# BORDER CROSSING MODELING AND ANALYSIS:

# A NON-STATIONARY DYNAMIC REALLOCATION METHODOLOGY FOR

# TERMINATING QUEUEING SYSTEMS

A Dissertation

by

HIRAM MOYA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2012

Major Subject: Industrial Engineering

Border Crossing Modeling and Analysis:

A Non-Stationary Dynamic Reallocation Methodology For Terminating Queueing Systems

# BORDER CROSSING MODELING AND ANALYSIS:

# A NON-STATIONARY DYNAMIC REALLOCATION METHODOLOGY FOR TERMINATING QUEUEING SYSTEMS

A Dissertation

by

HIRAM MOYA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

| | |
|---|---|
| Chair of Committee, | Guy L. Curry |
| Committee Members, | Antonio Arreola-Risa |
| | Richard M. Feldman |
| | Don T. Phillips |
| Head of Department, | César Malavé |

August 2012

Major Subject: Industrial Engineering

# ABSTRACT

Border Crossing Modeling and Analysis:
A Non-Stationary Dynamic Reallocation Methodology For Terminating Queueing Systems. (August 2012)
Hiram Moya, B.S., Texas A&M University; M.S., Texas A&M University
Chair of Advisory Committee: Dr. Guy L. Curry

The United States international land boundary is a volatile, security intense area. In 2010, the combined trade was $918 billion within North American nations, with 80% transported by commercial trucks. Over 50 million commercial vehicles cross the Texas/Mexico border every year, not including private vehicles and pedestrian traffic, between Brownsville and El Paso, Texas, through one of over 25 major border crossings called "ports of entry" (POE). Recently, securing our southwest border from terrorist interventions, undocumented immigrants, and the illegal flow of drugs and guns has dominated the need to efficiently and effectively process people, goods and traffic. Increasing security and inspection requirements are seriously affecting transit times. Each POE is configured as a multi-commodity, prioritized queueing network which rarely, if ever, operates in steady-state. Therefore, the problem is about finding a balance between a reduction of wait time and its variance, POE operation costs, and the sustainment of a security level.

The contribution of the dissertation is three-fold. The first uses queueing theory on the border crossing process to develop a methodology that decreases border wait times without increasing costs or affecting security procedures. The outcome is the development of the Dynamic Reallocation Methodology (DRM). Currently at the POE, inspection stations are fixed and can only inspect one truck type, FAST or Non-FAST program participant. The methodology proposes moveable servers that

once a threshold is met, can be switched to service the other type of truck. Particular emphasis is given to inspection (service) times under time-varying arrivals (demands).

The second contribution is an analytical model of the POE, to analyze the effects of the DRM. First assuming a Markovian service time, DRM benefits are evaluated. However, field data and other research suggest a general distribution for service time. Therefore, a Coxian k-phased approximation is implemented. The DRM is analyzed under this new baseline using expected number in the system, and cycle times.

A variance reduction procedure is also proposed and evaluated under DRM. Results show that queue length and wait time is reduced 10 to 33% depending on load, while increasing FAST wait time by less than three minutes.

To my wife
and children

# ACKNOWLEDGMENTS

I begin by offering my sincere appreciation to my committee members for serving in my advisory committee and for always insisting on excellence. In particular to Dr. Guy L. Curry because throughout my graduate school years, he has supported me in all my difficulties, and trusted me in many ways. I also appreciate the support from Dr. Don T. Phillips, his patience and feedback, and for allowing me to serve as a graduate research assistance in the Center of Excellence for Border Security Research. I want to thank Dr. Richard M. Feldman for believing in me, and helping me clarify all questions, regardless of how simple. And a warm thank you to Dr. Antonio (Tony) Arreola-Risa, for believing in me, and allowing me to participate in his academic business area.

To my family, I would like to say *gracias* because I owe my formation and the person that I am today. To my father †, that may he rest in peace; to my mother for her blessings and love; to all my brothers for their constant inspiration; and to all my nieces and nephews for their love and support, thank you all.

I also want to express my deepest gratitude to my wife who has supported me in all our endeavors, and for being a pillar of support for me during these years as our family has grown. To my children for their unwavering love. Thank you for being the reason I live. *¡Los quiero mucho!*

Most importantly, I want to thank God the Father, My Lord Jesus Christ, and the Holy Spirit for staying with me even when I do not deserve it, without your support I would not succeed. Thank you ever Virgin Mary for your constant intercessions and protection.

# NOMENCLATURE

| | |
|---|---|
| BOTA | The Bridge of the Americas |
| CBP | Customs and Border Protection |
| CT | Cycle Time |
| DHS | U.S. Department of Homeland Security |
| DRM | Dynamic Reallocation Methodology |
| FAST | Free and Secure Trade |
| MGE | Mixtures of Generalized Erlang (distributions) |
| NAFTA | North American Free Trade Agreement |
| NTAS | National Terrorism Advisory System |
| POE | Port of Entry |
| TTI | Texas Transportation Institute |
| TxDOT | Texas Department of Transportation |
| WIP | Work in Process |

# TABLE OF CONTENTS

Page

Page

Page

# LIST OF TABLES

# LIST OF FIGURES

FIGURE Page

# 1. INTRODUCTION: IMPORTANCE OF THE RESEARCH

This section introduces the significance of the United States' (U.S.) international commercial truck trade, which include getting goods to cross securely in trucks, cost efficiently and expeditiously across international border lines; and how it can be considered as a terminating queueing system problem. This section also presents the significance of the problem because of the volume, value and distance traveled of all the goods that this trade represents. And finally, concludes with a brief introduction of the research methodology, contribution and organization of the dissertation.

## 1.1 Background

Economies in the world are driven by the flow of goods and services to and from each country. As in many countries today, economic stability and growth of the U.S. is linked by its global supply chain with the rest of the world (Willis and Ortiz, 2004). All modes of transportation are used to get products, parts or supplies to their destinations throughout the world. But except for some air freight, all products that enter the U.S. by rail, ship, or truck, go through border inspection stations called "Ports of Entry" (POE). This flow of goods is a key component of the nation's economic engine, which in turn makes the U.S. international land boundary an economically significant, volatile, and security intense area. A significant part of the economic productivity is supported by the use use of sub-assembly and manufacturing plants in foreign countries. Canada and Mexico are neighboring countries that offer significant benefits over other countries because of their physical closeness. And in situations for example the automotive industry, where goods cross the border many times before becoming a final product, the benefits are more significant.

---

This dissertation follows the style of Management Science.

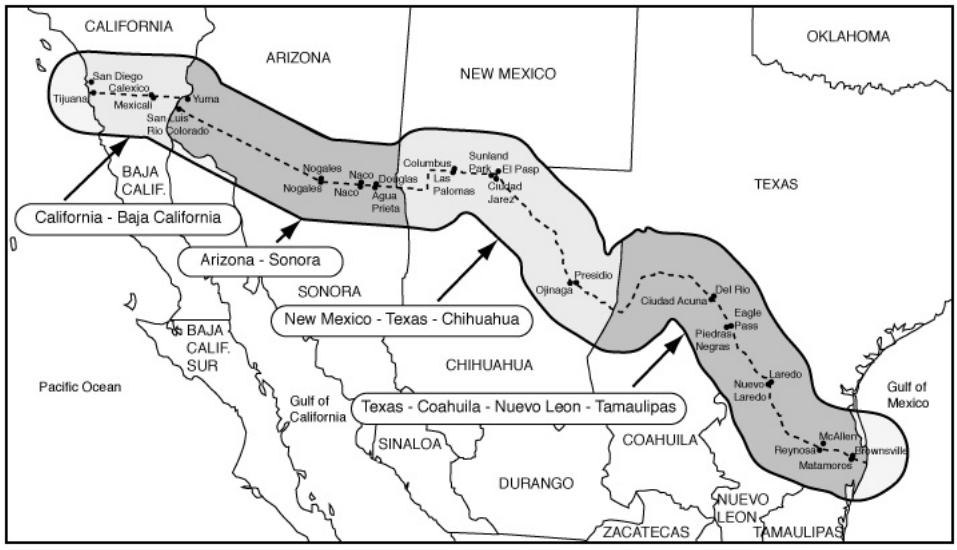



**Fig. 1.1.**: Twin Cities Along the U.S.-Mexico Border (WordPress, 2010)

Many manufacturing plants have been established in Canada, but twin plants, or maquiladoras as they are known in Mexico, offer benefits like inexpensive labor, and other tax benefits. The term comes from "maquila" which was considered as a measure of corn or oil that farmers would use to barter with millers in exchange of grinding services. Maquiladoras were first proposed after the U.S. ended the "Brazero" program in the 1970's when Mexico started the Border Industrialization Program (or Programa de Industrialización Fronteriza) (MacLachlan and Aguilar, 1998).

The program enticed foreign companies to invest in Mexico and establish manufacturing plants to make the goods from inside of the country. Within this program, companies that have operations in the U.S. established twin plants around the border region. The parts and products they work on cross the border several times in sub-assembly and final assembly before they continue down their own supply chain road to get distributed locally or shipped internationally to the end customers. The plant built in the U.S. would serve to supply, redeploy and backup the operations of products going into and coming out of the manufacturing plants in Mexico. The

other plant built in Mexico,would do the brunt of the manufacturing work with mostly local labor, imported parts and some local supplies. In particular, women labor proved to be an efficient and reliable work force for the maquiladoras. which fueled the growth of the border towns. Figure 1.1 identifies the major border cities that have seen manufacturing growth from maquiladoras (WordPress, 2010).

The economy growth in the area was amplified by the North American Free Trade Agreement (NAFTA). The treaty gives economic preferences and removes tariffs to products and materials from the member countries. Parts of NAFTA went into effect immediately, like the elimination of tariffs on most materials from member countries. Other parts of the treaty are implemented by stages, for example the ability of Mexican truckers to compete and deliver goods inside the U.S. territory. Overall, several recent economic indicators from public and private entities, such as the El Paso Regional Economic Development Corporation (2010), Deloitte Touche Tohmatsu and the U. S. Council on Competitiveness (2010), identify the advantages of business and manufacturing operations in Mexico and NAFTA countries in general. In effect, these twin manufacturing plants take advantage of:

- Proximity to the U.S.
- Lower labor costs associated with Mexican employees
- Mostly nonunion technically capable workforce
- Elimination of international trade tariffs
- Strong transportation infrastructure support

Even before, and with the implementation of NAFTA, international trade has strengthened the position of Mexico's northern border cities. Populations in towns like Ciudad Juárez has boomed since 1996, and "the trend is mirrored by other cities along the border" (Chavez, 2004). According to the National Institute of Statistics and geography (Instituto Nacional de Estadística y Geografía, INEGI), in January

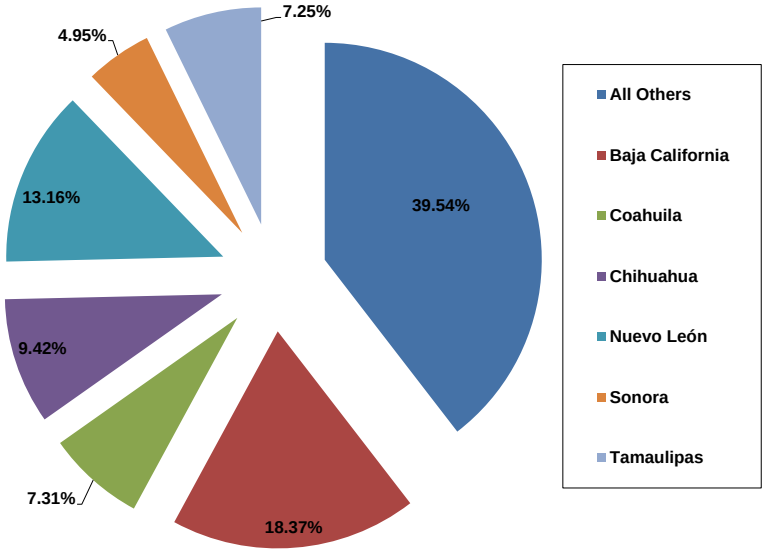



**Fig. 1.2.**: 2011 Percentage of Maquiladoras in Mexico by State (INEGI, 2011)

2011 there were 5,106 maquiladoras functioning in Mexico, and over 60% of them are in the northern border states of Mexico. Refer to Figure 1.2 for a detailed breakdown of maquiladoras by Mexican state, and in particular by U.S.-bordering mexican state (INEGI, 2011).

As mentioned before, all the goods that are imported by truck containers, get scanned and inspected at one of the 154 land based POEs when they enter the U.S. (U.S. Department of Homeland Security, 2011c). To get through the border, these trucks wait in line resembling a queue. The trucks send documentation in advanced of their cargo, and as they reach Mexican Customs checkpoint, they go through inspection. Once finished with Mexican authorities, trucks continue to U.S. inspection where their documentation is verified and the trucks get confirmed, scanned and inspected by various methods and technologies. And with over 3 million trucks with containers crossing the border every year, improvements in the bi-national supply chains will positively impact the economy (RITA, 2011).

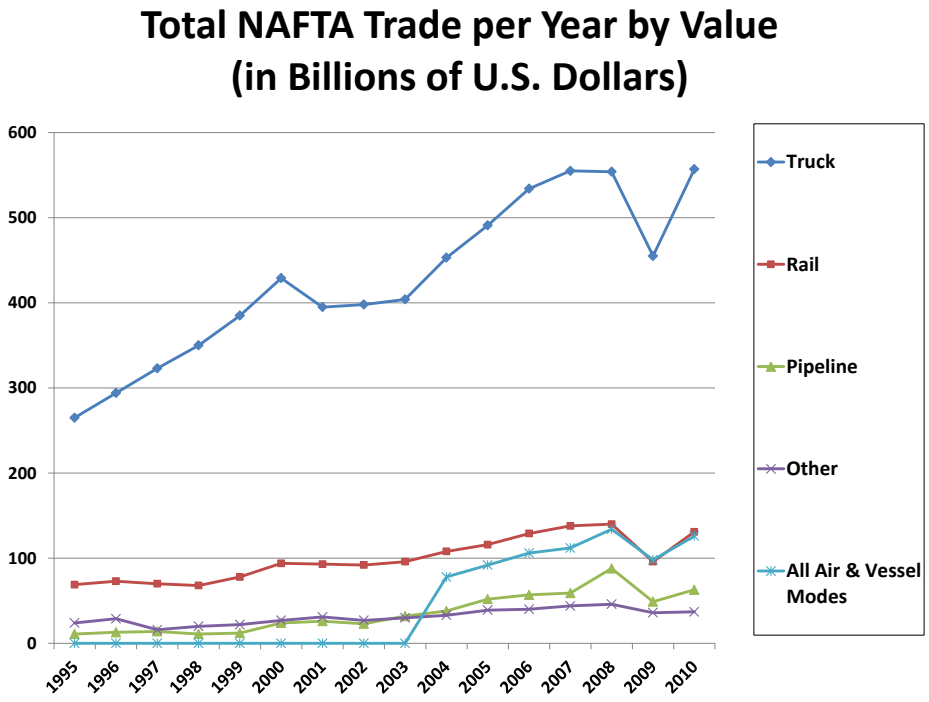



**Fig. 1.3.**: 1995-2010 NAFTA Trade by Mode (RITA, 2011)

## 1.2 Value of Trade

Except for some air freight, most goods are transported in containers through rail, ship, and trucks into and out of the U.S. And although traded goods come from all over the world, consistently since 1998, over 60% of the exports and imports comes from the top ten trade countries, with Canada and Mexico being the top two or in the top three economic trade partners of the U.S. According to data from the U.S. Census Bureau, in 2010, the combined trade was $918 billion within NAFTA countries. Canada, the U.S. number one global trading partner, accounted for $524.67 billion dollars; and Mexico, the number three global trading partner, produced $392.98 billion dollars in trade. by top ten countries, NAFTA trade is bigger than the trade with all of Asia of $787.34 and bigger than the European Union trade of $294.69 (U.S. Census, 2011).

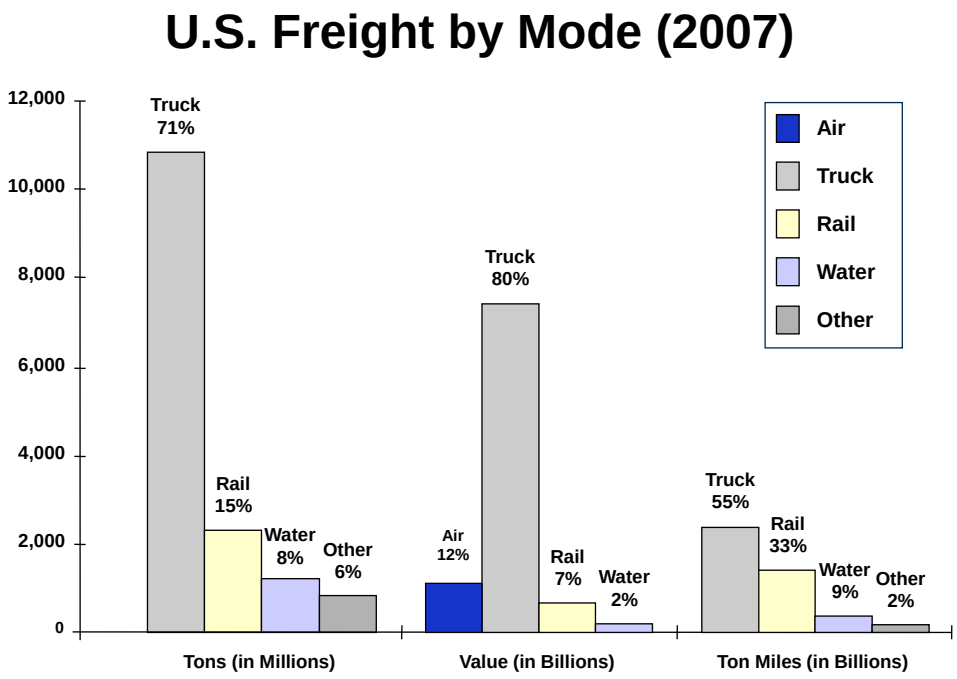



**Fig. 1.4.**: 2007 Freight by Mode Comparison (White, 2007)

Within the North America, freight is moved by trucks and trains, and the base transportation unit is the container. This container is typically a steel box of standard dimensions that carry most of the freight through the national highway system and railroads (Willis and Ortiz, 2004). Observe in Figure 1.3, with data from the U.S. Department of Transportation, and the Research and Innovative Technology Administration (RITA), that since 1995 truck shipping is the leading mode of transportation. And even with the economic downturn of 2009, 80% or more of the trade is consistently being handled by commercial trucks (RITA, 2011).

In his presentation, Dr. White, Schneider National Chair of Transportation & Logistics from the Georgia Institute of Technology, investigates these trends and the research direction of global supply chains and logistics. From Figure 1.4, and using detailed data from 2007, he compares all modes of transportation in the U.S., and by various measures, including tonnage, value and ton/millage being transported. His

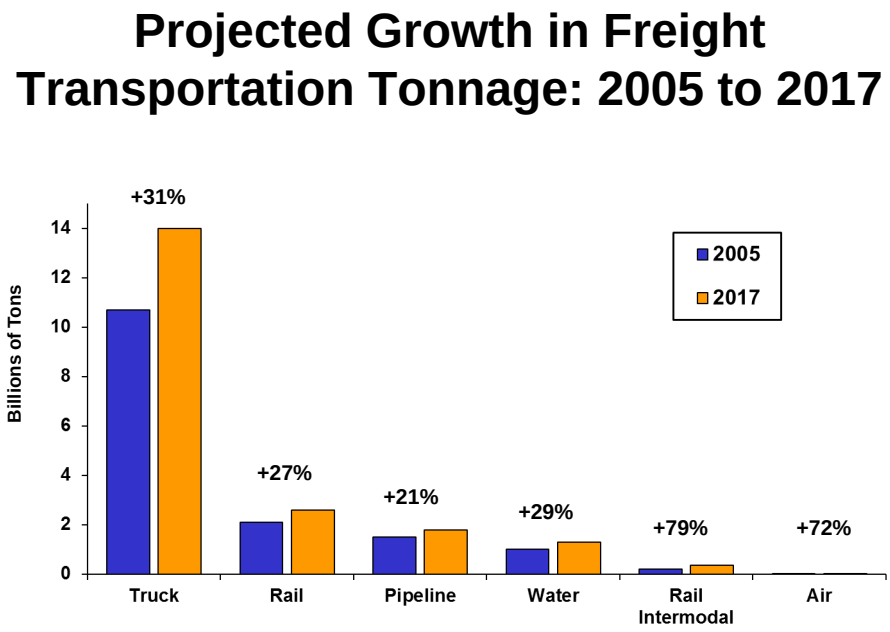



**Fig. 1.5.**: 2017 Projected Growth in Freight Transportation (White, 2007)

conclusion was that the trucking industry was not only "to remain dominant in the U.S. freight transportation mode" but this trend will increase (2007). Making truck based shipping the dominant transportation method by amount, distance, and value of items being shipped.

And the trend is likely to continue. In Texas alone, the commercial truck shipping business transports 85 billion tons yearly along the Texas/Mexico border, through one of over 25 major POEs, which handles over 5 million commercial vehicles crossings every year, not including private vehicles and pedestrian traffic (RITA, 2011; Texas A&M International University, 2011; U.S. Department of Homeland Security, 2011c).

Another motivating factor is the growth potential in this area. NAFTA has provisions that allow companies from the member countries to be able to transport goods and supplies across the road ways of the member countries. This means that

the recent activation of this provision, although by step wise implementation, will only generate greater opportunities to minimize transportation costs, and optimize the delivery of supplies even in the face of uncertain events or delays. And although there are certainly other methods of transportation, shipping companies like UPS, and FedEx are relying more and more on their truck fleet to deliver the packages domestically.

Dr. White also mentioned that according to the U.S. freight transportation forecast, the trend of using trucks to carry loads will continue to increase as it becomes more profitable for companies to use truckloads and less than truck loads to transport goods and packages. Observe from Figure 1.5, that the forecast expects to see a 31% increase in the use trucks to ship goods and supplies in tons, by the year 2017, which translates to an increase of 4 billion tons per year, much more than any other mode of transportation (White, 2007).

## 1.3 Research Objective, Contribution and Methodology

The main research objective is to use queueing theory on the border crossing process, and develop a methodology that could decrease cycle time, or border wait times at the POE, without increasing costs or affecting current security procedures. Currently the POE has a fixed number of inspection stations that serve only specific truck types. From a queueing theory perspective, the research focuses on the development of a non-stationary, congestion-based operational methodology that will improve throughput and other system performance measures on non-stationary, terminating queueing systems.

The next objective is to evaluate and analyze the performance of such a methodology using a Markovian assumption for inspection or service time. Afterwards, an analytical model of the POE is developed using an approximation of the inspection time with a mixture of generalized Erlang distributions, also called Coxian distributions. The objective is to set a more realistic baseline model for the POE. With this

new baseline, the methodology is re-tested using this approximation to assess and analyze the performance improvement.

### 1.3.1 Contribution of the Research

Contributions of the dissertation include, the non-stationary dynamic reallocation methodology, the analytical model of the POE with a general service time approximation, the analysis of the POE system that never reaches steady-state, and a variance reduction policy that also improves system performance. The basis of the application of this research is the U.S. bound border crossing process at every POE. This methodology improves system performance measures, while at the same time having little or no effect on either the cost or security procedures of U.S. Homeland Security - Customs and Border Protection (CBP) Agents. Emphasis is given to the dynamic reallocation of inspection (service) facilities and inspectors under time-varying arrivals (demands), via a transient queueing network analysis to evaluate throughput rates, queue size, cycle times and configuration effectiveness.

The specific problem is how to manage and optimize or balance the entire system, considering security constraints, cost, and supply chain flows. In such a way that the system can anticipate uncertainties in security threats, limitations in facilities, and unexpected increases and decreases in cargo shipments across the border, as observed by economic ups and downs.

To address these issues, the dissertation will consider the following research questions:

- Using queueing theory on the border crossing process, can a method be developed that could decrease cycle time, or border wait times at the POE, without increasing costs or affecting current security procedures?
- For a given POE, what is the current situation in resource utilization?

- Without major infrastructure investments, what are some improvements that can be identified from the analysis?
- Can an analytical model accurately represent the POE, and this methodology to verify the results?
- For the methodology developed, what analysis can be done to support expected improvements, and what are the expected results?
- What performance measures should be used to better assess the current system, its effectiveness and possible improvements?
- Given that the suggestions could be implemented, what are the expected improvements in the identified methodology in terms of performance measures?

The first question is the main contribution of this dissertation. The next two are considered in Section 5, the following two questions the subject of Section 6, while the last two questions are left for discussion in Section 7. In effect, and to answer these questions, the research focuses on modeling and analyzing the border crossing process at a single location, the Bridge of the Americas (BOTA) in El Paso, Texas, as a terminating non-stationary queueing system. The model will use empirical data for the scanning and inspection times. For inter-arrival rate of trucks, as in the work done by Zhang (2009) for the analysis done by at the northern border of non-commercial crossing, the initial assumption is that those inter-arrival times were independent and exponentially distributed. Furthermore, analytical methods as well as software as used to create an "as-is" model in order to benchmark the statistics on time in the system, utilization of secondary inspections, average queue length, variance of time in system, and other statistics. In addition, the research will propose a new congestion based policy to create a dynamic reallocation policy of servers to improve system performance. With performance measures defined, attention turns to an assessment of the current system, to be able to evaluate any proposed methodology

changes. Once the benchmarks are obtained and verified with on field data, the research will focus on improvements with or without major infrastructure changes. With the model, test scenarios will be created, and compared with existing research. Cases will be setup for low, medium, and high economic activity, and an increase in security from elevated to high and severe.

The main contribution to the academic body of knowledge is the proposed dynamic reallocation methodology to assess queueing systems with multiple diverging objectives. In addition, the POE inspection time is explored in detail, first with a Markovian assumption for service time, then with a general distribution approximation. When the methodology and improvements are demonstrated, one additional process based improvement is discussed to reduce variance by increasing the use of secondary inspections. Finally, future research can explore the application of this methodology in land and sea based POEs, and possible other geographical locations around the world.

### 1.3.2 Approach

This approach will combine analytical methods with empirical data to create an analytical model, in conjunction with a simulation model using the Arena simulation software in order to analyze the current, as-is situation at the selected POE. In order to characterize the stochastic nature of inspection time, a Coxian Phased approximation is used for the general service time distribution. Afterwards, the dynamic reallocation policy is developed, partially based on a modification and extension of the "Congestion-Based Staffing" policy used to analyze the non-commercial border crossing process in the northern border with Canada by Zhang (2009)in his recent paper. Similarly to this research effort, Zhang assumes that the customer's inter-arrival times are exponentially distributed and mutually independent. But unlike his work, the service times in this dissertation are not assumed to be exponentially distributed and mutually independent. Therefore, the model is not a clear cut Markovian multi-

server queue ($M/M/c$) where a stationary distribution of the queue length can be found.

In this research, performance measures based on inspection station (server) utilization and effectiveness are proposed, in order to set benchmarks. First, the focus is on creating a baseline simulation model in Arena, where the wait line queues, the scanning and inspection process, and the secondary inspections are represented. Once the model and benchmarks are validated, the plan is to create test instances for validation and "what-if" analysis to explore and find a balance between the identified key factors of wait time, cost and security level. In this effort, test cases will also be created using empirical data to assess changes in the proposed methodology.

To obtain empirical data, collaboration was established with ongoing research efforts by the Texas Transportation Institute (TTI) and the Industrial & Systems Engineering Department at Texas A&M University, where the PIs, Juan Villa and Dr. Melissa Tooley, conducted the Screening Scanning and Inspection Processes (SSIP) project and the Advanced Security Procedures at Border Crossing Points of Entry (ASBC) project. Involvement with this research effort became complementary; the SSIP focuses on the technology, and the ASBC looks at procedures, while this research looks at the modeling both analytically and via simulation. Any additional data needed, will be collected on field research and data collection in El Paso, Texas. With successful results, government agencies, large corporations, transportation companies, and the society in general will benefit from improvements in the flow of products within the North American free trade region. The population will benefit from lower costs of products, a safer environment, and increased productivity by reduction of wait times at our borders.

## 1.4 Organization of Dissertation

This dissertation is organized in 8 major sections. Section 2 explains the border crossing process, and the challenges from this type of queueing system. The next

section presents a review of the literature on the border crossing process from a Queueing theory perspective, as well as as part of the supply chain, and the research done on modeling the process and security considerations.

The contribution begins in Section 4 with the presentation of the Dynamic Reallocation Methodology (DRM), based on congestion policies for primary inspection stations. In addition, the collected empirical data is presented in this section, and analyzed for best distribution fitting.

Section 5 provides a deeper discussion and analysis of the POE queue with the Markovian assumption of the service time. As a proof-of-concept, the DRM is implemented with this Markovian assumption, and shows expected improvements versus the current as-is situation. The next section, drops the Markovian assumption, and continues the analysis of the POE, now with a general service time distribution. An analytical model is presented that considers a Coxian k-phased approximation for a general service time approximation. This model is compared with empirical data and serves as a baseline model.

Further on, Section 7 uses the analytical model in the previous section and implements the Coxian k-phased approximation into the simulation model that has the DRM. The effects of the DRM is again evaluated against the current situation, and results are shown. The results of the Coxian / DRM model are compared and contrasted with the Markovian / DRM model to highlight differences in the results. Lastly, the discussion turns to a variance reduction policy that can be used with the DRM, and is based on an expanded use of secondary inspections. Finally, the last section makes concluding remarks, and discusses future research.

# 2. THE COMMERCIAL BORDER CROSSING PROCESS

In this section, the process of getting good across and into the U.S. serves as a problem statement. The description of the physical characteristics and the the steps required to get goods, products or subassemblies from Mexico and into the U.S. serve to describe the limitations of the model that will be constructed. A thorough detail is necessary to evaluate and confirm the validity of the model, and establish it as a baseline for the current as-is situation, and be able to compare the proposed methodology to validate benefits and improvements.

The border crossing process is similar at all POEs around the nation. But for the purposes of this research, the focus will be in the El Paso / Ciudad Juárez / Santa Teresa area. This area is one of the worlds largest border communities, it houses the busiest activity of POEs (as a region) in North America, and has the biggest economic impact as far as dollar values of the shipments coming across each year (REDCO, 2010). Specifically, the research will focus on The Bridge of the Americas (BOTA) POE in El Paso, Texas, which is the only pedestrian, private vehicle, and commercial traffic toll-free bridge of North America.

In the El Paso / Ciudad Juarez / Santa Teresa area, there are approximately close to a million trucks crossing every year with goods and supplies. From REDCO's (2010) report, the number of trucks crossing the border with goods and supplies decreased 15% in 2001 to 612,938 trucks, which was directly attributed to the terrorist attacks of September 11. However, that was immediately reversed the following year with 694,868 trucks crossing, which translated into a 13% increase. Returning not only to a positive growth, but by 2008, the number of trucks increased again to 833,776, which is a 36% increase from the number of trucks crossing at the beginning of the decade. Thus, confirming what seems to be a continued reliable use of trucks as a means of transporting goods and supplies in the supply chain.

## 2.1 The Inherent Conflict of the Objectives

The specific problem is finding a balance between a rapid processing of commercial transient entities and Homeland Security's requirements to protect against terrorists and other activities. To handle these issues, there is a need for new methodologies that deal with the stochastic or random nature of the threats and worldwide commerce, while optimizing limited resources, and maintaining a set level of security. These objectives and requirements could be optimized independently; however, it cannot be guaranteed that the optimized solution for one specific condition will be the optimal for the others. Furthermore, it can be argued that maximizing throughput, which is a supply chain problem that searches for optimal methods to distribute and minimize delay in delivery of the goods during shipping, increases cost and decreases security. Also, minimizing cost, which includes personnel to inspect and manage the checkpoint, building and maintenance costs of infrastructure, etc., reduces throughput and security. And finally, efforts to maintain and guarantee a set level of security affects the transportation of incoming and outgoing cargo in the U.S. which increases cost and decreases throughput.

It the border community, all three main stake-holders for the border crossing process, DHS, the business community and the public at large, have similar objectives. These objectives are identified as key factors in border crossing process. The objectives are:

1. Minimize government cost associated with the inspection and scanning process.
2. Maximize economic throughput of north bound goods and supplies (i.e. Supply Chain Flows).
3. Maintain a level of inspection scrutiny and security at the POEs set forth by the National Terrorism Advisory System (NTAS).

Because these objectives seem to conflict with each other if they are optimized independently, their independent solutions would not optimize the other objectives

measures. This kind of problem is similar to a call center where you have customers waiting in a queue, and the objective is to minimize wait time, and maintain a certain service level. But in the case of cargo going across the border, the objective is not necessarily to optimize one single objective, but to maintain a balance between the identified key factors so that the performance indicators are within pre-determined ranges.

## 2.2 Security Concerns and Advisory Systems

The border crossing process is not just a logistics or supply chain problem, it is a national security issue as well. Transporting parts or supplies to manufacturing locations or distribution centers, has historically been reduced to a typical supply chain problem solved by logistics algorithms that optimize transportation routes and commodity movements. While moving a product to market quickly is very important to companies, costs and other factors are also carefully monitored. CBP Agents are tasked with the screening, scanning and inspection containers of goods across the U.S. borders. Before the turn of the century, customs enforcement in the transportation of freight was focused on contraband and drug enforcement. But since the events of September 11, 2001, a higher emphasis on security has been introduced in the transportation system.

Recently, securing the southwest border from undocumented immigrants, terrorist intervention and the illegal flow of drugs and guns, has dominated the need to efficiently and effectively process people, goods and traffic through the POEs. When the conditions are such, it is necessary to take additional steps in inspecting certain trucks or containers, that come through our borders. This implies that the cargo and the driver are inspected with more detailed at secondary inspection facilities. All of these decisions affect the time it takes for a shipment to come into the U.S. This also affects trucks and cargo containers behind in the queue. But increasing security and inspection requirements significantly affect transit times, which ulti-


**Fig. 2.1.**: Former Threat Levels of the Homeland Security Advisory System

mately impact the nation's economic engine. Today, current economic conditions require an effective border security program, as well as streamlined supply chains, effective transportation methods, low logistics costs, and higher throughput of parts and assemblies across the U.S. borders. Hence the difficulty of balancing objectives. In effect, optimizing security would tend to maximize security procedures, which intuitively would decrease throughput, increase costs or deteriorate overall system performance.

### 2.2.1 Homeland Security Advisory System

To deal with threats and manage national security, the Homeland Security Advisory System was created in 2002. Later on, several government agencies dealing with the nation's security were reorganized into the Department of Homeland Security (DHS) which was created on March 6, 2003. DHS managed the advisory system


**Fig. 2.2.**: The New National Terrorism Advisory System

in Figure 2.1, which indicates and explains the current risk of terrorist attacks to the nation. The levels range from "LOW" to "SEVERE" (U.S. Department of Homeland Security, 2011d). This advisory system is also very important to transportation firms. Whenever the advisory system moves to a higher level of security, inspections are more thorough and that translates to longer wait times at the POE. This also implies that more resources are tied up at the POEs limiting their use, and can have a negative impact in profits.

### 2.2.2 National Terrorism Advisory System

In April 21, 2011, DHS introduced the "National Terrorism Advisory System" (NTAS), which replaces the color-coded "Homeland Security Advisory System". According to DHS, "this new system will more effectively communicate information about terrorist threats by providing timely, detailed information to the public, government agencies, first responders, airports and other transportation hubs, and the private sector." The alert system will only be used when DHS has "credible information" about a particular threat. The new system only has "imminent threat" or "elevated threat" levels any of for the alerts it will publish.

The information icon can be placed in private and other public web-pages, as seen in Figure 2.2. The alerts now also include a "sunset provision indicating a

specific date when the alert expires" (2011d). The main difference is that there is no constant assessment of the current threat level for the U.S. In Homeland Security Advisory System, the level was at yellow at the time it ceased to be operated, and it did not provide specific information to the public regarding a specific threat. This new system acknowledges that we are in constant alert, and that specific information about an impending terrorist threat is more useful to the public.

## 2.3 Border Crossing Procedures

The process that empty or laden commercial vehicles follow to cross the border is almost the same at all POEs. And although NAFTA has eliminated or reduced the tariffs for goods traded among the three member countries, Canada, the U.S. and Mexico, the treaty is still not fully implemented. Getting goods, and supplies, across the border still requires documentation of the cargo being introduced. Starting June 2009, every individual entering the U.S. legally must present a valid passport, including all U.S. citizens. This new regulation is applicable in all air, sea and land POEs. CBP agents are not only looking to enforce tax and tariffs, but DHS' mission of securing the country from terrorists and the smuggling of illicit cargo including drugs, weapons and human trafficking.

Figure 2.3 contains the standard method that trucks and other shippers use to get goods into the U.S. But geographic limitations and infrastructure configurations set up the way the queues are formed. For commercial vehicles then, the process is summarized as follows: Commercial shipments from Mexico into the U.S. require going through three inspection stations: Mexican Customs Export Lot, U.S. Federal Compound and State Vehicle Safety Inspection Facility, except when the truck is empty then Mexican Customs Export Lot is omitted.

At the Mexican Export Lot, the Mexican Customs (Administración General de Aduana) conducts inspections consisting of a physical review of the cargo of randomly

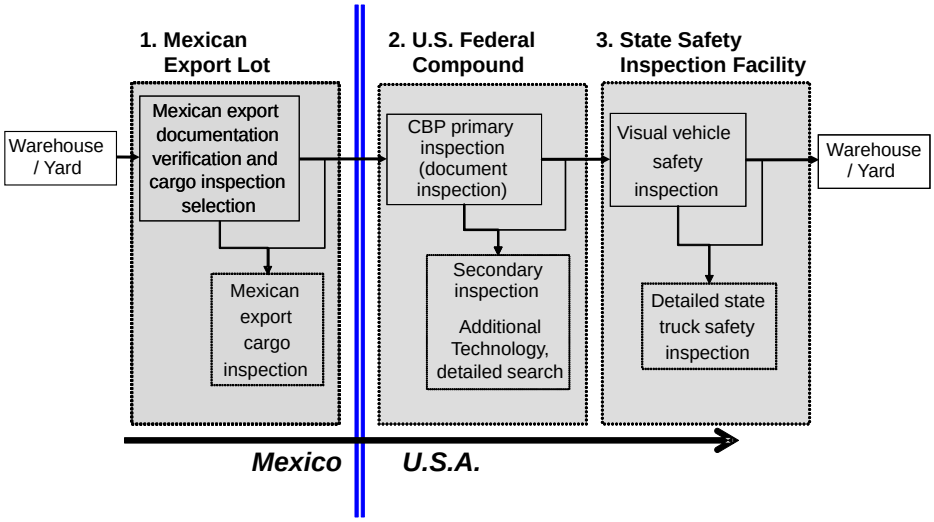



**Fig. 2.3.**: Schematic of the Border Crossing Process

selected outbound freight prior to its export. Shipments that are not selected proceed to the exit gate, cross the border, and continue on to the U.S. port of entry.

The truck proceeds into the U.S., and at the primary inspection booth, the truck driver presents documentation to the processing agent. The CBP inspector at the primary inspection booth uses a computer terminal to cross-check the basic information about the driver, vehicle, and load with information sent previously by the U.S. Customs broker, then makes a decision to refer the truck, driver, or load for a more detailed secondary inspection of any or all of these elements or releases the truck to the exit gate. A secondary inspection includes any inspection that the driver, freight, or conveyance undergoes between the primary inspection and the exit gate of the U.S. Federal Compound.

The Vehicle State Safety Inspection Station is where the state police inspect conveyances to determine whether they are in compliance with U.S. safety standards and regulations.

The process follows a basic queue that forms two lines composed of trucks and other commercial vehicles based on cargo and the type of documentation for customs

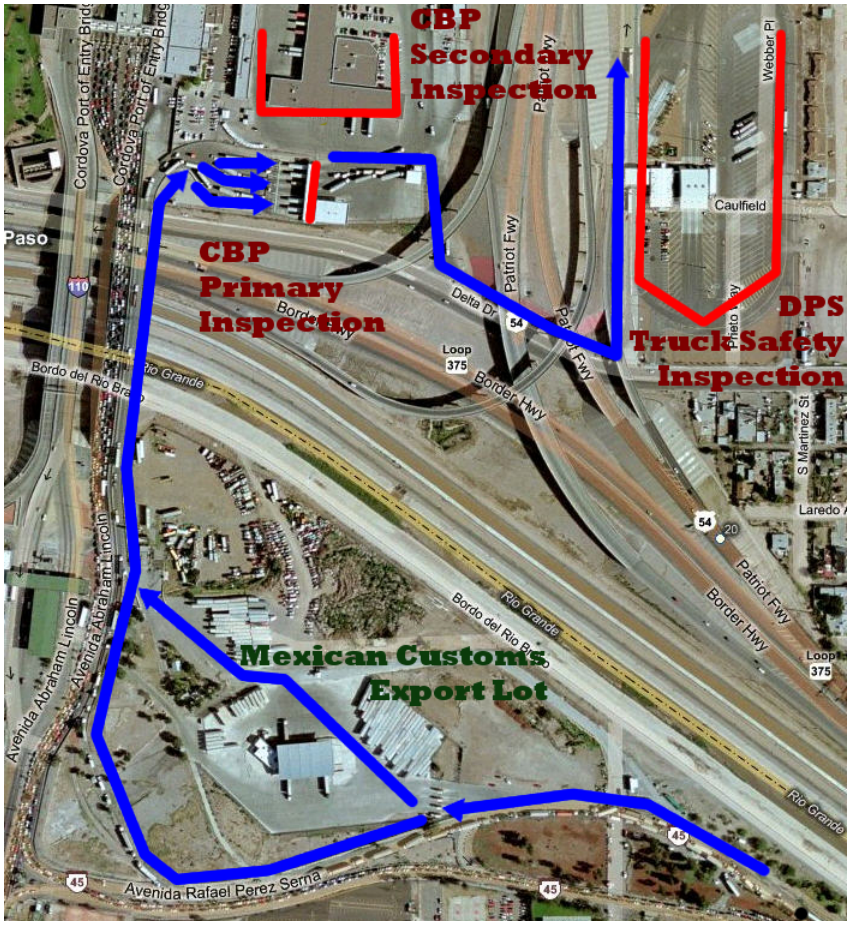



**Fig. 2.4.**: Bird's Eye View of The Bridge of The Americas (Google Maps, 2011)

inspection. The line is then serviced by inspection booths. And with the help of technology, a CBP Agent checks the documentation of the vehicle, and determines whether the vehicle and/or the driver need further inspection.

### 2.3.1 BOTA POE Hours of Operation

Observe in Figure 2.4 that at BOTA, the queue builds up behind the international bridge, and goes around and through the Mexican Export lot for a few miles down the road. Observe that the geographic restrictions, i.e. the bridge, and river, affect and limits the way the lines are formed. In rare or security sensitive occasions the

time it takes to cross the border maybe as long as 15 to 20 hours, while most of the time it takes about one to two hours U.S. Department of Homeland Security (2011a). Hours of operations are from 6:00 am to 6:00 pm Monday to Friday, and 6:00 am to 2:00 pm on Saturdays, with Non-FAST (explained in section 2.4) service starting at 8:00 am every work day, until the end of the service hours.

## 2.4 The Delay Problem

Long wait times for pedestrian, privately owned and commercial vehicles to cross the POE is not a recent problem. Before 9/11, wait times of about an hour was considered long, but now, wait times of over 4 hours might be expected. This delay per truck compounds the cost of trade for the U.S. And the long wait times are not exclusive of the southern border. In the Windsor-Detroit border crossing port, lines after 9/11 extended some 80 kilometers, and queues of four to ten kilometers remained common for several years thereafter, which prompted the government of Ontario and Canada to hire consultants to alleviate the problem (Li et al., 2005).

Economic incentives via lower taxes or cost of goods, are significant considerations that consumers and shippers have to weigh against the delay issues of crossing the border. For example, cigarettes taxes can be significantly different across jurisdictions, which gives Canadian consumers a choice to cross the border into the U.S. and shop (Chiou and Muehlegger, 2008).

Shipping companies, manufacturing plants, consumers, and the environment, have to all pay a "toll" for long delays, particularly for truck lines. Ferris (2000) observed that border crossing shopping is a consideration for consumer and companies that "value two consumption goods (goods that can and cannot he smuggled), leisure, and government services (provided through commodity taxes). However, today's reality also requires an increased focus on security to combat terrorist activities and illegal contraband.

### 2.4.1 The Free and Secure Trade (FAST) Program

To address security, and also the long lines, CBP under the U.S. Department of Homeland Security has established the Free and Secure Trade (FAST) program. Started after 9/11, the FAST program allows for expedited processing of certain commercial carriers. According to the press release, the FAST program "is designed to enhance the security and safety of North America, while also bolstering the economic prosperity of U.S., Canada, and Mexico, by aligning, to the maximum extent possible, their commercial processing programs" U.S. Department of Homeland Security (2003).

The FAST program is a "commercial clearance program for known low-risk shipments entering the U.S. from Canada and Mexico" U.S. Department of Homeland Security (2011b). The drivers, vehicles and cargo have to complete a background check and fulfill certain eligibility requirements in order to qualify for the program. But once registered, their inspection time is significantly shorter than those vehicles not pre-approved. If an inspector is dealing with FAST trucks, then only the FAST-approved vehicles will be in the queue in front of the inspection station. However, having all three components registered is not always the case. For instance, sometimes the cargo is not FAST certified, or the truck that the driver is using is not the usual one because of maintenance, or the original driver is absent. Moreover, there is a cost associated with using the FAST program, and some smaller freight companies do not see the economic benefit of registering with the FAST program.

Whatever the case may be, the wait time to cross the border must be balanced with security issues, and cost concerns. It should also be noted that there is a similar program for privately owned vehicles and their passengers, which is called Secure Electronic Network for Travelers Rapid Inspection (SENTRI). The SENTRI program works similarly to the FAST program, in that there is a dedicated "commuter lane" that offers shorter wait times to cross the border. The program also has an associated cost to it from the Mexican and American authorities.

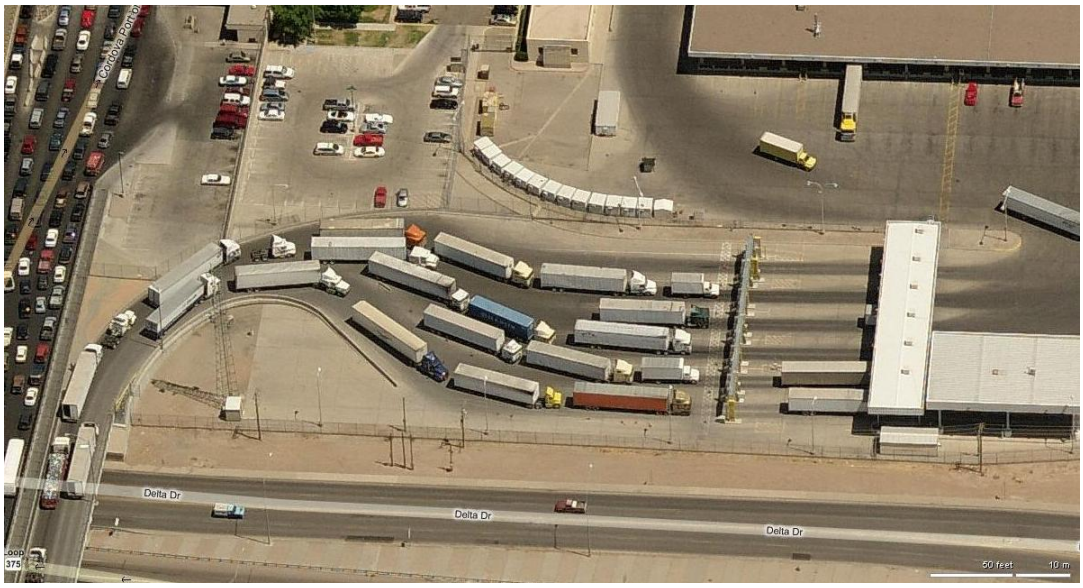



**Fig. 2.5.**: Close-up View of the POE at The Bridge of The Americas (Bing Maps, 2011)

From the expanded view of the BOTA POE in Figure 2.5, the border crossing process can be considered as two long queues, FAST and Non-FAST, feeding short multiple queues of up to five commercial trucks (customers), in front of each inspection station (server). The CBP Agent in the inspection station can only handle either FAST or Non-FAST vehicles. And the number of open inspection stations depends on the POE authorities. Whether opened or closed, each inspection station is static, that is, an inspection station does not switch between serving FAST and Non-FAST trucks. The current configuration has the top four inspection stations serving Non-FAST trucks while the bottom two inspection stations serve FAST trucks and cargo.

## 2.5 Summary

The commercial border crossing process is a multi step, secure sensitive process. Inbound traffic to the U.S. must comply with several requirements. And trucks that carry good and parts across the border must go through a POE to verify that the items carried are admissible into the U.S. After the terrorist events of 9/11 an higher emphasis on security is required. But since commerce is dependent on the throughput, and cost is closely monitored for government agencies, a balance of these objectives is necessary.

Commerce and other private individuals will still have to deal with delays when crossing the border. However, all parties involved realize the significance of disrupting trade. According to the U.S. Department of Homeland Security (2003), the FAST program was a good start because it "uses common risk-management principles, supply chain security, industry partnerships, and advanced technology to improve the efficiency of screening and clearing commercial traffic at ports of entry along the U.S./Canada and U.S./Mexico borders." However, the benefits have been unevenly distributed among the ports and companies (Bradbury, 2010). So any improvements to this problem will translate to significant positive impact to all the stake-holders, private citizens, the small and large businesses and the public sector, plus the economic benefits as a whole.

# 3. LITERATURE REVIEW

An initial literature review for a topic such as "border crossing" can lead to a wide variety of topics, including business management, infrastructure and traffic control, environmental issues, to political science research as well. However, this research focuses on topics that deal with basic research in queueing theory processes that relates to the process of getting admissible products across the border and into the U.S. in a safe and expeditious way.

Additional topics such as homeland security, also bring a variety of research areas, including screening and scanning technologies for the items and people that cross the border, to research of the process and procedures of crossing the border. They also include articles such as government documents, committee hearings, speeches, newspaper articles, opinion pages, and web-site reports. In some cases, these articles do not reflect findings of basic research but are important to acknowledge as many aspects of the border crossing process are impacted by national/local issues and current events. When necessary, some of these articles will be included in this literature review for their relevant information contained within.

Since this research is primarily focused on U.S. bound commercial traffic, the literature review will address topics that focus on research that models the border crossing process. This section is thus separated into sections of main interest in this research. The topics include:

1. Border crossing based models
2. Queueing theory approach to the border crossing process
3. Simulation research for border crossing process improvements
4. Variance reduction for POEs.

## 3.1 Border Crossing Models

The literature is extensive both academically and from government and private agencies in topics such as border crossing, border security, stochastic modeling and supply chain analysis. Particularly since September 11, 2001, a great deal of effort has been put in the securing the borders so that no threats come through and produce harm to the nation. But as the nation and other world economies come through economic cycles, a balance needs to be reached for population safety and economic prosperity.

This section considers research, industry and government agency models for the border crossing process. Many models have been used to focus on throughput, on cost, and more recently, on security. But the models differ significantly in their approach and scholarly background.

### 3.1.1 Widespread Approaches to Border Crossing Models

Given the discussion on the value and size of the commercial truck trade in Section 1.2 and the issues of supply chain delays at the POE inspection station in section 2.4, considerable research has been focused on addressing the issues arising from this process. For instance, the border crossing process has been studied from many perspectives, including:

- A congestion problem within a transportation systems formulation framework
- A logistics problem using supply chain management
- An optimization problem by applying operations research
- A systems management problem
- As a freight load problem

- An issue of national security and public policy by national laboratories and state agencies.

However, the focus primarily depends on the research group or agency tasked to look into solving or alleviating the problem, and mainly have focused on process improvement or technology implementation.

### 3.1.2 Border Crossing Models Based on Process Improvements

After 9/11, the member governments of NAFTA have increased restrictions on materials and people crossing the border, and border security has become a primary topic of research in the 2000 decade. Therefore today, the border crossing process and its security go hand in hand, and the associated research has had to address both topics. Research in this area has been addressed by both public and private agencies. For instance, government research institutions have focused research on analyzing the procedural problems and shortcomings of the border crossing system. One proposed solution includes the implementation of a coordination system to improve operations (Ojah et al., 2002). Their conclusion is that the underlying problem is there is no coordination in the planning and operations, and "as a result of this fundamental limitation, each of the public and private stake-holders plans and operates in ways that optimize their individual missions rather than the system as a whole."

Some of these solutions have already partly been implemented, such as the cross-border trusted travel programs, which facilitate land-border crossing of pre-screened low-risk travelers and commercial-truck drivers through exclusive dedicated lanes. In the case of commercial crossing, the FAST Driver Program affords expedited release to approved commercial truck drivers making fully-qualified FAST trips between the U.S and Canada or to the U.S. from Mexico. Bradbury (2010) made an assessment of FAST program along the CanadaU.S. border, and in addition to finding that the benefits have been unevenly distributed among the ports and companies, they

conclude that conclude that small to mid size firms are burden by cost and are unable to capitalize on the benefits. The authors's recommendations are all based on procedures, including "greater regulatory cooperation between Canada and the U.S. to reduce costly duplication and paperwork, and providing tax incentives or subsidies to small and medium-size firms as a means to increase the participation rate in the program." For additional detail on the FAST program, refer to Section 2.4.1.

Updates have also been the subject of research, considering the changes in the economic and political situation and with emphasis on the questions asked at the checkpoints (Villa, 2006). And with the increased focus on security in the past decade, the authors conclude that issues that still require attention are: the commercial border crossing process, information and data, interagency coordination, and binational cooperation. Of interest is one of the factors the authors identified that hinder the commercial border crossing process. They mention that "some supplemental inspections temporarily block primary inspection lanes," which is precisely the issue being address in Section 6.

Other authors have focused on the economic and environmental problems arising from the delays at the border crossing process; and how do they affect the congestion, the impact on air quality, and commerce in the region (Halvey, 2003). The authors identify several problems and classify them as high, medium and low priority. Among the high priority problems, they clearly identify that there are "unnecessary delays in queues at border stations, (i.e.) primary inspection booths." And just like other others, they have identified opportunities to improve inspection efficiencies and reduce wait times, except that the solution are either by using mostly technology and process changes or improvements, or are not fully developed.

### 3.1.3 Border Crossing Models Based on Technology Improvements

Usually in government activities, more money and/or technology is often seen as the solution to the problem. And the activities to secure the border and process the people and good into the U.S. is no exception.

Since the before the implementation of NAFTA, information technologies was already seen as a way to reduce delays at border POEs. In 1998, Nozick et al. created a simulation of a border crossing model that would assess the benefits of "information technologies to speed the processing of commercial vehicles at the border." Their conclusion was that information technologies can be a significant source of improvement to the inspection and processing of people and goods while reducing the amount of resources needed.

Focusing on public-policy, Villegas et al. (2006) researched policy options for land POEs. Their conclusion was that none of the presented options, "whether alone or in combination, has the potential to avoid conflicts between national security requirements that favor more detailed inspections and local traffic flow consideration that favor less detailed inspections." The options were:

- More primary screening
- More secondary inspection
- Higher use of specialized lanes.

Their conclusion supports our assertion that there is a need for a balance in the objectives, because optimizing a single objective independently does not result in a good overall answer.

With the focus on security and the practical operations and traffic flow configuration, Ojah et al. (2002) analyzed the benefits of coordination systems. The authors identified shortcomings in coordination at U.S. Mexico border POEs and recommended alternatives that would "improve operations and reduce congestion

and delay." These improvements are technology based and would require a pilot program for implementation.

As part of border technology driven research, Turnquist and Rawls (2010) proposed a multi-modal network model to assess the vulnerability of trade flow disruptions at one or more of the major bridges and tunnels that are the border crossing POEs. However, improvements in many of these cases imply a major investment associated with the new infrastructure.

Bracchi et al. (2006) created an "analytical modeling technique based on Layered Queuing Networks" to be able to predict the ability to meet certain performance goals with the use of technologies. Their results were used to research the use of machine readable travel documents (i.e., passports, visas, etc.), the use of biometric identifiers, and interactions among multiple information systems. In a hypothetical inspection system, their techniques are comparable to other studies that have used simulations extensively.

## 3.2 Queueing Theory Approach to the Border Crossing Process

At many POEs, arriving trucks can be considered as the customers of a queueing system with a non-stationary arrival process of different types of customers to multiple parallel servers. In most POEs that handle commercial traffic, there are two types of trucks, the ones that participate in the FAST program, and the the ones that do not. Again, please refer to Section 2.4.1 for a detailed description of the FAST program.

Understanding the border crossing process is critical to apply basic research. Of interest, is the analysis for pedestrian border crossing traffic addressed by Zhang (2009). In his research, Zhang used congestion based staffing policies to the pedestrian inspection process under steady-state conditions to maintain acceptable levels in the queue. The approach was to model the POE as a classical Markovian queueing model ($M/M/c$) to find appropriate staffing levels and meet service demand. And

by using congestion based staffing policies, his focus was on maintaining a certain queue length, instead of minimizing it, by opening and closing inspection booths.

Zhang's paper showed the benefit of servers that dynamically open and close for a single queueing system, and the flexibility that they offer. The formulation developed assumed that the customer inter-arrival times and service times are exponentially distributed and mutually independent. This assumption helps in the development of their benchmark congestion-based policy model to find a stationary distribution of the queue length. Unfortunately, each POE is configured as a multi-commodity, prioritized queueing network which rarely, if ever, operates in steady-state. Bell also addressed the use of servers in a classical Markovian ($M/M/2$) decision process where the servers could be removed and characterized the optimal policy by adjusting the number of working servers. Bell (1980).

In research of similar queueing systems, Whitt (2007) looked at the staffing problem in queueing service systems with time-varying demand. This research can be applied to call centers and other service operations, and his work looked at finding the optimal number of service agents as a function of time to maintain a certain level of service. However, call centers can open and close service operators with much greater flexibility than border crossing operations.

Cetin and List (2004) argues that when "IT systems or resources are shared among various processes or servers, the service times of these processes or servers become correlated." Furthermore, not recognizing such correlations in any type of model development can cause significant discrepancies or inaccuracies in the results. Their numerical examples included: Parallel Servers, Effects of Upstream Servers on Downstream Servers, and Sequential (Tandem) Servers. In all three cases, correlation was identified. Their conclusion supports our claim that although mathematically desirable, a Markovian service time for primary inspections is not realistic.

### 3.2.1 Transient Analysis of POEs

There are many papers dealing with the analysis of non-stationary queueing systems Choudhury et al. (1997), and most of them begin with the Chapman-Kolmogorov forward equations; however, the authors have not found many decision control problems using these formulations.

Margolius (2005) derived an "integral equation for the transient probabilities and expected number in the queue for the multi-server queue with Poisson arrivals, exponential service for time-varying arrival and departure rates, and a time-varying number of servers." The authors used an application of generating functions, but allowed the use of Markovian arrivals and service to develop the probability equations. This is a key element since Gross et al. (2008) explain that transient behavior discussion is most of the time restricted to $M/M/1/1$ and $M/M/1/\infty$, since "the mathematics becomes extremely complicated with the slightest relaxation of Poisson-exponential assumptions."

Other transient analysis of service environments, deal with stochastic supply chain research. Two-stage supply chains have also been the subject of several research papers. Lodree et al. (2004) considers customer response time minimization in a two-stage system facing stochastic demand. The random nature of customers is similar to that of the POEs, and the minimization of service is similar to the optimization inspection time. Their approach is to develop an expected cost function, and use a general demand distribution to find a close form optimal solution. Their results present "significant cost savings under certain assumptions when comparing solutions from the proposed model to the traditional newsvendor order/production quantity."

Kaminsky and Simchi-Levi (2003) researched a two-stage production-transportation model. The model features "capacitated production in two stages, and a fixed cost ... for transporting the product between the stages." But their solution methodology assumes non-speculative assumptions on production and transportation.

## 3.3 Simulation Research for Border Crossing Process Improvements

Simulation is used in a wide variety of applications, and just like the case of many projects or operations that are too expensive to test in a real environment, changes in the border crossing process is also an expensive and secure sensitive proposition, which simulation could serve to evaluate changes without having to commit to such expenses or undesirable results.

The FAST process has been the subject of modeling with some sample scenarios for numerical analysis. By applying Supply Chain Logistics, Chow (2006) did a simulation model based on the mapping of the cross-border transport chain, so that it could be used to identify and quantify strategic and operational choices in both the public and private sector.

In another effort to use simulation for high cost projects and to detect radiation, Nicol et al. (2006) used a simulation border crossing area model to track which vehicles move just when detected radiation changes. This research is also significant for security issues, and to stop inadmissible cargo.

Khoshons et al. (2006) creates a framework for the evaluation of commercial vehicle border pre-inspection systems. In their simulation model, the authors define and account for several measures of effectiveness, but use a case study of a hypothetical pre-inspection system. Their results showed an "increases in the efficiencies of border operations and increases in industry and agency participation."

Another simulation method for decision making is Goal programming, which is a method for multi-attribute decision making in the absence of uncertainty. This topic has been presented in many Operation Research books, such as Winston and Goldberg (2004) and Askin and Standridge (1993). Using goal programming Leung et al. (2006) presented a "preemptive goal programming model for multi-objective cross-border logistics problem, in which three objectives are optimized hierarchically." The model would adjusting the goal priorities, and give decision makers the options to make corrections are necessary.

## 3.4 Variance Reduction at POEs

Variance has been researched significantly in many statistical and quality control related areas. Not to mention that it is an important aspect of the effort put in Lean Manufacturing to make system improvement by reducing variance.

However, variance research in the border crossing process has not been explored significantly because the issue of security trumps any effort to reduce variance from letting the CBP Agents do a thorough job in ascertaining that the vehicle, whether commercial or private, is within the law to enter the U.S. But variance in the inspection process can lead to significant problems as presented by the Bullwhip effect. The bullwhip effect has the characteristic of increasing variability of orders up the supply chain. This effect has been studied in many textbooks including Nahmias (2008). A challenge is always finding ways to mitigate the bullwhip effect.

In a similar government agency environment, e.g. the U.S. Army, and in particular their repair unit, has the characteristic of having a queue of items to service. In this case, Phillips et al. (1999) observed that the operations closely resemble a job shop operation, and used statistics and statistical process control to evaluate the process and propose ways to reduce variance.

## 3.5 Summary

Research in the border crossing process has been the target mostly of public research institutions, which for the most part have focused on improving the process itself, or assessing and implementing some form of technology to make some sort of queue or system improvement. However, basic queueing theory research has only been recently applied to this process, which is in essence a terminating non-stationary queueing process.

The goal is to bring light to the border crossing process, and generate more research and interest into this very significant, economic important, and security

sensitive process. And with homeland security taking on a more prominent role in the nation's economic stability, expanding this research into sea-based, and air ports is the next natural step.

# 4. DYNAMIC REALLOCATION METHODOLOGY

This section introduces a methodology that focuses on improving the performance of the queueing system that forms at commercial truck POEs. Named the Dynamic Reallocation Methodology, or DRM, the methodology tracks the performance of the queueing system and when thresholds levels are met, reallocates resources based on predetermined service or performance levels. The major benefit of the methodology is the improvement of performance measures such as throughput and cycle time, without affecting security procedures or operations costs at the POE. The section begins by describing the POE queueing system and its characteristics. Then, an exposition of the DRM is presented. Finally, empirical data is evaluated and fitted to determine the best distribution for the service time. This section finishes with summary remarks and conclusions.

## 4.1 Introduction

The POE issues described in Section 2 are motivated by the behavior of commercial trucks in their process of crossing the border into the U.S. Wait time varies, and according to Rajbhandari et al. (2009) and Battelle/Texas Transportation Institute (2008), it can be significantly higher than officially reported on CBP's Border Wait Times web site (U.S. Department of Homeland Security, 2011a).

The scope of the discussion in this section will focus on the commercial traffic operations and the queues that are formed for traffic coming into the U.S. Commercial traffic includes all vehicles loaded or empty that enter the U.S. with the intention of importing goods and transporting them inside the country. CBP has implemented a program to help in the documentation of commercial traffic coming into the U.S., the FAST program. This program was discussed in Section 2.4.1. Currently, the POE has a fixed number of inspection stations that serve FAST vehicles, and another set of fixed servers inspect Non-FAST trucks. Given that the trucks will be served by

separate FAST and Non-FAST booths, the average length of the queue lines vary between truck types. To make improvements to the queue, without changing the security procedures, or adding to the cost by increasing the number of booths and agents servicing the trucks, we propose a methodology to implement a dynamic reallocation of servers, or server-line switching. This server reallocation can be activated by a number of policies including time, queue length, wait time, number of customers (trucks) in the system, etc.

## 4.2 System Characteristics of the Queueing Model

There are six basic characteristics of the queueing process, according to Gross et al. (2008), and we will use these to describe the queueing process that forms at the POEs. These characteristics are:

- Arrival pattern of customers
- Service pattern of servers
- Queue discipline
- System capacity
- Number of service channels
- Number of service stages.

### 4.2.1 Arrival Pattern of Customers

The hours of operation within the border crossing process at the BOTA POE were discussed in Section 2.3.1. Having specific hours of operation, that is a finite or natural opening and closing time, makes the process and the consequent model a terminating queueing system (Feldman and Valdez-Flores, 2010). Furthermore,

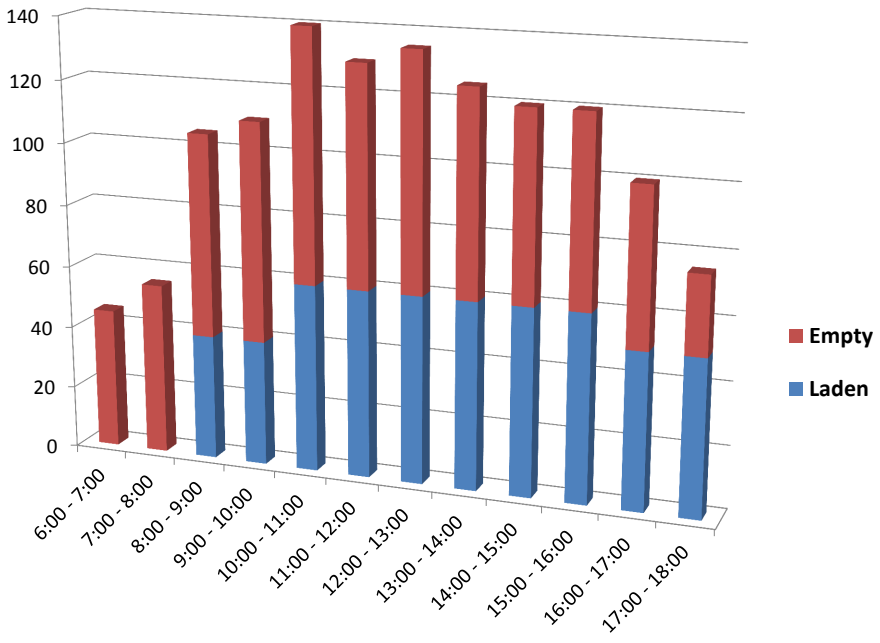



**Fig. 4.1.**: Arrival Rates at BOTA POE (Battelle/Texas Transportation Institute, 2008)

the arrival pattern for this terminating system is both type and time dependent. According to the Battelle/Texas Transportation Institute (2008) report, the arrival rate of empty trucks begins with 45 trucks per hour. Then it experiences an increase until peaking around the middle of the day at 80 trucks per hour, and then decreases to 25 trucks per hour until the system shuts down during the night allowing the queue to empty. See Figure 4.1 for the complete per hour arrival rate data.

Normally trucks that arrive at the POE begin forming a line. But there are two additional events that affect the line in a queue, balking and reneging. Balking occurs when "a customer decides not to enter the queue upon arrival." Furthermore, "a customer may enter the queue, but after a time lose patience and decide to leave." In this case, the customer is said to have reneged (Gross et al., 2008).

For the POE situation, balking is not likely to happen. Shipping companies are paid to transport goods across the border, and it is very unlikely that the driver decides not to get in line because the line is too long. Similarly, reneging is also very unlikely. This is because physically, a truck would have a very difficult time leaving the queue once it has joined. The sheer size of the trucks and the physical limitations of the roads are not conducive to trucks to be able to renege after waiting a while in the queue.

Using the northbound daily commercial crossings data, and assuming there is no balking or reneging, the total hourly number of northbound trucks crossings is an indicator of the arrival pattern because the system is terminating, and there are no trucks allowed to remain in the system over night. This arrival pattern will be used to determine a time-dependent exponential arrival process denoted by $\lambda(t)$. Also notice in Figure 4.1, that the arrival rates are classified by whether the truck is carrying a full or empty load. This distinction is denoted by $\iota$, with $\iota = \{empty, laden\}$. Finally, recall from Section 2.4.1 that U.S. bound commercial trucks have the option to participate in the FAST program. The additional index $\wp$ differentiates FAST program participation. The index $\wp = 1$ whenever the commercial truck, the load and the driver are participants of the FAST program, or simply FAST truck. On the other hand, $\wp = 2$ when any one of them is not certified as FAST program participants, or Non-FAST trucks. Notation wise, $\lambda_{\iota,\wp}(t)$ represents the time dependent arrival rate that is separated by truck load, and FAST program participation.

### 4.2.2 Service Pattern

To evaluate the service pattern, data was captured from February 2010 to April 2010 in BOTA by observing the service times by type and load. Considering that data collection is vital for validating a model, a letter to request the cooperation of DHS and TxDOT authorities in such a security related area was prepared and delivered. Please refer to Appendix A for the "Data gathering for scholarly dissertation" letter.

Trucks were classified into four categories for data collection:

- Laden FAST
- Empty FAST
- Laden Non-FAST
- Empty Non-FAST.

Even though DHS did not allow large data collection citing security concerns. Field observations yielded over 200 data points and at least 50 data points for each category. Please refer to Appendix B which has the raw data collected from the field observations, and to Section 4.4 for the data fitting analysis of the empirical observations.

Although, security procedures such as dog sniffing create dependencies, the use of Markovian service time to approximate the general distribution has been used by many, including the POE work by Zhang (2009). However, other authors started to use technology in order to capture the distribution of the wait times. For example the research done by McCord et al. (2010), who attempt to show a proof-of-concept in using technology as an approach for "an ongoing tool for collecting truck activity times at international crossings." Their hope is to be able to use technology to report "activity time data in indicating the distribution of direct crossing times."

Notation wise, the service times are denoted by $\mu_{(\iota,\wp)}$ with the same indexes $\iota$, representing load type, and $\wp$ describing FAST program participation. For our analysis, steady-state analysis cannot be used because the system rarely if ever becomes stationary, and terminates daily. In Section 5, we explore in detail the application and use of the Coxian 2-phase distribution as an approximation for the general distribution, as described by Curry and Feldman (2011) and Altiok (1996).

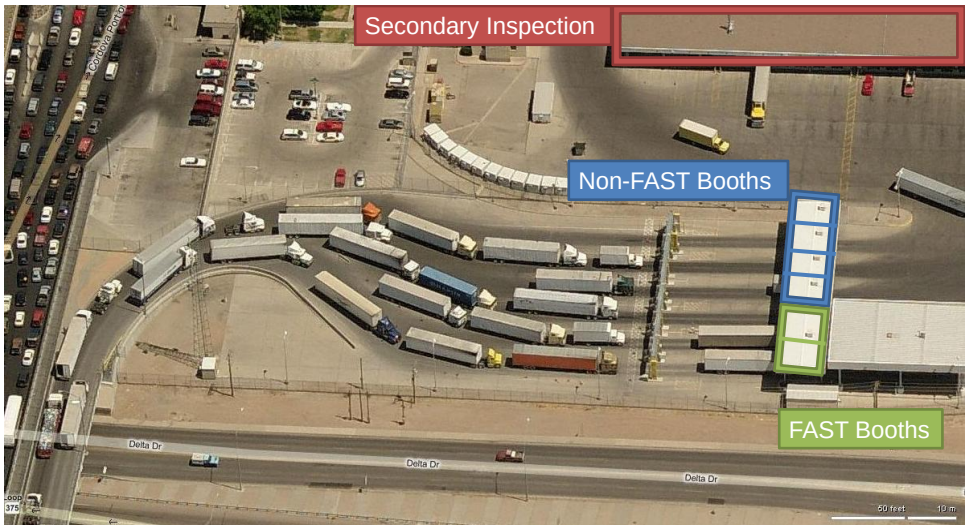



**Fig. 4.2.**: Inspection Booths by Type at BOTA POE (Bing Maps, 2011)

### 4.2.3 Queue Discipline and System Capacity

The queue discipline for all POEs is FIFO (first-in, first-out). That is, each truck arrives and joins in the queue, and when they finally reach the service area they are served at the inspection booth in the order they arrived. In other queueing systems there are two general service situations in priority disciples. These are *preemptive*, and *non-preemptive* cases. The preemptive case implies that different customers have different priorities Gross et al. (2008). However, at POEs the general priority case is non-preemptive because it is physically impossible to pick a truck from the line and have it inspected.

Observe from Figure 4.2, that there is a maximum service capacity of six inspection booths or stations, but not all are opened all the time. The decision to open or close each inspection station is a decision made by CBP, and their operations manager. Also, note from Figure 2.4 that the queues do not seem to have a limit, since the lines can, and have been known to go for miles down the road. Therefore, there is no limit in the waiting line or queue.

### 4.2.4 Number of Service Channels and Stages

The number of service channels refers to the number of parallel inspection booths that can service trucks simultaneously. Inspection Stations, just like commercial trucks are classified by participation in the FAST program. Also observe from Figure 4.2 that in the case of BOTA, there are normally four fixed Non-FAST inspection booths and two fixed FAST inspection booths in operation. Currently, by geographic and physical limitations, there is no possibility of expansion, unless a major construction project is undertaken. Hence, the number of service channels or inspection stations for FAST is two, and for Non-FAST is four.

Notation wise, the number of inspection stations, or capacity, is represented by $K_{\wp}$. Recall that the index $\wp = 1$ whenever the inspection station / commercial truck type is a participant of the FAST program, and $\wp = 2$ otherwise, that is servicing Non-FAST trucks. Since the DRM allows for servers to be reallocated, the number of inspection stations serving FAST trucks is either one, two or three at any point in time, that is $K_{\wp=1} = \{1, 2, 3\}$. And the number of servers inspecting $Non = FAST$ trucks is either three, four or five. In notation it is $K_{\wp=2} = \{3, 4, 5\}$.

Using the notation, the current situation at BOTA has $K_1 = 2$ and $K_2 = 4$. Adding more possible values for $K_{\wp}$ is necessary because the methodology allows for dynamic reallocations of servers. In the case of BOTA, two moveable servers are allowed with one originally serving FAST trucks, and the other originally serving Non-FAST trucks. Therefore with DRM, there is at least one FAST inspection booth open and at most three; and similarly, there are at least three Non-FAST and at most five Non-FAST inspection stations during normal hours of operation.

Because the current situation at BOTA is that the number of servers is finite and fixed, when the dynamic reallocation of servers takes place, the rule is that sum of the number of servers open must be less than the maximum physical number of

available servers. Such that when there is a reallocation, one type of server gains one server, and the other type looses a server. That is,

$$\sum_{\wp} K_{\wp} \leq MaxServers \tag{4.1}$$

where $MaxServers$ is the maximum number of servers for FAST and non-FAST, and in the case of BOTA $MaxServers = 6$, from Figure 2.5.

The number of service stages for the inspection process at the POE is two. Since there is the possibility of going to secondary inspection from the primary inspection booth. However, in the scope of this research, going to secondary inspection is of no consequence to our analysis. This is because at that time, and for all intents and purposes, the DHS Agent, or inspection station server, has finished inspecting the truck, and is ready for the next vehicle to approach the booth for inspection. Therefore, the number of service stages is omitted from the summary queue description.

### 4.2.5 Summary

The BOTA POE queueing system has distinct queueing characteristics. And for all queueing descriptions, the index $\iota$ represents load type by $\iota = \{laden, empty\}$. The index $\wp$, is used for FAST program participation for both the commercial trucks and the inspection stations. The value $\wp = 1$ represents FAST program participation and $\wp = 2$ is Non-FAST, which means that the truck or inspection station is not dealing with FAST paperwork.

In summary, the BOTA POE characteristics are:

- *Arrival Pattern:* Markovian arrival rate ($M_{t,\iota,\wp}$), but dependent on time $t$, the load type $\iota$, and whether the truck, load and driver participate in the FAST program $\wp$.

- *Service Pattern:* General service time ($G_{\iota,\wp}$) dependent on load $\iota$ and whether the inspection station services trucks that participate in the FAST program $\wp$.

- *Queue Discipline:* FIFO (First-In-First-Out).
- *System Capacity:* There is no real limit on the system capacity, thus, it is infinity.
- *Number of Service Channels:* Finite capacity and dependent on load ($K_\iota$). Where $K_1 = \{1, 2, 3\}$ and $K_2 = \{3, 4, 5\}$, according to the limitations of Condition 4.1.

Thus, the border crossing process can be described as a combination of two terminating, non-stationary queueing systems, which can be represented in Kendall's notation as

$$[\ (M_{(t,\iota,1)}\ /\ G_{(\iota,1)}\ /\ K_1)\ ,\ (M_{(t,\iota,2)}\ /\ G_{(\iota,2)}\ /\ K_2)\ ] \tag{4.2}$$

## 4.3 Dynamic Reallocation Methodology of Servers

As stated before, one of the contributions of this dissertation is the DRM. Recall as well, that the system is non-stationary and the servers are fixed in the sense that they only serve one type of truck. In this section the POE dynamics are described. Then, the DRM is characterized according to the non-stationary trigger or policy for a server switch, and the number of fixed and moveable servers in the system.

### 4.3.1 POE Process Dynamics

Recall from the description of the FAST process in Section 2.4.1 and Figure 2.5 that there are two main queues (FAST and Non-FAST) feeding several fixed inspections stations, or booths. And that these two queues feed short queues of up to five commercial trucks (customers) in front of each inspection station (server).

In order for a reallocation of a server to take place, the small queue in from of the inspection station must be allowed to empty. This is because, the servers are

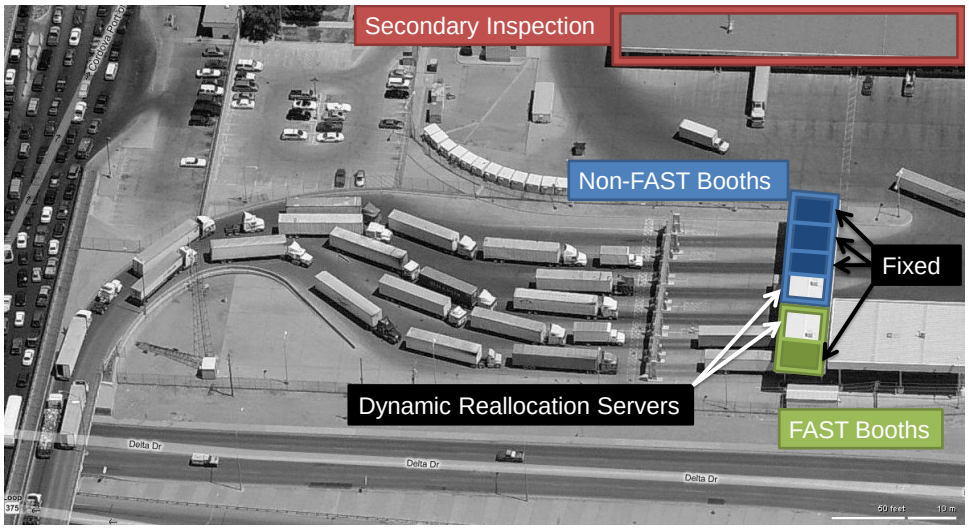



**Fig. 4.3.**: Dynamic Reallocation Inspection Booths at BOTA POE (Bing Maps, 2011)

dedicated, and thus, service only one type of truck at a time, either FAST or Non-FAST. Furthermore, because the trucks cannot easily maneuver between lanes, all FAST lanes must be adjacent to each other, and therefore all Non-FAST lanes should be adjacent to each other.

Therefore, the two servers that will no longer be fixed and subject to reallocation are, in the case of BOTA and from Figure 4.3, counting from top to bottom, server four and server five.

### 4.3.2 Dynamic Reallocation

The DRM incorporates servers, in this practical case, a CBP Agent at the primary inspection station booth of the POE, that is no longer fixed, and can be switched as an inspection station between FAST and Non-FAST trucks. So for an event where

there is a reallocation of a server from FAST to Non-FAST truck types, the capacity of servers would change by

$$\begin{aligned} K_1 &\Leftarrow K_1 - 1 \\ K_2 &\Leftarrow K_2 + 1 \end{aligned} \tag{4.3}$$

subject to the restrictions in Equation 5.1, and the allowed values of $K_1$ and $K_2$ in Condition 4.1. However, once the decision is made to switch a booth, it cannot be implemented until the small queue in front of the station being switched has been emptied. Once the local server queue is flushed, the Agent and POE can be switched to serve the other type of customer, noting that the time to switch is ignored.

Similarly, for a reallocation of servers from Non-FAST to FAST service, the capacity of servers would be

$$\begin{aligned} K_1 &\Leftarrow K_1 + 1 \\ K_2 &\Leftarrow K_2 - 1. \end{aligned} \tag{4.4}$$

In the two-customer type system of the commercial border crossing system, performing a dynamic reallocation or "server switch" takes advantage of the flexibility of moving a server to a more congested area. The Battelle/Texas Transportation Institute (2008) current state analysis report, shows that the BOTA POE has capacity for 120 trucks per hour, and from the stake holders meeting in the report, there is an expressed interest in reducing regular (Non-FAST) wait time to within an hour, and FAST processing to within 15 minutes. Therefore, we can develop a non-stationary dynamic reallocation policy in terms of a queue differential.

Thus, the non-stationary reallocation policy is a function $\delta(t)$ such that:

$$\delta(t) = \begin{cases} \text{F2R}, & QS_t(NonFAST) \ > QS_t(FAST) + Th_{F2R} \\ \text{R2F}, & QS_t(FAST) \ > QS_t(NonFAST) + Th_{R2F} \\ \text{No change}, & o.w. \end{cases} \tag{4.5}$$

where $QS_t$, represents the size of the queue at time $t$. $F2R$ is a call to dynamically reallocate a server from FAST to Non-FAST (regular); and $R2F$ is a call to dynamically reallocate a server from Non-FAST (regular) to FAST, for the allowed values of $K_1$ and $K_2$ in Formulation 5.1.

## 4.4 Data Fitting

Now, to determine whether the Coxian approximation is needed, this section conducts an evaluation of the field data gathered for this research in Appendix B, in order to find the best fitting distribution of the service time data. The data fitting was accomplished using the "Input Analyzer" tool of Arena's simulation software, version 12.000.00 - CPR 9, from Rockwell Automation Technologies. The results of the analyzer is printed verbatim following the data plot and corresponding best fitted distribution. The results include a distribution summary with the software selected best fit distribution.

There are also two measures of the distribution's fit to the data, the Chi Square and the Kolmogorov-Smirnov goodness of fit test. Kelton et al. (2002) mentions that these are the standard statistical hypothesis tests used to evaluate "whether a fitted theoretical distribution is a good fit to the data." In the following data fit calculations, the corresponding $p$-value which will always fall between 0 and 1, is the leading indicator of a distribution fit. The larger the $p$-value, the better the distribution "fits" the data analyzed. The authors also mention that corresponding $p$-values of less than 0.05 indicate that the distribution is not a very good fit. With the understanding that a "high $p$-value doesn't constitute 'proof' of a good fit - just a lack of evidence against it." When the data does not closely follow a distribution function, then the option of using a MGE or Coxian to approximate an empirical or general distribution is a favorable one. And from the research work of Curry and Feldman (2011), and Altiok (1996), calculations can determine the number of service phases needed for a general approximation of the service time for each type of truck and load combination.

Recall that in Section 4.2, that the POE inspects two types of commercial trucks. And to evaluate the service pattern, data was captured from February 2010 to April 2010 in BOTA by observing the service times by type and load. Over 200 data points of truck service time were classified into four categories:

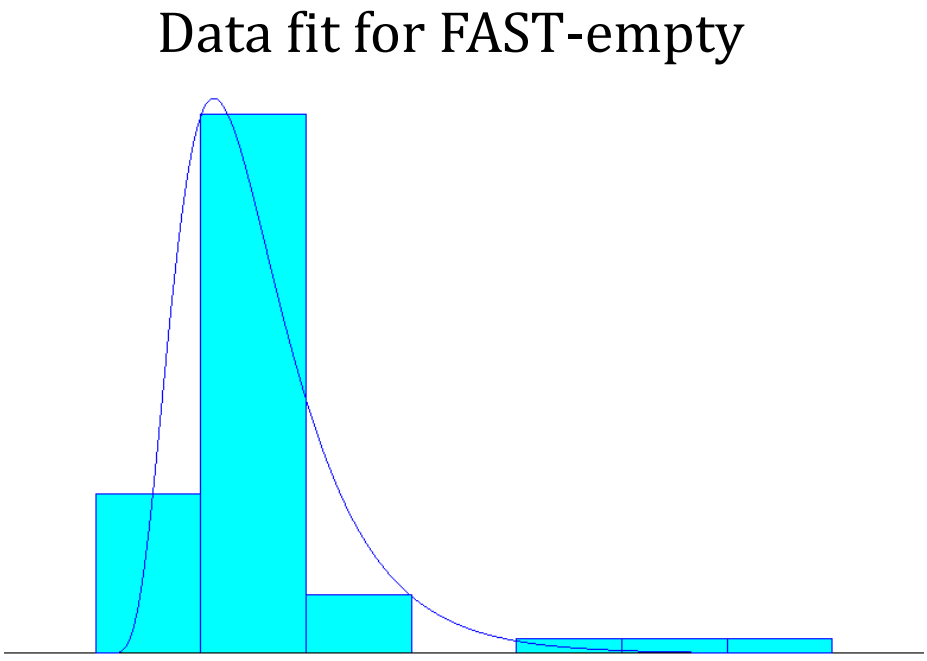



**Fig. 4.4.**: Data Fit Histogram and Distribution for FAST-empty Trucks

1. FAST truck and empty cargo or no cargo bed: FAST-empty
2. FAST truck with a loaded or partially loaded cargo: FAST-laden
3. Non-FAST truck and empty cargo or no cargo bed: Non-FAST-empty
4. Non-FAST truck with a loaded or partially loaded cargo: Non-FAST-laden.

Refer to Appendix B for the empirically observed data. The four cases, FAST-empty, FAST-loaded, Non-FAST-empty, Non-FAST-loaded, show the best fit distribution according to the "Input Analyzer" software. Results also include the corresponding $p$-value with data and histogram summaries.

### 4.4.1 Data Fit Case 1: FAST Truck - Empty Load

The first case is the one with the fastest service time data. FAST trucks with no load should have minimal inspection time, given that the truck, the driver and the load are all part of FAST, and in effect there is nothing to check other than the immigration papers of the driver.

With 55 data points, observe in this case from Figure 4.4 and the distribution summary below that the best distribution is a logNormal. In this case, the $p$-value for the Chi Square test is $< 0.005$, and would question the use of this distribution.

The Input Analyzer results for FAST Truck - Empty Load data:

```
Distribution Summary

Distribution: Lognormal
Expression: 0.14 + LOGN(0, 0)
Square Error: 0.028213

Chi Square Test
  Number of intervals = 3
  Degrees of freedom  = 0
  Test Statistic      = 6.77
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic = 0.155
  Corresponding p-value = 0.129


Data Summary

Number of Data Points = 55
Min Data Value        = 0.722
Max Data Value        = 6.54
Sample Mean           = 1.69
Sample Std Dev        = 1.07


Histogram Summary

Histogram Range     = 0.14 to 7
Number of Intervals = 7
```

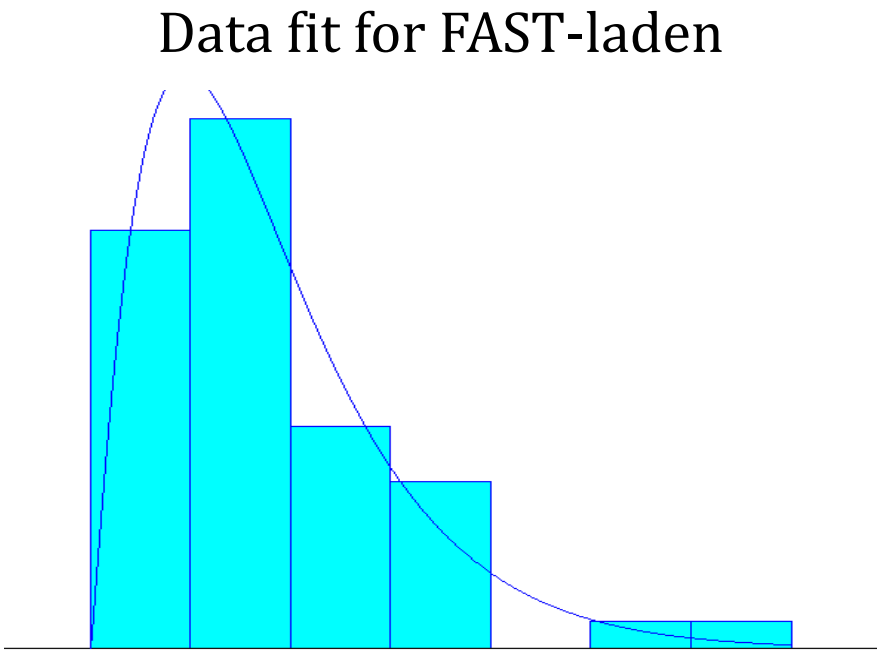



**Fig. 4.5.**: Data Fit Histogram and Distribution for FAST-laden Trucks

### 4.4.2 Data Fit Case 2: FAST Truck - Laden

With 50 data points, and Figure 4.5, the best distribution for FAST loaded trucks is a shifted Gamma distribution. This distribution has a Chi Square test $p$-value of 0.47. But the Kolmogorov-Smirnov Test only shows a corresponding $p$-value $> 0.15$.

The Input Analyzer results for FAST truck - Laden truck data:

```
Distribution Summary

Distribution: Gamma
Expression: 1 + GAMM(0, 0)
Square Error: 0.004561

Chi Square Test
  Number of intervals = 4
  Degrees of freedom  = 1
  Test Statistic      = 0.559
  Corresponding p-value = 0.47

Kolmogorov-Smirnov Test
  Test Statistic = 0.0778
```

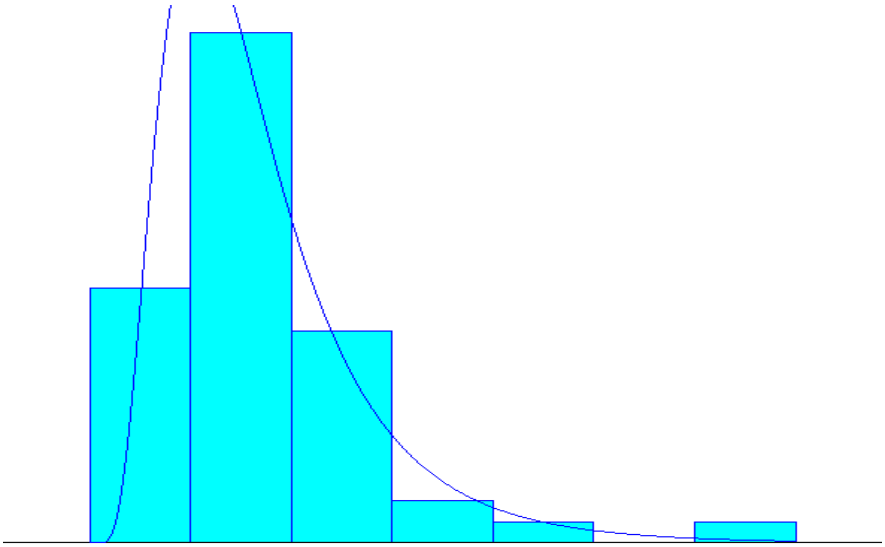



**Fig. 4.6.**: Data Fit Histogram and Distribution for Non-FAST-empty Trucks

```
  Corresponding p-value > 0.15


Data Summary

Number of Data Points = 50
Min Data Value        = 1.25
Max Data Value        = 8.24
Sample Mean           = 3.02
Sample Std Dev        = 1.49


Histogram Summary

Histogram Range     = 1 to 8.94
Number of Intervals = 7
```

### 4.4.3 Data Fit Case 3: Non-FAST Truck - Empty Load

In this case we have 50 data points. Figure 4.6 and the distribution summary below that the best distribution is a logNormal. Similar to the case of the FAST Truck - Empty Load data, the Non-Fast Truck - Empty Load data has a $p$-value for the Chi Square test of $< 0.005$, which would question the use of this distribution.

The Input Analyzer results for Non-FAST Truck - Empty Load data:

```
Distribution Summary

Distribution: Lognormal
Expression: LOGN(0, 0)
Square Error: 0.001272

Chi Square Test
  Number of intervals = 3
  Degrees of freedom  = 0
  Test Statistic      = 0.104
  Corresponding p-value < 0.005

Kolmogorov-Smirnov Test
  Test Statistic = 0.062
  Corresponding p-value > 0.15


Data Summary

Number of Data Points = 50
Min Data Value        = 0.751
Max Data Value        = 10.5
Sample Mean           = 2.66
Sample Std Dev        = 1.7


Histogram Summary

Histogram Range     = 0 to 11
Number of Intervals = 7
```

### 4.4.4 Data Fit Case 4: Non-FAST Truck - Laden

In the last case we have 53 data points. Figure 4.7 and the distribution summary below that the best distribution is a shifted Weibull. In this case the Chi Square test has a p-value of 0.047, even though the Kolmogorov-Smirnov Test shows a

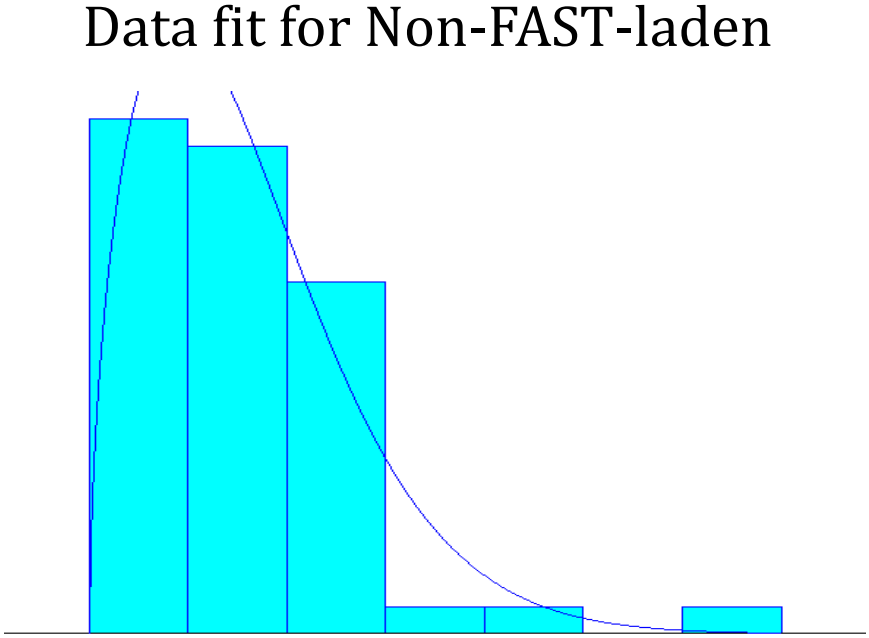



**Fig. 4.7.**: Data Fit Histogram and Distribution for Non-FAST-laden Trucks

corresponding $p$-value $> 0.15$. The discrepancy of $p$-values leads to questions in the use of the distribution for this data.

The Input Analyzer results for Non-FAST truck - Laden truck data:

```
Distribution Summary

Distribution: Weibull
Expression: 1 + WEIB(0, 0)
Square Error: 0.006982

Chi Square Test
  Number of intervals = 4
  Degrees of freedom  = 1
  Test Statistic      = 4.22
  Corresponding p-value = 0.0421

Kolmogorov-Smirnov Test
  Test Statistic = 0.0889
  Corresponding p-value > 0.15


Data Summary
```

```
Number of Data Points = 53
Min Data Value        = 1.1
Max Data Value        = 13.1
Sample Mean           = 3.95
Sample Std Dev        = 2.07


Histogram Summary

Histogram Range     = 1 to 14
Number of Intervals = 7
```

After the data fitting, three of the four cases did not produce an adequate $p$-values for the best fitted distribution. All $p$-vales were less than 0.05 except for the $p$-values for FAST-laden trucks, which was 0.47. This can be explained by the lack of outliers for this data type. With the empirical data set obtained, and the results of the data fitting, the use of a general distribution seems to be best distribution to describe the inspection or service times. The results also bring to light an important aspect of the service time data. There are some occasions where the inspection time takes significantly longer than normal. Data collection will include these occasional outliers which affect the fitting of a distribution. In Appendix B, the data set had an outlier which in one occasion was seven times longer than average. Since these outliers are significantly longer, the Input Analyzer cannot find a best fit distribution that accounts for these outliers with a high $p$-value.

## 4.5 Summary and Conclusions

This section presented the POE in terms of a queueing system, and characterized the DRM for use in a POE environment. The DRM is a proposed methodology that can be used in certain saturated POE systems where adding additional servers is not easily achieved. The methodology was developed considering the difficulty of adding additional resources and from the daily observations of the POEs with long queue lengths and excessive time spent waiting to transport goods across the border. After analyzing the POE in terms a combination of two queueing systems,

the methodology was presented as a viable method to improve performance measures such as throughout without affecting cost or security. Finally, empirical data was analyzed to better understand the service time parameters of the system, but the Input Analyzer was not able to find a best fit distribution with a high $p$-value and accounts for data outliers. Therefore, the conclusion is that an empirical distribution or a general service time distribution would more accurately describe the inspection service time.

While the focus of this research is on commercial traffic, and with the POE described in terms of queueing theory, the methodology can be implemented in any land-based POE, and also analyze the characteristics of traffic by privately owned vehicles or pedestrians. In addition, the methodology can also apply at other POEs. With some modification, the methodology and acceptable control policy, can also be considered for sea ports, given that the traffic is mainly container-based.

The discussion now turns to an analysis of the POE. First in Section 5, where for mathematical simplification, the POE is assumed to have a Markovian service time, and a simplified case of the DRM is presented. Later in Section 6, the service time is considered to be a general distribution, and applies the Coxian k-phase approximation as a better approximation of service time with an analytical and a simulated model of the POE queue.

# 5. ANALYSIS OF THE POE QUEUE I: DRM WITH A MARKOVIAN SERVICE TIME ASSUMPTION

Section 4 described the border crossing process as a queueing model, and detailed the Dynamic Reallocation Methodology (DRM). In this section, the research turns to analyzing the effects of the DRM on POE queues, assuming a Markovian service time distribution. This assumption implies that the inspection time of each truck will be considered independent of other truck inspection times.

Starting with some of the key characteristics of the POE queue, this section presents supporting arguments for assuming service (inspection) time independence. The next section presents and analytically solves a simplified case of the DRM with only one server per type and one moveable server. Afterwards the DRM model incorporates a Markovian service time assumption, and presents analysis of the results. Finally, the section concludes with highlights of the effects of the Markovian assumption and the effectiveness of the DRM with exponential service times, on the commercial border crossing process.

## 5.1 Introduction

Recall from Section 2 and Figure 2.3, that the POE is configured as a multi-commodity, prioritized queueing network which rarely, if ever, operates in steady-state. The commercial border crossing process is in its essence, a basic queueing process. It is composed of trucks and other commercial vehicles that form queues based on cargo and the type of documentation for CBP Agents. The line is then serviced by inspection booths. With the help of technology, CBP Agents check the documentation of the vehicle, and determine whether the vehicle and/or the driver need further inspection. To address long lines, CBP has implemented the FAST program. The FAST program allows for expedited processing of certain commer-

cial carriers under special conditions. Please refer to Section 2.4.1 for a detailed description of the FAST program.

Using exponential service times is widespread in papers dealing with the border crossing process, as discussed in Section 3.2. Haughton and Isotupa (2012) researched using "computer simulation study to predict the likely impacts of smoothing" with exponential arrivals and service times. Whitt (2007) also used approximation methods to help set the staffing requirements in service systems he was researching. His queueing model was a $M_t\ /GI\ /s_t + GI$. Their work is similar to this research of POEs in that the "model is difficult to analyze mathematically, so that the staffing problem is challenging. However, there is one special case that is amazingly tractable: the Markovian $M_t\ /M\ /s_t + M$ model in which $\theta = \mu$."

Furthermore, the non-stationary nature of the commercial truck arrival process to the POE, plus the hours of operations, is conducive to a study of the transient states of the system. Therefore, a Markovian assumption for service time is desirable mathematically, and would still provide insight to the processes. However, the queues of commercial vehicles vary per POE in hours of operation, congestion, capacity and other aspects such that the analysis of each border crossing process could be considered unique. In effect, the truck waiting line is fundamentally a queueing environment that does not behave in a way where the system can stabilize, or reach steady-state. In Section 6, a Coxian approximation for a general service time is discussed.

## 5.2 Markovian Service Time Assumption Analysis

This section proposes a DRM with the assumption that the inspection time, i.e. service time, is Markovian in nature. From Section 2.4, the border crossing delay problem is described as a combination of two terminating, non-stationary queueing system that can have very long queues. Each queueing system is formed by a truck type (FAST or Non-FAST) that can be laden or empty. So each queueing system

consists of a non-stationary arrival process of a single truck type that feeds multiple small server queues. Furthermore, each inspection station is dedicated to a truck type, and there is no overlap, i.e. one inspection station cannot service another truck type. Recall that Equation 4.2 describes the current queueing system of the border crossing. With a Markovian service time assumption, in Kendall's notation, the two terminating, non-stationary queueing systems can be expressed as

$$[\ (M_{(t,\iota,1)}\ /\ M_{(\iota,1)}\ /\ K_1)\ ,\ (M_{(t,\iota,2)}\ /\ M_{(\iota,2)}\ /\ K_2)\ ] \tag{5.1}$$

where $t$ denotes time dependency, $\iota$ reflects if the truck is loaded, and the last index separates the queues by FAST ($\wp = 1$) and Non-FAST ($\wp = 2$).

The Markovian assumption is not arbitrary or without support. There are common and significant characteristics of all POEs that should be considered when modeling and analyzing the system with a Markovian service time assumption. Particularly, the following POE characteristics support the assumption of Markovian service time:

- Lack of complete data on the service times.
- Previous research work in pedestrian POE was based on a Markovian service time, as noted in the research of Zhang (2009).
- Trucks entering the system follow a random arrival process, implying no relation between trucks.

These characteristics support the assumption that the service time can be considered Markovian. However in their work, Cetin and List (2004) mention that the service time should be not be considered independent since correlation exists, and "failure to recognize such correlations in model development may lead to significant inaccuracies." More detail is discussed in Section 6, and the data collected in Appendix B was analyzed to find the best-fit distribution in Section 4.4.

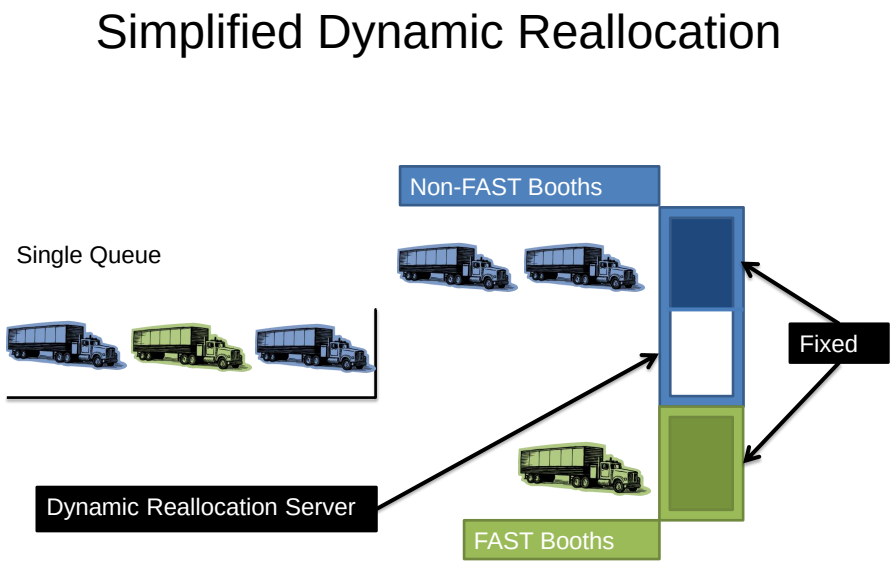



**Fig. 5.1.**: Simplified Border Crossing DRM Case Model

This section makes the Markovian assumption to facilitate the analysis, and to show that even with a simplification of the problem, and exponential service times, the POE with a DRM model becomes a complex system difficult to analyze. And regardless of complexity, an analytical approach brings understanding of the problem, and intuition to its behavior. In the following section the research turns to developing an analytical solution to the basic queueing system with DRM. This analysis will employ traffic intensity $\rho$, to assess the utilization of the system.

### 5.2.1 A Simplified DRM Base Case System

For a simplified queueing case with DRM, consider a non-stationary arrival process of two-customer types to a three-server queueing system, as depicted in Figure 5.1. Also assume that there is only one queue feeding short queues of up to five cus-

tomers, in front of each server. The outside inspectors serve a different customer type and the middle server is switchable, hence there are three servers in total. The first server can only handle FAST customers, the third server can only handle Non-FAST customers, but the middle, or second server can be switched between customer types as necessary. The control question is to determine which type of customer should be served by the second server as a function of the number of customers of each type within the system.

Recall that trucks were classified into four categories for data collection:

- Laden FAST
- Empty FAST
- Laden Non-FAST
- Empty Non-FAST.

If an inspector is dealing with FAST trucks, then only the FAST-approved vehicles will be in the short queue in front of the inspection station. Similarly to the DRM policy of Section 4, if there is an unusual buildup of Non-FAST vehicles, then a decision may be made to switch an inspector from FAST to Non-FAST truck types. However, once the decision is made, the server switch can not be implemented until the small queue in front of the station being switch has been emptied.

## 5.3 Analysis of the Simplified DRM Using Traffic Intensity

From Figure 5.2 and Equation 5.1, the arrival process is a non-stationary arrival process with the two arrival streams having mean time-dependent rate functions of $\lambda_{(\iota,1)}(t)$ and $\lambda_{(\iota,2)}(t)$. In this case, $\lambda_{(\iota,\wp)}(t)$ represents the mean non-stationary arrival rate at time $t$ for type $\wp$ customer and $\iota$ load, with indices $\wp = \{1, 2\}$ and $\iota = \{empty, laden\}$.

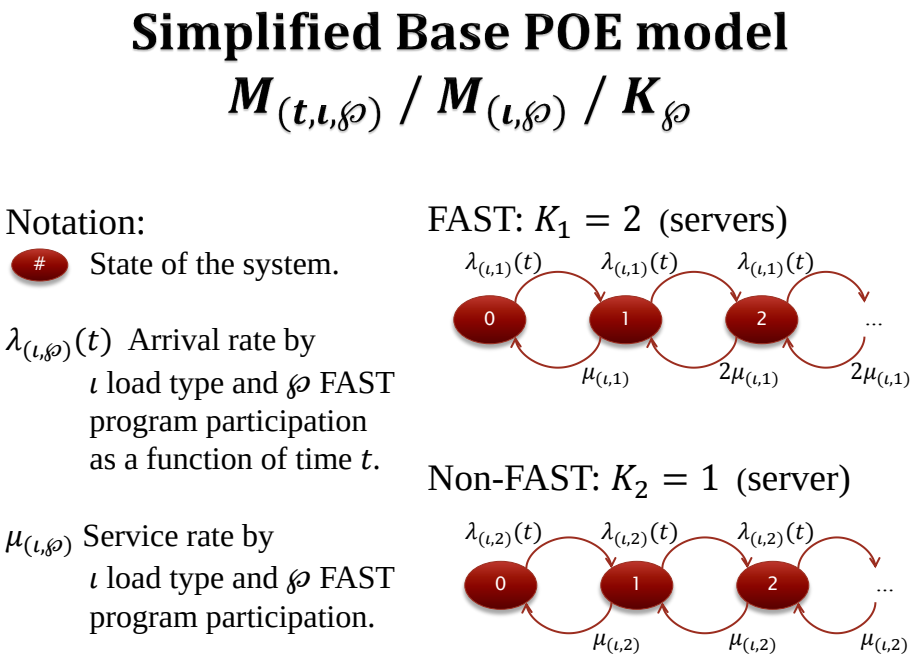



**Fig. 5.2.**: State Transition Diagram for a Simplified $M_{(t,\iota,\wp)}$ /$M_{(\iota,\wp)}$ /$K_{\wp}$ POE

One analytical approach to the model and a mathematical solution, follow the assumption that each customer type is independent of each other. Furthermore, the service rates and arrival rates, i.e. $\mu_{(\iota,1)}$ and $\lambda_{(\iota,1)}(t)$ have no relationship with the service or arrival rates of customer Type 2, that is $\mu_{(\iota,2)}$ and $\lambda_{(\iota,2)}(t)$, or with each other. Therefore, the Markovian assumption is justified by the lack of information and by the perceived "randomness" of the time it takes to inspect or service each truck at the POE.

In this simplified DRM queue model, all possible cases in the system can be enumerated, and evaluated independently. With only two customer types $\wp$, the DRM moveable server can begin in only one of two cases initial cases, either servicing FAST truck or servicing NON-FAST truck. And since the DRM server will end in one of two states after some time $t$, there are two ending cases for each initial situation.

Therefore, enumerating all possible scenarios from beginning to end, there are four possible cases:

1. The DRM server starts serving FAST customers $\wp = 1$, and ends serving FAST customers $\wp = 1$.

2. The DRM server starts serving FAST customers $\wp = 1$, and ends serving Non-FAST customers $\wp = 2$.

3. The DRM server starts serving Non-FAST customers $\wp = 2$, and ends serving FAST customers $\wp = 1$.

4. The DRM server starts serving Non-FAST customers $\wp = 2$, and ends serving Non-FAST customers $\wp = 2$.

All four cases can be solved by utilizing the traffic intensity $\rho$ as an indicator of where will the moveable server switch to, once the threshold is met. $\rho$, also known as utilization, is interpreted as the proportion of time that each server is busy or as the expected number of customers in service (Nahmias, 2008). Then if $\lambda$ and $\mu$ are independent of the number of customers in the system, then the utilization factor can be calculated as

$$\rho = \frac{\lambda}{c \times \mu}$$

given that $c$ is the number of identical servers. Normally, $\rho$ is bounded between 0 and 1, and $\rho < 1$ ensures that the queue does not grow to infinity. But $\rho$ can be greater than 1 for a short period, which is particularly the case for some POEs and some time dependent systems.

### 5.3.1 Arrival Rate Data for Simplified DRM

Recall from research by Battelle/Texas Transportation Institute (2008) and Figure 4.1 that in the case of the BOTA POE, the arrival data is time dependent and

is separated whether the truck is loaded or laden, and $\lambda_{(\iota,\wp)}(t)$ represents the arrival data per hour by load type $\iota$. The additional index denotes whether the truck participates in the FAST program, as described in Section 2.4.1. In the Simplified DRM case, the mean arrival rate is needed to obtain traffic intensity and compare $\rho$'s. Therefore, the first calculation

$$\overline{\lambda}_{(\iota,\wp)} = \frac{\sum^{n} \lambda_{(\iota,\wp)}(t)}{n}$$

is the arithmetic mean of the time dependent $\lambda_{(\iota,\wp)}(t)$ arrival rates.

According to Zietsman et al. (2006), and the work of Battelle/Texas Transportation Institute (2008), the research shows that the percentage of trucks, drivers and loads that participate in the FAST program is currently about 15%. Thus the values are

$$\overline{\lambda}_{(\iota,1)} = 0.15 \times \overline{\lambda}_{(\iota,\wp)}$$

and

$$\overline{\lambda}_{(\iota,2)} = 0.85 \times \overline{\lambda}_{(\iota,\wp)}$$

where $\overline{\lambda}_{(\iota,\wp)}$ is now used in calculating the traffic intensity $\rho$ by FAST or Non-FAST truck participation.

### 5.3.2 Comparison Approach for Simplified DRM

As presented in Figure 5.2, the mean service rate depends on the customer type and the type of load that the truck is carrying, which will be denoted by $\mu_{(\iota,\wp)}$, with $\wp = (1, 2)$ for FAST and Non-FAST, and load $\iota = (empty, laden)$. The data in Appendix B, is separated by load type and customer type. And the arithmetic mean of the data by truck and load type will be used as the exponential service time for each type of truck with load in the system. Notation wise, there is data to calculate

the mean for $\mu_{(laden,1)}$, $\mu_{(empty,1)}$, $\mu_{(laden,2)}$ and $\mu_{(empty,2)}$. Hence, the comparative equation for server utilization or traffic intensity becomes

$$\rho_{\wp} = \sum^{\wp} \frac{\overline{\lambda}_{(\iota,\wp)}}{c_{\wp} \times \mu_{(\iota,\wp)}} \tag{5.2}$$

where $c_{\wp}$ is the number of identical servers for type $\wp$ customers.

In summary, the Simplified DRM model has two initial possible conditions for where the switchable sever starts, and two options for where it ends. Therefore, four possible situations. The analysis of the Simplified DRM model involves an evaluation of the traffic intensity for both initial conditions when a reallocation decision point has been reached. Once $\rho_{\wp}$'s are calculated, then the initial condition with a higher $\rho_{\wp}$ will attract the switchable server. The decision criteria is explained in Proposition 5.3.1 as follows.

**Proposition 5.3.1** *In a simplified DRM environment, the maximum of the $\rho_{\wp}$'s will attract the switchable server once the reallocation condition $\chi_{\wp}$, has been met.*

**Proof** In a simplified DRM, let the reallocation condition $\chi_{\wp}$ be to call a switch server when

$$\chi_{\wp}\{Q_{\wp}(t) \geq x\} = true$$

where $Q_{\wp}(t)$ is the queue length at time $t$.

Suppose by contradiction that the DRM allocates the switchable server to the $\wp$ customer type with the smallest traffic intensity, called $\rho_{min}$. If $\wp \neq min$ customer type, then the queue length $Q_{\wp}$ will not decrease since $\rho_{\wp} = \frac{\lambda}{c_{\wp} \times \mu}$ because it depends on $c_{\wp}$, the number of identical servers. And the number of servers for $c_{\wp}$ will only decrease or stay the same. Therefore the reallocation condition $\chi_{\wp}$ will not be relieved, continue to call for a reallocation indefinitely, and thus a contradiction. ■

### 5.3.3 Numerical Solution for the Simplified DRM Case

The data parameters are calculated using the arithmetic mean of the collected data from Appendix B. Therefore the parameters are as follow: s will be used as the exponential service time for each type of truck in the system. Recall that notation wise, the arrival rates will be denoted as $\overline{\lambda}_{\iota,\wp}$ and service times will be denoted by $\mu_{(\iota,\wp)}$ as defined in section 5.3. Using the data from Appendix B, Table 5.1 presents a summary of the calculated values used for the DRM simplified case.

**Table 5.1**: Data used for the simplified base DRM case (in minutes).

| | **Type 1 FAST** | | **Type 2 Non-FAST** | |
|---|---|---|---|---|
| | Empty | Laden | Empty | Laden |
| | Empty | Laden | Empty | Laden |
| Arrival rate $\overline{\lambda}$ | 0.150 | 0.135 | 0.850 | 0.765 |
| Service rate $\mu$ | 1.6913 | 3.0163 | 2.6611 | 3.9490 |

The system can only start with the switchable server at either $\wp = 1$ or $\wp = 2$. So the system starts with two initial conditions. And whenever the reallocation condition $\chi_\wp$ is met at some time $t$, then two possible outcomes can occur, as identified in Section 5.3. Thus, four cases in total, and two possible initial states.

#### Cases 1 and 2

Beginning with the initial condition that the switchable server server starts serving FAST customer Type $\wp = 1$ in the DRM, and the reallocation condition $\chi_2$ is met. The calculations for utilization $\rho_1$, given the initial conditions, are

$$\rho_1 = \sum^{\iota} \frac{\overline{\lambda}_{(\iota,1)}}{2 \times \mu_{(\iota,1)}} = \frac{0.15}{2 \times 1.6913} + \frac{0.135}{2 \times 3.0163} = 0.0667 \ .$$

And the traffic intensity or utilization in the initial condition for customer Type $\wp = 2$ is

$$\rho_2 = \sum^{\iota} \frac{\overline{\lambda}_{(\iota,2)}}{\mu_{(\iota,2)}} = 0.5131 \ .$$

Therefore since $\rho_1 < \rho_2$, the reallocation server will switch to Type $\wp = 2$ as soon as the threshold is reached. Consequently, for the initial condition where the server starts at $\wp = 1$, only Cases 1 and 2 apply, and from the comparison of traffic intensity, Case 2 will occur.

### Cases 3 and 4

Similarly, if the initial condition calls for the switchable server would start at Type $\wp = 2$, and arbitrarily the reallocation condition $\chi_1$ is met, the calculations for utilization would be

$$\rho_2 = \sum^{\iota} \frac{\overline{\lambda}_{(\iota,2)}}{2 \times \mu_{(\iota,2)}} = 0.2566$$

and the utilization for customer Type $\wp = 1$ is

$$\rho_1 = \sum^{\iota} \frac{\overline{\lambda}_{(\iota,1)}}{\mu_{(\iota,1)}} = 0.1290.$$

Again, $\rho_1 < \rho_2$, and the DRM server will remain servicing Type $\wp = 2$. Therefore, for initial server condition starting at $\wp = 2$, Cases 3 and 4 apply, and only Case 4 will happens. In conclusion, regardless of where the reallocation server of the DRM starts, the server will switch to or remain at servicing Type $\wp = 2$ in the long run because $\rho_2$ *has higher traffic intensity.*

## 5.4 Discrete Event Simulation Model

To implement the DRM in the simulation model, input from the Battelle/Texas Transportation Institute (2008), the FHWA Office of Freight Management and Operations (2010) and Mexico Business Center (2010) from the the San Diego Regional Chamber of Commerce must be taken into account. These reports all agree that it is

desired to have the wait time of Non-FAST trucks reduced to within an hour. And according to the authors, the port has a suggested "capacity of approximately 120 trucks per hour." Consequently, it is desirable to have the Non-FAST queue be less than 120 trucks as compared to the FAST queue. This translates in Equation 4.5 to

$$Th_{F2R} \leq 120 \ .$$

Now, to determine the threshold for FAST trucks, consider DHS' stated objective to have wait time of FAST trucks be within 15 minutes. Therefore, the FAST queue should not be greater than 30 trucks. That is reflected by

$$Th_{R2F} \leq 30 \ .$$

For non-stationary analysis, the model was developed using Arena's simulation software, version 12.000.00 - CPR 9, from Rockwell Automation Technologies and Microsoft Visual Basic 6.5 version 1053. The POE process was modeled using the Arena simulation system for the process of queueing commercial trucks and servicing them at the inspection station. The DRM was coded using a combination of simulation software and Visual Basic code. Specifically, Visual Basic was used to code the logic behind minimum queue length identification, queue selection, and truck service times. The two programming languages interact through the built-in application programming interface in Arena that allows Visual Basic code to execute once a certain event is triggered.

Once the dynamic reallocation functionality was implemented, the model is the benchmark to verify that it behaves as the selected POE does using the data from several government agencies and sources Battelle/Texas Transportation Institute (2008). For comparison purposes, the model was configured without the dynamic reallocation capability. This configuration serves as the baseline model, that is, the current as-is situation.

To compare the level of effect of the dynamic reallocation policy, the arrival rate for the commercial trucks was increased by 10% to simulate a heavy load volume.

Similarly, the arrival rate was reduced by 10% for a light load volume. All three load volumes are combined with the use of the dynamic reallocation policy and without the policy.

### 5.4.1 Performance Measures

This section focuses on defining and identifying key performance measures that assess the model, and provide new metrics for the current POE environment. For the analysis of the methodology and the developed simulation model, queueing theory performance measures will be used and evaluated to compare the effectiveness of the methodology.

Security, cost and throughput are all important issues for the adequate and safe operations of the POE. However, the performance measures considered are specifically focused on throughput. Yet, the proposed methodology will have no impact on security or cost. Recall that there is no required change in the current staffing level or the infrastructure at the POE. Nor is there any need to change the security procedures when the CBP Agent is conducting the inspection process.

Performance measures are therefore selected with emphasis on throughput. These include:

- Average number of trucks in the system
- Average overall cycle time (in hours)
- Average Non-FAST cycle time (in hours)
- Average FAST Cycle Time (in hours)

And while other performance measures can be collected or calculates, these provide the information necessary to assess the effectiveness of the DRM under the Markovian arrival and service time assumption.

## 5.5 Results

The simulation program ran for 52 simulated weeks, or one year of simulated time, given that arrival rates are not the same every day, and replicated 26 times. As detailed in Section 5.4, the model ran with a "low, average, and high" arrival rates for comparison purposes. The following tables contain the results from the model, with and without the DRM, in conjunction with a different truck arrival load, that is light, normal and heavy.

Table 5.2 presents the results for the "average number of trucks in the system" and shows that none of the results overlap when comparing the baseline model with the DRM model by using a half width for a 95% confidence interval.

**Table 5.2**: Average number of trucks in the system for all arrival loads

| | Current as-is Model | | Dynamic Reallocation Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Light Load | 35.813 | 1.2615 | 31.190 | 0.7894 |
| Normal Load | 63.270 | 2.7734 | 44.457 | 1.0266 |
| Heavy Load | 109.320 | 7.8944 | 61.701 | 1.9321 |

In a different way of showing the average number of trucks in the system for all arrival loads, Table 5.3 contains the results in percentage improvement. Notice that higher arrival rates induces a bigger improvement of the DRM policy over the current as-is model. Yet, the average arrival rate improves the system's WIP by almost 30%.

After reviewing the results, the improvements are significant. Table 5.4 displays the results for average overall cycle time. Notice that the results are similar to those of the average number of trucks in the system in terms of percent improvement, and that the improvement column is also in units of time, but in minutes instead of hours. Similarly, Table 5.5 shows significant improvements.

**Table 5.3**: Performance improvement per average number of trucks in the system

| | Average Reduction | Percentage (%) |
|---|---|---|
| Light Load | 4.62 | 12.9% |
| Normal Load | 18.81 | 29.7% |
| Heavy Load | 47.62 | 43.6% |

**Table 5.4**: Results for average overall cycle time in hours on all arrival loads

| | Current as-is Model | | Dynamic Reallocation Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Light Load | 0.94719 | 0.03181 | 0.82128 | 0.01864 |
| Normal Load | 1.5014 | 0.06205 | 1.0502 | 0.02176 |
| Heavy Load | 2.3887 | 0.17153 | 1.3373 | 0.03793 |

**Table 5.5**: Performance improvement for average overall cycle time on all arrival loads in minutes

| | Average change (min) | Percentage (%) |
|---|---|---|
| Light Load | 7.55 | 13.29% |
| Normal Load | 27.07 | 30.05% |
| Heavy Load | 63.08 | 44.02% |

Notice in Table 5.5 that the average change improvement was switched to minutes for context.

In Tables 5.6 and 5.8, the trucks are separated by type. This allows for separate analysis on the effects of the DRM by FAST and Non-FAST trucks. Additionally, Table 5.7 presents the improvements of the DRM as a percentage over the base case. Similarly, the increase in Cycle Time for FAST trucks is noted as a percentage in Table 5.9.

**Table 5.6**: Results for average Non-FAST Cycle Time on all arrival loads

| | Current as-is Model | | Dynamic Reallocation Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Light Load | 1.1071 | 0.03642 | 0.95572 | 0.02153 |
| Normal Load | 1.7603 | 0.072 | 1.2224 | 0.02541 |
| Heavy Load | 2.8184 | 0.20263 | 1.5592 | 0.0443 |

**Table 5.7**: Performance improvement for average Non-FAST Cycle Time on all arrival loads

| | Average change (min) | Percentage(%) |
|---|---|---|
| Light Load | 9.08 | 13.67% |
| Normal Load | 32.27 | 30.56% |
| Heavy Load | 75.55 | 44.68% |

Notice that the significant benefits of Non-FAST trucks shown in Table 5.8, come over an increase in cycle time for FAST trucks as shown in Table 5.9. But although the percentages may look large, in terms of actual minutes, the time increased is small, and at most 3 minutes. Compare the cost of 3 minutes for FAST trucks with the improvement of 75 minutes for Non-FAST trucks, and the net gain is significant.

**Table 5.8**: Results for average FAST Cycle Time on all arrival loads

| | Current as-is Model | | Dynamic Reallocation Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Light Load | 0.06334 | 0.00771 | 0.07165 | 0.00948 |
| Normal Load | 0.06348 | 0.00692 | 0.09366 | 0.00881 |
| Heavy Load | 0.06418 | 0.00596 | 0.11303 | 0.01029 |

**Table 5.9**: Performance change for average FAST Cycle Time on all arrival loads

| | Average change (min) | Percentage (%) |
|---|---|---|
| Light Load | –0.5 | –13.12% |
| Normal Load | –1.81 | –47.54% |
| Heavy Load | –2.93 | –76.11% |

## 5.6 Summary and Conclusions

This section started the analysis of a POE with an assumption of Markovian service times, that is, the inspection times are memoryless, an has no effect on the next truck inspection. The assumption is a starting point when there is no information on the service time, or whether there may be some correlation between them, and allows for a simplified case model.

The analysis of the simplified case showed that for any given policy on when to switch or reallocate a dynamically server, once the threshold is met, the server will tend to stay or go to the type of server that has a higher utilization. This activity is explained by the fact that the server with higher utilization, either $\rho_1 < \rho_2$, or $\rho_2 < \rho_1$, indicates which type of server is much busier, and by consequence is in need

of more help. And of course, the goal is to balance the load. The numerical example in Section 5.3.3 also had the expected outcome.

The major complications to the decision analysis of this situation are: the non-stationary nature of the arrival process and the necessity of transient analysis. Other real world complications include the small queue in front of the inspection stations, that needs to be flushed before a server switch can be completed. But from the results, the straightforward conclusion is that the system observes an overall performance improvement for all load levels. In particular, the improvement in minutes is a reduction of over 25 minutes of wait time to cross the border without modifying any of the security procedures or adding any cost to attain the improvements at typical load levels. The benefits are even greater at heavy load volumes with wait time reductions of almost 45%.

But given the fixed number of total servers, the improvement comes at a small increase in the cycle time of FAST trucks. However the increase is minimal. Even though the percentage seems significant, it translates to less than a 3 minute increase in the average time to process FAST vehicles, while Non-FAST trucks can be reduced by over 30 minutes. And the effect of the dynamic reallocation policy is even more significant in a situation where the arrival rate load in greater than current levels. With infrastructure taking years to build, these policy improvements offer a way to better utilize current resources, without compromising any security procedures or adding any costs.

The section discussion also considered that when there is a policy that changes the type of customer that a server can handle, then the analytical model needs to cover both trajectories of the service that server is providing. This is one reason why the state space of an analytical model increases, when a server switch is allowed. In the next section the assumption that the service time is Markovian will no longer apply. In this case, other methods are needed to handle the even higher complexity of the model. Section 6 addresses the application of a Coxian k-phase approximation

as a better general distribution approximation of service time of the POE. Then, analysis turns to the POE queue, and a discussion of the different arrival loads for the POE and the effect of the DRM. Finally, a variance reduction policy is described and incorporated into the DRM for evaluation.

# 6. ANALYSIS OF THE POE QUEUE II: POE ANALYTICAL MODEL WITH COXIAN SERVICE TIME APPROXIMATION

This section continues the research of a two-queue terminating non-stationary queueing system that can be applied to the commercial border crossing process (BCP). The BCP is described in Section 2 and discusses the significance of balancing throughput, cost and security. Section 4 described the BCP as a queueing model, and introduced the Dynamic Reallocation Methodology (DRM) as a method to improve throughput without affecting other objectives. Also, analysis of empirical data in Appendix B suggested a that general service time distribution for inspection times would be better suited for modeling. Section 5 used a simplification assumption of Markovian service times to prove that the DRM would improve traffic intensity. It also illustrated a significant performance improvement for average overall cycle time on all arrival loads in Table 5.5, with minimal effects on FAST truck delays in terms of minutes as seen in Table 5.9.

The discussion continues on the analysis of the POE, but now with the service time being described by a general distribution. After the introductory arguments for a better approximation of service times, the focus turns to an analytical model and explores the state transition diagram of the POE, if a general distribution is employed. To characterize this behavior of service times, an approximation using Mixtures of Generalized Erlang (MGE) distributions is employed. MGE are also known as Coxian distributions. This exposition will illustrate the complexities and challenges of a POE model when a Coxian approximation is used to approach a general distribution. The section continues with an analytical model that incorporates a Coxian approximation for service times with an assumption of no short queues at the inspection station and implemented in Mathematica, a fully integrated environment

for technical computing. Finally a summary of the research and concluding remarks are made.

## 6.1 Introduction

Analysis of a queueing system tend to become mathematically more tractable when exponential distributions are used. A mathematical analysis of a simplified case model in Section 5.3.3, showed that the DRM will switch the moveable server to service or inspect the truck type $\wp$ that has a higher traffic intensity $\rho_{\wp}$, at the moment that the reallocation policy is met. Also in Section 5, a simulation of the POE modeled the current situation at the border crossing, and was compared with another model that incorporated the DRM, assuming the Markovian case for the inspection stations. The results showed that the benefits were significant, and in the case of heavy arrival rates, the benefits were over 46%. Refer to Tables 5.3 and 5.5 for details.

After an initial analysis of the border crossing process using a Markovian assumption for service times, the attention of the research now focuses on the POE with a general distribution for the service time. The discussion on this section deals with moving the modeling and analysis of the POE closer to a more realistic representation of the POE, in particular regarding inspection or service time. It is also about understanding the issues in creating an analytical model, and presenting the transient analysis. In this case, the DRM model has the Coxian phased approximation implemented for service time, so that a comparison can be made about improvements regarding the performance measures that were identified for the POE model.

Recall that the POE is fundamentally a queue and service environment, but it does not behave in a way where the system can reach a stable size, or steady-state. Additionally from Equation 4.2, the system is a two-queue environment that follows a general service time distribution. Yet there are common and significant characteristics of all POEs, that should be considered when modeling and analyzing

any potential improvement policy. In particular, POE characteristics that deal with service dependencies are:

- The POE never reaches steady-state.
- The inspection times (i.e. service times) are not independent, as observed in Section 4.4, and according to the research done by Cetin and List (2004).
- The inspection times can have dependencies based on security procedures, which are determined by the National Terrorism Advisory System as discussed in Section 2.2.1.

These characteristics serve as foundation for a general distribution of service time. And as previously mentioned, POE service times should be not be considered independent because according to Cetin and List (2004) correlation exists, and if ignored, the model model may not produce accurate results. Additionally, security procedures such as dogs coming to the lines to sniff and check for illegal cargo, create dependencies in the inspection time. Particularly when the security procedure calls for a freeze of a small number of trucks in the queue that have to wait for a close dog sniffing inspection for all trucks involved. In other occasions, DHS may have undisclosed information about attempts to bring contraband through the POE, and therefore there are more thorough inspections, and thus longer for only a few segment of trucks. This supports the description of the POE in Equation 4.2, as having a general distribution for inspection or service time.

## 6.2 Analytical Model with a Coxian-Phased Approximation of Service Times

After the analysis from Section 4.4, there was not a clear fit for a theoretical distribution. In this section, the implementation of the Coxian k-phased approximation for a general distribution will follow the work from Altiok (1996) and Curry and Feldman (2011), by using the moment-matching approximations strategy. In

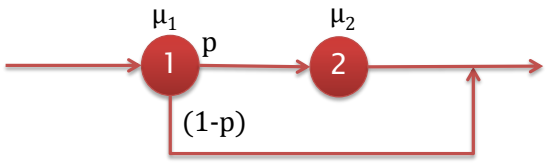



**Fig. 6.1.**: MGE 2-phased Transition Diagram with $\frac{1}{2} \leq C^2 < \infty$

their research, the LST (Laplace-Stieltjes transform) "of any distribution function can be approximated arbitrarily close by a rational function." The main problem, they explain, is the lack of a method to determine the parameters and structure of the phase-type distribution.

Consider the transition diagram from Figure 6.1, the parameters needed for a MGE-2 include $\mu_i$ (mean service time per phase $i$) and $p$ (probability of moving to the next phases), where the idea is to get arbitrarily close to the approximating distribution. In Section 7.2, the squared coefficient of variation, $C^2$, is discussed and the parameters are calculated using the the empirical data in order to be implemented in the simulation model. In the analytical model, the data is fitted to a 2-phase approximation to better handle the number of phases.

### 6.2.1 Research for Approximating Service Distributions

Authors like Ojah et al. (2002) and Bradbury (2010) coincide with Haralambides and Londono-Kent (2001) in their conclusion that "study of what actually happens at the border reveals significant time and cost inefficiencies in the border crossing process." And the difficulty in this matter is discussed in Section 3.2. However, the research done by Ashur et al. (2001) use the Erlang distribution because it "is frequently used in queueing systems to represent service-time distributions in discrete systems simulation."

Usually, phased-type distributions are used in models of stochastic characteristics. Altiok (1996) discusses the use of MGE, often called Coxian distributions, which have been used in these type of analyses. Additionally, these distributions have been used in the analysis of manufacturing, computer and communication systems. When using these distributions, they are characterized by phases, that is, spending an exponentially distributed amount of time in each phase, and the key is determining the number of phases needed and the phase. Curry and Feldman (2011) and Altiok (1996) elaborate on the approximation of service times using MGE, with more details to follow the discussion in Section 6.4.

Whitt (2007) also used approximation methods to help set the staffing requirements in service systems he was researching. His queueing model was a $M_t\ /GI\ /s_t+GI$. His work is similar to this research of POEs, in that the "model is difficult to analyze mathematically, so that the staffing problem is challenging. However, there is one special case that is amazingly tractable: the Markovian $M_t\ /M\ /s_t+M$ model in which $\theta = \mu$."

There are many papers dealing with the analysis of non-stationary queueing systems, e.g., Choudhury et al. (1997), Ong and Taaffe (1988), Margolius (2005), and Margolius (2007), and they almost always begin with the Chapman-Kolmogorov forward equations which will be used for the analytical model. However, there are not many decision control problems using these formulations. Most of the queue control literature, e.g., Adusumilli and Hasenbein (2010), Ata (2006), and references in the research work of Ata (2006), deal with steady-state results for stationary queueing systems.

### 6.2.2 Analytical Model Benefits

Mathematical analysis is the basis of many research studies. The aim is to characterize the system's behavior so that improvements can be validated, and in some cases, proven to work. Having the entire model described by a mathematical struc-

ture, for example the probability distribution for the number of jobs in the system ($p_n = \Pr\{N = n\}$) is very desirable, particularly in modeling queueing systems. Once that information is mathematically developed, the entire system can be characterized, and all the information about the its behavior can be computed.

When a model can be described mathematically, and an analytical solution to a mathematical model is available without being computationally inefficient, "it is usually desirable to study the model in this way, rather than via a simulation" (Law and Kelton, 1991). Since the POE is basically a queueing system, it makes sense to use queueing theory models to describe it. In fact, queueing theory was developed in order to provide models that predict and to describe the behavior of systems that provide a service for randomly behaving demands. Pioneered first by the work of Erlang, "The Theory of Probabilities and Telephone Conversations" in 1909, and continued by Molina (1922) and others. They became the basis of queueing theory, and also the basis of the approximation for the general distribution which will be used in this section.

However, it is not always possible to develop an analytical model that accurately reflects the real world environment. In some cases, the modeler has to make assumptions to be able to develop a closed-form solution or even an approximation of the system. Tractable queueing models require reasonable analytical assumptions and are generally based on the "forgetfulness" aspect of the exponential distribution. Simulation models offer a great deal of flexibility as they are generally able to describe with great detail almost any system. It can be used to validate models and approximations, and is "generally robust with respect to modeling distributional assumptions and allows for more realistic modeling of system interactions" (Curry and Feldman, 2011).

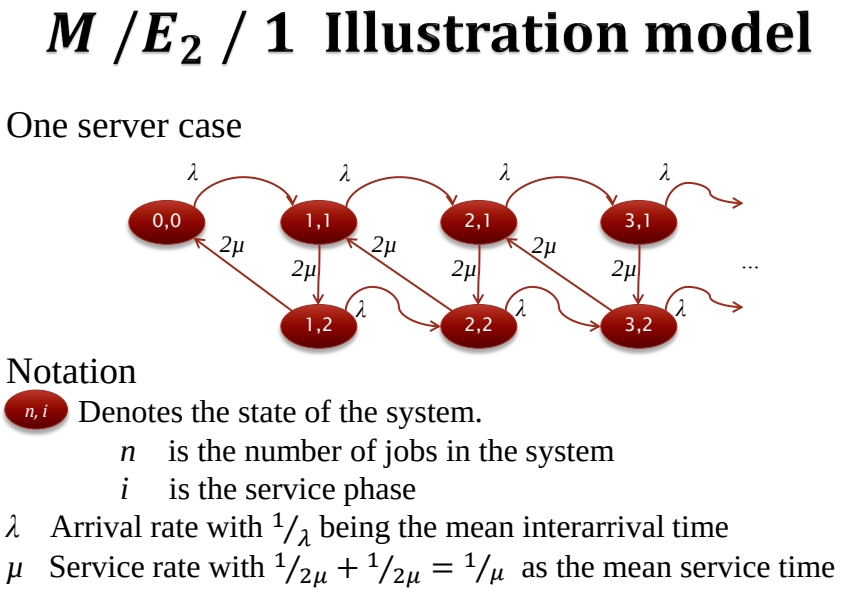


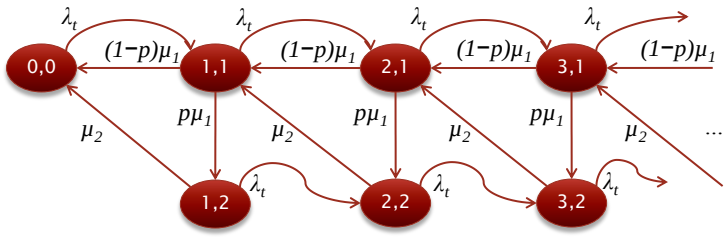

**Fig. 6.2.**: State Transition Diagram for $M_t\ /E_2\ /1$ Illustration Case

## 6.3 State Space for Coxian Approximations

The estimation of the service time by using the Coxian 2-phase distribution as an approximation method for a general distribution, will be implemented following the work done by Curry and Feldman (2011) and Altiok (1996). To illustrate the increased complexities of the Coxian, recall from Figure 5.2 in the Markovian state transition diagram that the service time is exponential and there is no DRM. In that scenario, there is only one phase to keep track of when accounting for the number of trucks in the system. In fact the number of trucks in the system is the number of states in the system. When an Erlang 2-phased distribution is considered for service time, such as in Figure 6.2 depicting an illustration model, the state transition diagram needs to account for another level or phase for each service state.

# $M_t\ /MGE_2\ /\ 1$ Analytical model

One server case




Notation

n, i Denotes the state of the system.

- $n$ is the number of jobs in the system
- $i$ is the service phase

$\lambda_t$ Time dependent arrival rate

$p$ Probability of going to the next phase

$\mu_i$ Service rate per phase $i$

**Fig. 6.3.**: State Transition Diagram for $M_t\ /MGE_2\ /1$ Analytical Model

Keeping track of phases is the characteristic of all MGE. MGE have phases in each service or arrival state and need to be tracked, depending on which process employs MGE. Observe in Figure 6.3, that the analytical model is employing MGE for the service time distribution, and in this case, there are two phases to keep track of. However, phases are not limited to just two. There are cases, as will be discussed in Section 7.2, where the calculations for the number of phases in the approximation turns out to be greater than two. In those situations, the state space doubles or triples, etcetera, in size and these additional phases of must also be tracked.

Recall that the analytical model will assume that the Coxian approximation will consists of two phases. To illustrate the state space, we have to keep track of the number of trucks in the system $n$, and the phase of the service for each identical server in the system $(i, j, k, ...)$. For example, the illustration model in Figure 6.3 with only one server, has the following state space.

$$
\begin{array}{l}
\{[\ (0)\ ]\ , \\
[\ (1,1)\ ,\quad (1,2)\quad ]\ , \\
[\ (2,1)\ ,\quad (2,2)\quad ]\ , \\
[\ (3,1)\ ,\quad (3,2)\quad ]\ , \\
[\ (4,1)\ ,\quad (4,2)\quad ]\ , \\
\qquad \vdots
\end{array}
$$

Where the first index represents the number in the system, and the following index is the current service phase.

In the case of two identical servers and continuing to assume only two phases, the state space would be as follows:

$$
\begin{array}{l}
\{[\ (0)\ ]\ , \\
[\ (1,1)\ ,\quad (1,2)\ ]\ , \\
[\ (2,1,1)\ ,\quad (2,1,2)\ ,\quad (2,2,2)\ ]\ , \\
[\ (3,1,1)\ ,\quad (3,1,2)\ ,\quad (3,2,2)\ ]\ , \\
[\ (4,1,1)\ ,\quad (4,1,2)\ ,\quad (4,2,2)\ ]\ , \\
[\ (5,1,1)\ ,\quad (5,1,2)\ ,\quad (5,2,2)\ ]\ , \\
\qquad\qquad \vdots
\end{array}
$$

Notice that after the empty state, the next two states do not need an additional phase index since there is only one job / entity / truck in the system.

In a three identical server situation, the state space grows to the following.

$$
\begin{array}{l}
\{[\ (0)\ ]\ , \\
[\ (1,1)\ ,\quad (1,2)\quad ]\ , \\
[\ (2,1,1)\ ,\quad (2,1,2)\ ,\quad (2,2,2)\ ]\ , \\
[\ (3,1,1,1)\ ,\quad (3,1,1,2)\ ,\quad (3,1,2,2)\ ,\quad (3,2,2,2)\ ]\ , \\
[\ (4,1,1,1)\ ,\quad (4,1,1,2)\ ,\quad (4,1,2,2)\ ,\quad (4,2,2,2)\ ]\ , \\
[\ (5,1,1,1)\ ,\quad (5,1,1,2)\ ,\quad (5,1,2,2)\ ,\quad (5,2,2,2)\ ]\ , \\
\qquad\qquad\qquad\qquad \vdots
\end{array}
$$

Notice here as well that after the zero state, the next two states do not need an additional phase index. And the next three state spaces only have two in the system, so only two indexes are needed to track the phases.

Using the MGE as the transient probability approximation method discussed in Section 6.4, requires the state space to double, or triple, or more in order to keep track of the phases of the problem. Depending on the number of phases needed to approximate the distribution, the the number of states could jump to several million or more.

### 6.3.1 State Transition Diagrams for Coxian Approximation

This section presents the the generator matrices that the analytical model will use, and they follow the state space description mentioned in the previous section. The generator matrices in Table 6.4, Table 6.5, and Table 6.6 move the state space from the top row to the left hand column with the corresponding arrival rate or service rate given a 2-phased MGE for a service time approximation.

The number of servers is important in creating the generator matrices. Figure 6.4 presents the generator matrix with only two identical servers, while Figure 6.5 shows the generator matrix for three identical servers, and Figure 6.6 is developed for four identical servers. Since the state space is dependent on the number of servers or inspection Agents as is the case in the POE, the calculations are based on the number of available inspection stations.

As illustration in Figure 6.4, take a starting point of being in state (0), this state can only move to state $(1, 1)$ with an arrival $\lambda_t$. If the originating state is (1,1), then one option is to move to state (0) with $(1 - p)\mu_1$. Another possibility is to move to $(1, 2)$ with $p\mu_1$. One last possibility is to move to state $(2, 1, 1)$ with an arrival $\lambda_t$, and so on. The structure of the matrices and use in the calculations of the analytical model is further discussed in Section 6.4.1.

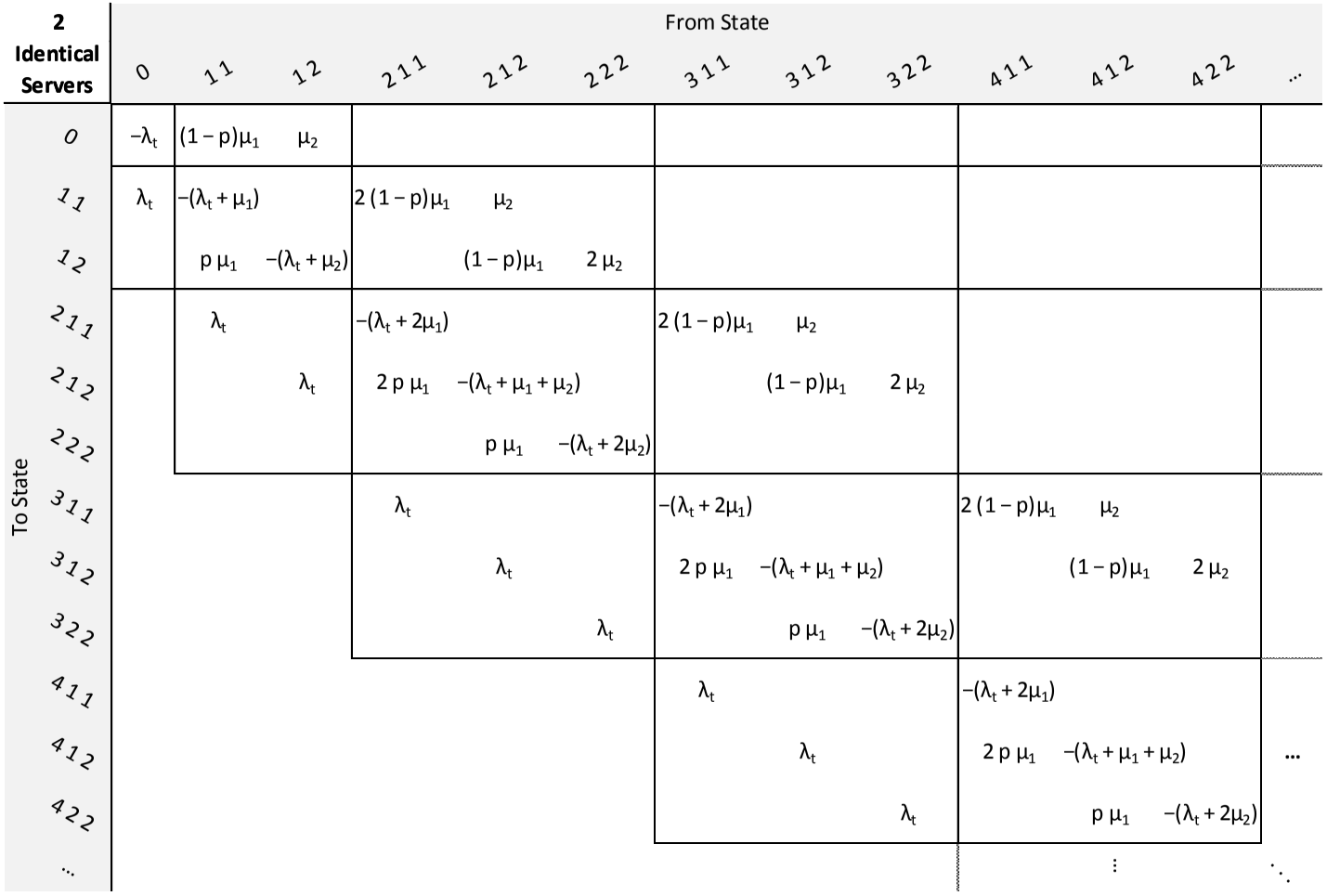



**Fig. 6.4.**: Coxian Generator Matrix for 2 Identical Servers

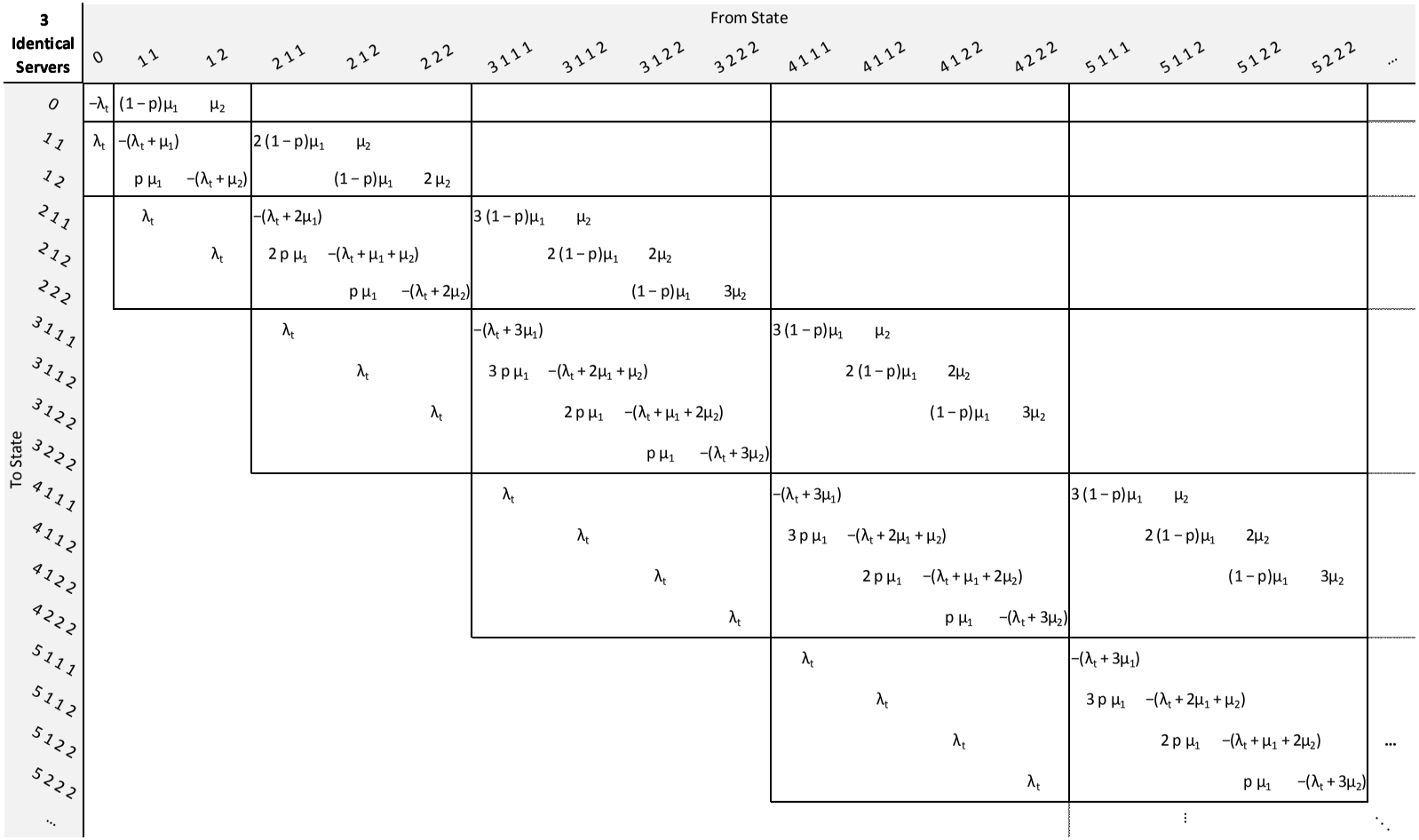



**Fig. 6.5.**: Coxian Generator Matrix for 3 Identical Servers

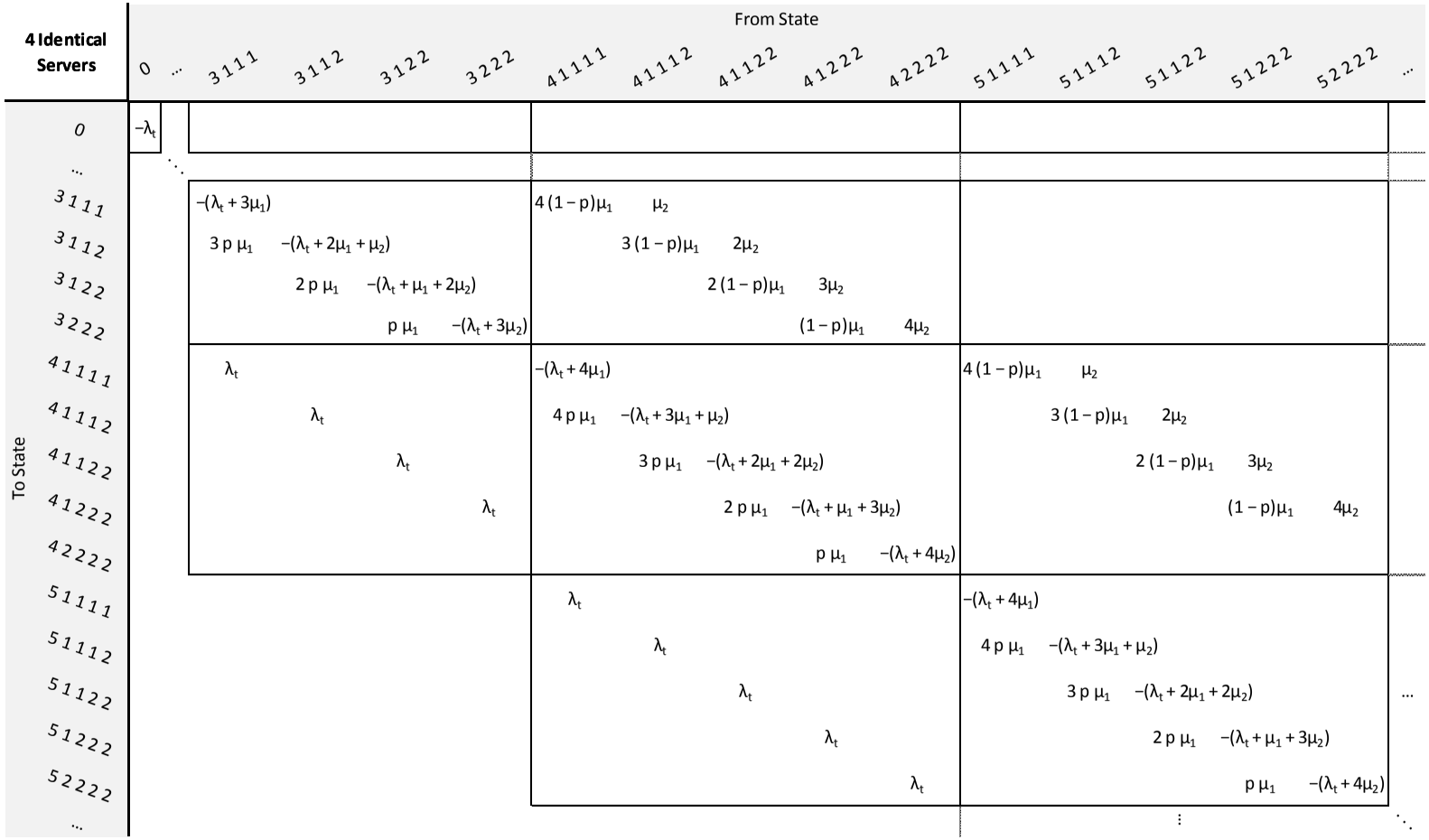



**Fig. 6.6.**: Coxian Generator Matrix for 4 Identical Servers

## 6.4 Approximation Method

An approximation method that uses the generator matrices is based on the standard results for Markov processes from Çınlar (1975). Thus, the approximation for the transient behavior as can be calculated by:

$$P_{t+\Delta t} = P_t + P_t \cdot G_t \cdot \Delta t \tag{6.1}$$

for $t \geq 0$ and a suitably small $\Delta t$.

However, the approximation in Equation 6.1 isn't an exact solution to $p_n(t) = \Pr\{N(t) = n\}$, and requires a small $\Delta t$ that changes time into small discrete time intervals, similar to the analysis performed in a discrete-event simulation model. Furthermore, in order to sum all the probabilities, the system needs to have a theoretical $n_{max}$, or expected maximum number in the system. This is illustrated in Figure 6.7, where there is an end to the generator matrix. The objective is to have a large enough $n_{max}$ so that when the probabilities are added, the probability of having more jobs / trucks in the system will be negligible and the sum would be within an acceptable range which approximately equals to one. As mentioned before, the analytical model will restrict the number of phases to two, in order to manage the total the number of states that are tracked.

To illustrate the cap in the state space and the 2-phased Coxian service time, consider the case of Non-FAST tucks where there are four identical servers or inspection stations, as shown in Figure 6.7. In this case, the state space looks similarly to those in Section 6.3, however there is finite maximum to the state space.

{[ (0) ] ,
[ (1,1) , (1,2) ] ,
[ (2,1,1) , (2,1,2) , (2,2,2) ] ,
[ (3,1,1,1) , (3,1,1,2) , (3,1,2,2) , (3,2,2,2) ] ,
[ (4,1,1,1,1) , (4,1,1,1,2) , (4,1,1,2,2) , (4,1,2,2,2) , (4,2,2,2,2) ] ,

$$
\begin{array}{l}
[\ (5,1,1,1,1)\ ,\quad (5,1,1,1,2)\ ,\quad (5,1,1,2,2)\ ,\quad (5,1,2,2,2)\ ,\quad (5,2,2,2,2)\ ]\ , \\
[\quad \vdots\ ,\quad \vdots\ ,\quad \vdots\ ,\quad \vdots\ ,\quad \vdots\ ]\ , \\
[\ (n_{max}-1,1,1,1,1)\ ,\quad (n_{max}-1,1,1,1,2)\ ,\quad (n_{max}-1,1,1,2,2)\ , \\
(n_{max}-1,1,2,2,2)\ ,\quad (n_{max}-1,2,2,2,2)\ ]\ , \\
[\ (n_{max},1,1,1,1)\ ,\quad (n_{max},1,1,1,2)\ ,\quad (n_{max},1,1,2,2)\ , \\
(n_{max},1,2,2,2)\ ,\quad (n_{max},2,2,2,2)\ ]\ \}
\end{array}
$$

Notice that the initial state spaces do not need all indices. This is the same situation as in the state space descriptions in Section 6.3, when there are less jobs in the system than available servers.

### 6.4.1 Generator Matrix Structure

To perform the calculations of the model, notice in Figure 6.8 that there is a repeating structure of the generator matrix identified by **A**, **B**, **C** sub-matrices that can be used to perform the calculations in order to obtain a new $P_{t+\Delta t}$. The analytical model takes advantage of this structure and enables the dynamic calculations of the probabilities and thus expected size of the system with an appropriate maximum size and $\Delta t$

Notice as well in Figure 6.8 that matrix **A** contains the arrival rates, which are time dependent, therefore in terms of calculations, matrix **A(t)** changes with time. Matrix **C** consists of the service rates, and matrix **B** completes the generator matrix. Similarly to matrix **A(t)**, matrix **B** contains time dependent arrival rates in $\lambda_t$, therefore for calculations, **B(t)** is also time dependent.

Finally, $\Delta t$ represents how small are the time step intervals to calculate the next $P_{t+\Delta t}$. A smaller $\Delta t$ gives more accurate results, but many more calculations to perform. So, along the generator sub-matrices **A(t)**, **B(t)** and **C**, and an appropriately small $\Delta t$, the operations can be performed iteratively.

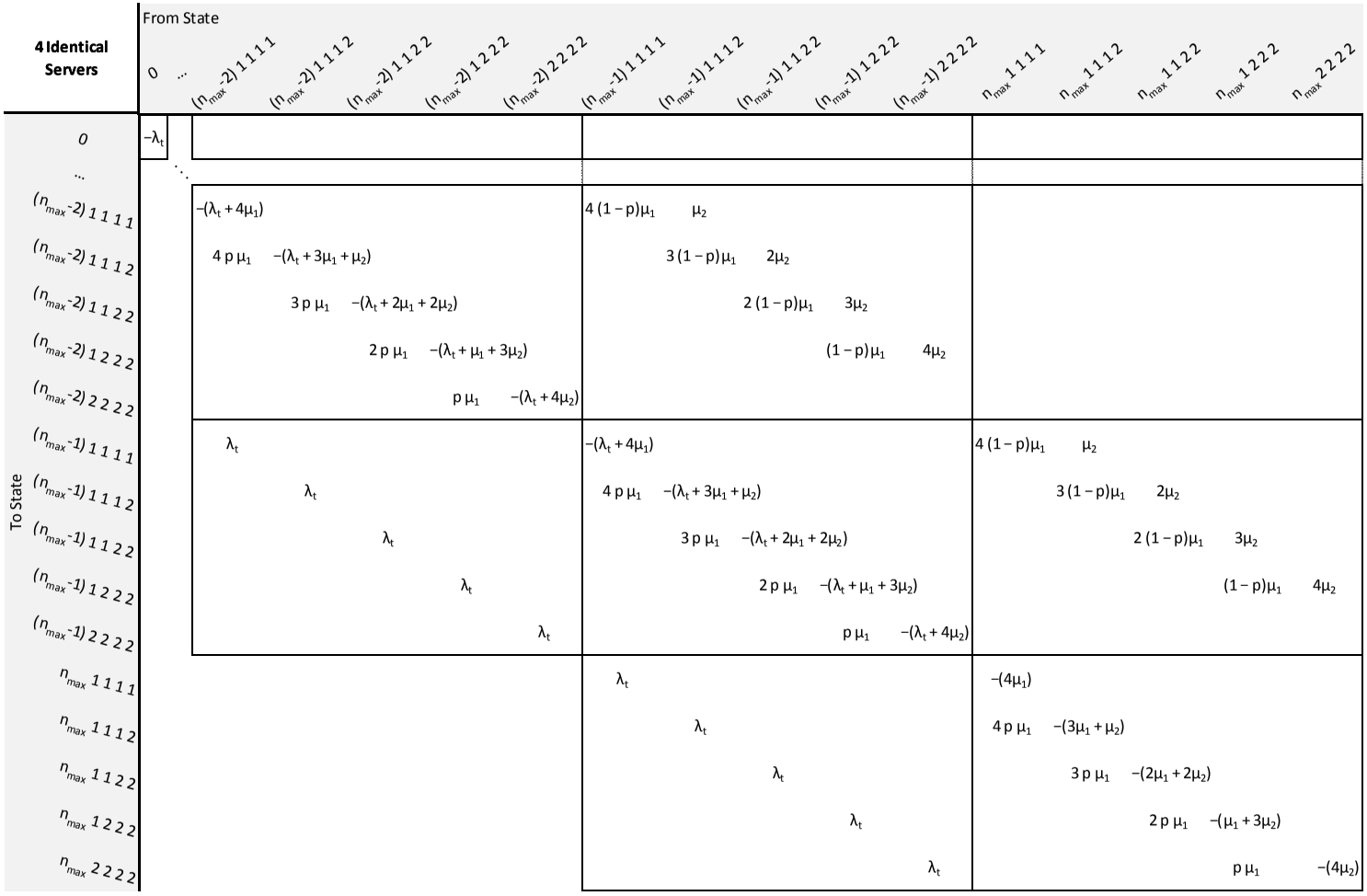



**Fig. 6.7.**: End of the Coxian Generator Matrix for 4 Identical Servers

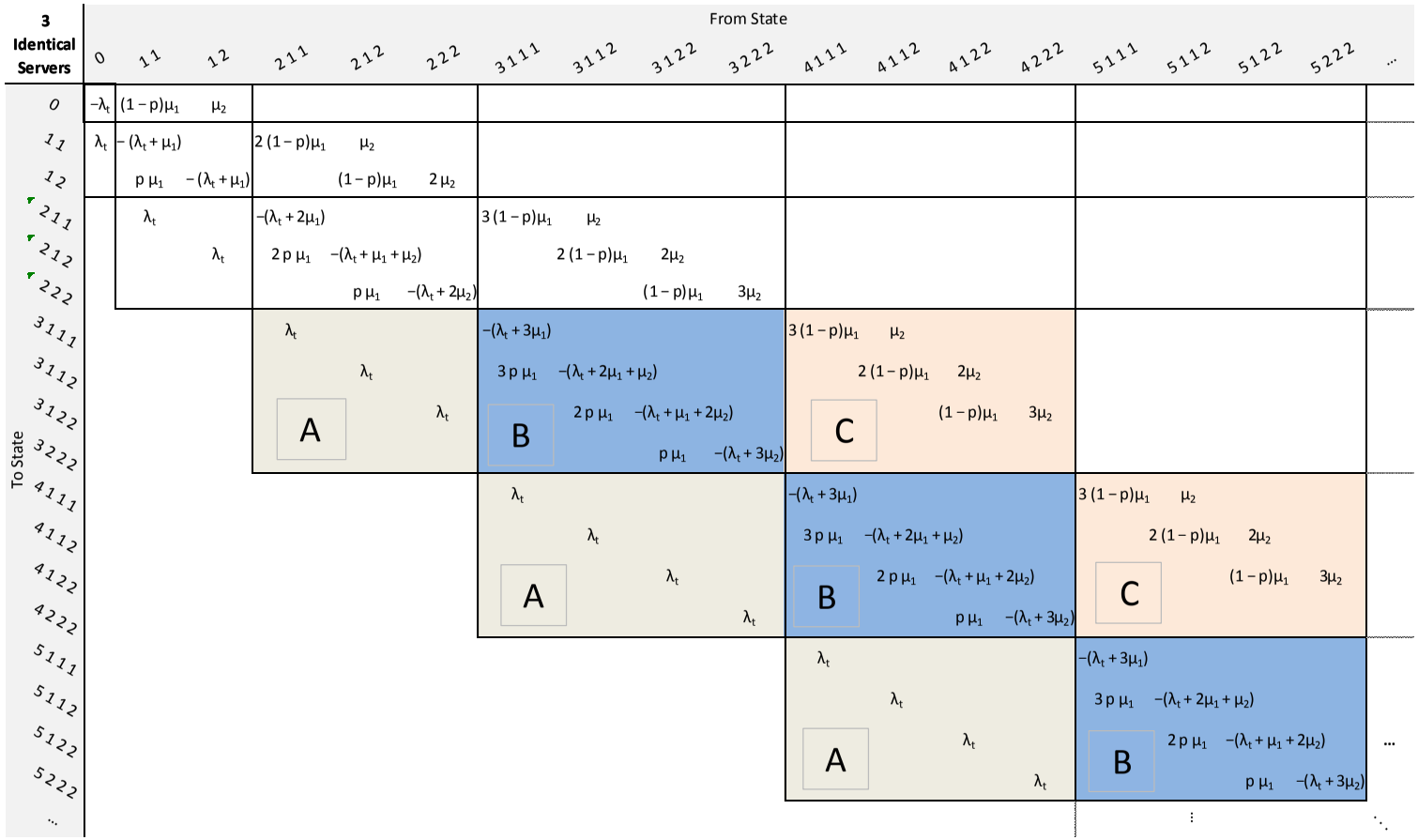



**Fig. 6.8.**: ABC Coded Coxian Generator Matrix for 3 Identical Servers

### 6.4.2 Transient Behavior Observations

The analytical model was developed in Mathematica, version number: 4.2.0.0, in the Windows platform. The program ran on a Dell Optiplex GX620 with a Pentium D 3.20 GHz processor, and 2.00 GB of RAM, on a Windows XP Professional platform.

For the analytical model, some critical parameters are needed to manage the run time of the program, if the parameters are too big or the $\Delta t$ too small, the the run time increases significantly. in the model, some of the critical parameters used are as follows:

dt = 0.001
Tmax = 16000*dt
K = 1000
sum1err = 0.01
probAccuracy = 0.000000001

In the model, *dt* is the time step for transient solution, and represents $\Delta t$ in the analytical model. $Tmax$ as the name implies, is the maximum analysis time in simulated hours. $K$ is the size of the maximum probability groups, which also represents the maximum expected size in the system. And $sum1err$ along with *probAccuracy* represent the model's internal error checking parameters to determine the accuracy levels for the probabilities at each time step.

A major aspect of the analytical model is that hundreds or thousands of simulation runs would be needed to approximate $p_n(t)$. Where as, one analytical model execution yields this result. Estimating $E[N_t]$ could be accomplished with fewer simulation runs, but the complete distribution would not be available. As shown in Figure 6.9, the result of the analytical model is the observation of the transient behavior with the complete distribution information for each $E[N_t]$ across time $t$.

Also notice in Figure 6.9, that the expected number of Non-FAST trucks in the POE is building up to around the noon hour, and then it starts to decrease, until

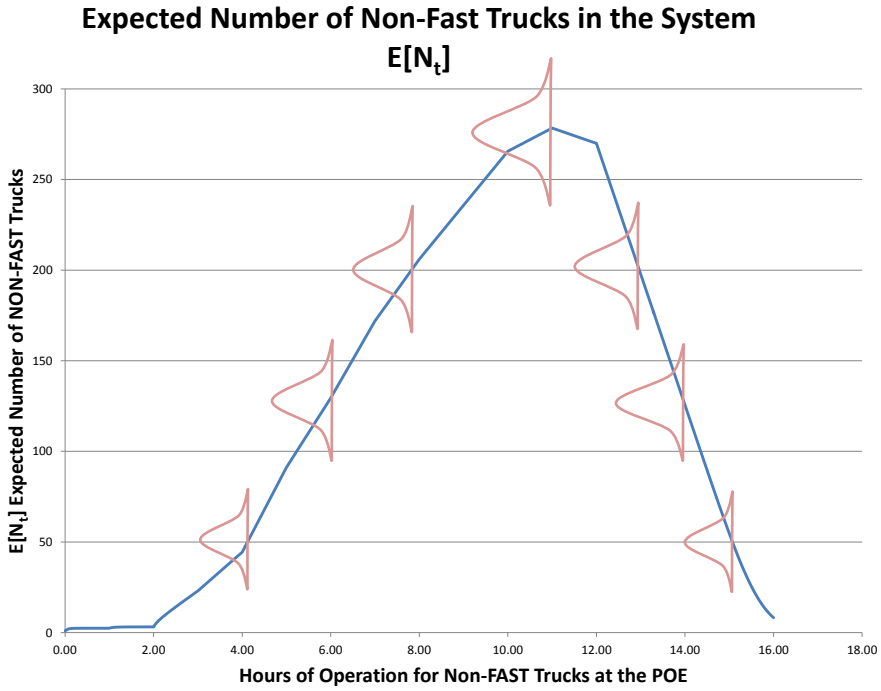



**Fig. 6.9.**: Expected Number of Non-FAST Trucks in the System

eventually the line is flushed. This is the same behavior observed in the field (Appendix B), and from the literature (Battelle/Texas Transportation Institute, 2008). Recalling that the model is separated into Non-FAST and FAST models for service time, the analytical model can now approximate the general service time distribution of the border crossing process at the POE and be used as a benchmark for the simulation model.

## 6.5 Summary and Conclusions

In this second part of the POE analysis, the discussion of service time distributions was continued by replacing the Markovian assumption of service time with a service time that is described with a general distribution. To characterize general service time behavior, an approximation method using Mixtures of Generalized Erlang

(MGE) or Coxian distributions is employed. The discussion turned to the benefits of using an analytical model, and continued with a description of the state space, and the state transition diagrams using a Coxian approximation. It was also noted that this method breaks the service time into phases, which increases the number of states.

The latter sections present the approximation method of the analytical model, and the structure of the generator matrices. The model takes advantage of the structure of the generator matrices to keep track of the probabilities and calculate the expected number of trucks in the system at every time step. Finally, the results showed that the transient behavior of the analytical model is analogous to the observed field data, and the literature on POEs, which will be useful in creating a benchmark for the simulation model.

Overall, there are significant conclusions that can be taken away from the research in this section, including:

- Field observations indicate that the service time is not Markovian.
- The analytical model can now approximate the general service time distribution of the border crossing process at the POE, using a Coxian or MGE distribution as shown by Curry and Feldman (2011) and Altiok (1996).
- The model results provide complete distribution information for each time step in the analysis.
- The transient analysis showed the same behavior that was observed in the field, and in the literature for the expected number of trucks in the system $E[N(t)]$.
- The analytical model can now be used in tandem with a simulation model for complicated analysis, such as the implementation of the DRM.

# 7. ANALYSIS OF THE POE QUEUE III: DRM SIMLUATION MODEL WITH COXIAN SERVICE TIME

This section converges the analysis work on the POE in Sections 5 and 6, as a two-queue terminating non-stationary queueing system that can be applied to the commercial border crossing process, as described in Section 2. This is accomplished by implementing the Dynamic Reallocation Methodology (DRM) from Section 4, in a simulation model that uses general service times, as a method to improve throughput without affecting cost or security procedures. The simulation model takes into account the work done in Section 5 that proved that the DRM would improve traffic intensity with a Markovian service time assumption, and incorporates the work done in Section 6 where this assumption was removed. In that section, the POE inspection time was described by a general service time approximation which used the Coxian or MGE distributions as described by Curry and Feldman (2011) and Altiok (1996). The approximation was implemented in an analytical model of the POE, where the transient analysis showed the same behavior that was observed in the field, and from the literature, for the expected number of trucks in the system, $E[N(t)]$.

The first section focuses on having a DRM implementation on a full POE model with a Coxian phased approximation for service times. The first step is to establish a basis for comparison, and the analytical model is used to establish a baseline simulation model. Once the baseline results are established for the POE, the next section step is to evaluate the effectiveness of the DRM against the current as-is situation with Coxian service times, using a variety of performance measures including CT and WIP for both FAST and Non-FAST trucks. For comparison purposes, the same basis of arrival rates are used as in the Markovian comparison, i.e. high, normal, and low. The results of the DRM and conclusions are immediately discussed.

This section concludes with a variance reduction policy that can be used in conjunction with the DRM policy. Following on the benefits of using the DRM, the

focus turns to reducing service time variability. And the proposal here is to reduce variance by using secondary inspections. The DRM / Coxian phased approximation of service times model gets modified to include a variance reduction procedure to observe results. Finally a summary of the research and concluding remarks are made.

## 7.1 Introduction

One of the main reason to do analytical solutions and a significant benefit, is that the entire information of the model behavior is captured when the complete activity distribution can be described. The main reason for simulation is "its ability to deal with very complicated models of corresponding complicated systems," however "simulation isn't quite paradise, either" (Kelton et al., 2002). Because, even with a very realistic simulation model, once the model runs it will produce a result, but this result is only be one realization of the distribution function behind the observed behavior. And in the case of the number of trucks in the system, it is only a realization of the probability distribution for the number of trucks.

To be able to obtain the entire information of the observed behavior, the simulation must have long simulated run time, and several replications of the simulation model, so that the combination of results generated from the simulation can describe the distribution function of the behavior under consideration. Fortunately, higher computing power and better simulation tools, allow complicated models today to have short run times, even with several replications.

Since the initial years of computers, simulation was seen as a companion tool to analytical modeling, and even as a way to develop and validate analytic models. According to Ignall et al. (1978), "the reason for doing so is to give the potential user of the analytic model confidence that it is a safe substitute for the more accurate simulation model." In this research a combination of analytical tools to accurately represent service time are implemented in a simulation model. This section will

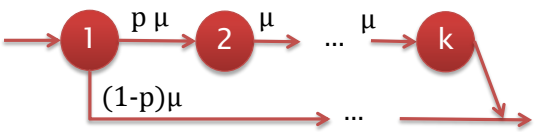



**Fig. 7.1.**: $k$-phased Generalized Erlang Transition Diagram

address the implementation of the Coxian approximation of service time in the simulation model that uses DRM to improve performance measures, regardless of the number of phases needed.

## 7.2 Phases for POE Coxian Service Time by Truck Type

With the results from the analytical model available, the next step is to incorporate the Coxian approximation method to the simulation model. In Section 5, the service time was exponential, and with the analytical model in Section 6, the inspection or service time was considered to follow a general distribution, represented by a Coxian approximation.

As previously indicated, the implementation of the Coxian k-phased approximation will follow the work from Altiok (1996) (pp. 52-57) and Curry and Feldman (2011) (pp. 89-90), by using the moment-matching approximations strategy. The approximation requires the first and second central moment about the origin, and the $C^2$, the squared coefficient of variation. Omitting the description other $C^2$ cases, if the $C^2 < 1$, the the approximation is as follows: Empirical studies in literature suggest that the first first two moments are adequate, because the third moment captures skewness of the distribution, when there is low variability. In this case and with $C^2 < 1$, the Coxian k-phased distribution in Figure 7.1 can be used to model

random phenomena. The parameters needed include $\mu$, mean service time, which is the same for all phases, and $p$, the probability of moving to the next phases.

Therefore, in the case of $C^2 < 1$ and given $E[X]$, and $C^2$, we have that a $k$ should be selected such that

$$\frac{1}{k} \leq C^2 \leq \frac{1}{k-1} \tag{7.1}$$

And once the $k$ is found, then the parameters of $\mu$ and $p$ are respectively given by:

$$1 - p = \frac{2kC^2 + k - 2\sqrt{k^2 + 4 - 4kC^2}}{2(C^2+1)(k-1)} \tag{7.2}$$

and

$$\mu = \frac{1 + (k-1)p}{E[X]} \tag{7.3}$$

Now, having the Coxian k-phased approximation defined, the following the calculations for the parameters of the approximation are based on the collected data in Appendix B. Notice as well that they are are separated by truck type $\wp$ and load status $\iota$, and the cases identified follow the same cases that were used in the data fitting analysis in Section 4.4.

### 7.2.1 Phases for Case 1: FAST Trucks - Empty Load

For case 1, the data calculations resulted in three Erlang phases for the service time approximation.

Mean = 1.691344848
Variance = 1.139770081
$C^2 = 0.398430818$
Since $1/3 \leq 0.243086207 \leq 1/2$:

$k = 3$

$1 - p = 0.093628994$

$p = 0.906371006$

$\mu = 1.663021006$

### 7.2.2 Phases for Case 2: FAST Trucks - Laden

In the case of FAST trucks with loads, the calculations produce five Erlang phases for the inspection time approximation.

Mean = 3.016342667

Variance = 2.211676849

$C^2 = 0.243086207$

Since $1/5 \leq 0.243086207 \leq 1/4$:

$k = 5$

$1 - p = 0.052066998$

$p = 0.947933002$

$\mu = 1.58859007$

### 7.2.3 Phases for Case 3: Non-FAST Trucks - Empty Load

When the data for Non-FAST, empty trucks was used, the calculations yielded three Erlang phases.

Mean = 2.661128

Variance = 2.890008079

$C^2 = 0.408100877$

Since $1/3 \leq 0.243086207 \leq 1/2$:

$k = 3$

$1 - p = 0.106893149$

$p = 0.893106851$

$\mu = 1.047004767$

### 7.2.4 Phases for Case 4: Non-FAST Trucks - Laden

Finally in the last case, Non-FAST loaded truck data produced four Erlang phases for service time approximation.

Mean = 3.949039623

Variance = 4.272448282

$C^2 = 0.27396421$

Since $1/4 \leq 0.243086207 \leq 1/5$:

$k = 4$

$1 - p = 0.031389257$

$p = 0.968610743$

$\mu = 0.989058759.$

Notice that almost for each type of truck-load data, the calculations resulted in a different number of phases: 3, 5, 3, and 4. And using a Coxian k-phased approximation for service time, and recalling the queue description from 4.2, the new queue description would be:

$$[\ (M_{(t,\iota,1)}\ /\ GE_{(\iota,1)}\ /\ K_1)\ ,\ (M_{(t,\iota,2)}\ /\ GE_{(\iota,2)}\ /\ K_2)\ ] \tag{7.4}$$

With arrival rates being time dependent $t$, and truck loads being described as $\iota =$ *laden* or *empty*.

Now that the parameters are calculated by truck type and by load, the service times using a Coxian k-phased approximation can be implemented. In the simulation model, the approximation, and the DRM policy was coded using Visual Basic code that interacts with the Arena simulation code. The simulation model reflects the queue description in Equation 7.4.

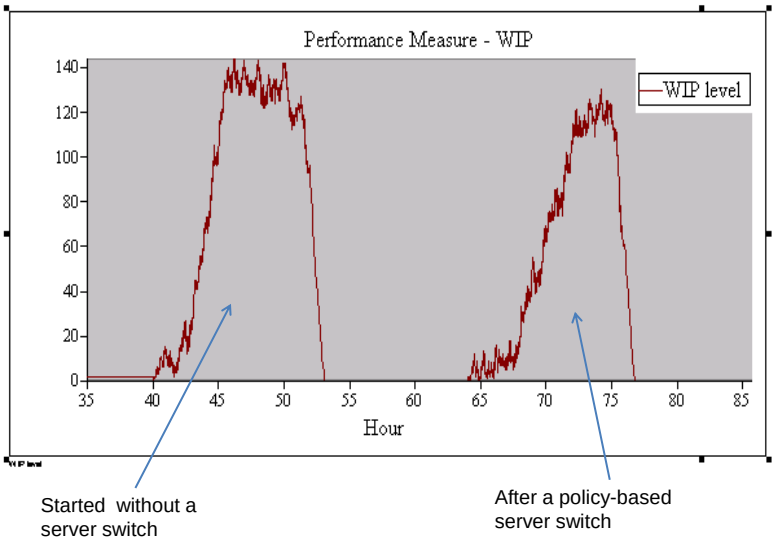



**Fig. 7.2.**: Work in Process Between a Single Work Period

## 7.3 New Base Case and DRM Comparison Approach

After implementation of the Coxian k-phased approximation, the focus turns on an evaluation strategy. First a new baseline is needed for comparison purposes. The analytical model provides the baseline comparison needed to confirm the simulation model. In Figure 7.2, the simulation model shows a realization of the work in process, which coincides with the behavior in the analytical model as seen in Figure 6.9.

Then the DRM model will be compared with the baseline case using three different arrival rates, light, medium and heavy, similar to the analysis of Section 4. This will allow a comparison of the current as-is situation with the improvements offered by DRM, when the Coxian approximation of service time is implemented without having to limit the MGE to two phases, no short queues, or separating the service times by load type $\iota$.

Afterwards, the analysis turns to the actual reallocation policy. Recall the reallocation policy in Equation 4.5, allows for setting a value of the threshold before requesting a reallocation. That is, the values for a reallocation from a FAST server

to a Non-FAST is set by $Th_{F2R}$, and the reallocation from a Non-FAST to FAST server is set by $Th_{F2R}$.

The expected result is given by the solution from the analytical model in Section 5.3, where the $\rho_k$ with higher utilization will draw the moveable server. In this case, the expectation is that the server moves to serving Non-FAST trucks, as seen in 5.3.3 because now there are four Non-FAST servers versus 2 FAST servers. Plus the server would stay there depending on the reallocation threshold, for a Non-FAST to FAST server switch, set by $Th_{F2R}$.

### 7.3.1 Stationary Versus Non-Stationary Policies

The current DRM strategy is to evaluate the system using non-stationary policies to decide when will a server switch will be called. In this case, a threshold value of queue size differential is set by the POE managers to determine the trigger for a server switch. This policy is non-stationary, however, there are many other policies, some stationary that may be defined by calling a server switch at a predetermined time of day. The question is not whether the DRM's non-stationary policies are better than stationary policies that do not require the tracking of the number in the queue, or the number in the small queues in front of the inspection stations, or the differences in queue size. This is because the DRM can be implemented with stationary policies. The question is whether stationary policies offer an alternative.

For analysis, consider in a two-queue server switch POE queueing system, as defined in Section 4.3, and with an objective of maintaining a maximum queue size differential, that is,

$$QS_t(NonFAST) \ > QS_t(FAST) + Th_{F2R}$$

for some $Th_{F2R}$ value. Similarly,

$$QS_t(FAST) \ > QS_t(NonFAST) + Th_{R2F}$$

for some other $Th_{R2F}$ value. The time dependent and changing nature of the arrival rates $\lambda_t$ and thus queue sizes, $QS_t(NonFAST)$ and $QS_t(NonFAST)$, ensures that the non-stationary policy will be able to call the server switch whenever necessary. On the other hand, a stationary policy may be effective today, but could be inadequate in another day with a different arrival realization of FAST and Non-FAST trucks, given the exponential nature of arrival rates, and the general service time distributions.

### 7.3.2 Results

Similarly to Section 5, the simulation model ran for each case 52 weeks, and was replicated 26 times. That translates to an approximated simulation time of one year, and replicated for 26 years. There were several computers used in running and debugging the simulation, but the same system was used to run and compare the results. The system was an Intel Core2 Duo (3.17 GHZ) using Windows 7 Enterprise edition, 64 bit, and 4 GB of RAM. The version of Arena was 12.0 CPR 9, licensed to Texas A&M University. The results were compiled by the model, and presented here as a summary.

**Table 7.1**: DRM and Coxian: Average Number of Trucks

| | Current as-is Model | | DRM and Coxian Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Light Load | 43.372 | 1.4176 | 38.496 | 1.1541 |
| Normal Load | 70.835 | 1.9817 | 63.219 | 1.6527 |
| Heavy Load | 155.26 | 13.047 | 105.85 | 4.5733 |

First, the new baseline model and the DRM model are compared, and this time, both have a Coxian k-phased approximation of the service time distribution. The results is as follows: Light Arrival Load case shows the smallest per average number

of trucks improvement. That is to be expected, however since there is a lighter arrival rate for all kinds of trucks. Noticed as well in Table 7.1 that the improvements, just like in the Markovian case, are more noticeable with higher arrival rates. That is, the case of a heavy arrival rate produces a more significant benefit in the use of the DRM, regardless of the type of service distribution.

**Table 7.2**: DRM and Coxian: Performance Improvement of Number of Trucks

| | Coxian | Percentage (%) | Markovian | Percentage (%) |
|---|---|---|---|---|
| Light Load | 4.876 | 11.24% | 4.62 | 12.9% |
| Normal Load | 7.616 | 10.75% | 18.81 | 29.7% |
| Heavy Load | 49.41 | 31.82% | 47.62 | 43.6% |

Also notice in Table 7.2 that the improvements are significantly lower when the Coxian k-phased approximation is used, instead of a Markovian service time. In fact, regardless of the arrival rate, and in agreement with the work by Cetin and List (2004), using a Markovian approximation for the service time underestimated the work in progress (Truck WIP) and cycle time (CT). This also produced an over-evaluation of the benefits by the DRM, whereas the Coxian approximation was at best, a more modest 32% with heavy arrival rates. Still very good, in the sense that with an increase in traffic, the benefits of DRM are more noticeable.

Table 7.3 presents the results for average overall cycle time. Again, here the results are also similar to those of the average number of trucks in the system in terms of percent improvement. Also notice that the CT improvements are in line with WIP, and as observed before, significantly better with higher arrival rates.

Similarly, Table 7.4 shows the following improvements in the CT. Notice that just like the WIP, the benefits of the Markovian case are not as noticeable, except for the heavy arrival rate case. In all cases, the calculated WIP and CT averages

**Table 7.3**: DRM and Coxian: Average Cycle Time with Coxian Approximation

| | Current as-is Model | | DRM and Coxian Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Light Load | 1.1489 | 0.03344 | 1.0146 | 0.02684 |
| Normal Load | 1.691 | 0.04306 | 1.5049 | 0.03437 |
| Heavy Load | 3.4217 | 0.28698 | 2.3103 | 0.09227 |

for the Coxian k-phased approximation case, do not overlap with a 95% confidence interval when comparing the baseline model with the dynamic reallocation model.

**Table 7.4**: DRM and Coxian: Performance Improvement of Average Cycle Time

| | Coxian change | Percentage | Markovian change | Percentage |
|---|---|---|---|---|
| Light Load | 0.1343 | 11.69 % | 0.1259 | 13.29% |
| Normal Load | 0.1861 | 11.01% | 27.07 | 30.05% |
| Heavy Load | 1.1114 | 32.48 % | 1.0513 | 44.02% |

Table 7.4 compares the differences from Table 7.3 and calculates a percentage difference. For example, from Table 7.3 under Heavy Load, $3.4217 - 2.3103 = 1.1114$. And as a percentage change, the calculation is, $1.1114/3.4217 = 0.3248$. Also notice that the comparative values for the Markovian case come from Table 5.4 and Table 5.5, but have been converted into hours. For example, in Table 5.4 under Light Load, the improvement was $0.94719 - 0.82128 = 0.12591$, and the percentage improvement come directly from Table 5.5, although that table showed the actual improvements in minutes for context.

One more performance measure is of importance, and that is the FAST truck CT. The reason is that there is a fixed number of servers, and for the benefits gained in Non-FAST CT, and WIP, there has to be some loss in FAST CT, and WIP. Notice as well that as Non-FAST $\rho_2$ decreases, FAST $\rho_1$ increases. Tables 7.5, and 7.6 illustrate the similar effect seen when the Markovian service time was analyzed. And as seen in the simplified case, the actual time delay increase is minimal.

**Table 7.5**: DRM and Coxian: Average FAST Cycle Time

| | Current as-is Model | | DRM and Coxian Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Light Load | 0.26462 | 0.02009 | 0.80266 | 0.10379 |
| Normal Load | 0.33182 | 0.02314 | 2.1917 | 0.1124 |
| Heavy Load | 0.54326 | 0.04086 | 3.7134 | 0.22171 |

**Table 7.6**: DRM and Coxian: Average FAST Cycle Time Increase

| | Coxian | Markovian |
|---|---|---|
| Light Load | -0.54 | –0.5 |
| Normal Load | –1.86 | –1.81 |
| Heavy Load | –3.17 | –2.93 |

The conclusion from this simulation exercise is that the Coxian k-phase approximation model seems to performs better than the Markovian assumption for service time. The benefits of the DRM are in any case significant, particularly, when the traffic increases. All performance measures show a similar result, and since there is a fixed number of servers, the cost is a small increase in the FAST truck CT.

## 7.4 Increased Use of Secondary Inspection for Variance Reduction

Lastly one additional improvement to the border crossing process can be achieved if the variance in the inspection times is addressed. Normally, in Lean Manufacturing principles address the reduction of waste. In addition to reducing waste, another principle is attacking the variance, and reducing it. These principles have brought major significant savings and process improvements in many industries throughout the world.

The border crossing process should not be excluded from such benefits when some of these principles are applied. And although this section only discusses a non-intrusive of security and brief method to reduce variance, the area should be subject of further research. To clarify, the objective should not be to time CBP Agents, to make sure that their inspections are below a certain threshold. Instead, this method proposes an increased use of secondary inspection stations to evaluate and inspect any trucks that are taking a significant time in primary inspection. The data collected in Appendix B shows some significant inspection time outliers of up to 14 minutes for a single inspection. Eliminating the few outliers would positively affect the throughput, and reduce the truck WIP and CT.

*Variance reduction procedure:* Once an inspection begins, normal inspection and security procedures take place for the truck. If the inspection takes longer than some threshold, say $\psi(\bullet)$, the truck would be automatically be sent to secondary inspection to complete inspection.

Implementation of the variance reduction procedure can take place in most POEs, where commercial traffic moves. Expected benefits of this additional procedure is observed in the following proposition.

**Proposition 7.4.1** *Use of secondary inspections stations after a threshold of* $\psi(\bullet)$ *time, reduces variance.*

**Proof** Let $P\{X \leq x\} = F(x)$ a be distribution function.

Now, a truncated distribution function is defined as:

$$P\{Y \leq x\} = \begin{cases} F(x) & \text{if } x < \psi(\bullet) \\ 1 & \text{if } x \geq \psi(\bullet) \end{cases}$$

The proposition is equivalent to $Var[X] > Var[Y]$ where $X = Y + Z$, and $Z$ is a random function of $x$, when $x \geq \psi(\bullet)$, otherwise 0.

Therefore,

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= E[(Y+Z)^2] - (E[Y+Z])^2 \\ &= E[Y^2 + 2YZ + Z^2] - (E[Y] + E[Z])^2 \text{ (by linearity)} \\ &= E[Y^2] + E[2YZ] + E[Z^2] - (E[Y])^2 - 2E[Y]E[Z] - (E[Z])^2 \\ &= E[Y^2] - (E[Y])^2 + E[2YZ] + E[Z^2] - 2E[Y]E[Z] - (E[Z])^2 \\ &= Var[Y] + E[2YZ] + E[Z^2] - 2E[Y]E[Z] - (E[Z])^2 \end{aligned}$$

Therefore, let $Tail(X) = E[2YZ] + E[Z^2] - 2E[Y]E[Z] - (E[Z])^2$

and we have

$$Var[X] = Var[Y] + Tail(X)$$

so,

$$Var[X] \geq Var[Y]$$

and if $Z \neq 0$ then

$$Var[X] > Var[Y]$$

■

### 7.4.1 Truncating the Probability Distribution

Consider an exponential distribution, for example the one in Figure 7.3 with a rate of 3.6 minutes. If the tail of the distribution were truncated and removed, say

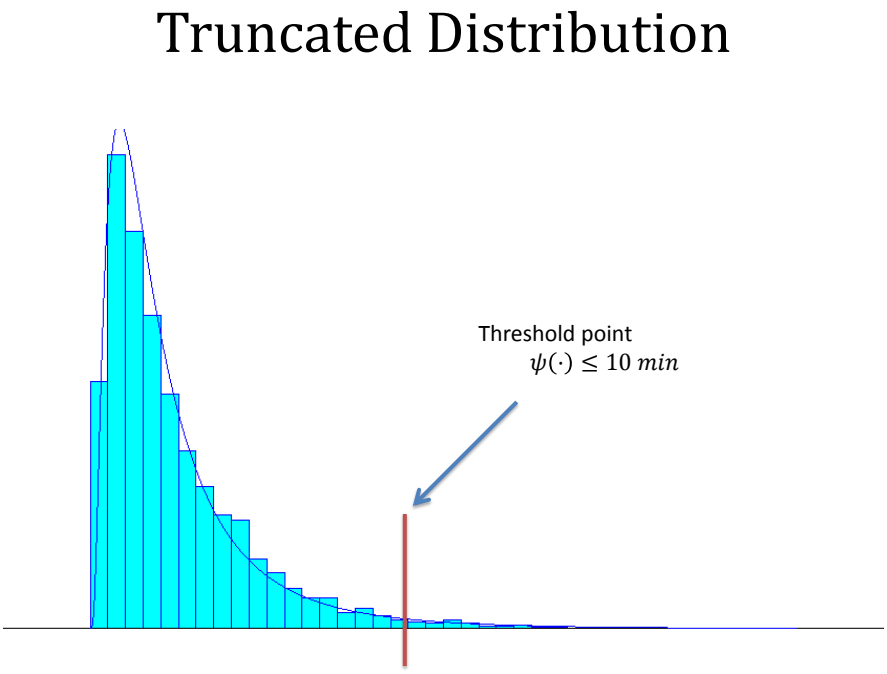



**Fig. 7.3.**: Example of an Exponential Distribution with Threshold $\psi(\bullet)$

at 10 minutes, then variance would be reduced, and by consequence, there would be an improvement in the mean.

For a numerical example, a shifted exponential is created with mean 3.6 and shift of 0.5 minutes. The generated data is sorted and all the points greater than 10 are removed, that is $\psi(\bullet) \leq 10min$. Then the data is introduced to the data fitting software of Arena, and the results are: mean of 3.35, and variance of 2.31. The reduction of the mean is not the objective, notice that the reduction in the variance is where the most benefit is received. The actual fitted data is seen in Figure 7.4, and the results are as follows:

```
 Distribution Summary

Distribution:        Erlang
Expression:        ERLA(1.68, 2)
```

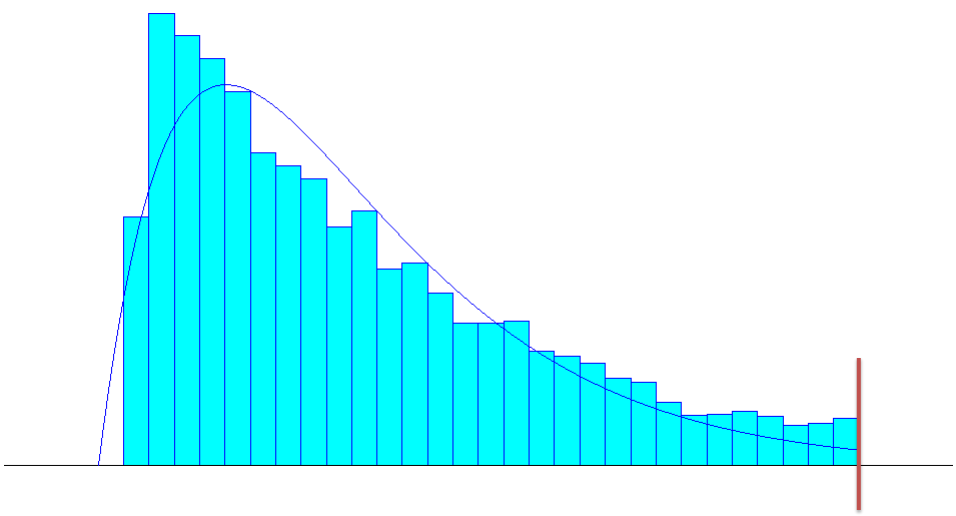



**Fig. 7.4.**: Example of a Truncated Exponential Distribution with $\psi(\bullet) \leq 10$ min.

```
Square Error:        0.001906

Chi Square Test
 Number of intervals        = 30
 Degrees of freedom         = 27
 Test Statistic             = 323
 Corresponding p-value        < 0.005

Kolmogorov-Smirnov Test
 Test Statistic             = 0.0366
 Corresponding p-value        < 0.01

       Data Summary
```

```
Number of Data Points        = 4645
Min Data Value               = 0.5
Max Data Value               = 10
Sample Mean                  = 3.35
Sample Std Dev               = 2.31

       Histogram Summary

Histogram Range            = 0 to 10
Number of Intervals        = 30
```

### 7.4.2 Results

As a final step, the variance reduction policy is incorporated to the DRM with a coxian k-phased approximation for service time model. Table 7.7 contains the results obtained when the model was run, comparing only the baseline model and a normal arrival rate. In this case, the light and heavy arrival rates were omitted since the impact of those arrival rates has already been established.

**Table 7.7**: DRM, Coxian and Variance Reduction: Normal Arrival Results

| | Current as-is Model | | DRM, Coxian and VR Model | |
|---|---|---|---|---|
| | Average | Half width | Average | Half width |
| Average Trucks | 70.835 | 1.9817 | 59.891 | 1.7147 |
| Average CT | 1.691 | 0.04306 | 1.4253 | 0.03703 |
| NonFAST CT | 1.9365 | 0.05003 | 1.3695 | 0.0283 |
| FAST CT | 0.33182 | 0.02314 | 1.7307 | 0.12948 |

Notice that the improvement in average number of trucks in the system, and the CT is better than the DRM model with Coxian service time alone, and from Tables 7.1, and 7.3 that WIP was 63.219, and CT was 1.5049. This was to be expected since, Proposition 7.4.1 proves the reduction in variance.

## 7.5 Summary and Conclusions

In this section, the research turned to the implementation and analysis of the DRM model with a Coxian k-phased approximation for service times. After the development of the analytical model in Section 6, it is now used as a baseline case or current as in situation at the POE. This allowed a comparison of the baseline scenario with the DRM in a simulation model using different arrival rates, i.e. high, medium, and low. The effectiveness of the DRM was evaluated using a variety of performance measures including CT and WIP for both FAST and Non-FAST trucks. In terms of Cycle Time improvement, results showed a more "realistic" performance measure improvements than the Markovian assumption, as seen in Table 7.4 and Table 7.6 with improvements up to 32% with a heavy arrival load. Similar results were seen in Table 7.2 regarding WIP.

In the later sections a variance reduction policy was presented, that proposes an increase use of secondary inspections. This policy was proven to decrease variance, and positively affect the system WIP and CT. Numerical results in Table 7.7 showed that the DRM / Coxian phased approximation of service times model modified to include a variance reduction procedure was also effective.

The research in this capstone section presents several key conclusions:

- The DRM can use stationary and non-stationary policies, but non-stationary policies are more flexible for time dependent arrival rates.
- Exponential service times for service times over estimate the benefits of DRM. This result coincides with the results in the work of Cetin and List (2004) where

he concluded that assuming Markovian service times would lead to significant discrepancies.

- The Coxian k-phased approximation produced a more realistic results for the DRM results. In hindsight, although the results are more modest, they can be justified.

- Variance reduction by using secondary inspections stations after a threshold of $\psi(\bullet)$ time, reduces overall system variance, and consequently improves performance measures.

# 8. CONCLUSIONS AND FUTURE RESEARCH

The motivation for this work arises from the daily symptoms of long border wait times at the POEs in cities between the U.S. and North American trading countries like Canada and Mexico. It also arises from the desire to improve system performance of traffic at border crossings, without increasing costs or modifying any security procedures. Although traffic crossing the border consists of pedestrian, private vehicles, and commercial traffic, the focus of this research targeted commercial traffic because it can be easily quantified with an economic impact, and thus serves as justification. The border crossing process for commercial vehicles basically consists of queues that split into multiple small queues of a fixed number of inspection stations that only open and close at certain hours. The trucks wait in the queues to be processed and inspected, and then finally cross into the U.S. at many of the land-based POEs.

## 8.1 Summary

This dissertation has accomplished the research objectives presented at the outset. The main objective was to use queueing theory on the border crossing process, and develop a methodology that could decrease cycle time, or border wait times at the POE, without increasing costs or affecting current security procedures. After presenting the economic significance of the international commerce shipped by commercial trucks through POEs in Section 1, and describing the border crossing process in Section 2, a non-stationary Dynamic Reallocation Methodology (DRM) for terminating queueing systems, such as the POEs, was developed in Section 4. Empirical data was also collected regarding the inspection or service times and evaluated with best fitted distributions.

The second objective was to analyze and verify the results of this methodology. Section 5 started the mathematical analysis with the assumption that the POE inspection time, i.e. service time, was Markovian in nature. After solving the problem

analytically, the results confirmed that the moveable server would tend to go to the system which had a higher utilization factor of truck Type $\wp$, $\rho_{\wp}$. Simulation results of the DRM with Markovian service times confirmed the findings as well.

Then in Section 6, the inspection or service time was considered to follow a general distribution, and was approximated using Coxian or MGE distributions described by Curry and Feldman (2011) and Altiok (1996). The section described the benefits of an analytical model, but had to assume out some POE modeling challenges, such as the split to small queues. However, transient behavior results of the analytical model coincided with empirical data (Appendix B), and current literature (Battelle/Texas Transportation Institute, 2008). Finally in Section 7, the Coxian approximation of service times was implemented in the DRM simulation model, to confirm the effectiveness of the DRM in a more realistic environment. The section also compared the current as-is situation, with the DRM simulation model that included the Coxian approximation. The results were positive for the methodology, with a reduction in Cycle Times of over one hour, as seen in Table 7.3. A comparison of the Markovian assumption and the Coxian approximation showed a more modest improvement in terms of percent changes, which coincides with the research of Cetin and List (2004) when he concluded that assuming Markovian service times would lead to significant discrepancies. Although an analytical model is the preferred method by most researchers, the main reason for simulation, as Kelton et al. said, is its ability to handle complicated models of complicated systems. Lastly, an add-on policy to the DRM, or the current system, was a variance reduction policy. This policy is a strategy to remove outliers in the inspection process. However, the policy requires the use of secondary inspection station, which would affect cost.

## 8.2 Conclusions

The following are highlights of the research work:

- The border crossing process is a significant part of the nations's economic engine.
- Traffic at the POEs is expected to grow.
- The commercial border crossing process is a multi step, secure sensitive process.
- Research in the border crossing process has been the target mostly of public research institutions, which for the most part have focused on improving the process itself, or assessing and implementing some form of technology.
- Basic queueing theory research has only been recently applied to this process.
- A Dynamic Reallocation Methodology (DRM) was developed that decrease cycle time, or border wait times at the POE, without increasing costs or affecting current security procedures.
- Analyzing the DRM model with a Markovian assumption of service time, indicates that the system observes an overall performance improvement for all load levels.
- The benefits are even greater at heavy load volumes with wait time reductions of almost 45%, with only marginal increase in FAST CT.
- Field observations indicate that the service time is not Markovian.
- The analytical model can now approximate the general service time distribution of the border crossing process at the POE, using a Coxian or MGE distribution as shown by Curry and Feldman (2011) and Altiok (1996).
- The analytical model results provide complete distribution information for each time step in the analysis.

- The transient analysis of the analytical model showed the same behavior that was observed in the field, and in the literature for the expected number of trucks in the system $E[N(t)]$.

- The analytical model can now be used in tandem with a simulation model for complicated analysis, such as the implementation of the DRM.

- The DRM can use stationary and non-stationary policies, but non-stationary policies are more flexible for time dependent arrival rates.

- When the Coxian k-phased approximation was implemented, the results for DRM produced a more realistic results, and in the case of heavy volume, the expected benefits are up to 32%.

- A more modest result coincides with the work of Cetin and List (2004) when he concluded that assuming Markovian Service times would lead to significant discrepancies.

- Analyzing the DRM model with a a Coxian approximation for service times also observed an overall system performance improvement for all arrival load levels.

- Variance reduction by using secondary inspections stations after a threshold of $\psi(\bullet)$ time, reduces overall system variance, and consequently improves performance measures.

## 8.3 Future Research and Applications

The following are two areas of particular interest for future research in border crossing, and related homeland security issues. The first one is to optimize the DRM with input from the appropriate stakeholders, that is CBP and DHS. Another option is to explore the comparison of the Markovian model of teh POE with the Coxian

approximation for POE service time which includes the small queues. The last consideration is to explore other performance measures, in particular those dealing with security, specially since an extension and application of the DRM can be to apply it in maritime POEs.

### 8.3.1 Optimize the DRM

This opportunity requires the input from appropriate homeland security-related agencies. The objective here is to go beyond the proof-of-concept, and through meetings and interviews, determine:

- Priorities and wait time limitations for FAST processing
- Incentives for increased adoption of the FAST pogrom, particularly for small shipping companies.
- Security parameters regarding the use of secondary inspections

This collaboration can provide input to the DRM, and allow the system to become a "smart system." The DRM can potentially aid POE administrators in making decision about POE dynamics. The benefits include:

- Increased the utilization of current resources
- Manage staffing levels effectively
- Improve system performance, i.e. truck wait time

### 8.3.2 Future Comparisons for the Coxian Approximation Using Generator Matrices

As analyzed in Section 6, the Generator Matrix becomes a very large 3-dimensional "matrix cube" that is required to hold all the probabilities $P_n(x)$ for the system. An

extension of the research is to compare other performance measures, for example expected WIP in the system, of a Markovian-assumed service time model and a Coxian-approximated service time model of the POE system.

Although the objective would not be to evaluate the DRM, but rather, the effectiveness of the Coxian approximation in POE environments, the results could give insight into better modeling strategies for service environments, both analytically and with simulation. Please refer to Appendix C for the generator matrices in a Markovian POE model that includes the small queues in front of the inspection stations.

### 8.3.3 Security Performance Measures

Another quick benefit includes the implementation of this research into maritime POEs. At the docks, where cargo ships come to unload, most the goods copme in containers as well. Once the ship arrives, and the unloading begins, a process very similar to current land-based, POEs begins. The benefits of the DRM could also be achieved at the maritime POE. Including the decrease of cycle time, without increasing costs or affecting current security procedures.

However, security is becoming more and more important, as terrorists and other national threats can affect the well being of the population. The National Strategy for Homeland Security U.S. Department of Homeland Security (2007) states that "Homeland Security is a concerted national effort to prevent terrorist attacks within the United States, reduce America's vulnerability to terrorism, and minimize the damage and recover from attacks that do occur." Included in the document is also the goal for Homeland security, which is "to prevent and disrupt terrorist attacks; protect the American people, critical infrastructure, and key resources; and respond to and recover from incidents that do occur." And although 100% security level is inherent in the goal, it is also unattainable. The document mentions that "despite our best deterrent and mitigation efforts, terrorist attacks and natural disasters will happen,

and we must work to minimize the consequences of their occurrence." But a detailed definition for security threat is not available, and other sources are ambiguous to help create a way to measure security.

Evaluating whether a security procedure is effective becomes an issue of tallying the number of successes obtained, compared to the overall set of security threats missed. Yet this is where the difficulty lies, because the true and accurate number attempts to enter the U.S. with inadmissible cargo in a given day is not known. Given the difficulty to objectively and accurately assess security, methods or procedures to compare a security measure versus its cost also becomes difficult. However, Statistical Decision Theory provides ways to test hypothesis and assess the significance of making mistakes when rejecting a shipment when it is admissible. This is based on the Type II error. A Type II error occurs when the researcher fails to reject a null hypothesis that is false.

The proposal is to define a new key performance measure: Rate of False Positives (RoFP). In terms of security at the POEs, the null and alternative hypothesis are:

> ($H_0$): The individual(s), and cargo contained are admissible into the United States.
>
> ($H_a$): The individual(s), and cargo contained are NOT admissible into the United States.

A False positive occurs at a secondary inspection. Here is where the null hypothesis, i.e. the individual and cargo is admissible, was rejected. But after further examination at the secondary inspection, it was determined that the individual and the cargo are indeed admissible.

It is important to note that the set of attempts,

$$\mathcal{A} = \left\{ \begin{array}{ll} a: & a \text{ is the number of attempts to bring introduce inadmisible} \\ & \text{items into the U.S. through the POE over time } t \end{array} \right\}$$

to bring into the U.S. inadmissible people or cargo across the border POEs over any time period $t$, i.e. daily, weekly, monthly or yearly, is unknown.

But the set of all attempts of everyone to cross the border

$$\mathcal{C} = \left\{ \begin{array}{ll} c: & c \text{ is the toal number of attempts to enter the U.S.} \\ & \text{through the POE over time } t \end{array} \right\}$$

is known, and the set of False Positives can be tallied. Thus, the information that is needed to develop a key performance measure is available, and can be defined as the rate of false positives in a given time basis (RoFP).

The main benefit of this performance measure is its usefulness in comparing the cost of current security procedures, with the economic impact of increased delays and lost productivity to the overall economy, both locally and for the nations involved. With this key performance measure, a different measure can be used to evaluate the effective use of secondary inspection, and another can be developed to assess the pre-screening process. All the performance measures can be used to assess the effectiveness of security, but without knowing exactly the rate of inadmissible attempts, the next best step would be to estimate it. With better information from the DHS, tests can be setup to improve security of the system as a whole.

## 8.4 Ending Remarks

In this research journey, there were many challenges and difficulties, but personal objectives did not change, which were to contribute to the body of knowledge with this dissertation, and to contribute and improve the border crossing environment. The hope was to accomplish both, which have been dearly needed for over 20 years. Thank you for reading this dissertation.

# REFERENCES


Adusumilli, K. M., J. J. Hasenbein. 2010. Dynamic admission and service rate control of a queue. *Queueing Systems* **66**(2) 131–154.

Altiok, T. 1996. *Performance Analysis of Manufacturing Systems*. Springer-Verlag, NY.

Ashur, S., J. Weissmann, S. Perez, A. J. Weissmann. 2001. Traffic simulation at international ports of entry - El Paso-Mexico case study. *Multimodal and Marine Freight Transportation Issues* (1763) 48–56. Transportation Research Record.

Askin, R. G., C. R. Standridge. 1993. *Modeling and analysis of manufacturing systems*. Wiley. URL http://books.google.com/books?id=L-xTAAAAMAAJ.

Ata, B. 2006. Dynamic control of a multiclass queue with thin arrival streams. *Operations Research* **54**(5) 876–892.

Battelle/Texas Transportation Institute. 2008. Commercial vehicle crossings at the Bridge of the Americas, El Paso, Texas. Tech. rep., The Texas Transportation Institue. http://tti.tamu.edu/documents/TTI-2008-1.pdf Retrieved July 21, 2010.

Bell, C. E. 1980. Optimal operation of an m-m-2 queue with removable servers. *Operations Research* **28**(5) 1189–1204.

Bing Maps. 2011. Map of zip code 79905 - bird's eye view. Web Publication. http://www.bing.com/maps/ Retreived on January 7, 2011.

Bracchi, P., B. Cukic, V. Cortellessa. 2006. Modeling the performance of border inspections with electronic travel documents. *ISSRE 2006:17th International Symposium on Software Reliability Engineering, Proceedings* (396) 237–244.

Bradbury, S. L. 2010. An assessment of the free and secure trade (FAST) program along the canada-us border. *Transport Policy* **17**(6) 367–380.

Cetin, M., G. F. List. 2004. Investigating effects of correlated service times on system performance - implications in border crossing operations. *Data and Information Technology* (1870) 70–76. Transportation Research Record.

Chavez, J. 2004. Localized effects of globalization: the case of Ciudad Juarez, Chihuahua, Mexico. *Urban Geography* **25**(2) 120–138.

Chiou, L., E. Muehlegger. 2008. Crossing the line: Direct estimation of cross-border cigarette sales and the effect on tax revenue. *B E Journal of Economic Analysis & Policy* **8**(1).

Choudhury, G. L., D. M. Lucantoni, W. Whitt. 1997. Numerical solution of piecewise-stationary m-t/g(t)/1 queues. *Operations Research* **45**(3) 451–463.

Chow, G. 2006. Applying supply chain logistics modeling to border security and efficiency. Annual Conference of the Transportation Association of Canada, Bureau of Intelligent Transportation Systems and Freight Security (BITSAFS). Also available online at http://www.tac-atc.ca/english/resourcecentre/readingroom/conference/conf2006/docs/pt005/chow.pdf.

Çınlar, E. 1975. *Introduction to stochastic processes*. Prentice-Hall. URL http://books.google.com/books?id=UNZQAAAAMAAJ.

Curry, G. L., R. M. Feldman. 2011. *Manufacturing Systems Modeling and Analysis*. 2nd ed. Springer-Verlag, Berlin.

Deloitte. 2010. 2010 global manufacturing competitiveness index. Tech. rep., Deloitte Touche Tohmatsu and the U. S. Council on Competitiveness. http://www.deloitte.com/assets/Dcom-Global/Local%20Assets/Documents/Manufacturing/DTT_Global_Manufacturing_Competiveness_Index_6_23_2010.pdf Retreived on March 27, 2011.

Erlang, A. K. 1909. The Theory of Probabilities and Telephone Conversations. *Nyt Tidsskrift for Matematik* **20**(B) 33–39.

Feldman, R. M., C. Valdez-Flores. 2010. *Applied Probability and Stochastic Processes*. 2nd ed. Springer-Verlag, Berlin.

Ferris, J. S. 2000. The determinants of cross border shopping: Implications for tax revenues and institutional change. *National Tax Journal* **53**(4) 801–824. Part 1 393WM.

FHWA Office of Freight Management and Operations. 2010. Commercial vehicle travel time and delay at u.s. border crossings. http://ops.fhwa.dot.gov/freight/freight_news/travel_time/travel_time_delay.htm, Accesed January 17, 2011.

Google Maps. 2011. Map of zip code 79905. Web Publication. http://maps.google.com/ Retreived on January 7, 2011.

Gross, D., J. F. Shortle, J. M. Thompson, C. M. Harris. 2008. *Fundamentals of queueing theory*. 4th ed. Wiley series in probability and statistics, Wiley, Hoboken, N.J.

Halvey, R. 2003. Border congestion, air quality, and commerce. Ian J. Bateman, Jan J. Batema, Linda Fernandez, Richard T. Carson, eds., *Both Sides of the Border*, *The Economics of Non-Market Goods and Resources*, vol. 2. Springer Netherlands, 281–304. http://dx.doi.org/10.1007/0-306-47961-3_14.

Haralambides, H. E., M. P. Londono-Kent. 2001. Supply chain bottlenecks: Border crossing inefficiencies between Mexico and the United States. *International Journal of Transport Economics* **XXXI**(2) 48–56.

Haughton, M., K. P. Sapna Isotupa. 2012. Scheduling commercial vehicle queues at a Canada/US border crossing. *Transportation Research Part E: Logistics and Transportation Review* **48**(1) 190 – 201. doi:10.1016/j.tre.2011.07.008. URL http://www.sciencedirect.com/science/article/pii/S1366554511001037. Select

Papers from the 19th International Symposium on Transportation and Traffic Theory.

Ignall, E. J., P. Kolesar, W. E. Walker. 1978. Using simulation to develop and validate analytic models - some case studies. *Operations Research* **26**(2) 237–253.

INEGI. 2011. Número de establecimientos activos según entidades federativas y municipios. Tech. rep., Instituto Nacional de Estadística y Geografía. http://dgcnesyp.inegi.org.mx/cgi-win/bdieintsi.exe/NIVR2501100070 Retreived on April 7, 2011.

Kaminsky, P., D. Simchi-Levi. 2003. Production and distribution lot sizing in a two stage supply chain. *IIE Transactions* **35**(11) 1065–1075.

Kelton, W. D., R. P. Sadowski, D. A. Sadowski. 2002. *Simulation with Arena*. McGraw-Hill series in industrial engineering and management science, McGraw-Hill. URL http://books.google.com/books?id=41eqQgAACAAJ.

Khoshons, M. K., C. C. Lim, T. Sayed. 2006. Simulation and evaluation of international border crossing clearance systems - a canadian case study. *Freight Analysis, Evaluation, and Modeling; Truck Transportation* (1966) 1–9. Transportation Research Record-Series.

Law, A. M., W. D. Kelton. 1991. *Simulation modeling and analysis*. McGraw-Hill series in industrial engineering and management science, McGraw-Hill. URL http://books.google.com/books?id=jeFQAAAAMAAJ.

Leung, S. C. H., Y. Wu, K. K. Lai. 2006. Cross-border logistics with fleet management: A goal programming approach. *Computers & Industrial Engineering* **50**(3) 263–272.

Li, K. W., J. Higginson, D. Friesen, J. K. Levy. 2005. Windsor-Detroit border crossing problem: Conflict analysis of the Schwartz report. *International Conference on Systems, Man and Cybernetics, Vol 1-4, Proceedings* (3957) 1132–1137.

Lodree, E., W. S. Jang, C. M. Klein. 2004. Minimizing response time in a two-stage supply chain system with variable lead time and stochastic demand. *International Journal of Production Research* **42**(11) 2263–2278.

MacLachlan, I., A. G. Aguilar. 1998. Maquiladora myths: Locational and structural change in Mexico's export manufacturing industry. *Professional Geographer* **50**(3) 315–331.

Margolius, B. H. 2005. Transient solution to the time-dependent multiserver poisson queue. *Journal of Applied Probability* **42**(3) 766–777.

Margolius, B. H. 2007. Transient and periodic solution to the time-inhomogeneous quasi-birth death process. *Queueing Systems* **56**(3-4) 183–194.

McCord, M. R., P. K. Goel, C. Brooks, P. Kapat, R. Wallace, H. Dong, D. E. Keefauver. 2010. Documenting truck activity times at international border crossings using redesigned geofences and existing onboard systems. *Transportation Research Record* (2162) 81–89.

Mexico Business Center. 2010. Project smart border 2010. Tech. rep., The San Diego Regional Chamber of Commerce. http://www.sdchamber-members.org/documents/ProjectSmartBorder20102.pdf Retrieved December 21, 2010.

Molina, E. C. 1922. The theory of probabilities applied to telephone trunking problems. *Bell System Technical Journal* **1**(2) 69–81.

Nahmias, S. 2008. *Production and operations analysis*. Irwin/McGraw-Hill series in operations and decision sciences, McGraw-Hill/Irwin. URL http://books.google.com/books?id=OfioPwAACAAJ.

Nicol, D. M., R. Tsang, H. Ammerlahn, M. Johnson. 2006. Detection of nuclear material at border crossings using motion correlation. *Proceedings of the 2006 Winter Simulation Conference, Vols 1-5* (2307) 536–544.

Nozick, L. K., G. F. List, M. A. Turnquist, T. L. Wu. 1998. Potential effects of advanced technologies at commercial border crossings. *Freight Transportation* (1613) 88–95. Transportation Research Record.

Ojah, M. I., J. C. Villa, W. R. Stockton, D. M. Luskin, R. Harrison. 2002. Truck transportation through border ports of entry: Analysis of coordination systems. Tech. Rep. TX-01/50-1XXA3038, Texas Transportation Institute, The Texas A&M University System, College Station, Texas 77843-3135. Also available on http://bordercross.tamu.edu/cpoe/shareddocs/50-1XXA3038.pdf.

Ong, K. L., M. R. Taaffe. 1988. Approximating nonstationary ph(t)/ph(t)/1/c queuing-systems. *Mathematics and Computers in Simulation* **30**(5) 441–452.

Phillips, D. T., J. F. Ayala, E. M. Kozak. 1999. Statistical process control for job shop and sustainment operations. *Industrial Engineering Solutions '99 Conference, Proceedings* (336) 110–116.

Rajbhandari, R., J. C. Villa, R. Aldrete-Sanchez. 2009. Expansion of the border crossing information system. Tech. rep., University Transportation Center for Mobility - Texas Transportation Institute, 3135 TAMU College Station, Texas 77843-3135. http://utcm.tamu.edu/publications/final_reports/Villa_08-30-15.pdf Retrieved July 7, 2010.

REDCO. 2010. El Paso/Juarez borderplex: Business barometers 2010. Tech. rep., El Paso Regional Economic Development Corporation. http://www.elpasoredco.org/files/FINALBar_2010.pdf Retreived on November 7, 2010.

RITA. 2011. North American Transborder Freight Data: Query Detailed Statistics. Web database, U.S. Department of Transportation, Research and Innovative Technology Administration (RITA), 1200 New Jersey Avenue, SE Washington,

DC 20590.
http://www.bts.gov/programs/international/transborder/TBDR_QA.html
Retrived April 27, 2011.

Texas A&M International University. 2011. Border trade data: Border crossings. Web database, Texas Center for Border Economic and Enterprise Development, 5201 University Boulevard Laredo, TX. 78041-1900. http://texascenter.tamiu.edu/texcen_services/truck_crossings.asp?framepg=datatruck Retrieved April 2, 2011.

Turnquist, M. A., C. Rawls. 2010. Multimodal network analysis and vulnerability assessment of US-Canadian trade in Lake Erie corridor. *Transportation Research Record* (2168) 9–16.

U.S. Census. 2011. Top trading partners - surplus, deficit, total trade. Tech. rep., Foreign Trade. http://www.census.gov/foreign-trade/top/ Retreived on April 17, 2011.

U.S. Department of Homeland Security. 2003. Free and secure trade (FAST) implementation on the US/Mexico border. Press Release. http://www.dhs.gov/xnews/releases/press_release_0309.shtm Retreived on August 7, 2009.

U.S. Department of Homeland Security. 2007. National strategy for homeland security. Tech. rep., Homeland Security Council. http://www.dhs.gov/xlibrary/assets/nat_strat_homelandsecurity_2007.pdf Retreived on July 21, 2009.

U.S. Department of Homeland Security. 2011a. Customs and border protection: Border wait times. Web Publication. http://apps.cbp.gov/bwt/ Retreived on January 2, 2011.

U.S. Department of Homeland Security. 2011b. Customs and border protection: Fast - free and secure trade program. Web Publication. http://www.cbp.gov/xp/cgov/trade/cargo_security/ctpat/fast/ Retreived on December 7, 2010.

U.S. Department of Homeland Security. 2011c. Locate a port of entry - air, land, or sea. Web Publication. http://cbp.dhs.gov/xp/cgov/toolbox/ports/ Retreived on April 27, 2011.

U.S. Department of Homeland Security. 2011d. National terrorism advisory system. Web Publication. http://www.dhs.gov/files/programs/ntas.shtm Retreived on May 1, 2011.

Villa, J. C. 2006. Status of the US-Mexico commercial border crossing process - analysis of recent studies and research. *Freight Analysis, Evaluation, and Modeling; Truck Transportation* (1966) 10–15. Transportation Research Record-Series.

Villegas, H., P. L. Gurian, J. M. Heyman, A. Mata, R. Falcone, E. Ostapowicz, S. Wilrigs, M. Petragnani, E. Eisele. 2006. Trade-offs between security and inspection capacity - policy options for land border ports of entry. *Security 2006* (1942) 16–22. Transportation Research Record.

White, C. C. III. 2007. Global supply chains & logistics: Trends & research directions. Georgia Institute of Technology. Presentation at Texas A&M University on December 17, 2007.

Whitt, W. 2007. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics* **54**(5) 476–484.

Willis, H. H., D. S. Ortiz. 2004. Evaluating the security of the global containerized supply chain. Tech. rep., RAND Corporation, Santa Monica, CA. Also available at http://www.rand.org/pubs/technical_reports/TR214.

Winston, W.L., J.B. Goldberg. 2004. *Operations research: applications and algorithms*. Thomson Brooks/Cole. URL http://books.google.com/books?id=tg5DAQAAIAAJ.

WordPress. 2010. Women's analysis of the U.S./Mexico border: Border violence. Tech. rep., La Mujer Obrera. http://lamujerobrera.wordpress.com/2010/11/08/border-violence/ Retreived on April 7, 2011.

Zhang, Z. G. 2009. Performance analysis of a queue with congestion-based staffing policy. *Management Science* **55**(2) 240–251.

Zietsman, J., J. C. Villa, T. L. Forrest, J. M. Store. 2006. Estimating truck emissions at the El Paso - Ciudad Juarez border. National Urban Freight Conference 2006. URL http://www.metrans.org/nuf/documents/Zietsman.pdf. Accesed on June 10, 2009.

# APPENDIX A

**DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING**

Dwight Look College of Engineering

TEXAS A&M UNIVERSITY

Dr. Guy L. Curry, Committee Chair for
Hiram Moya, PhD Candidate
Industrial & Systems Engineering Department
Texas A&M University
College Station, TX. 77843-3131

March 11, 2010

Re: Data gathering for scholarly dissertation

Department of Homeland Security and/or
Department of Public Safety

Dear Agent,

Hiram Moya is my student and a PhD candidate in the Department of Industrial & Systems Engineering here at Texas A&M University, and he is conducting research in modeling and analysis of the commercial border crossing process using a queuing systems approach.

Hiram is also supported by a research and education grant from the Department of Homeland Security (DHS), through the National Center for Border Security and Immigration from the University of Texas at El Paso, in his role as Graduate Research Assistant at the Texas A&M Center of Excellence Program for Border Security. The Program Director is Dr. Don Phillips, Chevron Professor from the department and also a member of Hiram's PhD committee. There are currently three projects at Texas A&M University in this effort. Two of which, the Screening, Scanning, and Inspection Procedures (SSIP) project and the Advanced Security Procedures at Border Crossing (ASBC) ports of entry project, Hiram is involved in. His dissertation would benefit from the empirical information, actual data, to verify the validity of the models that he is creating, which would ultimately benefit the projects funded by DHS.

Hiram has recruited the help of his brother, Rodolfo Moya, as a person who is responsible for collecting service time data. This implies recording from time to time in half hour to one hour intervals, the time it takes to inspect a truck at the primary port of entry booth. This will be done without interfering with the duly activities of inspecting the trucks entering the country.

If you have any questions please do not hesitate to contact me. Thank you for your cooperation.

Guy L. Curry

Guy L. Curry, PhD, PE
Professor and Director of Graduate Program
Phone: (979) 845-5576
Fax: (979) 458-4299
g-curry@tamu.edu

3131 TAMU
College Station, TX 77843-3131

Tel. 979.845.5531 Fax. 979.847.9005
http://ise.tamu.edu

# APPENDIX B

# DATA TABLES

Fast and laden data:

| Original | Mins | Sec | sec/100 | In Decimal Minutes | Mean | Variance | CSV |
|---|---|---|---|---|---|---|---|
| 8.14.43 | 8 | 14 | 43 | 8.237633 | 3.016343 | 2.211677 | 0.243086 |
| 3.17.92 | 3 | 17 | 92 | 3.292533 | | 2.211677 | |
| 4.47.02 | 4 | 47 | 2 | 4.783533 | | | |
| 5.08.80 | 5 | 8 | 80 | 5.141333 | | 1/5 ≤ | 0.243086 ≤ 1/4 |
| 4.00.01 | 4 | 0 | 1 | 4.0001 | | | |
| 2.59.76 | 2 | 59 | 76 | 2.990933 | | | So k= 5 |
| 2.54.66 | 2 | 54 | 66 | 2.9066 | | | 1-α = 0.052067 |
| 4.33.60 | 4 | 33 | 60 | 4.556 | | | α = 0.947933 |
| 7.38.83 | 7 | 38 | 83 | 7.641633 | | | $\mu_1$= 1.58859 |
| 4.18.38 | 4 | 18 | 38 | 4.3038 | | | |
| 3.15.05 | 3 | 15 | 5 | 3.2505 | | | |
| 3.29.27 | 3 | 29 | 27 | 3.486033 | | | |
| 5.05.91 | 5 | 5 | 91 | 5.092433 | | | |
| 5.00.10 | 5 | 0 | 10 | 5.001 | | | |
| 1.34.50 | 1 | 34 | 50 | 1.571667 | | | |
| 3.05.83 | 3 | 5 | 83 | 3.091633 | | | |
| 2.31.68 | 2 | 31 | 68 | 2.523467 | | | |
| 4.40.25 | 4 | 40 | 25 | 4.669167 | | | |
| 4.10.78 | 4 | 10 | 78 | 4.174467 | | | |
| 1.22.91 | 1 | 22 | 91 | 1.375767 | | | |
| 2.15.54 | 2 | 15 | 54 | 2.2554 | | | |
| 2.08.56 | 2 | 8 | 56 | 2.138933 | | | |
| 1.58.71 | 1 | 58 | 71 | 1.973767 | | | |
| 2.33.14 | 2 | 33 | 14 | 2.5514 | | | |
| 4.14.54 | 4 | 14 | 54 | 4.238733 | | | |
| 1.44.25 | 1 | 44 | 25 | 1.735833 | | | |
| 1.39.57 | 1 | 39 | 57 | 1.6557 | | | |
| 1.58.59 | 1 | 58 | 59 | 1.972567 | | | |
| 2.26.37 | 2 | 26 | 37 | 2.437033 | | | |
| 2.27.07 | 2 | 27 | 7 | 2.4507 | | | |
| 2.08.91 | 2 | 8 | 91 | 2.142433 | | | |
| 3.04.21 | 3 | 4 | 21 | 3.068767 | | | |
| 1.48.65 | 1 | 48 | 65 | 1.8065 | | | |
| 1.37.16 | 1 | 37 | 16 | 1.618267 | | | |
| 1.15.01 | 1 | 15 | 1 | 1.2501 | | | |
| 2.44.66 | 2 | 44 | 66 | 2.739933 | | | |
| 1.34.01 | 1 | 34 | 1 | 1.566767 | | | |
| 3.55.82 | 3 | 55 | 82 | 3.924867 | | | |
| 2.11.07 | 2 | 11 | 7 | 2.184033 | | | |
| 1.31.65 | 1 | 31 | 65 | 1.523167 | | | |
| 2.22.72 | 2 | 22 | 72 | 2.373867 | | | |
| 2.29.51 | 2 | 29 | 51 | 2.488433 | | | |
| 2.07.28 | 2 | 7 | 28 | 2.119467 | | | |
| 2.38.73 | 2 | 38 | 73 | 2.640633 | | | |
| 2.11.10 | 2 | 11 | 10 | 2.184333 | | | |
| 1.26.86 | 1 | 26 | 86 | 1.441933 | | | |
| 2.01.22 | 2 | 1 | 22 | 2.018867 | | | |
| 2.45.16 | 2 | 45 | 16 | 2.7516 | | | |
| 3.20.60 | 3 | 20 | 60 | 3.339333 | | | |
| 2.08.02 | 2 | 8 | 2 | 2.133533 | | | |

Fast and empty data:

| Original | Mins | Sec | sec/100 | In Decimal Minutes | Mean | Variance | CSV |
|---|---|---|---|---|---|---|---|
| 1.02.86 | 1 | 2 | 86 | 1.041933 | 1.691345 | 1.13977 | 0.398431 |
| 1.00.21 | 1 | 0 | 21 | 1.0021 | | 1.13977 | |
| 1.34.64 | 1 | 34 | 64 | 1.573067 | | | |
| 1.26.61 | 1 | 26 | 61 | 1.439433 | 1/3 ≤ | 0.398431 | ≤ 1/2 |
| 5.23.24 | 5 | 23 | 24 | 5.385733 | | | |
| 1.46.12 | 1 | 46 | 12 | 1.767867 | | So k= | 3 |
| 1.00.35 | 1 | 0 | 35 | 1.0035 | | 1-α = | 0.093629 |
| 1.23.81 | 1 | 23 | 81 | 1.391433 | | α = | 0.906371 |
| 1.14.04 | 1 | 14 | 4 | 1.233733 | | $\mu_1$= | 1.663021 |
| 0.59.02 | 0 | 59 | 2 | 0.983533 | | | |
| 0.43.55 | 0 | 43 | 55 | 0.722167 | | | |
| 0.55.60 | 0 | 55 | 60 | 0.922667 | | | |
| 0.58.25 | 0 | 58 | 25 | 0.969167 | | | |
| 1.26.88 | 1 | 26 | 88 | 1.442133 | | | |
| 1.21.70 | 1 | 21 | 70 | 1.357 | | | |
| 1.16.20 | 1 | 16 | 20 | 1.268667 | | | |
| 2.21.01 | 2 | 21 | 1 | 2.3501 | | | |
| 0.49.04 | 0 | 49 | 4 | 0.817067 | | | |
| 1.53.33 | 1 | 53 | 33 | 1.886633 | | | |
| 1.11.45 | 1 | 11 | 45 | 1.187833 | | | |
| 0.55.73 | 0 | 55 | 73 | 0.923967 | | | |
| 1.34.08 | 1 | 34 | 8 | 1.567467 | | | |
| 6.32.86 | 6 | 32 | 86 | 6.541933 | | | |
| 1.34.64 | 1 | 34 | 64 | 1.573067 | | | |
| 2.40.55 | 2 | 40 | 55 | 2.672167 | | | |
| 1.35.26 | 1 | 35 | 26 | 1.585933 | | | |
| 5.00.61 | 5 | 0 | 61 | 5.0061 | | | |
| 1.38.06 | 1 | 38 | 6 | 1.633933 | | | |
| 1.22.63 | 1 | 22 | 63 | 1.372967 | | | |
| 1.18.79 | 1 | 18 | 79 | 1.3079 | | | |
| 1.39.67 | 1 | 39 | 67 | 1.6567 | | | |
| 1.27.03 | 1 | 27 | 3 | 1.4503 | | | |
| 1.42.78 | 1 | 42 | 78 | 1.7078 | | | |
| 1.20.18 | 1 | 20 | 18 | 1.335133 | | | |
| 1.59.97 | 1 | 59 | 97 | 1.993033 | | | |
| 1.26.12 | 1 | 26 | 12 | 1.434533 | | | |
| 1.29.75 | 1 | 29 | 75 | 1.490833 | | | |
| 1.26.26 | 1 | 26 | 26 | 1.435933 | | | |
| 1.48.30 | 1 | 48 | 30 | 1.803 | | | |
| 1.32.96 | 1 | 32 | 96 | 1.542933 | | | |
| 2.26.93 | 2 | 26 | 93 | 2.442633 | | | |
| 1.53.96 | 1 | 53 | 96 | 1.892933 | | | |
| 1.31.22 | 1 | 31 | 22 | 1.518867 | | | |
| 3.02.34 | 3 | 2 | 34 | 3.036733 | | | |
| 1.10.12 | 1 | 10 | 12 | 1.167867 | | | |
| 1.30.66 | 1 | 30 | 66 | 1.5066 | | | |
| 1.52.77 | 1 | 52 | 77 | 1.874367 | | | |
| 1.08.14 | 1 | 8 | 14 | 1.134733 | | | |
| 0.56.08 | 0 | 56 | 8 | 0.934133 | | | |
| 1.22.07 | 1 | 22 | 7 | 1.367367 | | | |
| 1.33.87 | 1 | 33 | 87 | 1.5587 | | | |
| 0.57.13 | 0 | 57 | 13 | 0.9513 | | | |
| 1.09.77 | 1 | 9 | 77 | 1.1577 | | | |
| 1.11.73 | 1 | 11 | 73 | 1.190633 | | | |
| 1.30.80 | 1 | 30 | 80 | 1.508 | | | |

## Non-Fast and laden data:

| Original | Mins | Sec | sec/100 | In Decimal Minutes | Mean | Variance | CSV |
|---|---|---|---|---|---|---|---|
| 5.18.84 | 5 | 18 | 84 | 5.3084 | 3.94904 | 4.272448 | 0.273964 |
| 1.06.10 | 1 | 6 | 10 | 1.101 | | 4.272448 | |
| 2.05.07 | 2 | 5 | 7 | 2.084033 | | | |
| 2.04.37 | 2 | 4 | 37 | 2.070367 | | 1/4 ≤ | 0.273964 ≤ 1/3 |
| 2.24.00 | 2 | 24 | 0 | 2.4 | | | |
| 3.28.16 | 3 | 28 | 16 | 3.468267 | | | So k= 4 |
| 2.54.59 | 2 | 54 | 59 | 2.9059 | | | 1-α = 0.031389 |
| 3.19.39 | 3 | 19 | 39 | 3.320567 | | | α = 0.968611 |
| 3.20.58 | 3 | 20 | 58 | 3.339133 | | | $\mu_1$= 0.989059 |
| 1.57.17 | 1 | 57 | 17 | 1.9517 | | | |
| 2.17.71 | 2 | 17 | 71 | 2.290433 | | | |
| 1.56.89 | 1 | 56 | 89 | 1.942233 | | | |
| 5.29.66 | 5 | 29 | 66 | 5.489933 | | | |
| 5.49.78 | 5 | 49 | 78 | 5.824467 | | | |
| 3.39.29 | 3 | 39 | 29 | 3.6529 | | | |
| 5.08.64 | 5 | 8 | 64 | 5.139733 | | | |
| 5.31.90 | 5 | 31 | 90 | 5.525667 | | | |
| 13.04.51 | 13 | 4 | 51 | 13.07177 | | | |
| 5.43.20 | 5 | 43 | 20 | 5.718667 | | | |
| 1.58.99 | 1 | 58 | 99 | 1.976567 | | | |
| 2.31.05 | 2 | 31 | 5 | 2.517167 | | | |
| 5.21.56 | 5 | 21 | 56 | 5.3556 | | | |
| 2.53.05 | 2 | 53 | 5 | 2.883833 | | | |
| 5.36.51 | 5 | 36 | 51 | 5.6051 | | | |
| 5.36.51 | 5 | 36 | 51 | 5.6051 | | | |
| 2.44.46 | 2 | 44 | 46 | 2.737933 | | | |
| 3.39.05 | 3 | 39 | 5 | 3.6505 | | | |
| 2.45.23 | 2 | 45 | 23 | 2.7523 | | | |
| 5.10.38 | 5 | 10 | 38 | 5.170467 | | | |
| 4.38.07 | 4 | 38 | 7 | 4.634033 | | | |
| 3.45.41 | 3 | 45 | 41 | 3.7541 | | | |
| 3.26.48 | 3 | 26 | 48 | 3.438133 | | | |
| 2.01.99 | 2 | 1 | 99 | 2.026567 | | | |
| 4.50.44 | 4 | 50 | 44 | 4.837733 | | | |
| 6.10.04 | 6 | 10 | 4 | 6.167067 | | | |
| 9.25.83 | 9 | 25 | 83 | 9.424967 | | | |
| 4.16.28 | 4 | 16 | 28 | 4.269467 | | | |
| 4.06.29 | 4 | 6 | 29 | 4.1029 | | | |
| 5.42.79 | 5 | 42 | 79 | 5.7079 | | | |
| 2.39.99 | 2 | 39 | 99 | 2.6599 | | | |
| 2.28.54 | 2 | 28 | 54 | 2.472067 | | | |
| 3.50.79 | 3 | 50 | 79 | 3.841233 | | | |
| 2.03.81 | 2 | 3 | 81 | 2.0581 | | | |
| 7.02.18 | 7 | 2 | 18 | 7.035133 | | | |
| 3.58.54 | 3 | 58 | 54 | 3.972067 | | | |
| 3.36.47 | 3 | 36 | 47 | 3.6047 | | | |
| 4.39.12 | 4 | 39 | 12 | 4.6512 | | | |
| 2.17.22 | 2 | 17 | 22 | 2.285533 | | | |
| 3.24.73 | 3 | 24 | 73 | 3.4073 | | | |
| 1.51.93 | 1 | 51 | 93 | 1.8593 | | | |
| 1.47.18 | 1 | 47 | 18 | 1.785133 | | | |
| 2.41.18 | 2 | 41 | 18 | 2.685133 | | | |
| 3.45.97 | 3 | 45 | 97 | 3.7597 | | | |

Non-Fast and empty data:

| Original | Mins | Sec | sec/100 | In Decimal Minutes | | Mean | Variance | CSV |
|---|---|---|---|---|---|---|---|---|
| 1.52.26 | 1 | 52 | 26 | 1.869267 | 0.627044 | 2.661128 | 2.890008 | 0.408101 |
| 2.07.16 | 2 | 7 | 16 | 2.118267 | 0.294698 | | 2.890008 | |
| 2.26.30 | 2 | 26 | 30 | 2.436333 | 0.050533 | | | |
| 2.05.41 | 2 | 5 | 41 | 2.087433 | 0.329126 | | 1/3 ≤ | 0.408101 ≤ 1/2 |
| 1.08.10 | 1 | 8 | 10 | 1.134333 | 2.331102 | | | |
| 1.52.35 | 1 | 52 | 35 | 1.870167 | 0.62562 | | | So k= 3 |
| 2.27.70 | 2 | 27 | 70 | 2.457 | 0.041668 | | | 1-α = 0.106893 |
| 3.17.29 | 3 | 17 | 29 | 3.286233 | 0.390757 | 2.832208 | | α = 0.893107 |
| 1.57.87 | 1 | 57 | 87 | 1.9587 | 0.493405 | | | $\mu_1$= 1.047005 |
| 2.00.28 | 2 | 0 | 28 | 2.0028 | 0.433396 | 2.890008 | | |
| 0.54.96 | 0 | 54 | 96 | 0.9096 | 3.06785 | | | |
| 3.13.38 | 3 | 13 | 38 | 3.220467 | 0.31286 | | | |
| 4.32.52 | 4 | 32 | 52 | 4.538533 | 3.524651 | | | |
| 4.03.08 | 4 | 3 | 8 | 4.0508 | 1.931188 | | | |
| 6.24.48 | 6 | 24 | 48 | 6.4048 | 14.01508 | | | |
| 5.45.83 | 5 | 45 | 83 | 5.7583 | 9.592474 | | | |
| 10.32.72 | 10 | 32 | 72 | 10.54053 | 62.08503 | | | |
| 2.09.61 | 2 | 9 | 61 | 2.1561 | 0.255053 | | | |
| 0.54.41 | 0 | 54 | 41 | 0.9041 | 3.087147 | | | |
| 1.00.28 | 1 | 0 | 28 | 1.0028 | 2.750052 | | | |
| 1.09.77 | 1 | 9 | 77 | 1.1577 | 2.260296 | | | |
| 3.05.85 | 3 | 5 | 85 | 3.091833 | 0.185507 | | | |
| 1.16.62 | 1 | 16 | 62 | 1.272867 | 1.92727 | | | |
| 2.27.77 | 2 | 27 | 77 | 2.4577 | 0.041383 | | | |
| 3.17.85 | 3 | 17 | 85 | 3.291833 | 0.397789 | | | |
| 4.46.60 | 4 | 46 | 60 | 4.772667 | 4.458596 | | | |
| 1.39.47 | 1 | 39 | 47 | 1.6547 | 1.012897 | | | |
| 3.44.01 | 3 | 44 | 1 | 3.733433 | 1.149839 | | | |
| 0.45.12 | 0 | 45 | 12 | 0.7512 | 3.647825 | | | |
| 0.52.17 | 0 | 52 | 17 | 0.868367 | 3.213993 | | | |
| 2.00.94 | 2 | 0 | 94 | 2.0094 | 0.424749 | | | |
| 4.13.91 | 4 | 13 | 91 | 4.225767 | 2.448094 | | | |
| 1.28.15 | 1 | 28 | 15 | 1.468167 | 1.423157 | | | |
| 1.48.23 | 1 | 48 | 23 | 1.8023 | 0.737586 | | | |
| 3.25.57 | 3 | 25 | 57 | 3.422367 | 0.579484 | | | |
| 1.18.65 | 1 | 18 | 65 | 1.3065 | 1.835017 | | | |
| 2.20.99 | 2 | 20 | 99 | 2.343233 | 0.101057 | | | |
| 2.40.86 | 2 | 40 | 86 | 2.675267 | 0.0002 | | | |
| 2.12.89 | 2 | 12 | 89 | 2.2089 | 0.20451 | | | |
| 1.05.09 | 1 | 5 | 9 | 1.084233 | 2.486597 | | | |
| 1.41.53 | 1 | 41 | 53 | 1.688633 | 0.945746 | | | |
| 4.16.58 | 4 | 16 | 58 | 4.272467 | 2.596412 | | | |
| 2.59.00 | 2 | 59 | 0 | 2.983333 | 0.103816 | | | |
| 1.55.44 | 1 | 55 | 44 | 1.921067 | 0.547691 | | | |
| 2.32.84 | 2 | 32 | 84 | 2.541733 | 0.014255 | | | |
| 2.42.50 | 2 | 42 | 50 | 2.705 | 0.001925 | | | |
| 2.48.48 | 2 | 48 | 48 | 2.8048 | 0.020642 | | | |
| 2.48.83 | 2 | 48 | 83 | 2.8083 | 0.02166 | | | |
| 3.38.66 | 3 | 38 | 66 | 3.639933 | 0.95806 | | | |
| 1.23.28 | 1 | 23 | 28 | 1.386133 | 1.625611 | | | |

# APPENDIX C
# GENERATOR MATRICES FOR A POE MARKOVIAN MODEL WITH SMALL QUEUES

Because future transitions depend on which type of customer is being processed by two servers, the state space must include not only the customer type but also an indication of which customer type has the extra server, plus a tracking mechanism for the small queue size of the reallocating server. Thus, a state of the system will be denoted by $(i, j, \nu, \ell)$, where $\nu = 1$ indicates that two servers have been assigned to Type 1 customers and if $\nu = 2$ there are two servers assigned to Type 2 customers. And $\ell$ indicates the number of customers (trucks) in the small queue of the reallocating server (inspection station), that are waiting to be served (or inspected by CBP Agents).

The operational question is to determine when to assign two servers to Type 1 or Type 2 customers. Since we are assuming that all processes are exponential, this can be modeled as a Markov decision process in which we shall impose a control-limit type policy for the decision process.

This control-limit type policy is defined by functions $\tau_\nu(\,\cdot\,)$ for $\nu = 1, 2$. The $\tau_\nu$ function defines the decision to move the server that can be reallocated when the extra server is currently working on a Type $\nu$ customer. The decision control policy can be either stationary, i.e. at 11:00 am the server is reallocated; or it can be non-stationary, in the case the control policy is based on queue size differential, as implemented in Section 4.3. The term $\tau_\nu(j) = k$ indicates that the extra server should be changed to work on non-Type $\nu$ customers at a service epoch of a Type $\nu$ customer if there are $j$ customers of Type $\nu$ and $k$ customers of the other type in the system.

Such a control-limit policy insures that the decision problem reduces to a Markov decision problem, where the non-stationary Markov process is denoted as $\{Y_t; t \geq 0\}$

with generator matrix given by $G_t$ for $t \geq 0$. The problem is clearly a transient problem since within any realistic system, the dynamics are such that the queues build during most of the day until the inspection station closes for the day. The transient probabilities will be denoted by the vector $P_t$ for $t \geq 0$, where

$$\begin{aligned} P_t(i,j,\nu,\ell) &= Pr\{Y_t = (i,j,\nu,\ell) \mid Y_0 = (0,0,1,0)\}, \\ \text{for } t \geq 0,\ (i,j) \in \mathcal{N}^2,\ \nu &= \{1,2\},\ \ell = \{0,\cdots,5\}\ . \end{aligned} \tag{C.1}$$

Since the lines of commercial trucks have been seen to extend as far as 80 Kms (Li et al., 2005), the system does not have an effective limit on capacity. However, to develop a generator matrix, we need to cap the capacity, and truncate the state space to allow for a total of $n_{max}$ customers in the system. This also means that when the system is full, additional customers will not be allowed to enter because of the size cap of the generator matrix.

Therefore, the state space is:

$$\begin{aligned} \mathcal{E} = \{&(0,0,1,0) \mid (1,0,1,0),(1,0,1,1),(0,1,1,0) \mid \\ &(2,0,1,0),(2,0,1,1),(2,0,1,2),(1,1,1,0),(1,1,1,1),(0,2,1,0) \mid \\ &\cdots \mid (n_{max},0,1,5),(n_{max},0,1,4),(n_{max},0,1,3),(n_{max},0,1,2), \\ &(n_{max},0,1,1),(n_{max},0,1,0), \\ &(n_{max}-1,1,1,5),(n_{max}-1,1,1,4),(n_{max}-1,1,1,3),(n_{max}-1,1,1,2), \\ &(n_{max}-1,1,1,1),(n_{max}-1,1,1,0),\ldots,(0,n_{max},1,0)\} \\ &\cup \\ &\{(0,0,2,0) \mid (1,0,2,0),(0,1,2,0),(0,1,2,1) \mid \\ &(2,0,2,0),(1,1,2,0),(1,1,2,1),(0,2,2,0),(0,2,2,1),(0,2,2,2) \mid \\ &\cdots \mid (n_{max},0,2,0) \\ &\mid (n_{max}-1,1,2,0),(n_{max}-1,1,2,1), \\ &\ldots,(0,n_{max},2,5),(0,n_{max},2,4),(0,n_{max},2,3),(0,n_{max},2,2), \\ &(0,n_{max},2,1),(0,n_{max},2,0)\}\ . \end{aligned}$$

Notice there is a natural partition of the state space so that the probabilities, $P_t$, associated with this state space will also be partitioned. Thus, the probabilities $P_{k,\nu,\ell,t}$ are those probabilities associated with the states $\{(k,0,\nu,\ell),\cdots,(0,k,\nu,\ell)\}$, namely those states associated with a total of $k$ customers in the system with the movable server assigned to serve Type $\nu$ customers, and until the queue of the moveable server, $\ell$ is expunged. Therefore, allowing for a tracking of a moveable server, and a DRM.

**Generator Matrices**

The variable $\nu$ is the tasked with keeping track of the current status of the DRM server. This variable has four values, $\nu = \{0.5, 1, 1.5, 2\}$, and the meaning of each value is as follows:

- $\nu = 0.5$: The DRM server is in transition to servicing Type $\nu = 1$ trucks, but servicing $\nu = 2$ trucks until $\ell = 0$; that is, the DRM queue is expunged.
- $\nu = 1$: The DRM server is servicing Type $\nu = 1$ trucks, and can only be triggered to be $\nu = 1.5$
- $\nu = 1.5$: The DRM server is in transition to servicing Type $\nu = 2$ trucks, but servicing $\nu = 1$ trucks until $\ell = 0$; that is, the DRM queue is expunged.
- $\nu = 2$: The DRM server is servicing Type $\nu = 2$ trucks, and can only be triggered to be $\nu = 0.5$.

In the above description, "transition" means that the policy $\tau_\nu(\,\cdot\,)$ has been met, and the request has been placed to reallocate the DRM server to customer Type $\nu = 1$. Transient probabilities, $p_n(t) = \Pr\{N(t) = n | N(0) = 0\}$, are necessary interpretations of service time.

In the following generator matrices, the rates of $\mu_1^*$ and $\mu_2^*$ represent the service rates of customer Type $\nu = 1$ and $\nu = 2$ respectively, from the DRM server which is in the process of being reallocated. For a fixed value of the control-limit functions,

$\tau_1$ and $\tau_2$, the positive elements of the generator matrix for the resulting Markov process can be defined as follows:

The first generator matrix is for $\nu = 1$:

$$
G_t((i,j,1,k),(m,n,\nu,\ell)) = \begin{cases}
\mu_1 & \text{For } (i=1, k=0): \\
 & m=0,\ n=j,\ \nu=1,\ l=1 \\
\mu_1^* & \text{For } (i=1, k=1): \\
 & m=0,\ n=j,\ \nu=1,\ l=0 \\
\mu_1 & \text{For } (i>1): \\
 & m=i-1,\ n=j,\ \nu=1,\ l=k \\
\mu_1^* & \text{For } (i>1): \\
 & m=i-1,\ n=j,\ \nu=1,\ l=k-1 \\
\mu_2 & \text{For } (j>0): \\
 & m=i,\ n=j-1,\ \nu=1,\ l=k \\
\lambda_1(t) & \text{For } (i<n_{max},\ k<q_{max},\ j<\tau_1(i)): \\
 & m=i+1,\ n=j,\ \nu=1,\ l=k+1 \\
\lambda_1(t) & \text{For } (i<n_{max},\ k<q_{max},\ j\geq\tau_1(i)): \\
 & m=i+1,\ n=j,\ \nu=1.5,\ l=k+1 \\
\lambda_1(t) & \text{For } (i<n_{max},\ k=q_{max},\ j<\tau_1(i)): \\
 & m=i+1,\ n=j,\ \nu=1,\ l=k \\
\lambda_1(t) & \text{For } (i<n_{max},\ k=q_{max},\ j\geq\tau_1(i)): \\
 & m=i+1,\ n=j,\ \nu=1.5,\ l=k \\
\lambda_2(t) & \text{For } (j<n_{max}): \\
 & m=i,\ n=j+1,\ \nu=1,\ l=k
\end{cases} \tag{C.2}
$$

The following generator matrix is when $\nu = 1.5$:

$$
G_t((i,j,1.5,k),(m,n,\nu,\ell)) = \begin{cases}
\mu_1^* & \text{For } (i=1,\ k=1): \\
 & m=0,\ n=j,\ \nu=2,\ l=0 \\
\mu_1 & \text{For } (i \geq 1): \\
 & m=i-1,\ n=j,\ \nu=1.5,\ l=k \\
\mu_1^* & \text{For } (i>1,\ k>1): \\
 & m=i-1,\ n=j,\ \nu=1.5,\ l=k-1 \\
\mu_1^* & \text{For } (i>1,\ k=1): \\
 & m=i-1,\ n=j,\ \nu=2,\ l=0 \\
\mu_2 & \text{For } (j>0,\ m=i): \\
 & n=j-1,\ \nu=1.5,\ l=k \\
\lambda_1(t) & \text{For } (i<n_{max}): \\
 & m=i+1,\ n=j,\ \nu=1.5,\ l=k \\
\lambda_2(t) & \text{For } (j<n_{max}): \\
 & m=i,\ n=j+1,\ \nu=1.5,\ l=k
\end{cases} \tag{C.3}
$$

The next generator matrix is when $\nu = 2$:

$$
G_t((i,j,2,k),(m,n,\nu,\ell)) = \begin{cases}
\mu_1 & \text{For } (i>0): \\
 & m=i-1,\ n=j,\ \nu=2,\ l=k \\
\mu_2 & \text{For } (j=1,\ k=0): \\
 & m=i,\ n=0,\ \nu=2,\ l=k \\
\mu_2^* & \text{For } (j=1,\ k=1): \\
 & m=i,\ n=0,\ \nu=2,\ l=0 \\
\mu_2 & \text{For } (i<\tau_2(j),\ j>1): \\
 & m=i,\ n=j-1,\ \nu=2,\ l=k \\
\mu_2^* & \text{For } (i<\tau_2(j),\ j>1): \\
 & m=i,\ n=j-1,\ \nu=2,\ l=k-1 \\
\mu_2 & \text{For } (i\geq\tau_2(j),\ j>1): \\
 & m=i,\ n=j-1,\ \nu=0.5,\ l=k \\
\mu_2^* & \text{For } (i\geq\tau_2(j),\ j>1): \\
 & m=i,\ n=j-1,\ \nu=0.5,\ l=k-1 \\
\lambda_1(t) & \text{For } (i<n_{max}): \\
 & m=i+1,\ n=j,\ \nu=2,\ l=k \\
\lambda_2(t) & \text{For } (j<n_{max},\ k<q_{max}): \\
 & m=i,\ n=j+1,\ \nu=2,\ l=k+1 \\
\lambda_2(t) & \text{For } (j<n_{max},\ k=q_{max}): \\
 & m=i,\ n=j+1,\ \nu=2,\ l=k
\end{cases}
\tag{C.4}
$$

The last generator matrix is for $\nu = 0.5$:

$$
G_t((i,j,0.5,k),(m,n,\nu,\ell)) = \begin{cases}
\mu_1 & \text{For } (i>0): \\
 & \quad m=i-1,\ n=j,\ \nu=0.5,\ l=k \\
\mu_2 & \text{For } (j\geq 1): \\
 & \quad m=i,\ n=0,\ \nu=0.5,\ l=k \\
\mu_2^* & \text{For } (j=1,\ k=1): \\
 & \quad m=i,\ n=0,\ \nu=1,\ l=0 \\
\mu_2^* & \text{For } (j>1,\ k>1): \\
 & \quad m=i,\ n=j-1,\ \nu=0.5,\ l=k-1 \\
\mu_2^* & \text{For } (j>1,\ k=1): \\
 & \quad m=i,\ n=j-1,\ \nu=1,\ l=0 \\
\lambda_1(t) & \text{For } (i<n_{max}): \\
 & \quad m=i+1,\ n=j,\ \nu=0.5\ l=k \\
\lambda_2(t) & \text{For } (j<n_{max}): \\
 & \quad m=i,\ n=j+1,\ \nu=0.5\ l=k
\end{cases} \tag{C.5}
$$

for $t \geq 0$. All other elements of the generator matrix are zero, except for the diagonal elements which are such that the row sums are zero.

# VITA

Hiram Moya earned his Bachelor of Science degree in Industrial Engineering from Texas A&M University at College Station, Texas in 1996. After working in Accenture for 5 years, he founded and became the managing Partner of HMGroup LLP, while at the same time taking graduate level courses. In 2004, he received his Master of Science degree in Engineering Systems Management from Texas A&M University at College Station, Texas. His degree was earned taking online courses offered by the Industrial and Systems Engineering department at Texas A&M, with complimentary courses from the University of Texas at Dallas. Later he returned as a full time student and completed his Doctor of Philosophy in Industrial and Systems Engineering in 2012. His research interests include Operations Research, Optimization, Supply Chain Management, Queueing Theory, Applied Probability, and Homeland Security.

Dr. Hiram Moya may be reached at the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77840. His email address is Hiram@tamu.edu or Hiram.Moya@gmail.com.