ANALYTICAL APPROACH TO ESTIMATING AMHS PERFORMANCE IN 300MM FABS

A Dissertation
Presented to
The Academic Faculty

by

Dima Nazzal

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2006

ANALYTICAL APPROACH TO ESTIMATING AMHS PERFORMANCE IN 300MM FABS

Approved by:

Dr. Leon F. McGinnis, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Robert D. Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ron Billings
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Gunter P. Sharp
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Christiaan Paredis
School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: June 20, 2006

To Zaid,

my source of strength,

my love, my friend.

ACKNOWLEDGEMENTS

First I would like to express my deepest appreciation to my advisor, Leon McGinnis, without whom this thesis would not exist. Dr. McGinnis gave me his continuous guidance, support and encouragement. Thank you for helping me see the larger picture when I got lost in the details.

I want to thank my committee members, Dr. Ron Billings, Dr. Robert Foley, Dr. Chris Paredis, and Dr. Gunter Sharp for their valuable insights and direction in my thesis. I am also grateful to Dr. Doug Bodner for his guidance through the first years.

I thank my entire family, my mom, my brother, and especially my in-laws for being the most wonderful, supportive, and loving family anyone could ask for. Special thanks to my friends, overseas and in the United States. Many thanks to my friends at school Maga Khachatryan, Andy Johnson, and Lori Houghtalen.

Most of all, I could never thank my husband, Zaid Duwayri, enough for all of the patience, unconditional love and support he has shown me throughout this process. There is no way I could have done this without him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis proposes a computationally effective analytical approach to automated material handling system (AMHS) performance modeling for a simple closed loop AMHS, such as is typical in supporting a 300mm wafer fab bay. In this system, due to the significant impact of vehicle blocking, a straightforward queueing network model which treats the material handling system as a central server can be very inaccurate. On the other hand, discrete-event simulation can produce accurate assessments of the production performance, including the contribution by the automated material handling systems (AMHS). However, the corresponding simulation models are both expensive and time-consuming to construct, and require long execution times to produce statistically valid estimates. These attributes render simulation ineffective as a decision support tool in the early phase of system design, where requirements and configurations are likely to change often. We propose an alternative model that estimates the MHS performance considering the possibility of vehicle-blocking. Such models are useful in the design of vehicle-based AMHS and correctly estimate the throughput capacity and move request delay of the AMHS.

A probabilistic model is developed, based on a detailed description of AMHS operations, and the system is analyzed as an extended Markov chain. The model tracks the operations of all the vehicles on the closed-loop considering the possibility of vehicle-blocking. Steady-state analysis provides estimates of empty-vehicle flows, which are essential to accurately estimate other metrics such as the transport time and throughput capacity. The resulting large-scale model provided reasonably accurate estimates; however, it presented some computational challenges.

These computational challenges motivated the development of a second model that also analyzes the system as an extended Markov chain but with a much reduced state space because the model tracks the movement of a single vehicle in the system with additional assumptions on vehicle-blocking. This reduced-state model offers computationally fast, fairly accurate estimates of the AMHS throughput capacity.

Neither model is a conventional Markov Chain model because they combine the conventional Markov Chain analysis of the AMHS operations with additional constraints on AMHS stability and vehicle-blocking that are necessary to provide a unique solution to the steady-state behavior of the AMHS.

Based on the throughput capacity model, an analytical approach is developed to approximate the expected response time of the AMHS to move requests. The expected response times are important to measure the performance of the AMHS and for estimating the required queue capacity at each pick-up station. The derivation is not straightforward and especially complicated for multi-vehicle systems. The approximation relies on the assumption that the response time is a function of the distribution of the vehicles along the tracks and the expected length of the path from every possible location to the move request location.

The proposed analytical approach is novel because it models mutli-vehicle material handling systems considering practical issues that have not been previously addressed. Moreover, the semiconductor industry can benefit from such models because it proposes and demonstrates the capability of computationally fast and reusable analytic models that provide accurate and reliable estimates of AMHS performance necessary for the design stage.

CHAPTER 1

INTRODUCTION

Semiconductor technology is the building block of information technology, which touches most aspects of contemporary life. It is central to such areas as computers, communications systems, and consumer products. Moreover, the semiconductor industry has become a vital contributor to the world economy, with $227.5 billion in sales worldwide in 2005, as stated by the Semiconductor Industry Association (SIA[1]) press room (SIA report, 2005).

Within the past decade, the semiconductor industry has transitioned from 200mm to 300mm wafer fabrication. This transition is expected to produce 2.5 times more chips per wafer (Bonora and Feindel, 2001), at a cost 1.4 times more than the 200mm wafer (Jones, 2003).

The transition to automated material handling in the 300mm wafer era is required to maximize the productivity of the capital, to satisfy ergonomic restrictions caused by the weight and volume of 300mm wafer lot carriers, and to reduce the particle contamination and vibrational shocks on the wafers (Nadoli and Pillai, 1994). Furthermore, AMHS is at the center of full factory automation, which depends

[1] The SIA was created in 1977 by five innovators in the industry of microelectronics. Currently, the SIA has 15 board member companies. They are Advanced Micro Devices, Agere Systems, Altera, Analog Devices, Conexant Systems, Cypress Semiconductor, IBM, Intel, Intersil, LSI Logic, Micron Technology, Motorola, National Semiconductor, Texas Instruments and Xilinx. http://www.sia-online.org.

fundamentally on automated material handling systems (AMHS) as the medium through which factory operations control is implemented.

Most fabs use a bay layout (Cardarelli and Pelagagge, 1995), and each bay contains a group of process tools. During the manufacturing process, wafers make many trips between the bays, creating a large amount of "interbay" traffic. Wafers are manufactured in groups of twenty-five, called "lots" that travel together. Each lot has a route, which is a list of steps that must be performed on the wafers before completion of the final product. As a lot flows through the fab, it is processed by the process tools listed in its route.

An AMHS in a 300mm fab typically consists of an overhead track with vehicles, stockers, and load/unload ports for interactions with the process tools and storage units (stockers). A stocker is located in each bay to store work-in-process (WIP), when the process tool required for a lot is unavailable when the lot arrives to the bay.

In our research, we explore the use of queuing network models to approximate the performance of a discrete vehicle-based material handling system. These models for design and control of AMHS are scarce, and the semiconductor industry would benefit from the development and use of analytic modeling tools that have not been previously explored.

1.1. Problem domain

The process of semiconductor manufacturing typically consists of several stages, during which thousands of integrated circuits (ICs) are formed on a single wafer. The first stage produces a wafer, which is a thin round slice of a semiconductor material, usually silicon. Integrated circuits are formed on wafers in the wafer fabrication stage, which is performed inside the fab (cleanroom). The fabrication process involves a series of standard steps: diffusion, ion implantation, oxidation, depositions of metals and insulators, etching, and most importantly photolithography. These steps are repeated several times until the last layer is completed. The final stages of probing and assembly are then completed outside the cleanroom where the wafers are tested and sliced into single chips (dies). This research focuses on the stage of wafer production inside the cleanroom, which is the most complicated stage in terms of scheduling and routing, creating a rich opportunity for improvement.

1.2. Motivation

Constructing a 300mm fab is projected to cost $2-3 billion, (Jones, 2003), and the AMHS represents a significant share of the investment cost, as much as 3 to 5% of the total fab cost (Arzt and Bulcke, 1999). Even though the cost of AMHS is small relative to the total capital investment, AMHS performance is critical to achieving the planned return on investment (ROI) for the total capital investment. In other words, it is not AMHS cost, it is the AMHS impact on performance that is crucial. AMHS may introduce lot delays or cause tool idle time by failing to move lots in the planned and allotted time. In fact, according to the International Technology Roadmap for Semiconductors[2] (ITRS report, 2005), the key focus areas and issues for 2005 and

[2] The International Technology Roadmap for Semiconductors (ITRS) periodically lays out a technology plan to guide the semiconductor industry in the coming decades. The latest study is summarized in the ITRS 2005 annual report. It provides the current estimates for research and development that is required over the next decade to meet the historical numbers in performance growth, size reductions, cost, etc. http://public.itrs.net.

beyond is to increase throughput for AMHS, reduce average delivery times, and improve its reliability.

Fundamentally, the role of the AMHS is to serve the production system, and the two systems interact at the lot transfer points in the facility. Essentially, the AMHS should be able to perform the move requests generated by the production system. Once this basic level of service is achieved, AMHSs are distinguished from each other through their cost and performance. Tradeoffs are usually between these two measures. AMHS Performance usually is measured in terms of its throughput (number of moves per unit time) and response time; a MHS that responds faster to the move requests is generally more preferred. We say "generally" because it might be the case that a faster delivery will not affect the production cycle time, if the destination tool is not ready to start processing upon lot arrival. Material handling operations are non-value adding, so the goal is to minimize their cost while satisfying their performance requirements. The AMHS should be transparent, i.e. it should not constitute a bottleneck to the production system (McGinnis et al., 1992).

Estimating AMHS performance in IC fabs is difficult, because of the complexity of the systems. The International Technology Roadmap for Semiconductors (ITRS report, 2001) characterizes the AMHS as having several vehicles, operating on a network with loops, intersections, spurs, and short-cuts, serving many different pick-up/deposit stations. The movement requirements appear to be random, and although they exhibit some temporal correlations, these correlations are not strong enough to permit precise scheduling of the AMHS resources.

AMHS design is important because a reduction in material handling cycle times can reduce inventory cost through reduced Work in Process (WIP) and increase revenue through better on-time delivery performance and increased market share. Poor AMHS design and/or inefficient operations can drastically affect these key performance indicators. There is, therefore, a need to understand the role and evaluate the impact of AMHS on the production system performance.

Contemporary simulation technology can produce accurate assessments of fab production performance, including the contribution by the AMHS. However, the corresponding simulation models are both expensive and time-consuming to construct, and require long execution times to produce statistically valid estimates. These attributes render simulation ineffective as a decision support tool in the early phase of system design, where system configurations are likely to change often.

1.3. AMHS description

A typical 300mm AMHS has a spine layout, illustrated in Figure 1-1, with a central material handling spine and loops branching on both sides to serve production equipment. There are two distinct operating scenarios: (1) the spine and the loops are decoupled, and vehicles are dedicated to the spine or to one of the loops; or (2) the spine and the loops are integrated, and vehicles may move freely between them (Pillai et al. 1999). In this research, we consider only the first case, which means a wafer lot moving from a tool on one loop to a tool on a different loop must travel through the main spine, and will use three different vehicles, one in each loop and one in the spine. Automated storage units, referred to as stockers, are used to provide both temporary buffering for work-in-process and transfer between the bay and spine transport systems.

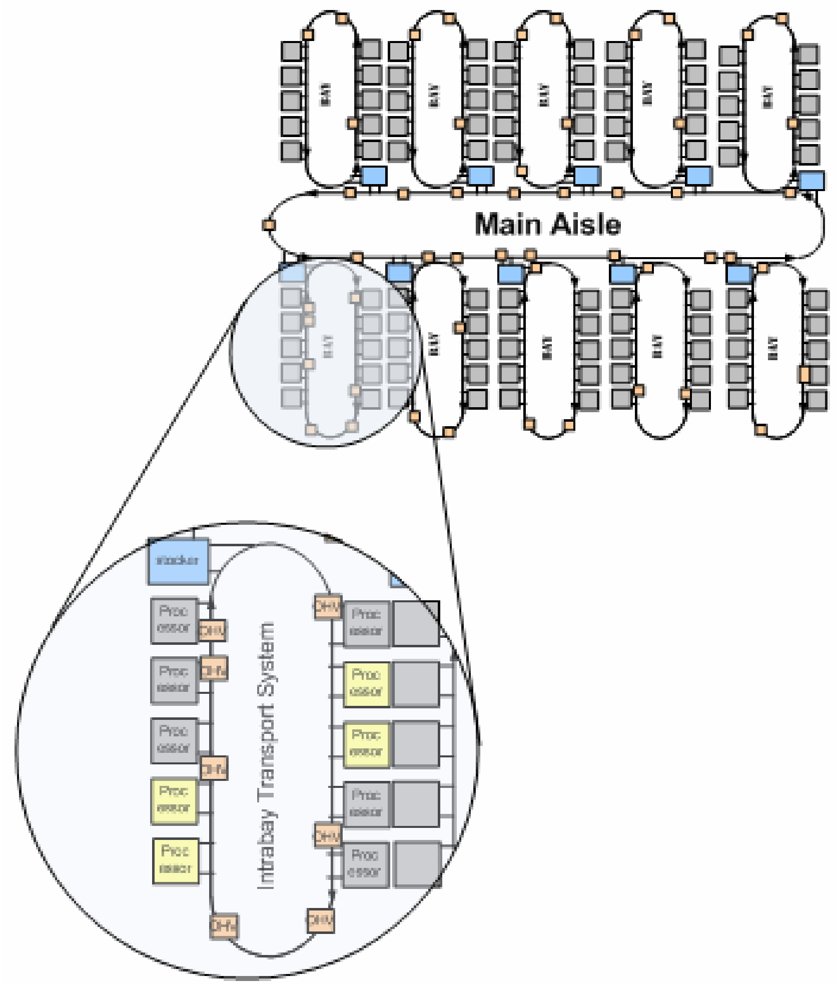


Figure 1.1 Closed-loop unidirectional interbay and intrabay AMHSs

Because of the space restrictions in the 300mm wafer fab bays, vehicle travel is on a unidirectional closed loop without the ability for vehicles to pass each other, even when a vehicle stops to drop-off/pick-up a lot at the input/output port of a process tool or a stocker (SEMATECH report, 1997). Thus, failure to carefully synchronize vehicle movements on a given loop can lead to significant amounts of vehicle blocking and the possibility of lot delay as well as induced vehicle idle time.

The closed loop overhead hoist transport (OHT) system serves the move requests originating from the stocker(s) and the processors (also referred to as production equipment or production tools) in the bay. Each machine —either stocker or tool has two load ports: an input port where loads are dropped off by the vehicle and an output port where loads are picked up by the vehicle to be delivered to their next destinations. Each port can accommodate one vehicle at a time. We use the term *station* to refer to the input and output ports of the machines. Thus, a loop serving M machines, denoted by m_i, consist of $s = 2m$ stations. Without loss of generality, we assume that the loops start at the stocker (m_1), and the vehicles' route is assumed to be m_1, m_2,.., m_i, m_{i+1},…, m_M, m_1, m_2,..

Machine m_i has two stations: the drop-off station s_i^d, and the pick-up station s_i^p. We model an OHT system configured as a simple loop, in which vehicles continuously travel the loop, when a vehicle approaches an m_i , it passes through the drop-off station s_i^d, then travels to the pick-up station s_i^p. Loads are served by vehicles based on the First-Encountered-First-Served (FEFS) rule. FEFS is a decentralized policy, first presented by Bartholdi and Platzman (1989). In FEFS, the vehicles are constantly circulating on the unidirectional loop. When an empty vehicle approaches an m_i, it inspects the output buffer, if there is a load (job) waiting at s_i^p, the vehicle picks it up, which requires time delay l for loading the job and then delivers it to its destination, say machine m_j, visiting machines m_{i+1}, m_{i+2},…, m_{j-1}, and finally the load's destination the drop-off station of m_j, denoted by s_j^d. The vehicle does not stop at machines m_{i+1}, m_{i+2},…, m_{j-1} unless it is blocked by other vehicles. If the output port s_i^p is empty, the vehicle travels to s_{i+1}^d, then inspects the output port s_{i+1}^p and so forth until it encounters a waiting load.

In order to estimate the throughput accurately, we need to estimate the blocking delays at each machine—either stocker or tool. The main objective of the models to be developed is to quantify the duration of this type of delay as a function of the layout of the transportation system, the demand rates, the speed of the vehicles and the number of vehicles circulating the loop.

1.4. Proposed analysis method

In this thesis, a queuing network type model is developed, based on a detailed description of AMHS operations, and the model is analyzed as an extended Markov chain. With this approach we are able to estimate both AMHS throughput and move request delays.

The analysis of AMHS is complex because:

1. The service rate of stations is state-dependent. A vehicle arriving at a station may stop for service to pick-up or drop-off loads, or may just pass through the station without stopping, or in the case of blocking, the vehicle might have to stop unnecessarily until the vehicle blocking its way moves on.
2. It is not possible to analyze each station independently. Stations have limited buffer capacity, and in many cases a station (including the track segment leading to it) cannot accommodate more than one vehicle at a time. In such systems, the number of vehicles queued at one station impacts the service rate of other stations.

We propose to model the vehicles' movements between stations as a discrete time Marko chain (DTMC). This keeps track of vehicles' loading/unloading and blocking; the discrete set of *states* simplifies the model by avoiding the continuous traveling process. We consider only those points in time when a vehicle is located at a station. Specifically, in the state space we do not consider the *state* of a vehicle traveling on a track segment in between stations.

We characterize a state by specifying the condition and location of vehicles in the network. The location of a vehicle specifies the station at which the vehicle is receiving service or arriving. The condition of the vehicle specifies whether it is loaded, empty, blocked while empty, blocked while loaded, or receiving service (picking-up or dropping-off a load). The system transition between the possible states is assumed to be Markovian, i.e., the next state of the system depends only on the current state, and not on the path taken to reach the current state.

The complexity of modeling a material handling system stems from its dependency on the move requests generated by the production equipment, but mostly from the nature of its operation where each service has two components: the response time to a move request that depends on the location of the vehicle, and the transport time that depends on the origin and destination of the move request. Therefore, the transitions between states in the Markov chain depends on the move requests generated by the machines and on the current state of the AMHS, and we need to match the AMHS state transitions with the generation of move requests by the machines. Some of the state transition probabilities are unknown, and hence the proposed *extended* Markov chain.

Figure 1-2 illustrates the modeling approach. The production and storage systems consist of physical elements: production tools, stockers, and products (lots), and of informational elements: products routes, and release rates. The material handling system consists of physical elements: vehicles, tracks, etc., and of informational elements: dispatching policies, and vehicle velocity. The interaction point between these two systems is the machines' loadports: in-port (drop-off station), and out-port (pick-up location).

In our approach, the production and storage systems description provides the from-to matrix for lot transport and the AMHS description provides the from-to travel time matrix and fleet size to the extended Markov chain queuing model. Based on the queuing model, we analyze the steady-state behavior of the vehicles by estimating the percentage of time vehicles spend at each station in each of their possible conditions. This analysis answers the following questions:

- Will the AMHS, on average, be able to handle the move requests imposed by the production and storage systems?
- If the system is feasible, what is the throughput capacity of the AMHS?
- How much of this capacity is lost to vehicle blocking?
- What is the utilization of the out-port station?
- This last question helps in designing the storage area at the machine.

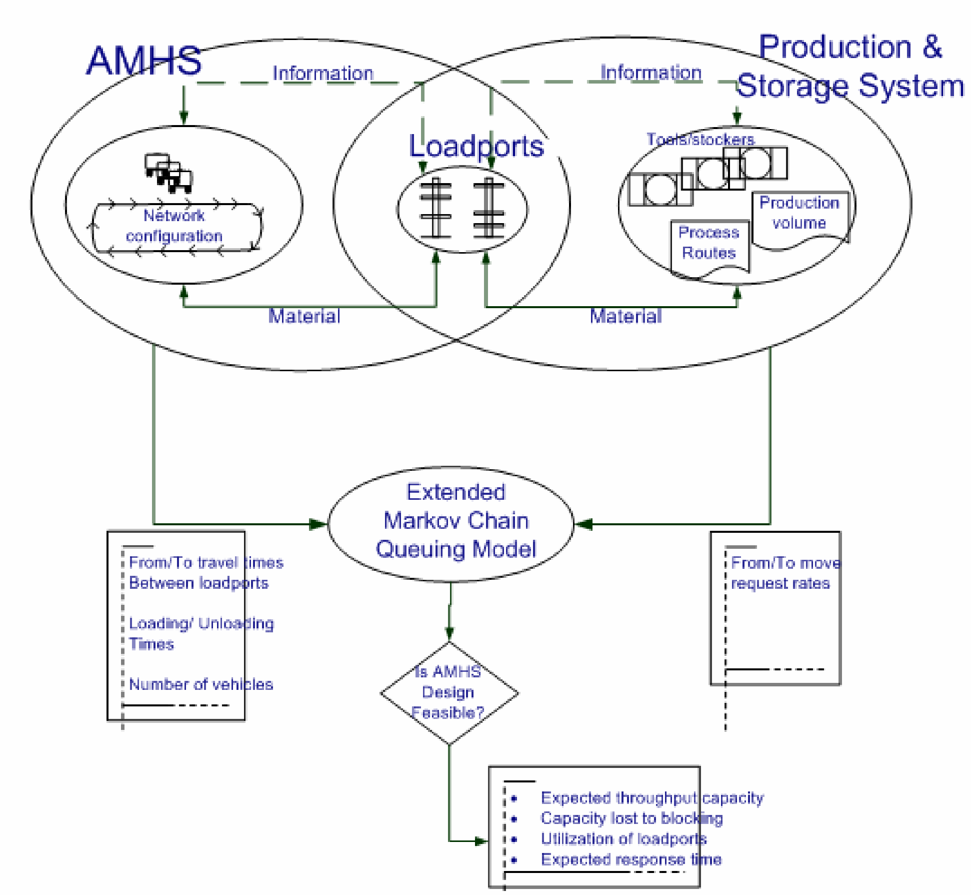


Figure 1.2 Modeling Framework

Two extended Markov chain models are developed in this thesis. The first model tracks the location and condition of the vehicles, simultaneously. Thus, a state will characterize the location and condition of *every* vehicle in the AMHS. This model is presented in Chapter three; the transition between the states is discussed along with the necessary conditions for coordinating the load drop-offs and pick-ups by the AMHS with the generation of the move requests by machines. Because the model tracks each vehicle, a modeling "trick" is employed to create a discrete-time Markov Chain (DTMC) representation of the state space and transitions. This model is quite accurate but the number of states grows exponentially with the problem size.

The second model discussed in Chapter four is an approximation of the first model. For the second model, we propose a different approach that creates one Markov chain for each vehicle separately, which reduces the size of the Markov chain drastically, but creates a challenge for modeling vehicle blocking. In addition to the issues we discuss in the first model, we demonstrate how to approximate vehicle-blocking probabilities for a single vehicle that operates on a multi-vehicle loop. The second model has two key advantages relative to the first: it has a much more compact state space, thus is more computationally tractable; and since it is a single-vehicle model, it does not require any modeling tricks to develop the DTMC model. However, there is a trade-off in terms of approximation error.

1.5. Thesis objective

The objective of this thesis is to develop computationally effective analytical models, useful in the design of vehicle-based AMHS to support semiconductor manufacturing and correctly estimate the throughput capacity and move request delay of the AMHS.

The developed analytical model answers the following questions:

- Given the AMHS design and production requirements, is the AMHS feasible?
- What is the number of vehicles that will provide the highest throughput capacity of the AMHS?
- What is the impact of changing the sequence of machines in the loop on the AMHS performance?

- What is the relationship between release rates and AMHS throughput capacity? In other words, what is the operating characteristics curve for the system?

This thesis makes a contribution to the research literature by proposing a novel approach to model multi-vehicle material handling systems that considers practical issues that have not been considered concurrently in the literature. First, we consider the state-dependent service rate of move request, whereas, in most analytical models of such systems, the material handling system is modeled by defining a "virtual" workstation between the processing tools in a product's route. The conventional approach assumes that the response time of the AMHS to a move request does not depend on the location of the load, nor on the vehicle distribution across the network. Second, we consider vehicle blocking and the resulting blocking delays in order to get good approximations of both the actual throughput of the AMHS and the average response time to move requests; an issue that is almost always ignored in the available analytical models.

The research is valuable to the semiconductor industry because it proposes and demonstrates the capability of computationally fast and reusable analytic models to provide accurate and reliable estimates of fab-level AMHS performance. These models are especially valuable for evaluating preliminary solutions to the AMHS design problem, where using simulation models is not a practical approach, because they take too long to develop, and require multiple lengthy executions to produce statistically valid estimates.

The organization of the thesis is as follows: Chapter two reviews literature on analytic and simulation-based models of the AMHS in semiconductor manufacturing, and previous research on queuing model of vehicle-based material handling systems. Chapters three and four present, respectively, the large-scale multi-vehicle extended Markov chain and the reduced-state extended Markov chain models. Both models are developed, discussed, and validated using discrete-event simulation. Chapter five presents an analytical approach to approximating the expected response time by the AMHS to move requests, the model is then validated using simulation. Chapter six evaluates the extended Markov chain model for throughput capacity estimation and the expected response time approximation using a detailed simulation model of international SEMATECH generic hypothetical fab.

CHAPTER 2

LITERATURE

2.1. Models of 300mm material handling

Throughout the literature, the importance of AMHS in 300mm wafer fabs has been repeatedly addressed, and research in this area can be broadly categorized into: (1) design optimization, targeting the guide path network layout design and calculating a feasible fleet size, and (2) performance evaluation of various AMHS methods or different AMHS configurations via simulation modeling.

2.1.1. Design optimization

Peters and Yang (1997) propose a network flow formulation to determine the number and location of shortcuts for the interbay transport system in a spine layout fab. The objective function minimizes the tradeoff between the increase in shortcut construction cost and the decrease in material handling costs. The authors also propose a space filling curve procedure to first determine the layout of the departments in the fab. The two procedures for determining the layout arrangement and material handling system design are embedded into an iterative steepest descent pair-wise interchange to solve the overall integrated problem.

Ting and Tanchoco (2000) propose an analytical procedure to construct a unidirectional circular layout for the interbay system in 300mm fabs under the assumption that not all the stockers can be connected by a simple loop. They describe the circular layout to consist of a central loop, shortcuts and loop additions; the final guide path connects all the stockers in the system. They propose a two-stage approach, in the first stage candidate layouts for the main loop are constructed using a heuristic that takes into account the construction costs and flow requirements. In the second stage, loop and shortcuts are added to connect the remaining stockers and eliminate flows on high-flow segments. A dynamic programming procedure is proposed to minimize the construction and operating cost in the second stage. In a later publication, Ting and Tanchoco (2001) use an analytical approach to develop a single and double rectilinear spine layout for overhead track layout design that connects the tools loadport to stocker loadports in an open ballroom layout bay for a 300mm fab. The objective function is to minimize the total loaded travel distances.

Steele (2002) proposes an algorithm to roughly estimate the performance of an automated material handling system during the design process. Each AMHS design is modeled as a network of nodes where a node may provide the transfer capability from/to wafer lot buffers, may enable vehicles to move to another branch of track, or may enable vehicles to recharge their batteries while waiting for a new move task. The algorithm computes the required loaded traffic flow, and the unloaded traffic flow. Using the total traffic flow, the AMHS is modeled as a network of queues. Finally, the algorithm can estimate the minimum number of vehicles required to deliver the required number of wafer lots and the average delivery times between each pair of source and destination nodes. The author applied the algorithm to a small-scale interbay material handling problem and compared the results of the algorithm to the results of a discrete event simulation. However, the algorithm assumes infinite capacity queues and thus does not consider blocking, as a result it is not sufficiently accurate to predict AMHS performance.

2.1.2. Simulation-based performance evaluation

To date, discrete event simulation has been the only methodology shown to give reliable estimates of fab-level AMHS performance. At the system design stage, however, large scale, high-fidelity simulation models are not a practical approach, because they take too long to develop, and require multiple lengthy executions to produce statistically valid estimates. Mackulak and Savory (2001) describe a study in which the AMHS experiments took over 250 hours of simulation time. Different approaches have been taken to overcome this problem. For example, Mackulak *et al.* (1998) propose developing a generic model that can be reconfigured according to the specific problem at hand, thereby reducing the model building time. Gaxiola and Mackulak (1999) describe the use of simple deterministic calculations in situations where the process requirements have not yet stabilized.

Pillai et al. (1999) discuss the issue of linking the interbay and intrabay tracks for a 300-mm fab layout. Rust et al. (2002) and Mackulak and Savory (2001) investigate the same problem by focusing on the impact of this decision on several AMHS performance measures.

Lin et al. (2003) propose using four different vehicle types to carry out the transport tasks from tool to tool. Type A vehicles move in an intrabay system and deliver the lots within the bay. Type B vehicles carry lots between the stockers. Type C vehicles carry lots from a tool in any bay to a stocker in the lot's destination bay. Type D vehicles move lots from a tool in any bay to a tool in any other bay. Three different transport methods using combinations of the four vehicle types were examined.

The conclusions in most of the simulation studies depend on the specifications of the fab being modeled, and thus do not constitute generic design guidelines. At the system design stage, therefore, large scale, high-fidelity simulation models are not a practical approach and system designers are limited in the range of alternatives they can expect to evaluate in detail.

2.2. Analytical models of material handling systems

In the literature, analytical models of AMHS are usually based either on deterministic optimization models or queuing models. The former fails to capture queuing in the system which is essential to accurately estimate the key performance measures. Often, in analytical factory modeling, the material handling system is modeled by defining a “virtual” workstation between the processing tools in a product’s routing. The delay associated with material handling is approximated by the processing time on this virtual workstation, which has a capacity equal to the number of vehicles available. This approach is appealing because it exploits well-understood queuing models. However, it has some inherent weaknesses. First, it assumes that the response time of the AMHS to a move request does not depend on the location of the load, nor on the vehicle distribution across the network. Second, it fails to capture the impact of vehicle-to-vehicle blocking, which, by consuming some of the available vehicle time, will degrade the capacity of the AMHS. Vis (2004) provides a survey of work in this area.

The objective of many of the analytic models of multi-vehicle systems is to estimate the minimum number of unit load capacity vehicles required to satisfy a given level of move requests. The fleet size is determined by the travel requirements for the system, which includes loaded and empty vehicle travel times, loading and unloading times, and blocking delays. Loaded travel can be directly estimated given the guidepath network layout, the production volumes, product mixes and routings. Estimating the empty travel requirements and blocking delays is more complicated because it depends on the system dynamic behavior, which is highly influenced by operational policies for the system like scheduling of loads and vehicle dispatching rules.

Maxwell and Muckstadt (1982) were the first to propose a transportation problem formulation to estimate the empty vehicle travel time. They compute the net flow at each pickup/drop-off location as the difference between the total number of loads delivered and the total number of loads picked up from that location. The empty vehicle travel between every pair of pickup/drop-off locations is assigned so as to minimize the total empty vehicle travel in the system.

Kuhn (1983) presents a more realistic estimation method to determine the empty vehicle travel, using a factoring method that considers the total number of loads delivered to a pickup/drop-off location as the number of empty trips originating from that location. The allocation of empty vehicles from one location to other locations is proportional to the total loads picked up at those other locations.

Egbelu (1987) compares four analytical methods for estimating the vehicles requirements, where each model uses assumptions concerning empty travel and vehicle blocking. The first method assumes that the empty travel is equal to the loaded travel. The second method includes blocking and idle time factors. The third method first computes the net flow into each P/D location, and then the total empty travel is computed by multiplying the average loaded traveled distance by the total net flow for all locations. The fourth method is based on the same reasoning used by Kuhn (1983) where the number of empty trips from location *i* to location *j* depends on the proportion of deliveries to *i* and the number of pickups from *j*. The four methods are compared to simulation results for a range of dispatching rules. Significant differences are reported between the simulation and the first three methods under most of the dispatching scenarios.

Tanchoco et al. (1987) propose a method based on the Computerized Analysis of Networks (CAN-Q) model; a queuing model, used for analyzing workflows through a manufacturing system based on the steady-state behavior. Empty vehicle travel is not considered in the model. The results for the minimum vehicle requirements from the CAN-Q model are compared to the simulation model results. The results indicate that CAN-Q underestimates the number of vehicles required. They conclude that this model should only be used as an approximation tool to get a lower bound on the vehicle requirements.

Mahadevan and Narendran (1990) propose an analytical method to estimate the number of vehicles required for a flexible manufacturing system (FMS), the main feature of their model is the consideration of routing flexibility and limited buffer capacity at the stations, by associating fixed probabilities with these events. However, the model does not consider empty vehicle travel. Later in Mahadevan and Narendran (1993), the empty vehicle flow is included in the total requirements under the assumption that vehicles do not deliver and pick up from the same pickup/drop-off location.

Malmborg (1990) proposes an approach to compute a lower bound and an upper bound on the empty vehicle travel. The lower bound is computed by a model similar to the one developed by Maxwell and Muckstadt (1983). The upper bound is computed from a similar model, except that the frequency of empty trips is based on total number of loads delivered at or pick up from each location, rather than on net flows. Furthermore, the objective is to maximize the empty vehicle trips, implying that after each vehicle delivers its load, it is routed to the farthest location. Malmborg argues that the actual empty travel is a convex combination of the two bounds, where the weights depend on the vehicle dispatching and load selection policies defined for the system. Malmborg also proposes a control-zone concept to model the vehicle-based system, in which one vehicle at a time is allowed to travel through a zone. The control-zone model is used to approximate the effects of vehicle blocking using an *M/M/n* queueing model to predict the performance of a system with *n* zones.

Sinriech and Tanchoco (1992) suggest a model that combines the system performance and the costs in the optimization model to determine the number of vehicles. The system performance is measured by the vehicles throughput. The approach taken in their study is a multi-criteria optimization model with two goals. First, the target values for these goals are calculated, then weights are defined for each goal. The objective function is to minimize the deviation from the target values

Rajotia et al. (1998) propose a similar model to the one developed by Maxwell and Muckstadt (1983), using the total flows instead of the net flows, and they introduce additional constraints on the empty flow from the drop-off to the pick-up locations for a single station. Empirical approximation of vehicle blocking and waiting times factors are incorporated into the vehicle requirements calculations.

Johnson and Brandeau (1995) develop and solve an analytic model for the design of a multi-vehicle system that carries loads from a central storage depot to a shop floor works centers. The objective is to determine which workstations to include in the network and how many vehicles are needed to service those workstations so as to maximize the benefit, defined as the savings in labor minus the cost of operating and acquiring the vehicles. Constraints are imposed on the waiting times, which are approximated using an M/G/c queueing model.

Bakkalbasi (1990) develops analytic models to approximate the empty vehicle travel times for the following dispatching rules: First Come First Served (FCFS), Closest Load First, Closest Load with Time Priority, and Furthest Load First. Srinivasan et al. (1994) develop an analytic model to determine the throughput capacity of a network with single vehicle and multiple vehicles, under a modified FCFS dispatching rule, the modified rule overrides the FCFS rule whenever a move request is present at the vehicle's current location.

Sharp and Liu (1990) propose a multi-commodity network model formulation to examine the cost and effectiveness of adding shortcuts to an exiting guide path network and spurs to workstations. The objective is to minimize the cost of constructing the shortcuts and spurs and the cost of vehicle travel and congestion delays. Before the model is formulated, cost functions that represent spur construction and vehicle time as a function of the total vehicle traffic at a workstation are developed assuming a Poisson process for the arrival of vehicles. The authors apply queueing models to estimate the waiting times for different types of vehicles at diverging and merging nodes in the network.

Johnson (2001) develops expressions for empty vehicle travel using FCFS policy and Nearest Vehicle Rule (NVR). Johnson (2001), Johnson and Brandeau (1994, 1995), and Kobza et al. (1998) analyze AMHS using *M/G/c* queuing models; these models give good approximations provided vehicle assignments are based on a First Come First Served (FCFS) discipline. However, queuing results deviate considerably from simulation results when the vehicle dispatching is system state-dependent, such as Nearest Vehicle Rule (NVR). Benjafaar (2002) presents a *G/G/*1 approximation to model a single-device MHS for selecting among alternative layouts to minimize the expected WIP in the system. In Johnson (2001), a queuing model is used to estimate the performance of a multi-vehicle AMHS with NVR Dispatching. Johnson first develops an approximation for the distribution of the empty vehicles among the stations, then uses an M/G/c model to estimate the waiting time of loads. The latter results tend to be inaccurate because of the assumption of state-independent service time. Curry et al. (2003) propose a more accurate service-dependent queuing network model that generates approximations that are close to the simulation results but the time to solve the analytic model grows exponentially with the number of vehicles.

Hodgson et al. (1987) have attempted to model single-vehicle systems using Markov decision processes. Due to the large number of states in even a relatively simple AGVS, several constraints were applied to make the Semi-Markov problem tractable. Srinivasan et al. (1994) propose a single-vehicle queuing model to estimate the throughput of the vehicle where the vehicle dispatching to move request is based on a modification of the FCFS rule. In Bozer, et al. (1994), the throughput approximation is used to estimate the waiting time of move requests at each station; their estimates are quite close to the simulation results. The authors propose an extension of their model to multi-vehicle systems by adjusting the travel times assuming that an AMHS that has *K* vehicles can be replaced by a single device that travels *K* times faster. Results indicate good throughput estimates but significant errors in waiting time estimates because congestion and blocking delays are not modeled.

Roeder et al. (2004) propose a simulation of a simplified closed queueing network to model intrabay AMHS in semiconductor manufacturing. The approach has fewer data requirements than an explicit detailed simulation model of the system. The authors use an information taxonomy to quantify the differences between the explicit AMHS simulation and the queueing network approximation. The approximation captures the movement of the vehicles, interaction of the vehicles with the machine loadports, and processing of lots at machines. Vehicle blocking is not modeled and the paper does not provide detailed empirical results. None of the above models consider blocking of vehicles due to the inability to pass each other, which is a signifcant portion of the travel time in systems where there are no offline-docking locations (spurs).

In short, the past literature offers accurate models of discrete vehicle-based material handling systems that are simulation-based, which is impractical given the development, execution and maintenance times needed to have an accurate representation of the system. In the context of the system we analyze in this thesis, the analytic models offered in the literature have one or more of the following shortcomings:

- assume that system operates in a deterministic manner and therefore rely on network flow problems to estimate the optimal fleet size.
- fail to consider empty vehicle travel, and thus their models are oversimplified and provide optimistic estimates of the fleet size requirements.

- are developed for single-vehicle systems, which has limited applications in practice.
- use central server models that oversimplify the service time of the material handling system.
- assume ample capacity for vehicles at the pick-up and drop-off stations and thus can analyze the stations independently.

For the material handling system described in this research, the move request arrival process is stochastic, multiple vehicles operate on the tracks, the pick-up and drop-off stations have finite capacity that leads to significant vehicle-blocking. Assuming deterministic flow rates fails to capture the inherent queuing in the system due to the high variability of the production system. The ample loadports capacity for vehicles ignore vehicle-blocking in multi-vehicle systems, an essential aspect of the system we study, where blocking and queuing of vehicles is not only possible but very likely. Modeling the AMHS as a single server oversimplifies the system because it assumes that every move request has the same response and travel time.

In short, analytical models developed under one or more of the above assumptions fail to represent the actual system with acceptable accuracy. Estimates for AMHS throughput capacity and/or response times based on such models will generate designs that deviate significantly from the actual system.

CHAPTER 3

MARKOV CHAIN MODEL OF VEHICLE-BASED CLOSED-LOOP AMHS

Contemporary simulation technology can produce accurate assessments of integrated circuit factory (fab) production performance, including the contribution by the automated material handling systems (AMHS). However, the corresponding simulation models are both expensive and time-consuming to construct, and require long execution times to produce statistically valid estimates. These attributes render simulation ineffective as a decision support tool in the early phase of system design, where requirements and configurations are likely to change often. In this paper, we describe an analytical approach to AMHS performance modeling for a simple closed loop AMHS, such as is typical in supporting a 300mm wafer fab bay. In this system, due to the significant impact of vehicle blocking, a straightforward queueing network model which treats the material handling system as a central server can be very inaccurate. We propose an alternative model that estimates the MHS performance considering the possibility of vehicle-blocking. While the resulting large-scale model presents some computational challenges, it promises reasonably accurate estimates with computation times that are acceptable in a design environment.

3.1. Introduction

Within the past decade, the semiconductor industry has transitioned from 200mm to 300mm wafer fabrication. This transition is expected to produce 2.5 times more chips per wafer (Bonora and Feindel, 1998), at a cost 1.4 times more than the 200mm wafer

(Jones, 2003). The shift to 300mm wafers is a challenging and expensive transition. Full factory automation, required to maximize the productivity of the capital, to satisfy ergonomic restrictions caused by the weight and volume of 300mm wafer lot carriers, and to reduce the particle contamination and vibrational shocks on the wafers (Nadoli and Pillai, 1994), depends fundamentally on automated material handling systems (AMHS) as the medium through which factory operations control is implemented.

Constructing a 300mm fab is projected to cost $2-3 billion, (Jones, 2003), and the AMHS represents a significant share of the investment cost, as much as 3 to 5% of the total fab cost (Arzt and Bulcke, 1999). As IC manufacturers drive to reduce manufacturing cycle times in 300mm fabs, the performance of AMHS becomes a critical factor. AMHS may introduce lot delays or cause tool idle time by failing to move lots in the planned and allotted time.

Estimating AMHS performance in IC fabs is difficult, because of the complexity of the systems. The International Technology Roadmap for Semiconductors (ITRS[3],) characterizes the AMHS as having several vehicles, operating on a network with loops, intersections, spurs, and short-cuts, serving many different pick-up/deposit stations. The movement requirements appear to be random, and although they exhibit some temporal correlations, these correlations are not strong enough to permit precise scheduling of the AMHS resources.

[3] The International Technology Roadmap for Semiconductors (ITRS) periodically lays out a technology plan to guide the semiconductor industry in the coming decades. The latest study is summarized in the ITRS 2003 annual report. It provides the current estimates for research and development that is required over the next decade to meet the historical numbers in performance growth, size reductions, cost, etc.

A typical 300mm AMHS has a spine layout, illustrated in Figure 3.1, with a central material handling spine and loops branching on both sides to serve production equipment. There are two distinct operating scenarios: (1) the spine and the loops are decoupled, and vehicles are dedicated to the spine or to one of the loops; or (2) the spine and the loops are integrated, and vehicles may move freely between them (Pillai et al. 1999). In this paper, we consider only the first case, which means a wafer lot moving from a tool on one loop to a tool on a different loop must travel through the main spine, and will use three different vehicles, one in each loop and one in the spine. Automated storage units, referred to as stockers, are used to provide both temporary buffering for work-in-process and transfer between the bay and spine transport systems.

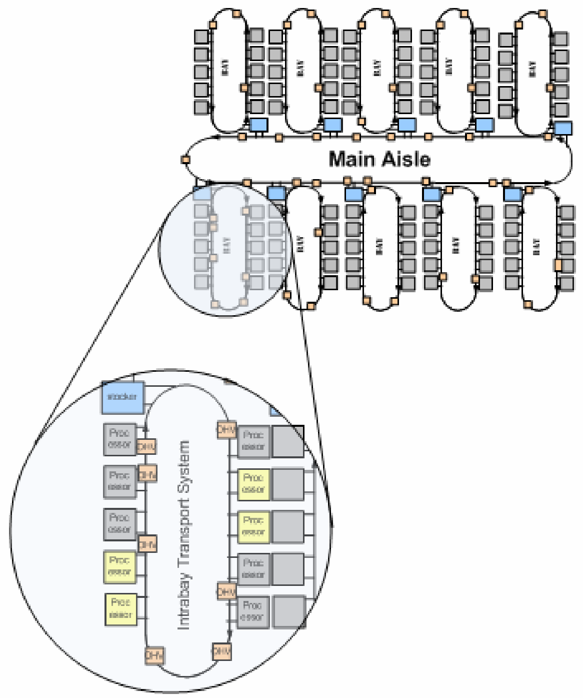


Figure 3.1 Closed-loop unidirectional interbay and intrabay AMHSs

Because of the space restrictions in the 300mm wafer fab bays, OHT vehicle travel is on a unidirectional closed loop without the ability for vehicles to pass each other, even when a vehicle stops to drop-off/pick-up a lot from the input/output port of a process tool or a stocker (SEMATECH report, 1997). Thus, failure to carefully synchronize vehicle movements on a given loop can lead to significant amounts of vehicle blocking and the possibility of lot delay as well as induced vehicle idle time.

3.2. Literature

3.2.1. Models of 300mm material handling

Throughout the literature, the importance of AMHS in 300mm wafer fabs has been repeatedly addressed, and research in this area can be broadly categorized into: (1) design optimization, targeting the guide path network layout design and calculating a feasible fleet size, and (2) performance evaluation of various AMHS methods or different AMHS configurations via simulation modeling.

In the area of design optimization, Peters and Yang (1997) propose a network flow formulation to determine the number and location of shortcuts for the interbay transport system in a spine layout fab. The objective function minimizes the tradeoff between the increase in shortcut construction cost and the decrease in material handling costs.

Ting and Tanchoco (2000) propose an analytical procedure to construct a unidirectional circular layout for the interbay system in 300mm fabs under the assumption that not all the stockers can be connected by a simple loop. First, candidate layouts for the main loop are constructed, and then, loop and shortcuts are added to connect the remaining stockers and eliminate flows on high-flow segments. In a later paper, Ting and Tanchoco (2001) use an analytical approach to develop a single and double rectilinear spine layout for overhead track layout design that connects tools to stockers in an open ballroom layout bay for a 300mm fab. The objective is to minimize the total loaded travel distances.

Steele (2002) proposes an algorithm to estimate the performance of an AMHS during the design process. Each design alternative is modeled as a network of nodes. The algorithm estimates the minimum number of vehicles required to deliver the required number of wafer lots and the average delivery times between each pair of source and destination nodes. The algorithm assumes infinite capacity queues and thus does not consider blocking.

To date, discrete event simulation has been the only methodology shown to give reliable estimates of fab-level AMHS performance. At the system design stage, however, large scale, high-fidelity simulation models are not a practical approach, because they take too long to develop, and require multiple lengthy executions to produce statistically valid estimates. Mackulak and Savory (2001) describe a study in which the AMHS experiments took over 250 hours of simulation time. Different approaches have been taken to overcome this problem. For example, Mackulak *et al.* (1998) propose developing a generic model that can be reconfigured according to the specific problem at hand, thereby reducing the model building time. Gaxiola and Mackulak (1999) describe the use of simple deterministic calculations in situations where the process requirements have not yet stabilized.

Pillai et al. (1999) discuss the issue of linking the interbay and intrabay tracks for a 300-mm fab layout. Rust et al. (2002) and Mackulak and Savory (2001) investigate the same problem by focusing on the impact of this decision on several AMHS performance measures.

Lin et al. (2003) propose using four different vehicle types to carry out the transport tasks from tool to tool. Type A vehicles move in an intrabay system and deliver the lots within the bay. Type B vehicles carry lots between the stockers. Type C vehicles carry lots from a tool in any bay to a stocker in the lot's destination bay. Type D vehicles move lots from a tool in any bay to a tool in any other bay. Three different transport methods using combinations of the four vehicle types were examined.

The conclusions in most of the studies depend on the specifications of the fab being modeled, and thus do not constitute generic design guidelines. At the system design stage, therefore, large scale, high-fidelity simulation models are not a practical approach and system designers are limited in the range of alternatives they can expect to evaluate in detail.

3.2.2. Analytical models of material handling systems

In the literature, analytical models of AMHS are usually based either on deterministic optimization models or queuing models. The former fails to capture queuing in the system which is essential to accurately estimate the key performance measures. Often, in analytical factory modeling, the material handling system is modeled by defining a "virtual" workstation between the processing tools in a product's routing. The delay associated with material handling is approximated by the processing time on this virtual workstation, which has a capacity equal to the number of vehicles available. This approach is appealing because it exploits well-understood queuing models. However, it has some inherent weaknesses. First, it assumes that the response time of the AMHS to a move request does not depend on the location of the load, nor on the vehicle distribution across the network. Second, it fails to capture the impact of vehicle-to-vehicle blocking, which, by consuming some of the available vehicle time, will degrade the capacity of the AMHS. Vis (2004) provides a survey of work in this area. Johnson (2001), Johnson and Brandeau (1994, 1995), and Kobza et al. (1998) analyze AMHS using M/G/c queuing models; these models give good approximations provided vehicle assignments are based on a First Come First Served (FCFS) discipline. However, queuing results deviate considerably from simulation results when the vehicle dispatching is system state-dependent, such as Nearest Vehicle Rule (NVR). Benjafaar (2002) presents a G/G/1 approximation to model a single-device MHS for selecting among alternative layouts to minimize the expected WIP in the system. In Johnson (2001), a queuing model is used to estimate the performance of a multi-vehicle AMHS with NVR Dispatching. Johnson first develops an approximation for the distribution of the empty vehicles among the stations, then uses an M/G/c model to estimate the waiting time of loads. The latter results tend to be inaccurate because of the assumption of state-independent service time. Curry et al. (2003) propose a more accurate service-dependent queuing network model that generates approximations that are close to the simulation results but the time to solve the analytic model grows exponentially with the number of vehicles. Bakkalbasi (1990) develops analytic models to approximate the empty vehicle travel times for the following dispatching rules: FCFS, Closest Load First, Closest Load with Time Priority, and Furthest Load First. Srinivasan et al. (1994) propose a single-vehicle queuing model to estimate the throughput of the vehicle where the vehicle dispatching to move request is based on a modification of the FCFS rule. In Bozer, et al. (1994), the throughput approximation is used to estimate the waiting time of move requests at each station; their estimates are quite close to the simulation results. The authors propose an extension of their model to multi-vehicle systems by adjusting the travel times assuming that an AMHS that has K vehicles can be replaced by a single device that travels K times faster. Results indicate good throughput estimates but significant errors in waiting time estimates because congestion and blocking delays are not modeled. Roeder et al. (2004) propose a simulation of a simplified closed queueing network to model intrabay AMHS in semiconductor manufacturing. The approach has fewer data requirements than an explicit detailed simulation model of the system. The authors use an information taxonomy to quantify the differences between the explicit AMHS simulation and the queueing network approximation. The approximation captures the movement of the vehicles, interaction of the vehicles with the machine loadports, and processing of lots at machines. Vehicle blocking is not modeled and the paper does not provide detailed empirical results. None of the above models consider blocking of vehicles due to the inability to pass each other, which is a signifcant portion of the travel time in systems where there are no offline-docking locations (spurs).

In this paper, we propose an analytic approach to evaluating AMHS performance in an IC fab. We consider vehicle-based systems, of which contemporary hoist-based overhead systems (OHT) are an example. We develop a queuing network type model, based on a detailed description of OHT operations, and propose to analyze the model as a large-scale Markov chain. With this approach we are able to estimate both AMHS throughput and move request delays. We model an OHT system configured as a simple loop, in which vehicles continuously travel the loop, and loads are served based on the First-Encountered-First-Served (FEFS) rule. FEFS is a decentralized policy, first presented by Bartholdi and Platzman (1989). In the FEFS, an empty OHV circulating the loop inspects the output buffer of a station (stocker or processor); if there is a lot waiting, the vehicle picks it up and delivers it to its destination. If the output buffer is empty, the vehicle travels to the next station and so forth until it encounters a waiting lot. An OHV carrying a lot (loaded/full) might pass other input and output buffers and experience delays if it has to wait while other vehicles drop-off or pick-up loads at those buffers. Our goal is to estimate these blocking delays in order to get a good approximations of both the actual throughput of the OHT system and the average response time to move requests. Bozer, et al. (1991) also use FEFS dispatching for a single-vehicle analytic model to approximate the throughput capacity of the vehicle. Our model differs because it is developed for multiple vehicles, where queuing and blocking of vehicles is possible.

3.3. Modeling approach

Our objective is to estimate the expected throughput capacity of the AMHS for a given set of input parameters expressed by the move requirements, the travel times, the layout of the stations on the AMHS closed loop track, and the fleet size. The analytical model provides estimates of a specific set of AMHS output variables that are essential to calculate the throughput capacity. These output variables are the proportion of time the vehicles spend traveling (empty and loaded), in service (loading and unloading), and being blocked (empty and loaded).

Rather than tracking the location of every vehicle while keeping a record of all the events that change the location and status of the vehicles, we focus on a subset of vehicle operations; operations that vehicles go through only at the drop-off and pick-up stations, eliminating the travel operations that occur on the track segments. Next, we enumerate the vehicle conditions relevant to our analysis. We choose to include the vehicle conditions that identify whether the vehicle has just arrived at the station (empty or loaded), is in service (loading/unloading), or is blocked (empty or loaded).

We then use a transition matrix to track the changes in the locations and the conditions of the vehicles. In the matrix, each possible location-condition-vehicle combination is identified as a state. The transition between the states is probabilistic and depends on the move requests rate, the number of vehicles, and the sequence of the machines on the AMHS loop.

Including the states when vehicles are located at stations simplifies the model, but creates a problem when the vehicles arrive at stations asynchronously. This problem is addressed by creating virtual locations to synchronize the vehicles' movements. For instance, if the travel time between two consecutive stations is half the loading/unloading time, we add one virtual station to each drop-off and pick-up station, and we make all the loading/unloading times equal to the travel times.

Assumptions on the arrival process of move requests allow us to analyze the transition between the states as a Markov chain. Some of the transition probabilities are known because they are easily calculated from the given problem parameters (such as the probability that a loaded vehicle will be dropping off its load at some station), but other transition probabilities are only partially determined by the problem parameters, but also influenced by the output variables that we are trying to calculate, specifically the arrival rate of empty-vehicles to stations.

Next, we present necessary conditions that ensure that the AMHS is able to meet the required throughput imposed by the machines. These conditions provide constraints on the unknown variables in the Markov chain. The resulting model is not a conventional

Markov chain, and hence the name Extended Markov Chain model, because it combines the Markov chain with a set of constraints that are necessary to solve for the unknown variables in the matrix and the AMHS output variables. The extended model provides a full rank system of equations that can be solved to give a unique set of estimates for the output parameters of the AMHS.

3.4. Closed-loop vehicle based AMHS

Figure 3.2 illustrates an example of a closed loop overhead transport system serving the stocker(s) and the processors (also referred to as production equipment or production tools) in the bay. Movement is unidirectional and multiple vehicles can be traversing the loop but they cannot pass each other.

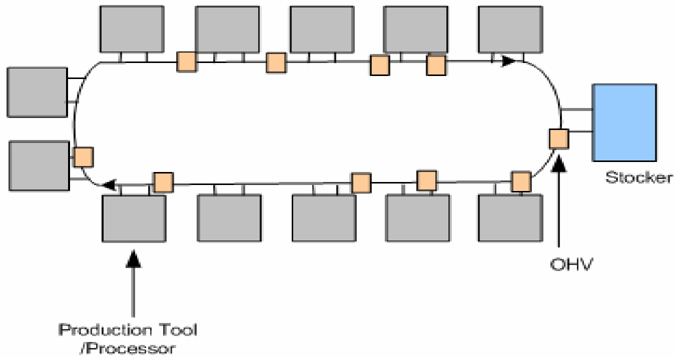


Figure 3.2 Unidirectional closed loop overhead transport system

In order to estimate the throughput accurately, we need to estimate the blocking delays at each machine—either stocker or tool. The main objective of the model is to quantify the duration of the blocking delay as a function of the layout of the transportation system, the demand rates, the speed of the vehicles and the number of vehicles circulating the loop.

Let $L(n)$ refer to the OHT directed loop with n vehicles. Let M be the set of machines in $L(n)$. Each machine has two load ports: an input port where loads are dropped off by the vehicle and an output port where loads are picked up by the vehicle to be delivered to their next destinations. Each port can accommodate one vehicle at a time. We use the term *station* to refer to the input and output ports of the machines. Thus, a loop serving *m* machines consist of $s = 2m$ stations. Let m_i denote machine *i*; then m_i has two stations: the drop-off station s_i^d, and the pick-station s_i^p. Figure 3.3 illustrates a network representation of the system in Figure 3.2 as a directed loop with stations represented as nodes, and the track segments as directed arcs.

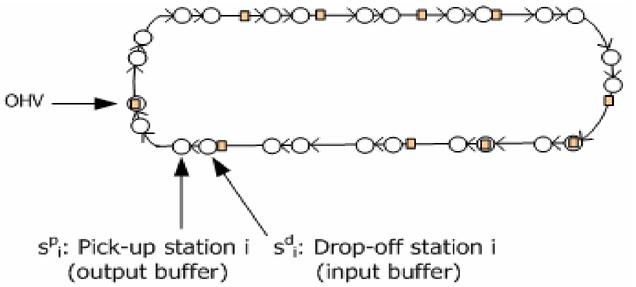


Figure 3.3 Network representation of the closed loop AMHS

3.4.1. Logic description

The vehicles are constantly circulating on the unidirectional loop. As an empty vehicle approaches m_i , it passes through the drop-off station s_i^d, then travels to the pick-up station s_i^p. If there is a load (job) waiting at s_i^p, the vehicle picks it up, which requires time delay *l* for loading the job and then delivers it to its destination, say machine m_j, visiting machines m_{i+1}, m_{i+2},…, m_{j-1}, and finally the load's destination the drop-off station of m_j, denoted by s_j^d. The vehicle does not stop at machines m_{i+1}, m_{i+2},…, m_{j-1} unless it is blocked by other vehicles. If the output port s_i^p is empty, the vehicle travels to s_{i+1}^d, then inspects the output port s_{i+1}^p and so forth until it encounters a waiting lot. Vehicles cannot travel on the same track segment

3.5. Notation

M: set of tools and stockers in the system.

m_i: machine i, which could be either a tool or a stocker, $m_i \in M$.

s_i^p : pick-up station of m_i, $m_i \in M$.

s_i^d : drop-off station of m_i, $m_i \in M$.

U_i : set of pick-up stations upstream of s_i^d .

D_i : set of drop-off stations downstream of s_i^d .

p_{ij} : probability that a load which is picked up from s_i^d is destined to s_j^d .

λ_i : mean arrival rate of move requests picked up from s_i^p .

Λ_i : mean arrival rate of move requests dropped at s_i^d .

r_i : probability that a loaded vehicle drops-off its load at s_i^d .

q_i : probability that an empty vehicle encounters a waiting load at s_i^p .

α_i^d : rate of loaded vehicles arrivals to s_i^d .

α_i^p : rate of loaded vehicles arrivals to s_i^p .

ε_i^d : rate of empty vehicles arrivals to s_i^d .

ε_i^p : rate of empty vehicles arrivals to s_i^p

θ: arrival rate of empty and loaded vehicles to stations.

3.6. Model assumptions

In manufacturing systems, demands for transportation depend on the jobs' release rates, and the routing sequences for jobs. The assumptions will be separated into system assumptions and mathematical assumptions. The system assumptions control the derivation of the analytical model while the mathematical sssumptions are necessary for model tractability.

System assumptions:

1. Move rquests are given as steady-state values per time period.
2. Flow is conserved at each machine and at the stocker (i.e. $\Lambda_i = \lambda_i, \forall i \in M$).
3. Each vehicle moves one load at a time operating under the FEFS rule.
4. Travel times, and loading and unloading times are deterministic.

Mathematical assumptions:

5. Move requests rates occur according to a Poisson process.

3.6.1. Disscussion of the assumptions

Assumption (1) is necessary because the analytical model is based on the steady-state behavior of the system. Even though the fab is a dynamic environment and release rates change frequently, the design could be based on the worst case (highest) expected release rate, or on the value a fab designer would like to base the AMHS design on. Assumption (2) is valid because every load delivered is also picked up, and so we can assume that the flow is conserved at every machine. In practice, this assumption is accurate for prcocessor tools but not necessarily for the stockers if the bay has multiple stockers and the same load can arrive to the bay and depart from two different stockers. This case can be included by small adjustments in the model but will not be pursued in this research.

Assumption (4) is reaonable for highly automated systems as in 300mm wafer fab and we can safely assume deterministic travel and loading/unloading times.

If the loaded vehicle arrivals at a station follow a Poisson process, then assumption (5) would be theoretically justified. It is an unlikely assumption since the travel times are expected to be determinisitic. However, if the utilization of processing tools is high, the departure process variability out of the processing tools will be elevated (Hopp and Spearman, 2000), and as a result the from-to matrix induces a level of variability that makes the assumption of Poisson move requests arrivals acceptable.

3.7. Markov chain model analysis

In this section, we develop a probabilistic model of the system *L*(*n*). We propose to analyze the system as a Markov chain. Since the move request (load) arrivals follow a Poisson process, by the PASTA (Poisson Arrivals See Time Averages) property, (Wolff, 1982), we assume that the instant at which a vehicle arrives to s_i^p is a random point in time (follows from assumptions 5), and the discrete Markov chain is embedded at the points in time when the vehicle just arrived or just finished service at a station. The Markovian property is explained by the fact that Poisson arrivals occur completely random in time. The system transition between the possible states is assumed to be Markovian, i.e., the next state of the system depends only on the current state, and not on the path taken to reach the current state.

A vehicle can experience two types of delays: (a) Queuing delays occur at pick-up and drop-off stations due to the time needed for the other vehicles to clear the station, as illustrated in Figure 3.3.

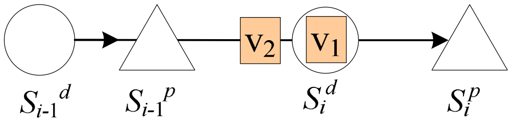


Figure 3.4 Queuing delay illustration

(b) Blocking delay is illstrated in Figure 3.4, and occurs when a vehicle has finished its service at its current station but cannot move because the downstream station does not have any space to accommodate it.

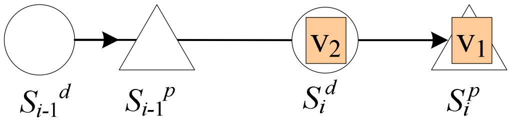


Figure 3.5 Blocking delay illustration

In this paper, we analyze systems that experience the second type; blocking delays. A common control mechanism in vehicle-based system does not allow vehicles to travel simultaneously on the same track segment (SEMATECH report, 2001), as a result queuing delays are not possible.

Due to the limited space for vehicles at stations, blocking of upstream stations is likely to occur. It is not possible to analyze each station independently because the number of vehicles queued at one station impacts the service rate of other stations. The analysis is complex because the service rate of stations is state-dependent, where the *state* is defined by every vehicle's location and condition. The location of a vehicle specifies the station at which the vehicle is receiving service or arriving. The condition of the vehicle specifies whether it is loaded, empty, blocked while empty, blocked while loaded, or receiving service (picking-up or dropping-off a load), and we use *f*, *e*, *b*, *k*, and *s* to denote each of these states, respectively.

In our model, we want to have a discrete set of *states* for the AMHS, simplifying the model by avoiding the continuous traveling process. We propose to consider only those points in time where a vehicle is located at a station. Specifically, in the state space we do not consider the *state* of a vehicle traveling on a track segment in between stations.

Although this assumption simplifies the model, it creates a problem when there are multiple vehicles, and the vehicles arrive at stations asynchronously, which raises the questions of how to determine what *state* the system is in, and what *state* it transitions to. If the travel time between every two consecutive locations as well as the loading and unloading times are all equal, we would not have this problem because the vehicles will always move synchronously. To overcome this problem, we create virtual locations in order to synchronize the vehicles' movements. For instance, if the travel time between two consecutive stations is half the loading/unloading time, we add one virtual station to each drop-off and pick-up station, and we make all the loading/unloading times equal to the travel times. In this case, if in some state *r*, one vehicle is at station s_i^d and needs to travel to s_i^p, while another vehicle is starting to unload at some other station s_j^d, $j \neq i$, in the next state, the first vehicle will be arriving to s_i^p and the other vehicle will be unloading at the virtual station s_j^d.

We characterize a state by specifying the condition and location of each vehicle. Each vehicle is defined by three characters:

$$(m_i, i = 1,2,..., M), (p, d), (e, f, b, k, s)$$

The first and second characters describe, respectively, the machine occupied by the vehicle and the station type (pick-up or drop-off), and the third character specifies the condition of the vehicle (empty, loaded, blocked/empty, blocked/loaded, and receiving service). For example, consider a system with two machines and two vehicles. There are two pick-up stations: s_1^p, s_2^p and two drop-off stations: s_1^d, s_2^d. State

$\{(1,p,s),(1,d,f)\}$ indicates that there are two vehicles at m_1, the first one is loading from s_1^p and the other is arriving loaded at s_1^d.

3.7.1. The transition probabilities

Consider the transition matrix **R**, which specifies the movement of the system between the states. The position and type of every vehicle (*e*, *f*, *b*, *k*, or *s*) determines the system transitions. An empty vehicle arriving to a drop-off station will certainly move to the next pick-up station, and so is the case for a loaded vehicle arriving to a pick-up station. However, a loaded vehicle approaching a drop-off station s_i^d will leave loaded if the load was not destined to that station, this happens with probability (1-r_i) which depends on the rate of moves destined to this station. Similarly, an empty vehicle approaching a pick-up station s_i^p will leave empty if there was no load waiting at s_i^p, this happens with probability 1-q_i which depends on the rate of moves originating at s_i^p and on the rate of empty vehicle arrivals to s_i^p. A vehicle receiving service at s_i^d must be dropping-off its load, and thus it will certainly move empty to s_i^p, and a vehicle receiving service at s_i^p must be picking a load and thus it will certainly move loaded to s_{i+1}^d. We demonstrate the transitions through the following example.

Example: consider a closed-loop OHT system with two vehicles (n=2) and five machines (10 pick-up and drop-off stations) (Figure 3.5). All stations and track segments have capacity for one vehicle, therefore if a vehicle is dropping-off a load at s_1^d the vehicle behind it has to wait at s_5^p until the first vehicle finished its drop-off.

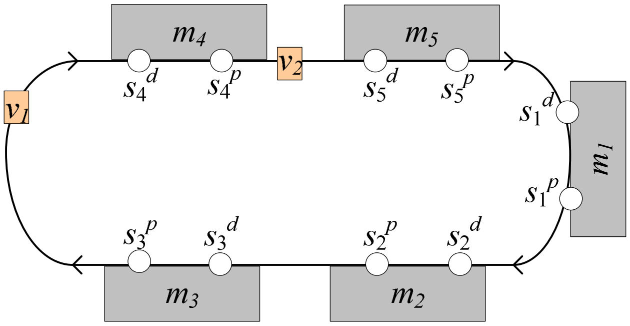


Figure 3.6 A 5-machine 2-vehicle example

Each state is defined by the string that specifies for each vehicle its condition and location. Consider the states transition diagram partially illustrated in Figure 3.6. State $\{(1,p,e),(1,d,e)\}$ indicates that there is one empty vehicle at s_1^p and one empty vehicle at s_1^d. The transitions from this state depend on whether the first vehicle finds a load at s_1^p, which happens with probability q_1. If the vehicle finds a load it starts receiving its service at s_1^p while the second vehicle is blocked and empty, and the system enters state $\{(1,p,s),(1,d,b)\}$. The first vehicle does not find a load waiting with probability 1-q_1, and in this case, the first vehicle moves empty to s_2^d while the second vehicle moves empty to s_1^p, and the system enters state $\{(1,p,e),(2,d,e)\}$. Consider the transition from state $\{(1,p,e),(2,d,f)\}$ to state $\{(1,p,s),(2,d,s)\}$, in the first state, one vehicle is arriving empty to s_1^p, and the other is arriving loaded to s_2^d, in the second state, both vehicles are receiving service at the same stations they were located. This transition happens if the first vehicle finds a load waiting at s_1^p with probability q_1, and the second vehicle drops off its load at s_2^d with probability r_2.

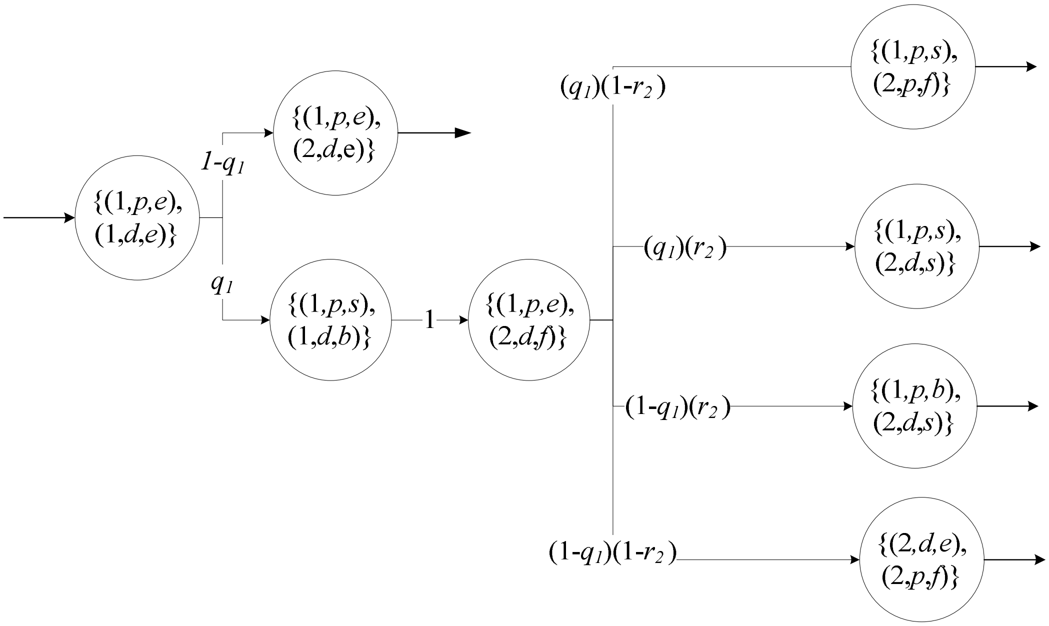

{(1,*p*,*e*), (1,*d*,*e*)}
$1\text{-}q_1$
q_1
{(1,*p*,*e*), (2,*d*,e)}
{(1,*p*,*s*), (1,*d*,*b*)}
1
{(1,*p*,*e*), (2,*d*,*f*)}
$(q_1)(1\text{-}r_2)$
{(1,*p*,*s*), (2,*p*,*f*)}
$(q_1)(r_2)$
{(1,*p*,*s*), (2,*d*,*s*)}
$(1\text{-}q_1)(r_2)$
{(1,*p*,*b*), (2,*d*,*s*)}
$(1\text{-}q_1)(1\text{-}r_2)$
{(2,*d*,*e*), (2,*p*,*f*)}

Figure 3.7 Part of the states transition diagram for the 5-machine 2-vehicle example

Now, consider the issue of asynchronous vehicle arrivals. Suppose the travel time from a loading station s_i^p to the downstream unloading station s_{i+1}^d is equivalent to the loading and unloading times but is double the internal travel time from s_i^p to s_i^d. We create a virtual station for each loading station and unloading station so that every vehicle movement is synchronized. In this case, suppose the system was in state $\{(1,p,e),(2,d,e)\}$, which is two empty vehicles one at s_1^p and the other at s_2^d. If the first vehicle finds a load waiting, it transitions to $(1,p,s)$ and since the other vehicle is empty at an input buffer, it will keep moving and transitions to $(2,p,e)$. Hence, the system enters state $\{(1,p,s),(2,p,e)\}$. Now, due to the unequal travel and loading times, we created a virtual loading station to which a vehicle moves to if it requires service, thus from this latest state $\{(1,p,s),(2,p,e)\}$, the system enters either state $\{(1,p,s),(2,p,s)\}$ or $\{(1,p,s),(2,d,e)\}$ depending on whether the second vehicle finds a load at s_2^p or not, respectively. This way every state transition requires the same amount of time, which in our example is the internal travel time from s_i^p to s_i^d . It is important to point out that during implementation and computer programming in order to generate the states and the transition matrix, we used a different notation for each state, wherein the stations are labeled consecutively, disregarding whether the station is a pick-up, drop-off or a virtual station. For example, in a system with five pick-up and five drop-off stations that requires five added virtual stations, there would be 15 unique stations labeled from 1 to 15. Therefore, during implementation, a state is defined by the string that specifies for each vehicle its condition and the unique station number where the vehicle is located.

3.7.2. The Markov chain steady-state analysis

Let $\boldsymbol{\upsilon}$={ υ_r} r=1, …, |R|, where v_r denotes the steady state probability of visiting state r. For a finite state, positive recurrent Markov Chain, the steady-state probabilities can be uniquely obtained by solving the square system of equations (Ross, 2000):

$$\mathbf{R}\boldsymbol{\upsilon} = \boldsymbol{\upsilon} \tag{3.1}$$

$$\sum_{\forall r \in R} \upsilon_r = 1 \tag{3.2}$$

The elements in the transition matrix are the transition probabilities between states, discussed earlier in Section 3.7.1. Some of these probabilities are unknown, specifically, the load-encountering probabilities. To see this, the probability that a loaded vehicle at some drop-off station will unload its load is determined entirely by the rate at which stations send loads to each other. On the other hand, the probability that an empty vehicle will pick-up a load from some pick-up station depends on the rate of move requests and also on the rate of empty-vehicle visits to pickup stations, which in turn depends on the number of vehicles in the system, and this relationship is not necessarily linear because of vehicle-blocking.

Each element r_{xy}, $x = 1,2,...|R|,\ y = 1,2,...,|R|$, denoting the transition probability from state *x* to state *y* in the transition matrix **R** can be written in terms of the unknown probabilities $\mathbf{q} = \{q_j\}$ *j*=1, …, *M* as follows:

$$r_{xy} = \prod_{j=1}^{M} \left(q_j^{\,b_{jxy}} (1-q_j)^{1-b_{jxy}} \right)^{h_{jxy}} a_{xy} \qquad x = 1,2,...|R|,\ y = 1,2,...,|R| \tag{3.3}$$

Where b_{jxy} and h_{jxy}, $j = 1,2,...,M$ take values of either 0 or 1, and a_{xy} $x = 1,2,...|R|,\ y = 1,2,...,|R|$ are computed from the problem parameters. The unknowns in expression (3.3) are the load encountering probabilities q_j, $j = 1,2,...,M$, all the other parameters depend on the states *x* and *y*.

The number of unknowns in equations (3.1) and (3.2) is |*R*| unknowns for the steady-state probabilities, plus |*M*| for the unknown load-encountering probabilities $\mathbf{q}$. In Section 3.7.3, we derive conditions that provide additional equations to solve for $\boldsymbol{\upsilon}$ and $\mathbf{q}$.

3.7.3. Conservation of vehicle flow

Let θ denote the arrival rate of vehicles to stations. Since the vehicles are circulating on a loop, the arrival rate θ is identical for all station and must satisfy:

$$\alpha_i^d + \varepsilon_i^d = \theta \qquad \forall i \in M \tag{3.4a}$$

$$\alpha_i^p + \varepsilon_i^p = \theta \qquad \forall i \in M \tag{3.4b}$$

α_i^d (α_i^p) denote the rate of loaded vehicles arrivals to s_i^d (s_i^p). ε_i^d (ε_i^p) denote the rate of empty vehicles arrivals to s_i^d (s_i^p). The α_i^d's values can be obtained by observing that the loads carried by a vehicle passing through s_i^d are those that originate from stations upstream of s_i^d for delivery to those downstream of and including s_i^d.

$$\alpha_i^d = \sum_{j \in U_i} \sum_{k \in D_i \bigcup s_i^d} \lambda_j p_{jk} \ , i \in M \tag{3.5}$$

Since we assume a simple closed loop, and that every load dropped off at s_i^d will be picked up from s_i^p, the flow of loaded vehicles is equal for every station, thus $\alpha_i^d = \alpha^d$, $\forall i \in M$ and the specific values can be calculated from:

$$\alpha^d = \sum_{j=2}^{M} \sum_{k=1}^{j} \lambda_j p_{jk} \tag{3.6}$$

In the cases of more complex network configurations with shortcuts and intersections, the flow of loaded vehicles is not the same for every station and the general expression from (3.5) would be more appropriate.

Λ_i denotes the rate of load arrivals to s_i^d, again because of conservation of flow at each machine, we have $\Lambda_i = \lambda_i, \forall i \in M$.

We can now calculate the probability that a loaded vehicle drops-off its load at s_i^d, denoted by r_i as:

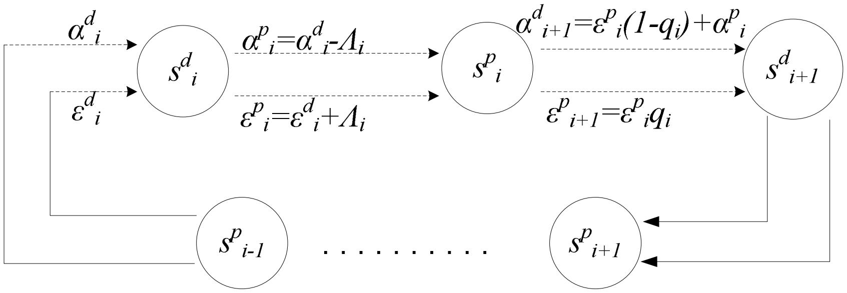
$$r_i = \frac{\Lambda_i}{\alpha^d}, \forall i \in M \tag{3.7}$$

Figure 3.7 illustrates how the empty and loaded vehicle arrival rates change as a vehicle travels between stations. Empty vehicles arriving to s_i^d stay empty as they move to s_i^p, while the loaded vehicles arriving to s_i^d drop-off their load at a rate Λ_i. Thus the rate of empty vehicle arrivals to s_i^p is:

$$\varepsilon_i^p = \varepsilon_i^d + \Lambda_i, \forall i \in M \tag{3.8}$$

The rate of empty vehicle arrivals to s_{i+1}^{d} depends on the probability that an empty vehicle at the upstream pick-up station s_i^p did not find a load waiting, denoted by $1-q_i$, thus:

$$\varepsilon_{i+1}^{d} = (1-q_i)\,\varepsilon_i^{p}\,, \forall i \in M \tag{3.9}$$



α^{d}_{i}
s^{d}_{i}
ε^{d}_{i}
$\alpha^{p}_{i}=\alpha^{d}_{i}-\Lambda_i$
$\varepsilon^{p}_{i}=\varepsilon^{d}_{i}+\Lambda_i$
s^{p}_{i}
$\alpha^{d}_{i+1}=\varepsilon^{p}_{i}(1-q_i)+\alpha^{p}_{i}$
$\varepsilon^{p}_{i+1}=\varepsilon^{p}_{i}q_i$
s^{d}_{i+1}
s^{p}_{i-1}
. . . .
s^{p}_{i+1}

Figure 3.8 Conservation of vehicles' flow

3.7.4. Stability conditions

For a stable system, the rate of pick-ups from s_i^p must equal the rate of move requests generated at s_i^p, which is λ_i. Also, the rate of pick-ups from s_i^p is the rate of empty vehicle arrivals to s_i^p (denoted by ε_i^p) multiplied by the probability of finding a load waiting denoted by q_i, thus:

$$q_i\varepsilon_i^p = \lambda_i \quad \forall i \in M \tag{3.10}$$

We now link the empty vehicle arrival rates to the Markov chain steady-state probabilities. Let E_i^p be the set of states where a vehicle arriving to pick-up station s_i^p is empty $i = 1,\ldots,M$. Let $v_{E_i^p}$ denote the stationary probability of visiting state set E_i^p, which, conceptually, is the steady-state probability of empty vehicle visits to s_i^p, and can be obtained from:

$$\upsilon_{E_i^p} = \sum_{r\in E_i^p} \upsilon_r, \forall i \in M \tag{3.11}$$

Since we have synchronized the movement of every vehicle, all transitions take the same amount of time, which is the least common denominator of all the possible time delays, denoted by $T_{\min}$. The rate of visits to each state is $\upsilon_{E_i^p} / T_{\min}$. The relationship between $\upsilon_{E_i^p}$ and ε_i^p is:

$$\frac{\upsilon_{E_i^p}}{T_{\min}} = \varepsilon_i^p, \forall i \in M \tag{3.12}$$

From (3.11) and (3.13), we get the following necessary stability conditions:

$$q_i = \frac{\lambda_i T_{\min}}{\upsilon_{E_i^p}} \ \forall i \in M \tag{3.13}$$

Combining equations (3.1), (3.2), (3.11), and (3.13) we have |*R*|+2|*M*| equations and |*R*|+2|*M*| unknowns, and we can find a unique solution to the system of equations and calculate the steady-state probability for every state, the blocking probabilities and other performance measures for the AMHS.

Proposition 1: for the system of nonlinear equations:

$$\begin{aligned} &\mathbf{Rv} = \mathbf{v} \\ &\sum_{\forall r\in R} v_r = 1 \\ &v_{E_i^p} = \sum_{r\in E_i^p} v_r, \forall i \in M \\ &q_i = \frac{\lambda_i T_{\min}}{v_{E_i^p}}, \forall i \in M \end{aligned}$$

There exists a unique solution for stationary probabilities **v** and the load-encountering probabilities **q** if the AMHS is stable (i.e. the AMHS can handle all the move requests within the planning horizon).

Proof:

The outline of the proof is as follows:

1. If the unknown transition probabilities in the transition matrix are known, then there exists a solution to the stationary probabilities of the Markov chain states, assuming that the AMHS stable; this follows from the fact that the Markov chain is finite, irreducible, and positive recurrent.
2. We then prove by contradiction that there cannot be two different sets of transition probabilistic for the same problem instance that will produce the same solution to the stationary probabilities of the Markov chain states.

Formal proof:

If the load-encountering probabilities q are known, and if the AMHS is stable, the Markov chain with transition matrix R is ergodic because it is finite and recurrent (Ross, 2000). For an ergodic Markov chain, there exists a unique solution to the stationary probabilities by solving the system of equations:

$$\begin{aligned}&\mathbf{Rv} = \mathbf{v}\\&\mathbf{v1} = \mathbf{1}\end{aligned}$$

Since the transition matrix R is a function of the unknown load-encountering probabilities **q**, we need to prove that there exists only one **q** for a given instance of the problem, where an instance is defined by $L(n)$, the directed loop with n vehicles, M the set of machines in $L(n)$, λ the arrival rate of move requests, and T_{min}, the minimum transition time between states. We now prove, by contradiction, that there is only one solution for **v** and **q.**

- Suppose there are two vectors $\mathbf{q^1}$ and $\mathbf{q^2}$, and two vectors $\mathbf{v^1}$ and $\mathbf{v^2}$ such that $\mathbf{q}^1 \neq \mathbf{q}^2, \mathbf{v}^1 = \mathbf{v}^2$, and both ($\mathbf{q^1}$, $\mathbf{v^1}$) and ($\mathbf{q^2}$, $\mathbf{v^2}$) satisfy the system of equations in Proposition 1.

For each machine *i*, the load encountering probability is obtained from the expression: $q_i = \dfrac{\lambda_i T_{\min}}{v_{E_i^p}} \quad \forall i$ where $v_{E_i^p} = \sum\limits_{r \in E_i^p} v_r$.

If for some machine *i*, there are two solution q_i^1 and q_i^2 where $q_i^1 \neq q_i^2$, then $v_{E_i^p}^1 \neq v_{E_i^p}^2$ but $v_{E_i^p} = \sum\limits_{r \in E_i^p} v_r$, therefore there is at least one state, say state *k*, $k \in E_i^p$ such that $v_k^1 \neq v_k^2$, which contradicts the initial assumption that there are two solutions such that $\mathbf{q}^1 \neq \mathbf{q}^2$ and $\mathbf{v}^1 = \mathbf{v}^2$. ■

- Suppose there are two vectors $\mathbf{q^1}$ and $\mathbf{q^2}$, and two vectors $\mathbf{v^1}$ and $\mathbf{v^2}$ such that $\mathbf{q}^1 = \mathbf{q}^2, \mathbf{v}^1 \neq \mathbf{v}^2$, and both ($\mathbf{q^1}$, $\mathbf{v^1}$) and ($\mathbf{q^2}$, $\mathbf{v^2}$) satisfy the system of equations in Proposition 1.

If $\mathbf{q}^1 = \mathbf{q}^2$, the transition matrix for both solutions is equal, thus $\mathbf{R}(\mathbf{q}^1) = \mathbf{R}(\mathbf{q}^2) = \mathbf{R}$. If $\mathbf{v}^1 \neq \mathbf{v}^2$, the square system of *linear* equations:

$$\mathbf{vR} = \mathbf{v}$$
$$\sum_{\forall r \in R} v_r = 1$$

has two different solutions. This contradicts the initial assumption that for an ergodic Markov chain, there is a unique solution to the steady-state probabilities v. ■

We have proved that if the AMHS is stable, and a solution to the system in proposition 1 exists, then this solution is unique. We defined stability as the ability of the vehicles to handle all the move requests within the planning horizon. We use enumeration of the number of vehicles (*n*) in order to find the solution. Let *B* denote the total number of vehicle buffers (locations vehicles can occupy). The maximum number of vehicles on the loop, n_{max}, has to satisfy:

$$n_{\max} \leq B - 1 \qquad (3.14)$$

otherwise, there would be a deadlock situation. We use the following algorithm to find the minimum number of vehicles for a stable AMHS:

1. Set the number of vehicles *n*=1.
2. Solve the system of equations in proposition 1:
    - If a solution exists, it is unique, exit the algorithm.
    - If a solution does not exist:
        - o If n=B-1, a solution does not exist, the AMHS cannot satisfy the move requirements, regardless of fleet size.
        - o If n<B-1, then set n=n+1, go back to step 2. ■

3.8. AMHS performance measures

The vehicle dispatching policy for the AMHS analyzed in this research dictates that vehicles should constantly circulate on the loop whether or not there are jobs waiting for pickup, and this implies that the vehicles are not dispatched to pick-up loads but rather “encounter” these loads as they circle the loop. This creates a dilemma when measuring the AMHS performance. For instance, the percentage of time vehicles are idle is not a meaningful measure since the vehicles are not dispatched to the jobs and in some situations all the vehicles can be traveling empty even though there are loads waiting to be picked up. Therefore, we choose to measure the performance using three metrics:

1. *AMHS utilization (ρ):* the percentage of time vehicles are dropping off/picking-up a load, and traveling loaded (including being blocked and loaded).

   Let S be the set of states where a vehicle is starting service (drop-off, or pick-up) at any station. In some sense, this measures the effectivity of the AMHS because it presents the percentage of time vehicles are actually carrying loads.

Let F be the set of states where a vehicle is arriving loaded to any station.

Let K be the set of states where a vehicle is blocked while loaded at any station.

Let υ_S , υ_F , and υ_K denote the stationary probability of visiting state sets S , F , and K , respectively. ρ can be estimated from:

$$\rho = \upsilon_S + \upsilon_F + \upsilon_K = \sum_{r \in S \cup F \cup K} \upsilon_r \tag{3.15}$$

2. *Percentage of time vehicles are blocked (β)*: Let B be the set of states where a vehicle is blocked while empty at any station, β can be estimated from:

$$\beta = \upsilon_K + \upsilon_B \sum_{r \in K \cup B} \upsilon_r \tag{3.16}$$

3. *Expected time between two consecutive empty vehicle arrivals to a pick-up station i, ($T_{E_i^p}$),* this measure is related to the throughput capacity of the AMHS. We need to distinguish here between throughput capacity, which is the number of move requests the AMHS can handle in a given period, and the AMHS throughput, which is the number of moves the AMHS actually does handle in a given period. As the time between two consecutive empty vehicle arrivals to each pick-up station decreases, the AMHS can handle more move requests.

$T_{E_i^p}$ is estimated from the visit ratios to state set $v_{E_i^p}$ as:

$$T_{E_i^p} = \frac{1}{\varepsilon_i^p} = \frac{T_{\min}}{v_{E_i^p}} \tag{3.17}$$

3.9. Numerical example

We use the layout in Figure 3.5 to compare the analytical model estimates of AMHS utilization, blocking, and the time between empty vehicles arrival to each pick-up station to values obtained from discrete event simulation. In Figure 3-5, we have 4 processing equipment (m_2 through m_5) and one stocker (m_1) for a bay that serves three products (p_a, p_b, p_c). The total arrival rate to the stocker is $\lambda = \lambda_a + \lambda_b + \lambda_c$ jobs per minute. Each job type is released and processed according to the routes given in Table 3.1. The resulting from-to matrix is presented in Table 3.2.

Table 3.1 Arrival rate and routing of products

Product Type	Routing	Jobs/Hour
p_a	1-2-3-4-1	12
p_b	1-3-5-1	6
p_c	1-4-2-1	4

Table 3.2 Flow matrix (Loads/hour)

From/To	1	2	3	4	5
1		12	6	4	
2	4		12		
3				12	6
4	12	4			
5	6				

In this example, loads delivered to the stocker exit the system. The processing times at each processing station are assumed to be deterministic, and the mean processing time is such that processors utilization is 60%. The distance from the loading station, s_i^p

to the unloading station s_{i+1}^d, is 15 feet. The distance from the unloading station s_i^d to the loading station s_i^p is 5 feet. Each vehicle travels at a speed of 60 ft/min, empty or loaded, and it takes 15 seconds to pick-up or drop-off a load.

We compare the interarrival times of empty vehicles at each pick-up station estimated from the analytic model and a simulation model, at two fleet sizes: two and three vehicles. We also compare the estimates of AMHS utilization and proportion of time vehicles are blocked. We used AutoMod simulation software to obtain simulated values for this performance measure based on 10 replications of 100 days each; we also made sure that the system reached steady state before we started collecting statistics. In the simulation model, move requests arrive at the stocker according to a Poisson process. However, move request arrivals at processor stations are the result of lot arrival and operation times at the processor station, i.e. we did not force them to follow a Poisson process.

The analytical and simulated expected interarrival time of empty vehicles are shown in Tables 3.3 and 3.4 at different combination of move requests arrival rate and fleet size. The absolute error represents the difference between the analytical and the average obtained from the simulation.

Table 3.3 Analytical and simulated expected times between empty vehicle interarrival times (sec.) n=2 vehicles λ=22 jobs/hr

Station	$T_{analytic}$	$T_{simulation}$	*Rel. error*
1	89.63	91.52	-0.02
2	105.78	107.21	-0.01
3	99.83	102.35	-0.02
4	105.85	107.18	-0.01
5	151.47	155.51	-0.03
Average abs. error			-0.02
Proportion of blocking (β)	6.9%	7.1%	-0.02
AMHS utilization(ρ)	48.9%	45%	0.07

Table 3.4 Analytical and simulated expected times between empty vehicle interarrival times (sec.) n=3 vehicles λ=22 jobs/hr

Station	$T_{analytic}$	$T_{simulation}$	*Rel. error*
1	54.36	56.14	-0.03
2	60.20	61.65	-0.02
3	58.16	60.06	-0.03
4	60.25	61.64	-0.02
5	73.43	75.16	-0.02
Average abs. error			-0.03
Proportion of blocking (β)	15.8%	15.1%	0.01
AMHS utilization(ρ)	33.4%	33.1%	0.05

The analytical model performs reasonably well from a design perspective with acceptable error percentages. Even though the processing times are deterministic, which was expected to weaken the Poisson assumption, the analytic results are still quite close to those from the simulation model. We also performed the same type of analysis when the processors utilization is 90%, the results were still close to those from the simulation model with no increase in the error percentages.

3.10. Summary and future work

In this paper, we have presented a Discrete Time Markov Chain model that can be used in assessing closed-loop AMHS performance. The model is novel because it considers vehicle-blocking without the need to include detailed AMHS operations. Experimental comparisons of the model generated results with detailed simulation for a small example problem produced acceptable error margins.

There are a number of issues to be explored in further research. One issue is the size of the state space of the Markov Chain, which may pose computational challenges. Although the Markov Chain is highly structured and sparse, it may be possible to reduce the number of states by eliminating those states that are not needed to estimate the key performance measures. A critical issue in state space dimensionality is the number of "places" where a vehicle can be, and particularly, the number of "virtual' places. One potential amelioration of the dimensionality problem is to modify the state transition probabilities to overcome the need for virtual stations that were used to account for the mismatched transition times.

The simple loop structure of the presented model is an issue as recent enhancements to OHT systems in practice violate this assumption. We believe the model can be extended to cover more general network configurations. The extension will require the empty vehicles to be routed probabilistically. With the FEFS policy used in the current model, empty vehicles are not dispatched to the loads but simply travel around the loop until they encounter a waiting load. It will be interesting to explore the impact of these probabilities on the AMHS performance and how machines will be affected differently depending on their location in the network.

Simulation remains a key resource for AMHS designers and analysts, but they also need effective and fast models to use in phases of concept development, design, and analysis prior to investing in high-fidelity simulation studies. The model presented here is a significant step toward meeting that need.

CHAPTER 4

REDUCED MARKOV CHAIN MODEL OF VEHICLE-BASED CLOSED-LOOP AMHS

4.1. Introduction and motivation

Chapter three presented an extended discrete time Markov chain model that can be used in assessing closed-loop AMHS performance. Experimental comparisons of the model generated results with detailed simulation results for a small example problem produced acceptable error margins. However, because the model tracks the movement of every vehicle in the system, the number of states grows exponentially with the problem size, and the model poses computational challenges. For instance, the number of states for a simple system of 5 machines (10 loading and unloading stations) served by two vehicles has 405 states, and increases to 3,270 states with three vehicles and to 17,700 with four vehicles.

This chapter presents an approximation of the first model that trades off accuracy for providing quick solutions for the large-scale real-life applications of AMHS. We propose a different modeling approach that creates one Markov chain for each vehicle separately; this reduces the size of the Markov chain drastically, but creates a challenge for modeling vehicle-blocking.

4.2. Modeling approach

The approach proposed here follows similar logic to the one presented in Chapter three but differs in that the Markov chain tracks one vehicle while assuming that there are *n* vehicles operating in the AMHS but the move requests are equally distributed among these vehicles.

The transition matrix tracks the changes in the locations and the conditions of one vehicle. In the matrix, each distinct possible location-condition pair is identified as a state (in the previous model, a state is identified by location-condition-vehicle combination because every vehicle was tracked). The transition between the states is probabilistic and depends on the move requests rate, the number of vehicles, the sequence of the machines on the AMHS loop, and on the possibility of vehicle-blocking. Assumptions on the arrival process of move requests allow us to analyze the transition between the states as a Markov chain.

Similar to the previous model, some of the transition probabilities are known because they are easily calculated from the given problem parameters (such as the probability that a loaded vehicle will be dropping off its load at some station), while other transition probabilities are only partially determined by the problem parameters, but also influenced by the output variables that we are trying to calculate, specifically the arrival rate of empty-vehicles to stations. We also have a new set of transition probabilities that will determine whether the vehicle will get blocked in the next state or not. We did not need these in the previous model because the state specifies where every vehicle is located.

Next, we present necessary conditions that ensure that the AMHS is able to meet the required throughput imposed by the machines. These conditions provide constraints on some of the unknown variables in the Markov chain.

The blocking probabilities that are introduced in this model are estimated by assuming that the probability that a vehicle gets blocked increases linearly with the number of vehicles.

One advantage of this model is that there is no need to create the virtual stations if the travel times and loading times are not equal in order to synchronize the vehicles' movements. However, this implies that unlike the previous model, the transition time between each pair of states is not equal. Therefore, the next step in the new approach is to approximate the transition time between the states, which will depend on the transition probabilities and the given travel and service times.

Again, the resulting model is not a conventional Markov chain, and has additional sets of unknown variables that were not present in the previous model. However, the new model has a significantly smaller state-space, which does not grow with the number of vehicles in the system. We call this model the *reduced-state Extended Markov Chain model*. The reduced-state extended model provides a full rank system of equations that can be solved to give a unique set of estimates for the output parameters of the AMHS.

4.3. Additional notation

p_i^p : probability that a vehicle is blocked by a vehicle occupying pick-up station

$s_i^p , m_i \in M$.

p_i^d : probability that a vehicle is blocked by a vehicle occupying drop-off station

$s_i^d , m_i \in M$.

C_r : expected time between two consecutive visits to state *r*.

$E(t_b)$: expected time a vehicle stays blocked from travel.

T_r: time from the instant the system enters state *r* until the instant it enters the next state.

π_r : proportion of time that a vehicle spends in state *r*.

4.4. Additional mathematical assumptions

In addition to the assumptions in Section 3.6, the reduced extended Markov chain model assumes that:

1. Loads are distributed evenly among the vehicles.
2. The probability that a vehicle is blocked by the downstream station depends on the steady-state probability that the downstream station is occupied by a vehicle which increases linearly with the number of vehicles in the system (fleet size).

Assumption (1) allows us to separate the Markov chain analysis of each vehicle, while assumption (2) simplifies the modeling of vehicle-blocking.

4.5. The reduced state extended Markov chain model

Similar to the detailed model in Chapter three, we propose to consider only those points in time where a vehicle is located at a station. We characterize a state by specifying the condition and location of the vehicle, which is defined by three characters:

$(m_i, i = 1,2,..., M), (p, d), (e, f, b, k, s)$

The first and second characters describe, respectively, the machine occupied by the vehicle and the station type (pick-up or drop-off), and the third character specifies the condition of the vehicle (empty, loaded, blocked/empty, blocked/loaded, and receiving service).

4.5.1. Transition probabilities

The reduced state model has an additional set of transition probabilities associated with the event that the vehicle gets blocked by the downstream station blocking. p_{i+1}^{d} denotes the probability that the vehicle will be blocked by a vehicle at s_{i+1}^{d} and cannot move from s_i^p to s_{i+1}^{d}, and let p_i^p denotes the probability that the vehicle will be blocked by s_i^p and cannot move from s_i^d to s_i^p. There are also the same transition probabilities that were in the detailed model; q_i the probability that an empty vehicle arriving at a pick-up station s_i^p will find a load waiting, and r_i the probability that a loaded vehicle arriving at drop-off station s_i^d will drop off its load at s_i^d. Based on these probabilities, we define the transition probabilities in the transition matrix **R** given by,

$$
\begin{array}{l|ccccccccccccccccccccccccc}
 & 1df & 1dk & 1ds & 1de & 1db & 1pf & 1pk & 1ps & 1pe & 1pb & 2df & 2dk & 2ds & 2de & 2db & \cdots & mdk & mds & mde & mdb & mpf & mpk & mps & mpe & mdb \\
\hline
1df & & \bar{r}_1 p_1^p & r_1 & & & \bar{r}_1\bar{p}_1^p & & & & & & & & & & & & & & & & & & & \\
1dk & & & & & & 1 & & & & & & & & & & & & & & & & & & & \\
1ds & & & & & p_1^p & & & & \bar{p}_1^p & & & & & & & & & & & & & & & & \\
1de & & & & & p_1^p & & & & \bar{p}_1^p & & & & & & & & & & & & & & & & \\
1db & & & & & & & & & 1 & & & & & & & & & & & & & & & & \\
1pf & & & & & & & p_2^d & & & & \bar{p}_2^d & & & & & & & & & & & & & & \\
1pk & & & & & & & & & & & & & & & & & & & & & & & & & \\
1ps & & & & & & & p_2^d & & & & \bar{p}_2^d & & & & & & & & & & & & & & \\
1pe & & & & & & & & q_1 & & \bar{q}_1 p_2^d & & & & \bar{q}_1\bar{p}_2^d & & & & & & & & & & & \\
1pb & & & & & & & & & & & & & & 1 & & & & & & & & & & & \\
2df & & & & & & & & & & & & \bar{r}_2 p_2^p & r_2 & & & & & & & & & & & & \\
2dk & & & & & & & & & & & & & & & & & & & & & & & & & \\
2ds & & & & & & & & & & & & & & & p_2^p & & & & & & & & & & \\
2de & & & & & & & & & & & & & & & p_2^p & & & & & & & & & & \\
2db & & & & & & & & & & & & & & & & & & & & & & & & & \\
\vdots & & & & & & & & & & & & & & & & \cdots & & & & & & & & & \\
mdf & & & & & & & & & & & & & & & & & \bar{r}_m p_m^p & r_m & & & \bar{r}_m\bar{p}_m^p & & & & \\
mdk & & & & & & & & & & & & & & & & & & & & & 1 & & & & \\
mds & & & & & & & & & & & & & & & & & & & & p_m^p & & & & \bar{p}_m^p & \\
mde & & & & & & & & & & & & & & & & & & & & p_m^p & & & & \bar{p}_m^p & \\
mdb & & & & & & & & & & & & & & & & & & & & & & & & 1 & \\
mpf & \bar{p}_1^d & & & & & & & & & & & & & & & & & & & & & p_1^d & & & \\
mpk & 1 & & & & & & & & & & & & & & & & & & & & & p_1^d & & & \\
mps & \bar{p}_1^d & & & & & & & & & & & & & & & & & & & & & & & & \\
mpe & & & & \bar{q}_m\bar{p}_1^d & & & & & & & & & & & & & & & & & & & q_m & & \bar{q}_m p_1^d \\
mpb & & & & 1 & & & & & & & & & & & & & & & & & & & & &
\end{array}
$$

where $\bar{x} = 1 - x$.

The transition matrix **R** specifies the movement of the vehicle between the states. The position and type of the vehicle (*e*, *f*, *b*, *k*, or *s*), the possibility of a vehicle blocking its path, and the presence of a load to be picked-up or dropped-off, determine the system transitions. In Figure 4.1 below, consider the transitions from state (i,d,e), an empty vehicle arriving to a drop-off station s_i^d will move empty to the next pick-up station s_i^p, entering state (i,p,e) if it was not blocked by another vehicle, which happens with probability $\overline{p_i^p}$. With probability p_i^p, the vehicle gets blocked and transitions to state (i,d,b). From state (i,p,e), the empty vehicle will leave empty if there was no load waiting at s_i^p, which happens with probability $\overline{q_i}$, and also if there is no vehicle at s_{i+1}^d blocking its way, with probability $\overline{p_{i+1}^d}$, thus, the vehicle moves from state (i, p, e) to state $(i+1, d, e)$ with probability $\overline{q_i}\,\overline{p_{i+1}^d}$, and to the blocked state (i, p, b) with probability $\overline{q_i}\, p_{i+1}^d$. However, if the empty vehicle encounters a load at s_i^p, it moves to state (i, p, s) and starts the pick-up process, with probability q_i. Similarly, a loaded vehicle arriving to a drop-off station s_{i+1}^d, a state denoted by $(i+1, d, f)$, will drop-off its load, thus entering state $(i+1, d, s)$ with probability r_{i+1}, or move to states $(i+1, p, f)$ and $(i+1, d, k)$ with probabilities $\overline{r_i}\,\overline{p_{i+1}^p}$ and $\overline{r_i}\, p_{i+1}^p$, respectively.

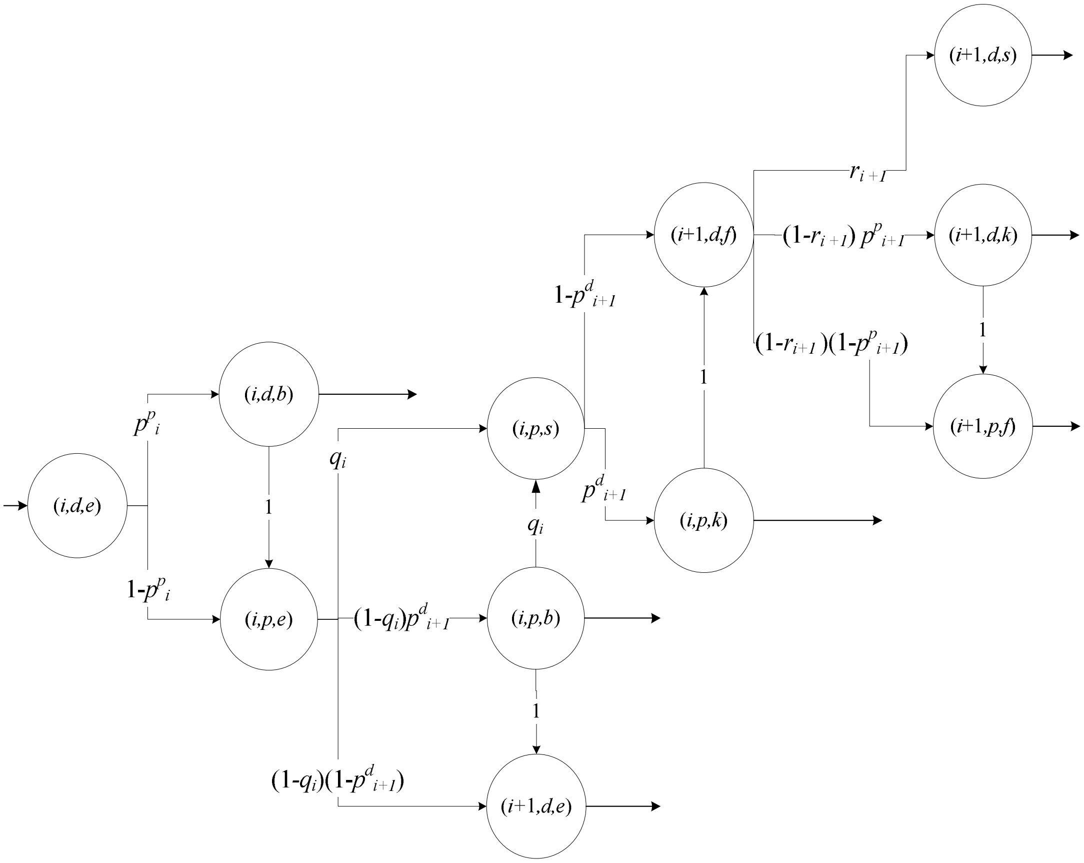


Figure 4.1 Part of the state transition diagram for the reduced-state extended Markov chain

An empty vehicle arriving to a drop-off station will move to the next pick-up station provided that there was no other vehicle blocking its way, and so is the case for a loaded vehicle arriving to a pick-up station. We assume that after a vehicle is blocked, it gets unblocked and moves to the downstream station with probability 1. The justification for this assumption is that a vehicle gets blocked because somewhere downstream another vehicle is in service, this vehicle can be either the one directly downstream from the blocked vehicle or several stations downstream given that there are vehicles occupying all the station in between. The vehicle does not get blocked twice at the same station because the vehicle that was in service has finished its job and will move, unblocking all the vehicles that were waiting behind it.

4.5.2. The reduced Markov chain steady-state analysis

Let C_r be the expected time between two consecutive visits to state *r*, *r*=1,…,|R|. Without loss of generality, assume that C is the expected time between two visits to some reference state, say state $(1, p, e)$. Let υ ={ υ_r} *r*=1, …, |R|, where υ_r denote the visit ratio to state *r*, which is the number of times the system visits state *r* between two successive visits to the reference state $(1, p, e)$, by this definition, $\upsilon_{(1,p,e)} = 1$. For a finite state, positive recurrent Markov Chain, the visit ratios can be uniquely obtained by solving the square system of equations (Ross, 2000):

$$\mathbf{R}\boldsymbol{\upsilon} = \boldsymbol{\upsilon} \tag{4.1}$$

$$\upsilon_{(1,p,e)} = 1 \tag{4.2}$$

The elements in the transition matrix are the transition probabilities between states, discussed earlier in Section 4.5.1. Some of these probabilities are unknown, specifically, the load-encountering probabilities $\mathbf{q} = \{q_i\}, i = 1,...,M$, and the blocking probabilities $\mathbf{p^d} = \{p_i^d\}, \mathbf{p}^p = \{p_i^p\}, i = 1,...,M$.

From equations (4.1), and (4.2), we get:

$$\upsilon_{(i,d,f)} = \upsilon_{(i-1,p,k)} + \overline{p_i^d}(\upsilon_{(i-1,p,f)} + \upsilon_{(i-1,p,s)}) \tag{4.3a}$$

$$\upsilon_{(i,d,k)} = p_i^p \overline{r_i} \upsilon_{(i,d,f)} \tag{4.3b}$$

$$\upsilon_{(i,d,s)} = r_i \upsilon_{(i,d,f)} \tag{4.3c}$$

$$\upsilon_{(i,d,e)} = \upsilon_{(i-1,p,b)} + \overline{q_{i-1}}\,\overline{p_i^d}\, \upsilon_{(i-1,p,e)} \tag{4.3d}$$

$$\upsilon_{(i,d,b)} = p_i^p(\upsilon_{(i,d,s)} + \upsilon_{(i,d,e)}) \tag{4.3e}$$

$$\upsilon_{(i,p,f)} = \upsilon_{(i,d,k)} + \overline{p_i^p}\,\overline{r_i}\,\upsilon_{(i,d,f)} \tag{4.4a}$$

$$\upsilon_{(i,p,k)} = p_{i+1}^d(\upsilon_{(i,p,f)} + \upsilon_{(i,p,s)}) \tag{4.4b}$$

$$\upsilon_{(i,p,s)} = q_i\,\upsilon_{(i,p,e)} \tag{4.4c}$$

$$\upsilon_{(i,p,e)} = \upsilon_{(i,d,b)} + \overline{p_i^p}(\upsilon_{(i,d,s)} + \upsilon_{(i,d,e)}) \tag{4.4d}$$

$$\upsilon_{(i,p,b)} = \overline{q_i}\,p_{i+1}^d\,\upsilon_{(i,p,e)} \tag{4.4e}$$

Combining (4.3d), (4.3e), and (4.4d), (4.4e), we get the following expression for $\upsilon_{(i,p,e)}$,

$$\upsilon_{(i,p,e)} = \upsilon_{(i,d,s)} + \overline{q_{i-1}}\,\upsilon_{(i-1,p,e)} \tag{4.5}$$

Note that in (4.3) and (4,4), the load-drop off probabilities at drop-off stations, $\mathbf{r} = \{r_i\}, i = 1,...,M$ are obtained from the move rates as was demonstrated in Section 3.7.3 using the expression:

$$r_i = \frac{\lambda_i}{\alpha^d}, \forall i \in M \tag{4.6}$$

4.5.3. Stability conditions

For a stable system, the expected number of loads delivered by a single vehicle to drop-off station s_i^d in a cycle of length C must equal the number of times a single vehicle enters state (i,d,s) (unloads at s_i^d) in the same period. Let n denote the number of vehicles on the loop, then:

$$\upsilon_{(i,d,s)} = \frac{\lambda_i C}{n}, \forall i \in M \tag{4.7}$$

During a cycle of length *C*, each vehicle makes $v_{(i,p,e)}$ empty trips to s_i^p. The vehicle picks up a load with probability q_i, and moves to s_{i+1}^d if it does not get blocked with probability p_{i+1}^d. Therefore, during a cycle, the expected number of loads picked up from s_i^p by each vehicle is $q_i v_{(i,p,e)}$. For a stable system, this should equal $C\lambda_i / n$; the number of move requests per vehicle in a time period of length *C*. Equating these terms we get the following necessary conditions for AMHS stability:

$$q_i = \frac{\lambda_i C}{v_{(i,p,e)} n} \quad \forall i \in M \tag{4.8}$$

From (4.7) and (4.8), we get:

$$v_{(i,d,s)} = q_i v_{(i,p,e)} \tag{4.9}$$

Substituting (4.9) in (4.5), we get:

$$v_{(i,p,e)} = \frac{\overline{q}_{i-1}}{\overline{q}_i} v_{(i-1,p,e)} \tag{4.10}$$

Continuing to express $v_{(i,p,e)}$ in terms of $v_{(i-1,p,e)}, v_{(i-2,p,e)} \cdots, v_{(1,p,e)}$, we get

$$v_{(i,p,e)} = \frac{\overline{q}_1}{\overline{q}_i} v_{(1,p,e)} \tag{4.11}$$

From (4.2), we have that $v_{(1,p,e)} = 1$ and thus,

$$v_{(i,p,e)} = \frac{\overline{q}_1}{\overline{q}_i}, \forall i \in M \tag{4.12}$$

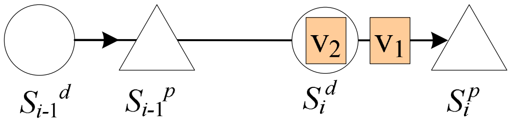
Substituting (4.12) in (4.8), and after some algebra we get:

$$q_i = \frac{\lambda_i C}{\overline{q}_1 n + \lambda_i C} \quad \forall i \in M \tag{4.13}$$

4.5.4. Vehicle blocking probabilities

A vehicle is blocked when it attempts to move to the downstream station but finds that station occupied by another vehicle. The downstream vehicle could be receiving service, traveling towards that station or is also blocked.

Proposition 4-1: if the distance between the drop-off station and pick-up station of a machine is less than the distance between two machines, a vehicle at a drop-off station cannot get blocked by a vehicle traveling towards the downstream machine's pick-up station.



Proof: consider the Figure above, assume that it is possible for *v₂* to be blocked at s_i^d by v_1 that is traveling to s_i^p. For this case to happen, the moment v_1 started leaving s_i^d, v_2 must have been traveling towards s_i^d on the segment between s_{i-1}^p and s_i^d, but this contradicts the initial control mechanism that does not allow a vehicle to travel towards a load port occupied by another vehicle. ■

Proposition 4-1 implies that a vehicle cannot be blocked at a drop-off station by a vehicle traveling towards the downstream pick-up station. The opposite, however, could happen, i.e. a vehicle at a pick-up station can get blocked by a vehicle traveling towards the downstream drop-off station.

Let $\upsilon_{(i,d)}$ and $\upsilon_{(i,p)}$ denote the visit ratios to the input buffer and the output buffer of machine *i*, respectively. By this definition:

$$\upsilon_{(i,d)} = \upsilon_{(i,d,f)} + \upsilon_{(i,d,k)} + \upsilon_{(i,d,s)} + \upsilon_{(i,d,e)} + \upsilon_{(i,d,b)} \qquad (4.14a)$$

$$v_{(i,p)} = v_{(i,p,f)} + v_{(i,p,k)} + v_{(i,p,s)} + v_{(i,p,e)} + v_{(i,p,b)} \tag{4.14b}$$

From (4.3), (4.4), (4.6), and (4.12), while keeping in mind that the arrival rate of loaded vehicles is equal for every drop-off station, and after some algebra we can express visit ratios as:

$$v_{(i,d,f)} = \frac{q_1}{r_1} \tag{4.15a}$$

$$v_{(i,d,k)} = \frac{q_1}{r_1}\overline{r}_i p_i^p \tag{4.15b}$$

$$v_{(i,d,s)} = \frac{q_i}{\overline{q}_i}\overline{q}_1 \tag{4.15c}$$

$$v_{(i,d,e)} = \overline{q}_1 \tag{4.15d}$$

$$v_{(i,d,b)} = \frac{p_i^p}{\overline{q}_i}\overline{q}_1 \tag{4.15e}$$

$$v_{(i,p,f)} = \frac{q_1}{r_1}\overline{r}_i \tag{4.16a}$$

$$v_{(i,p,k)} = (\frac{q_1}{r_1}\overline{r}_i + \frac{q_i}{\overline{q}_i}\overline{q}_1) p_{i+1}^d \tag{4.16b}$$

$$v_{(i,p,s)} = \frac{q_i}{\overline{q}_i}\overline{q}_1 \tag{4.16c}$$

$$v_{(i,p,e)} = \frac{\overline{q}_1}{\overline{q}_i} \tag{4.16d}$$

$$v_{(i,p,b)} = \overline{q}_1 p_{i+1}^d \tag{4.16e}$$

In expressions (4.15) and (4.16) we assume, without loss of generality, that the first machine in the sequence of machines 1,2,…,*M*, must receive loads (i.e. $r_1 > 0$). We can now express $\upsilon_{(i,d)}$ and $\upsilon_{(i,p)}$ as:

$$\upsilon_{(i,d)} = (1+\bar{r}_i p_i^p)\frac{q_1}{r_1} + (1+p_i^p)\frac{\overline{q_1}}{\overline{q_i}} \tag{4.17a}$$

$$\upsilon_{(i,d)} = (1+p_{i+1}^d)\frac{q_1}{r_1}\bar{r}_i + (1+q_i+p_{i+1}^d)\frac{\overline{q_1}}{\overline{q_i}} \tag{4.17b}$$

Recall that p_i^d is the steady-state probability that the vehicle is blocked and cannot move to s_i^d, and p_i^p denote the steady-state probability that the vehicle is blocked and cannot move to s_i^p. A vehicle is blocked by drop-off station s_i^d when there is another vehicle traveling to, receiving service, or also blocked at s_i^d. A vehicle is blocked by pick-up station s_i^p when there is another vehicle receiving service, or also blocked at s_i^p. We also assumed in Section 4.4 that the blocking probabilities increase linearly with the fleet size (*n*), since a vehicle gets blocked if any of the *n*-1 vehicles is occupying the downstream station, we can therefore estimate p_i^d and p_i^p from:

$$p_i^d = (n-1)\frac{\upsilon_{(i,d)}}{\sum_{\forall j \in M} \upsilon_{(j,d)} + \upsilon_{(j,p)}}, \forall s_i^d \tag{4.18a}$$

$$p_i^p = (n-1)\frac{\upsilon_{(i,p,s)} + \upsilon_{(i,p,k)} + \upsilon_{(i,p,b)}}{\sum_{\forall j \in M} \upsilon_{(j,d)} + \upsilon_{(j,p)}}, \forall s_i^p \tag{4.18b}$$

From (4.17) and (4.18)

$$p_i^d = (n-1)\frac{(1+\overline{r}_i p_i^p)\frac{q_1}{r_1} + (1+p_i^p)\frac{\overline{q_1}}{\overline{q_i}}}{\sum_{\forall j \in M}(1+\overline{r_j}(p_{j+1}^d + p_j^p) + \overline{r_j})\frac{q_1}{r_1} + (2+q_j+p_j^p+p_{j+1}^d)\frac{\overline{q_1}}{\overline{q_j}}} \tag{4.19a}$$

$$p_i^p = (n-1)\frac{\overline{r}_i p_{i+1}^d \frac{q_1}{r_1} + (q_i + p_{i+1}^d)\frac{\overline{q_1}}{\overline{q_i}}}{\sum_{\forall j \in M}(1+\overline{r_j}(p_{j+1}^d + p_j^p) + \overline{r_j})\frac{q_1}{r_1} + (2+q_j+p_j^p+p_{j+1}^d)\frac{\overline{q_1}}{\overline{q_j}}} \tag{4.19b}$$

The total number of unknowns we have is 4|*M*|+1:

- |*M*| unknowns for the visit ratios, $\mathbf{\upsilon_{(p,e)}} = \{\upsilon_{(i,p,e)}\}, i = 1,2,..., M$.
- |*M*| unknowns for the load-encountering probabilities at pick-up stations: $\mathbf{q} = \{q_i\}, i = 1,..., M$,
- |*M*| unknowns for the vehicle-blocking probabilities by drop-off stations: $\mathbf{p^d} = \{p_i^d\}, i = 1,..., M$.
- |*M*| unknowns for the vehicle-blocking probabilities by pick-up stations: $\mathbf{p}^p = \{p_i^p\}, i = 1,..., M$.
- One unknown for the cycle length *C*.

We still need another equation that links *C* with the visit ratios as will be shown in Section 4.5.5.

4.5.5. Transition times

Because this model tracks the movement of a single vehicle, it has an advantage over the earlier model developed in Chapter Three, because there is no need for the virtual stations in order to synchronize the vehicles' movement, when the travel times and loading times are not equal. However, this implies that unlike the previous model, the transition times between pairs of states are not equal. In this section, we develop expressions for the expected state transition time, which is the expected time from the instant the vehicle enters state *r* until the instant it enters the next state.

A discussion on the expected blocking times is necessary before we derive the expressions for the transition times. Let t_b denote the time that a vehicle is blocked, which is the time the vehicle blocking its ways needs before it finishes service (loading/unloading) or travel.

The time that a vehicle spends while blocked depends on the percentage of time the vehicle is traveling and the expected time the vehicle spends loading/unloading. As the number of vehicles increase, the proportion of the blocking time due to loading/unloading operations is less significant and is mostly caused by the traveling or other blocked vehicles. Therefore, we develop the expression to estimate the average blocking time that depends on the expected visit rate to each state in the Markov chain.

We should also consider that when a vehicle becomes blocked, the time it takes for the vehicle blocking its way to finish its service is not necessarily the entire service time. In fact, the average *remaining* service time, $E(S_r)$, of a vehicle as seen by a randomly arriving vehicle (Kleinrock, 1975) is:

$$E(S_r) = \frac{E(S)(C_s^2 + 1)}{2}$$

Where *E*(*S*) is the service time and C_s^2 is the coefficient of variation of service time. It is reasonable to assume that loading/unloading and travel times are all deterministic, $C_s^2 = 0$ and thus,

$$E(S_r) = \frac{E(S)}{2}$$

Taking this into consideration in addition to the earlier discussion of the blocking time, we can use the following expression to estimate the expected blocking time

$$E(t_b) = \frac{1}{2}.\left(\sum_{\forall i \in M} \frac{v_{(i,p,s)}(l - t_i)}{\sum\limits_{\forall i \in M} v_{(i,p,s)}} + \sum_{\forall i \in M} \frac{(v_{(i,d,e)} + v_{(i,d,f)})(t_{i-1,i} - t_i) + v_{(i,d,s)} l}{\sum\limits_{\forall i \in M} v_{(i,d,e)} + v_{(i,d,f)} + v_{(i,d,s)}} \right) \qquad (4.20)$$

where t_i is the travel time from the drop-off station to the pick-up station of machine *i*, $t_{i,i+1}$ is the travel time from the pick-up station of machine *i* to the drop-off station of machine *i*+1, and *l* is the loading/unloading time. The first term in expression (4.20) is the average blocking time caused by pick-up stations, and the second term is the average blocking time caused by drop-off stations. Our estimate for the average blocking time assumes that:

1. The expected blocking time is aggregated in one expression rather than separating the blocking time for each station; this is justified because neither the location nor the type of service is known for the vehicle that is causing the blocking. Recall that when a vehicle is blocked, the blocking is not necessarily caused by the vehicle occupying the immediate downstream station.
2. Loading/unloading time are larger than travel times from the drop-off to the pick-up station within the same machine, if this was not true, then a vehicle cannot get blocked by a loading/unloading vehicle.

From expressions (4.15)-(4.17), $E(t_b)$ can be written as

$$E(t_b) = \frac{1}{2} \sum_{\forall i \in M} \left(\frac{(l - t_i)\frac{q_i}{\overline{q}_i}}{\sum\limits_{\forall i \in M} \frac{q_i}{\overline{q}_i}} + \frac{(\frac{q_1}{r_1} + \overline{q}_i)(t_{i-1,i} - t_i) + \frac{q_i}{\overline{q}_i}\overline{q}_1 l}{\sum\limits_{\forall i \in M} \frac{q_1}{r_1} + \overline{q}_i + \frac{q_i}{\overline{q}_i}\overline{q}_1} \right) \qquad (4.21)$$

We can now develop the expressions for the state transition times, let T_r denote the time from the instant the vehicle enters state *r* until the instant it enters the next state. We can develop an expression for the cycle length *C* by considering the transition time from one state to the next. *C* was defined as the expected time between two successive visits to the reference state $(1, p, e)$, and can thus be obtained from:

$$C = \sum_{\forall r \in R} \upsilon_r E(T_r) \tag{4.22}$$

The terms $E(T_r)$, $r \in R$ can be determined based on the transition probabilities. For instance, consider state (i, p, e) ,which is an empty vehicle arriving at some pickup station s_i^p, the time that the vehicle spends in this state depends on the probability of encountering a load at s_i^p and on the probability of being blocked at s_i^p by s_{i+1}^d. Thus with probability q_i, the vehicle will find a load and the next state is (i, p, s); the transition time in this case is the loading time *l.* If the vehicle does not encounter a load but is blocked from moving to s_{i+1}^d, from the discussion above, we estimate this transition time to be $E(t_b)$, and this happens with probability $\overline{q_i}\, p_{i+1}^d$. However, with probability $\overline{q_i}\,\overline{p_{i+1}^d}$, the vehicle does not encounter a load and is not blocked from moving; the transition time in this case is the time to travel from s_i^p to s_{i+1}^d, denoted by $t_{i,i+1}$, thus the expected time spent in state (i, p, e) is:

$$E\left(T_{(i,p,e)}\right) = q_i l + \overline{q_i}\, p_{i+1}^d.E(t_b) + \overline{q_i}\,\overline{p_{i+1}^d}\, t_{i,i+1} \tag{4.23a}$$

Using the same logic, we can derive the expected transition time for the other possible states of a vehicle at s_i^p:

$$E\left(T_{(i,p,f)}\right) = p_{i+1}^d E(t_b) + \overline{p_{i+1}^d}\, t_{i,i+1} \tag{4.23b}$$

$$E\left(T_{(i,p,k)}\right) = t_{i,i+1} \tag{4.23c}$$

$$E\left(T_{(i,p,b)}\right) = t_{i,i+1} \tag{4.23d}$$

$$E\left(T_{(i,p,s)}\right) = p_{i+1}^{d} E(t_b) + \overline{p_{i+1}^{d}} t_{i,i+1} \tag{4.23e}$$

For a vehicle at s_i^d, we can also derive the expected transition times as follows:

$$E\left(T_{(i,d,e)}\right) = p_i^p E(t_b) + \overline{p_i^p} t_i \tag{4.24a}$$

$$E\left(T_{(i,d,f)}\right) = r_i l + p_i^p \overline{r_i} E(t_b) + \overline{r_i}\,\overline{p_i^p} t_i \tag{4.24b}$$

$$E\left(T_{(i,d,k)}\right) = t_i \tag{4.24c}$$

$$E\left(T_{(i,d,b)}\right) = t_i \tag{4.24d}$$

$$E\left(T_{(i,d,s)}\right) = p_i^p E(t_b) + \overline{p_i^p} t_i \tag{4.24e}$$

From (4.12), (4.15), (4.16), (4.22), (4.23), and (4.24), we have

$$C = \sum_{j=1}^{m} \left(\frac{\overline{r_j}}{r_1} \frac{q_1}{\overline{q_1}} + \frac{q_i}{\overline{q_i}} + 1 \right) . \left(E(t_b)(p_j^p + p_{j+1}^d) + t_j + t_{j-1,j} \right) + \left(\frac{r_j}{r_1} \frac{q_1}{\overline{q_1}} + \frac{q_i}{\overline{q_i}} \right) l \tag{4.25}$$

Combining equation sets (4.13), and (4.20) with equations (4.21) and (4.25), we have 3|*M*|+2 equations and 3|*M*|+2 unknowns, and we can find the solution to the system of equations and calculate the visit ratio to every state, and the blocking probabilities.

Proposition 4-2: Consider the system of nonlinear equations:

$$q_i = \frac{\lambda_i C}{\overline{q_1} n + \lambda_i C}, \forall i \in M$$

$$p_i^d = (n-1) \frac{(1 + \overline{r_i} p_i^p) \dfrac{q_1}{r_1} + (1 + p_i^p) \dfrac{\overline{q_1}}{\overline{q_i}}}{\sum\limits_{\forall j \in M} (1 + \overline{r_j} (p_{j+1}^d + p_j^p) + \overline{r_j}) \dfrac{q_1}{r_1} + (2 + q_j + p_j^p + p_{j+1}^d) \dfrac{\overline{q_1}}{\overline{q_j}}}$$

$$p_i^p = (n-1)\frac{\overline{r}_i p_{i+1}^d \frac{q_1}{r_1} + (q_i + p_{i+1}^d)\frac{\overline{q_1}}{\overline{q_i}}}{\sum_{\forall j \in M} (1+\overline{r}_j(p_{j+1}^d + p_j^p)+\overline{r}_j)\frac{q_1}{r_1} + (2+q_j+p_j^p+p_{j+1}^d)\frac{\overline{q_1}}{\overline{q_j}}}$$

$$E(t_b) = \frac{1}{2}\sum_{\forall i \in M}\left(\frac{(l-t_i)\frac{q_i}{\overline{q_i}}}{\sum_{\forall i \in M}\frac{q_i}{\overline{q_i}}} + \frac{(\frac{q_1}{r_1}+\overline{q}_i)(t_{i-1,i}-t_i)+\frac{q_i}{\overline{q_i}}\overline{q}_1 l}{\sum_{\forall i \in M}\frac{q_1}{r_1}+\overline{q}_i+\frac{q_i}{\overline{q_i}}\overline{q}_1} \right)$$

$$C = \sum_{j=1}^{m}\left(\frac{\overline{r_j}}{r_1}\frac{q_1}{\overline{q_1}}+\frac{q_i}{\overline{q_i}}+1\right).\left(E(t_b)(p_j^p+p_{j+1}^d)+t_j+t_{j-1,j}\right)+\left(\frac{r_j}{r_1}\frac{q_1}{\overline{q_1}}+\frac{q_i}{\overline{q_i}}\right)l$$

If the AMHS is stable (i.e. the AMHS can handle all the move requests within the planning horizon), there exists a unique solution for the load-encountering probabilities at pick-up stations **q**, the vehicle-blocking probabilities by drop-off stations, $\mathbf{p}^{\mathbf{d}}$, the vehicle-blocking probabilities by pick-up stations, $\mathbf{p}^p$, the expected blocking time $E(t_b)$, and the cycle length *C*, and this solution provides the unique solution to the steady-state visit ratios, $\boldsymbol{\upsilon}$ to states of the Markov chain **R**.

Proof

The outline of the proof is as follows:

1. If the unknown transition probabilities in the transition matrix are known, then there exists a solution to the stationary visit ratios to the Markov chain states, **v**, assuming that the AMHS is stable; this follows from the fact that the Markov chain is finite, irreducible, and positive recurrent.
2. We then prove by contradiction that for every set of unknowns in the system of equations we solve, there cannot be two different vectors that will generate the same

solution for **v**. The unknowns are the set of load-encountering probabilities **q**, the set of vehicle-blocking probabilities by drop-off stations, $\mathbf{p}^{\mathbf{d}}$, the set of vehicle-blocking probabilities by pick-up stations, $\mathbf{p}^{p}$, the expected blocking time $E(t_b)$, and the cycle length *C*.

This is accomplished in the following sequence:

- Starting with the cycle length *C*, we show that there cannot be two different values for *C* that will generate the same solution for **v**.
- We then show that there cannot be two sets of the load encountering probabilities **q** that will generate the same cycle length *C*. We combine this result with the previous one we draw the conclusion that two different vectors of **q** cannot generate the same **v**.
- Similarly, we show that there cannot be two sets of the blocking probabilities that will generate the same set of the load encountering probabilities **q**.
- Finally, we show that since the expected blocking time $E(t_b)$ is a function of the load encountering probabilities, and since we have already established that the load encountering probabilities have to be unique, we assert that $E(t_b)$ also must be unique.

Formal proof:

If the load-encountering probabilities **q**, the vehicle-blocking probabilities by drop-off stations, $\mathbf{p}^{\mathbf{d}}$, the vehicle-blocking probabilities by pick-up stations, $\mathbf{p}^{p}$, the expected blocking time $E(t_b)$, and the cycle length *C* are known, and if the AMHS is stable, the Markov chain with transition matrix **R** is ergodic because it is finite and recurrent (Ross, 2000). For an ergodic Markov chain, there exists a unique solution to the steady-state visit ratios by solving the system of equations:

$$R\upsilon = \upsilon$$
$$\upsilon_1 = 1$$

Since the transition matrix **R** is a function of $\mathbf{q}, \mathbf{p^d}, \mathbf{p}^p$, $E(t_b)$ and *C*, we need to prove that there exists only one $\mathbf{q}, \mathbf{p^d}, \mathbf{p}^p$, $E(t_b)$ and *C*, for a given instance of the problem, where an instance is defined by *L*(*n*), the directed loop with *n* vehicles, *M* the set of machines in *L*(*n*), $\boldsymbol{\lambda}$ the arrival rate of move requests, and **T**, the travel time matrix, and *l* the loading/unloading times. We now prove, by contradiction, that there is only one solution for $\boldsymbol{\upsilon}$, $\mathbf{q}, \mathbf{p^d}, \mathbf{p}^p$, $E(t_b)$ and *C*.

a) Suppose there are two scalars $C^1 \text{and} C^2$, and two vectors $\boldsymbol{\upsilon}^\mathbf{1}$ and $\boldsymbol{\upsilon}^\mathbf{2}$ such that $C^1 \neq C^2, \upsilon^\mathbf{1} = \upsilon^\mathbf{2}$, and both solutions satisfy the system of equations in the Proposition.

   From expression (4.7):

$$\upsilon_{(i,d,s)} = \frac{\lambda_i C}{n}, \forall i \in M \ ,$$

   If $C^1 \neq C^2$, then $\upsilon^1_{(i,d,s)} \neq \upsilon^2_{(i,d,s)}, \forall i \in M$, which contradicts the initial assumption that there are two solutions such that $C^1 \neq C^2, \upsilon^\mathbf{1} = \upsilon^\mathbf{2}$.

b) Suppose there are two vectors $\mathbf{q}^\mathbf{1}$ and $\mathbf{q}^\mathbf{2}$, and two vectors $\boldsymbol{\upsilon}^\mathbf{1}$ and $\boldsymbol{\upsilon}^\mathbf{2}$ such that $\mathbf{q}^\mathbf{1} \neq \mathbf{q}^\mathbf{2}, \upsilon^\mathbf{1} = \upsilon^\mathbf{2}$, and both solutions satisfy the system of equations in the Proposition.

   From expression (4.8), the load encountering probabilities are obtained from:

$$q_i = \frac{\lambda_i C}{\upsilon_{(i,p,e)} n}, \forall i \in M$$

Since we already established that $C^1 = C^2$, then if $\mathbf{q}^1 \neq \mathbf{q}^2$, there is at least one machine $i \in M$, such that $\upsilon^1{}_{(i,p,e)} \neq \upsilon^2{}_{(i,p,e)}$, which contradicts the initial assumption that there are two solutions such that $\mathbf{q}^1 \neq \mathbf{q}^2$ and $\upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$.

c) Suppose there are two vectors $\mathbf{p}^{p^1}$ and $\mathbf{p}^{p^2}$, and two vectors $\boldsymbol{\upsilon}^{\mathbf{1}}$ and $\boldsymbol{\upsilon}^{\mathbf{2}}$ such that $\mathbf{p}^{p^1} \neq \mathbf{p}^{p^2}, \upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$, and both solutions satisfy the system of equations in the Proposition.

We already established that for $\upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$, we must have $C^1 = C^2$ and $\mathbf{q}^1 = \mathbf{q}^2$. If $\mathbf{q}^1 = \mathbf{q}^2$, then from expression (4.12):

$\upsilon_{(i,p,e)} = \frac{\overline{q}_1}{\overline{q}_i}$, we must have $\upsilon^1{}_{(i,p,e)} = \upsilon^2{}_{(i,p,e)}, \forall i \in M$, combining this result along with expression (4.15e):

$$\upsilon_{(i,d,b)} = \frac{p_i^p}{\overline{q}_i} \overline{q}_1 ,$$

we conclude that if $\mathbf{p}^{p^1} \neq \mathbf{p}^{p^2}$, there is at least one machine $i \in M$, such that $\upsilon^1{}_{(i,d,b)} \neq \upsilon^2{}_{(i,d,b)}$, which contradicts the initial assumption that there are two solutions such that $\mathbf{p}^{p^1} \neq \mathbf{p}^{p^2}$ and $\upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$.

d) Similarly, suppose there are two vectors $\mathbf{p}^{d^1}$ and $\mathbf{p}^{d^2}$, and two vectors $\boldsymbol{\upsilon}^{\mathbf{1}}$ and $\boldsymbol{\upsilon}^{\mathbf{2}}$ such that $\mathbf{p}^{d^1} \neq \mathbf{p}^{d^2}, \upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$, and both solutions satisfy the system of equations in the Proposition.

We already established that for $\upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$, we must have $C^1 = C^2$ and $\mathbf{q^1} = \mathbf{q^2}$. If $\mathbf{q^1} = \mathbf{q^2}$, then from expression (4.12):

$\upsilon_{(i,p,e)} = \frac{\overline{q_1}}{\overline{q_i}}$, we must have $\upsilon^1{}_{(i,p,e)} = \upsilon^2{}_{(i,p,e)}, \forall i \in M$, combining this result along with expression (4.16b):

$$\upsilon_{(i,p,k)} = (\frac{q_1}{r_1}\overline{r_i} + \frac{q_i}{\overline{q_i}}\overline{q_1})p^d_{i+1}$$

We conclude that if $\mathbf{p}^{d^1} \neq \mathbf{p}^{d^2}$, there is at least one machine $i \in M$, such that $\upsilon^1{}_{(i,p,k)} \neq \upsilon^2{}_{(i,p,k)}$, which contradicts the initial assumption that there are two solutions such that $\mathbf{p}^{d^1} \neq \mathbf{p}^{d^2}$ and $\upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$.

e) We already established that for $\upsilon^{\mathbf{1}} = \upsilon^{\mathbf{2}}$, we must have $C^1 = C^2$, $\mathbf{q^1} = \mathbf{q^2}$, then from expression (4.21):

$$E(t_b) = \frac{1}{2}\sum_{\forall i \in M}\left(\frac{(l - t_i)\frac{q_i}{\overline{q_i}}}{\sum_{\forall i \in M}\frac{q_i}{\overline{q_i}}} + \frac{(\frac{q_1}{r_1} + \overline{q_i})(t_{i-1,i} - t_i) + \frac{q_i}{\overline{q_i}}\overline{q_1}l}{\sum_{\forall i \in M}\frac{q_1}{r_1} + \overline{q_i} + \frac{q_i}{\overline{q_i}}\overline{q_1}}\right)$$

we must have $E(t_b)^1 = E(t_b)^2$.

From (a), (b), (c), (d), and (e), we conclude that there cannot exist two solutions that satisfy the set of equations in the proposition such that

$$C^1 \neq C^2, \, or \; \mathbf{q^1} \neq \mathbf{q^2}, \, or \; \mathbf{p}^{p^1} \neq \mathbf{p}^{p^2}, \, or \; \mathbf{p}^{d^1} \neq \mathbf{p}^{d^2} \text{ or } E(t_b)^1 \neq E(t_b)^2 \, and \; \boldsymbol{\upsilon^1} = \boldsymbol{\upsilon^2} \blacksquare$$

We have proved that if the AMHS is stable, and a solution to the system in the proposition exists, then this solution is unique. We defined stability as the ability of the vehicles to handle all the move requests within the planning horizon. We use enumeration of the number of vehicles (*n*) in order to find the solution. Let *B* denote the total number of vehicle buffers (locations vehicles can occupy). The maximum number of vehicles on the loop, n_{max}, has to satisfy:

$$n_{\max} \leq B-1 \tag{4.26}$$

otherwise, there would be a deadlock situation. We use the following algorithm to find the minimum number of vehicles for a stable AMHS:

1. Set the number of vehicles *n*=1.
2. Solve the system of equations in proposition 1:
   - If a solution exists, it is unique, exit the algorithm.
   - If a solution does not exist:
     - If *n*=*B*-1, a solution does not exist, the AMHS cannot satisfy the move requirements, regardless of fleet size.
     - If *n*<*B*-1, then set *n*=*n*+1, go back to step 2.

4.5.6. Steady-state probabilities of the Markov chain

The previous analysis provides us with the relative frequency with which each state occurs in the embedded Markov chain. We can now find the proportion of time that a vehicle spends in each state. We defined υ_r as the relative frequency with which state *r* is visited, let π_r denote the proportion of time that a vehicle spends in state *r*, which can be obtained from:

$$\pi_r = \frac{\upsilon_r E(T_r)}{\sum_k \upsilon_k E(T_k)} \tag{4.27}$$

4.6. Numerical example

We use two layouts, L_1 and L_2, to compare the analytical model estimates of the average time between empty-vehicle arrivals to pick-up stations to values obtained from discrete-event simulation. Layout 1 (L_1) in Figure 4.2, is a bay that has one stocker (m_1) and 7 process tools (m_2 through m_8). Layout 2 (L_2) (Figure 4.3) has one stocker (m_1) and 14 process tools (m_2 through m_{15}). Both L_1 and L_2 serve five products (p_a, p_b, p_c, p_d, p_e). The total arrival rate to the stocker is $\lambda = \lambda_a + \lambda_b + \lambda_c + \lambda_d + \lambda_e$ jobs per minute. Each job type is released and processed according to the routes given in Table 4.1 and Table 4.2 for L_1 and L_2, respectively. The resulting from-to matrices are presented in Table 4.3 and Table 4.4.

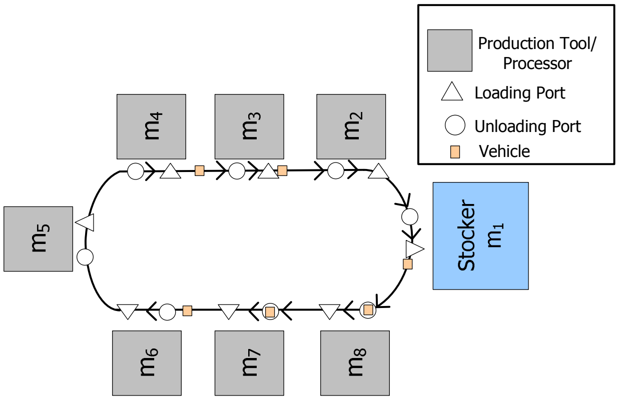


Figure 4.2 L1: An 8-machine example

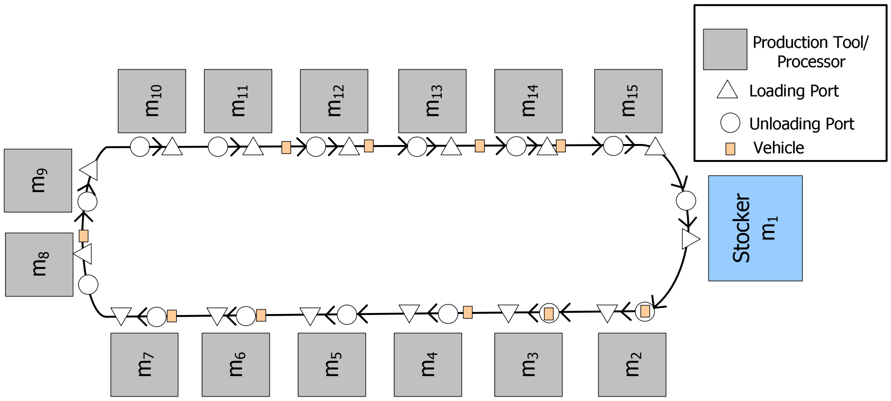


Figure 4.3 L2: A 15-machine example

Table 4.1 Arrival rate and routing of products for L1

Product Type	Routing	Jobs/Hour
p_a	1-2-3-4-1	12
p_b	1-3-5-7-1	6
p_c	1-4-8-1	4
p_d	1-7-6-1	3
p_e	1-3-5-8-1	2.4

Table 4.2 Arrival rate and routing of products for L2

Product Type	Routing	Jobs/Hour
p_a	1-2-4-10-12-1	12
p_b	1-3-15-1	6
p_c	1-5-11-14-1	4
p_d	1-7-9-6-1	3
p_e	1-3-5-8-13-1	2.4

Table 4.3 Flow matrix (Loads/hour) for L1

From/To	1	2	3	4	5	6	7	8
1		12	8.4	4			3	0
2			12					
3				12	8.4			
4	12							4
5							6	2.4
6	3							
7	6					3		
8	6.4							

Table 4.4 Flow matrix (Loads/hour) for L2

From/To	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		12	8.4		4		3								
2				12											
3					2.4										6
4										12					
5								2.4			4				
6	3														
7									3						
8													2.4		
9						3									
10												12			
11														4	
12	12														
13	2.4														
14	4														
15	6														

In both L_1 and L_2, loads delivered to the stocker exit the system. In L_1, the processing times at each process tool are assumed to be deterministic, and the mean processing time is such that processors utilization is 60%. The distance from the loading station, s_i^p to the unloading station s_{i+1}^d, is 15 feet. The distance from the unloading station s_i^d to the loading station s_i^p is 5 feet. Each vehicle travels at a speed of 60 ft/min, empty or loaded, and it takes 15 seconds to pick-up or drop-off a load.

In L_2, we tested the analytic model results for three different processing time distributions at processing stations: deterministic, exponential, and triangular; the parameters were chosen so that the average processing time is the same for all three cases, and the mean processing time is such that processors utilization is 60%. The distance from the loading station, s_i^p to the unloading station s_{i+1}^d, is 10 feet. The distance from the unloading station s_i^d to the loading station s_i^p is 5 feet. Each vehicle travels at a speed of 60 ft/min, empty or loaded, and it takes 15 seconds to pick-up or drop-off a load.

In both layouts, we compare the interarrival times of empty vehicles at each pick-up station estimated from the analytic model and a simulation model, at multiple fleet sizes starting with the minimum fleet size that can handle the expected move requests (3 vehicles for L_1, and 5 vehicles for L_2) up to the maximum fleet size that does not cause AMHS deadlock, 15 vehicles for L_1, and 29 vehicles for L_2. We used AutoMod simulation software to obtain simulated values for this performance measure based on 10 replications of 10 days each; we also made sure that the system reached steady state before we started collecting statistics. In the simulation model, move requests arrive at the stocker according to a Poisson process. However, move request arrivals at processor stations are the result of lot arrival and operation times at the processor station, i.e. we did not force them to follow a Poisson process.

For comparison purposes, we took the average interarrival time of empty vehicles at all the pick-up stations. The analytical and simulated expected interarrival time of empty vehicles are shown in Table 4-5 and Table 4-6 for L_1 and L_2, respectively. The relative error represents the difference between the analytical and the average obtained from the simulation.

Table 4.5 Analytical and simulated average expected time between two empty vehicle arrivals

Fleet size	$T_{analytic}$	$T_{simulation}$	*rel. error*
3	6.49	7.01	-7%
4	5.10	5.74	-11%
5	4.59	5.24	-12%
6	4.36	4.86	-10%
7	4.25	4.59	-7%
8	4.21	4.45	-5%

Table 4.5 Continued

Fleet size	$T_{analytic}$	$T_{simulation}$	*rel. error*
9	4.20	4.35	-3%
10	4.23	4.33	-2%
11	4.27	4.40	-3%
12	4.33	4.55	-5%
13	4.40	4.83	-9%
14	4.47	5.19	-14%
15	4.56	5.70	-20%

Table 4.6 Analytical and simulated average expected time between two empty vehicle arrivals

Fleet size	$T_{analytic}$	$T_{simulation}$ *(Determinstic)*	$T_{simulation}$ *(Exponential)*	$T_{simulation}$ *(Triangular)*	*Average rel. error*
5	8.19	9.48	8.88	9.50	13.5%
6	7.19	8.45	7.90	8.37	14.3%
7	6.86	7.81	7.23	7.78	11.0%
8	6.38	7.37	7.15	7.31	13.2%
9	6.19	7.05	6.57	7.00	11.2%
10	6.08	6.78	6.35	6.73	9.0%
11	6.02	6.58	6.15	6.47	6.4%
12	6.00	6.33	6.02	6.30	4.1%
13	5.99	6.22	5.95	6.14	1.8%
14	6.01	6.07	5.84	6.01	-0.4%
15	6.04	6.02	5.80	5.91	-2.2%
16	6.08	5.98	5.77	5.84	-3.7%
17	6.13	5.95	5.76	5.80	-4.7%
18	6.19	5.96	5.78	5.80	-5.4%
19	6.26	5.97	5.83	5.81	-6.1%
20	6.33	6.02	5.91	5.87	-6.1%
21	6.41	6.09	6.01	5.96	-5.9%
22	6.49	6.19	6.15	6.08	-5.4%
23	6.57	6.34	6.32	6.24	-4.1%
24	6.66	6.53	6.55	6.44	-2.2%
25	6.75	6.76	6.78	6.69	-0.1%
26	6.84	7.04	7.07	6.97	2.8%
27	6.94	7.39	7.44	7.33	6.6%
28	7.04	7.89	7.97	7.86	12.4%
29	7.14	8.87	8.92	8.85	24.6%

At the early design stages when the system requirements have not yet stabilized, testing alternative designs with simulation would be time-consuming and for those early phases the analytical model would be a good choice given that it performs reasonably well with acceptable error percentages. Based on the test results of the two numerical examples, the following conclusions can be drawn:

- The accuracy of the model deteriorates at low and high fleet sizes. This is due to the inability of the model to handle the complexity of estimating vehicle-blocking caused by chain-blocking (many vehicles blocking each other). It is expected that chain-blocking occurs when the number of vehicles is large since even when a single vehicle stops to perform service, many vehicles behind it will be blocked. When the number of vehicles is small, chain-blocking is likely to occur because the amount of loading and unloading per vehicle is high and vehicles make frequent stops.
- The processing times distribution does not seem to impact the analytical model Etimates of the expected throughput capacity of the AMHS.

4.7. Summary and future work

In this chapter, we have presented a reduced-state Discrete Time Markov Chain model that can be used in assessing closed-loop AMHS performance. The model deals with the computational challenges that were observed in the more detailed model discussed in Chapter three. The growth of the state space dimensionality is polynomial as opposed to the exponential growth of the previous model.

Experimental comparisons of the model generated results with detailed simulation for small and medium example problems produced acceptable error margins and the results are obtained very quickly. In fact, in simulation modeling, the execution time increases exponentially with the number of vehicles because of the drastic increase in the number of events that the model has to track. The analytical model execution time is a function of the number of stations only.

One issue that needs to be investigated is the value of the model in practice to AMHS designers and analysts. This can be done by solving a larger instance using the SEMATECH data set for a virtual fab and test the model accuracy.

CHAPTER 5

EXPECTED RESPONSE TIMES AT LOADING STATIONS OF VEHICLE-BASED CLOSED LOOP AMHS

5.1. Introduction and motivation

Chapters three and four presented two extended Markov chain models that can be used in estimating the throughput capacity of a closed-loop vehicle-based material handling system. Experimental comparisons of the analytical model with detailed simulation produced acceptable error margins with regard to this performance metric. This chapter uses the parameters estimated from the extended Markov chain model to derive an approximation of the expected response time by the material handling system to a move request from the production system. The response time is essentially the waiting time of loads at loading stations.

The expected response times are important because they impact fab-level performance metrics (such as the production cycle time) and are used to estimate the work-in-process (WIP) at the output buffers. If the response times are too large, the production cycle times are inflated by a non value-adding operation (material handling) and in this case, the performance of the AMHS could become a bottleneck, an unacceptable situation in general and for wafer fabs in particular. Moreover, having good estimates of the WIP levels helps support design decisions that set the capacity of the buffers for the processor tools and/or the stocker. Insufficient buffer capacity leads to an often-occupied buffer, and causes blocking and starvation of processor tools, also an unacceptable situation because it leads to under-utilization of the expensive production tools.

Response time to a load move request is defined as the time until an empty vehicle responds to the move request by traveling to, then picking up the load from the output buffer (loading station). The derivation of the expected response times is complicated by many factors. First, we must know the distribution of the vehicles on the loop at the time the load arrived. Second, there is a possibility that currently empty vehicles pick-up other loads at other stations before they reach the station for a newly arrived load. Third, we must consider the expected number of loads waiting ahead of the move request at the same station.

5.2. Modeling approach

The approach followed to derive the expected response time analyzes each pick-up station separately. We follow a load that just arrived at the pick-up station and condition on whether or not there are other loads waiting in the queue at the same station. The main details are in deriving the response time for the first-in-line load because its pick-up time depends on the location of the vehicles at the time of its arrival and the possibility of other loads being picked-up (dropped-off) from (to) other stations while the vehicles are traveling towards the waiting load's location.

The main idea is to consider each state, and to compute the average length of the path from each vehicle's location to the load's location, then condition on the state of a vehicle at the time the load arrived. The state-dependent response time is a random variable that has a probability function which is derived from the AMHS steady-state analysis developed in chapter four.

To account for having multiple vehicles, we must consider that the vehicle that eventually picks up the load is the one that uses the shortest time path among the paths the vehicles might take. We use the order statistic of the state-dependent response time and the adjusted probability function. Using the order statistic probability function, the expected response time at each station is estimated.

5.2.1. Additional assumptions

In addition to the assumptions in sections 3.6 and 4.4, the derivation of the expected response time is based on the following assumptions:

1. Jobs that queue at a machine requesting transportation are processed in FCFS order.
2. The number of loads at the other stations is irrelevant, the only relevant factor is whether or not there is a waiting load.
3. Vehicles are traveling independently and the correlation among the vehicles is ignored.

We need assumption (1) to make the expected response time approximation model tractable. Assumption (2) is based on the FEFS rule dictating that a load at some station is not picked up according to its order of arrival in the system but only when an empty vehicle *encounters* it. Assumption (3) simplifies the analysis because we can ignore the interaction among the vehicles. In the numerical tests based on simulation, assumption (3) will yield some errors in estimating the expected response time, and later we propose an approach to include correlation among the vehicles.

5.3. Relevant parameters from the extended Markov chain model

Suppose load x completed processing at machine i and joined the queue of loads at pick-up station s_i^p waiting for an empty vehicle. The waiting time of load x will be influenced by the following factors:

- The number of loads waiting ahead of x at s_i^p.
- The location and condition of each vehicle: vehicles' *states*.
- The transition probabilities between states.
- The possibility of loads waiting at other pick-up stations.
- The interarrival time between empty vehicles at s_i^p. If there are loads waiting ahead of x at s_i^p, it is important to know the expected time between pickups.

The extended Markov chain model developed in chapter four provides estimates of all the parameters that determine the above factors:

- The visit ratios of each state r, denoted by $v_r, r = 1,2,...,|R|$, where state r is expressed as r = (m, n, c), m = $machine \in \{1,2,...,M\}$, n = $station \in \{p,d\}$, c=$condition \in \{e,f,b,k,s\}$.
- The transition probabilities between the states, which are functions of:
    - load-drop off probabilities $\mathbf{r} = \{r_i\}, i = 1,...,M$,
    - load-encountering probabilities $\mathbf{q} = \{q_i\}, i = 1,...,M$, and
    - vehicle-blocking probabilities $\mathbf{p^d} = \{p_i^d\}, \mathbf{p}^p = \{p_i^p\}, i = 1,...,M$.
- The expected interarrival time between empty vehicles to each pick-up station $\mathbf{T} = \{T_i\}, i = 1,...,M$.

5.4. Expected response time model

5.4.1. Notation

s_i^p : pick-up station of machine *i*.

λ_i : arrival rate of loads to machine *i*.

q_i : probability that an empty vehicle encounters and picks up a load from machine *i*.

ϕ_i: probability that the vehicle arriving to s_i^p is empty.

S^P : set of states in which a vehicle picks-up a load.

E^P : set of states in which a vehicle is arriving empty at a pick-up station.

T_i: expected interarrival time between empty vehicles to station s_i^p.

T: expected interarrival time between vehicles (empty or loaded) to any station.

R_i : expected time until an empty vehicle arrives to s_i^p.

$R_i(L)$: expected response time for to a move request given that *L* loads are waiting in queue when the load arrives.

$p_i(L)$: probability that *L* loads are waiting in the queue at s_i^p.

$R_i(0)$; response time to the first-in-line load at s_i^p.

π_r: steady-state probability that a vehicle is in state *r*

t_{rk}: transition time of a vehicle from state *r* to state *k*.

$E(t_{(r)})$: time left before the vehicle leaves state *r*.

R_r^i: expected time until a vehicle arrives empty at s_i^p given that it is currently in state *r*.

τ_{ji}: probability that a vehicle picks-up a load from machine j and delivers it to some station other than and passing through s_i^p

$s_j = (j, p, s)$ a state in which a vehicle is picking-up a load at machine j.

$e_j = (j, p, e)$ a state in which a vehicle is empty at the pick-up station of machine j.

ω_{e_j}: expected time until a vehicle in state e_j arrives empty at s_i^p to pick-up the first waiting load.

5.4.2. Model derivation

The expected response time depends, among other factors, on the station at which the move request originates. Using a central server model, such as *M/G/c*, to approximate the response time to a move request at any station; a common approach in the literature for AMHS (Johnson, 2001), (Johnson and Brandeau, 1994), and (Kobza et al., 1998), is not appropriate for the AMHS analyzed in this research. First, the vehicles are constantly moving, and thus it is not obvious how to calculate an average trip time (necessary to estimate the distribution of the generalized service times). Second, because the response time for each move request depends on the station where it originates, there is a wide range of possible response time values and lumping them all under an average response will mask these differences. Third, for design purposes it is important to analyze each station separately to accurately estimate the required buffer capacity.

We will use the "tagged" load approach discussed in Bozer et al. (1994), where load x is the tagged load. The tagged load x arrives at the queue of pick-up station s_i^p, since loads are assumed to arrive according to a Poisson process, load x sees the steady-state distribution of loads at s_i^p. Let L denote the number of loads waiting at s_i^p at the time load x arrives. If L = 0, the response time to load x, denoted by R_i is the expected time until an empty vehicle arrives to s_i^p. If L > 0, R_i is the expected time until all L+1 loads are picked up plus the time left for the first load in queue to be picked up. Let $R_i(L)$ denote the response time for load x given that L loads are waiting in queue when x arrives, and let $p_i(L)$ denotes the probability that L loads are waiting in the queue at s_i^p, then the expected response time to load x at s_i^p is estimated by:

$$R_i = \sum_{L=0}^{\infty} R_i(L)P_i(L) \tag{5.1}$$

We first consider the case L > 0. For this case, $R_i(L)$ is the expected response time to the first-in-line load and the expected time for the remaining L-1 loads plus load x itself. The expected response time for each load other than the first-in-line is simply the expected interarrival time between empty vehicles to station s_i^p, denoted by T_i, thus:

$$R_i(L) = R_i(0) + L.T_i \tag{5.2}$$

Substituting Equation (5.2) into (5.1), we get:

$$\begin{aligned} R_i &= \sum_{L=0}^{\infty} (R_i(0) + L.T_i).P_i(L) \\ &= \sum_{L=0}^{\infty} R_i(0)P_i(L) + \sum_{L=0}^{\infty} L_i T_i P_i(L) \\ &= R_i(0) + T_i \sum_{L=0}^{\infty} LP_i(L) \end{aligned} \tag{5.3}$$

The term $\sum_{L=0}^{\infty} LP_i(L)$ in (5.3) is the WIP_i, the expected work-in-process at station s_i^p, and from Little's law, $WIP_i = \lambda_i R_i$ thus:

$$R_i = R_i(0) + T_i\lambda_i R_i$$

$$= \frac{R_i(0)}{1 - T_i\lambda_i} \tag{5.4}$$

Expression (5.4) is consistent with the expression developed in Bozer et al. (1994) for a single-vehicle system. λ_i is the arrival rate of loads to machine *i*, which is already given as a parameter for each problem instance. T_i is the expected interarrival time between empty vehicle visits to station s_i^p, which we estimate using the extended Markov chain model from chapter four. The only unknown is $R_i(0)$; the response time to the first-in-line load at s_i^p; its derivation is presented in Section 5.4.3.

5.4.3. Expected response time to the first-in-line load

In this section, we derive the expected response time to the first-in-line move request at pick-up station s_i^p, which is essentially the time until the first vehicle to respond enters state (i, p, e). We will next study $R_i(0)$ under two approaches. The first approach is quite simple and relies on the rate of empty vehicle arrivals to s_i^p. The second approach is more complex and is based on the expected length of the path from each vehicle's state to the waiting load location.

Approach 1

Suppose *T* is the expected time between two vehicle arrivals to s_i^p and ϕ_i is the probability that the arriving vehicle is empty. The expected time until a vehicle arrives at s_i^p from the moment the first-in-line load arrived is *T/2*, if this vehicle is empty, the load will be picked up, otherwise, the load will wait for the next vehicle to arrive, which takes *T* time units and if this second vehicle is empty, the load is picked up, otherwise, it has to wait for the next vehicle to arrive, and so forth. Based on this logic, we obtain the expected response time from:

$$R_i(0) = \frac{T}{2}\phi_i + (1-\phi_i)(2T\phi_i + (1-\phi_i)(3T\phi_i + (1-\phi_i)(4T\phi_i + (1-\phi_i)(.........)))$$

$$= \frac{T}{2}\phi_i + \sum_{z=1}^{\infty}(z+1)T\phi_i(1-\phi_i)^z \qquad (5.5)$$

We can show that $\sum_{z=1}^{\infty}(z+1)(1-\phi_i)^z = \frac{1-\phi_i^2}{\phi_i^2}$, and thus $R_i(0)$ is calculated from:

$$R_i(0) = T\left(\frac{\phi_i}{2} + \frac{1-\phi_i^2}{\phi_i}\right) \qquad (5.6)$$

T and ϕ_i are both obtained from the analysis results of the extended Markov chain model developed earlier in chapter four as follows. Recall from chapter four that $\upsilon_{(i,p,e)}$, and $\upsilon_{(i,p,f)}$ are the visit ratios to states (i,p,e) and (i,p,f), respectively, in a cycle of length *C*. Since the vehicle either arrives loaded or empty, then the arrival rate of vehicles is $\left(\upsilon_{(i,p,e)} + \upsilon_{(i,p,f)}\right)n/C$, where *n* is the number of vehicles. *T* can be obtained from:

$$T = \frac{C}{\left(\upsilon_{(i,p,e)} + \upsilon_{(i,p,f)}\right)n} \qquad (5.7)$$

And ϕ_i, the probability that the arriving vehicle is empty can be estimated from:

$$\phi_i = \frac{\upsilon_{(i,p,e)}}{\upsilon_{(i,p,e)} + \upsilon_{(i,p,f)}} \qquad (5.8)$$

Unfortunately, this approach is not supported by the numerical tests based on simulation, and as we show later, it yields significant errors in estimating the expected response time.

Approach 2

Consider the partial transition diagram in Figure 5.1. A transition probability and transition time is associated with each state transition. The second approach to estimating the time until a vehicle enters state (*i*, *p*, *e*), (i.e. a vehicle arrives empty to s_i^p), is to compute the *average* length of the path from each state to state (i, p, e). Later, because the AMHS has more than one vehicle, the vehicle that eventually picks-up the first-in-line load from s_i^p is the first one that arrives empty to s_i^p. We therefore need to use the order statistics of each expected path length. We divide this approach into two parts: deriving the expected path length from each state, and the probability associated with each path.

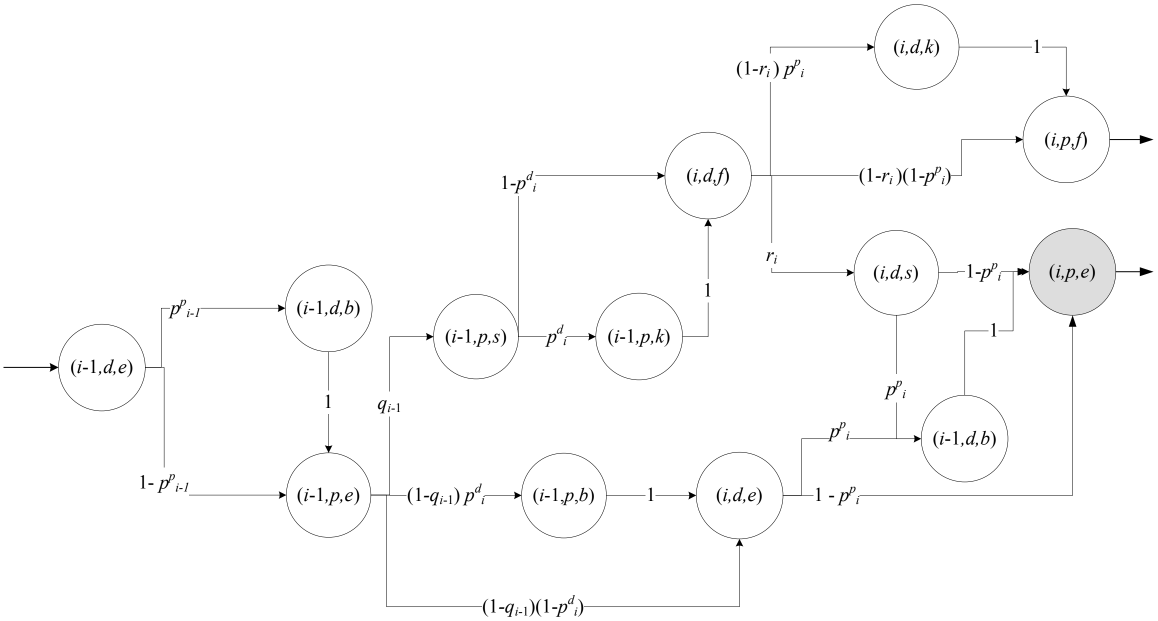


Figure 5.1 A partial transition diagram that demonstrates some of the paths leading to state (i, e, p)

Expected path length derivation

Let p_{rk} denote the transition probability from state r to state k, let t_{rk} denote the transition time from state r to state k. Let R_r^i denote the expected time until a vehicle arrives empty at s_i^p given that it is currently in state r , which we compute using

$$R_r^i = \sum_{\forall k \in R} p_{rk} (t_{rk} + R_k^i) \qquad (5.11)$$

The problem with this simplified approach arises when the vehicle enters a state that requires it to pick-up a load from a station other than s_i^p and deliver it to a station beyond s_i^p ; in this case the R_{r_i} values get inflated. For example, suppose that the vehicle is arriving empty at station s_{i-1}^p. If the vehicle does not find a load, it travels to s_i^d then to s_i^p and picks-up the waiting load. However, if the vehicle does find a load at s_{i-1}^p that is headed to a station beyond s_i^p , this particular vehicle will take a long time until it arrives empty at s_i^p and most likely, another vehicle will pick-up the waiting load, thus considering these pick-ups will inflate the expected response time from s_{i-1}^p to s_i^p .

We address this problem as follows: suppose that the vehicle under consideration is empty at machine *j*. If it finds a load at *j* that will not be dropped off before it arrives to s_i^p , the next closest vehicle, which arrives at *j* after T_j time units, becomes the candidate vehicle to pick-up the load at s_i^p , and if this vehicle also finds a load that will not be dropped off before it arrives to s_i^p , the next vehicle becomes a candidate and so forth.

Therefore, let S^P denote the set of states where a vehicle picks-up a load, and let τ_{ji} denote the probability that a vehicle picks-up a load from machine *j* and delivers it to a station passing through s_i^p, these τ_{ji} values can be computed from the release rates as:

$$\tau_{ki} = \sum_{j=i+1}^{k-1} \lambda_{kj} / \lambda_k , \forall k \in S^P \tag{5.12}$$

Let E^P denote the set of states where a vehicle is arriving empty at pick-up stations. We will now modify the expression for the expected response time for the states in E^P. Let $e_j = (j, p, e)$ refer to the state that a vehicle is empty at the pick-up station of machine *j*. Let $s_j = (j, p, s)$ be a state in which a vehicle picks-up a load at machine *j*.

q_j is the probability that an empty vehicle encounters and picks up a load from machine *j*. The vehicle in state e_j picks-up the load at s_i^p, if and only if, it does not find a load at *j*, (with probability $1 - q_j$) or it finds a load that will be dropped off before s_i^p, with probability $q_j(1-\tau_{ji})$, or it finds a load that will be dropped off after s_i^p, with probability $q_j \tau_{ji}$, and when this last case happens, this vehicle does not pick up the load and the same logic is repeated for the next vehicle that arrives empty at machine *j*, which happens after time delay *T*$_j$. For the first two cases, the expected time until the empty vehicle at machine *j* picks up the load from s_i^p is, donated by ω_{e_j}, is obtained from:

$$\omega_{e_j} = q_j (1-\tau_{ji})(t_{e_j s_j} + R_{s_j}^i) + \sum_{\forall k \in R / s_j} p_{e_j k} (t_{e_j k} + R_k^i) \tag{5.13}$$

Adding the last case yields the following approximation for $R_{e_j}^i$:

$$R_{e_j}^i = \omega_{e_j} + q_j \tau_{ji} (T_j + \omega_{e_j} + q_j \tau_{ji} (2T_j + \omega_{e_j} + q_j \tau_{ji} (3T_j + \omega_{e_j} + q_j \tau_{ji} (......))))$$

$$= \sum_{z=0}^{\infty} (\omega_{e_j} + zT_j)\left(q_j \tau_{ji}\right)^z$$

$$= \frac{\omega_{e_j}}{\left(1 - q_j \tau_{ji}\right)} + T_j \frac{q_j \tau_{ji}}{\left(1 - q_j \tau_{ji}\right)^2} \tag{5.14}$$

We can now estimate the expected response time by a vehicle coming from any state using expressions (5.11) and (5.15) as follows:

$$R_r^i = \begin{cases} \sum_{\forall k \in R} p_{rk} (t_{rk} + R_k^i) & r \in R / E^P \\ \dfrac{\omega_{e_j}}{\left(1 - q_j \tau_{ji}\right)} + T_j \dfrac{q_j \tau_{ji}}{\left(1 - q_j \tau_{ji}\right)^2} & r \in E^P \end{cases} \tag{5.15}$$

Where *j* refers to the station associated with state *r*, and τ_{ji}, and ω_{e_j}, are obtained from expressions (5.12), and (5.13), respectively.

For *M* machines, the number of R_r^i values that we need to compute is equal to 10*M*. From expression (5.15), we have 10*M* linear equations, one for each R_r^i, where $R_i^i = 0$. The equations in (5.15) provide a unique solution for R_r^i as long as the AMHS is stable. The uniqueness of the solution can be established since the coefficients of the unknown R_r^i are the transition probabilities of the discrete time Markov chain that models the vehicle's state transitions. It was already demonstrated in chapter four that if the AMHS is stable, the Markov chain is finite and irreducible, and hence the transition matrix has full rank (Ross, 2000).

Path probability derivation

We now derive the expression for π_r^i, which is the probability that a vehicle in state $r=(j,n,c)$ will pick-up the first-in-line load from s_i^p. Recall that π_r denotes the steady-state probability of a vehicle being in state *r* and was estimated using the extended Markov chain model developed in chapter four. State $r=(j,n,c)$ implies that the vehicle is currently at machine *j*. In order for a vehicle in state *r* to arrive empty at s_i^p, it must not pick-up loads on its way to s_i^p unless these loads will be dropped-off before the vehicle arrives to s_i^p. We thus obtain π_r^i from:

$$\pi_r^i = \pi_r \prod_{k=j}^{i-1}(1-q_k\tau_{ki}) \tag{5.16}$$

The term $\prod_{k=j}^{i-1}(1-q_k\tau_{ki})$ in expression (5.16) is the probability that the vehicle does not pick-up loads from the machines it encounters on its way to s_i^p unless they will be dropped off before s_i^p.

Since the first empty vehicle that reaches the load is the one that picks it up, we need the ordered values of the random variable R_r^i. Given that R_r^i, *r* = 1, 2, ..., |*R*| is the expected time to respond to a load at s_i^p starting from state *r*, with probability mass function π_r^i, then $R_{(1)}^i < R_{(2)}^i < ... < R_{(R)}^i$ is the order statistic. We must now take into account that there are *n* vehicles. Let $\pi_{(r)}^i(n)$ denote the probability that one of the *n* vehicles will come from state *r* to pick-up the load at s_i^p, which is the probability that one vehicle uses the path from state *r and* that none of the other *n*-1 vehicles is using this path or a shorter path:

$$\pi^{i}_{(r)}(n) = n\pi^{i}_{(r)}.\left(1 - \sum_{k=(1)}^{(r)} \pi^{i}_{k}\right)^{n-1} \frac{1}{N} \tag{5.17}$$

The first term in expression (5.17) is the probability that one out of *n* vehicles will arrive to pick-up the load starting from state *r*, the second term is the probability that none of the *n*-1 vehicles used a shorter time path than the one that starts at state *r*, and N is determined by the normalization condition that $\sum_{\forall r \in R} \pi^{i}_{(r)}(n) = 1$. The expected response time to the first-in-line load at s_i^p is:

$$R_i(0) = \sum_{(r)=1}^{(R)} \pi^{i}_{(r)}(n)\left(R^{i}_{(r)} + E(t_{(r)})\right) \tag{5.18}$$

$E(t_{(r)})$ denotes the time left before the vehicle leaves state *r*. Since travel times and loading/unloading times are deterministic, we estimate $E(t_r)$ from:

$$E(t_{(r)}) = \frac{t_r}{2} \tag{5.19}$$

Where t_r is the time from the moment a vehicle enters state *r* until it transitions to the next state. For instance, if state *r* is an empty (or loaded) travel from station *i* to station *i*+1, then $t_r = t_{i,i+1}$.

Substituting (5.18) into (5.4), we get the following expression for the expected response time to a load just arriving at s_i^p:

$$R_i = \frac{\sum_{r=1}^{R} \pi^{i}_{(r)}(n)\left(R^{i}_{(r)} + E(t_{(r)})\right)}{1 - T_i \lambda_i} \tag{5.20}$$

A numerical example is provided in the next section to test and compare the analytical and the simulation results.

5.5. Numerical example

We use layout L_2 from Section 4.5 to compare the analytical model estimates of the average response time to values obtained from discrete-event simulation. Layout 2 (L_2) re-illustrated in Figure 5.2 has one stocker (m_1) and 14 process tools (m_2 through m_{15}) and five products (p_a, p_b, p_c, p_d, p_e). The total arrival rate to the stocker is $\lambda = \lambda_a + \lambda_b + \lambda_c + \lambda_d + \lambda_e$ jobs per minute.

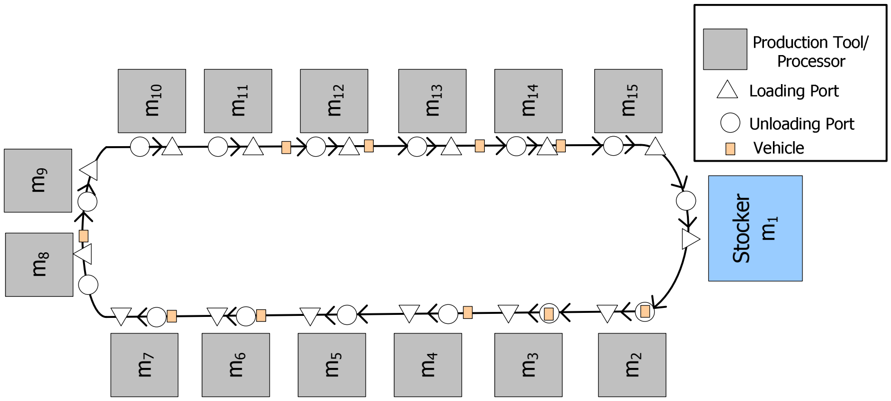


Figure 5.2 L2 A 15-machine example

We used the AutoMod simulation model to obtain the simulated values for the expected response time to move requests at each pick-up station. Simulation results are based on 10 replications and 10 days per replication. The analytical and simulated expected response times for each station are presented for three different fleet sizes: 6, 17, and 26 vehicles, shown in Tables 5.1, 5.2, and 5.3, respectively. The “Rel. error” column represents the relative difference between the analytical result and the sample mean obtained from simulation. In the simulation model, we experimented with two distributions for processing times at the processor tools: exponential and deterministic, and the comparisons with the analytical results are also reported in Tables 5.1-5.3.

Table 5.1 Analytical and simulated expected response time (secs) results for *n* = 6

Machine	*Analytical*		*Sim.*	*Rel. error*		*Sim.*	*Rel. error*	
	Approach 1	*Approach 2*	*(Exp.)*	*Approach 1*	*Approach 2*	*(Det.)*	*Approach 1*	*Approach 2*
1	46.0	81.6	87.0	-48%	-6%	100.2	-54%	-19%
2	61.3	93.8	104.1	-42%	-10%	110.4	-44%	-15%
3	64.4	99.6	106.9	-39%	-7%	96.9	-34%	3%
4	61.3	97.1	104.3	-42%	-7%	111.8	-45%	-13%
5	66.0	103.0	111.7	-43%	-8%	80.9	-18%	27%
6	68.6	104.6	117.2	-39%	-11%	112.0	-39%	-7%
7	68.6	103.1	115.0	-37%	-10%	109.3	-37%	-6%
8	69.0	101.1	115.1	-37%	-12%	110.9	-38%	-9%
9	68.6	95.8	114.5	-44%	-16%	116.8	-41%	-18%
10	61.3	105.8	99.4	-40%	6%	102.2	-40%	3%
11	67.8	112.2	113.4	-42%	-1%	111.5	-39%	1%
12	61.3	101.7	101.3	-39%	0%	108.8	-44%	-6%
13	69.0	107.2	118.5	-40%	-10%	117.0	-41%	-8%
14	67.8	102.3	114.6	-42%	-11%	119.3	-43%	-14%
15	66.3	108.2	106.2	-39%	2%	80.2	-17%	35%

Table 5.2 Analytical and simulated expected response time (secs) results for n = 17

Machine	*Analytical*		*Sim.*	*Rel. error*		*Sim.*	*Rel. error*	
	Approach 1	*Approach 2*	*(Exp.)*	*Approach 1*	*Approach 2*	*(Det.)*	*Approach 1*	*Approach 2*
1	12.0	18.7	17.1	-31%	10%	18.2	-39%	5%
2	13.8	19.9	18.8	-24%	6%	23.0	-24%	10%
3	14.2	19.6	19.6	-30%	0%	15.4	-38%	15%
4	13.8	19.2	19.7	-29%	-3%	22.4	-10%	25%
5	14.4	19.4	21.2	-33%	-8%	24.2	-36%	14%
6	14.8	18.9	21.7	-33%	-13%	21.3	-39%	22%
7	14.8	18.4	20.3	-29%	-9%	20.6	-31%	14%
8	14.8	19.4	19.4	-23%	0%	19.1	-28%	6%
9	14.8	18.5	19.1	-27%	-3%	13.7	-23%	3%
10	13.8	19.7	17.0	-19%	16%	18.7	1%	44%
11	14.7	19.6	18.7	-28%	5%	14.1	-21%	5%
12	13.8	18.8	18.1	-24%	4%	18.2	-2%	34%
13	14.8	19.1	19.5	-30%	-2%	17.8	-19%	5%
14	14.7	18.5	18.8	-25%	-2%	19.8	-18%	4%
15	14.4	18.7	17.8	-17%	5%	19.6	-27%	6%

Table 5.3 Analytical and simulated expected response time (secs) results for n = 26

Machine	*Analytical*		*Sim. (Exp.)*	*Rel. error*		*Sim. (Det.)*	*Rel. error*	
	Approach 1	*Approach 2*		*Approach 1*	*Approach 2*		*Approach 1*	*Approach 2*
1	12.0	11.2	10.9	-22%	3%	11.0	-22%	2%
2	13.8	11.7	12.0	-18%	-2%	8.9	8%	32%
3	14.2	11.5	12.1	-20%	-5%	11.2	-12%	3%
4	13.8	11.3	11.8	-21%	-4%	8.7	10%	30%
5	14.4	11.3	11.7	-11%	-3%	9.7	2%	16%
6	14.8	11.1	11.2	-14%	-1%	9.3	9%	19%
7	14.8	10.9	10.6	-13%	3%	8.8	15%	25%
8	14.8	11.7	10.6	-5%	10%	9.3	9%	25%
9	14.8	11.1	10.6	-2%	4%	8.8	15%	26%
10	13.8	11.8	10.1	-5%	17%	9.0	7%	32%
11	14.7	11.5	11.2	-8%	2%	10.3	-2%	12%
12	13.8	11.1	10.9	-13%	2%	9.8	-2%	14%
13	14.8	11.3	11.7	-12%	-4%	10.3	-2%	9%
14	14.7	11.0	11.3	-14%	-3%	9.5	6%	16%
15	14.4	11.2	11.3	-13%	0%	9.1	9%	23%

Figure 5.3 shows the average expected time to a move request averaged over all the stations obtained from the analytical model and the simulation model. “Sim-Exp.” refers to the model with exponential processing.

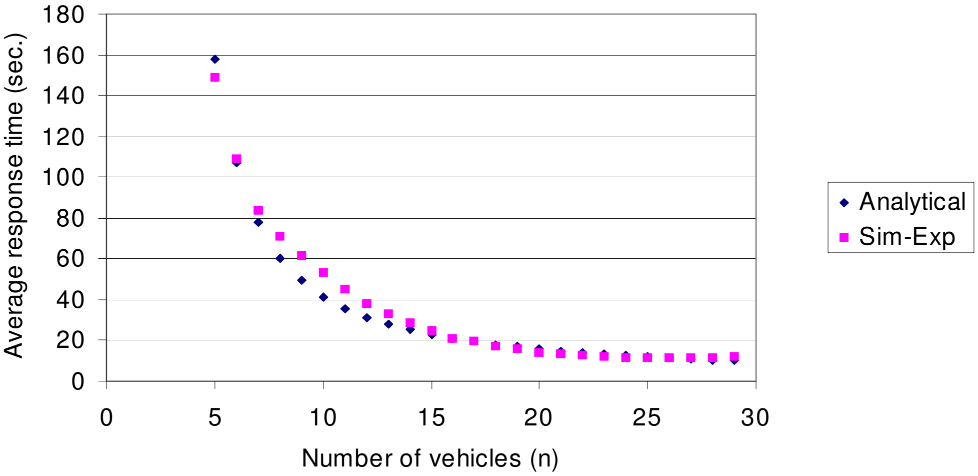


Figure 5.3 Average response time to a move request

We make the following observations based on the test results of the numerical example. The simple approach (approach 1) always underestimates the expected response time and yields significant errors. Recall that approach 1 estimates the response time by conditioning on whether or not the arriving vehicle is loaded, and we used the average interarrival time between two vehicles to do so. Vehicles are not evenly distributed around the loop, and taking the average interarrival time hides the variability in the arrival rate of vehicles. In fact, when the vehicles are displaying train-like behavior (vehicles are traveling close to each other), the variability in the interarrival time is very large.

Using approach 2, the analytical model estimates of the expected response time are reasonably accurate when the processor tools have exponential (rather than deterministic) processing times. This is expected since one of the main assumptions of the analytical model is that vehicles visit stations at random points in time. When the processing times are exponential, the randomness in vehicles' movement is somewhat justified because the arrival rate of move requests is exponential and as a result vehicles transition between states is random. When the vehicles' arrival process is Poisson, then we can assume that when load *x* arrives at the queue of a pick-up station, the remaining time of service for the first-in-line load is independent of the arrival time of load *x*. It also seems that this assumption is more important when estimating the response time than when estimating the throughput capacity.

The analytical model ignores the correlation between the vehicles' location and condition and to some extent this explains the deviation of the analytical model results from the simulation. More specifically, the vehicles are often clustered together and display train-like behavior, and when this happens, there is a correlation between the location of one vehicle and the rest of the fleet. The correlation increases when the proportion of empty travel increases because empty vehicles are more likely to cluster behind a vehicle in service. One would expect that as the number of vehicles increases, the AMHS utilization decreases, and the clustering behavior will increase. In fact, this will not be the case but rather two or more trains of vehicles will form. To see this, suppose there was a fleet of three vehicles that travel in a train and whenever this train approaches a loading station, the first vehicle picks up the waiting load and the next move request arrives at the station just after the third vehicle has passed. Now suppose that three more vehicles were added. Because of the time gap between the first vehicle and the fourth vehicle, it is more likely that by the time the fourth vehicle arrives at the loading station, a load has arrived, and this will break the train.

Ignoring the correlation between vehicles is also behind the considerable deviation of the analytical results from the deterministic processing time model. With deterministic processing, the randomness of move requests arrival process decreases but randomness would have led to breaking the trains of vehicles and reducing the correlation.

5.6. Summary and future work

In this chapter, we have presented an approach to derive the expected response time of the AMHS to a move request for each pick-up station separately. The response time depends on the location of the vehicles at the time of its arrival and the possibility of other loads being picked-up (dropped-off) from (to) other stations while the vehicles are traveling towards the waiting load's location.

The derivation is not straightforward and especially complicated for multi-vehicle system because the vehicle that picks up the load is not necessarily the closest to its location but it is the one that takes the shortest time path. We based our calculations on the expected length of the path from each vehicle's state to the load's location, then we conditioned on the state of a vehicle at the time the load arrived. For multi-vehicle systems, we assumed that the vehicles are moving independently on the loop and the state of a vehicle has no impact on the other vehicles' states.

Experimental comparisons of the model generated results with detailed simulation for one example problem produced acceptable error margins and the results are obtained very quickly. The model performs well mostly when there is a sufficient level of randomness in the arrival process of vehicles to pick-up and drop-off stations. Otherwise, low variability in the AMHS arrival process increases the correlation among the vehicles and increases the deviation of the analytic results from the simulation.

The model can be further improved if we incorporate the correlation among the vehicles into the model such that when a vehicle is in some state, the distribution of the other vehicles among the remaining states is adjusted. This enhancement of the current model will be pursued in future work in addition to conducting a study of the impact of machines' sequence along the loop on the expected response time. If the technological constraints allow the fab designer to allocate the processing equipment anywhere on the loop, achieving the right sequence might lead to significant improvement in performance metrics and cost savings.

One more issue that needs to be investigated, which is the topic of the next chapter is the value of the model in practice to AMHS designers and analysts. This will be done by solving a larger instance using international SEMATECH data set for a virtual 300mm fab and test the model accuracy.

CHAPTER 6

COMPUTATIONAL STUDY – SEMATECH FAB

The primary focus of this chapter is to evaluate the extended Markov chain model for throughput capacity estimation and the expected response time model using a detailed simulation model of a generic virtual fab, (International SEMATECH, 2001).

Ideally, the evaluation of the analytical models is based on the data set of an actual physical system. It is difficult, however, to get actual data, and therefore in this research we rely on simulation, under the assumption that the simulation model is equivalent to the real system. If the analytical model has a comparable accuracy to the simulation, we can assume that it also has a comparable accuracy to the real system.

6.1. SEMATECH fab model

The simulation model has two components: production and material handling. The production component describes the products, process routes, production tools and models the logical flow of material in the fab and the assignment of production equipment to products. The material handling component describes the layout, vehicles, stockers and models the physical material transport. The software used for simulation is AutoSched. AP (ASAP) 7.0 in conjunction with AutoMod 9.1; both are products of Brooks Automation Inc.

The SEMATECH fab has 24 bays and we apply the analytical models to one photolithography bay (bay seven) that holds a set of photolithography, inspection and measurement tools. This particular bay was selected because it has the highest percentage of intrabay traffic, because it holds the photolithography tools; the most expensive equipment in the fab with the highest utilization, and because there is a wide range of utilization values across the tools.

The modeled photolithography bay consists of four groups of production tools:

- Photolithography (Litho).
- Critical dimension (CD) measurement tools (Meas_CD).
- Overlay measurement tools (Meas_Overlay).
- Ply inspection tool (Insp_Ply).

Figure 6.1 illustrates the layout of bay 7 and Tables 6.1 and 6.2 summarize the details of the model. One product family fabrication is modeled in the simulation, which is SEMATECH’s 300mm aluminum process flow for 180nm technology with six metal layers, and 21 masks. For this single product family, ten products are constantly released into the facility and they all follow similar process plans (routings) with little differences during the photolithography operations. The wafers travel in lots of 25, and the release rate is 20,000 wafers/month (wpm) (800 lots/month).

Each lot visits the modeled bay six times throughout its production cycle. Each time a lot of 25 wafers enters bay seven it starts at the incoming-lots stocker (stocker 1) visits the processor tools following the sequence: *Litho –> Meas_Overlay –> Insp_Ply –> Meas_CD.* After the last step, the lot travels back to the outgoing-lots stocker (stocker 2), travels to other bays to undergo other operations then comes back to bay 7.

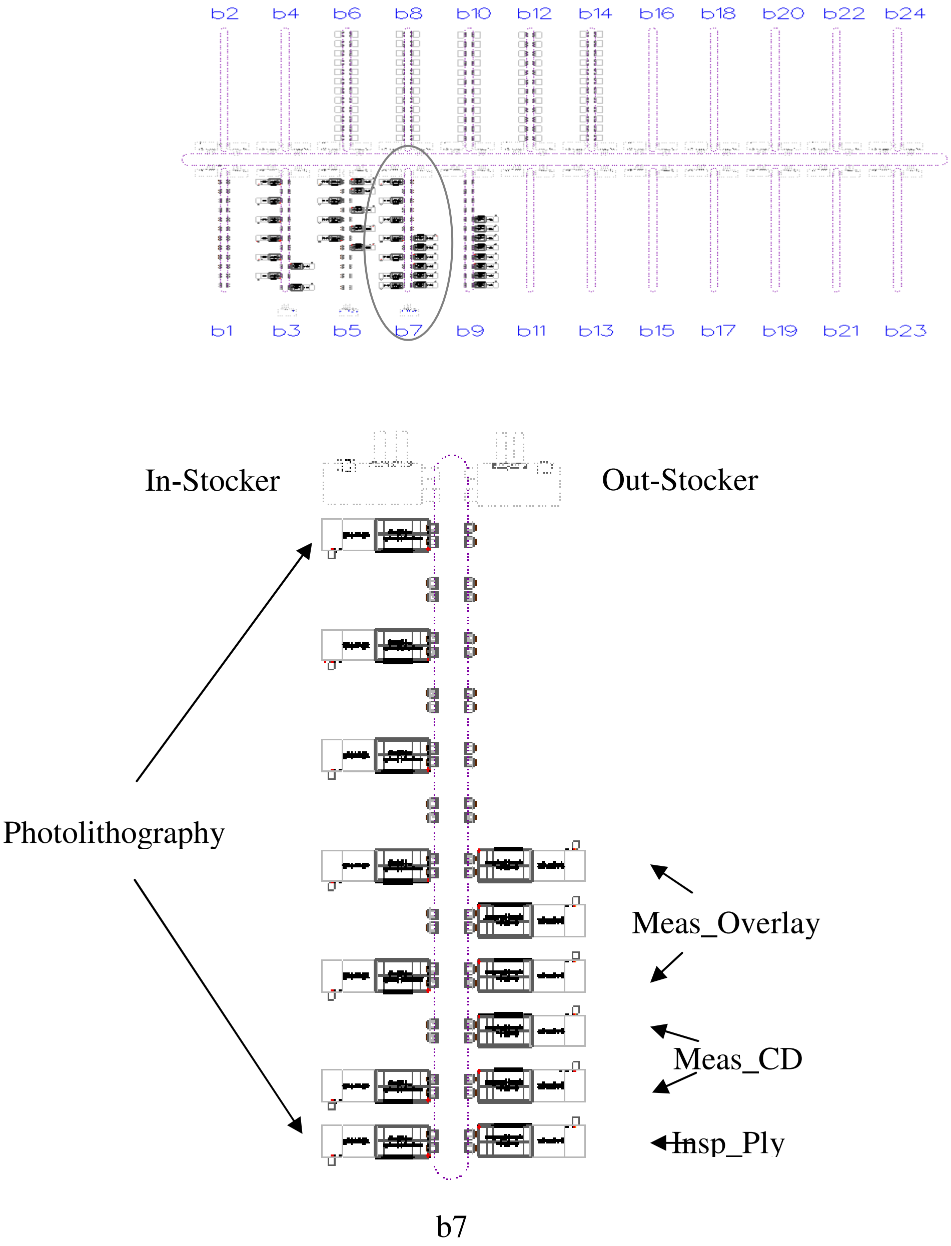


Figure 6.1 SEMATECH Fab: Photolithography bay 7

The modeled bay is 105 ft long and the aisle that separates the production equipment is five feet wide. There are two stockers that connect bay 7 to the interbay system, one is used to store the incoming lots to the bay and the other is for the outgoing lots. The AMHS intrabay loop of bay 7 has a series of 53 control points that divide the loop into zones. Vehicles circle the loop in search of work and only one vehicle can be at a control point at a time, if a vehicle stops at a control point for any reason, all the vehicles behind it stop at the preceding control points on the track. A vehicle cannot leave a control point unless it can claim the next one in the track, which implies that only one vehicle can be occupying the track segment between two control segments. These control points represent the loading/unloading stations for stockers and tools, or points where a vehicle may stop and wait until the vehicle blocking the control point ahead of it clears the way.

Table 6.1 Bay 7 information- processor tools

Production Equipment (Processor tools)			
Processor tool group	**Number of identical tools**	**Utilization**	**Distribution of processing times**
Photolithography (Litho)	7	60%	Deterministic
Meaure CD	2	25%	Deterministic
Meaure Overlay	3	32%	Deterministic
Insp. Ply	1	15%	Deterministic

Table 6.2 Bay 7 information- AMHS

AMHS	
Number of vehicles	3
Effective vehicle speed	3 *ft/sec*
Loading/unloading time	15 *sec*
Dispatching policy	FEFS

6.2. Experiments

This study is concerned with evaluating the analytical model at a wide range of operating scenarios. The performance metrics that will be compared are the AMHS throughput capacity and the expected response time to a move request. Before conducting the experiments, we did not know which factors will have an impact on the analytical results but we anticipated that the important factors might be those that control or affect the AMHS throughput capacity and response time. We decided to keep the processing time distributions, the process routings and the physical location of the equipment, unchanged and we investigated the impact of the factors listed in Table 6.3 at different levels. The factors we studied are:

1. *The average release rate of products (λ)*

The base setting for the average release rate is 20,000 wpm. Increasing this number will increase the utilization of the production equipment and the requirements from the AMHS.

2. *Single vs. multiple bottlenecks*

The modeled bay currently has the photolithography as the highest utilized tool group. We will create scenarios by changing the processing times of the other tool groups so that all the tools in the modeled bay are equally utilized.

3. *Capacity of the queue at each tool*

The analytical model assumes that the stockers and the processor tools have infinite queue capacity. Therefore, it is anticipated that this factor will impact the model performance. The stockers will always have ample capacity but we will experiment with two levels of the queue capacity at the processor tools.

4. AMHS fleet size

In the problems we studied in Chapters 4 and 5, it was obvious that the number of vehicles has a significant impact on the deviation of the analytical results from the simulation results. SEMATECH virtual fab has three vehicles in bay 7 and we will vary this factor between 3 and 6.

5. Vehicle travel velocity

The AMHS utilization (the percentage of loaded travel) will increase when the vehicles are slower and vice versa. The current value of this factor is 3 ft/sec and we consider two more levels.

Table 6.3 Levels of factors in the computational study

Factor	**Levels**
Release rate	15,000, 20,000, 22,500, and 24,000 *wpm*
Bottleneck tools	Single vs. multiple bottlenecks
Queue capacity	4, and 10 *lots*
AMHS Fleet size	3, 4, 5, and 6 *vehicles*
Vehicles' travel velocity	1, 3, and 4.5 *ft/sec*

6.2.1. Simulation model experiments

There are 192 unique combinations of the selected factors. The simulation model runs until it reaches steady-state before collecting the performance statistics. For each combination of the factors the simulation is executed with five replications, each runs for 100 days. For each combination we collect two metrics: the throughput capacity of the AMHS and the expected response time to a move request.

6.2.2. Analytical model experiments

The analytical model output will not be impacted by the queue capacity or the processing time of the processor tools. Recall that the inputs to the analytical model are the layout of the bay, the from-to release rates, the from-to travel times and the loading/unloading times. Therefore, the total number of analytical model experiments is 48.

6.3. Comparison results

Table 6.3 presents the differences in average throughput capacity and average response time at each factor combination. The columns labeled "*sim.*" represent the simulation model results and the columns labeled "*ana.*" refer to the analytic results. The columns labeled "*rel. error*" present the relative difference between the analytical result and the sample mean obtained from simulation. In this comparison, we take the average response time across all the stations for ease of demonstration and because, unlike the example in chapter five, the relative errors for each station were close so that taking the average does not mask the individual errors. When the vehicle speed is at the lowest level 1 *ft/sec*, the relative errors are different for different tool groups.

F_1 is the release rate factor, F_2 is the single (S) vs. multiple (M) bottleneck factor, F_3 refers to the queue capacity, F_4 refers to the AMHS fleet size, and F_5 refers to the vehicles' travel velocity.

Table 6.4 Simulation and analytical models outputs comparison

F_1	F_2	F_3	F_4	F_5	**Throughput (moves/month)**			**Response Time (seconds)**		
					sim.	***ana.***	***rel. error***	***sim.***	***ana.***	***rel. error***
24*K*	*S*	10	3	1	30,349	30,765	1%	117.0	105.3	-10%
24*K*	*S*	10	4	1	41,377	41,495	0%	83.3	71.3	-14%
24*K*	*S*	10	5	1	52,002	51,807	0%	63.5	54.8	-14%
24*K*	*S*	10	6	1	63,011	61,714	-2%	49.9	45.1	-10%
24*K*	*M*	10	3	1	30,349	30,765	1%	115.7	105.3	-9%
24*K*	*M*	10	4	1	41,377	41,495	0%	86.8	71.3	-18%
24*K*	*M*	10	5	1	52,002	51,807	0%	66.5	54.8	-18%
24*K*	*M*	10	6	1	63,011	61,714	-2%	50.1	45.1	-10%
24*K*	*S*	4	3	1	30,349	30,765	1%	120.7	105.3	-13%
24*K*	*S*	4	4	1	41,377	41,495	0%	89.0	71.3	-20%
24*K*	*S*	4	5	1	52,002	51,807	0%	67.8	54.8	-19%
24*K*	*S*	4	6	1	63,011	61,714	-2%	51.4	45.1	-12%
24*K*	*M*	4	3	1	30,349	30,765	1%	119.1	105.3	-12%
24*K*	*M*	4	4	1	41,377	41,495	0%	87.6	71.3	-19%
24*K*	*M*	4	5	1	52,002	51,807	0%	67.6	54.8	-19%
24*K*	*M*	4	6	1	63,011	61,714	-2%	49.7	45.1	-9%
24*K*	*S*	10	3	3	85,133	86,394	1%	34.0	30.5	-10%
24*K*	*S*	10	4	3	112,196	113,051	1%	29.0	23.2	-20%
24*K*	*S*	10	5	3	141,942	137,143	-3%	24.1	19.3	-20%
24*K*	*S*	10	6	3	173,086	158,954	-8%	20.0	16.9	-16%
24*K*	*M*	10	3	3	82,490	86,394	5%	36.2	30.5	-16%
24*K*	*M*	10	4	3	111,378	113,051	2%	30.4	23.2	-24%
24*K*	*M*	10	5	3	138,293	137,143	-1%	26.2	19.3	-26%
24*K*	*M*	10	6	3	169,896	158,954	-6%	20.8	16.9	-19%
24*K*	*S*	4	3	3	84,607	86,394	2%	33.6	30.5	-9%
24*K*	*S*	4	4	3	112,972	113,051	0%	29.1	23.2	-20%
24*K*	*S*	4	5	3	141,579	137,143	-3%	23.8	19.3	-19%
24*K*	*S*	4	6	3	171,172	158,954	-7%	20.3	16.9	-17%
24*K*	*M*	4	3	3	83,653	86,394	3%	35.7	30.5	-15%
24*K*	*M*	4	4	3	111,622	113,051	1%	30.1	23.2	-23%
24*K*	*M*	4	5	3	139,997	137,143	-2%	25.7	19.3	-25%
24*K*	*M*	4	6	3	170,911	158,954	-7%	20.8	16.9	-19%
24*K*	*S*	10	3	4.5	122,200	123,712	1%	24.3	20.5	-16%

Table 6.4 Continued

F_1	F_2	F_3	F_4	F_5	**Throughput (moves/month)**			**Response Time (seconds)**		
					sim.	***ana.***	***rel. error***	***sim.***	***ana.***	***rel. error***
24*K*	*S*	10	5	4.5	205,738	189,121	-8%	18.2	14.0	-23%
24*K*	*S*	10	6	4.5	248,441	215,653	-13%	15.7	12.6	-20%
24*K*	*M*	10	3	4.5	119,984	123,712	3%	24.7	20.5	-17%
24*K*	*M*	10	4	4.5	163,410	158,717	-3%	20.7	16.3	-21%
24*K*	*M*	10	5	4.5	206,947	189,121	-9%	17.0	14.0	-18%
24*K*	*M*	10	6	4.5	243,207	215,653	-11%	15.1	12.6	-16%
24*K*	*S*	4	3	4.5	122,200	123,712	1%	24.3	20.5	-16%
24*K*	*S*	4	4	4.5	164,217	158,717	-3%	21.5	16.3	-24%
24*K*	*S*	4	5	4.5	205,738	189,121	-8%	18.2	14.0	-23%
24*K*	*S*	4	6	4.5	248,441	215,653	-13%	15.7	12.6	-20%
24*K*	*M*	4	3	4.5	119,984	123,712	3%	24.7	20.5	-17%
24*K*	*M*	4	4	4.5	163,410	158,717	-3%	20.7	16.3	-21%
24*K*	*M*	4	5	4.5	206,947	189,121	-9%	17.0	14.0	-18%
24*K*	*M*	4	6	4.5	243,207	215,653	-11%	15.1	12.6	-16%
22.5*K*	*S*	10	3	1	30,574	31,006	1%	113.8	100.5	-12%
22.5*K*	*S*	10	4	1	41,191	41,732	1%	84.1	69.1	-18%
22.5*K*	*S*	10	5	1	51,674	52,040	1%	64.9	53.5	-17%
22.5*K*	*S*	10	6	1	63,252	61,943	-2%	49.0	44.3	-10%
22.5*K*	*M*	10	3	1	29,821	31,006	4%	112.5	100.5	-11%
22.5*K*	*M*	10	4	1	40,510	41,732	3%	85.2	69.1	-19%
22.5*K*	*M*	10	5	1	51,524	52,040	1%	66.1	53.5	-19%
22.5*K*	*M*	10	6	1	63,613	61,943	-3%	48.9	44.3	-9%
22.5*K*	*S*	4	3	1	30,452	31,006	2%	115.3	100.5	-13%
22.5*K*	*S*	4	4	1	41,528	41,732	0%	82.9	69.1	-17%
22.5*K*	*S*	4	5	1	52,277	52,040	0%	63.4	53.5	-16%
22.5*K*	*S*	4	6	1	63,160	61,943	-2%	49.1	44.3	-10%
22.5*K*	*M*	4	3	1	30,131	31,006	3%	110.9	100.5	-9%
22.5*K*	*M*	4	4	1	40,731	41,732	2%	86.5	69.1	-20%
22.5*K*	*M*	4	5	1	50,937	52,040	2%	67.1	53.5	-20%
22.5*K*	*M*	4	6	1	63,405	61,943	-2%	49.3	44.3	-10%
22.5*K*	*S*	10	3	3	85,772	87,076	2%	33.2	29.2	-12%
22.5*K*	*S*	10	4	3	115,090	113,704	-1%	27.9	22.5	-19%
22.5*K*	*S*	10	5	3	147,056	137,768	-6%	22.2	18.8	-15%
22.5*K*	*S*	10	6	3	172,250	159,553	-7%	19.9	16.5	-17%

Table 6.4 Continued

F_1	F_2	F_3	F_4	F_5	Throughput (moves/month)			Response Time (seconds)		
					sim.	*ana.*	*rel. error*	*sim.*	*ana.*	*rel. error*
22.5*K*	*M*	10	4	3	113,299	113,704	0%	29.4	22.5	-24%
22.5*K*	*M*	10	5	3	145,478	137,768	-5%	23.4	18.8	-20%
22.5*K*	*M*	10	6	3	173,228	159,553	-8%	20.2	16.5	-18%
22.5*K*	*S*	4	3	3	84,987	87,076	2%	33.4	29.2	-13%
22.5*K*	*S*	4	4	3	114,899	113,704	-1%	27.8	22.5	-19%
22.5*K*	*S*	4	5	3	143,596	137,768	-4%	23.4	18.8	-20%
22.5*K*	*S*	4	6	3	174,291	159,553	-8%	19.5	16.5	-15%
22.5*K*	*M*	4	3	3	84,947	87,076	3%	35.0	29.2	-17%
22.5*K*	*M*	4	4	3	113,107	113,704	1%	29.7	22.5	-24%
22.5*K*	*M*	4	5	3	143,865	137,768	-4%	24.5	18.8	-23%
22.5*K*	*M*	4	6	3	174,415	159,553	-9%	20.0	16.5	-17%
22.5*K*	*S*	10	3	4.5	123,844	124,691	1%	24.3	20.3	-17%
22.5*K*	*S*	10	4	4.5	165,515	159,635	-4%	20.8	16.2	-22%
22.5*K*	*S*	10	5	4.5	206,185	189,985	-8%	18.1	13.9	-23%
22.5*K*	*S*	10	6	4.5	251,255	216,469	-14%	15.5	12.5	-20%
22.5*K*	*M*	10	3	4.5	122,845	124,691	2%	24.0	20.3	-15%
22.5*K*	*M*	10	4	4.5	165,650	159,635	-4%	20.4	16.2	-21%
22.5*K*	*M*	10	5	4.5	206,977	189,985	-8%	17.1	13.9	-19%
22.5*K*	*M*	10	6	4.5	249,352	216,469	-13%	14.7	12.5	-15%
22.5*K*	*S*	4	3	4.5	123,202	124,691	1%	24.4	20.3	-17%
22.5*K*	*S*	4	4	4.5	166,042	159,635	-4%	21.2	16.2	-24%
22.5*K*	*S*	4	5	4.5	209,296	189,985	-9%	17.9	13.9	-22%
22.5*K*	*S*	4	6	4.5	248,397	216,469	-13%	15.7	12.5	-20%
22.5*K*	*M*	4	3	4.5	123,274	124,691	1%	23.9	20.3	-15%
22.5*K*	*M*	4	4	4.5	161,732	159,635	-1%	20.8	16.2	-22%
22.5*K*	*M*	4	5	4.5	205,088	189,985	-7%	17.6	13.9	-21%
22.5*K*	*M*	4	6	4.5	246,011	216,469	-12%	15.2	12.5	-18%
20*K*	*S*	10	3	1	31,126	31,408	1%	111.0	95.8	-14%
20*K*	*S*	10	4	1	41,975	42,127	0%	82.3	67.4	-18%
20*K*	*S*	10	5	1	53,135	52,428	-1%	62.9	52.5	-16%
20*K*	*S*	10	6	1	63,859	62,324	-2%	48.3	43.5	-10%
20*K*	*M*	10	3	1	30,336	31,408	4%	110.9	95.8	-14%
20*K*	*M*	10	4	1	40,761	42,127	3%	85.0	67.4	-21%
20*K*	*M*	10	5	1	52,004	52,428	1%	65.9	52.5	-20%

Table 6.4 Continued

F_1	F_2	F_3	F_4	F_5	**Throughput (moves/month)**			**Response Time (seconds)**		
					sim.	***ana.***	***rel. error***	***sim.***	***ana.***	***rel. error***
20*K*	*M*	10	6	1	63,524	62,324	-2%	48.5	43.5	-10%
20*K*	*S*	4	3	1	30,893	31,408	2%	110.5	95.8	-13%
20*K*	*S*	4	4	1	41,623	42,127	1%	81.2	67.4	-17%
20*K*	*S*	4	5	1	52,992	52,428	-1%	61.3	52.5	-14%
20*K*	*S*	4	6	1	64,228	62,324	-3%	47.1	43.5	-8%
20*K*	*M*	4	3	1	30,100	31,408	4%	110.1	95.8	-13%
20*K*	*M*	4	4	1	40,993	42,127	3%	86.1	67.4	-22%
20*K*	*M*	4	5	1	51,697	52,428	1%	65.2	52.5	-19%
20*K*	*M*	4	6	1	63,681	62,324	-2%	48.4	43.5	-10%
20*K*	*S*	10	3	3	87,532	88,214	1%	32.3	28.4	-12%
20*K*	*S*	10	4	3	117,395	114,792	-2%	27.1	22.1	-19%
20*K*	*S*	10	5	3	147,374	138,810	-6%	22.3	18.5	-17%
20*K*	*S*	10	6	3	177,886	160,553	-10%	18.9	16.3	-14%
20*K*	*M*	10	3	3	86,098	88,214	2%	33.6	28.4	-15%
20*K*	*M*	10	4	3	116,888	114,792	-2%	27.8	22.1	-21%
20*K*	*M*	10	5	3	145,078	138,810	-4%	23.9	18.5	-22%
20*K*	*M*	10	6	3	175,733	160,553	-9%	19.5	16.3	-16%
20*K*	*S*	4	3	3	87,251	88,214	1%	32.6	28.4	-13%
20*K*	*S*	4	4	3	116,437	114,792	-1%	27.3	22.1	-19%
20*K*	*S*	4	5	3	148,651	138,810	-7%	21.9	18.5	-16%
20*K*	*S*	4	6	3	178,548	160,553	-10%	18.7	16.3	-13%
20*K*	*M*	4	3	3	86,511	88,214	2%	33.8	28.4	-16%
20*K*	*M*	4	4	3	116,182	114,792	-1%	28.0	22.1	-21%
20*K*	*M*	4	5	3	146,554	138,810	-5%	23.3	18.5	-20%
20*K*	*M*	4	6	3	178,640	160,553	-10%	18.9	16.3	-14%
20*K*	*S*	10	3	4.5	128,092	126,322	-1%	22.9	19.5	-15%
20*K*	*S*	10	4	4.5	172,295	161,165	-6%	20.0	15.7	-21%
20*K*	*S*	10	5	4.5	213,811	191,425	-10%	17.4	13.6	-22%
20*K*	*S*	10	6	4.5	257,231	217,828	-15%	14.9	12.3	-18%
20*K*	*M*	10	3	4.5	126,307	126,322	0%	22.9	19.5	-15%
20*K*	*M*	10	4	4.5	170,551	161,165	-6%	19.5	15.7	-19%
20*K*	*M*	10	5	4.5	212,921	191,425	-10%	16.7	13.6	-19%
20*K*	*M*	10	6	4.5	258,802	217,828	-16%	13.9	12.3	-11%
20*K*	*S*	4	3	4.5	127,204	126,322	-1%	23.6	19.5	-17%

Table 6.4 Continued

F_1	F_2	F_3	F_4	F_5	**Throughput (moves/month)**			**Response Time (seconds)**		
					sim.	***ana.***	***rel. error***	***sim.***	***ana.***	***rel. error***
20*K*	*S*	4	4	4.5	172,300	161,165	-6%	20.3	15.7	-23%
20*K*	*S*	4	5	4.5	211,564	191,425	-10%	17.4	13.6	-22%
20*K*	*S*	4	6	4.5	260,642	217,828	-16%	14.3	12.3	-14%
20*K*	*M*	4	3	4.5	126,061	126,322	0%	23.1	19.5	-16%
20*K*	*M*	4	4	4.5	170,678	161,165	-6%	19.3	15.7	-19%
20*K*	*M*	4	5	4.5	210,215	191,425	-9%	16.8	13.6	-19%
20*K*	*M*	4	6	4.5	257,390	217,828	-15%	14.0	12.3	-12%
15*K*	*S*	10	3	1	31,852	32,213	1%	104.1	83.3	-20%
15*K*	*S*	10	4	1	43,192	42,917	-1%	74.8	60.9	-19%
15*K*	*S*	10	5	1	54,295	53,204	-2%	57.9	48.6	-16%
15*K*	*S*	10	6	1	65,837	63,086	-4%	44.6	40.9	-8%
15*K*	*M*	10	3	1	31,575	32,213	2%	101.1	83.3	-18%
15*K*	*M*	10	4	1	42,552	42,917	1%	81.9	60.9	-26%
15*K*	*M*	10	5	1	53,510	53,204	-1%	61.5	48.6	-21%
15*K*	*M*	10	6	1	64,993	63,086	-3%	44.9	40.9	-9%
15*K*	*S*	4	3	1	31,582	32,213	2%	105.2	83.3	-21%
15*K*	*S*	4	4	1	42,952	42,917	0%	76.0	60.9	-20%
15*K*	*S*	4	5	1	54,122	53,204	-2%	58.0	48.6	-16%
15*K*	*S*	4	6	1	65,776	63,086	-4%	42.8	40.9	-5%
15*K*	*M*	4	3	1	31,453	32,213	2%	102.0	83.3	-18%
15*K*	*M*	4	4	1	42,688	42,917	1%	82.2	60.9	-26%
15*K*	*M*	4	5	1	53,875	53,204	-1%	60.3	48.6	-19%
15*K*	*M*	4	6	1	65,416	63,086	-4%	43.7	40.9	-6%
15*K*	*S*	10	3	3	92,009	90,489	-2%	29.2	26.0	-11%
15*K*	*S*	10	4	3	124,512	116,968	-6%	24.0	20.7	-14%
15*K*	*S*	10	5	3	154,424	140,894	-9%	20.3	17.6	-13%
15*K*	*S*	10	6	3	184,933	162,551	-12%	17.1	15.6	-9%
15*K*	*M*	10	3	3	91,881	90,489	-2%	31.0	26.0	-16%
15*K*	*M*	10	4	3	122,782	116,968	-5%	25.4	20.7	-19%
15*K*	*M*	10	5	3	153,313	140,894	-8%	21.5	17.6	-18%
15*K*	*M*	10	6	3	186,369	162,551	-13%	17.3	15.6	-10%
15*K*	*S*	4	3	3	91,877	90,489	-2%	29.6	26.0	-12%
15*K*	*S*	4	4	3	123,401	116,968	-5%	24.4	20.7	-15%
15*K*	*S*	4	5	3	154,265	140,894	-9%	20.2	17.6	-13%

Table 6.4 Continued

F_1	F_2	F_3	F_4	F_5	**Throughput (moves/month)**			**Response Time (seconds)**		
					sim.	***ana.***	***rel. error***	***sim.***	***ana.***	***rel. error***
15*K*	*S*	4	6	3	186,077	162,551	-13%	16.8	15.6	-7%
15*K*	*M*	4	3	3	91,099	90,489	-1%	31.2	26.0	-17%
15*K*	*M*	4	4	3	122,608	116,968	-5%	25.3	20.7	-18%
15*K*	*M*	4	5	3	154,156	140,894	-9%	20.9	17.6	-16%
15*K*	*M*	4	6	3	183,782	162,551	-12%	17.7	15.6	-11%
15*K*	*S*	10	3	4.5	134,887	129,584	-4%	21.5	18.0	-16%
15*K*	*S*	10	4	4.5	181,249	164,226	-9%	18.6	14.8	-20%
15*K*	*S*	10	5	4.5	229,451	194,306	-15%	15.3	13.0	-15%
15*K*	*S*	10	6	4.5	270,181	220,547	-18%	13.5	11.8	-12%
15*K*	*M*	10	3	4.5	135,072	129,584	-4%	20.9	18.0	-14%
15*K*	*M*	10	4	4.5	181,020	164,226	-9%	18.1	14.8	-18%
15*K*	*M*	10	5	4.5	226,397	194,306	-14%	15.1	13.0	-14%
15*K*	*M*	10	6	4.5	272,899	220,547	-19%	12.1	11.8	-2%
15*K*	*S*	4	3	4.5	134,186	129,584	-3%	21.7	18.0	-17%
15*K*	*S*	4	4	4.5	180,052	164,226	-9%	18.8	14.8	-21%
15*K*	*S*	4	5	4.5	224,756	194,306	-14%	15.8	13.0	-18%
15*K*	*S*	4	6	4.5	272,347	220,547	-19%	13.4	12.5	-7%
15*K*	*M*	4	3	4.5	135,458	129,584	-4%	21.0	18.8	-10%
15*K*	*M*	4	4	4.5	181,093	164,226	-9%	18.0	15.6	-14%
15*K*	*M*	4	5	4.5	228,415	194,306	-15%	15.1	13.7	-9%
15*K*	*M*	4	6	4.5	274,536	220,547	-20%	12.3	12.5	1%

As was mentioned earlier, each simulation result is the average of the measured metric (response time or throughput capacity) taken over five replications. Examining the 90% confidence interval (C.I.), we observe that the ratio of the half width of the confidence interval to the mean for all the experiments, as illustrated in Figure 6.2, range mostly between 1.2% - 2.0%.

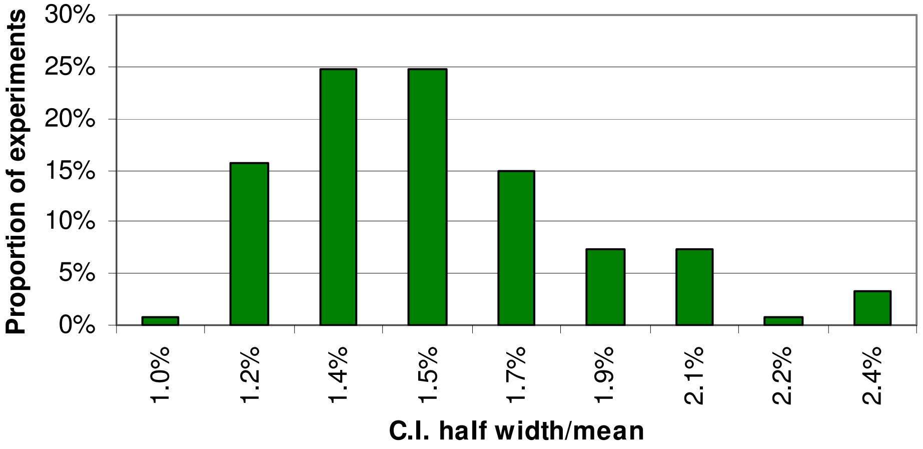


Figure 6.2 Simulation experiments half width results for 90% confidence

Now that we have established that the simulation results have reasonably narrow confidence intervals, we will compare the analytical result to the mean result of the simulation experiments. The comparison results are satisfactory and the error percentages are acceptable for both estimates of the throughput capacity and the average response time, despite the deterministic processing time distributions that violate the assumption of exponential arrival rate of loads at pick-up stations. The worst relative error in response time is -26% and in throughput capacity is -20%.

Figures 6.3 and 6.4 illustrate the frequency of the relative error percentages in the experiments for throughput capacity and average response time estimates, respectively.

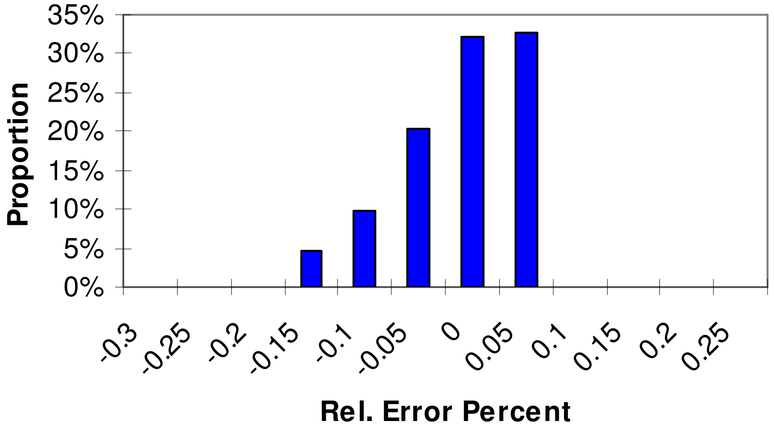


Figure 6.3 Distribution of the relative errors of throughput capacity estimates

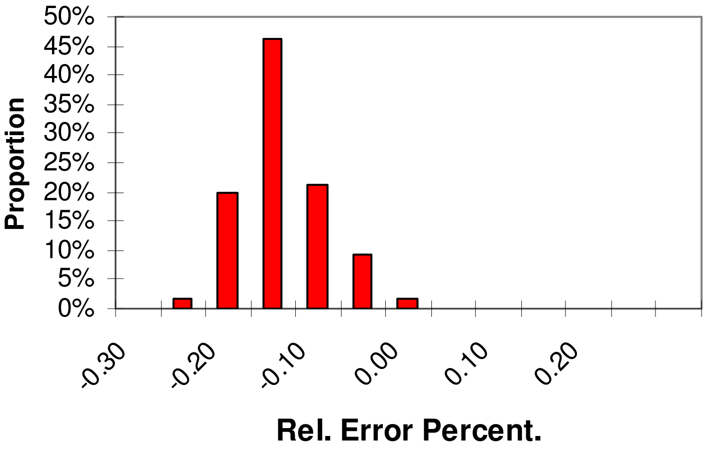


Figure 6.4 Distribution of the relative errors of average response time estimates

We observe that the relative errors in estimating the throughput capacity are mostly between -10% and 10% with few scenarios for which the errors ranged between -10% and -20%. For estimating the average response time, the analytical model mostly underestimates this metric; most of the scenarios generated errors ranging between -20% and -5%.

Figure 6.5 illustrates the impact of the individual factors on the relative error in estimating the throughput capacity.

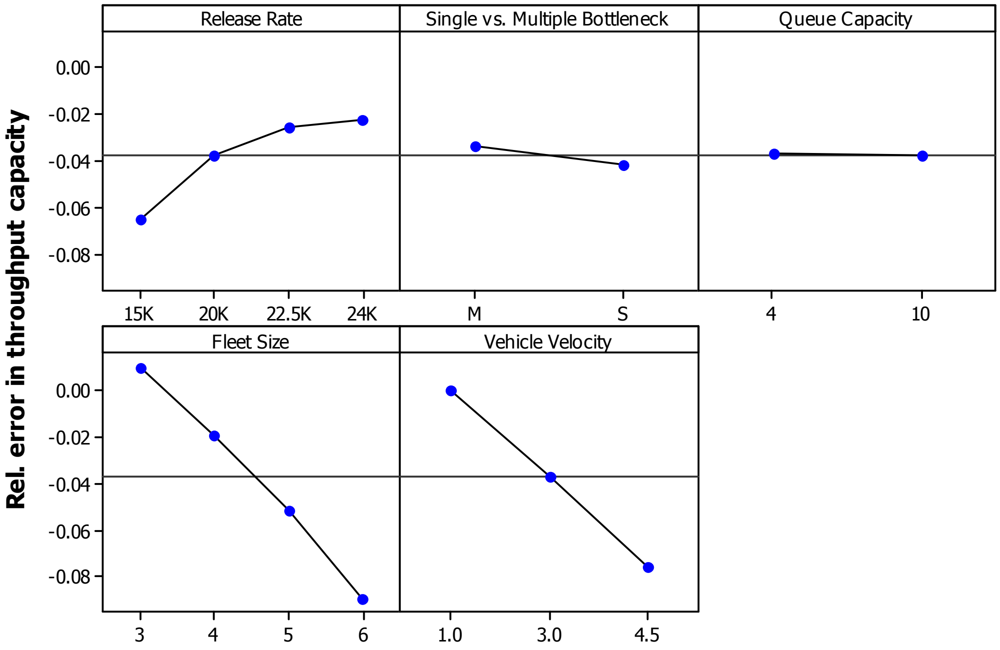


Figure 6.5 Effects of the individual factors on the relative errors in estimating the throughput capacity

Figure 6.5 shows that the analytical model accuracy in estimating the throughput capacity is influenced by both the fleet size and the vehicles' velocity. However, when examining the effect of the vehicles' velocity combined with the fleet size (Figure 6.6), we see the fleet size impacts the throughput capacity error only at high velocity values. In fact, the error is linear in the number of vehicles. Looking back at the reduced-state extended Markov chain model that was developed in chapter four, we made the assumption that the vehicle-blocking probabilities are linear in the number of vehicles. As a result, we observe that as the number of vehicles increase, the analytical model overestimates the percentage of time vehicles are blocked and this is possibly the source of underestimating the throughput capacity at large fleet sizes. To overcome this approximation error, we may consider a different relationship between the number of vehicles and the blocking probabilities. More specifically, the probability that a vehicle is blocked at some station *j* is the probability that there is a vehicle traveling to or receiving service at station *j*+1, or there that there is a vehicle traveling to station *j*+1 *and* another vehicle traveling to or receiving service at station *j*+2 and so forth; the number of terms is the number of vehicles minus one. For instance, for a four-vehicle fleet, the probability that a vehicle gets blocked at station *j*, denoted by p_b^j is:

$$p_b^j = \upsilon^{j+1} + (\upsilon_e^{j+1} + \upsilon_f^{j+1})\upsilon^{j+2} + (\upsilon_e^{j+1} + \upsilon_f^{j+1})(\upsilon_e^{j+2} + \upsilon_f^{j+2})\upsilon^{j+3} \qquad (6.1)$$

Where υ^{j+1} denotes the probability that a vehicle is traveling to or receiving service at station *j*+1 and $(\upsilon_e^{j+1} + \upsilon_f^{j+1})$ is the probability that a vehicle is traveling (empty or loaded) to station *j*+1. Preliminary tests of this alternative model at large fleet sizes indicate a reduction in the throughput capacity estimation errors.

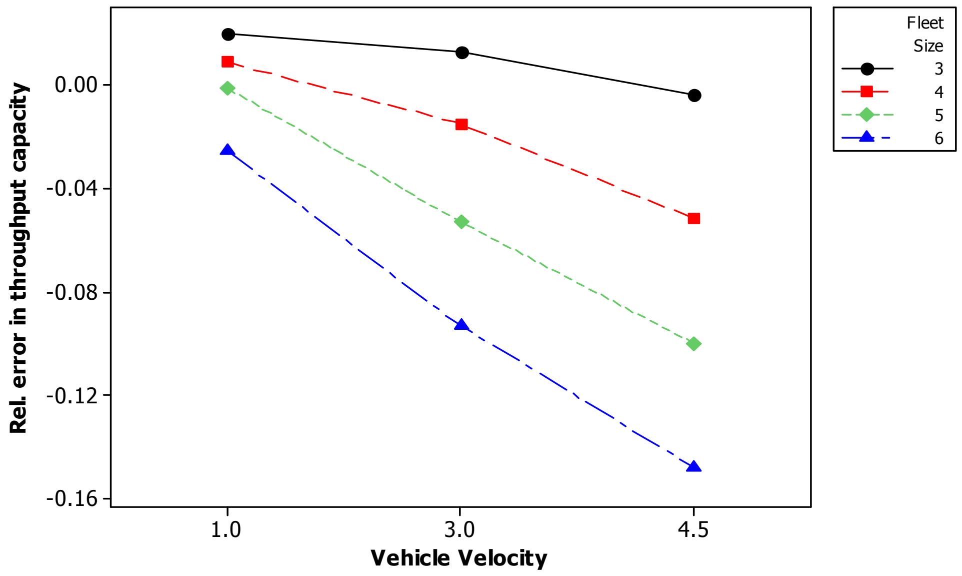


Figure 6.6 Interaction effect of vehicles' velocity and fleet size factors on error in estimating throughput capacity

Figure 6.7 illustrates the impact of the individual factors on the relative error in estimating the average response time. We observe that the analytical model accuracy in estimating the average response time is influenced by the release rate, fleet size and the vehicles' velocity. The analytical model consistently underestimates the response time, and we notice that although the negative difference becomes more significant as the fleet size increases from 3 to 4 to 5, but at 6 vehicles the estimates starts to get closer to the simulation result. The source of this behavior is the throughput capacity estimates. As we see in Figure 6.5, when the number of vehicles increases, the throughput capacity is underestimated. The response time is derived from, and inversely related to, the throughput capacity, and thus when the analytical model is underestimating the throughput capacity, the estimated response time tends to increase.

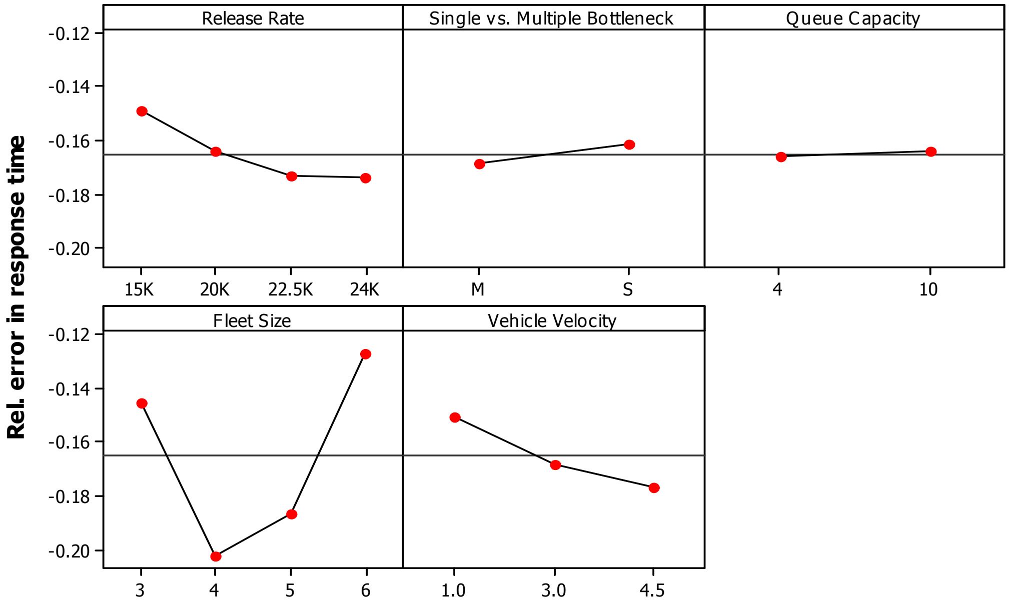


Figure 6.7 Effects of the individual factors on the relative errors in estimating the average response time

We can also argue that even though the relative errors can get large, the absolute errors remain reasonable, for instance when the relative error is -26%, the absolute error is -7 seconds. The designer or analyst who will be using the model will decide whether an error of 7 seconds is acceptable and is worth the significant time reduction achieved by avoiding building a more detailed simulation model. Moreover, when comparing two different designs using the analytic model, the results can be trusted to indicate which design is better even if they do not tell what exactly the performance will be. For instance, we compare two different fleet size scenarios while keeping the other factors at SEMATECH's original factor levels: vehicles are traveling at 3*ft/s*, the release rate is 20,000 wpm, all the processors have finite capacity queues, and the photolithography is the only bottleneck in the bay. If the number of vehicles is three, the expected response time to moves at one of the photolithography tools is 29 *sec*. (analytic estimate) vs. 33 *sec*. (simulation estimate), the relative error is -13%, but both estimates will recommend that the output buffer queue capacity should be at least one lot.

As mentioned earlier in this section, the errors across the machines were quite close except for when the vehicle speed is 1*ft/sec*. We observed that at this low speed, the impact of single vs. multiple bottlenecks factor has a significant impact on the error, and the analytical model works better when the processor tools are equally utilized, and not very well otherwise, regardless of the queue capacity, the release rates or the fleet size. In fact, as shown in Table 6.5, the simulation results are very different when all the tools are equally utilized from when they have different utilization values. The analytical results, however, are not affected by the utilization of the processor tools. For instance, the analytical estimate of the response time for Insp_Ply is 119*sec*, which would lead to a

6% relative error when compared with the multiple bottlenecks scenarios, and to a -17% relative error for the single bottleneck scenario. These differences were not observed at higher travel velocities and this is probably due to the lower utilization of the AMHS at higher speeds. We know from simple factory physics that when the utilization of a machine decreases, the impact of its performance variability decreases and thus at higher velocities, the impact of the AMHS delivery time variability has less impact on the arrival process to the processor tools, which in turn reduces the variability of the departures from the processor tools and so forth, leading to a lower variability all over the system.

The above discussion emphasizes that the analytical model for estimating the response time works best when there is a sufficient level of variability in the arrival process of the vehicles to pick-up or drop-off loads.

Table 6.5 An illustration of the different simulation model estimates of response time at different processor tools utilizations

Release rate	*Single vs. multiple*	*Queue capacity*	*Fleet size*	*Vehicle velocity*	Response time				
					Stocker	*Litho*	*Meas_Overlay*	*Insp_Ply*	*Meas_CD*
24*K*	*S*	10	3	1	97	116	140	143	84
24*K*	*M*	10	3	1	102	123	103	113	109

6.4. Comparison of dispatching rules

To validate the analytical model, we implemented FEFS dispatching policy for the modeled photolithography bay of SEMATECH simulation model. The original policy, however, is modified first come first served (ModFCFS); a centralized policy because, unlike FEFS, it partially depends on the local station information but also on the global waiting list of loads. Specifically, in ModFCFS dispatching, when an empty vehicle arrives at some machine, it searches for any waiting loads at that machine and if it finds a load it picks it up, but if there were no loads waiting at the current machine, the vehicle is assigned to the oldest load in the system even if there is a closer load that can be picked up and dropped off before the vehicle arrives to its assigned load. Although this rule is simple, it is more complicated implementation-wise than FEFS because of its dependency on global information.

Using simulation, we compare the AMHS performance metrics for FEFS and ModFCFS at SEMATECH's original factor levels: three vehicles each traveling at 3*ft/s*, the release rate is 20,000 *wpm*, all the processors have finite capacity queues, and the photolithography is the only bottleneck in the bay.

Simulation results indicate that the AMHS theoretical throughput capacity under ModFCFS policy is higher than under the FEFS policy by 7% (93,700 vs. 87,250 moves/month). The expected response time, however, is significantly lower under the FEFS policy as shown in Figure 6.8 that compares the expected response time at each station under each policy.

The throughput capacity is theoretical because almost certainly it would never be achieved; the current required throughput is around 800 moves/month. The system can never go up to 90,000 moves/month keeping the current bay configuration and production system specifications unchanged; the photolithography equipment currently has 60% utilization and increasing the release rate will require adding more equipment, which will change the configuration of the bay. Therefore, higher throughput capacity does not make the ModFCFS rule superior to the FEFS.

It might be counterintuitive that under FEFS, the AMHS has shorter response times but higher throughput capacity than under ModFCFS. The explanation is as follows: under ModFCFS, the vehicles are dispatched to the oldest load and thus each vehicle makes fewer stops in one loop traversal and so it travels faster and hence the higher throughput capacity. On the other hand, the vehicle might be ignoring loads that could have been picked up and dropped off before the vehicle reaches the oldest load and thus the longer waiting time for these loads inflates the average response time.

Examining the variability of response times, we notice that the FEFS rule has a slightly higher coefficient of variability at 0.55 versus 0.44 under the ModFCFS.

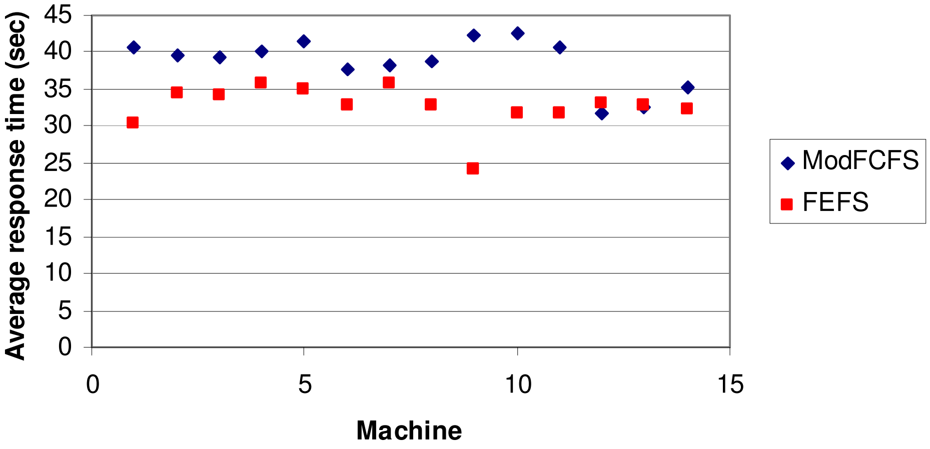


Figure 6.8 Comparison of the expected response time using ModFCFS vs. FEFS dispatching

6.5. Concluding remarks

The analytical model was tested on a realistic data set of SEMATECH hypothetical 300mm fab. The validation of the model was conducted through a comprehensive set of experiments over a wide range of values for the AMHS and the production system parameters that might influence the accuracy of the analytical results.

The comparison results indicated that the analytical model performs very well for estimating the throughput capacity and is reasonably accurate for estimating the expected response time.

The simple decentralized dispatching policy FEFS was compared to a more common centralized policy (ModFCFS), and for the numerical test, FEFS significantly outperformed ModFCFS by generating shorter response times, which leads to questioning the added value of smart yet more expensive and less robust dispatching.

The analytical model is superior to simulation in terms of development and execution times, and is a valuable tool to material handling system designer and fab analysts particularly at the early stages of design. Instead of relying on simulation, this model provides quick estimates of the AMHS operating parameters, and the required on-tool storage capacity as a starting point for building a more detailed simulation model.

CHAPTER 7

CONCLUSION

This research focuses on developing models for analysis and design of automated material handling systems specifically designed for the highly automated 300mm wafer fabrication facilities. An Extended Discrete Time Markov Chain model is developed that estimates the throughput capacity of the AMHS in terms of the number of moves per period that the AMHS can handle. The model considers multiple vehicles operating in simple closed loop configurations, and it considers vehicle-blocking without the need to include detailed AMHS operations.

Two models were developed to estimate the throughput capacity of the AMHS; the first model tracks the movement of every vehicle in the system. Although this model displayed very accurate results in comparison to simulation, the state space of the Markov Chain posed computational challenges for realistic implementations. This computational challenge was the motivation behind developing the second model that tracks a single vehicle with simplifying assumptions on vehicle-blocking.

A closely monitored AMHS performance metric is the expected response time of the AMHS to move requests at pick-up stations. An approach to derive this important metric was developed based on the expected time it takes an empty vehicle to arrive at the location of the waiting load by conditioning on the location of the vehicle at the moment the move request arrived. The derivation is not straightforward and especially complicated for multi-vehicle systems because the vehicle that picks up the load is not necessarily the closest to its location but it is the one that takes the shortest time path.

Experimental comparisons of the throughput capacity and response time models were conducted using a hypothetical 300mm fab from International SEMATECH. The validation of the model was conducted through a comprehensive set of experiments over a wide range of values for the AMHS and the production system parameters that might influence the accuracy of the results. The analytic results were consistent with the simulation results with reasonable error margins, particularly for the throughput capacity estimates. The response time comparisons were acceptable but the model performs better when the arrival rates of move requests is random and error rates are larger when arrival rates of move requests are deterministic.

This research is novel because it is the first to propose an analytical approach to model multi-vehicle material handling systems while considering several practical issues that have not been considered concurrently in the literature. First, we model the state-dependent service rate of move request, whereas, in most analytical models of such systems, the material handling system is modeled by defining a “virtual” workstation between the processing tools in a product’s route. Second, we consider vehicle blocking and the resulting blocking delays in order to get good approximations of both the actual throughput of the AMHS and the average response time to move requests; an issue that is almost always ignored in the available analytical models. Moreover, the response time derivation approach is unique because, unlike conventional models, we assume that the response time of the AMHS to a move request depends on the location of the load, and on the vehicles’ distribution across the network.

The specific contributions in this thesis include:

1. A robust analytic model of simple loop vehicle-based material handling systems that provides fast and accurate estimates of the steady-state performance measures. In the context of the system analyzed in this thesis, the proposed model is accurate for early design phases. There are no analytical models in the literature that simultaneously capture essential aspects of the AMHS: the limited available physical space for vehicles that leads to blocking of vehicles at stations, the inherent queuing in the AMHS due to variability, and the impact of empty vehicle travel.

2. A novel approach based on extended Markov chains, to modeling vehicle-blocking in mutli-vehicle material handling systems. In the rare cases that previously-developed analytic models considered the effect of blocking, it was included as a factor to inflate the capacity that was initially calculated without consideration to blocking. Additionally, the proposed approach holds promise for systems with configurations beyond simple loops and possibly with alternative dispatching policies besides FEFS.

3. A vehicle state-based approach that is also derived from the extended Markov chain model, to approximate material handling systems' response time. In the current literature, reasonably accurate approximations of response times are either developed for single-vehicle systems, which has limited application in practice, or for multi-vehicle systems that rely on modeling the AMHS as a single server, which oversimplifies and does not accurately represent the system described in this thesis. The proposed approximation goes further than single server models and is a significant step towards building more accurate response time approximations for configurations beyond those analyzed in this research.

Although simulation remains the key resource for AMHS design and analysis, the semiconductor industry is in need for analytical models such as the ones proposed in this research to provide an computationally fast and reasonably accurate approach to use in phases of concept development, design, and analysis prior to investing in high-fidelity simulation studies.

Future research

There are a number of issues to be pursued in further research. The analytic model is based on the assumption that machines have separate pick-up and drop-off stations. In some systems, pick-ups and drop-offs can be done at the same point; these systems require smarter control mechanisms. An extension to the model would be to include the possibility of vehicles picking up and dropping off from the same station. Another trivial extension of the proposed model is to relax the "balanced machines" assumptions; the number of pickups equals the number of drop-offs. This is a valid assumption for processor tools but it is not necessarily true for stockers if the bay has multiple stockers.

The current model does not consider a case where an unplanned move to the stockers takes place. For instance, if a lot is ready to be moved to another tool that has a full buffer, the model assumes that this lot will stay in the current tool, which could possibly lead to blocking the processing tool, where it is waiting. In reality, the lot would be moved temporarily to the stocker until its destination tool has space in its buffer. The analytical model may be adjusted to reflect these "induced moves"; the challenge is to estimate the fraction of lots that require this extra workload on the AMHS.

The response time model can be further improved if we incorporate the correlation among the vehicles into the model so that when a vehicle is in some state, the distribution of the other vehicles among the remaining states is adjusted. Currently, we assume that the vehicles are independent so that the location of a vehicle does not impact the other vehicles' locations. Relaxing this assumption would be to condition on the state and location of the closest vehicle to the load (the first in the train of vehicles) and say: if the first vehicle is in state *r* occupying station *j*, then the second vehicle is more likely to be occupying the upstream station *j*-1, and the third vehicle is also closely traveling upstream from the second vehicle and so forth. The average response time calculated using this approach is different from the one we obtain if we assume that each vehicle can be anywhere on the loop.

The response times at different stations are not necessarily equal and when technological constraints allow the fab designer to allocate the processing equipment anywhere on the loop, it would be interesting to look for the best arrangement of the machines along the loop in terms of reducing the response time. A computationally efficient and fairly accurate analytical model would be the preferred tool to aid the designer in evaluating different layouts and possibly finding the optimal arrangement of tools along the loop.

The analytical model is built for a simple loop track. While many systems are simple loops, there are also more general network configurations with shortcuts and spurs. The subject of future work would be to adjust the model for these complex configurations. A simple extension can be implemented for a system that has off-line spurs at each machine, so that a vehicle can be diverted into the station spur if there is a pick-up or drop-off at that station. The extension will require an adjustment to the transition probabilities of the vehicles' states. More complex configurations with shortcuts are more difficult to model because the extension will require the empty vehicles to be routed probabilistically to decide which track the vehicle takes at intersection points. With the FEFS policy used in the current model, empty vehicles are not dispatched to the loads but simply travel around the loop until they encounter a waiting load. It will be interesting to explore the impact of these routing probabilities on the AMHS performance and how machines will be affected differently depending on their location in the network.

We may also consider a modification to the FEFS rule that makes it smarter. Rather than make the vehicle pick-up the load it encounters immediately, the rule can be adjusted so that when an empty vehicle is closely followed by another empty vehicle, the load is not picked up because the second vehicle can pick it up and a blocking situation is avoided. This will be more difficult to analyze but investigating and testing this type of smarter dispatching is the subject of future research.

Many modern fabs are now installing a new concept of storage called "under track storage" (ITRS report, 2005), in which small storage units are added around the AMHS loop and serve as temporary buffers that are closer to the tools than the main stocker. A vehicle can pick-up the load and drop it at the closest under-rail storage location until the destination tool has a space in its buffer to receive the load, in which case possibly another vehicle picks-up the load and delivers it to its destination. For these systems, the analytical model can be used by including the storage locations as additional stockers, but the challenge is to estimate which loads and how often these loads need to be stored at these under-rail storages. This will also be a subject of future research.

REFERENCES

Arzt, T. and Bulcke, F., A new low cost approach in 200 mm and 300 mm AMHS. Semiconductor Fabtech, 1999, available online at http://www.fabtech.org.

Bakkalbasi, O., Flow path network design and layout configuration. Ph.D. Thesis, Georgia Institute of Technology, 1990.

Benjaafar, S., Modeling and analysis of congestion in the design of facility layouts. Management Science, 2002, 48, 679-204.

Bartholdi, J. J. III, and Platzman, L. K., Decentralized control of automated guided vehicles on a simple loop. IIE Transactions, 1989, 21, 76-81.

Bonora, T., and Feindel, D., New tools and fab demand 300mm automation optimization. Solid State Technology, 2001, 44, 87-89

Bozer, Y. A., Myeonsig, C., and Srinivasan, M. M., Expected waiting times in single-device trip-based material handling systems. European Journal of Operational Research, 1994, 75, 200-216.

Bozer, Y. A., Srinivasan, M. M., and Myeonsig, C., Tandem configurations for Automated Guided Vehicle systems and the analysis of Single vehicle loops. IIE Transactions, 1991, 23, 72-82.

Cardarelli G. and Pelagagge P.J., Simulation tool for design and management optimization of automated interbay material handling and storage systems for large wafer fab. IEEE Transactions on Semiconductor Manufacturing, 1995, 8, 1, 44-49.

Curry, G. L., Peters, B. A., and Lee, M., Queueing network model for a class of material-handling systems. International Journal of Production Research, 2003, 41, 3901-20.

Egbelu P.J, The use of non-simulation approaches in estimating vehicle requirement in an automated guided vehicle based transport system. Material Flow, 1987, 4, n1-2, 17-32.

Gaxiola, G., and Mackulak, G., Simulation analysis of a semiconductor handling and processing system: process instability can lead to wasted modeling efforts. Proceedings of the 31st Annual Summer Computer Simulation Conference, 1999, 137-142.

Goetz W. G., and Egbelu P.J, Guide path design and location of load pick-up/drop-off points for an automated guided vehicle system. International Journal of Production Research. 1990, 28, n5, 927-941.

Hodgson, T.J., King, R.E, Monteith, S.K.; Schultz, S.R., Developing control rules for an AGVS using Markov decision processes. Material Flow, 1987, 4, 1-2, 85-96.

Hopp W., and Spearman, M., Factory Physics, 2nd edition, McGraw-Hill, 2000.

ITRS report. The International Technology Roadmap for Semiconductors: Factory integration. 2001, available online at http://public.itrs.net.

Johnson, M. E., Modeling empty vehicle traffic in AGVS design. International Journal of Production Research, 2001, 39, 2615-33.

Johnson, M. E. and Brandeau, M. L, An analytic model for design and analysis of single-vehicle asynchronous material handling systems. Transportation Science, 1994, 28, 337-53.

Johnson, M. E. and Brandeau, M. L, Designing multiple-load automated guided vehicle systems for delivering material from a central depot. Transactions of the ASME Journal of Engineering for Industry, 1995, 117, 33-41.

Jones, S., 300 mm perceptions and realities. Semiconductor International, 2003, 26, 69-72.

Kleinrock, L. Queuing Systems. John Wiley & Sons, 1975.

Kobza, J. E., Yu-Cheng, S., and Reasor, R. J., A stochastic model of empty-vehicle travel time and load request service time in light-traffic material handling systems. IIE Transactions, 1998, 30, 133-42.

Kuhn, A., Efficient planning of AGVS by analytical methods, Proceeding of the 2nd international conference on automated guided vehicles systems, IFS Publications, 1983, p. 1-10.

Lin, J.T., Wang, F.K. G. and Wu, C.K., Connecting transport AMHS in a wafer fab. International Journal of Production Research, 2003, 41, 529-544.

Mackulak, G. and Savory, P., A simulation based experiment for comparing AMHS performance in a semiconductor fabrication facility. IEEE Transactions on Semiconductor Manufacturing, 2001, 14, 273-280.

Mackulak, G., Lawrence, F., and Colvin, T., Effective simulation model reuse: a case study for AMHS modeling. Proceedings of the 1998 Winter Simulation Conference, 979-984.

Mahadevan B., and Narendran T.T, Design of an automated guided vehicle-based material handling system for a flexible manufacturing system. International Journal of Production Research, 1987, 28, 9, 1611-1622.

Mahadevan, B., and Narendran, T.T., Estimation of number of Agvs For an FMS : an analytical model. International Journal of Production Research, 1993, 31, 7, 1655-1670.

Malmborg C.J., A model for the design of zone control automated guided vehicle systems. International Journal of Production Research, 1990, 28, 10, 1741-1758.

Maxwell W.L., and Muckstadt J. A., Design of Automated Guided Vehicle Systems. IIE Transactions, 1982, 14, n2, 114-124.

Nadoli, G. and Pillai, D., Simulation in automated material handling systems design for semiconductor manufacturing. Proceedings of the 1994 Winter Simulation Conference, 892–899.

Peters, B. A., and Yang, T., Integrated facility layout and material handling system design in semiconductor fabrication facilities. IEEE Transactions on Semiconductor Manufacturing, 1997, 10, 360-369.

Pillai, D., Quinn, T., Kryder, K., and Charlson, D., Integration of 300 mm fab layouts and material handling automation. Proceedings of the IEEE/CHMT Ninth International Electronic Manufacturing Technology Symposium, 1999, 23–26.

Rajotia S., Shanker K., and Batra, J.L., Determination of optimal AGV fleet size for an FMS. International Journal of Production Research, 1998, 36, 5, 1177-1198.

Roeder, T., Govind, N., and Schruben, L. A Queuing network approximation of semiconductor automated material handling systems: how much information do we really need? Proceedings of the 2004 Winter Simulation Conference, 1956-1961.

Ross, S., Introduction to Probability Models, 7th edition, Academic Press, 2000.

Rust, K., Wright, R., and Shopbell, M., Comparative analysis of 300 mm automated material handling systems. Proceeding of the Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM), 2002, 240-245.

SIA press release. 2006, Global Chip Sales Hit Record $227.5 Billion in 2005. Available online at http://www.sia-online.org/pre_release.cfm?ID=395.

SEMATECH Report. 1997, Global joint guidelines for 300mm semiconductor factories I300I/J300. Available online at International SEMATECH website http://www.sematech.org.

Sharp G. P., and Liu F. F., An analytical method for configuring fixed-path closed-loop material handling systems. International Journal of Production Research, 1990, 28, 4 757-783.

Sinriech D., and Tanchoco J. M, Intersection graph method for AGV flow path design. International Journal of Production Research, 1991, 29, 9, 1725-1732.

Sinriech, D., and Tanchoco J.M., An economic model For determining Agv fleet size. International Journal of Production Research, 1992, 30, 6, 1255-1268.

Srinivasan, M. M., Bozer, Y. A., and Myeonsig, C., Trip-based material handling systems: throughput capacity analysis. IIE Transactions, 1994, 26, 70-89.

Steele, J., An algorithm for estimating the performance of an automated material handling system for the semiconductor industry. Proceeding of the Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM), 2002, 229-234.

Tanchoco J.M., Egbelu P.J., and Taghaboni F., Determination of the total number of vehicles in an Agv-based material transport system. Material Flow, 1987, 4, 1-2, 33-51.

Ting, J. H., and Tanchoco, J. M, Unidirectional circular layout for overhead material handling systems. International Journal of Production Research, 2000, 38, 3913-3936.

Ting, J. H., and Tanchoco, J. M, Optimal bi-directional spine layout for overhead material handling systems. IEEE Transactions on Semiconductor Manufacturing, 2001, 14, 57-64.

Wolff, R. W., Poisson Arrivals See Time Averages. Operations Research, 1982, 30, 223-231.

Vis, I. F., Survey of research in the design and control of automated guided vehicle systems. European Journal of Operational Research, 2006, 170, 3, 677-709.

VITA

DIMA NAZZAL

Dima Nazzal was born in Damascus, Syria on November 15, 1976. After graduating from high school in 1993, she studied at the University of Jordan and obtained a Bachelors Degree in Industrial Engineering in 1998. In 1999, she joined the M.S. program in Industrial Engineering and Management Systems at the University of Central Florida and obtained the Masters Degree in 2001. In 2002, she joined the Ph.D. program in Industrial and Systems Engineering at Georgia Tech, she earned her Ph.D. in 2006.