Poolíng Queues with Discretionary Service Capacity

Mor Armony
Stern School of Business, New York University, New York, NY 10012, marmony@stern.nyu.edu

Guillaume Roels
UCLA Anderson School of Management, Los Angeles, CA 90095, groels@anderson.ucla.edu

Hummy Song
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, hummy@wharton.upenn.edu

Contrary to traditional queueing theory, recent case studies in health care and call centers indicate that pooling queues may not necessarily result in less expected work-in-process. In this paper, we propose that this phenomenon may arise when servers have some discretion over their choice of service capacity and are work-averse, i.e., bear a cost associated with either their expected workload or their degree of busyness. Under a pooled configuration, the servers’ incentives to invest in service capacity are indeed lower due to their ability to free ride and due to the greater operational efficiency of the pooled configuration. Moreover, the type and extent of work aversion matter; specifically, we find that dedicated configurations yield less expected work-in-process than pooled configurations when servers exhibit high degrees of workload aversion or low degrees of busyness aversion. We also find that busyness aversion tends to hurt more to the point that it could negate the operational benefits of queue pooling at their highest potential. Overall, our work suggests that service system designers may need to consider the servers’ type and extent of work aversion as well as their degree of capacity choice discretion before pooling their workload.

Key words: Queueing Theory, Game Theory, Behavioral Operations Management, Server Pooling

History: April 12, 2017

1. Introduction

Traditional queueing theory suggests that pooled queue configurations are more efficient than dedicated queue configurations (Smith and Whitt 1981). By allowing customers to be served by any available server rather than having them wait for a specific server to become available, pooled queue configurations help mitigate the negative effects of variability in arrivals, which leads to shorter waiting times for service, less expected throughput times, and less expected work-in-process (WIP) (Ata and Van Mieghem 2009, Gans et al. 2003).

However, recent empirical work suggests that, in some knowledge-intensive services, pooling queues may not necessarily yield these benefits in practice. Using data from an emergency department, Song et al. (2015) find that patients experience longer average wait times and lengths of stay when physicians are assigned patients under a pooled queueing system as opposed to a dedicated queueing system. In the context of a call center, Jouini et al. (2008) illustrate that moving from a pooled system to a team-based dedicated system leads to improvements in service rates and reductions in call-back rates. In both cases, financial incentives could not explain the result because servers were paid a fixed wage (personal communication with Hummy Song, May 2016; personal
communication with Oualid Jouini, May 2016). The authors suggest that these effects may arise from servers’ behavioral responses (i.e., servers’ intrinsic preferences) to the queue configuration, specifically, they suggest that servers may work more efficiently when in a dedicated queueing system due to an increased sense of ownership and responsibility over their customers. These case studies provide empirical support to findings from the economics and organizational behavior literatures, which propose that free riding (a.k.a. social loafing) is stronger in a pooled system relative to a dedicated system (Bendoly et al. 2010) since work is distributed among multiple agents in the former (Holmström 1982, Latané et al. 1979). Nevertheless, while the existing literature points to free riding as a potential explanation, it does not formally establish that causal link.

In this paper, we develop a game-theoretic model that explains why and when a pooled queue configuration may lead to more expected WIP than a dedicated queue configuration when servers are paid a fixed wage. We consider a two-server (pooled or dedicated) queueing system with Poisson arrivals and exponential service times. In line with the aforementioned practical applications (e.g., Song et al. (2015)) and as is typical in most knowledge-intensive services (Hopp et al. 2007), we assume that servers have discretion over their choice of service capacity (which we assume is not adjusted dynamically) and that they exhibit work aversion, i.e., that they incur a (psychological) cost as a function of the amount of work that is assigned to them. Like risk aversion, work aversion is an expression of preference (e.g., for less work) rather than suggestive of a behavioral bias.

We distinguish and consider two different types of work aversion: workload aversion and busyness aversion. Under workload aversion, servers are averse to high levels of workload. Servers’ sensitivity to varying levels of workload has been documented by a large body of empirical research (Berry Jaeker and Tucker 2016, Edie 1954, Hasija et al. 2010, KC 2013, KC and Terwiesch 2009, Oliva and Sterman 2001, Tan and Netessine 2014). Workload aversion may also arise from a sense of ownership and responsibility over jobs, such that servers exhibit empathy for the customers that are waiting to be processed. In essence, under workload aversion, servers internalize the associated holding costs (Shunko et al. 2017, Song et al. 2015, Valentine and Edmondson 2015). In contrast, busyness aversion captures the idea that servers may have a preference for idleness over busyness. Prior literature has modeled busyness aversion either as a fairness constraint (Armony and Ward 2010) or as an intrinsic preference (Doroudi et al. 2011, Gopalakrishnan et al. 2016, Wang and Zhou 2016, Zhan and Ward 2015). This latter type of work aversion is not necessarily selfish or lazy, but may simply characterize servers as wanting a break from work, especially when utilization is high. In sum, workload and busyness aversion have different focuses: Similar to the fill rate and stockout service levels in inventory control, workload aversion is focused on the jobs in process (e.g., customers), whereas busyness aversion is focused on the resources processing those jobs (i.e., servers).
With this framework of work aversion in mind, we address the following research questions: To what extent does work aversion affect the performance of a particular queue configuration? For a particular type of work aversion, which queue configuration leads to the lowest expected WIP, and how does it depend on the degree of servers’ discretion over the service capacity? If free riding makes a pooled configuration less efficient, how significant is this loss of efficiency and how does it depend on the type of work aversion?

Our findings suggest that work aversion may in fact explain why, in the aforementioned case studies (Jouini et al. 2008, Song et al. 2015), pooling may lead to an increase in expected WIP (or expected throughput time) relative to a dedicated system. Our analyses yield the following three results.

First, we find that there is a diminished incentive to invest in service capacity when subject to a pooled queue configuration. This can be attributed to two factors: the less salient need to invest in service capacity given the greater operational efficiency of a pooled system, and the servers' incentive to engage in free riding in a pooled system.

Second, we find that whether the dedicated system outperforms the pooled system depends on the type of servers’ work aversion as well as on their degree of discretion over capacity. Specifically, dedicated queue configurations yield less expected WIP than pooled queue configurations when servers exhibit high degrees of workload aversion. In contrast, dedicated queue configurations outperform pooled queue configurations in this way when servers exhibit low degrees of busyness aversion. Under either type of work aversion, we find that dedicated systems tend to outperform pooled systems when servers have more discretion over capacity.

Third, we find that not only does the direction of this effect depend on the type of work aversion at hand, but so does its magnitude. Specifically, while the loss of efficiency from choosing a suboptimal queue configuration (pooled versus dedicated) is mild under workload aversion, it can be very significant under busyness aversion, especially at high levels of utilization, which happens at low levels of work aversion, due to the dominance of the capacity cost over the work aversion cost. In fact, busyness aversion may hurt so much that it may annihilate the operational benefits of a pooled configuration at their highest potential.

These results have the following two managerial implications for the design of service systems:

- In systems in which servers have a high degree of discretion over service capacity, service system designers need to understand the type and magnitude of work aversion before choosing the queue configuration.

- Given that busyness aversion hurts more than workload aversion, if servers are busyness-averse, service providers should strive to either standardize processes to restrict the servers’ discretion over service capacity, or have them shift their focus from resources (i.e., themselves) to jobs (e.g.,
customers), e.g., by drawing their attention to the amount of workload or developing greater empathy towards customers.

The rest of the paper is organized as follows. Section 2 provides a literature review. Section 3 describes our model. Sections 4 and 5 consider the cases of workload aversion and busyness aversion, respectively. Section 6 compares operational performance under each of the different queue configurations and work aversion conditions. Section 7 concludes. Proofs and auxiliary results are provided in the Appendix.

2. Literature Review

There has been a growing body of work in operations management that compares pooled and dedicated queueing systems and finds that pooled systems may not always lead to more operationally efficient outcomes, contrary to what traditional queueing theory predicts (Smith and Whitt 1981). In what follows, we review prior work in this stream of literature and describe various factors that may lead one to question the benefits of pooling queues. Throughout this paper, we say that a queue configuration is more efficient than another if it leads to less expected WIP (or less expected throughput time); see Cao et al. (2016) for other comparison criteria such as probability of delay. We broadly categorize this work into two groups based on whether the choice of service capacity is assumed to be exogenous or endogenous.

We first consider the case in which service capacity is exogenously determined. Here, one reason for why pooling queues may not lead to operationally efficient outcomes is that when customers jockey between separate queues, they can mitigate the inefficiencies of dedicated queues (Rothkopf and Rech 1987).

Another condition under which dedicated queueing systems may outperform pooled queueing systems is when there are asymmetric service capacities across different servers. As is illustrated in Rubinovitch (1985), when servers have significantly different service capacities and traffic intensity is low, it may be more efficient to allocate less demand to the slow server. Depending on the mean delay in steady state, it may in fact be better to operate without the slow server altogether. One way to mitigate the possibility of such a scenario would be by employing intelligent flow control, which would allow the customer to choose between being served right away by a slow server or waiting to be served by a fast server (Loch 1998, Rubinovitch 1985).

Besides heterogeneity in service capacities, another factor that may mitigate the benefits of pooling is heterogeneity in customer types. Pooling customers with different service requirements or service time distributions introduces greater service variability, which in turn reduces system performance (Benjaafar 1995, Smith and Whitt 1981). In the context of call centers, van Dijk and van der Sluis (2008) find that a pooled system strictly dominates only when there is a single
call type as opposed to two different call types. They also find that with different call types, pooling is not only far less advantageous, but may also have a negative effect, the directionality of which depends on the mix ratio of the two types of calls and the traffic load. Considering health care settings, Joustra et al. (2010) and Saghafian et al. (2012) similarly conclude that streaming customers into dedicated queues may lead to improved performance when the different customer streams have different service requirements.

Next, we turn to factors that come into play when service capacity is endogenously determined. One important factor that may influence service capacity is the presence of financial incentives. When servers are paid per unit of service rendered, they are incentivized to compete for market share by adjusting their service capacity. Under certain demand allocation schemes, the competition for market share is higher in a dedicated system than in a pooled system. Building on the model by Kalai et al. (1992), Gilbert and Weng (1998) propose an allocation policy that routes customers to servers to balance expected waiting time. The authors illustrate that in this scenario, the competition for market share is more intense in a dedicated queueing system than in a pooled queueing system and will result in higher service capacity, and potentially less WIP, in the dedicated queue configuration. These findings extend to cases of multiple servers, as is shown in Choi et al. (2011). Cachon and Zhang (2007) further this stream of work by investigating linear and proportional allocation policies. In related work, Wee and Iyer (2011) provide a holding cost allocation scheme that incentivizes servers to increase their service rates for a given demand allocation scheme. The authors show that expected WIP is less in a pooled system than in a dedicated system under that scheme, when firms operate in a perfectly competitive market and earn zero profit.

The operational efficiency of pooled versus dedicated queueing systems can also be affected by the servers’ behavioral responses to such factors as salience of performance feedback, social comparisons, (Mas and Moretti 2009, Song et al. 2017), or work aversion, which is the focus of this paper.

Some work has begun to explore these factors in the context of pooled versus dedicated queue configurations. In a series of behavioral experiments, Shunko et al. (2017) find that servers work faster when performance feedback is made more salient by increasing the visibility into the length of the queue. Furthermore, the authors find that servers work faster in a dedicated queueing system, which they attribute to reduced free riding and more salient individual performance feedback. In a field setting of a call center, Jouini et al. (2008) similarly find that a team-based dedicated system outperforms the pooled system. The authors attribute this to the limited team size in the dedicated system allowing for better human resource management, which increases agents’ motivation and responsibility over customers in addition to enabling performance comparisons across teams. In a field setting of an emergency department, Song et al. (2015) find these results to hold even when
the dedicated system is individual-based as opposed to team-based and when there is the same level of full visibility into the length of the queue under either the dedicated or pooled system. They report that emergency department physicians increase their service rates when in a dedicated queueing system, and suggest that it is due to an increased sense of ownership over patients and resources. Building on this empirical work, Do et al. (2015) develop a set of theoretical models to examine the effects of free riding under the assumption that the service rate is decreasing in the number of servers.

In this paper, we focus on the aspect of work aversion specifically by assuming each of the conditions that prior work has formally shown to lead to greater operationally efficiencies given a pooled queueing system. Specifically, we assume that customers are randomly assigned to queues with no opportunity for jockeying, servers are symmetric, customers have homogeneous service requests, servers are paid a fixed wage independent of service rate, servers have complete information about their workload, and servers do not experience any benefit or loss from social comparisons. Our focus on work aversion is informed by practice and is motivated by the emergency department setting studied in Song et al. (2015), where the authors explicitly refer to dedicated queueing systems affording a higher level of ownership over jobs than do pooled queueing systems. This allows us to formalize this notion of job ownership, common in the organizational behavior literature (Campion et al. 1993, Hackman and Oldham 1976). Here, free riding emerges as a result of this behavioral phenomenon, not as an assumption as is the case in Do et al. (2015).

Another contribution of this work is that we separately consider two distinct forms of work aversion depending on whether servers are job-focused or resource-focused.

The first type of work aversion we consider is workload aversion. This is motivated by prior work that illustrates that servers make dynamic service capacity adjustments in response to varying levels of workload (for a recent review, see Delasay et al. (2016)). Workload aversion is, therefore, job-focused. Across a variety of service settings, prior work has shown that varying levels of workload may lead to increasing (Edie 1954), decreasing (Debo et al. 2008, Hasija et al. 2010), inverted U-shaped (KC 2013, KC and Terwiesch 2009, Tan and Netessine 2014), or N-shaped (Berry Jaeker and Tucker 2016) responses of service time, which are sometimes at the expense of quality (KC and Terwiesch 2009, Oliva and Sterman 2001). Workload aversion may also arise from servers’ sense of ownership and responsibility over the jobs that still remain in the queue. What is described as responsibility in Song et al. (2015) and Shunko et al. (2017) and as accountability in Valentine and Edmondson (2015) reflects this notion that servers care about the customers remaining in queue and seek to clear their workload.

The second type of work aversion we consider is busyness aversion. This is a resource-focused notion that is motivated by the idea that servers have a preference for idleness over busyness.
We find empirical support for busyness aversion, wherein servers work more slowly at high levels of workload, given that they can no longer overcome high levels of workload by speeding up (Berry Jaeker and Tucker 2016, KC and Terwiesch 2009). Thus far, busyness aversion has been modeled in one of two ways. First, Armony and Ward (2010) and Ward and Armony (2013) modeled it as a fairness constraint that seeks to balance the steady-state server idleness proportions across servers, arguing that a faster-server-first (FSF) policy may be perceived as unjust and lead to low levels of server satisfaction, which in turn may have negative implications for server retention. Second, other papers have modeled busyness aversion as an intrinsic preference, and this is the approach we adopt here. For example, Doroudi et al. (2011) and Gopalakrishnan et al. (2016) model server utility as the sum of an effort cost (a decreasing function of service rate) and a reward from idleness (an increasing function of the idle time experienced). Similarly, Wang and Zhou (2016) model a server’s capacity cost as discontinuous in effort, with a fixed cost incurred only when the server is busy. Zhan and Ward (2015) examine how joint staffing and compensation policies may affect service rates when idle time is incorporated into a server’s utility function, even if servers are paid only for the time they are busy.

In this paper, we relate these two extreme types of work aversion to the choice of queue configuration and examine why and when a dedicated queueing system may outperform a pooled queueing system, given the particular type of work aversion.

3. Model

3.1. Setting

Consider a queueing system with two symmetric servers. Customers (or jobs) arrive according to a Poisson process with rate $2\lambda > 0$. Servers simultaneously and non-cooperatively determine their service rates (capacities) in order to minimize their individual steady-state costs. Customer requests are random; specifically, we assume that service times are independent and exponentially distributed, and the mean service rate is determined by the server’s choice of capacity.

We consider two queueing system configurations: dedicated and pooled. Under the dedicated configuration we have two independent $M/M/1$ queues, where each server has his own infinite-buffer first-come-first-served (FCFS) queue and customers are routed randomly and uniformly between the two queues so that each of the queues has an arrival rate $\lambda$. Under the pooled configuration we have an $M/M/2$ system with a single infinite-buffer FCFS queue and an arrival rate $2\lambda$; when a customer arrives into an empty system, she is equally likely to be routed to either server.

---

1 We have also considered the case with three servers numerically, and our insights appear to remain the same.

2 We have also considered the case of servers adjusting their service rates dynamically as a function of their workload (as in George and Harrison (2001)) using simulations, and our insights appear to carry over to that setting. Details are available upon request.
Let $S_{\text{dedicated}}$ and $S_{\text{pooled}}$ be the sets of achievable states under dedicated and pooled configurations, respectively. Specifically, $S_{\text{dedicated}} = \{0, 1, 2, \ldots\}$ is the set of achievable states for one of the queues in the dedicated system, and $S_{\text{pooled}} = \{0, \{1, i = 1, 2\}, 2, 3, \ldots\}$, where $j = 0, 1, \ldots$ denotes the state where there are $j$ customers in the system, and $1_i$ denotes the state where there is only one customer in the system and this customer is being served by Server $i$. Let $P_{\text{dedicated}}(s; \mu)$ be the steady-state probability of being in state $s \in S_{\text{dedicated}}$ under a dedicated queue configuration with service rate $\mu$. Similarly, let $P_{\text{pooled}}(s; \mu_i, \mu_{-i})$ be the steady-state probability of one of the queues being in state $s \in S_{\text{pooled}}$ under a pooled queue configuration with service rates $(\mu_i, \mu_{-i})$, with the notation $-i = 3 - i$, representing the server other than Server $i$.

We are interested in how the different queue configurations affect the servers’ capacity investments and the resulting system performance. Note that under the pooled configuration, the servers’ choices of capacities are interdependent, because the system steady-state probabilities are a function of both servers’ choices of capacity. We use the concept of Nash equilibrium and, given the servers’ symmetry, restrict our attention to symmetric Nash equilibria. We describe the relevant costs next.

### 3.2. Cost Structure

We consider two additive cost components associated with each server: work-aversion cost and capacity cost.

**Work-Aversion Cost:** To define the work-aversion cost, denote by

- $h(s)$: A server’s work-aversion cost per unit of time in state $s$ of a dedicated system, $s \in S_{\text{dedicated}}$;
- $h_i(s)$: Server $i$’s work-aversion cost per unit of time in state $s$ of a pooled system, $s \in S_{\text{pooled}}$.

We consider two types of work aversion:

- Workload aversion: Under workload aversion, the server incurs a cost when the workload (i.e., the number of jobs or customers) in the system builds up. We consider a linear cost of workload. Accordingly, for the dedicated version we define $h(j) = w \cdot j$, with $w > 0$. For the pooled system we let $h_i(j) = w \cdot j/2$, $j \neq 1$, and $h_i(1) = h_{-i}(1) = w/2$. For comparison purposes, we assume the same workload-aversion intensity $w$ for both the dedicated and the pooled system.\(^3\)

\(^3\)This appears to be without loss of generality; for workload aversion, we show in Proposition 1 that there is no asymmetric equilibrium, and for busyness aversion we have conducted extensive numerical simulations and have never encountered asymmetric Nash equilibria.

\(^4\)We have also considered other forms of workload aversion, namely, one that allocates the full cost of holding a customer in service to the server who is serving that customer (resulting in higher capacity investments), one that allocates the cost proportionally to the service capacities (resulting in lower capacity investments), as well as quadratic functions of workload. Although we were not able to establish the existence of a Nash equilibrium in general, our numerical simulations on specific instances suggest that a symmetric equilibrium exists; moreover all of our performance analysis results remain consistent, with higher or lower capacity investments. Details are available upon request.
Busyness aversion: Under busyness aversion, the server incurs a cost when she is busy. Accordingly, \( h(s) = b > 0 \) when the server is serving a customer in the dedicated system in state \( s \in S_{\text{dedicated}} \), and \( h_i(s) = b \) when Server \( i \) is serving a customer in the pooled system in state \( s \in S_{\text{pooled}} \). In particular, \( h(0) = h_i(0) = h_{-i}(1_i) = 0 \), \( h_i(1_i) = h(j) = h_i(j) = b \), for all \( j \geq 2 \). For comparison purposes, we assume the same busyness-aversion intensity \( b \) for both the dedicated and the pooled system.

Capacity Cost: Let \( c(\mu_i) \) represent the cost per unit of time incurred by Server \( i \) of providing capacity \( \mu_i \). We consider a capacity cost function of the form \( c(\mu) = c \cdot \left( \frac{\mu}{\mu_0} \right)^k \), where \( \mu_0 \) may be interpreted as the base service rate, and the cost elasticity \( k \geq 1 \) captures the extent of the lack of discretion that the server has over the choice of capacity, similar to Hopp et al. (2007) (as we establish in Appendix EC.1).

To gain intuition into measuring the servers’ lack of discretion as \( k \), consider a very large value of \( k \). For \( \mu > \mu_0 \), the capacity cost becomes prohibitively high very quickly, so the server would not be tempted to choose \( \mu \gg \mu_0 \). On the other hand, if \( \mu \leq \mu_0 \), the capacity cost is fairly stable for any \( \mu \in [0, \mu_0] \); in this case the server is somewhat indifferent, in terms of capacity costs, in her choice of capacity for any \( \mu \in [0, \mu_0] \), but given the presence of work-related costs (see above), she would tend to choose \( \mu \) close to \( \mu_0 \). Hence, when \( k \) is very large, the server will tend to choose her capacity in the neighborhood of the base capacity \( \mu_0 \). In fact, when \( k \to \infty \), the server will choose her capacity at \( \mu_0 \), i.e., capacity is exogenous, as in standard queueing theory (Kleinrock 1975). In contrast, when \( k \) is small, the server has a high level of discretion over her choice of capacity and may choose a capacity that is significantly lower or higher than \( \mu_0 \) while avoiding a large impact on capacity costs. For simplicity, and without loss of generality, from hereon we normalize the cost parameters so that \( \mu_0 = 1 \) and \( c = 1 \).

Our capacity cost function thus captures, through the elasticity parameter \( k \), the degree of discretion in service capacity around a base capacity \( \mu_0 \). When \( k \approx 1 \), servers have a high degree of discretion, whereas when \( k \) is very large, servers have little or no discretion around the base capacity \( \mu_0 \). To illustrate this concept of service discretion in practice, consider the following examples. In a call center setting, filling out a standard survey is typically associated with limited discretion over the service capacity, whereas up-selling or engineering support is typically associated with a high degree of discretion (Hopp et al. 2007). Since up-selling activities or complex support is most likely going to be administered by skilled labor, we expect that, in this case, capacity discretion is correlated with labor skill level (Chase and Tansik 1983). In a health care setting, administrating a vaccine in a primary care clinic is often associated with limited discretion, whereas diagnosing a patient presenting to the emergency department with symptoms of abdominal pain has a much higher degree of discretion given that there is a variety of conditions associated with abdominal
pain and no standardized protocol is readily applicable. In this case, discretion over capacity may be correlated with the degree of standardization in the service process, which may be a function of the complexity of the “product” (Shostack 1987). Overall, we expect knowledge-intensive processes to be associated with greater level of discretion over service capacity (Hopp et al. 2007, 2009).

3.3. Capacity Choice Problem
We next present the servers’ capacity choice problems under the dedicated and pooled queue configurations, respectively. As a benchmark, we also consider a centralized system with pooled servers, in which a fictional centralized decision maker chooses the servers’ capacities to minimize the sum of their respective costs.

3.3.1. Dedicated Servers
For all $i$,

$$
\mu_i = \arg\min_{\mu_i \geq 0} C_i(\mu_i) := \begin{cases} 
  c(\mu_i) + \sum_{s \in S_{\text{dedicated}}} h_i(s) P_{\text{dedicated}}(s; \mu_i), & \text{if } \mu_i > \lambda \\
  c(\mu_i) + \sup_{s \in S_{\text{dedicated}}} h(s), & \text{otherwise.}
\end{cases}
$$

(1)

By Weierstrass’ Theorem, for the cost functions we consider here, there exists a capacity $\mu_i \geq 0$ minimizing $C_i(\mu_i)$ because $C_i(\mu_i)$ is continuous and coercive and the feasible set is closed. Denote by $\mu_{i,\text{dedicated}}^*$ the optimal capacity investments.

3.3.2. Pooled Servers
For all $i$,

$$
\mu_i = \arg\min_{\mu_i \geq 0} C_i(\mu_i; \mu_{-i}) := \begin{cases} 
  c(\mu_i) + \sum_{s \in S_{\text{pooled}}} h_i(s) P_{\text{pooled}}(s; \mu_i, \mu_{-i}), & \text{if } \mu_i + \mu_{-i} > 2\lambda \\
  c(\mu_i) + \sup_{s \in S_{\text{pooled}}} h_i(s), & \text{otherwise.}
\end{cases}
$$

(2)

For any $\mu_{-i}$, there exists a capacity $\mu_i \geq 0$ minimizing $C_i(\mu_i; \mu_{-i})$ because $C_i(\mu_i; \mu_{-i})$ is continuous and coercive and the feasible set is closed, by Weierstrass’ Theorem. We show in §4.1 and §5.1 that there exists an equilibrium under both workload aversion and busyness aversion. In case of multiple equilibria, we select the equilibrium that is Pareto-dominant.\(^5\) Denote by $(\mu_{i,\text{pooled}}^*, \mu_{i,\text{pooled}}^*)$ the (selected) equilibrium capacity investments.

3.3.3. Benchmark: Pooled Servers under Centralized Control
We consider the “first-best” capacity investment as a benchmark. Specifically, we consider a centralized system with pooled servers in which a fictional centralized decision maker chooses the servers’ capacities to minimize the sum of their respective costs. This system will obviously lead to smaller costs than the actual (decentralized) pooled configuration because it alleviates any negative effect that may be associated with free riding.

$$
(\mu_1, \mu_2) = \arg\min_{\mu_1 \geq 0, \mu_2 \geq 0} C(\mu_1, \mu_2) := \begin{cases} 
  c(\mu_1) + c(\mu_2) + \sum_{i=1}^2 \sum_{s \in S_{\text{pooled}}} h_i(s) P_{\text{pooled}}(s; \mu_i, \mu_{-i}), & \text{if } \mu_1 + \mu_2 > 2\lambda \\
  c(\mu_1) + c(\mu_2) + \sup_{i=1}^2 \sup_{s \in S_{\text{pooled}}} h_i(s), & \text{otherwise.}
\end{cases}
$$

(3)

\(^5\) Given that we focus on symmetric equilibria, there always exists a Pareto-dominant equilibrium; in the case of a tie, we always pick the equilibrium that leads to the largest capacity investment.
By Weierstrass’ Theorem, there exist capacities \((\mu_1, \mu_2) \geq (0, 0)\) minimizing \(C(\mu_1, \mu_2)\) because \(C(\mu_1, \mu_2)\) is continuous and coercive and the feasible set is closed. Denote by \((\mu_i^\text{centralized}, \mu_i^\text{centralized})\) the optimal capacity investments. Although this system is fictional, it will constitute a natural performance benchmark, as we discuss next.

3.4. Performance Metrics

In comparing between the various queue configurations we focus on two main performance measures: expected WIP and queue length. Expected WIP, denoted by \(L\), represents the total expected number of customers (or jobs) in the system. Expected queue length, denoted by \(Q\), refers to the expected number of customers (or jobs) in queue, not including those that are in service. For a dedicated queue, we have when \(\mu_i > \lambda\)

\[
L^{\text{dedicated}}(\mu_i) = \frac{\lambda}{\mu_i - \lambda},
\]

\[
Q^{\text{dedicated}}(\mu_i) = L(\mu_i) - \frac{\lambda}{\mu_i} = \frac{\lambda^2}{\mu_i(\mu_i - \lambda)},
\]

where \(L^{\text{dedicated}}(\mu_i)\) (\(Q^{\text{dedicated}}(\mu_i)\)) is the expected WIP (queue length) of Server \(i\) when she operates at capacity \(\mu_i\) in a dedicated queue.

For a pooled queue, we have that when \(\mu_1 + \mu_2 > 2\lambda\)

\[
L^{\text{pooled}}(\mu_1, \mu_2) = \frac{\lambda(\mu_1 + \mu_2)^3}{(\mu_1^2(\lambda + \mu_2) + \mu_2^2(\lambda + \mu_1))(\mu_1 + \mu_2 - 2\lambda)},
\]

and when \(\mu_1 = \mu_2 = \mu > \lambda\),

\[
L^{\text{pooled}}(\mu, \mu) = \frac{2\mu\lambda}{\mu^2 - \lambda^2},
\]

\[
Q^{\text{pooled}}(\mu, \mu) = \frac{2\lambda^3}{\mu(\mu^2 - \lambda^2)},
\]

where \(L^{\text{pooled}}(\mu_1, \mu_2)\) (\(Q^{\text{pooled}}(\mu_1, \mu_2)\)) is the expected WIP (queue length) in a pooled system when servers operate at capacities \(\mu_1\) and \(\mu_2\). See Appendix EC.2 for a derivation of these metrics.

When capacities are endogenous and equal to \(\mu\) across both pooled and dedicated configurations, we have

\[
\frac{2L^{\text{dedicated}}(\mu)}{L^{\text{pooled}}(\mu, \mu)} = 1 + \frac{\lambda}{\mu}.
\]

That is, for the same capacities, the benefit of pooling, as measured by the relative reduction in expected WIP, increases with utilization.

Although the pooled system under centralized control (introduced in §3.3.3) naturally leads to the lowest total costs, it is not clear a priori whether it may lead to more or less WIP. However, we will show in the sequel that, under both workload and busyness aversion, the total WIP is lower in the pooled system under centralized control than in either the dedicated system or the pooled system under decentralized control. Accordingly, the pooled system with centralized control will constitute a performance benchmark in terms of achievable WIP reduction.
4. Workload Aversion

In this section we characterize the equilibrium capacity investments and system performance when servers are workload-averse. Specifically, throughout this section, we consider the case where, for the dedicated system, \( h(j) = w \cdot j \), and for the pooled system, \( h_i(j) = w \cdot j/2 \), and \( h_i(1_i) = h_{-i}(1_i) = w/2 \). We first establish the existence and uniqueness of symmetric equilibrium service rates. We then discuss key properties of these equilibrium service rates across the different queue configurations. We finally compare system performance and identify the regime where the dedicated queue configuration outperforms the pooled configuration.

4.1. Existence and Uniqueness of Equilibria

We consider the existence and uniqueness of symmetric service rates under the dedicated and the pooled queue configurations, as well as under the centralized pooled benchmark.

4.1.1. Dedicated Servers

Plugging the functional form for the expected WIP, derived in (4), into (1), we obtain

\[
C_i(\mu_i) = \begin{cases} 
\mu_i^k + w \frac{\lambda}{\mu_i - \lambda} & \text{if } \mu_i > \lambda \\
\infty & \text{otherwise}
\end{cases}
\]

Because \( C_i(\mu) = \infty \) when \( \mu_i \leq \lambda \), the optimal capacity is greater than \( \lambda \) and therefore solves the following first-order optimality condition:

\[
C_i'(\mu_i) = k \mu_i^{k-1} - w \frac{\lambda}{(\mu_i - \lambda)^2} = 0.
\]

Because \( C_i''(\mu_i) > 0 \) when \( k \geq 1 \), there exists at most one \( \mu_i > \lambda \) such that \( C_i'(\mu_i) = 0 \), i.e., the optimal capacity is unique, and we denote it as \( \mu_i^{\text{dedicated}} = \mu_i^{\text{pooled}} = \mu_i^{\text{d}} \). In particular, when \( k = 1 \), \( \mu_i^{\text{dedicated}} = \lambda + \sqrt{\lambda} \).

4.1.2. Pooled Servers

Plugging the functional form for the expected WIP, derived in (6) into (2), we obtain

\[
C_i(\mu_i; \mu_{-i}) = \begin{cases} 
\mu_i^k + \frac{w}{2} \frac{\lambda(\mu_1 + \mu_2)^3}{(\mu_1 + \mu_2 + \lambda + \mu_1)(\mu_1 + \mu_2 - 2\lambda)^2} & \text{if } \mu_1 + \mu_2 > 2\lambda, \\
\infty & \text{otherwise}
\end{cases}
\]

Proposition 1. Under workload aversion, there exists a unique pure-strategy Nash equilibrium in the capacity choice game (2). The equilibrium is symmetric, i.e., \( \mu_1 = \mu_2 = \mu^{\text{pooled}} \), and the equilibrium capacity \( \mu^{\text{pooled}} \) is the unique solution \( \mu \) to

\[
C_i'(\mu; \mu) = k \mu_k^{k-1} - \frac{w}{2} \frac{\lambda(\mu_1 + \mu_2)}{(\mu_1 - \mu_2)^2} = 0,
\]

and \( \mu^{\text{pooled}} > \lambda \).

In particular when \( k = 1 \), solving (12) yields \( \mu^{\text{pooled}} = \sqrt{\lambda^2 + \frac{w\lambda}{4} + \frac{\lambda}{2} \sqrt{16w\lambda + w^2}} \).
4.1.3. Centralized Pooled Servers

Plugging the functional form for the expected WIP, derived in (6), into (3), we obtain:

\[ C(\mu_1, \mu_2) = \begin{cases} 
\mu_1^k + \mu_2^k + w \frac{\lambda(\mu_1 + \mu_2)}{(\mu_1^2(\mu_1 + \mu_2) + \mu_2^2(\mu_1 + \mu_2) + 2\mu_2)(\mu_1^2 + \mu_2^2 - 2\lambda)} & \text{if } \mu_1 + \mu_2 > 2\lambda, \\
\infty & \text{otherwise}.
\end{cases} \]

Because the cost function is strictly supermodular (see the proof of Proposition 1 for details) and symmetric, every optimal solution is symmetric (Topkis 1998, Theorem 2.7.5), i.e., \( \mu_1 = \mu_2 = \mu \) in every optimal solution. Moreover, \( \mu > \lambda \) for otherwise \( C(\mu, \mu) = \infty \). Accordingly, every optimal capacity \( \mu \) solves the first-order condition:

\[ \frac{\partial C(\mu, \mu)}{\partial \mu_i} = k\mu^{k-1} - w\lambda \frac{\lambda^2 + \mu^2}{(\mu^2 - \lambda^2)^2} = 0. \]  

(13)

Because \( \frac{d}{d\mu} \left( \frac{\partial C(\mu, \mu)}{\partial \mu_i} \right) > 0 \), there exists at most one \( \mu \in (\lambda, \infty) \) such that \( \frac{\partial C(\mu, \mu)}{\partial \mu_i} = 0 \), i.e., the optimal capacity is unique, and we denote it as \( \mu_{\text{centralized}} \). In particular, when \( k = 1 \), solving (13) yields \( \mu_{\text{centralized}} = \sqrt{\lambda^2 + \frac{w\lambda}{2} + \frac{\lambda}{2} \sqrt{8w\lambda + w^2}} \).

4.2. Comparison of Equilibrium Service Rates

We next compare the equilibrium service rates and will show that equilibrium capacity investments will be lower in a pooled queue configuration due to the following two factors:

- A pooled system is operationally more efficient, and this increased efficiency reduces the marginal returns to capacity investment (Yu et al. 2015);
- A pooled system couples the servers’ choices of service rates; that is, the total WIP depends on the service rates of all servers, thereby creating free riding (Holmström 1982).

Comparing (12) with (10) and (13) shows the two effects:

\[ C_i'(\mu; \mu) = k\mu^{k-1} - w \times \frac{\lambda}{(\mu - \lambda)^2} \times \frac{(\lambda^2 + \mu^2)}{(\mu + \lambda)^2} \times \frac{1}{2} \left( \frac{1}{2} \right). \]  

(14)

In particular, Equation (14) shows two sources of inefficiency in the derivative of the cost function in a decentralized pooled system, namely a term equal to \( (\lambda^2 + \mu^2)/(\mu + \lambda)^2 \leq 1 \), which appears in both the derivative of the cost function of the centralized and decentralized pooled systems, and therefore captures the greater efficiency of the pooled system, and a term equal to \( 1/2 \leq 1 \), which only appears in the derivative of the cost function of the decentralized pooled system and is due to free riding.

Comparing (10), (12), and (13) leads to the following ordering characterization.
Proposition 2. Under workload aversion,

\[ \mu^{\text{dedicated}} \geq \mu^{\text{centralized}} \geq \mu^{\text{pooled}}. \]

Hence, under workload aversion, because servers invest less in capacity in a pooled system than in a dedicated system, the expected WIP in a pooled system may end up being higher than the expected WIP in a dedicated system. Before investigating when that may happen, we first characterize how the distortion in capacity investments evolves as a function of the workload-aversion intensity. The next proposition shows that this distortion is the highest at some intermediate levels of workload aversion. Figure 1 illustrates the proposition.

Proposition 3. Under workload aversion, \( \mu^{\text{dedicated}}/\mu^{\text{pooled}} \) is quasiconcave in \( w \).

Because in (14), the distortion term due to the greater operational efficiency, \((\lambda^2 + \mu^2)/(\mu + \lambda)^2 \leq 1\), is increasing in \( \mu \) and because the distortion term due to free riding, \(1/2 \leq 1\), is constant, the total distortion in incentives to reduce workload between a dedicated system and a pooled system, measured by \((1/2) \times (\lambda^2 + \mu^2)/(\mu + \lambda)^2\), diminishes as \( \mu \) increases. Hence, for very large values of service rates \( \mu \) (i.e., for equilibrium service rates corresponding to very large workload-aversion intensities \( w \)), the incentives in a pooled system are very close to those in a dedicated system, and we expect \( \mu^{\text{dedicated}} \) and \( \mu^{\text{pooled}} \) to be similar.

On the other hand, when \( \mu \) tends to \( \lambda \), the workload-aversion costs tend to infinity in both the pooled and the dedicated systems. Accordingly, when the intensity of workload aversion \( w \) is very small, the incentives are also closely aligned between a pooled and a dedicated systems, in the sense that servers in either system want to invest as little as possible while preventing their workload-aversion costs from exploding.

It is thus only for intermediate values of workload aversion \( w \) that incentives due to both the greater operational efficiency of the pooled system and the presence of free riding are the most distorted; in that regime, servers in a pooled system have some leeway to exert significantly less effort than do servers in a dedicated system while containing their workload-aversion costs.

4.3. Performance Comparison

We next compare the equilibrium system performance under the two queue configurations. Our main result is that under workload aversion, a dedicated queue configuration leads to lower WIP than the pooled queue configuration when servers have a high degree of discretion over their service capacities and exhibit high levels of workload aversion:
Figure 1  Ratio of service capacities $\mu_{\text{dedicated}} / \mu_{\text{pooled}}$ as a function of $w$ under workload aversion ($\lambda = 0.5$, $k = 3$).

**Theorem 1.** Under workload aversion, there exists a threshold $W(k)$ such that $L_{\text{pooled}} \geq 2L_{\text{dedicated}}$ if and only if $w \geq \lambda^k W(k)$, with $W(k)$ continuous, $W'(k) \geq 0$, and $W(k) = 0$ for all $k \leq 2$.

In our discussion of this result, we separately evaluate the effects of workload aversion ($w$) and discretion over service capacity ($k$). The following two elements are helpful in understanding the effect of workload aversion:

- **Efficiency:** By (9), the benefit of pooling, for identical capacities $\mu$, increases in the utilization, $\lambda / \mu$. Utilization obtains its highest values when $\mu$ is small, which corresponds to low values of workload aversion $w$. Hence, the potential benefits of pooling are the highest at low levels of workload aversion.

- **Distorted incentives:** By Proposition 2, the capacity in the pooled configuration is lower than in the dedicated configuration because of greater operational efficiency and free riding. Because the distortion in service capacities is the largest at intermediate values of $w$ by Proposition 3, the downside of a pooled configuration tends to be the highest at intermediate values of $w$.

When $w$ is small, both effects are aligned, and indeed, we have that the pooled configuration results in lower expected WIP. As $w$ increases, the relative efficiency of the pooled system diminishes and the distortion in capacity investments increases by Proposition 3, therefore attenuating the benefit of a pooled system; beyond a certain point, the distortion in service capacities is so large that a dedicated system yields less expected WIP than a pooled system. Although the distortion in equilibrium service capacities starts to diminish beyond a certain point as $w$ keeps increasing, by Proposition 3, the operational benefits of a pooled system are so small in that regime that even a little distortion in incentives makes the dedicated system result in less expected WIP.

To interpret the effect of the capacity cost elasticity $k$ on the WIP comparison of Theorem 1, we re-express the condition $w \geq W(k)\lambda^k$ in terms of the original capacity costs, when $c \neq 1$ and
Because the capacities that minimize $C_i(\mu_i)$ and $C_i(\mu_i; \mu_{-i})$ also minimize the same cost functions multiplied by the constant $(\mu_i^k/c)$, the “un-normalized” workload-aversion intensity can be expressed as $\tilde{w} = w/(\mu_0^k/c)$. Hence, the condition can be expressed as $\tilde{w} (\frac{\mu_0}{\lambda})^k \geq W(k)$. The function $W(k)$ depends on no other problem parameters than $k$; although it cannot be expressed in closed form, it can easily be numerically evaluated for specific values of $k$. Figure 2 plots $W(k)$ on a logarithmic scale; consistent with Theorem 1, $W'(k) \geq 0$ and $W(k) = 0$ (i.e., $\ln(W(k)) = -\infty$) for all $k \leq 2$. On that logarithmic scale, the left-hand side of the condition, $\ln(\tilde{w}) + k \ln(\frac{\mu_0}{\lambda})$, is a linear function of $k$. Naturally, when $k \to \infty$, service capacities are exogenously given by $\mu_0$, and we recover the classical result that pooled configurations are more efficient (Kleinrock 1975).

Because $W(k)$ and $\tilde{w} (\frac{\mu_0}{\lambda})^k$ are continuous and because $\lim_{k \to \infty} W(k) > \lim_{k \to \infty} \tilde{w} (\frac{\mu_0}{\lambda})^k$ and $W(k) < \tilde{w} (\frac{\mu_0}{\lambda})^k$ for all $k \leq 2$, there is at least one crossing of the two functions when $k \geq 1$. Hence, for small values of $k$, i.e., when servers have a lot of discretion over their service capacity, as often happens in knowledge-intensive processes, a dedicated system leads to less expected WIP, i.e., $L_{\text{pooled}} \geq 2L_{\text{dedicated}}$; whereas for large values of $k$, i.e., when servers have limited discretion over their service capacity, as often happens in standardized processes, a pooled system leads to less expected WIP, i.e., $L_{\text{pooled}} \leq 2L_{\text{dedicated}}$. In Figure 2, $\ln(W(k))$ appears to behave almost linearly for all $k \gg 2$, so there may well be in general a unique crossing point $\hat{k}$ such that $L_{\text{pooled}} \leq 2L_{\text{dedicated}}$ if and only if $k \leq \hat{k}$. Although it is intricate to establish uniqueness of that crossing point, given the lack of closed form for $W(k)$, the key take-away from this analysis is that a pooled system tends to lead to more WIP when servers have significant discretion over their service capacity.

**Figure 2** Function $W(k)$ and one example of a function $\tilde{w} (\frac{\mu_0}{\lambda})^k$ when $\tilde{w} = 1$ and $\frac{\mu_0}{\lambda} = \frac{1}{\pi^r}$.
As shown in Proposition 2, lower capacity investments in a pooled system are the result of (i) an operational effect (since a pooled queue is more efficient, it is associated with lower returns on capacity), and (ii) a gaming effect, i.e., free riding. The next proposition shows that the first effect is not sufficient to lead to lower expected WIP under workload aversion: Without free riding, a pooled queue always leads to lower expected WIP.

**Proposition 4.** Under workload aversion, $L_{centralized} \leq 2L_{dedicated}$.

The following corollary provides justification for using the pooled queue configuration with centralized control as a benchmark. That is, the centralized pooled solution, in addition to (naturally) leading to the lowest total cost, also leads to the lowest expected WIP.

**Corollary 1.** Under workload aversion, $L_{centralized} \leq \min\{L_{pooled}, 2L_{dedicated}\}$.

Although a dedicated queue configuration may sometimes lead to less expected WIP than a pooled configuration, the expected queue length is always shorter under a pooled configuration.

**Proposition 5.** Under workload aversion, $Q_{pooled} \leq 2Q_{dedicated}$.

Hence, even though the capacity in the pooled system is lower than that in the dedicated system (Proposition 2), it is not sufficiently low to make queues longer on average. This suggests that the loss of efficiency introduced by free riding in pooled queue configurations remains moderate under workload aversion.

Figure 3 compares the expected WIP and queue length under various combinations of parameters $w$ and $k$. Consistent with our discussion of Theorem 1, the dedicated system tends to create less WIP when servers have high discretion over their service capacity (i.e., low value of $k$) and the pooled system tends to create less WIP when servers have low discretion over their service capacity (i.e., high value of $k$). For intermediate values of $k$, one should consider the degree of workload aversion: Dedicated systems tend to create less WIP when servers have a high degree of workload aversion. Consistent with Corollary 1, the centralized pooled system always leads to the least amount of expected WIP and constitutes therefore a benchmark. Finally, consistent with Proposition 5, the pooled system always leads to a shorter expected queue length.

5. **Busyness Aversion**

In this section we characterize the equilibrium investments and system performance under busyness aversion. Accordingly, throughout this section, we assume that each server incurs a constant cost rate $b > 0$ whenever they serve a customer and a cost rate of 0 otherwise. Specifically, in this case, $h(0) = h_i(0) = h_{-i}(1) = 0$, $h_i(1) = h(j) = h_{-i}(j) = b$, for all $j \geq 2$. As in the previous section, we first establish the existence of symmetric equilibrium service rates. We then discuss key properties
of these equilibrium service rates across the different queue configurations. Finally, we compare system performance and identify the regime where the dedicated queue configuration outperforms the pooled configuration.

5.1. Existence of Equilibria

We consider the existence of symmetric capacity equilibrium under the dedicated and the pooled queue configurations, as well as under the centralized pooled benchmark. While we are able to establish existence in all three cases, uniqueness may not be guaranteed.

5.1.1. Dedicated Servers

Under a dedicated queue configuration each server works independently, and the probability that Server \(i\) is busy is therefore equal to \(1 - P^{\text{dedicated}}(0) = \lambda / \mu_i\) when \(\lambda < \mu_i\). Thus, (1) reduces to

\[
C_i(\mu_i) = \begin{cases} 
\mu_i^k + b \frac{\lambda}{\mu_i} & \text{if } \mu_i > \lambda \\
\mu_i^k + b & \text{otherwise.}
\end{cases}
\]
Unlike the cost under workload aversion, the cost under busyness aversion is finite for all values of $\mu_i$, even when $\mu_i \leq \lambda$. Indeed, in this case, it is possible to have $\mu^i_{\text{dedicated}} = 0$.

Because $\mu^k + b$ is increasing, the optimal capacity is either equal to zero or is greater than $\lambda$; in the latter case, it must solve the first-order optimality condition:

$$C'_i(\mu_i) = k\mu_i^{k-1} - b\frac{\lambda}{\mu_i^2} = 0,$$

Let us denote by $\mu^*$ the solution to (15), which is unique because $\mu^k + b\frac{\lambda}{\mu_i}$ is strictly convex. Then, $\mu^i_{\text{dedicated}} = 0$ if either $C_i(0) \leq C_i(\mu^*)$ or if $\mu^* \leq \lambda$, and $\mu^i_{\text{dedicated}} = \mu^*$ otherwise. It can be checked that $C_i(0) > C(\mu^*) \Rightarrow \mu^* > \lambda$; hence $\mu^i_{\text{dedicated}} = 0$ if $C_i(0) \leq C_i(\mu^*)$, which is equivalent to $b \leq \lambda^k k^{-k}(1+k)^{1+k}$, and $\mu^i_{\text{dedicated}} = \mu^*$ otherwise. We summarize this statement with the following proposition (proof is omitted):

**Proposition 6.** Under busyness aversion, there exists an optimal service rate solving (1). The optimal service rate $\mu^i_{\text{dedicated}}$ is equal to zero if $b \leq \lambda^k k^{-k}(1+k)^{1+k}$ and to $\mu^* > \lambda$ when $b \geq \lambda^k k^{-k}(1+k)^{1+k}$, in which $\mu^*$ is the unique value of $\mu$ such that $C'_i(\mu^*) = 0$.

Hence, the optimal service rate $\mu^i_{\text{dedicated}}$ is uniquely defined when $b \neq \lambda^k k^{-k}(1+k)^{1+k}$ (when $b = \lambda^k k^{-k}(1+k)^{1+k}$, we set by convention $\mu^i_{\text{dedicated}} = \mu^*$). In particular, when $k = 1$, $\mu^i_{\text{dedicated}} = \sqrt{b\lambda}$ if $b \geq 4\lambda$ and $\mu^i_{\text{dedicated}} = 0$ if $b < 4\lambda$.

### 5.1.2. Pooled Servers

Using (2) and the steady-state probabilities $P^{\text{pooled}}(s; \mu_i, \mu_{-i})$ derived in Appendix EC.2, we obtain:

$$C_i(\mu_i; \mu_{-i}) = \begin{cases} \mu^k_i + b(1 - P^{\text{pooled}}(0; \mu_1, \mu_2) - P^{\text{pooled}}(1; \mu_1, \mu_2)) = \mu^k_i + b\lambda \frac{\lambda \mu_1 \mu_2}{\mu_1^2 \mu_2^2}, & \text{if } \mu_1 + \mu_2 > 2\lambda \\ \mu^k_i + b, & \text{otherwise.} \end{cases}$$

In general, $C_i(\mu_i; \mu_{-i})$ is not convex. For instance, if $\mu_{-i} < 2\lambda$, $C_i(\mu_i; \mu_{-i})$ may be increasing for $\mu_i \in [0, 2\lambda - \mu_{-i}]$ and be convex for $\mu_i \geq 2\lambda - \mu_{-i}$, potentially reaching a local maximum at $2\lambda - \mu_{-i}$. This is because of the difference in regimes since busyness-aversion costs are only affected by $\mu_i$ when $\mu_i \geq 2\lambda - \mu_{-i}$. Even if $\mu_{-i} \geq 2\lambda$, $C_i(\mu_i; \mu_{-i})$ may have up to two local minima (see Lemma C-14 for details). Furthermore, $C_i(\mu_i; \mu_{-i})$ is in general neither sub- nor supermodular. In fact, the best-response correspondence (or any selection thereof) may be increasing or decreasing, and it is in general discontinuous. Despite the lack of structure of the cost function, we next show that there always exists a symmetric equilibrium in the capacity choice game, using Tarski’s Intersection Point Theorem (Milgrom and Roberts 1994, Amir and De Castro 2013, Vives 1999). The proof relies on showing that the best response correspondence is quasi-increasing, which essentially means that any selection has only upward jumps.\(^6\)

\(^6\)Gopalakrishnan et al. (2016) establish the existence of a Nash equilibrium when $\mu_i$ is constrained to be larger than $\lambda$, for $i = 1, 2$. In that regime, the cost functions are quasi-convex and Kakutani’s Fixed Point Theorem applies.
PROPOSITION 7. Under busyness aversion, there exists a symmetric Nash equilibrium in the capacity choice game (2) such that either $\mu^\text{pooled}_i = \mu^\text{pooled} = 0$ or $\mu^\text{pooled}_i = \mu^\text{pooled} = \mu^*$, where $\mu^*$ is the unique solution to

$$C'_i(\mu; \mu) = k \mu^{k-1} - b \frac{\lambda}{\mu(\lambda + \mu)} = 0. \quad (17)$$

Moreover, $\mu^* > \lambda$ when $b \geq \lambda^k k^{-k} (1 + k)^{1+k}$.

In case there are multiple symmetric equilibria, we impose the Pareto-dominant equilibrium selection rule. That is, we choose the equilibrium $(\mu, \mu)$ that minimizes the total cost $C(\mu, \mu)$. In particular, when $k = 1$, it can easily be shown $(0, 0)$ is an equilibrium if $b \leq 8 \lambda$, $(\mu^*, \mu^*)$ is an equilibrium with $\mu^* = \left(-\lambda + \sqrt{\lambda^2 + 4b\lambda}\right)/2$ if $b \geq 2\lambda$. Hence, when $2\lambda \leq b \leq 8 \lambda$, there are multiple equilibria. According to the Pareto-dominance selection rule, $(0, 0)$ is the selected equilibrium if and only if $b \leq 2 + \sqrt{5}$.

5.1.3. Centralized Pooled Servers Using (3) and the steady-state probabilities derived in Appendix EC.2, we obtain:

$$C(\mu_1, \mu_2) = \begin{cases} 
  c(\mu_1) + c(\mu_2) + b(1 - 2P^\text{pooled}(0; \mu_1, \mu_2) - P^\text{pooled}(1; \mu_1, \mu_2) - P^\text{pooled}(1 - i; \mu_1, \mu_2)) \\
  = \mu_1^k + \mu_2^k + b\lambda \frac{(\mu_1 + \mu_2)(2\lambda + \mu_1 + \mu_2)}{(\mu_1^2 + \mu_2^2 + \mu_1 \mu_2 + \mu_1 \mu_2)} \\
  c(\mu_1) + c(\mu_2) + 2b
\end{cases}$$

if $\mu_1 + \mu_2 > 2\lambda$, otherwise.

Because the total cost is increasing when $\mu_1 + \mu_2 \leq 2\lambda$, the optimal capacity investments are either equal to zero or they are such that $\mu_1 + \mu_2 > 2\lambda$ and $\partial C(\mu_1, \mu_2)/\partial \mu_1 = \partial C(\mu_1, \mu_2)/\partial \mu_2 = 0$.

In the latter case, because $\partial C(\mu_1, \mu_2)/\partial \mu_1 - \partial C(\mu_1, \mu_2)/\partial \mu_2 > 0$ if and only if $\mu_1 > \mu_2$, we must have that $\mu_1 = \mu_2 = \mu$, where $\mu$ solves

$$\frac{\partial C(\mu_1, \mu_2)}{\partial \mu_i} = 0, \forall i \Leftrightarrow k \mu^{k-1} - b\lambda \frac{1}{\mu^2} = 0. \quad (18)$$

Since $d \left( \frac{\partial C(\mu, \mu)}{\partial \mu_i} \right)/d\mu < 0$, there exists a unique $\mu^*$ such that $\frac{\partial C(\mu, \mu)}{\partial \mu_i} = 0$. Hence, the optimal capacity investments are symmetric and are equal to either zero or $\mu^*$ provided that $\mu^* > \lambda$ and $C(\mu^*, \mu^*) \leq C(0, 0)$. It can easily be checked that $C(\mu^*, \mu^*) \leq C(0, 0) \Rightarrow \mu^* > \lambda$. Consequently, it is optimal to set $\mu^\text{centralized} = \mu^*$ if $C(\mu^*, \mu^*) \leq C(0, 0)$, i.e., if $b \geq \lambda^k k^{-k} (1 + k)^{1+k}$ and to set $\mu^\text{centralized} = 0$ if $b \leq \lambda^k k^{-k} (1 + k)^{1+k}$. The optimal capacity investment is therefore uniquely defined when $b \neq \lambda^k k^{-k} (1 + k)^{1+k}$ (when $b = \lambda^k k^{-k} (1 + k)^{1+k}$, we set by convention $\mu^\text{centralized} = \mu^*$). As a result, $\mu^\text{dedicated} = \mu^\text{centralized}$ as is formally stated in the next subsection.
5.2. Equilibrium Service Rates

Similar to the case with workload aversion, the pooled configuration results in lower capacity investments. While this was due to two factors under workload aversion, namely the greater operational efficiency of the pooled system and the presence of free riding, it is only due to the latter effect here. Indeed, the pooled configuration does not offer any operational benefit at reducing busyness. To see this, note that, under exogenous service rates $\mu$, the probability that a particular server is busy in a pooled configuration, i.e., $1 - (P(0; \mu, \mu) + P(1-i; \mu, \mu)) = \lambda/\mu$, is identical to the probability that a particular server is busy in a dedicated configuration, i.e., $1 - P(0; \mu)$. As a result, under busyness aversion, the service capacity under a dedicated configuration is identical to the service capacity under a centralized pooled queueing system.

Comparing (17) to (15) and (18) shows the free-riding effect when $\mu > \lambda$:

$$C_i'(\mu; \mu) = k\mu^{k-1} - \left(\frac{b\lambda}{\mu^2}\right) \times \left(\frac{\mu}{\lambda + \mu}\right).$$

(19)

The next proposition characterizes the effect of free riding on capacity investments for both interior and non-interior equilibria:

**Proposition 8.** Under busyness aversion,

$$\mu_{\text{dedicated}} = \mu_{\text{centralized}} \geq \mu_{\text{pooled}}.$$

The next corollary shows that the centralized pooled configuration always leads to lower expected WIP than the dedicated configuration or the decentralized pooled configuration. Hence, similar to the case with workload aversion, the centralized pooled configuration constitutes a natural benchmark.

**Corollary 2.** Under busyness aversion, in equilibrium, we have that

$$L_{\text{centralized}} \leq \min\{L_{\text{pooled}}, 2L_{\text{dedicated}}\}.$$

The next proposition shows that the distortion in capacity investment is the highest at the lowest values of busyness aversion. Figure 4 illustrates the proposition.

**Proposition 9.** Under busyness aversion, $\frac{\mu_{\text{dedicated}}}{\mu_{\text{pooled}}}$ is decreasing in $b$.

Note that the free-riding effect may also appear at non-interior equilibria, e.g., when $\mu_{\text{pooled}} = 0$ and $\mu_{\text{dedicated}} > 0$. Although comparing (19) to (14) may suggest that the distortion due to free-riding under busyness aversion $(1 - \mu/(\lambda + \mu))$ is smaller than that under workload aversion $(1 - 1/2)$, this is true only for interior equilibria (e.g., when $\mu > \lambda$), but not in general.
Note that the relative distortion in incentives to reduce busyness between a dedicated system and a pooled system \((1 - \mu/(\lambda + \mu))\) is decreasing in \(\mu\). Hence, for very large values of \(\mu\), which tend to arise when the intensity of busyness aversion \(b\) is very large, the extent of free riding is minimal, and we expect \(\mu^{\text{dedicated}}\) and \(\mu^{\text{pooled}}\) to be similar.

In contrast to workload aversion, which involved a very steep cost function as \(\mu\) tended to \(\lambda\), forcing servers to choose capacity strictly greater than \(\lambda\), under busyness aversion, the term reflecting the servers’ busyness aversion cost is bounded from above by \(b\). In particular, servers may choose zero capacity without seeing their cost exploding. Hence, under busyness aversion, servers in a pooled system can free ride as much as they want and choose a capacity that is arbitrarily close to zero without seeing their costs exploding; whereas under workload aversion, they are effectively constrained to choose a capacity that is sufficiently large to collectively cover the total demand \(2\lambda\). Under busyness aversion, the distortion in capacities \(\mu^{\text{dedicated}}/\mu^{\text{pooled}}\) will thus be the largest when the servers choose a capacity that is arbitrarily small, which will happen at low intensities of busyness aversion \(b\).

![Figure 4](image)

**Figure 4** Ratio of service capacities \(\mu^{\text{dedicated}}/\mu^{\text{pooled}}\) as a function of \(b\) under busyness aversion.

5.3. Performance Comparison

We next compare the equilibrium system performance under the two queue configurations. Our main result is that under busyness aversion a dedicated queue configuration leads to lower WIP than the pooled queue configuration when servers have a lot of discretion over their service capacities and exhibit low levels of busyness aversion:

**Theorem 2.** Under busyness aversion, there exists a threshold \(B(k)\) such that \(L^{\text{pooled}} \geq 2L^{\text{dedicated}}\) if and only if \(b < \lambda^k B(k)\).
In our discussion of this result, we separately evaluate the effects of busyness aversion \((b)\) and discretion over service capacity \((k)\). In order to evaluate the effect of the busyness-aversion intensity \(b\), consider the following two factors:

- Efficiency: Because the relative advantage of a pooled system at reducing workload, for given service capacities \(\mu > \lambda\), measured as \(2L_{\text{dedicated}}(\mu)/L_{\text{pooled}}(\mu, \mu) = 1 + \lambda/\mu\), increases in the utilization, the potential benefit of a pooled configuration is higher at higher levels of utilization, i.e., when \(b\) is smaller.

- Distorted incentives: By Proposition 8, the equilibrium capacity in the pooled configuration is lower than that in the dedicated configuration because of free riding. Because the distortion in service capacities is decreasing in \(b\) by Proposition 9, the downside of a pooled configuration will be higher at lower values of \(b\).

When \(b\) is small, the distortion in service capacities is so strong that it outweighs the greater efficiency of the pooled configuration. Hence, incentives dominate efficiency at low values of busyness aversion. In contrast, at high values of busyness aversion, the distortion in service capacities is so small that it is the small benefit of greater operational efficiency of the pooled configuration that dominates.

To interpret the effect of the cost elasticity \(k\) on the WIP comparison result in Theorem 2, we re-express the condition \(b < \lambda^k B(k)\) in terms of the original capacity costs, when \(c \neq 1\) and \(\mu_0 \neq 1\). Because the capacities that minimize \(C_i(\mu_i)\) and \(C_i(\mu_i; \mu_{-i})\) also minimize the same cost functions multiplied by the constant \((\mu_0^k/c)\), the “un-normalized” busyness-aversion intensity can be expressed as \(\tilde{b} = b/(\mu_0^k/c)\). Hence, the condition can be expressed as \(\tilde{b} \left( \frac{\mu_0}{\lambda} \right)^k < B(k)\). The function \(B(k)\) depends on no other problem parameters than \(k\); although it cannot be expressed in closed form, it can easily be evaluated, for specific values of \(k\). Naturally, when \(k \to \infty\), service capacities are exogenously given by \(\mu_0\), and we recover the classical result that pooled configurations are more efficient (Kleinrock 1975).

Because \(B(k)\) and \(\tilde{b} \left( \frac{\mu_0}{\lambda} \right)^k\) are continuous and because \(\lim_{k \to \infty} B(k) > \lim_{k \to \infty} \tilde{b} \left( \frac{\mu_0}{\lambda} \right)^k\), if \(B(1) > \tilde{b} \left( \frac{\mu_0}{\lambda} \right)\), there is at least one crossing of the two functions.\(^8\) Hence if \(B(1) > \tilde{b} \left( \frac{\mu_0}{\lambda} \right)\), we obtain that, similar to the case with workload aversion, for small values of \(k\), i.e., when servers have a lot of discretion over their service capacity, as often happens in knowledge-intensive processes, a dedicated system will lead to less WIP, i.e., \(L_{\text{pooled}} \geq 2L_{\text{dedicated}}\), whereas for large values of \(k\), i.e.,

\(^8\) When \(k = 1\), \(B(k) = (1/L + 1)^2\), in which \(L\) is the unique positive root of \(L^3 + 2L^2 - L - 1\), i.e., \(L \approx 0.801938\) and \(B(1) \approx 5.04892\). We consider the following three scenarios when \(k = 1\): If \(b/\lambda < 4\), then \(\mu_{\text{dedicated}} = 0\) (Proposition 6), which implies, by Proposition 8, that \(L_{\text{dedicated}} = L_{\text{pooled}} = \infty\). If \(4 \leq b/\lambda < 2 + \sqrt{5}\), then \(\mu_{\text{dedicated}} > 0\) (Proposition 6), but \(\mu_{\text{pooled}} = 0\); hence, \(L_{\text{dedicated}} < L_{\text{pooled}} = \infty\). Finally, if \(b/\lambda \geq 2 + \sqrt{5}\), both \(\mu_{\text{dedicated}} > 0\) and \(\mu_{\text{pooled}} > 0\); in that case, \(L_{\text{dedicated}} < L_{\text{pooled}} < \infty\).
when servers have limited discretion over their service capacity, as often happens in standardized processes, a pooled system will lead to less WIP, i.e., $L_{\text{pooled}} \leq 2L_{\text{dedicated}}$.

In Figure 5, $B(k)$ behaves almost linearly in $k$ whereas the function $\frac{b}{c} \left( \frac{\mu_0}{\lambda} \right)^k$ increases exponentially in $k$ given that $\mu_0 > \lambda$, so, if $B(1) > \frac{b}{c} \left( \frac{\mu_0}{\lambda} \right)$, there may well be in general a unique crossing point $\hat{k}$ such that $L_{\text{pooled}} \leq 2L_{\text{dedicated}}$ if and only if $k \leq \hat{k}$. Naturally, if $B(1) < \frac{b}{c} \left( \frac{\mu_0}{\lambda} \right)$, it is possible that the curves cross twice, and if $B(1) \ll \frac{b}{c} \left( \frac{\mu_0}{\lambda} \right)$, the curves may never cross. However, these latter cases are less relevant because, under those conditions (e.g., when $b/\lambda > 5$ and $k = 1$), the utilization in a dedicated system ($\lambda/\mu_{\text{dedicated}} = \lambda/\sqrt{b\lambda}$) would be less than 45%, and the associated expected WIP, $2L_{\text{dedicated}} = 2 \times \frac{\lambda/\mu}{1-\lambda/\mu}$, would be less than $2 \times 0.81$, which is at most 45% higher than the expected WIP under a pooled system, $L_{\text{pooled}}$, by (9). Given the small magnitudes of the expected WIPs, it is likely that, for these cases, the choice of queue configuration in practice will be based on other criteria than expected WIP.

Figure 5  Function $B(k)$ and one example of a function $\frac{b}{c} \left( \frac{\mu_0}{\lambda} \right)^k$ when $\frac{b}{c} = 1$ and $\frac{\mu_0}{\lambda} = \frac{1}{0.6}$.

Figure 6 compares the expected WIP and queue length under various combinations of parameters $b$ and $k$. Consistent with our discussion of Theorem 2, the dedicated system tends to create less WIP when servers have high discretion over their service capacity (i.e., low value of $k$) and the pooled system tends to create less expected WIP when servers have low discretion over their service capacity (i.e., high value of $k$). For intermediate values of $k$, one should consider the degree of busyness aversion: Dedicated systems tend to create less expected WIP when servers have a low degree of busyness aversion. Consistent with Corollary 2, the centralized pooled system always
leads to the least amount of expected WIP and therefore constitutes a natural benchmark. Finally, we observe that, unlike the case with workload aversion (Proposition 5), a pooled configuration does not always lead to a shorter average queue length under busyness aversion.

![Figure 6](image)

**Figure 6** Average Work-in-Process $L$ and Queue Length $Q$ under Busyness Aversion as a function of $b$ under a dedicated queueing system and a pooled queueing system.

6. Efficiency Analysis

In the previous two sections, we showed that free riding in pooled systems lowered their operational performance despite their operational benefits. In this section, we assess the significance of this loss of efficiency as it relates to the type of work aversion. Combining the results from our analysis with numerical simulations indicates that the loss of efficiency under busyness aversion is much more severe than that under workload aversion.

Before establishing this result, recall the following two results: The potential for a pooled configuration to reduce the expected WIP relative to a dedicated configuration is the highest at high levels of utilization, by (9). Furthermore, high levels of utilization (low capacity) tend to arise in
equilibrium with low intensity of work aversion \((w \text{ or } b)\), due to the dominance of the capacity cost versus the work aversion cost.

Under workload aversion, because \(L_{\text{pooled}} \leq 2L_{\text{dedicated}}\) when \(w\) is low (per Theorem 1), the pooled system remains the most efficient, despite the amount of free riding, when its potential benefits are the highest; hence, we conclude that not all operational benefits of a pooled system are lost due to incentive distortion under workload aversion. In contrast, under busyness aversion, because \(L_{\text{pooled}} \geq 2L_{\text{dedicated}}\) when \(b\) is low (per Theorem 2), it is the dedicated configuration that is the most efficient when the potential benefits of the pooled system are the highest. Hence, all operational benefits of the pooled system are lost due to free riding. Moreover, note that the pooled configuration has always the shortest expected queue length under workload aversion (per Proposition 5), but not necessarily under busyness aversion (see Figure 6). Taken together, these two results indicate that the loss of efficiency under busyness aversion is more severe than that under workload aversion.

To assess the relative extent of the loss of efficiency under either type of work aversion, we need a common benchmark, that would apply to both types of work aversion. Recall that the centralized pooled configuration leads to the lowest expected WIP under both types of work aversion (Corollaries 1 and 2). To provide a fair comparison between the performance associated with the two types of work aversion we need to find a way to relate between the values of the parameters that are not common to both, namely, \(w\) and \(b\). We adopt the following procedure: We first fix the values of the arrival rate \(\lambda\) and the capacity cost elasticity parameter \(k\). For any value of \(\mu > \lambda\), we identify the values of work aversion intensities \(w\) and \(b\) such that \(\mu_{\text{centralized}} = \mu\), i.e., \(\mu\) is the optimal capacity level in a centralized pooled system under either workload or busyness aversion, respectively. Now, given these values of \(w\) and \(b\), for each type of work aversion (workload or busyness), we compute the equilibrium capacity investments under the dedicated and the (decentralized) pooled configurations, respectively. We finally compute the expected WIP associated with the two queue configurations and the two forms of work aversion and display them as a function of \(\rho = \lambda/\mu = \lambda/\mu_{\text{centralized}}\), using the expected WIP of the centralized pooled system as a common benchmark. Because we express our results in terms of utilization, we arbitrarily set \(\lambda = 1\) and study the effect of various cost elasticities \(k\). See Figure 7.

Figure 7 leads to the following observations:

- Consistent with the discussion at the beginning of this section, the loss of efficiency (relative to the centralized pooled benchmark) appears much larger under busyness aversion than under workload aversion, especially at high levels of utilization.

- The performance gap between dedicated and pooled configuration under workload aversion is quite marginal; although there is a slight preference for a pooled configuration at high levels of utilization, both perform similarly.
In contrast, the performance gap between dedicated and pooled configuration under busyness aversion can be very substantial; adopting a pooled configuration instead of a dedicated configuration can be very costly at high levels of utilization under busyness aversion.

Consequently, service systems designers need to be careful in their choice of queueing service configuration when servers exhibit busyness aversion. Not only is the choice of queue configuration more sensitive under busyness than under workload aversion, the consequences of choosing a suboptimal queue configuration are also more severe. Based on Theorem 2, this is particularly the case when servers have high levels of discretion over their service capacities, as it often happens in knowledge-intensive processes, and when servers exhibit low degree of busyness aversion resulting in a high system utilization.

7. Conclusions

Although traditional queueing theory establishes that pooling queues results in better operational performance, recent case studies in health care and call centers suggest that it may not always be true in practice. In this paper, we develop a game-theoretic model to examine why and when a pooled queueing system may lead to more expected WIP than a dedicated queueing system.

Consistent with the relevant practical applications, we assume that servers are paid a fixed wage, have discretion over their service capacity, and exhibit varying levels of work aversion. We consider two distinct types of work aversion: workload aversion (i.e., aversion to high levels of workload) and busyness aversion (i.e., aversion to busyness or preference for idleness).

For those settings, our paper yields three key findings. First, in a pooled queueing system, servers have lower incentives to invest in service capacity. This is due to the higher levels of operational efficiency and the opportunity to free ride on others’ capacity investments. Second, whether a pooled queueing system leads to more expected WIP than a dedicated queueing system depends on the servers’ degree of discretion over service capacity and their type of work aversion. Specifically, we find that dedicated queueing systems outperform pooled queueing systems when servers have a high degree of discretion over their service capacity and when they exhibit either high levels of workload aversion or low levels of busyness aversion. Hence, while the classical result that pooled
systems lead to greater operational performance than dedicated systems holds true when servers have little discretion over their service capacity, as is the case in standardized routine processes, the result reverses when they have high discretion over their service capacity, as is the case in complex, knowledge-intensive processes. Third, we find that the loss of efficiency due to free riding is much more severe under busyness aversion than under workload aversion, to the extent that it could completely negate the potential benefits of queue pooling when they are the highest.

Although this work was motivated by practical case studies, it would be worthwhile to test the model’s predictions using lab experiments. Specifically, future work could leverage an experimental design to systematically vary levels of discretion over service capacity, workload aversion, and busyness aversion, respectively. This would allow us to gain a deeper understanding of how these three elements interact with one another.

Analytically, our work can be extended by considering servers with asymmetric costs and more general (convex, concave, S-shaped) work-aversion functions, which would generalize the linear (workload) and piecewise constant (busyness) forms considered in this paper. We could also consider a model combining workload and busyness aversion, though we numerically found no new insights in this more general setting beyond the insights derived by considering the two effects separately.

Given the significant growth of knowledge-intensive services in developed economies (Apte et al. 2008), it becomes increasingly relevant to understand how to efficiently manage these types of organizations. Our findings highlight several managerial implications for designers of service systems in which servers have discretion over their choice of service capacity, as is often the case for knowledge-intensive services. First, service system designers ought to understand the type and magnitude of work aversion that servers may exhibit. These considerations should be taken into account in deciding whether to adopt a pooled versus dedicated queueing system. Second, if servers exhibit busyness aversion, service system designers should take explicit measures to limit the extent of free riding by either restricting the servers’ discretion over service capacity (e.g., by standardizing processes) or shifting their preference for idleness to a preference for low workload (e.g., by shifting servers’ focus from themselves to their customers through greater customer empathy). How to best adjust servers’ levels of workload aversion and busyness aversion is another important topic that could be explored in future research.

References


Cao, P., S. He, J. Huang, Y. Liu. 2016. Dedicated or pooled: Designing queues for large service systems. Conference presentation.


This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.
Appendix

EC.1. Measuring Discretion

Hopp et al. (2007) study contexts where servers have discretion over their choice of capacity, in the sense that, if they spend more time with a customer, more value (revenue) is generated. Although their model considers dynamic capacity adjustments, it can be specialized to static capacity choices. Specifically, using their notations, let $f(\tau)$ be the value generated by spending an average of $\tau$ units of time per customer, which is assumed to be increasing and concave; see Figure EC.1. Let $\tau_{\text{min}}$ denote the minimum expected service time and $\tau_{\text{max}}$ be the amount of time that achieves 99% of the maximum value. Hopp et al. (2007) measure the degree of discretion over capacity as the following ratio

$$\frac{A}{B} = \frac{\int_{\tau_{\text{min}}}^{\tau_{\text{max}}} f(\tau) \, d\tau - \frac{1}{2} (\tau_{\text{max}} - \tau_{\text{min}}) f(\tau_{\text{max}})}{\frac{1}{2} (\tau_{\text{max}} - \tau_{\text{min}}) f(\tau_{\text{max}})},$$

which, in Figure EC.1, measures the ratio between the shaded area and the underlying triangle. The larger the shaded area ($A$), the more value servers can generate by spending time with customers.

Consider the following revenue function: $f(\tau) = c - c(1/\tau) = c \times (1 - \tau_{\text{min}}/\tau^k)$, in which $c(\mu)$ is defined as in §3.2, $\tau = 1/\mu$, and $\tau_{\text{min}} = 1/\mu_0$. In that case, $\tau_{\text{max}} = \tau_{\text{min}} \times 100^k$, and therefore the measure of discretion specializes to:

$$\frac{A}{B} = \frac{c \int_{\tau_{\text{min}}}^{100^k \times 100^k} \left(1 - \frac{1}{\tau_{\text{min}}^k}\right) d\tau - \frac{1}{2} \left(\tau_{\text{min}} \times 100^k - \tau_{\text{min}}\right) \times c \times 0.99}{\frac{1}{2} \left(\tau_{\text{min}} \times 100^k - \tau_{\text{min}}\right) \times c \times 0.99} = \frac{200}{99} \times \frac{k + (1 - k)100^k - 100^{1-k}}{(1 - k)\left(100^k - 1\right)} - 1,$$

and is decreasing in $k$. Hence, as $k$ increases, servers have less discretion.
EC.2. Balance Equations and Performance Metrics

EC.2.1. Dedicated Servers

We denote by $L(\mu_i)$ the expected WIP, $Q(\mu_i)$ the expected queue length, and $M(\mu_i)$ the expected squared WIP in a single-queue dedicated system with capacity $\mu_i$.

$$L(\mu_i) = \sum_{j=1}^{\infty} j P(j; \mu_i) = P(0; \mu_i) \sum_{j=1}^{\infty} j \left( \frac{\lambda}{\mu_i} \right)^j = \left( 1 - \frac{\lambda}{\mu_i} \right) \frac{\lambda}{\mu_i} \sum_{j=1}^{\infty} \left( \frac{\lambda}{\mu_i} \right)^j = \frac{\lambda}{\mu_i - \lambda}$$

$$Q(\mu_i) = L(\mu_i) - \frac{\lambda}{\mu_i} = \frac{\lambda^2}{\mu_i (\mu_i - \lambda)}$$

$$M(\mu_i) = \sum_{j=1}^{\infty} j^2 P(j; \mu_i) = \left( 1 - \frac{\lambda}{\mu_i} \right) \sum_{j=1}^{\infty} j^2 \left( \frac{\lambda}{\mu_i} \right)^j = \left( 1 - \frac{\lambda}{\mu_i} \right) \frac{\lambda}{\mu_i} \left( \frac{\lambda}{\mu_i} \right) \left( 1 + \frac{\lambda}{\mu_i} \right) \left( 1 - \frac{\lambda}{\mu_i} \right)^3 = \frac{\lambda (\mu_i + \lambda)}{(\mu_i - \lambda)^2}. \quad (A-1)$$

EC.2.2. Two Pooled Servers

We denote by $L(\mu_1, \mu_{-i})$ the expected WIP, $Q(\mu_1, \mu_{-i})$ the expected queue length, and $M(\mu_1, \mu_{-i})$ the expected squared WIP in a single-queue 2-server pooled system with capacities $\mu_1$ and $\mu_{-i}$. For notational brevity we fix $\mu_1$ and $\mu_{-i}$ and we denote by $P_s = P(s; \mu_1, \mu_{-i})$ for all $s \in S^{\text{pooled}}$.

The balance equations can be expressed as follows:

$$2 \lambda P_0 = \mu_1 P_1 + \mu_2 P_2$$

$$(2 \lambda + \mu_1) P_1 = \lambda P_0 + \mu_2 P_2$$

$$(2 \lambda + \mu_2) P_2 = \lambda P_0 + \mu_1 P_2$$

$$(2 \lambda + \mu_1 + \mu_2) P_2 = 2 \lambda (P_1 + P_2) + (\mu_1 + \mu_2) P_3$$

$$(2 \lambda + \mu_1 + \mu_2) P_j = 2 \lambda P_{j-1} + (\mu_1 + \mu_2) P_{j+1}, \forall j \geq 3,$$

which yields

$$P_1 = \frac{\lambda}{\mu_1} P_0 \quad (A-2)$$

$$P_2 = \frac{\lambda}{\mu_2} P_0 \quad (A-3)$$

$$P_j = \left( \frac{2 \lambda}{\mu_1 + \mu_2} \right)^j \frac{(\mu_1 + \mu_2)^2}{2 \mu_1 \mu_2} P_0, \forall j \geq 2. \quad (A-4)$$

Hence, if $2 \lambda < \mu_1 + \mu_2$,

$$P_0 = \left( 1 + \frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_2} + \sum_{j=2}^{\infty} \left( \frac{2 \lambda}{\mu_1 + \mu_2} \right)^j \frac{(\mu_1 + \mu_2)^2}{2 \mu_1 \mu_2} \right)^{-1}$$

$$= \left( 1 + \sum_{j=1}^{\infty} \left( \frac{2 \lambda}{\mu_1 + \mu_2} \right)^j \frac{(\mu_1 + \mu_2)^2}{2 \mu_1 \mu_2} \right)^{-1}$$

$$= \left( 1 + \frac{\lambda}{\mu_1 + \mu_2} \frac{(\mu_1 + \mu_2)^2}{2 \mu_1 \mu_2} \right)^{-1} \quad (A-5)$$

Therefore,

$$L(\mu_1, \mu_{-i}) = P_1 + P_2 + \sum_{j=2}^{\infty} j P_j$$
\[
L(\mu, \mu) = \frac{2\mu \lambda}{\mu^2 - \lambda^2}.
\]

Similarly,
\[
Q(\mu_1, \mu_2) = \sum_{j=3}^{\infty} (j-2)P_j
\]
\[
= P_0 \left( \sum_{j=3}^{\infty} (j-2) \left( \frac{2\lambda}{\mu_1 + \mu_2} \right)^j \left( \frac{\mu_1 + \mu_2}{2\mu_1 \mu_2} \right)^2 \right)
\]
\[
= P_0 \left( \sum_{j=1}^{\infty} j \left( \frac{2\lambda}{\mu_1 + \mu_2} \right)^j \frac{2\lambda^2}{\mu_1 \mu_2} \right)
\]
\[
= \frac{4\lambda^3(\mu_1 + \mu_2)}{\left( \mu_1^2(\lambda + \mu_2) + \mu_2^2(\lambda + \mu_1) \right)(\mu_1 + \mu_2 - 2\lambda)},
\]
and when \(\mu_1 = \mu_2 = \mu\), this expression simplifies to
\[
Q(\mu, \mu) = \frac{2\lambda^3}{\mu(\mu^2 - \lambda^2)}.
\]

Finally,
\[
M(\mu_1, \mu_2) = P_1 + P_2 + \sum_{j=2}^{\infty} j^2P_j
\]
\[
= P_0 \left( \sum_{j=1}^{\infty} j^2 \left( \frac{2\lambda}{\mu_1 + \mu_2} \right)^j \left( \frac{\mu_1 + \mu_2}{2\mu_1 \mu_2} \right)^2 \right)
\]
\[
= P_0 \left( \sum_{j=1}^{\infty} j \left( \frac{2\lambda}{\mu_1 + \mu_2} \right)^j \frac{2\lambda^2}{\mu_1 \mu_2} \right)
\]
\[
= \frac{\lambda(\mu_1 + \mu_2)^3(\mu_1 + \mu_2 + \lambda)}{\left( \mu_1^2(\lambda + \mu_2) + \mu_2^2(\lambda + \mu_1) \right)(\mu_1 + \mu_2 - 2\lambda)}
\]
and when \(\mu_1 = \mu_2 = \mu\), this expression simplifies to
\[
M(\mu, \mu) = \frac{2\mu \lambda}{(\mu - \lambda)^2}.
\]
EC.3. Proofs and Auxiliary Results

In this appendix, for each proposition or theorem that appears in the paper, we first present its proof and then present the supporting lemmas. To ensure consistency, all references to supporting lemmas are forward references.

EC.3.1. Workload Aversion

**Proof of Proposition 1.** The proof uses Lemma C-1. We first show that \( C_i(\mu_i; \mu_{-i}) \) is strictly supermodular. Considering the following derivative of \( C_i(\mu_i; \mu_{-i}) \) when \( \mu_1 + \mu_2 > 2 \lambda \):

\[
\frac{\partial C_i(\mu_i; \mu_{-i})}{\partial \mu_{-i}} = \frac{w \lambda^2 (\mu_1 + \mu_2)}{(\mu_1 + \mu_2 - 2\lambda)^3 (\lambda + \mu_2 + \mu_2^2 (\lambda + \mu_1))^2} \times \left( \mu_1^2 \mu_{-i} + \mu_2^3 (2\lambda^2 + 5\mu_{-i}^2) + 2\mu_1 \mu_{-i} (5\lambda^2 + 6\lambda \mu_2 + 5\mu_{-i}^2) \\
+ 2\mu_1^3 \mu_{-i} (-4\lambda^2 - 14\lambda_2 \mu_{-i} + 12\lambda_2^2 + 5\mu_{-i}^3) + \mu_2 \mu_{-i} (32\lambda^3 - 28\lambda^2 \mu_{-i} + 12\lambda \mu_{-i} + 5\mu_{-i}^3) \\
+ \mu_2^3 (-8\lambda^3 + 10\lambda^2 \mu_{-i} + \mu_{-i}^3) + 2\lambda^2 \mu_{-i}^5 \right),
\]

we obtain that \( C_i(\mu_i; \mu_{-i}) \) is strictly supermodular if and only if the function in parentheses is positive (Vives 1999). Defining \( \tilde{\mu}_i = \mu_i/\lambda \) and \( \tilde{\mu}_{-i} = \mu_{-i}/\lambda \) and using Lemma C-1 show that this is indeed the case.

Define, for any \( \mu_{-i} \), define the best-response correspondence \( \Phi_i(\mu_{-i}) = \arg \min_{\mu_i \geq 0} C_i(\mu_i; \mu_{-i}) \). The best-response correspondence is nonempty by Weierstrass’ Theorem because \( C_i(\mu_i; \mu_{-i}) \) is continuous, the feasible set \([0, \infty)\) is closed and \( \lim_{\mu_{-i} \rightarrow \infty} C_i(\mu_i; \mu_{-i}) = \infty \). By the strict supermodularity of \( C_i(\mu_i; \mu_{-i}) \), we thus obtain that \( \Phi_i(\mu_{-i}) \) is decreasing (Vives 1999, Theorem 2.3). In particular, for any \( \mu_{-i} \geq 0 \), \( \Phi_i(\mu_{-i}) \leq \Phi_i(0) \), and each server’s strategy set can be restricted to the compact interval \([0, \max \Phi_i(0)]\).

Hence, the game is strictly supermodular, and there exists a pure-strategy symmetric Nash equilibrium (Vives 1999, Theorem 2.5); moreover, all equilibria are symmetric (Vives 1999, Footnote 23). Hence, \( \mu_1 = \mu_2 = \mu \).

There is clearly no symmetric equilibrium such that \( \mu \leq \lambda \), because it would then result in an infinite cost; therefore \( \mu > \lambda \), and the equilibrium capacities solve (12). Because

\[
\frac{dC_i(\mu; \mu)}{d\mu} = \frac{d}{d\mu} \left( k \mu_{-i}^{k-1} - \frac{w \lambda (\lambda^2 + \mu^2)}{2 (\mu^2 - \lambda^2)^2} \right) = k(k-1) \mu_{-i}^{k-2} + \frac{\lambda \mu (3\lambda^2 + \mu^2)}{(\mu - \lambda)^2 (\mu + \lambda)^3} > 0 \quad (C-1)
\]

when \( k \geq 1 \) and \( \mu > \lambda \), \( C_i(\mu; \mu) \) is strictly increasing in \( \mu \) when \( \mu > \lambda \). As a result, there exists at most one \( \mu \) such that \( C_i(\mu; \mu) = 0 \). Because \( \lim_{\mu \rightarrow \lambda} C_i(\mu; \mu) = -\infty \) and because \( \lim_{\mu \rightarrow \infty} C_i(\mu; \mu) = 1 \) if \( k = 1 \) and \( \infty \) if \( k > 1 \), there exists exactly one value of \( \mu \), \( \lambda \), \( \mu < \lambda \), such that \( C_i(\mu; \mu) = 0 \). Hence, there exists only one symmetric equilibrium, and, as shown above, that symmetric equilibrium is the unique equilibrium. (See also Vives 1999, Remark 15, p. 34). \( \square \)

**Lemma C-1.** The function

\[
F(\mu_1, \mu_2) = \mu_1^3 \mu_2 + \mu_2^3 (2 + 5\mu_2^3) + 2\mu_1^3 \mu_2 (5 + 6\mu_2 + 5\mu_2^2) + 2\mu_1^3 \mu_2 (-4 - 14\mu_2 + 12\mu_2^2 + 5\mu_2^3) \\
+ \mu_1^2 \mu_2^2 (32 - 28\mu_2 + 12\mu_2^2 + 5\mu_2^3) + \mu_1 \mu_2^3 (-8 + 10\mu_2 + \mu_2^3) + 2\mu_2^5
\]

is positive over all \( \mu_1 \geq 0, \mu_2 \geq 0, \mu_1 + \mu_2 \geq 2 \).
Proof: Suppose that $\mu_2 \geq 2$; then $\mu_1 \geq \max\{0, 2 - \mu_2\} = 0$. Because, in that case,

$$(-4 - 14\mu_2 + 12\mu_2^2 + 5\mu_2^3) \geq (-4 - 14\mu_2 + 22\mu_2^2) \geq (-4 + 30\mu_2) \geq 56,$$

$$\quad (32 - 28\mu_2 + 12\mu_2^2 + 5\mu_2^3) \geq (32 - 28\mu_2 + 22\mu_2^2) \geq (32 + 16\mu_2) \geq 64,$$

$$\quad (-8 + 10\mu_2 + \mu_2^3) \geq (-8 + 14\mu_2) \geq 20,$$

we obtain that all coefficients of $\mu_1$ in $F(\mu_1, \mu_2)$ are nonnegative. Moreover, the constant term, $2\mu_2^3$, is positive. Hence, $F(\mu_1, \mu_2) > 0$ for all $\mu_1 \geq 0$ when $\mu_2 \geq 2$.

Suppose next that $\mu_2 < 2$; then $\mu_1 \geq \max\{0, 2 - \mu_2\} = 2 - \mu_2$. Observe that $\partial^4 F(\mu_1, \mu_2)/\partial \mu_1^4 \geq 0$ for all $\mu_1 \geq 0$ and $\mu_2 \geq 0$. Because

$$\frac{\partial^3 F(2 - \mu_2, \mu_2)}{\partial \mu_1^3} = 48(10 + 19\mu_2 + \mu_2^2 - 3\mu_2^3) > 48(10 + 19\mu_2 - 5\mu_2^2) > 48(10 + 9\mu_2) > 0,$$

we obtain that $\partial^3 F(\mu_1, \mu_2)/\partial \mu_1^3 > 0$ for all $\mu_1 \geq 2 - \mu_2$ and $2 > \mu_2 \geq 0$. Furthermore, because

$$\frac{\partial^2 F(2 - \mu_2, \mu_2)}{\partial \mu_1^2} = 8(40 + 48\mu_2 - 6\mu_2^2 - 12\mu_2^3 + 3\mu_2^4) > 8(40 + 48\mu_2 - 30\mu_2^2 + 3\mu_2^4)$$

$$\quad > 8(40 - 12\mu_2 + 3\mu_2^4) > 8(16 + 3\mu_2^4) > 0,$$

we obtain that $\partial^2 F(\mu_1, \mu_2)/\partial \mu_1^2 > 0$ for all $\mu_1 \geq 2 - \mu_2$ and $2 > \mu_2 \geq 0$. Furthermore, because

$$\frac{\partial F(2 - \mu_2, \mu_2)}{\partial \mu_1} = 16(10 + 6\mu_2 - 3\mu_2^2) > 160 > 0,$$

we obtain that $\partial F(\mu_1, \mu_2)/\partial \mu_1 > 0$ for all $\mu_1 \geq 2 - \mu_2$ and $2 > \mu_2 \geq 0$. Furthermore, because

$$F(2 - \mu_2, \mu_2) = 64 > 0,$$

we obtain that $F(\mu_1, \mu_2) > 0$ for all $\mu_1 \geq 2 - \mu_2$ and $2 > \mu_2 \geq 0$. □

Proof of Proposition 2. From (10), (12), and (13), we obtain

$$C'_i(\mu) = k\mu^{k-1} - h \frac{\lambda}{(\mu - \lambda)^2} \leq k\mu^{k-1} - h \frac{\lambda}{(\mu - \lambda)^2} \frac{\lambda^2 + \mu^2}{(\lambda + \mu)^2} = \frac{\partial C(\mu, \mu)}{\partial \mu},$$

$$\quad \leq k\mu^{k-1} - h \frac{\lambda}{(\mu - \lambda)^2} \frac{\lambda^2 + \mu^2}{(\lambda + \mu)^2} 1 = C'_i(\mu; \mu).$$

We showed in §4.1 that $C''_i(\mu) > 0$ and that $d \left( \frac{\partial C(\mu, \mu)}{\partial \mu} \right) / d\mu > 0$; moreover, $d (C'_i(\mu; \mu)) / d\mu > 0$ for all $\mu > \lambda$ by (C-1). The result then follows because $\mu^{\text{dedicated}}$, $\mu^{\text{centralized}}$, and $\mu^{\text{pooled}}$ are respectively roots of $C'_i(\mu)$, $\frac{\partial C(\mu, \mu)}{\partial \mu}$, and $C'_i(\mu; \mu)$. □

Proof of Proposition 3. In the proof, we denote $\mu^{\text{dedicated}} = \mu^d$ and $\mu^{\text{pooled}} = \mu^p$. Applying the Implicit Function Theorem to (10) and (12), we obtain that, when $\mu^d > \lambda$ and $\mu^p > \lambda$,

$$\frac{d\mu^d}{dw} = \frac{\lambda}{(\mu^d - \lambda)^2} (k - 1) (\mu^d)^{k-2} + 2w \frac{\lambda}{(\mu^d - \lambda)^3} = \frac{\lambda}{(\mu^d - \lambda)^2} (k - 1) (\mu^d)^{k-2} + 2w \frac{\lambda}{(\mu^d - \lambda)^3} = \frac{\mu^d}{w \left( k + \frac{\mu^d + \lambda}{\mu^d - \lambda} \right)}$$

$$\frac{d\mu^p}{dw} = \frac{\lambda (\lambda^2 + (\mu^p)^2)}{2(\mu^p)^2 - \lambda^2} (k - 1) (\mu^p)^{k-2} + w \lambda \mu^p \frac{(\mu^p)^2 + \lambda^2}{(\mu^p)^2 - \lambda^2}$$

$$\quad = \frac{\mu^p}{w \left( k + \frac{(\mu^p)^4 + \lambda^4 + 6(\mu^p)^2 \lambda^2}{(\mu^p)^4 - \lambda^4} \right)} \geq 0.$$
Hence,
\[
\frac{d}{dw} \left( \frac{\mu^d}{\mu^p} \right) \geq 0 \iff \mu^p \frac{d\mu^d}{dw} - \mu^d \frac{d\mu^p}{dw} \geq 0
\]
\[
\iff \frac{1}{k + \frac{\mu^d + \lambda}{\mu^d - \lambda}} \geq \frac{1}{k + \frac{(\mu^p)^3 + \lambda^4 + 6(\mu^p)^2 \lambda^2}{(\mu^p)^4 - \lambda^4}}
\]
\[
\iff \frac{(\mu^p)^4 + \lambda^4 + 6(\mu^p)^2 \lambda^2}{(\mu^p)^4 - \lambda^4} \geq \frac{\mu^d + \lambda}{\mu^d - \lambda}
\]
\[
\iff \mu^d \geq \frac{(\mu^p)^3 + 3\mu^p \lambda^2}{3(\mu^p)^2 \lambda + \lambda^3},
\]
(C-2)

Because
\[
\frac{d}{dp} \left( \frac{(\mu^p)^3 + 3\mu^p \lambda^2}{3(\mu^p)^2 \lambda + \lambda^3} \right) = \frac{3((\mu^p)^2 - \lambda^2)^2}{\lambda(3(\mu^p)^2 + \lambda^3)^2} > 0
\]
and because \(d\mu^p/dw \geq 0\), we obtain that the function \((\mu^p)^3 + 3\mu^p \lambda^2)/(3(\mu^p)^2 \lambda + \lambda^3)\) increases in \(w\). By (C-2), the function \(\mu^d/\mu^p\) crosses the function \((\mu^p)^3 + 3\mu^p \lambda^2)/(3(\mu^p)^2 \lambda + \lambda^3)\) at most once as \(w\) increases, and the crossing is from above. Hence, by (C-2), we obtain that \(\mu^d/\mu^p\) is quasiconcave in \(w\). \(\square\)

Proof of Theorem 1. The proof uses Lemmas C-2, C-3, and C-7. Define \(\Psi(L) - \frac{\sqrt{1 + 4L^2}}{2(L + 1)} + \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)^k \). Taking the derivative of \(\Psi(L)\) when it crosses zero, we obtain
\[
\Psi'(L)_{\Psi(L)=0} = \left( -\frac{4L - 1}{2(L + 1)^2 \sqrt{1 + 4L^2}} + k \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)^{k-1} \frac{4L - 1 - \sqrt{1 + 4L^2}}{2(L + 1)^2 \sqrt{1 + 4L^2}} \right)_{\Psi(L)=0}
\]
\[
= -\frac{4L - 1}{2(L + 1)^2 \sqrt{1 + 4L^2}} + \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)^{-1} \frac{\sqrt{1 + 4L^2}}{2(L + 1)^2 \sqrt{1 + 4L^2}}
\]
\[
= \frac{1}{2(L + 1)^2 \sqrt{1 + 4L^2}} \left( -4L + 1 + \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \frac{\sqrt{1 + 4L^2}}{1 + \sqrt{1 + 4L^2}} \left( 4L - 1 - \sqrt{1 + 4L^2} \right) \right).
\]

Note that \(\ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \leq 0\). Therefore, when \(4L - 1 - \sqrt{1 + 4L^2} < 0\), i.e., when \(L < 2/3\),
\[
-4L + 1 + \frac{\ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)}{1 + \sqrt{1 + 4L^2}} \left( 4L - 1 - \sqrt{1 + 4L^2} \right)
\]
\[
< -4L + 1 + \frac{\sqrt{1 + 4L^2}}{1 + \sqrt{1 + 4L^2}} \left( 4L - 1 - \sqrt{1 + 4L^2} \right)
\]
\[
= \frac{1}{1 + \sqrt{1 + 4L^2}} \left( -4L + 1 + (\sqrt{1 + 4L^2})^2 \right) = -4L(1 + L) \leq 0.
\]

On the other hand, when \(L \geq 2/3\), we obtain, using Lemma C-2,
\[
-4L + 1 + \frac{\ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)}{1 + \sqrt{1 + 4L^2}} \left( 4L - 1 - \sqrt{1 + 4L^2} \right)
\]
\[
\leq -4L + 1 + 2 \left( 4L - 1 - \sqrt{1 + 4L^2} \right)
\]
\[
= 4L - 1 - 2\sqrt{1 + 4L^2} < 0.
\]
Hence, $\Psi'(L) < 0$ when $\Psi(L) = 0$, that is, $\Psi(L)$ crosses zero at most once and the crossing is from above. Moreover, $\Psi(0) = 1/2 > 0$. Hence, there exists a threshold $\overline{L}(k)$ such that $\Psi(L) \geq 0$ if and only if $L \leq \overline{L}(k)$. Therefore, by Lemma C-3, $L_{\text{pooled}} \geq 2L_{\text{dedicated}} = 2L$ if and only if $L \leq \overline{L}(k)$. Using the Implicit Function Theorem, we obtain

$$
\overline{L}(k) = - \frac{\partial \Psi(L,k)}{\partial k} < 0,
$$

because $\left(1 + \frac{\sqrt{1 + 4L^2}}{2(L + 1)}\right) < 1$ and because $\Psi(L)$ crosses zero from above.

Define $\phi(L; k) \equiv e^{k+1}/(1 + L)^{k-1}$, and note that $\phi(L; k)$ is increasing in $L$. By Lemma C-7, $\phi(L; k) = k\lambda^k/\omega$. Since $\phi(.; k)$ is increasing, it is invertible; hence, $L_{\text{dedicated}} = \phi^{-1}(k\lambda^k/\omega; k)$. Hence, requiring that $L_{\text{dedicated}} < \overline{L}(k)$ is equivalent to requiring that $\phi^{-1}(k\lambda^k/\omega; k) < \overline{L}(k)$, or that $k\lambda^k/\omega < \phi(\overline{L}(k); k)$. Defining $W(k) \equiv k/\phi(\overline{L}(k); k)$, this is equivalent to requiring that $w/\lambda^k > W(k)$. Because $\phi(L; k)$ is continuous and $\overline{L}(k)$ is continuous by the Inverse Function Theorem, $W(k)$ is continuous. Moreover, because $\frac{\partial \phi(L,k)}{\partial k} = L^{1+k}(1 + L)^{1-k}(\ln(L) - \ln(1 + L)) < 0$, we obtain that

$$
W'(k) = \frac{1}{\phi(\overline{L}(k); k)} - \frac{k}{\phi(\overline{L}(k); k)^2} \left( \frac{\partial \phi(\overline{L}(k); k)}{\partial L} \overline{L}'(k) + \frac{\partial \phi(\overline{L}(k); k)}{\partial k} \right) \geq 0.
$$

Finally, when $k \leq 2$, because $(1 + \sqrt{1 + 4L^2})/(2(L + 1)) \leq 1$,

$$
\Psi(L; k) = - \sqrt{1 + 4L^2} + \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \geq - \frac{\sqrt{1 + 4L^2}}{2(L + 1)} + \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} = \frac{1 + 2L^2 - L\sqrt{1 + 4L^2}}{2(L + 1)^2} > 0,
$$

since $1 + 2L^2 > L\sqrt{1 + 4L^2}$ if and only if $1 + 4L^2 + 4L^2 > L^2 + 4L^4$ if and only if $1 + 3L^2 > 0$, which always holds. Hence, when $k \leq 2$, $L_{\text{pooled}} \geq 2L_{\text{dedicated}}$ for all $\lambda > 0$ and $w > 0$. \hfill \Box

**Lemma C.2.** When $L \geq 2/3$, the function

$$
F(L) \equiv \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right)^2 \sqrt{1 + 4L^2} - 2 \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \left( 1 + \sqrt{1 + 4L^2} \right)
$$

is nonnegative.

**Proof:** The proof uses Lemmas C-4, C-5, and C-6. Taking the derivative of $F(L)$ when it crosses zero yields:

$$
F'(L) \bigg|_{F(L)=0} = \frac{1}{(1 + L)^{\sqrt{1 + 4L^2}}} \left( 1 - 4L + 2\sqrt{1 + 4L^2} + 4L(1 + L) \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right) - 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \right) \bigg|_{F(L)=0}
$$

$$
= \frac{1}{(1 + L)^{\sqrt{1 + 4L^2}}}
$$

$$
x \left( 1 - 4L + 2\sqrt{1 + 4L^2} + 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \right) \frac{1 + \sqrt{1 + 4L^2}}{\sqrt{1 + 4L^2}} - 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)
$$

$$
= \frac{1}{(1 + L)^{\sqrt{1 + 4L^2}}}
$$

$$
x \left( 1 - 4L + 2\sqrt{1 + 4L^2} + 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \right) \frac{1}{\sqrt{1 + 4L^2}}
$$

$$
= \frac{L(1 + L)}{(1 + L)(1 + 4L^2)} \left( \frac{\sqrt{1 + 4L^2}}{L(1 + L)} \left( 1 - 4L + 2\sqrt{1 + 4L^2} \right) + 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \right).
$$
By Lemma C-4, there thus exists a threshold $\theta$ such that, when $F(L) = 0$, $F'(L) > 0$ if $2/3 \leq L < \theta$ and $F'(L) < 0$ if $L > \max\{\theta, 2/3\}$. Hence, when $L \geq 2/3$, $F(L)$ must cross zero at most twice as $L$ increases, first from below and then from above. However, because $F(2/3) = (16/3) \ln(5/4) - (5/3) \ln(2) \approx 0.0348563 > 0$, there is no crossing from below. By Lemma C-5, $\lim_{L \to \infty} F'(L) < 0$. Moreover, by Lemma C-6, $\lim_{L \to \infty} F(L) = 0$. Combining these two results shows that $F(L)$ tends to zero from above as $L$ tends to infinity, i.e., there is no crossing from above either. Hence, $F(L)$ never crosses zero as $L$ increases from $2/3$, and therefore it is always nonnegative. □

**Lemma C-3.** Under workload aversion, $L^{\text{pooled}} \geq 2L^{\text{dedicated}} = 2L$ if and only if

$$
\left(\frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)}\right)^k \geq \frac{\sqrt{1 + 4L^2}}{2(L + 1)}.
$$

**Proof:** The proof uses Lemma C-7. Using (7) and denoting $\mu^{\text{pooled}}$ by $\mu$ and $L^{\text{dedicated}}$ by $L$, we have:

$$
L^{\text{pooled}} \geq 2L^{\text{dedicated}} \iff \frac{2\mu\lambda}{\mu^2 - \lambda^2} \geq 2L
\iff 0 \geq L(\mu^2 - \lambda^2) - \mu\lambda
\iff \mu \leq \frac{\lambda}{2L} \left(1 + \sqrt{1 + 4L^2}\right) = \alpha(L)\lambda.
$$

By (C-1), $C'(\mu; \mu)$ is increasing in $\mu$. Hence, $\mu \leq \alpha\lambda$ if and only if $C'(\mu; \mu) \leq C'(\alpha\lambda; \alpha\lambda)$. Since $\mu$ solves $C'(\mu; \mu) = 0$ by Proposition 1, this is equivalent to requiring that $C'(\alpha\lambda; \alpha\lambda) \geq 0$. Hence, using (12), Lemma C-7, and the definition of $\alpha(L)$, we have

$$
\mu \leq \alpha(L)\lambda \iff C'(\alpha\lambda; \alpha\lambda) \geq 0
\iff -w(\lambda^3 + \lambda(\alpha\lambda)^2) + k(\alpha\lambda)^{k-1} 2(\lambda^2 - (\alpha\lambda)^2)^2 \geq 0
\iff -(1 + \alpha^2) + 2\frac{k\lambda^k}{\alpha} - 2 \alpha^{k-1} (1 - \alpha^2)^2 \geq 0
\iff -(1 + \alpha^2) + 2\frac{L^{k+1}}{(1 + L)^{k-1}} - 2 \alpha^{k-1} (1 - \alpha^2)^2 \geq 0
\iff -\frac{1 + \alpha^2}{(1 - \alpha^2)^2} + 2L^2\left(\frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)}\right)^{k-1} \geq 0
\iff -\frac{\sqrt{1 + 4L^2}}{1 + \sqrt{1 + 4L^2}} + \left(\frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)}\right)^{k-1} \geq 0
\iff -\sqrt{1 + 4L^2} + \left(\frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)}\right)^k \geq 0. \quad \Box
$$

**Lemma C-4.** When $L \geq 2/3$, the function

$$
G(L) = \frac{\sqrt{1 + 4L^2}}{L(1 + L)} \left(1 - 4L + 8\sqrt{1 + 4L^2}\right) + 8 \ln \left(\frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)}\right)
$$

crosses zero once only, and the crossing is from above.
Proof: First, note that
\[
\lim_{L \to 2/3} G(L) = \frac{9}{10} \left( \frac{25}{9} - \frac{80}{9} \ln \left( \frac{5}{4} \right) \right) \approx 0.714852 > 0
\]

\[
\lim_{L \to \infty} G(L) = \lim_{L \to \infty} \frac{\sqrt{1+4L^2} (1-4L)}{L(1+L)} + \lim_{L \to \infty} \frac{2(1+4L^2)}{L(1+L)} + \lim_{L \to \infty} \frac{8 \ln \left( \frac{1+\sqrt{1+4L^2}}{2(1+L)} \right)}{L(1+L)}
\]

\[
= \lim_{L \to \infty} \frac{\sqrt{\frac{1}{L^2} + 4 \left( \frac{1}{L} - 4 \right)}}{L(1+L)} + \lim_{L \to \infty} \frac{2 \left( \frac{1}{L^2} + 4 \right)}{L(1+L)} + \lim_{L \to \infty} 8 \ln \left( \frac{1+\sqrt{1+4L^2}}{2(1+L)} \right) = -8 + 8 + 0 = 0.
\]

Next, note that
\[
G'(L) = 0 \iff \frac{1}{L^2(1+L^2) \sqrt{1+4L^2}} \left( -1 - 10L - 12L^2 - 28L^3 - 2\sqrt{1+4L^2} \left( 1 - 2L - 8L^2 \right) \right) = 0
\]

\[
\iff 2\sqrt{1+4L^2} = \frac{-1 - 10L - 12L^2 - 28L^3}{1 - 2L - 8L^2}.
\]

The numerator of ratio in the right-hand side is nonpositive for all \( L \geq 0 \), whereas its denominator is nonpositive for all \( L \geq 1/4 \). Hence, the ratio is nonnegative for all \( L \geq 1/4 \), and therefore for all \( L \geq 2/3 \).

Consider any point \( \bar{L} \) such that \( G'(\bar{L}) = 0 \). At that point \( \bar{L} \), we have
\[
G''(\bar{L}) = \left( \frac{1}{L^2(1+L^2) \sqrt{1+4L^2}} \right) \times \left( -1 - 10L - 12L^2 - 2\sqrt{1+4L^2} \left( 1 - 2L - 8L^2 \right) \right)
\]

\[
+ \left( \frac{1}{L^2(1+L^2) \sqrt{1+4L^2}} \right) \times \left( -1 - 10L - 12L^2 - 2\sqrt{1+4L^2} \left( 1 - 2L - 8L^2 \right) \right)
\]

\[
= \left( \frac{1}{L^2(1+L^2) \sqrt{1+4L^2}} \right) \times \left( 4 \left( 1 + 6L + 8L^2 + 48L^3 \right) - 2\sqrt{1+4L^2} \left( 5 + 12L + 42L^2 \right) \right)
\]

\[
= \left( \frac{1}{L^2(1+L^2) \sqrt{1+4L^2}} \right) \times \left( 4 \left( 1 + 6L + 8L^2 + 48L^3 \right) - \frac{-1 - 10L - 12L^2 - 28L^3}{1 - 2L - 8L^2} \left( 5 + 12L + 42L^2 \right) \right)
\]

\[
= \left( \frac{1}{L^2(1+L^2) \sqrt{1+4L^2}} \right) \times \left( 9 + 78\bar{L} + 174\bar{L}^2 + 640\bar{L}^3 + 200\bar{L}^4 - 360\bar{L}^5 \right),
\]
in which the second and third equalities follow from using the fact that \( G(\bar{L}) = 0 \).

Because \( 1 - 2\bar{L} - 8\bar{L}^2 < 0 \) for all \( \bar{L} \geq 2/3 \), \( G''(\bar{L}) \leq 0 \) if and only if \( H(L) \equiv 9 + 78\bar{L} + 174\bar{L}^2 + 640\bar{L}^3 + 200\bar{L}^4 - 360\bar{L}^5 \geq 0 \). Because \( H''''(\bar{L}) < 0 \) for all \( \bar{L} \geq 0 \) and \( H''''(0) = 4800 > 0 \), we obtain that \( H''''(\bar{L}) \) is initially positive and then negative as \( \bar{L} \) increases from zero. Hence, \( H''''(\bar{L}) \) is initially increasing and then decreasing as \( \bar{L} \) increases. Because \( H'''(0) = 3840 > 0 \), we thus obtain that \( H'''(\bar{L}) \) is initially positive and then negative as \( \bar{L} \) increases from zero. Hence, \( H''(\bar{L}) \) is initially increasing and then decreasing as \( \bar{L} \) increases. Because \( H''(0) = 78 > 0 \), we thus obtain that \( H'(\bar{L}) \) is initially positive and then negative as \( \bar{L} \) increases from zero. Hence, \( H'(\bar{L}) \) is initially increasing and then decreasing as \( \bar{L} \) increases. Finally, because \( H(0) = 9 > 0 \), we thus obtain that \( H(\bar{L}) \) crosses zero at most once as \( \bar{L} \) increases from zero, and the crossing is from above. Hence, there exists a threshold \( \theta \) (potentially infinite) such that \( G''''(\bar{L}) < 0 \) if \( 2/3 \leq \bar{L} < \theta \) and \( G''''(\bar{L}) > 0 \) if \( \bar{L} > \max\{2/3, \theta \} \).

Because \( G'(\bar{L}) = 0 \) by definition, this implies that, when \( L \geq 2/3 \), \( G(L) \) has at most two stationary points, encountering, as \( L \) increases, first one local maximum and then one local minimum. Because \( G(2/3) > 0 \) and \( \lim_{L \to \infty} G(L) = 0 \), and, because \( G(1) < 0 \), we thus conclude that \( G(L) \) crosses zero exactly once and the crossing is from above. □
LEMMA C-5.

\[
\lim_{L \to \infty} 1 - 4L + 2\sqrt{1 + 4L^2} + 4L(1 + L) \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right) - 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) = -\frac{1}{2}
\]

Proof: First, note that

\[
\lim_{L \to \infty} 1 - 4L + 2\sqrt{1 + 4L^2} = \lim_{L \to \infty} \frac{\frac{1}{L} - 4 + 2\sqrt{\frac{1}{L^2} + 4}}{L^2} = \lim_{L \to \infty} \frac{\frac{(-1)}{L^2} + \frac{2}{L^3}\sqrt{\frac{1}{L^2} + 4}}{L^2} = \lim_{L \to \infty} 1 + \frac{2}{\sqrt{1 + 4L^2}} = 1, \quad (C-3)
\]

where we used L’Hospital’s rule. Similarly,

\[
\lim_{L \to \infty} 1 + 2L - \sqrt{1 + 4L^2} = \lim_{L \to \infty} \frac{\frac{1}{L} + 2 - \sqrt{\frac{1}{L^2} + 4}}{L^2} = \lim_{L \to \infty} \frac{\frac{(-1)}{L^2} - \frac{1}{L^3}\sqrt{\frac{1}{L^2} + 4}}{L^2} = \lim_{L \to \infty} 1 - \frac{1}{\sqrt{1 + 4L^2}} = 1. \quad (C-4)
\]

Finally,

\[
\lim_{L \to \infty} 4L(1 + L) \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right) - 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) = \lim_{L \to \infty} 4 \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right) - 8 \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)
\]

\[
= \lim_{L \to \infty} \frac{4\sqrt{1 + 4L^2} - 8\sqrt{1 + 4L^2}}{2(L + 1)^2} = \lim_{L \to \infty} -\frac{4L^2(1 + L)(1 + \frac{1}{L^2} + 4)L(1 + 2L - \sqrt{1 + 4L^2})}{(1 + 2L)(1 + 4L^2)(1 + \sqrt{1 + 4L^2})}
\]

\[
= \lim_{L \to \infty} -\frac{4 \left( \frac{1}{L} + 1 \right) \left( \frac{1}{L} + \sqrt{\frac{1}{L^2} + 4} \right)}{2 \times 4 \times 2} = -\frac{3}{2}
\]

in which we used L’Hospital’s rule and (C-4). Therefore, using (C-3) and the last equation, we obtain

\[
\lim_{L \to \infty} 1 - 4L + 2\sqrt{1 + 4L^2} + 4L(1 + L) \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right) - 8L(1 + L) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) = 1 - \frac{3}{2} = -\frac{1}{2}. \quad \square
\]

LEMMA C-6.

\[
\lim_{L \to \infty} \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right) \sqrt{1 + 4L^2} - 2 \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \left( 1 + \sqrt{1 + 4L^2} \right) = 0
\]
Proof: We have
\[
\lim_{L \to \infty} \ln \left( \frac{\sqrt{1 + 4L^2}}{2(L + 1)} \right) \sqrt{1 + 4L^2} - 2 \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) \left( 1 + \sqrt{1 + 4L^2} \right)
\]
\[
= \lim_{L \to \infty} \frac{2 + 8L^2 + \sqrt{1 + 4L^2} - 4L\sqrt{1 + 4L^2} + 8L(1 + L)}{4L(1 + L)} \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right)
\]
\[
= \lim_{L \to \infty} \frac{2 + 8L^2 + \sqrt{1 + 4L^2} - 4L\sqrt{1 + 4L^2} + 8L(1 + L)}{4L(1 + L)} \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right)
\]
\[
= \lim_{L \to \infty} \frac{8 - 4 \times 2 + 8 \times 0}{4} = 0,
\]
in which we used L'Hospital's rule. □

Proof of Proposition 4. The proof uses Lemma C-7. Denote $\mu^{\text{centralized}}$ by $\mu$ and $L^{\text{dedicated}}$ by $L$. Similar to the proof of Lemma C-3, we have
\[
L^{\text{centralized}} \geq 2L^{\text{dedicated}} \iff \mu \leq \frac{\lambda}{2L} \left( 1 + \sqrt{1 + 4L^2} \right) \equiv \alpha(L)\lambda
\]
We showed in §4.1 that $\partial C(\mu, \mu)/\partial \mu_i$ is increasing in $\mu$. Hence, $\mu \leq \alpha \lambda$ if and only if $\partial C(\mu, \mu)/\partial \mu_i \leq \partial C(\alpha \lambda, \alpha \lambda)/\partial \mu_i$. Since $\mu$ solves $\partial C(\mu, \mu)/\partial \mu_i = 0$ by (13), this is equivalent to requiring that $\partial C(\alpha \lambda, \alpha \lambda)/\partial \mu_i \geq 0$. Hence, similar to the proof of Lemma C-3, using (13), Lemma C-7, and the definition of $\alpha(L)$, we obtain
\[
\mu \leq \alpha(L)\lambda \iff -\frac{\sqrt{1 + 4L^2}}{(L + 1)} + \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)^k \geq 0.
\]
Since $(1 + \sqrt{1 + 4L^2})/(2(L + 1)) \leq 1$, we have for all $k \geq 1$:
\[
-\frac{\sqrt{1 + 4L^2}}{(L + 1)} + \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)^k \leq -\frac{\sqrt{1 + 4L^2}}{(L + 1)} + \left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right) < 0.
\]
Hence, when $k \geq 1$, $L^{\text{centralized}} \leq 2L^{\text{dedicated}}$ for all $\lambda > 0$ and $w > 0$. □

Lemma C-7. Under workload aversion, $L := L^{\text{dedicated}}$ satisfies
\[
L^{k+1} = \frac{k\lambda^k}{(1 + L)^{k-1}} \frac{w^k}{(w/k\lambda)^{k+1}}. \quad \frac{k\lambda^k}{(1 + L)^{k-1}} \frac{w^k}{(w/k\lambda)^{k+1}}
\]

Proof: From (4), we have that $L = \lambda/(\mu - \lambda)$. Hence, (10) can be re-expressed as $\mu^{k-1} = (w/(k\lambda))L^2$. Accordingly,
\[
L = \frac{\lambda}{\mu - \lambda} = \frac{\lambda}{(w/k\lambda)^{1/2}} \frac{L^{1/2}}{L^{1/2}} - \lambda = \frac{1}{(w/k\lambda)^{1/2}} L^{1/2} - 1
\]
\[
\iff L^{k+1} = \frac{w^k}{(w/k\lambda)^{k+1}} - L = 1
\]
\[
\iff \frac{w}{k\lambda} = \frac{(1 + L)^{k-1}}{L^{k+1}}. \quad \frac{(1 + L)^{k-1}}{L^{k+1}}
Proof of Proposition 5. The proof uses Lemmas C-8 and C-9. Denote \( f(\alpha) \doteqdot \alpha^3 Q - \alpha Q - 1 \). Because \( f(\alpha) \) has a unique positive root by Lemma C-8, denoted by \( \alpha(Q) \), because \( f(0) = -1 < 0 \), and because
\[
f \left( \frac{1 + \sqrt{1 + \frac{4}{Q}}}{2} \right) = \frac{1 + \sqrt{1 + \frac{4}{Q}}}{2} > 0,
\]
we have that \( 2\alpha(Q) \leq 1 + \frac{1 + \frac{4}{Q}}{2} \). Accordingly, and given that \( k \geq 1 \), we thus have
\[
\left( \frac{2\alpha(Q)}{1 + \sqrt{1 + \frac{4}{Q}}} \right)^k \leq \left( 4(\alpha(Q))^3 (1 + (\alpha(Q))^2) \right) \left( 1 + \sqrt{1 + \frac{4}{Q}} \right)^3
\]
in which the last two inequalities follow because \( \alpha(Q) \) is greater than \( 1/\sqrt{3} \), which corresponds to the minimum of \( f(\alpha) \) when \( \alpha \geq 0 \). Hence,
\[
\left( \frac{2\alpha(Q)}{1 + \sqrt{1 + \frac{4}{Q}}} \right)^k \leq \left( 4(\alpha(Q))^3 (1 + (\alpha(Q))^2) \right) \left( 1 + \sqrt{1 + \frac{4}{Q}} \right)^3
\]
for all \( k \geq 1 \), which is equivalent, by Lemma C-8, to \( Q_{\text{pooled}} \leq 2Q_{\text{dedicated}} \) for all \( k \geq 1 \). □

Lemma C-8. Under workload aversion, \( Q_{\text{pooled}} \geq 2Q_{\text{dedicated}} = 2Q \) if and only if
\[
\left( \frac{2\alpha(Q)}{1 + \sqrt{1 + \frac{4}{Q}}} \right)^k \geq \left( 4(\alpha(Q))^3 (1 + (\alpha(Q))^2) \right) \left( 1 + \sqrt{1 + \frac{4}{Q}} \right)^3,
\]
in which \( \alpha(Q) \) is the only positive root of \( \alpha^3 Q - \alpha Q - 1 = 0 \).

Proof: The proof uses Lemma C-9. Using (8) and denoting \( \mu_{\text{pooled}} \) by \( \mu \) and \( Q_{\text{dedicated}} \) by \( Q \), we have:
\[
Q_{\text{pooled}} \geq 2Q_{\text{dedicated}} \iff \frac{2\lambda^3}{\mu(\mu^2 - \lambda^2)} \geq 2Q
\]
\[
\iff 0 \geq Q\mu^3 - \lambda^2 Q\mu - \lambda^3
\]
\[
\iff 0 \geq Q \left( \frac{\mu}{\lambda} \right)^3 - Q \frac{\mu}{\lambda} - 1
\]
\[
\iff \mu \leq \alpha(Q)\lambda.
\]
Note that \( \alpha(Q) \) is the only positive root of \( f(\alpha) = \alpha^3 Q - \alpha Q - 1 = 0 \) since \( f(0) = -1 < 0 \), \( f'(\alpha) = Q(3\alpha^2 - 1) < 0 \) if and only if \( \alpha < 1/\sqrt{3} \) (when \( \alpha > 0 \)), and \( \lim_{\alpha \to \infty} f(\alpha) > 0 \).

By (C-1), \( C'(\mu; \mu) \) is increasing in \( \mu \). Hence, \( \mu \leq \alpha \lambda \) if and only if \( C'(\mu; \mu) \leq C'(\alpha \lambda; \alpha \lambda) \). Since \( \mu \) solves \( C'(\mu; \mu) = 0 \) by Proposition 1, this is equivalent to requiring that \( C'(\alpha \lambda; \alpha \lambda) \geq 0 \). Hence, using (12), Lemma C-9, and the definition of \( \alpha(Q) \), we have
\[
\mu \leq \alpha(Q)\lambda \iff C'(\alpha \lambda; \alpha \lambda) \geq 0
\]
\[
\iff -w(\lambda^3 + \lambda (\alpha \lambda)^2) + k (\alpha \lambda)^{k-1} 2((\alpha \lambda)^2 - \lambda^2)^2 \geq 0
\]
\[
\iff -(1 + \alpha^2) + 2 \frac{k \lambda^k}{w} \alpha^{k-1}(\alpha^2 - 1)^2 \geq 0
\]
We then obtain the result by replacing \((\alpha^2 - 1)^2 Q^2\) with \(1/\alpha^2\) given that \(\alpha\) solves \(\alpha^2 Q - \alpha Q - 1 = 0\). □

**Lemma C.9.** Under workload aversion, \(Q = Q_{\text{dedicated}}\) satisfies
\[
\frac{k\lambda^k}{w} = \frac{2k^3Q^2}{(1 + \sqrt{1 + \frac{2}{Q}})^{k-3}}.
\]

**Proof:** Denote \(\mu_{\text{dedicated}}\) by \(\mu\) and \(Q_{\text{dedicated}}\) by \(Q\). From (5), we have that \(Q = \lambda^2/|\mu(\mu - \lambda)|\), when \(\mu > \lambda\).
\[
Q = \frac{\lambda^2}{\mu(\mu - \lambda)} \iff Q\mu^2 - \mu\lambda Q - \lambda^2 = 0
\]
\[
\iff \mu = \frac{\lambda Q + \sqrt{\lambda^2Q^2 + 4\lambda^2Q}}{2Q} = \frac{\lambda}{2} \left(1 + \sqrt{1 + \frac{4}{Q}}\right).
\]
On the other hand, (10) can be re-expressed as \(\mu^{k-3} = (w/(k\lambda^3))Q^2\). Accordingly,
\[
\left(\frac{w}{k\lambda^3}\right)^{\frac{1}{k-3}} Q^{\frac{2}{k-3}} = \frac{\lambda}{2} \left(1 + \sqrt{1 + \frac{4}{Q}}\right) \iff \left(\frac{k\lambda^k}{w}\right)^{\frac{1}{k-3}} = \frac{2Q^{\frac{2}{k-3}}}{\left(1 + \sqrt{1 + \frac{4}{Q}}\right)\lambda} \iff \frac{k\lambda^k}{w} = \frac{2k^3Q^2}{(1 + \sqrt{1 + \frac{2}{Q}})^{k-3}}.\]

**EC.3.2. Busyness Aversion**

**Proof of Proposition 7.** The proof uses Lemmas C-10, C-14, and C-15. Because the function \(\mu^{k-1} - b/(\lambda + \mu)\) is increasing, it crosses zero at most once, and if it does, the crossing is from below. Hence, the function \(\mu \cdot (\mu^{k-1} - b/(\lambda + \mu))\) crosses zero at most once when \(\mu > 0\), and the crossing is from below. Let \(M = \sup\{\mu|b \cdot \mu/(\lambda + \mu) \leq 0\}\). Because \(\mu^{k-1} - b \cdot \mu/(\lambda + \mu)\) is continuous and because \(\lim_{\mu \to \infty} \mu^{k-1} - b \cdot \mu/(\lambda + \mu) = \infty\), we obtain that \(M < \infty\).

For any \(\mu_i > \max\{0, 2\lambda - \mu_{-i}\}\), we obtain from (16) that
\[
\frac{\partial C_i(\mu_i; \mu_{-i})}{\partial \mu_{-i}} = -\frac{b\lambda^2\mu_i(4\lambda\mu_{-i} + (\mu_i + \mu_2)^2)}{\mu_i^2(\lambda + \mu_2) + \mu_i^2(\lambda + \mu_1)} < 0.
\]
Hence, for any \(\mu_i > \max\{0, 2\lambda - \mu_{-i}\}\), \(C_i(\mu_i; \mu_{-i}) > \lim_{\mu_{-i} \to \infty} C_i(\mu_i; \mu_{-i}) = \mu_i^k + b \cdot \lambda/(\lambda + \mu_i)\). Because \(\mu_i^{k-1} + b \cdot \lambda/(\lambda + \mu_i) > b\) for all \(\mu_i > M\), we thus obtain that \(C_i(\mu_i; \mu_{-i}) > b = C_i(0; \mu_{-i})\) for any \(\mu_i > \max\{M, 2\lambda - \mu_{-i}\}\). Hence, there is no equilibrium such that \(\mu_i > \max\{M, 2\lambda\}\) because otherwise setting \(\mu_i\) to zero would result in a lower cost. Accordingly, one can restrict without loss of generality the strategy set of each server to the compact set \(D := [0, \min\{M, 2\lambda\}]\).

Let \(\Phi_i(\mu_{-i})\) be Server \(i\)'s best-response correspondence, i.e., \(\Phi_i(\mu_{-i}) = \arg\min_{\mu_i \in D} C_i(\mu_i; \mu_{-i})\). Because \(C_i(\mu_i; \mu_{-i})\) is continuous and the action sets are compact, \(C_i(\mu_i; \mu_{-i})\) attains its minimum on \(D\) by Weierstrass’ Theorem, and \(\Phi_i(\mu_{-i})\) is nonempty. Let \(\varphi_i(\mu_{-i})\) be the minimal or maximal selection of \(\Phi_i(\mu_{-i})\). We next show that \(\varphi_i(\mu_{-i})\) is quasi-increasing (Amir and De Castro 2013), i.e., that
\[
\limsup_{\mu_i \to \mu_{-i}} \varphi(i(\mu_{-i}) \leq \varphi_i(\mu_{-i}) \leq \liminf_{\mu_i \to \mu_{-i}} \varphi_i(\mu). \tag{C-6}
\]
Fix \( \hat{\mu}_i \). By Lemma C-14, \( C_i(\mu_i; \hat{\mu}_i) \) has at most two global minima, neither of which is equal to \( 2\lambda - \hat{\mu}_i \); hence, \( \Phi_i(\hat{\mu}_i) \) has one or two elements and \( 2\lambda - \hat{\mu}_i \notin \Phi_i(\hat{\mu}_i) \). We next consider four cases: (i) \( \Phi_i(\hat{\mu}_i) = \{ \mu_i, \overline{\mu}_i \} \) with \( 0 < \mu_i < \overline{\mu}_i \), (ii) \( \Phi_i(\hat{\mu}_i) = \{ \mu_i, \overline{\mu}_i \} \) with \( 0 < \mu_i < \overline{\mu}_i \), (iii) \( \Phi_i(\hat{\mu}_i) = \{ \hat{\mu}_i \} \) with \( \hat{\mu}_i > 0 \), and (iv) \( \Phi_i(\hat{\mu}_i) = \{ \hat{\mu}_i \} \) with \( \hat{\mu}_i = 0 \).

Case (i): \( \Phi_i(\hat{\mu}_i) = \{ \mu_i, \overline{\mu}_i \} \) with \( 0 < \mu_i < \overline{\mu}_i \). By Lemma C-14, having \( \mu_i > 0 \) implies that \( \hat{\mu}_i > 2\lambda \).

By Lemma C-14, \( \mu_i \) and \( \overline{\mu}_i \) are strict minima, i.e., there exists an \( \epsilon > 0 \) such that \( C_i(\mu_i; \hat{\mu}_i) < C_i(\mu_i; \mu_i) \) for all \( \mu_i \in B(\mu_i, \epsilon) \cap \mathcal{D} \), \( \mu_i \neq \mu_i \), and such that \( C_i(\overline{\mu}_i; \hat{\mu}_i) < C_i(\mu_i; \hat{\mu}_i) \) for all \( \mu_i \in B(\overline{\mu}_i, \epsilon) \cap \mathcal{D} \), \( \mu_i \neq \mu_i \), in which \( B(\mu_i, \epsilon) := \{ \mu_i : |\mu_i - \mu_i| < \epsilon \} \). From hereon, we fix that \( \epsilon \).

Define \( N(\mu_i, \epsilon/2) := \{ \mu_i : |\mu_i - \mu_i| \leq \epsilon/2 \} \cap \mathcal{D} \). Let \( V(\mu_i) := \min_{\mu_i \in N(\mu_i, \epsilon/2)} C_i(\mu_i; \hat{\mu}_i) \) and \( V(\mu_i) := \min_{\mu_i \in N(\mu_i, \epsilon/2)} C_i(\mu_i; \mu_i) \). Because \( C_i(\mu_i; \hat{\mu}_i) \) is continuous in \( (\mu_i; \hat{\mu}_i) \) and \( N(\mu_i, \epsilon/2) \) is compact, \( V(\mu_i) \) is continuous in \( \mu_i \), by the Maximum Theorem (Sundaram 1996, Theorem 9.14). Moreover, because \( \partial C_i(\mu_i; \hat{\mu}_i)/\partial \mu_i \) is continuous in \( (\mu_i; \hat{\mu}_i) \) and because \( \arg \min_{\mu_i \in N(\mu_i, \epsilon/2)} C_i(\mu_i; \hat{\mu}_i) = \{ \mu_i \} \), \( V(\mu_i) \) is differentiable at \( \mu_i = \hat{\mu}_i \) and \( V'(\hat{\mu}_i) = \partial C_i(\mu_i; \hat{\mu}_i)/\partial \mu_i \) (Milgrom and Segal 2002, Corollary 4 (iii)). Similarly, we obtain \( V'(\mu_i) \) is differentiable at \( \mu_i = \hat{\mu}_i \) and \( V'(\hat{\mu}_i) = \partial C_i(\overline{\mu}_i; \hat{\mu}_i)/\partial \mu_i \) by Lemma C-10, since \( \hat{\mu}_i > 2\lambda \) and \( \overline{\mu}_i > \mu_i > 0 \), we obtain:

\[
\frac{V(\mu_i) - V(\hat{\mu}_i)}{\mu_i - \hat{\mu}_i} = \frac{\partial C_i(\mu_i; \hat{\mu}_i)}{\partial \mu_i} - \frac{\partial C_i(\mu_i; \hat{\mu}_i)}{\partial \mu_i} - \frac{o \left( |\mu_i - \hat{\mu}_i| \right)}{\mu_i - \hat{\mu}_i}.
\]

Hence,

\[
\lim_{\mu_i \rightarrow \hat{\mu}_i} \frac{V(\mu_i) - V(\hat{\mu}_i)}{\mu_i - \hat{\mu}_i} = \lim_{\mu_i \rightarrow \hat{\mu}_i} \frac{o \left( |\mu_i - \hat{\mu}_i| \right)}{\mu_i - \hat{\mu}_i} = 0.
\]

By the Maximum Theorem, \( \Phi_i(\hat{\mu}_i) \) is upper semi-continuous (Sundaram 1996, Theorem 9.14). Hence, \( \exists \delta > 0 : |\mu_i - \hat{\mu}_i| < \delta \Rightarrow \Phi_i(\mu_i) \subset B(\mu_i, \epsilon/3) \cup B(\overline{\mu}_i, \epsilon/3) \).

For any \( \mu_i > \hat{\mu}_i \) such that \( |\mu_i - \hat{\mu}_i| < \delta \), suppose that \( \Phi_i(\mu_i) \cap B(\mu_i, \epsilon/3) \neq \emptyset \). By (C-7), when \( \mu_i > \hat{\mu}_i \), if \( \Phi_i(\mu_i) \cap B(\mu_i, \epsilon/3) \neq \emptyset \), then \( \Phi_i(\mu_i) \cap B(\overline{\mu}_i, \epsilon/3) \neq \emptyset \). On the other hand, suppose that \( \Phi_i(\mu_i) \cap B(\mu_i, \epsilon/3) = \emptyset \). Because \( \Phi_i(\mu_i) \subset B(\mu_i, \epsilon/3) \cup B(\overline{\mu}_i, \epsilon/3) \), we must thus have that \( \Phi_i(\mu_i) \subset B(\overline{\mu}_i, \epsilon/3) \). In either case, \( \Phi_i(\mu_i) \cap B(\overline{\mu}_i, \epsilon/3) \neq \emptyset \). Therefore, if \( \varphi(\mu_i) \) is defined as the maximal selection of \( \Phi_i(\mu_i) \), then \( \varphi(\mu_i) = \overline{\mu}_i = \lim inf_{\mu_i \rightarrow \hat{\mu}_i} \varphi_i(\mu_i) \). If \( \varphi(\mu_i) \) is defined as the minimal selection of \( \Phi_i(\mu_i) \), then \( \varphi(\mu_i) = \overline{\mu}_i \leq \lim sup_{\mu_i \rightarrow \hat{\mu}_i} \varphi_i(\mu_i) \). Summarizing both cases, we obtain that \( \varphi(\mu_i) \leq \lim inf_{\mu_i \rightarrow \hat{\mu}_i} \varphi_i(\mu_i) \). Similarly, one can show that \( \varphi(\mu_i) \geq \lim sup_{\mu_i \rightarrow \hat{\mu}_i} \varphi_i(\mu_i) \). Hence, (C-6) holds.
Case (ii): $\Phi_i(\tilde{\mu}_{-i}) = \{\tilde{\mu}_i, \overline{\mu}_i\}$ with $0 = \tilde{\mu}_i < \overline{\mu}_i$. Because $C_i(0; \mu_{-i}) = b$ for all $\mu_{-i}$, $\partial C_i(\tilde{\mu}_{-i}; \mu_{-i})/\partial \mu_{-i} = 0$. On the other hand, by (C-5), $\partial C_i(\overline{\mu}_{-i}; \mu_{-i})/\partial \mu_{-i} < 0$. Hence, $\partial C_i(\overline{\mu}; \mu_{-i})/\partial \mu_{-i} > \partial C_i(\mu; \mu_{-i})/\partial \mu_{-i}$. Applying the same argument as the one above shows that $\limsup_{\mu_{-i} \uparrow \overline{\mu}_{-i}} \varphi_i(\mu_{-i}) = 0 \leq \varphi_i(\tilde{\mu}_{-i}) \leq \overline{\mu}_i = \liminf_{\mu_{-i} \uparrow \overline{\mu}_{-i}} \varphi_i(\mu_{-i})$, i.e., (C-6) holds.

Case (iii): $\Phi_i(\tilde{\mu}_{-i}) = \{\hat{\mu}_i\}$ with $\hat{\mu}_i > 0$. Applying Corollary 4 (iii) by Milgrom and Segal (2002) shows that $C_i(\varphi_i(\tilde{\mu}_{-i}); \tilde{\mu}_{-i})$ is differentiable at $\tilde{\mu}_{-i}$ and its derivative equals $\partial C_i(\tilde{\mu}_{-i}; \tilde{\mu}_{-i})/\partial \mu_{-i}$. Applying the same argument as the one above shows that $\limsup_{\mu_{-i} \uparrow \tilde{\mu}_{-i}} \varphi_i(\mu_{-i}) = \hat{\mu}_i = \varphi_i(\tilde{\mu}_{-i}) = \hat{\mu}_i = \liminf_{\mu_{-i} \uparrow \tilde{\mu}_{-i}} \varphi_i(\mu_{-i})$, i.e., (C-6) holds.

Because $\varphi_i(\mu_{-i})$ is quasi-increasing and the strategy sets are compact, there exists a symmetric equilibrium (Milgrom and Roberts 1994, Corollary 1; Amir and De Castro 2013, Corollary 10; Vives 1999, p. 41).

By Lemma C-14, the symmetric equilibrium capacity investments are either equal to zero or such that $\mu_1 + \mu_2 > 2\lambda$. In the latter case, because $\mu_1 = \mu_2 = \mu^*$, $\mu^*$ must solve (17). Because the left-hand side of (17) is increasing in $\mu$, there exists only one $\mu^*$ satisfying (17).

We next show that $\mu^* > \lambda$ when $b \geq \lambda^k(k + 1)^{1+k}$. Because the left-hand side of (17) is increasing in $\mu$, showing that $\mu^* > \lambda$ is equivalent to showing that $C_i'(\lambda; \lambda) < 0$, i.e., that $k\lambda^k - b/2 < 0$. First, note that the function $(k/(1 + k))^{1+k}$ is nondecreasing, because its derivative is equal to $(k/(1 + k))^{1+k}(\ln (k/(1 + k)) + 1/k) \geq 0$ given the logarithm inequality $\ln(x) \geq 1 - 1/x$ for $x > 0$. Accordingly, when $b \geq \lambda^k(k + 1)^{1+k}$,

$$C_i'(\lambda; \lambda) = k\lambda^k - \frac{b}{2} \leq k \left(\frac{k}{1 + k}\right)^{1+k} - \frac{b}{2} \leq \lim_{k \to \infty} \left(\frac{k}{1 + k}\right)^{1+k} - \frac{b}{2} = b \left(\frac{1}{e} - \frac{1}{2}\right) < 0,$$

in which $e$ is Euler’s number, i.e., $e \approx 2.71828$. □

**Lemma C-10.** For any $\mu_{-i} > 2\lambda$, suppose that both $\mu_i$ and $\overline{\mu}_i$ globally minimize $C_i(\mu_i; \mu_{-i})$, defined in (16), with $0 < \mu_i < \overline{\mu}_i$. Then,

$$\frac{\partial C_i(\mu_i; \mu_{-i})}{\partial \mu_{-i}} > \frac{\partial C_i(\overline{\mu}_i; \mu_{-i})}{\partial \mu_{-i}}.$$

**Proof:** The proof uses Lemmas C-11, C-12, and C-13. Because $\mu_i$ and $\overline{\mu}_i$ minimize $C_i(\mu_i; \mu_{-i})$ in the interior of its domain and because $\mu_1 + \mu_2 > 2\lambda$, the first-order optimality conditions

$$C'_i(\mu_i; \mu_{-i}) = k\mu_i^{k-1} - b\lambda \left(\frac{\mu_i + \mu_2}{\lambda + \mu_2} + \frac{\mu_i + \mu_1}{\lambda + \mu_1}\right) \left(\frac{2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_1 + \mu_2)}{(\mu_i^2 + \mu_2^2)(\lambda + \mu_2)^2}\right) = 0,$$

are satisfied at both $\mu_i$ and $\overline{\mu}_i$. Therefore,

$$\frac{k}{b\lambda} \frac{\mu_i^{k-1}(\mu_i + \mu_{-i})(2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_1 + \mu_2))}{\mu_i^{k-1}(\mu_i + \mu_{-i})(2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_1 + \mu_2))} = \frac{\overline{\mu}_i^{k-1}(\mu_i + \mu_{-i})(2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_1 + \mu_2))}{\overline{\mu}_i^{k-1}(\mu_i + \mu_{-i})(2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_1 + \mu_2))},$$

$$\Rightarrow \frac{\mu_i^{k-1}(\mu_i + \mu_{-i})(2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_1 + \mu_2))}{\mu_i^{k-1}(\mu_i + \mu_{-i})(2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_1 + \mu_2))} = \frac{(\overline{\mu}_i^2(\lambda + \mu_{-i}) + \mu_{-i}^2(\lambda + \mu_i))^2}{(\overline{\mu}_i^2(\lambda + \mu_{-i}) + \mu_{-i}^2(\lambda + \mu_i))^2},$$

and

(C-9)
Suppose, for contradiction, that
\[
\frac{\partial C_i(\mu_i; \mu_{-i})}{\partial \mu_{-i}} \leq \frac{\partial C_i(\bar{\mu}_i; \mu_{-i})}{\partial \mu_{-i}},
\]
i.e., after taking the derivative of (16) with respect to \( \mu_{-i} \), that
\[
-b\lambda^2 \frac{\mu_i^2 (4\lambda \mu_{-i} + (\mu_i + \mu_{-i})^2)}{(\mu_i^2 + \mu_{-i}^2 + 2 \mu_i \mu_{-i} + \lambda + \mu_{-i})^2} \leq -b\lambda^2 \frac{\bar{\mu}_i^2 (4\lambda \mu_{-i} + (\bar{\mu}_i + \mu_{-i})^2)}{(\bar{\mu}_i^2 + \mu_{-i}^2 + 2 \bar{\mu}_i \mu_{-i} + \lambda + \mu_{-i})^2}.
\]
\[
\iff \frac{\mu_i^{k-1}(\bar{\mu}_i + \mu_{-i}) (2\lambda(\bar{\mu}_i - \mu_{-i}) + \mu_{-i}(\bar{\mu}_i + \mu_{-i}))}{\mu_i^{k-1}(\mu_i + \mu_{-i}) (2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_i + \mu_{-i}))} \geq \frac{\bar{\mu}_i(4\lambda \mu_{-i} + (\bar{\mu}_i + \mu_{-i})^2)}{\mu_i(4\lambda \mu_{-i} + (\mu_i + \mu_{-i})^2)}.
\]
\[
\iff F(\bar{\mu}_i, \mu_{-i}) \geq F(\mu_i, \mu_{-i}),
\]
in which \( F(\mu_i, \mu_{-i}) \) is defined in (C-10).

We next consider two cases, depending on whether (i) \( k < 1.03 \) and \( \mu_{-i} < 2.03\lambda \) or (ii) either \( k \geq 1.03 \) or \( \mu_{-i} \geq 2.03\lambda \). (i) Suppose first that \( k < 1.03 \) and \( \mu_{-i} < 2.03\lambda \). By Lemma C-12, \( \mu_i < \lambda < \bar{\mu}_i \). Hence, by Lemma C-13, \( F(\bar{\mu}_i, \mu_{-i}) < F(\mu_i, \mu_{-i}) \), a contradiction. Suppose next that either \( k \geq 1.03 \) or \( \mu_{-i} \geq 2.03\lambda \). Because \( F(\mu_i, \mu_{-i}) \) is nonincreasing in \( \mu_i \) when either \( k \geq 1.03 \) or \( \mu_{-i} \geq 2.03\lambda \) by Lemma C-11, and \( 0 < \mu_i < \bar{\mu}_i \), we must have \( F(\bar{\mu}_i, \mu_{-i}) < F(\mu_i, \mu_{-i}) \), a contradiction. \( \square \)

**Lemma C-11.** For any \( \mu_i > 0, \mu_{-i} > 2\lambda \), if either (i) \( k \geq 1.03 \) or (ii) \( \mu_{-i} \geq 2.03\lambda \), the function
\[
F(\mu_i, \mu_{-i}) = \frac{(\mu_i + \mu_{-i}) (2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_i + \mu_{-i}))}{\mu_i^{k}(4\lambda \mu_{-i} + (\mu_i + \mu_{-i})^2)}
\]
is decreasing in \( \mu_i \).

**Proof:** The proof uses Lemma C-18. We have
\[
\frac{\partial F(\mu_i, \mu_{-i})}{\partial \mu_i} \times \mu_i^+ \times (4\lambda \mu_{-i} + (\mu_i + \mu_{-i})^2)^2
\]
\[
= 4\lambda \mu_i (4\lambda \mu_i + (\mu_i + \mu_{-i})(\mu_i + 3\mu_{-i}))
\]
\[
- k(\mu_i + \mu_{-i}) (2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_i + \mu_{-i})) (4\lambda \mu_{-i} + (\mu_i + \mu_{-i})^2)
\]
\[
= G(\mu_i, \mu_{-i})
\]
and therefore \( F(\mu_i, \mu_{-i}) \) is nonincreasing in \( \mu_i \) if and only if \( G(\mu_i, \mu_{-i}) \leq 0 \).

(i) Because \( G(\mu_i, \mu_{-i}) \) is decreasing in \( \mu_{-i} \) by Lemma C-18, we obtain, for any \( \mu_{-i} > 2\lambda \),
\[
G(\mu_i, \mu_{-i}) < G(\mu_i, 2\lambda) = -4\lambda^4 \mu_i \left( 24(k-1) + 4(5k-6) \frac{\mu_i}{\lambda} + 2(3k-1) \left( \frac{\mu_i}{\lambda} \right)^2 + k \left( \frac{\mu_i}{\lambda} \right)^3 \right).
\]
The term in parentheses in the right-hand side is a cubic function of \( \mu_i/\lambda \), it is convex for all \( \mu_i \geq 0 \) and \( k \geq 1 \). When \( k \leq 1 + \sqrt{42}/6 \), it reaches a local minimum at \( \mu_i/\lambda = 2(1 - 3k + \sqrt{1 + 12k - 6k^2})/(3k) \geq 0 \); otherwise, it is always increasing for any \( \mu_i/\lambda \geq 0 \). Suppose first that \( k > 1 + \sqrt{42}/6 \). For any \( \mu_i > 0 \), we have

\[
G(\mu_i, \mu_{-i}) < G(0, 2\lambda) = 0.
\]

Suppose next that \( k \leq 1 + \sqrt{42}/6 \). For any \( \mu_i \geq 0 \), we have:

\[
G(\mu_i, \mu_{-i}) < G(\mu_i, 2\lambda) \leq G\left(2\lambda \frac{1 - 3k + \sqrt{1 + 12k - 6k^2}}{3k}, 2\lambda\right) = -4\lambda^4 \mu_i \frac{16(1 - \sqrt{1 - 6(k - 2)^2} + 6k(-3 - 2\sqrt{1 - 6(k - 2)^2} + k(6 + \sqrt{1 - 6(k - 2)^2})\right)}{27k^2}
\]

Define \( H(k) = (1 - \sqrt{1 - 6(k - 2)^2} + 6k(-3 - 2\sqrt{1 - 6(k - 2)^2} + k(6 + \sqrt{1 - 6(k - 2)^2})\right) \). Since

\[
H'(k) = 18 \left(1 - \sqrt{1 - 6(k - 2)^2} + k(4 + \sqrt{1 - 6(k - 2)^2})\right) \geq 18 \left(1 - \sqrt{1 - 6(k - 2)^2} + (4 + \sqrt{1 - 6(k - 2)^2})\right) = 18 \times 3 > 0,
\]

and since \( H(k) = 0 \) when

\[
k = 3/2 - \frac{1}{4\sqrt{3}} + \frac{3}{2} \sqrt{-32 + (111448 - 11376\sqrt{79})^2/12 + 2(13931 + 1422\sqrt{79})^2/6} = 1.0273259295534587,
\]

we conclude that \( H(k) > 0 \) for all \( k \geq 1.03 \). Hence, for any \( k \in [1.03, 1 + \sqrt{42}/6] \) and \( \mu_i \geq 0 \),

\[
G(\mu_i, \mu_{-i}) < G\left(2\lambda \frac{1 - 3k + \sqrt{1 + 12k - 6k^2}}{3k}, 2\lambda\right) = -4\lambda^4 \mu_i \frac{16}{27k^2} H(k) \leq 0.
\]

Summarizing both cases, we thus find that \( G(\mu_i, \mu_{-i}) < 0 \) for all \( \mu_i \geq 0 \), \( \mu_{-i} > 2\lambda \) and \( k \geq 1.03 \).

(ii) For this part, we include \( k \) in the arguments of the function \( G(\mu_i, \mu_{-i}) \) defined in (C-11). Because \( G(\mu_i, \mu_{-i}, k) \) is decreasing in \( \mu_{-i} \) by Lemma C-18 and because it is decreasing in \( k \), we obtain, for any \( \mu_{-i} \geq 2.03\lambda \) and \( k \geq 1 \),

\[
G(\mu_i, \mu_{-i}, k) \leq G(\mu_i, 2.03\lambda, k) \leq \lambda^5 \left(-1.51331 - 1.00385 \frac{H_i}{\lambda} + 15.4982 \left(\frac{H_i}{\lambda}\right)^2 - 16.4836 \left(\frac{H_i}{\lambda}\right)^3 - 4.03 \left(\frac{H_i}{\lambda}\right)^4\right).
\]

The term in parentheses is a quartic polynomial in \( \mu_i/\lambda \) with three stationary points: at \( \mu_i/\lambda = -3.60574 \) (local maximum), at \( \mu_i/\lambda = 0.0342819 \) (local minimum), and at \( \mu_i/\lambda = 0.503789 \) (local maximum) and tends to negative infinity when \( \mu_i/\lambda \to \infty \). Hence, for any \( \mu_i > 0 \),

\[
G(\mu_i, \mu_{-i}, k) < G(0.503789\lambda, 2.03\lambda, 1) \leq -0.452775\lambda^5 < 0.
\]

Hence, \( G(\mu_i, \mu_{-i}, k) < 0 \) for all \( \mu_i > 0 \), \( k \geq 1 \) and \( \mu_{-i} \geq 2.03\lambda \). \( \square \)
Lemma C-12. For any \( k < 1.03 \) and \( 2\lambda < \mu_- < 2.03\lambda \), suppose that both \( \mu_i \) and \( \pi_i \) globally minimize \( C_i(\mu_i; \mu_-), \) defined in (16), with \( 0 < \mu_i < \pi_i \). Then, \( \mu_i < \lambda < \pi_i \).

Proof: The proof uses Lemma C-14. Because \( C_i(0; \mu_-) = b \), if \( \mu_i \) and \( \pi_i \) are global minima, then \( C_i(\mu_i; \mu_-) \leq b \) and \( C_i(\pi_i; \mu_-) \leq b \). Because \( \mu_i \) and \( \pi_i \) minimize \( C_i(\mu_i; \mu_-) \) in the interior of its domain and because \( \mu_1 + \mu_2 > 2\lambda \), the first-order optimality conditions (C-8) are satisfied at both \( \mu_i \) and \( \pi_i \). Therefore, plugging (C-16) into (16), we obtain

\[
C_i(\mu_i; \mu_-) = \mu_i^k + b\lambda \left( \frac{2\lambda \mu_i + \mu^2_i + \mu \mu_-}{(\mu_i^2 (\lambda + \mu_-) + \mu^2_i (\lambda + \mu_-))} \right)
\]

\[
= \frac{\mu_i^k}{k} \left( b\lambda \left( \frac{(\mu_i + \mu_-)(\lambda + \mu_-)}{(\mu_i^2 (\lambda + \mu_-) + \mu^2_i (\lambda + \mu_-))} \right) \right) + \frac{2\lambda \mu_i + \mu^2_i + \mu \mu_-}{(\mu_i^2 (\lambda + \mu_-) + \mu^2_i (\lambda + \mu_-))}
\]

and similarly for \( C_i(\pi_i; \mu_-) \).

We define the function

\[
G(\tilde{\mu}_i, \tilde{\mu}_-, k) = \left( \tilde{\mu}_i (\tilde{\mu}_i + \tilde{\mu}_-)(1 + \tilde{\mu}_-) \right) \left( 2 \tilde{\mu}_i - \tilde{\mu}_- \right) + \tilde{\mu}_- (\tilde{\mu}_i + \tilde{\mu}_-)
\]

such that \( C_i(\mu_i; \mu_-) = bG(\tilde{\mu}_i, \tilde{\mu}_-, k) \) and \( C_i(\pi_i; \mu_-) = bG(\pi_i, \pi_-, \lambda, k) \). Hence, \( C_i(\mu_i; \mu_-) \leq b \) if and only if \( G(\tilde{\mu}_i, \tilde{\mu}_-, \lambda, k) \), and similarly for \( \pi_i \).

Because

\[
\frac{\partial G(\tilde{\mu}_i, \tilde{\mu}_-, k)}{\partial k} = -\tilde{\mu}_i (\tilde{\mu}_i + \tilde{\mu}_-)(1 + \tilde{\mu}_-) \left( 2 \tilde{\mu}_i - \tilde{\mu}_- \right) + \tilde{\mu}_- (\tilde{\mu}_i + \tilde{\mu}_-)
\]

and because

\[
\frac{\partial G(\tilde{\mu}_i, \tilde{\mu}_-, k)}{\partial \mu_-} = -\tilde{\mu}_i (1 + k)\tilde{\mu}_i^4 (1 + \tilde{\mu}_-) + \tilde{\mu}_i^3 \tilde{\mu}_- (4 + 3\tilde{\mu}_- + k(2 + 3\tilde{\mu}_-))
\]

\[
= \frac{\tilde{\mu}_i (\tilde{\mu}_i^2 (1 + \tilde{\mu}_-)^2 + \tilde{\mu}_i^2 (1 + \tilde{\mu}_-)^3)}{\tilde{\mu}_i (\tilde{\mu}_i^2 (1 + \tilde{\mu}_-)^2 + \tilde{\mu}_i^2 (1 + \tilde{\mu}_-)^3)}
\]

\[
< 0,
\]

and because \( k < 1.03 \) and \( \mu_- < 2.03\lambda \),

\[
G(\tilde{\mu}_i, \tilde{\mu}_-, k) \geq G(\tilde{\mu}_i, 2.03, 1.03) = \frac{1 + 1.99936 \tilde{\mu}_i + 3.14094 \tilde{\mu}_i^2 + 1.41717 \tilde{\mu}_i^3}{(1 + \tilde{\mu}_i + 0.735276 \tilde{\mu}_i^2)^2}
\]

In particular, \( G(\tilde{\mu}_i, 2.03, 1.03) \geq 1 \) if and only if

\[
\tilde{\mu}_i \left( -0.00642474 + 0.670386 \tilde{\mu}_i - 0.0533813 \tilde{\mu}_i^2 - 0.540631 \tilde{\mu}_i^3 \right) \geq 0.
\]
Because this quartic polynomial has four roots, namely at \(-1.16448, 0, 0.000958438,\) and 1.06478, and because the quartic term is negative, we obtain that, when \(\hat{\mu}_i \geq 0,\) \(G(\hat{\mu}_i, 2.03, 1.03) \geq 1\) if and only if \(\hat{\mu}_i \in [0.000958438, 1.06478].\) Hence, if \(C_i(\mu_i ; \mu_{-i}) \leq b,\) then \(\hat{\mu}_i \notin [0.000958438, 1.06478],\) and similarly for \(\overline{\mu}_i.\)

By Lemma C-14, \(C_i(\mu_i ; \mu_{-i})\) has at most two local minima. Because these two local minima are precisely \(\mu_i\) and \(\overline{\mu}_i,\) there must exist some local maximum in between. Because \(C_i(\mu_i ; \mu_{-i}) \geq b\) for all \(\mu_i \in [0.000958438, 1.06478],\) we obtain that \(\mu_i \leq 0.000958438\lambda\) and \(\overline{\mu}_i \geq 1.06478\lambda.\)

**Lemma C-13.** For any \(k < 1.03\) and \(\mu_{-i} > 2\lambda,\) suppose that both \(\mu_i\) and \(\overline{\mu}_i\) are local minima of \(C_i(\mu_i ; \mu_{-i}),\) defined in (16), with \(0 < \mu_i < \lambda < \overline{\mu}_i.\) Then, \(F(\mu_i , \mu_{-i}) > F(\overline{\mu}_i , \mu_{-i}),\) where \(F(\mu_i , \mu_{-i})\) is defined in (C-10).

**Proof:** The proof uses Lemmas C-15 and C-17. Suppose, to obtain a contradiction, that \(F(\mu_i , \mu_{-i}) \leq F(\overline{\mu}_i , \mu_{-i}).\) Define \(\hat{\mu}_i = \mu_i / \lambda\) and \(\breve{\mu}_{-i} = \mu_{-i} / \lambda.\) Moreover, define \(\tilde{F}(\mu_i, \mu_{-i})\) as identical to \(F(\mu_i, \mu_{-i})\) with \(\lambda = 1.\) Hence, using (C-10),

\[
\begin{align*}
& F(\mu_i , \mu_{-i}) \leq F(\overline{\mu}_i , \mu_{-i}) \\
\iff & \frac{(\mu_i + \mu_{-i})}{\mu_i^4(4\lambda\mu_i + (\mu_i + \mu_{-i})^2)} \leq \frac{(\overline{\mu}_i + \mu_{-i})}{\overline{\mu}_i^4(4\lambda\overline{\mu}_i + (\overline{\mu}_i + \mu_{-i})^2)} \\
\iff & \tilde{F}(\mu_i, \mu_{-i}) \leq \tilde{F}(\overline{\mu}_i, \mu_{-i}).
\end{align*}
\]

Because \(\overline{\mu}_i > 1,\) we obtain from Lemma C-17 that \(\tilde{F}(\mu_i, \mu_{-i}) < \tilde{F}(1, \mu_{-i}).\) Moreover, it can easily be checked that \(\tilde{F}(\mu_i, \mu_{-i})\) is increasing in \(k\) for all \(\mu_i \leq 1\) when \(\mu_{-i} > 2.\) Hence, because \(k \geq 1,\)

\[
\begin{align*}
\frac{(\mu_i + \mu_{-i})}{\mu_i^4(4\mu_i + (\mu_i + \mu_{-i})^2)} & \leq \tilde{F}(\mu_i, \mu_{-i}) \\
& < \tilde{F}(\overline{\mu}_i, \mu_{-i}) < \frac{(1 + \mu_{-i})(2(1 - \mu_{-i}) + \mu_{-i}(1 + \mu_{-i}))}{4\mu_{-i} + (1 + \mu_{-i})^2},
\end{align*}
\]

and therefore, combining the extreme ends of these inequalities,

\[
\begin{align*}
(\mu_i + \mu_{-i}) & \left(2(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_i + \mu_{-i})\right) (4\mu_{-i} + (1 + \mu_{-i})^2) \\
< (1 + \mu_{-i}) & \left(2(1 - \mu_{-i}) + \mu_{-i}(1 + \mu_{-i})\right) \mu_i (4\mu_i + (\mu_i + \mu_{-i})^2),
\end{align*}
\]

which is equivalent to requiring that the following function

\[
K(\mu_i, \mu_{-i}) = -\mu_i^3(1 + \mu_{-i})(2 - \mu_{-i} + \mu_{-i}^2) + \mu_i^2(2 + 9\mu_{-i} + 6\mu_{-i}^2 + \mu_{-i}^3 - 2\mu_{-i}^4) \\
+ \mu_i \mu_{-i} (-8 - 4\mu_{-i} + 11\mu_{-i}^2 - 2\mu_{-i}^3 - \mu_{-i}^4) + \mu_{-i}^2(\mu_{-i} - 2)(1 + 6\mu_{-i} + \mu_{-i}^2), \tag{C-12}
\]

is negative. Hence, we will obtain a contradiction if we can show that \(K(\mu_i, \mu_{-i}) \geq 0\) for all \(0 < \mu_i < \lambda\) and \(\mu_{-i} > 2\lambda.\)
This cubic polynomial $K(\tilde{\mu}, \tilde{\mu}_i)$ has at most three roots, at $\tilde{\mu}_i = 1$, and, if $\tilde{\mu}_i \leq 2.01635$, at

$$\tilde{\mu}_i^L(\tilde{\mu}_i) = \frac{4\tilde{\mu}_i + 3\tilde{\mu}_i^2 - \tilde{\mu}_i^3 - 2\tilde{\mu}_i\sqrt{5 + 12\tilde{\mu}_i - 3\tilde{\mu}_i^2 - 3\tilde{\mu}_i^3 + \tilde{\mu}_i^4 - \tilde{\mu}_i^5}}{2 + \tilde{\mu}_i + \tilde{\mu}_i^3}$$

$$\tilde{\mu}_i^R(\tilde{\mu}_i) = \frac{4\tilde{\mu}_i + 3\tilde{\mu}_i^2 - \tilde{\mu}_i^3 + 2\tilde{\mu}_i\sqrt{5 + 12\tilde{\mu}_i - 3\tilde{\mu}_i^2 - 3\tilde{\mu}_i^3 + \tilde{\mu}_i^4 - \tilde{\mu}_i^5}}{2 + \tilde{\mu}_i + \tilde{\mu}_i^3}.$$  

We next show that $\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i^R(\tilde{\mu}_i) < 1$. Indeed,

$$\tilde{\mu}_i^R(\tilde{\mu}_i) < 1 \iff 4\tilde{\mu}_i + 3\tilde{\mu}_i^2 - \tilde{\mu}_i^3 + 2\tilde{\mu}_i\sqrt{5 + 12\tilde{\mu}_i - 3\tilde{\mu}_i^2 - 3\tilde{\mu}_i^3 + \tilde{\mu}_i^4 - \tilde{\mu}_i^5} < 2 + \tilde{\mu}_i + \tilde{\mu}_i^3$$

$$\iff 2\tilde{\mu}_i\sqrt{5 + 12\tilde{\mu}_i - 3\tilde{\mu}_i^2 - 3\tilde{\mu}_i^3 + \tilde{\mu}_i^4 - \tilde{\mu}_i^5} < 2 - 3\tilde{\mu}_i - 3\tilde{\mu}_i^2 + \tilde{\mu}_i^3 + \tilde{\mu}_i^4$$

$$\iff 4\tilde{\mu}_i^2 (5 + 12\tilde{\mu}_i - 3\tilde{\mu}_i^2 - 3\tilde{\mu}_i^3 + \tilde{\mu}_i^4 - \tilde{\mu}_i^5) < (2 - 3\tilde{\mu}_i - 3\tilde{\mu}_i^2 + \tilde{\mu}_i^3 + \tilde{\mu}_i^4)^2$$

$$\iff -4 + 12\tilde{\mu}_i + 23\tilde{\mu}_i^2 + 26\tilde{\mu}_i^3 + 5\tilde{\mu}_i^4 + 9\tilde{\mu}_i^5 - 6\tilde{\mu}_i^7 - \tilde{\mu}_i^8 < 0.$$  

Because $\tilde{\mu}_i > 2$, we obtain

$$-4 + 12\tilde{\mu}_i + 23\tilde{\mu}_i^2 + 26\tilde{\mu}_i^3 + 5\tilde{\mu}_i^4 + 9\tilde{\mu}_i^5 - 6\tilde{\mu}_i^7 - \tilde{\mu}_i^8 < 0.$$  

Hence, $\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i^R(\tilde{\mu}_i) < 1$. Because the cubic term in $\tilde{\mu}_i$ in the expression $K(\tilde{\mu}, \tilde{\mu}_i)$ has a negative coefficient, $K(\tilde{\mu}, \tilde{\mu}_i) < 0$ for all $\mu_i > 1$ if $\tilde{\mu}_i > 2.01635$, and for all $\tilde{\mu}_i \in (\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i^R(\tilde{\mu}_i)) \cup (1, \infty)$ if $2 < \tilde{\mu}_i < 2.01635$.

By assumption, $\tilde{\mu} \leq 1$, so we can have $K(\tilde{\mu}, \tilde{\mu}_i) < 0$ only if $2 < \tilde{\mu}_i \leq 2.01635$, which we assume in the sequel. Under that condition, $K(\tilde{\mu}_i, \tilde{\mu}_i) < 0$ if $\tilde{\mu}_i \in (\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i^R(\tilde{\mu}_i)) \cup (1, \infty)$.

We next consider the function $H(\tilde{\mu}_i, \tilde{\mu}_i, k)$, which was defined in (C-15) in the proof of Lemma C-15, but extend its set of arguments to include $k$. It can easily be checked that, when $\tilde{\mu}_i > 2$, $H(\tilde{\mu}_i, \tilde{\mu}_i, k)$ is increasing in $k$. Similar to the proof of Lemma C-15, $H(\tilde{\mu}_i, \tilde{\mu}_i, k) > 0$ at every strict local minimum of $C_i(\tilde{\mu}_i; \tilde{\mu}_i)$ and $H(\tilde{\mu}_i, \tilde{\mu}_i, k) < 0$ at every strict local maximum. The proof of Lemma C-15 shows that, when $\tilde{\mu}_i > 2$, $H(\tilde{\mu}_i, \tilde{\mu}_i, k)$ is either decreasing and then increasing, or always increasing. Because $\tilde{\mu}_i$ and $\tilde{\mu}_i$ are local minima of $C_i(\tilde{\mu}_i; \tilde{\mu}_i)$, there must be some local maximum lying between these two minima; hence, $H(\tilde{\mu}_i, \tilde{\mu}_i, k)$ must cross zero twice, first from above and then from below, and the first crossing happens after $\tilde{\mu}_i$ and the second crossing happens before $\tilde{\mu}_i$. In particular, $H(\tilde{\mu}_i, \tilde{\mu}_i, k)$ is decreasing in $\tilde{\mu}_i$ for all $\tilde{\mu}_i \leq \tilde{\mu}_i$.

Because $\tilde{\mu}_i^L(\tilde{\mu}_i) \leq \tilde{\mu}_i$, and because $k \leq 1.03$, we must thus have

$$0 \leq H(\tilde{\mu}_i, \tilde{\mu}_i, k) \leq H(\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i, k) \leq H(\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i, 1.03).$$

It can be numerically checked that the univariate function $H(\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i, 1.03)$ is nonpositive and decreasing in $\tilde{\mu}_i$ when $\tilde{\mu}_i \in [2, 2.01635]$. Hence, $0 \leq H(\tilde{\mu}_i^L(\tilde{\mu}_i), \tilde{\mu}_i, 1.03) < H(\tilde{\mu}_i^L(2), 2, 1.03) = 0$, a contradiction. □
Lemma C-14. The function $C_i(\mu_i; \mu_{-i})$ defined in (16) has at most two local minima, and they are either equal to zero or greater than $2\lambda - \mu_{-i}$. Moreover, these minima are strict.

Proof: The proof uses Lemma C-15. First, note that, for any $\mu_{-i}$, $\lim_{\mu_i \to \infty} C_i(\mu_i; \mu_{-i}) = \infty$, and therefore all minima are equal to zero or interior.

Suppose that $\mu_{-i} > 2\lambda$. Then $C_i(\mu_i; \mu_{-i}) = \hat{C}_i(\mu_i; \mu_{-i})$ for all $\mu_i \geq 0$, in which $\hat{C}_i(\mu_i; \mu_{-i})$ is defined by (C-13). By Lemma C-15, $C_i(\mu_i; \mu_{-i})$ has at most two interior local minima and at most one when $k = 1$, and these minima are strict. From (C-8), we obtain that, when $k > 1$,

$$C_i'(0; \mu_{-i}) < 0 \iff k \mu_i^{k-1} \bigg|_{\mu_i = 0} - b \frac{(\lambda + \mu_{-i})(\mu_{-i} - 2\lambda)}{\lambda \mu_{-i}^2} < 0,$$

i.e., the function $C_i(\mu_i; \mu_{-i})$ is decreasing at $\mu_i = 0$; hence, zero is a strict local maximum and therefore, all minima are interior and greater than $2\lambda - \mu_{-i}$.

Suppose next that $\mu_{-i} \leq 2\lambda$. By (16), $C_i(\mu_i; \mu_{-i}) = \mu_i^k + b$ for all $\mu_i \in [0, 2\lambda - \mu_{-i}]$, and $C_i(\mu_i; \mu_{-i}) = \hat{C}_i(\mu_i; \mu_{-i})$ for all $\mu_i > 2\lambda - \mu_{-i}$. Note that $C_i(\mu_i; \mu_{-i})$ is continuous at $2\lambda - \mu_{-i}$ and increasing for all $\mu_i \in [0, 2\lambda - \mu_{-i}]$; in particular, zero is a strict local minimum. However, $C_i(\mu_i; \mu_{-i})$ is non-differentiable at $2\lambda - \mu_{-i}$ since

$$\lim_{\mu_i \to 2\lambda - \mu_{-i}} C_i'(\mu_i; \mu_{-i}) = k (2\lambda - \mu_{-i})^{k-1}$$

$$\geq \lim_{\mu_i \to 2\lambda - \mu_{-i}} C_i'(\mu_i; \mu_{-i}) = k (2\lambda - \mu_{-i})^{k-1} - \frac{4b \lambda^3 (\lambda + \mu_{-i})(2\lambda - \mu_{-i})}{((2\lambda - \mu_{-i})^2(\lambda + \mu_{-i}) + \mu_i^2(3\lambda - \mu_{-i}))^2}.$$

Note that $\lim_{\mu_i \to 2\lambda - \mu_{-i}} C_i'(\mu_i; \mu_{-i}) \geq 0$.

We consider two cases: If $\lim_{\mu_i \to 2\lambda - \mu_{-i}} C_i'(\mu_i; \mu_{-i}) \geq 0$, then $C_i(\mu_i; \mu_{-i})$ is non-decreasing at $2\lambda - \mu_{-i}$.

Because $\hat{C}_i(\mu_i; \mu_{-i})$ has at most one local maximum and one local minimum when $\mu_i > 2\lambda - \mu_{-i}$ by Lemma C-15, $C_i(\mu_i; \mu_{-i})$ has at most two local minima, one at zero and another one when $\mu_i > 2\lambda - \mu_{-i}$. Moreover by Lemma C-15, all minima are strict.

If on the other hand $\lim_{\mu_i \to 2\lambda - \mu_{-i}} C_i'(\mu_i; \mu_{-i}) < 0$, then $C_i(\mu_i; \mu_{-i})$ reaches a local maximum at $2\lambda - \mu_{-i}$. However, if $\lim_{\mu_i \to 2\lambda - \mu_{-i}} C_i'(\mu_i; \mu_{-i}) < 0$, then, by Lemma C-15, it must be that $\hat{C}_i(\mu_i; \mu_{-i})$ has only one local optimum greater than $2\lambda - \mu_{-i}$, and that local optimum is a local minimum. In that case, $C_i(\mu_i; \mu_{-i})$ has two local minima, one at zero and another one when $\mu_i > 2\lambda - \mu_{-i}$, and by Lemma C-15, both minima are strict. \qed

Lemma C-15. The function

$$\hat{C}_i(\mu_i; \mu_{-i}) = \mu_i^k + b \lambda \frac{2\mu_i + \mu_i^2 + \mu_i \mu_{-i}}{(\mu_i^2 (\lambda + \mu_{-i}) + \mu_i^2 (\lambda + \mu_{-i}))} \tag{C-13}$$

defined over all $\mu_i \geq 0$, has (i) at most three interior local optima, namely, a local minimum, a local maximum, and a local minimum as $\mu_i$ increases when $\mu_{-i} > 2\lambda$ and $k > 1$, and (ii) at most two interior local optima, namely, a local maximum and a local minimum as $\mu_i$ increases when either $\mu_{-i} \leq 2\lambda$ or $k = 1$. Moreover, all these optima are strict.
Proof: The proof uses Lemma C-16. Define \( \hat{\mu}_i \equiv \mu_i / \lambda \) for \( i = 1, 2 \). Differentiating (C-13), we obtain:

\[
\hat{C}'(\mu_i; \mu_{-i}) = k \lambda^{i-1} \hat{\mu}_i^{i-1} - b \lambda^{-1} \frac{(\hat{\mu}_1 + \hat{\mu}_2)(1 + \hat{\mu}_1)(2(\hat{\mu}_1 - \hat{\mu}_{-i}) + \hat{\mu}_{-i}(\hat{\mu}_1 + \hat{\mu}_2))}{(\hat{\mu}_1^2 (1 + \hat{\mu}_2) + \hat{\mu}_2^2 (1 + \hat{\mu}_1))^2}, \quad (C-14)
\]

Suppose that \( \hat{C}'(\mu_i; \mu_{-i}) = 0 \). Taking the second derivative, we obtain

\[
\hat{C}''(\mu_i; \mu_{-i}) \bigg|_{\hat{C}'(\mu_i; \mu_{-i})=0} = \left( k(k-1) \lambda^{i-1} \hat{\mu}_i^{i-2} + 2b \lambda^{-1} (1 + \hat{\mu}_{-i}) \right. \times \left( (\hat{\mu}_{-i} - 3) \hat{\mu}_i^4 + 3 \hat{\mu}_i^2 \hat{\mu}_{-i}^2 (1 + \hat{\mu}_{-i}) + 3 \hat{\mu}_i (\hat{\mu}_{-i} - 2) \hat{\mu}_{-i}^2 (1 + \hat{\mu}_{-i}) + \hat{\mu}_i^3 (1 + \hat{\mu}_{-i})(2 + \hat{\mu}_{-i}) \right) \frac{(\hat{\mu}_i^2 (1 + \hat{\mu}_1) + \hat{\mu}_1^2 (1 + \hat{\mu}_2))^2}{(\hat{\mu}_1^2 (1 + \hat{\mu}_2) + \hat{\mu}_2^2 (1 + \hat{\mu}_1))^3} \bigg) \left( \hat{\mu}_1 + \hat{\mu}_2 \right)(1 + \hat{\mu}_1)(2(\hat{\mu}_1 - \hat{\mu}_{-i}) + \hat{\mu}_{-i}(\hat{\mu}_1 + \hat{\mu}_2)) + \frac{2b \lambda^{-1} (1 + \hat{\mu}_{-i})}{(\hat{\mu}_1^2 (1 + \hat{\mu}_2) + \hat{\mu}_2^2 (1 + \hat{\mu}_1))^3} \times \left( (k-1)(\hat{\mu}_{-i} - 2) \hat{\mu}_i^4 + (1 + k) \hat{\mu}_i^4 (1 + \hat{\mu}_{-i})(2 + \hat{\mu}_{-i}) + \hat{\mu}_i \hat{\mu}_{-i}^3 (\hat{\mu}_{-i}(1 + k) - 6) + \hat{\mu}_i^3 \hat{\mu}_{-i}^2 (3 + 2 \hat{\mu}_{-i} + k(4 + 3 \hat{\mu}_{-i})) + 3 \hat{\mu}_i^2 \hat{\mu}_{-i}^2 (\hat{\mu}_{-i}(k + 1) - 2) - 4) \right) \bigg |_{\hat{\mu}_1 + \hat{\mu}_2 = \lambda \hat{\mu}_i (1 + \hat{\mu}_i)}.
\]

Hence, a stationary point \((\hat{\mu}_1, \hat{\mu}_{-i})\) is a strict local minimum if

\[
H(\hat{\mu}_i, \hat{\mu}_{-i}) = (k - 1)(\hat{\mu}_{-i} - 2) \hat{\mu}_i^4 + (1 + k) \hat{\mu}_i^4 (1 + \hat{\mu}_{-i})(2 + \hat{\mu}_{-i}) + \hat{\mu}_i \hat{\mu}_{-i}^3 (\hat{\mu}_{-i}(1 + k) - 6) + \hat{\mu}_i^3 \hat{\mu}_{-i}^2 (3 + 2 \hat{\mu}_{-i} + k(4 + 3 \hat{\mu}_{-i})) + 3 \hat{\mu}_i^2 \hat{\mu}_{-i}^2 (\hat{\mu}_{-i}(k + 1) - 2) - 4)
\]

is positive, and it is a strict local maximum if \(H(\hat{\mu}_i, \hat{\mu}_{-i}) < 0\).

Define \(G(\hat{\mu}_i, \hat{\mu}_{-i}) \equiv \partial H(\hat{\mu}_i, \hat{\mu}_{-i}) / \partial \hat{\mu}_i \). It can be checked that \(G(\hat{\mu}_i, \hat{\mu}_{-i})\) is convex, i.e., that \(\partial^2 G(\hat{\mu}_i, \hat{\mu}_{-i}) / \partial \hat{\mu}_i^2 \geq 0\) for all \(\mu_1, \mu_2 \geq 0\), with equality only if \(\mu_1 = \mu_2 = 0\).

The rest of the proof considers two cases separately, namely, when \(\hat{\mu}_{-i} > 2\) or not. Consider first the case where \(\hat{\mu}_{-i} > 2\). In that case,

\[
\frac{\partial G(\hat{\mu}_i, \hat{\mu}_{-i})}{\partial \hat{\mu}_i} \bigg|_{\hat{\mu}_i = 0} = 6 \hat{\mu}_{-i}^2 (\hat{\mu}_{-i}^2 (k + 1) - 2 \hat{\mu}_{-i} - 4) + 6 \hat{\mu}_{-i}^2 (2 \hat{\mu}_{-i}^2 - 2 \hat{\mu}_{-i} - 4) = 12 \hat{\mu}_{-i}^2 (1 + \hat{\mu}_{-i}) (\hat{\mu}_{-i} - 2) > 0.
\]

Because \(G(\hat{\mu}_i, \hat{\mu}_{-i})\) is convex in \(\hat{\mu}_i\), \(\partial G(\hat{\mu}_i, \hat{\mu}_{-i}) / \partial \hat{\mu}_i > 0\) for all \(\hat{\mu}_i \geq 0\). Hence, when \(\hat{\mu}_i > 0\), \(G(\hat{\mu}_i, \hat{\mu}_{-i})\) crosses zero at most once as \(\hat{\mu}_i\) increases, and the crossing is from below. Hence, when \(\hat{\mu}_{-i} > 2\), because \(G(\hat{\mu}_i, \hat{\mu}_{-i}) \equiv \partial H(\hat{\mu}_i, \hat{\mu}_{-i}) / \partial \hat{\mu}_i\), we thus obtain that \(H(\hat{\mu}_i, \hat{\mu}_{-i})\) is either first decreasing and then increasing in \(\hat{\mu}_i\) or monotone in \(\hat{\mu}_i\), starting from zero.

Suppose first that \(k > 1\). Because \(H(0, \hat{\mu}_{-i}) = (k - 1)(\hat{\mu}_{-i} - 2) \hat{\mu}_{-i}^4 > 0\) and \(\lim_{\hat{\mu}_{-i} \to \infty} H(\hat{\mu}_i, \hat{\mu}_{-i}) = \infty\), we obtain that \(H(\hat{\mu}_i, \hat{\mu}_{-i})\) crosses zero either twice or never, and if it does, it first crosses zero from above and then from below. Therefore, by Lemma C-16, \(\hat{C}'(\lambda \hat{\mu}_i; \lambda \hat{\mu}_{-i})\) may have up to three interior local optima, going through a local minimum, a local maximum, and then a local minimum as \(\hat{\mu}_i\) increases, and all optima are strict.

Suppose next that \(k = 1\). In that case, because \(H(0, \hat{\mu}_{-i}) = (k - 1)(\hat{\mu}_{-i} - 2) \hat{\mu}_{-i}^4 = 0\) and \(\lim_{\hat{\mu}_{-i} \to \infty} H(\hat{\mu}_i, \hat{\mu}_{-i}) = \infty\), \(H(\hat{\mu}_i, \hat{\mu}_{-i})\) crosses zero at most once when \(\hat{\mu}_i > 0\), and the crossing is from below,
and therefore by Lemma C-16,  \( \hat{C}_i(\lambda \hat{\mu}_i; \lambda \hat{\mu}_{-i}) \) may have up to two interior local optima when \( \hat{\mu}_i > 0 \), going through a local maximum and then a local minimum as \( \hat{\mu}_i \) increases, and all optima are strict.

Consider next the case where \( \hat{\mu}_{-i} \leq 2 \). Because \( (\hat{\mu}_{-i}^2 - 4\hat{\mu}_{-i} - 4) < 0 \) and \( (\hat{\mu}_{-i}^2 - \hat{\mu}_{-i} - 2) \leq 0 \) when \( \hat{\mu}_{-i} \leq 2 \), we obtain

\[
\frac{\partial G(\hat{\mu}_i, \hat{\mu}_{-i})}{\partial \hat{\mu}_i} \bigg|_{\hat{\mu}_i = \hat{\mu}_{-i} - \frac{2 - \hat{\mu}_{-i}}{2 + \hat{\mu}_{-i}}} = -\frac{12\hat{\mu}_i^2(2\hat{\mu}_{-i} + \hat{\mu}_{-i}^2 + k(\hat{\mu}_{-i}^2 - 4\hat{\mu}_{-i} - 4))}{2 + \hat{\mu}_{-i}}
\leq -\frac{12\hat{\mu}_i^2(2\hat{\mu}_{-i} + \hat{\mu}_{-i}^2 + (\hat{\mu}_{-i}^2 - 4\hat{\mu}_{-i} - 4))}{2 + \hat{\mu}_{-i}}
\leq \frac{24\hat{\mu}_i^2}{2 + \hat{\mu}_{-i}} \geq 0.
\]

Because \( G(\hat{\mu}_i, \hat{\mu}_{-i}) \) is strictly convex in \( \hat{\mu}_i > 0 \), this implies that \( G(\hat{\mu}_i, \hat{\mu}_{-i}) \) is increasing in \( \hat{\mu}_i \) for all \( \hat{\mu}_i > \hat{\mu}_{-i} - \frac{2 - \hat{\mu}_{-i}}{2 + \hat{\mu}_{-i}} \). Hence when \( \hat{\mu}_i \geq \hat{\mu}_{-i} - \frac{2 - \hat{\mu}_{-i}}{2 + \hat{\mu}_{-i}}, G(\hat{\mu}_i, \hat{\mu}_{-i}) \) crosses zero at most once as \( \hat{\mu}_i \) increases, and the crossing is from below. Hence, when \( \hat{\mu}_{-i} \leq 2 \), because \( G(\hat{\mu}_i, \hat{\mu}_{-i}) = \partial H(\hat{\mu}_i, \hat{\mu}_{-i})/\partial \hat{\mu}_i \), we obtain that \( H(\hat{\mu}_i, \hat{\mu}_{-i}) \) is either first decreasing and then increasing in \( \hat{\mu}_i \) or monotone in \( \hat{\mu}_i \), starting from \( \hat{\mu}_{-i} - \frac{2 - \hat{\mu}_{-i}}{2 + \hat{\mu}_{-i}} \). Because

\[
H(\hat{\mu}_i, \hat{\mu}_{-i}) = \frac{8\hat{\mu}_i^4(\hat{\mu}_{-i} - 2)(\hat{\mu}_{-i} - 2(1 - \sqrt{2}))(\hat{\mu}_{-i} - 2(1 + \sqrt{2}))}{(2 + \hat{\mu}_{-i})^3} \leq 0
\]

and \( \lim_{\hat{\mu}_i \to \infty} H(\hat{\mu}_i, \hat{\mu}_{-i}) = \infty \), we obtain that \( H(\hat{\mu}_i, \hat{\mu}_{-i}) \) crosses zero only once when \( \hat{\mu}_i \geq \hat{\mu}_{-i} - \frac{2 - \hat{\mu}_{-i}}{2 + \hat{\mu}_{-i}} \) and the crossing is from below. Therefore by Lemma C-16, when \( \hat{\mu}_i \geq \hat{\mu}_{-i} - \frac{2 - \hat{\mu}_{-i}}{2 + \hat{\mu}_{-i}}, \hat{C}_i(\lambda \hat{\mu}_i; \lambda \hat{\mu}_{-i}) \) may have up to two interior local optima, going through a local maximum and then a local minimum as \( \hat{\mu}_i \) increases, and all optima are strict. Finally, observe from (C-14) that, for all \( \hat{\mu}_i < \hat{\mu}_{-i} - \frac{2 - \hat{\mu}_{-i}}{2 + \hat{\mu}_{-i}}, \hat{C}_i(\mu_i; \mu_{-i}) > 0 \). Hence, when \( \hat{\mu}_i > 0, \hat{C}_i(\lambda \hat{\mu}_i; \lambda \hat{\mu}_{-i}) \) may have up to two interior local optima, going through a local maximum and then a local minimum as \( \hat{\mu}_i \) increases, and all these optima are strict. □

**Lemma C-16.** Let \( f(x) \) be a twice continuously differentiable function, and suppose that there exists a value \( \hat{x} \) such that \( f''(x) < \langle 0 \rangle \) for all \( x < \hat{x} \) such that \( f'(x) = 0 \) and \( f''(x) > \langle 0 \rangle \) for all \( x > \hat{x} \) such that \( f'(x) = 0 \). Then \( f(x) \) has at most two interior local optima, and both optima are strict.

**Proof.** We consider the case where \( f'(x) < 0 \) for all \( x < \hat{x} \) such that \( f'(x) = 0 \) and \( f''(x) > 0 \) for all \( x > \hat{x} \) such that \( f'(x) = 0 \); the other case is symmetric can be treated similarly.

First, we show that

\[
\forall x_1 < \hat{x} \text{ such that } f'(x_1) = 0 \Rightarrow f'(x) > 0, \forall x < x_1 \text{ and } f'(x) < 0, \forall x \in (x_1, \hat{x}). \tag{C-16}
\]

Suppose that there exists an \( x_1 < \hat{x} \) such that \( f'(x_1) = 0 \). To obtain a contradiction, suppose that there exists another point, \( x_2 \in (x_1, \hat{x}) \), such that \( f'(x_2) = 0 \). If there are more than one such point, we consider the smallest such point \( x_2 \) that is greater than \( x_1 \). Then, because \( f''(x_1) < 0 \) and \( f'(x) \) is continuous, we must have that \( f''(x_2) \geq 0 \), a contradiction. Similarly, we can obtain that

\[
\forall x_2 > \hat{x} \text{ such that } f'(x_2) = 0 \Rightarrow f'(x) < 0, \forall x \in (\hat{x}, x_2) \text{ and } f'(x) > 0, \forall x > x_2. \tag{C-17}
\]

Suppose first that there exists an \( x_1 < \hat{x} \) such that \( f'(x_1) = 0 \). By (C-16), \( x_1 \) is a strict local maximum.

We consider the three following cases.
• If \( f'(x) > 0 \) for all \( x > \hat{x} \), then because \( f'(x) < 0 \) for all \( x \in (x_1, \hat{x}) \) and \( f'(x) \) is continuous, we must have that \( f'(\hat{x}) = 0 \). Hence, \( x_1 \) and \( \hat{x} \) are the only stationary points. Because \( f'(x) < 0 \) for all \( x \in (x_1, \hat{x}) \) and \( f'(x) > 0 \) for all \( x > \hat{x} \), \( \hat{x} \) is a strict local minimum.

• If \( f'(x) < 0 \) for all \( x > \hat{x} \), then because \( f'(x) < 0 \) for all \( x \in (x_1, \hat{x}) \) and \( f'(x) \) is continuous, we must have that \( f'(\hat{x}) \leq 0 \). Hence, \( f(x) \) is nonincreasing for all \( x > x_1 \) and \( x_1 \) is the only local optimum.

• If there exists some \( x_2 > \hat{x} \) such that \( f'(x_2) = 0 \), then \( f'(x) < 0 \) for all \( x \in (\hat{x}, x_2) \) and \( f'(x) > 0 \) for all \( x > x_2 \). Because \( f'(x) \) is continuous, we obtain that \( f'(\hat{x}) \leq 0 \). Hence, \( f(x) \) is nonincreasing for all \( x \in (x_1, x_2) \), and \( x_1 \) and \( x_2 \) are the only optima. By (C-17), \( x_2 \) is a strict local minimum.

Suppose next that \( f'(x) > 0 \) for all \( x < \hat{x} \).

• If \( f'(x) > 0 \) for all \( x > \hat{x} \), then because \( f'(x) \) is continuous, we must have that \( f'(\hat{x}) \geq 0 \). Hence, \( f(x) \) is always nondecreasing, and there is no interior local optimum.

• If \( f'(x) < 0 \) for all \( x > \hat{x} \), then because \( f'(x) \) is continuous, we must have that \( f'(\hat{x}) = 0 \). Hence, \( \hat{x} \) is the only stationary point. Because \( f'(x) > 0 \) for all \( x < \hat{x} \) and \( f'(x) < 0 \) for all \( x > \hat{x} \), \( \hat{x} \) is a strict local maximum.

• If there exists some \( x_2 > \hat{x} \) such that \( f'(x_2) = 0 \), then \( f'(x) < 0 \) for all \( x \in (\hat{x}, x_2) \) and \( f'(x) > 0 \) for all \( x > x_2 \). Because \( f'(x) \) is continuous, we obtain that \( f'(\hat{x}) = 0 \). Hence, \( \hat{x} \) and \( x_2 \) are the only stationary points. Because \( f'(x) > 0 \) for all \( x < \hat{x} \) and \( f'(x) < 0 \) for all \( x > \hat{x} \), \( \hat{x} \) is a strict local maximum. By (C-17), \( x_2 \) is a strict local minimum.

Finally, suppose that \( f'(x) < 0 \) for all \( x < \hat{x} \). This case is symmetric to the case where \( f'(x) > 0 \) for all \( x < \hat{x} \). and can be treated similarly. Their argument is omitted for brevity. □

**Lemma C-17.** For any \( \mu_{-i} > 2\lambda \), the function \( F(\mu_i, \mu_{-i}) \) defined in (C-10) is decreasing in \( \mu_i \) for all \( \mu_i \geq \lambda \).

**Proof:** Let \( G(\mu_i, \mu_{-i}) \) defined as in (C-11). Therefore \( F(\mu_i, \mu_{-i}) \) is decreasing in \( \mu_i \) if and only if \( G(\mu_i, \mu_{-i}) < 0 \).

Because \( k \geq 1 \) and \( \mu_{-i} > 2\lambda \), we have, for all \( \mu_i \geq \lambda \),

\[
\frac{\partial^2 G(\mu_i, \mu_{-i})}{\partial \mu_i^2} = -4(-2\lambda \mu_{-i}(4\lambda + 3\mu_i + 4\mu_{-i}) + \mu_i (4\lambda^2 \mu_{-i} + 3\mu_{-i}(\mu_i + \mu_{-i})^2 + 2\lambda(3\mu_i^2 + 3\mu_{-i} + \mu_{-i}^2))) \\
\leq -4(-2\lambda \mu_{-i}(4\lambda + 3\mu_i + 4\mu_{-i}) + (4\lambda^2 \mu_{-i} + 3\mu_{-i}(\mu_i + \mu_{-i})^2 + 2\lambda(3\mu_i^2 + 3\mu_{-i} + \mu_{-i}^2))) \\
= -4\mu_i^2(6\lambda + 3\mu_{-i}) - 24\mu_i \mu_{-i}^2 + 4(4\lambda^2 \mu_{-i} + 6\lambda \mu_{-i}^2 - 3\mu_i^3) \\
\leq -4\lambda^2(6\lambda + 3\mu_{-i}) - 24\mu_i \mu_{-i}^2 + 4(4\lambda^2 \mu_{-i} + 6\lambda \mu_{-i}^2 - 3\mu_i^3) \\
= 4(-6\lambda^3 + \lambda^2 \mu_{-i} - 3\mu_i^3).
\]

Let \( K(\mu_{-i}) = (-6\lambda^3 + \lambda^2 \mu_{-i} - 3\mu_i^3). \) Because \( K(2\lambda) = -28\lambda^3 < 0 \) and, for all \( \mu_{-i} > 2\lambda \), \( K'(\mu_{-i}) = \lambda^2 - 9\mu_{-i}^2 < -35\lambda^2 < 0 \), we obtain that \( K(\mu_{-i}) < 0 \) for all \( \mu_{-i} > 2\lambda \), and therefore \( G(\mu_i, \mu_{-i}) \) is strictly concave in \( \mu_i \) for all \( \mu_i \geq \lambda \) and \( \mu_{-i} > 2\lambda \).

Next, we obtain, using the facts that \( k \geq 1 \) and \( \mu_{-i} > 2\lambda \),

\[
\frac{\partial G(\lambda, \mu_{-i})}{\partial \mu_i} = 4\mu_i \lambda(11\lambda^2 + 8\lambda \mu_{-i} + 3\mu_i^2) - 4\mu_{-i} \lambda(2\lambda^4 + 8\lambda^3 \mu_{-i} + 5\lambda^2 \mu_{-i}^2 + 4\lambda \mu_{-i}^3 + \mu_{-i}^4) \\
\leq 4\mu_i \lambda(11\lambda^2 + 8\lambda \mu_{-i} + 3\mu_i^2) - 4(2\lambda^4 + 8\lambda^3 \mu_{-i} + 5\lambda^2 \mu_{-i}^2 + 4\lambda \mu_{-i}^3 + \mu_{-i}^4) \\
\leq 4\mu_i \lambda(11\lambda^2 + 8\lambda \mu_{-i} + 3\mu_i^2) - 4\mu_i \lambda(2\lambda^4 + 8\lambda^3 \mu_{-i} + 5\lambda^2 \mu_{-i}^2 + 4\lambda \mu_{-i}^3 + \mu_{-i}^4).
\]
Because $G(\mu_i, \mu_{-i})$ is strictly concave in $\mu_i$ for all $\mu_i \geq \lambda$ and $\mu_{-i} > 2\lambda$, this implies that $G(\mu_i, \mu_{-i})$ is decreasing in $\mu_i$ for all $\mu_i \geq \lambda$ and $\mu_{-i} > 2\lambda$.

Finally, using the facts that $k \geq 1$ and $\mu_{-i} > 2\lambda$, observe that

$$G(\lambda, \mu_{-i}) = 4\lambda^2 \mu_{-i}(5\lambda^2 + 4\lambda \mu_{-i} + 3\mu_{-i}^2) - k(\lambda + \mu_{-i})(2\lambda^2 - \lambda \mu_{-i} + \mu_{-i}^2)(\lambda^2 + 6\lambda \mu_{-i} + \mu_{-i}^2)$$

$$\leq 4\lambda^2 \mu_{-i}(5\lambda^2 + 4\lambda \mu_{-i} + 3\mu_{-i}^2) - (\lambda + \mu_{-i})(2\lambda^2 - \lambda \mu_{-i} + \mu_{-i}^2)(\lambda^2 + 6\lambda \mu_{-i} + \mu_{-i}^2)$$

$$= -2\lambda^5 + 7\lambda^4 \mu_{-i} + 8\lambda^3 \mu_{-i}^2 + 10\lambda^2 \mu_{-i}^3 - 6\lambda \mu_{-i}^4 - \mu_{-i}^5$$

$$\leq -2\lambda^5 + 7\lambda^4 \mu_{-i} + 8\lambda^3 \mu_{-i}^2 + 10\lambda^2 \mu_{-i}^3 - 8\lambda \mu_{-i}^4$$

$$\leq -2\lambda^5 + 7\lambda^4 \mu_{-i} - 4\lambda^3 \mu_{-i}^2$$

$$\leq -2\lambda^5 - \lambda^4 \mu_{-i} < 0.$$

Because $G(\mu_i, \mu_{-i})$ is decreasing in $\mu_i$ for all $\mu_i \geq \lambda$, this implies that $G(\mu_i, \mu_{-i}) < 0$ for all $\mu_i \geq \lambda$ and $\mu_{-i} > 2\lambda$. □

**Lemma C-18.** For any $\mu_i \geq 0$ and $\mu_{-i} > 2\lambda$, the function

$$G(\mu_i, \mu_{-i}) = 4\lambda^2 \mu_{-i}(4\lambda \mu_i + (\mu_i + \mu_{-i})(\mu_i + 3\mu_{-i}))$$

$$- k(\mu_i + \mu_{-i})(2\lambda(\mu_i - \mu_{-i}) + \mu_{-i}(\mu_i + \mu_{-i})) \left(4\lambda \mu_{-i} + (\mu_i + \mu_{-i})^2\right)$$

is decreasing in $\mu_{-i}$.

**Proof:** We first show that $G(\mu_i, \mu_{-i})$ is concave in $\mu_{-i}$ when $\mu_i \geq 0$ and $\mu_{-i} \geq 2\lambda$:

$$\frac{\partial^2 G(\mu_i, \mu_{-i})}{\partial \mu_{-i}^2} = -8k\mu_i^3 - 4(2k\lambda + 9k \mu_{-i} - 8\lambda)\mu_i^2 - 24\mu_{-i}(2k \mu_{-i} + k \lambda - 3\lambda)\mu_i - 4k \mu_{-i}(5\mu_i^2 + 6\mu_{-i} - 12\lambda^2)$$

$$\leq -8\mu_i^3 - 4(2k\lambda + 18\lambda - 8\lambda)\mu_i^2 - 24\mu_{-i}(4\lambda + \lambda - 3\lambda)\mu_i - 4k \mu_{-i}(20\lambda^2 + 12\lambda^2 - 12\lambda^2)$$

$$= -8\mu_i^3 - 48\lambda \mu_i^2 - 48\lambda \mu_{-i} \mu_i - 80k \mu_{-i} \lambda^2$$

$$\leq 0.$$

We next show that $G(\mu_i, \mu_{-i})$ is decreasing in $\mu_{-i}$ at $\mu_{-i} = 2\lambda$ when $\mu_i \geq 0$:

$$\frac{\partial G(\mu_i, 2\lambda)}{\partial \mu_{-i}} = 4\lambda \mu_i(2\lambda + \mu_i)(18\lambda + \mu_i) - k(48\lambda^4 + 176\lambda^3 \mu_i + 96\lambda^2 \mu_i^2 + 20\lambda \mu_i^3 + \mu_i^4)$$

$$\leq 4\lambda \mu_i(2\lambda + \mu_i)(18\lambda + \mu_i) - (48\lambda^4 + 176\lambda^3 \mu_i + 96\lambda^2 \mu_i^2 + 20\lambda \mu_i^3 + \mu_i^4)$$

$$= -48\lambda^4 - 32\lambda^3 \mu_i - 16\lambda^2 \mu_i^2 - 16\lambda \mu_i^3 - \mu_i^4$$

$$< 0.$$

Because $G(\mu_i, \mu_{-i})$ is concave in $\mu_{-i}$, we obtain that $G(\mu_i, \mu_{-i})$ is decreasing in $\mu_{-i}$ for all $\mu_{-i} \geq 2\lambda$. □
Proof of Proposition 8. Proposition 6 shows that the optimal capacity investments in the dedicated systems are equal to \( \mu_1 = \mu_2 = 0 \) if \( b \leq \lambda k^{-k}(1 + k)^{1+k} \) and to \( \mu_1 = \mu_2 = \mu^* \) if \( b \geq \lambda k^{-k}(1 + k)^{1+k} \), and in case \( b = \lambda k^{-k}(1 + k)^{1+k} \), we set by convention \( \mu_{\text{dedicated}} = \mu^* \). Moreover, the discussion in §5.1.3 shows that \( \mu_{\text{centralized}} = \mu_{\text{dedicated}} \). Similarly, Proposition 7 shows that the symmetric equilibrium in the pooled configuration is either \( \mu_1 = \mu_2 = 0 \) or \( \mu_1 = \mu_2 = \mu^* \), where \( \mu^* \) is the unique solution to (17). Suppose first that \( \mu_{\text{pooled}} = 0 \), either because (0,0) is the unique equilibrium or because (0,0) is the selected equilibrium, i.e., \( C(0,0) < C(\mu^*, \mu^*) \). In that case, the result trivially holds true. Suppose next \( \mu_{\text{pooled}} = \mu^* \), either because there are two equilibria (0,0) and (\( \mu^*, \mu^* \)), and \( C(0,0) \geq C(\mu^*, \mu^*) \), or there is only one equilibrium (\( \mu^*, \mu^* \)). Suppose first that \( b \geq \lambda k^{-k}(1 + k)^{1+k} \), which, by Proposition 6, implies that \( \mu_{\text{dedicated}} \) solves (15). Moreover, \( \mu^* > \lambda \) by Proposition 7. In that case, from (15), (17), and (18), we obtain

\[
C_i' (\mu) = \frac{\partial C(\mu, \mu)}{\partial \mu_i} = k \mu_{i-1} - b \frac{\lambda}{\mu_i^2} \leq k \mu_{i-1} - b \frac{\lambda}{\mu_i \mu} = C_i (\mu; \mu).
\]

We showed in §5.1 that \( C_i'' (\mu) > 0 \), \( d \left( \frac{\partial C(\mu, \mu)}{\partial \mu_i} \right) / d \mu > 0 \) when \( \mu > \lambda \), and \( d (C_i' (\mu; \mu)) / d \mu > 0 \) for all \( \mu > \lambda \). The result then follows because \( \mu_{\text{dedicated}}, \mu_{\text{centralized}}, \) and \( \mu_{\text{pooled}} \) are respectively roots of \( C_i (\mu), \frac{\partial C(\mu, \mu)}{\partial \mu_i} \), and \( C_i (\mu; \mu) \).

Suppose next that \( b < \lambda k^{-k}(1 + k)^{1+k} \), which, by Proposition 6, implies that \( \mu_{\text{dedicated}} = 0 \), i.e., that \( C_i (0) < C_i (\mu) \) for all \( \mu > 0 \). In that case, the result can only hold if (i) either (0,0) is the only equilibrium in the pooled system or (ii) there are two equilibria in the pooled system, but the selected equilibrium according to the Pareto equilibrium-selection rule is (0,0). Suppose, for contradiction, that (0,0) is not an equilibrium in the pooled system. Because choosing a capacity of 0 is not a best response to the other server’s choosing a capacity of 0, there exists a \( \mu > 0 \) such that \( C_i (\mu; 0) < C_i (0; 0) \), i.e., \( \mu + 2b \lambda / \mu < b \), a contradiction because \( C_i (\mu) > C_i (0) \) for all \( \mu \), i.e., \( \mu + b \lambda / \mu > b \). Suppose next that there are two equilibria in the pooled system, namely (0,0) and (\( \mu^*, \mu^* \)). In that case, (0,0) is the Pareto-dominant equilibrium because \( C_i (\mu^*; \mu^*) = (\mu^*)^k + b \lambda / \mu^* = C_i (\mu^*) > C_i (0) = C_i (0; 0) \) for \( i = 1, 2 \). Hence, if \( \mu_{\text{dedicated}} = 0 \), then \( \mu_{\text{pooled}} = 0 \).

\[ \square \]

Proof of Proposition 9. In the proof, we denote \( \mu_{\text{dedicated}} = \mu^d \) and \( \mu_{\text{pooled}} = \mu^p \). Applying the Implicit Function Theorem to (15) and (17), we obtain:

\[
\frac{d \mu^d}{db} = \frac{\lambda}{(\mu^d)^2} \frac{1}{k(k-1)(\mu^d)^{k-2} + 2b \frac{\lambda}{(\mu^d)^2}} = \frac{\lambda}{(\mu^d)^2} \frac{1}{(k-1)(\mu^d)^{-1}b \frac{\lambda}{(\mu^d)^2} + 2b \frac{\lambda}{(\mu^d)^2}} = \frac{\mu^d}{b(k+1)}
\]

\[
\frac{d \mu^p}{db} = \frac{\mu^p (\lambda + \mu^p)}{b(k+1)}
\]

Hence,

\[
\frac{d}{db} \left( \frac{\mu^d}{\mu^p} \right) \leq 0 \iff \mu^p \frac{d \mu^d}{db} - \mu^d \frac{d \mu^p}{db} \leq 0
\]

\[
\iff \frac{1}{k+1} \leq \frac{\lambda + \mu^p}{k(\lambda + \mu^p) + \mu^p}
\]

\[
\iff k(\lambda + \mu^p) + \mu^p \leq (k+1)(\lambda + \mu^p),
\]

which always holds. \[ \square \]
Proof of Theorem 2. The proof uses Lemmas C-19 and C-20. By Proposition 7, the equilibrium capacity in the pooled system \( \mu_{\text{pooled}} \) is either zero or \( \lambda \) that satisfies (17), or when there are two symmetric equilibria, namely \((0,0)\) and \((\mu^*, \mu^*)\) where \( \mu^* > \lambda \) solves (17), but, according to the Pareto-dominance equilibrium selection rule, \( C(0,0) \leq C(\mu^*, \mu^*) \). Because, when \( \mu > \lambda \), 
\[
\frac{d}{d\mu} \frac{\partial C_i(\mu; \mu)}{\partial \mu_i} = k(k-1)\mu^{k-2} + b\lambda \frac{\lambda + 2\mu}{\mu^2(\lambda + \mu)^2} > 0,
\]
the first case occurs when \( \partial C_i(\lambda; \lambda)/\partial \mu_i > 0 \), i.e., when \( b < 2k\lambda^k \). The second case happens when there are two equilibria and \( C(0,0) \leq C(\mu^*, \mu^*) \), i.e., \( b \leq (\mu^*)^k + b\lambda / \mu^* \). Because \( \mu^* \) solves (17), \( b = k(\mu^*)^k(\lambda + \mu^*)/\mu^* \), and the condition associated with the second case can be equivalently expressed in terms of \( \mu^* \) as follows: \((((\mu^*)^2 - \lambda^2)/\mu^* \leq \lambda/k \). Because the left-hand side is an increasing function of \( \mu^* \) and because \( \mu^* \) is an increasing function of \( b \), by the Implicit Function Theorem applied to (17), the left-hand side is an increasing function of \( b \). Hence, the condition \( b \leq (\mu^*)^k + b\lambda / \mu^* \) can be equivalently expressed as setting an upper bound on \( b \). Summarizing both cases, we find that the equilibrium \( \mu_{\text{pooled}} = 0 \) can be sustained only when \( b \) is smaller than a threshold. If \( \mu_{\text{pooled}} = 0 \) and \( \mu_{\text{dedicated}} \leq \lambda \), \( L_{\text{pooled}} = 2L_{\text{dedicated}} = \infty \). If \( \mu_{\text{pooled}} = 0 \) and \( \mu_{\text{dedicated}} > \lambda \), then \( L_{\text{pooled}} = \infty > 2L_{\text{dedicated}} \).

Suppose next that \( b \) is sufficiently large that the equilibrium capacity \( \mu_{\text{pooled}} \) is equal to \( \mu^* > \lambda \) and solves (17). By Proposition 8, \( \mu_{\text{dedicated}} \geq \mu_{\text{pooled}} \), and therefore \( \mu_{\text{dedicated}} > \lambda \). By Proposition 6, \( b \geq \lambda^k k^{-k} (1+k)^{1+k} \) and the condition of Lemma C-20 holds. By Lemma C-19, denoting \( \mu_{\text{pooled}} \) by \( \mu \) and \( L_{\text{dedicated}} \) by \( L \), we have:

\[
L_{\text{pooled}} \geq 2L_{\text{dedicated}} \iff \Psi(L) = \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right)^k - \frac{2(1 + L)}{2L + 1 + \sqrt{1 + 4L^2}} \geq 0.
\]

Taking the derivative of \( \Psi(L) \) when it crosses zero, we have:

\[
\Psi'(L) = \left. \frac{1}{\sqrt{1 + 4L^2}} \right|_{\Psi(L) = 0} = \left( \frac{k}{L(1 + L)} \right)^k \left( \frac{-1 - L + \sqrt{1 + 4L^2}}{L(1 + L)} \right) + \frac{2(-1 + 4L + \sqrt{1 + 4L^2})}{(1 + 2L + 1 + \sqrt{1 + 4L^2})^2}
\]

\[
= 2 \frac{(-1 + 4L + \sqrt{1 + 4L^2}) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right) + (-3 + 2L + \sqrt{1 + 4L^2}) \ln \left( \frac{2(1 + L)}{1 + 2L + \sqrt{1 + 4L^2}} \right)}{\sqrt{1 + 4L^2}(1 + 2L + \sqrt{1 + 4L^2})^2 \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right)}.
\]

Note that the denominator is negative because

\[
\ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right) \leq 0 \iff 1 + \sqrt{1 + 4L^2} \leq 2 + 2L \iff 1 + 4L^2 \leq 1 + 4L^2 + 4L \iff L \geq 0.
\]
Hence, when \( \Psi(L) = 0, \Psi'(L) \geq 0 \) if and only if the numerator is nonpositive, i.e., if and only if

\[
\Upsilon(L) = (-1 + 4L + \sqrt{1 + 4L^2}) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right) + (-3 + 2L + \sqrt{1 + 4L^2}) \ln \left( \frac{2(1 + L)}{1 + 2L + \sqrt{1 + 4L^2}} \right)
\]

is nonpositive. Because, for all \( L > 0 \),

\[
\Upsilon'(L) = \frac{4(L + \sqrt{1 + 4L^2}) \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right) + 2(2L + \sqrt{1 + 4L^2}) \ln \left( \frac{2(1 + L)}{1 + 2L + \sqrt{1 + 4L^2}} \right)}{\sqrt{1 + 4L^2}} < 0,
\]

since

\[
\ln \left( \frac{2(1 + L)}{1 + 2L + \sqrt{1 + 4L^2}} \right) \leq 0 \iff 2(1 + L) \leq 1 + 2L + \sqrt{1 + 4L^2} \iff 1 \leq \sqrt{1 + 4L^2} \iff 4L^2 \geq 0,
\]

we obtain that \( \Upsilon(L) \) is decreasing and therefore, for any \( L > 0 \), \( \Upsilon(L) < \Upsilon(0) = 0 \). Hence, \( \Psi'(L) > 0 \) when \( \Psi(L) = 0 \) and \( L > 0 \), that is, \( \Psi(L) \) crosses zero at most once as \( L \) increases from zero and the crossing is from below. Moreover, \( \Psi(0) = 0 \) and \( \Psi'(0) = -k < 0 \). Hence, there exists a threshold \( \overline{L}(k) > 0 \) such that \( L_{\text{pooled}} \geq 2L_{\text{dedicated}} \) if and only if \( L_{\text{dedicated}} \geq \overline{L}(k) \). Applying the Implicit Function Theorem to \( \Psi(L, k) = 0 \), we obtain

\[
\overline{L}(k) = -\frac{\partial \Psi(\overline{L}(k), k)}{\partial L},
\]

because the denominator is positive given that \( \Psi(L) \) crosses zero from below and the numerator is negative because \( \ln \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right) < 0 \) when \( L > 0 \).

Define \( \phi(L; k) \equiv L^{k+1}/(1 + L)^{k+1} \), and note that \( \phi(L; k) \) is increasing in \( L \). By Lemma C-20, \( \phi(L; k) = k\lambda^k/b \). Since \( \phi(\cdot; k) \) is increasing, it is invertible; hence, \( L_{\text{dedicated}} = \phi^{-1}(k\lambda^k/b; k) \). Hence, requiring that \( L_{\text{dedicated}} \geq \overline{L}(k) \) is equivalent to requiring that \( \phi^{-1}(k\lambda^k/b; k) \geq \overline{L}(k) \), or that \( k\lambda^k/b \geq \phi(\overline{L}(k); k) \). Defining \( B(k) = k/\phi(\overline{L}(k); k) \), this is equivalent to requiring that \( b/\lambda^k \leq B(k) \). \( \square \)

**Lemma C-19.** Under busyness aversion, when \( \mu_{\text{pooled}} > \lambda \), \( L_{\text{pooled}} \geq 2L_{\text{dedicated}} = 2L \) if and only if

\[
\left( \frac{1 + \sqrt{1 + 4L^2}}{2(L + 1)} \right)^k \geq \frac{2(L + 1)}{2L + 1 + \sqrt{1 + 4L^2}}.
\]

**Proof:** The proof uses Lemma C-20. By Proposition 8, \( \mu_{\text{dedicated}} \geq \mu_{\text{pooled}} \), and therefore \( \mu_{\text{dedicated}} > \lambda \). By Proposition 6, \( b \geq \lambda^k k^{-k}(1 + k)^{1+k} \) and the condition of Lemma C-20 holds. Using (7) and denoting \( \mu_{\text{pooled}} \) by \( \mu \) and \( L_{\text{dedicated}} \) by \( L \), we have:

\[
L_{\text{pooled}} \geq 2L_{\text{dedicated}} \iff \frac{2\mu\lambda}{\mu^2 - \lambda^2} \geq 2L \iff 0 \geq L(\mu^2 - \lambda^2) - \mu \lambda \iff \mu \leq \frac{\lambda}{2L} \left( 1 + \sqrt{1 + 4L^2} \right) = \alpha(L)\lambda.
\]

By the proof of Proposition 7, \( C'_i(\mu; \mu) \) is increasing in \( \mu \). Hence, \( \mu \leq \alpha \lambda \) if and only if \( C'_i(\mu; \mu) \leq C'_i(\alpha \lambda; \alpha \lambda) \).

Since when \( \mu > \lambda \), \( \mu \) solves \( C'_i(\mu; \mu) = 0 \) by Proposition 7, this is equivalent to requiring that \( C'_i(\alpha \lambda; \alpha \lambda) \geq 0 \).

Hence, using (17), Lemma C-20, and the definition of \( \alpha(L) \), we have

\[
\mu \leq \alpha(L)\lambda \iff C'_i(\alpha \lambda; \alpha \lambda) \geq 0
\]
\[ \iff k\alpha^k \lambda^k - b \frac{1}{(1 + \alpha)} \geq 0 \]

\[ \iff \alpha^k \left( \frac{L}{1 + L} \right)^{k+1} (1 + \alpha) \geq 1 \]

\[ \iff \left( \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right)^k \left( \frac{L}{1 + L} + \frac{1 + \sqrt{1 + 4L^2}}{2(1 + L)} \right) \geq 1. \square \]

**Lemma C-20.** Under business aversion, when \( b \geq \lambda^k k^{-k} (1 + k)^{1+k} \), \( L_{\text{dedicated}} \) satisfies

\[ \left( \frac{L}{1 + L} \right)^{k+1} = \frac{k\lambda^k}{b} \]

**Proof:** From (15), we obtain that \( \mu_{\text{dedicated}} = \left( \frac{b\lambda}{k} \right)^{\frac{1}{k+1}} \). Hence, using (4), we obtain that

\[ L = \frac{\lambda}{\mu - \lambda} = \frac{\lambda}{\left( \frac{b\lambda}{k} \right)^{\frac{1}{k+1}} - \lambda} = \frac{1}{\left( \frac{b}{k\lambda} \right)^{\frac{1}{k+1}} - 1} \iff L \left( \frac{b}{k\lambda} \right)^{\frac{1}{k+1}} = 1 + L \iff \left( \frac{L}{1 + L} \right)^{k+1} = \frac{k\lambda^k}{b}. \square \]