70
6	371	8.35	213	8.86
7	152	3.02	187	2.74
8	128	3.64	170	4.46
9	109	13.73	168	19.52
10	102	11.33	168	15.68
11	109	13.11	168	18.59
12	111	1.90	156	2.46
13	203	8.39	144	8.16
Demand Weighted Lead Time	—	7.26	—	9.08

Table 1: Comparison of average lead time (days) and lot sizes (units) by part number as given by Q-LOTS and the best search results. The foot of the table gives the demand weighted average lead times.

	Lot Sizes From Q-LOTS	Lot Sizes From Simulation
Evaluation on Q-LOTS	8.57	10.66
Evaluation on Simulation	7.26	9.08

Table 2: Cross-validation results: weighted average lead times (days) predicted by Q-LOTS and the simulation for both sets of lot sizes.

corroborated. Next, Q-LOTS was used to try to determine the best lot sizes for the cell independent of the results from the simulation. Since the objective in the simulation study had been the minimization of average lead times, this was also used as the objective for Q-LOTS. More precisely, the objective function used was the demand weighted lead time for all the parts processed by the cell. Since the work-in-process for a system is given by production rate x lead time, this was also equivalent to minimizing the average number of items in process.

When the lot sizes produced by Q-LOTS are compared with the best lot sizes obtained by trial-and-error search on the simulation, the differences are substantial (Table 1). For example, Q-LOTS picks a lot size for part 5 that is twice the search value; the Q-LOTS choices for items 9, 10, and 11 are much smaller. Overall, Q-LOTS does better by almost 20 percent than the best search, largely because of its better performance on item 1 (a high volume item) and items 8-12.

In addition to running the Q-LOTS output on the simulation, we also tried to evaluate the search-lot sizes using the analytical model (Table 2). While the predictions of Q-LOTS and the simulation do not match exactly, they are very close

Scale Factor a	Average Lead Time (days)		WIP ($)	
	Q-LOTS	Simulation	Q-LOTS	Simulation
0.60	49.56	17.05	391,210	168,140
0.80	9.32	7.90	75,290	65,810
1.00	8.57	7.26	68,840	60,400
1.25	9.06	8.24	72,750	68,860
1.50	9.91	9.16	79,680	76,320
1.75	10.89	10.30	87,650	85,440
2.00	11.95	11.47	96,170	95,910

Table 3: Comparison of average lead time (days) and WIP($) predicted by Q-LOTS and simulation as the lot-size vector is scaled by a factor α.

with Q-LOTS showing a 20 percent advantage in both evaluations.

The results suggest that Q-LOTS found a slightly better solution than the search. However, it was possible that a better solution could exist (in the sense of performing better on the simulation). To thoroughly search the neighborhood of the Q-LOTS solution on the simulation would have been too time consuming; we tried a simpler alternative. We scaled the lot-size vector Q^* produced by Q-LOTS by a factor α ranging from 0.6 to 2.0 and entered the scaled lot sizes into the simulation and into Q-LOTS. The weighted average lead times for the cell for each of these vectors are given in Table 3. They represent the behavior of the simulation and the analytical model on a ray passing through Q^*. The table also shows predicted work-in-process on a cost basis which is slightly different from the unit basis mentioned earlier because of differences in costs across parts. The two measures for the two approaches are graphed in Figures 2 and 3.

The following observations were made from the parametric analysis:

— Q-LOTS corresponds fairly well with the simulation results at large lot sizes and more poorly at smaller lot sizes. However, the simulation becomes unstable in the congestion region. For example, a subsequent simulation run at $\alpha = 0.9$ produced an average lead time of 10.29 days and a WIP level of $84,210. Thus the behavior predicted by Q-LOTS at low α is not necessarily an overestimate.

— Q-LOTS appears to find the minimum quite well and generally indicates the qualitative behavior of the objective correctly.

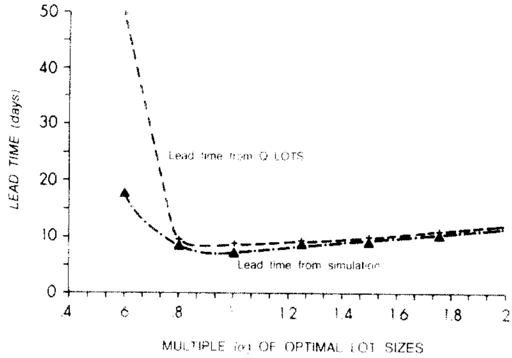

Figure 2: Comparison of average lead time (days) predicted by Q-LOTS and simulation as optimal lot sizes are scaled.

Application of the Methods

These methods can be applied in

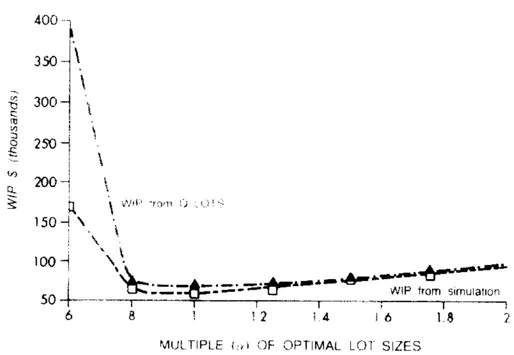

Figure 3: Comparison of WIP($) predicted by Q-LOTS and simulation as optimal lot sizes are scaled.

closed job shops which are multipart, multimachine manufacturing systems with repetitive batch production either to inventory or for assembly. The methods have potential uses at several levels.

At the scheduling level, the models can be used to devise optimal lot-size policies and to predict the performance of a facility for a given policy. Operationally, these policies can be used to fix lot sizes for batches when fixed batch sizes are desirable. Some limited tests at the detailed scheduling level suggest that even an order-launch approach with these lot sizes is quite successful. An equally important use of the models is to predict the lead times required to produce an order; this can be done using the expected queueing time for a batch given its route through the shop. Knowing these lead times permits the correct release of batches to the shop.

The implementation of such an approach would simply employ the lot sizes and lead times as inputs to a standard MRP system. The lot sizes should not be fixed; rather a range of say 0.8 to 1.5 times the lot size could be used as min-max limits. The advantage of this approach is that it is a top-down approach that requires very little administrative change or modification to existing systems. In addition, a shop-floor control system is not essential. It should be remembered that the models look at average characteristics of the shop and that these will change with the overall production mix. Thus, in principle, a detailed shop-floor system will improve performance; however, we conjecture that this improvement is not great especially relative to the cost of detailed control. Again, limited experiments suggest that it is no greater than 10 percent. A theoretical estimate is given by Kekre [1984].

The models can also be used in capacity and design decisions since they are essentially evaluation tools which predict performance. The parameters that might be considered are

- The number and size of machines at work centers,
- Overtime and shift policies,
- Routing of items (if alternatives are available),
- The choice of parts to be made in the cell or facility, and
- The operations or work centers that should be included in the cell.

In fact, the simulation has been used for many of these purposes for this cell. The analytical model has not, but in another paper [Karmarkar et al. 1984] we describe an application of Q-LOTS to capacity analysis in a different manufacturing system.

Summary

We have described a joint project between groups in industry and academia motivated by a common interest in the solution of certain manufacturing problems, although possibly with a differing sense of urgency. Our study provides strong support for the importance of lot-sizing techniques in shop performance and focuses attention on performance issues (lead time, WIP) that have been inadequately treated in the technical literature. Both parties have benefited; the analytical developments provide an understanding of the reasons for the observed phenomena as well as a fast numerical technique for analyzing such systems; the simulation provides validating evidence for the correctness of a complex model.

Acknowledgments

We appreciate the encouragement provided by John Barnes, Manager of Planning at the US Apparatus Division of Eastman Kodak Company and the support of Ed Sylvestre, Supervisor of the Analytical Services Division of Eastman Kodak, who created the initial opportunity for this collaboration. Support for development of Q-LOTS was provided by the Karres Group and by Case Hunter Inc.

APPENDIX

Using the notation in the text, an expression for the average waiting time in the system is developed assuming that M/M/1 model applied for this model,

λ = arrival rate = D/Q,

$\bar{x} = \frac{1}{\mu} =$ average processing time = $\tau + (Q/P)$,

$u = (D/P)$,

$\rho = \lambda\bar{x} = (D/Q)(\tau + Q/P) = (D\tau/Q) + (D/P)$,

T = the mean time in system.

The results for the M/M/1 model give

$$T = \frac{1}{\mu(1-\rho)} = \frac{\tau + (Q/P)}{1 - (D\tau/Q) - (D/P)}.$$

The stability condition $\rho < 1$ implies $(D\tau/Q) + (D/P) < 1$ which on rearrangement gives $Q > D\tau/(1 - (D/P))$, a lower bound on the lot size. A lower bound for T is given by

$$T \geqslant \frac{\tau}{(1-u)} + \frac{Q}{P(1-u)}$$

which is linear in Q, and is approached asymptotically as Q becomes large.

References

Burbidge, J. L. 1975, *The Introduction of Group Technology*, John Wiley and Sons, New York.

Buzacott, J. A. 1974, "On the optimal control of input to a job shop," Working Paper #74-004, Department of Industrial Engineering, University of Toronto, Toronto, Ontario.

Buzacott, J. A. and Shanthikumar, J. G. 1980, "Models for understanding flexible manufacturing systems," *AIIE Transactions*, Vol. 12, No. 4, pp. 339-350.

Karmarkar, U. S. 1983a, "Lot sizes, manufacturing lead times and utilization," Graduate School of Management, Working Paper No. QM8312, University of Rochester, Rochester, New York.

Karmarkar, U. S. 1983b, "Lot sizing and sequencing delays," Graduate School of Management, Working Paper No. QM8314, University of Rochester, Rochester, New York.

Karmarkar, U. S.; Kekre, S.; and Kekre, S. 1983, "Multi-item lot sizing and manufacturing leadtimes," Graduate School of Management, Working Paper No. QM8325, University of Rochester, Rochester, New York.

Karmarkar, U. S.; Kekre, S.; and Kekre, S. 1984, "Lot sizing in multi-item, multi-machine job shops," Graduate School of Management, Working Paper No. QM8402, University of Rochester, Rochester, New York.

Kekre, Sunder 1984, "The effect of the number of items on manufacturing lead time," Graduate School of Management, Working Paper No. QM8418, University of Rochester, Rochester, New York.

Koenigsberg, E. and Mamer, J. 1982, "The analysis of production systems," *International Journal of Production Research*, Vol. 20, No. 1, pp. 1-16.

Magee, J. F. 1956, "Guides to inventory policy, part II. Problems of uncertainty," *Harvard Business Review*, Vol. 34, No. 2 (March-April) pp. 103-116.

Sasser, W. E.; Clark, K.; Garvin, D.; Graham, M.; Jaikumar, R.; and Maister, D. 1982, *Cases in Operations Management*, Richard D. Irwin, Homewood, Illinois.

Seidmann, A. and Schweitzer, P. J. 1983, "Part selection policy for a flexible manufacturing cell feeding several production lines," Graduate School of Management, Working Paper No. QM8217 (October 1982, revised October 1983), University of Rochester, Rochester, New York.

Shanthikumar, J. G. and Buzacott, J. A. 1981, "Open queueing network models of dynamic job shops," *International Journal of Production Research,* Vol. 19, No. 3, pp. 255-266.

Solberg, J. J. 1977, "A mathematical model of computerized manufacturing systems," in *Proceedings of the Fourth International Conference on Production Research,* Tokyo, Japan.

Stecke, K. E. and Solberg, J. J. 1981, "Loading and control policies for a flexible manufacturing system," *International Journal of Production Research,* Vol. 19, No. 5, pp. 481-490.

Sugimori, Y.; Kusunoki, K.; Cho, F.; and Uchikawa, S. 1977, "Toyota production system and kanban system: Materialization of just-in-time and respect-for-human system," *International Journal of Production Research,* Vol. 15, No. 6, pp. 553-564.

Suri, R. 1983, "Robustness of queueing network formulae," *Journal of the Association for Computing Machinery,* Vol. 30, No. 3, pp. 564-594.

Zipkin, P. H. 1983, "Models for design and control of stochastic, multi-item batch production systems," Columbia Business School, Research Working Paper No. 496A, Columbia University, New York.

A letter from John C. Barnes, Manager, Planning, US Apparatus Division, Eastman Kodak Company, states: "We are now planning to establish lot sizes using the analytical model instead of the traditional EOQ approach. The confidence gained has allowed us to commit the manufacturing operation to accomplish a very substantial reduction in planned lead time and, therefore, inventory level. We anticipate . . . that our ability to fulfill schedule requirements will improve markedly as a result."