Quality and Quantity Modeling of a Production Line

Jongyoon Kim and Stanley B. Gershwin
kimjy@mit.edu gershwin@mit.edu
Massachusetts Institute of Technology

***Abstract* - During the past three decades, the success of the Toyota Production System has spurred research in the area of manufacturing systems engineering. Two research fields, productivity and quality, have been extensively studied and reported separately both in the manufacturing systems research literature and the practitioner, but there is a lack of research in the intersection of these areas. In addition to that, most studies on the relationship among manufacturing system design, quality and productivity are based on anecdotal evidence or qualitative reasoning that lack sound scientific quantitative foundations. This study tries to establish a scientific foundation to investigate how production system design and operation influence productivity and product quality by developing conceptual and computational models and performing experiments. By doing so, this study will show an important part of the way to produce high quality products with minimum cost.**

***Index Terms*— Quality, Quantity, Transfer Line**

I. INTRODUCTION

DURING the past three decades, the success of the Toyota Production System has spurred research in the area of manufacturing systems design. Numerous research papers have tried to explain the relationship between production system design and productivity, so that they can show ways to design factories to produce more products on time with less resources (people, material, space and others). On the other hand, topics in quality research have captured the attention of practitioners and researchers since the early 1980s. The recent popularity of Statistical Quality Control (SQC), Total Quality Management (TQM), and Six Sigma have emphasized the importance of quality.

These two fields, Productivity and Quality, have been extensively studied and reported separately both in the manufacturing systems research literature and the practitioner literature, but there is a lack of research in the intersection of these areas. The need for such work was recently described by authors from the GM Corporation based on their experience in industry [Inman at el., 2003]. In fact, all manufacturers must satisfy these two targets (high quality and low cost) at the same time to maintain their competitiveness in the recent tough market conditions.

Often high productivity and high quality are viewed as in conflict. Toyota Production System advocates admonish factory designers to combine inspections with operations. In the Toyota Production System, the machines are designed to detect abnormalities and to stop automatically whenever they occur. Also, operators are equipped with means of stopping the production flow whenever they note anything suspicious (they call this practice *Jidoka*). They argue that mechanical and human *Jidoka* prevents the waste that would result from producing a series of defective items. Therefore *Jidoka* is the means to improve quality and increase productivity at the same time [Shingo, 1989], [Toyota Motors Corporation, 1996]. But this statement is arguable: quality failures are often those in which the quality of each part is independent of the others. This is the case when the defect takes place due to common (or chance or random) causes of variations [Ledolter and Burrill, 1999]. In this case, there is no reason to stop a machine that has made a bad part because we have no reason to believe that stopping it will reduce the number of bad parts in the future. In this case, therefore, stopping the operation does not influence the quality but it reduces the productivity. On the other hand, when quality failures are those in which once a bad part is produced, all subsequent parts will be bad until the machine is repaired (due to special or assignable or systematic cause of variations) [Ledolter and Burrill, 1999], catching bad parts and stopping the machine as soon as possible is the best way to maintain high quality and improve productivity.

Non-stock production is another popular buzzword in manufacturing systems engineering. Some lean manufacturing professionals advocate reducing inventory on the factory floor since the reduction of work-in-process (WIP) reveals the problems in the production lines [Black, 1991]. Thus, it can help improve production quality. It is true in some sense: less inventory reduces the time between making a defect and identifying the defect. But it is also true that productivity would diminish significantly without the stocks [Burman at el., 1998]. Since there is tradeoff, there must be optimal stock levels that are specific to each manufacturing environment. In fact, Toyota recently changed their view on inventory and are trying to re-adjust their inventory levels [Fujimoto, 1999].

What is missing in the discussions of factory design, quality, and productivity is a quantitative model to show how they are inter-related. Most of the arguments about this are based on anecdotal evidence or qualitative reasoning that lack a sound scientific quantitative foundation. The research described here tries to establish scientific foundation to investigate *how production system design and operation influence productivity and product quality* by developing conceptual and computational models and performing experiments. By doing so this study will show an important part of the way to produce high quality products with minimum cost.

II. Research Objective

The objective of this research is to gain in-depth understanding to investigate *how manufacturing system design and operations simultaneously influence quality and productivity*. This will be done by developing conceptual and quantitative models and extensive experiments with the models. We will develop concepts and tools to design factories to produce high quality products and satisfy quantity and delivery requirements at minimum costs.

III. Research Method and Related Work

A. Quality Models

There are two extreme kinds of quality failures based on the characteristics of variations that cause the failures. In the quality literature, these variations are called *common* (or *chance* or *random*) cause variations and *assignable* (or *special* or *unusual*) cause variations [Montgomery, 1991]. The former are those in which the quality of each part is independent of the others. Such failures occur often when an operation is sensitive to environmental perturbations like temperature and humidity change or the operation uses a new technology that is difficult to control. This is inherent in the design of the process and cannot be removed. Such failures can be represented by independent Bernoulli random variables, in which a binary random variable indicating whether or not the part is good is chosen each time a part is operated on. A good part is produced with probability p, and a bad part is produced with probability $1 - p$. The occurrence of a bad part implies nothing about the quality of future parts, so no permanent changes can have occurred in the machine. For the sake of clarity, we call this as a *Bernoulli-type quality failure*. Most of quantitative literature on inspection allocation assume this kind of quality failures [Raz, 1986], [Lee and Unnikrishnan, 1998]. In this case, if bad parts are destined to be scrapped, it is useful to catch bad parts as soon as possible because the longer before they are scrapped, the more they consume the capacity of downstream machines. However, there is no reason to stop a machine that has produced a bad part due to this kind of failure.

Figure 1 shows the types of quality failures and types of variations. The quality failures due to assignable cause variations are those in which a quality failure only happens after a change occurs in the machine. In that case, it is highly likely that once a bad part is produced, many or all subsequent parts will be bad until the machine is repaired. Here, there is much more incentive to catch defective parts and stop the machine soon. In addition to minimizing the waste of downstream capacity, this strategy minimizes the further production of defective parts. For this kind of quality failure, there is no inherent measure of yield because the fractions of parts that are good and bad depend on how soon bad parts are detected and how quickly the machine is stopped for repair. In this, we will call this a *Markovian-type quality failure*. Most quantitative studies in Statistical Quality Control are dedicated to finding efficient inspection policies (sampling interval, sample size, and others) to detect this type of quality failure. [Woodall and Montgomery, 1999]. In reality, failures are mixtures of *Bernoulli-type quality failure* and *Markovian-type quality failure*. It can be argue that the quality strategy of the Toyota Production System [Monden, 1998], in which machines are stopped as soon as a bad part is detected is implicitly based on the assumption of the Markovian-type quality failure.

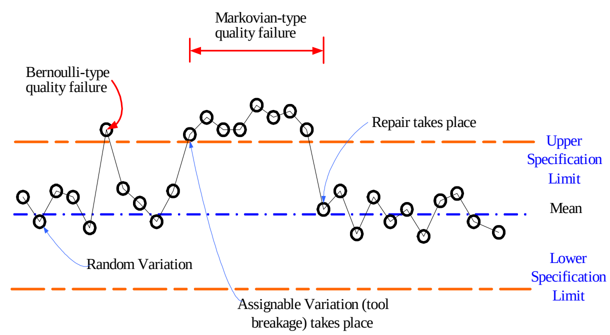

Figure 1: Types of Quality Failures

B. System Yield

System yield is defined here as the fraction of input to a system that is transformed into output of acceptable quality. This is an important metric because customers appreciate the quality of products only after all the manufacturing processes are done and the products are shipped. The system yield is a complex function of how the factory is designed and operated, as well as of the characteristics of the machines. Some of influencing factors include individual operation yields, inspection strategies, operation policies, buffer sizes, and other factors. Comprehensive approaches are needed to manage system yield effectively. This research aims to develop mathematical models to show how the system yield is influenced by these factors.

C. Response to Failures

The response to the event of a part failing inspection consists of two kinds of actions: one dealing with the part and the other dealing with the machine that produced the failing feature.

The possible actions on the parts include

- Immediate scrap
- Rework
  - Immediate
  - Deferred: The part is marked for rework later

A part should be scrapped if the rework cost would be substantial compared to the value of the part. It should be reworked otherwise. Reworking should take place immediately unless there is insufficient space on the factory floor for a rework station or for a queue of parts waiting to be reworked. This is precisely the case in an automobile assembly plant.

The possible actions on the machines include

- No action
- Immediate stop and repair

No action is appropriate if the failure is Bernoulli rather than Markovian. Since the defective part does not imply any greater likelihood of future bad parts from that machine, there is no reason to repair the machine. If the failure is Markovian, then the machine should be stopped and repaired immediately. Otherwise, it will continue to produce parts that will have to be scrapped or reworked.

D. Characterization of Machine Status

There are many possible ways to characterize the states of a machine based on the quality model discussed earlier. Figure 2 shows the proposed state transitions of a machine. In the model, the machine has three states.

- State 1: The machine is operating and producing good parts with probability of Y_1. Y_1 is typically close to 1. All the quality failures in this state are due to common cause variations.
- State –1: The machine is operating and producing good parts with probability of Y_{-1}. Y_{-1} is much less than 1. It may be 0. But the operator does not know that the machine is producing bad parts. All the quality failures in this state are due to assignable cause variations.
- State 0: The machine is not operating.

The machine also has two different failures (i.e. transition to failure states from state 1)

- Conventional failure: transition from state 1 to state 0. The machine stops producing parts due to failures like motor burnout.
- Quality failure: transition from state 1 to state –1. The machine stops producing good parts due to a failure like a severe tool wear.

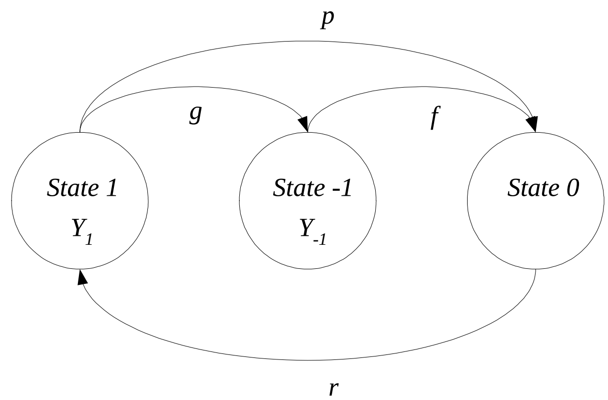

Figure 2: States of a Machine

When a machine is in state 1, it can fail due to a non-quality related event like a motor burning out and go to state 0 with transition probability rate *p*. After that an operator fixes it, so the machine goes back to state 1. Sometimes, due to an assignable cause, the machine begins to produce bad parts, so there is a transition from state 1 to state –1 with a probability rate of *g*. The machine can be stopped to be fixed by the operator when the operator knows that machine is producing bad parts. The transition from state –1 to state 0 occurs at probability rate *f*. Here for simplicity, we assume that whenever a machine is repaired, it goes back to state 1.

E. Quality Improvement Policies

System yield is a complex function of various factors such as inspection, individual operation yields, buffer size, and other operation policies. There are many ways to improve the system yield. Inspection policy is the one that has received the most attention in the literature. Research on inspection policy can be divided into optimizing inspection parameters at a single station and the inspection station allocation problem. The former issue has been investigated extensively in the Statistical Quality Control (SQC) literature [Wooddall and Montgomery, 1999]. Here, optimal SQC parameters such as sampling size and frequency are sought for an optimal balance between the inspection cost and the cost of quality. The other research looks for the optimal way to distribute inspection stations along production lines [Raz, T., 1986].

Improving individual operation yield is another important way to increase the system yield. Studies in this field try to stabilize the process either by finding root causes of variation and eliminate them or making the process insensitive to the external noise. The former topic has numerous qualitative researches in the field of Total Quality Management (TQM) [Besterfield et al., 2003] and Six Sigma [Pande and Holpp, 2002]. Quantitative research is more oriented toward the latter topic. Robust engineering

[Phadke, 1989] is an area that has gained substantial attention.

Inventory level reduction has been argued as one of the effective means to improve the system yield. Lots of lean manufacturing specialists have asserted that less inventory on the factory floor reveals the problems in the manufacturing lines more quickly and helps quality improvement activities [Alles et al., 2000], [Monden, 1998].

There also have been investigations to explain the relationship between plant layout design and quality [Cheng et al., 2000]. They argue that U-shaped lines are more advantageous than straight lines to produce higher quality products since there are more points of contact between operators. Also there is less material movement, and there are other reasons.

Although there are many possible ways to improve system yield, sticking to only one method will give marginal gains. The effectiveness of each method is greatly dependent on the details of a factory. Thus, there is need to show which method or which combination of methods is most effective in each case. The quantitative tools that will be developed from this research can help fulfill this need.

F. *Evaluation of Manufacturing Systems*

There are two different sets of tools for evaluating the performance of manufacturing systems. One is discrete event simulation and the other is analytic modeling. Simulation models have the potential for dealing with a larger class of systems than analytical methods but they are substantially slower since they need to generate large number of events to estimate performance measures of manufacturing system correctly [Law et al., 1999]. Most research in optimal buffer allocation has used simulation as an evaluation tool. Due to the low computation speed, only small fraction of the entire design space is searchable. Also, discrete-event simulations use random number generation to create events, so they give slightly different outputs for the same input. Therefore, it is difficult to determine the best directions for improvement since it is difficult to check whether the improvement in the performance is from the favorable random number streams or from the better input data. Figure 3 shows that a discrete-event simulation gives different throughput estimates for the same input each time it runs. It happens because the discrete-event simulation uses different random number streams to generate events. And from these events, it estimates the performance of a manufacturing system.

Analytic methods have different characteristics. They are substantially faster at evaluating designs and the same input always gives exactly the same output. Therefore it is convenient to search larger design space and easy to find reliable directions for design improvement [Gershwin and Schor, 2000]. But analytic models are difficult to develop, and they often require some approximation. There have been substantial volumes of research in analytic modeling of manufacturing systems [Dallery and Gershwin, 1992].

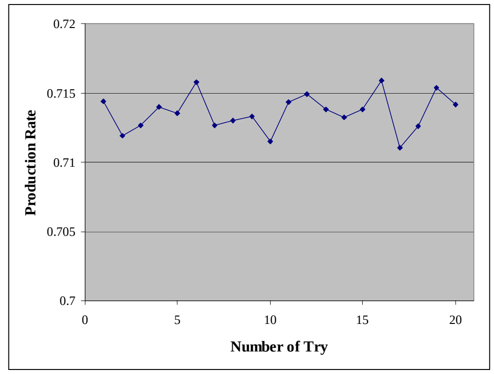

Figure 3: Variation in Production Rate Estimates from a Discrete-Event Simulation

IV. Research Progress Up to Date

A. *Continuous Model Flow Line*

A flow (or transfer) line is a manufacturing system with a very special structure. It is a linear network of service stations or machines ($M_1, M_2, ..., M_k$) separated by buffer storages ($B_1, B_2, ..., B_{k-1}$). Material flows from outside the system to M_1, then to B_1, then to M_2, and so forth until it reaches M_k after which it leaves. Figure 4 depicts a flow line. The rectangles represent machines and the circles represent buffers.

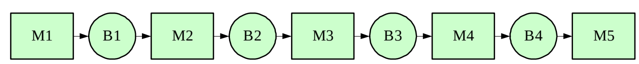

Figure 4: Five-Machine Flow Line

2-machine-1-buffer (2M1B) models should be studied first. Then a technique that divides long transfer line into multiple 2-machine-1-buffer models could be developed. Among the various modeling techniques for the 2M1B case, including deterministic, exponential, and continuous models, the continuous-material line modeling is used for this research because it can handle deterministic but different operation times at each operation. This is an extension of the continuous-material serial line modeling technique developed by Gershwin [Gershwin, 1994] by adding another machine failure state. Figure 5 shows the 2M1B continuous model where machines, buffer and discrete parts are treated as through valves, bathtub, and continuous fluid respectively.

We assume that an inexhaustible supply of workpieces is available upstream of the first machine in the line, and unlimited storage area is present downstream of the last machine. Thus, the first machine is never starved, and the last machine is never blocked. Also, failures are assumed to be operation dependent (ODF).

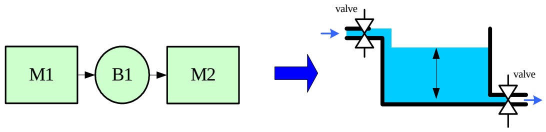

Figure 5: Two-Machine-One-Buffer Continuous Model

B. *Infinite Buffer/ Zero Buffer Case*

An infinite buffer case is a special 2M1B case in which the size of the Buffer (***B1***) is infinite. This is an extreme case where the first machine (***M1***) never suffers from blockage. Therefore it gives an upper limit of the production rate of the 2M1B model. A zero buffer case is the case in which there is no buffer between two machines. This is the other extreme case where blockage and starvation take place most frequently. Thus, it gives a lower limit of production rate of the 2M1B model. Figure 6 shows a typical shape of production rate (total or effective) curve as a function of buffer size. It increases as buffer size increases except special cases shown in section *F*. Mathematical expressions of *total production rate*, which is output rate of good and bad parts, and *effective production rate*, which is output rate of good parts, are derived for each case. Equation (1) and (2) are production rate and effective production rate of the infinite buffer case respectively.

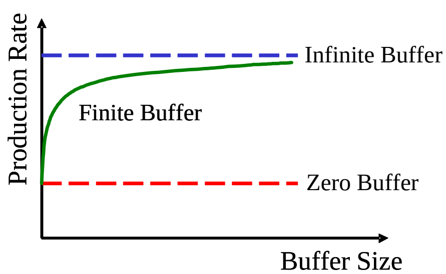

Figure 6: Infinite Buffer and Zero Buffer Case

$$PR_T = Min[\frac{\mu_1(1+g_1/f_1)}{1+(p_1+g_1)/r_1+g_1/f_1}, \frac{\mu_2(1+g_2/f_2)}{1+(p_2+g_2)/r_2+g_2/f_2}] \quad (1)$$

$$PR_E = \frac{f_1 f_2}{(f_1+g_1)(f_2+g_2)} PR_T \quad (2)$$

The accuracy of these equations has been tested by comparison with simulation result. A discrete event simulation was built in C++ for comparison purposes. Table 1 shows that the derived equations give good result.

Case #	PR(Analytic)	PR(Sim)	%Difference
1	0.9016	0.9018	-0.02%
2	0.8925	0.9018	-1.04%
3	0.9016	0.9018	-0.02%
4	0.8729	0.8750	-0.24%
5	0.8925	0.9018	-1.04%

Table 1: Infinite Buffer Case

Production rate and effective production rate of the zero buffer case are shown in Equation (3) and (4) respectively. Table 2 confirms that these equations are valid.

$$PR_T = \frac{Min[\mu_1, \mu_2]}{1+\frac{f_1(p_1+g_1)}{r_1(f_1+g_1)}+\frac{f_2(p_2+g_2)}{r_2(f_2+g_2)}} \quad (3)$$

$$PR_E = \frac{f_1 f_2}{(f_1+g_1)(f_2+g_2)} PR_T \quad (4)$$

Case #	PR(Analytic)	PR(Sim)	%Difference
1	0.8802	0.8802	0.00%
2	0.8211	0.8211	0.00%
3	0.8860	0.8860	0.00%
4	0.7748	0.7748	-0.01%
5	0.8207	0.8207	-0.01%

Table 2: Zero Buffer Case

C. *Finite Buffer Case*

1) *State Definition*

State of 2M1B case is initially defined as $(x, y, \alpha_1, \alpha_2)$ where each character represents

- *x*: total amount of material at B1
- *y*: amount of defective material at B1
- α_1: state of machine 1.
- α_2: state of machine 2.

2) *Internal State Transition Equations and Boundary Conditions*

When the buffer B1 is neither empty nor full, its level can rise or fall depending on the states of adjacent machines. Since it can change only a small amount during a short time interval, it is natural to use differential equations to describe its behavior. From the state definitions and modeling assumptions, 9 internal transition equations for the probability density function $f(x, y, \alpha_1, \alpha_2)$ are derived as follows:

$$\frac{\partial f(x,y,1,1)}{\partial t} = (\mu_2-\mu_1)\frac{\partial f(x,y,1,1)}{\partial x} + \frac{\mu_2 y}{x}\frac{\partial f(x,y,1,1)}{\partial y} - (p_1+g_1+p_2+g_2) f(x,y,1,1) + r_2 f(x,y,1,0) + r_1 f(x,y,0,1) \quad (5)$$

$$\frac{\partial f(x,y,1,0)}{\partial t} = p_2 f(x,y,1,1) - \mu_1\frac{\partial f(x,y,1,0)}{\partial x} + r_1 f(x,y,0,0) + f_2 f(x,y,1,-1) - (p_1+g_1+r_2) f(x,y,1,0) \quad (6)$$

$$\frac{\partial f(x,y,1,-1)}{\partial t} = g_2 f(x,y,1,1) + (\mu_2-\mu_1)\frac{\partial f(x,y,1,-1)}{\partial x} + \frac{\mu_2 y}{x}\frac{\partial f(x,y,1,-1)}{\partial y} - (p_1+g_1+f_2) f(x,y,1,-1) + r_1 f(x,y,0,-1) \quad (7)$$

$$\frac{\partial f(x,y,0,1)}{\partial t} = p_1 f(x,y,1,1) + \mu_2\frac{\partial f(x,y,0,1)}{\partial x} + \frac{\mu_2 y}{x}\frac{\partial f(x,y,0,1)}{\partial y} - (r_1+p_2+g_2) f(x,y,0,1) + r_2 f(x,y,0,0) + f_1 f(x,y,-1,1) \quad (8)$$

$$\frac{\partial f(x,y,0,0)}{\partial t} = p_1 f(x,y,1,0) + p_2 f(x,y,0,1) - (r_1+r_2) f(x,y,0,0) + f_2 f(x,y,0,-1) + f_1 f(x,y,-1,0) \quad (9)$$

$$\frac{\partial f(x,y,0,-1)}{\partial t} = p_1 f(x,y,1,-1) + g_2 f(x,y,0,1) + f_1 f(x,y,-1,-1) + \mu_2 \frac{\partial f(x,y,0,-1)}{\partial x} + \frac{\mu_2 y}{x}(x,y,0,-1) - (r_1+f_2) f(x,y,0,-1) \quad (10)$$

$$\frac{\partial f(x,y,-1,1)}{\partial t} = g_1 f(x,y,1,1) - (p_2+g_2+f_1) f(x,y,-1,1) + (\mu_2-\mu_1)\frac{\partial f(x,y,-1,1)}{\partial x} + (\frac{\mu_2 y}{x} - \mu_1)\frac{\partial f(x,y,-1,1)}{\partial y} + r_2 f(x,y,-1,0) \quad (11)$$

$$\frac{\partial f(x,y,-1,0)}{\partial t} = g_1 f(x,y,1,0) - \mu_1 \frac{\partial f(x,y,-1,0)}{\partial x} - \mu_1 \frac{\partial f(x,y,-1,0)}{\partial y} - (r_2+f_1) f(x,y,-1,0) + p_2 f(x,y,-1,1) + f_2 f(x,y,-1,-1) \quad (12)$$

$$\frac{\partial f(x,y,-1,-1)}{\partial t} = g_1 f(x,y,1,-1) + g_2 f(x,y,-1,1) - (f_1+f_2) f(x,y,-1,-1) + (\mu_2-\mu_1)\frac{\partial f(x,y,-1,-1)}{\partial x} + (\frac{\mu_2 y}{x} - \mu_1)\frac{\partial f(x,y,-1,-1)}{\partial y} \quad (13)$$

These transition equations are linear partial differential equations in *t*,*x*,*y* with coefficients that are nonlinear functions of *x* and *y*. Therefore this system is unlikely to be solved. But starvation and blockage of the machines occur independently of *y*. And if we know the starvation and blockage probabilities we can estimate total production rate and effective production rate of 2M1B. Therefore the state of 2M1B can be simplified as (x,α_1,α_2). After deriving all the equations again with the new state definition and setting $\partial f/\partial t = 0$, $\partial f/\partial x \to df/dx$ for the steady state condition, we have following equations:

$$0 = (\mu_2-\mu_1)\frac{df(x,1,1)}{dx} - (p_1+g_1+p_2+g_2) f(x,1,1) + r_2 f(x,1,0) + r_1 f(x,0,1) \quad (14)$$

$$0 = p_2 f(x,1,1) - \mu_1 \frac{df(x,1,0)}{dx} - (p_1+g_1+r_2) f(x,1,0) + f_2 f(x,1,-1) + r_1 f(x,0,0) \quad (15)$$

$$0 = g_2 f(x,1,1) + (\mu_2-\mu_1)\frac{df(x,1,-1)}{dx} + r_1 f(x,0,-1) - (p_1+g_1+f_2) f(x,1,-1) \quad (16)$$

$$0 = p_1 f(x,1,1) + \mu_2 \frac{df(x,0,1)}{dx} - (r_1+p_2+g_2) f(x,0,1) + r_2 f(x,0,0) + f_1 f(x,-1,1) \quad (17)$$

$$0 = p_1 f(x,1,0) + p_2 f(x,0,1) - (r_1+r_2) f(x,0,0) + f_2 f(x,0,-1) + f_1 f(x,-1,0) \quad (18)$$

$$0 = p_1 f(x,1,-1) + g_2 f(x,0,1) - (r_1+f_2) f(x,0,-1) + \mu_2 \frac{df(x,0,-1)}{dx} + f_1 f(x,-1,-1) \quad (19)$$

$$0 = g_1 f(x,1,1) - (p_2+g_2+f_1) f(x,-1,1) + (\mu_2-\mu_1)\frac{df(x,-1,1)}{dx} + r_2 f(x,-1,0) \quad (20)$$

$$0 = g_1 f(x,1,0) - \mu_1 \frac{df(x,-1,0)}{dx} - (r_2+f_1) f(x,-1,0) + p_2 f(x,-1,1) + f_2 f(x,-1,-1) \quad (21)$$

$$0 = g_1 f(x,1,-1) + g_2 f(x,-1,1) + (\mu_2-\mu_1)\frac{df(x,-1,-1)}{dx} - (f_1+f_2) f(x,-1,-1) \quad (22)$$

It is natural to assume an exponential form for the solution to the steady state density functions, since equations (14) to (22) are coupled ordinary linear differential equations. A solution of a form $f(x,\alpha_1,\alpha_2) = e^{\lambda x} G_1(\alpha_1) G_2(\alpha_2)$ is assumed. This form satisfies the transition equations if all of the following equations are met.

$$\{(\mu_2-\mu_1)\lambda - (p_1+g_1+p_2+g_2) G_1(1)G_2(1)\} + r_2 G_1(1)G_2(0) + r_1 G_1(0)G_2(1) = 0 \quad (23)$$

$$-\{\mu_1\lambda + (p_1+g_1+r_2)\} G_1(1)G_2(0) + p_2 G_1(1)G_2(1) + f_2 G_1(1)G_2(-1) + r_1 G_1(0)G_2(0) = 0 \quad (24)$$

$$\{(\mu_2-\mu_1)\lambda - (p_1+g_1+f_2)\} G_1(1)G_2(-1) + g_2 G_1(1)G_2(1) + r_1 G_1(0)G_2(-1) = 0 \quad (25)$$

$$\{\mu_2\lambda - (r_1+p_2+g_2)\} G_1(0)G_2(1) + p_1 G_1(1)G_2(1) + r_2 G_1(0)G_2(0) + f_1 G_1(-1)G_2(1) = 0 \quad (26)$$

$$p_1 G_1(1)G_2(0) + p_2 G_1(0)G_2(1) - (r_1+r_2) G_1(0)G_2(0) + f_2 G_1(0)G_2(-1) + f_1 G_1(-1)G_2(0) = 0 \quad (27)$$

$$\{\mu_2\lambda - (r_1+f_2)\} G_1(0)G_2(-1) + p_1 G_1(1)G_2(-1) + g_2 G_1(0)G_2(1) + f_1 G_1(-1)G_2(-1) = 0 \quad (28)$$

$$\{(\mu_2-\mu_1)\lambda - (p_2+g_2+f_1)\} G_1(-1)G_2(1) + g_1 G_1(1)G_2(1) + r_2 G_1(-1)G_2(0) = 0 \quad (29)$$

$$-\{\mu_1\lambda + (r_2+f_1)\} G_1(-1)G_2(0) + g_1 G_1(1)G_2(0) + p_2 G_1(-1)G_2(1) + f_2 G_1(-1)G_2(-1) = 0 \quad (30)$$

$$\{(\mu_2-\mu_1)\lambda - (f_1+f_2)\} G_1(-1)G_2(-1) + g_1 G_1(1)G_2(-1) + g_2 G_1(-1)G_2(1) = 0 \quad (31)$$

These are 9 equations with 7 unknowns ($\lambda, G_1(1), G_1(0), G_1(-1), G_2(1), G_2(0), \text{ and } G_2(-1)$). Thus, there are 7 independent equations and 2 dependent ones. But it is not easy to tell this from these equations.

If we divide equations (23) – (31) by $G_1(0)G_2(0)$ and define new parameters

$$\Gamma_i = p_i G_i(1)/G_i(0) - r_i + f_i G_i(-1)/G_i(0) = p_i Y_i - r_i + f_i Z_i \quad (32)$$

$$\Psi_i = -p_i - g_i + r_i G_i(0)/G_i(1) \quad (33)$$

$$\Theta_i = -f_i + g_i G_i(1)/G_i(-1) \quad (34)$$

then equations (23) – (31) can be rewritten as

$$\Gamma_1 + \Gamma_2 = 0 \quad (35)$$

$$-\mu_2 \lambda = \Gamma_1 + \Psi_2 \quad (36)$$

$$\mu_1 \lambda = \Gamma_2 + \Psi_1 \quad (37)$$

$$(\mu_1 - \mu_2)\lambda = \Psi_1 + \Psi_2 \quad (38)$$

$$(\mu_1 - \mu_2)\lambda = \Theta_1 + \Theta_2 \quad (39)$$

$$\mu_1 \lambda = \Gamma_2 + \Theta_1 \quad (40)$$

$$-\mu_2 \lambda = \Gamma_1 + \Theta_2 \quad (41)$$

$$(\mu_1 - \mu_2)\lambda = \Psi_2 + \Theta_1 \quad (42)$$

$$(\mu_1 - \mu_2)\lambda = \Psi_1 + \Theta_2 \quad (43)$$

After many mathematical manipulations the equations (35) – (43) become

$$-r_1 + \frac{\{(M + r_1)(\mu_1 N - 1) - f_1\}^2}{(f_1 - p_1)(\mu_1 N - 1)} - \frac{\{(p_1 + g_1 - f_1) + r_1(\mu_1 N - 1)\}\{(M + r_1)(\mu_1 N - 1) - f_1\}}{(f_1 - p_1)(\mu_1 N - 1)} = 0 \quad (44)$$

$$-r_2 + \frac{\{(-M + r_2)(\mu_2 N - 1) - f_2\}^2}{(f_2 - p_2)(\mu_2 N - 1)} - \frac{\{(p_2 + g_2 - f_2) + r_2(\mu_2 N - 1)\}\{(-M + r_2)(\mu_2 N - 1) - f_2\}}{(f_2 - p_2)(\mu_2 N - 1)} = 0 \quad (45)$$

where

$$p_1 Y_1 - r_1 + f_1 Z_1 = -(p_2 Y_2 - r_2 + f_2 Z_2) = M$$

$$\frac{1}{\mu_1}\left(1 + \frac{1}{Y_1 + Z_1}\right) = \frac{1}{\mu_2}\left(1 + \frac{1}{Y_2 + Z_2}\right) = N$$

Now all the equations and unknowns are simplified into 2 unknowns and 2 equations. By solving the equations (44) and (45) we can calculate the probability density functions. These equations are plotted in Figure 7. Equation (44) is represented as red lines and equation (45) is shown as blue lines. Intersections of the red lines and the blue lines are the solutions of the equations. Analytical solutions to these equations have been sought but were not found since these are high order equations where there no general solution exists.

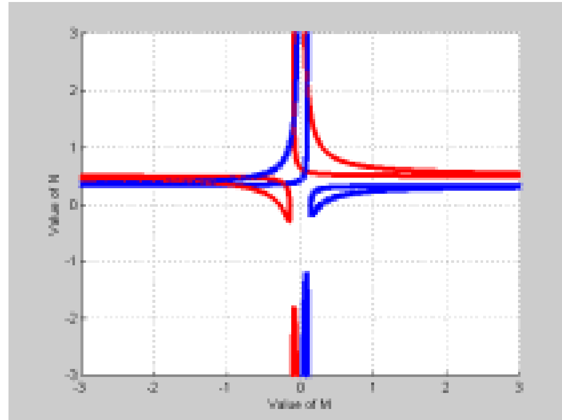

Figure 7: Plot of Equation (44) and (45)

A numerical solution approach is used to find roots of the equations. A special algorithm to find the solutions has been developed based on the characterizations of the curves. The number of roots depends on machine parameters. But it is found that there are only 3 roots when $\mu_1 = \mu_2$ regardless of other parameters. Therefore, a general expression of the probability density function is

$$f(x, \alpha_1, \alpha_2) = c_1 f_1(x, \alpha_1, \alpha_2) + c_2 f_2(x, \alpha_1, \alpha_2) + c_3 f_3(x, \alpha_1, \alpha_2) \quad (46)$$

where $f_1(x, \alpha_1, \alpha_2)$, $f_2(x, \alpha_1, \alpha_2)$, $f_3(x, \alpha_1, \alpha_2)$ are the roots of the equations (44) and (45).

Unknowns including c_1, c_2, c_3 and probability masses at the boundary can be calculated by solving boundary equations. For the $\mu_1 = \mu_2$ case, 22 boundary equations are derived:

$$\mu_1 f(0,1,0) = r_1 P(0,0,0) + p_2^b P(0,1,1) + f_2^b P(0,1,-1) \quad (47)$$

$$\mu_1 f(0,-1,0) = p_2^b P(0,-1,1) + f_2^b P(0,-1,-1) \quad (48)$$

$$\mu_2 f(N,0,1) = r_2 P(N,0,0) + p_1^b P(N,1,1) + f_1^b P(N,-1,1) \quad (49)$$

$$\mu_2 f(N,0,-1) = p_1^b P(N,1,-1) + g_2 P(N,0,1) + f_1^b P(N,-1,-1) \quad (50)$$

$$-(p_1 + g_1 + p_2^b + g_2^b) P(0,1,1) + r_1 P(0,0,1) = 0 \quad (51)$$

$$P(0,1,0) = 0 \quad (52)$$

$$g_2^b P(0,1,1) - (p_1 + g_1 + f_2^b) P(0,1,-1) + r_1 P(0,0,-1) = 0 \quad (53)$$

$$p_1 P(0,1,1) - r_1 P(0,0,1) + \mu_2 f(0,0,1) + f_1 P(0,-1,1) + r_2 P(0,0,0) = 0 \quad (54)$$

$$-(r_1 + r_2) P(0,0,0) = 0 \quad (55)$$

$$p_1 P(0,1,-1) - r_1 P(0,0,-1) + \mu_2 f(0,0,-1) + f_1 P(0,-1,-1) = 0 \quad (56)$$

$$g_1 P(0,1,1) - (f_1 + p_2^b + g_2^b) P(0,-1,1) = 0 \quad (57)$$

$$P(0,-1,0) = 0 \quad (58)$$

$$g_1 P(0,1,-1) + g_2^b P(0,-1,1) - (f_1 + f_2^b) P(0,-1,-1) = 0 \quad (59)$$

$$-(p_1^b + g_1^b + p_2 + g_2) P(N,1,1) + r_2 P(N,1,0) = 0 \quad (60)$$

$$p_2 P(N,1,1) - r_2 P(N,1,0) + \mu_1 f(N,1,0) + f_2 P(N,1,-1) + r_1 P(N,0,0) = 0 \quad (61)$$

$$g_2 P(N,1,1) - (p_1^b + g_1^b + f_2) P(N,1,-1) = 0 \quad (62)$$

$$-(r_1 + r_2) P(N,0,0) = 0 \quad (63)$$

$$g_1^b P(N,1,1) - (f_1^b + g_2 + p_2)P(N,-1,1) + r_2 P(N,-1,0) = 0 \quad (64)$$

$$-r_2 P(N,-1,0) + \mu_1 f(N,-1,0) + f_2 P(N,-1,-1) + p_2 P(N,-1,1) = 0 \quad (65)$$

$$g_1^b P(N,1,-1) + g_2 P(N,-1,1) - (f_1^b + f_2)P(N,-1,-1) = 0 \quad (66)$$

In addition to these, all the probability density functions and probability masses must satisfy the normalization equation.

D. *Performance Measures*

After finding all probability density functions and probability masses, we can calculate the average inventory in the buffer from

$$\overline{X} = \sum_{\alpha_1=-1}^{1} \sum_{\alpha_2=-1}^{1} [\int_0^N x f(x,\alpha_1,\alpha_2)dx + NP(N,\alpha_1,\alpha_2)] \quad (67)$$

The total production rate and effective production rate are calculated from equations (68) and (69) respectively:

$$PR_T = \sum_{\alpha_2=-1,1} \sum_{\alpha_1=-1,0,1} [\mu_2\{\int_0^N f(x,\alpha_1,\alpha_2)dx + P(N,\alpha_1,\alpha_2)\} + \mu_1 P(0,\alpha_1,\alpha_2)] \quad (68)$$

$$PR_E = \sum_{\alpha_1=-1,0,1} [\mu_2\{\int_0^N f(x,\alpha_1,1)dx + P(N,\alpha_1,1)\} + \mu_1 P(0,\alpha_1,1)] \quad (69)$$

The validity of the 2M1B continuous model when $\mu_1 = \mu_2$ has been checked through comparison with simulation. As Table 3 demonstrates, the results from the analytic model show good agreement with results from simulation.

Case #	PR(Anal)	PR(Sim)	Error	Inv(Anal)	Inv(Sim)	Error
1	0.806	0.808	-0.25%	2.500	2.619	-4.53%
2	0.855	0.858	-0.37%	25.000	24.883	0.47%
3	0.936	0.938	-0.23%	4.709	4.989	-5.60%
4	0.944	0.946	-0.22%	12.654	12.757	-0.81%
5	0.909	0.911	-0.19%	2.781	2.832	-1.81%
6	0.922	0.924	-0.24%	9.213	9.318	-1.13%
7	0.909	0.910	-0.07%	2.220	2.321	-4.39%
8	0.925	0.926	-0.18%	7.242	7.080	2.30%
9	0.840	0.843	-0.38%	20.020	20.149	-0.64%
10	0.763	0.767	-0.49%	4.983	5.110	-2.48%

Table 3: Simulation and Analytic Result Comparison

E. *Quality Information Feedback*

Factory designers and managers know that it is ideal to have inspection at every operation. However, it is also costly to inspect after each operation. As a result, factories are often designed so that multiple inspections are performed at a small number of stations. In this case, inspection at downstream operation can give feedback to upstream machines. (We call this *quality information feedback*). In this case, the yield of a line is a function of the sizes of buffers. This is because when buffers get larger, more material accumulates between an operation and the inspection of that operation. All such material will be defective if the Markovian type quality failure takes place. In other words, if buffers are larger, there tends to be more material in the buffers and consequently more material is defective. Therefore it takes longer time to have inspections after finishing operations. We can capture this phenomenon by modifying *f*.

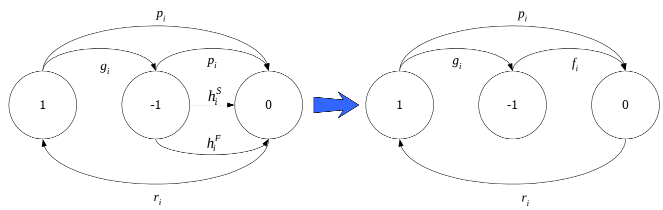

Figure 8: States of One Machine

In Figure 8, we show that there are three ways to have transition from state –1 to state 0.

- The machine fails by itself from conventional failure: p_i
- The machine is stopped by its own inspection: h_i^S
- The machine is stopped by inspections at downstream machines: h_i^F

Here, the machine can be stopped by inspections only when it does not fail by itself from conventional failures. Also, the machine can be stopped by inspection at downstream machine only if its own inspection does not identify the defect. Therefore

$$\begin{aligned} f_i\delta t &= f(p_i, h_i^S h_i^F)\delta t = p_i\delta t + (1-\delta t)\{h_i^S\delta t + (1-h_i^S\delta)h_i^F\delta t\} \\ f_i &= f(p_i, h_i^S h_i^F) = p_i + (1-\delta t)\{h_i^S + (1-h_i^S\delta)h_i^F\} \\ &Since\ \delta t \to 0, \quad f_i = p_i + h_i^S + h_i^F \end{aligned} \quad (70)$$

In the 2M1B model, p_1 is a characteristic of Machine 1 (M1). h_1^S is a function of inspection parameters such as sampling interval and sample size. $1/h_1^F$ is the mean time between making defects at M1 and identifying them at Machine 2 (M2) which is dependent on amount of inventory at Buffer (B1). Once M1 begins to produce bad parts, all subsequent parts are bad until the inspection at M 2 stops M1. When M2 stops M1, all parts in B1 and some of parts that M2 has processed are defective. Since $1/h_2^S$ is the mean time to identify defects at M2, $1/h_1^F$ is a sum of the *average time that material stays in B1* and $1/h_2^S$. *The average time that material stays in B1* can be estimated from the production rate of 2M1B and the average inventory. Since the average inventory is a function of *f* and *f* is also dependent on the average inventory, an iterative method is used to get these values.

F. *Insights from the Numerical Experimentation*

1) *Influence of Quality Feedback*

Having quality feedback means having more inspection. Therefore, machines tend to stop more frequently with quality feedback. As a result, the production rate of the line decreases but the effective production rate increases since added inspections prevent making defective parts. This phenomenon is shown in Figure 9. This can be interpreted as suggesting that stopping the line immediately when defective parts are found is a good way to improve quality and productivity at the same time if the quality failure is Markovian type. Note that in this case, both total production rate and effective production rate increase with buffer size, with or without quality feedback.

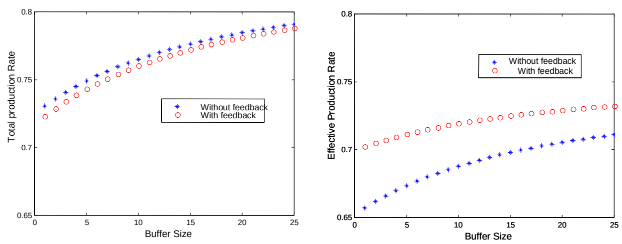

Figure 9: Production Rates with/without Quality Feedback

As explained in previous section, system yield is a function of buffer sizes if there is quality feedback. Figure 10 shows system yield decreasing as buffer size increases when there is quality feedback. This happens because when the buffer gets larger, more material accumulates between an operation and the inspection of that operation. All such material will be defective when the first machine is at state -1 but the inspection at the first machine does not find it. If there is no quality feedback, then system yield is independent of the buffer sizes. From this we are able to demonstrate that "*smaller stock improves quality*" which is widely believed, under the condition that quality failures are Markovian and quality feedback exists.

Typically, increasing the buffer size leads to higher effective production rate. This is shown in Figure 9. But under certain conditions, the effective production rate can actually decrease as buffer size increases. This can happen when

- The first machine produces bad parts frequently.
- The inspection at the first machine is poor or non-existent and inspection at the second machine is reliable.
- There is a quality feedback.
- The isolated production rate of the first machine is higher than that of the second machine.

Figure 10 shows a case in which a buffer size increase leads lower effective production rate due to rapid reduction of system yield.

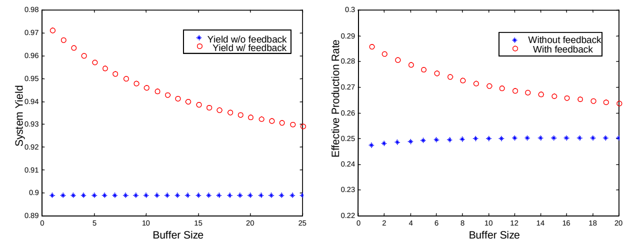

Figure 10: System Yield and Effective Production Rate

2) *How to Improve Quality*

There are two major ways to improve quality. One is to increase the yield of individual operations and the other is to perform rigorous inspection. Having extensive preventive maintenance on manufacturing equipment and using robust engineering techniques to stabilize operations have been suggested as tools to increase yield of individual operations. Both approaches increase the mean time to defect (MTTD) (i.e. decrease g of an operation). On the other hand, the inspection policy aims to detect bad parts as soon as possible and prevent their flow toward downstream operations. More rigorous inspection decreases the mean time to identify (MTTI) (i.e. increase f of an operation). It is natural to believe that using only on one kind of method to achieve a target quality level would not give the most cost efficient quality assurance policy. Figure 11 indicates that impact of individual operation stabilization on quality improvement decreases as the operation becomes more stable. It also shows that effect of improving inspection (MTTI) on quality decreases. Therefore, it is optimal to use a combination of both methods to reach to a target quality level.

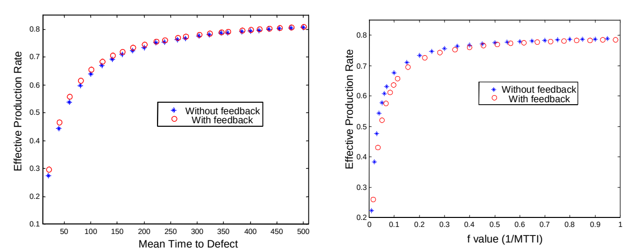

Figure 11: System Yield and Effective Production Rate

3) *How to Increase Effective Production Rate*

It is known that improving the stand-alone throughput of each operation and increasing buffer space are typical ways to increase production rate of manufacturing systems. If operations are apt to have quality failures, there may be other ways to increase the effective production rate of manufacturing systems: increasing the yield of each operation and conducting more extensive inspections. Stabilizing operations, thus improving the yield of individual operations, will increase effective throughput of a manufacturing system regardless of the type of quality failure. On the other hand, reducing the mean time to identify (MTTI) will increase effective production rate only if the quality failure is Markovian but it will decrease effective production rate if the quality failure is Bernoulli. In some situations, increasing inspection reliability is more effective than increasing buffer size to boost effective production rate. Figure 12 shows this. Also, in some other situations, increasing machine stability is more effective than increasing buffer size to enhance effective production rate. Figure 13 shows this phenomenon.

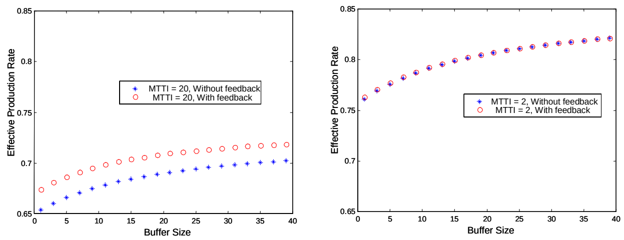

Figure 12: Mean Time to Identify and Effective Production Rate

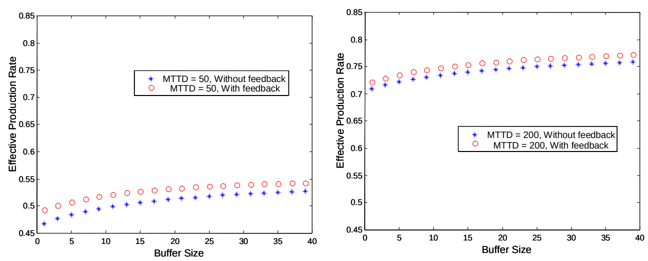

Figure 13: Quality Failure Frequency and Effective Production Rate

V. Future Research

The development of the 2 Machine 1 Buffer (2M1B) model is still under the progress. Solving boundary conditions with different machine speeds is expected to be done soon. This case is challenging since number of the roots of the internal transition equations depends on input parameters. As a result, more boundary conditions are to be met. After finishing the 2M1B model construction, decomposition techniques for longer line analysis with and without quality feedback will be developed. Rigorous validation process will follow. Extensive numerical experiments will be conducted to observe the behavior of flow lines with quality problems. Based on the characterization of the system behavior, recommendations for designing manufacturing systems to produce high quality products with minimum cost will be proposed.

References

[1] Alles, M., Amershi, A., Datar, S., and Sarkar, R., "Information and incentive effects of inventory in JIT production," *Management Science*, Vol. 46 No. 12 pp.1528 – 1544, December 2000.

[2] Besterfield, D. H., Besterfield-Michna, C., Besterfield, G., and Besterfield-Sacre, M., *Total quality management*, Prentice Hall, 2003.

[3] Black, J. T., *The design of the factory with a future*, McGraw-Hill, 1991

[4] Burman, M., Gershwin, S. B., and Suyematsu, C., "Hewlett-Packard uses operations research to improve the design of a printer production line," *Interfaces*, Vol. 28, No. 1, pp. 24 – 26, January – February, 1998.

[5] Cheng, C. H. Miltenburg, J., and Motwani, J., "The effect of straight and U shaped lines on quality", *IEEE Transactions on Engineering Management*, Vol. 47, No. 3, pp 321-334, August 2000

[6] Dallery, Y. and Gershwin, S. B., "Manufacturing flow line systems: a review of models and analytical results. Special issue on queuing models of manufacturing systems," *Queuing Systems Theory and Applications*, Vol 12, pp. 3 – 94, 1992

[7] Fujimoto, T., *The evolution of a manufacturing systems at Toyota*, Oxford University Press, 1999

[8] Gershwin, S. B., *Manufacturing systems engineering*, Prentice Hall, 1994

[9] Gershwin, S. B. and Schor, J. E. "Efficient algorithms for buffer space allocation," *Annals of Operations Research*, Volume 93, pp 117-144, 2000.

[10] Inman, R. R. Blumenfeld, D. E., Huang, N., and Li, J., (2003) "Designing production systems for quality: research opportunities from an automotive industry perspective", *International Journal of Production Research*, Vol. 41, No. 9, pp. 1953 – 1971.

[11] Law, A. M., Kelton, D. W., Kelton, W. D., and Kelton, D. M, Simulation modeling and analysis, McGraw-Hill, 1999

[12] Ledolter, J. and Burrill, C. W., (1999), Statistical quality control, John & Wiley

[13] Monden, *Toyota production system* – An integrated approach to Just-In-Time, EMP Books, 1998

[14] Montgomery, D. C., *Introduction to statistical quality control*, John Wiley & Sons, Inc, 1991

[15] Pande, P. and Holpp, L., What is six sigma? , McGraw-Hill, 2002

[16] Phadke, M., *Quality engineering using robust design*, Prentice Hall, 1989

[17] Raz, T., "A survey of models for allocating inspection effort in multistage production systems," *Journal of Quality Technology*, Vol. 18, No. 4, pp. 239 – 246, October 1986

[18] Shin, W. S., Mart, S. M., and Lee, H. F., "Strategic allocation of inspection stations for a flow assembly line: a hybrid procedure," *IIE Transactions*, Vol 27, pp. 707 – 715, 1995.

[19] Shingo, S., *A study of the Toyota production system from an industrial engineering viewpoint*, Productivity Press, 1989

[20] Toyota Motor Corporation, *The Toyota production system*, 1996

[21] Wooddall, W. H. and Montgomery, D. C., "Research issues and ideas in statistical process control", *Journal of Quality Technology*, Vol. 31, No. 4, pp 376-386, 1999