Proceedings of the 2004 Winter Simulation Conference
R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

OPTIMAL LOT-SIZING WITH CAPACITY CONSTRAINTS AND AUTO-CORRELATED INTERARRIVAL TIMES

S. T. Enns
Li Li

Dept of Mechanical and Manufacturing Engineering
University of Calgary
Calgary, AB T2N 1N4, CANADA

ABSTRACT

There have been recent advances in using queuing relationships to determine lot sizes that minimize mean flowtimes when multiple product types are being produced at capacity-constrained resources. However, these relationships assume lot interarrival times are independent, which is not the case in most manufacturing scenarios. This study examines the performance lot-sizing optimization relationships based on GI/G/1 relationships when lot interarrival times are auto-correlated. Simulation and response surface modeling are used to experimentally determine optimal lot sizes for a sample problem. The flowtimes for "optimal" lot sizes determined analytically are found to compare poorly with with the best flowtimes obtained experimentally. An approah is then developed that uses feedback during simulation to adjust parameters within queuing heuristics that support dynamic lot-size optimization. Performance using this approach compares well with the best performance obtained using the much more difficult experimental approach.

1 INTRODUCTION

The multi-item capacitated lot-sizing problem is important in batch production. Complex work flow configurations, changing demand levels and other factors mean that lot interarrival times are stochastic. When there are setup times, lot sizes that are too small will result in very high utilizations due to excessive machine time dedicated to setups. This will cause lot flowtimes to be higher than necessary. Lot sizes that are too large will mean that machines are dedicated to one part type for long periods of time. This means incoming lots of different part types will have to wait excessively long. This extra queue time also results in flowtimes higher than necessary. Therefore, a tradeoff exists even in the case of a single part type. In multiple part type scenarios there are also interactions effects between the different part types. Therefore finding the best combinations of lot sizes across all part types is difficult.

Analytical approaches to capacitated lot-size optimization have been developed using queuing relationships. However, these approaches all assume that lot interarrival times are independent. In other words, there is no auto-correlation in the arrival stream. In reality, this is not a good assumption. For example, if lots are being fed from one machine to another, the minimum interdeparture time from the source machine will be the lot processing time at the source machine. Under high utilization conditions, interdeparture time variability will be low and auto-correlation will be high. This means that the lot interarrival times to the destination machine will also be auto-correlated (i.e. they will not be independent). If there are several streams merging at the destination machine, lot interarrival time variability may increase but if each of the individual streams is auto-correlated the merged incoming stream will still not have independent interarrival times.

The objectives of this research are three fold. The first objective is to demonstrate the effects of auto-correlation and show why it cannot be ignored. The second objective is to develop a methodology for lot sizing that adjusts for auto-correlation effects. Dynamic performance feedback is used to facilitate this. The final objective is to demonstrate the feasibility and effectiveness of this new approach by comparing the results to those obtained through experimental methods.

2 LITERATURE REVIEW

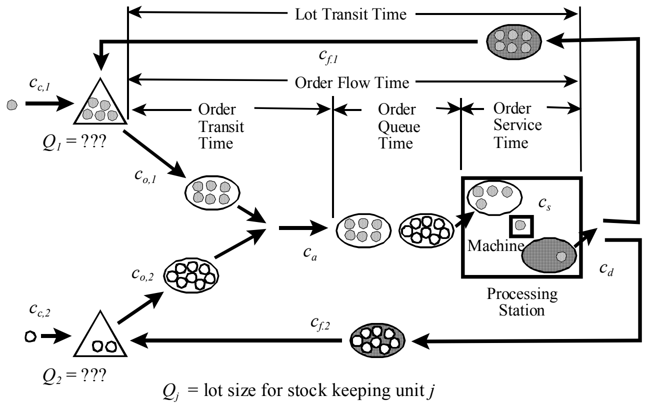
The single machine lot-sizing problem of most practical interest is one based on GI/G/1 queuing assumptions. However, it is well known that there is no closed form solution for queue time estimation. Most approximations are based on the mean and variance of the interarrival and service time distributions. Examples can be found in Whitt (1983) and Buzacott and Shanthikumar (1993). Although these approximations are generally presented in the context of having entities of single items in queue, they also apply when the entities are product lots.

Finding near, or approximately, optimal lot sizes to minimize flowtimes under GI/G/1 assumptions is important if we are to apply queuing analysis to real production problems, especially where networks of machines are involved. Lambrecht and Vandaele (1996) dealt with this problem for a single product type. Lot sizing considerations in their research included the time to accumulate lots as well as lot flowtimes at the machine. The solution procedure was based on a steepest decent search algorithm. A further contribution was the development of flowtime distributions. Lambrecht, Ivens and Vandaele (1998) extended the development of this lot-sizing approach, as part of a scheduling procedure called ACLIPS, to multiple product types moving through multiple machines. Fowler, et. al (2002) also investigated lot-size optimization in a multi-product, multi-stage facility by using queuing approximations and search techniques. Enns and Choi (2002) investigated lot-size optimization using GI/G/1 assumptions in an MRP environment. This study used a solution approach based on solving a set of first order differential equations. Auto-correlation was not explicitly considered in any of these studies.

3 MODELING A SIMPLE ENVIRONMENT

A simple problem environment was developed to experiment with lot-size optimization when interarrival times are correlated and there are capacity constraints. This problem environment is shown in Figure 1. There are two outlets at which individual customer orders are placed, with each outlet providing a different product. The customer order interarrival times at the outlets are assumed to have a coefficient of variation of $c_{c,j}$, where j is the stock keeping unit (SKU) or product type. These customer orders are batched until a quantity of Q_j orders have been received. The lot-size orders are then released and take some time to arrive at the capacity-constrained resource. This order placement delay has a coefficient of variation of $c_{o,j}$. When the orders are received, they are placed in queue at a single capacity-constrained machine with one processing stage. The merged arrival stream for the orders has a coefficient of variation of c_a. The machine manufactures the product to fill the lot-size orders on a on a first-come-first-serve (FCFS) basis. The processing of each lot-size order at the capacity constrained machine requires a setup time and a processing time for each unit in the order. The coefficient of variation of lot-size service times is designated as c_s. Once the lot-size order has been processed, the lot of units leaves the machine and is shipped to the outlet that placed the order. The coefficient of variation of lot interdeparture times is designated as c_d , while the coefficient of variation of lot transit times is shown as $c_{f,j.}$

The objective is to determine lot size quantities, Q_j, that will minimize the replenishment time, defined as the time from which a lot-size order is placed to the time it is received. If the lot sizes are too small, there will be too many setups incurred at the capacity-constrained machine. Utilization will be high and long queue times will result. If the lot sizes are too large, the machine will be committed to one item type for too long and other orders will have to wait longer than necessary, causing average flowtimes to increase. In this scenario, the lot delivery transit times are assumed to be independent of lot size and therefore have no effect on minimizing replenishment cycle times. The order placement delays do have an impact if these are considered to be stochastic. However, this impact is only on the queuing delay. As the variability of order placement delays increases, the variability of lot order arrivals to the queue will increase and this will cause queue times to increase. Therefore, optimal lot sizes are simply those that minimize the average weighted lot flowtimes at the capacity-constrained machine, defined as the sum of queue times and lot processing times.



Q_j = lot size for stock keeping unit j
$c_{c,j}$ = customer interarrival time coefficient of variation for j
$c_{o,j}$ = lot order delay time coefficient of variation for j
c_a = lot order interarrival time coefficient of variation
c_s = lot service time coefficient of variation
c_d = lot interdeparture time coefficient of variation
$c_{f,j}$ = lot delivery transit time coefficient of variation for j

Figure 1: Diagram of Experimental Scenario

This problem environment is designed to be simple. The intent is to allow analysis to be focussed on the performance effects of lot sizing with auto-correlated interarrival times. However, the problem environment can be easily extended to model more realistic scenarios. For example, lot order releases could be based on reorder points. In this case the reorder points would become another decision variable and performance would need to evaluated on both inventory levels, which are a function of the replenishment cycle times, and delivery performance measures.

In this research, the problems illustrated are based on the following set of assumptions. The mean demand rates at the item types at each outlet, D_1 and D_2, are assumed to be 44 and 50 units per period, respectively. The customer order interarrivals are described by a Poisson process, with each order quantity being 1. Therefore, the coefficients of variation for the order interarrival times, $c_{c,j}$, are 1. However, the lot-size order releases will have a lower coefficient of variation and will be auto-correlated if the lot size exceeds 1. The order placement delays are all assumed to be lognormal with a mean of 5 periods and a standard deviation of 1. Therefore, the $c_{o,j}$ values are 0.20.

The setup times, τ_1 and τ_2, are assumed to be 0.30 and 0.20 periods per lot, while the part processing rates, P_1 and P_2, are 120 and 140 units per period, respectively. The setup and processing times were considered to be deterministic since this simplifies the relationships later used for lot-size optimization. Relationships can also be used that treat setup and processing times as stochastic variables but adding this complexity does not contribute to the objectives of this research. The transit times for lot shipments from the manufacturer to the retail outlets are also all assumed to be lognormal with a mean of 5 periods and a standard deviation of 1. Therefore, the $c_{f,j}$ values are 0.20.

4 EXPERIMENTAL LOT-SIZE OPTIMIZATION

The first stage of analysis was to determine the optimal lot size combination, Q_1 and Q_2, experimentally using discrete-event simulation. A model was built using ARENA 5.0. (Kelton, et al., 2002) and experiments were run using the Central Composite Design (CCD) shown in Figure 2. The corner and axial points for this design were each run for two replications, while the center point was run for 10 replications. This resulted in a total of 26 runs. A warmup time of 100 periods was used to reach steady-state conditions and data collection continued for 40,000 periods in each run. Common random numbers were used as a variance reduction technique.

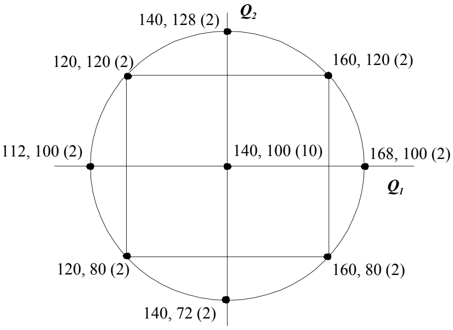


Figure 2: Central Composite Design

The average lot flowtime results at the capacity-constrained machine were analyzed using Design Expert® 6.0. A quadratic model, shown as Equation (1), fit the results nicely. Lack-of-fit was not significant at the 95% confidence level and the resulting model had an R^2 value of 93.44%.

$$W = 4.084 - 1.393(E-02)Q_1 - 2.269(E-02)Q_2 + 2.642(E-05)Q_1^2 + 6.749(E-05)Q_2^2 + 6.508(E-05)Q_1Q_2 \tag{1}$$

The response surface generated by this model is shown in Figure 3. It is obviously quite flat through some lines of orientation through the optimum, indicating there will be a variety of lot size combinations performing well. Figure 4 shows a contour plot of the surface around the optimal.

The optimizer in Design Expert® was used to determine the lot size combinations to minimize the mean flowtimes (Montgomery, 2001). These values were found to be 139 and 101, with a predicted minimum average flowtime of 1.97 periods per lot. This optimal lot size combination, was used in running five additional replications. The average flowtime was 1.966, the observed c_a value was 0.721, and the utilization was 0.918.

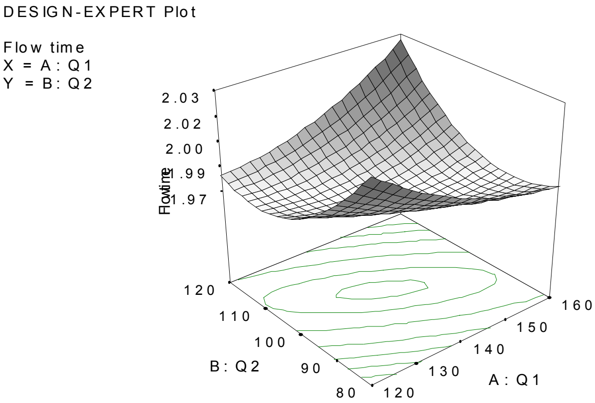


Figure 3: Response Surface for Flowtimes

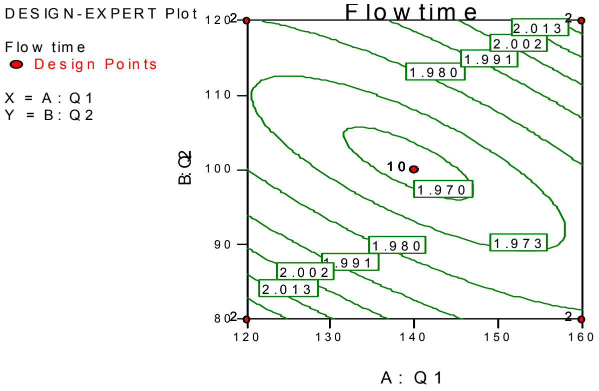


Figure 4: Contour Plot of Flowtimes

In addition, the actual lot interarrival time data was collected so that auto-correlation could be evaluated. Figure 5 shows a typical example of a correlogram created using Minitab 14 (Montgomery, 2001). It is obvious that the lot interarrival times are highly auto-correlated.

5 GI/G/1 LOT-SIZE OPTIMIZATION

Lot-sizing relationships to minimize mean lot flowtimes or queue times have been developed in previous research. However, these are based on restrictive assumptions about

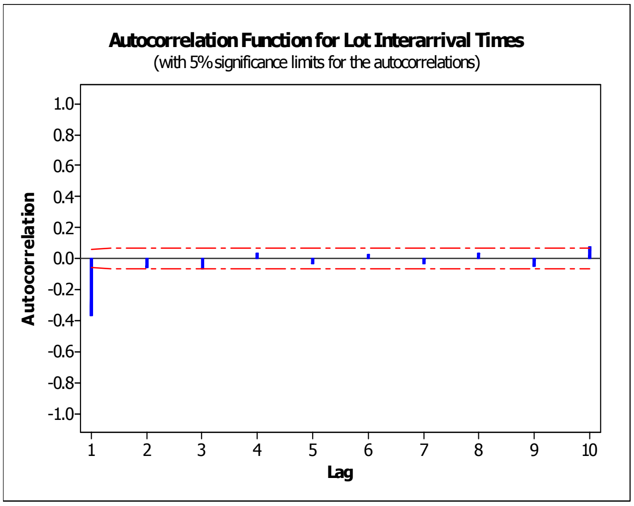


Figure 5: Correlogram of Lot Interarrival Times

the interarrival times. If general interarrival time distributions are used, they are usually based on the assumption of independent arrivals. This is clearly not the case in many applications involving lot sizes. Therefore, it is necessary to evaluate if using these relationships still provide reasonable flowtime estimates. Good estimates of flowtimes facilitate finding optimal lot sizes that minimize flowtimes.

When the lot interarrival time distribution is assumed to be general, it is usually satisfactory to describe it in terms of the first two moments, the mean and standard deviation. In this case, the following approximation is often suggested to estimate mean flowtimes, W, at a single machine (Whitt, 1983).

$$W_t = W_{q,t} + \bar{x}_t = \bar{x}_t \frac{\left(c_{a,t}^2 + c_{s,t}^2\right)}{2} \frac{\rho_t}{1-\rho_t} + \bar{x}_t \quad (2)$$

where W_q is the weighted mean time in queue, $\bar{x}$ is the weighted mean lot service time, c_a is the coefficient of variation for lot interarrival times, c_s is the coefficient of variation for lot service times and ρ is the machine utilization rate. The subscript t indicates values are for the current time, if the parameters of the system change through time. This relationship is based on steady-state GI/G/1 queuing assumptions, with interarrival times being independent.

When the entities in queue represent lots of parts, the weighted mean lot service time, including setup times, for n part types processed on the machine is given by the following,

$$\bar{x}_t = \frac{\sum_{j=1}^{n} \frac{D_j}{Q_{j,t}} \left[\tau_j + \frac{Q_{j,t}}{P_j}\right]}{\sum_{j=1}^{n} \frac{D_j}{Q_{j,t}}} \quad (3)$$

where j is the part type index, D_j is the demand rate, Q_j is the part type lot size, P_j is the part processing rate, and τ_j is the lot setup time.

The utilization rate, including setup times, is then given by the following.

$$\rho_t = \sum_{j=1}^{n} \left[\frac{D_j}{Q_{j,t}}\left(\tau_j + \frac{Q_{j,t}}{P_j}\right)\right] \quad (4)$$

This value is constrained to be less than 1 under steady-state conditions.

If it is assumed the lot setup times and part processing times are deterministic, the squared coefficient of variation for the lot service times is expressed as follows,

$$c_{s,t}^2 = \frac{\sum_{j=1}^{n} \frac{D_j}{Q_{j,t}} \left[\tau_j + \frac{Q_{j,t}}{P_j}\right]^2 \left(\sum_{j=1}^{n} \frac{D_j}{Q_{j,t}}\right)^{-1}}{\bar{x}_t^2} - 1 \quad (5)$$

The objective is to solve for the lot sizes, Q_j^*, that will minimize the lot flowtimes. This can be done by taking the partial differential of Equation (2) with respect to each Q_j, setting the resulting set of equations equal to 0 and then solving them simultaneously for Q_j^* (Enns and Choi, 2002). The resulting set of equations is shown as Equation (6).

$$\frac{\partial(W_t)}{\partial Q_{j,t}} = AA + BB + CC + DD = 0 \quad (6)$$

where,

$$AA = \frac{\left\{\left[\frac{2\rho_t\left(\frac{-D_j\tau_j}{Q_{j,t}^2}\right)\sum_{j=1}^{n}\frac{D_j}{Q_{j,t}} + \rho_t^{\,2}\frac{D_j}{Q_{j,t}^2}}{\left(\sum_{j=1}^{n}\frac{D_j}{Q_{j,t}}\right)^2}\right]\left(c_{a,t}^2 - 1\right)\right\} 2\left(1-\rho_t\right)}{\left[2\left(1-\rho_t\right)\right]^2},$$

$$BB = \frac{\left[\frac{-D_j}{Q_{j,t}^2}\left(\tau_j + \frac{Q_{j,t}}{P_j}\right)^2 + 2\frac{D_j}{Q_{j,t}P_j}\left(\tau_j + \frac{Q_{j,t}}{P_j}\right)\right] 2\left(1-\rho_t\right)}{\left[2\left(1-\rho_t\right)\right]^2},$$

$$CC = -\frac{\left\{\left[\frac{\rho_t^{\,2}}{\left(\sum_{j=1}^{n}\frac{D_j}{Q_{j,t}}\right)}\left(c_{a,t}^2 - 1\right) + \sum_{j=1}^{n}\frac{D_j}{Q_{j,t}}\left(\tau_j + \frac{Q_{j,t}}{P_j}\right)^2\right] 2\left(\frac{D_j\tau_j}{Q_{j,t}^2}\right)\right\}}{\left[2\left(1-\rho_t\right)\right]^2},$$

$$DD = \frac{\left[\frac{D_j}{Q_{j,t}P_j} - \frac{D_j}{Q_{j,t}^2}\left(\tau_j + \frac{Q_{j,t}}{P_j}\right)\right]\sum_{j=1}^{n} \frac{D_j}{Q_{j,t}} + \rho_t \frac{D_j}{Q_{j,t}^2}}{\left(\sum_{j=1}^{n} \frac{D_j}{Q_{j,t}}\right)^2} .$$

This set of equations can be readily solved using various software packages, such as the Solver in Excel®.

When the simulation was run with the optimal lot sizes based on the response surface model developed previously, a c_a value of 0.721 was observed. If the this value is used in Equation (6), the best lot sizes are found to be 159 and 158, with a predicted flowtime of 4.085. These lot sizes are significantly larger than the optimal lot sizes of 139 and 101 obtained experimentally in the previous section. Therefore, we conclude that violating the assumption of independence cannot be ignored when optimizing lot sizes with auto-correlated lot interarrival times.

While it is possible to obtain optimal lot sizes experimentally, this is impractical in real batch production environments. As well, the present analysis reveals it is insufficient to simply observe the lot interarrival coefficient of variation, c_a, and compute the optimal lot sizes using GI/G/1 relationships. Therefore, development of a practical approach that uses queuing relationships but takes auto-correlation into account is desirable.

6 DYNAMIC LOT-SIZE OPTIMIZATION

It is difficult to deal with the problem of auto-correlated data analytically. Auto-regressive models can be used to analyze the behaviour but queuing relationships that allow lot-size optimization with correlated data have not been developed. Therefore, an alternative approach is investigated. In this approach it is assumed the GI/G/1 relationships might prove satisfactory for lot-size optimization if c_a could be replaced by some other suitable parameter which is not actually the coefficient of variation of the interarrival times. In other words, it is assumed the form of Equation (2) is suitable and that Equation (6) could be used to find near optimal lot sizes if appropriate adjustments could be made to the parameter related to lot interarrival times, c_a. This adjusted parameter will be designated as c'_a.

The strategy is to use a dynamic feedback approach, implemented in a test bed where ARENA® is linked to Excel® through the use of Visual Basic for Applications® (VBA). This involves taking the terms in Equation (2) related to $W_{q,t}$, replacing c_a with c'_a, and rearranging them as shown in Equation (7).

$$c'_{a,t} = \sqrt{\frac{2W_{q,t}\left(1-\rho_t\right)}{\bar{x}_t \rho_t}} - c_{s,t} \qquad (7)$$

A dynamic estimate of queue time, $W_{q,t}$, can be obtained using exponential smoothing. Every time a lot is completed at the resource, the observed queue time for the lot is used to update $W_{q,t}$. This value is then fed over to the Excel® spreadsheet program where $W_{q,t}$ is plugged into Equation (7), along with the current values of $\bar{x}_t, \rho_t$ and $c_{s,t}$. The value of $c'_{a,t}$, which might be termed the implied lot interarrival time coefficient of variation, is then solved for. In other words, this adjusted coefficient of variation value is the one that would result in the observed flowtimes, given independent lot interarrival times. This implied lot interarrival time coefficient of variation, $c'_{a,t}$, is then used in solving for the current optimal lot sizes, $Q^*_{j,t}$, using Equation (6). As well, the values of $\bar{x}_t$, ρ_t, and $c_{s,t}$ are updated, based on the new $Q^*_{j,t}$ values. Equations (3)-(5) are used for this purpose. Finally, the new lot sizes are dynamically fed back to the ARENA® simulation program to determine the order quantity for any new order releases.

This feedback approach for dynamic lot sizing was applied to the previous problem. A smoothing constant of 0.05 was used and five replications were run, using a warmup period of 100 and data collection over 40,000 time units.

The average c'_a value used in lot size computations was 0.355, which is considerably less than the observed c_a of 0.695. Figure 6 shows a typical plot of the dynamic c'_a values through time. The time-average values of $Q^*_{i.t}$ were 120.65 and 139.95, respectively. Figure 7 shows a typical plot of the dynamic lot sizes. The average lot queue times, $W_{q,t}$, and lot flowtimes, W_t, were 0.767 and 2.018. The average utilization was 0.906.

The observed mean lot flowtimes using the dynamic lot sizing approach based on feedback were very close to the mean lot flowtimes obtained using the optimal lot sizes determined experimentally. The flowtime value of 2.018 obtained using dynamic lot sizing is around 2.6% higher than the flowtime value of 1.966 obtained using static lot sizes of 139 and 101. Therefore, it can be concluded that the dynamic lot sizing approach works well and is suitable to situations in which lot arrivals are auto-correlated.

7 CONCLUSIONS

The dynamic lot sizing approach presented in this paper is practical to implement and yields lot sizes that come very close to minimizing lot flowtimes. The key requirements are the ability to obtain the necessary flowtime feedback by monitoring shop floor performance and the ability to implement the appropriate optimization procedure.

More research is required to test the robustness of the approach and to apply it to multiple station scenarios. As well, it appears this approach is well suited to situations where demand may be non-stationary. Further testing is required to ensure the approach automatically adjusts the relative lot-sizing relationships adequately to accommodate time-varying demand levels.

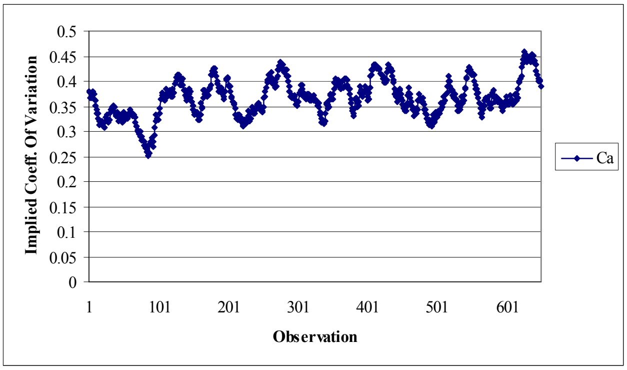


Figure 6: Implied Interarrival Time Coeff. of Variation

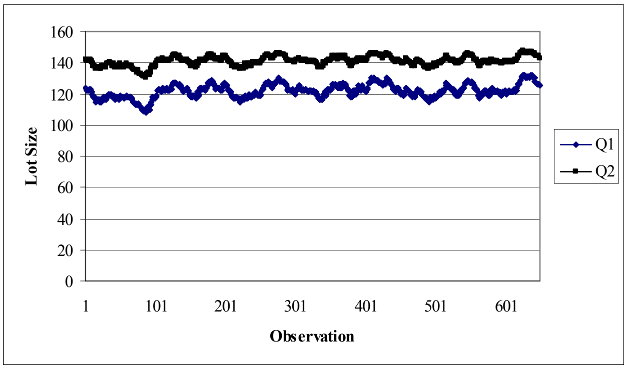


Figure 7: Dynamic Lot Sizes

ACKNOWLEDGMENTS

This research was supported in part by a grant from the National Science and Engineering Research Council (NSERC) of Canada.

REFERENCES

Buzacott, J.A., and G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*, Englewood Cliffs, New Jersy: Prentice-Hall.

Enns, S.T., and S. Choi. 2002. Use of GI/G/1 queuing approximations to set tactical parameters for the simulation of MRP systems, *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J.L. Snowdon and J. M. Charnes, 1122-1129. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Fowler, J.W., N. Phojanamongkolkij, J.K. Cochran and D.C. Montgomery. 2002. Optimal batching in a wafer fabrication facility using a multiproduct G/G/c model with batch processing, *International Journal of Production Research*, 40: 275-292.

Kelton, W.D., R.P. Sadowski and D.A. Sadowski. 2002. *Simulation with Arena*, 2nd Ed., New York: McGraw-Hill.

Lambrecht, M.R., and N.J. Vandaele. 1996. A general approximation for the single product lot sizing model with queueing delays, *European Journal of Operational Research*, 95: 73-88.

Lambrecht, M.R., P.L. Ivens, and N.J. Vandaele. 1998. ACLIPS: A Capacity and Lead Time Integrated Procedure for Scheduling, *Management Science*, 44: 1548-1561.

Montgomery, D.C. 2001. *Design and Analysis of Experiments*, 5th Ed., New York: John Wiley & Sons.

Whitt, W. 1983. The Queueing Network Analyzer*, The Bell Systems Technical Journal*, 62 (9): 2779-2813.

AUTHOR BIOGRAPHIES

SILVANUS T. ENNS is an Associate Professor at the University of Calgary, Canada. He received his PhD from the University of Minnesota. His research interests lie in the development of algorithms to support enhanced MRP performance as well as various aspects of job shop, batch production and supply chain modeling and analysis. His email address is `<enns@ucalgary.ca>`.

LI LI is completing her MSc degree at the University of Calgary, Canada. She received her BSc degree from Lanzhou University, People's Republic of China, in 1996. Her email address is `<lilil@ucalgary.ca>`.