Library Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Liping Liang
Name of Author (please print)

18/03/2004
Date (dd/mm/yyyy)

Title of Thesis: Determining Staffing Levels for a Product Inspection Process

Degree: Master of Science in Business Administration
2004

Department of Sauder School of Business
The University of British Columbia
Vancouver, BC Canada
Abstract

A company seeks a methodology to determine staff levels for its product inspection system. This company produces a type of product that requires 100% inspection before being shipped. The system is a complex tandem queue with machines or operators working in parallel at each station and having dynamic interaction between the inspection rates at two of the stations. The products to be inspected arrive at the inspection system stochastically, with the volume fluctuating considerably with time. However, the expected volume is partially predictable by time of day. If congestion occurs in the system, new products will not be able to enter and receive inspection. To ensure a reliable inspection system where the probability of system congestion is nearly zero, it is desired to determine the minimum staffing level required for different product volumes to avoid system congestions.

In this thesis, we investigate the staffing problem in two ways. First, we develop a simulation model for the inspection system. The simulation model is used for scenario and sensitivity analyses to calculate the service rate of the inspection system for various staffing levels and then to determine the minimum number of staff required to ensure system reliability for a range of product volumes. Next, we apply queuing theory to the staffing analysis, and show that the staffing levels determined by the queuing theory approximation are close to the simulation results.

The actual company and the nature of the products involved must remain confidential. Similarly, all data in this thesis have been scaled by an arbitrary number. However, the critical topology of the queuing system, the nature and demands of the problem and the methodology of their solution are easily portable to other environments.
# Table of Contents

Abstract.................................................................................................................. ii
Table of Contents...................................................................................................... iii
List of Tables ........................................................................................................... v
List of Figures .......................................................................................................... vi
List of Symbols and Abbreviations ......................................................................... vii
List of Symbols and Abbreviations ......................................................................... vii
Acknowledgements ................................................................................................ viii
1. Introduction ......................................................................................................... 1
   1.1 Background .................................................................................................... 1
   1.2 The Product Inspection Process ................................................................ 1
2. Methodology ........................................................................................................ 4
   2.1 Product Inspection Process as a Queuing System ...................................... 4
   2.2 System Performance Measure .................................................................... 4
   2.3 Assumptions ................................................................................................. 5
   2.4 Simulation Technique versus Queuing Theory .......................................... 5
3. Using Simulation to Determine Required Staffing Levels .............................. 7
   3.1 Process Mapping and Data Collection ....................................................... 7
   3.2 Data Analysis ............................................................................................... 8
      3.2.1 Service Times ....................................................................................... 8
      3.2.2 Machine Down Time .......................................................................... 9
   3.3 Simulation Model ......................................................................................... 10
      3.3.1 User Interface ...................................................................................... 10
      3.3.2 Program for Scenario Analysis .......................................................... 11
      3.3.3 Three Versions of Simulation Model ................................................. 12
      3.3.4 Model Validation ................................................................................ 13
   3.4 Scenario Analysis ......................................................................................... 14
      3.4.1 Determining System Service Rate ...................................................... 14
      3.4.2 Determining Required Staffing Levels .............................................. 17
      3.4.3 System Performance with a Shock in Product Arrival ..................... 19
   3.5 Sensitivity Analysis ...................................................................................... 21
      3.5.1 Approval Rate ..................................................................................... 21
      3.5.2 Operator Decision-making Time ....................................................... 22
      3.5.3 Type of Station 2 Service Time Distribution .................................... 22
3.5.4 Machine Down Time ................................................................. 23
4. Applying Queuing Theory in the Staffing Analysis ............................ 25
   4.1 Problem Description .............................................................. 25
   4.2 Station 2 Analysis ............................................................... 26
      4.2.1 Queuing Model Distributions with Decision Time Limit .......... 26
      4.2.2 Model Formulation ......................................................... 28
         4.2.2.1 Assumption .......................................................... 28
         4.2.2.2 Model: CPE Service Time, Multiple Servers ............... 29
      4.2.3 Numerical Results ....................................................... 31
      4.2.4 Accuracy of Approximation versus Original Distribution .... 33
   4.3 Station 3&4 Analysis ......................................................... 35
   4.4 Comparison of Queuing Theory Result with Simulation Result ....... 36
5. Conclusions ............................................................................ 40
   5.1 Summary ........................................................................... 40
   5.2 Future Research ................................................................. 40
Bibliography ............................................................................. 42
Appendix 1. Visual Basic Code for Simulation User Interface .............. 43
Appendix 2. Expected Total Time not Spent on Products because of Time-out 49
Appendix 3. Percentage of Time-out at Station 2 due to Decision Time Limit 51
List of Tables

Table 1. Three Versions of the Product inspection Simulation Model ........................................... 13
Table 2. Example of Scenario of Product Inspection System .................................................. 14
Table 3. Percentage of Time-out with Different Service Time Distributions ................................. 23
Table 4. Percentage of Time-out: Theoretical Result versus Simulation Result (CPE versus CPG) ................................................................................................................................. 32
Table 5. Percentage of Time-out: Theoretical Result versus Simulation Result (CPE versus CPG, Gamma with Different Shape Parameters) ................................................................. 34
Table 6. Scenario Analysis: Theoretical Result versus Simulation Result ...................................... 37
Table 7. Required Staffing Level: Theoretical Result versus Simulation Result ............................. 38
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Product Inspection Process at Company X</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Distribution of Station 2 Server’s Processing Time</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>Distribution of Station 3 Server’s Processing Time</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Distribution of Time between Breakdowns</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Distribution of Duration of Breakdown</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>Example of User Interface</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>Example of Excel File with Scenario Parameters</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>Example of Excel File with Scenario Output</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>System Service Rate with Different Number of Station 2 Operators (Two Station-3 machines)</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>System Service Rate with Different Number of Station 2 Operators (Three Station-3 machines)</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>System Service Rate with Different Decision Time Limit at Station 2</td>
<td>17</td>
</tr>
<tr>
<td>12</td>
<td>Comparison of Required Staffing Levels for Different Station-2 Time Limit</td>
<td>18</td>
</tr>
<tr>
<td>13</td>
<td>Sample Product Arrival Rate in Peak Hours</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td>Results of Product Shock Experiments</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>Effect of Change in Approval Rate on System Performance</td>
<td>21</td>
</tr>
<tr>
<td>16</td>
<td>Effect of Change in Operator’s Service Time on System Performance</td>
<td>22</td>
</tr>
<tr>
<td>17</td>
<td>Probability Density Functions of Two Gamma Distributions</td>
<td>23</td>
</tr>
<tr>
<td>18</td>
<td>Effect of Change in Time between Station-3 Machine Failures on System Performance</td>
<td>24</td>
</tr>
<tr>
<td>19</td>
<td>Distribution of Station-2 Effective Service Time with Decision Time Limit</td>
<td>27</td>
</tr>
<tr>
<td>20</td>
<td>Distribution of Station-2 Effective Waiting Time with Decision Time Limit</td>
<td>28</td>
</tr>
<tr>
<td>21</td>
<td>Distribution of Station-2 Effective Time in System with Decision Time Limit</td>
<td>28</td>
</tr>
<tr>
<td>22</td>
<td>Percentage of Time-out: Simulation Result versus Theoretical Result</td>
<td>32</td>
</tr>
<tr>
<td>23</td>
<td>Percentage of Time-out versus Number of Station 2 Servers for Different Product Volumes</td>
<td>33</td>
</tr>
<tr>
<td>24</td>
<td>Cumulative Probability Distribution of Different Statistical Distributions</td>
<td>34</td>
</tr>
<tr>
<td>25</td>
<td>Station 3 Arrival Rate: Simulation Result versus Theoretical Result</td>
<td>38</td>
</tr>
</tbody>
</table>
List of Symbols and Abbreviations

- $L$: Decision time limit of Station 2 Operators
- $W_q$: Waiting time of a product if there is decision time limit for Station 2 operators
- $W_s$: Service time if there is decision time limit
- $\Phi(L)$: Percentage of products processed within the decision time limit $L$
- $\Phi(L)$: Percentage of products not processed due to the decision time limit $L$
- $A$: Expected service time not spent on products because of time out
- $A_j$: Approval rate of Station 1
- $A_2$: Approval rate of Station 2
- $A_3$: Approval rate of Station 3
- $a$: Constant in the distribution of Station 2 operator's service time
- $X$: Effective service rate if there is decision time limit
- CPE: Constant plus exponential (distribution)
- CPG: Constant plus Gamma (distribution)
Acknowledgements

The completion of this thesis and the success of the project were made possible by the hard work and support of many people.

I would like to express my sincere appreciation to my advisor, Professor Derek Atkins, for being actively involved in all aspects of the project and providing very helpful advice to both the project and my thesis. I am always impressed by his profound knowledge and organizational skills.

I would also like to acknowledge Professor Martin L. Puterman, Director of the Centre for Operations Excellence (COE), for his help and encouragement in my pursuit of a higher education. In addition, I extend my sincere gratitude to Professor Anming Zhang, for his encouragement in my study, as well as his time and input as my thesis committee.

I would like to thank the Centre for Operations Excellence (COE) for giving me the extraordinary opportunity to work on an applied research project for my thesis. I extend my sincere thanks to the Director Industry, Mats Gerschman, for his feedback and input along the way. I would also like to thank all the students and staff from the COE who helped in this project for their hard work and suggestions.

Special thanks go to Bailey Kluczny, for his excellent work in the simulation development and other aspects of the project and Mehmet A. Begen, for his technical input, enthusiasm and support.

I offer many thanks to the company for sharing their challenging problem with the COE. Many staff shared their experience, and provided assistance and resources throughout the project. All their support was greatly appreciated.

Finally, I am always grateful to my family for their encouragement and support for my pursuit of a higher education.
1. Introduction

1.1 Background
This thesis describes the methodology used in the study of staffing requirements for a company’s product inspection operation. For reasons of confidentiality, the company will be called X. Company X manufactures a type of product that requires 100% inspection before being shipped. The inspection process for the product is a common tandem system. However, inspection requirements and operating characteristics make the system very complex. To improve the efficiency and effectiveness of the product inspection process, Company X asked the Centre for Operations Excellence to conduct this study.

Products arrive at the system stochastically, and there are considerable fluctuations in demand for inspection during the day. Although the number of machines is fixed, the number of operators can be changed in response to the product volume. Therefore, a main objective of this study was to provide a methodology for determining the minimum number of staff required and the optimal allocations of the staff for various product volumes. Another purpose of this study was to evaluate the capacity of the inspection system with the current machine level, and to make recommendations about the system capacity with different machine and operator levels. System capacity refers to the maximum number of products the system can inspect per time unit.

1.2 The Product Inspection Process
Product inspection is widely carried out by manufacturers to ensure product conformity and freedom from contamination. Products are usually inspected by machines and/or people.

The product of Company X requires 100% inspection, and the inspection is a tandem process with machines or operators working in parallel at each station. Figure 1 illustrates this process. Due to the uncertainty about whether or not a product should be approved, each inspection station is associated with an approval rate (percentage of products that a station is able to approve). Each product is first inspected at a cheaper station, which is set at a conservative approval level. Products approved at an inspection station will pass directly to shipping, while those not approved will go on to a more expensive and more rigorous station for further inspection. Station 1, which is at the beginning of the product inspection process, consists of machines with sufficient capacity for handling the maximum possible product volume experienced by the inspection system, so Station 1 will never cause a problem in the inspection process. Following
Station 1 is Station 2 consisting of a number of operators. There is a limit to the time each product is held at Station 2, which includes the time a product waits for inspection and the duration the product is inspected. If the inspection on a product is not completed or even not started when the time limit is reached, the product is treated as not being approved. It should be noted that the approval rate of Station 2 is the rate if there were no time limit. Station 3 consists of machines that are subject to frequent breakdowns and are also expensive equipment, thus this station has the risk of not being able to cope with the throughput and is vulnerable to becoming a potential bottleneck. At this stage, due to a finite buffer between Station 2 and Station 3 with space for holding only a limited number of products, the buffer will become full and the system has to stop processing new products. This state is called system congestion, and is considered to be a very expensive state to be in as effectively the capacity of the company is closed down. The intention is to avoid this if possible. For those products unable to be approved at Station 3, a fourth inspection station will be involved and each operator at this station inspects the product together with Station 3, which means the product will still occupy Station 3, so the processing time of Station 3 for this product becomes longer. This reduces the inspection capacity of Station 3. There is a final manual inspection for products not approved at Station 4, which is not an essential part of this study.

Although the inspection system is a common tandem type, the dynamics of the interaction between the inspection rates at stations 2 and 3, the time limit for each product at Station 2, the approval rate at stations 1 to 4, and the number of machines or operators at stations 1 to 4, as well as the size of buffers between Station 2 and Station 3 all make the inspection process very complex.

Station 1 has a fixed number of machines with sufficient capacity to handle the maximum possible product volume. Station 2 can have one up to a finite number of operators working in parallel; Station 3 consists of two or three machines; and Station 4, since each operator works together with a machine at Station 3, can have the number of operators from one to the number of Station 3 machines. When the number of Station 4 operators is less than the number of Station 3 machines, a machine with unapproved product has to wait for a Station 4 operator if all Station 4 operators are working with other machines. Therefore, the staffing analysis concerns the number of operators required for both Station 2 and Station 4, that is, the value of $N_2$ and $N_4$ in Figure 1.
Figure 1. Product Inspection Process at Company X
2. Methodology

2.1 Product Inspection Process as a Queuing System

The product inspection process at Company X can be modeled as a complex queuing system with the following characteristics. Firstly, it is composed of a few inspection stations with multiple servers working in parallel at each station, and the waiting space (buffer) between stations is finite. Secondly, there is a limit to each product’s total time at Station 2 (including the waiting time and the service time). Once the time limit is exceeded, no matter whether the product is still waiting for inspection or is being inspected, it has to pass to Station 3. As a result, only those products that have finished inspection and have been approved will not go to Station 3. In addition, machines at Station 3 may break down while they are working, resulting in a decrease in the capacity at that station. Finally, the product arrival is nonstationary, and the processing time of machines or operators at each station is neither deterministic nor exponentially distributed. Therefore, if we at least assume the product arrival follows a Poisson process for any particular time of day, then this inspection process can be regarded as an open queuing network with serial stations, where each station is an M/G/c/k queue, the product arrives at Station 1, and can leave the system at any station. The system capacity to be determined is in fact the service rate of the whole system.

For simplicity, we shall use server for a machine or operator, and will not distinguish them unless it is necessary.

2.2 System Performance Measure

In the staffing analysis, we needed to measure the system performance in order to determine the optimal staffing levels. We were told by Company X that the total processing time for each product is usually short as long as the inspection system is working, so the waiting time or the total time in system is not an appropriate system performance measure. Moreover, throughput rate is not good either because the throughput rate will always be equal to the product arrival rate if the system is working. Since the product will not be able to enter the inspection system if all the buffers in the system are full and thus the state of system congestion is undesirable, we decided that we only needed to choose staffing levels that ensure the system is not congested and that the products flow smoothly in the system. Therefore, we used the system uptime ratio as the system performance measure, which is defined as the percentage of time the product inspection system
has no congestion. If the system uptime ratio is close to 100%, the inspection system is considered reliable.

### 2.3 Assumptions

The following assumptions were made about the operation of the product inspection system:

- Each operator can work at either Station 2 or Station 4
- All operators are assumed to be full time equivalent (FTE), which means that there is no break in their work
- Product arrival at the inspection system follows a Poisson process
- Approval rate (the percentage of products approved at a station on average) is constant, but it varies with station
- All machines work independently, and machines of the same type have an identical breakdown pattern (including the duration of breakdown and the time between two successive breakdowns)

### 2.4 Simulation Technique versus Queuing Theory

Both queuing theory and simulation can be used to analyze queuing systems. Queuing theory is always the first choice if analytical models can be developed. However, it is very often that simulation has to be used because the queuing system is too complex to be analyzed theoretically.

Queuing theory provides the advantage of generating exact or near exact result, but its limitation is it can only be used to analyze relatively simple queuing systems. Simulation can be used to solve very complex queuing problems, but its drawback is people have to determine the run length, number of replications and simulation output. Another disadvantage to simulation is the considerable amount of time that has to be spent testing various scenarios if the number of scenarios is big; otherwise, the solution may not be optimal if the optimal scenario is not investigated. Moreover, since simulation does not give an exact solution due to the randomness in simulation runs, it would be difficult to determine which one is better when the results of two scenarios are very close. However, queuing theory can be used to determined the optimal or near optimal solution because mathematical optimization can be applied.
Due to the complexity of the product inspection system as discussed earlier, it is very difficult to develop an analytical model for analysis. Therefore, we decided to use simulation for the product inspection analysis. It also has the advantage of providing visual understanding of the system using the animation function.
3. Using Simulation to Determine Required Staffing Levels

The simulation analysis consisted of a series of steps, including process mapping, data collection and data analysis, simulation development and model validation, as well as scenario and sensitivity analyses.

After the simulation model was built, a number of scenarios were tested to determine the service rate of the system with different configurations of machine and operator levels. Since the service rate should be greater than the arrival rate to avoid system congestion, we chose the scenarios to evaluate by its service rate for each product volume, and compared the simulation results to determine the minimum required staffing level and its allocation to ensure system reliability for that product volume. Required staffing levels were then tested using product volume data provided by Company X. The sensitivity of system performance to various system parameters was then tested. Various scenarios were tested to reveal the location of system bottlenecks. Recommendations were made based on all the analytical results.

3.1 Process Mapping and Data Collection

Process mapping is a technique that allows one to see complex operational systems broken down into a sequence of activities and information flows. Process map generation was the initial step enabling us to analyze and eventually simulate the product inspection process. However, a great deal of information was required before any process maps could be created.

Information about the product inspection process was collected through extensive observations of the system and through numerous interviews with the directors and operators from Company X. Bringing all these pieces together helped us to develop a thorough understanding of the system’s processes and procedures, enabling us to create the preliminary version of the process map. The process map was finalized after we made a number of revisions on the preliminary version with the company personnel. It provided a detailed description of the activity and information flows of the inspection system, and enabled us to thoroughly understand the product inspection process and determine the various types of system-specific data to collect for developing the simulation model.

Data on machine and operator service times were collected manually in the system, while Company X provided data such as machine down time, product approval rates at each station,
decision time limit, buffer sizes, product volume, etc. The product approval rate for a station of the system is the percentage of products arriving at that station that are approved by the server at that station. For Station 2 with a decision time limit, the approval rate is the percentage of arriving products that a server would be able to approve if there were no decision time limit.

The accuracy of the process mapping was extensively tested through discussions and several presentations to the management of the company.

### 3.2 Data Analysis

We used the Input Analyzer of Arena 6.0 software to analyze the data on server's processing time and machine down time. Input Analyzer enables us to try various statistical distributions for the data. It also provides the best-fit distribution.

#### 3.2.1 Service Times

![Figure 2. Distribution of Station 2 Server's Processing Time](image1)

- **Distribution:** Gamma
- **Expression:** \( c + \text{GAMM}(\alpha, \beta) \)
- **Square Error:** 0.049342

![Figure 3. Distribution of Station 3 Server's Processing Time](image2)

- **Distribution:** Erlang
- **Expression:** \( c + \text{ERLA}(\lambda) \)
- **Square Error:** 0.004139
There are several types of machines and operators in the product inspection system, and the processing time of each was collected and analyzed. Figure 2 and Figure 3 show the fitted statistical distributions of two processing times. Parameter $c$ in a distribution expression is a constant. It is positive, and is different in different processing time distributions.

### 3.2.2 Machine Down Time

We analyzed the down time data of Station 3 machines in order to find the breakdown pattern. Figure 4 and Figure 5 show the distributions of time between breakdowns and duration of breakdown respectively. Parameter $c$ in a distribution expression is a positive constant, and is different in different distributions.
3.3 Simulation Model

A simulation model was developed for the product inspection system based on the process map, collected data and all other available information. It was built in the simulation software of Arena 6.0. Arena is very convenient for simulating queuing systems because it provides large amount of system-defaulted and user-specified statistics output that can be used to estimate system performance in terms of throughput rate, waiting time, service time and resource utilization. It also has animation function that helps visualize the dynamic changes in a queuing system.

To make the simulation model fully represent the logical flow in the real system, we built into the model important aspects such as machine breakdown, Station 2 decision time limit, and the buffers between Station 2 and Station 3.

The extensive animation features of Arena were used to check the operation of the system against the real life knowledge of the company management. The operation and results collection functions were adapted to make the animation feature highly usable to operating management.

3.3.1 User Interface

![User Interface](image)

Figure 6. Example of User Interface
A user interface was developed for the simulation model. It enables users who do not know the simulation software to view the current parameter settings in the simulation file and change any parameter to the desired value. For users who do know the simulation software, the user interface provides a very straightforward way to input model parameters without having to get into the details of the simulation model. Figure 6 shows an example of the user interface. It pops up when a user runs the simulation model. After the user sets all the parameters to the desired values and click the **Accept** button, the corresponding parameters in the simulation model will be updated, and the simulation model will then start running with the new parameters values.

The user interface was developed using Visual Basic in Arena 6.0. The Visual Basic code related to updating the parameters in the simulation model is attached in Appendix 1.

### 3.3.2 Program for Scenario Analysis

Although the Process Analyzer tool in Arena 6.0 can be used to run scenarios, setting up scenarios in Process Analyzer is very inconvenient. Since scenarios usually differ by the value of only one parameter, setting up scenarios would be very easy if users could use a spreadsheet and do some Copy and Paste operations. However, Process Analyzer does not recognize pasted values due to a bug in the Arena software, the value of each parameter has to be typed in one by one. When running a large number of scenarios, setting up the parameters in Process Analyzer would be very time-consuming. Therefore, we developed a Visual Basic program that is embedded in the simulation model to carry out scenario analysis.

This program reads the parameter values of a scenario from a number of scenarios in an Excel file, assign these values to the corresponding parameters in the simulation model for each scenario run, and write the output of each scenario into another Excel file. Users can define the statistics to output in the Visual Basic program. Reading and assigning parameter values of the simulation model follow the same way as the user interface does. The Excel files containing the parameter values of all scenarios and the output of scenarios are illustrated in Figure 7 and 8 below.

In the Excel file with scenario parameters, each column corresponds to a parameter whose value needs to be changed in a scenario, and each row contains all parameter values for a scenario. A number of scenarios can be set up in this file. In the Excel file with scenario output, each column
corresponds to one output item of a scenario, and each row contains all output of a scenario. The output of all scenarios is contained in one Excel output file.

Figure 7. Example of Excel File with Scenario Parameters

Figure 8. Example of Excel File with Scenario Output

3.3.3 Three Versions of Simulation Model

Three versions of the simulation model were developed and each was set up for a slightly different purpose. They are listed in Table 1.
### Table 1. Three Versions of the Product inspection Simulation Model

<table>
<thead>
<tr>
<th>Model Version</th>
<th>Description</th>
</tr>
</thead>
</table>
| Constant product volume & staffing - User Interface | • Easy to change system parameters  
• Constant level of staffing and product volume throughout the simulation period  
• Fully animated  
• Good for Ad Hoc Analysis  
• Good for demonstrations                                                                 |
| Constant product volume & staffing - Scenario Analysis | • Contains Visual Basic code to simplify scenario analysis  
• Constant level of staffing and product volume throughout the simulation period  
• No Animation  
• Requires accompanying input and output files                                                                 |
| Variable product volume & staffing                  | • Number of staff and product volume can vary over the course of the simulation run  
• Useful for analysis of different product volume patterns  
• Useful for analysis of different staff schedules  
• Fully animated  
• Requires development of schedules for staffing and product volume                                                                 |

### 3.3.4 Model Validation

A simulation model has to be validated before scenario and sensitivity analyses are conducted because an inadequate representation of the real system may mislead us to make a completely wrong conclusion. However, we were not able to validate the model against the real system due to lack of certain data. The only way we could do is to validate the model against itself. Testing the model against itself, or internal validation, is important to ensure the model behaves exactly as it is expected. The complexity of the simulation model made it very difficult to do so. Our first step was to check the logic flow of the simulation, then we designed scenarios to produce specific results. If the simulation outputs did not match the expectation, the model was debugged. We did this several times until we were confident that the model was an adequate representation of the real system. We worked extensively with the management at Company X to do this validation.
3.4 Scenario Analysis

Since the number of machines at Station 1 is fixed and Station 1 has sufficient capacity to handle the maximum possible product volume, the machine level of Station 1 is not under our consideration. Therefore, each scenario of the product inspection system is a combination of the product volume, and the number of machines or operators at inspection stations 2 to 4. Table 2 gives an example of one scenario of the system. Since the number of operators at Station 4 is at most the number of machines of Station 3, if there are $M$ product volumes, a maximum of $N$ operators at Station 2 and two machines at Station 3, then the total number of scenarios is $M*N*(1+2) = 3*M*N$.

<table>
<thead>
<tr>
<th>Product Volume</th>
<th>Station 2 Operators</th>
<th>Station 3 Machines</th>
<th>Station 4 Operators</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 products/time unit</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

The replication length of a simulation run was chosen to be the total operating time of the inspection system in a week so that the machine performance (breakdown) could be effectively simulated. Each scenario was run with 20 replications.

3.4.1 Determining System Service Rate

The service rate of the product inspection system is the maximum number of products the system is able to handle per time unit regardless of system congestion. It varies depending on the number of machines and operators working at each station. It also depends on the decision time limit at Station 2. To determine the system service rate, a constant product volume is specified for each simulation run. Because the output of the product manufacturing system is the input of the product inspection system, the constant product volume is set to be the designed maximum capacity of the manufacturing system and is much higher than normal: this ensures that the inspection system is always busy. The number of products processed is carefully measured. The system service rate equals the total number of products processed divided by the number of simulated time units. The simulation result is the system service rate corresponding to each combination of machines and operators at stations 2 to 4.

To compare the system service rate for different number of servers at a station, each time we fixed the number of servers at all other stations, and compared the service rate for the number of servers at this station from one to the maximum possible number.
3.4.1.a System service rate with different number of operators at a station

Figure 9 gives an example of comparing the system service rate for different number of operators at Station 2. The number of Station 2 operators investigated is from one to nine, and the number of Station 3 machines is two. It shows that the maximum system service rate with two Station 3 machines is 56 products per time unit. The line with diamond dots and the line with square dots describe the cases with one and two operators at Station 4 respectively. It can be seen that both lines change in the same pattern. It also indicates that beyond the same number of operators, which is eight, adding more operators to Station 2 makes very little difference to the system service rate, so the system does not need more than eight operators at Station 2. Moreover, the results enable us to check if the system service rate is increased by the same percentage when adding one more operator at Station 2 regardless of the number of servers at other stations.

3.4.1.b System service rate when adding one more Station 3 machine

When there are sufficient operators working in the product inspection system, a possible way to improve the system service rate is to add additional machines to certain stations. As described earlier, Station 3 is a potential system bottleneck, so scenarios with an additional machine added to this station were investigated, and system service rate was calculated.

Figure 10 compares the system service rate for different number of operators at stations 2 and 4 when the inspection system has 3 machines at Station 3. The line shows the system service rate
with two Station-4 operators, while the bars show the service rate with 3 Station-4 operators. Similar to the results in 3.4.1.a, the system is found to not require any more than 8 operators at Station 2, because the system service rate with 9 or more Station-2 operators is very close to that with 8 Station-2 operators. This figure shows that the system does not require any more than two Station-4 operators, since a third operator makes no improvement in the system service rate. It is also noted that with an additional Station 3 machine, fewer Station 2 operators would be required to achieve the same system service rate, and the maximum system service rate with an additional Station 3 machine is 77 products per time unit, a 36 percentage of increase over the service rate with two Station 4 machines.

![Figure 10. System Service Rate with Different Number of Station 2 Operators (Three Station-3 machines)](image)

3.4.1.c System service rate with different decision time limits

Another possible way to increase the system service rate is to increase the limit on the operator's decision-making time at Station 2. With the original decision time limit, about 11% of the inspection would be time-out even if there were plenty of operators working and no product had to wait for inspection. This is because the service time is longer than the time limit. To evaluate the improvement of the system service rate by increasing the time limit, a 30% increase over the original time limit was used. All the resource levels are the same as in 3.1.1.a. Two Station 4 operators were used. Figure 11 compares the system service rate for each of the two decision time limits considered. The line with square dots represents the system service rate with the original decision time limit, and the line with diamond dots represents the service rate with the new decision time limit.
With the same number of machines and operators at each station, a 30% increase in the Station 2 decision time limit results in an increase in the system service rate of anywhere from 16% to 35% (depending on the staffing level), and the maximum system service rate increases by 16%.

![Figure 11. System Service Rate with Different Decision Time Limit at Station 2](image)

3.4.2 Determining Required Staffing Levels

3.4.2.a Methodology

For a given level of product volume, the required resource level to ensure the reliability of the product inspection system is not necessarily the resource level that has a system service rate equal to the product volume. One reason for this is the fact that the system service rate of a resource level represents the maximum number of products this resource level is able to handle in a time unit, regardless of the number of products waiting in the queue or the possibility of system congestion. On the other hand, with a given level of product volume, the required resource level has to ensure that the probability of system congestion is almost zero. As well, when determining the system service rate, the inspection system is kept busy by maintaining a queue in the system using an excessively high arrival rate. In contrast, for lower arrival rates, machines or operators will sometimes be idle due to the randomness of product arrival.

With the system service rate determined for each configuration of operators at stations 2 and 4 in section 3.4.1.a, the simulation can be used to determine the minimum number of operators required and the optimal allocation of operators to the inspection stations to ensure system reliability over a range of product arrival rates. The product arrival rate and the staffing level were scheduled to be constant for each run. The arrivals were assumed to be a Poisson process.
because the product arrival pattern in the real system was unknown and Poisson arrival is a common assumption in queuing system analysis. Each staffing level that has a system service rate close to or higher than the arrival rate was investigated. The percentage of time the system has no congestion (uptime ratio) was determined to see if the uptime ratio is close to 1. The minimum staffing level that results in an uptime ratio close to 1 (> 99.9%) was determined to be the required staffing level for that arrival rate.

This methodology was applied to scenarios for product arrival rates within the whole possible range that was likely to occur during a day.

3.4.2.b Results

![Graph showing comparison of required staffing levels for different Station-2 time limits.](image)

Figure 12. Comparison of Required Staffing Levels for Different Station-2 Time Limit

The results give the required staffing level and allocation to ensure system reliability for each product arrival rate between 5 products per time unit and 40 products per time unit for the original Station-2 decision time limit and the time limit with 30% increase over the original limit). Figure 12 displays a comparison of these required staffing levels for the two decision time limits. It shows that a 30% increase in the Station-2 decision time limit consistently results in one fewer operator required.
3.4.2. Testing the required staffing levels

Due to significant variation in product volume throughout the day in the product inspection system, one might naturally ask whether the minimum staffing levels determined using a constant product arrival rate could ensure reliability in the real system. For this purpose, the staffing levels determined by the simulation model were tested using real data.

The test was carried out using three hour peak time product arrival data provided by Company X. The data describes product volume in small time intervals during peak hours as shown in Figure 13. The average product arrival rate during each hour interval was calculated from the data, and the number of operators required at each station for each hour period was determined using the required staffing levels obtained in the staffing level analysis. The real data was used to create the product arrival schedule in the simulation. The simulation was run for 20 repetitions, and the total number of congestions in the system was recorded. The simulation results showed there were no system congestions. Therefore, the system performance using recommended staffing levels with constant and fluctuating product rates is practically identical.

![Figure 13. Sample Product Arrival Rate in Peak Hours](image)

3.4.3 System Performance with a Shock in Product Arrival

The product inspection system may occasionally experience product volumes that are much higher than the typical average peak rate for short periods of time during peak hours. At such a high product volume, it is impossible for the system to work in a steady state since the maximum system service rate with current machine levels is lower than these product volumes. Therefore, it is worth investigating how long the system can work under such stress before congestions will likely occur.
It is assumed that no machine breakdowns occur during the periods with high product volume. The scenario was run with 100 repetitions using a typical average peak rate $N_\alpha$ as the average arrival rate. The required staffing level for this product rate was used for the entire run period. The scenario was set up as follows: an arrival rate of $\lambda'$ products per time unit for 50% of the length of the run period so that the simulation can attain a steady state; then an arrival rate of $\lambda''$ products per time unit for 10% of the length of the run period. This 10% is the maximum length of period the high product volume lasts in the real system. Then again $\lambda'$ products per time unit for the remaining 40% of the length of the run period in order to see how well the system can catch up. $\lambda''$ is the product volume used as a shock, and $\lambda'$ satisfies $90\% \lambda' + 10\% \lambda'' = N_\alpha$. The time between when the shock begins and when the first system congestion occurs is calculated. This duration is the length of time that the system can work under the $N_\alpha$ products per time unit arrival rate before system congestions occur. The result is that on average, the inspection system is able to handle a product volume of $N_\alpha$ products per time unit for 4.8 time units before congestions occur. The minimum and the maximum durations from this experiment are 3.1 and 7.1 time units respectively. In three of the experiments, no congestions occurred. The results of all 100 repetitions are plotted as a histogram in Figure 14. This experiment indicates that the inspection system cannot cope with such a high volume because congestions will occur; therefore, Company X should try to prevent such a high product volume from happening by spreading product arrivals.

![Figure 14. Results of Product Shock Experiments](image-url)
3.5 Sensitivity Analysis

The simulation model was also used to investigate the effects of altering certain parameters within the system. A technique known as sensitivity analysis was employed to quantify these effects. Sensitivity analysis is the process of systematically changing one parameter at a time within the model, in order to pinpoint the effects of changing that single variable. The parameters tested include approval rates at stations, the decision making time of operators at each station and Station 3 machine down time. All the analysis was conducted using a typical average peak rate, \( N_a \) products per time unit. The minimum required staffing level for this arrival rate was used for these scenarios.

3.5.1 Approval Rate

One of the parameters tested using sensitivity analysis was the approval rate: that is, the percentage of products arriving at a station of the product inspection system, which are approved by the machine or operator at that station. Since the approval rate at a station influences the product volume that is sent to the successive stations, it would be valuable to know how altering the approval rate at a station would affect the system as a whole. The approval rate of the last station was not investigated in the sensitivity analysis.

![Figure 15. Effect of Change in Approval Rate on System Performance](image)

Scenarios were run to test the effect of adding or reducing the approval rate by 10% at each of stations 1 to 3. All other system parameters were kept constant so the effects of changing the approval rate could be isolated. The effect of the change in approval rate on the number of system congestions was evaluated. Figure 15 compares this effect as a percentage of the congestions using the original approval rates, and shows the result for stations 1 to 3. It shows that a 10%
decrease in the Station 1 approval rate greatly increases the number of system congestions — in fact, congestions rise by nearly 20 times. It also reveals that the system performance is very sensitive to decreases in the Station 1 approval rate, and somewhat sensitive to decreases in the Station 2 approval rate, but not very sensitive to any change in the Station 3 approval rate. This result is reasonable because the Station 1 is closer to the beginning of the inspection process than stations 2 and 3, so its approval rate has bigger impact on the percentage of product processed and thus the system performance than the approval rates of stations 2 and 3.

3.5.2 Operator Decision-making Time
Another model parameter analyzed was the decision-making time for operators at a station. Using the same methodology as the analysis for approval rate, scenarios were run to test the effect on the number of system congestions with a 10% reduction or increase in the decision-making times of operators at stations 2 and 4. All other system parameters were kept constant. Figure 16 compares the results. For example, a 10% increase in the Station 2 decision-making time would result in a 144% increase in the number of system congestions. The figure reveals that the system performance is sensitive to the Station 2 decision-making time, but not sensitive to the Station 4 decision-making time.

![Figure 16. Effect of Change in Operator's Service Time on System Performance](image)

3.5.3 Type of Station 2 Service Time Distribution
Because of the limit to the time that each product is processed at Station 2, Station 2 service time determines the percentage of products not processed at this station due to time-out, and consequently the workload of Station 3 and the performance of the inspection system. It has been
shown in 3.5.2 that the mean of Station 2 service time has a big impact on the system performance. It is also worth to investigate if the type of the service time distribution greatly affects the percentage of time-out at Station 2. The distribution of Station 2 service time is a constant-plus-Gamma (CPG) distribution, with the shape parameter of the Gamma distribution equal to 1.14. To test a different distribution, the shape parameter is changed to be 2 and the mean of the distribution and the constant in the distribution are kept the same as before. Figure 17 depicts the difference between these two Gamma distributions. Using the same arrival rate and the same number of Station 2 servers, simulation was run for 20 replications, and the percentage of time-out at Station 2 was calculated. The result is provided in Table 3. With a small change in the type of service time distribution, the percentage of time-out changes greatly. Therefore, the percentage of time-out is very sensitive to the type of service time distribution.

![Figure 17. Probability Density Functions of Two Gamma Distributions](image)

**Table 3. Percentage of Time-out with Different Service Time Distributions**

<table>
<thead>
<tr>
<th>Shape Parameter</th>
<th>Percentage Time-out</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>1.14</td>
<td>12.7%</td>
</tr>
<tr>
<td>2</td>
<td>9.2%</td>
</tr>
</tbody>
</table>

**3.5.4 Machine Down Time**

As observed in the system, machine down time is an important factor of the system performance. The effect of different breakdown frequencies of various machines on system reliability was investigated.
Again, using the same methodology as the analysis for approval rate, scenarios were run to test the effect of a 10% reduction or increase in the time between machine failures. A 10% reduction in the time between machine failures is equal to a 10% increase in the frequency of machine failures. Similarly, a 10% increase in the time between machine failures is equal to a 10% reduction in the frequency of machine failures. All other system parameters were kept constant. Figure 18 describes the effect for machine at Station 3. A 10% decrease in the time between failures (10% increase in the frequency of machine failures) would result in a 40% increase in the number of system congestions. Therefore, the system performance is sensitive to the time between Station 3 machine failures (frequency of Station 3 machine failures).

![Figure 18. Effect of Change in Time between Station-3 Machine Failures on System Performance](image)

Figure 18. Effect of Change in Time between Station-3 Machine Failures on System Performance
4. Applying Queuing Theory in the Staffing Analysis

We have discussed in Section 2.4 the pros and cons of using queuing theory and simulation to analyze queuing systems. Although it is very difficult to purely use queuing theory to analyze a complex system like the product inspection system at Company X, queuing theory could still assist us in making decisions if it proved to be at least a reasonably approximate substitute.

4.1 Problem Description

In the staffing analysis by simulation, we sometimes encountered a problem of deciding the required staffing level at a station. For some product volumes, we used simulation to find two scenarios that both were suitable for the same product volume, and they differed by only one operator at either Station 2 or Station 4. The scenario with one more operator at either Station 2 or 4 has a smaller uptime ratio, which is unreasonable. It may also happen that the uptime ratios of these two scenarios are extremely close, with a difference less than 0.1%. This is due to the random factors in the simulation runs. We found that the machine down time differs in each scenario. This kind of result made it very difficult to make a choice between different staffing levels. Although it is discussed earlier that a queuing model is not easy to develop for the inspection system due to the complexity of the system, we want to use a queuing method to find an approximation for the inspection system, which could be used to help determine the required staffing level.

As described earlier, in the inspection system, Station 1 has sufficient capacity for any possible product volumes, so it is never a bottleneck; Station 2 cannot be a bottleneck because no product will be held up at this station due to the existence of decision time limit; Station 3 is a potential bottleneck because of the finite buffers between Station 2 and Station 3 and the possibility of machines down time; and each operator of Station 4 works with a Station 3 machine only when necessary. Therefore, the percentage of products timed out at Station 2 determines the product volume to Station 3; and the number of Station 4 operators can influence the service rate of Station 3 because if both Station 3 machines are held by products not approved and there is only one Station 4 operator available, then one Station 3 machine has to wait for this operator and thus the machine’s service time is prolonged. As a result, in order to avoid the congestion in front of Station 3, there is a tradeoff between the number of operators at stations 2 and 4.
Base on the above analysis, we focus our attention on Station 2 to Station 4, and use the volume of products not approved at Station 1 as the arrivals to Station 2.

4.2 Station 2 Analysis

First, we consider Station 2, which has a limit to each product's total time at this station (including the waiting time and the service time), and can have one up to an arbitrary number of operators working in parallel. We use $L$ to denote this time limit. Once the time limit $L$ is exceeded, no matter whether the product is still waiting for inspection or is being screened, it has to go to the next station. The arrival rate is calculated by $(1 - A_i)\lambda_1$, where $A_i$ and $\lambda_1$ are the approval rate and the arrival rate of Station 1, respectively. The operator's service time has a distribution as $a + \text{Gamma}(\alpha, \beta)$ with the density function as

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} (x-a)^{\alpha-1} e^{-\frac{x-a}{\beta}} \quad x \geq a,$$

where $a$ is a positive constant. It is noted that the shape parameter, $\beta$, is close to 1. A Gamma distribution with the shape parameter equal to 1 is an exponential distribution. Therefore, the service time distribution is approximately a constant plus exponential (CPE), which can be seen from the shape of the distribution shown in Figure 2.

We want to calculate the percentage of products that are not processed due to time-out because this value determines the product volume to Station 3, the potential bottleneck of the inspection system. Let $\Phi(L)$ denote the percentage of products inspected within the decision time limit $L$, then the percentage of products not inspected due to time-out is $\bar{\Phi}(L) = 1 - \Phi(L)$.

4.2.1 Queuing Model Distributions with Decision Time Limit

The existence of decision time limit (the limit to the total time in system) makes the service time distribution different from what it would be if there were no such limit. In order to distinguish from the original service time, we use the effective service time to refer to the actual service time when there is decision time limit. So are the actual waiting time and time in system. Let $W_q$ and $W_s$ denote a product's effective waiting time and effective service time if there is decision time limit; and $W_q'$ and $W_s'$ be the waiting time and service time if there is no decision time limit. $W_q$, $W_s$, $W_q'$ and $W_s'$ are random variables. $W_s$ and $W_s'$ have the following relationship.
1. If $W_q < L$, $W_s' \leq L - W_q$, then $W_s = W_s'$.

   Decision is made within time limit.

2. If $W_q < L$, $W_s' > L - W_q$, then $W_s = L - W_q$.

   $W_s$ is truncated. Decision cannot be made due to time-out.

3. If $W_q \geq L$, then $W_s' = 0$.

   Service is not started because the waiting time exceeds the time limit. No decision is made.

Therefore, the effective service time distribution is truncated, and its truncation point depends on the effective waiting time. It can be immediately seen that both the effective waiting time distribution is also defined on $[0, L]$.

Figures 19, 20 and 21 illustrate some possible distributions of the effective service time, waiting time and total time in system obtained from simulation. The arrival is a Poisson process, and the original service time has a density function as (1). It is clear that all three distributions are truncated at $L$, and this is the reason that both the effective service time and time in system distributions have a spike at $L$. The reason that there is a spike at time 0 in the effective waiting time distribution may be because of the time limit, some products leave Station 2 before the inspection is completed, which frees the operators and thus the successive products can receive inspection upon they arrive. The effective time in system has zero probability within 0 and $a$ (the constant in the original service time distribution) is because the effective time in system includes the effective service time and the effective waiting time, and if the waiting time is zero, the service time will be greater than $a$ according to the original service time distribution.

![Figure 19. Distribution of Station-2 Effective Service Time with Decision Time Limit](Image)
4.2.2 Model Formulation

4.2.2.1 Assumption

The following assumptions are made in the analysis:

- Arrivals to Station 2 follow a Poisson process with parameter $\lambda$
- $n$ operators are working in parallel at Station 2, and they are independent
- Operators' service times follow the same distribution
4.2.2.2 Model: CPE Service Time, Multiple Servers

Since the original service time distribution is approximately a constant plus exponential (CPE) distribution, we shall use a CPE distribution as an approximation for it in the analysis, then the service time has the density function as

\[ w_s(x) = \begin{cases} \mu e^{-\mu(x-a)} & x \geq a \\ 0 & x < a \end{cases} \]  

(2)

where

\[ \mu = \frac{1}{\alpha \beta} \]  

(3)

and \( a, \alpha, \beta \) are the parameters in the original service time distribution \( f(x) \). Based on the above assumptions, Station 2 is an M/G/n queuing system.

As shown in Section 4.2, the effective service time and waiting time are dependent on each other, and both are dependent on the time limit \( L \). Even if the original service time is exponentially distributed, it is impossible to obtain a closed form of the actual distribution for the effective service time and waiting time.

Consider the effective service rate, \( X \), so \( 1/X \) is the actual expected time spent on each product. Let \( A \) be the expected total time not spent on products because of time out, then

\[ \frac{1}{X} = \frac{1}{\mu} + a - A \]  

(4)

where \( \mu \) and \( a \) are the parameters in the function given by (2), and

\[ A = E(W_t - (L - W_q) | W_q < L) P(W_q < L) + \left( \frac{1}{\mu} + a \right) P(W_q \geq L) \]  

(5)

To simplify the analysis, we use the waiting time distribution of an M/M/n system as an approximation for the waiting time distribution here. Gross and Harris derived the cumulative distribution function and density function of the waiting time. For an M/M/n system with each server's service rate \( X \), the waiting time has a probability distribution as
\[ W_q(t) = \begin{cases} \frac{1 - \frac{n(\lambda / X)^n}{n! (n - \lambda / X)} P_0}{(n - 1)! (n - \lambda / X)} p_0 + W_q(0) & (t > 0) \\ \frac{1 - \frac{n(\lambda / X)^n}{n! (n - \lambda / X)} P_0}{(n - 1)! (n - \lambda / X)} p_0 + W_q(0) & (t = 0) \end{cases} \]

and a density function as

\[ w_q(t) = \begin{cases} \frac{1 - \frac{n(\lambda / X)^n}{n! (n - \lambda / X)} P_0}{(n - 1)! (n - \lambda / X)} p_0 X e^{-(\alpha X - \lambda) t} & (t = 0) \\ \frac{1 - \frac{n(\lambda / X)^n}{n! (n - \lambda / X)} P_0}{(n - 1)! (n - \lambda / X)} p_0 X e^{-(\alpha X - \lambda) t} & (t > 0) \end{cases} \]

where

\[ p_0 = \left[ \sum_{k=0}^{n-1} \frac{1}{k!} \left( \frac{\lambda}{X} \right)^k + \frac{1}{n!} \left( \frac{\lambda}{X} \right)^n \left( \frac{nX}{nX - \lambda} \right) \right]^{-1} \]

The density function of the effective service time is still given by (2), then \( A \) is calculated as

\[ A = \left[ \frac{1}{\mu} + \frac{X(\lambda / X)^n}{(n - 1)! (nX - \lambda) (nX - \mu - \lambda)} P_0 e^{-\mu(t - \alpha)} - \frac{((\alpha + 1) (nX - \lambda) + \mu) X (\lambda / X)^n}{(n - 1)! (nX - \lambda)^2} P_0 e^{-(\alpha X - \lambda) L} \right] \]

Comparing (9) with (4) gives

\[ \left[ \frac{1}{\mu} + \frac{X(\lambda / X)^n}{(n - 1)! (nX - \lambda) (nX - \mu - \lambda)} P_0 e^{-\mu(t - \alpha)} - \frac{((\alpha + 1) (nX - \lambda) + \mu) X (\lambda / X)^n}{(n - 1)! (nX - \lambda)^2} P_0 e^{-(\alpha X - \lambda) L} \right] \]

\[ \frac{\mu X (\lambda / X)^n}{(n - 1)! (nX - \lambda)^2 (nX - \mu - \lambda)} P_0 e^{-(\alpha X - \lambda) L} - \frac{1}{\mu} - \alpha \]

\[ \right] = 0 \]

\[ X t \]

1 Detailed calculation is given in Appendix 2.
The above nonlinear equation can be solved using mathematical software. We solved it using Excel Solver by treating the absolute value of the left-hand-side as a function to be minimized and \( X \) as the decision variable with nonnegative constraint. After the effective service rate \( X \) is obtained, the percentage of time-out, \( \Phi(L) \), can be calculated by the following formula\(^2\).

\[
\Phi(L) = P(W_s \geq L - W_q | W_q < L)P(W_q < L) + P(W_q \geq L) \\
= \left( 1 + \frac{\mu X(\lambda / X)^n}{(n-1)! (nX - \lambda)(nX - \mu - \lambda)} \right) p_0 e^{-\mu(L-a)} \\
- \frac{X(\lambda / X)^n}{(n-1)! (nX - \lambda)} p_0 e^{-\lambda L} \\
- \frac{\mu X(\lambda / X)^n}{(n-1)! (nX - \lambda)(nX - \mu - \lambda)} p_0 e^{-\lambda(L-a)} \\
* \left( 1 - \frac{X(\lambda / X)^n}{(n-1)! (nX - \lambda)} p_0 e^{-\lambda L} \right) + \frac{X(\lambda / X)^n}{(n-1)! (nX - \lambda)} p_0 e^{-\lambda(L-a)}
\]

4.2.3 Numerical Results

To evaluate the accuracy of the theoretical result for the percentage of time-out given by (11), we compare the numerical results with the simulation results where the original service time distribution is used. The simulation results were obtained by running 20 replications for each scenario. Table 4 shows this comparison using \( \mu = 8.4, \alpha = 0.3 \) and some common product volumes experienced at Station 2 as well as a number of servers. For each scenario, the average, minimum and maximum values of 20 replications are provided. In most cases, the range of each scenario result is small. Figure 22 also depicts this comparison.

Both Table 4 and Figure 22 indicate that generally the theoretical results slightly overestimate the percentage of time-out, and the difference between the theoretical results and the simulation results tends to decrease as the number of servers goes up. Overall, the theoretical result is either within the range of the simulation result or close to the maximum of the corresponding simulation result. Therefore, the queuing theory approximation might be good enough for estimating the percentage of time-out at Station 2, and it gives an upper bound for the number of Station-2 operators required to attain a target percentage of time-out.

\(^2\) Detailed calculation is given in Appendix 3.
Table 4. Percentage of Time-out: Theoretical Result versus Simulation Result
(CPE versus CPG)

<table>
<thead>
<tr>
<th>Product Volume</th>
<th># of servers</th>
<th>Simulation</th>
<th></th>
<th>Calculation</th>
<th>Relative Difference</th>
<th>Absolute Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>Min</td>
<td>Max</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>25.0%</td>
<td>21.47%</td>
<td>27.19%</td>
<td>27.1%</td>
<td>8.5%</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>12.4%</td>
<td>11.45%</td>
<td>13.47%</td>
<td>13.9%</td>
<td>12.0%</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>41.2%</td>
<td>38.86%</td>
<td>43.33%</td>
<td>40.6%</td>
<td>-1.4%</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>16.1%</td>
<td>15.07%</td>
<td>17.78%</td>
<td>19.0%</td>
<td>17.8%</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>12.0%</td>
<td>10.91%</td>
<td>13.07%</td>
<td>13.2%</td>
<td>10.2%</td>
</tr>
<tr>
<td>7.5</td>
<td>2</td>
<td>18.9%</td>
<td>17.35%</td>
<td>20.07%</td>
<td>22.6%</td>
<td>19.6%</td>
</tr>
<tr>
<td>7.5</td>
<td>3</td>
<td>12.7%</td>
<td>11.02%</td>
<td>14.17%</td>
<td>14.2%</td>
<td>11.6%</td>
</tr>
<tr>
<td>7.5</td>
<td>4</td>
<td>11.7%</td>
<td>10.24%</td>
<td>12.87%</td>
<td>12.5%</td>
<td>7.0%</td>
</tr>
<tr>
<td>7.5</td>
<td>5</td>
<td>11.4%</td>
<td>10.16%</td>
<td>12.40%</td>
<td>12.2%</td>
<td>6.9%</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>13.6%</td>
<td>12.45%</td>
<td>14.50%</td>
<td>15.6%</td>
<td>14.8%</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>11.8%</td>
<td>10.54%</td>
<td>12.71%</td>
<td>12.8%</td>
<td>8.6%</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>11.5%</td>
<td>10.43%</td>
<td>12.31%</td>
<td>12.3%</td>
<td>7.0%</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>11.4%</td>
<td>10.39%</td>
<td>12.21%</td>
<td>12.2%</td>
<td>7.0%</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>11.4%</td>
<td>10.46%</td>
<td>12.21%</td>
<td>12.2%</td>
<td>7.0%</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>11.4%</td>
<td>10.46%</td>
<td>12.21%</td>
<td>12.2%</td>
<td>7.0%</td>
</tr>
</tbody>
</table>

Using the calculations by (11), we obtain the relationship between the percentage of time-out and the number of servers at Station 2 for various product arrival rates as shown in Figure 23. The arrival rate of 30 is an extremely high rate that is never experienced by Station 2 in the inspection.
system, while others are some common product volumes to Station 2. The relationship in Figure 23 helps determine the number of Station-2 operators required for each product arrival rate given a target percentage of time-out. For example, if the target is the percentage of time-out does not exceed 20%, then the number of Station-2 operators required for the arrival rates 6, 7.5, 9, 10.5 and 30 are 2, 3, 3, 3, 6, respectively. This relationship also informs whether it is worth adding more operators to attain a minimum percentage of time-out. If adding one more operator makes very little difference in the percentage of time-out, then it is not worthwhile because it only reduces the utilization of operators. It can be seen that for the arrival rate 30, the percentage of time-out with 8 operators is very close to that with 9 operators, so this again justify the earlier result by simulation that the inspection system needs at most 8 Station-2 operators. Figure 23 also reveals that for normal arrival rates, Station 2 usually needs no more than 4 operators.

![Figure 23. Percentage of Time-out versus Number of Station 2 Servers for Different Product Volumes](chart.png)

4.2.4 Accuracy of Approximation versus Original Distribution

The Gamma distribution in the original service time distribution is very similar to an exponential distribution because the shape parameter, $\beta$, is close to 1. There comes a question that if it is still appropriate to use a CPE to approximate the service time distribution when the statistical distribution in the original service time distribution is not like an exponential. It has been shown in section 3.5.3 that the percentage of time-out is very sensitive to the type of Station 2 service time distribution. Figure 24 illustrates the similarity between an exponential distribution with a Gamma distribution having a shape parameter 1.14, and their difference from a Gamma
distribution having a shape parameter 2, where these distributions have the same mean. Table 5 shows a comparison of the simulation results for the CPG having the shape parameter of 2 and 1.14 with the theoretical result for the CPE, where the mean and the constant of three distributions are the same. The accuracy of approximation goes down as the shape parameter increases. It is clear that CPE is not a good approximation for a CPG if the Gamma distribution in the CPG is not like an exponential. Therefore, the accuracy of approximation is sensitivity to the type of service time distribution. However, since the theoretical result using a CPE approximation is generally greater than the simulation result for a CPG distribution, the theoretical results might give an upper bound for the percentage of time-out when the service time distribution is an arbitrary distribution and might be used as a worst case scenario.

![Figure 24. Cumulative Probability Distribution of Different Statistical Distributions](image)

**Table 5. Percentage of Time-out: Theoretical Result versus Simulation Result (CPE versus CPG, Gamma with Different Shape Parameters)**

<table>
<thead>
<tr>
<th>Shape Parameter</th>
<th>Simulation (CPG)</th>
<th>Theoretical (CPE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage Time-out</td>
<td>Percentage Time-out</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>Min</td>
</tr>
<tr>
<td>1.14</td>
<td>12.7%</td>
<td>11.0%</td>
</tr>
<tr>
<td>2</td>
<td>9.2%</td>
<td>8.5%</td>
</tr>
</tbody>
</table>
4.3 Station 3&4 Analysis

If we assume that the two Station-3 machines work independently, the average down time of a
machine is $D$ and the average time between breakdowns is $B$, the fraction of time that only one
machine is in operation while the other one breaks down is 8%, which is calculated using the
formula $(D/(B+D))*(B/(B+D))^2$ and the data for $B$ and $D$. This is not a negligible circumstance.
Similarly, the fraction of time that both machines are not working is $(D/(B+D))*(D/(B+D)) = 0.16\%$. Since this probability is small, we may ignore this circumstance to simplify our analysis.

Thus we assume that the inspection system has at least one Station 3 machine in operation at any
time. Notice that the assumption of independence is unlikely to be true in practice because one
breakdown will for certain place more load on the second machine. The question here is how
good can an approximation be, and is it good enough to give performance approximations
sufficient for the needs of system designers.

To ensure that the system uptime ratio is greater than 99.9%, it is required that one Station 3
machine be able to cope with the product arrivals from Station 2 when the other machine is down.
This in fact concerns the probability that the queue length in front of Station 3 exceeds the size of
the buffers between Station 2 and Station 3. Detailed analysis for this probability would consider
the transient behaviour at Station 3, but it is very complicated because the distribution of Station
3 service time is not exponential and Station 4 also needs to be taken into consideration. The fact
that the average down time is not short makes us consider the steady state of only one machine
working at Station 3. If we assume the product arrival to Station 3 is a Poisson process and the
service time of Station 3 (involving Station 4) has an exponential distribution, then it is an $M/M/1$
queue. Let $\lambda_3$ be the arrival rate to Station 3 and $\mu_3$ be the service rate of a Station 3 machine, $p_n$
be the probability that the queue length at Station 3 is $n$, and $N$ be the buffer size, then the
probability that the queue length at Station 3 is no more than the buffer size is given by

$$\sum_{n=0}^{N} p_n = \sum_{n=0}^{N} \left(1 - \frac{\lambda_3}{\mu_3}\right) \left(\frac{\lambda_3}{\mu_3}\right)^n = 1 - \left(\frac{\lambda_3}{\mu_3}\right)^{N+1}$$ (12)

Since a Station 4 operator works together with a Station 3 machine if there is a product not
approved by this machine, the actual service rate of one Station 3 machine without breakdown, $\mu_3$,
is calculated from:
\[
\frac{1}{\mu_3} = A_3 T_3 + \frac{1 - A_4}{\mu_1}
\]  

(13)

where \( A_3 \) is the approval rate of Station 3, \( T_3 \) is the average service time of a Station 3 machine without Station 4's involvement, and \( \mu_4 \) is the service rate of a Station-4 operator. It follows that the service rate of a Station 3 machine is 3.36 using the data for the parameters in (13).

To ensure the uptime ratio to be at least 99.9\%, the maximum arrival rate of Station 3 satisfies

\[
\sum_{n=0}^{N} p_n = 1 - \left( \frac{\lambda_3}{\mu_3} \right)^{N+1} \geq 99.9\%
\]

Using \( \mu_3 = 3.36 \) and \( N = 31 \), \( \lambda_3 \) is calculated by the above formula to be 2.71.

It can be seen that if the arrival rate of Station 3 is within 2.71, one Station 4 operator is enough to ensure an uptime ratio to be greater than 99.9\% because when one machine breaks down, the other machine in operation only needs one operator. However, if the arrival rate is greater than 2.71, then we must allocate two operators to Station 4 so that the products held up in front of Station 3 due to one machine’s breakdown can be processed more quickly after the machine goes back into operation.

### 4.4 Comparison of Queuing Theory Result with Simulation Result

Based on the above analysis, the procedure for determining the staffing levels at Station 2 and Station 4 is as follows.

1. Calculate the arrival rate of Station 2 by \((1 - A_1) \lambda_1\), where \( A_1 \) and \( \lambda_1 \) are the approval rate of Station 1 and the arrival rate of the inspection system, respectively.

2. Choose the smallest number of operators at Station 2, \( n \), using formulas (10) and (11) such that

\[
(1 - A_1) \lambda_1 \left[ (1 - \overline{\Phi}(L)) (1 - A_3) + \overline{\Phi}(L) \right] \leq 2.71
\]

(14)

where \( \overline{\Phi}(L) \) is the percentage of products timed out at Station 2.

3. If there is such \( n \) satisfying (14), then allocate \( n \) operators to Station 2 and one operator to Station 4; otherwise, choose an \( n \) such that adding one more operator does not improve
much on the percentage of time-out, $\Phi(L)$, and allocate $n$ operators to Station 2 and two operators to Station 4.

Table 6. Scenario Analysis: Theoretical Result versus Simulation Result

<table>
<thead>
<tr>
<th>Arrival Rate of System</th>
<th>Arrival Rate of Station 2</th>
<th># of Operators</th>
<th>% Time-out at Station 2</th>
<th>Arrival Rate of Station 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Simulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Average</td>
<td>Min</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>1</td>
<td>17.1%</td>
<td>14.5%</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>1</td>
<td>25.0%</td>
<td>21.5%</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2</td>
<td>12.4%</td>
<td>11.4%</td>
</tr>
<tr>
<td>15</td>
<td>4.5</td>
<td>1</td>
<td>32.6%</td>
<td>30.8%</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>1</td>
<td>41.2%</td>
<td>38.9%</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>2</td>
<td>16.1%</td>
<td>15.1%</td>
</tr>
<tr>
<td>25</td>
<td>7.5</td>
<td>2</td>
<td>18.9%</td>
<td>17.4%</td>
</tr>
<tr>
<td>25</td>
<td>7.5</td>
<td>3</td>
<td>12.7%</td>
<td>11.0%</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>3</td>
<td>13.6%</td>
<td>12.4%</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>4</td>
<td>11.8%</td>
<td>10.5%</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>5</td>
<td>11.5%</td>
<td>10.4%</td>
</tr>
<tr>
<td>35</td>
<td>10.5</td>
<td>3</td>
<td>14.5%</td>
<td>13.6%</td>
</tr>
<tr>
<td>35</td>
<td>10.5</td>
<td>4</td>
<td>12.0%</td>
<td>11.3%</td>
</tr>
<tr>
<td>35</td>
<td>10.5</td>
<td>5</td>
<td>11.5%</td>
<td>10.7%</td>
</tr>
<tr>
<td>40</td>
<td>12</td>
<td>4</td>
<td>12.5%</td>
<td>11.4%</td>
</tr>
<tr>
<td>40</td>
<td>12</td>
<td>5</td>
<td>11.6%</td>
<td>10.9%</td>
</tr>
<tr>
<td>40</td>
<td>12</td>
<td>6</td>
<td>11.4%</td>
<td>10.8%</td>
</tr>
</tbody>
</table>

Table 6 compares the scenario analysis results using the steps 1 and 2 above with the corresponding simulation results for a number of product volumes. The simulation results were obtained by running 20 replications for each scenario. For each scenario, the average, minimum and maximum values of 20 replications are provided. In most cases, the range of each scenario result is small. Generally, the theoretical result for the percentage of time-out is either greater than or within the range of the simulation result. Figure 25 compares the arrival rates of Station 3 calculated from the percentage of time-out given by the simulation and by the queuing
approximation. It is clear that the theoretical results are very close to the simulation result, and thus provides a reasonable approximation.

![Graph showing Station 3 Arrival Rate: Simulation Result versus Theoretical Result](image)

Figure 25. Station 3 Arrival Rate: Simulation Result versus Theoretical Result

Table 7 compares the number of operators required for Station 2 and Station 4 determined using the above method, together with the simulation result.

<table>
<thead>
<tr>
<th>Arrival rate of system</th>
<th># of Station 2 operators</th>
<th># of Station 4 operators</th>
<th>Total # of operators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation</td>
<td>Calculation</td>
<td>Calculation</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

Generally, the theoretical result is close to the simulation result. For some arrival rates, the theoretical analysis suggests to allocate one more operator to Station 2 instead of to Station 4 in contrast to the simulation result. However, as discussed in Section 4.1, the simulation result often
gives very close uptime ratios for two scenarios that differ by one operator at either Station 2 or Station 4. Therefore, the staffing levels determined by simulation may not be optimal, and the theoretical result may give a better result. We can also conclude that the queuing model is a good approximation to the complicated inspection system.
5. Conclusions

5.1 Summary
In this thesis we have applied simulation techniques and queuing theory to develop a methodology for determining the staffing levels for a product inspection system. First, a simulation model was built for the staffing analysis. System uptime ratio was chosen as the service level criteria for determining the required staffing levels. Required staffing levels for a range of product volumes were determined using the simulation model. A user interface and a program for scenario analysis were developed in Visual Basic to make it more convenient to use the simulation model.

In addition, we developed an analytical model using queuing theory and approximation. It was proved to be a reasonable approximation to the real inspection system. It also overcomes the problem of choosing optimal staffing levels by simulation when the uptime ratios of two scenarios are very close. The queuing model can be used to estimate the required staffing level for a given product volume. All results of the queuing model can be obtained using Microsoft Excel.

It should be noted that the percentage of time-out at Station 2 is sensitivity to the type of service time distribution and the mean of service time. Therefore, accurate data of Station 2 service time is very important in the staffing analysis and the queuing approximation is appropriate only if the Station 2 service time distribution is close to a constant-plus-exponential, but it may provide an upper bound for the required staffing levels.

As a result, we suggest that both the simulation model and the queuing model be used in implementation to help make a more accurate decision.

5.2 Future Research
To develop an analytical model for estimating the percentage of products timed-out at Station 2 due to time limit, we have used a constant-plus-exponential distribution to approximate the constant-plus-Gamma distribution of the service time. We have shown that the accuracy of approximation is sensitive to the type of the statistical distribution in the service time distribution. Since the percentage of time-out is also determined by the time limit and the constant in the
service time distribution, we could investigate the sensitivity of approximation accuracy to these factors.

We have used the result of steady state to estimate the probability of system congestion at Station 3. For only two machines at Station 3, the approximation is easy and is proved to be good. If there were more machines at Station 3, the approximation method might not work well. It would be interesting to analyze the transient behaviour of Station 3 with finite waiting space to obtain a closer result. Approximation for more machines could also be investigated and tested against the simulation result.
Bibliography


Appendix 1. Visual Basic Code for Simulation User Interface

- **Read parameter values from the simulation model:**
  (\text{Al} \text{ is a variable in the simulation model})

  Private Sub readpara()
  Dim myidx As Long

  myidx = s.SymbolNumber("\text{Al}") \text{ Determine index of parameter Al}
  UserFormParameter.TextBox1.value = s.VariableArrayValue(myidx)
  \text{ Assign the value of parameter Al to the TextBox1 of the user interface form}

  ... \text{ Read other parameters and assign the values to the user interface form}
  End Sub

- **Assign model parameters (variable) by user-specified new values:**

  Private Sub AcceptPara()
  Dim myidx As Integer

  myidx = s.SymbolNumber("\text{Al}") \text{ Determine index of parameter Al}
  s.VariableArrayValue(myidx) = para1
  \text{ Assign the value of variable para1 to parameter Al}

  ... \text{ Assign other parameters}
  End Sub

- **Assign model parameters (resource capacity) by user-specified new values:**

  Private Sub AcceptCapValue()
  Dim i, j As Integer


For i = 1 To capLevel1  'Assign the capacity of resource1 to be capLevel1
    s.ResourceCapacity(s.SymbolNumber("resource1." & i)) = 1
Next i

For i = capLevel1 + 1 To CAP1AMOUNT
    s.ResourceCapacity(s.SymbolNumber("resource1." & i)) = 0
Next i

...  'Assign other resources

End Sub

• **Update model parameters by user-specified new values:**

  Private Sub ModelLogic_RunBeginReplication()
      AcceptPara
  End Sub

• **Show user interface at the beginning of simulation run:**

  Private Sub ModelLogic_RunBeginSimulation()
      UserFormParameter.Show
  End Sub

• **Release memory at the end of simulation run:**

  Private Sub ModelLogic_RunEnd()
      CleanGlobalModel
  End Sub

• **Store values entered by user in variables:**

  Private Sub StorePara()
para(1) = Me.TextBox1.value
para(2) = Me.TextBox2.value

' Store other values

capLevel1 = Me.CBoxCapLevl.value

' Store other resource capacities

Unload Me
End Sub

- Read resource capacity from simulation model and show in the list box of user interface:

Private Sub LoadCapacity()
Dim i As Integer
Dim arrCapValue() As Variant
Dim capLevel1 As Integer

' Define other variables

capLevel1 = 0

' Initialize other variables

arrCapValue = ReadCapValue()

' Calculate resource capacity
For i = 1 To CAP1AMOUNT
Me.CBoxCapLevl.AddItem i
    capLevel1 = capLevel1 + arrCapValue(i)
Next i

Me.CBoxCapLevl.ListIndex = capLevel1 - 1

End Sub
'Read resource capacity from simulation model
Private Function ReadCapValue() As Variant
    Dim arrCapValue(1 To CAPAMOUNT)
    Dim i As Integer

    'read resource 1
    For i = 1 To CAP1AMOUNT
        arrCapValue(i) = s.ResourceCapacity(s.SymbolNumber("resource1." & i))
    Next i

    ' Read other resource capacities
    ReadCapValue = arrCapValue
End Function

- Accept values entered by user:
  Private Sub Accept_Click()
      StorePara
  End Sub

- Move cursor to the next TextBox after user presses Enter:
  Private Sub TextBox1_KeyPress(ByVal KeyAscii As MSForms.ReturnInteger)
      If KeyAscii = 13 Then
          Me.TextBox2.SetFocus
      End If
  End Sub
Private Sub TextBox12_KeyPress(ByVal KeyAscii As MSForms.ReturnInteger)
    If KeyAscii = 13 Then
        Me.Accept.SetFocus
    End If
End Sub

• **Activate & terminate user interface:**

Private Sub UserForm_Activate()
    InitializeGlobalModel
    readpara
    LoadCapacity
    Me.TextBox1.SetFocus
End Sub

Private Sub UserForm_Terminate()
    'CleanGlobalModel
End Sub

• **Define global variables:**

Public m As Model 'Arena model
Public s As SIMAN 'Arena SIMAN

Public para(10) As Double
Public capLevel1 As Integer

... 'Define other resource capacity variables

Public Const CAP1AMOUNT As Integer = 2
... 'Define other constants

'The mean of service time and machine down time
Public Const S2mean As Double =
Public Const S3duration As Double =
Public Const S3timebt As Double =

- **Initialize SIMAN object of simulation:**
  
  Public Sub InitializeGlobalModel()
     Set m = ThisDocument.Model
     Set s = m.SIMAN
  End Sub

- **Clean SIMAN object:**
  
  Public Sub CleanGlobalModel()
     Set s = Nothing
     Set m = Nothing
  End Sub
Appendix 2. Expected Total Time not Spent on Products because of Time-out

The expected total time not spent on products because of time-out, $A$, is given by:

$$ A = E(W_q - (L - W_q) \mid W_q < L) \, P(W_q < L) + \frac{1}{\mu} \, P(W_q \geq L) $$

The probability distribution and probability density function of the waiting time in an $M/M/n$ system with each server’s service rate $X$, are given by (6), (7) and (8). It follows that

$$ E(W_q - (L - W_q) \mid W_q < L) $$

$$ = \int_0^L \int_{-(L-t)}^0 (y - (L-t)) \, \mu e^{-\mu(y-a)} \, dy \, dW_q(t) $$

$$ = \left[ \int_{-(L-t)}^0 (y - (L-t)) \, \mu e^{-\mu(y-a)} \, dy \right] \left( 1 - \frac{n(\lambda / X)^n}{n!(n - \lambda / X)} \, p_0 \right) $$

$$ + \int_{-(L-t)}^0 W_q(t) \int_{-(L-t)}^0 (y - (L-t)) \, \mu e^{-\mu(y-a)} \, dy \, dt $$

$$ = (1 - \frac{n(\lambda / X)^n}{n!(n - \lambda / X)} \, p_0) \frac{1}{\mu} \, e^{-\mu(L-a)} $$

$$ + \int_{-(L-t)}^0 (\lambda / X)^n \, X e^{-(nX-\lambda)t} \, p_0 \int_{-(L-t)}^0 (y - (L-t)) \, \mu e^{-\mu(y-a)} \, dy \, dt $$

$$ + \int_{-(L-t)}^0 (\lambda / X)^n \, X e^{-(nX-\lambda)t} \, p_0 \int_{-(L-t)}^0 (y - (L-t)) \, \mu e^{-\mu(y-a)} \, dy \, dt $$

$$ = (1 - \frac{n(\lambda / X)^n}{n!(n - \lambda / X)} \, p_0) \frac{1}{\mu} \, e^{-\mu(L-a)} + \int_{-(L-t)}^0 (\lambda / X)^n \, X e^{-(nX-\lambda)t} \, p_0 \frac{1}{\mu} \, e^{-\mu(L-L-a)} \, dt $$

$$ + \int_{-(L-t)}^0 (\lambda / X)^n \, X e^{-(nX-\lambda)t} \, p_0 (a + \frac{1}{\mu} - L + t) \, dt $$

$$ = (1 - \frac{n(\lambda / X)^n}{n!(n - \lambda / X)} \, p_0) \frac{1}{\mu} \, e^{-\mu(L-a)} + \frac{X(\lambda / X)^n}{(n-1)!} \, p_0 \frac{1}{\mu} \, e^{-(nX-\lambda)t} \, e^{-(nX-\lambda)\mu} \, dt $$

$$ + \frac{X(\lambda / X)^n}{(n-1)!} \, p_0 \left[ (a + \frac{1}{\mu} - L) \int_{-(L-t)}^0 e^{-(nX-\lambda)t} \, dt + \int_{-(L-t)}^0 e^{-(nX-\lambda)t} \, dt \right] $$

$$ = (1 - \frac{X(\lambda / X)^n}{(n-1)!(nX-\lambda)} \, p_0) \frac{1}{\mu} \, e^{-\mu(L-a)} $$

$$ + \frac{X(\lambda / X)^n}{\mu(n-1)!(nX-\mu-\lambda)} \, p_0 \left( e^{-\mu(L-a)} - e^{-(nX-\lambda)(L-a)} \right) $$

49
\[
+ \frac{X(\lambda / X)^n}{(n-1)!} \left[ \frac{nX - \lambda + \mu}{\mu(nX - \lambda)^2} e^{-(nX - \lambda)(L-a)} - \frac{(a\mu + 1)(nX - \lambda) + \mu}{\mu(nX - \lambda)^2} e^{-(nX - \lambda)L} \right] \\
= \left( \frac{1}{\mu} + \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)(nX - \mu - \lambda)} \right) p_0 e^{-\mu(L-a)} \\
- \left( \frac{(a\mu + 1)(nX - \lambda) + \mu}{\mu(n-1)!(nX - \lambda)^2} X(\lambda / X)^n \right) p_0 e^{-(nX - \lambda)L} \\
- \frac{\mu X(\lambda / X)^n}{(n-1)!(nX - \lambda)^2(nX - \mu - \lambda)} p_0 e^{-(nX - \lambda)(L-a)} \\
\]

and

\[
P(W_1 < L) = 1 - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L} \\
\]

thus

\[
A = \left[ \left( \frac{1}{\mu} + \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)(nX - \mu - \lambda)} \right) p_0 e^{-\mu(L-a)} \\
- \left( \frac{(a\mu + 1)(nX - \lambda) + \mu}{\mu(n-1)!(nX - \lambda)^2} X(\lambda / X)^n \right) p_0 e^{-(nX - \lambda)L} \\
- \frac{\mu X(\lambda / X)^n}{(n-1)!(nX - \lambda)^2(nX - \mu - \lambda)} p_0 e^{-(nX - \lambda)(L-a)} \right] \\
* \left( 1 - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L} \right) \\
+ \left( \frac{1}{\mu} + a \right) \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L} \\
\]

\[
= \left[ \left( \frac{1}{\mu} + \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)(nX - \mu - \lambda)} \right) p_0 e^{-\mu(L-a)} \\
- \left( \frac{(a\mu + 1)(nX - \lambda) + \mu}{\mu(n-1)!(nX - \lambda)^2} X(\lambda / X)^n \right) p_0 e^{-(nX - \lambda)L} \\
- \frac{\mu X(\lambda / X)^n}{(n-1)!(nX - \lambda)^2(nX - \mu - \lambda)} p_0 e^{-(nX - \lambda)(L-a)} - \frac{1}{\mu} - a \right] \\
* \left( 1 - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L} \right) + \frac{1}{\mu} + a \\
\]

50
Appendix 3. Percentage of Time-out at Station 2 due to Decision Time Limit

The percentage of products timed-out at Station 2 due to the decision time limit, $\Phi(L)$, can be calculated by the following formula.

$$\Phi(L) = P(W_s \geq L - W_q | W_q < L)P(W_q < L) + P(W_q \geq L)$$

where the first conditional probability

$$P(W_s \geq L - W_q | W_q < L) = \int_0^L \int_{-\infty}^0 \mu e^{-\mu(y-a)} dy dW_q(t)$$

$$= \left[ \int_0^L \mu e^{-\mu(y-a)} dy \right] \left( 1 - \frac{n(\lambda / X)^n}{n!(n - \lambda / X)} p_0 \right) + \int_{-\infty}^0 \int_L^\infty \mu e^{-\mu(y-a)} dy dt$$

$$= \left( 1 - \frac{n(\lambda / X)^n}{n!(n - \lambda / X)} p_0 \right) e^{-\mu(L-a)} + \int_{-\infty}^0 \left( \frac{\lambda / X}{n-1} \right) X e^{-(nX-L)} \frac{p_0}{n!(n-1)!} \int_{-\infty}^0 \mu e^{-\mu(y-a)} dy dt$$

$$+ \int_0^L (\lambda / X)^n X e^{-(nX-L)} \frac{n!(n-1)!}{n!(n-1)!} \int_{-\infty}^0 \mu e^{-\mu(y-a)} dy dt$$

$$= \left( 1 - \frac{n(\lambda / X)^n}{n!(n - \lambda / X)} p_0 \right) e^{-\mu(L-a)} + \frac{X(\lambda / X)^n}{(n-1)!} \frac{p_0}{n!(n-1)!} \int_{-\infty}^0 e^{-(nX-L)} \mu e^{-\mu(y-a)} dy dt$$

$$+ \left( X(\lambda / X)^n \frac{(n-1)!}{(n-1)!} \frac{p_0}{(nX-L)} \int_{-\infty}^0 e^{-(nX-L)} \right)$$

$$= \left( 1 - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 \right) e^{-\mu(L-a)} + \frac{X(\lambda / X)^n}{(n-1)!(nX - \mu - \lambda)} p_0 e^{-\mu(L-a)}$$

$$- e^{-(nX-L)(L-a)} + \frac{X(\lambda / X)^n}{(n-1)!} p_0 \frac{1}{nX - \lambda} (e^{-(nX-L)(L-a)} - e^{-(nX-L)L})$$

$$= \left( 1 + \frac{\mu X(\lambda / X)^n}{(n-1)!(nX - \lambda)(nX - \mu - \lambda)} p_0 \right) e^{-\mu(L-a)} - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX-L)L}$$

$$- \frac{\mu X(\lambda / X)^n}{(n-1)!(nX - \lambda)(nX - \mu - \lambda)} p_0 e^{-(nX-L)(L-a)}$$

and the second probability
\[ P(W_q < L) = 1 - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L} \]

It follows that

\[ \Phi(L) = ((1 + \frac{\mu X(\lambda / X)^n}{(n-1)!(nX - \lambda)(nX - \mu - \lambda)} p_0 e^{-\mu(L-n)}) - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L})(1 + \frac{\mu X(\lambda / X)^n}{(n-1)!(nX - \lambda)(nX - \mu - \lambda)} p_0 e^{-\mu(L-n)}) \]

\[ \Phi(L) = ((1 - \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L}) + \frac{X(\lambda / X)^n}{(n-1)!(nX - \lambda)} p_0 e^{-(nX - \lambda)L}) \]