Optimal Capacity Conversion for Product Transitions Under High Service Requirements

Citation: Li, Hongmin, Stephen C. Graves, and Woonghee Tim Huh. "Optimal Capacity Conversion for Product Transitions Under High Service Requirements." M&SOM 16, no. 1 (February 2014): 46–60.

As Published: http://dx.doi.org/10.1287/msom.2013.0445

Publisher: Institute for Operations Research and the Management Sciences (INFORMS)

Persistent URL: http://hdl.handle.net/1721.1/90834

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Optimal Capacity Conversion for Product Transitions under High Service Requirements

Hongmin Li • Stephen C. Graves • Woonghee Tim Huh

W.P. Carey School of Business, Arizona State University
Tempe, Arizona 85287, USA

Sloan School of Management, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Sauder School of Business, University of British Columbia
Vancouver, BC, Canada, V6T 1Z2

We consider the capacity planning problem during a product transition in which demand for a new-generation product gradually replaces that for the old product. Capacity for the new product can be acquired both by purchasing new production lines and by converting existing production lines for the old product. Furthermore, in either case, the new product capacity is "retro-fitted" to be flexible, i.e., to be able to also produce the old product. This capacity planning problem arises regularly at Intel, which served as the motivating context for this research. We formulate a two-product capacity planning model to determine the equipment purchase and conversion schedule, considering (i) time-varying and uncertain demand, (ii) dedicated and flexible capacity, (iii) inventory and equipment costs, and (iv) a chance-constrained service level requirement. We develop a solution approach that accounts for the risk-pooling benefit of flexible capacity (a closed-loop planning approach) and compare it with a solution that is similar to Intel's current practice (an open-loop planning approach). We evaluate both approaches with a realistic but disguised example and show that the closed-loop planning solution leads to savings in both equipment and inventory costs, and matches more closely the service level targets for the two products. Our numerical experiments illuminate the cost tradeoffs between purchasing new capacity and converting old capacity, and between a level capacity plan versus a chase capacity plan.

Keywords: *Capacity Planning, Product Transition, Equipment Conversion, Flexible Capacity, Risk Pooling*

1 Introduction

Technology advances often drive frequent product upgrades in the high-tech industry. For example, Intel continually introduces new processors into the market, driven by its tick-tock cadence strategy where each "tick" represents a new generation of silicon technology and each "tock" represents a new product architecture (Shenoy and Daniel, 2006). As a result, a new microprocessor is introduced to the market each year and the company is constantly in transition, i.e., moving the production from one product to the next. In this paper, we study the capacity planning decisions during a product transition. During the transition period, the demand for the older generation product gradually ramps down whereas that for the new product ramps up. Therefore, companies need to

manage production according to this demand behavior. For each transition, many changes take place throughout the manufacturing process, often requiring upgrades or replacement of equipment; for example, at Intel, changes in process technology occur at wafer production (Fabs), as well as at Assembly and Test.

In comparison to product upgrades, equipment updates are less frequent since the same equipment can often be used for the production of multiple generations of products. However, during each product transition, the equipment has to be reconfigured to fit the production specifications of the new product. This may involve mechanically changing the equipment configurations (for example, replacing certain tools), relocating equipment across factories, as well as qualifying new equipment configurations for the production of existing and new products. Qualification at Intel follows the "copy exactly" process, which was developed to ensure fast and reliable technology transfers from development to mass production: "everything which might affect the process, or how it is run, is to be copied down to the finest detail ... data is collected at the process step output level on parameters ... and they are compared to results at the R&D site" (McDonald, 1998). In other words, each set of equipment has to match the target output generated from the development factory, and as a result, the qualification of equipment is a time-consuming, resource-intensive, and costly process. In a collaborative research project with Intel, we have examined the equipment decisions relating to product transitions that occur in an Assembly and Test production environment; the results are applicable to manufacturing companies facing similar challenges caused by product transitions. For example, in the automotive industry, the introduction of a new model often requires "retooling" of existing production lines (Bresnahan and Ramey, 1994; Leone and Bradley, 1982); thus a similar problem of capacity planning ensues.

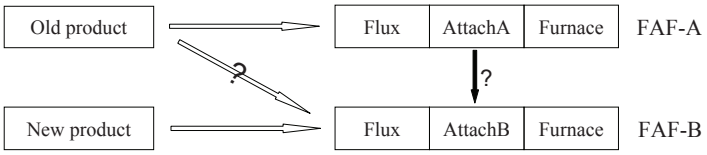


Figure 1: Equipment Options

As an example, we consider one equipment module in the assembly factory for Intel. The module consists of three tools and attaches an electronic chip to a substrate: (1) The flux tool applies flux to the substrate to remove oxides and contaminations from the surfaces, (2) the attachment tool places the chip die on the substrate and aligns it with the substrate, and (3) then, in a "reflow"

process, a furnace heats the substrate and forms a solder joint between the substrate and the chip. We refer to this module as the Flux-Attachment-Furnace or "FAF" module (to mask the actual name). A FAF module is configured to produce a single product or multiple products, where the attachment tool determines the configuration. In Figure 1, we illustrate two configurations of the FAF module: FAF-A has an attachment tool A that allows it to produce the older generation product, while FAF-B produces the new generation with attachment tool B. In this paper, we refer to each set of equipment of a particular configuration as a *line*, which is a combination of several tools and performs a sequence of production steps. The FAF line provides an illustration of equipment options during a transition. Often there might be multiple tools within a line that need to be replaced in a conversion; the mathematical model and method are applicable to these more general cases.

There are two ways to create capacity for the new generation product. First, the company can purchase new equipment from suppliers to set up a new FAF-B line. A new FAF-B line has all three of its subtools newly purchased from suppliers. Second, as the demand for the old product decreases, the need for FAF-A lines decreases, and the company may convert an existing FAF-A line to a FAF-B line by replacing the AttachA tool with an AttachB tool, as indicated by the solid arrow in Figure 1 (the question mark indicating that this is an option). In this case, the company only pays for a new AttachB tool instead of the entire line. The old line must be shut down for the duration of the conversion, which is about the same length as the lead time for the purchase and installation of a new line. In most cases, the conversion is irreversible. A manager has to decide upon an appropriate equipment conversion and procurement schedule during the capacity planning stage, which is the focus of this paper. We study these capacity planning decisions when a company faces stochastic and time-varying demand for two consecutive generations of products.

One option often used in practice is to "retro-fit" the new equipment so that it can also produce the old product, which effectively creates a flexible line. In Figure 1, we show this option with the hollow arrow extending from the old product to FAF-B. Although there are benefits from flexible capacity, the cost is high to retro-fit a line to be flexible across product generations (in a typical case, around $\$500,000$). This decision induces debate among the managers on whether the retro-fit investment cost is justified. Currently these decisions are made on an *ad hoc* basis. Therefore, our solution also enables the manager to quantify the value of retro-fitting.

The company determines the capacity plan nine to twelve months ahead of the actual demand based on the sales forecast. As is common in most capacity decisions, the fundamental tradeoff faced by the company is to meet service level requirements, and to control its equipment and

inventory costs in light of the demand uncertainty for both the old and new generation products. In this paper, given the time-varying and uncertain demand, we determine the optimal capacity plan: when to convert existing FAF-A lines to FAF-B lines, how many to convert, and when and how many new FAF-B lines to purchase.

The main contributions of the paper are to present, formulate and solve a capacity problem that is commonly faced by manufacturing companies experiencing product transitions. We develop a two-product joint capacity and production model with the following features: (i) time-varying and uncertain demand, (ii) dedicated as well as flexible capacity, (iii) inventory and equipment costs, and (iv) chance-constrained service level requirement. We develop a simple solution approach to the problem when production decisions are made before the realization of demand (an open-loop planning approach, which is a proxy for Intel's current practice) and a heuristic capacity solution to account for the risk-pooling benefit of flexible capacity (a closed-loop planning approach that postpones the production decision until after demand realization). We evaluate the model and solution approaches with a disguised Intel data set. The comparison of the two solutions suggests that accounting for risk pooling leads to savings in both equipment cost and inventory cost, in addition to providing more balanced service levels between the two products. Our solutions reveal two major tradeoffs in this capacity planning problem: one between the options of new purchase and conversion, and one between chase versus level capacity strategies. Whereas such tradeoffs have been studied in the literature, they have not been examined in the setting of a product transition.

In Section 2, we provide a survey of relevant literature. In Section 3, we describe in detail the assumptions and modeling choices made in the paper. In Section 4, we propose a two-step approach to solve the capacity problem, assuming production decisions are made before the realization of demand. In Section 5, we consider the impact of delaying the production decisions, and develop a heuristic solution to account for capacity flexibility. In Section 6, we apply these methods to a realistic example and test the performance of the methods. We conclude in Section 7.

2 Relationships to Prior Work

Literature on capacity planning has been vast and broad, and Luss (1982) and van Mieghem (2003) provide in-depth albeit dated reviews. We review research streams directly relevant to our work.

Capacity Expansions and Replacement. An early stream of research examines the tradeoff of capacity installation cost and economies of scale in capacity expansion (e.g., Manne 1961, Erlenkotter 1974, Bean and Smith 1985), to identify the timing and size of capacity expansions given

the demand growth. The capacity replacement literature (e.g., Yano 1984, Jones et al. 1991, Bean et al. 1985, 1994) considers equipment replacement due to aging, deterioration, and obsolescence. Rajagopalan (1998) combines replacement with expansion in a deterministic model where capacity can be purchased and disposed to meet the demand, and extends his model to include uncertain technology breakthroughs. In contrast, we focus on capacity expansion and conversion due to the introduction of a new product and product transition. In addition, we account for production and inventory planning considerations, along with capacity decisions.

Demand Uncertainty. Most capacity planning models in the literature do not address uncertainty. Among those that do, demand uncertainty is often modeled with a scenario-based approach (e.g., Eppen et al. 1989, Karabuk and Wu 2002, Swaminathan 2000, Ahmed and Sahinidis 2003, Huang and Ahmed 2009). The scenario-based method works well if demand can be represented by a small number of demand scenarios. However, as the number of possible scenarios increases, the computation burden grows rapidly. The problem exacerbates with time-varying demand and with the number of products and time periods under consideration.

Joint Capacity and Production Decisions. Capacity and production decisions are often decoupled in practice because they involve different scopes with different planning horizons. Thus, a capacity problem can be thought of a two-stage problem with recourse, consisting of (i) a capacity decision and (ii) a production decision. The second-stage problem is a multi-product *capacitated* production problem, and that alone has proven to be difficult since one has to consider the inventory and future capacity availability when determining the production quantities (see, for example Janakiraman et al. (2008)). The joint capacity and production problem is therefore more challenging. In simpler single-product, single-capacity settings of Angelus and Porteus (2002) and Bradley and Glynn (2002), capacity and production policies for minimizing a cost can be specified using thresholds. In this paper, we study a *two*-product, *two*-capacity, multi-period capacity problem with stochastic and time-varying demand under high service requirements, where service level requirements are exogenously imposed. We need to determine capacity purchase and capacity conversion in the presence of product-flexible capacity.

For a joint capacity and production problem with multiple products, Rajagopalan and Swaminathan (2001) have developed a heuristic procedure when demand is deterministic. With stochastic demand, Huh and Roundy (2005) assume lost sales and no inventory carry-over; hence there is no temporal dependence in their problem setup. Their approach is extended by Huh et al. (2006). All of these papers do not explicitly capture the inventory carry-over and the risk-pooling effect among products.

Product-flexible Capacity. In our problem, the equipment configuration for the new product can be made flexible to also produce the old product. Early work on capacity planning involving flexible capacity often ignores the risk-pooling benefit (e.g., Erlenkotter 1974). However, flexibility is a hedging tactic (van Mieghem, 2007) and the value of product-flexible capacity is commonly recognized but difficult to quantify. Fine and Freund (1990) study a single-period capacity problem in which both dedicated and flexible capacities are available, and provide shadow values of the capacity constraints. van Mieghem (1998) considers a similar problem and shows that investing in flexible capacity could be optimal even with perfectly positive correlation between the products because of the price differentials of the products. Bish and Wang (2004) extend the van Mieghem (1998) model to a continuous demand distribution and they include *ex post* pricing decisions. For multi-period capacity planning problems, Li and Tirupati (1994) allow both dedicated and flexible capacities to exhibit economies of scale and study the tradeoff with linear operating costs. They later extend the model to a special case of stochastic demand with a stationary and uniform demand distribution (Li and Tirupati, 1995), assuming probabilistic service level instead of shortage cost penalties, a modeling choice we adopt in this paper as well. Both of these papers ignore inventory fluctuations and backorders.

3 Problem Description and Assumptions

We consider a setting with two equipment configurations and two products, where the old configuration (Configuration A) is dedicated to produce the old product, and the new configuration (Configuration B) is flexible to produce both the old and the new products. The old and new products are referred to as products 1 and 2, respectively. We study a problem of planning capacity for a fixed horizon of T periods, and we assume that all the capacity decisions are made at the beginning of the horizon; this assumption is partly due to long procurement lead times for capacity adjustment, and is consistent with the current practice. (See, for example, Çakanyildirim et al. (2004).)

Initial Equipment Configuration. Our interest is the planning horizon during which one product transitions from one generation to the next. We assume that, at the beginning of the planning horizon, the existing Configuration-A capacity is sufficient for the demand of the old product. Since this demand is expected to decline during the horizon, we do not consider increasing the Configuration-A capacity. For the new configuration (Configuration-B), the company usually starts with a small number (often 1 or 2) of lines; these are often set up during the development

stage and used as a benchmark for any additional lines. We also assume zero inventory or backorders at the beginning of the horizon.

Capacity Decisions and Qualification Lead Times. We can acquire capacity for the new configuration by (i) purchasing new lines or (ii) converting existing old configuration lines to the new configuration, or both. In either case, it takes some time to assemble or convert the equipment and then to qualify the line, i.e., to install, test and tune the equipment such that they meet the performance standard. This lead time is approximately the same for both new purchase and conversion at Intel (in the motivating context); we assume the lead times are the same and denote the lead time by τ.

Let $Q_b(t)$ denote the number of new lines purchased that become *available for production* at time t, i.e., the company purchases the equipment from the suppliers at time $t-\tau$. Let $Q_c(t)$ be the number of conversions the company starts at time t (when the equipment under conversion can no longer produce the old product), and the converted equipment becomes available as Configuration-B capacity in period $t+\tau$. Let $K^A(t)$ and $K^B(t)$ be the available capacity (number of lines) of Configurations A and B, respectively, at time t. (The initial capacities, $K^A(0)$ and $K^B(0)$, are already given.) Then, capacity can be modeled by the following equations:

$$K^A(t) = K^A(t-1) - Q_c(t), \tag{3.1}$$

$$K^B(t) = K^B(t-1) + Q_b(t) + Q_c(t-\tau)\ . \tag{3.2}$$

Production Decisions. Our model includes production decisions and their inventory implications, as they are relevant within the planning horizon for the capacity decisions under study. Let $P_1^A(t)$ denote the production quantity of product 1 using Configuration A. Similarly, let $P_1^B(t)$ and $P_2^B(t)$ denote the production quantities of products 1 and 2, respectively, using Configuration B. Also, let r_1^A, r_1^B and r_2^B represent the corresponding production rates (units per machine hour), and let μ be the number of machine hours per line of capacity per time period. Then, the production quantities should satisfy the following capacity constraints:

$$P_1^A(t)/r_1^A(t) \leq \mu \cdot K^A(t)\ , \tag{3.3}$$

$$P_1^B(t)/r_1^B(t) + P_2^B(t)/r_2^B(t) \leq \mu \cdot K^B(t)\ . \tag{3.4}$$

At the time of capacity planning, one needs to have some understanding of future production decisions since they in turn affect capacity decisions. In the model presented in Section 4, we suppose that one determines a long-term production plan at the same time as capacity planning. Following the convention in Bertsekas (2000), we name this the "open-loop" planning model. This

loosely mimics Intel's current practice on capacity planning. We use this as a baseline and compare it to the closed-loop planning model that we present in Section 5. In this case we assume that one can delay the determination of the production plan until after demand has been realized.

Costs and Service Requirement. Based on Intel's current practice, we take a depreciation approach to account for cost related to capacity decisions. If c is the investment associated with a capacity decision, we attribute to each period the cost of $\delta \cdot c$, where δ is the depreciation rate per time period; this cost is charged for each period from the time the investment is made until the end of the planning horizon.

Let c_b and c_c denote the cost of purchasing one line of Configuration B and the cost of converting one line from Configuration A to B, respectively. Then, for a new line that becomes available for production in period t, the total depreciation cost associated with this line during the horizon is $\delta \cdot c_b \cdot [T - (t - \tau)]$ since this line must have arrived from the supplier τ periods before time t in order to go through the qualification process. Similarly, for any Configuration-A line that started conversion in period t, the total depreciation cost is $\delta \cdot c_c \cdot (T - t)$.

We model the service level as a probabilistic chance constraint for each product. Often, the service level requirement of a product is a hard constraint that is ultimately driven by the key customers of that particular product; thus it is product-specific. In addition, service level requirements at Intel are high (often between 95-99%) because stockout can cause major production disruptions and enormous loss to its OEM customers. Let $S_i(t)$ be a random variable representing the cumulative demand for product i by time t. We assume that the demand distributions are continuous. We let $CP_i(t)$ represent the cumulative production quantity for product i at the end of period t, which evolves according to

$$CP_1(t) = CP_1(t-1) + P_1^A(t) + P_1^B(t) \qquad \text{and} \tag{3.5}$$

$$CP_2(t) = CP_2(t-1) + P_2^B(t) \ . \tag{3.6}$$

We assume we have a target service level requirement for each product, given by η_i; and the service requirement is given as a chance constraint for each $i \in \{1, 2\}$:

$$Prob\left(S_i(t) \leq CP_i(t)\right) \geq \eta_i \ , \tag{3.7}$$

for each period t. We can have a time-dependent service requirement $\eta_i(t)$, but do not include this for expositional simplicity. This service level measure corresponds to the probability of "not stocking out" in each period. (Alternative service measures could be used as well. For example,

consider

$$\frac{E[\min\{S_i(t), CP_i(t)\}]}{E[S_i(t)]} \geq \hat{\eta} \qquad \text{or} \qquad Prob\left(\frac{\min\{S_i(t), CP_i(t)\}}{S_i(t)} \geq \zeta\right) \geq \check{\eta}$$

for appropriate values of $\hat{\eta}$, $\check{\eta}$ and ζ. In both cases, the left-hand-sides are nondecreasing in $CP_i(t)$, and one can carry out an analysis similar to the one presented in later sections.)

In addition, we note that $[CP_i(t) - S_i(t)]^+$ is the amount of inventory carried at the end of period t. Because the target service level is quite high, in this paper, we approximate $[CP_i(t) - S_i(t)]^+$ with $CP_i(t) - S_i(t)$. We recognize that the quality of this approximation depends on the tail behavior of the random variable $S_i(t)$. However, this simplification has been used in the inventory and production literature (Bitran and Yanasse, 1984; Bookbinder and Tan, 1988; Bitran and Leong, 1990). Then, we approximate the holding cost for product i in period t by $h_i \cdot (CP_i(t) - S_i(t))$, where h_i is the cost of holding one unit of product i for one period. For high values of η_i, we provide three arguments that our approximation is quite reasonable. First, if $S_i(t)$ is bounded above by $\overline{S}_i(t)$, then

$$\begin{aligned} E[CP_i(t) - S_i(t)]^+ - E[CP_i(t) - S_i(t)] &= E[S_i(t) - CP_i(t)]^+ \\ &= E[S_i(t) - CP_i(t) | S_i(t) > CP_i(t)] \cdot P[S_i(t) > CP_i(t)] \\ &\leq [\overline{S}_i(t) - CP_i(t)] \cdot (1 - \eta_i) \ . \end{aligned}$$

Second, if $S_i(t)$ is uniformly distributed, we can show that the relative error of the inventory amount is given by

$$\frac{(1-\eta_i)^2}{\eta_i^2} \tag{3.8}$$

which is only 0.28% when $\eta_i = 0.95$. (See Appendix A.1 for the proof of (3.8).) Note that this is a second-order quantity in terms of the product service level requirement. Third, if $S_i(t)$ is normally distributed with the coefficient of variation (standard deviation divided by mean) of 0.5, the relative error can be computed to be at most 1.06% when $\eta_i = 0.95$. (See Figure 6; the derivation of the relative error under normal distribution follows Chopra and Meindl (2003), Appendix 11C.)

4 Capacity Planning with Open-Loop Production

In this section, we study the capacity planning problem for a product transition described in Section 3, where both capacity and production decisions are determined at the beginning of the horizon, i.e., open-loop production. This is also referred to as the "static uncertainty" or "zero-order rule" production (Bookbinder and Tan, 1988; Charnes and Cooper, 1963). For this problem,

the optimization problem under demand uncertainty can be transformed into a linear programming formulation.

We state the following problem Γ_1, to determine the capacity purchase and conversion plan that minimizes the capacity investment costs and related inventory holding costs:

$$\Gamma_1 \equiv \min \sum_{t=1}^{T}[Q_c(t)\cdot\delta\cdot c_c\cdot(T-t)+Q_b(t)\cdot\delta\cdot c_b\cdot(\tau+T-t)] + E\sum_{t=1}^{T}\sum_{i\in\{1,2\}} h_i\cdot[CP_i(t)-S_i(t)]$$
$$\text{s. t.} \quad (3.1) \text{ to } (3.7) ,$$

where the expectation is taken over $S_i(t)$, for $t = 1,\ldots,T$. All of the decision variables are nonnegative. (In this paper, we assume that the number of purchases and conversions can be fractional as it makes it easier to illustrate the tradeoffs we explore; we could incorporate the integrality requirement into our formulation at the expense of increased computational effort.)

In the above formulation, the chance constraint (3.7) can be transformed to be linear by defining minimum cumulative production requirement, $s_i(t)$, as the smallest number such that

$$Prob(S_i(t) \le s_i(t)) = \eta_i , \qquad \text{for } i \in \{1,2\} . \tag{4.1}$$

Then, constraint (3.7) is equivalent to the following linear constraint:

$$CP_i(t) \ge s_i(t) . \tag{4.2}$$

Furthermore, we define $d_i(t) = s_i(t) - s_i(t-1)$ for each t, and then replace (3.5) to (3.7) with:

$$I_i(t) = I_i(t-1) + \sum_{j=A,B} P_i^j(t) - d_i(t) \tag{4.3}$$
$$I_i(t) \ge 0 , \tag{4.4}$$

where $I_i(t)$ represents $CP_i(t) - s_i(t)$, and we set $P_2^A(t) = 0$ for notational convenience. We can now decompose the inventory in Γ_1 into two parts

$$CP_i(t) - S_i(t) = [s_i(t) - S_i(t)] + [CP_i(t) - s_i(t)] = [s_i(t) - S_i(t)] + I_i(t) ,$$

where the first term in the rightmost expression represents the mismatch between the minimum cumulative production quantity and cumulative demand, corresponding to a safety stock, and the second term is additional inventory produced above the minimum cumulative production requirement. Since $I_i(t) = CP_i(t) - s_i(t)$ does not depend on the cumulative demand random variables, $I_i(t)$ can be considered a decision variable and can be viewed as additional inventory used to balance production.

Hence, we can write Γ_1 as a linear program below:

$$\Gamma_2 \equiv \min \quad \sum_{t=1}^{T}[Q_c(t)\cdot\delta\cdot c_c\cdot(T-t) + Q_b(t)\cdot\delta\cdot c_b\cdot(\tau+T-t)] + \sum_{t=1}^{T}\sum_{i\in\{1,2\}} h_i\cdot I_i(t)$$
$$\text{s. t.} \quad (3.1) \text{ to } (3.4), (4.1) \text{ and } (4.3) \text{ to } (4.4) ,$$

where the decisions variables are $Q_b(t)$, $Q_c(t)$, $P_1^A(t)$, $P_i^B(t)$, $K^A(t)$, $K^B(t)$, and $I_i(t)$.

We make a few remarks regarding the interpretation of the linear program Γ_2. First, it solves for the optimal capacity purchasing and conversion schedule that meets a set of *deterministic* demand $d_i(t)$'s with minimum cost. That is, we treat the production requirement $d_i(t)$ as if this is the deterministic demand quantities that the company has to produce (instead of considering demand as stochastic). Thus, if the company had an existing approach or software that solves the deterministic version of this problem, incorporating stochastic modeling of demand can easily be accommodated (assuming a chance-constrained service requirement).

Second, this formulation captures some benefit from flexible capacity; the company can use the same unit of flexible capacity to produce one product in one period and another product in the next period without adjusting capacities. This is particularly relevant when one product has decreasing demand while the demand of the other product increases over time. However, this formulation does not explicitly capture the benefit of risk pooling that may result from delaying production decisions; we explore this issue in the next section.

5 Capacity Planning with Closed-Loop Production

The formulation Γ_2 that we developed in Section 4 does not capture the risk-pooling benefit arising from closed-loop production and flexible capacity. For example, if the manager has overestimated the speed of product transition from the old to the new product, it is likely that, *ex post*, Configuration-A capacity has been converted to Configuration B too fast; nonetheless, since the Configuration-B capacity has been retro-fitted to be flexible, it can still be used to produce the old product. Yet, the capacity planning model Γ_2 does not account for this flexibility, namely the value from being able to dynamically allocate the flexible Configuration-B capacity.

One way to account for this benefit is to integrate the dynamic production allocation decisions into the capacity planning model, for instance by means of dynamic programming. The resulting dynamic programming formulation for production decisions, however, is computationally prohibitive to solve. The complexity arises from multiple compounding factors such as demand uncertainty, capacity constraints and flexible capacity. To dynamically determine production quantities, one has

to consider the current inventory position of both products, as well as the period-by-period capacity plan and demand uncertainties for both products over the remainder of the planning horizon. This remains a difficult problem. Evans (1967) establishes that the form of the optimal structure follows an order-up-to-vector policy, but his model does not consider the service level constraints, and it is difficult to obtain the state-dependent order-up-to vector. Hausman and Peterson (1972) take a heuristic approach for allocating capacity among multiple items including one that aims to equate the service levels among the products. More recently, Janakiraman et al. (2008) emphasize the theoretical and computational challenges of allocating limited capacity among multiple items and propose a weighted-balancing-rule heuristic (which allocates capacity based on weighted shortfalls for each product).

Our problem differs from that addressed in these papers: (i) we allow time-varying demand, accounting for the new product gradually replacing the old product, and (ii) our main focus is the more strategic decisions of capacity planning. We consider production decisions only to the extent that they are useful for informing these capacity decisions.

5.1 Single-Period Illustration

To present the basic idea of how we account for the risk pooling benefit of closed-loop production, we first consider a single-period problem involving flexible capacity. Since there is only one time period, we drop the time index. We assume that demand of the new product (product 2) has higher priority than the old product (product 1). Hence, the production decisions, for given fixed capacity K^A and K^B and given realized demands S_1 and S_2, are as follows:

$$P_1 \;=\; \min\left\{\mu r_1^A K^A + \frac{r_1^B}{r_2^B}(\mu r_2^B K^B - S_2)^+,\; S_1\right\} \quad \text{and} \quad P_2 \;=\; \min\left\{\mu r_2^B K^B,\; S_2\right\}\;, \tag{5.1}$$

where μ is the number of machine hours per line of capacity per time period, and r_1^A, r_1^B and r_2^B are the rates of production. Above, we note that any excess Capacity B can be allocated to product 1 after observing demand; due to the production rate difference, we need to multiply the term $(\mu r_2^B K^B - S_2)^+$ with the factor $\frac{r_1^B}{r_2^B}$. Then, the service level requirements become:

$$Prob\left(S_1 \le s_1 + \frac{r_1^B}{r_2^B}(s_2 - S_2)^+\right) \;\ge\; \eta_1 \qquad \text{and} \tag{5.2}$$

$$Prob(S_2 \le s_2) \;\ge\; \eta_2\;, \tag{5.3}$$

where we let $s_1 \equiv \mu r_1^A K^A$ and $s_2 \equiv \mu r_2^B K^B$.

Note that conditions (5.2) and (5.3) identify a set of conditions that quantities s_1 and s_2 must satisfy. We treat (s_1, s_2) as static quantities and refer to them as the *cumulative production target*

quantities, i.e., if we plan the capacities according to (s_1, s_2), we would be able to meet the actual demand with the required service level using this capacity plan. Note that (s_1, s_2) has a one-to-one correspondence with capacities (K^A, K^B). Thus, to find the optimal capacities, it suffices to find the pair (s_1, s_2) with the minimum cost while satisfying constraints (5.2) and (5.3).

Since η_2 is high, it is reasonable to replace $(s_2 - S_2)^+$ with $s_2 - S_2$, and we approximate (5.2) with

$$Prob\left(S_1 + \frac{r_1^B}{r_2^B} S_2 \leq s_1 + \frac{r_1^B}{r_2^B} s_2\right) \;\geq\; \eta_1 \;. \tag{5.4}$$

We note that (5.4) is more conservative than (5.2), i.e., any (s_1, s_2) satisfying (5.4) also satisfies (5.2). Furthermore, for any s_2 satisfying (5.3), the difference between (5.2) and (5.4) is bounded by

$$\begin{aligned}
&Prob\left(S_1 \leq s_1 + \frac{r_1^B}{r_2^B}(s_2 - S_2)^+\right) - Prob\left(S_1 + \frac{r_1^B}{r_2^B} S_2 \leq s_1 + \frac{r_1^B}{r_2^B} s_2\right) \\
&= \; Prob(S_2 > s_2) \cdot Prob\left(s_1 + \frac{r_1^B}{r_2^B}(s_2 - S_2) \leq S_1 \leq s_1 \mid S_2 > s_2\right) \\
&\leq \; (1 - \eta_2) \cdot Prob\left(s_1 - \frac{r_1^B}{r_2^B}(S_2 - s_2) \leq S_1 \leq s_1 \mid S_2 > s_2\right) \;.
\end{aligned}$$

For instance, if S_1 and S_2 are independent and uniformly distributed in $[0, \overline{S}_1]$ and $[0, \overline{S}_2]$, respectively, the conditional distribution, $(S_2 - s_2)|(S_2 > s_2)$, is uniformly distributed on $[0, \overline{S}_2 - s_2]$, where $\overline{S}_2 - s_2 \leq (1 - \eta_2)\overline{S}_2$ holds by (5.3). Then, the above expression is bounded above by

$$(1 - \eta_2) \cdot \left(\frac{r_1^B}{r_2^B}(1 - \eta_2)\overline{S}_2\right) \cdot \frac{1}{\overline{S}_1} \;=\; \frac{r_1^B}{r_2^B} \cdot \frac{\overline{S}_2}{\overline{S}_1} \cdot (1 - \eta_2)^2 \;,$$

which is a second-order quantity in terms of the new product service level requirement, η_2; see (3.8).

Based on the above approximation, to characterize the values of (s_1, s_2) satisfying (5.3) and (5.4), we replace these constraints with two linear constraints: $s_2 \geq F^{-1}(\eta_2)$ and $s_1 + \frac{r_1^B}{r_2^B} s_2 \geq G^{-1}(\eta_1)$, where F and G denote the distributions of S_2 and $S_1 + \frac{r_1^B}{r_2^B} S_2$, respectively. We define the feasible set

$$\begin{aligned}
\mathcal{S}(S_1, S_2) \;&=\; \{(s_1, s_2) \in \Re^+ \times \Re^+ \text{ satisfying (5.3) and (5.4)}\} \\
&=\; \left\{(s_1, s_2) \in \Re^+ \times \Re^+ \;\mid\; s_1 + \frac{r_1^B}{r_2^B} s_2 \geq G^{-1}(\eta_1) \;,\; s_2 \geq F^{-1}(\eta_2)\right\} \;.
\end{aligned} \tag{5.5}$$

Any capacity plan that corresponds to some $(s_1, s_2) \in \mathcal{S}(S_1, S_2)$ can achieve the desired service levels under a production rule in which meeting the new product demand takes priority over the

old product. Note that the problem of finding the value of (s_1, s_2) minimizing any linear objective function (for example, one similar to the single-period version of the objective function in Γ_2) subject to $\mathcal{S}(S_1, S_2)$ is a deterministic linear programming problem.

Lemma 1. *(a) $\mathcal{S}(S_1, S_2) \subseteq \Re^2_+$ is a convex set. Furthermore, if $(\hat{s}_1, \hat{s}_2) \in \mathcal{S}(S_1, S_2)$, then both $(s_1, \hat{s}_2)$ and $(\hat{s}_1, s_2)$ belong to $\mathcal{S}(S_1, S_2)$ for any $s_1 \geq \hat{s}_1$ and $s_2 \geq \hat{s}_2$.*

(b) The set of extreme points in $\mathcal{S}(S_1, S_2)$ is

$$\begin{cases} \left\{ \left(0, \frac{r_2^B}{r_1^B} G^{-1}(\eta_1)\right), \ \left(G^{-1}(\eta_1) - \frac{r_1^B}{r_2^B} F^{-1}(\eta_2), F^{-1}(\eta_2)\right) \right\} & \text{if } F^{-1}(\eta_2) \leq \frac{r_2^B}{r_1^B} G^{-1}(\eta_1), \\ \left\{ \left(0, F^{-1}(\eta_2)\right) \right\} & \text{otherwise.} \end{cases}$$

Proof. These results follow directly from (5.5). □

Lemma 1(a) implies that, from any $(\hat{s}_1, \hat{s}_2) \in \mathcal{S}(S_1, S_2)$, the set $\mathcal{S}(S_1, S_2)$ is unbounded in the direction of $(0, 1)$ or $(1, 0)$ as well as any of their convex combination. Also, from the boundaries of $\mathcal{S}(S_1, S_2)$, we observe a complementarity relationship between the cumulative production target quantities, s_1 and s_2: when a target quantity for one product is high, the target quantity for the other product can be low. Furthermore, by part (b) of the above lemma, the optimal solution to the linear program, if it exists, can be restricted to a line segment (convex combination of extreme points in the first case) or a single point (the second case).

We now compare the cumulative production target quantities studied in this section (closed-loop planning), which we denote by $(s_1^C, s_2^C) \in \mathcal{S}(S_1, S_2)$, and the minimum cumulative production requirement of Section 4 (open-loop planning), which we denote by (s_1^O, s_2^O) and is given by $Prob(S_1 \leq s_1^O) = \eta_1$ and $Prob(S_2 \leq s_2^O) = \eta_2$; these two constraints represent the counterpart of $\mathcal{S}(S_1, S_2)$ in the open-loop planning model. The feasible region for (s_1^O, s_2^O) can be specified with simple lower bound constraints on each of s_1^O and s_2^O, but the feasible region for (s_1^C, s_2^C) includes a linear equation involving both variables. In open-loop planning, since production of each product is decided before demand realization, s_1^O and s_2^O represent the minimum production quantities for old and new products, respectively; in closed-loop planning, s_1^C and s_2^C act as a *proxy* for production targets for the purpose of capacity planning, and the actual production quantities depend on demand realization.

The following results show that the feasible region in open-loop planning is *approximately* more restrictive than in closed-loop planning. More specifically, we show that the feasible region in open-loop planning is more restrictive than the constraints given in (5.2) and (5.3), and that it is more restrictive than a modified version of (5.4), where η_1 is replaced with $\eta_1\eta_2$ (at high service levels, $\eta_1\eta_2$ is a reasonable approximation of η_1).

Lemma 2. *If* (s_1, s_2) *satisfies* $Prob(S_1 \leq s_1) = \eta_1$ *and* $Prob(S_2 \leq s_2) = \eta_2$*, then it satisfies* (5.2) *and* (5.3)*. Furthermore, if* S_1 *and* S_2 *are independent, then* (s_1, s_2) *satisfies*

$$Prob\left(S_1 + \frac{r_1^B}{r_2^B} S_2 \leq s_1 + \frac{r_1^B}{r_2^B} s_2\right) \;\geq\; \eta_1 \eta_2 \;.$$

Proof. Clearly, $Prob(S_2 \leq s_2) = \eta_2$ implies (5.3). Also, $Prob(S_1 \leq s_1) = \eta_1$ implies (5.2) since $(s_2 - S_2)^+$ is nonnegative. Now, if S_1 and S_2 are independent, then by conditioning on the event $s_2 - S_2 \geq 0$, we obtain

$$\begin{aligned} Prob\left(S_1 + \frac{r_1^B}{r_2^B} S_2 \leq s_1 + \frac{r_1^B}{r_2^B} s_2\right) &= Prob\left(S_1 \leq s_1 + \frac{r_1^B}{r_2^B}(s_2 - S_2)\right) \\ &\geq Prob\left(S_1 \leq s_1\right) \cdot Prob\left(s_2 - S_2 \geq 0\right) \;=\; \eta_1 \eta_2 \;, \end{aligned}$$

as required. □

5.2 Multi-Period Formulation

A multi-period formulation with service level constraints is difficult to analyze even for a production-inventory control problem without any consideration of capacity issues. In this section, we present a heuristic approach based on the single-period solution from Section 5.1 by decomposing the multiple-period problem to a series of single-period problems.

More precisely, we consider a series of single-period problems, one for each $t \in \{1, 2, \ldots, T\}$. We aggregate demand in periods 1 through t, apply the notation of the static production quantities for the aggregated demand. Recall that $S_i(t)$ represents the cumulative demand for product i, i.e., $S_i(t) = D_i(1) + \cdots + D_i(t)$, where $D_i(\tau)$ corresponds to the demand of product i in period $\tau \in \{1, \ldots, t\}$. We define a set of static cumulative production target quantities $s_i(t)$ that meet the required service levels for $(S_1(t), S_2(t))$, i.e.,

$$(s_1(t), s_2(t)) \in \mathcal{S}(S_1(t), S_2(t)), \tag{5.6}$$

where $\mathcal{S}(S_1(t), S_2(t))$ is defined as in (5.5).

The following lemma formalizes an intuitive result that the cumulative production target quantities are increasing in time. It also shows that the feasible region restricted to each period eventually becomes the first case considered in Lemma 1(b). Let F_t and G_t denote the distributions of $S_2(t)$ and $S_1(t) + \frac{r_1^B}{r_2^B} S_2(t)$, respectively. The proof of Lemma 3 is based on the Central Limit Theorem, and appears in Appendix A.2. As time increases, the cumulative demand distributions eventually become dominated by the respective mean values (according to the Central Limit Theorem) and thus the first case in Lemma 1(b) holds for sufficiently large t.

Lemma 3. *(a)* $\mathcal{S}(S_1(t), S_2(t)) \supseteq \mathcal{S}(S_1(t+1), S_2(t+1))$ *for any* $t \geq 1$.

(b) Suppose that $\{D_i(t)|t = 1, 2, \ldots\}$ *is independent and identically distributed with mean* $\theta_i > 0$ *and standard deviation* σ_i*, for* $i \in \{1, 2\}$*, and* $D_1(t)$*'s and* $D_2(t)$*'s are independent. Then, there exists* $\overline{t}$ *such that for any* $t \geq \overline{t}$*, we have* $F_t^{-1}(\eta_2) \leq \frac{r_2^B}{r_1^B} G_t^{-1}(\eta_1)$.

(Part (b) of Lemma 3 can be extended to non-stationary demand settings.)

Now, having defined a feasible region for each $(s_1(t), s_2(t))$, we can introduce the objective function that represents the total equipment and holding costs. The optimization determines both the capacity plan and the cumulative production targets $(s_1(t), s_2(t))$ for all $t = 1, 2, \ldots, T$. Compared to the analysis in Section 4 where production is determined before demand realization and thus risk pooling is ignored, we incorporate capacity flexibility into our model through the use of the set $\mathcal{S}(S_1(t), S_2(t))$. Our heuristic approach is to use the solution to the deterministic optimization problem shown below:

$$
\begin{aligned}
\Gamma_3 \;\equiv\; \min \quad & \sum_{t=1}^{T} [Q_c(t) \cdot \delta \cdot c_c \cdot (T-t) \;+\; Q_b(t) \cdot \delta \cdot c_B \cdot (\tau + T - t)] \\
& + \sum_{t=1}^{T} [h_1 I_1(t) + h_2 I_2(t) + h_1 E(s_1(t) - S_1(t)) + h_2 E(s_2(t) - S_2(t))] \\
\text{s. t.} \quad & \text{(3.1) to (3.4), and (4.3) to (4.4)} , \\
& (s_1(t), s_2(t)) \in \mathcal{S}(S_1(t), S_2(t)) \qquad \text{for each } t,
\end{aligned}
$$

where the decisions variables are $Q_b(t)$, $Q_c(t)$, $P_1^A(t)$, $P_i^B(t)$, $K^A(t)$, $K^B(t)$ and $I_i(t)$, in addition to $s_1(t)$ and $s_2(t)$. In this formulation, we take $(s_1(t), s_2(t))$ as a proxy for required cumulative production quantities for products 1 and 2, and therefore as a proxy for the required capacities. Since the set $\mathcal{S}(S_1(t), S_2(t))$ defines a set of linear constraints (see equation (5.5)), problem Γ_3 is a linear program.

We can show that an optimal value of $(s_1(t), s_2(t))$ is a convex combination of the extreme points of $\mathcal{S}(S_1(t), S_2(t))$ even though $\mathcal{S}(S_1(t), S_2(t))$ is unbounded. See Lemma 1(b) for the characterization of extreme points. In other words, we can restrict the choice of $(s_1(t), s_2(t))$ to either the unique extreme point if there is only one extreme point in $\mathcal{S}(S_1(t), S_2(t))$, or otherwise on the line segment between two extreme points. This observation can be used to reduce the dimension of the problem. To show this result, we note that the holding cost h_1 is charged to both the "safety stock", $E(s_1(t) - S_1(t))$, and any quantity produced above the target quantity, $I_1(t) = CP_1(t) - s_1(t)$, and the sum of these quantities is invariant of the value of $s_1(t)$. Thus, $s_1(t)$ can be made as small

as possible without affecting optimality subject to the $(s_1(t), s_2(t)) \in \mathcal{S}(S_1(t), S_2(t)))$ constraint. A similar argument holds for the holding cost h_2 of product 2.

The difference between Γ_3 (closed-loop planning) and Γ_2 of Section 4 (open-loop planning) lies in constraints involving $(s_1(t), s_2(t))$, for each t. As illustrated in the single-period discussion (Section 5.1), the feasible region of Γ_2 is generally more restrictive than Γ_3 at high service levels. The results in Lemma 2 carry through in the multi-period formulation.

While the problem Γ_3 is still a deterministic optimization problem, our use of $(s_1(t), s_2(t))$ accounts for some risk pooling in closed-loop production using flexible capacity. Similar to the discussion in Section 5.1, the quantities $s_1(t)$ and $s_2(t)$ are essentially capacity measures that correspond to the cumulative capacity requirement of Configuration A and Configuration B production lines, respectively. The feasible region defined by $\mathcal{S}(S_1(t), S_2(t))$ accounts for the risk-pooling effect by coupling the capacity requirements of the two products. We note that the quantities $(s_1(t), s_2(t))$ are *approximate* values for the required capacity since we do not explicitly capture *dynamic* production adjustments that may be allowed in each period. However, treating them conveniently as cumulative "production" quantities in the *deterministic* formulation of problem Γ_3 allows us to solve for a capacity plan that is capable of meeting the service level requirement, a problem that is otherwise difficult to solve.

6 Application

In this paper, we have described two approaches for solving the capacity planning problem. The open-loop planning approach, shown in formulation Γ_2, may result in over-investment of capacity by ignoring the potential risk-pooling benefit from postponing the production (Section 4) while the closed-loop planning approach, based on formulation Γ_3, approximately captures the benefit of production postponement through a set of cumulative production target quantities (Section 5). In this section, we apply these two solution approaches to a case motivated by Intel. The case involves two product families with one family gradually replacing the other. The company wishes to learn what the capacity plan should be if they have a forecast that reflects the actual demand trend, given the uncertainty in their forecast process. Figure 2 shows the monthly demand forecast. In this example, the standard deviation for the monthly forecast error is estimated to be 30% of the mean demand for the new product and 20% for the old product. For the purpose of numerical investigation, we assume that the demand distributions are Normal and independent.

The equipment configurations, costs, and production rates are given in Table 1. The conversion

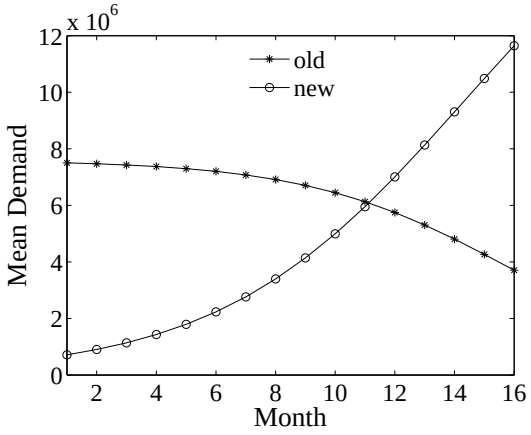


Figure 2: Forecast of Demand

Equipment Configuration	Tools	Tool Cost	Production Rate (units/hr)
A	a	$\$0.391M$	$r_1^A = 2450$
	b	$\$1.110M$	
	c	$\$0.425M$	
B	a	$\$0.391M$	$r_1^B = 3000$
	b′	$\$1.800M$	$r_2^B = 3200$
	c	$\$0.425M$	

Table 1: Equipment Costs and Production Rates

involves replacing tool b with tool b', which costs \$1.8 Million; a brand-new line of Configuration B (including tools a, b' and c) costs about \$2.62 Million. The production rate of the new product (r_2^B) is higher than the old product (r_1^B). In addition, the production rate for the old product is higher on Configuration B (r_1^B) than on Configuration A (r_1^A). The depreciation rate is 2% of the original purchase price of the equipment per month. The conversion lead time and the installation lead time for a brand new line are both one month. Service requirement is 98%. Machine hours per month are 720 (here, equipment down time is factored into the production rate). Holding cost per unit of product is based on the value-added by this production module, and is equal to \$0.01 per month for both products. In addition, we assume that at the beginning of the time horizon there are six Configuration-A lines and one Configuration-B line. As mentioned before, we treat the decision variables $Q_c(t)$ and $Q_b(t)$ as fractional quantities.

In Section 6.1, we demonstrate the advantage of the closed-loop planning approach over open-loop planning. We then examine the capacity plan details in Section 6.2 to uncover two key tradeoffs that affect the capacity paths over time.

6.1 Performance Comparison of the Two Approaches

In order to evaluate each capacity plan, we need to address in more detail how the production will be carried out, namely how we will simulate the production decisions for a given capacity plan. As mentioned earlier, the problem of finding the optimal production policy under service level constraints in a dynamic setting is known to be difficult. As a result, we use a dynamic production policy that is simple to describe and approximates the actual production practice. The policy is based on *a time-varying inventory target*: in each period, we decide on the production quantity that will bring the inventory position as close as possible to the target inventory level. (Note that the target inventory may not be reached because of capacity availability or excess carried-over inventory.) This production rule has some theoretical appeal because the optimal production policy under a cost-minimization setting (without service level constraints) follows this form (Evans, 1967). The target inventory level is based on a heuristic that takes advantage of the cumulative production target quantities $s_i(t)$, which are an output of the optimization problem for capacity planning. In particular, we set the target inventory, which we denote by $y_i(t)$, to the expected amount of inventory we would have in period t, which is $s_i(t) - E[S_i(t)]$ where $E[S_i(t)]$ is the mean cumulative demand for product i by time t and $s_i(t)$ is an output from the capacity planning model.

The following algorithm states the dynamic production policy used in the simulation: Let $X_i(t)$ be the realized demand for product i in period t and let $I_i(t)$ be the inventory or backordes (if negative) of product i. We use $LK^j, j = \{A, B\}$ to denote leftover capacity.

Algorithm 1 (Dynamic Production Policy)**.** *For each period t, we decide production using the following procedure:*

Step 1. Fill the backorders and current period demand of the new product (using Capacity B) and compute leftover Capacity B.

$$\begin{aligned} P_2(t) &\leftarrow \min\{[X_2(t) - I_2(t-1)]^+,\ \mu \cdot r_2^B(t) \cdot K^B(t)\}\ , \\ LK^B &\leftarrow [K^B(t) - (X_2(t) - I_2(t-1))/(\mu \cdot r_2^B(t))]^+. \end{aligned}$$

Step 2. Fill the backorders and current period demand of the old product (using Capacity B only *if Capacity A has been exhausted).*

$$\begin{aligned} P_1(t) &\leftarrow \min\{[X_1(t) - I_1(t-1)]^+,\ \mu \cdot r_1^B(t) \cdot LK^B + \mu \cdot r_1^A(t) \cdot K^A(t)\}\ , \\ LK^A &\leftarrow [K^A(t) - (X_1(t) - I_1(t-1))/(\mu \cdot r_1^A(t))]^+\ , \\ LK^B &\leftarrow \left\{LK^B - [X_1(t) - I_1(t-1) - \mu \cdot r_1^A(t) \cdot K^A(t))]^+/(\mu \cdot r_1^B(t))\right\}^+\ . \end{aligned}$$

Step 3. If $LK^B > 0$, produce to meet the inventory target of the new product $y_2(t)$ (using Capacity B).

$$\begin{aligned} P_2(t) &\leftarrow P_2(t) + \min\{y_2(t),\ \mu \cdot r_2^B(t) \cdot LK^B\}\ , \\ LK^B &\leftarrow [LK^B - y_2(t)/(\mu \cdot r_2^B(t))]^+\ . \end{aligned}$$

Step 4. If $LK^A > 0$ or $LK^B > 0$, produce to meet the inventory target of the old product $y_1(t)$ (again using Capacity B only *if Capacity A has been exhausted).*

$$P_1(t) \quad \leftarrow \quad P_1(t) + \min\{y_1(t),\ \mu \cdot r_1^B(t) \cdot LK^B + \mu \cdot r_1^A(t) \cdot LK^A\} \ .$$

Step 5. Update the inventory and repeat for the next time period.

$$\begin{aligned} I_1(t) \quad &\leftarrow \quad I_1(t-1) + P_1(t) - X_1(t) \ , \\ I_2(t) \quad &\leftarrow \quad I_2(t-1) + P_2(t) - X_2(t) \ . \end{aligned}$$

The dynamic production policy described in Algorithm 1 is not an optimal policy, which is complex to characterize and to compute. It is a reasonable dynamic policy used to evaluate the capacity plans.

We then simulate the monthly demand from month 1 to month 16 according to the demand forecast in Figure 2 and the estimated uncertainty. For each set of demand realization and each capacity plan, we allocate production according to the policy described above. The service level for each product at each time period is recorded by counting the percentage of time that demand is met from stock and we average the service levels over the 16 months. We repeat this simulation for $10,000$ demand sets. We report the average service level for each product, the average total cost, and the 95% confidence intervals of the cost. Tables 2 and 3 present the performance comparison as the target service level and the unit holding cost vary respectively. We evaluate the open-loop capacity plan using two policies: the static production policy (where production is determined at the beginning of the planning horizon according to Γ_2 and is not adjusted as demand is realized) and the dynamic production policy (following Algorithm 1). The closed-loop capacity plan is evaluated with the dynamic production policy only (since it is not meaningful to evaluate a capacity plan that considers closed-loop production with a static production policy).

We make several observations. First, the open-loop capacity plan with static production yields the targeted service levels; this is expected since the capacity plan derived under the assumption of open-loop production involves no approximation for the service levels. This serves as a validation for our implementation. Second, both capacity plans provide service levels that exceed the target under the dynamic production rule and the default holding cost $h = 0.01$ (Table 2). However, the closed-loop capacity plan demonstrates more "balanced" service levels between the old and the new products (i.e., a smaller gap between the service levels for the two products), This may be due to the fact that the closed-loop capacity plan considers dynamic capacity allocation, which is consistent with the dynamic production rule. Third, the closed-loop capacity plan results in lower total cost than the open-loop capacity plan. As the target service level changes from 95% to 99%, the savings

from the closed-loop plan increases from $\$0.44M$ (18% of the total cost) to $\$0.66M$ (21% of the total cost). Finally, under very low unit holding cost, as shown in Table 3 for cases $h = 0.001$ and $h = 0.0001$, the resulting service level for the old product using the open-loop capacity plan falls significantly below the target level whereas the closed-loop plan achieves the required service level. These observations indicate that incorporating closed-loop production into capacity planning is preferable to the capacity planning that ignores this. We have also tested the capacity plans using other reasonable production policies including one that exhausts the available capacity in each period, and the closed-loop planning approach consistently shows better performance than the open-loop planning approach, indicating that our observations are robust with respect to the choice of the production policy that we have used.

6.2 Capacity Plan

We now consider the optimal capacity paths obtained under both approaches. In our canonical example illustrated by Figure 3(d) ($h = 0.01$), incorporating closed-loop production leads to a delayed and smaller buildup of flexible capacity, as well as delayed and slightly fewer conversions of existing Configuration-A lines. This is further demonstrated in Figure 4. Table 4 provides a summary of the total number of conversions and new purchases, as well as the corresponding costs. (Note that these costs are derived from problems Γ_2 and Γ_3 and are not the simulated costs.)

It is no surprise that the consideration of closed-loop production reduces the overall capacity cost. In particular, the closed-loop planning solution delays and reduces purchases of new Configuration-B lines even as the unit holding cost varies (Figure 5). Therefore, ignoring the risk-pooling effect in closed-loop production results in a purchase plan that is too aggressive.

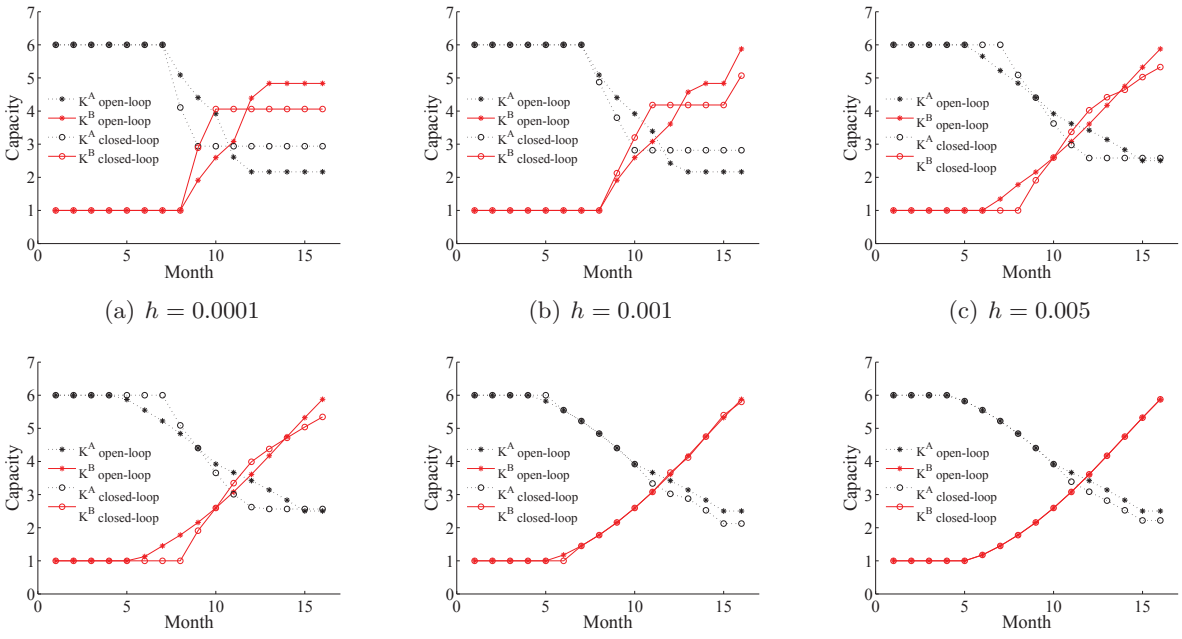
Incorporating closed-loop production consideration, however, could result in earlier *or* later conversions, as illustrated in Figure 3. There are two major tradeoffs in this capacity planning problem. (1) To build up Configuration B capacity, the company faces a choice between conversions and purchases. Given that Configuration-B capacity is flexible and can produce both products, it should be less risky to convert equipment under closed-loop production than open-loop production, arguing for more aggressive conversions under closed-loop production. (2) The second tradeoff is based on the contrast between the chase strategy (the capacity buildup follows the demand pattern gradually) and the level strategy (stable capacity is maintained through time except "quick" capacity changes). Allowing production adjustment after demand realization delays initial capacity investment because of the risk-pooling effect; but because of insufficient capacity or inventory later in time, it accelerates capacity build-up. These two tradeoffs impact the capacity path and

Target Service Level	Capacity Plan	Open-loop Planning		Closed-loop Planning
	Production	Static	Dynamic	Dynamic
99%	ave. SL 1	99.04%	99.99%	99.94%
	ave. SL 2	98.98%	99.23%	99.42%
	total cost ($M)	3.46([2.23, 4.73])	3.16([2.71, 3.42])	2.50([2.01, 2.78])
98%	ave. SL 1	98.00%	99.99%	99.87%
	ave. SL 2	98.00%	98.41%	98.83%
	total cost ($M)	3.14([1.93, 4.40])	2.85([2.43, 3.10])	2.28([1.81, 2.55])
97%	ave. SL 1	96.91%	99.99%	99.80%
	ave. SL 2	97.02%	97.53%	98.22%
	total cost ($M)	2.94([1.77, 4.20])	2.66([2.25, 2.90])	2.14([1.69, 2.42])
95%	ave. SL 1	94.85%	99.99%	99.60%
	ave. SL 2	95.05%	95.68%	96.87%
	total cost ($M)	2.67([1.55, 3.92])	2.40([2.02, 2.64])	1.96([1.54, 2.23])

Table 2: Performance Comparison ($h = 0.01$)

Holding cost	Capacity Plan	Open-loop Planning		Closed-loop Planning
	Production	Static	Dynamic	Dynamic
$h = 0.0001$	ave. SL 1	98.93%	99.77%	99.94%
	ave. SL 2	99.31%	88.13%	99.43%
	total cost ($M)	0.903([0.891, 0.916])	0.889([0.884, 0.894])	0.890([0.884, 0.893])
$h = 0.001$	ave. SL 1	98.84%	99.88%	99.99%
	ave. SL 2	99.05%	86.72%	99.31%
	total cost ($M)	1.16([1.04, 1.29])	1.04([0.99, 1.09])	1.03([0.98, 1.06])
$h = 0.005$	ave. SL 1	98.00%	99.99%	99.87%
	ave. SL 2	98.26%	97.70%	98.85%
	total cost ($M)	2.07([1.47, 2.70])	1.91([1.69, 2.04])	1.59([1.36, 1.73])
$h = 0.01$	ave. SL 1	98.00%	99.99%	99.87%
	ave. SL 2	98.00%	98.41%	98.83%
	total cost ($M)	3.14([1.93, 4.40])	2.85([2.43, 3.10])	2.28([1.81, 2.55])
$h = 0.1$	ave. SL 1	98.00%	99.99%	99.94%
	ave. SL 2	97.96%	98.51%	98.39%
	total cost ($M)	22.3([10.3, 35.0])	19.5([15.2, 22.0]	14.6([10.4, 17.1])
$h = 0.5$	ave. SL 1	98.00%	99.99%	99.95%
	ave. SL 2	97.96%	98.51%	98.50%
	total cost ($M)	107.7([47.3, 170.8])	93.3([72.1, 105.8])	69.6([49.0, 81.7])

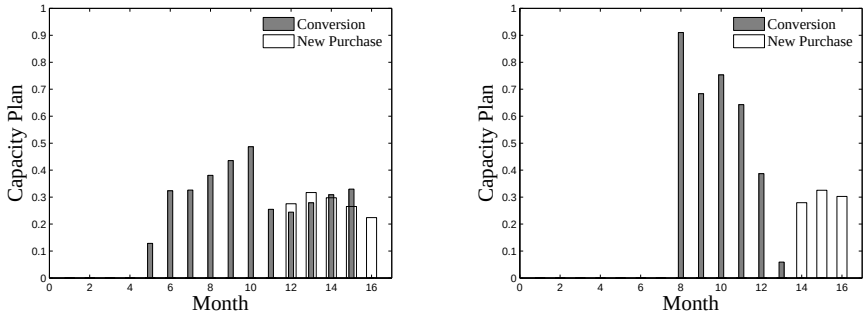
Table 3: Performance Comparison (Target service level is 98%)



(a) $h = 0.0001$ (b) $h = 0.001$ (c) $h = 0.005$

(d) $h = 0.01$ (default value) (e) $h = 0.1$ (f) $h = 0.5$

Figure 3: Capacity Path Comparison as Holding Cost Changes



(a) Open-loop Planning (b) Closed-loop Planning

Figure 4: Capacity Plan ($h = 0.01$)

Method	Open-loop Planning	Closed-loop Planning
Conversions	3.50	3.44
New Lines	1.38	0.91
Depreciation cost	$\$1.00M$	$\$0.91M$
Holding cost	$\$2.14M$	$\$1.63M$
Total Cost	$\$3.15M$	$\$2.54M$

Table 4: Capacity and Cost Comparison ($h = 0.01$)

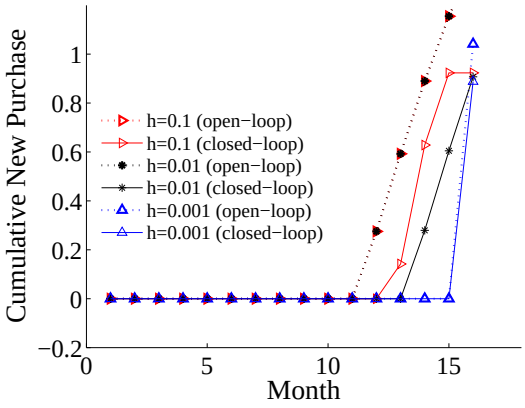


Figure 5: New Purchases

conversions in opposite ways. When the unit holding cost is low relative to equipment cost, a level strategy is favored (Figures 3(a) and 3(b)); when the unit holding cost is high, it is optimal to follow a chase strategy (Figures 3(e) and 3(f)).

7 Conclusion

We identify a crucial capacity planning problem in the context of a product transition, in which a new generation product replaces an older product. As the factory transitions the production from one product to the next, the capacity requirements also shift. In particular, capacity of the new equipment configuration needs to grow over time. The increasing capacity could be met by purchasing new lines, or by converting existing lines; the latter is the cheaper option. Capacity planning decides the schedule for equipment purchases and conversions, that satisfies a set of service level constraints under uncertain and time-varying demand. We develop two alternative approaches for determining the capacity plan: one in which we set the production decisions for the entire planning horizon before any demand is observed (which is a proxy for Intel's current approach) and one in which we approximate a dynamic production control policy. We show that although both methods generally meet the target service levels, the latter yields both lower equipment cost and lower inventory cost. We examine the performance of the methods as the holding cost varies, and as the target service level changes. This sensitivity analysis also reveals two major tradeoffs impacting the capacity decisions, namely, the choice between conversions and new purchases, and between a level versus chase strategy. These understandings are crucial for high-tech companies to make informed capacity decisions each time a product transition takes place. The method is shown to work well for Intel and we believe that it is applicable to many other manufacturing companies

who regularly manage product transitions.

Acknowledgement

The authors are grateful to Karl Kempf, Shamin Shirodkar, Dan Belton, Tom Rucker, Ajay Sevak, Naiping Keng and Jeff Pettinato from Intel Corporation for sharing their expert knowledge. We also thank Professor Michael Pinedo, as well as an anonymous Associate Editor and three anonymous reviewers for their constructive comments and suggestions.

References

Ahmed, S., and N. V. Sahinidis. 2003. Approximation Scheme for Stochastic Integer Programs Arising in Capacity Expansion. *Operations Research* 51 (3): 461–471.

Angelus, A., and E. L. Porteus. 2002. Simultaneous Capacity and Production Management of Short-Life-Cycle, Produce-to-Stock Goods under Stochastic Demand. *Management Science* 48 (3): 399–413.

Bean, J. C., J. R. Lohmann, and R. L. Smith. 1985. A Dynamic Infinite Horizon Replacement Economy Decision Model. *The Engineering Economist* 30 (2): 99–120.

Bean, J. C., J. R. Lohmann, and R. L. Smith. 1994. Equipment Replacement Under Technological Change. *Naval Research Logistics* 41 (2): 117–128.

Bean, J. C., and R. L. Smith. 1985. Optimal Capacity Expansion Over an Infinite Horizon. *Management Science* 31 (12): 1523–1532.

Bertsekas, D. P. 2000. *Dynamic Programming and Optimal Control, Second Edition*. Belmont, MA: Athena Scientific.

Bish, E. K., and Q. Wang. 2004. Optimal Investment Strategies for Flexible Resources, Considering Pricing and Correlated Demands. *Operations Research* 52 (6): 954–964.

Bitran, G. R., and T. Y. Leong. 1990. Deterministic Approximations to Co-Production Problems With Service Constraints. *MIT Working paper 3071-89-MS*.

Bitran, G. R., and H. H. Yanasse. 1984. Deterministic Approximations to Stochastic Production Problems. *Operations Research* 32 (5): 999–1018.

Bookbinder, J. H., and J. Tan. 1988. Strategies for the Probabilistic Lot-Sizing Problem with Service-Level Constraints. *Management Science* 34 (9): 1096–1108.

Bradley, J., and P. Glynn. 2002. Managing Capacity and Inventory Jointly in Manufacturing Systems. *Management Science* 48 (2): 273–288.

Bresnahan, T., and V. A. Ramey. 1994. Output Fluctuations at the Plant Level. *The Quarterly Journal of Economics* 109 (3): 593–624.

Çakanyildirim, M., R. Roundy, and S. Wood. 2004. Optimal Machine Capacity Expansions with nested Limitations under Demand Uncertainty. *Naval Research Logistics* 51 (2): 217–241.

Charnes, A., and W. W. Cooper. 1963. Deterministic Equivalents for Optimizing and Satisficing under Chance Constraints. *Operations Research* 11 (1): 18–39.

Chopra, and Meindl. 2003. *Supply Chain Management: Strategy, Planning and Operations, Third Edition*. New Jersey: Pearson Prentice Hall.

Eppen, G., R. Martin, and L. Schrage. 1989. A Scenario Approach to Capacity Planning. *Operations Research* 37 (4): 517–527.

Erlenkotter, D. 1974. Dynamic Programming Approach to Capacity Expansion With Specialization. *Management Science* 21 (3): 360–368.

Evans, R. V. 1967. Inventory Control of a Multiproduct System with a Limited Production Resource. *Naval Research Logistics Quarterly* 14 (2): 173–184.

Fine, C. H., and R. M. Freund. 1990. Optimal Investment in Product Flexible Manufacturing Capacity. *Management Science* 36 (4): 449–466.

Hausman, W. H., and R. Peterson. 1972. Multiproduct Production Scheduling for Style Goods with Limited Capacity, Forecast Revisions and Terminal Delivery. *Management Science* 18 (7): 370–383.

Huang, K., and S. Ahmed. 2009. The Value of Multistage Stochastic Programming in Capacity Planning Under Uncertainty. *Operations Research* 57 (4): 893–904.

Huh, W. T., and R. O. Roundy. 2005. A Continuous-Time Strategic Capacity Planning Model. *Naval Research Logistics* 52:329–343.

Huh, W. T., R. O. Roundy, and M. Cakanyildirim. 2006. A General Strategic Capacity Planning Model under Demand Uncertainty. *Naval Research Logistics* 53:137–150.

Janakiraman, G., M. Nagarajan, and S. Veeraraghavan. 2008. Simple Policies for Capacitated Multi-Product Inventory Systems. *Working paper*.

Jones, P. C., J. L. Zydiak, and W. J. Hopp. 1991. Parallel Machine Replacement. *Naval Research Logistics* 38:351–365.

Karabuk, S., and S. Wu. 2002. Coordinating strategic capacity planning in the semiconductor industry. *Opeations Research* 51 (6): 839–849.

Leone, R., and S. Bradley. 1982. Federal energy policy and competitive strategy in the US automobile industry. *Annual Review of Energy* 7:61–106.

Li, S., and D. Tirupati. 1994. Dynamic Capacity Expansion Problem with Multiple Products: Technology Selection and Timing of Capacity Additions. *Operations Research* 42 (5): 958–976.

Li, S., and D. Tirupati. 1995. Technology choice with stochastic demands and dynamic capacity allocation: A two-product analysis. *Journal of Operations Management* 12:239–258.

Luss, H. 1982. Operations research and capacity expansion problems: A survey. *Operations Research* 30 (5): 907–947.

Manne, A. S. 1961. Capacity Expansion and Probabilistic Growth. *Econometrica* 29 (4): 632–649.

McDonald, C. J. 1998. The Evolution of Intel.s Copy EXACTLY! Technology Transfer Method. *Intel Technoogy Journal*.

Rajagopalan, S. 1998. Capacity Expansion and Equipment Replacement: A Unified Approach. *Operations Research* 46 (6): 846–857.

Rajagopalan, S., and J. M. Swaminathan. 2001. A Coordinated Production Planning Model with Capacity Expansion and Inventory Management. *Management Science* 47 (11): 1562–1580.

Shenoy, S. R., and A. Daniel. 2006. Intel architecture and silicon cadence: the catalyst for industry innovation. *ftp://download.intel.com/software/pdf/IAandSiliconCadence.pdf* .

Swaminathan, J. M. 2000. Tool capacity planning for semiconductor fabrication facilities under demand uncertainty. *European Journal of Operational Research* 120:545–558.

van Mieghem, J. A. 1998. Investment Strategies for Flexible Resources. *Management Science* 44 (8): 1071–1078.

van Mieghem, J. A. 2003. Capacity Management, Investment, and Hedging: Review and Recent Develop-

ments. *Manufacturing and Service Operations Management* 5 (4): 269302.

van Mieghem, J. A. 2007. Risk Mitigation in Newsvendor Networks: Resource Diversification, Flexibility, Sharing, and Hedging. *Management Science* 53 (8): 12691288.

Yano, C. A. 1984. On the Equivalence of an Equipment Replacement Problem and a Facility Location Problem. *Technical Report 84-27. IOE Dept., Univ. of Michigan, Ann Arbor.*.

A Appendix

A.1 Additional Argument for Section 3

Uniform Demand: Proof of (3.8)

Note that the relative error of the inventory quantity, denoted by ϕ, is given by

$$\begin{aligned}\phi &= \frac{E[CP_i(t)-S_i(t)]^+ - E[CP_i(t)-S_i(t)]}{E[CP_i(t)-S_i(t)]^+} = \frac{E[S_i(t)-CP_i(t)]^+}{E[CP_i(t)-S_i(t)]+E[S_i(t)-CP_i(t)]^+} \\ &= \frac{1}{\frac{E[CP_i(t)-S_i(t)]}{E[S_i(t)-CP_i(t)]^+}+1}\,.\end{aligned}$$

Suppose that $S_i(t)$ is uniformly distributed on $[\alpha,\beta]$. Since $CP_i(t)$ is the η_i-th fractile of $S_i(t)$, we have

$$\begin{aligned}E[S_i(t)] &= \frac{\alpha+\beta}{2} = \alpha+\frac{\beta-\alpha}{2} \\ CP_i(t) &= \alpha+\eta(\beta-\alpha) \\ E[S_i(t)-CP_i(t)]^+ &= (\beta-\alpha)\cdot\frac{(1-\eta_i)^2}{2}\end{aligned}$$

Then,

$$\begin{aligned}\frac{E[CP_i(t)-S_i(t)]}{E[S_i(t)-CP_i(t)]^+} &= \frac{\alpha+\eta(\beta-\alpha)-\alpha-\frac{\beta-\alpha}{2}}{(\beta-\alpha)\cdot\frac{(1-\eta_i)^2}{2}} \\ &= \frac{\eta(\beta-\alpha)-\frac{\beta-\alpha}{2}}{(\beta-\alpha)\cdot\frac{(1-\eta_i)^2}{2}} = \frac{\eta-\frac{1}{2}}{\frac{(1-\eta_i)^2}{2}} = \frac{2\eta-1}{(1-\eta_i)^2}\,.\end{aligned}$$

Therefore,

$$\phi = \frac{1}{\frac{2\eta-1}{(1-\eta_i)^2}+1} = \frac{1}{\frac{2\eta-1}{(1-\eta_i)^2}+\frac{(1-\eta_i)^2}{(1-\eta_i)^2}} = \frac{1}{\left[\frac{\eta_i^2}{(1-\eta_i)^2}\right]} = \frac{(1-\eta_i)^2}{\eta_i^2}\,.$$

Normal Demand

Figure 6 shows the expected relative error in our inventory approximation.

A.2 Proof of Lemma 3

Proof. Part (a) follows directly from (5.5) and an observation that $\mathcal{S}(S_1(t),S_2(t))$ is stochastically increasing in t.

For Part (b), the Central Limit Theorem implies that the distributions of $[S_2(t)-t\theta_2]/\sqrt{t\sigma_2^2}$ and $[S_1(t)+\rho S_2(t)-t(\theta_1+\rho\theta_2)]/\sqrt{t\sigma_1^2+t\rho\sigma_2^2}$ converge in distribution to the standard normal

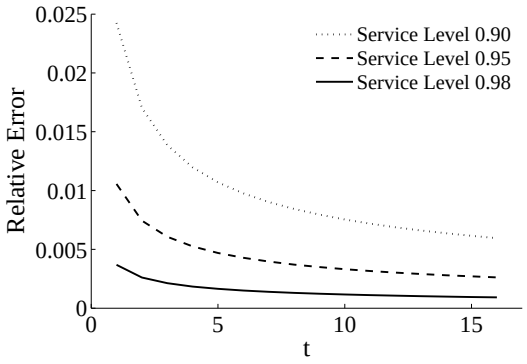


Figure 6: The relative error of the expected inventory quantity approximation $E[S_i(t) - CP_i(t)]^+ / E[CP_i(t) - S_i(t)]^+$ as a function of t when $S_i(t)$ is normally distributed with mean $t\mu$ and standard deviation $\sqrt{t}\sigma$. Note $\sigma/\mu = 0.5$.

distributions, where $\rho = \frac{r_1^B}{r_2^B}$. Let $\Phi^{-1}(\eta)$ represent the inverse of the standard normal distribution corresponding to η. Then, for any $\epsilon > 0$, we have

$$t\theta_2 + \sqrt{t\sigma_2^2}\Phi^{-1}(\eta_2 - \epsilon) \;\leq\; F_t^{-1}(\eta_2) \;\leq\; t\theta_2 + \sqrt{t\sigma_2^2}\Phi^{-1}(\eta_2 + \epsilon)$$

and also

$$\frac{t(\theta_1 + \rho\theta_2) + \sqrt{t\sigma_1^2 + t\rho\sigma_2^2}\Phi^{-1}(\eta_1 - \epsilon)}{\rho} \;\leq\; \frac{G_t^{-1}(\eta_1)}{\rho} \;\leq\; \frac{t(\theta_1 + \rho\theta_2) + \sqrt{t\sigma_1^2 + t\rho\sigma_2^2}\Phi^{-1}(\eta_1 + \epsilon)}{\rho},$$

for sufficiently large t. Since the bounds of $F_t^{-1}(\eta_2)$ consist of a linear term $t\theta_2$ and lower ordering terms, and the bounds of $\frac{G_t^{-1}(\eta_1)}{\rho}$ consist of a linear term $t(\theta_2 + \theta_1/\rho)$ and lower order terms, where $\theta_2 > 0$, we conclude that $F_t^{-1}(\eta_2) < \frac{G_t^{-1}(\eta_1)}{\rho}$ when t is sufficiently large. □