9-2010

Contracting for Infrequent Restoration and Recovery of Mission-Critical Systems

Sang-Hyun Kim

Morris A. Cohen
University of Pennsylvania

Serguei Netessine

Recommended Citation

Kim, S., Cohen, M. A., & Netessine, S. (2010). Contracting for Infrequent Restoration and Recovery of Mission-Critical Systems. *Management Sciecne, 56* (9), 1551-1567. http://dx.doi.org/10.1287/mnsc.1100.1193

Contracting for Infrequent Restoration and Recovery of Mission-Critical Systems

Abstract
Firms that rely on functioning mission-critical equipment for their businesses cannot afford significant operational downtime due to system disruptions. To minimize the impact of disruptions, a proper incentive mechanism has to be in place so that the suppliers provide prompt restoration and recovery services to the customer. A widely adopted incentive mechanism is performance-based contracting (PBC), in which suppliers receive compensation based on realized system uptime. A key obstacle is that disruptions occur infrequently, making it very expensive for a supplier to commit the necessary resources for recovery because they will be idle most of the time. In this paper, we show that designing a successful PBC creates nontrivial challenges that are unique to this environment. Namely, because of the infrequent and random nature of disruptions, a seemingly innocuous choice of performance measures used in contracts may create unexpected incentives, resulting in counterintuitive optimal behavior. We compare the efficiencies of two widely used contracts, one based on sample-average downtime and the other based on cumulative downtime, and identify the supplier's ability to influence the frequency of disruptions as an important factor in determining which contract performs better. We also show that implementing PBC may create high agency cost when equipment is *very reliable*. This counterintuitive situation arises because the realized downtimes from which the customer might intuit about the supplier's capacity investment are highly uncertain when there are not many samples of downtimes, i.e., when disruptions occur rarely.

Keywords
service outsourcing, supply chain, after-sales support, maintenance–repairs, disaster recovery

Disciplines
Operations and Supply Chain Management

Contracting for Infrequent Restoration and Recovery of Mission-Critical Systems

Sang-Hyun Kim

Yale School of Management, Yale University, New Haven, CT 06520

sang.kim@yale.edu

Morris A. Cohen • Serguei Netessine • Senthil Veeraraghavan

The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

cohen@wharton.upenn.edu • netessine@wharton.upenn.edu • senthilv@wharton.upenn.edu

Abstract

Firms that rely on functioning mission-critical equipment for their businesses cannot afford significant operational downtime due to system disruptions. To minimize the impact of disruptions, a proper incentive mechanism has to be in place so that the suppliers provide prompt restoration and recovery services to the customer. A widely adopted incentive mechanism is performance-based contracting (PBC), in which suppliers receive compensation based on realized system uptime. A key obstacle is that disruptions occur infrequently, making it very expensive for a supplier to commit the necessary resources for recovery since they will be idle most of the time. In this paper we show that designing a successful PBC creates nontrivial challenges that are unique to this environment. Namely, due to the infrequent and random nature of disruptions, a seemingly innocuous choice of performance measures used in contracts may create unexpected incentives, resulting in counterintuitive optimal behavior. We compare the efficiencies of two widely-used contracts, one based on sample-average downtime and the other based on cumulative downtime, and identify the supplier's ability to influence the frequency of disruptions as an important factor in determining which contract performs better. We also show that implementing PBC may create high agency cost when equipment is *very reliable*. This counterintuitive situation arises since the realized downtimes from which the customer might intuit about the supplier's capacity investment is highly uncertain when there are not many samples of downtimes, i.e., when disruptions occur rarely.

1 Introduction

Whether they are caused by an earthquake, a hazardous oil spill, a simple power failure or a random machine breakdown, unexpected disruptions of mission-critical operations can lead to dramatic consequences. In some cases, such disruptions may cost firms millions of dollars even if they last only a few hours or even minutes (see Sheffi 2007 for numerous examples). While firms put much effort in preventing such events from happening, perfect prevention is often impossible or economically infeasible to achieve, especially when the firm is not in control of the source of the disruption (as is the case, for example, of natural disasters). Therefore, contingency planning is essential; however severe an initial disruption may be, its impact can be significantly reduced if an affected system is quickly restored to its normal operating condition through carefully thought-out recovery action plans and prior deployment of resources.

Not surprisingly, disaster recovery/restoration services constitute a significant portion of many industries (see Disaster Recovery Journal, www.drj.com, for numerous examples). For instance, firms such as Sungard, HP, and IBM, offer recovery and business continuity services for IT equipment, where the market is estimated to be $4.2B in 2006 and growing at about 7% per year (Frauenheim 2003). The list of events covered by the HP business recovery services is long and includes fires, accidents, sabotage, chemical spills and power anomalies, to name just a few.[1] As another example, Clean Harbors Inc. manages environmental emergency responses or disaster recovery on land and water, from cleanup and removal of a single mercury bottle to a large-scale multiphase containment and cleanup of a coastal oil spill. The company booked more than $1B in revenues in 2008.[2] Another example is large maintenance and repair services that exist in industries spanning semiconductor manufacturing, aerospace, defense, medical equipment, and others, where equipment is complex and the consequences of breakdowns are severe. In the aerospace and defense industry alone, the revenues generated from these services are in excess of $100B (Wall Street Journal 2009).

While the above examples are drawn from very distinct industries, in this paper we focus on at least three similarities that unite these examples. First, events leading to operational disrup-

[1] See http://www.hp.com/sbso/services/recoverall.html.
[2] Source: annual report.

tions are random and infrequent. Indeed, earthquakes and oil spills occur very rarely, and even a complex equipment such as an airplane engine does not fail or require maintenance very often. Second, resources needed to restore the affected system quickly cannot be procured on the spot and therefore have to be deployed far in advance. In most cases described above, the restoration process requires sophisticated machinery, spare parts, and extensive personnel training. Third, and most importantly, restoration and recovery services are typically outsourced. For example, Clean Harbors Inc. is subcontracted by the Environmental Protection Agency (EPA) to perform cleanup of hazardous materials; Sungard, HP, and IBM lend their expertise when computing equipment goes down; Boeing offers maintenance and repairs of its aircraft to its airline customers.

When a supply chain is decentralized as in these examples, implementing a contingency plan requires coordination among distinct organizations. The key question we want to answer is: how can proper incentives be structured in such a situation? For example, Intel requires its suppliers to respond within 15 minutes to a failure of its semiconductor manufacturing equipment (Harrington 2006). How can Intel make sure that a promise to restore its equipment quickly is fulfilled by the supplier? After all, it is Intel, not the supplier, who bears the direct consequences of the failure and has more urgency; on the other hand, it is challenging and costly for the supplier to dedicate resources for fast diagnosis and repair at every customer location because the supplier has limited capacity and because equipment failures occur infrequently.

In investigating incentives for restoration/recovery services, we focus on performance-based contracting (PBC) which is gaining wide acceptance as an effective instrument for providing such incentives across commercial and government supply chains. Under PBC, compensation to a supplier is based on realized service outcomes such as equipment uptime or response time which are directly related to the value created by the customer through the operation of the system. For example, many service level agreements that Internet service providers offer stipulate financial remedies for a failure to deliver promised network uptime, which depends on how fast a problem can be resolved after it is detected and how frequently such problems arise (Stanbury 2004). PBC is even more pervasive at federal government agencies because of a major initiative led by the White House that has been in effect in recent years (see Government Accountability Office 2002 and Office of Management and Budget 2003). There, PBC is required to be used to the maximum extent possible for the procurement of services. Based on this requirement, EPA has issued

guidance (EPA 2003) regarding outsourcing of cleanup activities that states: "by linking payment to performance measures, PBCs offer potential advantages to the government. In the interests of minimizing costs while expediting the reduction of risks to public health and the environment, EPA is committed to working with the lead federal agencies in applying a performance basis in cleanup activities." As another example, the U.S. Department of Defense (DoD) has initiated a policy called Performance-Based Logistics that mandates all of its outsourced logistics and services, including restoration and recovery services, be performance-based (DoD 2003). The following excerpt from a U.S. Army equipment service contract (Army Material Command 2006) illustrates how a typical PBC contract is set up: "If a complete critical system remains inoperative and cannot perform the scheduled workload due to a product malfunction (system downtime) through no fault or negligence of the Government for a period of 24 contiguous hours, the Contractor shall grant a credit to the Government for each half-hour of downtime."

Although there is some evidence suggesting that PBC successfully incentivizes suppliers to meet required performance goals (Geary 2006), there is a flip side: there are significant financial risks that accompany such contracts. Since service outcomes are inherently random, fluctuations in the suppliers' contractual income streams that depend on such outcomes are inevitable. Indeed, one of the principal motivations for the customer to adopt PBC is to transfer the risk of output uncertainty to the supplier in the form of contract payment uncertainty. To many suppliers, this uncertainty is a source of great concern, as they a prefer predictable cash flow. For example, it is possible under PBC for a supplier to make a negative profit if an unforeseen problem such as spare parts shortage contributes to a long delay in completing a required restoration service. Such aversion to revenue risk makes implementing PBC inefficient, because suppliers demand a risk premium as a condition for entering into a PBC arrangement. In the airline industry, for instance, high risk premiums demanded by maintenance service providers have become an important sticking point (Sobie 2007, Oliver Wyman 2007).

In this paper, we use a principal-agent contracting framework to construct a stylized model of an outsourcing environment in which operational disruptions occur infrequently and where the customer cannot write the supplier's capacity investment decision directly into the contract but instead must incentivize the supplier using PBC. The customer in our model has the goal of maximizing profit while limiting system downtime following a disruption event to a target level. The

main theme we explore in this paper is the interaction between an external condition (infrequent and random system disruptions) and the ensuing internal uncertainty (random service completion times realized after disruptions), as well as their combined effect on the efficiency of PBC.

From our interactions with practitioners in the aerospace, defense, and high-tech manufacturing industries, we have found that the majority of the performance-based contracts used for restoration/recovery services fall into two categories, depending on how performance is defined: a contract based on cumulative downtimes (hereafter, CC) and a contract based on the sample-average of downtime (hereafter, AC). Under the former, the supplier is penalized for the total system downtime within a contract period, whereas under the latter the supplier is penalized for the total downtime divided by the number of disruption occurrences. Both are designed to achieve the same goal, namely, to incentivize the supplier to reserve a high level of service capacity so that restoration can be completed quickly upon system failure.

We show that there are unique challenges in the environments described above. Because disruptions are infrequent, any signal about the supplier's capacity investment that the customer might intuit from the supplier's delivery of service (which is the basis of contract payments) is likely to be highly uncertain. The following are the specific major insights from our analysis:

1. We show that contracts based on two seemingly similar performance measures described above, CC and AC, create completely different incentive structures that may induce very different responses from the supplier in terms of his capacity decision. For example, one would expect that, for a given downtime penalty, the supplier will build more capacity when the system is more prone to failure. This is indeed the case for CC, but under AC the supplier's optimal capacity investment is non-monotone in system failure rate.

2. When the supplier cannot affect the frequency of disruptive events (e.g., when system failures are due to natural disasters), we find that AC generally brings higher efficiency than CC does. In contrast, when the supplier can reduce frequency of disruptions (e.g., by investing into equipment reliability), CC is often preferred because its effectiveness in incentivizing the supplier to improve reliability is greater than that of AC.

3. When meeting a downtime target is the pressing goal for the customer (which happens whenever disruptions are sufficiently rare), we show that the efficiency of PBC is worst when

disruptions almost never occur, e.g., when equipment is *most reliable*. This situation arises because information about the supplier's performance is severely limited when there are few opportunities to perform, i.e., when the system almost never fails. Therefore, high equipment reliability does not always mean good news – when fast restoration is as important as high reliability, infrequent failures may make it challenging to implement PBC successfully.

The rest of the paper is organized as follows. After a brief survey of related literature, we lay out modeling assumptions in Section 3 and proceed to model analysis in Sections 4 through 6. In Section 7 we discuss consequences of relaxing some of the assumptions in the base model, including a scenario in which the supplier can determine frequency of disruptions as well as service capacity. Finally, in Section 8, we summarize the major findings and discuss future directions for research.

2 Literature Review

Our model applies the principal-agent analysis framework to a service outsourcing environment. The most closely related service operations literature concerned with modeling delays and waiting times has traditionally focused on queuing problems, i.e., settings in which server utilization is relatively high. There, economic decisions are made on how to manage congestion, either by adding more servers or by changing the rate of service. Examples of articles that consider principal-agent relationships with significant queuing effects include Gilbert and Weng (1998), Plambeck and Zenios (2003), Ren and Zhou (2008), Hasija et al. (2008), and Lu et al. (2009). Some of these and related papers model system behaviors in the heavy-traffic regime, i.e., when server utilization approaches one. In contrast, our model considers the opposite end of the spectrum, namely, where utilization is close to zero; the service capacity in our model is generally idle except when responding to infrequent disruptions, and hence queuing for service is not an issue. The "demand" in our problem context means infrequent but high-impact disruptions that incur large opportunity costs. As a result, a high level of service capacity must be maintained to reduce the impact of such disruptions.

Service parts inventory management is an area in which rare equipment failures drive managerial decisions, just as in our model. Sherbrooke (1992), Muckstadt (2005), and Cohen et al. (1990) give an overview of theory and applications of this stream of research. For the most part, this literature does not consider contracting or incentives. Exceptions occur in Kim et al. (2007, 2009), who study

contracting issues in after-sales product support outsourcing. The current paper is very distinct from these works since, here, we highlight the challenges associated with PBC that arise due to the infrequent nature of equipment failures and the role of performance metric specifications, a topic that has not been explored before. Furthermore, in the current paper we focus on the provision of restoration services rather than on inventory management.

Our model is based on the "moral hazard" principal-agent framework, relying on the assumption that the parties contract on a commonly observable performance measure which is a noisy indicator of the agent's action, as is the case with PBC. However, there are two major features of our model that are quite distinct from what can be found in the traditional contracting models such as the ones presented in Laffont and Tirole (1993), which provides a comprehensive overview of procurement contracting theory. Namely, we model an environment in which performance outcomes are *intermittent* and *randomly* realized. There are very few papers in economics that consider a low-frequency environment. An exception is Abreu et al. (1991), in which the role of the review period length in a repeated partnership game is investigated. Although our model affirms some of the insights that they have found, we derive a richer set of findings under the assumptions that reflect real-world service outsourcing practices. In particular, many of the results in this paper are driven by the complex interaction between an endogenous uncertainty (random service times) and an exogenous uncertainty (random system disruptions), an operational detail that creates significant and interesting dynamics. Our model also adds a layer of complexity to the classical principal-agent model as it allows for a situation in which an agent's performance outcome may not materialize (i.e., if the system does not fail, there is no opportunity to service it). This creates nontrivial contracting issues that, to the best of our knowledge, have not received attention in the literature.

Finally, our paper is related to the literature on supply chain disruption management. We refer to Kleindorfer and Saad (2005) and Sheffi (2005) for a general review, and Tomlin (2006) for a recent article in this area of research. While much of the literature focuses on preventing disruptions, contingency planning in a decentralized supply chain is also recognized as an important aspect of risk management. Our paper contributes to the latter stream. To summarize, we believe that our paper is the first to address the issue of outsourcing restoration/recovery services to mitigate low-probability, high-impact disruptions.

3 Model Assumptions

In the remainder of the paper, we use the term "equipment failure" in place of "system disruption" because the former frames the problem in a clearer context. A risk-neutral customer ("she") derives utility from continued usage of equipment, which is subject to random breakdowns. When a breakdown occurs, the equipment needs to be restored to working condition as quickly as possible. As the customer lacks technical expertise, she delegates the control of all restoration activities (which we call a *service*) including diagnostics, parts replacement, and repairs and testing, to a single supplier ("he"), who is risk-averse. Such one-to-one relationships are commonly observed in practice, especially when a government organization such as the U.S. Army outsources maintenance service of customized equipment to a contractor. The customer and the supplier establish a contractual relationship for the duration of one time period (e.g., a year) whose length is normalized to one. In practice the length of the contracting period is typically tied to the annual budgeting process and cannot be easily extended. In the beginning of the period the customer offers a contract to the supplier. In response, the supplier decides how much he should invest in service capacity. The supplier's investment is unobservable and non-contractible. Over the length of the contract period, random equipment failures occur, triggering service activities by the supplier.[3] At the end of the period, the customer assesses the supplier's performance based on service completion times, and payments are made according to the agreed contract terms.

3.1 Equipment Failure Process and Capacity Decision

Equipment failures occur according to a Poisson process with a rate λ, which is assumed to be common knowledge. Let N be the random variable representing the number of equipment failures within the period. In this paper we consider low λ values, i.e., λ near zero up to a single digit (recall that this scale is in relative to the contract length which is normalized to one). Aerospace and defense contractors and telecommunications companies routinely observe such low failure rates for mission-critical equipment, while outages due to natural disasters are even rarer. In the main part of the paper we assume that λ is exogenously determined, as is the case when the failures

[3] Since we are concerned with restoration/recovery services, *force majeure* does not apply to the contracts considered in our model, as an unforseeable event (equipment failure) triggers the supplier's action, not disrupts it.

occur due to events like natural disasters. We relax this assumption later in Section 7.2.

Each incident of equipment failure triggers a service process performed by the supplier which may include traveling to the customer's site, diagnosing the failure, shipping the necessary parts and/or repairing failed parts, installing replacement parts, and testing the equipment. Let S_i be the service completion time or, equivalently, the equipment downtime for the i^{th} failure. We refer to it throughout the paper as either the *service time* or the *downtime*. We assume that $\{S_i\}$ are i.i.d. with a rate $1/\mu$, where μ is the service *capacity* set by the supplier at the beginning of the period. We assume that μ remains unchanged throughout the period. Although situations arise in which the capacity level can be dynamically altered, we focus on the cases in which it is too costly or impractical to do so. Repair facility purchase, employee training, and process re-design are examples of capacity decisions which require large up-front investments and cannot easily be adjusted, because either the commitment cannot be reversed or the impact of the decision is only realized after a long time. We use the superscript $*$ to denote the supplier's optimal response (i.e., capacity choice) to a contract. Since $\{S_i\}$ are i.i.d., we drop the subscript i unless it is needed for clarity. We use the convention that μ is bounded below by $\underline{\mu}$, which we interpret as the default level of capacity that the supplier already possesses and hence can provide with zero investment. We assume that $1/\underline{\mu} \ll 1/\lambda$, i.e., the maximum expected service completion time is much shorter than the mean time between equipment failures. For example, while a network server may go down once or twice a year, repair standards as low as several hours are common (Cohen et al. 2006). This assumption on the scale difference not only reflects reality but also simplifies our analysis, since it allows us to approximate λ to be constant even though equipment failure interarrival times depend on how fast services are completed. This assumption is consistent with what is found the service parts management literature (see Muckstadt 2005).[4]

Additionally, we assume that the expected service time resulting from the default capacity level $\underline{\mu}$ is unacceptably long to the customer. As a result, the customer wants the supplier to expand capacity beyond $\underline{\mu}$, which requires extra investment for the supplier. For simplicity, we assume that this cost is linear in μ with a unit cost c such that the total investment is equal to $c(\mu - \underline{\mu})$. Other

[4] In addition, the same assumption makes it unlikely to encounter a situation in which an initiated service is not completed by the end of the period, since the probability that a failure occurs within a time interval of order $1/\mu$ before the end of the period is negligible.

papers on service operations frequently assume linear service capacity cost (for example, see Allon and Federgruen 2007). While costly, increased capacity reduces expected service time.

We further assume that the variability of service time, defined as the coefficient of variation $v(\mu)$ of S, possesses the following three properties: it does not increase when capacity increases ($v'(\mu) \leq 0$), there is a non-increasing rate of variability reduction ($v''(\mu) \geq 0$), and $|v'(\underline{\mu})|$ is bounded from above. This assumption generalizes the typical construct found in the queuing literature in which the service time distribution is often assumed to be exponential, which fixes the coefficient of variation to a constant. By imposing both $dE[S \mid \mu]/d\mu < 0$ and $v'(\mu) \leq 0$, we are able to clearly distinguish between a "good" state (i.e., fast service time and low variability) when μ is high and a "bad" state (i.e., slow service time and high variability) when μ is low, a distinction that is made in most principal-agent models (this is analogous to imposing the monotone likelihood ratio property; see Milgrom 1981). Without this assumption, it is no longer clear if increasing capacity is beneficial to the supplier and ultimately to the customer, needlessly complicating the main insights we obtain in this paper. While we acknowledge that there may be situations where $v(\mu)$ increases (such as when capacity increase is associated with adopting new, untested technology), the opposite is more commonly found in practice as a result of factors such as economies of scale and learning, which make service times more predictable. In the following analysis, we will frequently revisit the special case in which $v(\mu)$ is constant since it allows for tractable analysis.

3.2 Contracting

We assume that the customer cannot directly observe the supplier's capacity choice μ and therefore cannot directly contract on it. This is a reasonable assumption since most suppliers exert a multitude of discretionary efforts (such as decisions on spare parts inventory investment, repair depot staffing, training, transportation methods, etc.) that are too difficult or costly to monitor. Because she cannot contract on μ, the customer enforces her service requirement via a performance-based contract such that the compensation T to the supplier is tied to an agreed-upon performance measure (e.g., equipment downtime), generically denoted by X. In this paper, we analyze linear contracts that have the payment form $T = w - pX$, in which w is the fixed payment independent of the realized performance X, and p is the penalty rate for each unit of X. This contract form is motivated by a convention observed in many industries where a fixed pool of money (w) is reserved

for the supplier that only gets subtracted in proportion (constant penalty rate p) to the realization of performance (X). While linear contracts are known to be suboptimal compared to nonlinear, state-contingent contracts, they are easier to implement and hence widely adopted in practice. For this reason, we focus only on linear contracts in this paper.

We consider two types of linear contracts that differ by the definition of performance measure that enter into the contracts. The first is based on cumulative downtime $\sum_{i=1}^{N} S_i$ and is referred to as a CC (cumulative-performance contract). The second is based on average downtime $(\sum_{i=1}^{N} S_i)/N$ and is referred to as an AC (average-performance contract). A precise expression of the performance measure X used in each contract will be introduced in Section 5. The majority of contract terms that are encountered in practice fall into these two categories. Contracts such as the U.S. Army example in the Introduction are of the CC type, but AC-type contracts are also observed in practice. For example, a service level agreement by a voice/data service provider (whose name is not revealed due to confidentiality) defines the target IP service restoration time on which financial remedy is based as "an *average* service restoration interval of 4.0 hours for each circuit measured on a per circuit." However, there is little understanding as to which type of contract should be used under which circumstance. A main theme of this paper is to compare the consequences of implementing these two contracts.

It should be noted that "average" in AC refers to the sample-average of $\{S_i\}$, as opposed to the time-average, which in fact applies to CC. This last statement follows from the fact that we have normalized the length of the contracting period to one: the time-average of total downtime is equal to cumulative downtime since the former is obtained from dividing the latter by one. In this sense, a comparison between CC and AC can also be viewed as a comparison between two different methods of averaging the supplier's performance. As we will find out in Section 5, this seemingly innocuous choice for evaluating the supplier's performance may lead to surprisingly different results.

The risk-averse supplier is assumed to have a mean-variance utility function that depends on stochastic compensation T that he receives from the customer:

$$u(\mu) = E[T \mid \lambda, \mu] - \eta \mathrm{Var}[T \mid \lambda, \mu] - c(\mu - \underline{\mu}). \tag{1}$$

The parameter η is the coefficient of risk aversion. A larger η corresponds to higher risk aversion

and $\eta = 0$ represents risk neutrality. We employ the risk aversion assumption to capture the supplier's desire to avoid revenue risks posed by PBC. The inefficiency arising from such aversion to risk is expressed in the second term of (1) and is called the *risk premium* (see Gollier 2001, p. 20), denoted by $\psi \equiv \eta \text{Var}[T \mid \lambda, \mu]$. Our close work with supplier organizations in various industries reveals that revenue risk is an issue that most concerns them about PBC. At the same time, they express the need to quantify the risk premium for contract negotiation purposes. The assumption of risk aversion reflects such concerns. The mean-variance function captures the basic expected revenue vs. revenue risk tradeoff for the supplier, and is widely adopted in the recent operations management literature (for example, see Tomlin 2006, Van Mieghem 2007, and Kim et al. 2007). On the other hand, we assume in the main part of the paper that the customer is risk neutral on the grounds that she represents a larger enterprise, such as the DoD, that is less sensitive to cash flow risk. We relax this assumption in Section 7.3, where we investigate an alternative situation where the customer is more averse to financial risk.

After being offered contract terms, the supplier chooses capacity μ^* to maximize his utility (1). Anticipating the supplier's choice of capacity, the customer decides on contract terms that (a) induce the supplier to voluntarily choose μ^* that satisfies the customer's objective (defined in the next subesection) and (b) ensure that the supplier participates in the contractual relationship. The second requirement is expressed in the individual rationality (IR) constraint $u(\mu^*) \geq \underline{u}$, where $\underline{u}$ denotes the supplier's reservation utility, i.e., the level of utility that he obtains if he opts out of the contract. We normalize $\underline{u}$ to zero throughout this paper, because doing so changes no qualitative insights. As is typical in most principal-agent models, the IR constraint will turn out to be binding in all cases we analyze. As a consequence, the customer's maximized expected profit will be equal to the supply chain's maximum profit, since the supplier is left with zero utility. Both profits will be denoted as Π.

3.3 The Customer's Objective

We assume that the customer earns r per unit time while equipment is functioning. As the total revenue is proportional to the equipment uptime $1 - \sum_{i=1}^{N} S_i$, the customer's expected profit is $\Pi = r(1 - E[\sum_{i=1}^{N} S_i \mid \lambda, \mu^*]) - E[T \mid \lambda, \mu^*]$, where T denotes payment to the supplier. Note that expected uptime $1 - E[\sum_{i=1}^{N} S_i \mid \lambda, \mu^*]$ is equivalent to expected *availability*, the fraction of equipment uptime

relative to the contract length, since the latter is normalized to one. Therefore, the customer maximizes her profit by maximizing the expected availability.

However, availability maximization criterion alone does not capture the important concern of the practitioners that uniquely exists in the environment that we consider, namely, when the failure frequency λ is small. To illustrate this point, suppose that equipment almost never fails within a contracting horizon, i.e., $\lambda \approx 0$. Such a scenario is quite plausible if equipment failure occurs due to very rare events such as earthquakes. Then, the expected equipment availability is guaranteed to be near 100% as restoration/recovery service is unlikely to be requested. As a result, it is optimal not to motivate the supplier to increase capacity beyond the default level $\underline{\mu}$ since doing so only adds to the cost (i.e., payment to the supplier) with little increase in revenue. But what if equipment does fail? After all, there is always a positive probability, albeit small, that it will. In such a case the customer will experience an unacceptably long outage, handicapped by the low service capacity.

Most service-providing organizations try to avoid this undesirable situation by focusing not only on availability, which is an aggregate measure directly tied to their profitability, but also on individual service experience. Namely, they set a standard on the delivery of *each* service instance. The most common way is to set the standard is to impose a maximum on the expected service time, such that:

$$E[S_i \,|\, \mu^*] = 1/\mu^* \leq s_I. \tag{STC}$$

The *service time target* s_I is assumed to satisfy $s_I < 1/\underline{\mu}$, consistent with our earlier assumption that the default capacity $\underline{\mu}$ is inadequate for the customer. (The subscript I denotes "individual" service instance.) STC stands for *service time constraint*. Real-world examples of this constraint, e.g., terms like "target turnaround time of 12 hours", are commonly found in service contracts of companies like Sungard, HP, and others. In general, the target s_I may be a function of λ. For analytical tractability, however, we will focus on a special case where s_I is independent of λ in the main part of the paper (we relax this assumption in Section 7.1).[5] Letting $\mu_I \equiv 1/s_I$, we can rewrite STC as the lower bound constraint on capacity chosen by the supplier: $\mu^* \geq \mu_I$. Combined

[5] We believe that the assumption that the service time target s_I is independent of λ is, at least in a limited range of λ, reasonable in many practical settings. This is based on the obeservation that many service contracts stipulate a rather arbitrary but a convenient value for the target, such as "24 hours" or "3 days", regardless of the estimated failure frequency.

with the assumptions above, the target capacity μ_I satisfies $\mu_I > \underline{\mu} \gg \lambda$. The service time target s_I, or equivalently the capacity target μ_I, reflects the customer's willingness to accept performance risk (as opposed to financial risk, which we discussed in Section 3.2): the lower the s_I, the more the customer wishes to avoid downtime and receive a faster service. We note that imposing a service time constraint is ubiquitous not only in the practice of restoration/recovery service outsourcing but also in other service settings, such as call center operations (see, for example, Gans et al. 2003, Gurvich et al. 2005, and Milner and Olsen 2006, where a constraint on "average speed of answer", or ASA, is defined similarly as STC).

In sum, motivated by real-world practices, we assume that the customer's objective is to maximize the expected profit $r(1 - E[\sum_{i=1}^{N} S_i \,|\, \lambda, \mu^*]) - E[T \,|\, \lambda, \mu^*]$ subject to STC.[6] The constraint may or may not bind depending on the parameter values. As will become clear below, the solution behavior critically depends on whether STC binds at optimum.

4 First-Best

No efficiency is lost if the customer can contract directly on the supplier's capacity choice μ. Under such a complete observability assumption, a fixed payment contract $T = w$ that guarantees the supplier's participation achieves the first-best. The solution is straightforward. If $\lambda > c\mu_I^2/r$, STC does not bind at optimum. Then the customer imposes $\mu^{FB} = \sqrt{r\lambda/c}$ and offers a fixed-payment contract $w^{FB} = c(\sqrt{r\lambda/c} - \underline{\mu})$. If, on the other hand,

$$\lambda \leq c\mu_I^2/r, \tag{2}$$

[6] An alternative modeling choice is to drop STC and define the customer's expected utility as

$$U = r(1 - E[\textstyle\sum_{i=1}^{N} S_i \,|\, \lambda, \mu^*]) - E[T \,|\, \lambda, \mu^*] - G(\mu^*),$$

where $G(\mu^*)$ denotes the customer's disutility caused by a long service time after a failure incident when the supplier chooses μ^*. $G(\mu^*)$ is significantly large when μ^* is close to $\underline{\mu}$, but it decreases and converges to zero as μ^* increases. This extra term plays a role similar to STC, i.e., it ensures that the supplier chooses large enough capacity. (In fact, they are equivalent if we set $G(\mu) = \phi(\mu_I - \mu)$, where ϕ is the shadow price associated with STC.) As a practical matter, however, managers at customer organizations have little idea on how to estimate the functional form of $G(\mu^*)$, whereas performance requriements such as STC are routinely found in service contracts. In this paper we take a descriptive approach and employ the service time constraint, reflecting how practitioners view their contracting problem.

the STC binds at optimum and the customer sets $\mu^{FB} = \mu_I$ and $w^{FB} = c(\mu_I - \underline{\mu})$. In both cases, the resulting expected profit for the customer is equal to $\Pi^{FB} = r(1-\lambda/\mu^{FB}) - c(\mu^{FB} - \underline{\mu})$, leaving zero surplus for the supplier.

As is apparent from the expression in (2), STC binds at optimum whenever the equipment failures are relatively rare (small λ) compared to a fixed revenue-to-cost ratio r/c and a fixed service time target $s_I = 1/\mu_I$. Therefore, the condition (2) formalizes our assertion in Section 3.3: when the chance of encountering a failure is sufficiently small, the customer's desire to ensure an acceptable outcome of the service for that rare event takes precedence over maximizing profitability. We believe that the instances that satisfy (2) are routinely observed in many real-world situations.[7] Notice that the optimal capacity μ^{FB} is independent of λ if (2) is satisfied but increases in λ otherwise (the first being a direct consequence of our assumption that s_I is constant; we study a more general case in Section 7.1). This observation suggests that, depending on how frequently equipment fails, the customer may want to offer a qualitatively different set of incentives to the supplier when the former outsources the restoration/recovery services. We investigate this further in the next section.

5 Supplier's Capacity Decision

Having analyzed the first-best benchmark case, we now return to the general model with non-contractible μ. In this section, we characterize the supplier's capacity decision after he is offered contract terms from the customer. We present derivations of optimal capacities for CC and AC in separate subsections (Sections 5.1 and 5.2) as they require substantially distinct analyses. In Section 6 we study the optimal contract design problem of the customer, who takes into account the supplier's capacity decision analyzed in this section.

[7] As an example to support this, let us assume that contract duration is 100 days and the customer enforces the service time target of 12 hours, a 40% improvement over the default average of 20 hours. These values correspond to $\mu_I = 200$ and $\underline{\mu} = 120$. Suppose, conservatively, that the revenue rate is so high that daily revenue $0.01r$ is comparable to total capacity investment $c(\mu_I - \underline{\mu}) = 80c$, so let $0.01r = 80c$. Then the condition (2) is satisfied whenever the equipment fails on average 5 times or less, a very likely scenario in practice.

5.1 Optimal Capacity Under Cumulative-Performance Contract

We first consider CC, under which the supplier performance is evaluated based on the cumulative downtime $\sum_{i=1}^{N} S_i$. Since the number of arrivals N is Poisson-distributed, $\sum_{i=1}^{N} S_i$ is a compound Poisson random variable. The supplier, who is given a contract $T = w - p\sum_{i=1}^{N} S_i$ under CC, determines his optimal capacity as follows.

Lemma 1 *The supplier's utility (1) is concave under CC. Define*

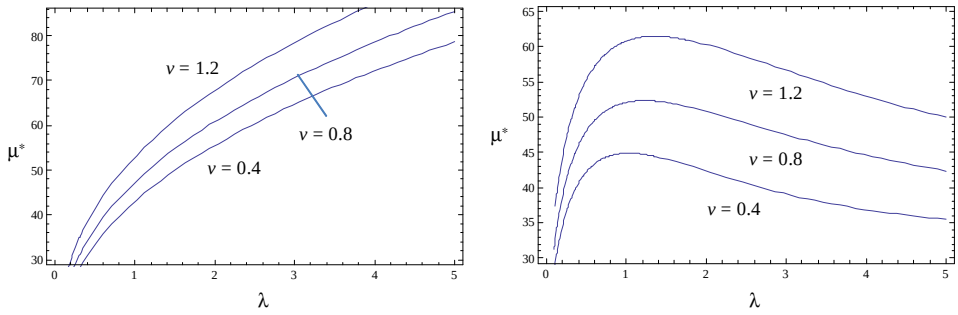
$$\theta(\mu) \equiv v(\mu)^2 - \mu v(\mu) v'(\mu) \geq 0. \tag{3}$$

The supplier chooses $\mu^* > \underline{\mu}$ that satisfies the optimality condition $p\mu^{-2} + 2\eta p^2 \left(1 + \theta(\mu)\right) \mu^{-3} = c/\lambda$, provided that p is sufficiently large to admit an interior solution. Moreover, $\partial\mu^*/\partial c < 0$, $\partial\mu^*/\partial p > 0$, $\partial\mu^*/\partial\eta > 0$, and $\partial\mu^*/\partial\lambda > 0$.

The quantity $\theta(\mu)$ in (3) is the normalized rate at which the supplier's revenue risk is reduced (i.e., changes in $\mathrm{Var}[T \,|\, \lambda, \mu]$) as a result of increasing μ by one unit. Under the assumptions made regarding the shape of $v(\mu)$, it can be easily verified that $\theta'(\mu) \leq 0$, which implies that it becomes more difficult to reduce the revenue risk as capacity grows larger (i.e., decreasing marginal scale).

We can explain the supplier's optimal capacity choice μ^* in response to variations in parameters c, p, and η as follows. The supplier's incentive to invest in capacity is higher when (a) the unit capacity cost is lower ($\partial\mu^*/\partial c < 0$), (b) the performance incentive is higher ($\partial\mu^*/\partial p > 0$), or (c) the supplier is more risk-averse ($\partial\mu^*/\partial\eta > 0$). The first two results are intuitive. The third result arises from the fact that increased capacity leads to reduced service time uncertainty, as $\mathrm{Var}[S \,|\, \mu] = v(\mu)^2 (E[S \,|\, \mu])^2$ decreases in μ. In other words, a risk-averse supplier can hedge against revenue risk by increasing capacity μ. This effect becomes more pronounced the more risk-averse the supplier is.

The key result from Lemma 1 is that μ^* is an increasing function of λ, as shown in Figure 1(a). There are two reasons. First, given a fixed penalty rate p, more equipment failures imply a greater expected total penalty for the supplier under CC, as the contract stipulates that he loses money for each minute the equipment is down. To avoid such losses, the supplier increases capacity μ. Second, since the revenue risk increases with higher λ (since $\mathrm{Var}[\sum_{i=1}^{N} S_i \,|\, \lambda, \mu]$ is proportional



(a) μ^* under *CC* for fixed *p* (b) μ^* under *AC* for fixed *p*

Figure 1: Example showing how μ^* varies as a function of λ under CC and AC. The coefficient of variation $v(\mu)$ of S is constant in these examples.

to λ), the risk-averse supplier seeks to avoid this uncertainty by increasing capacity even further, because doing so reduces the uncertainty in his performance. Combined, these two effects induce the supplier to choose higher capacity with more equipment failures.

5.2 Optimal Capacity Under Average-Performance Contract

The average downtime is defined as $\widehat{S} \equiv (\sum_{i=1}^{N} S_i)/N \,|\, N > 0$. Note that this is equal to the sample mean estimator for $\{S_i\}$. The condition $N > 0$ is necessary since average downtime is undefined when $N = 0$. The performance measure under AC is, then, $\widehat{S}\mathbf{1}(N > 0)$, which quantifies the performance as zero if $N = 0$ but $(\sum_{i=1}^{N} S_i)/N$ otherwise ($\mathbf{1}(\cdot)$ is an indicator variable). The supplier's capacity choice under AC with $T = w - p\widehat{S}\mathbf{1}(N > 0)$ is specified as follows.

Lemma 2 *Define* $\Delta(\lambda) \equiv \frac{1}{e^{\lambda}-1} \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} \frac{1}{n}$. *The supplier's utility is concave under AC. The supplier chooses* $\mu^* > \underline{\mu}$ *that satisfies the first-order condition* $p\mu^{-2} + 2\eta p^2[e^{-\lambda} + \Delta(\lambda)\theta(\mu)]\mu^{-3} = c/(1 - e^{-\lambda})$, *provided that p is sufficiently large to admit interior solutions. Moreover,* $\partial\mu^*/\partial c < 0$, $\partial\mu^*/\partial p > 0$, *and* $\partial\mu^*/\partial\eta > 0$. *Also,* $\partial\mu^*/\partial\lambda > 0$ *for* $\lambda \sim 0$ *but* $\partial\mu^*/\partial\lambda < 0$ *for sufficiently large* λ *for which* $e^{-\lambda}/\Delta(\lambda) \approx 0$.

Note that the ratio $e^{-\lambda}/\Delta(\lambda)$, which appears in the last part of the lemma, converges quickly to zero as λ increases (see Table 1 in Appendix A). Comparing with Lemma 1, we see that the major difference between CC and AC is on how the supplier reacts to changes in λ with regard to his capacity decision. Recall from Lemma 1 that $\partial\mu^*/\partial\lambda > 0$ under CC. When AC is used,

however, the supplier reacts in a completely different and unexpected manner: $\partial\mu^*/\partial\lambda$ exhibits non-monotonicity. See Figure 1(b) for examples demonstrating this behavior. Consider first the case where λ is sufficiently large so that $\Pr(N = 0) \approx 0$. Since the supplier is compensated based on the average downtime $(\sum_{i=1}^{N} S_i)/N$, his expected contract payment under a fixed penalty rate p is independent of how many failures occur, which is in sharp contrast to what we observed under the CC case. At the same time, the supplier benefits from *sampling variance reduction*: the higher λ, the more failures are likely, and hence, the variance of the sample mean estimator $(\sum_{i=1}^{N} S_i)/N$ decreases as more samples of performance realizations are collected. As a result, the supplier becomes less concerned about his revenue risk and is more willing to gamble by choosing lower capacity (i.e., $\partial\mu^*/\partial\lambda < 0$ when λ is high). Notice that in this case the sign of $\partial\mu^*/\partial\lambda$ is opposite of that under CC. From this discussion, we witness the first evidence that CC and AC can lead to very different consequences depending on equipment characteristics, represented by the failure rate λ, even though both contracts are designed to achieve the same goal: to give the supplier incentives to reserve a high level of capacity.

The insight just described regarding the sign of $\partial\mu^*/\partial\lambda$ no longer holds when $\lambda \sim 0$.[8] That is, given that there is a high chance that equipment never fails, the supplier reacts in the opposite way, i.e., he *increases* capacity in response to higher λ. The reason is that there is little benefit of sampling variance reduction when $\lambda \sim 0$. Instead, the supplier under AC mimics the behavior under CC, as the two performance measures become indistinguishable near $\lambda = 0$: $\sum_{i=1}^{N} S_i$ and $\widehat{S}\mathbf{1}(N > 0)$ both converge to $S_1\mathbf{1}(N = 1)$. We call this mimicking behavior around $\lambda = 0$ under AC a *no-failure effect*. We emphasize that this is a unique property that manifests itself when performance realizations are rare and random, a situation that has been overlooked in the principal-agent literature.

[8] $\lambda \sim 0$ denotes the limit in which the terms of order λ^2 and higher can be dropped in the first-order condition of the supplier's problem in Lemma 2: see the proof of Lemma 2 in Appendix D.

6 Contracting Efficiencies

In this section, we study the customer's contract design problem. For both CC and AC, the customer chooses the pair of contract terms (w, p) that solves the optimization problem

$$\max_{w,p} \; r(1 - E[\textstyle\sum_{i=1}^{N} S_i \,|\, \lambda, \mu^*]) - E[T \,|\, \lambda, \mu^*] \;\text{ subject to }\; E[S \,|\, \mu^*] \leq 1/\mu_I \text{ and } u(\mu^*) \geq 0, \qquad (4)$$

anticipating the optimal supplier response μ^*, as specified in Lemmas 1 and 2. We denote the solution of this program by a superscript $j \in \{CC, AC\}$, representing each contract. As is well known from the principal-agent literature, the first-best efficiency cannot be achieved with PBC when risk aversion is present. The inefficiency relative to the first-best, i.e., $\Pi^{FB} - \Pi^{j}$, is created by PBC's role in transforming performance risk into the supplier's financial risk. In the following analysis, our main interest is in investigating how the optimal penalty rates (p^{CC} and p^{AC}) and the contracting inefficiencies ($\Pi^{FB} - \Pi^{CC}$ and $\Pi^{FB} - \Pi^{AC}$) behave as a function of the important environmental characteristic, namely, the equipment failure rate λ.

Similar to what we observed in Section 4 for the benchmark case, the solution behaviors turn out to be quite different across the case where STC does not bind at optimum (which happens if λ is sufficiently large) and the case when the constraint binds at optimum (which happens if λ is sufficiently small). A sufficient condition for the latter case is (2). For ease of exposition, we first present an analysis of the non-binding case in Section 6.1 and then turn to the discussion of the binding case in Section 6.2, where counterintuitive results are derived. Thus, we move forward in the reliability spectrum, from a scenario in which failures are moderately infrequent to a scenario in which failures are rare, characterized by the sufficient condition $\lambda \leq c\mu_I^2/r$.

6.1 Case 1: Equipment Fails at Moderate Infrequency ($\lambda > c\mu_I^2/r$)

We first note that the condition $\lambda > c\mu_I^2/r$, which ensured that STC does not bind at optimum in the benchmark case, no longer guarantees the same when capacity is not observable and hence the customer cannot contract directly on it. Regardless, we observe numerically that non-binding solution is obtained whenever λ is sufficiently larger than $c\mu_I^2/r$. In such cases, unfortunately, analytical specification of the optimal contract terms is intractable, partly due to the implicit

nature of the optimal capacity expressions appearing in Lemmas 1 and 2. In particular, concavity of the optimization problem or sensitivity results with respect to λ cannot be easily established. However, numerical experiments consistently show that the solution behavior under CC is rather straightforward: the optimal penalty rate p^{CC} and the contracting inefficiency $\Pi^{FB} - \Pi^{CC}$ are both increasing functions of λ. This result is in line with intuition. With more frequent failures the total downtime is expected to be longer, so the profit-maximizing customer attempts to increase the equipment uptime by incentivizing the supplier to reserve higher service capacity, which is achieved by imposing a larger penalty rate p^{CC}. In other words, the customer offers a high-powered incentive to compensate for the loss of total uptime with quick service times. The deviation of the profit from the first-best, $\Pi^{FB} - \Pi^{CC}$, then increases with λ since the larger penalty rate generates greater fluctuations in the supplier's income stream, forcing the supplier to request higher risk premium.

Under AC, however, the shapes of p^{AC} and $\Pi^{FB} - \Pi^{AC}$ as a function of λ are more convoluted and do not show a consistent pattern. In particular, numerical examples show that $\Pi^{FB} - \Pi^{AC}$ may exhibit non-monotonicity; whereas it increases in λ under some parameter combinations, in general, it may decrease after an initial increase. This observation suggests two key facts. First, the above reasoning that the customer incentivizes the supplier to provide quick service times to compensate for the loss of total uptime only provides an incomplete picture; there is another force that counters this effect. Second, this counteractive force is stronger under AC than under CC. We identify this force in the next section, where we consider the case in which STC binds at optimum.

6.2 Case 2: Equipment Fails Rarely ($\lambda \leq c\mu_I^2/r$)

When λ is sufficiently small such that the condition (2) is satisfied, i.e., $\lambda \leq c\mu_I^2/r$, a complete analytical description of the solution behavior can be obtained, as long as the following condition is satisfied:

$$\theta(\mu) + \mu\theta'(\mu) \geq 0 \quad \text{for } \mu \geq \underline{\mu}. \tag{5}$$

Along with (2), the condition (5) is sufficient to ensure that STC binds at optimum. It is trivially satisfied when $v(\mu)$ is constant, as is the case if the distribution of service time is exponential or gamma with a constant shape parameter. The literature makes such an assumption frequently (for example, see Ata and Shneorson 2006). Note that (5) is a sufficient condition that is not tight, in

that STC will continue to bind in many instances in which the condition is violated.

When STC binds at optimum, analysis is simplified since, in that case, the optimal induced capacity should be equal to a constant, μ_I. (In Section 7.1, we check the robustness of the results obtained in this subsection by allowing the capacity target μ_I to vary with λ.) In the following proposition we specify the equilibrium solutions under CC and AC.

Proposition 1 *Suppose that the conditions (2) and (5) hold. Under CC, the customer offers the penalty rate* $p^{CC} = 2c\mu_I^2\left[\lambda\left(1+\sqrt{1+8\eta c\mu_I V^{CC}(\lambda)}\right)\right]^{-1}$, *where* $V^{CC}(\lambda) = [1+\theta(\mu_I)]/\lambda$. *Under AC, the customer offers the penalty rate* $p^{AC} = 2c\mu_I^2\left[(1-e^{-\lambda})\left(1+\sqrt{1+8\eta c\mu_I V^{AC}(\lambda)}\right)\right]^{-1}$, *where* $V^{AC}(\lambda) = \frac{e^{-\lambda}+\Delta(\lambda)\theta(\mu_I)}{1-e^{-\lambda}}$. *As a response, the supplier chooses* $\mu^{CC} = \mu^{AC} = \mu_I$ *and is left with zero utility. The resulting expected customer profit is* $\Pi^j = r(1-\lambda/\mu_I) - c(\mu_I - \underline{\mu}) - \psi^j$, *where* $\psi^{CC} = \frac{c\mu_I}{2}\left(\frac{1+v(\mu_I)^2}{1+\theta(\mu_I)}\right)\left(1-2\left(1+\sqrt{1+8\eta c\mu_I V^{CC}(\lambda)}\right)^{-1}\right)$ *for* $j = CC$ *and* $\psi^{AC} = \frac{c\mu_I}{2}\left(\frac{e^{-\lambda}+\Delta(\lambda)v(\mu_I)^2}{e^{-\lambda}+\Delta(\lambda)\theta(\mu_I)}\right)\left(1-2\left(1+\sqrt{1+8\eta c\mu_I V^{AC}(\lambda)}\right)^{-1}\right)$ *for* $j = AC$.

When STC binds, the contracting inefficiency is completely captured by the risk premium ψ^j, that is, $\Pi^{FB} - \Pi^j = \psi^j$. This is because the optimal capacity $\mu^j = \mu_I$ does not deviate from the first-best level $\mu^{FB} = \mu_I$, thereby leaving the portion of Π^j other than ψ^j unchanged from the first-best quantity. (In contrast, when the constraint does not bind, as in Case 1 above, risk premium is a major, but not the only contributor to the inefficiency because in that case the optimal capacity deviates from the first-best level, i.e., $\mu^{FB} \neq \mu_I$. Hence, in that case, $\Pi^{FB} - \Pi^j \neq \psi^j$.) The changes in p^j and ψ^j, $j \in \{CC, AC\}$, with respect to parameters c, μ_I, and η are qualitatively the same across the contracting scenarios. Under both contracts, we can show that: (i) $\partial p^j/\partial c > 0$, $\partial p^j/\partial \mu_I \geq 0$, $\partial p^j/\partial \eta < 0$, and (ii) $\partial \psi^j/\partial c > 0$, $\partial \psi^j/\partial \mu_I \geq 0$, $\partial \psi^j/\partial \eta > 0$. In other words, the customer has to offer higher incentive (p^j) to induce the target capacity μ_I as (a) the unit capacity cost goes up, (b) the service time constraint is tightened, and (c) the supplier becomes less risk-averse. The last result stems from the fact that a less risk-averse supplier is more willing to take a chance on a fortuitous performance outcome, i.e., realization of shorter service time. Similarly, efficiency loss in the supply chain, as measured by the risk premium ψ^j, increases as (a) the unit capacity cost goes up, (b) the service time constraint is tightened, and (c) the supplier becomes more risk-averse. These results are in line with intuition. Next, we state how p^j and ψ^j change with λ.

Proposition 2 *Suppose that the conditions (2) and (5) hold. Then*

(i) $\partial p^{CC}/\partial\lambda < 0$*, whereas* $\partial p^{AC}/\partial\lambda < 0$ *for* $\lambda \sim 0$ *but* $\partial p^{AC}/\partial\lambda > 0$ *for sufficiently large* λ *for which* $e^{-\lambda}/\Delta(\lambda) \approx 0$*. Moreover,* $\lim_{\lambda\to 0} p^{CC} = \lim_{\lambda\to 0} p^{AC} = \infty$*.*

(ii) $\psi^{CC} \geq \psi^{AC}$ *if* $V^{CC}(\lambda) \geq V^{AC}(\lambda)$*. Moreover,* $\partial\psi^{CC}/\partial\lambda < 0$ *and* $\partial\psi^{AC}/\partial\lambda < 0$*, while* $\lim_{\lambda\to 0}\psi^{CC} = \lim_{\lambda\to 0}\psi^{AC} = \frac{c\mu_I}{2}\left(\frac{1+v(\mu_I)^2}{1+\theta(\mu_I)}\right)$*.*

The solution behaviors near $\lambda = 0$ are explained in detail in Appendix B. In the following, we examine the rest of the results one by one.

Changes in optimal penalty rates as a function of λ. We observe that, while p^{CC} is monotonically decreasing in λ, p^{AC} decreases initially but goes up as failures occur more frequently. Such behaviors reflect the supplier's differing responses to changes in λ under the two contracts, as discussed in Section 5. Under CC, as we found out, the supplier tends to choose higher capacity when more equipment failures are likely; the customer can then utilize this voluntary action to reach the target capacity μ_I without providing a strong contractual incentive, i.e., without offering a high penalty rate p. However, the same logic does not hold when AC is used, precisely because the supplier's capacity choice μ^* in response to λ exhibits non-monotonicity, as explained in Section 5.2. Therefore, $\partial p^{AC}/\partial\lambda$ is non-monotonic as was $\partial\mu^*/\partial\lambda$ but in the opposite direction, since, again, the customer's goal is to induce the target capacity μ_I. This counterintuitive solution behavior is a direct consequence of the no-failure effect.

Relative magnitudes of risk premiums. We find that ψ^{CC} and ψ^{AC} are not equal in general. In fact, we can gain greater insight into when one contract leads to a more efficient outcome than the other by considering a special case in which the coefficient of variation $v(\mu)$ of the service time S is constant (denoted simply as v). In this special case, ψ^{CC} and ψ^{AC} differ only by $V^{CC}(\lambda) = (1+v^2)/\lambda$ and $V^{AC}(\lambda) = (e^{-\lambda} + \Delta(\lambda)v^2)/(1-e^{-\lambda})$, which are in fact the squares of the coefficients of variation contained in the performance measures $\sum_{i=1}^{N} S_i$ and $\widehat{S}\mathbf{1}(N>1)$, respectively. Comparing the two quantities, we can state the following result.

Corollary 1 *Assume that the conditions (2) and (5) hold and that* $v(\mu)$ *does not vary in* μ*. Let* $v \equiv v(\mu)$ *and define* $\omega(\lambda) \equiv \sqrt{\frac{(1-e^{-\lambda})/\lambda - e^{-\lambda}}{\Delta(\lambda) - (1-e^{-\lambda})/\lambda}}$*. Then* $\psi^{CC} \geq \psi^{AC}$ *iff* $v \leq \omega(\lambda)$*.*

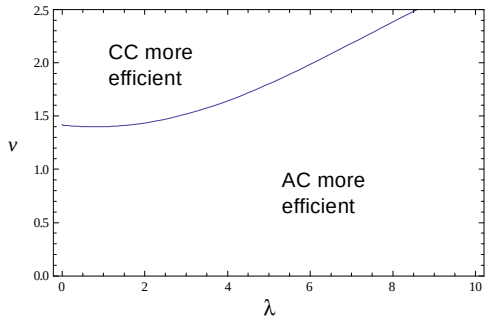


Figure 2: Regions on the (λ, v) plane where one contract dominates the other, when $v = v(\mu)$ is a constant.

According to Corollary 1, the customer (and hence the supply chain) can achieve better efficiency with AC than with CC if and only if v, the variability of service time, is less than or equal to $\omega(\lambda)$. Numerical plotting shows that $\omega(\lambda)$ is a convex function that has a unique minimizer at $\lambda = 0.79$ with $\omega(\lambda) = 1.4$ at that point (see Figure 2). Therefore, AC always performs better than CC whenever $v < 1.4$, regardless of λ. Considering that the coefficient of variation of 1.4 is a very large number for most well-known distributions, we conclude that AC is preferable unless the service time S exhibits extremely large variability. Numerical experiments lead to a similar conclusion even if we allow $v(\mu)$ to decrease, as doing so in fact has a larger impact on reducing ψ^{AC} than ψ^{CC}.

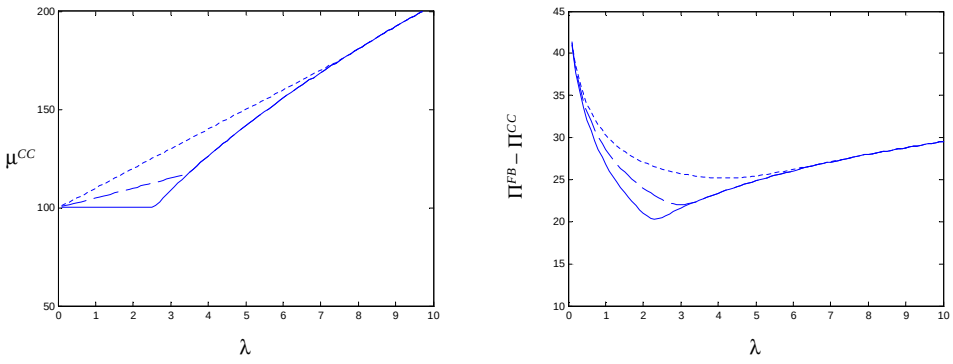
What drives AC to be more efficient than CC in most reasonable cases? In brief, the performance measure used under CC typically contains more variability than does the measure under AC, since the latter, the sample-average measure, effectively removes the uncertainty stemming from stochastic failures through the division by N. However, N is a random number, so the division actually introduces a noise which is negligible if the service time S contains modest level of variability but becomes magnified otherwise. This amplification of the extra noise, which CC is free from, pushes AC to become more inefficient than CC if the variability $v(\mu)$ is sufficiently large. For detailed explanation, refer to Appendix C.

Changes in risk premiums as a function of λ. It is shown in part (ii) that risk premiums ψ^{CC} and ψ^{AC}, and hence the efficiency loss in the supply chain, decrease in λ. See the solid lines in Figure 3(b) and Figure 4 for illustrations (in the figures, the condition $\lambda \leq c\mu_1^2/r$ corresponds to $\lambda \leq 2$). The two drivers of this result are intricately related. First, the optimal capacity level

μ_I attained under the condition (2) is free of λ. Then, in the absence of any confounding factors originating from λ, reductions in ψ^{CC} and ψ^{AC} are solely due to reductions in the performance measure variabilities $\sqrt{V^{CC}(\lambda)}$ and $\sqrt{V^{AC}(\lambda)}$, which are decreasing functions of λ. This is a direct consequence of *risk pooling*: the more i.i.d. service time samples are collected, the less variable the supplier's performance is. As a consequence, the supplier faces less revenue risk, and, hence, smaller risk premium. (Note that this identifies "another force" that we mentioned in the discussion of non-monotone behavior of $\Pi^{FB} - \Pi^{AC}$ in Case 1; risk pooling plays a greater role under AC than under CC, contributing to the non-monotonicity. See Figure 4).

This observation leads us to the following conclusion: under the condition that the service time constraint binds at optimum, which is guaranteed to happen for small λ that satisfies (2), contracting efficiency is *worst* when the equipment almost never fails. The intuition is as follows. If equipment fails rarely, the customer's primary concern is not further increasing uptime-driven profitability (which is already close to maximum) but making sure that a downtime, if it is realized, is not exceedingly long. Hence, it is optimal for the customer to induce the supplier to reserve the minimum, constant target level of capacity. With the mean service time fixed at the inverse of the target capacity level, then, the only remaining consequence of reducing λ is that the risk pooling effect is diminished, since with fewer failures, the supplier has fewer opportunities to perform and present signals about his capacity decision. As a result, the variability of performance realization becomes larger with smaller failure frequency, thereby creating a larger contracting inefficiency. Therefore, when failures occur due to imperfect equipment reliability, firms face the following dilemma: while firms value high reliability, it may become very inefficient to contract with a supplier to deliver fast restoration/recovery services for reliable equipment.

Summarizing, we find that frequency of equipment failures is an important factor in determining how efficiently PBC can be implemented. In particular, one may encounter a situation in which higher equipment reliability leads to less efficient contracting. This happens when failures are so rare that the customer's concern for ensuring a minimum service time requirement outweighs her profitability (i.e., the service time constraint binds at optimum).



(a) Equilibrium capacity under *CC*

(b) Contracting inefficiency under *CC*

Figure 3: Equilibrium capacity μ^{CC} and contracting inefficiency $\Pi^{FB} - \Pi^{CC}$ as a function of λ. In these examples, $r/c = 5000$, $v = 0.5$, and the target capacity $\mu_I(\lambda)$ of the constraint $\mu^* \geq \mu_I(\lambda)$ is assumed to be linear function with $\mu_I(\lambda) = 100 + \mu_1\lambda$. In both (a) and (b), the solid horizontal line, the dashed line, and the dotted line correspond to $\mu_1 = 0$, $\mu_1 = 5$, and $\mu_1 = 10$, respectively. The values of λ where the kinks are located in (a) represent the points at which the constraint starts to bind at optimum as λ decreases.

7 Discussion of Assumptions and Extensions

Thus far, we have presented the analysis based on the set of assumptions we made in Section 3. We have kept the assumptions as simple as possible in order to clearly identify the drivers of the results we have found. In reality, there are situations that require altering our assumptions. In this section we briefly discuss the impact of relaxing some of them.

7.1 Non-Constant Service Time Target

We have assumed throughout the analysis that the service time target s_I in STC, or equivalently, the capacity target μ_I in the constraint $\mu^* \geq \mu_I$, is a constant. Although many service contracts appear to suggest that this assumption is reasonable (see Footnote 5), in general, the target may vary with equipment failure frequency λ. If it does, we expect μ_I to increase with λ, as the profitability-conscious customer would prefer quicker service time to compensate for the loss of equipment uptime due to more failures. In this subsection we relax the constant target assumption and see if, in particular, the counterintuitive finding that contracting efficiency decreases with equipment reliability for small λ continues to hold. To this end, we conduct numerical experiments assuming that the capacity target linearly increases with λ, such that $\mu_I(\lambda) = \mu_0 + \mu_1\lambda$. The results

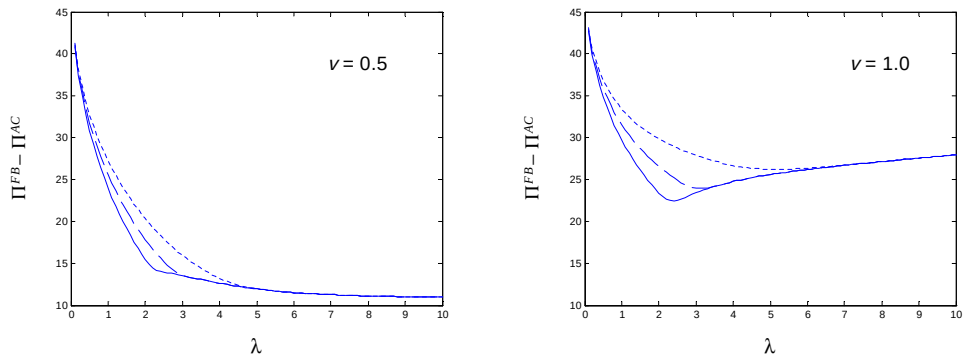


Figure 4: Contracting inefficiency $\Pi^{FB} - \Pi^{AC}$ under AC as a function of λ, for $v = 0.5$ and $v = 1$. All other parameter values for the three lines (solid, dashed, dotted) are the same as those of Figure 3.

are summarized in Figures 3 and 4 (compare the dashed and the dotted lines with the solid line). We see there that our finding does not change qualitatively; although the effect is more muted compared to the constant target assumption, it is clear that contracting inefficiency continues to decrease in λ as long as λ is sufficiently small so that STC binds at optimum. In addition, as observed in Section 6.1, when the constraint does not bind, the inefficiency increases in λ under CC but it may increase or decrease under AC. Hence, our analysis in Section 6 is robust to the constant service time target assumption.

Although we chose the simple linear form $\mu_I(\lambda) = \mu_0 + \mu_1\lambda$ as an example, we note that it actually has practical interpretation. Many managers in service-providing organizations find it difficult, if not impossible, to guess the opportunity cost of equipment downtime r. (For example, what is the monetary value of losing the war because an aircraft was down due to a defective part?) Hence, it is difficult to represent these organizations as profit-maximizers since r is unknown. As an alternative, it is common in the literature (see, for example, Sherbrooke 1968) to assume that their objective is to minimize expected cost subject to a constraint on expected equipment availability. In fact, specifying a target availability is a wide-spread practice; terms like "target availability of 95%" or "target availability of 99%" are easily spotted in many service contracts. As specifying a target availability is equivalent to specifying a target on the total downtime during the contract length, the availability constraint can be expressed as $E[\sum_{i=1}^{N} S_i \,|\, \lambda, \mu^*] = \lambda/\mu^* \leq 1/\mu_T$, or equivalently, $\mu^* \geq \mu_T\lambda$, where μ_T is a constant. Comparing this expression with the service time constraint

$\mu^* \geq \mu_I(\lambda) = \mu_0 + \mu_1\lambda$, we see that the capacity targets in the two constraints become equivalent if $\mu_1 = \mu_T$ and if λ is sufficiently large so that $\mu_0 \ll \mu_1\lambda$. Therefore, the linear form of $\mu_I(\lambda)$ can be considered as an approximation of the two service targets that the cost-minimizing customer requires: it represents the minimum service time target when λ is small, whereas it represents an availability target when λ is large.

7.2 Endogenous Failure Rate

In the base model, we treated the failure rate λ as an exogenous variable. Although assuming λ to be beyond the supplier's control is reasonable in many situations (such as when equipment failures occur after natural disasters or in the Clean Harbors example mentioned in the Introduction), there are also situations in which the supplier is able to lower the frequency of failures, for example, by improving equipment reliability. It turns out that endogenizing the failure rate along with capacity presents many analytical challenges (see Kim et al. 2009). However, with a simple model extension, we find interesting results that refine the insights that we obtained previously.

Let us assume that the supplier has two failure rate choices at the outset: low (L) or high (H), with $\lambda_L < \lambda_H$. Discrete failure rates may arise, for example, when the supplier elects to retrofit equipment or software, which results in a jump in reliability. Choosing the low failure rate λ_L (choosing higher reliability) requires an additional investment amount $K \geq 0$. Therefore, the supplier's utility can be redefined as $U_t(\mu) \equiv u_t(\mu) - K\mathbf{1}(t = L)$, $t \in \{L, H\}$, where $\mathbf{1}(\cdot)$ denotes the indicator variable and $u_t(\mu)$ is the utility function from (1) with λ replaced by λ_t. Presented with contract terms (w, p), the supplier does the following: compute the optimal capacity levels μ_L^* and μ_H^* that maximize $U_L(\mu)$ and $U_H(\mu)$, respectively, compare $U_L(\mu_L^*)$ and $U_H(\mu_H^*)$, and choose λ_H if and only if $U_L(\mu_L^*) \leq U_H(\mu_H^*)$. When does the supplier choose λ_L, the higher reliability? The following lemma answers this question for the special case where $v(\mu)$ is constant.

Lemma 3 *Let $\underline{\underline{p}} = \max\{\underline{p}_L, \underline{p}_H\}$, where $\underline{p}_L$ and $\underline{p}_H$ are the minimum penalty rates under λ_L and λ_H, respectively, beyond which the supplier is induced to choose his capacity above the default level $\underline{\mu}$. Suppose that $v(\mu)$ is constant and consider $p \geq \underline{\underline{p}}$.*

(i) Under CC or under AC with $\lambda_t \sim 0$, there exists at most one $p^\dagger \geq \underline{\underline{p}}$ such that the supplier chooses λ_H if $p \leq p^\dagger$ and λ_L otherwise.

(ii) Under AC with λ_t sufficiently large for which $e^{-\lambda_t}/\Delta(\lambda_t) \approx 0$, the supplier always chooses λ_H.

In Lemmas 1 and 2 we observed that the supplier's optimal capacity choice exhibits markedly different behaviors across CC and AC as λ varies. This difference is reflected in his failure rate choice, as illustrated in parts (i) and (ii) of Lemma 3. Consider CC first. Under this contract, the supplier has an incentive to lower the failure rate (choose higher reliability). This is because less frequent failures subtract the number of downtime realizations, hence reducing the total downtime and the resulting penalty. The only hindrance in choosing λ_L is the investment cost K. If it is too large, reduction of failure rate is too costly to implement. Part (i) of the lemma states this tradeoff and shows that the supplier tends to lower the failure rate if the penalty rate p is high enough to justify the investment. The same result holds for AC when λ_L and λ_H are very small, since AC and CC converge to each other in the low-λ limit.

By contrast, when λ_L and λ_H are sufficiently large (such that the condition in part (ii) of the lemma is satisfied), AC may provide an opposite incentive that leads the supplier to prefer more frequent failures, i.e., lower reliability. This happens because sampling variance reduction is in full effect; with more failures (more samples of service times) the supplier's performance outcome under AC is less variable due to averaging, increasing his utility. In addition, choosing λ_L incurs an extra cost K which can be avoided if λ_H is chosen. Hence, there is no reason to choose λ_L in this scenario. Therefore, the supplier may or may not be incentivized to lower the failure rate depending on which of CC and AC is used and whether the failures are very rare or not.

Lower failure rate is beneficial to the customer since her revenue is proportional to equipment uptime, which increases as failures occur less often. The insights from Lemma 3 suggest that there may be circumstances where CC is a better contract to use, in contrast to our earlier observation that AC is generally preferred when failures are driven by an exogenous process. The difference is due to the fact that CC gives the supplier greater incentives to lower the failure rate than AC does. Figure 5 illustrates this point. Recall from the discussion below Corollary 1 that AC is superior to CC whenever $v < 1.4$ and when the supplier had no control over the failure rate. If the supplier can influence the failure rate, the examples in Figure 5 show that the opposite can be true even with $v = 1$; while AC is still more efficient than CC for very small λ_L and λ_H, the situation is

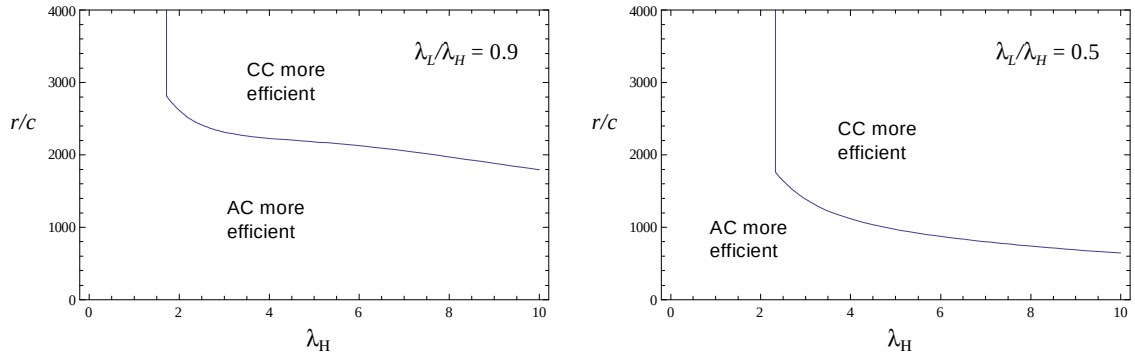


Figure 5: Regions on the $(\lambda_H, r/c)$ plane where one contract dominates the other, for two ratio values $\lambda_L/\lambda_H = 0.9$ and $\lambda_L/\lambda_H = 0.5$. $v = 1$ is assumed in both examples.

reversed for intermediate values of λ_L and λ_H with sufficiently high revenue-to-cost ratio r/c. In addition, CC becomes more dominant as the supplier is able to reduce the failure rate by a wider margin (as the ratio λ_L/λ_H decreases). Therefore, the intuitive conclusion from the previous section that the benefit of risk pooling points to AC as the preferred contract ceases to be true when the supplier controls not only the service capacity but also the frequency of equipment failures. This is an important reminder that there is no "one-size-fits-all" when it comes to choosing the right performance metric in the PBC environment. In particular, whether the supplier can influence the frequency of disruptions should be an important factor in the decision.

7.3 Risk-Averse Customer

We have assumed that the supplier is averse to financial risk but the customer is not, on the grounds that the supplier, as a smaller entity, is more susceptible to fluctuations in cash flow. This is a reasonable assumption in the majority of bilateral relationships in which the customers are typically larger and more diversified (such as semiconductor manufacturers or government agencies) than the supplier of customized services who often rely heavily on a single customer, as is the case for many defense contractors. This assumption is also consistent with our modeling construct where the customer has a superior market power as the "principal" who determines the terms of contracts.

However, there may be cases where the situation is reversed. To gain insights on how the results that we obtained thus far change in such a case, let us assume that the customer is risk-averse with the coefficient of risk aversion η_c and the supplier is risk-neutral. The additional term that

appears in the customer's utility function, $\psi_c = \eta_c(r^2\text{Var}[\sum_{i=1}^{N} S_i \mid \lambda, \mu] + p^2\text{Var}[X \mid \lambda, \mu])$, where $X = \sum_{i=1}^{N} S_i$ for CC and $X = \widehat{S}\mathbf{1}(N > 0)$ for AC, represents the degree to which she is reluctant to participate in the trade because of the financial risk posed by PBC. This quantity represents the contracting inefficiency. There are a few notable changes in the results under this setup. First, the condition (2), which led STC to bind under both CC and AC previously, no longer guarantees to do the same under AC. However, the constraint continues to bind for sufficiently small λ. For simplicity's sake, let us consider only such cases. It is found that the optimal penalty rates p^{CC} and p^{AC} are both decreasing in λ, in contrast to our earlier finding that p^{AC} exhibits non-monotonicity (see (i) of Proposition 2). This reminds us that such a distinct feature was a consequence of a risk-averse supplier's opportunistic behavior in choosing capacity (see Section 5.2); since the party who makes capacity decision, the supplier, is now risk-neutral, this feature is no longer present.

How contracting inefficiency (represented by ψ_c) changes as a function of λ, the other central focus of this paper, turns out to be similar to what we observed earlier. Under the same conditions specified in Proposition 2, we can analytically show that, as before, ψ_c^{CC} decreases in λ (proof is omitted). A similar pattern is observed for ψ_c^{AC} numerically, as long as STC is binding. This is quite remarkable considering that there is an additional source of risk that tends to increase with λ which did not matter to a risk-neutral customer: the revenue that depends on uncertain equipment uptime (the first term in the expression of ψ_c above). To a risk-averse customer, this revenue becomes more volatile as failures occur more frequently, and the increase of this risk works counter to the benefit of reduced risk through pooling, the effect that we observed previously. The fact that ψ_c^{CC} and ψ_c^{AC} continue to decrease with λ indicates that the latter dominates even in this case. Therefore, we conclude that our earlier counterintuitive result is quite robust; no matter who is risk-averse – the supplier, the customer, or both – high equipment reliability begets low contracting efficiency.

8 Conclusion

In this paper we study issues arising from performance-based contracting for restoration and recovery services, which are essential when minimizing the impact of disruptions for mission-critical operations. Despite a large volume of literature on service contracting, surprisingly little attention

has been directed to outsourced services in an environment characterized by low-frequency, high-impact events such as equipment failures. With increasing use of PBC for service outsourcing in both the commercial and the government sectors, analyzing the merits and pitfalls of PBC in such an environment provides important managerial guidelines to practitioners who face unique issues that arise in those arenas. In addition, we contribute to the ongoing discussion regarding the best performance metrics to be used in PBCs. In his testimony to the Committee on Homeland Security, the Chief Procurement Officer of the Department of Homeland Security stated that "Commercial organizations told the Panel that implementing the [PBC] can be difficult, particularly in identifying the appropriate performance standards to measure" (Department of Homeland Security 2008). Our work aims to help firms in this regard.

We find that one prominent source of inefficiency when contracting in this environment is the low rate of system disruptions. Since disruptions of mission-critical systems are relatively rare, the customer has few opportunities to observe signals about a supplier's choice of service capacity through repeated realizations of the supplier's performance, namely, service completion times. In an extreme scenario, a disruption may not occur at all within a contracting period, revealing no signals about the supplier's capacity choice. With limited information about the supplier's decision, it becomes costly to provide a high-powered incentive via PBC. This implies that, counter to our intuition, implementing PBC may be least efficient when the equipment is most reliable. This happens, in particular, when equipment failures are so rare that the customer's primary concern is ensuring a minimum downtime target when a failure actually occurs. Under such a circumstance, firms face a dilemma if they value both fast restorations and high reliability: while it is crucial that the customer resolve any disruptive event as quickly as possible, it may become very inefficient to contract with a supplier to achieve that objective when the equipment does not fail often.

This analysis provides a theoretical support to the argument that PBC should be implemented with discretion. Currently, there is a major policy shift in the government sector which advocates a complete switch to PBC for all service acquisitions (Office of Management and Budget 2003). Our study reveals, however, that there may be situations (such as when disruptions are rare) in which PBC creates potentially very high agency costs. In such cases, it might be prudent to consider in-sourcing or expending significant effort to continuously monitor a supplier's capacity investments.

We also highlighted other nontrivial issues that arise in the settings characterized by infrequent disruptions by analyzing two widely used contracts that function identically but which actually yield quite different incentive effects. One contract is based on cumulative system downtime, and the other is based on sample-average of downtime. Although both motivate the supplier to invest in service capacity, they also create very different incentive structures to a risk-averse supplier, which in turn affect the way optimal contracts are designed. For example, the optimal penalty rate may change non-monotonically in the equipment failure rate if sample-average downtime is used as the basis of the supplier performance evaluation. To the best of our knowledge, this unexpected feature has not been studied in the contracting literature. We also compare the relative efficiencies of the two contracts, and find that the contract based on sample-average downtime is superior in most practical situations if disruptions are driven by an exogenous process such as natural disasters. If the supplier also controls the frequency of disruptions, however, the contract based on cumulative downtime may be preferred instead.

The model we propose in this paper is not without shortcomings, as we have made a number of simplifying assumptions that permit tractable analysis and highlight important features. For example, some of the effects driven by risk aversion may be reduced if the supplier serves a large number of customers, thanks to risk pooling. Although additional details would sharpen the managerial insights, we believe that our model captures the most important aspects of our problem setting, namely, using PBC for rarely requested restoration and recovery services, and they serve as useful guidelines to practitioners. As for the future directions of our research, we envision many ways in which our model can be extended. One promising idea is to fully account for the multi-indentured structure of equipment and investigate the effects of having heterogenous failure processes for different components. Another fruitful direction would be to extend this model to a repeated setting. In the government sector, for example, PBC is sometimes augmented with contract renewals, which provide an added incentive to the supplier to invest in capacity as past performance determines whether the contract is renewed. We believe that our model paves the way for analyzing this practice as well. Finally, considering how product design impacts after-sales service performance would allow us to view the issues analyzed in this paper from a product lifecycle planning perspective.

References

[1] Abreu, D., P. Milgrom, D. Pearce. 1991. Information and timing in repeated partnerships. *Econometrica*, 59(6), 1713-1733.

[2] Allon, G., A. Federgruen. 2007. Competition in service industries. *Operations Research*, 55(1), 37-55.

[3] Army Material Command. 2006. Performance Work Statement (PWS) samples: ADPE Maintenance. http://www.amc.army.mil/amc/rda/rda-ac/pbsc/amcac-adpe-maint.doc.

[4] Ata, B., S. Shneorson. 2006. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science*, 52(11), 1778-1791.

[5] Bolton, P., M. Dewatripont. 2005. *Contract Theory*. MIT Press, Cambridge, MA.

[6] Cohen, M. A., N. Agrawal, V. Agrawal. 2006. Winning in the aftermarket. *Harvard Business Review*, 84(5), 129-138.

[7] 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.

[8] Department of Defense. 2003. Department of Defense Directive 5000.1. http://www.acq.osd.mil/ie/bei/pm/ref-library/dodd/d50001p.pdf.

[9] Department of Homeland Security. 2008. Statement of Thomas W. Essig, Chief Procurement Officer, Department of Homeland Security. http://homeland.house.gov/SiteDocuments/20080508101017-04748.pdf.

[10] Environmental Protection Agency. 2003. OSWER Guidance 9272.0-21. http://www.epa.gov/fedfac/pdf/performance_based.pdf.

[11] Frauenheim, E. 2003. Disaster industry finds silver lining. *CNET News* (May 1), http://news.cnet.com/Disaster-industry-finds-silver-lining/2100-1011_3-999364.html?tag=mncol;txt.

[12] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2), 79-141.

[13] Geary, S. 2006. Ready for combat. *DC Velocity*, 4(7), 75-80.

[14] Gilbert, S. M., Z. K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: the principal-agent perspective. *Management Science*, 44(12), 1662-1669.

[15] Gollier, C. 2001. *The Economics of Risk and Time*. MIT Press, Cambridge, MA.

[16] Government Accountability Office. 2002. Guidance needed for using performance-based service contracting. Report GAO-02-1049.

[17] Gurvich, I., M. Armony, A. Mandelbaum. 2005. Service level differentiation in call centers with fully flexible servers. *Management Science*, 54(2), 279-294.

[18] Harrington, L. 2006. Getting service parts logistics up to speed. *Inbound Logistics* (November). http://www.inboundlogistics.com/articles/features/1106_feature02.shtml.

[19] Hasija, S., E. J. Pinker, R. A. Shumsky. 2008. Call center outsourcing contracts under information asymmetry. *Management Science*, 54(4), 793-807.

[20] Kim, S.-H., M. A. Cohen, S. Netessine. 2007. Performance contracting in after-sales service supply chains. *Management Science*, 53(12), 1843-1858.

[21] Kim, S.-H., M. A. Cohen, S. Netessine. 2009. Reliability or inventory? Analysis of product support contracts in the defense industry. Working paper.

[22] Kleindorfer, P. R. and G. H. Saad. 2005. Managing disruption risks in supply chains. *Production and Operations Management*, 14(1), 53-68.

[23] Laffont, J.-J., J. Tirole. 1993. *A Theory of Incentives in Regulation and Procurement*. Boston, MA. MIT Press.

[24] Lebedev, N. N. 1972. *Special Functions & Their Applications*. Dover, Mineola, New York.

[25] Lu, L. X., J. A. Van Mieghem, R. C. Savaskan. 2009. Incentives for quality through endogenous routing. *Manufacturing & Service Operations Management*, 11(2), 254-273.

[26] Milgrom, P. R. 1981. Good news and bad news: representation theorems and applications. *Bell Journal of Economics*, 12, 380-391.

[27] Milner, J., T. Olsen. 2006. Service Level Agreements in Call Centers: Perils and Prescriptions. *Management Science*, forthcoming.

[28] Muckstadt, J. A. 2005. *Analysis and Algorithms for Service Parts Supply Chains*. Springer, New York.

[29] Office of Management and Budget. 2003. Performance-based service acquisition: contracting for the future. http://www.whitehouse.gov/omb/procurement/0703pbsat.pdf.

[30] Oliver Wyman. 2007. Airlines have not yet realized the full benefits of new MRO supplier relationships. http://www.oliverwyman.com/ow/pdf_files/AAD07-MROSurvey-SupplierRelations.pdf.

[31] Plambeck, E. L., S. A. Zenios. 2003. Incentive efficient control of a make-to-stock production system. *Operations Research*, 51(3), 371-386.

[32] Ren, Z. J., Y.-P. Zhou. 2008. Call center outsourcing: coordinating staffing levels and service quality. *Management Science*, 54(2), 369-383.

[33] Shapiro, C., J. E. Stiglitz. 1984. Equilibrium unemployment as a worker discipline device. American Economic Review, 74(3), 433-444.

[34] Sheffi, Y. 2005. *The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage*. MIT Press, Cambridge, MA.

[35] Sherbrooke, C. C. 1968. METRIC: A Multi-Echelon Techniques for Recoverable Item Control. *Operations Research*, 16, 122-141.

[36] Sherbrooke, C. C. 1992. *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*. Wiley & Sons, New York.

[37] Slay, F. M. et al. 1996. Optimizing spares support: the aircraft sustainability model. Logistics Management Institute Report AF501MR1.

[38] Sobie, B. 2007. Maintenance for low-cost carriers: outer limits. *Airline Business*, 23(10), 46-53.

[39] Stansbury, T. 2004. Choose the right partner. *Communications News*, 41(12), 28-29.

[40] Tomlin, B. 2006. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science*, 52(5), 639-657.

[41] Van Mieghem, J. A. 2007. Risk mitigation in newsvendor networks: resource diversification, flexibility, sharing, and hedging. *Management Science*, 53(8), 1269-1288.

[42] *The Wall Street Journal*. 2009. GE's focus on services faces test. (March 3).

Technical Appendix to: Contracting for Infrequent Restoration and Recovery of Mission-Critical Systems

Sang-Hyun Kim

Yale School of Management, Yale University, New Haven, CT 06520,

sang.kim@yale.edu

Morris A. Cohen • Serguei Netessine, and Senthil • Veeraraghavan

The Wharton School, University of Pennsylvania, Philadelphia, PA 19104,

cohen@wharton.upenn.edu • netessine@wharton.upenn.edu • senthilv@wharton.upenn.edu

A Tables and Figures

λ	1	2	3	4	5	6	7	8	9	10
$\Delta(\lambda)$	0.767	0.577	0.433	0.330	0.258	0.208	0.172	0.147	0.128	0.113
$e^{-\lambda}/\Delta(\lambda)$	0.480	0.235	0.115	0.056	0.026	0.012	0.005	0.002	0.001	4×10^{-4}

Table 1: Numerical values of $\Delta(\lambda)$ and $e^{-\lambda}/\Delta(\lambda)$ for $\lambda = 1, ..., 10$.

B Solution Behavior in the $\lambda \to 0$ Limit

In the $\lambda \to 0$ limit, STC binds at optimum because the condition (2) is trivially satisfied. Both p^{CC} and p^{AC} approach infinity in this limit, as stated in part (i) of Proposition 2; when equipment failures are extremely unlikely the supplier has little incentive to invest in capacity because the chance of his being penalized for poor service time realization is very small. To convince the supplier otherwise and to induce the target capacity μ_I, the customer must threaten him with a very high penalty rate. However, risk premiums under the two contracts converge to the same *finite* number $\frac{c\mu_I}{2}\left(\frac{1+v(\mu_I)^2}{1+\theta(\mu_I)}\right)$ in the $\lambda \to 0$ limit. In fact, this counterintuitive result is more general than what our model allows for, i.e., it continues to hold even if the failure process is not Poisson, as proved in the following proposition.

Proposition B.1 *Let* $\{Y_i\}$ *be arbitrary but i.i.d. random variables representing the failure interarrival time and* $F(\cdot \mid \lambda)$ *be their cdf when the arrival rate is* λ*. Suppose that* $F(\cdot \mid \lambda)$ *satisfies* $\lim_{\lambda \to 0} F(\cdot \mid \lambda) = 0$*. Let* ψ_Y^j*,* $j \in \{CC, AC\}$ *be the risk premium under either CC or AC. Then* $\lim_{\lambda \to 0} \psi_Y^{CC} = \lim_{\lambda \to 0} \psi_Y^{AC} = \frac{c\mu_I}{2}\left(\frac{1+v(\mu_I)^2}{1+\theta(\mu_I)}\right)$.

The intuition behind Proposition B.1 is as follows. In the vicinity of $\lambda = 0$, that is, when it is highly unlikely that an equipment failure occurs within the contracting period, the customer faces the following situation. Since the chance is high that the supplier's service is not required, even a fairly large penalty rate will not convince the supplier to invest in capacity. Therefore, the customer has to provide a very high contractual incentive (large p) in order to ensure that the supplier reserves the target capacity μ_I, as we showed above for Poisson failures. At the same time, however, uncertainty in the supplier's performance $\mathrm{Var}[X \mid \lambda, \mu_I]$, where $X = \sum_{i=1}^{N} S_i$ for CC and $X = \widehat{S}\mathbf{1}(N > 0)$ for AC, approaches zero since the supplier does not get a chance to reveal his ability to perform if there is no equipment failure. Since the risk premium combines these two effect, i.e., $\psi = \eta p^2 \mathrm{Var}[X \mid \lambda, \mu_I]$, a tension exists between p that goes to infinity and $\mathrm{Var}[X \mid \lambda, \mu_I]$ that goes to zero. Remarkably, Proposition B.1 states that a middle ground is chosen between these two opposing forces in the $\lambda \to 0$ limit, and that this asymptote depends neither on the supplier's risk aversion coefficient η or the equipment failure process.

To put this last result into a perspective, we compare it to the analysis in Abreu et al. (1991). One of the implications of the analysis in Abreu et al. (1991) is that it becomes infinitely expensive in the $\lambda \to 0$ limit to implement an incentive-compatible fixed-price contract in a repeated setting, which is known to allow for achieving the first-best solution if the discount rate is close to one.[9] Although a side-by-side comparison between our model and that of Abreu et al. (1991) is not possible because our model assumes a single interaction between the customer and the supplier, we find evidence from Proposition B.1 that PBC offers a unique advantage of containing cost even in extreme situations, as an upper bound on ψ exists. Given that repeated interactions can only improve the efficiency of a contract, as is well known in the contracting literature, this advantage over

[9] Rather than showing that supply chain cost approaches infinity, Abreu et al. (1991) show that an incentive-compatible fixed-price contract that satisfies a budget constraint does not exist if the agent's action is evaluated too frequently. Note that frequent action evaluation (i.e., short period length) in their model is equivalent to infrequent product failures in our model, in that they assume that the signal frequency is fixed, whereas we assume that it is the period length that is fixed.

a fixed-price contract would become even more pronounced in a setting with repeated interactions. Thus, this result advocates the use of a performance-based contract over a fixed-price contract (i.e., the contract that is independent of the performance outcome) in high-reliability environments, if outsourcing is required.

C Comparing Efficiencies of CC and AC When $v(\mu)$ Is Constant

What drives AC to be more efficient when the service time S does not vary too much? Why is CC more efficient in some cases? The key to answering these questions lies in examining in detail $V^{CC}(\lambda)$ and $V^{AC}(\lambda)$, which determine relative magnitudes of risk premiums (see Proposition 1 for expressions of the two quantities). These two quantities are in fact squares of the coefficients of variations (CV) of the two performance measures $\sum_{i=1}^{N} S_i$ and $\widehat{S}\mathbf{1}(N>0)$ (as opposed to v, which is the CV of S). It is instructive to write them in the following way:

$$V^{CC}(\lambda) = \frac{1}{\lambda} + \frac{1}{\lambda}v^2 \qquad \text{and} \qquad V^{AC}(\lambda) = \frac{e^{-\lambda}}{1-e^{-\lambda}} + \frac{\Delta(\lambda)}{1-e^{-\lambda}}v^2. \tag{6}$$

As we can see from these expressions, each $V^j(\lambda)$ is separated into terms that are either independent or dependent of v^2. The identities of the independent (first) terms are revealed by rewriting them as

$$\frac{1}{\lambda} = \frac{\lambda}{\lambda^2} = \frac{\mathrm{Var}[N]}{(E[N])^2} \qquad \text{and} \qquad \frac{e^{-\lambda}}{1-e^{-\lambda}} = \frac{e^{-\lambda}(1-e^{-\lambda})}{(1-e^{-\lambda})^2} = \frac{\mathrm{Var}[\mathbf{1}(N>0)]}{(E[\mathbf{1}(N>0)])^2}.$$

In other words, they are the squares of the CVs that originate from uncertainty in the number of equipment failures N. However, they manifest themselves in different forms for the two contracts: while it is the CV of N that enters into $V^{CC}(\lambda)$, it is the CV of $\mathbf{1}(N>0)$ that enters into $V^{AC}(\lambda)$. The latter comes from the no-failure effect of AC. $\mathrm{Var}[N]$ is present in CC because uncertainty in N is one of the two components of the total variance in cumulative downtime (see (8) in Appendix D), whereas under AC, $\mathrm{Var}[N]$ is eliminated through division of $\sum_{i=1}^{N} S_i$ by N, leaving the no-failure effect as the only residual of uncertainty from N. As intuition suggests, the variability of CC turns out to be greater than that of AC when only the first terms of $V^{CC}(\lambda)$ and $V^{AC}(\lambda)$ in (6) are compared: it can be shown that $1/\lambda - e^{-\lambda}/(1-e^{-\lambda}) > 0$.

This, however, does not tell the entire story because we have not taken into account the in-

teractions of N with S that are present in the second terms of $V^{CC}(\lambda)$ and $V^{AC}(\lambda)$ in (6). The interactions occur because variabilities in performance measures are also impacted by how many samples are collected, i.e., by N. It turns out that there is more variability in AC with regard to these interactions because division of $\sum_{i=1}^{N} S_i$ by a random variable N introduces more noise than when it is not divided by N, as is the case under CC. We confirm this insight by showing that the difference of the second terms in $V^{CC}(\lambda)$ and $V^{AC}(\lambda)$ is negative: $1/\lambda - \Delta(\lambda)/(1-e^{-\lambda}) < 0$, which follows from the property (iv) of $\Delta(\lambda)$ in Lemma D.1.

Combined, the sign of $V^{CC}(\lambda) - V^{AC}(\lambda)$ is ambiguous. However, we can infer from (6) and the preceding arguments that $V^{CC}(\lambda) > V^{AC}(\lambda)$ if v is sufficiently small but $V^{CC}(\lambda) < V^{AC}(\lambda)$ otherwise. This result answers the questions that we posed above, as the risk premium ψ^j, and hence the supply chain efficiency, is completely determined by $V^j(\lambda)$ in the constant $v(\mu)$ case: AC is more efficient when v is relatively small but the reverse is true if v is large. See Figure 3 that divides the (λ, v) space in terms of relative efficiency of the two contracts. A similar argument can be made for the case where $v(\mu)$ is allowed to vary, although it is more complicated than what we have presented here. The basic insight, however, remains the same.

D Proofs and Auxiliary Results

Proof of Lemma 1. The mean and the variance of a compound Poisson variable is evaluated as (see Ross 1996, pp. 82-89)

$$E[\textstyle\sum_{i=1}^{N} S_i \,|\, \lambda, \mu] = E[N \,|\, \lambda] E[S \,|\, \mu] = \lambda/\mu, \tag{7}$$

$$\mathrm{Var}[\textstyle\sum_{i=1}^{N} S_i \,|\, \lambda, \mu] = \mathrm{Var}[N \,|\, \lambda](E[S \,|\, \mu])^2 + E[N \,|\, \lambda]\mathrm{Var}[S \,|\, \mu] = \lambda\left(1 + v(\mu)^2\right)/\mu^2. \tag{8}$$

The supplier utility under CC (with $T = w - p\sum_{i=1}^{N} S_i$) is $u(\mu) = w - p\lambda/\mu - \eta p^2 \lambda\left(1 + v(\mu)^2\right)/\mu^2 - c(\mu - \underline{\mu})$. Differentiating,

$$u'(\mu) = p\lambda/\mu^2 + 2\eta p^2 \lambda\left(1 + \theta(\mu)\right)/\mu^3 - c.$$

Observe that

$$\theta'(\mu) = v(\mu)v'(\mu) - \mu[v'(\mu)]^2 - \mu v(\mu)v''(\mu) \leq 0.$$

Hence, $u''(\mu) < 0$, i.e., the supplier's utility maximization problem with CC is concave. The sensitivity analysis results can be found from implicit differentiations, which we omit. The same results for $u(\mu)$ can be shown analogously. ■

Proof of Proposition 1. We show the solution of the CC case only. The solution of the AC case is obtained similarly. The customer's contract design problem is reduced to the cost minimization problem

$$\min_p \ \Psi(p) \equiv r\lambda/\mu^* + c(\mu^* - \underline{\mu}) + \eta p^2 \lambda \left(1 + v(\mu^*)^2\right)/(\mu^*)^2 \ \text{ subject to } \mu^* \geq \mu_I$$

as the IR constraint $u(\mu^*) \geq 0$ binds at optimum, by an appropriate selection of w. We can invert μ^* found from the first-order condition in Lemma 1 with respect to p, using the monotonicity relation $\partial\mu^*/\partial p > 0$. Thus the optimal penalty rate that induces the supplier to choose μ is

$$p(\mu) = \frac{-1 + \sqrt{1 + 8\eta c\mu\left(1 + \theta(\mu)\right)/\lambda}}{4\eta\left(1 + \theta(\mu)\right)/\mu} = \frac{2c\mu^2}{\lambda\left(1 + \sqrt{1 + 8\eta c\mu\left(1 + \theta(\mu)\right)/\lambda}\right)}.$$

For notational convenience, let us suppress the argument μ in $p(\mu)$, $\theta(\mu)$, and $v(\mu)$. Observe that

$$\begin{aligned}\frac{\partial}{\partial\mu}\left(\frac{1+v^2}{(1+\theta)^2}\right) &= \frac{2vv'(1+\theta) - 2(1+v^2)\theta'}{(1+\theta)^3} = \frac{2vv'\left(1+v^2-\mu vv'\right) - 2(1+v^2)(vv' - \mu(v')^2 - \mu vv'')}{(1+\theta)^3} \\ &= \frac{2\mu(v')^2 + 2\mu v(1+v^2)v''}{(1+\theta)^3} \geq 0.\end{aligned}$$

Combining this result with (5), we find that the risk premium $\psi = \eta p^2 \lambda\left(1+v^2\right)/\mu^2$ (from the expression of $\Psi(p)$ above) is increasing in μ for $\mu \geq \underline{\mu}$:

$$\begin{aligned}\frac{16\eta}{\lambda}\psi'(\mu) &= 16\eta^2 \frac{d}{d\mu}\left((p/\mu)^2\left(1+v^2\right)\right) = \frac{d}{d\mu}\left(\left(-1 + \sqrt{1 + 8\eta c\mu\left(1+\theta\right)/\lambda}\right)^2 \frac{1+v^2}{(1+\theta)^2}\right) \\ &= \frac{8\eta c}{\lambda}\left(1 + \theta + \mu\theta'\right)\left(1 - \frac{1}{\sqrt{1 + 8\eta c\mu\left(1+\theta\right)/\lambda}}\right)\frac{1+v^2}{(1+\theta)^2} \\ &\quad + \left(-1 + \sqrt{1 + 8\eta c\mu\left(1+\theta\right)/\lambda}\right)^2 \frac{d}{d\mu}\left(\frac{1+v^2}{(1+\theta)^2}\right) \geq 0\,.\end{aligned}$$

In addition, $r\lambda/\mu + c(\mu - \underline{\mu})$ increases in $\mu \geq \mu_I$ when (2) is satisfied. Thus, the customer cost $\Psi = r\lambda/\mu + c(\mu - \underline{\mu}) + \psi$ increases in μ in the feasible region. Since our goal is to find the minimum

of Ψ, which is increasing in μ, in the feasible region $[\mu_I, \infty)$ on the μ-domain, the cost minimizer is found at the left-most boundary, i.e., $\mu = \mu_I$. Hence, the service time constraint binds at the optimal solution. The equilibrium solutions p^{CC} and ψ^{CC} are found by substituting $\mu = \mu_I$ in $p(\mu)$ and $\psi(\mu)$. ■

In the following auxiliary lemma, we evaluate the mean and the variance of the performance measure $\widehat{S}\mathbf{1}(N > 0)$ that are used to prove Lemma 2.

Lemma D.1

$$\begin{aligned} E[\widehat{S}\mathbf{1}(N>0) \,|\, \lambda, \mu] &= \Pr(N>0) E[S \,|\, \mu] = (1 - e^{-\lambda})/\mu \quad \textit{and} \qquad (9) \\ Var[\widehat{S}\mathbf{1}(N>0) \,|\, \lambda, \mu] &= \Pr(N>0) \left(\Pr(N=0)(E[S \,|\, \mu])^2 + \Delta(\lambda)\, Var[S \,|\, \mu] \right) \\ &= (1 - e^{-\lambda})[e^{-\lambda} + \Delta(\lambda) v(\mu)^2]/\mu^2, \qquad (10) \end{aligned}$$

where $\Delta(\lambda) \equiv \frac{1}{e^{\lambda}-1} \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} \frac{1}{n}$ *has the following properties: (i)* $\Delta'(\lambda) < 0$*, (ii)* $\lim_{\lambda \to 0} \Delta(\lambda) = 1$*, (iii)* $\lim_{\lambda \to \infty} \Delta(\lambda) = 0$*, (iv)* $\Delta(\lambda) > (1 - e^{-\lambda})/\lambda$*, and (v)* $\frac{d}{d\lambda} \left(e^{-\lambda}/\Delta(\lambda) \right) < 0$.

Proof of Lemma D.1. For notational convenience, let us suppress the conditional arguments (λ, μ). First, we prove the following intermediate results.

Lemma D.2

$$\begin{aligned} E[\widehat{S}] &= E[S] = 1/\mu, \qquad (11) \\ Var[\widehat{S}] &= \Delta(\lambda)\, Var[S] = \Delta(\lambda) v(\mu)^2/\mu^2, \qquad (12) \end{aligned}$$

Proof. Let $M(t) \equiv E[e^{t\widehat{S}}]$ be the moment generating function for $\widehat{S}$. Then

$$\begin{aligned} M(t) &= E[e^{t(\sum_{i=1}^{N} S_i)/N} \,|\, N > 0] = \frac{1}{\Pr(N>0)} \sum_{n=1}^{\infty} E[e^{t(\sum_{i=1}^{N} S_i)/N} \,|\, N = n] \Pr(N = n) \\ &= \frac{1}{1 - e^{-\lambda}} \sum_{n=1}^{\infty} E\left[e^{t(\sum_{i=1}^{n} S_i)/n} \right] \frac{\lambda^n e^{-\lambda}}{n!} = \frac{1}{1 - e^{-\lambda}} \sum_{n=1}^{\infty} \left(E[e^{tS_i/n}] \right)^n \frac{\lambda^n e^{-\lambda}}{n!}, \end{aligned}$$

where the last equality follows from independence of $\{S_i\}$. Differentiating,

$$\begin{aligned}
M'(t) &= \frac{1}{1-e^{-\lambda}} \sum_{n=1}^{\infty} n \left(E[e^{tS_i/n}] \right)^{n-1} E\left[\frac{S_i}{n} e^{tS_i/n} \right] \frac{\lambda^n e^{-\lambda}}{n!}, \\
M''(t) &= \frac{1}{1-e^{-\lambda}} \sum_{n=1}^{\infty} n(n-1) \left(E[e^{tS_i/n}] \right)^{n-2} \left(E\left[\frac{S_i}{n} e^{tS_i/n} \right] \right)^2 \frac{\lambda^n e^{-\lambda}}{n!} \\
&\quad + \frac{1}{1-e^{-\lambda}} \sum_{n=1}^{\infty} n \left(E[e^{tS_i/n}] \right)^{n-1} E\left[\frac{S_i^2}{n^2} e^{tS_i/n} \right] \frac{\lambda^n e^{-\lambda}}{n!}
\end{aligned}$$

The first and second moments are

$$\begin{aligned}
E[\widehat{S}] &= M'(0) = \frac{E[S]}{1-e^{-\lambda}} \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} = E[S], \\
E[\widehat{S}^2] &= M''(0) = \frac{(E[S])^2}{e^{\lambda}-1} \sum_{n=1}^{\infty} \left(1 - \frac{1}{n} \right) \frac{\lambda^n}{n!} + \frac{E[S^2]}{e^{\lambda}-1} \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} \frac{1}{n} = (E[S])^2 + \frac{\mathrm{Var}[S]}{e^{\lambda}-1} \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} \frac{1}{n},
\end{aligned}$$

which together yield $\mathrm{Var}[\widehat{S}] = E[\widehat{S}^2] - (E[\widehat{S}])^2 = \left(\frac{1}{e^{\lambda}-1} \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} \frac{1}{n} \right) \mathrm{Var}[S]$. ■

Next, we prove (9) and (10). Let $I \equiv \mathbf{1}(N > 0)$. Note that, using (11) and (12), $E[\widehat{S}I \,|\, I = 0] = 0$, $E[\widehat{S}I \,|\, I = 1] = E[\widehat{S}] = E[S]$, $\mathrm{Var}[\widehat{S}I \,|\, I = 0] = 0$, and $\mathrm{Var}[\widehat{S}I \,|\, I = 1] = \mathrm{Var}[\widehat{S}] = \Delta(\lambda)\mathrm{Var}[S]$. The mean is

$$E[\widehat{S}I] = E[E[\widehat{S}I \,|\, I]] = \Pr(I = 1) E[S].$$

To compute the variance, first observe that

$$\begin{aligned}
\mathrm{Var}[E[\widehat{S}I \,|\, I]] &= E[E[\widehat{S}I \,|\, I]^2] - (E[E[\widehat{S}I \,|\, I]])^2 = E[E[\widehat{S}I \,|\, I]^2] - (E[\widehat{S}I])^2 \\
&= \Pr(I = 1)(E[\widehat{S}I \,|\, I = 1])^2 + \Pr(I = 0)(E[\widehat{S}I \,|\, I = 0])^2 - (\Pr(I = 1) E[S])^2 \\
&= \Pr(I = 1)(E[S])^2 - (\Pr(I = 1))^2 (E[S])^2 = \Pr(I = 0) \Pr(I = 1)(E[S])^2,
\end{aligned}$$

where we have used the results obtained above. Therefore,

$$\begin{aligned}
\mathrm{Var}[\widehat{S}I] &= \mathrm{Var}[E[\widehat{S}I \,|\, I]] + E[\mathrm{Var}[\widehat{S}I \,|\, I]] \\
&= \Pr(I = 0) \Pr(I = 1)(E[S])^2 + \Pr(I = 1)\Delta(\lambda)\mathrm{Var}[S] \\
&= \Pr(I = 1) \left(\Pr(I = 0)(E[S])^2 + \Delta(\lambda)\mathrm{Var}[S] \right).
\end{aligned}$$

Finally, we show the properties of $\Delta(\lambda)$. The following facts are useful:

$$\begin{aligned}
\Delta'(\lambda) &= -\frac{e^\lambda}{(e^\lambda-1)^2}\sum_{n=1}^{\infty}\left(\frac{\lambda^n}{n!}\frac{1}{n}\right)+\frac{1}{e^\lambda-1}\sum_{n=1}^{\infty}\left(\frac{\lambda^{n-1}}{n!}\right) \\
&= -\frac{e^\lambda}{(e^\lambda-1)^2}\sum_{n=1}^{\infty}\left(\frac{\lambda^n}{n!}\frac{1}{n}\right)+\frac{1}{e^\lambda-1}\frac{1}{\lambda}\sum_{n=1}^{\infty}\left(\frac{\lambda^n}{n!}\right) = -\frac{e^\lambda}{(e^\lambda-1)^2}\sum_{n=1}^{\infty}\left(\frac{\lambda^n}{n!}\frac{1}{n}\right)+\frac{1}{\lambda} \\
&= -\frac{\Delta(\lambda)}{1-e^{-\lambda}}+\frac{1}{\lambda}, \qquad (13)
\end{aligned}$$

$$\begin{aligned}
\sum_{n=1}^{\infty}\frac{\lambda^n}{n!}\frac{1}{n} &= \lambda+\sum_{n=2}^{\infty}\frac{\lambda^n}{n!}\frac{1}{n} \\
&> \lambda+\sum_{n=2}^{\infty}\frac{\lambda^n}{(n+1)!} = \lambda+\frac{1}{\lambda}\sum_{n=3}^{\infty}\frac{\lambda^n}{n!} \\
&= \lambda+\frac{1}{\lambda}\left(e^\lambda-1-\lambda-\frac{\lambda^2}{2}\right) = \frac{1}{\lambda}\left(e^\lambda-1-\lambda+\frac{\lambda^2}{2}\right), \qquad (14)
\end{aligned}$$

and

$$e^{-\lambda} \le 1-\lambda+\lambda^2/2, \qquad (15)$$

which can be shown as follows. Let $\xi(\lambda) \equiv 1-\lambda+\lambda^2/2-e^{-\lambda}$. Then $\xi'(\lambda) = -1+\lambda+e^{-\lambda}$ and $\xi''(\lambda) = 1-e^{-\lambda}$. Since $\xi'(0)=0$ and $\xi''(\lambda) \ge 0$, we have $\xi'(\lambda) \ge 0$. But this implies $\xi(\lambda) \ge 0$ since $\xi'(0)=0$.

(i) Differentiating $\Delta(\lambda)$, we obtain

$$\begin{aligned}
\Delta'(\lambda) &= -\frac{e^\lambda}{(e^\lambda-1)^2}\sum_{n=1}^{\infty}\left(\frac{\lambda^n}{n!}\frac{1}{n}\right)+\frac{1}{\lambda} \\
&< -\frac{e^{2\lambda}-e^\lambda-\lambda e^\lambda+\lambda^2e^\lambda/2}{\lambda(e^\lambda-1)^2}+\frac{1}{\lambda} = -\frac{e^{2\lambda}-e^\lambda-\lambda e^\lambda+\lambda^2e^\lambda/2-e^{2\lambda}+2e^\lambda-1}{\lambda(e^\lambda-1)^2} \\
&= -\frac{e^\lambda-\lambda e^\lambda+\lambda^2e^\lambda/2-1}{\lambda(e^\lambda-1)^2} = -\frac{e^\lambda(1-\lambda+\lambda^2/2-e^{-\lambda})}{\lambda(e^\lambda-1)^2} \le 0,
\end{aligned}$$

where the first and second inequalities follow from (14) and (15), respectively.

(ii) By l'Hopital's rule,

$$\lim_{\lambda\to 0}\Delta(\lambda) = \lim_{\lambda\to 0}\frac{\sum_{n=1}^{\infty}\frac{\lambda^n}{n!}\frac{1}{n}}{e^\lambda-1} = \lim_{\lambda\to 0}\frac{\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{n!}}{e^\lambda} = \lim_{\lambda\to 0}\frac{1+\sum_{n=2}^{\infty}\frac{\lambda^{n-1}}{n!}}{e^\lambda} = 1.$$

(iii) Applying l'Hopital's rule n times,

$$\lim_{\lambda\to\infty}\Delta(\lambda)=\sum_{n=1}^{\infty}\frac{1}{n!}\frac{1}{n}\left(\lim_{\lambda\to\infty}\frac{\lambda^n}{e^{\lambda}-1}\right)=\sum_{n=1}^{\infty}\frac{1}{n!}\frac{1}{n}\left(\lim_{\lambda\to\infty}\frac{n!}{e^{\lambda}}\right)=\sum_{n=1}^{\infty}\frac{1}{n}\left(\lim_{\lambda\to\infty}\frac{1}{e^{\lambda}}\right)=0.$$

(iv) The lower bound of $\Delta(\lambda)$ follows from (13) and part (i).

(v) Since

$$\frac{\Delta(\lambda)}{e^{-\lambda}}=\frac{1}{1-e^{-\lambda}}\sum_{n=1}^{\infty}\frac{\lambda^n}{n!}\frac{1}{n},$$

we see that

$$\begin{aligned}\frac{\partial}{\partial\lambda}\left(\frac{\Delta(\lambda)}{e^{-\lambda}}\right) &= \frac{1}{(1-e^{-\lambda})^2}\left((1-e^{-\lambda})\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{n!}-e^{-\lambda}\sum_{n=1}^{\infty}\frac{\lambda^n}{n!}\frac{1}{n}\right)\\ &\geq \frac{1}{(1-e^{-\lambda})^2}\left((1-e^{-\lambda})\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{n!}-e^{-\lambda}\sum_{n=1}^{\infty}\frac{\lambda^n}{n!}\right)\\ &= \frac{1}{(1-e^{-\lambda})}\left(\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{n!}-1\right)=\frac{\left(e^{\lambda}-1\right)/\lambda-1}{1-e^{-\lambda}}>0,\end{aligned}$$

where the last inequality comes from $e^{\lambda}-1>\lambda$ for $\lambda>0$.

■

Proof of Lemma 2. The proofs for all results in the lemma are analogous to those of Lemma 1, except for the last result concerning the sign of $\partial\mu^*/\partial\lambda$. Suppose that λ is close to zero such that terms of order λ^2 and above can be dropped. Then the first-order condition is approximated as

$$c=\frac{p}{\mu^2}(1-e^{-\lambda})+\frac{2\eta p^2}{\mu^3}\left(e^{-\lambda}+\Delta(\lambda)\theta(\mu)\right)(1-e^{-\lambda})\approx\frac{p}{\mu^2}\lambda+\frac{2\eta p^2}{\mu^3}\left(1+\theta(\mu)\right)\lambda,$$

where we have used (ii) of Lemma D.1. This result identical to the first-order condition in Lemma 1 for CC. Hence, $\partial\mu^*/\partial\lambda>0$ for small λ. On the other hand, if λ is sufficiently large so that $e^{-\lambda}/\Delta(\lambda)\approx 0$ (see (v) of Lemma D.1 and Table 1), $1-e^{-\lambda}\approx 1$ and the first-order condition becomes

$$c=\frac{p}{\mu^2}(1-e^{-\lambda})+\frac{2\eta p^2}{\mu^3}\left(e^{-\lambda}+\Delta(\lambda)\theta(\mu)\right)(1-e^{-\lambda})\approx\frac{p}{\mu^2}+\frac{2\eta p^2}{\mu^3}\Delta(\lambda)\theta(\mu).$$

Since $\Delta'(\lambda) < 0$ by (i) of Lemma D.1, it is clear from this expression that $\partial\mu^*/\partial\lambda < 0$. ■

Proof of Proposition 2.

(i) Rewriting p^{CC} and p^{AC} derived in Proposition 1,

$$p^{CC} = 2c\mu_I^2 \left(\lambda + \sqrt{\lambda^2 + 8\eta c\mu_I \left(1+\theta(\mu_I)\right)\lambda}\right)^{-1} \text{ and} \tag{16}$$

$$p^{AC} = 2c\mu_I^2 \left((1-e^{-\lambda}) + \sqrt{(1-e^{-\lambda})^2 + 8\eta c\mu_I \left(e^{-\lambda} + \Delta(\lambda)\theta(\mu_I)\right)(1-e^{-\lambda})}\right)^{-1}. \tag{17}$$

$\partial p^{CC}/\partial\lambda < 0$ is clear from (16). Since performance measures under CC and AC converge near $\lambda = 0$ (see the discussion below Lemma 2), $p^{AC} \to p^{CC}$ as $\lambda \to 0$, so $\lim_{\lambda\to 0} \partial p^{AC}/\partial\lambda < 0$. On the other hand, if λ is sufficiently large so that $e^{-\lambda}/\Delta(\lambda) \approx 0$ (see Table 1), (17) can be approximated as $p^{AC} \approx 2c\mu_I^2 \left(1 + \sqrt{1 + 8\eta c\mu_I \Delta(\lambda)\theta(\mu_I)}\right)^{-1}$, from which we find that $\partial p^{AC}/\partial\lambda > 0$ since $\Delta'(\lambda) < 0$ by (i) of Lemma D.1. To show $\lim_{\lambda\to 0} p^{CC} = \lim_{\lambda\to 0} p^{AC} = \infty$, notice that term-by-term comparison of the denominators of (16) and (17) reveals that $p^{CC} < p^{AC}$, since $1 - e^{-\lambda} < \lambda$, $e^{-\lambda} < 1$, and $\Delta(\lambda) < 1$. Retaining only the terms up to $O(\lambda)$, we see that $p^{AC} \to p^{CC}$ in the $\lambda \to 0$ limit. Moreover, $\lim_{\lambda\to 0} p^{CC} = \infty$ is clear from (16).

(ii) $\partial\psi^{CC}/\partial\lambda < 0$ is clear from the expression of ψ^{CC} in Proposition 1. To show $\partial\psi^{AC}/\partial\lambda < 0$, let $\varphi(\lambda) \equiv \frac{e^{-\lambda}+\Delta(\lambda)v(\mu_I)^2}{e^{-\lambda}+\Delta(\lambda)\theta(\mu_I)}$ be the multiplicative factor that appears in ψ^{AC} (see the same proposition). Define $\chi \equiv -\mu_I v(\mu_I) v'(\mu_I) \geq 0$. Note that

$$\varphi'(\lambda) = \frac{d}{d\lambda}\left(1 + \frac{\chi}{e^{-\lambda}/\Delta(\lambda) + v(\mu_I)^2}\right)^{-1} < 0,$$

by the property (v) of Lemma D.1. Using this and $\Delta'(\lambda) < 0$, $\partial\psi^{AC}/\partial\lambda < 0$ follows. Notice that $\lim_{\lambda\to 0} \varphi(\lambda) = \varphi_c$, where $\varphi_c \equiv \left(1 + v(\mu_I)^2\right)/\left(1 + \theta(\mu_I)\right)$ is the multiplicative factor that appears in ψ^{CC}. Together with $\varphi'(\lambda) < 0$, this implies $\varphi(\lambda) < \varphi_c$. With the latter, the stated condition $V^{CC}(\lambda) \geq V^{AC}(\lambda)$, or $\left(1 + \theta(\mu_I)\right)/\lambda \geq \left(e^{-\lambda} + \Delta(\lambda)\theta(\mu_I)\right)/(1 - e^{-\lambda})$, implies $\psi^{CC} \geq \psi^{AC}$, as can be verified from their respective expressions. The $\lambda \to 0$ limit is immediate from the same expressions.

■

Proof of Lemma 3. For notational convenience, let $u^* \equiv u(\mu^*)$, $u_t^* \equiv u(\mu_t^*)$, and $U_t^* \equiv U_t(\mu_t^*)$.

(i) Under CC or under AC with $\lambda_t \sim 0$, the first-order condition for μ_t corresponding to λ_t is (or is approximated as, in the case of AC, since AC mimics CC if $\lambda_t \sim 0$)

$$\frac{p}{\mu_t^2} + \frac{2\eta p^2}{\mu_t^3}\left(1+v^2\right) = \frac{c}{\lambda_t},$$

from Lemma 1. For each $t \in \{L, H\}$, the supplier's utility at μ_t^* that solves this optimality condition is $U_t^* = w - p\lambda_t/\mu_t^* - \eta p^2\left(1+v^2\right)\lambda_t/(\mu_t^*)^2 - c(\mu_t^* - \underline{\mu}) - K\mathbf{1}(t=L)$. Their difference is $U_H^* - U_L^* = -\zeta(p) + K$, where

$$\zeta(p) \equiv p\left(\frac{\lambda_H}{\mu_H^*} - \frac{\lambda_L}{\mu_L^*}\right) + \eta p^2\left(1+v^2\right)\left(\frac{\lambda_H}{(\mu_H^*)^2} - \frac{\lambda_L}{(\mu_L^*)^2}\right) + c(\mu_H^* - \mu_L^*).$$

Hence, $U_H^* \geq U_L^*$, i.e., the supplier chooses λ_H, if and only if $K \geq \zeta(p)$. Note that, from the first-order condition above for a generic λ,

$$\frac{\lambda}{\mu^*} = c\left(\frac{p}{\mu^*} + \frac{2\eta p^2}{(\mu^*)^2}\left(1+v^2\right)\right)^{-1} \text{ and } \frac{\lambda}{(\mu^*)^2} = c\left(p + \frac{2\eta p^2}{\mu^*}\left(1+v^2\right)\right)^{-1},$$

both of which are increasing in μ^* for fixed p. On the other hand, $\mu_L^* < \mu_H^*$ since $\partial\mu^*/\partial\lambda > 0$ for fixed p, according to Lemma 1. Therefore, we have $\lambda_L/\mu_L^* < \lambda_H/\mu_H^*$ and $\lambda_L/(\mu_L^*)^2 < \lambda_H/(\mu_H^*)^2$ and conclude that $\zeta(p) > 0$. Using these results and applying the envelope theorem, we have $\frac{d}{dp}\left(U_H^* - U_L^*\right) = \frac{\partial}{\partial p}\left(U_H^* - U_L^*\right) = -\frac{\partial}{\partial p}\zeta(p) = -\left(\frac{\lambda_H}{\mu_H^*} - \frac{\lambda_L}{\mu_L^*}\right) - 2\eta p\left(1+v^2\right)\left(\frac{\lambda_H}{(\mu_H^*)^2} - \frac{\lambda_L}{(\mu_L^*)^2}\right) < 0$. Because of this monotonicity, $U_H^* - U_L^*$ crosses zero at most once, i.e., $p^\dagger$ that satisfies $\zeta(p^\dagger) = K$ is unique if it exists. Suppose that $K < \zeta(\underline{\underline{p}})$. Then $U_H^* - U_L^*$ starts from a negative value at $p = \underline{\underline{p}}$ and becomes more negative as p increases. Hence, the supplier always chooses λ_L in this case (the statement in (i) of the lemma is true by setting $p^\dagger = \underline{\underline{p}}$). On the other hand, if $K \geq \zeta(\underline{\underline{p}})$, there may be a value (which we have shown to be unique) $p^\dagger \geq \underline{\underline{p}}$ for which $u_H^* - u_L^*$ crosses zero from positive to negative. In this case, the supplier chooses λ_H if $p \leq p^\dagger$ and λ_L if $p > p^\dagger$.

(ii) Fix p and λ. Under AC with λ sufficiently large for which $e^{-\lambda}/\Delta(\lambda) \approx 0$, the supplier's utility at μ^*, which satisfies the first-order condition in Lemma 2, is approximated as $u^* \approx$

$w - p/\mu^* - \eta p^2 \Delta(\lambda) v^2/(\mu^*)^2 - c(\mu^* - \underline{\mu})$. By the envelope theorem, $du^*/d\lambda = \partial u^*/\partial\lambda \approx -\eta p^2 \Delta'(\lambda) v^2/(\mu^*)^2 > 0$, implying that $u_H^* - u_L^*$ increases as the distance between λ_L and λ_H becomes larger. Since $U_t^* = u_t^* - K\mathbf{1}(t = L)$ and $u_H^* - u_L^* \to 0$ as $\lambda_H - \lambda_L \to 0$, $U_H^* - U_L^* \to K \geq 0$ as $\lambda_H - \lambda_L \to 0$ while $U_H^* - U_L^*$ increases as $\lambda_H - \lambda_L$ becomes larger. In other words, $U_H^* - U_L^* > K \geq 0$ for any $\lambda_L < \lambda_H$ regardless of p. Therefore, the supplier always chooses λ_H.

■

Proof of Proposition B.1. The CC and AC converge to one another when $\lambda \sim 0$ since their performance measures become indistinguishable, as $\sum_{i=1}^{N} S_i \sim S_1\mathbf{1}(N = 1)$ and $\widehat{S}\mathbf{1}(N > 1) \sim S_1\mathbf{1}(N = 1)$. Let us consider AC with $\lambda \sim 0$, for which $T_a \approx w - pS_1\mathbf{1}(N = 1)$. Since the period length is normalized to one, $\mathbf{1}(N = 1) = \mathbf{1}(Y_1 < 1)$. Thus, $T \approx w - pS_1\mathbf{1}(Y_1 < 1)$. By the law of total variance (proof is similar to that of Lemma D.1), it can be shown that

$$\begin{aligned} E[T \,|\, \lambda, \mu] &\approx w - pF(1 \,|\, \lambda)E[S \,|\, \mu], \\ \mathrm{Var}[T \,|\, \lambda, \mu] &\approx p^2 F(1 \,|\, \lambda)\left([1 - F(1 \,|\, \lambda)](E[S \,|\, \mu])^2 + \mathrm{Var}[S \,|\, \mu]\right) \approx p^2 F(1 \,|\, \lambda)\left((E[S \,|\, \mu])^2 + \mathrm{Var}[S \,|\, \mu]\right), \end{aligned}$$

where the last approximation is valid since $[F(1 \,|\, \lambda)]^2$ is negligible when $\lambda \sim 0$. Compare these expressions to their Poisson counterparts for λ around zero (where the terms of order only up to $O(\lambda)$ in (9) and (10) are retained):

$$\begin{aligned} E[T \,|\, \lambda, \mu] &= w - p(1 - e^{-\lambda})E[S \,|\, \mu] \approx w - p\lambda E[S \,|\, \mu], \\ \mathrm{Var}[T \,|\, \lambda, \mu] &= p^2(1 - e^{-\lambda})\left(e^{-\lambda}(E[S \,|\, \mu])^2 + \Delta(\lambda)\mathrm{Var}[S \,|\, \mu]\right) \approx p^2\lambda\left((E[S \,|\, \mu])^2 + \mathrm{Var}[S \,|\, \mu]\right). \end{aligned}$$

We see that they have the same forms, the only difference being that $F(1 \,|\, \lambda)$ is substituted by λ. Hence, the analysis of AC with Y is equivalent to that with the Poisson failures. Therefore, when $\lambda \sim 0$,

$$\psi_Y^{AC} \approx \frac{c\mu_I}{2}\left(\frac{1 + v(\mu_I)^2}{1 + \theta(\mu_I)}\right)\left(1 - 2\left(1 + \sqrt{1 + 8\eta c\mu_I \frac{1 + \theta(\mu_I)}{F(1 \,|\, \lambda)}}\right)^{-1}\right),$$

which is obtained from ψ^{AC} in Proposition 1 with $1 - e^{-\lambda} \approx \lambda$ replaced by $F(1 \,|\, \lambda)$ and $e^{-\lambda} \to 1$

and $\Delta(\lambda) \to 1$. $\lim_{\lambda \to 0} \psi_Y^{AC} = \frac{c\mu_I}{2} \left(\frac{1+v(\mu_I)^2}{1+\theta(\mu_I)} \right)$ follows after letting $\lambda \to 0$, since $\lim_{\lambda \to 0} F(\cdot \,|\, \lambda) = 0$ by assumption. ■