SINGLE SIMULATION BUFFER OPTIMIZATION

Christoph Roser, Masaru Nakano, and Minoru Tanaka
Toyota Central Research and Development Laboratories
Nagakute, Aichi 480-1192 JAPAN
croser@robotics.tytlabs.co.jp
nakano@robotics.tytlabs.co.jp
tanaka@robotics.tytlabs.co.jp

Abstract

The performance of manufacturing systems can be adjusted by allocation buffer into the manufacturing system. Buffer will improve the performance of manufacturing systems by improving the utilization of the constraints; yet buffer will also increase the makespan and the work in progress. Due to the complex nature of the systems, buffer allocation is usually difficult to optimize. This paper presents a prediction model of the effect of buffer based on the shifting bottleneck detection and a blocking and starving analysis. The prediction model is used to optimize the buffer allocation using only a single simulation.
Keywords: Buffer Allocation, Optimization, Manufacturing Systems, Theory of Constraints

1 INTRODUCTION

The performance of manufacturing systems can be adjusted by allocation buffer into the manufacturing system. Buffer will improve the performance of manufacturing systems by improving the utilization of the constraints by reducing the aggrandizement of random effects. However, buffer will also increase the makespan and the work in progress. Due to the complex nature of the systems, buffer allocation is usually difficult to optimize.

There is a large body of research related to buffer allocation. Most of the methods are based on building a metamodel requiring numerous repetitions, for example by using simulated annealing and genetic algorithms (Spinellis 1999; Spinellis 2000a; Spinellis 2000b), neural networks (Altiparmak 2002), gradient based searches (Gershwin 2000; Levantesi 2001; Schor 1995), or tabu searches (Shi 2002). However, in industry it is usually difficult to obtain the large number of replications needed to implement the model, and the use of these methods is inefficient. Other approaches are based on a functional approximation and evaluation (Enginarlar 2001; Enginarlar 2002) and knowledge based methods (Vouros 1998), or combinations of analytical and simulation based methods (Nakano 2000).

This paper analyzes the behavior of the short term bottlenecks in a manufacturing system. A shifting bottleneck detection method is used to determine level of constraint of the machines onto the system. The idle times of all machines is also analyzed and the cause of the idle time is determined. Subsequently, a general prediction model is established to estimate the effect of buffer onto the system performance. This prediction model can optimize an example system using only a single simulation. The optimization results are compared to the results of a commercial optimization software.

2 EXAMPLE SYSTEM

The presented method will be demonstrated using a complex simulation example, consisting of 7 machines M1 to M7 in a branched system as shown in Figure 1. The first machine M1 receives material from an unlimited supply and is therefore never starved. The last machine M7 delivers to an unlimited demand and is therefore never blocked. All machines have exponentially distributed cycle times and therefore a large variation. 11 different buffer locations are considered, and buffer of an initial capacity of one have been added to all locations. These buffer are named BMx and AMx for buffer before and after machine Mx. There are two different part types A and B in a ratio of 2:1. All parts pass through machines M1 and M2, and M6 and M7. Parts A pass through machines M3 and M4, and parts B pass through machine M5.

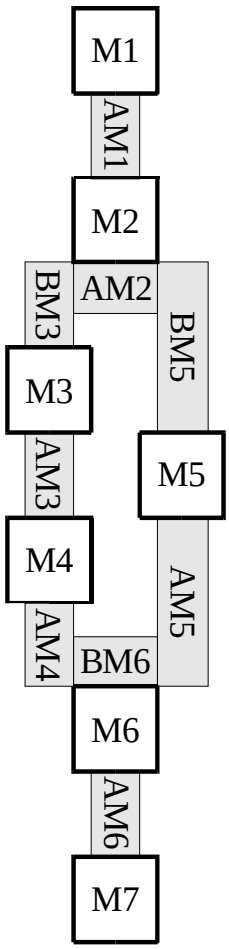

Figure 1: Example system layout

The system described above has been simulated for 20,000 seconds, of which a warming up period of 4000s has been removed. The system produced one part every 3.36 minutes, or 17.87 parts per hour. There were on average 9.97 parts in the system for a makespan of 33.48 minutes. Table 1 shows the mean processing times and utilizations of the system.

Machine	Mean Processing Time (min)	Utilization
M1	1.5	44.73%
M2	2.6	76.64%
M3	3.2	71.92%
M4	2.0	46.00%
M5	3.0	22.52%
M6	1.4	42.89%
M7	1.6	47.22%

Table 1: Mean processing times and utilizations

3 BUFFER EFFECT PREDICTION MODEL

The buffer prediction model uses the bottleneck probability and the blocking and starving analysis to estimate the effect of the buffer onto the system. The following sections give a brief overview of the methods, and a more detailed description can be found in other publications (Roser 2003).

3.1 Shifting Bottleneck Detection

The shifting bottleneck detection method will be able to detect and monitor the shifting momentary bottleneck of a production system, and also determine the average bottleneck over a selected period of time (Roser 2002a; Roser 2002b). The underlying idea is that the longer a machine is working without interruption, the more likely it is that this machine constrains the performance of other machines. Therefore, at any given time the machine with the longest uninterrupted active period is the momentary bottleneck at this time. Figure 2 visualizes the method using a simple example consisting of only two machines. At the beginning, M1 has the longer active period and is the bottleneck machine. Later, however, M2 has the longer active period and becomes the bottleneck machine. During the overlap between the current bottleneck period and the subsequent bottleneck period the bottleneck shifts from M1 to M2.

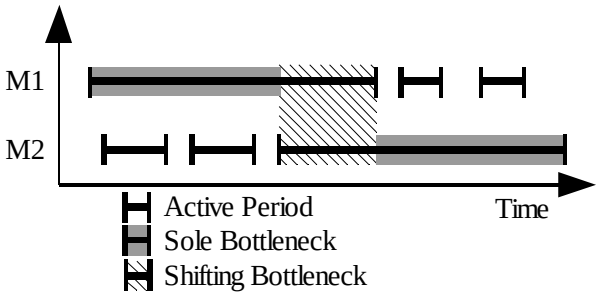

Figure 2: Shifting bottlenecks example

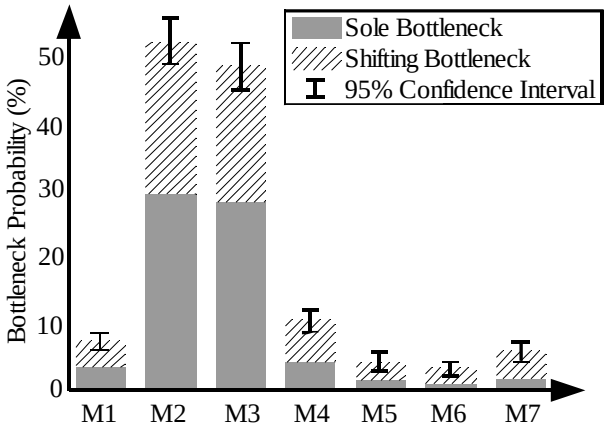

Figure 3: Example bottleneck probability

The shifting bottleneck method detects and monitors the momentary bottleneck at any instant of time. Subsequently, the bottleneck probability can be estimated as the percentage of time a machine is a bottleneck. The

example has been analyzed, and it was found that machines M2 and M3 were the main bottlenecks as shown in Figure 3.

3.2 Blocking and Starving Analysis

There are also four possible modes how a buffer can affect another machine. A buffer can provide either additional parts or spaces. Usually, parts are given to starved machines downstream, and spaces are provided to blocked machines upstream. However, a buffer may also relieve a blocked machine indirectly by providing parts to another machine, or relieve a starved machine indirectly by providing spaces to another machine. Therefore, to understand the buffer it is crucial to understand the causes of the blocking and starving and the path to the causes. The blocking and starving analysis determines the cause of every idle period of every machine.

Figure 4 shows the results of the blocking and starving analysis of the example for machines M5 and M6, where the width of the line represents the percentage of the starved and blocked times that pass through this path. M5 is mostly blocked by M6, and sometimes by M7. M5 is starved by machine M2. However, occasionally M3 or M4 block M2, which in turn starves M5. Therefore, M5 is starved indirectly by M3. Machine M6 is blocked by machine M7.

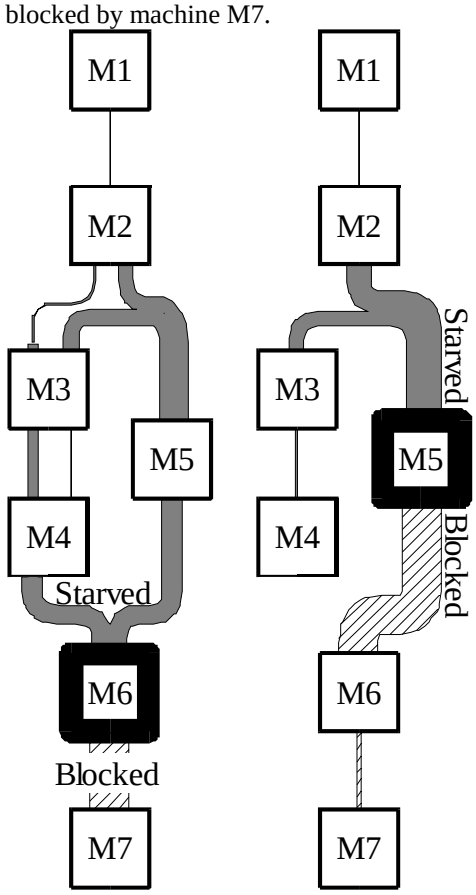

Figure 4: Blocking and starving analysis for machines M5 and M6

However, the starving of machine M6 is more complex. Naturally, machine M6 is starved by machine M4, which in turn is starved by M3 and M2. Machine M6 is also starved by M5, which in turn is starved by M2. However, M5 is also starved indirectly by M3, which in turn is blocked by M4. Therefore, some of the starving times of M6 are caused by M3 or M4 blocking M2, therefore starving M5 and subsequently starving M6! This means, that a buffer BM3 before machine M3 affects machine M6 in two ways: Parts in the buffer BM3 reduce the starving of M6 through M4. But empty spaces in BM3 also reduce the starving of M6 through M2 and M5! Understanding such a behavior is crucial for the estimation of the buffer BM3, or any general buffer prediction.

3.3 Buffer Effect Prediction Model

The prediction of the effect of a buffer onto the system is based on a number of steps. (1) The buffer configuration for which the performance of the manufacturing system is to be evaluated is selected. Buffer capacity may be added to already existing buffers or new buffers may be created at buffer locations without a current buffer. This selection may be random, may be based on a optimization algorithm, or may be picked by a human operator.

(2) The mean number of parts in a buffer is measured for existing buffers, and the additional available parts and empty spaces due to the buffer increase in step (1) is determined. If a buffer location of interest does not yet have any buffer capacity, the mean number of parts or empty spaces have to be estimated using other methods, as for example the ratio of upstream bottlenecks to downstream bottlenecks. The mean number of parts in a buffer represents the number of parts available to relieve starving, whereas the difference of the mean number of parts to the total capacity represents the mean number of empty spaces available to relieve blocking. For example, a buffer that is always full cannot provide empty spaces against blocking, and a buffer that is always empty cannot provide additional parts against starving.

(3) This step combines the information gained from the starving and blocking analysis with the mean number of parts and empty spaces determined in step (2). The mean number of additional parts or additional empty spaces in a buffer is multiplied with the percentage of the time additional parts or spaces would affect a machine. This represents the additional number of parts or spaces available at a machine due to the increase in a buffer capacity. I.e. if the mean number of parts in a buffer is increased by two, and parts in this buffer affect another machine 50% of the time, then this buffer represents one additional part available for this machine. This has to be done for all possible combinations of machines and buffer increases, distinguishing if the buffer increase would reduce blocking or if the buffer increase would reduce starving of the machine.

(4) This step sums up the results of step (3), to determine the total number of additional parts available in front of a machine to reduce starving, and the additional number of empty spaces available after a machine to reduce blocking. This has to be done for every machine.

(5) This step estimates the additional time a machine could work due to the additional parts and spaces available in front of and after the machine. The additional number of parts in front of a machine from step (4) is multiplied with the mean processing time of the machine to determine how much longer the machine could work due to the additional parts in front of the machine. Similarly, the additional number of empty spaces after the machine is multiplied with the mean processing time of the machine to determine how much longer the machine could work due to the additional spaces after the machine. This has to be done for every machine.

(6) This step estimates the reduction in the blocking and starving time of a machine based on the additional time a machine could work and the distribution of the blocking and starving times. For this, the probability distribution of the starving and blocking times of the machines are needed. It is usually difficult to match these distribution to a commonly used probability distribution, but this is not necessary. However, a probability distribution based on the measured starving and blocking times is sufficient. It has to be determined what percentage of the starving times could be avoided if the machine could work more due to additional parts. This is simply the mean value of the starving distribution for the range between zero and the maximum additional time due to additional parts, divided by the mean of the entire starving distribution. Similarly, the percentage of the blocking times that could be avoided is simply the mean value of the blocking distribution for the range between zero and the maximum additional time due to additional spaces, divided by the mean of the entire blocking distribution. This represents the reduction of the times the machine is blocked or starved, and subsequently the reduction in the overall time the machine is idle. Again, these evaluations have to be done for all machines.

(6) The previous step determined the possible improvement in the idle time for each machine due to the additional buffers. However, this improvement of a machine does not necessarily turn into an improvement of the entire system. The system improves only if the bottleneck machines improve. Therefore, the bottleneck probability of the shifting bottleneck detection method defines which part of the machine improvement will yield a system improvement. For example if the total time between parts of a machine would be reduced by 10% due to a reduced idle time, and the bottleneck probability of this machine would be 50%, then the overall improvement of the time between parts for the system due to the improvement of this machine would be 5%. If the bottleneck probability of the machine is zero, then the system will not improve regardless of the possible machine improvement. This is very similar to a throughput sensitivity analysis (Roser 2002c). (7) The new work in progress is estimated based on the buffer increase and the mean number of parts in the buffer. The makespan is predicted based on the work in progress estimate and the predicted throughput.

4 OPTIMIZATION

The buffer prediction model can also easily be used to optimize a manufacturing system. Two different optimization approaches are described, one using a single step optimization, based on only a single simulation. The other approach uses a multi-step optimization, where the prediction model is used for a local optimization, after which a new simulation verifies the system and is itself optimized again. The system has also been optimized with a commercial software product for verification purposes.

The prediction model for the example system has been created as described above. The automatic analysis software *TopQ Analyzer* provides the results in an easy-to-understand MS Excel file including graphical output. This Excel file also contains the buffer prediction model, predicting the throughput, production rate, makespan, work in progress, and the total buffer capacity.

To optimize the system an utility function is needed to create a trade off between the throughput, makespan, and the work in progress. Other variables can also be included, as for example the total buffer capacity. Furthermore, the buffer capacities of the individual buffers have to be constrained to be positive integers. Other constraints as for example a minimum production rate or a maximum WIP can also be added.

The profit utility function was the sum of the frequency times 3,000 [$*hour] minus the work in progress times 50 [$], the makespan times 100 [$/min], and the total number of Buffer Spaces times 5 [$]. The cost was adjusted such that the cost of the initial system was zero, i.e. a total of $49,686.61 was subtracted.

The buffer allocation of the manufacturing system can now easily be optimized to maximize the utility function subject to the constraints. As the prediction model allows the rapid comparison of many different buffer allocations, a wide variety of optimization methods can be used, as for example a gradient based method or a genetic algorithm. A good description of a wide range of optimization methods can be found in (Nemhauser 1994).

The optimizations itself was performed using the Solver add-in included in MS Excel, maximizing the profit subject to the buffer capacities being nonnegative integers. All 9 buffer locations have been included in the optimization, with no upper limit on the buffer capacity.

4.1 Single Simulation Optimization

In a single step optimization, the buffer allocation of the manufacturing system is optimized to determine the buffer allocation with the maximum utility. Optimizing a system similar to the example in Figure 1, the prediction recommended the adding of 66 buffer spaces to the existing 13 buffer spaces to maximize the profit utility.

As the prediction model allows the fast comparison of multiple settings, the Solver was able to compare approximately 14,000 different buffer configurations in about 200 seconds in order to find the optimal solution. The predicted optimum showed a profit increase of $8,832 per hour, which was very close to the actually measured profit increase of the system of $11,474 per hour despite the large changes in the system, as for example a 20% increase in the production rate, a 50% increase in the WIP, or a 80% increase in the buffer capacity. Table 2 shows the comparison of the initial measured performance, the predicted optimal performance, and the actually measured performance of the optimized system. It seems that most predictions were quite accurate. Only the work in progress (WIP) and the makespan had larger errors, where the predicted WIP and makespan increase was double the measured increase. Nevertheless the overall measured profit differs only by 20% from the predicted profit, and the overall prediction algorithm seems to work quite well.

Performance	Initial	Predicted	Measured
Time Per part (s)	3.36	2.76	2.70
Frequency (1/h)	17.87	21.75	22.22
WIP	9.97	19.54	15.97
Makespan (s)	33.48	53.89	43.12
Buffer Spaces	13	75	75
Profit Increase ($)	0.00	8,832.61	11,474.63

Table 2: Comparison of initial, predicted, and measured system performance

4.2 Multi-Step Optimization

Alternatively, it is also possible to use a multi step optimization, where the prediction model is used for a local area optimization, after which a new simulation verifies the results. The results of the new simulation are then used for a subsequent optimization step. This is repeated until no further improvement is possible. Figure 5 shows a multi step optimization for a system similar to Figure 1, where the step size is limited to 15 buffer spaces per buffer. The simulation quickly reached an optimal plateau after 4 steps, and no further improvement was possible after step 13. The optimal plateau showed a profit utility increase of $12,000 per hour, surpassing the results of the commercial optimization software. Table 3 shows a comparison of the performance of the initial system with the best step (4) and the last step (13) of the multi step optimization.

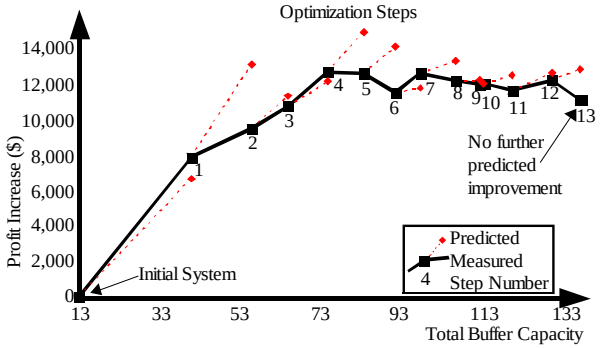

Figure 5: Multi Step Buffer Optimization

Step Number	0 Initial	4 Best	13 Last
Time Per part (s)	3.36	2.63	2.67
Frequency (1/h)	17.87	22.84	22.44
WIP	9.97	18.40	18.36
Makespan (s)	33.48	48.34	49.09
Buffer Spaces	13	75	138
Profit Increase ($)	0.00	12,715.38	11,105.44

Table 3: System performance for different optimization steps

4.3 Commercial Simulation Software

For comparison the model has also been optimized using an evolutionary optimizer which was supplied with the Extend simulation software used to simulate the system. The example optimization has the additional constraint of limiting the buffer capacity to a maximum of 100 and also does not include AM5 due to limitations in the number of variables.

The optimizer compared 108 different settings, simulating a total of 993 configurations over the period of 32 hours. The commercial analyzer increased the same significant buffer as the presented prediction model. However, the commercial analyzer also increased other buffer, which did not seem to have a significant effect on the system, and subsequently the returned optimal solution was inferior to the solution found by the presented prediction model. The resulting buffer capacities of buffer with small effects also appeared to be somewhat random due to the very small differences of the performance for different buffer sizes. The returned model had a profit of only $10,157 compared to the $11,474 profit of the model selected by the presented prediction model. Table 4 compares the initial buffer settings with the returned optimum of the single step optimization, the best step (4) and the last step (13) of the multi step optimization, and the results of the commercial evolutionary optimizer. Table 5 shows a comparison of the performance of the initial system with the optimal system according to the evolutionary optimizer.

Buffer	AM1	AM2	BM3	AM3	AM4	BM5	AM6
Initial	1	1	1	1	1	1	1
Single Step	3	1	17	34	1	1	16
Multi Step (4)	3	3	38	16	1	1	7
Multi Step (13)	5	3	99	16	1	1	7
Com-mercial	5	18	15	28	23	7	8

Table 4: Buffer allocation comparison showing Initial system, Single Step optimum, Multi step best and last result, and commercial evolutionary simulation optimum (AM5 and BM6 excluded because unchanged)

Performance	Initial	Optimal
Time Per part (s)	3.36	2.69
Frequency (1/h)	17.87	22.31
WIP	9.97	19.27
Makespan (s)	33.48	51.84
Buffer Spaces	13	124.00
Profit Increase ($)	0.00	10,467.57

Table 5: Results of the commercial optimization software

5 IMPLEMENTATION

The method has been implemented in a software analysis tool for the TOPQ simulation engine. A screenshot of the software is shown in Figure 6. Besides a thorough statistical analysis and a bottleneck detection, this software also produces a starving and blocking analysis and a complete prediction model as a MS Excel worksheet. The constraints and the utility function can easily be added into the Excel model, which then can be optimized using the Solver plug-in included in MS Excel. The software is currently used by selected companies of the TOYOTA group.

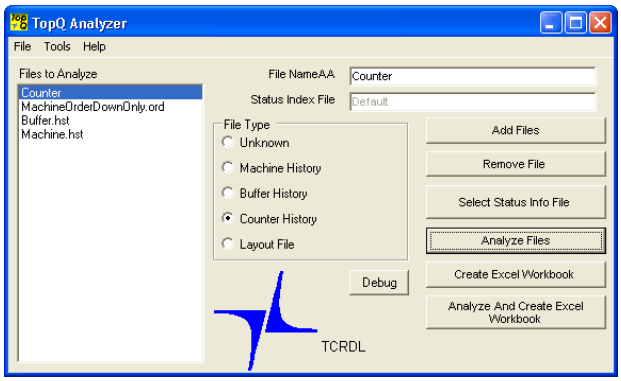

Figure 6: TopQ Analyzer Screenshot

6 CONCLUSIONS

This paper describes a prediction model to estimate the effect of increased buffer capacity onto the system performance based on only a single simulation. There are two main steps to this method. The first step determines the bottleneck probabilities of the machines in the system based on the active periods. The second step analyzes the causes of the idle (starving or blocking) periods for all machines, and determines which buffer locations would reduce the idle time, creating a prediction model.

This prediction model is used to optimize an example system using only a single simulation. Despite a total increase by 66 buffer spaces, the predicted and the measured performance are very similar. The optimization results of the prediction model is compared to the results of a commercial optimization software, finding a better performance in the presented prediction model results despite requiring only a fraction of the computation power. The optimization using the prediction model took only about 10 minutes, including initial simulation, prediction model creation, and optimization, but compared 14,000 buffer configurations. The evolutionary simulation optimizer, however, took about 32 hours and compared only 993 configurations.

This prediction model therefore allows for an easy and quick optimization of manufacturing systems, cost-effectively allocating buffer to reduce the detrimental temporary effect of other machines onto the primary bottleneck.

References

Altiparmak, Fulya, Dengiz, Berna, and Bulgak, Akif A. (2002). "Optimization of Buffer Sizes in Assembly Systems Using Intelligent Techniques." *Winter Simulation Conference*, San Diego, CA., USA, 1157-1162.

Enginarlar, Emre, Li, Jingshan, and Meerkov, Semyon M. (2001). "A Potpourri on the Theme of Lean Buffering." *Third Aegean International Conference on "Design and Analysis of Manufacturing Systems"*, Tinos Island, Greece.

Enginarlar, Emre, Li, Jingshan, Meerkov, Semyon M., and Zhang, Rachel Q. (2002). "Buffer Capacity for Accommodating Machine Downtime in Serial Production Lines." *International Journal of Production Research*, 40(3), 601-624.

Gershwin, Stanley B., and Schor, James E. (2000). "Efficient Algorithms for Buffer Space Allocation." *Annals of Operations Research*, 93, 117-144.

Levantesi, R., Matta, A., and Tolio, T. (2001). "A new algorithm for Buffer Allocation in Production Lines." *Third Aegean International Conference on "Design and Analysis of Manufacturing Systems"*, Tinos Island, Greece.

Nakano, Masaru, and Ohno, Katsuhisa. (2000). "An Integrated Analytical/Simulation Approach for Economic Design of an AGV System." *Journal of*

the Operations Research Society of Japan, 43(3), 382-395.

Nemhauser, G. L., Rinnooy Kan, A. H. G., and Todd, M. J. (1994). *Optimization*, Elsevier Science, Amsterdam.

Roser, Christoph, Nakano, Masaru, and Tanaka, Minoru. (2002a). "Detecting Shifting Bottlenecks." *International Symposium on Scheduling*, Hamamatsu, Japan, 59-62.

Roser, Christoph, Nakano, Masaru, and Tanaka, Minoru. (2002b). "Shifting Bottleneck Detection." *Winter Simulation Conference*, San Diego, CA, USA, 1079-1086.

Roser, Christoph, Nakano, Masaru, and Tanaka, Minoru. (2002c). "Throughput Sensitivity Analysis using a single simulation." *Winter Simulation Conference*, San Diego, CA, USA, 1087-1094.

Roser, Christoph, Nakano, Masaru, and Tanaka, Minoru. (2003). "Buffer Allocation Model based on a Single Simulation." *Winter Simulation Conference*, New Orleans, Louisiana, USA, 1238-1246.

Schor, James E. (1995). "Efficient Algorithms for Buffer Allocation," Masters of Science, Massachusetts Institute of Technology, Boston.

Shi, Leyuan, and Men, Shuli. (2002). "Optimal Buffer Allocation in Production Lines." *Submitted to IIE Transactions*.

Spinellis, Diomidis D., and Papadopoulos, Chrissoleon T. (1999). "Explore: A Modular Architecture for Production Line Optimisation." *Proceedings of the 5th International Conference of the Decision Science Institute*, Athens, Greece, 1446-1449.

Spinellis, Diomidis D., and Papadopoulos, Chrissoleon T. (2000a). "A Simulated Annealing Approach for Buffer Allocation in Reliable Production Lines." *Annals of Operations Research*, 93, 373.

Spinellis, Diomidis D., and Papadopoulos, Chrissoleon T. (2000b). "Stochastic Algorithms for Buffer allocation in Reliable Production lines." *Mathematical Problems in Engineering*, 5, 441-458.

Vouros, G. A., and Papadopoulos, H. T. (1998). "Buffer Allocation in unreliable production lines using a knowledge based system." *Computers and Operations Research*, 25(12), 1055-1067.