

INVENTORY CONTROL FOR HIGH TECHNOLOGY CAPITAL EQUIPMENT
FIRMS

by

Hari Shreeram Abhyankar

B.S. Mathematics
B.S. Economics
Purdue University. 1992

M.S. Industrial Engineering
Purdue University. 1994

Submitted to the
Sloan School of Management
in partial fulfillment of the requirement for the degree of

Doctor of Philosophy in Management
at the

Massachusetts Institute of Technology

February 2000

© Massachusetts Institute of Technology (2000)
All rights reserved

Signature of Author _____
MIT Sloan School of Management
September 15, 1999

Certified by _____
Stephen C. Graves
Abraham J. Siegel Professor of Management
Thesis Supervisor

Accepted by _____

INVENTORY CONTROL FOR HIGH TECHNOLOGY CAPITAL EQUIPMENT FIRMS

by

Hari Shreeram Abhyankar

Submitted to the Sloan School of Management
on September 15, 1999, in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Operations Management

Abstract

Many firms within the high technology capital equipment sector are faced with a situation where effective inventory management is a rather complex and possibly most critical factor to their long-term profitability. Within this thesis we discuss the development of two decision support tools that address some of the unique aspects of the situation faced by Teradyne, Inc., one of the largest suppliers of semiconductor test equipment for the. We also discuss our implementation experiences and develop a framework that calls for a closer interaction with industry, which in our case has provided the motivation and laboratory for this research.

In the first part we discuss our involvement with Teradyne over the course of the past four years. We highlight some of the problems faced by firms that operate in a manner and environment similar to Teradyne. We highlight two of the key problems that we chose for study and discuss their importance to Teradyne. We develop a framework that was used to develop good research problems that had an immediate practical impact. We believe that in the current era of limited public sector funding for fundamental research, our framework may provide some guidance for conducting research projects with greater real world applicability.

In the second part we present a single product inventory model subject to non-stationary demand. We develop exact, as well as approximate performance measures, for this system and develop a relevant optimization problem. Many firms face environments where the underlying demand is non-stationary and there is little visibility of this non-stationary nature. In Teradyne's case this is possibly the most critical problem. We believe that our research provides some insight into the viability of a model that we implemented at Teradyne and permits us to fine-tune the model for greater benefit. From our work we are able to assess the role of intermediate-decoupling inventories in non-stationary demand environments. We believe that our model could also serve as a decision support tool in configuring finished goods inventories as well as intermediate-decoupling inventories in practice.

In the third part we present a robust, computationally efficient methodology to determine the base stocks for components in assemble-to-order environments. This is a rather generic problem faced by many firms within the high technology sector. We present a computationally efficient procedure that outperforms an equal-allocation-policy as well as other heuristic policies that are often used in practice. To this end we believe that our work has significant practical implications.

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

Acknowledgments

I wish to express my deepest gratitude to Prof. Stephen Graves for his patience, guidance, and encouragement, over my tenure at MIT. I wish to thank Dr. Don Rosenfield for his insightful comments and for his help in developing my teaching skills over the past few years. I also wish to thank Prof. Yashan Wang for his suggestions and support.

I owe a great deal to my family and in particular to my wife Deepali without whose patience, the completion of this thesis would have been impossible. I also wish to thank my parents for their support and encouragement.

I wish to thank the folks at Teradyne for providing me with a laboratory to test the ideas contained in this thesis. In particular I wish to thank Jim Wood for his guidance to find the two projects undertaken at Teradyne and the numerous insightful discussions that we had over the past five years. The folks at ICD including Steve Petter, Asa Siggens, Dennis Mauriello, and Jim Desimone were instrumental in facilitating the implementation of the ideas that served as a basis for part II of this thesis.

Finally I wish to thank my colleagues Brian Tomlin, Prof. Sharon Novak, Sean Willems, Amit Dhadwal, and Hemant Taneja for their support over the past few years. I owe a debt of gratitude to Constance Emmanuel for her kind words of encouragement. I wish to thank Sharon Cayley for her guidance. I also want to thank Vivan Mirchandani for listening to my concerns over the past few years.

Table of Contents

1 Introduction.....	14
1.1 Problem context	14
1.2 Performance evaluation and analysis of a single product inventory model subject to non-stationary demand.....	17
1.3A computationally efficient procedure to set base stocks in assemble-to-order environments.....	21
2 Introduction.....	29
2.1 Background information about Teradyne.....	29
2.2 Background information regarding ICD and FDY	30
3 A Process Flow For ICD And FDY	32
3.1 The bill-of-materials structure	32
3.2 The flow of information.....	32
3.3 The master scheduling process	33
3.3.1 MPS process assumptions/facts.....	33
3.3.2 Consequences of the MPS planning process.....	34
3.4 The order procurement process.....	35
3.5 ICD's business environment.....	36
3.6 Diagnosis of FDY's environment	37
3.7 Diagnosis of ICD's environment	38
3.7.1 Discussion	41
3.8 Other key problems for further study	41
3.8.1 Problem 1: Vendors and non-stationarity	42

3.8.2 Problem 2: Effective contracts under allocation	42
4 Description Of The Process Implemented At ICD	43
4.1.1 Stationarity of product options	43
4.1.2 The shape of the cost accrual profile	43
4.1.3 How much of a ramp should one prepare for?	44
4.1.4 Option level target fill rate determination.....	45
4.2 A description of our policy	45
4.2.1 Configuration of the two inventories.....	45
4.2.2 Mapping back to the physical inventory.....	47
4.2.3 Dynamics of the process	48
4.3 The clear need for research.....	49
4.4 Preliminary performance evaluation of our planning strategy.....	50
4.5 A generic strategy for conducting applicable research.....	50
4.6 Conclusion	51
5 Introduction.....	53
6 Modeling Framework.....	54
6.1 Serial line representation	54
6.1.1 Stage related assumptions	54
6.2 Description of demand.....	54
6.3 The inventory control policy.....	54
6.4 Definition of a recurring cycle	56
6.4.1 The low to high transient period.....	57
7 An Optimization Problem.....	71

7.1 Objective function	71
7.1.1 Holding costs.....	71
7.1.2 Objective function	72
7.2 Constraints	72
8 Numerical Experiments And Discussion.....	74
8.1 Overview.....	74
8.1.1 Motivation for selecting the parameter values	74
8.2 Results and discussion.....	75
8.2.1 Tradeoff between TC* per unit and L* versus p for various values of q	75
8.2.2 Sensitivity of TC* per unit and L* versus various λ_H/λ_L	77
8.2.3 Sensitivity of TC* per unit and L* for different cost functions	79
8.2.4 TC* versus various values of L.....	82
9 Summary And Opportunities For Further Work	84
9.1 Deterministic service times.....	84
9.2 More general demand patterns	85
9.3 Deterministic rate change predictability	85
9.4 Modeling expediting capability.....	86
9.5 Modeling market share loss due to stock outs	86
9.6 Permitting stock-outs at the intermediate decoupling inventory.....	86
10 Appendix I.....	88
10.1.1 Sub-problem 1	88
10.1.2 Sub-problem 2	88
10.1.3 Sub-problem 3	88

10.1.4 Sub-problem 4	88
11 Introduction.....	90
12 Model Development.....	91
12.1 Notation:	94
12.2 The related optimization problem:	95
12.3 Discussion	95
12.4 The embedded queueing system.....	96
12.5 Approximations	98
12.5.1 A surrogate for dealing with the $G/D/\infty$ queue	98
12.5.2 The $G/D/m$ queue with superimposed renewal process arrivals	99
12.5.3 Superimposition of renewal processes.....	100
12.5.4 The response time for an end-item order	101
12.5.5 Cases where multiple copies of components are required	102
12.6 The approximation based formulation:.....	103
12.6.1 Discussion:	103
13 Test Models.....	106
13.1 Heuristics for comparison	106
13.1.1 The equal allocation policy	106
13.1.2 Three other heuristics.....	107
13.2 The development of problem instances	108
13.2.1 Model structure characteristics.....	109
13.2.2 Other end-item characteristics	109
13.2.3 Component characteristics	109

13.2.4 Other system characteristics.....	110
13.2.5 Determining the σ_k and μ_k parameters	110
13.2.6 Performance criteria.....	110
13.2.7 Data for the base case for two problems	111
13.2.8 Discussion	113
13.3 Results.....	114
13.4 Some observations.....	115
13.5 An alternate method to evaluate the results	119
13.6 Sensitivity analysis	121
13.7 Discussion	122
14 Conclusion, Extensions And Room For Further Analysis	124
15 References.....	133

1 Introduction

1.1 Problem context

Many firms within the high technology capital equipment sector are subject to a rather unique set of challenges when it comes to materials management. Some critical aspects of the situation that they face can include long procurement lead-times for raw materials, short assembly lead-times, extreme volatility in demand, individually customized orders for products, and short product life cycles. These factors make materials management rather difficult. Traditional inventory control methods often do not take some of these critical aspects into account.

This thesis consists of three sections. In the first section we present a detailed description and diagnosis of the situation faced by Teradyne Inc., a leader in the high technology semiconductor test equipment sector. We present a model/process that was implemented to address the demand volatility faced by Teradyne. We also propose a generic framework within which to conduct new research in a manner that may lead to a greater benefit to both practitioners as well as researchers within the field.

In the second section we develop an inventory control method to address the demand volatility faced by such a firm. Specifically we represent the materials pipeline by a three stage serial line. The first stage represents the external supplier with the longest lead-time, the second stage represents an intermediate inventory, and the third stage represents the finished goods inventory. Our goal is to understand both the role of an intermediate decoupling inventory that is configured to absorb the demand volatility, as well as the role of safety stock within this context. For this model we develop both exact and approximate performance

measures and develop an optimization problem to set both the finished goods base stock levels, as well as the location of the intermediate-decoupling inventory.

In the third section we address the assembly nature of the situation faced by such a firm. As discussed in the first paragraph, end-items are assembled from subsets of components per customer specifications. At Teradyne testers are assembled from PCBs (printed circuit boards) per customer specifications. The assembly lead-times for the testers are negligible relative to the procurement lead-times for the PCBs. We propose a materials management process under which inventories of PCBs are maintained and testers are assembled-to-order per customer specifications. For this system we develop a heuristic procedure to set the base stocks for the PCBs. We also benchmark the quality of the solutions that result from our heuristic against other candidate policies using a detailed simulation study.

Over the course of the past four years we undertook two separate projects at Teradyne. The first project evolved through a consultation with foundry east (FDY), one of Teradyne's PCB assembly divisions. The FDY division assembles-to-order roughly half of all of the PCBs that then get assembled into finished testers by one of several other divisions. The FDY division is not decoupled from the overall production system in the sense that they do not produce PCBs to stock. Rather they assemble PCBs to order per customer specifications from raw material procured from outside vendors. Since there is no formal hedging process at the PCB level and since there is a great deal of uncertainty in terms of customer requirements at the finished goods level, the FDY division has endured a great deal of chaos caused by raw material stock outs. Since the stock out of a 1-cent component can delay the production of a PCB and ultimately a million-dollar tester, it is crucial to effectively manage the PCB

inventory. Our first engagement led to the development of a computationally efficient heuristic to manage the PCB inventory. For the second project we worked with Teradyne's Industrial Consumer Division (ICD). ICD is Teradyne's largest and most profitable finished goods assembly division. Many of the devices that are tested using ICD's testers are ultimately assembled in consumer products such as disk drives, stereos, VCRs, etc. If one were to draw a process map of the supply chain for any one of these consumer products from the raw material stage to the finished product, then ICD's test equipment would be used in one of the most upstream operations. As a result, the bullwhip effect¹ is quite intense and leads to extreme volatility in the demand for the products of ICD. During the course of this project we developed and implemented a materials management process that explicitly takes this extreme demand volatility into account. Preliminary data indicates that the process has led to considerably better responsiveness to their customers. However the academic implications of the underlying model were not well understood. To this end we developed a stylized version of the situation in an effort to both better understand the performance of our process as well as to make improvements to the process that was implemented.

The rest of this chapter discusses the last two parts of the thesis in greater detail. The objective is to position the problems with respect to previous research and to articulate why the problems are worth studying.

¹ The bullwhip effect corresponds to the amplification of the variance in demand that is observed in supply chains.

1.2 Performance evaluation and analysis of a single product inventory model subject to non-stationary demand

As discussed in the introductory section, many firms within the high-technology capital-equipment sector are subject to highly volatile demand, long procurement lead-times for components, and little visibility of the evolution of demand over time. When these three situational characteristics coexist they can lead to a great deal of chaos to such organizations in the absence of effective materials management strategies.

In this section we develop a stylized situation of the situation faced by ICD in which demand alternates between low and high periods in accordance with a two-state recurrent Markov chain. Each state is completely described by a single parameter, the mean rate of demand for a Poisson process. We assume that a finished goods inventory is managed using a state-dependant base stock policy, and we also assume that an intermediate inventory is calibrated using a state-independent base stock policy to decouple the materials pipeline at a certain point in time. This description of the system represents a case where all components with lead-times exceeding a target value are managed using state-independent base stock policies with base stocks set assuming a maximal reasonable rate of demand such as in Simpson (1958). The components with lead-times smaller than the target are then managed using state-dependent base stock policies. Based on our observations the length of the longest component lead-time is roughly equal to the length of an underlying period (of high or low demand). It is clear that under this system it is possible to either have too little inventory when the state changes from the low demand state to the high demand state or too much inventory when the reverse situation takes place. The effect of being ill positioned in the low to high transient period can lead to substantial loss in market share and thus it is necessary to

proactively plan for these rate changes. We propose the use of an intermediate-decoupling inventory to absorb the upward rate changes that take place in such an environment. This inventory is to be strategically located at a point in time that results in system-wide minimal inventory holding cost and provides a suitable amount of protection during this transient period. Implicitly we assume that for such a system there will be a length of time during which there is insufficient FGI to meet the higher rate of demand. However at the termination of this length of time the material released from the intermediate-decoupling inventory will then bring the FGI inventory position to an appropriate level. We refer to the period beginning with such a rate change and ending with the next possibility of a rate change as a transient period. It is this length of time during which there is insufficient material that is of critical importance to this type of a system. This is due to the following anecdotal observation: Stocking out during an upward shift in demand can lead to a significant loss in market-share that can persist for a significantly long period of time.

For our model we develop exact and approximate fill rate expressions for the transient period, develop and test an optimization problem that jointly seeks to minimize the holding costs per unit time subject to constraints on the low period, the high period, and the low to high transient period. We evaluate numerous test formulations based on various combinations of the key problem parameters and gather insights that could have a significant implication on improvement efforts for real world applications.

Researchers have been focusing on developing more realistic demand models for non-stationary demand situations for the past four decades. However their analyses have focused primarily on the demand volatility. In practice a situation where long procurement lead-times, demand volatility and little to no visibility of how demand evolves has only recently become a

reality. Therefore we believe that traditional work within the field has only recently begun to focus on such settings. In the following paragraphs we discuss some of the key papers on non-stationary demand inventory models.

Iglehart and Karlin (1962) develop a discrete time model for a system where the demand process can be completely characterized by a finite state Markov chain. In each period the current state characterizes the one period density of demand. The system is operated using an (S, s) policy. In the paper the authors develop a rather complex computational technique to determine the optimal policy parameters for a linear holding cost setting.

Hillestad and Carrillo (1980) and Hillestad (1982) develop an inventory model based on a non-homogeneous Poisson demand process for military applications. They assume that the instantaneous intensity function for the demand process is known and develop optimization problems to set the base stock levels for a variety of replenishment lead-time distributions.

Johnson and Thompson (1975) prove the optimality of a myopic inventory policy for the case with zero lead-times and when demand occurs according to a Box-Jenkins process.

Graves (1997) develops a model for a integrated moving average process of order $(0,1,1)$ for which an exponentially-weighted moving average provides an optimal forecast. This paper is unique as it combines an underlying forecasting model with a base stock policy. Two of the key findings from this paper are that one requires substantially more safety stock when demand is non-stationary and that the relationship between lead-times and the required safety stock is convex. This is one of few papers that highlights the connection between inventory investments and non-stationary demand

Jennings et al. (1996) develop approximate procedures to determine the required number of servers when demand occurs according to non-stationary renewal processes. Similar to the work of Hillestad and Carrillo the authors assume that the demand evolution is completely specified.

Song and Zipkin (1993) model a single-product, single-stage inventory system subject to a Markov-modulated Poisson demand process. For this system they derive some characteristics of the optimal policies and develop algorithms to compute them. Song and Zipkin extend this work to two echelon depot-retailer systems (1992, 1996). In the first paper they assume that both stages operate under state-independent base stock policies and in the second paper they permit the depot to operate under a state-dependant base stock policy. In both papers they provide procedures to compute the exact steady state performance measures. In many regards our work is most similar to this stream of work with some important distinctions. In our work the intermediate inventory is somewhat analogous to the depot and the finished goods inventory is analogous to the retailer. Under suitable simplifying assumptions we make the external lead-time, i.e., the delivery lead-time from the supplier to the depot a decision variable. Furthermore in our model the finished goods inventory is managed using a state-dependant base stock policy and the intermediate inventory is managed using a state-independent base stock policy. In our work, for an approximations-based formulation we provide a method to determine the optimal position for the intermediate inventory, and the state-dependent base stock levels for the finished goods inventory.

In conclusion we develop a model where demand occurs according to state-dependent Poisson process which depends upon an underlying Markov chain. We study a system with two states, however the extension to multiple states is straightforward. For this system we

provide both exact and approximate performance measures. The approximate performance measures allow us to pose an optimization problem that permits us to explicitly understand the role of a decoupling inventory in non-stationary demand environments. Based on various combinations of the key parameter values we are able to understand better the roles of the intermediate inventory as well as the finished goods inventory in such a setting.

1.3 A computationally efficient procedure to set base stocks in assemble-to-order environments

Effective inventory control in assembly systems has become a problem of ever-increasing practical relevance. This is partly due to the fact that there has been a substantial increase in the number of manufacturing firms that provide custom built products from a set of components that they procure from outside vendors.

Teradyne procures electronic components from outside vendors. These components are assembled into printed circuit boards (PCB) and finally several different boards are assembled into a tester. These testers are assembled to customer specifications. Based on previous work done at Teradyne we suggested that they use base stock policies to manage their PCB supply. By this we do not mean to suggest that they actually assemble the boards to stock, but rather that they plan replenishment orders for electronic components in the form of board kits. In the discussion that follows we consider boards to be components and testers to be end-items.

In some regards the situation faced by Teradyne is very similar to the situation faced by the personal computer (PC) manufacturers. Both Teradyne as well as the PC manufacturers provide their customers with custom built products that are assembled-to-order from components that are procured through outside vendors. Surely, there are several other

examples of firms that operate in a similar manner. There has been a recent resurgence in the study of this and related problems as is evident through the academic literature that has been produced in the past few years.

At a high level we could model such a situation using a two-level bill-of-material. The first level would be identified with the end-items and the second with the PCBs. These PCBs may be unique to a particular end-item or common across several end-items. Moreover, the assembly of an end-item requires the availability of all of its constituent PCBs. Teradyne faces a situation where the replenishment lead-times for the PCBs are much longer than the time required to assemble the end-items (roughly a week for assembly and a range of 10 weeks to 60 weeks for the procurement of components). Due to this aspect of the situation under consideration, we assume that this assembly time is negligible within the context of planning component inventories. We are not attempting to address the issue of detailed scheduling but rather the issues of inventory planning in isolation. A reasonable strategy for such a firm (often observed in practice) is to maintain sufficient component inventories to meet a desired customer service target. In other words such a firm could procure these components to stock and assemble end-items as per customer requests. Based on these observations we characterize performance on the basis of the percent of orders that can be immediately filled from the component inventories (the fill rate). The demand for end-items in such an environment is often stochastic in nature. In the particular division of Teradyne that we studied, the volume of end-items sold is on the order of a 100-150 testers/year. In such a case using a point process description for the demand process could be quite reasonable. In such an environment there is clearly a need to hold component safety stocks to provide the desired service. Based on a study of weekly demand data we noticed that the ratio

of the standard deviation of weekly demand to the mean weekly demand falls within a range between .5 and 4 which does not permit us to assume that the underlying demand processes are Poisson processes. Typically there is a constraint on the system-wide safety stock that such a firm would hold.

Through a series of approximations we develop an optimization problem with an objective of determining the base stock levels for the components that seeks to minimize an upper bound on the expected waiting times for the end-items subject to a budget constraint on the steady state unallocated expected total inventory. We conjecture that a solution that minimizes this bound will result in good fill rates for the end-items. We test this conjecture through a benchmark simulation study in which we compare our heuristic to several alternative policies. We conclude that our method outperforms other candidate policies and is thus effective in meeting our objective.

The problem described above is in no sense new. Both researchers and practitioners have attempted to address the issue under a variety of settings. The key difficulty in analyzing such a system in an exact analytical fashion is that an end-item assembly requires the simultaneous availability of all of its constituent components and the fact that the component availabilities are not independent. This problem is very difficult to analyze even for a single end-item in isolation, unless one makes very restrictive assumptions. In reality one has several end-items to contend with, making this a truly daunting task.

The goal of this work was to determine an effective strategy to set the base stock levels in practice. Rather than developing an exact analysis we elected to use an approach based on as many approximations as were needed. Due to this aspect of our method of analysis we cannot guarantee that the solutions thus generated are optimal; however, in our

test problems it is evident that the quality of solutions generated seems very close to optimal. In practice components can vary considerably on the basis of some key characteristics such as: unit cost, replenishment lead-time, and the number of distinct end-items that use them. Furthermore, if there is component commonality between end-items, it is possibly more cost effective to pool the risk associated with each end-item demand stream when setting the component safety stock levels rather than independently buffering each stream. We propose a model that captures these interdependencies in a fairly simple manner. The effectiveness of the model is then determined through simulation studies.

Early work in this arena dealt with the demonstration of risk pooling due to component commonality. Collier (1982) studied a two-echelon bill of material structure and compared the case of complete component commonality versus the case of no component commonality. In this work the author demonstrated that there is a decrease in safety stock as we move from no commonality to complete commonality. In this paper the author defined a metric based on the number of distinct end-items that use a component. Then by using a version of the Markov inequality the author demonstrated the aggregate safety stock reduction that results by replacing different components that are used in multiple end-items with a single component that could be used in all of the relevant end-items. In this paper the author does not distinguish components on the basis of their value, or provide a methodology to set optimal service levels for the components.

Baker (1985) and Baker et al. (1985) extended the above model to include both common and unique components. The authors compared a two end-item, two component system without commonality to a two end-item, three component system with the end-items sharing a common component. Their analysis demonstrated that there was in fact a risk

pooling effect with the common component; however in moving to the latter situation the safety stock for the unique component increased. This analysis, however, does not seem to extend easily to either more end-items or more components.

More recently Song et al. (1996) derived the exact waiting time distribution in a two-echelon system where the components are made-to-stock while the end-items are made-to-order. In order to make their analysis tractable they assume Poisson arrival processes for the end-items and exponential replenishment lead-times for the components. The exponential replenishment lead-time assumption is not suitable within our context. The work in this paper is primarily for performance evaluation, as their goal was to derive the exact form of the waiting time distribution for end-items. They do however present (but do not test) an iterative procedure for determining the minimal base stock levels for the components that meet a desired service level objective for the end-items. In doing so they are in fact able to relate the service levels at the component level to the service level at the end-item level. However, their model assumes that all component costs are identical.

Hopp and Spearman (1993) suggest a methodology that could be used to set safety lead-times for purchased components. In this paper the authors suggest a methodology for determining the safety lead-times for purchased components in an assemble-to-order environment. A key simplifying assumption in their analysis is to assume that the replenishment lead-times for the components are independent normally distributed random variables. They provide two formulations for this situation and note that managers may not be able to grasp such formulations.

In a paper by Ettl et al. (1996) the authors model a general multi-level bill of material as a queueing network. Each component is managed using a 1-for-1 replenishment policy.

The authors assume that the demands for end-items follow compound Poisson processes and the replenishment lead-times for the components possess arbitrary distributions. The authors formulate a non-linear program with an objective of minimizing the on hand plus WIP inventory subject to end-item service level constraints. The authors provide a conjugate gradient based algorithm for determining the optimal base stock levels for the components. In the paper the authors are able to determine the relationship between the base stock levels and the end-item service levels as they assume compound Poisson demand processes. In our approach we assume more general demand processes which do not permit us to determine an analogous relationship between base stocks and service levels. We formulate a related problem in which the waiting time for end-items is used as a surrogate for the end-item service level in an approximate fashion. We believe that this approach is of value as it captures a wider range of demand processes making it more robust for practical applications.

Song (1998) presents a computational method to determine the end-item fill rates for a multi-product assembly system. The end-items are assembled from different sets of components. The component replenishment lead-times are assumed to be deterministic and the demand processes for the end-items are assumed to follow independent Poisson processes. The focus of this paper is performance evaluation and the issue of optimal base stock determination is not addressed.

Zhang (1997) presents a discrete time multi-item inventory system where end-items are assembled from different sets of components. The author assumes that component inventories are maintained using periodic-review order-up-to policies and the demand for the different types of end-items occurs according to a multivariate normal distribution. In some regards the proposed model is unique in that it permits correlation between end-items within

but not between periods. The author also develops bounds on performance as the exact formulations are computationally cumbersome to deal with.

Gallien and Wein (1999) present a method to set component safety lead-times for a single-item, make-to-stock assembly system with stochastic procurement lead-times for components and assuming that demand occurs according to a Poisson process. This paper deserves special mention because it is one of the few papers to our knowledge that provides a closed form solution to the problem at hand. Their work differs from our work primarily because their objective is to determine the optimal safety lead-times that tradeoff inventory holding costs and backorder costs due to shortages. In our approach we wish to determine the component base stock levels that maximize the individual end-item fill rates subject to a budget constraint on the expected on-hand uncommitted inventory.

Glasserman and Wang (1999) present a simple and effective inventory control policy for a multi-item stochastic assembly system with capacitated suppliers. This work builds on earlier work by the same authors (Glasserman and Wang (1998)) in which the authors use asymptotic methods to develop explicit performance measures.

In summary we develop a heuristic procedure to set the base stock levels for a multi-item multi-component assemble-to-order inventory system with general independent renewal processes to model end-item demands and deterministic replenishment lead-times for the components. Through a simulation study we conclude that our heuristic outperforms a number of other candidate heuristics.

Part I: Teradyne Case Study

2 Introduction

In this chapter we summarize our experiences at Teradyne. In this section we provide some background information regarding two of the key divisions at Teradyne; namely the Industrial Consumer Division (ICD) and Foundry East (FDY). We had the opportunity to complete two separate projects (one with each of these two divisions) over the past few years. In section 3 we provide a process flow of how these two divisions operate individually and interact with one another. This is followed by a diagnosis of their business environment and the problems that result. We then address some of the key problems that Teradyne has faced over the last few years and provide an overview of the two research problems that we chose for study. In section 4 we discuss a methodology that led to a successful implementation of a materials strategy and the role of research to fine-tune this strategy.

2.1 Background information about Teradyne²

Alex d'Arbeloff and Nick DeWolf founded Teradyne, Inc. in 1960; they met while studying as undergraduates at MIT. Alex and Nick foresaw that testing would become a bottleneck to high-volume production of electronic components unless the tasks performed by technicians and laboratory instruments could be automated. The pair rented space above Joe and Nemo's hot dog stand, on the corner of Kingston and Summer Streets in downtown Boston. It was a location they both could walk to from their homes, and convenient to public transportation for all employees. Teradyne's first product was a logic-controlled go/no-go diode tester, the D133. It was introduced in 1961, at a time when semiconductor

² Source: Teradyne website

manufacturers used sophisticated laboratory instruments and slow manual equipment in the factory. Teradyne followed with products for testing resistors and transistors.

In 1966, Teradyne introduced an integrated circuit tester, the J259. It was the first tester to use a minicomputer to control a series of test steps, and it launched the automatic test equipment (ATE) industry. Over the next 25 years, Teradyne focused on expanding its semiconductor test markets and extending its business into new markets that leverage the company's technology, customer relationships, and marketing expertise. By the early 1970s, Teradyne's product line-up included ATE dedicated to memory devices and test systems for electronic subassemblies (printed circuit boards and backplanes). Teradyne also had established itself as a supplier of commercial backplane connection systems. By the end of the decade, Teradyne had a division supplying telecommunications test products, including an automated system for testing telephone subscriber lines. In 1987, the company introduced the first analog VLSI test system, the A500, leading the market then, as Teradyne does today, in the testing of integrated devices that provide the interface between the analog world and digital data.

Teradyne has grown almost 115-fold, from sales in 1971 of \$13 million to sales in 1998 of \$1.5 billion.

2.2 Background information regarding ICD and FDY

The ICD division manufactures and markets systems that test linear and mixed-signal devices. Linear and mixed-signal devices function in a diverse group of commercial products, including personal computer disk drives, stereos, wireless phone systems, VCRs, camcorders, and automobiles. ICD is now focusing on four future markets: wireless, multimedia (the synthesis of television and personal computers), mass storage (disk drives), and the

automotive/industrial telecommunications market. The ICD division is one of Teradyne's largest and most profitable divisions. The cost of one of their testers can run well over \$1 million.

A tester is assembled from sets of printed circuit boards (PCB) as well as other hardware such as a workstation, test head and a mechanical assembly. Most of the PCBs are assembled at two of Teradyne's divisions that they refer to as foundries. FDY is one such division located in Boston. The other such division foundry west (FDW) is located in California. Both FDY and FDW serve most of the finished-goods divisions at Teradyne.

At present the Catalyst family of products represent ICD's flagship product line. These testers sell for an average price of roughly \$1.5 million. To gain some further sense into the scale of the problem that we addressed we provide the following additional information. The Catalyst represents the largest fraction of ICD's revenues, and ICD is the largest (from a revenue and profitability perspective) division of Teradyne. Thus in many regards we had an opportunity to make a very significant impact on Teradyne's operations as a whole.

3 A Process Flow For ICD And FDY

3.1 The bill-of-materials structure

The bill of material has three key levels, an option level, a PCB level, and a piece part level. Different sets of piece parts are assembled into PCBs that are then tested and then assembled into options. Customers' request customized testers by selecting a set of appropriate options. Several different options together with other material such as workstations, test heads, and mechanical assemblies are then put together to form a tester. Each tester is assembled per customer specification and thus requires different options and in turn different sets of boards and components. The ICD division is responsible for assembling the options and other material into a tester while the FDY division assembles the boards from piece parts.

The overall product structure in terms of the number of distinct end-items that can be produced, the number of distinct PCBs and the number of distinct components has an hourglass structure. There are roughly 10,000 distinct components, 200-300 distinct PCBs, and on the order of $^{100}C_{50}$ possible end-items (roughly 100 options and an average of 50 options per tester).

3.2 The flow of information

The ICD division is responsible for maintaining an option level master production schedule (MPS). Using MRP logic these options are gross exploded into a time phased PCB level requirements schedule used by the FDY division to assemble boards. There is very little lot sizing done at the board level partly due to short set-up times for the boards and also since there are no significant capacity constraints.

The FDY divisions as well as the ICD division operate as assemble-to-order divisions. The FDY division holds virtually no safety stocks of PCBs and the ICD division does not assemble finished testers in advance of customer demand. FDY does hold some component safety stocks but the policies to set the stocks do not formally take the demand fluctuations into account.

In this regard the information flow is a top down information flow, i.e., the product divisions such as ICD dictate exactly which PCBs the foundry divisions assemble and when they assemble them.

3.3 The master scheduling process

If one were to view the MPS at any point in time it would be apparent that it stretches out well over a year. This is because the cumulative lead-times (from component procurement to final assembly) are well over a year. The ICD master-scheduling group is responsible for maintaining the MPS. This entails adding new systems at the end of the MPS, rescheduling material if misalignments have occurred (the material required does not fall within the required time period), and adding or deleting miscellaneous material that was not suitably planned for.

3.3.1 MPS process assumptions/facts

- The management policy to date has been to plan the material pipeline in a manner, which assumes that the current demand rate will persist forever into the future.
- This rate is then revised periodically based on market trends.

- A single planning bill³ has been used to fill the material pipeline. Some of the key reasons for using such a planning bill are as follows:
 - The cumulative manufacturing lead-time (the lead-time for the longest lead-time component procured from an outside vendor plus the internal assembly lead-times) exceeds the customer lead-time (the delivery lead-time requested by customers).
 - The customers order completely customized systems.
- It was evident that the planning bill that had been used was an initial bill created prior to the introduction of the Catalyst. The bill was not revised over time to reflect the historical usage of the particular options.
- In addition to this common planning bill, miscellaneous time based inventories of additional options are maintained at a few points out in time.
- The master schedulers determine the timing and quantities of the miscellaneous options that they plan in somewhat of an ad-hoc manner, based on their experience and judgement.

3.3.2 Consequences of the MPS planning process

- The system is always in a reactive mode, i.e., coping with problems as they occur (rather than strategically proactively planning for them)
- Delivery performance to customers has been poor.
- A great deal of expediting takes place to address material shortages.

³ In this context a planning bill represents a “average” system based on historical usage.

- There is also a constant rescheduling of the MPS causing a great deal of chaos in the organization as a whole (from ICD to FDW and FDY as well as their vendors).

The testers in the MPS fall into one of three categories: open, identified or booked.

An open system is one that has not been allocated to a customer, an identified system is identified with a potential customer, and a booked system is a firm order placed by a customer.

3.4 The order procurement process

Marketing personnel, either through direct contact with potential customers or through market analysis, identify potential customers for testers. If contact has been made with a customer, they request a tentative product specification for a tester from the customer. Such a tester is henceforth referred to as an identified tester. A tentative due date is also (when possible) obtained from the customer. An open tester from the MPS that falls within the appropriate period of time (closest to the potential due date) is assigned to the customer. If no such open tester is available, then either a new tester is added to the MPS causing a lot of chaos on the foundry division if the requisite piece part inventory is not available or the due date is negotiated. In any case as the tester rolls closer in time, the customer either books it or the customer does not commit to the tester turning it back into an open tester. If the tester books, it typically does not book as initially specified, i.e., the initial product specification as provided by the potential customer changes. This causes a great deal of rescheduling as unnecessary options have to be removed and previously unplanned options have to be located in the MPS.

The cumulative lead-time is on the order of a year while typically a customer is identified well within the cumulative lead-time. Thus the MPS has to be maintained

containing a suitable number of testers as well as miscellaneous options; the planning of which has to be done well in advance of customer demand.

3.5 ICD's business environment

The demand environment for ICD is very volatile. The mean demand rate per week in a down period can be on the order of a few testers. However this rate can more than double with little to no visibility. We hypothesize that a key reason for this type of volatility could be due to the presence of a very pronounced bullwhip effect. ICD falls at one of the upstream most positions within their respective technology supply chains if one were to refer to a computer or a VCR as a true end-product that requires ICD's testers. A typical up period can last for 1-2 quarters, which could be followed by a 1-2 quarter down period. This sort of volatility has caused a great deal of chaos on the ICD-FDY/FDW production system. The following figure shows the aggregate sales for the ICD division for the past five years (across all product lines; the actual revenues have been masked for confidentiality reasons)

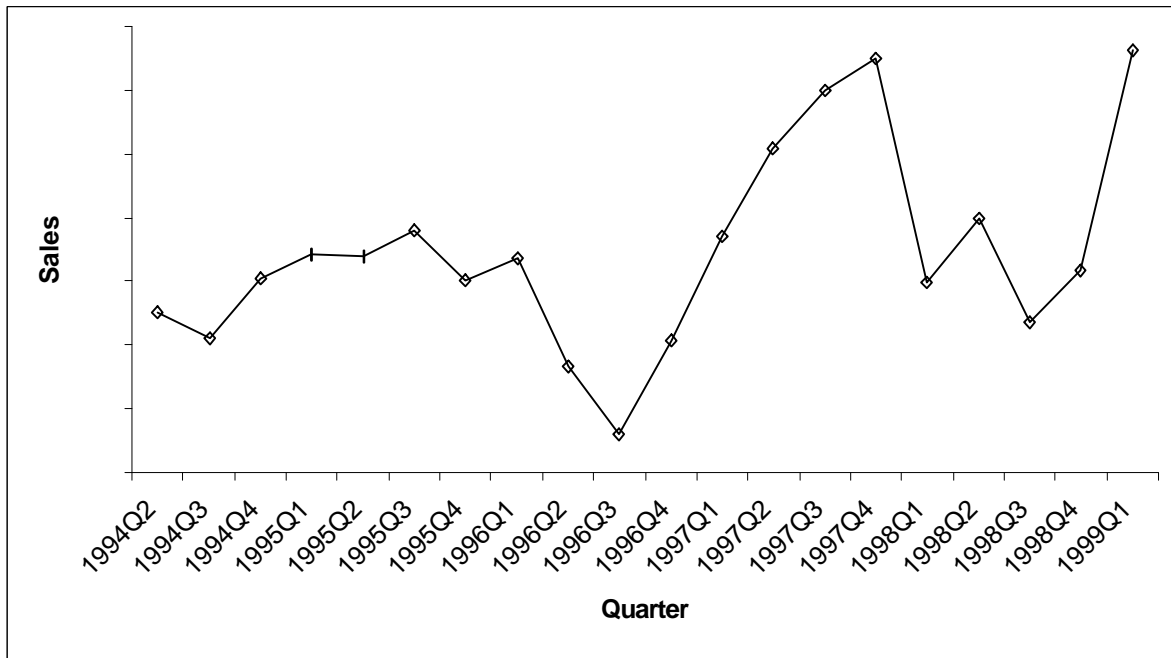


Figure 1: Quarterly Sales Data (Q2 1994 - Q1 1999)

3.6 Diagnosis of FDY's environment

As described in the previous section FDY obtains its production schedule from ICD's MPS using MRP logic. Due to the great deal of rescheduling at the MPS level together with unanticipated option requirements, FDY is often left unprepared to handle all of the PCB requests from ICD. They are unprepared largely due to lack of the appropriate piece part raw material inventory needed to build the PCBs that are associated with the options. When such a piece part is unavailable, a PCB cannot be assembled and in turn this delays the production and ultimate delivery of a finished tester. This sort of a situation can occur even in a stable demand period due to customer order changes. Based on the hourglass shape of the overall product structure, the PCB level represents a strategic hedging point. We proposed a policy where a PCB inventory would be maintained effectively decoupling ICD and FDY's production systems. This would create a window of time to address piece part stockouts that

are inherently occurring. The issue was that there was no prevailing methodology that permitted the setting of base stocks for the PCBs in such a dependent demand assemble-to-order environment. We addressed this issue by developing a heuristic methodology to set these PCB base stock levels. The resulting work is presented in part III of this thesis.

3.7 Diagnosis of ICD's environment

Due to the inherent nature of the demand environment that ICD faces coupled with their policy of planning based on the current level, no formal hedging policy was in place to address the demand volatility. Furthermore the planning bill used by the planners to introduce new open systems was typically the bill created by new product planners at the time when the particular tester line was introduced. Planners often relied on their current experience of option level shortages to plan miscellaneous new options or to make adjustments to the miscellaneous option inventories in the MPS. In this respect the process had been reactive, i.e., options that had run out in the past or were “hard” to obtain through rescheduling were over buffered, while options that had sporadic use were typically not even planned. Very little attention had been paid to collect and utilize demand histories for each of the options. The demand uncertainty that ICD faces can be characterized as having three different attributes: time based – when will a customer require a system, option based – what will a customer require, and level based – what is the aggregate demand rate at the tester level. Based on these observations together with a comprehensive study of the demand data for their mature flagship product (The Catalyst family of products), we proposed a planning process to address each of these sources of uncertainty. Prior to proceeding to our proposed method we present some summary data for the Catalyst systems. The first Catalyst system shipped during the first quarter of 1998. Since then it has replaced their prior flagship product family

(the A-5 series family of products). There was little demand history for the Catalyst as it is a relatively new product. We looked at the combined demand history for the A-5 series and the Catalyst from 2Q 1994 to 1Q 1999 (the sales data in figure 1 on page 36 is a representative summary of this data). A Catalyst system's planning bill consists of a set of options (also referred to as top-level line items). The total number of distinct options at the time of this study was roughly 175. Roughly 150 Catalysts had shipped from the introduction of the product through 4Q 1999 (this formed our data set). The average number of distinct options across our data set of 150 Catalysts was around 50. We noticed that though the aggregate number of Catalysts shipped showed drastic fluctuations from quarter to quarter (such as those seen in figure 1), the frequency-of-use for most of the individual options appeared to be stationary over a suitable time horizon. In the next two figures we present the option level frequency-of-use data. In figure 2 we show the percent of the total number of distinct options by frequency-of-use. To generate the data for figure 3 we took the product of the unit cost for each of the options and multiplied it by the frequency-of-use for that option. The sum of all these costs gives us the average value of a Catalyst (we refer to this as the average material use value). We then created several frequency-of-use buckets, i.e., grouped all those options whose frequency-of-use falls within a particular range such as .1-.2, etc.. We then took the sum of average material use values for each of the options in a range, and divided this quantity by the average value of a Catalyst.

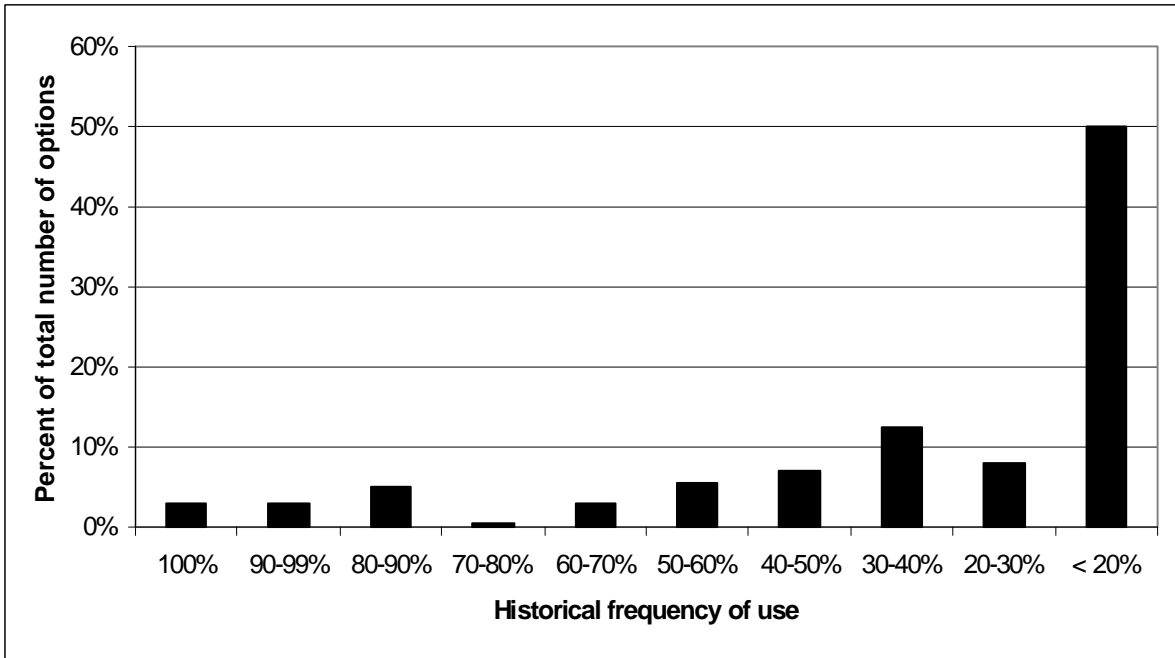


Figure 2: Percent of total number of distinct options versus historical frequency of use

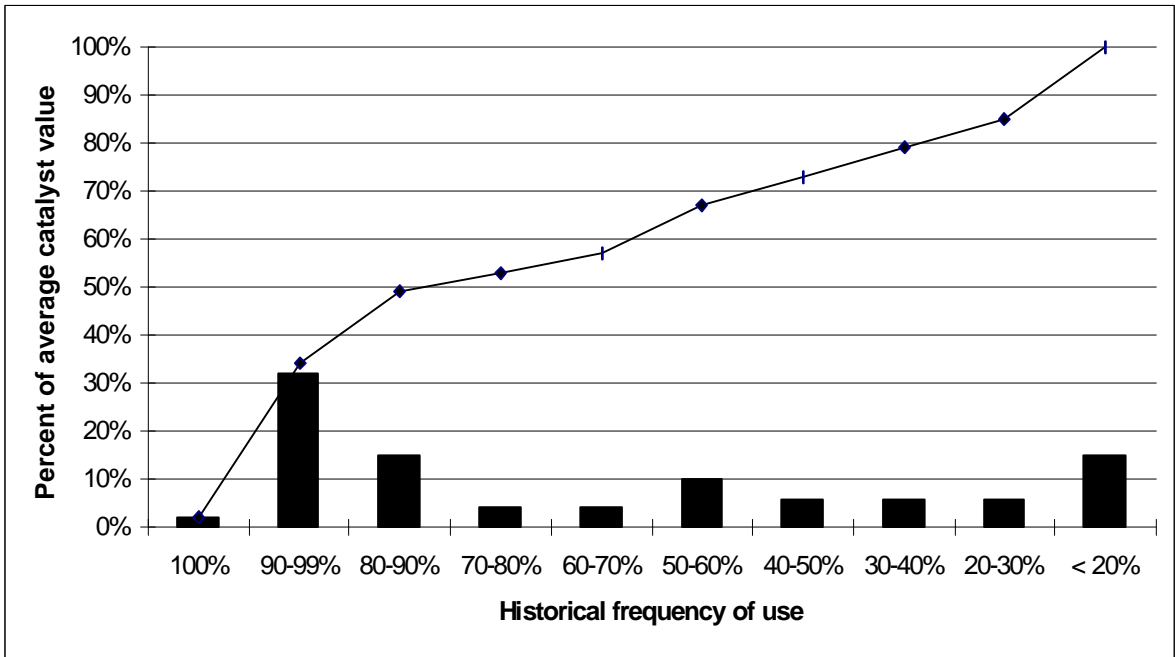


Figure 3: Percent of total number of distinct options versus historical frequency of use

3.7.1 Discussion

From figure 2 it is clear that most Catalysts shipped appear to be quite disparate from one another. However by looking at figure 3 we see that roughly 72% of the cost of an average system consists of options that are used 40% of the time or more. This suggests that there is a sufficient amount of commonality between the systems to tie the aggregate sales levels to the option level usage. We also see that the least frequently-used options (those with frequencies-of-use less than or equal to 20%) account for about 15% of the value of an average Catalyst. These less frequently used options cannot be directly linked to the aggregate sales volumes. We discuss the use of this data in section 4.2.1. Additional data suggests that there is in fact a 80/20 rule in effect, i.e., 20% of the options represent 80% of the costs. As a part of our miscellaneous recommendations we suggested they outsource the assembly and materials management of the option with the highest extended dollar usage (average use multiplied by the cost per unit).

In the next section we outline our solution approach together with a discussion of the hurdles that we had to overcome in order for ICD to accept and implement our proposed process. We also developed a stylized version of the situation to further understand the theoretical consequences of our process as well as to fine-tune the process. In part II of this thesis we present this stylized version that has permitted us to assess some of the qualitative tradeoffs between the decision variables from our process.

3.8 Other key problems for further study

As outlined in the previous two sections we selected two of the primary problems faced by ICD and FDY for further study. In this section we present two other problems that could also be relevant to Teradyne.

3.8.1 Problem 1: Vendors and non-stationarity

The type and size of a vendor plays a key role in their ability to react to sudden swings in business from their customers. For example, a mom and pop shop with 50 employees producing specialized power supplies may have to double its workforce if its capacity is constrained and the demand placed on it doubles. However when the demand falls back to the original levels, it will have excess capacity that it may not be able to sustain. On the other hand a large supplier of commodity ICs may welcome the opportunity to sell more components in a ramp. Clearly the non-stationary nature of Teradyne's demand has different consequences for different suppliers. Developing effective strategies for materials management should take the supplier characteristics into account.

3.8.2 Problem 2: Effective contracts under allocation

Some vendors provide only a limited amount of products to potential customers largely due to capacity constraints. This sort of a situation is coined as "being on allocation" by buyers. Developing effective contracts under this setting is an important problem that could be studied by researchers with vast consequences for many firms in industry.

4 Description Of The Process Implemented At ICD

We developed a process to address the three key sources of demand uncertainty outlined in section 3.7, i.e., level based, mix based, and time based. Prior to doing so we conducted a detailed study of the demand history as well as the relevant material costs as they pertain to ICD's flagship product line. Some of our observations were as follows

4.1.1 Stationarity of product options

While the aggregate tester demand volumes varied quite drastically over time the number of units of the options (or fractions of units of the options) per tester was quite stable for a large percent of the options (there were a few notable exceptions). The remaining options fell into a low frequency of use category that could not be tied directly to the number of testers shipped in any given period of time.

4.1.2 The shape of the cost accrual profile

The lead-time cost accrual profile represents the cumulative value of the components purchased as a function of time. We noticed that such a profile was relatively flat and then it spiked up rather quickly as it approached $t = 0$ which implies that the long lead-time components represent a small portion of the total material cost of a tester. To this end we picked the point of inflection closest to time $t = 0$ as a strategic hedging point to stock additional testers in order to meet sudden upward spikes in demand. The following figure (figure 4) provides a representative lead-time cost-accrual profile

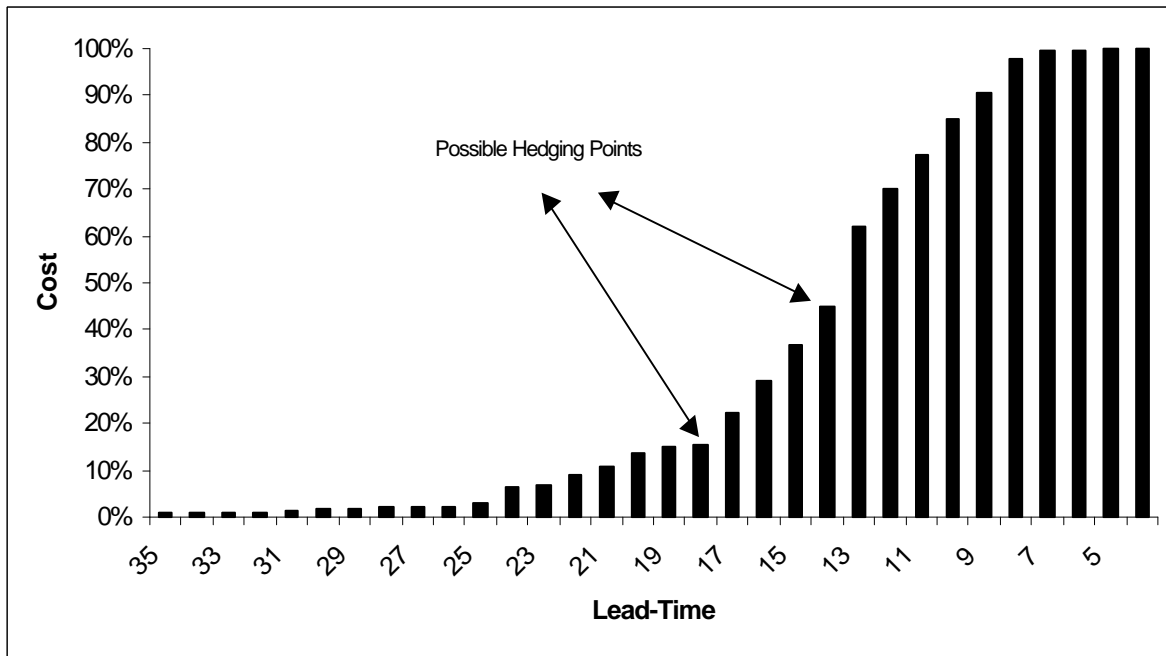


Figure 4: A Sample Lead-Time Cost Accrual Profile (for an average Catalyst)

4.1.3 How much of a ramp should one prepare for?

Traditional forecasting models were known to be inadequate to fit ICD's demand profile due largely to the extreme demand volatility. Thus a different approach which called for expert judgement was undertaken. During this process the divisional VP, the VP in charge of Teradyne's operations, marketing directors as well as materials managers with the aid of decision tree came to a consensus judgement to set the maximum reasonable demand rate to plan for. When setting this level the panel of experts took into account the impact of not meeting a ramp. ICD's market can be characterized as one with a few very large competitors each trying to capture the other's market-share. The impact of not meeting a ramp can (and has in the past) lead to a significant loss of market share which can have serious consequences.

4.1.4 Option level target fill rate determination

A study of the data determined that the fill rate at the option level was anywhere between 5%-90% depending on the particular option. To this end due to a sense of the product-form nature of end-item versus component fill rates in assembly systems, and the implied costs as a basis of fill rates (using a constraint on the system wide inventory investment), a 97% fill rate was proposed as a target fill rate for all of the relevant options. The set of all options can be partitioned into two subsets. In the first subset we have those options for which a customer order either requires one unit of the option or does not require the option at all. In the second subset (roughly 20% of the options) we have those options for which customers can require multiple units of the particular option. For the first subset we fit a binomial random variable to the data to characterize the usage across several systems. For the second subset we fit a normal distribution to characterize the usage across several systems.

4.2 A description of our policy

4.2.1 Configuration of the two inventories

We configured the hedging point, henceforth to be referred to as the intermediate-decoupling inventory, assuming this maximum reasonable demand rate. For discussion purposes let m represent the length of the material pipeline, i.e., the point in time (in the future) where new systems are added. Here m equals the internal assembly lead-time plus the lead-time for the longest lead-time component. We assume that the intermediate decoupling inventory is located L units of time out in the future. We configured this inventory at the maximal run rate, i.e., $\lambda_H(m-L)$ (assuming that the maximal run rate accounts for variability in

the demand process during a stable period). Here λ_H is the maximal run rate, and $(m-L)$ is the replenishment lead-time for the intermediate-decoupling inventory.

We then used a 97% fill rate together with a characterization of option-level demand for the frequently used options (with historical frequencies of use exceeding 20%) to plan for $\lambda_H(m-L)$ testers. For the infrequently used options we planned either 1 or 2 units for each option. The FGI is maintained at a base stock level corresponding to $\lambda_{\text{current}}L$ units.

To illustrate the mechanics of the inventory system observe the following diagram

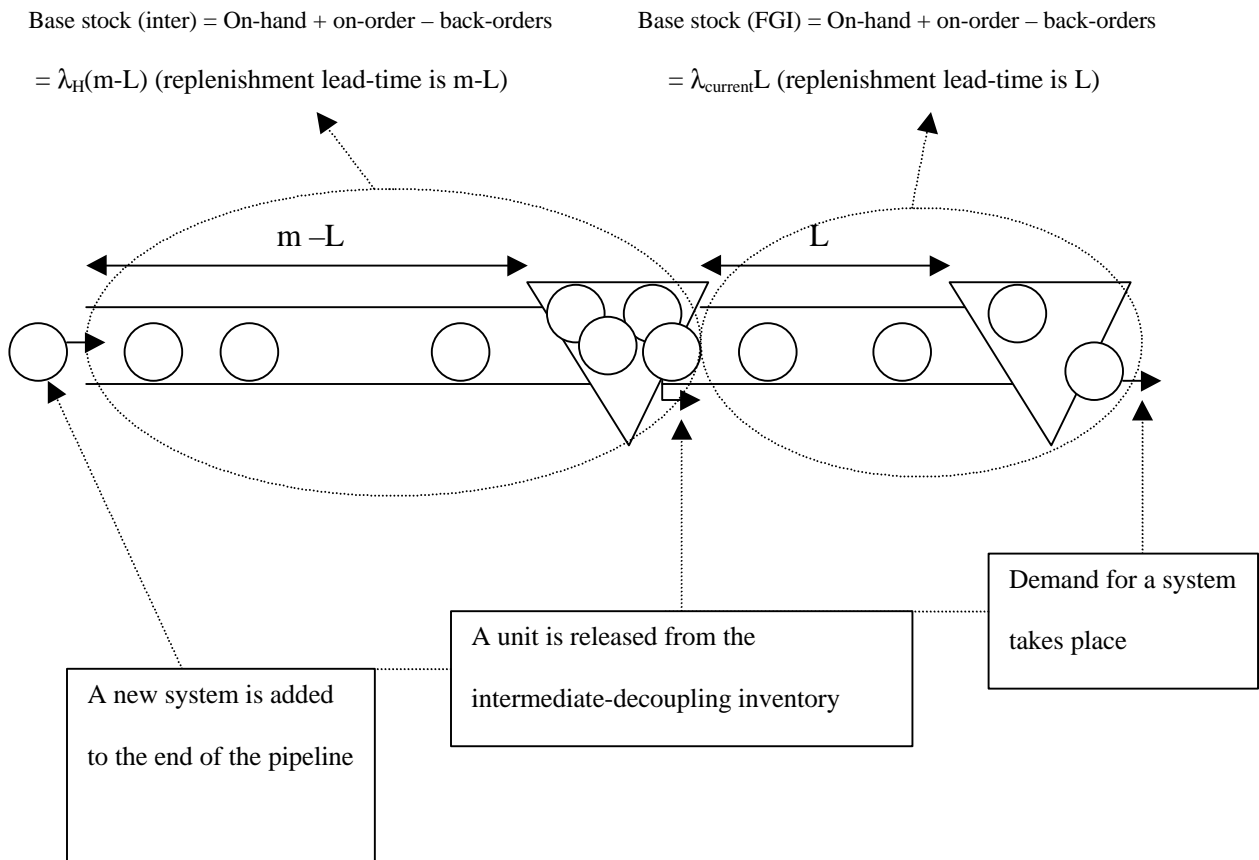


Figure 5: Mechanics of inventory policy

The triangle on the extreme right represents the FGI, the other triangle represents the intermediate-decoupling inventory. The pipeline going into the intermediate-decoupling inventory represents the pipeline inventory between the end of the material pipeline to the intermediate-decoupling inventory and the pipeline between the two inventories represents the in-transit material between the intermediate-decoupling inventory and the FGI. The circles denote individual systems.

Note that the above diagram represents all situations except for the transient situations. During the high to low transient period, systems will not be released into the intermediate-decoupling inventory to the FGI inventory pipeline until the appropriate low period base stock level is reached. At the beginning of a low to high transient period a batch order will be released from the intermediate-decoupling inventory en route to the FGI with a batch size such that the appropriate base stock level for the high period will be attained L time units later.

4.2.2 Mapping back to the physical inventory

As was discussed previously, the contents of each circle are identical, e.g., they represent a common planning bill. As a planning bill moves closer in time, i.e., from left to right in figure 5, it triggers orders for components with outside vendors. Consider a production system with $m = 16$ weeks, $L = 4$ weeks, $\lambda_H = 4$ systems/week, $\lambda_D = 2$ systems/week, $z = 2$ (protection level). Assume that the current demand period is a low demand period. Then the number of units on-hand plus on-order for the intermediate-decoupling inventory will be $(4*(16-4) + 2*(4*(16-4))^{1/2})^4 \approx 62$ and the number of units on-

⁴ The base stock level assuming Poisson demand and using a normal approximation

hand plus on order for the FGI will be $2*4+2*(2*4)^{1/2} \approx 14$. Now consider a component with a replenishment lead-time equal to 9 weeks. Until a planning bill (circle in figure 5) crosses the lead-time for the component, 9 weeks away from the FGI, the component is not ordered. So in effect we can think of the planning bills in the pipeline as pallets, until the pallets crosses the replenishment lead-time for our component, the component is not “added” to the pallet. Now this “adding” is not literal adding, but rather “virtual” adding as it represent the placement of an order with the vendor for our component. Thus although there are 62 pallets either on-hand or on-order between the end of the pipeline and the intermediate-decoupling inventory, only those pallets beyond the 9 week threshold have the component in question “added” to it.

4.2.3 Dynamics of the process

The intermediate-decoupling inventory is sized in a manner to respond to a sudden demand-rate change. Management decides that the rate has changed, then they authorize the release of additional material from the intermediate-decoupling inventory to ramp the FGI to a new base stock that reflects the new demand rate. We use the following example to clarify the release decision. Consider a system with one option, call it option I. Assume that a Catalyst will use 1 unit of option I with probability p_1 . Assume for convenience that the FGI is maintained at a run-rate of 2 units per week. For convenience assume that at the time the demand-rate change takes place the FGI base stock level for option I = $\text{argmin}(j \mid \text{binomial}(2*L,j,p_1) \geq .97)$. Here $\text{binomial}(x, y, z)$ corresponds to a binomial CDF, with x equal to the number of trials, y equal to the number of successes, and z corresponds to the probability of a success on any given trial.. Now assume that the demand rate changes and the appropriate run-rate for the FGI becomes 4. Thus the base stock for option I now increases by $\text{argmin}(j \mid$

$\text{binomial}(4*L,j,p_1) \geq .97) - \text{argmin}(j \mid \text{binomial}(2*L,j,p_1) \geq .97)$. Thus, the FGI places a replenishment order for the amount on the intermediate-decoupling inventory. We assume that the intermediate-decoupling inventory has been sized so that it can respond to such a request. Thus, when a demand rate changes, the FGI base stock will be restored to its new target after L time units. In order to do this, we need to size the intermediate-decoupling inventory assuming the high demand rate and assuming a high fill rate, e.g., 97%. If the rate changes from the current level to a lower level, then the same calculation is done and the excess material is de-expedited to the intermediate-decoupling inventory. The effect of this de-expediting is somewhat unclear at the piece-part level. This is due to piece-part lot sizing and the effect of order cancellation requests with ones vendors. Suppose that the target base stock level for option I goes from 20 to 15, thus there are 5 excess units of option I either on-hand at the FGI or on order between the intermediate-decoupling inventory and the FGI. If all 5 units are on-hand then de-expediting will have no effect as all requisite component inventory has already been purchased and delivered. However if say 2 of the 5 units were on-order and closer to the intermediate-decoupling inventory than the FGI, the de-expediting may cancel the purchase of some of the piece-part inventory corresponding to the 2 extra units.

4.3 The clear need for research

In setting up this policy we were unable to characterize the tester service level (even without assuming the assembly nature of the system) that could be expected from the FGI. As a consequence we arbitrarily chose a 97% fill rate target for the intermediate-decoupling inventory and hoped that this would provide a satisfactory service level over the ramp period. To this end the performance measures developed in section II become directly applicable to gain some quantitative insight into the expected performance of this policy. Furthermore we

selected the intermediate-decoupling inventory hedging point by a back of the envelope analysis of the cost accrual profile. We envision using the optimization model developed in section II to perhaps revise this hedging point.

4.4 Preliminary performance evaluation of our planning strategy

The timing of the implementation of our planning strategy could not have been any more opportune. Soon after our planning strategy went into effect, a ramp in demand after nearly a yearlong slump abruptly took place. The demand rate exceeded the maximal rate that was set by management; however the intermediate-decoupling inventory was sufficient to capture nearly all of the additional demand.

Anecdotally it also seems as though master schedulers have had little difficulty to find the requisite options in the MPS. In the past nearly every order caused some chaos, as at least one option was not available in the required time window within the MPS.

The incremental inventory investment to permit the formation of such an intermediate inventory has not been substantial. This is primarily due to the fact that the MPS had already contained a sufficiently high percentage of this requisite inventory primarily in the form of expensive options that had been over planned through the use of an outdated planning bill.

4.5 A generic strategy for conducting applicable research

The mode of operation that we used to go from concept to implementation has permitted us to quickly provide benefit to our industrial customer and has also fostered the development of new research. Our approach has three principal steps; namely, academic consulting, research, and revision. During the first step an academic's version of a back of the envelope study is conducted during which a tentative model/strategy is developed possibly using simulation to speed up the development. During the second step this model/strategy is

then studied through a more detailed analytical research phase where we attempt to gain insights into our model/strategy. Finally in the last step we attempt to revise our model/strategy using the insights gained in step two. For example the work from section II could be used to revise the base stock quantities and/or the location of the decoupling inventory to improve the performance of the process that we put into place.

The benefit of this model of operation in conducting research is that industrial customers often require a shorter lead-time for the development of a solution to their impending issues/problems than the required lead-time for developing new research. Thus by providing an implementable solution based on insights gained from preliminary models, one is able to create a forum for future research. The preliminary models developed in the academic consulting phase are perhaps more sophisticated than those that might be provided from other sources.

4.6 Conclusion

In this section we have presented the diagnosis of an actual production system that has permitted us to develop and apply new research in a manner that has been beneficial both from an application standpoint as well as a research standpoint. We hope that the generic process discussed in section 4.5 could foster a richer interaction between academics and industry.

Part II: Performance Evaluation and
Analysis of a Single Product
Inventory Model Subject to Non-
Stationary Demand

5 Introduction

In this section we develop a single product inventory model subject to non-stationary demand. As outlined in the introductory section the goal of this work to develop a stylized version of the situation faced by firms such as the ICD division at Teradyne in order to gain insight into the role of a decoupling inventory as well as a finished goods inventory. As was previously discussed, these firms face a situation where demand is very volatile, the lead-times for raw materials are very long, and there is little visibility of the evolution of demand.

This section is organized as follows. In section 6 we present the modeling framework. In section 7 we present a math program that jointly seeks to determine the location of the intermediate-decoupling inventory and the FGI base stock levels. In section 8 we present the solutions of the different problem instances that were solved and the insight gained. In section 9 we discuss some possible extensions and directions for future work.

6 Modeling Framework

6.1 Serial line representation

We represent the material pipeline by a two-stage inventory system. The first stage represents an intermediate decoupling inventory that orders material from an outside supplier. The second stage represents a finished goods inventory that orders material from the intermediate inventory. There are two relevant pipeline inventories as well: one between the supplier and the intermediate decoupling inventory and one between the intermediate decoupling inventory and the finished goods inventory.

6.1.1 Stage related assumptions

We assume that each stage is uncapacitated and that the external supplier is completely reliable. Excess demand at each of the stages is backlogged.

6.2 Description of demand

The external demand evolves according to an underlying discrete time, recurrent Markov chain with two states, with self-transitions. Each state is completely described by a single parameter, the mean rate of demand for a Poisson process. We assume that there is no advance warning of the rate change.

6.3 The inventory control policy

The intermediate decoupling inventory is calibrated assuming that the high rate of demand. That is, the base stock level is set to make the probability of a stock out at this stage suitably small, assuming Poisson demand with mean equal to the high rate of demand. We assume that any stockouts that take place at the intermediate-decoupling inventory will be handled through expediting. Thus we assume that stockouts at the intermediate-decoupling

inventory are not a factor in calculating the fill rates of the FGI. This assumption is similar to the “maximum reasonable demand” assumption from Simpson (1958).

The finished goods inventory follows a state-dependant base stock policy (1-for-1 replenishment) with a base stock level for the high period and a base stock level for the low period.

When the rate changes from low to high, the FGI places a batch order to bring its inventory position up-to the high-period base stock level. When the rate changes from high to low, we let the FGI return to the low-period base stock level. Thus, the next several demands are not replenished so as to bring the inventory position to the low-period base stock level.

Here L (equal to the replenishment lead-time for the FGI) is a key parameter as it directly effects the responsiveness of the system in the face of low to high transitions in demand. In the next sections we develop expressions to understand the relationship between the FGI fill rates, L , and the FGI base stock levels.

Notation

λ_t	The instantaneous demand rate at time t (units/week). The high rate is denoted by λ_H , and the low rate is denoted by λ_D
OH[FGI]	The averaged expected on-hand finished goods inventory over a cycle
OH[Inter]	The averaged expected on-hand intermediate inventory over a cycle
OH[Pipe1]	The averaged expected in-transit inventory from the supplier to the intermediate inventory over a cycle
OH[Pipe2]	The averaged expected in-transit inventory from the intermediate inventory to the finished goods inventory over a cycle

L	The distance in time units between the intermediate inventory and the finished goods inventory, expressed in weeks (need not be an integer number of weeks)
M	The length of a period (in weeks) before a transition takes place, for the underlying Markov chain for demand rate transitions
S_H	The high period finished goods inventory base stock level
S_D	The low period finished goods inventory base stock level
S_I	The intermediate inventory base stock level
P	Probability of a high period to a low period transition
Q	The probability of a low period to a high period transition

6.4 Definition of a recurring cycle

Recall that the FGI is maintained using a state-dependant base stock policy. We assume that the underlying period length (m) is suitably long so that by the end of the high to low transient period, the target base stock level for the low period is reached, i.e., m is the length of time between transitions. As was stated earlier the inventory control policy during the high to low transient period is such that the first $S_H - S_D$ demands are ignored. Here we are simply making the assumption that by the end of this transient period, at least $S_H - S_D$ demands have occurred (over one period that has length m).

With this assumption we define a cycle starting with a transition from a high period to a low period and ending just prior to the next such transition. By defining a cycle in this manner a cycle will always begin with a high to low transient period followed by a geometrically distributed number of low periods, a low to high transient period, and a geometrically distributed number of high periods. We proceed by analyzing the performance of each of these sub-periods.

Claim 1: The expected length of a cycle is given by $m^*(2 + (1-p)/p + (1-q)/q)$

Proof: A cycle begins with a period containing a high to low transient period then the distribution of the number of additional low periods is given by $f(x=i) = q(1-q)^i, i = 0,1,2,\dots$, and $E[x] = (1-q)/q$. This is followed by a transition to a high period and the additional number of high periods has the same distribution but with parameter p rather than q

6.4.1 The low to high transient period

As discussed earlier, for the first L units of time during this period, S_D will be the base stock level. The instant when L units of time elapse, a batch order will arrive from the intermediate-decoupling inventory to bring the base stock level up-to S_H . In the following claims we develop the key performance measures for this system.

Claim 2: The fill rate during the first L units of time of the low to high transient period is given by:

$$\sum_{n=0}^{\infty} \left[\frac{1}{n} \left[\sum_{i=1}^{\min(S_D, n)} E[1_{Z_i | N=n} | N(L) = n] \right] \frac{(I_H L)^n}{n!} e^{-I_H L} \right]$$

Where

$$E[1_{Z_i | N=n} | N(L) = n] = \begin{cases} \sum_{j=0}^{S_D-i} \int_0^L \frac{(I_D(L-z))^j}{j!} e^{-I_D(L-z)} \frac{(L-z)^{i-1} z^{n-i} n!}{L^n (n-i)! (i-1)!} dz, & i \leq S_D \\ 0, & i > S_D \end{cases}$$

Proof:

We begin by conditioning on the number of arrivals $N = n$ during the first L units of time. Define the indicator random variable $1_{Z_i | N=n}$ as follows:

$$1_{Z_i | N(L)=n} = \begin{cases} 1 & \text{If } i\text{th demand is filled from stock} \\ 0 & \text{Otherwise} \end{cases} \Bigg|_{N(L)=n}$$

Then

$$P\{1_{Z_i | N(L)=n} = 1\} = P\{\text{No more than } S_D \text{ arrivals in } (t_i - L, t_i)\}$$

where t_i is the arrival epoch of the i th demand during the first L units of time. This follows since the on-hand inventory level at time t_i depends only upon the cumulative number of demands over the preceding L units of time⁵. Now this implies that

$$P\{1_{Z_i | N=n} = 1\} = \begin{cases} P\{\text{No more than } S_D - i \text{ arrivals in } L - t_i \text{ units of time}\}, & i \leq S_D \\ 0, & i > S_D \end{cases}$$

This follows by construction since we are considering the i th arrival that falls within the first L time units of the low to high transient period. Now this arrival is not delayed if there were no more than $S_D - i$ arrivals during the final $L - t_i$ time units of the immediately preceding low period. Now the density function for the arrival epochs t_i given n arrivals in L time units ($N(L) = n$), is given by (this is a standard result, a reference to which may be found in chapter 2 of Gallager (1996)):

$$f_{t_i}(x | N(L) = n) = \frac{x^{i-1} (L-x)^{n-i} n!}{L^n (n-i)! (i-1)!}, 0 \leq x \leq L$$

We substitute $L - z$ for x to get

$$\therefore P\{1_{Z_i | N(L)=n} = 1\} = E[1_{Z_i | N(L)=n}] = \int_0^L \sum_{j=0}^{S_D-i} \frac{(I_D(L-z))^j}{j!} e^{-I_D(L-z)} \frac{(L-z)^{i-1} z^{n-i} n!}{L^n (n-i)! (i-1)!} dz$$

$$\text{The fill rate} = \sum_{n=0}^{\infty} \left[\frac{1}{n} \left[\sum_{i=1}^{\min(S_D, n)} E[1_{Z_i | N=n} | N = n] \right] \frac{(I_H L)^n}{n!} e^{-I_H L} \right]$$

$$\text{Here } E[1_{Z_i | N=n}] = \sum_{j=0}^{S_D-i} \int_0^L \frac{(I_D(L-z))^j}{j!} e^{-I_D(L-z)} \frac{(L-z)^{i-1} z^{n-i} n!}{L^n (n-i)! (i-1)!} dz$$

This expression is computationally time consuming to evaluate so we develop a discrete time approximation as follows:

Claim 3: A discrete time approximation to the fill rate during the first L units of time of the low to high transient period is given by:

$$\sum_{n=1}^B (1/n) (I_H \Delta)^n (1 - I_H \Delta)^{r-n} \left[\sum_{i=1}^{\min(S_D, n)} \sum_{l=i}^{r-S_D+i-1} \binom{l-1}{i-1} \binom{r-l}{n-i} \sum_{j=0}^{S-i} \binom{r-l}{j} (I_D \Delta)^j (1 - I_D \Delta)^{r-l-j} + \sum_{l=r-S+i}^{r-n+i} \binom{l-1}{i-1} \binom{r-l}{n-i} \right]$$

Proof:

Subdivide the first L units of time from the low to high transient period as well as the last L time units of the preceding low period into time-intervals of size Δ (chosen so that the probability of more than 1 arrival in a time-interval of length Δ is suitably small). Let $r = L/\Delta$. Now condition on the number of demands $N(L) = n$ during the first L units of time of the low to high transient period as before. We again focus on the arrivals during the first L units of time of the low to high transient period. We select a value of B so that the $\Pr(N(L) > B)$ is suitably small.

Observe that

$${}^6\Pr(i^{\text{th}} \text{ demand occurs in the } l^{\text{th}} \text{ time interval} \mid N(L)) = \binom{l-1}{i-1} \binom{r-l}{n-i} / \binom{r}{n}$$

Based on this observation we first determine

$$\Pr(i^{\text{th}} \text{ demand not delayed} \mid i^{\text{th}} \text{ demand in the } l^{\text{th}} \text{ time interval, } N(L) = n) =$$

⁵ On-hand inventory at $t = S_D$ -demand $[t-L, t)$

Pr(No more than $S_D - i$ demands in the last $r - l$ time intervals of the preceding low period |

i^{th} demand in the l^{th} time interval, $N(L) = n) =$

$$\left\{ \begin{array}{l} \sum_{j=0}^{S_D-i} \binom{r-l}{j} (I_D \Delta)^j (1 - I_D \Delta)^{(r-l-j)}, \quad i \leq S_D, l < r - S_D + i \\ 1, \quad i \leq S_D, r - S_D + i \leq l \leq r - n + i \\ 0, \quad i \geq S_D \end{array} \right\}$$

Consider first our partition of the relevant values of l . If $l \geq r - S_D + i$, then there can be at most $S_D - i$ time-intervals in the preceding low period that are relevant (if the l^{th} arrival occurs in the $(r - S_D + i)^{\text{th}}$ time-interval then there will be a residual $r - (r - S_D + i)$ or $S_D - i$ time intervals in the preceding low period that are relevant), and thus at most $S_D - i$ arrivals in the preceding r time-intervals. This implies that if $l \geq r - S_D + i$ then the i^{th} demand will not be delayed with probability 1. Now if $l \leq r - S_D + i$; this condition ensures that there are a sufficient number of time-intervals to “fit” the n arrivals. Finally, $l \geq i$ since the i^{th} arrival cannot be any time-interval numbered lower than i . Now we uncondition on the position of the i^{th} demand to obtain :

$$\sum_{i=1}^{\min(S_D, n)} \sum_{l=i}^{r-S_D+i-1} \frac{\binom{l-1}{i-1} \binom{r-l}{n-i}}{\binom{r}{n}} \sum_{j=0}^{S_D-i} \binom{r-l}{j} (I_D \Delta)^j (1 - I_D \Delta)^{r-l-j} + \sum_{l=r-S_D+i}^{r-n+i} \frac{\binom{l-1}{i-1} \binom{r-l}{n-i}}{\binom{r}{n}}$$

Finally we uncondition on the number of arrivals $N(L) = n$ to get

⁶ This expression is the discrete time analog of the conditional density function for the demand epoch for the i^{th} demand obtained by conditioning on the number of arrivals.

$$\sum_{n=1}^B (1/n) \binom{r}{n} (\mathbf{I}_H \Delta)^n (1 - \mathbf{I}_H \Delta)^{r-n} \left[\sum_{i=1}^{\min(S_D, n)} \sum_{l=i}^{r-S_D+i-1} \frac{\binom{l-1}{i-1} \binom{r-l}{n-i}}{\binom{r}{n}} \sum_{j=0}^{S_D-i} \binom{r-l}{j} (\mathbf{I}_D \Delta)^j (1 - \mathbf{I}_D \Delta)^{r-l-j} \right] \\ + \sum_{l=r-S_D+i}^{r-n+i} \frac{\binom{l-1}{i-1} \binom{r-l}{n-i}}{\binom{r}{n}} \Bigg], \text{ which simplifies to the desired expression}$$

This computation can be done somewhat efficiently in Maple. Note that if we compute the products of the various binomial coefficients and store them in an array this computation could be done only once for a whole set of test problems.

We used simulation to assess the quality of this approximation. For each instance of the problem the simulation required the maximum permissible number of replications to obtain a standard error strictly less than .01 and our approximation agreed to two significant digits in every case. Furthermore each simulation took roughly 30 minutes, while for moderate sized problems the discrete approximation takes no more than 1 minute.

Claim 4: A closed form approximate expression for the average safety factor over the first L units of time of the low to high transient period is given by:

$$\left(\frac{2}{3}\right) \frac{-(LI_H)^{3/2} + (LI_D)^{3/2} + 3S_D(LI_H)^{1/2} - 3S_D(LI_D)^{1/2}}{(\mathbf{I}_H - \mathbf{I}_D)L}$$

Proof:

Define $\mathbf{I}_D L + z\sqrt{\mathbf{I}_D L}$, $\mathbf{I}(t) = \mathbf{I}_D(L-t) + \mathbf{I}_H t$

Note that the demand over $(t-L, t)$ has a Poisson distribution with mean and variance given by $\lambda(t)$.

Then,

$z(t) = (S_D - I(t))/\sqrt{I(t)}$, can be interpreted as the instantaneous protection level at time t .

Thus the average protection level over the interval can be obtained as follows

$$\bar{z}(t) = 1/L \int_0^L z(t) dt = \left(\frac{2}{3} \right) \frac{-(LI_H)^{3/2} + (LI_D)^{3/2} + 3S_D(LI_H)^{1/2} - 3S_D(LI_D)^{1/2}}{(I_H - I_D)L}$$

Evaluating the standard normal distribution at $\bar{z}(t)$ yields the desired approximation.

We tested this approximation across a set of 30 test problems. We assume that

$S_D = \lambda_D L + z(\lambda_D L)^{1/2}$ for the test problems. We selected four values of z namely .9773, .95,

.90, .85, fixed $\lambda_H = 1.2$, $\lambda_D = .7$, and varied S_D over the set of values given by {6,8,10,..20},

the value of L and hence was chosen so that target value of z was met (L chosen so that for a

value of S_D from {6,8,10,..20}, the target fill rate, e.g., either .9773 or .95 or .9 or .85 was

met. This is just to ensure that the values of S_D fall in the set {6,8,10,..20}). Across this set of

test cases our closed form approximation results in an average absolute error of 1.5%. The

reason for selecting this approximation rather than one that determines the average fill rates

directly was that it led to a closed form approximation.

Claim 5: The expected on hand FGI inventory and the expected number of backorders during the first L units of time of the low to high transient period is given by

$$\begin{aligned} E[OH] &= 1/L \sum_{j=0}^{S_D-1} (S_D - j) \int_0^L \left(\frac{(I_D(L-t) + I_H t)^j}{j!} \exp(-(I_D(L-t) + I_H t)) dt \right) \\ &= 1/L \sum_{j=0}^{S_D-1} \frac{(S_D - j)}{j!(I_H - I_D)} \left[e^{-I_D L} \sum_{i=0}^j \frac{(I_D L)^i}{i!} - e^{-I_H L} \sum_{i=0}^j \frac{(I_H L)^i}{i!} \right] \end{aligned}$$

$$E[BO]=1/L \sum_{j=0}^{S_D-1} \frac{(S_D - j)}{(I_H - I_D)} \left[e^{-I_D L} \sum_{i=0}^j \frac{(I_D L)^i}{i!} - e^{-I_H L} \sum_{i=0}^j \frac{(I_H L)^i}{i!} \right] - S_D + L \left(\frac{I_D + I_H}{2} \right)$$

Proof:

Let time $t \in [0, L]$; here 0 denotes the start of the low to high transient period.

Now the expected on-hand inventory at time t is given by

$$\sum_{j=1}^{S_D-1} (S_D - j) \frac{(I_D(L-t) + I_H t)^j}{j!} \exp(-(I_D(L-t) + I_H t))$$

This follows from standard arguments, i.e., there are k units on-hand if there were $S_D - k$ demands over the immediately preceding period of length L . We now integrate over the relevant range of t , i.e., $[0, L]$ and divide by the length of the time period L to obtain

$$1/L \sum_{j=1}^{S_D-1} (S_D - j) \int_0^L \left(\frac{(I_D(L-t) + I_H t)^j}{j!} \exp(-(I_D(L-t) + I_H t)) dt \right)$$

Now

$$\int_0^L \left(\frac{(I_D(L-t) + I_H t)^j}{j!} \exp(-(I_D(L-t) + I_H t)) dt \right) = \frac{1}{j!(I_H - I_D)} \left[e^{-I_D L} \sum_{i=0}^j \frac{(I_D L)^i}{i!} - e^{-I_H L} \sum_{i=0}^j \frac{(I_H L)^i}{i!} \right]$$

This follows by repeatedly applying integration by parts to the expression on the left-hand side. Thus we obtain the desired expression in the statement of the claim. The expected backorders expression can be obtained in a similar manner as follows:

$$E[BO] = 1/L \sum_{j=S_D+1}^{\infty} (j - S_D) \int_0^L \left(\frac{(I_D(L-t) + I_H t)^j}{j!} \exp(-(I_D(L-t) + I_H t)) dt \right)$$

$$= E[OH] - E[NI]$$

$$\text{Where } E[NI] = \text{"Expected Net Inventory"} = \frac{1}{L} \int_0^L S_D - I_D(t) dt$$

Thus,

$$E[BO] = 1/L \sum_{j=0}^{S_D-1} \frac{(S_D - j)}{(I_H - I_D)} \left[e^{-I_D L} \sum_{i=0}^j \frac{(I_D L)^i}{i!} - e^{-I_H L} \sum_{i=0}^j \frac{(I_H L)^i}{i!} \right] - S_D + L \left(\frac{I_D + I_H}{2} \right)$$

These expressions are not time consuming to compute, nevertheless, they are perhaps too complex to be used in an optimization problem.

If a simpler expression is required for the expected on-hand inventory we can use the standard approximation to expected on-hand inventory levels that treats backorders as negative inventory.

Claim 6: The expected on-hand FGI during the low to high transient period obtained by treating backorders as negative inventory is given by:

$$E[OH_{FGI, low \rightarrow high}] = \frac{1}{m} \left[\frac{L^2 (I_H - I_D)}{2} + L(S_D - I_H L) + (m - L)(S_H - I_H L) \right], \quad S_D \geq I_H L$$

A refinement for the case where the expected on-hand can be less than zero is given

$$E[OH_{FGI, low \rightarrow high}] = \frac{1}{m} \left[\frac{S_D}{2I_H} (S_D - I_D L) + (m - L)(S_H - I_H L) \right],$$

The expected on - hand inventory can go below zero if $S_D \leq I_H L$

Proof: The following two expected on-hand inventory diagrams represent each of the above two cases. The expressions can be obtained by determining the areas under the respective functions.

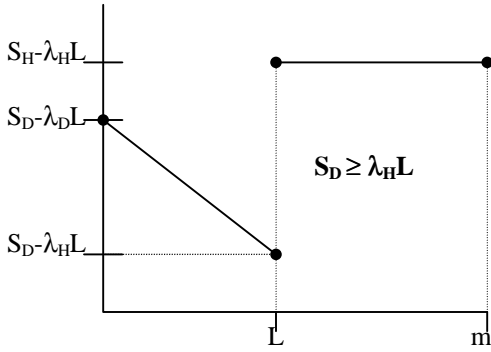


Figure 6: Low to high $E[\text{FGI}]$, $S_D \geq \lambda_H L$

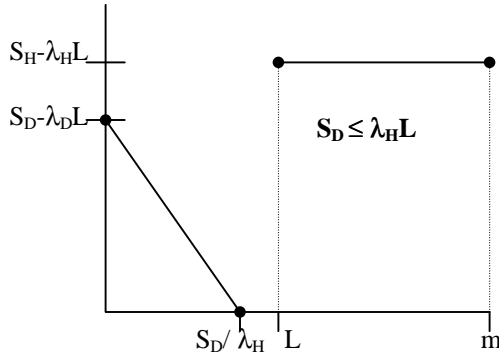


Figure 7: Low to high $E[\text{FGI}]$, $S_D \leq \lambda_H L$

Claim 7: The expected on-hand FGI during the high to low transient period obtained by treating backorders as negative inventory is given by:

$$E[\text{OH}_{\text{FGI}, \text{high} \rightarrow \text{low}}] = \frac{1}{m} \left[\frac{L^2}{2} (\mathbf{I}_H - \mathbf{I}_D) + L(S_H - \mathbf{I}_H L) + \frac{(S_H - S_D)^2}{2\mathbf{I}_D} + (m - L)(S_D - \mathbf{I}_D L) \right]$$

Proof: As with the previous claim the proof follows from the following diagram

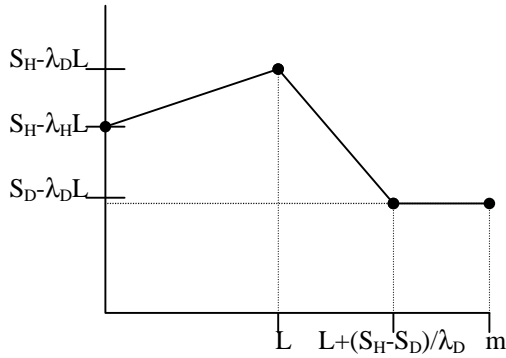


Figure 8: High to low $E[\text{FGI}]$

Claim 8: The expected on-hand FGI during a high period, and a low period obtained by treating backorders as negative inventory are given by:

$$E[OH_{FGI,high}] = (S_H - I_H L), E[OH_{FGI,low}] = (S_D - I_D L)$$

Proof: During a high (low) period the expected on-hand inventory (treating backorders as negative inventory) is constant and has a value of $S_H - \lambda_H L$ ($S_D - \lambda_D L$)

Claim 9: The expected Pipe 2 inventory (intermediate decoupling inventory to FGI pipeline) during the high to low transient period obtained by treating backorders as negative inventory is given by:

$$E[OH_{Pipe2,high \rightarrow low}] = \left\{ \frac{1}{m} \left[\begin{aligned} & \left(\frac{S_H - S_D}{I_D} \right)^2 \frac{I_H}{2} + I_H \left(\frac{S_H - S_D}{I_D} \right) \left(L - \left(\frac{S_H - S_D}{I_D} \right) \right) \\ & + \frac{1}{2} \left(L - \left(\frac{S_H - S_D}{I_D} \right) \right)^2 (I_H - I_D) + I_D \left(L - \left(\frac{S_H - S_D}{I_D} \right) \right)^2 \\ & + \left(\frac{S_H - S_D}{I_D} \right)^2 \frac{I_D}{2} + I_D \left(\frac{S_H - S_D}{I_D} \right) \left(L - \left(\frac{S_H - S_D}{I_D} \right) \right) \\ & + \left(m - L - \left(\frac{S_H - S_D}{I_D} \right) \right) I_D L \end{aligned} \right\} \\ \text{When } \left(\frac{S_H - S_D}{I_D} \right) \leq L$$

A refined version explicitly accounting for negative expected on-hand inventory is given by:

$$E[OH_{Pipe2,high \rightarrow low}] = \frac{1}{m} \left[\frac{(I_H + I_D)L^2}{2} + \left(m - L - \left(\frac{S_H - S_D}{I_D} \right) \right) I_D L \right], \text{ When } \left(\frac{S_H - S_D}{I_D} \right) \geq L$$

Proof: The following two diagrams provide representation of these two situations.

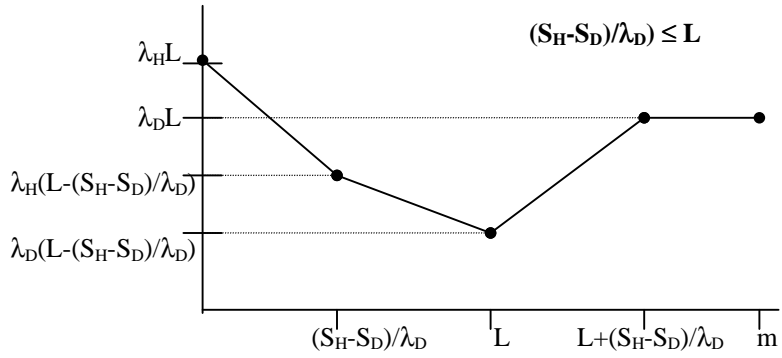


Figure 9: High to low E[Pipe 2], $(S_H - S_D) / \lambda_D \leq L$

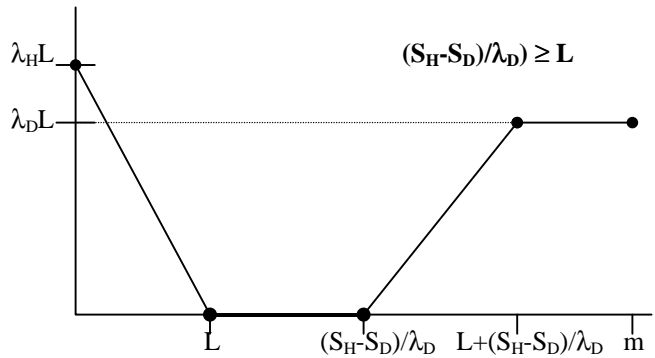


Figure 10: High to low E[Pipe 2], $(S_H - S_D) / \lambda_D \geq L$

Observe Figure 9, the first linear segment corresponds to the period of time where replenishment orders are not placed to bring the base stock to the appropriate low-period level. The next linear segment corresponds to the period of time where replenishment orders are placed in a 1-for-1 fashion. The third linear segment corresponds to the period of time where both 1-for-1 replenishment orders are placed and where the orders placed during the previous segment begin to fill-in the pipeline. The fourth segment corresponds to the period of time after the pipeline has been filled so that the expected pipeline inventory returns to $\lambda_D L$ (here L is the replenishment lead-time)

Claim 10: The expected on-hand Pipe 2 inventory during the low to high transient period obtained by treating backorders as negative inventory is given by:

$$E[OH_{Pipe\ 2, low \rightarrow high}] = \frac{1}{m} \left[\frac{(I_H - I_D)L^2}{2} + L(S_H - S_D + I_D L) + (m - L)I_H L \right]$$

Proof: The following diagram provides a representation of this situation:

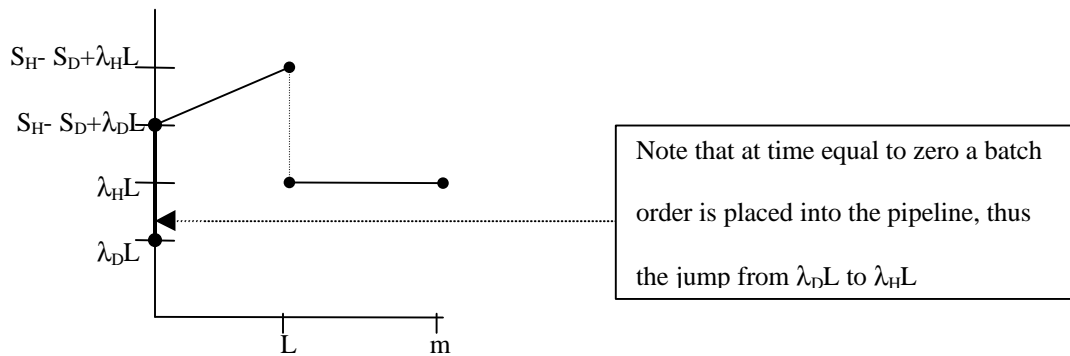


Figure 11: Low to high $E[Pipe\ 2]$

Claim 11: The expected on-hand Intermediate decoupling inventory over a cycle, obtained by treating backorders as negative inventory is given by:

$$E[OH_{Inter,Cycle}] = \frac{1}{m^*(2 + (1-p)/p + (1-q)/q)} \left[(m-L)^2(I_H - I_D) + \frac{m}{p}(S_I - I_H(m-L)) + \left(\frac{m(1-q)}{q} + L \right) (S_I - I_D(m-L)) + (m-L)(S_I - S_H + S_D + I_H(m-L)) \right]$$

Proof: As before we construct an expected on-hand inventory diagram and determine the areas under the respective functions. The following two diagrams represent the two transient periods (low to high and high to low)

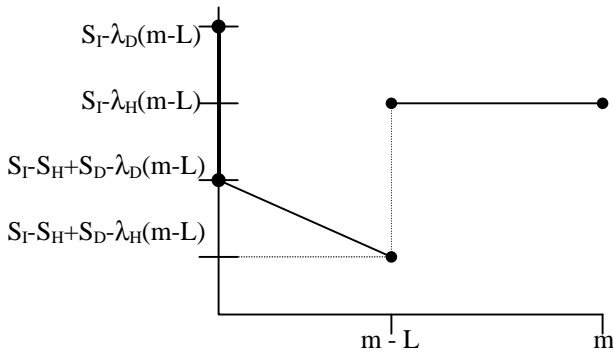


Figure 12: E[Inter], low to high

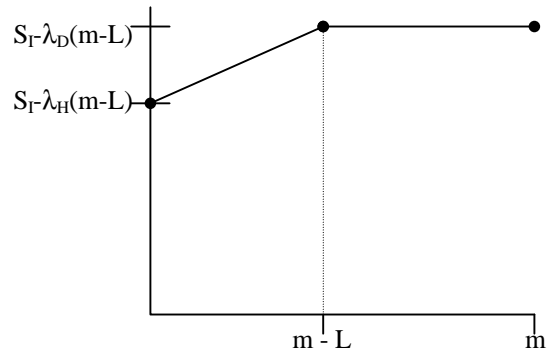


Figure 13: E[Inter], high to low

During a high period and a low period the expected inventory levels will be equal to $S_I \cdot \lambda_H(m-L)$, $S_I \cdot \lambda_D(m-L)$ respectively, where $(m-L)$ is the replenishment lead-time. By putting together the expression for the expected on-hand inventory values and after some simplification we obtain the expression in the claim.

Claim 12: The expected Pipe 1 (supplier to intermediate-decoupling inventory pipeline) inventory over a cycle, obtained by treating backorders as negative inventory is given by:

$$E[OH_{Pipe1,Cycle}] = \frac{1}{m*(2 + (1-p)/p + (1-q)/q)} \left[(m-L)^2(I_H - I_D) + \frac{m}{q}I_D(m-L) + \left(\frac{m(1-p)}{p} + L \right) I_H(m-L) + (m-L)(S_H - S_D + I_D(m-L)) \right]$$

Proof: As before we construct an expected on-hand inventory diagram and determine the areas under the respective functions. The following two diagrams represent the two transient periods (low to high and high to low)

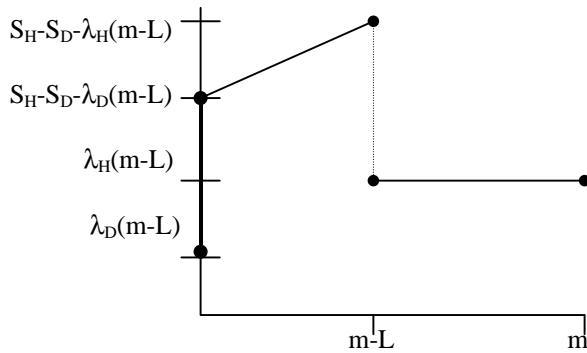


Figure 14: E[Pipe 1], low to high

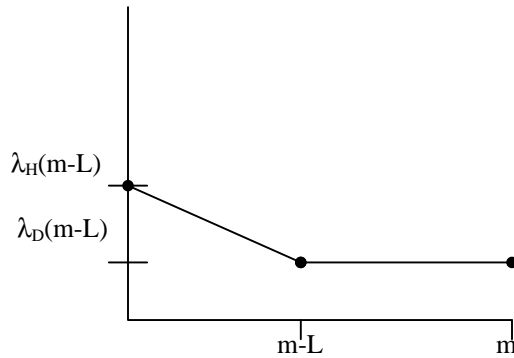


Figure 15: E[Pipe 1], high to low

The fill rates for the FGI during the high and low periods are constant and depend only upon the values of S_H, S_D respectively.

7 An Optimization Problem

7.1 Objective function

The objective of our problem is to minimize the total expected inventory costs per unit time. Due to the nature of the system we have to define four relevant sub-problems by partitioning the feasible region of the problem. In claim 6 we see that the expected FGI depends on whether $S_D \leq \lambda_H L$ or $S_D \geq \lambda_H L$. The other such condition effects the Pipe 2 inventory; the conditions are given by $(S_H - S_D) / \lambda_D \leq L$ or $(S_H - S_D) / \lambda_D \geq L$. As is evident in the figures for claim 6, these conditions permit us to accurately account for the cases where the expected on-hand inventory becomes negative. To this end for each case (set of parameter values) we solve four problems and then choose the solution that minimizes the objective value.

7.1.1 Holding costs

Define $f(L)$ to be the holding cost rate per unit per unit time for a unit of inventory located a time period of length L out in time. For a corresponding real system this corresponds to the holding cost of all material with lead-times exceeding L units of time. For convenience we assume that $L \geq l$ (this lower bound l could correspond for instance to the internal assembly lead-time). Then the following expressions define the relevant holding costs for each location L in time:

$$C_{\text{FGI}} = f(l)$$

$$C_{\text{Inter}} = f(L)$$

$$C_{\text{Pipe1}} = \frac{1}{m-L} \int_L^m f(x) dx$$

$$C_{\text{Pipe2}} = \frac{1}{L-l} \int_l^L f(x) dx$$

Here C_{FGI} , C_{Inter} , $C_{\text{Pipe 1}}$, $C_{\text{Pipe 2}}$ represent the holding cost rates for the finished goods inventory, the intermediate inventory, the supplier to intermediate pipeline inventory, and the intermediate inventory to the finished goods inventory respectively. Note that for each of the pipeline inventories we use the average holding cost rate.

7.1.2 Objective function

Per our discussion we define four separate sub-problems. The objective function for all of the problems is the minimization of the expected inventory holding cost per unit time (expected holding cost over a cycle divided by the expected cycle length) across all four relevant locations. The objective functions for sub-problem 1 as well as the definitions for the other sub-problems are in appendix 1.

7.2 Constraints

We assume that there is a constraint on the low period fill rate, the high period fill rate, and the average fill rate for the high period containing the low to high transient sub-period. Finally there are upper and lower bound constraints on the value of L . These constraints can be written as follows

$$\frac{(S_D - I_D L)}{\sqrt{I_D L}} \geq t_D$$

$$\frac{(S_H - I_H L)}{\sqrt{I_H L}} \geq t_H$$

$$\frac{(m-L)}{m} \frac{(S_H - I_H L)}{\sqrt{I_H L}} + \frac{L}{m} \left(\frac{2}{3} \right) \frac{-(LI_H)^{3/2} + (LI_D)^{3/2} + 3S_D (LI_H)^{1/2} - 3S_D (LI_D)^{1/2}}{(I_H - I_D)L} \geq t_{D \rightarrow H}$$

$$l \leq L \leq m$$

$$S_H, S_D \geq 0$$

Note that the third constraint corresponds to the constraint on the average fill rate obtained during the low to high transient period. The first term corresponds to the sub-period of length (m-L) starting with the instant where the base stock level has been brought up-to the appropriate high period level. The second term corresponds to the expression developed in claim 4 which allows us to approximately specify the fill rate during the first L time units of the low to high transient period. Since we consider a continuous review system with Poisson arrivals and 1-for-1 replenishment by PASTA the τ value implies an appropriate stock-out probability (this is a normal approximation to the demand over the past L units of time)

We define τ_H, τ_D as the fill rate targets for the high and low periods respectively.

8 Numerical Experiments And Discussion

8.1 Overview

We solved a total of 90 instances of the problem (360 sub-problems) by varying the following key parameters

p : .1, .3, .5, .7, .9

q : .1, .3, .5, .7, .9

λ_H/λ_D : 1.2, 1.4, 1.6, 1.8, 2.0, and 2.2

$\tau_H = \tau_D$: 1.6, 1.8, 2.0, 2.4 & $\tau_{D \rightarrow H} = 1.4$

$f(L)$: $1000L^{-5}$, $1000L^{-1}$, $1000L^{-2}$, $1000L^{-3}$ (different holding cost functions)

By doing these experiments we were able to better understand the role of decoupling inventories and finished goods inventories in non-stationary demand environments. For the problems that we solved simple rounding of the non-integral solutions lead to an average increase of .1% in the total cost value over the lower bound. This suggests that the integer solutions thus obtained are very close to the local optimal solutions to these problems. By doing a careful study of the formulation we have concluded that the problem is not a convex program. We tried several different starting points for our test cases and noted that the same solution was obtained. This suggests that the solutions found are possibly the global optimal solutions to these problems.

8.1.1 Motivation for selecting the parameter values

The range of cost functions covers most of the lead-time cost-accrual profiles that we observed in practice. The relative demand rates also correspond to those that we observed in practice. The fill rate targets cover a range of fill rates between 94.5% and 99.1% which

again fall between the values that are typically set in practice. As we had not collected any data on realistic values for p , and q we selected a range of values that would cover the set of all possible values in a sufficient manner.

8.2 Results and discussion

In this section we present graphical representations of our results and discuss the intuition gained from each of the experiments

8.2.1 Tradeoff between TC^* per unit and L^* versus p for various values of q

We varied the values of p and q and solved the problems. Figures 16 and 17 depict the sensitivity of the objective function value and the optimal value of L and TC^* per unit (TC^* per unit = $TC^*/E[\text{demand over a cycle}]$) to p for various values of q . The other values for these problem instances were: $\lambda_H = 12$, $\lambda_D = 7.5$, $m = 30$, $\tau_H = \tau_D = 1.6$, $\tau_{D \rightarrow H} = 1.4$

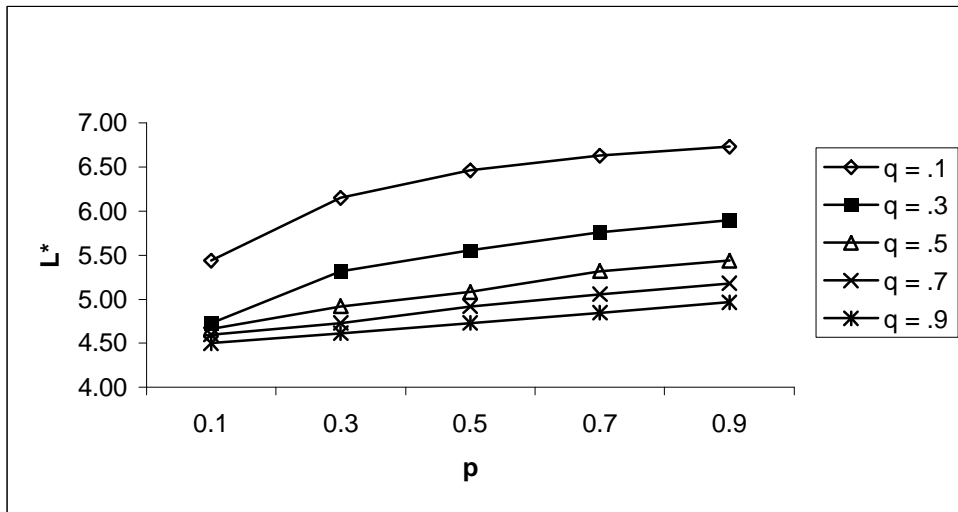


Figure 16: L^* Versus p for various values of q

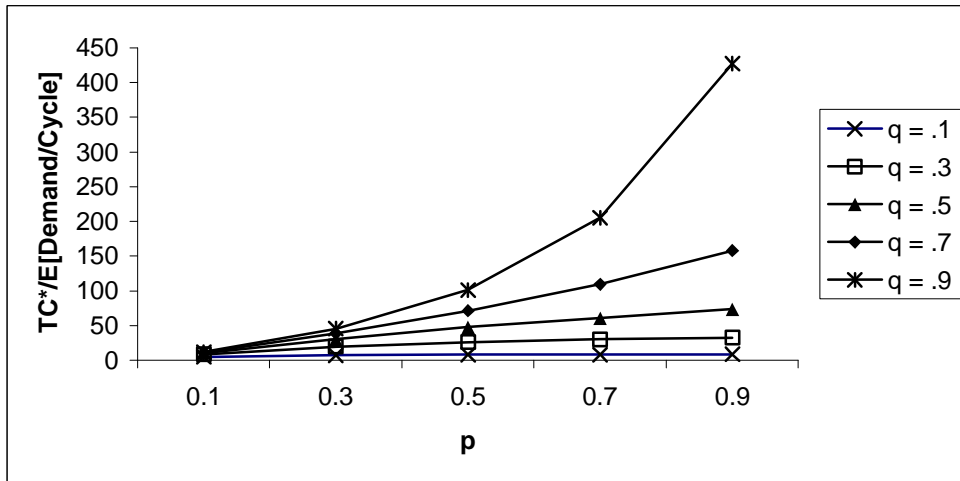


Figure 17: TC^* per unit Versus p for various values of q

Discussion

Increasing p results in a higher L^* for all values of q and increasing q results in a lower L^* for all values of p . As p increases, for any fixed value of q the expected on-hand intermediate decoupling inventory during a cycle increases as there are fewer high periods relative to a fixed expected number of low periods during a cycle. Now the intermediate decoupling inventory is maintained using a state independent base stock policy that would result in higher on-hand inventory levels at all times. Thus to counter this a higher value of L^* would be better. Conversely as q increases, for any fixed value of p the expected on-hand intermediate decoupling inventory during a cycle decreases as there are fewer low periods relative to high periods and the expected on-hand inventory at the intermediate decoupling inventory will be relatively lower. Thus a lower value of L^* will be better. The relationship between p (or q) and TC^* per unit appears to be an increasing and convex in p . As q increases, the rate of increase in costs increases, with the cost increases being nearly linear in p (for small

values of q), and non-linear in for $q = .9$. The number of low periods is strictly convex and decreasing in q as the graph below shows

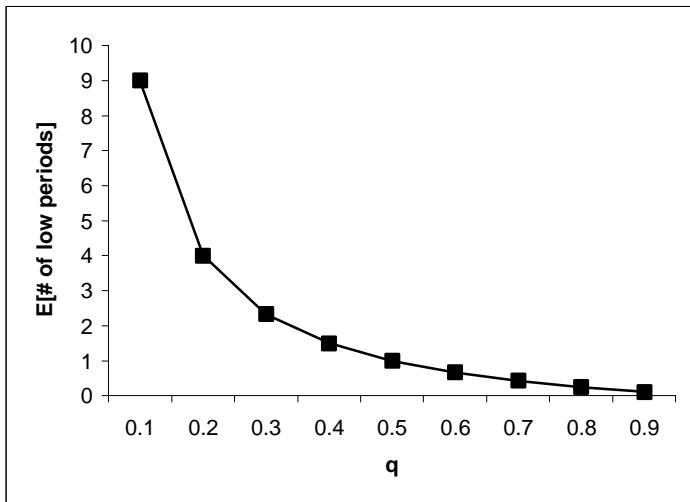


Figure 18: The expected number of low periods as a function of q

Thus for low values of q there will be considerably more low periods than for considerably higher values of q . So low values of q will result in considerably more low periods than high periods resulting in higher values of L^* and lower costs.

8.2.2 Sensitivity of TC^* per unit and L^* versus various λ_H/λ_L

For this set of experiments we wanted to assess the impact of more drastic business cycles as captured through the ratio of high to low demand rates on the optimal costs as well as the optimal location of the intermediate inventory. Other parameter values for these experiments were: $m = 30$, $\tau_H = \tau_D = 1.6$, $\tau_{D \rightarrow H} = 1.4$

Figures 19 and 20 provide a graphical representation of these experiments. The y axis in these graphs is the value of TC^* per unit.

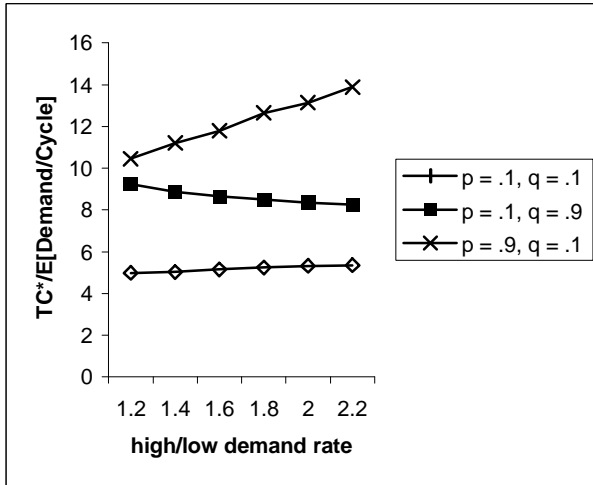


Figure: 19a

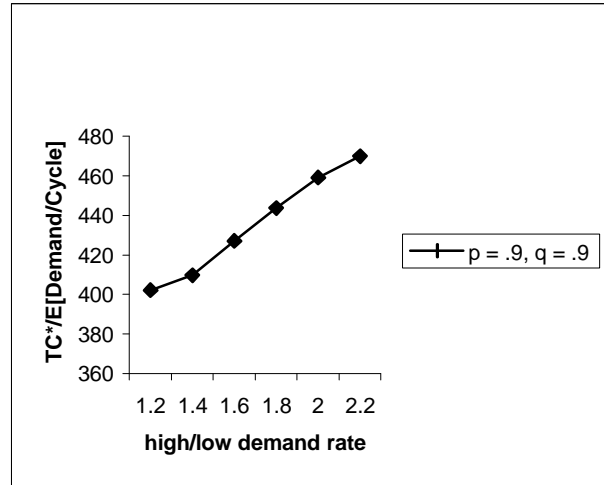


Figure 19b

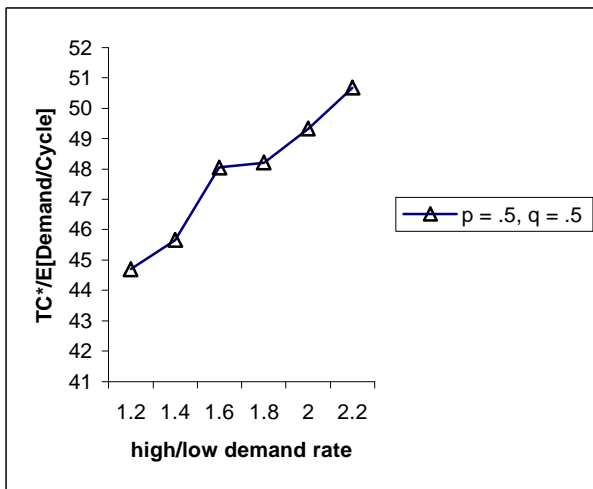


Figure 19c

Figures 19a,b,c: TC* per unit Versus high/low demand rates

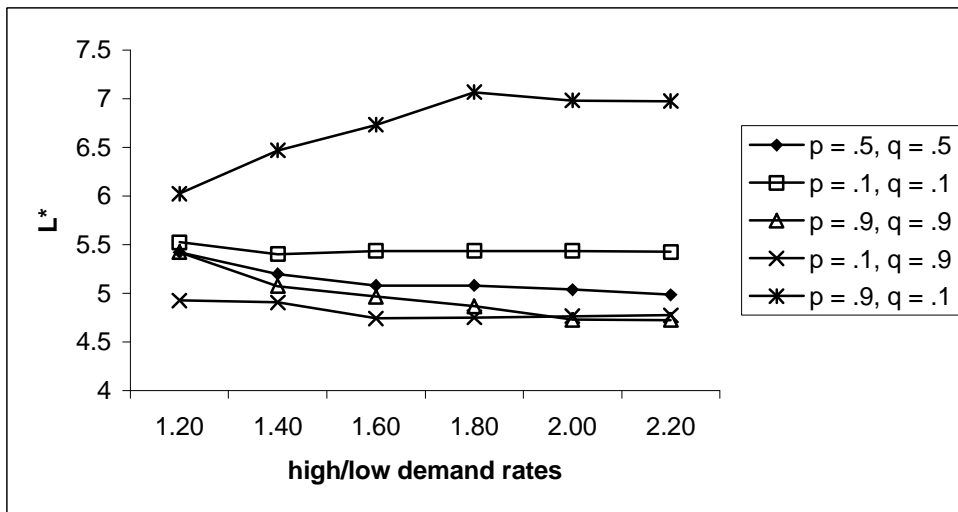


Figure 20: L^* Versus high/low demand rates

Discussion

For all cases the optimal costs are increasing in the relative rates and the costs increase linearly. Based on figure 20 as one moves from the $p = .9, q = .1$ case to the $p = .1, q = .9$ case the sensitivity of L^* relative to the high to low ratio decreases with a near constant value for L^* in the $p = .1, q = .9$ case.

8.2.3 Sensitivity of TC^* per unit and L^* for different cost functions

In this set of experiments we attempt to model lead-time reduction efforts through the use of different cost functions. For all of the cost functions the value of the finished goods inventory is fixed. We vary the shapes of the cost functions to represent cases where the negotiations with the suppliers of expensive components leads to lead-time reductions of these components thus pushing the bulk of the costs closer to zero.

For each of these cost functions we solved the problems for five (p,q) pairs. The results are represented in figures 21 and 22 below (the other parameter values were $\lambda_H = 12$, $\lambda_D = 7.5$, $m = 30$, $\tau_H = \tau_D = 1.6$, $\tau_{D \rightarrow H} = 1.4$)

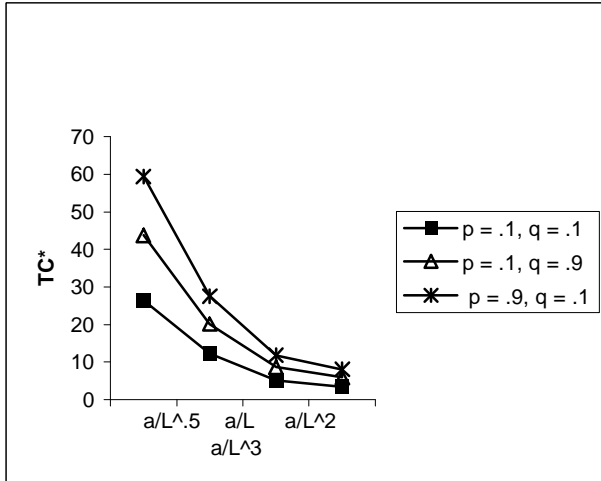


Figure 21a

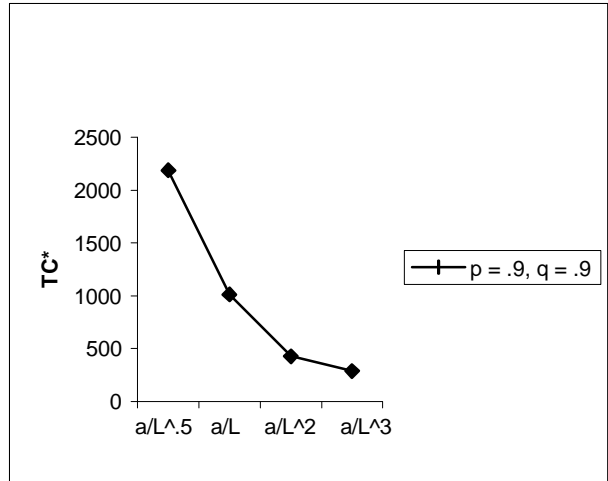


Figure 21b

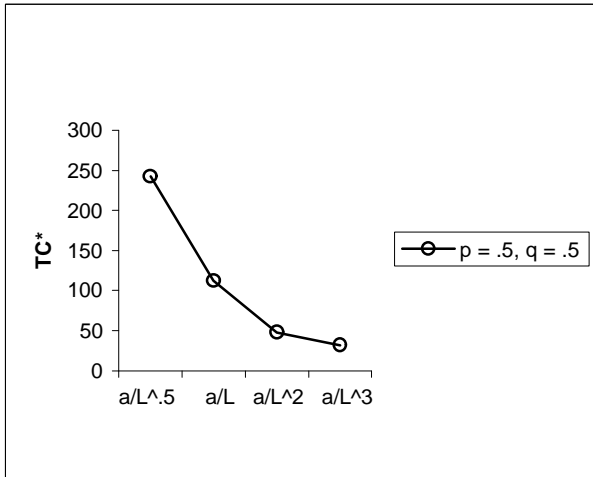


Figure 21c

Figures 21a,b,c: TC* per unit for different cost functions and values of (p,q)

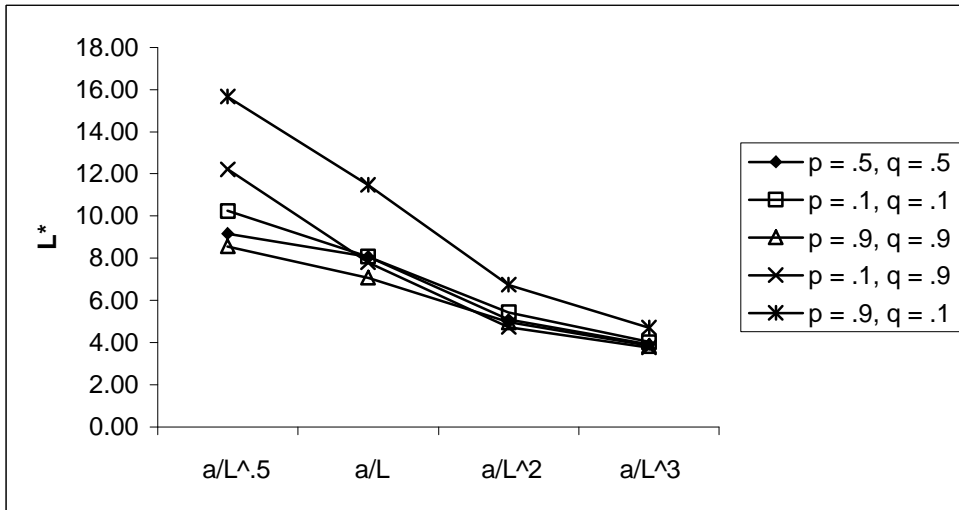


Figure 22: L^* for different cost functions and values of (p,q)

Discussion

Based on figure 21 it is fairly clear that a lead-time reduction effort as discussed in the previous paragraph can have a significant impact on costs. Somewhat surprisingly the percent cost reduction obtained in going from one to the next consecutive cost function seems to be nearly independent of the values of (p, q) suggesting that this strategy would be beneficial under all circumstances. Furthermore the cost reductions are significant ranging anywhere from about 33% to 80% depending upon the cost profile that fits the situation under consideration. From figure 22 we notice that value of L^* depends both on the values of (p,q) and the particular cost function that fits the situation with the ranges in the L^* values being the greatest for $a/L^{.5}$ and the least for a/L^3 . This discussion suggests that using an accurate representation of costs is critical in determining suitable values for L .

8.2.4 TC* versus various values of L

In this series of experiments we attempt to better understand the role of a decoupling inventory by fixing the location of the inventory and solving for the optimal values of S_H , and S_D . Figure 23 represents this situation for $p = .5$, $q = .5$ (the general shape of the curves is the same for other values of p and q ; for this case we assumed that $m = 30$, $\lambda_H = 12$, $\lambda_D = 7.5$, $\tau_H = \tau_D = 1.6$, $\tau_{D \rightarrow H} = 1.4$)

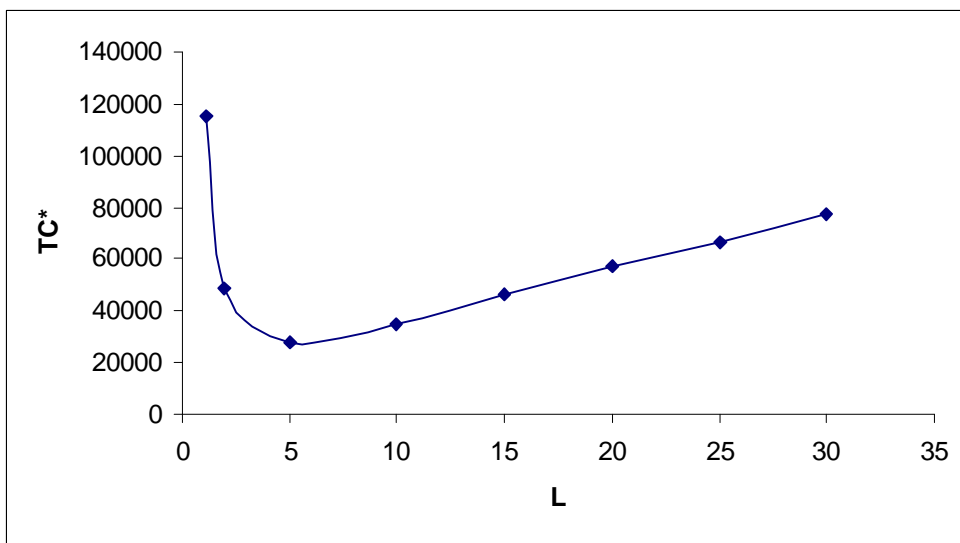


Figure 23: TC* Versus fixed values of L

Discussion

TC* appears to be convex in L with a steep decrease in the initial values followed by a nearly linear increase after the point of inflection. This figure adds further evidence that intermediate decoupling inventories are critical in these types of situations. For a range of relative demand rates (between 1.2 and 2.2) we solved the problems by fixing L at its lower bound and compared this to the optimal objective value (by letting L vary). Across this set of test problems the savings from having an optimally configured intermediate decoupling

inventory was between 71% and 81% (with an average of 78%), suggesting that decoupling inventories play a key role in this sort of a setting.

9 Summary And Opportunities For Further Work

We present a model that characterizes the situation faced by some firms in the high technology capital equipment sector. These firms are subjected to a great deal of demand volatility without visibility and long procurement lead-times for components. For this context we demonstrate the importance of a decoupling inventory that permits the firm to better react to this sort of demand volatility. We present an approximation-based formulation to determine the optimal decoupling inventory location as well as the low and high period base stock levels. In order to do so we make the key assumption that the decoupling inventory is stocked in a manner to handle a “maximum reasonable demand” as in Simpson (1958) and then ignore stockouts due to extraordinary demand. Based on the framework developed in this section the following extensions could be incorporated to handle other situations that may arise in practice

9.1 Deterministic service times

In certain situations customers permit a period of length w to fill an order. An order filled within w units of time from its arrival instant is not delayed. We can incorporate this situation rather easily into our discrete time approximation as follows.

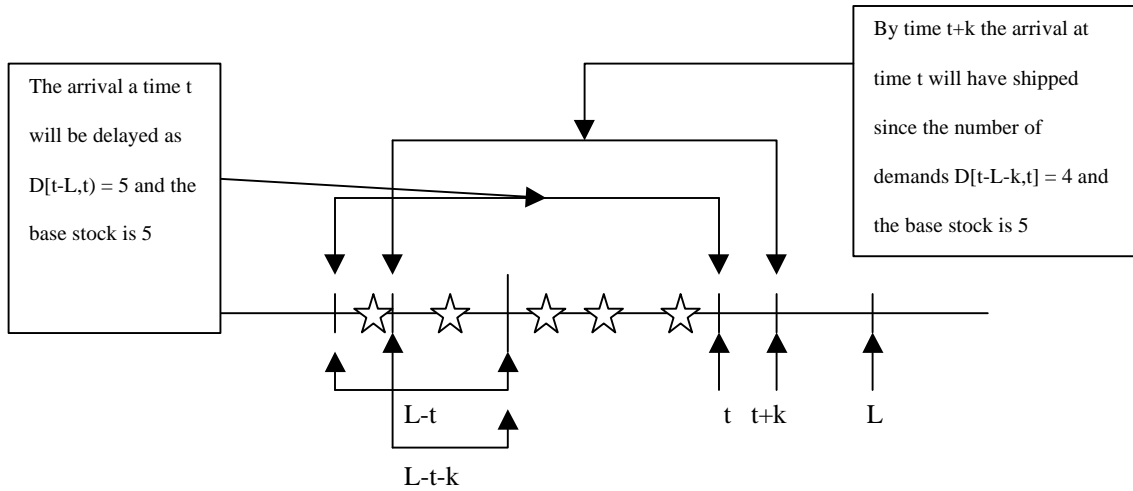
For the sake of discussion assume that an order arrives at time t .

$$\Pr(\text{order filled within } w \text{ time units}) = 1 - \Pr(\text{order not filled within } w \text{ time units})$$

$$\Pr(\text{order not filled within } w \text{ time units}) = \Pr(\text{at least } S_D \text{ units on-order in } (t - r + w, t))$$

Then an equivalent closed form expression for this case can be developed as follows:

Consider a case where $S_D = 5$. In the following figure a star represents a demand



Thus the effective arrival rate observed by the arrival at time t is given by

$$\lambda(t) = \lambda_D(L-t-k) + \lambda_{HT}, \text{ when } t+k \leq L, \text{ and } \lambda(t) = \lambda_H(L-k), \text{ when } t+k \geq L$$

Thus we can again determine the average value of z by integrating over these two regions.

9.2 More general demand patterns

We could create a more complex demand process by making the number of self-transitions, i.e., the high to high period or the low to low transitions depend on the number of periods spent in the previous low or the previous high state respectively. This one state dependence could be incorporated by appropriately changing a few of the terms in the objective functions.

9.3 Deterministic rate change predictability

Assuming a deterministic time advance notice of the rate change of a period of time of length γ is equivalent to locating the intermediate inventory γ units of time closer to the FGI.

9.4 Modeling expediting capability

If we assume that during the low to high transient period the time required to adjust to the high rate is a discrete random variable bounded by the distance between the two inventories, we could condition on this random variable to extend our discrete time approximation to this case.

9.5 Modeling market share loss due to stock outs

Consider a case where potential demand is lost due to poor service. One way to capture this could be to make the effective demand rate during the high period dependent on the low to high transient period fill rate. One issue that would arise is the determination of the penalty cost of losing a portion of the demand and how long such a market share loss may persist. Perhaps this extension is of the greatest practical interest as a market share loss due to stockout is a reality, furthermore anecdotal evidence suggests that overcoming such a market share loss can take a substantially long period of time.

9.6 Permitting stock-outs at the intermediate decoupling inventory

We could extend the model to permit stock-outs at the intermediate decoupling inventory that would delay the batch order that brings the FGI to the suitable base stock level. Two assumptions could be made regarding the mechanics of this system. Under the first assumption if a stockout takes place then there would be no partial release of the batch, i.e., wait until the entire batch becomes available then release from the intermediate-decoupling inventory to the FGI. Under the second assumption if a stockout takes place one would release any available material immediately and wait until the additional requisite material becomes available to make subsequent releases. We could also inflate the value of L by some

fixed quantity to account for any delays that could take place due to stockouts at the intermediate-decoupling inventory. This would be similar to the approximation made by Sherbrooke (1968).

10 Appendix I

10.1.1 Sub-problem 1

For sub-problem 1 we impose the following additional constraints

$$S_D \leq I_H L$$

$$\frac{S_H - S_D}{I_D} \leq L$$

10.1.2 Sub-problem 2

For sub-problem 2 we impose the following additional constraints

$$S_D \geq I_H L$$

$$\frac{S_H - S_D}{I_D} \geq L$$

10.1.3 Sub-problem 3

For sub-problem 2 we impose the following additional constraints

$$S_D \leq I_H L$$

$$\frac{S_H - S_D}{I_D} \geq L$$

10.1.4 Sub-problem 4

For sub-problem 2 we impose the following additional constraints

$$S_D \geq I_H L$$

$$\frac{S_H - S_D}{I_D} \geq L$$

Part III: A computationally efficient procedure
to set base stocks in assemble-to-
order environments

11 Introduction

In this section we consider a situation where end-items are assembled from a set of components in an assemble-to-order fashion. These components are either made-to-stock or procured from outside vendors. We assume that the component inventories are maintained using independent one-for-one replenishment policies. Components can either be unique to a specific end-item or common across several end-items and can also be differentiated on the basis of their costs and their replenishment lead-times, which are assumed to be deterministic. We assume that the arrival processes for the end-items are independent renewal processes. We also assume that there is a constraint on the system-wide safety stock. Under such a setting there is no prevailing methodology to set the base stock levels for the component inventories. In this section we address this issue by formulating an appropriate optimization problem. The solution to this optimization problem provides us with the base stock numbers for the components. Through a simulation study we compare the quality of our solutions to the quality of the solutions obtained from a few alternate heuristic policies. For the sample problems studied, our method outperforms the other heuristics under a wide variety of circumstances confirming the effectiveness of our approach.

This section is organized as follows. In section 12 we develop our heuristic method. In section 13 we present the results from our simulation study. In section 14 we discuss some possible extensions and directions for future work.

12 Model Development

As noted in the introductory section we provide a robust decision support tool that is computationally tractable. Based on our observations at Teradyne we consider a situation where the end-item demand processes are independent renewal processes and the replenishment lead-times for the components are deterministic. The following graph provides a summary of a sample of option level (150 options) weekly demand data over a period of 48 weeks. During this period of time demand was stationary. As the graph indicates, the ratios of the standard deviation of weekly demand to the square root of the mean weekly demand range between .04 and 2.2 (for a Poisson random variable this ratio has to equal 1).

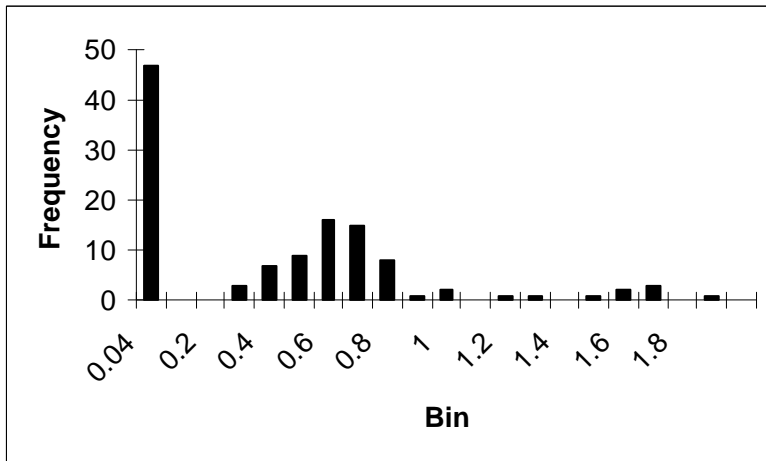


Figure 24: Histogram of the ratio of the standard deviation to the square root of the mean

Teradyne's component replenishment lead-times are fairly predictable; the primary form of uncertainty in their system is due to the unpredictable nature of customer demand. These assumptions make an exact representation of the situation quite difficult, as there is no Markovian aspect to our system. The exact models discussed in section 1.3 assume either simple or compound Poisson processes used to describe the end-item demand processes

and/or assume exponential (or Erlang) replenishment lead-times for the components. We formulate a simple approximation-based approach that is easy to implement.

In this section we provide an exact formulation of the relevant optimization problem. We then point out the areas that are difficult to analyze in an exact fashion. In the next section we discuss the approximations that were used.

Referring back to Teradyne's product structure that we discussed in Part I of the thesis, we reiterate one of the key aspects of their system. End-items are assembled to order from a set of PCBs. It is not possible to assemble the end-items in advance of customer demand as customers order completely customized systems. In this setting our method provides a way to set the PCB base stock levels that attempts to provide good performance from a type II service level standpoint.

To reiterate, we make the following key assumptions

- We consider a system where demands for end-items arrive according to independent renewal processes.
- Each end-item is assembled from a set of components, the inventories of which are controlled using independent 1-for-1 replenishment policies. From Teradyne's perspective this assumption is not unreasonable, as components correspond to PCBs here and there are minimal economies of scale in PCB assembly.
- For a given order, if all of its constituent components are available, then this order is filled immediately. Otherwise the order waits till all of its requisite components become available.
- We assume that orders are processed in a FCFS fashion. If for an end-item order a proper subset of all of its constituent components is available, then we commit

these components to the end-item order. This is equivalent to a case of no cannibalization, i.e., end-item orders cannot “steal” allocated components. This assumption is similar in spirit to the virtual allocation assumption made by Graves (1996).

- We further assume that there is a budget constraint on the system-wide safety stock. Under this scenario we wish to determine the optimal base stock levels for the components that will maximize some service performance criterion.

Here we could consider two alternate performance criteria:

- Maximize the system-wide type II service level (steady state fraction of all end-item demands that are met from stock)
- Determine the stocking policy that results in a Pareto optimal vector of type II service levels across all end-items. Such a stocking policy will have the following property: There is no other stocking policy that provides a higher type II service level for at least one of the end-items without providing a lower type II service level for any of the other end-items.

At Teradyne the policy is to give equal service to orders regardless of the order type, i.e., it is not acceptable to give very good service to an inexpensive order and poor service to a more expensive order or vice versa. At the same time the company is also concerned with its system-wide type II service level. Due to these reasons we focus on these two criteria in order to judge the quality of our solutions.

We define the following parameters and variables that will be used in our formulations

12.1 Notation:

j, J	The index for the end-items, and the number of end-items
k, K	The index for components, and the number of components
$\Omega(j)$	This is an index set to describe the set of components used in end-item j . For example if end-item j uses component k then $k \in \Omega(j)$
$\Phi(k)$	This is a similar index set that describes the set of end-items that use component k .
a_{jk}	The number of units of component k that go into end-item j .
m_k	The base stock level for component k .
τ_k	The replenishment lead-time for component k
λ_j	The external arrival rate of end-item j
λ_0	The total system-wide arrival rate across all end-items
$scv(j), scv(k)$	The squared coefficient of variation for the arrival process of end-item j and the induced arrival process squared coefficient of variation for component k .
c_k	The cost per unit for component k
B	A limit on the expected value of the system-wide safety stock for unallocated components.
h	The holding cost rate.

$EW_j(m_1, m_2, \dots, m_K)$ The expected waiting time for end-item j given that the base stock levels are m_k .

$EW_k(m_k)$ The expected waiting time for component k given a base stock level of m_k

12.2 The related optimization problem:

We are now ready to formulate an optimization problem to determine the base stock levels

$$P: \text{Min} \sum_{j=1}^J \left(\frac{I_j}{I_0} \right) EW_j(m_1, \dots, m_K)$$

Subject to :

$$m_k \geq t_k \sum_{j=1}^J a_{jk} I_j \quad \forall k = 1, \dots, K \quad (3.1)$$

$$h \left(\sum_{k=1}^K c_k (m_k - t_k \sum_{j=1}^J a_{jk} I_j) \right) \leq B \quad (3.2)$$

$$m_k \text{ Integer}$$

12.3 Discussion

We use this formulation to determine the component base stock levels that minimize the time-average waiting time across all end-items for a given value of system-wide safety stock. Constraint (3.1) assures that the base stock is greater than the expected number of units on order for each component; in effect, constraint (3.1) assures that the safety stock for each component is non-negative. Constraint (3.2) puts a bound on this system-wide safety stock. In this formulation we are not explicitly relating the service levels at the component level to the

service levels at the end-item level, which is difficult to do in our setting. Observe constraint (3.2), the left-hand side of the constraint determines the expected on-hand inventory treating backorders as negative inventory. However since the mean demand over lead-time is unaffected by shortages of other components (we commit inventory to an order without taking into account the availability of the other components that make up the order) the left-hand side provides a measure of the excess unallocated inventory. We are also implicitly assuming that a materials manager could provide us with the value of B .

One of the most common problems faced by practitioners in such an environment is one in which the objective is to determine the component inventory levels to meet pre-specified fill rate targets. Often there is at least an implicit budget constraint on the expected inventory levels. For our system the relationship between end-item fill rates and component base stock levels is not well understood. We conjecture that minimizing the time-average system-wide response time will result in good fill rate performance. For an $M/M/m$ system both the probability that an arriving customer observes a full system (unfill rate) and the expected waiting time are increasing in the utilization level. Thus by minimizing the waiting time we minimize the utilization level which in turn should result in good fill rates. In section 13.1, we explain our use of a $G/D/m$ queue as a surrogate to deal with the $G/D/\infty$ queue to represent the system under study.

12.4 The embedded queueing system

In classical inventory theory there are numerous instances in which one can model inventory systems with simple queueing models. A continuous-review base stock policy in an environment with renewal process arrivals and general service times and FCFS service could be represented by a $G/G/\infty$ queue. However for such a representation to be valid we need to

permit order crossing (or we could ignore the fact that it occurs). In our system we assume deterministic replenishment lead-times and so order crossing does not occur. Let us assume that the base stock level is B , then the event of a backorder is equivalent to having more than B customers in the system. We could then infer the expected number of back orders in the system by taking expected value across all states in which the number of customers in the system exceeds B ; then by applying Little's law we could determine the expected response time of the system. This is the approach taken in Sherbrooke [1968] and Graves [1985]. In these papers, the authors assume that the arrival process is Poisson (or compound Poisson), which leads to an $M/G/\infty$ queue. For this queue by Palm's theorem we know the exact form of the steady state distribution of the occupancy of the system. Our system when viewed from the perspective of any given component is a $G/D/\infty$ queue. For this system (unless we assume that $G = M$ or E_r) it is a non-trivial matter to get an analogous closed form expression. We address this difficulty in section 12. Our system is considerably more complicated than the situation described in the last paragraph due to the following reasons: An end-item order has to wait until all of its components become available. Consider a single end-item that uses two components. The arrival of an end-item order is equivalent to the simultaneous arrival to two queueing systems (one for each of the constituent components). So to begin with we are considering $G/D/\infty$ systems with correlated arrivals. To determine the expected waiting time for an end-item we have to determine the expected value of the maximum waiting time for each of the components that the end-item uses. To make matters worse, we are trying to model a system with multiple distinct end-items, so we are in reality superimposing correlated renewal processes. Although, it is theoretically possible to determine the interarrival time

distribution of a superimposed renewal process, the resulting process is not a renewal process unless each of the component processes is Poisson (see Cinlar [1972]).

For these reasons we have to resort to approximations. In the next section we discuss the approximations that we use.

12.5 Approximations

12.5.1 A surrogate for dealing with the $G/D/\infty$ queue

In the previous section we discussed the relationship between simple queues and related inventory systems. Recall that for inventory systems under continuous review using base stock policies, a situation where there are more customers in the system than the base stock level corresponds to a backorder situation. With suitable restrictions we could in theory compute the expected backorders in the system and then compute the expected time required to process the request for a component. However with a general distribution G this is not an easy task. As an approximation we replace a $G/D/\infty$ queue with a $G/D/m$ queue where m is the base stock level. After this modification, the existence of a queue in the modified system is similar to having more than m orders in the system for the $G/D/\infty$ representation. We can relate this modification to our inventory control policy. Under the control policy, replenishment orders are placed the instant that a unit of demand arrives. This would be the case if we used the $G/D/\infty$ representation. However, with the $G/D/m$ representation a replenishment order is placed only after a component becomes available. Since replenishment orders are delayed, they can cause future demands to be delayed and this would lead to an overestimated expected waiting time. This representation is accurate in terms of the waiting times beginning with an instant when a server is free to the next instant when

there are m orders in the system. Then it overestimates the waiting times until the next instant when a server becomes free. For this reason it is clear that the accuracy of the approximation decreases as the utilization increases, since these instants of empty queues are less frequent with higher utilizations than with lower utilizations.

12.5.2 The G/D/m queue with superimposed renewal process arrivals

Given that there is no Markovian element to this queueing system we have to resort to using approximations. From our discussion in the last section we desire a closed form expression for the waiting time in queue for each of the components. Based on some testing⁷ we decided to use the following approximations from Whitt (1993)

$$EW_q(G/D/m) = \left(\frac{c_a^2}{2} \right) EW_q(M/M/m) \quad (3.3)$$

Here c_a^2 is the scv for the arrival process, EW_q is the expected waiting time in queue, and $EW_q(M/M/m)$ is the expected queueing time for an M/M/m queue with the same parameters.

We also had some flexibility in choosing the $EW_q(M/M/m)$ expression. For our optimization problem we desire a closed form (differentiable) expression for this quantity. Although this queue can be analyzed exactly, the resulting expression is not differentiable in m . To this end we used the following closed form approximation developed by Sasasekawa which appears in Whitt (1993)⁸:

⁷ We experimented with a simple heavy traffic approximation that did not perform as well in our simulation experiments.

⁸ Refer to claims 1 and 2 in appendix I for a proof of convexity (treating m as a continuous variable) of the Sasekawa approximation and conditions under which 3.3 is convex respectively.

$$EW_q(M / M / m) = t(r^{\sqrt{2(m+1)-1}})/(m(1-r))$$

where I is the arrival rate, t is the mean service time, and $r = It / m$

12.5.3 Superimposition of renewal processes

Whitt (1983) provides a method for approximating the superimposition of independent renewal processes with a single renewal process. Consider a component k that is used in multiple end-items. We compute the scv for component k 's queue using the following equations (these are equations 10-12 from Whitt (1983)):

$$v_k = \left(\frac{\sum_{i:i \in \Phi(k)} I_i^2}{\left(\sum_{i:i \in \Phi(k)} I_i \right)^2} \right)^{-1} \quad (3.4)$$

$$w_k = [1 + 4 * (1 - r_k)^2 (v_k - 1)]^{-1} \quad (3.5)$$

$$ascv(k) = (1 - w_k) + w_k \left(\sum_{i:i \in \Phi(k)} scv(i) * (I_i / \sum_{l:l \in \Phi(k)} I_l) \right) \quad (3.6)$$

Equations (3.4) and (3.5) serve as intermediate steps in the computation of the scv for a superimposition of several renewal processes. In equation 3.6 $ascv(k)$ denotes this aggregate scv for such a process. Whitt [1983] provides some motivation for these equations. These equations are developed using the asymptotic method in which the scv is a convex combination of the individual scv's weighed by their relative arrival frequencies (the

individual arrival rates divided by the cumulative arrival rate) and the stationary interval method.

12.5.4 The response time for an end-item order

A second area of difficulty arises due to the assembly nature of our system, i.e., an end-item can be assembled only when all of its component parts become available. We begin by ignoring the dependence between the component replenishment processes (the dependence exists since subsets of components are simultaneously required for end-item orders). We let $W_k(m_k)$ be the (non-negative) random variable representing the steady state amount of time that an end-item order j waits for component k at an arrival epoch and let $\Omega(j)$ denote the index set for the components that are required for end-item j . The steady state expected total amount of time that the end-item order j waits has the following form:

$$E[\text{MAX}_{k \in \Omega(j)}(W_k(m_k))]$$

Here the expected waiting time for an end-item is the expectation of the maximum waiting time required to obtain all of its constituent components. For an arbitrary instance of the problem this expression is not easy to evaluate. This is due partly to the fact that there is no easy way to determine $W_k(m_k)$. Moreover as pointed out in the last paragraph, the $W_k(m_k)$ are not independent. An alternate approach may be to assume that they are distributed according to a mathematically convenient distribution. At any rate we propose an alternative based on the following observation.

$$E[\text{MAX}_{k \in \Omega(j)}(W_k(m_k))] \leq \sum_{k \in \Omega(j)} E[W_k(m_k)] \quad (3.7)$$

The above property is trivial to show for non-negative random variables⁹. However, this bound need not be tight in general. It is easy to construct sequences of numbers that result in a very weak bound. We determine the validity of using this bound through the assessment of the quality of our solutions from the simulation study (the details of which are provided in section 13). Clearly, this is a possible place for refinement. However, the intent of this work is to stimulate approximation-based approaches to this problem that are as robust as possible, so we feel that this representation is adequate for the time being.

12.5.5 Cases where multiple copies of components are required

In practical problems there are instances where multiple copies of the same component are required by an end-item order. Consider a particular end-item that uses a_k copies of a component, let $W_k(m_k, a_k)$ denote the random variable representing the steady state amount of time required to get a_k copies of component k at an arrival epoch. For these cases we use the following approximation based on the bound in the preceding section

$$E[\text{MAX}_{k \in \Omega(j)}(W_k(m_k, a_k))] \leq \sum_{k \in \Omega(j)} a_k E[W_k(m_k)]$$

We can motivate this bound as follows, if on average it takes $E[W_k(m_k)]$ time to get a single copy of component k , then on average it will take no more than $a_k E[W_k(m_k)]$ time to get a_k copies of the component. Suppose that a_k copies of component k are required, then we would multiply both the scv and the mean arrival rate by a_k (of the single component case). This method of multiplying the scv by a_k can be found in Whitt (1983).

⁶ Refer to claim 3 in appendix I

12.6 The approximation based formulation:

Upon applying the previously discussed approximations we can restate the formulation as:

$$P': \text{Min} \sum_{j=1}^J \left(\frac{I_j}{I_0} \right) \sum_{k \in \Omega(j)} a_{jk} EW_k(m_k)$$

Subject to :

$$m_k \geq t_k \sum_{j=1}^J a_{jk} I_j \quad \forall k$$

$$h \left(\sum_{k=1}^K c_k (m_k - t_k \sum_{j=1}^J a_{jk} I_j) \right) \leq B, \quad m_k \geq 0$$

$$\text{where } EW_k(m_k) = \left(\frac{a_k \text{scv}(k)}{2} \right) t_k (r_k^{\sqrt{2(m_k+1)}-1}) / (m_k (1 - r_k))$$

12.6.1 Discussion:

In this representation we effectively treat the component queues as independent queues. As discussed earlier, P' is a convex program if the objective function is convex as all constraints are linear. In appendix I we show that the Sasekawa approximation for $EW(M/M/m)$ is convex. The convexity of $EW_k(m_k)$ depends both on the properties of $EW(M/M/m)$ as well as the properties of scv for the component queue. We note that the approximation for the scv of a component queue depends on mp , which is a function of m . In general P' is not a convex program (in appendix I we provide an explicit range over which it is a convex program).

In this formulation we have relaxed the integrality constraints that appear in the exact formulation. With the integrality constraints P' is a mixed integer non-linear program which is difficult to solve. In order to create integral solutions some amount of rounding has to be

done. In our examples there are instances where the budget constraint is violated by the rounded solution. To obtain the rounded solution we use simple rounding. Upon substituting in the values of the rounded solution into the budget constraint, three things can happen: the left-hand side could be less than or equal to or greater than the budget constraint. If the left-hand side is less than the budget constraint, we add to the multiple use components with the highest utilization values until a feasible solution is achieved. If the left-hand side is greater than the budget constraint, we reduce the single use components with the lowest utilization values. In making these adjustments we ensure that we do not move too far away from the unrounded solution in a Euclidean distance sense. There are several alternate-rounding procedures that we could conceive of. We could begin with the solution to the NLP and round down which will result in a left hand side value strictly less than the budget constraint; we could then add to the base stock levels to build up to a feasible solution adding only to multiple use components. Alternately we could round down and then formulate a integer nonlinear program with an objective of minimizing either the regular Euclidean distance or the where used weighted Euclidean distance from the rounded solution subject to the linear budget constraint. In any event we leave these as open issues to be addressed at a later time. In any case the value of rounded solution was rarely too far from the value of the unrounded solution so this may be a relatively insignificant issue. In a real world application a minor violation of the budget constraint may not be significant.

We solved our NLP formulations using the MINOS solver in GAMS. We also used the standard solver available in Microsoft Excel and noted that the two solvers provided the same solutions. The solution time from either of the solvers was very short (less than a

second). The simulations however took on the order of 20 minutes for each instance of the problem, making sensitivity analysis very time consuming.

13 Test Models

In order to test the effectiveness of our approach we constructed some simple test problems and compared our methodology with a policy where the component base stocks are set in a manner that provides an equal protection level on a component by component basis. In the case of the smaller problem that was studied, we also attempted to assess the overall quality of our solution through an intelligently devised, fairly exhaustive interval search.

13.1 Heuristics for comparison

13.1.1 The equal allocation policy

In the remainder of this paper we refer to the equal protection level policy as the equal allocation policy (EAP). Let μ_k and σ_k denote the mean and standard deviation of the demand over lead-time for component k . Under the EAP we set the base stock level B_k for component k such that

$$m_k = \mu_k + z\sigma_k$$

Here in effect we assume that the component demands over their respective replenishment lead-times are normally distributed. Under the EAP we use a common value of z across all of the components. The EAP policy is commonly used in practice. If we assume that demand is normally distributed then the EAP ensures that the probability of a stockout is the same across all components. In a way we are trying to determine a methodology that outperforms this policy. It is quite reasonable to conjecture the existence of alternate policies that outperform the EAP policy based on the following observations.

- The unavailability of a component used in multiple end-items could potentially cause several different types of end-items to wait whereas the unavailability of a component

used in a single end-item could potentially cause only one type of end-item to wait. In such a setting it may be prudent to let the z value for a component used in multiple end-items exceed the z value for a component used in relatively fewer places.

- We could have a very cheap component that causes us to hold very expensive components on hand and the converse. In such an instance we may wish to use a fairly large z value for the cheap component.

Our formulation explicitly takes these aspects of the situation into account through the objective function coefficients.

13.1.2 Three other heuristics

For Problem 2 (4 end-items, 12 components) we tested three other heuristics. The heuristics were constructed by altering the objective function of the math program, and leaving the constraints the same.

We conjecture that the component queue utilization levels serve as a proxy for the component fill rates. For an M/M/m system the relationships between the utilization level and the probability of an arriving demand not being met from stock are well known. Specifically as the utilization level decreases both these quantities decrease. However since we have an assembly system the end-item fill rates are an unknown function of the component fill rates. In heur 1 we use a linear functional form as a surrogate for the end-item fill rates while in heur 2 we use a product form. In essence we use these functions of the component utilization levels to serve as proxies for the unknown end-item utilization levels.

heur 1: Minimize the sum of the utilization levels over all components

$$\text{Objective Function : } \sum_j \sum_k a_{jk} r_k$$

heur 2: Minimize the sum of the products of the utilization levels

$$\text{Objective Function: } \sum_j \prod_k a_{jk} r_k$$

To develop heur 3 we make the assumption that the end-item fill rates have a product form. However rather than using the utilization levels from the queues, we use the product of the scaled safety stocks as a surrogate for the fill rates. Consider an end-item assembled from two components. Suppose we make the assumption that the mean demands over lead-time for the two components is 20, 200 units respectively. Suppose further that both the components cost the same. Then 1 additional unit will result in 5% more safety stock for the first component and .5% more safety stock for the second component. Since the costs are the same and we assume a product form, intuitively 1 additional component of the first type should provide a better end-item fill rate than 1 additional component of the second type.

heur 3: Maximize the sum of the products of the standard deviation of demand over lead-time scaled safety stock

$$\text{Objective Function: } \sum_j \prod_k a_{jk} (m_k - s_k) / s_k$$

- The first two heuristics could be thought of as first-moment based heuristics as they only involve the first moments of the arrival processes.
- The last heuristic requires the determination of the standard deviation of the demand for each component over its respective lead-time (which requires simulation). This approximation is similar to the one used by Hopp and Spearman (1993)

13.2 The development of problem instances

In order to test the robustness of our approach we had to design problem instances that spanned the potential variety of characteristics that could exist in a real world problem. The following discussion highlights some of these characteristics that we considered.

13.2.1 Model structure characteristics

By model structure we are referring to the bill of material (BOM) structure for the end-items. In order to create such structures we considered the following characteristics: component commonality between end-items and the total number of distinct components used by an end-item. The structures constructed represent a somewhat stylized but not significantly different representation of Teradyne's end-item and PCB structure. The goal was to determine a representative set of sample products that would span product structures observed in practice.

13.2.2 Other end-item characteristics

We considered end-item demand processes such that a range of variability was incorporated into our models. Specifically we considered cases of high, medium, and low levels of variability in these demand processes. Here low refers to $scv's \leq 1$, medium refers to $scv's \in (1,2]$, and high refers to $scv's \in [5,10]$ for the demand processes. Again we have made an effort to capture sufficient variability to cover virtually all of Teradyne's end item demand streams as well as to study the extreme case of very high SCVs.

We do point out that there is a key simplification of the real problem in that the real demand processes tend to be non-stationary but the goal is to address the stationary demand case first.

13.2.3 Component characteristics

We also made sure that after creating such a structure we had sufficient distinct components in our model in order to provide for a sufficient variety of components on the

basis of cost, replenishment lead-time, and the number of distinct end-items that use a component.

13.2.4 Other system characteristics

The upper bound on the system-wide safety stock is highly company dependent. We conjectured that as this value is increased the room to improve system performance should decrease, relative to the EAP. In order to capture a wide variety of real world protection levels we set the value of B corresponding to a range of z values between 1 and 2 (for the EAP). We believe that Teradyne wishes to operate in the 1.3 and 1.7 z value range.

13.2.5 Determining the σ_k and μ_k parameters

Since the end-item demand processes are arbitrary renewal processes (or the superimposition of renewal processes) it is not easy to determine the standard deviation of the component lead-time demands (σ_k) analytically. We use simulation to determine these values. We only have to run this simulation once for each problem scenario. The mean demand over lead-time μ_k for component k is simply the sum of the arrival rates for the end-items using component k multiplied by the replenishment lead-time for component k. These parameters are not needed by our proposed method, but are needed in order to implement the EAP and heur 4.

13.2.6 Performance criteria

We used the expected type II service level both on an order-by-order basis as well as for the whole system in order to compare our method to the EAP method. The type II service level is the percent of the end-item demands that are met from stock, i.e., incur no

shortage of components. In Teradyne’s context this is in fact a key metric as the unavailability of components causes a tremendous amount of chaos.

13.2.7 Data for the base case for two problems

	A	B	C	D
1	1	1	1	
2	1	1		
3	1			1
4			1	1
5				1
6	1			

Table 1: Problem 1 data

	A	B	C	D
1	1	1	1	1
2	1	1	1	1
3	1		1	1
4	1	1	1	
5	1			
6				1
7	1			
8	1			
9		1		
10		1		
11	1			
12			1	

Table 2: Problem 2 data

In the above tables the columns with the labels A-D represent the end-items and rows with the labels 1-6 (1-12) represent components. A value of 1 in the (i,j)th position represents a case where end-item j uses component i. These tables define the problem structure for our two problems. We created several problem instances using these structures. In the data that follows we provide the base case scenarios used with these two problem structures. Data for additional problems analyzed can be found in appendix II.

Problem 1a data:

End-item demand processes

End-Item	Interarrival-time ¹⁰
A	Exponential(.1) [scv = 1]
B	Uniform(.05,.15) [scv= .083]
C	Erlang(.1,3) [scv = .33]
D	Erlang(.1,5) [scv = .2]

Table 3: Problem 1 demand processes low SCVs

Component costs and replenishment lead-times

Component	Cost/unit	Lead-time
1	100	10
2	500	20
3	100	5
4	100	30
5	500	12
6	100	5

Table 4: Problem 1 Component Data

End-item demand processes

End	Interarrival time
A	Gamma(.2,.5) [scv = 5]
B	Gamma(.25,.4) [scv = 4]
C	Gamma(.125,.8) [scv = 8]
D	Gamma(.1,1) [scv = 10]

Table 5a: Problem 2a data - High SCVs

Component costs and replenishment lead-times

Com	Cost	Lea
1	10	5
2	20	10
3	100	5
4	150	15
5	20	5
6	10	10
7	500	5
8	1000	20
9	50	5
10	20	15
11	200	25
12	500	5

Table 6: Problem 2 Component Data

End	Interarrival time
A	Gamma(1,.1) [scv = 1]
B	Gamma(1.25,.08) [scv = .8]
C	Gamma(.625,.16) [scv = 1.6]
D	Gamma(.5,.2) [scv = 2]

Table 5b: Problem 2b data - Medium SCVs

End	Interarrival time
A	Gamma(2,.05) [scv = .5]
B	Gamma(2.5,.04) [scv = .4]
C	Gamma(1.25,.08) [scv = .8]
D	Gamma(1,.1) [scv = 1]

Table 5c: Problem 2c data - Low SCVs

¹⁰ Exponential(a) corresponds to an exponential distribution with mean a, Gamma(a,b) corresponds to Gamma distribution with mean ab and variance ab², Uniform(a,b) corresponds to a uniform distribution on the interval [.05,.15], and Erlang(a,b) corresponds to an Erlang distribution of order b with mean a

13.2.8 Discussion

Problem 1

To create problem 1a we considered a case where we had 4 cheap components fitting two possible lead-time profiles, i.e., short or long and 2 expensive components with the same type of lead-time profiles. Given this set of components we created a cheap component that is used in 3 places (component 1), two cheap component used in 2 places (components 3,4), a single-use cheap component (component 6), a single-use expensive component (component 5) and a multiple use expensive component (component 2).

Problem 2

The motivation for creating problem 2 was to consider the impact of our methodology for somewhat larger sized problems. With this problem we decided to create two unique and one multiple-use component for each of the following attribute combinations: (cheap, short lead-time), (cheap, long lead-time), (expensive, short lead-time), (expensive, long lead-time). This yields a total of 12 components as represented in problem 2's problem structure matrix. We selected the gamma distribution, as it can be used for an arbitrary range of scv values.

Additional problems studied

Fixing the problem structures as depicted in Tables 1 and 2 we constructed several other problem instances that are tabulated below [problems using problem 1's where-used structure will be labeled problem 1(letter) and the same holds for problem 2]

Problem number	Description
Problem 1b	More than one copy of a given component can be required by the end-items
Problem 2b	Medium variability in the arrival process
Problem 2c	Low variability in the arrival process

Table 7: Description of additional problems

13.3 Results

This section is organized as follows: We begin by presenting the results of the NLPs that were solved for problems 1a and problem 2a. We discuss these results. The results of the NLP are inputs to the simulation model. We then present the results of our simulation study for problem 1a, 1b, problems 2a-2c. This is followed by the description of an interval search that we performed for problem 1a in order to assess the overall quality of our solution. We then discuss these results.

We begin with the NLP outputs for problems 1a and 2a. The numbers presented for each component are the safety stock values expressed as a multiple of the respective standard deviation of the demand over lead-time. Under the EAP we would pick a common value of z . In contrast, for our method, the value of z is allowed to vary across the components. In these tables #WU refers to the number of distinct end-items that use a particular component, RLT refers to the replenishment lead-time, and the z value in the first row is the budget constraint value obtained as follows: As noted earlier for each model we varied the budget constraint value B so that it corresponded to a z value range between 1 and 3 (1 and 2 for problem 2).

For example a z value of z_1 corresponds to a budget constraint value of $B = .2(z_1 \sum_{k \in K} s_k)$

(here .2 is an arbitrarily selected holding cost rate).

Problem 1 a			z value								
Component	\$/unit	#WU	RLT	1	1.2	1.4	1.6	1.8	2	2.5	3
1	100	3	10	1.5	1.8	2	2.3	2.5	2.8	3.3	3.8
2	500	2	20	.8	.9	1.1	1.3	1.5	1.6	2.1	2.5
3	100	2	5	1.4	1.6	1.9	2.1	2.4	2.5	3.1	3.5
4	100	2	30	1.8	2.2	2.4	2.8	3.1	3.7	4.2	5.1
5	500	1	12	.8	1	1.2	1.4	1.6	1.8	2.6	3
6	100	1	5	1.3	1.6	1.7	2	2.1	2.3	2.6	3.1

Table 8: Problem 1a, component base stocks as multiples of standard deviations of component demands over respective lead-times(post rounding)

Component	\$/unit	#WU	RLT	1	1.5	2
1	10	4	5	1.4	1.8	2.1
2	20	4	10	1.7	2.1	2.5
3	100	3	5	1.2	1.6	2.1
4	150	3	15	1.2	1.7	2.2
5	20	1	5	1.9	2.6	3.2
6	10	1	10	1.6	2	2.5
7	500	1	5	1	1.6	2.2
8	1000	1	20	.8	1.3	1.8
9	50	1	5	1.7	2.2	3
10	20	1	15	2.2	2.9	3.6
11	200	1	25	1.2	1.7	2.3
12	500	1	5	1	1.4	1.9

Table 9: Same as table 8 but for problem 2a

13.4 Some observations

- In both problems we notice that expensive components have lower protection levels than cheaper components. In table 8 look at the values for components 2 and 5 and in table 9 look at the values for components 7,8,12. This is not surprising given that the respective objective function coefficients are relatively smaller than the objective function coefficients for the cheaper components.
- Components that are used in multiple end-items do not necessarily have higher protection levels associated with them; observe table 9 components 1 and 2 versus components 5 and 10. A reasonable conjecture may be that the stockout of these components could

potentially hold up more end-item types than stockout for items specific to only one end-item. Our solutions do not reflect this fact (however, lead-times may also play a role here).

As mentioned in the last section we used the type II service level as a metric to evaluate both our proposed methodology as well as the EAP. We compare the various heuristics by using simulation to determine the type II service levels.

In tables 10 and 11 we present these results on an end-item by end-item basis (A-D) for problem 1 as well as for the aggregate system (sys.). In tables 12,13 we present analogous results for problem 2. The data presented in tables 10-13 corresponds to the difference between the approximated mean type II service level for our policy and the EAP (positive values represent a net benefit). For example the number 16.9 in the first row of table 10 under the column for component C refers to a case where our policy provides on average a 16.9% better type II service level than the EAP. The columns in these tables represent the estimated mean of the type II service level for our policy. In all of the tables the cells marked with a (*) correspond to statistically insignificant data as in these cases the .95 confidence interval crossed zero. For all of the data in the tables the first two digits are statistically significant (as all relevant standard errors are strictly less than .001).

Z	A	B	C	D	Sys.
1	5.7	1.7 (*)	16.9	8.1	8.1
1.2	5	0.6 (*)	17.3	8.4	7.8
1.4	3.4	-0.1 (*)	12.2	7	5.6
1.6	2.5	-0.80 (*)	9.8	5.2	4.2
1.8	1.2 (*)	-1.1(*)	5.9	3.7	2.4
2	0.1 (*)	-1.4(*)	4.3	2.2	1.3

Table 10: Percent improvement (loss) in Type II service; our policy vs EAP, problem 1a

z	A	B	C	D	Sys.
1	8.4	-0.3 (*)	12.7	18.3	9.8
1.2	7.4	-1.5(*)	10.5	16.8	8.3
1.4	4.9	-1.9	8	15	6.4
1.6	3.8	-2.3	5.3	12.7	4.9
1.8	2.1	-2.1	3.7	9.1	3.3
2	1 (*)	-2.1	1.9	5.9	1.7

Table 11: Percent improvement (loss) in Type II service; our policy vs EAP, problem 1b

z	Heuristic	A	B	C	D	Sys.
1	EAP	4.1	21.6	6.7	12.6	11.3
	heur1	37.52	-16.56	33.79	48.17	25.71
	heur2	26.89	26.28	21.61	-19.65	13.75
	heur3	29.01	28.32	23.61	-18.47	15.31
1.5	EAP	2.2	14.2	3	6.3	6.4
	Heur1	13.76	57.03	35.36	40.58	36.95
	heur2	41.57	37.65	31.73	-9.73	25.31
	heur3	43.61	39.43	33.6	-8.84	26.72
2	EAP	0.8 (*)	7	0.3 (*)	2.3	2.6
	heur1	19.36	62.51	44.74	45.96	43.44
	heur2	45.51	26.79	23.36	-4.36	22.87
	heur3	47.43	28.41	25.11	-3.8	24.13

Table 12: Percent improvement (loss) in Type II service. Problem 2a

z		A	B	C	D	sys.
1	EAP	10.3	35.6	13.7	24.9	21.2
	heur1	41.52	-6.74	38.72	57.21	32.67
	heur2	39.94	37.14	40.88	-8.68	27.29
	heur3	41.89	38.9	42.43	-8.05	28.52
1.5	EAP	7.4	21.4	9.7	14.2	13.2
	heur1	15.1	62.83	39.96	46.23	41.18
	heur2	51.01	44.24	44.43	-3.48	34.04
	heur3	52.84	45.8	45.96	-3.12	35.18
2	EAP	4.3	9.9	4.1	6.5	6.2
	heur1	19.77	64.43	45.19	48.39	44.61
	heur2	59.08	46.05	43.64	-1.33	36.87
	heur3	60.74	47.57	45.16	-1.12	37.97

Table 13: Percent improvement (loss) in Type II service. Problem 2b

z		A	B	C	D	Sys.
1	EAP	7.4	35	15.1	27	21.3
	heur1	41.5	-6.08	39.82	59.03	33.56
	heur2	41.27	38.64	44.26	16.57	35.18
	heur3	42.96	40.23	45.66	17.44	36.32
1.5	EAP	3.3	21.3	9.9	15.1	12.4
	heur1	59.47	-1.75	53.75	63.97	43.87
	heur2	54.51	43.11	55.38	-2.15	37.72
	heur3	56.13	44.58	56.66	-1.89	38.68
2	EAP	1.6	9.4	3.5	6.8	5.4
	heur1	17.81	64	45.44	47.64	43.79
	heur2	61.55	47.05	51.54	-0.55	39.93
	heur3	63.07	48.48	52.91	-0.43	40.91

Table 14: Percent improvement (loss) in Type II service. Problem 2c

Note that our method outperforms the EAP in nearly all cases. In the few cases that the EAP outperforms our method it does only slightly better and only on one end-item (for example observe table 11, end-item B). We observe that for end-item B the EAP can do about 2% better, but it performs considerably worse for the other end-items as well as at a system level. Notice also that our method performs relatively better for systems with lower demand process variability (problems 2b, 2c) than for those with much higher demand process variability (problem 2a). As we pointed out earlier the SCVs from problem 2b, 2c are representative of those seen at Teradyne and thus we expect that our method will work well for their situation.

Heur1 – heur3 perform considerably worse than our method for most cases. There are a few cases (for example look at end-item D in table 12, for the $z = 1$ case) in which these heuristics perform considerably better for one of the four end-items, but in these cases they perform considerably worse for the other end-items as well as at a system level. In order to gain further insight into these heuristics, consider the unit costs for each of the end-items. End-items A-D cost 2000, 250, 780, and 30 units respectively. When heur2 and heur3 perform better they only do so for end-item D, the cheapest end-item. There are also a few cases in which heur1 does better (for example look at end-item B, in table 12, for the $z = 1$ case). However in these cases heur1 performs considerably worse for the other end-items as well as at a system level. Notice that unlike our method these heuristics do not perform monotonically better relative to our method as the right hand side of the budget constraint is increased. This is clearly a desirable feature for a heuristic, since if it holds it would imply that higher budget constraints result in better service (in relative sense). This does seem to hold when we compare the EAP to our method. Notice that our method performs better for

almost all cases. However as the budget constraint is increased, the relative improvement decreases. It appears as though $heur1 - heur3$ allocate more of the budget constraint to the relatively cheaper end-items. The result is as anticipated, i.e., better service for the cheaper end-items and worse service for the more expensive end-items. This is perhaps the reverse of how a company might wish to operate, as it would result in poorer service to customers that purchase more expensive products over those that purchase cheaper products.

13.5 An alternate method to evaluate the results

In single product 1-1 replenishment inventory models the base stock is often set using the formula $B = \mu + z\sigma$. If the demand over lead-time is normally distributed then we can characterize the type II service level using a standard normal table for any value of z . We did some rudimentary tests to determine whether the induced component demands over their respective lead-times are normal; this does seem to be the case, but for full verification we would need to do additional data analysis. If we assume that the component demand processes are normally distributed, then an upper bound on the system-wide type II service level is given by the z value chosen. In the EAP we use a common z value across all components. This z value then provides us with an upper bound on the system-wide achievable type II service level. For example in the EAP a z value of 1 will provide us with an upper bound type II system-wide service of .8413 (obtained from a standard normal table). This is a strong (unachievable) upper bound for aggregate system-wide type II service since the end-items in our examples use multiple components. We would expect that the gap between simulated performance using the EAP and this upper bound would be worse for end-items using more components than end-items that use fewer components (this is in fact true in our examples). However, this argument establishes the fact that this bound is an unachievable

target on a system-wide basis for any policy. We can then use this bound as follows: Determine the difference between our policy and the EAP and determine the difference in performance between the EAP and the bound and take the ratio. This quantity is the percent of the maximal possible improvement that is achieved by using our method. In the tables below we present this data. The columns E(u) and h(u) correspond to the simulated type II service level for the system. The column titled Up provides the upper bound in each case and the % gap filled column provides the ratio of improvement $(100*(h(u)-E(u))/(Up-E(u)))$.

Problem 1a

	E(u)	h(u)	Up	% gap filled
1	70	78.1	84.1	57
1.2	76	83.8	88.4	63
1.4	82.8	88.4	91.9	62
1.6	87.4	91.6	94.5	59
1.8	91.8	94.2	96.4	52
2	94.5	95.8	97.7	41

Table 15: Problem 1a percent gap filled

Problem 1b

	E(u)	h(u)	Up	% gap filled
1	68.6	78.4	84.1	63
1.2	76.4	84.7	88.4	69
1.4	82.7	89.1	91.9	70
1.6	87.6	92.5	94.5	71
1.8	91.5	94.8	96.4	67
2	94.6	96.3	97.7	55

Table 16: Problem 1b percent gap filled

Problem 2a

	E(u)	h(u)	Up	% gap filled
1	57.9	69.2	84.1	43
1.5	77.3	83.7	93.3	40
2	89.7	92.3	97.7	32

Table 17: Problem 2a percent gap filled

Problem 2b

	E(u)	h(u)	Up	% gap filled
1	57.8	79	84.1	81
1.5	77.8	91	93.3	85
2	89.9	96.1	97.7	79

Table 18: Problem 2b percent gap filled

Problem 2c

	E(u)	h(u)	Up	% gap filled
1	59	80.3	84.1	85
1.5	78.9	91.3	93.3	86
2	90.9	96.3	97.7	79

Table 19: Problem 2c percent gap filled

Notice that by analyzing performance in this manner we see quite dramatic improvements further strengthening the validity of our approach¹¹. Notice that as the value of the budget constraint is increased, the room for improvement decreases, and thus the percent gap filled should decrease (which is indeed the case in our results). From these results we can conclude that at least for the problems that we studied, our method captures most of the possible improvement in fill rates at a system level.

13.6 Sensitivity analysis

As our approach is a heuristic approach it is necessary to do an exhaustive interval search to assess the global quality of our solution. In order to do this we have to intelligently perform interval searches. Such an analysis was carried out for problem 1a for a particular value of B. Starting with our solution the following types of interval searches were performed at a $z = 1$ budget constraint level.

- Cheap Vs. expensive.
- Cheap Vs. cheap.
- Expensive Vs. expensive.

For each of the above categories we considered pairs of items such as a cheap item and an expensive item. We proceeded by either increasing or decreasing the base stock level for one of the items and offset this by respectively decreasing or increasing the base stock level for the other item in a manner that preserved the budget constraint. So for example consider two items labeled 1 and 2, for convenience assume that item 1 costs \$10/unit and item 2 costs \$50/unit. Beginning with our original solution we could perform a search by

¹¹ We could lend further credibility to these results by constructing confidence intervals for these ratios.

increasing the base stock level for item 1 by 5 units and decrease the base stock level for item 2 by 1 unit. Such pair wise searches were conducted for all of the above categories. We were unable to find any solutions that were strictly better with respect to at least one end-item and no worse with respect to the other end-items. On a system level the best solution found under the searches yielded only a 1% improvement in the aggregate type II service level. This observation leads us to believe that our solution lies along a relatively flat region of the unknown expected type II service level surface for this test problem.

This sort of an analysis could be performed for other problems and/or for other values of z , however for each search step a simulation has to be performed which makes the overall search time consuming.

13.7 Discussion

As conjectured, the potential benefit from our approach diminishes as the budget constraint value is increased. This is a fairly intuitive conjecture that we tested through our models. To justify the value of our efforts we would argue that most real world systems operate somewhere within the $z=1$ to $z=2$ range. Within this range our methodology performed quite well across all of our test models. In some of our models we observed one end-item that did not do better under our method (at times worse) relative to the EAP [tables 10 and 11 column for end-item B]. However, the potential benefits from the other end-items as well as the aggregate benefit seem to outweigh the potential service level loss for the one product. We can compare the effects of variability on the effectiveness of our approach. Observe the data in Tables 12-14 (the arrival processes become less and less variable as we go from problem 2a to 2c). We also point out that our method does not require the

determination of the standard deviation of the demand over lead-time for the components which is a difficult task.

14 Conclusion, Extensions And Room For Further Analysis

In this paper we have provided a simple heuristic methodology for setting the base stock levels for components in an assemble-to-order environment subject to stochastic demand. Our methodology explicitly accounts for the differences in component attributes such as unit costs, lead-times, and the number of distinct end-items that use a particular component. Based on our simulation studies we see that the methodology yields benefits for different levels of variability in the arrival processes. Hence, we feel that our methodology is fairly robust. This is a significant attribute of our methodology as it increases the viability of its application to real world problems. Furthermore the simplicity of our approach has made it easy to explain to potential users.

The key analytical difficulty is encountered when determining the expected response time for an end-item order. We address this through a simple (but potentially weak) bound on this unknown performance measure that we use as a surrogate. Stronger bounds may improve the performance of our method. An alternative may be to develop approximate expressions for this quantity in a manner similar to the methods presented in Whitt (1993).

We may wish to pursue alternate procedures to create feasible solutions after rounding the solutions to the relevant math programs. Another avenue for further thought may be to carefully study the convexified version of the problem that we discuss in our formulation section.

Based on some observations replacing the term $(1-\rho_k)^2$ by $1-\rho_k^2$ in the final superimposition (of renewal processes) equation does not change the value of the expected queueing time expression given by (3.3) for almost all parameter values but it ensures

convexity for a much wider range of cases. Based on the work in Albin [1983] this modification still satisfies the required properties of a weighting function. In the same paper the author lists another weighting function, the use of which will ensure the convexity of the formulation. We do not pursue this issue further at this point but leave it as an open issue to address in subsequent work based on these ideas.

We have assumed 1-for-1 replenishment policies for components. This may not be reasonable if there are significant fixed ordering costs. We could explore batch service queues to address this issue.

Stochastic lead-times could be addressed using this method if we ignore the possibility of order crossing. Based on our discussion in section 12.4, if we simply ignore the order crossing that could take place, our approach would be even more approximate. The efficacy of using the method ignoring order crossing (when it exists) is to be determined.

We could potentially extend this model to incorporate non-stationary demand processes using the machinery developed in a recent paper by Jennings et al. (1996). In this paper the authors develop approximations for multi-server queues subject to non-stationary arrival processes. In particular they provide a closed form expression for the probability that an arriving customer sees all servers busy. This is equivalent to the case where there is one end-item that uses one component in which the aforementioned probability corresponds to the time average type II service level. It is conceivable that we could develop approximate expressions similar to those developed for our model together with the expressions from this paper to formulate an analogous math program to address the non-stationary demand case.

Finally, we could consider extending this model to cases with a more complex bill of material. An option may be to iteratively solve such problems level by level, but this thought needs considerably more work

Appendix II

Claim1: The Sasekawa approximation for EW(M/M/m) is a convex function of m (treating m as a continuous variable)

Proof:

Let

$f(m) = t \frac{r^{\sqrt{2m+2}-1}}{m(1-r)}$, here $r = \frac{Kt}{m}$. After some algebra this expression can be written as

$$f(m) = f_1(m)f_2(m); \quad \text{with} \quad f_1(m) = \left(\frac{K}{m}\right)^{\sqrt{2m+2}} \quad \text{and} \quad f_2(m) = \frac{t}{(m-K)} \quad \text{and} \quad K = It$$

Note that

$f''(m) = f_1''(m)f_2(m) + 2f_1'(m)f_2'(m) + f_2''(m)f_1(m)$ So a sufficient condition for the convexity of $f(m)$ would be the following :

$f_1(m), f_2(m), f_1''(m), f_2''(m) > 0, f_1'(m), f_2'(m) < 0 \quad \forall m > K$. We proceed by showing that these conditions hold.

Since $m > K$, $f_1(m), f_2(m) > 0$

Let us begin by working with $f_2(m)$.

$$f_2'(m) = \frac{-t}{(m-K)^2}. \text{ Since } m > K \Rightarrow f_2'(m) < 0 \quad \text{and} \quad f_2''(m) = \frac{2t}{(m-K)^3}$$

Since $m > K \Rightarrow f_2''(m) > 0$

$$\text{Now consider } f_1'(m) = \left(\frac{K}{m}\right)^{\sqrt{2m+2}} \left[\left(\frac{-1}{m}\right)(2m+2)^{1/2} + \left(\ln \frac{K}{m}\right)(2m+2)^{-1/2} \right].$$

Since $m > K \Rightarrow \ln \frac{K}{m} < 0 \Rightarrow f_1'(m) < 0$

$$\text{and } f_1''(m) = \left(\frac{K}{m}\right)^{\sqrt{2m+2}} \left[\left(\frac{-1}{m}\right)(2m+2)^{1/2} + \left(\ln \frac{K}{m}\right)(2m+2)^{-1/2} \right]^2 +$$

$$\left(\frac{K}{m}\right)^{\sqrt{2m+2}} \left[-\left(\ln \frac{K}{m}\right)(2m+2)^{-3/2} - \left(\frac{1}{m}\right)(2m+2)^{-1/2} \right]$$

$$\left[+\left(\frac{1}{m^2}\right)(2m+2)^{1/2} - \left(\frac{1}{m}\right)(2m+2)^{-1/2} \right]$$

In order to see that this quantity is positive note

$$\text{that } P(m) = \left(\frac{K}{m}\right)^{\sqrt{2m+2}} \left[\left(\frac{-1}{m}\right)(2m+2)^{1/2} \right]^2$$

$$\left[+\left(\ln \frac{K}{m}\right)(2m+2)^{-1/2} \right]$$

is clearly non - negative and

$$Q(m) = \left(\frac{K}{m}\right)^{\sqrt{2m+2}} \left[-\left(\ln \frac{K}{m}\right)(2m+2)^{-3/2} - \left(\frac{1}{m}\right)(2m+2)^{-1/2} - \right]$$

$$\left[\left(\frac{1}{m}\right)(2m+2)^{-1/2} + \left(\frac{1}{m^2}\right)(2m+2)^{1/2} \right]$$

can be rewritten as

$$\text{R.H.S.} = \left(\frac{K}{m}\right)^{\sqrt{2m+2}} (2m+2)^{-3/2} \left[-\left(\ln \frac{K}{m}\right) - \left(\frac{2}{m}\right)(2m+2) + \left(\frac{1}{m^2}\right)(2m+2)^2 \right]$$

$$= \left(\frac{K}{m}\right)^{\sqrt{2m+2}} (2m+2)^{-3/2} \left[\left(\frac{4m+4}{m^2}\right) + \ln \frac{m}{K} \right]. \text{ Since } m > K \Rightarrow \ln \frac{m}{K} > 0$$

$$\therefore Q(m) > 0 \Rightarrow f_1''(m) > 0$$

Which proves our claim

The approximate expected queueing time expression for the G/D/m queue is the product of Sasekawa's approximation and the scv of the arrival process divided by 2. If a component is used in only one end-item then the scv of the arrival process is a parameter independent of the number of servers at any queue. The only time the scv for the arrival process depends on the number of servers is if we have a superimposed arrival process to a queue as described in section 3. We call the scv for the superimposed process ascv. The following claim provides a sufficient condition under which ascv is both strictly decreasing as well as convex.

Claim 2: The following condition is a sufficient condition for ascv(k) (the scv for the superimposed process) to be both strictly decreasing and convex (treating the number of servers as continuous):

$$\sum_{i \in \Phi(k)} scv(i) * (I_i / \sum_{i \in \Phi(k)} I_i) \geq 1 \quad \text{and } v_i \geq 1 \quad \text{and } r_k = \sum_{i \in \Phi(k)} I_i / m_k \leq 2/3$$

Proof:

To reiterate, we are focusing on the case where multiple distinct end-items use a common component.

For convenience we repeat the definitions of the equations used for the superimposed process.

$$v_k = \left(\frac{\sum_{i:i \in \Phi(k)} I_i^2}{\left(\sum_{i:i \in \Phi(k)} I_i \right)^2} \right)^{-1} \quad (3.4)$$

$$w_k = [1 + 4 * (1 - r_k)^2 (v_k - 1)]^{-1} \quad (3.5)$$

$$ascv(k) = (1 - w_k) + w_k \left(\sum_{i:i \in \Phi(k)} scv(k) * (I_i / \sum_{i:i \in \Phi(k)} I_i) \right) \quad (3.6)$$

We proceed by assuming that the sign of the w_k term in (3.6) is positive.

As in the proof for claim 1 let $K=\lambda\tau$. We can express (3.5) as a function of m (the number of servers) after dropping the dependence on the subscripts and letting $v' = v-1$ as:

$$w(m) = [1 + 4v'(1 - K/m)^2]^{-1}. \quad w(m) > 0 \text{ by observation; furthermore}$$

$$w'(m) = \left(\frac{-8Kv'(m-K)}{m^3} \right) [1 + 4v'(1 - K/m)^2]^{-2} < 0 \text{ by observation as well. And}$$

$$w''(m) = \left(\frac{128K^2v'^2(1 - K/m)^2}{m^4} \right) [1 + 4v'(1 - K/m)^2]^{-3} + \frac{-8Kv'm + 24Kv'(m-K)}{m^4} [1 + 4v'(1 - K/m)^2]^{-2}$$

After some simplification we see that the second term is positive iff $m > 3K/2$ which is equivalent to the condition $\rho < 2/3$. We have conducted a more careful computational study to assess the quality of this bound and it turns out that it is very tight. Within this range of parameter values, $ascv(m) > 0$, $ascv'(m) < 0$, and $ascv''(m) > 0$ and therefore the claim follows.

Claim3: The Whitt-Sasekawa approximation for $EW(G/D/m)$ is a convex function over the region specified in claim2.

Proof: We mimic the argument in the proof of claim 1 by letting $f_1(m) = \text{ascv}(m)$ and $f_2(m) = EW(M/M/m)$ and the claim follows.

Claim 4: For an arbitrary set of non-negative random variables X_k

$$E[\text{MAX}(X_k)] \leq \sum_{k \in \{1,2,\dots,n\}} E[X_k]$$

Proof: Suppose that we jointly generate m vectors of the n random variables. We have not assumed that the X_k are independent therefore we cannot assume that they can be generated independently. For clarity let us place these m vectors in a $m \times n$ matrix. Now the column average of column j provides us with an estimate of the mean of X_k . Form an additional column by taking the largest entry from each row. The average of these m entries provides us with an estimate of the L.H.S. of the above expression. Let a_{ij} denote the entries from this matrix, let α_j denote the column average for column j , let γ_i denote the i^{th} entry in the additional column, and let γ denote the average of these entries:

$$\begin{aligned} \mathbf{g}_i &\leq \sum_{j=1}^n a_{ij} \forall i. \text{ This follows since } a_{ij} \geq 0 \forall i,j \\ \Rightarrow \frac{1}{m} \sum_{i=1}^m \mathbf{g}_i &\leq \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^n a_{ij} \\ \Rightarrow \mathbf{g} &\leq \sum_{j=1}^n \mathbf{a}_j \end{aligned}$$

15 References

- Albin, S.L., "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues", *Operations Research*, (1983), 1133-1162
- Baker, K.R., "Safety Stocks and Component Commonality", *Journal of Operations Management*, 6,1 (1985), 13-23
- _____, M.J. Magazine AND L.W. Nuttle, "The Effect of Commonality on Safety Stock in a Simple Inventory Model", *Management Sci.*, 32,8 (1986), 982-988
- Cinlar, E., Superimposition of Point Processes. In *Stochastic Point Processes: Statistical Analysis, Theory and Application*, 1972, pp. 549-606, P. A. W. Lewis (ed.). John Wiley & Sons, New York.
- Collier, D.A., "Aggregate Safety Stock Levels and Component Part Commonality", *Management Sci.*, 28, 11 (1982), 1296-1303
- Ettl, M., G.E. Feigin, G.Y. Lin AND D.D. Yao, "A Supply Network Model with Base stock Control and Service Requirements", unpublished paper, May 1996
- Gallager, R.G., *Discrete Stochastic Processes*, Kluwer Academic Publishers, 1996
- Gallien J., L.M. Wein, "A Simple and Effective Component Procurement Policy for Stochastic Assembly Systems", Working paper, MIT Sloan School of Management, 1998
- Glasserman P., Y. Wang, "Fill-Rate Bottlenecks in Production-Inventory Networks", *MSOM*, 1,1 (1999), 62-76
- Glasserman P., Y. Wang, "Leadtime-Inventory Trade-Offs in Assemble-To-Order Systems", *Operations Research*, 46, 6 (1998), 858-871
- Graves, S.C., "A Multiechelon Inventory Model for a Repairable Item With One-For-One Replenishment", *Management Sci.*, 31,10 (1985), 1247-1256
- _____, "A Multiechelon Inventory Model with Fixed Replenishment Intervals", *Management Sci.*, 42, 1, (1996), 1-18
- _____, "A Single-Item Inventory Model for a Non-Stationary Demand Process", *MSOM*, 1,1 (1999), 50-61
- Hausman W.H., H.L. Lee, AND A. X. Zhang, "Joint Demand Fulfillment Probability in a Multi-item Inventory System with Independent Order-up-to Policies", *EJOR*, 109 (1998), 646-659

- Hopp, W.J. AND M.L. Spearman, "Setting Safety Lead-times for Purchased Components in Assembly Systems", *IIE Transactions*, 25, 2 (1993), 2-11
- Hillestad, R.J., "Dyna-METRIC: Dynamic Multi-Echelon Technique for Recoverable Item Control", Rand Corporation report # R-2785-AF, July 1982
- Hillestad, R.J., M.J. Carrillo, "Models and Techniques for Recoverable Item Stockage when Demand and the Repair Process are Nonstationary – Part I : Performance Measurement", Rand Corporation Report # N-1482-AF, May 1980
- Jennings, O.B., A. Mandelbaum, W.A. Massey AND W. Whitt, "Server Staffing to Meet Time-Varying Demand", *Management Sci.*, 42, 10 (1996), 1383-1394
- Law, A.M. AND W.D. Kelton, *Simulation Modeling & Analysis*, McGraw-Hill, New York, 1991
- Simpson, K.F., "In-Process Inventories", *Operations Research*, 6 (1958), 863-873
- Song, J.-S., S.H. Xu, AND B . Liu, "Order-Fulfillment Performance Measures in an Assemble-to-Order System with Stochastic Lead-times", unpublished paper, April 1996
- Song, J.-S., "On the Order Fill Rate in a Multi-Item Bas-Stock Inventory System", *Operations Research*, 46, 6 (1998), 831-845
- Song, J.-S., P.H. Zipkin, "Inventory Control in a Fluctuating Demand Environment", *Operations Research*, 41, 2 (1993), 351-370
- Song, J.-S., P.H. Zipkin, "Evaluation of Base-Stock Policies in Multiechelon Inventory Systems with State-Dependent Demands. Part II: State-Dependent Depot Policies", *Naval Research Logistics*, 43 (1996), 381-396
- Sox C.R., L.J Thomas, J. O. McClain, "Coordinating Production and Inventory to Improve Service", *Management Science*, 43, 9(1996) 1189-1197
- Sherbrooke, C.C., "METRIC: A Multi-Echelon Technique for Recoverable Item Control.", *Operations Research*, 34, (1968), 122-141
- Whitt, W., "The Queueing Network Analyzer", *The Bell System Technical Journal*, 62, 9 (1983), 2779-2813
- _____, "Approximations for the GI/G/m Queue", *Production and Operations Management*, 2, 2 (1993), 114-161
- Zhang, A.X., "Demand Fulfillment Rates in an Assemble-To-Order System with Multiple Products and Dependent Demands", *Production and Operations Management*, 6, 3 (1997), 309-324