Roser, C., Nakano, M., Tanaka, M., 2004. Holistic Analysis of Manufacturing Systems: Conclusions from understanding the interactions, in: European Simulation Multiconference. Magdeburg, Germany.

HOLISTIC ANALYSIS OF MANUFACTURING SYSTEMS: CONCLUSIONS FROM UNDERSTANDING THE INTERACTIONS

Christoph Roser
Masaru Nakano
Minoru Tanaka
Digital Engineering Laboratory
Toyota Central Research and Development Laboratories
Nagakute, Aichi, 480-1192, JAPAN
E-mail: croser@robotics.tytlabs.co.jp

KEYWORDS

Manufacturing System Simulation, Bottleneck Detection, Buffer Optimization, Holistic Analysis,

ABSTRACT

Manufacturing systems are crucial for the Industry. A frequent objective is to improve the throughput of the system, which can be done by carefully adding or removing buffers in the system or by improving the machine performances. The Toyota Central Research Laboratories have recently developed a number of simulation based methods to understand and predict the behavior of a manufacturing system. These methods include a reliable quantitative bottleneck detection, resulting in a machine sensitivity analysis including a prediction model of the effect of machine changes, and a blocking and starving analysis, resulting in a prediction model of the effects of buffers and a subsequent buffer optimization. The novel idea of these methods is a holistic view of the manufacturing system, i.e. understanding the system by analyzing the relations between the machines instead of analyzing machines independently. This allows a much better understanding of the manufacturing system, and enables the optimization of the manufacturing system using only a single simulation. These methods are being evaluated at selected companies of the Toyota group, and have also generated great interest in academia and industry. This paper provides a framework of the holistic manufacturing analysis and a summary of the developed methods.

INTRODUCTION

The understanding and optimization of manufacturing systems is a frequently researched subject in discrete event simulation (Boesel et al. 2001; Fu et al. 2000) (Azadivar 1999; Swisher et al. 2000). Yet, most methods analyze the simulation results in an atomistic view, and study each machine separately. However, if the correlated manufacturing system is broken into its independent machines before analysis, it is difficult to later combine the analysis of the machines into a understanding of the entire system. The presented methods here all use a holistic approach by studying the interactions between the machines in order to understand the system. Subsequently, the methods using this holistic approach yield a much better understanding of the manufacturing system.

Applying the holistic approach to the active times of the machines evaluates the shifting bottleneck in the system. There are numerous definitions as to what constitutes a bottleneck (Lawrence and Buss 1995). Within this paper, we define a bottleneck as a stage in a production system that has the largest effect on slowing down or stopping the entire system. Finding the bottleneck is no trivial task, and (Cox and Spencer 1997) for example simply recommends that '… the best approach is often to go to the production floor and ask knowledgeable employees …'. Furthermore, although most manufacturing systems usually have one main bottleneck, in all but the simplest applications bottlenecks are not static but rather shift between different machines (Lawrence and Buss 1994; Moss and Yu 1999). There are existing methods to determine the bottleneck, as for example based on the utilization (Law and Kelton 2000), the queue length, mathematical approaches (Chiang et al. 1998; Chiang et al. 2002; Kuo et al. 1996), or disjunctive graphs (Adams et al. 1988). However, these methods all have serious shortcomings in usability or accuracy or both. The presented shifting bottleneck detection method is based on a detailed analysis of the relationships of the working times of the different machines and provides a very accurate and quantitative measure of the constraints in the machines.

Applying the holistic approach to the idle times of the machines enhances the understanding of the blocking and starving relations in the system and allows a better understanding of the effect of buffers. An excellent discussion of the effect of buffers can be found by Conway et al (Conway et al. 1988) and others (Brittan 1996; Caramanis et al. 2001). 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 and

Papadopoulos 1999a; Spinellis and Papadopoulos 2000a; Spinellis and Papadopoulos 2000b), neural networks (Altiparmak et al. 2002), gradient based searches (Gershwin and Schor 2000; Levantesi et al. 2001; Schor 1995), or tabu searches (Shi and Men 2002). Other approaches are based on a functional approximation and evaluation (Enginarlar et al. 2001; Enginarlar et al. 2002) and knowledge based methods (Vouros and Papadopoulos 1998), or combinations of analytical and simulation based methods (Nakano and Ohno 2000). The presented buffer allocation method is based on a detailed analysis of the blocking and starving relationships in manufacturing systems and provides a usable prediction model of the effect of buffers in a manufacturing system, subsequently allowing the optimization of the system using only a single simulation.

A NEW FRONTIER: HOLISTIC MANUFACTURING SYSTEM ANALYSIS

The newly developed methods described below all use a holistic approach to manufacturing system analysis compared to the atomistic view of the current simulation analysis methods. The holistic methods analyze not only independent machines, but the interaction between machines. The Merriam-Webster online dictionary defines holistic and atomistic as follows:

> ***holistic****: relating to or concerned with wholes or with complete systems rather than with the analysis of, treatment of, or dissection into parts*
>
> ***atomistic****: characterized by or resulting from division into unconnected or antagonistic fragments*

The Status Quo: Atomistic Analysis

What is called generally “a simulation” consists of many different steps, as for example building a model, verifying the model, simulating, collecting data, and finally analyzing the collected data. A manufacturing system is a network of interconnected entities, as for example machines, buffers, workers, or AGV. These entities affect each other and interact with each other. To understand the system it is crucial to analyze these interactions.

However, the status quo of current manufacturing simulation analysis is usually an atomistic analysis, completely ignoring the dependencies and interactions and achieving only a fraction of the possible conclusions which are hidden in the simulation data. For example, in a standard analysis the working times of a machine is summed up to calculate the utilization, with utter disregard of the relation of the individual working times to the other machines. Similar, the idle times are summed up to extract the percentage of the time a machine is idle, again with complete disregard of the reasons a machine is idle, which is usually caused by another entity in the system.

This approach is visualized in Figure 1, where the manufacturing system is first simulated, and then broken into its individual elements for an atomistic analysis. Naturally, based on the statistical data of the independent elements it is very difficult to determine valid conclusions for the correlated system. While these results are needed to understand a manufacturing system, they represent only a fraction of the wealth of knowledge contained in the simulation data. Furthermore, as the analysis is only based on an atomistic view, it is very difficult or almost impossible to estimate how a change of one of the entities would affect the other entities and ultimately the entire system, i.e. it is very difficult to make holistic conclusions from the atomistic analysis.

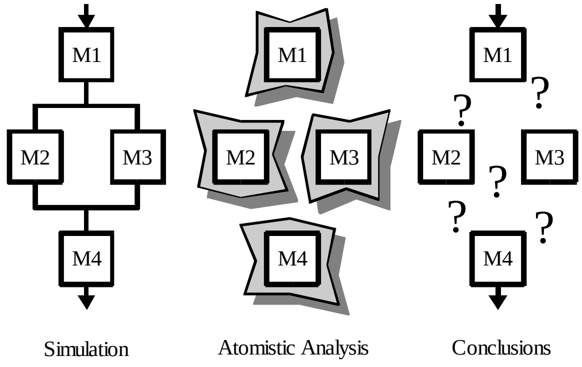

Figure 1: Atomistic Analysis Process

Holistic Analysis

The analysis presented here goes beyond the standard and is fundamentally different from the standard output analysis. The holistic analysis analyzes the interactions between the machines and statistically represents the correlations between the machines in order to understand the system. This results in a much better understanding of the system, and also allows for a prediction of the system performance based on the changes in the different entities.

This approach is visualized in Figure 2, where after the simulation of the manufacturing system, the complex interactions of the system are analyzed to obtain a holistic understanding of the system. Subsequently, valuable information about the system is obtained and it is easy to make valid conclusions for the correlated system.

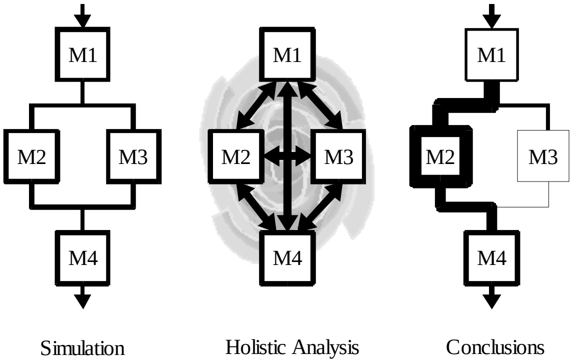

Figure 2: Holistic Analysis Process

For example the shifting bottleneck detection method as described in the next section compares the uninterrupted active (working) times of the different machines, and at any given time the machine with the longest uninterrupted active period is the bottleneck for this time. Thus the system is analyzed from a holistic point of view by comparing the different machines at different times. Subsequently, it is possible to measure the level of constrains of the machines and to make a valid prediction of the effect of a change in the machines onto the entire system.

Another example is the blocking and starving analysis as described later in this paper. The blocking and starving analysis investigates every single idle period, i.e. blocked and starved period, and tries to determine the cause of this idle period, and the path between the idle machine and the cause of the idle machine. Therefore this used a holistic view by analyzing the blocking and starving interactions between the machines. Subsequently, it was possible to make a valid prediction of the effect of a change in a buffer onto the entire system.

It can be seen that this holistic analysis offers great promise to the simulation analyst by providing valuable information about the system, allowing valid conclusions about the behavior of the system if the system changes. The current research describes only two holistic approaches based on the working times and the blocking and starving analysis, however, it is certainly possible to develop more holistic methods to answer questions about the system behavior. Of course, the holistic methods also include the results of the standard analysis, which is necessary to support the estimation of the system behavior.

While holistic methods use a slightly more complex analysis than atomistic methods, the presented methods are implemented in a analysis software and do not require an additional effort by the analyst. Overall, an holistic approach promises to obtain much more information about a system than an atomistic approach ever could, and further research in holistic analysis methods will be very useful for the industry.

SHIFTING BOTTLENECK DETECTION METHOD

The presented 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 based on the duration the machines are active without interruption. Subsequently, a sensitivity analysis of the machines with respect to the throughput can be performed and a basic prediction model established. More details for the different sub-methods can be found in previous publications (Roser et al. 2002a; Roser et al. 2002b; Roser et al. 2002c; Roser et al. 2002d; Roser et al. 2003b).

Holistic Analysis Of The Active Duration

The presented method is based on a holistic analysis of the duration a processing machine is active without interruption. All possible machine states are grouped into two groups, being either active states or inactive states. Similar can also be done for workers, AGV, or any other entity in the system that may cause a delay. A state is active whenever the machine may cause other machines to wait. For example working on one part may cause a subsequent idle machine to wait for the completion of the part, or a machine under repair may block previous machines. A state is inactive if the associated machine is not active but instead waiting for the completion of another task, for example the arrival of a part or service, or for the removal of a part. Similar definitions can be made for any entity in a manufacturing system, as for example workers or AGV, or any entity in a discrete event system in general. Figure 3 shows an example of the active (work, repair, tool change) and inactive (waiting) states of one machine during a brief period of a simulation. The active periods without interruption are shown.

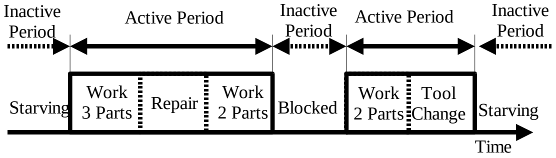

Figure 3: Active Periods of Machine During Simulation

The bottleneck detection method compares the durations of the active periods of the different machines. If the analysis is based on simulation data or historical data, it is possible to determine the durations of all active periods for all machines. However, if the analysis is used for real time monitoring, the future is unknown and the durations of the active periods are known only until the present. In this case, the active duration is measured until the present and may be updated if further information becomes available with time.

The Momentary Shifting Bottleneck

The underlying idea of the method is that at any given time the machine with the longest uninterrupted active period is the momentary bottleneck at this time. The overlap of the active period of a bottleneck with the previous or subsequent bottleneck represents the shifting of the bottleneck from one machine to another machine. In an interconnected production system, machines block and starve each other. If a machine is active, it is neither starved nor blocked. The longer a machine is active without interruption, the more likely it is that this machine blocks or starves other machines in the production system. The machine with the longest uninterrupted active period therefore has the biggest impact onto starving or blocking the other machines, therefore being the largest constraint a.k.a. the largest bottleneck. The following method describes how to determine which machine of a production system is the sole or part of a shifting bottlenecks at any time *t*.

If at time *t* no machines are active, then there is no bottleneck. If one or more machines are active at the time *t*, the machine with the longest active period at the time *t* is the momentary bottleneck machine, and the active period of this machine is the current bottleneck period. It is also necessary to find the previous and subsequent bottleneck machines before and after the current bottleneck period. The previous bottleneck machine is the machine with the longest active period just prior to the beginning of the current bottleneck period. Similarly, the subsequent bottleneck machine is the machine with the longest active period just after the end of the current bottleneck period.

The shifting of the bottleneck from the previous bottleneck machine to the current bottleneck machine happens during the overlap of the previous and the current bottleneck periods. Similarly, the shifting of the bottleneck from the current bottleneck machine to the subsequent bottleneck machine happens during the overlap of the current and the subsequent bottleneck periods. During the overlaps between the bottleneck periods no machine is the sole bottleneck, instead the bottleneck shifts between the two machines. If a bottleneck machine is not shifting, then this machine is the sole and only bottleneck at this time. Of course, if there are no other machines active just prior or after the current bottleneck period, then there is no overlap and subsequently no shifting bottleneck. Using this method, it can be determined at any given time if a machine is a non-bottleneck, a shifting bottleneck, or a sole bottleneck. This method allows the detection of the bottleneck, where and when the previous bottleneck was shifting to the current bottleneck, and where and when the current bottleneck is shifting to the next bottleneck.

Figure 4 visualizes the method using a simple example consisting of only two machines. The figure shows the active periods of the machines over a short period of time. At the selected time t, both machines M1 and M2 are active. Yet, as M1 has the longer active period, M1 is the bottleneck machine for the time t. As there is no machine active before the current bottleneck period, there is no overlap and no shifting at the beginning of the current bottleneck period. However, at the end of the current bottleneck period, M2 is active and has the longest active period. Therefore, the subsequent bottleneck machine is M2. During the overlap between the current bottleneck period and the subsequent bottleneck period the bottleneck shifts from M1 to M2. Now, M2 is the bottleneck machine. Similarly, at the end of the bottleneck period of M2, the bottleneck shifts back to M1. Processing all available data using this method shows at what time which machine is the bottleneck machine, when the bottleneck is shifting, and when there is no bottleneck at all. Therefore, it is possible to detect and monitor the bottleneck at all times.

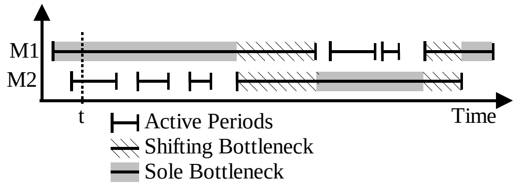

Figure 4: Shifting Bottlenecks

The Bottleneck Probability

The above method detects and monitors the momentary bottleneck at any instant of time. However, in many cases it may be of interest not to investigate an instant of time but rather a period of time. This section describes how to compare different machines with respect to the bottleneck over a period of time. To determine the bottleneck during a period of time the available data is analyzed and the momentary bottlenecks are determined over the selected period of time. Next, the percentage of time a machine is the sole bottleneck machine and the percentage of the time a machine is part of a shifting bottleneck is measured for the selected period of time.

Figure 5 visualizes this method using the example with two machines as shown in Figure 4. The percentages of the machines being the sole bottleneck or the shifting bottleneck have been measured over the period of time shown in Figure 4. The larger the percentages, the larger is the effect of the respective machine onto slowing down or stopping the system. M1 is the sole bottleneck more often than M2, and is also involved in a number of shifting operations. M2 is the smaller constraint, i.e., a secondary bottleneck, having being the sole bottleneck for a smaller percentage of time. The graph below shows the overall effect of the machines in terms of being the bottleneck over a period of time by plotting the sum of the machines being the bottleneck or shifting. Overall, an improvement of the throughput of M1 would yield a larger overall improvement of the system throughput than an improvement of M2, as M1 is the primary bottleneck during the selected period of time.

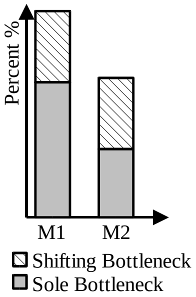

Figure 5: Average Bottleneck over Period of Time

Machine Sensitivity Analysis

The shifting bottleneck detection method as described above determined the sole and shifting bottlenecks at any given time during the simulation. The sensitivity analysis enhances this approach by analyzing the events of which the bottleneck periods consist of. Figure 6 shows the detailed per event analysis of the example used in Figure 4. The example includes three types of events, namely machine M1 working, machine M2 working, and machine M2 under repair. Each of the active periods shown in Figure 4 consists of one or more of these events. The sole and shifting bottleneck periods are underlined grey and hatched respectively, while non-bottleneck periods are grayed out, as they do not affect the throughput.

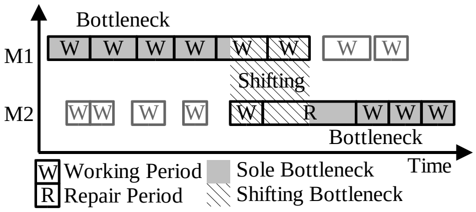

Figure 6: Bottleneck Events

The bottleneck periods limit the overall manufacturing system throughput, and the bottleneck periods consist of the different actions of the machines. Therefore, the actions of the bottleneck machines during the bottleneck periods determine the overall system throughput. Knowing the sole and shifting bottleneck periods and the events therein, the percentage contribution of the variables of the machines to the throughput can be calculated easily.

Figure 7 shows the percentages of the time each of the three events was a sole or shifting bottleneck for the example shown in Figure 6. Machine M1 working contributed with 45% sole and 20% shifting bottlenecks the largest part of all sole and shifting bottleneck periods, and therefore has the largest effect onto the throughput. Machine M2 working and machine M2 repair contributed smaller percentages, and therefore the throughput is less sensitive to these two variables. These values represent the relative effect of a change in the variables towards the overall throughput. For example if machine M1 Working would be improved by a small amount, between 45 and 65% of this improvement would benefit the overall system throughput. Therefore, these sensitivity values allow the prediction of the system performance of a changed system as described in the next section.

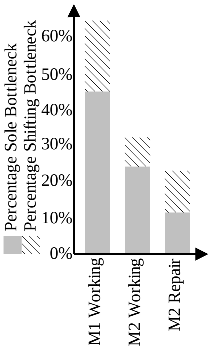

Figure 7: Sensitivity Analysis

Machine Prediction Model

The sensitivity analysis using the shifting bottleneck method determines the percentage effect of each machine state (working, repair, tool change,) onto the overall throughput. This allows the prediction of the effect of a change in a machine variable (working time, repair time, tool change time …) onto the overall throughput. Note that the method distinguishes between the effect due to sole bottlenecks and due to shifting bottlenecks. A sole bottleneck is the only bottleneck at this time in the system, and an improvement of the sole bottleneck events will improve the throughput. However, if there is a shifting bottleneck, then it is not sure which machine actually is the true bottleneck, and an improvement of the shifting bottleneck events may or may not improve the overall throughput. Therefore, the lower and upper limits of the expected percentage change of the system performance can be calculated based on the percentage of the time a machine state was a sole bottleneck or a sole or shifting bottleneck.

However, one shortcoming of sensitivity analysis and gradient-based methods in general is that they are only strictly true at the system for which the sensitivity has been measured. As the system variables change, the system changes, and subsequently the sensitivity changes. The larger the system changes the larger the uncertainty of the prediction. This is illustrated in Figure 8 for the above example. As the machine M1 working contributes between 65 and 85% of the bottlenecks, a reduction of the working time of M1 to zero would theoretically reduce the mean time between parts by 65 to 85%. However, it is to be expected that as the working time of M1 decreases, M1 becomes less likely to be a bottleneck and other machines will become a bottleneck, and the true performance improvement will be less than the expected performance improvement for larger changes.

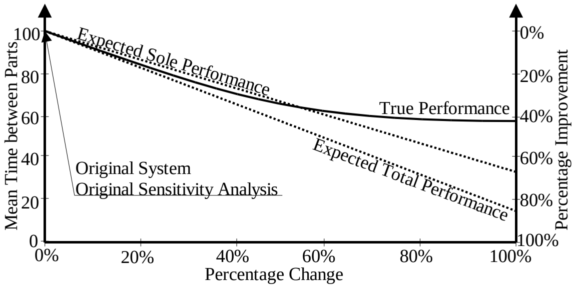

Figure 8: Performance Prediction

This sensitivity analysis and performance prediction can then be used to form the base of a manufacturing system optimization to allow the rapid evaluation of manufacturing system alternatives for a local optimization. For an overview of optimization techniques please see (Nemhauser et al. 1994) for general optimization techniques, and (Andradottir 1998; Fu 2001; Swisher et al. 2000) for simulation optimization methods.

BUFFER ALLOCATION MODEL

Buffers improve the system throughput by reducing the idle time (blocking and starving) of the machines. Therefore, to understand the buffers it is crucial to understand the blocking and starving of the machines, the causes thereof, and, most important the path to the causes and the buffer locations in between. This method analyzes every starving or blocking occurrence of every machine in the simulation, and finds the cause of the starving and blocking, and, more important, the buffer locations on the path between the idle machine and the cause thereof. The time a possible buffer location is part of a path is determined for each machine.

There are also four possible modes how a buffer can affect another machine as shown in Table 1. A buffer can provide either additional parts or spaces. Usually, parts are given to starved machines downstream (Mode I), and spaces are provided to blocked machines upstream (Mode IV). However, a buffer may also relieve a blocked machine indirectly by providing parts to another machine (Mode II), or relieve a starved machine indirectly by providing spaces to another machine (Mode III).

Table 1: Effect Modes of Buffers on Machines

Effect Modes	Machine: Starved	Machine: Blocked
Buffer: Provide Parts	I	II
Buffer: Provide Spaces	III	IV

Holistic Starving And Blocking Analysis

To find the cause of an idleness of a machine, an algorithm has been developed that follows the cause from machine to machine or buffer until the cause of the idle period has been found. While depending on the detail of the available data, there may some ambiguity, the following set of rules provide a good estimate for the search of the cause of an idle period. In this algorithm, it is also assumed that the loading time of parts to and from a machine is negligible and that a buffer is always between two machines and two machines only (the branched system in the example uses a transfer machine to realize the branches).

A machine is always either active (A), blocked (B) or starved (S). The definition of active includes not only working machines, but also machines under repair or performing a tool change. The cause of a block can always be found by following a machine downstream. If the downstream machine is also blocked or the buffer is full, continue following the block downstream. If the downstream machine is active, or the downstream buffer is not full, the cause of the block has been found. If the downstream machine is starved, it is necessary to turn around and find the cause of the starving period. An overview of the 5 possible situations is given in Figure 9.

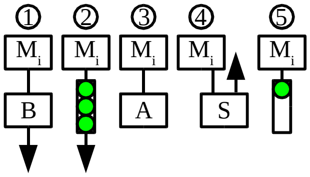

Figure 9: Situations with Blocking of Machine M_i

For starved machines, there are also a total of 5 possible situations, determining the next machine/buffer in the search for the cause of the starved machine. The cause of a starving situation is always sought upstream. If the upstream machine is also starved, or the upstream buffer is empty, continue searching for the cause upstream. If the upstream machine is active or the upstream buffer is not empty, the cause of the starving situation has been found. If the upstream machine is blocked, it is necessary to turn around and find the cause of the block for the upstream machine. An overview of the situations is given in Figure 10.

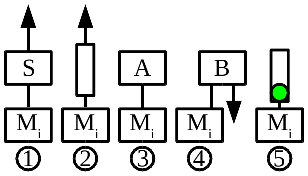

Figure 10: Situations with Starving of Machine M_i

Using this set of rules, it is possible to find the cause of an idle period for all idle periods of all machines, and to determine the period of time a buffer location was part of the path to the cause of an idle machine. This allows the conclusion of the effect of a buffer onto the different machines.

An example system with 7 machines has been analyzed, and the causes of the blocking and starving of the machines has been established. Figure 11 presents the results for machines M3 and M5 in graphical form, showing the path of the starves (cross-hatched) and blocks (diagonal-hatched) from machine M3 and machine M5 to the machine causing the starve or block. The width of the path represents the fraction of the starves/blocks following this path.

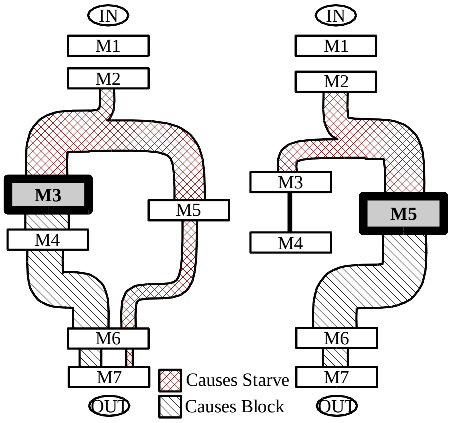

Figure 11: Causes of Blocking and Starving of Machines M3 and M5

In the simulated example, machine M3 is blocked 6.1% of the time. Whenever machine M3 is blocked, the path to the cause of the block leads to the next downstream machine M4 (100% of the blocked time). However, M4 itself is rarely the cause of the block. Most the paths continue to machine M6 (78% of the blocked time), and M7 (46% of the blocked time). Therefore, a buffer increase before machine M7 affects the blocking of machine M3 46% of the time. Machine M3 is also starved for 5.8% of the time. The path to the cause of the starve splits, with 38% of the starving periods caused by machine M2, and 62% following to machine M5. From machine M5 the paths continue to M6 (27% of the starving time), and from there to M7 (15% of the starving time).

The causes of the starving and blocking of machine M5 can be traced similarly, with the path to the cause of the blocks continue to machine M6 (99% of the blocked time) and M7 (57% of the blocked time). The path to the cause of the starves splits towards M2 (55% of the starved time), and M3 (40% of the starved time), continuing to M4 (7% of the starved time). The path to the causes of the blocked and starved periods has to be analyzed for all machines to estimate the effect of buffers.

The path between the idle machines and the cause thereof allows an estimation of the effect of buffers. Only buffers in these path affect the machines. Furthermore, there are different modes in which a buffer can affect a machine as discussed in Table 1. For example in Figure 11, if the buffer before machine M3 can provide parts, the starving of M3 is reduced (Mode I). At the same time, if the buffer can provide additional spaces, the starving of machine M5 is also reduced (Mode III). The buffer before machine M6 has an especially interesting effect, as it not only reduces the blocking of machine M3 by providing spaces (Mode IV), but also reduces starving on the very same machine M3 by providing spaces to machine M5 (Mode III).

Buffer Prediction Model

The same buffer can have different effects depending on the number of parts and the number of spaces provided to the machines in the system. Therefore, the first step is to estimate the mean number of parts in a buffer, and subsequently the mean number of additional parts and the mean number of additional free spaces if a buffer is increased. There are a number of methods available in the literature, most of them based on a decomposition approach (Bouhchouch et al. 1993; Dallery and Frein 1989; Spinellis and Papadopoulos 1999b). This paper uses a estimation of the mean number of parts based on the shifting bottleneck detection approach, but the reader may choose any suitable method of his/her choice, as long as the additional number of parts and free spaces can be estimated based on the change in a buffer size.

The next step estimates the number of additional parts available in front of a machine to reduce starving and the additional number of spaces available after a machine to reduce blocking. This estimation is based on the additional number of parts and free spaces available in all buffers, and the effect of the buffer into the machines for the four modes

After estimating the number of additional parts and free spaces available for each machine, the possible reduction in the time per part of the machines can be estimated. Each additional part available in front of the machine allows the machine to work longer by avoiding starving periods. Similarly, each available free space after the machine allows the machine to work longer by avoiding blocking periods. The maximum additional time that can be worked depends on the additional number of parts, spaces, and the mean cycle time needed to produce one part. For example, if there would be one additional part available in front of machine M3, then machine M3 with a cycle time of 140s could avoid starving periods up to 140s completely and reduce all remaining starving periods by 140s.

The mean time that can be reduced therefore depends on the distribution of the starving and blocking times of the machines, and the probability density function of the starving time distribution and the probability density function of the blocking time distribution are needed to estimate the reduction in the idle times of the machines

The mean reduced idle time can be calculated by integrating the probability density functions multiplied by the time span between the time 0 and the upper limit defined by the cycle time and the additional number of parts or spaces. The mean waiting time of the entire distribution can be calculated by setting the upper limit of the integral to infinite. The ratio of these two integrals is the percentage reduction of the waiting time. Combining this percentage reduction with the percent of the time a machine is starved or blocked gives the overall percentage reduction of the mean starving time per part and the mean blocking time per part. The total percentage reduction in the time between parts for a machine is the sum of the percentage reduction of the starving times and blocking times.

The above equations estimate the possible reduction in the time between parts for all machines based on the additional number of parts before and free spaces after the machine and the blocking and starving time distributions. However, this estimation does not yet take the complex interactions in the system into account, and the predicted machine improvement may not be realized because other machines continue to block and starve this machine. The transition from a possible machine improvement to the actual system improvement depends on the bottlenecks and is described below.

The previous step estimated the improvement in the machine performances based on the change in the buffers. However, this improvement may not be realized because other machines continue to block or starve this machine. To estimate the system improvement based on the individual machine improvements, the contribution of the individual machines to the system performance has to be determined, i.e. which machines constrain the system and by how much. This estimation is based on the bottleneck probability as described above. While the bottleneck probability distinguishes between sole (unique) bottlenecks and shifting bottlenecks (bottlenecks in the process of changing from one machine to another), this method uses the sole bottleneck probability only.

The bottleneck probability of a machine describes what effect a percentage improvement of the time between parts of a machine would have on the percentage improvement of the time between parts of the system. The improvement of the system is simply the sum of the individual machine improvements weighted by the bottleneck probability. To get from the initial time between parts of the system to the improved time between parts of the system simply reduce the initial time by the percentage reduction. This predicted time per part for increased buffers can then easily be used to predict other system performance measures like the make span or the work in progress.

The prediction model for the effects of buffers has been validated extensively for a number of different systems and buffers (Roser et al. 2003a). For example, Figure 12 shows the comparison of the predicted time per part to the measured time per part for a buffer BM3 located before machine M3 in the same manufacturing system as shown in Figure 11. The continuous line shows the measured data including the 95% confidence intervals, and the dotted lines shows the predicted system performance. The predicted performance follows the measured data very nicely. The overall root mean squared error RMSE was only 0.24s.

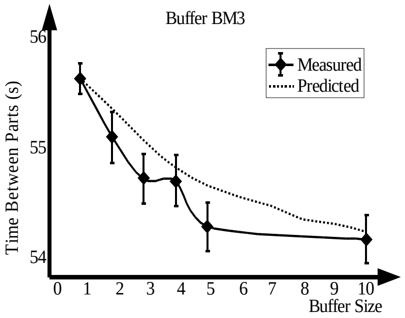

Figure 12: Performance Prediction for Buffer BM3

Buffer 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.

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 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 et al. 1994).

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 11, the prediction recommended the adding of 66 buffer spaces to the existing 13 buffer spaces to maximize the profit utility. The predicted optimum showed a profit increase of $9,000 per hour, which was very close to the actually measured profit increase of $12,000 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. The result was also close to the results of a commercial optimization software based on repetitive simulations, which returned an optimum profit of $10,000 per hour, but required 700 times longer to evaluate the result.

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 13 shows a multi step optimization for a system similar to Figure 11, 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.

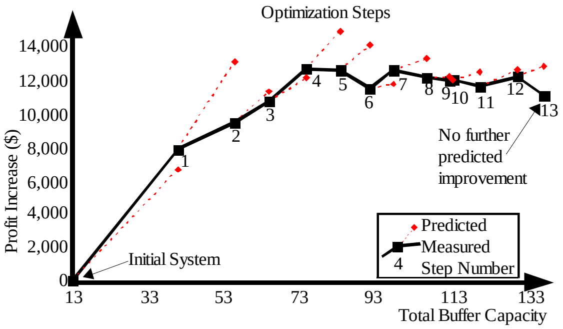

Figure 13: Multi Step Buffer Optimization

CONCLUSIONS

This paper summarizes a medley of methods recently developed by the Toyota Central Research and Development Laboratories related to the analysis of manufacturing system simulations. Unique to all the methods is the underlying holistic approach of analyzing not only the entities of the manufacturing system but also the interactions between these entities. All of the methods have been validated and implemented in an automated analysis software used by some companies in the Toyota group. This software creates an easy to use MS Excel file containing all the data and prediction models for optimizations.

The bottleneck detection method determines the momentary shifting bottleneck based on a holistic comparison of the machine working times. This provides a quantitative measure of the constraints, allowing the subsequent use of the bottleneck probability for a sensitivity analysis and a prediction model. Overall, this method greatly enhances the understanding of the system by quantitatively determining the primary and secondary bottlenecks and the non-bottlenecks.

The buffer prediction model is based on a detailed holistic analysis of the blocking and starving situations in the manufacturing system, determining the cause of every idle period in the system. This allows for a subsequent prediction model of the effect of buffers onto the system and a optimization of the buffer allocations. This prediction model greatly reduces the time needed to understand and improve the effect of buffers in the system.

Further research includes the development of additional methods in the promising field of holistic simulation analysis. By statistically analyzing the interactions in a manufacturing system or any discrete event system in general, a much deeper understanding of the system can be obtained. If the relations between the system entities is understood, the effect of changes in one entity can be estimated, greatly benefiting the industry by improving their competitiveness.

REFERENCES

Adams, J., Balas, E., and Zawack, D. 1988. "The Shifting Bottleneck Procedure for Job-Shop Scheduling". *Management Science*, 34(3), 391-401.

Altiparmak, F., Dengiz, B., and Bulgak, A. A. 2002. "Optimization of Buffer Sizes in Assembly Systems Using Intelligent Techniques". *Winter Simulation Conference*, ed. E. Yucesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 1157-1162, San Diego, CA., USA.

Andradottir, S. 1998. "A Review of Simulation Optimization Techniques". *Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 151-158, Washington, DC, USA.

Azadivar, F. 1999. "Simulation Optimization Methodologies". *Winter Simulation Conference*, ed. P. A. Farrington, D. T. Nembhard, D. T. Sturrock, and G. W. Evans, 93-100, Phoenix, AZ, USA.

Boesel, J., Jr., R. O. B., Glover, F., Kelly, J. P., and Westwig, E. 2001. "Future of Simulation Optimization". *Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 1466-1469, Arlington, Virginia, USA: Institute of Electrical and Electronics Engineers.

Bouhchouch, A., Frein, Y., and Dallery, Y. 1993. "A Decomposition Method for the Analysis of Tandem Queueing Networks with Blocking Before Service". *Queueing Networks with Finite Capacity*, 97-112.

Brittan, D. 1996. "When Bad Things Happen to Good Factories". *Technology Review*.

Caramanis, M., Pan, H., and Anli, O. 2001. "Is there a Trade off between Lean and Agile Manufacturing? A Supply Chain Investigation". *Third Aegean International Conference on "Design and Analysis of Manufacturing Systems"*, ed., Tinos Island, Greece.

Chiang, S.-Y., Kuo, C.-T., and Meerkov, S. M. 1998. "Bottlenecks in Markovian Production Lines: A Systems Approach". *IEEE Transactions on Robotics and Automation*, 14(2), 352-359.

Chiang, S.-Y., Kuo, C.-T., and Meerkov, S. M. 2002. "c-Bottlenecks in Serial Production Lines: Identification and Application". *Mathematical Problems in Engineering*.

Conway, R. W., Maxwell, W., McClain, J. O., and Thomas, L. J. 1988. "The role of work-in-process inventory in serial production lines". *Operations Research*, 36(2), 229-241.

Cox, J. F. I., and Spencer, M. S. 1997. "*The Constraints Management Handbook"*. Boca Raton, Florida: CRC Press - St. Lucie Press.

Dallery, Y., and Frein, Y. 1989. "A Decomposition Method for the Approximation Analysis of Closed Queueing Networks with Blocking". *Queueing Networks with Blocking*, 193-215.

Enginarlar, E., Li, J., and Meerkov, S. M. 2001. "A Potpourri on the Theme of Lean Buffering". *Third Aegean International Conference on "Design and Analysis of Manufacturing Systems"*, ed., Tinos Island, Greece.

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

Fu, M., Andradottir, S., Carson, J. S., Glover, F., Harrell, C. R., Ho, Y.-C., Kelly, J. P., and Robinson, S. M. 2000. "Integrating Optimization and Simulation: Research and Practice". *Winter Simulation Conference*, ed. P. A. Fishwick, K. Kang, J. A. Joines, and R. R. Barton, 610-615, Orlando, Florida, USA.

Fu, M. C. 2001. "Simulation Optimization". *Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, Arlington, Virginia, USA: Institute of Electrical and Electronics Engineers.

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

Kuo, C.-T., Lim, J.-T., and Meerkov, S. M. 1996. "Bottlenecks in Serial Production Lines: A System-Theoretic Approach". *Mathematical Problems in Engineering*, 2, 233-276.

Law, A. M., and Kelton, D. W. 2000. "*Simulation Modeling & Analysis"*. McGraw Hill.

Lawrence, S. R., and Buss, A. H. 1994. "Shifting Production Bottlenecks: Causes, Cures, and Conundrums". *Journal of Production and Operations Management*, 3(1), 21-37.

Lawrence, S. R., and Buss, A. H. 1995. "Economic Analysis of Production Bottlenecks". *Mathematical Problems in Engineering*, 1(4), 341-369.

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"*, ed., Tinos Island, Greece.

Moss, H. K., and Yu, W. B. 1999. "Toward the Estimation of Bottleneck Shiftiness in a Manufacturing Operation". *Production and Inventory Management Journal*, 40(2), 53-58.

Nakano, M., and Ohno, K. 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"*. Amsterdam: Elsevier Science.

Roser, C., Nakano, M., and Tanaka, M. 2002a. "Detecting Shifting Bottlenecks". *International Symposium on Scheduling*, ed., 59-62, Hamamatsu, Japan.

Roser, C., Nakano, M., and Tanaka, M. 2002b. "Shifting Bottleneck Detection". *Winter Simulation Conference*, ed. E. Yucesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 1079-1086, San Diego, CA, USA.

Roser, C., Nakano, M., and Tanaka, M. 2002c. "Throughput Sensitivity Analysis using a single simulation". *Winter Simulation Conference*, ed. E. Yucesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 1087-1094, San Diego, CA, USA.

Roser, C., Nakano, M., and Tanaka, M. 2002d. "Tracking Shifting Bottlenecks". *Japan-USA Symposium on Flexible Automation*, ed., 745-750, Hiroshima, Japan.

Roser, C., Nakano, M., and Tanaka, M. 2003a. "Buffer Allocation Model based on a Single Simulation". *Winter Simulation Conference*, ed. S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, 1238-1246, New Orleans, Louisiana, USA.

Roser, C., Nakano, M., and Tanaka, M. 2003b. "Constraint Management in Manufacturing Systems". *International Journal of the Japan Society of Mechanical Engineering, Series C, Special Issue on Advanced Scheduling*, 46(1), 73-80.

Schor, J. E. 1995. Efficient Algorithms for Buffer Allocation, Masters of Science, Massachusetts Institute of Technology, Boston.

Shi, L., and Men, S. 2002. "Optimal Buffer Allocation in Production Lines". *Submitted to IIE Transactions*.

Spinellis, D. D., and Papadopoulos, C. T. 1999a. "Explore: A Modular Architecture for Production Line Optimisation". *Proceedings of the 5th International Conference of the Decision Science Institute*, ed. D. K. Despotis and C. Zopounidis, 1446-1449, Athens, Greece.

Spinellis, D. D., and Papadopoulos, C. T. 1999b. "Production Line Buffer Allocation: Genetic Algorithm versus Simulated Annealing". *Second International Aegean Conference on the Analysis and Modelling of Manufacturing Systems*, ed., 89-101, Tinos, Greece: University of the Aegean, Department of Business Administration.

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

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

Roser, C., Nakano, M., Tanaka, M., 2004. Holistic Analysis of Manufacturing Systems: Conclusions from understanding the interactions, in: European Simulation Multiconference. Magdeburg, Germany.

Swisher, J. R., Hyden, P. D., Jacobson, S. H., and Schruben, L. W. 2000. "A Survey of Simulation Optimization Techniques and Procedures". *Winter Simulation Conference*, ed. P. A. Fishwick, K. Kang, J. A. Joines, and R. R. Barton, 119-128, Orlando, Florida, USA.

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.

AUTHOR BIOGRAPHIES

CHRISTOPH ROSER is a researcher at the Toyota Central Research and Development Laboratories, Nagoya, Japan. Current research topics are simulation output analysis, bottleneck detection and risk and sensitivity analysis. He graduated from the Fachhochschule Ulm, Germany with a diploma in automation engineering, and received his Ph.D. in Mechanical Engineering at the University of Massachusetts, Amherst, studying flexible and robust design methods. His email address is:
croser@robotics.tytlabs.co.jp

MASARU NAKANO is a principal researcher at Toyota Central Research and Development Laboratories, Inc. (TCRDL) in Japan. He joined TCRDL in 1980 and since then has studied in the field of robotics, computer vision, and manufacturing system design. He is currently research leader and manager of the Digital Engineering Laboratory. He received his B.S. and M.S. in operations research from Kyoto University, and his Dr. Eng. in manufacturing systems from the Nagoya Institute of technology. His email address is
nakano@robotics.tytlabs.co.jp

MINORU TANAKA is a researcher at the Toyota Central Research and Development Laboratories, Nagoya, Japan. He graduated from Matsue Technical College, joined TCRDL in 1985, and specialized in the field of robotics, manufacturing system modeling, and discrete event simulation. He is a member of the Japan Society for Precision Engineering. His email address is
tanaka@robotics.tytlabs.co.jp