

The Pennsylvania State University
The Graduate School
Harold and Inge Marcus Department of Industrial and Manufacturing Engineering

**WORKFORCE PLANNING MODELS FOR
DISTRIBUTION CENTER OPERATIONS**

A Thesis in
Industrial Engineering
by
Athul Gopala Krishna

© 2016 Athul Gopala Krishna

Submitted in Partial Fulfillment
of the Requirements
for the Degree of
Master of Science

May 2016

The thesis of Athul Gopala Krishna was reviewed and approved* by the following:

Vittaladas V. Prabhu
Professor of Industrial and Manufacturing Engineering
Thesis Adviser

A. Ravi Ravindran
Professor of Industrial and Manufacturing Engineering

Janis P. Terpenney
Professor of Industrial and Manufacturing Engineering
Peter and Angela Dal Pezzo Department Head of Industrial and Manufacturing Engineering

*Signatures are on file in the Graduate School.

ABSTRACT

Customer order fulfillment at distribution centers (DC) is increasingly necessitated by innovative strategies to maximize operational performance that are primarily driven by cost and service level under supply chain variability. In order to better understand the trade-offs, in this thesis, a generic computational model is developed to estimate forklift travel times for DCs with any arbitrary floor space and loading docks. In particular, travel times are modelled as random variables and the moments of the probability distribution of travel times are estimated and used as inputs to analytical queueing model and discrete event simulation model. Results show that the analytical and simulation models are within 3% under different demand scenarios. These models are used to determine the impact of workforce capacity on key performance measures such as Truck Processing Time (TPT) and Labor Hours Per Truck (LHPT). The workforce capacity for different demand scenarios is determined using three different approaches - Target Utilization Level, Square Root Staffing (SRS) rule (adapted from call center staffing) and Optimization. The result from these models indicate that adapting workforce capacity to match varying demand can reduce cost by 18% while maintaining desired service level.

Keywords: Distribution Center; Decision Model; Workforce Capacity; Simulation

TABLE OF CONTENTS

List of Tables.....	vii
List of Figures.....	viii
Acknowledgements.....	x
CHAPTER 1 INTRODUCTION.....	1
1.1 Distribution Center Operations.....	1
1.2 Literature Summary and Research Motivation.....	2
1.3 Overview.....	3
1.4 Organization.....	4
CHAPTER 2 LITERATURE REVIEW.....	6
2.1 Warehouse Design and Performance Analysis.....	6
2.1.1 Travel Time Models for Warehouse Design.....	6
2.1.2 Performance Analysis under Class-based Storage Policy.....	7
2.1.3 Performance Analysis of AS/RS.....	7
2.1.4 Performance Analysis of Order Picking System.....	8
2.2 Queueing Models in Production Systems.....	8

2.2.1	Little's Law.....	9
2.2.2	Variability.....	9
2.2.3	Queueing System.....	10
2.2.4	Queueing Notation and Performance Measures.....	10
2.2.5	Kendall's Notation.....	12
2.2.6	M/M/1 Queueing Model.....	12
2.2.7	G/G/1 Queueing Model.....	13
2.2.8	M/M/m Queueing Model.....	13
2.2.9	G/G/m Queueing Model.....	13
2.3	Service Engineering.....	14
2.3.1	Operational Regimes.....	14
2.3.2	Square Root Staffing Rule.....	15
2.4	Chapter Summary and Conclusions.....	15
CHAPTER 3 MODEL AND ASSUMPTIONS.....		16
3.1	Travel Time Computational Model.....	16
3.2	Queueing Model.....	18
3.3	Performance Measures.....	20
3.4	Chapter Summary and Conclusions.....	21

CHAPTER 4	PERFORMANCE ANALYSIS.....	22
4.1	Performance Analysis using Queueing Model.....	22
4.2	Workforce Capacity Analysis using Queueing Model.....	23
4.3	Chapter Summary and Conclusions.....	25
CHAPTER 5	SQARE ROOT STAFFING RULE AND DC WORKFORCE CAPACITY...26	
5.1	Workforce Capacity Analysis using SRS rule.....	26
5.2	Chapter Summary and Conclusions.....	28
CHAPTER 6	SIMULATION MODEL AND ANALYSIS.....	30
6.1	Simulation Model.....	30
6.1.1	Performance Analysis using Simulation Model.....	31
6.2	Simulation-based Optimization Model.....	33
6.2.1	Multi-Objective Optimization.....	34
6.2.2	Pattern Frontier Optimization.....	35
6.3	Chapter Summary and Conclusions.....	38
CHAPTER 7	SUMMARY AND CONCLUSIONS.....	39
REFERENCES.....		41
APPENDIX.....		43

LIST OF TABLES

Table 4.1	Queueing Model Output.....	22
Table 4.2	Workforce Capacity Model Output.....	24
Table 6.1	Simulation Model Output.....	31
Table 6.2	Optimal Workforce Capacity.....	34
Table 6.3	Optimal TPT, LHPT and Operating Cost.....	35

LIST OF FIGURES

Figure 1.1	DC Functions.....	1
Figure 1.2	Workforce Planning Models for DCs.....	3
Figure 3.1	Skewness of Travel Time Distribution against L and W of DC (view 1).....	16
Figure 3.2	Skewness of Travel Time Distribution against L and W of DC (view 2).....	17
Figure 3.3	Skewness for Travel Time Distribution with Distance from Center Dock.....	18
Figure 3.4	DC Layout.....	18
Figure 3.5	Outbound Queueing Model.....	19
Figure 4.1	Comparison - TPT v/s Arrival Rate.....	23
Figure 4.2	Comparison - Workforce Capacity v/s TPT.....	24
Figure 4.3	Comparison - Workforce Capacity v/s LHPT.....	25
Figure 5.1	DC Workforce Level from SRS rule.....	26
Figure 5.2	Workforce Capacity for Demand Scenarios.....	27
Figure 5.3	Workforce Level for Demand Scenarios.....	28
Figure 6.1	Simulation Model.....	30
Figure 6.2	Cycle Time - Analytical v/s Simulation Output.....	32
Figure 6.3	WIP Level - Analytical v/s Simulation Output.....	32

Figure 6.4	Workforce Utilization - Analytical v/s Simulation Output.....	33
Figure 6.5	Optimal Workforce Capacity.....	35
Figure 6.6	Optimal Scenarios - TPT v/s Operating Cost.....	36
Figure 6.7	Optimal Scenarios - Operating Cost v/s LHPT.....	36
Figure 6.8	Optimal Scenarios - Workforce Capacity.....	37
Figure 6.9	Optimal Scenarios - TPT v/s LHPT.....	37

ACKNOWLEDGEMENTS

I am extremely grateful to my adviser, Prof. Vittaldas V. Prabhu, who expressed immense trust in me and motivated me to work smarter and harder with confidence.

I am thankful to Prof. A. Ravi Ravindran for his contributions in improving the quality of the research.

I am thankful to Prof. Jeffrey Smith, Department of Industrial and Systems Engineering, Auburn University, for his expert guidance during the research.

I am thankful to my friends, Kai-Wen Tien and Rakshith Badrinath, for their wholehearted support during the research.

DEDICATION

to my mother, Asha

CHAPTER 1

INTRODUCTION

1.1 Distribution Center Operations

Supply chain variability can be caused by product seasonality, batch production and transportation, product consolidation or value added processing. Distribution Centers (DCs) buffer the material flow in supply chains to accommodate this variability. Incoming items brought to the DC are unloaded at the receiving docks (receiving) and put into storage (storing). Outgoing items are retrieved from storage (order picking), processed and shipped to customers through the shipping docks (shipping). Resources such as space, labor, and equipment are allocated to different DC functions following organizational policies to achieve desired operational performance in terms of capacity, throughput and service at minimum cost. DCs can adapt to varying demand by adjusting workforce capacity to meet desired service level. The different DC functions are shown in Figure 1.1, [1], [2].

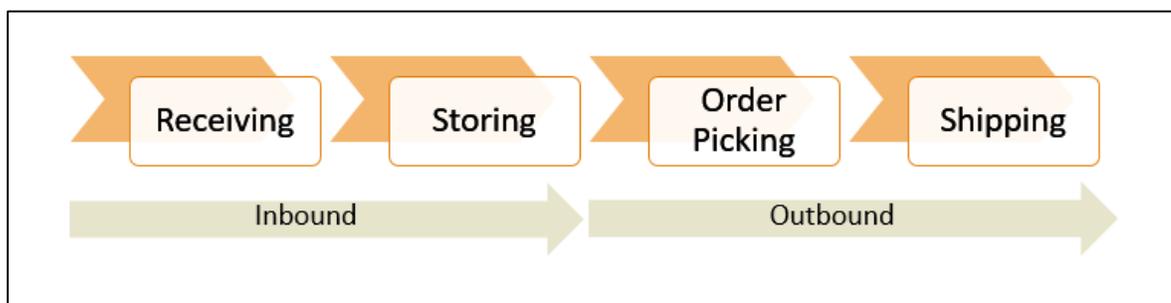


Figure 1.1. DC Functions

1.2 Literature Summary and Research Motivation

Several authors had investigated the effects of warehouse design and control on operational performance and developed analytical and simulation models for performance analysis. Pandit and Palekar (1993) investigated the effects of warehouse design on response time and suggested a method for optimal design based on response time, [3]. Chew and Tang (1999) presented a travel time model that evaluates performance of an order picking system with consideration to order batching and storage allocation strategies, [4]. Graves et al. (1977) evaluated warehouse performance for several sequencing and class based storage policies using continuous analytical models and discrete evaluation procedures, [5]. Bozer and White (1990) modelled the performance of an end-of-aisle order picking system by deriving analytical expressions and developed a design algorithm to determine the optimal configuration, [6]. Koster (1994) proposed a modeling and approximate analysis method for a pick-to-belt order picking system based on Jackson network modeling and analysis, [7]. Lee (1997) presented an analytical method for the stochastic analysis of a unit load AS/RS, [8]. Hur et al. (2004) presented an analytical model for the stochastic analysis of a unit load AS/RS without assuming any specific distribution for the travel time of the S/R machine, [9].

The existing literature is heavily focused on warehouse design & control using travel time models and performance analysis in terms of throughput, resource utilization and storage strategy for order picking systems. As labor cost can be a significant component of variable cost in DCs, it is evident that there is a need for research to explore the impact of workforce capacity on DC performance by considering cost and service level trade-off for varying demand scenarios.

1.3 Overview

The thesis develops a decision model that integrates critical operational performance measures to evaluate DC performance and workforce capacity policies under varying demand scenarios.

The framework for workforce planning models is shown in Figure 1.2.

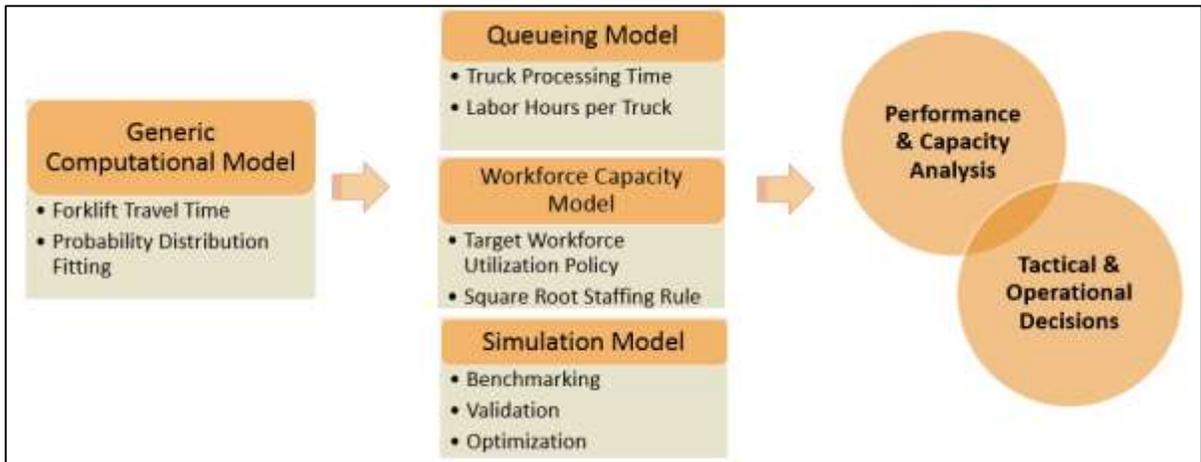


Figure 1.2: Workforce Planning Models for DCs

Generic Computational Model

Labor cost can be a significant component of variable cost in DCs owing to variability in demand. The desired service level in DCs are also to be met for different demand scenarios. In order to better understand the trade-offs between cost and service level, a large scale, non-automated, rectangular DC is analyzed and a generic computational model with length (L), width (W), aisle-width (A) and number of docks (N) as parameters is developed to estimate forklift travel times in DCs. The skewness profile of forklift travel time distribution against L & W and distance from center dock is investigated. The travel times are modelled as random

variables and the moments of the probability distribution of travel times are estimated and used as inputs to analytical queueing model and discrete event simulation model.

Analytical and Simulation Model

The analytical and simulation models are used to determine the impact of workforce capacity on key performance measures such as Truck Processing Time (TPT) and Labor Hours Per Truck (LHPT). Results show that the analytical and simulation models are within 3% under different demand scenarios. The workforce capacity for different demand scenarios is determined using three different approaches - Target Utilization Level, Square Root Staffing (SRS) rule (adapted from call center staffing) and Optimization. The target utilization level is taken as 70% for DC workforce capacity analysis. The SRS rule is applied for efficiency driven, quality driven and quality efficiency driven operational regimes in DC. Multi-objective weighted optimization and pattern frontier optimization are used to determine optimal workforce capacity scenarios in DC. The result from these models indicate that adapting workforce capacity to match varying demand can reduce cost by 18% while maintaining desired service level.

1.4 Organization

The thesis is organized as follows:

Literature review on warehouse design, control, performance evaluation & operations, queueing models in production systems and service engineering in call centers is presented in

chapter 2. The travel time computational model, investigations on skewness profile for forklift travel time distributions and the DC queueing model are explained in chapter 3. The performance analysis results obtained from the queueing model is provided in chapter 4. The workforce capacity estimation results by the application of SRS rule is produced in chapter 5. The benchmarking and validation results from the discrete event simulation model are detailed in chapter 6. The conclusions arrived at from the research are presented in chapter 7.

CHAPTER 2

LITERATURE REVIEW

2.1 Warehouse Design and Performance Analysis

2.1.1 Travel Time Models for Warehouse Design

Pandit and Palekar (1993) investigated the effects of warehouse design on response time and suggested a method for optimal design based on response time. The multi-vehicle material handling system in an automated rectangular warehouse was modelled as an M/G/K queueing system. An expression for waiting time in the queue was approximated after computing first and second moments of the service time distribution. The optimal layout design for the warehouse was determined by modelling a nonlinear integer programming problem that minimizes response time and solving it by Generalized Reduced Gradient method. It was concluded that the effect of door location on optimal layout design was significant. Using a simulation model, it was concluded that congestion does not have a significant effect on response time, [3].

Chew and Tang (1999) presented a travel time model that evaluates performance of an order picking system with consideration to order batching and storage allocation strategies. The order picking system for a rectangular warehouse under general item location assignment was modelled as an $E_k/G/c$ queueing system. The exact probability mass function that characterize the tour of an order picker was determined. The first and second moments of expected travel time, its bounds and approximation was derived from the model as a performance measure. The expected bounds and approximation for turnover time was also derived as a service level

measure. The travel time model was validated using simulation. It was concluded that skewed item location performed better than uniform item location for a given batch size, [4].

2.1.2 Performance Analysis under Class-based Storage Policy

Graves et al. (1977) evaluated warehouse performance for several sequencing and class based storage policies using continuous analytical models and discrete evaluation procedures. Operating performance was measured in terms of expected one-way travel time, expected interleave time and expected round-trip time. It was established that significant travel time reduction was obtained from dedicated storage over random storage. It was also determined that class based storage with relatively few classes yields travel time reductions comparable to dedicated storage, [5].

2.1.3 Performance Analysis of Order Picking Systems

Bozer and White (1990) modelled the performance of an end-of-aisle order picking system by deriving analytical expressions and developed a design algorithm to determine the optimal configuration, [6].

Koster (1994) proposed a modeling and approximate analysis method for a pick-to-belt order picking system based on Jackson network modeling and analysis. It evaluated performance measures such as system throughput, picker utilization and average number of bins in the system based on factors such as the speed and length of the conveyor, the number of picking stations and the number of picks per station, [7].

2.1.4 Performance Analysis of AS/RS

Lee (1997) presented an analytical method for the stochastic analysis of a unit load AS/RS. A unit load AS/RS was modelled as an M/M/1 queueing system with two queues (storage and retrieval) and two different service modes (single command and dual command). The performance of AS/RS was evaluated using analytical expressions for system throughput, S/R machine utilization and turnaround time, [8].

Hur et al. (2004) presented an analytical model for the stochastic analysis of a unit load AS/RS without assuming any specific distribution for the travel time of the S/R machine. A unit load AS/RS was modelled as an M/G/1 queueing system with two queues (storage and retrieval) and two different service modes (single command and dual command). The performance of AS/RS was evaluated using analytical expressions for the probability distribution and the expected number of commands waiting in the storage and retrieval queues, [9].

2.2 Queueing Models in Production Systems

Wallace and Mark (2001) defines a production system as “an objective-oriented network of processes through which entities flow.” The “objective” is to generate revenue; “process” include usual physical and ancillary production processes; “entities” comprise parts produced/processed in the system and information used to control the system; “flow” describes how materials and information are processed; production lines connect individual processes and the overall system to create a “network” of interacting parts, [12].

2.2.1 Little's Law

Littles' law (1964), provides a fundamental relationship between three long-term average measures of performance of any production system and can be applied in queue length calculations, cycle time measures and inventory planning.

$$WIP = TH \times CT$$

where,

WIP: Work in Process, the inventory between start and end points of a product routing

TH: Throughput, the average output of a production process per unit time

CT: Cycle time, the average time from the release of a job at the beginning of the routing until it reaches an inventory point at the end of the routing (i.e., the time a part spends as WIP)

2.2.2 Variability

Variability is the “quality of non-uniformity of a class of entities”, [12]. It has a significant impact on the performance of any production system and is a critical measure in production management. It can be caused by the process owing to cycle time fluctuation, outage or setup change or by flow during transfer of parts from one station to the other.

Coefficient of Variation (CV) is a relative measure of variability of a random variable and is denoted by,

$$CV = \frac{\sigma}{t}$$

where,

σ : Standard Deviation

t: Mean

A random variable can have three classes of variability: (i) Low Variability (LV), if CV is less than 0.75; (ii) Moderate Variability (MV), if CV is between 0.75 and 1.33; (iii) High Variability (HV), if CV is greater than 1.33.

2.2.3 Queueing System

High levels of variability cause waiting and waiting time is mostly the largest portion of cycle time in a production system. “The science of waiting” is called queueing theory, [12]. A queueing system comprise of an arrival process, service process and a queue. The part can arrive individually or as a batch and can have identical or different characteristics. Inter-arrival time and process time can be constant or random. The parts can be processed by a single machine or several machines in parallel following a queueing discipline such as first-come-first-served (FCFS), earliest due date (EDD), shortest process time (SPT) or last-come-first-served (LCFS). The queue space can be finite or infinite.

2.2.4 Queueing Notation and Performance Measures

The following parameters are assumed to be known to apply queueing theory and analyze the performance of a system, [12]:

r_a = Rate of arrivals in parts per unit time

$$\begin{aligned}
 t_a &= \left. \begin{array}{l} \text{Average time between arrivals (in minute)} \\ \frac{1}{r_a} \end{array} \right\} \\
 c_a &= \text{Arrival CV (Coefficient of Variation)} \\
 m &= \text{Number of machines available for a process} \\
 t_e &= \text{Mean effective process time} \\
 c_e &= \text{CV of effective process time}
 \end{aligned}$$

The performance of a queueing system is characterized by the following parameters, [12]:

$$\begin{aligned}
 u &= \left. \begin{array}{l} \text{Utilization} \\ \frac{r_a}{m} t_e \end{array} \right\} \\
 CT_q &= \text{Expected waiting time in queue for a process} \\
 CT &= \left. \begin{array}{l} \text{Expected time for a process} \\ CT_q + t_e \end{array} \right\} \\
 WIP &= \left. \begin{array}{l} \text{Average work-in-process level at process} \\ TH \times CT \end{array} \right\} \\
 WIP_q &= \left. \begin{array}{l} \text{Expected WIP in queue} \\ r_a \times CT_q \end{array} \right\} \\
 c_d &= \left. \begin{array}{l} \text{Departure CV} \\ \sqrt{c_e^2 u^2 + c_a^2 (1 - u^2)} \end{array} \right\}
 \end{aligned}$$

2.2.5 Kendall's Notation

A queueing system is described by Kendall's notation as A/B/m/b.

where,

A: Inter-arrival time distribution

B: Process time distribution

m: Number of machines available for a process

b: Queue space

A and B can be a constant (deterministic) distribution denoted by D, exponential (Markovian) distribution denoted by M or general distribution denoted by G.

2.2.6 M/M/1 Queueing Model

M/M/1 queueing model assumes an exponential inter-arrival time and process time with a single machine for the process that follows a FCFS queueing discipline having unlimited queue space.

$$\text{WIP (M/M/1)} = \frac{u}{1 - u}$$

Using Little's law,

$$\text{CT (M/M/1)} = \frac{\text{WIP (M/M/1)}}{r_a} = \frac{t_e}{1 - u}$$

$$\text{CT}_q \text{ (M/M/1)} = \text{CT (M/M/1)} - t_e = \frac{u}{1 - u} t_e$$

$$WIP_q(M/M/1) = r_a \times CT_q(M/M/1) = \frac{u^2}{1-u}$$

2.2.7 G/G/1 Queueing Model

G/G/1 queueing model assumes a general inter-arrival time and process time with a single machine for the process that follows a FCFS queueing discipline having unlimited queue space.

By Kingman's (1961) equation,

$$CT_q(G/G/1) = \left(\frac{c_a^2 + c_e^2}{2} \right) \left(\frac{u}{1-u} \right) (t_e)$$

2.2.8 M/M/m Queueing Model

M/M/m queueing model assumes an exponential inter-arrival time and process time with m parallel machines for the process that follows a FCFS queueing discipline having unlimited queue space.

By Sakasegawa's (1977) approximation,

$$CT_q(M/M/m) = \frac{u\sqrt{2(m+1)}-1}{m(1-u)} t_e$$

2.2.9 G/G/m Queueing Model

G/G/m queueing model assumes a general inter-arrival time and process time with m parallel machines for the process that follows a FCFS queueing discipline having unlimited queue space.

$$CT_q (G/G/m) = \left(\frac{c_a^2 + c_e^2}{2} \right) \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \right) (t_e)$$

2.3 Service Engineering

Service Engineering “develops scientifically-based design principles and tools that support and balance service quality, efficiency and probability, from conflicting perspectives of customers, servers and managers”, [13]. Contact Centers are service organizations for customers to request service via the phone, fax, e-mail, chat or other tele-communication channels. Call Centers are a type of contact centers that primarily serve phone calls. Call center operations can be viewed as queueing systems.

2.3.1 Operational Regimes

It is imperative to have the appropriate balance between service quality and efficiency. The operational regimes for a queueing system can be a Quality Driven (QD) with emphasis on service quality over efficiency, Efficiency Driven (ED) with emphasis on efficiency over service quality or Quality Efficiency Driven (QED) regime with trade-off between quality and efficiency.

2.3.2 Square Root Staffing (SRS) Rule

Operational preferences can be set by determining service capacity relative to service demand.

The Square Root Staffing (SRS) rule is applied in call centers to determine the appropriate staffing levels (n) for an offered load and quality of service and is approximated as:

$$n = R + \beta\sqrt{R}$$

where,

R: Offered Load, the amount of work that arrives in the system in unit time

β : Quality of Service parameter

The value of β reflects the service-level and operational efficiency trade-off and signifies the operational regime of the system. Larger the β value, better the service level, [13], [14].

2.4 Chapter Summary and Conclusions

- The existing literature is heavily focused on warehouse design & control using travel time models and performance analysis in terms of throughput, resource utilization and storage strategy for order picking systems.
- It is evident that there is a need for research to explore the impact of workforce capacity on DC performance by considering cost and service level trade-off for varying demand scenarios.

CHAPTER 3

MODEL AND ASSUMPTIONS

3.1 Travel Time Computational Model and Analysis

A generic computational model with length (L), width (W), aisle-width (A) and number of docks (N) as parameters is developed to estimate forklift travel times in DCs. The model is capable to evaluate DCs with a size between 60,000 sq.ft (small scale) and 500,000 sq.ft. (large scale) and up to 45 docks. A probability distribution is fit to the forklift travel times and moments of the distribution is provided as input to the DC queueing model, (see Section 3.2).

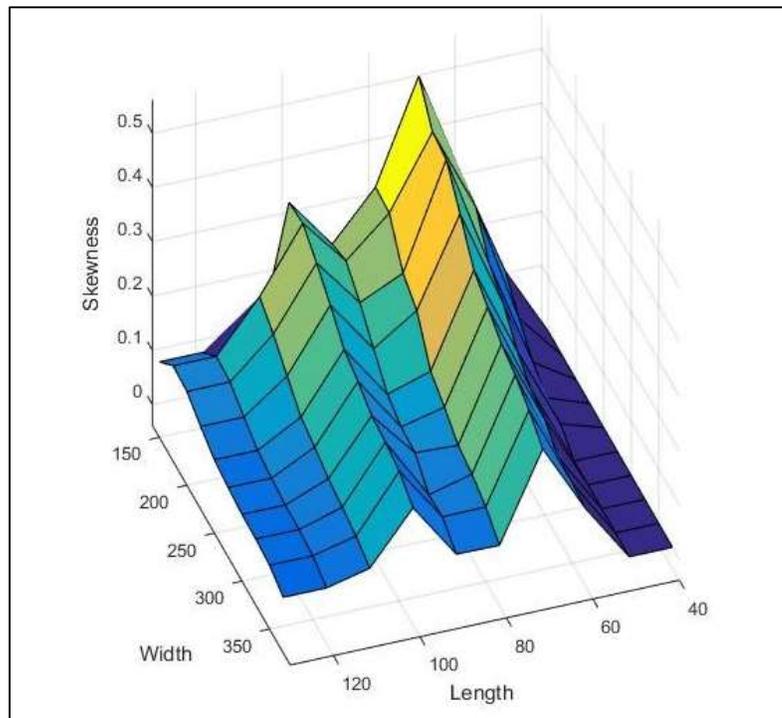


Figure 3.1: Skewness of Travel Time Distribution against L and W of DC (view 1)

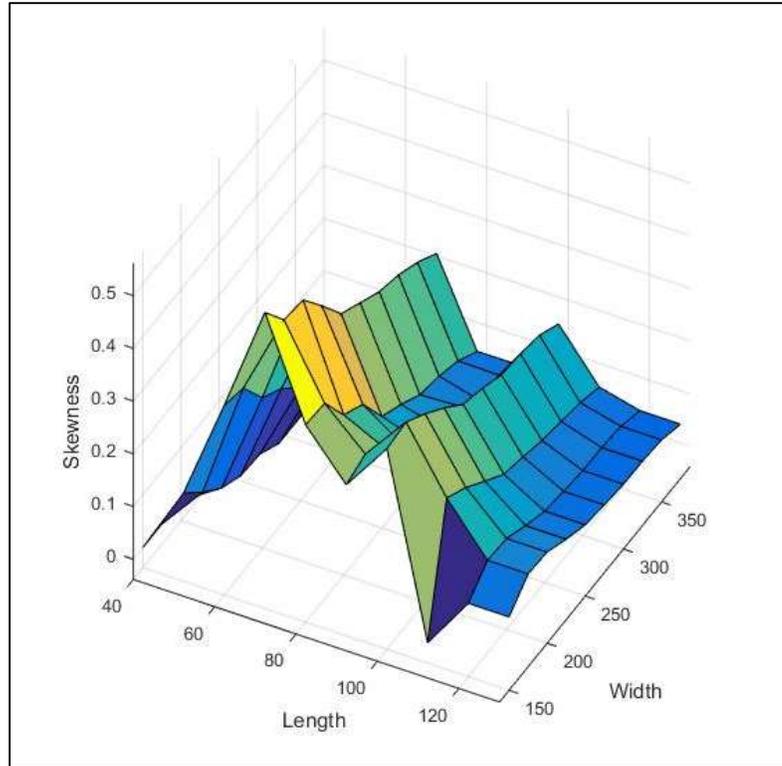


Figure 3.2: Skewness of Travel Time Distribution against L and W of DC (view 2)

The skewness profile of forklift travel time distribution against L and W is investigated. It is observed that for a given W, skewness profile is represented by a series of peaks and valleys with decreasing value as L increases. Similarly, for a given L, skewness profile remains nearly constant as W increases. The skewness profiles are shown in Figure 3.1 and 3.2.

The skewness for forklift travel time distributions fluctuate significantly from dock 1 (end dock) and becomes negligible towards dock 23 (center dock) and remains negligible from the center dock to dock 45 (end dock) as shown in Figure 3.3.

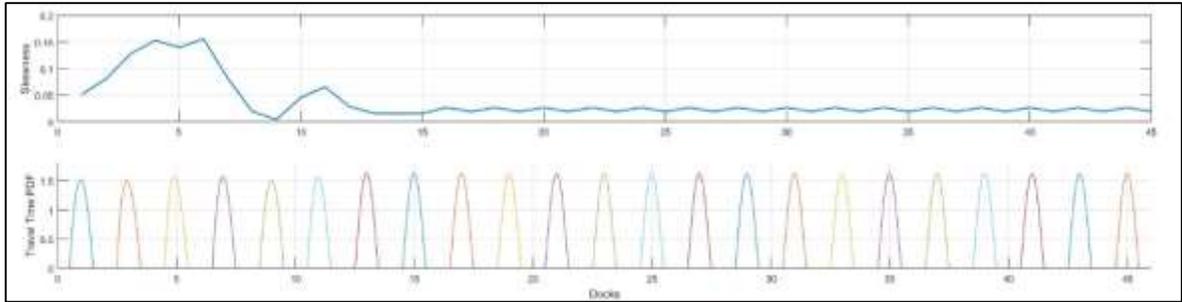


Figure 3.3: Skewness for Travel Time Distribution with Distance from Center Dock

3.2 Queueing Model

A large scale, non-automated, rectangular DC is analyzed. A rectangular shape is the optimal geometrical shape for storing rectangular units such as pallets, [10]. The storage locations are characterized by single-deep racks and drive-in racks. The racks are arranged back-to-back, to form a block, parallel to the dock of the DC, such that space between blocks form aisles. The blocks are arranged in a rectangular grid to form a network of aisles through which material handling devices travel. The DC layout is shown in Figure 3.4, [3].

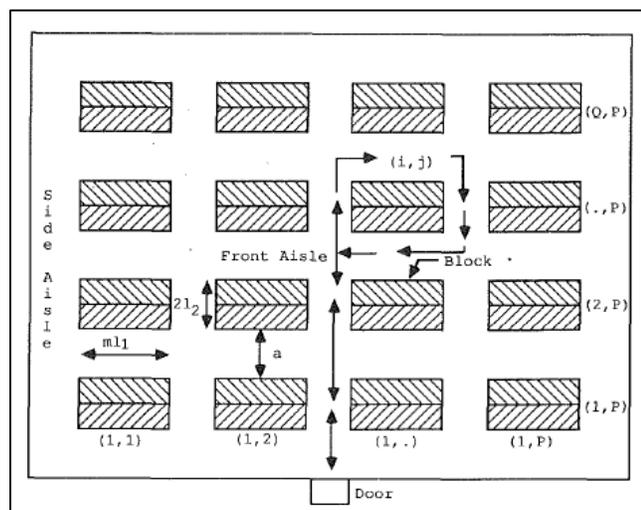


Figure 3.4: DC Layout, [3]

The outbound operations in the DC begins by “accumulation” of customer orders from storage area as pallet loads in the dock to form a truck load. The pallet loads then undergo a “wrapping” process to secure items in the pallet after accumulation. After wrapping, pallet loads are “inspected” to verify that items in the pallets have completed all processes before loading. Pallets are then “loaded” onto the truck for shipping. The process time associated with accumulation, wrapping, inspection and loading are denoted by Accumulation Time (AT) or Forklift Travel Time, Wrapping Time (WT), Inspection Time (IT) and Loading Time (LT) respectively. The process time for a truck load, denoted by Truck Processing Time (TPT), is the sum of AT, travel time taken by forklifts to retrieve a truck load from storage area to the dock; WT, time taken by wrapping team to wrap a truck load; IT, time taken by inspection team to inspect a truck load; and LT, time taken by loading team to load a truck load to a truck for shipping; $TPT = AT + WT + IT + LT$. The operational productivity is expressed by a metric called Labor Hours Per Truck (LHPT). The movement of pallets are done using forklifts and pallet jacks. It is assumed that forklifts perform only accumulation operations and pallet jacks are used by wrapping team, inspection team and loading team to perform wrapping, inspection and loading operations respectively. Fork lifts, pallet jacks, wrapping team, inspection team and loading team are shared resources in the DC. The queuing model is shown in Figure 3.5.

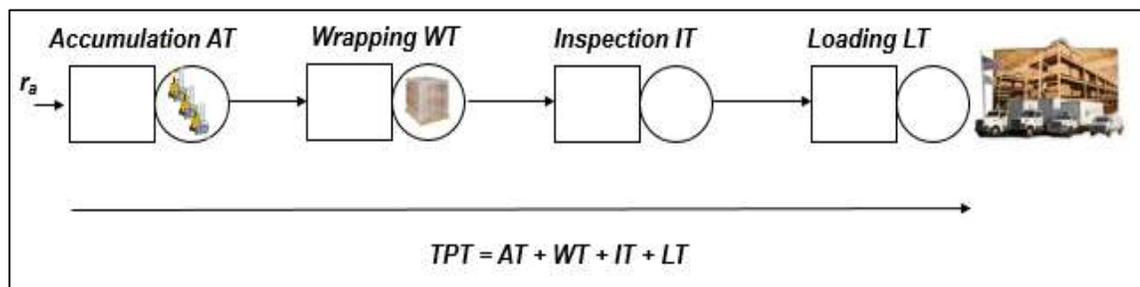


Figure 3.5: Outbound Queuing Model

The outbound operations in the DC can be modelled as an open network queueing system with a First Come First Serve (FCFS) queue discipline and an infinite queue space. The customer orders for accumulation arrive at a rate r_a in the form of palletized truck loads. Forklifts, wrapping team, inspection team and loading team act as servers. The system is characterized by a series of G/G/m queues for accumulation, wrapping, inspection and loading with general (non-exponential) inter-arrival and process time distributions. The model facilitates development and evaluation of workforce capacity policies for the DC for several demand scenarios by providing a performance estimate for each policy-scenario combination.

3.3 Performance Measures

The following parameters are provided as input to the queueing model: r_a - rate of arrivals in truck load per unit time; t_a - Average time between arrivals (in minute); c_a - Arrival CV; c_d - Departure CV; m - Workforce level in a process; t_e - Mean effective process time (in minute); c_e - CV of effective process time, [12].

The performance of the queueing system is characterized by the following parameters and are considered as the output from the queueing model: u - Workforce Utilization; CT_q - Expected waiting time in queue for a process (in minute); CT - Expected time for a process (in minute); TPT - Truck Processing Time (in minute); WIP - Average work-in-process level at process (in truck load); WIP_q - Expected WIP in queue (in truck load); ATL - Average Truck Load; LHPT - Labor Hours Per Truck, [12].

3.4 Chapter Summary and Conclusions

- A generic computational model, with length (L), width (W), aisle-width (A) and number of docks (N) as parameters, is developed to estimate forklift travel times in DCs, which is capable to evaluate DCs with a size between 60,000 sq.ft. (small scale) and 500,000 sq.ft. (large scale) and up to 45 docks. The travel times are modelled as random variables and the moments of travel time distribution is used as input to analytical queueing model and discrete event simulation model of DC.
- The analytical queueing model facilitates development and evaluation of workforce capacity policies for the DC for several demand scenarios by providing performance estimates, particularly in terms of TPT and LHPT, for each policy-scenario combination.

CHAPTER 4

PERFORMANCE ANALYSIS

4.1 Performance Analysis using Queueing Model

Table 4.1: Queueing Model Output

OUTPUT					
Process	Process Time		WIP Levels		Workforce Utilization
	Waiting Time	Cycle Time	WIP Waiting	WIP	
Accumulation	0.30	295.54	0.01	12.76	50.98%
Wrapping	36.21	89.68	1.56	3.87	76.94%
Inspection	43.81	164.03	1.89	7.08	86.49%
Loading	3.82	155.51	0.16	6.71	65.48%
TPT		704.77	ATL	30.42	
LHPT		357.34			

The performance of outbound operations of the DC is analyzed for a given arrival rate, process time and workforce level using the queueing model. The queueing model output is given in Table 4.1. The TPT and LHPT is determined as 704.77 minutes and 357.34 respectively. It is observed that the expected waiting time in queue and workforce utilization for wrapping and inspection process is high (36.21 & 43.81 minutes and 76.94% & 86.49% respectively).

The performance of the DC is analyzed for several demand scenarios that range from an arrival rate of 0.01(low) to 0.09 (high) for a given process time and workforce level. It is observed that TPT / LHPT remains almost constant till an arrival rate of 0.04 (medium demand) after which it increases sharply as workforce utilization approaches 100% as shown in Figure 4.1.

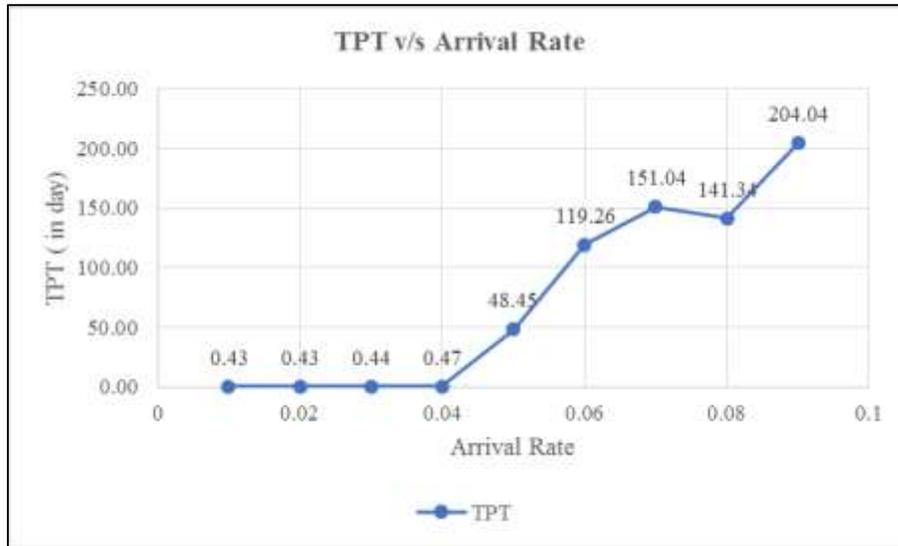


Figure 4.1: Comparison - TPT v/s Arrival Rate

4.2 Workforce Capacity Analysis using Queueing Model

The queueing model is used to determine the workforce capacity at different stages of operations to meet desired performance levels. The workforce utilization becomes an input to the queueing model and workforce level at accumulation, wrapping, inspection and loading process is the output from the queueing model. All other input and output parameters remains same as that of the initial model.

The workforce capacity of outbound operations of the DC is analyzed for a given arrival rate and process time at 70% target workforce utilization for all the process using the model. The workforce capacity model output is given in Table 4.2. The total workforce level (M) is determined as 37 for the DC with a TPT of 664.22 minutes and LHPT of 317.40. It is observed that the expected waiting time in queue for wrapping is high under the present target workforce utilization.

Table 4.2: Workforce Capacity Model Output

OUTPUT					
Process	Process Time		WIP Levels		Workforce Level
	Waiting Time	Cycle Time	WIP Waiting	WIP	
Accumulation	7.17	302.40	0.31	13.05	18
Wrapping	20.49	73.96	0.88	3.19	3
Inspection	8.56	128.78	0.37	5.56	7
Loading	7.39	159.08	0.32	6.87	9
TPT		664.22	ATL	28.67	37
LHPT		317.40	M		

The workforce capacity of the DC is analyzed for several demand scenarios that range from an arrival rate of 0.01 (low) to 0.09 (high) for a given process time at 70% target workforce utilization for all the process. It is observed that TPT decreases steadily as M increases and LHPT increases linearly with M as shown in Figure 4.2 and 4.3.

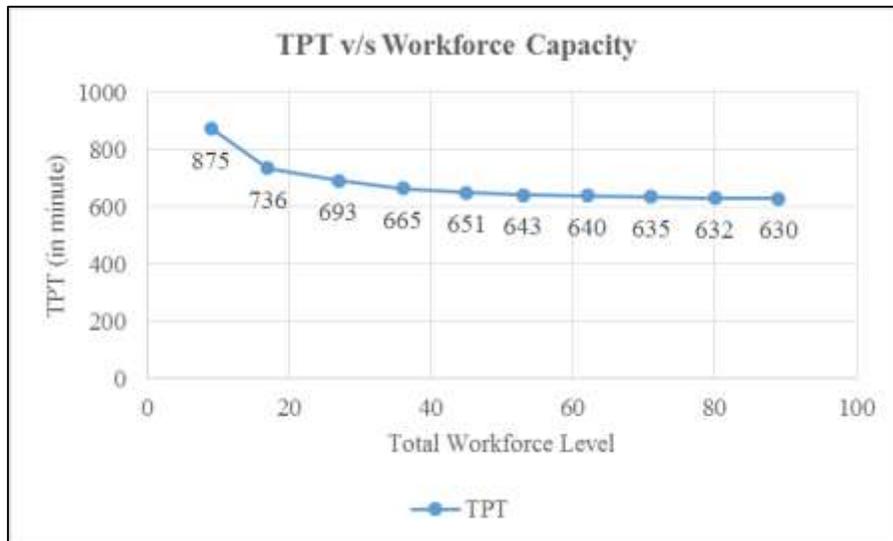


Figure 4.2: Comparison - TPT v/s Workforce Capacity

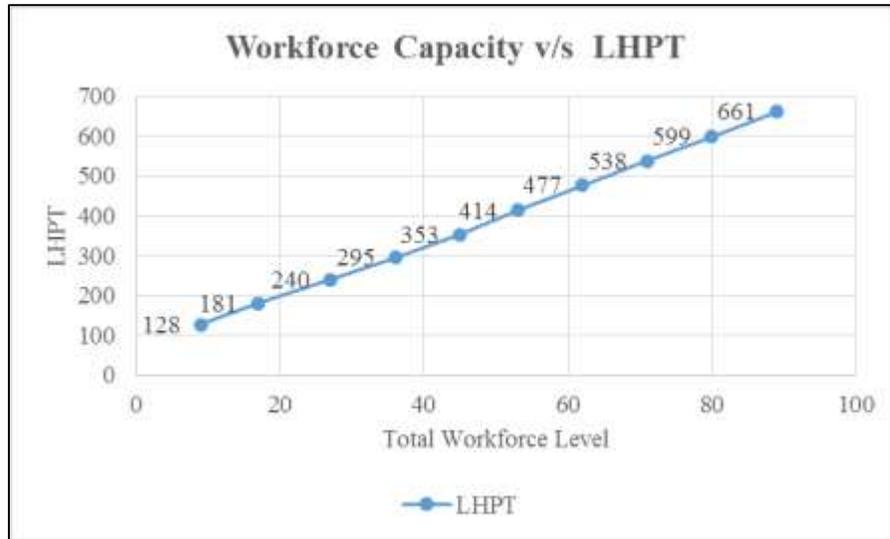


Figure 4.3: Comparison - Workforce Capacity v/s LHPT

4.3 Chapter Summary and Conclusions

- The performance of the DC is analyzed for several demand scenarios that range from an arrival rate of 0.01(low) to 0.09 (high) for a given process time and workforce level using the analytical queueing model. It is observed that TPT / LHPT remains almost constant till an arrival rate of 0.04 (medium demand) after which it increases sharply.
- The workforce capacity of the DC is analyzed for several demand scenarios that range from an arrival rate of 0.01 (low) to 0.09 (high) for a given process time at 70% target workforce utilization for all the process. It is observed that TPT decreases steadily as M increases and LHPT increases linearly with M.

CHAPTER 5

SQUARE ROOT STAFFING (SRS) RULE AND DC WORKFORCE CAPACITY

5.1 Workforce Capacity Analysis using SRS rule

The SRS rule is applied in call centers to determine the appropriate staffing levels for an offered load (R) and Quality of Service, (β) and is approximated as $R + \beta\sqrt{R}$, where, R is the amount of work that arrives in the system in unit time. The value of β signifies the operational regime of the system and can be a Quality Driven (QD) with emphasis on service quality over efficiency, Efficiency Driven (ED) with emphasis on efficiency over service quality or Quality Efficiency Driven (QED) regime with trade-off between quality and efficiency. Larger the β value, better the service level, (see Section 2.3), [14].

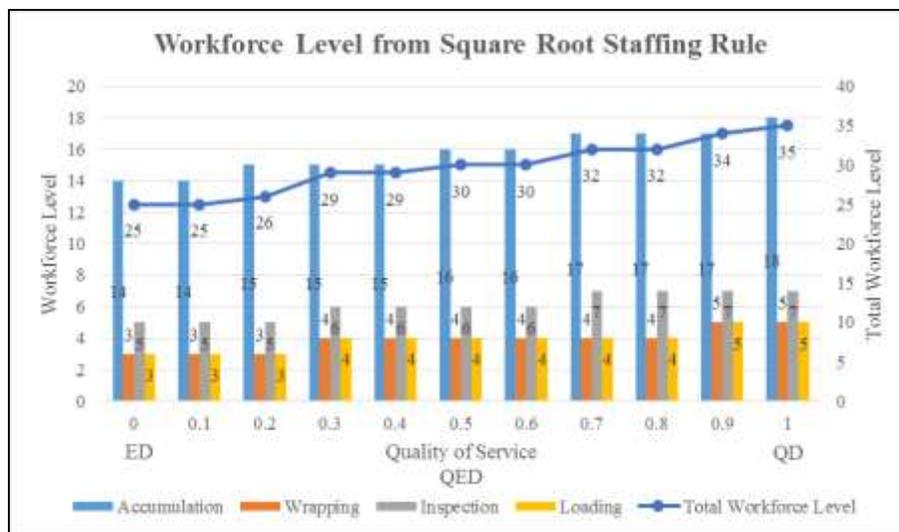


Figure 5.1: DC Workforce Level from SRS rule

The workforce capacity for the outbound operations of the DC is computed by the application of SRS rule for ED ($\beta = 0$), QD ($\beta = 1$) and QED ($\beta = 0.5$) operational regimes. It is observed that as quality of service increases, total workforce level increases as shown in Figure 5.1.

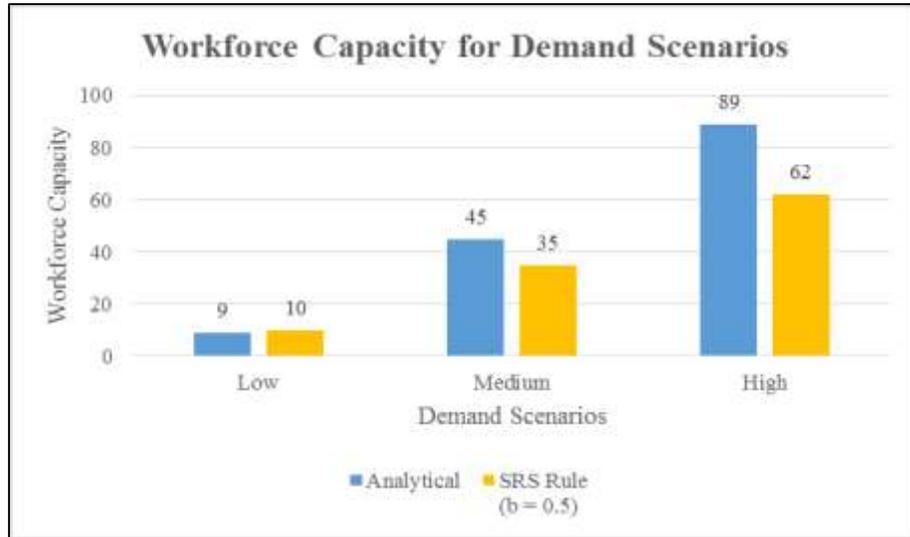


Figure 5.2: Workforce Capacity for Demand Scenarios

The workforce capacity for low, medium and high demand scenarios under analytical queueing model and SRS rule ($\beta = 0.5$) is determined. It is observed that workforce capacity determined from analytical queueing model and SRS rule are comparable only for a low demand scenario as shown in Figure 5.2.

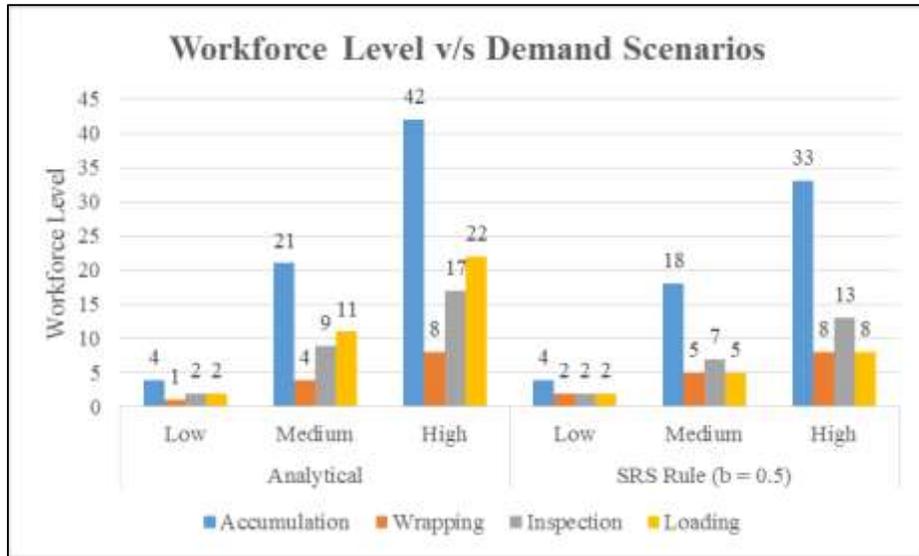


Figure 5.3: Workforce Level for Demand Scenarios

The workforce level at accumulation, wrapping, inspection and loading for low, medium and high demand scenarios under analytical queueing model and SRS rule ($\beta = 0.5$) is also determined. It is observed that workforce level for accumulation, wrapping, inspection and loading determined from analytical queueing model and SRS rule are comparable only for a low demand scenario as shown in Figure 5.3.

5.2 Chapter Summary and Conclusions

- DC workforce capacity is estimated by adapting SRS rule used for call center staffing. It is observed that as quality of service increases, total workforce level increases.

- The workforce capacity estimated using the analytical queueing model and SRS rule is comparable only for a low demand scenario.

- The workforce level for accumulation, wrapping, inspection and loading determined from analytical queueing model and SRS rule is comparable only for a low demand scenario.

6.1.1 Performance Analysis using Simulation Model

The performance of the DC is analyzed using the simulation model. The model is run for 500 hours with 30 replications after a warm-up period of 100 hours. The simulation model output is given in Table 6.1. The TPT and LHPT is determined as 683.39 minutes and 335.56 respectively. It is observed that the expected waiting time in queue and workforce utilization for wrapping and inspection process is high (32.30 & 28.64 minutes and 78.75% & 86.00% respectively), (see Appendix A.1).

Table 6.1: Simulation Model Output

OUTPUT					
Process	Process Time		WIP Levels		Workforce Utilization
	Waiting Time	Cycle Time	WIP Waiting	WIP	
Accumulation	0.05	294.99	0.002	12.67	50.76%
Wrapping	32.30	55.00	1.40	2.36	78.75%
Inspection	28.64	120.19	1.24	5.16	86.00%
Loading	0.26	152.28	0.01	6.54	65.36%
TPT		683.39	ATL	29.2	
LHPT		335.56			

The results from simulation model is compared with that of the analytical queueing model. The analytical and simulation model outputs for cycle time, WIP level and workforce utilization are comparable. The TPT and LHPT is determined with -3.13% and -6.49% error respectively from the simulation model with reference to the analytical model. It is observed that error levels for accumulation and loading are within acceptable limits though that of wrapping and inspection are outside limits as shown in Figure 6.2 - 6.4.

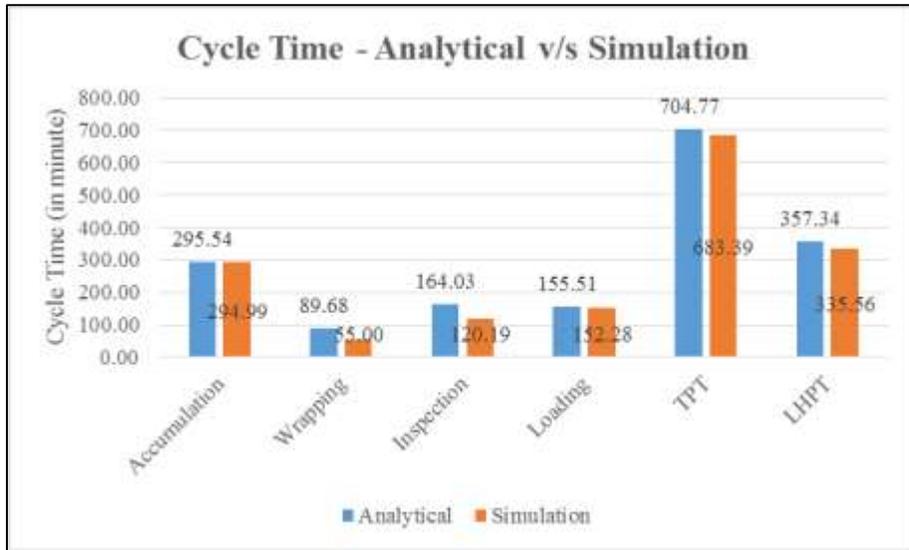


Figure 6.2: Cycle Time - Analytical v/s Simulation Output

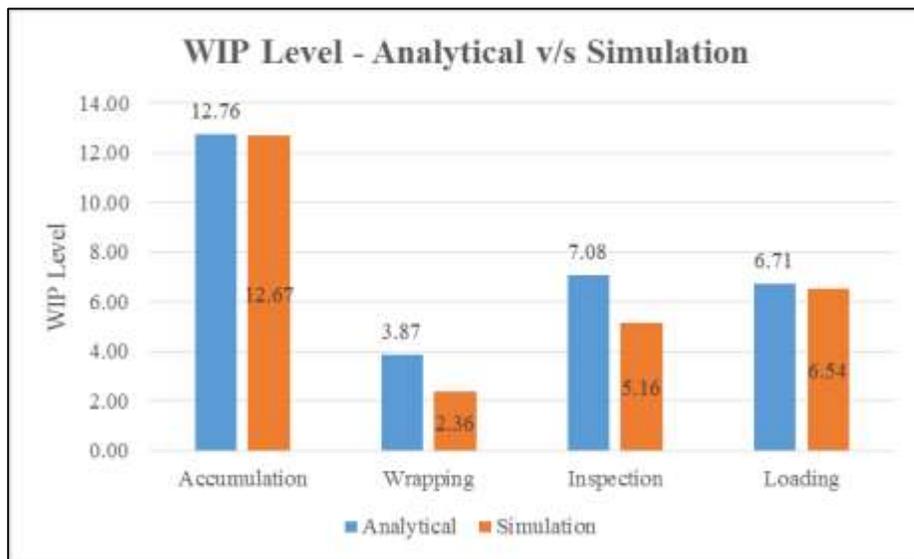


Figure 6.3: WIP Level - Analytical v/s Simulation Output

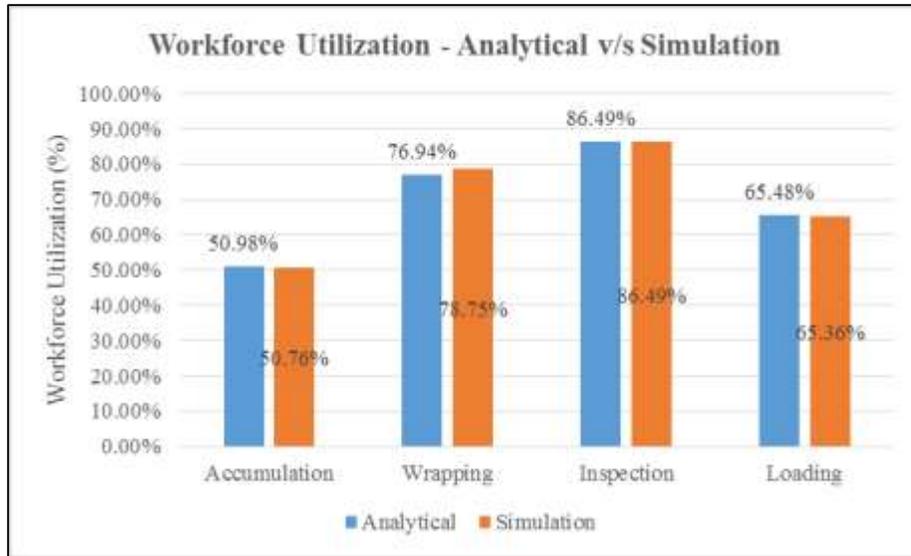


Figure 6.4: Workforce Utilization - Analytical v/s Simulation Output

6.2 Simulation-based Optimization Model

A simulation-based multi-objective optimization model is developed from the initial simulation model using OptQuest for Simio add-in to determine the optimal workforce capacity for a target workforce utilization policy by minimizing TPT and operating cost subject to constraints. The model is run for 500 hours after a warm-up period of 100 hours with 30 replications for 100 scenarios.

The optimal workforce capacity that minimizes TPT and operating cost is determined by multi-objective weighted optimization and pattern frontier optimization.

Minimize: TPT and Operating Cost

subject to:

$$1 \leq \text{Forklift} \leq 25$$

$$1 \leq \text{Wrapping Team} \leq 6$$

$$1 \leq \text{Inspection Team} \leq 8$$

$$1 \leq \text{Loading Team} \leq 10$$

$$u \geq 0.70$$

6.2.1 Multi-Objective Weighted Optimization

OptQuest optimizes across all responses considering each response's weight value to determine a single optimal solution. The optimal workforce capacity that minimizes TPT and operating cost is determined by multi-objective weighted optimization as given in Table 6.2 - 6.3 and shown in Figure 6.5, (see Appendix A.2), [15].

Table 6.2: Optimal Workforce Capacity

Optimal Workforce Capacity				
Accumulation	Wrapping	Inspection	Loading	Total Workforce Level
18	4	7	10	39

Table 6.3: Optimal TPT, LHPT & Operating Cost

TPT	LHPT	Operating Cost
641.04	294.44	\$ 5,502.30

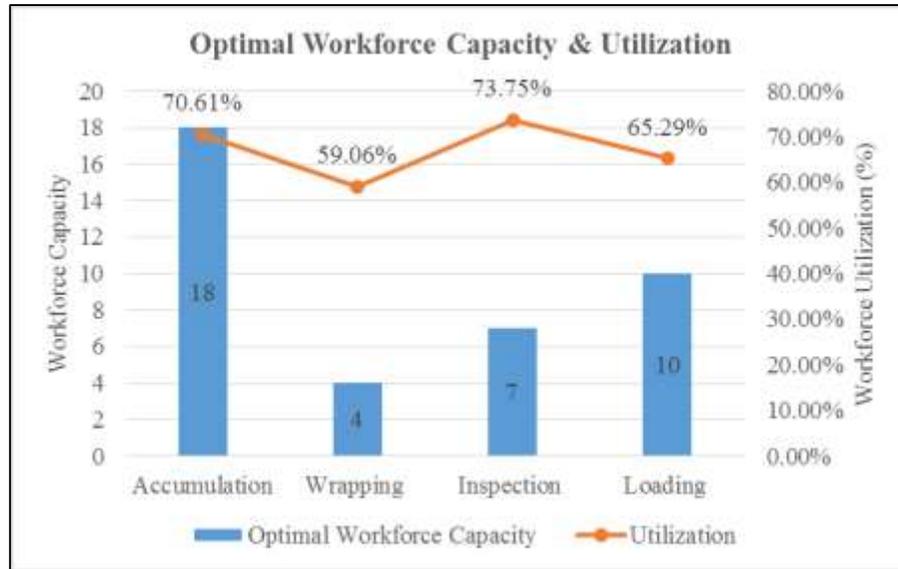


Figure 6.5: Optimal Workforce Capacity

6.2.2 Pattern Frontier Optimization

OptQuest optimizes across all responses and finds the set of scenarios that are optimal, rather than a single optimal solution based on weights. The optimal scenarios for workforce capacity that minimizes TPT and operating cost is determined by pattern frontier optimization. The optimal scenarios indicate that a 9% reduction in TPT increases the operating cost by 18% as shown in Figure 6.6 - 6.9, (see Appendix A.3), [15].

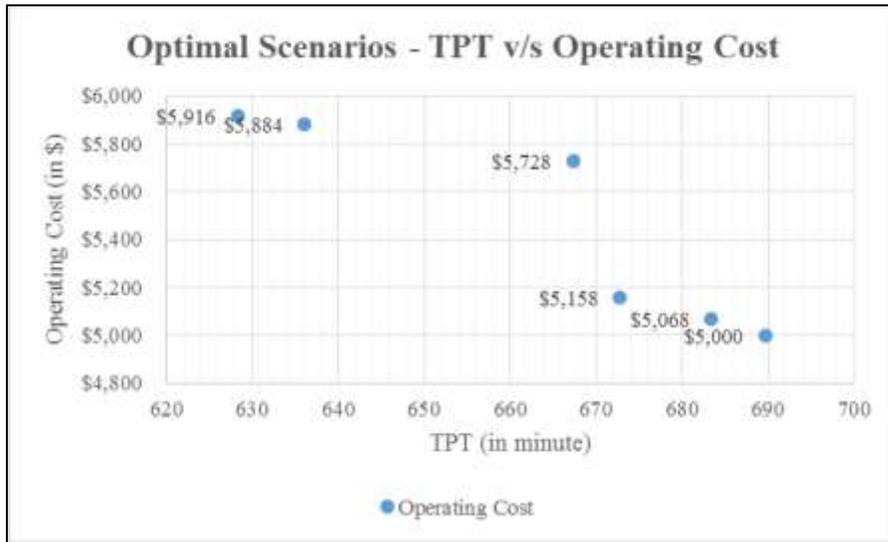


Figure 6.6: Optimal Scenarios - TPT v/s Operating Cost



Figure 6.7: Optimal Scenarios - Operating Cost v/s LHPT

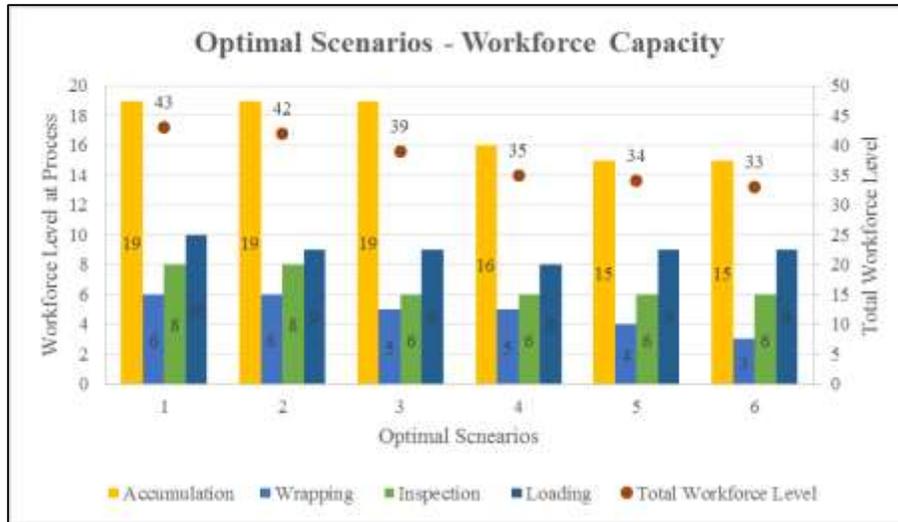


Figure 6.8: Optimal Scenarios - Workforce Capacity

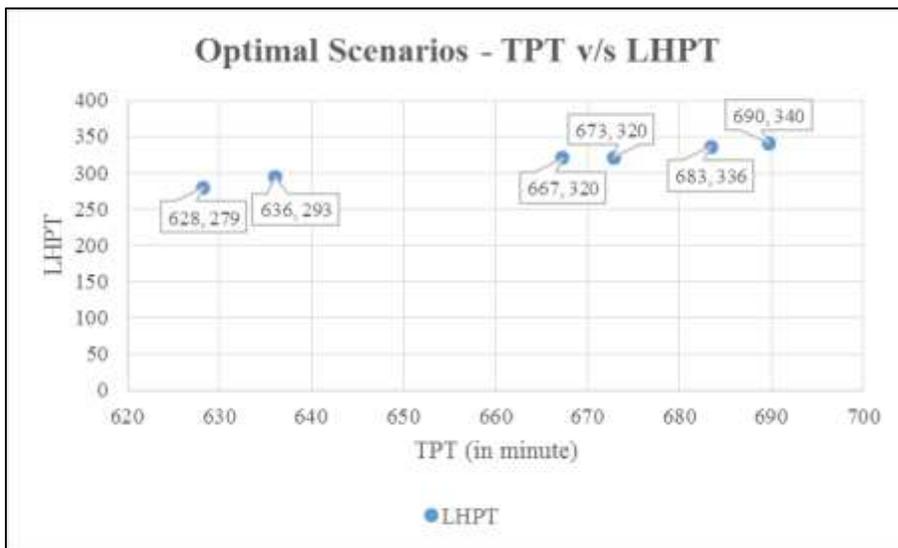


Figure 6.9: Optimal Scenarios - TPT v/s LHPT

It is observed that there are six optimal scenarios with TPT , LHPT and operating cost that range between (628.24 minutes, 340 labor hours, \$5915.95) and (689.66 minutes, 279 labor hours, \$5000.07) corresponding to a workforce capacity that range between 43 to 33.

6.3 Chapter Summary and Conclusions

- The analytical models are benchmarked and validated by a discrete event simulation model developed using Simio and its optimization tools. The TPT and LHPT is determined with -3.13% and -6.49% error respectively from the simulation model with reference to the analytical model.
- The optimal workforce capacity that minimizes TPT and operating cost is determined by multi-objective weighted optimization and pattern frontier optimization using the simulation model.
- It is observed that there are six optimal scenarios with TPT , LHPT and operating cost that range between (628.24 minutes, 340 labor hours, \$5915.95) and (689.66 minutes, 279 labor hours, \$5000.07) corresponding to a workforce capacity that range between 43 to 33.
- The optimal scenarios indicate that a 9% reduction in TPT increases the operating cost by 18%.

CHAPTER 7

SUMMARY AND CONCLUSIONS

A decision model that integrates critical operational performance measures is developed to evaluate DC performance and workforce capacity policies under varying demand scenarios. Industrial practitioners can use this model for tactical and operational decisions in DCs.

- The existing literature is heavily focused on warehouse design & control using travel time models and performance analysis in terms of throughput, resource utilization and storage strategy for order picking systems. As labor cost can be a significant component of variable cost in DCs, it is evident that there is a need for research to explore the impact of workforce capacity on DC performance by considering cost and service level trade-off for varying demand scenarios.
- A generic computational model, with length (L), width (W), aisle-width (A) and number of docks (N) as parameters, is developed to estimate forklift travel times in DCs, which is capable to evaluate DCs with a size between 60,000 sq.ft. (small scale) and 500,000 sq.ft. (large scale) and up to 45 docks. The travel times are modelled as random variables and the moments of travel time distribution is used as input to analytical queueing model and discrete event simulation model of DC.

- The analytical queueing model facilitate development and evaluation of workforce capacity policies for the DC for several demand scenarios by providing performance estimates, particularly in terms of TPT and LHPT, for each policy-scenario combination.
- DC workforce capacity is estimated by adapting SRS rule used for call center staffing. The workforce capacity estimated using the analytical queueing model and SRS rule are comparable only for a low demand scenario.
- The analytical models are benchmarked and validated by a discrete event simulation model developed using Simio and its optimization tools. The TPT and LHPT is determined with -3.13% and -6.49% error respectively from the simulation model with reference to the analytical model.
- The optimal workforce capacity that minimizes TPT and operating cost is determined by multi-objective weighted optimization and pattern frontier optimization using the simulation model. The optimal scenarios indicate that a 9% reduction in TPT increases the operating cost by 18%.

REFERENCES

- [1] Jinxiang Gu, Marc Goetschalckx, Leon F. McGinnis, 2007, “Research on warehouse operation: A comprehensive review”, *European Journal of Operational Research* 177, 1-21
- [2] B. Rouwenhorst a, B. Reuter, V. Stockrahm, G.J. van Houtum, R.J. Mantel, W.H.M. Zijm, 2000, “Warehouse design and control: Framework and literature review”, *European Journal of Operational Research* 122, 515-533.
- [3] R. Pandit, U.S. Palekar, 1993 “Response time considerations for optimal warehouse layout design”, *Journal of Engineering for Industry* 115, 322-328.
- [4] Chew, E.P., Tang, L.C., 1999, “Travel time analysis for general item location assignment in a rectangular warehouse”, *European Journal of Operational Research* 112, 582–597.
- [5] S.C. Graves, W.H. Hausman, L.B. Schwarz, 1977, “Storage retrieval interleaving in automatic warehousing systems”, *Management Science* 23 (9), 935-945.
- [6] Y.A. Bozer, J.A. White, “Design and performance models for end-of-aisle order picking systems”, *Management Science* 36 (7), 852-866.
- [7] de Koster, R., 1994, “Performance approximation of pick-to-belt order picking systems”, *European Journal of Operational Research* 72 (3), 558–573.
- [8] Lee, H.S., 1997, “Performance analysis for automated storage and retrieval systems”, *IIE Transactions* 29, 15–28.
- [9] Hur, S., Lee, Y.H., Lim, S.Y., Lee, M.H., 2004, “A performance estimating model for AS/RS by M/G/1 queueing system”, *Computers and Industrial Engineering* 46, 233–241.

- [10] Berry, J. R., 1968, "Elements of Warehouse Layout," The Int. J. of Prod. Res., Vol. 7, No. 2, pp. 105-121.
- [11] Medina, L., 2009, "A Simulation Approach for the Predictive Control of a Distribution Center", M.S. Thesis, The Pennsylvania State University, U.S.A.
- [12] Wallace J. Hoop, Mark L. Spearman, 2001, "Factory Physics: Foundations of Manufacturing Management", McGraw Hill Higher Education, Second Edition.
- [13] Technion Israel Institute of Technology, The William Davidson Faculty of Industrial Engineering and Management, <http://ie.technion.ac.il/serveng>
- [14] Avishai Mandelbaum, Sergey Zeltyn, 2007, "Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers", Advances in Services Innovations, 17 - 45.
- [15] Renee M. Thiesing, C. Dennis Pegden, 2013. "Recent Innovations in Simio" In Proceedings of the 2013 Winter Simulation Conference.

A.3 Pattern Frontier Optimization Simulation Output

Scenario	Name	Status	Replications	Completed	Partials/Capacity	Warp/Capacity	Inspector/TeamCapacity	Loading/TeamCapacity	TTT (Min. + OperatingCost (\$M))	LHFT	FCap/Nb	ITCap/Nb	TCap/Nb	LTCap/Nb	
003	Completed	20	10 of 20	25	6	8	8	20	875.100	8046.39	176.388	30.229	38.838	82.724	84.278
020	Completed	20	10 of 20	26	6	8	8	20	828.242	8915.95	176.312	30.829	38.734	82.615	84.088
027	Completed	20	10 of 20	23	4	8	8	20	838.98	8362.88	186.682	31.172	38.281	84.422	85.343
044	Completed	20	10 of 20	28	6	8	8	8	836.885	8882.78	203.268	31.728	38.368	85.225	85.225
026	Completed	20	10 of 20	24	8	8	8	8	881.28	8302.58	207.262	31.828	46.422	82.847	71.288
025	Completed	20	10 of 20	24	8	8	8	20	862.762	8746.23	117.128	32.724	38.308	86.127	88.087
018	Completed	20	10 of 20	22	5	8	8	8	864.424	8388.87	205.024	34.222	42.888	84.277	78.814
040	Completed	20	10 of 20	26	5	8	8	8	867.286	8727.42	118.747	36.882	46.888	85.724	72.081
081	Completed	20	10 of 20	21	5	8	8	8	867.886	8883.1	118.328	36.878	47.278	88.874	72.187
061	Completed	20	10 of 20	23	5	8	8	8	888.228	8727.87	118.862	38.872	47.181	88.822	72.188
018	Completed	20	10 of 20	26	5	8	8	8	877.278	8282.87	122.888	38.328	46.888	84.788	80.228
046	Completed	20	10 of 20	28	8	8	8	8	878.212	8622.88	126.6	38.818	47.822	85.822	72.407
032	Completed	20	10 of 20	27	3	8	8	8	878.846	8228.38	125.228	32.862	38.008	88.816	71.886
017	Completed	20	10 of 20	27	5	8	8	8	878.144	8322.4	122.748	31.888	46.878	85.188	78.822
074	Completed	20	10 of 20	26	4	8	8	8	878.214	8822.23	121.228	21.284	38.288	86.788	72.084
080	Completed	20	10 of 20	23	4	8	8	8	878.888	8788.31	128.328	38.822	38.248	88.822	72.887
042	Completed	20	10 of 20	28	5	8	8	8	878.626	8222.88	138.862	38.388	47.882	88.248	81.424
065	Completed	20	10 of 20	28	5	8	8	8	878.622	8888	128.872	38.348	46.878	82.787	81.887
097	Completed	20	10 of 20	20	4	8	8	8	878.886	8882.11	128.842	38.178	38.888	88.148	72.288
018	Completed	20	10 of 20	21	5	8	8	8	877.72	8842.88	121.888	40.824	47.871	88.48	81.788
088	Completed	20	10 of 20	17	4	8	8	8	878.881	8282.88	121.228	32.724	38.248	88.282	81.888
085	Completed	20	10 of 20	17	4	8	8	20	878.287	8482.71	121.071	24.822	38.288	88.882	85.228
087	Completed	20	10 of 20	26	4	8	8	8	878.812	8282.78	122.288	38.878	38.281	88.228	72.277
078	Completed	20	10 of 20	28	4	8	8	8	881.284	8482.84	124.128	38.128	38.888	88.288	81.881
084	Completed	20	10 of 20	28	4	8	8	8	881.677	8782.48	124.882	37.888	38.478	88.488	72.821
079	Completed	20	10 of 20	28	2	8	8	8	882.844	8884.18	122.87	38.388	38.478	88.487	72.228
091	Completed	20	10 of 20	25	4	8	8	8	882.287	8884.48	126.128	38.878	38.228	88.722	71.887
077	Completed	20	10 of 20	17	4	8	8	8	888.828	8222.81	148.842	28.287	38.488	88.888	81.878

Scenario	Name	Status	Replications	Completed	Partials/Capacity	Warp/Capacity	Inspector/TeamCapacity	Loading/TeamCapacity	TTT (Min. + OperatingCost (\$M))	LHFT	FCap/Nb	ITCap/Nb	TCap/Nb	LTCap/Nb
062	Idle	20	10 of 20	24	3	8	8	8	888.862	8886.8	122.822	28.847	38.228	72.888
047	Idle	20	10 of 20	25	3	8	8	8	888.804	8782.13	121.888	28.124	38.721	72.864
030	Idle	10	10 of 10	15	3	8	8	8	888.882	8888.87	84.882	28.812	38.822	71.888
048	Idle	20	10 of 20	15	8	8	8	8	888.98	8124.13	84.842	47.822	38.531	81.878
040	Idle	18	10 of 18	28	2	8	8	8	881.828	8486.46	88.488	28.278	37.182	72.222
028	Idle	18	10 of 18	24	5	8	8	8	715.117	8122.51	88.888	47.822	38.888	88.878
024	Idle	18	10 of 18	24	3	8	8	8	717.688	8122.36	88.878	28.124	38.884	71.887
018	Idle	20	10 of 20	17	4	8	8	7	1022.81	10412.7	74.888	18.122	38.272	88.284
094	Idle	20	10 of 20	13	4	8	8	8	1425.87	8881.23	97.124	88.874	38.888	78.248
034	Idle	20	10 of 20	12	3	8	8	8	1888	1842.15	88.788	24.288	38.228	88.224
098	Idle	20	10 of 20	12	3	8	8	8	1782.13	1888.8	88.888	24.271	38.128	72.128
022	Idle	20	10 of 20	18	5	8	8	8	2078.46	1872.3	78.444	47.824	38.428	88.881
001	Idle	20	10 of 20	13	4	8	8	8	2125.42	1188.4	88.881	58.882	38.287	88.288
023	Idle	20	10 of 20	17	4	8	8	8	2221.71	1822.2	74.788	88.818	38.881	88.888
095	Idle	20	10 of 20	11	3	8	8	8	2847.25	1782.1	88	88.478	24.788	78.228
033	Idle	24	24 of 24	13	2	8	8	8	3288.1	2212.1	88.188	88.878	28.888	81.281
020	Idle	20	10 of 20	16	2	8	8	8	3447.28	2842.5	78.818	88.888	32.228	81.881
096	Idle	20	10 of 20	11	2	8	8	8	3484.21	2324.5	88	88.882	28.784	81.218
032	Idle	16	16 of 16	13	3	8	8	8	4438.17	2188.6	88	88.814	88	84.248
089	Idle	20	10 of 20	23	3	8	8	8	4827.28	2882.2	55.222	28.288	88	88.222
037	Idle	29	29 of 29	22	2	8	8	8	4874.82	2724.5	57.188	88	88	85.218
090	Idle	20	10 of 20	25	2	8	8	8	4728.82	4848.2	11.882	88	32.228	88
077	Idle	12	12 of 12	8	4	8	8	8	7025.28	2848.9	88	37.222	81.882	51.884
061	Idle	15	15 of 15	20	4	8	8	8	7841.34	8187.6	41.728	38.878	43.787	88
097	Idle	14	14 of 14	8	4	8	8	8	7781.88	3888.9	88	37.422	88	47.48
030	Idle	18	18 of 18	6	2	8	8	8	8884.88	2122.8	88	55.822	81.814	78.124