Setting planned leadtimes in customer-order-driven assembly systems

Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Setting Planned Leadtimes in Customer-Order-Driven Assembly Systems

Zümbül Atan, Ton de Kok, Nico P. Dellaert, Richard van Boxel, Fred Janssen

To cite this article:
Zümbül Atan, Ton de Kok, Nico P. Dellaert, Richard van Boxel, Fred Janssen (2016) Setting Planned Leadtimes in Customer-Order-Driven Assembly Systems. Manufacturing & Service Operations Management 18(1):122-140. http://dx.doi.org/10.1287/msom.2015.0565

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 18, No. 1, Winter 2016, pp. 122–140
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2015.0565

Setting Planned Leadtimes in Customer-Order-Driven Assembly Systems

Zümbül Atan, Ton de Kok, Nico P. Dellaert
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, 5600 MB, Eindhoven, Nethlerlands {z.atan@tue.nl, a.g.d.kok@tue.nl, n.p.dellaert@tue.nl}

Richard van Boxel, Fred Janssen
ASML, 5504DR Veldhoven, Netherlands {richard.van.boxel@asml.com, fred.janssen@asml.com}

We study an assembly system with a number of parallel multistage processes feeding a multistage final assembly process. Each stage has a stochastic throughput time. We assume that the system is controlled by planned leadtimes at each stage. From these planned leadtimes the start and due times of all stages can be derived. If a job finishes at a particular stage and has to wait before the start of the next job(s), a holding cost proportional to the waiting time is incurred. A penalty cost proportional to the lateness is incurred when the last stage of the final assembly process finishes after its due time. The objective is to determine planned leadtimes for each individual stage, such that the expected cost of a customer order is minimized.

We derive the recursive equations for the tardiness and earliness at all stages and an exact expression for the expected cost. We discuss the similarity between these expressions and those for serial inventory systems. Based on this observation and a conjecture related to the generalized Newsvendor equations, we develop an iterative heuristic procedure. Comparison with a numerical optimization method confirms the accuracy of the heuristic. Finally, we discuss an application of the model to a real-life case, showing the added value of a system-wide optimization of planned leadtimes compared to current practice.

Keywords: planned leadtimes; customer-order-driven; assembly systems
History: Received: April 2, 2015; accepted: August 8, 2015. Published online in *Articles in Advance* November 25, 2015.

1. Introduction

The aim of any supply chain is to provide timely delivery of products in the most cost-effective way. Attaining this goal is a particularly difficult challenge for companies in the capital goods industry because of their low and highly unpredictable demand, low production volumes, long procurement leadtimes, huge capacity requirements, and low inventory turnovers (Hicks 2004, Hicks and Pongcharoen 2006).

The semiconductor industry shares some of the challenges and characteristics of the capital goods industry. As a technology enabler, the semiconductor industry has grown rapidly from its formation in 1960 to a $305.6 billion market in 2013 (Rosso 2014). A well-known company contributing to this growth is ASML (originally ASM Lithography). Founded in 1984 and headquartered in Veldhoven, the Netherlands, ASML is the world's leading provider of lithography systems with €5.86 billion in sales in 2014 (ASML 2014). These are complex and expensive systems that are used in the production of integrated circuits and microchips.

As a supplier for the semiconductor industry, ASML's demand pattern has a strong relationship with the demand for semiconductors. Historically, the semiconductor industry has been volatile, with periods of rapid growth followed by downturns. This implies a fluctuating demand for ASML's systems. ASML's customers include all of the world's leading chip manufacturers, who use ASML's systems to manufacture a wide range of different chips. ASML constantly improves the capabilities of its lithography systems, allowing the customers to make smaller, faster, and more energy-efficient chips. ASML adapts its technologies to its customers' requirements by innovating its systems. As a result, the fluctuating demand, together with the risk of technology obsolescence, makes it infeasible to hold inventories of fully manufactured systems.

ASML faces uncertainties in the supply chain and its manufacturing processes. The company procures components and submodules from approximately 700 different suppliers. Careful component and submodule inventory management and intense communication between ASML and its suppliers (how to reschedule component and submodule orders) ensure that at the moment of order release all components and submodules are available for a timely start of their assembly into the main modules of the lithography systems. Both the assembly of submodules into

modules and the assembly of modules into systems involve complicated processes and extensive functional testing, which together require a large number of different operations.

High product complexity implies variability in ASML's operations throughput times. As order release is planned such that assembly of submodules can start on their planned start date, ASML decouples the uncertainty in the manufacturing process from uncertainties in demand and supplies. Thus, after order release the only uncertainty to be taken into account is the throughput time uncertainty of the various phases in the assembly process.

Holding safety stocks and planning for safety times are two techniques that companies use to absorb uncertainties in demand, procurement, and manufacturing. At ASML, safety times are preferred since each type of system has its own specific modules. In fact, each module is associated with an end product. Because of the uniqueness of the modules, keeping safety stock is not an option. Another reason is the dynamic nature of the semiconductor industry. If a module is kept in stock and a redesign is executed, the module must either undergo rework or become obsolete. Both outcomes are costly and undesirable. Another problem is the need to produce and store the modules in cleanrooms with a controlled level of contamination. These rooms are expensive to build and maintain. Therefore, keeping safety module stocks implies high holding costs.

Multiple papers investigate the use of safety times versus safety stocks and provide guidelines on which technique works better under different circumstances. The simulation study by Whybark and Williams (1976) suggests using safety times instead of safety stocks when uncertainties in demand and supply are mostly due to timing rather than quantity. In another simulation study, Molinder (1997) concludes that using safety times instead of safety stocks results in lower costs when the variabilities in demand and leadtime are high at the same time. Yano (1987b) argues that when all units in a batched order are produced at the same time, the safety stock needs to be as large as the batch size. Therefore, using safety times should be a preferred strategy, especially when the batch size is large.

ASML and other companies using safety times as a buffering technique face the challenging problem of determining *planned leadtimes* for their processes. The planned leadtime is the sum of the average leadtime and the safety time. The difficulty of the problem arises from the interactions among multiple processes. The tardiness of one process might imply delays in the subsequent processes and, eventually, late delivery of products. In this paper, we study the problem of optimizing the planned leadtimes. Motivated by ASML's manufacturing environment, we consider an assembly system that consists of multiple processing stages each delivering subassemblies (or modules) to its succeeding processing stage, eventually yielding the final product, which in this case is the lithography system. The throughput times at all stages are stochastic. The system incurs holding costs from the start of the process until delivery of the system to the customer and a penalty cost for late delivery of the final product. Our objective is to find the planned leadtimes of all the stages so that the sum of holding and penalty costs is minimized.

As acknowledged in Yano (1987b) and Axsäter (2005), it is difficult to obtain exact solutions for this problem, especially for large assembly systems. We contribute to the literature by proposing an iterative heuristic procedure that relies on a conjecture related to the generalized Newsvendor equations. We compare our heuristic with a procedure based on the well-known nonlinear optimization method of Davidon-Fletcher-Powell (Press et al. 2007). Our results indicate that the heuristic performs extremely well, with an average percentage cost difference from the Davidon-Fletcher-Powell (DFP) method of only 1.33%.

At ASML, the leadtimes are planned using a decomposition approach. For each planned stage, throughput time data are collected over a period of time and their mean and standard deviation are computed. The planned leadtime of each stage is set to a fixed percentile of the normal distribution with the computed mean and standard deviation.

Our contribution to ASML practice and thereby to the practice of order-driven manufacturing is as follows: First, we show empirical validity of our model by comparing the actual on-time delivery of two lithography systems and the most complicated main module with the on-time delivery percentage according to the model. Second, we show that the decomposition approach described above can be improved by considering the overall process from release of submodules to the delivery of the systems to the customer. Our optimization method yields overall cycle time reductions of 10%–11% compared to ASML's decomposition method. When all cleanrooms are occupied, the throughput reduction of 10%–11% yields an increase in output of 11%–12%.

In the remainder of this section, we provide a review of studies related to the problem of setting planned leadtimes. We then describe ASML's manufacturing environment and the procedure used at ASML for setting planned leadtimes. In §2, we explain the derivation of the average cost expression. We detail the development of the iterative heuristic procedure in §3. We analyze the performance of our heuristic in §4. The application of this at ASML is

discussed in §5. We provide concluding remarks and future research directions in §6.

1.1. Literature Review

The problem of determining the optimal planned leadtimes for general multistage production systems is recognized to be very difficult in terms of obtaining exact solutions. Earlier work focuses mostly on single-stage systems and systems with specific structures.

One of the earliest studies analyzing the single-stage problem is by Weeks (1981), who establishes the equivalence between the Newsvendor problem and the problem of planned leadtime optimization. Subsequently, Matsuura and Tsubone (1993), Matsuura et al. (1996), and Buzacott and Shanthikumar (1994) develop single-stage models. The model by Matsuura and Tsubone (1993) is suitable for material requirements planning (MRP) systems and considers the trade-off between work-in-process inventories and variations in capacity requirements while determining the planned leadtime. The model by Matsuura et al. (1996) is for multioperation jobs and, similar to Matsuura and Tsubone (1993), it takes the trade-off between work-in-process inventories and capacity requirement variations on a bottleneck job into consideration. Buzacott and Shanthikumar (1994) compare the effectiveness of safety times versus safety stocks in single-stage MRP-controlled manufacturing systems. They conclude that safety times are only preferable to safety stocks when future required shipments over the leadtime can be accurately forecast. Otherwise, holding safety stocks is a better strategy to cope with changes in customer demands.

Earlier work on setting planned leadtimes in multiechelon supply chains includes Yano (1987a), Gong et al. (1994), and Yano (1987c). Yano (1987a) studies two-stage serial production systems and, assuming quasi-convexity of the cost function, develops an algorithm to solve it. Yano (1987a) generalizes this algorithm to serial systems with more than two locations. Studying the same system as Yano (1987a), Gong et al. (1994) show that the problem of determining the optimal planned leadtimes is equivalent to that of determining the optimal base-stock levels in serial inventory systems. Therefore, the well-known algorithm by Clark and Scarf (1960) can be used to find the optimal planned leadtimes. Yano (1987c) studies the same problem as Yano (1987a) and Gong et al. (1994) for one-warehouse, two-retailer distribution systems. The author suggests two heuristic policies that rely on the optimal solutions for the decoupled serial systems.

The first research on planned leadtimes for assembly systems is by Yano (1987b), who considers a two-component assembly process with stochastic component production/procurement and stochastic assembly processing times. The problem is to determine when to produce/procure the components and when to start the assembly process. On the one hand, inventory holding costs are charged if components wait until the assembly process starts and if the finished product waits to be shipped to the customer. On the other hand, if the final product is available after the promised delivery date, penalty costs are charged per time unit late. The problem is formulated as a nonlinear program. Yano shows that the cost function is not convex for all leadtimes but has some properties that can be used when solving the problem numerically. Yano concludes that the objective function is not well-behaved even for a simple assembly system with two components and argues that the general case with more than two components will be even less well-behaved. Yano acknowledges that we must resort to heuristics for assembly systems with an arbitrary number of components.

After the pioneering work by Yano (1987b), multiple researchers tackled the problem of planned leadtime optimization for assembly systems with stochastic leadtimes. Hopp and Spearman (1993) consider an assembly system in which multiple components are purchased and then assembled. They develop an approximate procedure to solve the problem of leadtime determination. Song et al. (2001) develop a recursive heuristic method for due date planning of all the components and the final assembly process. The method does not minimize the cost but allows the system to meet specified service targets. Axsäter (2005) studies a multiechelon assembly system with the objective of choosing the starting times of different processes to minimize the total expected holding and penalty costs. The author suggests an approximate decomposition technique that does not perform well for systems with more than two echelons. Chauhan et al. (2009) consider a single-period model for a multicomponent assembly process where the component procurement times are random variables with known distributions. The author develops an approximate procedure to determine the release dates of all the components. Another stream of research views the planned leadtimes as a tactical decision to capture the trade-off between resource requirements and Work in Process in the light of demand uncertainty (Chhaochhria and Graves 2013; Teo et al. 2011, 2012).

Our work contributes to the literature by providing a fast and accurate heuristic procedure that outperforms existing approaches and works for assembly systems with two and more echelons. The heuristic is based on an important conjecture that the optimal planned leadtimes solution satisfies so-called generalized Newsvendor equations (Diks and de Kok 1998). As outlined in §6, our heuristic has the potential to be extended to distribution systems and even multiechelon systems with general structures.

1.2. ASML Case Study

ASML produces lithography systems, which are used in the production of integrated circuits and microchips. Lithography systems are assembled from modules. ASML's manufacturing process necessitates timely deliveries of system components by external and internal suppliers to start the final assembly of the systems. The external suppliers provide the components and submodules and the internal suppliers are in charge of producing the modules.

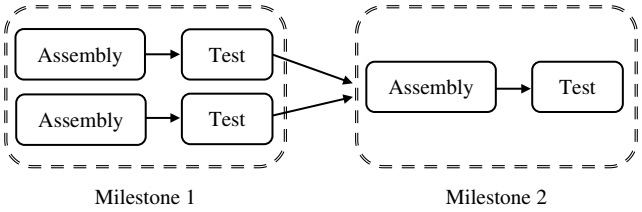
In this study we focus on one module and two systems. The module is the most complex module in a lithography system. We refer to the analyzed systems as Systems A and B. System A was introduced at the end of 2010 and its successor, System B, was introduced at the beginning of 2012. Next, we explain ASML's module production and final assembly processes and summarize its current practice for setting planned leadtimes.

1.2.1. Manufacturing Processes. Modules are produced by assembling the components and submodules in work centers. Each work center is responsible for the production of one module or submodule type. The module production itself is a complex process requiring multiple operations in series and for some work centers in parallel, too. An operation is defined to be the most basic step, requiring one to four hours for completion. The production of one module from each type requires the completion of 100–150 operations. Although the execution of these operations is based on standard procedures, many adjustments may be required during execution based on the specifications demanded by customers and the results of intermediate tests.

To manage the complex assembly process of the modules, operations are clustered into so-called *milestones*. The total processing time of a milestone is 2–20 days. The notion of milestones is mainly used for planning purposes. The leadtimes are planned *not* for the individual operations *but only* for the milestones.

The operations constituting a milestone are performed by the same workforce inside the same work center. Each workforce consists of 5–25 employees with expertise in performing the required operations. As the plan for the assembly of a system is made, the workforce is informed about the planned start time of its milestone. Each milestone starts either at its planned start time or later if the previous milestones are not completed on time. Early completion of a milestone does not permit for an early start of the consecutive milestone. This will require additional planning for getting the whole team of employees ready to perform their operations earlier than planned. ASML planners require the milestone workforces to stick to their designated starting times to avoid any complications that might result from rescheduling.

Figure 1 Production of the Most Complex Module



In this study we analyze two ASML systems: Systems A and B. Both systems are assembled from seven modules, six of which are manufactured in their own work centers and each of which has its own planned leadtime. The most complex module requires the longest production time. At ASML, the operations of this complex module are grouped into two distinct milestones and a planned leadtime is calculated for each milestone. The first milestone is used for assembly and test of two similar submodules in parallel. The second milestone, performed by a different workforce in a different work center, concerns the assembly and test of the most complex module (Figure 1). After all seven modules are produced, the final assembly of Systems A and B can start. The final assembly process takes place in rooms called *cabins*, which are specifically designed for the final assembly process. Together with the work centers, the cabins are located in the cleanrooms. The final assembly and the subsequent tests constitute three milestones. As explained above, different workforces are responsible for the completion of each milestone. Assembled and tested systems are packed and shipped to customers. The average number of operations required for the completion of the final assembly and test processes is approximately 800. We refer to Figure 2 for a schematic overview of Systems A and B's production processes. Among the module production, final assembly, test, and packing processes, only the first three account for the uncertainty in the production leadtimes. The packing times do not depend on the system type; they are fairly standard and stable. Therefore, we exclude the packing process from our analysis.

1.2.2. Calculation of Planned Leadtimes. We now explain ASML's current method for setting the milestone planned leadtimes.

Each milestone consists of multiple operations. The average time required for the completion of an operation is called the *operation leadtime*. In addition to its operation leadtime, each operation requires an additional time because of the uncertainties related to production. This additional time is called the *buffer time*.

At ASML, each milestone has an *internal* and an *external* cycle time. Consider a milestone with N operations. Let t_n and b_n be the operation leadtime and

Figure 2 Systems A and B Production Processes

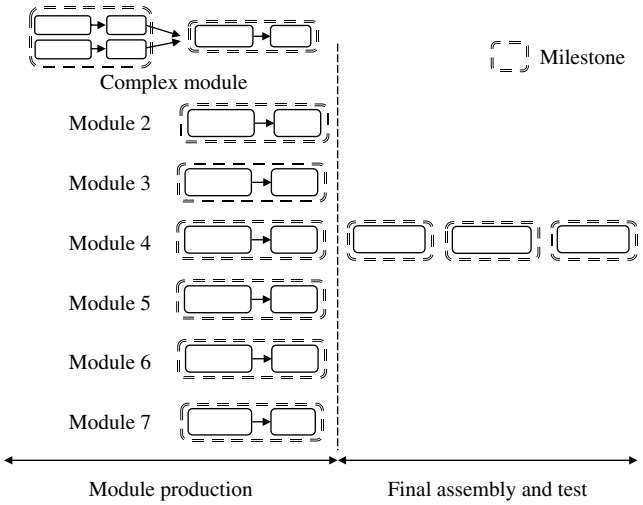
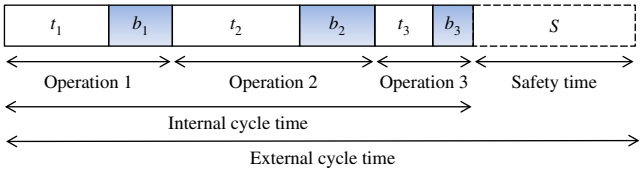


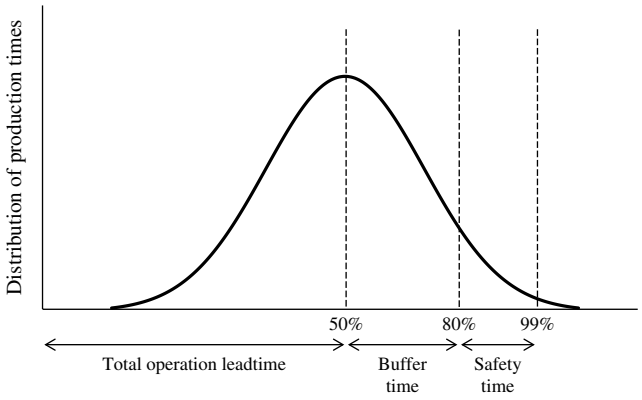
Figure 3 (Color online) Internal and External Milestone Cycle Times



buffer time for operation $n \in \{1, 2, \ldots, N\}$, respectively. The milestone's internal cycle time is calculated as $\sum_{n=1}^{N}(t_n + b_n)$. Hence, the internal cycle time includes the operations' leadtimes and the buffer times within the operations. According to ASML's strategy, every milestone's internal cycle time should ensure 80% delivery reliability. The difference between the internal and external cycle times is the additional time needed to cope with possible disturbances. This additional time is referred to as *safety time*. Let S represent the safety time of a milestone with N operations. Then the external cycle time is calculated as $S + \sum_{n=1}^{N}(t_n + b_n)$. (Refer to Figure 3.) ASML's strategy suggests that after adding the safety time, each milestone should reach 99% delivery reliability. Note that the buffer and safety time are not scientific but ASML notions. In the literature safety time equals $S + \sum_{n=1}^{N} b_n$.

The current method of calculating the buffer and safety times of a module at ASML relies on the data of the last M production instances. After removing outliers, the production times are plotted on a histogram and a Normal distribution is fitted. The mean of the distribution is assumed to represent the total operation leadtime of all the operations, i.e., $\sum_{n=1}^{N} t_n$. The time between the 50th percentile and the 80th percentile is set as the total buffer time for all the operations, i.e., $\sum_{n=1}^{N} b_n$, and the time between the 80th

Figure 4 Calculation of Milestone Buffer and Safety Times



percentile and the 99th percentile is set as the safety time, i.e., S (Figure 4).

The *planned leadtime* of a milestone is set to the milestone's external cycle time. The planned leadtimes are communicated to the responsible work centers. More detailed plans regarding the allocation of the total buffer time to operations are developed *manually* by experienced planners at the work centers, and allocation does not rely on a structured methodology.

Clearly, the current method for determining the planned leadtimes poses some problems. Most importantly, the current method does not take the interrelated nature of module production processes into account. As suppliers for the final assembly, the module production processes need synchronized on-time deliveries. Setting the right module leadtimes prevents delayed start of the final assembly, while setting the right final assembly and test leadtimes is crucial for meeting promised due dates. Delayed deliveries result in high penalty costs and loss of future demands. This is why ASML is interested in a more integrated approach for determining the milestone leadtimes.

2. Formulation of the Model

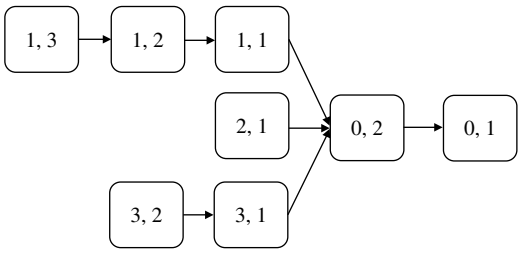
We consider an assembly system consisting of M multistage processes delivering subassemblies to a multistage final assembly process whose index is 0.[1] Each process m, $m \in \{0, 1, \ldots, M\}$, consists of N_m stages. We refer to process m's stage j as stage (m, j). We number the stages in all the processes in decreasing order. This means that stage $(m, j+1)$ is the predecessor of stage (m, j). An example is given in Figure 5.

We define τ_{mj} as the throughput time of stage (m, j).[2] Throughput times at each stage are ran-

[1] In terms of ASML terminology, stage and subassembly correspond to milestone and module, respectively.

[2] If not stated otherwise, all the definitions are valid for all $j \in \{1, 2, \ldots, N_m\}$ and $m \in \{0, 1, \ldots, M\}$.

Figure 5 **An Example with $M = 3$, $N_0 = 2$, $N_1 = 3$, $N_2 = 1$, and $N_3 = 2$**



dom variables with known continuous distributions. Throughput times at different stages are independent.

The system incurs a marginal holding cost h_{mj} from the moment the production at stage (m, j) starts until the final product is delivered to the customer. Typically the holding cost, in the setting discussed here, consists of labor and material costs. We define H_{mj} as the local holding cost per unit time at stage (m, j). For all $m \in \{1, 2, \ldots, M\}$ and $j \in \{1, 2, \ldots, N_m\}$, we have $H_{mj} = \sum_{i=j}^{N_m} h_{mi}$, and for $m = 0$ and $j \in \{1, 2, \ldots, N_0\}$, we have $H_{0j} = \sum_{m=1}^{M} \sum_{i=1}^{N_m} h_{mi} + \sum_{i=j}^{N_0} h_{0i}$. We assume that $H_{mj+1} < H_{mj}$, $\forall\, m \in \{0, 1, \ldots, M\}$. This implies that within each process, as we move from one stage to another, value is added and, therefore, the local unit holding costs increase. We also assume that $\sum_{m=1}^{M} H_{m1} < H_{0N_0}$. This ensures that the local unit holding cost of the first final assembly stage is more than the sum of the local unit holding costs of the subassemblies' final stages. In fact, these two constraints imply that all the marginal holding costs are positive. In addition to the holding costs, the system incurs a penalty cost p per unit time late for delivery to the customer.

Without loss of generality, we set the planned start time of the final assembly process to 0. The objective is to determine the planned leadtimes, T_{mj}, for all the stages such that the sum of expected holding and penalty costs are minimized. These times can be used for determining the planned start times at all the stages.

If the production at stage $(m, j+1)$ finishes earlier than the planned start time at stage (m, j), production at stage (m, j) starts exactly at the planned time. Otherwise, stage $(m, j+1)$ is tardy, in which case the production at stage (m, j) is delayed and starts as soon as the production at stage $(m, j+1)$ is completed. The assumption of holding back production when the previous stage finishes early is quite common in the literature (Yano 1987a, Axsäter 2005) and is also consistent with ASML's manufacturing plan. The main motivation behind the holding-back policy is that the holding costs increase as more value is added to the products. Therefore, cost effectiveness calls for waiting until the planned start time and holding inventory of less costly products, instead of moving forward and running the risk of holding expensive products (Kanet and Christy 1984). At ASML the holding-back policy is followed since the operations within different stages (milestones) are executed by different workforces and every workforce acts according to the plan communicated to them well in advance. Changing the plan would imply extensive rescheduling.

2.1. Earliness and Tardiness Expressions

In this section, we derive the process equations that relate the earliness and tardiness of adjacent stages. For this purpose, we define the following state variables:

F_{mj} = actual finish time of stage (m, j),
W_{-m} = waiting time of process m until the start of process 0 because of the tardiness of other processes,
W_0 = waiting time of process 0 because of the tardiness of processes 1 to M,
E_{mj} = earliness of stage (m, j),
L_{mj} = tardiness of stage (m, j).

Let T_{mj} be the planned leadtime of stage (m, j) and $x^+ = \max\{0, x\}$. Given that the planned start time of the final assembly is 0, the planned production finish time at stage (m, j), $\forall\, j \in \{1, 2, \ldots, N_m\}$ and $\forall\, m \in \{1, 2 \ldots, M\}$, is $-\sum_{i=1}^{j-1} T_{mi}$. Therefore, the earliness and tardiness of stage (m, j) are calculated as $E_{mj} = (-\sum_{i=1}^{j-1} T_{mi} - F_{mj})^+$ and $L_{mj} = (F_{mj} + \sum_{i=1}^{j-1} T_{mi})^+$. The planned finish time at stage $(0, j)$, $\forall\, j \in \{1, 2, \ldots, N_0\}$, is $\sum_{i=j}^{N_0} T_{0i}$. The earliness and tardiness of this stage are calculated as $E_{0j} = (\sum_{i=j}^{N_0} T_{0i} - F_{0j})^+$ and $L_{0j} = (F_{0j} - \sum_{i=j}^{N_0} T_{0i})^+$.

For the final assembly process to start, the production of all subassemblies should be completed. If all the subassemblies are on time or early, the final assembly process starts at time 0. Otherwise, the start time is delayed by an amount equal to the maximum tardiness of the subassembly processes. Therefore, we have $W_0 = \max_{1 \le m \le M} L_{m1}$. The waiting time of subassembly m because of the tardiness of other subassemblies is $W_{-m} = (\max_{n \in \{1,2,\ldots,N_0\}, n \ne m} L_{n1} - L_{m1})^+$.

We now introduce alternative expressions for the earliness and tardiness at each stage to provide recursive expressions that relate the values of these performance measures for consecutive stages.

The instantaneous delivery of raw materials to subassembly processes implies that these processes can always start at the planned time, which is $-\sum_{i=1}^{N_m} T_{mi}$ for subassembly m. Therefore, the earliness and tardiness of the first stage, i.e., (m, N_m), depend only on the planned leadtime and the throughput time of that stage. We have $E_{mN_m} = (T_{mN_m} - \tau_{mN_m})^+$ and $L_{mN_m} = (\tau_{mN_m} - T_{mN_m})^+$.

On the other hand, the earliness and tardiness of the subsequent production stage, i.e., $(m, N_m - 1)$,

also depend on the tardiness of stage (m, N_m). Since production is held back if the previous stage finishes earlier than planned, the earliness of stage (m, N_m) does not affect the earliness or the tardiness of $(m, N_m - 1)$. Therefore, L_{mN_m}, E_{mN_m-1} and L_{mN_m-1} are related as $E_{mN_m-1} = (T_{mN_m-1} - \tau_{mN_m-1} - L_{mN_m})^+$ and $L_{mN_m-1} = (L_{mN_m} + \tau_{mN_m-1} - T_{mN_m-1})^+$.

Based on the same intuition, we can write recursive expressions to relate the earliness and tardiness of the stages in the subassembly processes. Given that $L_{mN_m+1} = 0$, $\forall\, j \in \{1, 2, \ldots, N_m\}$ and $\forall\, m \in \{1, 2, \ldots, M\}$, we have $E_{mj} = (T_{mj} - \tau_{mj} - L_{mj+1})^+$ and $L_{mj} = (L_{mj+1} + \tau_{mj} - T_{mj})^+$.

As discussed above, the start of the final assembly process is delayed for W_0 time units. Therefore, for the first stage of the final assembly process, i.e., stage $(0, N_0)$, we have $E_{0N_0} = (T_{0N_0} - \tau_{0N_0} - W_0)^+$ and $L_{0N_0} = (W_0 + \tau_{0N_0} - T_{0N_0})^+$. For $(0, j)$, $\forall\, j \in \{1, 2, \ldots, N_0 - 1\}$, the recursive expressions for the earliness and tardiness are $E_{0j} = (T_{0j} - \tau_{0j} - L_{0j+1})^+$ and $L_{0j} = (L_{0j+1} + \tau_{0j} - T_{0j})^+$.

Given that the throughput times at all stages are random variables, E_{mj} and L_{mj}, $\forall\, j \in \{1, 2, \ldots, N_m\}$ and $\forall\, m \in \{0, 1, \ldots, M\}$, are random variables as well. Let $\mathbb{E}[\,\cdot\,]$ denote the expectation of a random variable.

2.2. Cost Expression

Next, we write the total expected cost of the system. Let $\mathbf{T}$ be the vector of all planned leadtimes, i.e., $\mathbf{T} = \{T_{mj}\}_{j=1,\,m=0}^{N_m,\,M}$, and $C(\mathbf{T})$ the total expected cost of the system as a function of $\mathbf{T}$. From the moment production at stage (m, j) starts until the final product is delivered to the customer, the system incurs a marginal holding cost h_{mj}. The product is delivered to the customer at time $\sum_{i=1}^{N_0} T_{0i} + L_{01}$ and production starts at stage (m, j) at time $F_{mj+1} + E_{mj+1}$. In addition, a penalty cost p is charged per unit tardy to the customer. The sum of the expected holding and penalty costs is

$$C(\mathbf{T}) = \mathbb{E}\Bigg[\sum_{m=0}^{M}\sum_{j=1}^{N_m} h_{mj}\Bigg(\sum_{i=1}^{N_0} T_{0i} + L_{01} - (F_{mj+1} + E_{mj+1})\Bigg) + pL_{01}\Bigg]. \quad (1)$$

We can write the production start time at stage (m, j) as the sum of the start time of process m and the time elapsed until the production at stage (m, j) starts. For subassembly process m, the start time is $-\sum_{i=1}^{N_m} T_{mi}$ and the final assembly starts at time W_0. The elapsed time from the start of these processes until the production at stage (m, j) starts is $\sum_{i=j+1}^{N_m}(\tau_{mi} + E_{mi})$. Therefore, we can express the production start times as follows:

$$F_{mj+1} + E_{mj+1} = -\sum_{i=1}^{N_m} T_{mi} + \sum_{i=j+1}^{N_m}(\tau_{mi} + E_{mi}),$$

$$m \in \{1, 2, \ldots, M\},\ j \in \{1, 2, \ldots, N_m\}$$

$$F_{0j+1} + E_{0j+1} = W_0 + \sum_{i=j+1}^{N_0}(\tau_{0i} + E_{0i}), \quad j \in \{1, 2, \ldots, N_0\}.$$

Using these expressions and the equality $H_{01} = \sum_{m=0}^{M}\sum_{j=1}^{N_m} h_{mj}$, we can rewrite $C(\mathbf{T})$ as follows:

$$C(\mathbf{T}) = \mathbb{E}\Bigg[\sum_{m=1}^{M}\sum_{j=1}^{N_m} h_{mj}\Bigg(\sum_{i=1}^{N_0} T_{0i} + \sum_{i=1}^{N_m} T_{mi} - \sum_{i=j+1}^{N_m}(\tau_{mi} + E_{mi})\Bigg) + \sum_{j=1}^{N_0} h_{0j}\Bigg(\sum_{i=1}^{N_0} T_{0i} - W_0 - \sum_{i=j+1}^{N_0}(\tau_{0i} + E_{0i})\Bigg) + (H_{01} + p)L_{01}\Bigg]. \quad (2)$$

Our objective is to find the vector $\mathbf{T}$ that minimizes the total expected cost $C(\mathbf{T})$. As acknowledged by other authors who have studied similar problems, this problem cannot be solved to optimality since the cost function is not well-behaved. In the next section, we propose a solution approach that generates quite accurate results.

3. Solution Approach

The main idea behind our approach is to decompose the assembly system into M serial systems, one for each subassembly. For that purpose, we write the cost expression from serial system m's perspective. We use the following equality, which states the time elapsed from the start of subassembly process m until the start of the final assembly process in two alternative ways.

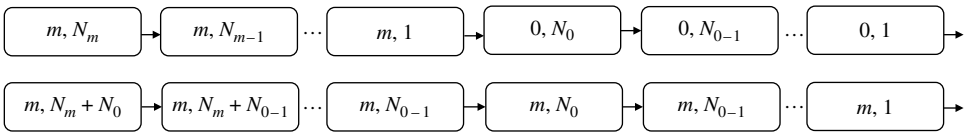
$$\sum_{i=1}^{N_m} T_{mi} + W_0 = \sum_{i=1}^{N_m}(\tau_{mi} + E_{mi}) + W_{-m}$$

This equality implies that the maximum tardiness, i.e., W_0, equals $\sum_{i=1}^{N_m}(\tau_{mi} + E_{mi}) + W_{-m} - \sum_{i=1}^{N_m} T_{mi}$. Substituting W_0 in (2), we rewrite $C(\mathbf{T})$ from serial system m's perspective as follows:

$$C(\mathbf{T}) = \mathbb{E}\Bigg[\sum_{\substack{n=1\\ n\neq m}}^{M}\sum_{j=1}^{N_n} h_{nj}\Bigg(\sum_{i=1}^{N_0} T_{0i} + \sum_{i=1}^{N_n} T_{ni} - \sum_{i=j+1}^{N_n}(\tau_{ni} + E_{ni})\Bigg) + \sum_{j=1}^{N_m} h_{mj}\Bigg(\sum_{i=1}^{N_0} T_{0i} + \sum_{i=1}^{N_m} T_{mi} - \sum_{i=j+1}^{N_m}(\tau_{mi} + E_{mi})\Bigg) + \sum_{j=1}^{N_0} h_{0j}\Bigg(\sum_{i=1}^{N_0} T_{0i} - \Bigg(\sum_{i=1}^{N_m}(\tau_{mi} + E_{mi}) + W_{-m} - \sum_{i=1}^{N_m} T_{mi}\Bigg) - \sum_{i=j+1}^{N_0}(\tau_{0i} + E_{0i})\Bigg) + (H_{01} + p)L_{01}\Bigg]. \quad (3)$$

Serial system m, $\forall\, m \in \{1, 2, \ldots, M\}$, has $N_m + N_0$ stages. We renumber the stages so that stage (m, j)

Figure 6 **Renumbering the Stages**



becomes $(m, j + N_0)$, $\forall j \in \{1, 2, \ldots, N_m\}$ and stage $(0, j)$ becomes stage (m, j), $\forall j \in \{1, 2, \ldots, N_0\}$. Therefore, serial system m starts with stage $(m, N_m + N_0)$ and ends with stage $(m, 1)$. The top and bottom pictures in Figure 6 represent the old and new numbering, respectively.

We retain the recursive expressions for the earliness and tardiness with a slight modification at stage (m, N_0), which was originally the first stage of the final assembly process, i.e., stage $(0, N_0)$. The production start time at this stage is related to the tardiness of the other subassemblies. We link serial system m with the other serial systems by considering the tardiness caused by the other subassemblies. This link is reflected by the random variable W_{-m}. Given that $L_{m(N_m+N_0+1)}$ is 0, the earliness and tardiness expressions for the stages in serial system m are summarized in Table 1.

Using the new numbering for the stages in serial system m, we can combine the second and third lines of the cost function (3) into one. The resulting cost expression is as follows:

$$C(\mathbf{T}) = \mathbb{E}\Bigg[\sum_{\substack{n=1\\n\neq m}}^{M}\sum_{j=1}^{N_n} h_{nj}\Bigg(\sum_{i=1}^{N_0} T_{0i} + \sum_{i=1}^{N_n} T_{ni} - \sum_{i=j+1}^{N_n}(\tau_{ni} + E_{ni})\Bigg) + \sum_{j=1}^{N_m+N_0} h_{mj}\Bigg(\sum_{i=1}^{N_m+N_0} T_{mi} - \sum_{i=j+1}^{N_m+N_0}(\tau_{mi} + E_{mi}) - W_{-m}\mathbf{1}_{j\le N_0}\Bigg) + (H_{m1} + \hat{p})L_{m1}\Bigg]. \quad (4)$$

Here, $H_{m1} = \sum_{j=1}^{N_m+N_0} h_{mj}$ and $\hat{p} = p + \sum_{\substack{n=1\\n\neq m}}^{M}\sum_{j=1}^{N_n} h_{nj}$. In addition, $\mathbf{1}_{j\le N_0}$ is the indicator function, which equals 1 if $j \le N_0$ and 0 otherwise.

Note that the recursion in Table 1 resembles the well-known Clark-Scarf backorder recursion. In addition, the cost expression for serial system m has a similar form as the cost for serial inventory systems (refer to Zipkin 2000). Here, backorders correspond to the tardiness and the demand random variable is replaced by the throughput time. The only difference is that stage (m, j), $\forall j \in \{1, 2, \ldots, N_0\}$, has an extra random variable W_{-m}. In fact, the coupling between the serial systems is due to the random variable W_{-m}, and our decomposition idea is based on the observation that from serial system m's point of view, W_{-m} is exogenous. The exact correspondence between the *serial* planned leadtime problem and the *serial* multistage inventory optimization problem is shown by Gong et al. (1994).

For divergent multiechelon inventory systems, Diks and de Kok (1998) show that we can find cost-optimal base-stock policies by recursively solving so-called generalized Newsvendor equations. Obviously, this implies the same for the serial systems in Clark and Scarf (1960) and the assembly systems in Rosling (1989). It is not obvious whether such a result holds for the planned leadtime problem discussed in this paper, which is more complicated because of the mutual impact of the multistage assembly processes on each other through W_{-m}. Unlike the situation discussed by Rosling (1989) with constant leadtimes, we cannot coordinate the system such that an equivalent serial system emerges. Fundamentally, the exogenous stochastic throughput times do not allow for that. Despite this fact, our exploratory research suggests that we can formulate generalized Newsvendor equations from which the planned leadtimes can be determined. For a two-echelon assembly system consisting of multiple parallel single-stage processes feeding a single-stage process, we formulate this statement as a theorem and prove it in the appendix. In addition, we prove that the leadtimes determined through Newsvendor equations are unique. Additional numerical experimentation for a large set of multistage systems, part of which we present in the subsequent sections, supports our conjecture.

For the sake of simplicity we identify each stage (m, j) with a unique number s. We associate 0 with stage $(0, 1)$. Thus, we consider a convergent assembly system with, say, S stages. With each stage s we

Table 1 **Recursive Expressions for Earliness and Tardiness for Serial System m**

Stage(s)	Earliness	Tardiness
$j \in \{1, 2, \ldots, N_0 - 1, N_0 + 1, \ldots, N_m + N_0\}$	$(T_{mj} - \tau_{mj} - L_{mj+1})^+$	$(L_{mj+1} + \tau_{mj} - T_{mj})^+$
(m, N_0)	$(T_{mN_0} - \tau_{mN_0} - (L_{mN_0+1} + W_{-m}))^+$	$((L_{mN_0+1} + W_{-m}) + \tau_{mN_0} - T_{mN_0})^+$

can associate a random variable W_{-s} which is the delay after stage s caused by other predecessors of stage s's immediate successor. Note that, in our system, we have $W_{-s} \equiv 0$, except for the last stages of the subassembly processes, i.e., stage $(m, 1)$, $\forall m \in \{1, 2, \ldots, M\}$. In general convergent assembly systems, each stage s whose immediate successor has more than one predecessor has a nonzero W_{-s}. To formulate our conjecture, we introduce a set of events A_s. An element w of the probability space is an element of A_s if and only if under w, (i) stage s starts in time, (ii) if the immediate successor of stage s starts in time then the final stage finishes in time, and (iii) the tardiness of stage s exceeds W_{-s}, and the final stage, stage 0, is late. Informally, A_s represents the set of all events in which stage s is to be blamed for the tardiness of the system. It is easy to see that the sets A_s, $\forall s \in \{1, 2, \ldots, S\}$, are mutually exclusive and the union of these events consists of all w in the probability space for which the final stage is late. The generalized Newsvendor equation conjecture is as follows:

Conjecture 1. *The optimal planned leadtimes policy satisfies*

$$P(A_s) = \frac{h_s}{p + H_0}, \quad \forall s \in \{1, 2, \ldots, S\}. \tag{5}$$

This conjecture is the most important driver for the accuracy and speed of the heuristic. An immediate consequence of Conjecture 1 is that the probability of on time delivery equals the classical Newsvendor fractile $p/(p + H_0)$ ($p/(p + H_{01})$ with the original notation). The definition of A_s involves mutually dependent events, implying, at first sight, great computational complexity. Yet the introduction of the random variables W_{-s} enables the derivation of tractable expressions. We compute these expressions using so-called two-moment approximations. The key observation made earlier is that assuming the W_{-s} are known, the system to be considered for computation of $P(A_s)$ is a serial system, consisting of the stages on the path starting from stage s until stage 0. From this we can also see that *given* W_{-s} the optimal planned leadtimes can be computed recursively starting with the computation of T_0 (originally T_{01}) from the equation $P(A_0) = h_0/(p + H_0)$. To compute $P(A_s)$ we associate with each stage s a subsystem B_s of the original serial system, i.e., the system starting with stage s. Denote the immediate successor of stage s as $s - 1$. Remember that L_s is the tardiness of stage s. Then we have the following equations:

$$\begin{aligned}
P(A_s) &= P(\text{final stage of system } B_{s-1} \text{ is in time and } L_s \geq W_{-s} \text{ and final stage of system } B_s \text{ is late}),\\
&= P(\text{final stage of system } B_{s-1} \text{ is in time and } L_s \geq W_{-s}) - P(\text{final stage of system } B_{s-1} \text{ is in time and } L_s \geq W_{-s} \text{ and final stage of system } B_s \text{ is in time}),\\
&= P(\text{final stage of system } B_{s-1} \text{ is in time}) \cdot P(L_s \geq W_{-s}) - P(L_s \geq W_{-s} \text{ and final stage of system } B_s \text{ is in time}),\\
&= P(\text{final stage of system } B_{s-1} \text{ is in time}) \cdot P(L_s \geq W_{-s}) - P(\text{final stage of system } B_s \text{ with conditional tardiness } L_s \text{ given } L_s \geq W_{-s} \text{ is in time}) \cdot P(L_s \geq W_{-s}).
\end{aligned}$$

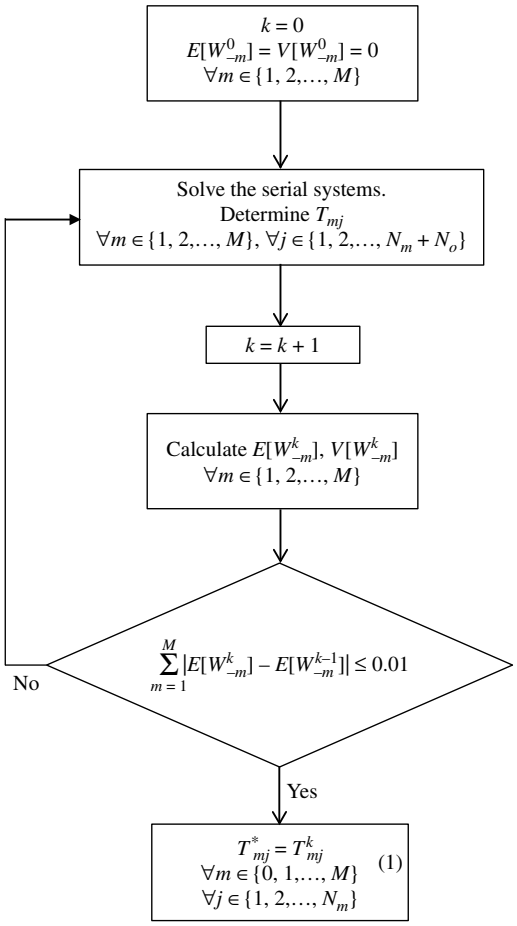
In the above derivation we used the independence of (i) the events in system B_{s-1} and (ii) L_s and W_{-s}. Thus, the probabilities $P(A_s)$ can be computed from the expressions for serial subsystems with exogenous delays W_{-s}. However, we do not know the random variables W_{-s}. To deal with this we propose an iterative approach, which starts with $W_{-s} \equiv 0$, $\forall s \in \{1, 2, \ldots, S\}$. As the exogenous delays are 0, the *optimal* planned leadtimes are *minimal*. With these minimal planned leadtimes we compute the associated W_{-s}, which should be *maximal*. With the maximal W_{-s} we compute *maximal optimal* planned leadtimes. By repeated application of this scheme we get an ordered series of W_{-s} that approaches some limit from above and below. Although our reasoning is not much more than intuitive, it is supported by numerical evidence.

The above reasoning enables us to develop an iterative heuristic procedure. In this heuristic, we rely on two-moment fits. Therefore, we do not even need the exact distribution of W_{-m} but only the first two moments. The details are explained in the next section.

3.1. Iterative Heuristic Procedure

In our iterative solution approach, we initially set the values of W_{-m}, $\forall m \in \{1, 2, \ldots, M\}$ to 0. We then rely on Conjecture 1 to determine the values of the decision variables, i.e., T_{mj}, for all the stages in all the serial systems. We apply the proven technology of two-moment mixed Erlang fits to approximate the distributions of $\tau_j + L_{j+1}$, $\forall j \in \{1, 2, \ldots, N_m + N_0\} - N_0$, and $\tau_{N_0} - L_{N_0+1} - W_{-m}$, i.e., the sum of tardiness and throughput times for all the stages (Diks and de Kok 1999). Next, we consider the original system, and by setting the planned leadtimes to the values obtained from the decomposed system, we determine the first two moments of W_{-m}, $\forall m \in \{1, 2, \ldots, M\}$. For this purpose we use two-moment approximations based on the recursive determination of $\max_{1 \leq n \leq M, n \neq m} L_{n1}$ from the maximum of two random variables (Whitt 1982). Using these moments and relying on the two-moment mixed Erlang fits to approximate the distributions of the sum of W_{-m}, tardiness and throughput

Figure 7 Flowchart Representation of the Iterative Heuristic Procedure



(1)Here we use the original notation for the stages.

times, we go back to the serial systems and recompute the corresponding values of the decision variables. Therefore, at each iteration, given the first two moments of W_{-m}, we solve the serial systems for the planned leadtimes. Using this solution, we recompute the first two moments of W_{-m}. Let $\mathbb{E}[W^k_{-m}]$ be the expected values of W_{-m} after iteration k. The algorithm stops when $\sum_{m=1}^{M} |\mathbb{E}[W^k_{-m}] - \mathbb{E}[W^{k-1}_m]| \le 0.01$. Refer to Figure 7 for a flow chart of the iterative heuristic procedure. Our numerical experiments show that by setting the initial value of W_{-m} to 0, the stopping criterion is satisfied, on average, in 5.32 iterations. As mentioned above, if $\mathbb{E}[W^0_{-m}] = 0$, $\forall m \in \{1, 2, \ldots, M\}$, the next iteration results in the maximum possible values of $\mathbb{E}[W_{-m}]$. In every iteration, the expectations get closer and, eventually, the stopping criterion is satisfied.

4. Numerical Experiments

To develop some intuition about the solution structure and performance of the iterative heuristic and to compare it with ASML's fractile method, we solve a set of test problems. We resort to a numerical optimization technique as a benchmark. We use the DFP method as described in Press et al. (2007). This method is one of the earliest and *most effective* quasi-Newton methods (Hashamdar and Ibrahim 2010). It simultaneously generates conjugate directions and constructs approximations of the inverse Hessian matrix, providing monotonically improving approximations to the exact solution. To test the validity of our results, we also built a simulation model that uses the planned leadtimes, the distributions of throughput times, and the cost parameters as inputs. The outputs are the service levels and the average costs.

Using ASML's SAP (systems applications and products) system, we gathered actual leadtime data for the milestones of Systems A and B. Our cycle time data for Systems A and B consist of 128 and 11 samples, respectively.[3] In addition, we were provided with the data on the value added by each milestone (echelon holding costs). In our numerical analysis we initially fix a fractile and use ASML's method to calculate the planned leadtimes. We then simulate the system with these planned leadtimes to determine the corresponding on-time delivery performance. Next, we exploit the relationship between the Newsvendor fractile and the calculated on-time delivery probability to estimate the penalty cost associated with late deliveries. Finally, we use our iterative heuristic procedure and the DFP method to calculate the planned leadtimes. We simulate both results to test their validity.

One might argue that the amount of data for System B is not sufficient for validation. However, our model validation does not concern standard statistical modeling, where the model is estimated from the data. We use the actual leadtime data to determine the sample average and standard deviation. Then our model computes the service level and compares this to the observed service level for the same set of systems. Given the mathematical complexity of the model, it would be very unlikely to find accurate model estimates for both systems unless the model is a good description of reality. In fact, our service level estimates for both systems are within 2% of the actual values.

For each fractile value we report the percentage cycle time reduction, CT_{red}, and the percentage cost reduction, TC_{red}, ASML would achieve if our heuristic were used instead of the fractile method. Table 2 summarizes our results. For both systems the 85-percentile corresponds to the actual values for which we obtained good validation results. The heuristic results in significant reductions in the cycle times and total costs, especially if ASML targets high on-time probability.

[3] Because of confidentiality reasons, we cannot report the data.

Table 2 **Comparison of ASML's Fractile Method and Iterative Heuristic Procedure**

	System A		System B	
Percentile	CT_{red}	TC_{red}	CT_{red}	TC_{red}
80	9.18	12.53	8.52	17.73
85	10.63	14.94	10.08	20.46
90	12.68	17.91	12.13	23.52
95	15.85	21.91	14.87	27.10
99	21.17	27.81	19.93	32.22

The average percentage cycle time and cost differences between the DFP method and the heuristic are 1.45% and 1.33%, respectively. The heuristic terminates, on average, in 5.1 iterations. We validate our analytical cost and on-time delivery probability figures by simulating the systems with the planned leadtimes suggested by all the methods (ASML's method, heuristic, DFP). The average and standard deviation of the percentage difference between the costs (service levels) calculated analytically and by simulation are 0.31% and 0.15% (0.83% and 0.65%), respectively.

5. Detailed ASML Analysis

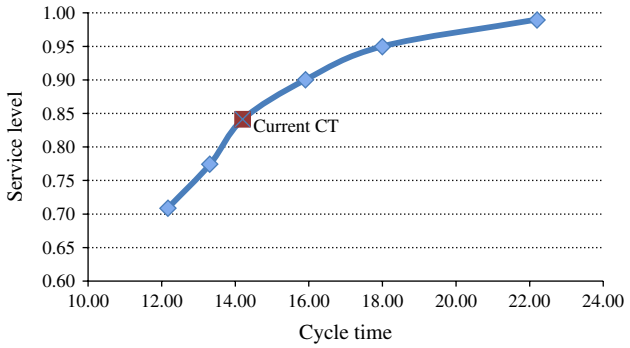
The iterative heuristic was developed with the intention of solving ASML's planned leadtime problem. Given the complexity of the analysis of the assemble-to-order planned leadtime model, it is impractical to explain the mathematical principles to the people responsible for setting the planned leadtimes. This is why, to gain credibility at ASML, we empirically validated the model for the most complex module and Systems A and B and, to our knowledge, we report the first results on the empirical validity of the planned leadtime model. In this section, we explain our findings on ASML's current practice of planning the leadtimes and provide the details of our validation. In addition, we explain the benefits of implementing the suggested solutions.

5.1. Analysis of the Most Complex Module

Using ASML's SAP system, we gathered 112 *actual* leadtime data for two milestones of the most complex module. In addition, we gathered data for the *planned* leadtimes of these milestones. The planned leadtimes in the SAP system differs considerably from those suggested by the fractile procedure. This implies that employees on the shop floor do not follow the recommendations based on the fractile method. Therefore, the fractile method does not provide a relevant benchmark for this module.

To compare ASML's current performance with the solution suggested by the iterative heuristic procedure, we construct an efficient frontier based on the relationship between the cycle time, i.e., the total planned leadtime, and the service level (Figure 8).

Figure 8 **(Color online) Efficient Frontier-Service Level vs. Cycle Time**



To our surprise, we found that the planned leadtime data in the SAP system, as decided by people in the most complex module's work center, yields a solution on the efficient frontier generated using our iterative heuristic, indicating that our solutions are in line with the understanding of the people in the work center. In Table 3, we summarize the planned cycle times (CT), on-time delivery percentages, holding costs, and total costs for the solutions from ASML's SAP system, the DFP method, and the Newsvendor (NV) method, i.e., the iterative heuristic. We present results both from discrete event simulation (Sim) and our analytical procedures (Anly).

The planned leadtimes solution from SAP is considered as the benchmark solution. We normalize the holding costs of the benchmark solution at 100%. The DFP and NV solutions are remarkably close to the solution found by the ASML operators. A detailed analysis on the planned leadtimes of the milestones suggest that for the first milestone our method's planned leadtime (3.4 days) is 10.5% shorter than the current ASML plan (3.8 days). On the other hand, for the second milestone our method's planned leadtime (10.5 days) is 1.9% longer than the current ASML plan (10.3 days).

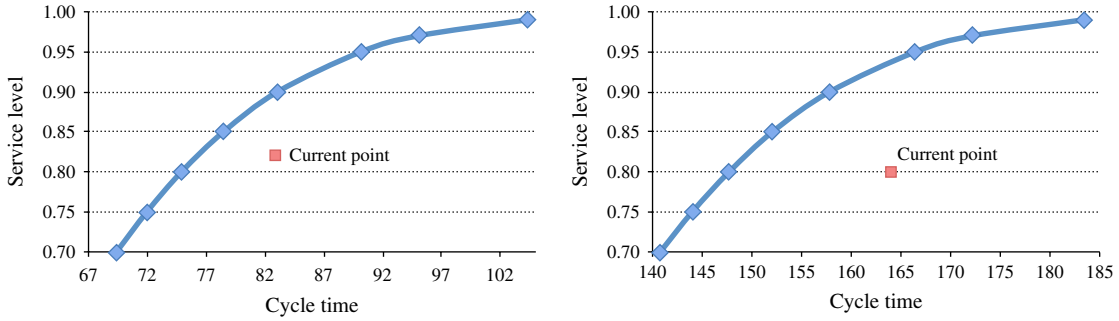
5.2. Optimization of the System Planned Leadtimes

The SAP data for Systems A and B indicate that the fractile chosen for determining safety leadtimes is 85%, resulting in actual service levels of only 0.82

Table 3 **Performance of Alternative Solution Procedures for the Most Complex Module**

		On-time delivery %		Holding cost (%)		Total cost (%)	
Case	CT (*days*)	Sim	Anly	Sim	Anly	Sim	Anly
ASML	14.1	82.1	82.7	100.0	100.4	148.4	149.2
DFP	14.2	82.1	82.8	99.7	99.3	148.1	147.3
NV	13.9	81.4	82.1	95.6	96.0	147.5	149.1

Figure 9 **(Color online) Efficient Frontier (Cycle Time vs. Service Level)**



(a) System A (b) System B

and 0.81, with associated overall planned leadtimes of 86 days and 165 days, respectively. For both systems, we use our heuristic results to draw the efficient frontiers in Figure 9.

Figure 9 suggests that unlike in the most complex module, the current cycle times for Systems A or B are not on the efficient frontiers. Therefore, using our Newsvendor method to calculate the planned leadtimes provides an attractive alternative to the fractile method.

Table 4 presents the results of the alternative solution procedure using the same notation as in Table 3. As the ASML solution yields a simulated on-time delivery percentage of 82.3%, we set the penalty costs accordingly using the Newsvendor fractile. In this case, we normalize the DFP solution's holding costs for earliness at 100%, as we find that the DFP solution yields the lowest overall cost. For all performance data presented we find that our analytical expressions perform quite well. The Newsvendor solution, is quite close to the DFP solution in performance and in overall planned throughput time, yielding a cycle time reduction of 11.8% over the current planned leadtimes. Typically, cycle time reduction is targeted as a means to reduce inventory capital tied up in work in progress (*WIP*). We find that *WIP* can be reduced by about 7%. This represents a substantial addition to ASML's current profits.

Goldratt (1997) claims that the safety buffer in project networks with uncertain activity durations should be positioned at the end of the project. The conceptual approach discussed in Goldratt (1997) is termed Critical Chain. Noting the similarity between project networks and the ASML system assembly, we explore the validity of this claim. We start from the Newsvendor solution and assume all module planned leadtimes remain unchanged, but the planned leadtimes for the final assembly and testing milestones are set equal to zero. We then modify the last stage's safety buffer such that the 82.3% on time delivery target is met. We find that the resulting solution differs very little in performance from the Newsvendor solution, which shows robustness of the optimal solution. This also follows from the fact that the DFP and Newsvendor solutions have a relatively large buffer at the last milestone, representing about 24% of the overall cycle time. The Critical Chain (CC) solution for the final assembly and testing phases has a safety time buffer at the last stage representing 52% of the overall cycle time.

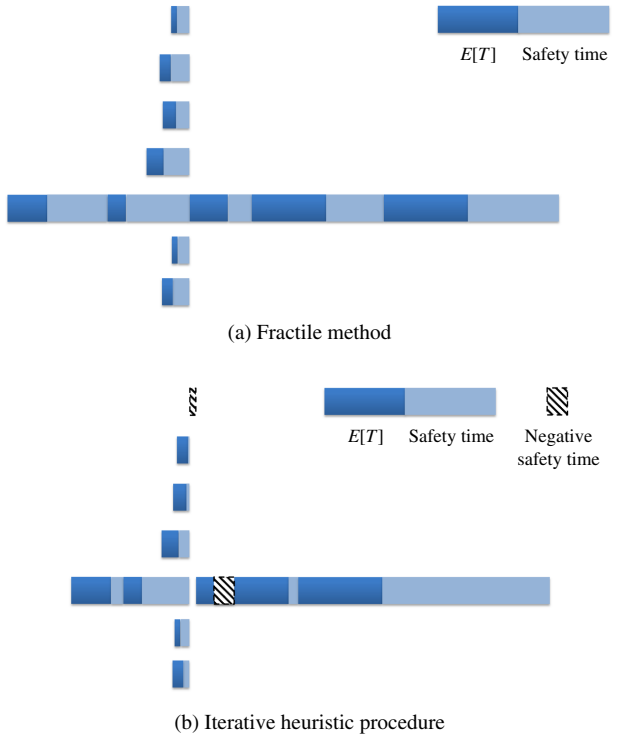
Taking the Critical Chain concept to the extreme, we also consider the solution with planned leadtimes equal to zero for all milestones, but the last one. We find a serious deterioration of the performance, as we lose 5% of our 7% reduction in capital tied up in work in progress. This shows the importance of coordinating the module assembly processes and a proper trade-off between earliness and tardiness costs.

Based on these results, ASML planners are convinced that our Newsvendor method can reduce the overall System A cycle time. The next question they have is related to the detailed plan: Are there any

Table 4 **Performance of Alternative Solution Procedures for System A**

			On-time delivery %		Holding cost (%)		Total cost (%)		Work in progress (%)	
Case	CT (*days*)	CT_{red} (%)	Sim	Anly	Sim	Anly	Sim	Anly	Sim	Anly
ASML	85.8	0	82.3	82.6	130.9	130.4	168.9	166.5	0.0	0.0
DFP	75.7	12	82.7	82.4	100.0	99.8	142.0	142.0	7.3	7.3
NV	76.5	11	82.7	82.3	100.7	100.5	142.8	143.3	7.1	7.1
Critical chain from final	76.2	11	82.8	81.8	100.7	100.0	142.8	140.4	7.2	7.2
Critical chain for system	68.8	20	82.7	82.1	120.6	120.1	168.6	165.7	2.4	2.5

Figure 10 (Color online) Structure of the Planned Leadtimes



(a) Fractile method

(b) Iterative heuristic procedure

significant differences in the detailed plans when the current method and the proposed method are compared? To answer this question, we show the structure of the planned leadtimes obtained for System A using ASML's method and the Newsvendor method (Figure 10). ASML's method results in positive safety times at all milestones. The total cycle time is dictated by the most complex module (5th module) and the final assembly and test milestones. On the other, the detailed plan generated by our method is quite different. Our method suggests that at the start of the final assembly process, the 1st module has a planned leadtime of zero, thereby becoming the *drum* of the assembly process. When we compare the throughput times and the added values of all the module milestones, we see that the value added per unit throughput time is the highest for the first module. Therefore, our method, which also takes the earliness cost into account, suggests a solution with a planned leadtime of zero for the first module. The other modules are planned such that they have a limited impact on the first module-final assembly interaction. Furthermore, as Figure 10 suggests, our method results in a substantial mean earliness at the last milestone of the final assembly process. This is consistent with both the claims of Goldratt (1997) and the finding for multiechelon systems that the majority of the slack should be at the most downstream stage (de Kok and Fransoo 2003).

In Table 5 we present our results for System B, which was introduced in 2012 and represents the next generation of ASML's lithography systems. This system has substantially longer and more erratic assembly and testing times and thus a substantially longer overall cycle time. We find that the DFP solution does not meet the target on-time delivery of 81.1%, set by the on-time delivery performance of the currently used 85%-fractile solution. We modify the planned leadtime of the last milestone to meet the required performance. This modified solution is used as the benchmark: the earliness costs are normalized at 100%. The cycle time reduction from the modified DFP solution is about 11%. This yields a 9.6% reduction in working capital. The Newsvendor solution yields a capital reduction of 8.4%, while meeting the required on-time delivery without need for modification. For this system Critical Chain performs better than for System A, but still much worse than the Newsvendor solution. Again we observe robustness of the Critical Chain solution for the final assembly and testing milestones. The structure of the planned leadtimes for System B is similar to the one for System A in Figure 10.

Our results show that we can reduce the overall cycle times by 11%. Apart from the impact of cycle time reduction on working capital, in the ASML situation there is another important impact: capacity increase. As explained in §1.2 each system is assembled and tested in its own cleanroom and occupies this cleanroom until delivery to the customer. The number of available cleanrooms limits the *WIP* measured in number of systems. Little's formula tells us that the capacity of ASML in terms of output per week is less than the number of cleanrooms divided by the cycle time in weeks. A cycle time reduction of 11% yields an increase in this capacity limit by 12%. As ASML's demand for lithography systems is volatile, it has gone through periods of high demand, where the number of cleanrooms was the binding constraint. This implies that in times of high demand, the 11% cycle time reduction resulting from our optimization yields an effective output increase of 12%. To give an indication of the financial impact of such an increase, consider that the company sold €5.86 billion worth of systems in 2014. Thus the cycle time reduction from our planned leadtime optimization would yield an additional revenue of €0.70 billion, assuming fully occupied cleanrooms throughout the year. Clearly, this is the best case scenario, but the order of magnitude underlines the contribution of our research in this case.

5.3. Discussion of Implementation

The study outlined in this paper is motivated by the problem of setting the planned leadtimes at ASML. As

Table 5 Performance of Alternative Solution Procedures for System B

			On-time del. %		Holding cost (%)		Total cost (%)		Work in progress (%)	
Case	CT (*days*)	CT_{red} (%)	Sim	Anly	Sim	Anly	Sim	Anly	Sim	Anly
ASML	165.0	0	81.1	80.8	157.5	157.5	192.3	191.9	0.0	0.0
DFP	144.3	13	78.6	79.0	93.5	93.5	146.2	145.7	10.7	10.7
DFP(modified)	146.25	11	81.0	81.4	100.0	100.0	145.6	145.3	9.6	9.6
NV	148.6	10	80.9	81.1	107.5	107.3	153.2	152.5	8.4	8.4
Critical chain from final	148.9	10	81.1	81.1	108.4	108.3	153.5	153.1	8.2	8.2
Critical chain system	141.5	14	81.2	81.2	119.4	119.3	163.5	164.5	6.4	6.4

discussed in §1.2.2, ASML currently relies on a fractile method for setting the milestone leadtimes. This method has been used for many years. There have been multiple initiatives to question its validity but none have resulted in an alternative method. According to ASML's business engineers the current method is problematic because it does not take the interrelated nature of module production processes into account. As suppliers for the final assembly, the module production processes need synchronized on-time deliveries. Our collaboration with ASML on this problem started in 2012.

The research outlined in this paper was conducted together with ASML's business engineers, planners, and a group of operators. After gathering information, we developed the model in §2, which reflects ASML's current manufacturing environment. The results, as explained in §§4 and 5, were presented to the business engineers and senior management.

First of all, we discussed our finding regarding the most complex module. Our analysis revealed that, for this module, ASML's fractile plan is not followed. The planned leadtimes are determined by experienced shop floor workers. ASML's managers do not interfere with the current way this work center is operating since they find the realized leadtimes satisfactory. It is impossible for the workers in the complex module's work center to come up with a model which explains their intuition developed over many years. Our analysis shows that their intuition results in a solution, which is on the efficient frontier generated using our model. This was a major contribution to the acceptance of our model and software tool developed. Now ASML has a model, which is in line with the understanding of very experienced people in the most complex module's work center.

Next, we communicated the main results of our analysis for Systems A and B in the form of the efficient frontiers versus the solutions from the ASML fractile approach. We shared our finding that the overall cycle times of Systems A and B can be reduced. The detailed plans as depicted in Figure 10 are explained to the planners. The detailed milestone modeling showed economic considerations behind the clustering of milestones: some milestones should not have safety buffers, implying that these milestones should be clustered with their successor. The message that overall cycle time could be reduced substantially (11%) while maintaining the same on-time delivery performance or even higher, and support for decisions on milestone clustering led to the decision to use our planned leadtime optimization tool. Another driver of this decision is ASML's interest on the reduction of *WIP* capital.

ASML has had an active cycle time reduction program for years. Satisfied with our promising results, a separate workstream has been added to this program with an overall cycle time reduction target of more than 11%. Toward this objective, our model and the software tool has been used for detailed analysis of the leadtimes in the factories since the second half of 2014. Over the last year, in a pilot run for a limited number of systems, the planned leadtimes have been set as suggested by our model. The described dynamics of our model, together with other initiatives and interpretations, have already resulted in cycle time reductions achieving the target, although it is difficult to specify the reduction percentage that is solely due to our model.

Looking forward, we made a mutual agreement with ASML to continue our collaboration on this topic. Initially, the lessons learned from this study will be translated into a better setting for other systems. This setting is planned to accommodate our suggestions on milestone clustering. As mentioned above, ASML has other initiatives to reduce the cycle times. Before a company-wide implementation, we agreed to adjust/extend our model and software tool to accommodate these initiatives.

6. Conclusions and Future Research Directions

In this paper, we studied an assembly system that consists of a number of parallel multistage processes feeding a multistage final assembly process. Each stage has a stochastic throughput time and is controlled by planned leadtimes. All stages incur a cost for early completion and a lateness cost when the customer order is delivered late. The objective is to determine the planned leadtimes to minimize

the sum of expected earliness and lateness costs. We developed an iterative heuristic that relies on a conjecture related to the generalized Newsvendor functions. Comparing our results with those from a numerical optimization method (Davidon-Fletcher-Powell), we concluded that our heuristic performs well with a percentage cost error of 1.33%. This heuristic was developed with the intention of solving the planned leadtime problem of ASML, the world's leading provider of lithography systems for the semiconductor industry. The data revealed that the most complex module's production was already performed according to our suggestion. However, the same was not true for the systems. Using the planned leadtimes proposed by our heuristic, the system cycle times can be reduced by 10%–11%. We conclude that, if ASML produces according to our proposal, during peak demand periods, on average an 11%–12% increase in the output can be achieved. This implies, for a year like 2014, a year with very high demand, an additional revenue of €0.70 billion.

As a future research direction, our heuristic can be extended to solve for the leadtimes of divergent systems since there are no imbalance issues as there is no allocation problem. It might be possible to use the same approach to obtain optimality equations for general networks under planned leadtime control. Material availability may also be incorporated based on the delay distribution under some material control policy and demand process. The effectiveness of the heuristic also suggests an investigation of whether the Clark and Scarf (1960) decomposition also holds for assembly systems and beyond.

Appendix. Proof of Conjecture 1 for Two-Echelon Assembly Systems

We consider an assembly system consisting of M single-stage processes feeding a single-stage final process whose index is 0. Let $i \in \{0, 1, 2, \ldots, M\}$ be the index of the ith process. Let τ_i be the throughput time of process $i \in \{0, 1, 2, \ldots, M\}$. Assume that $\{\tau_i\}_{i=0}^{M}$ are mutually independent. Let T_i denote the planned leadtime of process $i \in \{0, 1, 2, \ldots, M\}$. We define h_i and H_i as the marginal and local holding costs of process $i \in \{0, 1, 2, \ldots, M\}$ and p as the penalty cost per unit time late for delivery to the customer. Note that $h_i = H_i$, $\forall i \in \{1, 2, \ldots, M\}$ and $\sum_{i=0}^{M} h_i = H_0$. As in §2.1, we define E_i and L_i as the earliness and lateness of process $i \in \{0, 1, 2, \ldots, M\}$. $W_{-i} = (\max_{j\in\{1,2,\ldots,M\},\, j\neq i} L_j - L_i)^+$ is the waiting time of process i until the start of process 0 because of the tardiness of other processes. W_0 is the waiting time of process 0 because of the tardiness of processes 1 to M. We define $\bar{C}(\mathbf{T})$ as the cost of an arbitrary system assembled from process i's perspective. Hence, $C(\mathbf{T}) = \mathbb{E}[\bar{C}(\mathbf{T})]$. Using Equation (3), $\bar{C}(\mathbf{T})$ can be written as

$$\bar{C}(\mathbf{T}) = \sum_{\substack{j=1 \\ j\neq i}}^{M} h_j(T_j + T_0) + h_i(T_i + T_0) + h_0(T_i + T_0 - (\tau_i + E_i) - W_{-i}) + (H_0 + p)L_0. \quad (6)$$

Using the above notation, we rewrite Conjecture 1 for two-echelon assembly systems as a theorem.

THEOREM 2. *The optimal planned leadtimes* $\{T_i^*\}_{i=0}^{M}$ *satisfy the following Newsvendor equations*:

1. $P(\tau_0 > T_0^*) = h_0/(p + H_0)$
2. $P(\max_{j\in\{1,2,\ldots,M\},\, j\neq i}(\tau_j - T_j^*)^+ < (\tau_i - T_i^*)^+,\ \tau_0 < T_0,\ (\tau_i - T_i^*)^+ + \tau_0 - T_0^* > 0) = h_i/(p + H_0),\ \forall i \in \{0, 1, 2, \ldots, M\}$.

PROOF. We initially establish the equality $P(L_0^* > 0) = H_0/(p + H_0)$. We perturb the optimal solution $\{T_j^*\}_{j=0}^{M}$. For $\epsilon > 0$, define $\hat{T}_0 = T_0^* + \epsilon$ and $\hat{T}_j = T_j^*$, $\forall j \in \{1, 2, \ldots, M\}$. With these definitions of $\{\hat{T}_j\}_{j=0}^{M}$, we can associate the random variables $\hat{E}_j$, $\hat{L}_j$ and $\hat{W}_{-j}$, $\forall j \in \{0, 1, \ldots, M\}$. We have the following identities $\forall j \in \{1, 2, \ldots, M\}$; $\hat{E}_j = E_j^*$, $\hat{L}_j = L_j^*$, and $\hat{W}_{-j} = W_{-j}^*$. Using these, we find that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^{M}) - \bar{C}(\{T_j^*\}_{j=0}^{M}) = \left(\sum_{j=0}^{M} h_j\right)\epsilon + (p + H_0)(\hat{L}_0 - L_0^*).$$

We consider 3 cases regarding the relative magnitudes of L_0^* and ϵ; (a) $L_0^* = 0$, (b) $0 < L_0^* \leq \epsilon$, and (b) $L_0^* > \epsilon$.

(a) $L_0^* = 0$: In this case we have $\hat{L}_0 = 0$ implying that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^{M}) - \bar{C}(\{T_j^*\}_{j=0}^{M}) = \left(\sum_{j=0}^{M} h_j\right)\epsilon$$

(b) $0 < L_0^* \leq \epsilon$: In this case $0 \leq L_0^* - \hat{L}_0 \leq \epsilon$ implying that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^{M}) - \bar{C}(\{T_j^*\}_{j=0}^{M}) = \left(\sum_{j=0}^{M} h_j\right)\epsilon - (p + H_0)(L_0^* - \hat{L}_0)$$

(c) $L_0^* > \epsilon$: Using the fact that for some i, $L_0^* = (L_i^* + W_{-i}^* + \tau_0 - T_0^*)^+$ and $\hat{L}_0 = (\hat{L}_i + W_{-i}^* + \tau_0 - \hat{T}_0)^+ = (\hat{L}_i + W_{-i}^* + \tau_0 - \hat{T}_0 - \epsilon)^+$, we find that $\hat{L}_0 = L_0^* - \epsilon$ implying that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^{M}) - \bar{C}(\{T_j^*\}_{j=0}^{M}) = \left(\sum_{j=0}^{M} h_j\right)\epsilon - (p + H_0)\epsilon.$$

Combining these three cases, we get

$$\begin{aligned}
&\bar{C}(\{\hat{T}_j\}_{j=0}^{M}) - \bar{C}(\{T_j^*\}_{j=0}^{M}) \\
&= \mathbf{1}_{\{L_0^*=0\}}\left(\left(\sum_{j=0}^{M} h_j\right)\epsilon\right) + \mathbf{1}_{\{0<L_0^*\leq\epsilon\}}\left(\left(\sum_{j=0}^{M} h_j\right)\epsilon - (p+H_0)(L_0^* - \hat{L}_0)\right) \\
&\quad + \mathbf{1}_{\{L_0^*>\epsilon\}}\left(\left(\sum_{j=0}^{M} h_j\right)\epsilon - (p+H_0)\epsilon\right).
\end{aligned}$$

Taking the expectations with respect to all possible events and dividing the expected cost difference by ϵ we find

$$\begin{aligned}
&\frac{|C(\{\hat{T}_j\}_{j=0}^{M}) - C(\{T_j^*\}_{j=0}^{M})|}{\epsilon} \\
&= \mathbb{E}[\mathbf{1}_{\{L_0^*=0\}}]\left(\sum_{j=0}^{M} h_j\right) + \mathbb{E}\left[\mathbf{1}_{\{0<L_0^*\leq\epsilon\}}\left(\left(\sum_{j=0}^{M} h_j\right) - (p+H_0)\frac{(L_0^* - \hat{L}_0)}{\epsilon}\right)\right] \\
&\quad + \mathbb{E}[\mathbf{1}_{\{L_0^*>\epsilon\}}]\left(\left(\sum_{j=0}^{M} h_j\right) - (p+H_0)\right).
\end{aligned}$$

Given that for $0 < L_0^* \le \epsilon$ we have $0 \le L_0^* - \hat{L}_0 \le \epsilon$, the following inequality is satisfied

$$\mathbb{E}[\mathbf{1}_{0<L_0^*\le\epsilon}]\Big(\sum_{j=0}^{M} h_j + (p+H_0)\Big) \ge \frac{|C(\{\hat{T}_j\}_{j=0}^M) - C(\{T_j^*\}_{j=0}^M)|}{\epsilon} - \mathbb{E}[\mathbf{1}_{L_0^*=0}]\Big(\sum_{j=0}^{M} h_j\Big) - \mathbb{E}[\mathbf{1}_{L_0^*>\epsilon}]\Big(\sum_{j=0}^{M} h_j - (p+H_0)\Big).$$

As $\{T^*\}_{j=0}^M$ is the optimal solution and the left-hand side tends to zero as ϵ goes to zero, we have

$$P(L_0^*>0)\Big(\sum_{j=0}^{M} h_j - (p+H_0)\Big) + P(L_0^*=0)\sum_{j=0}^{M} h_j = 0,$$

$$P(L_0^*>0)\Big(\sum_{j=0}^{M} h_j - (p+H_0)\Big) + (1-P(L_0^*>0))\sum_{j=0}^{M} h_j = 0,$$

$$\sum_{j=0}^{M} h_j - (p+H_0)P(L_0^*>0) = 0,$$

$$P(L_0^*>0) = \frac{H_0}{p+H_0}.$$

Hence, the required equality is established.

Next, relying on the same idea, we prove the Newsvendor equations. We consider process i. For $\epsilon > 0$, we define $\hat{T}_0 = T_0^* + \epsilon$, $\hat{T}_i = T_i^* - \epsilon$, $\hat{T}_j = T_j^*$, $\forall\, j \in \{1, 2, \ldots, M\}$, $j \ne i$. We derive an expression for $\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M)$ by considering 3 cases and multiple subcases. The cases are based on the relative magnitudes of E_i^* and ϵ; (a) $E_i^* = 0$, (b) $0 < E_i^* \le \epsilon$, and (c) $E_i^* > \epsilon$.

(a) $E_i^* = 0$: It follows that $\hat{E}_i = 0$ and $\hat{L}_i = L_i^* + \epsilon$. Now, we consider 3 subcases based on the relative magnitudes of W_{-i}^* and ϵ; (a.1) $W_{-i}^* = 0$, (a.2) $0 < W_{-i}^* \le \epsilon$, (a.3) $W_{-i}^* > \epsilon$.

(a.1) $W_{-i}^* = 0$: This implies that $\hat{W}_{-i} = 0$. We write $\hat{L}_0$ as follows

$$\hat{L}_0 = (\hat{L}_i + \hat{W}_{-i} + \tau_0 - \hat{T}_0)^+ = (L_i^* + \epsilon + 0 + \tau_0 - T_0^* - \epsilon)^+ = (L_i^* + \tau_0 - T_0^*)^+ = L_0^*.$$

Thus, we have

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j\Big)\epsilon.$$

(a.2) $0 < W_{-i}^* \le \epsilon$: This implies that $\hat{W}_{-i} = 0$. We write $\hat{L}_0$ and $L_0^* - \hat{L}_0$ as follows

$$\hat{L}_0 = (\hat{L}_i + \hat{W}_{-i} + \tau_0 - \hat{T}_0)^+ = (L_i^* + \epsilon + 0 + \tau_0 - T_0^* - \epsilon)^+ = (L_i^* + \tau_0 - T_0^*)^+$$

$$L_0^* - \hat{L}_0 = (L_i^* + W_{-i}^* + \tau_0 - T_0^*)^+ - (L_i^* + \tau_0 - T_0^*)^+.$$

Then, it follows that $0 \le L_0^* - \hat{L}_0 \le \epsilon$ and the cost difference is

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j\Big)\epsilon + h_0(W_{-i}^* - \hat{W}_{-i}) - (p+H_0)(L_0^* - \hat{L}_0).$$

(a.3) $W_{-i}^* > \epsilon$: This implies that $\hat{W}_{-i} = W_{-i}^* - \epsilon$. We write $\hat{L}_0$ as follows

$$\hat{L}_0 = (\hat{L}_i + \hat{W}_{-i} + \tau_0 - \hat{T}_0)^+ = (L_i^* + \epsilon + W_{-i}^* - \epsilon + \tau_0 - T_0^* - \epsilon)^+ = (L_i^* + W_{-i}^* + \tau_0 - T_0^* - \epsilon)^+.$$

Within this subcase we consider three sub-subcases; (a.3.1) $L_0^* = 0$, (a.3.2) $0 < L_0^* \le \epsilon$ and (a.3.3) $L_0^* > \epsilon$.

(a.3.1) $L_0^* = 0$: Then $\hat{L}_0 = 0$, from which it follows that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j + h_0\Big)\epsilon.$$

(a.3.2) $0 < L_0^* \le \epsilon$: Then $\hat{L}_0 = 0$, from which it follows that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j + h_0\Big)\epsilon - (p+H_0)L_0^*.$$

(a.3.3) $L_0^* > \epsilon$: Then $\hat{L}_0 = L_0^* - \epsilon$, from which it follows that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j + h_0\Big)\epsilon - (p+H_0)\epsilon.$$

(b) $0 < E_i^* \le \epsilon$: In this case we have the following identities and inequalities: $L_i^* = 0$, $0 < \hat{L}_i \le \epsilon$, $\hat{E}_i = 0$, $0 < L_0^* - \hat{L}_0 \le \epsilon$ and $0 < W_{-i}^* - \hat{W}_{-i} \le \epsilon$. Then, it follows that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j\Big)\epsilon + h_0(E_i^* - \hat{E}_i) + h_0(W_{-i}^* - \hat{W}_{-i}) - (p+H_0)(L_0^* - \hat{L}_0).$$

(c) $E_i^* > \epsilon$: In this case we have the following identities: $\hat{E}_i = E_i^* - \epsilon$, $\hat{L}_i = L_i^* = 0$, $\hat{W}_{-i} = W_{-i}^*$, $\hat{E}_j = E_j^*$ and $\hat{L}_j = L_j^*$, $\forall\, j \in \{1, 2, \ldots, M\}$, $j \ne i$. Then, it follows that

$$\hat{L}_0 = (\hat{L}_i + \hat{W}_{-i} + \tau_0 - \hat{T}_0)^+ (L_i^* + W_{-i}^* + \tau_0 - (T_0^* + \epsilon))^+.$$

Based on the relative magnitudes of L_0^* and ϵ we consider three subcases: (c.1) $L_0^* = 0$, (c.2) $0 < L_0^* \le \epsilon$, and (c.3) $L_0^* > \epsilon$.

(c.1) $L_0^* = 0$: Then $\hat{L}_0 = 0$, from which it follows that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j + h_0\Big)\epsilon.$$

(c.2) $0 < L_0^* \le \epsilon$: Then $\hat{L}_0 = 0$, from which it follows that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j + h_0\Big)\epsilon - (p+H_0)L_0^*.$$

(c.3) $L_0^* > \epsilon$: Then $\hat{L}_0 = L_0^* - \epsilon$, from which it follows that

$$\bar{C}(\{\hat{T}_j\}_{j=0}^M) - \bar{C}(\{T_j^*\}_{j=0}^M) = \Big(\sum_{j=1,j\ne i}^{M} h_j\Big)\epsilon + h_0(E_i^* - \hat{E}_i) + h_0(W_{-i}^* - \hat{W}_{-i}) - (p+H_0)(L_0^* - \hat{L}_0) = \Big(\sum_{j=1,j\ne i}^{M} h_j + h_0\Big)\epsilon - (p+H_0)\epsilon.$$

Now we have covered all possible cases. This implies that

$$
\begin{aligned}
&\bar{C}(\{\hat{T}_j\}_{j=0}^M)-\bar{C}(\{T_j^*\}_{j=0}^M)\\
&=\mathbf{1}_{\{E_i^*=0,\,W_i^*=0\}}\Big(\sum_{j=1,\,j\neq i}^{M}h_j\Big)\epsilon+\mathbf{1}_{\{E_i^*=0,\,0<W_i^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon\\
&\quad+h_0(W_{-i}^*-\hat{W}_{-i}-\epsilon)-(p+H_0)(L_0^*-\hat{L}_0)\Big)\\
&\quad+\mathbf{1}_{\{E_i^*=0,\,W_i^*>0,\,L_0^*=0\}}\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon\\
&\quad+\mathbf{1}_{\{E_i^*=0,\,W_i^*>0,\,0<L_0^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon-(p+H_0)L_0^*\Big)\\
&\quad+\mathbf{1}_{\{E_i^*=0,\,W_i^*>0,\,L_0^*>\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon-(p+H_0)\epsilon\Big)\\
&\quad+\mathbf{1}_{\{0<E_i^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon+h_0(E_i^*-\hat{E}_i-\epsilon)\\
&\qquad+h_0(W_{-i}^*-\hat{W}_{-i})-(p+H_0)(L_0^*-\hat{L}_0)\Big)\\
&\quad+\mathbf{1}_{\{E_i^*>\epsilon,\,L_0^*=0\}}\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon\\
&\quad+\mathbf{1}_{\{E_i^*>\epsilon,\,0<L_0^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon-(p+H_0)\hat{L}_0\Big)\\
&\quad+\mathbf{1}_{\{E_i^*>\epsilon,\,L_0^*>\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon-(p+H_0)\epsilon\Big).
\end{aligned}
$$

Rearranging the terms, dividing by ϵ, and taking the expectations, we find

$$
\begin{aligned}
&\Big|\frac{C(\{\hat{T}_j\}_{j=0}^M)-C(\{T_j^*\}_{j=0}^M)}{\epsilon}-\mathbb{E}[\mathbf{1}_{\{E_i^*=0,\,W_i^*=0\}}]\Big(\sum_{j=1,\,j\neq i}^{M}h_j\Big)\\
&\quad-\mathbb{E}[\mathbf{1}_{\{E_i^*=0,\,W_i^*>0,\,L_0^*=0\}}]\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)-\mathbb{E}[\mathbf{1}_{\{E_i^*=0,\,W_i^*>0,\,L_0^*>\epsilon\}}]\\
&\quad\cdot\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)-(p+H_0)\Big)-\mathbb{E}[\mathbf{1}_{\{E_i^*>\epsilon,\,L_0^*=0\}}]\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\\
&\quad-\mathbb{E}[\mathbf{1}_{\{E_i^*>\epsilon,\,L_0^*>\epsilon\}}]\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)-(p+H_0)\Big)\Big|\\
&=\frac{1}{\epsilon}\Big|\mathbb{E}\Big[\mathbf{1}_{\{E_i^*=0,\,0<W_i^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon\\
&\qquad+h_0(W_{-i}^*-\hat{W}_{-i}-\epsilon)-(p+H_0)(L_0^*-\hat{L}_0)\Big)\Big]\\
&\quad+\mathbb{E}\Big[\mathbf{1}_{\{E_i^*=0,\,0<W_i^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon\\
&\qquad+h_0(W_{-i}^*-\hat{W}_{-i}-\epsilon)-(p+H_0)(L_0^*-\hat{L}_0)\Big)\Big]\\
&\quad+\mathbb{E}\Big[\mathbf{1}_{\{0<E_i^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon+h_0(E_i^*-\hat{E}_i-\epsilon)\\
&\qquad+h_0(W_{-i}^*-\hat{W}_{-i})-(p+H_0)(L_0^*-\hat{L}_0)\Big)\Big]\\
&\quad+\mathbb{E}\Big[\mathbf{1}_{\{E_i^*>\epsilon,\,0<L_0^*\le\epsilon\}}\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\epsilon-(p+H_0)\hat{L}_0\Big)\Big]\Big|.
\end{aligned}
$$

As all the terms on the right-hand side are bounded and the intervals approach zero probability mass, we find that the right-hand side of the above equality goes to zero as ϵ approaches 0. We also note that because of the optimality of $\{T_j^*\}_{j=0}^M$, we have $\lim_{\epsilon\to 0}((C(\{\hat{T}_j\}_{j=0}^M)-C(\{T_j^*\}_{j=0}^M))/\epsilon)=0$. Then it follows that

$$
\begin{aligned}
&P(E_i^*=0,\,W_i^*=0)\Big(\sum_{j=1,\,j\neq i}^{M}h_j\Big)\\
&\quad+P(E_i^*=0,\,W_i^*>0,\,L_0^*=0)\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\\
&\quad+P(E_i^*=0,\,W_i^*>0,\,L_0^*>0)\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)-(p+H_0)\Big)\\
&\quad+P(E_i^*>0,\,L_0^*=0)\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)\\
&\quad+P(E_i^*>0,\,L_0^*>0)\Big(\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)-(p+H_0)\Big)=0.
\end{aligned}
$$

After rearrangement of the terms we get

$$
\begin{aligned}
&(p+H_0)(P(E_i^*=0,W_i^*>0,L_0^*>0)+P(E_i^*>0,L_0^*>0))\\
&=\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)(P(E_i^*=0,W_i^*>0,L_0^*=0)\\
&\qquad+P(E_i^*=0,W_i^*>0,L_0^*>0))\\
&\quad+\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)(P(E_i^*>\epsilon,L_0^*=0)+P(E_i^*>0,L_0^*>0))\\
&\quad+\Big(\sum_{j=1,\,j\neq i}^{M}h_j\Big)P(E_i^*=0,W_i^*=0).
\end{aligned}
$$

The sum of the probabilities with $(\sum_{j=1,\,j\neq i}^{M}h_j+h_0)$ as a multiplier can be simplified to

$$
\begin{aligned}
&P(E_i^*=0,\,W_i^*>0,\,L_0^*=0)+P(E_i^*=0,\,W_i^*>0,\,L_0^*>0)\\
&\quad+P(E_i^*>0,\,L_0^*=0)+P(E_i^*>0,\,L_0^*>0)\\
&=P(E_i^*>0)+P(E_i^*=0,\,W_{-i}^*>0)=1-P(E_i^*=0,\,W_{-i}^*=0).
\end{aligned}
$$

Using this simplification we write

$$
\begin{aligned}
&(p+H_0)(P(E_i^*=0,\,W_i^*>0,\,L_0^*>0)+P(E_i^*>0,\,L_0^*>0))\\
&\quad=\Big(\sum_{j=1,\,j\neq i}^{M}h_j+h_0\Big)-h_0P(E_i^*=0,\,W_{-i}^*=0).
\end{aligned}
$$

Next we note that the following identity holds

$$
\begin{aligned}
P(L_0^*>0)&=P(E_i^*>0,\,L_0^*>0)+P(E_i^*=0,\,L_0^*>0)\\
&=P(E_i^*>0,\,L_0^*>0)+P(E_i^*=0,\,W_{-i}^*>0,\,L_0^*>0)\\
&\quad+P(E_i^*=0,\,W_{-i}^*=0,\,L_0^*>0).
\end{aligned}
$$

Using this identity and the equality $\sum_{j=1,\, j\neq i}^{M} h_j + h_0 = H_0 - h_i$ we write the optimality equation as

$$\begin{aligned}&(p+H_0)(P(L_0^*>0)-P(E_i^*=0, W_{-i}^*=0, L_0^*>0))\\&\quad = H_0 - h_i - h_0 P(E_i^*=0, W_{-i}^*=0).\end{aligned}$$

Furthermore, we have

$$\begin{aligned}&P(E_i^*=0, W_{-i}^*=0, L_0^*>0)\\&\quad = P(E_i^*=0, W_{-i}^*=0, \tau_0>T_0^*)\\&\qquad + P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0)\\&\quad = P(E_i^*=0, W_{-i}^*=0)P(\tau_0>T_0^*)\\&\qquad + P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0).\end{aligned}$$

The later equality is due to independence of τ of $E_i^*=0$ and W_{-i}^*. Substituting the latest equality into our optimality equation, we find

$$\begin{aligned}&(p+H_0)\big[P(L_0^*>0)-P(E_i^*=0, W_{-i}^*=0)P(\tau_0>T_0^*)\\&\qquad - P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0)\big]\\&= H_0 - h_i - h_0 P(E_i^*=0, W_{-i}^*=0)\\&\quad \cdot((p+H_0)P(\tau_0>T_0^*)-h_0)P(E_i^*=0, W_{-i}^*=0)\\&= -(p+H_0)P(L_0^*>0) - h_i + H_0\\&\qquad + (p+H_0)P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0).\end{aligned}$$

Using the equality $-(p+H_0)P(L_0^*>0)+H_0=0$, we obtain the following set of equations $\forall i\in\{1,2,\ldots,M\}$.

$$\begin{aligned}&((p+H_0)P(\tau_0>T_0^*)-h_0)P(E_i^*=0, W_{-i}^*=0)\\&\quad = -h_i + (p+H_0)P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0).\end{aligned}\tag{7}$$

The above set of equations are solved if

$$P(\tau_0>T_0^*) = \frac{h_0}{p+H_0}$$

$$P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0) = \frac{h_i}{p+H_0}.$$

Note that $P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0)$ equals the probability that process i finishes late, and it is the process that causes process 0 to start late, and process 0 has a throughput time shorter than its planned leadtime and the system is delivered late. This event is donated as the event that process i causes lateness. The event $\tau_0 > T_0$ is defined as the event that process 0 causes lateness. Thus we find that $P(\text{process } i \text{ causes system lateness}) = h_i/(p+H_0)$, $\forall i\in\{0,1,\ldots,M\}$. Furthermore, the events are mutually exclusive and together make up all possible events for which the system is delivered late. Thus, we have $P(\text{system is delivered late}) = \sum_{i=0}^{M}(h_i/(p+H_0)) = H_0/(p+H_0)$.

For sake of completeness, we note that

$$\begin{aligned}&P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0)\\&\quad = P\Big(\max_{j\in\{1,2,\ldots,M\},\, j\neq i}(\tau_j - T_j^*)^+ < (\tau_i - T_i^*)^+,\\&\qquad \tau_0 < T_0, (\tau_i - T_i^*)^+ + \tau_0 - T_0^* > 0\Big). \quad \square\end{aligned}$$

Finally, with the following lemma we prove the uniqueness of the solution.

Lemma 3. *The solution to the optimality equations in Theorem 2 is unique.*

Proof. If we sum the optimality Equations (7) $\forall i\in\{1,2,\ldots,M\}$ we obtain

$$\begin{aligned}&((p+H_0)P(\tau_0>T_0^*)-h_0)\sum_{i=1}^{M}P(E_i^*=0, W_{-i}^*=0)\\&= -\sum_{i=1}^{M}h_i + (p+H_0)\sum_{i=1}^{M}P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0).\end{aligned}\tag{8}$$

Note that we have

$$\begin{aligned}&\sum_{i=1}^{M}P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0)\\&\quad = P(L_0^*>0) - P(\tau_0>T_0^*) = \frac{H_0}{p+H_0} - P(\tau_0>T_0^*).\end{aligned}$$

Therefore, the equality (8) becomes

$$\begin{aligned}&((p+H_0)P(\tau_0>T_0^*)-h_0)\sum_{i=1}^{M}P(E_i^*=0, W_{-i}^*=0)\\&= -\sum_{i=1}^{M}h_i + (p+H_0)\left(\frac{H_0}{p+H_0} - P(\tau_0>T_0^*)\right)\\&= h_0 - (p+H_0)P(\tau_0>T_0^*).\end{aligned}$$

As $\sum_{i=1}^{M}P(E_i^*=0, W_{-i}^*=0)<1$, for any solution $\{T_j^*\}_{j=1}^{M}$, we must have $P(\tau_0>T_0^*) = h_0/(p+H_0)$. This implies the equality $P(E_i^*=0, W_{-i}^*=0, \tau_0\le T_0^*, L_0^*>0) = h_i/(p+H_0)$. Hence, the solution to the optimality equations is unique and gives a global minimum. $\square$

References

ASML (2014) ASML Annual Report 2013. Accessed February 18, 2015, http://www.asml.com/asml/show.do?ctx=49757.

Axsäter S (2005) Planning order releases for an assembly system with random operation times. *OR Spectrum* 27(2):459–470.

Buzacott J, Shanthikumar J (1994) Safety stock versus safety time in MRP controlled production systems. *Management Sci.* 40(5):1678–1689.

Chauhan S, Dolgui A, Proth J-M (2009) A continuous model for supply planning of assembly systems with stochastic component procurement times. *Internat. J. Production Econom.* 120(2): 411–417.

Chhaochhria P, Graves S (2013) A forecast-driven tactical planning model for a serial manufacturing system. *Internat. J. Production Res.* 51(23–24):6860–6879.

Clark J, Scarf H (1960) Optimal policies for a multi-echelon inventory problem. *Management Sci.* 6(4):475–490.

de Kok T, Fransoo J (2003) Planning supply chain operations: Definitions and Comparison of Planning Concepts. de Kok T, Graves S, eds. *Handbooks in Operations Research and Management Science* (Elsevier, Amsterdam), 597–677.

Diks E, de Kok T (1998) Optimal control of a divergent multi-echelon inventory system. *Eur. J. Oper. Res.* 111(1):75–97.

Diks E, de Kok T (1999) Computational results for the control of a divergent N-echelon inventory system. *Internat. J. Production Econom.* 59(1–3):327–336.

Goldratt E, ed. (1997) *Critical Chain* (North River Press, Great Barrington, MA), 1–246.

Gong L, de Kok T, Ding J (1994) Optimal leatimes planning in a serial production system. *Management Sci.* 40(5):629–632.

Hashamdar H, Ibrahim Z (2010) A new method of dynamic analysis structures by using advance mathematical. *Asian J. Appl. Sci.* 3(3):186–196.

Hicks C (2004) A genetic algorithm tool for designing manufacturing facilities in the capital goods industry. *Internat. J. Production Econom.* 90(2):199–211.

Hicks C, Pongcharoen P (2006) Dispatching rules for production scheduling in the capital goods industry. *Internat. J. Production Econom.* 104(1):154–163.

Hopp W, Spearman M (1993) Setting safety leadtimes for purchased components in assembly systems. *IIE Trans.* 25(2):2–11.

Kanet J, Christy D (1984) Manufacturing systems with forbidden early order departure. *Internat. J. Production Res.* 22(1):41–50.

Matsuura H, Tsubone H (1993) Setting planned lead times in capacity requirements planning. *J. Oper. Res. Soc.* 44(8):809–816.

Matsuura H, Tsubone H, Kanezashi M (1996) Setting planned lead times for multi-operation jobs. *Eur. J. Oper. Res.* 88(2): 287–303.

Molinder A (1997) Joint optimization of lot-sizes, safety stocks and safety lead times in an MRP system. *Internat. J. Production Res.* 35(4):983–994.

Press W, Teukolsky S, Vetterling W, Flannery B, eds. (2007) *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, New York), 490–522.

Rosling K (1989) Optimal inventory policies for assembly systems under random demand. *Oper. Res.* 37(4):565–579.

Rosso D (2014) Semiconductor industry posts record sales in 2013. *Semiconductor Indust. Assoc.* (February 3), http://www.semiconductors.org/news/2014/02/03/global_sales_report_2013/semiconductor_industry_posts_record_sales_in_2013/.

Song D, Earl C, Hicks C (2001) Stage due date planning for multistage assembly systems. *Internat. J. Production Res.* 39(9): 1943–1954.

Teo C, Bhatnagar R, Graves S (2011) Setting planned lead times for a make-to-order production system under master schedule smoothing. *IIE Trans.* 43(6):399–414.

Teo C, Bhatnagar R, Graves S (2012) An application of master schedule smoothing and planned lead time control. *Production Oper. Management* 21(2):211–223.

Weeks J (1981) Optimizing planned lead times and delivery dates. *21th Ann. Conf. Proc., Amer. Production and Inventory Control Soc.*, 177–188.

Whitt W (1982) Approximating a point process by a renewal process, I: Two basic methods. *Oper. Res.* 30(1):125–147.

Whybark D, Williams J (1976) Material requirements planning under uncertainty. *Decision Sci.* 7(4):595–606.

Yano CA (1987a) Setting planned leadtimes in serial production systems with tardiness costs. *Management Sci.* 33(1):95–106.

Yano CA (1987b) Stochastic leadtimes in two-level assembly systems. *IIE Trans.* 19(4):371–378.

Yano CA (1987c) Stochastic leadtimes in two-level distribution-type networks. *Naval Res. Logist.* 34(6):831–843.

Zipkin P, ed. (2000) *Foundations of Inventory Management* (Irwin/McGraw-Hill, New York).