Improving and Maintaining the Operational Efficiency of a Semiconductor Equipment Manufacturing Warehouse

by

Hui Ni Grace Fong

B.Eng. Materials Science & Engineering (2014)

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Master of Engineering in Manufacturing

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

Author .... Signature redacted ....
Hui Ni Grace Fong
Department of Mechanical Engineering
August 10, 2015

Certified by .... Signature redacted ..
Stephen C. Graves
Abraham J. Siegel Professor of Management Science
Thesis Supervisor

Accepted by .... Signature redacted ....
David E Hardt
Chairman, Committee for Graduate Student

Improving and Maintaining the Operational Efficiency of a Semiconductor Equipment Manufacturing Warehouse

by

Hui Ni Grace Fong

Submitted to the Department of Mechanical Engineering
on August 10, 2015, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Manufacturing

Abstract

The present work addresses an operational inefficiency problem at a semiconductor equipment manufacturing warehouse, Varian Semiconductors Associates and Equipment (VSEA). This problem is important because if unresolved, the warehouse is unable to meet the part delivery time target of 24 hours during the busy period. The downstream effects of the late part delivery are delayed production schedules and in the worst case scenario, a missed shipment to the customer, which is very costly. In order to improve the efficiency of the warehouse so as to consistently deliver parts on time, the picking efficiency needs to be enhanced. Parts are primarily picked from two types of storage locations - GL shelves and Vertical Lift Modules (VLMs). The picking efficiency can be improved by the simultaneous reduction in workload on GL and improvement in the VLM picking efficiency. The first part of this thesis focuses on improving the picking inefficiency at the VLMs by employing a more efficient picking method. From our time study, we find that the pick-and-consolidate (parallel picking) is more efficient than pick-and-pass (sequential picking). The average makespan time savings per order by pick-and-consolidate is 8% (20 minutes). The second part of this paper discusses what is required to maintain a high VLM picking efficiency. New metrics to measure the workload distribution of the VLMs and the average flow time per order are proposed. Three dynamic slotting methods that maintain a balanced workload distribution across the VLMs without the need for periodic review are also examined. The methods are evaluated based on how balanced is the workload distribution across the VLMs and the cost of implementation.

Thesis Supervisor: Stephen C. Graves
Title: Abraham J. Siegel Professor of Management Science

Acknowledgments

I would like to express my heartfelt gratitude to the faculty of the Master of Engineering in Manufacturing program for giving me the opportunity to pursue my Master at MIT.

I am deeply privileged to have the chance to further my learning at Varian Semiconductors Equipment and Associates, a business unit of Applied Materials Inc. I would like to thank Dan Martin, my supervisor, for his mentorship and encouragements throughout the project. I am especially grateful to him for giving my team the flexibility to select a project that interests us most. I would also like to thank Thomas Faulkner, Ronald Dognazzi, Gaetano Peritore, Mark Goodhue, Robbie Roberts, Anthony Lafata, Maria Doyle and James Firth and everyone else at VSEA for supporting us throughout our project and making this project possible.

I would like to thank Professor Stephen C Graves for his mentorship and support. His guidance and advisory has made sure that not only a good solution is provided technically, but the solution is also practical and can be implemented within the company. I would also like to thank Professor Sean Willems for his valuable feedback for our project.

I would like to thank Applied Materials (Singapore) for their support and sponsorship for my Master at MIT.

I am very grateful to work with my teammates, Stephen Racca and Paramveer Singh Toor, who have made our collaboration fruitful and enjoyable. I would like to thank Stephen Racca for his constant support. I cannot even begin to articulate how much I have learned from him over the past one year.

I would like to thank all my friends who have made this journey a memorable and fulfilling one. I would like to thank Ng Yee Sian for helping me develop a simulation program for my thesis work at the initial phases of our project.

Last but not least, I would like to thank my family for their love and endless support along my life's journey. Without them, I would not have come this far.

Contents

List of Figures

List of Tables

This page intentionally left blank

Chapter 1

Introduction

1.1 Problem context

In the semiconductor manufacturing equipment industry, late deliveries to the customers are costly. In order for customer orders to be delivered on time, semiconductor equipment manufacturers need an efficient flow of material from the warehouse to the production floor. However, maintaining an efficient material flow is a challenge for companies in this industry because of the extremely high number of stock keeping units (SKUs) in the warehouse. This is not surprising since semiconductor capital equipment is highly complex and made up of at least a hundred thousand of parts.

In this paper we examine a warehouse operational inefficiency problem at a semiconductor capital equipment manufacturing facility, Varian Semiconductor Equipment and Associates (VSEA). VSEA manufactures ion implantation equipment and also provides spare parts to its customers. For this company, as for many other warehouses that manages a high amount of stock keeping units (SKUs), maintaining an efficient material flow and delivering parts on time from the warehouse to the production floor are key issues.

At VSEA, the target part delivery time from the warehouse to the production floor is 24 hours. However, during the busy periods, parts are not delivered on time to the production floor as the warehouse struggles to keep up with the high demand of parts required by the production floor and their customers. During the last busy period between November 2014 and January 2015, the time to deliver the parts to the production floor was above the target of 24 hours. The average time to completely pick and consolidate an order was 74 hours. In addition, up to 40% of the orders were not completed within the goal of 24 hours.

Meeting the target part delivery time is critical for the following reasons. First, if parts are not delivered on time, a machine lay-down is delayed. This may cause major backups in the production line, which can further lead to missed shipments to their customers. A missed shipment is highly unfavorable as the company has to expedite the missed shipment. Furthermore, each shipment is worth at least a million dollars and a missed shipment may mean lost in potential sales revenue. Moreover, the industry is primarily customer driven with a few key customers. VSEA maintains its competitive advantage by providing a high service level, which includes delivering orders on time. The greatest risk of a missed shipment is unhappy customers and ultimately, losing the customer.

Second, delivering parts on time minimizes the work-in process (WIP) inventory on the production floor. For example, if a machine has been laid down and additional parts to continue building the machine do not arrive on time, then the machine is lying idle on the production floor. WIP inventory is not only costly, but also takes up valuable space on the production floor.

Third, VSEA has set the part delivery time goal to be 24 hours so as to be more responsive to changes in customers' requirements. Customers routinely make configuration changes and VSEA mitigates the risk of such changes by pushing the actual production to the latest possible date. If the goal was set to a value above 24 hours, VSEA would face a greater risk of changes in production build orders. For any changes in an order, the order has to be re-picked from the warehouse, which delays the build schedule on the production floor.

Last but not least, besides meeting the part delivery time goal, maintaining consistency in part delivery times is also critical. This helps to make delivery times predictable, which provide the production floor a precise idea of when they can expect parts to be delivered and thus allow them to develop more accurate production plans. During the busy period, it is known that production schedulers order parts earlier in an attempt to compensate for the anticipated late deliveries. Therefore, maintaining predictable delivery times also prevents the production floor from ordering parts way ahead of lay-down date, which reduces WIP inventory on the production floor.

In summary, our work focuses on improving the operational efficiency of the warehouse at VSEA where the objective is to consistently deliver parts on time to the production floor during the busy period. In this paper, efficiency is measured by the time taken to deliver the order to the production floor. The process of fulfilling warehouse orders involves five steps: releasing the orders to the pickers, picking, consolidation of parts arriving from three distinct storage locations, staging for shipping and finally, shipping. Through our preliminary studies, we identified that the primary operational inefficiency stemmed from picking.

In particular, the picking inefficiency mainly originates from two distinct storage locations namely, GL and Vertical Lift Module (VLM). Whereas parts in GL are manually picked from shelves, a VLM is an automated storage and order picking system that bring parts to the pickers. There are currently five VLM pods in the warehouse and each pod has one operator picking from the VLMs (refer to chapter 2 for a more detailed explanation of the storage locations in the warehouse). More importantly, picking from VLMs is faster than picking from GL shelves because pickers picking from VLMs do not need to walk along aisles or climb ladders to pick from shelves. Furthermore, the expected pick rate of the VLMs is 60 picks per hour, which is three times faster than that of GL. The management bought the VLMs precisely because they wanted to improve the picking efficiency of the warehouse so as to deliver parts on time. However, the picking efficiency of the warehouse has been falling short of expectations. The next section examines the causes of inefficient picking at the warehouse.

1.2 Analysis of Problem and Approach

A warehouse order requires picks from three picking locations before consolidation of the order can begin. From table 1.1, picking from GL is the bottleneck of the picking process. During the busy period, it took 50 hours on average to complete picking parts for an order from GL, whereas picking parts for the order from the VLMs required an average of 42 hours. The total average values are obtained from 42,881 historical orders. Orders that require more than 3 days to complete, are eliminated from the 3 day cut off data; and orders that require more than 1 day are eliminated from the 1 day cutoff. For example, the average time to pick orders from the VLMs, that are delivered within 24 hours, is 7 hours. The percentage of orders that are not picked from the VLMs within 24 hours is 23%.

Table 1.1: Time Study results from busy period. Cutoffs ignore parts outside longer then stated in that area and percent excluded shows the percentage of order that was excluded. The amount of time total pick time is assumed to be $Max(PickTime) + Consolidation\ time$

	Total Average (hours)	3 Day Cutoff (hours)	1 Day Cutoff (hours)	3 Day % Excluded	1 Day % Excluded
VLM	41.9	11.8	7	13	23
GL	49.5	19.1	9.4	16	40
RK	27.9	12.7	8.2	8	20
CO80	24.3	3.1	1.1	11	15
Total Pick Time	73.8	22.2	10.5		

Figure 1-1 summarizes the reasons associated with the picking inefficiencies, which resulted in the late delivery of parts. The two major causes of picking inefficiency are a low number of VLM picks and an inefficient VLM system. It is worth noting that the low number of VLM picks means that parts, which could have been picked from the VLMs, are picked from GL instead. It is desirable to pick more from the VLMs rather than GL because picking is more efficient on the VLMs than on GL. An individual can pick 60 items per hour on the VLMs, but only 20 on GL.

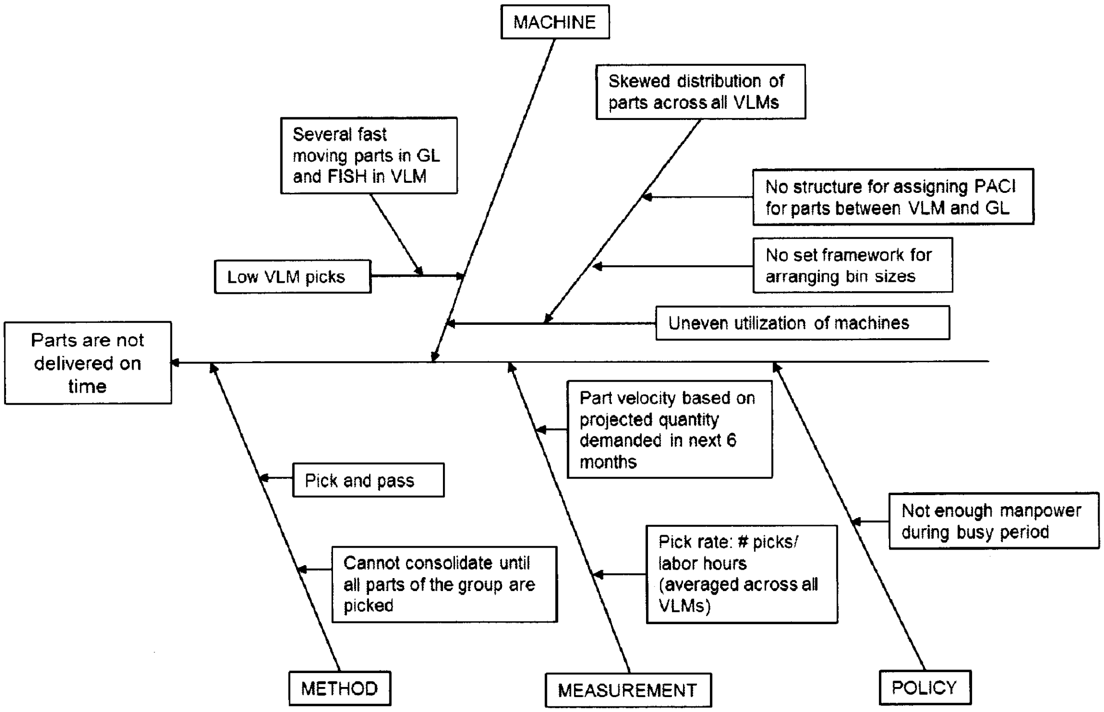


Figure 1-1: Fishbone analysis used for problem identification

First, from figure 1-2, the increase in number of VLM picks is slower than that of GL picks during the busy period between December 2014 and January 2015. Moreover, more parts are picked from GL than the VLMs. This is partly because of a high number of slow moving parts in the VLM. More than 18% of parts in the VLMs were not picked in the past one year and were not projected to be picked in next six months. Furthermore, some parts in the VLMs have not been moved since 2012 and these parts are known as 'FISH' (First In Still Here) in the company. Typically, these FISH parts are spare parts for older product models and thus low in demand. Therefore these FISH parts do not contribute to number of VLM picks and occupy valuable real-estate. On the other hand, there is a large number of fast moving parts in the GL picking area.

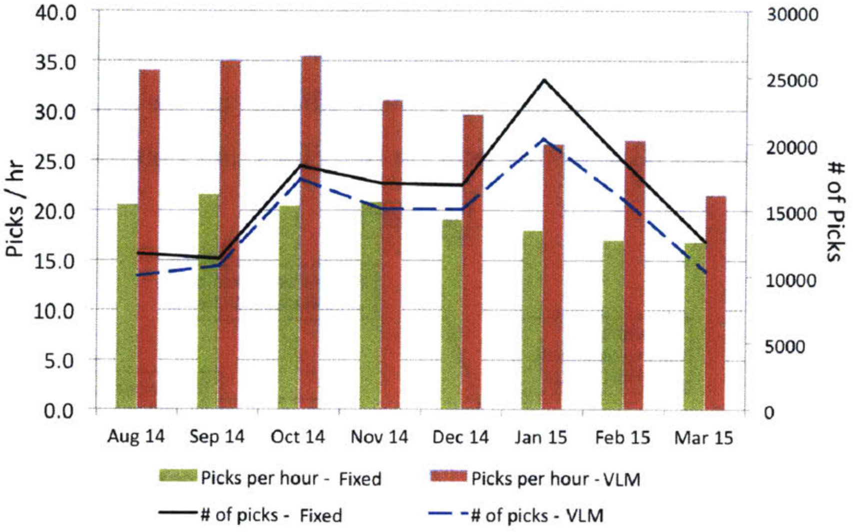


Figure 1-2: Picking rates and total picks from 9/14/2014 to 3/15/2015. Pick rates correspond to columns and left axis and number of picks corresponds to the overlaid lines and the right axis. The fixed picks are picks from manual picking locations i.e. GL and High Racks.

Second, the VLM pick rates shown in Figure 1-2 are steadily declining. In addition, the highest pick rate displayed was 35 picks per hour a mere 58% of the original estimate pick rate of 60 picks per hour. The current operation of the VLMs is inefficient because the workload distribution across the five VLM pods is skewed, primarily towards the first pod (see figure 1-3). Next, the picking strategy also relates to how fast an order can be completely picked from the VLMs. At VSEA, there are two different picking strategies: pick-and-pass, and pick-and-consolidate. Pick-and-pass is identified as a less efficient method of picking parts than pick-and-consolidate. However, pick-and-pass is practiced more often than pick-and-consolidate.

In order to resolve the bottleneck problem at GL, we propose to reduce the workload on GL by increasing the workload on the VLMs, while ensuring that the picking time at the VLMs is minimized. In essence, the objective of reducing the workload on GL is to ensure that an order can be completed faster at a low cost. This can be achieved by transferring fast moving parts from GL to VLMs. However, the current operation

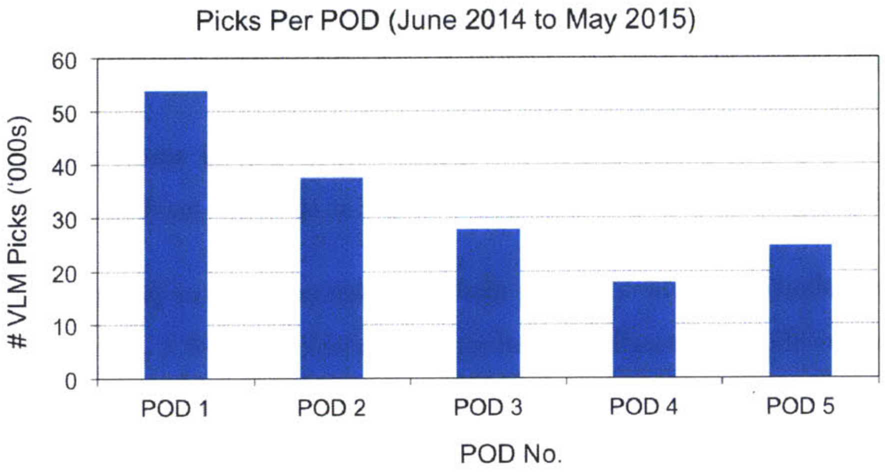


Figure 1-3: Total picks per POD from June 2014 - May 2015 units in thousands. Note: POD 1 has 2 times the picks of POD 3

of the VLMs is inefficient. Unless the VLMs are operating efficiently, the improvement in time taken to complete picking orders by reducing the workload on GL is minimal. In order to ensure that the picking efficiency of the warehouse is improved by more than marginal so that parts are consistently delivered within 24 hours during the busy periods, it is necessary to also address the operational inefficiency of the VLMs.

1.3 Summary of Contributions

Our present work addresses the operational inefficiency of the warehouse through enhancing the picking efficiency. The proposed solution is to increase the number of VLM picks and reduce the workload at the GL shelves. This is because picking from VLMs is a more efficient process than picking from shelves. However, the current operational inefficiency at the VLMs limits the effectiveness of our proposed solution. Therefore, it is of interest to improve the operational efficiency of the VLMs. This is achieved by balancing the workload across the VLMs and employing an efficient picking method.

The three main contributions of our work are:

1. A practical solution that was implemented at VSEA to increase the number of VLM picks and reduce the workload at the GL shelves
2. An analysis of three methods to balance the workload across the five VLMs, including a final recommendation to VSEA as to which method is most feasible
3. An evaluation of which picking method – pick-and-pass or pick-and-consolidate is more efficient, including a final recommendation to VSEA as to which method should be employed during the busy periods

They are addressed in three separate thesis:

The first thesis [16] goes in greater depth explaining the problem identification process. It focuses on moving and identifying 2,052 FISH parts (parts that are not picked in past one year and are not projected to be picked in next 6 months) to GL area in the most efficient manner possible. It also further identifies fast moving parts in the GL area that could be fit into VLM and allocating these fast movers to specific VLMs to balance the workload distribution of VLMs. It also identifies the fast moving GL parts that could not be fit into VLMs so that they could be put into easily accessible slots in GL (also known as golden picking zones). Finally, the thesis discusses how the implementation process was carried out.

The second thesis [15] concentrates on balancing the workload distribution of all five VLMs to maximize the throughput of the VLMs. The thesis further evaluates three different part allocation strategies: snake frequency balancing, randomization, and order grouping. The snake frequency balancing strategy distributes parts in VLMs based on their pick frequencies; the randomization strategy randomly distributes parts amongst the VLMs; and the order grouping strategy arranges parts that are ordered together, evenly among all five pods. A discussion on which strategy gives the greatest improvement in balancing the workload across the VLMs and pick time savings is provided.

The long delivery times are addressed by the simultaneous reduction in workload on GL and improvement in the picking efficiency at the VLMs. The first part of this thesis focuses on improving the picking inefficiency at the VLMs by employing a more efficient picking method. A time study is performed to evaluate which is a more efficient picking method, pick-and-pass (a sequential picking process that does not require VLM consolidation) or pick-and-consolidate (a parallel picking process that requires VLM consolidation). By employing a more efficient picking method, the average makespan per order at the VLMs can be reduced. The second part of this paper discusses what is required to maintain a high VLM picking efficiency. New metrics to measure the workload distribution of the VLMs and the average flow time per order are proposed. Three dynamic slotting methods are also examined in terms of how even is the workload distribution across the VLMs and the costs involved in implementing the solution within the company.

1.4 Thesis Organization

Chapter 2 presents an overview of the semiconductor industry and the company. It also details the current operations of the warehouse. Chapter 3 examines the picking performances of the two picking strategies – pick-and-pass and pick-and-consolidate. The chapter concludes by providing a final recommendation that pick-and-consolidate is more efficient and thus should be employed during the busy period. Chapter 4 discusses the sustainability of maintaining an efficient VLM through employing effective performance measurement tools and balancing the workload across the five VLMs via dynamic slotting. The conclusion and future work are provided in chapter 5.

This page intentionally left blank

Chapter 2

Background Information

2.1 Semiconductor Industry Overview

The semiconductor industry is a key enabler and driver for technological progress. Our smartphones, computers and electronic gadgets are running at greater processing speeds, larger memories with additional functions. In particular, advancements in sensing modules in terms of greater data storage capacity, enhanced data transmission and processing capabilities, have enabled the creation and development of Internet of Things (IoT) devices which consumers and businesses use to monitor, track and process data. In fact the IoT market is driven by cloud computing and the growing interconnectedness of machines and electronic devices [1].

The semiconductor value chain is summarized in Figure 2-1. The industry can be categorized into several semiconductor industries: semiconductor materials and equipment suppliers, semiconductor foundries, integrated device manufacturers (IDMs), semiconductor electronic design and manufacturing services and original equipment manufacturers (OEMs) [2]. OEMs produce electronic devices and hardware that are sold to end users and examples of OEMs include Apple, Dell, Cisco and Seagate. Semiconductor electronic design and manufacturing services include electronics manufacturing service (EMS) and original design manufacturer (ODM) which test, manufacture and distribute electronic components for OEMs. IDMs are semiconductor companies that design, manufacture and sell integrated circuit devices. The key IDM players are Intel, Samsung Electronics, Qualcomm, Micron Technology and SK Hynix [3]. Semiconductor foundries focus solely on mass producing chips and the top two players in this sector are Taiwan Semiconductor Manufacturing Company (TSMC) and United Microelectronics (UMC) [4]. Semiconductor manufacturing is a highly complex process and requires specialized equipment. The major semiconductor capital equipment manufacturers include Applied Materials, ASML, Tokyo Electron, KLA-Tencor and Lam Research [5]. Essentially, semiconductor capital equipment manufacturers' customers are semiconductor foundries and IDMs.

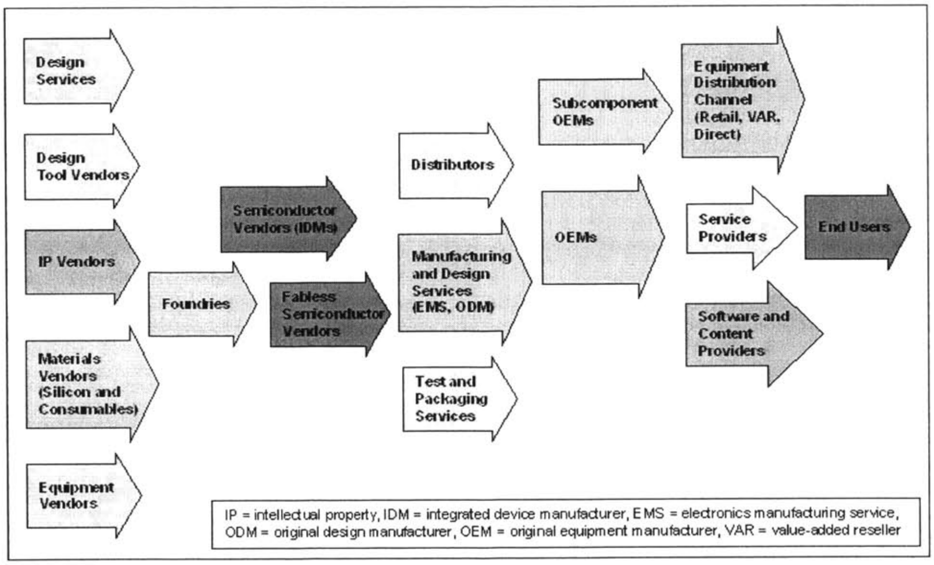


Figure 2-1: Semiconductor value chain, Applied Materials fits into the chain in the bottom left under equipment vendors

In 2014, the worldwide semiconductor market revenue was \$340.3 billion [6] and the global semiconductor capital equipment spending was \$65.3 billion [7].

2.2 Varian Semiconductor Equipment Associates Background

Varian Semiconductor Equipment and Associates (VSEA) is a wholly owned subsidiary of Applied Materials[1] based in Gloucester, Massachusetts that specializes in designing and manufacturing ion implantation[2] equipment. VSEA was incorporated in 1999 and acquired by Applied Materials in 2011.

The four main product lines are medium current, high current, high energy and ultra-high dose. A detailed description of the different production lines can be found in previous research work [8, 9]. These products are sold in low volume (not more than 300 tools a year) but highly customizable. Thus, there are a high number of stock-keeping units (SKUs), over 20,000 SKUs. There are three warehouses known as Buildings 5, 70 and 80. The majority of the parts are stored in Building 80. Building 5 is the production floor and main building and it has two primary storage locations, the Supermarket (SMKT), which store parts required for sub-assemblies, and Module (MOD) which houses larger components and material kits required to build a module. The bulkiest parts are stored in Building 70; these parts include enclosures, large paneling and other protective equipment typically only installed once the machine is delivered. Improving the operations in Building 80 will be the primary focus of this thesis.

---

[1] Applied Materials, Inc., founded in 1967 in Santa Clara, California, is the leading producer of semiconductor equipment, services and software for the global semiconductor, flat panel display, solar photovoltaic and related industries [4]. It offers a wide portfolio of products such as chemical vapor deposition, physical vapor deposition, etch and ion implantation.

[2] Ion implantation is the fundamental process to fabricate semiconductor devices. It is a process of introducing dopants in the wafer. Ions from an ion source are electrostatically accelerated to high energy (10 – 500 keV) and implanted onto the wafer target. The ion implantation process is characterized by the dose and penetration of the dopant and these characteristics are dependent on the ion beam current.

2.3 Current Operations of the Warehouse

The workload in the warehouse fluctuates significantly on a daily basis as shown in figure 2-2. The days that have near zero number of picks correspond to the weekends as the warehouse has a 5-day week. The mean daily workload is 500 picks per day.

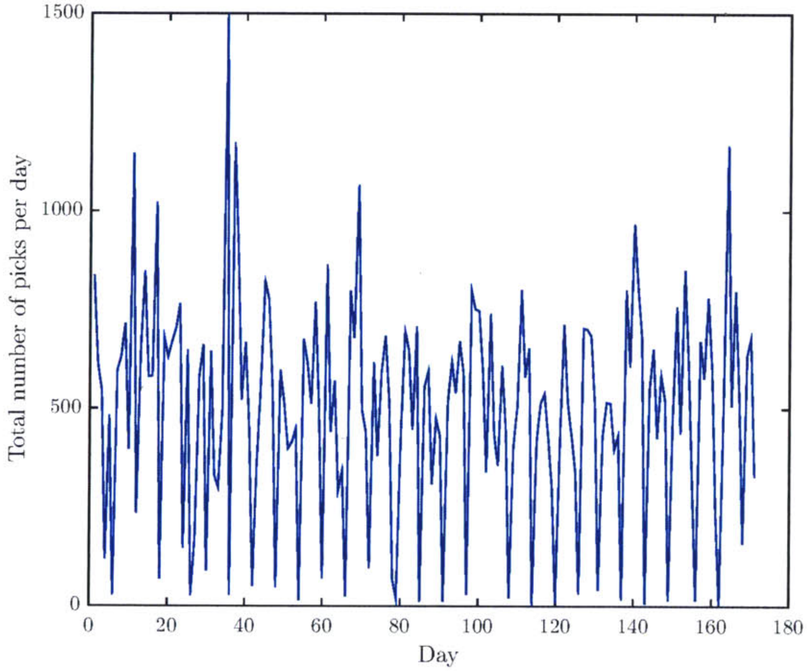


Figure 2-2: Workload daily plot between January and July 2015

Material flows through the warehouse through three main stages starting from receiving, picking and consolidation to shipping. Each step is being monitored by scanning bar-codes. The information is managed using a software known as SAP Extended Warehouse Management (EWM). Parts are first received at the receiving area, sized accordingly and if new, assigned a default Put Away Control Indicator (PACI)[3], and then put away into their respective storage locations. When an order is released to the

---

[3] A PACI tells EWM where to start looking for a open bin during the part put-away search, in the VLMs a sequential numerical order is followed (e.g. PACI of VL01 results in a put-away search sequence of VL01-01→ VL01-02 → VL01-03 → VL02-01 → ... etc.)

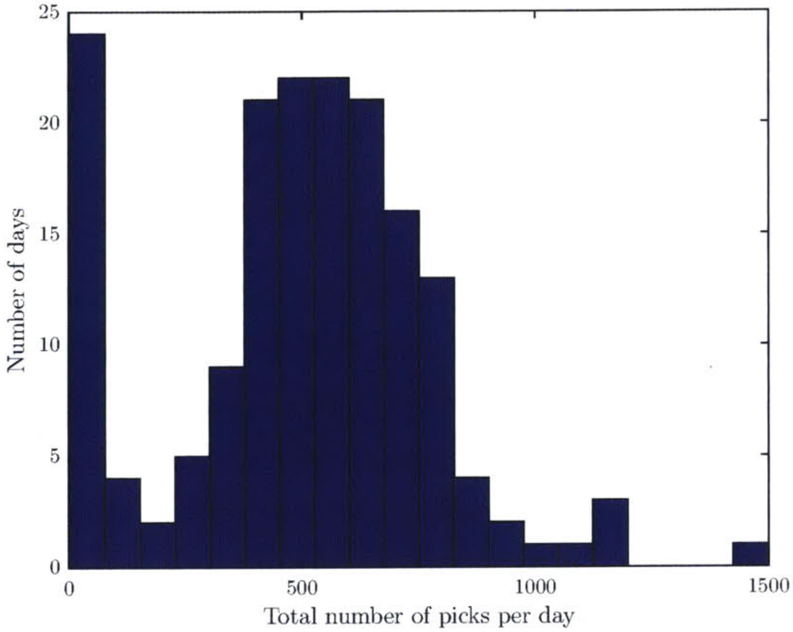


Figure 2-3: Workload histogram

warehouse, parts are picked from their respective storage locations. After the picks have been completed, they are consolidated[4] in the consolidation area. The consolidated pallets or boxes of parts are moved to the loading dock for shipping.

There are three main types of warehouse orders namely sales, production and transfers. A production order can either be classified as shop order or z-pick. A shop order is a list of assemblies to be built with a detailed bill of materials required for each assembly. Parts required from the warehouse are ordered 24 hours in advance of a laydown[5] using z-pick kit codes. Whereas production orders are typically delivered to the assembler on the production floor in the Main Building, transfers are parts issued from the warehouse to replenish the supermarket inventory[6] in the Main Building. Sales orders are replacement parts shipped directly to the customer [10, 9]. The focus

---

[4] Consolidation involves both physically grouping parts as well as electronic grouping of the Handling Units (HUs) of each part. In EWM the parts are grouped in a hierarchal structure using one as a master HU.

[5] Laydown signifies the start of building a new machine. Typically 3-4 laydowns occur per week.

[6] The supermarket inventory is an inventory of parts for sub-assemblies that are assembled into sub-modules of the ion implantation equipment.

of this project is on improving the flow of material from picking to consolidation and two key aspects that affect the material flow are where the parts are picked from i.e. part storage locations and how they are picked i.e. picking methods.

2.3.1 Storage Locations

Materials in the warehouse are picked from three distinct storage locations namely High Rack, GL and Vertical Lift Module (VLM) as illustrated in Figure 2-4. High Rack contains bulk parts which are picked with the aid of cherry pickers and forklifts. Parts in GL are less bulky and manually picked from shelves. VLMs are automated storage and order picking systems where they mechanically eject trays for parts to be manually picked from (see Figure 2-5a). A VLM is made up of vertically arranged trays for storage, a delivery lift platform mechanism and a computerized control system. A tray is automatically brought to the access area (picking zone) upon request from the software and a location indicator guides the picker on the part to be picked. The operator picks or replenishes the stock and then the tray is returned to its original position after confirmation. Figure 2-6 highlights the ergonomic differences in picking from VLM and GL; the golden zone is the optimal picking location where the most efficient picks are made.

At VSEA a single VLM machine is known as a bay and a pod comprises of a set of bays grouped together and integrated as one entire setup (see Figure 2-5b). While the term 'VLM' can be used to describe either a pod or a bay, 'VLM' refers to a pod in the remaining parts of this thesis in accordance with the terminology used at the company. The terms 'VLM' and 'pod' are used interchangeably in this thesis. There are five pods at VSEA and one operator is assigned to every pod. Four of the pods (Pods 1, 2, 3 and 5) have three bays each and one pod (Pod 4) has two bays. There are 40 – 50 trays per bay and each tray (96" x 24") can fit 144 BA04 type (smallest sized) bins [7]. There are seven distinct bin sizes (see Table 2.1) and a total

---

[7] 24 bins along the length and six bins along the width of one tray

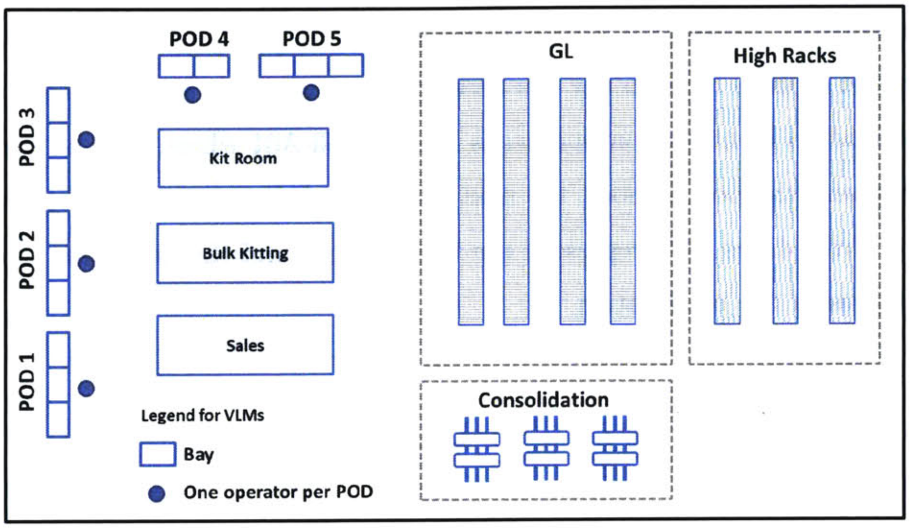
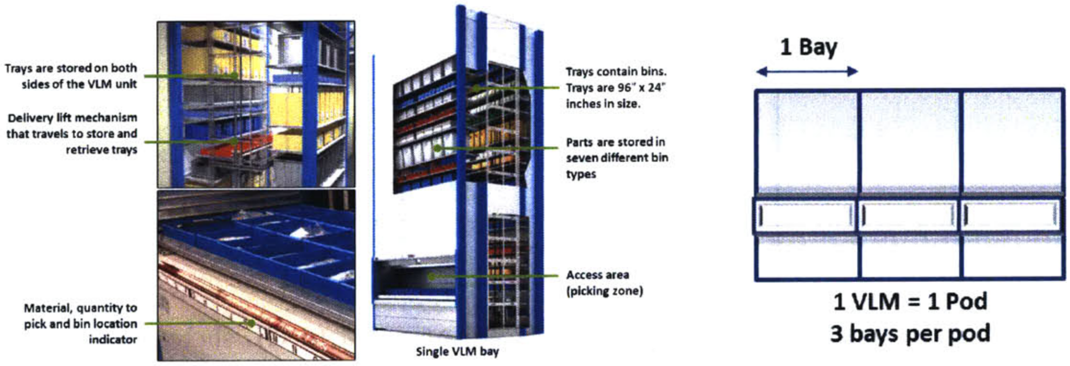


Figure 2-4: Warehouse Floor plan shows various picking areas, Kitting rooms, and consolidation areas.



(a) (b)

Figure 2-5: (a) Description of a single VLM Bay; (b) VLM terminology: VLM, pod and bays

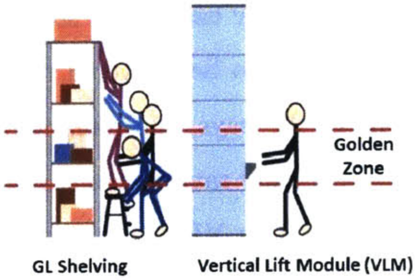


Figure 2-6: Shows the ergonomic differences between GL and VLM picking and the optimal picking zone

of approximately 25,000 bins across all five pods. The number of bins, types of bins and number of trays for each bay can be uniquely configured. An example of a VLM storage location identification number is VL01-02-03-A01 where the first two digits '01' represent the pod number, the next two digits '02' represent the bay number, the next three digits '03' represent the tray number and the last three characters 'A01' represent the bin location.

$$\underbrace{VL01}_{\text{POD \#}} - \underbrace{02}_{\text{Bay \#}} - \underbrace{03}_{\text{Tray \#}} - \underbrace{A01}_{\substack{\text{Grid}\\ \text{Location}}} \tag{2.1}$$

Table 2.1: Bin sizes used in the VLMs for part storage

Bin type	Width	Length	Height
BA04	4	5.5	4
BB04	4	11	4
BC04	8	11	4
BD04	8	16.5	4
BF04	8	33	4
BG04	12	16.5	4
BG08	12	16.5	8

There are a couple of advantages of VLM as compared to GL shelves which have led to a push for increasing the number of VLM picks at the warehouse in recent months. The primary benefit of picking from VLM is that the travel time associated with walking along aisles to pick the parts from shelves is eliminated since the VLM bring the parts to the pickers. Typically each VLM setup has multiple bays so as to reduce the waiting time for the parts to be delivered to the access area [11]. While the picker is picking from one of the bays, the remaining bays search and retrieve the next part to be picked. This time savings translates into higher pick rates at the VLM than GL. The pick rate from VLM is estimated to be 60 picks per hour (approximately a minute for one pick) while GL averages 20 picks per hour based on historical data. In fact, the VLM manufacturer claims that the ideal pick rate of the VLMs is 100 picks per hour. The higher pick rate is critical for delivering parts on time since the faster parts can be picked, the sooner orders can be consolidated to be ready for shipping.

In addition, unlike picking from the shelves of GL which may require pickers to climb ladders, picking from VLMs is safer since the parts are brought to the picker at the access area. Last but not least, VLMs have a greater storage capacity of up to 85% than GL shelves since parts are stored vertically in VLMs.

2.3.2 Pick-Consolidate Flowchart and Picking Methods

The pick-consolidate process is summarized in Figure 2-7. The warehouse first receives orders either from the production floor or sales and a designated employee releases the parts periodically as pick waves to the warehouse pickers. The parts are then picked in parallel from the three storage areas. VLM consolidation involves grouping the parts picked from different VLMs that belong to the same consolidation group. Depending on the picking method employed, VLM consolidation may or may not be required. A consolidation group is a group of parts that has to be delivered as one package. For production orders, a consolidation group typically corresponds to z-pick kit codes or shop orders; for sales orders, a consolidation group typically corresponds to GS kits. Similarly GL consolidation and High Rack consolidation involves putting together parts of the same consolidation group at their respective consolidation areas. Once all the parts of a consolidation group from the three different picking areas have been picked, they are packaged together in a consolidation area and finally staged for trucking.

There are two VLM picking methods namely, pick and pass and pick and consolidate. In the former method, parts are picked sequentially and a tote is passed along the pods as shown in Figure 2-8a. Parts for a consolidation group are put together in a single tote as the tote is passed along and hence, no VLM consolidation is required. In contrast in the latter case, parts of a consolidation group are picked in parallel from the pods separately as illustrated in Figure 2-8b and thus arrive to the VLM consolidation racks in up to five different totes (since there are five pods). VLM

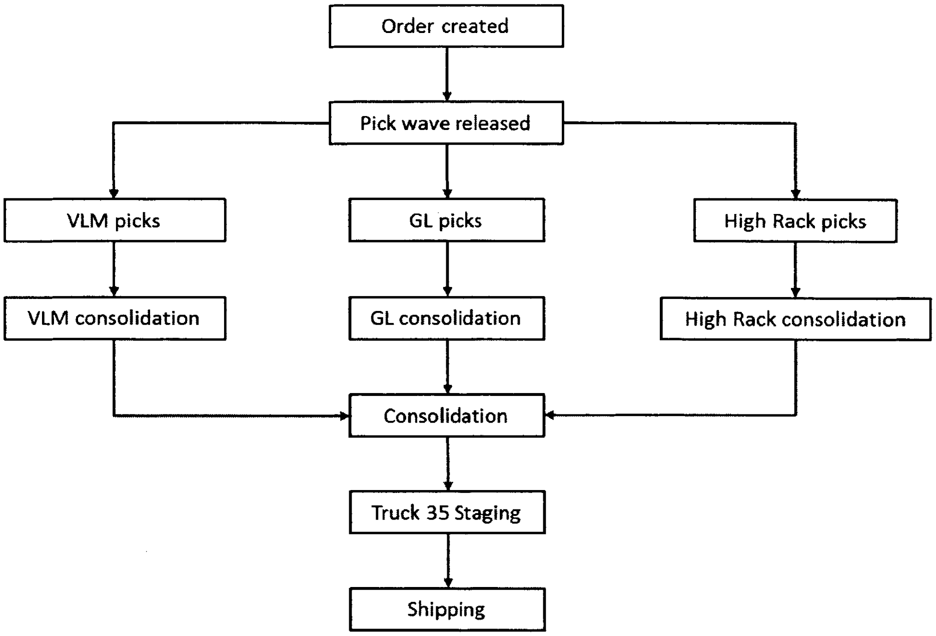


Figure 2-7: Pick-consolidate flow chart

consolidation is required for pick-and-consolidate.

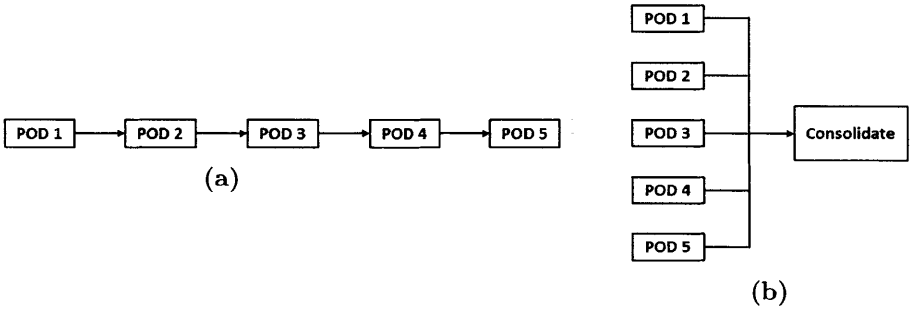


Figure 2-8: Picking strategies (a) Pick and pass; (b) Pick and consolidate

At VSEA, an employee is in charge of releasing the orders to be picked in waves and she determines whether to pick-and-pass or to pick-and-consolidate from the VLMs. Pick-and-consolidate is carried out by first dividing the list of parts of an order into five different waves defined by pods, followed by releasing one wave at a time. In contrast, pick-and-pass does not involve splitting the list of parts of an order into the five different pods, but simply releasing the list at one go.

Chapter 3

VLM Picking Strategies

3.1 Introduction

VLMs are gaining traction among warehouses because they improve the picking efficiency. Picking efficiency can be defined as pick rate (average number of picks per labor hour) or the average makespan to complete picking all the items of an order. The increase in picking efficiency is desirable because it is expected to improve the service level of a warehouse as well as reduce labor cost, which is a major component of the total warehouse operating cost. According to Tompkins et al. [12], order picking accounts for up to 55% of warehouse total operating costs. The service level of a warehouse is typically measured by whether orders are delivered on time and how long it takes to deliver an order. However, the extent of improvement in picking efficiency is partly dependent on the picking strategy.

At present, the main problem is that VSEA, and by extension many companies, do not know which picking strategy would give them the greatest improvement in picking efficiency. When the managers at VSEA first bought the VLMs, a pick-and-pass strategy was recommended by the VLM manufacturer. However, after implementing the pick-and-pass strategy for about a year, the warehouse noticed that bottlenecks were developed using the pick-and-pass method and totes were accumulated at the start of each VLM station. In addition, the warehouse claimed that the picking efficiency had only improved marginally. This motivated the warehouse to use an alternative picking strategy, pick-and-consolidate. Since they started to pick-and-consolidate in 2014, the warehouse observed that there were near zero accumulation of totes formed at the start of each VLM station. They further asserted that the average time taken to complete picking an order via pick-and-consolidate was shorter than pick-and-pass. However, there is no time study performed to compare the pick timings between pick-and-pass and pick-and-consolidate. There is thus a need to validate the claim that pick-and-consolidate is faster than pick-and-pass as well as examine the time savings to complete picking an order via pick-and-consolidate as compared to pick-and-pass.

In addition, it was postulated that the picking strategy might be one of the causes of the poor service levels of the warehouse during the busy periods i.e. parts are not delivered from the warehouse to the production floor within the 24 hours goal. Therefore, a key question to address in this section is – *which picking strategy should be employed during the busy period?*

Another area of concern that needs to be addressed is the effect of pick-and-consolidate on the additional time required to consolidate the totes arriving from up to five different VLMs. Unlike pick-and-consolidate, because the items of an order are consolidated in one tote for pick-and-pass, there is no VLM consolidation step. In contrast, for pick-and-consolidate, since items of an order are picked in parallel from up to five VLMs, a VLM consolidation step is required to put together the totes arriving from different VLMs.

Despite the additional VLM consolidation step required for pick-and-consolidation as compared to pick-and-pass, it is worth noting that the labor requirements for pick-and-pass and pick-and-consolidate are the same – five pickers at the VLMs and one person in charged of consolidation. At VSEA, the VLM consolidation step is performed by the employee who is in charge of consolidating the parts arriving from the three distinct picking areas (VLMs, GL and High Racks). Therefore, only one person works on consolidation (VLM consolidation and consolidating the parts arriving from different storage locations) and no additional manpower is needed for the VLM consolidation step for pick-and-consolidate.

The goal of this section is to first evaluate two distinct picking strategies, pick-and-pass and pick-and-consolidate in terms of i) the time taken to complete picking an order and ii) the time needed to consolidate the order after picking. The paper then examines under what circumstances, one picking strategy is more appropriate than the other. A recommendation to VSEA on which picking method should be employed during the busy period is provided at the end of this section.

3.2 Literature Review

There is an extensive research on order picking systems. Order picking systems focus on various aspects, including storage assignment (where parts are stored), picking layout (how many picking zones there should be) and picking policy (picking method and how orders are released i.e. in batches or individually). A comprehensive review on order picking systems is provided by Koster et al. [13]. While the paper presents picking policies for "pickers-to-parts" systems (a system where pickers walk along aisles to pick items), no picking policies were discussed for automated storage and retrieval systems (AS/RS) systems. Yet, picking policies for "pickers-to-parts" systems such as progressive zoning and synchronized zoning are employed in AS/RS systems. For example, progressive zoning is similar to the pick-and-pass method employed by VSEA as orders picked in one picking zone (one VLM station) is passed on to the other zones for completion; and synchronized zoning is similar to pick-and-consolidate as orders are picked from different zones in parallel.

Furthermore, in the research of AS/RS systems, there is a vast number of papers on

AS/RS systems related to crane-in-aisle types of AS/RS. However, only two papers [11, 14] to date deal with VLMs and their focus is on modeling the throughput of different VLM designs and configurations.

In summary, although there are many studies on improving picking productivity, most work focus on non-AS/RS systems i.e. picking from shelves or racks. Moreover, the primary research focus of AS/RS systems is on the operating modes of the systems and there is no work on storage assignment (as it relates to in-aisle picking), picking layout and picking policy of these systems. To our knowledge, there is no study that examines the impact of picking strategy on picking efficiency for AS/RS systems, in particular VLMs. The main contribution of this section is an analysis on the effect of picking methods, specifically pick-and-pass and pick-and-consolidate, on the picking efficiency of VLMs.

3.3 Hypothesis

It is proposed that the pick-and-consolidate strategy is a faster picking method than pick-and-pass i.e. the average time to complete picking an order from the VLMs is shorter. This is because pick-and-pass has much waiting time and delays because of starvation and blocking i.e. pods downstream have to wait for picks to be completed upstream the picking line. Unlike pick-and-pass, for pick-and-consolidate, the pickers work independently of each other and thus the waiting time due to starvation and blocking is eliminated.

A simplified setup that compares the mean time to complete picking an order between both picking strategies is presented here (refer to table 3.1). We first assume that an order size of 100 parts is to be picked equally from all five pods i.e. 20 parts are picked from each pod. We also assume an average pick rate of 30 parts per hour, which is the estimated average VLM pick rate at VSEA's warehouse. The time taken to pick 20 parts at one pod is thus 40 minutes. For pick-and-pass, picking starts from pod 1 and is sequentially passed to pod 2, pod 3, pod 4 and pod 5. Assuming that only one order is picked and there is no queue in the picking line, the time taken to complete picking the order is 200 minutes. In contrast, for pick-and-consolidate, picking from each pod occurs in parallel. The time taken to complete picking the order is 40 minutes, which is shorter than that of pick-and-pass. In summary, the time taken to complete picking an order for pick-and-pass is longer because it is a summation of the picking time at each pod, whereas that of pick-and-consolidate is determined by the longest picking time at the slowest pod.

Table 3.1: Simplified setup comparing pick-and-pass and pick-and-consolidate

Picking method	Pick-and-pass	Pick-and-consolidate
Number of parts to be picked from each pod	20	20
Picking time at each pod (minutes)	40	40
Total picking time (minutes)	200	40

The VLM consolidation step for pick-and-consolidate primarily involves scanning bar-codes of different handling units that belong to the same order, so as to integrate the different handling units into a single handling unit. For example, if an order of 100 parts is picked equally from five VLMs, there will be five handling units (or five totes) and each handling unit contains 20 parts. The completed handling units will be placed on the VLM consolidation rack to prepare for VLM consolidation. When all the five handling units are present on the VLM consolidation rack, the employee, who is in charged of consolidation, consolidates the order by scanning the bar-codes of the five handling units and integrating them into a single handling unit. In this example, a total of six scans are required i.e. five scans for the five handling units and one scan to integrate them into a single handling unit. A point to note is that the number of handling units corresponds exactly to the number of pods the order is picked from; and the number of scans required scales with the number of handling units as shown in table 3.2. The number of parts in each handling unit is dependent on where the parts are stored in the VLMs. On the other hand, for pick-and-pass, there is only one handling unit and the order is consolidated as the handling unit is passed from one pod to another. Thus no bar-code scan is required. If the order also requires parts from GL and High Racks, the employee then re-consolidates the parts arriving from VLMs, GL and High Racks. This second consolidation step is performed for both pick-and-pass and pick-and-consolidate.

Table 3.2: VLM consolidation for pick-and-consolidate

Number of pods an order is picked from	Number of handling units	Number of bar-code scans required	VLM consolidation time (s)
1	1	0	0
2	2	3	45
3	3	4	60
4	4	5	75
5	5	6	90

It is postulated that the additional time required to consolidate the totes arriving from up to five pods is minimal as compared to the longer picking time for pick-and-pass. This is because the additional VLM consolidation time merely involves up to six scans of bar codes. The scan rate by the employee in charged of consolidating orders is approximately 15 seconds per handling unit (three bar-code scans are required per handling unit). Therefore, the maximum additional VLM consolidation time for pick-and-consolidate is estimated to be 90 seconds.

3.4 Experiment

A time study comparing pick-and-pass and pick-and-consolidate is performed using historical data of VSEA's warehouse. This section first explains the metric used for comparing the picking performances between pick-and-pass and pick-and-consolidate. The method used to retrieve and analyze the historical data, including any assumptions made, is then detailed.

3.4.1 Metric for comparing picking performances

There are a couple of possible metrics to compare the picking performances between pick-and-pass and pick-and-consolidate. It is worth noting that these metrics measure the makespan to pick an order, rather than the labor time or requirements to pick an order.

Average pick time: Average time taken to complete picking an order. However, it does not account for the number of parts picked and the number of pods that the order is picked from. The average pick time is expected to scale with larger order sizes and higher number of pods that the order is picked from.

Average pick time per part: Average time taken to pick an item of an order. It is calculated by taking the total pick time of an order divided by the total number of picks of the order. However, it does not account for the number of pods that the order is picked from.

Average pick time per part per pod: Average time taken to pick an item of an order from a pod. It is calculated by taking the total pick time of an order divided by the total number of picks of the order and the number of pods that the order is picked from. The metric accounts for both the number of picks and number of pods that the order is picked from.

Because of the variations in order sizes and number of pods that the order is picked from, it is important to normalize the average pick time. Therefore, the metric used to analyze the data and compare the picking performances between both picking methods is average pick time per part per pod.

3.4.2 Data processing method

We first retrieved log files of the VLM picking tasks performed between January and July 2015 from the Extended Warehouse Management (EWM) system. As mentioned in chapter 2, a list of parts that are packaged together forms a consolidation group. The data is then categorized into unique and date-specific consolidation groups: each group contains a list of parts that belong to the same consolidation group and are released for picking on the same date. It is worth noting that picks of a unique consolidation group may be released over a few days because of parts shortages. Any pick is released only when there is inventory for the part in the warehouse. In the remaining sections of this chapter, the term 'order' is used to refer to a unique and date-specific consolidation group, so as to be coherent with the academic publications in this field.

In this experiment, we only examined orders which have more than one pick and the parts are picked from more than one pod. Orders, which only have one pick or comprise of parts which are picked from a single pod, are eliminated from the analysis because the average pick time per pod is estimated to be similar regardless of the picking method.

For each order, the following data are retrieved from the log files:

- Product ID of the part to be picked
- Time when the pick was released in terms of date and time of the day
- Time when the part was picked in terms of date and time of the day
- Storage location of the part in terms of pod 1, pod 2, pod 3, pod 4 and pod 5

First, we define the picking method for every order and divide the list of orders into the two categories. We consider an order to be picked by the pick-and-pass method if the times when the picks were released are exactly the same for all the picks in the order. On the other hand, if the times when the picks were released are different, we consider the order to be picked by the pick-and-consolidate method. This method of categorizing corresponds with the warehouse operations at VSEA: for pick-and-pass, an order is released in a single wave, thus the picks are released at a single timing.

In contrast, for pick-and-consolidate, the picks are grouped by pod storage locations and unique waves are released for each pod storage location group. The waves are typically released sequentially and thus the database records different timings when the waves are released.

3.4.3 Pick time of an order

The pick time of an order picked by pick-and-pass and pick-and-consolidate are calculated using two different approaches.

For pick-and-pass, the pick time of an order, t (minutes) is the time between when the first pick was released and when the last item was picked, as shown in equation 3.1.

$$t = t_2 - t_1 \tag{3.1}$$

Where t_1 is the time when the first pick is released for picking and t_2 is the time when the last item is picked.

Since the picking process for pick-and-consolidate is parallel, the pick time for this picking method is the maximum pick time from a pod. The pick time for every part of an order is first determined. The pick time is the difference between when the pick was released and when the item was picked. The parts of the order is then categorized based on where they were picked from i.e. pod location. For each pod, the maximum pick time of the list of picks that were picked from the pod is then determined. Finally, the pick time of the order t is determined by the bottleneck pod. Table 3.3 illustrates the method to determine the pick time of four orders picked by pick-and-consolidate.

The total number of parts picked, a and total number of unique pods, b are also obtained for each order. The pick time per part per pod of an order, P (minutes)

Table 3.3: Method to calculate pick time of an order, t for pick-and-consolidate in minutes

Order	Max pick time (min)					Pick time of order (min)
	Pod 1	Pod 2	Pod 3	Pod 4	Pod 5	
1	97	0	159	245	264	264
2	109	157	181	255	290	290
3	105	158	182	266	367	367
4	0	241	0	0	305	305

is

$$P = \frac{t}{ab} \tag{3.2}$$

The average pick time per part per pod over n orders, $\bar{P}$ in minutes is

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{3.3}$$

Where i represents the ith order.

3.4.4 Method for calculating the concurrent workload per order

When an order is being picked, the amount of workload that occurs simultaneously will affect the time taken to complete picking the order. If the workload is heavy, the pick time is raised; if the workload is light, the pick time is lowered. The workload when an order is being picked is evaluated based on the number of concurrent picks and concurrent orders. The higher the number of concurrent picks and orders, the heavier is the workload, vice versa.

The method of calculating the average amount of concurrent workload per order is explained below. The waves corresponding to an order is first identified. A wave is a group of picks that are released together all at once and typically, a batch of orders is released in one wave. For pick-and-pass, an order is picked from only a single wave. A pick-and-pass wave is made of picks of one or more orders. In contrast, for pick-and-consolidate, an order is picked from multiple waves, depending on the number of pods the order is picked from. A pick-and-consolidate wave is made of picks of one or more orders that originate from a particular pod (see table 3.4).

The number of concurrent picks for an order is calculated by summing up all the picks of the waves that are picked concurrently with the order. The number of concurrent orders for an order is calculated by counting the number of orders in the waves that are picked concurrently with the order. The methods for calculating the concurrent workload for both pick-and-pass and pick-and-consolidate are presented below:

Pick-and-consolidate: Referring to table 3.4, an order that is picked from five pods will be picked from five separate waves – wave 1, wave 2, wave 3, wave 4 and wave 5. Each pick-and-consolidate wave corresponds to a particular pod storage location and contains a batch of orders. For instance, all the parts in wave 1 are picked from pod 1 and the wave has a batch of 10 orders. The total number of picks in wave 1 is 20. The total number of concurrent picks is the sum of all the picks in waves 1 – 5 i.e. $20 + 15 + 25 + 10 + 10 = 80$; and the total number of concurrent orders is the sum of all the orders in waves 1 – 5 i.e. $10 + 5 + 3 + 2 + 5 = 25$.

Table 3.4: Example for calculating concurrent picks and concurrent orders for an order that is picked from five pods by pick-and-consolidate.

Wave	Storage location	Number of picks in wave	Number of orders in wave
1	Pod 1	20	10
2	Pod 2	15	5
3	Pod 3	25	3
4	Pod 4	10	2
5	Pod 5	10	5

Pick-and-pass: For an order that is picked by pick-and-pass, the order is picked from only one wave. If the wave consists of 10 orders, the total number of concurrent orders is 10. If the total number of picks in the wave is 100 i.e. the total number of picks from these 10 orders is 100, the total number of concurrent picks is 100.

3.4.5 Method for grouping orders based on order size

The order size also affects the flow time of an order. In order to determine the effect of order size on the average pick time per part per pod, we categorize the orders into three categories: i) small-sized orders ≤ 5 picks, ii) $5 <$ medium order ≤ 10 picks and iii) large order > 10 picks. After splitting the orders into their picking methods as described in section 3.4.2, we further divide the pick-and-pass and pick-and-consolidate orders based on their order sizes. The order size for an order is determined by the total number of picks in the order.

3.5 Results & Discussion

The differences between pick-and-pass and pick-and-consolidate in terms of the picking process, method for releasing the wave and VLM consolidation process, are summarized in table 3.5.

Table 3.5: Summary of differences between pick-and-pass and pick-and-consolidate

Picking method	Pick-and-consolidate	Pick-and-pass
Picking process	Parallel	Sequential
Wave release grouping method	Grouped by pod storage location; can have multiple orders in a single wave	Grouped by order; can have multiple orders in a single wave
VLM consolidation	Yes	No
VLM consolidation time	Scanning bar-codes which take up to 90s for five handling units	–

A summary of the historical VLM picking tasks between January and July 2015 is shown in table 3.6. The total number of orders with multiple picks and from multiple pods is 10,145, out of which 3,628 orders were picked by pick-and-pass and 6,517 orders were picked by pick-and-consolidate. There were 22,584 picks in total for pick-and-pass and 49,835 picks in total for pick-and-consolidate.

Table 3.6: Summary of historical VLM picking tasks between January and July 2015

Total number of orders	21,083
Total number of orders with single picks	9,313
Total number of orders with multiple picks from single pod	1,625
Total number of orders with multiple picks and from multiple pods	10,145
Total number of picks	85,554
Total number of picks (minus single picks and multiple picks-single pod)	72,419

3.5.1 Average picking performance

Table 3.7 presents the average picking performances of both picking methods. On average, the time required to complete picking an order by pick-and-consolidate is shorter than pick-and-pass by 3%. The average pick times of pick-and-consolidate and pick-and-pass are 3 h 44 min and 3 h 51 min respectively. The average pick time of pick-and-consolidate for an average of 8 picks is unexpectedly high as we expect the pick time to be in the order of 1 – 2 hours. The percentage difference in average pick times between pick-and-pass and pick-and-consolidate is also surprisingly small as we expect the difference to be greater. There are a few reasons:

First, the time required to complete picking an order depends on when the parts for the order are released for picking and when the part is actually picked. With reference to figure 3-1, most of the picks that are released between 5 am and 7 am are fulfilled only after 3 hours. By extension, this means that the time to complete picking the order is about 3 hours. This makes sense since the first few picks in the early morning correspond to picks that were released in the afternoon the day before. It is worth noting that most of the VLM picks are carried out within the first shift of the warehouse i.e. between 6 am and 3 pm. Hence, orders that are released between 5 am and 7 am are only picked once the backlog has been cleared. In particular, since the pickers begin work at 6 am, orders that are released around 5 am are typically picked only after 6 am .

Second, the pickers have other responsibilities such as put away, cycle count[1] and

---

[1] Cycle count is the task of counting the inventory to ensure that the quantity of a part in a bin

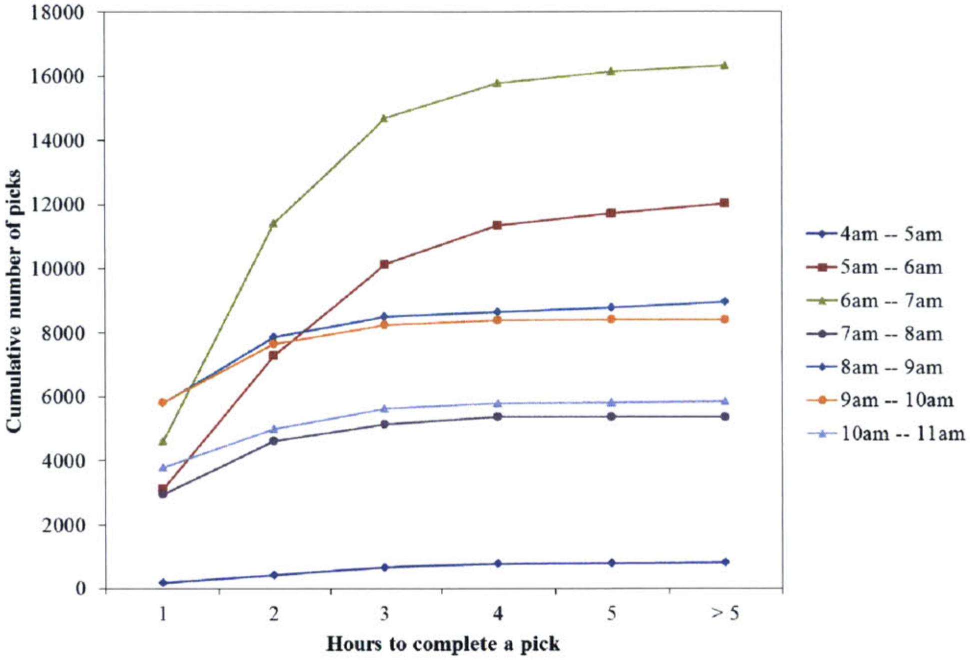


Figure 3-1: Cumulative number of picks vs. hours to complete a pick for picks that are released at different times of the day before 11am

packing sales kits. Moreover, the work schedule of pickers are relatively flexible as they can move between picking and performing other warehouse tasks. The pickers typically return to picking either when they have completed the other warehouse tasks or when they are informed of the release of additional picks[2]. This flexibility causes picks to be delayed, and is estimated to be delayed up to an hour. Despite the flexible work schedule, the pickers place picking as their top priority and they usually complete their picks before moving on to other warehouse tasks.

Moreover, we assume that there are always five pickers picking in parallel and at the same point in time. However, in reality, some pickers might move on to perform other warehouse tasks and then return to continue picking. For instance, although the pickers at pods 1 and 2 are picking the order, the pickers at pods 3 and 4 are involved with other warehouse tasks such as cycle count and put away instead of

---

tallies with that in the EWM system.

[2] It is worth noting that at the warehouse of VSEA, pickers and other warehouse staff often communicate among each other of what needs to be done.

picking. Even if the orders for pick-and-consolidate are released at the same point in time, the items for the orders may not necessarily be picked at the same point in time.

In addition, because the delivery time goal is set as 24 hours, the main motivation for picking is to complete the orders within 24 hours. As long as the warehouse is meeting the delivery time goal, there is currently no incentive to pick an order faster. In particular, for production orders, the orders should be completed by 3 pm on the same day as when the order is released for picking. The 3 pm timing is preferred so as to ensure that the parts are delivered from the warehouse to the production floor within the next day by 7 am. Typically, the production picks are released in the morning around 10 am (after sales picks have been completed). If the production orders have to be completely picked by 3 pm, the allowable pick time is about 4 – 5 hours. Because of the lack of incentive to pick faster than 4 – 5 hours, the average pick time for pick-and-consolidate is high.

Last but not least, part shortages and missing parts in the VLMs also contribute to the longer pick times. For instance, stock for replenishment that is received by the warehouse in the morning can only be picked from the VLMs once it has been put away into the VLMs. If the stock is only put away into the VLMs by the afternoon and a pick for the part is released in the morning (since the EWM system records stock availability once it has been received at the receiving area of the warehouse), the pick time for the part is also extended. Therefore, the pick times also depends on when the stock for replenishment is put away into the VLMs.

The combination of all the above-mentioned reasons may further amplify the time taken to complete picking an order. Therefore, the average pick time of 3 h 44 min calculated from historical data is higher than the expected 1 – 2 hours needed to complete picking an order. By extension, the percentage difference in average pick times between pick-and-pass and pick-and-consolidate becomes relatively low as well.

The average pick time per part per pod is 18 min 24 s for pick-and-consolidate and 22 min 56 s for pick-and-pass. For pick-and-consolidate, it is assumed that the time needed to consolidate the totes arriving from different pods is 15 s per pod (or per tote). The resultant average pick time per part per pod for pick-and-consolidate is 18 min 39 s. In contrast, for pick-and-pass, the order is consolidated as the totes are passed along the VLMs. As mentioned in the hypothesis, the key advantage of pick-and-consolidate is that orders are picked in parallel and can therefore be picked simultaneously from different VLMs. On the other hand, pick-and-pass is a sequential process and each VLM has to wait on the previous VLM. If the distribution of picks across the VLMs is not balanced, the time to complete picking an order is further delayed due to long waiting times between VLMs.

Table 3.7: Overall picking performances of pick-and-pass and pick-and-consolidate between January and July 2015. Time results are based on per order.

Picking method	Pick-and-consolidate	Pick-and-pass
Number of orders	6,517	3,628
Average pick time	3 h 44 min	3 h 51 min
Average pick time per part per pod	18 min 24 s	22 min 56 s
Average number of pods picked from	3	2
Average number of picks	8	6
Consolidation time per handling unit (s)	15	0
Average pick time per part per pod + consolidation time	18 min 39 s	22 min 56 s
Average number of concurrent picks	87	49
Average number of concurrent orders	38	10

Moreover, the average workload when an order is being picked by pick-and-consolidate is heavier than that by pick-and-pass. At the warehouse, the average daily workload fluctuates significantly and the workload also varies within a day. Therefore, in order to determine the average workload when an order is being picked, we measure the number of picks and orders that have to be fulfilled concurrently. The higher the number of concurrent picks and the higher the number of concurrent orders, the busier the workload is at the time of picking the order. From table 3.7, the orders that were picked by pick-and-consolidate had a larger number of concurrent picks and orders. One would thus expect that the average pick time per part per pod would be longer for a busier workload. However, the results reveal that despite the busier workload, pick-and-consolidate has a better average picking performance than pick-and-pass.

Comparing the average picking performances, we can conclude that our hypothesis – pick-and-consolidate is a faster picking method than pick-and-pass – is valid. More importantly, despite the heavy workload, the average time to complete an order by pick-and-consolidate remains low, and is shorter than pick-and-pass.

3.5.2 Service levels of picking methods

Figure 3-2 presents the service level of both picking methods. The service level is the proportion of orders that have been completely picked within a given time. The points of inflection of both curves occur at three given times namely, 8 hours, 15 hours and 20 hours. In general, pick-and-consolidate has a higher proportion of completed orders than pick-and-pass within the specified time, albeit the difference is marginal. In particular, a noticeable higher proportion of orders are completed between 8 and 15 hours for pick-and-consolidate as compared to pick-and-pass. A better picking service level at the VLM helps ensure that orders are delivered on time from the warehouse to the production floor. Therefore, comparing the service levels, pick-and-consolidate is the preferred picking method.

3.5.3 Effects of order size and number of pods

From preliminary analysis of the picking tasks in the warehouse, we found that there are variations in the order sizes and number of pods an order is picked from. It is thus also of interest to perform a robustness evaluation of the picking methods by examining the effect of order size and number of pods an order is picked from on the average picking performances of pick-and-pass and pick-and-consolidate. The results

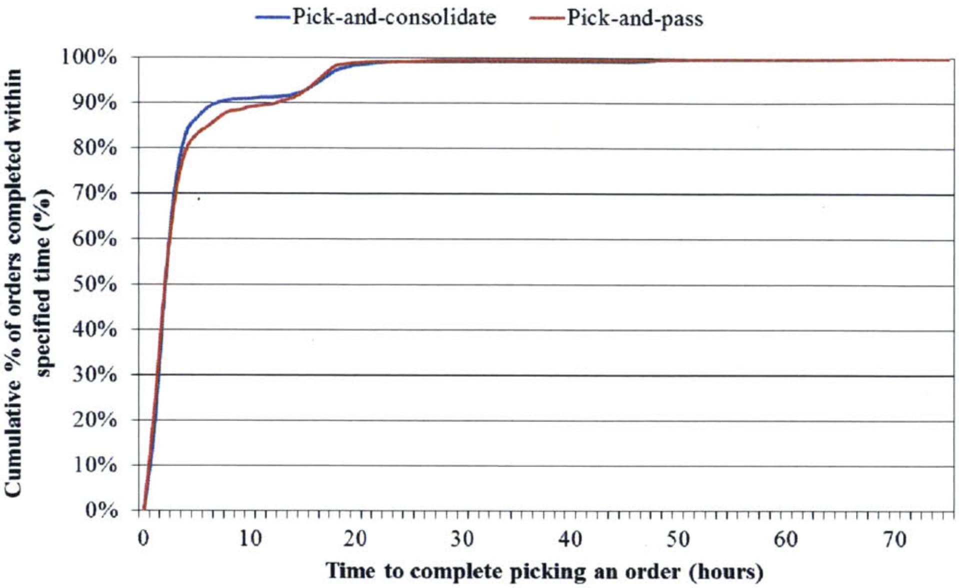


Figure 3-2: Effect of order size on the picking performances of pick-and-consolidate and pick-and-pass

show that pick-and-consolidate has better picking performances than pick-and-pass for most scenarios.

Effect of order size

With reference to figure 3-3, pick-and-consolidate has a lower average pick time per part than pick-and-pass for all order sizes. The average time savings gained by using pick-and-consolidate are 91s, 137s and 33s for small-, medium- and large-sized orders. In this case, we also assume that the time to scan an additional tote is 15s.

Effect of number of pods

To analyze the effect of number of pods on the picking performances of both picking methods, we use the average pick time per part metric instead of pick time per part per pod because the number of pods have been taken into account. We assume that the time to scan an additional tote is 15s. The additional times required for consolidating

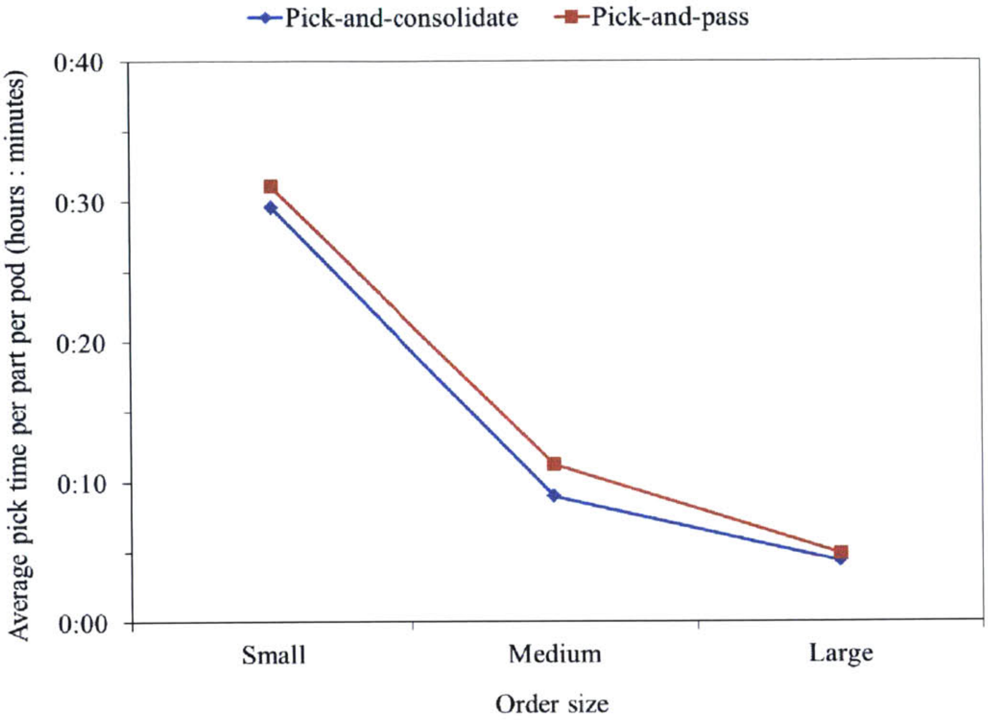


Figure 3-3: Effect of order size on the picking performances of pick-and-consolidate and pick-and-pass. The order sizes are defined as follows: small order ≤ 5; $5 <$ medium order ≤ 10; large order > 10.

the orders arriving from 2, 3,4 and 5 pods are 45, 60, 75 and 90s respectively. Figure 3-4 shows that the average pick time per part for pick-and-consolidate is lower than pick-and-pass if the order is picked from more than two pods. The time savings is greatest if the number of pods an order is picked from is 3. For orders that are picked from only two pods, the average pick time per part for pick-and-pass is 44s faster than pick-and-consolidate. In general, for pick-and-pass, the waiting time between pods increases with the number of pods an order is picked from. Thus, it is reasonable to expect that the picking performances of pick-and-pass to be comparable to pick-and-consolidate, if not marginally better.

While we expect that the time savings gained by pick-and-consolidate should increase with the number of pods, the results show otherwise. A possible explanation is that the average workload for orders that are picked by pick-and-consolidate is greater than that by pick-and-pass. Therefore, the average pick time per part by pick-and-consolidate is raised and the time savings are reduced.

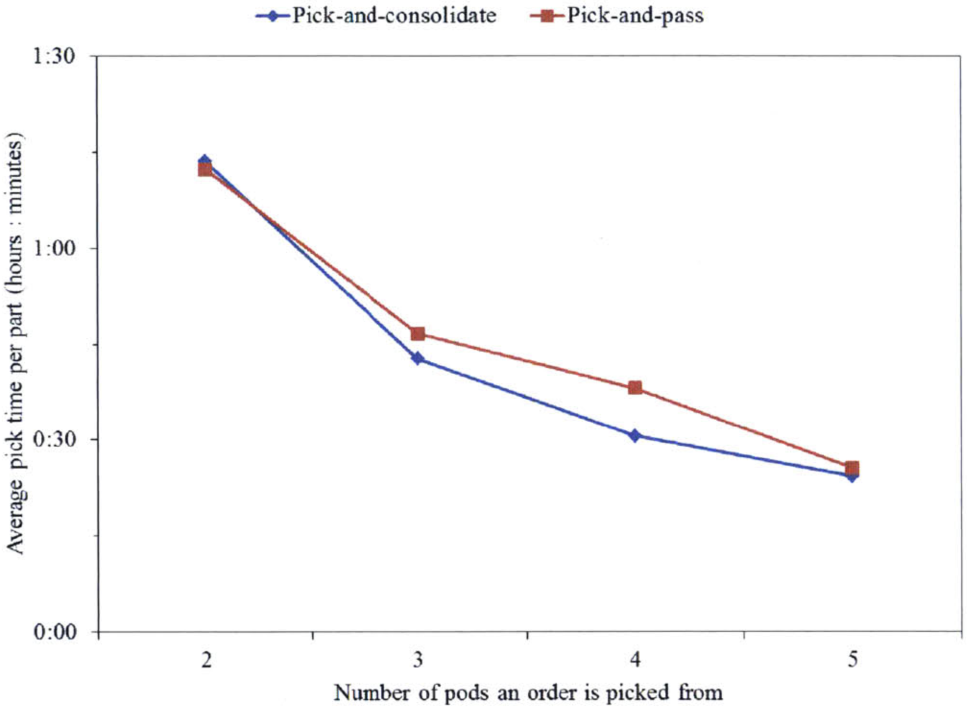


Figure 3-4: Effect of number of pods an order is picked from on the picking performances of pick-and-consolidate and pick-and-pass

3.5.4 Estimation of time savings by pick-and-consolidate

Comparing the average picking performances of both picking methods, it can be concluded that pick-and-consolidate is the better picking method in terms of providing a lower average pick time per part pod and a better picking service level. The next question of interest to warehouse managers is how much time savings can be gained by pick-and-consolidate? This section provides an estimated time savings of picking an order if pick-and-consolidate was employed instead of pick-and-pass.

With reference to figures 3-5 and 3-6, the distribution of pick times of both picking methods are skewed towards the left. Majority of the orders for both picking methods are completed within 8 hours, but there are few orders that have a very long pick time per part of up to 70 hours and 75 hours for pick-and-consolidate and pick-and-pass respectively.

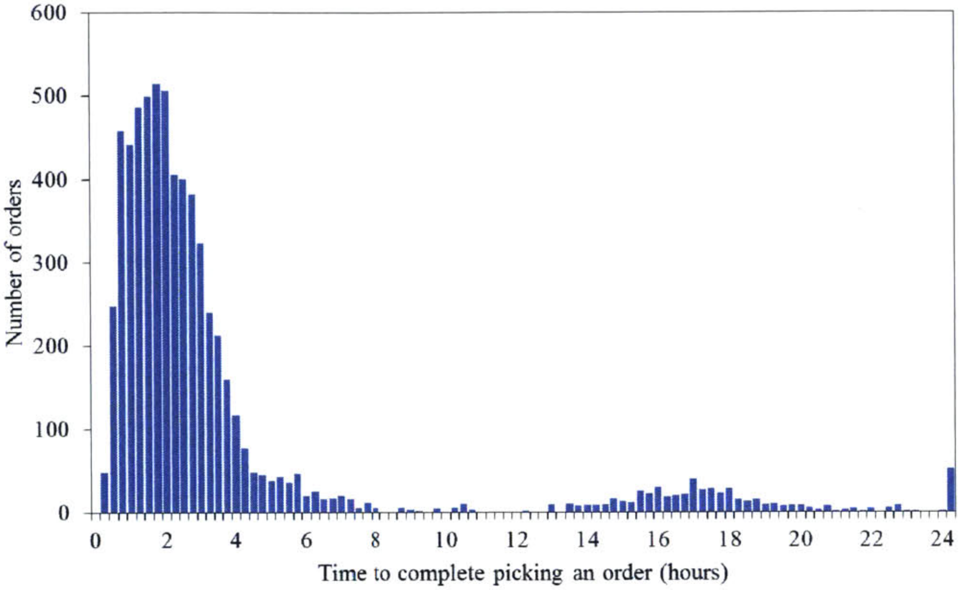


Figure 3-5: Distribution of pick times of orders that were picked by pick-and-consolidate

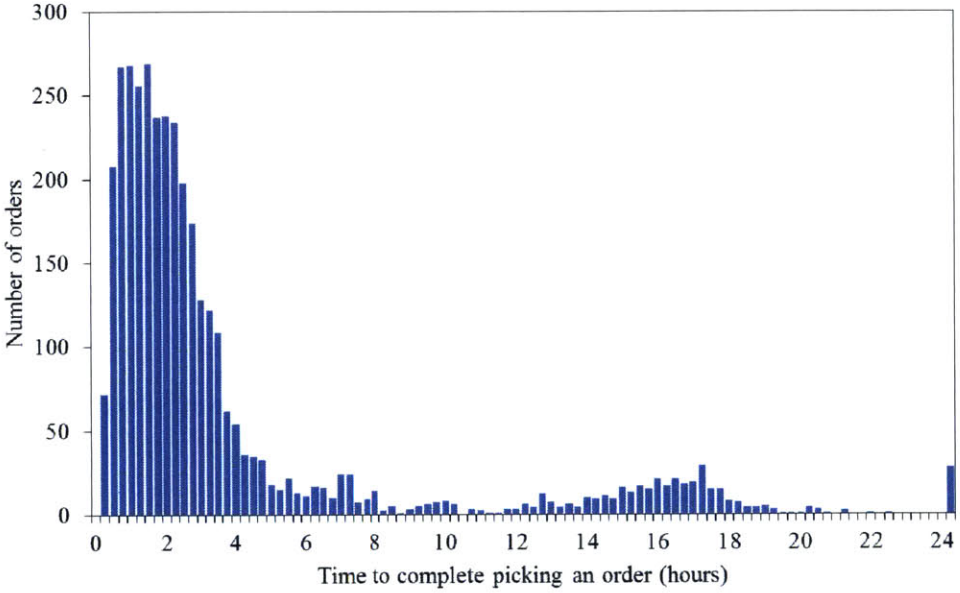


Figure 3-6: Distribution of pick times of orders that were picked by pick-and-pass

There are a couple of reasons for the anomaly long pick time per part per pod. First, the order was released for picking in the second shift when there are limited labor resources. Typically, there are no pickers specifically assigned to the VLMs. The few who are working on the second shift handle a variety of warehouse tasks such as put away, receiving and picking from all the different picking areas. Therefore, any picks that have been released at 3 pm and onwards may not get picked until the next morning during the first shift. Second, part shortages also cause delays in picking. Third, between end of December and early January, the warehouse faced manpower issues and they had insufficient manpower at the VLMs. It was claimed that only one or two pickers were working at the VLMs. Moreover, this time period was the busy period and the warehouse had a high number of VLM picks. There is thus a need to narrow down the data we are analyzing in order to minimize the influence of the few orders that have a high pick time per part per pod on the average result.

The entire set of data is then narrowed down to orders that are completed within 8 hours. The time of 8 hours is selected because i) the sample size is sufficiently large and ii) most of the VLM picks occur during the first shift, which is 8 hours long. First, with reference to figure 3-2, the proportion of orders fulfilled within 8 hours for pick-and-consolidate and pick-and-pass are relatively high at around 90%. This ensures that majority of the data is still captured for analysis. The larger the sample size, the more accurate is the representation of reality. Second, the 8-hours time to complete picking an order is also a reasonable value. Although the target part delivery time is 24 hours, the allowed time to complete picking an order should be defined by when the warehouse is active. The warehouse is most active during the first shift between 7 am and 3 pm, which is an 8-hour time frame. Moreover, the first shift has more labor resources and typically has five employees allocated to work at the VLMs.

The results for orders completed within 8 hours are presented in table 3.8. The average pick time per part per pod is 11 and 14 minutes for pick-and-consolidate and pick-and-pass respectively. The average pick times are 2 h 7 min and 2 h 4 min for pick-

Table 3.8: Picking performances of orders that were completed within 8 hours by both pick-and-pass and pick-and-consolidate methods. The orders were picked between January and July 2015. Time results are based on per order.

Picking method	Pick-and-consolidate	Pick-and-pass
Number of orders	5,914	3,200
Average pick time	2 h 7 min	2 h 4 min
Average pick time per part per pod	11 min	14 min
Average number of pods picked from	3	3
Average number of picks	7	6
Average number of concurrent picks	86	46
Average number of concurrent orders	38	10

and-consolidate and pick-and-pass respectively. The difference between the average pick times is small because pick-and-pass is carried out only during the slow periods, while pick-and-consolidate is carried out during the busy periods. Therefore, for pick-and-pass, the amount of waiting time due to starvation and blocking is reduced and the average pick time is not as high as we would expect.

Table 3.9: Estimation of makespan time savings per order. The estimated time savings per order by pick-and-consolidate is 252 − 238 = 20 min.

Picking method	Pick-and-consolidate	Pick-and-pass
Average pick time per part per pod (minutes)	11	14
Average number of pods picked from	3	3
Average number of picks	7	6
Calculated average pick time	231 min	252 min
Calculated average VLM consolidation time	60 s (3 pods)	–
Average pick-consolidate time	232 min	252 min

The calculations for estimating the makespan time savings per order for pick-and-consolidate are shown in table 3.9. Based on the average number of picks and average number of pods picked from, the average pick times are 231[3] and 252 minutes[4]

---

[3] Average pick time for pick-and-consolidate = Average pick time per part per pod × Average number of pods × Average number of picks =11 × 3 × 7 = 231 min

[4] Average pick time for pick-and-pass = Average pick time per part per pod × Average number of pods × Average number of picks =14 × 3 × 6 = 252 min

for pick-and-consolidate and pick-and-pass respectively. For pick-and-consolidate, an additional minute is required for VLM consolidation. Comparing the average pick-consolidate times between both picking methods, the average makespan time savings per order is 20 minutes. In addition, in the future when the workload across the VLMs is balanced [15, 16], the number of pods an order is picked from will increase from 3 to 5. The estimated future makespan time savings per order by employing pick-and-consolidate over pick-and-pass is 27 and 34 minutes if the order has to be picked across 4 and 5 pods respectively.

3.6 Summary

A time study is performed to compare the picking efficiencies of pick-and-consolidate and pick-and-pass. The picking efficiency is defined by the average makespan to complete picking an order. The picking performances of both picking strategies are compared using the average pick time per part per pod metric. This is because the metric takes into account the number of picks and number of pods an order is picked from. Both factors are positively correlated to the time to complete picking an order and thus the pick time has to be normalized. Since the average pick time per part per pod by pick-and-consolidate is shorter than pick-and-pass, by extension, the average flow time to complete picking an order is lower for pick-and-consolidate. Pick-and-pass has a longer flow time because of starvation and blocking i.e. the pickers downstream have to wait for the picks to be completed upstream the line. Moreover, the warehouse management's concern that VLM consolidation is a time consuming process is invalid. It is estimated that the time to scan a tote is 15s and the maximum additional time required to consolidate an order arriving from up to five VLMs is only 90s. In summary, our results support our hypothesis that pick-and-consolidate is a more efficient picking method than pick-and-pass.

The recommended picking method to VSEA is thus pick-and-consolidate. If an order is picked by pick-and-consolidate instead of pick-and-pass, the expected makespan time savings per order is 20 minutes (or 8%). Furthermore, once the workload across the VLMs is balanced in the near future, an order is expected to be picked from 4 to 5 pods, rather than from 1 to 3 pods. The time savings gained from using pick-and-consolidate is thus expected to increase further.

This page intentionally left blank

Chapter 4

Maintaining the Operational Efficiency of the Warehouse

In the other two theses [15, 16], we identified the problem and developed solutions to address the poor on time delivery performance of the warehouse during busy periods. This chapter reviews our proposed and implemented solutions and evaluates what is required to maintain a high service level of the warehouse (i.e. consistently meeting the target part delivery time of 24 hours) and high picking efficiency of the VLMs. We first define the key metrics to measure the performances of the warehouse and the VLMs. Then, we discuss how to maintain a high number of VLM picks and a high VLM picking efficiency. Finally, we apply our framework for maintaining a highly efficient VLM system to future new VLMs installed.

4.1 Key Performance Metrics

The current metrics used by the company to evaluate the performance of the warehouse are inadequate. In this section, we analyze the effectiveness and limitations of existing performance metrics and propose new additional performance measurements that should be adopted by the company.

4.1.1 Warehouse goal

The goal of the warehouse is to consistently deliver completed orders on time to the production floor and their customers i.e. parts must be delivered within the target delivery time of 24 hours.

4.1.2 Warehouse service level

The service level of the warehouse is currently measured by the average delivery time of orders of the same category. The production floor has three types of warehouse orders namely, MOD, gas box and facilities, and universal end station. The delivery time for a type of order is typically determined by the average of the orders of that type. However, this method of measurement does not reveal the number of orders that meet the delivery time goal. For instance, if one order took 8 hours to be completed and another order took 26 hours to be completed, the average delivery time of these two orders would be 17 hours. The 17 hours delivery time reported on their performance sheet would not reveal the fact that one order did not meet the delivery time goal. Therefore, in accordance to the warehouse goal, it is proposed that the service level of the warehouse should be measured in terms of either the number (or percentage) of completed orders that have been delivered within the target delivery time of 24 hours, or the maximum delivery time of the set of orders of a particular type.

4.1.3 Picking efficiency

The picking efficiency of a picking zone is currently measured by the pick rate. The pick rate is defined as the number of picks per labor hour, which is calculated by equation 4.1.

$$\text{Pick rate} = \frac{\text{Total number of picks from all the VLMs}}{\text{Total picking labor hours}} \tag{4.1}$$

A limitation of this metric is that the picking labor hours are manually reported by the operators, rather than tracked using a software. As a result, the picking labor hours reported might include put-away time and idle time between orders. Therefore, the VLM pick rate tends to appear lower than it is in reality – the target pick rate is 60 picks per hour, but the current average pick rate is only at 35 picks per hour. While a software application is a preferred method of measuring the pick rate since it is more accurate, setting up a new software can be costly.

The main problem of measuring picking efficiency of the VLMs by pick rate is that the metric does not relate to the time required to complete picking an order. For example, an order of 100 parts have to be picked and there are two possible scenarios: i) the order is picked only from one pod and ii) the order is picked equally from five pods. Assuming that the time to pick a part is one minute, then the time to complete picking the order of 100 parts in the first scenario is 100 minutes. On the other hand, the time to complete picking the order in the latter scenario is 20 minutes. However, the pick rates of both scenarios are the same (pick rate equals to 60 picks per hour). The pick rate indicates whether the flow of picking from the VLMs is lean through accounting for the time taken to pick the items, bag and label, but does not provide information on how fast an order is completely picked.

In order to relate the picking efficiency of the VLMs to the speed at which orders are completed, it is recommended to measure the average make span for each order. In addition, because the flow time of an order is strongly correlated to the distribution of workload across the VLMs [15, 16], it is recommended to measure the workload distribution across the five pods. These two metrics are further explained below.

Average makespan per order

The more efficient the VLMs are, the shorter the average flow time of each order, and the shorter the total delivery time[1]. While the picking efficiency of the VLMs is critical, because the primary goal of the warehouse is to deliver orders within 24 hours, the picking efficiency of the entire warehouse is the most important. Hence, the average makespan per order should not only be measured at the VLMs, but also at the GL and High Racks. By comparing the average flow times per order at the VLMs, GL and High Racks, the bottleneck of the warehouse picking process can be identified. If the average flow time per order at the VLMs is the highest, picking from the VLMs is the bottleneck. This implies that there is a need to further improve the picking efficiency of the VLMs. On the other hand, if the average flow time per order at the GL is the highest, picking from GL is the bottleneck. In order to alleviate the bottleneck at GL, the workload on the GL has to be reduced and this can be achieved by transferring fast moving parts from the GL to the VLMs. This method of resolving the bottleneck at GL is briefly discussed in chapter 1 and further detailed in [16].

Workload distribution of VLMs

The workload distribution across the five VLMs provides insights on how balanced the workload is across the five pods. The metric is important because it closely relates to the time required to complete picking an order. The picking time of an order is higher for an unbalanced scenario as compared to a balanced scenario. If the current workload across the five pods is skewed, picking in parallel from different pods is limited. The value of maintaining a balanced workload across the five pods is elaborated in [16].

The workload of a pod can be determined by the cumulative expected number of picks

---

[1] The delivery time is the duration between when the warehouse receives an order and when the completed order is delivered to the production floor or customer.

of all the parts stored in the pod over a certain time period. In our work, we define the expected pick frequency of a part as the total number of picks over the last one year plus projected number of picks in the next six months. Because VSEA produces at low volumes, one year of historical picks is a reasonable time period for the number picks for slower moving parts to be accumulated over a year. In addition, from our preliminary analysis, we found that the number of picks of most parts only change marginally between one year and 1.5 years measurement time periods. Next, the projected number of picks in the next six months accounts for the future demand of the part. If the future demand of the part declines, the contribution of the part to the cumulative pick frequency of the pod becomes smaller as well. However, the accuracy of the projected number of picks in the next six months depends on the forecast accuracy. Therefore, if a more accurate projected number of picks is required, the future picks can be projected over a shorter time period, say within the next two or three months instead of six months.

4.1.4 Workload of picking locations

The workload of each picking location is measured by the number of picks at the picking location over some fixed period of time. This is an effective metric to compare the workloads between picking locations.

4.2 Methodology to Sustain a High Number of VLM Picks

As mentioned in chapter 1, the bottleneck of the pick-consolidate process is picking from GL, which ultimately results in the late deliveries to the production floor. One aspect of our work is focused on reducing the workload on GL and increasing the number of VLM picks in order to alleviate the bottleneck. This is achieved by moving fast moving parts from the GL to the VLMs. However, because there is limited space in the VLMs, it is necessary to remove FISH[2] parts in order to ensure that the VLMs have sufficient storage capacity for fast moving parts. The methodologies used to determine which fast moving parts are to be moved from the GL to the VLM and which slow moving parts are to be moved from the VLM to the GL are detailed in this paper [16]. The paper also presents the results of the implementation of our recommendations on the warehouse floor.

After implementation of our proposed idea, the follow up question that needs to be addressed is: *how often does the warehouse need to review the fast moving parts in the GL and FISH parts in the VLM in order to sustain the high number of VLM picks?*

A review period of six months is recommended. This is a reasonable estimate because as mentioned above, the pick frequencies of parts require time to be accumulated. In particular, for the FISH parts, this is also a reasonable estimate since an ion implantation equipment model is typically sold in the market for two to three years. Once the initial list of 2,052 FISH parts that have been identified in [16] have been removed from the VLMs, majority of the remaining parts in the VLMs should correspond to those required to manufacture the equipment that are currently sold in the market. Minority of the parts are spare parts for older equipment models. We expect that the subsequent lists of FISH parts will be significantly smaller than the current list because FISH parts take time to accumulate.

4.3 Methodology to Sustain an Efficient VLM system

The two important elements to maintain an efficient VLM system are:

---

[2] First In Still Here (FISH) parts are parts that have not been picked over the last one year and have no projected picks in the next six months.

1. **Dynamic slotting**

   The key to operating an efficient VLM system is to have a balance workload distribution across the pods and this is primarily attained by utilizing an effective slotting technique. From our preliminary analysis [15], the existing slotting method used at VSEA is ineffective. At present, parts are put away based on the Put Away Control Indicator (PACI)[3] assigned to each part. The problem is that there is no proper framework on the PACI assignments for the parts. As a result, the distribution of picks are skewed towards the first pod and the distribution of workload across the five pods is unbalanced.

   In this paper [15], static and dynamic slotting methods to balance the workload across the pods are examined. Three static PACI assignment methods that require periodic reviews of the PACI assignments for all parts in the VLMs are first presented. The static slotting methods are then extended to dynamic slotting systems that balance the workload as each incoming part is put away into the VLMs. The main benefit of a dynamic slotting system is that unlike a static slotting method, the dynamic setup does not require constant reviewing and updating, which is time consuming and labor intensive. It is thus of interest to further examine dynamic slotting methods for maintaining a balanced workload distribution across the five pods so as to ensure that the VLMs are running efficiently. This is detailed in section 4.3.1.

2. **Employing an efficient picking method**

   From chapter 3, it was concluded that the more efficient picking method is pick-and-consolidate, a parallel picking process. Moreover, for this picking method, the time required to consolidate totes coming from up to five pods is minimal, with each tote requiring an additional 15s scan time. Therefore, pick-and-consolidate should be employed as much as possible.

---

[3] A PACI tells EWM where to start looking for a open bin during the part put-away search, in the VLMs a sequential numerical order is followed (e.g. PACI of VL01 results in a put-away search sequence of VL01-01→ VL01-02 → VL01-03 → VL02-01 → ... etc.)

4.3.1 Dynamic Slotting Methods

Slotting is the activity of determining the most appropriate storage location for each item in the VLMs. In our work, the objective of slotting is to maximize the VLMs' picking efficiency by balancing the workload across the five pods. Re-slotting is performed when the picking efficiency of the VLMs is declining or when the workloads across the five pods become skewed. In the industry, the common practice for re-slotting parts involves periodically reviewing the workload distribution across the VLMs – in our work, we call this static slotting. If the distribution progressively becomes skewed, parts have to be manually transfered from one VLM to another in order to balance the workload. Static slotting is not only labor intensive, but the approach is more of a reactive one rather than a preventive one i.e. the solution is implemented only after the problem has occurred. In this section, we propose to implement a preventive solution – dynamic slotting. Unlike static slotting, dynamic slotting focuses on determining the most appropriate storage location for each item so as to minimize both the imbalance in workload across the five pods and the imbalance in workload across the bays within a pod. By employing a dynamic slotting solution, we postulate that re-slotting the VLMs (manually moving parts to re-balance the workload) would be kept to the minimum.

At VSEA, the warehouse can receive hundreds of parts on any given day, most of which need a PACI assignment. The first part of this section looks at dynamic slotting using the EWM software at VSEA. The implementation strategy involves changing the software logic in the mainframe. To determine which strategy works best in this situation, logical flow charts were created to determine the programing requirements and cost estimates. At present, the PACI of a part is based purely on pod assignment. However, having only a pod assignment is inadequate; a bay assignment is also important for ensuring that the workload across the bays within a pod is balanced since this relates to the pick rate of the VLMs. In this section, referring to the flow diagrams in figures 4-1, 4-3 and 4-2, a method is considered to be fully implemented if the PACI assignment includes both pod and bay assignments. The method is considered partially implemented if only pod assignments are allowed. We also propose an alternative dynamic slotting method that is not based on PACI assignment, but rather the cumulative expected pick frequencies of the pods.

Randomization

This method uses a random number generating function which can generate a number for the pod assignment between 1 and 5. If bay assignment is feasible, a random number for the bay assignment is subsequently generated; otherwise, only the pod is assigned to the part and an empty bin for the part is found based on the sequential search sequence that is currently in use. For instance, if the PACI assignment is pod 2, an empty bin is found by beginning the search from VL02-01, VL02-02, VL02-03, VL03-01, VL03-02 etc.

Order Grouping

The order grouping method is by far the most computationally intensive strategy to implement. This would require knowledge of the most frequently ordered groups of parts and the current location of every part in the group.

Snake Frequency Balancing

In [15], the snake frequency balancing method, which uses pick frequency to assign parts to pods in a sequential order, is discussed. In the static case, the snake frequency balancing method first sorts the parts in descending order of expected pick frequency, followed by assigning parts to pods and bays in an alternating balancing fashion i.e. 1, 2, 3, 4, 5, 5, 4, 3, 2, 1, 1, 2 etc. However, the static slotting approach would be excessively complicated to implement into a dynamic solution. In order to capture the essence of the strategy in an implementable manner, the logic diagram shown in

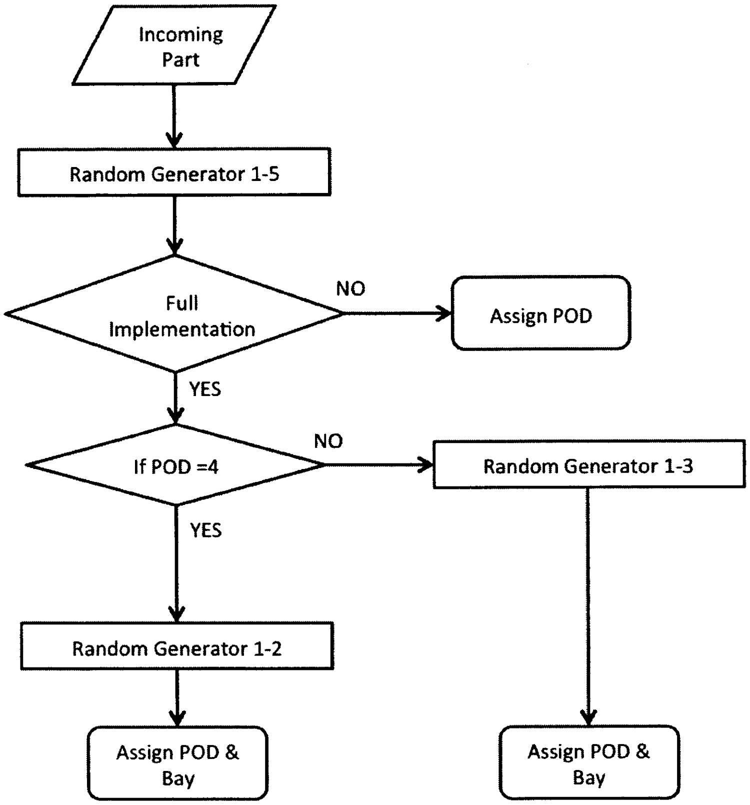


Figure 4-1: The logic required to implement the Randomization strategy into the software

Figure 4-3 is created. This method is in the middle ground in terms of complexity and implementability. The focus of this is to put away incoming parts into the pod and bay that have the lowest utilization i.e. lowest cumulative expected pick frequency. This is a workload balancing dynamic slotting method that involves calculating the current workload distribution of the pods. In figure 4-3, the workload distribution is measured based on the total number of picks in each pod for the last six months. The sequence of pods will then be arranged in ascending order from most to least utilized and this order will be used as the put-away search sequence starting from the lowest utilized pod as the first pod assignment. If the bay assignment is possible (a full implementation is possible), then each bay within the chosen pod will be ranked in terms of total number of picks in the last six months and the bay. Similarly, the bay search sequence starts from the lowest utilized bay as the first bay assignment, followed by the second lowest utilized bay.

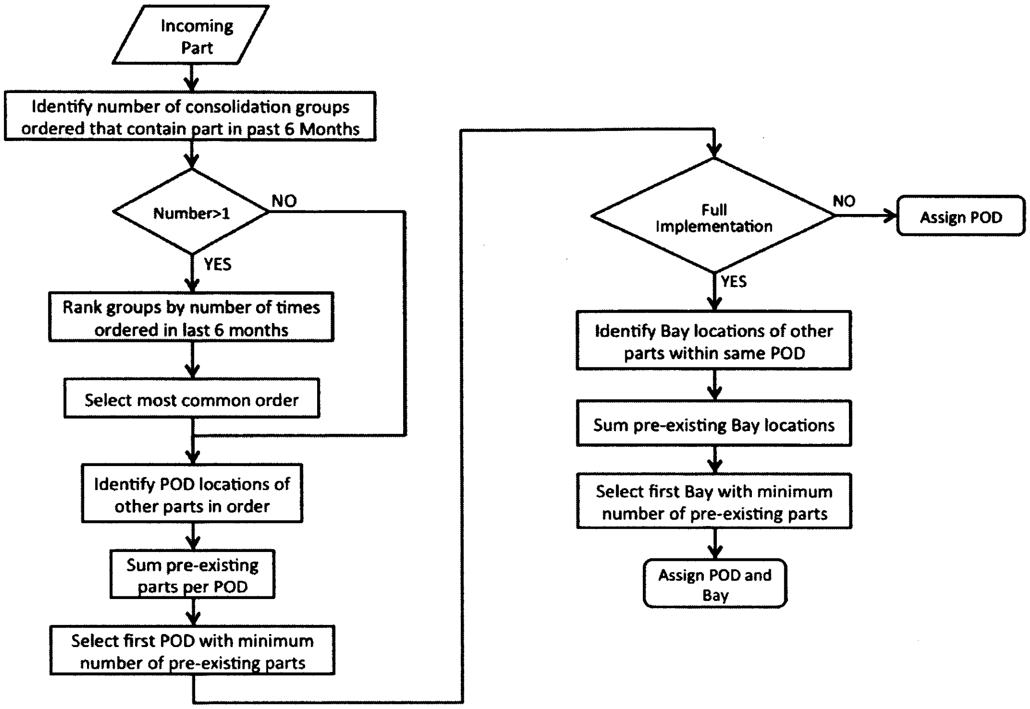


Figure 4-2: The logic required to implement the Order Grouping strategy in the software

Figure 4-4 presents an alternative logic to execute the workload balancing dynamic slotting method. While it assigns a part to the least utilized pod and bay (similar to the snake frequency balancing method in that sense), this approach does not require a pod search sequence and bay search sequence. First, a suitable bin size of the incoming part is determined. The cumulative expected pick frequencies of pods that have the suitable bin size are then evaluated. In order to balance the workload across the pods, the part is set to be put away to the pod which have the lowest expected workload. Within the pods, the part is then put away into the bay that has the suitable bin size and lowest expected workload. Within the bays, the part is randomly assigned a specific bin location of the specified bin type. A point to note is that the mainframe of this logic works entirely different from the existing system. Whereas the existing system finds for suitable empty bins only after the pod is assigned i.e. a pod search sequence is required, the logic presented in figure 4-4 first filters pods and bays that have suitable empty bins for the part and then provides for a pod assignment i.e. no pod and bay search sequences are needed.

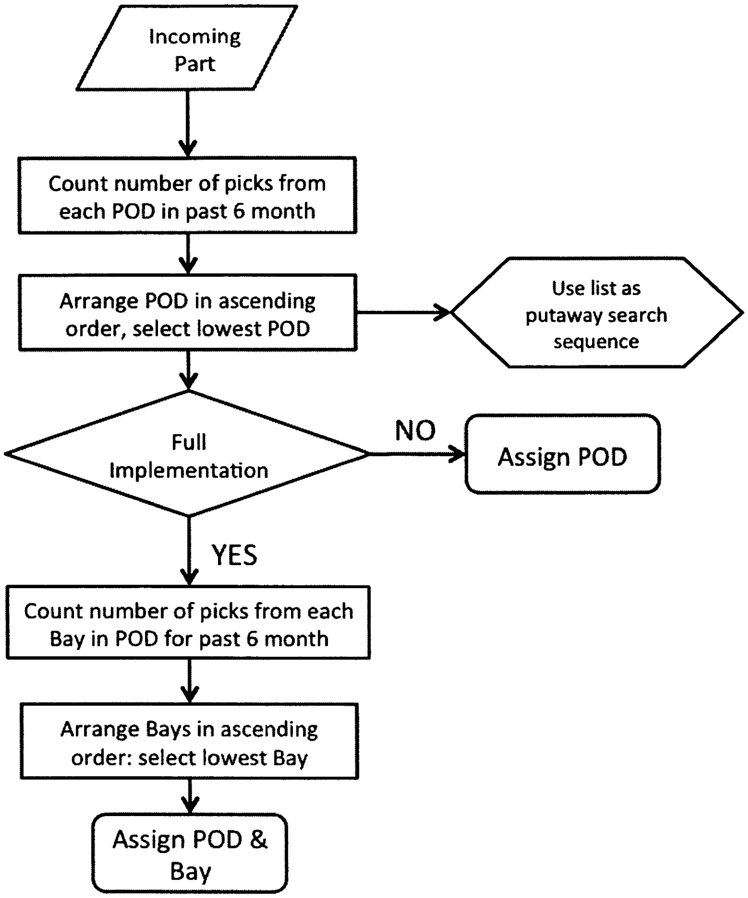


Figure 4-3: The logic required to implement the Snake Frequency Balancing strategy into the software. This workload balancing dynamic slotting method requires a pod search sequence and bay search sequence.

For the workload balancing slotting strategy, the main advantage of using dynamic slotting over the static slotting is that the search sequences of the pods and bays are dynamic i.e. it changes according to the workload of each pod or bay. Static slotting methods typically have a fixed search sequence for the pods. For instance, in the snake frequency balancing static slotting strategy, once the PACI has been defined, say it starts from pod 3, then the software is programmed to search for the empty bins starting from VL03-01, VL03-02, VL03-03, VL04-01, VL04-02, VL05-01 etc. As a result, if pod 5 has the next lowest utilization, second to pod 3, the part will be placed in pod 4 if there is an empty bin in pod 4 instead of pod 5. Therefore, the static slotting method is a less effective method in ensuring that the workload is

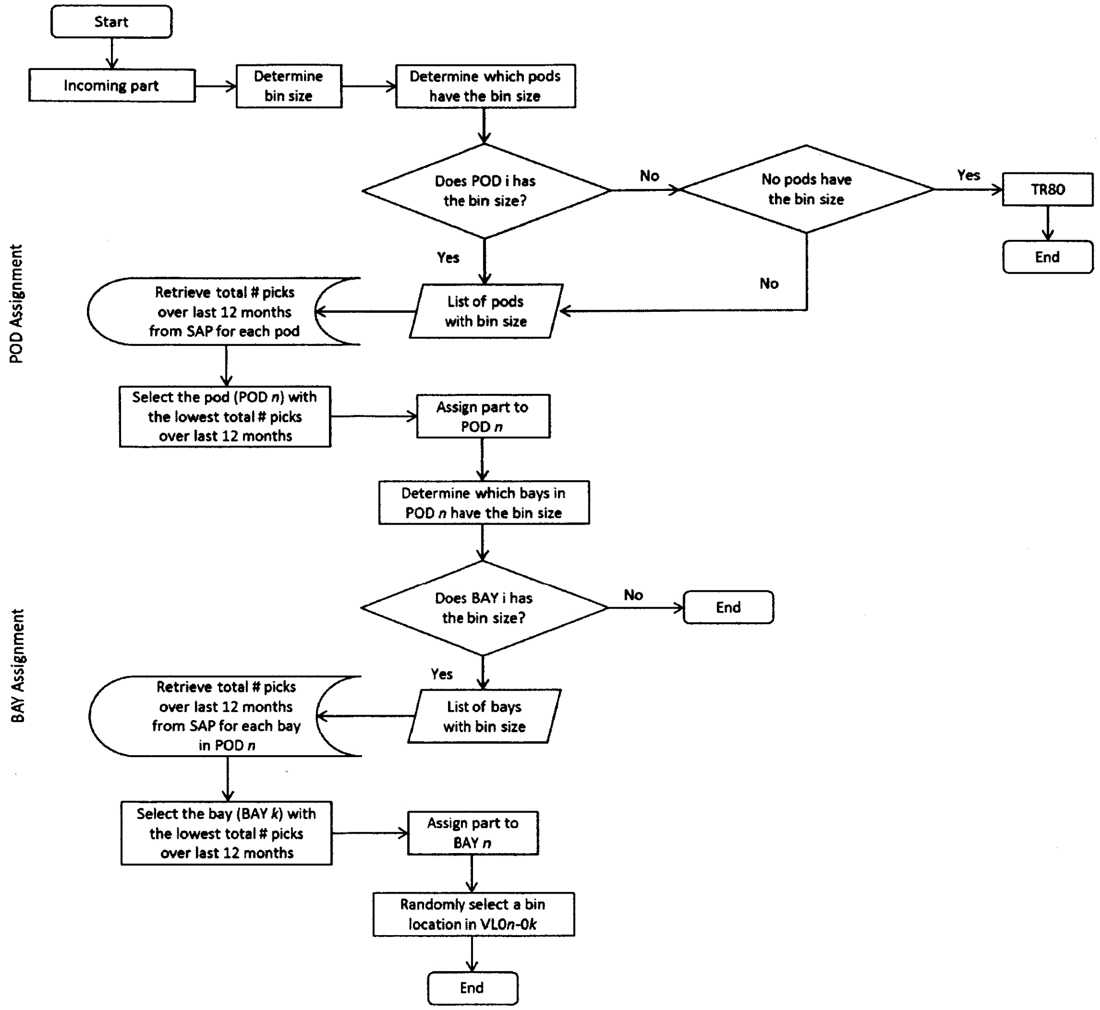


Figure 4-4: Flow chart of an alternative workload balancing dynamic slotting method that does not require a pod search sequence and bay search sequence.

evenly distributed across the five pods.

However, the proposed workload balancing dynamic slotting methods do not account for the part's demand. If the demand of an incoming part is low and is placed into the pod with the lowest utilization, the increase in utilization of the pod is marginal. This becomes a problem if there is a high volume of incoming parts that are low in demand – the pod runs out of spare capacity to store parts that are high in demand in future. As a result, the pod may remain as the lowest utilization and in this case, the workload will not be balanced evenly across the pods. Future work can thus be focused on developing a more robust non-PACI dynamic slotting method which accounts for the part's demand.

4.3.2 Cost-Benefit Analysis of PACI-based Methods for Dynamic Slotting

A cost-benefit analysis of the dynamic slotting methods is performed in this section based on the following criteria:

Programming costs The logic flow charts were submitted to the IT department to get a cost estimation for implementing the PACI-based dynamic slotting methods. There are two quotations provided: i) for full implementation and ii) for partial implementation. Currently, PACIs are for a specific pod however, improving the resolution to a specific bay would be helpful to balancing the workload within each pod. The full implementation costs include the costs of implementing both the pod and bay assignments. The partial implementation costs include the cost of using the current system with only the pod assignments.

Complexity The complexity of a new dynamic slotting method is measured as easy, medium, or hard, based on the number of changes that are required from the original setup.

% Pick time savings The percentage pick time savings is measured over a 9-month period and is calculated based on the difference in number of picks between the slowest pod in the original slotting method and the slowest pod in the dynamic slotting method. It is worth noting that the slowest pod has the highest number of picks. For example, in [15], the highest number of picks is 46,500 in pod 1 for the original method and the highest number of picks is 30,000 in pod 2 for the randomization method. The % time savings between randomization and original is the relative difference, 35.5% [15].

% Improvement $\bar{\sigma}^2_{pod}$ per order The percentage improvement in the pod variance per order quantifies how balanced the workload is across the five pods for each dynamic slotting method is per order. They are taken relative to the original workload distribution and averaged over a 9-month period [15].

% Improvement $\bar{\sigma}^2_{bay}$ per order The percentage improvement in the bay variance per order quantifies how balanced the workload is across the 3 bays within a pod for each dynamic slotting method is per order. They are taken relative to the original workload distribution and averaged over a 9-month period [15].

% Improvement $\bar{\sigma}^2_{pod}$ per day The percentage improvement in the pod variance per day quantifies how balanced the workload is on a daily basis across the five pods for a particular dynamic slotting method, relative to the original workload distribution and averaged over a 9-month period [15].

% Improvement $\bar{\sigma}^2_{bay}$ per day The percentage improvement in the bay variance per day quantifies how balanced the workload is on a daily basis across the three bays within a pod for each dynamic slotting method, relative to the original workload distribution and averaged over a 9-month period [15].

In [15], the last five metrics are determined for the three slotting methods (randomization, snake frequency balancing and order grouping) for the static case. In the static case, specific pod and bay are assigned for each part. In this section, we extend the results obtained in the static case to the dynamic methods.

The programming costs scale with the number of days required to develop, configure and test the new software. The daily programming cost ranges between $\$800-\$1,000$, in this estimate the upper limit was used. The number of days required to develop the new dynamic slotting software depends on two factors – i) whether a full implementation or partial implementation is desired and ii) the complexity of the dynamic slotting method. From Table 4.1, the cost difference between a full and partial implementation is \$3,000. It is estimated that three more days are needed to develop the software application for a full implementation rather than partial implementation.

Table 4.1: Implementation cost estimate from IT department using logic diagrams. The partial Implementation is the cost of implementing the system without modifying the PACI and using only a pod assignment.

Strategy	Full Implementation ($)	Partial Implementation ($)	Complexity
Randomization	28,000	25,000	Easy
Snake	33,000	30,000	Medium
Order Grouping	38,000	35,000	Hard

We highly recommend a full implementation because bay assignments are critical in maintaining a balanced workload across the bays within a pod, which is important for balancing the workload for each picker. By balancing the workload across the bays within a pod, the pick rate can be improved. If a partial implementation is put in place, the consequence is that picks will not be evenly spread across the bays in a pod and the machines could become the bottleneck. If all the parts are located in a single bay, the worker will need to wait for the tray to arrive between picks. Therefore, the costs of the dynamic slotting methods are compared based on the full implementation costs as shown in table 4.2.

Depending on the complexity of the dynamic slotting methods, the time needed to develop a new dynamic slotting software application varies between 10 and 20 days. Typically, configuration of the new software takes 5 days long. An additional 10 days is added due to testing required prior to actual program implementation. From table 4.2, the order grouping method is the most complex and the randomization method is the least complex. In general, each of the three strategies requires a new basic software application that dynamically assigns a PACI to a part. However, both the snake frequency balancing and order grouping methods are slightly more complex because they require additional programs to support the fundamental pod and bay assignments. The snake frequency balancing method requires an additional program to be written that calculates the utilizations of the pods and bays; and the order grouping method requires a separate program that classify which parts are frequently ordered together by the production floor and customers.

Table 4.2: Summary of benefits from each order strategy the % Pick Time Savings is taken over 9 months.

Dynamic slotting method	Randomization	Snake	Order Grouping
Complexity	Easy	Medium	Hard
Cost (\$)	28,000	33,000	38,000
% Pick Time Savings	35.5	35.3	17.5
% Improvement σ^2_{pod} per Order	70.1	69.3	75.3
% Improvement σ^2_{bay} per Order	33.4	33.1	54.8
% Improvement σ^2_{pod} per Day	87.0	85.0	76.7
% Improvement σ^2_{bay} Per Day	87.1	87.4	35.6

The complexity differences among the strategies equate to an approximate price difference of $\$5,000$ between randomization and snake methods, and between snake and order grouping methods. Consequently, the randomization strategy is the cheapest to implement and the order grouping method is the most expensive.

Comparing the improvements in the workload distributions among the three methods, we find that the randomization method gives the highest percentage pick time savings over the 9-month period of 35.5%. In most instances, the randomization method performs better than the snake frequency balancing method, except for the percentage improvement in σ^2_{bay} per day, the latter method performs marginally better (by 0.3%) than the former method. Although the order grouping method shows the greatest improvements in the variances among pods and bays when measured on a per order basis, it has the weakest performance on a daily basis. Because majority of the orders at VSEA have a single order line, the effectiveness of the order grouping method in maintaining a balance workload distribution on a daily basis is limited. Thus, the order grouping method has the lowest percentage pick time savings over the 9-month period of 17.5%. The order grouping method may be more effective if majority of the orders are large-sized i.e. having several order lines (preferably greater than five order lines).

From the cost-benefit analysis summarized in table 4.2, the randomization strategy provides the most benefit at the lowest cost. The least well suited to the company is the order grouping method because most of the orders are small-sized. In addition, developing the software for the order grouping method is highly complex. The final software would also require a high computing power consumption to determine which group the part belongs to for each received part.

The snake frequency balancing method is also a possible option. Although the performance of the snake frequency balancing method is not as good as the randomization method, the differences in performances between both methods are marginal and we consider the overall performance of the snake frequency balancing method to be relatively good. However, for the snake frequency balancing method, an estimated $5,000 of investment is required to create the program to measure the utilization of the pods and bays. The key question here is – *is the additional investment worth it?*

While one may immediately see this as an additional benefit for monitoring how balanced is the workload across the pods and bays, it is worth noting that this can be done easily by retrieving the total number of picks per pod and per bay from EWM. We expect that it takes no longer than a couple of minutes to obtain the data. Hence, the additional $5,000 is not justified if it is seen as a method to monitor the workload distribution of the pods and bays.

However, the additional $5,000 can be justified if it is intended to see a faster improvement in the workload balance across the pods and bays. Since the workload distribution is currently very skewed towards pod 1, even if a dynamic slotting strategy is implemented, a period of time is needed for the workload to be re-balanced. We expect that the snake frequency balancing method will attain this balanced workload faster than randomization because it puts away a part based on the part's expected pick frequency. A fast moving part will be put away into the bay that has the lowest expected utilization[4] and this will raise the workload of the bay. Therefore, we expect that the workload will become more balanced within a shorter period of time if the

---

[4] The expected utilization of a pod or bay is measured based on the cumulative expected number of picks i.e. the sum of the expected number of picks of all the parts in the pod.

snake frequency balancing method is used.

We thus recommend the company to adopt the snake frequency balancing dynamic slotting method primarily because the workload imbalance issue can be resolved within a shorter period of time. Moreover, the predicted workload balance for this strategy is comparable to the randomization approach.

4.4 Future New VLMs

VSEA is considering installing new VLMs in the near future. In the past when they first installed the VLMs at the warehouse, they simply put in as much inventory into the VLMs without considering the workload distributions across the VLMs, as well as placed FISH parts into the VLMs. The poor slotting method has resulted in the unbalanced workload across the five pods and the skewed distribution of picks, which limited the picking efficiency of the VLMs. A framework is developed to ensure that new VLMs that are installed in future will be set up to run efficiently.

For future new VLMs, it is recommended that the company performs the following two steps:

1. Identify fast moving parts. The company should first determine the expected pick frequencies of the parts that are intended to be placed in the VLMs and categorized the parts in terms of their expected pick frequencies.
2. Evenly distribute the parts within the same category across each pod and each bay.

The purpose of sorting the parts and distributing them evenly is to ensure that the workload is balanced across the pods and bays. This is crucial to ensure that all the VLMs are utilized and enables parallel picking to be effective. Once the new VLMs have been set up, the company has to adopt the above-mentioned methodologies to maintain a high number of VLM picks and an efficient VLM system.

4.5 Summary

In order for the warehouse to meet their delivery time goal of 24 hours, the warehouse needs to first establish a couple of new performance metrics in order to provide an actual representation of the warehouse performance and account for any workload unbalance among the VLMs. The two new metrics recommended are:

1. Service level of the warehouse should be measured in terms of either the number (or percentage) of completed orders that has been delivered within the target delivery time of 24 hours, or the maximum delivery time of the set of orders of a particular type.
2. Workload distribution of pods. The workload of a pod can be determined by the cumulative expected pick frequency of all the parts stored in the pod over a certain time period.

Next, in order to sustain a high number of VLM picks, FISH parts in the VLMs have to be removed and fast moving parts have to be transferred from the GL to the VLMs. A review period of six months is recommended.

In order to maintain a highly efficient VLM system, an efficient picking method i.e. pick-and-consolidate should be employed, especially during the busy periods. In addition, dynamic slotting should be employed because it is a more efficient method to balance the workload across the VLMs than the present static slotting techniques. Unlike static slotting, dynamic slotting eliminates the need for re-slotting since it is a self-balancing system.

Comparing the improvements in the workload distributions, randomization and snake frequency balancing methods are the better methods. On the other hand, because majority of the orders are single lines, the order grouping method is the least well suited for VSEA. The main advantage of the snake frequency balancing method is that the time needed to balance the current skewed distribution of workload of the VLMs is shorter than the randomization method. This justifies the additional \$5,000 required to implement the snake frequency balancing method as opposed to the randomization method. We thus recommend the company to adopt the snake frequency balancing dynamic slotting method to balance the workload across the pods and bays.

Finally, if the company wants to install new VLMs in the near future, it is recommended that the pick frequencies of the parts are first identified and categorized based on part velocity, followed by evenly distributing the parts within the same category across each pod and each bay. This will prevent a skewed distribution of picks, which is the main problem faced by company at present.

This page intentionally left blank

Chapter 5

Conclusion and Future Work

Our work is focused on improving the operational efficiency of the warehouse so as to consistently deliver parts on time from the warehouse to the production floor. If parts are not delivered on time, machine lay-downs on the production floor are delayed and the likelihood of a missed shipment increases. Because late shipment is costly to the company, it is critical for the warehouse to meet the target delivery time of 24 hours. However, during the busy periods, the warehouse could not keep up with the target delivery time. The long delivery times are addressed by the simultaneous reduction in workload on GL and improvement in picking efficiency at the VLMs. There are two ways to improve the picking efficiency: i) balancing the workload across the VLMs and ii) using an efficient picking method. Our proposed solutions and their potential benefits are summarized below.

Strategy 1: The workload on GL is reduced by increasing the number of VLM picks. This is accomplished by moving fast moving parts from the GL to the VLMs. The primary motivation for increasing the number of VLM picks is that it is more efficient to pick from the VLMs than GL. Whereas the pick rate of the VLMs is 60 picks per hour, GL is only 20 picks per hour. Additionally, FISH parts have to be removed from the VLMs in order to make room for fast moving parts.

Benefit 1: The VLM workload is expected to increase by 13% as fast moving GL parts are moved to the VLMs [16].

Strategy 2: We recommend to balance the workload across the VLMs by using a dynamic slotting technique – the snake frequency balancing method. The dynamic slotting framework is a preventive approach that constantly seeks to minimize the workload imbalance across the pods and bays. This eliminates the need for periodically re-slotting the VLMs, which is labor intensive. Both the snake frequency balancing and randomization methods show comparable results in terms of achieving a balanced workload distribution across the pods and bays on a daily basis and over a 9-month period. However, we recommend the snake frequency balancing method. The primary advantage of the snake frequency balancing method is that the current VLM workload imbalance can be resolved within the shortest period of time, at a relatively low cost (it costs only $5,000 more than the randomization method) and ease of implementation.

Benefit 2: A 35% time savings can be obtained by using the snake frequency balancing method. Taking into account the 13% increase in VLM picks, the resultant time saving is 23% [15]. The additional time savings will also provide additional time for workers to perform cycle counts and put aways.

Strategy 3: We recommend to employ the pick-and-consolidate method during the busy periods. Pick-and-consolidate is more efficient than pick-and-pass. Because each pod is working independently of each other for pick-and-consolidate, there is no starvation and blocking. In contrast, because the pickers downstream the picking line has to wait for picks to be completed upstream, there is the problem of starvation and blocking for pick-and-pass.

Benefit 3: The expected makespan time savings per order for pick-and-consolidate is 20 min (or 8%). Furthermore, once the workload across the VLMs is balanced in the near future, an order is expected to be picked from four to five pods, rather than from one to three pods. The time savings gained from using pick-and-consolidate is thus expected to increase further.

Additional recommendations to maintain the operational efficiency of the warehouse and monitor the picking efficiency at the VLMs are also provided below.

1. We propose three new additional performance metrics to effectively evaluate the performance of the warehouse and VLMs. In order to effectively measure the service level of the warehouse, the company should report the percentage of completed orders within the 24 hours goal instead of the average delivery times. The flow time of an order should also be measured to evaluate the picking efficiency of the warehouse. This metric closely relates to how fast an order can be completely picked. Finally, in order to monitor the performance of the VLMs, the workload distributions of the VLMs should be measured as well.

2. In order to maintain a high number of VLM picks, we also recommend to review the list of FISH parts and the list of fast moving GL parts every six months. This is a reasonable estimate because the pick frequencies of parts require time to be accumulated. Whereas slower moving parts should be retained in the GL, fast moving parts should be kept in the VLMs. This ensures a higher picking efficiency since picking from VLMs is more efficient than picking from GL.

5.1 Future Work

In our work, we find that about 50% of the fast moving GL parts cannot be moved into the VLMs because they are either too large in size or the lot size is too huge. In the latter case, the size of the parts is relatively small and able to fit into the bins of the VLMs. However, because the order quantities of these parts are large, the consequence is that large sized VLM bins or multiple medium sized VLM bins are required to store the parts. A proposed solution is to reduce the lot sizes of these parts. Future work could examine what is the optimal lot size of these parts and also analyze the impact of reducing the lot sizes of these parts on inventory management of the VLMs.

During our analysis, a noticeable amount of part shortages resulted in late deliveries. The ordering methodologies such as the lot sizes and review periods of the inventory in the warehouse should be examined.

The present evaluation, on which slotting method is the most effective in balancing the workload, is based purely on historical data. Future work could focus on evaluating the actual performance of the slotting method once the method is implemented so as to determine the effectiveness of the dynamic slotting method.

While at present the goal of the warehouse is to ensure that orders are delivered on time within 24 hours to the production floor, a possible longer term goal is to reduce the target delivery time from 24 hours to 8 hours. By reducing the time it takes to deliver parts from the warehouse to the production floor, the assembly lead time of the ion implantation tool can be reduced.

Bibliography

[1] S. Benerjea. (2015, January) Semiconductor stock outlook. [Online]. Available: http://www.zacks.com/commentary/37436/semiconductor-stock-outlook

[2] (2005, May) How to respond to changes in the semiconductor value chain. [Online]. Available: https://www.gartner.com/doc/480913

[3] L. Spencer. (2015, April) Global semiconductor market hit $340 billion in 2014: Gartner. [Online]. Available: http://www.zdnet.com/article/global-semiconductor-market-hit-340-billion-in-2014-gartner/

[4] (2015, April) Worldwide semiconductor foundry market grew 16.1 percent in 2014, according to final results by gartner. [Online]. Available: http://www.gartner.com/newsroom/id/3027717

[5] (2014, April) Worldwide semiconductor manufacturing equipment spending declined 11.5 percent in 2013, according to final results by gartner. [Online]. Available: http://www.gartner.com/newsroom/id/2701117

[6] (2015, April) Gartner says worldwide semiconductor sales expected to reach $354 billion in 2015, a 4 percent increase from 2014. [Online]. Available: http://www.gartner.com/newsroom/id/3034518

[7] (2015, January) Gartner says worldwide semiconductor capital spending to increase 0.8 percent in 2015. [Online]. Available: http://www.gartner.com/newsroom/id/2961017

[8] S. Jain, "Assembly lead time reduction in a semiconductor capital equipment plant through improved material kitting," Master's thesis, Massachusetts Institute of Technology, September 2014.

[9] A. S. Bhadauria, "Production lead time reduction in a semiconductor capital equipment manufacturing plant through optimized testing protocols," Master's thesis, Massachusetts Institute of Technology, September 2014.

[10] Y. Wu, "Operations improvement in a semiconductor capital equipment manufacturing plant: Component level and assembly level inventory management," Master's thesis, Massachusetts Institute of Technology, September 2012.

[11] R. D. Meller and J. F. Klote, "A throughput model for carousel/vlm pods," *IIE Transactions*, vol. 36, no. 8, pp. 725–741, August 2010.

[12] J. A. Tompkins and J. D. Smith, *The warehouse management handbook.* Tompkins press, 1998.

[13] R. de Koster, T. Le-Duc, and K. J. Roodbergen, "Design and control of warehouse order picking: A literature review," *European Journal of Operational Research*, vol. 182, no. 2, pp. 481 – 501, 2007.

[14] G. Dukic, T. Opetuk, and T. Lerher, "A throughput model for a dual-tray vertical lift module with a human order-picker," *International Journal of Production Economics*, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925527315001243

[15] S. D. Racca, "Improving operational efficiency of a semiconductor equipment manufacturing warehouse through effective utilization of vertical lift modules," Master's thesis, Massachusetts Institute of Technology, 77 Mass Ave, Cambridge MA, 02139, September 2015.

[16] P. S. Toor, "Improving operational efficiency of a semiconductor equipment manufacturing warehouse through strategic allocation of parts," Master's thesis, Massachusetts Institute of Technology, September 2015.