Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

CAPACITY AND BACKLOG MANAGEMENT IN QUEUING-BASED SUPPLY CHAINS

Edward G. Anderson

McCombs School of Business
Management Department
The University of Texas at Austin
Austin, TX 78712, U.S.A.

Douglas J. Morrice

McCombs School of Business
MSIS Department
The University of Texas at Austin
Austin, TX 78712, U.S.A.

ABSTRACT

In this paper, we model and analyze a type of two-stage serial supply chain often found in service sector and make-to-order manufacturing industries. The chain holds no finished goods inventory at either stage. Rather, processing occurs only after an order is received and backlogs are managed solely by adjusting capacity. We model this supply chain using a tandem queuing model. Our analysis considers the impact of changes in first stage lead-time and capacity adjustment time on backlog, waiting time, and capacity variances at both stages. The results can be used to support the argument for better coordination across stages in these types of supply chains.

1 INTRODUCTION

Supply chains, particularly in service sector industries such as insurance, financial services, and health care, and in make-to-order manufacturing, can often be modeled using queuing models. In this paper, we present and analyze the properties of one such model. In particular, we consider a two-stage serial supply chain that holds no finished goods inventory at either stage. Rather, processing occurs only after an order is received and backlogs are managed solely by adjusting capacity. Capacity in our queuing-based model is represented by the number of servers at each stage.

Although the model is an abstraction of reality, it is contains general characteristics that we have observed in practice in the aforementioned industries. For example, average lead-time is often managed to some target. This is achieved by adjusting capacity to manage backlog so that the target lead-time is maintained. Changes in backlog trigger capacity changes but the change in capacity may not be instantaneous because, for example, capacity may be personnel who may take time to be brought up to speed or released.

Since we are changing the capacity to achieve a target average lead-time, the main statistics of interest will be variances. More specifically, we will analyze variances in the waiting times, backlogs (or queue lengths), and capacities for both stages for different model parameter levels. In service sector industries, consistency tends to be highly valued by the end customer (see, for example, Fitzsimmons and Fitzsimmons 1998). This is another reason why we focus on the variance metrics.

Although analytical results exist in the queuing theory literature for certain models that are designed to control backlog via capacity adjustment (see, for example, Rosberg et al. 1982 and Chen et al. 1994), no analytical results exist for the model and the metrics that we consider other than the related work of Anderson and Morrice (1999, 2000, 2001). However, even these three reference do not contain analytical results for the waiting time variances which are of great importance to customer service. For this reason, we resort to simulation. We will relate our results to the results in these references later in the paper.

2 DESCRIPTION OF THE SIMULATION MODEL

Figure 1 contains a picture our Arena model. The top flow diagram represents the main simulation model of the supply chain. Jobs arrive, are processed through two stages (each with infinite buffer capacity), and then depart from the system. The arrival process is intended to represent aggregate demand viewed from a supply chain perspective. Thus, jobs arrives in a batch at the beginning of each time period where the batch size is normally distributed with mean μ and standard deviation σ (the result is rounded to the nearest integer). The additional create nodes at each server initially load the queues at each server to a backlog level that matches the target lead-time (we will refer to this as the target backlog). Note, it is not uncommon, especially in service industries, to have a positive target backlog. Positive backlogs are viewed as a way to keep workers busy and often bolster the financial evaluations of service firms because having some backlog can be an indicator of future profitability.

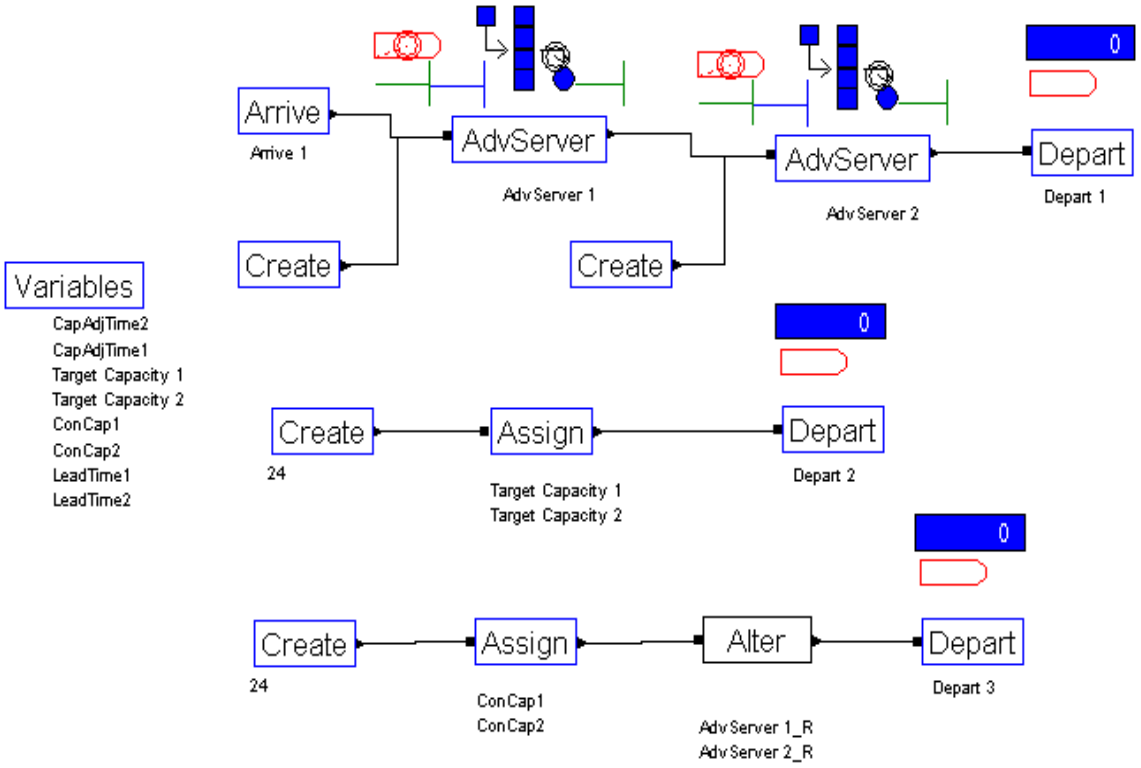

Figure 1: Arena Simulation Model

The middle flow diagram in Figure 1 is a logical process designed to calculate, periodically, the *target* (or *desired*) capacity at each stage based on the backlog and the target lead-time. More specifically, at time t (≥ 0), the target capacity at stage i (i.e., the variable Target Capacity i) (i=*1,2*) equals,

$$\frac{(\#\text{ in Queue at Stage } i) + (\#\text{ in Service at Stage } i)}{\text{LeadTime}i} \quad (1)$$

The variable LeadTime$i$ is the target lead-time at stage i.

The bottom flow diagram in Figure 1 is a logical process that periodically calculates the *actual* capacity at each stage. This is a two-step calculation in which a continuous capacity level is calculated at time t for stage i (i.e., the variable ConCap$i$) to equal,

$$\text{ConCap}i + \left(\frac{\text{Target Capacity } i - \#\text{ of Servers at Stage } i}{\text{CapAdjTime}i} \right) \quad (2)$$

Roughly speaking, the variable CapAdjTime$i$ is the average time it takes to adjust the actual capacity (i.e., the # of Servers at Stage i) to the target capacity. Thus, we will refer to this parameter as the capacity adjustment time. More precisely, (1/CapAdjTime$i$) is the amount by which the gap between target capacity and actual capacity closes each time (2) is evaluated. Expression (2) is designed to model the property that the actual capacity lags target capacity. Since the capacity change can only be integer, the ALTER block alters the # of Servers at Stage i by the difference

$$\text{ConCap}i - \#\text{ of Servers at Stage } i$$

rounded to the nearest integer value.

3 DESIGN OF EXPERIMENTS

The base time unit used in the simulation model is hours but most state changes occur on a daily basis. Again, this is sufficient to represent the aggregate behavior from a supply chain perspective. It is also sufficient to represent supply chain level decision making. For example, supply chain level staffing decisions are not likely to be done more than once per day. To mimic such daily decisions, we recalculate capacities (i.e., (1) and (2)) every 24 hours.

For all simulation scenarios consider, we assume that the demand arrives in batches once per day at the beginning of each day. The mean batch size is μ is set at 20 per day. Regardless of the value we choose for μ, we manipulate capacity in order to match supply with demand and to maintain a target lead-time. Thus, the specific value that we have chosen for μ is relatively unimportant. What is important is the variation about the target levels.

We set the variance in the demand so that the coefficient of variance is 10%. This is a low to moderate value that is common in aggregated data. The theoretical results in Anderson and Morrice (2001) suggest that variance in the demand is a scaling constant in the variances of the backlog and capacity. This follows since the system is approximately linear away from the boundary level of zero backlog. Thus, the same general patterns appear in the variance results from the simulation for different levels of the demand variance. Larger demand variance just masks these patterns in more noise.

The initial capacity, i.e., the initial number of servers, is set to 20. Each job takes 24 hours to process (processing time is assumed to be deterministic). In other words, supply is set equal to average demand.

The two main parameters of interest in this study are the target lead time and the capacity adjustment time at stage 1. We focus on the first stage parameters only because they impact both stages. We use these results to draw conclusions about the necessity for coordination of management policies across different stages of a supply chain.

The base scenario for target lead-time is 10 days of backlog at each stage; the base scenario for capacity adjustment time is 20 days at each stage. It is more common that capacity adjustment is longer than the target lead time especially when the capacity resources are people. This is reflected in our base scenario. Holding all other parameters fixed, we will vary target lead-time on stage 1 from 10 to 30 days in increments of 10 days. We do the same for the capacity adjustment time at stage 1, holding all other parameters at the base case levels.

After initially loading the queue at each service stage to a backlog that matches the target lead-time, the predominant initial effects can be mitigated by collecting observations after the initially loaded jobs have exited the system. For the worst case, with target lead time of 30 days at the first station and a target lead-time of 10 days at the second station, all initially loaded jobs will have exited the system by 960 hours. Thus, we start collecting statistics after 1000 hours (the truncation point) and run the simulation for a total of 20000 hours (i.e., 19000 hours of data are used to estimate the variance statistics). Time series plots reveal that the data look stable after the truncation point. To be able to compare the results statistically, we make 30 replications of each scenario.

4 ANALYSIS

The results from the five different scenarios that we consider are given in Table 1. For each scenario label, "LT" stands for target lead-time and the two numbers that follow stand for the target lead-times (in days) for stages 1 and 2, respectively. The letters "CAT" in the scenario label represent capacity adjustment time and the two numbers that follow this term are the capacity adjustment times (in days) for stages 1 and 2, respectively. For each scenario, we calculate the average of the standard deviations from 30 replications for the number in queues at stages 1 and 2, the waiting time in the queues at stages 1 and 2, and capacity at stages 1 and 2. Invoking the central limit theorem, we compute the halfwidth ("HW") for a t-distribution based confidence interval for each average in each scenario.

Table 1 results can be summarized as follows:

1. As the target lead-time at stage 1 increases:
   a. The standard deviation of the number in queue 1 increases significantly (confidence intervals do not overlap).
   b. The standard deviation of the number in queue 2 decreases significantly.
   c. The standard deviation of the waiting time in queue 1 increases significantly.
   d. The standard deviation of the waiting time in queue 2 decreases significantly.
   e. The standard deviation of capacity at stage 1 decreases significantly.
   f. The standard deviation of capacity at stage 1 decreases significantly.
2. As the capacity adjustment time as stage 1 increases:
   a. The standard deviation of the number in queue 1 increases significantly.
   b. The standard deviation of the number in queue 2 does not change significantly (overlapping confidence intervals).
   c. The standard deviation of the waiting time in queue 1 increases significantly.
   d. The standard deviation of the waiting time in queue 2 does not change significantly.
   e. The standard deviation of capacity at stage 1 does not change significantly.
   f. The standard deviation of capacity at stage 2 does not change significantly.

The statistically significant cases in 1a, b, e, f and 2 a, c match with the analytical results in Anderson and Morrice (2001). In particular, these results show that reducing target lead-time at stage 1 reduces queue length (or backlog) variance at stage 1 but increases capacity variance resulting in a variance trade-off at stage 1. Furthermore, the reduction in stage 1 target lead-time increases backlog and capacity variance at stage 2 resulting in a variance trade-off across the two stages. The later trade-off indicates the need for coordination across the stages of the supply chain.

The significant results in cases 1c, d support a conjecture in Anderson and Morrice (2001) that standard deviations in the queue lengths move in the same direction as the standard deviations of the waiting times. While this may seem to be an intuitive result since queue lengths and waiting times often move in the same direction, it is

Table 1: Statistics for Stage 1 Target Lead-Times and Capacity Adjustment Times

Scenario	Statistics	Number in Que 1	Number in Que 2	Wait Time in Que 1	Wait Time in Que 2	Stage 1 Capacity	Stage 2 Capacity
LT10_10 CAT20_20	Average of StDev	7.27	7.25	10.61	9.11	0.52	0.56
	HW of 95% CI	0.33	0.64	0.22	0.54	0.02	0.03
LT20_10 CAT20_20	Average of StDev	8.61	4.40	11.08	6.11	0.42	0.42
	HW of 95% CI	0.34	0.33	0.20	0.36	0.01	0.02
LT30_10 CAT20_20	Average of StDev	9.66	3.52	11.80	5.12	0.38	0.38
	HW of 95% CI	0.37	0.28	0.21	0.36	0.02	0.02
LT10_10 CAT10_20	Average of StDev	6.06	6.71	8.89	9.17	0.55	0.52
	HW of 95% CI	0.21	0.42	0.10	0.36	0.02	0.02
LT10_10 CAT30_20	Average of StDev	8.30	7.39	11.75	8.91	0.51	0.58
	HW of 95% CI	0.45	0.72	0.32	0.56	0.02	0.04

not immediately evident because capacity is changing in order to maintain a certain target lead-time at each stage.

5 CONCLUSIONS

Queuing models and simulation are useful for analyzing the complex behavior of certain service and make-to-order manufacturing supply chains. As part of our future research, we plan to consider a number of extensions to this work. In particular, we plan to extend these results to more complex supply chain networks, include additional stochastic elements in the model, and analyze more complex decision rules for controlling the supply chain.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Red McCombs Graduate School of Business at the University of Texas at Austin for this research.

REFERENCES

Anderson, E. G. and D. J. Morrice. 2001. Capacity and Backlog Management in Service-Oriented Supply Chains. Working Paper, The University of Texas at Austin.

Anderson, E. G. and D. J. Morrice. 2000. A Simulation Game for Service-Oriented Supply Chain Management *J. of Production Oper. Man.* **9** 40-55.

Anderson, E. G. and D. J. Morrice. 1999. A Simulation Model to Study the Dynamics in a Service-Oriented Supply Chain. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nemhard, D. T. Sturrock, G. W. Evans, 742-748. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Chen, H., P. Yang, D.D. Yao. 1994. Control and Scheduling in a Two-station Queueing Network: Optimal Policies and Heuristics. *Queueing Sys.* **18** 301-332.

Fitzsimmons, J.A., and M.J. Fitzsimmons 1998. *Service Management: Operations, Strategy, and Information Technology*. Boston: Irwin/McGraw-Hill.

Rosberg, Z., P.P. Varaiya, J.C. Walrand. 1982. Optimal Control of Service in Tandem Queues. *IEEE Trans. on Auto. Control* **27** 600-610.

AUTHOR BIOGRAPHIES

EDWARD G. ANDERSON is an Assistant Professor of Operations Management at the University of Texas McCombs School of Business. He received his doctorate from the Massachusetts Institute of Technology and his bachelor's degree in electrical engineering and history from Stanford University. His research interests include supply chain management, outsourced product development, knowledge management, and system dynamics. He has published articles in such journals as *Management Science, Production and Operations Management,* and *The Systems Thinker*. Dr. Anderson won the prestigious Wickham Skinner Early-Career Research Award from the Production and Operations Management Society. He sits on the editorial review board of *Production and Operations Management*. Professor Anderson has consulted with Ford, Dell, Hewlett-Packard, Frito-Lay, and Atlantic-Richfield. Prior to his academic work, he was a product design engineer at the Ford Motor Company, from which he was granted three U.S. patents. His email address is <Edward.Anderson@bus.utexas.edu>.

DOUGLAS J. MORRICE is an Associate Professor in the MSIS Department at The University of Texas at Austin. He has a BA, Honours in Operations Research from Carleton University in Ottawa, Canada. He holds an M.S. and a Ph.D. in Operations Research and Industrial Engineering from Cornell University. His research interests include operations simulation design, modeling, and analysis. Dr. Morrice is a member of INFORMS, POMS, and CLM. He served as the Secretary for the INFORMS College on Simulation (1994-1996) and was Co-Editor of the Proceedings of the 1996 Winter Simulation Conference. His email address is <morrice@mail.utexas.edu>.