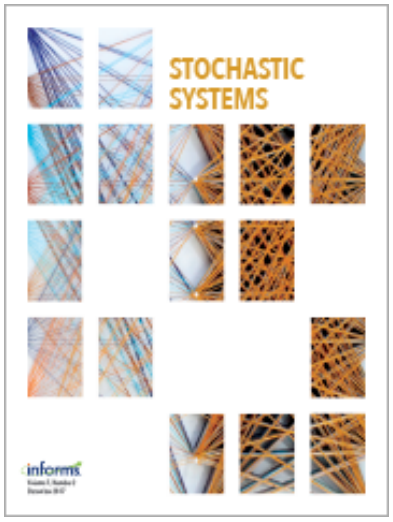


Stochastic Systems

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

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

Noa Zychlinski, Itai Gurvich

To cite this article:
Noa Zychlinski, Itai Gurvich (2026) The Production of Service: A Workload View of Complementarity and Substitution. Stochastic Systems

Published online in Articles in Advance 05 Jun 2026

. https://doi.org/10.1287/stsy.2025.0111

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Articles in Advance, pp. 1–21
ISSN 1946-5238 (online)

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

Noa Zychlinski,[a,*] **Itai Gurvich**[b]

[a] Faculty of Data and Decision Sciences, Technion–Israel Institute of Technology, Haifa 3200003, Israel; [b] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208

*Corresponding author

Contact: noazy@technion.ac.il, https://orcid.org/0000-0002-5125-3089 (NZ); i-gurvich@kellogg.northwestern.edu (IG)

Received: April 17, 2025
Revised: November 20, 2025; March 29, 2026
Accepted: April 2, 2026
Published Online in Articles in Advance: June 5, 2026

https://doi.org/10.1287/stsy.2025.0111

Abstract. We study optimal coordination in quality-driven rework systems, where each station contributes to an item's quality and the likelihood of rework depends on these contributions through a production function. Our model links quality targets to processing times via Brownian hitting times and aggregates quality into rework probability. By setting quality targets, a planner jointly shapes processing times and rework 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, whereas 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 using published operational statistics from industrial maintenance settings, in which equipment undergoes repair followed by preventive service while offline. The results show that the optimal design is robust to parameter misspecification, achieves substantial cost savings relative to benchmark policies, and can be estimated from standard operational data.

Funding: Partial financial support was provided by the Israel Science Foundation [Grant 277/21] and the Bernard M. Gordon Center for Systems Engineering at the Technion.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsy.2025.0111.

Keywords: processing networks • rework systems • multi-stage process design • complementarity and substitutability • stochastic modeling

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. Longer processing times improve output quality but require more capacity; at the same time, they reduce rework, making the net effect on workload ambiguous.

Rework 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 rework—representing a major share (Schiffauerova and Thomson 2006). In asset-intensive industries, such as construction and maintenance, direct rework costs typically account for 4%–6% of total project expenditures, and they can reach 12%–15% in complex or poorly coordinated projects, contributing to delays and cost overruns (Love and Li 2000, Hwang et al. 2009). In healthcare, similar challenges arise; in the United States, 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 Ghabowen 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 (Jencks et al. 2009). These figures show that rework 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 that we introduce in this paper and the questions that 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 step. The more effort allocated to the processing step, the better the quality of the product is 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 mapping from the quality of output—a vector that includes quality measures for each step of the process—and the likelihood of rework: 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 that 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 multiple distributions when two stations operate—affects the arrival rate.

In our model, illustrated in panel (a) of Figure 1, the planner chooses the workload (effort) levels $w = (w_1, w_2)$ at two processing stations. The arrival rate depends endogenously on the stations' effort levels through the rework mechanism: that is, $\lambda = \lambda(w)$. The presence of rework makes the attainable workload region nonlinear. This nonlinearity reflects the endogenous relationship between effort and workload, and it forms the basis for the optimization problems analyzed in the paper.

Although grounded in a dynamic quality-evolution model, the rework probability takes the simple form

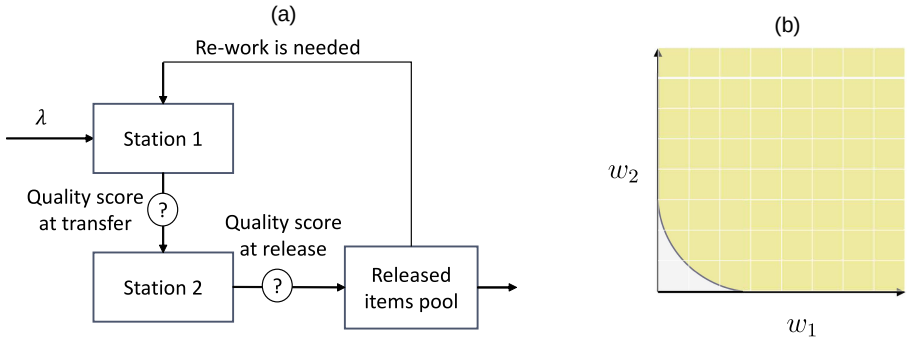
$$\log(\textit{probability of rework}) = -\text{Aggregated Quality}.$$

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 rework.

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

Figure 1. A Two-Station Model with Rework (a) and the Corresponding Attainable Workload Region (b)



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 rework. 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 and 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 rework, 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.

1.1. 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 it 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, whereas thorough postrepair 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 is. 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 bears all of the workload) are never optimal as long as there exists some complementarity. Systems with perfect substitution generally exhibit *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 and Varian 2010, chapter 4). There, with perfect substitution in the utility function, one has only corner solutions. Because the rework probability is a nonlinear 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 capacity is increasing in the complementarity; for other capacity levels, 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 nonobvious ways as capacity changes.

In Online Appendix B, we present three variants of the base model that demonstrate the flexibility and robustness of our results. These include *processing complementarity* (Online Appendix B.2), where the effectiveness—the quality-improvement rate—of the second station depends on the final score from the first station; *rework within a finite time window* (Online Appendix B.3); and *rework while in Station 2* (Online Appendix B.6). 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 as well as showcase their usefulness through the insights that they provide.

Organization. We start with a brief review of the relevant literature in Section 2. The single-station model described in Section 3 serves as an important building block. We introduce substitution/complementarity in the two-station model in Section 4. This section contains the key results of the paper as they pertain to the characterization of the attainable workload region. Section 5 draws on this characterization to study process optimization. Section 6 contains a concentrated discussion of the operational/managerial implications of our results. In Section 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 Online Appendix.

2. Literature Review

Coordinating 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 as well as to the literature on tandem stations with rework.

2.1. 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, Batt and Terwiesch 2012, Staats and Gino 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 rework. Such state-dependent service times also make workload forecasting 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 rework, and cross-stage complementarities.

Our work is fundamentally concerned with discretionary services, in which working procedures are not necessarily prespecified; employees or managers 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. Recently, Zychlinski et al. (2026) used a Brownian motion (BM) to model the evolution of a patient's health scores and determine optimal call-in thresholds in a hybrid hospital setting. 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 rework, 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, it is used to design the *item-specific* baseline around which further dynamic fine-tuning can be done.

2.2. Queues with Rework

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 queueing systems with rework has grown in recent years, partly because of their relevance in healthcare settings, where rework naturally arises in the form of patient readmissions—patients whose condition deteriorates after discharge and must return for additional treatment. and Shi et al. (2021) study operational decision making in such systems by explicitly modeling individual patient progression. Motivated by healthcare quality-improvement interventions, Chan et al. (2025) analyze control strategies aimed at reducing return probabilities in stylized queueing systems where customers may require multiple service episodes. Hybrid healthcare systems, in which in-person visits may be followed by unsuccessful virtual care and subsequent returns, are examined in Zychlinski (2024) and Liu et al. (2023). Recently, Zychlinski (2026) studied a queueing model of a mental health system with rework, where patients first receive group therapy and if it proves ineffective, transition to individual therapy.

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 multistage 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, which is 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 rework likelihood in a unified way.

Finally, models of complementarity and substitution have 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 rework. 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 item type. Items arrive following a renewal process with arrival rate λ. These are so-called "index arrivals" relative to which rework is measured.

The evolution of the quality score is modeled as a BM, whose drift captures the quality-improvement rate, and its standard deviation is the extent of randomness in improvement. The model captures item dynamics through the network. Because hitting times and rework probabilities admit closed-form expressions, it maps service-content decisions to outcomes in a tractable way.

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

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

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

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

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

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

so that $e^{-\nu}$ is the base probability of rework. The baseline probability p_b is close to one (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 zero for more complicated cases, where randomness is inherent.

Conditional on the problem not being—terminally and deterministically—solved, rework depends on the random evolution of the quality score. The postrelease (or discharge) score follows a BM $\mathcal{B}_t^{pr}$ (pr = postrelease), which starts at l and has a nonnegative drift $\eta(l) = \gamma \times l$ that is linear in l. This relation reflects the fact that the greater the time is spent on an item, the less likely this item is to require rework. The linear choice is not arbitrary; it arises as a special case of the multistation model that we introduce in Section 4.

The diffusion coefficient of this BM is $\sigma_{pr} > 0$. Although the improvement rate is positive, randomness allows the quality score to hit negative levels; it reaches *rework* 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 rework likelihood is then the probability that the positive-drift motion $\mathcal{B}_t^{pr}$ *starting at* l hits zero 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—and consequently, $p(l)$ increases—as the quality-improvement variability increases. The probability of rework accounting for base rework probability is

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

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

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

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

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

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

This expression implicitly handles the case where the initial score a (for the first visit) exceeds 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

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

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

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

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

We also observe that at optimality, the likelihood of rework $p(l^*) = e^{-(\nu+\varrho\gamma(l^*)^2)} = e^{-(\nu+(y^*)^2)}$, where $y^* = arg\min_{y \geq 0} y$ $[1 - e^{-(\nu+y^2)}]^{-1}$ 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 rework 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 increasing 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.

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 rework 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.

We extend the notation from the previous section by adding 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 $l_1 + l_2$; this value is the initial condition for the subsequent evolution of quality. After release from Station 2, the postdischarge 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 postrelease improvement or deterioration.

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

$$p(l) = e^{-(\nu + \varrho\eta(l)(e \cdot l))}, \qquad N(l) = [1 - p(l)]^{-1}, \tag{2}$$

where $e^{-\nu}$ is the base probability of no rework. For tractability, we assume that the rework processing follows the same protocol as the initial service so that the transfer targets l do not vary across visits.

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 rework. In all three scenarios, an item arrives at Station 1 at $t = 0$ with score $\mathcal{B}_0^1 = 0$; the quality level at each stage serves as the initial condition for the subsequent phase. In particular, the score at transfer from Station 1 to Station 2 is l_1, and the postdischarge process starts at the discharge level $l_1 + l_2$, ensuring continuity of the quality trajectory across stages. In Scenario 1 ($l_1 = 3.75, l_2 = 0$), the item is released at $t \approx 3$. Its quality deteriorates, and rework 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, rework occurs only after release from Station 2; see Online Appendix B.6 for a variant where a return to Station 1 can happen *while* in Station 2.

4.1. Complementarity and Substitution

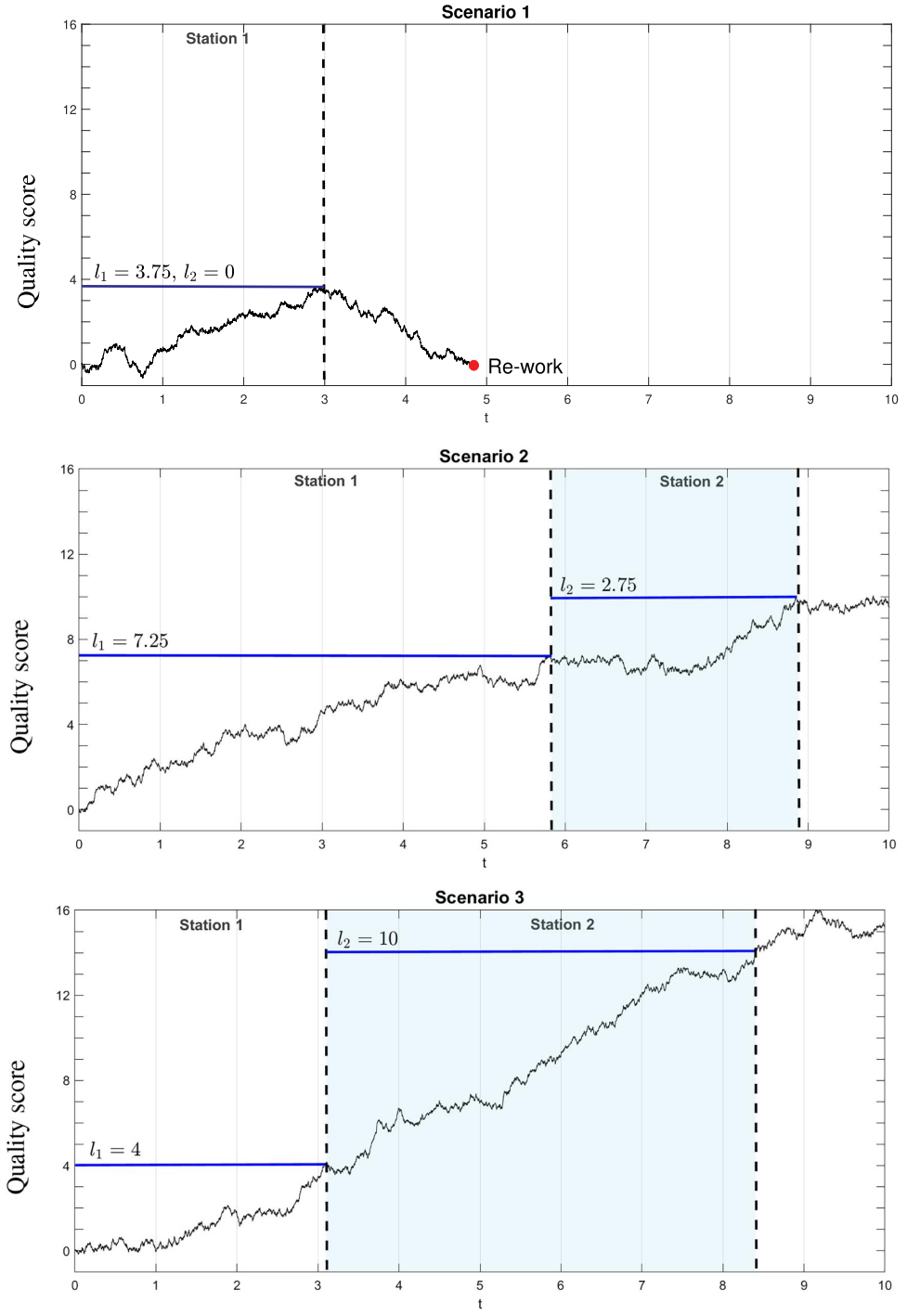
The efforts in both stations can be substitutive or complementary. Informally, the two stations are complements if as we increase l_1, the incremental effect of l_2 increases. Conversely, they are substitutes if the incremental effect of l_2 is independent of l_1, so the two efforts contribute additively 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 postrelease dynamics that govern rework 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 1967, Sato and Koizumi 1973, Stern 2011):

$$\eta(l) = (\alpha l_1^{\beta} + (1 - \alpha) l_2^{\beta})^{\frac{1}{\beta}}, \quad 0 < \alpha < 1, \quad \beta \leq 1, \tag{3}$$

Figure 2. Sample Paths Illustrating the Evolution of the Quality Score and the Timing of Rework



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 (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 that it captures that the total score is mapped into the rework 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 Section 3 taking $\gamma = 1$ there.

Station Labeling. Assuming $\alpha \geq 1/2$ without loss of generality, Station 1 ($\alpha \geq 1/2$, decision variable l_1) is the *primary station*, whereas Station 2 is the *secondary station*. When $\alpha > 1/2$, the primary station is more effective, contributing more than the secondary one.

4.2. 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]

$$W_1(l) = \frac{\lambda}{\theta_1}[l_1 N(l) - a], \quad W_2(l) = \lambda \frac{l_2}{\theta_2} N(l). \tag{4}$$

We define the *rate-normalized* workload as

$$W^r_{1,a}(l) = l_1 N(l) - a \quad \Rightarrow \quad W^r_1(l) = W^r_{1,a}(l) + a = l_1 N(l), \quad W^r_2(l) = l_2 N(l). \tag{5}$$

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

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

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

$$\Gamma = min_{y \geq 0}\ y\left[1 - e^{-(\nu + y^2)}\right]^{-1}. \tag{6}$$

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

$$e^{-(\nu + y^2)} = (1 + 2y^2)^{-1}. \tag{7}$$

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

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

where

$$f_{\alpha,\beta}(w_1) = \begin{cases} w_1 \mathbb{h}^{-1}_{\alpha,\beta}\left(\dfrac{\Gamma^2}{w_1^2}\right), & w_1 \leq w_1^0 \\ 0, & w_1 > w_1^0, \end{cases} \tag{8}$$

as well as

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

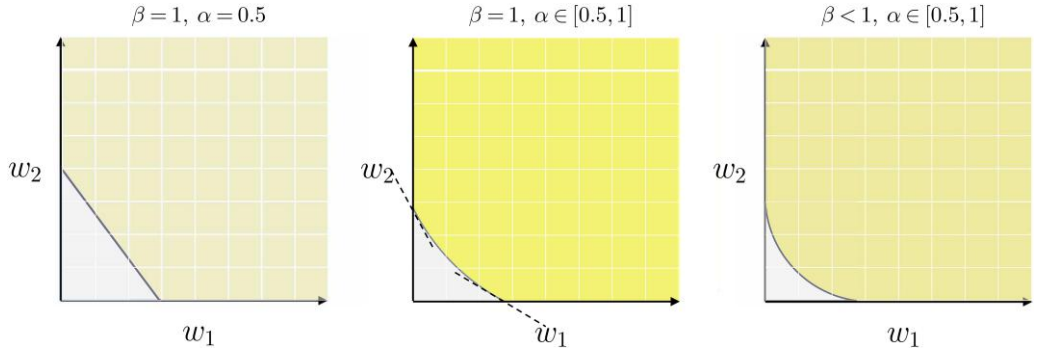
and Γ is as in (5). Furthermore, $\mathcal{W}$ can be expressed equivalently as

$$\mathcal{W} = \{(w_1, w_2) \geq 0 : \eta(w)(e \cdot w) \geq \Gamma^2/\varrho\}. \tag{9}$$

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

Figure 3. The Set for $\mathcal{W}$ Different Values of α, β



Notes. In the center panel, the boundary is convex but "hits" the axes with a strictly negative derivative (the tangents in black represent these derivatives). In the right panel, the tangents at the points where $f_{\alpha,\beta}$ meets the axes are the axes themselves.

concave), implying that the product $\eta(w)(e \cdot w)$ is log concave and that its upper contour sets are convex. Any positive aggregation function $\eta(\cdot)$ that is concave would also yield a convex feasible region W.

The characterization of the attainable workload region extends naturally to any number of stations n as shown in Online 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 $f_{0.5,1}(w_1) = \Gamma\sqrt{2/\varrho} - w_1$ for $w_1 \le w_1^0 = \Gamma\sqrt{2/\varrho}$. The set $\mathcal{W}$ in this case has the simple linear boundary shown in the left panel 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, $f_{\alpha,\beta}$ is as in (7).

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 $\gamma_1 = 2\alpha > 0$, $\gamma_2 = \frac{1}{2(1-\alpha)} > 0$.

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 $(l, 0)$, whereas 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 nonnegligible use of both stations. This is established in the next section.

4.3. 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 rework is $p(l) = e^{-(\nu + \varrho\eta(l)(e \cdot l))}$, with reworked items returning to Station 1. In the two-station setting, station-level quality contributions aggregate via the CES function $\eta(l) = (\alpha l_1^\beta + (1-\alpha) l_2^\beta)^{1/\beta}$. Online Appendix A summarizes all main notation and symbols used throughout the paper.

For completeness, Online 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 rework frequencies. The Online Appendix also includes a simulation demonstrating that the proposed estimation procedure accurately recovers these parameters even in moderate sample sizes under the assumed model specification.

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

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

Recalling (5), 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:

$$\begin{aligned} \min_{l \in \mathbb{R}_+^2} \quad & (c_1/\theta_1) \cdot W_1^r(l) + (c_2/\theta_2) \cdot W_2^r(l) \\ \text{s.t.} \quad & (\lambda/\theta_1) \cdot W_1^r(l) \le C_1, \\ & (\lambda/\theta_2) \cdot W_2^r(l) \le C_2. \end{aligned} \tag{10}$$

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

We consider the case where $C_1 = C_2 = \infty$ in (10). The resulting optimization problem is given by

$$V^* = \min_{l \in \mathbb{R}_+^2} \left(\frac{c_1}{\theta_1} W_1^r(l) + \frac{c_2}{\theta_2} W_2^r(l) \right). \tag{11}$$

5.1.1. 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, whereas $1/\theta_1$ 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

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

quantifies the *relative cost* of Stations 1 and 2. According to our labeling, Station 1 is the primary one, and thus, it is 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:

$$\overline{V}^*(\mathcal{R}^c) := \frac{\theta_1}{\lambda c_1} V^* = \min_{l \ge 0} \left(W_1^r(l) + \mathcal{R}^c W_2^r(l) \right) \quad \text{(cost-normalized problem)}. \tag{12}$$

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. Although we have no explicit resource budget constraints at this point, they are implicit through the rework; 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* (12) (*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*

$$\left(z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)}\right) = -\frac{1}{\mathcal{R}^c} \tag{13}$$

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* $l_2^* = z_0 y^*/\sqrt{\mathbb{h}_{\alpha,\beta}(z_0)}$, *where* y^* *are as in (7).*

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 rework 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 (6). The optimization problem adjusts the decision to the cost and substitution parameters so that the likelihood of rework is the same across all types (i.e., all combinations of cost and substitution) that have the same base rework parameter ν. This invariance follows from the exponential form of the return probability and the resulting first-order condition, which pin down the optimal value of $\varrho\,\eta(l)(e \cdot l)$ as a function of ν alone.

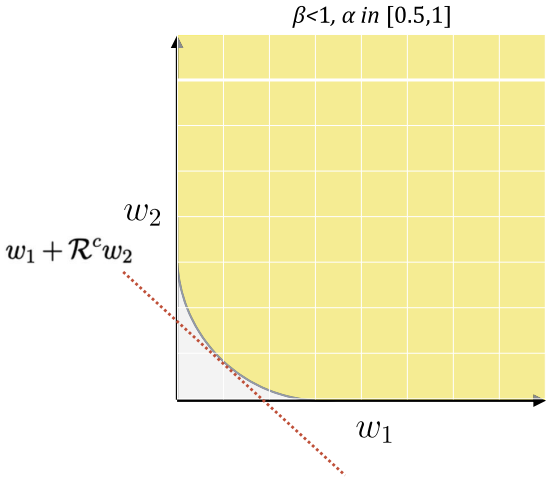
Problem (12) is a minimization of a linear function over a convex region. At the optimal solution (if attained at a smooth boundary), it must be the case that the cost function is a tangent to the boundary. In Proposition 1, we saw that with $\beta < 1$, the axes are the tangents to the function $f_{\alpha,\beta}$ 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 nontrivial. 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 interior solution applies only within the middle regime where both stations are active, whereas 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 center panel of Figure 3, where the tangents at the axes illustrate these corner solutions).

Theorem 2 (Operating Regimes). *The optimal solution to* (12) *is as follows.*

- *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* $0 < \mathcal{R}^c_{LB} \le \mathcal{R}^c_{UB} < \infty$ *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* y^* *in (7);*
  - *for* $\mathcal{R}^c \in (\mathcal{R}^c_{LB}, \mathcal{R}^c_{UB})$: $l^*(\mathcal{R}^c)$ *is as characterized in Lemma 2; and*
  - *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 (7).*

Figure 4. The Attainable Workload Region for $\beta < 1$ and the Cost Objective Tangent



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 postrelease 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 Stations 1 and 2. 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.

Figure 5 is a numerical illustration of the optimal decisions in Theorem 2. It shows symmetry around a threshold value, which is explicitly characterized in part (i) of Proposition 2. For completeness, we note that the qualitative behavior of the solution extends to $\beta \le 0$, 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.

5.1.2. 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. Online Appendix D grounds these results in an order-of-magnitude calibration based on published operational statistics from industrial maintenance settings for assets, such as wind turbines, generators, and other heavy equipment. The calibrated parameters produce consistent qualitative magnitudes aligned with observed practices, where repair is followed by preventive service while equipment remains offline. From an operational perspective, complementarity versus substitution can be diagnosed by how improvements in one stage affect the marginal impact of effort in the other stage. Online Appendix C provides an estimation procedure for this interaction from data. This distinction has direct design implications; complementarity favors balanced effort across stages, whereas substitution supports concentrating effort in the more cost-effective stage. Online 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.

Proposition 2 (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 $\mathcal{R}_0^c := (2(1-\alpha)+1)/(2\alpha+1) = -\left(z - \frac{2\mathbb{h}_{\alpha,\beta}(z)}{\mathbb{h}'_{\alpha,\beta}(z)}|_{z=1}\right)^{-1}$, $l_1^*(\beta, \mathcal{R}_0^c) = l_2^*(\beta, \mathcal{R}_0^c)$ (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 substitution 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 $\overline{V}^*(\beta, \mathcal{R}^c)$ is decreasing in β (with all else equal, greater complementarity means a larger value function).*

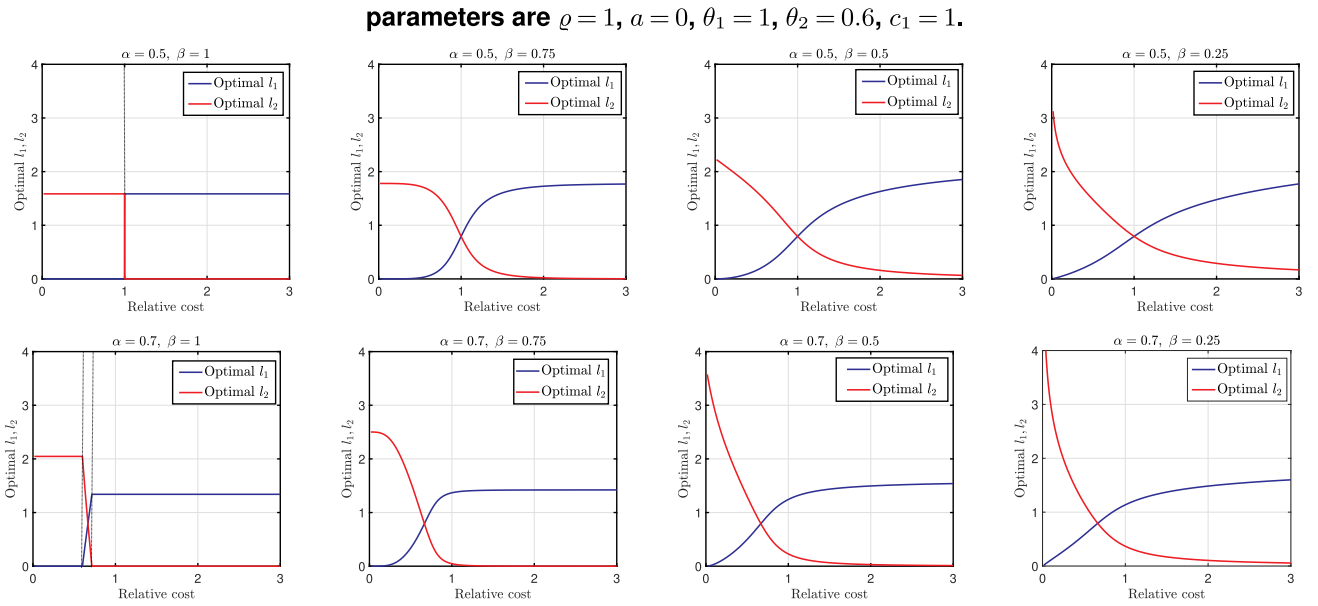
Figure 6 is an illustration of Proposition 2. At the simplest level, complementarity effectively introduces a constraint on the network; to benefit from the effort in one station, one must make greater effort in the other station. In turn, the higher the complementarity is (the smaller the value of β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

More subtle is the way that optimal decisions differ on both sides of the threshold $\mathcal{R}_0^c$. When $\alpha = 1/2$ and $\mathcal{R}_0^c = 1$, part (ii) of Proposition 2 stipulates that if $\mathcal{R}^c < \mathcal{R}_0^c$, the greater the complementarity is (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 station.

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	$\beta = 1$	$R^c \ge R^c_{UB}$	$l_1 = l_1^0$ (constant)
II. Station 2 only	$\beta = 1$	$R^c \le R^c_{LB}$	$l_2 = l_2^0$ (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; as complementarity increases ($\beta \downarrow$), effort allocation across stations becomes more balanced

Figure 5. The Solution Structure in Theorem 2: Optimal l_1, l_2 as a Function of Relative Cost $\mathcal{R}^c$



Note. The parameters are $\varrho = 1$, $a = 0$, $\theta_1 = 1$, $\theta_2 = 0.6$, and $c_1 = 1$.

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).

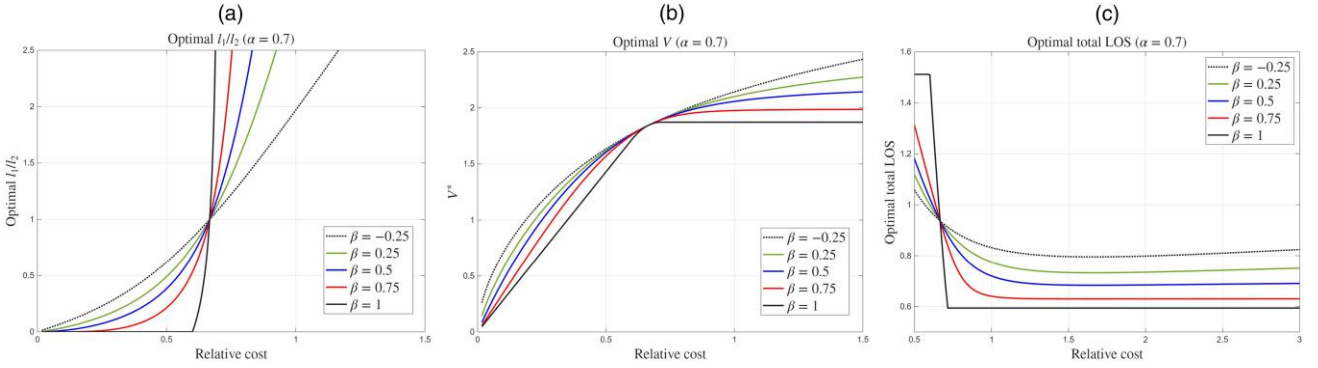
In Figure 6, $\alpha = 0.7$ so that threshold is per part (i) of Proposition 2 equal to 2/3. Complementarity keeps the stations as balanced as possible—as extracting value from one station necessitates keeping some effort in the other station. Accordingly, we see that as we move away from the symmetry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

5.1.3. Total Processing Time. Panel (c) of Figure 6 presents the total processing time (i.e., including all visits and both stations) for different values of β. Because per item type, the rework 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

Figure 6. Optimal l_1/l_2, V, and Total Length of Stay for Different Values of β



Notes. The parameters are $\varrho = 1$, $\theta_1 = 1$, $\theta_2 = 0.6$, and $c_1 = 1$. (a) Optimal l_1/l_2 ($\alpha = 0.7$). (b) Optimal V ($\alpha = 0.7$). (c) Optimal total length of stay ($\alpha = 0.7$).

station increases, the processing time initially decreases, but then, it may increase as Station 2 becomes prohibitively expensive; this latter increase is more pronounced when complementarity is high. 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) rework is constant within an item type, this is the "sweet spot." It achieves the optimal processing time and rework likelihood.

5.2. The Capacitated Problem

When capacity is finite ($C_1, C_2 < \infty$ in Problem (10)), the choice set is an intersection of the attainable workload region and the subspace defined by the constraints; see Figure 7.

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

$$C \in \mathcal{C}_\beta = \{C_1, C_2 : C_2^r \geq f_{\alpha,\beta}(C_1^r)\}, \tag{14}$$

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 (Optimal Capacitated Solution). *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 (10) is given by $l^{b,*}(\tilde{\mathcal{R}}^c)$, where*

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

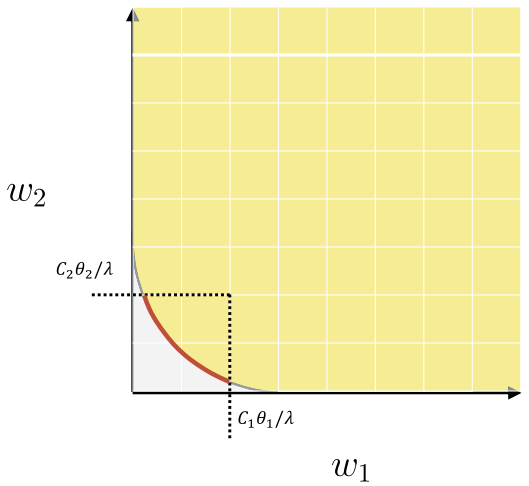
with

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

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 $\Gamma/\sqrt{\mathbb{h}_{\alpha,\beta}(1)} = \Gamma/\sqrt{2\varrho}$. The following is a direct consequence of Proposition 2.

Figure 7. A System with a Single Item Type and Finite Capacities C_1 and C_2



Note. The optimal solution to the cost optimization problem lies on the marked curved boundary portion.

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

5.2.1. 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 Online Appendix D.1 illustrate this mechanism. When repair capacity decreases from 10 to 6 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 p_k denote the fraction of arrivals of type k so that the arrival rate of type k items is $\lambda^k = p_k \lambda$ with $\sum_{k\in[K]} p_k = 1$. 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. We represent $\vec{l}$ as an element of $\mathbb{R}_+^{2K}$.

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

$$\mathcal{W}^\times := \prod_{k\in[K]} \mathcal{W}_k \tag{15}$$

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 $\mathbf{W}_i^r(\vec{l}) = (W_i^{r,1}(l^1), \dots, W_i^{r,K}(l^K))$ 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 $\mathcal{C}_{\boldsymbol{\beta}}$ 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 $\beta_0^k \le \beta_1^k$ for all $k \in [K]$.

In words, the greater the complementarity is between the stations, the greater the capacity is 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 multitype constrained optimization problem is, therefore,

$$\begin{aligned} \min_{\vec{l} \in \mathbb{R}_+^{2K}} \quad & (c_1/\theta_1) \cdot \mathbf{W}_1^r(\vec{l}) + (c_2/\theta_2) \cdot \mathbf{W}_2^r(\vec{l}) \\ \text{s.t.} \quad & (\lambda/\theta_1) \cdot \mathbf{W}_1^r(\vec{l}) \le C_1, \quad (\lambda/\theta_2) \cdot \mathbf{W}_2^r(\vec{l}) \le C_2. \end{aligned}$$

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

$$\begin{aligned} \min_{\mathbf{w}} \quad & (c_1/\theta_1) \cdot \mathbf{w}_1 + (c_2/\theta_2) \cdot \mathbf{w}_2 \\ \text{s.t.} \quad & (\lambda/\theta_1) \cdot \mathbf{w}_1 \le C_1, \quad (\lambda/\theta_2) \cdot \mathbf{w}_2 \le C_2, \\ & \mathbf{w} \in \mathcal{W}^\times. \end{aligned} \tag{16}$$

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 the station's capacity constraint is binding.

Theorem 4 (Multitype Capacitated Solution). *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 $\overline{l}^k(\cdot)$ denote the solution for type k items in Theorem 2 as a function of the relative cost. Then, the unique solution $\vec{l}^*$ to (16) satisfies*

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

where

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

and $\kappa = (\kappa_1^*, \kappa_2^*)$ are such that $\kappa_i^* \left(C_i - \mathcal{L}_i(\vec{l}^*) \right) = 0$, $i = 1, 2$. If $C_1 = \infty$, the corrected cost $\tilde{\mathcal{R}}_k^c$ 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 $\tilde{\mathcal{R}}_k^c$ 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 $\overline{l}_a^*(\mathcal{R}_a^c) > \overline{l}_b^*(\mathcal{R}_b^c)$ might have $\overline{l}_a^*(\tilde{\mathcal{R}}_a^c) < \overline{l}_b^*(\tilde{\mathcal{R}}_b^c)$. 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.

5.3.1. Maintenance Example. Consider a system that handles multiple categories of equipment that differs 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 because 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.

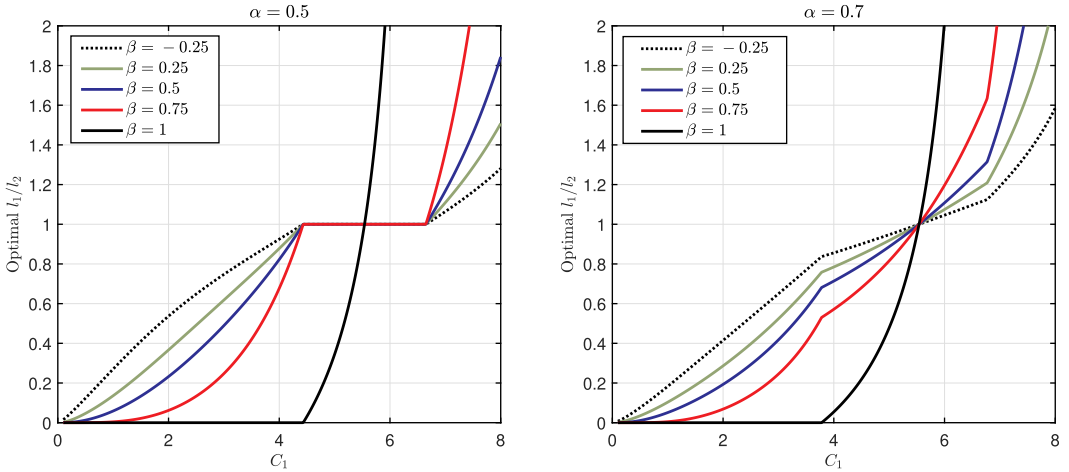
For more on the role of complementarity and substitution, Figure 8 provides two numerical examples. We consider $K = 5$ different levels of substitution/complementarity. When there is enough capacity, all items receive their required effort mix. As capacity becomes scarce, processing time in Station 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 is (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 stay shorter 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 of the required processing in Station 2. Indeed, when $C_1 \le 4.5$ (if $\alpha = 0.5$) and $C_1 \le 3.8$ (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 $C_1 \ge 6.5$ (if $\alpha = 0.5$) and $C_1 \ge 6.8$ (if $\alpha = 0.7$), $l_2^{k,*} = 0$ under perfect substitution; the entire service is provided in Station 1.

In Online Appendix B.4, 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 $\mathcal{L}(\vec{l}) = (\lambda/\theta_1) \cdot \mathbf{W}_1^r(\vec{l}) + (\lambda/\theta_2) \cdot \mathbf{W}_2^r(\vec{l}) \le C$. In that setting, the interesting question is how the total capacity is optimally allocated between the stations. In Online Appendix B.5, we consider the throughput maximization problem.

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 rework, reflecting the trade-offs captured by our model. In such systems, the planner must decide how much effort to

Figure 8. Optimal l_1^k/l_2^k, $k \in [5]$, Where Each k Corresponds to a Different Value of β



Notes. The parameters are $\varrho = 1$, $C_2 = \infty$, $\theta_1 = 1$, $\theta_2 = 0.6$, $\lambda = 1$, and $c_1 = c_2 = 1$. For $\alpha = 0.5$ in this example, capacity is binding if $C_1 \le 10$ and/or $C_2 \le 2.8$; for $\alpha = 0.7$, it is binding at $C_1 \le 9.7$ or $C_2 \le 1.16$.

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 postacute 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.

1. Rework, Workload, and the Operations Frontier

A fundamental feature of the model—and the reality that 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/rework. 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 is strictly convex regardless of the level of complementarity, except in the case of perfectly symmetric contribution ($\alpha = 1/2$), where $\beta = 1$ renders the boundary linear.

From a practical perspective, the nontrivial 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.

2. Operating Regimes and the Effect of Rework

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.

3. 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 toward 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.

4. 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 toward Station 2.

5. 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 Online Appendix 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 rework through substitution and complementarity. Optimizing effort allocation 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 rework and its associated cost. We introduce a model that

operationalizes item characteristics and subsequently, study how the latter determines the allocation of both stations' efforts.

The baseline model can be expanded to include additional 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 rework reveals information about the item. It then makes sense to have different transfer targets in subsequent (i.e., postindex) 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. Building in this baseline, real-time performance can be improved by dynamically adjusting the transfer targets in response to observed workload. The optimal control would introduce state-dependent discharge or transfer actions that are perturbations of the optimal fluid decisions.

Acknowledgments

The authors sincerely thank Editor Rami Atar, the anonymous Associate Editor and three reviewers for their insightful and constructive feedback, which has greatly strengthened this work.

Endnotes

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

[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 $a - l_1^*$.

References

Arrow KJ, Chenery HB, Minhas BS, Solow RM (1961) Capital-labor substitution and economic efficiency. *Rev. Econom. Statist.* 43(3):225–250.

Batt RJ, Terwiesch C (2012) Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, Wharton School, University of Pennsylvania, Philadelphia.

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

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

Carvalho VM, Tahbaz-Salehi A (2019) Production networks: A primer. *Annual Rev. Econom.* 11(1):635–663.

Chan TCY, Huang SY, Sarhangian V (2025) Dynamic control of service systems with returns: Application to design of postdischarge hospital readmission prevention programs. *Oper. Res.* 73(4):2242–2263.

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

Clark JR, Huckman RS (2012) Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Sci.* 58(4):708–722.

Dai JG, Weiss G (1996) Stability and instability of fluid models for reentrant lines. *Math. Oper. Res.* 21(1):115–134.

Fujiy BC, Ghose D, Khanna G (2022) Production networks and firm-level elasticities of substitution. Working paper, World Bank, Washington, DC.

Harrison JM (2003) Correction: Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probab.* 13(1):390–393.

Harrison JM, Van Mieghem JA (1997) Dynamic control of Brownian networks: State space collapse and equivalent workload formulations. *Ann. Appl. Probab.* 7(3):747–771.

Henningsen A, Henningsen G, Literáti G (2021) Econometric estimation of the "constant elasticity of substitution" function in R: The micEconCES package. Hashimzade N, Thornton MA, eds. *Handbook of Research Methods and Applications in Empirical Microeconomics* (Edward Elgar Publishing, Cheltenham, UK), 596–640.

Hopp WJ, Iravani S, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Hwang B-G, Thomas SR, Haas CT, Caldas CH (2009) Measuring the impact of rework on construction cost performance. *J. Construction Engrg. Management* 135(3):187–198.

Jencks SF, Williams MV, Coleman EA (2009) Rehospitalizations among patients in the medicare fee-for-service program. *New England J. Medicine* 360(14):1418–1428.

Jiang HJ, Hensche M (2023) Characteristics of 30-day all-cause hospital readmissions, 2016–2020. HCUP Statistical Brief No. 304, Agency for Healthcare Research and Quality, Rockville, MD.

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

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

Koesler S, Schymura M (2015) Substitution elasticities in a constant elasticity of substitution framework—Empirical estimates using nonlinear least squares. *Econom. Systems Res.* 27(1):101–121.

Kum Ghabowen I, Epane JP, Shen JJ, Goodman X, Ramamonjiarivelo Z, Zengul FD (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* 12(7):750.

Kumar PR (1993) Re-entrant lines. *Queueing Systems* 13(1):87–110.
Liu N, Wang S, Zychlinski N (2023) Channel management in outpatient care: Implications of telemedicine and transportation support. Preprint, submitted April 12, http://dx.doi.org/10.2139/ssrn.4383199.
Love PED, Li H (2000) Quantifying the causes and costs of rework in construction. *Construction Management Econom.* 18(4):479–490.
Madan KC (2000) An M/G/1 queue with second optional service. *Queueing Systems* 34(1):37–46.
Mahaboob B, Venkateswarlu B, Sankar JR (2017) Estimation of parameters of constant elasticity of substitution production functional model. *IOP Conf. Ser. Materials Sci. Engrg.* 263(1):042121.
Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic Theory*, vol. 1 (Oxford University Press, New York).
Netessine S, Zhang F (2005) Positive vs. negative externalities in inventory management: Implications for supply chain design. *Manufacturing Service Oper. Management* 7(1):58–73.
Sato K (1967) A two-level constant-elasticity-of-substitution production function. *Rev. Econom. Stud.* 34(2):201–218.
Sato R, Koizumi T (1973) On the elasticities of substitution and complementarity. *Oxford Econom. Papers* 25(1):44–56.
Schiffauerova A, Thomson V (2006) A review of research on cost of quality models and best practices. *Internat. J. Quality Reliability Management* 23(6):647–669.
Shi P, Helm JE, Deglise-Hawkinson J, Pan J (2021) Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Oper. Res.* 69(6):1842–1865.
Staats BR, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Sci.* 58(6):1141–1159.
Stern DI (2011) Elasticities of substitution and complementarity. *J. Productivity Anal.* 36(1):79–89.
Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Sci.* 60(6):1574–1593.
Varian HR (2010) *Intermediate Microeconomics: A Modern Approach* (W. W. Norton & Company, New York).
Wang X, Huang R, Gao J, Debo LG (2021) Managing discretionary services: A review and research opportunities. *J. Systems Sci. Systems Engrg.* 30(1):1–16.
Wilson SF, Shorten B, Marks RMI (2005) Costing the ambulatory episode: Implications of total or partial substitution of hospital care. *Australian Health Rev.* 29(3):360–365.
Zychlinski N (2024) Managing queues with reentrant customers in support of hybrid healthcare. *Stochastic Systems* 14(2):167–190.
Zychlinski N (2026) An operational view on managing mass trauma events. *Manufacturing Service Oper. Management* 28(1):76–99.
Zychlinski N, Mendelson G, Daw A (2026) Optimal call-in policies under travel-induced risk: Application to hybrid hospitalization. *Queueing Systems* 110(1):6.