ABSTRACT

ASHBAUGH, BRADLEY KENNETH. Extending the Kingman General Arrival, General Service Queuing Approximation to Overhead Allocation and Service Pricing. (Under the direction of Michael G. Kay and Robert B. Handfield).

In addition to recording transactions, accounting systems are used to evaluate performance, help determine what to produce, allocate overhead, and determine break-even pricing. Most accounting systems are well suited to calculate average costs but are less successful in determining marginal costs and most ignore arrival and process variability effects. Only time-driven activity-based costing attempts to measure process variability, and no established system measures the costs of arrival variability. From queuing theory, it is known that variability affects work-in-process inventory (WIP) for products and average wait times for services. Higher variability causes higher WIP and higher average wait times for a given resource utilization level. This results in higher inventory costs for products and lower customer satisfaction for customers.

This dissertation continues the work of using allocated clearing functions to account for variability in production. Using a two product, three resource production facility, clearing functions are used to charge for WIP as resources become congested. This system is compared to a classic linear programming overhead allocation schema that preserves the optimal product mix solution. In solving this problem, a new way to derive clearing functions that makes fitting empirical data and generating additional piecewise linear clearing function segments easier was developed and is employed. Next, the problem is extended to multiple, parallel resources. Finally, the clearing functions are reformulated to work with pure service systems. Two applications are investigated. First, a generic service system with both high
variability and low variability customers is used to determine capacity needs and resulting cost differentials. Second, an MRI facility with both scheduled patients and emergency room patients is modeled to determine how much more capacity is needed for emergency room patients to maintain a required level of service versus scheduled patients. Per patient capacity needs are then used to justify differential pricing between the different patient types.

The results of this research show that WIP costs can be integrated into some existing linear programming based overhead allocation methods to generate better cost estimates of resource values. Customer wait time models can be used to better allocate extra capacity to customers based on their arrival and service variabilities. These better cost estimates can be used to justify variance reduction efforts, better identify customers or products that are not profitable, and use market-based pricing to change customer behavior or at least cover the additional costs based on their increased variability.
Extending the Kingman General Arrival, General Service Queuing Approximation to Overhead Allocation and Service Pricing.

by
Bradley Kenneth Ashbaugh

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Industrial Engineering

Raleigh, North Carolina
2019

APPROVED BY:

Robert Handfield
Co-chair of Advisory Committee

Michael Kay
Co-chair of Advisory Committee

Richard Wysk

Thom Hodgson
DEDICATION

This dissertation is dedicated to

*My wife who worked very hard to help me reach my dream*

*My son who pushed me to finish*

*My mother and father who helped me love school*

*My friends who encouraged me*

*Math students, some of whom have learned to share my passion*
BIOGRAPHY

Bradley Ashbaugh was born in Los Angeles, California as the son of a Drama / English / History high school teaching and World War II veteran father and a Math / Computer high school and later community college teaching mother. At age 3 he moved to San Clemente, California where he grew up loving competitive swimming, water polo, and surf riding. The love of the water and the beach paid off as he became a lifeguard at San Clemente, Doheny, and San Onofre State Beaches. In addition to saving hundreds of lives, the job was important as it enabled him to pay his way through college.

He went to San Diego State University starting in Aerospace Engineering. He later switched to Mechanical Engineering when he saw deep budget cuts in the aerospace industry with the fall of the Soviet Union. While at SDSU, he was fortunate to become involved in student government where he had the opportunity to serve as President of the Associated Engineering Student Council leading and participating in College of Engineering wide events. He was honored to be inducted into Tau Beta Pi, Pi Tau Sigma, and the Order of Omega honor societies as well as receive a university award for Outstanding Contribution to Student Government. He also went on to receive a minor in Economics and was an intern for the City of San Diego Traffic Engineering Department. Despite successfully graduating, something was amiss. He loved engineering and economics but lacked intuition and passion for machine design. He had a talent for Operations Management and Computer Programming, so he decided to enter graduate school in a major that combined Engineering, Operations, and Economics.

Having previously visited as a student government representative from SDSU, he was excited to enroll at NC State University in the Master’s program in Industrial Engineering.
He enjoyed his classes and was incredibly lucky to be assigned as teaching assistant to IE 498: Senior Design. Working with Dr. William Smith and Dr. Wilbur Meier to help student teams solve real industry problems was his dream assignment. He found that he had an ability to quickly see through problems at industry facilities which helped earn him the Outstanding Teaching Assistant award from Industrial Engineering. He was also fortunate to meet his future wife, Maria, a Furniture Manufacturing and Management degree recipient from NC State University and jointly an Interior Design degree holder from Meredith College.

Upon graduating with a Master’s of Industrial Engineering, he went to work at SABRE Decision Technologies as a consultant. He improved and automated business processes at travel industry related customers including Carlson Wagonlit, American Express, Hilton Hotels, INTRAV, Renaissance Cruises, and Grand Circle Travel using proprietary software that connected to reservations systems and mainframe computers. His work locations included Dallas, TX; Chicago, IL; Phoenix, AZ; Newport News, VA; Fort Lauderdale, FL; Saint Louis, MO; Boston, MA; Paris, France; and one year between Alton, Brighton, and London in the United Kingdom. Despite being successful in designing custom solutions and auto-booking systems, SABRE wanted to move him from a lead design role to a person brought in to fix projects already in trouble. He found a better work environment at Renaissance Cruises where he simplified and automated processes to help meet 800% capacity growth in less than three years without adding people to the departments he helped. Unfortunately, the terrorist attacks on the World Trade Center in New York City caused the cruise line to shut down and liquidate.
After a time-out for family and a return to saving lives in California, he enrolled in the PhD program of Industrial and Systems Engineering. Once again he found a dream assignment in teaching BUS 370: Operations Management for the Poole College of Management at NC State University. Buoyed by good teaching evaluations, he also taught BUS 472: Operation Planning & Control Systems, then MBA 541: Supply Management for the Master’s in Global Luxury Management dual degree program, MBA 540: Principles of Operations and Supply Chain Management for the Jenkins MBA Program and BUS 340: Information Systems Management. He also taught BUS 343: Operations Management, BUS 250: Data Analysis for Business Decisions, and MBA 680: Operations Management at Meredith Collage. In 2016 - 2018, he taught BUS 370: Operations Management and M 298: Doing Business in China as part of the Poole Study Abroad Summer Program in Shanghai, China (M 298 was changed to M 380: Doing Business Globally in 2018). Finally, he was hired by both Wake Forest University and the SKEMA School of Business to teach BEM 241: Production and Operations Management and ECBUS 3600 Logistics and Trade in the US, respectively. In all, he was blessed to be instructor-of-record for forty course sections at NC State University, seven course sections at Meredith College, two courses at Wake Forest University, and one course at the SKEMA Business School during completion of his degree.
ACKNOWLEDGMENTS

I have been fortunate to learn from amazing professors. My journey in Industrial Engineering started with Mr. Clarence Smith and Dr. Steve Roberts offering me a teaching assistantship so I could afford Graduate School. I took classes from Dr. Richard Bernhard and Dr. William Smith my first semester and they became my earliest mentors. I am a better businessperson from my experience with Dr. Smith and I am a better academic from Dr. Bernhard. I was fortunate to also witness the teaching styles of Dr. Thom Hodgson, Dr. Russell King, the late Dr. Wilbur Meier, and Dr. James Wilson. My Accounting instructor, Dr. Gilroy Zuckerman was the first to tell me that I had a talent for teaching on the university level. This education served me well in my consulting and business career.

Upon returning to NC State University for my PhD, I was lucky to get to learn from Dr. Paul Cohen, Dr. Yahya Fathi, Dr. Ola Harrysson, Dr. Julie Ivy, Dr. Dave Kaber, Dr. Michael Kay, Dr. Yuan-Shin Lee, Dr. Reha Uzsoy, Dr. Richard Wysk, and Dr. Robert Young in addition to Drs. Bernhard, Hodgson, King, and Wilson. Although the subject matter varied among courses, I try to bring lessons I have learned from them to be a better instructor.

I also want to thank Dr. David Baumer from the Poole College of Management for giving me the opportunity to teach Operations Management at NC State and the other instructors and administrators who have shown me nothing but kindness including Dr. Steven Allen, Dr. Brian Ashford, the late Dr. Cecil Bozarth, Dr. Al Chen, Mr. Carlton Daniels, Dr. Shannon Davis, Dr. Bartley Danielson, Dr. Julie Earp, Mr. Donovan Favre, Ms. Sherry Fowler, Ms. Tracy Freeman, Mrs. Ellen Frost, Dr. Robert Handfield (again), Mr. Steven Hayes, Dr. Hans Sebastian Heese, Dr. Sarah Khan, Dr. Neal Parker, Mr. Richard Podurgal, Dr. Christian Rossetti, Mr. Robert Sandruck, Dr. Jeffery Stonebraker, Dr. Richard Warr, Dr.
Donald Warsing, Dr. Ira Weiss, and Dr. Gilroy Zuckerman (again). A special thank-you to Mr. Donovan Favre, Dr. Christian Rossetti, Dr. Donald Warsing, and the late Dr. Cecil Bozarth for letting me use their course notes to help make my teaching more effective.

Additional thanks to Dr. Mohamed Desoky at the SKEMA Business School, Dr. Haresh Gurnani and Dr. Mark Evans at Wake Forest University’s School of Business, and Dr. Jeff Langenderfer and Dr. Mary Jane Lenard at Meredith College’s School of Business for entrusting me to teach their students. It has been a pleasure to teach at each of these institutions and I hope I have inspired their students to use Operations Management tools and techniques to make their future businesses better.
# TABLE OF CONTENTS

**LIST OF TABLES** ........................................................................................................... x

**LIST OF FIGURES** ......................................................................................................... xi

CHAPTER 1: Introduction ........................................................................................................... 1

CHAPTER 2: Previous Related Work ................................................................................... 15
  2.1 Accounting Research ................................................................................................. 15
  2.2 Linear Programming Research .................................................................................... 17
  2.3 Medical Applications .................................................................................................. 18

CHAPTER 3: Overhead Allocation Using Clearing Functions ............................................... 20
  3.1 Overhead Allocation Using Traditional Clearing Function Formulation ................. 21
  3.2 Where the Kefeli (2011) Model Goes Wrong .............................................................. 27
  3.3 Allocating Untraceable Overhead ............................................................................... 28
  3.4 Allocating Traceable Overhead .................................................................................. 33
  3.5 Allocating Overhead with Product Interdependencies ............................................. 44
  3.6 Conclusions ................................................................................................................ 46

CHAPTER 4: Alternate Clearing Function Derivation .......................................................... 50

CHAPTER 5: Allocating Overhead Using Clearing Functions with Parallel Resources ..... 59
  5.1 Deriving Clearing Functions with Symmetrical Parallel Machines ....................... 59
  5.2 A Numerical Example Using Symmetric Parallel Machines ................................... 61
  5.3 Extending Multiple Parallel Machines ...................................................................... 62

CHAPTER 6: Overhead Allocation Using Asymmetric Parallel Machines or Different
  Customer Arrivals ............................................................................................................. 63
  6.1 Asymmetric Parallel Machines ................................................................................... 63
  6.2 Using Asymmetric Parallel Machines to Model Different Product Mixes ............... 63
  6.3 Incorporating Average Customer Waiting Time Costs in Overhead Allocation for
      Service Systems .......................................................................................................... 65
  6.4 Using Geometric Power Series to Allocate Capacity .............................................. 66

CHAPTER 7: A Medical Example Using Two Patient Types ............................................. 68
7.1 MRI Numerical Example .......................................................... 69
7.2 Monte Carlo Simulation ...................................................... 71
7.3 Conclusions ............................................................................ 73

REFERENCES 75
LIST OF TABLES

Table 3-1: Numerical Example Common Coefficients ............................................................. 21
Table 3-2: Clearing Function Segments .................................................................................. 21
Table 3-3: Model Output from CF Formulation ..................................................................... 24
Table 3-4: Duels of the Optimal Product Mix for Both the LP and CF Formulation .......... 25
Table 3-5: Contribution Margin of the Optimal Product Mix for Both LP and CF Model .... 36
Table 3-6: Optimal Values Between CF Models ................................................................. 41
Table 3-7: Comparison of CF Formulation Duels ................................................................. 41
Table 3-8: Comparison of Duels Between LP and CF Formulations with Product Interdependencies ................................................................................................................. 44
Table 4-1: Comparison of 3 Segment versus 20 Segment Clearing Function .................... 55
Table 4-2: Machine 1 Enhanced Excel Clearing Function ..................................................... 55
Table 4-3: Machine 2 Enhanced Excel Clearing Function ..................................................... 56
Table 4-4: Machine 3 Enhanced Excel Clearing Function ..................................................... 57
Table 5-1: LP Output for Symmetric Parallel Machines ....................................................... 62
Table 7-1: Overhead Allocation Summary by Method and Patient Type ............................ 71
Table 7-2: Monte Carlo Simulation Results .......................................................................... 72
LIST OF FIGURES

Figure 1-1: WIP vs Utilization ................................................. 5
Figure 3-1: Graph of Clearing Function Segments ........................................ 22
Figure 4-1: 20 Segment Clearing Function ............................................. 52
Figure 4-2: Chapter 3 Model Replaced with 20 Segment Clearing Function .......... 54
Figure 4-3: 20 Segment Clearing Function with High Variability in Machine 2 ........ 58
Figure 5-1: Graph of Clearing Functions with 1 through 5 Identical, Parallel Machines .... 61
Figure 7-1: MRI Time Estimates from Advanced Imaging in Missoula, MT ............... 68
Figure 7-2: Monte Carlo Simulation Model for MRI Usage .................................. 72
CHAPTER 1: Introduction

What is the cost of variability? This is a simple question that does not often have an easy answer. We know from the Law of Variability (Hopp and Spearman 2001) that the more variability to a process, the less efficient the process. Variability can be observed in many forms. There can be variability in worker productivity, machine performance, quality, product variety, incoming materials, tooling, environmental factors, employee absenteeism, learning curves, training, education, physical ability, morale, machine dependability, and other factors. There is variability of customer arrivals, expectations, degree of customization, tastes, and in the medical profession, drug efficacy, anatomical variations, ability to fight infections, and attitude. This dissertation provides a methodology that answers the question: how do we account for this variability when it comes to pricing or overhead allocation?

For those not well versed in Queuing Theory, higher arrival variability manifests itself as lumper arrivals (Hopp and Spearman 2001). That is, there may be longer periods between arriving customers and when customers do arrive, they tend to come close together. From the employer perspective, it may mean paying employees for long periods where they have no customers to serve. But when customers arrive, they can be so tightly bunched that lines can quickly form discouraging some customers from buying. Higher service variability means that there will be more customers needing extensive time to be served. When a customer that needs a long time to serve is the only one in the establishment, there is not problem. However, if a server is stuck serving a customer that needs a long time to be served when lots of other customers arrive, the result might be customers leaving before making their transaction. Think about a grocery store with only one cashier and finding the customer in the front of the line has three carts full of items and a folder full of coupons. Customers
trying to quickly purchase a few items are likely to look elsewhere to buy their products. The same phenomena can be observed in factories when larger orders reach the factory floor, when batches of parts arrive at a machine, a factory recovering after a machine breakdown, and even in the loading and unloading of trucks. In each case, this variability can immediately mean either paying employees for no work, losing potential sales due to short-term long lines, or having to pay employees overtime should the backups occur at the end of shift. Secondary effects include needing longer lead times to fill orders which can decreases customer service, increase inventory, or require more surge capacity. Additionally, increasing lead time reduces forecasting accuracy as forecasts are less accurate farther in the future.

Production planning, product costing, and overhead allocation are all important functions of both manufacturing and service organizations. Profitable enterprises are set pricing or reduce costs such that both fixed and variable overhead can be fully absorbed in the products or services rendered. However, competitive pressures often tempt management to limit capacity in order to reduce overhead but at a cost of increased WIP, customer waiting, or system cycle time. In addition, when there is arrival or process variability, manufacturing and service systems operate less efficiently. Because many organizations measure product or service profitability through their accounting systems, it is important to choose a system with an overhead allocation schema that is aligned with the optimal product mix problem. If the system was also sensitive to the effects of variability, managers would have a powerful tool to direct process improvement resources to opportunities that have greater positive impact on operational performance and profitability.

In many cases, different accountings systems do not tell the same story when using them to allocate overhead, price new products, or even determine product profit margins.
There have been several attempts that work well depending on the number of products, the number of resources used, and the number of resource constraints. A core problem with accounting systems is that they are usually designed to assign costs and profits once production and sales have already occurred. They are less reliable as a means of planning and pricing. Provided is a quick summary of some strengths and weaknesses of several managerial accounting allocation schemes.

1. Full absorption costing is the most traditional accounting system and has been used for over a century. It is not without flaws or detractors. Overhead is applied to products via a primary overhead driver. In many applications this driver is direct labor. A weakness of this approach is that it often drives attention to reducing the driver inputs to a product or facility in order to reduce the overhead assigned. The factory may have the exact same total labor charges but be able to shift large overhead charges to other products or facilities by reducing the quantity of labor considered direct, should direct labor be the overhead driver. Relatively, this kind of accounting system provides little incentive to reduce overhead versus the incentive to shift overhead to others.

2. Direct costing, or marginal cost accounting, does not assign fixed overhead to products and services. If one were to look at fixed overhead costs as strictly sunk costs, this makes some sense. However, this makes capital budgeting difficult as investment may not be recaptured with direct costing and has been blamed for the failure of several large companies including Braniff Airways.

3. Activity-based costing is an improvement over full absorption accounting in that it allocates overhead based on multiple overhead drivers. This is especially useful in
planning. The additional detail makes the system more complex and has another drawback of not always being aligned with or recommending the optimal product mix, especially when average costs and marginal costs diverge.

4. Throughput accounting works very well at recommending the most profitable production plan when there is a single constrained resource capping production. Although throughput accounting can be very useful, it is not generally accepted accounting for use in financial accounting or to determine transfer pricing.

5. Linear programming allocation systems will allocate overhead while preserving the optimal product mix no matter how many constrained resources are in a facility. However, traditional linear programming allocation systems require that resources be utilized 100% in order for them to have shadow prices and the ability to assign overhead to them. It is a well-known result from Queuing Theory that is very difficult to achieve 100% utilization, especially when there is any arrival variability or process variability. The more variability at a resource, the larger the work-in-process inventory and the longer cycle time will be as seen in Figure 1-1.
Finally, time-driven activity-based costing captures some of the process variability by charging for various sources of different process times and creating time equations based on which factors are present. This richer breakdown of processing time is certainly useful. However, arrival variability is ignored. The M/D/1 queuing model, despite having fully deterministic (and constant) service times, still shows the effect of just arrival variability, albeit at half of the rate of the traditional M/M/1 queue.

Because these traditional and alternative accounting systems fail to fully account for both arrival and process variability, they are likely to overestimate resource and system output. One common way to prevent this is for planning to cap available capacity at 80% of theoretical capacity for service systems and 85% of theoretical capacity for manufacturing resources. There are at least three potential problems with this approach. One is that the
threshold may be wrong for a particular process or resource. If variability is low, output will be arbitrarily constrained. Conversely, if variability is high, the adverse effects of high WIP and cycle time will add costs and complexity to the system while still missing output goals. A second potential problem is a missed opportunity to focus resources on eliminating process variability and for scheduling to help reduce arrival variability. If the accounting is content with output based on 80 – 85% utilization, why strive for better? A third potential problem is a wasted opportunity to price products, services, or customers better. Products or services with more variability cost the company more and should not be subsidized by lower variability offerings unless there is a compelling business case. Also, consider a company with two different major customers. A customer that brings in steady business costs the company less than a customer with infrequent but urgent requests. A better understanding of how each customer affects costs can lead to better pricing or even turning away unprofitable customers.

The first part of this dissertation is devoted to finding a way to incorporate variability and queuing effects into the allocation of fixed overhead to products. It will be assumed that the two products have similar arrival and process variability. The first model used will be a two product, three machine production system using the classic clearing function models proposed by Asmundsson et al. (2009). A second model will use a clearing function developed by a new derivation from the Kingman G/G/1 queuing approximation that is useful in generating finer resolution piecewise linear clearing functions. This higher resolution is useful to show how much of the theoretical capacity will likely be used given that WIP will increase as utilization of the theoretical capacity increases. At some point, additional WIP costs will exceed the value of additional production. This will indicate what
level of capacity utilization the system is likely to support economically. In addition, the model will calculate shadow prices for resources employed and many will be nonzero even when utilization is less than 100%. This will allow managers to better determine pricing, capital budgeting for new capacity, and a way to estimate the potential to increase profits through variability reduction initiatives. In addition, it may be possible to observe practical bottlenecks from resources with higher theoretical capacity but higher variability than other resources. A third model will use clearing functions derived from G/G/m queuing where single machines can be replaced with multiple identical parallel machines.

A business case for the first half of this dissertation follows. Imagine a simple production shop where there are three machines. One machine is shared between two products while the other two machines are dedicated to each product. The shared machine could be a drilling machine with little to no setup between products. The dedicated machines may be shapers that are calibrated to the specific shapes of the two products. The products have different machine needs and sell for different prices. However, it is desired that the overhead for this machine shop be fully assigned to its products. Chapter 3 shows how this allocation should take place, when accounting for different arrival variability and process variability is resource dependent, not product dependent. Chapter 4 will provide the derivation and implementation of an alternative clearing function formulation. Finally, the model will be extended to multiple parallel machines in Chapter 5.

The second half of this dissertation proposes a way to allocate overhead and determine pricing when customers have different arrival and / or cause different process variability. It will use variability and steady state queuing to help determine how much extra capacity is needed and how to allocate any extra capacity between products, customers,
departments, or in this hypothetical case, patients. The model will have a resource shared between two products, customers, or departments each with different arrival and process variability. A two-step optimization problem using clearing functions to determine how much of the resource capacity, both used and unused, should be applied to each product, customer, or department. Multiple parallel resources serving both customers will also be tested versus dedicated resources to see if and where dedicated resources might produce superior results.

A compelling business case for is easy to construct. Imagine a hospital with its many high value resources such as operating rooms, diagnostic equipment, and specialized physicians. How should it price its services? Certainly, services should be priced to at least cover their direct costs. To be fully profitable, they need to price to also cover their fixed overhead costs. If the pricing is not competitive for any service, specialty groups can form and siphon off patients forcing other hospital services to charge more potentially making additional services uncompetitive. One way to try to stay competitive is to make sure these high value resources are all highly utilized. However, we know from queuing that highly utilized resources generate long waits and crowded waiting rooms. We also know that the more variability involved in either arrival or processing time increases line length and wait times.

Many hospitals have expensive diagnostic equipment such as MRI machines. There may be appointments booked in advance for patients needing non-urgent imaging. On the other hand, often there will be unpredictable and extremely time-sensitive demand from the Emergency Room (ER). From queuing, we know that there needs to be some extra capacity built into the system whenever there is arrival variability and / or process variability. The higher the variability, the faster WIP starts to accumulate and, by Little’s Law, so does cycle
Both scheduled patients and ER patients inherently require some process variability. According to Advanced Imaging, MRI scan times range from 20 minutes to one hour depending on the body part to be scanned. Patients may be slow to load, there may be stops to clean up from messy patients, tracer dye might be slow to reach the desired imaging sites, and patients can inadvertently move requiring rescans. In a hospital setting, patients with long wait times can result in decreased customer satisfaction at best and at worst cancellations, loss of future business, diminished patient outcomes, and even patient death. Therefore, despite the large capital costs associated with high technology diagnostic equipment, it may be prudent to have some planned unused capacity built into a schedule. Two questions arise: how much extra capacity do we need and how do we allocate the costs of this extra capacity to the patients?

One method is to apply direct costs to all patients using the equipment and share the extra capacity costs evenly across all patients based on averages. For example, if the MRI averages 20 patients over a 24-hour period, charge each patient 20/24 or 5/6 of an hour for the MRI’s overhead costs. However, we know that we need more additional capacity to serve ER patients than those regularly scheduled due to the higher arrival and process variability these patients have. Charging all patients the same for the overhead of the MRI can have an adverse effect on the hospital when there are specialty MRI facilities that do not need to be available for emergency patients. The specialty MRI facility can schedule to reduce variability thereby removing the need for as much extra capacity and therefore charge less while maintaining greater customer service. An interesting intuition: throughput accounting would only consider the preemption of scheduled patients by ER patients if the ER patients generated more revenue per minute on the MRI.
Another option is to charge scheduled patients overhead based on just the time they use the machine. All the overhead from the extra capacity can be applied to the ER patients. In practice we would see MRI bills from scheduled appointments much lower than the same MRI from ER patients. Let us return to the example where the MRI averages 20 patients in a 24-hour period. Let us further stipulate that there are 5 ER patients, on average, each day. The 15 scheduled patients would just be charged their time on the machine. Each ER patient would be charged their time plus 1/5 of the total unused average time left in the day. If each MRI averaged 1 hour, each emergency room patient will be charged for 1.8 hours. Now imagine that total time on the MRI machine averaged 30 minutes per patient. Now the ER patients will be paying for the 30 minutes plus 14/5 hours of the unused capacity, a total 3.3 hours or 198 minutes. This would help the hospital price scheduled MRI’s competitive with specialty MRI centers. However, the competition can cry foul and with good reason. The hospital is now subsidizing its scheduled MRI service with ER patients.

Chapter 6 of the dissertation proposes a third way to allocate the capacity. Using the steady state queuing results derived from the Kingman VUT approximation for G/G/1 queuing, we can determine how much capacity should be scheduled for non-urgent patients and how to allocate unused capacity between patients for pricing purposes. This MRI example has some features different from steady state queuing. Steady state queuing uses first-in first-out (FIFO) scheduling. A hospital will certainly prioritize the MRI machine to critical ER patients. However, the model should give insight and help develop intuition in scheduling and pricing high value resources, including MRI machines and operating rooms.

Other uses of this dissertation chapter extend to a graphic design firm with a steady workload stream from one customer and a much more varied and unpredictable demand from
another customer. Extra capacity will need to be set aside for the second customer and there
needs to be a fair and rational way to assign capacity costs. The analysis will also be
beneficial to a company that has two versions of the same product. The first version might
have a steady production stream while the second might be an older model that the firm
refuses to terminate support for. The model should provide better estimates of the true cost of
the second, older version and see if it should be retired if its marginal costs exceed the
marginal revenue of the line. Returning to health care, the model could provide capacity and
costing guidance for a physician or nurse with responsibilities for two or more departments
within a hospital.

Finally, simulation will be used to make sure that models yield appropriate results.
Using steady-state queuing is useful to help make the math less complicated. However, it is
important to note that startup effects will make the results less accurate when period length is
too short. Simulation will help provide guidance for setting minimum period lengths where
the steady state results prove more reliable.

Definitions:

\[ WIP \] total work-in-process inventory

\[ WIP_q \] work-in-process inventory waiting

\[ C_S \] customers / parts in system (waiting and being served)

\[ CT_q \] steady state waiting time

\[ \rho \] utilization = arrival rate / total service rate

\[ \lambda \] arrival rate

\[ \mu \] service rate
\( \sigma \)  standard deviation of arrival rate
\( \sigma_\mu \)  standard deviation of service rate
\( m \)  number of identical, parallel resources
\( c_a \)  coefficient of variation for arrivals
\( c_e \)  coefficient of variation for service
\( t_e \)  average time to perform service on a single resource

Clearing function definitions for our Part 1 model

\[
\text{CF} \quad \text{Clearing function}
\]
\[
i = 1, \ldots, \hat{i} \quad \text{product type index}
\]
\[
n = 1, \ldots, N \quad \text{machine index}
\]
\[
c = 1, \ldots, \hat{C} \quad \text{segment of the piecewise linear clearing function}
\]
\( p_i \)  contribution margin (profit) of one unit of product \( i \)
\( w_i \)  cost of holding one unit of product \( i \) as \( WIP \) for one time period
\( k_{in} \)  time required to process one unit of product \( i \) on machine \( n \)
\( C^n \)  theoretical capacity (in time units) of machine \( n \)
\( \alpha^n_c \)  the slope of CF section \( c \) on machine \( n \)
\( \beta^n_c \)  the intercept of CF section \( c \) on machine \( n \)
\( X_i \)  the number of units produced of product \( i \)
\( W_{in} \)  the average units of \( WIP \) for product \( i \) on machine \( n \)
\( Z_{in} \)  the fraction of maximum output that product \( i \) uses on machine \( n \)

Summary of research objectives
• Determine the economic meaning of the dual price of congested resources in clearing function applications.

• Fix research related to using the dual values of congested resources of clearing function formulations to allocate fixed overhead to products.

• Prove or disprove that clearing functions can be used to allocate fixed, untraceable overhead to products congruent with previously proposed overhead allocation by Kaplan and Thompson (1971) Rule 1 based on linear programming models.

• Prove or disprove that the allocation of traceable overhead using clearing functions is compatible with the LP overhead allocation schema proposed by Kaplan and Thompson (1971) Rule 2.

• Verify that overhead allocation when there are product sales or production interdependencies is compatible with Kaplan and Thompson’s schema.

• Transition the model to parallel machines. Derive clearing functions using G/G/m approximations and use them with Kaplan and Thompson (1971) overhead allocation rules.

• Use the dummy product, customer, or department approach to model systems where parallel machines have different service times or different variability.

• Use the results from different variability parallel machines to model customers with different variability in arrival rates. Determine overhead allocation to the different customers and use this to influence pricing for the different customers.

• Use clearing functions with service systems noting that work in process may not have a specific value so cycle time should be used instead.
• Verify clearing function models with simulation to see if the results are consistent and to determine period length necessary for steady state queuing derived clearing functions to be reasonable.
CHAPTER 2: Previous Related Work

This dissertation owes much of its start to the end of Kefeli’s (2011) dissertation, specifically Chapter 7 where he attempts to align clearing functions with Kaplan and Thompson (1971) use of linear programming and dual variables to allocate overhead. Unfortunately, an ill-timed algebraic mistake rendered his solution inadequate. However, much of the research is retraced as a basis for this dissertation.

The other main inspirations behind this work are Hopp and Spearman (2001) that introduced me to the Kingman approximation for G/G/1 queuing systems and applies lessons from this equation to help understand how variability impacts production processes, Asmundsson et al. (2009) where allocated clearing functions are developed, and Aucamp (1984) where the importance of marginal versus average costs are explained. The rest of this section is split among the various fields intersecting for the topics in this dissertation.

2.1 Accounting Research

Several textbooks were read to fully understand the basics of the various cost accounting and managerial accounting methods. My introduction to accounting is from Burch (1994). However, it was recommended that an updated classic Zimmerman (2009) text would explain full absorption costing and direct costing. Zimmerman was instrumental in understanding how better cost allocation can improve management outcomes. Because much of the first section is devoted to using clearing functions with linear programming, Kaplan’s (1989) textbook was thoroughly read to look for more insights about applications of linear programming in the field of accounting, especially in overhead allocation. Curiously, the 3rd edition of the textbook, Kaplan (1998) removes the linear programming and replaces the
chapters with activity-based costing as if the optimization models no longer existed. Other manuscripts devoted to using linear programming to improve accounting include Wright (1968) and Colantoni et al. (1969), the latter being an important predecessor to the Kaplan and Thompson (1971) paper featured prominently in Chapter 3.

The late 1980’s introduced a new type of accounting: throughput accounting, the accounting system envisioned by Goldratt (1986). A good summary of the technique is by Corbett (1998) with other texts by Smith (2000) and Noreen et al. (1995). Throughput accounting and linear programming techniques aligning the profit mix solution provide the same guidance for managers when there is a single binding constraint to production, but throughput accounting, using a greedy algorithm of ordering products based on contribution margin per unit time on a single constrained resource, fails to support the optimal product mix when there are multiple constrained resources.

Johnson and Kaplan (1991) provides a takedown of traditional full absorption accounting and how it can lead to bad production decisions. A good explanation of activity-based costing and its relationship with other accounting systems may be found in Kaplan and Cooper (1998). Driving activity-based costing is the realization that for many products, direct labor is declining relative to the total cost of the product. Combined with advances in computing power and the ability to identify better drivers for overhead allocation, activity-based costing is useful in identifying products that may look profitable but are actually subsidized by other products when using older techniques like direct labor hour allocation. Proponents of throughput accounting gleefully point to instances where using the more advanced activity-based costing methodology fail to support the optimal, most profitable product mix. The problem with activity-based costing is its use of average costs instead of
marginal costs. Ironic given that this is a fundamental part of what makes the Kaplan and Thompson (1971) formulation successful. An article in Kaplan’s (1990) compilation contains the first article I have found that acknowledges variability and queuing need to be a part of capacity and production planning from an accounting perspective. However, no tangible techniques were proposed.

Later, Kaplan and Anderson (2007) create a new accounting paradigm with time-driven activity-based costing. In it, process variability is attempted to be captured by time equations that can be used to price products, services, and customers better. Arrival variability is still not captured very well unless a more random customer requires additional resources than are needed to supply existing customers. In this case, several of the texts will allocate additional resource to only the more random customer without considering a steady customer still requires more capacity that the sum of service times due to slight variations in arriving orders.

2.2 Linear Programming Research

Linear programming has a rich history in production and operations. The optimal product mix problem has been used for years to tell production managers what to produce when limited by man or machine or raw materials. As rich as this history is, this paper is devoted to using linear programming to help with accounting and pricing. Kaplan has the most research devoted to using linear programming in the accounting field. Aucamp (1984) reminds accountants that average costs may not provide meaningful results when resources are in various states of utilization. An extra hour on a resource can be close to free if the operator and machine would otherwise be idle at regular wage rates. That extra hour will cost
much more if it requires overtime. Finally, in a 24/7 operation, another hour of a fully
constrained resource is the full value of what could be produced. The same machine,
therefore, can have multiple different potential costs of capacity.

Moghaddam and Michelot (2009) offer an interesting take on joint inputs to allocate
pricing to different products, in this case derivatives from oil refining.

A good example of adding combining costs and queuing is Banker et al. (1988). They
formulate a problem with \( m \) products sharing a production line and show how the nonlinear
effects of queuing as utilization approaches 100% causes WIP costs to increase dramatically
as each product is added to the production line. Buss Lawrence and Kropp (1994) combine
M/M/1 queuing and production planning to show how congestion effects reduce profitability
after a certain utilization level. They mathematically determine ranges beyond which there
are diseconomies of scale depending on how much congestion a production line encounters.

Clearing functions to model queuing effects for product mix optimization has been employed
to gain insight by several authors. Karmarkar (1989) provides the basis for the clearing
functions used in this dissertation, although alternative linear models exist, for example
Graves (1986). Asmundsson et al. (2009) further refines the problem to the allocated clearing
function approach used here where there can be multiple products on a multiple machine
production environment.

2.3 Medical Applications

Based on the MRI example used in the introduction, a brief look at accounting
systems for high capital cost equipment in the medical industry was investigated.

Unfortunately, until recently, papers on the subject proved less in-depth than what is
available for production. Schwartz (1985) writes about various factors influencing the price of MRI’s but does not consider variability in either process time or patient arrivals. A broader paper on cost accounting for the more general field of radiology was attempted by Camponovo (2004). Although more advanced cost accounting models are considered such as activity-based costing, variability is not. This paper reads less as an authority on how to integrate more advanced accounting methods to better cost and manage radiology departments, but as a plea for help in setting up better methods. Gavirneni and Kulkarni (2014) consider the economic costs of waiting in determining optimal bid prices for concierge medicine where patients willing to pay more can enjoy shorter lines. This work was inspired by Kleinrock (1967) who used bidding combined with queuing to determine the value of line position.

An active area of research for accounting in the medical industry uses time-driven activity-based costing. Yun et al. (2016) provides an implementation strategy and benefits to applying time-driven activity-based costing to emergency rooms. Curiously, one of the main benefits according to the authors is to identify and eliminate excess capacity. It is curious because emergency rooms are known for having high arrival, diagnostic, and process variability versus more specialized medicine. Much of this dissertation demonstrates problems when highly variable resources are driven at high capacities. Anzai et al. (2017) use time-driven activity-based costing to cost abdomen and pelvic CT scans. To get away from congestion effects due to queuing, practical capacities were calculated at 80-85% of the theoretical capacities. Also, despite using three patient types, in-patient, out-patient, and emergency department, the analysis just calculated process times. There was no discussion or consideration for waiting times due to arrival variability.
CHAPTER 3: Overhead Allocation Using Clearing Functions

Kefeli (2011) started this work as Chapter 7 of his PhD dissertation. It presents an overhead allocation scheme for a two product, three machine production process with variability. It attempts to show how the Kaplan and Thompson (1971) overhead allocation scheme could be applied to this system and calculate results from this allocation. When adding variance into the system, allocating the fixed overhead did not retain the optimal product mix solution from before the overhead allocation process. Attempting to prove this work, a key mathematical error was found. Going back to the Kaplan and Thompson (1971) paper, it was determined where Dr. Kefeli’s derivation went wrong and Kaplan and Thompson’s Rule 1 for allocating untraceable overhead was proved to be compatible with the clearing functions of Asmundsson et al. (2009).

Both Kaplan and Thompson (1971) and Uzsoy and Kefeli (2011) allocate fixed overhead based on the dual prices of the production resources. Both models have production capacities limiting sales. That is, all that is capable of being produced by the system will be sold at the price offered. The model from Kaplan and Thompson (1971) does not use the clearing functions used by Kefeli (2011) and thus fails to show the effects of WIP costs to production planning. However, it will be proved that the Kaplan-Thompson approach can still be used to allocate fixed overhead to the system without changing the optimal product mix and this will be shown using Kefeli (2011)’s high variance model.

A paper and presentation were prepared from this work, where the introduction and initial idea was written by Kefeli. The successful model and proof for Rule 1, the counter examples to Rule 2, the successful use of clearing functions on cross-subsidized products, and the presentation are from this author.
3.1 Overhead Allocation Using Traditional Clearing Function Formulation

Assume a firm has $N = 3$ resources (machines 1, 2 and 3) that produce $i = 2$ products (products 1 and 2). We use $\hat{C} = 3$ segments to represent the piecewise linearized CFs for each resource. Let $X_1$ and $X_2$ represent the number of units of products 1 and 2 to be produced in the period. The parameters of the numerical example are given in Table 3-1 and 3-2.

*Table 3-1: Numerical Example Common Coefficients*

<table>
<thead>
<tr>
<th>$k_{in}$</th>
<th>Machine 1</th>
<th>Product 1</th>
<th>Product 2</th>
<th>Capacity, $C^n$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Machine 1</td>
<td>3</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Machine 2</td>
<td>5</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Machine 3</td>
<td>0</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Per unit profit, $p_i$</td>
<td>1</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per unit WIP cost, $w_i$</td>
<td>0.1</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 3-2: Clearing Function Segments*

<table>
<thead>
<tr>
<th>Segment</th>
<th>Machine 1</th>
<th>Machine 2</th>
<th>Machine 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Machine 1</td>
<td>Machine 2</td>
<td>Machine 3</td>
</tr>
<tr>
<td>$\alpha^n$</td>
<td>1</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>$\beta^n$</td>
<td>0</td>
<td>2.0</td>
<td>9</td>
</tr>
</tbody>
</table>

A graph of the clearing functions for the three machines is provided in Figure 3-1.
Note that for all machines, the intercept of the last segment of the CF, $\beta^n_C$, is equal to its respective $C^n$ in the LP formulation, e.g., $\beta^1_1 = C^1 = 9$. This implies that the maximum throughput of the CF is equal to the capacity limit of the LP at high WIP levels (utilizations).

The classic formulation for a constrained resource linear programming optimization model is:

$$\text{Maximize } X_1 + 0.5X_2$$

Subject to:
The corresponding model using the clearing functions defined above is:

Maximize \((X_1 + 0.5X_2) - 0.1(W_{11} + W_{12} + W_{13} + W_{21} + W_{22} + W_{23})\)

Subject to:

\[
\begin{align*}
3X_1 &\leq (1)(3)W_{11} + (0)Z_{11} \\
3X_1 &\leq (0.4)(3)W_{11} + (2)Z_{11} \\
3X_1 &\leq (0)(3)W_{11} + (9)Z_{11} \\
5X_1 &\leq (1)(5)W_{12} + (0)Z_{12} \\
5X_1 &\leq (0.7)(5)W_{12} + (1.6)Z_{12} \\
5X_1 &\leq (0)(5)W_{12} + (10)Z_{12} \\
0X_1 &\leq (1)(0)W_{13} + (0)Z_{13} \\
0X_1 &\leq (0.6)(0)W_{13} + (1.6)Z_{13} \\
0X_1 &\leq (0)(0)W_{13} + (5)Z_{13} \\
2X_2 &\leq (1)(2)W_{21} + (0)Z_{21} \\
2X_2 &\leq (0.4)(2)W_{21} + (2)Z_{21} \\
2X_2 &\leq (0)(2)W_{21} + (9)Z_{21} \\
0X_2 &\leq (1)(0)W_{22} + (0)Z_{22} \\
0X_2 &\leq (0.7)(0)W_{22} + (1.6)Z_{22} \\
0X_2 &\leq (0)(0)W_{22} + (10)Z_{22} \\
4X_2 &\leq (1)(4)W_{23} + (0)Z_{23} \\
4X_2 &\leq (0.6)(4)W_{23} + (1.6)Z_{23} \\
4X_2 &\leq (0)(4)W_{23} + (5)Z_{23}
\end{align*}
\]

\[
\begin{aligned}
Z_{11} + Z_{21} &= 1 \\
Z_{12} + Z_{22} &= 1 \\
Z_{13} + Z_{23} &= 1 \\
\end{aligned}
\]

\[X_1, X_2, W_{11}, W_{12}, W_{13}, W_{21}, W_{22}, W_{23}, Z_{11}, Z_{12}, Z_{13}, Z_{21}, Z_{22}, Z_{23} \geq 0\]
Using the standard Solver in Microsoft Excel, the optimal values of the main decision variables are identical in both formulations - both formulations yield \( X_1^* = 2 \) and \( X_2^* = 1.25 \). This congruence of results will, of course, not always occur given different constraints or variable coefficients. Should there be large penalties for WIP, it is natural that the LP model that ignores WIP costs should yield a different result than one that accounts for WIP costs.

The optimal objective function value for the LP formulation is \( G_{LP}^* = 2.625 \) and for the CF formulation is \( G_{CF}^* = 1.625 \). The difference in the objective functions values stems from the additional WIP cost included in the CF formulation. The optimal values of the decision variables in the CF formulation are given in Table 3-3.

\[ X_i^* \]

Before moving to overhead cost allocation, we can examine the optimal values of the dual variables associated with the capacity constraints in both models. We denote the dual variables of the machines that correspond to constraints in the LP model by \( \omega_n^{LP} \), and those corresponding to the Z constraints in the CF model by \( \omega_n^{CF} \) (the other duals are all zero, which is a function of the CF setup). The optimal values of these dual variables are in Table 3-4.

### Table 3-3: Model Output from CF Formulation

<table>
<thead>
<tr>
<th></th>
<th>Product 1</th>
<th>Product 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_{in}^* )</td>
<td>Machine 1</td>
<td>3.889</td>
</tr>
<tr>
<td></td>
<td>Machine 2</td>
<td>2.400</td>
</tr>
<tr>
<td></td>
<td>Machine 3</td>
<td>0.000</td>
</tr>
<tr>
<td>( Z_{in}^* )</td>
<td>Machine 1</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>Machine 2</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Machine 3</td>
<td>0.000</td>
</tr>
<tr>
<td>( X_i^* )</td>
<td>2</td>
<td>1.25</td>
</tr>
</tbody>
</table>
Table 3-4: Duels of the Optimal Product Mix for Both the LP and CF Formulation

<table>
<thead>
<tr>
<th>Machine</th>
<th>$\omega_{n}^{LP}$</th>
<th>$\omega_{n}^{CF}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine 1</td>
<td>0</td>
<td>0.250</td>
</tr>
<tr>
<td>Machine 2</td>
<td>0.200</td>
<td>1.204</td>
</tr>
<tr>
<td>Machine 3</td>
<td>0.125</td>
<td>0.171</td>
</tr>
</tbody>
</table>

The result reveals an inherent advantage of the CF model over the LP model in its representation of limited resource capacity. LP duality theory implies that a dual variable will only take a non-zero value when the associated capacity constraint is tight at optimality. In other words, for the LP formulation to yield positive dual prices for capacity, utilization of that resource should be equal to one, implying that there is no benefit to the objective function from adding capacity to resources that are not fully utilized. On the other hand, the CF formulation’s optimal solution frequently results in some of the throughput constraints being tight at optimality (except in cases of degeneracy). In our example, machine 1’s utilization is given by:

$$\rho = \frac{\sum_{i=1}^{2} k_{i}X_{i}^{*}}{C^{i}} = \frac{(3)(2)+(2)(1.25)}{9} = \frac{8.5}{9} = 0.94 = 94\%$$

which is less than 100%, suggesting positive slack in the capacity constraints. Hence the optimal LP solution yields $\omega_{1}^{LP} = 0$, implying that additional capacity at machine 1 would be of no benefit to the system. On the other hand, as evidenced by the positive dual price obtained in the CF formulation, i.e., $\omega_{1}^{CF} = 0.250$, additional capacity for machine 1 should improve the objective function value and hence be beneficial for the system. Machine 1 is a near–bottleneck for the system in our numerical example, while machines 2 and 3 are full bottleneck machines. In fact, a closer look at the machine constraint duals shows that near-
bottleneck machine 1 has a larger dual price than the fully constrained machine 3 (\(\omega_1^{CF} = 0.250\) versus \(\omega_3^{CF} = 0.171\)).

Another difference between the models is the nonlinear nature of the cost of WIP. Queuing suggests that the average WIP in a queuing system will increase nonlinearly with the utilization. Thus, one would expect the dual price of capacity, the marginal benefit of additional capacity, to increase nonlinearly with utilization, as Kefeli et al. (2011) have shown to be the case. Classical parametric analysis of linear programs shows, in contrast, that for constraints similar to those in our LP formulation, the objective function value is a piecewise linear function of the right-hand side of the constraint, implying piecewise linear rather than nonlinear benefit of additional capacity. To observe this effect, note in our table of dual values that in the optimal LP solution, each unit of machine 2, which is used exclusively by product 1, is 60% more valuable than machine 3 (\(\omega_2^{LP} = 0.2\) and \(\omega_3^{LP} = 0.125\)) making machine 2 2.2 times more valuable than machine 3 (\(\omega_2^{LP}b = 2\) and \(\omega_3^{LP}b = 0.625\)). In contrast, in the CF formulation, machine 2 is \(\approx 7\) times as valuable as machine 3 (\(\omega_2^{CF} = 1.204\) versus \(\omega_3^{CF} = 0.171\)). This is partly due to the shape of the CF of machine 2, which exhibits a higher congestion effect than that of machine 3 (represented in the slopes of the CFs). But the shape of the CF is not the only reason for the difference. The network structure of the machine routings has an effect on the value of dual prices, i.e., the dual price of one machine depends on the utilization and dual price of other machines that share the same product with it. Nevertheless, given the increase in the dual price of machine 2 relative to that for machine 3, product 1 should absorb more overhead cost than product 2.
3.2 Where the Kefeli (2011) Model Goes Wrong

Kefeli (2011, Chapter 7) overhead allocation scheme does not generate the same results as this Kaplan-Thompson (1971) application. It is suspected that the dissertation chapter over costs capacity utilization. Kefeli proposes calculating production costs $V_{in}$ and $V_i$ based on the following relationships:

$$V_{in} = k_{in}X_i\omega_n \quad \text{and} \quad V_i = \sum_{n=1}^{N} V_{in}$$

Kefeli (2011) defines a new variable $v_i$ as the relative cost of producing $X_i$ units of product $i$ to the cost of producing all other products. This relative cost $v_i$ is defined by:

$$v_i = \frac{V_i}{\sum_{i=1}^{N} \sum_{n=1}^{N} k_{in}X_i\omega_n}$$

Where $0 \leq v_i \leq 1$. $v_i$ is to represent “the fraction of the total overhead cost allocated to $X_i$ units of product $i$.” Thus, given the relative costs $v_i$ for each product $i$, the amount of overhead $H_i$ allocated to product $i$ is given by:

$$H_i = Hv_i$$

Kefeli (2011) invokes the strong duality theorem of LP optimality to introduce the heart of where the analysis goes awry. While it is true “that in an optimal solution, the optimal values of the primal and dual problems will be equal, implying that marginal profit should equal marginal cost at optimality,” the following formula is incorrect:

$$\sum_{n=1}^{N} k_{in}X_i\omega_n = p_iX_i$$
The real relationship between the primal and dual at optimality is \( pX^* = \omega^* b \). By simple algebra substitution, the Kefeli (2011) dissertation model seems to assert that at optimality, \( \sum_{i=1}^{N} k_i X_i = b \) which is clearly not the case for this example. The transpose of the \( b \) vector in this case is (0,…,0,1,1,1), not the sum of machine times \( k_i \) multiplied by the number of \( X_i \) produced. It is no surprise, then, that the model fails to maintain the optimal product mix for many of the problems where this method has been applied.

3.3 Allocating Untraceable Overhead

Having solved the mathematical models to optimality and calculated the dual variables associated with the production resources, we now proceed to cost allocation. We assume that we have \( H = 1 \) units of total untraceable overhead cost to allocate to our optimal product mix of 2 units of product 1 and 1.25 units of product 2. Following Kaplan and Thompson’s (1971) Rule 1 for allocating untraceable overhead, we define the ratio of overhead costs to variable contribution

\[ k = \frac{H}{G^*} \quad \text{where} \quad G^* = px^* = \omega^* b \]

obtaining

\[ k_{LP} = \frac{1}{2.625} = \frac{8}{21} \approx 0.381 \]

and

\[ k_{CF} = \frac{1}{1.625} \approx 0.615 \]

Note that when \( k \geq 1 \) overhead costs exceed variable contribution meaning that the results of the optimal product mix are incapable of fully absorbing the overhead because the products
can only absorb $G$ units. The difference in variable contributions is due to the CF model’s explicit representation of WIP holding costs, which are omitted in the LP model. To allocate the one unit of untraceable overhead, the coefficients are changed using $k$ so the new objective function is, using matrix notation,

$$\text{Maximize } (p - k \omega^* A)x$$

Due to complimentary slackness at optimality, either $\omega^* A_j = p_j$ or $x_j = 0$ for each column $j$ in matrix $A$. Our equation can thus be rewritten, after some algebra,

$$\text{Maximize } (1 - k)px$$

We can calculate the new coefficients of the objective function of the LP formulation as follows:

$$p_1 = \left(1 - \frac{8}{21}\right)(1) = \frac{13}{21}, \quad p_2 = \left(1 - \frac{8}{21}\right)(0.5) = \frac{13}{42}$$

The LP objective function that allocates untraceable overhead according to Rule 1 is thus:

$$\text{Maximize } \frac{13}{21}x_1 + \frac{13}{42}x_2$$

subject to the same equations for the basic LP problem as in section 3.1. Likewise, we can use the same methodology to calculate the new objective function for the CF:

$$p_1 = (1 - 0.615)(1) = 0.385, \quad p_2 = (1 - 0.615)(0.5) = 0.1925, \quad w = (1 - 0.615)(-0.1) = -0.0385$$

Therefore, we get:

$$\text{Maximize } 0.385x_1 + 0.1925x_2 - 0.0385\left(W_{11} + W_{12} + W_{13} + W_{21} + W_{22} + W_{23}\right)$$

subject to the same constraints as the CF formulation in section 3.1. Running both the LP and CF formulations, we see that none of the values for the decision variables changed in either formulation.
Proof: The decision variables remain unchanged because both the feasible regions are unchanged, and the new objective functions are parallel to the original objective functions. This is clearly shown by the equation above where the new objective functions are the old objective functions scaled by the constant \((1 - k)\).

The LP formulation now shows a net profit of 1.625 and the CF formulation 0.625, each reflective of the one unit of untraceable overhead allocated between \(X_1\) and \(X_2\). Kaplan and Thompson show that the amount of fixed overhead applied to each unit of production of the LP formulation is \(k_{LP} \omega^* A\).

\[
k_{LP} \omega^* A = \frac{8}{21} \begin{bmatrix} 0 & 1 & 1 & 0 \ 5 & 8 & 3 & 2 \ 0 & 0 & 5 & 0 \ 4 & 4 & 0 & 4 \end{bmatrix} \frac{8}{21} \begin{bmatrix} 4 \ 21 \ 21 \ 0 \end{bmatrix}
\]

Multiplying by the optimal values of \(x\) from the primal, we obtain the total fixed overhead absorbed by each product: \(X_1 = \frac{8}{21} \cdot 2 = \frac{16}{21}, \ X_2 = \frac{4}{21} \cdot \frac{5}{4} = \frac{5}{21}\)

which, of course, sum to one unit.

Before allocating fixed overhead for the CF formulation, we must first recognize that the dual variables from the LP formulation are different from the dual variables for the CF formulation. Recall that from strong duality,

\[
px^* = \omega^* b
\]

where \(px^*\) is the profit vector per unit (contribution margin) multiplied by the total units of each product produced at optimality. The right-hand side, \(\omega^* b\) is the value of capacity at optimality. For the original LP formulation, the dual variables represent the value per unit of additional resource. Recall from our table of dual values that each additional unit of machine 2 is worth up to 0.20 of value. Similarly, each additional unit of machine 3 is worth up to
0.125. We can calculate the value of capacity from the original LP formulation by multiplying 0.20 by 10 units of the machine 2 used plus 0.125 times 5 units of the machine 3 used. Not surprisingly, the value of capacity equals the planned profit of 2.625.

The duals of the CF formulation represent something different. These duals are the value of capacity for each of the machines, not on a per unit basis. A nice feature of the CF formulation is that the equations of the primal problem can be rearranged such that only the $Z_{in}$ variable constraints have nonzero right-hand side values (which happen to be 1) for the $b$ vector. Another nice feature is that the only interdependencies of the products are in the $Z_{in}$ variable constraints. Therefore, the value of capacity is simply $\omega^*b = \omega^*$ for this CF formulation. We can allocate fixed untraceable overhead to products by allocating overhead based on each machine’s capacity value and the percentage of capacity value generated by each product. Because the only equations that intermingle the products are the $Z_{in}$ decision variables and that the columns on the $A$ matrix representing these $Z_{in}$ variables are trivial, we take advantage of symmetry in the $A$ matrix with respect to the $Z_{in}$ variables and apply $k_{CF}\omega^*A$ to get the overhead allocated to each machine.

$$
k_{CF}\omega^*A = (0.615) \begin{bmatrix} 0.25 & 1.204 & 0.171 \\ 0.154 & 0.741 & 0.105 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} =
\begin{bmatrix} 0.154 & 0.741 & 0.105 & 0.154 & 0.741 & 0.105 \end{bmatrix}
$$

We multiply the overhead allocated to each machine by the $Z_{in}$ decision variables assigned to each product to obtain the overhead allocated to each product.
Because $k_{CF} \omega^* A x^* = k_{CF} \omega^* b$ at optimality, we can further simplify our method. For the CF formulation, because $b = 1$, when there are no product sales or production interdependencies the calculation simplifies to

\[
\text{Fixed overhead allocated to product } X_i = k_{CF} \sum_n Z_{in} \omega_{n,CF}^* 
\]

where the $Z_{in}$ variables show the proportion of machine $n$ overhead to be applied to each product $i$. Returning to our numerical example, we see that

\[
\begin{align*}
X_1 \text{ total fixed overhead absorbed} & = 0.615 \left( \frac{2}{3} \right) (0.250) + (1)(1.204) + (0)(0.171) \approx 0.843 \\
X_2 \text{ total fixed overhead absorbed} & = 0.615 \left( \frac{1}{3} \right) (0.250) + (0)(1.204) + (1)(0.171) \approx 0.156
\end{align*}
\]

which is only slightly different to our previous result due to rounding errors. Notice this overhead allocation can proceed directly from the initial problem.

Another way to allocate the amount of fixed overhead applied to each product is to recognize that the overhead applied to each product is the per unit overhead applied multiplied by the number of units produced at optimality, or $k \omega^* A x^*$. Due to complementary slackness at optimality, the equation is reduced to simply $k p x^*$. For the LP formulation,

\[
\begin{align*}
X_1 \text{ total fixed overhead absorbed} & = \frac{8}{21} (1)(2) = \frac{16}{21} \\
X_2 \text{ total fixed overhead absorbed} & = \frac{8}{21} (0.5)(\frac{5}{4}) = \frac{5}{21}
\end{align*}
\]
Calculating the overhead allocated to each product in the CF formulation is more complicated. There is a key distinction between the LP and CF results – the new CF formulation scales the coefficients of WIP charges by \((1 - k_{CF})\). However, a closer look at the results of the new CF formulation shows that throughput is reduced by 1.615 units while the total WIP charges due to scaling are reduced by 0.615 units. Subtracting the difference, we are left with the 1 unit of fixed untraceable overhead that we allocated to products \(X_1\) and \(X_2\). The missing WIP charges are fully accounted for in the equivalent lost throughput of the new formulation. Recall that in our CF formulation, WIP levels are objective function decision variables. We can determine how the fixed overhead is allocated to products using a slightly more complex procedure.

\[
\text{Overhead absorbed by product } X_i = k_{CF} \left( p_i X_i^* - \sum_n w_i W_{in}^* \right)
\]

Numerically in our example, the overhead absorbed by each product is calculated as follows:

\[
\begin{align*}
X_1 \text{ total fixed overhead absorbed} &= 0.615 \left( (1)(2) - (0.1)(3.889 + 2.400 + 0) \right) \approx 0.843 \\
X_2 \text{ total fixed overhead absorbed} &= 0.615 \left( (0.5)(1.25) - (0.1)(2.292 + 0 + 1.417) \right) \approx 0.156
\end{align*}
\]

Notice this overhead allocation also can proceed directly from the initial problem. There is no need to set up the second CF linear program to determine overhead allocation. One just needs the results of the first CF formulation.

3.4 Allocating Traceable Overhead

Satisfied with the allocation of untraceable overhead using Kaplan Thompson Rule 1, the next step is allocating traceable overhead. Recall that the duals are important as they provide limits on how much overhead can be applied to each product. Should any dual be
reduced below zero, the solution to either formulation will no longer remain optimal. (In fact, the dual’s problem structure explicitly forbids this.) If a machine’s marginal contribution becomes negative, the solution would redirect the production mix to favor products using machines that have positive contribution, barring interdependencies of the demand for each product.

Kaplan Thompson Rule 2 allocates overhead that can be traced directly to resources among the products. The traceable overhead assigned to a resource \( n \) is divided by the amount of the resource quantity \( b_n \) and assigned to variable \( B_n \). The traceable overhead is then applied to the resource using the following rule:

\[
\forall n \text{ such that } \omega_n' \geq B_n, \text{ assign a traceable overhead charge of } B_n' = B_n
\]
\[
\forall n \text{ such that } \omega_n' < B_n, \text{ assign a traceable overhead charge of } B_n' = \omega_n'
\]

The new objective function, in matrix notation, is

\[
\text{Maximize } (p - B'A)x
\]

This rule enforces nonnegative duals by limiting the amount of traceable overhead that can be applied to any resource. Any remaining traceable overhead is added to the pool of untraceable overhead which is then allocated by Rule 1. To apply Rule 1, Kaplan and Thompson define a new ratio of overhead costs to variable contribution, \( k^* \) using the following:

\[
k^*_LP = \frac{H}{(p - B'A)x^*}
\]

where \( H \) is the remaining common overhead, or more specifically, the sum of the untraceable overhead and any traceable overhead exceeding the marginal cost of the resource from the original problem. Taking the capacity value route, we can calculate \( k^* \) as:
\[ k_{LP}^* = \frac{H}{(\omega^* - B')b} \]

Notice that the denominator stays positive based on Rule 2. The combined objective function in matrix notation, after some algebra, is

\[
\text{Maximize } \left( (p - B'A)(1 - k_{LP}^*) \right)x
\]

The CF version will be different. Recall that the duals in the CF formulation represent the total marginal value of the resource, not the marginal value per unit of additional capacity. Therefore, \( B_n \) is now the total traceable overhead to be applied to a resource, not the per unit overhead. The \( B' \)’s are still calculated by Rule 2. For Rule 1, we were able to use the relationship \( \omega^* A = p \) to calculate the new coefficients of the objective function. For Rule 2, it is unclear what corresponds to \( B'A \) in the CF formulation. Recall that the dual variables from the CF formulation are associated with the \( Z_{in} \) constraints which allocate the time used by product \( i \) on resource \( n \). If we apply the traceable overhead based on the \( Z_{in} \) variables, we still have a problem. We need to use the results of the initial CF formulation to determine the new coefficients of the new objective function. WIP charges are still a function of resource usage based on the product mix. Leaving the product decision variables free to recalculate, the revised objective function for the CF model is then

\[
\text{Maximize } \sum_i \left( p_i - \frac{1}{X_i} \sum_n Z_{in}^* B_n' \right) X_i - \sum_i w_i \sum_n W_{in}
\]

The \( k^* \) calculation for the CF formulation looks more complex using contribution margin because we have to allocate the overhead to the specific resource, to the products that use the resource, in the proportion the products use it, and account for the WIP charges. However, the calculation is just the remaining common overhead over the remaining profit. Because the
formulation recalculates, the original optimal values are marked with * and the new optimal values are marked with o.

\[
k_{CF}^* = \frac{H}{\sum_i \left( p_i - \frac{1}{X_i} \sum_n Z_{in}^* B_{n}^* \right) X_i^o - \sum_i w_i \sum_n W_{in}^o}
\]

Using the remaining capacity value is an easier way to calculate \( k_{CF}^* \). WIP is already “baked into” the duals and the \( b \) parameters are all equal to 1, yielding:

\[
k_{CF}^* = \frac{H}{\sum_n \left( \omega_n^* - B_n^* \right)}
\]

Finally, the objective function with the combined traceable and untraceable overhead is

\[
\text{Maximize } \sum_i \left( p_i - \frac{1}{X_i} \sum_n Z_{in}^* B_{n}^* \right) \left( 1 - k_{CF}^* \right) X_i - \sum_i w_i \left( 1 - k_{CF}^* \right) \sum_n W_{in}
\]

Returning to the numerical example, instead of allocating untraceable overhead as before, it is determined that the overhead is really fixed and assignable to each machine. Recall our dual variables at optimality and calculate the capacity value for each machine reprinted in Table 3-5.

\textit{Table 3-5: Contribution Margin of the Optimal Product Mix for Both LP and CF Model}

<table>
<thead>
<tr>
<th>Machine</th>
<th>( \omega_{LP}^n b )</th>
<th>( \omega_{CF}^n (b = 1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine 1</td>
<td>0</td>
<td>0.250</td>
</tr>
<tr>
<td>Machine 2</td>
<td>2.000</td>
<td>1.204</td>
</tr>
<tr>
<td>Machine 3</td>
<td>0.625</td>
<td>0.171</td>
</tr>
</tbody>
</table>

For this example, say our 1 unit of overhead is traceable to machine 2. Because 1 unit of traceable overhead is less than the value of machine 2 capacity for both the LP and CF
model, we should see that both formulations are able to allocate the traceable overhead to the products that use machine 2 and that the optimal product mix is maintained.

Let’s begin with the LP formulation. \( 1/b_2 = 0.1 = B_2' \). Recall we get \( A \) from our machine constraints:

\[
\begin{align*}
3X_1 + 2X_2 &\leq 9 \quad \text{Machine 1 constraint} \\
5X_1 &\leq 10 \quad \text{Machine 2 constraint} \\
4X_2 &\leq 5 \quad \text{Machine 3 constraint}
\end{align*}
\]

Note that only product 1 gets produced by machine 2. So, for the LP formulation, our new objective function is calculated by

\[
\text{for } X_1 \rightarrow 1 - 0.1(5) = 0.5, \quad \text{for } X_2 \rightarrow 0.5 - 0.1(0) = 0.5
\]

Our new objective function for the LP formulation is thus

\[
\text{Maximize } 0.5X_1 + 0.5X_2
\]

subject to the machine capacity constraints and nonnegative values for our decision variables constraints. The new value of the objective function is 1.625 and the values of \( X_1 \) and \( X_2 \) remain 2 and 1.25 respectfully. The new duals are \( \omega^o_1 = 0, \ \omega^o_2 = 0.1, \ \text{and } \omega^o_3 = 0.125 \). The overhead absorbed is

\[
\begin{align*}
X_1 \text{ total fixed overhead absorbed } &= (B'A)_1^o = (0.1 \times 5) \times 2 = 1 \\
X_2 \text{ total fixed overhead absorbed } &= (B'A)_2^o = (0.1 \times 0) \times 1.25 = 0
\end{align*}
\]

using the product contribution margin. Knowing that all of the overhead of machine 2 gets absorbed by only product 1, the overhead absorbed by product 1 can be calculated as just \( B'b \), or the difference \( \omega^*b - \omega^o b = 0.2(10) - (0.1) (10) = 1 \).

The revised CF formulation requires the use of the optimal values of the product variables \( X_i \). The new coefficients for the product variables are calculated as follows:
\[
\text{coefficient for } X_1 \rightarrow \left( p_1 - \frac{1}{X_1} B'_2 Z^*_{12}\right) = \left( 1 - \frac{1}{2} (1)(1) \right) = 0.5
\]

\[
\text{coefficient for } X_2 \rightarrow \left( p_2 - \frac{1}{X_2} B'_2 Z^*_{22}\right) = \left( 0.5 - \frac{1}{1.25} (1)(0) \right) = 0.5
\]

Noting that the WIP variables are unchanged, the new objective function is thus:

Maximize \(0.5X_1 + 0.5X_2 - 0.1(W_{11} + W_{12} + W_{13} + W_{21} + W_{22} + W_{23})\)

subject to constraints in equations machine clearing function constraints, the \(Z\) constraints, and the nonnegativity constraints. The results in this case are predictable. The decision variables all remain unchanged with the new objective function value of 0.625, one unit less than the original CF formulation. The new duals are \(\omega^o_1 = 0.250\), \(\omega^o_2 = 0.204\), and \(\omega^o_3 = 0.171\). The total traceable overhead absorbed can be calculated as

\[
X_1 \text{ total fixed overhead absorbed} = \left( \frac{1}{X_1} B'_2 Z^*_{12}\right) X^o_1 = \left( \frac{1}{2} (1)(1) \right) 2 = 1
\]

\[
X_2 \text{ total fixed overhead absorbed} = \left( \frac{1}{X_2} B'_2 Z^*_{22}\right) X^o_2 = \left( \frac{1}{1.25} (1)(0) \right) 1.25 = 0
\]

or using the duals, given that the decision variables remain unchanged, by

\[
X_1 \text{ total fixed overhead absorbed} = (\omega^*_{1} - \omega^o_{1}) Z^*_{11} + (\omega^*_{2} - \omega^o_{2}) Z^*_{12} + (\omega^*_{3} - \omega^o_{3}) Z^*_{13}
\]
\[= (0.250 - 0.250) \frac{2}{3} + (1.204 - 0.204)1 + (0.171 - 0.171)0 = 1
\]

\[
X_2 \text{ total fixed overhead absorbed} = (\omega^*_{1} - \omega^o_{1}) Z^*_{21} + (\omega^*_{2} - \omega^o_{2}) Z^*_{22} + (\omega^*_{3} - \omega^o_{3}) Z^*_{23}
\]
\[= (0.250 - 0.250) \frac{1}{3} + (1.204 - 0.204)0 + (0.171 - 0.171)1 = 0
\]

The next step is to add some untraceable overhead to the problem and see how each formulation performs. Suppose our untraceable overhead is 0.4 units. For the LP formulation, we first calculate \(k^{*}_{LP}\) which is 0.4 divided by the objective function value after applying the
traceable overhead (which is 1.625). Therefore, $k^*_{\text{LP}} = 16/65$. We use this result and equation (29) to calculate the new coefficients

\[
\text{coefficient for } X_1 \rightarrow (1 - 0.1(5))\left(1 - \frac{16}{65}\right) = \frac{49}{130} = p'_1
\]

\[
\text{coefficient for } X_2 \rightarrow (0.5 - 0.1(0))\left(1 - \frac{16}{65}\right) = \frac{49}{130} = p'_2
\]

Our new objective function for this combined LP formulation is thus

\[
\text{Maximize } \frac{49}{130} X_1 + \frac{49}{130} X_2
\]

subject again to the machine capacity constraints and product nonnegativity constraints. The new value of the objective function is 1.225 and the values of $X_1$ and $X_2$ remain 2 and 1.25 respectfully. The total fixed overhead absorbed by product, both traceable and untraceable, is

\[
X_1 \text{ total fixed overhead absorbed } = \left(p_1 - p'_1\right)X_1^* = \left(1 - \frac{49}{130}\right)2 = \frac{81}{65} = 1.246
\]

\[
X_2 \text{ total fixed overhead absorbed } = \left(p_2 - p'_2\right)X_2^* = \left(0.5 - \frac{49}{130}\right)1.25 = \frac{2}{13} = 0.154
\]

using the product contribution margins.

Allocating 0.4 units of untraceable overhead in addition to the 1 unit of traceable overhead for the CF formulation first requires the calculation of a new $k^*_{\text{CF}}$. Recall the value of the objective function after allocating the traceable overhead in the CF formulation is 0.625. Therefore, $k^*_{\text{CF}} = 0.4/0.625 = 0.64$. We calculate the coefficients in the objective function by using our equation and our newly calculated $k^*_{\text{CF}}$

\[
\text{coefficient for } X_1 \rightarrow \left(p_1 - \frac{1}{X_1}B_2'Z_{i2}\right)\left(1 - k^*_{\text{CF}}\right) = \left(1 - \frac{1}{2}(1)(1)\right)(1 - 0.64) = 0.18
\]

\[
\text{coefficient for } X_2 \rightarrow \left(p_2 - \frac{1}{X_2}B_2'Z_{i2}\right)\left(1 - k^*_{\text{CF}}\right) = \left(0.5 - \frac{1}{1.25}(1)(0)\right)(1 - 0.64) = 0.18
\]

\[
\text{coefficient for } w_i \rightarrow (w_i)(1 - k^*_{\text{CF}}) = (0.1)(1 - 0.64) = 0.036
\]
The new objective function is thus

\[
\text{Maximize } 0.18X_1 + 0.18X_2 - 0.036\left(W_{11} + W_{12} + W_{13} + W_{21} + W_{22} + W_{23}\right)
\]

subject to the machine clearing function constraints, the \(Z\) constraints, and the nonnegativity constraints. As expected, our new objective function value is 0.225 and the decision variables are unchanged. The overhead absorbed by each product is more easily calculated using the duals:

\[
X_1 \text{ total fixed overhead absorbed } = \left(\omega_1^* - \omega_1^v\right)Z_{11}^* + \left(\omega_2^* - \omega_2^v\right)Z_{12}^* + \left(\omega_3^* - \omega_3^v\right)Z_{13}^*
\]

\[
= \left(0.25 - 0.09\right)\frac{2}{3} + (1.2044 - 0.0736)1 + (0.1708 - 0.0615)0 \approx 1.2374
\]

\[
X_2 \text{ total fixed overhead absorbed } = \left(\omega_1^* - \omega_1^v\right)Z_{21}^* + \left(\omega_2^* - \omega_2^v\right)Z_{22}^* + \left(\omega_3^* - \omega_3^v\right)Z_{23}^*
\]

\[
= \left(0.25 - 0.09\right)\frac{1}{3} + (1.2044 - 0.0736)0 + (0.1708 - 0.0615)1 \approx 0.1626
\]

We can use fractions to get more exact answers.

It would certainly be nice if we could prove that Kaplan and Thompson’s Rule 2 works for all CF formulations. However, it is easier to provide a counterexample that shows that it does not. This numerical example for the LP formulation still provides the same product mix even when the coefficients for the products approach zero for this problem so long that the fixed traceable overhead applied to product 1 does not increase beyond a tipping point. Let us try to assign 1.16 units of overhead to machine 2. A quick look at the duals would show that 1.204 > 1.16 so the formulation should maintain the optimal product mix.

\[
\begin{align*}
\text{coefficient for } X_1 & \rightarrow \left(\frac{p_1}{X_1} - \frac{1}{X_1}B_2 Z_{12}^*\right) = \left(1 - \frac{1}{2}(1.16)(1)\right) = 0.42 \\
\text{coefficient for } X_2 & \rightarrow \left(\frac{p_2}{X_2} - \frac{1}{X_2}B_2 Z_{22}^*\right) = \left(0.5 - \frac{1}{1.25}(1)(0)\right) = 0.5
\end{align*}
\]

\[
\text{Max}\left(0.42X_1 + 0.5X_2 - 0.1\left(W_{11} + W_{12} + W_{13} + W_{21} + W_{22} + W_{23}\right)\right)
\]
subject to the machine clearing function, \( Z \), and nonnegativity constraints. Unfortunately, this analysis gives a new optimal product mix. The old values of the decision variables are compared with the new values in Table 3-6 below:

\[ \text{Table 3-6: Optimal Values Between CF Models} \]

<table>
<thead>
<tr>
<th></th>
<th>( W^*_n )</th>
<th>( Z^*_n )</th>
<th>( X^*_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W^*_n ) Machine 1</td>
<td>3.89</td>
<td>0.67</td>
<td>2.00</td>
</tr>
<tr>
<td>Machine 2</td>
<td>2.40</td>
<td>1.00</td>
<td>1.25</td>
</tr>
<tr>
<td>Machine 3</td>
<td>0.00</td>
<td>1.42</td>
<td>1.07</td>
</tr>
<tr>
<td>( Z^*_n ) Machine 1</td>
<td>2.07</td>
<td>0.33</td>
<td>1.00</td>
</tr>
<tr>
<td>Machine 2</td>
<td>1.07</td>
<td>0.00</td>
<td>1.07</td>
</tr>
<tr>
<td>Machine 3</td>
<td>1.51</td>
<td>1.42</td>
<td>1.25</td>
</tr>
</tbody>
</table>

The new duals differ from those expected and are shown in Table 3-7:

\[ \text{Table 3-7: Comparison of CF Formulation Duals} \]

<table>
<thead>
<tr>
<th></th>
<th>Expected (( \omega^{CF}_n ))</th>
<th>( \omega^{CF'}_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine 1</td>
<td>0.250</td>
<td>0.250</td>
</tr>
<tr>
<td>Machine 2</td>
<td>0.044</td>
<td>0.045</td>
</tr>
<tr>
<td>Machine 3</td>
<td>0.171</td>
<td>0.171</td>
</tr>
</tbody>
</table>

Solver found a solution with an objective function value slightly higher than the one expected based on how Rule 2 was thought to operate. This is due to the WIP charges and the way the products’ WIP interact on the shared resource machine 1. Solver found a solution where producing less provided a higher objective function based on the overhead adjusted coefficient for product 1. The reduction in WIP costs was more valuable to the maximized objective function than the lost throughput from product \( X_1 \) with its traceable overhead reduced coefficient. The new solution fails to absorb the traceable overhead, however. Recall the calculation of overhead absorbed by product gives us
\[
X_1 \text{ total fixed overhead absorbed } = \left( \frac{1}{X_1} B_2^T Z_{12}^* \right) X_1^* = \left( \frac{1}{2} (1.16)(1) \right) 1.07 \approx 0.62
\]

\[
X_2 \text{ total fixed overhead absorbed } = \left( \frac{1}{X_2} B_2^T Z_{22}^* \right) X_2^* = \left( \frac{1}{1.25} (1.40)(0) \right) 1.25 = 0
\]

Using the duals to calculate how much overhead is absorbed by the new solution does not work given that the decision variables have changed. We see that our traceable fixed overhead absorbed is 0.62 units which is less than our expected value of 1.16 units.

A first hypothesis as to why Rule 2 fails under some circumstances is that this simple clearing function model may be introducing error due to trying to represent a curved clearing function with a small number of line segments. The solution to the piecewise linear clearing function linear program will be a corner point solution. At each corner point, the model will most underestimate WIP charges. Perhaps if more segments were used, it would be more difficult to find counterexamples where Rule 2 does not maintain the optimal product mix for a CF problem.

To test this hypothesis, the CF formulation problem was rerun using 20 segments instead of the current 3. It turns out that using a larger number of line segments to represent the curved clearing function caused the Rule 2 allocation problem to find optimal solutions with reduced output at smaller amounts of traceable overhead applied. This result forced a rejection of the hypothesis and forced a reexamination the mechanics of the Rule 2 problem. Going back to the economic view of this problem, marginal costs of the resources equal the marginal value of the resources at optimality. Rule 1 always works because the objective function after adjusting for untraceable overhead is always parallel with the original objective function and the constraint space remains unchanged. In Rule 1, both the per unit product profit contribution coefficients and the WIP charges for the CF formulation are scaled with
the reduced WIP charges being fully compensated by lower throughput at the optimal solution. Rule 2 causes the slope of the original objective function to change. For the LP formulation, Rule 2 explicitly gives the limits on how much the slope can change and still maintain the optimal solution of the original product mix problem. For the CF formulation, the marginal value of the product(s) with the traceable overhead applied is reduced but the WIP contributions to marginal costs remain the same. If enough segments are added (or if a column generation approach was used) to represent the curved clearing function and if the product allocated traceable overhead shares a bottleneck or near bottleneck with other products, any change in slope of the objective function will change the optimal solution. One possible explanation is that the structure of the clearing function formulation does not allow us to disaggregate the duals to all the affected coefficients. Unlike the LP formulation, the duals in the CF formulation are not on a per unit basis. Even though WIP charges are built into the CF formulation duals, there is no obvious way to allocate the traceable overhead back to variables beyond the product variables.

There are three distinct ways to move forward. One is to only apply overhead allocation when we are confident that the decision variables of optimal solution will not change. Even the sensitivity analysis of the basic Excel Solver gives us this capability. Trying to allocate 1.16 units of traceable overhead to machine 2 was going to be just greater than 1.158 (0.579 times 2) that Solver said was going to cause a tipping point in the solution. Another way is to ignore the new solution and calculate overhead absorption based on the original duals. In a way, the new solution is a sham given that the reduced output necessary to make the WIP charges less does not fully allocate our traceable overhead, which in this case is a difference of 0.54 units of overhead. Because we cannot guarantee that the new CF
optimization will maintain the same product mix after applying Rule 2, the final way to move forward is to reject Rule 2 for CF formulation problems.

3.5 Allocating Overhead with Product Interdependencies

So far, we have considered only capacity constraints. Sometimes there are sales or production interdependencies between products. For example, product 1 may use the same stock of a raw material as product 2 where product 1 uses 2/3 of the stock and product 2 uses the other third. Or there might be a sales dependency where for each unit of product 2 sold, two units of product 1 are sold. Let’s add this constraint and see how each problem formulation performs:

\[-X_1 + 2X_2 \leq 0\]

We proceed to solve the both formulations and find that the new optimal values of the product variables are $X_1^* = 2$ and $X_2^* = 1$. The new objective function value for the LP formulation is $G_{LP}^* = 2.5$ and for the CF formulation is $G_{CF}^* = 1.604$. There are, of course, new duals, displayed in Table 3-8.

<table>
<thead>
<tr>
<th>Machine 1</th>
<th>Machine 2</th>
<th>Machine 3</th>
<th>Interdependency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega_{n}^{LP}$</td>
<td>$\omega_{n}^{CF}$</td>
<td>$\omega_{n}^{CF}$</td>
<td>$\omega_{n}^{CF}$</td>
</tr>
<tr>
<td>0</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>0.25</td>
<td>1.354</td>
<td>0</td>
<td>0.075</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 3-8: Comparison of Duels Between LP and CF Formulations with Product Interdependencies
Kaplan and Thompson show that the fixed overhead applied to each unit produced in the LP formulation with product interdependencies is still $kω^*A$. For our example, keeping the amount of fixed overhead to allocate at one unit, the calculation is

$$\text{Per unit overhead absorbed} = k_{LP}ω^*A = \left( \begin{array}{c} 1 \\ 2.5 \end{array} \right) \left[ \begin{array}{cccc} 0 & 0.25 & 0 & 0.25 \\ 5 & 0 & 4 & 0 \\ -1 & 2 & \end{array} \right] = \left[ \begin{array}{c} 0.4 \\ 0.2 \end{array} \right]$$

Multiplying by the optimal values of $x$ from the primal, we obtain the total fixed overhead absorbed by each product. For product 1, the fixed overhead absorbed is $2 \times 0.4 = 0.8$ units. Similarly, for product 2 the fixed overhead absorbed is $1 \times 0.2 = 0.2$ units. We can also calculate our fixed overhead absorbed more directly using the $kpx^*$ method where we obtain identical results.

$$X_1 \text{ total fixed overhead absorbed} = \left( \frac{1}{2.5} \right)(1)(2) = 0.8$$

$$X_2 \text{ total fixed overhead absorbed} = \left( \frac{1}{2.5} \right)(0.5)(1) = 0.2$$

Turning our attention to the CF formulation, the introduction of sales or production interdependencies violates the structure of $X_i = k_{CF} \sum_n Z_{in}ω^*_{nCF}$. We return to the more general $kω^*A$ formulation and find that the calculations proceed the same way as the general untraceable overhead case only there are two more columns (representing $X_1$ and $X_2$) and one more row (representing the interdependency constraint) in the $A$ matrix.

$$k_{CF}ω^*A = \left( \frac{1}{1.604} \right) \left[ \begin{array}{c} 0.250 \\ 1.354 \\ 0.0 \\ 0.075 \end{array} \right] \times \left[ \begin{array}{ccccccccc} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ -1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] = \approx \left[ \begin{array}{ccccccccc} 0.047 & 0.093 & 0.156 & 0.075 & 0.156 & 0.156 & 0.844 & 0 \\ -1.604 \end{array} \right]$$
Again, we multiply only the variables that are assigned to each product. Multiplying by $X_i$ and the $Z_{in}$ decision variables, we obtain

$$\text{Fixed overhead allocated to } X_1 = \begin{bmatrix} -0.047 & 0.156 & 0.844 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 2/3 \\ 1 \\ 0 \end{bmatrix} \approx 0.854$$

$$\text{Fixed overhead allocated to } X_2 = \begin{bmatrix} 0.093 & 0.156 & 0.844 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1/3 \\ 0 \\ 1 \end{bmatrix} \approx 0.145$$

Allocation can also be done using the $kpx^*$ method which yields only slightly different results due to rounding errors where our spreadsheet returns identical results.

$$X_1 \text{ total fixed overhead absorbed} = \frac{1}{1.604} \left( (1)(2)-(0.1)(3.889+2.400+0) \right) \approx 0.855$$

$$X_2 \text{ total fixed overhead absorbed} = \frac{1}{1.604} \left( (0.5)(1)-(0.1)(1.667+0+1.000) \right) \approx 0.145$$

3.6 Conclusions

Through a simple numerical example, we find that the dual characteristics of different mathematical models describing production alternatives greatly influence results obtained in cost allocation methods proposed in the literature. The clearing function model that tracks WIP and incorporates its congestive effects on the system provides richer dual information. This dual information, when used to allocate overhead costs, allows the allocation of certain overhead costs to near-bottleneck machines that operate at high utilization levels but are still below 100%, contributing to deterioration in system performance. The shape of the clearing function depends on the amount of variability in the production system, allowing at least a first-order representation of the effects of variability in the cost allocation schemes.
Results proved beneficial for allocating untraceable overhead. Using the duals, we can allocate untraceable overhead directly. We found and proved that Kaplan and Thompson’s Rule 1 worked as expected for the clearing function formulations. However, the model was unable to consistently replicate the optimal product mix when allocating traceable overhead using Kaplan and Thompson’s Rule 2. Finally, we added the complexity of product interdependencies to our clearing function example and found that it works the same way as the methods Kaplan and Thompson proposed overhead allocation with basic linear programming product mix problems.

The results of our numerical example suggest that allocating all fixed, untraceable overhead based on the proposed scheme is likely to be beneficial only for a subset of firms. The assumption that all units are sold certainly does not apply to many firms. Additionally, the firm incurs many costs that can be allocated using more appropriate drives, as advocated by activity-based costing due to the failure of Kaplan and Thompson’s Rule 2 for allocating traceable overhead with clearing functions. However, a thorough understanding of Kaplan and Thompson’s Rule 2 clearly shows how activity-based costing can distort the optimal product mix when the average costs exceed marginal costs for resources or setups. An interesting possibility would be to allocate certain overhead costs, such as production supervision and work in process costs, based on the proposed scheme. The rationale for this is that much of the time of production supervisors is likely to be spent managing bottleneck and near-bottleneck resources, and the accumulation of WIP on the shop floor is determined to a large degree by the load at bottlenecks and near-bottlenecks. Hence a costing scheme that charges WIP and supervision costs to products may well motivate managers to redesign
products and processes. Or, they might give priority to products that do not place additional loads on the bottleneck and near-bottleneck resources.

One problem with the numerical example provided is that it assumes that variability is a function of only the resource. Real production environments may have products that have very different arrival or service variabilities. Should this be the case, a more complicated clearing function may be needed as total variability (and therefore WIP accumulation) will be a function of both utilization and of product mix.

Another issue with this clearing function model has to do with the allocation of time on a shared resource where the total utilization is less than 100%. This numerical example lacked sufficient segments, so resources could be at 100% utilization. We know that at steady state, 100% utilization causes WIP to rapidly accumulate should there be any variability. Also, the initial conditions of the optimization problem can influence the proportion of the shared resource allocated to each product. It allocates all used capacity to Product 1 and Product 2. However, it will either allocate all unused capacity to Product 1 or all unused capacity to Product 2 depending on the initial answer of the optimization. The reality is that the unused capacity should be apportioned with some assigned to each product. A power series representation is useful to fully allocate costs of this unused capacity. This insight is an important foundation for Chapter 6.

In summary, Kaplan and Thompson’s Rule 2 proved not to be compatible with clearing functions. It was first thought that the lack of clearing function segments used as the clearing functions in the paper allowed there to be counterexamples where the rule was not capable of replicating anticipated results for Rule 2. Using only three segments to represent the clearing functions for each resource means that the clearing functions allowed significant
underestimating of WIP. Re-deriving clearing functions (work presented in Chapter 4) allowed for the easy creation of more segment-defined clearing functions which allowed for greater detail. What was found was that Rule 2 only worked before because of the inaccuracies of the clearing function representation. Clearing functions with more segments showed more lack of alignment with our expected Rule 2 results. This derivation and these findings were presented at the ISERC 2014 Conference & Expo Ashbaugh (2014), only the output rates were changed to 2 orders of magnitude higher. Because WIP is a function of only utilization and variability, the results were parallel to these. The next section details the new derivation and will be used in a numerical model.
CHAPTER 4: Alternate Clearing Function Derivation

A new derivation for clearing functions can be obtained starting with the Kingman approximation for general arrival, general service queues (the $VUT$ formula) Hopp and Spearman (2001) of the steady state average waiting time in a G/G/1 queue.

$$CT_q = \left(\frac{\left(\frac{\sigma_\lambda}{\lambda}\right)^2}{2} + \frac{\left(\frac{\sigma_\mu}{\mu}\right)^2}{2}\right) \left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{\mu}\right)$$

Using Little’s Law,

$$I = RT \rightarrow WIP_q = \lambda CT_q$$

the average steady state WIP waiting for service can be calculated.

$$WIP_q = \frac{\left(\frac{\sigma_\lambda}{\lambda}\right)^2 + \left(\frac{\sigma_\mu}{\mu}\right)^2}{2} \left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{\mu}\right) = \frac{\left(\frac{\sigma_\lambda}{\lambda}\right)^2 + \left(\frac{\sigma_\mu}{\mu}\right)^2}{2} \left(\frac{\rho^2}{1-\rho}\right)$$

But

$$WIP = WIP_q + \rho = \frac{\left(\frac{\sigma_\lambda}{\lambda}\right)^2 + \left(\frac{\sigma_\mu}{\mu}\right)^2}{2} \left(\frac{\rho^2}{1-\rho}\right) + \rho$$

To make the derivation clearer,

$$V = \frac{c^2_a + c^2_e}{2} = \frac{\left(\frac{\sigma_\lambda}{\lambda}\right)^2 + \left(\frac{\sigma_\mu}{\mu}\right)^2}{2}$$

Substituting,
We see an important result. We can determine the WIP based on only utilization and variability. We do not need to know either the arrival rate or service rate to calculate WIP, just the relative utilization. If our arrival variability and service variability are constant between products, then we can calculate total WIP just by knowing the total utilization of the resource, not the relative mix of products using the resource.

We can calculate the slope of the WIP to utilization by taking the derivative with respect to utilization, \( \rho \). The slope is calculated using the quotient rule as follows:

\[
Slope = \frac{\partial}{\partial \rho} \left( \frac{(V-1)\rho^2 + \rho}{1 - \rho} \right) = \frac{(1-\rho)(2V\rho - 2\rho + 1) + (V - \rho^2 + \rho)}{(1 - \rho)^2} = \frac{-V\rho^2 + 2V\rho + \rho^2 - 2\rho + 1}{(1 - \rho)^2}
\]

The clearing function plots capacity multiplied by utilization, \( \rho \), on the y-axis and WIP on the x-axis. Therefore, we need to adjust our slope to the following

\[
Slope_{cr} = \left( \frac{(1-\rho)^2}{-V\rho^2 + 2V\rho + \rho^2 - 2\rho + 1} \right) \times \text{Capacity}
\]

An example of a clearing function is provided where the variance constant \( V \) reduces to 1 as in the case of the familiar M/M/1 case. The capacity in this clearing function is 10-time units and the WIP is calculated for every \( \frac{1}{2} \) time units of capacity used, shown graphically in Figure 4-1.
To use clearing functions, we need to know both the slope and intercepts of line segments tangent to our clearing function curve. Recall intercepts can be calculated by

\[ y_{\text{int}} - y_i = m(x - x_i) \rightarrow y_{\text{int}} - (\text{Capacity})\rho = \text{Slope}_{CF} (0 - \text{WIP}) \]

\[ y_{\text{int}} = (\text{Capacity})\rho - (\text{Slope}_{CF} \text{ (WIP)}) \]

So, using our formulas for the slope and WIP, we can use algebraic cancellation to arrive at a modest equation for the intercept.
\[ y_{\text{int}} = (\text{Capacity}) \rho - \frac{(\text{Capacity})(1 - \rho)^2}{-V \rho^2 + 2V \rho + \rho^2 - 2\rho + 1} \left( \frac{(V - 1) \rho^2 + \rho}{1 - \rho} \right) \]

\[ y_{\text{int}} = (\text{Capacity}) \rho - \frac{(\text{Capacity})(1 - \rho)(V \rho^2 - \rho^2 + \rho)}{-V \rho^2 + 2V \rho + \rho^2 - 2\rho + 1} \]

\[ y_{\text{int}} = (\text{Capacity}) \rho - \frac{(\text{Capacity})(V \rho^2 - V \rho^3 + \rho^3 - 2\rho^2 + \rho)}{-V \rho^2 + 2V \rho + \rho^2 - 2\rho + 1} \]

\[ y_{\text{int}} = \frac{(\text{Capacity})V \rho^2}{-V \rho^2 + 2V \rho + \rho^2 - 2\rho + 1} \]

In the special case of the M/M/1 queue, our y-intercept reduces to just

\[ y_{\text{int}} = (\text{Capacity}) \rho^2 \]

We now have an easy way to calculate clearing function line segments for use in piecewise linear LP models. Each piecewise linear segment representation has a slope and a y-intercept. If we know what the variation constant is, we can easily calculate the slopes. The y-intercepts are calculated based on the amount and percentage of capacity used.

This clearing function formulation has several advantages versus the clearing function formation proposed by Asmundsson et al. (2009). First, it is easier to count WIP than to count hours of WIP. Second, this formulation has advantages in generating clearing functions fitting empirical data. If theoretical capacity is known, the variability constant can be obtained with a curve fitting to the G/G/1 WIP formula. This may prove to be a major improvement over the methodology employed by Kacar (2012). Third, having more segments will give better answers to determine when adding additional load to a resource increases WIP costs greater than any increase in throughput contribution, or, when adding more production reduces total profit.
Returning to the numerical example in Chapter 3, we can use this new derivation to generate richer clearing functions with many more segments. If all three machines were essentially M/M/1 servers, our clearing functions would look like Figure 4-2:

![Figure 4-2: Chapter 3 Model Replaced with 20 Segment Clearing Function](image)

These clearing functions were generated using the same theoretical capacities as before but can now use 20 piecewise linear segments. If we used these clearing functions, the solution would change as WIP would increase as utilization increased forcing a solution where not all capacity is used. The utilizations (and throughput) would be higher if WIP charges were lower. A comparison is displayed in Table 4-1.
## Table 4-1: Comparison of 3 Segment versus 20 Segment Clearing Function

<table>
<thead>
<tr>
<th>Utilization Function</th>
<th>3 Segment Paper CF</th>
<th>20 Segment CF</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>2</td>
<td>1.76</td>
</tr>
<tr>
<td>X2</td>
<td>1.25</td>
<td>0.91</td>
</tr>
<tr>
<td><strong>Utilization Machine 1</strong></td>
<td>94.4%</td>
<td>78.9%</td>
</tr>
<tr>
<td><strong>Utilization Machine 2</strong></td>
<td>100%</td>
<td>88%</td>
</tr>
<tr>
<td><strong>Utilization Machine 3</strong></td>
<td>100%</td>
<td>72.73%</td>
</tr>
<tr>
<td><strong>Total Profit</strong></td>
<td>$2.625</td>
<td>$1.87</td>
</tr>
</tbody>
</table>

The clearing functions used in Excel are provided in Tables 4-2, 4-3, and 4-4:

## Table 4-2: Machine 1 Enhanced Excel Clearing Function

<table>
<thead>
<tr>
<th>Capacity</th>
<th>9</th>
<th>Variability</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine 1 Clearing Function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Active = 1</th>
<th>Utilization</th>
<th>Cap Used</th>
<th>WIP</th>
<th>Slope</th>
<th>Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.45</td>
<td>0.052632</td>
<td>8.1225</td>
<td>0.0225</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.9</td>
<td>0.111111</td>
<td>7.29</td>
<td>0.09</td>
</tr>
<tr>
<td>1</td>
<td>0.15</td>
<td>1.35</td>
<td>0.176471</td>
<td>6.5025</td>
<td>0.2025</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>1.8</td>
<td>0.25</td>
<td>5.76</td>
<td>0.36</td>
</tr>
<tr>
<td>1</td>
<td>0.25</td>
<td>2.25</td>
<td>0.333333</td>
<td>5.0625</td>
<td>0.5625</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>2.7</td>
<td>0.428571</td>
<td>4.41</td>
<td>0.81</td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
<td>3.15</td>
<td>0.538462</td>
<td>3.8025</td>
<td>1.1025</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>3.6</td>
<td>0.666667</td>
<td>3.24</td>
<td>1.44</td>
</tr>
<tr>
<td>1</td>
<td>0.45</td>
<td>4.05</td>
<td>0.818182</td>
<td>2.7225</td>
<td>1.8225</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>4.5</td>
<td>1</td>
<td>2.25</td>
<td>2.25</td>
</tr>
<tr>
<td>1</td>
<td>0.55</td>
<td>4.95</td>
<td>1.222222</td>
<td>1.8225</td>
<td>2.7225</td>
</tr>
<tr>
<td>1</td>
<td>0.6</td>
<td>5.4</td>
<td>1.5</td>
<td>1.44</td>
<td>3.24</td>
</tr>
<tr>
<td>1</td>
<td>0.65</td>
<td>5.85</td>
<td>1.857143</td>
<td>1.1025</td>
<td>3.8025</td>
</tr>
<tr>
<td>1</td>
<td>0.7</td>
<td>6.3</td>
<td>2.333333</td>
<td>0.81</td>
<td>4.41</td>
</tr>
<tr>
<td>1</td>
<td>0.75</td>
<td>6.75</td>
<td>3</td>
<td>0.5625</td>
<td>5.0625</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>7.2</td>
<td>4</td>
<td>0.36</td>
<td>5.76</td>
</tr>
<tr>
<td>1</td>
<td>0.85</td>
<td>7.65</td>
<td>5.666667</td>
<td>0.2025</td>
<td>6.5025</td>
</tr>
<tr>
<td>1</td>
<td>0.9</td>
<td>8.1</td>
<td>9</td>
<td>0.09</td>
<td>7.29</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>8.55</td>
<td>19</td>
<td>0.0225</td>
<td>8.1225</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9</td>
<td>infinite</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>
**Table 4-3: Machine 2 Enhanced Excel Clearing Function**

<table>
<thead>
<tr>
<th>Capacity</th>
<th>10</th>
<th>Variability</th>
<th>1</th>
<th>Machine 2 Clearing Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active = 1</td>
<td>Utilization</td>
<td>Cap Used</td>
<td>WIP</td>
<td>Slope</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.5</td>
<td>0.052632</td>
<td>9.025</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>0.111111</td>
<td>8.1</td>
</tr>
<tr>
<td>1</td>
<td>0.15</td>
<td>1.5</td>
<td>0.176471</td>
<td>7.225</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>2</td>
<td>0.25</td>
<td>6.4</td>
</tr>
<tr>
<td>1</td>
<td>0.25</td>
<td>2.5</td>
<td>0.333333</td>
<td>5.625</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>3</td>
<td>0.428571</td>
<td>4.9</td>
</tr>
<tr>
<td>1</td>
<td>0.35</td>
<td>3.5</td>
<td>0.538462</td>
<td>4.225</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>4</td>
<td>0.666667</td>
<td>3.6</td>
</tr>
<tr>
<td>1</td>
<td>0.45</td>
<td>4.5</td>
<td>0.818182</td>
<td>3.025</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>5</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>1</td>
<td>0.55</td>
<td>5.5</td>
<td>1.222222</td>
<td>2.025</td>
</tr>
<tr>
<td>1</td>
<td>0.6</td>
<td>6</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>1</td>
<td>0.65</td>
<td>6.5</td>
<td>1.857143</td>
<td>1.225</td>
</tr>
<tr>
<td>1</td>
<td>0.7</td>
<td>7</td>
<td>2.333333</td>
<td>0.9</td>
</tr>
<tr>
<td>1</td>
<td>0.75</td>
<td>7.5</td>
<td>3</td>
<td>0.625</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>8</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>1</td>
<td>0.85</td>
<td>8.5</td>
<td>5.666667</td>
<td>0.225</td>
</tr>
<tr>
<td>1</td>
<td>0.9</td>
<td>9</td>
<td>9</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>9.5</td>
<td>19</td>
<td>0.025</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
<td>infinite</td>
<td>0</td>
</tr>
</tbody>
</table>
The newer clearing functions allow a much more realistic accounting of WIP and effective capacity, especially when subject to variability. A more interesting problem would be to have significantly different variability inherent to each resource. We would expect our solution to change with lower projected output from higher variability resources. Returning to our example, let us look at the plot when the highest capacity resource, machine 2, is now the resource with the most variability. Graphically, this can be seen in Figure 4-3.
Figure 4-3: 20 Segment Clearing Function with High Variability in Machine 2

Notice that the variability of machine 2 causes steady state WIP to increase faster than machine 1, despite machine 1 having lower theoretical capacity. Depending on the cost of WIP, we could see a result where the throughput constraining resource actually has the highest theoretical capacity.
CHAPTER 5: Allocating Overhead Using Clearing Functions with Parallel Resources

Having proved the ability to use clearing functions with an established overhead allocation schema for untraceable fixed overhead, the next challenge is to extend the research into parallel resource systems. There are different approaches to modeling parallel machines. One approach is to go back to multichannel queuing and derive clearing function segment slopes and intercepts based on the G/G/m approximations. A different approach is to set up variables for each product on each machine so that they don’t interfere with one another. For example, if we have a product 1 and product 2 sharing a machine, we can create a dummy product 3 and product 4 for a parallel machine and use the same contribution margins. The total of product 1 produced will be the sum of product 1 and the dummy product 3. This second approach will be used in Chapter 6.

5.1 Deriving Clearing Functions with Symmetrical Parallel Machines

Once again, we go back using Factory Physics, this time equation 8.28 in Hopp and Spearman (2001), as a start to derive clearing functions for G/G/m queues.

\[ CT_q(G / G / m) = \left( \frac{c_a^2 + c_r^2}{2} \right) \rho^{(m+1)^{-1}} \frac{1}{m(1 - \rho)} \]  

where \( t_c = \frac{m\rho}{\lambda} \)

Using Little’s Law, \( I = RT \) we note that \( WIP_q = \lambda CT_q \). Combining terms and again, substituting

\[ V = \frac{c_a^2 + c_r^2}{2} = \left( \frac{\sigma_a}{\lambda} \right)^2 + \left( \frac{\sigma_r}{\mu} \right)^2 \]
we see that

\[ WIP_q (G / G / m) = \{V\} \left( \frac{\rho^{\frac{m}{m+1}}}{1 - \rho} \right). \]

Further, noting that \( WIP = WIP_q + \frac{\lambda}{\mu} \) but \( \rho = \frac{\lambda}{m \mu} \) so \( \frac{\lambda}{\mu} = m \rho \) in the multichannel model.

This leaves \( WIP = \frac{V \rho^{\frac{m}{m+1}}}{1 - \rho} + m \rho = \frac{V \rho^{\frac{m}{m+1}} + m \rho - m \rho^2}{1 - \rho} \). Note that WIP is now a function of variability \( V \), utilization \( \rho \), and number of machines \( m \). Assuming that the number of machines available during our time period remains fixed, we can take the derivative with respect to utilization to determine the change in steady state WIP due to changes in utilization. Using the Quotient Rule, we obtain:

\[
\frac{\partial \rho}{\partial WIP} = \frac{V \sqrt{2(m+1)} \left( \rho^{\frac{m}{m+1}} - \rho^{\frac{m}{m+1}} \right) + m - 2m \rho + m \rho^2 + V \rho^{\frac{m}{m+1}}}{(1 - \rho)^2}
\]

Like our single server model, the clearing function is made by swapping the x and y axis and multiplying by the theoretical capacity, \( mC \). The resultant clearing functions are graphed in Figure 5-1.
These multi-server models may be used with the existing optimization models to allocate overhead. An added benefit of this analysis is it allows us to see the impact of additional machines on a production system. Using these multi-server clearing functions, we can easily see how much additional throughput can be expected by adding a parallel server. In addition, we can easily see the costs of the added overhead to see if the additional capacity is worth it.

5.2 A Numerical Example Using Symmetric Parallel Machines

One advantage of the symmetrical parallel machine approach is that it allows us to experiment with adding machines to decrease congestion and therefore WIP and cycle time. Suppose a machine shop employs symmetric parallel machines but has variability twice that as M/M/m. Suppose each unit takes 2.5 hours on a machine and that the shop runs for 8 hours per day. If per unit profits are 0.5 per unit and WIP costs are 0.1 unit per unit, the model
calculates the economic output at different numbers of machines used and is summarized in Table 5-1.

**Table 5-1: LP Output for Symmetric Parallel Machines**

<table>
<thead>
<tr>
<th>Number Machines</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective Value</td>
<td>0.75</td>
<td>1.86</td>
<td>3.07</td>
<td>4.35</td>
<td>5.63</td>
</tr>
<tr>
<td>Optimal Output</td>
<td>2.17</td>
<td>4.98</td>
<td>7.47</td>
<td>10.61</td>
<td>13.26</td>
</tr>
<tr>
<td>WIP at Optimal Output</td>
<td>3.48</td>
<td>6.31</td>
<td>6.65</td>
<td>9.52</td>
<td>9.91</td>
</tr>
<tr>
<td>Capacity Used</td>
<td>5.42</td>
<td>12.46</td>
<td>18.67</td>
<td>26.51</td>
<td>33.14</td>
</tr>
<tr>
<td>Percent Capacity Used</td>
<td>67.7%</td>
<td>77.8%</td>
<td>77.8%</td>
<td>82.9%</td>
<td>82.9%</td>
</tr>
</tbody>
</table>

Should the output need to be 9 units per day, which is obtainable in a no to low variability environment as 24hours/2.5hours/unit is 9.6 units of theoretical capacity, there would need to be four machines based on the variability of the system, WIP costs, and per unit profit for this product.

5.3 Extending Multiple Parallel Machines

Using symmetric multiple parallel machines does not change the original problem from Chapters 3 and 4. There may be some additional iterations to determine the appropriate number of machines before finally generating the duals and using them to allocate overhead. The clearing functions work the same way as before, in that the \( Z_{in} \) variables will output the percentage of time each product will be assigned to the combined resource cluster. From an Excel point of view, once the formulas for intercept, WIP, and capacity are entered, the mechanics remain the same provided the number of machines stays constant within a problem.
CHAPTER 6: Overhead Allocation Using Asymmetric Parallel Machines or Different Customer Arrivals

The final approach to extending the Kingman Approximation is using asymmetric parallel machines or modeling service systems with different customer arrival type. The models should prove similar in that modeling with be an iterative process as the average arrival and service rates are dependent on the product mix or customer type.

6.1 Asymmetric Parallel Machines

One advantage this approach is that it allows parallel machines each with different variability and therefore different clearing function representations. Product interdependencies can be considered as well as WIP interdependencies. For example, there might be a strategic decision to make the WIP, or by extension, hours of WIP, on each machine to be equal.

6.2 Using Asymmetric Parallel Machines to Model Different Product Mixes

Clearing functions are typically based on the machine variability based on the Kingman VUT equation. The $V$ is calculated using both the coefficient of variation for the arrivals and the coefficient of variation for the service process. Therefore, we will need more information than just theoretical capacity and $V$ to generate our clearing functions. Recall that

$$V = \frac{\left(\frac{\sigma_\lambda}{\lambda}\right)^2 + \left(\frac{\sigma_\mu}{\mu}\right)^2}{2}$$
In previous chapters, product mix variation was not assumed to affect the clearing functions. The arrival variabilities were assumed to be similar and the process variability was assumed to be machine specific, not dependent on product. One way around this assumption is to use dummy parallel machines and then link them together similar to the classic linear programming cutting stock problem. Another is to use a nonlinear model where capacity allocation are the decision variables and waiting time or loss function minimization is the objective function. The model is thus:

\[
\min \quad X_1W_{q1} + X_2W_{q2} \\
\text{Subject to :}
\]

\[
\sum_{i=1}^{n} C_n \leq C_T \forall C_n \quad \text{Capacity Constraint}
\]

\[
C_n \geq 0 \forall C_n \quad \text{No capacities can be negative}
\]

\[
W_{q}\ell = \left(\frac{c_a^2 + c_e^2}{2}\right) \left(\frac{X_T c_i}{C_i} \right) 1\ell
\]

Which is a variation of the Kingman G/G/1 approximation.

The decision variables in the model are the C\(_n\)'s, or the capacity to allocate each product if the resource could be split between products. The odd thing about this model is the X\(_n\)'s and C\(_T\), or theoretical capacity, are known. The output of the model tells us how much of the theoretical capacity is needed for each product or customer type.
6.3 Incorporating Average Customer Waiting Time Costs in Overhead Allocation for Service Systems

Because service systems often do not have physical WIP that can be priced, a better system might be to cost cycle time. This costing should remain congruent to the production models where WIP is a decision variable and the magnitude of WIP comes from clearing functions.

Recall that the WIP calculating clearing functions were derived from the Kingman VUT equation for G/G/1 queues, the average time the customer waits for service is:

\[
CT_q = \left( \frac{\left( \frac{\sigma^2}{\lambda} \right)^2 + \left( \frac{\sigma}{\mu} \right)^2}{2} \right) \left( \frac{\rho}{1 - \rho} \right) t_e
\]

We can add a cost to waiting time instead of a WIP cost for service systems. WIP in a service system is really line length. My experience is that people are willing to tolerate longer lines that move quickly compared to shorter lines that do not. The clearing function is easier to represent but no longer has the convenient feature that only relative utilization and variability are important. Like when using asymmetric parallel machines, now we must also track the average service time at the resource, the combined service standard deviation, the combined arrival standard deviation, and the average arrival rate. The average arrival rate and average service rate will be dependent on the product mix. Should there be two customer arrival streams with different probability distributions but share the same service distribution, this makes the problem a little easier which is our MRI problem in Chapter 7.
6.4 Using Geometric Power Series to Allocate Capacity

A curious result in Chapter 3 forms a basis for understanding how two product types or two customer types can share a resource. In our Chapter 3 example, there are 9 hours of capacity on the shared machine where product 1 uses 6 hours and product 2 uses 2.5 hours. Because the Simplex LP method uses point solutions, it is possible that there be two equally optimal solutions where the $Z_{11}$ and $Z_{12}$ are $\{2/3, 1/3\}$ or $\{13/18, 5/18\}$. If the contribution margins minus WIP costs were identical, we could allocate capacity using a geometric series.

In this case, we would expect that product 1 would use up 2/3 of the remaining 1/18 of unused capacity while product 2 would use up 5/18 of the remaining 1/18. There would be 1/18 squared remaining after the first pass of overhead allocation. The two series would look like this:

$$\frac{2}{3} + \frac{2}{3} \times \left(\frac{1}{18}\right) + \frac{2}{3} \times \left(\frac{1}{18}\right)^2 + \frac{2}{3} \times \left(\frac{1}{18}\right)^3 + \ldots + \frac{2}{3} \times \left(\frac{1}{18}\right)^n$$

and

$$\frac{5}{18} + \frac{5}{18} \times \left(\frac{1}{18}\right) + \frac{5}{18} \times \left(\frac{1}{18}\right)^2 + \frac{5}{18} \times \left(\frac{1}{18}\right)^3 + \ldots + \frac{5}{18} \times \left(\frac{1}{18}\right)^n$$

Recalling that convergence for a geometric power series is $\frac{a}{1-r}$

where $r = \frac{1}{18}$, $a_1 = \frac{2}{3}$, and $a_2 = \frac{5}{18}$ our convergence points will be

$$\frac{2}{3} = \frac{12}{17}$$

and

$$\frac{5}{18} = \frac{5}{17}$$

which allows us an easy method to fully allocate the capacity costs to the different products. Generically, for a two product or
two customer type system, \( r = 1 - (a_1 + a_2) \) where \( a_1 \) is the capacity directly used by product or customer type 1, \( a_2 \) is the capacity directly used by product or customer type 2, and \( r \) is the capacity not directly used by either product or customer type. The \( r \) is really the extra capacity required by the variability to keep WIP and/or cycle time at a more economical level.
CHAPTER 7: A Medical Example Using Two Patient Types

Returning to the MRI example, if both scheduled patients and ER patients have the same average service time on the MRI, we just need to track the impact of the different arrival distributions to the overall wait times of each patient type. Recall that ER patient may preempt scheduled patients because waiting for these patients may be life threatening. Therefore, we would expect the cost of waiting for these patients to be larger than the scheduled patients. These costs are going to make it less ideal to take on too many patients despite the lost revenue of potential idle MRI time, depending of course as to the availability of alternative imaging locations. So far, we have used two streams of patients with different

<table>
<thead>
<tr>
<th>MRI FAQ From Advanced Imaging in Missoula, Montana</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>How long does a MRI scan take?</strong></td>
</tr>
<tr>
<td>The length of the exam depends on the type of study being performed.</td>
</tr>
<tr>
<td>• MRI of the Brain … 20-45 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Orbits … 20-35 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the TMJ … 45-60 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Soft Tissue Neck … 25-35 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Cervical Spine … 20-35 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Upper Extremity … 20-45 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Thoracic Spine … 25-45 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Chest … 25-45 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Abdomen … 25-45 minute scan time.</td>
</tr>
<tr>
<td>• MRI MRCP … 50-60 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Lumbar Spine … 20-35 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Pelvis … 20-35 minute scan time.</td>
</tr>
<tr>
<td>• MRI of the Lower Extremity … 20-35 minute scan time.</td>
</tr>
<tr>
<td>• MRI Run Off … 50-60 minute scan time.</td>
</tr>
<tr>
<td>• MRI Arthrogram … 30-60 minute scan time.</td>
</tr>
</tbody>
</table>

*Figure 7-1: MRI Time Estimates from Advanced Imaging in Missoula, MT*
arrival variabilities to use two dummy servers where the total server time of dummy server 1 plus dummy server 2 equals the total time for the server system. Each patient stream has a different cost of waiting. The output of the model divides the server time so that the capacity is allocated to each dummy server. Using the arrival variability and server variability based on the capacity allocation, unused capacity can be allocated to the different patient types reflective of their impact on waiting times. This is tested to see how useful the two-dummy server model is at allocating capacity between the patient streams and using MRI scan times from Figure 7-1.

7.1 MRI Numerical Example

As promised, there will be two arrival streams of patients to the MRI machine. Scheduled patients will typically arrive between 10 minutes early to 10 minutes late for their appointments. Penalties can be used to enforce these time limits. Scheduled patients are to arrive on the hour such that there are 8 scheduled patients during the 8-hour day shift. Emergency room patients will arrive more randomly using an exponential distribution where 3 patients should arrive to the MRI center during the 8-hour day shift. Because emergency room patients have a higher risk of dying while waiting for the imaging service, they get priority over scheduled patients but will not pull a scheduled patient out of a scan in-progress. Based on data from Advanced Imaging, a triangular distribution will be used where the minimum scan is 20 minutes, the maximum scan is 60 minutes, and the most likely scan time is 40 minutes. This yields a mean of 40 minutes, a variance of 66.67 minutes, and a coefficient of variation of 0.041667. The VUT equation is thus:
\[ CT_q = \left( \frac{0.080527 + 0.041667}{2} \right) \left( \frac{0.91667}{1-0.91667} \right)^{(40)} \]

Which gives us an average wait time of approximately 53.8 minutes, or total system time just over 93.8 minutes. On average, 40 minutes will not be used on the MRI in an 8-hour day shift which equates to a 91.667% utilization. There are two methods to allocate this extra capacity to patients:

Method 1: Use an optimization model where the objective function is to minimize total wait time or total loss function based on waiting and priority of all customers. This is the same model from section 6.2. An advantage of this method is that the allocation method is it reduces total waiting time of all patients. It will penalize emergency room patients more than scheduled patients based on a higher volume of scheduled patients. The average wait in this system is 83 minutes. A smaller number of emergency patients means that a larger penalty can be applied to each emergency room patient than scheduled patients to lower total patient waits or loss function values. In this scenario, the share of this unused capacity assigned to emergency room patients is 64% while the share of unused capacity assigned to scheduled patients is 36%. If we could price MRI scans based on the number of minutes of capacity allocated, each scheduled patient scan would cost their actual scan time plus 1.8 minutes while each emergency room patient would be charged actual scan time plus 8.5 minutes, based on the large difference in arrival variabilities.

Method 2: Use an optimization model where the objective function is to equalize average per patient wait time between customer types. This adds one constraint to the model used in section 6.2: \( W_{q1} = W_{q2} \) for this two-patient type model. This method caused more capacity to be used by the emergency room patients. Now 89.4% of extra capacity is allocated to emergency room patients and 10.6% is allocated to
scheduled patients. 11.9 minutes of spare capacity should be allocated to each emergency room patient and 0.53 minutes should be allocated to each scheduled patient. The average wait time for all patients increases to 139 minutes, however. A summary of results for the two methods where the loss function for both patient types is simply average time waiting is provided in Table 7.1.

Table 7-1: Overhead Allocation Summary by Method and Patient Type

<table>
<thead>
<tr>
<th>Allocation</th>
<th>Scheduled Patients</th>
<th>Emergency Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minutes capacity per 8 hours</td>
<td>334.4</td>
<td>145.6</td>
</tr>
<tr>
<td>Allocated capacity utilization</td>
<td>95.6%</td>
<td>82.4%</td>
</tr>
<tr>
<td>Percentage of unused capacity</td>
<td>36%</td>
<td>64%</td>
</tr>
<tr>
<td>Per unit unused capacity (minutes)</td>
<td>1.8</td>
<td>8.5</td>
</tr>
<tr>
<td>Average wait time (minutes)</td>
<td>41.18</td>
<td>195.27</td>
</tr>
</tbody>
</table>

Method 2

<table>
<thead>
<tr>
<th>Allocation</th>
<th>Scheduled Patients</th>
<th>Emergency Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minutes capacity per 8 hours</td>
<td>324.24</td>
<td>155.76</td>
</tr>
<tr>
<td>Allocated capacity utilization</td>
<td>98.7%</td>
<td>77.0%</td>
</tr>
<tr>
<td>Percentage of unused capacity</td>
<td>10.6%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Per unit unused capacity (minutes)</td>
<td>0.53</td>
<td>11.9</td>
</tr>
<tr>
<td>Average wait time (minutes)</td>
<td>139.8</td>
<td>139.8</td>
</tr>
</tbody>
</table>

Notice that small shifts in the overall allocation of capacity between the patient types causes large swings in wait times of each type. Also note that the allocated capacity utilization is much higher for the lower variability patient type. Variability reduces the effective capacity a resource can run without line length increasing at a nonlinear rate. It is no surprise that scheduled patients generate lower waiting times despite higher effective utilization than emergence room patients in this analytic model.

7.2 Monte Carlo Simulation

Finally, no model should go without thorough testing. Simple Monte Carlo simulations will be run comparing the results of the G/G/m and asymmetric parallel machine models to see how well they perform with simple arrival and process distributions using the
Arena 14.7 simulation software. Triangular distributions were used for scheduled patients and the actual MRI scanning time. Emergency room patients arrive with an exponential interarrival time averaging 160 minutes (about three ER patients per 8-hour day shift). Additionally, emergency room patients are given priority when entering the MRI process versus scheduled patients. Figure 7-2 provides a diagram of the model:

![Diagram of the model](image)

**Figure 7-2: Monte Carlo Simulation Model for MRI Usage**

The simulation was run over a 1680-hour period (seventy 24-hour days). Results of the simulation are presented in Table 7-2:

<table>
<thead>
<tr>
<th>Patient Type</th>
<th>Time in Line</th>
<th>Time Being Served</th>
<th>Total System Time</th>
<th>Max in System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergency Room</td>
<td>27.85</td>
<td>40.20</td>
<td>68.05</td>
<td>5</td>
</tr>
<tr>
<td>Scheduled Patient</td>
<td>78.59</td>
<td>39.97</td>
<td>118.56</td>
<td>10</td>
</tr>
</tbody>
</table>

We instantly see that the emergency room patients have much shorter wait times due to their priority versus scheduled patients. Total average wait time in the simulation was lower than from the Kingman Approximation analytic solution at 65.15 minutes. Our MRI machine utilization was 90.94% and our average number of total patients waiting was only 1.48 but was as high as 9. The simulation results show a more accurate representation of the MRI wait times and MRI utilization, but it is difficult to suggest an overhead or unused capacity.
allocation schema from the simulation results. Part of this is due to the use of priority 
sequencing to the MRI versus a standard first-in-first-out for the analytic model. A possible 
method of capacity allocation is to discount based on average wait times. 88.3% of the total 
waiting time is forced onto the scheduled patients while only 11.7% onto the emergency 
patients. Should unused capacity be allocated as an inverse function of additional wait time 
suffered by the patients, we would see 88.3%, or approximately 11.8 minutes per patient, of 
the unallocated capacity assigned to emergency MRI users and 11.7%, or 0.59 minutes per 
patient, allocated to scheduled patients. These results match favorably to Method 2’s analytic 
solution.

7.3 Conclusions

Using the Kingman G/G/1 Approximation allows for costing variability to products, 
to resources, and to customers or customer types. Although there are many accounting 
systems used today to price products, allocate overhead, and evaluate performance, none put 
a price on variability for more complex systems. Hopefully determining this price of 
variability can be useful in justifying variability reduction efforts, determining which 
customers and / or products are profitable, and finding market-based solutions to change 
behavior or at least be adequately compensated by higher variability products or customers. 
In the MRI example, the model can be rerun should management feel one more scan per day 
by scheduled patients can deliver additional revenue. What they will find in this model is 
adding just one more customer increases customer waiting by an order of magnitude. Should 
the hospital be a monopoly, they might be able to get away with this terrible customer service 
due to lack of alternatives. If the wait time is long enough and the location not sufficiently
remote, customers may choose to drive over an hour to reduce wait times to more reasonable levels. Seeing how much extra burden on customer service emergency room patients can be, seeing price advantages to scheduled patients may soften the frustration due to long potential waits. At least the methods created here can provide a logical way to allocate typically unused capacity to the different patient types.
REFERENCES


design.” IIE Transactions Vol. 26 No. 4, July 1994: 36-49.


Chakravorty, S. S., Atwater, J. B. “Bottleneck management: theory and practice.” Production

Chakravorty, S. S., Atwater, J. B. “How theory of constraints can be used to direct

Chakravorty, S. S., Sessum, J. L. “Developing effective strategies to prioritize set-up
reduction in a multi-machine production system: A throughput approach.” International

Charnes, A., Cooper W. W., and Ijiri, Y. “Breakeven Budgeting and Programming to


Colantoni, C. S., Manes R. P., Whinston A. “Programming, Profit Rates and Pricing


Deng, S., Yano, C. A. “Joint Production and Pricing Decisions with Setup Costs and

Gavirneni, S. and Kulkarni, V. “Concierge medicine: applying rational economics to health


Kee, R. “Integrating Activity-Based Costing with the Theory of Constraints to enhance production-related decision making.” *Accounting Horizons* Vol. 9 No. 4, 1995: 48-61.


Nanfang, C., Kaijun, L. “Optimization of a supply chainsystem based on TOC.” *IEEE 2009*


APPENDIX
Multiple Parallel Machine Clearing Functions used to calculate for Table 6.1

<table>
<thead>
<tr>
<th>Slope</th>
<th>Intercept</th>
<th>WIP</th>
<th>Capacity</th>
<th>Capacity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>6.578588</td>
<td>0.036446</td>
<td>0.055263</td>
<td>0.4</td>
<td>5%</td>
</tr>
<tr>
<td>5.445378</td>
<td>0.134454</td>
<td>0.122222</td>
<td>0.8</td>
<td>10%</td>
</tr>
<tr>
<td>4.524462</td>
<td>0.2818</td>
<td>0.202941</td>
<td>1.2</td>
<td>15%</td>
</tr>
<tr>
<td>3.764706</td>
<td>0.470588</td>
<td>0.3</td>
<td>1.6</td>
<td>20%</td>
</tr>
<tr>
<td>3.130435</td>
<td>0.695652</td>
<td>0.416667</td>
<td>2</td>
<td>25%</td>
</tr>
<tr>
<td>2.596026</td>
<td>0.953642</td>
<td>0.557143</td>
<td>2.4</td>
<td>30%</td>
</tr>
<tr>
<td>2.142631</td>
<td>1.242472</td>
<td>0.726923</td>
<td>2.8</td>
<td>35%</td>
</tr>
<tr>
<td>1.756098</td>
<td>1.560976</td>
<td>0.933333</td>
<td>3.2</td>
<td>40%</td>
</tr>
<tr>
<td>1.425626</td>
<td>1.908689</td>
<td>1.186364</td>
<td>3.6</td>
<td>45%</td>
</tr>
<tr>
<td>1.142857</td>
<td>2.285714</td>
<td>1.5</td>
<td>4</td>
<td>50%</td>
</tr>
<tr>
<td>0.901252</td>
<td>2.692629</td>
<td>1.894444</td>
<td>4.4</td>
<td>55%</td>
</tr>
<tr>
<td>0.695652</td>
<td>3.130435</td>
<td>2.4</td>
<td>4.8</td>
<td>60%</td>
</tr>
<tr>
<td>0.521971</td>
<td>3.600533</td>
<td>3.064286</td>
<td>5.2</td>
<td>65%</td>
</tr>
<tr>
<td>0.376963</td>
<td>4.104712</td>
<td>3.966667</td>
<td>5.6</td>
<td>70%</td>
</tr>
<tr>
<td>0.258065</td>
<td>4.645161</td>
<td>5.25</td>
<td>6</td>
<td>75%</td>
</tr>
<tr>
<td>0.163265</td>
<td>5.22449</td>
<td>7.2</td>
<td>6.4</td>
<td>80%</td>
</tr>
<tr>
<td>0.091024</td>
<td>5.845765</td>
<td>10.48333</td>
<td>6.8</td>
<td>85%</td>
</tr>
<tr>
<td>0.040201</td>
<td>6.512563</td>
<td>17.1</td>
<td>7.2</td>
<td>90%</td>
</tr>
<tr>
<td>0.010013</td>
<td>7.229036</td>
<td>37.05</td>
<td>7.6</td>
<td>95%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Slope</th>
<th>Intercept</th>
<th>WIP</th>
<th>Capacity</th>
<th>Capacity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>7.735018</td>
<td>0.015908</td>
<td>0.101369</td>
<td>0.8</td>
<td>5%</td>
</tr>
<tr>
<td>7.265679</td>
<td>0.089509</td>
<td>0.207894</td>
<td>1.6</td>
<td>10%</td>
</tr>
<tr>
<td>6.680445</td>
<td>0.245115</td>
<td>0.322566</td>
<td>2.4</td>
<td>15%</td>
</tr>
<tr>
<td>6.026945</td>
<td>0.496862</td>
<td>0.448509</td>
<td>3.2</td>
<td>20%</td>
</tr>
<tr>
<td>5.342435</td>
<td>0.851288</td>
<td>0.589378</td>
<td>4</td>
<td>25%</td>
</tr>
<tr>
<td>4.656711</td>
<td>1.308986</td>
<td>0.749674</td>
<td>4.8</td>
<td>30%</td>
</tr>
<tr>
<td>3.992671</td>
<td>1.866316</td>
<td>0.935134</td>
<td>5.6</td>
<td>35%</td>
</tr>
<tr>
<td>3.368826</td>
<td>2.517081</td>
<td>1.153288</td>
<td>6.4</td>
<td>40%</td>
</tr>
<tr>
<td>2.790095</td>
<td>3.253974</td>
<td>1.414298</td>
<td>7.2</td>
<td>45%</td>
</tr>
<tr>
<td>2.268816</td>
<td>4.069725</td>
<td>1.732302</td>
<td>8</td>
<td>50%</td>
</tr>
<tr>
<td>1.805804</td>
<td>4.957911</td>
<td>2.127634</td>
<td>8.8</td>
<td>55%</td>
</tr>
<tr>
<td>1.401337</td>
<td>5.91348</td>
<td>2.630717</td>
<td>9.6</td>
<td>60%</td>
</tr>
<tr>
<td>Slope</td>
<td>Intercept</td>
<td>WIP</td>
<td>Capacity</td>
<td>Capacity %</td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>--------</td>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>7.93296</td>
<td>0.006566</td>
<td>0.15044</td>
<td>1.2</td>
<td>5%</td>
</tr>
<tr>
<td>7.749509</td>
<td>0.049583</td>
<td>0.303299</td>
<td>2.4</td>
<td>10%</td>
</tr>
<tr>
<td>7.452758</td>
<td>0.164306</td>
<td>0.460996</td>
<td>3.6</td>
<td>15%</td>
</tr>
<tr>
<td>7.046916</td>
<td>0.38609</td>
<td>0.626361</td>
<td>4.8</td>
<td>20%</td>
</tr>
<tr>
<td>6.542256</td>
<td>0.747517</td>
<td>0.802855</td>
<td>6</td>
<td>25%</td>
</tr>
<tr>
<td>5.955802</td>
<td>1.274909</td>
<td>0.994843</td>
<td>7.2</td>
<td>30%</td>
</tr>
<tr>
<td>5.310296</td>
<td>1.985377</td>
<td>1.20796</td>
<td>8.4</td>
<td>35%</td>
</tr>
<tr>
<td>4.631962</td>
<td>2.885269</td>
<td>1.449652</td>
<td>9.6</td>
<td>40%</td>
</tr>
<tr>
<td>3.94769</td>
<td>3.970425</td>
<td>1.730018</td>
<td>10.8</td>
<td>45%</td>
</tr>
<tr>
<td>3.282319</td>
<td>5.228106</td>
<td>2.063143</td>
<td>12</td>
<td>50%</td>
</tr>
<tr>
<td>2.656594</td>
<td>6.640026</td>
<td>2.469318</td>
<td>13.2</td>
<td>55%</td>
</tr>
<tr>
<td>2.08608</td>
<td>8.185721</td>
<td>2.978927</td>
<td>14.4</td>
<td>60%</td>
</tr>
<tr>
<td>1.581019</td>
<td>9.845616</td>
<td>3.639667</td>
<td>15.6</td>
<td>65%</td>
</tr>
<tr>
<td>1.146907</td>
<td>11.6034</td>
<td>4.530971</td>
<td>16.8</td>
<td>70%</td>
</tr>
<tr>
<td>0.78547</td>
<td>13.4476</td>
<td>5.795764</td>
<td>18</td>
<td>75%</td>
</tr>
<tr>
<td>0.495791</td>
<td>15.37258</td>
<td>7.719822</td>
<td>19.2</td>
<td>80%</td>
</tr>
<tr>
<td>0.275387</td>
<td>17.37904</td>
<td>10.96987</td>
<td>20.4</td>
<td>85%</td>
</tr>
<tr>
<td>0.121141</td>
<td>19.47446</td>
<td>17.54596</td>
<td>21.6</td>
<td>90%</td>
</tr>
<tr>
<td>0.030076</td>
<td>21.67373</td>
<td>37.44815</td>
<td>22.8</td>
<td>95%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Slope</th>
<th>Intercept</th>
<th>WIP</th>
<th>Capacity</th>
<th>Capacity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>7.979242</td>
<td>0.00286</td>
<td>0.200162</td>
<td>1.6</td>
<td>5%</td>
</tr>
<tr>
<td>7.901114</td>
<td>0.027471</td>
<td>0.401529</td>
<td>3.2</td>
<td>10%</td>
</tr>
<tr>
<td>7.748335</td>
<td>0.105773</td>
<td>0.605837</td>
<td>4.8</td>
<td>15%</td>
</tr>
<tr>
<td>7.506812</td>
<td>0.278924</td>
<td>0.815403</td>
<td>6.4</td>
<td>20%</td>
</tr>
<tr>
<td>Slope</td>
<td>Intercept</td>
<td>WIP</td>
<td>Capacity</td>
<td>Capacity %</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>-------</td>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td>7.166479</td>
<td>0.595073</td>
<td>1.033273</td>
<td>8</td>
<td>25%</td>
</tr>
<tr>
<td>6.723414</td>
<td>1.10529</td>
<td>1.263452</td>
<td>9.6</td>
<td>30%</td>
</tr>
<tr>
<td>6.181907</td>
<td>1.857541</td>
<td>1.511259</td>
<td>11.2</td>
<td>35%</td>
</tr>
<tr>
<td>5.555601</td>
<td>2.889598</td>
<td>1.783858</td>
<td>12.8</td>
<td>40%</td>
</tr>
<tr>
<td>4.867041</td>
<td>4.222574</td>
<td>2.091091</td>
<td>14.4</td>
<td>45%</td>
</tr>
<tr>
<td>4.145364</td>
<td>5.857095</td>
<td>2.446807</td>
<td>16</td>
<td>50%</td>
</tr>
<tr>
<td>3.422569</td>
<td>7.773542</td>
<td>2.871077</td>
<td>17.6</td>
<td>55%</td>
</tr>
<tr>
<td>2.729331</td>
<td>9.936424</td>
<td>3.394083</td>
<td>19.2</td>
<td>60%</td>
</tr>
<tr>
<td>2.09149</td>
<td>12.30159</td>
<td>4.063329</td>
<td>20.8</td>
<td>65%</td>
</tr>
<tr>
<td>1.527976</td>
<td>14.82419</td>
<td>4.958071</td>
<td>22.4</td>
<td>70%</td>
</tr>
<tr>
<td>1.050371</td>
<td>17.46558</td>
<td>6.221061</td>
<td>24</td>
<td>75%</td>
</tr>
<tr>
<td>0.663785</td>
<td>20.19818</td>
<td>8.137915</td>
<td>25.6</td>
<td>80%</td>
</tr>
<tr>
<td>0.368501</td>
<td>23.00823</td>
<td>11.3752</td>
<td>27.2</td>
<td>85%</td>
</tr>
<tr>
<td>0.161871</td>
<td>25.8972</td>
<td>17.93284</td>
<td>28.8</td>
<td>90%</td>
</tr>
<tr>
<td>0.040129</td>
<td>28.88269</td>
<td>37.81072</td>
<td>30.4</td>
<td>95%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Slope</th>
<th>Intercept</th>
<th>WIP</th>
<th>Capacity</th>
<th>Capacity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>7.992633</td>
<td>0.001318</td>
<td>0.250066</td>
<td>2</td>
<td>5%</td>
</tr>
<tr>
<td>7.956575</td>
<td>0.01564</td>
<td>0.500763</td>
<td>4</td>
<td>10%</td>
</tr>
<tr>
<td>7.87416</td>
<td>0.068456</td>
<td>0.753292</td>
<td>6</td>
<td>15%</td>
</tr>
<tr>
<td>7.728007</td>
<td>0.198761</td>
<td>1.009476</td>
<td>8</td>
<td>20%</td>
</tr>
<tr>
<td>7.501032</td>
<td>0.459465</td>
<td>1.271896</td>
<td>10</td>
<td>25%</td>
</tr>
<tr>
<td>7.178144</td>
<td>0.91609</td>
<td>1.544119</td>
<td>12</td>
<td>30%</td>
</tr>
<tr>
<td>6.74899</td>
<td>1.642302</td>
<td>1.831044</td>
<td>14</td>
<td>35%</td>
</tr>
<tr>
<td>6.211236</td>
<td>2.711459</td>
<td>2.139436</td>
<td>16</td>
<td>40%</td>
</tr>
<tr>
<td>5.573497</td>
<td>4.184697</td>
<td>2.47875</td>
<td>18</td>
<td>45%</td>
</tr>
<tr>
<td>4.856652</td>
<td>6.098023</td>
<td>2.862461</td>
<td>20</td>
<td>50%</td>
</tr>
<tr>
<td>4.092515</td>
<td>8.452617</td>
<td>3.310283</td>
<td>22</td>
<td>55%</td>
</tr>
<tr>
<td>3.319664</td>
<td>11.2125</td>
<td>3.852047</td>
<td>24</td>
<td>60%</td>
</tr>
<tr>
<td>2.577485</td>
<td>14.31132</td>
<td>4.534916</td>
<td>26</td>
<td>65%</td>
</tr>
<tr>
<td>1.90032</td>
<td>17.66641</td>
<td>5.437816</td>
<td>28</td>
<td>70%</td>
</tr>
<tr>
<td>1.313506</td>
<td>21.19534</td>
<td>6.703177</td>
<td>30</td>
<td>75%</td>
</tr>
<tr>
<td>0.832109</td>
<td>24.8303</td>
<td>8.616298</td>
<td>32</td>
<td>80%</td>
</tr>
<tr>
<td>0.462057</td>
<td>28.52765</td>
<td>11.84344</td>
<td>34</td>
<td>85%</td>
</tr>
<tr>
<td>0.202755</td>
<td>32.27251</td>
<td>18.38422</td>
<td>36</td>
<td>90%</td>
</tr>
<tr>
<td>0.050196</td>
<td>36.0806</td>
<td>38.23824</td>
<td>38</td>
<td>95%</td>
</tr>
</tbody>
</table>