4-22-2011

A Theory of Constraints Service Systems Improvement Method: Case of the Airline Turnaround Problem

Steven C. Ellis
Florida International University, ellisc@fiu.edu

DOI: 10.25148/etd.FI11051601

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

A THEORY OF CONSTRAINTS SERVICE SYSTEMS IMPROVEMENT METHOD: CASE OF THE AIRLINE TURNAROUND PROBLEM

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

INDUSTRIAL AND SYSTEMS ENGINEERING

by

Steven C. Ellis

2011

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Steven C. Ellis, and entitled A Theory of Constraints Service Systems Improvement Method: Case of the Airline Turnaround Problem, having been approved with respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Chin-Sheng Chen

Shih-Ming Lee

Christos Koulamas

Ronald Giachetti, Major Professor

Date of Defense: April 22, 2011

The dissertation of Steven C. Ellis is approved.

Dean Amir Mirmiran
College of Engineering and Computing

Interim Dean Kevin O'Shea
University Graduate School

Florida International University, 2011

DEDICATION

I dedicate this work to my family. Without their sacrifice, understanding, support, and most of all love, the completion of this work would not have been possible.

ACKNOWLEDGMENTS

I wish to thank the members of my committee for their support and patience as I labored to organize this work. Our associations have been rewarding both personally and professionally. I gratefully acknowledge Dr. Ron Giachetti for his support and advice, Marcelo Alvarado, Debra Vandermeer, and Monica Tremblay for the support and encouragement. Thanks to Mike Anderson at the Fort Lauderdale airport for the data and access to ramp operations to be able to complete the case study.

I have found my coursework throughout the curriculum and instruction program to be stimulating and thoughtful, providing me with grist for the intellectual mill of ongoing process improvement; enough to last a lifetime of investigation and reflection.

ABSTRACT OF THE DISSERTATION

A THEORY OF CONSTRAINTS SERVICE SYSTEMS IMPROVEMENT METHOD:

CASE OF THE AIRLINE TURNAROUND PROBLEM

by

Steven C. Ellis

Florida International University, 2011

Miami, Florida

Professor Ronald Giachetti, Major Professor

This dissertation develops a process improvement method for service operations based on the Theory of Constraints (TOC), a management philosophy that has been shown to be effective in manufacturing for decreasing WIP and improving throughput. While TOC has enjoyed much attention and success in the manufacturing arena, its application to services in general has been limited. The contribution to industry and knowledge is a method for improving global performance measures based on TOC principles. The method proposed in this dissertation will be tested using discrete event simulation based on the scenario of the service factory of airline turnaround operations. To evaluate the method, a simulation model of aircraft turn operations of a U.S. based carrier was made and validated using actual data from airline operations. The model was then adjusted to reflect an application of the Theory of Constraints for determining how to deploy the scarce resource of ramp workers. The results indicate that, given slight modifications to TOC terminology and the development of a method for constraint identification, the Theory of Constraints can be applied with success to services. Bottlenecks in services must be defined as those processes for which the process rates

and amount of work remaining are such that completing the process will not be possible without an increase in the process rate. The bottleneck ratio is used to determine to what degree a process is a constraint. Simulation results also suggest that redefining performance measures to reflect a global business perspective of reducing costs related to specific flights versus the operational local optimum approach of turning all aircraft quickly results in significant savings to the company. Savings to the annual operating costs of the airline were simulated to equal 30% of possible current expenses for misconnecting passengers with a modest increase in utilization of the workers through a more efficient heuristic of deploying them to the highest priority tasks. This dissertation contributes to the literature on service operations by describing a dynamic, adaptive dispatch approach to manage service factory operations similar to airline turnaround operations using the management philosophy of the Theory of Constraints.

TABLE OF CONTENTS

CHAPTER		PAGE
1	INTRODUCTION…….	1
	Background/Personal Motivation……..	1
	The Focus of the Theory of Constraints……	4
	Contribution and Significance……..	8
	Research Method……..	10
	Organization…….	12
2	LITERATURE REVIEW…….	13
	The Theory of Constraints……	13
	Research on Turnaround Operations……	19
	Research on Measurement of System Status……	20
	Literature Review Summary…….	23
3	THE IMPROVEMENT METHOD……	25
	Defining a Bottleneck in Services……..	26
	The Logic of the Method……	29
	Inputs Needed for the Logic of the Method……..	29
	Characteristics of a Service Factory Wherein This Method Will Apply……..	30
	Long term vs. Short Term View……	33
4	CASE STUDY DESCRIPTION AND MODEL FORMULATION……	35
	Mapping the Service System of Aircraft Turns……	35
	Overview of this Case Study……	35
	Scope of the study……	43
	Narrative of the Process of Turning a Flight……	45
	Data Collection for the Turnaround Process…….	48
	Characterizing this Service System……..	59
	All Flights are NOT of Equal Importance……	61
	The As-Is Model……	62
	Validation of the As-Is Model……..	70
	Applying the Theory of Constraints to Formulate the To-Be Model……..	75
	Determining the Number of Ramp Workers Assigned to a Task……	84
5	ANALYSIS AND DISCUSSION…….	90
	Performance Measures and Terms……...	90
	Number of Ramp Workers Assigned to Each Task…….	90
	Paired Samples Statistics of Each Model…….	92
	Comparing Costs per Flight Controlling for Hub Destination Flights Only……	99

Optimistic Turn Times…….. 100
Critically Late Flights…… 100
Cost this Flight…….. 100
Man Minutes and Worker Compensation……. 101

6 CONCLUSIONS…… 102
Summary……. 102
Contributions and Significance…….. 103
Requirements for the Method…… 103
Limitations……. 104
The Theory of Constraints is Applicable to Services…… 105
Service Improvement Method…… 106
Future Research…….. 107

REFERENCES…….. 109

VITA……. 113

LIST OF TABLES

TABLE	PAGE
1 - Isomorphism between TOC and PERT/CPM…….	22
2 - Processes, Datasets, and Observed Data…….	58
3 - Flight Attributes, Descriptions, and Mathematical Expressions……	65
4 - Variables and their Expressions for the As-Is Model…….	68
5 - Processing Rates Given the Number of Ramp Workers…….	69
6 - Decision Tree for Bottleneck Ratios……...	87
7 - Confidence intervals for the Mean Savings Given Worker Compensation……	101

LIST OF FIGURES

FIGURE	PAGE
1 - The Outline of TOC from (Mabin and Balderstone, 2000)…….....	3
2 - Integrated TOC and Six Sigma Framework (Ike Ehie 2005)…….....	17
3 - Service Process Matrix…….....	18
4 - Improvement Method Using the Bottleneck Ratio…….....	28
5 - Process Map of the Turnaround Process Contemplated in this Study…….....	46
6 - Distribution of Arrival Lateness…….....	50
7 - Gantt Chart of Work Schedule of Workers…….....	51
8 - Gantt Chart of Scheduled Turns Used in the Model…….....	63
9 - Modeling Arrival Lateness…….....	64
10 - Comparison of Turn Times…….....	72
11 - Actual vs. Simulated Arrival Lateness…….....	73
12 - Logic for the To-Be Arena Model in the Airline Case Study…….....	76
13 - Part 1 of the Logic for Selecting the Number of Ramp Workers…….....	84
14 - Part 2 of the Logic for Selecting the Number of Ramp Workers…….....	85
15 - Part 3 of the Logic for Selecting the Number of Ramp Workers…….....	86
16 - Number of Workers per Task…….....	91
17 - Difference in Task Processing Times…….....	93
18 - Costs per Flight Controlling for Hub Destination Flights Only…….....	99

LIST OF ACRONYMS

ETA	Estimated Time of Arrival
ATA	Actual Time of Arrival
ETD	Estimated Time of Departure
ATD	Actual Time of Departure
SFM	Synchronous Flow Manufacturing
TOC	Theory of Constraints
TSA	Transportation Safety Administration
CCR	Capacity Constraint Resource

1 INTRODUCTION

1.1 Background/Personal Motivation

The Theory of Constraints (TOC) has enjoyed a place of prominence in the minds of practitioners and challenged academics since its appearance in mainstream manufacturing thought in the mid 1980s. Though not a true theory, but rather a management philosophy artfully enunciated by its chief proponent and author, Eliyahu Goldratt, its application nonetheless, has been shown to improve cycle times, increase service levels, and decrease inventory levels in all manner of manufacturing industries (Mabin and Balderstone 2000). I was introduced to TOC in the Executive MBA program of Florida International University in 1995.

With its advent, TOC presented a revolutionary, intuitive way of thinking about production scheduling whose implementation would prove superior to MRP, the then currently accepted method. TOC requires a company to first define and then relentlessly focus on its goals and objectives, the policies that enable or constrain the attainment of those goals and objectives, and the measurements used by management to incentivize operations. The theory of constraints has also come to be known as "Management by Constraints", Synchronous Flow Manufacturing (SFM), and Synchronous Production (SP).

The theory of constraints was shown to have its origins in old and well-known lessons from project management (Dan Trietsch 2005). Dan Trietsch also suggests that Goldratt's main contribution was packaging in a way that captured the North American

manager's interest. The debate as to how much of the credit for the principles in the theory of constraints should be given to Goldratt is not at the heart of this research.

To illustrate the new improvement method for services based on the TOC, a case study of the aircraft turnaround process will be used. Airline ground operations consist of a series of service activities that must be done in some order such as deplaning inbound passengers, cleaning the plane, performing a safety check, and then boarding outbound passengers. The rate of processing for each of these activities varies due to the involvement of human workers and variability in the situations they face. In the teachings of Goldratt, problems arise because of the combination of two phenomena – dependent events and statistical fluctuations. In manufacturing, these are evidenced by two negative outcomes – decrease in throughput and an increase in work in process (WIP) or inventories. In services such as airline turns, the evidence of the combination of these two phenomena is seen as delays and the expenses that arise from those delays. I have personal experience with this process having spent two years in the employ of the now defunct Braniff Airlines and a year with Virgin Atlantic Airways while an engineering undergraduate student in the mid 80's doing and supervising exactly the things that I have studied in this research.

TOC has evolved to become a comprehensive framework for decision making consisting in operations strategy tools, a performance measurement system, and a set of thinking processes used to overcome resistance to change. As can be seen in Figure 1, TOC as a management philosophy has evolved to include Operations Strategy Tools with applications in Production Management, Distribution Management, and Project Management. A performance measurement system is defined adapted to the approach

and thinking processes are mapped to overcome resistance to adoption and implementation of the approach. Figure 1 shows the Five Focusing Steps as part of the Operations Strategy Tools. These are the steps that will be used to construct the method for process improvement in this study.

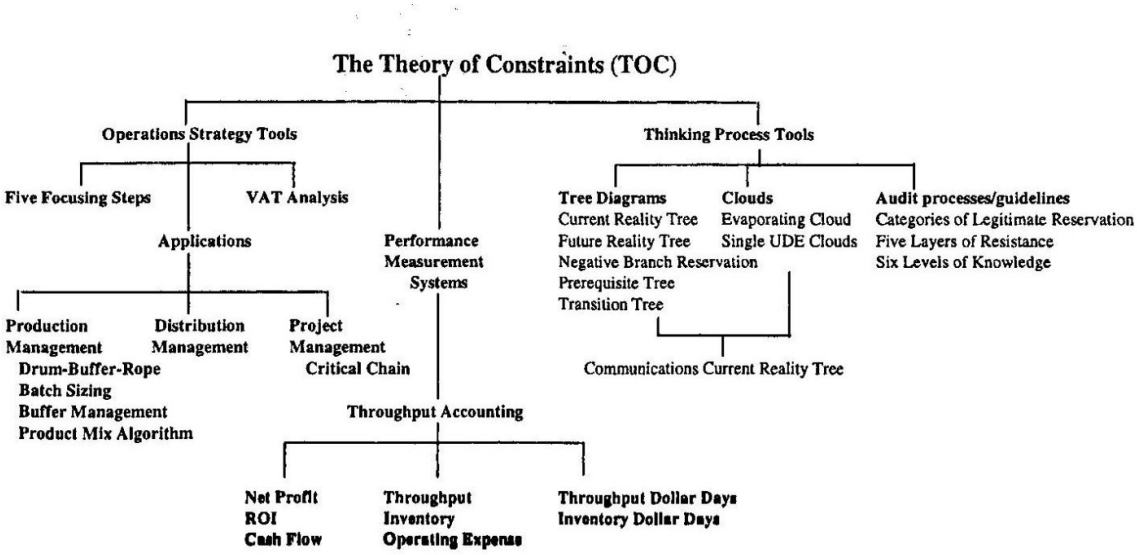

Figure 1 - The Outline of TOC from (Mabin and Balderstone, 2000)

The US Bureau of Labor Statistics, in its report on annual employment averages, puts the number of service-providing jobs in the United States at over 112 Million. (US Bureau of Labor Statistics Website 2011). This compares with a figure of almost 130 Million jobs in goods-producing industries suggesting that a large percentage (42%) of the country is employed in environments that produce no tangible goods. However, despite the large number of workers employed in goods-producing jobs, the percent of GDP that comes from manufacturing is relatively small. With 2009 GDP at $14.12 Trillion and manufacturing only accounting for $1.58 Trillion, the importance of industries other than manufacturing to US GDP can be appreciated. (Bureau of Economic Analysis - US Dept. of Commerce 2011)

Can this popular management philosophy which has been shown to produce positive results in manufacturing be applied successfully to service operations where there can be many constraints and where invisible constraints are constantly shifting? If TOC can be applied to service environments, what is the heuristic for managing the process? Service operations are sometimes managed by roving human supervisors whose job it is to make sure that people are busy. Being busy is equated with productivity. Defining productivity as bringing a company closer to its goals suggests that this function of keeping people busy does not necessarily equate to being productive. Since service operations are managed to operational goals by people with limited vision of the status of the entire system, it follows that sub-optimization will occur.

Motivated by my life experience working with the airlines and a fascination with the simplicity and elegance of the TOC approach that has yielded such remarkable results in manufacturing industries, I have undertaken to adapt the TOC to services in general and airline operations specifically to determine how TOC can be used successfully to improve service performance measures.

1.2 The Focus of the Theory of Constraints

As the terminology of the Theory of Constraints was developed for manufacturing, it has not been applied with as much frequency to service industries and service's less-visible processes. The implementation of the five focusing steps of process improvement of the TOC in manufacturing is straightforward and easy to understand.

The five focusing steps of the Theory of Constraints are:

1. Indentify the system constraint.
2. Exploit the constraint.
3. Subordinate all else to the constraint.
4. Elevate the constraint.
5. If the constraint is "broken", return to step 1.

In manufacturing, application of the TOC focusing steps is well documented, but in services we must begin by asking the question, "How shall a bottleneck be defined?" How shall it be exploited and the rest of the operation subordinated to its needs? These are some of the questions addressed in this research.

Goldratt's definition of a bottleneck is: "Any resource whose capacity is less than or equal to the demand placed on it". A bottleneck or constraint in a manufacturing environment can be easy to spot - look for pile of work-in-process in front of a resource. But this method for identifying a bottleneck can scarcely apply in the world of services as there are no piles of work-in-process. Finding an appropriate method for identifying bottlenecks and responding to them is one of the main objectives of this study because the TOC is based on the fundamental idea that system performance is limited by a bottleneck. By identifying and removing the bottleneck, system performance improves. The removal of one bottleneck means another bottleneck takes its place, so the process is repeated.

One of the key tenets of the TOC is the notion that dependent business activities can be thought of as links in a chain and complex operations can be thought of as networks of chains. Treating all of the links as if they were of equal importance leads to improvement efforts aimed at increasing the strength of each link. This might be appropriate if we were considering the weight of the chain where an increase at any link

would increase the weight of the chain. However, chains are not usually measured by their weight, but by their strength, which suggests that the strength of a chain is measured only at the weakest link. Therefore, to improve the strength of the chain, one must only focus on improving the strength of the weakest link - the bottleneck.

This research develops a bottleneck ratio measure, based on the principles of TOC, to determine what activities are constraints in a service operation. This simple ratio compares the amount of work remaining for an activity such as a flight given a certain number of workers and a known amount of time left to accomplish that activity. If the ratio is less than one, then there is less minute's worth of work than the minutes left to accomplish it – like having 20 minutes worth of work and 30 minutes in which to do it. If the ratio is positive, there is more work than can be done in the time remaining given the rate of processing corresponding to the number of workers assigned to do that work – an example would be a job that requires 30 man-minutes of labor to accomplish with only 25 minutes to accomplish it and only one worker to do it. The dissertation explains why this is a suitable way to identify otherwise invisible constraints in service, and how it can help management identify and address performance issues based on the dispatching of human resources.

The service environment in which we will apply TOC is one of multiple, simultaneous activities done by human workers with all of their variations in efficiency, going on in proximate locations each one of which has the potential, due to process variation and assignable causes of variation, to become a constraint to the accomplishment of the objective of the company which is usually on-time departure. The larger goal of the company, to make money, is seldom on the minds of the workers who,

on the whole, labor for personal reasons. Thus bottlenecks come and go and tardiness propagates through the system. It is possible that, due to a delay on some task for whatever reason, a particular process could become a constraint and those involved in the process not realize that this process is what is or will be causing a delay later. This is especially true if the delay occurs on one of the tasks performed early in the flowchart for all activities. For this reason, bottlenecks in services can be said to be invisible. Invisible bottlenecks can only be seen from a global perspective by someone with a perspective of the work remaining to be done for all pending activities in the system, the status of each activity across the system, and the rate at which the activity is progressing. The problem for service operations is the identification of such bottlenecks or constraints in the system absent such markers available to manufacturing industries as inventory piled in front of a resource. Take the example of a worker who must move 100 boxes from point A to point B. The quantity of work to be processed is known. The rate at which the worker can accomplish the task can also be known or at least estimated using observational studies. Therefore, the time required to accomplish the task can be calculated. If a time for accomplishment is fixed, then a bottleneck ratio can be calculated. This ratio would warn supervisory personnel that the process was beginning to lag behind such as if the worker above received a 5-minute phone call halfway through the process. The amount of work to be completed has not changed, therefore, the casual observer might not immediately recognize that, due to the loss of 5 minutes of potential process time, the bottleneck ration is much closer to 1 indicating that the activity is at risk for becoming impossible to accomplish in the time remaining before the deadline.

Once a resource is identified as a bottleneck, a means of responding to that knowledge is needed. There are several possible responses. One is to increase the work rate of the individuals doing the work. This is simply done with a shouted, "Hurry up!" In an environment where work rates are less variable, where some part of the service delivery system operates at a fixed speed, the option to hurry up is not as effective as adding more identical human resources to accomplish the task quicker. An example from the airline turnaround process is the process of downloading bags from the hold of an aircraft using a belt loader – a movable conveyer belt that runs at a constant speed moving bags from the bin to the worker at the bottom who stacks them on baggage carts.

1.3 Contribution and Significance

An intellectual contribution of this study is a method for applying an adaptation of the Theory of Constraints to the environment of a service factory. The term "service factory" is used to describe a service operation in which there is a low degree of interaction with the customer and customization of the service provided combined with a low degree of labor intensity. The application will require certain characteristics to exist in the service environment and as such this method could be applied to operations similar to that studied in this dissertation of airport ground operations. Similar service factories such as the construction industry, some aspects of agriculture, and some restaurant operations would benefit from the application of the method. The method is true to the intent of TOC in that it seeks to identify the system's constraint and then improve some objective function based on management of that constraint. Once the constraint is no longer a constraint, the method will automatically seek out the next constraint and seek to improve that using the five focusing steps.

1.3.1 A Recontexting of TOC for Services

The concept of a bottleneck will be adapted to service operations in such a way that an improvement method can be proposed for managing bottlenecks. Other terms in TOC will not change in their intent. Objective functions will also be discussed as applied to services.

1.3.2 A Generic Improvement Method for Applying the TOC to Service Environments

An improvement method is proposed that uses the principles of TOC to dispatch resources to address that which matters most at the time in keeping with the philosophy of TOC. The generic model will require the existence of a coordinating system with a global view of the state of the various tasks in the system such that assignment of resources will address the accomplishment of global objectives rather than local optimums.

1.3.3 Implementation of the Method in a Simulation

A "bottleneck ratio" will be defined and, in combination with a measure of urgency, used to rank the various tasks in a system. The method for identifying bottlenecks in a system using the bottleneck ratio and then managing the resources available to the system accordingly as a means of improving the objective function will be modeled in simulation software. The logic flow will be discussed as well as the limitations of such a system.

1.3.4 Application of the Method to a Case Study with Real Data

As an example of a service operation, the process of turning an aircraft at an international airport in the United States will be modeled in Arena - a discrete event

modeling software. Real data on flight arrivals, departures, passenger loads, and other flight attributes will be used to develop the model. Stochastic variation of various flight attributes will be applied to the input data. The objective function will be defined and the simulation will be run to verify that the method of assigning resources improves the objective function. A comparison of the operation based on operational objectives versus the operation based on financial objectives will show that profitability can be improved by adopting the method to manage constraints.

1.3.5 Experimental Demonstration of Efficacy

This is a significant contribution to theory and practice. Once a suitable software and hardware package were designed, the method would be simple to implement requiring virtually no training on the part of those that will be managed by the system and supplanting much of the guesswork of those supervisors that currently manage the workers. Decisions about what to do next would always be made with the company's highest priorities in mind such as profitability through cost avoidance rather than by the personal priorities of workers. A demonstration of the applicability of this method to the specific service of aircraft turnaround will be used to validate these claims. If TOC can be applied successfully to a service environment run by people without special training or knowledge, it would be of great benefit in terms of profitability, throughput, and ease of implementation.

1.4 Research Method

Since this dissertation is exploratory in nature, an empirical research methodology will be used. The case study method coupled with simulation will be the primary

methodologies used to gain insight into the behavior of service systems. According to (Lockamy III and Spencer 1998), a case study is an empirical inquiry that:

- Investigates a contemporary phenomenon within its real life context; when
- The boundaries between phenomenon and context are not clearly evident; and in which
- Multiple sources of evidence are used.

Site visits to various airports were made to collect the needed data. Interviews with airline executives in charge of airport operations and airline safety were conducted over several years to explore dependencies in the operation and quantify process times. A case study was written using the logic of the service process and the process times from the actual process. A method for applying TOC to services was generated. The process was mapped to determine the precedence relationships in the operation. Through observational studies of actual turns at various airports, data was gathered to build an as-is model of the operation of a particular airline. Process times and distributions were calculated based on the data gathered for the various processes that comprise the model. A model of the operation was created in Arena, discrete-event modeling software. The as-is model was validated to assure that the model closely resembles the process it represents. Once validated, the model will be modified to reflect the philosophy of TOC – its objective function, its method of management, and its response priorities. Both the as-is model and the to-be model outputs will then be compared to see if management by TOC can be successfully applied to service factories. The efficacy of the method for improving performance measures and improving profitability was tested using a computer simulation. If the outputs of the simulation suggest that management by such a

system improves performance measures, then further research is warranted both in terms of increasing the nuances of the simulation to provide richer detail and in adjusting the parameters of the decision making heuristic to optimize performance measures.

1.5 Organization

Following this introductory chapter, chapter 2 will include a review of the literature on TOC as it has been applied to services with light coverage of application to manufacturing. The literature will also be reviewed for other process improvement methods which will then be compared to TOC. Chapter 3 will begin with a discussion of the aircraft turnaround process in general. A process map of the service system will be given. Following the map, a discussion of the logic that must go into making the simulation of the process in ARENA will be presented. The model will then be run with the input data and distributions devised from the data collection phase to demonstrate validity. In the second half of Chapter 3, the method for managing the system based on TOC will be presented and the model will be adjusted to reflect this method for dispatching human resources. In Chapter 4, output results for both the as-is and to-be models will be compared and discussion of the results will be presented. Chapter 5 will contain the conclusions of the study and suggestions for future research on the topic.

2 LITERATURE REVIEW

2.1 The Theory of Constraints

The Theory of Constraints was originally proposed in, "The Goal" (Goldratt and Cox 2004). It is an intuitive management philosophy developed by Eliyahu Goldratt in the mid 1980's. TOC principles suggest that management begin process improvement by identifying the constraint in an operation and then focus process improvement efforts on that constraint to improve the process.

Bottlenecks arise in that both manufacturing and service processes due to the combination of two phenomena: dependent events and statistical fluctuations. "Dependent events" are activities with precedence relationships, that is, one activity must be done before another. "Statistical fluctuations" refers to the fact that process times vary from batch to batch or unit to unit.

Constraints in a system are also known as bottlenecks. Bottlenecks can be processes, machines with limited production capacity, policies, or practices that limit the company from achieving whatever its desired outcomes are. By reducing the impact of bottlenecks in an operation, substantial improvements in throughput and reductions in WIP can be realized. Therefore, the relatively simple approach of the TOC is to identify the bottleneck, manage it, manage the rest of the operation according to its needs, and then take whatever measures are necessary to break the bottleneck, if possible, such that is no longer a constraint in the system. If successful, one seeks out the inevitable new constraint and repeats the process.

2.1.1 Origins of TOC

The Theory of Constraints had a humble beginning when, in the late 1970‘s, a neighbor of Eliyahu Goldratt, an Israeli physicist, asked him for assistance in creating a scheduling program to increase the output of his chicken coop factory. The resulting software package, known as “Optimized Production Timetables” (OPT) scheduling software was the first practical application of TOC. From this simple scheduling software TOC has evolved into a set of management tools encompassing production, logistics, and problem solving and thinking tools (Watson, Blackstone and Gardiner 2007). This business philosophy captured the attention of practitioners with the publishing of “The Goal” in 1984. Be it remembered that Goldratt, owing to the target audience of “The Goal”, did not present his creation as an academician through peer-reviewed publication, but as an entrepreneur through the medium of a business novel. The success of the book prompted Goldratt to leave the software business and establish himself as a business educator. (Stanley C. Gardiner 2007) Nevertheless, in this seminal work Goldratt borrowed from previous operational wisdom and put forth many common-sense philosophies that have come to be lumped together and known collectively as the Theory of Constraints or TOC. The TOC’s application has been documented to produce favorable results when applied to the manufacturing environment. (Mabin and Balderstone 2000)

2.1.2 The Five Focusing Steps of TOC

At the core of the TOC lie the five focusing steps designed to help management discover a system's constraint and manage the entire system accordingly (Goldratt and Cox 2004). Dan Trietsch reminds us that there is an implied Step 0 here - we must first "select an objective function and decide how to measure it" (Dan Trietsch 2005).

Step 0 – select an objective function and decide how to measure it
Step 1 - Identify the system constraint or bottleneck
Step 2 – Exploit the constraint
Step 3 - Subordinate all else to the constraint.
Step 4 - Elevate the constraint.
Step 5 – If the constraint is broken, go back to step 1 and find the new constraint.

2.1.3 TOC in Manufacturing

Much research has been done on the Theory of Constraints (TOC) in manufacturing applications. Successful implementations have been documented in (Umble, Umble and Murakami 2006), (Bolander and Taylor 2000), (Cook 1994), and (Miller 2000). Other research has concluded that the successful findings are generalizable to other types of organizations, particularly their operations aspects (Mabin and Forgeson 2003), (Gupta and Boyd 2008).

The gains in the manufacturing industry are well documented. (Watson, Blackstone and Gardiner 2007) cite Mabin and Balderstone's comprehensive review in 2000 of publicly disclosed benefits from adoption of the TOC philosophy. Improvements of an order of magnitude were not uncommon. Their findings include:

- A 70% mean reduction in order to delivery lead time from a sample of 32 observations with more than 75% reporting a reduction greater than 50%

- A 65% mean reduction in manufacturing cycle time based on 14 observations
- A 49% mean reduction in inventory from a sample of 32 observations
- A 63% mean increase in throughput/revenue, excluding one outlier of +600% at Lucent Technologies, from a sample of 22, 5 of which increased revenue +100%.
- A 44% mean improvement in due date performance from a sample of 13.

2.1.4 TOC and JIT

Sale and Inman have conducted a survey of 45 Indian companies comparing the performance and the change in performance of companies reporting TOC adoption, those reporting JIT adoption, those reporting to have adopted both, and those reporting to have adopted neither (traditional manufacturing). They indicate that the greatest performance and improvement in performance accrued to adopters of TOC. They have further reported that the idea that the combining of the two philosophies (JIT and TOC) may result in a synergy or performance higher than either one alone, was not substantiated. They could not find support to this as low number of firms (6) in the mixed methods category may somewhat limit the generality of the results (Sale and Inman 2003)

2.1.5 TOC and Six Sigma

Ehie and Sheu propose an integrated TOC/Six Sigma framework to combine the benefits of Six Sigma and TOC. They review the two strategies and the process involved in each strategy before suggesting a way of integrating them. The two methods and their integration are shown in Figure 2. Six Sigma deals with defect reduction and, as adopted by Motorola, would lead to 3.4 parts per million defects. The strategy involves

the use of statistical tools for gaining knowledge needed to achieve better, faster and less expensive products and services. The way to integrate the two is to use TOC to identify the constraint in the system and then use the DMAIC principles of Six Sigma to improve the output of that constraint in accordance with step four, "elevate the constraint", of the five focusing steps of TOC. Six Sigma should not be used indiscriminately to improve all processes in a system as this leads to local optimums.

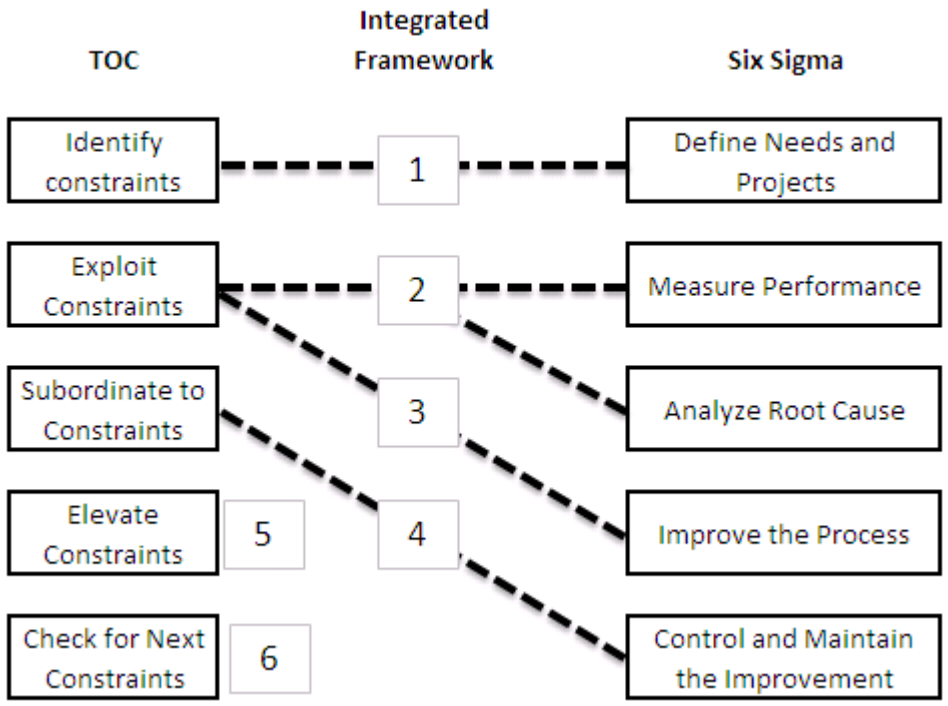

Figure 2 – Integrated TOC and Six Sigma Framework (Ike Ehie 2005)

2.1.6 TOC Research with Respect to Services

In (Schmenner 1986) a service process matrix is proposed relating the degree of interaction and customization to the degree of labor intensity. This service process matrix is shown in Figure 3.

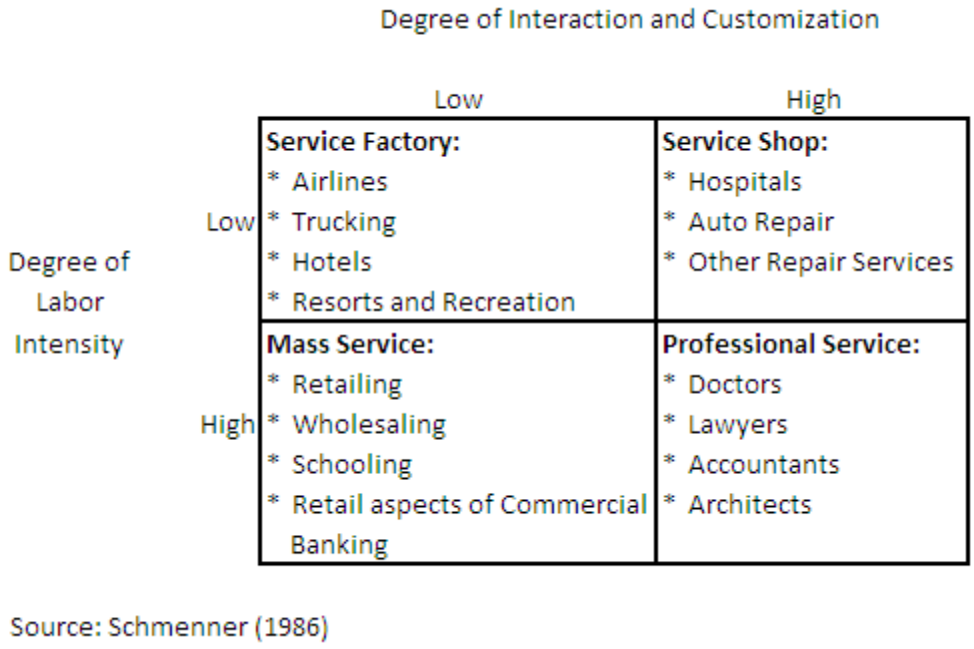

Figure 3: Service Process Matrix

Applying TOC to service operations presents a set of challenges. Application to services requires that we adapt concepts and translate vocabulary from a world of inventory, machines in series, fixed capacities, and production lines to an environment of workers who can work at variable speeds, switch jobs almost instantaneously, and whose output is more difficult to measure. Siha focused on the translation of the TOC vocabulary for use in the service industry (Siha 1999). Siha suggests that "inventory" is unused service such as a room in a hotel, a space in a truck, and a seat on a flight; that inventory in services is physical in nature. "Throughput" is the money generated from selling the service. Operating expense follows the standard TOC definition of the money spent to turn inventory into throughput. He also proposed a classified model for TOC applications based on Schmenner's classification of services shown in Figure 3.

In manufacturing, a buffer frequently consists of a quantity of parts or inputs to a process stacked immediately upstream of the resource. The buffer protects the resource from interruptions in the flow through upstream processes. Service buffers thus are harder to observe and quantify than manufacturing buffers because constraints are not as easily discerned and labeled as such. A bottleneck in a service operation will be defined as any process in which the amount of time required to complete the process, given the current process rate, is greater than the time remaining before the due date. It can be assumed that a resource, say a worker, has a limited capacity to accomplish the amount of work given to him in the amount of time remaining prior to the time the work must be completed, and that he operates in relative isolation from the rest of the agents in the system, he may become a bottleneck . Perhaps such a worker is equipped with means to communicate with others in the system such as a walkie-talkie or telephone. Then there must be times when, due to the irregular arrival pattern of the work he must accomplish, his capacity to accomplish that work is insufficient in the time remaining to him before there is a detrimental delay. In essence, he unknowingly becomes a constraint in the global system. He is a constraint with respect to the effort to accomplish some global objective function. More than one such bottleneck may exist at a time. It is through awareness on the part of operational controllers, of the existence of such constraints that tactical decisions could be made quickly to exploit, subordinate, and elevate according to the TOC before too much time is lost.

2.2 Research on Turnaround Operations

Much work has been done analyzing passenger boarding, as the passenger boarding process comprises a significant part of the larger turnaround process - one in

which there is substantial variation. Modeling interferences among passengers as they board the aircraft using linear programming has been used to suggest a preferred boarding pattern if the airline adopts the practice of issuing reserved seats (Basargan 2007). In contrast to Basargan's work, Ferrari et. al put special emphasis on disturbances, such as a certain number of passengers not following their boarding group but boarding earlier or later. The surprising result was that the typical back-to-front boarding strategy becomes improved when passengers do not board with their assigned group (Ferrari and Nagel 2005).

Computer simulation has been researched for contingency planning for turnaround operations at airline hubs (Adeleye and Chung 2006). The essence of their research experimental design was to conduct an analysis of the effects of altering the baggage upload delay across seven different levels. Baggage upload delay was defined as the time between the end of offload and the start of upload compared to the base model in which baggage upload was initiated 40 minutes before scheduled departure.

(Van Landeghem and Beuselinck 2002) used simulation analysis of different boarding patterns and operating strategies to suggest ways to improve the existing system.

2.3 Research on Measurement of System Status

The latest technology for tracking bags and even passengers are Radio Frequency Identification (RFID) tags that can be inserted inexpensively into boarding passes and bag tags. Read by radio frequency "readers" placed at key points in the airport and on the ramp, these devices enable management to know where every passenger and every bag is

in real time. Thorne, et.al analyzed the impact of RFID on the aircraft turnaround process and concluded that ID technologies provide a mechanism for obtaining automated visibility of physical processes, but the information systems and business processes for sharing this information between partners are essential (Thorne, Barrett and McFarlane 2007).

The other piece of technology necessary to implement a TOC-based solution involves the communication to management of the timestamp for certain activities in the turnaround process. An Aircraft Turnaround Monitoring System has been proposed and tested by Wu using PDAs and GPRS technology at the Sydney airport in Australia to improve situational awareness of loading supervisors and operational controllers (Wu 2008)

As "The Goal" was written principally for a manufacturing audience, many of the concepts and terms included therein deal with the challenges of operating a manufacturing plant. Machines are fixed in series. Capacity is more rigid and dictated by the individual machines in a process, by policy, and by other ways of advancing work that limit processing time.

This line of thinking, applied to airport operations, leads to the question of long term vs. short term constraints. A long term constraint could be identified by the symptom of repeated delays resulting from a persistent lack of capacity in a given area. If such frequent disruptions with respect to cycle time can be assigned to a particular resource, this may be evidence that that resource is a constraint in the system.

Bottlenecks are relative; that is, a resource of a given capacity is only a bottleneck if the associated resources upstream and downstream have more capacity. Once a bottleneck's capacity is elevated with respect to the other bottlenecks in the system, it ceases being the bottleneck.

2.3.1 Isomorphism between TOC and PERT/CPM (Dan Trietsch 2005)

Table 1 - Isomorphism between TOC and PERT/CPM

Language from TOC	Language from PERT/CPM
Step 0 - Select an objective function; determine how to measure it.	The objective is to complete a project on time and within budget.
Step 1 - Identify the bottleneck or constraint.	Identify the critical path. (Bottleneck resources will always be on a critical path.
Step 2 - Exploit the constraint.	Focus managerial attention on the critical path to ensure that no time is wasted on any of the activities of the critical path.
Step 3 - Subordinate all non-constraint resources to the operation of the constraint.	Manage activities not on the critical path in such a way that they do not delay the project. One conservative approach is to start all non-critical activities as early as possible
Step 4 - Elevate or improve the constraint to increase its capacity.	Through project "crashing" or the deliberate application of additional resources to critical activities, the length of the critical path may be reduced where it is financially advantageous.
Step 5 - If the constraint is broken, return to Step 1.	When crashing is applied, additional critical paths may be created or the critical path may change completely. Either way, crashing requires focusing on the evolving critical path; an iterative process.

Isomorphism between TOC and PERT/CPM is shown in Table 1 (Dan Trietsch 2005). In PERT, a weighted average of three time estimates is used for project planning. These are the optimistic, most likely, and pessimistic estimates of the time required to complete some task. From these three, an "expected" time is calculated using a beta distribution which allots a weight of 1/6 to the pessimistic, 2/3 to the most likely, and 1/6 to the optimistic times to calculate the expected time for an activity. This is the time used for estimating, in advance, the duration of the various activities in the project. Goldratt claims that, "the uncertainty existing in every project is the underlying main cause for most problems. Now, we see that people are not blind to it and they do add a lot of safety in their planning" (Goldratt 1997). Since uncertainty is unavoidable in project management, management must focus its efforts on how they deal with uncertainty and the application of safety time to improve processes. A key to the critical chain approach is how uncertainty is managed in estimating completion times.

2.4 Literature Review Summary

Though the originality of the theory of constraints may be cause for debate, there is solid evidence that managing a process according to its bottlenecks will improve performance measures. Gains in manufacturing have included mean reductions in order to delivery lead time, mean reduction in manufacturing cycle time, mean reduction in inventory, mean increase in throughput/revenue, and mean improvement in due date performance. Despite the fact that this approach has met with considerable success in manufacturing, the number of applications to services in the literature is sparse. One of the contributions of this dissertation is to focus on a method for identifying bottlenecks in services so that TOC can be applied.

The theory of constraints has been compared to other improvement methods such as JIT, Six Sigma, and TQM and has proven robust at improving performance measures of increasing throughput and decreasing WIP. TOC has also been paired successfully with other improvement methods in an attempt to develop synergies. Successful pairings used the theory of constraints to identify the bottlenecks and other process improvement strategies such as six Sigma or TQM to improve the performance of the bottleneck.

Modeling of aircraft turnaround operations has been conducted by several but with an eye toward decreasing turnaround times on all flights - an operational improvement. The focus of the research has ranged from the control of the waiting time between downloading and uploading bags to the pattern in which passengers board the aircraft. The use of RFID tags has been proposed to decrease the number of lost bags and improve real-time knowledge of the location of all bags. Radiofrequency systems have been proposed to timestamp key events in the process; once again with an eye toward improving turnaround times across all flights. However, no general system has been proposed for applying the theory of constraints to a service factory to attain global objectives.

3 THE IMPROVEMENT METHOD

The purpose of this research is to design a method for improving performance in a pure service environment through the implementation of a method that assigns and dispatches workers to various jobs based on the global objectives of the company. The method developed in this research will be based upon principles from TOC. The improvement method proposed in this research assumes process times based on the number of workers assigned to a job. Process rates will be faster if more workers are assigned. The method also assumes that workers are a scarce resource and that it is the intention of management to do as much as possible with as few workers as possible to minimize costs associated with employing workers. The method also assumes diminishing returns as the number of workers approaches a maximum. Such a maximum could be constrained by job environment limitations such as space in which to do the job or number of stations that can be manned for a given job. The addition of workers beyond this maximum would not increase the rate at which the job was completed.

The arrival pattern of jobs, though anticipated, is subject to uncertainty. Normally, the estimated arrival pattern of the jobs allows for the scheduling of workers to accomplish the jobs. However, in the face of uncertainty, system improvement is possible through adapting the schedule according to actual work patterns. These workers can be said to be "on the bench". Whenever a worker completes any task, that worker returns to the bench. The system can then reassign that worker to another task according to the logic of the method.

The method also assumes an environment in which all jobs are not equal. The assumption is that some jobs bear late penalties that will be incurred if the job is not

finished by some critical due date. Construction projects large and small, architectural projects, landscaping projects, and engineering design projects among others frequently have such penalties for late delivery of a job.

An assumption in this research is that the rate at which a job is done is constant. Obviously, this is a simplification of real life especially in an environment in which work is accomplished by human labor. The ability of human workers to speed up the rate at which they are processing units of work adds a layer of complexity that will be reserved for future research. Therefore, for purposes of this model the process rate will vary only with the number of workers. The key output of the method is the number of workers assigned to each task to be able to complete the tasks in a manner that follows the objectives of the company with respect to cost avoidance and lateness.

3.1 Defining a Bottleneck in Services

In the manufacturing world, a bottleneck is defined as any resource whose capacity is less than or equal to the demands placed upon it. In the service world that definition must be manipulated to be actionable. Drawing a parallel between a machine that accomplishes a set amount of work in a certain amount of time and a human being who also accomplishes a set amount of work in a certain amount of time the notion of a constraint can be adapted.

A constraint in this service environment will be defined as a resource, particularly a human resource, who does not have the capacity to accomplish the work before him in the time allotted for that work.

This definition of a bottleneck can reduce the measurement to a simple ratio. With a nod to Goldratt's terminology, this ratio will be called the “bottleneck ratio”. The bottleneck ratio will have as its numerator the amount of work remaining in all of the tasks that must be accomplished prior to push back. The Bottleneck Ratio is given by Equation 1.

$$\frac{amount\ of\ work\ remaining\ (in\ minutes)}{amount\ of\ time\ remaining\ before\ the\ due\ date\ (in\ minutes)} \qquad \text{Eq. 1.}$$

Work remaining = # units to be processed x process rate given the # of workers Eq. 2.

The "amount of work remaining" in equation 1 represents the total amount of time that will be required to process outstanding units. It consists of the outstanding number of units to be processed and the assumed rate of processing given the # of resources assigned to the job. Notice that the units in the numerator are given in units of time. However, what is really known is the remaining number of units to be processed. Conversion from the number of units to an amount of time can be performed using the process rate that corresponds to the number of resources in the form of workers assigned to accomplish the job. It is assumed that the outstanding number of units to be processed is not variable; therefore, only the process rate may be changed in the numerator. The "amount of time remaining before the due date” can be measured by subtracting the current clock time from the time associated with the due date.

Once both the numerator and the denominator are calculated, the ratio may be calculated. Values for this ratio less than one suggest that the process rate is adequate for the accomplishment of the task by the due date. An example would be having a

calculated value of 20 minutes worth of work and 30 minutes to do it in. Values for the ratio in excess of one suggest that the process rate is inadequate for the accomplishment of the task at the due date. An example would be having a calculated value of 40 minutes worth of work and only 30 minutes to do it in. A ratio greater than one is a signal to the system that additional resources must be assigned to this job to increase the process rate if the job is to be completed by the due date. Notice that the bottleneck ratio alone does not determine how many workers are assigned to a job. It merely calculates the number of workers that would be required to finish the job on time. It is the bottleneck ratio in combination with the concept of urgency with respect to the objective function that will be used to assign scarce resources.

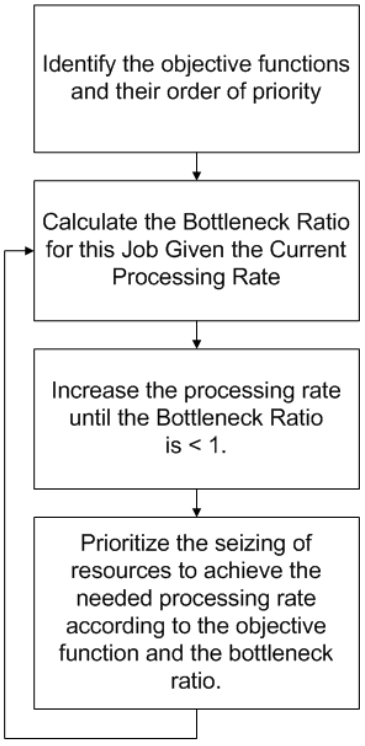

Figure 4 - Improvement Method Using the Bottleneck Ratio

3.2 The Logic of the Method

The logic of the method for managing scarce resources according to the bottleneck ratio to the accomplishment of the objective function is shown in Figure 4. which shows the decision logic's three basic parts: the assignment of importance based upon the objective function, a loop for determining the process rate needed to accomplish the work prior to a due date, and the command line that combines the two into both a quantitative and qualitative action of increasing assigned resources to increase the process rate to the desired level to accomplish the work on time.

3.3 Inputs Needed for the Logic of the Method

The logic of the method shown in Figure 4 requires the following input information:

1. Known amounts of work. This could be in the form of the number of identical units to be processed in the job assuming that the job consists of processing identical units or some measure quantifying the amount of time a single server would require to complete the work.
2. The process rate per unit given the number of workers assigned. Typically, the more workers that are assigned to a job, the faster the process rate per unit.
3. The maximum number of workers allowed per job. This assumes a maximum exists beyond which either more cannot be assigned due to some other constraint such as working space or diminishing returns due to constrained working space.
4. A time due date by which the job must be completed.
5. A monetary impact for not finishing the job by the due date.

3.4 Characteristics of a Service Factory Wherein This Method Will Apply

For this method to be put into practice, the service process would have to have certain characteristics that would lend themselves to effective implementation. The logic and calculations of the model were designed for environment of series processing of tasks with precedence relationships. Such an environment would also require close proximity between tasks. The arrival pattern of the jobs and the expected duration of the jobs or volume of the work involved would have to be known in advance.

There are other such service systems that can be modeled using this approach and that can benefit from this system to dispatch workers. One very large example comes from the construction industry. A contractor will take on multiple jobs that frequently bear penalties for late completion. The contractor has access to a labor pool comprised of subcontractors. These can be employed and released on a job by job basis. Assigning more workers to a job increases the rate at which the job is processed decreasing the risk of incurring arbitrary penalties associated with lateness. The contractor would benefit from a system that constantly updates the status of all jobs and allows him the ability to re-task subcontractors to satisfy financial and operational goals. The arrival pattern of work is stochastic in nature due to the fact that many jobs have precedence relationships with other jobs. Another example is found in agriculture. Given a fruit orchard to be harvested in which the amount of work is known a priori given the acreage or number of trees in the grove that must be picked, the average work rate of the laborers, and the importance of different varieties of the crop to be harvested. The penalty would come from spoilage in the event that the optimal time for harvesting passed. Another example comes from the management of a restaurant in which some customers are more important

than others. Assigning additional crew to service an important party would diminish losses from poor service associated with timeliness.

3.4.1 Known Quantities of Work per Activity and Processing Rates

For the logic of the method to work the number of units to be processed as well as the processing rate per unit must be known before hand. This could also be in the form of piecework wherein the known process rates per piece given the number of workers was known or, in the event of a different kind of work other than piecework the process rate for the job given the number of workers was known. Either way, to be able to calculate the bottleneck ratio, the amount of work remaining must be calculable.

Processing rates must be calculated beforehand. Logically, the processing rate given n+1 workers should be faster than the process rate given n workers. In the logic as it is, the improvements need not be linear. In some applications, a maximum would probably be reached beyond which the process would see a diminution in the process rate stemming from having too many workers that would conceivably slow down the process.

3.4.2 Flexible Resources and Proximity

This logic shown in the method simplifies some of the realities of real implementation. For example, adding additional workers happens instantaneously in the model. They "hit the ground running" and immediately alter the process rate by their presence. In a real application, the addition of another worker would not be instantaneous nor would the time required for the worker to be present be constant. Time required to get from where the worker is to where the worker is assigned would impact the real results also.

3.4.3 Global System and Heuristic to Dispatch Workers

This method would apply in an environment where traditionally workers have been rewarded for doing a good job on what they were assigned to. Frequently however, the intent of doing one's job well results in sub-optimization wherein one work center seeks to improve its own performance measures and by so doing decreases the performance measures of the global system. The true benefit of the bottleneck ratio method is that all work is done according to global priorities. Therefore the most important thing that could be done is always assigned first.

3.4.4 Frequency of Pushing the Button

The worker who notifies the system that she is available for reassignment by "pushing the button" of whatever device is being used to communicate with the system is essentially being released to the bench -- a holding place where workers who have completed some task await reassignment. This is a virtual place not an actual place. The frequency with which buttons pushed throughout the system would depend on the number of workers, and the number of simultaneous tasks and the duration of those tasks as well as the number of workers assigned to each task. The more often workers press their buttons signaling that they are available for reassignment, the more the system can achieve global optimization. The less the buttons are pressed, the more the system tends to resemble a traditional system where workers are infrequently assigned to specific tasks.

3.5 Long term vs. Short Term View

Manufacturing companies who embrace TOC look for long term constraints rather than temporary constraints. Long term constraints in manufacturing are usually the

result of process design and management policy rather than day-to-day operations. Due to routings for the individual components of various finished goods, what is a constraint for one order configuration might not be a constraint for another order configuration. Thus, to be able to identify a long term constraint requires a certain level of stability in the system with respect to the demand for the system's resources.

A temporary bottleneck is one in which the resource has enough capacity to handle the load placed upon it over time but perhaps not over the short term spikes in demand or in the face of certain configurations of demand aggregation. The implied element of time comes into play here. In other words, over the short term, such a resource may not have the capacity to accomplish what is expected, hence it is deemed a bottleneck, but over the long term as demand is aggregated, it does have enough capacity and is therefore not a bottleneck. It is the comparison of the demand for its output of a resource to its actual output capacity that determines whether it is a bottleneck.

Goldratt and Cox (1984) suggest four simple qualitative ways to identify bottlenecks in a manufacturing plant:

1. Rely upon the experience of management in knowing which resource is usually the cause of missing parts.
2. The resources most often investigated by expeditors or those who must rush late orders.
3. Piles of work-in-process inventory in front of a resource.
4. Aggregate the routings and demand quantities in the database and analyze the resulting loads on each resource in the system.

These somewhat primitive means for identifying a bottleneck may be sufficient if one is looking for a long-term bottleneck; one that is a constraint despite small perturbations in flow occasioned by the statistical fluctuations of typical human/machine interaction. They are not useful for identifying temporary bottlenecks that cause

relatively small perturbations. Such temporary bottlenecks require measurement protocols and sensing devices and systems that exceed the capabilities of the naked eye to discern.

4 CASE STUDY DESCRIPTION AND MODEL FORMULATION

4.1 Mapping the Service System of Aircraft Turns

The service process that will serve as the laboratory for this study is that of the turnaround process of a major airline at an international airport in the United States of America. All the activities involved in processing an aircraft while it is on the ground are called "ground operations" in the industry. The turnaround process, simply put, consists of all those activities that must be performed to complete service for one flight and begin another: passengers must deplane from inbound flight and luggage must be offloaded the luggage and passengers for the outbound flight can be processed. The aircraft must also be cleaned, serviced, fueled, and prepared for its next flight. All of these tasks must be completed by relatively unskilled labor within a time period of 40 to 75 minutes on average using equipment and procedures that haven't changed much in the past twenty-five years.

4.2 Overview of this Case Study

The airline studied runs its own ground operations in Ft. Lauderdale International Airport and all workers participating in the processes under study are paid and supervised by the airline. This includes all ramp, counter, and gate personnel. Consequently, the company is completely responsible for all aspects of ground operations including the functions of ticketing, baggage handling, gate operations, and cleaning of the aircraft. The only significant sub-process in this research that is outsourced to a third party will be the fueling of the aircraft. This completeness of managerial control was favorable to the research as it allows for access to all levels of management practice and pertinent operations data.

Current operations performance measures at the airlines focuses on the number of flights that depart late each week and the total number of minutes that they are late. These are presented in histograms to management and used to assess operation's effectiveness. These measures encourage management to get every flight out on time so that performance measures for the week will improve. While this is accurate, it is not in the best interest of the company to elevate this measure and hold supervisors responsible for this metric as the principle measure of their effectiveness nor is it in the best interest of the company to use it to manage operations as the on-time departure status of all flights are not equally important. Other priorities eclipse this measure of success for operations and improve the true goal of a company – to make money.

This system is characterized by a fairly accurate foreknowledge of the work to be accomplished during the turn in the form of a known number of pieces of luggage or "bags" as they are called in the industry for both the inbound and outbound flights. Not only is the number of bags and their storage locations on the inbound flight known, but once the passengers have checked in and weight and balance calculations have been accomplished, the assignment of bags to front and rear bins is also known for the outbound flight. Using average process times gathered from observational studies, the rate of processing these bags can be estimated given the number of workers assigned to the task.

Each of the activities involves a setup process in which equipment will be positioned, bin doors opened, and baggage tugs and carts positioned to collect the downloaded luggage. The setups in this model are not trivial as they involve the

movement of belt loaders, baggage tugs, tow bars, pushback tractors, and carts full of stacked luggage.

The skill level of the workers involved in the process is very low. Their main activity is comprised of muscling around people's luggage and moving machinery. Therefore, the human resources required to accomplish the tasks are fairly interchangeable; they can be re-tasked easily using modern communication equipment such as radios, and pager type devices. This interchangeability of workers is one of the keys to the success of this model. The key characteristic is that employees are flexible and can be reassigned dynamically.

Another important characteristic of this kind of operation, an operation that relies heavily on human workers, is that employees can "speed-up" for short durations and work at a faster pace. Logically, a faster pace cannot be sustained for extended periods of time without deterioration in the quality of the output and the morale of the worker. These characteristics are not the focus of this study and therefore the ability of workers to work at faster rates for short durations will not be included in this model but will be left to future research of this model. While workers will be assumed to work at a constant rate as individuals, the process rate will be variable through the addition of more workers.

The process of an aircraft turn is never accomplished the same way twice; each turn is unique. Even the game plan of the process will be different for different airports, and constraints can shift from activity to activity during the turn as the flexible resources that accomplish the work either speed up or slow down or are moved and retasked. This shifting of the constraint is at the heart of this model and is necessary for the model to

have meaning. It will be shown that various departments or tasks can become temporary bottlenecks when the capacity to accomplish the task is insufficient for the amount of time remaining in which the task must be accomplished. The definition of a constraint in this service system will be discussed further on.

There is a lack of a global view of the process. For example, what motivates a ramp supervisor to give priority to one flight over another? This is the key consideration when we think of a global view of the process.

Lacking some sort of computer-based dashboard system that gives the status of each operation of each flight forces management to treat all operations with the same urgency. Treating all flights as equal will lead to sub-optimization of the process and the true objective of the airline – "to make money now and in the future" will suffer.

Current measures call attention to the number of flights that depart late and the total number of minutes that they are late. Such measures tend to suggest that by getting any flight out on time that the performance measure will improve. While this is correct, it is not in the best interest of the company to use this measure or to use it to manage operations as the on-time departure status of all flights are not equally important.

If we assume that a resource, say a baggage room employee, has a limited capacity to accomplish the amount of work given to him in the amount of time remaining prior to the time the work must be completed, and that he operates in relative isolation from the rest of the agents in the system (with a walkie-talkie at best), then there must be times when, due to the irregular arrival pattern of the work he must accomplish, his capacity to accomplish that work is insufficient in the time remaining to him before there

is a detrimental delay. In essence, he unknowingly becomes a constraint in the global system. This can happen with bag room workers, ticket counter agents, TSA agents, the boarding agent, fueler, and any other agent that must process the entities of the system (the passengers and the bags) prior to push-back. More than one such bottleneck may exist at a time. It is through awareness on the part of operational controllers, of the existence of such constraints that tactical decisions could be made quickly to exploit, subordinate, and elevate according to the TOC before too much precious turn-time is lost.

4.2.1 The Environment of airport ground operations

The environment of airport ground operations adds complexity to the problem due to the following: the work of a "turn" must be accomplished by a certain clock time. Though this deadline is known for each, the start time for the operation is not as inbound flights can be delayed or arrive early. Also, the processing of people and luggage during a turn is accomplished in batches. The batches are directly correlated with the arrival pattern of the passengers and luggage to those areas where they will be processed. Inbound luggage is downloaded as a single batch as are inbound passengers. However, outbound luggage and passengers have a different arrival pattern and so are usually processed in multiple batches.

4.2.2 The State of the Art with Respect to Technology

What is the system that is used today to manage the process? Most airlines still accomplish aircraft turns using the same technology that has been in use for the past 40 years. They use the same baggage tugs, the same baggage carts, belt loaders, aircraft push-back tractors, tow bars, catering trucks, jetways, and even aircraft that have been common for decades. It would be very costly to innovate with respect to the equipment

used since the company that did innovate would have to either change the technology of the entire network, or use and maintain parallel systems in the event that certain interfaces were not brought up to the latest technology.

When we speak of a system to manage the process, we think of a centralized computer database that receives inputs of where the passengers are in the process and where their luggage is in the process. The status of the system is monitored by an automated or human system with the power to intervene and redeploy assets to accomplish the goals of the company. This is possible using either laser and UPC symbols or, more reliably, RFIDs in boarding passes and luggage tags. So embedded, and with sensors at key points along the routes possible from ticket counter to gate and at key checkpoints in the movement of the baggage this technology allows a central computer to know where its RFIDs are in the process and manage intervention much more efficiently if needed.

4.2.3 The Current System of Management

Obviously there is some system in use in the current system. While it is not a centralized system, nor a computerized system, there is a system. It is a manual system made up of expectations, assignments, standard operating procedures, and inertia in the minds of the agents and their supervisors responsible for the accomplishment of the turn. They have been charged with accomplishing certain isolated, measurable tasks by certain times for each flight. Such agents may be well motivated to accomplish that which has been assigned to them; however, their actions may lead to suboptimal results. The lack of a centralized system that has the capability of determining which of the areas of work is falling behind, leads to individual agents seeking to do that which has been assigned to

them in such a way as to garner praise or, conversely, avoid censure for the apparent lack of efficiency or timeliness of their work. This leads to what is called local optimums in TOC jargon as each resource seeks to have its numbers look good without regard to the impact they have on the global system.

4.2.4 Batching

The first observation that affects all others is that where there is work that must be done that involves processing multiple entities in a similar fashion, that work is accumulated in holding areas then subsequently processed in batches. This is particularly true of bags and passengers.

Passengers arrive to board the aircraft at the gate area and wait in a queueing area for the signal to gather their belongings and board the plane. Each of the passengers will have to be processed briefly as they show their boarding passes to the gate agent prior to entering the jet bridge to the plane. This is done individually or in small groups, nevertheless, the boarding of the plane can be said to be a batch process.

Checked bags are processed by ticket counter agents individually or in small batches corresponding to families that check in together, then sent down a conveyor belt to be processed individually by TSA prior to being sent on to the bag room where they are sorted by destination and flight number. This accumulation and aggregation in the bag room creates the batches that the airline will have to manager during the turn. The batch size is limited by the carrying capacity or volume of the baggage cart, i.e., a standard baggage cart can only hold so many bags. This number too varies with the size of the bags in the batch. There is also a limitation on the number of baggage carts that

can safely be towed by a baggage tug due to the increased and less predictable turning radius occasioned by additional carts. The suggested maximum number of baggage carts that should be towed is four.

The next key observation that will impact the simulation is that, for two of the key elements of the turn - passengers and bags, while we can manage the start time for the processing of the batches, we cannot determine the end time of the completion of the batch due to the random arrival pattern of these entities. There are numerous reasons why a passenger will come running down the concourse too late to meet the cutoff for boarding the plane. Some of these reasons are the fault of the airline; others are the fault of the customer. Nevertheless, despite the reason, there are times when the gate agent must apologize to the customer but inform them that they have officially missed the flight. The same thing happens to the luggage of the passengers. To be able to be processed, it must arrive at the plane before the plane pushes back.

In summary, we have an arrival pattern of items that must be processed at key processing points of the gate area and the bag room. These items comprise the batches that must be processed during the turn. This is batch processing with known start times, but, in the case of passengers, due to the upper tail of the distribution of the arrival rate of the items, in some instances the ending of the processing of the batch is not signaled by the arrival of the last item, but rather by time. In the case of bags, a similar cutoff by time rather than the arrival of the last bag assigned to a flight will determine when work ceases and the process ends.

4.2.5 Operating from Standard Operating Procedures

Currently, airline ramp workers have a plan in their minds when they are deployed to a particular gate to accomplish all of the tasks required to turn the aircraft that might arrive at that gate. This could be thought of as the standard operating procedures for that airline. Nevertheless, when the aircraft actually arrives, in many instances that plan is followed only loosely. There is a parallel here between the turnaround operation and a play executed by a professional football team. Every play begins with a plan and every player knows what they should do when the ball is hiked. Nevertheless, not everything happens according to plan in a football play in the NFL despite flawless execution. Similarly, since the entire turnaround process is executed by human beings, the turnaround process is rife with process variation and assignable cause variation. It is this very variation that lends meaning to this research and the development of the heuristic that leads to improved performance measures and monetary savings.

For changes to occur in the task list of a worker, they would have to receive verbal instructions from a ramp supervisor to go and assist with another flight on another gate. They assume that they are to stay with a flight until that flight leaves so, logically, when one task is completed the entire group moves on to the next task on that flight. This practice leads to sub optimization with respect to global objectives, but it is easy to manage.

4.3 Scope of the study

The scope of this research will be limited to measuring, simulating, and managing those activities that take place in the narrow window between the time the inbound aircraft sets the parking brakes upon arrival at the gate and the time the pilot releases the

parking brakes before being pushed back from the gate. This time window was chosen because it is in this interval that management has the most control of the process and must make operational decisions that will determine if the flight pushes on time or is delayed. Data gathering and subsequent analysis will take place within this window and ignore the influence of the rest of the processes that must be accomplished to process passengers and luggage. Within this window, all activities, both planeside (on the ramp), and in the cabin of the aircraft will be considered for analysis. It is understood that this is an artificial boundary and that many other operations activities at the airport have an impact on the turnaround time. Examples of activities that will not be included in this study include: passenger ticketing and check-in at the ticket counter, security processing by the Transportation Safety Administration (TSA), pre-staging of ramp equipment, readiness of gate and ramp personnel, staging of outbound luggage and freight on the ramp prior to arrival of the aircraft, and the appropriate organization of outbound passengers in the gate area. Despite their impact on the turnaround process, these activities will not be considered in this study.

This research is simulation driven. The very nature of simulation is always a simplification of real life and thus its conclusions must be weighed carefully. That is, this study will capture data from current airline ground operations to construct a model of daily operations using discrete event modeling software. The model, once validated and verified, will be run to determine if applying TOC in this environment will improve operations performance measures. The output of the model should suggest to airline management responsible for ground operations a management philosophy and resource assignment method that will improve performance measures.

The unique characteristics of this service system will be used to define the needs of a system that can identify a bottleneck and thus provide helpful information to management. The aircraft turnaround process is a service system that could be said to be pure service in that there is no good that can be inventoried and sold to the customer (disregarding the negligible sales of food and drinks in the cabin). The main thing the customer wants and pays for is to be moved from one place to another with his appurtenances.

4.4 Narrative of the Process of Turning a Flight

When an aircraft arrives at the assigned gate, the first priority is to park it. This is known in the industry as marshalling the flight. This process requires three people to accomplish. The marshal uses orange wands or flashlights to guide the plane in and signal the pilot when to stop so that the door of the plane can be aligned with the jetway. Two others act as "wing walkers" to ensure that the plane does not hit any stationary objects and that no vehicle drives under the wing while the plane is parking. Once the plane has stopped and the brakes are set, the clock begins on the turnaround process.

This setting of the brakes marks the Actual Time of Arrival (ATA). The scheduled time of arrival is known as the Estimated Time of Arrival (ETA).

As can be seen in

Figure 5, three operations happen in parallel: cabin operations, fueling operations, and ramp operations. The initial activities of cabin operations consist of positioning the jetway over the left front door of the aircraft, opening the door, deplaning passengers from the aircraft, and cleaning of the aircraft.

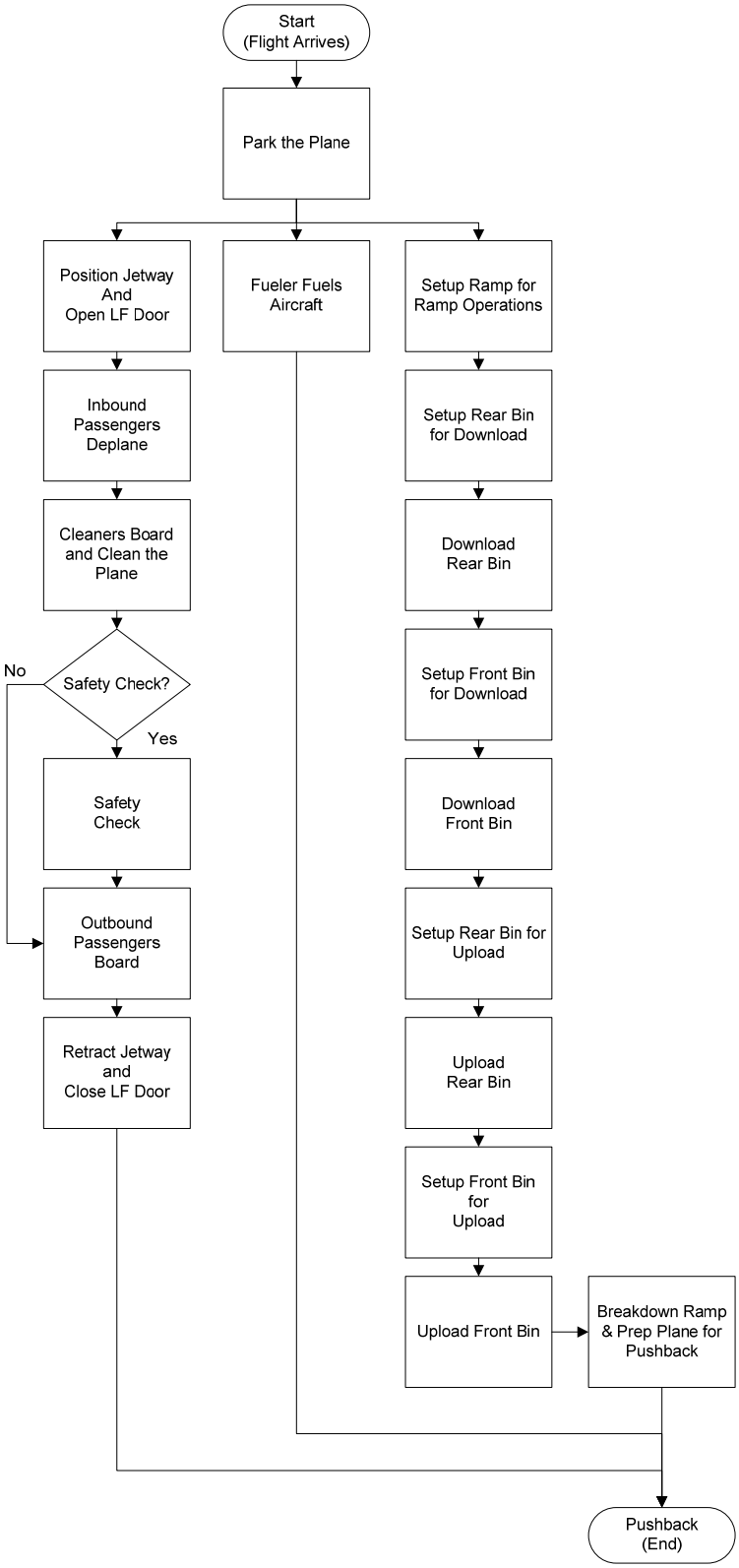

Figure 5: Process Map of the Turnaround Process Contemplated in this Study

Once cleaned, the aircraft may require a safety check depending on where it came from and where it is going. If a flight is coming from or destined to an international location, or if this is the first flight of the day, the aircraft requires a safety check. This procedure requires several individuals to check every seat and seat pocket, the lavatories and galleys for potentially harmful items and contraband. Following the safety check, the passengers may board, however, boarding is usually delayed until thirty minutes prior to departure so that people don't get too restless on the plane. This is the one activity on the flight that is delayed in this way; all others are done as soon as they can be. Once all passengers are boarded and their carry-on luggage stowed, the left front door is closed and the jetway retracted. This concludes cabin operations for the purposes of this model.

Fueling operations take place in parallel to cabin operations. A third party contractor arrives either in a fueling truck or towing a solar-powered pump to connect the underground fuel supply to the wing tanks. Following a brief setup, the fuel is pumped into the wing and center tanks at a fairly constant rate. Given the pounds of fuel to be uploaded, the time required for fueling can be fairly accurately estimated. Once uploaded a brief teardown of the fueling equipment is required.

Ramp operations are done in parallel with cabin operations and fueling operations. For the purposes of this model, ramp operations will consist in four main activities only. These four activities are also done in the order given here: 1) downloading the inbound bags from the rear bin, 2) downloading the bags from the front bin, 3) uploading the outbound bags into the rear bin, and 4) uploading the outbound bags into the front bin. Other activities comprise ramp activities but are not included in this

model since they are either insignificant in duration (like chocking the wheels), or are not on the critical path of activities for pushback (like adding potable water to the plane).

4.5 Data Collection for the Turnaround Process

Several trips were made to the Ft. Lauderdale airport to observe the process in general and gather data on the start times and completion times of certain key activities associated with aircraft turns at one of the major carriers operating at that airport. Airport identification was obtained to be able to get close to the operation and observe. Access was granted to the terminal area where passengers queue up and then board the aircraft. A high definition digital video camera was used to film aircraft turns from the window in the gate area overlooking the right side of the aircraft where all of the process takes place. The videos were time-stamped which allows careful review of the videos to ascertain both absolute times and elapsed times associated with the various steps of the process. The videos were analyzed for patterns in the arrival of work and the resources that were assigned to handle that work, and the arrival of specialized equipment key tasks such as fueling, catering, potable water service, push-back tractor, and lavatory service. This process allowed for review and was more productive than attempting to timestamp all relevant activities by hand in real time.

It was soon discovered that the scheduled gate time for many of the flights at the Ft. Lauderdale airport were excessively long on the whole. This is because Ft. Lauderdale is the major hub for the airline being studied. Consequently, gate turnaround times are scheduled to be longer to accommodate the variability in arrival times from the outstations that feed into Ft. Lauderdale. Scheduled turnaround times of 90 minutes and longer are not uncommon. The longer turn times allow for significant periods of idle

workers, machines and gate space which serves to confound the research. If turnaround times were shortened to the minimums at the hub, then the airline would incur increased costs associated with having responsibility for leaving passengers behind due to misconnection with another flight on the same carrier. These costs take the form of hotel accommodations, meals, and rebooking on subsequent flights. Consequently, it was determined that to be able to study turns that truly reduced the turn time down to the amount of time necessary to accomplish the tasks required for the turn, flights at the outstations of the airline would have to be captured. Outstations such as Ft. Meyers, Tampa, and Orlando are airports served from the hub that generally turn the plane right back around and send it back to the hub. The risk of leaving connecting passengers behind is not a factor at such stations, therefore the turnaround time is minimized to levels representative of the time it actually takes to accomplish the work without forced idle time of any of the resources. Such airports proved better for observation of quick turns as their scheduled ground times were in the range of 30 to 50 minutes.

4.5.1 Dataset 1 - Daily Operations Report Worksheet

Dataset 1 consists of a data file supplied by the airline listing certain performance measures for all flights of that airline at all stations for the month of January 2010. It is titled, "Daily Operations Report Worksheet". Once rows of incomplete data were removed, the remaining data represented 4,340 flights. Data points in this set of data included:

- Flight Number
- Tail Number
- Departure Date
- Inbound Departure City
- Inbound Scheduled Time of Arrival

- Inbound Actual Time of Arrival
- Scheduled Turn Time
- Actual Turn Time
- Difference Turn Time
- Departure Airport
- Destination Airport
- Scheduled Time of Departure
- Actual Time of Departure
- Departure Delay
- Scheduled Time of Arrival at Destination City
- Actual Time of Arrival at Destination City
- Arrival Delay at Destination City
- Type of Aircraft

This dataset was most useful in determining the average lateness with respect to ETA. The lateness for all flights was calculated by subtracting the estimated or scheduled time of arrival from the actual time arrival. The resulting column was converted to a histogram as shown in Figure 6.

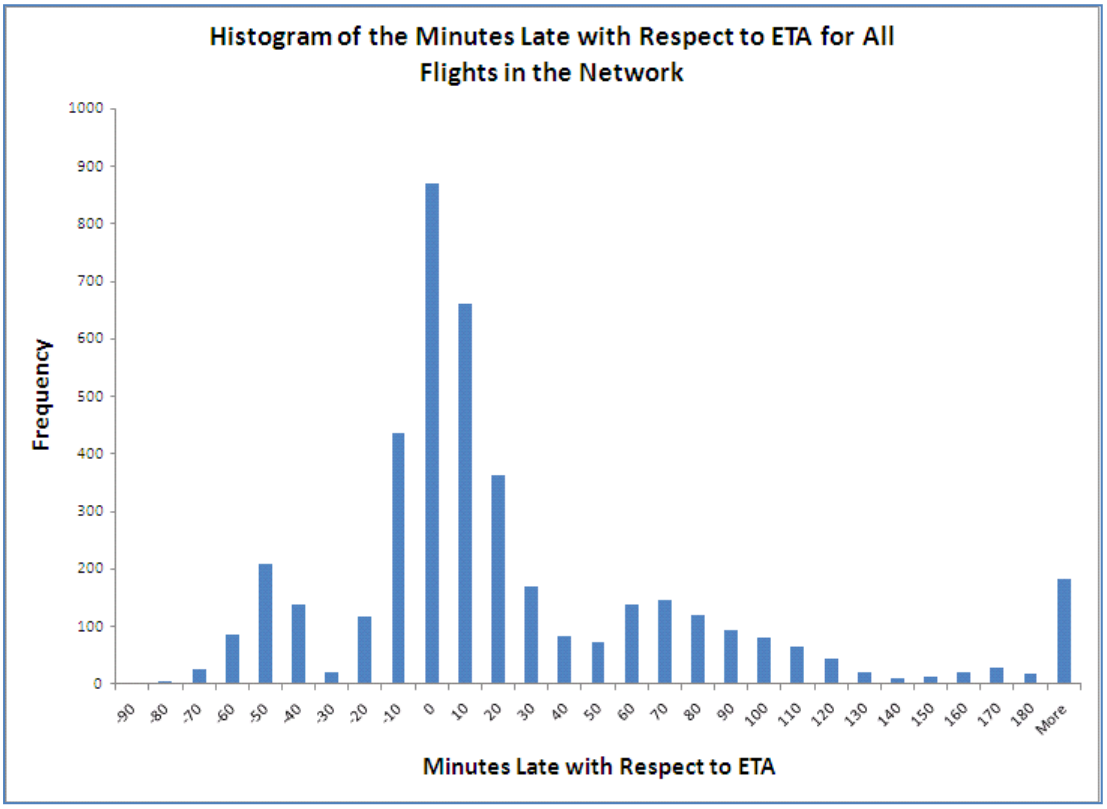

Figure 6 – Distribution of Arrival Lateness

Minimum lateness in the data was a flight that was actually 2:46 early. Maximum lateness was from a flight 18:30 late. The average lateness is 24.8 minutes; however, as the upper limit for lateness is infinite but the lower limit is not, this figure is heavily influenced by the outliers to the far right of zero. Note the positive skewness in the distribution of 3.48. Most flights arrive around the time they are scheduled. The median of the lateness is 3 minutes and the mode of the lateness is zero.

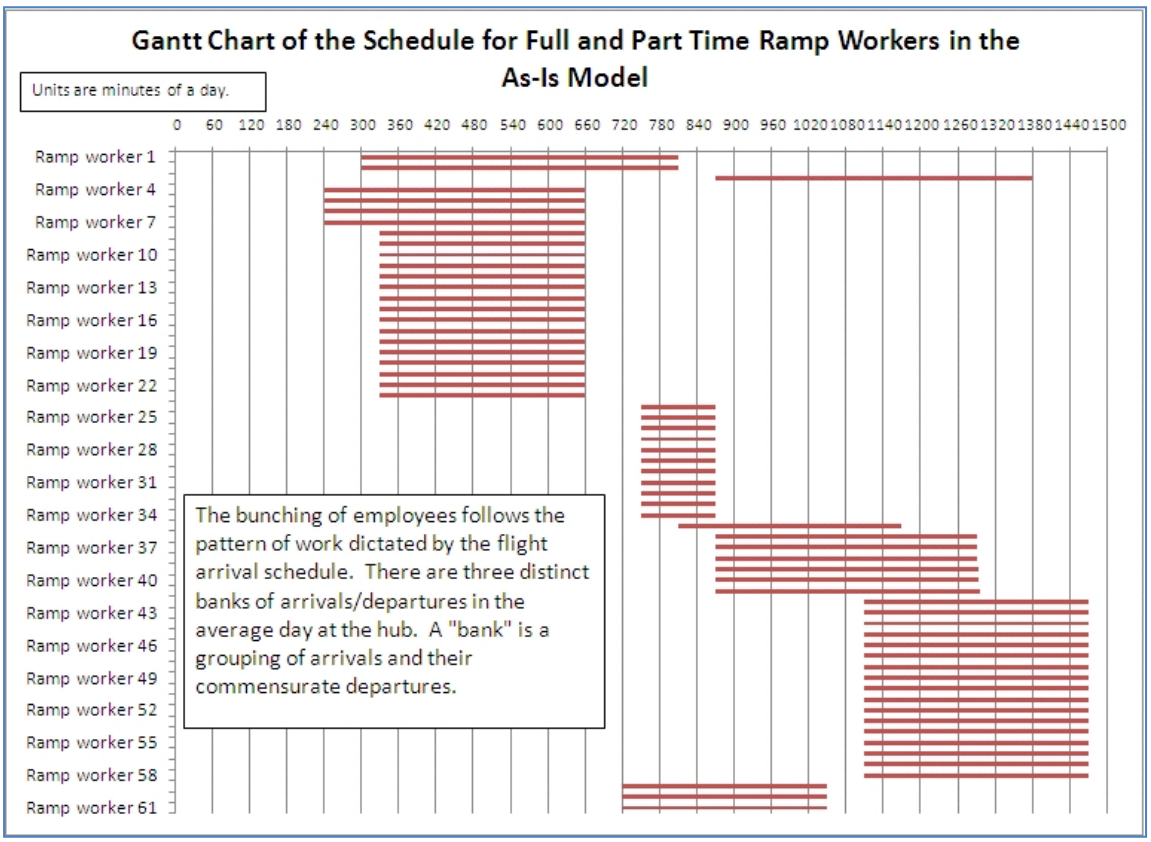

Figure 7 - Gantt Chart of Work Schedule of Workers

4.5.2 Dataset 2

Dataset 2 consists of a data file supplied by the airline listing the daily work schedule of all ramp supervisors, operations personnel, full-time ramp workers, and part-time ramp workers. The basic data from this file included time to clock in and out and

job assignment and was used to schedule the workers to accommodate the flight turnaround schedule. This file supplies the upper limit for the number of workers that are available at any given time. Figure 7 shows a Gantt chart of the work schedules. The pattern of work reflects the arrival pattern of the flights which come in three waves or "banks" in a day.

4.5.3 Dataset 3 – Data Collection Trip to Orlando

Dataset 3 came from an observational study from Orlando, Florida conducted in February 2008. The trip was made to observe and capture key characteristics associated with turnaround operations at the Orlando International Airport as this airport's operations were among the best in the network with respect to on-time performance, a traditional measure. Specifically, the relationship between carry-on baggage and boarding time was to be studied. Five graduate students from the FIU College of Business, assisted in the study. Each observer was given a stopwatch and a checklist containing key points at which time stamps were to be captured. Ten flights were observed and the data was compiled into dataset 3.

For these 10 flights, the beginning and ending time of the boarding process were captured as well's the number of passengers to deplane or disembark. The average time per passenger to deplane the aircraft was 4.6 seconds with a standard deviation of 1.4 seconds. A similar calculation was made with respect to the boarding process. The average passenger required 6.3 seconds to board the aircraft with the standard deviation of 2.1 seconds. The average load factor for these 10 flights was 78% and the average time required to board the aircraft was 13.7 minutes with a standard deviation of 4.1 minutes. Fueling the aircraft took an average of 9.0 minutes with a standard deviation of

3.0 minutes. The short fueling times can be attributed to the specific destinations for these ten flights.

4.5.4 Dataset 4

Dataset 4 was compiled from one month's worth of flight information summaries from the Fort Lauderdale airport. These summaries consist of forms that must be filled out by gate agents either during a turnaround or immediately following the departure of the plane. Their main purpose was to capture notes with respect to gate operations and in the event of departure delays, be a means of ascertaining the reason for the delay. Of all of the data used in this research, this data set had the most number of omissions and approximations since the people filling it out 1) had an interest in looking good on paper and 2) were not incentivized for accuracy of the data. Many timestamps are rounded in such a way as to suggest that the data was noted sufficiently after the fact as to have been approximated. Nevertheless, this data set provides insights from the point of view of the gate agent in the concourse.

This data set provides insight into the amount of time required to clean the aircraft. The average time required to clean the aircraft by a team of four cleaners was 9.3 minutes with the standard deviation of 3.5 minutes. Time to cater the aircraft was also noted; however, as catering is no longer a function of this airline it was not considered in the model. The current practice is to cater the flight once in the morning and bring a board nothing more than the ice at the outstations. The only other measures of any importance from these data sets were gross measures of the time required to deplane the aircraft and board the aircraft. But since the actual number of passengers per

flight was missing from the data, this data set could not be used to corroborate the data collected in dataset 3.

4.5.5 Dataset 5

Dataset 5 was compiled from observation of eleven flights at four different airports in Florida. This is the most detailed data set used in the study. A high definition digital video camera was used to capture the entire operation as seen from the vantage point of the second story window in the concourse. Flights were filmed from just before arrival at the gate and the setting of the brakes till slightly after pushback from the gate. The camera featured an elapsed time display. The resulting videos were then observed to capture key time markers for all aspects of ground operations on the right side of the aircraft on the observed flights.

Data captured from these videos includes all aspects of the ramp setup upon arrival of an aircraft. The initial timestamp was made at the setting of the brakes upon arrival. Timestamps were captured of the chocking of the wheels, the connection of ground power and ground air-conditioning equipment, the placement of safety cones in front of the engines and at the rear of the aircraft and the time at which the jetway coupled to the aircraft and the front door was opened.

The next set of operations observed dealt with the set up for rear bin operations. Timestamps were captured from the opening of the rear bin, the positioning of a belt loader to download bags from the rear bin in the positioning of baggage carts around the bottom of the belt loader to receive the bags that would be downloaded from the rear bin. In each case the beginning and ending of these tasks was captured.

The next set of observations dealt with set up for the front bin operations. Once again, timestamps were made for the beginning and ending of the opening of the front bin, and the positioning of a belt loader to receive the bags downloaded from the front bin, and the positioning of baggage carts to receive those downloaded bags.

The next four timestamps all dealt with fueling operations. A timestamp was made of the arrival of the fuel truck. Following his particular setup operation, a timestamp was made of the time at which the fuel hose was connected to the aircraft. The timestamps of fueler disconnect and the departure of the fuel truck from the side of the aircraft completes this section.

In the next section, timestamps were captured reflecting when the cabin cleaning crew ascended the jetway stairs and when they descended those same stairs having completed the cleaning of the aircraft. It was realized, that the elapsed time between these two timestamps does not accurately reflect the time required to clean the aircraft as the cleaning crew will often go up the stairs and wait at the top of the jetway out of sight of an observer until all passengers deplane and they can board the aircraft to begin their operations. Nevertheless, good data on cleaning times was to be had from dataset 3.

In the next section, timestamps associated with several necessary but minor activities and that comprise each turn were captured. The first was the connection of the pushback tractor and tow bar. As this process is usually done by a select few who are certified to pushback aircraft, the activity is not on any of the paths limiting the pushback time. Like so many of the smaller activities such as the placement of safety cones and the chocking of the landing gear which are done during a turnaround connecting the tow bar

and tractor was considered to be inconsequential to this analysis of the process. This same logic applies to water and lavatory services provided for the aircraft by specialized equipment by operators dedicated to that function. Last in this section was the beginning and ending timestamp associated with the pilot walk around ground inspection.

The next section contains timestamps essential to this research involving rear bin download operations. The first timestamps captured the time at which the first bag arrived at the bottom of the belt loader and when the last bag arrived at the bottom of the belt loader. Notes with respect to the number of times the baggage carts had to be adjusted and the time associated with that adjustment follow. A second viewing of the videos yielded a count of the number of bags downloaded from the rear bin. From these timestamps the time required per bag for downloading can be calculated based on the number of workers used in the operation.

Rear bin upload operations follow. This particular process is slightly different than the others due to the arrival pattern of the outbound bags. Since all passengers must check in at least one hour before the flight departs, customarily the majority of the outbound bags are already stacked in baggage carts and positioned in the gate area prior to the arrival of the aircraft. This large batch of bags usually comprises the bulk of the work that must be done to upload the rear bin. However, at some hub airports, it is not uncommon to receive multiple batches of bags from connecting flights whose passengers make a connection with this one. The size of these batches of connecting bags is usually less than five and the time required to process them is almost negligible. Therefore, only the main upload of the outbound bags for the front and rear bins will be considered in this study. Nevertheless, timestamps were captured of the time of arrival of the first batch of

bags, the number of bags in that batch with a note as to the number of baggage carts required to carry the bags. The beginning of the upload operation and the ending of the upload operation of this first batch of bags complete the measurement of this operation.

The next two sections involve the downloading of inbound bags from the front bin and the uploading of the outbound bags to the front bin and are very similar to rear bin operations. As their timestamps are identical, they will not be repeated here.

The final section of observational data was concerned with all of those activities required to break down the ramp and disconnect ground support equipment prior to pushback. Timestamps were made of when the signal was given to disconnect ground power, when the ground power unit was properly disconnected and stowed, the beginning and ending of the removal of the belt loader used in the rear bin in the beginning and ending of the removal of the belt loader used for the front bin, the closing of the front and rear bin doors, and the disconnection of ground air-conditioning units. In one case a portable jet engine had to be connected to the aircraft to provide pressurized air to start the engines. The last timestamp involved the removal of the jetway and the pushback of the aircraft.

4.5.6 Distributions and Process Times Based on the Real data

Specific real-life flight numbers will be used with their respective attributes to measure performance in the system. These flights have known origins and destinations, estimated times of arrival and departure, passenger capacity, and fuel requirements. Since these attributes are known, it is not necessary to simulate them. Other variables, such as the average time required to process a single bag or the time required to board a

single passenger were determined from observational study. The summary of processes, distribution means, and standard deviations is given in Table 2.

Table 2 - Processes, Datasets, and Observed Data

Process	Dataset	N	Mean (min)	Std. Dev. (min)
Time to position jetway and open the door	Dataset 5	11	1.317	0.322
Time per passenger to disembark	Dataset 3	10	0.077	0.023
Time to clean the aircraft	Dataset 4	118	9.300	3.500
Time to perform a safety check on inbound or outbound international flights	Trial Run	1	8.700	NA
Time per passenger to board the aircraft	Dataset 4	130	0.232	0.117
Time to setup for fueling the aircraft	Dataset 5	11	10.650	4.850
Time to fuel the aircraft	Dataset 5	11	16.300	2.367
Time to setup ramp operations	Dataset 5	11	2.700	1.800
Time to setup for rear bin download operations	Dataset 5	11	5.433	2.433
Time to download bags from the rear bin (regardless of inbound bag count)	Dataset 5	11	6.517	4.617
Time to download inbound bags from rear bin in seconds/bag	Dataset 5	11	0.107	0.030
Time to setup for front bin download operations	Dataset 5	11	3.000	1.800
Time to download bags from the front bin	Dataset 5	4	1.617	0.457
Time to upload bags to the rear bin	Dataset 5	11	5.983	3.917
Time to upload bags to the front bin	Dataset 5	11	1.422	1.322
Time to break down the ramp in preparation for pushback	Dataset 5	11	7.750	4.750

4.6 Characterizing this Service System

Several things characterize the service system used in this study. These characteristics are what make the application of Bottleneck Ratio Method relevant. Things that characterize the service system in this study are:

1. Management has complete control over the dispatching of workers as they are company employees. This completeness of managerial control was favorable to the research as it allows for access to all levels of management practice and pertinent operations data.
2. Management has foreknowledge of the amount of work to be accomplished on both the inbound and outbound activities in the form of number of bags to be loaded and where they are to be loaded. Given the ETA, management can forecast the time in which the work must be accomplished. Given the ATA, that forecast becomes a certainty.
3. There is a setup function for each of the four major activities. Each of the activities involves a setup process in which equipment will be positioned, bin doors opened, and baggage tugs and carts positioned to collect the downloaded luggage.
4. Due to the low skill level required to accomplish the tasks, there is a flexibility of the common resources of ramp workers; they can work on any flight. The skill level of the workers involved in the process is very low. Their main activity is comprised of muscling around passenger's luggage and moving equipment. Therefore, the human resources required to accomplish the tasks are fairly interchangeable; they can be easily re-tasked. Using modern communication

equipment such as radios, and pager type devices, ramp workers could easily be dispatched to higher priority activities. This resource flexibility is one of the keys to the success of this model. Knowing when to reassign them and to what task is the key to the model.

5. There is high process variability. The process of an aircraft turn is never accomplished the same way twice; each turn is unique. Even the game plan of the process will be different for different airports, and constraints can shift from activity to activity during the turn as the flexible resources that accomplish the work either speed up or slow down or are moved and re-tasked. This shifting of the constraint is at the heart of this model and is necessary for the model to have meaning. It will be shown that various departments or tasks can become temporary bottlenecks when the capacity to accomplish the task is insufficient for the amount of time remaining in which the task must be accomplished.
6. Management has a limited global view across all flights on the ground at a time. There is a lack of a global view of the process. For example, what motivates a ramp supervisor to give priority to one flight over another? This is the key consideration when we think of a global view of the process.
7. For operations as they are in the as-is version, the number of flights that depart late as well as the magnitude of the lateness are both measured.
8. The work of a "turn" must be accomplished by a certain clock time. Though this deadline is known for each turn, the start time for the operation is not as inbound flights can be delayed or arrive early.

9. The processing of people and luggage during a turn is accomplished in batches. The batches are directly correlated with the arrival pattern of the passengers and luggage to those areas where they will be processed. Inbound luggage is downloaded as a single batch as are inbound passengers. Outbound luggage and passengers have a different arrival pattern and so are usually processed in multiple batches.
10. The number of workers on a particular flight ranges from zero to six. Zero is never actually used even on long turnarounds but is used to initiate the variable. The number six was chosen as the upper limit since assigning more than six would lead to diminishing returns as workers begin to interfere with each other in the cramped spaces in which they work.
11. Up to six workers, the rate at which the work gets accomplished is faster with the addition of each worker.

4.7 All Flights are NOT of Equal Importance

One of the major expenses for the airline is the expense of compensating passengers who miss their connections on an airline. The airline in this study spends over $12M annually on meals, hotels, transportation and delivery of passenger luggage for those passengers that can claim that it is the airline's fault that they missed their connections at a hub airport. The average is $380 per misconnecting passenger. Consequently, it is particularly important that flights that are destined for hubs depart on time to minimize the possibility of arriving late at their destinations to avoid the costs of compensating misconnected passengers. Contrast this with the consequences of a late arrival in a non-hub outstation. The lateness of such a flight means that passengers and

those that have gone to collect them are slightly inconvenienced, but the airline has no negative monetary consequence excluding the loss of good will. Therefore, the to-be model must find a way to pay special attention to these potential money-losing flights.

Flights bound for hubs carry connecting passengers must receive special attention. However, on-time departure is important for all flights of the airline to maintain a reputation in the customer base and reliability of flight schedules across the network. Additionally, if one takes the long view, the aircraft that is bound for an outstation early in the day must soon return since each aircraft flies an average 5 flights in a day. There is a domino effect if a flight is late early in the day that has repercussions throughout the day. So, the dispatch policy cannot be merely to send six workers to every flight destined for a hub. In some cases this would be overkill and would take away resources from other flights where they are also needed to leave on time.

4.8 The As-Is Model

"Make everything as simple as possible, but not simpler." -Albert Einstein

The model was created based on 36 actual flights scheduled in and out of the Ft. Lauderdale airport. This represents a day's worth of turns from the airline in the study. A Gantt chart of the schedule of arrival and departure for the flights is shown in Figure 8. The darker bars in the chart represent flights that are not outbound to one of the hubs serviced by this airline. Lighter bars represent flight turns that are destined for hub locations. These flights carry connecting passengers and must therefore receive special attention. Nevertheless, on-time departure is important for all flights to maintain a reputation in the customer base and reliability of flight schedules across the network.

These 36 flights were replicated 365 times to simulate a year's worth of doing the same flights every day. The total number of flights in the simulation is thus 13,140.

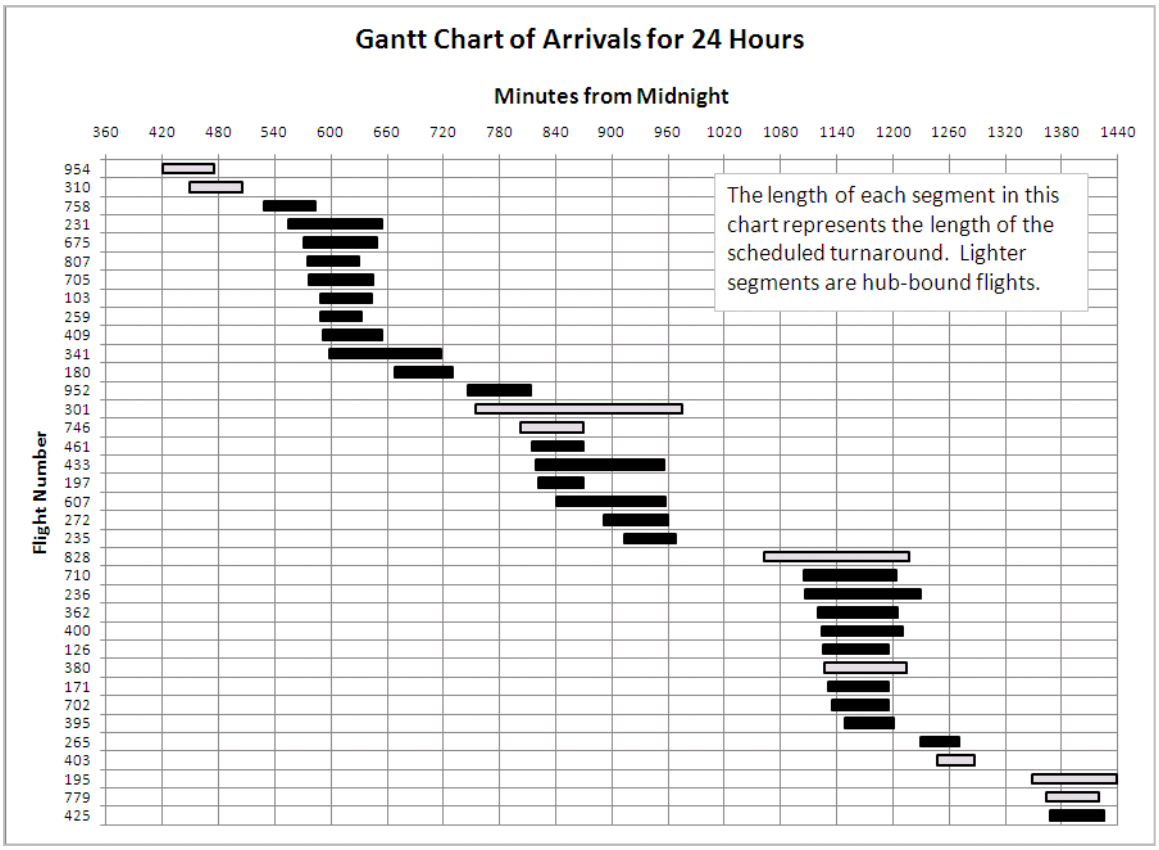

Figure 8 – Gantt Chart of Scheduled Turns Used in the Model

4.8.1 Modeling Variation

Variation for several attributes of the flights were simulated using random number generators within a given distribution prior to execution of the model using StatPlus, a statistical add-in for Microsoft Excel. For example, the number of minutes late with respect to arrival for the flights was modeled using data from Dataset 1 as shown previously in Figure 6. This dataset for which N = 3,463 contained data points representing the number of minutes late for every flight in the system for a given month. The input analyzer tool of Arena was invoked to determine the equation with the best fit

for modeling the data. Possible distributions include: Beta, Empirical, Erlang, Exponential, Gamma, Log Normal, Normal, Triangular, Uniform, and Weibull. For this particular distribution, shown in Figure 9, the log normal function was chosen as having the best fit to the data. The expression of the line that would be used to represent the data was given as "30 less than a lognormal distribution with a mean of 54.2 and a standard deviation of 57.8". The "-30" models those times that flights arrive early or before their scheduled ETA. The square error for this distribution with respect to the actual data is 0.0122. A Chi-Square test is run automatically in Arena to determine goodness of fit. Using 36 intervals and 33 degrees of freedom, the chi-square test produced a test statistic of 729 with a corresponding p-value of < 0.005 suggesting a good approximation of the data with this log-normal distribution.

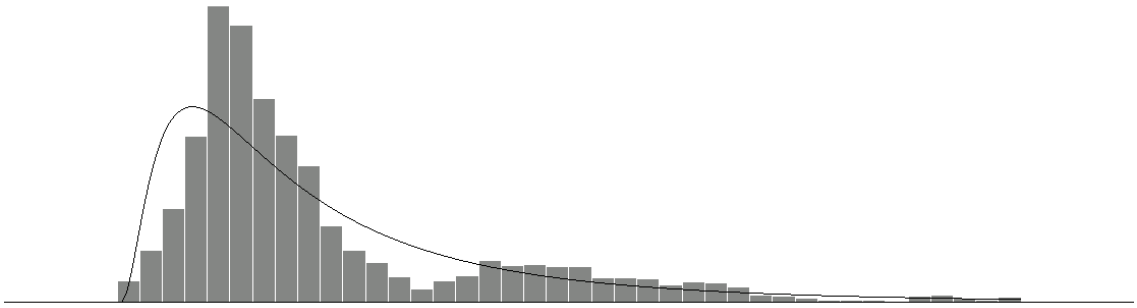

Figure 9 - Modeling Arrival Lateness

Once the equation modeling the phenomenon is known, the Excel add-in of StatPlus can generate random numbers based on the equation. These numbers are then added to the scheduled time of arrival to simulate the actual time of arrival which would then include earliness or lateness.

Inbound passenger counts were simulated using from the load factor distribution supplied by the airline multiplied by passenger capacity of the specific aircraft. The data

was found to follow a normal distribution with a mean of 87% and a standard deviation of 6%. Of course, all values in excess of 100% are truncated back to 100% as a flight cannot operate with too many passengers. Again, StatPlus was used to generate random numbers needed to model passenger loads. The remaining flight attributes and their distributions used in the model are shown in Table 3.

Table 3 – Flight Attributes, Descriptions, and Mathematical Expressions

Flight Attribute	Description	Distribution
Day number	Day of the year starting Jan. 1	Numbers 1-365 generated in Excel
Inbound Flight Number	Distinguishes a unique flight	Text field taken from airline schedule
Passenger Capacity	Passenger capacity for that flight given the type of aircraft used	Airbus 319 – 148 Airbus 320 – 178 Airbus 321 – 218
City of Origin	Used to determine whether or not a safety check is required	Text field taken from airline schedule
Inbound International	Used to determine number of bags each passenger will bring	Text from airline schedule. Total bags this flight calculated from Inbound Passenger Count using 0.9 bag/domestic passenger and 1.3 bags/international passenger.
Absolute Estimated Time of Arrival	Number of minutes from midnight, Jan 1 that the flight is scheduled to arrive	Taken from airline schedule.
Minutes Late	Random number of minutes the flight will be late	Generated by subtracting 20 from a log normal distribution with a mean of 4 and a standard deviation of 0.41.
Absolute Actual Time of Arrival	The time at which the flight will actually arrive in the system	Based on the ETA and the Minutes late.
Inbound Passenger Count	Random number generated to represent the number of passengers on the inbound flight	Generated from load factor distribution given by the airline multiplied by passenger capacity. Expression used: PaxCapacity *Normal(.87, .06). Outputs were truncated if they exceeded 100%.
Inbound Bags Rear Bin	Number of pieces of luggage in the rear bin of the inbound flight	Generated using uniform distribution from 65% to 99% of total bags this flight.
Inbound Bags Front Bin	Number of pieces of luggage in the front bin of the inbound flight	Inbound Bags Total minus Inbound Bags Rear Bin.
Safety Check	One or zero indicated whether	If inbound from or outbound to an

	the flight needs a safety check	international location, safety check =1
Fuel Required	Number of pounds of fuel scheduled to be uploaded.	Taken from historical data for these destinations
City Destination	Used to determine if the flight is bound for a hub	From flight schedule
Outbound International	Used to calculate number of bags per passenger for outbound flight.	From flight schedule. Total bags this flight calculated from Outbound Passenger Count using 0.9 bag/domestic passenger and 1.3 bags/international passenger.
Outbound Passenger Count	Number of passengers on this outbound flight	Generated from load factor distribution given by the airline multiplied by passenger capacity. Expression used: PaxCapacity *Normal(.87,.06). Outputs were truncated if they exceeded 100%.
Outbound Bags Rear Bin	Number of pieces of luggage to be uploaded to the rear bin of the outbound flight	Generated by multiplying the number of total Outbound Bags by a uniform distribution from 65% to 99%.
Outbound Bags Front Bin	Number of pieces of luggage to be uploaded to the rear bin of the outbound flight	Outbound Bags Total minus Outbound Bags Rear Bin.
Estimated Time of Departure	Scheduled time the plane should push back	Taken from flight schedule
Destination Hub	Dummy variable to signify which flights are going to ACY, ORD, DTW, or MYR airports (hubs)	1 = hub, 0 = non-hub
Connect Passengers	Random percentage of Outboard Passenger Count that will be making connections	Varies by destination. FLL 40%, DTW 15%, ACY 10%, ORD 5%, MYR 10% of Outbound Passenger Count.
Monetary Penalty	Cost to the airline if this flight is critically late	Calculated value at $380 per connecting passenger.

In the as-is model, the policy of the airline is to assign three workers to each gate. These workers are to service any flight that arrives at the gate with the mandate of getting all flights out on time. This could be considered a push system. The available number of workers was modeled at 26 for both the as-is and to-be models. The rate of work varies only with the individual exertion of the workers rather than the number of workers

assigned to a task. For purposes of modeling, we will assume that workers work at a constant rate, i.e., the processing rate per bag remains constant. Each flight is treated to be of equal importance. There are nine gates in this simulation, however, in the event that a tenth plane lands and the gates are all occupied, that plane is sent to a remote gate in an adjacent concourse rather than be left out on the runway. Consequently the number of gates will be modeled without constraint on the number of gates. Each team of three is given a single belt-loader to work with.

Entities or flights are created in the simulation and assigned the attributes of the corresponding flight. The entity is then cloned twice to yield three identical copies of the flight for parallel processing. The processes of the cabin operations, fueling operations and ramp operations are then conducted in parallel just as they are in the real world. The emphasis is on the ramp workers since that is where the heuristic in this study is applied. Consequently, the cabin operations are simplistic as are the fueling operations. Cabin operations consist of four activities: 1) Inbound passengers deplaning the aircraft, 2) Cleaning the aircraft, 3) Safety check if needed, and 4) Outbound passengers boarding the aircraft. Fueling operations are modeled given the amount of fuel that is to be uploaded on the flight. These amounts were collected from the records of the airline so that in the simulation, the time required to upload the fuel truly reflects the time spent on a flight to this outbound destination.

These activities are modeled using simple expressions based on the number of inbound and outbound passengers. Table 4 gives a listing of all of the expressions used to model the as-is version of the simulation. Due to the number of human beings involved in the process, there is much variation in the turning of an aircraft making

validation a challenge. Additionally several of the sources of data were collected by individuals who had an interest in looking good on paper or who did not take the care needed for accuracy as they might have. Nevertheless, comparisons with the available real data will show that the model is a good approximation of the reality of the turn.

Table 4 – Variables and their Expressions for the As-Is Model

Variable Names in the As-Is Model	Expression Used in As-Is Model
Position jetway and open door	Triangular (.75, 1, 2) minutes
Deplane inbound passengers	(InPaxCount*5) seconds
Clean cabin of aircraft	NORM(882,473) seconds
Safety check	Triangular (6, 8, 12) minutes
Board Outbound Passengers	OutPaxCount*9 seconds
Close Door and Retract Jetway	Triangular (.75, 1, 2)
Fueling the Aircraft	NORM(640,291)+(FuelReq*.09) seconds
Setup Ramp	NORM(326,146) seconds
Gate Checks	Triangular(0,3,6) minutes
Setup Rear Bin for Download	Triangular(1, 2, 4) minutes
Rear Bin Download	InBagsRB*(1/RateDownGiven3) minutes
Front Bin Setup for Download	Triangular(1, 2, 4) minutes
Front Bin Download	InBagsFB*(1/RateDownGiven3) minutes
Rear Bin Setup for Upload	Triangular (1, 2, 4) minutes
Rear Bin Upload	OutBagsRB*(1/RateUpGiven3) minutes
Front Bin Setup for Upload	Triangular (1, 2, 4) minutes
Front Bin Upload	OutBagsFB*(1/RateUpGiven3) minutes
Front Bin Teardown	Triangular (1, 2, 3) minutes
Setup for Pushback	NORM(4,2)

It is obvious that workers can work at varying speeds, however, for purposes of the model, the rate at which an employee can process a piece of luggage will be modeled as constant. Through observational study, the rate at which bags are processed was approximated according to the Table 5.

Table 5 - Processing Rates Given the Number of Ramp Workers

Number of Ramp Workers	Processing Rate in Minutes/Bag	Processing Rate in Bags/Minute
For six ramp workers	.0733 min/bag	14 bags/min
For five ramp workers	.0850 min/bag	12 bags/min
For four ramp workers	.0917 min/bag	11 bags/min
For three ramp workers	.1067 min/bag	9 bags/min
For two ramp workers	.1667 min/bag	6 bags/min
For one ramp worker	.5000 min/bag	2 bags/min
For zero ramp workers	100000 min/bag	0 bags/min

The work of the ramp consists of four main activities: 1) downloading and inbound bags from the rear bin; 2) downloading the inbound bags found in the front bin; 3) uploading the outbound bags in the rear bin; 4) uploading the outbound bags in the front bin. Each of these four activities involves a setup of equipment in the form of positioning a belt loader, baggage carts, and baggage tugs. For purposes of the model, these four activities will be treated as if they are dependent events and that the order will not change. Downloading the bags from the rear bin will always be the first activity and uploading bags to the front bin will always be the last activity. In addition to these four main activities that comprise the bulk of the work done on the ramp during an aircraft turn, there is a brief setup and tear down function that involves the other support equipment necessary for the turn such as ground power units, ground air conditioning units, and the placing of various safety markers around the plane.

Once cabin operations are completed, fueling is completed, and all ramp operations are completed including the removal of ground power units, ground air conditioning units, and safety devices, the flight is almost ready to be pushed back. A tow bar must be connected to the nose gear and a tractor connected to the tow bar to execute the maneuver. This set of activities is all lumped together and is known as the setup for pushback.

At this point the jetway is retracted and when the pilot is ready, he gives a signal to initiate the pushback of the aircraft. The act of releasing the brakes of the nose gear marks the end of the turn. At this point in the model, a timestamp is made of the current operating time and all performance measures are written to a file. Chief among these is whether or not the flight is "critically late". Critically late will be defined as any flight that departs 30 minutes or more beyond its estimated time of departure. This arbitrary figure was given by the airline as a typical maximum value that a flight at a destination hub would wait for connecting passengers who were inbound delayed. If connecting passengers would not arrive at the gate ready to board within 30 minutes of ETD at the hub, they would be left behind to be able to accommodate the needs of the passengers already on board. Being critically late only has negative repercussions if the flight is bound for a hub. If the flight is not bound for a hub being critically late may negatively impact the company image, but they will not have immediate financial impact on the cost associated with misconnecting passengers.

4.9 Validation of the As-Is Model

The first priority is to verify and validate the model to assure that the outputs of the model correspond to the process being modeled. Verification of the model consists in

ensuring that the model functions as intended. Verification ensures that the model contains all of the components required to represent what is being modeled. Verification is also concerned with making sure the model is bug free.

The process of validating a model consists of ensuring that the model represents reality. (Adeleye and Chung 2006) discuss verification and validation of a similar model used to represent aircraft turnaround procedures. They suggest that there are two stages in the validation process: face validity and statistical validity. Face validity is when a domain expert who understands the system being modeled and the intent of the model provides a critical appraisal of whether or not the model is suitable. The model used in this dissertation was created in collaboration with airport operations executives and an additional industrial engineer. Though it is a simplification, as is common to all models, it can be said that it has face validity. Statistical validity requires the use of statistics to compare the model performance and the real system performance given identical input parameters. Comparison of means tests are typically applied to corroborate statistical validity. Tests for statistical validity follow.

4.9.1 *t*-test for the Differences in the Mean Between Scheduled and Actual Turn Times

A two sample t-test for the difference between the percentages of flights in each of the bins of the histogram shown in Figure 10 was run. Formally stated:

1. H_0: There is no difference between the actual system and model system times for the difference in actual and scheduled turn times.
2. H_a: There is a difference between the actual system and model system times.
3. $\alpha = 5\%$.
4. Test statistic for the t test was -0.001 (p-value of .999)
5. Conclusion: there is no evidence to suggest that the null hypothesis is false. Therefore, we assume the models to be similar.

The outputs from both systems and the scheduled times are seen in Figure 10.

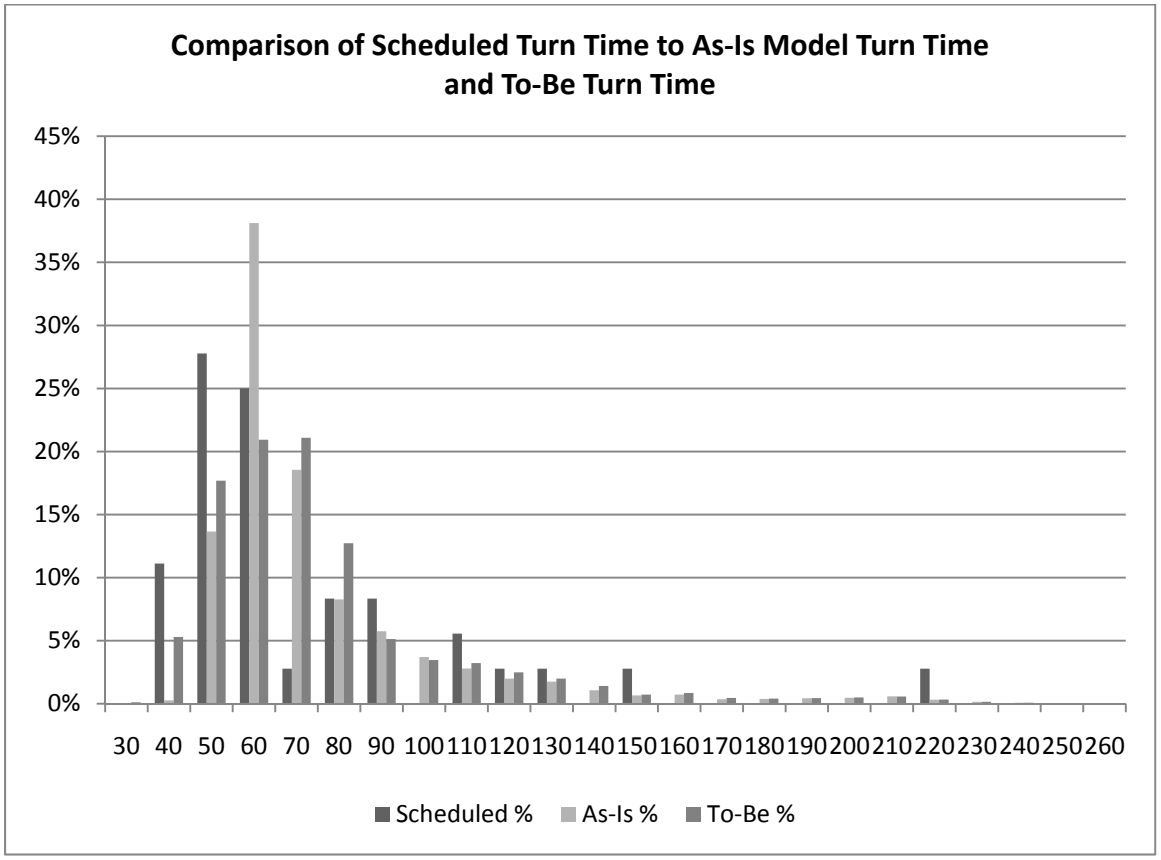

Figure 10 - Comparison of Turn Times

4.9.2 Minimum and Maximum Values from Observed and Simulated Outputs

The minimum value for scheduled turn times was 40 minutes. The maximum value for scheduled turn times was 221 minutes. The minimum value for turn times output by the as-is model was 46 minutes and the maximum was 251 minutes. These represent a 14% and 15% variation from the scheduled and model times respectively; an acceptable variation.

4.9.3 Comparing Simulated versus System Arrival Delays

When comparing the percentage of flights that arrive a certain number of minutes late, it can be seen in Figure 11 that the simulated and actual values used in the model are similar. This is a comparison of the percentages that fall in each of the bins of the histogram as the samples sizes were of quite different sizes. Visual observation of the histograms reveals patterns in the actual data that do not exist in the model data such as incidences of flights that arrive significantly early or considerably late. Such anomalies were not generated by the distribution used to simulate lateness for the as-is model.

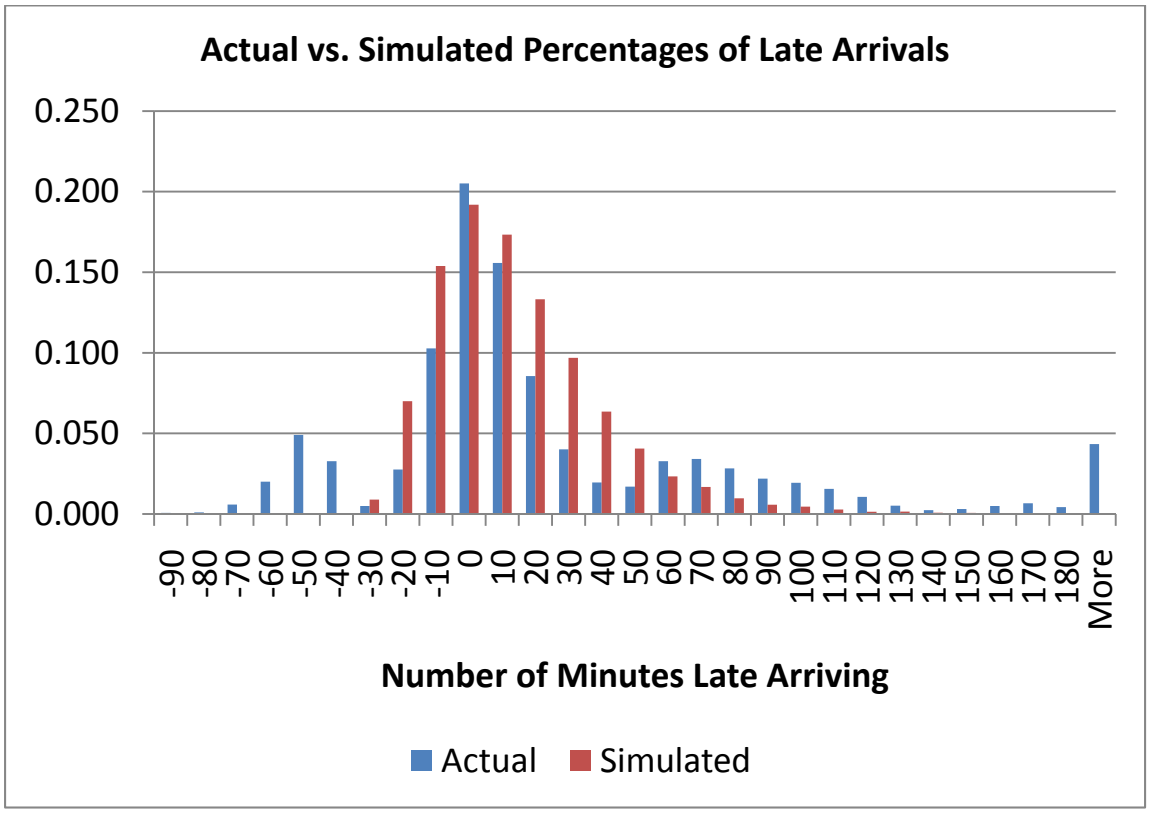

Figure 11 –Actual vs. Simulated Arrival Lateness

A paired t-test run on the output data from the as is model compared with actual data from 3,400 flights from January 2010 suggests that there is no evidence to support the hypothesis that the difference in means is other than zero.

Formally:

1. H_0: There is no difference between the mean differences between the actual system and the model system times for the number of minutes late arriving.
2. H_a: There is a difference between the actual system and model system times.
3. $\alpha = .05$
4. p-value of 0.9945; cannot reject the null hypothesis.

The high p-value suggests that we do not have enough evidence to reject the null hypothesis. This confirms that the model, as created, is verified to produce results corresponding to actual data.

4.9.4 *t*-test of the Differences in Simulated and Actual Turn Times

A t-test conducted on the difference between the simulated turn times and the actual turn times using a year's worth of flights from the simulation run of the as-is model and flights recorded at the FLL airport suggests that there is no difference between the simulated and actual turn times. The null hypothesis is that the means are the same. The alternative hypothesis is that they are not. Formally stated:

1. H_0: There is no difference between the actual system and model system times.
2. H_a: There is a difference between the actual system and model system times.
3. $\alpha = .05$
4. t-critical for a two-tailed test = 1.375
5. p-value of 0.169, cannot reject the null hypothesis.

The high p-value suggests that we fail to reject the null hypothesis that the means are the same. This is evidence that the model can produce a similar mean difference between the actual turn time and the scheduled turn time. The p-value of 0.169 is not sufficient

evidence to reject the null hypothesis at the 5% level of significance; however the significance of the p-value is not trivial. The difference can be attributed to the simplifications in the model.

4.10 Applying the Theory of Constraints to Formulate the To-Be Model

The main manipulation of the as-is model will be the inclusion of a method for dynamic and adaptive dispatching of ramp workers to tasks with high priorities while simultaneously seeking to accomplish all tasks with a minimum number of workers to minimize labor costs, minimize the number of flights that are late, and minimize the sum of that lateness. The heuristic proposed in this study has as its purpose the deploying and subsequent redeploying of human resources to various tasks using the principles of the Theory of Constraints as a guide. Based on its proper use, the heuristic should improve the objective function and also improve performance measures on other more traditional measures like number of flights with late departures. The heuristic will focus only on a portion of the as-is model; that portion involving ramp workers. Figure 12 captures the flow of logic of the method as it relates to the case study of airline ramp operations.

The idea of urgency assigned at the end of the top row in the flowchart is not generally part of TOC. The term "urgency" is being used to denote those jobs that impact the objective function and that are also possible to accomplish prior to the due date. The logic flow shown in Figure 12 addresses two objectives specific to airline turnaround operations: one financial and the other operational. In the first part of the logic wherein decision nodes inquire as to the monetary risk associated with missing the deadline for this job and whether or not completing the job by the deadline is a possibility, the system is determining the priority with which workers should be selected for this job regardless

of the number of workers that will be assigned to the job. This is the financial part of the method and assures that when it comes time for assigning workers that if this job bears a financial penalty for non-completion by the deadline, and the amount of work remaining for that job can be completed by the deadline, the job will have a high priority when it comes to seizing workers in the event that workers are scarce and the system must choose between different jobs for the assignment of scarce workers. In this objective is in keeping with the main objective of most companies to make money now and in the future and is true to the TOC.

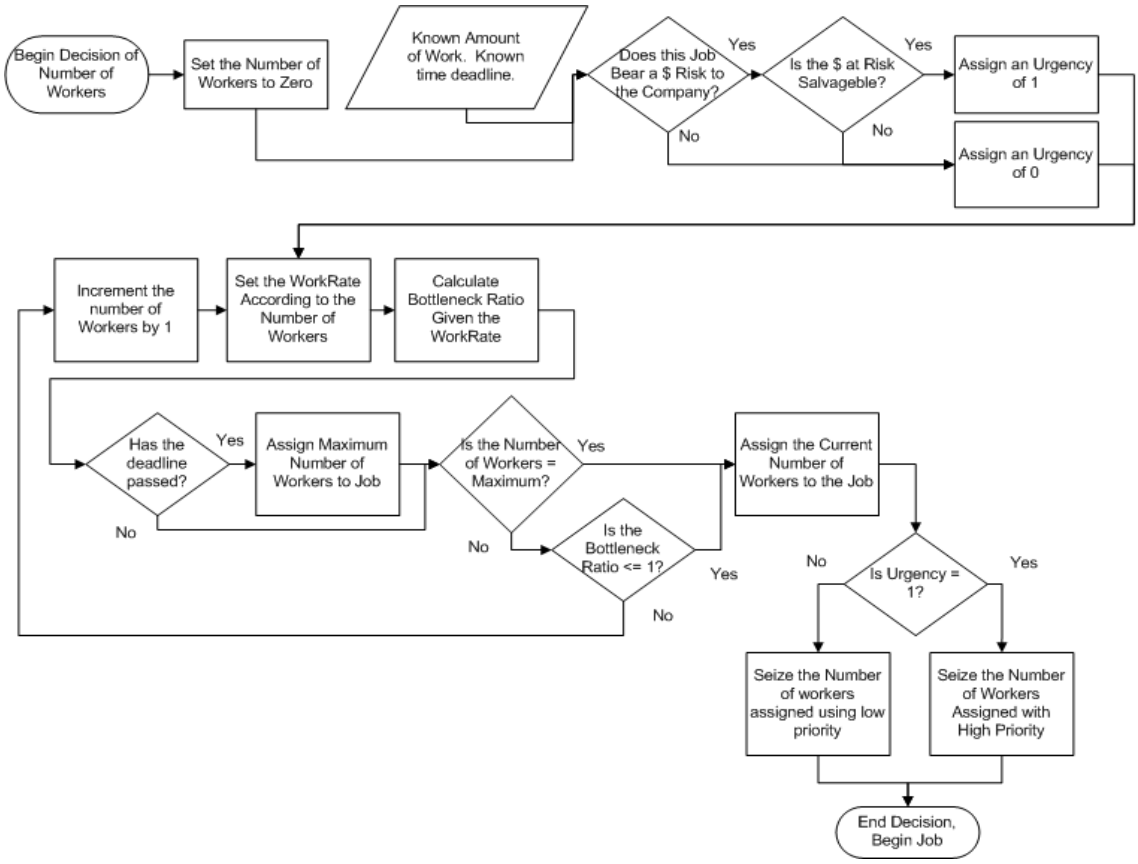

Figure 12 - Logic for the To-Be Arena Model in the Airline Case Study

4.10.1 Determining the Number of Workers Required per Task

Following the assignment of urgency to this job in the beginning of the decision loop is reached wherein the lowest number of workers possible is assigned. Recall that the number of workers was initialized to the minimum at the beginning of the logic. Given the number of workers assigned to the job, the process rate at which units will be processed given that particular number of workers can be assigned also. All of the necessary ingredients for calculating the bottleneck ratio are now assembled.

In the next module the bottleneck ratio is calculated. The numerator is calculated by multiplying the number of units remaining to be processed by the rate at which they will be processed given the number of workers assigned in the previous decision loop. Obviously, on the first run through the loop, since the number of workers was assigned to zero, the bottleneck ratio will be infinitely large.

Next, control passes to a decision node that checks to see whether or not the deadline has passed. This logic is necessary to deal with those situations in which the job is just past the deadline; therefore, the bottleneck ratio is negative. Since any negative number is less than or equal to one, were it not for the insertion of this logic, zero workers would be assigned to a job that was passed its deadline and the job would never be finished. So, the query is made as to whether or not the deadline is already passed. If the answer is yes, the maximum number of workers that can be assigned to this job are assigned to this job. Other possible assignments are possible here at the discretion of management. It would be sufficient to say that if the deadline had passed, the system would have to decide how to handle it. In the version shown above, the maximum number of workers is assigned in keeping with the second and third business objective to

diminish the number of jobs that are late and diminish the sum of the lateness across all jobs. Control then passes to the following decision node with a yes. That decision node as to whether or not the maximum number of workers has been assigned is the exit point that places an upper limit on the number of workers that can be assigned. If the maximum number of workers has not been reached, control is passed to the next decision node which inquiries whether or not the bottleneck ratio is less than or equal to one. If the bottleneck ratio is less than or equal to one, then the preceding logic has determined the minimum number of workers needed to accomplish the task by the deadline. This satisfies the operational needs of the system and dispatches just enough workers to a job to get it out on time barring unforeseen variation. If the bottleneck ratio calculates to a value greater than or equal to one, then the current number of workers is insufficient to accomplish the task by the deadline. Control is passed to a module that increments the number of workers by one and loops back to the N-way decision node for the number of workers and the process is repeated. Eventually, either the bottleneck ratio will be less than or equal to one or the maximum number of workers will be reached. Either way, control now passes to an assignment node in which the number of workers is assigned to this job.

4.10.2 Dispatching a Certain Number of Workers with a Certain Priority

The last part of the decision refers back to the initial processing in the logic which assigned an urgency of one or zero. It is assumed that there will be occasions in which more than one job will require workers. In such situations, priority should be given to those jobs that bear a financial risk to the company as opposed to those jobs for which lateness is merely an operational inconvenience in nature. This was accomplished in the

initial part of the model with the assignment of a zero or one for the job attribute of urgency. Consequently, the number of workers needed to accomplish this job by the deadline is then seized, given the priority according to the calculation of urgency.

4.10.3 Summary of the Logic of the Method

In summary, given a known amount of work, known job process rates given the number of workers, a deadline, and a limit on the number of workers that can be assigned to a job, this method can both determine the number of workers required to accomplish the work by the deadline and assigned them a priority if it comes down to the system selecting between jobs for scarce resources.

An even more nuanced version of the logic could be used in the event that the system had to choose between jobs that had both monetary penalties associated and the potential of not meeting the deadline. Rather than assign a simple priority of high or low, the actual monetary cost could be used in that variable in the system would then choose the job with a higher monetary cost for not meeting the deadline.

In the model a single job is considered. There are no dependent events in the drawing. However, the same logic would apply to a series of jobs wherein the decision had to be made in front of each job as to how many workers would be assigned to that job. If there are jobs in series, and the deadline is given for the accomplishment of all jobs, and the logic of the bottleneck ratio method is followed, that statistical fluctuations in the process times for the work would tend toward increasing the number of workers needed on subsequent tasks rather than decreasing that number. All of the logic in the method also assumes that only series processing is possible and that the addition of more

workers will accelerate one of the sub processes but never result in parallel processing that would effectively change the precedence relationships.

4.10.4 Change in Objective Function

Modification will begin by the adoption of a new objective function. Whereas in the as-is model, the emphasis was placed on the number of flights that push back after the estimated time of departure and the histogram of the lateness across all flights in the system, the new priority will be to avoid costs. Essentially, the as-is model was operations-focused and treated all flights as equal in importance. The to-be model will adopt a new objective function in line with Goldratt's original philosophy from "The Goal" wherein he states that the true goal of any company is…to make money. Choosing profit as the highest objective function for this business operation leads to a change in the prioritization of planes that bear a monetary risk to the company. As mentioned earlier, these are the flights that are destined for the hub airports where connecting passengers run the risk of missing their connections if departure is delayed beyond a certain value. The value used in this model will be 30 minutes. The average calculated cost of missing a connection on this airline has shown to be $380 per passenger misconnected with expenses totaling $12M per year.

The knee-jerk reaction would be to shoot beyond the mark here and assign the maximum number of workers to any flight that had a monetary risk associated with it. While this would minimize the costs associated with passengers misconnecting at hubs due to ramp operations, it would do little for customer loyalty due to the number of non-hub-bound flights that would necessarily be neglected and depart late inconveniencing passengers by simple late departures and arrivals. Better to assign a minimum yet

sufficient number of workers to a task to be able to finish it on time to allow other resources to work on other important tasks. For purposes of both the as-is and to-be models, the maximum number of workers that can possibly be assigned to work a flight will be six. Due to the limited space in which activities must take place, such as in the bins of the aircraft, more than six workers would merely get in each other's way and not decrease the processing time of the units of work.

It is soon obvious that there are multiple conflicting priorities to be addressed in the construction of the to-be model. This principle is not new and can be found in current Management Science textbooks such as (Taylor III 2007). The principle is a linear programming method called “goal programming” and recognizes that, at time, multiple priorities, or goals are in conflict with each other. The hierarchy of priority will dictate what should be done first, second, and last. Frequently, satisfying a high priority will mean that a lower priority will be sub-optimized. With this in mind, the second priority of the to-be model will be that of minimizing the number of flights that depart late, and the third priority will be the minimization of the sum of the lateness across all flights.

4.10.5 Signaling Availability of Ramp Workers

The system contemplated in the to-be model requires a simple telecommunication device for each of the ramp workers. This device will be referred to as “the red button”: it is an input/output device that can send a signal to a receiver and receive some sort of text message similar to the pagers of the 1990’s. It sends a signal to the receiver system and its implied message is that the worker that pressed the red button has completed the assigned task and is now available for reassignment to another work area. The receiver system is connected to a computer running a program with a version of the heuristic

described below at its core to dispatch the workers to the most important task at that moment. The receiver system has a global view of all of the flights and the status of each with respect to how much work remains to be done on each flight and how much time remains to be able to accomplish that work. The receiver must include a type of decision-making system that could then determine the activity with the highest priority and return signal to the worker indicating where the workers should go for his next assignment. In this way, the system is constantly paying attention to the highest priority activities.

4.10.6 Prioritizing Classes of Activities

In the to-be model, three different priorities will be used. The highest priority, a priority of one, will be assigned to the task of marshalling aircraft in to the gate. A priority of two will be assigned to those activities involving flights that are destined to hub airports. A priority of three will be assigned to those activities for flights that are not destined to hub airports. The priority of three will also be used for those flights which are destined to hub airports but where, due to their late arrival, it would be impossible to turn the flight in the time remaining before estimated time of departure. As nothing can be done to salvage the on-time departure of such a flight, it will be lumped in with those flights which are not bound for hub airports and assigned a priority of three.

4.10.7 The Bench

When a worker pushes the red button indicating that he is available for reassignment, he will be considered to be "on the bench". This is a virtual holding area for ramp workers who are idle and does not represent an actual bench. Therefore, following the marshaling of an aircraft to the gate, the three ramp workers involved in the

process will be released to the bench, meaning that they are available when needed. As each one of them has a red button, they would each press that button indicating that the marshaling had been completed and would promptly received a text message indicating where the next assignment was. This idea of releasing workers to the bench following the completion of every task allows the system to always have access to its workers.

In the modeling software used to create and run the simulation, resources such as ramp workers can be seized with different priorities. This is a key part of the model. As nine different gates are going simultaneously, and each gate has four main activities as mentioned earlier, these tasks will be starting and concluding frequently. At the conclusion of each, as the worker is released to the bench, the system is constantly adjusting so that the most important activities are done first.

If the number of bags remaining in the rear bin is known and the number of workers assigned to process those bags is also known as well as the rate at which that number of workers can process bags, then a simple calculation of the number of bags times 1/(work rate in minutes per bag given the number of ramp workers) yields the amount of time necessary to process the remaining bags. For purposes of the model, when a bottleneck ratio is calculated, the assumption is made that the current number of ramp workers assigned to the task will continue throughout any subsequent tasks on that flight. For example if the system determines that only three ramp workers are needed to download the bags from the rear bin on a flight, it has done so having calculated that three workers should be sufficient for all remaining tasks. Stochastic variation in process times will often lead to situations where original calculations are no longer valid. At such

points the number of ramp workers is incremented so that the flight is not delayed. The denominator will consist of the amount of time remaining until pushback.

4.11 Determining the Number of Ramp Workers Assigned to a Task

Prior to each of the four activities that comprise ramp operations, a sub-model is run to determine the number of workers required for that activity to be able to get the flight out on time as shown in Figure 13. This is where the theory of constraints comes into play. The logic of the sub-model will now be explained. An attribute corresponding to the number of ramp workers assigned to that activity is initialized to zero.

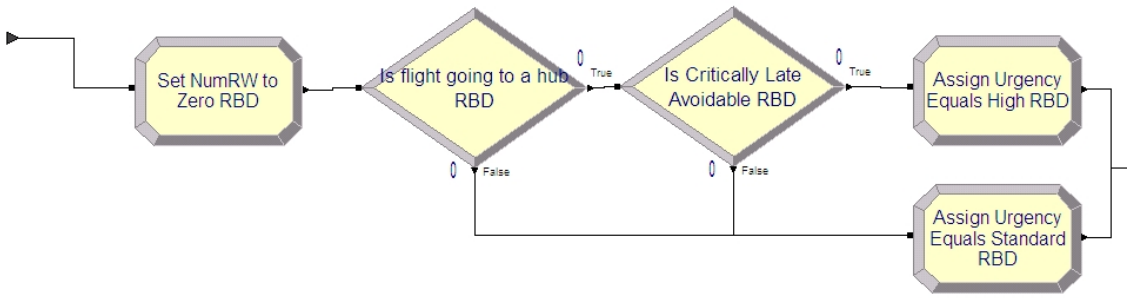

Figure 13 - Part 1 of the Logic for Selecting the Number of Ramp Workers

The following module, a decision node, checks to see whether or not the flight is destined for a hub airport. If false, the flight is assigned an urgency of zero. Urgency is an attribute that can take on the values of one or zero. If the decision node returns true, then the entity passes to yet another decision node. In this decision node, the system checks to see whether or not it is possible to get the flight out before it is critically late. If the answer to this decision node is false then the flight is once again assigned an urgency of zero in keeping with the prior explanation that even though a flight is destined for a hub, if it is impossible to get it out on time, it will be treated like a normal flight to a non-

hub destination. If this decision node returns true, then urgency is assigned a one. In summary, this piece of logic only allows flights that are destined for a hub and for which it is possible to accomplish the work prior to the flight being classified as critically late to be assigned the higher priority of one for urgency. This attribute to "urgency" will come into play later in the sub-model when it is time to seize ramp workers to accomplish the activity contemplated in the sub model. Activities on flights with an urgency of one will have a higher priority in getting the next available ramp workers than a flight that has an urgency of zero.

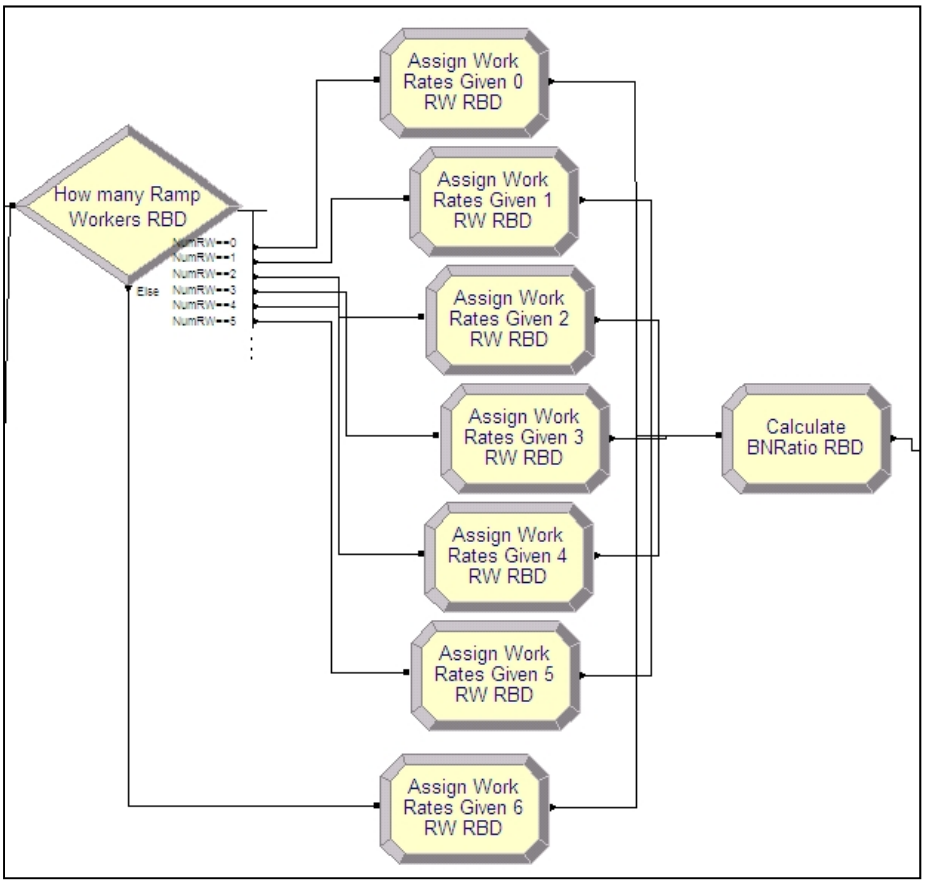

Figure 14 – Part 2 of the Logic for Selecting the Number of Ramp Workers

Once urgency has been established a decision node checks the assigned value in the attribute for number of ramp workers. As this number has been initialized to zero,

control passes to an assignment module in which a work rate is assigned based upon the number of ramp workers as can be seen in Figure 14. Control then passes to a node that calculates the bottleneck ratio given the work rate given the number of ramp workers. Since work rates get faster in bags per minute as the number of employees increases starting at zero implies an impossibly long delay in accomplishing this task and all subsequent tasks if there are zero workers.

In Figure 15, the logic required to determine the number of ramp workers for a given task is concluded. The first decision node is whether or not the flight is already late. If the ETD has passed, the bottleneck ratio will be negative which, of course is less than one. A blanket decision is made that if any flight is late already, that flight will have six workers assigned to it.

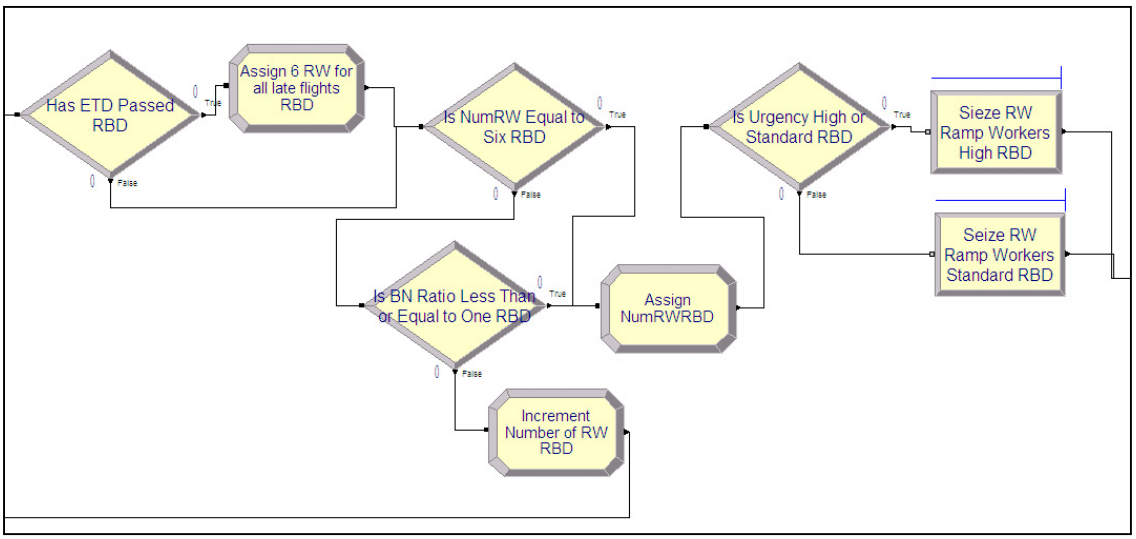

Figure 15 - Part 3 of the Logic for Selecting the Number of Ramp Workers

The following decision that would asks whether or not the number of ramp workers is six. If true, the number of ramp workers assigned to the task is finalized as well as the rate at which all subsequent bags will be processed given this number of

workers. Control then passes to a decision node that recalls the urgency previously assigned. If the urgency is one in the ramp workers are seized with a higher priority than if the urgency is zero. In this manner, priority is given to those flights that are destined for a hub which are still doable.

If the estimated time of departure has not passed and the number of ramp workers is not six, control is passed to a decision node which compares the calculated bottleneck ratio to one.

Table 6 - Decision Tree for Bottleneck Ratios

State	Decision
Bottleneck Ratio < 1	The time required to complete the job is less than the time allotted for its completion suggesting that the current process rate is sufficient.
Bottleneck Ratio = 1	The process rate is marginally capable; however, any disruption in the process or variability in process rate could lead to a situation of the bottleneck ratio > 1. See below.
Bottleneck Ratio > 1	The time required to complete the job exceeds the time allotted for its completion suggesting that the current process rate is too low. The process rate must be increased by incrementing the number of resources assigned to the job.

As can be seen in Table 6, if the bottleneck ratio is less than 1, the number of ramp workers is incremented by one and the entity is sent back to the system to see if that number of ramp workers will be sufficient to accomplish the tasks in the allotted time. In this manner, control loops back to the system until either the bottleneck ratio is less than

one suggesting a sufficient number of workers to accomplish the task in the remaining time, or the number of ramp workers reaches the maximum allowed at which point control would then exit the module and assign the maximum number of ramp workers to the task.

After determining the number of ramp workers sufficient to complete the impending task and all tasks that follow it, a setup function is performed for that task. The following module then processes the bags for that task at the rate prescribed in the model just discussed. Once all the bags for this task have been processed, all ramp workers assigned to the task will press their respective red buttons and return to the bench. This process will be reiterated four times, once for each of the four principal activities for handling baggage.

Using the bottleneck ratio and the definition of a constraint as it applies to a service such as this, the constraints to on-time performance are mathematically identified and prioritized. Exploiting the constraint suggests either attention by management or mandate that keeps the workers busy and focused on the accomplishment of the task until it is finished. Step three of the Theory of Constraints, "subordinate all else to the bottleneck", is accomplished by the system automatically as activities that are constraints receive higher priority for seizing ramp workers. If those ramp workers are insufficient to the task, additional ramp workers are assigned in keeping with step four, elevate the constraint. The iterative nature of the model automatically complies with step five, if the constraint is broken, go back to step one. Thus, with this logic and model, a system that is true to the theory of constraints as originally proposed by Goldratt has been developed and will now be tested to see if it produces the expected results of first decreasing the

costs to the airline associated with misconnecting passengers and improving other performance measures such as on-time performance.

5 ANALYSIS AND DISCUSSION

After the completing the as-is model and validating it, the model was adjusted to reflect a heuristic based on the principles of the Theory of Constraints. Both the as-is model and the to-be model were run replicating a year's worth of flights or 365 day's worth of flights times 36 flights per day equals 13,140 flights. All of the attributes of each flight as well as performance measures were collected in a spreadsheet for analysis using SPSS software to determine if the objectives of the model were met.

5.1 Performance Measures and Terms

Traditional performance measures will be used to determine the effectiveness of the method. Actual turn time is the difference between actual time of departure and actual time of arrival. Scheduled turn time or estimated turn time is the difference between estimated time of departure and estimated time of arrival.

A new term, optimistic turn time, is defined as estimated time of departure minus actual time of arrival. This measure addresses the idea that no matter what time the plane actually arrives, there is hope to be able to turn it around by the estimated time of departure. The difference between the actual time of arrival and this estimated time of departure could be thought of as what is hoped to be accomplished.

5.2 Number of Ramp Workers Assigned to Each Task

The maximum number of ramp workers in both models is fixed. The assigned capacity for both the as-is model and to-be model is 26 ramp workers. The utilization of the ramp workers is higher in the to-be model than it is in the as-is model; that is, they do more work. These workers were not assigned by schedule; consequently, the

measurements are with respect to a fixed capacity capable of handling the maximum need.

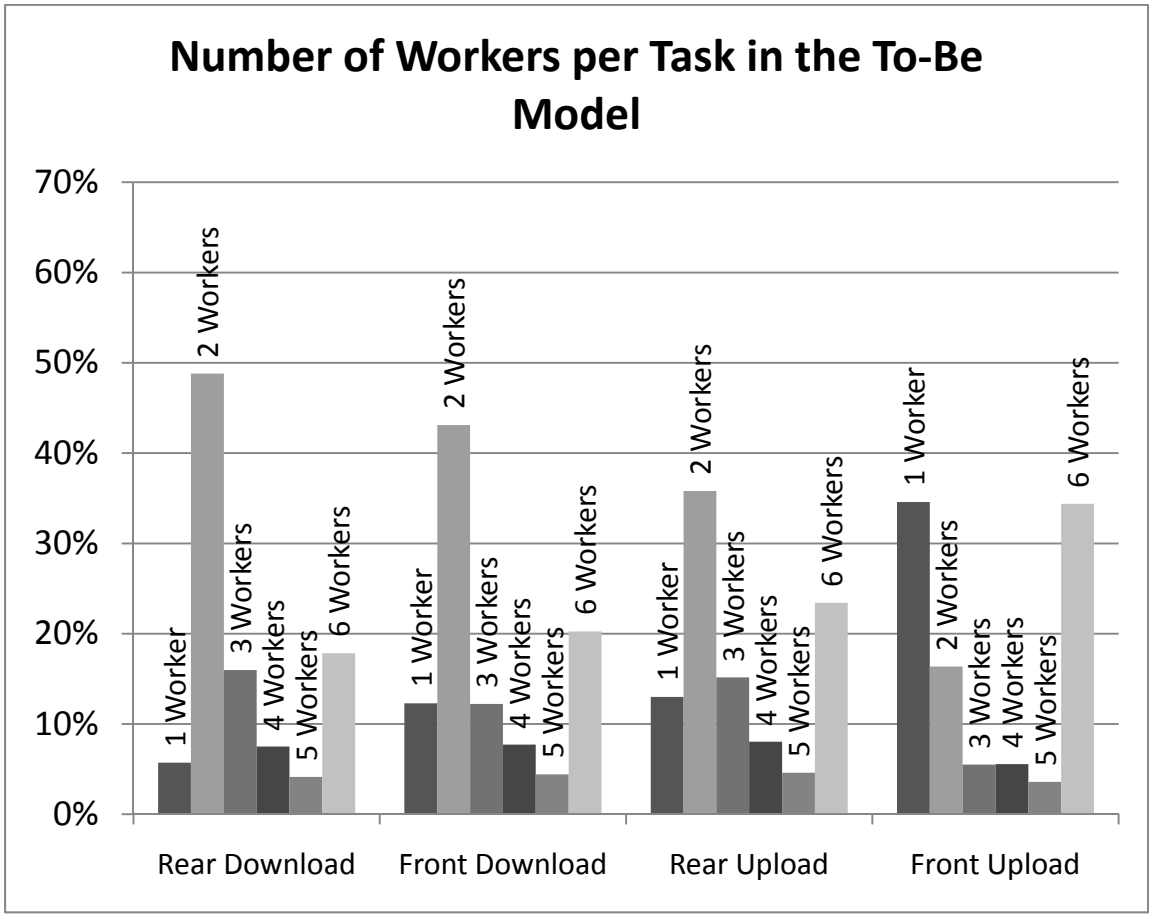

Figure 16 - Number of Workers per Task

In the as-is model, a constant number of ramp workers was assigned for all tasks. Each of the four main tasks had three workers assigned to it. In the to-be model, the number of ramp workers varies per task. In Figure 16, the number of ramp workers assigned to each task in the to-be is shown in a histogram. It is immediately obvious that less than three workers on the initial tasks of downloading the rear bin, downloading the front bin, and uploading the rear bin is common. This reflects the logic of the model that always attempts to accomplish each task with the minimum number of workers

necessary. The fact that two workers are frequently sufficient to accomplish the initial tasks suggests that the use of three workers in the as-is model is an application of too many resources.

Also of note is the fact that the use of six workers increases across the four tasks in series; increasing from 18% of all initial tasks up to 34% of all final tasks. As statistical fluctuations in process times eat away at the time remaining to accomplish the job, more and more jobs will require the maximum number of workers to be able to finish by the due date.

5.3 Paired Samples Statistics of Each Model

In the paired sample statistics that follow, a comparison will be made in measures of central tendency and variation from using the heuristic developed in this model for managing the process. Each of the pairs will now be briefly discussed to ascertain whether the model accomplishes what it sets out to accomplish and whether the outputs are as expected.

5.3.1 Comparing the Time Required to Process Bins

Insights gained from Figure 16 helped to make clear the differences seen in Figure 17 for the process times for each task. The frequent use of only one or two workers for a given task makes the process times in the to-be model longer than their counterparts in the as-is model.

In Figure 17, the average process times for each of the four main tasks are compared between the as-is model and to-be model. The average difference of 4.7 minutes between the time to download the rear bin in the as-is model and the

corresponding time to download the rear bin in the to-be model was tested for statistical significance using a two-sample t-test

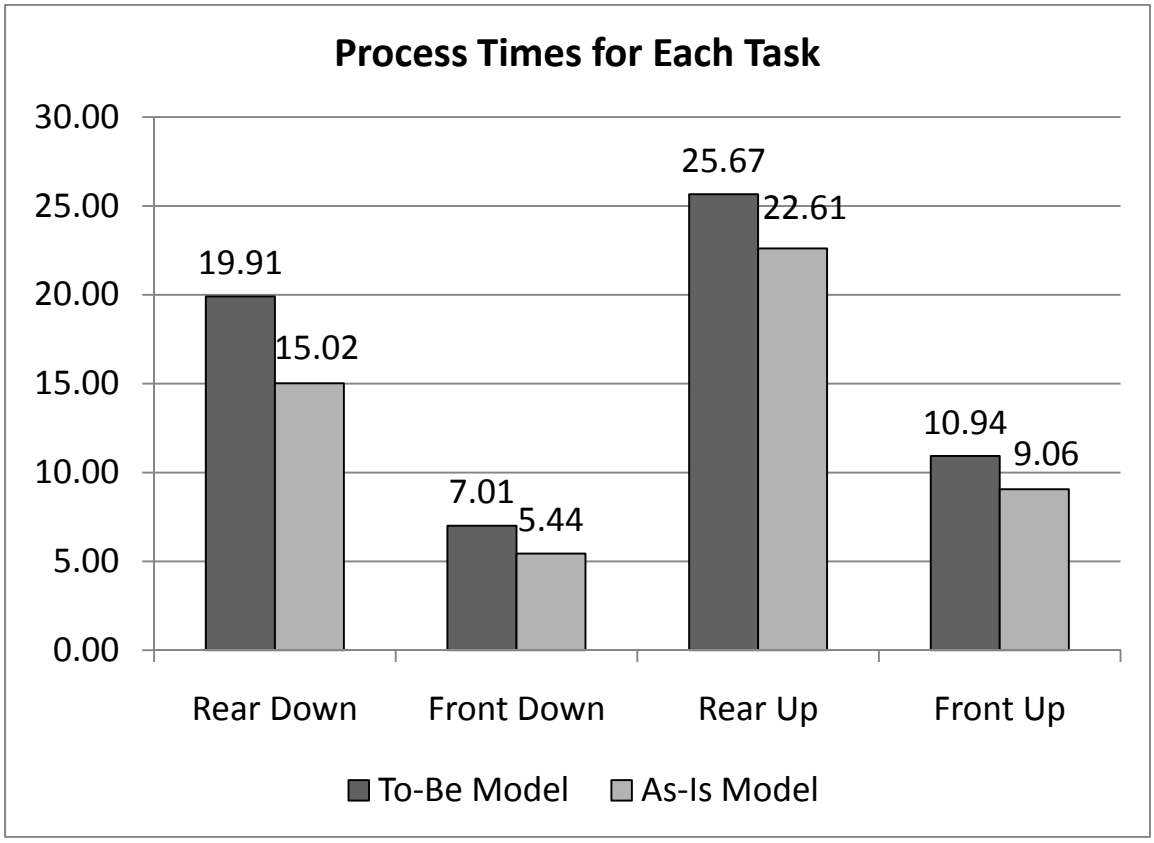

Figure 17 - Difference in Task Processing Times

Formally stated:

Hypothesis test results:
μ_1 : mean of As-Is Model for Downloading the Rear Bin
μ_2 : mean of To-Be Model for Downloading the Rear Bin
$\mu_1 - \mu_2$: mean difference
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$
(with pooled variances)

The test shows that there is sufficient evidence to reject the null hypothesis and conclude that the time to unload the front bin is significantly different between the as-is and to-be systems (p-value < 0.001). This suggests that in the as-is model, due to the

constant application of three workers, the process is shorter. This result is not unexpected as the downloading of the rear bin is the first of the four activities and consequently the variability in processing has not consumed the time buffers associated with all remaining operations. Additionally, in 55% of all flights only one or two workers are assigned to the initial task of downloading the rear bin.

5.3.2 Comparing the Time Required to Download the Front Bin

Referring again to Figure 17, the average difference of 1.6 minutes between the time to download the front bin in the as-is model and the corresponding time to download the front bin in the to-be model was tested for statistical significance using a two-sample t-test. Formally stated:

Hypothesis test results:
μ_1 : mean of As-Is Model for Downloading the Front Bin
μ_2 : mean of To-Be Model for Downloading the Front Bin
$\mu_1 - \mu_2$: mean difference
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$
(with pooled variances)

The test shows that there is sufficient evidence to reject the null hypothesis and conclude that the time to unload the front bin is significantly different between the as-is and to-be systems (p-value = 0.000). This suggests that in the as-is model, due to the constant application of three workers, with its commensurate faster process times, processing time is shorter. This result is not unexpected as the downloading of the front bin is the second of the four activities and consequently the variability in processing has not consumed the time buffers associated with all remaining operations. Again, in 55% of all flights either one or two workers is assigned to the task of downloading the front bin.

This result is in keeping with the arguments of the previous section that in the early stages of the turnaround process, while none of the time buffers inherent in the time remaining to accomplish the remaining jobs have been consumed, less than three workers are assigned resulting in slower processing times when compared with the as-is model where the number of workers is constant.

5.3.3 Comparing the Time Required to Upload the Rear Bin

An experiment to determine the significance of the difference between the as-is and to-be models with respect to the time required to upload the rear bin yields a p-value of zero. The null hypothesis that there is no difference in the means is rejected due to the low p-value suggesting that the difference here is also statistically significant. Formally stated:

Hypothesis test results:
μ_1 : mean of As-Is model for uploading the front bin
μ_2 : mean of To-Be model for downloading the front bin
$\mu_1 - \mu_2$: mean difference
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$
(with pooled variances)

The trend continues. When considering all flights, the method will move workers away from flights that have no monetary risk associated with them and apply them where there is a monetary risk. Additionally, the bottleneck ratio method only assigns the number of workers necessary to accomplish the task by the due date which, in many cases, means the individual process times may be longer, however, the work can and will be accomplished by the minimum number of workers. Thus, it is possible that average flights will have operations that are longer in the to-be model then in the as-is model.

5.3.4 Comparing the Time Required to Upload the Front Bin

Refer again to Figures 16 and 17. The final of the four tasks, uploading the front bin, displays certain characteristics that are worth mentioning. The average difference between the times in the two-be model and the as-is model is 1.8 minutes (p-value = 0). Thus, we have evidence that the difference is significant statistically. A review of the input data reveals that this is one of the smallest tasks of the four. So much so, that in 35% of all flights only one worker is assigned to this task.

Thus, in all cases, the amount of time required to process the job on the to-be model was greater than the same amount of time required for the as-is model. The as-is model, with its emphasis on the traditional measure of diminishing actual turn time would appear to be superior by these measures. However, since these measures do not comprise the objective function of the bottleneck ratio method, we must withhold judgment until the complete picture is painted.

5.3.5 Actual Turn Times

A comparison of sample statistics between the as-is model and to-be model follows. The first comparison is between the time required for the actual turn time in the as is model versus the actual turn time in the to-be model. In the as is model, the actual turn time was 79.63 minutes with a standard deviation of 30.1 minutes. The actual turn time for that to-be model was 80.4 minutes with the standard deviation of 31.9. A two-sample t-test testing the null hypothesis that the means are the same yields a p value of .044. At the 5% level of significance this is sufficient to reject the null hypothesis in favor of the alternative hypothesis that the means are not the same.

This result is not unexpected and suggests that the bottleneck ratio method places higher priority on flights that have monetary risk for the company rather than traditional measures such as actual turn time. The difference nevertheless was slight; less than one minute.

5.3.6 Optimistic Turn Times

Due to the nature of this measurement which was previously defined as the estimated time of departure minus the actual time of arrival (ETD-ATA) and the fact that the input data for both the as-is model and the to-be model was the same, this measurement is identical in both models and cannot be used for comparison.

5.3.7 Comparing the Percentage of Flights that are Critically Late

Recall that the definition of "critically late" is that the flight pushes back from the gate 30 minutes or more after the estimated or scheduled time of departure (ETD). In some instances, due to the lateness of the arrival of the inbound flight, it is impossible to avoid the outbound flight being critically late. Nevertheless, since the inputs into both models were the same, this affects both the as-is model and the to-be model comparison can be made between both models with respect to whether or not a flight is critically late.

Eighteen percent of all flights in the as-is model left the gate having been characterized as being critically late. That is, they left more than 30 minutes beyond their scheduled departure times. The standard deviation with respect to this proportion of flights categorized as critically late with .39. The same measure in the to-be model yielded an average proportion of 12% with a standard deviation of .33. A two-sample

test to determine whether the two proportions are equal based on these statistics yields a p value of zero based upon as z statistic of a -13.6.

Thus we have significant statistical evidence to suggest that the bottleneck ratio method decreases the number of flights that can be classified as being critically late across all flights whether or not they bear monetary risk to the company or not. This effect is in spite of the fact that individual jobs can take longer as seen in the previous four sections of analysis.

5.3.8 Comparing Costs per Flight across All Flights

The average expense associated with hub bound flights that depart critically late in the as-is model is $177.65 with a standard deviation of $956. Recall that this value stems from the idea that hub-bound flights carry connecting passengers who, if they are critically late, will have to be accommodated by the airline responsible for the late arrival. In the two-be model the average cost associated with hub bound flights that depart critically late is $122.88 with a standard deviation of $800.6.

A two sample t-test based on the hypothesis that the two means are the same yields a p value of essentially zero based upon a t statistic of 5.03. The difference of $55 per flight is therefore statistically significant and represents successful implementation of the method in reducing costs to the airline associated with critically late departures.

It can be seen that the average turn time in the to-be model is two minutes longer. This does not suggest that the to-be model is less efficient, but rather that the prioritization of flights destined for hubs is causing flights not destined for hubs to receive less attention than they might in the as is model. Plus, logically, it would take

longer to process the average flight because ramp workers were being assigned to destination hubs.

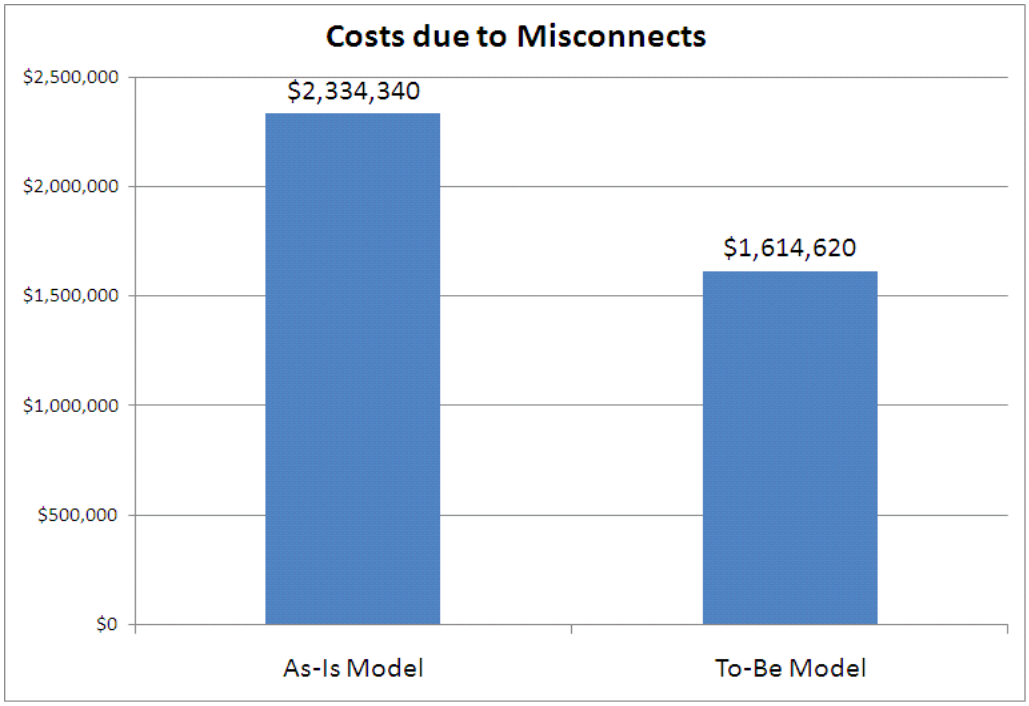

Figure 18 - Costs per Flight Controlling for Hub Destination Flights Only

5.4 Comparing Costs per Flight Controlling for Hub Destination Flights Only

Filtering the data by whether or not the flight is bound for a hub destination demonstrates the effectiveness of the bottleneck ratio method in accomplishing its purpose of reducing costs for the company. In Figure 18, we see that in the traditional model, with its level production strategy, costs associated with misconnecting passengers amount to slightly more than $2.3 million. Application of the bottleneck ratio method reduces those expenses to $1.6 million. This represents a 30.8% improvement.

5.5 Optimistic Turn Times

The optimistic turn time is defined in the model as the difference between the actual time of arrival and the expected time of departure. In other words this is the time that management hopes to be able to turn the flight around. As both the actual time of arrival and the expected time of departure were known with certainty at the beginning of the run for both the to-be and the as-is model, the means are identical.

5.6 Critically Late Flights

There is a 6% difference in the proportion of flights that are critically late in favor of the to-be model. This difference is due to the fact that more attention is paid to flights as they get increasingly closer to their estimated time of departure. These critically late flights are destined for both hubs and non-hubs. Therefore, this measure alone does not suggest an achievement of the first priority of making more money. It does comply with the second priority which is to decrease the number of flights that are late.

5.7 Cost this Flight

The key output of the cost associated with misconnecting passengers is significantly decreased in the to-be model. The average cost per flight for the to-be model was $122.8 and the average cost per flight in the as-is model was $177.65. This represents a decrease in approximately $54.00 per flight across all flights whether destined for hubs or not. The projected costs associated with critically late flights destined for hubs in the as-is model was $2,334,340. The projected costs associated with critically late flights under the to-be model was $1,614,620. The difference between the two models with respect to costs is approximately $719,000 in favor of the to-be model.

5.8 Man Minutes and Worker Compensation

As mentioned previously, the number of ramp workers in both models as identical and fixed at 26. However, a calculation of the number of man-minutes applied to each flight reveals a difference of approximately 22.5 man-minutes per flight. The 95% confidence interval for the mean difference in man-minutes is 21.87 and 23.12. If ramp workers earn $10 an hour of this represents an increase labor expense of an average $3.67 per flight for the to-be model. This results in an annual cost of $48,224.

Table 7 – Confidence intervals for the mean savings given worker compensation

Amount Earned by workers per Hour	Savings given a 95% confidence interval of between 21.87 to 23.12 man-minutes.
$8	Between $2.92 and $3.08
$9	Between $3.28 and $3.47
$10	Between $3.65 and $3.85

In a year, this gross savings equates to between $591K and $846K. Subtracting average costs of the increased labor yields a net savings of from $542,287 to $797,992 in the case study.

5.8.1 Net Reduction in the Number of Critically Late Flights across all Flights

In the as-is model, 18.5% of all flights for the year left the airport thirty minutes or more late. In the to-be model, only 12.5% of all flights left the airport thirty minutes late or more. The net reduction in the number of flights that left critically late was 6%. This operational measure was the second of the two priorities that were addressed by the bottleneck ratio method.

6 CONCLUSIONS

6.1 Summary

A classical problem in a large service factory environment is being able to focus on those tasks that are of greatest importance to the overall objectives of the company. If there are choices to be made in the assignment of scarce resources and there is no system in place that takes a global view, then all possible solutions must be arrived at as some form of local optimization. If the method for assigning resources relies on the real-time perceptions of roving supervisory personnel, then the resulting configuration of asset deployment would logically be sub-optimal based on the limited vision of decision makers.

Management by constraints has been shown to be effective in increasing throughput and decreasing work in process in manufacturing environments. The identification of constraints is fairly straightforward and consists in identifying that resource whose capacity is less than or equal to the demand placed upon it. Capacities are usually fixed. To adapt the principles of management by constraints to service operations requires a precise definition with respect to capacity constraints. The definition of a constraint in a pure service environment was given as "any process for which the time remaining to accomplish the amount of work assigned is insufficient given the processing rate." If the assumption is made that the processing rate of a worker is relatively fixed and that increases in the processing rate may only be accomplished by the addition of workers, then a system must be identified that can dispatch such workers taking into account all of the jobs in the system. Given the above definition of a system constraint, a system that relies on the continued, deliberate efforts of human workers is a

system in which any and all processes could be or become constraints. All that would be required for a process to become a constraint would be a slowing down of the processing rate or the rate at which the process is advanced to the point that the time remaining for the accomplishment of the process is insufficient even at the fastest possible work rate. Thus, in a service factory in which multiple processes are being accomplished simultaneously, there is a possibility for process is to fall behind their schedules requiring intervention if they are to be put back on schedule. If intervention is required, then such intervention should be done using a system with a global view. The method and heuristic put forth in this dissertation is just such a system.

6.2 Contributions and Significance

TOC can be applied to services. The concept of a bottleneck was adapted to service operations in such a way that an improvement method could be proposed for managing them. A method for applying that adaptation of the Theory of Constraints to the environment of a service factory was developed. The use of the heuristic based on the bottleneck ratio to determine which activities should be done next was shown to produce significant results and performance measures when compared against a system focused on local optimums.

6.3 Requirements for the Method

For the logic of the method to work the quantity of work per activity must be known beforehand. This could either be in the form of piecework wherein the known process rates per piece given the number of workers was known or, in the event of a different kind of work other than piecework the process rate for the job given the number

of workers was known. Either way, to be able to calculate the bottleneck ratio, the amount of work remaining must be calculable.

Processing rates must be calculated beforehand. Logically, the processing rate given n+1 workers should be faster than the process rate given n workers. In the logic as it is the improvements need not be linear. In some applications a maximum would probably be reached beyond which the process would see a diminution in the process rate as having too many workers which would slow down the process.

6.4 Limitations

All of the logic in the method also assumes that only series processing is possible and that the addition of more workers will accelerate one of the sub processes but never result in parallel processing. In reality though, the addition of more workers frequently results in the option of doing more than one thing at once and this ability to add workers who focus on the next sub-process rather than the one at hand is not contemplated in the model.

One weakness of such a system would be evident if employees were work-averse or lazy. Following the completion of one job, rather than press the button and send the signal that they are available for the next job they could simply delay thus avoiding their next assignment. Adoption of a system similar to the one proposed in this dissertation would also require an adjustment in the incentive program and compensation schedule of the workers. For example, if workers are paid by the hour, there would be no incentive to cooperate with the system by pushing the button and notifying the system of one's

availability. Such an action would result in more work with no increase in compensation. Such discussion is beyond the scope of this dissertation.

Implementation of the bottleneck ratio method would have a cultural impact on workers. A worker who is working for the satisfaction of seeing a job through to its completion and a sense of accomplishment that comes from meeting a deadline would have to adjust to a new system whereby he participated in many varied sub processes all for the greater good and profitability of the company but none of which were seen through to completion. New means of identifying excellence in the workforce would have to be devised.

Frequently in service operations, the current culture of work assigns certain people to certain jobs on which they usually work from start to finish. They get a sense of satisfaction and completion when the the job is completed. This is felt as a job well done. Moving to an environment where individual workers may work several parts of various jobs and never see any particular job through to completion due to being constantly moved around to satisfy the needs of the system could have a psychological effect on their work. Another topic of study is the effect of gaming the system if workers are only interested in avoidance of exertion to earn hourly wages. Which worker will press the button first when the task is completed? If a task is completed by four workers and two are then reassigned somewhere else, who goes and who stays?

6.5 The Theory of Constraints is Applicable to Services

The results from the simulation indicate that the theory of constraints can be applied to services with certain adaptations to the terms and implementation of the

philosophy. The non-intuitive results from the case study in which an airline was demonstrated to save money while apparently worsening certain performance measures is one of the hallmarks in moving to the TOC world. Local optimization is supplanted by global optimization and reducing performance measures on individual flights resulted in improvement across all flights.

6.6 Service Improvement Method

Use of the bottleneck ratio to make decisions with respect to worker dispatch allows the company to prioritize important jobs. It simultaneously prioritizes for profitability, the true objective function of any business, and is a simple means for prioritizing tasks as they approach their deadlines.

An Airline that is willing to replace a system in which workers always try to get their own flight out on time (local optimization) with a system that allows for a global system to dispatch workers stands to benefit from the reduced costs associated with passengers who fail to make connections in Airline destination hubs. The airline in this simulation was shown to experience a decrease in costs of 30% of what is an onerous expense.. The technology to implement such a system is neither expensive nor technologically advanced.

Application of the bottleneck ratio in this service environment was made possible by knowing in advance the amount of work required for each task in the form of number and location of inbound bags and location of outbound bags. Work rates were also sampled and known a priori and therefore the status of the system could be ascertained "at the touch of a button". An important consideration in adoption of the system similar

to this one would be the frequency with which workers checked in by pressing the buttons. If the individual tasks took too long then the system would not adapt quickly enough to serve the purposes of the system therefore the relative frequency with which the system updates would move along a spectrum from local optimization toward global optimization. Modern technology in the form of RFIDs, laser scanners, and bar codes, could yield an even more sensitive system.

6.7 Future Research

Adding more detailed nuances to the model would yield deeper insights into the true behavior of service operations. The inclusion of more and more constraints such as worker scheduling, resource scheduling, and interactions with other agents, customers, and agencies that are part of the delivery of the service would yield insights into their effect on the throughput of the system and where improvement efforts should be focused. This is in keeping with step five of the theory of constraints wherein, if a bottleneck is broken, management must return to step one and identified the new bottleneck. In an environment that relies heavily upon human cooperation on the part of both airline employees and customers bottlenecks are constantly shifting. One aspect of the model in this study that was not applied would be that of preemption are taking workers off a task while the task is not yet finished. This layer of complexity, if added to the model would be interesting to study. Additionally, no penalties were made in this model for distance between work areas and all workers were able to move freely and virtually instantaneously to other tasks.

Current culture of working at the airport assigns certain people to certain flights which they usually work from start to finish. They get a sense of satisfaction and

completion when the flight pushes back. This is felt as a job well done. Moving to an environment where individual workers may work several flights and never pushed back any due to being constantly moved around to satisfy the needs of the system could have a psychological effect on their work. Another topic of study is the effect of gaming the system if ramp workers are only interested in avoidance of exertion. Which ramp worker will press the button first when the task is completed? If a task is completed by four workers and two are then reassigned somewhere else, who goes and who stays? Such problems, old yet new, would have to be addressed under the new system. Such evolution of problems motivated Einstein to say, "No problem can be solved from the same level of consciousness that created it."

REFERENCES

Adeleye, Sanya, and Christopher Chung. "A Simulation Based Approach for Contingency Planning for Aircraft Turnaround Operation System Activities in Airline Hubs." *Journal of Air Transportation*, 2006: Vol. 11, No. 1.

Anderson, Mike, interview by S. Christopher Ellis. *Director of Safety* (May 24, 2010).

Basargan, Massoud. "A linear programming approach for aircraft boarding strategy." *European Journal of Operationa Research*, 2007: Vol. 183 394-411.

Boaz, R., and M. Starr. "Synchronized Manufacturing as in OPT: from Practice to Theory." *Computers and Industrial Engineering*, 1990: 585-600.

Bolander, S. F., and S. G. Taylor. "Scheduling Techniques: a comparison of logic." *Production and Inventory Management Journal*, 2000: Vol 41, No. 1, 1-5.

Boyd, Mahesh C. Gupta and Lynn H. "Theory of constraints: a theory for operations management." *International Journal of Operations and Production Management Vol. 28 No. 10*, 2008: 991-1012.

Bramorski, Tom, Manu S. Madan, and Jaideep Motwani. "Application of the Theory of Constraints in Banks." *The Banker's Magazine*, January/February 1997: 53-59.

Bureau of Economic Analysis - US Dept. of Commerce. April 28, 2011. http://www.bea.gov/industry/gdpbyind_data.htm (accessed April 28, 2011).

Cook, D. P. "A simulation comparison of traditional JIT and TOC manufacturing systems in a flow-shop with bottlenecks." *Production and Inventory Management Journal*, 1994: Vol. 35, No. 1 73-78.

Dan Trietsch, Isom. "Why a Critical Path by Any Other name Would Smell less Sweet? Towards a Holistic Approach to Pert/Cpm." *Project Management Institute*, 2005: Vol. 36. No. 1, 27-36.

Dave Nave. "How to Compare Six Sigma, Lean and the Theory of Constraints." *Quality Progress*, March 2002: 73-78.

Eisenhardt, Kathleen M. "Building Theories from Case Study Research." *Academy of Management Review*, Oct. 1989 Vol. 14, No. 4: 532-550.

Ferrari, Pieric, and Kai Nagel. "Robustness of Efficient Passenger Boarding Strategies for Airplanes." *Transportation Research Record: Journal of the Transportation Research Board*, 2005: Vol. 1915, No. 1 44-54.

Gillespie, Monty W., Mike C. Patterson, and Bob Harmel. "TOC Beyond Manufacturing." *Industrial Management*, 1999: Nov/Dec 41, 6 22-25.

Goldratt, E. *Critcal Chain.* Great Barrington: North River Press, 1997.

Goldratt, Eliyahu, and J. Cox. *The Goal.* Great Barrington: North River Press, 2004.

Gupta, Mahesh C., and Lynn H. Boyd. "Theory of constraints: a theory for operations management." *International Journal of Operations and Production Management*, 2008: 991-1012.

Gupta, Mahesh, and Joseph Kline. "managing a community mentla health agency: A Theory of Constraints based framework." *Total Quality Management*, 2008: Vol 19, No. 3 281-294.

Ike Ehie, Chwen Sheu. "Integrating six sigma and theory of constraints for continuous improvement: a case study." *Journal of Manufacturing Technology Management*, 2005: Vol. 16 No. 5, 542-553.

Kumar, Sameer, Kevin L. Johnson, and Steven T. Lai. "Performance improvement possiblities within the US airline industry." *International Journal of Productivity and Performance Management*, 2009: Vol. 58 No. 7 pp. 694-717.

Lockamy III, A., and M. S. Spencer. "Performance measurement in a theory of constraints environment." *International Journal of Production Research*, 1998: Vol. 36, No 8 pgs. 2045-2060.

Lubitsh, Guy, Christine Doyle, and John Valentine. "The impact of theory of constraints (TOC) in an NHS trust." *The Journal of Management Development*, 2005: Vol. 24, No. 2 116-131.

Mabin, Victoria J., and Steven J. Balderstone. *The World of the Theory of Constraints.* Boca Raton, FL: CRC Press, 2000.

Mabin, Victoria, and S. Forgeson. "Harnessing resistance: using the theory of constratints to assist change management." *Training*, 2003: 168-191.

Miller, B. "Applying TOC in the real world." *IIE Solutions*, 2000: Vol. 32, No. 5 49-53.

Moss, Hollye K. "Improving Service Quality with the Theory of Constraints." *Journal of Academn of Business and Economics*, 2007.

Motwani, Jaideep, Donald Klein, and Raanan Harowitz. "The theory of constraints in service: part 2 - examples from health care." *Managing Service Quality*, 1996: Vol. 2, No. 2 pp.30-34.

Motwani, Jaideep, Donald Klein, and Raanan Harowitz. "The theory of constraints in services: part 1 - the basics." *Managing Service Quality* 6, no. 1 (1996): 53-56.

Olson, Christopher T. "The Theory of Constraints: Application to a Service Firm." *Production and Inventory Management Journal:*, 1998: Second Qtr. 1998; 39,2 55-59.

Reid, Richard A. "Applying the TOC five-step focusing process in the service sector, A banking subsystem." *Managing Service Quality*, 2007Vol. 17, No. 2, 209-234.

Sale, M. L., and R. A. Inman. "Survey based comparison of performance and change in performanc of firms using raditional manufacturing, JIT, and TOC." *IJPR*, 2003: Vol. 41, No. 4, 829-844.

Schmenner, R. W. "How can service businesses survive and prosper?" *Sloan Managment Review*, Spring Vol. 27 (3) 1986: 21-32.

Siha, Samia. "A classified model for applying the theory of constaints to service organizations." *Managing Service Quality*, 1999: Vol. 9, No. 4 255-264.

Simons, Jr., MAJ Jacob V., and Jr., LTCOL Richard I Moore. "Improving Logistics Flow Using the Theory of Constaints." *Military Logistics*, 1992: 14-18.

"Southwest Airlines Fact Sheet." 12 31, 2009. http://www.southwest.com/about_swa/press/factsheet.html#fleet.

Spencer, Michael S., and Samuel Wathen. "applying the Theory of Consttraints'Process management Techique to an Administrative Fiunction at Stanley Furniture." *National Productivity Review*, 1994: 379-385.

Stanley C. Gardiner, John H. Blackstone Jr. and Lorraine R. Gardiner. "The Evolution of a management philosophy: The theory of constraints." *Journal of Operations Management*, 2007: 387-402.

Swets, Nancy. *Rockwell Training Manual.* Pittsburgh: Rockwell, 2010.

Taylor III, Bernard W. *Introduction to Management Science 9e.* Upper Saddle River, New Jersy: Pearson, 2007.

Thorne, Alan, Dan Barrett, and Duncan McFarlane. "Impact of RFID on Aircraft Turnaround Processes." *IEMS*, 2007.

Umble, M, E Umble, and S Murakami. "Implementing the Theory of Constraints in a traditional Japanese manufacturing environment: The case of Hitachi Tool Engineering." *International Journal of Production Research*, 2006: 1863.

US Bureau of Labor Statistics Website. April 28, 2011. ftp://ftp.bls.gov/pub/suppl/empsit.ceseeb1.txt (accessed April 28, 2011).

Van Landeghem, H., and A. Beuselinck. "Reducing passenger boarding time in airplanes: a simulation based approach." *European Journal of Operations Research*, 2002: 294-308.

Van leeuwen, Pim, and Cees Witteveen. "Temporal Decoupling and Determining Resource Needs of Autonomous Agents in the Airport Turnaround Process." *2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* Milan, Italy, 2009. vol. 2, pp.185-192.

Victoria J. Mabin, Steven J. Balderstone. "The performance of the theory of constraints methology." *International Journal of Operations and Production Management*, 2003: 568-595.

Watson, Kevin J., John H. Blackstone, and Stanley C. Gardiner. "The evolution of a management philosophy: The theory of constraints." *Journal of Operations Management*, 2007: 387-402.

Wu, Cheng-Lung. "Monitoring Aircraft Turnaround Operations: Framework Development, Application and Implications for Airline Operations." *Transportation Planning and Technology*, 2008: 215-228.

Wu, Cheng-Lung, and Robert E. Caves. "Aircraft operational costs and turnaround efficiency at airports." *Journal of Air Transport management*, 2000: Vol. 6 201-208.

VITA

STEVEN C. ELLIS

Jan. 4, 1959	Born, Miami, Florida
April 17, 1992	B. S. Mechanical Engineering Florida International University Miami, Florida
1992 – 2001	Manager of Distributor Development Cummins Engine Company, Inc. Miramar, Florida
Dec. 5, 1997	Master of Business Administration Florida International University Miami, Florida
2001 – Present	Instructor – Decision Sciences Florida International University Miami, Florida
Spring 2011 (Projected Graduation)	PhD Student in Industrial and Systems Engineering Florida International University Miami, Florida

TEACHING AND SERVICE AWARDS

Fall 2008 – Master of International Business Program, FIU – Best Professor

Fall 2008 – Professional MBA Program, FIU – Best Course for Corporate Simulation

Spring 2009 – Master of International Business Program, FIU – Best Professor

Fall 2009 – Master of International Business Program, FIU – Best Professor

2004 – 2005 – Outstanding Service Award, FIU College of Business Administration

Spring 2010 – Master of International Business Program, FIU – Best Course

Spring 2010 – Master of International Business Program, FIU – Best Professor

Summer 2010 – Master of International Business Program, FIU – Best Course

Summer 2010 – Master of International Business Program, FIU – Best Professor