

Reliability or Inventory? Analysis of Product Support Contracts in the Defense Industry

Sang-Hyun Kim

Yale School of Management, Yale University, New Haven, CT 06511, sang.kim@yale.edu

Morris A. Cohen

The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, cohen@wharton.upenn.edu

Serguei Netessine

INSEAD, Boulevard de Constance, 77305 Fontainebleau, France, serguei.netessine@insead.edu

July 17, 2010

Abstract

In the defense industry, traditional sourcing arrangements for after-sales support of weapons systems (“products”) have centered around physical assets. Typically, the customer, a military service, would pay the supplier of maintenance services in proportion to the resources consumed, such as spare parts, that are needed to maintain the product. In recent years, we have witnessed the emergence of a new service contracting strategy called Performance-Based Logistics. Under such a performance-based contract (PBC), the basis of supplier compensation is actual realized uptime of the product. The goal of this paper is to compare the inefficiencies arising under the traditional consumption-based contract (CBC) and PBC. In both cases, the customer sets the contract terms, and as a response, the supplier sets the base-stock inventory level of spares as well as invests in increasing product reliability. We find that PBC provides stronger incentives for the supplier to invest in reliability improvement, which in turn leads to savings in acquiring and holding spare product assets. Moreover, the efficiency of PBC improves if the supplier owns a larger portion of the spare assets. Our analysis advocates usage of PBC in favor of CBC and supports a DoD recommendation for transforming suppliers into total service providers.

1 Introduction

The importance of after-sales product support in the defense industry cannot be understated. There, only about 28% of a weapon system’s total ownership cost is attributed to development and procurement, whereas the costs to operate, maintain, and dispose of the system account for the remaining 72% [14]. Given that the U.S. Department of Defense’s (DoD) annual budget for operations and maintenance amounted to nearly \$70B in 2007 (a 25% increase over 2006), it is not surprising that the manufacturers of military aircraft, engines, and avionics equipment (e.g., Boeing, GE, Honeywell, Lockheed Martin, Pratt & Whitney, and Rolls-Royce) have developed competitive strategies for the provision of service parts and repair/maintenance services.

Traditionally, many after-sales contractual relationships for mature products in the defense industry were governed by consumption-based contracts (CBC), such as the time and material (T&M) contracts, that specified the unit prices of the service parts and other consumable resources that need to be pre-positioned in order to satisfy a required service level, such as product availability. However, increasing pressure on the DoD to reduce spending as well as growing dissatisfaction with the level of after-sales support from key suppliers have led to reevaluation of these arrangements. In recent years, a novel strategy for aligning interests in the after-sales service supply chains has emerged: Performance-Based Logistics. Its premise is simple: instead of paying suppliers for parts, labor, and other services consumed to provide after sales support, the compensation is based on the actual availability of the product realized by the customer. The key idea behind such performance-based contracts (PBC)¹ is to align the incentives of all parties by tying suppliers’ compensation to the same service value that the customer cares about. After several pilot studies, the DoD mandated the implementation of PBC for all new system acquisition programs beginning in 2003 [10]. Initial reports support the view that PBC improve product availability: the U.S. Navy’s implementation of Performance-Based Logistics for its fleet of F/A-18 E/F fighter jets, for example, has resulted in an availability increase from 67% to 85%, while a similar effort has seen the material availability of Aegis guided missile cruisers rise from 62% to 94% [12].

The ultimate goal of PBC—providing incentives to suppliers to attain high product availability

¹In the remainder of the paper we use the generic term PBC in lieu of Performance-Based Logistics or PBL, which is a name specifically attributed to the DoD program. Examples of PBC are found in the commercial industry as well, and the insights found in this paper are equally applicable to those practices.

at a lower cost—can be achieved through a variety of actions. Examples include service parts deployment across multiple stocking locations, R&D effort to improve product reliability, investment in capacity for scheduled/unscheduled maintenance activities, and parts cannibalization. In this paper, our focus is on the trade-off and interaction between two such actions: investment in spare assets and in product reliability.² Industry practitioners in both government agencies and commercial enterprises identify these two actions as their most important strategic decisions.³ Reflecting this view, a recent DoD Guidebook [11] designates reliability as one of the three essential elements (along with asset availability and product maintainability) that enable mission capability. Given these considerations, we aim to address the following research questions: How does PBC differ from CBC in motivating suppliers to improve reliability and to manage the inventory of spares, which are major sources of the DoD’s expenditure? What kind of inefficiencies arise under these two contracts? Does the ownership structure of the spare assets (by the customer or by the supplier) affect the answers to these questions?

In this paper, we develop a stylized economic model that draws upon two distinct bodies of literature. We employ the classical repairable service parts inventory management model to represent repair and maintenance processes. This model is further enriched by a novel feature which has not been previously considered in the literature: endogenous product reliability improvement effort. By introducing this new decision variable, which has always been assumed to be exogenous in the previous literature, we demonstrate a new perspective for the management of after-sales service assets. The relationship between the customer and the supplier is modeled using a sequential game formulation, in which the customer sets the terms of the contract in order to minimize her total cost subject to a minimum product availability requirement. The supplier’s goal is to set the profit-maximizing levels of reliability and spares inventory given these contract terms. We allow for an arbitrary allocation of spare inventory ownership between the customer and the supplier, and compare the impacts of employing two types of contracts. Under CBC the supplier is compensated for the consumed resources (spare units, labor, and other materials), and under PBC the compen-

²Investment in repair capability in order to reduce response time is another important factor that impacts product availability. As we will demonstrate shortly, our analysis is minimally impacted by treating repair capability rather than product reliability as a variable to be controlled by the supplier.

³We are grateful to the many participants of the Wharton Service Supply Chain Thought Leaders’ Forum for bringing this issue to our attention. See <http://opim.wharton.upenn.edu/fd/forum/>.

sation is based on product availability. These two contracting approaches are widely adopted in practice, yet there is still an ongoing debate in the industry about the relative merits of the two. For example, the Government Accountability Office has expressed doubts about the superiority of PBC on several occasions [15] despite a general consensus among practitioners that PBC brings significant benefits.

In our model we assume that the availability target can be achieved by two means: investment in spares inventory or investment in product reliability. We find that CBC results in inefficiencies such that the supplier invests less in reliability and more in the inventory of spares than an integrated firm would. Compared to CBC, we demonstrate that PBC incentivizes the supplier to achieve the product availability target by investing more in reliability and simultaneously achieving savings in inventory investment. As a direct consequence, contracting efficiency is higher under PBC than under CBC. We also find that the allocation of spare asset ownership between the customer and the supplier affects efficiency of the two contracts in an opposite way. Namely, under PBC, the supplier invests more in reliability and less in inventory as his share of asset ownership increases, whereas under CBC, the opposite occurs. This unexpected contrast between CBC and PBC is revealed by our analysis of the subtle interactions between operationally significant variables, namely product reliability and inventory, in the complex service support environment that includes many different cost elements, such as repair cost, spare product holding costs that depend on the condition of a product, and reliability improvement cost.

One of the major conclusions of our paper is that the maximum benefit of PBC is realized only when spare assets are fully owned by the supplier and, moreover, the channel is coordinated with a complete asset transfer under PBC. While this conclusion provides clear policy guidance, implementing this idea is not trivial. Indeed, contrary to what our results advocate in this paper (i.e., transfer asset ownership to supplier), the prevailing industry practice is for the customer to own spare assets while the supplier decides on the stocking level of spares and recommends to the customer a budget of spares acquisitions to achieve these levels. We suspect that this ownership/decision structure is largely a relic of pre-PBC practice and of the fact that customers, especially government agencies such as the armed forces, are historically reluctant to cede control of their assets to third parties due to the fear of mismanagement and the potentially catastrophic costs of product downtime. While this is understandable, our findings indicate that such resistance

may actually be an impediment to achieving the full benefits of the PBC strategy. Thus, our analysis suggests that there are significant benefits for transforming military suppliers into total service providers who assume complete control of service functions, including asset ownership, and one of the key goals of our paper is to draw attention of senior service support managers to the importance of understanding incentives created by different contracting structures.

Although we focus primarily on the defense industry to motivate this paper, it is worth mentioning that there are many other application areas. PBCs are widely adopted outside the military under different names: in the technology sector they are known as Service Level Agreements and in commercial aviation they are known as Power by the HourTM, a term that is copyrighted by Rolls-Royce. The rest of the paper is organized as follows. After a brief survey of the related literature in Section 2, we provide our modeling assumptions and formulation in Section 3. In Section 4 we present analysis of CBC and PBC and a comparison between them. This is followed by Section 5, in which we consider the consequences of relaxing some of the basic assumptions we make in our analysis. Section 6 concludes our investigation with a summary of major findings and areas of future research.

2 Literature Review

Our study presents a game-theoretic model applied to a service parts inventory management problem. Sherbrooke [29] introduces the classical METRIC model for service parts (repairables) in the 1960's which led to numerous multi-echelon, multi-indentured inventory model extensions. In METRIC, the repair process for each part is represented by an $M/G/\infty$ queueing system, and the decision is to optimize the number of spares in stock given an *exogenous* part failure rate. Over the years the METRIC model and related models inspired by non-military applications have become the basis for a number of decision support systems that are currently used in both commercial and military settings (see, for example, [6], [8]). Despite the large volume of literature in this field, the issues of contracting and outsourcing have remained largely unaddressed. We use a simplified version of the repairable model in order to minimize complexity arising from the game-theoretic aspects of the model.

One of the novel features of our paper is endogenizing the product failure rate which, to the best

of our knowledge, has never been attempted in the service parts inventory management literature. This allows us to model the interaction between reliability improvement and inventory level decisions made by the supplier, the main focus of this paper. In this respect our model has a connection to the controlled queue literature, including recent papers by Ren and Zhou [27], Hasija et al. [17], Lu et al. [25], and Baiman et al. [2]. A follow-up of this paper by Kim [20] also considers endogenous reliability decision but in a quite different problem context.

To represent the contractual relationship between the customer and the supplier, we employ modeling approaches commonly found in the existing supply chain contracting papers. Many papers in this research area have emerged over the years, primarily focusing on the retailing industry. See Cachon [4] for the summary of this literature. The contracts we analyze in this paper fall under the class of contracts found in this stream of literature, but our paper is distinguished in that we focus on the current practices found in the context of the after-sales support business in the defense industry. Our model is closely related to the multitasking literature (e.g., [18], [13]), in which the agent controls more than one action (reliability and inventory in our case). While our paper is grounded on the ideas originating from the economics tradition, it is enriched by a faithful representation of industry practices as well as our focus on operationally significant variables and the specific recommendations that we offer in managing them.

Although not in great quantity, there are papers that discuss incentives and contracting in the defense industry. Early papers include Cummins [9] and Rogerson [28], and more recently, Kang et al. [19] propose a decision-support model that can help support PBC relationships by trading off reliability and maintenance tasks. While the last paper investigates a similar problem context as ours, it does not present an economic analysis where incentives play a central role. The work that is most related to this paper is Kim et al. [21], who consider how cost reduction and performance incentives interact under a general contracting arrangement that includes PBC when significant cost uncertainty is present, while ignoring asset ownership issues. The theme of this paper is quite different since we focus mainly on reliability improvement and its interaction with inventory management decisions under varying asset ownership structures. Another related paper is Kim et al. [22], who specifically study the contracting challenge arising from the infrequent nature of product failures, as they provide severely limited information about the supplier's effort to maintain equipment. A recent work by Kim [20] provides yet another perspective by considering

a game-theoretic situation arising from a multi-indenture structure of the service supply chain. While these works do not provide a comprehensive picture of the complex dynamics created by PBC in after-sales support environments, they include complementary analyses of different aspects of the problem and thus provide insights that are relevant to practitioners.

In summary, the analytical contributions of our paper are two-fold. First, we endogenize reliability improvement decisions in a classical repairable inventory management model and, for the first time, study the interaction between reliability and inventory. Second, we study and compare two frequently used contractual arrangements (CBC and PBC), evaluate their inefficiencies, and identify the factors that cause them. From a managerial perspective, our paper sheds light on how performance-based incentives can lead to reliability improvement and on the role of supply chain restructuring in achieving an efficient solution.

3 Model

A risk-neutral customer owns and operates a fleet of N identical products, whose continued usage is disrupted by random product failures. A failed unit is immediately sent to the supplier for a repair, while a working spare unit is pulled from the inventory, if one exists. The supplier performs three kinds of activities to support the customer’s fleet of products: (1) repairs defective units, (2) manages spare product inventory, and (3) manages product reliability. The duration of the contracting relationship between the customer and the supplier is normalized to one. The failures occur at a rate $\lambda \equiv E[\Lambda]$, where Λ is the total number of product failures within the contracting horizon. It takes ℓ_j amount of time to repair the j^{th} failure. The expected repair lead time $l \equiv E[\ell_j]$, or equivalently, the repair rate $1/l$, is assumed to be fixed and not impacted by the supplier’s effort (e.g., repairs are always performed at the maximum speed). In contrast, we assume that the Mean Time Between Failures (MTBF) $1/\lambda$, a measure of product reliability, can be increased by the supplier’s effort. In this paper we represent the reliability improvement effort by the normalized MTBF $\tau \equiv (\lambda l)^{-1}$, and henceforth will refer to it simply as reliability.⁴ The range in which τ can vary is assumed to be between $\underline{\tau}$ and $\bar{\tau}$. The lower limit $\underline{\tau}$ represents the existing level of reliability,

⁴To be precise, τ represents the inverse of the expected load in a repair facility. With l assumed to be constant, varying τ is equivalent to treating λ as the decision variable. The analysis, however, could have been carried out with repair capacity $1/l$ as a supplier decision variable or with both λ and l as decision variables.

whereas $\bar{\tau}$ is the theoretical upper limit of reliability that can be achieved.

For simplicity, we only consider a single indenture level for the product, i.e., spares inventory is managed at the product level. In practice, inventory to support maintenance and repair operations primarily consists of parts at different indenture and echelon levels and at different locations; see Section 5 for further discussion. Likewise, we assume that contracting happens at the product level and not the line replacement unit or component level. This is consistent with many PBC programs implemented in practice, for example most airplane engines in both military and commercial situations are contracted this way, and some recent systems, such as F-35 Lightning II, are contracted at the full product level. However, we acknowledge that many existing PBCs are written at the component level which we do not capture in this model.

The customer moves first as a Stackelberg leader by offering a contract, either CBC or PBC, that influences the supplier's simultaneous decision on product reliability τ and the stocking level s of spare products. Because spare products are repairable items, i.e., they are repaired upon failures and returned to the system instead of being scrapped, the quantity s (the number of spares that are produced initially) remains constant after its value is chosen. Therefore, at any given moment, there are $N + s$ products in the system. Without loss of generality, we assume $s = 0$ at the outset. By assuming that s is the supplier's choice, we only consider the case of Vendor Managed Inventory (VMI), which is the prevailing practice in after-sales product support environments. We assume that the ownership of spare assets is split between the customer and the supplier by introducing the parameter $\delta \in (0, 1]$ that represents the fraction of spares owned by the supplier. Therefore, $(1 - \delta)s$ and δs are the spares quantities belonging to the customer and the supplier, respectively. (For a reason that we describe in detail in Section 4.2, we do not include the special case $\delta = 0$, under which the customer owns the entire set of spares.) In practice, these units are physically separated as the customer stocks her portion of spares at her "retail" site (e.g., base) while the supplier holds them at his location (e.g., depot). Holding costs are incurred by the two parties in proportion to the number of spare units each owns. Throughout the paper we treat δ as an exogenous parameter (for example, the customer and the supplier agrees to split the spare assets 50-50), in order to reflect on the spectrum of ownership structures observed in practice and to highlight the consequences of varying the ownership allocation.

3.1 Repair Process and Performance Measurement

To model the repair process, we adhere to the standard assumptions in the classical service parts inventory management literature (see, for example, [26]). The repair facility is modeled as an M/G/ ∞ queue. Product failures occur according to a Poisson process, and the failed product is replaced by a working unit from the spares inventory, if one is available. Otherwise, a backorder occurs. A one-for-one base stock inventory policy is used for replacement of defective units: each failed product immediately undergoes a repair that takes a random amount of time with a general distribution function. Note that the Poisson failure process is not an exact representation since, in general, the failure rate in the closed-loop repair cycle (i.e., repaired units are restored back to the system) depends on the number of deployed units that are in working condition. However, this model is a good approximation as long as $\lambda l = 1/\tau \ll N$ is satisfied, which is true in most environments where products fail relatively infrequently. This is indeed a standard assumption in the service parts management literature, including the paper by Sherbrooke [29] who first introduced the METRIC model.

The Poisson failure assumption allows the application of Palm's Theorem, which postulates that the steady-state inventory on-order $O(\tau)$, the number of units that are being repaired at a random point in time, is Poisson-distributed with mean $\lambda l = 1/\tau$. (As noted earlier, it is only the product of λ and l that plays a role, and thus our analysis equally applies to a setting in which the supplier controls the repair leadtime. However, we do not explicitly consider this case in this paper in order to focus the discussion on the impact of PBC on reliability.) Two important random variables are on-hand inventory I and backorder B , which are related to $O(\tau)$ and s by $I | \tau, s = (s - O(\tau))^+$ and $B | \tau, s = (O(\tau) - s)^+$, where $(\cdot)^+ \equiv \max\{0, \cdot\}$. There is a one-to-one correspondence between the performance measure of our interest, the expected product availability $E[A | \tau, s]$, and the expected backorder: $E[A | \tau, s] = 1 - E[B | \tau, s]/N$. Consistent with the common practice, we assume that the customer faces an explicit service requirement $E[A | \tau, s] \geq \alpha$ (e.g., expected availability should be 95% or more), which can be translated into the backorder constraint $E[B | \tau, s] \leq \beta$.⁵

⁵An alternative to imposing the availability requirement is to assume that the customer receives a revenue stream from product use and chooses the optimal expected backorder level by maximizing its expected profit. The solution in this case would resemble the classical newsvendor solution which trades off the unit revenue with the unit cost of backorder. However, for most applications that we have in mind (e.g., military equipment) it is difficult, if not impossible, to estimate unit revenue. For this and other reasons, most related models in the service supply chain

It is important to note that our analysis rests on the assumption that the system reaches steady-state, which is standard in the repairables inventory literature. To be more precise, let $B(t)$ be the number of backorders logged at time t . It maps to the number of products that are missing from the fleet at time t since a backorder occurs only when there is no spare in the inventory to replace those units. Therefore, the cumulative product downtime is equal to $\int_0^1 B(t)dt$ (recall that the contract duration is normalized to one). The realized availability A is then $A = \frac{1}{N} \left(1 - \int_0^1 B(t)dt\right)$, i.e., the fraction of cumulative product uptime $1 - \int_0^1 B(t)dt$ against the maximum time that can be achieved at full capacity (N products times the contract duration of one). Although it is $\int_0^1 B(t)dt$, not the steady-state random variable B , that determines the realized performance outcome A , we are able to use B instead since in our model all decisions are made ex-ante and only the expectations matter; in steady-state, $E \left[\int_0^1 B(t)dt \mid \tau, s \right] = E[B \mid \tau, s]$.

While these are the standard assumptions in the literature, the discrete nature of the Poisson distribution in $O(\tau)$ limits our ability to obtain insights into the game-theoretic problem that we set out to analyze, as it handicaps our ability to obtain analytically tractable expressions. To circumvent this difficulty, we conduct an asymptotic analysis by treating $O(\tau)$ and s as continuous variables and restricting attention to situations in which N is sufficiently large and τ is sufficiently small so that

$$1/N \ll \underline{\tau} < \bar{\tau} \lesssim 0.1 \tag{1}$$

is satisfied. This condition holds, for instance, if the fleet size is large. For example, $N = 200$ and $\tau = 0.1$ imply that an average of 10 products out of 200 are being repaired at any point in time and condition (1) is satisfied. In the range of τ defined by (1), we can apply the Normal approximation of $O(\tau)$ (with $E[O(\tau)] = \text{Var}[O(\tau)] = 1/\tau$), which yields very accurate evaluations of $E[B \mid \tau, s]$ and $E[I \mid \tau, s]$, the quantities of managerial interest (see [31], pp. 205-209 for extensive discussion of the Normal approximation in this setting).

To this end, let ϕ and Φ be the pdf and the cdf of the standard Normal distribution. Define $\bar{\Phi}(\cdot) \equiv 1 - \Phi(\cdot)$. In addition, let $f(x) \equiv \phi(x)/\bar{\Phi}(x)$ be the hazard function and $L(x) \equiv \phi(x) - x\bar{\Phi}(x)$

literature were developed using the availability requirement framework (see [26], [30]) and we follow this convention.

be the loss function. The normal z -statistic for a given τ and s is

$$z \equiv (s - E[O(\tau)]) / \sqrt{\text{Var}[O(\tau)]} = (s - 1/\tau) / \sqrt{1/\tau} = \sqrt{\tau}s - 1/\sqrt{\tau}. \quad (2)$$

Hence, $s = 1/\tau + z/\sqrt{\tau}$. The expected backorder and the expected inventory on-hand are, respectively, $E[B | \tau, s] = L(z)/\sqrt{\tau}$ and $E[I | \tau, s] = (z + L(z))/\sqrt{\tau}$. Note that the expression for $E[I | \tau, s]$ contains the negative domain of s , but its effect is inconsequential under (1).

3.2 Cost Structures

We assume that the customer and the supplier are subject to the following costs:

- $K(\tau)$: cost of improving reliability τ ,
- κ : cost of repairing a defective product per unit time,
- c : cost of producing a unit of spare,
- h_g : cost of carrying a functional product per unit time,
- h_b : cost of carrying a defective product per unit time.

All of these parameters are assumed to be public knowledge. $K(\tau)$ represents the dollar amount of investment in research and development or engineering changes required to improve reliability to τ . It is the supplier who incurs this cost. We assume that $K(\tau)$ is increasing and convex, i.e., $K'(\tau) > 0$, $K''(\tau) > 0$. Convexity is a reasonable assumption since the most efficient improvement opportunities will be exploited first from among many technological and process choices. Furthermore, we assume that $K'''(\tau) > 0$,⁶ $K(\underline{\tau}) = 0$, and $\lim_{\tau \rightarrow \bar{\tau}} K(\tau) = \lim_{\tau \rightarrow \bar{\tau}} K'(\tau) = \infty$. Hence, $\underline{\tau}$ can be interpreted as the baseline reliability that the supplier can provide without incurring the extra cost $K(\tau)$, while it becomes prohibitively expensive to achieve the theoretical upper bound $\bar{\tau}$.

We assume that a cost κ is incurred per unit time while a unit is being repaired. The source of such a variable cost may include labor, repair equipment rental, and electricity. While a repair

⁶The assumption that marginal cost is convex increasing is frequently employed in the economics literature to facilitate analytical tractability, as is in our paper. For example, see [23].

may incur a fixed cost as well, focusing on the variable cost κ is without loss of generality since the constant expected repair lead time assumption allows us to absorb any fixed cost into κ . Since the total expected duration of repairs over the contracting horizon is $\lambda l = 1/\tau$, the expected repair cost is κ/τ . (Equivalently, from a steady-state perspective, the repair cost is proportional to the expected number of units being repaired at a random point in time, i.e., $\kappa E [O(\tau)] = \kappa/\tau$.) This cost is borne by the supplier since we assume that all repairs are performed by him.

We assign two different values for the holding cost, h_g and h_b , each corresponding to the state that a product is in: at any given time, a product is either functional (“good” unit) or defective (“bad” unit). Good units include those deployed in the fleet and the spares stored in the inventory, while the bad units are those undergoing repairs in the repair facility. The value of a good unit is higher than that of a bad unit, since a bad unit cannot generate the same services that the good one does. (In an open market, a bad unit can only receive a scrap value whereas a good unit receives a full value.) As the major portion of the holding cost is an opportunity cost of capital that is proportional to the current product value, $h_b < h_g < c$.⁷

The holding costs incurred by the customer and the supplier are proportional to the number of the products each owns, and thus they depend on the parameter δ that represents the supplier’s portion of the total spare assets. In this paper we adopt the convention that the products are indistinguishable as long as they are in the same state (good or bad). In other words, product ownership is independent of the serial number attached to each product. For example, if a product that was initially in the fleet is sent to the repair facility and is replaced by a spare, the latter becomes the customer’s property. Under this assumption, the number of good units that the customer owns at any given moment is equal to $N - (O(\tau) - s)^+ + (1 - \delta)(s - O(\tau))^+$; if a backorder occurs ($O(\tau) > s$) and consequently availability is less than 100%, then inventory is empty and there are $N - (O(\tau) - s)$ good units (all in the fleet), whereas if availability is 100% then there are N good units in the fleet as well as $s - O(\tau)$ good units in the inventory, of which the customer owns a fraction $1 - \delta$. Similarly, the number of bad units that the customer owns is equal to $(O(\tau) - s)^+ + (1 - \delta) \min\{O(\tau), s\}$; if a backorder occurs then the customer’s property includes $O(\tau) - s$ units “missing” from the fleet as well as $(1 - \delta)s$ spares, all of which are in the repair

⁷While in general c , h_g , and h_b may depend on τ , we believe that such dependence is a second-order effect, so we assume these cost parameters are constant.

facility, whereas if the inventory is nonempty (and hence availability is 100%) then only $1 - \delta$ fraction of the $O(\tau)$ units being repaired is owned by the customer. The expected total holding costs for the customer and the supplier are then

$$\begin{aligned} H(\tau, s) &\equiv h_g(N - E[B | \tau, s] + (1 - \delta)E[I | \tau, s]) + h_b(E[B | \tau, s] + (1 - \delta)E[\min\{O(\tau), s\}]), \\ \eta(\tau, s) &\equiv \delta h_g E[I | \tau, s] + \delta h_b E[\min\{O(\tau), s\}]. \end{aligned}$$

Using the identities $E[\min\{O(\tau), s\}] = s - E[I | \tau, s]$ and $E[B | \tau, s] = E[O(\tau)] - s + E[I | \tau, s]$, we can rewrite $H(\tau, s)$ and $\eta(\tau, s)$ as

$$H(\tau, s) = h_g N - (h_g - h_b)E[O(\tau)] + (h_g - \delta h_b)s - \delta(h_g - h_b)E[I | \tau, s], \quad (3)$$

$$\eta(\tau, s) = \delta h_b s + \delta(h_g - h_b)E[I | \tau, s]. \quad (4)$$

As expected, the system-wide expected holding cost is equal to $H(\tau, s) + \eta(\tau, s) = h_g E[N + s - O(\tau)] + h_b E[O(\tau)]$, i.e., only $O(\tau)$ units that are in repair out of the total product population $N + s$ are subject to the lower unit holding cost h_b .

Adding all cost components described thus far, the total expected internal costs of the customer and the supplier are, respectively,

$$\Psi(\tau, s) \equiv H(\tau, s), \quad (5)$$

$$\psi(\tau, s) \equiv K(\tau) + \kappa/\tau + cs + \eta(\tau, s). \quad (6)$$

Note that the supplier's production cost is cs , not δcs , because the stocking level s is the supplier's discretionary choice and hence he has to bear the full cost of production. In other words, dividing asset ownership does not imply that the customer subsidizes the production cost; the division occurs after the supplier completes production of s units, and therefore, it is reflected only in the holding costs $H(\tau, s)$ and $\eta(\tau, s)$.

In the remainder of the paper, we make the following technical assumptions regarding the cost parameters which ensure that the problem is well-defined and allow us to focus on the most

interesting and managerially relevant cases. They are:

$$\underline{\tau} < 1/\beta, \tag{7}$$

$$\underline{\tau}^2 K'(\underline{\tau}) < \kappa - (h_g - h_b), \tag{8}$$

$$\kappa + c + h_b < (1/\beta)^2 K'(1/\beta), \tag{9}$$

$$2(h_g - h_b) < \underline{\tau}^3 K''(\underline{\tau}) \tag{10}$$

Assumptions (7)-(9) together ensure that in no circumstances we consider in our analysis is it optimal to have $\tau = \underline{\tau}$ or $s = 0$ or both, thereby allowing us to focus on practically more relevant situations. All of these assumptions offer quite reasonable interpretations. (7) simply says that the availability target is sufficiently high so that investment in reliability and inventory should be considered. (8) and (9) state that the net benefit of improving reliability is sufficiently high to justify the cost of such an investment but is not too high to the extent that investing only in reliability, and not in inventory, is optimal.⁸ Finally, without (10), it is possible in a decentralized setting that an overinvestment in reliability and/or inventory, i.e., the combination that leads to higher availability than the required target level, is optimal in equilibrium. While this represents a nontrivial and unexpected situation, we believe it is of less importance from a managerial perspective compared to the main message we aim to deliver in this paper. So we sidestep such a case by imposing (10).

3.3 Contracts

At the beginning of the contractual relationship the customer offers to the supplier a contract that defines the payment to the supplier, denoted by T . Anticipating the supplier's optimal response τ^* and s^* , the customer determines compensation terms that would minimize her total cost $E[T | \tau^*, s^*] + \Psi(\tau^*, s^*)$ subject to the availability constraint while making sure that the supplier participates in the trade. In response, the supplier sets τ^* and s^* that maximize his expected profit

⁸Reliability improvement brings the following benefits and costs on the units *already in the system*: savings in repair cost (κ) and a holding cost increase resulting from converting a bad unit to a good unit ($-h_g + h_b$). The net is $\kappa - (h_g - h_b)$, as it appears in (8). Note that (8) implicitly assumes that this net savings is positive, i.e., repair cost is significant. This assumption allows us to focus on interior solutions instead of the corner solutions, streamlining our analysis. In addition, improved reliability lessens the need to produce a brand new spare unit, thereby achieving a savings amount of $c + h_g$, the sum of the product cost and the holding cost. In total, the per unit net savings resulting from reliability improvement is $\kappa - (h_g - h_b) + c + h_g = \kappa + c + h_b$, the quantity that appears in (9).

$E[T | \tau, s] - \psi(\tau, s)$. As the names suggest, the customer pays the supplier based on the amount of resources consumed for repair and maintenance activities under CBC, whereas it is the performance outcome (availability) that is the basis of compensation under PBC.

Consistent with the common assumption found in the majority of papers in the supply chain contracting literature, we assume that the supplier's decision variables, τ and s , are observable but not directly contractible, i.e., the customer cannot specify the desired levels of τ and s in a contract. That s is observable is motivated by the CBC practice in which the supplier reveals his chosen stocking quantity when he bills the customer for compensation (this is similar to the well-studied setup in a retailing context where a manufacturer offers a price-based contract knowing that a retailer's order quantity will be revealed later; see, for example, [24], [3]). Under PBC, on the other hand, the supplier does not have to report s , since it is only the realized performance outcome (availability) that constitutes the basis of compensation. Hence, observability of s under PBC is less of an issue from a modeling perspective. However, we make the consistent assumption across the two contracting cases in order to facilitate fair and meaningful comparisons. Non-contractibility assumption reflects the reality that accurately verifying the supplier's decisions is typically prohibitively expensive in a multi-echelon, multi-indenture environment typically found in practice, where the stocking levels and reliability data for tens of thousands of different parts need to be recorded in real time and analyzed to allow for an audit. As we abstract away from these complexities, observability/non-contractability assumption offers the best balance in representing the real-world practices in a stylized model that we present here, and it allows us to focus on the main theme of comparing the incentive roles provided by the two widely-used contracts. Likewise, we assume that the two intermediate random variables that influence availability, the cumulative repair time $\sum_{j=1}^{\Lambda} \ell_j$ and the cumulative backorders $\int_0^1 B(t)dt$, are also observable, as they need to be assessed under CBC and PBC in order to pay or charge the supplier.

In this paper we focus on the following linear functions for the payments defined by the two contracts: $T = w + ps + r \sum_{j=1}^{\Lambda} \ell_j$ for CBC and $T = w - v \int_0^1 B(t)dt$ for PBC. $w \geq 0$ is a lump-sum payment, $p \geq 0$ is the unit price for each unit of spare product the supplier produces, $r \geq 0$ is the compensation rate per unit time for repairing each defective unit, and $v \geq 0$ is the penalty rate for the realized backorders. In expectation, $E[T | \tau, s] = w + ps + r/\tau$ for CBC and

$E[T | \tau, s] = w - vE[B | \tau, s]$ for PBC.⁹ The parameter p is to be interpreted as a *reservation price* that the customer uses as a lever to incentivize the supplier to secure a desired level of spares. While charging a penalty rate v for backorders is equivalent to applying a bonus rate for availability, a penalty-based PBC is more commonly observed in defense industry practices so we adopt that approach in our analysis. Despite somewhat restrictive linear forms, the two payment functions defined above are quite general as they encompass many well-known contracts. One of them is the time and material (T&M) contract, which was widely used in the defense industry before the introduction of Performance-Based Logistics. Under the traditional T&M contract, the supplier generates profit through the markups on each consumed resource, i.e., the margins $p - c$ and $r - \kappa$ for spares and time-based consumptions. Therefore, the T&M contract (which can be viewed as a special case of the Cost Plus contract) is in fact CBC with $w = 0$, $p > c$, and $r > \kappa$. In addition, both CBC and PBC reduce to the traditional Fixed Price contract when $w > 0$ and $p = r = v = 0$. A simpler version of PBC called Power by the Hour is also observed in the commercial airline industry, for which the compensation is proportional to the product uptime; in this case, $w = vN$ and $v > 0$. In our analysis we mention these special cases whenever appropriate.

Note that, from a cursory look, it is not fair to compare CBC and PBC as defined above since the former has two contract parameters whereas the latter has only one, indicating that the customer may have more flexibility with the former. As we will find out later, however, this seeming disadvantage of PBC does not present a handicap.

To summarize, the customer's problem can be written as (*SB* for second-best)

$$(SB) \quad \min_{\mathcal{P}} \quad E[T | \tau^*, s^*] + \Psi(\tau^*, s^*) \quad \text{subject to } E[B | \tau^*, s^*] \leq \beta, \\ E[T | \tau^*, s^*] - \psi(\tau^*, s^*) \geq \underline{u}, \text{ and } (\tau^*, s^*) \in \arg \max \{E[T | \tau, s] - \psi(\tau, s)\},$$

where T , Ψ , and ψ are defined above and $\mathcal{P} = \{w, p, r\}$ for CBC and $\mathcal{P} = \{w, v\}$ for PBC. The first constraint in (*SB*) represents the customer's availability requirement, expressed in terms of the upper limit on the expected backorder (thanks to one-to-one mapping between the two performance measures). The last two constraints, the individual rationality (IR) and the incentive compatibility

⁹The former obtains since $\sum_{j=1}^{\Lambda} \ell_j$ is a compound Poisson variable, whose mean is $E[\Lambda]E[\ell_j] = \lambda l = 1/\tau$. The latter follows from the fact that $E\left[\int_0^1 B(t)dt \mid \tau, s\right] = E[B | \tau, s]$ in steady-state.

(IC) constraints, describe that the supplier's participation is ensured and that he decides (τ^*, s^*) to maximize his expected profit given the contract parameters p , r , and v . It is important to recognize that the supplier is not subject to the same backorder constraint that the customer faces, and as a result, the customer has to use contract terms as a lever to influence the supplier's decisions in order to satisfy the constraint. The supplier's constant reservation utility level \underline{u} represents the profit that he can generate in an outside opportunity, and without loss of generality, we assume that its value is sufficiently high so that the lump-sum payment w is always nonnegative. (SB) can be simplified after recognizing that the (IR) constraint is always binding at the optimum by adjusting w accordingly. Then the problem is reduced to

$$(SB) \quad \min_{\mathcal{P} \setminus \{w\}} C(\tau^*, s^*) \equiv \underline{u} + \Psi(\tau^*, s^*) + \psi(\tau^*, s^*) \quad \text{s.t.} \quad E[B | \tau^*, s^*] \leq \beta, \quad (\tau^*, s^*) \in \arg \max u(\tau, s),$$

where we used the notation $u(\tau, s) \equiv E[T | \tau, s] - \psi(\tau, s)$ for the supplier's expected profit. Therefore, the problem becomes that of minimizing the total expected supply chain cost under the constraints that the backorder target should be met and the reliability and inventory levels are optimally set by the supplier.

4 Analysis

In this section we determine the optimal contract terms and the equilibrium solutions for τ and s . We begin with the benchmark first-best case in which the customer and the supplier are assumed to be one entity and thus contracts are unnecessary. We then analyze the equilibrium outcomes of CBC and PBC, and finally compare their merits and limitations.

4.1 The First-Best Benchmark

To establish the benchmark, we first analyze the case in which the customer and the supplier are assumed to be one integrated firm minimizing its total cost subject to the availability requirement (first-best or *FB*):

$$(FB) \quad \min_{\tau, s} \underline{u} + \Psi(\tau, s) + \psi(\tau, s) \quad \text{subject to} \quad E[B | \tau, s] \leq \beta.$$

(\underline{u} is included here to permit a fair comparison with the second-best cases; in this setting, it is interpreted as the transaction cost of merging the two parties.) The solution is stated in the following proposition. The following two quantities appear frequently in the analysis below and are defined here for convenience:

$$\zeta(\tau) \equiv L^{-1}(\beta\sqrt{\tau}), \quad (11)$$

$$\Gamma(\tau) \equiv \frac{\kappa + c + h_b}{\tau^2} + \frac{c + h_g}{2\tau^{3/2}} f(\zeta(\tau)). \quad (12)$$

Note that both are decreasing functions. The first-best optimal values of τ and s are characterized as follows.

Proposition 1 (*First-best*) *The backorder constraint binds at the optimum. The integrated firm chooses $\tau^{FB} > \underline{\tau}$ and $s^{FB} = (\tau^{FB})^{-1} + (\tau^{FB})^{-1/2} \zeta(\tau^{FB}) > 0$ where τ^{FB} is uniquely determined from the equation $\Gamma(\tau) = K'(\tau)$.*

Under the assumptions (7)-(9), the integrated firm finds it optimal to invest in both reliability and spares inventory in order to satisfy the specified backorder target β . The quantity $\Gamma(\tau)$ represents the firm's marginal benefit of improved reliability; hence, the first-order condition $\Gamma(\tau) = K'(\tau)$ points to the optimal level at which the marginal benefit is equal to the marginal cost of improving reliability. At this level, τ^{FB} , the optimal stocking quantity s^{FB} is determined from the backorder constraint, which is shown to bind at the optimum. Later in our analysis τ^{FB} and s^{FB} will be compared against their counterparts under CBC and PBC.

4.2 Consumption-Based Contract

Under CBC, the supplier is compensated in proportion to the resources consumed for repair and maintenance services provided to the customer. The supplier chooses the optimal reliability and spares stocking levels τ and s in response to the contract terms p and r , which are the unit prices for spares quantity and repair times. As noted in Section 3.3, the expected payment under CBC is $E[T | \tau, s] = w + ps + r/\tau$. The following lemma shows that the customer should consider only a limited range of price p .

Lemma 1 *Under CBC, a finite feasible solution of (\widetilde{SB}) exists only if $c + \delta h_b < p < c + \delta h_g$.*

The lower and upper bounds of p specified in the lemma, $c + \delta h_b$ and $c + \delta h_g$, represent the minimum and the maximum costs incurred by the supplier to own a spare product (depending on whether the product is functional or not). If the supplier is paid an amount smaller than the minimum, i.e., $p \leq c + \delta h_b$, then it is optimal for the supplier to choose $s = 0$. In this case, however, it is impossible to satisfy the backorder constraint in the range of parameters we assume, i.e., (7) and (9). On the other hand, if $p \geq c + \delta h_g$, the supplier attempts to produce as many spares as possible ($s \rightarrow \infty$), since he is fully covered for the ownership cost regardless of the product's working condition and at the same time gets paid an extra amount. Anticipating these optimal choices by the supplier, the customer should restrict the price p in the range specified in the lemma.

The lemma implies that, interestingly, a feasible solution cannot be obtained in the limit $\delta \rightarrow 0$ that represents full customer asset ownership. This result is a by-product of our model assumption that it is the supplier who has the decision rights; if the supplier does not own any part of spares inventory and hence does not bear costs associated with it, he becomes indifferent to the consequences of his stocking decision. In practice such a scenario is somewhat unlikely since the customer organizations usually place safeguards against the suppliers' misuse of contract terms. (We do, however, often hear about exorbitant amount of charges submitted to the government for reimbursements.) Since this extreme case pushes the limit of the stylized nature of our model, we do not consider it in our paper.

Assuming that the customer offers the unit price in the range specified in Lemma 1, we now study how the supplier optimally responds to the proposed contract terms p and r .

Lemma 2 (*Supplier's optimal response under CBC*) *Suppose that the condition in Lemma 1 is satisfied. Define $a \equiv \kappa - \underline{\tau}^2 K'(\underline{\tau}) - \delta(h_g - h_b) > 0$ and $b \equiv \kappa - \underline{\tau}^2 K'(\underline{\tau}) > a$ and let $\hat{\tau}$ be the unique solution of*

$$\frac{\kappa + c + \delta h_b - r - p}{\tau^2} + \frac{\delta(h_g - h_b)}{2\tau^{3/2}} \phi(z^*) = K'(\tau), \quad (13)$$

where $z^* = \Phi^{-1}\left(1 - \frac{c + \delta h_g - p}{\delta(h_g - h_b)}\right)$. The supplier chooses τ^* and $s^* = (\tau^*)^{-1} + (\tau^*)^{-1/2} z^*$ as follows:

- (i) If $0 \leq r \leq a$ then $\tau^* = \hat{\tau} > \underline{\tau}$ for all $p \in (c + \delta h_b, c + \delta h_g)$.
- (ii) If $a < r < b$ then there exists $\bar{p}(r) \in (c + \delta h_b, c + \delta h_g)$ such that $\tau^* = \hat{\tau} > \underline{\tau}$ if $p \in (c + \delta h_b, \bar{p}(r))$ and $\tau^* = \underline{\tau}$ if $p \in [\bar{p}(r), c + \delta h_g)$.

(iii) If $r \geq b$ then $\tau^* = \underline{\tau}$ for all $p \in (c + \delta h_b, c + \delta h_g)$.

Furthermore, $\frac{\partial \tau^*}{\partial p} \leq 0$, $\frac{\partial s^*}{\partial p} > 0$, $\frac{\partial \tau^*}{\partial r} \leq 0$, where the equalities hold if and only if $\tau^* = \underline{\tau}$, and $\frac{\partial s^*}{\partial r} = 0$.

Several interesting observations are made from the lemma. First, it is clear that, in order to motivate the supplier to invest in reliability improvement, the customer should not offer large prices for either p or r or both. Second, the T&M contract, which we defined in Section 3.3 as CBC with positive margins $p - c > 0$ and $r - \kappa > 0$ for both spares and time-based resources, never incentivizes the supplier to improve reliability, as is evident from the condition stated in part (iii). (Although Lemma 2 does not enforce the condition $w = 0$ for the T&M contract, the result is intact as long as the margins are positive.) These results are direct consequences of the unique incentive structure inherent in the after-sales support contracting environment. Namely, for the supplier who provides repair and maintenance services and gets compensated for the consumed resources to support them (as has been the prevailing business model in the defense industry), his business grows if the products are *less* reliable: the more frequently the products fail, the higher the supplier's revenue under CBC. While this may benefit the supplier, it has a negative impact on product availability and consequently on the customer's ability to generate value through product use. Therefore, increasing the compensation rates for the consumed resources only exacerbates this skewed incentive as the supplier earns higher margins for a higher rate of product failures. This insight is summarized by the sensitivity results in the lemma. With higher values of p and r , the supplier lowers investment in reliability ($\partial \tau^* / \partial p \leq 0$ and $\partial \tau^* / \partial r \leq 0$). On the other hand, higher p induces the supplier to increase the stocking level s ($\partial s^* / \partial p > 0$), whereas he is indifferent to r ($\partial s^* / \partial r = 0$). The conflicting dual roles of spares unit price p is particularly noteworthy: increasing p induces higher stocking quantity but lower reliability. These results indicate that, in order to produce high product availability (achieved by high levels of reliability and inventory), the customer has to offer a small value for r and a sufficiently large value for p .

Armed with the insights into the supplier's optimal response, now we turn to the customer's contract design problem and the optimal solution that emerges in equilibrium. From (\widetilde{SB}) , we see that the customer's problem under CBC is equivalent to choosing the optimal values for the contract terms p and r so as to minimize the total expected supply chain cost $C(\tau^*, s^*) = \underline{u} +$

$\Psi(\tau^*, s^*) + \psi(\tau^*, s^*)$ subject to the backorder constraint $E[B | \tau^*, s^*] \leq \beta$, where τ^* and s^* are determined as described in Lemma 2. The equilibrium solution, denoted by the superscript C , is specified as follows.

Proposition 2 (*Optimal CBC*) *In equilibrium the backorder constraint binds. The customer offers $r = 0$ and $p = c + \delta h_b + \delta (h_g - h_b) \Phi(\zeta(\tau^C))$, where $\tau^C > \underline{\tau}$ is the equilibrium reliability chosen by the supplier that is uniquely determined from the equation*

$$\frac{\kappa - \delta (h_g - h_b) \Phi(\zeta(\tau))}{\tau^2} + \frac{\delta (h_g - h_b)}{2\tau^{3/2}} \phi(\zeta(\tau)) = K'(\tau). \quad (14)$$

We find that, under the assumption (7) that the repair cost is significant enough to justify an investment in reliability improvement (thereby leading to $\tau^C > \underline{\tau}$), the customer who employs CBC should not pay the supplier based on time-based consumptions: r should be set to zero. Therefore, although we began with a general contract form that includes two contract parameters $p \geq 0$ and $r \geq 0$, only p turns out to be a useful lever that enables satisfaction of the backorder constraint. If we relax (7), on the other hand, we may have $r > 0$ but it requires the equilibrium reliability to be $\tau^C = \underline{\tau}$. Either way, r is ineffective in incentivizing reliability improvement. Given that many existing support contracts in the defense industry include time-based compensations, this result is quite striking. Proposition 2 strongly suggests that such a practice impedes the supplier's motivation to improve product reliability and therefore should be suppressed when reliability is a concern. The spare unit price p , on the other hand, is an important (and the only) instrument under CBC that makes it possible to achieve high availability through investment in spares inventory, although it does not promote reliability improvement, either.

4.3 Performance-Based Contract

Next, we analyze PBC. As in the previous section, we start by identifying the feasible range of the contract term.

Lemma 3 *Under PBC, a finite feasible solution of (\widetilde{SB}) exists only if $v > c + \delta h_b$.*

Notice that, unlike what we found in Lemma 1 for CBC, there is no upper bound on the contract term v to ensure feasibility of the solution. This provides the first hint at the qualitative difference

between CBC and PBC. The supplier's optimal response to the contract terms is as follows.

Lemma 4 (*Supplier's optimal response under PBC*) *Suppose that the condition in Lemma 3 is satisfied. The supplier chooses $\tau^* > \underline{\tau}$ which is a unique solution of the equation*

$$\frac{\kappa + c + \delta h_b}{\tau^2} + \frac{v + \delta(h_g - h_b)}{2\tau^{3/2}} \phi(z^*) = K'(\tau), \quad (15)$$

where $z^* = \Phi^{-1}\left(1 - \frac{c + \delta h_g}{v + \delta(h_g - h_b)}\right)$, and sets $s^* = (\tau^*) + (\tau^*)^{-1/2} z^* > 0$. Furthermore, $\frac{\partial \tau^*}{\partial v} > 0$ and $\frac{\partial s^*}{\partial v} > 0$.

We see a stark contrast between CBC and PBC from the sensitivity results. Recall from Lemma 2 that increasing the unit price p induces the supplier to choose a higher spares stocking level s but lower reliability τ . In contrast, the backorder penalty v induces the supplier to increase both τ and s . Thus, Lemmas 2 and 4 highlight the key difference between CBC and PBC. Namely, the two comparable terms in these contracts, p and v , induce opposite reactions from the supplier with respect to reliability improvement decision. This difference arises from the relationship between availability, the performance measure that the customer ultimately wants to increase, and the two intermediate outcomes that each contract term is designed to evaluate, namely the spares inventory and the backorders, respectively for p and v . Availability can be increased in different ways: higher reliability, more spares, or a combination of both. Under CBC only one component of this mix (i.e., spares inventory) receives attention, whereas under PBC both do, as the backorder penalty is reduced by higher levels of both reliability and spares inventory.¹⁰ CBC does contain an additional contract term r that influences the supplier's reliability decision, but it does not complement the shortcoming of p since increasing it goes counter to the direction that the customer intends: reliability is reduced with higher r .

Overall, we infer from these sensitivity results that PBC is superior to CBC in incentivizing the supplier to improve product reliability. However, since availability is a function of both reliability

¹⁰ Although it sounds intuitive that higher backorder penalty v induces the supplier to choose higher spares inventory, this result does not necessarily hold in all circumstances. As a matter of fact, violation of the condition (10) may break down the monotonicity. This can be understood from the expression $s^* = (\tau^*)^{-1} + (\tau^*)^{-1/2} z^*$ that appears in Lemma 4: while the backorders can be reduced by having more units in stock (higher z^*), the supplier actually cuts down the necessary amount of units by increasing reliability (higher τ^*). The net effect is unclear, but the condition (10) ensures that the former effect dominates the latter.

and inventory, it is still unclear if PBC leads to lower cost than CBC does. We answer this question in the next subsection. As a prerequisite, we derive the equilibrium outcome under PBC. The solution approach is similar to that of Proposition 2. That is, we solve the optimization problem (\widetilde{SB}) by incorporating the supplier's optimal responses τ^* and s^* as specified in Lemma 4.

Proposition 3 (*Optimal PBC*) *In equilibrium the backorder constraint binds. The customer offers $v = (c + \delta h_g) / \overline{\Phi}(\zeta(\tau^P)) - \delta(h_g - h_b)$, where τ^P is the equilibrium reliability chosen by the supplier that is uniquely determined from the equation*

$$\Gamma(\tau) - (1 - \delta) \left(\frac{h_b}{\tau^2} + \frac{h_g}{2\tau^{3/2}} f(\zeta(\tau)) \right) = K'(\tau). \quad (16)$$

4.4 Comparisons of the Contracts

Having analyzed the structures of optimal contracts and the equilibrium outcomes under CBC and PBC, we are now in a position to compare relative performances of each contracting approach and see how they fare against the first-best benchmark. This is summarized in the following proposition. Here, we use the notations C^C , C^P , and C^{FB} to represent the customer's (equivalently, the supply chain's) expected cost in equilibrium for CBC, PBC, and the first-best cases.

Proposition 4 *In equilibrium, $\tau^C < \tau^P \leq \tau^{FB}$, $s^C > s^P \geq s^{FB}$, and $C^C > C^P \geq C^{FB}$. Furthermore, first-best can be achieved under PBC with $\delta = 1$ but not under CBC.*

The insights we have gained from our discussions of CBC and PBC above point to PBC's superiority in promoting reliability improvement, as stated in the proposition. Additionally, however, the proposition demonstrates that it is not only reliability that PBC brings an advantage of—compared to CBC, it also lowers inventory. Therefore, a win-win situation marked by higher reliability and lower inventory can be attained through PBC. This reflects the fundamental relationship between reliability and inventory: they are *substitutes* in achieving a given level of availability. In other words, less frequent product failures lessens the need to maintain a large stock of spares and other physical resources. See Figure 1 for a schematic illustration of this relationship. The proposition also reveals that contracting efficiency, measured by the inverse cost ratios $(C^C/C^{FB})^{-1}$ and $(C^P/C^{FB})^{-1}$, is higher under PBC. This is quite intuitive given what we have learned. As products

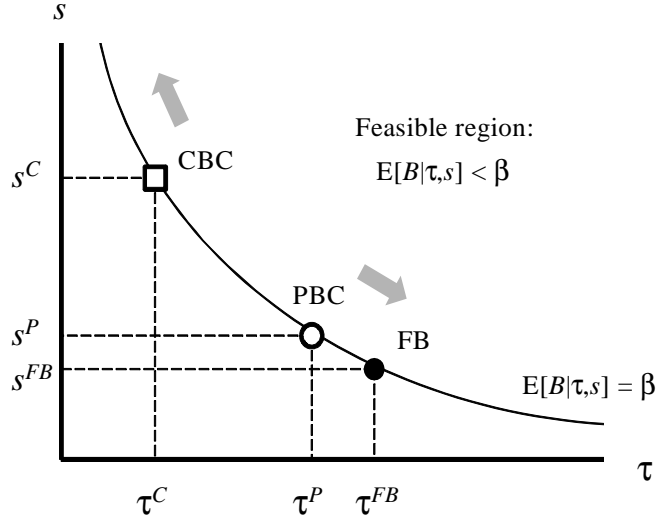


Figure 1: Substitutable relationship between product reliability (τ) and inventory (s) with respect to a fixed availability target, expressed in terms of the backorder constraint. The optimal combinations of τ and s in equilibrium under each contracting scenario are marked in the diagram. The arrows represent the direction to which the optimal combination moves as δ increases.

fail are less frequently, there is a smaller need for the resources that are used to counter the adverse effects of failures, resulting in cost savings. Although it is costly to improve reliability, the contract terms under PBC brings “more bang for the buck”, and therefore, contributes more to the savings.

Another important insight from Proposition 4 is that the spares asset ownership structure, represented by the parameter δ , impacts system efficiency differently across the two contracting cases. In particular, first-best can be achieved under PBC in the special case $\delta = 1$, i.e., when the supplier owns the entire spare assets. Under CBC, by contrast, first-best can never be achieved. This observation makes it clear that incentives between the customer and the supplier are better aligned under PBC, and moreover, a transfer of asset ownership to the supplier facilitates it. Perfect incentive alignment is attained because of the combination of two factors: (a) a complete ownership transfer forces the supplier absorb the entire cost existing in the supply chain and (b) PBC effectively converts the stochastic performance outcome, i.e., realized availability, into financial consequences for the supplier. As a result, the supplier bears the full risks of product downtimes and the associated loss of value, as the integrated supply chain would.

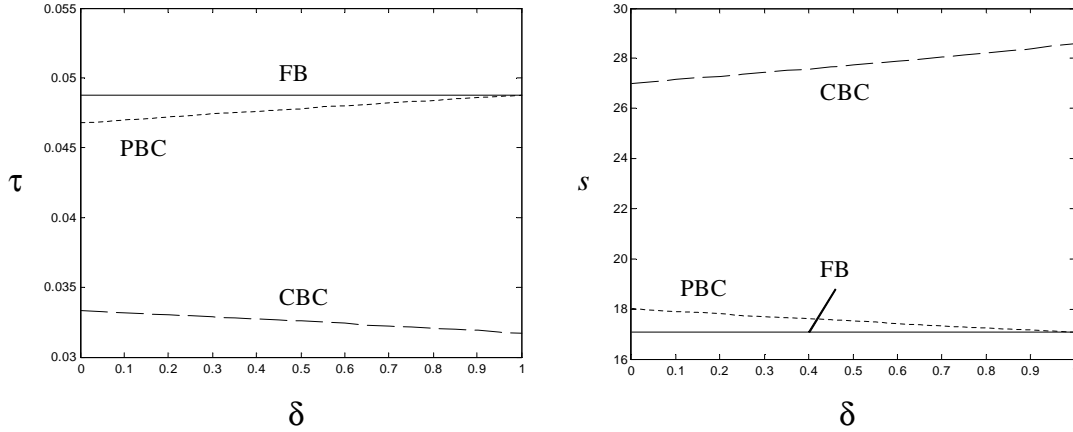


Figure 2: An example showing the changes in the equilibrium levels of τ and s under CBC and PBC as a function of δ .

This argument suggests that PBC is not as effective when the supplier only has a partial ownership of assets ($\delta < 1$). Indeed this is what we find. Under PBC, lowering δ from one results in lower τ , higher s , and higher supply chain cost—in other words, all equilibrium numbers move away from the first-best levels. Interestingly, we find the opposite behavior under CBC: reliability becomes worse, inventory goes up, and the supply chain cost increases with a *larger* value of δ , i.e., as the supplier’s share of asset ownership becomes larger, not smaller. (This is numerically verified; an analytical proof is intractable.) See Figure 2 for an illustration. Intuitively this happens under CBC since, as the supplier’s profitability is eroded by a higher ownership cost, he may compensate for the loss by letting the products fail often and increasing the revenue originating from resource consumption. This observation again highlights the contrasting incentive structures that are present under CBC and PBC.

Summarizing, we find that spare asset ownership allocation plays an important role in effective management of repair and maintenance services outsourcing. If a situation dictates that CBC has to be implemented, then it is best that the customer to retain a majority share of assets. (Recall from our discussion in Section 4.2, however, that full asset ownership brings unpredictable consequences under our model assumptions.) On the other hand, organizations considering switching to PBC can maximize the benefit by transferring as many assets as possible, thereby converting the supplier into a total service provider who not only delivers requested services but also actively manages

physical resources that are needed to support them.

5 Discussion of Modeling Assumptions

In order to highlight the main issues of interest, we have made several simplifying assumptions throughout the paper. In this section, we outline possible consequences of relaxing some of these assumptions. First, in the paper, we treat each spare product as an integrated “kit” instead of an assembled product consisting of many different parts. In reality, contracts are often enforced at the subsystem or part level (e.g., PBC is often awarded for an engine or an avionics subsystem). An explicit model of subsystems raises the issue of how to break down the availability requirement for the final product into the requirements for each component which leads to a stocking policy for each item. While the algorithms to solve this problem are well-known (“marginal analysis”; see [30]), a game-theoretic analysis of the setup in which there are multiple suppliers of an assembly system presents a new layer of complexities. For example, one component supplier may decide to free ride on another if it is difficult to establish which component has caused the system to fail. Such gaming scenarios are beyond the scope of this paper, in which the focus is on the interaction between reliability and inventory decisions and the implication of ownership structure. We have, however, embarked on an extension of this paper considering a multi-indenture supply chain structure with less emphasis on ownership issues and more on interactions among multiple suppliers [20].

In addition, while focusing on the trade-off between investment in reliability and service parts management, we ignored several other important aspects of the contractual relationships prevalent in the defense industry. Perhaps the most important aspect is the long-term nature of most of these relationships, which is partially driven by the fact that there is a single monopolistic customer and very few potential system suppliers. We found that, in practice, in addition to explicit contractual terms (such as those based on availability), customers often evaluate their suppliers based on a variety of other metrics which are used to award contract renewals. A natural modeling framework for such practice is a repeated game, which introduces additional methodological challenges but points to another direction for future research. Another issue that can be investigated under our modeling framework is the possibility of double moral hazard due to reckless usage of equipment by the customer. Last but not least, practitioners we communicated with expressed interest in

formalizing insights from stylized economic models into a decision support tool that can aid the negotiation process between customers and suppliers. Clearly, this is an important and difficult problem that requires an explicit model of the multi-echelon, multi-indentured structure of the military supply chain, a direction we wish to pursue in the future. Since the goal of our paper is to gain insights into the principal trade-offs rather than to accurately model every aspect of the problem, we believe that our stylized approach is appropriate as a means to supply practitioners with insights to a problem that they consider to be of utmost importance.

6 Conclusion

In this paper we propose a stylized economic model to evaluate the trade-off between investing in reliability improvement and in spare assets under two contracts that are commonly observed in after-sales support for complex equipment. The motivation for our research comes from the new contracting strategy, Performance-Based Logistics, which is gaining wide acceptance in the defense industry today. Performance-based contracts are designed to replace more conventional consumption-based contracts in an attempt to better align the incentives of customers and suppliers. However, even several years after the Performance-Based Logistics strategy has been announced, significant confusion surrounds the implementation of it. Our conversations with many suppliers to the Department of Defense indicate that they face difficulties estimating the costs and benefits of PBC, whereas this was relatively straightforward under CBC, when suppliers were paid for each resource consumed to support the repair and maintenance activities.

Our theoretical model suggests that CBC is not as effective as PBC in incentivizing suppliers to invest in reliability improvements. Instead, under CBC, suppliers tend to meet the availability target by increasing the inventory of spares. Under PBC, on the other hand, the supplier achieves the availability target by improving reliability as well as by increasing the size of the spares inventory. In general, both contracts result in inefficiencies manifested in less reliable products and more inventory than the first-best solution prescribes. Compared to CBC, however, PBC enables a potential win-win scenario where the products are more reliable and a lower inventory investment is needed.

Moreover, we found that successful implementation CBC and PBC depends crucially on the

asset ownership structure. Our analysis shows that the optimal ownership structures under these two contracting approaches are the opposite: under CBC, it is best if the customer retains the majority of spare assets, whereas under PBC, a full transfer of ownership to the supplier is recommended, if it is viable. Under PBC, incentives between the two parties are better aligned with ownership transfer if the supplier fully internalizes the cost of maintaining the physical resources that are used to support after-sales services, leading to the maximum levels of product reliability and savings in material use. Therefore, our analysis advocates giving suppliers full ownership responsibility and thereby transforming them into total service providers. When this is done, our model suggests that PBC will achieve the first-best solution, thus coordinating the supply chain. Practical implementation of our policy recommendation will not be straightforward, however, since many military customers believe that asset ownership protects them from mismanagement by the supplier and endows them with more control over fleet availability. Despite such difficulties, we see evidence that customer organizations are moving towards increased levels of asset ownership transfer to their suppliers. For example, in a case we are familiar with, a foreign military service is currently in negotiation with one of its U.S. aircraft suppliers to transfer the title of its spares assets.

This paper contributes to the literature by highlighting the role of product reliability and its relationship with inventory in an after-sales product support environment, where a natural misalignment of incentives is present. That is, a supplier who gets compensated for the resources required to provide support services increases profitability by having more frequent product failures. We believe that this paper represents the first attempt at modeling endogenous reliability improvement in an after-sales service context, and we hope that it will spur further research of this important aspect of service operations.

Finally, our study generates at least three simple hypotheses that naturally lend themselves to empirical examinations. We predict that PBC will result in greater product reliability, lower inventory, and lower contracting cost than CBC does. Further, we predict that PBC with a larger share of inventory owned by the supplier will result in greater efficiency. It remains to be seen whether these predictions hold in practice since a host of other issues is at play. However, in a separate related study that analyzes data provided by an commercial aircraft engine manufacturer, we have confirmed empirically that the use of PBC significantly increases mean time between

unscheduled repairs, which we interpret as a proxy for reliability [16]. This provides evidence that confirms one of the hypotheses generated from our analysis in this paper. Work is ongoing to test the remaining hypotheses and we hope that our theoretical work in this paper will spur more empirical research in this area.

References

- [1] Baiman, S., P.E. Fischer, M.V. Rajan. 2001. Performance measurement and design in supply chains. *Management Science*, 2001, 47(1), 173-188.
- [2] Baiman, S., S. Netessine, R. Saouma. 2010. Informativeness, incentive compensation and the choice of inventory buffer. *The Accounting Review*, forthcoming.
- [3] Cachon, G. P. and M. A. Lariviere. 2001. Contracting to assure Supply: How to share demand forecasts in a supply chain. *Management Science*, 47(5), 629-646.
- [4] Cachon, G. P. 2003. Supply chain coordination with contracts. *Handbooks in Operations Research and Management Science: Supply Chain Management*. eds. Graves, S. and T. de Kok. North Holland.
- [5] Cachon, G. P. 2004. The Allocation of inventory risk in a supply chain: Push, pull, and advance-purchase discount contracts. *Management Science*, 50(2), 222-238.
- [6] Cohen, M. A., P. R. Kleindorfer, H. L. Lee. 1989. Near-optimal service constrained stocking policies for service parts. *Operations Research*, 37(1), 104-117.
- [7] Cohen, M. A., N. Agrawal, V. Agrawal. 2006. Winning in the aftermarket. *Harvard Business Review*, 84(5), 129-38.
- [8] Cohen, M. A., P. Kamesam, P. Kleindorfer, H. Lee, A. Tekerian. 1990. OPTIMIZER: A multi-echelon inventory system for service logistics management. *Interfaces*, 20(1), 65-82.
- [9] Cummins, J. M. 1977. Incentive contracting for national defense: A problem of optimal risk sharing. *Bell Journal of Economics*, 8, 168-185.

- [10] Department of Defense. 2003. Department of Defense Directive 5000.1. <http://www.dtic.mil/whs/directives/corres/html/50001.htm>.
- [11] Department of Defense. 2005. DoD Guide for achieving reliability, availability, and maintainability. <http://www.dote.osd.mil/reports/RAMGuide.pdf>.
- [12] Geary, S. 2006. Ready for combat. *DC Velocity*, 4(7), 75-80.
- [13] Gibbons, R. 2005. Incentives between firms (and within). *Management Science*, 51(1), 2-17.
- [14] Government Accountability Office. 2003. Best practices: Setting requirements differently could reduce weapon systems' total ownership costs. GAO-03-57. <http://www.gao.gov/new.items/d0357.pdf>.
- [15] Government Accountability Office. 2008. Improved analysis and cost data needed to evaluate the cost-effectiveness of Performance Based Logistics. GAO report, GAO-07-234. <http://www.gao.gov/products/GAO-09-41>.
- [16] Guajardo, J., M. Cohen, S.-H. Kim, Netessine. 2010. Impact of performance-based contracting on product reliability: An empirical analysis. Working paper, University of Pennsylvania.
- [17] Hasija, S., E. Pinker, R. Shumsky. 2008. Call center outsourcing contracts under information asymmetry. *Management Science*, 54(4), 793-807.
- [18] Holmström, B., P. Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7, 24-52.
- [19] Kang, K., K. H. Doerr, U. Apte, M. Boudreau. 2005. Decisions support models for valuing improvements in component reliability and maintenance. Working paper, Naval Postgraduate School.
- [20] Kim, S.-H. 2010. Incentives in multi-indenture service supply chains. Working paper, Yale University.
- [21] Kim, S.-H., M. A. Cohen, S. Netessine. 2007. Performance contracting in after-sales service supply chains. *Management Science*, 53(12), 1843-1858.

- [22] Kim, S.-H., M. A. Cohen, S. Netessine, S. Veeraraghavan. 2010. Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science*, forthcoming.
- [23] Laffont, J.-J., J. Tirole. 1986. Using cost observation to regulate firms. *Journal of Political Economy*, 94(3), 614-641.
- [24] Lariviere, M. A., E. L. Porteus. 2001. Selling to the newsvendor: An analysis of price-only contracts. *Manufacturing & Service Operations Management*, 3(4), 293-305.
- [25] Lu, X. L., J. A. Van Mieghem, R. C. Savaskan. 2009. Incentives for quality through endogenous routing. *Manufacturing and Service Operations Management*, 11(2), 254-273
- [26] Muckstadt, J. A. 2005. *Analysis and Algorithms for Service Parts Supply Chains*. Springer.
- [27] Ren, Z. J., Y.-P. Zhou. 2008. Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2), 369-383.
- [28] Rogerson, W. P. 1994. Economic incentives and the defense procurement process. *Journal of Economic Perspectives*, 8(4), 65-90.
- [29] Sherbrooke, C. C. 1968. METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, 16, 122-141.
- [30] Sherbrooke, C. C. 2004. *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*. Springer.
- [31] Zipkin, P. H. 2000. *Foundations of Inventory Management*. McGraw-Hill.

Appendix

A Mathematical Preliminaries

In the proofs, we use the following conventions related to Normal approximation. To circumvent the conceptual difficulty of having negative s , we will regard 0 as the lower bound on s and, as a consequence, we *define* the lower bound on z as $\underline{z} \equiv (0 - 1/\tau)/\sqrt{1/\tau} = -1/\sqrt{\tau}$. This definition does not cause a problem because all quantities of our interest on the domain $(-\infty, -1/\sqrt{\tau})$ are insignificant in the range of τ defined in (1). Thus,

$$\sqrt{\tau}\phi(\underline{z}) \simeq 0, \quad \Phi(\underline{z}) \simeq 0. \quad (\text{A.1})$$

These approximations require us to use the following conventions

$$L(\underline{z}) = \phi(\underline{z}) - \underline{z}\bar{\Phi}(\underline{z}) = [\sqrt{\tau}\phi(\underline{z}) + \bar{\Phi}(\underline{z})] / \sqrt{\tau} \simeq 1/\sqrt{\tau} = -\underline{z}, \quad (\text{A.2})$$

$$\Phi^{-1}(0) \simeq \underline{z} = -1/\sqrt{\tau}, \quad (\text{A.3})$$

in order to be consistent with (A.1). Throughout the paper we replace the notation \simeq with an equality, with an understanding that some of them represent approximations. The following results are stated here for convenience.

Lemma A.1 *For τ satisfying (1),*

$$(i) \quad \frac{\partial}{\partial s} E[B | \tau, s] = -\bar{\Phi}(z) < 0, \quad \frac{\partial}{\partial \tau} E[B | \tau, s] = -\frac{\phi(z)}{2\tau^{3/2}} - \frac{\bar{\Phi}(z)}{\tau^2} < 0,$$

$$(ii) \quad \frac{\partial}{\partial s} E[I | \tau, s] = \Phi(z) > 0, \quad \frac{\partial}{\partial \tau} E[I | \tau, s] = -\frac{\phi(z)}{2\tau^{3/2}} + \frac{\Phi(z)}{\tau^2} > 0.$$

In the next lemma we state the property of an arbitrary distribution that has an increasing failure rate (IFR) property (and hence has general applicability beyond the model presented in this paper)¹¹ that becomes useful in the proofs.

¹¹The quantity $\omega(x)$ defined in Lemma A.2 arises frequently in game-theoretic supply chain models (for example, see [5]).

Lemma A.2 Let X be a random variable with an IFR property whose pdf g is differentiable and vanishes at both extremes of its support $[\underline{y}, \bar{y}]$, where $\underline{y} = -\infty$ and $\bar{y} = \infty$ are permitted. Let G be the cdf of X and $\bar{G}(\cdot) \equiv 1 - G(\cdot)$. Then $\omega(y) \equiv g(y)E[(X - y)^+]/[\bar{G}(y)]^2 \leq 1$.

Proof. Note that IFR means $\frac{d}{dy} \left(\frac{g(y)}{\bar{G}(y)} \right) = (g'(y)\bar{G}(y) + [g(y)]^2) / [\bar{G}(y)]^2 \geq 0$, which in turn implies

$$-g'(y)/g(y) \leq g(y)/\bar{G}(y). \quad (\text{A.4})$$

It can easily be shown that $\omega'(y) \geq 0$ (a similar result for a distribution exhibiting increasing *generalized* failure rate was shown in Cachon 2004). To derive the upper bound, we only need to show $m \equiv \lim_{y \rightarrow \bar{y}} \omega(y) \leq 1$. Since m is of 0/0 form, we apply l'Hopital's rule to obtain:

$$\begin{aligned} m &= \lim_{y \rightarrow \bar{y}} \frac{g(y)E[(X - y)^+]}{[\bar{G}(y)]^2} = \lim_{y \rightarrow \bar{y}} \frac{g'(y)E[(X - y)^+] - g(y)\bar{G}(y)}{-2g(y)\bar{G}(y)} = \frac{1}{2} - \frac{1}{2} \lim_{y \rightarrow \bar{y}} \frac{g'(y)E[(X - y)^+]}{g(y)\bar{G}(y)} \\ &\leq \frac{1}{2} + \frac{1}{2} \lim_{y \rightarrow \bar{y}} \frac{g(y)E[(X - y)^+]}{[\bar{G}(y)]^2} = \frac{1}{2} + \frac{m}{2}, \end{aligned}$$

where we have used (A.4) to prove the inequality. $m \leq 1$ follows by rearranging both sides. ■

B Proofs and Auxiliary Results

Proof of Proposition 1. From (3)-(6), we can write the Lagrangian of (FB) as $\mathcal{L}^{FB}(\tau, s) = \underline{u} - \theta\beta + h_g N + K(\tau) + (\kappa - h_g + h_b)/\tau + (c + h_g)s + \theta E[B | \tau, s]$, where θ is the Lagrangian multiplier. We first show that the the assumptions (7) and (9) require $\theta > c + h_g$ at the optimum. Suppose, to the contrary, $\theta \leq c + h_g$. Differentiating \mathcal{L}^{FB} with respect to s (see Lemma A.1) yields $\frac{\partial \mathcal{L}^{FB}}{\partial s} = c + h_g - \theta \bar{\Phi}(z)$ and $\frac{\partial^2 \mathcal{L}^{FB}}{\partial s^2} = \theta \sqrt{\tau} \phi(z) > 0$. With $\theta \leq c + h_g$, $\frac{\partial \mathcal{L}^{FB}}{\partial s} > 0$ and therefore it is optimal to choose $s = 0$. Consequently, the Lagrangian becomes $\mathcal{L}^{FB}(\tau, 0) = \underline{u} - \theta\beta + h_g N + K(\tau) + (\kappa - h_g + h_b + \theta)/\tau$. If $\kappa - h_g + h_b + \theta \leq 0$, it is optimal to choose $\tau = \underline{\tau}$ since $\mathcal{L}^{FB}(\tau, 0)$ increases in τ . However, at $\tau = \underline{\tau}$ and $s = 0$ the backorder constraint becomes $E[B | \underline{\tau}, 0] = 1/\underline{\tau} \leq \beta$, violating (7). If $\kappa - h_g + h_b + \theta > 0$, on the other hand, the optimal τ is found from the first-order condition $\tau^2 K'(\tau) = \kappa - h_g + h_b + \theta$. Since the backorder constraint requires $E[B | \tau, 0] = 1/\tau \leq \beta$ or $\tau \geq 1/\beta$, the optimal τ should satisfy $\kappa - h_g + h_b + \theta = \tau^2 K'(\tau) \geq (1/\beta)^2 K'(1/\beta)$, as the function $\varphi(\tau) \equiv \tau^2 K'(\tau)$ is increasing. However, rearranging the terms in this inequality and applying (9)

leads to $\theta \geq (1/\beta)^2 K'(1/\beta) - \kappa + h_g - h_b > c + h_g$, which contradicts the assumption $\theta \leq c + h_g$. Hence, we should have $\theta > c + h_g$ in order to be consistent with the assumptions (7) and (9).

Note that $\theta > c + h_g > 0$ implies that the backorder constraint binds, i.e., $E[B | \tau, s] = L(z)/\sqrt{\tau} = \beta$, at the optimum. By setting $\partial \mathcal{L}^{FB}/\partial s = 0$, we find that the optimality condition for s is independent of τ : $z = z(\theta) \equiv \Phi^{-1}\left(1 - \frac{c+h_g}{\theta}\right)$. Hence, by considering the (τ, z) space instead of the (τ, s) space after applying the z -transform (2), we see that the optimal solution lies on the horizontal line $z = z(\theta)$. As a result, the original problem of finding the optimal (τ, s) is transformed to the problem of finding the optimal τ with the restriction $z = z(\theta)$ (call it $\tau(\theta)$) and identifying θ that satisfies the binding backorder constraint $L(z(\theta))/\sqrt{\tau(\theta)} = \beta$. To this end, fix θ and let $\tilde{\mathcal{L}}^{FB}(\tau)$ be the reduced Lagrangian with $z = z(\theta)$. Using the relations $s = 1/\tau + z/\sqrt{\tau}$ and $L(z) = \phi(z) - z\bar{\Phi}(z)$, we can write $\tilde{\mathcal{L}}^{FB}(\tau)$ as

$$\tilde{\mathcal{L}}^{FB}(\tau) = \underline{u} - \theta\beta + h_g N + K(\tau) + (\kappa + c + h_b)/\tau + \theta\phi(z(\theta))/\sqrt{\tau}.$$

Differentiating this yields

$$\begin{aligned} \frac{d\tilde{\mathcal{L}}^{FB}}{d\tau} &= K'(\tau) - \frac{\kappa + c + h_b}{\tau^2} - \frac{\theta\phi(z(\theta))}{2\tau^{3/2}}, \\ \frac{d^2\tilde{\mathcal{L}}^{FB}}{d\tau^2} &= K''(\tau) + \frac{2(\kappa + c + h_b)}{\tau^3} + \frac{3\theta\phi(z(\theta))}{4\tau^{5/2}} > 0. \end{aligned}$$

In addition, we have $\lim_{\tau \rightarrow \bar{\tau}} \tilde{\mathcal{L}}^{FB}(\tau) = \infty$ and

$$\left. \frac{d\tilde{\mathcal{L}}^{FB}}{d\tau} \right|_{\tau=\underline{\tau}} = K'(\underline{\tau}) - \frac{\kappa + c + h_b}{\underline{\tau}^2} - \frac{\theta\phi(z(\theta))}{2\underline{\tau}^{3/2}} < -\frac{c + h_g}{\underline{\tau}^2} - \frac{\theta\phi(z(\theta))}{2\underline{\tau}^{3/2}} < 0,$$

where the first inequality comes from (8). Since $\tilde{\mathcal{L}}^{FB}(\tau)$ is convex and eventually increases after an initial decrease at $\tau = \underline{\tau}$, there is a unique global minimum $\tau(\theta) > \underline{\tau}$ that solves the first-order condition $\frac{d\tilde{\mathcal{L}}^{FB}}{d\tau} = 0$ for any given $\theta > c + h_g$. It remains to show that there is a unique value of θ that solves the binding backorder constraint $L(z(\theta))/\sqrt{\tau(\theta)} = \beta$ and verify that this value satisfies the condition $\theta > c + h_g$. For notational convenience, define $b(\theta) \equiv L(z(\theta))/\sqrt{\tau(\theta)}$. Via implicit

differentiation of the first-order condition $\frac{d\tilde{\mathcal{L}}^{FB}}{d\tau} = 0$ with respect to θ ,

$$\left(K''(\tau) + \frac{2(\kappa + c + h_b)}{\tau^3} + \frac{3\theta\phi(z(\theta))}{4\tau^{5/2}} \right) \tau'(\theta) = \frac{L(z(\theta))}{2\tau^{3/2}},$$

from which it is clear that $\tau'(\theta) > 0$. Combining this with $L'(z) < 0$ and $z'(\theta) > 0$, we conclude that $b(\theta)$ is decreasing. In addition, $\lim_{\theta \rightarrow c+h_g} b(\theta) = 1/\tau(c+h_g)$ (see (A.2)) and $\lim_{\theta \rightarrow \infty} b(\theta) = 0$. In other words, $b(\theta)$ decreases from $1/\tau(c+h_g)$ to 0 as θ increases from $c+h_g$ to infinity. So there is a unique $\theta > c+h_g$ that solves $b(\theta) = \beta$ if $1/\tau(c+h_g) > \beta$. To see that this sufficient condition is satisfied under the model assumptions, let $\theta \rightarrow c+h_g$ in the first-order condition $d\tilde{\mathcal{L}}^{FB}/d\tau = 0$ to obtain $[\tau(c+h_g)]^2 K'(\tau(c+h_g)) = \kappa + c + h_b < (1/\beta)^2 K'(1/\beta)$, where the inequality is from (9). Since $\varphi(\tau) = \tau^2 K'(\tau)$ is an increasing function, $\varphi(\tau(c+h_g)) < \varphi(1/\beta)$ implies $\tau(c+h_g) < 1/\beta$, which is the desired condition. Finally, the optimality condition for τ in the proposition, $\Gamma(\tau) = K'(\tau)$, is obtained by inverting $z(\theta) = \Phi^{-1}\left(1 - \frac{c+h_g}{\theta}\right)$ and $L(z(\theta))/\sqrt{\tau} = \beta$ to get $\theta = (c+h_g)/\bar{\Phi}(z)$ and $z = L^{-1}(\beta\sqrt{\tau})$, and substituting them to the first-order condition $d\tilde{\mathcal{L}}^{FB}/d\tau = 0$. ■

Proof of Lemma 1. The supplier's expected profit under CBC is $u(\tau, s) = E[T | \tau, s] - \psi(\tau, s) = w - K(\tau) + (r - \kappa)/\tau + (p - c - \delta h_b)s - \delta(h_g - h_b)E[I | \tau, s]$. Differentiating this with respect to s , (see Lemma A.1) we have $\frac{\partial u}{\partial s} = p - c - \delta h_b - \delta(h_g - h_b)\Phi(z)$ and $\frac{\partial^2 u}{\partial s^2} = -(v + \delta(h_g - h_b))\sqrt{\tau}\phi(z) < 0$, i.e., $u(\tau, s)$ is concave in s for a fixed τ . Suppose $p \leq c + \delta h_b$. Then $\frac{\partial u}{\partial s} < 0$ and therefore it is optimal to choose $s = 0$. Then $u(\tau, s)$ becomes $u(\tau, 0) = w - K(\tau) + (r - \kappa)/\tau$, which can be shown to be quasiconcave. Suppose $r \geq \kappa - \underline{\tau}^2 K'(\underline{\tau})$. (Recall $\kappa - \underline{\tau}^2 K'(\underline{\tau}) > 0$ by (8)) Then $\left. \frac{\partial u(\tau, 0)}{\partial \tau} \right|_{\tau=\underline{\tau}} \leq 0$, and therefore, it is optimal for the supplier to choose $\tau = \underline{\tau}$. In this case, however, the backorder constraint $E[B | \underline{\tau}, 0] = 1/\underline{\tau} \leq \beta$ is inconsistent with the assumption (7). Hence, the customer should offer $0 \leq r < \kappa - \underline{\tau}^2 K'(\underline{\tau})$. Then the supplier optimally chooses $\tau^* > \underline{\tau}$ that solves the equation $\tau^2 K'(\tau) = \kappa - r$. This solution should satisfy the backorder constraint $E[B | \tau^*, 0] = 1/\tau^* \leq \beta$. Since $1/\beta \leq \tau^*$ and $\varphi(\tau) \equiv \tau^2 K'(\tau)$ is increasing, we have $(1/\beta)^2 K'(1/\beta) \leq (\tau^*)^2 K'(\tau^*) = \kappa - r$. Together with (9), this implies $\kappa + c + h_b < \kappa - r$ or $r < -(c + h_b) < 0$, again leading to an infeasible value of r ; as long as $r \geq 0$, the backorder constraint is not satisfied in this case, either. Hence, we conclude that a feasible solution does not exist if $p \leq c + \delta h_b$. Next, suppose $p \geq c + \delta h_g$. Then $\frac{\partial u}{\partial s} \geq \delta(h_g - h_b)\bar{\Phi}(z) > 0$, implying that the supplier chooses $s \rightarrow \infty$. So a finite solution does not

exist if $p \geq c + \delta h_g$. This argument leaves $c + \delta h_b < p < c + \delta h_g$ as the only range in which a finite feasible solution may exist. ■

Proof of Lemma 2. As shown in the proof of Lemma 1, the supplier's expected profit $u(\tau, s) = w - K(\tau) + (r - \kappa)/\tau + (p - c - \delta h_b)s - \delta(h_g - h_b)E[I|\tau, s]$ is concave in s and has the derivative $\frac{\partial u}{\partial s} = p - c - \delta h_b - \delta(h_g - h_b)\Phi(z)$. With $\delta > 0$ and $c + \delta h_b < p < c + \delta h_g$, the profit-maximizing s is found in the interior. The first-order condition $\partial u/\partial s = 0$ yields $z^* = \Phi^{-1}\left(1 - \frac{c + \delta h_g - p}{\delta(h_g - h_b)}\right)$, which is independent of τ . Substituting $s^* = 1/\tau + z^*/\sqrt{\tau}$, $u(\tau, s)$ becomes $\tilde{u}(\tau) \equiv u(\tau, s^*) = w - K(\tau) + (r + p - \kappa - c - \delta h_b)/\tau - \delta(h_g - h_b)\phi(z^*)/\sqrt{\tau}$. Note that $\lim_{\tau \rightarrow \bar{\tau}} \tilde{u}(\tau) = -\infty$. Differentiating $\tilde{u}(\tau)$,

$$\begin{aligned}\tilde{u}'(\tau) &= -K'(\tau) - \frac{r + p - \kappa - c - \delta h_b}{\tau^2} + \frac{\delta(h_g - h_b)\phi(z^*)}{2\tau^{3/2}}, \\ \tilde{u}''(\tau) &= -K''(\tau) + \frac{2(r + p - \kappa - c - \delta h_b)}{\tau^3} - \frac{3\delta(h_g - h_b)\phi(z^*)}{4\tau^{5/2}}.\end{aligned}$$

Let $\hat{\tau}$ be the solution of the first-order condition $\tilde{u}'(\tau) = 0$. Multiplying this condition by $2/\hat{\tau}$ and adding it to $\tilde{u}''(\tau)$, we get

$$\begin{aligned}\tilde{u}''(\hat{\tau}) &= -K''(\hat{\tau}) - \frac{2K'(\hat{\tau})}{\hat{\tau}} + \frac{\delta(h_g - h_b)\phi(z^*)}{4\hat{\tau}^{5/2}} \leq -\frac{2(h_g - h_b)}{\hat{\tau}^3} - \frac{2K'(\hat{\tau})}{\hat{\tau}} + \frac{h_g - h_b}{4\sqrt{2\pi}\hat{\tau}^3} \\ &= -\frac{2K'(\hat{\tau})}{\hat{\tau}} - \left(2 - \frac{1}{4\sqrt{2\pi}}\right) \frac{h_g - h_b}{\hat{\tau}^3} \approx -\frac{2K'(\hat{\tau})}{\hat{\tau}} - (1.9) \frac{h_g - h_b}{\hat{\tau}^3} < 0,\end{aligned}$$

where the first inequality follows from $\delta \leq 1$ and the upper bound on the standard normal pdf, i.e., $\phi(z) \leq (2\pi\tau)^{-1/2}$. Suppose that the solution is in the interior, i.e., $\underline{\tau} < \hat{\tau} < \bar{\tau}$. Then $\tilde{u}''(\hat{\tau}) < 0$ implies that any interior critical point, if it exists, should be a maximizer. Let us consider two cases: $\tilde{u}'(\underline{\tau}) > 0$ and $\tilde{u}'(\underline{\tau}) \leq 0$. If $\tilde{u}'(\underline{\tau}) > 0$ then $\hat{\tau} > \underline{\tau}$ since $\tilde{u}(\tau)$ initially increases and approaches $-\infty$ as $\tau \rightarrow \bar{\tau}$. Therefore $\hat{\tau}$ is a unique maximizer since more than one maximizer requires at least one interior minimizer (as $\tilde{u}(\tau)$ is continuous), which contradicts our earlier observation that any interior critical point should be a maximizer. Now suppose $\tilde{u}'(\underline{\tau}) \leq 0$. Again, if an interior critical point exists, then it should be a maximizer. But this means a minimizer should exist to the left of the maximizer, since $\tilde{u}(\tau)$ initially decreases. This leads to a contradiction, and therefore, no interior critical point exist in this case; $\tilde{u}(\tau)$ decreases monotonically if $\tilde{u}'(\underline{\tau}) \leq 0$. Then $\tilde{u}(\tau)$

is maximized at $\tau = \underline{\tau}$. Summarizing, the supplier uniquely chooses $\tau^* = \underline{\tau}$ if $\tilde{u}'(\underline{\tau}) \leq 0$ and $\tau^* = \hat{\tau} > \underline{\tau}$ if $\tilde{u}'(\underline{\tau}) > 0$, where $\hat{\tau}$ is obtained from the first-order condition (13).

To establish a connection between this result and the three cases stated in the lemma, observe that, for $p_{\min} \equiv c + \delta h_b < p < c + \delta h_g \equiv p_{\max}$ and a fixed τ ,

$$\begin{aligned}\bar{m}(\tau) &\equiv \lim_{p \rightarrow p_{\min}} \tilde{u}'(\tau) = -K'(\tau) - \frac{r - \kappa}{\tau^2}, \\ \underline{m}(\tau) &\equiv \lim_{p \rightarrow p_{\max}} \tilde{u}'(\tau) = -K'(\tau) - \frac{r - \kappa + \delta(h_g - h_b)}{\tau^2} < \bar{m}(\tau), \\ \frac{\partial \tilde{u}'(\tau)}{\partial p} &= -\frac{1}{\tau^2} - \frac{z^*}{2\tau^{3/2}} = -\frac{1}{2\tau^2} - \frac{s^*}{2\tau} < 0.\end{aligned}$$

At $\tau = \underline{\tau}$, $\bar{m}(\underline{\tau}) = (b - r)/\underline{\tau}^2$ and $\underline{m}(\underline{\tau}) = (a - r)/\underline{\tau}^2$, where a and b are defined in the lemma. If $0 \leq r \leq a$, we have $\bar{m}(\underline{\tau}) > \underline{m}(\underline{\tau}) \geq 0$. Since $\tilde{u}'(\underline{\tau})$ decreases in p from $\bar{m}(\tau) > 0$ to $\underline{m}(\tau) \geq 0$, $\tilde{u}'(\underline{\tau}) > 0$ for all $p \in (p_{\min}, p_{\max})$, and therefore, $\tau^* = \hat{\tau} > \underline{\tau}$. Next, assume $a < r < b$. Then $\bar{m}(\underline{\tau}) > 0$ and $\underline{m}(\underline{\tau}) < 0$, implying that there exists $\bar{p}(r) \in (p_{\min}, p_{\max})$ such that $\tilde{u}'(\underline{\tau}) > 0$ for $p \in (p_{\min}, \bar{p}(r))$ and $\tilde{u}'(\underline{\tau}) \leq 0$ for $p \in [\bar{p}(r), p_{\max})$. As we found above, $\tau^* = \hat{\tau} > \underline{\tau}$ in the former case and $\tau^* = \underline{\tau}$ in the latter case. $\bar{p}(r)$ is determined from the equation $\tilde{u}'(\underline{\tau}) = 0$. Finally, assume $r \geq b$. Then we have $0 \geq \bar{m}(\underline{\tau}) > \underline{m}(\underline{\tau})$, which implies that $\tilde{u}'(\underline{\tau})$ remains in the negative region as p increases from p_{\min} to p_{\max} . So $\tilde{u}'(\underline{\tau}) \leq 0$ for all $p \in (p_{\min}, p_{\max})$. Then by the finding above, $\tau^* = \underline{\tau}$.

To prove the stated sensitivity results, first note that $\frac{\partial z^*}{\partial p} = [\delta(h_g - h_b)\phi(z^*)]^{-1} > 0$ and $\frac{\partial z^*}{\partial r} = 0$. Using these and via implicit differentiation of the first-order condition $\tilde{u}'(\hat{\tau}) = 0$, we get

$$\frac{\partial \hat{\tau}}{\partial r} = -\frac{1}{\hat{\tau}^2} < 0 \quad \text{and} \quad \frac{\partial \hat{\tau}}{\partial p} = \left(\frac{1}{\hat{\tau}^2} + \frac{z^*}{2\hat{\tau}^{3/2}} \right) \frac{1}{\tilde{u}''(\hat{\tau})} = \left(\frac{1}{2\hat{\tau}^2} + \frac{s^*}{2\hat{\tau}} \right) \frac{1}{\tilde{u}''(\hat{\tau})} < 0.$$

Also,

$$\frac{\partial s^*}{\partial r} = \frac{1}{\sqrt{\hat{\tau}}} \frac{\partial z^*}{\partial r} = 0 \quad \text{and} \quad \frac{\partial s^*}{\partial p} = -\left(\frac{1}{2\hat{\tau}^2} + \frac{s^*}{2\hat{\tau}} \right) \frac{\partial \hat{\tau}}{\partial p} + \frac{1}{\sqrt{\hat{\tau}}} \frac{\partial z^*}{\partial p} > 0.$$

In addition, it is easy to verify $\frac{\partial \tau^*}{\partial r} = \frac{\partial \tau^*}{\partial p} = \frac{\partial s^*}{\partial r} = 0$ and $\frac{\partial s^*}{\partial p} > 0$ when $\tau^* = \underline{\tau}$. ■

The following auxiliary lemma is used in the proof of Proposition 2.

Lemma B.1 *Under CBC, one of the following three outcomes emerges in equilibrium, along with*

the condition $L(z^*)/\sqrt{\tau^*} = \beta$: (i) $\tau^* > \underline{\tau}$ that solves (13) with $r = 0$, (ii) $\tau^* > \underline{\tau}$ that solves $(\tau^*)^2 K'(\tau^*) = \kappa - h_g + h_b$ and (13), or (iii) $\tau^* = \underline{\tau}$.

Proof. As shown in Lemma 2, the supplier chooses either $\tau^* = \underline{\tau}$ or $\tau^* = \hat{\tau} > \underline{\tau}$ as specified in (13) with $s^* = 1/\tau^* + z^*/\sqrt{\tau^*}$, where $z^* = \Phi^{-1}\left(1 - \frac{c+\delta h_g - p}{\delta(h_g - h_b)}\right)$. We consider these two cases separately as they require different analyses.

First suppose that $\tau^* = \hat{\tau} > \underline{\tau}$ in equilibrium. The corresponding backorder constraint is $E[B|\hat{\tau}, s^*] = L(z^*)/\sqrt{\hat{\tau}} \leq \beta$. According to Lemma 2, $\tau^* = \hat{\tau} > \underline{\tau}$ is possible only in the following domain in the (r, p) space:

$$\{\{0 \leq r \leq a\} \cap \{c + \delta h_b < p < c + \delta h_g\}\} \cup \{\{a < r < b\} \cap \{c + \delta h_b < p < \bar{p}(r)\}\}, \quad (\text{B.1})$$

where $\bar{p}(r)$ is the boundary value of p for a fixed r over which $\tau^* = \underline{\tau}$ is optimal. Note that

$$\frac{\partial E[B|\hat{\tau}, s^*]}{\partial r} = -\frac{L(z^*)}{2\hat{\tau}^{3/2}} \frac{\partial \hat{\tau}}{\partial r} > 0, \quad (\text{B.2})$$

$$\frac{\partial E[B|\hat{\tau}, s^*]}{\partial p} = -\frac{L(z^*)}{2\hat{\tau}^{3/2}} \frac{\partial \hat{\tau}}{\partial p} - \frac{\bar{\Phi}(z^*)}{\sqrt{\hat{\tau}}} \frac{\partial z^*}{\partial p} = \frac{(1/\hat{\tau} + s^*) L(z^*)}{4\hat{\tau}^{5/2} (-\tilde{u}''(\hat{\tau}))} - \frac{\bar{\Phi}(z^*)}{\delta(h_g - h_b) \sqrt{\hat{\tau}} \phi(z^*)}, \quad (\text{B.3})$$

where we have used the results found in the proof of Lemma 2. The expression of $\tilde{u}''(\hat{\tau})$ is also found there. To determine the sign of $\frac{\partial}{\partial p} E[B|\hat{\tau}, s^*]$, the following two intermediate results are needed.

First observe that

$$\begin{aligned} -\tilde{u}''(\hat{\tau}) &= K''(\hat{\tau}) + \frac{2(\kappa + c + \delta h_b - r - p)}{\hat{\tau}^3} + \frac{3\delta(h_g - h_b)\phi(z^*)}{4\hat{\tau}^{5/2}} \\ &> K''(\hat{\tau}) + \frac{2(\underline{\tau}^2 K'(\underline{\tau}) + c + \delta h_b - p)}{\hat{\tau}^3} + \frac{3(c + \delta h_g - p)}{4\hat{\tau}^{5/2}} z^* \\ &= K''(\hat{\tau}) + \frac{2\underline{\tau}^2 K'(\underline{\tau})}{\hat{\tau}^3} + \frac{5c + 8\delta h_b - 3\delta h_g - 5p}{4\hat{\tau}^3} + \frac{3(c + \delta h_g - p)}{4\hat{\tau}^2} s^*, \end{aligned}$$

where we have used $L(z^*) = \phi(z^*) - z^* \bar{\Phi}(z^*) = \phi(z^*) - \frac{c+\delta h_g - p}{\delta(h_g - h_b)} z^* > 0$ and $r < \kappa - \underline{\tau}^2 K'(\underline{\tau})$ in (B.1) to establish the inequality. The last expression is obtained by substituting $z^* = \sqrt{\hat{\tau}} s^* - 1/\sqrt{\hat{\tau}}$ and arranging the terms. From the assumptions (10) and $K'''(\tau) > 0$, we have $\hat{\tau}^3 K''(\hat{\tau}) > \underline{\tau}^3 K''(\underline{\tau}) \geq$

$2(h_g - h_b) \geq 2\delta(h_g - h_b)$. Applying this, we get

$$-\tilde{u}''(\hat{\tau}) > \frac{2\underline{\tau}^2 K'(\underline{\tau})}{\hat{\tau}^3} + \frac{5(c + \delta h_g - p)}{4\hat{\tau}^3} + \frac{3(c + \delta h_g - p)}{4\hat{\tau}^2} s^* > \frac{c + \delta h_g - p}{4\hat{\tau}^2} \left(\frac{1}{\hat{\tau}} + s^* \right). \quad (\text{B.4})$$

Note that the last expression is positive since $p < c + \delta h_g$ under the conditions defined in Lemma 2. Next, from Lemma A.2, we get

$$L(z^*) \leq \left(\frac{c + \delta h_g - p}{\delta(h_g - h_b)} \right)^2 \frac{1}{\phi(z^*)}. \quad (\text{B.5})$$

Substituting (B.4) and (B.5) in (B.3), it is straightforward to show $\frac{\partial}{\partial p} E[B|\hat{\tau}, s^*] < 0$. Summarizing, we have $\frac{\partial}{\partial r} E[B|\hat{\tau}, s^*] > 0$ and $\frac{\partial}{\partial p} E[B|\hat{\tau}, s^*] < 0$, indicating that the feasible region for the backorder constraint $E[B|\hat{\tau}, s^*] \leq \beta$, if it exists within the domain of the (r, p) space under consideration, (B.1), should be in the upper left corner of that domain. If the feasible region is not in this domain then $\tau^* = \underline{\tau}$ in equilibrium, the case that we analyze below. Note that $\lim_{p \rightarrow c + \delta h_b} E[B|\hat{\tau}, s^*] > \beta$ (as shown in the proof of Lemma 1). This limit suggests that for any fixed r within (B.1), the feasible range of p is bounded below by $\underline{p}(r)$, the value of p that solves $E[B|\hat{\tau}, s^*] = \beta$, and above by either $c + \delta h_g$ or $\bar{p}(r)$, depending on whether or not $r \leq a = \kappa - \underline{\tau}^2 K'(\underline{\tau}) - \delta(h_g - h_b)$. In particular, the limit $\lim_{p \rightarrow c + \delta h_g} E[B|\hat{\tau}, s^*] = 0$ implies that the feasible region $E[B|\hat{\tau}, s^*] \leq \beta$ overlaps with (B.1) at least for $r \leq a$. In addition, parts (ii) and (iii) of Lemma 2 imply that as r approaches its upper limit $b = \kappa - \underline{\tau}^2 K'(\underline{\tau})$, $\bar{p}(r)$ approaches its lower limit $c + \delta h_b$ (as defined in (B.1)). Therefore, $\lim_{r \rightarrow b} E[B|\hat{\tau}, s^*] = \lim_{p \rightarrow c + \delta h_b} E[B|\hat{\tau}, s^*] > \beta$. This means that there is $r_{\max} < b$ at which $\underline{p}(r_{\max})$ coincides with $\bar{p}(r_{\max})$, i.e., the backorder constraint binds at $(r_{\max}, \bar{p}(r_{\max}))$. Hence, we have narrowed the feasible region from (B.1) to

$$\{ \{0 \leq r \leq a\} \cap \{ \underline{p}(r) < p < c + \delta h_g \} \} \cup \{ \{a < r \leq r_{\max}\} \cap \{ \underline{p}(r) < p < \bar{p}(r) \} \} \quad \text{s.t.} \quad \underline{p}(r_{\max}) = \bar{p}(r_{\max}). \quad (\text{B.6})$$

Given (B.6), the cost-minimizing price p for a fixed r should be found either (a) as p approaches its upper bound $c + \delta h_g$ or $\bar{p}(r)$, (b) at the lower bound $\underline{p}(r)$, or (c) in between the two bounds, as long as $r < r_{\max}$. (By definition, the lower and upper bounds converge at $r = r_{\max}$ and therefore there is no space in between.) For $0 \leq r \leq a$, case (a) never arises since the customer's expected

cost $C(\hat{\tau}, s^*) = \underline{u} + h_g N + K(\hat{\tau}) + (\kappa - h_g + h_b)/\hat{\tau} + (c + h_g)s^*$ approaches ∞ as $r \rightarrow c + \delta h_g$. For $a < r < r_{\max}$, case (a) leads to $\tau^* = \underline{\tau}$, the case we analyze below. If case (b) is true, the backorder constraint binds at the optimum. If case (c) is true, then there is an interior optimal value of p , i.e., $\frac{\partial C}{\partial p} = 0$ at this point. Differentiating $C(\hat{\tau}, s^*)$ with respect to r and p ,

$$\frac{\partial C}{\partial r} = \left(K'(\hat{\tau}) - \frac{\kappa - h_g + h_b}{\hat{\tau}^2} \right) \frac{\partial \hat{\tau}}{\partial r}, \quad (\text{B.7})$$

$$\frac{\partial C}{\partial p} = \left(K'(\hat{\tau}) - \frac{\kappa - h_g + h_b}{\hat{\tau}^2} \right) \frac{\partial \hat{\tau}}{\partial p} + (c + h_g) \frac{\partial s^*}{\partial p}. \quad (\text{B.8})$$

Since $\frac{\partial \hat{\tau}}{\partial p} < 0$ and $\frac{\partial s^*}{\partial p} > 0$ as we found in Lemma 2, from (B.8), we see that $\frac{\partial C}{\partial p} = 0$ requires $K'(\hat{\tau}) - (\kappa - h_g + h_b)/\hat{\tau}^2 > 0$ for the induced $\hat{\tau}$. But we see from (B.7) that this and $\frac{\partial \hat{\tau}}{\partial r} < 0$ together imply $\frac{\partial C}{\partial r} < 0$, i.e., our chosen value of r is not optimal and we should increase it. Repeating the same argument for each r and the corresponding range of p in the feasible region (B.6), we encounter two possibilities at the end: the cost-minimizing p for a given r is found either (A) at the lower bound $\underline{p}(r)$ or (B) at or beyond the upper bound $\bar{p}(r)$, as long as $a < r \leq r_{\max}$. By definition the backorder constraint binds in case (A). Having identified where the cost-minimizing p is found for each r , let us now find the cost-minimizing r . Like p , there are three candidate points: $r = 0$, $r \in (0, r_{\max})$, or $r = r_{\max}$. If $r = 0$ is optimal, then $\frac{\partial C}{\partial r} \geq 0$ at that point so from (B.7) we should have $K'(\hat{\tau}) - (\kappa - h_g + h_b)/\hat{\tau}^2 \leq 0$. But this implies $\frac{\partial C}{\partial p} > 0$ (see (B.8)), indicating that for $r = 0$ to be optimal p should be at its lower bound $\underline{p}(0)$, where the backorder constraint binds; *this is the case (i) stated in the lemma*. If optimal r is found in the interior, on the other hand, $\frac{\partial C}{\partial r} = 0$ and therefore $K'(\hat{\tau}) - (\kappa - h_g + h_b)/\hat{\tau}^2 = 0$, which again leads to $\frac{\partial C}{\partial p} > 0$. Therefore, another candidate for the optimal solution is found with $r > 0$, at which $K'(\hat{\tau}) - (\kappa - h_g + h_b)/\hat{\tau}^2 = 0$, and $p = \underline{p}(r)$, i.e., the backorder constraint binds; *this is case (ii) in the lemma*. Finally, if $r = r_{\max}$ is optimal, we have $\tau^* = \underline{\tau}$ at that point since $\underline{p}(r_{\max}) = \bar{p}(r_{\max})$. We now turn to the next case $\tau^* = \underline{\tau}$.

Suppose that $\tau^* = \underline{\tau}$ in equilibrium. In this case the feasible region is characterized by the intersection of $r > a$, $p \geq \bar{p}(r)$, and $E[B | \underline{\tau}, s^*] \leq \beta$. Notice that the backorder constraint $E[B | \underline{\tau}, s^*] \leq \beta$ is independent of r since s^* is a function of p only and $\underline{\tau}$ is fixed. Therefore, there is a single value p_{\min} that solves $E[B | \underline{\tau}, s^*] = \beta$, regardless of r . Thus, $E[B | \underline{\tau}, s^*] \leq \beta$ can be expressed as $p \geq p_{\min}$, since $E[B | \underline{\tau}, s^*]$ decreases in p and $\lim_{p \rightarrow c + \delta h_g} E[B | \underline{\tau}, s^*] = 0$. So the feasible region in this case

is

$$\{r > a\} \cap \{p \geq \max\{p_{\min}, \bar{p}(r)\}\}$$

In this region, $\tau^* = \underline{\tau}$. In addition, the customer's cost $C(\underline{\tau}, s^*) = \underline{u} + h_g N + K(\underline{\tau}) + (\kappa - h_g + h_b)/\underline{\tau} + (c + h_g)s^*$ increases in p but remains unchanged in r , as can be easily proved. Therefore, for any given r , $C(\underline{\tau}, s^*)$ has the smallest value at the lower bound $\max\{p_{\min}, \bar{p}(r)\}$. If $\bar{p}(r) \leq p_{\min}$ then the optimal p is p_{\min} that satisfies the binding constraint; *this is case (iii) stated in the lemma*. If $p_{\min} < \bar{p}(r)$, on the other hand, the constraint does not bind at the lower bound $\bar{p}(r)$ as $E[B | \underline{\tau}, s^*] < \beta$ at that point. We prove by contradiction that the customer's cost is not minimized in this case. Suppose otherwise, i.e., there exists r such that the backorder constraint does not bind at $\bar{p}(r)$ and the customer's cost is minimized at $(r, \bar{p}(r))$. For this to be true, $\bar{p}(r)$ should be a local minimizer, i.e., $\left. \frac{\partial}{\partial p} C(\hat{\tau}, s^*) \right|_{p \rightarrow \bar{p}(r)^-} \leq 0$ and $\left. \frac{\partial}{\partial p} C(\underline{\tau}, s^*) \right|_{p \rightarrow \bar{p}(r)^+} \geq 0$. The latter was already verified above. If the former were true, then from (B.8), we should have $K'(\hat{\tau}) - (\kappa - h_g + h_b)/\hat{\tau}^2 > 0$ for p near from but less than $\bar{p}(r)$. But this implies $\left. \frac{\partial}{\partial r} C(\hat{\tau}, s^*) \right|_{p \rightarrow \bar{p}(r)^-} < 0$ from (B.7), so the chosen r is not optimal and we should choose a higher value. Repeating the same argument, we end up at the highest r allowed for $p < \bar{p}(r)$, i.e., r_{\max} . So the customer's cost should be minimized at r_{\max} . However, we showed earlier that the backorder constraint binds at r_{\max} as $\underline{p}(r_{\max}) = \bar{p}(r_{\max})$; this contradicts our assertion that the constraint does not bind at the cost-minimizing point $(r_{\max}, \bar{p}(r_{\max}))$.

After exhausting all possibilities, we conclude that the three cases stated in the lemma are the only candidates for the optimal solution. ■

Proof of Proposition 2. That the backorder constraint binds in equilibrium follows directly from Lemma B.1. Binding constraint reduces the original two-dimensional problem to a single-dimensional one, as the equation $L(z)/\sqrt{\tau} = \beta$ establishes a one-to-one correspondence between τ and s . Writing this condition as $z = L^{-1}(\beta\sqrt{\tau}) = \zeta(\tau)$ and substituting it in the customer's expected cost expression yields the reduced cost function $\tilde{C}(\tau) = \underline{u} + h_g N + K(\tau) + (\kappa + c + h_b)/\tau + (c + h_g)\zeta(\tau)/\sqrt{\tau}$. As expected, this function is minimized at the first-best solution τ^{FB} that we derived in Proposition 1. In the proof of the same proposition we also showed that $\tilde{C}(\tau)$ is convex. (Note that $\tilde{C}(\tau)$ is equivalent to the Lagrangian \mathcal{L}^{FB} with an appropriate choice of the Lagrangian multiplier θ .) It remains to choose the optimal solution among the three candidates we identified

in Lemma B.1 that produces τ^* with the lowest cost $\tilde{C}(\tau^*)$. Consider each case in Lemma B.1.

(i) With $r = 0$, (13), and $z^* = \Phi^{-1}\left(1 - \frac{c + \delta h_g - p}{\delta(h_g - h_b)}\right) = \zeta(\tau)$, we get $p = c + \delta h_b + \delta(h_g - h_b)\Phi(\zeta(\tau))$

and

$$\tau^2 K'(\tau) = \kappa - \delta(h_g - h_b) \left(\Phi(\zeta(\tau)) - \frac{\sqrt{\tau}}{2} \phi(\zeta(\tau)) \right). \quad (\text{B.9})$$

Note that $\Phi(\zeta(\tau)) - \frac{\sqrt{\tau}}{2} \phi(\zeta(\tau)) = 1 - \bar{\Phi}(\zeta(\tau)) - \frac{\sqrt{\tau}}{2} \phi(\zeta(\tau)) < 1$. Hence, the solution of (B.9), which we call τ_1 , satisfies $\tau_1^2 K'(\tau_1) > \kappa - \delta(h_g - h_b)$.

(ii) Optimal τ for this case, called τ_2 , is given by the stated condition $\tau^2 K'(\tau) = \kappa - (h_g - h_b)$. It is clear that $\tau_2^2 K'(\tau_2) = \kappa - (h_g - h_b) < \kappa - \delta(h_g - h_b) < \tau_1^2 K'(\tau_1)$, where the last inequality is from (i) above.

(iii) In this case we have $\tau^* = \underline{\tau}$. From (8), $\underline{\tau}^2 K'(\underline{\tau}) < \kappa - (h_g - h_b)$.

Combining the inequalities we derived above, we have $\varphi(\underline{\tau}) < \varphi(\tau_2) < \varphi(\tau_1)$, where $\varphi(\tau) = \tau^2 K'(\tau)$. Since $\varphi(\tau)$ is increasing, this implies $\underline{\tau} < \tau_2 < \tau_1$. Next, we show $\tau_1 < \tau^{FB}$. The optimality condition (B.9) can be rewritten as

$$\Gamma(\tau_1) - \left(\frac{(1 - \delta)h_b + p}{\tau_1^2} + \frac{(1 - \delta)h_g + p}{2\tau_1^{3/2}} f(\zeta(\tau_1)) \right) = K'(\tau_1). \quad (\text{B.10})$$

The derivative of the customer's cost is $\tilde{C}'(\tau) = K'(\tau) - \Gamma(\tau)$. Evaluating this at τ_1 using (B.10), it is immediate that $\tilde{C}'(\tau_1) < 0$. Since $\tilde{C}(\tau)$ is convex and minimized at τ^{FB} , $\tilde{C}'(\tau_1) < 0$ implies $\tau_1 < \tau^{FB}$. Therefore, we have $\underline{\tau} < \tau_2 < \tau_1 < \tau^{FB}$ and among the three candidate equilibrium outcomes (i)-(iii) above, (i) produces the lowest customer cost at τ_1 . The optimality condition (14) is obtained by rearranging (B.9). ■

Proof of Lemma 3. The supplier's expected profit under PBC is $u(\tau, s) = E[T | \tau, s] - \psi(\tau, s) = w - K(\tau) - \kappa/\tau - (c + \delta h_b)s - \delta(h_g - h_b)E[I | \tau, s] - vE[B | \tau, s]$. Differentiating this with respect to s (see Lemma A.1), we get $\frac{\partial u}{\partial s} = v - c - \delta h_b - (v + \delta(h_g - h_b))\Phi(z)$ and $\frac{\partial^2 u}{\partial s^2} = -(v + \delta(h_g - h_b))\sqrt{\tau}\phi(z) < 0$, i.e., for a fixed τ , the supplier's expected profit is concave in s . Suppose $v \leq c + \delta h_b$, contrary to the condition stated in the lemma. Then $\frac{\partial u}{\partial s} < 0$ and hence it is optimal to choose $s = 0$. Then $u(\tau, s)$ becomes $\tilde{u}(\tau) \equiv u(\tau, 0) = w - K(\tau) - (v + \kappa)/\tau$,

which is concave and approaches $-\infty$ as $\tau \rightarrow \bar{\tau}$. Its derivative evaluated at $\tau = \underline{\tau}$ is $\tilde{u}'(\underline{\tau}) = -K'(\underline{\tau}) + (v + \kappa)/\underline{\tau}^2 > (v + h_g - h_b)/\underline{\tau}^2 > 0$, where the first inequality comes from (8). Hence, $\tilde{u}(\tau)$ is maximized at an interior point $\hat{\tau} > \underline{\tau}$ that solves the equation $\tau^2 K'(\tau) = v + \kappa$. This solution should satisfy the backorder constraint $E[B|\hat{\tau}, 0] = 1/\hat{\tau} \leq \beta$. Since $1/\beta \leq \hat{\tau}$ and $\varphi(\tau) \equiv \tau^2 K'(\tau)$ is increasing, we have $(1/\beta)^2 K'(1/\beta) \leq \hat{\tau}^2 K'(\hat{\tau}) = v + \kappa$. Together with (9), this implies $\kappa + c + h_b < v + \kappa$ or $v > c + h_b > c + \delta h_b$, which contradicts our initial assumption. Therefore, the backorder constraint is inconsistent with $v \leq c + \delta h_b$. ■

Proof of Lemma 4. As shown in the proof of Lemma 3, the supplier's expected profit $u(\tau, s) = w - K(\tau) - \kappa/\tau - (c + \delta h_b)s - \delta(h_g - h_b)E[I|\tau, s] - vE[B|\tau, s]$ is concave in s and has the derivative $\frac{\partial u}{\partial s} = v - c - \delta h_b - (v + \delta(h_g - h_b))\Phi(z)$. With $v > c + \delta h_b$, the profit-maximizing s is found in the interior. The first-order condition $\frac{\partial u}{\partial s} = 0$ yields $z^* = \Phi^{-1}\left(1 - \frac{c + \delta h_g}{v + \delta(h_g - h_b)}\right)$, which is independent of τ . With $s^* = 1/\tau + z^*/\sqrt{\tau}$, the supplier's expected profit becomes $\tilde{u}(\tau) \equiv u(\tau, s^*) = w - K(\tau) - (\kappa + c + \delta h_b)/\tau - (v + \delta(h_g - h_b))\phi(z^*)/\sqrt{\tau}$. Note that $\lim_{\tau \rightarrow \bar{\tau}} \tilde{u}(\tau) = -\infty$. Differentiating $\tilde{u}(\tau)$,

$$\begin{aligned}\tilde{u}'(\tau) &= -K'(\tau) + \frac{\kappa + c + \delta h_b}{\tau^2} + \frac{(v + \delta(h_g - h_b))\phi(z^*)}{2\tau^{3/2}}, \\ \tilde{u}''(\tau) &= -K''(\tau) - \frac{2(\kappa + c + \delta h_b)}{\tau^3} - \frac{3(v + \delta(h_g - h_b))\phi(z^*)}{4\tau^{5/2}} < 0,\end{aligned}$$

which shows that $\tilde{u}(\tau)$ is concave. Evaluating $\tilde{u}'(\tau)$ at $\tau = \underline{\tau}$ and using (8) and the condition $v > c + \delta h_b$ in Lemma 3, we have

$$\tilde{u}'(\underline{\tau}) = -K'(\underline{\tau}) + \frac{\kappa + c + \delta h_b}{\underline{\tau}^2} + \frac{(v + \delta(h_g - h_b))\phi(z^*)}{2\underline{\tau}^{3/2}} > \frac{c + h_g - (1 - \delta)h_b}{\underline{\tau}^2} + \frac{(c + \delta h_g)\phi(z^*)}{2\underline{\tau}^{3/2}} > 0.$$

Since $\tilde{u}(\tau)$ is a concave function that initially increases at $\tau = \underline{\tau}$ and approaches $-\infty$ as $\tau \rightarrow \bar{\tau}$, we conclude that it is maximized at $\hat{\tau} > \underline{\tau}$ which is uniquely determined from the first-order condition $\tilde{u}'(\hat{\tau}) = 0$, as written in (15). To obtain sensitivity results, first note that

$$\frac{\partial z^*}{\partial v} = \frac{c + \delta h_g}{(v + \delta(h_g - h_b))^2} \frac{1}{\phi(z^*)} > 0.$$

Using this result and via implicit differentiation of the first-order condition $\tilde{u}'(\hat{\tau}) = 0$, we can show

that

$$\frac{\partial \hat{\tau}}{\partial v} = \frac{L(z^*)}{2\hat{\tau}^{3/2}} (-\tilde{u}''(\hat{\tau}))^{-1} > 0.$$

To show $\frac{\partial s^*}{\partial v} > 0$, we first derive two intermediate results. Observe

$$\begin{aligned} -\tilde{u}''(\hat{\tau}) &= K''(\hat{\tau}) + \frac{2(\kappa + c + \delta h_b)}{\hat{\tau}^3} + \frac{3(v + \delta(h_g - h_b))\phi(z^*)}{4\hat{\tau}^{5/2}} \\ &> K''(\hat{\tau}) + \frac{2(\kappa + c + \delta h_b)}{\hat{\tau}^3} + \frac{3(c + \delta h_g)}{4\hat{\tau}^{5/2}} z^* = K''(\hat{\tau}) + \frac{8\kappa + 5c + 8\delta h_b - 3\delta h_g}{4\hat{\tau}^3} + \frac{3(c + \delta h_g)}{4\hat{\tau}^2} s^* \\ &> \frac{8\kappa + 5(c + \delta h_g)}{4\hat{\tau}^3} + \frac{3(c + \delta h_g)}{4\hat{\tau}^2} s^* = \frac{4\kappa + c + \delta h_g}{2\hat{\tau}^3} + \frac{3(c + \delta h_g)}{4\hat{\tau}^2} \left(\frac{1}{\hat{\tau}} + s^* \right) > \frac{c + \delta h_g}{4\hat{\tau}^2} \left(\frac{1}{\hat{\tau}} + s^* \right), \end{aligned}$$

where we have used $L(z^*) = \phi(z^*) - z^*\bar{\Phi}(z^*) = \phi(z^*) - \frac{c + \delta h_g}{v + \delta(h_g - h_b)} z^* > 0$ in the first inequality and

(10) in the second. Also from Lemma A.2, we obtain

$$-L(z^*) + \left(\frac{c + \delta h_g}{v + \delta(h_g - h_b)} \right)^2 \frac{1}{\phi(z^*)} \geq 0$$

Using these two results, we can verify

$$\begin{aligned} \frac{\partial s^*}{\partial v} &= \frac{\partial}{\partial v} \left(\frac{1}{\hat{\tau}} + \frac{z^*}{\sqrt{\hat{\tau}}} \right) = -\frac{1}{2\hat{\tau}} \left(\frac{2}{\hat{\tau}} + \frac{z^*}{\sqrt{\hat{\tau}}} \right) \frac{\partial \hat{\tau}}{\partial v} + \frac{1}{\sqrt{\hat{\tau}}} \frac{\partial z^*}{\partial v} \\ &= -\frac{1}{2\hat{\tau}} \left(\frac{1}{\hat{\tau}} + s^* \right) \frac{L(z^*)}{2\hat{\tau}^{3/2}} (-\tilde{u}''(\hat{\tau}))^{-1} + \frac{c + \delta h_g}{(v + \delta(h_g - h_b))^2} \frac{1}{\sqrt{\hat{\tau}}\phi(z^*)} > 0. \end{aligned}$$

■

Proof of Proposition 3. We first show that the backorder constraint binds at the optimum.

First observe that, for $\tau^* = \hat{\tau} > \underline{\tau}$ and $s^* > 0$ found in Lemma 4,

$$\frac{\partial E[B | \hat{\tau}, s^*]}{\partial v} = \frac{\partial}{\partial v} \left(\frac{L(z^*)}{\sqrt{\hat{\tau}}} \right) = -\frac{L(z^*)}{2\hat{\tau}^{3/2}} \frac{\partial \hat{\tau}}{\partial v} - \frac{\bar{\Phi}(z^*)}{\sqrt{\hat{\tau}}} \frac{\partial z^*}{\partial v} < 0,$$

since $\frac{\partial \hat{\tau}}{\partial v} > 0$ and $\frac{\partial z^*}{\partial v} > 0$, as we showed in the proof of the lemma. Combined with $\lim_{v \rightarrow c + \delta h_b} E[B | \hat{\tau}, s^*] > \beta$ and $\lim_{v \rightarrow \infty} E[B | \hat{\tau}, s^*] = 0$, which are straightforward to show, $\frac{\partial}{\partial v} E[B | \hat{\tau}, s^*] < 0$ implies that the feasible region for the backorder constraint $E[B | \hat{\tau}, s^*] \leq \beta$ can be expressed as $v \geq v_{\min}$, where $v_{\min} > c + \delta h_b$ solves $E[B | \hat{\tau}, s^*] = L(z^*)/\sqrt{\hat{\tau}} = \beta$. Second, differentiating the customer's expected cost $C(\hat{\tau}, s^*) = \underline{u} + h_g N + K(\hat{\tau}) + (\kappa - h_g + h_b)/\hat{\tau} + (c + h_g)s^*$ and substituting the supplier's

optimal response $\hat{\tau}$ given by the first-order condition (15) yields

$$\begin{aligned}\frac{\partial C}{\partial v} &= \left(K'(\hat{\tau}) - \frac{\kappa - h_g + h_b}{\hat{\tau}^2} \right) \frac{\partial \hat{\tau}}{\partial v} + (c + h_g) \frac{\partial s^*}{\partial v} \\ &= \left(\frac{c + h_g - (1 - \delta) h_b}{\hat{\tau}^2} + \frac{v + \delta (h_g - h_b)}{2\hat{\tau}^{3/2}} \phi(z^*) \right) \frac{\partial \hat{\tau}}{\partial v} + (c + h_g) \frac{\partial s^*}{\partial v} > 0,\end{aligned}$$

since $\frac{\partial \hat{\tau}}{\partial v} > 0$ and $\frac{\partial s^*}{\partial v} > 0$. This monotonicity implies that the $C(\hat{\tau}, s^*)$ is minimized at the smallest feasible value of v , i.e., it is optimal to set $v^P = v_{\min}$, at which the backorder constraint binds ($L(z^*)/\sqrt{\hat{\tau}} = \beta$). The equilibrium values v^P and τ^P , determined from (16), are obtained by combining the optimality conditions (15), $L(z^*)/\sqrt{\hat{\tau}} = \beta$, and $z^* = \Phi^{-1} \left(1 - \frac{c + \delta h_g}{v + \delta (h_g - h_b)} \right)$. ■

Proof of Proposition 4. In all three cases (FB, CBC, and PBC) the backorder constraint binds in equilibrium, i.e., $L(z)/\sqrt{\tau} = \beta$. With this restriction, the customer's expected cost becomes $\tilde{C}(\tau) = \underline{u} + h_g N + K(\tau) + (\kappa + c + h_b)/\tau + (c + h_g)\zeta(\tau)/\sqrt{\tau}$, which is convex and minimized at τ^{FB} (see the proof of Proposition 2 for more details). Its derivative is $\tilde{C}'(\tau) = K'(\tau) - \Gamma(\tau)$. Substituting the optimality conditions for CBC and PBC in (B.10) and (16), it is easy to verify $\tilde{C}'(\tau^C) < \tilde{C}'(\tau^P) \leq \tilde{C}'(\tau^{FB}) = 0$, which implies $\tau^C < \tau^P \leq \tau^{FB}$. The next result follows immediately from the binding constraint, as $\tau^C < \tau^P \leq \tau^{FB}$ and $\beta = L(z^C)/\sqrt{\tau^C} = L(z^P)/\sqrt{\tau^P} = L(z^{FB})/\sqrt{\tau^{FB}}$ imply $z^C > z^P \geq z^{FB}$, and from $s^* = 1/\tau^* + z^*/\sqrt{\tau^*}$, we have $s^C > s^P \geq s^{FB}$. The next result $C^C > C^P \geq C^{FB}$ is implied by convexity of $\tilde{C}(\tau)$, $\tilde{C}'(\tau^{FB}) = 0$, and $\tau^C < \tau^P \leq \tau^{FB}$. Finally, with $\delta = 1$, the optimality conditions (B.10) and (16) become $\Gamma(\tau^C) - 0.5 (\tau^C)^{-3/2} pf(\zeta(\tau^C)) = K'(\tau^C)$ and $\Gamma(\tau^P) = K'(\tau^P)$, respectively, indicating the first-best is achieved under PBC but not under CBC. ■