Quality/Quantity Modeling and Analysis of Production Lines Subject to Uncertainty

Phase I, Final Report

Presented to General Motors R & D Center

Irvin C. Schick, Stanley B. Gershwin, and Jongyoon Kim
Laboratory for Manufacturing and Productivity
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307

May 10, 2005

Abstract

During the past three decades, the success of the Toyota Production System (TPS) and other production system design methodologies have spurred much research in manufacturing systems engineering. However, although productivity and quality have each been extensively studied, there has been little research on how they interact. This report describes work done under GM funding to analyze how production system design, quality, and productivity are inter-related in production networks, and to develop analytical as well as numerical methods that make it possible to evaluate, compare, and optimize the performance of competing designs. More specifically, it summarizes simulated, analytical, and numerical results concerning the effect of separating inspection from operation, as a function of machine, buffer, and inspection station parameters; control policy; production system topology, and other factors.

In this report, a taxonomy is presented for quality failures, and a class of stochastic models is described to represent a realistic subset of those failures. Quality control mechanisms are formulated, including scrapping of defective parts as well as an information feedback scheme that results in taking offending machines down for maintenance. Pertinent performance measures are defined, including mean total production rate, good (i.e. not defective) production rate, yield, in-process inventory, and lead time.

Analytical results based upon a continuous-material approximation of a two-stage line are presented, and these results are validated by comparison with discrete-event simulation. It is shown, in particular, that production rate is not always a monotonically increasing function of buffer capacity when quality inspection and information feedback are present; that taking a machine down for maintenance as soon as a single quality defect is detected (*jidoka*) is not always the optimal policy, and that system parameters must be taken into account in determining the best course of action following the detection of one or more defects.

Further simulation results include the investigation of the effect of the number and placement of inspection stations, of machine and buffer parameters, of inspection policies, of control policies (scrapping and information feedback), and of production line topology on selected performance measures. In particular, it is shown that performance can be quite sensitive to the placement of inspection stations, so that a few well-placed inspection stations will result in better performance than many poorly-placed ones.

Contents

1 Introduction 4

2 Modeling and Taxonomy 5

2.1 Production Systems . . . 5

2.1.1 Machines . . . 6

2.1.2 Buffers . . . 11

2.1.3 Topology . . . 12

2.2 Quality Failures . . . 13

2.2.1 Internal vs. External Failures . . . 14

2.2.2 Dynamic Characteristics . . . 14

2.2.3 Correlation and Stacking . . . 16

2.3 Inspection . . . 18

2.3.1 Destructive vs. Non-Destructive Testing . . . 18

2.3.2 Domain and Frequency . . . 19

2.3.3 Accuracy and Declaration of a Defective Part . . . 19

2.3.4 Declaration of an Impaired Machine . . . 20

2.4 Control Policies . . . 20

2.4.1 Action on Defective Parts . . . 21

2.4.2 Action on Impaired Machines . . . 22

2.4.3 Action on People . . . 22

2.5 Measures of System Performance . . . 23

2.5.1 Quantity and Quality . . . 23

2.5.2 Inventory . . . 24

2.5.3 Production Lead Time . . . 25

3 Analytical Results 25

3.1 Mathematical Models . . . 25

3.1.1 Single Machine Model . . . 25

3.1.2 The 2-Machine-1-Buffer Continuous Model . . . 28

3.1.3 The Infinite Buffer Case . . . 28

3.1.4 The Zero Buffer Case . . . 29

3.1.5 The 2-Machine-1-Finite-Buffer Line . . . 30

3.1.6 Quality Information Feedback . . . 30

3.1.7 Insights From Numerical Experimentation . . . 31

3.2 Effectiveness of the Jidoka Stopping Policy . . . 35

3.2.1 Motivation . . . 35

3.2.2 Single Machine . . . 37

3.2.3 Analysis of a Two-Machine System . . . 39

3.2.4 Effectiveness of Jidoka . . . 40

3.3 Modeling and Analysis of Long Flow Lines with Quality and Operational Failures . . . 43

3.3.1 Motivation and Literature Review . . . 43

3.3.2 Ubiquitous Inspection . . . 44

3.3.3 Remote Quality Information Feedback . . . 47

3.3.4 Multiple Quality Information Feedback 48

4 Simulation Studies 50

4.1 Location and Number of Inspection Stations 50
4.2 Control Policies 52
4.2.1 Scrapping 52
4.2.2 Quality Information Feedback 55
4.3 Production System Topology 59

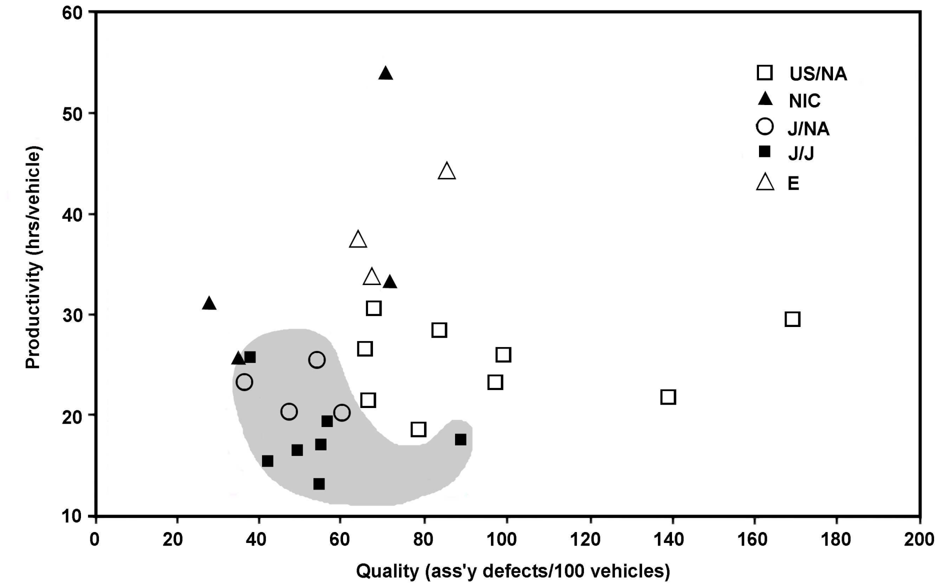
5 Conclusion 62

5.1 Summary 62
5.2 Current Research 63
5.2.1 Exploration of Phenomena 63
5.2.2 Validation of model assumptions 63
5.2.3 GM factory case study 64
5.3 Further Research 64
5.3.1 Non-Instantaneous Inspection 64
5.3.2 Buffer Delays and Failures 64
5.3.3 Analytical Models Of Scrapping 64
5.3.4 Non-FIFO Buffer Discipline 65
5.3.5 Rework 65
5.3.6 Analytic Modeling of More Elaborate Failure Dynamics 65
5.3.7 Preventive Maintenance 66
5.3.8 More General Phenomena, Including Loops, Batches, and Setups 66
5.3.9 Integrated Models Of Material Flow Control And Inspection 66
5.3.10 Decomposition Without Reducing Machine Models to Two States 66
5.3.11 Two-Machine, One-Buffer Lines with Good and Bad Parts Modeled Distinctly in the Buffer 66
5.3.12 More General Inspection Strategies 67

1 Introduction

Manufacturers strive to satisfy two requirements while minimizing cost. These are *quality* (every important characteristic of every manufactured item must satisfy certain specifications) and *quantity* (a specified number of items must be produced within a specified time interval). The task of satisfying these requirements is complicated by the presence of uncertainty (changing machine characteristics, unknown machine state, imprecise observations of machine output, etc.) as well as changing conditions (variable demand, lead time, inventory constraints, product mix, etc.).

Production system designers must make quality-motivated decisions such as choosing operation parameters (for example, speeds and feed rates in metal-cutting), the locations of inspection stations, and the actions to take in response to failed inspections. They must also make quantity-motivated decisions such as the sizes of buffers and the structures of production systems. Precise mathematical models have not been available for all relevant performance measures, but even when some have been formulated, quality-motivated design choices have historically been evaluated with quality-focused models, and quantity-motivated design choices with quantity-focused models. Precise models to *simultaneously* predict the interaction of quantity- and quality-related design choices on quantity and quality performance are needed, but do not exist.



***Source*: IMVP World Assembly Plant Survey, 1989**

Figure 1: Productivity vs. Quality in the Assembly Plant, Volume Producers, 1989. The abscissa expresses quality in terms of number of defects per 100 vehicles, so that left corresponds to higher quality; likewise the ordinate expresses quantity in terms of hours spent per vehicle, so down corresponds to higher productivity

Considerable empirical and anecdotal evidence has shown that quality and quantity performance need not be conflicting objectives. The historical data in Figure 1 show that a number of automotive companies have succeeded in achieving both high quality and high productivity (Womack, Jones, and

Roos 1990). Many lessons and design rules have been extracted from such studies, but predictive models are lacking.

During the past three decades, the success of the Toyota Production System (TPS) (Monden 1983) has spurred much research in manufacturing systems design. Numerous research papers have explored the relationship between production system design and productivity, so as to formulate ways of designing factories that manufacture more products, on time, and more economically (in terms of labor, material, and space). At the same time, topics in quality research have captured the attention of practitioners and academics alike since the early 1980s. The recent popularity of Statistical Quality Control (SQC) (Woodall and Montgomery 1999), Total Quality Management (TQM) (Besterfield, Besterfield-Michna, Besterfield, and Besterfield-Sacre 2003), and Six Sigma (Pande and Holpp 2002) demonstrate the importance given to quality by the manufacturing community.

The two issues discussed here, productivity and quality, have been extensively studied and reported separately, both in the manufacturing systems research literature and in the practitioner literature. However, there has been little research on their *relationship*. The need for such work was recently described by General Motors researchers, based upon their practical experience (Inman, Blumenfeld, Huang, and Li 2003). As this testimony demonstrates, it is necessary to satisfy both criteria simultaneously for any manufacturer to remain competitive.

The work described here brings to bear the tools of stochastic systems, operations research, and statistics on the problem of designing manufacturing systems that simultaneously meet high standards of productivity and quality. Markov models are used to represent unreliable machines, Bayesian decision theory provides the framework for formulating optimal control policies, and both analytical and numerical methods are employed to develop well-defined, quantitative tools for evaluating and comparing candidate designs, and choosing optimal policies and system parameters.

2 Modeling and Taxonomy

This section describes the components of a production system, the characteristics of quality failures, the process of quality inspection, and the actions taken as a result of inspection.

2.1 Production Systems

Consider a production system composed of a sequence of *machines*, possibly separated by interstage *buffers*, and possibly followed by *inspection stations* (Figure 2).

Each machine may be *operational* or *down*. When it is operational, the machine is capable of producing a part, though external conditions may prevent it from doing so. If it does produce a part, then it removes an unprocessed part from the buffer immediately upstream and, having processed it, moves it to the buffer immediately downstream. The part it produces may be defective, but this is not known until the part is inspected. When the machine is down, it is not capable of producing a part.

A machine that is operational needs an unprocessed part to work on. If such a part is temporarily unavailable (i.e. if the buffer immediately upstream is empty), the machine is said to be *starved*. Similarly, a machine that is operational needs space to discharge the part it has processed. If such space is temporarily unavailable (i.e. if the buffer immediately downstream is full), the machine is said to be *blocked*. A machine that is operational only produces a part when it is neither starved nor blocked.

Interstage buffers are passive devices that accumulate parts produced by upstream machines when downstream machines are down, and supply downstream machines with parts when upstream machines

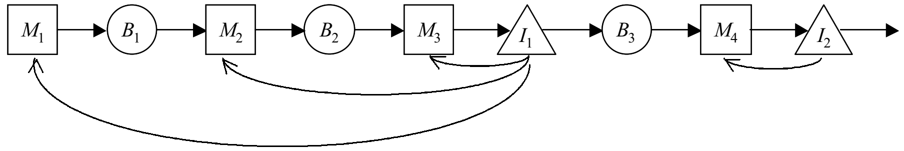


Figure 2: Example of a production line: M_i are machines, B_i are interstage buffers, and I_i are inspection stations. The thin lines indicate material flow; the bold lines indicate which inspection station inspects the work of which machine.

are down. Buffers act to *partially decouple* sequences of machines, and thus increase the total production rate. However, they accumulate in-process inventory, which may be undesirable in certain circumstances.

Finally, inspection stations test parts produced by a given set of upstream machines to *detect defects*; however, the test may be unreliable and thus may result in incorrect classification.

2.1.1 Machines

Each machine in a production system is modeled as a memoryless (Markovian) subsystem. It switches among operational and down states. When in an operational state, the machine produces parts if it is neither starved, nor blocked, though the parts it produces may be defective. When in a down state, the machine does not produce parts. During any given cycle, a machine can only switch from an operational state to a down state if it is working on a part.

In isolation, a machine is relatively easy to model and analyze. When embedded in a production line, however, it interacts with other components of the production line and such interactions make it difficult to formulate mathematical models that are both exact and of reasonably low dimensionality (Gershwin 1987, Gershwin 1994). Consequently analytical approximations as well as simulation techniques are utilized to determine the performance of production lines composed of multiple machines, interstage buffers, and inspection stations.

A machine may have multiple operational and down states. These states may correspond to physical aspects of a machine, or they may be abstractions (*auxiliary* states) that make it possible to model the behavior of machines with Markov chains. To illustrate these two options, we now describe some machine models.

A two-state machine model A very simple machine model is illustrated in Figure 3. Here, there are two states. The operation or cycle time is fixed. When in state U, the machine produces parts. An operational failure (e.g. a blown fuse or a burnt motor) takes the machine to state D. When in state D, the machine does not produce parts. Given that the machine is in state U, the probability p that it switches to state D during any given cycle is the reciprocal of the *Mean Time between Failures* (MTBF). Given that the machine is in state D, the probability r that it switches to state U during any given cycle is the reciprocal of the *Mean Time to Repair* (MTTR).

The dotted line that goes from U to D in Figure 3 represents a command to take the machine down for quality maintenance. As we shall see in Section 2.4.2, this is known as *information feedback* and corresponds to corrective action on a machine thought to be responsible for quality defects. Information

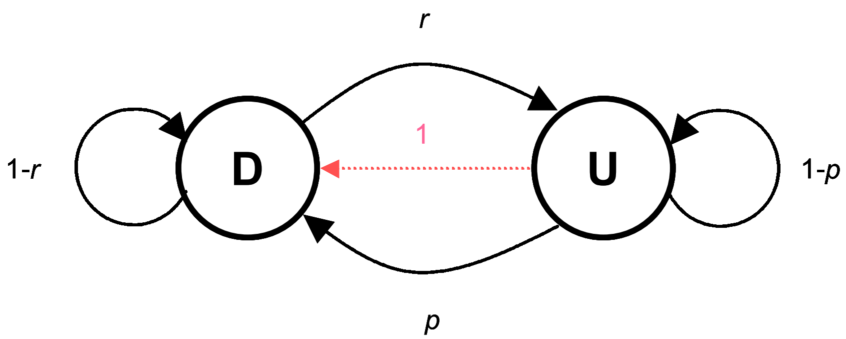


Figure 3: Example of a Markov model for a machine with a single operational state and a single down state. U denotes the operational state, and D denotes the down state.

feedback is a control signal originating *outside* the machine proper. It preempts the nominal behavior of the machine. That is, the machine switches from state U to state D with probability 1.

If the machine is in state U at the end of a cycle, it produces a part. It is possible to superimpose a *quality process* on top of this model. For example, given that the machine is in state U at the end of a cycle, the part may be good with a certain probability, and defective otherwise. In other words, given that a part is produced, its quality is determined by a coin toss.

A single operational state means that the rate at which the machine is producing defective parts is constant. This is not a particularly interesting case, since the purpose of this effort is in part to determine when the machine is in need of maintenance. Moreover, a single down state means that operational and quality repairs are all performed at once. This is unlikely to be a realistic assumption for most types of machines. Thus, more complex models are needed.

A three-state machine model Figure 4 represents a three-state machine, in which two states are operational and one is down. The difference between the two operational states is the rate at which defective parts are produced in each. H denotes a *high-yield state*, i.e. a state in which the machine produces parts most of which are good. L denotes a *low-yield state*, i.e. a state in which the machine produces defective parts at a higher rate.

As before, the dotted lines that go from H and L to D in Figure 4 represent information feedback, i.e. a command to take the machine down for quality maintenance. Information feedback preempts the nominal behavior of the machine. That is, the machine switches from states H or L to state D with probability 1.

A transition from H to L may be due, for example, to the aging or breakage of a tool. For this reason, there is typically no direct transition from L back to H; for the machine to return to a high-yield state from a low-yield state, it needs to be taken down for maintenance.

Having a single down state means that operational and quality repairs are all performed at once when the machine is in state D, whether the machine got there because of an operational failure, or because it received information feedback. In particular, this means that if a machine suffers an operational failure while in the low-yield state, it will come back up in the high-yield state. Once again, this is not always realistic: when a fuse blows, the repair person will identify the problem and correct it, but will generally

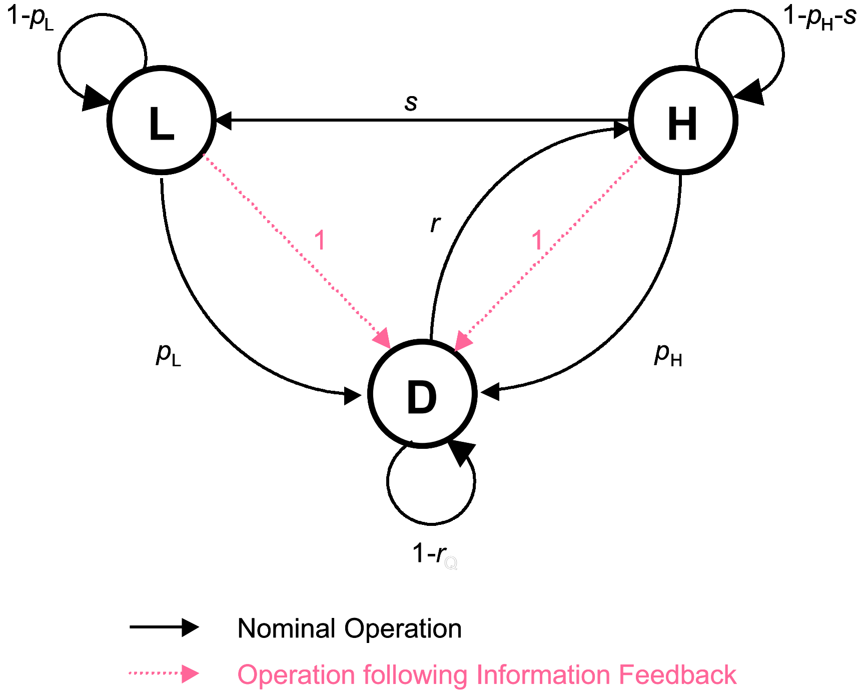


Figure 4: Example of a Markov model for a machine with two operational states and a single down state. Here, H denotes a high-yield operational state, L denotes a low-yield operational state, and D denotes the down state.

not perform a full inspection and thus will usually not discover that the tool has aged. In other words, in this model, an operational repair will buy us a "free" quality repair as well. The simplicity of this model makes it useful for analytical approaches, and it forms the basis of the results in Section 3. For the simulation studies presented in Section 4, however, we formulate a more complex model that can account for the realistic behavior of a repair crew.

A five-state machine model Figure 5 represents a five-state machine, in which two states are operational and three are down. As before, the difference between the two operational states is the rate at which defective parts are produced in each. H denotes a high-yield state, i.e. a state in which the machine produces parts most of which are good. L denotes a low-yield state, i.e. a state in which the machine produces defective parts at a higher rate. The difference between the three down states is that D_Q corresponds to *quality maintenance*, D_H corresponds to the repair of an operational failure that has occurred while the machine was in state H (at the end of which the machine will return to state H), and, D_L corresponds to the repair of an operational failure that has occurred while the machine was in state L (at the end of which the machine will return to state L).

As before, the dotted lines that go from the various operational and down states to the quality maintenance state D_Q in Figure 5 represent information feedback, i.e. a command to take the machine down for quality maintenance. Information feedback preempts the nominal behavior of the machine. That is, the machine switches from states H, L, D_H, or D_L to state D_Q with probability 1.

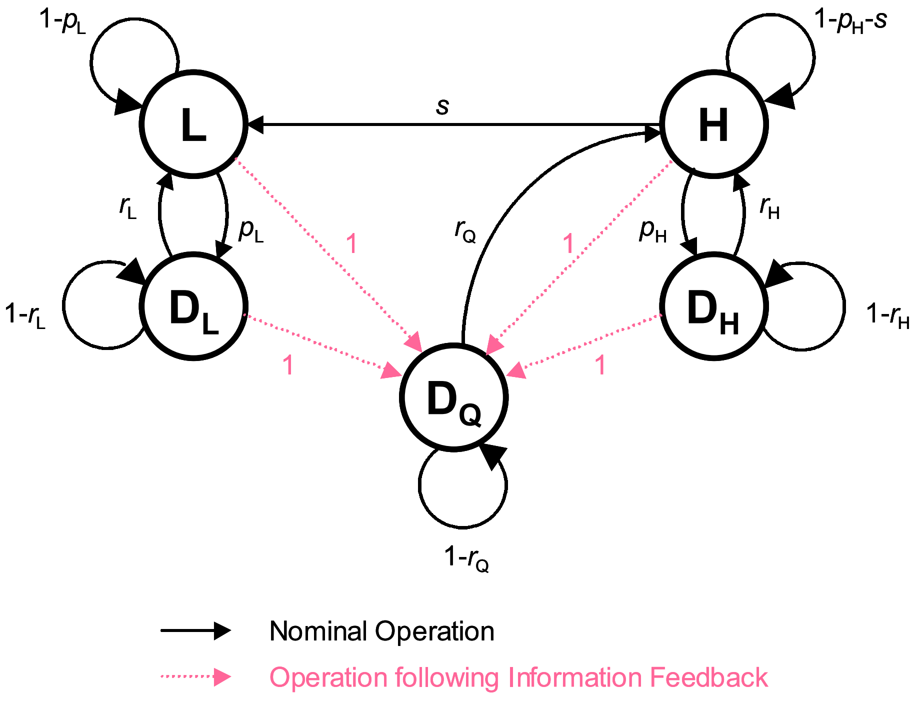


Figure 5: Example of a Markov model for a machine with two operational states and three down states. Here, H denotes a high-yield operational state, L denotes a low-yield operational state, D_Q denotes quality maintenance, D_H denotes an operational failure that has occurred while the machine was in state H, and D_L denotes an operational failure that has occurred while the machine was in state L.

The additional complexity of the five-state machine (as compared to the three-state machine) is offset by the fact that now, quality repairs are not obtained "for free" when operational repairs are performed. However, there is still a problem: if the machine is taken down for quality maintenance during an operational repair, the latter is cut short. This is because an *immediate* transition from D_H or D_L to D_Q is forced by the reception of information feedback, artificially shortening the time spent in the operational repair states. Thus, this time one gets a "partially free" operational repair when the machine is taken down for quality maintenance. Once again, this is not always realistic, and we finally introduce a model capable of properly distinguishing among operational and quality outages.

A seven-state machine model Figure 6 represents a seven-state machine, in which two states are operational and five are down. As in the five-state machine, the difference between the two operational states is the rate at which defective parts are produced in each. H denotes a high-yield state, i.e. a state in which the machine produces parts most of which are good. L denotes a low-yield state, i.e. a state in which the machine produces defective parts at a higher rate. Three of the down states are also identical to those in the five-state machine: D_Q corresponds to quality maintenance, D_H corresponds to the repair of an operational failure that has occurred while the machine was in state H (at the end of which the machine will return to state H), and, D_L corresponds to the repair of an operational failure that has occurred while the machine was in state L (at the end of which the machine will return to state L). In addition to these, however, there are now the two additional states D'_H and D'_L. These correspond to operational down states *after reception of information feedback*. In other words, if a machine is undergoing an operational repair and receives information feedback, it continues with the operational repair until it has been completed, and only then proceeds to quality maintenance.

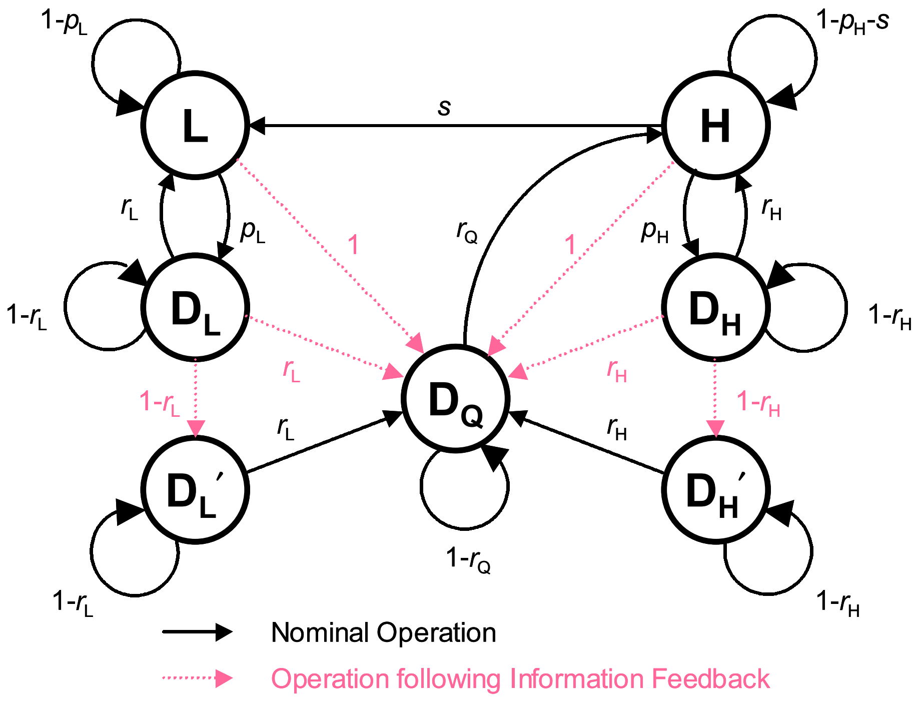


Figure 6: Example of a Markov model for a machine with two operational states and five down states. Here, H denotes a high-yield operational state, L denotes a low-yield operational state, D_Q denotes quality maintenance, D_H denotes an operational failure that has occurred while the machine was in state H, and D_L denotes an operational failure that has occurred while the machine was in state L. The states D'_H and D'_L respectively denote operational down states following reception of information feedback.

As before, the dotted lines that go from the various operational and down states to the quality maintenance state D_Q in Figure 6 represent information feedback, i.e. a command to take the machine down for quality maintenance. Information feedback preempts the nominal behavior of the machine. That is, the machine switches from states H or L to state D_Q with probability 1. If, however, the machine is in either D_H or D_L when information feedback is received, then there are two possibilities: either the operational repair is completed during the current cycle (which happens with probability r_H and r_L, respectively) and the machine moves on to the quality maintenance state D_Q; or else the operational repair is not completed, in which case it continues in the states D'_H or D'_L, respectively. These states are identical to D_H and D_L, except that the machine moves to D_Q rather than to H or L when the operational repair is completed. Because of the memoryless property, continuing an operational repair in D'_H or D'_L is indistinguishable from staying in D_H or D_L—except for the state entered after the next outward transition.

In the seven-state machine model, neither operational repairs, nor quality maintenance is obtained "for free." Rather, operational repairs and quality maintenance are performed in sequence.

Auxiliary states The states D'_H and D'_L in the seven-state machine model can be considered auxiliary states, in that their main function is to make it possible to "remember" that information feedback has been received, even though Markovian models are memoryless. There are also other cases where auxiliary states are useful.

For example, suppose that the distribution of time spent in a particular state (say, operational repair) is not geometric. This would mean that the probability of completing the repair during any given cycle is not constant but a function of the time since the failure occurred—a realistic situation when, for instance, it is unlikely that a repair will take a very short or a very long time. Such a system is not memoryless, and modeling it as a Markov chain may require the addition of auxiliary states (Howard 1971).

In addition to the use of auxiliary states to represent time distributions, it is also possible to construct arbitrarily complex machine models to account for phenomena such as the following:

- A variety of operational failures, each with different repair characteristics (time distribution, operational state reached after the repair is completed);
- The gradual aging of a machine tool and the changing characteristics of the operation and/or failures;
- Proactive control policies in which the behavior of a machine is predicted and maintenance is scheduled before a catastrophic failure occurs.

The construction of complex Markovian models for individual machines presents no conceptual difficulties, but factors and behaviors such as those outlined above must be taken into consideration for a model to represent accurately the properties of realistic production systems. Furthermore, the analysis can get very complicated when machines with complex models are embedded in production systems where they interact with each other through buffers.

2.1.2 Buffers

Buffers are passive interstage storage locations. They provide space for parts produced by upstream machines when a downstream machine is down, thus delaying the onset of blockage; likewise, they provide unprocessed parts to downstream machines when an upstream machine is down, thus delaying the onset of starvation. In this manner, they act to partially decouple sequences of machines and thus increase the production rate (Gershwin 1994).

However, when a machine occasionally switches to a state in which defective parts are produced, and such a switch must be detected as soon as possible, extra buffer capacity may be harmful because it may delay inspection—and hence, detection. The role of buffer capacity on good production rate and other performance measures under different inspection and control policies is one of the matters that this project has investigated, and is described in this report.

In a mathematical model of a production system, buffer capacity may be finite or infinite. The assumption of infinite buffers totally decouples machines and makes it possible to use powerful methods such as Jackson networks. However, in real life floor space is typically at a premium, and even very large buffers are rare. Furthermore, inventory is often undesirable, and thus large buffers are not often used. Also, in cases where a bottleneck is downstream of the buffer, an infinite buffer is *not* a good approximation of a very large buffer.

It is worth noting that large buffers do not mitigate the effects of a significant mismatch in the mean production rates of neighboring machines or sequences or machines. If, for example, the first of two machines is more productive than the second, then the buffer will generally be nearly full; furthermore, no matter how large it is made, it will still be nearly full. The usefulness of a large buffer is in mitigating the effects of high *variance* in production rate, not of a mismatch in the *mean* production rates.

Since buffers are passive devices, their models are often simple. However, there are some instances where modeling buffers presents some challenges. For example, it is usually assumed that buffers obey a *First In First Out* (FIFO) service rule. Other service rules may complicate matters significantly if it is necessary to keep track of individual parts in order to assess their quality and thereby determine the states of the machines that processed them. Similarly, it is usually assumed that buffers are distinct and of fixed capacity; situations where buffers share space, and capacity is allocated dynamically, may complicate the analysis significantly. The work presented here is limited to distinct buffers obeying FIFO service.

2.1.3 Topology

The simplest production systems are linear. Figure 2 illustrates such a system. In a linear system, unprocessed parts enter the system at machine M_1, which is assumed never to be starved. To model interruptions in delivery by suppliers, one can add an auxiliary machine (model) to the beginning of the line; an operational failure in this machine represents the event in which parts are unavailable to the production system due to a delivery problem upstream.

As material moves through the production system, machine M_i takes a part from buffer B_{i-1}, processes it, and passes it on to buffer B_i.

At the end of a k-machine system, parts leave the system at machine M_k, which is assumed never to be blocked. To model variability in removing parts from the production system (e.g. fluctuating sales), one can add an auxiliary machine (model) to the end of the line; an operational failure in this machine represents the event in which space to discharge finished parts is unavailable to the production system due to a delivery problem downstream.

At specific locations in the production line, between a machine and the buffer that follows it, are inspection stations. The time these stations take to inspect each part is assumed in this report to be negligible compared to production cycle time.

If an inspection station determines a part to be defective (whether correctly or not—see Section 2.3), it may scrap it locally (i.e. remove it from the production system right there) or mark it for scrapping or reworking downstream after the part has left the production system. In other words, conservation of material may or may not be satisfied at an inspection station, depending on whether or not scrapping is performed locally. If parts are reworked at a rework station, the location of the rework station influences both the action taken upon detection of a defect, and the placement of the inspection station.

Each machine is associated with a particular feature of a part; each inspection station is responsible for inspecting the features associated with a particular subset of upstream machines. The subset of machines for which a given inspection station is responsible is referred to as its *inspection domain*. It is assumed here that the inspection stations are non-overlapping and collectively exhaustive. Thus, the

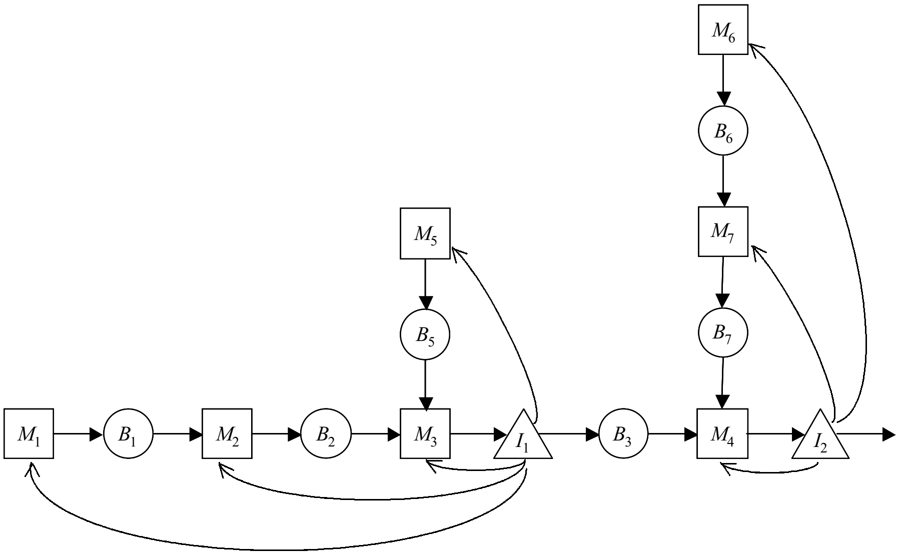


Figure 7: Example of a non-linear production system with feeder lines. Material enters the system at machines M_1, M_5, and M_6. Machines M_3 and M_4 are assembly systems. The inspection domains of inspection stations I_1 and I_2 are indicated by the bold lines.

first inspection station in the production system inspects the features associated with all the machines upstream of it; each subsequent inspection station inspects the features associated with all the machines downstream of the inspection station immediately preceding it. It is assumed that one inspection station is always located after the final machine in the system, so that the feature associated with each machine in the production system is inspected at some point.

More complicated topologies may be required to model real-life production systems. In view of the current pressures to modularize and outsource, for example, it was deemed necessary to analyze production systems composed of a short main line with multiple feeder lines. Figure 7 shows such a production system. Parts enter the system at multiple locations; it is assumed that the first machine in each branch is never starved. Production systems with multiple exit points were not considered; only assembly stations were modeled.

Even more complex production systems may include loops; such systems were not analyzed in the project reported here.

2.2 Quality Failures

This section describes a taxonomy for potentially realistic (even if analytically complex or even intractable) models spanning different types of quality failures and defect processes.

We begin by defining quality failures. This can be done in at least two ways:

- *Out-of-spec.* A quality failure may be defined as the event of a feature not being within specifications.

- *Quality loss function.* A quality loss function may describe the influence of a given feature on the performance characteristics of the part.

The first definition is assumed here.

2.2.1 Internal vs. External Failures

Quality failures may be due to external causes or internal causes. The latter include failures that occur at the machines inside the production line under study—e.g. a drill bit breaks or a paint nozzle gets clogged, resulting in a product that is sub-standard.

External failures occur outside the production system proper, perhaps in a supplier's factory, or during transportation. Examples of such failures include instances where parts delivered by the supplier are damaged, out-of-sequence, incorrect or mislabeled, or where raw material is of poor quality or incompatible with the production process.

Like delays in delivery or removal, external quality failures too can be modeled as occurring at auxiliary machines at the entry point(s) of the production system. Consequently, external quality failures are not treated separately here. Note, however, that actions taken pursuant to the detection of external failures may differ significantly from actions taken after internal failures are detected.

2.2.2 Dynamic Characteristics

The incidence of quality failures (defects) is often reported as an aggregate measure, such as the fraction of defective parts out of the total (i.e. the *yield*—see Section 2.5). However, our research has established that the dynamic (time) characteristics of defects have a pronounced effect on system performance. Specifically, the dependence or independence of individual occurrences of failures across time can be given precise probabilistic definitions. The following terms were coined in the course of the research described here, and are not traditionally used in the manufacturing literature. They are intended to convey the stochastic properties of each type of failure.

Bernoulli quality failures Quality failures that occur independently of each other are generally referred to in the traditional quality literature as "random," "common," or "chance" failures. The presence of a failure in one part says nothing about the likelihood of occurrence of a failure in any other part. For example, a pre-existing defect in an unmachined part may cause a quality failure which is specific to that part only. Such failures may also be associated with difficult operations or those that are based on developing technology.

The first graph in Figure 8 represents Bernoulli quality failures. The horizontal axis is time. Dots above the line indicate good parts; dots below the line indicate bad parts. (Periods when the machine is down are suppressed.) In this graph, most of the parts produced by the machine are good, but some are bad. The occurrence of each bad part is completely independent of previous bad parts.

Persistent quality failures In some cases, a quality failure occurs when a state change takes place in a machine, e.g. a tool breaks. Such failures are generally referred to in the literature as "systematic," "special," "assignable," or "unusual" failures. Once a failure occurs, every subsequent part will be bad until the machine is repaired.

The second graph in Figure 8 represents persistent quality failures. At first, the machine only produces good parts. Then a state change occurs, and the machine begins to produce bad parts only.

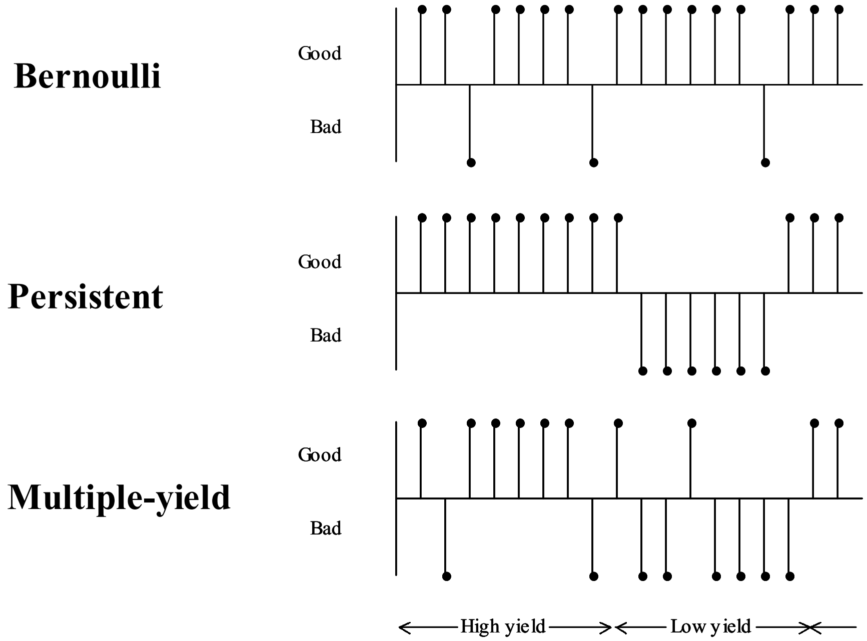


Figure 8: Simulated examples of Bernoulli, persistent, and multiple-yield quality failures.

After some time, another state change occurs, and (following a down period, which is omitted) the machine returns to producing good parts only.

Multiple-yield failures There may be cases where failures occur independently but at different rates, depending on the state in which the machine finds itself. Specifically, the machine may produce defective parts with a certain small probability p when it is in good working order; when it is in need of adjustment, however, it may produce defective parts with a certain probability $q > p$.

The third graph in Figure 8 represents multiple-yield quality failures. At first the machine is in a high-yield state, and produces bad parts in Bernoulli fashion, but at a low rate. It then switches to a low-yield state, in which it also produces bad parts in Bernoulli fashion, but this time at a higher rate. Ultimately (following a down period, which is omitted) it switches back to the high-yield state.

Table 1 indicates the correspondence between these quality failure types and real quality failures that may occur in the context of automotive manufacturing.

Given these definitions, one may justifiably ask if a given feature must be associated with only one type of quality failure, or if it can be associated with several at different times. One may also wonder how one can reliably identify the type of quality failure from empirical observations and/or engineering considerations. These are statistical issues that have not been addressed in the work performed so far. However, there is nothing fundamentally difficult about them: straight-forward hypothesis tests are sufficient to identify, to within statistical certainty, the underlying process that has created a sample of

Type of operation	Bernoulli	Persistent	Multiple-yield
Body	Sheet metal surface contamination	Wrong type sheet metal	Uneven quality sheet metal
Machining	Loose tool	Broken tool	Tool wear-out
Spot welding	Pulled wrong trigger	Bad cooling water temperature	Workers with different levels of experience
Assembly	Part misfeed	Jammed feeder	Misaligned feeder
Paint	Surface dirt	Improper paint mix	Nozzle dirt build-up

Table 1: Quality failures in an automotive plant classified according to their dynamic properties.

data (i.e. a particular realization of the process).

Individual or batch quality failures The Markovian assumption imposes the memoryless property on the model: at any time, future transitions given the current state are independent of the past history of the model—i.e. the precise sequence of states that the model may have traversed on its way to the present state, and the time since earlier transitions.

In some cases, however, the memoryless property may not hold. For example, an entire batch of mislabelled or incorrect parts may be installed before the problem is discovered. In this case, at any given point in time during a run of defective parts, the time remaining until the end of the run depends on the time elapsed since the run began, so that the system is not memoryless. However, it is still possible to model such a situation with an underlying *hidden* Markov chain, with states *Good Batch* and *Bad Batch*. Although the state space would grow and the analysis may become more complex, there is no qualitative difference between such a model and one in which batches are not considered—i.e. it is assumed that each batch is of size one.

Time-varying quality failure rates The simplest Markovian models typically assume fixed transition rates among its states, and fixed yields in each state. However, this may not adequately reflect the behavior observed in the real system. For example, an inexperienced worker may become more skilled over time, so that yield may increase; tool wear may cause more frequent failures over time, i.e. aging may degrade the yield; the buy rate may be bad on Mondays and Fridays, due to increased absenteeism and inexperienced replacement workers; likewise, the buy rate may be bad early and late in each shift, due to warm-up periods and tiredness.

Such varying system parameters can be modeled, by once again having recourse to an underlying *hidden* Markov process. This process may have a discrete or continuous state space, which governs the numerical value of such system parameters as state transition probabilities and yields. For example, underlying the machine model may be a diffusion process, where the drift represents aging. Figure 9 shows the relationship between a decreasing yield (i.e. increasing probability of producing defective parts) and the incidence of defects in the parts that are produced.

2.2.3 Correlation and Stacking

By definition, a specific quality failure is associated with a specific feature of a part. When there are multiple sources of quality failures, certain additional factors must be taken into consideration. First,

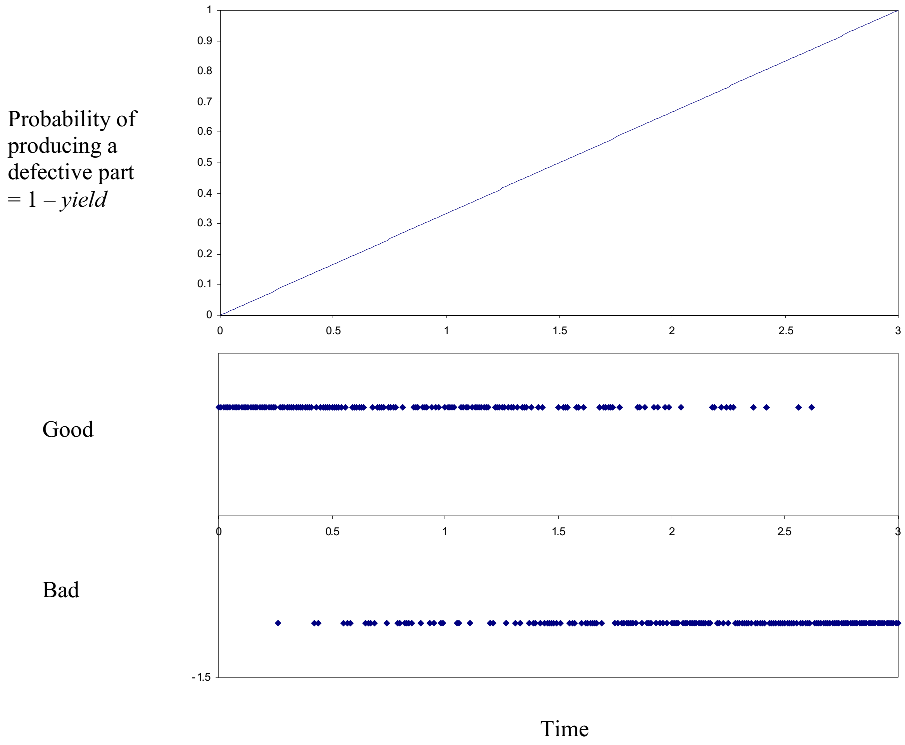


Figure 9: Simulated example of the reduction of yield due to an aging tool. The top graph shows yield decreasing over time; the bottom graph shows good parts being produced at a decreasing rate, and bad parts being produced at an increasing rate.

distinct quality failures in a given part may or may not be correlated with each other depending on whether or not the distinct features to which they correspond are interrelated in the part or in the production process. Second, the sequence in which features are processed by machines may lead to correlation among quality failures in different features, so that quality failures may suffer from variation stack-up, or benefit from variation stack-down.

Bias (mean-shift) correlation When a single machine performs several tasks, or several tools are mounted on a single head, it is possible that a single misalignment could result in a consistent shift across several different features.

Variance correlation When a single machine performs several tasks, or several tools are mounted on a single head, it is possible that a single source of imprecision (e.g. a loose arm) could result in several different features being out of specification, though not necessarily in the same direction.

Cumulative effects (Stacking) In a sequence of operations, it is possible that an upstream quality failure results in the malfunctioning of downstream operations as well, or that a downstream failure results in the corruption of the product of upstream operations. For example, if two holes are drilled side by side, a misalignment in the first might render the second defective, and vice versa. More seriously, in a two-step process such as first drilling a coarse hole and then refining it, an error in the location of the coarse hole could result in the breakage of the fine adjustment tool. Thus, multiple quality failures may occur due to a single root cause even when operations are performed by physically distinct machines.

Note that it is also possible for a downstream operation to compensate for an upstream quality failure and correct it. The aggravation of an upstream quality failure by a downstream quality failure is called *stacking up*. Conversely, the mitigation of an upstream quality failure by a downstream quality failure is called *stacking down*.

Defects in distinct features are assumed to be independent in the work presented here.

2.3 Inspection

Inspection stations are responsible for assessing the quality of processed parts. They are located between machines and the buffers that immediately follow them.

Placing an inspection station after each and every machine ("ubiquitous inspection") may allow the immediate detection and isolation of quality failures, simplifying root-cause traceability and minimizing the waste of downstream production capacity. This may be very desirable, but inspection stations are expensive, in that they consume floor space, machinery, and labor. Therefore it is necessary to choose the number and location of inspection stations carefully, so as to place them as sparsely as possible while meeting quality goals. In our analysis, therefore, we allow for fewer inspection stations than machines in the production system. In such cases, a given inspection station may be responsible for inspecting the work of several machines.

Inspection stations perform certain test on parts, in order to detect quality defects; these tests may, however, be inaccurate. When a defect is detected, the inspection station may cause some action to be taken on the part, and/or some action to be taken on the machine thought to be responsible for the quality defect in the part.

2.3.1 Destructive vs. Non-Destructive Testing

Testing for failures in manufactured parts can follow different strategies, depending both on the characteristics of the failures one is trying to detect, and on economic and other considerations. In particular, testing may or may not have a lasting effect on the part.

- *Non-Destructive Testing* does not have any lasting effect on the part. As a result, inspecting every part is an alternative to consider.

- *Destructive testing* results in the complete loss of the part. In this case, inspecting every part would result in zero production; thus, it is necessary to sample the parts, and only inspect those parts that are in the sample.

When quality failures are systematic, i.e. persistent or multiple-yield, it might be possible to make a business case for periodic destructive testing in order to detect failures reliably. When quality failures are Bernoulli, however, destructive testing cannot be justified since the discovery of a failure in one part does not provide any information about the presence or absence of failures in other parts; the fraction of failures among non-destroyed parts will be identical to the fraction of failures in the population as a whole.

2.3.2 Domain and Frequency

Each machine is assumed to work on a particular feature or set of features of the parts; each inspection station is assumed to work on the features associated with a certain set of machines, called the *inspection domain* of the inspection station.

Distinct inspection stations may or may not inspect distinct features; in our work so far, we have assumed that each machine's work is inspected by one station only. Furthermore, inspection domains can overlap, or they can be disjoint; in our work so far, we have assumed that inspection domains are disjoint. What this means is that the first inspection station inspects the features associated with all machines upstream of it, and each subsequent station inspects the features associated with all the machines between the inspection station immediately upstream of it and the machine it immediately follows.

It is assumed that if a production system contains any inspection stations, then one of the inspection stations is always placed at the very end of the system. Thus, the inspection stations collectively inspect the work of all the machines in the production system.

In our work so far, we have assumed that all parts are inspected. It is also possible to inspect only some parts, either randomly (at some chosen rate), or periodically (with a given period). Note that when defects are associated with a machine state change (as in the cases of persistent or multiple-yield quality failures), inspecting a sample of parts instead of all parts may be reasonable, although this approach would necessarily delay the detection of the state change and would thus degrade performance. However, when defects are independent (Bernoulli), sampling would not be a reasonable approach: when only a sample of parts are inspected, defective parts that happen not to be in the sample will be missed, and if the defects are Bernoulli, the fraction of defective parts among those inspected will be the same as the fraction of defective parts among those not inspected. Defective parts that are not inspected cannot be scrapped or marked for reworking, and they may thus escape into the marketplace, causing customer dissatisfaction and other undesirable consequences.

2.3.3 Accuracy and Declaration of a Defective Part

The first stage in the inspection process is the detection of quality failures. Since the inspection station is potentially unreliable, we prefer to use the term *declaration* rather than *detection*. This is because the latter term connotes an authoritative judgement, whereas the former is only an *assessment* of part quality.

Given a part, the inspection station runs a series of tests on it pertaining to the features processed by the machines in the station's inspection domain. A separate test is run for each feature. Each test returns one of two results: good or defective. Tests are potentially inaccurate, meaning that a good part may be declared to be defective or vice versa. This inaccuracy is characterized by a false positive (Type I error) probability and a false negative (Type II error) probability. Note that since each test is associated with a feature, and each feature with a machine, there are as many false negative and false positive probabilities as there are machines—even though the actual testing is performed at the inspection stations.

There is a trade-off between the stringency of quality testing and the cost of testing in terms of scrapping good parts or missing bad parts. Hence, the design of an optimal testing strategy requires the achievment of a careful balance between Type I and Type II errors. Very stringent quality testing would catch most or all the quality failures, but at the cost of many false positives. When the detection of a failure results in scrapping a part or stopping the production line, false positives may be very expensive.

On the other hand, very relaxed quality testing would seldom result in unwarranted scrapping or stopping of the line, but this would be at the cost of many false negatives. Defective parts that are not identified as such consume downstream production capacity, and, if missed entirely, would result in defective products and customer dissatisfaction.

In our analysis, we keep track of the true quality characteristics of a part, in order to be able to calculate such performance measures as the rate at which defective parts reach the marketplace or the rate at which good parts are reworked or discarded. However, in reality, there is typically no *direct* information as to the quality of a part outside of the results of the tests performed by the inspection stations. (Indirect information may be obtained from customer complaints, warranty repairs, etc.) In other words, on the factory floor, parts are treated as good or defective only on the basis of the inspection results.

2.3.4 Declaration of an Impaired Machine

In addition to declaring each part to be good or defective, an inspection station may also decide whether or not a given machine in its inspection domain is in need of quality maintenance. In other words, the purpose of inspection is not only to catch defective parts, but also to signal that adjustments or corrections in the production system itself are needed, so that fewer or no defective parts will be produced in the immediate future.

A machine is considered to be in need of quality maintenance if there is evidence that it has entered a state in which it produces a larger fraction of defective parts than is desirable. This evidence is obtained by inspecting the parts processed by that machine, and inferring the state of the machine from the results of inspection.

For example, the Toyota Production System philosophy of *jidoka* requires that a machine be taken down for maintenance as soon as a single defect has been detected. This may be a good strategy when quality failures are persistent (or multiple-yield with a significant difference in yields) and inspection is highly accurate, but not otherwise. (See Section 3.2.)

Alternatively, a decision rule can be formulated based on a moving window of parts, e.g. "declare the machine to be in need of maintenance if n parts have been declared defective among the most recent m inspected parts."

When a machine is thought to need maintenance, *information feedback* is sent to that machine, i.e. a signal is sent that overrides the nominal behavior of the machine and takes it into the quality maintenance state. (See Section 2.4.2.)

2.4 Control Policies

The purpose of inspection is, of course, to trigger action. The declaration of quality defects may elicit actions in several different areas: on the part thought to be defective, on the machine thought to be responsible for the defect, and on the workforce responsible for that section of the production system.

Decisions as to the type and timing of action must be contingent upon cost defined in the broadest possible sense: cost of materials, cost of processing, cost of maintenance, cost of customer dissatisfaction, and so on. Actions can be taken reactively or pro-actively.

- *Pro-active action.* If inspection reveals a trend towards degrading quality, it may be possible to schedule preventive maintenance in order to prevent further degradation and potentially a catastrophic failure. This would only be possible if the system is non-stationary and the underlying changes can be modeled so as to allow the forecasting of performance deterioration.

- *Reactive action.* When a quality failure is detected, some action will be taken, ranging from simply noting that a failure has occurred, to stopping the production line for repair. This action may be immediate, or it may be delayed, depending on the type of action and the circumstances prevailing.

2.4.1 Action on Defective Parts

When inspection station I_i declares a part to be good, the part is moved into buffer B_j. (Because of the imperfect accuracy of the test, the part may in truth be defective.) When inspection station I_i declares a part to be defective, the part is either scrapped immediately (i.e. never makes it into buffer B_j), or it is marked as defective and moved into buffer B_j. In the latter case, it is assumed that the part will be either scrapped or reworked after it has left the production system. (Once again, a part that is scrapped or marked as defective may in truth be good.)

The advantage of scrapping locally is that parts believed to be defective are immediately removed from the system; thus, they no longer consume such resources as buffer space and machine capacity. This is especially useful when a bottleneck is present downstream, as the strain on the bottleneck is somewhat relieved by reducing the rate at which jobs arrive at it.

At the same time, scrapping locally is often not possible because material removal, storage, and transportation equipment must be on hand to handle the part to be scrapped. If a part is marked for subsequent scrapping downstream, the benefits of scrapping to the production line itself are no longer realized. Even though it is marked as defective, the part will continue to move through the production system, consuming resources as it goes.

Indeed, it is often highly impractical to scrap parts at all, particularly when they are expensive and represent a significant sunken cost in terms of raw material consumed and labor or machine capacity already expended. In such cases, parts are marked for reworking, and may have to go through some or all the operations again. Although this ensures that defective parts are not released into the marketplace, no savings of production line capacity are realized in this case.

The location of the rework station influences both the action taken upon detection of a defect, and the placement of the inspection station. If defective parts are scrapped locally, it is best to scrap them as soon as possible, i.e. to locate the inspection station as close as possible to the machine whose work it inspects. This may not always be possible, however, due the floor space constraints, availability of material handling equipment, or other factors. Likewise, if defective parts are reworked locally, correcting the defect may require some additional on-line operations, temporarily altering the operation characteristics of that machine. Alternatively, corrective work may be performed off-line, with the part taken out of the production line, reworked, and re-introduced into the line at the same point. That does not introduce any changes into the operation characteristics of the machine. In both cases, inspection and reworking are performed at roughly the same location, possibly imposing certain constraints on the choice of location. In other cases, particularly when the part is large (e.g. a partially completed car) and floor space is at a premium, defective parts are taken to a remote location where they are reprocessed. In such cases, the placement of inspection stations may not be constrained by the requirement to co-locate a reworking station.

In a line where parts are inexpensive and some amount of wasteage is acceptable, the declaration of defects may be used to compile quality statistics. In such a line, it may not be individual failures that trigger action, but rather the frequency of failures. Of course this assumes that the defective parts would eventually be identified and eliminated downstream.

2.4.2 Action on Impaired Machines

When an inspection station declares a machine to be in need of quality maintenance, it sends information feedback to the machine. This signals to the machine or to its operators that it must be taken into the quality maintenance state. Taking a machine down for servicing is clearly not a decision that should be made lightly: a cost will invariably be involved in making this decision, including not only the cost of the service call but also lost productivity due to the unscheduled outage.

Information feedback preempts the nominal operation of the machine; in other words, whatever state the machine may have remained in or moved to were it not for information feedback, the decision to go into quality maintenance overrides the dynamics of the machine itself.

Since inspection may take place far downstream of a machine, some time may elapse between the occurrence of a state change in the machine, and the detection of that change by the inspection station. In the meantime, the machine may have undergone one or more state changes of its own accord. What exactly happens to the machine once it receives information feedback thus depends on the model of the machine: it may be taken down for maintenance immediately (at the end of the current cycle), or it may remain in an operational repair state and only move on to the quality maintenance state after the repair has been completed. (See Section 2.1.1.)

Another issue to consider—in view of the distance between the machine and the inspection station that monitors its performance—is the fact that a potentially significant number of defective parts may have already been produced by the time the machine is taken down for quality maintenance. For example, suppose that there are some FIFO buffers between the machine and the inspection station, and suppose that they collectively have n parts in them at the time the machine begins to produce defective parts; barring any other machine failures, it will take n machining cycles for the first defective part to reach the inspection station. If the quality failure mechanism is persistent, furthermore, then the n parts behind it will all be defective. This means that once information feedback has been sent to the machine, it should receive no further commands until all the parts already in the pipeline have drained out. We call this the *dead period*. Experiments show that the dead period has a significant influence on the total production rate of the machine, since the machine is not taken down for maintenance during a dead period, and dead periods can be of significant duration; the longer or more frequent the dead periods, the fewer the opportunities to take an operational machine down. In this respect, dead periods compensate to some extent for decision rules that are not conservative enough. For example, taking a machine down for quality maintenance as soon as a single part is declared to be defective (*jidoka*) may result in degraded performance in some cases; dead periods may mitigate this effect, but they will do so at different rates depending on the distance between the machine and the inspection station that monitors it.

2.4.3 Action on People

The detection and analysis of quality failures can affect the people involved in the production line.

Traceability Factors such as line topology (especially the existence of alternate routes), location of inspection stations, and the complexity of operations will have a significant effect on the ease or difficulty with which the root causes of quality failures are isolated, diagnosed, and corrected. Ideally, the detection and analysis of quality failures should help locate their root causes.

Motivation toward better quality The detection of quality failures provides an information feedback loop that will help better planning, motivate process improvements, and bring about a rise in product quality.

Learning Likewise, the detection and analysis of quality failures provides an information feedback loop that will allow personnel to develop a better understanding of the manufacturing processes. This will result in faster diagnosis and correction over time.

2.5 Measures of System Performance

There is no single performance measure on the basis of which a complex manufacturing system can be adequately summarized. Very diverse quantities reveal different aspects of the performance of a production system, and which of these aspects is the most important will often depend on external factors.

For example, inventory and lead time are positively correlated, so that high inventory tends to cause long lead times. If a production system manufactures off-the-shelf items, it may be desirable to have sufficient inventory available for customers to buy, even if production lead time is long as a consequence; by contrast, if a production system manufactures custom—or even partly customized—products, it may be important to have low lead times, implying that inventory must be kept to a minimum.

Similarly, high production rates can sometimes be achieved only at the expense of high in-process inventory. A production system that manufactures products that are in great demand may choose to tolerate a large inventory provided that a high enough production rate is achieved; by contrast, a production system that manufactures products that are perishable (not only in a literal sense, but also in the sense that new models quickly make old models obsolescent) may prefer to keep inventory levels very low, even if that limits the production rate.

Some important measures of production system performance are defined below.

2.5.1 Quantity and Quality

Two aspects of manufacturing that are of particular interest to this project are *quantity* and *quality*. These have been studied separately by past researchers, but the ways in which they interact is the main contribution of the work reported here.

Quantity: Total Production Rate A measure of quantity in the performance of a production line is the mean rate at which processed parts emerge from the end of the line, measured in parts per cycle or parts per time unit. When it is measured in parts per cycle, it is often called the *efficiency* of the line. One way to interpret efficiency is as the probability that a processed part emerges from the end of the line during any given cycle.

At steady-state, the average rate at which each machine produces parts is the same; this is because more productive machines will tend to become blocked or starved more often than less productive machines. Thus, production rate could be measured at any machine along the line.

For reasons that will become clear in the next paragraph, we rename the performance measure described above as the *total production rate*.

Quality: Good and Bad Production Rates The rate at which processed parts emerge from the production system does not tell the whole story. The parts produced by the line may be of good quality (i.e. up to specifications), or of bad quality (i.e. defective). Thus, we may be tempted to define the *good production rate* as the rate at which good parts emerge from the production system, and the *bad production rate* as the rate at which defective parts do so.

However, for workers on the factory floor, the only source of information as to the quality of produced parts is the results of inspection. Since inspection is not perfectly accurate, a part declared to be good may be defective, and a part declared to be defective may be good. Hence, we must tighten the above definitions as follows: we define the *good production rate* as the rate at which good parts that are *known to be good* emerge from the production system, and *bad production rate* as the rate at which defective parts *known to be defective* do so.

Note that this information is only available to the "omniscient analyst" and not to the workers on the factory floor, whose knowledge is limited by the results of the inspection process. For them, it is more appropriate to define the *apparent good production rate* as the rate at which parts *thought to be good* emerge from the production system, and the *apparent bad production rate* as the rate at which parts *thought to be defective* do so.

Depending on the characteristics of the production system and the product, either good or bad production rate may be the more important performance measure. If, for example, the product is inexpensive, it may be acceptable to strive for a very high good production rate even if that comes at the expense of a high bad production rate; by contrast, if the product is very valuable and reworking it is expensive, then it may be preferable to minimize bad production rate, even if that means that the highest possible good production rate is not achieved.

Quality: Miss and Waste Rates Since good and bad production rates are defined on the basis of a *correct* classification on the part of the inspection station(s), it is necessary to define measures that reflect mistakes made by the inspection process.

We define the *miss rate* as the rate at which bad parts that were *declared to be good* emerge from the production system, and *waste rate* as the rate at which good parts that were *declared to be bad* do so.

Depending on external factors, these measures may be given different degrees of importance. For example, a bad part that is declared to be good will be released into the marketplace, and may cause defective products to reach the customer; this may entail waranty repairs, product recalls, ill will, and even injured customers and ensuing litigation. Thus, it may be desired to keep the miss rate at the lowest possible level for certain key features (or components) of the product. By contrast, in the case of expensive features that do not present significant risk, it may be desirable to minimize the waste rate even if that is at the expense of reducing other measures of performance.

2.5.2 Inventory

Common wisdom holds that inventory is bad; advocates of lean manufacturing are especially emphatic on this count. However, some level of inventory may be desirable to achieve a particular production rate.

Thus, it is necessary to have a measure by which to assess the amount of inventory in the production system, and to determine if that amount is justified based on the rest of the system's performance.

The *mean buffer level* of a particular buffer B_i in the production system is the statistical mean (or alternatively the long-term average) number of parts in the buffer during a given cycle.

2.5.3 Production Lead Time

The amount of time elapsed between the entry of a part (or of the first component thereof to enter the production system) and the emergence of the finished product from the production system is referred to as the *production lead time* (and sometimes also *cycle time*; in the present report, however, that term is reserved for the duration of machine operations). This measure is undefined for parts that are scrapped.

Production lead time does not have a simple relationship to production rate: parts may emerge from the production system at a very high rate, but they may have stayed inside the production system for a very long time. Production lead time is strongly correlated with such system characteristics as cycle time and number of stages, and such performance measures as in-process inventory.

3 Analytical Results

In this section, we describe results obtained from analytical modeling of manufacturing systems. Analytical models are those that are constructed from mathematical relations like equations. They are desirable in that when they are solvable, they can produce numerical results quickly. In addition, their construction and evaluation can provide insight into the behavior of the systems we study. Their disadvantage is that they are often difficult or even impossible to solve. In that case, simulation models like those described elsewhere in this report should be used.

The results reported here are a summary of those of (Kim 2004) and Kim and Gershwin (2005).

3.1 Mathematical Models

3.1.1 Single Machine Model

There are many possible ways to characterize a machine for the purpose of simultaneously studying quality and quantity issues. Here, we model a machine as a discrete state, continuous time Markov process. Material is assumed continuous, and μ_i is the speed at which Machine i processes material while it is operating and not constrained by anything else in the system. It is a constant, in that μ_i does not vary with time or, when the machine is part of a larger system, it does not depend on the repair state of the other machine(s) or the buffer level(s).

Figure 10 shows the assumed state transitions of a single machine with persistent-type quality failures. In the model, the machine has three states:

- State 1: The machine is operating and producing good parts.

- State -1: The machine is operating and producing bad parts.

- State 0: The machine is not operating.

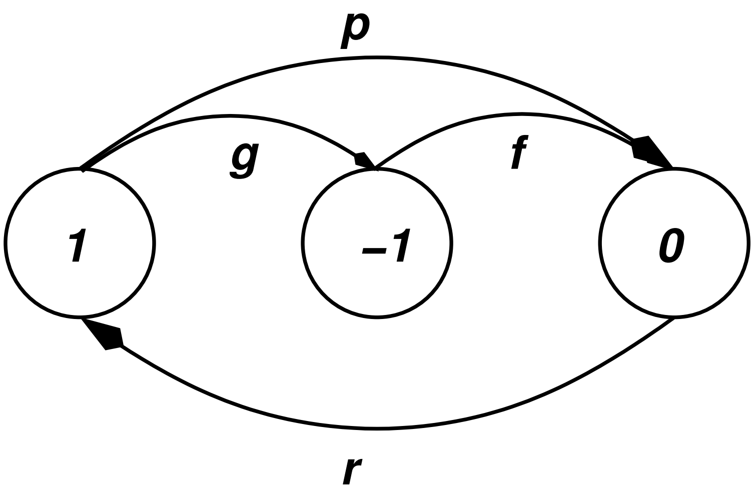


Figure 10: Three-state machine model.

(Note that when the machine is in state -1, it is producing defective parts but the operator does not know that yet; as soon as inspection reveals that defective parts are being produced, the machine is taken to state 0 for quality maintenance.)

The machine therefore has two different failure modes (i.e. transitions to failure states from state 1):

- *Operational 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 (and starts producing bad parts) due to a failure like a sudden tool damage.

This is a simplification of the model in Figure 4.

When a machine is in state 1, it can fail due to a non-quality-related event. It goes to state 0 with transition probability rate p, which is the reciprocal of the *Mean Time to Fail* (MTTF). After that an operator fixes it, and the machine goes back to state 1 with transition rate r, the reciprocal of the *Mean Time to Repair* (MTTR). 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 g. Here g is the reciprocal of the *Mean Time to Quality Failure* (MTQF). A more stable operation leads to a larger MTQF and a smaller g.

The machine, when it is in state -1, can stop for two reasons: it may experience the same kind of operational failure as it does when it is in state 1; and the operator may stop it for repair when he learns that it is producing bad parts. We assume that the transition from state -1 to state 0 occurs at probability rate $f = p + h$ where h is the reciprocal of the *Mean Time To Detect* (MTTD). A more reliable inspection leads to a shorter MTTD and a larger f. (The detection can take place elsewhere, for example at a remote inspection station.) Note that this implies that $f > p$. All the indicated transitions are assumed to follow exponential distributions.

A larger $h = f - p$ means that MTTD is smaller and that fewer bad parts escape detection. Therefore we refer to inspection stations with larger h as more reliable or more accurate; and those with small h as poor. Note also that we are only modeling misses or false negatives in this section; we are not modeling false alarms or false positives. That is, we do not consider the possibility of declaring a good part bad.

The assumption of exponential distributions is made for computational convenience. Like all such engineering assumptions, it should be judged on the predictive power of the model rather than its fidelity to reality. It may differ from reality where we have quality information feedback (Section 3.1.6). In that case, the time until the quality failure is detected is proportional to the amount of material in the buffer when the quality failure occurs, and that amount is not exponentially distributed.

Here, for simplicity, we assume that whenever a machine is repaired, it goes back to state 1. It makes the mathematics easier to do, but it comes at a cost of possible inaccuracy. This assumption is discussed in Section 2.1.1.

Single machine analysis To determine the production rate of a single machine, we first determine the steady-state probability distribution. This is calculated based on the probability balance principle: when a system is in steady state, the probability rate of leaving a state is the same as the probability rate of entering that state. We have

$$(g+p)P(1) = rP(0) \tag{1}$$

$$fP(-1) = gP(1) \tag{2}$$

$$rP(0) = pP(1) + fP(-1) \tag{3}$$

The probabilities must also satisfy the normalization equation:

$$P(0) + P(1) + P(-1) = 1 \tag{4}$$

The solution of Equations (1)–(4) is

$$P(1) = \frac{1}{1 + (p+g)/r + g/f} \tag{5}$$

$$P(0) = \frac{(p+g)/r}{1 + (p+g)/r + g/f} \tag{6}$$

$$P(-1) = \frac{g/f}{1 + (p+g)/r + g/f} \tag{7}$$

The *total production rate*, including good and bad parts, is

$$P_T = \mu(P(1) + P(-1)) = \mu\frac{1 + g/f}{1 + (p+g)/r + g/f} \tag{8}$$

The *effective production rate*, the production rate of good parts only, is

$$P_E = \mu P(1) = \mu\frac{1}{1 + (p+g)/r + g/f} \tag{9}$$

(This quantity is also called the *good production rate*.) Since there is no scrapping, the rate at which parts enter the system is equal to the rate at which parts leave the system, so that the *yield* is

$$Y = \frac{P_E}{P_T} = \frac{P(1)}{P(1) + P(-1)} = \frac{f}{f+g} \tag{10}$$

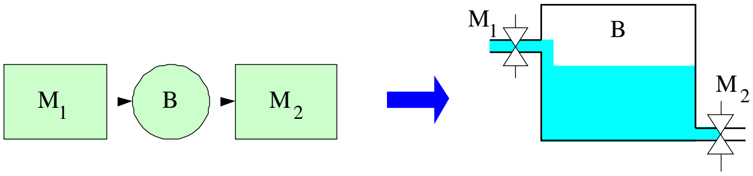


Figure 11: The two-machine-one-buffer continuous model.

3.1.2 The 2-Machine-1-Buffer Continuous Model

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.

We first study 2-machine-1-buffer (2M1B) models. (A decomposition technique that divides a long transfer line into multiple 2-machine-1-buffer models has been developed; see Section 3.3 and Kim (2004).) Decomposition for systems in which quality issues are not considered is described in Gershwin (1994).) Among the various modeling techniques for the 2M1B case, including deterministic, exponential, and continuous models, the continuous material line model 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 of Gershwin and Schick (1980) and Gershwin (1994) by adding another machine failure state. Figure 11 shows the 2M1B continuous model where the machines, buffer and discrete parts are represented as valves, a tank, and a continuous fluid.

We assume that an inexhaustible supply of workpieces is available upstream of the first machine in the line, and an 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). That is, transitions from 1 to -1 or 0 are only allowed when the machine is operational, and not blocked or starved.

Finally, we assume that each machine works on a different feature. For example, the two machines may be making two different holes. We do not consider cases where the both machines work on the same hole, in which the first machine does a roughing operation and the second does a finishing operation. This allows us to assume that the failures of the two machines are independent.

3.1.3 The Infinite Buffer Case

The infinite buffer case is a special 2M1B line in which the size of the buffer (B) is infinite. This is an extreme case in which the first machine (M_1) never suffers from blockage. To derive expressions for the total production rate and the effective production rate, we observe that when there is infinite buffer capacity between two machines (M_1, M_2), the two machines are in some sense decoupled, so that the total production rate of the 2M1B system is the minimum of the total production rates of M_1 and M_2. The total production rate of machine i is given by Equation (8), so the total production rate of the 2M1B system is

$$P_T^{\infty} = \min\left[\frac{\mu_1(1+g_1/f_1)}{1+(p_1+g_1)/r_1+g_1/f_1}, \frac{\mu_2(1+g_2/f_2)}{1+(p_2+g_2)/r_2+g_2/f_2}\right] \tag{11}$$

The probability that machine M_i does not add defects (non-conformities) is

$$Y_i = \frac{P_i(1)}{P_i(1) + P_i(-1)} = \frac{f_i}{f_i + g_i} \tag{12}$$

Since there is no scrapping or reworking in the system, the system yield is

$$Y = Y_1 Y_2 = \frac{f_1 f_2}{(f_1 + g_1)(f_2 + g_2)} \tag{13}$$

As a result, the effective production rate is

$$P_E^{\infty} = \frac{f_1 f_2}{(f_1 + g_1)(f_2 + g_2)} P_T^{\infty} \tag{14}$$

The effective production rate evaluated from Equation (14) has been compared with a discrete-event, discrete-part simulation and has shown good agreement.

As indicated in Section 3.1.1, the detection of quality failures due to machine M_1 need not occur at that machine. For example, the inspection of the feature that M_1 works on could take place at an inspection station at or just after M_2, and this inspection could trigger a repair of M_1. (This is *quality information feedback*. See Section 3.1.6.) In that case, the MTTD of M_1 (and therefore f_1) will be a function of the amount of material in the buffer. We return to this important case in Section 3.1.6.

3.1.4 The Zero Buffer Case

In the zero buffer case, there is no buffer space between the machines. Whenever one of the machines stops, the other one is also stopped. When both of them are working, the production rate is $\min[\mu_1, \mu_2]$. This is the other extreme case, in which blockage and starvation take place most frequently. The details of the analysis appear in Kim (2004) and Kim and Gershwin (2005). We summarize the results here.

We define

$$p_i^b = p_i \frac{\min(\mu_1, \mu_2)}{\mu_i}, \; g_i^b = g_i \frac{\min(\mu_1, \mu_2)}{\mu_i}, \text{ and } f_i^b = f_i \frac{\min(\mu_1, \mu_2)}{\mu_i} \tag{15}$$

This is because we have assumed that failures are operation-dependent. As a consequence, the rates of failure must be reduced for the faster machine. The reduction of p_i to p_i^b is explained in detail in Gershwin (1994) and Gershwin and Schick (1980). We must reduce g_i and f_i for the same reasons.

The total production rate is then

$$P_T^0 = \frac{\min[\mu_1, \mu_2]}{1 + \dfrac{f_1^b(p_1^b + g_1^b)}{r_1(f_1^b + g_1^b)} + \dfrac{f_2^b(p_2^b + g_2^b)}{r_2(f_2^b + g_2^b)}} \tag{16}$$

and the effective production rate is

$$P_E^0 = \frac{f_1^b f_2^b}{(f_1^b + g_1^b)(f_2^b + g_2^b)} P_T^0 \tag{17}$$

3.1.5 The 2-Machine-1-Finite-Buffer Line

The two-machine line is the simplest non-trivial case of a production line. In the existing literature on the performance evaluation of systems in which quality is not considered, two-machine lines are used in decomposition approximations of longer lines. (See Gershwin (1994).)

We analyze this system by extending the continuous material two-machine approach of Gershwin and Schick (1980). As in the earlier work, the goal of the analysis is to calculate the average production rate and buffer level. It is more complex because first, each machine has three states rather than two, and second, there are now two different production rates: the total production rate and the production rate of good material.

State definition The state of the 2M1B line is defined as (x, α_1, α_2) where

- x: the total amount of material in buffer B, $0 \leq x \leq N$,
- α_1: the state of M_1. ($\alpha_1 = -1, 0,$ or 1),
- α_2: the state of M_2. ($\alpha_2 = -1, 0,$ or 1)

The parameters of machine M_i are $\mu_i, r_i, p_i, f_i, g_i$ and the buffer size is N.

Model development As in Gershwin and Schick (1980), we analyze this system by finding its steady-state probability distribution. This consists of a set of density functions $f(x, \alpha_1, \alpha_2)$ and a set of probability masses $P(0, \alpha_1, \alpha_2)$ and $P(N, \alpha_1, \alpha_2)$. Since α_1 and α_2 each take three values, there are 9 density functions and 18 masses. We therefore derive and solve 9 differential equations (considerably more than the 4 differential equations in the earlier work) and numerous boundary conditions .

As usual, the solution of the differential equations requires the solution of a set of simultaneous polynomial equations. New complexity arises from the fact that the number of roots—and therefore the number of boundary conditions—depends on the values of the parameters.

Validation Numerous cases were compared with simulation. The average absolute value of the error in the effective production rate was under 1% and it was under 2% for the average inventory.

3.1.6 Quality Information Feedback

Factory designers and managers know that it is ideal to have inspection after every operation, but it is often costly to do this. As a result, many factories are designed so that multiple inspections are performed at a small number of stations. In this case, downstream inspection at can detect bad features made by upstream machines. When that happens, the inspection signals that the machine that made the bad feature should be repaired. We call this *quality information feedback*.

We focus here on quality information feedback in 2M1B systems when M_1 produces defective features but does not have inspection, and M_2 has inspection that can detect bad features made by M_1. In this situation, as we demonstrate below, the yield of a line is a function of the size of the buffer. This is because when the buffer gets larger, more material can accumulate between an operation at M_1 and the inspection of that operation at M_2. All such material will be defective if a persistent quality failure takes place. In other words, if buffer is larger, there tends to be more material in the buffer and consequently more material is defective. We can approximate this phenomenon with the adjustment of the transition probability rate f of M_1 from state -1 to state 0.

In Kim (2004) and Kim and Gershwin (2005) we describe a method for using the two-machine model described in Section 3.1.5 to analyze quality information feedback. To use that model, we approximate the transition time from state -1 to state 0 in Machine M_1 as exponentially distributed. The rate is given by

$$f_1 = p_1 + h_1$$

where $1/h_1$ (the mean time of transitions due to the detection of a quality defect) is proportional to the average amount of inventory in the buffer. This creates a computational problem: since the average buffer level is not known, it is impossible to choose h_1. We solve this problem by iterating: we guess h_1; then evaluate the average buffer level; then we use the average buffer level to recalculate h_1; and we repeat. Experience shows that this procedure converges reliably and rapidly.

We compared the effective production rate and the average inventory from the analytic model with a simulation. Numerous cases were evaluated and the average absolute value of the errors in P_E and $\overline{x}$ estimates were about 1% and under 4%, respectively.

3.1.7 Insights From Numerical Experimentation

In this section, we perform a set of numerical experiments to provide intuitive insight into the behavior of production lines with inspection and quality information feedback. In Figures 12–15, we compare pairs of systems that are identical except for a quality feedback loop in one of them. In both systems, only M_1 can make bad parts, and M_1 can inspect its own parts. M_1's failure behavior is persistent, as described in Section 3.1.1. In one system of each pair (*with feedback*), M_2 also inspects the parts; in the other (*without feedback*) M_2 does no inspection. In both, M_1 is taken down for service when a bad part is observed.

We consider only a system with feedback in Figures 16–18.

Beneficial buffer case The system with feedback has more inspection events than the system without feedback. This causes the machines to stop more frequently. As a result, the total production rate of the line is lower. However, the effective production rate can be greater since additional inspections reduce the making of defective parts. This phenomenon is shown in Figure 12. Note that the total production rate P_T without quality information feedback is consistently higher than P_T with quality information feedback regardless of buffer size; the opposite is true for the effective production rate P_E. Also, both the total production rate and the effective production rate increase with buffer size, with or without quality information feedback.

Even though a larger buffer increases both total and effective production rates in this case, it decreases yield. As explained in Section 3.1.6, the system yield is a function of the buffer size if there is quality information feedback. Figure 13 shows the system yield decreasing as the buffer size increases when there is quality information 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 in state -1 and the first machine does not detect the quality defects. This is a case in which *a smaller buffer improves quality (i.e. the yield),* which is widely believed to be generally true. If there is no quality information feedback, then the system yield is independent of the buffer size (and is substantially lower).

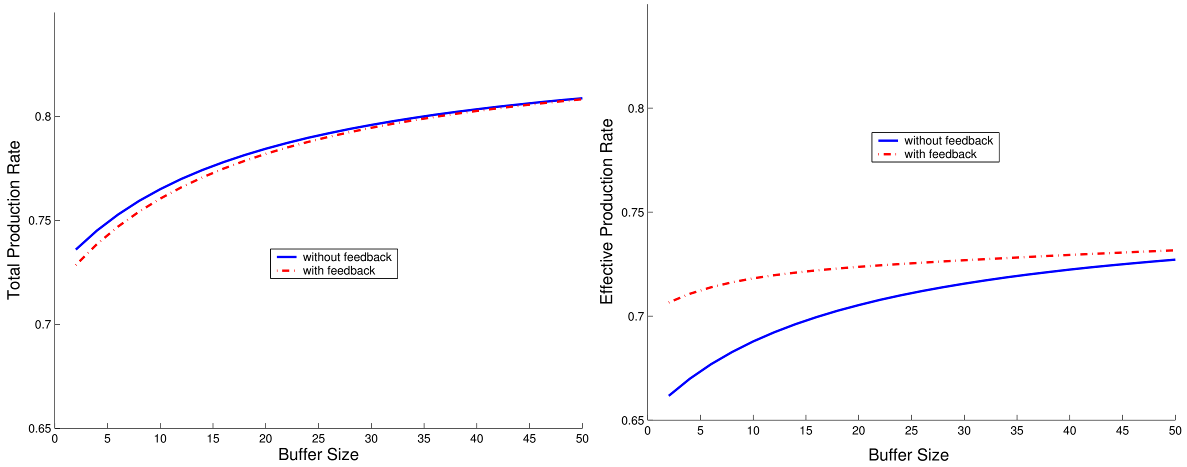


Figure 12: Production rates with/without Quality Information Feedback, for a case where larger buffer size helps performance.

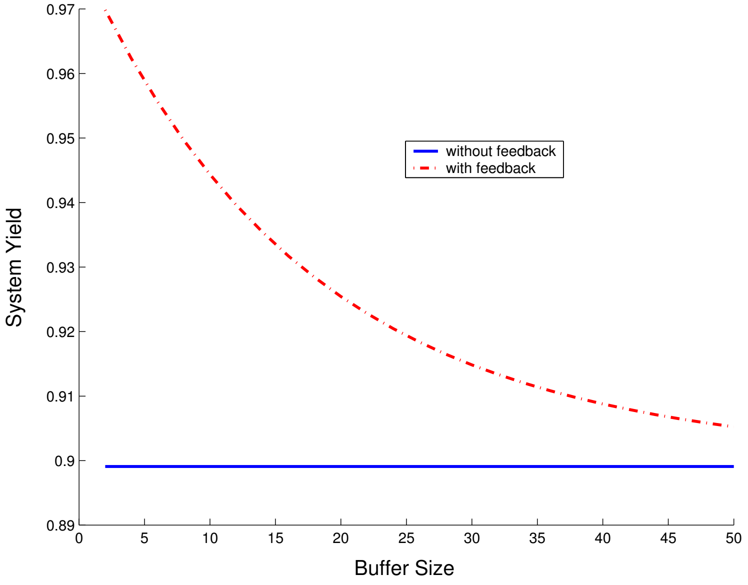


Figure 13: System yield as a function of buffer size.

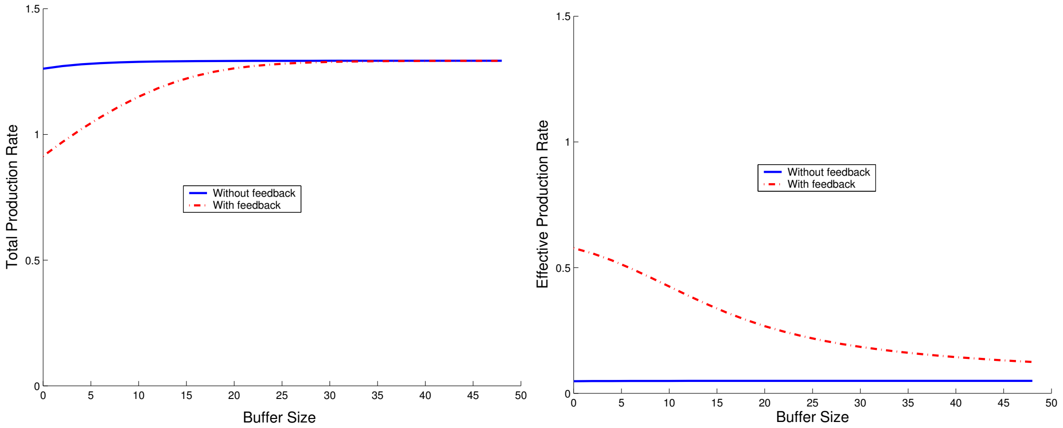


Figure 14: Total production rate and effective production rate for a case where larger buffer size hurts performance.

Harmful buffer case From experience with models that do not consider quality, we might expect that increasing the buffer size leads to a higher effective production rate. This is the case in Figure 12. But under certain conditions, the effective production rate can actually decrease as buffer size increases.

Figure 14 shows a case in which a buffer size increase leads to a lower effective production rate. Total production rate, by contrast, monotonically increases as buffer size increases.

This behavior is exhibited in cases where

- The first machine produces bad parts frequently: this means g_1 is large.
- The inspection at the first machine is poor or non-existent. That is, the first machine lets many bad parts go without declaring them bad. The inspection at the second machine is much better. This means $h_1 \ll h_2$ or $f_1 - p_1 \ll f_2 - p_2$.
- There is quality information feedback.
- The isolated production rate of the first machine is higher than that of the second machine. That is,

$$\mu_1 \frac{1 + g_1/f_1}{1 + (p_1 + g_1)/r_1 + g_1/f_1} > \mu_2 \frac{1 + g_2/f_2}{1 + (p_2 + g_2)/r_2 + g_2/f_2}$$

When the first machine goes from state 1 to state -1, it starts producing bad parts but this not detected at the first machine. As a consequence, the bad parts enter the buffer as the good parts are drawn out of it by the second machine. As long as there are good parts in the buffer, the second machine cannot detect the quality failure in M_1 and therefore cannot signal that repair is needed. The repair of M_1 can only take place after the buffer contains only bad parts.

Because the isolated production rate of the first machine is higher than that of the second machine, the buffer is usually full. As a consequence, there is a long delay between the failure and the time when all the good parts are out of the buffer. When the failure is detected, the buffer has only bad parts, and it is probably close to full.

If the buffer is large enough, increasing the buffer size produces only a small increase in total production rate. However, since the buffer fills with bad parts almost every time M_1 goes to the -1

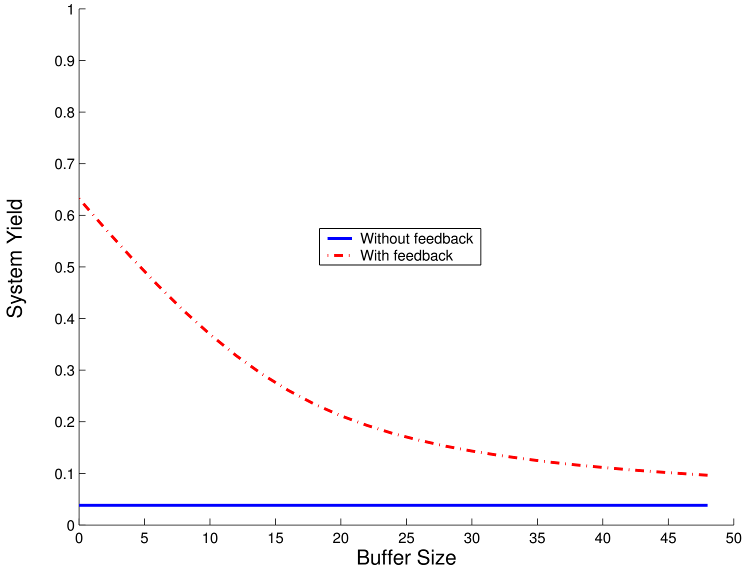


Figure 15: System yield as a function of buffer size.

state, a larger buffer produces more bad parts. Consequently, a larger buffer reduces the production rate of good parts in this system.

The system yield for this case is shown in Figure 15. Note that the yield decreases dramatically as the buffer size increases. In this case, the decrease of the system yield is more than the increase of the total production rate so that the effective production rate monotonically decreases as buffer size gets bigger.

How to improve quality in a line with persistent quality failures There are two major ways to improve quality. One is to increase the yield of individual operations and the other is to improve the accuracy of the inspection station. Performing 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 Quality Failure (MTQF) (i.e. decrease g). On the other hand, the inspection policy aims to detect bad parts as soon as possible and prevent their flow toward downstream operations. Improving inspection decreases the mean time to detect (MTTD) (i.e. increases h and therefore increases f).

There is no reason to use only one of these approaches to achieve a target quality level. Figure 16 indicates that the impact of individual operation stabilization on the system yield decreases as the operation becomes more stable. It also shows that effect of improving inspection (MTTD) on the system yield decreases. Therefore, it is best to use a combination of both methods to improve quality.

How to increase productivity Improving the stand-alone throughput of each operation and increasing the buffer size are typical ways to increase the production rate of manufacturing systems. If operations are apt to have quality failures, however, there may be other ways to increase the effective production rate: increasing the yield of each operation and conducting more accurate inspections. Stabilizing operations, thus improving the yield of individual operations, will increase the effective throughput of a

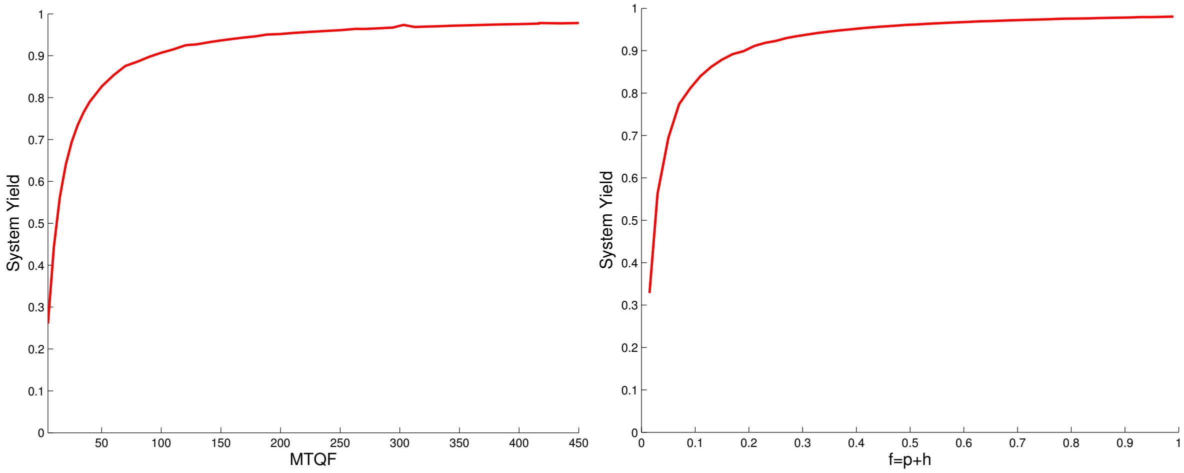


Figure 16: Quality improvement by decreasing the frequency of quality errors, and by decreasing the mean time to detect a quality failure.

manufacturing system regardless of the type of quality failure. On the other hand, reducing the MTTD will *increase* the effective production rate if the quality failure is persistent but it will *decrease* the effective production rate if the quality failure is Bernoulli. This is because the quality of each part is independent of the others when the quality failure is Bernoulli. Therefore, stopping the line does not reduce the number of bad parts in the future.

In a situation in which machines produce defective parts frequently and inspection is poor, increasing inspection reliability is more effective than increasing buffer size to boost the effective production rate. Figure 17 shows this. Also, in other situations in which machines produce defective parts frequently and inspection is reliable, increasing machine stability is more effective than increasing buffer size to enhance effective production rate. Figure 18 shows this phenomenon.

3.2 Effectiveness of the Jidoka Stopping Policy

3.2.1 Motivation

Many tools in the Toyota Production System have been widely used in the design and operation of manufacturing systems in the automotive industry. *Jidoka*, the principle of stopping an operation as soon as a single problem is observed and preventing the production of defective items, is fundamental to the Toyota Production System (Monden 1983). In the Toyota Production System, equipment is designed to detect abnormalities and to stop automatically and immediately whenever they occur. Operators at assembly lines are provided means of stopping the production flow (*andon* cords) whenever they note anything unusual.

Experts in the Toyota Production System argue that the *jidoka* practice has brought several benefits. One is that it motivates *kaizen* (continuous improvement) since operators can clearly see the painful outcome of producing defects: the line stoppage. It is easier to find the root cause of a problem right after the problem takes place. Through the use of systematic ways of resolving problems (e.g. asking “Why?” five times) which are widely accepted in Toyota, operators' learning speed accelerates (Fujimoto 1999). It has been known that operators' learning can significantly improve productivity and

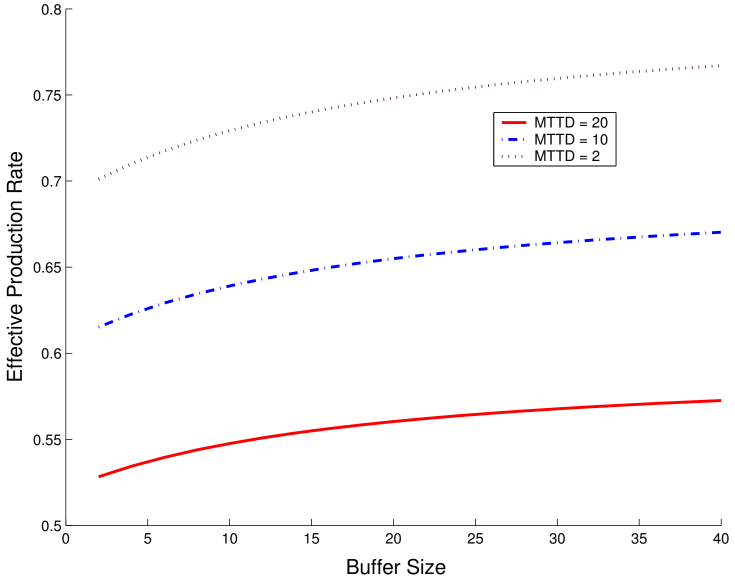


Figure 17: Relative effectiveness of increasing inspection reliability vs. increasing buffer size in improving the production rate: mean time to detect and effective production rate.

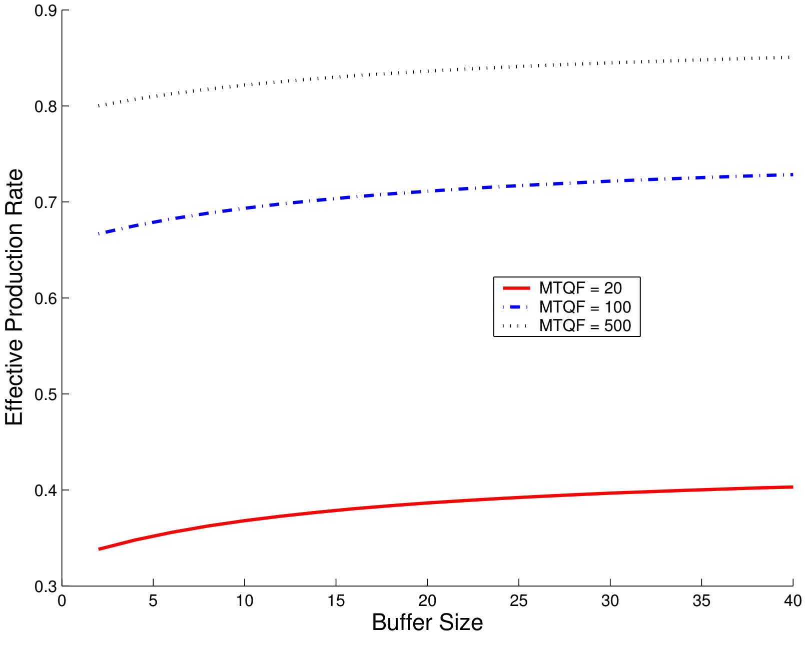


Figure 18: Relative effectiveness of increasing machine reliability (i.e. decreasing quality failure frequency) vs. increasing buffer size in improving the production rate: mean time to detect and effective production rate.

quality (Sandberg 1995).

The other benefit of *jidoka* that Toyota Production System advocates claim is that it prevents the waste that would result from producing a series of defective items. Therefore, *jidoka* is considered to be a means to improve quality and increase productivity at the same time (Toyota Motor Corporation 1996, Monden 1983). It is this aspect of the *jidoka* policy that we investigate here. When quality failures are such that once a bad part is produced, all subsequent parts will be bad until the machine is repaired, and when inspection is very accurate, catching bad parts and stopping the machine as soon as possible is the best way to maintain high quality and productivity. This is the case with breakage of thread, which caused the invention of the *jidoka* practice a century ago at Toyota Spinning & Weaving (Togo and Waterman 1993). On the other hand, when a defect is generated from common cause variations so that the occurrence of a bad part implies nothing about the quality of future parts, there is no benefit to stopping a machine that has made a bad part because there is no reason to believe that future parts will be bad; thus stopping the machine would not reduce the number of bad parts in the future. In this case, therefore, stopping the operation does not improve quality and it reduces productivity by reducing working time.

In reality, most machines have multiple-yield quality failures in a sense that quality failures occur independently but at different rates depending on what state the machine is in. When a machine is in good shape and operating without any assignable cause variations (i.e., when it is in control), it may produce a defective part with a very small probability, not because of any change in the machine but because of random perturbations. However, when a machine is operating under an assignable cause variation (i.e., it is out of control), it is likely that many of the parts that will be produced will be bad. In this situation, the optimal stopping policy is, therefore, to stop the machine only if the machine is out of control. But in many cases it is not easy to tell whether the machine is out of control or in control (in other words, whether a quality failure is from a random variation or an assignable cause variation) with a single sample. Matters get even more complicated when the inspection is not perfectly accurate.

Here, we investigate when stopping a machine after the observation of a single bad part is better than waiting until two consecutive bad parts are seen. We do not determine when *jidoka* is optimal; rather we find conditions under which it is certainly not optimal.

3.2.2 Single Machine

State definition People adopt *jidoka* assuming that inspection is 100% accurate and that all the defects are from persistent quality failures. In that case, it is clear that *jidoka* improves quality and productivity at the same time. But when there are Bernoulli quality failures or multiple-yield quality failures, there is no guarantee that subsequent parts will be defective after finding a defect. In this case, it may be better to stop a machine when the machine produces two defective parts in a row, for example, since it is not likely to have two Bernoulli quality failures consecutively.

We extend the model of Section 3.1.1 to analyze *jidoka*. In this multiple-yield model, the states have the following meanings:

- State 1: The machine is in good shape and operating without any assignable cause variations. It may produce defective parts with probability $1 - \pi(1)$, which is close to 0, due to random variations.

- State -1: The machine is operating under an assignable cause variation and is producing bad parts with probability $1 - \pi(-1)$, which is larger than $1 - \pi(1)$.

- State 0: The machine is not operating.

Therefore, $\pi(1)$ is the yield of a machine when the machine is in state 1 and $\pi(-1)$ is the yield of the machine when it is in state -1. A system that has persistent quality failures is a special case where $\pi(1) = 1$ and $\pi(-1) = 0$.

When the machine is either in state 1 or -1, it can stop for two reasons: operational failures with probability rate p^j and quality failures with probability rate q^j $(j = -1, 1)$. Here, $1/q^j$ is the *Mean Time to Stop due to Quality failures* (MTSQ) which depends on the frequency of quality failures, inspection accuracy, and machine stopping policies. Since we assume that the occurrence of operational failures is independent of machine states, $p^1 = p^{(-1)} = p$.

The state transition diagram is as shown in Figure 10, except that the operational failure rate p is replaced by the rate s. Here, p is still the rate that operational failures occur while the machine is in state 1, but $q_1 = s - p > 0$ is the rate that the inspection station (erroneously) decides that the machine should undergo quality repair when it is in state 1. As earlier, $q_{-1} = h = f - p > 0$ is the rate that the inspection station (correctly) decides that the machine should undergo quality repair when it is in state -1.

Modeling of stopping policies It is not easy to determine whether the stopping policy of *jidoka* is optimal since we would need to compare the performance of all possible stopping policies[1] But proving that it is not optimal for a given case is easy because it can be done by giving an example of a better policy. Therefore, the question that we seek to answer is "Under what conditions would stopping with one defect (*jidoka*) not be optimal?" For the sake of simplicity, we assume that inspection is perfect and we consider two different stopping policies:

- *Policy 1*: Stop a machine when a bad part is produced.
- *Policy 2*: Stop a machine when two bad parts are produced consecutively.

Stopping policy 1 is *jidoka*. If $\pi(1) = 1$ (i.e. if the machine never makes a bad part when it is state 1), stopping a machine immediately after it produces a defect is better than stopping with two consecutive defects. In the case of general multiple-yield quality failures, however, the performance of the two stopping policies depends on many factors.

Modeling of stopping policy 1 The probability of making a bad part when machine i (M_i) is in state j is $1 - \pi_i(j)$. As a result, the MTSQ ($=1/q_i^j$) is the same as the mean time for a quality failure to occur. We have found that

$$q_i^j = \mu_i(1 - \pi_i(j)), \quad j = -1, 1 \tag{18}$$

As a result, the transition rates from state 1 to state 0 and state -1 to state 0 are given by $s_i = p_i + \mu_i(1 - \pi_i(1))$ and $f_i = p_i + \mu_i(1 - \pi_i(-1))$ respectively.

Modeling of stopping policy 2 For stopping policy 2, the MTSQ is the expected time for two quality failures to occur in a row. MTSQ can be estimated by solving the *expected time to absorption problem* (Bertsekas and Tsitsiklis 2002) which is illustrated in Figure 19.

The states in Figure 19 have the following meanings:

[1] Determining the optimal policy is one of the goals of the current research.

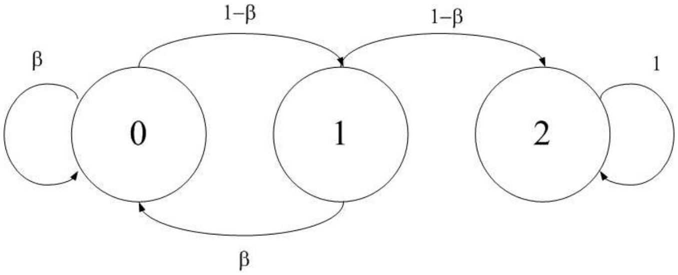


Figure 19: MTSQ estimation through an *expected time to absorption* problem.

- *State 0*: Neither of the last two parts was defective.
- *State 1*: The most recent part was defective and the one before that was good.
- *State 2*: Both of the last two parts were defective.

In the figure, β is the probability of producing a good part and it depends on whether the machine is in state 1 or -1. When two bad parts are produced in a row, the machine is stopped according to stopping policy 2. Therefore, state 2 is an absorbing state. From the value of the mean time to reach state 2, we find

$$q_i^j = \frac{\mu_i(1-\pi_i(j))^2}{2-\pi_i(j)} \tag{19}$$

Therefore,

$$\begin{aligned} s_i &= p_i + \frac{\mu_i(1-\pi_i(1))^2}{2-\pi_i(1)} \\ f_i &= p_i + \frac{\mu_i(1-\pi_i(-1))^2}{2-\pi_i(-1)} \end{aligned} \tag{20}$$

To analyze a single-machine system, we use Equations (1)–(8) with these parameters. We find that the effective production rate, the production rate of good parts, is

$$P_E = \mu[\pi(1)P(1) + \pi(-1)P(-1)] = \mu\frac{\pi(1) + \pi(-1)g/f}{1 + (s+g)/r + g/f} \tag{21}$$

A numerical analysis is described below.

3.2.3 Analysis of a Two-Machine System

The analysis of the state dynamics is exactly like that of the two-machine line of Section 3.1.2, and the total production rate P_T is the same. The only difference is that now the effective production rate is a function of $\pi(1)$ and $\pi(-1)$.

A mathematical model for the two-machine system was solved and validated through comparison with discrete event simulation (Kim 2004). Agreement with simulation was good.

μ	r	p	g	$\pi(1)$	$\pi(-1)$
1	0.2	0.01	0.01	0.997	0

Table 2: Base machine parameters.

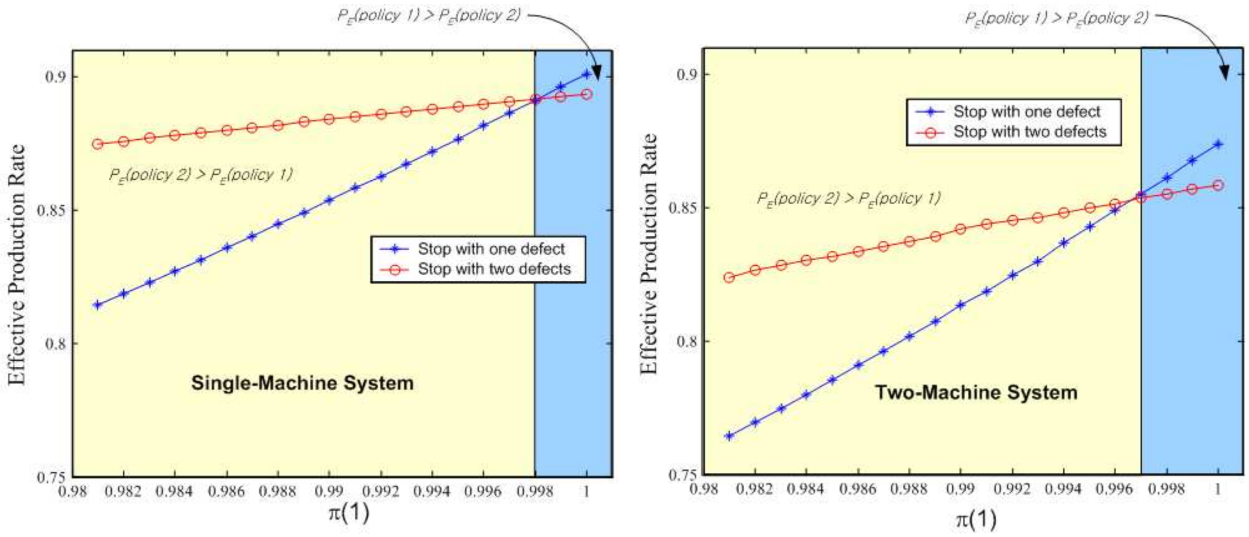


Figure 20: Effectiveness of stopping policies vs. $\pi(1)$.

3.2.4 Effectiveness of Jidoka

The effective production rates of a single-machine system and a two-machine system with two different stopping policies are compared with varying machine parameters to see under what operating conditions stopping with one defect (*jidoka*) is effective.

We should note that the comparison of the two policies does not answer the question "Under what conditions is stopping with one defect optimal?" It only shows some operating conditions in which stopping with one defect is *not* optimal. But this numerical experiment gives a good idea about the conditions under which stopping with one defect would be a reasonable policy. The base machine parameters used for the numerical experiments are shown in Table 2.

Figure 20 shows the effective production rates of a single-machine system and a two-machine system with the two stopping policies by varying $\pi_i(1)$ $(i = 1, 2)$. As the figure indicates, the effectiveness of the stopping policies depends significantly on $\pi_i(1)$. Stopping policy 1 is better than stopping policy 2 only when $\pi_i(1)$ is very close to 1 (i.e. $\pi_i \geq 0.997$ for the two-machine system). This is a case in which a machine seldom produces a defect unless it is operating under an assignable cause variation since there is very little random variation in the operation.

The rate of transitions from the in-control state (state 1) to the out-of-control state (state -1) is given by g. Figure 21 shows the impact of g on the relative performance of the two stopping policies. Note that the influence of g on the the relative performance of each stopping policy seems to be smaller than that of $\pi(1)$.

Stopping policy 1 is better when g is large, since a large g means more frequent transitions to the out-of-control state from the in-control state; thus, when a bad part is detected, it is likely that the machine has been in the out-of-control state.

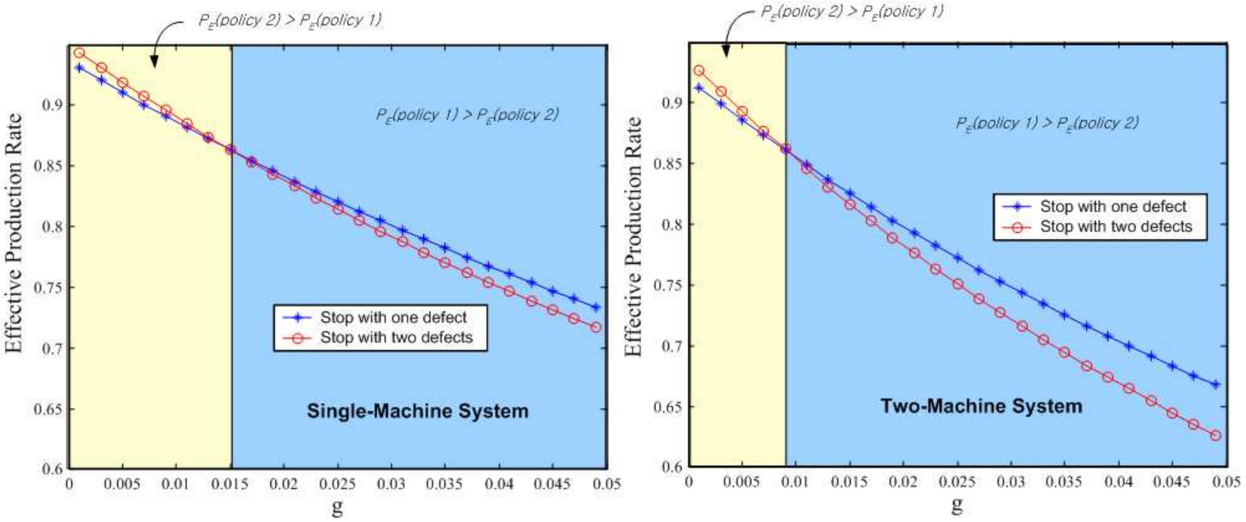


Figure 21: Effectiveness of stopping policies vs. g.

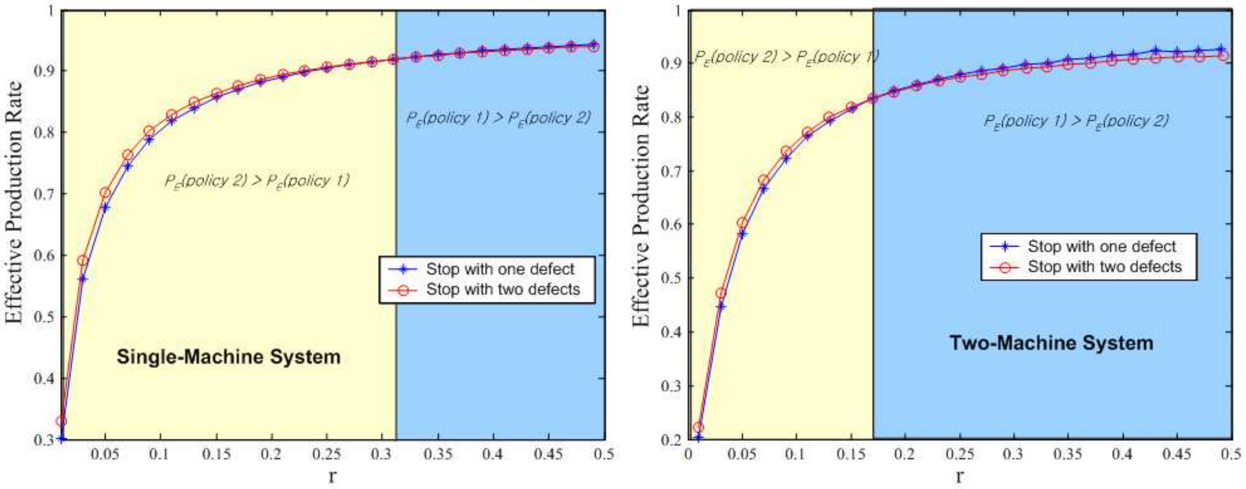


Figure 22: Effectiveness of stopping policies vs. repair rate.

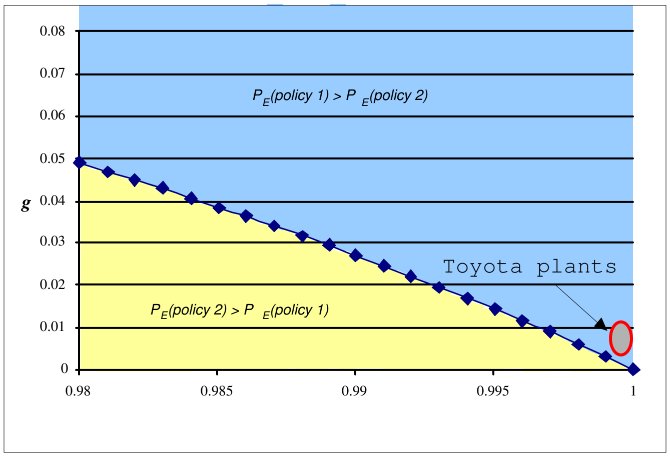


Figure 23: Comparison of stopping policy and operation range of Toyota plants.

Quicker repair means a reduction of capacity loss caused by the inappropriate stoppage of a machine. Therefore, stopping with one defect outperforms stopping with two defects in a row when r is large, which is shown in Figure 22. More numerical experiments show that a higher $\pi_i(1)$ leads to a higher value of r at which the two stopping policies give the same effective production rate. Comparison with Figures 20 and 21 reveals that the impact of r on the relative effectiveness of the two stopping policies is weaker than $\pi(1)$ and g.

Numerical experiments show that the relative performances of a single-machine systems and two-machine systems with the two stopping policies are less sensitive to other machine parameters (p_i, f_i, μ_i, and $\pi_i(-1), i = 1, 2$).

Figure 23 illustrates the domain of $\pi(1)$ and g where 'stopping with one defect' outperforms 'stopping with two defects' in a two-machine system. Two machine parameters $\pi(1)$ and g are used since these are the two major factors that the effective production rate is sensitive to. Standard process capability used at Toyota plants is $C_p = 1.33$, which means that operations have a yield which is more than 99.99% (Monden 1998). If we assume that a typical value of r at factories in the automotive industry is around 0.2, and that g is usually less than 0.01, this operating condition is in the domain where 'stopping with one defect' is better as shown in Figure 23. In other words, it seems that the jidoka stopping policy is close to optimal at Toyota plants under the assumption that inspection is accurate.

Note that all the results described here are based on the assumption of perfect inspection. When inspection is not perfect, stopping with one defect is less likely be optimal since it is not even certain whether the machine actually produced a nonconformity. We conclude that jidoka is likely to be optimal when factories are operating under desirable conditions (e.g. high process capability ($C_p > 1$), infrequent occurrence of assignable causes, and short repair time). However, more research is needed to determine the influence of other realistic factors such as imperfect inspection and the influence of workers' learning curves.

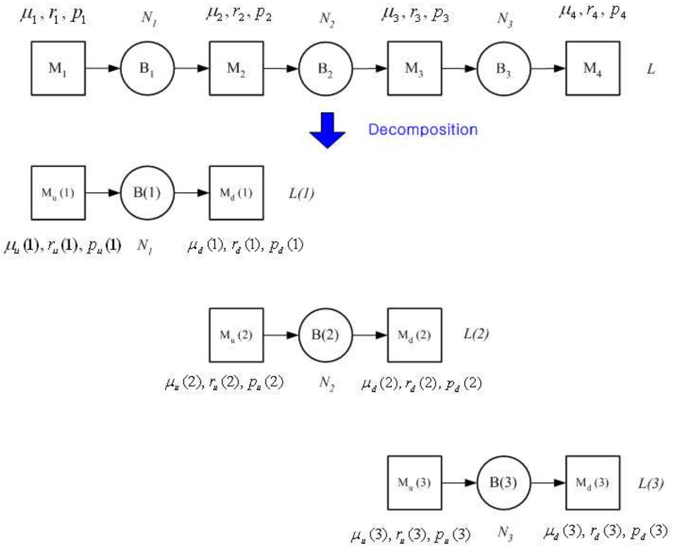


Figure 24: Decomposition of a four-machine line into three two-machine lines.

3.3 Modeling and Analysis of Long Flow Lines with Quality and Operational Failures

3.3.1 Motivation and Literature Review

The analysis of a line with more than two machines is more difficult than the analysis of a two-machine system since it leads to a higher dimensional state space and increases the number of boundary conditions. The *decomposition* technique is a widely used approximation technique that decomposes a k-machine line L into a set of $k-1$ two-machine lines $L(i)$ $(i = 1, 2, ..., k-1)$. Each line $L(i)$ is composed of an upstream machine $M_u(i)$ and a downstream machine $M_d(i)$, separated by a buffer $B(i)$. This decomposition is illustrated in Figure 24 for a four-machine line.

The use of decomposition techniques for the analysis of long flow line was proposed by Zimmern (1956) for machines with operation-dependent failures and by Sevast'yanov (1962) for machines with time-dependent failures. Both authors used the continuous model and considered only the case of homogeneous lines in which all machines have the same repair rates. For the analysis of the discrete model of long homogeneous lines, approximate decomposition equations were proposed by Gershwin (1987). The decomposition equations derived by Gershwin were efficiently solved by the DDX algorithm, which was formulated by Dallery, David, and Xie (1988). (See Gershwin 1994.)

The principle of the decomposition is that the behavior of the material flow in buffer $B(i)$ closely matches that of the flow in buffer B_i of line L. Machine $M_u(i)$ represents the part of the line L upstream of B_i and machine $M_d(i)$ represents the part of the line L downstream from B_i.

The decomposition techniques for continuous-material long lines is more complex than for other models (e.g. the discrete material, discrete deterministic processing time long line) since it allows machines to operate at different speeds, and the speeds of machines are slowed down due to partial blockage and partial starvation. (See Gershwin 1994.) A decomposition technique for a continuous long line with different operation speeds and operation dependent failures was proposed by Glassey and Hong (1993). The Accelerated DDX algorithm (ADDX), which was formulated by Burman (1995), converges faster

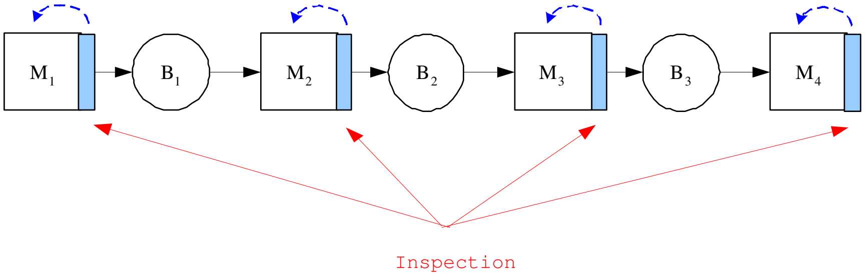


Figure 25: Ubiquitous inspection.

and gives more accurate estimates than Glassey and Hong's algorithm.

3.3.2 Ubiquitous Inspection

Many different kinds of long manufacturing lines with quality and operational failures can be analyzed: different topologies, different quality failures, different inspection policies, and so on. However, there has been no analytical model of these in the literature. Here, we take a first, fundamental research step in building analytic models of long manufacturing lines with quality and operational failures, by focusing on three cases. The analysis of various kinds of long manufacturing lines is a promising topic for future research.

First, we analyze the *ubiquitous inspection case* illustrated in Figure 25. The assumptions of the case are as follows:

- Each machine has both operational failures and quality failures.
- Each operation works on different features. Thus, quality failures at one operation do not influence the quality of other operations.
- The inspection at machine i (M_i) can detect defective features made only by M_i, and not other machines.
- There is no scrapping or reworking in the line; defective parts are marked, and scrapped or reworked later.

Transformation technique Since machines with quality and operational failures have five parameters, μ_i, r_i, p_i, g_i, and f_i, the analysis of a k-machine line with decomposition techniques requires equations for $10(k-1)$ pseudo-machine parameters $(\mu_u(i), r_u(i), p_u(i), g_u(i), f_u(i), \mu_d(i), r_d(i), p_d(i), g_d(i)$ and $f_d(i)$ $(i = 1, 2 \ldots k-1))$ and an efficient algorithm to solve them.

For the ubiquitous inspection case, *quality failures at an operation do not influence the quality of other operations* because of the assumption that each operation works on different features. As a result, g_i is independent of other machines' parameters. Therefore, we have

$$g_u(i) = g_i$$

$$g_d(i) = g_{i+1} \tag{22}$$

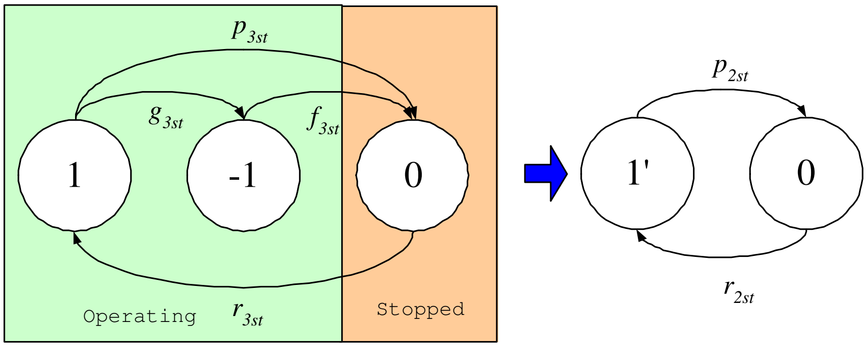


Figure 26: Three-state-machine and corresponding two-state-machine.

Another fundamental assumption of the model is that inspection can only identify bad features made locally. Therefore, for each decomposed line $L(i)$, $(i = 1, 2, ...k-1)$ the incoming parts from upstream machines are treated as non-defective since the inspections at the decomposed line $L(i)$ cannot detect defective parts from the upstream machines. In addition, outgoing defective parts from $L(i)$ are not detected by the inspections at downstream machines. Thus, f_i is also independent of other machines' parameters. Therefore, we have

$$f_u(i) = f_i$$

$$f_d(i) = f_{i+1} \tag{23}$$

Equations (22) and (23) are a set of $4(k-1)$ equations. The remaining $6(k-1)$ equations are needed for the determination of $\mu_u(i)$, $\mu_d(i)$, $r_u(i)$, $r_d(i)$, $p_u(i)$, and $p_d(i)$, which are the parameters for machines with operational failures only. To determine these parameters, we propose that three-state-machines (state 1, state -1, and state 0) can be approximated by two-state-machines (state 1' and state 0), as depicted in Figure 26.

In Figure 26, two up states (state 1 and state -1) of the three-state-machine are consolidated into one up state (state 1') of the two-state-machine. (Kim 2004) gives simulation results in which two-state machines had similar down-time behavior to that of three-state machines; and where two-machine-one-buffer lines with two-state-machines (chosen as described in the following) had very similar performance measures as lines with three-state machines.

For a three-state-machine in isolation, the probability of a machine being at each state is

$$P_{3st}(1) = \frac{1}{1 + (p_{3st} + g_{3st})/r_{3st} + g_{3st}/f_{3st}}$$

$$P_{3st}(0) = \frac{(p_{3st} + g_{3st})/r_{3st}}{1 + (p_{3st} + g_{3st})/r_{3st} + g_{3st}/f_{3st}}$$

$$P_{3st}(-1) = \frac{g_{3st}/f_{3st}}{1 + (p_{3st} + g_{3st})/r_{3st} + g_{3st}/f_{3st}} \tag{24}$$

On the other hand, for a two-state-machine in isolation, the probability of a machine being at each state is

$$P_{2st}(1') = \frac{r_{2st}}{p_{2st} + r_{2st}}$$

$$P_{2st}(0) = \frac{p_{2st}}{p_{2st} + r_{2st}} \tag{25}$$

The probability of state 1' of the two-state-machine is the sum of the probability of state 1 and state -1 of the three-state-machine. Therefore,

$$P_{3st}(1) + P_{3st}(-1) = P_{2st}(1') \tag{26}$$

From Equations (24), (25), and (26), we have

$$p_{2st} = \frac{f_{3st}(p_{3st} + g_{3st})}{f_{3st} + g_{3st}} \tag{27}$$

As a result, the three-state machine can be approximated by the two-state machine with machine parameters $\mu_{2st} = \mu_{3st}$, $r_{3st} = r_{2st}$, and Equation (27).

The equivalent two-state-machine gives the total production rate and average inventory. But the effective production rate should be estimated indirectly since the two-state-machine can not tell the difference between 'good' state and 'bad' state. Since there is no scrapping in the system, the yield of a machine is

$$\frac{P_{3st}(1)}{P_{3st}(1) + P_{3st}(-1)} = \frac{f_{3st}}{f_{3st} + g_{3st}}$$

For multiple machine lines, the system yield is a product of the individual yields. Thus, the effective production rate can be calculated by multiplying the system yield by the total production rate.

Analysis procedure The four-machine ubiquitous inspection case, presented in Figure 25, can be analyzed by using the following procedure:

- *Step 1*: Calculate the system yield

$$Y_{sys} = \left(\frac{f_1}{f_1 + g_1}\right)\left(\frac{f_2}{f_2 + g_2}\right)\left(\frac{f_3}{f_3 + g_3}\right)\left(\frac{f_4}{f_4 + g_4}\right)$$

- *Step 2*: Transform the original line L with 3-state-machines into an equivalent line L' with 2-state-machines by setting
  - ⋆ $\mu'_i = \mu_i$, $r'_i = r_i$, $p'_i = \frac{f_i(p_i + g_i)}{f_i + g_i}$ $(i = 1, 2, 3, 4)$.
  - ⋆ $N'_i = N_i$ $(i = 1, 2, 3)$.

- *Step 3*: Calculate the total production rate and average inventory levels for B_i $(i = 1, 2, 3)$ of the 2-state-machines line L' from the ADDX algorithm.

- *Step 4*: Evaluate the effective production rate by multiplying the system yield by the total production rate.

The same procedure can be used for the analysis of a general k-machine line.

Performance evaluation Fifty cases were generated, and numerical results from this procedure were compared with simulation. The errors are displayed in Table 3.

	P_T	P_E	WIP	Inv_1	Inv_2	Inv_3
Average absolute error (%)	0.37	0.64	3.41	5.46	4.51	2.34

Table 3: Average absolute errors in the long line analysis with ubiquitous inspection.

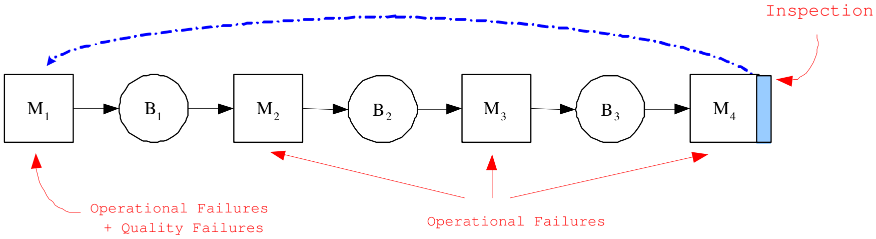


Figure 27: Remote Quality Information Feedback.

3.3.3 Remote Quality Information Feedback

Introduction The second long manufacturing line analysis case is a line with four machines and remote Quality Information Feedback, as illustrated in Figure 27. This is an extension of the two-machine-one-buffer (2M1B) quality information feedback model to a longer line (Kim 2004). It is an approximation of a real situation where operations in the manufacturing line are reliable in terms of quality, whereas incoming raw material causes major quality problems, and the defect in the raw material can only be identified at the end of line.

The assumptions of the remote QIF model are as follows:

- The first machine (M_1) has both operational failures and quality failures.
- The other machines have only operational failures.
- The only inspection is located at the end of line and it can detect defects made by M_1.

Solution method The four-machine remote QIF case is an extension of a 2M1B line with quality information feedback. Therefore, we can use a similar procedure to that used for 2M1B with quality information feedback: adjustment of transition probability rate of M_1 from state -1 to state 0 (i.e., adjusting f_1, which is referred to as f_1^q when it is adjusted) as explained in Kim (2004). After f_1 is adjusted, remote QIF can be treated like ubiquitous inspection so that a similar solution method is used.

The four-machine remote QIF case shown in Figure 27 can be analyzed by using the following procedure:

- *Step 1*: Estimate WIP ($=Inv_1 + Inv_2 + Inv_3$) to get an initial estimate of f_1^q.
- *Step 2*: Adjust f_1^q by using the quality information feedback method described in Section 3.1.6 and Kim (2004). The average inventory that is used in the calculation of f_1 is $Inv_1+Inv_2+Inv_3$.
- *Step 3*: Calculate the system yield

$$Y_{sys} = \left(\frac{f_1}{f_1+g_1}\right)\left(\frac{f_2}{f_2+g_2}\right)\left(\frac{f_3}{f_3+g_3}\right)\left(\frac{f_4}{f_4+g_4}\right).$$

	P_T	P_E	WIP	Inv_1	Inv_2	Inv_3
Average absolute error (%)	0.52	1.02	3.47	5.16	5.56	2.43

Table 4: Average absolute errors in long line analysis with remote QIF.

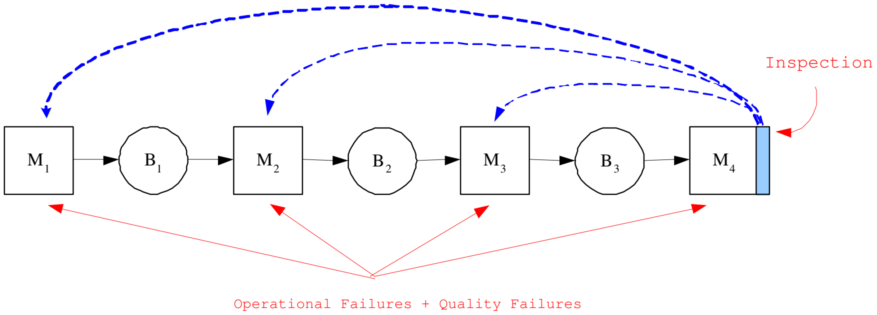


Figure 28: Multiple Quality Information Feedback.

- *Step 4*: Transform the original line L with 3-state-machines into an equivalent line L' with 2-state-machines by setting
  - $\star$ $\mu_i' = \mu_i$, $r_i' = r_i$ $(i = 1, 2, 3, 4)$, $p_1' = \frac{f_1^q(p_1 + g_1)}{f_1^q + g_1}$, $p_i' = \frac{f_i(p_i + g_i)}{f_i + g_i}$ $(i = 2, 3, 4)$
  - $\star$ $N_i' = N_i$ $(i = 1, 2, 3)$.

- *Step 5*: Use the ADDX algorithm to calculate the total production rate (P_T) and average inventory at each buffer B_i $(i = 1, 2, 3)$.

- *Step 6*: Estimate the effective production rate (P_E) by multiplying the total production rate by the system yield.

- *Step 7*: If new P_T, P_E, Inv_i are close enough to their previous values, then stop. Otherwise go to *Step 2* and repeat the procedure.

Performance evaluation Fifty cases were generated, and numerical results from this procedure were compared with simulation. The errors are displayed in Table 4.

3.3.4 Multiple Quality Information Feedback

Due to the cost of inspection stations, factories are often designed in a way that multiple inspections are performed at a small number of stations. The inspection stations are usually located at the end of a line to guarantee that outgoing parts are defect-free. This is a typical example of a multiple Quality Information Feedback case, which is illustrated in Figure 28. The assumptions of the model are as follows:

- All the machines have both operational failures and quality failures.

- The only inspection is located at the end of line, and it can detect defects made by any of the machines (M_1, M_2, M_3, and M_4).

	P_T	P_E	WIP	Inv_1	Inv_2	Inv_3
Average absolute error (%)	0.55	1.95	8.62	5.97	4.75	2.76

Table 5: Average absolute errors in long line analysis with multiple QIF.

- Each operation works on different features. Quality failures at a given operation do not influence the quality of other operations.
- There is no scrapping in the line; defective parts are marked and reworked later.

Solution method The four-machine multiple QIF case shown in Figure 28 is an extension of the remote QIF case in the sense that multiple Quality Information Feedback loops exist. Therefore, we can repeat the same procedure that is used for the remote QIF case for each of the loop.

The four-machine multiple QIF case can be analyzed by using the following procedure:

- *Step 1*: Estimate the average inventory of each buffer (Inv_1, Inv_2, and Inv_3).
- *Step 2*: Adjust f_i^q $(i = 1, 2, 3)$ by using the procedure summarized in Section 3.1.6 and described in detail in Kim (2004). In this case, f_3^q is a function of Inv_3; f_2^q is a function of $Inv_2 + Inv_3$; and f_1^q is a function of $Inv_1 + Inv_2 + Inv_3$;
- *Step 3*: Calculate the system yield

$$Y_{sys} = \left(\frac{f_1}{f_1 + g_1}\right)\left(\frac{f_2}{f_2 + g_2}\right)\left(\frac{f_3}{f_3 + g_3}\right)\left(\frac{f_4}{f_4 + g_4}\right).$$

- *Step 4*: Transform the original line L with 3-state-machines into an equivalent line L' with 2-state-machines by setting
  - ⋆ $\mu_i' = \mu_i$, $r_i' = r_i$, $p_i' = \dfrac{f_i^q(p_i + g_i)}{f_i^q + g_i}$ $(i = 1, 2, 3, 4)$.
  - ⋆ $N_i' = N_i$ $(i = 1, 2, 3)$.
- *Step 5*: Use the ADDX algorithm to calculate the total production rate (P_T) and average inventory at each buffer B_i $(i = 1, 2, 3)$.
- *Step 6*: Estimate the effective production rate (P_E) by multiplying the total production rate by the system yield.
- *Step 7*: If the new P_T, P_E, Inv are close enough to their previous values, then stop. Otherwise go to *Step 2* and repeat the procedure.

Performance evaluation Fifty cases were generated, and numerical results from this procedure were compared with simulation. The errors are displayed in Table 5.

4 Simulation Studies

This section summarizes the results of extensive simulation experiments designed to provide qualitative insight into the relationship between quality, quantity, and system design parameters. Because these results were obtained by simulation, they are not restricted to the limited classes of systems described in Section 3. Furthermore, they are based on a model that represents a larger and more realistic set of phenomena, including machines with multiple down states, inaccurate inspection, and non-linear topologies.

In particular, it addresses such issues as the influence of the number and placement of inspection stations; of machine and buffer parameters such as yield and capacity; of inspection characteristics such as false negative and false positive probabilities; of control policies such as scrapping and information feedback; and finally of production system topology, on system performance measures.

The results presented here are far from exhaustive. They are intended to highlight some of the more interesting and important conclusions that these simulations have permitted us to draw.

4.1 Location and Number of Inspection Stations

It is reasonable to believe that a larger number of inspection stations is better than a smaller number. However, it is important to qualify this statement. Certain performance measures depend crucially on the locations of the inspection stations, and it is even possible to do worse by adding inspection stations, if they are poorly placed.

Inspection stations affect performance in ways that are sometimes contradictory. For example, if one is only allowed a small number of inspection station (due to cost, floor space, or other restrictions), then the stations are necessarily widely interspersed throughout the production system, which means that the distance between a machine and the station that inspects its work can be large. (Note that "distance" is measured both in terms of physical stages—machines and buffers—and parts—buffer capacity. For example, if a large FIFO buffer with a high mean level is placed between a machine and the station that inspects its work, parts processed by the machine will generally take a long time to pass through the buffer and reach the inspection station.) This means that a state change that causes the machine to begin producing many defective parts will be detected after a significant delay, and many defective parts will have been produced during that time interval. Consequently, inspection will not be as effective in preventing the production of defective parts as it would have been if the station were closer to the machine. This suggests that a larger number of inspection stations might be preferable in some cases.

On the other hand, when an inspection station declares a machine to be in need of quality maintenance and sends it information feedback, the machine is taken down; a machine that is taken down for servicing will not produce parts for a certain amount of time. As a result, information feedback will reduce the total production rate. Dead periods (i.e. periods of time when information feedback is suspended to allow defective parts already in the system to drain out) reduce the opportunity for an inspection station to take a machine down for servicing, and thus tend to increase the total production rate. (Whether or not the good production rate changes in the appropriate direction depends on system parameters—see Section 4.2.) This suggests that longer dead periods (i.e. fewer inspection stations and a larger distance between a machine and the station that inspects its work) might sometimes be preferable.

For these reasons, performance measures such as production rate are not monotonic functions of the number of inspection stations. In particular, a larger number of inspection stations can be harmful if the

Number of stations	Number of sets of possible locations	Best good production rate
0		0.100400
1	1	0.299664
2	14	0.330765
3	91	0.345702
4	364	0.353942
5	1001	0.359044
6	2002	0.363403
7	3003	0.365687
8	3432	0.367533
9	3003	0.368477
10	2002	0.369800
11	1001	0.370766
12	364	0.371974
13	91	0.372056
14	14	0.372437
15	1	0.372901

Table 6: Number of combinations of sets of locations for each number of inspection stations, and best achievable good production rate.

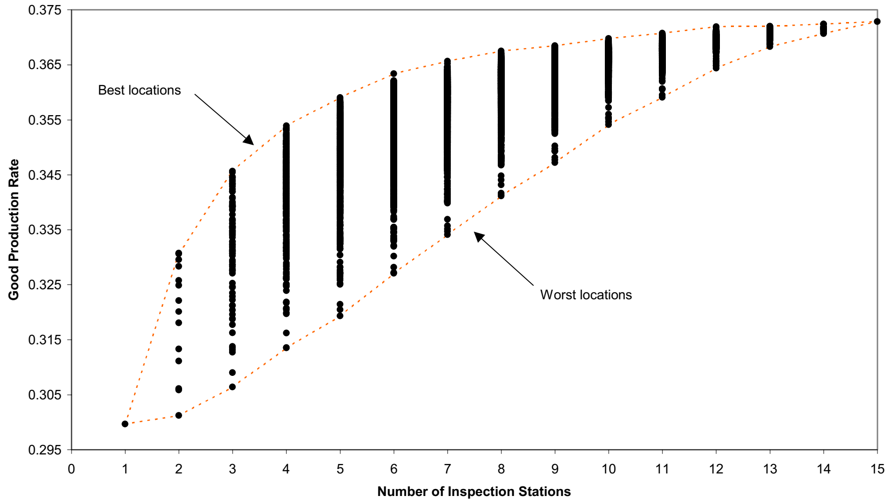


Figure 29: Effect of the number and locations of inspection stations on the good production rate, for a production line composed of 15 five-state machines, 14 buffers, and $1, 2, ..., 15$ inspection stations.

stations are clustered together so that some machines are too frequently taken down for maintenance while others produce many defective parts before they receive information feedback. This is evident in Figure 29, where the good production rate of a 15-machine line is plotted against the number of inspection stations. For each number of stations, there are potentially many combinations of locations to place them. Bearing in mind the fact that one of the inspection stations is always located after the 15th machine, i.e. at the end of the production line, if there are k inspection stations altogether, then there are "14 choose $k-1$" possible combinations of locations for the remaining $k-1$. This quantity is given by:

$$\begin{pmatrix} 14 \\ k-1 \end{pmatrix} = \frac{14!}{(k-1)!(15-k)!}$$

The number of sets of locations corresponding to each number of inspection stations appears in Table 6, together with the best achievable good production rate for that number of inspection stations. Each dot in Figure 29 represents a particular combination of locations. The dotted line at the top goes through the best combination of locations for each number of inspection stations, and the dotted line at the bottom goes through the worst. Although both lines would appear to be monotonically increasing, it is important to note that many combinations of locations for some higher numbers of inspection stations lead to worse good production rates than the best combination of locations for lower numbers. For example, the highest good production rate that can be achieved with 3 stations is 0.346 ppc (parts per cycle). Although the highest good production rate that can be achieved with 4 stations is 0.354 ppc, for 234 of the 364 combinations of locations—or 64%—the good production rate is actually *lower* than the best achievable rate with 3 inspection stations. For 5 inspection stations, 279 of the 1001 combinations of locations—or 28%—provide a good production rate below the best achievable rate with 3 inspection stations.

Although the range between best and worst will of course change depending on machine and buffer parameters, the fact is that a higher number of inspection stations does not always give better results. The locations of inspection stations are as important as their number.

4.2 Control Policies

Two types of action (control policies) were analyzed: scrapping (which is an action on parts) and information feedback (which is an action on machines). Representative results for both policies are presented below.

4.2.1 Scrapping

The principal benefit of scrapping locally is that defective parts are removed from the system, freeing up downstream machine and buffer capacity. Furthermore, lower buffer levels downstream allow faster response to future quality failures—not only for the machine that made the bad parts that were scrapped, but for all downstream machines. As a result, good parts are produced at a higher rate, lead times are shorter, and buffer levels (a measure of in-process inventory) are lower.

Figure 30 illustrates the effect of local scrapping on the good production rate of a production line composed of 5 three-state machines. There are two inspection stations, one of which always immediately follows the last machine; the other can be placed after either of the remaining four machines. The rightmost bar corresponds to no scrapping, providing a good production rate of 0.605 ppc. Counting from the left, the first bar corresponds to locating the free inspection station after the first machine, with

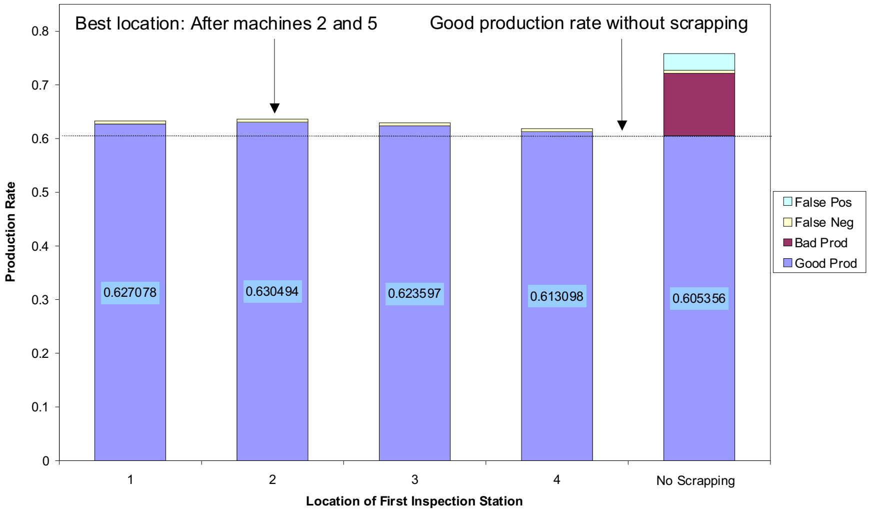


Figure 30: Effect of local scrapping on the good production rate of a production line composed of 5 three-state machines and two inspection stations. The rightmost bar corresponds to no scrapping. The four bars to its left correspond to locating one of the inspections station at machine $1, 2, 3$, or 4. (The other is at machine 5.)

a good production rate of 0.627 ppc. The second bar corresponds to locating the free inspection station after the second machine, and so on. The best location (in terms of maximizing good production rate) is achieved when the free inspection station is located after the second machine, providing 0.630 ppc. Note that the bad production rate and the waste (false positive) rate are both zero when there is scrapping. This is because the production rates only count the parts that emerge from the last machine in the line; parts that are known or thought to be bad are scrapped and therefore do not appear in these bars. By comparison, the scrap rates corresponding to the first four bars vary between 0.062 ppc and 0.135 ppc.

Figure 31 illustrates the effect of local scrapping on the production lead time of the same line. As before, there are two inspection stations, one of which always immediately follows the last machine. The production lead time when there is no scrapping is 157.3 cycles. (Note that the capacities of the four buffers are each 60; travel time through these buffers accounts for the significant cycle time.) The best production lead time is achieved when the free inspection station is located after the third machine, with 127.4 cycles. Much of this reduction is due to the drop in buffer utilization that results from scrapping, as we shall discuss next. Note that the best good production rate and the best production lead time are not necessarily achieved by the same inspection station location; this highlights the importance of choosing the optimization criterion carefully.

Figure 32 illustrates the effect of local scrapping on the mean buffer levels in a production line composed of 15 three-state machines. There are three inspection stations, located after machines 5, 10, and 15 (and before buffers 5, 10, and 15). The clear bars correspond to no scrapping, and the solid bars correspond to local scrapping at the inspection stations. When there is no scrapping, since the machines and buffers in this example are identical, the mean buffer levels decrease monotonically as one moves from the first buffer to the last one. This well-recognized behavior is due to the fact that a long sequence of identical machines and buffers is less efficient than a similar but shorter sequence.

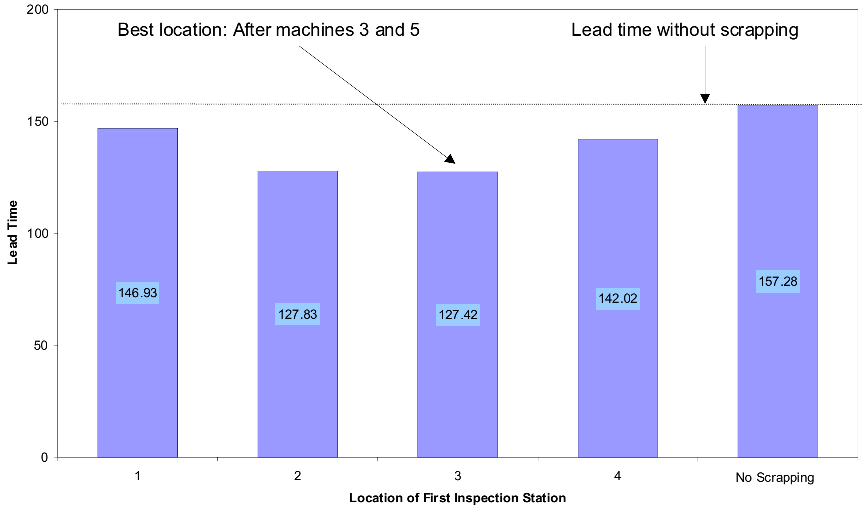


Figure 31: Effect of local scrapping on the production lead time of a line composed of 5 three-state machines and two inspection stations. The rightmost bar corresponds to no scrapping. The four bars to its left correspond to locating one of the inspections station at machine $1, 2, 3$, or 4. (The other is at machine 5.)

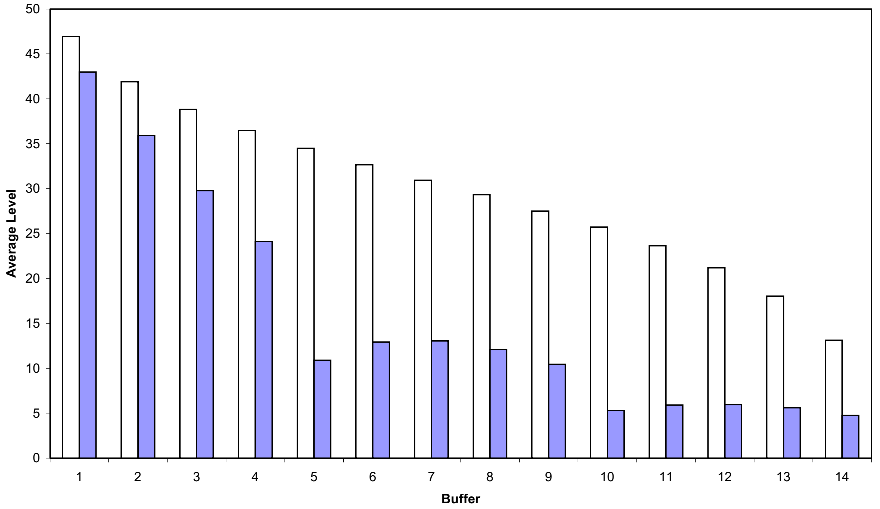


Figure 32: Effect of local scrapping on the mean buffer levels in a line composed of 15 three-state machines and three inspection stations. The clear bars correspond to no scrapping, and the solid bars correspond to local scrapping at the inspection stations which are located after machines 5, 10, and 15.

For example, the sub-line composed of machines 1–5 (and the buffers between them) is more efficient than the sub-line of machines 1–6; by contrast, the sub-line of machines 6–15 is less efficient than the sub-line of machines 7–15. Consequently, the mean level of buffer 6 is higher than the mean level of buffer 7. Scrapping results in the reduction of mean buffer levels across the board, because parts are physically removed from the line. This reduction is most pronounced in the buffer immdiately following the inspection station where scrapping occurs. It is interesting to note, however, that mean buffer levels are not monotonically decreasing immediately after an inspection station. This behavior is consistent with what we have observed in production lines without inspection where there are bottlenecks due to cycle time (i.e. machines with longer cycle times than their neighbors). Although the intuition underlying it eludes us for now, the fact that both scrapping and a cycle time bottleneck result in the same behavior is satisfying, since, from the point of view of buffers downstream of an inspection station where scrapping is performed, the mean rate at which parts emerge from the station is reduced when the scrap rate increases.

4.2.2 Quality Information Feedback

The purpose of Quality Information Feedback is to take down for maintenance a machine that has switched to a state where defective parts are produced at a higher rate than usual. In the realistic situation where a machine cannot spontaneously switch from a low-yield to a high-yield state (since, for example, a worn tool cannot repair itself), this is the only way to ensure that good parts are produced at an acceptable rate. Indeed, if a machine has a zero-yield state, the steady-state good production rate in the absence of information feedback is exactly zero. Thus, information feedback would be essential in that case.

In general, information feedback is most beneficial under the following conditions:

- The difference between the yield in normal performance (e.g. the high-yield state) and impaired performance (e.g. the low-yield state) is significant. If this were not the case, there would be little to be gained from taking the machine down for quality maintenance. Indeed, the loss in productivity due to the forced down-time would outweigh the gain in yield achieved by the quality repair. By contrast, if the yield is much lower when the machine is impaired than in normal operation, then although the loss in productivity incurred by the forced-down time would cause a drop in total production rate, the improved yield achieved in normal performance would lead to a net increase in the good production rate.

- Mean buffer levels are low. If this were not the case, then an inspection station that is not adjacent to the machine whose work it is monitoring would only receive evidence that a state change has occurred at the machine after a considerable delay, since the first defective part produced by the machine would have to make its way through a large quantity of in-process inventory. As a result of this delay, the machine would produce many defective parts before it receives information feedback from the inspection station and is taken down for quality maintenance. As a result, the improvement in good production rate due to information feedback would be minimal. There are two reasons that cause the mean level of a buffer to be low: the buffer capacity may be low, and/or upstream machines may be less efficient than downstream machines.

To illustrate the second point, we note the well-known fact that production rate is a monotonically increasing function of buffer capacity when information feedback is not present. When information feedback is used, however, this is not necessarily the case: good production rate sometimes decreases

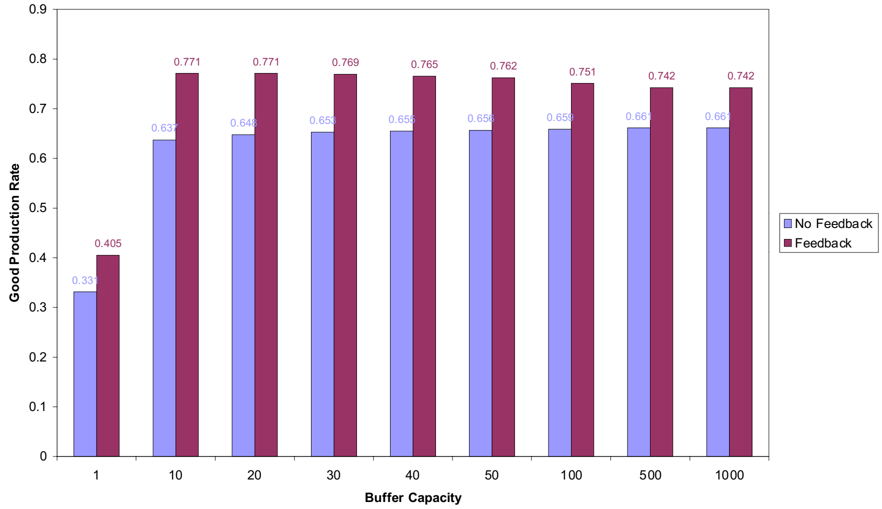


Figure 33: Good production rate as a function of buffer capacity in a two-machine production line, with and without information feedback. Production rate is a monotonically increasing function of buffer capacity when information feedback is not used, but not when information feedback is present.

monotonically, and sometimes it reaches a maximum and then begins to decrease. Figure 33 shows good production rate as a function of buffer capacity in a two-machine production line, and illustrates the latter case. The left-hand bars correspond to good production rate when information feedback is not present, and they increase monotonically and asymptotically. By contrast, the right-hand bars correspond to good production rate when information feedback is present; in this case, a maximum is reached in the neighborhood of a buffer capacity of 10–20, after which the good production rate decreases asymptotically. (Compare this with the similar results obtained analytically in Section 3.1.7.)

When an inspection station declares a machine to be in need of quality maintenance and sends it information feedback, the machine is taken down. A machine that is taken down for servicing will not produce parts for a certain amount of time. As a result, information feedback will reduce the total production rate. However, as long as the inspection process is reasonably accurate, machines will generally be taken down for good cause, namely because they have entered states where they produce more defective parts than is desirable. Thus, taking the machine down will reduce the bad production rate; moreover, since the machine will come back up in a good (high-yield) state after servicing, the good production rate will rise. This is illustrated in Figure 34, which shows the good and bad production rates for a seven-state machine in isolation (i.e. not embedded in a production system) in two situations: information feedback turned off and on. Going from no inspection to inspection with information feedback, the total production rate drops from 0.98 ppc to 0.95 ppc; however, the good production rate increases from 0.73 ppc to 0.92 ppc. The bad production rate drops from 0.24 ppc to 0.01 ppc. (The difference between the sum of good and bad and the total is due to miss and waste rates; there is no scrapping in this example.)

Figure 35 illustrates the effect of information feedback on the mean buffer levels in a line composed of 15 three-state machines and three inspection stations located after machines 5, 10, and 15 (and before buffers 5, 10, and 15). The clear bars correspond to no information feedback, and the solid

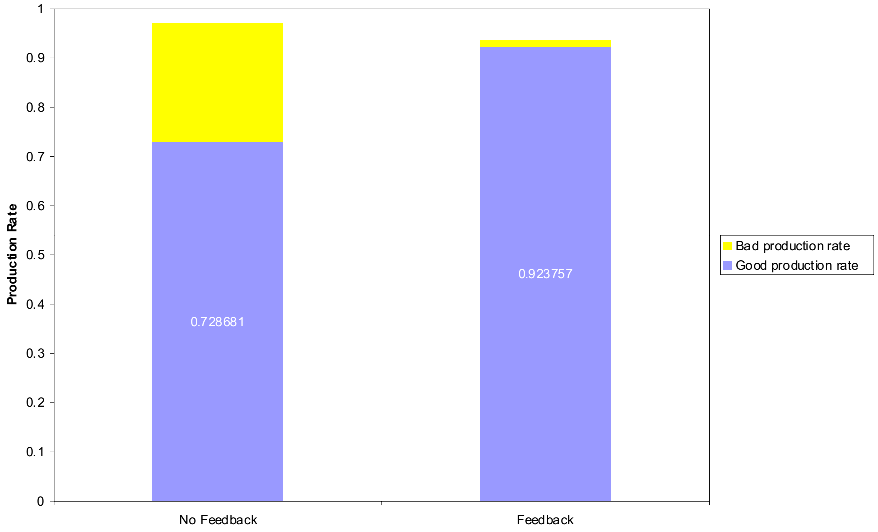


Figure 34: The good and bad production rates for a seven-state machine in isolation with and without information feedback.

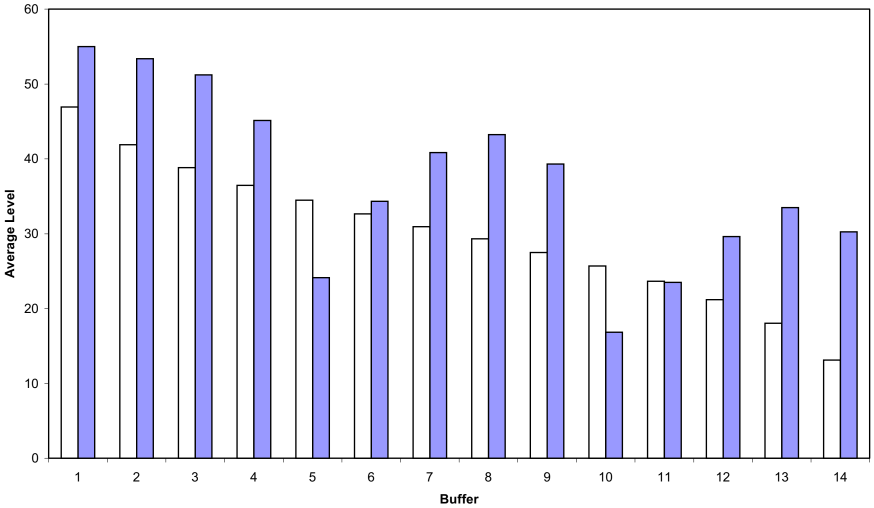


Figure 35: Effect of information feedback on the mean buffer levels in a line composed of 15 three-state machines and three inspection stations. The clear bars correspond to no information feedback, and the solid bars correspond to information feedback from the inspection stations which are located after machines 5, 10, and 15.

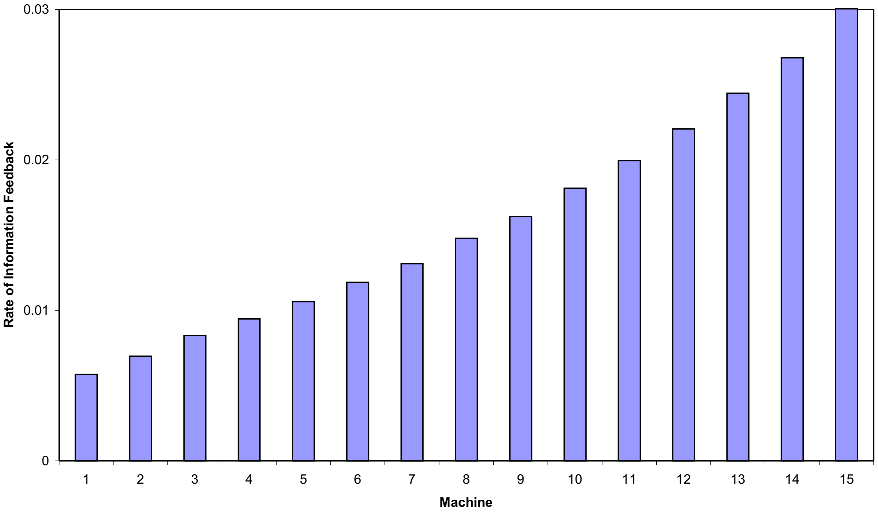


Figure 36: Rate at which information feedback is received by each machine, in a line composed of 15 three-state machines and one inspection station at the very end.

bars correspond to information feedback. In the absence of information feedback, since the machines and buffers in this example are identical, the mean buffer levels decrease monotonically as one moves from the first buffer to the last one, as seen in Figure 32. We see in Figure 35 that the main effect of information feedback is to increase the mean level of most of the buffers. To explain this behavior, note first that even machines that are identical in every respect are not taken down for quality maintenance at the same rate. This is because of the dead periods: all things being equal, the further a machine is from the inspection station that monitors its work, the longer the dead periods typically become. As a result, machines far from an inspection station are less affected by information feedback than machines near the station.

Figure 36 shows the rate at which information feedback is received by each machine, measured as signals per cycle. For clarity, there is only one inspection station, after the last machine. The information feedback rate decreases monotonically with distance to the inspection station. This means that the performance hit (in terms of reduced production rate, not quality) taken by the machines decreases as they get further from the inspection station. Since this effect is not symmetrical (each inspection station only looks at machines *upstream* of it), the net result is that the machines in each inspection domain become less efficient (but higher-yield) as we move downstream within the domain.

This explains why mean buffer levels rise virtually across the board. The few buffers whose mean level decreases in the presence of information feedback generally immediately follow inspection stations. In these cases, the drop is due to the difference between the high rate at which the immediately upstream machine receives information feedback from the adjacent inspection station, and the much lower rate at which the immediately downstream machine receives information feedback from the distant downstream inspection station. (The examples in this section do not feature scrapping.)

4.3 Production System Topology

Many manufacturing systems have topologies more complex than tandem lines, particularly due to current pressures to modularize and outsource. For this reason, we modeled and analyzed production systems composed of a short main line with multiple feeder lines.

Figures 37–40 show the specific topologies we compared. Figure 37 is a linear production system, in which segments that can be separated due to the ability to modularize are coded with different colors. The line segments composed of machines 1–5, 6–10, and 11-15 are each assumed to be independent of the others—i.e. the operations of machines 6–10 can be performed on a given part whether or not those of machines 1–5 have already been performed on that part, etc. It is also assumed that these segments can be interconnected at certain machines, namely machines 6, 11, and 16. Thus, in Figure 37, segment 1–5 is connected to machine 6, segment 6–10 to machine 11, and segment 11–15 to machine 16. In Figure 38, segments 1–5 and 6–10 are both connected to machine 11, and segment 11–15 to machine 16. In Figure 39, segments 1–5 and 11–15 are both connected to machine 16, while segment 6–10 is connected to machine 11. Finally, in Figure 40, segments 1–5, 6–10, and 11–15 are all connected to machine 16.

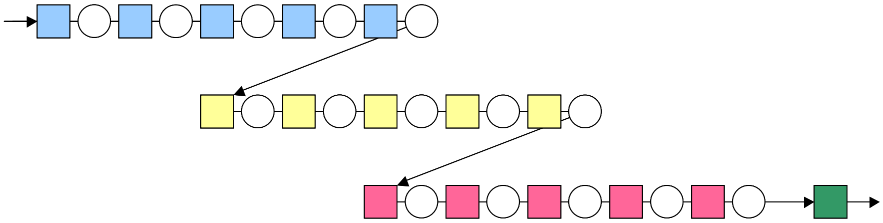


Figure 37: Linear topology. The potential for modularization is indicated by different colors.

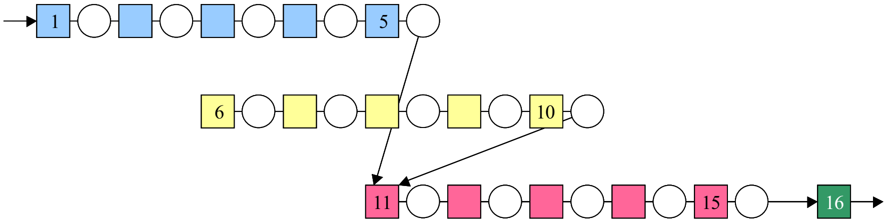


Figure 38: Non-linear topology. Further potential for modularization is indicated by different colors.

It was first assumed that there are no quality failures at all. For each topology, production rate, in-process inventory, and production lead time (defined as the longest amount of time any portion of a finished part spent in the system) were compared, given the following scenarios:

- All machines and buffers are identical;

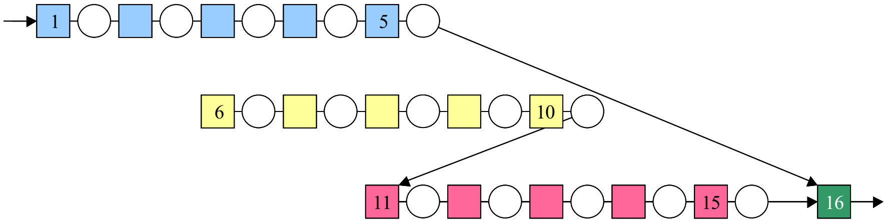


Figure 39: Non-linear topology. Further potential for modularization is indicated by different colors.

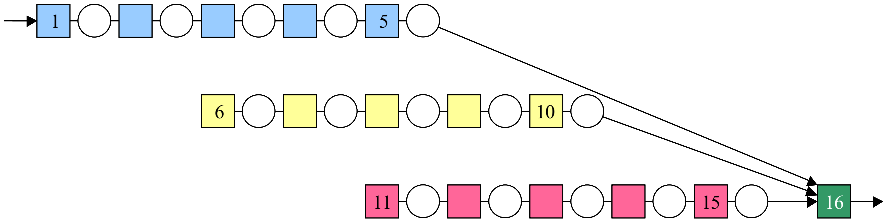


Figure 40: Star topology.

- All machines and buffers are identical, except for a bottleneck at machine 6;
- All machines and buffers are identical, except for a bottleneck at machine 5.

The difference between the two cases involving bottlenecks is that in the first, the bottleneck is at the beginning of a feeder line, whereas in the second, it is at the end. In both cases, the bottleneck was severe enough that upstream buffers were usually full and downstream buffers were usually empty.

For any given scenario, the production rate was found to be highly insensitive to topology. However, in-process inventory and production lead time were very sensitive.

No bottleneck Mean buffer levels decrease monotonically along a production line when all machines and buffers are identical. The longer the line, the more pronounced is the decrease. Thus, for the linear topology in Figure 37, mean buffer levels decreased by 50% from the first to the fourteenth buffer. By contrast, for the star topology in Figure 40, the reduction was less than 10%. Note that the contents of different buffers can have very different economic values; thus, the reduction in the *value* of in-process inventory may be even larger, particularly if downstream buffers contain more valuable inventory than upstream buffers.

Production lead time is a function of both line length and in-process inventory, since in FIFO buffers, a part spends as much time in a buffer as there are parts ahead of it. In the absence of a bottleneck, and for the system parameters utilized, the influence of line length was found to be more important than mean buffer levels. Thus, the production lead time in the star topology was found to be 46% of the production lead time in the linear topology.

Bottleneck at the beginning of a feeder line A bottleneck at the beginning of a feeder line will tend to starve the machines downstream of it. As a result, the effect of the bottleneck will "move" to the assembly machine where the starved feeder rejoins the rest of the line. An assembly machine that is a bottleneck, on the other hand, will tend to block other feeder lines that are connected to it. Thus, in the first three topologies discussed here, buffers 1–5 were full most of the time, and all other buffers were empty. In the star topology, however, both non-starved feeders connected to the assembly machine were blocked, resulting in nearly full buffers in the segment 1–5 and 11–15. In other words, the amount of in-process inventory was essentially doubled.

Furthermore, because of the starved segments where buffers were generally empty, the length of production line segments was much less influential in determining production lead time than the amount of inventory. As a result, production lead time was almost the same for all four topologies, since the length of blocked segments was the same in every case (five buffers).

Bottleneck at the end of a feeder line The fact that a bottleneck in a feeder line "moves" to the assembly machine to which the feeder is connected has the most pronounced influence when the bottleneck is at the end of the feeder. This is because the feeder with the bottleneck is also blocked in this case. As a result, while only four buffers were usually full in the linear topology, nine buffers were full in Figure 38 (buffers 1–4 and 6–10), and fourteen buffers were full in the other two cases (all the buffers except for buffer 5, which immediately follows the bottleneck). Thus, the total amount of in-process inventory increased by a factor of 3.5 (though the value of inventory may have changed by a different amount).

As a consequence of the many full buffers in all the topologies other than the linear one, production lead time was least in the linear topology; the production lead time of the star topology was 25% higher in the star topology than in the linear topology.

These results show that although a short main line with multiple feeder lines may be desirable for external reasons (e.g. geography), the performance of a production system with such a topology may be significantly adversely influenced by a bottleneck in one of the feeder lines. Furthermore, if feeder lines represent production systems that are located at different facilities or that even belong to other corporate entities, they may be difficult to control.

Quality failures Finally, we analyzed the four topologies described above for the case where quality failures are present. We assumed seven-state machines with a yield of 99% in the high-yield state, and 90% in the low-yield state. While total production rate is very insensitive to topology, good production rate may exhibit some slight sensitivity to topology for a given number of inspection stations. For example, Figure 41 shows the best achievable good production rate for 1, 2, or 3 inspection stations, as well as in the case of no inspection at all. When there is no inspection, the good production rate is highly insensitive to topology, as seen earlier. When a small number of inspection stations are available, the good production rate can be weakly sensitive to topology, because different topologies lead to different proximities to inspection stations; in other words, in some topologies, distances between the machines and the inspection stations that monitor their work are shorter than in others, resulting in more effective inspection.

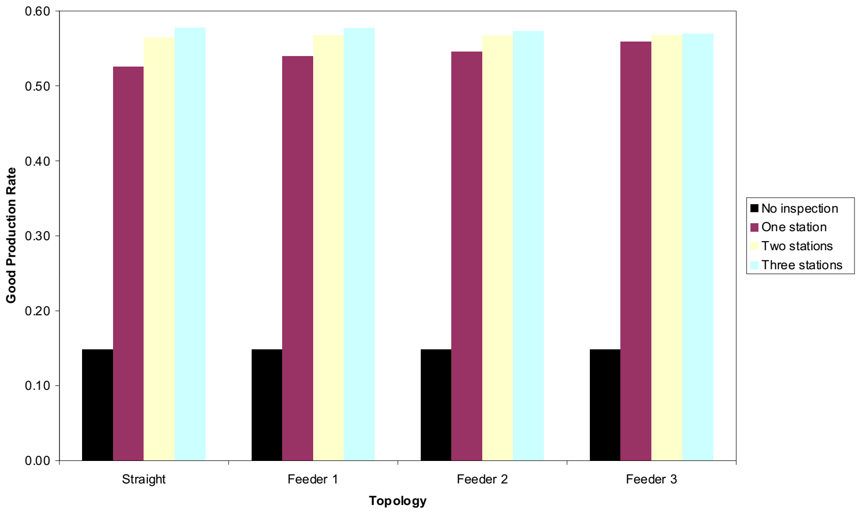


Figure 41: Good production rate for different topologies and different numbers of inspection stations. The leftmost bar in each set corresponds to no inspection; other bars indicate the best achievable good production rate for each number of inspection stations.

5 Conclusion

5.1 Summary

Although productivity and quality have each been extensively studied in the manufacturing engineering literature, there has been little research on how they interact. The work reported here represents a first attempt at analyzing how production system design, quality, and productivity are inter-related in production networks, and developing analytical as well as numerical methods that make it possible to evaluate, compare, and optimize the performance of competing designs.

This report presents the results of work using analytical, numerical, and simulation techniques to investigate the effect of separating inspection from operation, as a function of machine, buffer, and inspection station parameters; control policy; production system topology, and other factors.

A taxonomy was presented for quality failures, and a class of stochastic models was described to represent a realistic subset of those failures. Markov models were formulated for machines subject to both operational and quality failures. Quality control mechanisms were formulated, including scrapping of defective parts as well as a quality information feedback scheme that results in taking offending machines down for maintenance. Pertinent performance measures were defined, including mean total production rate, good production rate, yield, in-process inventory, and lead time.

Analytical results based upon a continuous-material approximation of a two-stage line were presented. These results have been validated by comparison with discrete-event simulation. Further simulation results were presented for more complex production systems including more realistic machine models, inaccurate inspection stations, non-linear topologies, arbitrary locations of inspection stations, and other factors.

It was shown that total production rate and good production rate do not always behave in the same manner as system parameters are varied. For example, although total production rate is known to be a monotonically increasing function of buffer capacity, good production rate is not always monotonic in buffer capacity when quality inspection and information feedback are present.

Furthermore, while improvement in production rate may be obtained through different means—e.g. more reliable machines, more accurate inspection, larger buffers—such improvements are more sensitive to some system parameters than to others. Thus, it is necessary to exercise care in choosing where to invest money in order to obtain the greatest and most cost-effective marginal improvement in good production rate.

It was also shown that taking a machine down for maintenance as soon as a single quality defect is detected (*jidoka*) is not always the optimal policy, and that system parameters must be taken into account in determining the best course of action following the detection of one or more defects.

Other results presented here include the investigation of the effect of the number and placement of inspection stations, of machine and buffer parameters, of inspection policies, of control policies, and of production line topology on selected performance measures. In particular, it was shown that performance can be quite sensitive to the placement of inspection stations, so that a few well-placed inspection stations will result in better performance than many poorly-placed ones.

It was further shown that in-line scrapping can significantly improve performance, if economically justifiable. Finally, it was shown that although production rate is not sensitive to topology, in-process inventory and production lead time are very sensitive. Thus, external factors that call for short main lines with long feeder lines, such as modularization and outsourcing, must be weighed against system performance in making system design decisions.

5.2 Current Research

The research described below is currently underway, supported by General Motors.

5.2.1 Exploration of Phenomena

Analytical, numerical, and simulation techniques are being used to investigate important issues in a way that will help us to identify important phenomena and form preliminary hypotheses about their behavior. Examples include:

- The interaction between buffers, inspection location and accuracy, and production system performance;
- The development of optimal inspection rules based on Bayes risk methods in the context of our models and methods; and
- A wider variety of production topologies, including systems with rework, repair, scrap, assembly/disassembly, etc.

5.2.2 Validation of model assumptions

Data (time-stamped machine events and inspection verdicts) will be analyzed to validate various versions of the multi-state quality model and its assumptions (such as the Markovian property, and distinct high yield vs. low yield states). Transition probabilities will be deduced that fit available quality and throughput data provided by GM.

General Motors will identify plant, process, and data availability, and provide available data to support analysis and model development.

5.2.3 GM factory case study

A model will be developed for the GM-supplied plant layout and process, to take advantage of available data. That model will be applied (using simulation, if mathematically intractable) to identify improvement opportunities in the current process.

General Motors will identify plant, process, and data availability, and provide available data to support analysis and model development.

5.3 Further Research

We believe that the area that we have described is an important one, which will attract many follow-on research contributions. In this section, we briefly describe some areas that promise to provide practical results or deep insights, or that expand our framework for future research.

5.3.1 Non-Instantaneous Inspection

In all our models, we have assumed that inspection takes no time. As a consequence, time plays no role in deciding where inspection should occur. Our conclusions may change, however if it does take time; and how they change will depend on whether inspections of different features at the same inspection station can occur in parallel, or whether they must be done sequentially. If they can be done in parallel, there is incentive to group them; otherwise it may be better to spread them out. This is initially a simulation task.

5.3.2 Buffer Delays and Failures

In all our models, we have assumed that the travel time in buffers is zero. That is, we assume that parts instantaneously move from the upstream machine to the last position in the buffer queue. As a consequence, our models overestimate production rate and underestimate lead times. In addition, they overestimate yield since the additional travel delay causes a greater lag between production and detection of bad parts, when inspection is not ubiquitous.

We have also assumed that buffers are perfectly reliable. This is a reasonable assumption when buffers are simple, passive devices. However, when they include material handling equipment such as robots, they are as prone to failure as anything else in a factory. There has been some work on unreliable buffers (Lipset, Til, and Sengupta 1999; Burman 1995), but none on their impact on quality.

5.3.3 Analytical Models Of Scrapping

It should not be difficult to add scrapping to the two-machine, one buffer model of Section 3. At the same time that Machine 1 makes the transition from state 1 to state -1, the buffer level instantaneously drops to zero. This will change the differential equations and boundary conditions of the probability distribution, but they should not become much more difficult to solve because they will remain linear.

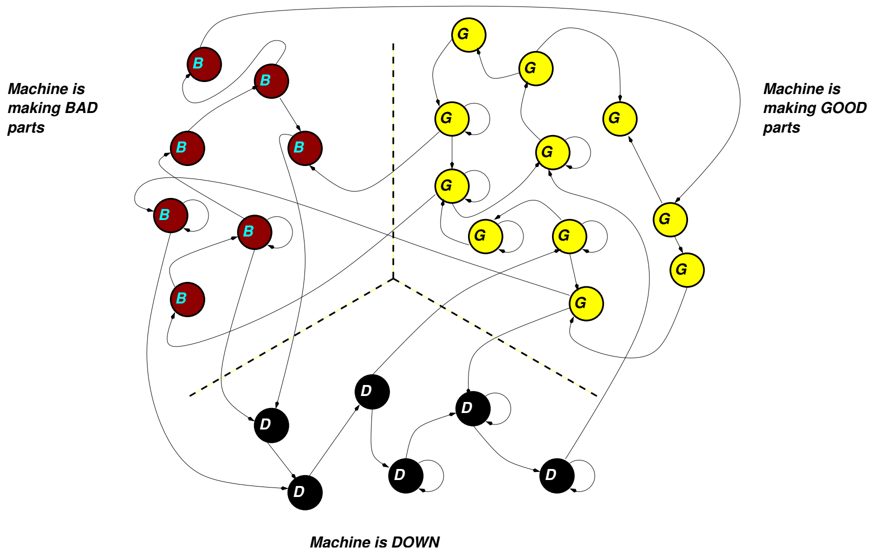


Figure 42: Machine model with general transition structure.

5.3.4 Non-FIFO Buffer Discipline

We have assumed that parts are processed first-in-first-out in all our simulations. This is a reasonable default assumption, but other disciplines, such as LIFO, may provide some benefit. For example, if parts are processed LIFO, there is minimal buffer delay for some parts. This may make it possible to catch persistent failure transitions faster. However, since other parts experience a longer delay, this strategy may be risky. Careful analysis and simulation experimentation may determine whether LIFO or another policy is always, sometimes, or never better than FIFO.

5.3.5 Rework

The rework of parts that fail inspection is done at many factories. This feature should be included in both simulation and analytical models. It is not clear how adding rework to a line affects any of the observations that have been described in this report.

5.3.6 Analytic Modeling of More Elaborate Failure Dynamics

We have investigated only a limited kind of failure dynamics—abrupt transitions from high-yield states to low-yield states with a transition time that is distributed exponentially. In reality, there are many different ways in which quality can degrade. For example, the transition times do not have to be exponentially distributed. Non-exponential transition times can be studied using Markov process models by adding states. Figure 42 is a representation of a system with very general transition behavior. There are good production states, bad production states, and down states. By designing the transition graph appropriately, a wide range of failure dynamics can be analyzed. Also, yield can degrade gradually, as in Figure 9 Over time, the fraction of parts that are bad increases.

A further generalization would be to associate one or more physical quantities (such as hole diameter or location, or surface finish) with each of the good and bad states of Figure 42. It would not be the values of the quantities that would be associated with the state; it would be their probability distributions.

The measurement of these quantities, when combined with the knowledge of the failure dynamics, could provide a more precise and more accurate estimate of the quality state of the machine.

5.3.7 Preventive Maintenance

The more elaborate quality failure dynamics described above provides an opportunity for preventive maintenance. The estimate of the quality state of the machine can be used to make an estimate of how many more operations can be performed before the machine produces parts of unacceptable quality. The machine can then be scheduled for maintenance before any bad parts are actually produced (with high probability).

5.3.8 More General Phenomena, Including Loops, Batches, and Setups

Material flows in factories are not limited to tandem lines or even assembly trees. There are often rework loops, reentrant loops, pallet/fixture loops, assembly following temporary disassembly, etc. In addition, items do not always travel one by one; sometimes it is preferable to create batches or lots (especially when there are operations that involve chambers, such as ovens). Batches create complexities when they are not permanent, that is, when the same group of parts do not travel together in the whole process. Setups, in which the time needed by a machine to operate on a part A after a part B is greater than after another part A, is another cause of scheduling and performance estimation complexity.

All of these phenomena have a bearing on both quality and quantity. They are under active investigation in the quantity literature; we are not aware of any quality-related research in them.

5.3.9 Integrated Models Of Material Flow Control And Inspection

Models of loops can also be used to represent some kinds of material flow control (Frein, Mascolo, and Dallery 1994; Dallery and Liberopoulos 2000; Gershwin 2000). Thus, the analysis of quality using models of material flow in some kinds of loops can be used to model quality in systems with material flow control. This can lead to the unified modeling and design of manufacturing systems taking into account a very large set of important items.

5.3.10 Decomposition Without Reducing Machine Models to Two States

We have seen that two-state machine models can sometimes be good approximations of three-state models (Section 3.3.2) and that this can be used to simplify the calculations associated with decomposition. However, this is not likely to work in all cases, especially when machine models become more elaborate. It would be useful to extend the decomposition to deal with the full complexity hinted at in Section 5.3.6.

5.3.11 Two-Machine, One-Buffer Lines with Good and Bad Parts Modeled Distinctly in the Buffer

In the models described in Section3, we do not model good and bad parts separately in the buffers. This choice was made to simplify the models; if good and bad are modeled separately, then the probability distribution has two continuous variables, and this means that the internal transition equations are partial differential equations. While that choice was appropriate for a first approach to the issues we are studying, another look at the models with good and bad parts modeled explicitly may still be fruitful.

5.3.12 More General Inspection Strategies

We have studied only a limited range of possible inspection strategies. In particular, we have only considered 100% inspection. This is sometimes the best strategy, but sometimes it is preferable to sample, and future research should be devoted to sampling in the quality/quantity context. Second, we have only considered systems in which each feature is inspected only once. Sometimes, certain features are inspected more than once, and this may have an impact on the issues we have considered. Third, we have not at all considered destructive testing. Destructive testing is appropriate sometimes, and it should be studying within the framework we have been developing.

Acknowledgements

The authors gratefully acknowledge the stimulating discussions and valuable input of Dennis Blumenfeld, Ninja Huang, Robert Inman, Jingshan Li, Samuel Marin, and Jonathan Owen.

References

Bertsekas, D. P. and J. N. Tsitsiklis (2002). *Introduction to Probability*. Athena Scientific.

Besterfield, D. H., C. Besterfield-Michna, G. Besterfield, and M. Besterfield-Sacre (2003). *Total quality management*. Englewood Cliffs, NJ: Prentice Hall.

Burman, M. H. (1995). *New Results in Flow Line Analysis*. Ph. D. thesis, Massachusetts Institute of Technology, Operations Research Center.

Dallery, Y., R. David, and X.-L. Xie (1988). An efficient algorithm for analysis of transfer lines with unreliable machines and finite buffers. *IEEE Transactions on Automatic Control 34*, 943–953.

Dallery, Y. and G. Liberopoulos (2000). Extended kanban control system: combining kanban and base stock. *IIE Transactions on Design and Manufacturing 32*.

Frein, Y., M. D. Mascolo, and Y. Dallery (1994). On the design of generalized kanban control systems. *International Journal of Operations and Production Management 15*(9), 158–184.

Fujimoto, T. (1999). *The evolution of a manufacturing systems at Toyota*. Oxford University Press.

Gershwin, S. B. (1987). An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Operations Research 35*(2), 291–305.

Gershwin, S. B. (1994). *Manufacturing Systems Engineering*. Englewood Cliffs, NJ: Prentice-Hall. For corrections, see `http://web.mit.edu/manuf-sys/www/gershwin.errata.html`.

Gershwin, S. B. (2000). Design and operation of manufacturing systems — the control-point policy. *IIE Transactions 32*, 891–906.

Gershwin, S. B. and I. C. Schick (1980). Continuous model of an unreliable two-machine material flow system with a finite interstage buffer. Report LIDS-R-1039, MIT Laboratory for Information and Decision Systems.

Glassey, C. R. and Y. Hong (1993). Analysis of behaviour of an unreliable n–stage transfer line with (n-1) interstage buffers. *International Journal of Production Research 31*(3), 519–530.

Howard, R. A. (1971). *Dynamic Probabilistic Systems*. New York: Wiley.

Inman, R. R., D. E. Blumenfeld, N. Huang, and J. Li (2003). Designing production systems for quality: research opportunities from and autombile industry perspective. *International Journal of Production Research 41*(9), 1953–1971.

Kim, J. (2004). *Designing Production System for Quality and Quantity*. Ph. D. thesis, Massachusetts Institute of Technology.

Kim, J. and S. B. Gershwin (2005). Integrated quality and quantity modeling of a production line. *OR Spectrum*. to appear.

Lipset, R., R. V. Til, and S. Sengupta (1999). Steady-state performance analysis of serial transfer lines subject to machine and buffer failure. *IEEE Transactions on Automatic Control 44*(2), 319–325.

Toyota Motor Corporation (1996). *The Toyota production system*.

Monden, Y. (1983). *Toyota Production System: Practical approach to production management*. Atlanta: Industrial Engineering and Management Press.

Monden, Y. (1998). *Toyota production system — An integrated approach to Just-In-Time*. EMP Books.

Pande, P. and L. Holpp (2002). *What is six sigma?* McGraw-Hill.

Sandberg, A. (1995). *Enriching production*. Avebury Publishing Company.

Sevast'yanov (1962). Influence of storage bin capacity on the average standstill time of a production line. *Theory of Probability Application 7*, 419–438.

Togo, Y. and W. Waterman (1993). *Against all odds: the story of the Toyota Motor Corporation and the family that created it*. New York: St. Martin's.

Womack, J. P., D. T. Jones, and D. Roos (1990). *The machine that changed the world: The story of lean production*. Rawson Associates.

Woodall, W. H. and D. C. Montgomery (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology 31*(4), 376–386.

Zimmern, B. (1956). Etudes de la propagation des arrets aleatoires dans les chaines de production. *Review Statististical Applications 4*, 85–104.