Abstract

Manufacturing throughput time reduction can be a daunting task due to the many factors that influence it and their complex interactions. However, there are basic principles that, if applied correctly, can be used to reduce manufacturing throughput time. This paper presents a conceptual framework that illustrates these principles. The framework illustrates the factors that influence manufacturing throughput time, the actions that can be taken to alter each factor, and their interactions. The framework is detailed enough to provide guidance to the industry practitioner on how to reduce manufacturing throughput time, while being general enough to apply to most manufacturing situations.

Keywords: Throughput Time, Throughput Time Reduction, Lead Time Reduction, Quick Response Manufacturing

1. Introduction

Manufacturing throughput time is defined as the length of time between the release of an order to the factory floor and its receipt into finished goods inventory or its shipment to the customer. Reductions in manufacturing throughput time can generate numerous benefits, including lower work-in-process and finished goods inventory levels, improved quality, lower costs, and less forecasting error (because forecasts are for shorter time horizons). More importantly, reductions in manufacturing throughput time increase flexibility and reduce the time required to respond to customer orders. This can be vital to the survival and profitability of numerous firms, especially those experiencing increased market pressures for shorter delivery lead times of customized product.

Many firms are struggling in their attempts to reduce manufacturing throughput time, and the factor changes that can reduce manufacturing throughput time are not always understood (Suri et al. 1996). While manufacturing throughput time reduction can indeed be a daunting task due to the many factors that influence it and their complex interactions, there are basic principles that, when applied correctly, can be used to reduce manufacturing throughput time.

To apply the principles correctly, the basic factors that determine manufacturing throughput time must be clearly understood. This paper first uses a simple hypothetical manufacturing system to illustrate the basic factors that determine manufacturing throughput time and explain why each factor occurs. This tutorial could be used to train workers in these basic concepts. The paper then presents a conceptual framework that illustrates the factors that influence manufacturing throughput time, the actions that can be taken to alter each factor, and their interactions. Because customers are concerned about the response time to their order and because the minimum order size can be for a single part/product, the focus throughout this paper will be on the manufacturing throughput time per part (MTTP).

Information obtained from case studies of lead time reduction efforts at four different plants (see Johnson and Wemmerlöv 1998 and Johnson 1999 for published versions of two of the case studies), previous research on throughput time reduction factors (see Table 1), and queuing theory principles were used to construct the framework. The framework is detailed enough to provide guidance to the industry practitioner on how to reduce MTTP while being general enough to apply to most manufacturing situations.

The tutorial on factors contributing to MTTP is presented next. The MTTP reduction framework is then presented, and the factor changes that will reduce each component of MTTP are discussed. The paper concludes with some general guidelines on focusing efforts to reduce MTTP.

2. Understanding the Factors Determining MTTP

2.1 Processing Time

Consider a simple manufacturing system consisting of two workstations (a workstation is either a
machine or a workbench where a worker performs the job, Cox and Blackstone 1998) that manufacture parts X and Y. Both parts must first go through workstation 1 (WS-1) and then through workstation 2 (WS-2). The processing time per part is 10 minutes at each workstation, move time between stations is instantaneous, parts arrive one at a time to the workstation, and no variability in arrival rates or processing time exists. Under these conditions, it would be possible to sequence the arrivals to the workstation so the next part doesn’t arrive until the current part is finished. As Figure 1a illustrates, if X and Y are processed consecutively, the MTTP for each part type is the sum of the processing times at each station for a total of 20 minutes. Given the current state of technology used to produce the parts, 20 minutes is the minimum MTTP possible, and it is a perfect system. Any increase in the processing time per part would increase the MTTP by the same amount.

<table>
<thead>
<tr>
<th>Factor</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setup time</td>
<td>Burgess, Morgan, and Vollmann (1993)</td>
</tr>
<tr>
<td></td>
<td>Flynn (1987)</td>
</tr>
<tr>
<td></td>
<td>Garza and Smunt (1991)</td>
</tr>
<tr>
<td></td>
<td>Hopp and Spearman (2001)</td>
</tr>
<tr>
<td></td>
<td>Leu, Russell, and Huang (1993)</td>
</tr>
<tr>
<td></td>
<td>Morris and Tersine (1990)</td>
</tr>
<tr>
<td></td>
<td>Suresh and Meredith (1994)</td>
</tr>
<tr>
<td></td>
<td>Wemmerlov (1992)</td>
</tr>
<tr>
<td></td>
<td>Yang and Jacobs (1992)</td>
</tr>
<tr>
<td>Processing time per part</td>
<td>Suresh and Meredith (1994)</td>
</tr>
<tr>
<td>Move time</td>
<td>Hopp and Spearman (2001)</td>
</tr>
<tr>
<td></td>
<td>Leu, Russell, and Huang (1993)</td>
</tr>
<tr>
<td></td>
<td>Morris and Tersine (1990)</td>
</tr>
<tr>
<td></td>
<td>Shafer and Meredith (1993)</td>
</tr>
<tr>
<td></td>
<td>Suresh and Meredith (1994)</td>
</tr>
<tr>
<td></td>
<td>Shafer and Charnes (1995)</td>
</tr>
<tr>
<td>Production batch size</td>
<td>Ang and Willey (1984)</td>
</tr>
<tr>
<td></td>
<td>Hopp and Spearman (2001)</td>
</tr>
<tr>
<td></td>
<td>Karmarkar (1987)</td>
</tr>
<tr>
<td></td>
<td>Karmarkar, Kekre, and Kekre (1985)</td>
</tr>
<tr>
<td></td>
<td>Shafer and Charnes (1993)</td>
</tr>
<tr>
<td></td>
<td>Suresh and Meredith (1994)</td>
</tr>
<tr>
<td></td>
<td>Yang and Jacobs (1992)</td>
</tr>
<tr>
<td>Transfer batch size</td>
<td>Ang and Willey (1984)</td>
</tr>
<tr>
<td></td>
<td>Hopp and Spearman (2001)</td>
</tr>
<tr>
<td></td>
<td>Jacobs and Bragg (1988)</td>
</tr>
<tr>
<td>Arrival variability</td>
<td>Hopp and Spearman (2001)</td>
</tr>
<tr>
<td></td>
<td>Jensen, Malhotra, and Philipoom (1996)</td>
</tr>
<tr>
<td></td>
<td>Morris and Tersine (1990)</td>
</tr>
<tr>
<td></td>
<td>Suresh and Meredith (1994)</td>
</tr>
<tr>
<td></td>
<td>Garza and Smunt (1991)</td>
</tr>
<tr>
<td></td>
<td>Hopp and Spearman (2001)</td>
</tr>
<tr>
<td></td>
<td>Leu, Russell, and Huang (1993)</td>
</tr>
<tr>
<td></td>
<td>Moily and Stinson (1987)</td>
</tr>
<tr>
<td></td>
<td>Suresh and Meredith (1994)</td>
</tr>
<tr>
<td>Resource utilization and/or resource availability</td>
<td>Ang and Willey (1984)</td>
</tr>
<tr>
<td></td>
<td>Athersmith and Crokall (1974)</td>
</tr>
<tr>
<td></td>
<td>Burgess, Morgan, and Vollmann (1993)</td>
</tr>
<tr>
<td></td>
<td>Hopp and Spearman (2001)</td>
</tr>
<tr>
<td></td>
<td>Morris and Tersine (1989)</td>
</tr>
</tbody>
</table>
2.2 Production and Transfer Batch Sizes

Production batch sizes (that is, the number of parts of the same type processed before the workstation is set up to process a different part) and transfer batch sizes (the number of parts moved at the same time to the next workstation) of one unit are often unrealistic due to machine setup times and material handling constraints, respectively. Further realism can thus be incorporated into the example by first increasing both the production and transfer batch size for each part to 10 units. Under these conditions, each part spends 100 minutes at each station for a total MTTP of 200 minutes (see Figure 1b). Each part incurs only 20 minutes of actual processing time. The remaining 180 minutes is either time a part spends waiting for its turn to be processed at a workstation, or time the part spends waiting for the remaining parts in the batch to be processed so the batch can be moved. These wait times are sometimes referred to as wait-in-batch and wait-to-batch times, respectively (Hopp and Spearman 2001), or collectively as the wait-for-lot time (MPX 1996). The wait-for-lot time incurred by each part in this case is linearly related to the size of the production and transfer batches used. This causes MTTP to also increase in a linear fashion as production and transfer batch sizes increase.

2.3 Setup and Move Time

Further realism can be entered into the hypothetical system by requiring a setup time of 40 minutes before each batch is processed and including a 15 minute batch move time between machines. If no other changes are made to the process, MTTP increases by 95 minutes, causing the total MTTP to increase to 295 minutes (see Figure 1c). Any further increases in setup and move time would directly increase MTTP by the same amount.

2.4 Variability

Assuming the same production cycle continuously repeats in the examples shown in Figures 1a, 1b, and 1c, no idle time will exist at either station once the system fills with work (that is, once WS-2 starts processing the first part), causing the steady-state utilization to be 100%. This can only happen in a system with no variability. Because such systems don’t exist in reality, variability is introduced and examined in the hypothetical system.

Variability can be a result of either controllable or random variation (Hopp and Spearman 2001). Controllable variation is a result of decisions made and includes such things as differences in the processing time of different parts due to design differences, differences in wait-for-batch time due to production and transfer batch size decisions, and so on. In contrast, random variation is a result of events beyond our immediate control. This includes such things as natural variation in process time for the same type of part due to unplanned machine downtime or differences in machines, operators, or material; variation in the time between arrivals to each workstation, etc. Regardless of the type, variability generates the possibility that a batch of parts arriving to the workstation will find the workstation still busy processing a previous batch. When this happens, the new batch must join the queue and wait its turn for processing.

For example, suppose in Figure 1c that variability caused the processing time for the batch of X at WS-1 to be 110 minutes instead of 100 minutes, and at WS-2 to be 90 minutes instead of 100 minutes. In addition, the batch of Y arrives at WS-1 at 130 minutes, which is 10 minutes earlier than planned. The impact on MTTP is shown in Figure 1d. The early arrival of the batch of Y to WS-1 coupled with the extended batch processing time of X at WS-1 caused an initial wait time of 20 minutes for the batch of Y at WS-1. This wait time is called queue time. The extended batch processing time of X at WS-1 also delayed the arrival of the batch of Y to WS-2 by 10 minutes (when compared to Figure 1c), which caused 10 minutes of idle time between the completion of X at WS-2 and the start of Y. The net result is an MTTP for X that is the same as in Figure 1c (i.e., 295 minutes), but an MTTP for Y that is 20 minutes longer than in Figure 1c (i.e., 315 minutes instead of 295 minutes).

Increases in variability cause queue size and its associated queue time to increase. For example, suppose variability caused the batch of Y to arrive at the same time as the batch of X, but all other conditions are the same as in Figure 1d. As Figure 1e shows, the MTTP for X remains unchanged at 295 minutes, but the MTTP for Y has now increased by the additional 130 minutes of queue time for a total MTTP of 445 minutes. When variability of all kinds is considered, queuing theory indicates that queue size and
Figure 1
Impact of Batch Size, Setup Time, Move Time, and Variability on MTTP

\[
\begin{align*}
&\text{PX} = \text{Processing time for X} & \text{PY} = \text{Processing time for Y} \\
&S_X = \text{Setup time for X} & S_Y = \text{Setup time for Y} \\
&M_X = \text{Move time for X} & M_Y = \text{Move time for Y} \\
&\text{WS-1} = \text{Workstation 1} & \text{WS-2} = \text{Workstation 2} \\
&Q_X = \text{Queue time for X} & Q_Y = \text{Queue time for Y} \\
&I = \text{Idle time} & \text{Fork} = \text{Forklift}
\end{align*}
\]
its corresponding queue time increases at an increasing rate as the standard deviation or coefficient of variation of interarrival and/or processing time increases (see Figure 2).

In assembly or joining operations, variability can also cause a part to arrive at a workstation before its mate(s). When this happens, wait time can occur, even though the workstation is available for setup and processing of the part. Although this waiting time is often included as part of queue time, it is also sometimes referred to as wait-to-match time (Hopp and Spearman 2001).

2.5 Utilization

Variability has less impact on queue time when workstation utilization is low than when workstation utilization is high. When utilization is low and significant slack workstation capacity exists, it is fairly easy for a batch to arrive when the workstation is idle and be processed immediately. However, as utilization increases and less slack capacity is available, it becomes more difficult for a batch to arrive when the workstation is idle. This increases the probability that the batch must join the queue, resulting in longer queue times and MTTP.

For example, suppose batches of parts arrive to a single workstation on average every 10 hours, each batch contains 10 parts, and the average batch processing time is 6 hours. However, due to variability, the actual interarrival and processing times deviate from the average. The actual interarrival and processing times for four batches of different parts arriving to this workstation are shown in Figure 3a. In this case, the workstation is idle when each batch arrived, the average utilization is 60%, and the average MTTP is 6 hours. In Figures 3b, 3c, and 3d, the average utilization of the workstation is increased to 70%, 80%, and 90%, respectively, by decreasing the average time between batch arrivals to 8.6, 7.5 hours, and 6.7 hours, respectively, while simultaneously keeping the absolute deviations from the average interarrival time for each batch the same as in Figure 3a. As shown, the increases in utilization caused the batch of X to incur queue time of 0.4 hours in Figure 3b; batches of X and Y to incur queue times of 1.5 and 1.0 hours, respectively, in Figure 3c; and batches of X and Y to incur queue times of 2.3 and 2.6 hours, respectively, in Figure 3d. If each part cannot leave the station until the entire batch has been processed, these queue times caused the average MTTP to increase at an increasing rate with successive increases in utilization.

The magnitude of the impact that utilization and variability have on MTTP will vary from system to system. However, queuing theory indicates the general pattern of results shown in Figure 3 holds for all systems, namely that queue time and its associated MTTP increase at an increasing rate as utilization increases (see Figure 4). Furthermore, queue time and MTTP at a workstation with high variability will increase faster as utilization increases than will queue time and MTTP at a workstation with low variability.

2.6 Factor Interactions

The preceding discussion indicates that MTTP is equal to the sum of the processing, setup, move, queue, wait-in-batch, wait-to-batch, and wait-to-match times. Because queue, wait-in-batch, wait-to-batch, and wait-to-match times all involve waiting, and because actions to reduce one type of waiting may also reduce other forms of waiting, they are collectively referred to as waiting time in the MTTP reduction framework. Reductions in MTTP thus require reductions in one or more of these components. While setup time, processing time per part, and move time are independent of each other (i.e., a reduction in move time does not affect setup time or processing time per part, and so on), changes in any of these three compo-
Figure 3

Impact of Utilization on MTTP

(a) Batch of W arrives

Average batch interarrival time = 10.0 hrs.
Average batch processing time = 6.0 hrs.
Average utilization = 60%
Average MTTP = 6.0 hrs.

(b) Batch of W arrives

Average batch interarrival time = 8.6 hrs.
Average batch processing time = 6.0 hrs.
Average utilization = 70%
Average MTTP = 6.1 hrs.

(c) Batch of W arrives

Average batch interarrival time = 7.5 hrs.
Average batch processing time = 6.0 hrs.
Average utilization = 80%
Average MTTP = 6.6 hrs.

(d) Batch of W arrives

Average batch interarrival time = 6.7 hrs.
Average batch processing time = 6.0 hrs.
Average utilization = 90%
Average MTTP = 7.2 hrs.
ponents can affect waiting time (Hyer and Wemmerlöv 2002). Consequently, one way to reduce waiting time is to manipulate the other three components of MTTP.

For example, if the average processing time per part is reduced to 5 minutes for each part type at each workstation in Figure 1e while all other conditions remain the same, Y would only wait 100 minutes at WS-1 and the MTTP would be 295 minutes (see Figure 5). Reducing batch processing time by 100 minutes for each part (i.e., 50 minutes at station 1 and 50 minutes at station 2) in this case actually caused a 150 minute reduction in MTTP for Y due to the additional impact on waiting time at WS-1.

3. Manufacturing Throughput Time Reduction Framework

3.1 Overview of Framework

Figure 6 presents the MTTP reduction framework. The framework can be described as a flowchart with five columns. Column 1 lists the objective of the framework as the reduction in MTTP. Column 2 presents the components of MTTP. Setup time is the sum of the times spent setting up all workstations required to process the part through the production system. Processing time is the sum of the times spent processing a part at each workstation required in the production routing for the part. Move time is the sum of times spent moving a part between each workstation in the production routing for the part. Waiting time is the sum of the times spent waiting at the queue, wait-in-batch, wait-to-batch, and wait-to-match times at all workstations in the production routing for the part. Waiting time is usually the largest of the four components, accounting for as much as 90% of manufacturing lead time in some systems (Houtzeel 1982). Column 3 illustrates the factors that will reduce each component. Column 4 specifies actions that will alter each factor shown in column 3, and column 5 presents important changes that might be required to enable some of the actions shown in column 4. The feasibility of accomplishing some of the actions and changes shown in columns 4 and 5 are directly related to the type of production layout used (i.e.,

\[\text{MTTP}_X = 195 \text{ min.} \]
\[\text{MTTP}_Y = 295 \text{ min.} \]
job shop/functional layout, cellular layout, or product layout/assembly line). The issue of layout choice will be included in the following discussion where appropriate.

Based on these definitions, one or more of these four components must be reduced in order to reduce MTTP; by following the flowchart from left to right, actions that will reduce each component can be identified. This flowchart is intended to provide a structured way to examine the types of actions that can be taken to reduce MTTP and the relationships between these actions. The following sections briefly discuss how to reduce each component of MTTP.

3.2 Setup Time Reduction

Column 3 of Figure 6 indicates that setup time reductions can be accomplished by reducing the time per setup and/or the number of setups. Time per setup can be reduced by purchasing equipment with short setup times, improving setup procedures, dedicating workstations to families of parts with similar setup requirements so that common fixtures can be used and developed, and/or by using family scheduling to group batches that have common setup requirements. Workstation dedication and family scheduling can also reduce the number of setups required. Further information on improving setup procedures can be found in works by Steudel and Desruelle (1992) and Shingo (1985).

3.3 Processing Time per Part Reduction

Column 3 of Figure 6 indicates that reductions in processing time per part can be accomplished by reducing the number of operations required, reducing the processing time per operation, and/or reducing scrap and rework. The number of operations per part may be reduced through the adoption of new technology that allows a single operation to do what was previously done by several operations, or by redesigning the part so that fewer operations are required. Processing time per operation can be reduced by redesigning the part to require less processing, incorporating faster technology to process the part (if available), or dedicating labor to a family of parts with similar processing requirements. Labor dedication allows the workers processing the parts to become more familiar with a smaller family of parts, thus potentially reducing the amount of time spent reading blueprints, setting machine speeds, performing quality inspections while the parts are on the machine, and so on.

The best way to reduce scrap and rework is to improve raw material quality to prevent defective material from entering the system, and to improve equipment capabilities, processes, and procedures to prevent scrap and rework from happening in the first place. Implementing poka-yoke (fail-safe) devices can be especially beneficial in this respect. Using one-piece flow (or very small transfer batches) can also reduce scrap and rework because defective parts can be quickly detected at the next operation. One-piece flow is often impractical in a job shop/functional layout due to the increased material handling, production control, scheduling, and/or information systems requirements such a change would entail. In contrast, one-piece flow can often be used in a cellular or product-oriented layout with little impact on the same requirements. As a last resort, increased inspection of the parts to identify defective units and prevent them from being transferred to the next operation can be used to improve scrap and rework.

3.4 Move Time Reduction

Column 3 of Figure 6 indicates that reductions in move time can be accomplished by reducing either the time required per move or the number of moves required. The time required per move can be reduced by increasing the speed of the material handling equipment (which may not be possible due to safety implications), or by reducing the move distance required. If the speed of the material handling system is increased through the installation of conveyors or other automated handling equipment, it is questionable how realistic this option would be when a job shop/functional layout is used. While move distance can sometimes be reduced by reorganizing the equipment to optimize the material handling between departments in a job shop/functional layout, the amount of reduction is greater if the equipment performing sequential operations on a part is grouped to form manufacturing cells.

If a job shop or functional layout is currently being used, the number of moves requiring material handling equipment can often be reduced by grouping workstations performing sequential operations
Figure 6
Manufacturing Throughput Time per Part (MTTP) Reduction Framework

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective</td>
<td>Component Changes That Will Reduce MTTP</td>
<td>Factors That Will Change Component</td>
<td>Actions That Will Alter Factors in Column 3</td>
</tr>
<tr>
<td>Reduce Setup Time</td>
<td>Reduce Time/Setup</td>
<td>Purchase Equipment With Short Setup Time</td>
<td>Increase Material Handling Capacity (if constrained) or Increase Workstation Capacity (if constrained) or Reduce Need for These Systems</td>
</tr>
<tr>
<td>Reduce Manufacturing Throughput Time Per Part (MTTP)</td>
<td>Reduce # Setup</td>
<td>Improve Procedures</td>
<td>Group Equipment Performing Sequential Operations</td>
</tr>
<tr>
<td>Reduce Move Time Per Part</td>
<td>Reduce Scrap and/or Rework</td>
<td>Dedicate Equipment</td>
<td>Increase Production Control, Scheduling and/or Information System Capabilities (if needed) or Reduce Need for These Systems</td>
</tr>
<tr>
<td>Reduce Processing Time Per Part</td>
<td>Reduce Number of Operations</td>
<td>Family Scheduling</td>
<td>Optimize Current Layout</td>
</tr>
<tr>
<td>Reduce Move Time Per Part</td>
<td>Reduce Time/Operation</td>
<td>One-piece Flow</td>
<td>Group Equipment Performing Sequential Operations</td>
</tr>
<tr>
<td>Reduce Waiting Time Per Part</td>
<td>Reduce Time/Move</td>
<td>Improve Procedures</td>
<td>Increase Material Handling Capacity (if constrained) or Increase Workstation Capacity (if constrained) or Reduce Need for These Systems</td>
</tr>
<tr>
<td>Reduce Production Batch Size</td>
<td>Reduce Number of Moves</td>
<td>Improve Equipment Capability</td>
<td>Increase Production Control, Scheduling and/or Information System Capabilities (if needed) or Reduce Need for These Systems</td>
</tr>
<tr>
<td>Reduce Transfer Batch Size</td>
<td>Reduce Production Variability</td>
<td>Improve Raw Material Quality</td>
<td>Reduce Time Required</td>
</tr>
<tr>
<td>Reduce Setup Time, Processing Time/Part, and/or Move Time (See Column 2)</td>
<td>Standardize Part Design</td>
<td>Increase Raw Material Quality</td>
<td>Cross-Train Workers</td>
</tr>
<tr>
<td>Reduce Arrival Variability</td>
<td>Dedicate Labor and Equipment</td>
<td>Increase Inspection</td>
<td>Increase Equipment Pooling</td>
</tr>
<tr>
<td>Reduce Workstation Utilization</td>
<td>Stabilize Batch Sizes</td>
<td>Improve Technology</td>
<td></td>
</tr>
<tr>
<td>Increases Resource Access</td>
<td>Improve Preventative Maint.</td>
<td>Reduce Processing Variability (See Column 2)</td>
<td></td>
</tr>
<tr>
<td>Reduce Number of Queues</td>
<td></td>
<td>Reduce Processing Variability (See Column 2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control Order Releases</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Improve Coordination</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Increase Time Available</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce Time Required</td>
<td></td>
</tr>
</tbody>
</table>
into manufacturing cells. In some cases, technological improvements that allow more sequential operations to be done by a single machine can achieve the same result (for example, a CNC milling machine is used to perform the operations previously done by several machines).

3.5 Waiting Time Reduction

3.5.1 Overview

Column 3 of Figure 6 indicates that reductions in waiting time can be accomplished by reducing setup time, processing time per part, move time, production batch sizes, transfer batch sizes, processing time variability, arrival variability, resource utilization, and/or the number of queues. It can also be reduced by increasing access to resources. Reductions in setup time, processing time per part, and move time have already been discussed. The remaining factor changes will be discussed in the following sections.

3.5.2 Production Batch Size Reduction

Production batch size reduction is often the easiest and most cost-effective way to reduce waiting time and MTTP in most plants. Not only does it reduce the wait-for-lot time for the part in question, but it also reduces queuing time for parts in other batches as well. For instance, consider the example in Figure 1e. The average processing time per part for X is 110 / 10 = 11 minutes at WS-1 and 90 / 10 = 9 minutes at WS-2. Because only one part in the batch is processed at a time, 9 parts are always waiting, resulting in a wait-for-lot time of 11 * 9 = 99 minutes at WS-1 and 9 * 9 = 81 minutes at WS-2. The queue time for Y is 150 minutes at WS-1 and the wait-for-lot time at both WS-1 and WS-2 is (100 / 10) * 9 = 90 minutes. This produces a MTTP for X and Y of 295 minutes and 445 minutes, respectively. In contrast, if production and transfer batch sizes are reduced to 5 parts for both X and Y but all other conditions remained the same, the wait-for-lot time for X is reduced to 11 * 4 = 44 minutes at WS-1 and 9 * 4 = 36 minutes at WS-2 (see Figure 7). Queue time for Y drops to 95 minutes at WS-1 and the wait-for-lot time at both WS-1 and WS-2 drops to (50 / 5) * 4 = 40 minutes. The net result of the batch size reduction is that MTTP is reduced to 195 minutes for X and 290 minutes for Y.

To reduce batch sizes, the plant needs to implement a policy to schedule production of smaller batches. However, if demand stays constant, smaller batch sizes increase the number of setups required. As the number of setups increases and more of the available capacity is used for setups, workstation utilization increases, which causes queues to grow. Eventually, the increased queues negate any benefit to be obtained from batch size reduction and MTTP increases rapidly (see Figure 8). Reducing setup time would allow further batch size and MTTP reduction.

**Figure 7**
Impact of Batch Size Reduction on MTTP Compared to Figure 1e
When batches are transferred between workstations by forklift, handcart, or another similar conveyance device, batch size reduction also increases the number of trips required. The increased number of trips raises the utilization of the forklift, which causes increased queuing. If utilization increases enough, the increased queues counteract any benefit to be obtained from batch size reduction, and MTTP increases rapidly in the same manner as previously described for the impact of batch size reduction on setup time.

Batch size reduction also increases the number of different batches of product on the shop floor at any one time, which may increase the load on the production control, scheduling, and/or information systems.

Based on this discussion, if MTTP is to be reduced through batch size reduction, one or more of the following changes are often required (see Column 5 in Figure 6):

1. Workstation capacity must be increased (if capacity is constrained) or setup times reduced.
2. Material handling capacity must be increased (if capacity is constrained) or the workstations required to process a batch be consolidated so that material handling equipment is not needed as often.
3. The capabilities of the production control, scheduling, and/or information systems must be increased (which may include increases in both labor and computer capacity) to handle the increased requirements or the need for these systems reduced.

If production is performed using a job shop/functional layout, the spatial separation of workstations and labor resources required to produce the batch of parts will likely require increases in workstation and material handling capacity and production control, scheduling, and/or information systems capabilities as batch sizes are reduced. In contrast, if cells are formed, workstations and labor are dedicated to families of parts and grouped in close proximity. This dedication and grouping reduces setup time and often allows the parts to be transferred between workstations by hand or by small conveyors, thus eliminating the need for forklifts and other material handling equipment. Cells reduce the amount of centralized scheduling required because only the cell must be scheduled rather than each workstation. Tracking of parts is less because the parts are either in one of the cells or the order hasn’t been started yet. Finally, reduced scheduling and tracking requirements may reduce the amount of computer information system capacity needed (if a computerized information system was used) and the amount of time needed to enter data, maintain the system, etc. Thus, converting a job shop/functional layout to a cellular layout would likely allow batch size reduction without corresponding increases in machine capacity, material handling, production control, scheduling, and information system capacity/capabilities. In fact, the use of cells may result in less need for these systems, even though batch sizes are reduced.

3.5.3 Transfer Batch Size Reduction

If production batch sizes cannot be reduced, waiting time can still be reduced through the use of transfer batches smaller than the production batch size. For example, suppose in Figure 1e that setup times cannot be reduced below 40 minutes, which prevents production batch size reductions. However, material handling capacity is such that transfer batches of five parts could be used. The impact of this change is illustrated in Figure 9. As shown, the transfer batch size reduction reduced the wait-to-batch time for the first transfer batch of X and Y at WS-1 to $11 \times 4 = 44$ and $10 \times 4 = 40$ minutes, respectively. This allowed these transfer batches to be moved to WS-2 earlier than in Figure 1e, and because WS-2 was idle, it could begin processing the transfer batches immediately. Thus, the first transfer batch of X was being processed at WS-2 at the

![Figure 8](image-url)  
**Figure 8**  
MTTP vs. Batch Size. Note: Graph constructed using standard queuing theory formulas.
same time as the second transfer batch for X was being processed at WS-1. Similar results occurred for Y. Even if both transfer batches must be combined before leaving WS-2, the net result is a reduction in MTTP for X of 295 – 240 = 55 minutes and for Y of 445 – 395 = 50 minutes when compared to Figure 1e.

Transfer batch size reduction has the same implications for material handling capacity, production control, scheduling, and information system capabilities as those previously discussed for batch size reduction, but it does not influence the number of setups required if all transfer batches of the same production batch are processed consecutively before parts of a different type are processed. Transfer batch size reduction also has less impact on material handling capacity, production control, scheduling, and information system capacity if manufacturing cells are used versus a job shop layout.

### 3.5.4 Processing Time Variability Reduction

Variability in processing time comes from several sources: variance in setup time for a workstation, variance in the processing time per part, variance in the size of the batch processed, and variance due to unplanned downtime and repair of the workstation. Reducing any of these sources of variability will reduce processing time variability and, consequently, waiting time as well. Grouping similar jobs based on part family affiliation, dedicating equipment and labor to these part families, and/or standardizing part design will help reduce the variance associated with setup times and processing time per part. Stabilizing or establishing similar batch sizes for all jobs in the family will help reduce variance associated with batch size differences. Improvements in preventive maintenance will help reduce variance associated with unplanned downtime and repair of the workstation.

### 3.5.5 Arrival Variability Reduction

Reductions in arrival variability will also reduce waiting time. Arrival variability is more complex than processing variability and is dependent on the variability of new orders released directly to the workstation, as well as the departure variability from any upstream workstations that feed the station in question. When workstation utilization is high, each job is extremely likely to arrive when the workstation is busy and, consequently, is likely to have to join the queue. As a result, the departure variability from the

---

**Figure 9**

Impact of Transfer Batch Reduction on MTTP Compared to Figure 1e

- $P_{X_i}$ = Processing time for X, transfer batch $i$
- $P_{Y_i}$ = Processing time for Y, transfer batch $i$
- $S_X$ = Setup time for X
- $S_Y$ = Setup time for Y
- $M_{X_i}$ = Move time for X, transfer batch $i$
- $M_{Y_i}$ = Move time for Y, transfer batch $i$
- $Q_{X_i}$ = Queue time for X, transfer batch $i$
- $Q_{Y_i}$ = Queue time for Y, transfer batch $i$
- WS-1 = Workstation 1
- WS-2 = Workstation 2
- Fork = Forklift
workstation is primarily dependent on the processing variability at the station. In contrast, when workstation utilization is low, the workstation is idle a substantial portion of the time and each job arriving to the station is more likely to find the station idle. In this case, variability in the time between arrivals tends to directly impact departure variability. In addition, departure variability is reduced as the number of identical copies of the resource at the station increase (Hopp and Spearman 2001, p263). This is a direct result of resource pooling. More will be said about this impact in section 3.5.7.

Regardless of the utilization level, any changes that reduce variability in the time between arrivals or in the actual processing at the workstation will reduce departure variability. Processing variability has already been discussed. Variability in the time between the arrivals of new orders can be reduced through the use of controlled order release mechanisms. Such mechanisms stabilize the production schedule by releasing new orders to the workstation when the queue reaches a set level. For assemblies that require two or more components to start production of the job, any changes in production control that improve the coordination of the arrival of the components will also reduce arrival variability.

3.5.6 Workstation Utilization Reduction

As discussed in section 2.5, wait time is heavily influenced by workstation utilization. Workstation utilization can be defined as “the total workstation time required per period divided by the total workstation time available per period.” In this framework, the total workstation time required per period is equal to the sum of the times spent setting up the workstation, processing parts, waiting for labor to become available, and waiting for the equipment to be repaired. This is similar to the definition used in queuing packages like MPX (MPX 1996). The total workstation time available per period is equal to the sum of the times each identical unit of the resource at the workstation is available to be used. Thus, for example, if the workstation has two semi-automated machines operated by a single worker, each machine is available eight hours per day, and on average a total of two hours are spent setting up the machines, ten hours are spent processing parts, one hour is spent waiting for labor, and unplanned downtime equals one-half hour each day, the average workstation utilization = (2 + 10 + 0.5 + 1) / (8 + 8) = 84.4%. Workstation utilization will decrease if the total time required per period is reduced proportionately more than the total time available per period is reduced, or if the total time required per period is increased proportionately less than the total time available per period is increased.

The time available per period can be increased by adding equipment if capacity is machine constrained, adding workers (and possibly extra shifts) if capacity is worker constrained, and reducing absenteeism. The capacity or time required can be reduced by reducing the arrival rate of jobs to the workstation (which will reduce output), and/or by reducing setup time, processing time per part, equipment downtime, scrap and rework, and delays due to unavailability of workers. Reducing delays due to unavailability of workers may require adding additional workers (which also increases capacity), re-assigning worker responsibilities to better balance the load, or cross-training workers to handle multiple tasks. In the case of cross-training, workers can float to the workstation or resource experiencing the most delays. This will reduce the utilization of equipment, but it will not necessarily increase the overall average worker utilization because it may simply change when and which worker is idle, rather than the total amount of idle time. If this occurs, resource availability is increased without increasing utilization, and wait time goes down.

3.5.7 Increase Resource Access

Figure 6 indicates that waiting time can also be reduced by increasing access to resources. While resource access can be increased by purchasing equipment, hiring workers, working overtime, etc., the intent of this factor is to increase resource access without incurring these additional costs. Using cross-trained workers and increasing equipment pooling can sometimes accomplish both of these goals. Using cross-trained workers has previously been discussed and will not be mentioned further.

To understand how equipment pooling can increase resource access and reduce waiting time, consider the case where parts A and B both require a milling operation (as well as other operations not requiring a mill) and this operation can be done on either Mill 1 or Mill 2. However, the two mills are currently located in different areas of the plant and
Mill 1 is dedicated to the production of A and Mill 2 is dedicated to the production of B. While this may have its advantages, it does create the possibility that Mill 1 is starved for work due to variability in demand, variability in processing times at previous stations, workstation downtime at a previous station, etc., while Mill 2 has a queue of B waiting for processing. Thus, B incurs waiting time even though a mill with the required capabilities is currently sitting idle in another area of the plant. This would not happen if the two mills were pooled (i.e., resource pooling is increased) by locating them in close proximity and feeding them with a common queue of work. Whenever a mill in the pool becomes idle, it would begin processing the next job in the queue. This can reduce waiting time and MTTP for A and B, provided the increase in equipment pooling doesn’t increase setup times, processing times, move times, variability, etc., to the point where the impact of these increases overcomes any potential waiting time reduction resulting from the pooling increase. Due to the complex interaction of such factor changes, queuing theory or simulation models are often required to determine if MTTP would be reduced through increases in equipment pooling.

3.5.8 Reduce Number of Queues

The final way to reduce waiting time is to reduce the number of queues by increasing the number of successive operations that the same worker or machine performs. For example, suppose a metal part requires several different milling, drilling, and tapping operations and these operations are currently done on three different machines made specifically for that purpose. At each machine, the part may have to join a queue to wait its turn for processing. In contrast, if all these operations can be done on a CNC milling machine, the queues between operations are eliminated. The elimination of wait time will reduce MTTP, provided any increase in setup and processing time resulting from the use of the CNC milling machine rather than the specialized equipment is less than the amount of time the part normally spends waiting. Similarly, cross-training workers to perform multiple assembly tasks that were previously done by separate workers will reduce MTTP, provided any increase in task time resulting from the loss of specialization is less than the waiting time eliminated.

4. Conclusion

Manufacturing throughput time reduction can often be a daunting and confusing task due to the large number of factors that can be changed and the interactions between them. This paper provides a brief tutorial that illustrates the basic factors that determine MTTP and explains why each factor impact occurs. This tutorial can be used to educate workers on these basic concepts. The paper also presents a conceptual framework that illustrates the actions that can be taken to reduce each factor, and the relationships between them. This framework provides an easy-to-use tool that managers can use to determine a course of action to reduce MTTP in their own plants.

Because all of the actions listed in column 4 of the framework (Figure 6) can be used to reduce MTTP, manufacturing plants striving to reduce throughput time must decide where to focus their efforts. The answer to this question will vary between plants, but the following guidelines can be offered:

1. Production and transfer batch size reductions offer the largest potential for MTTP in most plants. If the plant has a job shop/functional layout in place, significant reductions in batch size may require conversion to manufacturing cells (for reasons listed in column 5 of the framework). Converting to cells may also reduce move time per part, processing time per part, processing variability, and arrival variability, further reducing MTTP beyond that achieved through production and transfer batch size reductions alone.

2. High workstation utilization is a major contributor to long MTTP, especially in cases where variability is high. If variability cannot be reduced, workstation utilization must be reduced to lower throughput times. In general, workstation utilization levels in the 75–80% range may be required on critical resources to keep MTTP low (Suri 1998).

3. Many causes of long MTTP are a result of policies and procedures implemented in the past that are used to control production batch sizes, transfer batch sizes, workstation utilization, resource access, and so on. With the current emphasis on short manufacturing throughput times, the relationships illustrated in the throughput time reduction framework...
and the principles discussed in this paper should be used to evaluate whether these decisions should be changed.

A version of the framework presented in this paper has been used by engineers and managers at a sheet metal products fabrication plant to guide their MTTP reduction efforts. The plant's engineering manager stated that many of the relationships contained in the flowchart are concepts they have been trying to teach their employees. They found that the flowchart organizes these concepts and relationships in a way that is easy for the employees to understand. The framework has also been used to teach manufacturing throughput time reduction concepts to university students enrolled in production/operations management classes. These experiences indicate the framework is a useful tool for understanding the actions that can be taken to reduce manufacturing throughput time per part and the relationships between them and for guiding throughput time reduction efforts in real manufacturing plants.

References


**Author’s Biography**

Danny J. Johnson is an assistant professor of operations management at the College of Business at Iowa State University. He holds a BS in business administration from Moorhead State University and an MBA and a PhD in operations management from the University of Wisconsin–Madison. Prior to obtaining his BS, he worked for eight years in the service sector. Dr. Johnson’s research interests are in the design, implementation, operation, and management of quick response manufacturing systems and the problems faced by firms as they attempt to develop and use these systems to improve key performance measures. He has conducted and assisted with studies on the implementation of cellular manufacturing systems in industry, and two case studies from these research projects have been published as chapters in books on cellular manufacturing. He has also published articles on cellular manufacturing in the *International Journal of Production Research* and *Production and Operations Management*. He is certified in production and inventory management by the Educational Society for Resource Management (APICS) and is a member of the Educational Society for Resource Management, the Decision Sciences Institute, and the Production and Operations Management Society.