The Production of Service: A Workload View of Complementarity and Substitution

Noa Zychlinski[1], Itai Gurvich[2]
[1] Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa 3200003, Israel
[2] Kellogg School of Management, Northwestern University, 2211 Campus Dr, Evanston, IL 60208
noazy@technion.ac.il, i-gurvich@kellogg.northwestern.edu

We study optimal coordination in quality-driven re-work systems, where each station contributes to an item's quality and the likelihood of re-work depends on these contributions through a production function. Our model links quality targets to processing times via Brownian hitting times and aggregates quality into re-work probability.

By setting quality targets, a planner jointly shapes processing times and re-work rates, determining overall workload. We characterize the attainable workload region and use it to minimize effort costs in two-station networks. The optimal policy depends on (i) complementarity versus substitutability of quality contributions, (ii) each station's quality impact, and (iii) relative operational costs.

We identify three operating regimes: two corner regimes (one active station) and a network regime (both active). As complementarity increases, effort shifts toward the more cost-effective station, while finite capacity modifies this threshold, pushing effort toward the less constrained station. The analysis highlights how complementarity shapes efficiency trade-offs and capacity requirements.

To illustrate the model's practical relevance, we calibrate it to industrial maintenance operations in which equipment undergoes repair followed by preventive service while offline. The results show that the optimal design is robust, efficient, and estimable from operational data.

Key words: Stochastic modeling, queueing systems, processing networks, station integration

1. Introduction

Workflows are a result of a design process. The design determines the activities or tasks to be performed and the effort exerted in each one. Once the workflow is set, one can turn to questions that are central to queueing theory, such as the optimization of waiting time. This paper concerns the design stage.

Our premise is that processing time and/or effort affects the quality of the outcome. A longer processing time and better quality outputs require more capacity; however, they are likely to reduce re-work, making the overall effect on workload more ambiguous.

Re-work imposes substantial financial and operational burdens across sectors. In manufacturing and other production systems, analyses of the "cost of quality" indicate that

5–15% of sales are absorbed by quality-related costs, with internal failure costs—including re-work—representing a major share (Schiffauerova and Thomson 2006). In asset-intensive industries such as construction and maintenance, direct re-work costs typically account for 4–6% of total project expenditures and can reach 12–15% in complex or poorly coordinated projects, contributing to delays and cost overruns (Hwang et al. 2009, Love and Li 2000). In healthcare, similar challenges arise: in the U.S., hospital readmissions number about 3.8 million annually (13% of adult discharges) (Jiang and Hensche 2023), with an average cost of $16,037 per readmission (Kum G. et al. 2024). Preventable readmissions alone cost the U.S. healthcare system tens of billions of dollars each year—about $17 billion in Medicare acute-care payments (Medicare and Medicaid Statistical Supplement 2007). These figures show that re-work is widespread and costly, highlighting the importance of understanding how design choices and effort allocation affect performance.

The output quality of a service or production process depends on an aggregation of efforts throughout the process. The optimal coordination of efforts—for example, the one that minimizes effort or capacity costs—depends on the efficiency of the different steps (the capacity needed per improvement in the step's output quality) and their effectiveness (the impact of quality improvement in an individual step on the overall process output).

The ideas we introduce in this paper, and the questions we address, are relevant to various processing networks where there is a clear notion of repair and maintenance activities. In such systems, the planner must decide how much effort to allocate to each activity (e.g., preventative maintenance). In production systems, software development, and project management, a processing step is often followed by a quality assurance (QA) step. The more effort allocated to the processing step, the better the quality of the product at transfer. The operator must distribute the efforts among the stations by considering the properties (cost, capacity, etc.) of the different steps in the process.

To study these questions we develop a model that stipulates

(A) a relationship between a targeted quality of output and the processing time, and

(B) a relationship between the quality of output—a vector that includes quality measures for each step of the process—and the likelihood of re-work, hence the total arrival rate.

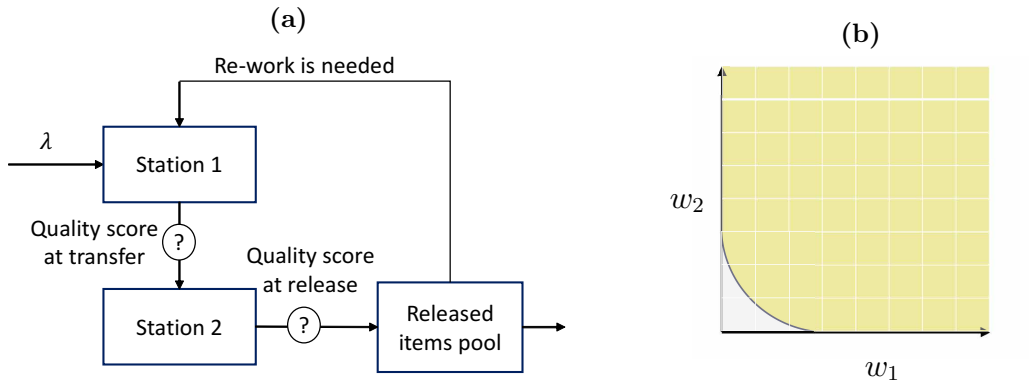
We use a threshold model: An item is released (or transferred to the next step in the process) when its state reaches a threshold. We capture an item's state via a quality "score"

that serves as an aggregation of different quality measurements. The individual score evolution in each step is described as a stochastic process. The properties we focus on are hitting time statistics—averages and probabilities—that determine processing times in each station. The threshold—the score at which the item is released or transferred—is the design decision. Once set, it determines the processing time distribution. This distribution—or, distributions in plural when two stations operate—affects the arrival rate.

To introduce the baseline setting, consider the case where arrivals and service rates are exogenous and there is no re-work. The planner decides how to distribute the total workload $w = (w_1, w_2)$ between two processing resources. Let the average service time be normalized to one hour, and let the hourly arrival rate be λ; the incoming work is therefore λ hours of work per hour. Feasibility requires that all incoming work be processed, namely $w_1 + w_2 \geq \lambda$, which yields a linear attainable workload region.

In our model, illustrated in Figure 1(a), the arrival rate depends endogenously on the stations' effort levels through the re-work mechanism, that is, $\lambda = \lambda(w)$. The presence of re-work makes the attainable workload region nonlinear. This nonlinearity reflects the endogenous relationship between effort and workload and forms the basis for the optimization problems analyzed in the paper.

Figure 1 **A two-station model with re-work (a) and the corresponding attainable workload region (b).**



While there is an underlying dynamic evolution model leading to our probability of re-work, the final outcome is a simple and practical relationship:

$$\log(\text{probability of re-work}) = -\text{Aggregated Quality}.$$

We model the aggregated quality as a function that combines the quality contributions of the different stations into a single measure, which in turn determines the probability of re-work.

The aggregated-quality model captures the interaction between stations through their degree of *substitution* or *complementarity*. Intuitively, when stations are substitutes, additional effort in one can compensate for less effort in the other, whereas when they are complements, the effectiveness of one station increases with the effort devoted to the other.

We formalize these interactions in Section 4 using the *constant-elasticity-of-substitution* (CES) family, a well-established framework in the economics literature for modeling the interplay between production factors (Carvalho and Tahbaz-Salehi 2019). By varying the CES parameters, we can flexibly represent different degrees of substitutability or complementarity between stations and study how these relationships affect optimal system design.

To the best of our knowledge, this is the first application of a CES formulation in a service-operations setting with quality-driven re-work. Existing operations studies typically assume additive or separable relationships between process stages, where improvements in one step do not directly affect the marginal effectiveness of another. In contrast, the CES structure allows us to *quantitatively vary* the degree of cross-station substitution and complementarity, capturing nonlinear interdependencies that prior models abstract away. This formulation bridges economic production theory with operational design, providing a unified analytical tool for studying how interdependencies across stages shape both workload and efficiency.

Unlike in economics, where the CES function represents how inputs such as labor and capital combine to produce output and revenue, our operational setting *endogenizes* these "market" effects: the quality targets chosen at each station influence both processing times and the probability of re-work, jointly determining total workload and cost. The CES specification is therefore the right lens here—it captures the nonlinear way in which joint effort across service stages translates into effective quality, while remaining analytically tractable and interpretable. In this way, output quality—and hence the trade-off between effort and workload—is fully internalized in our model.

Maintenance Example. The concepts of substitution and complementarity arise naturally in maintenance and reliability settings. Consider two sequential activities within a maintenance process: *repair* followed by *preventive service*. Repair represents corrective actions that restore the asset to an operational state—such as replacing a failed component, realigning a gearbox, or fixing an electrical malfunction. Preventive service then takes place while

the equipment remains offline and involves inspections, lubrication, recalibration, and other conditioning tasks that enhance long-term reliability and reduce the likelihood of future failures. If these activities are substitutes, more extensive repair can compensate for less preventive service effort, while thorough post-repair maintenance may delay the need for the next repair intervention. If they are complements, the effectiveness of one depends on the other: a well-executed repair increases the impact of subsequent preventive service, whereas insufficient repair limits the benefit of additional preventive effort. We return to this example throughout the paper—first to illustrate the theoretical mechanisms and later to demonstrate how the model can be calibrated and interpreted.

We establish the following:

1. **Model and workload characterization.** We build our model on components (A) and (B) above and show how the decisions—the targeted quality levels—can be mapped to workload decisions. We then show that the workload feasibility set is convex; its characterization highlights the effects of complementarity on its structure (Figure 3).

   We prove that the higher the complementarity coefficient the smaller the set of attainable workloads. This implies, in particular, that the larger the complementarity, the more constrained is any cost minimization problem.

2. **The structure of optimal designs.** We show that when minimizing processing costs, corner solutions (where one station has zero effort and the other makes all the effort) are never optimal as long as there exists some complementarity. Systems with perfect substitution generally experience *three operating regimes*: two where only one station works, and one, where the solution is interior, and both stations are used.

   This structure stands in contrast to the well-studied problem of maximizing CES utilities under budget constraints (see, e.g., Mas-Colell et al. 1995, Exercise 3.C.6; Varian 2010, Chapter 4). There, with perfect substitution in the utility function, one has only corner solutions. Because the re-work probability is a non-linear function of the process quality, the solution space in our model is richer.

   We show that whether the optimal station efforts are increasing or decreasing in the level of substitution depends on where the cost stands relative to a "symmetry" cost threshold. This threshold does not itself depend on the level of substitution.

3. **The capacitated explicit solution.** The characterization of the optimal solution when capacity is constrained has an informative structure—the effect of complementarity on the optimal effort mix is mediated through the capacity levels. At certain capacity levels, the "shadow-price" of the capacity level is increasing in the complementarity and for others it decreases. When there are multiple types of items that "compete" for the limited capacity, the optimal mixture of workload in each station changes in non-obvious ways as capacity changes.

In Appendix B, we present three variants of the base model that demonstrate the flexibility and robustness of our results. These include *processing complementarity* (Appendix B.2), where the effectiveness—the quality improvement rate—of the second station depends on the final score from the first station; *re-work within a finite time window* (Appendix B.3); and *quadratic cost structures* (Appendix B.4). We further show that these variants, as well as resource-sharing and throughput-maximization problems, can be formulated and solved within the attainable workload framework.

This paper does not aim to be the final word or embody a comprehensive framework for arbitrary networks. Rather, it is intended to advance key notions and initial results, and showcase their usefulness through the insights they provide.

Organization. We start with a brief review of the relevant literature in §2. The single-station model described in §3 serves as an important building block. We introduce substitution/complementarity in the two-station model in §4. This section contains the key results of the paper as they pertain to the characterization of the attainable workload region. §5 draws on this characterization to study process optimization. §6 contains a concentrated discussion of the operational/managerial implications of our results. In §7 we provide some concluding remarks and suggest a few directions for future research. All proofs, model variants, parameter estimation procedures, and calibration details appear in the appendix.

2. Literature Review

Coordinating and balancing the optimal effort of the two stations is an instance of determining the *anatomy* of the service: defining its parts and their interaction. The question is one of design. Rather than taking the service content at each step as fixed, we optimize it to meet network-level goals. The anatomy of a service refers to the collection of its components, the content of each, and their interaction.

Our work speaks to the operations management literature on modeling and control of service times, and to that on tandem stations with re-work.

Discretionary services. Classical models in queuing theory assume that service times are fixed and independent of the system state. Empirical evidence, however, shows that flexible processing times are prevalent in service, healthcare and transportation systems (Chen et al. 2001, Staats and Gino 2012, Batt and Terwiesch 2012, Tan and Netessine 2014). Empirical studies in healthcare settings illustrate several behavioral mechanisms consistent with these ideas. Berry Jaeker and Tucker (2017) document an "N-shaped" relationship between hospital occupancy and patient length of stay, showing that service speed initially increases with workload but eventually declines once staff reach saturation, highlighting a speed–quality trade-off. Similarly, Kc and Terwiesch (2012) find that intensive care units discharge patients earlier when occupancy is high, which increases subsequent readmissions—direct evidence of state-dependent effort and quality-driven re-work. Such state-dependent service times also make workload forecast difficult. Finally, Clark and Huckman (2012) show that hospitals benefit from complementarities across related clinical services, where cospecialization improves quality performance. Together, these studies provide empirical motivation for our modeling assumptions regarding state-dependent effort, quality-driven re-work, and cross-stage complementarities.

Our work is fundamentally concerned with discretionary services, in which working procedures are not necessarily pre-specified; employees or management get to determine the service time (or effort) and quality given to each customer. Hopp et al. (2007) is an early modeling work studying a controlled queueing model that allows the operator to decide how much time to allocate to a service. Much of the work on discretionary services concerns single-station services (Wang et al. 2021). Here we explore the discretionary allocation of effort across process steps/stations.

Our analysis is also related to the Brownian processing-network literature, which develops workload representations and control formulations for stochastic flow systems (Harrison and Van Mieghem 1997, Harrison 2003). In those models, the workload is obtained as a linear transformation of queue lengths that satisfies flow-balance constraints and serves as a sufficient statistic for dynamic control under heavy traffic. By contrast, our attainable workload region arises from endogenous relationships between quality, effort, and re-work, yielding a convex workload space. Rather than modeling state-space collapse and dynamic

control, we study how these endogenous quality interactions determine the feasible workload configuration at the design stage and its associated cost trade-offs.

Our focus in this paper is on the design of the system to determine the baseline service content. Our work takes a macro view of service times. We consider the initial design of the system—the allocation of work between two stations. Our quality progression model is not, in this paper, used to develop adaptive control for individual items but rather to design the *item-specific* baseline around which further dynamic fine-tuning can be done.

Queues with re-work. Kumar (1993) studies re-entrant lines in a manufacturing setting with several machines and buffers where items visit some machines more than once at different processing stages. A single class $M/G/1$ queue with a second optional service is studied by Madan (2000). Fluid models for re-entrant lines were analyzed by Dai and Weiss (1996), who derived stability conditions for different scheduling policies.

Interest in queues with re-work has increased because of its instantiation in healthcare as readmission: patients whose condition deteriorates after discharge and must return for additional service/treatment. Shi et al. (2021) and Nambiar et al. (2020) addressed operational decision making by explicitly modeling individual patient progression. Motivated by healthcare quality improvement interventions, Chan et al. (2024) studied control strategies for reducing return probabilities in stylized queueing systems where customers can return for additional service episodes. The balance and/or coordination of efforts across process steps in a healthcare setting was studied by Armony et al. (2018) and Bavafa et al. (2022).

The mix of items in our paper is fixed and given. Our focus is on introducing a simple but principled model of effort coordination in a multi-stage processing network. We cede some level of granularity, relative to some papers cited above, in order to understand substitution/complementarity and their operational role in a processing network.

We model quality evolution as a dynamic process of improvement and random deterioration; this allows us, in a tractable way, to map service-content decisions to outcomes. This tractability stems from modeling the quality evolution as a Brownian motion (BM), characterized by its mean improvement speed (drift) and variability (diffusion coefficient). The BM formulation provides a coherent and analytically convenient representation of gradual improvement and potential deterioration, linking processing-time decisions to re-work likelihood in a unified way.

Finally, models of complementarity and substitution appeared before in the operations management literature. Netessine and Zhang (2005) consider externalities in supply chains and show how these depend on whether retailers' stocking decisions are complementary or substitutes. In our case, risk is endogenized via re-work. substitution/complementarity between medical procedures/resources is also common Within the healthcare economic literature (for instance, Wilson et al. 2005).

3. The Single-Station Model

For clarity, we first construct the model for a single type of items. Items arrive following a renewal process with arrival rate λ. These are so-called "index arrivals" relative to which re-work is measured.

The evolution of the quality score is modeled as a BM, whose drift captures the quality-improvement rate, and its standard deviation the extent of randomness in improvement. The model captures the dynamics of items through the network, allows us to map service-content decisions to outcomes in a tractable way, since measurements such as processing time and hitting probability have closed-form expressions.

The processing time depends on the discharge threshold l, which is a design variable. The processing time is the time it takes a positive-drift BM $\mathcal{B}_t$ to hit a target l starting at a for the first visit and 0 for subsequent re-visits, with drift $\theta > 0$, and diffusion coefficient $\sigma > 0$:

$$\tau(a,l) = \inf\{t \geq 0 : \mathcal{B}_t = l \mid \mathcal{B}_0 = a\}.$$

Once an item's quality score reaches level l, the item leaves the station. Given a choice of l, the expected processing time and its variance are

$$m(a,l) = \mathbb{E}[\tau(a,l)] = (l-a)/\theta, \quad \text{and} \quad \mathbb{V}ar[\tau(a,l)] = (l-a)\sigma^2/\theta^3. \tag{1}$$

After an item is released, there is a baseline probability p_b that the initial processing resolves the problem permanently, making the likelihood of re-work negligible. Define:

$$\nu := -\log(1-p_b),$$

so that $e^{-\nu}$ is the base probability of no re-work. The baseline probability p_b is close to 1 (ν is large) for simpler cases, where there is little or no randomness, and service success

is essentially guaranteed. Conversely, p_b is close to 0 for more complicated cases, where randomness is inherent.

Conditional on the problem not being—terminally and deterministically—solved, re-work depends on the random evolution of the quality score. The post-release (or discharge) score follows a BM $\mathcal{B}_t^{pr}$ ($pr =$ post release), which starts at l, and has a non-negative drift $\eta(l) = \gamma \times l$ that is linear in l. This relation reflects the fact that the greater the time spent on an item, the less likely this item is to require re-work. The linear choice is not arbitrary; it will arise as a special case of a multi-station model that we later introduce.

The diffusion coefficient of this BM is $\sigma_{pr} > 0$. While the improvement rate is positive, randomness allows the quality score to hit negative levels; it reaches the *re-work* level 0 at $\tau_r = \inf\{t \geq 0 : \mathcal{B}_t^{pr} = 0\}$. At this point, the item returns to the station for additional processing. The re-work likelihood is then the probability that the positive-drift motion $\mathcal{B}_t^{pr}$, *starting at* l, hits 0 in finite time:

$$\Pr\{\tau_r < \infty | B_0^{pr} = l\} := e^{-\varrho\eta(l)l} = e^{-\varrho\gamma l^2},$$

where $\varrho := 2/\sigma_{pr}^2$; ϱ decreases (consequently $p(l)$ increases), when the quality-improvement variability increases. The probability of re-work, accounting for base re-work probability, is

$$p(l) = 1 - p_b \Pr\{\tau_r < \infty | B_0^{pr} = l\} := e^{-(\nu + \varrho\eta(l)l)} = e^{-(\nu + \varrho\gamma l^2)}, \tag{2}$$

The re-work probability decreases as l increases (hence, the average service time increases), aligning with empirical evidence on the relationship between processing time and re-work likelihood. Notably, this re-work probability, derived from our modeling framework, is consistent with the relationship ($\log p(l) = -\varrho\gamma l^2$, which parallels logit estimation commonly used in the empirical literature (Kc and Terwiesch 2009, Carey 2015); see more on estimation in §C.

Once returned for re-work, the item must be brought back to quality score l before being released again. The number of processing visits is then

$$N(l) = [1 - p(l)]^{-1} = \left[1 - e^{-(\nu + \varrho\gamma l^2)}\right]^{-1};$$

$N(l) \geq 1$ and $N(l) - 1$ is the number of returns (re-work visits). The total workload, given a release-score threshold l, is then

$$W(l) = \lambda \times (m(a, l) + m(0, l) \times [N(l) - 1]) = \lambda \times \left[\frac{l}{\theta} \times \left[1 - e^{-(\nu + \varrho\gamma l^2)}\right]^{-1} - \frac{a}{\theta}\right].$$

Implicit in this expression is a stipulation of the action if the initial score a (for the first visit) is greater than the optimal target l^*. In that case, we treat the item as requiring no processing at all.

LEMMA 1. *The workload $W(l)$ is convex in the release score l and the minimizer l^* is the unique strictly positive solution to the equation*

$$e^{-(\nu+\varrho\gamma l^2)} = (1+2l^2\varrho\gamma)^{-1}.$$

The optimal workload is given by

$$W^* = \frac{\lambda}{\theta}\left(\frac{\Gamma_\nu}{\sqrt{\varrho\gamma}} - a\right),$$

where $\Gamma_\nu = \min_{y\geq 0} y\left[1-e^{-(\nu+y^2)}\right]^{-1}$

Notice that the maximal throughput to this single station is given by

$$\lambda^* = \theta\left(\frac{\Gamma_\nu}{\sqrt{\varrho\gamma}} - a\right)^{-1}$$

We also observe that, in optimality, the likelihood of re-work $p(l^*) = e^{-(\nu+\varrho\gamma(l^*)^2)} = e^{-(\nu+(y^*)^2)}$ where $y^* = arg\min_{y\geq 0} y\left[1-e^{-(\nu+y^2)}\right]^{-1}$ depends only on ν. That is, the *optimal* solution adjusts the choice of l, so that the likelihood of return depends only on the baseline re-work probability of the item.

In systems where the arrival rate is exogenous and fixed, the load—the arrival rate multiplied by the service time— is obviously monotone in the service time. In our model, however, the workload is minimized at l^*—it is decreasing up to that point and increases thereafter[1]. The mathematical implication is that the mapping from workloads to targets l *is not invertible*: given a workload level $w > W(l^*)$, there can be two distinct values of l that achieve it. As stated, at $w^* = W(l^*)$ there is a unique value l (namely l^*) that achieves w^*. Nevertheless, a mapping from w to l can be defined by imposing a choice:

$$l(w) = \min\{l \geq 0 : W(l) = w\}.$$

For the two-station model, we map the space of service-time decisions to the space of workloads. The mapping is not one-to-one but, as here, the mapping between the two is well defined at optimality.

[1] If $\eta(l)$ does not depend on l, then $l^* = 0$ would minimize the workload; that is providing *any* service is sub-optimal. Therefore, one must allow for $\eta = \eta(l)$ to grow with l to have non-trivial solutions.

REMARK 1 (MODELING CHOICE). Starting from a dynamic model of improvement and deterioration, the BM formulation provides a unified and analytically tractable framework that links design decisions to operational outcomes. It yields closed-form expressions for hitting times, which translate discharge thresholds directly into processing-time distributions, and it connects these to re-work probabilities through a consistent dynamic mechanism. This structure enables an explicit characterization of the attainable workload region and facilitates the analysis of complementarity and substitution. The BM thus serves as both a mathematically convenient and conceptually coherent abstraction for modeling gradual improvement and potential deterioration up to a release threshold.

4. Beyond a Single Station: Complementarity and Substitution

We proceed to a two-station setting with Stations 1 and 2 and a single item type. In Section 5.3, we extend the model to incorporate multiple item types with different transfer thresholds. In the context of our Maintenance Example, this setting represents the joint modeling of *repair* followed by *preventive service* as coordinated stages of a maintenance process, where preventive activities are performed while the equipment remains offline after repair.

To the notation from the previous section, we add a subscript to identify the station: for example, θ_i is the BM drift of Station $i \in \{1,2\}$, and l_i is the target score for station i. We use l and w for the vectors (l_1, l_2) and (w_1, w_2), respectively. The mean processing time is $(l_1 - a)/\theta_1$ in Station 1 in the first visit, l_1/θ_1 in subsequent visits to this station, and l_2/θ_2 in Station 2.

Once processing at both stations is completed, an item's total quality score at discharge is $l_1 + l_2$. After release from Station 2, the post-discharge quality evolves as a BM $\mathcal{B}_t^{pr}$ that starts at this discharge level, $e \cdot l = l_1 + l_2$, with drift $\eta(l_1, l_2) \geq 0$ (defined below) and diffusion coefficient $\sigma_{pr} > 0$. Hence, the drift $\eta(l_1, l_2)$ captures how the efforts at the two stations interact—through substitution or complementarity—to influence the rate of post-release improvement or deterioration.

The resulting return probability and expected number of visits are then given by

$$p(l) = e^{-(\nu + \varrho\eta(l)(e\cdot l))}, \qquad N(l) = [1 - p(l)]^{-1}, \tag{3}$$

where $e^{-\nu}$ is the base probability of no re-work.

Figure 2 illustrates the evolution of the quality score during an item's stay in the network. It visualizes the stochastic evolution of quality and the timing of re-work. In all three scenarios, an item arrives at Station 1 at $t = 0$ with score $\mathcal{B}_0^1 = 0$. In Scenario 1 ($l_1 = 3.75, l_2 = 0$), the item is released at $t \approx 3$. Its quality deteriorates and re-work starts at $t \approx 5$. In Scenario 2 ($l_1 = 7.25, l_2 = 2.75$), the item is transferred to Station 2 at $t \approx 6$ and leaves Station 2 at $t \approx 9$. In Scenario 3 ($l_1 = 4, l_2 = 10$), the item is transferred to Station 2 at $t \approx 3$ and leaves Station 2 at $t \approx 8.5$. In our base model, for tractability, re-work occurs only after release from Station 2; see Appendix B.7 for a variant where a return to Station 1 can happen *while* in Station 2.

Complementarity and Substitution. The efforts in both stations can be substitutive or complementary. Informally speaking, the two stations are complements if, as we increase l_1, the incremental effect of l_2 increases. Conversely, they are substitutes if the incremental effect of l_2 is independent of l_1, so that the two efforts contribute additively to the outcome.

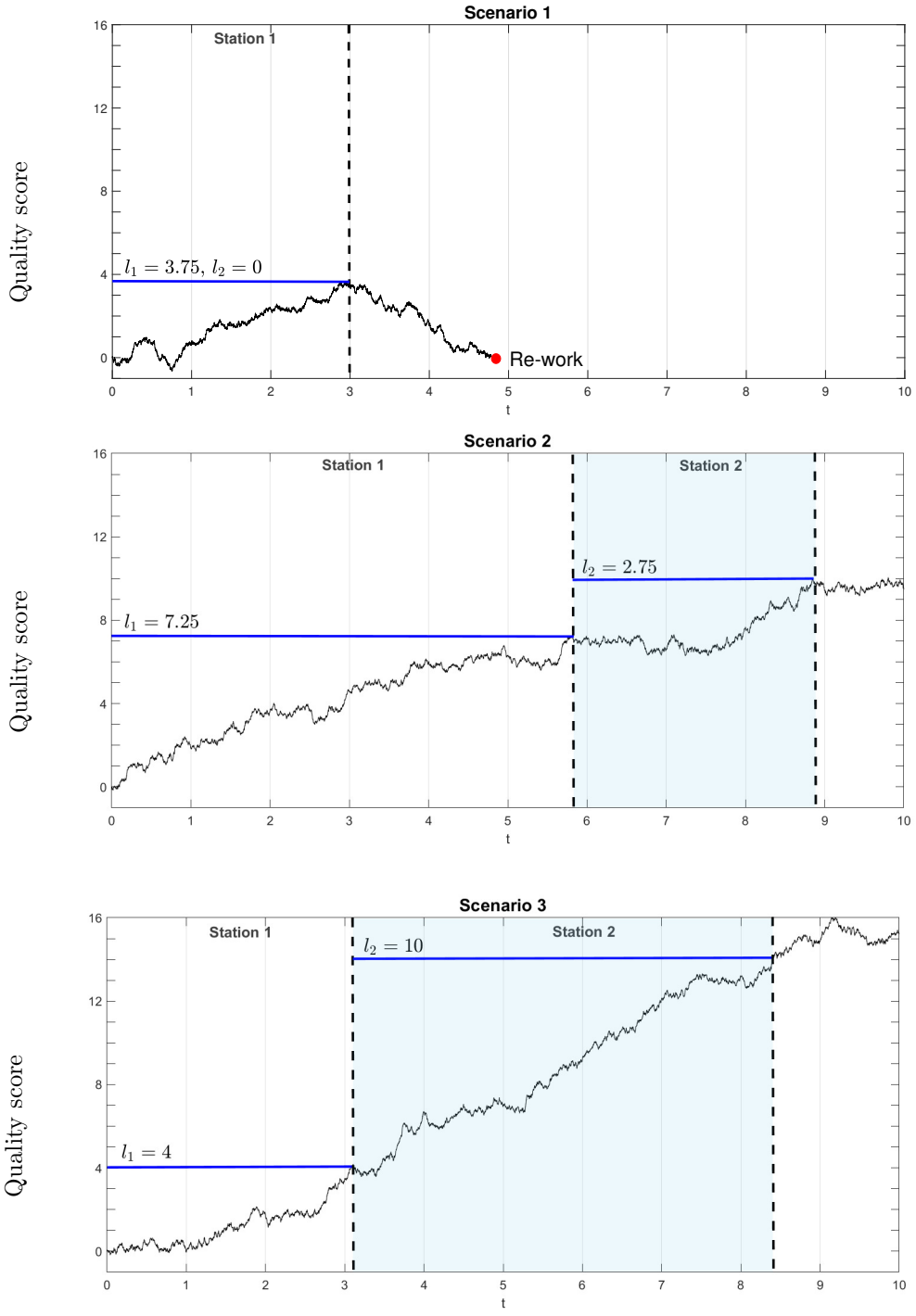
Because our goal is to study the *design* of a two-stage service process, we do not impose ex ante whether both stations must operate or only one. Instead, the optimal allocation of work across stations is determined endogenously, and the resulting design may involve effort in one or both stations depending on costs and complementarity. Complementarity and substitution, in our model, refer to how station-level efforts interact *after* discharge through the post-release dynamics that govern re-work and hence the system's effective workload.

To flexibly capture a range of such interactions, we use the CES family (Arrow et al. 1961, Sato and Koizumi 1973, Sato 1967, Stern 2011):

$$\eta(l) = (\alpha l_1^{\beta} + (1-\alpha) l_2^{\beta})^{\frac{1}{\beta}}, \quad 0 < \alpha < 1, \quad \beta \leq 1, \tag{4}$$

where α and $1-\alpha$ are the relative weights given to Stations 1 and 2, respectively, and β is the scaling parameter that captures the degree of substitution or complementarity. With $\beta = 1$, $\eta(l) = \alpha l_1 + (1-\alpha) l_2$, representing perfect substitution; as $\beta \downarrow -\infty$, $\eta(l) \to \min\{l_1, l_2\}$, representing perfect complementarity. In between, as $\beta \downarrow 0$, we obtain the so-called Cobb–Douglas structure, $\eta(l) \to l_1^{\alpha} l_2^{1-\alpha}$ (see Mas-Colell et al. 1995, Exercise 3.C.6). The CES family has long been used to capture complementarity and substitution since the seminal work of Arrow et al. (1961) and remains the basis for empirical specification

Figure 2 **Sample paths illustrating the evolution of the quality score and the timing of re-work.**



(e.g., Koesler and Schymura 2015, Mahaboob et al. 2017, Henningsen et al. 2021, Fujiy et al. 2022). It is through $\eta(l)$, and the substitution or complementarity it captures, that the total score is mapped into the re-work probability.

Observe that a single station is the same as taking $\alpha = 1$. In this case, $\eta(l) = l$, which is consistent with our developments in §3 taking $\gamma = 1$ there.

Station labeling. Assuming $\alpha \geq 1/2$ without loss of generality (stations can be relabeled otherwise), station 1 ($\alpha \geq 1/2$, decision variable l_1) is the *primary station*, while station 2 is the *secondary station*. When $\alpha > 1/2$, the primary station is more effective, contributing more than the secondary one.

4.1. The Attainable Workload Region

In this section, we characterize the attainable workload region in the two-station setting. Beyond having its own intrinsic value, this characterization lays the foundations for solving the effort allocation problems in a structured informative way.

Given a choice $l = (l_1, l_2)$, the workload in stations 1 and 2 is given by [2]

$$W_1(l) = \frac{\lambda}{\theta_1}\left[l_1 N(l) - a\right], \quad W_2(l) = \lambda \frac{l_2}{\theta_2} N(l). \tag{5}$$

We define the *rate-normalized* workload as

$$W^r_{1,a}(l) = l_1 N(l) - a \quad \Rightarrow \quad W^r_1(l) = W^r_{1,a}(l) + a = l_1 N(l), \quad W^r_2(l) = l_2 N(l), \tag{6}$$

This is the workload if $\lambda = \theta_1 = \theta_2 = 1$ (hence "rate-normalized"). The actual workload is then a linear transformation of the rate-normalized workload. The vector $(W^r_1(l), W^r_2(l))$ takes values in the set

$$\mathcal{W} = \{(w_1, w_2) \geq 0 : \exists l = (l_1, l_2) \geq 0, \text{ s.t. } w^r_1 = l_1 N(l), \ w^r_2 = l_2 N(l)\}. \tag{$\mathcal{W}$-region}$$

Theorem 1 is a key result. Recall that, by construction, $\alpha \geq 1/2$ and that $\beta \leq 1$. Also, we let Γ_ν be the minimal rate-normalized workload in the single-station case with $\varrho\gamma = 1$, that is

$$\Gamma_\nu = min_{y \geq 0} \ y\left[1 - e^{-(\nu + y^2)}\right]^{-1}. \tag{7}$$

The minimizer, y^*, is the unique solution to

$$e^{-(\nu + y^2)} = (1 + 2y^2)^{-1}. \tag{8}$$

[2] As in the single-station case, we assume that an item is "rejected" if $a \geq l^*_1 + l^*_2$; it is sent directly to the second station if $l^*_1 \leq a < l^*_1 + l^*_2$. Such item types become dedicated second station traffic in which case the single-station model is to be used with initial score $a - l^*_1$.

THEOREM 1. *The attainable workload set $\mathcal{W}$ is convex and can be written equivalently as*

$$\mathcal{W} = \{(w_1, w_2) \geq 0 : \quad w_2 \geq f_{\alpha,\beta}(w_1)\},$$

where

$$f_{\alpha,\beta}(w_1) = \begin{cases} w_1 \mathbb{h}^{-1}_{\alpha,\beta}\left(\frac{\Gamma_\nu^2}{w_1^2}\right), & w_1 \leq w_1^0, \\ 0, & w_1 > w_1^0 \end{cases}, \tag{9}$$

as well as

$$\mathbb{h}_{\alpha,\beta}(z) := \varrho\left(\alpha + (1-\alpha)z^\beta\right)^{\frac{1}{\beta}}(z+1), \quad w_1^0 := \Gamma_\nu \frac{1}{\sqrt{\mathbb{h}_{\alpha,\beta}(0)}},$$

and Γ is as in (7). Furthermore, $\mathcal{W}$ can be expressed equivalently as

$$\mathcal{W} = \{(w_1, w_2) \geq 0 : \eta(w)(e \cdot w) \geq 1\}. \tag{10}$$

Theorem 1 establishes that the attainable workload region $\mathcal{W}$ is convex and characterizes its boundary. Intuitively, this convexity reflects the trade-off between the workloads at the two stations: an increase in one station's effort can partially offset a decrease in the other while maintaining the same overall system performance. The specific shape of this region depends on the parameters α and β, which determine the degree of substitutability or complementarity between the two stations. Proposition 1 and Figure 3 further illustrate how these parameters shape the attainable region and influence the boundary's curvature.

REMARK 2 (ROBUSTNESS OF CONVEXITY). The convexity of the attainable workload region W derives from the properties of the aggregation function $\eta(l)$, which combines the quality contributions across stations. As shown in the final part of the proof of Theorem 1, convexity follows from the fact that $\eta(l)$ is concave (and in particular, log-concave), implying that the product $\eta(w)(e \cdot w)$ is log-concave and its upper contour sets are convex. Hence, any positive aggregation function $\eta(\cdot)$ that is concave would also yield a convex feasible region W. By contrast, if $\eta(\cdot)$ were non-concave—for example, exhibiting increasing returns to scale—the corresponding region could become non-convex, as effort reductions at one station might amplify rather than offset the workload at another. Therefore, convexity is not specific to the CES family but stems from the general diminishing-returns property of the aggregation function.

The characterization of the attainable workload region extends naturally to any number of stations n, as shown in Appendix B.1.

EXAMPLE 1. A special but informative case has $\alpha = 1/2$ (stations are equally important) and $\beta = 1$ (they are perfect substitutes). In this case, $\mathbb{h}_{0.5,1}(z) = \frac{1}{2}\varrho(1+z)^2$, so that $f_{0.5,1}(w_1) = \Gamma_\nu\sqrt{2/\varrho} - w_1$ for $w_1 \le w_1^0 = \Gamma_\nu\sqrt{2/\varrho}$. The set $\mathcal{W}$ in this case has the simple linear boundary shown in the left plot of Figure 3 ∎

Next, we characterize the boundary $f_{\alpha,\beta}$ of the set $\mathcal{W}$. This will be instrumental in our study of optimization problems over $\mathcal{W}$. In Proposition 1, below, $f_{\alpha,\beta}$ is as in (9).

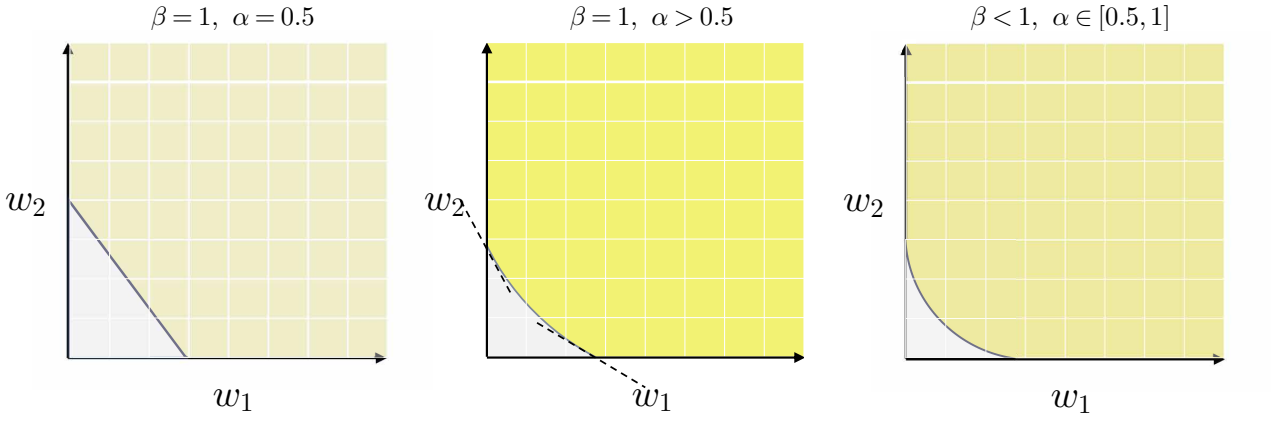
PROPOSITION 1. *The derivative of $f_{\alpha,\beta}(w_1)$ satisfies the following:*

$$\lim_{w_1\uparrow w_1^0} f'_{\alpha,\beta}(w_1) = \begin{cases} 0, & \text{if } \beta < 1, \\ -\gamma_1, & \text{if } \beta = 1, \end{cases} \quad \text{and} \quad \lim_{w_1\downarrow 0} f'_{\alpha,\beta}(w_1) = \begin{cases} -\infty, & \text{if } \beta < 1, \\ -\gamma_2, & \text{if } \beta = 1, \end{cases}$$

where $\gamma_1 = 2(\Gamma_\nu/w_1^0)^2 > 0$, $\gamma_2 = 1/(1-\alpha) > 0$.

Proposition 1 stipulates that for any $\beta < 1$, the boundaries of the convex set "asymptote" smoothly to the axes. This implies that, for linear objective functions, there are no optimal solutions of the form $(0, d)$ or $(l, 0)$ while we might have some for $\beta = 1$. Indeed, from Example 1 we know that with $\beta = 1$ and $\alpha = 1/2$, the boundary is linear. Figure 3 illustrates the three possible scenarios. Importantly, when $\alpha > 1/2$ perfect substitution in drift ($\beta = 1$) could still be consistent with an optimal solution that requires the non-negligible use of both stations. This is established in the next station.

Figure 3 **The set $\mathcal{W}$ for different values of α, β. In the center, the boundary is convex but "hits" the axes with a strictly negative derivative (the tangents in black represent these derivatives). On the right, the tangents at the points where $f_{\alpha,\beta}$ meets the axes are the axes themselves.**



Model summary. Before turning to the analysis, we briefly summarize the main modeling assumptions for clarity. Items arrive at rate λ and evolve independently. While in Station i, an item's quality follows a BM with drift θ_i and diffusion σ_i, and service ends when its quality score reaches a threshold l_i. After release, quality continues to evolve with drift $\eta(l)$ and diffusion σ_{pr}, and the probability of re-work is $p(l) = e^{-(\nu+\varrho\eta(l)(e\cdot l))}$, with re-worked items returning to Station 1. In the two-station setting, station-level quality contributions aggregate via the CES function $\eta(l) = (\alpha l_1^\beta + (1-\alpha) l_2^\beta)^{1/\beta}$. Appendix A summarizes all main notation and symbols used throughout the paper.

For completeness, Appendix C describes how the model's parameters can be estimated from operational data: the quality-evolution parameters (θ_i, σ_i) are inferred from service-time observations, and the substitution/complementarity parameters (α, β) are estimated from realized re-work frequencies. The appendix also includes a simulation demonstrating that the proposed estimation procedure accurately recovers these parameters even in moderate sample sizes.

5. Processing Costs and Optimization

An item's handling cost reflects the resources allocated to it. The more resources deployed, the higher the cost per unit of time. In other words, c_1 and c_2—the cost per item per unit of time in Stations 1 and 2, respectively—would be greater for items that are more important, present a higher risk or require more resources.

Recall that under a given decision pair l an item (re)visits each station $N(l)$ times in expectation; this includes the index visit and any subsequent visits. The handling cost at each station is then given by

$$\begin{aligned}\text{Station 1 cost} &= \lambda \frac{c_1}{\theta_1}\left(l_1 N(l) - a\right) = \lambda \frac{c_1}{\theta_1} W_1^r(l),\\ \text{Station 2 cost} &= \lambda \frac{c_2}{\theta_2} l_2 N(l) = \lambda \frac{c_2}{\theta_2} W_2^r(l),\end{aligned}$$

Recalling (6), the total journey cost for an item is, therefore,

$$\text{Station 1 cost} + \text{Station 2 cost} = \lambda \frac{c_1}{\theta_1} W_1^r(l) + \lambda \frac{c_2}{\theta_2} W_2^r(l)$$

With capacity constraints C_1 and C_2 in Stations 1 and 2, respectively, we arrive at the (constrained) optimization problem:

$$\begin{aligned} \min_{l\geq 0} \quad & (c_1/\theta_1)\cdot W_1^r(l) + (c_2/\theta_2)\cdot W_2^r(l) \\ \text{s.t. } & (\lambda/\theta_1)\cdot W_1^r(l) \leq C_1, \\ & (\lambda/\theta_2)\cdot W_2^r(l) \leq C_2. \end{aligned} \tag{11}$$

The solution for this capacity-constrained problem builds in a fundamental way on that of the unconstrained problem so we start with the latter.

5.1. The Uncapacitated Problem

This is the case where $C_1, C_2 = \infty$ in (18). In this case, the optimization problem decomposes into K single-type problems. Fixing one type (and omitting the subscript k from notation), we have the optimization problem

$$V^* = \min_{l\in\mathbb{R}^2_+} \lambda\left(\frac{c_1}{\theta_1}W_1^r(l) + \frac{c_2}{\theta_2}W_2^r(l)\right). \tag{12}$$

Relative marginal cost (relative cost, in short). The ratio c_1/θ_1 is *the marginal cost of quality improvement in Station 1:* c_1 is the resource consumption cost while $1/\theta_1$ is the time during which this effort is exerted. This is an effectiveness measure that is Station-1-focused. Similarly, c_2/θ_2 is an effectiveness measure that is Station 2-focused.

The measure

$$\mathcal{R}^c := \frac{c_2/\theta_2}{c_1/\theta_1},$$

then quantifies the *relative cost* of Stations 1 and 2. According to our labeling, Station 1 is the primary one and thus more effective ($\alpha > 1/2$); it can, however, be either cheaper or costlier than Station 2.

The optimal solution is identical to that of the cost-normalized problem,

$$\bar{V}^*(\mathcal{R}^c) := \frac{\theta_1}{\lambda c_1}V^* = \min_{l\geq 0}\left(W_1^r(l) + \mathcal{R}^c W_2^r(l)\right) \quad \text{(cost-normalized problem)} \tag{13}$$

We write $(l^*)(\mathcal{R}^c)$ where needed to make the dependence of decisions on the relative cost explicit.

The result below—which identifies a relationship between l_1^* and l_2^* that depends on α and β— is familiar from budget-constrained production-quantity maximization with CES

production functions. While, at this point, we have no explicit resource-budget constraints, these are implicit through the re-work; our problem is, informally, a dual problem where we minimize resource consumption while satisfying a production constraint.

LEMMA 2. *If there exists an interior optimal solution to* (13)*, i.e., with* $l^* > 0$*, it must be of the form*

$$l_2^* = z_0\, l_1^*,$$

where $z_0 = z_0(\mathcal{R}^c) > 0$ *is the unique solution to*

$$\left(z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)}\right) = -\frac{1}{\mathcal{R}^c}, \tag{14}$$

and $\mathbb{h}_{\alpha,\beta}(\cdot)$ *is as in Theorem 1. Given* z_0*,* $l_1^* = y^*/\sqrt{\mathbb{h}_{\alpha,\beta}(z_0)}$ *and* $l_2^* = \Gamma_\nu z_0/\sqrt{\mathbb{h}_{\alpha,\beta}(z_0)}$ *where* Γ_ν *and* y^* *are as in* (7) *and* (8)*, respectively.*

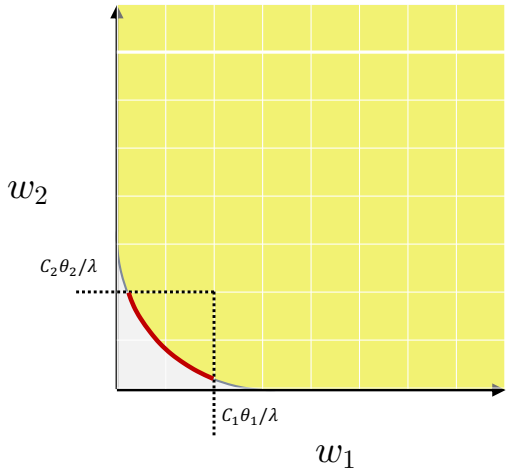
With $\alpha = 0.5$ and $\beta = 1$—recall Example 1—the attainable workload region has a linear boundary. The slope of that linear boundary is -1. Hence, minimizing the linear cost function $w_1 + \mathcal{R}^c w_2$ yields only corner solutions with the exception of the case where $\mathcal{R}^c = 1$ (where there are infinitely many solutions).

Under the optimal solution l^*, the likelihood of re-work depends only on ν and not on the relative costs. As in Lemma 1, $p(l^*) = e^{-(\nu+\varrho\gamma(l^*)^2)} = e^{-(\nu+(y^*)^2)}$ where y^* is as in (8). The optimization problem adjusts the decision to the cost and substitution parameters so that the likelihood of re-work is the same across all types (i.e., all combination of cost and substitution) that have the same base re-work parameter ν.

The problem (13) is a minimization of a linear function over a convex region. At the optimal solution (if attained at a smooth boundary), it must be the case that the cost function is a tangent to the boundary. In Proposition 1 we saw that with $\beta < 1$, the axes are the tangents to the function $f_{\alpha,\beta}$ at the points where it meets the axes. This means that—with the exception of the trivial case that $\mathcal{R}^c = 0$ (or $\mathcal{R}^c = \infty$)—there will be only interior optimal solutions where both $w_1, w_2 > 0$; see Figure 4. With $\beta = 1$, the tangents at both meeting points with the axes (recall Figure 3) are non-trivial. Thus, there is a value of $\mathcal{R}^c$, for which the cost function is a tangent to $f_{\alpha,\beta}$ at $w_1 = 0$ and another cost where the cost function is tangential to $f_{\alpha,\beta}$ at $w_1 = w_1^0$.

Lemma 2 characterizes the interior solution corresponding to the middle regime, where both stations are active. Theorem 2 extends that result by showing that, for $\beta = 1$, the

Figure 4 **The attainable workload region for $\beta < 1$ and the cost objective tangent; $\beta < 1$, $\alpha \in [0.5, 1]$**



interior solution applies only within the middle regime where both stations are active, while the first and last regimes correspond to corner solutions where one station is inactive. In those cases, the optimal solution shifts to a corner regime in which only one station operates, because the interior condition of Lemma 2—requiring both $l_1^*, l_2^* > 0$—cannot be satisfied when the cost function is tangent to the boundary along an axis (see the middle panel of Figure 3, where the tangents at the axes illustrate these corner solutions).

THEOREM 2 **(operating regimes)**. *The optimal solution to (13) is:*

- *If $\beta < 1$ (partial substitution or complementarity), there exists only the unique interior optimal solution characterized in Lemma 2.*
- *If $\beta = 1$ (substitution), there exists $0 < \mathcal{R}^c_{LB} \leq \mathcal{R}^c_{UB} < \infty$, such that the optimal solution $l^*(\mathcal{R}^c)$ is:*
  - *For $\mathcal{R}^c \in [0, \mathcal{R}^c_{LB}]$: $l_1^*(\mathcal{R}^c) = 0$ and $l_2^0 := l_2^*(\mathcal{R}^c) = y^*/\sqrt{\varrho(1-\alpha)}$ with $\bar{y}$ in (8).*
  - *For $\mathcal{R}^c \in (\mathcal{R}^c_{LB}, \mathcal{R}^c_{UB})$: $l^*(\mathcal{R}^c)$ is as characterized in Lemma 2.*
  - *For $\mathcal{R}^c \in [\mathcal{R}^c_{UB}, \infty)$: $l_2^*(\mathcal{R}^c) = 0$ and $l_1^0 := l_1^*(\mathcal{R}^c) = y^*/\sqrt{\varrho\alpha}$ with y^* in (8).*

This theorem identifies the existence of three *operating regimes*. As long as there is some complementarity ($\beta < 1$), it is optimal to have some effort in each of the two process steps. When the post-release drift is linear in the efforts—there is perfect substitution in the drift ($\beta = 1$)—there are three distinct operating regimes: one where no processing is optimally happening in station 1, one where no processing is happening in station 2, and one where some processing is required in both. Thus, with $\beta = 1$, the exact operating regime is determined by the relative cost.

Table 1 summarizes the three operating regimes characterized by Theorem 2, indicating which stations are active and how the optimal thresholds behave across substitution levels and relative costs.

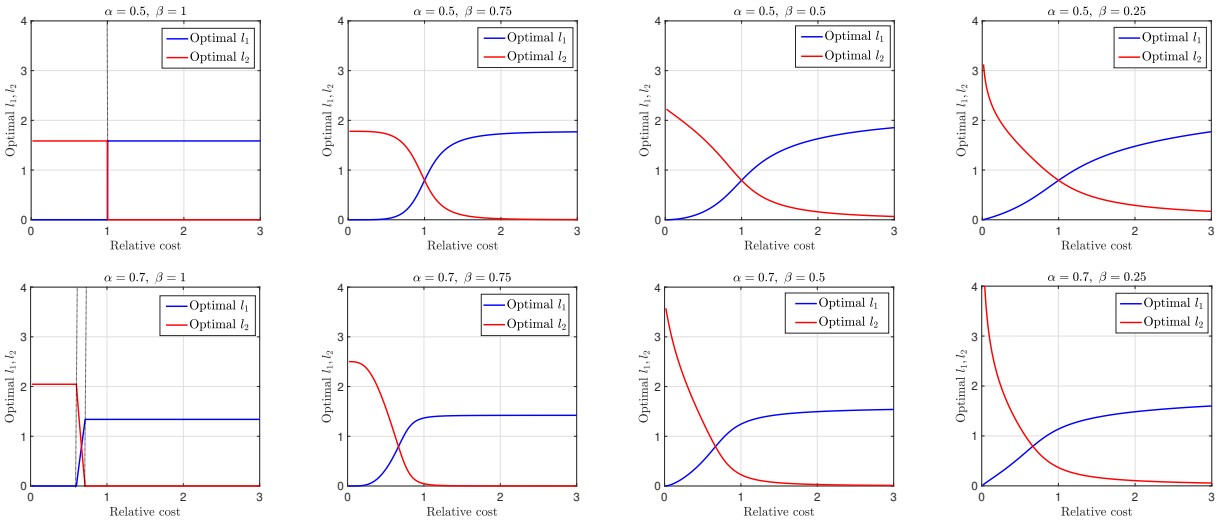
Table 1 Operating regimes characterized by Theorem 2.

Regime	Complementarity level (β)	Relative cost range (R^c)	Threshold behavior with R^c
I. Station 1 only	$\beta = 1$	$R^c \geq R^c_{UB}$	$l_1 = l_1^0$ (constant)
II. Station 2 only	$\beta = 1$	$R^c \leq R^c_{LB}$	$l_2 = l_2^0$ (constant)
III. Both active	$\beta < 1$ or ($\beta = 1$ and $R^c_{LB} < R^c < R^c_{UB}$)	—	l_1 decreases and l_2 increases; stronger complementarity ($\beta \downarrow$) $\rightarrow$ more balanced effort

Figure 5 is a numerical illustration of the optimal decisions in Theorem 2. It shows symmetry around a threshold value, explicitly characterized in Proposition 2(i). For completeness, we note that the qualitative behavior of the solution extends to $\beta \leq 0$, where increasing complementarity (smaller β) accentuates the curvature of the response and shifts effort toward the more effective station. The limiting case $\beta \to -\infty$ corresponds to perfect complementarity, in which both stations must contribute equally.

Maintenance Example. Theorem 2 shows that when repair and preventive service are strong complements, both stages optimally receive positive effort, whereas under near substitution, resources are allocated primarily to the more cost-effective stage. Appendix D grounds these results in an order-of-magnitude calibration based on industrial maintenance operations for assets such as wind turbines, generators, and other heavy equipment. The calibrated parameters produce consistent qualitative patterns and magnitudes aligned with observed practices, where repair is followed by preventive service while equipment remains offline. Appendix E complements the analytical results with numerical experiments demonstrating robustness: the optimal design remains stable under moderate parameter variation, and the integrated policy yields substantial cost savings relative to decentralized benchmarks.

Figure 5 **The solution structure in Theorem 2: Optimal l_1,l_2 as a function of relative cost $\mathcal{R}^c$. The parameters are $\varrho = 1$, $a = 0$, $\theta_1 = 1$, $\theta_2 = 0.6$, $c_1 = 1$.**



PROPOSITION 2 **(the complementarity–relative cost interaction)**. *Let* $l^*(\beta, \mathcal{R}^c)$ *be the optimal solution as a function of the complementarity and the relative cost. The following then holds:*

(i) *At* $\mathcal{R}_0^c := (2(1-\alpha)+1)/(2\alpha+1) = -\left(z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)}\Big|_{z=1}\right)^{-1}$, $l_1^*(\beta, \mathcal{R}_0^c) = l_2^*(\beta, \mathcal{R}_0^c)$ *(both stations' decisions are identical). Moreover,* $l^*(\beta, \mathcal{R}_0^c)$ *does not depend on* β.

(ii) *For all* $\mathcal{R}^c < \mathcal{R}_0^c$, *the ratio* $l_1^*(\beta, \mathcal{R}^c)/l_2^*(\beta, \mathcal{R}^c)$ *is decreasing in* β *(with all else equal, greater complementarity means a smaller ratio). For all* $\mathcal{R}^c > \mathcal{R}_0^c$, *the ratio* $l_1^*(\beta, \mathcal{R}^c)/l_2^*(\beta, \mathcal{R}^c)$ *is increasing in* β *(with all else equal, greater complementarity means a larger ratio).*

(iii) *The optimal cost function* $\bar{V}^*(\beta, \mathcal{R}^c)$ *is decreasing in* β *(with all else equal, greater complementarity means a larger value function).*

Figure 6 is an illustration of Proposition 2. At the simplest level, complementarity effectively introduces a constraint on the network; to benefit from the effort in one station, one must make greater effort in the other. In turn, the higher the complementarity (the smaller the value of β), the greater the optimal cost.

More subtle is the way that optimal decisions differ on both sides of the threshold $\mathcal{R}_0^c$. When $\alpha = 1/2$, $\mathcal{R}_0^c = 1$, item (ii) of the proposition stipulates that if $\mathcal{R}^c < \mathcal{R}_0^c$, the greater the complementarity (with the relative cost fixed)—not only will more work be done (optimally) at the cheaper second station—but that *relatively* more is done in the

second station: the relative decrease in the first station is dominated by the relative increase in the second.

The explicitly identified value of the threshold shows that the range of costs where this happens shrinks as the first station becomes more dominant (α grows).

In Figure 6, $\alpha = 0.7$, so that threshold is, per part (i) of Proposition 2, equal to $2/3$. Complementarity keeps the stations as balanced as possible—as extracting value from one station necessitates keeping some effort in the other. Accordingly, we see that as we move away from the symmetry point, it is the types with smaller β that are the slowest to change (in relative terms).

This result implies that when different item types share the same relative cost but differ in their substitution level, it is sufficient to know on which side of the threshold the relative cost falls to effectively allocate efforts between the stations.

Improving Station 2 to render it a better substitution for Station 1 would have different effects on different items. Two item types that share the same relative cost, might receive relatively more effort in one station depending on where items' costs stand relative to their respective thresholds.

Total Processing Time. Figure 6(c) presents the total (i.e., including all visits and both stations) for different values of β. Because, per item type, the re-work likelihood is *at optimality* constant, the total processing time is proportional to a single visit length (except the first visit if $a > 0$). As the cost of the second station increases, the processing time initially decreases but may then increase as Station 2 becomes prohibitively expensive—this latter increase is more substantial if complementarity is substantial. Like the ratio l_1/l_2, the effect of complementarity is different (and reversed) on the left and right of the threshold.

Importantly, there is a point of relative cost (on the right of the threshold) where the processing time is minimized. Recalling that (optimal) re-work is constant within an item type, this is the "sweet spot" – it achieves the optimal processing time and re-work likelihood.

5.2. The Capacitated Problem

When capacity is finite, $C_1, C_2 < \infty$ in Problem (18), the choice set is an intersection of the attainable workload region and the sub-space defined by the constraints; see Figure 7.

Figure 6 **Optimal l_1/l_2, V, and total length of stay for different values of β. The parameters are $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, $c_1 = 1$.**

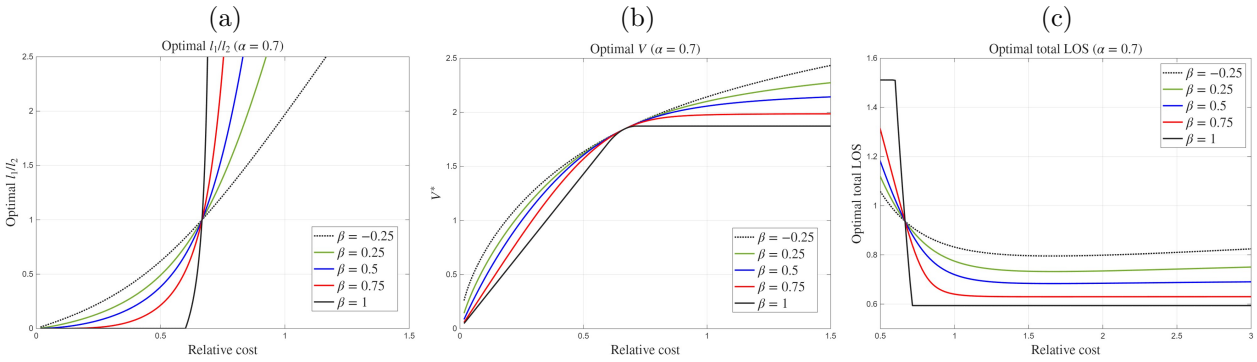
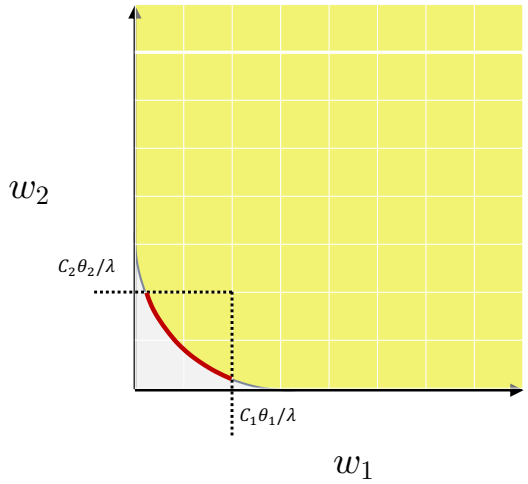


Figure 7 **A system with a single item type and finite capacities C_1 and C_2. The optimal solution to the cost optimization problem lies on the boundary portion marked in red.**



Let us define the rate-normalized capacities

$$C_i^r = C_i\theta_i/\lambda, \quad i = 1, 2.$$

For the constrained optimization problem to be feasible, it must be that

$$C \in \mathcal{C}_\beta = \{C_1, C_2 : C_2^r \geq f_{\alpha,\beta}(C_1^r)\}, \tag{15}$$

where $f_{\alpha,\beta}$ is as in Theorem 1. The attainable set shrinks with β (Lemma 8): $\mathcal{C}_{\beta_2} \subseteq \mathcal{C}_{\beta_1}$ if $\beta_1 \geq \beta_2$ with all else $(\lambda, \theta_1, \theta_2)$ fixed.

Theorem 3 characterizes the optimal capacitated solution, which is based upon the uncapacitated solution.

THEOREM 3 **(constrained service anatomy).** *Fix $C \in \mathcal{C}_\beta$. Let $l^{b,*}(\mathcal{R}^c)$ be the solution to the unconstrained problem when the relative cost is $\mathcal{R}^c$. Then, the optimal solution l^* to (17) is given by $l^{b,*}(\widetilde{\mathcal{R}}^c)$ where*

$$\widetilde{\mathcal{R}}^c = \mathcal{R}^c \left(\frac{c_2+\kappa_2}{c_2}\right) / \left(\frac{c_1+\kappa_1}{c_1}\right),$$

with

$$\kappa_1 = \max\left\{c_1(|f'_{\alpha,\beta}(C_1^r)|\mathcal{R}^c - 1), 0\right\}, \quad \kappa_2 = \max\left\{c_2\left(\frac{1}{|f'_{\alpha,\beta}(f^{-1}_{\alpha,\beta}(C_2^r))|\mathcal{R}^c} - 1\right), 0\right\},$$

and either they are both infinite—if $C_1^r = f^{-1}_{\alpha,\beta}(C_2^r)$ and $f'_{\alpha,\beta}(C_1^r) = -1/\mathcal{R}^c$—or only one of them is strictly positive. Finally, κ_i is decreasing in C_i.

The Lagrange multipliers κ_1, κ_2 capture the value of increasing the capacity. Because they are explicitly specified, we can analyze the effect of complementarity/substitution on them. To that end, let C^0 be such that $f_{\alpha,\beta}(C^0) = C^0$. Recall that this point does not depend on β and is given by $\Gamma/\sqrt{\mathbb{h}_{\alpha,\beta}(1)} = \Gamma/\sqrt{2}$. The following is a direct consequence of Proposition 2.

LEMMA 3. *Fix β and C_1, C_2, which are feasible at $(\beta-\delta, \beta+\delta)$, for some $\delta > 0$.*

1. *If $C_1^r < C^0$, then κ_1 increases with β, while $\tilde{R}^c$ and l_1^*/l_2^* decrease with β; the monotonicities reverse if $C_1^r \geq C^0$.*
2. *If $C_2^r < C^0$, then κ_2 decreases with β, while $\tilde{R}^c$ and l_1^*/l_2^* increase with β; the monotonicities reverse if $C_2^r \geq C^0$.*

This result has an intuitive interpretation that is best understood by considering the extreme case of complementarity. As $\beta \downarrow -\infty$, $\eta(l) \to \min\{l_1, l_2\}$. Thus, if $l_1 < l_2$, it is infinitely more valuable to increase l_1 than it is to increase l_2. The best is when the two are equal. Extending this logic informally to any $\beta < 1$, the intuition is that we want to "balance" the line. The greater the substitution, the more we can realize the aforementioned value because it is costless to transfer work from station 2. If, however, there is strong complementarity, in order to extract value in Station 1, we must keep some work in Station 2. Hence, we get less value from increasing the capacity (κ_1 is increasing in β). When $C_1 > C^0$, this logic is reversed.

When the capacity in Station 1 is small, the more substitution we have, the more we want to transfer work to Station 2 (hence, we "make" Station 2 cheaper by reducing $\widetilde{\mathcal{R}}^c$). Complementarity, on the other hand, would mean keeping Station 2 expensive to make sure we keep enough work in Station 1.

Maintenance Example. When repair capacity is limited, the system must rely more heavily on preventive service to maintain quality, increasing the overall workload. If repair and preventive service are strong complements, however, this adjustment is less effective: restricted repair capacity reduces the value of subsequent preventive work and leads to sharper performance deterioration. The calibrated results in Appendix D.1 illustrate this mechanism. When repair capacity decreases from ten to six workers, the system compensates by increasing preventive effort, but this adjustment yields smaller gains under strong complementarity. In complementary systems, the effectiveness of preventive service depends on sufficiently thorough repair; thus, limited capacity at the repair stage reduces the overall benefit of both stages. By contrast, when the stages are more substitutable, effort can be shifted more flexibly toward preventive service, resulting in smaller efficiency losses.

5.3. Multiple Types

In many practical systems, items differ in their characteristics, leading to heterogeneity in costs, processing rates, and complementarity levels. In this section we generalize the analysis to incorporate multiple types of items flowing through the two stations. When resources are scarce, the different item types "compete" over them. Let $[K] = \{1, \dots, K\}$ be the set of item types; we expand the notation by adding the superscript k for Type k. Thus, for example, a^k is the initial quality of type-k jobs, θ_i^k is the rate of improvement for type-k jobs in Station $i \in \{1, 2\}$, $\alpha^k \in [0, 1]$ is the station-importance coefficient for type k items, and β^k is the substitution/complementarity coefficient.

Let the total arrival rate to the system be λ, and let $\mathbf{p}_k$ denote the fraction of arrivals of type k, so that the arrival rate of type k items is $\lambda^k = \mathbf{p}_k \lambda$ with $\sum_{k\in[K]} \mathbf{p}_k = 1$. Each type k is characterized by a pair of station thresholds $l^k = (l_1^k, l_2^k)$. We denote by $\vec{l} = (l^1, \dots, l^K)$ the collection of all type-specific threshold pairs, which we represent as a stacked $2K$-dimensional decision vector.

The attainable workload region for type k is

$$\mathcal{W}_k := \left\{ (w_1^k, w_2^k) \geq 0 : \ \exists\ l^k = (l_1^k, l_2^k) \geq 0 \text{ s.t. } w_1^k = l_1^k N(l^k),\ w_2^k = l_2^k N(l^k) \right\},$$

where each $\mathcal{W}_k$ is convex and has the structure derived in Theorem 1 and Proposition 1, with the corresponding α^k and β^k. A vector (of pairs) $\mathbf{w} = (\vec{w}^1, \dots, \vec{w}^K)$ belongs to

$$\mathcal{W}^\times := \prod_{k\in[K]} \mathcal{W}_k, \tag{16}$$

if $\vec{w}^k \in \mathcal{W}_k$ for each k.

Given a decision vector $\vec{l}$, we let $\mathcal{L}_i(\vec{l})$, $i=1,2$, denote the total load on Station i:

$$\mathcal{L}_i(\vec{l}) = (\lambda/\theta_i) \cdot \mathbf{W}_i^r(\vec{l}),$$

where $\mathbf{W}_i^r(\vec{l}) = (W_i^{r,1}(l^1), \ldots, W_i^{r,K}(l^K))$ is the K-dimensional vector of workloads contributed by all item types at Station i. The set of feasible capacity vectors $C=(C_1,C_2)$ is defined as

$$\mathcal{C}_{\boldsymbol{\beta}} = \{C \geq 0 : \ \exists\ \vec{l} \geq 0 \text{ s.t. } \mathcal{L}_i(\vec{l}) \leq C_i,\ i=1,2\},$$

where the dependence on $\boldsymbol{\beta} = (\beta^k, k \in [K])$ is made explicit.

As in the case of a single type, the set $\mathrm{tr}\mathcal{C}_{\boldsymbol{\beta}}$ is "increasing" in the vector $\boldsymbol{\beta}$ of complementarity/substitution levels under the standard partial order: $\mathcal{C}_{\boldsymbol{\beta}_0} \subseteq \mathcal{C}_{\boldsymbol{\beta}_1}$ for $\boldsymbol{\beta}_0 = (\beta_0^1, \ldots, \beta_0^K)$ and $\boldsymbol{\beta}_1 = (\beta_1^1, \ldots, \beta_1^K)$ such that $\beta_0^k \leq \beta_1^k$ for all $k \in [K]$.

In words, the greater the complementarity between the stations, the greater the capacity that is needed to serve everyone. Scarce capacity introduces an interaction between item types. One must prioritize the capacity usage of the two stations, and one expects that all item types will be allocated less *aggregate* effort, with some "suffering" more than others. Beyond aggregation, scarce capacity also alters the effort allocation between the stations, and this reallocation differs across item types depending on their characteristics.

The total cost for K types of items with different characteristics is

$$V(\vec{l}) = (\lambda c_1/\theta_1) \cdot \mathbf{W}_1^r(\vec{l}) + (\lambda c_2/\theta_2) \cdot \mathbf{W}_2^r(\vec{l}),$$

where c_i/θ_i (for $i=1,2$) denotes the K-dimensional vector $(c_i^1/\theta_i^1, \ldots, c_i^K/\theta_i^K)$ and $\mathbf{W}_i^r(\vec{l}) = (W_i^{r,1}(l^1), \ldots, W_i^{r,K}(l^K))$ is the vector of workloads contributed by all item types at Station i. The decision variable is $\vec{l} = (l^1, \ldots, l^K)$.

The multi-type constrained optimization problem is therefore

$$\begin{aligned} \min_{\vec{l} \geq 0} \quad & (c_1/\theta_1) \cdot \mathbf{W}_1^r(\vec{l}) + (c_2/\theta_2) \cdot \mathbf{W}_2^r(\vec{l}) \\ \text{s.t. } & (\lambda/\theta_1) \cdot \mathbf{W}_1^r(\vec{l}) \leq C_1, \quad (\lambda/\theta_2) \cdot \mathbf{W}_2^r(\vec{l}) \leq C_2. \end{aligned} \tag{17}$$

This problem can be equivalently expressed using the space of attainable workloads $\mathcal{W}^\times$ as

$$\begin{aligned} \min_{\mathbf{w}} \quad & (c_1/\theta_1) \cdot \mathbf{w}_1 + (c_2/\theta_2) \cdot \mathbf{w}_2 \\ \text{s.t. } & (\lambda/\theta_1) \cdot \mathbf{w}_1 \leq C_1, \quad (\lambda/\theta_2) \cdot \mathbf{w}_2 \leq C_2, \\ & \mathbf{w} \in \mathcal{W}^\times. \end{aligned} \tag{18}$$

Theorem 4 establishes that the solution structure for each type remains as in the single-type problem, but with a scaled cost ratio determined by the type-specific holding costs and by which station's capacity constraint is binding.

THEOREM 4 **(multi-type service anatomy)**. *Fix capacities $C_1, C_2 \in \mathcal{C}_{\boldsymbol{\beta}}$. Let $\mathcal{R}_k^c = (c_2^k/\theta_2^k)/(c_1^k/\theta_1^k)$ denote the baseline relative cost for type k items. Let $\bar{l}^k(\cdot)$ denote the solution for type k items in Theorem 2 as a function of the relative cost. Then the unique solution $\vec{l}^*$ to (17) satisfies*

$$l^{k,*} = \bar{l}^k(\widetilde{\mathcal{R}}_k^c), \quad k \in [K],$$

where

$$\widetilde{\mathcal{R}}_k^c = \mathcal{R}_k^c \left(\frac{c_2^k + \kappa_2}{c_2^k}\right) \Bigg/ \left(\frac{c_1^k + \kappa_1}{c_1^k}\right),$$

and $\kappa = (\kappa_1^*, \kappa_2^*)$ are such that $\kappa_i^*\Big(C_i - \mathcal{L}_i(\vec{l}^*)\Big) = 0$, $i = 1, 2$. If $C_1 = \infty$, the corrected cost $\widetilde{\mathcal{R}}_k^c$ is decreasing in C_2, and it is increasing in C_1 if $C_2 = \infty$.

The dual variables κ_1 and κ_2 for the resource constraints scale the effective cost ratios up or down. If the capacity of Station 1 is strictly binding but Station 2 has positive slack (so $\kappa_2 = 0$), the relative cost is reduced for all item types. The first (second) constraint, if binding (hence $\kappa_1 > 0$ ($\kappa_2 > 0$)), decreases the processing time in Station 1 (Station 2) of all item types.

With $\beta^k < 1$, the optimal processing time in Station 1 (Station 2) for type k declines as the corresponding constraint becomes more binding. With $\beta^k = 1$, the scaled relative cost can drive certain item types to a regime where it is optimal to perform all processing in only one of the two stations. This is because $\beta^k = 1$ supports three operating regimes, and there exist effective second-station costs for which it is optimal to use only a single station; recall Theorem 2 and see the example further below.

For given $\mathcal{R}_k^c$, θ_1^k, and θ_2^k, the optimal processing time in Station 1 decreases more (relative to the unconstrained baseline) for those item types with smaller (larger) marginal Station 1 (Station 2) costs, c_1^k (c_2^k). This is because $\widetilde{\mathcal{R}}_k^c$ effectively represents a discount in the relative Station 2 (Station 1) cost. For this same reason, two item groups, say a and b, with unconstrained solutions satisfying $\bar{l}_a^*(\mathcal{R}_a^c) > \bar{l}_b^*(\mathcal{R}_b^c)$ might have $\bar{l}_a^*(\widetilde{\mathcal{R}}_a^c) < \bar{l}_b^*(\widetilde{\mathcal{R}}_b^c)$. This means that the "discount" is more significant for type a. With $c_1^a = c_2^b$ and $\mathcal{R}_a^c = \mathcal{R}_b^c$, such a "swap" can still occur because of differences in substitution/complementarity. For a group

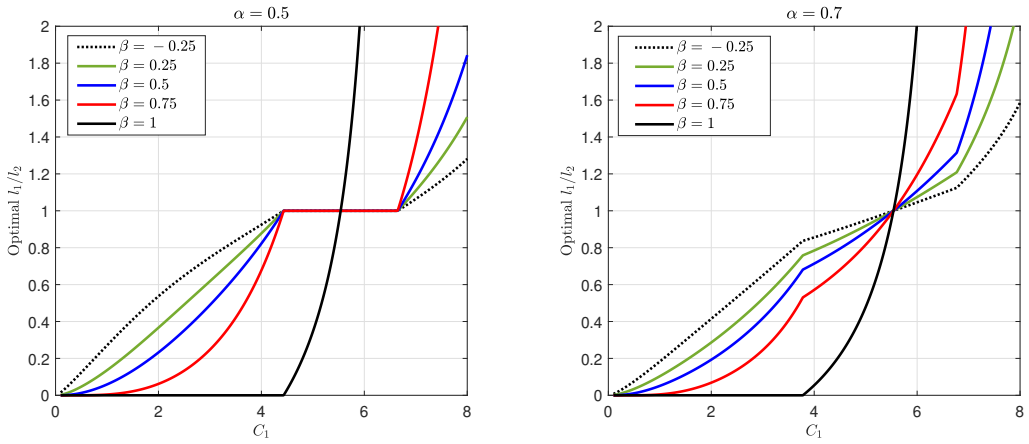
with a higher degree of substitution (i.e., $\beta_a > \beta_b$), a large Station 2 discount shifts more effort toward Station 2.

Maintenance Example. Consider a system that handles multiple categories of equipment that differ in maintenance costs and in how strongly repair and preventive service complement each other. When capacity is scarce, items with stronger complementarity between the two stages (lower β) retain relatively more repair effort, since cutting repair capacity sharply reduces the effectiveness of subsequent preventive service. By contrast, for items that are more substitutable, the system can shift a greater share of work toward preventive service without a substantial loss in performance.

For more on the role of complementarity and substitution, Figure 8 provides two numerical examples. We consider $K = 5$ different levels of substitution/complementarity. When there is enough capacity, all items receive their required effort mix. As capacity becomes scarce, processing time in Station 1 decreases and processing time in Station 2 increases. When $C_1 = 5.5$, the efforts in both stations are the same: $l_1^{k,*}/l_2^{k,*} = 1$ for all $k \in [5]$ (i.e., regardless of the value of β^k). When $C_1 < 5.5$, the stronger the complementarity (smaller β^k), the larger the ratio $l_1^{k,*}/l_2^{k,*}$ is. That is, items with stronger complementarity are prioritized (stay longer in Station 1 and less in Station 2) over items with a smaller level of complementarity. This is because it is easier, for types with greater substitution (larger β^k), to compensate for the shorter processing time in Station 1 by moving effort to Station 2. Furthermore, only under substitution can we reduce $l_1^{k,*}$ to zero and provide all the required processing in Station 2. Indeed, when $C_1 \leq 4.5$ (if $\alpha = 0.5$) and $C_1 \leq 3.8$ (if $\alpha = 0.7$), we see that $l^{k,*} = 0$ under perfect substitution ($\beta^k = 1$). Under complementarity, we must keep the items, even for a short period of time, in Station 1. Note that when $C_1 \geq 6.5$ (if $\alpha = 0.5$) and $C_1 \geq 6.8$ (if $\alpha = 0.7$), $l_2^{k,*} = 0$ under perfect substitution—the entire service is provided in Station 1.

In Appendix B.5 we consider the case where resources can be shared between the stations. In this case, the individual-station constraints are replaced with a total consumption constraint $\mathcal{L}(\vec{l}) = (\lambda/\theta_1) \cdot \mathbf{W}_1^r(\vec{l}) + (\lambda/\theta_2) \cdot \mathbf{W}_2^r(\vec{l}) \leq C$. In that setting, the interesting question is how the total capacity is optimally allocated between the stations. In Appendix B.6 we consider the throughput maximization problem.

Figure 8 **Optimal l_1^k/l_2^k, $k \in [5]$, where each k corresponds to a different value of β. The parameters are $\varrho = 1$, $C_2 = \infty$, $\theta_1 = 1$, $\theta_2 = 0.6$, $\lambda = 1$, $c_1 = c_2 = 1$. For $\alpha = 0.5$ in this example, capacity is binding if $C_1 \leq 10$ and/or $C_2 \leq 2.8$; for $\alpha = 0.7$, it is binding at $C_1 \leq 9.7$ or $C_2 \leq 1.16$.**



6. Synthesis: The Operations Frontier

The key forces that our model seeks to capture are present in processing networks where there is a clear notion of repair and maintenance activities. Such systems often involve sequential tasks—*repair followed by preventive maintenance while the equipment remains offline*—where greater investment in the first stage restores functionality, and additional preventive work in the second stage enhances long-term reliability and reduces future re-work, reflecting the trade-offs captured by our model. In such systems, the planner must decide how much effort to allocate to each activity. Examples include manufacturing (where the second station represents quality assurance) and services (such as healthcare, where the second station might represent post-acute care follow-up). Common to these examples is that effective operation requires allocating effort across stations and item types while accounting for process characteristics such as cost and capacity.

The model abstracts from the specifics of particular settings to focus on the role of, and interaction between, cost and substitution-complementarity. The goal of this section is to highlight and consolidate key insights from the model that, we believe, are robust across various settings.

Re-work, Workload, and the Operations Frontier. A fundamental feature of the model – and the reality it seeks to capture – is that greater effort (longer processing time) results in a higher immediate workload but has the potential to reduce total workload by minimizing

returns/re-work. As such, the focus (both managerially and mathematically) should be on workload rather than immediate service times.

The boundary of the attainable workload region is the *operations frontier* for such systems and represents the trade-off between immediate and long-term workload. This trade-off, regardless of the level of complementarity, is strictly convex – except in the case of perfectly symmetric contribution ($\alpha = 1/2$).

From a practical perspective, the non-trivial nature of this trade-off implies that capacity exchanges are not "one-for-one." Reducing capacity by Δ at one station may require adding more than Δ capacity at the second station to maintain the same performance level.

Operating Regimes and the Effect of Re-work. There are three broad regimes: two "corner" regimes where only one station is utilized, and an interior *network* regime where both stations are (optimally) used for processing.

The optimal operating regime depends on item parameters, and as a result, different regimes may coexist within the same system. For items processed exclusively at Station 1, the release threshold from the system coincides with the transfer threshold. Conversely, items processed solely at Station 2 should be directed there immediately upon arrival.

For items exhibiting (even minimal) complementarity between stations, both stations are required. Corner regimes arise only for item types where the stations are perfect substitutes. Even in these cases, unless the relative costs are significantly skewed toward one station, the optimal solution lies in the interior regime.

Cost and Substitution Interaction: A Threshold Determines the Most Effective Station. The level of complementarity influences the optimal allocation of effort between stations. For "low" relative costs (Station 1 to Station 2), greater complementarity leads to more effort being allocated to Station 1. Conversely, for "high" relative costs, greater complementarity shifts effort towards Station 2. "Low" and "high" costs are defined relative to a threshold determined by relative contribution and substitution/complementarity levels.

This threshold marks the point where the "most effective" station transitions from Station 1 to Station 2. Below this threshold, increasing interdependence between stations (through greater complementarity) results in more effort being directed to Station 1. Above the threshold, additional effort is allocated to Station 2.

The Effect of Capacity: Shifting the Threshold. Finite capacity alters the relative costs between stations. A tighter constraint at Station 1 raises its relative cost. Our results

suggest that capacity adjustments will shift the balance of effort, depending on where item types fall relative to the adjusted threshold. For instance, if capacity at Station 1 is reduced, effort for item types near the original unconstrained threshold will shift towards Station 2.

Expanding the Model's Coverage. Where the specifics of a setting can be incorporated as constraints—such as minimum processing requirements at Station 1—the attainable workload region remains a valuable tool for studying the optimal network configuration. As indicated in §C, it is feasible to estimate the model parameters using standard processing-network data to inform design decisions.

7. Concluding Remarks and Directions for Future Research

The two-station integration model is, in general terms, a service network design problem. The anatomy of the network translates into the allocation of work across networked stations. We take a modeling approach that captures the individual quality evolution through the network and the way in which the station-wise quality determines re-work through substitution and complementarity. Optimizing the service anatomy in this case means setting the transfer targets (processing times) or efforts invested in each station to minimize the total cost of operating the system; this also implies minimizing re-work and its associated cost. We introduce a model that *operationalizes* item characteristics and subsequently study how the latter determine the allocation of both stations' efforts.

The baseline model can be expanded to include some realistic features. In our model, the transfer target l does not vary from one visit to the next. In some cases, the very event of re-work reveals information about the item. It then makes sense to have different transfer targets in subsequent (i.e., post-index) visits. The model can be expanded to capture more elaborate process protocols/networks.

The workload allocation is the *first-order* or "fluid" optimization problem. As is the case with the fluid baseline in various queueing control problem, our solution can serve as a stepping stone for the development of dynamic control policies. Our allocation of efforts between the stations provides a baseline. Using it, real-time performance can be improved by reconciling the transfer targets with the workload. The optimal control would introduce state-dependent discharge or transfer actions that are perturbations of the optimal fluid decisions.

References

Armony, M., C.W. Chan, B. Zhu. 2018. Critical care capacity management: Understanding the role of a step down unit. *Production and Operations Management* **27**(5) 859–883. 8

Arrow, K.J., H.B. Chenery, B.S. Minhas, R.M. Solow. 1961. Capital-labor substitution and economic efficiency. *The Review of Economics and Statistics* 225–250. 13

Batt, R.J., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. *The Wharton School, the University of Pennsylvania, Philadelphia, PA* **19104**. 7

Bavafa, H., L. Örmeci, S. Savin, V. Virudachalam. 2022. Surgical case-mix and discharge decisions: Does within-hospital coordination matter? *Operations Research* **70**(2) 990–1007. 8, 44

Berry Jaeker, J.A., A.L. Tucker. 2017. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* **63**(4) 1042–1062. 7

Bertsekas, D.P. 1997. *Nonlinear Programming*. Taylor & Francis. 61, 63

Bhattacharya, R.N., E.C. Waymire. 2009. *Stochastic Processes with Applications*. SIAM. 43

Carey, K. 2015. Measuring the hospital length of stay/readmission cost trade-off under a bundled payment mechanism. *Health economics* **24**(7) 790–802. 10

Carvalho, V.M., A. Tahbaz-Salehi. 2019. Production networks: A primer. *Annual Review of Economics* **11** 635–663. 4

Chan, T.C.Y, S.Y. Huang, V. Sarhangian. 2024. Dynamic control of service systems with returns: Application to design of postdischarge hospital readmission prevention programs. *Operations Research* . 8

Chen, C., Z. Jia, P. Varaiya. 2001. Causes and cures of highway congestion. *IEEE Control Systems Magazine* **21**(6) 26–32. 7

Clark, J.R., R.S. Huckman. 2012. Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Science* **58**(4) 708–722. 7

Committee, American Wind Energy Association (AWEA) Operations & Maintenance. 2017. Chapter 8: Condition-based maintenance. Operations & Maintenance Recommended Practices RP 801 et seq., American Wind Energy Association. URL https://cleanpower.org/wp-content/uploads/2024/06/AWEA-OM-RP-Chapter-8_Condition-Based-Maintenance.pdf. Accessed November 2025. 52

Dai, J.G., G. Weiss. 1996. Stability and instability of fluid models for reentrant lines. *Mathematics of Operations Research* **21**(1) 115–134. 8

Facts, Wind Energy – The. 2024. Commissioning, operation and maintenance for wind turbines. https://wind-energy-the-facts.org/commissioning-operation-and-maintenance.html. Accessed November 2025. 52

Fujiy, B.C., D. Ghose, G. Khanna. 2022. Production networks and firm-level elasticities of substitution. Tech. rep., Technical report, Working Paper. 14

Harrison, J.M. 2003. Brownian models of open processing networks: Canonical representation of workload. *The Annals of Applied Probability* **13**(1) 390–393. 7

Harrison, J.M., J.A. Van Mieghem. 1997. Dynamic control of brownian networks: State space collapse and equivalent workload formulations. *The Annals of Applied Probability* 747–771. 7

Hashimoto, E.M., E.M.M. Ortega, G.M. Cordeiro, V.G. Cancho, I. Silva. 2023. The re-parameterized inverse gaussian regression to model length of stay of COVID-19 patients in the public health care system of Piracicaba, Brazil. *Journal of Applied Statistics* **50**(8) 1665–1685. 49

Henningsen, A., G. Henningsen, G. Literáti. 2021. Econometric estimation of the" constant elasticity of substitution" function in R: The miceconCES package. *Handbook of Research Methods and Applications in Empirical Microeconomics*. Edward Elgar Publishing, 596–640. 14

Hopp, W.J., S. Iravani, G.Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61–77. 7

Hwang, B.-G., S.R. Thomas, C.T. Haas, C.H. Caldas. 2009. Measuring the impact of rework on construction cost performance. *Journal of construction engineering and management* **135**(3) 187–198. 2

Jiang, HJoanna, Molly Hensche. 2023. Characteristics of 30-day all-cause hospital readmissions, 2016-2020. *HCUP Statistical Brief* **304**. 2

Kc, D., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498. 10

Kc, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50–65. 7

Koesler, S., M. Schymura. 2015. Substitution elasticities in a constant elasticity of substitution framework–empirical estimates using nonlinear least squares. *Economic Systems Research* **27**(1) 101–121. 14

Kum G., Iwimbong, J.P. Epane, J.J. Shen, X. Goodman, Z. Ramamonjiarivelo, F.D. Zengul. 2024. Systematic review and meta-analysis of the financial impact of 30-day readmissions for selected medical conditions: A focus on hospital quality performance. *Healthcare*, vol. 12. MDPI, 750. 2

Kumar, P.R. 1993. Re-entrant lines. *Queueing Systems* **13**(1) 87–110. 8

Love, P.E.D., H. Li. 2000. Quantifying the causes and costs of rework in construction. *Construction management & economics* **18**(4) 479–490. 2

Madan, K.C. 2000. An M/G/1 queue with second optional service. *Queueing Systems* **34**(1) 37–46. 8

Mahaboob, B., B. Venkateswarlu, J.R. Sankar. 2017. Estimation of parameters of constant elasticity of substitution production functional model. *IOP Conference Series: Materials Science and Engineering*, vol. 263. IOP Publishing, 042121. 14

Mas-Colell, A., M.D. Whinston, J.R. Green. 1995. *Microeconomic Theory*, vol. 1. Oxford university press New York. 5, 13

Medicare and Medicaid Statistical Supplement. 2007. Centers for Medicare and Medicaid Services. 2

Nambiar, S., M. Mayorga, M. Capan. 2020. Resource allocation strategies under dynamically changing health conditions. *Wrking paper* . 8

Netessine, S., F. Zhang. 2005. Positive vs. negative externalities in inventory management: Implications for supply chain design. *Manufacturing & Service Operations Management* **7**(1) 58–73. 9

Sato, K. 1967. A two-level constant-elasticity-of-substitution production function. *The Review of Economic Studies* **34**(2) 201–218. 13

Sato, R., T. Koizumi. 1973. On the elasticities of substitution and complementarity. *Oxford Economic Papers* **25**(1) 44–56. 13

Schiffauerova, A., V. Thomson. 2006. A review of research on cost of quality models and best practices. *International Journal of Quality & Reliability Management* **23**(6) 647–669. 2

Shi, Pengyi, Jonathan E Helm, Jivan Deglise-Hawkinson, Julian Pan. 2021. Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Operations Research* **69**(6) 1842–1865. 8

Staats, B.R., F. Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management Science* **58**(6) 1141–1159. 7

Stern, D.I. 2011. Elasticities of substitution and complementarity. *Journal of Productivity Analysis* **36**(1) 79–89. 13

Tan, T.F., S. Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593. 7

U.S. Bureau of Labor Statistics. 2024. Occupational employment and wage statistics (oews), 2024 tables for soc 51-9061 and 49-9041. https://www.bls.gov/oes/. Accessed November 2025. 52

Varian, H.R. 2010. *Intermediate Microeconomics: A Modern Approach*. W. W. Norton & Company, New York. 5

Wang, X., R. Huang, J. Gao, L.G. Debo. 2021. Managing discretionary services: A review and research opportunities. *Journal of Systems Science and Systems Engineering* 1–16. 7

Whitmore, G.A. 1975. The inverse gaussian distribution as a model of hospital stay. *Health Services Research* **10**(3) 297. 49

Whitmore, G.A. 1979. An inverse gaussian model for labour turnover. *Journal of the Royal Statistical Society: Series A (General)* **142**(4) 468–478. 49

Wilson, S.F., B. Shorten, R.M.I Marks. 2005. Costing the ambulatory episode: Implications of total or partial substitution of hospital care. *Australian Health Review* **29**(3) 360–365. 9

Online Appendix

Appendix A: Notation

Conventions. Scalars are set in italic; 2D vectors (per type) use arrows. Superscript k indexes item types; subscript $i \in \{1,2\}$ indexes stations. The vector $e = (1,1)$.

Table 2: Main symbols and notation used in the paper.

Symbol	Type	Meaning / Where Used
Core primitives		
a	scalar	Initial quality score on first visit (single item).
λ	scalar	Exogenous base arrival rate to the network.
l_i	scalar	Threshold (release/transfer target) at Station i.
$l = (l_1, l_2)$	2D vector	Threshold pair for a single item (two-station model).
w_i	scalar	Rate-normalized workload at Station i.
$w = (w_1, w_2)$	2D vector	Workload pair for a single item (two-station model).
θ_i	scalar	Drift (quality improvement rate) at Station i.
σ_i	scalar	Diffusion coefficient at Station i.
σ_{pr}	scalar	Post-release diffusion coefficient.
$\mathcal{B}_t^i$	process	In-station quality BM at Station i (positive drift).
$\mathcal{B}_t^{pr}$	process	Post-release quality BM (drift $\eta(l)$, start $e{\cdot}l$).
Single-station objects		
l^*	scalar	Optimal threshold in the one-station setting.
$W_i^r(l)$	scalar	Rate-normalized workload at Station i for threshold(s) l.
τ	r.v.	Hitting time to threshold (inverse Gaussian under BM).
Post-release and re-work		
$e{\cdot}l$	scalar	Total discharge score $l_1 + l_2$ (starting level post-release).
$p(l)$	function $\to$ scalar	Re-work probability $p(l) = \exp\big(-(\nu + \varrho\,\eta(l)\,(e{\cdot}l))\big)$.
$N(l)$	function $\to$ scalar	Expected number of visits $N(l) = [1 - p(l)]^{-1}$.

Continued on next page

Symbol	Type	Meaning / Where Used
ν	scalar	Baseline log-scale parameter in the re-work probability; $e^{-\nu}$ is the base probability of no re-work.
ϱ	scalar	$\varrho := 2/\sigma_{pr}^2$
Two-station aggregation (CES) and geometry		
$\alpha \in (0,1)$	scalar	Station-importance weight (CES).
$\beta \leq 1$	scalar	CES substitution/complementarity parameter (β=1 subst.; $\beta \downarrow -\infty$ complementarity).
$\eta(l)$	function $\to$ scalar	CES aggregator: $(\alpha l_1^\beta + (1-\alpha) l_2^\beta)^{1/\beta}$.
$\mathcal{W}$	set	Attainable workload region in $\mathbb{R}_+^2$.
$f_{\alpha,\beta}(\cdot)$	function	Boundary of $\mathcal{W}$: $w_2 \geq f_{\alpha,\beta}(w_1)$.
$\mathbb{h}_{\alpha,\beta}(\cdot)$	function	Auxiliary function in boundary characterization.
w_1^0	scalar	Right-endpoint of nontrivial boundary segment of $\mathcal{W}$.
Optimization, costs, and capacity		
c_i	scalar	Holding/effort cost rate at Station i.
$V(l)$ or $V(\vec{l})$	scalar	Total expected cost objective.
$\mathcal{R}^c$	scalar	Relative cost ratio (effective Station 2 vs. Station 1).
$\widetilde{\mathcal{R}}^c$	scalar	Corrected cost ratio (accounts for duals κ).
C_i	scalar	Capacity at Station i.
$\mathcal{L}_i(\vec{l})$	scalar	Total load on Station i: $(\lambda/\theta_i)\, W_i^r(\vec{l})$.
κ_i	scalar	Dual variable for capacity at Station i.
Multi-type extension		
$k \in [K]$	index	Type index, $[K] = \{1, \ldots, K\}$.
$l^k = (l_1^k, l_2^k)$	2D vector	Thresholds for type k.
$\vec{l} = (l^1, \ldots, l^K)$	$2K$-vector	Stacked decision vector over all types.
$\theta_i^k,\ \sigma_i^k$	scalars	Drift and diffusion for type k at Station i.
$\alpha^k,\ \beta^k$	scalars	CES weight and parameter for type k.

Continued on next page

Symbol	Type	Meaning / Where Used
$\vec{w}^k = (w_1^k, w_2^k)$	2D vector	Workload pair for type k.
$\mathbf{w} = (\vec{w}^1, \ldots, \vec{w}^K)$	$2K$-vector	Stacked workload vector across types.
$\mathbf{W}_i^r(\vec{l})$	K-vector	$(W_i^{r,1}(l^1), \ldots, W_i^{r,K}(l^K))$, type workloads at Station i.
$\mathcal{W}_k$	set	Attainable workload set for type k.
$\mathcal{W}^\times$	set	Product set $\prod_{k\in[K]} \mathcal{W}_k$.
$\boldsymbol{\beta} = (\beta^1, \ldots, \beta^K)$	K-vector	Type-wise substitution/complementarity parameters.
$\mathcal{C}_{\boldsymbol{\beta}}$	set	Feasible capacity vectors under $\boldsymbol{\beta}$.

Appendix B: Variants and Extensions

To simplify notation, throughout the extensions we assume that $\nu = 0$.

B.1. Attainable Workload Region for n Stations

Consider the case of n stations and let $(l_1, \ldots, l_n)$ denote the transfer target from each of the n stations. The drift and diffusion coefficient of each Station i, $i = 1, \ldots, n$ are θ_i and σ_i. The constant elasticity of substitution (CES) production function for n stations is:

$$\eta(l) = \left[\sum_{i=1}^{n} \alpha_i l_i^{\beta}\right]^{\frac{1}{\beta}}, \quad 0 < \alpha_i < 1, \quad \beta \leq 1, \tag{19}$$

where $\sum_{i=1}^{n} \alpha^i = 1$; we use henceforth α for the probability vector $\alpha_1, \ldots, \alpha_n$. The rate-normalized workload in Station i is $W_i^r(l) = l_i N(l)$. Consider here the case where the initial state in the first visit is 0 as in subsequent visits.

The final part of Theorem 1 shows that the attainable workload region is given by

$$\mathcal{W} := \{w \in \mathbb{R}_+^n : \eta(w)(e \cdot w) \geq 1\}.$$

This, as for $n = 2$, is a convex region. This is formally argued within the proof of Theorem 1.

B.2. Processing Complementarity

The analysis so far focused on the case where the two process steps can have *outcome* complementarity, but there is no processing complementarity. That is, the improvement rate in one station does not depend on the quality score reached at the other station.

In this section we consider the other extreme: processing complementarity/substitution. First, we assume that process steps are perfect outcome substitutes and have the same contribution to the return probability ($\alpha = 0.5$ and $\beta = 1$). The number of visits is then

$$N_{PC}(l) = \left[1 - e^{-\frac{1}{2}(e \cdot l)^2}\right]^{-1}.$$

We introduce process dependence by having a non-constant drift in Station 2, a drift $\theta_2(l_1)$ that depends on the final score at completion in Station 1. We use $\theta_2(l_1) = \gamma + y l_1$. This is clearly *not* as general or flexible as what we allowed in outcome complementarity; however, it provides initial tractability. We then explore numerically variations on this structure.

Notice that if $y = 0$ and $\gamma > 0$ (so that, $\theta_2 \equiv \gamma > 0$), we restore perfect processing substitution as in our base model. If $\gamma = 0$ and $y > 0$, then $\theta_2(l_1)$ is strictly increasing, and the two steps are strong complements in the sense that one cannot commence processing in Station 2 unless some processing is done in step 1 ($l_1 > 0$).

It will be useful, for clarity, to fix $\theta_1 = 1$. In that case, l_1 is not only the target but also the mean processing time in Station 1.

The objective function takes the form

$$V(l) = \left(c_1 l_1 + \frac{c_2}{\theta_2(l_1)} l_2 \right) N_{PC}(l) \ .$$

Instead of considering l as the decision, let us define

$$\tau_1 = l_1, \quad \tau_2 = l_2/\theta_2(\tau_1).$$

Then, with $\tau = (\tau_1, \tau_2)$, and (re)defining

$$N_{PC}(\tau) = \left[1 - e^{-\frac{1}{2}(\tau_1 + \theta_2(\tau_1)\tau_2)^2} \right]^{-1},$$

we write

$$V(\tau) = (c_1 \tau_1 + c_2 \tau_2) N(\tau).$$

This makes clear that the optimization problem with processing dependence is the same as one with outcome dependence but where the decisions are the processing times rather than the target, and where the processing times impact returns through the sum $\tau_1 + \theta_2(\tau_1)\tau_2$. We define rate-normalized workloads as before

$$W_1(\tau) = \tau_1 N_{PC}(\tau), \text{ and } W_2(\tau) = \theta_2(\tau_1)\tau_2 N_{PC}(\tau).$$

The following is a characterization of the attainable workload region.

LEMMA 4. *Suppose that $\theta_2(\tau_1) = \gamma + y\tau_1$ with either $\gamma = 0$, $y > 0$ or $\gamma > 0$, $y = 0$. In either case, the region*

$$\mathcal{W} = \{(w_1, w_2) \geq 0 : \exists \tau \quad s.t.\ W_1(\tau) = \tau_1 N_{PC}(\tau), \quad W_2(\tau) = \theta_2(\tau_1)\tau_2 N_{PC}(\tau)\},$$

is equivalently given by

$$\mathcal{W} = \{(w_1, w_2) \geq 0 : \quad w_2 \geq f_{\alpha,\beta}(w_1)\}$$

where

$$f_{\alpha,\beta}(w_1) = \begin{cases} a - b w_1, & \text{if } w_1 \leq b/a, \\ 0, & \text{otherwise,} \end{cases}$$

where a, b are constants given explicitly in terms of γ, y and Γ from (7).

Thus, in either extreme case for our model, the attainable workload region is a polyhedron (as in the case of outcome complementarity with $\beta = 1$ and $\alpha = 1/2$). The difference is in the objective function. Given $w = (w_1, w_2) \in \mathcal{W}$, let $\tau = \tau(w)$ be such that $w_1 = \tau_1 N_{PC}(\tau)$ and $w_2 = \tau_2 N_{PC}(\tau)$. Then, the objective function is given by

$$c_1 w_1 + c_2 \frac{w_2}{\theta_2(\tau_1(w))}.$$

In the case that $\gamma > 0$, $y = 0$, we have $\theta_2(\cdot) \equiv \gamma$ and thus, a linear objective function. Therefore, there exists a threshold of cost ratio where the optimal solution is of the form $(0, w_2)$ when the cost ratio is below the threshold and of the form $(w_1, 0)$ beyond the threshold.

If, instead $\theta_2(\tau_1(w)) = y\tau_1(w)$, the objective function is non-linear. Moreover, because $w_1 = \tau_1 N(\tau) \geq \tau_1$, for any $w_2 > 0$

$$c_1 w_1 + c_2 \frac{w_2}{\theta_2(\tau_1(w))} \geq c_1 w_1 + c_2 \frac{w_2}{y w_1} \uparrow \infty, \text{ as } w_1 \downarrow 0.$$

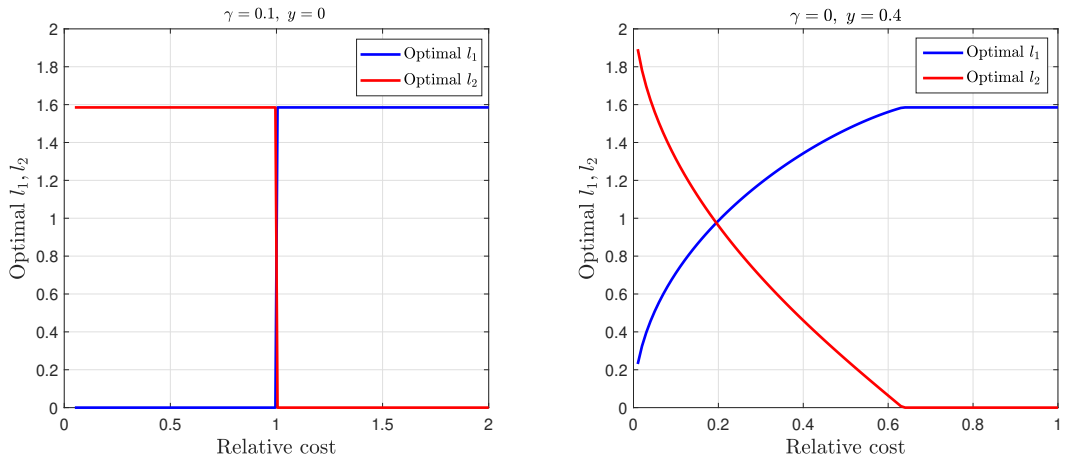
In this scenario, then, we cannot have $(0, w_2)$ (for any $w_2 > 0$) as an optimal solution. Recalling that $w_2 \geq \max\{a - bw_1, 0\}$ for any $w \in \mathcal{W}$, we further have that for any such w, the optimal cost is lower bounded by

$$\begin{cases} c_1 w_1 + c_2 \frac{a}{y w_1} - \frac{b}{y}, & \text{if } w_1 < a/b \\ c_1 \frac{a}{b}, & \text{otherwise} \end{cases}$$

The first line is minimized at $w_1 = \sqrt{\frac{c_2}{c_1}\frac{a}{y}}$ at the value $2\sqrt{c_1 c_2 a/y} - b/y$. For all c_2 large enough, this is greater than $c_1 a/b$. We arrive, therefore, at the conclusion that an optimal solution of the form $(w_1, 0)$ does exist in this case.

Figure 9 is a summary of these derivations. It shows that, in both cases, at least one corner solution exists. This suggests that process complementarity of this form is somewhat weaker that outcome complementarity in that it does not force an optimal solution where both stations have non-negligible workload.

Figure 9 Optimal solution with $\theta_2(l_1) = \gamma + y l_1$. The parameters are $\varrho = 1$, $\theta_1 = 1$ and $c_1 = 1$.

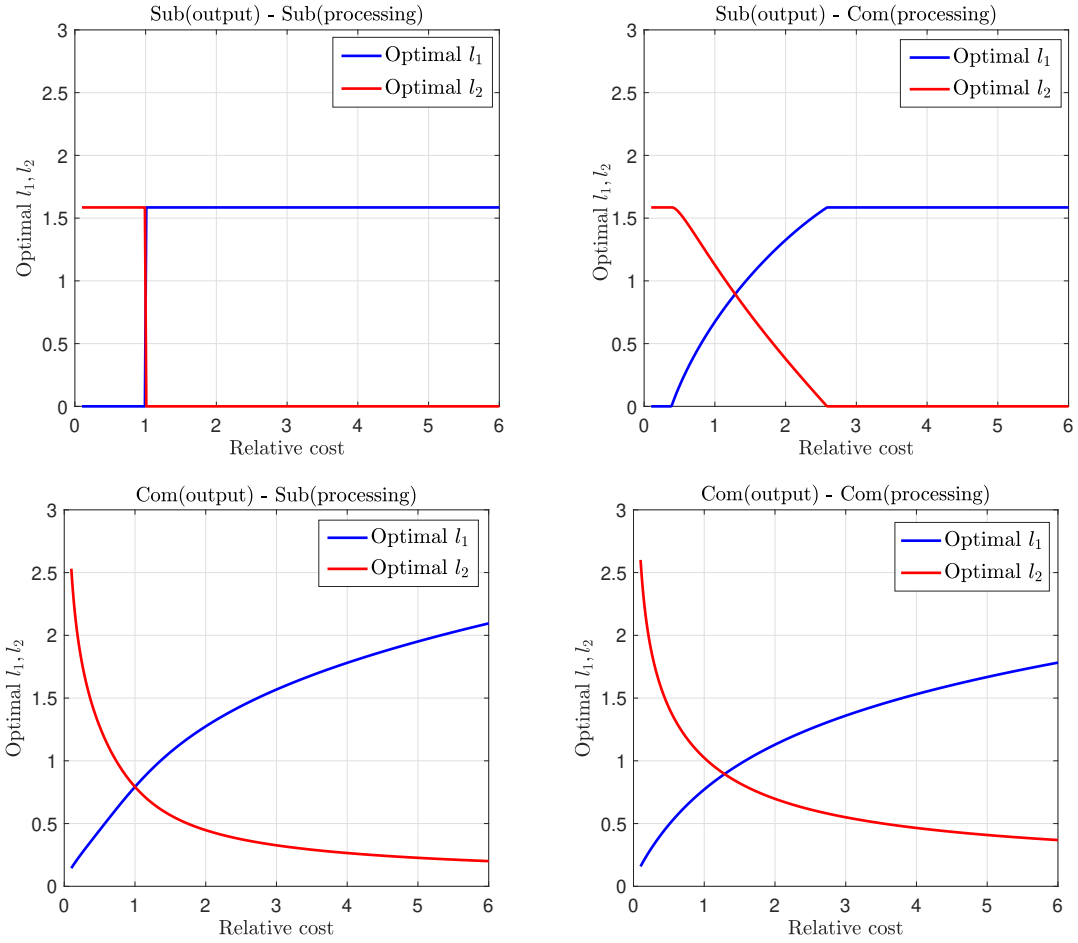


Let us consider some numbers to substantiate this claim. We restrict our attention here to the case where $\theta_2(l_1) = \gamma + y l_1$. The output complementarity is determined by the value of β, while the processing

complementarity is determined by the value of y. We consider four combinations: output and processing substitution; output and processing complementarity; output substitution and processing complementarity; and output complementarity and processing substitution. Figure 10 is a visualization of the optimal decisions.

We observe that the processing substitution/complementary mitigates the effect of the output substitution/complementarity. Under output substitution and processing complementarity (top right plot), the optimal l_1 and l_2 can be zero, while under output complementarity and processing substitution (bottom left plot), these cannot be zero. This suggests that outcome substitution/complementarity has a stronger effect than processing substitution/complementarity over the total level of substitution/complementarity.

Figure 10 **Optimal l_1, l_2 for different outcome/processing substitution/complementarity. The parameters are $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, $c_1 = c_2 = 1$; for the outcome substitution/complementarity: $\alpha = 0.5$, $\gamma = 0.1$, $\beta = 1$ and $\beta = 0.25$; for the processing substitution/complementarity: $y = 0$ and $y = 0.1$.**



Numerical results for the mixed model $\gamma, y > 0$ with multiple types and finite capacity show that the optimal solution exhibits similar structural properties to the case of outcome substitution/complementarity.

B.3. Incorporating Re-Work Probability Up to a Predefined Time

The re-work probability, as defined in (2), is the probability of a positive-drift BM, $\mathcal{B}_t^{pr}$, starting at l_1+l_2 to hit 0 in finite time. In this section we study the case where the re-work probability is the probability of $\mathcal{B}_t^{pr}$ to hit 0 in a predefined time $T>0$.

The first time passage distribution at 0 of $\mathcal{B}_t^{pr}$ that starts at l_1+l_2 has the possibly defective PDF (see, e.g., Bhattacharya and Waymire 2009, Chapter 1, page 27):

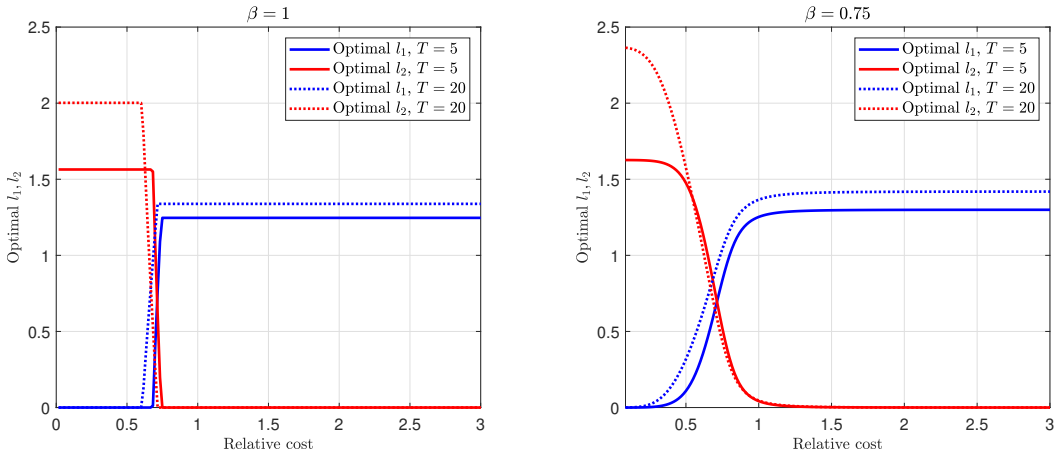
$$f_p(t)=\frac{l_1+l_2}{\sqrt{2\pi}\sigma_{pr}t^3}\exp\left(-\frac{(-l_1-l_2-\eta(l)t)^2}{2\sigma_{pr}^2 t}\right).$$

The re-work probability $p_T(l)$ as a function of T can hence be calculated by

$$p_T(l)=\int_0^T f_p(t)\mathrm{d}t.$$

Figure 11 illustrates the solution structure for different values of T. We derived it by numerically calculating $p_T(l)$, and solving (12) when replacing $p(l)$ with $p_T(l)$. We observe that both stations' efforts increase with T. This is because for a fixed effort mix, the re-work probability increases with T; to overcome this, the discharge threshold l_1+l_2 is increased. The solution structure, however, remains the same as our base model with $p(l)$.

Figure 11 **Solution structure when the re-work probability is up to time T. The parameters are $\varrho=1$, $\theta_1=1$, $\theta_2=0.6$, $c_1=1$, $\alpha=0.7$.**



The results are basically identical to our base case and this can be explained by alluding to the attainable workload region foundation of our earlier results. Specifically, let us define, as before, $N_T(l)=(1-p_T(l))^{-1}$ and $W_i(l)=l_iN(l)$. Fixing $z\in[0,\infty)$, we have that, with the constraint that $w_2/w_1=z$,

$$w_1\geq F_p(z):=\min_{l_1} l_1N_T(l_1,zl_1). \tag{20}$$

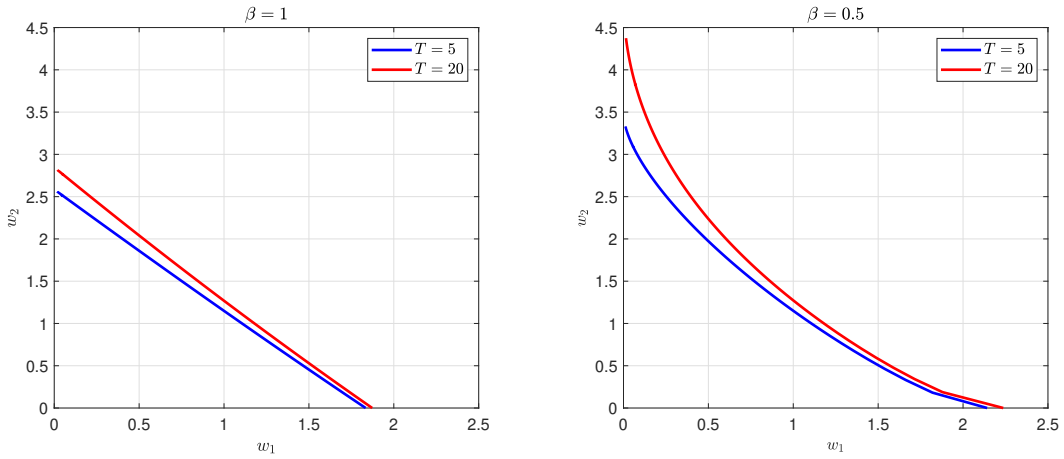
This function plays the role of $\Gamma/\mathbb{h}_{\alpha,\beta}(z)$ in our baseline model. The constraint on w_2 is $w_2\geq w_1F_p^{-1}(w_1)$.

Thus, the attainable workload region is given by

$$\mathcal{W}:=\{(w_1,w_2)\geq 0:\quad w_2\geq w_1F_p^{-1}(w_1)\}.$$

This feasible region is depicted in Figure 12 for $\alpha = 0.7$ and two different values of β. As before with $\beta = 1$, the feasible region "hits" the axis at a non-trivial angle; it asymptotes to the axes for $\beta < 1$. As in the baseline model, we end up with three cost regions when $\beta = 1$ and only interior solutions if $\beta < 1$; see Figure 11.

Figure 12 Feasibility region for different T.



B.4. Incorporating Quadratic Holding Costs and Variance in Processing Times

In this section we explore the solution under a quadratic holding cost structure, which brings up the recovery variance of coefficient, σ_1^2 and σ_2^2, in Stations 1 and 2, respectively. This variant is inspired by Bavafa et al. (2022).

Recall that $\tau_1(0, l_1)$ and $\tau_2(l_1, l_1 + l_2)$ correspond to the (random) length of stay in Stations 1 and 2, respectively. Consider the following holding cost functions:

$$C_1(l_1) = c_1 \left(\tau_1(0, l_1)\right)^2, \quad C_2(l_2) = c_2 \left(\tau_2(l_1, l_1 + l_2)\right)^2,$$

for some $c_1, c_2 > 0$. Standard results for the hitting time of Brownian motions yield

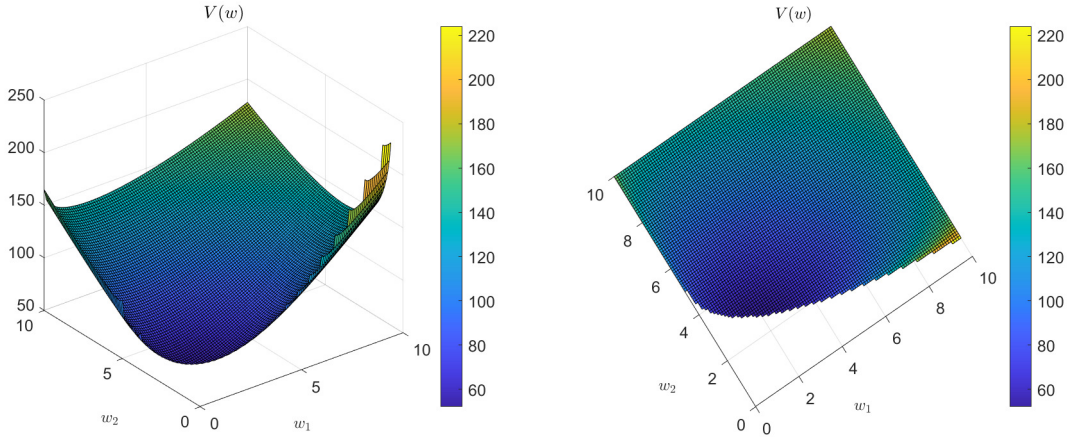
$$\mathbb{E}\left(\tau_1(a, l_1)\right)^2 = \frac{l_1 - a}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1 - a}{\theta_1}\right)^2; \quad \mathbb{E}\left(\tau_1(0, l_1)\right)^2 = \frac{l_1}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1}{\theta_1}\right)^2; \quad \mathbb{E}\left(\tau_2(l_1, l_1 + l_2)\right)^2 = \frac{l_2}{\theta_2^3}\sigma_2^2 + \left(\frac{l_2}{\theta_2}\right)^2.$$

The total journey cost for an item, as a function of the decision $l = (l_1, l_2)$, is, therefore, given by

$$\begin{aligned} V(l) &= c_1 \left[\frac{l_1 - a}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1 - a}{\theta_1}\right)^2\right] + (N(l) - 1)c_1 \left[\frac{l_1}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1}{\theta_1}\right)^2\right] + N(l)c_2 \left[\frac{l_2}{\theta_2^3}\sigma_2^2 + \left(\frac{l_2}{\theta_2}\right)^2\right] \\ &= c_1 \left[\frac{l_1 - a}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1 - a}{\theta_1}\right)^2\right] - c_1 \left[\frac{l_1}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1}{\theta_1}\right)^2\right] + N(l)\left(c_1 \left[\frac{l_1}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1}{\theta_1}\right)^2\right] + c_2 \left[\frac{l_2}{\theta_2^3}\sigma_2^2 + \left(\frac{l_2}{\theta_2}\right)^2\right]\right) \\ &= N(l)\left(c_1 \left[\frac{l_1}{\theta_1^3}\sigma_1^2 + \left(\frac{l_1}{\theta_1}\right)^2\right] + c_2 \left[\frac{l_2}{\theta_2^3}\sigma_2^2 + \left(\frac{l_2}{\theta_2}\right)^2\right]\right) - c_1 \left(\frac{a}{\theta_1^3} - \frac{a^2 - 2l_1 a}{\theta_1^2}\right). \end{aligned}$$

Similarly to the linear-cost case, the workload set can be constructed, so that $v(w)$ is the objective function in terms of the rate-normalized workload. Figure 13 illustrates the objective function $V(w)$ as a function of

Figure 13 **Solution structure for quadratic costs for different coefficient variances. The parameters are** $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, $c_1 = 1$, $\alpha = 0.7$, $\beta = 0.5$, $\sigma_1 = \sigma_2 = 1$.



w_1 and w_2. The right plot is a top-down view that makes evident the convex boundary of the attainable workload region.

Figure 14 illustrates the solution structure when $\sigma_1/\sigma_2 = 4$ and when $\sigma_1/\sigma_2 = 1/4$. The variability has the expected effect of increasing the cost. That is, if the variability in Station 1 grows (with all else remaining the same), more work will be allocated to Station 2. The results, however, do not fundamentally change compared to our baseline linear cost case.

Figure 14 **Solution structure for quadratic costs for different coefficient variances. The parameters are** $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, $c_1 = 1$, $\sigma_1 = 1$, $a = 0$.

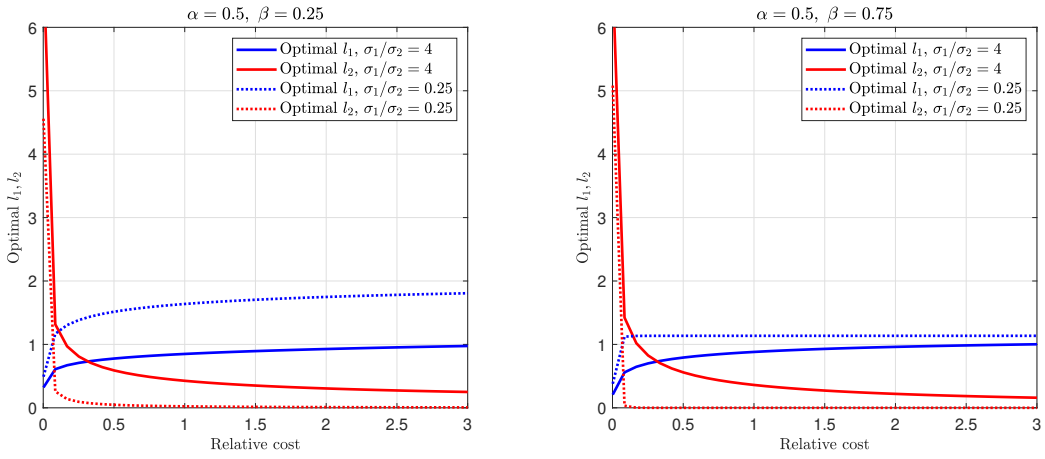
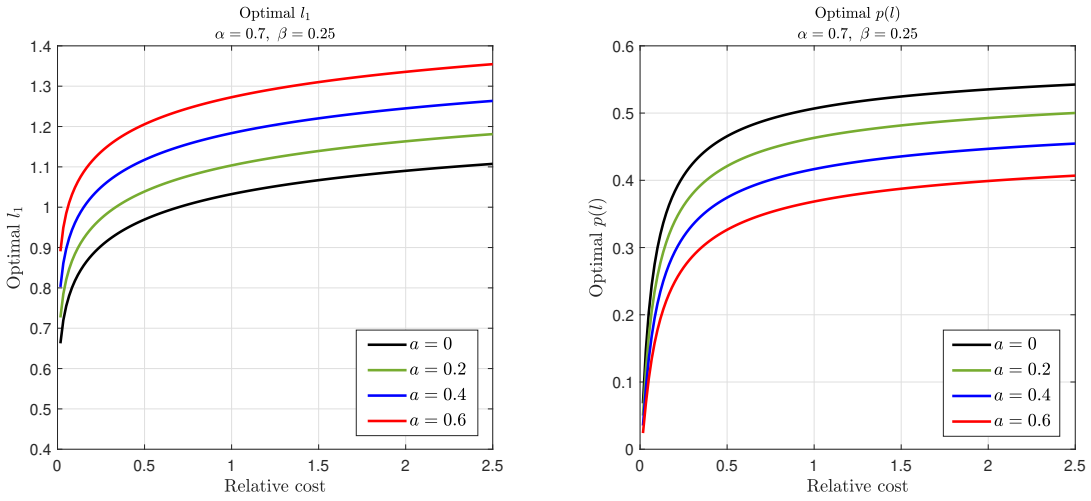


Figure 15 illustrates the optimal transfer threshold l_1 and the optimal re-work probability $p(l)$ for different initial scores. We observe that as the initial score improves, the transfer threshold to Station 2 increases. This is because the cost function is $(l_1 - a)^2$, and its derivative is $2(l_1 - a)$. The higher the value of a, the smaller the cost derivative; hence, you might actually increase l_1 further in the first station and do less in

the second station to save on costs. Furthermore, the re-work probability decreases with an improved initial score; as items arrive with a better quality score, we can provide a less costly and more effective service, resulting in a smaller likelihood of re-work.

Figure 15 **Optimal l_1 and $p(l)$ for different initial scores. The parameters are $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, $c_1 = 1$, $\sigma_1 = 1$.**



B.5. Limited Capacity with Resource Sharing

In many systems some resources can be shared between the stations. In this case, we face a constraint of the form

$$\mathcal{L}(\vec{l}) = (\lambda/\theta_1) \cdot W_1^r(l) + (\lambda/\theta_2) \cdot W_2^r(l) \leq C$$

We, therefore, want to simultaneously set the the stations' effort mix and the resource sharing between the stations.

Let C denote the total resource amount for both stations. We consider the following resource sharing problem:

$$\begin{aligned} \min_{\vec{l} \geq 0} \quad & (\lambda c_1/\theta_1) \cdot W_1^r(l) + (\lambda c_2/\theta_2) \cdot W_2^r(l) \\ \text{s.t.} \quad & (\lambda/\theta_1) \cdot W_1^r(l) + (\lambda/\theta_2) \cdot W_2^r(l) \leq C. \end{aligned} \tag{21}$$

For a vector (of pairs) $\mathbf{w} = ((w_1^1, w_2^1), \ldots (w_1^K, w_2^K))$, let $\mathbf{w}_1 = (w_1^1, \ldots, w_1^K)$ be the sub-vector that has the first coordinate from each pair and similarly define $\mathbf{w}_2$ for the second coordinate. Then, (21) is equivalently formulated as

$$\begin{aligned} \min \quad & (\lambda c_1/\theta_1) \cdot \mathbf{w}_1 + (\lambda c_2/\theta_2) \cdot \mathbf{w_2} \\ \text{s.t.} \quad & (\lambda/\theta_1) \cdot \mathbf{w}_1 + (\lambda/\theta_2) \cdot \mathbf{w}_2 \leq C, \\ & \mathbf{w} \in \mathcal{W}^{\times}, \end{aligned} \tag{22}$$

Since all input must be processed, some capacity levels C might be infeasible. The set of feasible capacity vectors C is given by

$$\mathcal{C}_\beta = \{C \geq 0 : \exists \vec{l} \geq 0, \text{ s.t. } \mathcal{L}(\vec{l}) \leq C\}.$$

Lemma 5 characterized the optimal effort mix and the optimal resource sharing for a given total capacity C.

LEMMA 5 **(optimal resource sharing)**. *Fix capacity $C \in \mathcal{C}_\beta$. Let $\bar{l}_k^*(\cdot)$ be the solution for type k items in Theorem 2, as a function of the relative cost. Then, the unique solution $\vec{l}^*$ to (21) is given by*

$$l_k^* = \bar{l}_k^*(\widetilde{\mathcal{R}}_k^c),\ k \in [K],$$

where

$$\widetilde{\mathcal{R}}_k^c = \mathcal{R}_k^c \left(\frac{c_2^k + \kappa}{c_2^k}\right) / \left(\frac{c_1^k + \kappa}{c_1^k}\right),$$

and κ is the minimal non-negative value for which $\mathcal{L}(\vec{l}^*)(\widetilde{\mathcal{R}}^c) \le C$. The optimal capacity allocation to each station is $C_1^* = (\lambda/\theta_1) \cdot W_1^r(l^*),\ C_2^* = (\lambda/\theta_2) \cdot W_2^r(l^*)$.

Proving Lemma 5 is argued identically to Theorem 4.

The dual variable κ for the resource constraint scale up/down the relative cost. If $c_1 < c_2$, the relative cost is shrunk for all item groups. The factor by which it is shrunk depends on the baseline costs c_1^k, c_2^k. This, in turn, diverts more effort to Station 2, the more expensive one. If $c_1^k > c_2^k$, however, the relative cost is scaled up for all item groups. This, in turn, diverts more effort to Station 1, the more expensive one. If $c_1^k = c_2^k$, the relative cost remains intact.

Important is that—because the finite capacity is mapped to a perturbed price $\widetilde{\mathcal{R}}^c$ and because l_1 is increasing and $\widetilde{\mathcal{R}}^c$ and l_2 is decreasing—the capacity allocated to the two stations is not increasing in total capacity. Instead, as the total capacity increases when $c_1 > c_2$, less of it is allocated to Station 1 and more of it is allocated to station 2. This is illustrated in Figure 16. The right plot shows the optimal capacity allocation for Stations 1 and 2 vs. the available capacity C. When there is ample capacity, most of the capacity (90%) is allocated to Station 2 – the less costly one. However, as capacity becomes scarce, more capacity is allocated to Station 1 and less to Station 2. When $c_1 < c_2$, it is the other way around.

B.6. Throughput Maximization

Throughput maximization is a special case of an optimization problem over $\mathcal{W}$. Specifically,

$$\begin{aligned} \lambda^* = \max_{w_1, w_2, \lambda}\ & \lambda \\ \text{s.t. } & \lambda w_1/\theta_1 \le C_1; \quad \lambda w_2/\theta_2 \le C_2; \\ & w_1, w_2 \in \mathcal{W}. \end{aligned}$$

It is immediate that, at optimality,

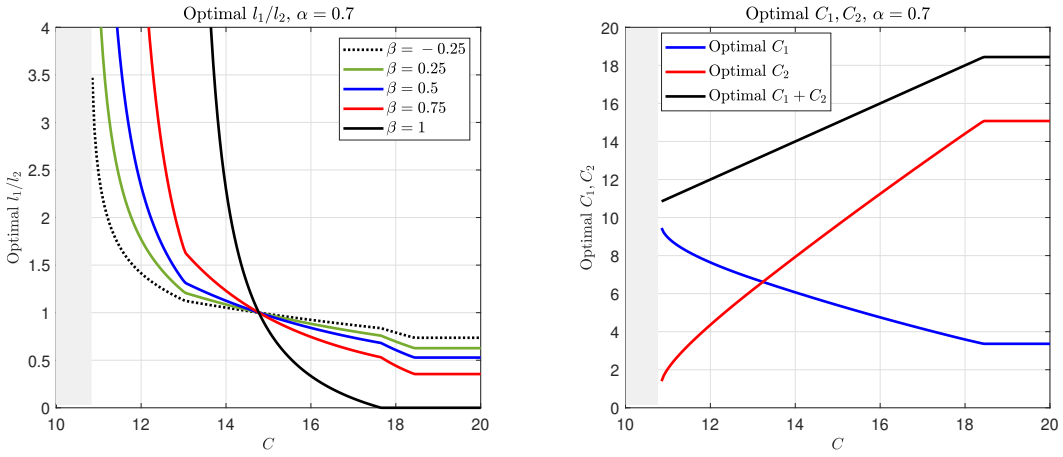
$$\lambda^* = \min\left\{\frac{C_1\theta_1}{w_1^*}, \frac{C_2\theta_2}{w_2^*}\right\} = \max_{w \in \mathcal{W}} \min\left\{\frac{C_1\theta_1}{w_1}, \frac{C_2\theta_2}{w_2}\right\}.$$

This is a problem of maximizing the concave function

$$g(w) = \min\left\{\frac{C_1\theta_1}{w_1}, \frac{C_2\theta_2}{w_2}\right\}.$$

over the convex set $\mathcal{W}$; it is, hence, a convex optimization problem. Per Lemma 8, the feasible set shrinks as β decreases. In turn, the throughput is monotonically decreasing in β. Stated verbally, this shows that the

Figure 16 **Optimal resource sharing as a function of the available capacity C for $k \in [5]$, where each k corresponds to a different value of β. The left plots present l_1^k/l_2^k, and the right plots present the optimal capacity allocation C_1^* and C_2^*. The parameters are $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, $\lambda_k = 1$, $c_1 = 3$, $c_2 = 1$.**



model and its math are consistent with prior intuition: the greater the substitution, the more throughput we should be able to handle, because substitution allows flexibility to spread out the load.

To state this formally, with multiple types we make the assumption that the mix is fixed and it is the total volume that changes. That is, $\lambda^k = \Lambda a^k$, where $a^k \geq 0$ and $\sum_{k\in[K]} a^k = 1$.

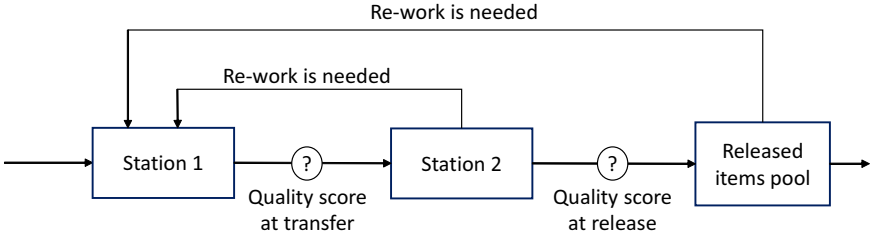
$$\begin{aligned} \lambda^* = \max_{\mathbf{w},\lambda} \ & \lambda \\ \text{s.t. } & \lambda \sum_{k\in[K]} \frac{a^k}{\theta_1^k} w_1^k \leq C_1; \quad \lambda \sum_{k\in[K]} \frac{a^k}{\theta_2^k} w_2^k \leq C_2, \\ & \mathbf{w} \in \mathcal{W}^\times. \end{aligned}$$

LEMMA 6. *Given $C_1, C_2 \geq 0$, the maximal throughput can be found by solving a convex optimization problem. The optimal solution is increasing in β: the greater the substitution, the greater the throughput.*

B.7. Incorporating Re-Work from Station 2 to Station 1.

Some fallback in Station 2 can be handled within Station 2's environment. Some cases, however, require intensive intervention in Station 1. In this section we analyze the case where quality deterioration and re-work can occur while in Station 2.

Figure 17 **Model illustration with re-work from Station 2.**



Specifically, if item's quality score hits some re-work threshold $a \leq 0$ while in Station 2, the item is returned to Station 1. In other words, items stay in Station 2 until their quality score either reaches $l_1 + l_2$ (when they are released) or zero (when they are returned); this takes, in expectation,

$$m_2(l) := \frac{l_2(1 - e^{\varrho l_1 \theta_2}) + l_1(1 - e^{-\varrho l_2 \theta_2})}{\theta_2 \left(e^{-\varrho l_2 \theta_2} - e^{\varrho l_1 \theta_2}\right)}.$$

We now distinguish between two types of re-work: The first occurs after Station 2; the expected number of these returns remains $N(l)$.

The second type of re-work occurs while processing in Station 2; the expected number of returns from Station 2 is $N_2(l) := [1 - p_2(l)]^{-1}$, where

$$p_2(l) = \frac{e^{-\varrho l_2 \theta_2} - 1}{e^{-\varrho l_2 \theta_2} - e^{\varrho l_1 \theta_2}}$$

is the re-work probability from Station 2 – the probability of hitting zero before hitting level $l_1 + l_2$.

The total costs of Stations 1 and 2 are then given by

$$\begin{aligned} \text{Station 1 cost} &= c_1 \; m_1(l) \left[N(l) + N_2(l) - 1\right]; \\ \text{Station 2 cost} &= c_2 \; m_2(l) \left[N(l) + N_2(l)\right]. \end{aligned}$$

We are, therefore, looking to solve

$$V^* = \inf_{l \geq 0} V(l, d) = \inf_{l \geq 0} \; \text{Station 1 cost} \; + \text{Station 2 cost}. \tag{23}$$

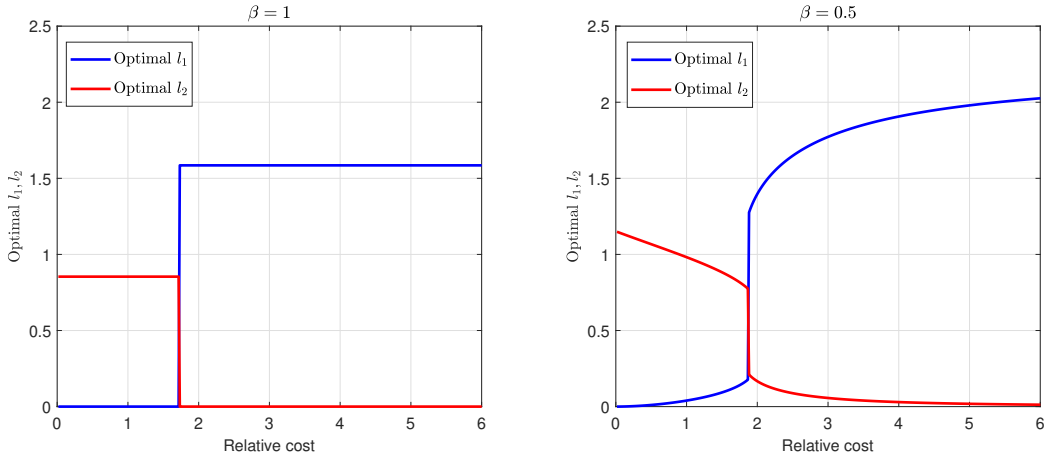
Figure 18 illustrates the solution structure when the re-work threshold from Station 2 is $a = -1$. We derived it by numerically solving (23). In this example and many others we examined, the solution structure is consistent with the one derived in §4, demonstrating the solution robustness. Moreover, we see that in contrast to the base model and its variants that we considered, having $\beta < 1$ does not guarantee that (l_1, l_2) changes smoothly with cost. Instead, there is a jump at a cost-threshold cost as in the case where $\beta = 1$ and $\alpha = 0.5$. This is certainly specific to our construction of this variant; nevertheless, it showcases the richness of the type of behavior we would observe as we introduce additional features.

Appendix C: Data Requirements and Parameter Estimation for Model Implementation

Estimating the quality-evolution parameters $\boldsymbol{\theta}, \boldsymbol{\sigma}$ from service-time samples. While our model stipulates that quality evolves as a positive-drift Brownian motion, the drift and variance parameters can be estimated without directly observing the quality path. Instead, we can rely on station-level processing-time data. Under our model, processing times follow an inverse Gaussian distribution—a distribution widely used for modeling and estimating service times (see Whitmore 1975, 1979 and the recent Hashimoto et al. 2023).

The required data include the processing times of items at each station and for each visit. Such data are typically recorded systematically in operational databases. From these data, we can estimate the parameters of the inverse Gaussian distribution, which correspond to l_i/θ_i and $(l_i/\sigma_i)^2$ for each station i. These parameters can be obtained using standard maximum-likelihood estimation methods.

Figure 18 **Solution structure when there are returns from Station 2 to Station 1. The parameters are** $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, $c_1 = 1$, $\alpha = 0.5$, $a = -1$.



In practice, the quality thresholds l_i can be obtained either directly or indirectly. They may be directly observed, for example, when quality or performance scores are recorded at transfer or release, or inferred from design-level targets used in engineering or service planning. When l_i is not directly observed, it can be estimated jointly with θ_i and σ_i by fitting the inverse-Gaussian likelihood to the observed processing-time data, treating l_i as a latent parameter inferred from the same data. Because l_i enters both the mean and the shape parameters of the inverse-Gaussian distribution, any measurement error in l_i propagates to the estimated drift and diffusion parameters (θ_i, σ_i). However, since the inverse-Gaussian likelihood depends on l_i only through the ratios l_i/θ_i and l_i/σ_i, a small multiplicative error in l_i can be absorbed by a roughly proportional rescaling of θ_i and σ_i. Consequently, such error primarily affects the overall scale of these parameters rather than their ratio θ_i/σ_i, leaving the process dynamics and identification largely unchanged. Substantial mismeasurement, however, would proportionally rescale predicted processing times and re-work probabilities, as these quantities depend directly on the magnitude of l_i.

Subsequently, we estimate the quality score at transfer or release to build estimates for θ_i and σ_i^2.

Estimating the substitution parameters α, β from realized re-work. The parameters of the CES production function, α and β, can be estimated via maximum likelihood using item-level data. Assume we have a dataset of m items, indexed by j, each possibly generating multiple visits. For item j, let N_j denote the observed number of visits until final completion (no further re-work). Under our model, all visits for the same item share the same re-work probability $p_{\alpha,\beta}(l_1, l_2)$, so that N_j follows a geometric distribution with success probability $1 - p_{\alpha,\beta}(l_1, l_2)$. Hence, the likelihood contribution of item j is

$$L_j(\alpha, \beta) = (1 - p_{\alpha,\beta}(l_1, l_2))\, p_{\alpha,\beta}(l_1, l_2)^{N_j - 1},$$

and the total likelihood is $L(\alpha, \beta) = \prod_{j=1}^{m} L_j(\alpha, \beta)$. The corresponding log-likelihood is

$$\ell(\alpha, \beta) = \sum_{j=1}^{m} \Big[\log(1 - p_{\alpha,\beta}(l_1, l_2)) + (N_j - 1) \log(p_{\alpha,\beta}(l_1, l_2)) \Big].$$

Maximizing this likelihood yields consistent estimates for α and β while properly accounting for the dependence across visits generated by the same item. Synthetic simulation experiments, described below, assess the finite-sample precision of the maximum-likelihood estimator and confirm that it accurately recovers the true parameters even in moderate sample sizes.

If, in addition to binary re-work indicators, data on the time elapsed between release and re-work were available, the estimation procedure could be extended to use this richer information. Under our model, post-release quality evolves as a BM with drift $\eta(l_1, l_2)$ and diffusion coefficient σ_{pr}, so the time until re-work corresponds to the first-hitting time of this process, which follows an inverse Gaussian distribution. While our current approach infers these post-release parameters indirectly from re-work probabilities, observing actual return times would allow them to be estimated directly from their likelihood function, thereby providing a dynamic validation of the post-discharge process.

Synthetic Precision Check. To evaluate the accuracy and precision of the item-level maximum likelihood estimator, we conducted a synthetic simulation calibrated to the maintenance setting. Each replication simulated m independent items, where item j's thresholds $(l_{1,j}, l_{2,j})$ were generated with independent variation in their total effort $l_{1,j} + l_{2,j}$ and in the relative mix between Station 1 and Station 2, so that overall scale and allocation effects were varied separately rather than by sampling $l_{1,j}$ and $l_{2,j}$ independently.

For each item j, the number of visits N_j was drawn from a geometric distribution with success probability $1 - p_j$, where

$$p_j = p_{\alpha,\beta}(l_{1,j}, l_{2,j}) = \exp\{-\eta_{\alpha,\beta}(l_{1,j}, l_{2,j})(l_{1,j} + l_{2,j})\}$$

denotes the re-work probability defined in (3). The true parameter values (α_0, β_0) were fixed at four representative combinations corresponding to strong substitution ($\beta_0 = 1$) and three levels of complementarity $(0.5, 0, -0.5)$. For each design we ran 300 replications with sample sizes $m \in \{500, 1000\}$. Table 3 reports the resulting mean number of visits and the bias and root-mean-square error (RMSE) of the estimators $\hat{\alpha}$ and $\hat{\beta}$.

The results indicate that both parameters are estimated with high precision even in moderate samples ($m = 500$), and that bias and RMSE decline further as the sample size increases. Across all designs, the bias in α is negligible ($|\text{bias}| < 0.01$), while the bias in β remains below 0.08 at $m = 500$ and below 0.05 at $m = 1000$, with RMSEs typically under 0.10.

Appendix D: Parameter Calibration for Maintenance Example

To provide an empirical grounding for the stylized *Maintenance Example*, we consider large-scale industrial and facility maintenance tasks such as those performed on power-generation turbines, compressors, and other heavy-plant equipment. These systems involve two sequential activities: an initial *repair stage* followed by *preventive service* performed while the equipment remains offline.

Station 1 (Repair) represents corrective actions that restore functionality— for instance, replacing a failed generator bearing, realigning a gearbox, or repairing electrical components. **Station 2 (Preventive service)** represents inspection and tuning activities— such as lubrication, recalibration, torque checks, or component balancing—that enhance long-term reliability and reduce the probability of future re-work.

Table 3 Simulation-based evaluation of the item-level MLE under per-item thresholds.

Design (α_0, β_0)	m	Mean visits / item	$\hat{\alpha}$		$\hat{\beta}$	
			Bias	RMSE	Bias	RMSE
(0.60, 1.00)	500	1.230	-0.0085	0.0408	-0.0713	0.1272
(0.60, 1.00)	1000	1.229	-0.0049	0.0319	-0.0502	0.0917
(0.60, 0.50)	500	1.286	0.0022	0.0396	0.0311	0.1568
(0.60, 0.50)	1000	1.288	-0.0013	0.0272	0.0089	0.1027
(0.60, 0.00)	500	1.395	0.0019	0.0297	0.0084	0.1090
(0.60, 0.00)	1000	1.394	-0.0002	0.0209	0.0100	0.0755
(0.55, -0.50)	500	1.542	-0.0006	0.0349	0.0059	0.1408
(0.55, -0.50)	1000	1.540	-0.0006	0.0237	0.0024	0.0904

Industry studies and field reports on wind-turbine operation and maintenance indicate that the scope and time requirements of these activities vary widely across sites. Facts (2024) notes that regular service campaigns often involve full-day maintenance windows that include inspection, cleaning, calibration, and documentation tasks, while Committee (2017) highlights that corrective interventions—when a specific fault has been identified—typically achieve a faster improvement in the asset's operational condition per hour of work, even though they may target a smaller subset of components.

Anchoring to these observations, we set the quality-improvement rates (Brownian drifts) to

$$\theta_1 = 0.15 \text{ h}^{-1} \text{ (repair)} \quad \text{and} \quad \theta_2 = 0.10 \text{ h}^{-1} \text{ (preventive service)},$$

reflecting that corrective actions directly address known issues and thus improve the quality state more rapidly per hour, whereas preventive activities yield slower but more comprehensive gains that sustain reliability over time.

For labor costs, we use U.S. Bureau of Labor Statistics (2024) medians—repair tasks corresponding to *industrial-machinery mechanics* (SOC 49-9041) and preventive-service tasks corresponding to *equipment or quality-control technicians* (SOC 51-9061)— augmented by a conservative 40–60% overhead. This yields $c_1 \in [\$40, \$50]$/hr and $c_2 \in [\$35, \$45]$/hr; therefore, we set $c_1 = 45$ and $c_2 = 40$.

We take $\alpha = 0.4$ and explore $\beta \in \{1, 0.5, 0.25, -4, -25\}$ to span the range from substitution to complementarity between repair and preventive service. Here, $\alpha = 0.4$ indicates that greater weight is placed on Station 2 (preventive service) in the aggregated-quality function: while repair is essential to restore basic functionality, preventive service has a larger marginal influence on long-term reliability and re-work reduction. These values reflect realistic maintenance durations for large assets, confirming that the comparative statics are plausible for real operations.

Under these inputs, the optimality conditions in Theorem 2 produce the order-of-magnitude solutions summarized in Table 4.

Table 4 Order-of-magnitude calibration for the maintenance example.

β	Interaction type	(l_1^*, l_2^*)	Optimal mean durations (hours)
1	Full substitution	(1.7723,0)	(11.8, 0)
0.5	Mild complementarity	(0.9435,0.6744)	(6.3, 6.7)
0.25	Moderate complementarity	(0.8957,0.7718)	(6, 7.7)
-4	Strong complementarity	(0.8087,0.78)	(5.4, 7.8)
-25	Full complementarity	(0.7957,0.79)	(5.3, 7.9)

D.1. Finite-Capacity Calibration

To examine how limited preventive-service capacity affects the optimal design, we extend the calibrated maintenance example to a setting with a finite number of preventive-service workers. We consider two representative capacity levels for the preventive-service stage, $C_1 \in \{6, 10\}$, and five degrees of interaction between preventive service and repair: $\beta \in \{1, 0.5, 0.25, -4, -25\}$. The corresponding optimal thresholds and mean durations are reported in Table 5.

The results illustrate how the capacity of the repair stage shapes the system's overall performance. When repair capacity is tight ($C_1 = 6$), the system relies more heavily on preventive service to sustain quality, as reflected in the longer preventive durations across all interaction levels (β values). Under strong or full complementarity ($\beta = -4$ and -25), limited repair capacity substantially reduces the effectiveness of preventive service, resulting in particularly long processing times at Station 2. When capacity increases ($C_1 = 10$), both stations operate more efficiently, and the preventive durations shorten accordingly. Overall, these patterns confirm that capacity constraints amplify performance losses in complementary systems, whereas their effect is less pronounced when the interaction between stages approaches substitution.

Table 5 Calibrated optimal thresholds and mean durations (hours) under finite repair capacity.

β	Optimal thresholds (l_1^*, l_2^*)		Mean durations (hours)	
	$C_1 = 6$	$C_1 = 10$	$C_1 = 6$	$C_1 = 10$
1.0	(0.2146, 1.27)	(0.3580, 1.15)	(1.4, 12.7)	(2.4, 11.5)
0.5	(0.2146, 1.38)	(0.3580, 1.21)	(1.4, 13.8)	(2.4, 12.1)
0.25	(0.2146, 1.46)	(0.3580, 1.24)	(1.4, 14.6)	(2.4, 12.4)
–4	(0.2146, 4.44)	(0.3580, 2.44)	(1.4, 44.4)	(2.4, 24.4)
–25	(0.2146, 5.43)	(0.3580, 3.03)	(1.4, 54.3)	(2.4, 30.3)

Appendix E: Additional Numerical Experiments

This section presents additional numerical experiments designed to illustrate the robustness and practical implications of the analytical results. All numerical results were obtained using standard MATLAB routines; the solution of a single instance requires under 0.01 seconds on a laptop CPU.

E.1. Sensitivity to Parameter Misspecification

Table 6 presents the results of a sensitivity analysis. Each parameter was independently perturbed by ±10% around the calibrated baseline reported in Appendix D, while all other parameters were held fixed. The columns labeled "Δ Mean Duration (hours,%)" report the absolute and relative changes in the average processing time at each station relative to the baseline solution (5.4 hours for Station 1 and 7.8 hours for Station 2; see Table 4). Across all scenarios, the variation in mean duration remained below 2.5%, indicating that the optimal design and operating regime are highly robust to moderate parameter misspecification.

Table 6 Sensitivity of the optimal design to parameter misspecification. Each parameter is varied by ±10% holding others fixed. Rows for c_i and θ_i are consolidated as they enter only via c_i/θ_i and R^c.

					Δ Mean Duration (hours, %)	
Parameter	**Baseline**	**Change**	l_1^*	l_2^*	**Station 1**	**Station 2**
α	0.40	±10%	0.79–0.82	0.77–0.79	0.1 (1.8%)	0.10 (1.3%)
β	−4	±10%	0.8–0.81	0.78–0.79	0.03 (0.6%)	0.05 (0.6%)
c_1/θ_1	300	±10%	0.79–0.83	0.77–0.8	0.13 (2.5%)	0.15 (1.9%)
c_2/θ_2	400	±10%	0.79–0.83	0.77–0.8	0.13 (2.5%)	0.15 (1.9%)

E.2. Comparison with Other Benchmark Policies

Although our proposed policies are analytically derived, proved to be optimal, and admit closed-form expressions, it is instructive to compare its performance with other intuitive benchmark policies that are often used in practice. We consider two such benchmarks that reflect simple, decentralized decision rules.

Cost-Priority Policy. This benchmark prioritizes effort allocation to the more cost-effective stage—that is, the station with the smaller effective cost ratio c_i/θ_i. Such a rule captures cost-driven decision making guided by immediate efficiency considerations rather than joint optimization.

Decentralized Policy. In this benchmark, each station determines its own quality threshold as if it were operating independently, ignoring the interaction captured by the complementarity or substitution parameter β. This represents an uncoordinated design, where local optimization occurs at the station level rather than through a system-wide perspective. In practice, this corresponds to departments or teams setting their own operating targets without accounting for cross-stage effects.

Table 7 compares the optimal integrated design with the two benchmark policies across different levels of substitutability and complementarity, as captured by the parameter β. When $\beta < 0$, the benchmarks that

deactivate one stage become *degenerate*, as effective quality collapses to zero when only a single station operates. In contrast, the integrated policy consistently delivers superior outcomes across all interaction regimes, achieving substantial cost savings relative to both benchmarks. These results underscore the operational value of jointly optimizing both thresholds and highlight the importance of coordinated decision-making within an integrated framework.

Table 7 Comparison with benchmark policies. "Degenerate" indicates that the benchmark collapses under $\beta < 0$, yielding $\eta = 0$.

	Objective value			Cost saving (%)	
β	Optimal	Benchmark 1	Benchmark 2	Benchmark 1	Benchmark 2
1	743.3	743.3	1,117.5	0.0%	33.5%
0.5	772.8	1,346	1,588	42.6%	51.3%
0.25	773.7	6881.5	3347.2	88.8%	76.9%
-4	775.3	Degenerate	819.8	–	5.4%
-25	775.3	Degenerate	847.6	–	8.5%

Appendix F: Proofs of Main Results

This section contains the proof of the main results (propositions and theorems). Proof of all lemmas appear in §G.

Proof of Theorem 1. We will prove this through explicit characterization of the constrains that define the set $\mathcal{W}$. We break the space into "rays":

$$\mathcal{W}_z = \left\{ w_1, w_2 \in \mathcal{W}, \quad \frac{w_2}{w_1} = z \right\}.$$

For $w_1, w_2 \in \mathcal{W}_z$,

$$\begin{aligned} w_1 &\geq \min_{l_1 \geq 0} \; l_1 N(l_1, z l_1); \\ w_2 &= (w_1)\, z \geq z \min_{l_1 \geq 0} \; l_1 N(l_1, z l_1). \end{aligned} \tag{24}$$

Let us expand the right-hand side of (28)

$$\begin{aligned} l_1 N(l_1, z l_1) &= l_1 \left[1 - e^{-\kappa - \varrho \eta(l_1, z l_1) l (z+1)}\right]^{-1} \\ &= l_1 \left[1 - e^{-\kappa - \varrho l_1^2 \left(\alpha + (1-\alpha) z^\beta\right)^{\frac{1}{\beta}} (z+1)}\right]^{-1} \\ &= \frac{1}{\sqrt{\mathbb{h}_{\alpha,\beta}(z)}} \sqrt{\mathbb{h}_{\alpha,\beta}(z)} l_1 \left[1 - e^{-\kappa - l_1^2 \mathbb{h}_{\alpha,\beta}(z)}\right]^{-1}, \end{aligned}$$

where

$$\mathbb{h}_{\alpha,\beta}(z) := \varrho \left(\alpha + (1-\alpha) z^\beta\right)^{\frac{1}{\beta}} (z+1).$$

Changing variables by defining $y = \sqrt{\mathbb{h}_{\alpha,\beta}(z)}l_1$, yields the following

$$w_1 = \min_{l_1 \geq 0} \; l_1 N(l_1, z l_1) = \frac{1}{\sqrt{\mathbb{h}_{\alpha,\beta}(z)}} \min_{y \geq 0} y \left[1 - e^{-\nu - y^2}\right]^{-1} = \frac{\Gamma_\nu}{\sqrt{\mathbb{h}_{\alpha,\beta}(z)}},$$

where

$$\Gamma_\nu = \min_{y \geq 0} y \left[1 - e^{-\kappa - y^2}\right]^{-1}.$$

Notice that Γ_ν does not depend on α, β, or z.

Transformation then gives us

$$w_1^2 = \frac{\Gamma_\nu^2}{\mathbb{h}_{\alpha,\beta}(z)} \quad \Rightarrow \quad \mathbb{h}_{\alpha,\beta}(z) = \frac{\Gamma_\nu^2}{w_1^2} \quad \Rightarrow \quad z = \mathbb{h}_{\alpha,\beta}^{-1}\left(\frac{\Gamma_\nu^2}{w_1^2}\right),$$

and thus,

$$w_2 = z\left(w_1\right) = \left(w_1\right) \mathbb{h}_{\alpha,\beta}^{-1}\left(\frac{\Gamma_\nu^2}{w_1^2}\right).$$

Since $w_2 \geq 0$, the set $\mathcal{W}$ can now be rewritten as

$$\mathcal{W} = \left\{ w_1, w_2 \geq 0 : \; w_2 \geq w_1 \mathbb{h}_{\alpha,\beta}^{-1}\left(\frac{\Gamma_\nu^2}{w_1^2}\right) \right\}.$$

Since Γ_ν is independent in z, $\mathcal{W}$ is a convex set provided that the function

$$f_{\alpha,\beta}(w_1) = \left(w_1\right) \mathbb{h}_{\alpha,\beta}^{-1}\left(\frac{\Gamma_\nu^2}{w_1^2}\right)$$

is convex.

We will prove this convexity by showing that $f''_{\alpha,\beta}(w_1) \geq 0$. The following representation will be useful

$$g(z) = \frac{\Gamma_\nu}{\sqrt{\mathbb{h}_{\alpha,\beta}(z)}}, \qquad m(z) = \frac{z}{\sqrt{\mathbb{h}_{\alpha,\beta}(z)}},$$

so that

$$f_{\alpha,\beta}(w_1) = m\left(g^{-1}(w_1)\right).$$

First, we introduce the following auxiliary lemma.

LEMMA 7. *The function $g(z) = \Gamma_\nu / \sqrt{\mathbb{h}_{\alpha,\beta}(z)}$ is monotone decreasing in z for $\alpha \geq 0.5$.*

The first and second derivatives of $f_{\alpha,\beta}$ are

$$f'_{\alpha,\beta}(w_1) = m'\left(g^{-1}(w_1)\right) \frac{1}{g'\left(g^{-1}(w_1)\right)},$$

and

$$\begin{aligned} f''_{\alpha,\beta}(w_1) &= m''\left(g^{-1}(w_1)\right) \left(\frac{1}{g'\left(g^{-1}(w_1)\right)}\right)^2 + m'\left(g^{-1}(w_1)\right) \left(-\frac{g''\left(g^{-1}(w_1)\right)}{g'\left(g^{-1}(w_1)\right)^3}\right) \\ &= \left(g'\left(g^{-1}(w_1)\right)\right)^{-3} \left[m''\left(g^{-1}(w_1)\right) g'\left(g^{-1}(w_1)\right) - m'\left(g^{-1}(w_1)\right) g''\left(g^{-1}(w_1)\right)\right]. \end{aligned}$$

Per Lemma 7, $g'(z) < 0$; therefore, it suffices to prove that the second-line term in brackets is negative; namely,

$$m''(z)g'(z) - m'(z)g''(z) \leq 0, \quad \forall z \geq 0.$$

This is what we do next.

Direct differentiating yields that

$$m''(z)g'(z) - m'(z)g''(z) = \frac{\mathcal{L}(z)}{4z^2(1+z)^3\left(\alpha+(1-\alpha)z^\beta\right)^{2+\frac{1}{\beta}}}, \tag{25}$$

where

$$\mathcal{L}(z) = -z^{2\beta}(1-\alpha)^2 - z^2\alpha^2 + 2z^\beta(1-\alpha)\alpha\left(-1-z(1+z)+\beta+z(2+z)\beta\right).$$

Since the denominator in (25) is positive, it suffices to show that $\mathcal{L}(z)$ is negative for all $z \geq 0$. Since $\mathcal{L}(0) = 0$, we need to show that $\mathcal{L}'(z) \leq 0$ to conclude that $\mathcal{L}(z) \leq 0$ for all $z \geq 0$.

Direct differentiation gives

$$\mathcal{L}'(z) = -2\beta z^{-1+2\beta}(1-\alpha)^2 + 2z^{-1+\beta}\left(\beta^2(1+z)^2 - z(1+2z) + \beta(-1+z+z^2)\right)\alpha(1-\alpha) - 2z\alpha^2.$$

In turn,

$$\mathcal{L}'(z) \leq a(z) := -2\beta z^{-1+2\beta}(1-\alpha)^2 + 2z^{-1+\beta}\left(\beta^2(1+z)^2 - z(1+2z) + \beta(-1+z+z^2)\right)\alpha(1-\alpha) - 2z\alpha^2.$$

Since $a(z) \leq 0$, if and only if, $z^{1-\beta}a(z) \leq 0$, for all $z \geq 0$, we proceed with

$$z^{1-\beta}a(z) = -2\beta z^\beta(1-\alpha)^2 + 2\left[(\beta^2+\beta-2)z^2 + (2\beta^2+\beta-1)z + \beta^2 - \beta\right]\alpha(1-\alpha) - 2z^{2-\beta}\alpha^2.$$

Since $\beta \leq 1$, we get that $\beta^2+\beta-2 \leq 0$, and $\beta^2-\beta \leq 0$, so that

$$z^{1-\beta}a(z) = -2\beta z^\beta(1-\alpha)^2 + 2\left[(\beta^2+\beta-2)z^2 + (2\beta^2+\beta-1)z + \beta^2 - \beta\right]\alpha(1-\alpha) - 2z^{2-\beta}\alpha^2.$$

Now, if $2\beta^2+\beta-1 \leq 0$ (equivalently, $\beta \leq 0.5$), then $z^{1-\beta}a(z) \leq 0$, and we are done. It remains to consider the case where $\beta > 0.5$. In this case,

$$z^{1-\beta}a(z) \leq -2\left(\beta z^\beta(1-\alpha)^2 - (2\beta^2+\beta-1)z\alpha(1-\alpha) + z^{2-\beta}\alpha^2\right).$$

It can easily be verified by differentiation that the right-hand side is decreasing in $\alpha \in [0.5, 1]$, when $2\beta^2 + \beta - 1 \geq 0$. We can then take $\alpha = 0.5$ (recall that $\alpha \geq 0.5$ by assumption), and get that

$$z^{1-\beta}a(z) \leq -0.5\left(\beta z^\beta - (2\beta^2+\beta-1)z + z^{2-\beta}\right).$$

Dividing further by z, we seek to prove that

$$-0.5\left(\beta z^{\beta-1} - (2\beta^2+\beta-1) + z^{1-\beta}\right) \leq 0,$$

or equivalently,

$$\beta z^{\beta-1} + z^{1-\beta} - 2\beta^2 - \beta + 1 \geq 0. \tag{26}$$

Define

$$h(z) = \beta z^{\beta-1} + z^{1-\beta}.$$

Then, $h(z)$ has a unique minimum at

$$z^* = \beta^{-\frac{1}{2(\beta-1)}}, \quad \text{and} \quad h(z^*) = 2\sqrt{\beta}.$$

Thus, to prove (26), it suffices to show that

$$2\sqrt{\beta} - 2\beta^2 - \beta + 1 \geq 0.$$

Since this is a decreasing function in β, which equals 0 at the upper boundary, $\beta = 1$, we are done.

We conclude that $\mathcal{L}(z) \leq 0$, $\forall z \geq 0$, and that the function $f_{\alpha,\beta}(w_1)$ is convex. Note that in case where $\beta < 1$ (strictly), $\beta^2 + \beta - 2 < 0$, and $\beta^2 - \beta < 0$, so that $z^{1-\beta}a(z) < 0$ (strictly), and therefore, $f_{\alpha,\beta}(w_1)$ is strictly convex.

Lastly, we prove (10) for n stations; $n = 2$ is then a special case.

Specifically, let $(l_1, \ldots, l_n)$ denote the transfer target from each of the n stations. The service cost, drift and diffusion coefficient of each Station i, $i = 1, \ldots, n$ are c_i, θ_i and σ_i.

The constant elasticity of substitution (CES) production function for n stations is:

$$\eta(l) = \left[\sum_{i=1}^{n} \alpha_i l_i^{\beta}\right]^{\frac{1}{\beta}}, \quad 0 < \alpha_i < 1, \quad \beta \leq 1, \tag{27}$$

where $\sum_{i=1}^{n} \alpha^i = 1$; we use henceforth α for the probability vector $\alpha_1, \ldots, \alpha_n$.

The workload in Station i is

$$W_i(l) = \lambda \frac{l_i}{\theta_i} N(l), \text{ and } W_i^r(l) = l_i N(l).$$

As before, we first break the space into "rays":

$$\mathcal{W}_z = \left\{ w \in \mathbb{R}_+^n \in \mathcal{W}, \quad \frac{w_n}{w_1} = z_n \right\}.$$

where $z = (1, z_2, \ldots, z_n)$.

For $w \in \mathcal{W}_z$,

$$\begin{aligned} w_1 &\geq \min_{l_1 \geq 0} \; l_1 N(l_1 z); \\ w_n &= w_1 z_n \geq z_n \min_{l_1 \geq 0} \; l_1 N(l_1 z), \end{aligned} \tag{28}$$

where $l_1 z$ is the vector $l_1 z = (l_1, l_2 z_2, \ldots, l_n z_n)$.

As in the proof of Theorem 1 we then have

$$w_1 = \min_{l_1 \geq 0} \; l_1 N(l_1, z l_1) = \frac{1}{\sqrt{\mathbb{h}_{\alpha,\beta}(z)}} \min_{y \geq 0} y \left[1 - e^{-\nu - y^2}\right]^{-1} = \frac{\Gamma_\nu}{\sqrt{\mathbb{h}_{\alpha,\beta}(z)}},$$

where

$$\Gamma_\nu = \min_{y \geq 0} y \left[1 - e^{-\kappa - y^2}\right]^{-1}.$$

and

$$\mathbb{h}_{\alpha,\beta}(z) := \varrho \left(\alpha_1 + \sum_{i=2}^{n} \alpha_i z_i^{\beta}\right)^{\frac{1}{\beta}} \left(1 + \sum_{i=2}^{n} z_i\right).$$

The attainable workload set in terms of w_1 and the multipliers z is then given by those $w_1, z_2, \ldots, z_n \geq 0$ such that

$$\frac{1}{w_1} - \sqrt{\mathbb{h}_{\alpha,\beta}(z_2, \ldots, z_n)} \leq 0,$$

Substituting $w_n/w_1 = z_n$, we get

$$\mathcal{W} := \left\{ w \in \mathbb{R}^n_+ : \frac{1}{w_1} - \sqrt{\mathbb{h}_{\alpha,\beta}(w_2/w_1, \ldots, w_n/w_1)} \leq 0 \right\}.$$

Basic manipulation (extracting w_1) gives

$$\sqrt{\mathbb{h}_{\alpha,\beta}(w_2/w_1, \ldots, w_n/w_1)} = \frac{1}{w_1}\sqrt{\eta(w)(e \cdot w)},$$

which then gives

$$\mathcal{W} = \left\{ w \in \mathbb{R}^n_+ : \frac{1}{w_1} - \frac{1}{w_1}\sqrt{\eta(w)(e \cdot w)} \leq 0 \right\} = \left\{ w \in \mathbb{R}^n_+ : \eta(w)(e \cdot w) \geq 1 \right\},$$

as stated.

Finally, we prove that $\mathcal{W}$ is a convex subset of $\mathbb{R}^n_+$.

The production function $\eta(w)$ is concave in $w \in \mathbb{R}^n_+$ for any probability vector α and $\beta \leq 1$. It is, in particular, log-concave. The function $(e \cdot w)$ is trivially concave and hence log-concave. The function $\eta(w)(e \cdot w)$ is the product of two log-concave functions is itself log-concave; it is, in particular, quasi-convex. The contour set $\mathcal{W} = \{w \in \mathbb{R}^n_+ : \eta(w)(e \cdot w) \geq 1\}$ is then a convex set. Q.E.D.

Proof of Proposition 1. Direct differentiation gives

$$f'_{\alpha,\beta}(w_1) = \mathbb{h}^{-1}_{\alpha,\beta}\left(\frac{\Gamma^2}{w_1^2}\right) - \frac{2\Gamma^2}{w_1^2 \mathbb{h}'_{\alpha,\beta}\left(\mathbb{h}^{-1}_{\alpha,\beta}\left(\frac{\Gamma^2}{w_1^2}\right)\right)}.$$

As $w_1 \uparrow w_1^0$, $z(w_1) = \mathbb{h}^{-1}_{\alpha,\beta}(\Gamma^2/w_1^2) \downarrow 0$, by definition of w_1^0.

Since $w_1^0 = \Gamma/\sqrt{\mathbb{h}_{\alpha,\beta}(0)} > 0$, we have

$$\lim_{w_1 \uparrow w_1^0} f'_{\alpha,\beta}(w_1) = -\frac{2\Gamma^2}{(w_1^0)^2} \lim_{z \downarrow 0} \frac{1}{\mathbb{h}'_{\alpha,\beta}(z)}.$$

Therefore, we proceed by focusing on $\lim_{z\downarrow 0} \mathbb{h}'_{\alpha,\beta}(z)$.

Recall that,

$$\begin{aligned}
&\mathbb{h}_{\alpha,\beta}(z) := \varrho\left(\alpha + (1-\alpha)z^\beta\right)^{\frac{1}{\beta}}(z+1);\\
&\mathbb{h}'_{\alpha,\beta}(z) = \left(\alpha + (1-\alpha)z^\beta\right)^{\frac{1}{\beta}} + \left(\alpha + (1-\alpha)z^\beta\right)^{\frac{1}{\beta}-1} z^{\beta-1})(z+1)(1-\alpha).
\end{aligned}$$

Therefore,

$$\lim_{z\downarrow 0} \mathbb{h}'_{\alpha,\beta}(z) = \alpha^{\frac{1}{\beta}} + \alpha^{\frac{1}{\beta}-1}(1-\alpha)z^{\beta-1} = \begin{cases} \infty, & \text{if } \beta < 1; \\ 1, & \text{if } \beta = 1, \end{cases}$$

where the first equality if because $\frac{1}{\beta} - 1 \geq 0$, and the second equality is because $z^{\beta-1} \to \infty$ when $\beta < 1$, and $z^{\beta-1} = 1$ when $\beta = 1$.

We conclude that

$$\lim_{w_1 \uparrow w_1^0} f'_{\alpha,\beta}(w_1) = \begin{cases} 0, & \text{if } \beta < 1; \\ -\frac{2\Gamma^2}{(w_1^0)^2} := -\gamma_1, & \text{if } \beta = 1. \end{cases}$$

Next, we similarly consider the limit as $y \downarrow 0$. In this case, $z(y) \uparrow \infty$. It can be verified that

$$\frac{\mathbb{h}'_{\alpha,\beta}(z)}{z} \to 2(1-\alpha)^{\frac{1}{\beta}}, \quad \text{as } z \uparrow \infty.$$

Since $\mathbb{h}_{\alpha,\beta}(z(w_1)) = \Gamma^2/w_1^2$, we get that

$$f'_{\alpha,\beta}(w_1) = \mathbb{h}^{-1}_{\alpha,\beta}\left(\frac{\Gamma^2}{(w_1)^2}\right) - \frac{2}{(w_1)^2 \mathbb{h}'_{\alpha,\beta}\left(\mathbb{h}^{-1}_{\alpha,\beta}\left(\frac{\Gamma^2}{w_1^2}\right)\right)} = z(w_1) - \frac{2\mathbb{h}_{\alpha,\beta}(z(w_1))}{\mathbb{h}'_{\alpha,\beta}(z(w_1))}.$$

Let us focus on the second term

$$\lim_{w_1 \downarrow 0} \frac{2\mathbb{h}_{\alpha,\beta}(z(w_1))}{\mathbb{h}'_{\alpha,\beta}(z(w_1))} = \lim_{z \uparrow \infty} \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)} = \lim_{z \uparrow \infty} \frac{2\mathbb{h}_{\alpha,\beta}(z)/z}{\mathbb{h}'_{\alpha,\beta}(z)/z} = 1 + z + \frac{\alpha(1+z)}{2z^{\beta}},$$

where the last equality is because

$$\begin{aligned}\frac{\mathbb{h}_{\alpha,\beta}(z)}{z} &= (1-\alpha)^{\frac{1}{\beta}}(1+z)\left(\frac{\alpha}{(1-\alpha)z^{\beta}} + 1\right) \\ &= (1-\alpha)^{\frac{1}{\beta}}(1+z) + (1-\alpha)^{\frac{1}{\beta}}\alpha\left(\frac{1+z}{z^{\beta}}\right).\end{aligned}$$

In turn,

$$\lim_{w_1 \downarrow 0} \frac{f'_{\alpha,\beta}(w_1)}{z(w_1)} = -\lim_{z \uparrow \infty} 0.5\alpha\left(\frac{1+z}{z^{\beta}}\right) = \begin{cases} 0.5\alpha, & \text{if } \beta = 1; \\ -\infty, & \text{if } \beta < 1, \end{cases}$$

and, therefore,

$$\lim_{y \downarrow 0} f'_{\alpha,\beta}(w_1) = -\infty, \text{ when } \beta < 1.$$

For the special case $\beta = 1$, it is not enough that

$$\lim_{y \downarrow 0} \frac{f'_{\alpha,\beta}(w_1)}{z(w_1)} = 0.5\alpha,$$

since we want to show that $f'_{\alpha,\beta}(w_1) \to$ constant. For $\beta = 1$,

$$\begin{aligned}\frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)} &= \frac{2(1+z)(\alpha + (1-\alpha)z)}{1 + 2(1-\alpha)z} = \frac{(1+z)\left(\frac{\alpha}{(1-\alpha)z} + 1\right)}{\frac{1}{2(1-\alpha)z} + 1} \\ &= (1+z) + \frac{\alpha}{1-\alpha} + o(1), \text{ as } z \uparrow \infty,\end{aligned}$$

so that,

$$\begin{aligned}f'_{\alpha,\beta}(w_1) &= z(w_1) - \frac{2\mathbb{h}_{\alpha,\beta}(z(w_1))}{\mathbb{h}'_{\alpha,\beta}(z(w_1))} \\ &= z(w_1) - (1 + z(w_1)) - \frac{\alpha}{1-\alpha} - o(1) = -\frac{1}{1-\alpha} - o(1),\end{aligned}$$

or, stated equivalently,

$$f'_{\alpha,\beta}(w_1) \to -\frac{1}{1-\alpha} \quad \text{as } w_1 \downarrow 0, \quad \text{when } \beta = 1.$$

Q.E.D.

Proof of Theorem 2. Recall our optimization problem

$$\min_{w_1,w_2\geq 0} w_1+\mathcal{R}^c w_2$$
$$\text{s.t. } w\in\mathcal{W}.$$

Then (see for example, Bertsekas 1997, Proposition 2.1.2), w^* is an optimal point if and only if

$$(1,\mathcal{R}^c)'(w-w^*)\geq 0,\quad \forall w\in\mathcal{W}.$$

We will prove two claims:

1. If $\beta=1$, there exists $\mathcal{R}^c$ small enough, for which $w^*=(0,w_2^0)$. If $\beta=1$, there also exists $\mathcal{R}^c$ large enough, for which $w^*=(w_1^0,0)$.
2. If $\beta<1$, the optimal solution always has $w_1,w_2>0$.

We start with $\beta=1$. Take as a candidate for optimality $w^*=(0,w_2^0)$; then, for each $w\in\mathcal{W}$,

$$w_1+\mathcal{R}^c\left(w_2-w_2^0\right)\geq 0. \tag{29}$$

If $w_2\geq w_2^0$, there is nothing to verify, so let us assume $w_2<w_2^0$. Recall that

$$w_2>f_{\alpha,\beta}(w_1)=w_1^2\mathbb{h}_{\alpha,\beta}^{-1}\left(\frac{\Gamma^2}{w_1^2}\right).$$

Hence, (29) holds, in particular, if

$$w_1+\mathcal{R}^c\left(f_{\alpha,\beta}(w_1)-w_2^0\right)=w_1-\mathcal{R}^c\left(w_2^0-f_{\alpha,\beta}(w_1)\right)\geq 0.$$

Recall that for $\beta=1$, the derivative is in $[-\gamma_1,-\gamma_2]$, where $\gamma_1,\gamma_2\in(0,\infty)$, hence,

$$w_1-\mathcal{R}^c\left(w_2^0-f_{\alpha,\beta}(w_1)\right)\geq w_1-\mathcal{R}^c\gamma_2 w_1.$$

This expression is, in turn, positive for all $w_1\leq w_1^0$ provided that $\mathcal{R}^c\gamma_2\leq 1$.

For $w_1>w_1^0$,

$$w_1+\mathcal{R}^c\left(w_2-w_2^0\right)\geq w_1^0-\mathcal{R}^c w_2^0,$$

which holds as long as $\mathcal{R}^c\leq w_1^0/w_2^0$.

Combining the above, we have that $w^*=(0,w_2^0)$ is optimal for small enough $\mathcal{R}^c$. The argument for large $\mathcal{R}^c$ is similar and omitted.

For $\beta<1$, let us again consider $w^*=(0,w_2^0)$; we show that it cannot be optimal. Proceeding in the same way as before, notice that given w_1, $w_1+\mathcal{R}^c(w_2-w_2^*)$ is minimized at $f_{\alpha,\beta}(w_1)$. We will show that for any $\mathcal{R}^c>0$, there exists w_1 small enough, such that

$$w_1+\mathcal{R}^c\left(f_{\alpha,\beta}(w_1)-w_2^0\right)<0.$$

Recall that $w_2^0=f_{\alpha,\beta}(0)$ and that $f'_{\alpha,\beta}(w_1)\to-\infty$, as $w_1\downarrow 0$ because $\beta<1$. Thus, given $\mathcal{R}^c$ small enough, such that $f'_{\alpha,\beta}(w_1)\leq -2/\mathcal{R}^c$, then

$$w_1+\mathcal{R}^c\left(f_{\alpha,\beta}(w_1)-w_2^0\right)\leq w_1-\mathcal{R}^c\frac{2}{\mathcal{R}^c}w_1=-w_1<0.$$

The argument is similar for $y=(w_1^0,0)$, and is omitted. We, therefore, showed that there will never be a corner solution when $\beta<1$.

Note that with the exception of the case where $\beta=1$ and $\alpha=0.5$, the optimal solution is unique. This follows from the strict convexity at the end of Theorem 1's proof. Q.E.D.

Proof of Proposition 2. Let us start from item (iii). Per Lemma 8, the function f is increasing in β for each w_1. In particular, the set $\mathcal{W}_{\beta_1} \subseteq \mathcal{W}_{\beta_2}$ if $\beta_1 < \beta_2$. In turn,

$$\bar{V}(\beta_2, \mathcal{R}^c) := \min_{w \in \mathcal{W}_{\beta_2}} (w_1 + \mathcal{R}^c w_2) \leq \min_{w \in \mathcal{W}_{\beta_1}} \lambda (w_1 + \mathcal{R}^c w_2) =: \bar{V}_{\beta_1}.$$

We turn to (i). Per Lemma 2, the optimal ratio z^* is achieve at the point where

$$-\left(z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)}\right) = \frac{1}{\mathcal{R}^c}.$$

At $z = 1$ (where $w_1 = w_2$), the left hand-side evaluates to

$$-\left(z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)}\right)\Big|_{z=1} = \frac{2\alpha + 1}{2(1-\alpha)+1},$$

which does not depend on β.

At $z = 1$, $w_1 = \Gamma/\sqrt{\mathbb{h}_{\alpha,\beta}(1)} = \Gamma/\sqrt{2\varrho}$ and $l_2^* = l_1^*$. Thus, $\eta(l^*) = l_1^* = l_2^*$, so that l^* solves

$$W_1^r(l) = l_1 N(l_1, l_1) = l_1 \frac{1}{1 - e^{-2\varrho l_1^2}} = \frac{\Gamma}{\sqrt{2\varrho}},$$

which is an equation that does not depend on β.

Finally, we turn to (ii). Again, we use the fact that the optimal ratio z^* uniquely solves

$$g_\beta(z) := -\left(z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)}\right) = \frac{1}{\mathcal{R}^c}.$$

First, we note that because f is convex, the right-hand side (which is $-f'$) is decreasing in z. Thus, the inverse g^{-1} is also monotone decreasing in its argument, hence monotone increasing in $\mathcal{R}^c$: the larger $\mathcal{R}^c$ is, the larger z^* is. In particular, $z^* \geq 1$ for $\mathcal{R}^c \geq \mathcal{R}_0^c$ and $z^* \leq 1$ for $\mathcal{R}^c \leq \mathcal{R}_0^c$.

All we need to show now is that z^* is increasing in β when $\mathcal{R}^c \leq \mathcal{R}_0^c$ (which by the above means $z^* > 1$) and decreasing for $\mathcal{R}^c \geq \mathcal{R}_0^c$.

Differentiation gives

$$\frac{\partial}{\partial \beta} g_\beta(z) = \frac{-2(1-\alpha)\alpha(z+1)^2 z^{\beta+1}\log(z)}{\left((1-\alpha)(2z+1)z^\beta - \alpha z\right)^2} = \begin{cases} > 0, & \text{if } z < 1 \\ < 0, & \text{if } z > 1, \\ = 0, & \text{if } z = 1. \end{cases}$$

We conclude, as stated, that $g_\beta(z)$ is increasing in β for $z < 1$ (for $\mathcal{R}^c \leq \mathcal{R}_0^c$) and decreasing otherwise. Q.E.D.

Proof of Theorem 3. The existence of an optimal solution w^* is guaranteed by the linearity of the objective function and the convexity of the set $\mathcal{W} \cup \{(w_1, w_2) : (\lambda/\theta_1)w_1 + (\lambda/\theta_2)w_2\}$. With the exception of the case $\beta = 1$, $\alpha = 1/2$, $\mathcal{R}^c = 1$, this solution is unique.

Notice that for $w_1 \leq w_1^0$, the optimal solution must be of the form $(w_1, f_{\alpha,\beta}(w_1))$. The total cost is $w_1 + \mathcal{R}^c f_{\alpha,\beta}(w_1)$ and has the derivative $1 + \mathcal{R}^c f'_{\alpha,\beta}(w_1)$. If $\mathcal{R}^c f'_{\alpha,\beta}(w_1) \leq -1$–equivalently $\mathcal{R}^c |f'_{\alpha,\beta}(w_1)| - 1 \geq 0$— the cost is decreasing in w_1, so that the optimal solution is $(C_1^r, f_{\alpha,\beta}(C_1^r))$. Notice that this is the optimal unconstrained solution at the cost $\widetilde{\mathcal{R}}^c$, where $|f'_{\alpha,\beta}(C_1)| = \frac{1}{\widetilde{\mathcal{R}}^c}$. The solution κ_1 to $\mathcal{R}^c \frac{c_1}{c_1 + \kappa_1} = \frac{1}{|f'_{\alpha,\beta}(C_1^r)|}$ is as stated. It is now by construction that with this choice of κ_1 and κ_2, the optimal solution is $(C_1^r, f_{\alpha,\beta}(C_1^r))$. A similar argument is repeated for κ_2.

It is immediate now that κ_1, κ_2 are infinite if $f'_{\alpha,\beta}(C_1^r) = f'_{\alpha,\beta}(f^{-1}(C_2^r)) = 1/\mathcal{R}^c$. Notice that if $\kappa_1 > 0$, then $|f'_{\alpha,\beta}(C_1^r)| > 1/\mathcal{R}^c$; hence, it must be also that $|f'_{\alpha,\beta}(f^{-1}_{\alpha,\beta}(C_2^r))| > 1/os$ and, in turn, that $\kappa_2 = 0$. Indeed, if $|f'_{\alpha,\beta}(C_1^r)| > 1/\mathcal{R}^c$, then $f'_{\alpha,\beta}(C_1^r) < -1/\mathcal{R}^c$ and $(f'_{\alpha,\beta})^{-1}(1/\mathcal{R}^c) > C_1^r$. Moreover, $|f'_{\alpha,\beta}(f^{-1}_{\alpha,\beta}(C_2^r))| \leq 1/\mathcal{R}^c$ implies similarly that $(f'_{\alpha,\beta})^{-1}(1/\mathcal{R}^c) \leq f^{-1}_{\alpha,\beta}(C_2^r)$. Combining these two yields that $C_1^r < f^{-1}_{\alpha,\beta}(C_2^r)$, or equivalently—because $f_{\alpha,\beta}$ and $f^{-1}_{\alpha,\beta}$ are decreasing functions—that $f_{\alpha,\beta}(C_1^r) > C_2^r$, which means that $(C_1^r, C_2^r) \notin \mathcal{W}$ and contradicts the feasibility of this capacity levels. Q.E.D.

Proof of Theorem 4. The existence of an optimal solution follows from the linearity of the objective function and the convexity of the set $\mathcal{W}^{\times}$. The existence of Lagrange multipliers as stated then follows from standard necessary conditions for convex optimization with inequality constraints; see, e.g., (Bertsekas 1997, Proposition 3.1).

By definition, the optimal constrained value is the same as the optimal unconstrained value with the corrected relative cost $\widetilde{\mathcal{R}}_k^c$. If $C_1 = \infty$, as C_2 increases, the optimal solution places more (or the same, if C_2 is not binding) work in Station 2. Because the corrected cost $\widetilde{\mathcal{R}}_k^c$ either goes up for all types or down for all types, if it goes up for all types then the consumption in Station 2 goes down, contradicting the increased usage. This is argued identically for C_1. Q.E.D.

Appendix G: Proof of Lemmas

Proof of Lemma 1. We prove the lemma for the case that $\nu = 0$. The proof readily extends.

$$W(l) = \frac{\lambda}{\theta} l \left[1 - e^{-\varrho\gamma l^2}\right]^{-1} - \lambda \frac{a}{\theta}.$$

Direct differentiation gives

$$W''(l) = \frac{\lambda}{\theta} \frac{l\gamma\varrho\left(-3 + 2l^2\gamma\varrho Coth\left(l^2\gamma\varrho/2\right)\right)}{-1 + Cosh\left(l^2\varrho\gamma\right)} > 0.$$

The inequality follows because $Cosh(u)$ (the hyperbolic cosine) satisfies $Cosh(u) \geq 1$, for all $u \geq 0$ and because $Coth(u)$ (the hyperbolic cotangent of u) satisfies that $2uCoth(u/2) \geq 4$, for all $u > 0$. We, therefore, have that $W(l)$ is convex in l. The optimal solution then satisfies the first order condition, $W'(l) = 0$, which corresponds to $e^{l^2\gamma\varrho} = 1 + 2l^2\gamma\rho$, as stated.

We turn to the explicit expression of the optimal workload. Changing variable $y = \sqrt{\varrho\gamma}l$, we write

$$W^* = \min_l W(l) = \frac{\lambda}{\theta}\left(\min_{y\geq 0}\left\{\frac{y}{\varrho\gamma}\left[1 - e^{-y^2}\right]^{-1}\right\} - a\right).$$

Defining $\Gamma_\nu = \min_{y\geq 0} y\left[1 - e^{-y^2}\right]^{-1}$ we have the stated result. Q.E.D.

Proof of Lemma 2. If there is an interior optimal solution, it is on the boundary where $w_2 = f_{\alpha,\beta}(w_1)$. The objective function takes the form $w_1 + \mathcal{R}^c f_{\alpha,\beta}(w_1)$ and, by the first order conditions, the optimum satisfies

$$f'_{\alpha,\beta}(w_1^*) = -\frac{1}{\mathcal{R}^c}, \tag{30}$$

which has a unique solution. This is because f is strictly convex for $w_1 \le w_1^0$ if $\beta < 1$ or if $\beta = 1$ and $\alpha > 1/2$; see the proof of Theorem 1. The strict convexity then implies that $f'' < 0$ and, in turn, that f' is a strictly monotone decreasing function.

Recall that

$$f'_{\alpha,\beta}(w_1) = z(w_1) - \frac{2\mathbb{h}_{\alpha,\beta}(z(w_1))}{\mathbb{h}'_{\alpha,\beta}(z(w_1))}.$$

Hence, $z(w_1)$ is the unique solution z

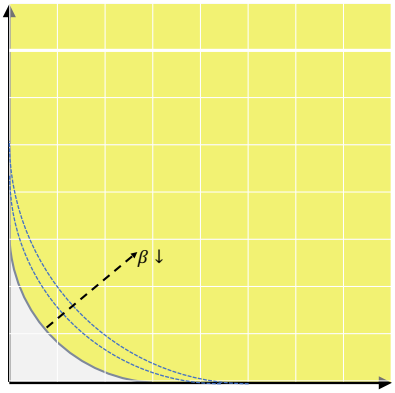
$$z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)} = -\frac{1}{\mathcal{R}^c}.$$

Q.E.D.

Lemma 8, depicted in Figure 19, proves invaluable in our optimization analysis. It established that stronger complementarity (smaller β) shrinks the feasibility region.

LEMMA 8 **(complementarity shrinks the feasibility region)**. *The function $f_{\alpha,\beta}(\cdot)$ is decreasing in β; that is, for each $w_1 \ge 0$, $f_{\alpha,\beta_1}(w_1) \ge f_{\alpha,\beta_2}(w_1)$ if $\beta_2 \ge \beta_1$. In turn, given a convex function $g(\cdot) : \mathbb{R}^2_+ \to \mathbb{R}$, the optimal value $g^* = \min_{w \in \mathcal{W}} g(w)$ is larger for smaller β (greater complementarity).*

Figure 19 The attainable workload region shrinks as complementarity increases (β decreases).



Proof of Lemma 8. Recall that $\mathbb{h}_{\alpha,\beta}(z) = \eta(1,z)(1+z)$, where $\eta(l_1,l_2) = (\alpha l_1^\beta + (1-\alpha) l_2^\beta)^{\frac{1}{\beta}}$. We will first show that $\eta(l_1,l_2)$ and, in turn, $\mathbb{h}_{\alpha,\beta}(z)$ are increasing in β. From direct differentiation it follows that

$$\begin{aligned}
\frac{\partial}{\partial\beta}\eta(l_1,l_2) &= \frac{1}{\beta}\eta(l_1,l_2)\left[\frac{(1-\alpha)l_2^\beta \ln(l_2) + \alpha l_1^\beta \ln(l_1)}{\eta^\beta(l_1,l_2)} - \ln\left(\eta(l_1,l_2)\right)\right] \\
&= \frac{1}{\beta}\eta^{1-\beta}(l_1,l_2)\left[\alpha l_1^\beta \ln(l_1) + (1-\alpha) l_2^\beta \ln(l_2) - \eta^\beta(l_1,l_2)\ln\left(\eta(l_1,l_2)\right)\right] \\
&= \frac{1}{\beta^2}\eta^{1-\beta}(l_1,l_2)\left[\alpha l_1^\beta \ln(l_1^\beta) + (1-\alpha) l_2^\beta \ln(l_2^\beta) - \left[\alpha l_1^\beta + (1-\alpha) l_2^\beta\right]\ln\left(\alpha l_1^\beta + (1-\alpha) l_2^\beta\right)\right] > 0,
\end{aligned}$$

where the inequality comes from the convexity of the function $x\ln(x)$.

Thus, $\mathbb{h}_{\alpha,\beta}(z) = \eta(1,z)(1+z)$ is increasing in β. In particular, $\mathbb{h}^{-1}_{\alpha,\beta}(\cdot)$ is decreasing in β and we may conclude that, for each w_1, $f_{\alpha,\beta}(w_1) = w_1 \mathbb{h}^{-1}_{\alpha,\beta}(\Gamma^2/w_1^2)$ is decreasing in β as stated.

Lastly, since $f_{\alpha,\beta}(w_1)$ is decreasing in β, the feasibility set is increasing in β. Minimizing any convex function $g(\cdot)$ over a smaller set, yields a larger optimal value. Thus, the minimum is larger for smaller values of β. Q.E.D.

Proof of Lemma 3. Notice that if $C_1^r < C^0$, it must be the case that $\widetilde{\mathcal{R}}^c < \mathcal{R}_0^c$ where $\mathcal{R}_0^c$ is as in Proposition 2 and does not depend on β. In turn, by item (ii) there, the ratio is decreasing in β (the optimal ratio $z^* = l_2^*/l_1^*$ is increasing). Moreover, the ratio z^* is greater than 1 in this range. Since $w_2^*(z)$ is increasing in z, and because $f_{\alpha,\beta}(w_1)$ is increasing in β, we then have that w_2^* is increasing in β. In turn, the corresponding cost must be decreasing in β. From here we can conclude that κ_1 must be increasing in β. The argument is the same for the other parts of the lemma. Q.E.D.

Proof of Lemma 6. First, let us rewrite this problem equivalently as a minimization one:

$$\begin{aligned}&\min_{\mathbf{w},\lambda} 1/\lambda\\ &\text{s.t. } \frac{1}{\lambda} \geq \frac{\sum_{k\in[K]} \frac{a^k}{\theta_1^k} w_1^k}{C_1}; \quad \frac{1}{\lambda} \geq \frac{\sum_{k\in[K]} \frac{a^k}{\theta_2^k} w_2^k}{C_2},\\ &\quad \mathbf{w} \in \mathcal{W}^{\times}.\end{aligned}$$

An optimal solution must satisfy

$$\frac{1}{\lambda} = \max\left\{ \frac{\sum_{k\in[K]} \frac{a^k}{\theta_1^k} w_1^k}{C_1}, \frac{\sum_{k\in[K]} \frac{a^k}{\theta_2^k} w_2^k}{C_2} \right\} =: \mathcal{G}(w). \tag{31}$$

Notice that because the maximum operations preserves convexity, $\mathcal{G}(w)$ is a convex function. Since $\mathcal{W}^{\times}$ is a convex set, w^* can be found by solving the convex problem

$$\begin{aligned}&\min_{\mathbf{w}} \mathcal{G}(w)\\ &\text{s.t. } \mathbf{w} \in \mathcal{W}^{\times}.\end{aligned}$$

Once solved, λ^* is then given by (31). Q.E.D.

Proof of Lemma 4. Let us consider first the case that $\gamma = 0$, $y > 0$. In this case,

$$w_1 \geq \min \tau_1 N_{PC}(\tau) = \frac{1}{1+y\tau_2} \min_{\tau_1} \tau_1(1+y\tau_2)\left[1 - e^{-\frac{1}{2}\tau_1^2(1+y\tau_2)^2}\right]^{-1} = \frac{\Gamma}{1+y\tau_2},$$

where $\Gamma := \min_{x\geq 0} x\left[1 - e^{-\frac{1}{2}x^2}\right]^{-1}$. By construction, $W_2(\tau)/W_1(\tau) = y\tau_2$; therefore, we have the constraint $w_1 \geq \frac{\Gamma}{1+w_2/w_1}$, from which follows that $w_2 \geq (\Gamma - w_1) =: f_{\alpha,\beta}(w_1)$, as stated.

For sufficiency, suppose that $w_2, w_1 \geq 0$ satisfy the inequality $w_1 + w_2 \geq \Gamma$. We need to show that there exists τ_1, τ_2 such that $w_1 = \tau_1 N_{PC}(\tau)$ and $w_2 = y\tau_1\tau_2 N_{PC}(\tau)$. To this end, define

$$\tau_1 = \frac{w_1}{N}, \quad \tau_2 = \frac{w_2}{y w_1},$$

and let N be a solution (if one exists) to

$$N = \left[1 - e^{-\frac{1}{2}(\tau_1+\theta_2(\tau_1)\tau_2)^2}\right]^{-1} = \left[1 - e^{-\frac{1}{2}\left(\frac{w_1+w_2}{N}\right)^2}\right]^{-1}. \tag{32}$$

If such, a solution N that we have shown for which w_1, w_2 are indeed feasible and induced by the above choice of τ_1, τ_2. To see that (32) has indeed a solution for $w_1, w_2 \geq 0$ with $w_1 + w_2 \geq \Gamma$, let us rewrite this equation as

$$\frac{1}{N}\left[1 - e^{-\frac{1}{2}\left(\frac{w_1+w_2}{N}\right)^2}\right]^{-1} = 1. \tag{33}$$

Since the function $g(x) = x(1-e^{-x^2}) \to 0$ as $x \downarrow 0$, the left-hand side of (33) diverges as $N \uparrow \infty$. This left-hand side is a continuous function of N. To show that a solution exists, we then only have to show that there exists a choice of N (for this w_1, w_2) such that the left-hand side is smaller or equal to 1.

Multiplying and dividing this left-hand side we have

$$\frac{1}{w_1+w_2}\frac{w_1+w_2}{N}\left[1 - e^{-\frac{1}{2}\left(\frac{w_1+w_2}{N}\right)^2}\right]^{-1}.$$

Let

$$y^* := \arg\min_{y \geq 0} y\left[1 - e^{-\frac{1}{2}y^2}\right]^{-1};$$

This y^* is unique follows the proof of Lemma 1; set $N^0 = \frac{w_1+w_2}{y^*}$. Recalling that $\Gamma := \min \arg\min_{y\geq 0} y\left[1 - e^{-\frac{1}{2}y^2}\right]^{-1}$, we then have with this choice of N, that the left-hand side equals

$$\frac{\Gamma}{w_1+w_2} \leq 1,$$

where the last inequality follows from our assumption that $w_1 + w_2 \geq \Gamma$. We conclude then that at $N = N^0$,

$$\frac{1}{N}\left[1 - e^{-\frac{1}{2}\left(\frac{w_1+w_2}{N}\right)^2}\right]^{-1} \leq 1.$$

The fact that $\frac{1}{N}\left[1 - e^{-\frac{1}{2}\left(\frac{w_1+w_2}{N}\right)^2}\right]^{-1} \uparrow \infty$ as $N \uparrow \infty$, establishes the existence of N^* that solves (32), as required. This concludes the proof of sufficiency.

The proof for the case where $\gamma > 0$, $y = 0$ is similar and omitted. Q.E.D.

Proof of Lemma 7. We start by proving that $\mathbb{h}_{\alpha,\beta}(z)$ is increasing in z. We have

$$\mathbb{h}'_{\alpha,\beta}(z) = \left(-z^\beta(\alpha-1)+\alpha\right)^{\frac{1}{\beta}}\left(1 + \frac{(1+z)(-1+\alpha)}{z(-1+\alpha-z^{-\beta}\alpha)}\right),$$

where the first term is positive since $0.5 \leq \alpha \leq 1$ implies that

$$\alpha \leq -z^\beta(\alpha-1) \leq 0.5z^\beta + \alpha.$$

The second term is positive since both denominator and numerator in the quotient are negative for $\alpha \geq 0.5$; specifically,

$$-z^\beta + \alpha - 1 \leq -z^\beta\alpha + \alpha - 1 \leq -0.5z^\beta + \alpha - 1 \leq 0.$$

Finally, because $\mathbb{h}_{\alpha,\beta}(z)$ is increasing in z, $\sqrt{\mathbb{h}_{\alpha,\beta}(z)}$ is also increasing in z, and $1/\sqrt{\mathbb{h}_{\alpha,\beta}(z)}$ is decreasing in z. Q.E.D.