Analysis of Constraint Location in a Lean Facility

Trumone Sims and Hung-da Wan
Center for Advanced Manufacturing and Lean Systems
and Mechanical Engineering Department
University of Texas at San Antonio
San Antonio, Texas 78249, U.S.A.

Abstract

Theory of Constraints, with its five steps for constraint elimination, is viewed as a continuous improvement process. This paper demonstrates the theory's potential use in the early planning stages, during the design of a manufacturing system. The constraint resource and its location can be pre-determined with the constraint's location being anywhere in the system. Having the constraint in different locations (i.e., upstream or downstream) can result in different behaviors of the system. The impact on a matured lean system with moving assembly line needs to be investigated further. In this paper, a lean manufacturing system with a series of tandem moving assembly lines and single piece flow will be used to analyze different management philosophies on constraints. Through discrete event simulation, the manufacturing system constraint will be moved to different locations to determine what effect the location of the constraint has on throughput.

Keywords
Theory of Constraints, Constant Work-In-Process (CONWIP), Continuous Moving Conveyors, Effective Utilization Analysis

1. Introduction

1.1 Motivation and Research Objectives
The manufacturing system output is a function of the whole system, not just individual processes. When we view our system as a whole, we realize that the system output is a function of the weakest link. The weakest link of the manufacturing system is the constraint. Consequently, there needs to be focus on the coordination of efforts to optimize the overall system, not just individual processes [1]. When a system matures in lean implementation, the main constraint becomes less obvious. However, the impact of performance of constraining resources in a lean system, especially with moving assembly lines, is still evident. This research attempts to investigate the impact of location of constraints in a system. To facilitate that, the concept of “constraint” should be reviewed first.

Theory of Constraints (TOC) is a well-known methodology for systems improvement that includes principles and practice guidelines that can be adopted by practitioners [2]. The famous novel for operations management, The Goal, written by Eli Goldratt [3] caught the attention of process improvement professionals onto this methodology. From this book, a five focusing steps (5FS) was brought out: 1) Identify the System Constraint, 2) Decide How to Exploit the Constraint, 3) Subordinate Everything Else, 4) Elevate the Constraint, and 5) Go Back to Step 1, but beware of “Inertia.” The optimization process begins in step 2 of the 5 step continuous improvement process. Step 2 is deciding how to exploit the manufacturing system’s constraint, where exploit means to get the most from the constraining element without additional investment [1].

Constraints are often referred to as bottleneck, which limits the performance of the whole system. By exploiting the constraint, we strive to maximize the utilization of the capability of the constraining component as it currently exists. In other words, TOC urges to rethink what we can do to get the most out of this constraint without committing to
potentially expensive changes or upgrades and implement in a short period of time [4]. Constraints can be both external and internal. External constraints are often beyond the control of management because they are market driven. External or market constraints affect demand, they influence product mix, which in turn affects resource utilization [5].

Lean manufacturing and TOC were two methodologies developed independently in the past. Based on Toyota Production System (TPS) and Just-in-Time (JIT) concept, the lean principles aims at eliminating waste to the maximum extent in order to improve the flow of the value stream [6]. Both Lean and TOC have proven themselves effective in productivity improvement in the past couple of decades. Nevertheless, some of the principles do not seem to agree totally between the two. Some related literature exist, such as the investigation of various configurations of pull production systems, including Kanban, Constant Work-In-Process (CONWIP) and different hybrid configurations that integrate the former two, while varying the location of the dominant constraint [7]. Watson et al. [2] discussed briefly about the differences between JIT and TOC and also pointed out that there is a need for more supporting literature.

1.2 Research Approach
This paper will explore the effects of the dominant constraint location on a manufacturing system that consists of paced moving assembly lines and individual automatic stations. Using simulation modeling, the performance of different scenarios can be compared.

A manufacturing system will be modeled following two different management philosophies. The first philosophy represents the typical manufacturing scenario where the constraint resource is located at the beginning of the assembly line. The downstream resources are allowed to run at a faster cycle time. The manufacturing system that follows this philosophy will be called Model A.

The second philosophy has the constraint resource at the end of the assembly line. This philosophy recommends that your nonconstraints have sprint capacity, that is, the capacity to produce product at faster rates than your constraint operation; thus, if an upstream nonconstraint operation experiences downtime for some reason, when it begins producing again, it should still have the ability to produce product at a fast enough rate to resupply the constraint buffer before the constraint buffer runs dry [8]. Model B will represent a manufacturing system following this philosophy, which is also called over speed. The key metric used for evaluating the models is throughput. Rockwell Arena simulation software was used to model the assembly operations.

In short, the two models compared in this paper are:
- **Model A**: Constraint located at the beginning of the assembly line.
- **Model B**: Constraint located at the end of the assembly line.

2. Automotive Plant as Choice for Study
In this paper, a real manufacturing company was used in the development of a simulation study. The manufacturing system being studied is the assembly operations in an automotive plant. There are several tandem, balanced, paced production lines. The plant produces two product families, which will be referred to as Part 1 and Part 2. Each production line has a fixed CONWIP level, but the level of the individual lines is not identical. The lines are decoupled from each other with Work-In-Process (WIP) buffers.

This system is considered a highly matured lean manufacturing system, and the buffer levels are typically well limited. As a result, the performance of constraints can bring a significant impact to the whole system. Figure 1 is a layout of the assembly operations.
2.1 Study Performance Measure Selection

It has been stated that performance analysis of production lines strive to evaluate their performance measures as a function of a set of system parameters. The most commonly used performance measures follow [9]:

- Throughput
- Average inventory levels in buffers
- Downtime probabilities
- Blocking probabilities at bottleneck workstations
- Average system flow times (also called manufacturing lead times).

TOC states that the performance of the weakest link determines the performance of the whole chain. During this analysis, the constraint will be located at the beginning and then at the end of the manufacturing system. Even though the constraint bottleneck rate will be the same for both models, the system’s throughput will change because of process variation and downtime. Throughput has been selected as the system performance measure.

By design, most manufacturing production lines provide for a limited amount of work-in-process inventory [10]. The manufacturing system being studied is no exception. Each assembly line is located between two buffers. Both buffers have two set points; one stops the assembly line while the other restarts it. The upstream buffer set points are called short and short reset, while the downstream buffer set points are full and full reset. The buffer set points control the average inventory levels in the buffer, therefore this measure will not be tracked during the study.

Downtime will be restricted to only one of the individual stations. The assembly lines will use another technique to account for downtime called the Utilization Method. This method will be discussed in greater detail in the next section.

Model A has the constraint located at the beginning of the manufacturing system with all of the downstream resources running at a faster pace. Since downtime for the assembly lines will be modeled without the actual assembly line stopping, the probability of being blocked will be small since it will be a function of the automatic stations’ downtime. System B has the constraint located at the end of the manufacturing system; therefore there are
Sims and Wan

no opportunities for the constraint to be blocked. Because of these reasons blocking probabilities at bottleneck workstations will not considered a performance measure.

The structure of the data collected for the study does not allow the accurate tracking of average system flow time. Therefore, flow time will not be considered a performance measure.

3. Model Construction

The plant runs a two shift operation with scheduled overtime. For this study only one shift with 10 hours of production will be modeled. The simulation run contains a warm-up period that allows the system to fill with parts. The replication length and hours per day are set to 18 hours. All of the model resources will follow a capacity schedule which matches the current day shift 10 hour production schedule.

3.1 Verification Using Quick Effective Utilization Analysis

Model verification was performed by inspecting the model statistics to verify that a proper flow of entities was maintained and by observing the model’s animation evolution to ensure that the manufacturing lean rules were being followed through the individual automatic stations, such as single piece flow and first in first out sequencing of parts. Figure 2 shows the logic for this section of the model.

Figure 2 - Individual stations simulation logic

Calculating the practical production rate, \( r_b^p \), for each of the production lines will the method in which downtime will be accounted for on continuous moving conveyors. The steps involved in calculating the practical production rate are covered in literature describing the Effective Utilization Method, which can be used for determining a manufacturing system’s constraint [11].

\[
W = r_b^p X T_0^P
\]  
\[
U = \frac{r_b^P}{r_b}
\]

A quicker and rougher technique for calculating the practical production rate is to use the Quick Utilization Method which will now be presented. From equation 1, which is Little’s law, we are able to calculate the practical production rate \( r_b^P \) [12]. The CONWIP level, \( W \), for the production line is obtained by waiting for a planned stop to occur, such as a break or lunch period, and then counting the number of parts on the line. To determine the minimum practical lead time, a small sample size of parts is selected to track through the production line. Start at the beginning of the line and record the time the part enters the line and the sequence/job number associated with the part. Walk the part through the production line and record the time the same part exits the line. Subtract the start time from the stop time to calculate the lead time. The goal is to include the time durations for minor stoppages on the production line in the lead time. The time durations that the production line spends in a blocked or starved state should not be included in the lead time calculation. Then calculate the sample set average lead time; this value is the minimum practical lead time, \( T_0^P \). Table 1 shows a data sheet for the Frame 2 / Final 1 production line. Several
columns in the table are left blank. The blank columns would have been used if the production line stopped for long
durations of time (i.e. breaks, lunch, and excessive downtime).

<table>
<thead>
<tr>
<th>FR2_FN1</th>
<th>Seq. #</th>
<th>Start T1</th>
<th>Stop T1</th>
<th>Start T2</th>
<th>Stop T2</th>
<th>Time 1</th>
<th>Time 2</th>
<th>T0 - Lead Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>8:42</td>
<td>9:10</td>
<td>-</td>
<td>-</td>
<td>0:28</td>
<td>-</td>
<td>0:28</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>8:43</td>
<td>9:11</td>
<td>-</td>
<td>-</td>
<td>0:28</td>
<td>-</td>
<td>0:28</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>8:44</td>
<td>9:12</td>
<td>-</td>
<td>-</td>
<td>0:28</td>
<td>-</td>
<td>0:28</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>8:45</td>
<td>9:14</td>
<td>-</td>
<td>-</td>
<td>0:29</td>
<td>-</td>
<td>0:29</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>8:46</td>
<td>9:15</td>
<td>-</td>
<td>-</td>
<td>0:29</td>
<td>-</td>
<td>0:29</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8:47</td>
<td>9:16</td>
<td>-</td>
<td>-</td>
<td>0:29</td>
<td>-</td>
<td>0:29</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>8:48</td>
<td>9:17</td>
<td>-</td>
<td>-</td>
<td>0:29</td>
<td>-</td>
<td>0:29</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>8:49</td>
<td>9:18</td>
<td>-</td>
<td>-</td>
<td>0:29</td>
<td>-</td>
<td>0:29</td>
</tr>
</tbody>
</table>

CONWIP = 26 Frames

Using the data from Table 1, the practical production rate can be calculated.

\[
26 \text{ units} = r_b^p \times 28 \text{ minutes}
\]

\[
r_b^p = 0.929 \text{ units/minute}
\]

With the practical production rate and bottleneck rate known, equation 2 can be used to calculate the utilization of
the production line.

\[
U = \frac{0.929}{0.983} = 94.5\%
\]

One of the advantages of using the \( r_b^p \) is the ease of accounting for the production line’s downtime in the simulation
model. By simply multiplying the \( r_b^p \) times the assembly line pitch we get the conveyor’s velocity which can
directly be entered into the simulation data module without the need for additional modules to model production line
stoppages.

The Quick Utilization Method should be used in situations where downtime data is not available for the production
line, during verification of the simulation model or when the production line being modeled has very little impact on
the simulation results.

3.2 Validation Using the Utilization Method
The validation was carried out by running 8 replications of the model and comparing the average output from the
model with the average output from 8 production days. Figure 3 shows the simulation logic for the production lines
after the individual automatic stations.
The utilization for each of the production lines under study is shown in Table 2. Notice the difference between the utilization values calculated using the Quick Utilization Method and the regular Utilization Method for the Frame 2 / Final 1 production line. The data used for the regular Utilization Method calculations came from a larger sample size that covered multiple days and shifts.

4. Simulation Study
The purpose of the experiment is to determine one of the following three possibilities that is the output from two discrete event simulation runs using Arena software.

1. Model A produces a higher throughput than Model B
2. Model A and Model B throughputs are the same
3. Model B produces a higher throughput than Model A

There are five moving conveyor assembly lines and three automatic stations. The conveyors will be modeled to run at a constant practical production rate. The system’s layout will remain constant.

4.1 Description and execution of the study procedure
The resource cycle times will be adjusted to create the two management philosophies. To determine the cycle time for the various conveyors, the model will start with the current physical limitations of the manufacturing system. The lowest cycle time for all the conveyors in the system, which is 60 seconds, is used as the cycle time for the fastest conveyor in the model. From the fastest conveyor, the rest of the conveyor cycle times are increased by increments of 1 second. The individual conveyor cycle times are converted to conveyor velocities by dividing the...
assembly line pitch by the cycle time. To account for conveyor downtime, the conveyor velocities will be multiplied times the lowest utilization rate of all the assembly lines, which is 96%.

The assembly lines will operate as continuous moving conveyors. This system characteristic means that the full and full reset buffer controls will not be required to be modeled.

4.2 Execution of the study procedure
The manufacturing system is simulated for 50 one-shift periods. Two different management philosophies are modeled with system throughput being the performance measure. In each of the models there is a model mix of two assemblies (i.e., two product types). The model buffering strategy for short state (starving) matches that of the actual production operations. Model A represents a system where the constraint is located at the beginning of the production line while Model B represents the constraint being located at the end.

The turnover station will operate at a constant cycle time of 52 seconds. The align/adjust resources will operate with random process cycle times, see Table 3. The belt elevator will operate at a constant cycle time of 27.3 seconds. Only the align/adjust work center resources will experience random failures.

<table>
<thead>
<tr>
<th>Table 3: Discrete Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part A Cumulative Probabilities</td>
</tr>
<tr>
<td>Part A Station Cycle Time (seconds)</td>
</tr>
<tr>
<td>Part B Cumulative Probabilities</td>
</tr>
<tr>
<td>Part B Station Cycle Time (seconds)</td>
</tr>
</tbody>
</table>

5. Results Summary
A t-Test: two-sample assuming unequal variance was performed on fifteen throughput values taken from each model to determine if the simulation results are statistically significant. The p-value was determined to be less than 5%, therefore the null hypothesis was rejected and the means are accepted as not being the same. The results can be found in Table 4.

<table>
<thead>
<tr>
<th>Table 4: Overall average of 50 simulated scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
</tr>
<tr>
<td>Average Throughput (units/minute)</td>
</tr>
<tr>
<td>Average Number Produced</td>
</tr>
</tbody>
</table>

The overall averages of the simulation results demonstrate that Model B produces a significantly higher throughput than Model A. In other words, locating the system constraint at the end of the manufacturing system coupled with running an “over speed” setup has benefits.

Based on our observations, having the bottleneck at the beginning may be beneficial for systems with higher buffers and decoupled segments throughout the system, because it controls WIP level and reduces chance of blocking the constraining recourse. For a matured lean manufacturing system, especially with synchronized moving assembly line, having the bottleneck at the end, i.e., the “over speed” setup, actually is more beneficial. One of the benefits is that the constraint cannot be blocked, since there is no other process downstream. Another benefit is that running with over speed forces the work in process (WIP) level to increase or event saturate in the manufacturing system. It has been well documented that increasing the system’s WIP level will also increase the system’s throughput until the critical WIP level has been achieved [12]. A third benefit is that implementing the over speed philosophy in the planning stages of design of a manufacturing system will require little to no additional resources while simultaneously improving the manufacturing system’s performance.
6. Conclusion

In this paper, the authors attempt to investigate the impact of location of constraints in a manufacturing system. Having the constraint in different locations (i.e., upstream or downstream) can affect the dynamics of the system and hence results in different performance level. The level of WIP allowed in system also plays an important role. The focus of this paper is on the impact of constraint locations on matured lean system with moving assembly line.

A simulation study has been carried out on a real manufacturing system with a series of tandem moving assembly lines and single piece flow. Fifty different scenarios were tested, and the results reveal that having the constraint at the end of the system, i.e., the “over speed” setup, can make the lean system more productive. The main reason concluded by the authors is that the over speed setup makes the lean system saturated with limited WIP and thus reduces chance of starving the constraining resource.

This research can be further extended by investigating the impact of different WIP levels allowed in the system, as well as considering other manufacturing system configuration. The considerations of Lean and TOC simultaneously can contribute to a better paradigm of manufacturing systems design.

References


