TRANSPORTATION SYSTEMS RELIABILITY ANALYSIS:
MODELING TRAFFIC SUBJECT TO INCIDENTS APPLYING
QUEUEING THEORY

By

MARCELO RICARDO FIGUEROA CANDIA

A dissertation submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Industrial and Systems Engineering
Written under the direction of
Melike Baykal-Gürsoy
And approved by

______________________________

______________________________

______________________________

______________________________

New Brunswick, New Jersey
October, 2018
ABSTRACT OF THE DISSERTATION

Transportation Systems Reliability Analysis:
Modeling Traffic Subject to Incidents Applying Queueing Theory

by MARCELO RICARDO FIGUEREA CANDIA

Dissertation Director:
Melike Baykal-Gürsoy

This dissertation investigates different aspects of the reliability of a transportation system subject to random events by means of applied queueing theory. We use queueing models and fundamental principles of stochastic analysis, to shine light on the behavior and properties of a traffic system that is subject to random deteriorations of the quality of service as indicated by vehicular travel time. For this, we discuss stationary and transient analytical solutions for the number of customers and service-time distributions of Markov-modulated service rates queues. We also study analytically the effect that different random components of the system have in the completion time of a single trip. We are able to give closed form solutions to accommodate a number of different combinations of random variables as inputs of the system, as well as to give analytical insights on the asymptotic behavior when abnormally slow customers show up. We validate and calibrate the analytical models using incident reports and weather conditions as sources of traffic deterioration. We generate measures of the performance of the system with explicit dependencies on traffic and incident parameters, avoiding the use of costly simulation. We make use of these measures for optimizing risk-averse route choice problems. We expect our models contribute to the design and operation of management tools for roadway traffic and incident mitigation that can lead to safer and more efficient movement of people, goods, and other resources.
Dedication

To Carla, my beloved wife and best friend.
Acknowledgments

This dissertation document is a symbol of the culmination of six years of work, and it feels strangely monumental to be closing these covers after turning so many pages, day after day. But all due dates eventually come, and futures revisited constantly in the imagination actually come to be. A sense of void will be inevitable, at least for some time.

My biggest gratitude is without a doubt to Carla, my life companion. This adventure of pursuing a doctoral degree was a joint project of ours; we were going to make it together, against any odds. This was difficult from the start, and we have been working for this for many more than six years. Look where we are now! Thanks for the sweat and tears, for being with me through all the hardships, for letting me borrow your incredible mind for making tough decisions, for finding so many exciting things to do, for the awesome times you have shared with me at Rutgers, for the beautiful memories, trips, meals, and slow walks. Thanks for being with me every day, you make me so happy.

I am also deeply indebted to my parents, Dante and Patricia. They are the reason of my existence in all the senses of the word. I thank you for your constant dedication to me and my brothers during our childhood, for instilling in us the values that make us who we are, and for providing us with a safe, fun, and stimulating environment as we grew to become men. We inherited so many great things from you, in particular the passion and drive to pursue doctoral degrees. You have supported me in many ways during this tough years, thanks for trusting my dreams. I also thank my grandparents, who maybe never imagined the things that we would accomplish, but were absolute necessary conditions for this dissertation to have been ever written. Big thanks to my two brothers, Leonardo and Mauricio, for being such cool guys and providing me
with a niece and nephews for my continuous amusement.

I am very honored to thank here my academic advisor Dr. Melike Baykal-Gürsoy. I am forever grateful for all the time, energy, and enthusiasm that you have constantly poured into this work. Thank you for everything you taught me during these years, it has been incredible to see you work up close. I admire your intellectual prowess as much as I admire your unwavering smile; you have this special charm for motivating people around you. Thank you for your generosity in sharing your knowledge, for your time in so many wonderful work meetings and interesting conversations, and for being there to support me in every aspect of my life as a Rutgers student. I am so glad to have had you as my advisor.

Many thanks to the members of my dissertation committee, Professors Susan Albin, David Coit, Peter J. Jin, and Tuğrul Özel, for their time and advice; your comments have been critical for the improvement of this work. Thanks also to all of the professors whom I had the pleasure of taking classes with, specially at the ISE Department and RutCOR; I learnt so many invaluable things during my years here. Thanks also to the ISE staff for making our lives easier, to all of my classmates and friends, to the Aresty student assistants, to the Alpha Pi Mu honor society, to my band (G.I.S), and to the many furry little friends that endured me taking pictures of them during all this years in the wonderful Busch Campus. I will miss you all.

I thank Peter J. Jin, Eric Gonzales, Shadi Tahvildar-Zadeh, and Mert Gürbüzbalaban for many fruitful discussions and helpful comments. Peter J. Jin also kindly shared traffic data. This research was partly funded by the Tayfur Altiok Scholarship of the Industrial & Systems Engineering Department, Rutgers University. I specially thank the Chilean Government through the National Commission for Scientific and Technological Research (CONICYT) and the Chilean Fulbright Commission for sponsoring my Ph.D. studies.
## Table of Contents

Abstract of the Dissertation ........................................ ii

Dedication ....................................................................... iii

Acknowledgments .......................................................... iv

List of Tables .................................................................. ix

List of Figures .................................................................. x

1 Introduction .................................................................. 1

1.1 Literature Review ..................................................... 3

1.1.1 Transient Analysis of a Repairable Queueing System ..... 3

1.1.2 The Completion Time of an Interrupted Process .......... 5

1.1.3 Travel Time and Traffic Volume Reliability ................. 7

1.1.4 Route Optimization: Risk Averse Shortest Path ........... 9

2 Transient Analysis of a Repairable Queueing System ........ 12

2.1 PDE System .......................................................... 13

2.2 Governing PDE ...................................................... 16

2.3 Laplace Transform Treatment .................................... 19

2.3.1 Special Case: Total Failure System ($\mu' = 0$) ......... 20
## 2.4 Moment Calculation

- **2.4.1 First Moment** .................................................. 22
- **2.4.2 Laplace Transform of \( k \)-th Moment** ............... 32

## 2.5 Series Expansion Treatment ..................................... 35

## 2.6 Future Work .......................................................... 38

- **2.6.1 Remaining Direct Calculations** ............................ 38
- **2.6.2 Alternative Approaches** ................................... 38
- **2.6.3 Applications** .................................................. 45

## 3 The Completion Time of an Interrupted Process ............... 46

- **3.1 Queueing Model** .................................................. 47
- **3.2 Review of Completion Time Analysis** ....................... 48
- **3.3 Non Exponential Service Time Distributions** ............. 60
  - **3.3.1 Deterministic Service Time** ............................. 60
  - **3.3.2 Uniform Service Time** ................................... 62
  - **3.3.3 Triangular Service Time** ................................ 65
  - **3.3.4 Degenerate Triangle: \( c = a \)** ......................... 68
  - **3.3.5 Degenerate Triangle: \( c = b \)** ......................... 69
  - **3.3.6 Shifted Exponential Service Time** ..................... 70
- **3.4 On the Preempt-RESTART Service Discipline** ........... 72
- **3.5 Asymptotics of The Completion Time** ....................... 77
  - **3.5.1 Repeat Different/REPLACE** ............................. 78
  - **3.5.2 Repeat Identical/RESTART** ............................. 79
- **3.6 Insensitivity of the Stationary Mean Measures to How Interruptions Affect Servers** .................................. 92
- **3.7 Conclusions and Future Work** ............................... 94
4 Travel Time and Traffic Volume Reliability 96

4.1 Notation and Analytic Model ........................................ 97
  4.1.1 Full Travel Time Distribution .................................. 99
  4.1.2 Specializing the Service Time Distribution .................... 100

4.2 Model Implementation ............................................... 103
  4.2.1 Validation of Model Assumptions ................................. 106
  4.2.2 Calibration ...................................................... 108
  4.2.3 FREEVAL-RL .................................................... 111

4.3 Results ............................................................. 112
  4.3.1 Reliability Measures .............................................. 112
  4.3.2 Scenario Analysis ................................................ 116
  4.3.3 Across Sensor Model Validation ................................. 122
  4.3.4 Model Specialization ............................................. 124

4.4 Conclusions and Future Work ....................................... 126

5 Route Optimization: Risk Averse Shortest Path 128

5.1 The Shortest Path Problem .......................................... 129

5.2 Modeling Frameworks ................................................ 132
  5.2.1 Dynamic Programming ............................................ 132
  5.2.2 Stochastic Programming ......................................... 134

5.3 Implementations ..................................................... 142
  5.3.1 Simple Shortest Path Problem .................................. 144
  5.3.2 Dynamic Programming .......................................... 145

5.4 Conclusions and Future Work ....................................... 155
  5.4.1 Stochastic Programming Approach ............................... 155
List of Tables

3.1 Expected Completion Time Exponential Service Time . . . . . . . 58
3.2 Worpitzky Triangle $W(n, k)$ . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

4.1 Percentiles for travel times distribution’s tail . . . . . . . . . . . 114
4.2 Quantiles for Travel Time Distribution’s Tail: Varying $\alpha$ . . . . 118
4.3 Reliability Indices for Travel Time Distribution: Varying $\alpha$ . . . . 119
4.4 Quantiles for Travel Time Distribution’s Tail: Varying $f$ . . . . . . 121
4.5 Reliability Indices for Travel Time Distribution: Varying $f$ . . . . . 121
4.6 Average Sensor Model Validation Direction SE . . . . . . . . . . . 123
4.7 Average Sensor Model Validation Direction WN . . . . . . . . . . . 123
4.8 Average Validation for Other Models . . . . . . . . . . . . . . . . . . 123
4.9 Specializing Model by Day of the Week . . . . . . . . . . . . . . . . . 124

5.1 Data for Test Network $\mathcal{G}($N, $\mathcal{A}_o$) . . . . . . . . . . . 144
5.2 Solution for DP Equations of Look-Ahead Formulation . . . . . . 150
5.3 Total Cost Comparison of SPP . . . . . . . . . . . . . . . . . . . . . . . 153
5.4 Total Cost for Test Network . . . . . . . . . . . . . . . . . . . . . . . . . 155
List of Figures

2.1 Rate Transition Diagram ............................. 14
2.2 Transient Expectation E[X(t)] ........................... 27
2.3 Convergence Dominance for Transient Expectation E[X(t)] .... 29
3.1 Sample Path of Service System for a Customer Arrival Under $G^1$ 50
3.2 Sample Path of Service System for a Customer Arrival Under $G^2$ 53
3.3 Completion Time Distribution under Exponential Service Requirement 58
3.4 Deterministic Distribution ................................ 61
3.5 Uniform Service Time Distribution ........................... 63
3.6 Completion Time Distribution under Uniform Service Requirement 65
3.7 Triangular Service Time Distribution .......................... 66
3.8 Completion Time Distribution under Triangular Service Requirement 67
3.9 Degenerate Triangular Distribution, $c = a$ ........................ 68
3.10 Degenerate Triangular Distribution, $c = b$ ..................... 70
3.11 Shifted Exponential Distribution ............................. 71
4.1 Triangular Service Time Distribution: Application .................. 101
4.2 Uniform Service Time Distribution: Application .................. 102
4.3 Shifted Exponential Service Time Distribution: Application ........ 103
4.4 Map of Interstate 894 ........................................... 104
Chapter 1

Introduction

Increased traffic volume on existing roadways leads, inevitably, to a rise in congestion, longer delays, a decrease in traffic flow, greater fuel consumption, and potentially negative environmental and health impacts [74]. The U.S. Department of Transportation (USDOT) has estimated that the total annual cost of delays in rural and urban areas is approximately $1 trillion [61]. While congestion during peak hours is commonplace, congestion at other times may be caused by irregular occurrences. Nonrecurrent congestion may be due to significant fluctuations in demand, special events, work zones, traffic accidents, disabled vehicles, natural causes (e.g., adverse weather conditions), and spilled loads and hazardous materials. The nonrecurrent congestion is a significant contributor to the total delay or vehicle travel time [51, 85]. Well over half of nonrecurring traffic delays in urban areas, and nearly all delays in rural areas are attributed to work zones and accidents that affect the capacity of the roadway [86].

One of the most important performance measures of a transportation system along with traffic volume is the travel time experienced by travelers. However, the travel time can be highly variable in the face of nonrecurrent events that lead to congestion. The recurrent events such as morning or evening rush hour are expected, hence travelers might leave early or late to reach their destination. But, nonrecurrent events
are random in nature, and thus travelers are not able to prepare or plan for them in advance, contributing in the end to overall customer dissatisfaction. Travel time reliability is a measure of travel time variability due to nonrecurrent events [73, 15]. Estimating the travel time (un)reliability is an important task for traffic planners since it gives measures directly associated with the quality of service experienced by travelers. Moreover, it is important to calculate statistics of the travel time (un)reliability function and, if possible, approximate the entire travel time distribution. In the same vein obtaining reliability estimates for traffic volume can directly inform planners about the quality of service of the system.

The use of simulation to determine travel time and traffic volume distributions is costly if one is to include a representative set of scenarios. The results are in general difficult to transfer and the simulation needs to be repeated many times to provide reasonable statistical inference capabilities. Ideally, the direct use of travel time data and volume data would shed light on the distributions underlying a traffic system but the availability of such data is generally poor. Indeed, for travel time, probe vehicles are the most reliable source of data, but their use is heavily limited to experimental settings due to high costs.

This dissertation deals with modeling a transportation systems’ traffic reliability by means of applied queueing theory. We investigate advanced queueing models to study properties of traffic such as travel time and traffic volume when the system is subject to random deteriorations of the quality of service. We consider both the stationary and transient distributions of the queueing systems and validate them against empirical traffic distributions of travel time and traffic volume. This general methodology allows us to discuss the reliability of the traffic systems that are subject to random deteriorations while avoiding the use of costly simulation.

The focus of our approach is modeling the travel-time and traffic volume (un)reliability
under nonrecurrent congestion due to accidents and adverse weather conditions. We propose a methodology to calculate the analytical approximations of travel time and traffic volume distributions (and hence unreliability measures) that can be calibrated using data readily available from detector stations and that explicitly takes into account the stochastic nature of traffic deteriorating events. These analytical models will help us investigate scenarios that are not supported by the calibration data. The mathematical framework answers an important question: *How the travel time and traffic volume variability depend on the frequency, duration, and severity of random incidents?* Furthermore, we will incorporate the reliability measures into a route optimization scheme in order to reduce congestion and extra cost of delay and pollution. Ultimately, all these models can support traffic control and incident management tools that can lead to safer, more efficient movement of people, goods, and other resources.

1.1 Literature Review

In this section we present an overview of the relevant literature to situate the research in its context. We mainly go through works that belong to the queueing theory, traffic modeling, and/or operations research.

1.1.1 Transient Analysis of a Repairable Queueing System

The study of queueing systems with randomly occurring service deteriorations has found interest from practitioners in many application areas such as production and manufacturing systems, inventory and supply chains, and other service systems such as traffic, and call centers. Responding to this interest, the queueing research community has developed several models that focus on such systems with a number of different variations. In the case of multi-server queue systems with random deteriorations of the service some relevant work has been carried out by Keilson and Servi [44]

Some early transient solutions for simple queueing systems where given by Grassmann and Abate and Whitt in [1, 32], for Markovian queueing systems and in particular the M/M/1 queue via Laplace transforms. The non-stationary behavior of a queueing model for studying traffic flow was given by Heidemann [36]. Recent studies have advanced the study of the transient behavior of queueing systems with customer abandonment due to queue length [67], time-dependent rates [68], and queues subject to disasters together with customer impatience [88].

The traffic flow interrupted by roadway incidents or weather conditions can be modeled as a multi-server queue operating in a multi-state Markovian environment that affects the service process. Particularly relevant are the works by Baykal-Gürsoy and Duan [7], Baykal-Gürsoy and Xiao [9], and Baykal-Gürsoy et al. [8]. In this context an incident is defined as any occurrence that affects capacity of the roadway, thus affecting the service to deteriorate, and may be due to spilled loads, hazardous materials, work zones, traffic accidents, disabled vehicles, or natural causes (e.g., adverse weather conditions).

We are interested in transient solutions of an infinite-server queueing system with Markov-modulated service rates, such as the one described in [9]. The use of stationary distributions can be shown to be useful to describe stationary/long-term traffic conditions (see Chapter 4), and it is expected that the transient solutions will also contribute to the modeling of transient traffic effects, such as queue buildup and dissipation from which reliability metrics can be readily calculated.
We follow an approach similar to [1], in the sense that we apply Laplace Transforms to try to obtain the transient distributions in closed form. Our system however, which will be described in Chapter 2, is more similar to the system considered in [88]. The solution of the stationary case given in [9] gives us useful information: as time goes to infinity, the solution of the transient system will be given in terms of Kummer functions.

1.1.2 The Completion Time of an Interrupted Process

For applications in computer science and telecommunications it is of interest to study the time it takes for a processor to complete a task of random size when the service can be interrupted or slowed down by the occurrence of random failures or higher priority task arrivals to processor sharing systems. This line of research on interrupted service times that leads to our study started in the sixties with the work of Gaver [31] who investigates the completion time distribution of a task experiencing complete service breakdowns during which times the server stops processing the task. Either the interruption can be postponed until the current task is completed or the current task is preempted whenever an interruption occurs and then the server decides on a preemption scheme/discipline. The preemption scheme determines the course of action for the server when a task has been interrupted. Gaver [31] considers postponable interruptions in addition to the following types of preemption strategies: (1) Preemptive *RESUME*, once the interruption is cleared the server continues with the unfinished task from where it left. (2) Preemptive *REPLACE* (repeat different), the server discards the unfinished task to revisit later and selects a different task. (3) Preemptive *RESTART* (repeat identical), the server starts the unfinished task from the beginning. The author derives the Laplace Stieltjest Transform (LST) of the completion time distribution under each interruption/recovery type by counting the
number of repeats until task completion. In [31], the service times are generally distributed while the interruptions arrive as a Poisson process with generally distributed down times. Nicola [62] considers a mixture of interruption types affecting a single server as Poisson arrivals. He obtains the LST of task completion time distribution under various types of interruptions. Kulkarni et al. [48] study a server affected by a Markov modulated environment in which the service rate in each environmental state is different, i.e., the service is deteriorated. They obtain the LST of the completion time distribution under each type of recovery schemes using renewal arguments. Furthermore the authors assume that associated to each environmental state a recovery scheme is fixed as either preemptive RESUME or preemptive RESTART discipline. They generalize their results to the semi-Markovian environment in [49]. Nicola et al. [63] analyze a single server queue with Markov modulated service process, and show that the queue has a block M/G/∞ structure and provide a procedure to evaluate the moment generating function for the stationary distribution of the number of jobs in the system.

For the complete service breakdown case, Gaver [31] is the first to notice that under the RESTART strategy the first two moments of the task completion time may not always exist. Later Fiorini et al. [28] and Sheahan et al. [83] show that under the RESTART strategy and the same exponential up time assumption, the total time it takes to execute a task not including failures follows a power tailed distribution even when the task service time has exponential tail. Asmussen et al. [5] further extend the asymptotic analysis of the RESTART case in Sheahan et al. [83] to more general up time and task time distributions. They notice that the relationship between the up time and task time distributions play an important role impacting the distribution of completion times. Jelenković and Tan [40] independently study the same RESTART strategy and approach the analysis by first proving that the number of restarts is
power tailed, and then using large deviation theory show that the completion times also have power law distribution irrespective of how heavy or light the distributions of task times and up times may be. Jelenković and Tan [41] extend these results to analyze further how a certain functional relationship between the tail distributions of the up time and task time distributions impacts the distributions of the number of restarts and the completions times.

In this dissertation, we study the completion time distribution of a task processed by a server experiencing service interruptions. During these interruptions the server works at a lower service rate. Firstly, we derive the LST of the completion time distributions under both REPLACE and RESTART service disciplines using counting arguments. The approach that we present here yields more detailed results than Kulkarni et al. [49] for our specific cases. Secondly, we show the asymptotic behavior of these distributions and prove that under the RESTART service discipline the completion time has power tail. Finally, using Little’s law, we compare the stationary system size and system time distributions of customers in a two server state M/MSP/$\infty$ queue, and in a M/G/$\infty$ queue in which each server independently of the others can change its state as a two-state Markov process.

### 1.1.3 Travel Time and Traffic Volume Reliability

During the last decade, there has been increasing attention given to the study of travel time reliability [38]. While majority of research has been devoted to the value of travel time reliability [15, 64], the calculation of statistical indices of travel time unreliability still has prominence in the research community due to their direct relation to the level of service (LOS) provided by traffic systems [16, 34, 50, 71].

Since the Federal Highway Administration (FHWA) has defined travel-time reliability as “the consistency or dependability in travel times as measured from day-to-day
or across different times of a day” [73], it has been natural to relate travel time unreliability to travel time variability. One of the most common approaches has been to try to determine some statistical information on the travel time variability in order to compute an index. Examples include the 90th and 95th travel time percentiles, the Buffer Index (BI), the Planning Time Index (PTI), and the Percentage of Trips on Time (PTT) [73]. These indices are very useful since they are easy to communicate and to relate to customer satisfaction.

To calculate any of the travel time variability indices some information of the travel time distribution needs to be known. In fact, the Institute for Transportation Research & Education reported that “...reliability is best described by creating holistic pictures like probability density functions (PDFs) and their associated cumulative density functions (CDFs)” [38]. In this vein some approaches use simulation to come up with the distributions [90] while many others try to adjust analytical distributions to the empirical data [34, 71], and still others use extensive GPS, cell phone, and other sources of data in order to create real-time indices. These distributions have been shown to be unimodal, bimodal or even more complex depending on the mixture of conditions that gave rise to the empirical data. At the same time, the vast number of conditions affecting travel time variability has given rise to specialized models in which the adjusted distributions are differentiated according to their Level of Service (LOS), and according to the causes of variability [50].

There are few works that determine overall travel-time distributions analytically while incorporating all LOS conditions and all sources of travel-time variability. Some approaches include the works presented in [45, 72], both of which employ Markov Chain methods. However, ideally we would want to have models that do not require the use of vehicle-following data for calibration.

The model that we propose in Chapter 4 uses simple analytical results from a
particular queueing system and gives accurate approximations of the traffic travel-time behavior subject to random degradations. These results can be readily calibrated or adjusted to give closed form solutions for various traffic settings. The model describes the time it takes for a customer (a traveler) to transit through a service system that can be in either one of two states: (1) a normal full-speed state and (2) a slower degraded state. These two states alternate according to a stochastic process, and the resulting travel time is a random variable that holds all the variability inherent to the process.

1.1.4 Route Optimization: Risk Averse Shortest Path

We present here some relevant literature on the Shortest Path Problem and some of its variations to place our study in its context.

The Shortest Path Problem (SPP) has provided a long standing line of research that (most notably) stemmed from a seminal work by Dijkstra [21]. As originally formulated, the SPP deals with finding a path of minimum length between two given nodes on a weighted graph, which formally means to identify subsets of arcs in the graph that conform weight-minimal paths from an origin to a destination node. Aggregation of arc weights is usually done additively.

A more general version of the SPP considers time-dependent and/or random weights. If the time dependency is deterministic the problem is usually called the Dynamic Shortest Path Problem (DSPP). If, on the other hand, part of the problem is randomized the problem is called Stochastic Shortest Path Problem (SSPP). These settings are particularly appropriate for modeling transportation activities such as vehicle routing, fleet management, delivery services, etc., which are inherently time-sensitive. In this context, the weight function usually measures the travel time of the arcs (road sections) and the objective is to find shortest travel time paths, which is
at the core of the problem we present in Chapter 5.

For these non-static models the optimal solutions cannot consist of simple paths anymore, but rather have to be expressed as optimal routing policies which dictate optimal decisions to make under time-varying circumstances. Some instances of the DSPP are still solvable using some classic SPP algorithms [93], but this is not generally the case as some decisions that would be sub-optimal for the SPP (such as waiting at a node or making a backward movement) can be part of optimal solutions of DSSPs and SSPPs [6, 78].

Purely dynamic settings do not take into account the rather frequent unexpected changes in traffic conditions that vehicles experience, and it has long been made clear that taking expected values of the random variables in the problem data to reduce a DSSPP into a DSPP leads to overly optimistic paths [54]. Because of this the literature has mostly turned its attention to mixed-stochastic approaches for the SPP and related problems [29, 57], giving rise to the Dynamic and Stochastic Shortest Path Problem (DSSPP). The dynamic and stochastic components are used to encode (but not exclusively): the possibility of having to change a decision [13], recurrent congestion [29, 52], non-recurrent congestion [89, 35], and the availability of real time information from intelligent transportation systems (ITS) [46]. Discrete-time Markov chains are a common technique to model the dynamic-stochastic processes, although continuous-time random variables and queueing theory are also used. For random travel times, discrete distributions are most prevalent to ensure computational tractability.
Expectations and variances are used in the objective functions to aggregate the random effects, and utility functions are commonly employed to establish preference of the random variables and model risk-aversion [80]. If calibrated travel time random variables like the ones developed in Chapter 3 are employed for the vulnerable links in a road network, a realistic risk-averse shortest travel time path problem can be defined and studied.
Chapter 2

Transient Analysis of a Repairable Queueing System

In this chapter we show several advances towards analytically solving the partial differential equations that govern the transient behavior of an infinite-server queueing system with Markov-modulated service rates. We show the connections to the stationary behavior of such queues that was reported by Baykal-Gürsoy and Xiao [9]. When possible, we also look for analytical convergence results which can help determine the validity of using stationary solutions as approximations to the transient behavior of the system.

The objective that we have in mind and that motivates this study is to determine if the resulting analytical forms yield any insights to applied problems such as real-time traffic congestion estimation, traffic reliability management, and traffic control.

We start by describing in section 2.1 the queueing system and the differential equations that govern the transient system along with what it is known in the literature for the stationary solution. In section 2.2, we show how to transform the governing partial differential equation (PDE) system into a single hyperbolic second order PDE to facilitate the treatment of the equations. In section 2.3, we detail our
advances in the use of Laplace transform and we give some formal solutions in transformed domains for particular cases of the system. Section 2.4 shows a methodology for calculating all the moments of the transient solution, we give formal relations in the frequency domain and show explicitly the transient (time-dependent) expected value of the number of customers in the system. Additionally we give preliminary results on the rate of convergence of the transient mean to the stationary solution in terms of the model parameters. In section 2.5, we open the frequency domain PDE by using series expansions to obtain another formal solution in the form of a well determined infinite system of equations for the transformed time-dependent probability mass function of the number in the system. We finalize with section 2.6 where we discuss future work and some guidelines as to how the study should proceed.

2.1 PDE System

We consider a service system with an infinite number of servers subject to random interruptions of exponentially distributed durations. During interruptions, all servers work at lower efficiency compared to their normal functioning state. The service rate of each server is $\mu > 0$ in the absence of interruptions, decreases to $\mu' \geq 0$ at the arrival of an interruption, and recovers back to $\mu$ at the clearance of the interruption. If $\mu' > 0$, then the system is called a partial failure system, and if $\mu' = 0$ it is called a total failure system. We assume that interruptions arrive according to a Poisson process with rate $f$, and the repair time is exponentially distributed with rate $r$. The customer arrivals are in accordance with a homogeneous Poisson process with intensity $\lambda$. The interruption and customer arrival processes and the service and repair times are all assumed to be mutually independent. The stochastic process $\{X(t), U(t)\}$ describes the state of the system at time $t$, where $X(t)$ is the number of customers in the system at time $t$, and $U(t)$ is the status of the system. If at
time $t$, the system is experiencing an interruption, then $U(t)$ is equal to $F$ (failure), otherwise, $U(t)$ is $N$ (normal). Keep in mind that the failures considered in this proposal are partial failures in the sense that in general all servers continue to work under deteriorated service rate. The system is said to be in state $(i, F)$, if there are $i$ customers in the system in which the servers all work at a lower rate, while the system is said to be in state $(i, N)$ if there are $i$ customers in the system that is functioning as normal. We denote the transient probability of the system being in state $(i, F)$ at time $t$ by $P_{iF}(t)$ and the transient probability of the system being in state $(i, N)$ at time $t$ by $P_{iN}(t)$. The transition rate diagram of the system described is given in Figure 2.1.

![Rate Transition Diagram](image)

Figure 2.1: Rate Transition Diagram

Let $G(z, t)$ denote the probability generating function of the system for $|z| \leq 1$, a complex variable, and $t \geq 0$, i.e.

$$G(z, t) = \sum_{i=0}^{\infty} (P_{iN}(t) + P_{iF}(t)) z^i. \quad (2.1)$$

$G(z, t)$ can be decomposed into

$$G(z, t) = G_N(z, t) + G_F(z, t), \quad (2.2)$$

where $G_N(z, t)$ represents the partial generating function for the normal state and $G_F(z, t)$ is the partial generating function for the failure state. Explicitly, $G_N(z, t) = \sum_{i=0}^{\infty} P_{iN}(t) z^i$ and $G_F(z, t) = \sum_{i=0}^{\infty} P_{iF}(t) z^i$. 
The following are the boundary conditions of the system:

1. Since $G(z,t)$ is a probability generating function we set a normalization condition $G(1,t) = 1$, $\forall t$.

2. We arbitrarily assume that the system initializes in normal state with zero customers, that is $(X(0), U(0)) = (0, N)$. For this we set $P_{0N}(0) := 1$, $P_{iN}(0) := 0$ for all $i > 0$, and $P_{IF}(0) := 0$ for all $i \geq 0$. Also, this means that $G(z,0) = 1$.

The global balance equations of the system in Figure 2.1 yield the PDE system shown below.

\[
\frac{\partial}{\partial t} G_{N}(z,t) = -\mu (z-1) \frac{\partial}{\partial z} G_{N}(z,t) + [\lambda (z-1) - f] G_{N}(z,t) + r G_{F}(z,t), \quad (2.3)
\]

\[
\frac{\partial}{\partial t} G_{F}(z,t) = -\mu' (z-1) \frac{\partial}{\partial z} G_{F}(z,t) + [\lambda (z-1) - r] G_{F}(z,t) + f G_{N}(z,t). \quad (2.4)
\]

We know a couple of things related to these equations. Consider a simple system with no interruptions, i.e. $r = f = 0$, governed by equation (2.5).

\[
\frac{\partial}{\partial t} G(z,t) = -\mu (z-1) \frac{\partial}{\partial z} G(z,t) + [\lambda (z-1)] G(z,t). \quad (2.5)
\]

For this system the transient solution can be found in [43] and is given by

\[
G(z,t) = \exp \{ (z-1)(1 - e^{-\mu t}) (\lambda/\mu) \}. \quad (2.6)
\]

Our service-modulated system reduces to this simplified model if $\mu' = \mu$, and also if $f = 0$. If $r = 0$ the system will converge in a finite amount of time to such a simplified system. To avoid this simplification we assume henceforth that both $f$ and $r$ are positive quantities (that is $f > 0$ and $r > 0$).

For the steady state case as $t \to \infty$, the solution of system (2.3)-(2.4) is given in terms of Kummer functions as shown in equation (2.7) (this result was obtained by
Keilson and Servi [44] and Baykal-Gürsoy and Xiao in 2004 [9]).

\[
G(z) = \exp \left[ \frac{\lambda}{\mu} (z - 1) \right] \cdot \left[ \frac{f \mu' + r \mu}{\mu' (f + r)} \right] \cdot M \left( \frac{f}{\mu}, \frac{f}{\mu} + \frac{r}{\mu'}, \left( \frac{\lambda}{\mu'} - \frac{\lambda}{\mu} \right) (z - 1) \right) \\
+ \frac{r (\mu' - \mu)}{\mu' (f + r)} \cdot M \left( \frac{f}{\mu}, \frac{f}{\mu} + \frac{r}{\mu'} + 1, \left( \frac{\lambda}{\mu'} - \frac{\lambda}{\mu} \right) (z - 1) \right)
\]

(2.7)

We now show how system (2.3)-(2.4) can be reduced to a single second order PDE by the method of characteristics.

## 2.2 Governing PDE

In order to solve the above PDE system, we introduce the notation \(u(z, t) := G_N(z, t)\), \(v(z, t) := G_F(z, t)\). In order to avoid cluttering, and as long as it remains clear, we will use \(u\) and \(v\) instead of \(u(z, t)\) and \(v(z, t)\), respectively. System (2.3)-(2.4) becomes:

\[
u_t - \mu u_z = -[\lambda(1 - z) + f]u + rv, \quad (2.8)
\]

\[
v_t - \mu' v_z = -[\lambda(1 - z) + r]v + fu. \quad (2.9)
\]

Let, \(1 - z = e^{-s}\), so that \(s = -\ln(1 - z)\), \(z = 0\) implies \(s = 0\), and \(z = 1\) implies \(s = \infty\). Noting that \(u_z = u_s/(1 - z)\), from equations (2.8)-(2.9) we get

\[
u_t - \mu u_s = -[\lambda e^{-s} + f]u + rv, \quad (2.10)
\]

\[
v_t - \mu' v_s = -[\lambda e^{-s} + r]v + fu. \quad (2.11)
\]

By writing the Lagrange-Charpit equations for equation (2.10):

\[
\frac{dt}{1} = \frac{ds}{-\mu},
\]

which, parametrically, gives us:

\[
s(t) = -\mu. \quad (2.12)
\]
Solving (2.12), and working analogously from equation (2.11) for \(v\), we obtain:

\[ s_1(t) = -\mu t + s_1(0), \]
\[ s_2(t) = -\mu' t + s_2(0). \]

This motivates the introduction of two new independent variables,

\[ \xi = \frac{s + \mu t}{\mu - \mu'} > 0, \]
\[ \eta = \frac{s + \mu' t}{\mu - \mu'} > 0. \]

giving

\[ t = \xi - \eta, \]

from which \(t = 0\) implies \(\xi = \eta\), and

\[ s = \mu \eta - \mu' \xi, \]

from which \(s = 0\) implies \(\xi = (\mu / \mu')\eta > \eta\).

The partial derivatives with respect to the new variables are

\[ u_\eta = -u_t + \mu u_s = (\lambda e^{-s} + f)u - rv, \quad \text{(2.13)} \]
\[ v_\xi = v_t - \mu' v_s = -(\lambda e^{-s} + r)v + fu. \quad \text{(2.14)} \]

By taking another derivative with respect to \(\xi\) of the first equation and substituting \(rv = (\lambda e^{-s} + f)u - u_\eta\) we get

\[ u_{\eta \xi} = (\lambda e^{-s} + f)u_\xi - (\lambda e^{-s} + r)u_\eta + \lambda e^{-s}(\lambda e^{-s} + f + r + \mu')u \]
\[ = a(s)u_\xi - b(s)u_\eta + c(s)u. \quad \text{(2.15)} \]

with \(a(s) = \lambda e^{-s} + f\), \(b(s) = \lambda e^{-s} + r\), and \(c(s) = \lambda e^{-s}(\lambda e^{-s} + f + r + \mu')\).
Chain rule gives that $u_\xi = u_t - \mu' u_s$, taking a partial derivative with respect to $\eta$ we obtain

$$u_{\eta \xi} = (u_t)_{\eta} - \mu'(u_s)_{\eta}$$

$$= -u_{tt} + \mu u_{st} + \mu' u_{st} - \mu \mu' u_{ss}. \quad (2.16)$$

Finally, combining (2.15) and (2.16) we get

$$-u_{tt} + (\mu + \mu')u_{st} - \mu \mu' u_{ss} = a(s)(u_t - \mu' u_s) + b(s)(-u_t + \mu u_s) + c(s)u$$

$$= (a(s) - b(s))u_t - (\mu' a(s) - \mu b(s))u_s + c(s)u.$$  

Substituting for $a(s)$, $b(s)$, and $c(s)$, we obtain

$$-u_{tt} + (\mu + \mu')u_{st} - \mu \mu' u_{ss} = (f - r)u_t - (\lambda e^{-s}(\mu' - \mu) + \mu' f - \mu r)u_s + \lambda e^{-s}(\lambda e^{-s} + f + r + \mu')u.$$  

(2.17)

Returning to $z$ with $z = 1 - e^{-s}$, we have,

$$u_s = u_z(1 - z),$$

$$u_{ss} = u_{zz}(1 - z)^2 - u_z(1 - z),$$

substituting in (2.17) gives

$$-u_{tt} + (\mu + \mu')u_{zt}(1 - z) - \mu \mu'(u_{zz}(1 - z)^2 - u_z(1 - z))$$

$$= (f - r)u_t - (\lambda (1 - z)(\mu' - \mu) + \mu' f - \mu r)u_z(1 - z) + \lambda(1 - z)(\lambda(1 - z) + f + r + \mu')u,$$

that is equivalent to

$$-u_{tt} + u_{zt}(\mu + \mu')(1 - z) - u_t(f - r) - u_{zz}\mu'(1 - z)^2 +$$

$$u_z(1 - z)(\mu' + \lambda(1 - z)(\mu' - \mu) + \mu' f - \mu r) - (\lambda^2(1 - z)^2 + \lambda(f + r + \mu')(1 - z))u = 0.$$  

(2.18)
This is an hyperbolic PDE. To see this, call \( A(z, t) = -1, \) \( B(z, t) = (\mu + \mu')(1 - z), \) and \( C(z, t) = -\mu\mu'(1 - z)^2 \) (that are the coefficients of the \( u_{tt}, u_{zt}, \) and \( u_{zz} \) terms, respectively). Then, the discriminant \( B^2 - 4AC \) is such that:

\[
B^2 - 4AC = (\mu + \mu')^2(1 - z)^2 - 4\mu\mu'(1 - z)^2 \\
= (1 - z)^2[(\mu + \mu')^2 - 4\mu\mu'] \\
= (1 - z)^2[\mu^2 - 2\mu\mu' + \mu'^2] \\
= (1 - z)^2(\mu - \mu')^2 \\
> 0,
\]

which makes the equation hyperbolic.

### 2.3 Laplace Transform Treatment

We now change to the frequency domain through the Laplace transform, in order to transform our PDE into an ODE. Taking the Laplace transform \( \hat{u}(z, \tau) = \int_0^\infty e^{-\tau t}u(z, t)dt \) in (2.18) we have:

\[
-\tau^2\hat{u} - \tau u(z, 0) - u_t(z, 0) + (\tau \hat{u}_z - u_z(z, 0))(\mu + \mu')(1 - z) - (\tau \hat{u} - u(z, 0))(f - r) - \hat{u}_{zz}\mu\mu'(1 - z)^2 + \\
\hat{u}_z(1 - z)(\mu\mu' + \lambda(1 - z)(\mu' - \mu) + \mu'f - \mu r) - (\lambda^2(1 - z)^2 + \lambda(f + r + \mu')(1 - z))\hat{u} = 0.
\]

Now, due to the initial conditions and the boundary equations we have:

\[
\begin{align*}
 u(z, 0) &= G_N(z, 0) = P_{0N} + \sum_{i=1}^{\infty} P_{iN}(0)z^i = 1 + 0 = 1, \\
 u_z(z, 0) &= \sum_{i=1}^{\infty} iP_{iN}(0)z^{i-1} = 0, \\
 u_t(z, 0) &= P_{0N}'(0) + P_{1N}'(0)z + \sum_{i=2}^{\infty} P_{iN}'(0)z^i = -(\lambda + f) + \lambda z + 0 = -(\lambda + f) + \lambda z.
\end{align*}
\]
Hence we have

\[-(\tau^2 \dot{u} - \tau + (\lambda + f) - \lambda z) + \tau \dot{u}_z(\mu + \mu')(1 - z) - (\tau \dot{u} - 1)(f - r) - \dot{u}_{zz}\mu \mu'(1 - z)^2 + \hat{u}_z(1 - z)(\mu \mu' + \lambda(1 - z)(\mu' - \mu) + \mu' f - \mu r) - (\lambda^2 (1 - z)^2 + \lambda(f + r + \mu')(1 - z)) \dot{u} = 0.\]

Reordering the terms we obtain,

\[-\tau^2 \dot{u} + \tau - \lambda(1 - z) - f + \tau \dot{u}_z(\mu + \mu')(1 - z) - \tau \dot{u}(f - r) + f - r - \dot{u}_{zz}\mu \mu'(1 - z)^2 + \hat{u}_z(1 - z)(\mu \mu' + \lambda(1 - z)(\mu' - \mu) + \mu' f - \mu r) - (\lambda^2 (1 - z)^2 + \lambda(f + r + \mu')(1 - z)) \dot{u} = 0,
\]

which leads to,

\[-\hat{u}_{zz}\mu \mu'(1 - z)^2 + \hat{u}_z(1 - z)\left(\mu \mu' + \lambda(1 - z)(\mu' - \mu) + \mu' f - \mu r + \tau(\mu + \mu')\right) - \hat{u}\left(\lambda^2 (1 - z)^2 + \lambda(f + r + \mu')(1 - z) + \tau^2 + \tau(f - r)\right) = -\tau + \lambda(1 - z) + r.\quad (2.19)
\]

Note that for \(x = 1 - z\), we have \(\hat{u}_z(z, \tau) = -\hat{u}_x(x, \tau)\) and \(\hat{u}_{zz}(z, \tau) = \hat{u}_{xx}(x, \tau)\). Then by substituting \((1 - z) = x\) and reordering, we get

\[\hat{u}_{xx}\mu \mu' x^2 + \hat{u}_x\left(x^2 \lambda(\mu' - \mu) + x(\mu \mu' + \mu' f - \mu r + \tau(\mu + \mu'))\right) + \hat{u}\left(x^2 \lambda^2 + x \lambda(f + r + \mu') + \tau^2 + \tau(f - r)\right) = -\lambda x + \tau - r,\quad (2.20)
\]
or the alternative form,

\[\hat{u}_{xx} + \hat{u}_x\left(\frac{\lambda(\mu' - \mu)}{\mu \mu'} + \frac{1}{x} \frac{\mu \mu' + \mu' f - \mu r + \tau(\mu + \mu')}{\mu \mu'}\right) + \hat{u}\left(\frac{\lambda^2}{\mu \mu'} + \frac{\lambda (f + r + \mu')}{x \mu' \mu} + \frac{\tau (\tau + f - r)}{x^2 \mu' \mu}\right) = -\lambda \frac{1}{\mu \mu'} x + \frac{1}{x^2} \tau - r.\quad (2.21)
\]

### 2.3.1 Special Case: Total Failure System (\(\mu' = 0\))

The special case with \(\mu' = 0\), which corresponds to a full breakdown system, leads to a formal solution for the transient behavior of the system. Here (2.20) becomes,

\[\hat{u}_x\left(-x^2 \lambda \mu + x(-\mu r + \tau \mu)\right) + \hat{u}\left(x^2 \lambda^2 + x \lambda(f + r) + \tau^2 + \tau(f - r)\right) = -\lambda x + \tau - r,
\]
Which can be written as a linear ODE with variable coefficients:

\[ \dot{u}_x x \mu (-x \lambda - r + \tau) + \dot{u} (x^2 \lambda^2 + x \lambda (f + r) + \tau^2 + \tau(f - r)) = -\lambda x + \tau - r \]

\[ \Rightarrow \dot{u}_x + \frac{x^2 \lambda^2 + x \lambda (f + r) + \tau^2 + \tau(f - r)}{x \mu (x \lambda + r - \tau)} = \frac{1}{x \mu}, \]

\[ \Rightarrow \dot{u}_x + \frac{x \lambda + r - \tau)(x \lambda + f + \tau) + 2 \tau(\tau + f - r) - fr}{x \mu (x \lambda + r - \tau)} = \frac{1}{x \mu}, \]

\[ \Rightarrow \dot{u}_x + \frac{\left[ \frac{x \lambda + f + \tau}{x \mu} + \frac{2 \tau(\tau + f - r) - fr}{x \mu (x \lambda + r - \tau)} \right]}{\frac{x \lambda + r - \tau}{x \mu}} = \frac{1}{x \mu}. \] (2.22)

We define the integrating factor for the ODE (2.22) as:

\[ M(x, \tau) = \exp \left( \int \frac{-\lambda}{\mu} \frac{f + \tau}{x} \frac{1}{x \mu} \frac{2 \tau(\tau + f - r) - fr}{x \mu} \frac{1}{x(x \lambda + r - \tau)} \cdot dx \right) \]

\[ = \exp \left( -\frac{1}{\mu} \int \frac{x}{x \lambda + r - \tau} \left( \lambda x + (f + \tau) \log |x| + 2 \tau(\tau + f - r) - fr \frac{\log |x| - \log |x \lambda + r - \tau|}{r - \tau} \right) \cdot dx \right) \]

\[ = \exp \left( -\frac{\lambda}{\mu} x \frac{f + \tau}{\mu} \log |x| \frac{2 \tau(\tau + f - r) - fr}{\mu(r - \tau)} \log |x| + \frac{2 \tau(f + \tau - r) - fr}{\mu(r - \tau)} \log |x \lambda + r - \tau| \right) \]

\[ = \exp \left( -\frac{\lambda}{\mu} x \right) \cdot \frac{f + \tau}{\mu} \frac{2 \tau(\tau + f - r) - fr}{\mu(r - \tau)} \cdot \frac{\log |x| - \log |x \lambda + r - \tau|}{r - \tau} \]

Then, formally, the solution of (2.22) is given by equation (2.23).

\[ \hat{u}(x, \tau) = 1/M(x, \tau) \cdot \int_{x_0}^{x} \frac{1}{\mu y} M(y, \tau) dy + C/M(x, \tau). \] (2.23)

### 2.4 Moment Calculation

We devote our efforts now to the calculation of the moments of the number of customers \(X(t)\). We show a recursive definition for the factorial moments, explicit transforms for the expectation and variance, and an explicit time-dependent expression for the expectation. We also study the dominant terms in the convergence of the expectation to the stationary value.
2.4.1 First Moment

To obtain the first moment, we need to evaluate the first derivative of the time-dependent generating function $G(z, t)$ with respect to $z$ at $z = 1$, i.e.,

$$
\frac{\partial}{\partial z} G(z, t) \bigg|_{z=1} = \frac{\partial}{\partial z} (G_N(z, t) + G_F(z, t)) \bigg|_{z=1} = (u_z(z, t) + v_z(z, t)) \bigg|_{z=1}.
$$

Equations (2.8) and (2.9) evaluated at $z = 1$ are given in their full form as

$$
u_t(1, t) = -fu(1, t) + rv(1, t),
$$

$$
v_t(1, t) = -rv(1, t) + fu(1, t).
$$

Which yields,

$$
u(1, t) = \frac{1}{f + r} \left( C_1( f e^{-t(f+r)} + r ) - C_2 r ( e^{-t(f+r)} - 1 ) \right),
$$

$$
v(1, t) = \frac{1}{f + r} \left( C_2 ( r e^{-t(f+r)} + f ) - C_1 f ( e^{-t(f+r)} - 1 ) \right).
$$

We have assumed that $u(0, 0) = 1$, thus $u(1, 0) = 1$ and $v(1, 0) = 0$, providing two boundary equations

$$
f + r = C_1(f + r),
$$

$$
0 = C_2(r + f),
$$

implying $C_1 = 1$ and $C_2 = 0$. Hence equation (2.24) and (2.25) become

$$
u(1, t) = \frac{f}{f + r} e^{-t(f+r)} + \frac{r}{f + r} = \frac{1}{f + r} \left( f e^{-t(f+r)} + r \right),
$$

$$
v(1, t) = -\frac{f}{f + r} e^{-t(f+r)} + \frac{f}{f + r} = \frac{1}{f + r} \left( -f e^{-t(f+r)} + f \right).
$$

Now, using Laplace transform on (2.26) and (2.27) we obtain:

$$
\hat{u}(1, s) = \frac{1}{f + r} \left( \frac{f}{s + f + r} + \frac{r}{s} \right),
$$

$$
\hat{v}(1, s) = \frac{1}{f + r} \left( \frac{-f}{s + f + r} + \frac{f}{s} \right).
$$
Gathering terms we obtain,

\begin{align*}
  u_{t_z} + \mu u_z - \mu (1-z) u_{zz} &= \lambda u - \lambda (1-z) u_z - fu_z + rv_z, \\
  v_{t_z} + \mu' v_z - \mu' (1-z) v_{zz} &= \lambda v - \lambda (1-z) v_z - rv_z + fu_z.
\end{align*}  

Using Laplace transform:

\begin{align*}
  s^2 \hat{u}_z(1,s) - u_z(1,0) + (\mu + f) \hat{u}_z(1,s) - \lambda \hat{u}(1,s) - r \hat{v}_z(1,s) &= 0, \\
  s^2 \hat{v}_z(1,s) - v_z(1,0) + (\mu' + r) \hat{v}_z(1,s) - \lambda \hat{v}(1,s) - f \hat{u}_z(1,s) &= 0.
\end{align*}  

Taking \( z = 1 \) and gathering terms we get:

\begin{align*}
  u_{t_z}(1,t) + (\mu + f) u_z(1,t) - \lambda u(1,t) - rv_z(1,t) &= 0, \\
  v_{t_z}(1,t) + (\mu' + r) v_z(1,t) - \lambda v(1,t) - fu_z(1,t) &= 0.
\end{align*}  

Using Laplace transform:

\begin{align*}
  s^2 \hat{u}_z(1,s) - u_z(1,0) + (\mu + f) \hat{u}_z(1,s) - \lambda \hat{u}(1,s) - r \hat{v}_z(1,s) &= 0, \\
  s^2 \hat{v}_z(1,s) - v_z(1,0) + (\mu' + r) \hat{v}_z(1,s) - \lambda \hat{v}(1,s) - f \hat{u}_z(1,s) &= 0.
\end{align*}  

Replacing the transforms (2.28) and (2.29), using the fact that \( u_z(1,0) = 0 \) and \( v_z(1,0) = 0 \) we get:

\begin{align*}
  s^2 \hat{u}_z(1,s) + (\mu + f) \hat{u}_z(1,s) - \lambda \frac{1}{f + r} \left( \frac{f}{s + f + r} + \frac{r}{s} \right) - r \hat{v}_z(1,s) &= 0, \\
  s^2 \hat{v}_z(1,s) + (\mu' + r) \hat{v}_z(1,s) - \lambda \frac{1}{f + r} \left( -\frac{f}{s + f + r} + \frac{f}{s} \right) - f \hat{u}_z(1,s) &= 0.
\end{align*}  

Gathering terms we obtain,

\begin{align*}
  \hat{u}_z(1,s)(s + (\mu + f)) - r \hat{v}_z(1,s) &= \frac{\lambda}{f + r} \left( \frac{f}{s + f + r} + \frac{r}{s} \right), \\
  \hat{v}_z(1,s)(s + (\mu' + r)) - f \hat{u}_z(1,s) &= \frac{\lambda}{f + r} \left( -\frac{f}{s + f + r} + \frac{f}{s} \right).
\end{align*}  

These equations can be written as:

\begin{equation}
  \begin{bmatrix}
    s + \mu + f & -r \\
    -f & s + \mu' + r
  \end{bmatrix}
  \begin{bmatrix}
    \hat{u}_z(1,s) \\
    \hat{v}_z(1,s)
  \end{bmatrix}
  = 
  \frac{\lambda}{f + r}
  \begin{bmatrix}
    f \\
    -f
  \end{bmatrix}
  \begin{bmatrix}
    s + f + r \\
    s + f + r
  \end{bmatrix}
  \begin{bmatrix}
    r \\
    \frac{r}{s}
  \end{bmatrix}.
\end{equation}  

of which its solution is given as

\begin{equation}
  \begin{bmatrix}
    \hat{u}_z(1,s) \\
    \hat{v}_z(1,s)
  \end{bmatrix}
  = 
  \frac{\lambda}{(s + \mu + f)(s + \mu' + r) - fr}
  \begin{bmatrix}
    s + \mu' + r & r \\
    f & s + \mu + f
  \end{bmatrix}
  \begin{bmatrix}
    f \\
    -f
  \end{bmatrix}
  \begin{bmatrix}
    s + f + r \\
    s + f + r
  \end{bmatrix}
  \begin{bmatrix}
    r \\
    \frac{r}{s}
  \end{bmatrix}.
\end{equation}
O’Cinneide and Purdue [65], computed a general form for the transient moments of an M/M/∞ queue in a random environment. Here we have obtained the explicit form for our case.

We can simplify each component separately,

\[ \hat{u}_z(1, s) = \frac{\lambda}{(s + \mu + f)(s + \mu' + r)} - fr \left[ (s + \mu' + r) \left( \frac{f}{s + f + r} + \frac{r}{s} \right) + r \left( \frac{-f}{s + f + r} + \frac{f}{s} \right) \right] \]

\[ = \frac{\lambda}{(s + \mu + f)(s + \mu' + r)} - fr \left[ (s + \mu) \left( \frac{f}{s + f + r} + \frac{r}{s} \right) + r \left( \frac{f + r}{s} \right) \right] \]

\[ \hat{v}_z(1, s) = \frac{\lambda}{(s + \mu + f)(s + \mu' + r)} - fr \left[ f \left( \frac{f}{s + f + r} + \frac{r}{s} \right) + (s + \mu + f) \left( \frac{-f}{s + f + r} + \frac{f}{s} \right) \right] \]

\[ = \frac{\lambda}{(s + \mu + f)(s + \mu' + r)} - fr \left[ (s + \mu) \left( \frac{-f}{s + f + r} + \frac{f}{s} \right) + f \left( \frac{f + r}{s} \right) \right] \]

\[ \hat{u}_z(1, s) = \frac{\lambda}{(s + \mu + f)(s + \mu' + r)} - fr \left[ 1 \cdot \left( \frac{(s + \mu')(s + r)}{s + f + r} + r \right) \right] \]

\[ \hat{v}_z(1, s) = \frac{\lambda}{(s + \mu + f)(s + \mu' + r)} - fr \left[ 1 \cdot \left( \frac{f(s + \mu)}{s + f + r} + f \right) \right] \]

The Laplace transform of the transient expectation is given by the sum of (2.32) and
(2.33):

\[
\hat{u}_z(1, s) + \hat{v}_z(1, s) = \frac{\lambda}{(s + \mu + f)(s + \mu' + r) - fr} \cdot \frac{1}{s} \cdot \left[ \frac{(s + \mu')(s + r) + f(s + \mu)}{s + f + r} + f + r \right]
\]

\[
= \frac{\lambda}{(s + \mu + f)(s + \mu' + r) - fr} \cdot \frac{1}{s} \cdot \left[ \frac{(s + f + r)^2 + \mu f + \mu'(s + r)}{s + f + r} \right]
\]

\[
= \frac{\lambda}{(s + \mu + f)(s + \mu' + r) - fr} \cdot \frac{1}{s} \cdot \left[ (s + f + r) + \frac{\mu f + \mu'(s + r)}{s + f + r} \right].
\]

\[ (2.34) \]

It is known that, if the limit \( \lim_{t \to \infty} f(t) \) exists, then \( \lim_{s \to 0} s F(s) = \lim_{t \to \infty} f(t) \), where \( F(s) \) is the Laplace transform of \( f(t) \). Since the limiting distribution with finite moments has already been obtained by Baykal-Gürsoy [9] we can obtain the stationary expected value as:

\[
\lim_{t \to \infty} E[X(t)] = \lim_{s \to 0} s \cdot (\hat{u}_z(1, s) + \hat{v}_z(1, s))
\]

\[
= \lim_{s \to 0} \frac{\lambda}{(s + \mu + f)(s + \mu' + r) - fr} \cdot \left[ (s + f + r) + \frac{\mu f + \mu'(s + r)}{s + f + r} \right]
\]

\[
= \frac{\lambda}{(\mu + f)(\mu' + r) - fr} \left[ (f + r) + \frac{\mu f + \mu'r}{f + r} \right],
\]

\[ (2.35) \]

which of course matches the result in [9].

Consider now the polynomial in the denominator of the first factor in (2.34), and call \( s_1 \) and \( s_2 \) the roots of the corresponding equation, (2.36):

\[
(s + \mu + f)(s + \mu' + r) - fr = 0.
\]

\[ (2.36) \]

Equation (2.34) can be written as

\[
E[X(s)] = \frac{\lambda}{(s - s_1)(s - s_2)} \cdot \frac{1}{s} \cdot \left[ s + (f + r + \mu) + \frac{f(\mu - \mu')}{s + f + r} \right]
\]

\[
= \lambda \left[ \frac{1}{(s - s_1)(s - s_2)} + \frac{f + r + \mu'}{s(s - s_1)(s - s_2)} + \frac{f(\mu - \mu')}{s(s + f + r)(s - s_1)(s - s_2)} \right].
\]
Using partial fraction expansion and gathering terms:

\[
\frac{E[X(s)]}{\lambda} = \frac{1}{s - s_1 (s_1 - s_2)} \left[ 1 + \frac{1}{s_1} \left( f + r + \mu' + \frac{f(\mu - \mu')}{s_1 + f + r} \right) \right] \\
+ \frac{1}{s - s_2 (s_2 - s_1)} \left[ 1 + \frac{1}{s_2} \left( f + r + \mu' + \frac{f(\mu - \mu')}{s_2 + f + r} \right) \right] \\
- \frac{1}{s + f + r} \cdot \frac{f(\mu - \mu')}{(f + r)(s_1 + f + r)(s_2 + f + r)} \\
+ \frac{1}{s} \cdot \frac{1}{s_1 s_2} \left[ f + r + \mu' + \frac{f(\mu - \mu')}{f + r} \right].
\]

Finally, inverting the Laplace transform:

\[
E[X(t)] = \frac{\lambda}{(s_1 - s_2)} \left[ 1 + \frac{1}{s_1} \left( f + r + \mu' + \frac{f(\mu - \mu')}{s_1 + f + r} \right) \right] e^{s_1 t} \\
+ \frac{\lambda}{(s_2 - s_1)} \left[ 1 + \frac{1}{s_2} \left( f + r + \mu' + \frac{f(\mu - \mu')}{s_2 + f + r} \right) \right] e^{s_2 t} \\
- \frac{\lambda f(\mu - \mu')}{(f + r)(s_1 + f + r)(s_2 + f + r)} e^{-(f+r)t} \\
+ \frac{\lambda}{s_1 s_2} \left[ f + r + \mu' + \frac{f(\mu - \mu')}{f + r} \right].
\]

Figure 2.2 shows the transient expectation obtained in equation (2.37) for parameters \( \lambda = 8, \mu = 0.4, \mu' = 0.3, r = 0.04, \) and \( f = 0.2. \) The solid and dashed red lines correspond to the components \( u_z(1,t) \) and \( v_z(1,t), \) respectively; the solid black line is \( E[X(t)], \) and the dotted line is the stationary expectation given by the limit in equation (2.35). This system has, relatively, both low service and repair rates.

Let us see \( s_1 \) and \( s_2 \) explicitly. Opening the polynomial in (2.36) we have

\[
(s + \mu + f)(s + \mu' + r) - fr = s^2 + (f + r + \mu + \mu')s + \mu \mu' + f \mu' + r \mu,
\]

that provides the two roots as

\[
s_1 = \frac{-(f + r + \mu + \mu') + \sqrt{(f + r + \mu + \mu')^2 - 4(\mu \mu' + f \mu' + r \mu)}}{2}, \quad (2.38)
\]

\[
s_2 = \frac{-(f + r + \mu + \mu') - \sqrt{(f + r + \mu + \mu')^2 - 4(\mu \mu' + f \mu' + r \mu)}}{2}. \quad (2.39)
\]
It holds that $s_1$ and $s_2$ are both negative real numbers. In fact the radicand in (2.38) and (2.39) is non-negative:

$$(f + r + \mu + \mu')^2 - 4(\mu\mu' + f\mu' + r\mu) = (f + r - \mu + \mu')^2 + 4f(\mu - \mu') \geq 0.$$

It is clear that $s_2$ is trivially negative, and the negativity of $s_1$ follows from the fact that $(f + r + \mu + \mu')^2 - 4(\mu\mu' + f\mu' + r\mu) \leq (f + r + \mu + \mu')^2$. Indeed:

$$0 \leq (f + r + \mu + \mu')^2 - 4(\mu\mu' + f\mu' + r\mu) \leq (f + r + \mu + \mu')^2$$

$$\Rightarrow \sqrt{(f + r + \mu + \mu')^2 - 4(\mu\mu' + f\mu' + r\mu)} \leq (f + r + \mu + \mu')$$

$$\Rightarrow - (f + r + \mu + \mu') + \sqrt{(f + r + \mu + \mu')^2 - 4(\mu\mu' + f\mu' + r\mu)} \leq 0$$

$$\Rightarrow \frac{-(f + r + \mu + \mu') + \sqrt{(f + r + \mu + \mu')^2 - 4(\mu\mu' + f\mu' + r\mu)}}{2} = s_1 \leq 0$$

Finally, it is clear that $s_1 \geq s_2$.

For the case in which $\mu' = \mu$, $s_1$ reduces to $-\mu$ and $s_2$ reduces to $-(f + r + \mu)$, as
shown below

\[
s_1 = \frac{-(f + r + \mu + \mu') + \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')}}{2} \\
= \frac{-(f + r + 2\mu) + \sqrt{(f + r)^2}}{2} \\
= \frac{-(f + r + 2\mu) + (f + r)}{2} \\
= -\mu,
\]

\[
s_2 = \frac{-(f + r + \mu + \mu') - \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')}}{2} \\
= \frac{-(f + r + 2\mu) - \sqrt{(f + r)^2}}{2} \\
= \frac{-(f + r + 2\mu) - (f + r)}{2} \\
= -(f + r + \mu).
\]

Thus, in this case, the first moment given in (2.37) reduces to:

\[
E[X(t)] = \frac{\lambda}{(s_1 - s_2)} \left[ 1 + \frac{1}{s_1} (f + r + \mu) \right] e^{st} + \frac{\lambda}{s_1 s_2} (f + r + \mu) \\
= \frac{\lambda}{f + r} \left[ 1 - \frac{(f + r + \mu)}{\mu} \right] e^{-\mu t} + \frac{\lambda}{\mu (f + r + \mu)} (f + r + \mu) \\
= \frac{\lambda}{\mu} - \frac{\lambda}{\mu} e^{-\mu t} \\
= \frac{\lambda}{\mu} (1 - e^{-\mu t}),
\]

that coincides with limit \( \lim_{z \to 1} \frac{\partial}{\partial z} G(z, t) \), when \( G(z, t) \) is the transient solution of the simple system as shown in equation (2.6).

We will check to identify the dominant term in the convergence of (2.37), namely the term that yields the slowest convergence. It is clear that this term is one with the rate closest to zero.
When looking for such conditions we focus on the parameters of the model that would be more realistically controllable from a practical standpoint. The service rate $\mu$, of the system is usually hard to change in comparison with the reduced service rate $\mu'$, which could be improved in order to get it closer to full service speed. In the same vein, the repair rate $r$ can usually be improved as is constrained by budgets and thus flexible up to a certain extent. The failure rate $f$, on the other hand, is usually an exogenous input of the model and planners have little power on performing modifications to the system that would yield changes in this parameter.

In particular here, and preliminarily, we look for conditions on $\mu'$ that determine an ordering of the convergence rates $s_1$, $s_2$, and $-(f + r)$, in (2.37).

Figure 2.3 shows $s_1$ and $s_2$ as functions of $\mu'$, and $-(f + r)$ for the same system described before with $\mu = 0.4$, $r = 0.04$, $f = 0.2$, while we vary $0 \leq \mu' \leq 0.3$. We notice in this case that there is a cutoff $\mu'$-value for which $s_1(\mu')$ and $-(f + r)$ exchange places as the dominant term.

Figure 2.3: Convergence Dominance for Transient Expectation $E[X(t)]$

Under certain conditions on the parameters $\mu$, $\mu'$, $r$, and $f$ we are able to determine an ordering for $s_1$, $s_2$ and $-(f + r)$. Proposition 1 shows an example of how some specific conditions indeed yield a complete ordering of the convergence rates. A more complete case-study than the one shown in Proposition 1 is warranted, however.
Proposition 1. Assume $\mu > r$. Then statements (S1)-(S3) hold.

(S1) $\mu' \geq \frac{f\mu}{\mu - r}$ implies $s_2 \leq s_1 \leq -(f + r) \leq 0$

(S2) $\mu' \leq \frac{f\mu}{\mu - r}$ implies $s_2 \leq -(f + r) \leq s_1 \leq 0$

(S3) There are no conditions under which $-(f + r) \leq s_2 \leq s_1 \leq 0$.

Proof. We assume henceforth in this proof that $\mu > r$. We start by proving statement (S3), and we do this by contradiction. Assume that $-(f + r) \leq s_2$, which is $-(f + r) \leq \frac{-(f + r + \mu + \mu') - \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')}}{2}$.

Simple algebra yields

$$(f + r) - (\mu + \mu') \geq \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')}, \quad (2.40)$$

which in particular implies that necessarily $(f + r) - (\mu + \mu') \geq 0$.

Squaring both sides of the inequality (2.40) and simplifying we obtain

$$\mu'(\mu - r) \geq f\mu, \quad (2.41)$$

and since we assumed that $\mu > r$, from (2.41) we obtain the necessary condition

$$\mu' \geq \frac{f\mu}{\mu - r}. \quad (2.42)$$

But condition (2.42) leads to a contradiction with the fact that $(f + r) - (\mu + \mu') \geq 0$. Indeed, since $\mu > r$ and both $\mu$ and $r$ are positive quantities (for any system with practical purposes), we can write $\mu = (1 + \alpha)r$ for some $\alpha > 0$. From here, condition (2.42) yields,

$$\mu' \geq \frac{f(1 + \alpha)r}{(1 + \alpha)r - r} = \frac{f(1 + \alpha)r}{\alpha r} = \frac{f(1 + \alpha)}{\alpha} = \frac{f}{\alpha} + f \quad (2.43)$$
At the same time \((f + r) - (\mu + \mu') \geq 0\) implies
\[ f - \alpha r \geq \mu'. \tag{2.44} \]

Then, combining (2.43) and (2.44), which is
\[ f - \alpha r \geq \mu' \geq f, \]
we have that
\[ -\alpha r \geq \frac{f}{\alpha}, \]
which in turn leads to the contradiction
\[ -\alpha^2 r \geq f. \]

Since \(-\alpha^2 r\) is negative it cannot be larger than \(f\) which is a positive number.

Let us now look at statement (S1). First it is clear that \(\mu' \geq f\mu/(\mu - r)\) implies
\(\mu'(\mu - r) \geq f\mu\), working backwards from (2.40) we obtain
\[ (f + r - \mu + \mu')^2 + 4f(\mu - \mu') \leq ((\mu + \mu') - (f + r))^2. \tag{2.45} \]

Since the left-hand side is non-negative we obtain
\[ \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')} \leq |(\mu + \mu') - (f + r)| \]
which implies that either
\[ \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')} \leq (\mu + \mu') - (f + r) \tag{2.46} \]
or
\[ (\mu + \mu') - (f + r) \leq -\sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')} \tag{2.47} \]
Since (2.47) leads to the same contradiction that we showed for statement (S3), then (2.46) has to hold true, and after some reordering we obtain \(s_1 \leq -(f + r)\) which
Finally let us see statement (S2). Assume that \( \mu' \leq f\mu/(\mu - r) \). From here \( \mu'(\mu - r) \leq f\mu \) and working backwards as before we obtain

\[
((\mu + \mu') - (f + r))^2 \leq (f + r - \mu + \mu')^2 + 4f(\mu - \mu').
\]

Taking the square root we have that

\[
|((\mu + \mu') - (f + r)| \leq \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')}.
\]

which is the same as,

\[-\sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')} \leq (\mu + \mu') - (f + r) \leq \sqrt{(f + r - \mu + \mu')^2 + 4f(\mu - \mu')}.
\]

This last proposition can be rearranged to \( s_2 \leq -(f + r) \leq s_1 \), which completes the proof.

\[\square\]

This analysis needs to be completed to determine if there is any order to be found in the convergence rates for the cases \( \mu = r \) and \( \mu < r \). We have the preliminary conjecture that under the case \( \mu < r \) we are not able to give an ordering to \( s_1 \) and \( s_2 \) versus \( -(f + r) \).

Similar results to the ones shown in Proposition 1 are expected when looking for bounds of \( r \) that yield orderings of the convergence rates. We know a priori that this imposes conditions on the relationships between \( \mu' \) and \( f \).

### 2.4.2 Laplace Transform of \( k \)-th Moment

The same procedure that we have used to obtain the first moment allows us to calculate any other order factorial moments. We illustrate with the second factorial
moment which is obtained by taking the second derivative of \( G(z,t) \) with respect to \( z \) and evaluating at \( z = 1 \), i.e.,

\[
\frac{\partial^2}{\partial z^2} G(z,t) \Bigg|_{z=1} = \frac{\partial^2}{\partial z^2} (G_N(z,t) + G_F(z,t)) \Bigg|_{z=1} = (u_{zz}(z,t) + v_{zz}(z,t))\bigg|_{z=1} \quad (2.48)
\]

We start by differentiating (2.30) and (2.31) with respect to \( z \):

\[
\begin{align*}
    u_{tzz} + \mu' u_{zz} + \mu' (1 - z) u_{zzz} &= \lambda u_z + \lambda u_z - \lambda (1 - z) u_{zz} - fu_{zz} + rv_{zz}, \\
v_{tzz} + \mu' v_{zz} + \mu' (1 - z) v_{zzz} &= \lambda v_z + \lambda v_z - \lambda (1 - z) v_{zz} - rv_{zz} + fu_{zz}.
\end{align*}
\]

Taking \( z \) to 1 and gathering terms we obtain

\[
\begin{align*}
    u_{tzz}(1,t) + (2\mu + f)u_{zz}(1,t) - 2\lambda u_z(1,t) - rv_{zz}(1,t) &= 0, \\
v_{tzz}(1,t) + (2\mu' + r)v_{zz}(1,t) - 2\lambda v_z(1,t) - fu_{zz}(1,t) &= 0.
\end{align*}
\]

Using Laplace transform gives

\[
\begin{align*}
    s\hat{u}_{zz}(1,s) - u_{zz}(1,0) + (2\mu + f)\hat{u}_{zz}(1,s) - 2\lambda \hat{u}_z(1,s) - r\hat{v}_{zz}(1,s) &= 0, \\
    s\hat{v}_{zz}(1,s) - v_{zz}(1,0) + (2\mu' + r)\hat{v}_{zz}(1,s) - 2\lambda \hat{v}_z(1,s) - f\hat{u}_{zz}(1,s) &= 0.
\end{align*}
\]

Reordering and using the fact that \( u_{zz}(1,0) = v_{zz}(1,0) = 0 \), we get

\[
\begin{align*}
    (s + 2\mu + f)\hat{u}_{zz}(1,s) - r\hat{v}_{zz}(1,s) &= 2\lambda \hat{u}_z(1,s), \\
    (s + 2\mu' + r)\hat{v}_{zz}(1,s) - f\hat{u}_{zz}(1,s) &= 2\lambda \hat{v}_z(1,s),
\end{align*}
\]

that can be written in matrix form as

\[
\begin{bmatrix}
    s + 2\mu + f & -r \\
    -f & s + 2\mu' + r
\end{bmatrix}
\begin{bmatrix}
    \hat{u}_{zz}(1,s) \\
    \hat{v}_{zz}(1,s)
\end{bmatrix}
= 2\lambda
\begin{bmatrix}
    \hat{u}_z(1,s) \\
    \hat{v}_z(1,s)
\end{bmatrix},
\]

hence giving the solution

\[
\begin{bmatrix}
    \hat{u}_{zz}(1,s) \\
    \hat{v}_{zz}(1,s)
\end{bmatrix}
= \frac{2\lambda}{(s + 2\mu + f)(s + 2\mu' + r) - fr}
\begin{bmatrix}
    s + 2\mu' + r & r \\
    f & s + 2\mu + f
\end{bmatrix}
\begin{bmatrix}
    \hat{u}_z(1,s) \\
    \hat{v}_z(1,s)
\end{bmatrix},
\]

\( (2.49) \)
and
\[
\hat{u}_{zz}(1, s) = \frac{2\lambda}{(s + 2\mu + f)(s + 2\mu' + r)} - fr \left[(s + 2\mu' + r)\hat{u}_z(1, s) + r\hat{v}_z(1, s)\right],
\]
and
\[
\hat{v}_{zz}(1, s) = \frac{2\lambda}{(s + 2\mu + f)(s + 2\mu' + r)} - fr \left[(s + 2\mu + f)\hat{v}_z(1, s)\right].
\]

From here, the second moment is fully defined in terms of the first moment, but more importantly the structure of system (equation (2.49)) allows us to calculate any higher order moment. The detail are shown in Proposition 2.

**Proposition 2.** Denote by \(\hat{u}_z(k)(1, s)\) and \(\hat{v}_z(k)(1, s)\) for \(k \geq 1\), as the Laplace transforms of the \(k\)-th partial derivative with respect to \(z\) of \(u(z, t)\) and \(v(z, t)\) evaluated at \(z = 1\), respectively.

It is clear that for \(k \geq 2\) the Laplace transform of \(k\)-th factorial moment, which is defined as,
\[
E\left[\prod_{i=0}^{k-1}(X(s) - i)^k\right] = E[X(s)(X(s)-1)(X(s)-2)\cdots(X(s)-(k-1))] = \hat{u}_z(k)(1, s) + \hat{v}_z(k)(1, s),
\]
(2.50)

\[
\begin{bmatrix}
\hat{u}_z(k)(1, s) \\
\hat{v}_z(k)(1, s)
\end{bmatrix} = \frac{k\lambda}{(s + k\mu + f)(s + k\mu' + r)} - fr \begin{bmatrix}
s + k\mu' + r & r \\
f & s + k\mu + f
\end{bmatrix} \begin{bmatrix}
\hat{u}_z(k-1)(1, s) \\
\hat{v}_z(k-1)(1, s)
\end{bmatrix}.
\]
(2.51)

Proposition 2 is proved inductively by following the procedure shown for the second moment.

In particular, the Laplace transform of the transient variance can be calculated from the transforms of the first two factorial moments as:
\[
\text{Var}[X(s)] = \hat{u}_z(2)(1, s) + \hat{u}_z(1, s) + \hat{u}_z(1, s) + \hat{v}_z(1, s) - (\hat{u}_z(1, s) + \hat{u}_z(1, s))^2.
\]
2.5 Series Expansion Treatment

Another approach to the problem is to try to solve for the individual probabilities $P_i(t)$ in the frequency domain, which is the same as opening the function $\hat{u}(z, \tau)$ using a series expansion. Let $\hat{C}_i(\tau) = \int_0^\infty e^{-\tau t} P_i(t) dt$. Then, explicitly $\hat{u}(z, \tau) = \sum_{i=0}^\infty \hat{C}_i(\tau) z^i$, that has the derivatives shown in (2.52) and (2.53).

\[
\hat{u}_z(z, \tau) = \sum_{i=1}^\infty \hat{C}_i(\tau) i z^{i-1}, \quad (2.52)
\]

\[
\hat{u}_{zz}(z, \tau) = \sum_{i=2}^\infty \hat{C}_i(\tau) i(i-1) z^{i-2}. \quad (2.53)
\]

Substituting these expansions in equation (2.19), which was:

\[
-\hat{u}_{zz} \mu \mu' (1 - z)^2 + \hat{u}_z (1 - z) \left( \mu \mu' + \lambda (1 - z) (\mu' - \mu) + \mu' f - \mu r + \tau (\mu + \mu') \right) - \hat{u} \left( \lambda^2 (1 - z)^2 + \lambda (f + r + \mu') (1 - z) + \tau^2 + \tau (f - r) \right) = -\tau + \lambda (1 - z) + r, \quad (2.54)
\]

gives

\[
-\sum_{i=2}^\infty \hat{C}_i(\tau) i(i-1) z^{i-2} \mu \mu' (1 - z)^2 + \sum_{i=1}^\infty \hat{C}_i(\tau) i z^{i-1} (1 - z) \left( \mu \mu' + \lambda (1 - z) (\mu' - \mu) + \mu' f - \mu r + \tau (\mu + \mu') \right) - \sum_{i=0}^\infty \hat{C}_i(\tau) z^i \left( \lambda^2 (1 - z)^2 + \lambda (f + r + \mu') (1 - z) + \tau^2 + \tau (f - r) \right) = -\tau + \lambda (1 - z) + r.
\]

Let us denote

\[
A_1(\tau) := \mu \mu' + \lambda (\mu' - \mu) + \mu' f - \mu r + \tau (\mu + \mu'), \quad (2.55)
\]

and

\[
A_2(\tau) := \lambda (f + r + \mu') + \tau^2 + \tau (f - r). \quad (2.56)
\]
Then we have,

\[-\sum_{i=2}^{\infty} \hat{C}_i(\tau)i(i - 1)z^{i-2}\mu'\mu'(1 - z)^2 +\]

\[\sum_{i=1}^{\infty} \hat{C}_i(\tau)iz^{i-1}(1 - z)(A_1(\tau) - z\lambda(\mu' - \mu)) -\]

\[\sum_{i=0}^{\infty} \hat{C}_i(\tau)z^i\left(A_2(\tau) + \lambda^2(1 - z)^2 - z\lambda(f + r + \mu')\right) = -\tau + \lambda(1 - z) + r,\]

where we clean the summations to write

\[-\mu'(1 - 2z + z^2)\sum_{i=2}^{\infty} \hat{C}_i(\tau)i(i - 1)z^{i-2} +\]

\[(1 - z)(A_1(\tau) - z\lambda(\mu' - \mu))\sum_{i=1}^{\infty} \hat{C}_i(\tau)iz^{i-1} -\]

\[\left(A_2(\tau) + \lambda^2(1 - z)^2 - z\lambda(f + r + \mu')\right)\sum_{i=0}^{\infty} \hat{C}_i(\tau)z^i = -\tau + \lambda - \lambda z + r,\]

and expand it into

\[-\mu'\sum_{i=2}^{\infty} \hat{C}_i(\tau)i(i - 1)z^{i-2} + 2\mu'\sum_{i=2}^{\infty} \hat{C}_i(\tau)i(i - 1)z^{i-1} - \mu'\sum_{i=2}^{\infty} \hat{C}_i(\tau)i(i - 1)z^i +\]

\[A_1(\tau)\sum_{i=1}^{\infty} \hat{C}_i(\tau)iz^{i-1} - (A_1(\tau) + \lambda(\mu' - \mu))\sum_{i=1}^{\infty} \hat{C}_i(\tau)iz^i + \lambda(\mu' - \mu)\sum_{i=1}^{\infty} \hat{C}_i(\tau)iz^{i+1} -\]

\[A_2(\tau)\sum_{i=0}^{\infty} \hat{C}_i(\tau)z^i + \lambda(f + r + \mu')\sum_{i=0}^{\infty} \hat{C}_i(\tau)z^{i+1} -\]

\[\lambda^2\sum_{i=0}^{\infty} \hat{C}_i(\tau)z^i + 2\lambda^2\sum_{i=0}^{\infty} \hat{C}_i(\tau)z^{i+1} - \lambda^2\sum_{i=0}^{\infty} \hat{C}_i(\tau)z^{i+2} = -\lambda z + (\lambda - \tau + r).\]

Then, gathering terms with respect to powers of \(z\), we can identify three boundary equations (2.57)-(2.59),

\[-2\mu'\hat{C}_2(\tau) + A_1(\tau)\hat{C}_1(\tau) - (A_2(\tau) + \lambda^2)\hat{C}_0(\tau) = \lambda - \tau + r, \quad (2.57)\]

\[-6\mu'\hat{C}_3(\tau) + 2(2\mu' + A_1(\tau))\hat{C}_2(\tau) - (A_2(\tau) + \lambda^2 + A_1(\tau) + \lambda(\mu' - \mu))\hat{C}_1(\tau) +\]

\[\lambda(f + r + \mu' + 2\lambda)\hat{C}_0(\tau) = -\lambda, \quad (2.58)\]
\[-12 \mu' \lambda C_4(\tau) + 3(4 \mu' + A_1(\tau)) \hat{C}_3(\tau) - (A_2(\tau) + \lambda^2 + 2(A_1(\tau) + \lambda(\mu' - \mu) + \mu')) \hat{C}_2(\tau) + \lambda(f + r + 2\mu' - \mu + 2\lambda) \hat{C}_1(\tau) - \lambda^2 C_0(\tau) = 0, \]  

(2.59)

and a general equation (2.60),

\[-\mu' \lambda C_{i+2}(\tau)(i + 1)(i + 2) + (i + 1)(2i \mu' + A_1(\tau)) \hat{C}_{i+1}(\tau) - (A_2(\tau) + \lambda^2 + i(A_1(\tau) + \lambda(\mu' - \mu) + (i - 1)\mu')) \hat{C}_i(\tau) + \lambda(f + r + i\mu' - (i - 1)\mu + 2\lambda) \hat{C}_{i-1}(\tau) - \lambda^2 \hat{C}_{i-2}(\tau) = 0. \]

(2.60)

As a normalization equation, from (2.28), we have:

\[\hat{u}(1, \tau) = \sum_{i=0}^{\infty} \hat{C}_i(\tau) = P\{\text{system is working normal}\} = \frac{1}{\tau} \cdot \frac{r}{f + r} + \frac{1}{\tau + f + r} \cdot \frac{f}{f + r}. \]

(2.61)

Also, using equations (2.32) and (2.52), and evaluating correspondingly we have:

\[\hat{u}_z(1, \tau) = \sum_{i=1}^{\infty} \hat{C}_i(\tau) i = \frac{\lambda}{(\tau + \mu + f)(\tau + \mu' + r) - fr} \cdot \frac{1}{\tau} \cdot \left[ \frac{(\tau + \mu)(\tau + r)}{\tau + f + r} + r \right]. \]

(2.62)

It is clear that additional equations similar to (2.62) can be obtained by calculating the components of higher moments of $X(t)$ following the procedure shown in section 2.4.2.

These system of equations (2.57)-(2.62) is well defined since it is linear in terms of the transformed probabilities $\hat{C}_j(\tau), j = 0, \ldots, \infty$ and all equations are linearly independent.

Finally, since the stationary solution of this queueing system is known [9], we have, from the Final Value Theorem that:

\[\lim_{\tau \to 0} \tau \hat{C}(\tau) = \lim_{t \to \infty} P_{iN}(t) := P_{iN}. \]  

(2.63)
2.6 Future Work

Even though we have been able so far to obtain many different pieces of the queue transient solution we have still to determine the full time-dependent distribution of the number of customers in the system $X(t)$, and characterize this solution in terms of known distributions.

Continuing from the work shown before we recognize the following areas and leads that require or allow for further study.

2.6.1 Remaining Direct Calculations

One straightforward piece to add to the study is the complete characterization of the rates of convergence of the transient expectation. This corresponds to a generalization of Proposition 1 to include the cases $\mu \leq r$, as well as characterizations of the dominance in terms of other controllable parameters of the system (say the repair rate $r$).

Once this generalization is complete, and in the same vein, we want to see if the dominance results in the convergence of the first moment shed any light on the convergence of higher order moments. For this we will study how the recurrence equations that define the higher order moments change the convergence rates.

Additionally we can invert the Laplace transforms of higher order moments, in particular we can make explicit the variance $\text{Var}[X(t)]$, and hopefully in general for any moment.

2.6.2 Alternative Approaches

Some alternative approaches can be used to attempt the solution of the transient number of customers in the system $X(t)$. 
The first of this approaches we already mentioned in the last section of Series Expansion Treatment 2.5. This involves solving the infinite system of equations (2.57)-(2.62). This, in particular, can be done in one of the following two ways:

1. Solving the recursive equations directly. Although laborious, we could find some success in trying to solve the system recursively by using the recurrence in the general equation (2.60). This, of course, involves the use of series since we have an infinite system of equations.

2. Finding a special structure. We could rearrange our system of equations to a special form with a known solution. For example, if our system was a particular case of a Toeplitz system of equations then we would have the means to solve it directly [19]. Also, we may have some success by separating the system into two or more subsystems that could be solved independently. For example, we can write the system (2.57)-(2.62) with a \( \tau \)-dependent part and a constant part separately as shown in equation (2.64), where \( A(\tau) \) and \( K \) are infinite-dimension matrices, \( K \) is constant, \( \hat{C}(\tau) \) is a column vector of the transformed probabilities \( \hat{C}_i(\tau), i = 0, \ldots, \infty \), and \( b(\tau) \) is the \( \tau \)-dependent right hand side from equations (2.57)-(2.62).

\[
[A(\tau) + K]\hat{C}(\tau) = b(\tau)
\]

System (2.64) is rather sparse (5-diagonal for the most part) and it could be provided with some helpful structure.

A second approach corresponds to work with a transformed version of the defining time-dependent PDE (2.18) in which a particular exponential form has been extracted. This is motivated by the transient solution of the unmodulated case given in equation (2.6) and the transient solution for the modulated case given in equation (2.7). Notice here that the time dependent generating function in (2.6) converges to the generating function of a Poisson distribution.
Since we expect that the transient distribution will have the stochastic decomposition property with one of the components being a Poisson process, we conjecture that extracting a priori an exponential form from the generating function could lead to a simpler PDE to solve. Proposition 3 shows the resulting structure of the defining PDE after extracting a factor of the form \( s(z, t) = \exp\{(z - 1)(1 - e^{-\mu t})(\lambda/\mu)\} \).

**Proposition 3.** Consider the reordering of (2.18) shown in equation (2.65),

\[
- u_{tt} + u_{zt}(\mu + \mu')(1 - z) - u_{zz}\mu\mu'(1 - z)^2 + \\
- u_t(f - r) + \\
\mu u_z(1 - z)\left(-\lambda(1 - z)(1 - \mu'/\mu) + (\mu' + \mu'/\mu f - r)\right) + \\
- \lambda u(1 - z)\left(\lambda(1 - z) + (f + r + \mu')\right) = 0, \tag{2.65}
\]

And define \( s(z, t) = \exp\{(z - 1)(1 - e^{-\mu t})(\lambda/\mu)\} \). The substitution \( u(z, t) = w(z, t)s(z, t) \) in (2.65) allows to extract the exponential form \( s(z, t) \) as a factor resulting in

\[
- w_{tt} + w_{zt}(\mu + \mu')(1 - z) - w_{zz}\mu\mu'(1 - z)^2 + \\
- w_t\left(-\lambda(1 - z)(e^{-\mu t}(1 - \mu'/\mu) + (1 + \mu'/\mu)) + (f - r)\right) + \\
\mu w_z(1 - z)\left(-\lambda(1 - z)(e^{-\mu t}(1 - \mu'/\mu) + (1 + \mu'/\mu)) + (\mu' + \mu'/\mu f - r)\right) + \\
- \lambda w(1 - z)\left(f(1 - \mu'/\mu)(1 - e^{-\mu t}) + 2r\right) = 0. \tag{2.66}
\]
Proof. Elementary calculus shows that

\[ s_t = \lambda e^{-\mu t}(z - 1)s, \]
\[ s_{tt} = \lambda e^{-\mu t}(z - 1)(s_t - \mu s) \]
\[ = \lambda e^{-\mu t}(z - 1)(\lambda e^{-\mu t}(z - 1) - \mu)s, \]
\[ s_z = \frac{\lambda}{\mu}(1 - e^{-\mu t})s, \]
\[ s_{zz} = \frac{\lambda}{\mu}(1 - e^{-\mu t})s_z \]
\[ = \left(\frac{\lambda}{\mu}\right)^2 (1 - e^{-\mu t})^2s, \]
\[ s_{zt} = \lambda e^{-\mu t}(s_z(z - 1) + s) \]
\[ = \lambda ea^{-\mu t}(\lambda(1 - e^{-\mu t})(z - 1) + \mu)s \]
\[ = \lambda (e^{-\mu t}s + (1 - e^{-\mu t})s_t/\mu), \]

and that, for \( u(z, t) = w(z, t)s(z, t), \)

\[ u_t = w_t s + ws_t \]
\[ = (w_t + \lambda \mu e^{-\mu t})s, \]
\[ u_{tt} = w_{tt} s + 2w_t s_t + ws_{tt} \]
\[ = w_{tt} s + 2w_t \lambda(z - 1)e^{-\mu t}s + \lambda e^{-\mu t}(z - 1)(\lambda e^{-\mu t}(z - 1) - \mu)s \]
\[ = (w_{tt} + 2w_t \lambda(z - 1)e^{-\mu t} + \lambda e^{-\mu t}(z - 1)(\lambda e^{-\mu t}(z - 1) - \mu))s, \]
\[ u_z = w_z s + ws_z \]
\[ = (w_z + \frac{\lambda}{\mu}w(1 - e^{-\mu t}))s, \]
\[ u_{zz} = w_{zz} s + 2w_z s_z + ws_{zz} \]
\[ = (w_{zz} + 2w_z \lambda/\mu(1 - e^{-\mu t}) + w\left(\frac{\lambda}{\mu}\right)^2 (1 - e^{-\mu t})^2)s, \]
\[ u_{zt} = w_{zt} s + w_t s_z + w_z s_t + ws_{zt} \]
\[ = (w_{zt} + \frac{\lambda}{\mu}(1 - e^{-\mu t}) + w_z e^{-\mu t}(z - 1) + \lambda\mu e^{-\mu t}(\lambda(1 - e^{-\mu t})(z - 1) + \mu))s. \]
Replacing in \((2.18)\) we get,

\[
-w_{tt} - 2w_t \lambda (z - 1)e^{-\mu t} - w \lambda e^{-\mu t}(z - 1)(\lambda e^{-\mu t}(z - 1) - \mu) \\
+ \left(w_{zt} + \lambda \mu(1 - e^{-\mu t}) + w_2 \lambda e^{-\mu t}(z - 1) + w \lambda \mu e^{-\mu t}(\lambda(1 - e^{-\mu t})(z - 1) + \mu)(\mu + \mu')(1 - z)\right) \\
- (w_t + w \lambda (z - 1)e^{-\mu t})(f - r) \\
- \left(w_{zz} + 2w z \lambda \mu(1 - e^{-\mu t}) + w \lambda^2 (1 - e^{-\mu t})^2 \mu \mu'(1 - z)^2\right) \\
+ (w_z + \lambda \mu w(1 - e^{-\mu t}))(1 - z)(\mu \mu' + \lambda(1 - z)(\mu' - \mu) + \mu' f - \mu r) \\
-w(\lambda^2 (1 - z)^2 + \lambda(f + r + \mu')(1 - z)) = 0,
\]

where we gather terms,

\[
w \left(-\lambda e^{-\mu t}(1 - z)(\lambda e^{-\mu t}(1 - z) + \mu) + \lambda \mu(\mu + \mu')(1 - z)e^{-\mu t}(-\lambda(1 - e^{-\mu t})(1 - z) + \mu) + (f - r)\lambda(1 - z)e^{-\mu t}\right) \\
+ w \left(-\lambda \mu'(1 - z)^2(\lambda \mu')^2(1 - e^{-\mu t})^2 + \lambda \mu(1 - e^{-\mu t})(1 - z)(\mu \mu' + \lambda(1 - z)(\mu' - \mu) + \mu' f - \mu r)\right) \\
+ w \left(-\lambda^2 (1 - z)^2 - \lambda(f + r + \mu')(1 - z)\right) \\
+w_t \left(2\lambda(1 - z)e^{-\mu t} + (\mu + \mu')(1 - z)\lambda \mu(1 - e^{-\mu t}) - (f - r)\right) \\
- w_{tt} \\
+w_t \left(2\lambda(1 - z)e^{-\mu t} + (\mu + \mu')(1 - z)\lambda \mu(1 - e^{-\mu t}) - (f - r)\right) \\
- w_{tt} \\
+w_z \left(-\lambda \mu'(1 - z)^2 + w_{zt}(\mu + \mu')(1 - z) = 0.\right)
\]

Factoring out \(\lambda\) and \((1 - z)\) from the \(w\) term, and \((1 - z)\) from the \(w_z\) term we get,

\[
w(1 - z)\lambda \left(-e^{-\mu t}(\lambda e^{-\mu t}(1 - z) + \mu) + \lambda \mu(\mu + \mu')(1 - z)e^{-\mu t}(-\lambda(1 - e^{-\mu t})(1 - z) + \mu) + (f - r)e^{-\mu t}\right) \\
+ w(1 - z)\lambda \left(-\mu \mu'(1 - z)^2(\lambda \mu')^2(1 - e^{-\mu t})^2 + \lambda (1 - e^{-\mu t})(\mu \mu' + \lambda(1 - z)(\mu' - \mu) + \mu' f - \mu r)\right) \\
+ w(1 - z)\lambda \left(-\lambda(1 - z) - (f + r + \mu')\right) \\
+w_t \left(2\lambda(1 - z)e^{-\mu t} + (\mu + \mu')(1 - z)\lambda \mu(1 - e^{-\mu t}) - (f - r)\right) \\
- w_{tt} \\
+w_z(1 - z)\left(-\lambda \mu'(1 - z)^2 + w_{zt}(\mu + \mu')(1 - z) = 0.\right)
\]

Factoring out \(\lambda\) and \((1 - z)\) from the \(w\) term, and \((1 - z)\) from the \(w_z\) term we get,
which can be simplified to,

\[
\lambda w(1-z)\left(-f(1-\mu'/\mu)(1-e^{-\mu t})-2r\right)+w_t\left(\lambda(1-z)(e^{-\mu t}(1-\mu'/\mu)+(1+\mu'/\mu))-f+r\right)-w_{tt}+
\]

\[
w_z(1-z)\left(\lambda(1-z)(e^{-\mu t}(\mu'-\mu)-(\mu'+\mu)) + \mu\mu' + \mu'f - \mu r\right) - w_{zz}\mu(1-z)^2 + w_{zt}(\mu + \mu')(1-z) = 0,
\]

and finally reordered to,

\[
-w_{tt} + w_{zt}(\mu + \mu')(1-z) - w_{zz}\mu(1-z)^2 -
\]

\[
w_t\left(-\lambda(1-z)(e^{-\mu t}(1-\mu'/\mu) + (1 + \mu'/\mu)) + (f - r)\right) + 
\]

\[
\mu w_z(1-z)\left(-\lambda(1-z)(e^{-\mu t}(1-\mu'/\mu) + (1 + \mu'/\mu)) + (\mu + \mu'/\mu)f - r\right) - 
\]

\[
\lambda w(1-z)\left(f(1-\mu'/\mu)(1-e^{-\mu t}) + 2r\right) = 0.
\]

The resulting equation (2.66), which is also a second-order hyperbolic PDE with variable coefficients, does not appear particularly simpler than the original equation (2.18), but the lower order coefficients are more symmetrical. In this sense, we hope that this transformed PDE could yield a solution that we have not found using the original equation. Here, we could also try to use continued fractions to solve the PDE as shown in [88].

A third approach consists on reducing the original system of two PDEs (2.3)-(2.4) to a single PDE with a similar approach as the one shown in section 2.2, but introducing the variable change shown in equations (2.67)-(2.68).

\[
\tilde{u}(z,t) := \frac{G_N(z,t) + G_F(z,t)}{2},
\]

\[
\tilde{v}(z,t) := \frac{G_N(z,t) - G_F(z,t)}{2}.
\]
Notice that this is just a variable change from what we had before since,

\[ u(z, t) = G_N(z, t) = \tilde{u}(z, t) + \tilde{v}(z, t), \]

\[ v(z, t) = G_F(z, t) = \tilde{u}(z, t) - \tilde{v}(z, t). \]  

(2.69)  

(2.70)

In our current approach we write the PDE in terms of \( u(z, t) \) and try to solve for this function. Once obtained, we would need to calculate \( v(z, t) \) (which is relatively easy), to then and ultimately determine \( G(z, t) = u(z, t) + v(z, t) \) (equation (2.2)). This two-step approach is motivated by the solution methodology used in [9].

On the other hand, the proposed variable change is motivated by the idea of writing the system directly in terms of the quantity we are looking for: \( G(z, t) \). We conjecture that writing the PDE in terms of \( \tilde{u}(z, t) \) instead of \( u(z, t) \) could yield a form more suited for finding an analytical solution. Moreover, we hope that we could even decouple the system (2.8)-(2.9) in its analogous version without, and solve the equations independently instead of having to write a single governing PDE. If this is not possible we would use the method of characteristics again to write the PDE.

A fourth and last fundamentally different approach is solving the system of PDE that comes from the global balance equations (2.3)-(2.4) numerically, solving the second-order hyperbolic PDE (2.18) numerically, or solving the infinite system of equations (2.57)-(2.62) numerically. Any of this approaches would call for specialized and maybe even tailored numerical methodologies for time-dependent probability distributions.

Although the numerical solutions to the transient behavior would be valuable and worth reporting to the academic community, we expect to use such solutions to inform us and find clues on the analytic solution of this problem.
2.6.3 Applications

An important part of this study is to study applications in traffic modeling that can be informed by the transient behavior of these queues.

As it will be shown in Chapter 4, the use of queues to model a traffic system that is subject to random deteriorations of the road capacities can greatly simplify the calibration and simulation efforts incurred by practitioners to determine traffic characteristics such as travel-time reliability metrics. We have already shown the advantage that queueing models have for stationary, long term, and/or averaged-out traffic conditions, and we expect that the results on the transient behavior of queues can also be of great value to aid the traffic management and planning community.

Among other aspects of traffic systems that we want to study are

- Traffic volume reliability and system resiliency to incidents,
- Queue buildup and dissipation times,
- Distributional characterizations of traffic volume changes.
Chapter 3

The Completion Time of an Interrupted Process

In this Chapter we present an analytical model of the travel-time distribution that will be used later in Chapter 4. There, the analytical travel-time distributions are calibrated against data to obtain travel time reliability estimates that consider explicitly the dependencies on failure characteristics of the traffic corridors, such as deterioration rates and repair rates. For the sake of completeness we review in this Chapter some materiel developed by Dr. Zhe Duan that was developed while he was working for his Ph.D. at the Industrial and Systems Engineering Department at Rutgers University, under the supervision of Dr. Melike Baykal-Gürsoy.

Here we deal with a similar system to the one presented in Chapter 2, which is an infinite-server queue with Markovian arrival rates and Markov-modulated service rates. Here, however, we assume the perspective of a server and we try to determine the amount of time elapsed before a customer can be processed. We call this time as completion time. This change in point of view makes certain aspects of the previous system transparent for the purposes of this particular study. For instance, the arrival
rate to the system is irrelevant to determine the completion time of a single customer if service is guaranteed.

We continue, in Section 3.1 by presenting the queueing model in detail and introducing the corresponding notation. In Section 3.2 we motivate the problem and we show the results of Dr. Duan including derivations. From here we go to Section 3.3, in which we developed closed-form expressions for some special cases that are motivated by applications. Finally, in Section 3.7 we discuss how the research can be extended, and what we propose as our future contributions to the study.

3.1 Queueing Model

As in Chapter 2, we consider an unreliable service system with no queuing (infinite servers), in which the system from time to time experiences partial failure where all servers work at a reduced speed. Upon the completion of repair, all servers resume their normal speed. We call the periods that the system works with normal speed as up periods, and the periods that the system works with low speed as down periods. It is so that the service state follows an alternating renewal process of up and down periods. In the case that the service times are generally distributed, we assume that any state change reinitiates the service for current customers with the same service requirement distribution. This last consideration corresponds to the preemption discipline called repeat different or repeat replace, since the service requirement is re-sampled at each service-speed change. This system can be classified as an M/G/∞ queue with two service speeds.

We denote by $F_S(t)$ the cumulative distribution function for service requirement of each customer, with probability density function $f_S(t)$, Laplace transform $L_S(s)$, and mean service requirement $1/\mu$. The service time requirement random variable $S$
is generally distributed.

The up period duration random variable $U$ is exponentially distributed with mean $1/f$. However, the down period duration $D$ is generally distributed with CDF $F_D(t)$, pdf $f_D(t)$, Laplace transform $L_D(s)$, and mean duration $1/r$. Notice that $f$ and $r$ can be interpreted as mean failure and repair rates of the system, respectively.

When technical reasons require its definition we will denote by $Y$ the remaining down time duration after the system has already spent some time in a deteriorated state. It is clear that $Y$, following $D$, is generally distributed.

Finally, without loss of generality the system’s normal service speed is considered to be 1, and when a failure happens, the service speed drops to $\alpha$ with $0 < \alpha < 1$. Notice that since $\alpha > 0$ this kind of system is called partial failure system.

The main objective is to calculate a closed form expression for the distribution of the completion time random variable $T$. Although a general procedure to calculate this distribution in the frequency domain could already be found in the literature in the works by Kulkarni et al. [48, 49], Duan & Baykal-Gürsoy found a novel and much simpler counting argument that could be used for this case. The results and arguments are detailed in Section 3.2.

### 3.2 Review of Completion Time Analysis

We show now in Proposition 4 the resulting completion time distribution in the frequency domain and the accompanying argument in the proof.

**Proposition 4.** Consider a two-service-speed $M/G/\infty$ queue as described in section 3.1. Then, the completion time random variable $T$ has a distribution with Laplace
transform given in equations (3.1)-(3.2).

\[
L_T(s) = E[e^{-sT}]
= \frac{1}{1 - V(s)} \left\{ \begin{array}{c}
\frac{r}{r + f} \left( L_S(s + f) + \frac{r}{s + f} [1 - L_S(s + f)] \left[ L_S(s/\alpha) - \int_0^\infty e^{-s/\alpha t} f_S(t) F_D(t/\alpha) dt \right] \right) + \\
\frac{f}{r + f} \left( L_S(s + f) + \frac{f}{s + f} [1 - L_S(s + f)] \left[ L_S(s/\alpha) - \int_0^\infty e^{-s/\alpha t} f_S(t) F_D(t/\alpha) dt \right] \right) \cdot \\
\left[ L_Y(s) - \int_0^\infty e^{-st} f_Y(t) F_S(\alpha t) dt \right] \right\} + \frac{f}{r + f} \left[ L_S(s/\alpha) - \int_0^\infty e^{-s/\alpha t} f_S(t) F_Y(t/\alpha) dt \right],
\end{array} \right.
\]

(3.1)

where,

\[
V(s) = \frac{f}{s + f} [1 - L_S(s + f)] \cdot \left[ L_D(s) - \int_0^\infty e^{-st} f_D(t) F_S(\alpha t) dt \right].
\]

(3.2)

The proof of Proposition 4 is constructive and is detailed below.

**Proof.** A first key observation is that the completion time of a customer is calculated differently depending on the customer’s arrival epoch. Let us call \(G_1\) as the event that an arrival happens during an up period, and \(G_2\) as the event that an arrival happens during a down period. In particular, by renewal arguments it holds that:

\[
P\{G_1\} = \frac{r}{f + r}, \quad P\{G_2\} = \frac{f}{f + r}.
\]

(3.3)

The Laplace transform of the conditional completion time can be calculated separately for each one of these events. And from these, the Laplace transform of the unconditional travel time distribution can be calculated.

We start by studying the completion time of a customers arrival under \(G_1\).

Consider a customer that arrives during an up period, and denote the subsequent up and down periods as \(U_i\) and \(D_i\), respectively, for \(i = 1, 2, 3, \ldots\). The service requirement on each period is denoted as \(S_i\), \(i = 1, 2, 3, \ldots\). Figure 3.1 shows a
sample path of how the system could evolve. We assume the customer arrives at time zero, and we denote with the crosses on the time-axis some possible departure times (realizations of the completion time).

![Service Speed vs Time Diagram](image)

**Figure 3.1: Sample Path of Service System for a Customer Arrival Under $G^1$**

Define the following events:

- $A_n$, for $n = 0, 1, 2, \ldots$. There are $n$ complete up and down periods, and another incomplete up period in the completion time. The second cross in Figure 3.1 is an example of event $A_2$. Notice that the occurrence of event $A_n$ implies that necessarily:
  
  1. $S_{2i-1} > U_i$, for all $i = 1, 2, \ldots, n$,
  
  2. $\frac{1}{\alpha} S_{2i} > D_i$, for all $i = 1, 2, \ldots, n$,
  
  3. $S_{2n+1} < U_{n+1}$.

- $E_n$, for $n = 0, 1, 2, \ldots$. There are $n + 1$ complete up and $n$ complete down periods, and another incomplete down period in the completion time. The first cross in Figure 3.1 is an example of event $E_1$. Here, necessarily:
  
  1. $S_{2i-1} > U_i$, for all $i = 1, 2, \ldots, n$,
2. $\frac{1}{\alpha}S_{2i} > D_i$, for all $i = 1, 2, \ldots, n$,

3. $S_{2n+1} > U_{n+1}$, for all $i = 1, 2, \ldots, n$,

4. $\frac{1}{\alpha}S_{2n+2} < D_{n+1}$.

Since clearly $S_i$’s, $U_i$’s, and $D_i$’s are all independent and respectively identically distributed,

$$P \{ A_n | G^1 \} = P^n \{ S > U \} \cdot P^n \{ S > \alpha D \} \cdot P \{ S < U \}, \quad (3.4)$$

$$P \{ E_n | G^1 \} = P^{n+1} \{ S > U \} \cdot P^n \{ 1/\alpha S > D \} \cdot P \{ 1/\alpha S < D \}. \quad (3.5)$$

Notice that the probabilities defined in (3.4)-(3.5) were obtained by simple enumeration of the number of system state transitions. Similarly, the conditional completion time $\{ T | G^1 \}$ under each event stated above will be,

$$\{ T | G^1 \} = \begin{cases} 
\sum_{i=1}^n (U_i + D_i) + S_{2n+1}, & \text{event } A_n, \\
\sum_{i=1}^{n+1} U_i + \sum_{i=1}^n D_i + \frac{1}{\alpha}S_{2n+2}, & \text{event } E_n,
\end{cases} \quad (3.6)$$

Now, we use an indicator variable to specify the conditional travel time Laplace transform under some event $H$ as

$$1\{ H \} = \begin{cases} 
1, & \text{event } H, \\
0, & \text{otherwise}
\end{cases}$$

then,

$$E[e^{-sT}1\{ A_n \}|G^1] = E[e^{-sT}|A_n,G^1] \cdot P\{ A_n | G^1 \}$$

$$= (E[e^{-sU}|S > U]P\{ S > U \})^n \left( E[e^{-sD}|S > \alpha D]P\{ S > \alpha D \} \right)^n E[e^{-sS}|S < U]P\{ S < U \}, \quad (3.7)$$

where $s$ is a complex number with positive real part, $U \sim \text{Exp}(f)$, $S$ denotes the generally distributed service time random variable under normal conditions, and $D$
is the generally distributed down period random variable.

It holds that

\[
E\left[e^{-st}\mid S > U\right]P\{S > U\} = \frac{f}{s+f}[1 - LS(s+f)],
\]

\[
E\left[e^{-sd}\mid S > \alpha D\right]P\{S > \alpha D\} = LD(s) - \int_{0}^{\infty} e^{-st} f_D(t) F_S(\alpha t) dt,
\]

\[
E\left[e^{-sS}\mid S < U\right]P\{S < U\} = LS(s+f)
\]

from where (3.7) becomes:

\[
E\left[e^{-sT}1\{A_n\}\mid G^1\right] = \left\{\frac{f}{s+f}[1 - LS(s+f)]\right\}^n \left\{LD(s) - \int_{0}^{\infty} e^{-st} f_D(t) F_S(\alpha t) dt\right\}^n LS(s+f).
\]

(3.8)

For the case \(E_n\), similar to the case \(A_n\), we have:

\[
E\left[e^{-sT}1\{E_n\}\mid G^1\right] = E\left[e^{-sT}\mid E_n, G^1\right] \cdot P\{E_n\mid G^1\}
\]

\[
= (E\left[e^{-su}\mid S > U\right]P\{S > U\})^{n+1} (E\left[e^{-sD}\mid S > \alpha D\right]P\{S > \alpha D\})^n E\left[e^{-s^{\frac{1}{\alpha}}S}\mid S < \alpha D\right]P\{S < \alpha D\},
\]

where the last term is,

\[
E\left[e^{-s^{\frac{1}{\alpha}}S}\mid S < \alpha D\right]P\{S < \alpha D\} = LS\left(\frac{s}{\alpha}\right) - \int_{0}^{\infty} e^{-s^{\frac{1}{\alpha}}t} f_S(t) F_D\left(\frac{t}{\alpha}\right) dt,
\]

giving,

\[
E\left[e^{-sT}1\{E_n\}\mid G^1\right] = \left\{\frac{f}{s+f}[1 - LS(s+f)]\right\}^{n+1} \left\{LD(s) - \int_{0}^{\infty} e^{-st} f_D(t) F_S(\alpha t) dt\right\}^n \cdot \left[LS\left(\frac{s}{\alpha}\right) - \int_{0}^{\infty} e^{-s^{\frac{1}{\alpha}}t} f_S(t) F_D\left(\frac{t}{\alpha}\right) dt\right].
\]

(3.9)

Then, from (3.8) and (3.9), we obtain the Laplace transform of the conditional completion time, given that the customer arrives during an up period as shown in (3.10).

\[
E \left[e^{-sT}\mid G^1\right] = \sum_{n=0}^{\infty} E \left[e^{-sC}\mid A_n, G^1\right] P\{A_n\mid G^1\} + \sum_{n=0}^{\infty} E \left[e^{-sC}\mid E_n, G^1\right] P\{E_n\mid G^1\}
\]

\[
= \frac{1}{1 - V(s)} \cdot \left\{LS(s+f) + \frac{f}{s+f} \cdot [1 - LS(s+f)] \cdot \left[LS\left(\frac{s}{\alpha}\right) - \int_{0}^{\infty} e^{-s^{\frac{1}{\alpha}}t} f_S(t) F_D\left(\frac{t}{\alpha}\right) dt\right]\right\},
\]

(3.10)
where
\[
V(s) = \frac{f}{s + f} \left[1 - L_S(s + f)\right] \cdot \left[L_D(s) - \int_0^\infty e^{-st} f_D(t) F_S(\alpha t) dt\right].
\]

Consider now a customer under \(G^2\), that is who arrives during a down period. Figure 3.2 shows a sample path of a run for the system for a customer arriving during a down period.

![Sample Path of Service System for a Customer Arrival Under \(G^2\)](image)

The probability density function of the remaining down time \(Y\) is,
\[
f_Y(t) = r[1 - F_D(t)], \quad t > 0,
\]
and its Laplace transform is,

\[ L_Y(s) = \frac{r}{s}[1 - L_D(s)]. \]  

(3.13)

We write the Laplace transform of the conditional travel time under each event as,

\[ E[e^{-sT}1|A_0|G^2] = E[e^{-sT}|A_0,G^2] \cdot P\{A_0|G^2\} = L_S(s/\alpha) - \int_0^\infty e^{-\frac{s}{\alpha}t} f_S(t) F_Y(t/\alpha) dt, \]

(3.14)

\[ E[e^{-sT}1|E_0|G^2] = E[e^{-sT}|E_0,G^2] \cdot P\{E_0|G^2\} = L_S(s + f) \cdot \left[ L_Y(s) - \int_0^\infty e^{-st} f_Y(t) F_S(\alpha t) dt \right], \]

(3.15)

\[ E[e^{-sT}1|A_n|G^2] = E[e^{-sT}|A_n,G^2] \cdot P\{A_n|G^2\} \]

\[ = \left[ L_Y(s) - \int_0^\infty e^{-st} f_Y(t) F_S(\alpha t) dt \right] \cdot \left[ L_D(s) - \int_0^\infty e^{-st} f_D(t) F_S(\alpha t) dt \right]^{n-1} \cdot \left[ \frac{f}{s + f} (1 - L_S(s + f)) \right]^n \left[ L_S(s/\alpha) - \int_0^\infty e^{-s/\alpha t} f_S(t) F_D(t/\alpha) dt \right], \]

(3.16)

\[ E[e^{-sT}1|E_n|G^2] = E[e^{-sT}|E_n,G^2] \cdot P\{E_n|G^2\} \]

\[ = \left[ L_Y(s) - \int_0^\infty e^{-st} f_Y(t) F_S(\alpha t) dt \right] \cdot \left[ L_D(s) - \int_0^\infty e^{-st} f_D(t) F_S(\alpha t) dt \right]^n \cdot \left[ \frac{f}{s + f} (1 - L_S(s + f)) \right]^n \cdot L_S(s + f). \]

(3.17)

Using (3.14)-(3.17) we obtain the Laplace transform of the conditional completion times as,

\[ E[e^{-sT}|G^2] = \frac{1}{1-V} \cdot \left[ L_Y(s) - \int_0^\infty e^{-st} f_Y(t) F_S(\alpha t) dt \right] \cdot \left\{ \frac{f}{s + f} [1 - L_S(s + f)] \cdot \left[ L_S(s/\alpha) - \int_0^\infty e^{-s/\alpha t} f_S(t) F_D(t/\alpha) dt \right] + L_S(s + f) \right\} + \left[ L_S(s/\alpha) - \int_0^\infty e^{-s/\alpha t} f_S(t) F_Y(t/\alpha) dt \right]. \]

(3.18)

Finally, by combining the conditional completion times (3.10) and (3.18) using the corresponding probabilities \( P\{G^1\} \) and \( P\{G^2\} \), the unconditional travel time Laplace transform is obtained as shown in (3.1)-(3.2). 

\[ \square \]
Some special cases of interest are detailed in Corollaries 1-3. These are stated without proof since they are direct specializations.

**Corollary 1** (Exponential Down Periods). Consider the queueing system described in Section 3.1, and assume that the down period duration $D$ is exponentially distributed with mean $1/r$. Then, the Laplace transform of the completion time distribution given in (3.1)-(3.2) reduces to (3.19)-(3.20).

$$L_T(s) = E[e^{-sT}]$$

$$= E[e^{-sT}|G^1]P\{G^1\} + E[e^{-sT}|G^2]P\{G^2\}$$

$$= \frac{r}{f + r} \cdot \frac{1}{1 - V(s)} \left( L_S(s + f) + \frac{f}{s + f} \left[ 1 - L_S(s + f) \right] L_S\left( \frac{s + r}{\alpha} \right) \right) +$$

$$\frac{f}{f + r} \cdot \frac{1}{1 - V(s)} \left( L_S\left( \frac{s + r}{\alpha} \right) + \frac{f}{f + r s + f} \left[ 1 - L_S\left( \frac{s + r}{\alpha} \right) \right] L_S(s + f) \right),$$

(3.19)

with,

$$V(s) = \frac{rf \left[ 1 - L_S(s + f) \right] \left[ 1 - L_S\left( \frac{s + r}{\alpha} \right) \right]}{(s + f)(s + r)},$$

(3.20)

and with mean

$$E[T] = \frac{1}{r} \left( 1 - L_S(r/\alpha) \right) \left[ 1 - \frac{r}{r + f} L_S(f) \right] + \frac{1}{f} \left( 1 - L_S(f) \right) \left[ 1 - \frac{f}{r + f} L_S(r/\alpha) \right]$$

$$\frac{L_S(r/\alpha) + L_S(f) - L_S(f)L_S(r/\alpha)}{L_S(r/\alpha) + L_S(f) - L_S(f)L_S(r/\alpha)}.$$ 

(3.21)

**Corollary 2** (Exponential Service Time). Consider the queueing system described in Section 3.1, and assume that the service requirement is exponentially distributed with mean $1/\mu$ and $L_S(s) = \mu/(s + \mu)$. Then, the Laplace transform of the completion time distribution given in (3.1)-(3.2) reduces to (3.22)-(3.23).

$$L_T(s) = E\left[ e^{-sT} \right]$$

$$= \frac{f}{(f + r)} \left\{ \frac{\mu \alpha [s + \mu \alpha - r(1 - L_D(s + \mu \alpha))]}{(s + \mu \alpha)^2} + \frac{1}{1 - V} \left\{ \frac{r}{(f + r)} \frac{\mu [s + \mu \alpha + f(1 + \alpha)(1 - L_D(s + \mu \alpha))]}{(s + f + \mu)(s + \mu \alpha)} + \frac{f}{(f + r)} \frac{fr \mu \alpha [1 - L_D(s + \mu \alpha)]^2}{(s + f + \mu)(s + \mu \alpha)^2} \right\} \right\}$$

(3.22)
where,
\[ V = \frac{f}{s + f + \mu} \cdot L_D(s + \mu \alpha). \] (3.23)

**Corollary 3** (Exponential Down Periods and Service Time). *Consider the queueing system described in Section 3.1, assume that the down period duration \( D \) is exponentially distributed with mean \( 1/r \), and that the service requirement is exponentially distributed with mean \( 1/\mu \) and \( L_S(s) = \mu/(s + \mu) \). Then, the Laplace transform of the completion time distribution given in (3.1)-(3.2) reduces to (3.24).

\[ L_T(s) = E[e^{-sT}] \]
\[ = E[e^{-sT}|G^1]P\{G^1\} + E[e^{-sT}|G^2]P\{G^2\} \]
\[ = \frac{r}{f + r} \cdot \frac{\mu(s + r + f \alpha + \mu \alpha)}{(s + f + \mu)(s + r + \mu \alpha) - fr} + \frac{f}{f + r} \cdot \frac{\mu(s \alpha + f \alpha + r + \mu \alpha)}{(s + f + \mu)(s + r + \mu \alpha) - fr} \]
\[ = \frac{\mu}{f + r} \cdot \frac{r(s + f + r + \mu \alpha) + f\alpha(s + f + r + \mu)}{(s + f + \mu)(s + r + \mu \alpha) - fr} \] (3.24)

There are a couple of additional items to mention on the particular case show in Corollary 3. First, by having both the service time and down periods exponentially distributed makes this an M/M/\( \infty \) queueing system with Markov modulated service speeds that are independently changing for each customer.

Secondly, from (3.24), the completion time \( T \) distribution can be written as shown in equation (3.25),
\[ T = \frac{r}{f + r}T_1 + \frac{f}{f + r}T_2. \] (3.25)

The density functions for \( T_1 \) and \( T_2 \) can be obtained in closed form by inverting the Laplace transforms as shown in (3.26)-(3.27),
\[ f_{T_1}(t) = \frac{\mu(s_1 + f \alpha + r + \mu \alpha)}{s_1 - s_2} e^{s_1 t} - \frac{\mu(s_2 + f \alpha + r + \mu \alpha)}{s_1 - s_2} e^{s_2 t}, \] (3.26)
\[ f_{T_2}(t) = \frac{\mu(s_1 \alpha + f \alpha + r + \mu \alpha)}{s_1 - s_2} e^{s_1 t} - \frac{\mu(s_2 \alpha + f \alpha + r + \mu \alpha)}{s_1 - s_2} e^{s_2 t}, \] (3.27)
where \( s_1 \) and \( s_2 \) are the roots of
\[ (s + f + \mu)(s + r + \mu \alpha) - fr = 0. \] (3.28)
Equation (3.28) is the same as equation (2.36) from Chapter 2, Section 2.4, and we have already shown there that these two roots are both real and negative, and that \( s_1 > s_2 \).

The expected completion time is then derived from the generating function given in (3.24) as

\[
E[T] = \frac{1}{\mu} + \frac{f(1 - \alpha)}{f + r} \cdot \frac{f + r + \mu}{\mu(f\alpha + r + \mu\alpha)},
\]

(3.29)

Moreover, by the use of Little’s theorem, we can obtain the expected number of customers in the system as

\[
E[N] = \frac{\lambda}{\mu} + \frac{\lambda f(1 - \alpha)}{f + r} \cdot \frac{f + r + \mu}{\mu(f\alpha + r + \mu\alpha)}.
\]

(3.30)

The result coincides with Baykal-Gürsoy and Xiao [9], of the M/MSP/\( \infty \) queue.

It is worth mentioning that the generating function of the completion time for the complete breakdown case (which can be obtained by setting \( \alpha = 0 \)) coincides with Gaver’s preemptive-repeat-different interruption case [31]. In fact, since in this case all services start at up periods, we only need to look at our result for the group of customers \( G^1 \).

Figure 3.3 shows an example of the time domain completion time distribution for exponential service time given in equations (3.26)-(3.27). Here we consider \( 1/f = 5 < 40 = 1/r \) time units, which corresponds to a system in which failures occur more frequently than repairs. The mean service requirement is \( 1/\mu = 40 \), and we vary \( \alpha \) to 0.2, 0.4, 0.6, and 0.8.

It is clear that as \( \alpha \) decreases the mass of the distribution is increasingly shifted towards the right-tail. In particular, Table 3.1 shows the expected completion times for different values of \( \alpha \) calculated using equation (3.30), in time units. The expected completion time increases as the server capacity drops to a smaller proportion \( \alpha \).
Figure 3.3: Completion Time Distribution under Exponential Service Requirement

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E[T]$</td>
<td>183.57</td>
<td>127.50</td>
<td>102.85</td>
<td>88.96</td>
</tr>
</tbody>
</table>

Table 3.1: Expected Completion Time Exponential Service Time
Renewal arguments provide the probability of system being in up period in steady state as \( r/(r + f) \), and the probability of encountering the system in down period as \( f/(r + f) \). Then, since the mean completion time in an up period is the mean service time requirement, \( 1/\mu \), and correspondingly \( 1/(\alpha \mu) \) for down periods, the mean completion time in the steady state can be written as the mixture:

\[
SSMM := \frac{r}{r + f/\mu} + \frac{f}{r + f/\alpha \mu} \tag{3.31}
\]

where SSMM stands for Steady-State Mixture Mean.

Proposition 5 gives the conditions under which we can determine if the steady state mixture under or over-estimates the expected value given in (3.21).

**Proposition 5.** Call \( C = \frac{\alpha \mu}{2\alpha \mu + \alpha r + f} \) and \( \mathcal{L} = L_S(f) L_S(r/\alpha) - L_S(r/\alpha) - L_S(f) \) with \( \mathcal{L} \neq 0 \). Then, equivalence (3.32) holds.

\[
E[T] \leq SSMM \iff \mathcal{L} \leq \frac{C}{rf} \left( r^2 (L_S(f) - 1) + f^2 (L_S(r/\alpha) - 1) - 2rf \right) \leq 0. \tag{3.32}
\]

**Proof.** Call \( C := \frac{\alpha \mu}{2\alpha \mu + \alpha r + f} \) and \( \mathcal{L} := L_S(f) L_S(r/\alpha) - L_S(r/\alpha) - L_S(f) \). Considering the mean completion time in equation (3.21) we can calculate the difference with the steady-state mixture mean (SSMM) to be as shown in (3.33),

\[
E[T] - SSMM = \frac{\alpha \mu}{rf} \left( \frac{r^2 (1 - L_S(f)) + 2rf + f^2 (1 - L_S(r/\alpha))) + (2\alpha \mu + \alpha r + f) \mathcal{L}}{-(r + f)\alpha \mu} \right). \tag{3.33}
\]

Now, we know that \( \mathcal{L} \) is non-positive. In fact, since \( L_S(f) \) is bounded in \([0, 1]\), and \( L_S(r/\alpha) \) is in particular non-negative we have that

\[ L_S(f) L_S(r/\alpha) \leq L_S(r/\alpha), \]

from where is clear that

\[ L_S(f) L_S(r/\alpha) - L_S(r/\alpha) - L_S(f) = \mathcal{L} \leq -L_S(f) \leq 0. \]
Then, we have that the difference in the left-hand side of (3.33) is non-positive if and only if the numerator in the right-hand side is non-positive. The latter condition translates to (3.34),

$$L_S(f) \left( C_r f + 1 \right) + L_S(r/\alpha) \left( C_{r/\alpha}^f + 1 \right) \geq C \frac{(r + f)^2}{rf} + L_S(f)L_S(r/\alpha).$$

which can be shown to be equivalent to,

$$\mathcal{L} \leq C \frac{r^2(L_S(f) - 1) + f^2(L_S(r/\alpha) - 1) - 2rf}{rf}.\]

3.3 Non Exponential Service Time Distributions

In this section we again assume exponential down times, and we show the explicit Laplace transforms of the completion time distributions for a number of service time requirement distributions $S$ (much like in Corollary 3). Some choices are of analytical interest while others are motivated by applications of the completion time model to travel time reliability that will be investigated in Chapter 4.

Notice that equations (3.19)-(3.20) from Corollary 1 show the expression for the completion time distribution when the service time requirement distribution has still not been decided.

3.3.1 Deterministic Service Time

If the service time variable is not random, it can be assumed to follow a deterministic distribution as shown in Figure 3.4.

Corollary 5 shows the resulting completion time distribution for the choice of uniform service time requirement. This is again a corollary of Proposition 4.
Corollary 4 (Deterministic Service Time). Consider the queueing system described in Section 3.1, and assume that the down period duration \( D \) is exponentially distributed with mean \( 1/r \). Let the service requirement be deterministic \( S = \mu \). Then the Laplace transform of the completion time distribution is given as shown in equation (3.35).

\[
L_T(s) = \frac{r}{(f + r)} \left( e^{-\mu(s+f)}(s + f) \frac{[(s + r + f) - f e^{-\mu/\alpha(s+r)}]}{[(s + f)(s + r) - rf [1 - e^{-\mu(s+f)}] [1 - e^{-\mu/\alpha(s+r)}]]} \right) + \frac{f}{(f + r)} \left( e^{-\mu/\alpha(s+r)}(s + r) \frac{[(s + r + f) - re^{-\mu(s+f)}]}{[(s + f)(s + r) - rf [1 - e^{-\mu(s+f)}] [1 - e^{-\mu/\alpha(s+r)}]]} \right). \tag{3.35}
\]

Proof. We write \( f_S(t) \) in terms of the Dirac Delta function as shown in equation (3.36).

\[
f_S(t) = \delta(t - \mu), \tag{3.36}
\]

with Laplace transform

\[
L_S(s) = e^{-\mu s}. \tag{3.37}
\]

\[
L_S(s + f) = e^{-\mu(s+f)} \tag{3.38}
\]

\[
L_S\left(\frac{s + r}{\alpha}\right) = e^{-\mu/\alpha(s+r)} \tag{3.39}
\]
From here we obtain $V(s)$ from equation (3.20) as

$$V(s) = \frac{rf [1 - e^{-\mu(s+f)}] [1 - e^{-\mu/\alpha(s+r)}]}{(s+f)(s+r)},$$

(3.40)

from where,

$$\frac{1}{(1 - V(s))} = \frac{(s + f)(s + r)}{[(s + f)(s + r) - rf [1 - e^{-\mu(s+f)}] [1 - e^{-\mu/\alpha(s+r)}]].}$$

(3.41)

Call from here on,

$$L_1(s) = rL_S(s+f)+ \frac{fr}{s+f} [1 - L_S(s + f)] L_S \left( \frac{s + r}{\alpha} \right) + fL_S \left( \frac{s + r}{\alpha} \right) + \frac{fr}{s+r} \left[ 1 - L_S \left( \frac{s + r}{\alpha} \right) \right] L_S(s+f)$$

(3.42)

In the deterministic case, this becomes,

$$L_1(s) = \frac{e^{-\mu(s+f)}(s + f) [(s + r + f) - fe^{-\mu/\alpha(s+r)}]}{(s + r)(s + f)} + \frac{e^{-\mu/\alpha(s+r)}(s + r) [(s + r + f) - re^{-\mu(s+f)}]}{(s + r)(s + f)}$$

(3.43)

Finally, replacing (3.41) and (3.43) in (3.19), gives (3.35), which ends the proof.

\[\square\]

### 3.3.2 Uniform Service Time

In the somewhat extreme case in which no information is known about the regular service time requirement distribution except for a range of customer requirements, then the use of a uniform distribution for the service requirement can be warranted. Let $[a, b]$ be the service time requirement range, and assume that $S$ is a uniform random variable in the interval $[a, b]$, $0 < a < b$, that is $S \sim \mathcal{U}(a, b)$. Figure 3.5 shows a diagram of such a distribution. Here the mean service time is $\frac{1}{\mu} = \frac{(a+b)}{2}$.

Corollary 5 shows the resulting completion time distribution for the choice of uniform service time requirement. This is again a corollary of Proposition 4.

**Corollary 5** (Uniform Service Time). Consider the queueing system described in Section 3.1, and assume that the down period duration $D$ is exponentially distributed
with mean $1/r$. Let the service requirement be uniformly distributed with $S \sim \mathcal{U}(a, b)$. If we denote $A = b - a$, $h_1(s) = e^{-a(s+f)} - e^{-b(s+f)}$, and $h_2(s) = e^{-a/(s+r)} - e^{-b/(s+r)}$, then the Laplace transform of the completion time distribution is given as shown in equation (3.44).

\[
L_T(s) = \frac{r}{(f + r)} \left[ (s + f)h_1(s) [A(s + r)(s + r + f) - \alpha fh_2(s)] \right] + \frac{f}{(f + r)} \left[ \alpha(s + r)h_2(s) [A(s + f)(s + f + r) - rh_1(s)] \right].
\] (3.44)

The proof is constructive but elementary, and is shown below.

**Proof.** First we write $f_S(t)$ in terms of unit step functions $u(t)$ as shown in equation (3.45).

\[
f_S(t) = \begin{cases} 
\frac{1}{b - a} , & t \in [a, b] \\
0 , & \text{otherwise}
\end{cases} = \frac{1}{b - a} [u(t - a) - u(t - b)] 
\] (3.45)

Clearly, the Laplace transform of the service time requirement distribution is

\[
L_S(s) = \frac{1}{b - a} \left[ \frac{1}{s} e^{-as} - \frac{1}{s} e^{-bs} \right] = \frac{e^{-as} - e^{-bs}}{(b - a)s}. 
\] (3.46)
Then we can derive $V(s)$ from equation (3.20) as

$$V(s) = \frac{rf [1 - L_S(s + f)] [1 - L_S(\frac{s+r}{\alpha})]}{(s + f)(s + r)}$$

$$= \frac{rf [A(s + f) - h_1(s)] [A(s + r) - \alpha h_2(s)]}{A^2(s + f)^2(s + r)^2}.$$ 

It is clear that,

$$\frac{1}{1 - V(s)} = \frac{A^2(s + f)^2(s + r)^2}{[A^2(s + f)^2(s + r)^2 - rf [A(s + f) - h_1(s)] [A(s + r) - \alpha h_2(s)]]}. \quad (3.47)$$

Let,

$$L_1(s) = rL_S(s + f) + \frac{fr}{s + f} [1 - L_S(s + f)] L_S \left(\frac{s + r}{\alpha}\right) + fL_S \left(\frac{s + r}{\alpha}\right)$$

$$+ \left[1 - L_S \left(\frac{s + r}{\alpha}\right)\right] L_S(s + f).$$

Simple algebra leads to,

$$L_1(s) = r\frac{h_1(s)}{A(s + f)} + \frac{fr}{s + f} \left[1 - \frac{h_1(s)}{A(s + f)}\right] \frac{h_2(s)}{A(s + r)} + f\left[1 - \frac{\alpha}{A(s + r)}\right] \frac{h_2(s)}{A(s + f)}$$

$$= \frac{r(s + f)h_1(s) [A(s + r)(s + r + f) - \alpha h_2(s)]}{A^2(s + f)^2(s + r)^2} + \frac{\alpha f(s + r)h_2(s) [A(s + f)(s + f + r) - rh_1(s)]}{A^2(s + f)^2(s + r)^2}. \quad (3.48)$$

Finally, replacing (3.47) and (3.48) in (3.19), gives (3.44), which ends the proof.

Figure 3.6 shows examples of the full completion time distribution for uniform service requirement $S \sim U(30, 90)$, with the rest of the parameters being the same as in the previous example ($1/f = 5$ and $1/r = 40$). The service speed drop $\alpha$ varies again from 0.2 to 0.8. The Laplace transform is inverted numerically by using a quotient-difference algorithm developed by De Hoog et al. [20] in a version implemented with accelerated convergence for the continued fraction expansion by Hollenbeck [37].

It is noteworthy that for values of $\alpha$ closer to $\alpha = 1$ the resulting completion time distribution is similar to the service time distribution but, part of the mass is transferred to the right-tail. As with the case for exponential service requirement, as $\alpha$ decreases the distribution becomes flatter, and the probabilities of encountering large completion times increase.
Figure 3.6: Completion Time Distribution under Uniform Service Requirement

### 3.3.3 Triangular Service Time

We switch our attention now to a triangular service time requirement distribution $S$, which is appropriate for unimodal service requirements. Consider a triangular distribution on the interval $[a, b]$ with $a < b$, and vertex at $c$, $a < c < b$. Here $a$ and $b$ denote the minimum and maximum service time requirement, respectively. Figure 3.7 shows a diagram of a triangular distribution. Notice that $f_S(c) = 2/(b - a)$ holds, and that the mean service requirement is $\mu = \frac{1}{3}(a + b + c)$.

In Corollary 6 we find the explicit completion time distribution for triangular service time requirement.

**Corollary 6 (Triangular Service Time).** Consider the queueing system described in Section 3.1, and assume that the down period duration $D$ is exponentially distributed with mean $1/r$. Let the service requirement have a triangular distribution over the interval $[a, b]$, with vertex $c$ such that $a < c < b$. Denote $A = b - a$, $B_1 = c - a$, $B_2 = b - c$, $D = AB_1B_2$, and the functions $g_1(s) = B_2e^{-a(s+f)} - Ae^{-c(s+f)} + B_1e^{-b(s+f)}$, ...
and \( g_2(s) = B_2e^{-a/(s+r)} - Ae^{-c/(s+r)} + B_1e^{-b/(s+r)} \). Then, the Laplace transform of the completion time distribution is given in (3.49).

\[
L_T(s) = \frac{r}{(f+r)} \frac{2(s+f)g_1(s)[D(s+r)^2(s+f+r) - 2\alpha^2 fg_2(s)]}{D^2(s+f)^3(s+r)^3 - rf[D(s+f)^2 - 2g_1(s)]D(s+r)^2 - 2\alpha^2 g_2(s)]} + \frac{f}{(f+r)} \frac{2\alpha^2(s+r)g_2(s)[D(s+f)^2(s+f+r) - 2rg_1(s)]}{D^2(s+f)^3(s+r)^3 - rf[D(s+f)^2 - 2g_1(s)]D(s+r)^2 - 2\alpha^2 g_2(s)]}.
\]

(3.49)

Proof. We can use unitary step functions to describe \( f_S(t) \) as shown in (3.50).

\[
f_S(t) = \frac{2}{b-a} \frac{(t-a)u(t-a)}{c-a} - \frac{2}{b-a} \frac{(t-c)u(t-c)}{b-c} + \frac{2}{b-a} \frac{(t-b)u(t-b)}{b-c} = \frac{2}{AB_1B_2} \left( B_2(t-a)u(t-a) - A(t-c)u(t-c) + B_1(t-b)u(t-b) \right)
\]

(3.50)

with Laplace transform given as

\[
L_S(s) = \frac{2}{AB_1B_2} \frac{B_2e^{-as} - Ae^{-cs} + B_1e^{-bs}}{s^2}.
\]

(3.51)

We calculate \( V(s) \) from equation (3.20),

\[
V(s) = \frac{rf[D(s+f)^2 - 2g_1(s)]D(s+r)^2 - 2\alpha^2 g_2(s)]}{D^2(s+f)^3(s+r)^3},
\]

and

\[
\frac{1}{1-V(s)} = \frac{D^2(s+f)^3(s+r)^3}{D^2(s+f)^3(s+r)^3 - rf[D(s+f)^2 - 2g_1(s)]D(s+r)^2 - 2\alpha^2 g_2(s)]}.
\]

(3.52)
Define now $L_1(s)$ as in Corollary 5. For this case we have:

$$L_1(s) = \frac{2r(s+f)g_1(s)}{D^2(s+f)^3(s+r)^3} \left[ D(s+r)^2(s+f+r) - 2\alpha^2 sg_2(s) \right] + \frac{2\alpha^2 f(s+r)g_2(s)}{D^2(s+f)^3(s+r)^3} \left[ D(s+f)^2(s+f+r) - 2rg_1(s) \right].$$

Replacing (3.52) and (3.53) in (3.19), gives (3.49).

Once again we show an example of the resulting completion time distribution for a system with $1/f = 5$, $1/r = 40$, but now a triangular service requirement distribution on the interval $[a, b] = [30, 90]$ with vertex in $c = 60$. The $\alpha$ parameter varies from 0.2 to 0.8 and the Laplace transform in (3.49) is inverted numerically. Figure 3.8 shows the resulting completion time distributions.

![Figure 3.8: Completion Time Distribution under Triangular Service Requirement](image)

Similarly as in the uniform case, for $\alpha$ close to 1 the distribution resembles the service time distribution but with a heavy right-tail, and as $\alpha$ drops the mass of the distribution is concentrated in the tail.
The distributions obtained in Corollaries 5 and 6 have been tested in vehicular traffic travel time applications, in which the service time requirement is interpreted as a speed requirement, and the completion time is the full travel time distribution when considering random deterioration of the road conditions. These Laplace transforms can be inverted numerically to obtain travel time distributions in the time domain.

\subsection{Degenerate Triangle: $c = a$}

Consider a triangular service time distribution where $c = a$, that is, a rectangular triangle on the left side. This is a limit case from the distribution presented in the previous Section. Figure 3.9 shows a diagram of such distribution. Notice that $f_S(a) = 2/(b - a)$ to have total mass of one.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure3.9.png}
\caption{Degenerate Triangular Distribution, $c = a$}
\end{figure}

Corollary 7 shows the explicit completion time distribution for triangular service time requirement with $c = a$. The expression is the same as in Equation (3.49), with only changes in the notation.

\textbf{Corollary 7 (Triangular Service Time $c = a$).} Consider the queueing system described in Section 3.1, and assume that the down period duration $D$ is exponentially distributed with mean $1/r$. Let the service requirement have a triangular distribution over the interval $[a, b]$, with vertex $c$ such that $c = a$. Denote $A = b - a$
and \( D = A^2 \), and the functions \( g_1(s) = (A(s + f) - 1)e^{-a(s+f)} + e^{-b(s+f)} \), and \( g_2(s) = (A/\alpha(s + r) - 1)e^{-a/\alpha(s+r)} + e^{-b/\alpha(s+r)} \). Then, the Laplace transform of the completion time distribution is given in (3.49).

**Proof.** We use unitary step functions to describe \( f_S(t) \).

\[
f_S(t) = \frac{2}{(b-a)^2} [-(t-a)u(t-a) + (b-a)u(t-a) + (t-b)u(t-b)]
\]

Call \( A = b-a \), and \( D = A^2 \), then

\[
f_S(t) = \frac{2}{D} [-(t-a)u(t-a) + Au(t-a) + (t-b)u(t-b)]
\]

Then, directly we obtain

\[
L_S(s) = \frac{2}{D} \left[ -\frac{e^{-as}}{s^2} + A\frac{se^{-as}}{s^2} + \frac{e^{-bs}}{s^2} \right] = \frac{2}{D} \frac{(As-1)e^{-as} + e^{-bs}}{s^2}.
\]  

(3.54)

From here,

\[
L_S(s + f) = \frac{2g_1(s)}{D(s + f)^2},
\]  

(3.55)

and,

\[
L_S \left( \frac{s + r}{\alpha} \right) = \frac{2\alpha^2g_2(s)}{D(s + r)^2}.
\]  

(3.56)

Then, following analogously as in the general triangle case the final Laplace transform is given by (3.49). Notice that for \( a = 0 \) we have \( A = b \), \( D = b^2 \), \( g_1(s) = b(s + f) - 1 + e^{-b(s+f)} \), and \( g_2(s) = b/\alpha(s + r) - 1 + e^{-b/\alpha(s+r)} \).

\[
3.3.5 \text{ Degenerate Triangle: } c = b
\]

The case when \( a < b = c \), is analogous to the \( a = c \) case. In this case the distribution has a shape as shown in Figure 3.10.
Corollary 8 shows the explicit completion time distribution for triangular service time requirement with \( c = b \). The expression is the same as in Equation (3.49) with only changes in the notation, and the proof is analogous to the case \( c = a \) so it is omitted.

**Corollary 8** (Triangular Service Time \( c = b \)). Consider the queueing system described in Section 3.1, and assume that the down period duration \( D \) is exponentially distributed with mean \( 1/r \). Let the service requirement have a triangular distribution over the interval \([a, b]\), with vertex \( c \) such that \( c = b \). Denote \( A = b - a \) and \( D = A^2 \), and the functions \( g_1(s) = e^{-a(s+f)} - (A(s + f) + 1)e^{-b(s+f)} \), and \( g_2(s) = e^{-a/\alpha(s+r)} - (A/\alpha(s + r) + 1)e^{-b/\alpha(s+r)} \). Then, the Laplace transform of the completion time distribution is given in (3.49).

### 3.3.6 Shifted Exponential Service Time

A final case again motivated by applications in travel time reliability is the case of shifted-exponential service time requirement distribution. The shifted exponential density function is defined in equation (4.5), where \( \eta \) is the rate parameter, \( c \) is the right-shift applied to a regular exponential distribution, and \( u(\cdot) \) is the heaviside step function. Figure 4.3 shows a diagram of such distribution. The mean of the shifted exponential distribution is the shift \( c \) plus the reciprocal of \( \eta \). If the mean time requirement is already known to be \( 1/\mu \), then we would require that \( c + 1/\eta = 1/\mu \),
which forces \( \eta = \mu/(1 - \mu c) \) adding the condition \( 1 - \mu c > 0 \), or \( c < 1/\mu \).

\[
f_S(t) = \eta u(t - c)e^{-\eta(t-c)}
\]  

(3.57)

Corollary 9 gives the expression for the Laplace transform of the completion time distribution for the shifted exponential service time requirement case.

**Corollary 9 (Shifted Exponential Service Time).** Consider the queueing system described in Section 3.1, and assume that the down period duration \( D \) is exponentially distributed with mean \( 1/r \). Let the service requirement follow a shifted exponential distribution with rate parameter \( \eta = \mu/(1 - \mu c) \), and right-shift \( c \). Then the Laplace transform of the completion time distribution is given as shown in equation (3.58).

\[
L_T(s) = \frac{r}{(f + r)} \frac{\mu e^{-c(s+f)}(s + f)[(s + r + f)(s + r + \alpha u) - \alpha f\mu e^{-\eta(s+r)}]}{[(s + f)(s + r + \alpha u)(s + f + \mu) - rfW(s)]} + \frac{f}{(f + r)} \frac{\alpha \mu e^{-\eta(s+r)}(s + r)[(s + r + f)(s + f + \mu) - r\mu e^{-c(s+f)}]}{[(s + f)(s + r + \alpha u)(s + f + \mu) - rfW(s)]}
\]  

(3.58)

for

\[
W(s) := [s + f + \mu(1 - e^{-c(s+f)})] [s + r + \alpha \mu(1 - e^{-\eta(s+r)})].
\]

**Proof.** The service time density function is,

\[
f_S(t) = \mu e^{-\mu(t-c)}u(t-c),
\]
with laplace transform,

\[ L_S(s) = \mu e^{-cs}/s + \mu, \]

From here

\[ L_S(s + f) = \mu e^{-c(s+f)}/s + f + \mu, \]

and,

\[ L_S \left( \frac{s + r}{\alpha} \right) = \alpha \mu e^{-c/\alpha(s+r)}/s + r + \alpha \mu. \]

Then,

\[ V(s) = \frac{rf [s + f + \mu(1 - e^{-c(s+f)})] [s + r + \alpha \mu(1 - e^{-c/\alpha(s+r)})]}{(s + f)(s + r)(s + r + \alpha \mu)(s + f + \mu)}. \]

Call \( W(s) := [s + f + \mu(1 - e^{-c(s+f)})] [s + r + \alpha \mu(1 - e^{-c/\alpha(s+r)})]. \) From here we obtain that,

\[ \frac{1}{(1 - V(s))} = \frac{(s + f)(s + r)(s + r + \alpha \mu)(s + f + \mu)}{(s + f)(s + r)(s + r + \alpha \mu)(s + f + \mu) - rfW(s)}. \]

Simple calculations yield,

\[ L_1(s) = \frac{r \mu e^{-c(s+f)}(s + f) [(s + r + f)(s + r + \alpha \mu) - \alpha f \mu e^{-c/\alpha(s+r)}]}{(s + f)(s + f + \mu)(s + r)(s + r + \alpha \mu)} + \frac{\alpha f \mu e^{-c/\alpha(s+r)}(s + r) [(s + r + f)(s + f + \mu) - r \mu e^{-c(s+f)}]}{(s + r + \alpha \mu)(s + f)(s + f + \mu)(s + r)}. \]

Replacing in (3.19) gives the result.

\[ \square \]

### 3.4 On the Preempt-\textit{RESTART} Service Discipline

We study here a particular case that yields some interesting closed form results in the Laplace domain for general up period and down period random variables. We make the assumption that the service requirement is sampled once, and every time a state-change occurs the same service requirement is repeated. Then, we have a system under the Preempt-\textit{RESTART} Service Discipline.
Since the service requirement is always the same, say $S = t$, the completion time given in equations (3.6) and (3.11) can be rewritten for an up period arrival and down period arrival as shown in Equations (3.59) and (3.60), respectively.

$$\{T|G^1, S = t\} = \begin{cases} 
\sum_{i=1}^{n} (U_i + D_i) + t, & \text{event } A_n, \\
\sum_{i=1}^{n+1} U_i + \sum_{i=1}^{n} D_i + \frac{t}{\alpha}, & \text{event } E_n,
\end{cases} \quad (3.59)$$

$$\{T|G^2, S = t\} = \begin{cases} 
\sum_{i=1}^{n} D_i + \sum_{i=1}^{n} U_i + \frac{t}{\alpha}, & \text{event } A_n, \\
\sum_{i=1}^{n+1} D_i + \sum_{i=1}^{n} U_i + t, & \text{event } E_n,
\end{cases} \quad (3.60)$$

Theorems 1–2 generalize the results in [28, 83] for the complete breakdown case to our partial breakdown case under the RESTART preemption strategy. Theorem 1 shows the Laplace transform of the completion time conditioned on the task size, and Theorem 2 shows the expected completion time conditioned on the task size. The results are stated for general up and down durations with CDFs $F_U(t)$ and $F_D(t)$, respectively, and Laplace transforms $L_U(s)$ and $L_D(s)$, respectively.

**Theorem 1.** The conditional Laplace transform of the completion time distribution for a task of duration $t$ is

$$L_T(s|t) \equiv E[e^{-sT}|S \approx t] = \frac{r e^{-(s+f)t} \left[ 1 + \frac{f}{s+r} (1 - e^{-(s+r)t/\alpha}) \right] + f e^{-(s+r)t/\alpha} \left[ 1 + \frac{r}{s+f} (1 - e^{-(s+f)t}) \right]}{(r+f) \left[ 1 - \frac{r f}{(s+r)(s+f)} (1 - e^{-(s+f)t})(1 - e^{-(s+r)t/\alpha}) \right]].$$

**Theorem 2.** The expected completion time for a task of length $t$ is

$$E[T|S \approx t] = \frac{\left( te^{-ft} + (1 - e^{-ft})/f \right)(r + f(1 - e^{-rt/\alpha}))}{(r+f)(1 - (1 - e^{-ft})(1 - e^{-rt/\alpha}))} + \frac{\left( \frac{t}{\alpha} e^{-rt/\alpha} + (1 - e^{-rt/\alpha})/r \right) (f + r(1 - e^{-ft}))}{(r+f)(1 - (1 - e^{-ft})(1 - e^{-rt/\alpha}))} \quad (3.62)$$

**Proof.** (Theorem 1)

We aim to obtain the conditional Laplace transform of the completion time distribution for a service time requirement of size $t$, so we set $S = t$. From here, the
completion time for customers arriving during an up period given in Equation (3.6) now becomes:

\[
\{T | G^1, S = t\} = \begin{cases} 
\sum_{i=1}^{n} (U_i + D_i) + t, & \text{event } A_n, \\
\sum_{i=1}^{n+1} U_i + \sum_{i=1}^{n} D_i + \frac{t}{\alpha}, & \text{event } E_n,
\end{cases} \quad n = 0, 1, 2, \ldots (3.63)
\]

with probabilities,

\[
P\{A_n | G^1, S = t\} = F_U(t)^n F_D(t/\alpha)^n F_U(t), \quad (3.64)
\]

\[
P\{E_n | G^1, S = t\} = F_U(t)^n F_D(t/\alpha)^n F_D(t/\alpha). \quad (3.65)
\]

Then, we have that the conditional Laplace transform of the completion time under event \(G^1\) and \(A_n\) for \(S = t\) is

\[
L_T(s | S = t, G^1, A_n) = L_U(s | t)^n L_D(s | t/\alpha)^n e^{-st} F_U(t)^n F_D(t/\alpha)^n F_U(t)
\]

For the union of events \(A_n\), \(n = 0, 1, 2, \ldots\),

\[
L_T \left( s | S = t, G^1, \bigcup_{n=0}^{\infty} A_n \right) = \sum_{n=0}^{\infty} L_U(s | t)^n L_D(s | t/\alpha)^n e^{-st} F_U(t)^n F_D(t/\alpha)^n F_U(t)
\]

\[
= \frac{e^{-st} F_U(t)}{1 - L_U(s | t) L_D(s | t/\alpha) F_U(t) F_D(t/\alpha)} \quad (3.66)
\]

Similarly,

\[
L_T(s | S = t, G^1, E_n) = (L_U(s | t) L_D(s | t/\alpha) F_U(t) F_D(t/\alpha))^n L_U(s | t) F_U(t) e^{-st/\alpha} F_D(t/\alpha),
\]

and,

\[
L_T \left( s | S = t, G^1, \bigcup_{n=0}^{\infty} E_n \right) = \frac{L_U(s | t) F_U(t) e^{-st/\alpha} F_D(t/\alpha)}{1 - L_U(s | t) L_D(s | t/\alpha) F_U(t) F_D(t/\alpha)} \quad (3.67)
\]

Summing up (3.66) and (3.67), we obtain the conditional Laplace transform of the completion time for customers arriving during an up period:

\[
L_T \left( s | S = t, G^1 \right) = \frac{e^{-st} F_U(t) + L_U(s | t) F_U(t) e^{-st/\alpha} F_D(t/\alpha)}{1 - L_U(s | t) L_D(s | t/\alpha) F_U(t) F_D(t/\alpha)}. \quad (3.68)
\]
For customers arriving during a down period, we interchange up and down periods in the notation for $A_n, E_n, n = 0, 1, \ldots$. The conditional completion time from Equation (3.11) becomes

$$\{T|G^2, S = t\} = \begin{cases} 
\sum_{i=1}^{n} D_i + \sum_{i=1}^{n} U_i + t, & \text{event } A_n, \quad n = 0, 1, 2, \ldots \\
\sum_{i=1}^{n+1} D_i + \sum_{i=1}^{n} U_i + t, & \text{event } E_n,
\end{cases} \quad (3.69)$$

with probabilities,

$$P\{A_n|G^2, S = t\} = F_U(t)^n F_D(t/\alpha)^n \overline{F_D(t/\alpha)}, \quad (3.70)$$

$$P\{E_n|G^2, S = t\} = F_U(t)^n F_D(t/\alpha)^{n+1} \overline{F_U(t)}. \quad (3.71)$$

From here,

$$L_T(s|S = t, G^2, A_n) = \left(L_U(s|t)L_D(s|t/\alpha)F_U(t)F_D(t/\alpha)\right)^n e^{-st/\alpha} \overline{F_D(t/\alpha)},$$

and,

$$L_T(s|S = t, G^2, E_n) = \left(L_U(s|t)L_D(s|t/\alpha)F_U(t)F_D(t/\alpha)\right)^n L_D(s|t/\alpha)F_D(t/\alpha)e^{-st} \overline{F_U(t)}.$$

Taking the union of events,

$$L_T\left(s|S = t, G^2, \bigcup_{n=0}^{\infty} A_n\right) = \frac{e^{-st/\alpha} \overline{F_D(t/\alpha)}}{1 - L_U(s|t)L_D(s|t/\alpha)F_U(t)F_D(t/\alpha)},$$

and,

$$L_T\left(s|S = t, G^2, \bigcup_{n=0}^{\infty} E_n\right) = \frac{L_D(s|t/\alpha)F_D(t/\alpha)e^{-st} \overline{F_U(t)}}{1 - L_U(s|t)L_D(s|t/\alpha)F_U(t)F_D(t/\alpha)},$$

and combining all events under $G^2$,

$$L_T\left(s|S = t, G^2\right) = \frac{e^{-st/\alpha} \overline{F_D(t/\alpha)} + L_D(s|t/\alpha)F_D(t/\alpha)e^{-st} \overline{F_U(t)}}{1 - L_U(s|t)L_D(s|t/\alpha)F_U(t)F_D(t/\alpha)}. \quad (3.72)$$

Combining Laplace transforms for events $G^1$ and $G^2$ from (3.68) and (3.72), and
Then it can be shown that, the events from Equation (3.59), and the probabilities from Equations (3.64)–(3.65). As in Theorem 1, we condition the service time to a task length (Theorem 2).

Proof. (Theorem 2)

As in Theorem 1, we condition the service time to a task length \( S = t \). Define the events from Equation (3.59), and the probabilities from Equations (3.64)–(3.65). Then it can be shown that,

\[
E[T \cdot 1\{A_n\}|S = t, G^1] = E[T|A_n, S = t, G^1] P\{A_n|G^1\} \\
= n(F_U(t)F_D(t/\alpha))^n (E[U] + E[D]) \overline{F_U(t)} + t(F_U(t)F_D(t/\alpha))^n \overline{F_U(t)},
\]

\[
E[T \cdot 1\{E_n\}|S = t, G^1] = (E[U] + E[D]) F_U(t) \overline{F_D(t/\alpha)} \cdot n(F_U(t)F_D(t/\alpha))^n \\
+ (E[U] + t/\alpha) F_U(t) \overline{F_D(t/\alpha)} (F_U(t)F_D(t/\alpha))^n.
\]

From here,

\[
E[T \cdot 1\{\cup_{i=0}^\infty A_n\}|S = t, G^1] = \frac{(E[U] + E[D]) \overline{F_U(t)}F_U(t)F_D(t/\alpha)}{(1 - F_U(t)F_D(t/\alpha))^2} + \frac{t \overline{F_U(t)}}{1 - F_U(t)F_D(t/\alpha)}.
\]  \hspace{1cm} (3.73)

\[
E[T \cdot 1\{\cup_{i=0}^\infty E_n\}|S = t, G^1] = \frac{(E[U] + E[D]) F_U(t)^2 \overline{F_D(t/\alpha)}F_D(t/\alpha)}{(1 - F_U(t)F_D(t/\alpha))^2} + \frac{(E[U] + t/\alpha) F_U(t) \overline{F_D(t/\alpha)}}{1 - F_U(t)F_D(t/\alpha)}.
\]  \hspace{1cm} (3.74)

Summing up (3.73) and (3.74) and simplifying we obtain

\[
E[T|S = t, G^1] = \frac{t \overline{F_U(t)} + \frac{t}{\alpha} F_U(t) \overline{F_D(t/\alpha)}}{1 - F_U(t)F_D(t/\alpha)} + \frac{F_U(t)}{1 - F_U(t)F_D(t/\alpha)} (E[U] + E[D]F_D(t/\alpha)).
\]  \hspace{1cm} (3.75)
Consider now the events from Equation (3.60), and the probabilities in (3.70)–(3.71).

Then it holds that,

\[ E[T \cdot 1\{A_n\} | S = t, G^2] = n(F_U(t)F_D(t/\alpha))^n (E[U] + E[D]) \overline{F_D(t/\alpha)} + t/\alpha (F_U(t)F_D(t/\alpha))^n \overline{F_D(t/\alpha)}, \]

\[ E[T \cdot 1\{E_n\} | S = t, G^2] = (E[U] + E[D]) \overline{F_U(t)} n(F_U(t)F_D(t/\alpha))^n + (t + E[D]) F_D(t/\alpha) \overline{F_U(t)}(F_U(t)F_D(t/\alpha))^n. \]

Taking the union of events we get,

\[ E[T \cdot 1\{\bigcup_{i=0}^{\infty} A_n\} | S = t, G^2] = \frac{(E[U] + E[D]) F_D(t/\alpha) \overline{F_D(t/\alpha)}}{(1 - F_U(t)F_D(t/\alpha))^2} + \frac{t/\alpha \overline{F_D(t/\alpha)}}{1 - F_U(t)F_D(t/\alpha)}, \]

(3.76)

\[ E[T \cdot 1\{\bigcup_{i=0}^{\infty} E_n\} | S = t, G^2] = \frac{(E[U] + E[D]) F_D(t/\alpha)^2 \overline{F_U(t)} F_U(t)}{(1 - F_U(t)F_D(t/\alpha))^2} + \frac{(t + E[D]) F_D(t/\alpha) \overline{F_U(t)}}{1 - F_U(t)F_D(t/\alpha)}. \]

(3.77)

and summing up (3.76) and (3.77) we obtain,

\[ E[T | S = t, G^2] = \frac{\frac{t}{\alpha} F_D(t/\alpha) + tF_D(t/\alpha) \overline{F_U(t)}}{1 - F_U(t)F_D(t/\alpha)} + \frac{F_D(t/\alpha)}{1 - F_U(t)F_D(t/\alpha)} (E[U]F_U(t) + E[D]). \]

(3.78)

Finally, combining (3.75) and (3.78) with the probabilities (3.3), replacing the appropriate exponential distributions for up and down time durations, and simplifying, we obtain (3.62).

\[ \square \]

3.5 Asymptotics of The Completion Time

In this section we show two results that help us understand the asymptotic behavior of the completion time distribution. This amounts to studying the characteristics of the distribution’s right-tail. The right-tail defines the probability of observing abnormally large completion times, and it is calculated as the complementary cumulative
distribution, as shown in (3.79).

\[ \bar{F}_T(t) = \mathbb{P}(T > t) = 1 - F_T(t) = 1 - \int_0^t f_T(t) dt. \] (3.79)

### 3.5.1 Repeat Different/REPLACE

The first result is for the exponential down periods case. We calculate the Laplace transform of the right-tail in closed form as shown in Corollary 10.

**Corollary 10** (Laplace Transform of the Right Tail). Consider the queueing system described in Section 3.1, and assume that the down period duration \( D \) is exponentially distributed with mean \( 1/r \). The Laplace transform of the right-tail of the completion time distribution is given as shown in equation (3.80).

\[ \mathcal{L}[\bar{F}_T](s) = \frac{f(1 - L_r(s))[s + f + r(1 - L_f(s))]}{(f + r)[(s + f)(s + r) - rf(1 - L_f(s))(1 - L_r(s))]} + \frac{rL_f(s)(1 - L_f(s))}{s + r} \] (3.80)

**Proof.** As before, call \( L_f(s) = L_S(s + f) \) and \( L_r(s) = L_S\left(\frac{s+r}{s}\right) \), then, from (3.19),

\[ L_T(s) = \frac{1}{(f + r)(1 - V(s))} \left( rL_f(s) \left(1 + \frac{f}{s + r} [1 - L_r(s)] \right) + fL_r(s) \left(1 + \frac{r}{s + f} [1 - L_f(s)] \right) \right) \] (3.81)

with,

\[ V(s) = \frac{rf(1 - L_f(s))(1 - L_r(s))}{(s + f)(s + r)}. \] (3.82)

From its definition from Equation (3.79), and properties of the Laplace transform for integrals, we have that

\[ \mathcal{L}[\bar{F}_T](s) = \frac{1}{s} - \frac{\mathcal{L}[f_T](s)}{s} = \frac{1 - L_T(s)}{s}. \] (3.83)

Then

\[ \mathcal{L}[\bar{F}_T](s) = \frac{(1 - V(s))(f + r) - \left( rL_f(s) \left(1 + \frac{f}{s + r} [1 - L_r(s)] \right) + fL_r(s) \left(1 + \frac{r}{s + f} [1 - L_f(s)] \right) \right)}{(f + r)(s - sV(s))}, \] (3.84)
which can be simplified to,
\[
\mathcal{L}[\hat{F}_T](s) = \frac{f(s + f)(1 - L_f(s)) + 2r f(1 - L_f(s))(1 - L_r(s)) + r(s + r)(1 - L_f(s))}{(f + r)(s + f)(s + r)(1 - V(s))}.
\]
(3.85)

Substituting $V(s)$ in the denominator, and simplifying leads to (3.80).

3.5.2 Repeat Identical/RESTART

For this case we are able to explicitly determine the asymptotic classification of the completion time distribution. We establish under what circumstances the completion time has a power tail. We start by giving a definition.

Consider a r.v. $X$ with support contained in $[0, +\infty)$ and distribution $F_X(t)$, and consider the following function:
\[
\phi(\theta; X) := \int_0^\infty e^{\theta t}dF_X(t).
\]
(3.86)

Call then $\theta_{\min}(X) := \sup\{\theta | \phi(\theta; X) < \infty\}$. Using this definition we can classify any distribution function on having finite range, a light tail, an exponential tail, or a heavy tail. This corresponds to the cases of finite support, $\theta_{\min}(X) = \infty$, $0 < \theta_{\min}(X) < \infty$, and $\theta_{\min}(X) = 0$, respectively [83]. We write $\theta_{\min}$ when there is no ambiguity on the r.v. $X$.

Theorem 3 shows conditions under which the completion time distribution is power tailed.

**Theorem 3.** Let $U$ and $D$ be exponentially distributed with failure rate $f$ and repair rate $r$, and call $\Delta = \min\{f, r/\alpha\}$. Assume that the service-time requirement (task size) has an exponential tail, that is: $0 < \theta_{\min}(S) = \mu < \infty$, and denote $\varepsilon := \theta_{\min}(S)/\Delta = \mu/\min\{f, r/\alpha\}$. Then,
\[
E[T^m] = \infty, \quad \forall \, m \geq \varepsilon,
\]
(3.87)
and the completion time has a power tail,

\[ F_T(t) \sim \frac{c}{t^\varepsilon}. \]  (3.88)

The fact that we compare the moment power \(m\) to the quantity \(\varepsilon = \mu / \min\{f, r/\alpha\}\) comes form the fact that the system is only stable if, on average, the task can be completed during normal conditions (which on average last for a time of \(1/f\)), and at the same time can be completed on an \(\alpha\)-slower deteriorated server that on average remains in that condition for a time of \(1/r\). The details of this derivation can be seen in the proof of Theorem.

Proof. As in Fiorini et al. [28], we calculate the moments of the completion time distribution to assess the heaviness of the tail. We assume here that the task size has an exponential tail, which amounts to \(\theta_{\min}(S) = \mu\), and denote \(\Delta = \min\{f, r/\alpha\} > 0\) and \(\Delta^+ = f + r/\alpha - \Delta = \max\{f, r/\alpha\} > 0\). Let \(U\) and \(D\) be exponentially distributed with failure rate \(f\) and repair rate \(r\), respectively. Their distributions are

\[ F_U(t) = 1 - e^{-ft}, \quad F_D(t) = 1 - e^{-rt}. \]

Notice that the task-size-conditioned Laplace transforms can be expressed as in (3.89)–(3.90).

\[ L_U(s|t) = E[e^{-sU|t} > U] = \int_0^t e^{-sx} dF_U(x|t) = \int_0^t e^{-sx} \frac{f e^{-fx}}{F_U(t)} dx = \frac{f}{s + f} \frac{1 - e^{-(s+f)t}}{1 - e^{-ft}}, \]  \hspace{1cm} (3.89)

\[ L_D(s|t/\alpha) = E[e^{-sD|t} > D] = \int_0^{t/\alpha} e^{-sx} dF_D(x|t/\alpha) = \int_0^{t/\alpha} e^{-sx} \frac{re^{-rx}}{F_D(t/\alpha)} dx = \frac{r}{s + r} \frac{1 - e^{-(s+r)t/\alpha}}{1 - e^{-rt/\alpha}}. \]  \hspace{1cm} (3.90)

From Theorem 1, the conditional Laplace transform of the completion time in Equation (3.61) becomes,

\[ L_T(s|t) = \frac{re^{-(s+f)t} \left[ 1 + \frac{f}{s + r} (1 - e^{-(s+r)t/\alpha}) \right] + fe^{-(s+r)t/\alpha} \left[ 1 + \frac{r}{s + f} (1 - e^{-(s+f)t}) \right]}{(r + f) \left[ 1 - \frac{rf}{(s + r)(s + f)} (1 - e^{-(s+f)t}) (1 - e^{-(s+r)t/\alpha}) \right]}, \]
where we call,

\[ h_r(s,t) := \frac{f}{s + r}(1 - e^{-(s+r)t/\alpha}), \]
\[ h_f(s,t) := \frac{r}{s + f}(1 - e^{-(s+f)t}), \]

obtaining,

\[ L_T(s|t) = \frac{1}{r + f} \cdot \frac{re^{-(s+f)t}(1 + h_r(s,t)) + f e^{-(s+r)t/\alpha}(1 + h_f(s,t))}{1 - h_r(s,t)h_f(s,t)} \] (3.91)

To calculate the \( m \)-th moment \( E[T^m] \) of the completion time \( T \), we resort to the derivatives of the Laplace transform of the completion time distribution evaluated at \( s = 0 \) as shown in Equation (3.92).

\[ E[T^m] = (-1)^m \frac{d^m}{ds^m} L_T(s) \bigg|_{s=0}. \] (3.92)

In particular, however, we calculate the \( m \)-th partial derivative with respect to \( s \) of the conditioned Laplace transform for the completion time in (3.91), evaluate at \( s = 0 \), and un-condition with respect to the task size whenever is possible.

Consider the following function definitions and notation for the partial derivatives of the conditional Laplace transform:

1. Main term for the \( n \)-th derivative:

\[ L_T(s; n|t) := \frac{1}{r + f} \cdot \frac{re^{-(s+f)t}(1 + h_r(s,t)) + f e^{-(s+r)t/\alpha}(1 + h_f(s,t))}{1 - h_r(s,t)h_f(s,t)}, \quad n = 0, 1, 2, \ldots \]

Notice how \( L_T(s; 0|t) := L_T(s|t) \).

2. First recurrent factor of higher-order term for \( n \)-th derivative:

\[ \nu_n(s, t) := \frac{re^{-(s+f)t} + \frac{f}{\alpha^n} e^{-(s+r)t/\alpha}}{r + f}, \quad n = 1, 2, \ldots \] (3.93)

It holds that \( \frac{\partial^n}{\partial s^n} \nu_1(s, t) = (-1)^n t^n \nu_{n+1}(s, t) \).
3. Recurrent rational exponential functions and terms:

\[ \theta(s, t; a, b) := \frac{e^{-(s+a)t/b}}{1 - e^{-(s+a)t/b}}, \quad (3.94) \]

It can be shown that for \( n \geq 1 \),

\[ \frac{\partial^n}{\partial s^n} \theta(s, t; a, b) = (-1)^n \left( \frac{t}{b} \right)^n \sum_{i=1}^{n+1} W(n, i) \frac{e^{-i(s+a)t/b}}{(1 - e^{-(s+a)t/b})^i} \]

\[ := (-1)^n \left( \frac{t}{b} \right)^n \sum_{i=1}^{n+1} W(n, i) \theta_i(s, t; a, b), \quad (3.95) \]

where \( W(n, i) \) denotes the \( n \)-th row and \( i \)-th column’s Worpitzky number from the Worpitzky triangle [59], as defined in (3.96).

\[ W(n, i) = \frac{1}{i} \sum_{j=0}^{i} (-1)^{i-j} \binom{i}{j} j^n, \quad n \geq 1, \quad 1 \leq i \leq n. \quad (3.96) \]

The first seven rows of the Worpitzky triangle are given in Table 3.2. It is easy to show that there is an equivalent representation of the partial derivatives in Equation (3.95) written in terms of the Eulerian numbers.

\[
\begin{array}{lcccccccc}
\hline
n \backslash k & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 1 & & & & & & \\
2 & 1 & 1 & & & & & \\
3 & 1 & 3 & 2 & & & & \\
4 & 1 & 7 & 12 & 6 & & & \\
5 & 1 & 15 & 50 & 60 & 24 & & \\
6 & 1 & 31 & 180 & 390 & 360 & 120 & & \\
7 & 1 & 63 & 602 & 2100 & 3360 & 2520 & 720 & \\
\end{array}
\]

Table 3.2: Worpitzky Triangle \( W(n, k) \)
In order to simplify the formulas we also give the following notation,

\[ \Theta(s, t; a, b) := \frac{t}{b} \cdot \frac{e^{-(s+a)t/b} - 1}{1 - e^{-(s+a)t/b}} - \frac{1}{s + a} \]

\[ = \frac{t}{b} \theta(s, t; a, b) - \frac{1}{s + a}. \quad (3.97) \]

From here, similarly, for \( n \geq 0, \)

\[ \frac{\partial^n}{\partial s^n} \Theta(s, t; a, b) = (-1)^n \left( \frac{t^{n+1}}{b^{n+1}} \sum_{i=1}^{n+1} W(n, i) \theta_i(s, t; a, b) - \frac{n!}{(s + a)^{n+1}} \right) \]

\[ := \Theta_{n+1}(s, t; a, b). \quad (3.98) \]

Finally, we call,

\[ \omega_n(s, t) := \Theta_n(s, t; f, 1) + \Theta_n(s, t; r, \alpha), \quad n = 1, 2, \ldots \quad (3.99) \]

and by definition, \( \frac{\partial^n}{\partial s^n} \omega_1(s, t) = \omega_{n+1}(s, t). \)

Now we calculate the first derivative of the conditioned Laplace transform in Equation
\[
\frac{\partial}{\partial s} L_T(s|t) = \frac{\partial}{\partial s} L_T(s; 0|t)
\]
\[
= -t \left[ re^{-(s+f)t}(1 + h_r(s, t)) + \frac{f}{\alpha} e^{-(s+r)t/\alpha}(1 + h_f(s, t)) \right]
\]
\[
+ \frac{r}{(r + f)(1 - h_r(s, t)h_f(s, t))^2} \left( \frac{f}{\alpha} e^{-(s+r)t/\alpha}(1 + h_f(s, t)) \right)
\]
\[
+ re^{-(s+f)t} \frac{\partial}{\partial s} h_r(s, t) + \frac{f}{\alpha} e^{-(s+r)t/\alpha} \left( \frac{\partial}{\partial s} h_f(s, t) + \frac{\partial}{\partial s} h_r(s, t) h_f^2(s, t) \right)
\]
\[
= -tL_T(s; 1|t) + \nu_1(s, t) \frac{\partial}{\partial s} \left( h_r(s, t) h_f(s, t) \right)
\]
\[
+ \frac{s + f}{r + f} \theta(s, t; f, 1) \left( \Theta(s, t; r, \alpha) + h_r(s, t) h_f(s, t) \Theta(s, t; f, 1) \right) \frac{h_r(s, t) h_f(s, t)}{(1 - h_r(s, t) h_f(s, t))^2}
\]
\[
+ \frac{s + r}{r + f} \theta(s, t; r, \alpha) \left( \Theta(s, t; f, 1) + h_r(s, t) h_f(s, t) \Theta(s, t; r, \alpha) \right) \frac{h_r(s, t) h_f(s, t)}{(1 - h_r(s, t) h_f(s, t))^2}
\]
\[
:= -tL_T(s; 1|t) + \nu_1(s, t) \frac{\partial}{\partial s} \left( h_r(s, t) h_f(s, t) \right)
\]
\[
+ \left[ \frac{s + f}{r + f} \theta(s, t; f, 1) A(s, t) + \frac{s + r}{r + f} \theta(s, t; r, \alpha) B(s, t) \right] \frac{h_r(s, t) h_f(s, t)}{(1 - h_r(s, t) h_f(s, t))^2}.
\]

(3.100)

It can be shown that,
\[
\frac{\partial}{\partial s} \left( h_r(s, t) h_f(s, t) \right) = h_r(s, t) h_f(s, t) \omega_1(s, t).
\]

(3.101)

Then, calling,
\[
u(s, t) := \frac{h_r(s, t) h_f(s, t)}{(1 - h_r(s, t) h_f(s, t))^2},
\]
we have from (3.100),
\[
\frac{\partial}{\partial s} L_T(s|t) = -tL_T(s; 1|t) + u(s, t) \left[ \nu_1(s, t) \omega_1(s, t) + \frac{s + f}{r + f} \theta(s, t; f, 1) A(s, t) + \frac{s + r}{r + f} \theta(s, t; r, \alpha) B(s, t) \right].
\]

(3.102)

Moreover,
\[
\frac{\partial}{\partial s} L_T(s; n|t) = -tL_T(s; n+1|t) + u(s, t) \left[ \nu_{n+1}(s, t) \omega_1(s, t) + \frac{s + f}{r + f} \theta(s, t; f, 1) A(s, t) + \frac{s + r}{r + f} \theta(s, t; r, \alpha) B(s, t) \right].
\]

(3.103)
To calculate the higher order partial derivatives of $L_T(s|t)$ we have to use the general Leibniz rule on the second term in (3.102). For this we have to study the partial derivatives of the factor $u(s,t)$, and the functions $A(s,t)$ and $B(s,t)$. We start with $u(s,t)$. From (3.101) we observe that,

$$h_r(s,t)h_f(s,t) = C(t)e^{\int \omega_1(s,t)ds},$$

(3.104)

where $C(t) := rf e^{-t(f+r/\alpha)}$.

Consider now $t$ fixed and the two following one-dimensional functions:

$$g(x(s)) := C(t)e^{\int x(s)ds}, \quad f(y) := \frac{y}{(1-y)^2}$$

(3.105)

Then we have that,

$$h_r(s,t)h_f(s,t) = g(w_1(s,t)),$$

(3.106)

$$u(s,t) = f(g(w_1(s,t))),$$

(3.107)

And $u_1(s,t)$ is a nested composite function.

The derivatives of composite functions can be written using Faà di Bruno’s formula [25] as shown in Equation (3.108):

$$\frac{d^n}{dx^n} f(g(x)) = \sum_{k=1}^{n} f^{(k)}(g(x)) \cdot B_{n,k}\left(g'(x), g''(x), \ldots, g^{(n-k+1)}(x)\right),$$

(3.108)

where $B_{n,k}$ are the incomplete (or partial) exponential Bell polynomials [11, 55] defined on Equation (3.109).

$$B_{n,k}(x_1, x_2, \ldots, x_{n-k+1}) = \frac{n!}{j_1!j_2!\cdots j_{n-k+1}!} \left(\frac{x_1}{1!}\right)^{j_1} \left(\frac{x_2}{2!}\right)^{j_2} \cdots \left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{j_{n-k+1}},$$

(3.109)
where the sumation is carried over all sequences $j_1, j_2, j_3, \ldots, j_{n-k+1}$ of non-negative integers such that conditions (3.110)-(3.111) are satisfied.

\[
j_1 + j_2 + \cdots + j_{n-k+1} = k; \tag{3.110}
\]
\[
j_1 + 2j_2 + 3j_3 + \cdots + (n - k + 1)j_{n-k+1} = n. \tag{3.111}
\]

The $n$-th complete exponential Bell polynomial is given by the sum of the incomplete polynomials as shown in Equation (3.112).

\[
B_n(x_1, \ldots, x_n) = \sum_{k=1}^{n} B_{n,k}(x_1, x_2, \ldots, x_{n-k+1}) \tag{3.112}
\]

The first five complete exponential Bell polynomials are shown below.

\[
B_0 = 1,
\]
\[
B_1(x_1) = x_1,
\]
\[
B_2(x_1, x_2) = x_1^2 + x_2,
\]
\[
B_3(x_1, x_2, x_3) = x_1^3 + 3x_1x_2 + x_3,
\]
\[
B_4(x_1, x_2, x_3, x_4) = x_1^4 + 6x_1^2x_2 + 4x_1x_3 + 3x_2^2 + x_4,
\]
\[
B_5(x_1, x_2, x_3, x_4, x_5) = x_1^5 + 10x_2x_1^3 + 15x_2^2x_1 + 10x_3x_1^2 + 10x_3x_2 + 5x_4x_1 + x_5.
\]

The derivatives of composite functions in Equation (3.108) are similar to the complete exponential Bell polynomials, where the variables $x_1, \ldots, x_n$ are replaced by the successive derivatives of the inner function $g(x)$, and the coefficients of the polynomials are factored by the derivatives of the outer function. Since $u(s, t) = f(g(w_1(s, t)))$ is a double composite function, its partial derivatives are given by nested Bell polynomials.

From (3.99) we know that the partial derivatives of the innermost function $\omega_1(s, t)$ with respect to $s$ are,

\[
\frac{\partial^n}{\partial s^n} \omega_1(s, t) = \omega_{n+1}(s, t). \tag{3.113}
\]
The partial derivatives of the first composite function \( g(w_1(s, t)) \) are calculated using (3.108) where the variable coefficients are the derivatives of \( g(x(s)) = C(t)e^{\int x(s)ds} \) evaluated in \( \omega_1(s, t) \), and the incomplete Bell polynomials are evaluated on the partial derivatives of \( \omega_1(s, t) \) from Equation (3.113). The first derivatives are given below, we omit the functional dependencies of \( \omega_n(s, t) \) to unclutter the notation.

\[
G_0(s, t) := g(w_1) = h_r(s, t)h_f(s, t) = h_r(s, t)h_f(s, t)B_0,
\]

\[
G_1(s, t) := \frac{\partial}{\partial s} h_r(s, t)h_f(s, t) = h_r(s, t)h_f(s, t)\omega_1 = h_r(s, t)h_f(s, t)B_1(\omega_1),
\]

\[
G_2(s, t) := \frac{\partial^2}{\partial s^2} h_r(s, t)h_f(s, t) = h_r(s, t)h_f(s, t)(\omega_1^2 + \omega_2) = h_r(s, t)h_f(s, t)B_2(\omega_1, \omega_2),
\]

\[
: \quad G_n(s, t) := \frac{\partial^n}{\partial s^n} h_r(s, t)h_f(s, t) = h_r(s, t)h_f(s, t)B_n(\omega_1, \omega_2, \ldots, \omega_n).
\]

The similarities of the derivatives to the complete Bell polynomials are due to the exponential nature of \( g(x(s)) \), which makes the successive derivatives of the outer function self-similar, and hence it can be factored out from the coefficients of the partial polynomials. Notice that derivatives of low order for \( g(x(s)) \) may also be easily calculated recursively following from Equation (3.101).

The derivatives of the outermost function \( f(y) \) are given in (3.114).

\[
\frac{d^n}{dy^n} f(y) = \frac{d^n}{dy^n} \frac{y}{(1-y)^2} = \frac{n!}{(1-y)^2} \frac{y + n}{(1-y)^n}.
\]

(3.114)

With this we give the general form of the partial derivatives of \( u(s, t) \) where we drop
the functional dependencies of $G_n(s,t)$ to unclutter the notation:

$$u_{n+1}(s,t) := \frac{\partial^n}{\partial s^n} u(s,t)$$

$$= \frac{\partial^n}{\partial s^n} h_r(s,t)h_f(s,t) \left( 1 - h_r(s,t)h_f(s,t) \right)^2$$

$$= \frac{\partial^n}{\partial s^n} f(g(w_1(s,t)))$$

$$= \sum_{k=1}^{n} f^{(k)}(g(w_1(s,t))) \cdot B_{n,k} \left( G_1, G_2, \ldots, G_{n-k+1} \right)$$

$$= \sum_{k=1}^{n} k! \left( k + h_r(s,t)h_f(s,t) \right) \left( 1 - h_r(s,t)h_f(s,t) \right)^{k+2} \cdot B_{n,k} \left( G_1, G_2, \ldots, G_{n-k+1} \right), \quad n = 1, 2, \ldots$$

(3.115)

We denote $u_1(s,t) := u(s,t)$.

The derivatives of $A(s,t)$ and $B(s,t)$ are also written in terms of Bell polynomials. This is due to the fact that, from Equation (3.104), and by definition both functions are composite functions:

$$A(s,t) = \Theta(s,t;r,\alpha) + h_r(s,t)h_f(s,t)\Theta(s,t;f,1)$$

$$= \Theta(s,t;r,\alpha) + C(t)e^{\int_0^t \omega_1(s,t)ds} \Theta(s,t;f,1)$$

$$= \Theta(s,t;r,\alpha) + C(t)e^{\int_0^t (\Theta_n(s,t;f,1) + \Theta_n(s,t;r,\alpha))ds} \Theta(s,t;f,1),$$

$$B(s,t) = \Theta(s,t;f,1) + h_r(s,t)h_f(s,t)\Theta(s,t;r,\alpha)$$

$$= \Theta(s,t;f,1) + C(t)e^{\int_0^t (\Theta_n(s,t;f,1) + \Theta_n(s,t;r,\alpha))ds} \Theta(s,t;r,\alpha).$$

Denote by $\Theta_{1:k}(s,t;a,b)$ the ordered list $\Theta_1(s,t;a,b), \ldots, \Theta_k(s,t;a,b)$. It can be shown by induction that,

$$\frac{\partial^n}{\partial s^n} A(s,t) = \Theta_{n+1}(s,t;r,\alpha) + h_r(s,t)h_f(s,t) \sum_{k=0}^{n} \binom{n}{k} B_{n-k+1}(\Theta_{1:(n-k+1)}(s,t;f,1))B_k(\Theta_{1:k}(s,t;r,\alpha)), $$

(3.116)

$$\frac{\partial^n}{\partial s^n} B(s,t) = \Theta_{n+1}(s,t;f,1) + h_r(s,t)h_f(s,t) \sum_{k=0}^{n} \binom{n}{k} B_{n-k+1}(\Theta_{1:(n-k+1)}(s,t;r,\alpha))B_k(\Theta_{1:k}(s,t;f,1)).$$

(3.117)
Where the summations are binomial expansions on the indices of $\Theta_k(s, t; a, b)$, with the first index shifted to the right by one. Getting back to Equation (3.102), we had that
\[
\frac{\partial}{\partial s} L_T(s|t) = -tL_T(s; 1|t) + u_1(s, t) \left[ v_1(s, t)\omega_1(s, t) + \frac{s + f}{r + f} \theta(s, t; f, 1)A(s, t) + \frac{s + r}{r + f} \theta(s, t; r, \alpha)B(s, t) \right].
\]
(3.118)

It can be shown from successive differentiation of Equation (3.118) and considering (3.103) that the $n$-th partial derivative of the conditioned Laplace transform of the completion time is given by the expression in Equation (3.119).
\[
\frac{\partial^n}{\partial s^n} L_T(s|t) = (-t)^n L_T(s; n|t) + \sum_{k=1}^{n} (-t)^{n-k} \frac{\partial^{k-1}}{\partial s^{k-1}} \left[ u_{n-k+1}(s, t)\omega_1(s, t) + A(s, t) \left( \frac{s + f}{r + f} \theta(s, t; f, 1) \right) + B(s, t) \left( \frac{s + r}{r + f} \theta(s, t; r, \alpha) \right) \right],
\]
(3.119)

where the partial derivatives in the right hand side are calculated with the General Leibniz rule, which is given in (3.120) for the product of 3 functions $f_1(x)$, $f_2(x)$, $f_3(x)$:
\[
\frac{d^n}{dx^n} (f_1(x)f_2(x)f_3(x)) = \sum_{k_1 + k_2 + k_3 = n} \frac{n!}{k_1!k_2!k_3!} \frac{d^{k_1}}{dx^{k_1}} f_1(x) \frac{d^{k_2}}{dx^{k_2}} f_2(x) \frac{d^{k_3}}{dx^{k_3}} f_3(x),
\]
(3.120)

where the sum extends over all triplets $(k_1, k_2, k_3)$ of non-negative integers that are such that $\sum_{t=1}^{3} k_t = n$.

Finally, the $m$-th moment of the completion time $T$ is obtained by unconditioning (3.119) with respect to the task size $t$ as shown in Equation (3.121), whenever the improper integral converges.
\[
E[T^m] = (-1)^m \int_{0}^{\infty} \left( \frac{\partial^m}{\partial s^m} L_T(s|t) \bigg|_{s=0} \right) dF_S(t)
\]
\[
= (-1)^m \mu \int_{0}^{\infty} \left( \frac{\partial^m}{\partial s^m} L_T(s|t) \bigg|_{s=0} \right) e^{-\mu t} dt.
\]
(3.121)

To study the convergence of the integral in (3.121) we look at the asymptotic behavior of the different terms and factors in (3.119) evaluated at $s = 0$. We will use the
definition of asymptotic equivalence given in (3.122). Recall also that we denote 
\[ \Delta = \min\{f, r/\alpha\} > 0 \] 
and \[ \Delta^+ = f + r/\alpha - \Delta = \max\{f, r/\alpha\} > 0. \]

\[ \psi_1(t) \sim \psi_2(t) \] if and only if 
\[ \lim_{t \to \infty} \frac{\psi_1(t)}{\psi_2(t)} = 1. \] (3.122)

We classify according to the asymptotic order.

1. Constant factors:

(a) Main factor of the first term: \( L_T(0; n|t) \)

\[
L_T(0; n|t) \sim C_1(n) := \begin{cases} 
1, & f < r/\alpha, \\
1/\alpha^n, & f > r/\alpha, \quad n = 0, 1, 2, \ldots \\
\frac{1}{2}(1 + 1/\alpha^n), & f = r/\alpha 
\end{cases}
\]

(b) Recurrent factor \( \Theta_n(0; t; a, b) \):

\[
\Theta_0(0; t; a, b) = \frac{t}{b} \frac{e^{-at/b}}{1 - e^{-at/b}} - 1/a \sim -1/a,
\]

\[
\Theta_n(0; t; a, b) = (-1)^{n-1} \left( \frac{t^n}{b^n} \sum_{i=1}^{n} W(n, i) \frac{e^{-i-at/b}}{(1 - e^{-at/b})^i} - \frac{(n-1)!}{a^n} \right) \sim (-1)^n \frac{(n-1)!}{a^n}, \quad n = 2, 3, \ldots
\]

(c) Recurrent factor \( \omega_n(0; t) \):

\[
\omega_n(0; t) = \Theta_n(0; t; f, 1) + \Theta_n(0; t; r, \alpha) \sim (-1)^n (n-1)! \left( \frac{1}{f^n} + \frac{1}{r^n} \right), \quad n = 1, 2, \ldots
\]

(d) Recurrent factors \( A(s, t) \) and \( B(s, t) \), and derivatives: from (3.116)–(3.117), and due to the continuity of the Bell polynomials, we have that for every 
\( n \geq 0 \) there exist constants \( C_A(n) > 0 \) and \( C_B(n) > 0 \) such that,

\[
\frac{\partial^n}{\partial s^n} A(s, t) \sim (-1)^{n+1} C_A(n),
\]

\[
\frac{\partial^n}{\partial s^n} B(s, t) \sim (-1)^{n+1} C_B(n).
\]

2. Polynomial factors with exponential damping:
(a) Recurrent factor \( \nu_n(0|t) \):

\[
\nu_n(0|t) \sim e^{-\Delta t \cdot C_\nu(n)} := e^{-\Delta t \cdot \frac{r}{r+f}} \begin{cases} 
\frac{r}{r+f}, & f < r/\alpha, \\
\frac{f}{(r+f)a^\alpha}, & f > r/\alpha, \quad n = 1, 2, \ldots \\
\frac{1}{r+f} \left( r + \frac{f}{a^\alpha} \right), & f = r/\alpha
\end{cases}
\]

(b) Recurrent factors \( \theta_k(0; t; a, b) \) and derivatives of \( \theta(0; t; a, b) \):

\[
\theta_k(s, t; a, b) \sim e^{-k \cdot at/b},
\]

\[
\frac{\partial^n}{\partial s^n} \theta(s, t; a, b) \bigg|_{s=0} = (-1)^n \left( \frac{t}{b} \right)^n \sum_{i=1}^{n+1} W(n, i) \theta_i(0; t; a, b) \sim (-1)^n \left( \frac{t}{b} \right)^n e^{-at/b}, \quad n = 1, 2, \ldots
\]

(c) Accompanying factors of \( A(0, t) \) and \( B(0, t) \), and derivatives:

\[
\frac{\partial^n}{\partial s^n} \left( \frac{s + a}{r + f} \theta(s, t; a, b) \right) \bigg|_{s=0} \sim \frac{1}{r + f} \left[ n + a(-1)^{n-1} \frac{t^{n-1}}{b^{n-1}} \right] e^{-at/b}, \quad n = 1, 2, \ldots
\]

3. Exponential growth factors: recurrent factor \( u_n(0, t) \).

Again due to the continuity of the Bell polynomials, for every \( n \in \mathbb{N} \) there exists a constant \( C_G(n) > 0 \) such that \( G_n(0, t) \sim (-1)^n C_G(n) \), and since the dominant term in (3.115) is for \( k = n \), there exists a constant \( C_B(n - 1) > 0 \) such that \( B_{n-1, n-1}(G_1) \big|_{s=0} \sim (-1)^{n-1} C_B(n - 1) \). From here, given that,

\[
\frac{(n + h_r(0, t)h_f(0, t))}{(1 - h_r(0, t)h_f(0, t))} \sim (n + 1) e^{(n+2)\Delta t},
\]

we obtain that,

\[
u_n(0, t) \sim (-1)^{n-1} n! C_B(n - 1) \cdot e^{(n+1)\Delta t}, \quad n = 1, 2, \ldots
\]

Notice that \( u_n(0, t) \) is the only factor that has an asymptotic behavior that is not decaying nor constant for large \( t \), and also that the asymptotic order increases with \( n \), which is the order of the partial derivative.

It is easy to see from the first derivative in Equation (3.102) that \( \frac{\partial}{\partial s} L_T(s|t) \big|_{s=0} \sim -Ce^{\Delta t} \), for some constant \( C > 0 \), and that the expected Completion time \( E[T] \) only
exists when $\Delta = \min\{f, r/\alpha\} < \mu$. For higher order derivatives in Equation (3.119) we only have to observe last term in the summation, which will yield the higher order terms overall. Additionally, since the partial derivatives of $u(s, t)$ dominate the asymptotic behavior, we only have to look at the term coming from the Leibniz rule for which $u(s, t)$ has the partial derivative of order $n - 1$ and the remaining factors are not differentiated. Then, for $n \geq 1$,

$$
\frac{\partial^n}{\partial s^n} L_T(s|t) \bigg|_{s=0} \sim u_n(0, t) \left[ v_1(0, t) \omega_1(0, t) + A(0, t) \left( \frac{r}{r+\gamma} \theta(0, t; f, 1) \right) + B(0, t) \left( \frac{r}{r+\gamma} \theta(0, t; r, \alpha) \right) \right] \sim C(n)e^{n\Delta t},
$$

(3.123)

where $C(n)$ is a constant that depends on $n$. Then, from (3.121), the $m$-th moment $E[T^m]$ is not finite whenever $m\Delta - \mu \geq 0$ which is $m \geq \mu/\Delta = \mu/\min\{f, r/\alpha\} = \varepsilon$, and the completion time distribution is power-tailed [33], that is, there exists a positive constant $c$ such that

$$
\overline{F_T(t)} \to \frac{c}{t^\varepsilon}.
$$

(3.124)

3.6 Insensitivity of the Stationary Mean Measures to How Interruptions Affect Servers

Several different queueing systems with service degradation can be defined by using the completion time model described above as a generally distributed service time. Here we discuss models with Poisson arrivals with parameter $\lambda$ and multiple servers.

With exponential up and down periods, the process controlling the service rate is a two-state continuous-time Markov chain, independent of the arrival process, and is considered as the external environment. If all servers are controlled simultaneously, which means that interruptions occur system-wide then the queue is said to have a Markovian service process (MAP). On the other hand, similar interruption processes may affect each server independently.
In the case of finitely many servers, system-wide partial failures, and exponential task times, the system becomes the M/MSP/c queue analyzed in [7] with two service states. The independent server breakdown case is studied by Mitrany & Avi-Itzhak [58].

In the case of infinitely many servers, system-wide partial failures, and exponential task time distribution, the system coincides with the M/MSP/∞ queue considered in [9]. Baykal-Gürsoy and Xiao [9] show that the steady-state number in the system is the sum of two independent random variables: a Poisson r.v. representing the stationary number of customers in an uninterrupted M/G/∞ system, and a randomized Poisson r.v. representing the extra customers accumulated during interruptions. Then, the mean steady-state number of customers in the system is derived as

$$E[N] = \frac{\lambda}{\mu} + \frac{\lambda f(1 - \alpha)}{f + r} \cdot \frac{f + r + \mu}{\mu(f\alpha + r + \mu\alpha)}. \tag{3.125}$$

Assuming that there are no jockeying between the servers, one can analyze the independent server interruption case similar to an M/G/∞ system as will be discussed below.

When Markovian service interruptions arrive independently to each of the infinitely many servers, a job joining the system experiences exactly the same task completion time derived in Corollary 1. Each task completion time is independent and identically distributed. Hence, this system becomes an M/G/∞ system with the service time equal to the completion time, $T$, that is given in the frequency domain by Equations (3.19) and (3.20).

Clearly, the stationary number of customers in the M/G/∞ queue is Poisson distributed with parameter $\rho = \lambda \cdot E[T]$. Hence, the expected number of customers in the system, $N$, using the expected completion time from Equation (3.30) is given as in (3.125) (see in [9] Eq. (3.6)).

The expected system time for this system, i.e., the job completion time, in turn
coincides with the mean system time for the M/MSP/∞ queue obtained via Little’s law from Equation (3.125) as \( E[T] = E[N]/\lambda \). For the M/MSP/∞ queue in [9], however, the steady state variance of the number in the system is not equal to its expected value, which has to be the case in the M/G/∞ setting with independent server failures and our completion time service time model.

### 3.7 Conclusions and Future Work

We study the task completion time of a server experiencing randomly occurring service deterioration. Focusing on preempt-replace and preempt-repeat service recovery disciplines following each service interruption, we derive the Laplace transform of the task completion times using counting arguments. We observe that in general the resulting distributions are difficult to obtain explicitly in the time domain and one has to resort to numerical inversion of the transforms. For the specific case of exponential down period and exponential task time case we can determine the exact form of the completion time. Furthermore, we show how the steady-state mixture can be compared against the expected completion time, which can be useful for applications in which a simpler steady state model is preferred.

The asymptotic analysis demonstrate that in the preempt-repeat service discipline even when the task time distribution has exponential tail, the completion time distribution may have power tail. Moreover, the connection between the expected task completion time presented in this study and the expected system time for the M/MSP/∞ queue studied by Baykal-Gürsoy and Xiao [9], reveals that the first order moments at an infinite server queue in random environment are insensitive to how the random environment affects the servers.

As future work we aim to obtain some analytical properties of the tail of the distribution, such as the asymptotic classification and the decay behavior. Additionally
the connection between the expected number of customers in our model, and the M/MSP/$\infty$ queue studied by Baykal-Gürsoy and Xiao [9], suggests the existence of a deeper property for this family of queueing models which ought to be explored.
Chapter 4

Travel Time and Traffic Volume Reliability

As mentioned in the Introduction, one of the main contributions of this dissertation is creating a bridge between queueing models and traffic theory. With this objective in mind, from this chapter onwards we turn our attention to applications of the queueing models to vehicle-traffic planning and operations.

The present chapter, in particular, builds upon the completion time distribution results presented in Chapter 3 that are used to obtain estimates of the reliability of the travel time of a vehicle transversing an incident-prone corridor. Travel time reliability estimates are used for traffic planning and budget management, and inform decision makers on weak links on the network and appropriate prioritization.

For this application we interpret the service requirement distribution $S$ as a travel time requirement under normal conditions, which has to do with the random speed that an arriving vehicle carries when entering the corridor.

We start in Section 4.1 describing the particular version of the completion time
model (Chapter 3) that will be used in this application. Section 4.2 shows the validation strategy and calibration methodology using traffic data on traffic volume and occupancy. In Section 4.3 we demonstrate the calculation of reliability indices and how to perform scenario analysis for planning purposes. We discuss the results and derive some conclusions in Section 4.4, and finalize with Section ?? where we comment on some pending tasks for future research.

4.1 Notation and Analytic Model

Our model represents the travel time of a single vehicle traveling on a short corridor that is subject to random quality of service degradations. To this end, we consider a job that arrives to a service system in which there is no waiting time to start service. We assume that once the job arrives it is processed at the current service speed that could be in either one of two different speeds [9, 10].

Since the speed of the vehicle may depend on the driver, state regulations, and roadway conditions, due to the randomness in the chosen speeds, we assume that each vehicle takes a random amount of time to traverse the corridor. Under normal traffic conditions (under neither recurrent nor nonrecurrent congestion), the travel time requirement, $S$, is distributed randomly with $F_S(t)$ as the cumulative distribution function (CDF), that is, $F_S(t) = P\{S \leq t\}$, $f_S(t)$ as the corresponding probability density function (PDF), and $L_S(s)$ as the Laplace transform of $S$, that is, $L_S(s) = E[e^{-sS}]$. We define the mean travel time under normal conditions as $\mathbb{E}[S] = 1/\mu$. Conversely, $\mu$ denotes the rate of job completion, or service rate under normal conditions. Notice that the range of speeds that will be observed when different vehicles utilize the road will be then ultimately determined by the choice of this travel time distribution.

When an incident happens, the vehicle’s speed drops to $\alpha$ times the normal speed with $0 < \alpha < 1$. Hence, the mean travel time under normal conditions, $1/\mu$, will
increase to $1/\alpha \mu$, a higher level under incident conditions, with $\mu' = \alpha \mu$ representing the job completion rate under incidents.

These degradations in the quality of service happen randomly, and we assume that whenever the system is working properly (corresponding to normal traffic conditions) it stays in this state for an exponentially distributed amount of time with mean $1/f$. Here $f$ represents the rate of incidents. We call up periods the periods of time in which the system is continuously working without degradation. Similarly, we call down periods the time periods in which the roadway is experiencing a degradation of service. For the purpose of this paper we assume that the down periods also have a duration that is exponentially distributed with mean duration $1/r$, where $r$ is the repair rate. This type of service process is said to be Markov modulated [63]. We verify in section 4.2 that the exponential distribution assumption for the up and down time durations is valid for our data.

We mention here that one technical assumption in our model requires that if a quality of service change occurs while a vehicle is being served the corresponding travel time requirement needs to be resampled and a new travel time must be added to the travel time already expanded. This assumption will only make an impact when the travel time distribution under normal conditions is on the same scale as the mean up and down times. This scenario, however is unlikely to be realized in a realistic setting, since usually travel times for any traffic-homogenous road section are much faster than the times between incidents or incident durations.

Our main goal is to estimate the (un)reliability of the travel times experiencing such random incidents. One of the (un)reliability measures that we will consider is the probability that the travel time, $T$, exceeds a given value, i.e., $\bar{F}(t) = 1 - F_T(t) = P\{T > t\}$. In order to obtain this travel time (un)reliability measure, we will first obtain the Laplace transform of the travel time distribution. The details of the
4.1.1 Full Travel Time Distribution

In this section, we present the Laplace transform, \( L_T(s) \), of the full travel time, \( T \). While under some very special distributional assumptions the Laplace transform can be inverted analytically. In general it is necessary to resort to numerical Laplace inversion methods [39, 2].

Equations (4.1)-(4.2) present the Laplace transform of the travel time distribution for the aforementioned system in terms of the Laplace transform of \( S \), \( L_S(s) \):

\[
L_T(s) = \frac{r}{f + r} \cdot \frac{1}{1 - V(s)} \left( L_S(s + f) + \frac{f}{s + f} \left[ 1 - L_S(s + f) \right] L_S \left( \frac{s + r}{\alpha} \right) \right) + \frac{f}{f + r} \cdot \frac{1}{1 - V(s)} \left( L_S \left( \frac{s + r}{\alpha} \right) + \frac{f}{f + r} \frac{r}{s + r} \left[ 1 - L_S \left( \frac{s + r}{\alpha} \right) \right] L_S(s + f) \right),
\]

with

\[
V(s) = \frac{rf \left[ 1 - L_S(s + f) \right] \left[ 1 - L_S \left( \frac{s + r}{\alpha} \right) \right]}{(s + f)(s + r)}.
\]  

The two terms in (4.1) are weighted by the probabilities that a traveler arrives to the system during an up period, \( r/(f + r) \), or arrives during a down period, \( f/(f + r) \). The multipliers of these probabilities are the Laplace transforms of the travel time of a traveler who arrives during an up or down period, respectively. The above structure of \( L_T(s) \) informs us that these two travel times are statistically independent.

It is worth mentioning that the arguments employed to obtain the analytical results that are used in this paper allow for a more general result than the one shown in equations (4.1)-(4.2) for exponentially distributed down periods. In fact, a generally distributed service system with general down periods and partial breakdowns can be analyzed, as shown in Chapter 3. Additionally, our analytical results replicate the results Kulkarni et al. [49] for a preemptive-repeat different service discipline, but the
derivations and rationale behind the solution techniques are much simpler than the techniques presented in the cited studies.

4.1.2 Specializing the Service Time Distribution

By selecting a specific distribution for $S$, the travel-time under normal conditions, the Laplace transform in (4.1)-(4.2) can be fully described. Of course, if the Laplace transform of the selected travel time distribution has a closed form expression then we directly obtain a closed form for the full travel time distribution $T$. This, however, while being very practical, does not limit the use of non-parametrical distributions for $S$ that can be dealt with numerically.

The distribution of $S$ can be calibrated to meet the traffic characteristics of a specific road link. We show this procedure later in section 4.3.

**Triangular Service Time**

As confirmed by a number of empirical studies, travel time during free flow conditions shows a unimodal distribution with most vehicles traveling at free-flow speed and progressively smaller proportions traveling either slower or faster than the free flow speed [34, 71]. In this vein we propose the use of a triangular service time distribution on a travel time interval $[a, b]$ with $a < b$, and vertex at $c$, $a \leq c \leq b$. Here $a$ denotes the minimum and $b$ denotes the maximum travel time. Figure 4.1 shows a diagram of a triangular distribution. Notice that $f_S(c) = 2/(b - a)$ holds and that the mean service requirement is $1/\mu = (a + b + c)/3$. A triangular distribution has the advantage of yielding its Laplace transform in closed form while being general enough to be calibrated to a variety of traffic conditions.

Setting the parameters $A = b - a$, $B_1 = c - a$, $B_2 = b - c$, $D = AB_1B_2$, and
the functions $g_1(s) = B_2 e^{-a(s+f)} - A e^{-c(s+f)} + B_1 e^{-b(s+f)}$, and $g_2(s) = B_2 e^{-a/s(s+r)} - A e^{-c/s(s+r)} + B_1 e^{-b/s(s+r)}$, the Laplace transform of the full travel time distribution is given as shown in equation (4.3).

$$L_T(s) = \frac{r}{(f+r)} \left[ \frac{2(s+f)g_1(s)[D(s+r)^2(s+f+r) - 2\alpha^2f g_2(s)]}{D^2(s+f)^3(s+r)^3 - rf [D(s+f)^2 - 2g_1(s)] [D(s+r)^2 - 2\alpha^2g_2(s)]} \right] + \frac{f}{(f+r)} \left[ \frac{2\alpha^2(s+r)g_2(s)[D(s+f)^2(s+f+r) - 2rg_1(s)]}{D^2(s+f)^3(s+r)^3 - rf [D(s+f)^2 - 2g_1(s)] [D(s+r)^2 - 2\alpha^2g_2(s)]} \right].$$

(4.3)

The transform in (4.3) has negative poles, i.e., it is stable, and can be inverted numerically as it will be shown in section 4.3.

**Uniform Service Time**

If no information is known about the travel time without incidents except for a range of vehicle speeds, then the use of a uniform distribution can be warranted, yielding an even simpler transformed travel time distribution. Let $[a, b]$ be the travel time interval obtained from the speed range and the length of the road link, and assume that $S$ is a uniform random variable in the interval $[a, b]$, $0 < a < b$, that is $f_S(t) \sim \mathcal{U}(a, b)$. Figure 4.2 shows a diagram of such distribution. Here the mean
service time is $1/\mu = (a + b)/2$.

$$f_s(t)$$

\[
\frac{1}{b-a}
\]

\[a\]

\[b\]

\[t\]

Figure 4.2: Uniform Service Time Distribution: Application

Denoting $A = b - a$, $h_1(s) = e^{-a(s+f)} - e^{-b(s+f)}$, and $h_2(s) = e^{-\eta/\alpha(s+r)} - e^{-b/\alpha(s+r)}$, the Laplace transform of the travel time distribution is given in equation (4.4).

\[
L_T(s) = \frac{r}{(f+r)} \left[ \frac{(s + f)h_1(s)[A(s + r)(s + r + f) - \alpha fh_2(s)]}{A^2(s + f)^2(s + r)^2 - rf[A(s + f) - h_1(s)][A(s + r) - \alpha h_2(s)]} + \right] +
\]

\[
\frac{f}{(f+r)} \left[ \frac{\alpha(s + r)h_2(s)[A(s + f)(s + f + r) - rh_1(s)]}{A^2(s + f)^2(s + r)^2 - rf[A(s + f) - h_1(s)][A(s + r) - \alpha h_2(s)]} \right].
\]

(4.4)

This transform has again negative poles and can be inverted numerically.

**Shifted Exponential Service Time**

Another more realistic alternative for the service time requirement (travel time) distribution is a shifted exponential model. This is motivated by the study of the empirical travel time distributions observed in the vehicular traffic data, as it will be validated in Section 4.3 of results.

The shifted exponential density function is defined in equation (4.5), where $\eta$ is the rate parameter, $c$ is the right-shift applied to a regular exponential distribution, and $u(\cdot)$ is the heaviside step function. Figure 4.3 shows a diagram of such distribution.
The mean of the shifted exponential distribution is the shift $c$ plus the reciprocal of $\eta$. If the mean time requirement is already known to be $1/\mu$, then we would require that $c + 1/\eta = 1/\mu$, which forces $\eta = \mu/(1 - \mu c)$ adding the condition $1 - \mu c > 0$, or $c < 1/\mu$.

$$f_S(t) = \eta u(t - c)e^{-\eta(t-c)} \quad (4.5)$$

Figure 4.3: Shifted Exponential Service Time Distribution: Application

The Laplace transform of the density function in Equation (4.5) is $L_S(s) = \mu e^{-cs}/(s + \mu)$, and the full travel time distribution in the Laplace domain for this case is given by equation (4.6).

$$E[e^{-sT}] = \frac{r}{f + r} \frac{\eta e^{-c(s+f)}(s + f)[(s + r + f)(s + r + \alpha \eta) - \alpha f \eta e^{-c/\alpha (s+r)\eta}]}{[(s + f)(s + r)(s + r + \alpha \eta)(s + f + \eta) - rfW(s)]} +$$

$$\frac{f}{f + r} \frac{\alpha \eta e^{-c/\alpha (s+r)\eta}(s + r)[(s + r + f)(s + f + \eta) - r \eta e^{-c(s+f)}]}{[(s + f)(s + r)(s + r + \alpha \eta)(s + f + \eta) - rfW(s)]}, \quad (4.6)$$

where $W(s) := [s + f + \eta(1 - e^{-c(s+f)\eta})] [s + r + \alpha \eta(1 - e^{-c/\alpha (s+r)})].$

### 4.2 Model Implementation

The implementation of our model for the analysis of an actual traffic corridor requires minimal statistical information about (1) the vehicle’s travel time under non-congested traffic conditions and (2) the events that deteriorate/deviate the traffic from these conditions. In this section we validate our model’s technical assumptions
using vehicular traffic data and also show how it can be reasonably calibrated using readily available data averages.

We use traffic data from an 8.5-mile freeway segment from Interstate 894 in Milwaukee, WI. Figure 4.4, shows a composite map of the corridor.

![Figure 4.4: Map of Interstate 894](image)

The traffic data is reported by thirty-five detector stations located near or at interchanges with eighteen detectors in the South-East direction and seventeen in the West-North direction. The average spacing between detector stations is half a mile. Every sensor gives minute-timestamped speed, volume, and occupancy data for a 14-month period from January 1st 2008 to February 28, 2009. Speed data is reported in miles per hour, volume in vehicles per hour, and occupancy in fractions of an hour of active vehicle presence. The speed data is truncated (censored) at the local speed limit of either 55 or 60 miles per hour. The proportion of truncated speed data ranges from 1% to 70%, with an average of 51%. We recalculate these truncated values by using the deterministic algorithm given by Dailey in [17] for speed calculations using
single inductance loop measurements of volume and occupancy. We discard any data point with post-processed speed above 85 miles per hour. On average we discard 3% of the data at this last stage.

As sources of traffic deterioration we use processed reports of incidents and historical adverse weather conditions. The original incident data comes from reports of both the local police authorities and the State Traffic Operations Center (STOC) of the Wisconsin DOT. The data includes the start and end times of each incident, together with the identification of the nearest detector station.

We obtain precipitation data from the Climate Data Online system of the National Climatic Data Center of the National Oceanic and Atmospheric Administration (NOAA) [60]. This data is given as the hourly amount of precipitation in hundredths of inches recorded at the Milwaukee Mitchell International Airport weather station, during the same time period as the traffic data. The data also include the information on snow days and days with fog and/or thunderstorms.

The model requires knowledge of two statistical parameters about the traffic characteristics (either $\mu$ and $\alpha$, or equivalently $\mu$ and $\mu'$), the description of the travel time distribution under normal conditions, and two more parameters about the frequency and duration of incidents and/or adverse weather events ($f$ and $r$). Since we have detailed information about the traffic conditions and incidents for specific sections of I-894 (separated by the detector stations) we are able to validate the model for relatively short sections of the road.

Finally, because we seek to approximate non-recurrent traffic congestion, after analyzing the traffic patterns we restrict the data to non-rush hour times of the day on weekdays, that is: Monday through Friday at 5am-7am, 9am-4pm, and 6pm-00am.
4.2.1 Validation of Model Assumptions

We are building here an aggregated yearly model and hence we consider long weather events in the data, and we show how the assumptions of the model hold in this case. These assumptions, however, are still valid and robust in other time-scales, and the model can be both calibrated and/or validated for smaller time-scales as needed.

In order to proceed we define up and down periods for the traffic system: we label as an up period any time interval from our selected time windows for which there are (1) no incident reports and (2) no adverse weather reports. The distribution of the travel time during the up periods will give us information on how the system behaves under non-congested conditions. Similarly, but not conversely, we define as a down period any time interval with reported traffic incidents and/or serious weather events (snow, and over half an inch of hourly rain). The deterioration frequency $f$ and the restoration frequency $r$ can be obtained from the average duration of these intervals. Let $\bar{U}$ denote the mean up time duration of the system and $\bar{D}$ the mean down time duration. Then $f = 1/\bar{U}$ and $r = 1/\bar{D}$.

For demonstration purposes we select a southbound half a mile stretch between Cleveland and Oklahoma avenues. This is one of the most incident prone sections of the road with 137 recorded incidents in the 14-month period of the study. Figures 4.5(a) and 4.5(b) show probability-scaled histograms of the durations (in days) of the up and down periods for this road section, respectively, as well as the corresponding exponential distributions fitted to the data.

The high $R$ squared values (in this case 0.960 and 0.969 for up and down time periods, respectively) are robust to changes in both the selected section of the road and definitions of the down periods. This means that the assumption of exponential up and down time durations is well supported by the data for this sensor. This assumption is also valid for all other sensors with an average $R^2$ value of 0.969, and
Figure 4.5: Probability-Scaled Event Duration Histograms and Adjusted Exponential Distributions: (a) Up Period Durations, (b) Down Period Durations.

In our example we obtain the mean up time duration as $\bar{U} = 1/f = 30.59 \text{ [hour]}$ and the mean down time duration as $\bar{D} = 1/r = 19.55 \text{ [hour]}$. The failure frequency $f$ and the repair frequency $r$ correspond to the rate parameters of the adjusted exponential distributions. We note here that the mean up and down time durations are in the scales of days, meaning that on average incidents occur every few days and that on average restoration of the service may take almost an entire day. The later is because we are including weather effects as sources of disruptions. At the same time, since we will consider a half-mile section of the road, travel times will be on the scales of fractions of minutes and this means that many customers in the system will experience disrupted conditions. We stress here again that we are studying the effect of non-recurring disrupted conditions, and hence we are not including here time periods with recurrent congestion such as peak-hours.

From the data in figures 4.5(a) and 4.5(b), it can be shown that the fit of an exponential distribution on the smaller time scales without considering the longer
weather events also yields $R^2$ values above 0.75.

### 4.2.2 Calibration

In order to model traffic conditions the simplest parameters to obtain directly are $1/\mu$ and $1/\mu'$, which are the mean travel times for the normal and deteriorated traffic conditions, respectively. The required travel times are approximated from the speed data by “walking the speed field” of the specific road section [66], which yields sufficiently good travel time estimations for our statistical analysis. In order to obtain $1/\mu$, we use the mean travel time during up periods of the system and, similarly, $1/\mu'$ is the mean travel time for down periods of the system as described before.

Figures 4.6(a) and 4.6(b) show probability-scaled travel time histograms for up and down periods respectively in the half a mile test road section. The mean travel time is $1/\mu = 29.60 \text{ [sec]}$ during the up times and it is $1/\mu' = 32.67 \text{ [sec]}$ during the down times. As expected $1/\mu < 1/\mu'$, that is: the mean travel time increases under deteriorated traffic conditions. Since the length of the corridor is fixed, the mean travel times $1/\mu$ and $1/\mu'$ provide the mean travel speeds under normal and deteriorated traffic conditions, respectively. For the 0.5 mile corridor, this translates to speeds of $61.80 \text{ [mi/hr]}$ and $55.08 \text{ [mi/hr]}$, respectively.

Notice that these travel-time averages are usually readily available information even if travel-time data is not available.

The final step to complete the implementation of our analytical model is to select a distribution to the travel times, $S$, under the normal traffic conditions which correspond to our up-periods (see Figure 4.6(a)). We show here the case of selecting a triangular and shifted exponential parametric distributions, which yield closed form of the Laplace transforms, as shown in section 4.1.2. If non-parametric distributions are to be used they can be transformed numerically to the frequency domain in order
Figure 4.6: Probability-Scaled Travel Time Histograms and Mean Travel Time: (a) Up periods, (b) Down periods.

to be used in equations (3.19)-(3.20).

Figure 4.7 shows the probability-scaled histogram of the travel times for the up periods and adjusted triangular and shifted exponential distributions. The parameters for the triangular distribution in the case $a < c < b$ are $a = 21.67$ [sec], $c = 27.41$ [sec], and $b = 38.06$ [sec]. For the $a = c < b$ case the parameters are $a = c = 23.54$ [sec] and $b = 35.49$ [sec]. The shifted exponential distribution has parameters $\eta = 1/14.98$ [1/sec] and $c = 25.23$ [sec]. We also adjust a Log-Normal distribution as it is widely used to model travel time distributions [71].

The best fit in this case corresponds to the general triangular distribution with an $R^2$ value of 0.989. The Akaike information criterion (AIC), however, indicates that the triangular distribution with $c = a$ may be a good candidate too since it holds the minimum value of $\text{AIC} = 5.78 \times 10^5$, against $\text{AIC} = 6.55 \times 10^5$ for the triangular model, $\text{AIC} = 6.93 \times 10^5$ for the shifted exponential model, and $\text{AIC} = 7.37 \times 10^5$ for the Log-Normal model. Triangular and shifted exponential distributions are then appropriate candidates to select as travel-time distributions under normal traffic conditions. The information needed for their parameters, we emphasize, does not have to come from
Once properly defined, the travel time distribution in equations (3.19)-(3.20) can be inverted numerically to the time domain. In particular, De Hoog et al.’s quotient-difference algorithm for numerical Laplace inversion [19] is very effective in our case when implemented with accelerated convergence for the continued fraction expansion that is developed by Hollenbeck [37]. It is worth mentioning that no single method for Laplace transform inversion is guaranteed to give good results as this will depend greatly on the specific application [24]. Some alternatives are presented by Abate & Whitt who give two algorithms for numerical inversion of Laplace transforms of probability distributions [2].

The resulting travel time distribution can then be compared against the pooled empirical travel time data for validation. The pooled travel time for our test road section is depicted in the probability-scaled histogram in Figure 4.8. The travel time distribution resulting from inverting the Laplace transform in (3.19)-(3.20) while using the aforementioned parameters is also shown.
Figure 4.8: Pooled Travel Time Distribution and Analytical Travel Time

The resulting distribution fits very well to the empirical travel time distribution reproducing up to a certain extent the heavier right tail of the distribution. The resulting $R^2$ value is 0.923 in this case. The triangular model is not necessarily the best model for every sensor, in Section 4.3 we show the resulting best model for each sensor.

4.2.3 FREEVAL-RL

For comparison purposes we also calibrate and run standard FREEVAL-RL computational engine simulations [75]. Since the simulator can only handle periods of 6 hours at a time and within a single calendar year we run simulations for the periods 5am-7am, 9:30am-3:30pm, and 6pm-00am separately for the year 2008. For the Seed Demand Day [75], which is the date for which volume data will be provided as a basis, we select a date in the month of March 2008 without adverse weather conditions and no other incident reports. We have 3 (three) 4-lane HCM segments of 1/6-mile in length, with level terrain, no ramp metering, and jam density approximated to 190 vehicles per mile per lane from the data observed. The Capacity Drop in the Queue
Discharge Mode is set to 5%. For every lane the free flow speed is set to 65 miles per hour. The worksheets of seed demand, weather, and incidents are generated from our data to fit the format of the FREEVAL-RL inputs. We create 200 scenarios with demand, weather and incident characteristics in proportion to the observed data.

4.3 Results

Once the model is calibrated experiments can be carried out by adjusting the parameters allowing sensitivity studies, what-if scenarios, and long-term traffic deterioration studies. We start by showing some travel time reliability measures that be directly obtained from our model.

4.3.1 Reliability Measures

Figure 4.9(a) shows the time domain travel time distribution we obtained in section 4.2.1 (zoomed-in version of the distribution in Figure 4.8), Figure 4.9(b) shows the corresponding cumulative distribution function.

From here we can immediately obtain the mean travel time ($E[T]$) and the standard deviation ($\sigma_T$) as measures of travel time reliability. Percentiles such as the usual 95th or 90th percentiles can also be immediately calculated. Other reliability indices that can be readily obtained are defined below.

Buffer Index: fraction of time that customers should add to their average travel time to ensure on-time arrival at the 95% level, i.e.,

$$BI = \frac{95\text{th percentile travel time} - \text{average travel time}}{\text{average travel time}} \quad (4.7)$$

Planning Time Index: factor of free flow time that a traveler should consider to
(a) Analytical travel time distribution PDF

(b) Analytical travel time distribution CDF

Figure 4.9: Full Travel Time Distribution: (a) Probability Density Function, (b) Cumulative Distribution Function.

ensure on-time arrival, i.e.,

$$\text{PTI} = \frac{95\text{th percentile travel time}}{\text{free-flow travel time}} \quad (4.8)$$

The definitions of these and other indices, such as the Percentage of Trips on Time, can be found in the report by the Federal Highway Administration [73].

We also look at the tail of the distribution to find the percentage of travelers who experience abnormally high travel times. Let $F_T(t)$ denote the CDF of the travel time distribution, that is $F_T(t) = P\{T \leq t\} = P\{\text{travel time} \leq t\}$, then the tail of the distribution is $\bar{F}_T(t) := 1 - F_T(t) = P\{\text{travel time} > t\}$.

Table 4.1 shows some time-percentiles $t$ of the tail of the travel time distribution for some selected percentage levels. The second column (Data Percentile) shows the percentiles obtained from the reconstructed travel-time data, the third column (FREEVAL-RL Percentile) shows the corresponding values obtained from the scenarios run in the FREEVAL-RL computational engine [75], as mentioned in Section 4.2.3. After running the scenarios the travel time cumulative distribution function is sampled from the Reliability Analysis Summary Report, the relative error with
respect to the data percentiles is shown in percentages in the column at the right. Lastly, the fifth column (Model Percentile), shows the percentiles predicted by our model and again to the right the percentage relative error with respect to the data percentiles. For our model we mention also that the predicted mean travel time corresponds to 30.5 seconds with a standard deviation of 3.99 seconds.

\[
\bar{F}_T(t) = P\{T > t\} \%
\]

<table>
<thead>
<tr>
<th>Percentiles t [sec]</th>
<th>Data</th>
<th>FREEVAL-RL</th>
<th>error %</th>
<th>Model</th>
<th>error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>22.21</td>
<td>23.21</td>
<td>4.5</td>
<td>24.44</td>
<td>10.04</td>
</tr>
<tr>
<td>90</td>
<td>24.41</td>
<td>26.18</td>
<td>7.25</td>
<td>25.51</td>
<td>4.51</td>
</tr>
<tr>
<td>80</td>
<td>27.08</td>
<td>28.15</td>
<td>3.95</td>
<td>26.92</td>
<td>-0.59</td>
</tr>
<tr>
<td>70</td>
<td>27.5</td>
<td>31.56</td>
<td>14.76</td>
<td>28.01</td>
<td>1.85</td>
</tr>
<tr>
<td>60</td>
<td>30.12</td>
<td>32.4</td>
<td>7.57</td>
<td>29.07</td>
<td>-3.49</td>
</tr>
<tr>
<td>50</td>
<td>30.78</td>
<td>34.56</td>
<td>12.28</td>
<td>30.14</td>
<td>-2.08</td>
</tr>
<tr>
<td>40</td>
<td>30.98</td>
<td>35.79</td>
<td>15.52</td>
<td>31.23</td>
<td>0.81</td>
</tr>
<tr>
<td>30</td>
<td>31.6</td>
<td>37.8</td>
<td>19.62</td>
<td>32.44</td>
<td>2.66</td>
</tr>
<tr>
<td>20</td>
<td>32.83</td>
<td>39.78</td>
<td>21.17</td>
<td>33.91</td>
<td>3.29</td>
</tr>
<tr>
<td>10</td>
<td>35.03</td>
<td>40.5</td>
<td>15.61</td>
<td>35.92</td>
<td>2.54</td>
</tr>
<tr>
<td>5</td>
<td>36.31</td>
<td>42.13</td>
<td>16.03</td>
<td>37.56</td>
<td>3.44</td>
</tr>
</tbody>
</table>

Table 4.1: Percentiles for travel times distribution’s tail

We see immediately from Table 4.1 (fifth column, last row), that the 95th travel time percentile corresponds to 37.56 seconds, while the 90th percentile is 35.92 seconds. Also, 95% of the travelers will experience a travel time longer than 24.44 seconds, 50% of the travelers should expect to experience a travel time longer than 30.14 seconds, and that 5% of the drivers would take longer than 37.56 seconds to transverse the corridor. The average percentage difference from our model to the data
is 2.09% with the biggest percentage difference of 10.04% occurring for the fastest customers. As a general trend, our model tends to slightly overestimate the travel time. The FREEVAL-RL simulation, on the other hand, shows increasing differences with respect to the data percentiles when estimating the travel times of the slowest customers in the system (average difference 12.57%). Overall we observe that our model follows more closely the data percentiles than the FREEVAL-RL simulation does.

Figure 4.10 shows the graph of $F_T(t)$, which is the function from which any percentile or quantile of interest can be obtained.

Figure 4.10: Tail of the Cumulative Distribution Function: $F_T(t)$

BI is 0.2327 for this example, which means that a driver needs to allot a 23% buffer to the average travel time to ensure being on time with 95% confidence. Also, PTI has a value of 1.2687, which indicates that a driver that wants to be on time 95 percent of the times has to multiply the free-flow travel time by a factor of 1.3.
4.3.2 Scenario Analysis

We show now how the parameters of the model can be varied to obtain travel time reliability estimates for traffic conditions for which we do not have data. We start by showing the impact on the travel time of the severity of the traffic deterioration. To this end, we vary the $\alpha$ parameter, which is the ratio of expected travel time under deteriorated and normal traffic conditions. Equation (4.9) shows the calculation of $\alpha$ for our test example in section 4.2.1.

$$\alpha = \frac{1}{\mu} = \frac{\mu'}{\mu} = \frac{110.17}{121.60} = 0.8890.$$  \hfill (4.9)

We will see the effect of $\alpha$ dropping to 0.8, 0.7, 0.6, and 0.5. This is equivalent to keeping $\mu$ fixed, and having the expected speed under deteriorated conditions drop from its original value of $55.08 \text{ [mi/hr]}$ to $44.07 \text{ [mi/hr]}$, $38.56 \text{ [mi/hr]}$, $33.05 \text{ [mi/hr]}$, and $27.54 \text{ [mi/hr]}$, respectively.

In figure 4.11 we show how the travel time probability density function changes with the reduction of $\alpha$. We also plot the corresponding values of $1/\mu$ and $1/\mu'$. We observe that as the value of $\alpha$ drops the distribution becomes bimodal with some proportion of the travelers experiencing larger travel times. The distribution tends to become disconnected for very low values of $\alpha$. Notice additionally, that as the value of $\alpha$ decreases the tail of the distribution becomes thicker.

Figure 4.12 shows the tail cumulative distributions $\bar{F}_T(t)$ for the same cases presented in Figure 4.11, for values of $\alpha$ of 0.8, 0.7, 0.6, and 0.5. From here we see the expected proportion of travelers that will experience travel times larger than a specific amount of seconds. It is apparent that with decreasing values of $\alpha$ (which corresponds to increasingly deteriorating traffic conditions) the proportion of travelers that will experience long travel times is increasing.

Table 4.2 shows some of the quantiles of the tail distribution $\bar{F}_T(t)$. We also include the quantiles of the original calibrated model for which $\alpha = 0.8890$. A dash indicates
Figure 4.11: Travel Time Distribution PDF: (a) $\alpha = 0.8$, (b) $\alpha = 0.7$, (c) $\alpha = 0.6$, (d) $\alpha = 0.5$. 
Figure 4.12: Tail of the Cumulative Distribution Function, $\bar{F}_T(t)$: Varying $\alpha$

percentages lower than 0.113%. Table 4.3 shows the reliability indices calculated for each one of the experiments.

<table>
<thead>
<tr>
<th>$t$ [sec]</th>
<th>$1/\mu = 29.60$</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.8890$</td>
<td>55.03</td>
<td>1.42</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\alpha = 0.8$</td>
<td>62.35</td>
<td>8.34</td>
<td>0.12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\bar{F}_T(t, \alpha)$ [%]</td>
<td>$\alpha = 0.7$</td>
<td>64.03</td>
<td>22.75</td>
<td>2.23</td>
<td>0.12</td>
</tr>
<tr>
<td>$\alpha = 0.6$</td>
<td>63.98</td>
<td>36.77</td>
<td>14.55</td>
<td>0.98</td>
<td>-</td>
</tr>
<tr>
<td>$\alpha = 0.5$</td>
<td>64.04</td>
<td>39.06</td>
<td>34.46</td>
<td>14.60</td>
<td>2.16</td>
</tr>
</tbody>
</table>

Table 4.2: Quantiles for Travel Time Distribution’s Tail: Varying $\alpha$

Observe that always more than 50% of the travelers are expected to have travel times longer than the mean travel time during up periods $1/\mu$, with the proportion being higher than 60% for $\alpha \leq 0.8$. For $\alpha = 0.5$ more than 14% of the travelers are expected to take more than $2/\mu$ seconds, and at least 2.16% are expected to take longer than 70 seconds to complete the trip. Notice also from Table 4.3 that,
as expected, all un-reliability indices are increasing with decreasing values of the parameter $\alpha$, and that for the case with $\alpha = 0.5$ BI indicates that a 65% buffer needs to be added to the mean travel time to ensure on-time arrival at the 95% level of confidence. Also, PTI indicates that the driver should expect to travel more than double the distance at free flow speed to ensure on time arrival at the same confidence level.

Now we want to show the effect that an increment in the frequency of incidents can have in the travel time distribution. Hence, we change the parameters of our experiment to achieve similar scales for the travel time during up times ($1/\mu$) and the incident duration and the repair times. Consider a roadway corridor of 30 miles, which is 60 times longer than our previous example. If we use the speeds as before, then the time parameters ($1/\mu$, $1/\mu'$, $a$, $b$, and $c$) are scaled by a factor of sixty and are now read in minutes. For example $1/\mu$ is now 29.60 [min] and $1/\mu' = 32.67$ [min] (notice that the $\alpha$ parameter remains the same under this scaling). We select now a mean down time duration $\bar{D} = 1/r$ of 30 minutes, which is in the same order of magnitude of the travel times. From here we calculate the travel time distribution for three different values of the up time duration (or, reciprocally, failure frequency): $1/f = 240$ [min], $1/f = 120$ [min], and $1/f = 30$ [min]. Note that in the most extreme

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\mathbb{E}[T]$ [sec]</th>
<th>$\sigma_T$ [sec]</th>
<th>95th [sec]</th>
<th>BI</th>
<th>PTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8890</td>
<td>30.47</td>
<td>4.00</td>
<td>37.56</td>
<td>0.23</td>
<td>1.27</td>
</tr>
<tr>
<td>0.8</td>
<td>31.85</td>
<td>5.27</td>
<td>41.74</td>
<td>0.31</td>
<td>1.41</td>
</tr>
<tr>
<td>0.7</td>
<td>33.88</td>
<td>7.38</td>
<td>47.70</td>
<td>0.41</td>
<td>1.61</td>
</tr>
<tr>
<td>0.6</td>
<td>36.60</td>
<td>10.45</td>
<td>55.58</td>
<td>0.52</td>
<td>1.88</td>
</tr>
<tr>
<td>0.5</td>
<td>40.37</td>
<td>15.06</td>
<td>66.73</td>
<td>0.65</td>
<td>2.25</td>
</tr>
</tbody>
</table>

Table 4.3: Reliability Indices for Travel Time Distribution: Varying $\alpha$
case the failures occur at the same rate as the repairs, which is one expected failure and one expected repair every 30 minutes.

Figure 4.13 shows the travel time density functions for these three cases and the corresponding tail distributions.

![Travel Time PDF](a) Travel time PDF  
![Tail of CDF: $\bar{F}_T(t)$](b) Tail of CDF: $\bar{F}_T(t)$

Figure 4.13: Effect of Changes in Failure-Frequency: (a) Analytical Travel Time PDF, (b) Tail Distribution: $\bar{F}_T(t)$

In Figure 4.13(a) we first observe that the distributions have thick tails. This is because the scale of the frequency of incidents is similar to the scale of the travel times, which allows the effect of the incidents on the quality of traffic to be more drastic. Secondly, we observe that the more frequent the incidents the bigger the part of the mass of the distribution that goes to the tail, indicating that such systems become more and more unstable with higher probabilities of abnormally large travel times.

As expected, from Figure 4.13(b) we observe thicker tails for the more incident prone scenarios. Table 4.4 shows some quantiles of interest of the tail distribution $\bar{F}_T(t)$ for all three cases, and Table 4.5 shows the reliability indices. The quantiles roughly correspond to multiples of the expected travel time under normal conditions $1/\mu$. Again, a dash indicates a percentage lower than 0.03%.
From the quantiles in Table 4.4, we first observe that in every case more than 50% of the travelers are expected to take more time to complete the trip than what it would take if normal traffic conditions alone were considered. Now, in the second case when the repairs happen four times more frequently than the failures ($4 \cdot 1/f = 1/r$, which is the same as $r = 4f$), there is about a 8.5% probability of taking a travel time larger than $2/\mu [\text{min}]$, which is double the expected time under normal conditions. Compare this to the 29% chance of taking more than double the normal travel time in the third case where $1/f = 1/r$. Lastly in the third case: more than 11% of the travelers should expect to take more than $3/\mu [\text{min}]$ to complete the trip, 4.38% will take more than two hours which is close to $4/\mu [\text{min}]$, and 1.71% should expect to take more than five times the travel time under normal conditions.

From Table 4.5 we can observe that the un-reliability indices are again increasing, which is consistent with the decrease in quality of service. Standard deviation grows
rapidly, and the 95th percentile moves drastically to the right. For the last case, which is the worst system, BI gives a value larger than one. This indicates that the mean travel time is smaller than half the 95th percentile. Also, the value of PTI is almost 4, which means that a driver that expects to get on time at the 95% confidence level should picture a trip 4 times longer than the original one (that is a 120-mile trip) to be traveled at free flow speed.

4.3.3 Across Sensor Model Validation

For each sensor, out of the choices for the uncongested travel time distribution $f_S(t)$, the most appropriate model varies greatly. Out of the 35 sensors there were two for which no model was able to fit the pooled travel time data with an $R^2$ higher than 0.55.

Tables 4.6 and 4.7 show the selected model for each sensor along with the $R^2$ value for the full travel time calculated model for directions SE and WN, respectively. We add also the fit of a Log-Normal distribution to the pooled travel time empirical distribution. We obtain relatively good results with averages of 0.86 and 0.84 $R^2$ for the SE and WN directions, respectively. On average, our models explain about 5% more of the variation in the data than the Log-Normal model. We also show here the Akaike information criterion calculated for each model and for the Log-Normal model. This information measure is calculated as shown in equation (4.10), where $k$ is the number of parameters in the mode and $L$ is the likelihood of the model given the calibration data.

$$AIC = 2k - 2\ln (L).$$

This criterion helps identifying higher quality models by relative comparisons, and it favors more parsimonious models. The smallest values of the AIC indicate that the quality of the model may be higher. We see in Tables 4.6 and 4.7 that our models
tend to have an AIC value which is in general at least one order of magnitude below the Log-Normal model. The high AIC values are due to the large amount of data points available for the calibration (in the order of two hundred fifty thousand).

<table>
<thead>
<tr>
<th>$f_T(t)$</th>
<th>$R^2$</th>
<th>Log-Normal $R^2$</th>
<th>AIC $f_T(t)$</th>
<th>AIC Log-Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.862</td>
<td>0.821</td>
<td>$3.63 \times 10^5$</td>
<td>$1.49 \times 10^6$</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6: Average Sensor Model Validation Direction SE

<table>
<thead>
<tr>
<th>$f_T(t)$</th>
<th>$R^2$</th>
<th>Log-Normal $R^2$</th>
<th>AIC $f_T(t)$</th>
<th>AIC Log-Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.848</td>
<td>0.786</td>
<td>$6.18 \times 10^5$</td>
<td>$1.56 \times 10^6$</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.7: Average Sensor Model Validation Direction WN

It is important to mention that there are other distributions besides Log-Normal that provide good fits to the empirical travel time distribution, this has been shown in numerous articles including the work by Rakha et al. [71]. Table 4.8 shows the across-sensor average $R^2$ for models besides Log-Normal, and also average AIC values.

<table>
<thead>
<tr>
<th>Model</th>
<th>Average $R^2$</th>
<th>Average AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>0.739</td>
<td>$1.56 \times 10^6$</td>
</tr>
<tr>
<td>Log-Normal</td>
<td>0.788</td>
<td>$1.53 \times 10^6$</td>
</tr>
<tr>
<td>Generalized Extreme Value</td>
<td>0.860</td>
<td>$1.50 \times 10^6$</td>
</tr>
<tr>
<td>Log-Logistic</td>
<td>0.865</td>
<td>$1.49 \times 10^6$</td>
</tr>
<tr>
<td>t-Location Scale</td>
<td>0.907</td>
<td>$1.51 \times 10^6$</td>
</tr>
</tbody>
</table>

Table 4.8: Average Validation for Other Models

t-Location Scale distribution provides consistently the best fits to the empirical
travel time data, however once the distribution is adjusted there it is not possible to calculate the travel time reliability for scenarios not supported by the data. Log-Logistic shows the smallest AIC values on average, although the differences in AIC values are not pronounced.

### 4.3.4 Model Specialization

If the adequate data is available, then the model can be specialized to specific months of the year, days of the week or times of the day. We show here how the model changes when using only the data for winter, and separating the days of the week into Weekends, Monday, Midweek (Tuesday through Thursday), and Fridays. We work with the same sensor as in the previous sections. The seasonal constraint yields new values of $1/r = 29.05$ [hour] and $1/f = 24.52$ [hour].

Figure 4.14 shows the resulting calculated full travel time distributions $f_T(t)$ for each day of week category, and their respective fits. Notice that for weekend and monday the best model was the general triangle for the non-congested travel time distribution choice, whereas for midweek and fridays the triangle with $a = c < b$ was the best model.

Table 4.9 shows a summary of relevant parameters and reliability metrics for the aforementioned four models, and the $R^2$ for a calibrated Log-Normal model.

<table>
<thead>
<tr>
<th>Model</th>
<th>$1/\mu$ [sec]</th>
<th>$\alpha$</th>
<th>$E[T]$ [sec]</th>
<th>$\sigma_T$ [sec]</th>
<th>$f_T(t)$</th>
<th>$R^2$</th>
<th>Log-Normal $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekend</td>
<td>Triangle</td>
<td>28.15</td>
<td>0.8337</td>
<td>30.21</td>
<td>4.41</td>
<td>0.9289</td>
<td>0.8970</td>
</tr>
<tr>
<td>Monday</td>
<td>Triangle</td>
<td>30.31</td>
<td>0.9048</td>
<td>31.25</td>
<td>4.10</td>
<td>0.8096</td>
<td>0.8985</td>
</tr>
<tr>
<td>Midweek</td>
<td>Triangle $c = a$</td>
<td>30.10</td>
<td>0.9013</td>
<td>31.21</td>
<td>5.80</td>
<td>0.9588</td>
<td>0.8976</td>
</tr>
<tr>
<td>Friday</td>
<td>Triangle $c = a$</td>
<td>29.37</td>
<td>0.8315</td>
<td>30.12</td>
<td>4.44</td>
<td>0.8814</td>
<td>0.8956</td>
</tr>
</tbody>
</table>

Table 4.9: Specializing Model by Day of the Week
Figure 4.14: Travel Time PDF: (a) Weekend, (b) Monday, (c) Midweek, (d) Friday.
We observe that our model predicts higher average travel-times for Mondays and Midweeks (31.25 [sec] and 31.21 [sec] against 30.21 [sec] and 30.12 [sec]), and that the Midweek model shows the highest deviation ($\sigma_T = 5.8$ [sec]). The Monday and Friday models have worst fits according to the $R^2$ values, and for the Monday model the Log-Normal model is significantly better explaining about 9% more of the variation in the data.

### 4.4 Conclusions and Future Work

We have presented an analytical modeling tool and calibration strategy to approximate the travel time distribution of a road segment that is subject to random degradations of service. By focusing on off-peak hours of the day we have effectively developed a model that incorporates the stochasticity of traffic degradation due to incidents and weather events. The analytical approach allows us to obtain the full travel time distribution, from which a variety of un-reliability indices can be readily calculated.

By comparing with travel time data, and incident and weather reports we have observed that the assumptions of our analytical model are reasonable, and that its application to a travel time reliability study is justified. Moreover, the resulting travel time distributions approximate very well the empirical data, while requiring only some percentiles and average values for calibration.

Additionally, the travel time reliability measures obtained are in line with the FREEVAL-RL computation engine simulations without the need of the extensive manual calibration or simulation running times. Standard un-reliability indices can be calculated directly from the full analytical travel time distribution. Our methodology also requires minimal statistical information for calibration and posterior scenario analysis of road segments.
As future work we consider numerical implementations that allow the use of service time requirement distributions \(f_S(t)\) that do not have a closed form Laplace transform (e.g. Log-Normal distribution). For this we need to be able to reliably and efficiently approximate the Laplace transform for truncated versions of the distributions.

Additionally, our ability to calculate full travel time distributions opens an opportunity to better inform the estimation of the value of travel time reliability. From current literature in the subject [15, 64], it is common to see that the calculations of the value of travel time reliability is separated in two main components: (1) the selection of the travel time reliability metrics that will be considered in the valuation (2) the calibration of cost estimates according to empirical evidence on user preferences. Since we are able to calculate the entire distribution of the travel time for any parameter describing the vulnerability of the corridors we should be able to study the effect of incidents in the user preferences for travel time reliability according to the expected utility. For example we could study the effect of increasing the frequency of incidents on the user preferences, or how the level of deterioration of the road \(\alpha\) affects the utility that a user perceives.
Chapter 5

Route Optimization: Risk Averse Shortest Path

In this Chapter we outline a second application of the analytical queueing system results to vehicle traffic modeling. While the application in Chapter 4 is aimed at traffic planing and long-term reliability considerations, here the focus is on the operational side as we intend to inform a driver on the shortest travel-time routing choices given a particular risk-aversion level. We make use of the completion-time distributions obtained in Chapter 3 and discussed in Chapter 4 for informing the travel-time distribution. The addition of risk-aversion is motivated by the sensitivity in the routing of emergency, priority, and first response services, where the modeling of different levels of risk aversion and/or stochastic dominance requirements could be of particular interest.

We present two different modeling approaches. The first approach is a pure dynamic programming model in which we use our traffic-calibrated travel time distributions, and risk aversion is considered \textit{a posteriori}. This approach, while analytically simple, is difficult to treat computationally due to the \textit{curse of dimensionality}. A
second and more technically challenging approach is through the use of stochastic programming. This approach is analytically richer and as it will be seen is naturally suited to consider risk-aversion directly in the formulation by means of measures of risk. We want to study the advantages or disadvantages of using stochastic programming as the modeling framework, considering: flexibility in the modeling, computational tractability, and the analytical traceability of the results. Analytical convergence analysis of the solution methodology for both techniques will be included, as well as other analytical insights on the tailored numerical methodologies or heuristics. The two models are expected to yield directionally similar insights on the effects of risk aversion to the shortest-travel-time route choice problem.

The chapter is structured as follows: Section 5.1 gives further details on the SPP and more general related problems. Section 5.2 introduces two particular modeling frameworks of our interest including dynamic programming in subsection 5.2.1 and stochastic programming in subsection 5.2.2. Finally, Section 5.4 gives some details on the modeling approach we are proposing for each of the frameworks.

### 5.1 The Shortest Path Problem

We provide a classical formulation for the SPP and describe how it has been extended in the literature to include dynamic and stochastic settings, as well as the effects of stochastic disruptions. Many of these formulations are motivated by transportation applications where a planner, manager, and/or ultimately a driver needs to minimize the total or expected travel time of a trip. We mention that since the direction of travel is usually restricted in transportation networks, all versions of the SPP treated here are specified for directed graphs.

We start by describing the underlying network. Let $\mathcal{N}$ denote a finite set of nodes
(destinations and/or road intersections) where routing decisions have to be made, and \( A \subseteq \mathcal{N} \times \mathcal{N} \) denote a set of directed arcs between the nodes. When modeling a transportation network, the set \( A \) includes the arc \( a := (n, n') \) if and only if there exists a road that permits vehicles to travel directly from node \( n \) to node \( n' \) (in that direction). Of course, given any arc in \( A \) the arc to travel in the opposite direction could also exist, but this is not necessary nor guaranteed. The directed graph \( G(\mathcal{N}, A) \) defined by these two sets represents the road structure, and it is assumed to be a weakly connected graph. The network is provided with a deterministic real valued weighting function \( l : A \rightarrow \mathbb{R} \) which measures a quantity of interest for every arc (usually distance, travel time, or monetary cost). Two of the nodes in the network are special and are labeled as the origin and destination nodes.

The SPP is then to identify subsets of arcs in \( A \) which conform \( l \)-minimal paths from the origin to the destination node, where the aggregation of the weights of each arc on the path is usually done additively. It is clear that such \( l \)-minimizing (henceforth minimizing) paths are not necessarily unique.

Depending on how the weighting function is specified, different problems arise and different strategies can be used to solve the SPP. For settings where \( l \) produces no negative cycles (but is otherwise general) an efficient classical solution is obtained by the Bellman-Ford algorithm [12], which is a progressive relaxation technique that relies on improving total weight upper bounds. For the non-negative case, \( l : A \rightarrow \mathbb{R}_+ \), the greedier Dijkstra’s algorithm is more effective [21]. Both algorithms actually solve a more general problem which is the \( l \)-shortest path tree that includes paths to every node in the network. Finally, if for every arc \( a \) in the network the weight function returns the constant \( l(a) = 1 \), then the problem is called the shortest spanning tree, and the total weight of each path corresponds to the number of arcs that compose the path. For a more detailed account on the classical SPP the interested reader is
referred to the book by Korte and Vygen [47].

A more general version of the SPP considers time-dependent and/or random weight functions. If the time dependency is deterministic the problem is usually called the Dynamic Shortest Path Problem (DSPP). If, on the other hand, part of the problem is randomized the problem is called Stochastic Shortest Path Problem (SSPP). These settings are particularly appropriate for modeling transportation related activities such as vehicle routing, fleet management, delivery services, etc., which are inherently time-sensitive. In this context, the weight function usually measures the travel time of the arcs (road sections) and the objective is to find shortest travel time paths.

It is clear that for these non-static models the optimal solutions cannot consist of simple paths anymore, but rather have to be expressed as optimal routing policies which dictate optimal decisions to make under time-varying circumstances. Some instances of the DSPP are still solvable using some classic SPP algorithms [93], but this is not generally the case partly because these algorithms rely on an optimality principle which no longer holds for the DSPP and SSPP (SPP-optimal paths are composed of optimal sub-paths). Moreover, while it is perfectly clear for the travel time SPP that waiting at a node or making a backward movement is sub-optimal, this is no longer the case for general DSSPs and SSPPs. Indeed, these last two possibilities (or conversely restrictions) have to be considered explicitly in the model. For models that allow waiting at a node the reader is referred to the work by Azaron and Kianfar [6], for backward movements see the work by Sever et al. [78].

As mentioned in the literature review, purely dynamic settings do not take into account the rather frequent unexpected changes in traffic conditions that vehicles experience, and the literature has mostly turned its attention to mixed-stochastic approaches for the SPP and related problems [29, 57], giving rise to the Dynamic
and Stochastic Shortest Path Problem (DSSPP). The dynamic and stochastic components are used to encode the effects of (but not exclusively), the possibility of having to change a decision [13], recurrent congestion [29, 52], non-recurrent congestion [89, 35], and the availability of real time information from intelligent transportation systems (ITS) [46]. Discrete-time Markov chains are a common technique to model the dynamic-stochastic processes, although continuous-time random variables and queueing theory are also used. Additionally, strong independency assumptions often have to be made. For random travel times, discrete distributions are most prevalent to ensure computational tractability.

It has long been observed that taking expected values of the random variables in the problem data to reduce a DSSPP into a DSPP leads to overly optimistic paths [54]. So, some researchers use expectations and variances are used in the objective functions to aggregate the random effects, while others employ utility functions to establish preference among the random variables.

5.2 Modeling Frameworks

Here we describe the dynamic programming and the stochastic programming frameworks that we will use to model the problem. These are particular cases of a Dynamic and Stochastic Shortest Path Problem (DSSPP).

5.2.1 Dynamic Programming

Among the DSSPPs we consider a model that explicitly accounts for randomly occurring disruptions of the network. These disruptions or incidents cause arc capacity reductions or, reciprocally, travel time increases in the form of delays, and are included to represent the effects of non-recurrent congestion and unexpected deteriorated traffic conditions in general.
Models with random incidents usually consider a (relatively small) number of congested states for the arcs in the network, which transition according to some type of discrete or continuous-time Markov Chain [27, 6, 78]. Markov Decision Processes (MDP) and Dynamic Programming (DP) are hence usual choices for modeling and solving the problem, respectively [35, 79]. An arc is considered to be in recovered or normal state when the governing Markov Chain arrives to an appropriately labeled state in which the capacity is at its fullest and/or the travel time delay is zero. Incidents can be restricted to a subset of vulnerable arcs [79] or they can affect all arcs in the network [27], but in any case each vulnerable arc has independent incidents and they do not propagate explicitly [35]. While some random incidents are considered to be fully independent of the travel time [6], some recent studies have considered a more realistic assumption in which the disruptions are in turn dependent on the current travel time conditions of the arcs [78, 79].

An important choice in the modeling is the level of incident awareness that the navigating agent will be given. Assuming that some form of ITS data is available in real time, some models limit the incident vision to: only the adjacent nodes [6], some pre established but not all nodes in the network [35, 79], and all nodes in the network [6].

While some of these models can be solved using classic DP techniques, as in the work by Sever et al. [78] where the MDP is solved by a backward recursion algorithm, most models are too complex to obtain optimal solutions even for small problem instances. Hence, the use of heuristics and approximations has become an important research field which has called for a lot of attention. Fu et al. (2006) compiled a survey on heuristic algorithms for transportation applications of the shortest path problem [30]. For approximate DP applications we can cite the works by Zhang et al. [92]
and Sever et al. [79], and for a proposition of a unifying framework for the modeling of transportation and logistics applications via approximate dynamic programming (ADP) the reader is referred to the article by Powell et al. [70].

5.2.2 Stochastic Programming

We want to examine a Multi-Stage Risk Averse Stochastic Programming setting of the Dynamical Shortest Path Problem (DSPP) with both stochastic capacity disruptions and stochastic constraints to the routing of a single vehicle from an origin to a destination node on a transportation network. A multi-stage stochastic programming setting is believed to be appropriate to model the dynamical routing of a vehicle through the network, while also providing a natural context to adopt risk measures and non-deterministic constraints.

Stochastic programming is a general framework for optimization problems involving random variables and/or random parameters. A distinguishing feature is that the random nature of the underlying process being modeled is explicitly considered in the formulation without resorting to \textit{a-priori} reductions such as averages or best/worst-case scenarios. As a modeling framework, and due to the myriad of random effects and probabilistic considerations encountered in applications, it is different from other programming fields such as nonlinear-optimization, where general and somewhat unified theories have been developed for broad classes of problems. On the contrary, each family of stochastic programs usually requires the elaboration of some tailored mathematical sub-theory and the development of ad hoc solution methodologies. Stochastic programs are in general hard to solve and computational feasibility is a constant worry.
Classical applications in the field are portfolio management problems, inventory problems, and resource allocation problems. On the other hand, among the stochastic programming settings that have been rigorously studied (and that are of our interest) we count chance constrained problems, two- and multi-stage problems, and stochastic dominance constrained problems [42].

For an extensive theoretical-oriented reference on stochastic programming we give the book by Shapiro et al. [82]. This reference is particularly relevant to our study, and will be extensively used in the remainder of the Chapter.

We follow this section with three relevant stochastic programming topics that will help to situate our work: multi-stage stochastic programming (section 5.2.2), the theory of coherent measures of risk (section 5.2.2), and advances on mixed-integer stochastic programming (section 5.2.2). We will discuss how stochastic programming is a natural environment for modeling some aspects of the DSSPP and show some advances found in the literature as applied to this and similar problems.

Two-Stage and Multi-Stage Stochastic Programming

We present first a simple linear two-stage stochastic program written in terms of cost. Consider the problem given by equations (5.1)–(5.2).

\[
\begin{align*}
\min _{x \in \mathbb{R}^n} & \quad c^T x + \mathbb{E}[Q(x, \xi)] \\ 
\text{s.t. } & \quad Ax = b, \quad x \geq 0,
\end{align*}
\]

(5.1)

where \(Q(x, \xi)\) is the optimal value of the problem

\[
\begin{align*}
\min _{y \in \mathbb{R}^m} & \quad q^T y \\ 
\text{s.t. } & \quad T x + W y = h, \quad y \geq 0.
\end{align*}
\]

(5.2)

Here \(\xi := (q, h, T, W)\) could be entirely or partly random to model any part of the problem with random coefficients. The optimization in (5.1) is called first-stage prob-
lem with decision variables $x$, data $A$ and $b$, and cost vector $c$. The second-stage problem (5.2), has decision variables $y$ and depends both on the decision variables of the first-stage problem $x$ and random data $\xi$. The expectation in the objective function of (5.1) is placed to account for the randomization of $\xi$.

The rationale behind this type of modeling comes from planning situations in which decisions have to be made (here deciding $x$ and $y$) but only $x$ can be decided a priori, as the choice of an optimal $y$ will depend on unobserved random effects. Once the randomness $\xi$ has been observed $y$ needs to be decided. One example is pre-resource allocation for later shipment: the first-stage decision is how to best pre-allocate resources in depots placed on a number of locations to be shipped later. At a posterior time, demand levels for different depots are informed and how much to ship from each depot has to be decided; this is the second-stage decision. The random demands are unknown at the time when the first-stage decision is made, but are known when the second-stage decision is made. In general this approach is well suited for problems with a planning stage and a posterior operational stage. The decision process is schematized in (5.3).

$$\text{decision}(x) \rightsquigarrow \text{observation}(\xi) \rightsquigarrow \text{decision}(y) \quad (5.3)$$

One important technical property of the problem is the level of recourse, which has to do with the feasibility of the second-stage problem. The problem is said to have complete recourse if the system $Wy = h - Tx$, $y \geq 0$ has a solution for every instance of the right-hand side vector $h - Tx$. Also, the problem is said to have relatively complete recourse if for every feasible first-stage decision variable (this is $x \geq 0$ such that $Ax = b$) the feasible set of the second-stage problem is non-empty for almost everywhere $\omega \in \Omega$, with $\Omega$ being the sample space of the probability space where $\xi$ is defined. There are other levels of recourse which will not be discussed in the current version of this paper (the interested reader can consult [82]).
The level of recourse has implications on the possibility of solving the two-stage problem via some classical methodologies based on duality theory. In general, the more complete the recourse, the easiest is to solve the problem. In particular, however, complete and relatively complete recourse are (quite) non-trivial properties since they clearly do not hold true for general random data \( \xi = (q, h, T, W) \). When they do hold, the matrices \( T \) and \( W \) will have very special properties such as being deterministic, invertible, with sparse identity-like structures, etc.

An application of a two-stage stochastic problem to a capacitated routing problem can be found in the work of Ferris and Ruszczyński [27], where the first-stage decision variables correspond to marginal flows that have to be assigned to the network under normal conditions, and the second-stage variables are marginal flows to be assigned once a change to one of many possible (random) failure states is observed.

Liu et al. [53] model a network infrastructure retrofit-resource allocation problem as a two-stage stochastic problem for minimizing cost and travel time and improving infrastructure robustness. The first stage is how to decide where to allocate the resources under a hard budget, and the second stage comprises the evaluation of monetary loss once a destructive emergency has occurred, as well as the flow assignment to the network to minimize travel time delay costs.

Shen et al. [84] apply a two-stage model to a vehicle routing planning problem for resource delivery under large-scale emergencies. The first stage corresponds to a planning phase where emergency delivery routes are designed. The second stage, which is operational and models the time after the emergency has occurred, deals with deciding optimal deliveries of the resources, while making adjustments to the routes set in the first stage. This work, also gives a good example of an explicitly \textit{chance-constrained} model, in which the probability of satisfying the flow constraints
is constrained to be above a certain threshold.

The idea behind two-stage modeling can be extended to *multi-stage stochastic programming*, where there is now a finite number \( S \) of decision stages which have to be planned while random information is progressively revealed. This is schematized in (5.4).

\[
\text{decision}(x_1) \rightsquigarrow \text{observation}(\xi_2) \rightsquigarrow \text{decision}(x_2) \rightsquigarrow \cdots \rightsquigarrow \text{observation}(\xi_S) \rightsquigarrow \text{decision}(x_S).
\]

(5.4)

Here, \( \xi_i \) for \( i = 2, \ldots, S \) is in general random data which is unknown at the time decisions \( x_1, \ldots, x_{i-1} \) have to be made, but that is available (observed) at the time the posterior decisions \( x_i, \ldots, x_S \), respectively, are made. The linear version can be written in analogously as in (5.1)--(5.2), or also as a single stochastic program.

Multi-stage modeling leads to complex nested formulations that mandate the use of relatively small number of stages. Efficient solution techniques are restricted to cases in which the random data follows discrete distributions and the dynamic random process can be decomposed into *scenario trees*.

One application of multi-stage stochastic programming to a transportation design can be found in the work by Ukkusuri and Patil [91]. The multiple stages represent investment decision that are made progressively to add capacity to the network as the demand is observed in time.

**Coherent Measures of Risk**

The variability introduced by random effects is very important in situations where the law of large numbers does not apply, namely when the experiment or situation modeled can only be replicated a small number of times or even only once (such as with high investment decisions). In this cases using expected values to quantify the
aggregate costs of the model could not be appropriate. In this sense, the concept of risk measure was introduced to quantify deviations from the expected values in random situations. A risk measure is a mapping from a random variable to the real numbers, where the random variable represents some kind of cost of making and investment (or reward depending on the setting). In a successful effort to establish a set of desirable properties for risk measures, Artzner et al. [4] provided in 1999 a set of four axioms that define what they call coherent measures of risk. The axioms establish that to be coherent, a measure of risk should be convex and monotonic, and it should have a translation property and a positive homogeneity property. The convexity establishes that risk does not increase under diversification of investments, and the monotonicity means that if a random outcome of an investment is always (almost surely) better than another, then the risk measure should prefer (in the appropriate economical sense) the first investment. The translation property makes shifts on the random variable’s outcomes to be reflected as the same shift of the risk-measure value, and finally the positive homogeneity property ensures that a change in scales (such as a currency conversion or unit change) is properly and proportionally reflected in the risk measure. Additionally, a risk measure is called law invariant if, roughly, its value depends only on the distribution of the random variable and not on the random variable itself [82].

Some well known examples of coherent law-invariant risk measures are the expected value (which is risk neutral), mean upper semi-deviation, and average value-at-risk. Also, any convex combination of coherent measures of risk is in turn a coherent measure of risk.

Models of stochastic programming are well suited to be combined with coherent risk measures. And risk-averse versions of two-stage and multi-stage stochastic
problems can be constructed as shown in the works by Miller and Ruszczyński for two-stage problems [56] and Philpott and de Matos on dynamic sampling algorithms for multi-stage stochastic programming with risk aversion [69]. In multi-stage or dynamical (time-dependent) scenarios a property called time consistency is usually included in the requirements of coherent risk measures. This property is similar to the optimality principle of the static SPP mentioned in section 5.1, and establishes that “...a decision maker needs to be constantly concerned about optimizing his or her decisions for the remaining portion of the time horizon. That is, current optimal decisions must look to the future, rather than the past” [14]. The reader is also referred to [81] for a treatment of multi-stage time consistent law invariant risk measures.

A fairly recent work by Shapiro [80] deals with the relations of risk averse multi-stage stochastic programming and other multi-stage stochastic programming settings, namely the minimax and nested formulations.

Mixed-Integer Stochastic Programming

When DP equations can be written for the DSSPP, the combinatorial aspects of the problem can be absorbed by the state and action space definitions, and the use of integer programming can be avoided. When this is not possible, some integer or binary variables have to be explicitly used in the formulation (for example to model the inclusion or not of a link in a path). Rather than developing specialized mathematical theories for mixed-integer stochastic programming, the mixed-integer stochastic programming research community generally draws from integer programming (IP) theory to adapt/generalize the stochastic programming to non-continuous settings [77, 82]. As expected, the addition of this particular new layer of mathematical theory complicates greatly the already hard stochastic programs, and the possibility of finding efficient optimization algorithms depends heavily on the particular structural
characteristics of each problem.

Stougie and Vlerk [87] give an early general literature review on stochastic integer programming, including general theory for the subject, some of the classical solving algorithms, such as decomposition and convexification techniques, and some approaches to specific problems. Around the same time, Schultz et al. gave a survey on mixed integer two-stage problems [76]. An important conclusion is that heuristics can play a big role on the solution of this type of problems when they are asymptotically optimal; this, because optimization algorithms are only practical when (1) only the first-stage of the problem is mixed-integer or (2) when the second-stage is simple enough that it can be evaluated with relative ease.

Ahmed et al. [3] use a mixed-integer multi-stage setting for a general capacity expansion problem. The successive stages correspond to random information revelation epochs and the decisions are progressive capacity expansion investments. The random parameters vary according to finite discrete distributions, which allows decomposition into scenario trees. They develop specialized solution strategies, such as tight linear programming (LP) relaxations and specialized lower bounds, that take extensive advantage of the particular problem structure.

In 2005, Sen [77] reviewed in a book chapter a series of algorithms for stochastic mixed-integer programming (MIP) models. The survey begins by covering some general principles on decomposition methods, mixed-integer duality theory, and disjunctive programming. Then, the attention is turned to many decomposition techniques for mixed-integer two-stage problems depending on the first and second stage characteristics (binary first-stage, arbitrary second-stage, binary second-stage, MIP second-stage, etc.). The effort is to obtain convex value functions for the second-stage IP. The survey ends with scenario decomposition techniques for multi-stage mixed-integer problems.
5.3 Implementations

The underlying traffic network is modeled as a finite, directed, and weakly connected graph \(G(\mathcal{N}, \mathcal{A}_o)\), where \(\mathcal{N}, \mathcal{N} := |\mathcal{N}|\), denotes a finite set of nodes (destinations and/or road intersections) where routing decisions have to be made, and \(\mathcal{A}_o \subseteq \mathcal{N} \times \mathcal{N}\) denotes a set of directed arcs between the nodes. The set \(\mathcal{A}_o\) includes the arc \(a := (n, n')\) if and only if there exists a road that permits vehicles to travel directly from node \(n\) to node \(n'\) (in that direction). Of course, given any arc in \(\mathcal{A}_o\) the arc to travel in the opposite direction could also exist, but this is not necessary nor guaranteed. We distinguish two nodes that we call the origin (O) node, \(n_O\), and the destination (D) node, \(n_D\). We assume there exists at least one O-D path in the directed network, but we do not need to assume that the graph is connected nor strongly connected. Figure 5.1 shows an example of such an underlying graph with \(N = 6\), \(n_O = n_1\), and \(n_D = n_6\). There are 3 different O-D paths in this example.

For a node \(n \in \mathcal{N}\) we denote the set of direct successors or the out-neighborhood of \(n\) by \(\mathcal{N}^+(n)\). This is the set of nodes reachable in one step from \(n\). The size of this set is called the out-degree of \(n\) and it is denoted \(\deg^+(n) := |\mathcal{N}^+(n)|\). Denote by \(\mathcal{A}^+(n) \subseteq \mathcal{A}_o\) the set of outgoing arcs from a node \(n \in \mathcal{N}\), this is:

\[
\mathcal{A}^+(n) := \{a \in \mathcal{A} \text{ such that } a = (n, n'), \text{ with } n' \in \mathcal{N}^+(n)\}.
\]

Clearly \(|\mathcal{A}^+(n)| = \deg^+(n)|\).
The travel time status for any arc in the network \( a \in \mathcal{A}_o \) transitions between normal (N) and failed (F) states according to a two-state CTMC with repair frequency \( r_a \) and failure frequency \( f_a \). The failed state represents a degraded travel time regime. The mean travel time in the normal state is \( 1/\mu_a \) and in the failed state \( 1/\mu'_a \), where \( \mu_a > \mu'_a = \alpha_a \mu_a \) with \( 0 < \alpha_a < 1 \). This system was studied in detail in the works by Figueroa & Baykal-Gürsoy, who gave the Laplace transform of the travel time distribution as well as the explicit moments. We denote by \( T_a \) the random variable describing the travel time for arc \( a \in \mathcal{A}_o \), in particular we know \( \mathbb{E}[T_a] \) explicitly.

Let \( s_a \) denote the random travel time status of an arc \( a \in \mathcal{A}_o \). We write \( s_a \in \{N, F\} \) to indicate this (more formally: \( s_a(\omega) \in \{N, F\} \), where \( \omega \) is an outcome from the state space \( \Omega \)). We define the random travel time status vector \( S(A^+(n)) \) of the outgoing arcs from \( n \) as an arbitrary but fixed ordering of the statuses of the arcs in \( A^+(n) \). When necessary, \( S(A^+(n), \omega) \) will denote a particular outcome so that \( S(A^+(n), \omega) \in \{N, F\}^{\text{deg}^+(n)} \). Additionally, \( S(A^+(n), \omega)(i) \) will denote the \( i \)-th component of the vector, with \( i = 1, \ldots, \text{deg}^+(n) \), and, for \( a' \in A^+(n) \) we will have \( S(A^+(n), \omega)(a') \) denoting the travel time status of the arc \( a' \). We denote by \( T(s_a(\omega)) \) the travel time for arc \( a \) under the observed travel time status \( s_a(\omega) \) as if it was going to stay constant for the entire trip of arc \( a \). In our case, \( T(s_a(\omega)) \in \{1/\mu_a, 1/\mu'_a\} \).

We give some data for a test system for the network \( \mathcal{G}(N, \mathcal{A}_o) \) with the parameters in Table (5.1). This system is calibrated using travel time data and incident data with similar scales from a road system in Milwaukee, WI as described in Figueroa & Baykal-Gürsoy. The travel time is measured in minutes, while the failure and repair rates are in the scales of hours as it happens when weather conditions dominate the states of the system. Notice how, due to the effect of random incidents, the expected travel time \( \mathbb{E}[T_a] \) is always larger than the mean travel time under normal conditions.
1/\mu_a. We give also for every arc \( a \) in the network the calculated *upper semi deviation* measure of risk, which is \( r_{USD}[T_a] = \mathbb{E}[(T_a - \mathbb{E}[T_a])^+] \).

<table>
<thead>
<tr>
<th>Arc ( a )</th>
<th>( 1/\mu_a ) [min]</th>
<th>( \alpha_a )</th>
<th>( 1/r_a ) [hr]</th>
<th>( 1/f_a ) [hr]</th>
<th>( \mathbb{E}[T_a] ) [min]</th>
<th>( r_{USD}[T_a] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>5</td>
<td>0.9</td>
<td>1</td>
<td>20</td>
<td>5.04</td>
<td>3.41</td>
</tr>
<tr>
<td>2-3</td>
<td>10</td>
<td>0.8</td>
<td>5</td>
<td>5</td>
<td>11.45</td>
<td>6.55</td>
</tr>
<tr>
<td>2-4</td>
<td>22</td>
<td>0.6</td>
<td>1</td>
<td>20</td>
<td>23.20</td>
<td>9.95</td>
</tr>
<tr>
<td>3-4</td>
<td>10</td>
<td>0.8</td>
<td>5</td>
<td>5</td>
<td>11.45</td>
<td>6.55</td>
</tr>
<tr>
<td>3-5</td>
<td>18</td>
<td>0.8</td>
<td>10</td>
<td>2</td>
<td>22.32</td>
<td>9.12</td>
</tr>
<tr>
<td>4-5</td>
<td>10</td>
<td>0.9</td>
<td>2</td>
<td>10</td>
<td>10.33</td>
<td>5.83</td>
</tr>
<tr>
<td>5-6</td>
<td>5</td>
<td>0.9</td>
<td>1</td>
<td>20</td>
<td>5.04</td>
<td>3.41</td>
</tr>
</tbody>
</table>

Table 5.1: Data for Test Network \( G(N, A_0) \)

### 5.3.1 Simple Shortest Path Problem

The first implementations correspond to a simple Shortest Path Problem. Here, the random variables \( T_a \), are deterministic and considered as a weight \( w_a \). Figure 5.3.1 shows a weighted network.

![Weighted Graph](image-url)
5.3.2 Dynamic Programming

We use a modeling approach that is related to the studies in [78, 79] which are based on Markov Decision Processes (MDP). Our general strategy will be that of progressive refinement of the underlying network by maintaining a computationally tractable state-spaces and overall complexity.

Risk-Neutral Formulations

We write three risk-neutral dynamic programming (DP) formulations with increasing levels of detail. The first two are based directly on the network $G(N, A)$, while the last one is based on an augmented network $G(N \cup N_a, A)$. That we describe later.

The most basic version that we use as a benchmark is just the Shortest Path Problem (SPP) for the network $G(N, A)$ considering as arc-weights (travel times) the expected travel time calculated using the distributions obtained by Figueroa & Baykal-Gürsoy. This amounts to reducing $T(s_a) := \mathbb{E}[T_a]$.

This problem is solved easily by Dijkstra’s algorithm, or in a DP setting by backward recursion. The state space of the system is the set of nodes $N$, the action space is the the set of arcs with feasible actions $A^+(n)$ for state $n$. The immediate cost of taking action $a \in A^+(n)$ is the expected travel time $\mathbb{E}[T_a]$. The value of node $n$, $v(n)$, corresponds to the minimum travel-time to go until node $n_D$. The DP equations are given by (5.5).

$$v(n) = \begin{cases} 
\min_{a=(n,n')} \{ \mathbb{E}[T_a] + v(n') \} & , n \neq n_D \\
0 & , n = n_D 
\end{cases} \quad (5.5)$$

The system is calibrated with the data in Table 5.1 as shown in Figure 5.3, where the arc weights correspond the the expected travel time values for each arc.
The solution, calculated by backward recursion, is given in Figure 5.4. Adjacent to each node we write its value, and the shortest path is marked with red color.

Figure 5.4: Solution for SPP with $T(s_a) = E[T_a]$, Example 1

As it is expected a change in, for example, the degradation level of a particular link can result in another optimal path being chosen. While keeping everything else constant we drop the value of $\alpha_{3,5}$ from 0.9 to 0.6, which means that if before the road worked at 90% capacity whenever an incident occurred, now the road works at a capacity of 60%. This results in a change in the expected travel time, which rises from 20.99 [min] to 23.00 [min], which causes the solution to change as shown in Figure 5.5.

By setting other fixed values to the random variable realizations $T(s_a)$, several risk-neutral versions of the shortest path problem can be modeled:
Figure 5.5: Solution for SPP with $T(s_a) = \mathbb{E}[T_a]$, Example 2

- Optimistic SPP: $T(s_a) := 1/\mu$
- Robust SPP: $T(s_a) := 1/\mu'$
- Steady State Expectation SPP: $T(s_a) := \frac{r}{r + f} \cdot \frac{1}{\mu} + \frac{f}{r + f} \cdot \frac{1}{\mu'}$
- Risk-Averse SPP: $T(s_a) := \mathbb{E}[T_a] + r_{USD}[T_a] = \mathbb{E}[T_a] + \lambda \cdot \mathbb{E}[(T_a - \mathbb{E}[T_a])^+]$

Figures 5.6-5.8 show the solution paths for these versions of the problem, as well as the total travel time cost.

Figure 5.6: Solutions for Optimistic SPP: 38.00, and Robust SPP: 46.11

Figure 5.7: Solution for Steady State SPP: 42.74
The second risk-neutral formulation considers the dynamics of the system in a more explicit way, as we incorporate into the state of the system the observed status of the outgoing arcs which can be found in one of their two states. It is assumed that once the arc is observed, the conditions do not change until arriving to the endpoint node.

We recall the notation of set of arc status for the outgoing arcs of a given node. Consider node $n_1$ in Figure 5.1, we have that, as $n_2$ is the only successor, $S(A^+(n_1)) = (s_{1,2})$, and in particular $S(A^+(n_1), \omega)$ is either $N$ or $F$. On the other hand node $n_2$ has two successors, so we write $S(A^+(n_2))$ as $(s_{2,3}, s_{2,4})$, here $S(A^+(n_2)) \in \{N,F\}^2 = \{(N,N), (N,F), (F,N), (F,F)\}$.

We define then the state of the system as the current node being visited, $n$, plus the travel time status of the outgoing arcs in $A^+(n)$. Let $\mathcal{X}$ denote the set of all possible states of the system. Then, denoting by $x$ a particular state, equation (5.6) gives the formal definition.

$$x = (n, S(A^+(n), \omega)), \quad n \in \mathcal{N}, \quad \omega \in \Omega. \quad (5.6)$$

The action space and the definition of the value of a state are the same as before. The feasible actions from state $x = (n, S(A^+(n), \omega))$ is given by the set:

$$\mathcal{U}(x) = \mathcal{U}(n, S(A^+(n), \omega)) = A^+(n).$$

We denote by $u(x) \in \mathcal{U}(x)$ a specific action in the action space.
In order to write DP equations we need to have information on the probability of finding an arc \( a \) in a specific status \( s_a \). For this we make use of the stationary probabilities for the MC that modulates the service rate. Specifically, for an arc \( a \in A_o \), the probabilities \( \mathbb{P}\{s_a = s\} \) with \( s \in \{N, F\} \) can be calculated as:

\[
\mathbb{P}\{s_a = N\} = \frac{r_a}{f_a + r_a}, \quad \mathbb{P}\{s_a = F\} = \frac{f_a}{f_a + r_a}.
\]

Consider \( x = (n, S(A^+(n), \omega)) \in \mathcal{X} \). Then, we write the following DP equations:

\[
v(x) = \min_{\substack{a=(n,n') \in A(n) \setminus A^+(n), \, a' \in A^+(n')}} \left\{ T(s_a(\omega)) + \sum_{\nu \in \Omega} v(n', S(A^+(n'), \nu)) \cdot \prod_{a' \in A^+(n')} \mathbb{P}\{s_{a'} = S(A^+(n'), \nu)(a')\} \right\},
\]

(5.7) if \( n \neq n_D \), and \( v(x) = 0 \) for \( n = n_D \).

Table 5.2 and Figure 5.9 show the solution for this look-ahead formulation in which we observe the status of the outgoing arcs. Notice that the solution consists of a shortest path policy, that in general gives different recommendations according to observed realizations of the random variables (or upcoming travel time conditions).

![Figure 5.9: Solution for Look-Ahead SPP for \( G(N, A_o) \)](image)

For the last risk-neutral model we introduce a number of intermediate arc-nodes that we denote by the set \( N_a \). This augmenting set nodes subdivides each arc \( a \in A_o \)
Table 5.2: Solution for DP Equations (5.7) with data in Table 5.1
Figure 5.10: Augmented Graph $G(N \cup N_a, A)$ according to the CTMC of arc $a_{3,4}$ from Figure 5.1. The state and the actions of the system are defined as in the previous model.

Consider two consecutive arcs, $a'$ and $a''$, in the augmented network stemming from the same arc $a$ in $A_0$ in the original network. Since the travel-time status for both arcs is governed by the same CTMC, knowing the observed travel time status of $a'$ and its corresponding travel time before reaching $a''$, we can calculate the transient transition probabilities of status change. Let $t$ be the time taken to transverse arc $a'$ under its observed travel time status $s' := s_{a'}(\omega)$, $\omega \in \Omega$. We denote the probability of finding arc $a''$ in status $s'' = s_{a''}(\nu)$, $\nu \in \Omega$, by $p_{s',s''}(t)$. Of course, with only two possible travel time states we only have four possible transition probabilities, which are shown explicitly in Equation (5.8).

$$P(t) := \begin{pmatrix} p_{N,N}(t) & p_{N,F}(t) \\ p_{F,N}(t) & p_{F,F}(t) \end{pmatrix} = \begin{pmatrix} \frac{r}{f+r} + \frac{f}{f+r}e^{-(f+r)t} & f \frac{f}{f+r} - \frac{r}{f+r}e^{-(f+r)t} \\ \frac{r}{f+r} - \frac{r}{f+r}e^{-(f+r)t} & f \frac{f}{f+r} + \frac{r}{f+r}e^{-(f+r)t} \end{pmatrix}. \quad (5.8)$$

When no conditioning information on the travel time status is available because of arc independence (whenever two consecutive arcs in the network stem from different arcs in the original network) we can use the probabilistic approach as before making use of the stationary probabilities of encountering a specific outcome for the travel
For a state $x = (n, S(A^+(n), \omega)) \in \mathcal{X}$ we again define its value $v(x)$ as the expected travel time to go from node $n$ under travel time conditions $S(A^+(n), \omega)$ to the destination node $n_D$. Equations (5.9)–(5.11) give the DP equations.

Consider $x = (n, S(A^+(n), \omega)) \in \mathcal{X}$. Then, for $n \in \mathcal{N} \setminus (\mathcal{N}_{a, \text{end}} \cup \{n_D\})$ we write,

$$v(x) = \min_{a = (n,n') \in A^+(n)} \left\{ T(s_a(\omega)) + p_{s_a(\omega),N} \cdot v(n', S(A^+(n')) = N) + p_{s_a(\omega),F} \cdot v(n', S(A^+(n')) = F) \right\}$$

(5.9)

On the other hand, for $n \in \mathcal{N}_{a, \text{end}}$ and denoting $n' := N^+(n)$,

$$v(x) = T(s_{n,n'}(\omega)) + \sum_{\nu \in \Omega} v(n', S(A^+(n'), \nu)) \cdot \prod_{a' \in A^+(n')} \mathbb{P}\{s_{a'} = S(A^+(n'), \nu)(a')\} .$$

(5.10)

Notice that for any arc $a \in A$, the probabilities $\mathbb{P}\{s_a = s\}$ with $s \in \{N,F\}$ can be calculated as

$$\mathbb{P}\{s_a = N\} = \frac{r_a}{f_a + r_a}, \quad \mathbb{P}\{s_a = F\} = \frac{f_a}{f_a + r_a} .$$

Finally, for $n = n_D$, for which $A^+(n_D) = \emptyset$, we set

$$v(x) = 0 .$$

(5.11)

Figure 5.11 shows the shortest path policy for the augmented look-ahead SPP model with $K = 3$ subdivisions. Here $v(1, N) = 41.60 \text{[min]}$, $v(1, F) = 42.14 \text{[min]}$, and we define $v(1) := \frac{r_1}{f_1 + r_1} v(1, N) + \frac{f_1}{f_1 + r_1} v(1, F) = 41.63 \text{[min]}$, which is the overall cost.

Figure 5.12 shows the convergence of $v(1)$ for the same example as the number of subdivisions $K$ increases. It is clear that there is a limit on the level of detail added by further subdivisions.

In Table 5.3 we compare the total costs of all the presented models for the same example.
Figure 5.11: Solution for Augmented Look-Ahead SPP, $K = 3$ Subdivisions

Figure 5.12: Convergence of Total Cost $v(1)$ as $K$ Increases

<table>
<thead>
<tr>
<th>Method</th>
<th>Optimistic</th>
<th>Look-Ahead</th>
<th>Augmented $K = 2$</th>
<th>Augmented $K = 60$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost [min]</td>
<td>38.00</td>
<td>41.56</td>
<td>41.61</td>
<td>41.65</td>
</tr>
<tr>
<td>Method</td>
<td>Steady State</td>
<td>$\mathbb{E}[T]$</td>
<td>Robust</td>
<td>Risk Averse ($\lambda = 0.2$)</td>
</tr>
<tr>
<td>Cost [min]</td>
<td>42.74</td>
<td>43.32</td>
<td>46.11</td>
<td>48.15</td>
</tr>
</tbody>
</table>

Table 5.3: Total Cost Comparison of SPP
Finally, we test our methods in a 16-node network from the works of Sever et al. [78]. The network has 4 out of 24 vulnerable (bi-directional) arcs. Figure 5.13 shows a diagram of the network.

Figure 5.13: Test-Network with $\alpha = 1/3$ for Vulnerable Arcs

Figure 5.14 shows the solutions for the Look-Ahead and Augmented $K = 5$ SPP.

Figure 5.14: Test-Network Solutions for Look-Ahead SPP and Augmented ($K = 5$) SPP

Table 5.4 shows the total costs of the solutions of the SPPs.
<table>
<thead>
<tr>
<th>Method</th>
<th>Optimistic Look-Ahead Aug $K = 5$</th>
<th>$E[T]$</th>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost [min]</td>
<td>24.00</td>
<td>25.22</td>
<td>25.22</td>
</tr>
</tbody>
</table>

Table 5.4: Total Cost for Test Network

5.4 Conclusions and Future Work

The representation of a risk prone corridor by a two-state CTMC has been validated by the authors, and hence the modeling of a SPP with Markov-modulated arc-weights is reasonable. The calculation of the solutions is computationally tractable in the small scales tested, and it is expected to be scalable. In this sense, the addition of risk aversion and/or more refined travel time calculations including variability is warranted.

We outline now the general characteristics of the dynamic programming model and the stochastic programming model. In both cases a detailed and formal formulation is left as future research, as well as the detailed solution strategies/algorithms, and computational considerations and implementations.

It is important to highlight that in both cases we will make use of the completion-time distributions from Chapters 3 and 4 for informing the travel-time distribution of any vulnerable link.

5.4.1 Stochastic Programming Approach

We propose to intensively study multi-stage stochastic programming formulations for the DSSPP, with the hope that they will allow for more flexibility on the inclusion of risk aversion and the use of chance constraints than existing formulations available in the literature. We will present a general modeling structure which will be specialized if tractability mandates it.
Here we also consider the finite, directed, and weakly connected graph $\mathcal{G}(\mathcal{N}, \mathcal{A})$. A non-negative discrete time-dependent random variable $q_a(t)$ is assigned to each arc to model the travel time in the network following the dynamics of recurrent congestion. We will write $q_a(t, \omega)$, where $\omega$ is an event in a sample space whenever there is a need to specify a random outcome instead of the random variable. On a first approximation we assume that these random variables are independent from each other and we postpone the calibration details to a later stage of the study. We assume that a pre-established set of arcs in the network (which range from none to all) is prone to stochastic incidents generating non-recurrent congestion. We explore the possibility of modeling these random variables with Markov chains with a number of disrupted states. The travel time delay imposed by these incidents is then added to the recurrent congestion. It is worth mentioning that there are approaches stemming from queueing theory to model both types of congestion at the same time. Indeed, it has been proposed that the distributions obtained by Baykal-Gürsoy and Xiao [9] can be appropriate to model traffic congestion states under incidents, which will be validated as part of this study.

In the multi-stage setting each stage corresponds to the decision of which arc to follow from the current node. The starting node is $n_1$ and the first-stage decision is to choose the second node to visit. At each state the random travel times of the arcs are realized and a decision for a consecutive node needs to be made. Assuming that at stage $s$ we are in node $n$ with $n \neq n_N$, we define binary decision variables for this stage as

$$x^s(n') = \begin{cases} 
1, & \text{if the node } n' \text{ is visited next,} \\
0, & \text{otherwise.} 
\end{cases}$$

(5.12)
Here, $n'$ is in $\text{SUCC}(n)$, where $\text{SUCC}(n)$ denotes the set of successor nodes of node $n$. From the destination node $n_N$, however, we set $x^s(n') = 0$ for any successor node $n'$ in $\text{SUCC}(n_N)$.

In a first approximation we are looking for a travel-time-shortest cycle-free O-D path in the network, and hence we do not allow backtracking or waiting at any node, even if the network structure allows it. The only exception to these restrictions is the possibility of waiting at the destination node $n_N$ at termination. We emphasize that these three restrictions on paths (cycle-free, no backtracking, and no node waiting) will be relaxed if it proves feasible in the advancement of this research. Using these assumptions we know that $N - 1$ is an upper bound on the number of stages, since this is the maximum possible arc-size of a cycle-free O-D path. If the network itself is acyclic the number of stages can be reduced by using the length of the the longest O-D path, which can be obtained by any shortest path algorithm in linear time. Let $L + 1$ be the node-length of this longest path if the graph is acyclic or plus infinity otherwise ($+\infty$). Then, the number of stages to be considered can be bounded by $S \leq \min\{N - 1, L\}$.

Each stage of the multi-stage model will have a recourse cost (travel time), which will be ultimately collected in the objective function of the first-stage problem adjusted by some risk aversion measure. For this, we will follow and adapt the approaches found in [56] and [69] on nested risk-aversion measures. The incorporation of chance constraints to model reliability concerns will also be evaluated. We will look at the possibility of adding chance constraints on each stage or only on the first-stage of the problem.

We will aim to develop specialized scenario decomposition methods for our DSSPP.
For this purpose, we will have to deal with the convexification of each stage problem and the development of efficient bounding procedures on the approximated problems. By means of formal analyses we will first concentrate on developing theoretically sound techniques and establishing their computational complexity/efficiency. On a second stage of the research, we will focus on the tractability of the problem, by elaborating efficient optimization algorithms and also acceptable approximation alternatives such as specialized heuristics.

We also want to study the structure of the optimal or sub-optimal solutions if there is any to be found. We hope that sensitivity analysis, as well as carefully planned experiments, will reveal some underlying patterns. An example of such patterns can be found in the work by Sever et al. [78], where their examples of the so called online($n$) and hybrid($n$) routing policies appear to be nested according to the level of information detail given to the agent (Figure 3 on the paper: the higher the information level, the more instructions the routing policy gives, but all previous instructions are still in effect). Is this indeed a property of the solutions? Can it be proved formally? Other questions that may arise are: are there problem instances that allow for a priori include/exclude certain arcs/paths decisions? Are there problem instances in which disjoint arcs/paths can be ranked according to some travel-time preference?

These and other research questions will be addressed in detail with the objective of determining if a multi-stage stochastic programming formulation is appropriate to model a DSSPP with probabilistic constraints. While many parts of the study will be analytical in nature, many others will consist of computational experiments and simulations. The study we propose intends to give a complete overall first view of the subject that will allow further investigation.
Bibliography


User guide for the freeway reliability methodology computational engine. Institute 
for Transportation Research and Education, University of North Carolina.


Handbooks in operations research and management science, volume 12, chapter 9, 

shortest path problems: Hybrid routing policies considering network disruptions. 

Dynamic Shortest Path Problem with Travel-Time-Dependent Stochastic Disrup-
tions: Hybrid Approximate Dynamic Programming Algorithms with a Clustering 
Approach.

[80] Shapiro, A. (2012a). Minimax and risk averse multistage stochastic program-


programming: modeling and theory. SIAM.


