Using Simulation to Understand and Optimize a Lean Service Process

Kumar Venkat
Surya Technologies, Inc.
4888 NW Bethany Blvd., Suite K5, #191
Portland, OR 97229
kvenkat@suryatech.com

Wayne W. Wakeland
Systems Science Ph.D. Program
Portland State University
Portland, OR 97207

Keywords: Discrete event simulation, lean service, optimizing lean systems.

Abstract

This paper describes the application of discrete event simulation to understand and optimize a lean service process. Simulation is being used increasingly in the design and improvement of lean manufacturing systems. We now apply simulation to the emerging notion of lean service. We use the case study of a lean auto repair facility to demonstrate the significant role that simulation can play in the design of a cost-effective system. This lean service system eliminates queues by carefully scheduling appointments. A consequence of this type of system is that customers may sometimes need to wait for a considerable time outside the system before the start of their appointments. Our simulations show that the "time to appointment" can be optimized in conjunction with other metrics such as utilization of repair technicians and work in process. Simulation can clarify the exact nature of the tradeoff between customer satisfaction and cost-effective delivery of service. Our simulations also show that perturbations introduced by customers arriving late to their appointments can be absorbed with minimal impact provided there is some slack in the system. This finding may help to ameliorate one of the primary concerns regarding the lean service model.

INTRODUCTION

Lean thinking [1] is a systematic approach to developing business processes with the aim of doing more with less while coming as close as possible to providing customers exactly what they want. It is already the dominant paradigm in manufacturing today. It provides a way to specify value, determine the best sequence for value-creating steps, perform these activities without interruption when a customer requests them, and continually improve the process.

The key to lean manufacturing is to compress time by eliminating waste and let customers pull the product as needed [1]. A number of standard lean manufacturing tools exist, including value-stream mapping. In the last several years, simulation has emerged as a complementary tool for the design and improvement of lean manufacturing processes [2, 3].

The complexity and high cost of modern manufacturing systems necessitate the use of formal models of the system to support management decisions [4]. Discrete event simulation models are often needed for a detailed performance evaluation of a complex manufacturing system.

Simulation is a technique for modeling dynamics that can augment static value-stream analysis [5]. For example, simulation can be used to estimate the effectiveness of alternative configurations of a lean business process prior to actual implementation.

Lean concepts are now beginning to be applied to service activities [6, 7]. Going beyond lean production, "lean consumption" [8] calls for solving the customer's problem completely by ensuring that all products and services work well together without wasting the customer's time. The idea is to increase profits by delivering exactly what customers want, when and where they want it. Performance and productivity are measured from the customer's perspective. Lean service processes have been implemented in applications such as writing insurance policies, providing technical support for computer users, and car repair.

Lean service differs from lean manufacturing in one key respect. While customers provide the pull that

activates both kinds of lean systems, customers are more intimately involved in many lean service processes. It is expected that customers will quickly learn their role in lean service, and will embrace the opportunity to better address their needs [8]. However, the task of evaluating and optimizing the performance of these systems can be more complicated than in the manufacturing case.

Our goal in this work is to demonstrate the significant role that simulation can play in the design of a lean service system. As lean concepts take hold in service industries, we believe that simulation will become increasingly useful for understanding of the processes and optimizing performance. Simulation can often clarify the exact nature of the tradeoff between customer satisfaction and cost-effective delivery of service, and allow the service provider to choose the right level of resources.

PROBLEM DEFINITION

We use a published case study of a car repair facility as an example of lean service and lean consumption [8]. GFS, a Portuguese automobile dealer group, has implemented a lean car repair process by removing many wasteful steps. The lean techniques include prediagnosing every car repair whenever possible, scheduling jobs to eliminate queues, and standardizing repair processes. Customers and vehicles move faster through the lean system. The reduced waste and increased speed have translated to a 30 percent reduction in the company's cost per repair. Customers have also benefited from reduced prices and less wasted time. Most repair jobs are done right the first time.

Figure 1 illustrates the lean car repair process from the perspective of serving each customer. When a customer calls to make an appointment, a service consultant attempts to prediagnose the problem by phone, creates a repair plan, and orders parts. When the customer arrives later at the appointed time, a service consultant confirms the diagnosis. The customer can typically authorize the repair right then, and leave with a loaner car. When the repair is completed, the customer can return the loaner and drive home in his or her own car. The workflow has been smoothed by carefully scheduling customer arrivals, separating jobs according to their complexity, and delivering prekitted parts and tools to technicians just as needed.

The case study in [8] presents the ideal scenario for the lean car repair process. Our purpose in this study is to look beyond the ideal case and evaluate the performance of the system under various resource levels and unanticipated events. The process as shown in Figure 1 is a slightly modified version of the original process, reflecting some of the issues that we want to study.

In Figure 1, boxes enclosed in solid lines are actual steps in the original lean process. Each step includes an average processing or elapsed time as specified in [8]. Boxes enclosed in dashed lines represent delays where no work is done. On the customer side, there is a waiting time until the start of the appointment. This is typically unavoidable and occurs outside the system. The customer waits again outside the system, with use of the loaner car, for the actual repair to be completed. On the provider side, unexpected delays can occur if the customer arrives late to the appointment or if parts are not delivered on time.

Ideally, neither the customer nor the car will have to wait within the repair facility, but queues can form at various points if the workflow is disturbed and becomes uneven for any reason. We show two possible places in the repair process where queues can form as a result unexpected delays.

The horizontal double arrows in Figure 1 indicate points of contact between the customer and the provider. This illustrates how customers tend to be deeply involved at various points in a lean service process.

A customer's service experience is influenced by two key performance metrics (assuming that the repair job itself is done right):

- Time to appointment (how long the customer has to wait after the phone call for the appointed time).
- Repair completion time (how long it takes to complete the repair after the customer's car has been dropped off).

For the provider, we have identified two additional performance metrics:

- Utilization of repair technicians (average fraction of repair technicians that are busy).
- Number of loaners out at any time.

Time to appointment and repair completion time should be as small as possible. Both of these metrics contribute to customer satisfaction.

Utilization of repair technicians should be as high as possible, since each technician, along with his or

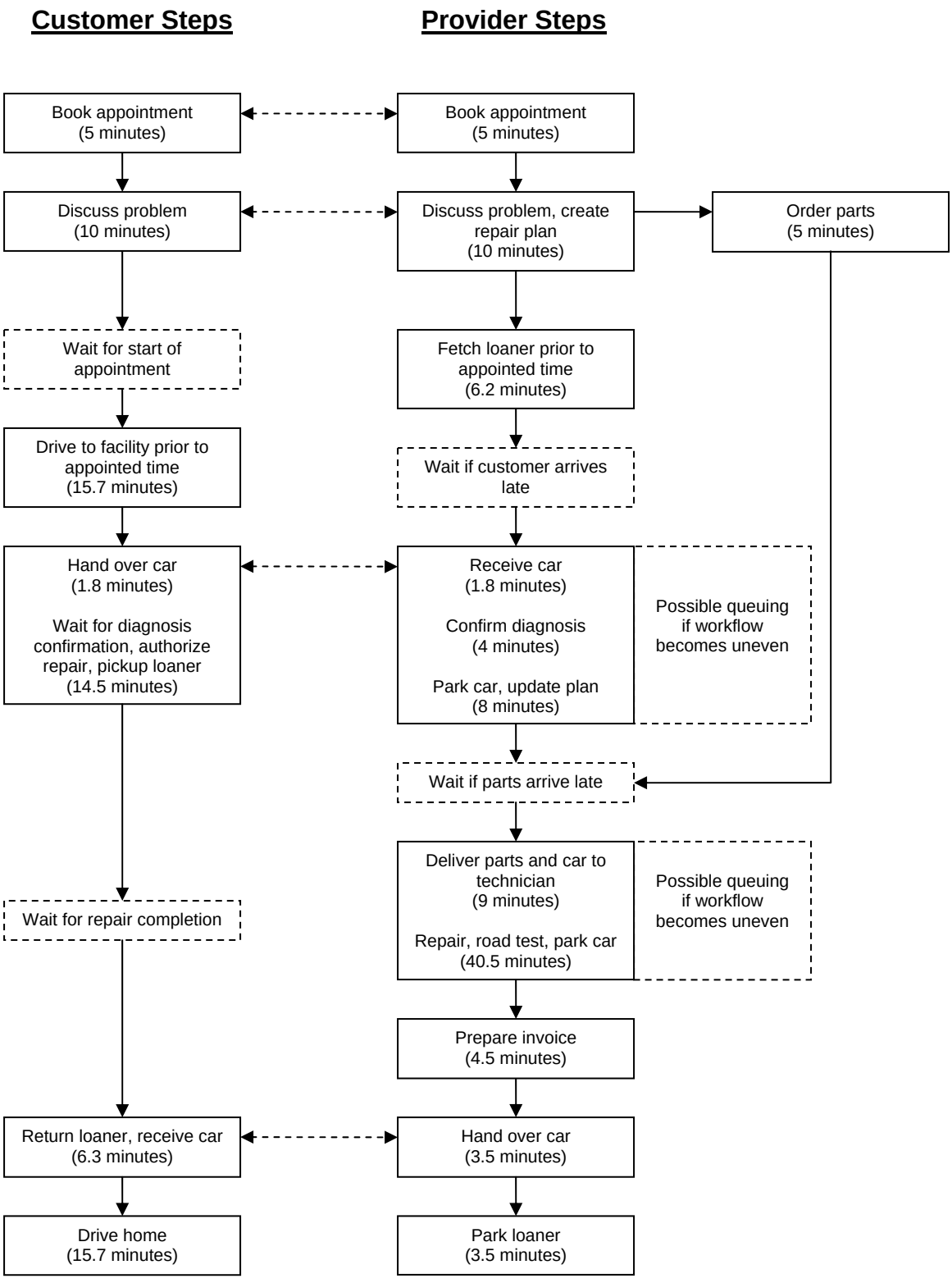

Figure 1. Lean car repair process (adapted from [8])

her equipment and space, is a resource that the provider must pay for. Although the repair facility uses other staff and resources, such as service consultants, we focus our attention on repair technicians since they are the sole creators of customer value.

The number of loaners out should generally be as small as possible (other things being equal), since there is a cost associated with each free loaner provided to a customer. The number of loaners out also quantifies work in process, since each customer gets a loaner.

Our problem in this study is two-fold:

- Given an expected rate at which appointment requests arrive, identify the right level of the critical resource, such that resource utilization is quite high and customer satisfaction is also likely to be high.
- Explore the performance impact of perturbations such as late arrival of customers and late delivery of parts.

METHOD

We have converted the process flowchart in Figure 1 into a discrete event simulation model using the Arena simulation software [9]. Since published information about this repair process is limited, we have made a number of reasonable assumptions in order to proceed with the simulation.

Our key assumptions are as follows:

- We model the inter-arrival time (IAT) for appointment requests as exponentially distributed. We use an IAT of 7.5 minutes for the baseline case.
- The average repair time is specified as 40.5 minutes in [8]. We model the inherent customer-to-customer variation in service time as a triangular distribution, with a minimum of 22.5 minutes, maximum of 76.5 minutes, and mode of 40.5 minutes. We assume that the service time can be estimated accurately when appointments are booked, so this variation is not uncertainty.
- Repair technicians, including space and equipment, are the critical resource for scheduling. Other resources are set at a high enough level.
- Each customer gets a free loaner while his or her car is being serviced. We assume that a sufficient number of loaners is available or that additional loaners can be obtained as needed.
- Appointments are always made such that the repair can be completed the same day. When a customer calls, the earliest available appointment is for the next business day. Appointments are based strictly on the availability of a repair technician for the estimated repair time.
- The repair facility operates 7 days a week, 10 hours a day.
- To model the extreme case of customers arriving late for their appointments, we use a triangular distribution with a minimum of 30 minutes, maximum of 90 minutes, and mode of 60 minutes to represent their tardiness.
- In cases where the parts may be delivered late, we model the *total* parts delivery time as a triangular distribution with a minimum of 0, maximum of 1200 minutes (two business days), and mode of 600 minutes.

We have tested the simulation model extensively under various conditions. As expected, all queue lengths and waiting times within the repair facility are zero when the workflow is not disturbed. As part of the validation step, we have confirmed that the total (average) time spent on each repair job by the customer and the provider match the reference behavior specified in Figure 1, under a constant service time. We have also performed stress tests and sensitivity analysis on the model, which are described as part of the results in the next section.

We have run each simulation for 30 simulated days after a warm-up period of 60 simulated days. Each data point in the graphs below was obtained by averaging the performance figures from 20 replications of a particular simulation.

RESULTS

We present our main simulation results in this section, followed by an interpretation and discussion in the next section. In all cases, we have measured performance and sensitivity as functions of the resource level (number of repair technicians) and demand from customers (rate of appointment requests).

Figure 2 illustrates how the instantaneous utilization of repair technicians varies with the number

of technicians for two different inter-arrival times. The utilization starts out at 1 when there is more than enough work for the technicians. As the resource level is increased, fewer technicians remain busy on average. When the inter-arrival time is doubled from 7.5 minutes to 15 minutes, the system exhibits significantly more slack, as expected.

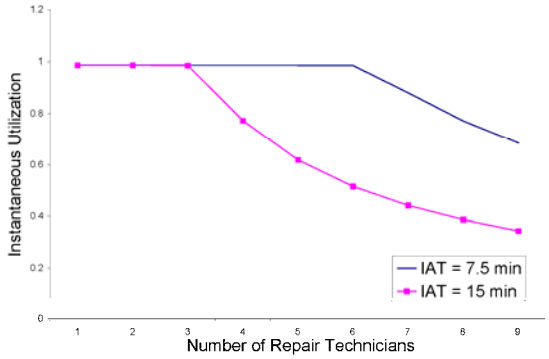

Figure 2. Instantaneous utilization of repair technicians

Figure 3 shows that the average time to appointment declines as more repair technicians are added. It finally reaches a constant value at the point where there are enough technicians and the only remaining constraint is our assumption that the earliest appointment would be for the next day. The time to appointment is consistently shorter for the longer inter-arrival time.

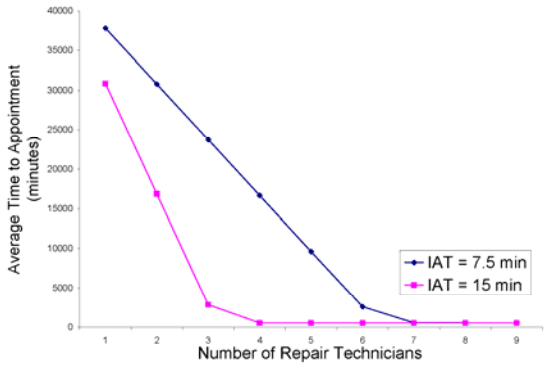

Figure 3. Average time to appointment

Figure 4 shows the rise in the average number of loaners out as the number of repair technicians is increased. The number of loaners also indicates the work in process, which saturates as the resource level is increased beyond a useful point. A shorter inter-arrival time results in a larger amount of work in process.

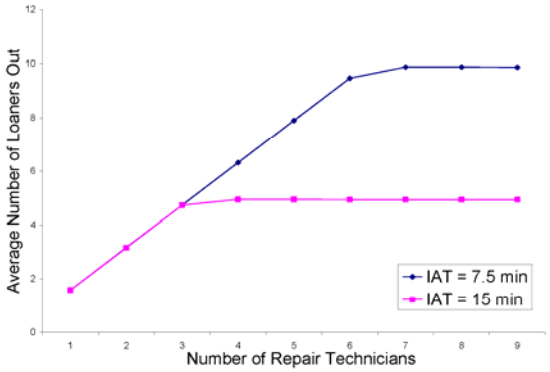

Figure 4. Average number of loaners out

The average repair completion time for the baseline case is constant regardless of resource level, as indicated in Figure 5. When the smooth workflow is disturbed through late arrival of customers, the repair completion time is longer, but converges with the baseline case when the resource level is increased sufficiently.

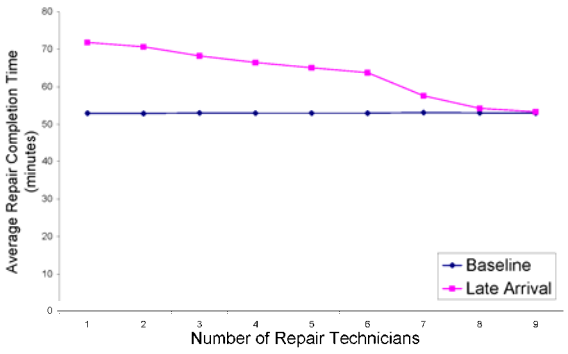

Figure 5. Average repair completion time (Baseline vs. Late arrival; IAT = 7.5 min.)

Figure 6 displays additional results for the late arrival scenarios. The average number of loaners out (work in process) is higher as the workflow becomes uneven. As in the previous figure, it converges with the baseline case when there are enough repair technicians. Note that the scheduling of future appointments is not modified when the workflow becomes uneven.

We then run a similar experiment by enabling possible delays in parts delivery, but without late arrival of customers. Figures 7 and 8 show that the

response of the system is very different compared to late arrival of customers.

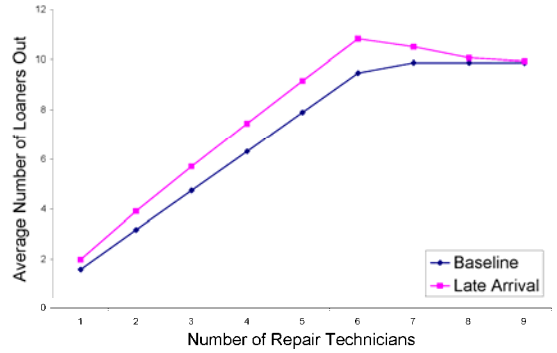

Figure 6. Average number of loaners out (Baseline vs. Late arrival; IAT = 7.5 min.)

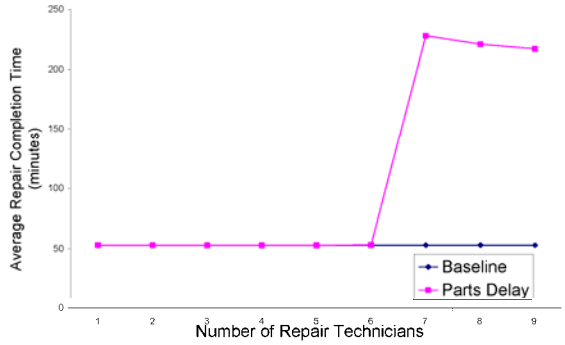

Figure 7. Average repair completion time (Baseline vs. Parts delay; IAT = 7.5 min.)

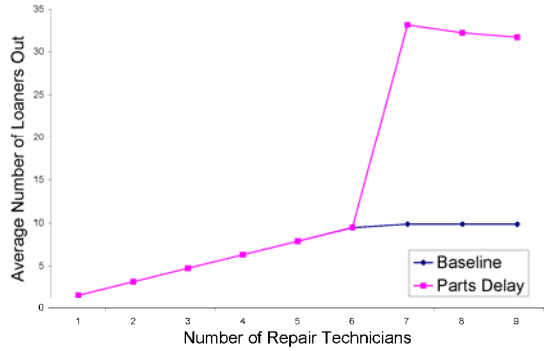

Figure 8. Average number of loaners out (Baseline vs. Parts delay; IAT = 7.5 min.)

Delay in parts delivery is exposed only when the resource level is high enough, in which case the effect on system performance is much more dramatic than when customers arrive late. The actual distribution of the parts delivery delay does not change the basic trend here. We found that both triangular and uniform distributions yield similar results.

DISCUSSION

A lean service process, such as the car repair example discussed in this paper, works smoothly through careful scheduling of appointments, highly accurate work estimates, and preordering parts. A side-effect of this approach is that customers may have to wait outside the system until the start of the appointment. The modeling and simulation here show that the time to appointment is a performance metric that can be optimized in conjunction with other metrics, such as utilization of repair technicians and work in process, to maximize customer satisfaction.

Figures 2 and 3 suggest that an optimal number of repair technicians can be determined for the anticipated rate of appointment requests, such that an acceptable utilization of technicians can be maintained while keeping customers reasonably satisfied. For an inter-arrival time of 7.5 minutes, about seven repair technicians can reduce the time to appointment to its minimum level while utilizing over 85% of the technicians' time.

At seven repair technicians, the average number of loaners or work in process begins to saturate (Figure 4), indicating that there will be no further gain to the provider from adding more technicians. Thus, the optimal number of repair technicians is about seven.

From Figures 5 and 6, it is clear that late arriving customers throw off the careful scheduling. This results in non-zero queue lengths and waiting times inside the repair facility and increases the average repair completion time. At eight or more technicians, there is enough slack in the system to absorb the perturbations without degrading performance – at this point, the late arrival curve converges with the baseline case.

This addresses a typical concern with lean service systems. What if customers do not arrive "just in time"? This particular lean system appears to be fairly resilient with respect to late arrival of customers. Based on the performance characteristics in Figures 5 and 6, the problem can be minimized or eliminated with the right level of resources.

Note that the horizontal curve for the baseline in Figure 5 implies that the average repair completion time is constant regardless of the number of

technicians, because the scheduling eliminates queuing. This is an important property of a lean service process.

When delays are introduced in the delivery of parts, the effect is visible only when the parts arrive later than customers. As the number of repair technicians increases, the time to appointment decreases, and this increases the chances of customers arriving sooner than parts. This occurs at about seven repair technicians in our simulations (Figures 7 and 8). The additional repair time includes both queuing delay and parts delay. At eight or more technicians, the queuing delay is reduced due to the additional service resources. Thus, both the average repair completion time and the average number of loaners decrease after reaching a peak. The system is clearly very sensitive to delays in parts delivery.

CONCLUSION

We have demonstrated in this paper the significant role that simulation can play in the analysis and optimization of a typical lean service process. The system can be optimized and tuned given an expected rate of appointment requests and possible perturbations. It is difficult to see how this could be accomplished without resorting to modeling and simulation. Simulation can also enhance understanding and insight by revealing important characteristics and properties of the system.

There are still more perturbations and constraints that could be applied to our car repair example. They include failure and downtime of resources, limited numbers of loaner cars and service consultants (thus treating both as additional resources to be analyzed and optimized), and inaccurate initial diagnosis. When these issues are addressed in the future, we expect that they will further highlight the need for simulation in the development of lean service processes.

REFERENCES

[1] Womack, J.P. and D.T. Jones. 1996. *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*. Simon & Schuster, New York, NY.

[2] Schroer, B.J. 2004. “Simulation as a Tool in Understanding the Concepts of Lean Manufacturing.” *Simulation*, Vol. 80, Issue 3, March: 171-175.

[3] Adams, M., P. Componation, H. Czarnecki, and B.J. Schroer. 1999. “Simulation as a Tool for Continuous Process Improvement.” In *Proceedings of the 1999 Winter Simulation Conference.* IEEE, Piscataway, NJ, 766-773.

[4] Fowler, J.W. and O. Rose. 2004. “Grand Challenges in Modeling and Simulation of Complex Manufacturing Systems.” *Simulation*, Vol. 80, Issue 9, September: 469-476.

[5] Dennis, S., B. King, M. Hind, and S. Robinson. 2000. “Applications of Business Process Simulation and Lean Techniques in British Telecommunications Plc.” In *Proceedings of the 2000 Winter Simulation Conference.* IEEE, Piscataway, NJ, 2015-2021.

[6] Swank, C.K. 2003. “The Lean Service Machine.” *Harvard Business Review*, October: 123-129.

[7] Nallicheri, N., T.C. Bailey, and J.S. Cade. 2004. “The Lean, Green Service Machine.” *Strategy+Business Resilience Report*, November 18.

[8] Womack, J.P. and D.T. Jones. 2005. “Lean Consumption.” *Harvard Business Review*, March: 59-69.

[9] Kelton, W.D., R.P. Sadowski, and D.T. Sturrock. 2004. *Simulation with Arena*. McGraw-Hill, New York, NY.