Design of an aggregated marketplace under congestion effects: Asymptotic analysis and equilibrium characterization

Ying-Ju Chen* Costis Maglaras† Gustavo Vulcano‡

July 28, 2008

Abstract

We study an aggregated marketplace where potential buyers arrive and submit requests-for-quotes (RFQs). There are n independent suppliers modelled as $M/GI/1$ queues that compete for these requests. Each supplier submits a bid that comprises of a fixed price and a dynamic target leadtime, and the cheapest supplier wins the order as long as the quote meets the buyer's willingness to pay.

We characterize the asymptotic performance of this system, and subsequently extract insights about the equilibrium behavior of the suppliers and the efficiency of this market. We show that supplier competition results into a mixed-strategy equilibrium phenomenon and is significantly different from the centralized solution. We propose a compensation-while-idling mechanism that coordinates the system: each supplier gets monetary compensation from other suppliers during his idle periods. This mechanism alters suppliers' objectives and implements the centralized solution at their own will.

Keywords: aggregated marketplace, service competition, asymptotic analysis

1 Introduction

Business-to-Business (B2B) online markets constitute an Internet-based solution that aggregates business entities interested in buying and selling related goods or services from one another. They provide automatic transactions, lower searching costs, and increased process effectiveness and efficiency. Despite the turbulent dynamics of the developing B2B environment, several e-markets have

*University of California, 4121 Etcheverry Hall, Berkeley, CA 94720, chen@ieor.berkeley.edu

†Columbia Business School, 409 Uris Hall, 3022 Broadway, New York, NY 10027, c.maglaras@gsb.columbia.edu

‡New York University, 44 West 4th street, KMC 8-76, New York, NY 10012, gvulcano@stern.nyu.edu

survived and have exhibited a high degree of success (e.g., see Laseter and Bodily [2004]). The worldwide spread of e-markets is confirmed by a recent release about B2B transactions in China hitting US$ 168.9 billion during 2007, representing a 25.5% increase from 2006.[1]

B2B online market is well-suited for industrial procurement. Interested (sometimes, subscribed) buyers arrive at the e-marketplace and submit request-for-quotes (RFQs) that include product specification information, desired quality level and quantity, and target leadtime. Potential suppliers (maybe pre-qualified) bid for these orders, and the bids are ranked according to a scoring function that depends on the quote attributes. The order is then awarded to the most "desirable" supplier. Each transaction is typically run as a reverse auction. This aggregated marketplace raises several interesting practical and theoretical questions: How does the system dynamics evolve as suppliers compete for each potential order? How should these suppliers determine their bidding strategies? Is the market efficient under the competitive behavior? If not, is it possible to find a coordination scheme that aligns the suppliers' incentives? How can we implement it?

From a modeling viewpoint, an aggregated marketplace is a system that operates in a stochastic environment and handles high volumes of orders that are routed in real-time to the winning supplier in each case. As a starting point of addressing the aforementioned questions, we develop a stylized stochastic model where potential buyers arrive according to a Poisson process and a number of independent suppliers (modeled as $M/GI/1$ queues) compete for these requests. Each supplier submits a bid that comprises a fixed price and a target leadtime that depends on his own queue status. Each buyer uses her scoring function to compute the net utility associated with each supplier's bid, and awards the order to the lowest-quote supplier in order to maximize her own surplus (provided that it is nonnegative). Delay quotations are state-dependent, and as a result the input stream of orders that get awarded to one of the cheapest suppliers at any point in time is itself state-dependent. At each supplier, orders are processed in a First-In-First-Out manner. The suppliers' capacities (i.e., service rates) are common knowledge; nevertheless, a supplier's queue length is privately observed by himself but unknown to other suppliers. Suppliers compete for these orders by selecting prices in order to maximize their own long-run average revenues. Each supplier faces an economic tradeoff: a high price will lead to high revenues per order, but will reduce the total number of orders awarded, which will cause excessive idleness and implicit revenue loss; a low price will result in many awarded orders and large backlogs, that, in turn, will cause the

[1]Reported by Xinhua News Agency on January 9th, 2008.

long delay quotations thus increasing the full cost of the respective bid.

Despite the stylized nature of the above problem, the resulting dynamical system is still difficult to analyze due to the state dependent nature of the order input stream, the dynamic order routing policy, and the supplier competitive behavior. The most direct tool for analyzing the system behavior is simulation, but this offers little insight on how to study the supplier pricing game. Our approach, therefore, is to adopt some form of approximate analysis that leads to tractable characterizations of suppliers' behavior. These approximations are motivated by an intuitive economic argument, whose robustness is supported by the resulting supplier pricing game equilibrium.

The first problem addressed in this paper is that of performance analysis assuming that the price vector is given. The solution of this problem allows suppliers to evaluate their revenues given their prices as well as the prices of the competitors, which is an essential subroutine in the equilibrium analysis of the supplier pricing game. So, given a specified price vector, the first set of results of this paper characterize the behavior of the marketplace using an asymptotic analysis that focuses on settings where the potential demand and the supplier processing capacities grow large simultaneously. This asymptotic analysis is motivated by the following observation: If this market were served by a unique supplier (modeled as an $M/GI/1$ queue as well), then it would be economically optimal for this supplier to set the price that induces the so-called "heavy-traffic" operating regime; i.e., rather than assuming that the system is operating in the heavy traffic regime, as is often done, this result provides a primitive economic foundation that this regime emerges naturally since it optimizes the system-wide revenues (Besbes [2006] and Maglaras and Zeevi [2003]). Specifically, if Λ is the market size, then the above result says that the economically optimal price is of the form $p^* = \bar{p}+\pi/\sqrt{\Lambda}$, where π is the price that induces full resource utilization in the absence of any congestion; the result is proved in an asymptotic sense as the market size and processing capacity of the server grow large.

Based on this observation, the starting point of our analysis is to write the supplier prices as perturbations around the price $\bar{p}$ of the form $p_i = \bar{p} + \pi_i/\sqrt{\Lambda}$. Treating Λ as a scaling parameter that will grow large, we can view the system of original interest as an element of a sequence of systems that we will approximate asymptotically. §3.1 motivates this modeling substitution, §3.2 describes the above sequence of systems, and §3.3 - 3.4 derive the asymptotic behavior of this sequence of systems. Proposition 1 derives a transient behavior property of this system using an appropriate deterministic fluid model. Proposition 2 establishes a state-space collapse result,

whereby the suppliers' behavior is asymptotically coupled and can be described as a function of the aggregate (system-wide) workload process; we further characterize this workload process by the state-dependent arrival process and a state-independent service time process. This state space collapse result also implies that a supplier is able to know his competitors' bids by simply observing awaiting orders in his own buffer. Finally, Theorem 1 establishes the weak convergence for the one-dimensional workload process to a reflected Ornstein-Uhlenbeck process, with a reflection potentially away form zero depending on the suppliers' prices. The latter implies the potentially surprising property that the aggregate workload process can never drain even though some of the suppliers may be idling.

The second problem is the characterization of the revenue streams of each supplier (Lemmas 1 and 2) and the analysis of the resulting pricing game. Lemma 3 derives the centralized solution, i.e., the pricing choice that would maximize the system-wide revenues. Lemma 4 shows that the pricing game does not admit a pure strategy equilibrium, and the online appendix specifies the structure of the supporting mixed strategy equilibria where suppliers randomize over their pricing decisions. The online appendix also proves that the (second-order) efficiency loss of the decentralized solution can be arbitrarily large.

The third problem we address is that of specifying an intuitive mechanism that would "coordinate" the marketplace, in the sense that all suppliers will self-select to price according to the centralized solution. Propositions 3 and 4 specify such a transfer pricing mechanism that compensates suppliers during idleness periods, which aligns the incentives of the competing suppliers and recovers the performance of the centralized solution. It is worth noting that our approximate analysis of the supplier game is internally consistent in the sense that the lower order price perturbations that essentially capture the supplier pricing game do not become unbounded, but rather stay finite. In essence, all suppliers choose to operate in the asymptotic regime we propose, both under the market mechanism and the coordinating mechanism.

To recapitulate, the main contributions of the paper are three: 1) the formulation and characterization of the equilibrium solution of the supplier pricing game through the use of asymptotic approximations that are amenable to a tractable analysis and lead to structural insights about the behavior of the aggregated marketplace; 2) the derivation of the limit model, which is of separate interest; and, 3) the proposal of a market mechanism that is intuitive, implementable, and it coordinates the suppliers' behavior.

Literature review: Our paper touches on three related bodies of literature. The first focuses on the economics of queues, the second on competition models in queueing contexts, and the third develops machinery for approximating the behavior of complex queueing models such as the ones we study in this paper.

The literature that studies the pricing of queues dates back to Naor [1969]; see the monograph by Hassin and Haviv [2002] for a review of this literature. The economic model that we use is inspired by that introduced by Mendelson [1985]; there is one type of potential customers that arrive according to a Poisson arrival process, each having a private valuation that is an independent draw from a general distribution and a delay sensitivity parameter that is common across all customers. Mendelson and Whang [1990] extended that model to multiple customer types, where each type is characterized by its own valuation distribution and delay sensitivity parameter. The above papers study congestion pricing schemes for a single server system that maximize the social welfare, and where customers decide on whether to join the system based on the steady-state delay cost as opposed to the real-time delay cost they will actually incur based on the backlog ahead of them at the time of their arrival to the system.

The system analyzed in this paper looks at a customer population that as in Mendelson [1985] differs in their valuations but have a common delay sensitivity, but in contrast to the above set of papers it is served by multiple competing revenue maximizing suppliers (servers). Suppliers compete by supplying bids that comprise of a static (state-independent) price and a state-dependent delay. The revenue maximization pricing problem for a single server queue with state-dependent information was studied in Mendelson [1985] (for an $M/M/1$ system). The use of real-time delay information results into an arrival process with state-dependent rate, which complicates the solution of the pricing problem. In addition, the presence of multiple suppliers that compete for the arriving orders further complicates the analysis of the behavior of the marketplace under any given price vector, and of the supplier pricing game.

Our use of asymptotic approximations and heavy traffic analysis to study the supplier pricing game is motivated by the results of Maglaras and Zeevi [2003] and Besbes [2006], mentioned earlier, that showed that in large scale systems the heavy traffic regime is the one induced by the revenue maximizing price. Our paper implicitly assumes the validity of the heavy traffic regime (as opposed to proving it as in the two papers above) in deriving its asymptotic approximation, but the equilibrium pricing behavior of the competing suppliers seems to lend support to this assumption in

the sense that no supplier wishes to price in a way that would deviate from that operating regime.

Our derivation of the limit model makes heavy use of the work by Mandelbaum and Pats [1995] on queue with state dependent parameters, and of the framework developed by Bramson [1998] for proving state space collapse results. We also use results from Williams [1998] and Ata and Kumar [2005] in our analysis. A paper that studies a problem that is related to ours is Stolyar [2005], however, the analysis therein does not cover the type of problem we are interested in, mainly due to the presence of the suppliers' pricing decisions that introduce a non-zero cost (potentially negative) to the arriving customer even when the system is empty.

In the context of queueing models with pricing and service competition, starting from the early papers by Luski [1976] and Levhari and Luski [1978], customers are commonly assumed to select their service provider on the basis of a "full cost" that consists of a fixed price plus a waiting cost. In both Luski [1976] and Levhari and Luski [1978], competition is modeled in a duopoly setting where firms behave as $M/M/1$ queueing systems. More recent variants of Levhari and Luski [1978] within the $M/M/1$ duopoly framework include Li and Lee [1994] and Armony and Haviv [2003], and an extensive survey of this research stream in provided by Hassin and Haviv [2002]. There are also a few papers that generalize the model characteristics to incorporate other scenarios. For example, Loch [1991] study a variant of Luski's model in which the providers are modeled as symmetric $M/GI/1$ systems. Lederer and Li [1997] generalize Loch [1991] for arbitrary number of service providers. Cachon and Harker [2002] and So [2000] analyze customer choice models more general than always choosing the lowest cost supplier, although confined to the $M/M/1$ case. Allon and Federgruen [2007] treat the price and waiting time cost as separate firm attributes that can be traded off differently by each arriving customer. The restriction of confining to $M/M/1$ firms and linear demand rates was relaxed in their follow-up paper Allon and Federgruen [2006].

More recently, Johari and Weintraub [2008] study two types of contractual agreements in oligopolistic service industries with congestion effects: Service Level Guarantees (SLG), where each firm is responsible for investing to guarantee a congestion level for the users consistent with the standard, and Best Effort (BE), where firms provide the best possible service given their infrastructure. Gurvich and Allon [2008] appeal to an asymptotic analysis to study a competitive game of a queueing model, and propose a general recipe for relating the asymptotic outcome to that of the original system. They show that the pricing decisions and service level guarantees result in respectively first-order and second-order effects on the suppliers' payoffs. Notably, the overall

approach of Gurvich and Allon [2008] is similar to ours; however, our model falls outside their scope, and our economic motivation for the asymptotic analysis is different. It is worth mentioning that while all the aforementioned competitive models in queues assume a static measure of the waiting time standard (usually the expected value or some percentile of the steady state distribution), we characterize the equilibrium behavior based on the asymptotic performance of the system under *dynamic* delay quotations.

In terms of the characterization of market equilibrium, we follow Varian [1980] and Baye et al. [1992] that provide a systematic approach to characterize the mixed-strategy equilibria when no pure-strategy equilibrium can be found. This approach is typically used to study rent-seeking, political contests, R&D competition, etc., in which every individual makes non-cooperative inputs but typically the reward is fixed ex ante and is given to the individual with the highest input. Our analysis extends this research stream to the case when the aggregate penalty/reward function also depends on the individual decisions (suppliers' pricing decisions in our context). Our derivation of the coordination mechanisms follows the standard approach that aligns individuals' incentives by internalizing the externality an individual brings to the system (see, e.g., Mas-Colell et al. [1995]). Our contribution is to identify the appropriate externality through our formulation of the suppliers pricing game and to find a coordination mechanism that is easily implementable.

The remainder of the paper is organized as follows. In §2, we describe the model details. §3 derives the asymptotic characterization of the marketplace behavior, and §4 characterizes the equilibrium behavior of the supplier pricing game. §5 offers some concluding remarks. The proofs of propositions and theorems are in the appendix, and the proofs of auxiliary results are relegated to the online appendix.

Notation: The notation $f(x) = o(g(x))$ refers to the fact $f(x)/g(x) \to 0$, and $f(x) = g(x)$ to the fact that $f(x) = g(x) + o(\sqrt{x})$ or $f(x) = g(x) + o(1/\sqrt{x})$. Similarly, $f(x) = o_p(g(x))$ if $f(x)/g(x) \to 0$ in probability. We say that $f(x) = O(g(x))$ if $f(x)/g(x) \to a$, for a constant $a > 0$ and define $f(x) = O_p(g(x))$ as the counterpart of convergence in probability. For any integer $k > 0$, and for any $y \in \mathbb{R}^k$, we define $|y| := \max\{|y_j|, j = 1, ..., k\}$ as its maximum norm. For any function $f : \mathbb{R}_+ \to \mathbb{R}^k$ and constant $L > 0$, we define $||f(\cdot)||_L := \sup_{0 \le t \le L} |f(t)|$. For any sequence $\{a_r, r \in \mathbb{N}\}$, $a_r \to \alpha$ if for any $\delta > 0$, there exists r_δ such that $|\alpha - a_r| < \delta$, whenever $r > r_\delta$. The function $\phi(\cdot)$ denotes the standard normal density, and $\Phi(\cdot)$ its corresponding c.d.f.

2 Model

We consider an aggregated marketplace that functions as follows. The market is served by a set of suppliers $\mathcal{N} = \{1, \ldots, n\}$ that can produce one type of good. Potential buyers arrive according to a homogeneous Poisson process with rate Λ, and submit requests-for-quotes (RFQs). Each RFQ corresponds to the procurement of one unit of that good. Each supplier submits a bid that comprises a price and a target leadtime component for each such RFQ, and the order is awarded to the "cheapest" supplier in a way that will be made precise below. Each supplier selects his initial investment in processing capacity, and subsequently competes for the business brought by the arriving buyers.

Suppliers: There are n suppliers. Each supplier i is modeled as an $M/GI/1$ queue with an infinite capacity buffer managed in a First-In-First-Out fashion. Service times at supplier i follow a general distribution with processing rate $\mu_i^{'}$ (and mean equal to $1/\mu_i^{'}$) and standard deviation σ_i. Let $\hat{\mu} := \sum_{i \in \mathcal{N}} \mu_i^{'}/\Lambda$ be the (normalized) aggregated service rate of the market.

Order arrivals: Potential buyers arrive to the marketplace according to a Poisson arrival process with intensity Λ. Each buyer has a private valuation v for her order that is an independent draw from a general, and continuously differentiable distribution $F(\cdot)$. Buyers are delay sensitive and incur a cost c per unit of delay. Thus, buyers are homogeneous with respect to delay preferences, and heterogeneous with respect to valuations (though symmetric across the common c.d.f. $F(\cdot)$). A buyer that arrives at time t initiates an RFQ process to procure one unit of good. The service time associated with an order depends on the selected supplier.

Market mechanism: Suppliers compete for this request by submitting bids that comprise a price p_i and a target leadtime $d_i(t)$. We assume that the price component of the bid is state-independent, i.e., supplier i always submits the same price bid p_i for all orders. The leadtime component of the bid submitted by each supplier i is state-dependent and equals the expected time it would take to complete that order; cf. (2) later on. We are implicitly assuming that the supplier always submits a truthful estimate of that expected delay d_i, which could be supported by long-term reputation arguments, or by the existence of an auditor in the aggregated marketplace.

On their end, buyers are price and delay sensitive, and for each supplier i they associate a "full cost" given by $p_i + cd_i(t)$, where the delay sensitivity parameter c is assumed to be common for all buyers. Upon reception of the bids, the buyer awards her order to the lowest cost supplier,

provided that her net utility is positive, i.e., $v \geq \arg\min_{i\in\mathcal{N}}\{p_i + c\,d_i(t)\}$; otherwise, the buyer leaves without submitting any order. Whenever a tie occurs, the order is awarded by randomizing uniformly among the cheapest suppliers.[2]

Given vectors $p = (p_1, \ldots, p_n)$, and $d(t) = (d_1(t), \ldots, d_n(t))$, the instantaneous rate at which orders enter this aggregated market is given by

$$\lambda(p, d(t)) = \Lambda \bar{F}(\min_i\{p_i + c\,d_i(t)\}).$$

Focusing on the right-hand-side of the above expression, we note that the buyers' valuation distribution $F(\cdot)$ determines the nature of the aggregate demand rate function.[3] Let $x = \min_i p_i$ and, with slight abuse of notation, write $\lambda(x)$ in place of $\lambda(p, 0)$, where 0 is the vector of zeros. We further define $\epsilon(x) = -\frac{\partial\lambda(x)}{\partial(x)}\frac{x}{\lambda(x)}$. We will make the following economic assumption.

Assumption 1. *$\lambda(p, 0)$ is elastic in the sense that $\epsilon(x) > 1$ for all price vectors p in the set $\{p : 0 \leq \lambda(p, 0) \leq \sum_{i=1}^{n} \mu_i^{'}\}$ and $x = \min_i p_i$.*

The above assumption implies that in the absence of delays a decrease in price would result in an increase in the market-wide aggregated revenue rate $p \cdot \lambda(p, 0)$.[4] Of course, this would increase the utilization levels of the suppliers, leading to increased congestion and delays, which, in turn, would moderate the aggregate arrival rate $\lambda(p, d(t))$.

Let $A(t)$ be the cumulative number of orders awarded to all the competing suppliers up to time t,

$$A(t) = N\left(\Lambda \int_0^t \bar{F}(\min_{i\in\mathcal{N}}\{p_i + c\,d_i(s)\})ds\right),$$

where $N(t)$ is a unit rate Poisson process. To represent the cumulative number of orders for each individual supplier, let

$$\mathcal{J}(t) \equiv \{i \in \mathcal{N} : p_i + cd_i(t) \leq p_j + cd_j(t), \forall j \in \mathcal{N}\}$$

[2]We could also allow other tie-breaking rules, and it can be verified that our results are not prone to the specific choice of tie-breaking rules.

[3]For example, if $v \sim U[0, \Lambda/\alpha]$, then the demand function is linear, of the form $\lambda(x) = \Lambda - \alpha\,x$, where $x = \min_i p_i + cd_i(t)$; if $v \sim \text{Exp}(\alpha)$, then the demand is exponential, with $\lambda(x) = \Lambda\, e^{-\alpha x}$.

[4]In the absence of congestion effects and assuming that there exists a central planner that could select a common price p and an aggregate capacity $\hat{\mu}$ under a linear capacity cost $h\hat{\mu}$ and the arrival rate Λ, the solution to the problem $\max_{p,\hat{\mu}}\ \{p\lambda(p, 0) - h\Lambda\hat{\mu}\ :\ 0 \leq \lambda(p, 0) \leq \Lambda\hat{\mu}\}$ results in a capacity decision that satisfies the above assumption.

as the set of the cheapest suppliers at time t. Further, define $\Xi_{\mathcal{J}(t)}$ as the random variable that assigns the orders uniformly amongst the cheapest suppliers. That is, $\Xi_{\mathcal{J}(t)} = i$ with probability $\frac{1}{|\mathcal{J}(t)|}$ if $i \in \mathcal{J}(t)$, where $|\mathcal{J}(t)| > 0$ is the cardinality of $\mathcal{J}(t)$, and $\Xi_{\mathcal{J}(t)} = i$ with zero probability otherwise. To facilitate exposition, it is easiest to assume that $\mathcal{J}(t)$ and $\Xi_{\mathcal{J}(t)}$ are defined as continuous processes for all $t \geq 0$, even if no actual arrival occurs at that time. This allows us to write the cumulative number of orders awarded to supplier i, denoted by $A_i(t)$, as

$$A_i(t) = \int_0^t \mathbb{1}\{\ \Xi_{\mathcal{J}(s)} = i\} dA(s),$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Note also that $A(t) = \sum_{i \in \mathcal{N}} A_i(t)$.

Supplier dynamics: Let $Q_i(t)$ denote his number of jobs in the system (i.e., in queue or in service) at time t, and $T_i(t)$ denote the cumulative time that supplier i has devoted into producing orders up to time t, with $T_i(0) = 0$. Let $Y_i(t)$ denote the idleness incurred by supplier i up to time t. Note that $T_i(t) + Y_i(t) = t$ for each supplier i; moreover, $Y_i(t)$ can only increase at time t where the queue $Q_i(t)$ is empty. Let $S_i(t)$ be the number of supplier i's service completions when working continuously during t time units, and $D_i(t) = S_i(T_i(t))$ be the cumulative number of departures up to time t. The production dynamics at supplier i are summarized in the expression:

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t). \tag{1}$$

Given this notation and using the PASTA (Poisson arrivals see time averages) property,

$$d_i(t) = \frac{Q_i(t) + 1}{\mu_i^{'}}, \tag{2}$$

gives the expected sojourn time of the new incoming order if it gets awarded to supplier i. Under our modeling assumptions, supplier i will therefore bid $(p_i, d_i(t))$, where $d_i(t)$ is given by (2).

We assume that the capacity vector $\mu^{'} \equiv \{\mu_i^{'}\}$'s is common knowledge. Each supplier knows his own system queue length $Q_i(t)$, but is not informed about his competitors' queue lengths. This paper studies three problems for the market model described above:

1. *Performance analysis for a given p:* Given a fixed price vector p and a vector of processing capacities $\mu^{'}$, the first task is to characterize the system performance, i.e., to characterize the behavior of the queue length processes $Q_i(t)$ at each supplier, and calculate the resulting revenue streams for each supplier. A supplier's long-run revenue is

$$\Phi_i(p_i, p_{-i}) \equiv p_i \cdot \lim_{t \to \infty} \frac{S_i(T(t))}{t}, \tag{3}$$

where $p_{-i} \equiv (p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n)$ denotes other suppliers' price decisions. Our goal, therefore, is to analyze the performance of relevant system dynamics that leads to a tractable representation of these long-run average revenues.

2. *Characterization of market equilibrium:* The above problem serves as an input to studying the competitive equilibrium that characterizes the supplier pricing game, which is a one-shot game where each supplier selects its static price. The assumption that suppliers are uninformed about the queue lengths of the competitors raises the issue of information incompleteness; however, as we will show, the suppliers' competitive behavior is insensitive to this assumption. Consequently, we adopt the Nash equilibrium as our solution concept, which is appropriate for games with *complete information*. Given the revenue specified in (3), a Nash equilibrium $\{p_i^*\}$ requires that $p_i^* = \arg\max_{p_i} \Phi_i(p_i, p_{-i}^*)$, $\forall i \in \mathcal{N}$.

3. *Market efficiency and market coordination:* Characterize the efficiency loss:

$$\max_{\{p_i\}} \left\{ \sum_{i \in \mathcal{N}} \Phi_i(p_i, p_{-i}) \right\} - \sum_{i \in \mathcal{N}} \Phi_i(p_i^*, p_{-i}^*),$$

i.e., the difference between the performance of a system where a central planner would control the pricing decision of each supplier (the first best solution) and of the market equilibrium. If the market equilibrium is inefficient, we would like to specify a simple market mechanism that coordinates the market and achieves the first best solution identified above. Such a mechanism could specify, for example, the rules according to which orders are allocated and payments are distributed among the suppliers.

3 Asymptotic analysis of marketplace dynamics

This section focuses on the first problem described in §2. Despite the relatively simple structure of the suppliers' systems and the customer/supplier interaction, it is still fairly hard to study their dynamics due to the state-dependent delay quotations and the dynamic order routing rule. Our approach is to develop an approximate model for the market dynamics that is rigorously validated in settings where the market potential and processing capacities of the various suppliers are large. In this regime, the market and supplier dynamics simplify significantly, and are essentially captured through a tractable one-dimensional diffusion process. This limiting model offers many insights about the structural properties of this market, and provides a vehicle within which we are able

to analyze the supplier game and the emerging market equilibrium. This is pursued in the next section.

3.1 Background: Revenue maximization for an $M/M/1$ monopolistic producer

As a motivation for our subsequent analysis, this section will summarize some known results regarding the behavior of a monopolistic supplier modeled as an $M/M/1$ queue that offers a product to a market of price and delay sensitive customers. The supplier posts a static price and dynamically announces the prevailing (state-dependent) expected sojourn time for orders arriving at time t, which is given by $d(t) = (Q(t) + 1)/\mu$. The assumptions on the customer purchase behavior are those described in the previous section. Given p and $d(t)$, the instantaneous demand rate into the system at time t is given by $\lambda(t) = \Lambda \bar{F}(p + cd(t))$. The supplier wants to select p to maximize his long-run expected revenue rate.

It is easy to characterize the structure of the revenue maximizing solution in settings where the potential market size Λ and the processing capacity μ grow large. Specifically, we will consider a sequence of problem instances indexed by r, where $\Lambda^r = r, \quad \mu^r = r\mu$; that is, r denotes the size of the market. The characteristics of the potential customers, namely their price sensitivity c and valuation distribution $F(\cdot)$, remain unchanged along this sequence. Let $\hat{p} = \arg\max p\bar{F}(p)$ and $\bar{p}$ be the price such that relation $\Lambda^r \bar{F}(\bar{p}) = \mu^r$ holds; i.e., neglecting congestion effects, $\hat{p}$ is the price that maximizes the revenue rate and $\bar{p}$ is the price that induces full resource utilization; both of these quantities are independent of r. Assumption 1 implies that $\hat{p} < \bar{p}$ (or equivalently that $\Lambda^r \bar{F}(\hat{p}) > \mu^r$) thus accentuating the potential tradeoff between revenue maximization and the resulting congestion effects. Besbes [2006] showed that the revenue maximizing price, denoted by $p^{*,r}$, is of the form

$$p^{*,r} = \bar{p} + \pi^*/\sqrt{r} + o(1/\sqrt{r}), \tag{4}$$

where π^* is a constant independent of r. Moreover, the resulting queue lengths are of order $\sqrt{r}$, or in a bit more detail the normalized queue length process $\tilde{Q}^r(t) = Q^r(t)/\sqrt{r}$ has a well defined stochastic process limit as $r \to \infty$. Since the processing time is itself of order $1/r$, the resulting delays are of order $1/\sqrt{r}$. Therefore, the delays are moderate in absolute terms (of order $1/\sqrt{r}$) but significant when compared to the actual service time (of order $1/r$).[5] If the supplier can select the price and capacity μ, the latter assuming a linear capacity cost, then the optimal capacity

[5] This is an extension of the result in Maglaras and Zeevi [2003] to a system with state-dependent delay information.

choice is indeed such that $\hat{p} < \bar{p}$ (where $\bar{p}$ is determined by μ), i.e., making the above regime the "interesting" one to consider. Finally, we note that the above results also hold for the case of generally distributed service times (Besbes [2006]).

3.2 Setup for asymptotic analysis

Given a set of suppliers characterized by their prices and capacities p_i, μ_i', we propose the following approximation:

1. Define the normalized parameters $\mu_i = \mu_i'/\Lambda$ for all suppliers i.
2. Define $\bar{p}$ to be the price such that $\Lambda \bar{F}(\bar{p}) = \sum_{i \in \mathcal{N}} \mu_i'$. Define $\pi_i = \sqrt{\Lambda}(p_i - \bar{p})$ so that the prices p_i can be represented as $p_i = \bar{p} + \pi_i/\sqrt{\Lambda}$.
3. Embed the system under consideration in the sequence of systems indexed by r and defined through the sequence of parameters:

$$\Lambda^r = r, \quad \mu_i^r = r\mu_i, \ \forall i \in \mathcal{N}, \quad c^r = c, \quad \text{and} \quad v \sim F(\cdot), \tag{5}$$

and prices given by $p_i^r = \bar{p} + \pi_i/\sqrt{r}$ for all i.

Given the preceding discussion, one would expect that the market may operate in a manner that induces almost full resource utilization, and where the underlying set of prices take the form assumed in item 3) above; this would be true if the market was managed by a central planner that could coordinate the supplier pricing and capacity decisions. The approach we propose to follow is to embed the system we wish to study in the sequence of systems indexed by r and described in (5), and subsequently approximate the performance of the original system with that of a limit system that is obtained as $r \to \infty$, which is more tractable. Note that for $r = \Lambda$ in (5), where Λ denotes the market size of the potential order flow as described in the previous section, we recover the exact system we wish to study. If Λ is sufficiently large, then the proposed approximation is expected to be fairly accurate.

The remainder of this section derives an asymptotic characterization of the performance of a market that operates under a set of parameters (p, μ') that are embedded in the sequence (5).

3.3 Transient dynamics via a fluid model analysis

The derivation of the asymptotic limit model (specifically, Proposition 2) will show that the following set of equations

$$\bar{Q}_i(t) = \bar{Q}_i(0) + \bar{A}_i(t) - \bar{D}_i(t), \forall i \in \mathcal{N}, \tag{6}$$

$$\bar{A}_i(t) = \int_0^t \mathbb{1}\left\{\Xi_{\mathcal{J}(s)} = i\right\} d\ N\left(\bar{F}(\bar{p})s\right), \forall i \in \mathcal{N}, \tag{7}$$

$$\bar{D}_i(t) = \mu_i \bar{T}_i(t), \forall i \in \mathcal{N}, \tag{8}$$

$$\int_0^t \bar{Q}_i(s) d\bar{Y}_i(s) = 0, \forall i \in \mathcal{N}, \tag{9}$$

$$\bar{T}_i(t) + \bar{Y}_i(t) = t, \forall i \in \mathcal{N}, \tag{10}$$

$$\bar{W}(t) = \sum_{i \in \mathcal{N}} \frac{\bar{Q}_i(t)}{\mu_i}. \tag{11}$$

captures the market's transient dynamics over short periods of length $1/\sqrt{r}$. This subsection studies the transient evolution of (6)-(11) starting from arbitrary initial conditions.

In (6)-(11), the processes $\bar{Q}, \bar{A}, \bar{D}, \bar{T}, \bar{Y}$ are the fluid analogues of Q, A, D, T, Y defined in Section 2, and $\bar{W}$ are the fluid analog of the system workload W. Equation (7) indicates how the arrivals are routed to these servers: An arrival walks away if her valuation is sufficiently low; otherwise, she joins server i based on the routing rule specified in Section 2. From Equation (7), the orders get awarded to the various suppliers at a rate $\Lambda\bar{F}(\bar{p}) = \sum_{i \in \mathcal{N}} \mu_i'$, i.e., $\bar{F}(\bar{p}) = \sum_{i \in \mathcal{N}} \mu_i$ (as indicated by the aggregate counting process $N(\bar{F}(\bar{p})t)$). Equation (9) demonstrates the non-idling property: $\bar{Y}_i(t)$ cannot increase unless $\bar{Q}_i(t) = 0$. Finally, Equation (11) establishes the connection between the total workload and the queue lengths of each supplier.

The next Proposition establishes that starting from any arbitrary initial condition, the transient evolution of the market (as captured through (6)-(11)) converges to a state configuration where all suppliers are equally costly in terms of the full cost of the bids given by (price + c $\times$ delay). This is, of course, a consequence of the market mechanism that awards orders to the cheapest supplier(s), until their queue lengths build up so that their full costs become equal. Simultaneously, expensive suppliers do not get any new orders and therefore drain their backlogs until their costs become equal to that of the cheapest suppliers. From then on, orders are distributed in a way that balances the load across suppliers. This result is robust with respect to the tie-breaking rule that one may apply when multiple suppliers are tied in terms of their full cost.

Proposition 1. *Let $\bar{Q}, \bar{A}, \bar{D}, \bar{T}, \bar{Y}, \bar{W}$ be the solution to (6)-(11) with $\max\{|\bar{Q}(0)|, |\bar{W}(0)|\} \leq M_0$ for some constant M_0. Then for all $\delta > 0$, there exists a continuous function $s(\delta, M_0) < \infty$ such that*

$$\max_{i,j \in \mathcal{N}} \left| \left(\pi_i + c\frac{\bar{Q}_i(s)}{\mu_i} \right) - \left(\pi_j + c\frac{\bar{Q}_j(s)}{\mu_j} \right) \right| < \delta, \ \forall s > s(\delta, M_0).$$

3.4 State-space collapse and the aggregate marketplace behavior

The next result shows that the transients studied above appear instantaneously in the natural time scale of the system, and as such the marketplace dynamics appear as if all suppliers are equally costly at all times.

We use the superscript r to denote the performance parameters in the r-th system, e.g., $A_i^r(t)$, $S_i^r(t)$, $T_i^r(t)$, and $Q_i^r(t)$. The (expected) workload (i.e., the time needed to drain all current pending orders across all suppliers) is defined as $W^r(t) = \sum_{i \in \mathcal{N}} \frac{Q_i^r(t)}{\mu_i^r}$.

Motivated by the discussion in §3.1 we will optimistically assume (and later on validate) that the supplier queue lengths are of order $\sqrt{r}$, and accordingly define re-scaled queue length processes for all suppliers according to

$$\tilde{Q}_i^r(t) = \frac{Q_i^r(t)}{\sqrt{r}}. \tag{12}$$

The corresponding re-scaled expected workload process is given by $\tilde{W}^r(t) = \sqrt{r} W^r(t) = \sum_{i \in \mathcal{N}} \frac{\tilde{Q}_i^r(t)}{\mu_i}$.

Define $\tilde{Z}^r(t) = \bar{\pi} + \bar{c}\tilde{W}^r(t)$, where $\bar{\pi} = \frac{\sum_{i \in \mathcal{N}} \pi_i}{n}$, $\bar{c} = \frac{c}{n}$. $\tilde{Z}^r(t)$ can be regarded as a proxy for the average of the second-order terms of suppliers' bids since $\tilde{Z}^r(t) = \frac{1}{n} \sum_{i \in \mathcal{N}} \left(\pi_i + c\frac{\tilde{Q}_i^r(t)}{\mu_i} \right)$. Note that the first-order term, $\bar{p}$, is common for all suppliers, and can be omitted while comparing suppliers' bids.

Proposition 2. (State Space Collapse) *Suppose $\pi_i + c\frac{\tilde{Q}_i^r(0)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(0)$ in probability, $\forall i \in \mathcal{N}$. Then, for all $\tau > 0$, for all $\epsilon > 0$, as $r \longrightarrow \infty$,*

$$\mathbb{P}\left\{ \sup_{0 \leq t \leq \tau} \max_{i,j \in \mathcal{N}} \left| \left(\pi_i + c\frac{\tilde{Q}_i^r(t)}{\mu_i} \right) - \left(\pi_j + c\frac{\tilde{Q}_j^r(t)}{\mu_j} \right) \right| > \epsilon \right\} \longrightarrow 0,$$

$$\mathbb{P}\left\{ \sup_{0 \leq t \leq \tau} \max_{i \in \mathcal{N}} \left| \left(\pi_i + c\frac{\tilde{Q}_i^r(t)}{\mu_i} \right) - \tilde{Z}^r(t) \right| > \epsilon \right\} \longrightarrow 0.$$

The proof applies the "hydrodynamic scaling" framework of Bramson [1998], introduced in the context of studying the heavy-traffic asymptotic behavior of multi-class queueing networks. Our

model falls outside the class of problems studied in Bramson [1998], but as we show in appendix, his analysis can be extended to address our setting in a fairly straightforward manner.

3.5 Limit model and discussion

Proposition 2 shows that the supplier behavior can be inferred by analyzing an appropriately defined one-dimensional process $\tilde{Z}^r(t)$ that is related to the aggregated market workload. This also implies that although each supplier only observes his own backlog, he is capable of inferring the backlog (or at least the full cost) of all other competing suppliers.

The next theorem characterizes the limiting behavior of the one-dimensional process $\tilde{Z}^r(t)$, and as a result also that of $\tilde{W}^r(t)$ and $\tilde{Q}^r(t)$.

Theorem 1. (Weak Convergence) *Suppose* $\pi_i + c\frac{\tilde{Q}_i^r(0)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(0)$ *in probability,* $\forall i \in \mathcal{N}$. *Then* $\tilde{Z}^r(t)$ *weakly converges to a reflected Ornstein-Unlenbeck process* $\tilde{Z}(t)$ *that satisfies*

$$\tilde{Z}(t) = \tilde{Z}(0) - \gamma c \int_0^t \tilde{Z}(s)ds + \tilde{U}(t) + \frac{c\sqrt{\sigma^2 + \hat{\mu}}}{\hat{\mu}} B(t), \tag{13}$$

where $\tilde{U}(0) = 0$, $\tilde{U}(t)$ *is continuous and nondecreasing, and* $\tilde{U}(t)$ *increases only when* $\tilde{Z}(t) = \hat{\pi} \equiv \max_{i\in\mathcal{N}} \pi_i$; $B(t)$ *is a standard Brownian motion. The parameters are* $\gamma = f(\bar{p})/\bar{F}(\bar{p})$, *and* $\sigma \equiv \sqrt{\sum_{i\in\mathcal{N}} \sigma_i^2}$. *In addition,* $\tilde{W}^r(t) \Rightarrow \frac{1}{\bar{c}}(\tilde{Z}(t) - \bar{\pi})$, *and* $\tilde{Q}_i^r(t) \Rightarrow \frac{\mu_i}{c}(\tilde{Z}(t) - \pi_i)$, $\forall i \in \mathcal{N}$.

The process $\tilde{U}(t)$ is the limiting process of $\tilde{U}^r(t) \equiv \frac{c}{\hat{\mu}} \sum_{i\in\mathcal{N}} \mu_i \tilde{Y}_i^r(t)$ (defined in the proof of Theorem 1), which can be regarded as the aggregate market idleness of the system.

This theorem characterizes the limiting marketplace behavior under a given price vector p. The market exhibits a form of "resource pooling" across suppliers. Given that $\tilde{Z}(t) \geq \hat{\pi}$, where $\hat{\pi} \equiv \max_{i\in\mathcal{N}} \pi_i$, it follows that $\tilde{W}(t) \geq \frac{1}{\bar{c}} \max_{i\in\mathcal{N}}(\pi_i - \bar{\pi}) := \zeta$. This says that unless all the suppliers submit the same price bid, the aggregate workload in the marketplace will always be strictly positive and some suppliers will never incur any idleness. The intuition for this result is the following. When the queue of the most expensive supplier(s) gets depleted, and this supplier(s) starts to idle, momentarily the imbalance between the aggregate arrival rate and service rate force suppliers to build up their queue lengths instantaneously. Consequently, suppliers that price below $\hat{\pi}$ never deplete their queue lengths asymptotically.

To summarize, in the limit model, the suppliers' queue length processes follow from $\tilde{Q}_i(t) =$

$\frac{\mu_i}{c}(\tilde{Z}(t) - \pi_i),\ \forall i \in \mathcal{N}$, where $\tilde{Z}(t)$ is defined through (13). The next section will use this result as an input to study the supplier pricing game.

A numerical example: To demonstrate the system dynamics, we consider a system with two $M/M/1$ servers, delay sensitivity parameter $c = 0.5$, and arrival rate of buyers $\Lambda = 1$. The valuation v of each customer is assumed to follow an exponential distribution with mean $\frac{1}{10}$. The aggregate and individual service rates are respectively $\hat{\mu} = e^{-1.3}$, $\mu_1 = 0.8\hat{\mu}$, and $\mu_2 = 0.2\hat{\mu}$. Moreover, suppose the price parameters are $\pi_1 = -1, \pi_2 = -2$, and therefore $\bar{\pi} = \frac{\pi_1+\pi_2}{2} = -1.5$. The workload trajectories for $r = 30$ and $r = 80$, depicted in Figures 1 and 2 respectively, show that the cumulative time that the workload process spends below the respective boundaries η^r is small; the above analysis proves that this time is asymptotically negligible.

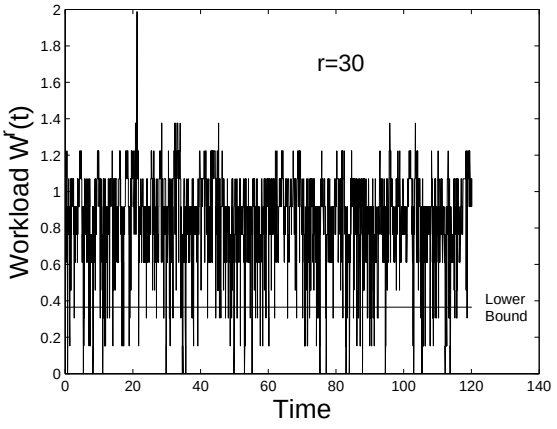

Figure 1: An instance of workload process when $r = 30$.

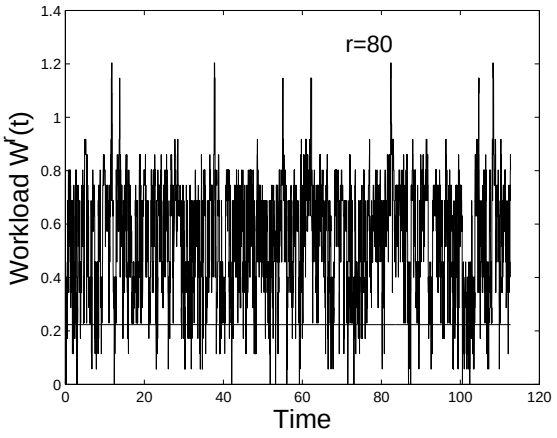

Figure 2: An instance of workload process when $r = 80$.

4 Competitive behavior and market efficiency

In this section, we discuss the equilibrium behavior of suppliers in the pricing game described in Section 2. We use the performance characterization of Section 3 to study the supplier pricing game. This is done by casting the suppliers' prices as "small" deviations around the market clearing price $\bar{p}$. This results in a pricing game with respect to these price deviations that is simplified due to the structural results of the previous section.

We first derive the suppliers' payoffs and focus on the analysis of the second-order revenues.

We then briefly discuss the centralized solution that maximizes the aggregate payoffs. Next, we characterize the non-cooperative behavior of suppliers under the competitive environment. Finally, we propose a coordination scheme that achieves the aggregate payoff under the centralized solution, and describe how this coordination scheme can be implemented in the original system.

4.1 Suppliers' payoffs

Let $R_i^r(t)$ denote supplier i's cumulative revenue. Since the pricing is static, supplier i earns $R_i^r(t) = (\bar{p} + \frac{\pi_i}{\sqrt{r}})S_i^r(t - Y_i^r(t))$, where $S_i^r(\cdot)$ is the counting service completion process, and $Y_i^r(t)$ is the cumulative idleness process for supplier i. The next lemma shows that the "first-order" revenues of the suppliers only depend on the first-order price term $\bar{p}$ and the service rates $\{\mu_i\}$'s.

Lemma 1. $\frac{R_i^r(t)}{r} \to \bar{p}\mu_i t$, *as* $r \to \infty, \forall i \in \mathcal{N}$.

The characterization in Lemma 1 is too crude, and oversimplifies the suppliers' decision making problem for purposes of studying their equilibrium behavior. To study the supplier pricing game we will focus on the second order correction around $R_i^r(t)$ defined as $r_i^r(t) \equiv \frac{1}{\sqrt{r}}(R_i^r(t) - r\bar{p}\mu_i t), \forall i \in \mathcal{N}$. The limiting processes of these corrected terms are characterized in the following lemma.

Lemma 2. $r_i^r(t) \Rightarrow \mu_i\pi_i t + \bar{p}\sigma_i B_{s,i}(t) - \mu_i\bar{p}\tilde{Y}_i(t)$, *as* $r \to \infty$, $\forall i \in \mathcal{N}$, *where* $\tilde{Y}_i(t)$ *is the limiting process of* $\tilde{Y}_i^r(t)$ *as* $r \to \infty$.

Define $r_i(t) := \mu_i\pi_i t + \bar{p}\sigma_i B_{s,i}(t) - \mu_i\bar{p}\tilde{Y}_i(t)$, $\forall i \in \mathcal{N}$. Instead of using the revenue functions $\{\Phi_i(p_i, p_{-i})\}$'s defined in (3), we will study the supplier pricing game based on their (second-order) revenues given by

$$\Psi_i(\pi_i, \pi_{-i}) \equiv \lim_{t\to\infty} \frac{r_i(t)}{t} = \mu_i(\pi_i - \bar{p}\mathbb{E}[\tilde{Y}_i(\infty)]), \tag{14}$$

where $\pi_{-i} \equiv (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$. Define h_i as the (steady-state) proportion of the market idleness incurred by supplier i, i.e.,

$$\mathbb{E}[\tilde{Y}_i(\infty)] = h_i\mathbb{E}[\tilde{U}(\infty)], \tag{15}$$

where $\mathbb{E}[\tilde{U}(\infty)]$ is the long-run average of the local time process of $\tilde{Z}(t)$ specified in Theorem 1.

Dividing (13) by t and letting $t \to \infty$, we obtain that $\mathbb{E}[\tilde{U}(\infty)] = \lim_{t\to\infty} \frac{\tilde{U}(t)}{t} = \gamma c\mathbb{E}[\tilde{Z}(\infty)] = \gamma c\beta\frac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)}$, where

$$\beta = (\frac{c(\hat{\mu}^2 + \sigma^2)}{2\gamma\hat{\mu}^2})^{1/2}, \tag{16}$$

and the closed-form expression follows from the fact that the reflected Orstein-Uhlenbeck process $\tilde{Z}(t)$ has the stationary distribution as a truncated Normal random variable ([Browne and Whitt, 2003, Proposition 1]).[6] Note that we do not derive the closed-form expressions of $\{h_i\}$'s and, as we will see below, they are not needed for our equilibrium analysis.

Let $J = \{j|\pi_j = \hat{\pi}\}$, where $\hat{\pi} \equiv \max_{i\in\mathcal{N}} \pi_i$, denote the set of the most expensive suppliers (allowing for ties). It is easy to see that $h_j = 0$, if $j \notin J$, and $\sum_{k\in J} \mu_k h_k = \frac{\hat{\mu}}{c}$. Therefore, from (14) and (15), we obtain the suppliers' second-order long-run average revenue functions as follows:

$$\Psi_i(\pi_i, \pi_{-i}) = \begin{cases} \mu_i\pi_i - \mu_i\bar{p}h_i\gamma c\beta\frac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)}, & \text{if } i \in J, \\ \mu_i\pi_i, & \text{otherwise.} \end{cases} \tag{17}$$

4.2 Centralized system performance

In the centralized version of the system, a central planner makes the price decisions $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_n)$ in order to maximize the total aggregated revenue:

$$\max \quad \left\{ \sum_{i\in\mathcal{N}} \mu_i\pi_i - \bar{p}\gamma\hat{\mu}\beta\frac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)} : \ \pi_i \le \hat{\pi} \right\}, \tag{18}$$

where we have applied $\sum_{i\in J} \mu_i h_i = \frac{\hat{\mu}}{c}$ to combine all the penalties imposed on the most expensive suppliers.

For convenience we define $\mathcal{L}(\hat{\pi}) \equiv \bar{p}\gamma\hat{\mu}\beta\frac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)}$ as the revenue loss that the system suffers if $\hat{\pi}$ is the highest price offered. The optimal pricing decisions are summarized in the following lemma:

Lemma 3. *In a centralized system, all prices π_is are equal. The optimal static price is $\pi^C := \arg\max_\pi[\hat{\mu}\pi - \mathcal{L}(\pi)]$.*

4.3 Competitive equilibrium

In a decentralized (competitive) system, each supplier is maximizing his own payoff, $\Psi_i(\pi_i, \pi_{-i})$, in a non-cooperative way: $\max_{\pi_i} \ \Psi_i(\pi_i, \pi_{-i})$. Recalling the definition of $\mathcal{L}(\hat{\pi})$, we can rewrite the supplier's payoff in (17) as

$$\Psi_i(\pi_i, \pi_{-i}) = \mu_i\pi_i - \mu_i h_i \frac{c}{\hat{\mu}}\mathcal{L}(\hat{\pi})\mathbb{1}\{i \in J\}] \tag{19}$$

[6] We can verify that using their notation, the $\tilde{Z}(t)$ process corresponds to the following parameters: $a = \gamma c, m = 0$, and the process has only a left reflecting barrier $\hat{\pi}$.

Although a standard approach is to look for a pure-strategy Nash equilibrium, the next result shows that none exists:

Lemma 4. *There exists no pure-strategy equilibrium for the supplier pricing game with the payoff function given in (19).*

The reason for not having any pure-strategy equilibrium is intuitively due to the discontinuity of suppliers' revenue functions. This creates an incentive for the cheap suppliers to increase their prices all the way to $\hat{\pi}$; however, they would also avoid to reach $\hat{\pi}$ when themselves become the most expensive and incur a discontinuous penalty.

The full characterization of the competitive equilibrium is provided in the online appendix, where we show that the suppliers randomize over certain range of prices in equilibrium, and each supplier receives an expected payoff lower than what he would obtain under the centralized solution. This implies that the centralized solution Pareto dominates all decentralized equilibria. Thus, implementing a coordination scheme results in no conflict of interests, even though the suppliers may be ex ante heterogeneous with respect to service rates. We also illustrate the competitive equilibrium behavior through representative numerical examples in the online appendix. We further show that as the market size grows, the competitive behavior among the suppliers may result in an unbounded efficiency loss. This demonstrates a significant inefficiency due to the market7 mechanism and motivates the search for a coordination scheme.

4.4 Coordination scheme

4.4.1 Sufficient condition for coordination

We will first study the suppliers' behavior if they were "forced" to share the penalty, or revenue loss, that arises due to the market idleness. Under this scheme, supplier i's payoff is

$$\Psi_i^{PS}(\pi_i, \pi_{-i}) \equiv \mu_i \pi_i - \frac{\mu_i}{\hat{\mu}} \mathcal{L}(\max_{j \in \mathcal{N}} \pi_j), \tag{20}$$

where the superscript PS refers to *penalty sharing* according to the service rates. The first term $\mu_i \pi_i$ is the gross revenue supplier i earns by serving customers, and the second term $\frac{\mu_i}{\hat{\mu}} \mathcal{L}(\max_{j \in \mathcal{N}} \pi_j)$ corresponds to his penalty share that is proportional to his service rate μ_i. Note that this scheme is budget-balanced, i.e., no financing from outside parties is required. Let $\left\{\pi_i^{PS}\right\}' s$ denote the

equilibrium prices under this sharing scheme. Under this sharing scheme, the centralized solution can be achieved.

Proposition 3. *Under the penalty sharing schemes that satisfy (20),* $\{\pi_i^{PS} = \pi^C, \forall i \in \mathcal{N}\}$ *is the unique equilibrium.*

Under the *PS* scheme, a supplier's objective is in fact an affine function of the aggregate revenue (18). Hence, this coordination mechanism eliminates the wrong incentives of suppliers, regardless of the number of suppliers and their service rates. Since in both competitive and the coordinated equilibria the suppliers' expected payoffs are proportional to their service rates, all suppliers have a natural incentive to join.

4.4.2 "Compensation-while-idling" mechanism that achieves coordination

The natural question is whether we can implement a penalty-sharing mechanism based on observable quantities. We now show that this is achievable through an appropriate set of transfer prices between suppliers when one or more suppliers are idling.

Let η_{ij} be the transfer price per unit of time from supplier i to supplier j when supplier j is idle in the limit model, with $\eta_{ii} = 0$. The second-order revenue process for a supplier i under this compensation scheme becomes

$$\tilde{r}_i^{PS}(t) = \tilde{r}_i(t) + \tilde{\delta}_i(t), \tag{21}$$

where

$$\tilde{r}_i(t) \equiv \mu_i \pi_i t + \bar{p}\sigma_i B_{s,i}(t) - \mu_i \bar{p}\tilde{Y}_i(t), \tag{22}$$

$$\tilde{\delta}_i(t) \equiv \sum_{j\in\mathcal{N}, j\neq i} \eta_{ji}\tilde{Y}_i(t) - \sum_{j\in\mathcal{N}, j\neq i} \eta_{ij}\tilde{Y}_j(t). \tag{23}$$

According to (22), the three terms correspond to his revenue by serving customers. In (23), $\tilde{\delta}_i(t)$ corresponds to the net transfers for supplier i : $\sum_{j\in\mathcal{N}, j\neq i} \eta_{ji}\tilde{Y}_i(t)$ is the compensation he receives from other suppliers during the idle period, and $\sum_{j\in\mathcal{N}, j\neq i} \eta_{ij}\tilde{Y}_j(t)$ is the cash outflow to other suppliers while compensating their idleness.

Given (21), supplier i's long-run average revenue can be expressed as

$$\begin{aligned}\tilde{\Psi}_i(\pi_i, \pi_{-i}) &\equiv \lim_{t\to\infty} \frac{1}{t}[\tilde{r}_i(t) + \tilde{\delta}_i(t)] \\ &= \mu_i\pi_i - \mu_i\bar{p}\mathbb{E}[\tilde{Y}_i(\infty)] + \sum_{j\in\mathcal{N}, j\neq i} \eta_{ji}\mathbb{E}[\tilde{Y}_i(\infty)] - \sum_{j\in\mathcal{N}, j\neq i} \eta_{ij}\mathbb{E}[\tilde{Y}_j(\infty)].\end{aligned} \tag{24}$$

The next proposition specifies a set of transfer prices that implement the PS scheme.

Proposition 4. *The transfer prices*

$$\eta_{ij} = \frac{\mu_i\mu_j}{\hat{\mu}}\bar{p},\ \forall i \neq j, i, j \in \mathcal{N}, \quad \textit{and} \quad \eta_{ii} = 0,\ \forall i \in \mathcal{N}, \tag{25}$$

implement the PS rule, i.e., $\tilde{\Psi}_i(\pi_i, \pi_{-i}) = \Psi_i^{PS}(\pi_i, \pi_{-i})$.

The transfer prices proposed in Proposition 4 essentially eliminate the imbalance between the current share of the market idleness incurred by an individual supplier and his required share ($\frac{\mu_i}{\hat{\mu}}\mathcal{L}(\max_{j\in\mathcal{N}}\pi_j)$). Given these transfer prices, every supplier's objective is aligned with the centralized system (i.e., the objective is $\Psi_i^{PS}(\pi_i, \pi_{-i})$ in (20)), and thus all suppliers are induced to set prices equal to π^C.

Proposition 4 shows that we are able to achieve coordination since given any chosen prices, we can align the suppliers' objectives with the planner's objective. To implement this compensation scheme in the original system, we can simply request each supplier make transfers according to (25). Note that the coordination scheme is independent of the the static prices $\{\pi_i\}$'s; it only requires the information of the service rates $\{\mu_i\}$'s.[7]

Our approximate analysis of the supplier game is internally consistent in the sense that the lower order price perturbations that essentially capture the supplier pricing game do not become unbounded that would indicate a desire to deviate from the operating regime that supports our limiting model. In essence, all suppliers choose to operate in the regime that we proposed in Section 3, both under the originally proposed market mechanism and the coordinating mechanism derived in this section.

[7] In this paper we do not directly address the issue of whether the suppliers will truthfully report their capacity/ service rates. In a high volume setting such as the one analyzed in this paper, the market maker observes the orders routed to each supplier, and can, indirectly through the supplier's delay bid, compute the supplier's capacity to within an error that will be of lower order than the relevant quantities. This allows the market maker to accurately set transfer prices as if she knew the true service rates.

5 Conclusion

In this paper we study an oligopolistic model in which suppliers compete for the buyers that are both price and delay averse. We apply both fluid and diffusion approximations to simplify the multi-dimensional characteristics of the decoupled suppliers into a single-dimensional aggregated problem. Specifically, we establish the "state space collapse" result in this system: the multi-dimensional queue length processes at the suppliers can be captured by a single-dimensional workload process of the aggregate supply in the market, which can be expressed explicitly as a reflected Ornstein-Unlenbeck process with analytical expressions. Based on this aggregated workload process, we derive the suppliers' long-run average revenues and show that the supplier competition results in a price randomization over certain ranges, whereas under the centralized control suppliers should set identical and deterministic prices.

To eliminate the inefficiency due to the competition, we propose a novel compensation-while-idling mechanism that coordinates the system: each supplier gets monetary transfers from other suppliers during his idle periods. This mechanism alters suppliers' objectives and implements the centralized solution at their own will. The implementation only requires a set of static transfer prices that are independent of the suppliers' prices and the queueing dynamics such as the current queue lengths or the cumulative idleness. Its simplicity is an appealing feature to be considered for practical implementations.

Appendix. Proofs of main results

This appendix gives the proofs of the main results, while proofs of technical lemmas introduced in the sequel are relegated to the online appendix.

Proof of Proposition 1

Define $\kappa_i(t) = \pi_i + c\frac{\bar{Q}_i(t)}{\mu_i}, \forall i \in \mathcal{N}$ as the total cost supplier i submits for the buyer at epoch t, and $\kappa(t) = [\kappa_1(t), ..., \kappa_n(t)]$. Let $\Upsilon(s) \subseteq \mathcal{N}$ be the set of suppliers tied as the cheapest at time s and $\Upsilon^c(s) \equiv \mathcal{N} \setminus \Upsilon(s)$ be its complement. By the continuity of $\{\kappa_i(t)\}$'s (κ_i is continuous in $\bar{Q}_i(t)$ and $\bar{Q}_i(t)$ is continuous in t), we have $\Upsilon(s) \subseteq \Upsilon(t),\ \forall s \leq t$. In words, once a supplier's proposed cost ties the lowest one, his cost stays the lowest.

Define the function

$$g(\kappa(t)) = \max_{i,j\in\mathcal{N}} (\kappa_i(t) - \kappa_j(t)) = \max_{i\in\mathcal{N}} \kappa_i(t) - \min_{j\in\mathcal{N}} \kappa_j(t).$$

Note that $g(\kappa(t))$ is nonnegative, and $g(\kappa(t)) = 0$ if and only if $\kappa_i(t) = \kappa_j(t),\ \forall i, j \in \mathcal{N}$. When $g(\kappa(t)) > 0$, the arrivals will not be routed to the most expensive supplier at time t, and the cheapest supplier accumulates customers. We will show that $g(\kappa(t))$ can be used as a Lyapunov function to prove that $|\kappa_i(t) - \kappa_j(t)| \to 0$ as $t \to \infty$, and subsequently conclude the statement of the proposition.

We now prove that there exists a lower bound for the depreciation rate of $g(\kappa(t))$. To this end, we will bound $\frac{d}{dt}\left(\max_{i\in\mathcal{N}} \kappa_i(t)\right)$ and $\frac{d}{dt}\left(\min_{j\in\mathcal{N}} \kappa_j(t)\right)$ separately. We first note that as long as not all suppliers receive new orders, the aggregate arrival rate, that equals $\sum_{j\in\mathcal{N}} \mu_j$, is always higher than the aggregate service rate of servers that tie as the cheapest ones. Recall that the orders are routed only to these cheapest suppliers according to (7). Thus, from the imbalance between the arrival and the service completion in (6) and (8), the queue lengths of these cheapest suppliers increase as time goes by. Moreover, once a supplier becomes the cheapest before the state space collapse occurs, his queue length grows unambiguously. On the contrary, those suppliers who do not receive any customer only depreciate their queues and decrease their $\{\kappa_i(t)\}$'s (according to the non-idling property (9)). This implies $\frac{d}{dt}\left(\min_{j\in\mathcal{N}} \kappa_j(t)\right) \geq 0$.

From the above discussions, it suffices to bound $\frac{d}{dt}\left(\max_{i\in\mathcal{N}} \kappa_i(t)\right)$. Let $k \in \Upsilon(t)$ denote one of the cheapest suppliers at epoch t. We then have

$$\dot{g}(\kappa(t)) \geq \dot{\kappa}_k(t) = \frac{c}{\mu_k} \frac{\mu_k}{\sum_{i\in\Upsilon(t)} \mu_i} \left[\sum_{i\in\mathcal{N}} \mu_i - \sum_{i\in\Upsilon(t)} \mu_j\right] \geq \frac{c}{\sum_{i\in\mathcal{N}} \mu_i - \min_{j\in\mathcal{N}} \mu_j} \left[\sum_{i\in\mathcal{N}} \mu_i - \sum_{i\in\Upsilon(t)} \mu_j\right],$$

where $\sum_{i\in\mathcal{N}} \mu_i - \sum_{i\in\Upsilon(t)} \mu_j$ is the imbalance between the aggregate arrival rate and aggregate service rate for those cheapest suppliers, and $\frac{\mu_k}{\sum_{i\in\Upsilon(t)} \mu_i}$ is the proportion of customers routed to supplier k. From $\sum_{i\in\mathcal{N}} \mu_i - \sum_{i\in\Upsilon(t)} \mu_j \geq \min_{j\in\mathcal{N}} \mu_j$ as long as $\Upsilon(t) \neq \mathcal{N}$, we then have $\dot{g}(\kappa(t)) \geq \frac{c \min_{j\in\mathcal{N}} \mu_j}{\sum_{i\in\mathcal{N}} \mu_i - \min_{j\in\mathcal{N}} \mu_j}$. Note that this lower bound is independent of epoch t and is strictly positive; the tie-breaking rule does not matter due to the continuity of queue length processes.

Let $\triangle\kappa$ be such that $\pi_l + c\frac{\bar{Q}_l(0)}{\mu_l} + \triangle\kappa = \max_{i\in\mathcal{N}} \{\pi_i + c\frac{\bar{Q}_i(0)}{\mu_i}\}$. In words, $\triangle\kappa$ is the amount of cost that supplier l has to add so that his $\kappa_l(0)$ ties the maximum initial cost. Because the imbalance of proposed costs is decreasing in t at a guaranteed rate, the $\kappa_i(t)$'s will become equal before time $s^*(\bar{Q}(0), \bar{W}(0)) = \frac{\triangle\kappa(\sum_{i\in\mathcal{N}} \mu_i - \min_{j\in\mathcal{N}} \mu_j)}{c \min_{j\in\mathcal{N}} \mu_j}$. Thus, $g(\kappa(t)) = 0,\ \forall t \geq s^*$.

The next step needed is to bound $\triangle\kappa$ as a function of M_0. Call it $\triangle(M_0)$. Then for any $\delta > 0$, $s^*(\delta, M_0) \leq s^*(M_0) = \frac{\triangle(M_0)(\sum_{i\in\mathcal{N}}\mu_i - \min_{j\in\mathcal{N}}\mu_j)}{c\min_{j\in\mathcal{N}}\mu_j}$. This completes the proof for the convergence of fluid model. □

Proof of Proposition 2

The proof builds on the framework advanced by Bramson [1998]. To facilitate exposition we will follow Bramson [1998] very closely, stating and addressing the differences and the required modifications.

A brief sketch of the proof is as follows. Step 1. *Hydrodynamic limits:* Lemmas 5-9 will establish that appropriately scaled processes that focus on the system behavior over order $\frac{1}{\sqrt{r}}$ time intervals satisfy the fluid equations (6)-(11). Step 2. *Convergence of diffusion scaled processes:* Combining the above with Proposition 1, we will establish that the supplier queue length process is close to the "balanced" state configuration for all time t; i.e., the short transient digressions are not visible in the natural time scale of the system.

Step 1:

The hydrodynamic scaling we adopt here is the following (cf. [Bramson, 1998, Equation (5.3)]): For $m = 1, 2, ..., \lfloor\sqrt{r}\tau\rfloor$,

$$\begin{aligned}
Q^{r,m}(t) &= \frac{1}{\sqrt{r}}Q^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m),\\
A^{r,m}(t) &= \frac{1}{\sqrt{r}}[A^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) - A^r(\frac{1}{\sqrt{r}}m)],\\
D^{r,m}(t) &= \frac{1}{\sqrt{r}}[D^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) - D^r(\frac{1}{\sqrt{r}}m)],\\
T^{r,m}(t) &= T^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) - T^r(\frac{1}{\sqrt{r}}m),\\
Y^{r,m}(t) &= Y^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) - Y^r(\frac{1}{\sqrt{r}}m),\\
X^{r,m}(\cdot) &= (Q^{r,m}(\cdot), A^{r,m}(\cdot), D^{r,m}(\cdot), T^{r,m}(\cdot), Y^{r,m}(\cdot)),
\end{aligned}$$

where Q^r, A^r, D^r, T^r, and Y^r are all n-dimensional vectors. Note that for cumulative processes $A^r(\cdot), D^r(\cdot), T^r(\cdot)$, and $Y^r(\cdot)$, the associated hydrodynamic scale processes $A^{r,m}(t), D^{r,m}(t)$, and $T^{r,m}(t), Y^{r,m}(t)$, account for the cumulative changes between $[\frac{1}{\sqrt{r}}m, \frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m]$. On the contrary, $Q^{r,m}(t)$ only keeps record of the values at epoch $\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m$. The scaling $\frac{1}{\sqrt{r}}$ is used to ensure that the scaled processes admit meaningful limits as $r \to \infty$.

We now analyze the processes introduced above and provide probabilistic bounds. This essentially is equivalent to [Bramson, 1998, Proposition 5.1]. For ease of notation, let us define $Q^{r,m}(t) = \{Q_i^{r,m}(t)\}$. The routing indicator is $\mathbf{I}_i(Q^{r,m}(t)) = \begin{cases} 1 & \text{if } \Xi_{\mathcal{J}(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)} = i, \\ 0 & \text{otherwise.} \end{cases}$. Note that $\sqrt{r}Q_i^{r,m}(t) = Q_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)$ is the queue length of supplier i at epoch $\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m$. Therefore,

$$\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} = \bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{Q_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)+1}{\mu_i^r}$$

is simply the total cost submitted by supplier i since $\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_i^r}$ is the delay quotation. We further define

$$\mathbf{J}_i^r(Q_i^{r,m}(t)) = \mathbb{P}\left\{v \geq \bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i}\right\}, \forall i \in \mathcal{N}$$

as the probability that at time epoch $\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m$, the valuation of a new arrival, v, exceeds the total cost submitted by supplier i. The following lemma establishes the probability bounds.

Lemma 5. *Fix $\epsilon > 0$, $L > 0$, and $\tau > 0$. For r large enough,*

$$\mathbb{P}\left\{\max_{m<\sqrt{r}\tau}\max_{i\in\mathcal{N}}||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L > \epsilon\right\} \leq \epsilon, \tag{26}$$

$$\mathbb{P}\left\{\max_{m<\sqrt{r}\tau}\max_{i\in\mathcal{N}}||D_i^{r,m}(t) - \mu_i T_i^{r,m}(t)||_L > \epsilon\right\} \leq \epsilon, \tag{27}$$

$$\mathbb{P}\left\{\max_{m<\sqrt{r}\tau}\max_{i,j\in\mathcal{N}}||\left(\pi_i + c\frac{Q_i^{r,m}(t)}{\mu_i}\right) - \left(\pi_j + c\frac{Q_j^{r,m}(t)}{\mu_j}\right)||_L > \epsilon\right\} \leq \epsilon. \tag{28}$$

The next lemma shows that the hydrodynamic processes are "nearly Lipschitz continuous."

Lemma 6. *Fix $\epsilon > 0$, $L > 0$, and $\tau > 0$. There exists a constant N_0 such that for r large enough,*

$$\mathbb{P}\left\{\max_{m<\sqrt{r}\tau}\sup_{t_1,t_2\in[0,L]}|X^{r,m}(t_2) - X^{r,m}(t_1)| > N_0|t_2 - t_1| + \epsilon\right\} \leq \epsilon. \tag{29}$$

The above two lemmas imply that the measure of "ill-behaved" events is negligible for the hydrodynamic scale processes. Let N_0 denote the constant required in Lemma 6 and focus on the complement of these ill-behaved events. We can choose a sequence $\epsilon(r)$ decreasing to 0 sufficiently

slowly such that the inequalities (28), (27), and (29) still hold. For such a sequence $\epsilon(r)$, we define

$$
\begin{aligned}
K_0^r &= \left\{ \max_{m<\sqrt{r}\tau} \sup_{t_1,t_2\in[0,L]} |X^{r,m}(t_2) - X^{r,m}(t_1)| \le N_0|t_2 - t_1| + \epsilon(r) \right\}, \\
K_1^r &= \left\{ \begin{array}{c} \max_{m<\sqrt{r}\tau} \max_{i\in\mathcal{N}} ||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L \le \epsilon(r); \\ \max_{m<\sqrt{r}\tau} \max_{i\in\mathcal{N}} ||D_i^{r,m}(t) - \mu_i T_i^{r,m}(t)||_L \le \epsilon(r) \\ \max_{m<\sqrt{r}\tau} \max_{i,j\in\mathcal{N}} || \left(\pi_i + c\frac{Q_i^{r,m}(t)}{\mu_i}\right) - \left(\pi_j + c\frac{Q_j^{r,m}(t)}{\mu_j}\right) ||_L \le \epsilon(r); \end{array} \right\}, \\
K^r &= K_0^r \cap K_1^r.
\end{aligned}
$$

Note that the set K^r contains those events that possess our desired properties. The next step is to show that when r is sufficiently large, we can restrict our attention to these well-behaved events. This is parallel to [Bramson, 1998, Proposition 5.2, Corollary 5.1]. and follows from Lemmas 5 and 6.

Lemma 7. $\mathbb{P}(K^r) \to 1$, *as* $r \to \infty$.

Let $\mathbb{E}$ be the space of functions $x : [0, L] \to \mathbb{R}^5$ that are right continuous and have left limits. Define $\mathbb{E}' = \{x \in \mathbb{E} : |x(0)| < M_0, |x(t_2) - x(t_1)| < N_0|t_2 - t_1|, \forall t_1, t_2 \in [0, L]\}$, where M_0 is a fixed constant. Let $\mathbb{E}_0^r = \{X^{r,m}(\cdot, \omega), m < \sqrt{r}\tau, \omega \in K^r\}$ denote the sets of well-behaved events and $\mathbb{E}_0 = \{\mathbb{E}_0^r, r \in \mathbb{R}_+\}$ is the collection of these sets. We next show that the set of candidate hydrodynamic limits is "dense" in the state space: when r is sufficiently large, the vector of processes $X^{r,m}(\cdot, \omega)$ is close to some cluster point of $\mathbb{E}_0$. A function $\widehat{X}$ is said to be a cluster point of the functional space $\mathbb{E}_0$ if for all $\delta > 0$, there exists a $X_\delta \in \mathbb{E}_0$ such that $||\widehat{X} - X_\delta||_L < \delta$.

Lemma 8. *Fix* $\epsilon > 0, L > 0$, *and* $\tau > 0$. *There exists a sufficiently large* $r(\epsilon)$ *such that for all* $r > r(\epsilon)$, *for all* $\omega \in K^r$ *and for all* $m = 1, 2, ..., \lfloor\sqrt{r}\tau\rfloor$, $||X^{r,m}(\cdot, \omega) - \widehat{X}(\cdot)||_L < \epsilon$, *for some* $\widehat{X}(\cdot) \in E_0 \cap E'$ *with* $\widehat{X}(\cdot)$ *being a cluster point of* E_0.

The above lemma follows closely from [Bramson, 1998, Proposition 6.1]. Given that all hydrodynamic scale processes have a close-by cluster point, it remains to study the behavior of these cluster points. The next step proves that all these cluster points are in fact solutions to the deterministic fluid equations (6)-(11) (cf. [Bramson, 1998, Proposition 6.2]).

Lemma 9. *Fix* $L > 0$ *and* $\tau > 0$. *Let* $\widehat{X}(\cdot)$ *be an arbitrary cluster point of* E_0 *over* $[0, L]$. *Then* $\widehat{X}(\cdot)$ *satisfies the fluid equations (6)-(10).*

Step 2:

Combining the above results with Proposition 1, we will now establish the desired state space collapse property for the supplier queue length processes. Let us first fix constants $\eta, \xi, \epsilon > 0$. By Lemma 7, there exists a sufficiently large $r(\xi) > 0$ such that $P(K^r) > 1 - \xi$ for all $r > r(\xi)$.

Take $L = s(\epsilon, M_0) + 1$ where $s(\cdot, \cdot)$ was defined in Proposition 1. Let $r(\eta)$ be sufficiently large such that $\frac{L}{\sqrt{r}} < \eta$ whenever $r > r(\eta)$. Now we consider the diffusive scaled processes in the time interval $[\frac{L}{\sqrt{r}}, \tau]$. For all $\varsigma \in [\frac{L}{\sqrt{r}}, \tau]$, let $m_r(\varsigma) = \min\{m \in \mathbb{N}_+ : m < \sqrt{r}\varsigma < m + L\} = \max\{\lceil \sqrt{r}\varsigma - L \rceil, 0\}$ and $\tau_r'(\varsigma) := \sqrt{r}\varsigma - m_r(\varsigma)$. Thus, $\varsigma = \frac{1}{\sqrt{r}}\left[\tau_r'(\varsigma) + m_r(\varsigma)\right]$. Straight-forward algebra shows that $\tilde{Q}_i^r(\varsigma) = \frac{1}{\sqrt{r}} Q_i^r(\varsigma) = \frac{1}{\sqrt{r}} Q_i^r\left(\frac{1}{\sqrt{r}}\tau_r'(\varsigma) + \frac{1}{\sqrt{r}} m_r(\varsigma)\right) = Q_i^{r, m_r(\varsigma)}(\tau_r'(\varsigma))$. From the definition of $\tau_r'(\varsigma)$, if $r > r(\eta)$, for all $\varsigma \in [\frac{L}{\sqrt{r}}, \tau]$, we have

$$\tau_r'(\varsigma) = \sqrt{r}\varsigma - m_r(\varsigma) = \sqrt{r}\varsigma - \max\{\lceil \sqrt{r}\varsigma - L \rceil, 0\} \geq \sqrt{r}\varsigma - (\sqrt{r}\varsigma - L - 1) = L - 1 = s(\epsilon, M_0).$$

This implies that the convergence of fluid scale process can be applied at time $\varsigma \in [\frac{L}{\sqrt{r}}, \tau]$ when $r > r(\eta)$. Moreover, $m_r(\varsigma) \leq \sqrt{r}\varsigma \leq \sqrt{r}\tau$ by construction. Therefore, the convergence of hydrodynamic scale processes is valid here for all $\varsigma \in [\frac{L}{\sqrt{r}}, \tau]$ as well.

We now verify that the imbalance between $\pi_i + c\frac{\tilde{Q}_i^r(\varsigma)}{\mu_i}$ and $\pi_i + c\frac{\tilde{Q}_i^r(\varsigma)}{\mu_i}$ is upper bounded (note that the additional terms $\{\frac{1}{\mu_i^r}\}$'s in the suppliers' bids have been taken into account in Lemma 5 through (28)):

$$\begin{aligned}
& \max_{i,j \in \mathcal{N}} \left| \left(\pi_i + c\frac{\tilde{Q}_i^r(\varsigma)}{\mu_i}\right) - \left(\pi_j + c\frac{\tilde{Q}_j^r(\varsigma)}{\mu_j}\right) \right| \\
\leq & \max_{i,j \in \mathcal{N}} \left\{ \begin{array}{r} \left| \left(\pi_i + c\frac{Q_i^{r,m_r(\varsigma)}(\tau_r'(\varsigma))}{\mu_i}\right) - \left(\pi_i + c\frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i}\right) \right| + \left| \left(\pi_i + c\frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i}\right) - \left(\pi_j + c\frac{\widehat{Q}_j(\tau_r'(\varsigma))}{\mu_j}\right) \right| \\ + \left| \left(\pi_j + c\frac{\widehat{Q}_j(\tau_r'(\varsigma))}{\mu_j}\right) - \left(\pi_j + c\frac{Q_j^{r,m_r(\varsigma)}(\tau_r'(\varsigma))}{\mu_j}\right) \right| \end{array} \right\} \\
\leq & 2 \max_{i \in \mathcal{N}} \left| \left(\pi_i + c\frac{Q_i^{r,m_r(\varsigma)}(\tau_r'(\varsigma))}{\mu_i}\right) - \left(\pi_i + c\frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i}\right) \right| \\
& + \max_{i,j \in \mathcal{N}} \left| \left(\pi_i + c\frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i}\right) - \left(\pi_j + c\frac{\widehat{Q}_j(\tau_r'(\varsigma))}{\mu_j}\right) \right|.
\end{aligned}$$

From Lemmas 5 and 8, for fixed L, τ, ξ, η, and $\epsilon > 0$, there exists a sufficiently large $r(L, \tau, \xi, \eta, \epsilon) > \max\{r(\xi), r(\eta)\}$ such that for all $r > r(L, \tau, \xi, \eta, \epsilon)$,$\omega \in K^r$, for all $\varsigma \in [\frac{L}{\sqrt{r}}, \tau]$, we have

$$\max_{i \in \mathcal{N}} \left| \left(\pi_i + c\frac{Q_i^{r,m_r(\varsigma)}(\tau_r'(\varsigma))}{\mu_i}\right) - \left(\pi_i + c\frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i}\right) \right| \leq \frac{\epsilon}{3},$$

and $\max_{i,j\in\mathcal{N}} \mid \left(\pi_i + c\frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i}\right) - \left(\pi_j + c\frac{\widehat{Q}_j(\tau_r'(\varsigma))}{\mu_j}\right) \mid \leq \frac{\epsilon}{3}$. This implies that $\max_{i,j\in\mathcal{N}} \mid \left(\pi_i + c\frac{\tilde{Q}_i^r(\varsigma)}{\mu_i}\right) - \left(\pi_j + c\frac{\tilde{Q}_j^r(\varsigma)}{\mu_j}\right) \mid \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$. Note that $P(K^r) \to 1$ as $r \to \infty$, i.e., the state space collapse holds in probability (the additional term $\frac{1}{\mu_i^r}$ is of order $\frac{1}{r}$ and therefore does not affect the result as demonstrated in Lemma 5). We can then choose appropriate $L, \tau, \xi,\ \eta$, and ϵ such that the proposition is established. □

Proof of Theorem 1

We start with a sketch of the proof, and then provide the details for each part.

Step 1. Write down the equation of $W^r(t)$. Recall that the workload process is $W^r(t) = \sum_{i\in\mathcal{N}} \frac{Q_i^r(t)}{\mu_i^r}$, where $Q_i^r(t) = Q_i^r(0) + A_i^r(t) - S_i^r(T_i^r(t))$, and the scaled workload process $\tilde{W}^r(t) = \sqrt{r}W^r(t)$. We first apply strong approximations ([Glynn, 1990, Theorem 5]) on the cumulative arrival process routed to each supplier and the service requirement processes, and then the state space collapse result to derive the expression of $W^r(t)$ as a function of the total arrival process $A^r(t)$.

Step 2. Fluid model properties of cumulative idleness and aggregate market demand. Let $Y_i^r(t) = t - T_i^r(t)$ denote the market idleness, where $T_i^r(t)$ is the cumulative work completion for supplier i up to time t, and $\Lambda^r(t)$ be the aggregate arrival rate into the market. We will prove that $Y_i^r(t) \to 0$, in probability, u.o.c., $\forall i \in \mathcal{N}$, and $\frac{\Lambda^r(t)}{r} \to (\sum_{i\in\mathcal{N}} \mu_i)t$, *u.o.c.*

Step 3. Use strong approximations and oscillation inequalities to bound $\tilde{U}^r(t)$ and $\tilde{W}^r(t)$. In this step, we approximate $\tilde{W}^r(t)$ using the reflection maps (φ, ψ) (similar to the framework in Williams [1998]) and apply Lemma 7 in Ata and Kumar [2005] to bound the market idleness $\tilde{U}^r(t)$ and workload $\tilde{W}^r(t)$.

Step 4. Establish the weak convergence of $\tilde{Z}^r(t)$. Having established the convergence for individual idleness process, we now focus on the process $\tilde{W}^r(t)$. We will follow Mandelbaum and Pats [1995] and use Gronwall's inequality to bound the difference of $\tilde{W}^r(t)$ from another process that has the "desirable" limit, and then apply convergence together lemma to establish the weak convergence. The weak convergence of $\tilde{Q}^r(t)$ follows immediately from the state space collapse result and the convergence together lemma.

Complete Proof of Theorem 1

Step 1:

From Proposition 2, we have $\pi_i + c\frac{\tilde{Q}_i^r(t)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(t) + o_p(1),\ \forall i \in \mathcal{N}$. In the sequel we can simply focus on those events in which state space collapse arises. Recall that the queue length process can be rewritten as $Q_i^r(t) = Q_i^r(0) + A_i^r(t) - S_i^r(t - Y_i^r(t))$, where $S_i^r(\cdot)$ is the cumulative service completions when the underlying random service times are i.i.d. with mean $\frac{1}{r\mu_i}$ and standard deviation $\frac{\sigma_i}{r}$. Then,

$$\begin{aligned} & \pi_i + c\frac{\tilde{Q}_i^r(t)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(t) + o_p(1), \\ \Rightarrow \quad & \pi_i + c\frac{\tilde{Q}_i^r(0)}{\mu_i} + c\frac{A_i^r(t)}{\sqrt{r}\mu_i} - \frac{cS_i^r(t - Y_i^r(t))}{\sqrt{r}\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(t) + o_p(1). \end{aligned}$$

Rearranging the above equation, we see that

$$\frac{A_i^r(t)}{\sqrt{r}} = \frac{\mu_i}{c}[\bar{\pi} + \bar{c}\tilde{W}^r(t)] + \frac{S_i^r(t - Y_i^r(t))}{\sqrt{r}} - \frac{\mu_i\pi_i}{c} - \tilde{Q}_i^r(0) + o_p(1),$$

and therefore the aggregate market demand $A^r(t) \equiv \sum_i A_i^r(t)$ satisfies

$$\frac{A^r(t)}{\sqrt{r}} = \frac{\sum_{i\in\mathcal{N}}\mu_i}{c}[\bar{\pi} + \bar{c}\tilde{W}^r(t)] + \sum_{i\in\mathcal{N}}\frac{S_i^r(t - Y_i^r(t))}{\sqrt{r}} - \frac{\sum_{i\in\mathcal{N}}\mu_i\pi_i}{c} - \sum_{i\in\mathcal{N}}\tilde{Q}_i^r(0) + o_p(1). \tag{30}$$

Next we focus on the demand and service time processes. A strong approximation ([Glynn, 1990, Theorem 5]) for the Poisson process $N(\cdot)$ allows us to rewrite $A^r(t) = N(\Lambda^r(t))$ as

$$A^r(t) = \Lambda^r(t) + B_a(\Lambda^r(t)) + O(\log rt), \tag{31}$$

where $\Lambda^r(t) = \int_0^t \lambda^r(\min_{i\in\mathcal{N}}[\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}\tilde{Q}_i^r(s)+1}{r\mu_i}])ds$ is the arrival rate of aggregate market demand, and $B_a(\cdot)$ is a standard Brownian motion and the subscript "a" stands for "arrival."

Similarly, a strong approximation of the renewal process $S_i^r(\cdot)$ gives that

$$S_i^r(t - Y_i^r(t)) = \mu_i^r t - \mu_i^r Y_i^r(t) + \sqrt{r}\sigma_i B_{s,i}(t - Y_i^r(t)) + O(\log r[t - Y_i^r(t)]), \tag{32}$$

where $B_{s,i}(\cdot)$ is the standard Brownian motion associated with supplier i's service time distribution, and the subscript "s" stands for "service." Note that the service rate is of order r, and therefore the Brownian motion is scaled by $\sqrt{r}$.

From Proposition 2 we have that $\min_{i\in\mathcal{N}}\left\{\ \pi_i + c\frac{\tilde{Q}_i^r(t)}{\mu_i} + c\frac{1}{\mu_i^r}\right\} = \bar{\pi} + \bar{c}\tilde{W}^r(t) + o_p(1),\ \forall t$, and $P(\frac{\tilde{W}^r(t)}{\sqrt{r}} > 0) \to 0$, uniformly in t. Therefore, the arrival rate can be expressed as

$$\lambda^r\left(\min_{i\in\mathcal{N}}\left\{\bar{p} + \ \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}\tilde{Q}_i^r(s)+1}{r\mu_i}\right\}\right) = \lambda^r\left(\bar{p} + \frac{\bar{\pi}+\bar{c}\tilde{W}^r(s)}{\sqrt{r}} + o_p(\frac{1}{\sqrt{r}})\right)$$

$$= \lambda^r(\bar{p}) - \frac{\lambda^r(\bar{p})'}{\sqrt{r}}[\bar{\pi}+\bar{c}\tilde{W}^r(s)] + o_p(\sqrt{r}) = \sum_{i\in\mathcal{N}}\mu_i^r - \sqrt{r}(\sum_{i\in\mathcal{N}}\mu_i)\frac{f(\bar{p})}{\bar{F}(\bar{p})}[\bar{\pi}+\bar{c}\tilde{W}^r(s)] + o_p(\sqrt{r}),$$

where the second equality follows from a first-order Taylor expansion at $p=\bar{p}$ and the last equality follows from the fact that $\bar{p}$ induces full resource utilization and the definition of $\lambda^r(\cdot)$. With this expression, the cumulative rate of aggregate market demand $\Lambda^r(t)$ becomes

$$\Lambda^r(t) = (\sum_{i\in\mathcal{N}}\mu_i^r)t - \sqrt{r}(\sum_{i\in\mathcal{N}}\mu_i)\frac{f(\bar{p})}{\bar{F}(\bar{p})}\int_0^t[\bar{\pi}+\bar{c}\tilde{W}^r(s)]ds + o_p(\sqrt{r}t). \tag{33}$$

Combining (30), (31), (32), and (33), and for $\gamma \equiv \frac{f(\bar{p})}{\bar{F}(\bar{p})}$, we obtain

$$\frac{1}{\sqrt{r}}\left\{(\sum_{i\in\mathcal{N}}\mu_i^r)t - (\sum_{i\in\mathcal{N}}\mu_i)\gamma\int_0^t[\bar{\pi}+\bar{c}\tilde{W}^r(s)]ds\right\} + \frac{B_a(\Lambda^r(t))}{\sqrt{r}} + o_p(1)$$
$$= \frac{\sum_{i\in\mathcal{N}}\mu_i}{c}[\bar{\pi}+\bar{c}\tilde{W}^r(t)] + \frac{\sum_{i\in\mathcal{N}}\mu_i^r}{\sqrt{r}}t - \frac{\sum_{i\in\mathcal{N}}\mu_i^r}{\sqrt{r}}Y_i^r(t) + \sum_{i\in\mathcal{N}}\sigma_i B_{s,i}(t-Y_i^r(t)) - \frac{\sum_{i\in\mathcal{N}}\mu_i\pi_i}{c} - \sum_{i\in\mathcal{N}}\tilde{Q}_i^r(0).$$

We can rearrange the equation as follows:

$$\frac{\sum_{i\in\mathcal{N}}\mu_i}{c}[\bar{\pi}+\bar{c}\tilde{W}^r(t)] = -(\sum_{i\in\mathcal{N}}\mu_i)\gamma\int_0^t[\bar{\pi}+\bar{c}\tilde{W}^r(s)]ds + \sqrt{r}\sum_{i\in\mathcal{N}}\mu_i Y_i^r(t) + \frac{\sum_{i\in\mathcal{N}}\mu_i\pi_i}{c} + \sum_{i\in\mathcal{N}}\tilde{Q}_i^r(0)$$
$$+ \frac{B_a(\Lambda^r(t))}{\sqrt{r}} - \sum_{i\in\mathcal{N}}\sigma_i B_{s,i}(t-Y_i^r(t)) + o_p(1).$$

Recall that $\tilde{Z}^r(t) = \bar{\pi} + \bar{c}\tilde{W}^r(t)$, $\hat{\mu} \equiv \sum_{i\in\mathcal{N}}\mu_i$, and $\tilde{Y}_i^r(t) = \sqrt{r}Y_i^r(t), \forall i \in \mathcal{N}$. From the assumption of initial queue lengths $\pi_i + c\frac{\tilde{Q}_i^r(0)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(0), \forall i\in\mathcal{N}$, we obtain $\tilde{Q}_i^r(0) = \frac{\mu_i}{c}(\tilde{Z}^r(0) - \pi_i)$, and therefore that $\sum_{i\in\mathcal{N}}\tilde{Q}_i^r(0) = \frac{\hat{\mu}}{c}\tilde{Z}^r(0) - \frac{\sum_{i\in\mathcal{N}}\mu_i\pi_i}{c}$.

Using these substitutions, we get that

$$\frac{\hat{\mu}}{c}\tilde{Z}^r(t) = -\hat{\mu}\gamma\int_0^t\tilde{Z}^r(s)ds + \sum_{i\in\mathcal{N}}\mu_i\tilde{Y}_i^r(t) + \tilde{Z}(0) + [\frac{B_a(\Lambda^r(t))}{\sqrt{r}} - \sum_{i\in\mathcal{N}}\sigma_i B_{s,i}(t-Y_i^r(t))] + o_p(1).$$

Defining $\tilde{U}^r(t) = \frac{c}{\hat{\mu}}\sum_{i\in\mathcal{N}}\mu_i\tilde{Y}_i(t)$ as the "total market idleness," we conclude that

$$\tilde{Z}^r(t) = \tilde{Z}(0) - \gamma c\int_0^t\tilde{Z}^r(s)ds + \tilde{U}^r(t) + \frac{c}{\hat{\mu}}\left[\frac{B_a(\Lambda^r(t))}{\sqrt{r}} - \sum_{i\in\mathcal{N}}\sigma_i B_{s,i}(t-Y_i^r(t))\right] + o_p(1). \tag{34}$$

Step 2:

In this section, we will establish the convergence of scaled copies of $\Lambda^r(t)$ and $Y_i^r(t)$ according to

$$\frac{\Lambda^r(t)}{r} \to \hat{\mu}t, \text{ in probability, } u.o.c. \tag{35}$$

$$Y_i^r(t) \to 0, \text{in probability, } u.o.c., \forall i \in \mathcal{N}. \tag{36}$$

First we prove (35). From (33), we have $\frac{\Lambda^r(t)}{r} = \hat{\mu}t - \frac{\hat{\mu}\gamma}{\sqrt{r}}\int_0^t \tilde{Z}^r(s)ds + o_p(\frac{1}{\sqrt{r}})$, where the last term vanishes in the limit. From Proposition 2, we know that for any fixed constant C, $P(\sup_{t\le T}\tilde{W}^r(t) > C) \to 0$, and hence $P(\sup_{t\le T}\frac{\tilde{W}^r(t)}{\sqrt{r}} > \epsilon) \to 0$, $\forall \epsilon > 0$. Moreover, $\frac{\hat{\mu}\gamma}{\sqrt{r}}\int_0^t \tilde{Z}^r(s)ds = \frac{\hat{\mu}\gamma}{\sqrt{r}}[\bar{\pi}t + c\int_0^t \tilde{W}^r(s)ds] \to 0$,in probability, u.o.c., as $r \to \infty$, because the first term inside the brackets is a constant and the integral is uniformly bounded by $\epsilon\sqrt{r}$. Therefore, $\frac{\Lambda^r(t)}{r} \to \hat{\mu}t$, in probability, *u.o.c.*, which completes the proof of (35).

Next we show (36) by contradiction. Suppose that there exists a supplier i such that $Y_i^r(t) \nrightarrow 0$ u.o.c., then we can find a constant $\epsilon_1 > 0$ and a sequence (r_k, t_k) such that $r_k \to \infty$ as $k \to \infty$ and $Y_i^{r_k}(t_k) > \epsilon_1$, $\forall k$. Along this sequence, due to the nonnegativity of $\tilde{Y}_j^r(t)$'s and the scaling $\sqrt{r}$ for $\tilde{Y}_i^r(t)$, we obtain $\tilde{U}^{r_k}(t_k) > \sqrt{r_k}(c\mu_i\epsilon_1/\hat{\mu})$. Thus, as $k \to \infty$, $\tilde{U}^{r_k}(t_k) \to \infty$. From (34), this implies that $\tilde{Z}^{r_k}(t_k) \to \infty$, and therefore that $\tilde{W}^{r_k}(t_k) \to \infty$. This, however, contradicts the fact that for all $T > 0$, $\epsilon > 0$, there exists a constant C such that $P(\sup_{t\le T} W^{\tilde{r}}(t) > C) < \epsilon$. This completes the proof of (36).

Step 3:

We can rewrite (34) as $\tilde{Z}^r(t) = \tilde{Z}^r(0) - \gamma c\int_0^t \tilde{Z}^r(s)ds + \tilde{U}^r(t) + \tilde{V}^r(t)$, where $\tilde{V}^r(t) = \frac{c}{\hat{\mu}}[\frac{B_a(\Lambda^r(t))}{\sqrt{r}} - \sum_{i\in\mathcal{N}}\sigma_i B_{s,i}(t - Y_i^r(t))]$. Recalling that $\frac{\Lambda^r(t)}{r} \to \hat{\mu}t$ and $Y_i^r(t) \to 0$, we can apply the invariance principle of Brownian motion and convergence together lemma to obtain

$$\frac{B_a(\Lambda^r(t))}{\sqrt{r}} \stackrel{D}{=} B_a(\frac{\Lambda^r(t)}{r}) \Rightarrow B_a(\hat{\mu}t) \stackrel{D}{=} \sqrt{\hat{\mu}}B_a(t),$$

$$\sum_{i\in\mathcal{N}}\sigma_i B_{s,i}(t - Y_i^r(t)) \Rightarrow \sum_{i\in\mathcal{N}}\sigma_i B_{s,i}(t) \stackrel{D}{=} \sigma B_s(t),$$

where $\sigma \equiv \sqrt{\sum_{i\in\mathcal{N}}\sigma_i^2}$ and B_a, B_s are independent standard Brownian motions. Thus, $\tilde{V}^r(t) \Rightarrow \frac{c}{\hat{\mu}}[\sqrt{\hat{\mu}}B_a(t) - \sigma B_s(t)] \stackrel{D}{=} \frac{c}{\hat{\mu}}\sqrt{\sigma^2 + \hat{\mu}}B(t)$, where $B(t)$ is a standard Brownian motion and the second expression follows again from the invariance principle. Since $\tilde{Y}_i^r(t)$ is nonnegative, non-decreasing, and continuous, we have $\tilde{U}^r(t)$ is continuous, non-decreasing, and $\tilde{U}^r(t) \ge 0$.

Now we quote a technical lemma:

Lemma 10. *Let ϵ^r be a real-valued sequence such that $\epsilon^r \to 0$ as $r \to \infty$, and $\zeta \equiv \frac{1}{\bar{c}} \max_i (\pi_i - \bar{\pi})$. Then, $\int_0^t \mathbb{1}\{|\tilde{W}^r(s) - \zeta| > \epsilon^r\} d\tilde{U}^r(s) \to 0$ in probability, u.o.c.*

Motivated by the result in Lemma 10, we will rewrite $\tilde{U}^r(t)$ in two parts as follows:

$$\tilde{U}^r(t) = \underbrace{\int_0^t \mathbb{1}\{|\tilde{W}^r(s) - \zeta| > \epsilon^r\} d\tilde{U}^r(s)}_{\tilde{U}^r_\epsilon(t)} + \underbrace{\int_0^t \mathbb{1}\{|\tilde{W}^r(s) - \zeta| \leq \epsilon^r\} d\tilde{U}^r(s)}_{\tilde{U}^r_\zeta(t)}.$$

Note that $\tilde{U}^r_\zeta(\cdot), \tilde{U}^r_\epsilon(\cdot)$ are nonnegative, continuous, and nondecreasing. Moreover, $\tilde{U}^r_\epsilon(t) \to 0$, as $r \to \infty$ in probability, u.o.c., and $\sup_{t \leq T} |\tilde{U}^r_\zeta(t) - \tilde{U}^r(t)| \to 0$ in probability. Define now the process $V(t) = \frac{c}{\hat{\mu}} \sqrt{\sigma^2 + \hat{\mu}^2} B(t)$, and consider the auxiliary process $\tilde{R}(t)$ defined as follows: $\tilde{R}(t) = \tilde{R}(0) - \gamma c \int_0^t \tilde{R}(s) ds + V(t) + U(t)$, where $U(t)$ is continuous and nondecreasing, $U(0) = 0$, and $U(t)$ increases only when $\tilde{W}(t)$ hits ζ, i.e., only when $\tilde{R}(t) = \bar{\pi} + \bar{c}\zeta = \max_{i \in \mathcal{N}} \pi_i \equiv \hat{\pi}$. Note also that $\hat{\pi}$ can be regarded as the lower reflecting barrier of the limiting process of $\tilde{R}(t)$. We later on show that this coincides with the limit of $\tilde{U}^r(t)$. By construction, $\tilde{R}(t)$ has the behavior of the hypothesized limit for $\tilde{Z}^r(t)$ specified in Theorem 1.

Let $\tilde{X}(t) = \tilde{R}(t) - \hat{\pi}$. Then,

$$\tilde{X}(t) = \tilde{X}(0) - \gamma c \int_0^t [\tilde{X}(s) + \hat{\pi}] ds + V(t) + U(t) \equiv \varphi(\hat{X}(t)), \tag{37}$$

where $\varphi(\cdot)$ is the reflection operator (Mandelbaum and Pats [1995]) and $\hat{X}(t)$ is defined by $\hat{X}(t) = \tilde{X}(0) - \gamma c \int_0^t [\tilde{X}(s) + \hat{\pi}] ds + V(t)$.

The remaining goal is to show that $\tilde{Z}^r(t)$ converges to $\tilde{R}(t)$. Recall that $\tilde{Z}^r(t) = \tilde{Z}^r(0) - \gamma c \int_0^t \tilde{Z}^r(s) ds + \tilde{V}^r(t) + \tilde{U}^r_\zeta(t) + \tilde{U}^r_\epsilon(t)$, $\tilde{Z}^r(t) = \tilde{Z}^r(0) - \gamma c \int_0^t \tilde{Z}^r(s) ds + \tilde{V}^r(t) + \tilde{U}^r_\zeta(t) + \tilde{U}^r_\epsilon(t)$,and define $\tilde{H}^r(t) = \tilde{Z}^r(t) - \hat{\pi}$ and its primitive process $\hat{H}^r(t)$ as follows: $\hat{H}^r(t) = \hat{H}^r(0) - \gamma c \int_0^t [\tilde{H}^r(s) + \hat{\pi}] ds + \tilde{V}^r(t)$, and $\tilde{H}^r(t) = \hat{H}^r(t) + \tilde{U}^r(t)$.

The key remaining element of the proof is to show that $\tilde{H}^r(s) \Rightarrow \tilde{X}(t)$, which will imply the weak convergence of $\tilde{Z}^r(t)$ to $\tilde{R}(t)$. From Lemma 10 and Proposition 2, for any sequence $\epsilon^r > 0$, s.t. $\epsilon^r \to 0$ as $r \to \infty$, there exists a sequence $\delta^r > 0$ where $\delta^r \downarrow 0$ and $\bar{r}$ large enough such that for all $r > \bar{r}$, we have $P(\inf_{s \leq t} \tilde{W}^r(s) \geq \zeta - \epsilon^r) = 1 - \delta^r$. Let $\Omega(\epsilon^r)$ be the set of sample paths for which $\inf_{s \leq t} \tilde{W}^r(s) \geq \zeta - \epsilon^r$. Note that $P(\Omega(\epsilon^r)) = 1 - \delta^r \to 1$ as $r \to \infty$.

In the sequel we will use Lemma 7 of Ata and Kumar [2005]. We first observe that $\forall \omega \in \Omega(\epsilon^r)$, $\tilde{W}^r(t) \geq \zeta - \epsilon^r \Leftrightarrow \tilde{Z}^r(t) \geq \hat{\pi} - \bar{c}\epsilon^r \Leftrightarrow \tilde{H}^r(t) \geq -\bar{c}\epsilon^r$, and, similarly, $\tilde{W}^r(s) > \zeta + \epsilon^r$ implies $\tilde{H}^r(s) > \bar{c}\epsilon^r$. If we define $H_1^r(t) = \tilde{H}^r(t) + \bar{c}\epsilon^r$, $H_2^r(t) = \hat{H}^r(t) + \bar{c}\epsilon^r$, and focus on the event $\omega \in \Omega(\epsilon^r)$, then $H_1^r, H_2^r, \tilde{U}^r$ satisfy the following conditions:

$$
\begin{aligned}
&H_1^r(t) = H_2^r(t) + \tilde{U}^r(t),\\
&H_1^r(t) \geq 0,\\
&\tilde{U}^r(\cdot) \text{ nondecreasing; } \tilde{U}^r(0) = 0,\\
&\int_0^t \mathbb{1}\{H_1^r(s) > 2\bar{c}\epsilon^r\} d\tilde{U}^r(s) = 0,
\end{aligned}
$$

where the last equation follows from a simple transformation $\tilde{H}^r(s) > \bar{c}\epsilon^r \Leftrightarrow H_1^r(s) > 2\bar{c}\epsilon^r$. Since $\tilde{H}^r, \hat{H}^r, \tilde{U}^r$ all are right continuous and have left limits in $[0,T]$, $\forall T > 0$, these processes $H_1^r, H_2^r, \tilde{U}^r$ fit the conditions of Lemma 7 in Ata and Kumar [2005]. Thus, if we define $\psi(X(t)) = \sup\{X(s)^- : 0 \leq s \leq t\}$ where $X^- = \max(0, -X)$, and $\varphi(X) \equiv X - \psi(X)$ is the regulated process of X, we obtain from Lemma 7 of Ata and Kumar [2005] that $\psi(H_2^r(t)) \leq \tilde{U}^r(t) \leq \psi(H_2^r(t)) + 2\bar{c}\epsilon^r$.

Recall that $\psi(\cdot)$ is nonincreasing and Lipschitz continuous with unity constant, and $H_2^r(t) = \hat{H}^r(t) + \bar{c}\epsilon^r$. It follows that

$$
\psi(\hat{H}^r(t)) - \bar{c}\epsilon^r \leq \tilde{U}^r(t) \leq \psi(\hat{H}^r(t)) + 2\bar{c}\epsilon^r, \tag{38}
$$

which in turn implies $\varphi(\hat{H}^r(t)) - 2\bar{c}\epsilon^r \leq \tilde{H}^r(t) \leq \varphi(\hat{H}^r(t)) + \bar{c}\epsilon^r$.

Step 4:

Subtracting $\hat{H}^r(t)$ by the auxiliary process $\hat{X}(t)$, we obtain

$$
\begin{aligned}
\hat{H}^r(t) - \hat{X}(t) &= -\gamma c \int_0^t [\tilde{H}^r(s) - \tilde{X}(s)] ds + V^r(t) - V(t) + o_p(1),\\
\Rightarrow |\hat{H}^r(t) - \hat{X}(t)| &\leq \gamma c \int_0^t |\tilde{H}^r(s) - \tilde{X}(s)| ds + |V^r(t) - V(t) + o_p(1)|,
\end{aligned}
$$

where the last inequality follows from the triangle inequality. Recall that $\varphi(\cdot)$ is Lipschitz continuous and let K denote the Lipschitz constant. Thus, using (38), this inequality can be rewritten as $|\hat{H}^r(t) - \hat{X}(t)| \leq \gamma c K \int_0^t |\hat{H}^r(s) - \hat{X}(s)| ds + |V^r(t) - V(t) + o_p(1)| + 2\gamma c^2 \epsilon^r t$.

From the strong approximation for $V^r(t)$, we have that for any sequence $v^r > 0, v^r \downarrow 0$, there exists r large enough such that

$$
\sup_{t \leq T} |V^r(t) - V(t)| \leq v^r, \quad \text{w.p. } 1 - \theta^r, \tag{39}
$$

where $\theta^r \downarrow 0$ as $r \to \infty$. In the sequel we concentrate on sample paths in $\Omega(\epsilon^r)$ for which (39) holds, i.e., we consider only $\omega \in \Omega(\epsilon^r, v^r) \equiv \Omega(\epsilon^r) \cap \Omega(v^r)$, where $\Omega(v^r)$ is the set of sample paths for which $\sup_{t \le T} |V^r(t) - V(t)| \le v^r$. The measure of $\Omega(\epsilon^r, v^r)$ is more than $1 - \delta^r - \theta^r$. Then,

$$\begin{aligned} |\hat{H}^r(t) - \hat{X}(t)| &\le \gamma c K \int_0^t |\hat{H}^r(s) - \hat{X}(s)| ds + v^r + 2\gamma c^2 \epsilon^r t \\ &\le (v^r + 2\gamma c^2 \epsilon^r T) e^{\gamma c K T}, \end{aligned} \tag{40}$$

where the second inequality follows from Gronwall's inequality and the fact that $t \le T$.

From the Lipschitz continuity of $\varphi(\cdot)$, (37), and (40), we get that $|\tilde{H}^r(t) - \tilde{X}(t)| \le K(v^r + 2\gamma c^2 \epsilon^r T) e^{\gamma c K T} + 2\bar{c}\epsilon^r$. Let $r \to \infty$, sequences $\epsilon^r, v^r, \delta^r, \theta^r$ all vanish and therefore we conclude that

$$\sup_{t \le T} |\tilde{H}^r(t) - \tilde{X}(t)| \to 0 \quad \text{in probability.} \tag{41}$$

From (41), the fact that $\tilde{Z}^r(t) = \tilde{H}^r(t) + \hat{\pi}$, and the convergence together theorem, it follows that $\tilde{Z}^r(t)$ converges to $\tilde{R}(t)$, which is the desired result identified in Theorem 1. The weak convergence of the queue length processes follows immediately from the convergence of $\tilde{Z}^r(t)$ and the state space collapse result of Proposition 2. □

Proof of Proposition 3

From the *PS* rule, we can define $\Pi_i^{PS}(\pi_i, \pi_{-i}) = \mu_i \pi_i - \frac{\mu_i}{\hat{\mu}} \mathcal{L}(\max_{j \in \mathcal{N}} \pi_j) = \mu_i [\pi_i - \frac{1}{\hat{\mu}} \mathcal{L}(\max_{j \in \mathcal{N}} \pi_j)]$ as the revenue for supplier i when other suppliers select π_{-i}. Note that supplier i always benefits from raising his price infinitesimally if it has not been the highest. Thus, in equilibrium all suppliers must submit the same price. The derivative of Π_i^{PS} with respect to π_i becomes $\frac{\partial \Pi_i^{PS}(\pi_i, \pi_{-i})}{\partial \pi_i} = \mu_i [1 - \frac{1}{\hat{\mu}} \mathcal{L}'(\max_{j \in \mathcal{N}} \pi_j) \mathbb{1}\{\pi_i = \max_{j \in \mathcal{N}} \pi_j\}]$.

Recall from Lemma 3 that $\pi^C = \arg\max_\pi \{\hat{\mu}\pi - \mathcal{L}(\pi)\}$ and therefore $1 - \frac{1}{\hat{\mu}} \mathcal{L}'(\pi^C) = 0$. Imposing symmetry and plugging $\pi_j = \pi^C, \forall j \in \mathcal{N}$, we obtain $\frac{\partial \Pi_i^{PS}(\pi_i, \pi_{-i})}{\partial \pi_i}|_{\pi_j = \pi^C, \forall j \in \mathcal{N}} = \mu_i \left[1 - \frac{1}{\hat{\mu}} \mathcal{L}'(\pi^C)\right] = 0$, where the second equality follows from the definition of π^C. Therefore, every supplier setting price π^C is the symmetric equilibrium under this sharing rule. The uniqueness follows from the strict convexity of $\mathcal{L}(\cdot)$. □

Proof of Proposition 4

Given the transfer prices $\{\eta_{ij} = \frac{\mu_i\mu_j}{\hat{\mu}}\bar{p}, i, j \in \mathcal{N}\}$, supplier i's second-order revenue becomes

$$\begin{aligned}
\tilde{r}_i^{PS}(t) &= \mu_i\pi_i t + \bar{p}\sigma_i B_{s,i}(t) - \mu_i\bar{p}\tilde{Y}_i(t) + \sum_{j\in\mathcal{N},j\neq i}\eta_{ji}\tilde{Y}_i(t) - \sum_{j\in\mathcal{N},j\neq i}\eta_{ij}\tilde{Y}_j(t) \\
&= \mu_i\pi_i t + \bar{p}\sigma_i B_{s,i}(t) - \bar{p}\frac{\mu_i}{\hat{\mu}}\sum_{j\in\mathcal{N}}\mu_j\tilde{Y}_j(t) \\
&= \mu_i\pi_i t + \bar{p}\sigma_i B_{s,i}(t) - \mu_i\bar{p}\frac{1}{c}\tilde{U}(t),
\end{aligned}$$

where we recall that $\tilde{U}(t) = \frac{c}{\hat{\mu}}\sum_{j\in\mathcal{N}}\mu_j\tilde{Y}_j(t)$. Thus, supplier i's long-run average second-order revenue is $\tilde{\Psi}_i(\pi_i,\pi_{-i}) = \mu_i\pi_i - \mu_i\frac{\bar{p}}{c}\mathbb{E}[\tilde{U}(\infty)] = \mu_i\pi_i - \frac{\mu_i}{\hat{\mu}}\mathcal{L}(\max_{j\in\mathcal{N}}\pi_j) = \Psi_i^{PS}(\pi_i,\pi_{-i})$. □

Proof of Lemma 3

Since the second term in (18) is independent of the lower static prices $\pi_i, i \notin J$, at optimality these π_is should all be equal, i.e., $\pi_i = \hat{\pi},\ \forall i \in \mathcal{N}$. Thus, the problem reduces to a single-parameter maximization problem: $\max_{\hat{\pi}}\ \left\{\hat{\mu}\hat{\pi} - \bar{p}\gamma\hat{\mu}\beta\frac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)}\right\}$. Recall that $\frac{\phi(z)}{1-\Phi(z)}$ is the hazard rate of standard normal distribution and hence it is increasing and convex ([Zeltyn and Mandelbaum, 2005, Section 5]).[8] Thus, π^C is well-defined and unique. □

References

G. Allon and A. Federgruen. Service competition with general queueing facilities. Working paper, Kellogg School of Management, Northwestern University. Forthcoming in *Operations Research*, 2006.

G. Allon and A. Federgruen. Competition in service industries. *Operations Research*, 55:37–55, 2007.

M. Armony and M. Haviv. Price and delay competition between two service providers. *European Journal of Operational Research*, 147:32–50, 2003.

B. Ata and S. Kumar. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Annals of Applied Propability*, 1: 331–391, 2005.

[8] We thank Ramandeep Randhawa for bringing this reference to our attention.

M. Baye, D. Kovenock, and C. De Vries. It takes two to tango: equilibria in a model of sales. *Games and Economic Behavior*, 4:493–510, 1992.

O. Besbes. Revenue maximization for a queue that announces real-time delay information. Working paper, Graduate School of Business, Columbia University, 2006.

M. Bramson. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems*, 30:89–148, 1998.

S. Browne and W. Whitt. Piecewise-linear diffusion processes. In *Advances in Queueing*, pages 463–480. CRC Press, 2003.

G. Cachon and P. Harker. Competition and outsourcing with scale economies. *Management Science*, 48:1314–1333, 2002.

P. Glynn. Diffusion approximations. In D. P. Heyman and M. J. Sobel, editors, *Handbooks in OR & MS*, volume 2. North-Holland, 1990.

I. Gurvich and G. Allon. Pricing and dimensioning competing large-scale service providers. Working paper, Columbia Business School, 2008.

R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems.* Kluwer Academic Publishers, Boston, MA, 2002.

R. Johari and G. Weintraub. Competition and contracting in service industries. Working paper, Columbia Business School, 2008.

T. Laseter and S. Bodily. Strategic indicators of B2B e-marketplace financial performance. *Electronic Markets*, 14:322–332(11), 2004.

P. Lederer and L. Li. Pricing, production, scheduling and delivery -time competition. *Operations Research*, 45:407–420, 1997.

D. Levhari and I. Luski. Duopoly prcing and waiting lines. *European Economic Review*, 11:17–35, 1978.

L. Li and Y. Lee. Pricing and delivery-time performance in a competitive environment. *Management Science*, 40:633–646, 1994.

C. Loch. *Pricing in markets sensitive to delay.* Ph.D. dissertation, Stanford University, Stanford, CA, 1991.

I. Luski. On partial equilibrium in a queueing system with two servers. *The Review of Economic Studies*, 43:519–525, 1976.

C. Maglaras and A. Zeevi. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science*, 49:1018–1038, 2003.

A. Mandelbaum and G. Pats. State-dependent queues: approximations and applications. In F. Kelly and R. Williams, editors, *Stochastic Networks*, volume 71 of *Proceedings of the IMA*, pages 239–282. North-Holland, 1995.

A. Mas-Colell, M. Whinston, and J. Green. *Microeconomic Theory.* Oxford University Press, USA, 1995.

H. Mendelson. Pricing computer services: Queueing effects. *Communications of ACM*, 28:312–321, 1985.

H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research*, 38:870–883, 1990.

P. Naor. On the regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

K. So. Price and time competition for service delivery. *Manufactuing and Service Operations Management*, 2:392–409, 2000.

A. L. Stolyar. Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Science*, 19:141–189, 2005.

H. Varian. A model of sales. *American Economic Review*, 70:651–659, 1980.

A. Watts. On the uniqueness of equilibrium in Cournot oligopoly and other games. *Games and Economic Behavior*, 13(2):269–285, 1996.

R. J. Williams. An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Systems*, 30:5–25, 1998.

S. Zeltyn and A. Mandelbaum. Internet supplement to "Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue". *Queueing Systems*, 51:361–402, 2005.

Online Appendix for

"Design of an aggregated marketplace under congestion effects:

Asymptotic analysis and equilibrium characterization"

Proofs of auxiliary results

In this online appendix, we provide proofs for the auxiliary results for the asymptotic analysis (Section 3) and equilibrium characterization (Section 4).

A. Proofs of auxiliary lemmas in Sections 3 and 4

Proof of Lemma 1

Recall that using the strong approximation theorem ([Glynn, 1990, Theorem 7]), the service time process of supplier i can be expressed as follows:

$$S_i^r(t) = \mu_i^r(t - Y_i^r(t)) + \sqrt{r}\sigma_i B_{s,i}(t - Y_i^r(t)) + O(\log rt),$$

where $B_{s,i}$ is the associated Brownian motion with standard deviation σ_i, and Y_i^r is the corresponding idleness process. Recall that $R_i^r(t) = (\bar{p} + \frac{\pi_i}{\sqrt{r}})S_i^r(t)$. Thus, as $r \to \infty$,

$$\frac{R_i^r(t)}{r} = \frac{1}{r}(\bar{p} + \frac{\pi_i}{\sqrt{r}}) \left[r\mu_i(t - Y_i^r(t)) + \sqrt{r}\sigma_i B_{s,i}(t - Y_i^r(t)) + O(\log rt)\right] \to \bar{p}\mu_i t,$$

after removing the lower-order terms. □

Proof of Lemma 2

Recalling the expression $R_i^r(t) = (\bar{p} + \frac{\pi_i}{\sqrt{r}})S_i^r(t)$ from Lemma 1, we have

$$\begin{aligned}
& \frac{1}{\sqrt{r}}\{R_i^r(t) - r\bar{p}\mu_i t\} \\
= \; & \frac{1}{\sqrt{r}}\left\{(\bar{p} + \frac{\pi_i}{\sqrt{r}}) \left[\mu_i^r t - \mu_i^r Y_i^r(t) + \sqrt{r}\sigma_i B_{s,i}(t - Y_i^r(t)) + O(\log rt)\right] - r\bar{p}\mu_i t\right\} \\
= \; & \frac{1}{\sqrt{r}}\{\sqrt{r}[\mu_i\pi_i t + \bar{p}\sigma_i B_{s,i}(t - Y_i^r(t)) - \mu_i\bar{p}\tilde{Y}_i^r(t)] \\
& + \sigma_i\pi_i B_{s,i}(t - Y_i^r(t)) - \mu_i\pi_i\tilde{Y}_i^r(t) + O(\log rt)\}.
\end{aligned}$$

From $Q_i^r(t) = Q_i^r(0) + A_i^r(t) - S_i^r(t - Y_i^r(t))$, the idleness process can be represented as $Y_i^r(t) = t - [S_i^r]^{-1}\left(Q_i^r(0) + A_i^r(t) - Q_i^r(t)\right)$, where $[S_i^r]^{-1}$ is well-defined since $S_i^r(\cdot)$ is monotonically increasing. Since $A_i^r(t)$, $Q_i^r(t)$ both converge according to Theorem 1, the convergence $\tilde{Y}_i^r(t) = \sqrt{r}Y_i^r(t)$ to the limiting process $\tilde{Y}_i(t)$ then follows from Theorem 1 and the convergence together theorem. Moreover, since $Y_i^r(t) \to 0$ as $r \to \infty$ from the proof of Theorem 1, we get that
$r_i^r(t) = \frac{1}{\sqrt{r}}(R_i^r(t) - r\bar{p}\mu_i t) \Rightarrow \mu_i\pi_i + \bar{p}\sigma_i B_{s,i}(t) - \mu_i\bar{p}\tilde{Y}_i(t)$, as $r \to \infty$. □

Proof of Lemma 5

Our first goal is to show that the scaled arrival process $A_i^{r,m}(t)$ increases if and only if $\mathbf{I}_i(Q^{r,m}(t)) = 1$ and $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le v$, where v is the customer's willingness to pay. To this end, we shall consider two events $\{\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le \bar{p}+\frac{\pi_j}{\sqrt{r}}+c\frac{\sqrt{r}Q_j^{r,m}(t)+1}{r\mu_j}\}$ and $\{\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le v\}$. We recall the definition of $Q^{r,m}(t)$ and establish the following equivalence:

$$\begin{aligned} & \bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{\sqrt{r}Q_j^{r,m}(t)+1}{r\mu_j} \qquad (42)\\ \Leftrightarrow\ & \bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_i^r} \le \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_j^r}\\ \Leftrightarrow\ & \bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_i^r} \le \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_j^r}, \end{aligned}$$

where we recall $Q_i^{r,m}(t) = \frac{1}{\sqrt{r}}Q_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)$ and $\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_i^r}$ is simply the expected delay a buyer encounters when joining the queue i at epoch $\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m$. Thus, (42) implies that the total cost submitted by supplier i is less than that by supplier j.

Similarly, if $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le v$, we obtain

$$\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_i^r} = \bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le v.$$

Therefore, if $\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le \bar{p}+\frac{\pi_j}{\sqrt{r}}+c\frac{\sqrt{r}Q_j^{r,m}(t)+1}{r\mu_i}$, $\forall j \ne i$, and $\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le \bar{p}+\frac{\pi_j}{\sqrt{r}}+c\frac{\sqrt{r}Q_j^{r,m}(t)+1}{r\mu_i}$, $\forall j < i$, and $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le v$, then $A_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)$, the arrival process routed to server i at time epoch $\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m$, increases. But this implies that $A_i^{r,m}(t) = \frac{1}{\sqrt{r}}\left[A_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) - A_i^r(\frac{1}{\sqrt{r}}m)\right]$ also increases since $A_i^r(\frac{1}{\sqrt{r}}m)$ is unchanged. Therefore, the scaled arrival process $A_i^{r,m}(t)$ increases if and only if $\mathbf{I}_i(Q^{r,m}(t)) = 1$ and $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \le v$. This

implies that

$$r \max_{i\in\mathcal{N}} ||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L \leq ||N(t)-t||_{Lr},$$

where $N(t)$ represents the counting process of the arrivals and we have applied

$$0 \leq r\int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du \leq Lr, \forall t \in [0,L].$$

We can now establish the probability bound based on the above inequality:

$$\begin{aligned}
& \mathbb{P}\left(\max_{m<\sqrt{r}\tau}\max_{i\in\mathcal{N}} ||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L > \epsilon\right) \\
\leq & \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} \mathbb{P}\left\{||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L > \epsilon\right\} \\
\leq & \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} \mathbb{P}\left\{\frac{1}{r}||N(t)-t||_{Lr} > \epsilon\right\} \\
\leq & \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} \frac{\epsilon}{L^2 r} \\
\leq & \lfloor\sqrt{r}\tau\rfloor \frac{1}{L^2 r}\epsilon,
\end{aligned}$$

where in the third second inequality we have applied [Bramson, 1998, Proposition 4.3].

The proof for the departure process is similar to the above argument, and therefore we only present the major steps here. Consider the term $D_i^{r,m}(t) - \mu_i T_i^{r,m}(t)$. From the definition of hydrodynamic scaling, we have

$$\begin{aligned}
& D_i^{r,m}(t) - \mu_i T_i^{r,m}(t) \\
= & \frac{1}{\sqrt{r}}[D_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) - D_i^r(\frac{1}{\sqrt{r}}m)] - \mu_i^r\left[T_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) - T_i^r(\frac{1}{\sqrt{r}}m)\right] \\
= & \frac{1}{\sqrt{r}}\left\{S_i^r(T_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)) - S_i^r(T_i^r(\frac{1}{\sqrt{r}}m)) - \mu_i^r T_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) + \mu_i^r T_i^r(\frac{1}{\sqrt{r}}m)\right\} \\
= & \frac{1}{\sqrt{r}}\left\{[S_i^r(T_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)) - \mu_i^r T_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)] - [S_i^r(T_i^r(\frac{1}{\sqrt{r}}m)) - \mu_i^r T_i^r(\frac{1}{\sqrt{r}}m)]\right\},
\end{aligned}$$

where $S_i^r(\cdot)$ is the service completion process. Therefore, The probability bound

$$\begin{aligned}
& \mathbb{P}\left\{\max_{m<\sqrt{r}\tau}\max_{i\in\mathcal{N}}||D_i^{r,m}(t)-\mu_i T_i^{r,m}(t)||_L>\epsilon\right\} \\
\leq\ & \mathbb{P}\left(\max_{m<\sqrt{r}\tau}\max_{i\in\mathcal{N}}\left\{\begin{array}{c}||\frac{1}{\sqrt{r}}\left[S_i^r(T_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m))-\mu_i^r T_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)\right]||_L \\ +||\frac{1}{\sqrt{r}}[S_i^r(T_i^r(\frac{1}{\sqrt{r}}m))-\mu_i^r T_i^r(\frac{1}{\sqrt{r}}m)]||_L\end{array}\right\}>\epsilon\right) \\
\leq\ & \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor}\sum_{i\in\mathcal{N}}\mathbb{P}\left\{\begin{array}{cc}\frac{1}{\sqrt{r}}||\left[S_i^r(T_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m))-\mu_i^r T_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)\right]||_L & \\ +\frac{1}{\sqrt{r}}||[S_i^r(T_i^r(\frac{1}{\sqrt{r}}m))-\mu_i^r T_i^r(\frac{1}{\sqrt{r}}m)]||_L & >\epsilon\end{array}\right\} \\
\leq\ & n\lfloor\sqrt{r}\tau\rfloor C_1\frac{\epsilon}{r},
\end{aligned}$$

where C_1 is an appropriate constant independent of r.

Finally, we consider the imbalance of $\pi_i+c\frac{Q_i^{r,m}(t)}{\mu_i}$ and $\pi_j+c\frac{Q_j^{r,m}(t)}{\mu_j}$. The imbalance can be expressed as

$$\begin{aligned}
& \pi_i+c\frac{Q_i^{r,m}(t)}{\mu_i}-\left(\pi_j+c\frac{Q_j^{r,m}(t)}{\mu_j}\right) \\
=\ & \pi_i+c\frac{\frac{1}{\sqrt{r}}Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)}{\mu_i}-\left(\pi_j+c\frac{\frac{1}{\sqrt{r}}Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)}{\mu_j}\right) \\
=\ & \pi_i+\sqrt{r}c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_i}-\left(\pi_j+\sqrt{r}c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_j}\right)-\frac{c}{\sqrt{r}}(\frac{1}{\mu_i}-\frac{1}{\mu_j}) \\
=\ & \sqrt{r}\left(\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_i}\right)-\sqrt{r}\left(\bar{p}+\frac{\pi_j}{\sqrt{r}}+c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_j}\right) \\
& +\frac{c}{\sqrt{r}}(\frac{1}{\mu_i}-\frac{1}{\mu_j}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
& ||\pi_i+c\frac{Q_i^{r,m}(t)}{\mu_i}-\left(\pi_j+c\frac{Q_j^{r,m}(t)}{\mu_j}\right)||_L \\
\leq\ & \sqrt{r}||\left(\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_i}\right)-\left(\bar{p}+\frac{\pi_j}{\sqrt{r}}+c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_j}\right)||_L \\
& +\frac{c}{\sqrt{r}}|\frac{1}{\mu_i}-\frac{1}{\mu_j}|.
\end{aligned}$$

Note that $\left(\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_i}\right)-\left(\bar{p}+\frac{\pi_j}{\sqrt{r}}+c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_j}\right)$ is simply the imbalance of the total cost submitted by suppliers i and j. The probability bound can then be obtained

as follows:

$$
\begin{aligned}
& \mathbb{P}\left\{ \max_{m<\sqrt{r}\tau} \max_{i,j\in\mathcal{N}} || \left(\pi_i + c\frac{Q_i^{r,m}(t)}{\mu_i}\right) - \left(\pi_j + c\frac{Q_j^{r,m}(t)}{\mu_j}\right) ||_L > \epsilon \right\} \\
\leq & \mathbb{P}\left\{ \begin{array}{c} \max_{m<\sqrt{r}\tau} \sqrt{r} \max_{i,j\in\mathcal{N}} || \left(\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_i}\right) - \left(\bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_j}\right) ||_L \\ > \epsilon - \frac{c}{\sqrt{r}}|\frac{1}{\mu_i} - \frac{1}{\mu_j}| \end{array} \right\} \\
\leq & \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} \mathbb{P}\left\{ \begin{array}{c} \max_{i,j\in\mathcal{N}} || \left(\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_i}\right) - \left(\bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{r\mu_j}\right) ||_L \\ > \frac{1}{\sqrt{r}}\left(\epsilon - \frac{c}{\sqrt{r}}|\frac{1}{\mu_i} - \frac{1}{\mu_j}|\right) \end{array} \right\} \\
\leq & \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} n\mathbb{P}\left\{ \frac{1}{r}||N(t)-t||_{Lr} > \frac{1}{\sqrt{r}}\left(\epsilon - \frac{c}{\sqrt{r}}|\frac{1}{\mu_i} - \frac{1}{\mu_j}|\right)\right\} \\
\leq & C_2\lfloor\sqrt{r}\tau\rfloor\frac{\sqrt{r}}{r}\left(\epsilon - \frac{c}{\sqrt{r}}|\frac{1}{\mu_i} - \frac{1}{\mu_j}|\right),
\end{aligned}
$$

where the third inequality follows from a similar argument to bound the routing process, and C_2 is an appropriate constant that is independent of r. This bound is arbitrarily small when r is sufficiently large and ϵ is small. This completes the proof. □

Proof of Lemma 6

The proof is similar to that of [Bramson, 1998, Proposition 5.2], and therefore we only discuss the main steps. Consider first $A^{r,m}(t)$. Without loss of generality, let us assume $t_2 > t_1$. We have that

$$
\begin{aligned}
A_i^{r,m}(t_2) - A_i^{r,m}(t_1) &= \frac{1}{\sqrt{r}}[A_i^r(\frac{1}{\sqrt{r}}t_2 + \frac{1}{\sqrt{r}}m) - A_i^r(\frac{1}{\sqrt{r}}m)] - \frac{1}{\sqrt{r}}[A_i^r(\frac{1}{\sqrt{r}}t_1 + \frac{1}{\sqrt{r}}m) - A_i^r(\frac{1}{\sqrt{r}}m)] \\
&= \frac{1}{\sqrt{r}}[A_i^r(\frac{1}{\sqrt{r}}t_2 + \frac{1}{\sqrt{r}}m) - A_i^r(\frac{1}{\sqrt{r}}t_1 + \frac{1}{\sqrt{r}}m)],
\end{aligned}
$$

and therefore

$$
A_i^{r,m}(t_2) - A_i^{r,m}(t_1) = \frac{1}{\sqrt{r}} N\left(\int_{\frac{1}{\sqrt{r}}t_1+\frac{1}{\sqrt{r}}m}^{\frac{1}{\sqrt{r}}t_2+\frac{1}{\sqrt{r}}m} \Lambda^r \mathbf{I}_i(Q^r(t))\mathbb{P}(v \geq \min_{j\in\mathcal{N}}\left\{\bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(t+m)+1}{\mu_j^r}\right\})dt\right),
$$

where $N(\cdot)$ is the counting process of a unit-rate Poisson. Note that the arrival rate is bounded $\Lambda^r \mathbf{I}_i(Q^r(t))\mathbb{P}(v \geq \min_{j\in\mathcal{N}}\left\{\bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(t+m)+1}{\mu_j^r}\right\}) \leq r$. Therefore, the mean of $A_i^{r,m}(t_2) - A_i^{r,m}(t_1)$ is $\frac{1}{\sqrt{r}} \times r \times (\frac{1}{\sqrt{r}}t_2 - \frac{1}{\sqrt{r}}t_1) = t_2 - t_1$. This implies that $A_i^{r,m}(t)$ is nearly Lipschitz continuous with unit constant. Similarly, we can also show that the scaled processes $\{D_i^{r,m}(t), i \in \mathcal{N}\}$ are also nearly Lipschitz continuous with Lipschitz constants $\{\mu_i\}$'s. This parallels [Bramson, 1998, Proposition 5.2] and hence the details are omitted.

We now consider $\{Q_i^{r,m}(t), i \in \mathcal{N}\}$. Recall that the queueing dynamics $Q_i^r(t) = Q_i^r(0) + A_i^r(t) - D_i^r(t)$,$\forall i \in \mathcal{N}$, we obtain that

$$\begin{aligned} |Q_i^{r,m}(t_2) - Q_i^{r,m}(t_1)| &= |\frac{1}{\sqrt{r}}[Q_i^r(\frac{1}{\sqrt{r}}t_2 + \frac{1}{\sqrt{r}}m) - Q_i^r(\frac{1}{\sqrt{r}}t_1 + \frac{1}{\sqrt{r}}m)]| \\ &\leq \frac{1}{\sqrt{r}}|A_i^r(\frac{1}{\sqrt{r}}t_2) - A_i^r(\frac{1}{\sqrt{r}}t_1)| + \frac{1}{\sqrt{r}}|D_i^r(\frac{1}{\sqrt{r}}t_2) - D_i^r(\frac{1}{\sqrt{r}}t_1)|. \end{aligned}$$

From the above discussions on A_i^r and D_i^r, $\{Q_i^{r,m}(t), i \in \mathcal{N}\}$ are also nearly Lipschitz continuous with constant $\max_{i\in\mathcal{N}}\{\mu_i\} + 1$. The Lipschitz continuity of $T_i^{r,m}(t)$ and $W_i^{r,m}(t)$ can be established similarly. □

Proof of Lemma 8

Although the routing in our model depends on the queue length processes $\{Q_i^{r,m}(t), i \in \mathcal{N}\}$, the Lipschitz continuity of $X_i^{r,m}$ from Lemma 7 and the definition of K^r lead to this lemma immediately according to [Bramson, 1998, Proposition 4.1]. □

Proof of Lemma 9

This follows from [Bramson, 1998, Proposition 6.2]. The idea is to approximate the cluster point $\widehat{X}(\cdot)$ by a sequence of $\{X^{r,m}(\cdot)\}$, and then take r to the infinity. Specifically, let us consider, for example, the queue length dynamics:

$$\widehat{Q}_i(t) = \widehat{Q}_i(0) + \int_0^t \widehat{\Lambda}\mathbf{I}_i(\widehat{Q}(u))du - \widehat{D}_i(t), \forall i \in \mathcal{N}.$$

We now show that this equation is indeed satisfied by the cluster point $\widehat{X}(\cdot)$. We first find a sequence of $\{X^{r,m}(\cdot)\}$ that converges to $\widehat{X}(\cdot)$. We then obtain

$$\begin{aligned} & |\widehat{Q}_i(t) - \widehat{Q}_i(0) + \int_0^t \widehat{\Lambda}\mathbf{I}_i(\widehat{Q}(u))du - \widehat{D}_i(t)| \\ \leq\ & |\widehat{Q}_i(t) - Q_i^{r,m}(t)| + |\widehat{Q}_i(0) - Q_i^{r,m}(0)| \\ & + |\widehat{D}_i(t) - D_i^{r,m}(t)| + |\int_0^t \mathbf{I}_i(\widehat{Q}(u))du - \int_0^t \mathbf{I}_i(Q^{r,m}(u))du| \\ & + |Q_i^{r,m}(t) - Q_i^{r,m}(0) + \int_0^t \Lambda\mathbf{I}_i(Q^{r,m}(u))du - D_i^{r,m}(t)| \\ =\ & |\widehat{Q}_i(t) - Q_i^{r,m}(t)| + |\widehat{Q}_i(0) - Q_i^{r,m}(0)| + |\widehat{D}_i(t) - D_i^{r,m}(t)| + |\int_0^t \mathbf{I}_i(\widehat{Q}(u))du - \int_0^t \mathbf{I}_i(Q^{r,m}(u))du|. \end{aligned}$$

Since $\mathbf{I}_i(\cdot)$ is a continuous function, $|\int_0^t \mathbf{I}_i(\widehat{Q}(u))du - \int_0^t \mathbf{I}_i(Q^{r,m}(u))du|$ is bounded. Moreover, all other terms are bounded by construction, and therefore $|\widehat{Q}_i(t) - \widehat{Q}_i(0) + \int_0^t \widehat{\Lambda}\mathbf{I}_i(\widehat{Q}(u))du - \widehat{D}_i(t)| \to 0$ as $r \to \infty$. Likewise, other fluid equations can be verified as well. □

Proof of Lemma 10

First we will show that there exists a constant r_0 such that $\tilde{W}^r(s) \geq \zeta - \epsilon^r$ in probability, $\forall r > r_0$. The proof is by contradiction.

We let $m = \arg\max_{j\in\mathcal{N}} \pi_j$ to denote the supplier with the highest static price and $i \notin \arg\max_{j\in\mathcal{N}} \pi_j$ in the sequel. In the following all the inequalities refer to the inequalities "in probability," and we omit this indication for convenience. Suppose $\tilde{W}^r(s) < \zeta - \epsilon^r$, then

$$\sum_{j\neq m} \frac{\tilde{Q}_j^r(s)}{\mu_j} < \zeta - \epsilon^r - \frac{\tilde{Q}_m^r(s)}{\mu_m} \leq \zeta - \epsilon^r. \tag{43}$$

Now since ϵ^r is given, we can define $\eta^r = \frac{c}{2(n-2)}\epsilon^r > 0$ and $\eta^r \downarrow 0$ because $\epsilon^r \downarrow 0$. If for all r we can find a pair $i, j \in \mathcal{N}$ such that $\pi_j + \frac{\tilde{Q}_j^r(s)}{\mu_j} \leq \pi_i + \frac{\tilde{Q}_i^r(s)}{\mu_i} - \eta^r$, then letting $r \to \infty$ we have found a sequence that contradicts Proposition 2. Thus from now on we assume that

$$\pi_j + \frac{\tilde{Q}_j^r(s)}{\mu_j} > \pi_i + \frac{\tilde{Q}_i^r(s)}{\mu_i} - \eta^r, \ \forall j \neq i, m; i, j \in \mathcal{N}. \tag{44}$$

Combining Equations (43) and (44), we obtain

$$\begin{aligned} \zeta - \epsilon^r \quad &> \quad \sum_{j\neq m} \frac{\tilde{Q}_j^r(s)}{\mu_j} \geq (n-1)\frac{\tilde{Q}_i^r(s)}{\mu_i} + \frac{1}{c}\sum_{j\neq i,m}(\pi_i - \pi_j) - \frac{n-2}{c}\eta^r \\ &\Rightarrow \quad \frac{\tilde{Q}_i^r(s)}{\mu_i} \leq \frac{1}{n-1}[\zeta - \epsilon^r - \frac{1}{c}\sum_{j\neq i,m}(\pi_i - \pi_j) + \frac{n-2}{c}\eta^r]. \end{aligned}$$

The proposed cost by supplier i is upper bounded by:

$$\begin{aligned} \pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i} &\leq \pi_i + \frac{c}{n-1}\left\{\frac{1}{c/n}\left[\pi_m - \frac{1}{n}\sum_{k\in\mathcal{N}}\pi_k\right] - \epsilon^r - \frac{1}{c}\sum_{j\neq i,m}(\pi_i - \pi_j) + \frac{n-2}{c}\eta^r\right\} \\ &\leq \pi_i + \frac{c}{n-1}\left[\frac{n}{c}\pi_m - \frac{1}{c}\sum_{k\in\mathcal{N}}\pi_k - \frac{1}{c}\sum_{j\neq i,m}(\pi_i - \pi_j)\right] - \frac{c}{n-1}\left(\epsilon^r - \frac{n-2}{c}\eta^r\right) \end{aligned}$$

By our choice of η^r, the last term is strictly negative. The other terms can be combined such that

$$\begin{aligned}\pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i} &\leq \frac{1}{n-1}[(n-1)\pi_i + n\pi_m - \sum_{j\in\mathcal{N}}\pi_j - (n-2)\pi_i + \sum_{j\neq i,m}\pi_j] - \frac{c}{2(n-1)}\epsilon^r \\ &= \pi_m - \frac{c}{2(n-1)}\epsilon^r.\end{aligned}$$

Since the queue length can never be negative, we conclude that

$$\pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i} \leq \pi_m + c\frac{\tilde{Q}_m^r(s)}{\mu_m} - \frac{c}{2(n-1)}\epsilon^r,$$

which contradicts Proposition 2 if we let $r \to \infty$. since costs proposed by supplier i and m are different. Thus, $\tilde{W}^r(s) \geq \zeta - \epsilon^r,\ \forall r \geq r_0$.

The next thing is to show that $\tilde{Q}^r(s) > 0$ in probability if $\tilde{W}^r(s) \geq \zeta + \epsilon^r$. We again prove this by contradiction. Suppose that there exists i such that $\tilde{Q}_i^r(s) = 0$. If $i \notin \arg\max_{j\in\mathcal{N}}\pi_j$, then

$$\pi_m + c\frac{\tilde{Q}_m^r(s)}{\mu_m} \geq \pi_m > \pi_i = \pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i},$$

which contradicts the state space collapse while $r \to \infty$. So it suffices to consider the case $\tilde{Q}_m^r(s) = 0$.

Recall the definition of η^r and apply state space collapse, we have

$$\begin{aligned}\pi_j \quad +c\frac{\tilde{Q}_j^r(s)}{\mu_j} \leq &\quad \pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i} + \frac{\eta^r}{c}, \forall j \neq m, \\ \Rightarrow \qquad &\frac{\tilde{Q}_j^r(s)}{\mu_j} \leq \frac{1}{c}(\pi_i - \pi_j) + \frac{\tilde{Q}_i^r(s)}{\mu_i} + \frac{\eta^r}{c}, \forall j \neq m.\end{aligned}$$

Note that $\tilde{Q}_m^r(s) = 0$ implies $\zeta + \epsilon^r \leq \tilde{W}^r(s) = \sum_{j\neq m}\frac{\tilde{Q}_j^r(s)}{\mu_j}$, and therefore

$$\begin{aligned}\zeta + \epsilon^r &\leq \frac{1}{c}\sum_{j\neq i,m}(\pi_i - \pi_j) + (n-1)\frac{\tilde{Q}_i^r(s)}{\mu_i} + \frac{(n-2)}{c}\eta^r, \\ \Rightarrow \frac{\tilde{Q}_i^r(s)}{\mu_i} &\geq \frac{1}{n-1}[\zeta - \frac{1}{c}\sum_{j\neq i,m}(\pi_i - \pi_j)] + \epsilon^r - \frac{(n-2)}{c}\eta^r.\end{aligned}$$

The cost proposed by supplier i is lower bounded by

$$\pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i} \geq \pi_i + \frac{c}{n-1}[\frac{1}{c/n}\pi_m - \frac{1}{c/n}\frac{1}{n}\sum_{k\in\mathcal{N}}\pi_k - \frac{n-2}{c}\pi_i + \frac{1}{c}\sum_{j\neq i,m}\pi_j] + c(\epsilon^r - \frac{(n-2)}{c}\eta^r).$$

After some manipulation, the above inequality can be rewritten as

$$\pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i} \geq \pi_m + \frac{c}{2}\epsilon^r = \pi_m + c\frac{\tilde{Q}_m^r(s)}{\mu_m} + \frac{c}{2}\epsilon^r,$$

where the last equality follows from $\tilde{Q}_m^r(s) = 0$. This, however, contradicts Proposition 2, and hence $\tilde{Q}^r(s) > 0$ if $\tilde{W}^r(s) \geq \zeta + \epsilon^r$ in probability. □

B. Competitive equilibrium of the suppliers' pricing game

Recall that in a decentralized system, each supplier is maximizing his own payoff in a non-cooperative way. We first prove Lemma 4, which states that no pure-strategy equilibrium exists.

Proof of Lemma 4

Suppose that there exists a pure-strategy Nash equilibrium. If in equilibrium not all π_i's are the same, then there exists a supplier j with $\pi_j < \hat{\pi}$. She will be better off by increasing his price π_j a little bit since his revenue $\mu_j\pi_j$ is strictly increasing in π_j. Thus, the only possibility is that prices π_i's are all equal. But then, again, price undercutting by any supplier is a profitable deviation. □

In the following we characterize the equilibrium behavior. We will separate our discussion into two cases, depending on whether suppliers are endowed with homogeneous or heterogeneous service rates.

Homogeneous service rate case

We first consider the case where the service rates are the same across suppliers, i.e., $\mu_i = \mu_j \equiv \mu,\ \forall i, j \in \mathcal{N}$, and focus on symmetric equilibria. Define $\pi^* = \arg\max_\pi[\mu\pi - \mathcal{L}(\pi)]$ and $\Psi^* \equiv \mu\pi^* - \mathcal{L}(\pi^*)$. Note that we are charging all the idling penalty to a single supplier. In this way, a price π^* guarantees a lower bound for the payoff $\Psi_i(\pi_i, \pi_{-i})$. Hence, Ψ^* is the payoff that a supplier can guarantee herself to obtain, i.e., his *minmax* level. We further let $\underline{\pi} := \Psi^*/\mu = \pi^* - \mathcal{L}(\pi^*)/\mu < \pi^*$ and observe that choosing price $\pi < \underline{\pi}$ is a dominated strategy. Thus, $\underline{\pi}$ can be regarded as a lower bound of suppliers' rational pricing strategies.

Let $G(\pi)$ denote the mixing cumulative probability distribution of a supplier's pricing strategy π. The next proposition characterizes the structure of these mixing probabilities.

Proposition 5. *With homogeneous rates, there exists a unique symmetric equilibrium in which all suppliers randomize continuously over* $[\underline{\pi}, \pi^*]$*, and every supplier gets* Ψ^**. The randomizing distribution is* $G(\pi) = [\frac{\mu(\pi-\underline{\pi})}{\mathcal{L}(\pi)}]^{1/(n-1)}, \forall \pi \in [\underline{\pi}, \pi^*]$.

Proof. We will follow Baye et al. [1992] to prove this result. Let $\bar{s}$ and $\underline{s}$ denote respectively the essential upper and lower bounds of $G(\cdot)$'s support. That is, $G(\pi) = 0, \forall \pi < \underline{s}, 0 \leq G(\pi) < 1,\ \forall \pi \in (\underline{s}, \bar{s})$, and $G(\pi) = 1, \forall \pi \geq \bar{s}$. If the support is not bounded above, we can simply set $\bar{s} = \infty$. We further let $\alpha(\pi)$ denote the mass probability a supplier puts on price π.

First we observe that $\underline{s} \geq \underline{\pi},\ \forall i \in \mathcal{N}$. To see this, if a supplier chooses π^*, his expected payoff is at least $\mu\pi^* - \mathcal{L}(\pi^*) = \Psi^* = \mu\underline{\pi}$. Thus, choosing any price $\pi < \underline{\pi}$ yields an expected payoff no more than $\mu\pi < \mu\underline{\pi} = \Psi^*$, and therefore is strictly dominated. Moreover, $\alpha(\underline{s}) = 0$, i.e., no supplier is expecting to be the most expensive while placing $\underline{s}$. We verify the latter by contradiction: Suppose $\alpha(\underline{s}) > 0$, and that a supplier selects $\underline{s}$. He will become the most expensive supplier (and therefore shares the penalty) when all other suppliers also select $\underline{s}$, which occurs with probability $[G(\underline{s})]^{n-1}$. When this occurs, all n suppliers share the penalty equally, and therefore the supplier's expected payoff is

$$\mu\underline{s} - \frac{1}{n}\mathcal{L}(\underline{s})[G(\underline{s})]^{n-1} < \mu\underline{s}. \tag{45}$$

Since $G(\underline{s})$ is strictly positive, we are subtracting a positive amount in the LHS. Now suppose that a supplier deviates and chooses a price $\underline{s} - \epsilon$ while other suppliers set prices at $\underline{s}$. Thus, his expected payoff is $\mu(\underline{s} - \epsilon)$ (without any penalty incurred). When ϵ is small enough, $\mu(\underline{s} - \epsilon)$ is strictly higher than the LHS of (45). This implies that the supplier will always intend to undercut the price by a sufficiently small ϵ.

We now define $\Psi_i(\pi)$ as the expected payoff of supplier i if he places price π, and Ψ_i^e his equilibrium payoff. We claim that the suppliers get equal payoffs in equilibrium, i.e., $\Psi_i^e = \Psi_j^e = \mu\underline{s},\ \forall i, j \in \mathcal{N}$. The proof is as follows. Under a mixed-strategy equilibrium, a supplier should get the same payoff from all strategies on which he places positive probabilities. Hence $\Psi_i^e = \Psi_i(\underline{s}) = \mu\underline{s},\ \forall i \in \mathcal{N}$.

Next, we characterize the support. Suppose that $\alpha(\bar{s}) > 0$. Then transferring the weight $\alpha(\bar{s})$ to a price just below it will be a profitable deviation. Therefore $\alpha(\bar{s}) = 0$. Given that, when a supplier places price $\bar{s}$, with probability one he will be the sole most expensive supplier and hence

carries the entire market idleness. Thus, if $\bar{s} \neq \pi^*$, his payoff would be

$$\Psi_i(\bar{s}) = \mu\bar{s} - \mathcal{L}(\bar{s}) < \mu\pi^* - \mathcal{L}(\pi^*) = \mu\underline{\pi} \leq \mu\underline{s},$$

where the strict inequality follows from that π^* is the unique maximizer of $\mu\pi - \mathcal{L}(\pi)$. This leads to a contradiction, since $\Psi_i^e = \mu\underline{s},\ \forall i \in \mathcal{N}$. Thus, $\bar{s} = \pi^*$. Finally, since

$$\mu\underline{s} = \Psi_i^e = \Psi_i(\pi^*) = \mu\pi^* - \mathcal{L}(\pi^*) = \mu\underline{\pi},$$

we conclude that $\underline{s} = \underline{\pi}$.

Having characterized the support, we now show that the suppliers will randomize continuously over the entire support. We first claim that there is no point mass in $(\underline{\pi}, \pi^*)$. Suppose this is not true. Then there exists a price $\pi \in (\underline{\pi}, \pi^*)$ such that each supplier puts a point mass on π. Following the argument in Lemma 10 of Baye et al. [1992], it pays for a supplier to transfer a ϵ-neighborhood mass above π to a δ-neighborhood below π, and thus this cannot be an equilibrium.

Now we show that G is strictly increasing in the entire support. If the claim is false, there must exist an interval (π_1, π_2) such that $G(\pi_1) = G(\pi_2)$. Since $[G(\pi)]^{n-1}$ (i.e., the probability that a supplier quoting π becomes the most expensive supplier) does not change in (π_1, π_2), by moving a small mass from slightly below π_1 to π_2, a supplier is strictly better off. That is, the supplier charges a higher price π, with a smaller probability of becoming the most expensive. This contradicts the assumption that $G(\pi)$ is an equilibrium strategy, and therefore G must be strictly increasing.

Combining all above, in a symmetric equilibrium suppliers randomize continuously in the support. Finally, we derive the closed-form expression for $G(\pi)$. Since $G(\pi)$ is strictly increasing in the entire support, π is involved in supplier i's equilibrium strategy. Therefore a supplier should get the same payoff for all such π's (otherwise he would not be willing to play the equilibrium strategy). Thus, For all $\pi \in [\underline{\pi}, \pi^*]$,

$$\Psi_i^e = \mu\underline{s} = \Psi_i(\pi) = \mu\pi - \mathcal{L}(\pi)[G(\pi)]^{n-1}, \forall \pi \in [\underline{\pi}, \bar{\pi}^*],$$

where the last equality follows from that with probability $[G(\pi)]^{n-1}$ supplier i will become the most expensive one. Rearranging the above equation, we obtain the expression for $G(\pi)$, which completes the proof. □

Note that the range over which the price is randomized is completely determined by the individual's problem. In all generic cases, no tie of the highest static price may occur. In other

words, the market idleness process is contributed by only one supplier. Moreover, any tie of two prices takes place with probability zero, which is in contrast to the centralized system where prices are always equal. Therefore, our homogeneous service model suggests that *price dispersion can be regarded as a sign of incoordination.* Also note that in equilibrium the expected payoff of a supplier is identical to the case where he carries the entire market idleness, and hence he receives on average the minmax level. Competitive behavior drives away the possibility of extracting extra revenues.

Having characterized the competitive equilibrium, we now turn to the market efficiency issue. Define $\Pi^C \equiv \max_{\hat{\pi}} \{\hat{\mu}\hat{\pi} - \mathcal{L}(\hat{\pi})\}$ as the aggregate (second-order) revenue under the centralized solution and Π^* as the aggregate revenue among suppliers in the unique competitive equilibrium. The next proposition shows that the efficiency can be arbitrarily low when the market size explodes.

Proposition 6. *Suppose that the service rates are homogeneous. For any given aggregate service rate $\hat{\mu}$, for any given constant M, there exists a sufficiently large number N_M such that $|\Pi^C - \Pi^*| > M$, $\forall n > N_M$.*

Proof. Given $\hat{\mu}$ (and other relevant parameters such as c and σ), $\mathcal{L}(\pi)$ is a known function and therefore Π^C is a fixed constant independent of n, the number of suppliers. From Proposition 5, $\Pi^* = n\Psi^*$, where $\Psi^* = \max_\pi[\mu\pi - \mathcal{L}(\pi)] = \max_\pi[\frac{\hat{\mu}}{n}\pi - \mathcal{L}(\pi)]$. The function $\mathcal{L}(\pi)$ is strictly convex and positive, and therefore Ψ^* can be obtained via the first-order condition. When n is sufficiently large, the maximizer $\pi^* = \arg\max_\pi[\frac{\hat{\mu}}{n}\pi - \mathcal{L}(\pi)]$ is arbitrarily small. Thus, for a given number M, we obtain that there exists a number such that $\pi^* < \min\{-\frac{M-\Pi^C}{\hat{\mu}}, \frac{\Pi^C}{\hat{\mu}}\}$ whenever $n > N_M$. When $n > N_M$, the difference between the aggregate revenues under the centralized and decentralized systems is then

$$|\Pi^C - \Pi^*| = \Pi^C - \Pi^* = \Pi^C - n\max_\pi[\frac{\hat{\mu}}{n}\pi - \mathcal{L}(\pi)] > \Pi^C - n\frac{\hat{\mu}}{n}\pi^* = -\hat{\mu}\pi^* + \Pi^C > M.$$

This completes the proof. □

Proposition 6 shows that as the market size grows, the competitive behavior among the suppliers may result in an unbounded efficiency loss. This demonstrates a significant inefficiency due to the market mechanism and it therefore calls for the need of the coordination scheme, as we investigate in Section 4.4. Note also that the first-order aggregate revenues of the centralized solution and the competitive equilibrium coincide; nevertheless, this is by construction of the asymptotic regime specified in Section 3.

Heterogeneous service rate case

Now we consider the scenario where suppliers are endowed with different service rates. We again first define a global maximizers π_1^*, π_2^*,..., π_m^*, if supplier i ($1 \leq i \leq m$) is the one who proposes the highest price solely; i.e., we define $\pi_i^* = \arg\max_{\pi_i}[\mu_i \pi_i - \mathcal{L}(\pi_i)]$ and $\Psi_i^* \equiv \mu_i \pi_i^* - \mathcal{L}(\pi_i^*)$ as the global maximum revenue that supplier i can achieve as $J = \{i\}$. Next we let $\underline{\pi}_i = \Psi_i^*/\mu_i$ and recall that choosing price $\pi < \underline{\pi}_i$ is a dominated strategy for supplier i. The following proposition characterizes the relevant properties of an equilibrium needed for our purpose. $G_i(\cdot)$ denotes the mixing distribution that supplier i adopts in equilibrium.

Proposition 7. *Suppose suppliers are endowed with heterogeneous service rates. Then in a competitive equilibrium,*

- *All $G_i(\cdot)$'s have the same left endpoint (denoted by $\underline{s}$) of their supports. Moreover, $\underline{s} \geq \max_{i \in \mathcal{N}} \underline{\pi}_i$.*
- *Suppliers' expected payoffs are proportional to their service rates $\{\mu_i\}$'s.*
- *If $n = 2$ and $\mu_1 > \mu_2$, then there exists a unique equilibrium in which supplier i's revenue is $\mu_i \underline{\pi}_1, i = 1, 2$. The equilibrium mixing probabilities are respectively*

$$G_2(\pi) = \frac{\mu_1(\pi - \underline{\pi}_1)}{\mathcal{L}(\pi)}, \forall \pi \in [\underline{\pi}_1, \pi_1^*], \quad G_1(\pi) = \frac{\mu_2(\pi - \underline{\pi}_1)}{\mathcal{L}(\pi)}, \forall \pi \in [\underline{\pi}_1, \pi_1^*),$$

and $G_1(\pi_1^*) = 1$. $G_1(\pi) = G_2(\pi) = 0, \forall \pi \leq \underline{\pi}_1$.

Proof. Let $\underline{s}_i$ and $\bar{s}_i$ denote respectively the essential lower and upper bounds of $G_i(\cdot)$'s support, and $\alpha_i(\pi)$ is the probability mass supplier i puts at point π. Let $\Psi_i(\pi, G_{-i})$ be the expected payoff of supplier i when he plays π and other suppliers adopt the mixing probabilities G_{-i}. Let Ψ_i^e be his expected payoff in equilibrium.

Since quoting a price below $\underline{\pi}_i$ is a dominated strategy for supplier i, we have $\underline{s}_i \geq \underline{\pi}_i$. Moreover, the lower bounds of supports for $G_i(\cdot)$'s must be equal. If this is not the case, a supplier with the lowest $\underline{s}_i$ would refuse to put any positive weight on prices between his lower bound and the highest lower bound $\max_{j \in \mathcal{N}} \underline{s}_j$. Combining the above, we know that $\underline{s}_i := \underline{s} \geq \max_{j \in \mathcal{N}} \underline{\pi}_j, \forall i \in \mathcal{N}$.

Now we claim that if a supplier plays $\underline{s}$, with probability 1 he will not be the most expensive supplier. That is, there exist i, j such that $\alpha_i(\underline{s}) = \alpha_j(\underline{s}) = 0$. If this is not true, then transferring some weight to a neighborhood just below $\underline{s}$ will be profitable for some supplier.

Since at least two suppliers put zero mass at $\underline{s}$, and every supplier may place $\underline{s}$ in equilibrium, we have

$$\Psi_i(\underline{s}, G_{-i}) = \mu_i \underline{s} - \mathcal{L}(\underline{s}) \times 0 = \mu_i \underline{s},$$

which results in $\Psi_i^e = \mu_i \underline{s},\ \forall i \in \mathcal{N}$.

We now focus on the two-supplier case. The following lemma provides some relations of π_1^*, π_2^* and $\underline{\pi}_1, \underline{\pi}_2$, which are needed for characterizing the equilibrium mixing probabilities.

Lemma 11. *If $\mu_1 > \mu_2$, then $\pi_1^* > \pi_2^*$ and $\underline{\pi}_1 > \underline{\pi}_2$.*

Proof. Let us define $\Psi^*(\mu) = \max_\pi \{\mu\pi - \mathcal{L}(\pi)\}$. Consider two service rates μ_1 and μ_2, and assume without loss of generality that $\mu_1 > \mu_2$. By definition the maximizer for supplier with μ_1 is π_1^*, i.e., $\Psi^*(\mu_1) = \mu_1\pi_1^* - \mathcal{L}(\pi_1^*)$. From the optimal condition, we have

$$\begin{aligned} \Psi^*(\mu_2) &= \max_\pi \{\mu_2\pi - \mathcal{L}(\pi)\} \geq \mu_2\pi_1^* - \mathcal{L}(\pi_1^*) = \mu_1\pi_1^* - \mathcal{L}(\pi_1^*) + \pi_1^*(\mu_2 - \mu_1) = \Psi^*(\mu_1) + \pi_1^*(\mu_2 - \mu_1) \\ &\Leftrightarrow \quad \Psi^*(\mu_2) \geq \Psi^*(\mu_1) + \pi_1^*(\mu_2 - \mu_1), \forall \mu_1, \mu_2. \end{aligned}$$

Also, from the Envelope Theorem, $\frac{\partial \Psi^*(\mu_1)}{\partial \mu_1} = \pi_1^*$. Note that the above inequality holds for arbitrary pair of μ_1, μ_2, and hence $\Psi^*(\mu)$ is convex in μ.

Since $\Psi^*(\mu)$ is convex, its derivative $\pi_i^*(\mu)$ is increasing in μ, and therefore $\pi_1^* \geq \pi_2^*$. The strict monotonicity follows immediately from the first-order condition of $\max_\pi \{\mu\pi - \mathcal{L}(\pi)\}$. Moreover, $\underline{\pi}(\mu) = \Psi^*(\mu)/\mu$ can be regarded as the average of the derivative over $[0, \mu]$, i.e., $\frac{\Psi^*(\mu)}{\mu} = \frac{1}{\mu}\int_0^\mu \frac{\partial \Psi(v)}{\partial v} dv$, by the strict monotonicity of $\pi^*(\mu)$, it is also monotonic. Therefore, $\underline{\pi}_1 := \underline{\pi}(\mu_1) > \underline{\pi}_2 := \underline{\pi}(\mu_2)$. □

We now return to the proof of Proposition 7. Note that the upper bounds of the supports for both G_1 and G_2 must be the same. We argue by contradiction: If this is not true, then we have either $\bar{s}_1 > \bar{s}_2$, or $\bar{s}_2 > \bar{s}_1$. Consider the first case. Note that when supplier 1 places a price $\pi \in (\bar{s}_2, \bar{s}_1]$, he will carry the entire market idleness. If $\pi_1^* < \bar{s}_1$, then transferring some distribution weight in $(\bar{s}_2, \bar{s}_1]$ downwards to $\max(\pi_1^*, \bar{s}_2)$ is a profitable deviation for supplier 1. If $\pi_1^* > \bar{s}_1$, then he would like to transfer all the weight in $(\bar{s}_2, \bar{s}_1]$ upwards to π_1^*. Thus $\bar{s}_1 > \bar{s}_2$ is impossible. Similarly, one can show that the other case never occurs as well. Therefore $\bar{s}_1 = \bar{s}_2 = \bar{s}$.

Given that there are only two suppliers, both suppliers do not put a point mass on the common upper bound simultaneously, otherwise a mass transfer from $\bar{s}$ to its lower neighborhood will be profitable for either supplier. Thus, at most one supplier puts probability zero on $\bar{s}$. Suppose $\alpha_1(\bar{s}) = 0$. Then

$$\Psi_2(\bar{s}, G_{-2}) = \mu_2\bar{s} - \mathcal{L}(\bar{s}) \leq \Psi_2^* = \mu_2\underline{\pi}_2 < \mu_2\underline{\pi}_1 \leq \mu_2\underline{s},$$

which contradicts the equilibrium condition. Hence, from the strict inequality in between the extremes, $\alpha_1(\bar{s}) > 0$ and consequently $\alpha_2(\bar{s}) = 0$. Moreover,

$$\Psi_1(\bar{s}, G_{-1}) = \mu_1\bar{s} - \mathcal{L}(\bar{s}) \leq \mu_1\pi_1^* - \mathcal{L}(\pi_1^*) = \mu_1\underline{\pi}_1 \leq \mu_i\underline{s} = \Psi_1(\underline{s}, G_{-1}),$$

and the only chance for equalities to hold is that $\bar{s} = \pi_1^*$ and $\underline{s} = \underline{\pi}_1$. This determines the common support.

In the interior, no hole and no point mass should be placed for both suppliers, otherwise one can construct a profitable weight transfer. Thus, both G_1, G_2 increase continuously in $[\underline{\pi}_1, \pi_1^*)$. Finally, since in equilibrium a supplier should obtain the same expected payoff at any point in $[\underline{\pi}_1, \pi_1^*]$, we have

$$\begin{aligned} \Psi_1(\pi, G_{-1}) &= \mu_1\pi - G_2(\pi)\mathcal{L}(\pi) = \mu_1\underline{\pi}_1 = \Psi_1(\underline{\pi}_1, G_{-1}), \forall \pi \in [\underline{\pi}_1, \pi_1^*], \\ &\Rightarrow \quad G_2(\pi) = \frac{\mu_1(\pi - \underline{\pi}_1)}{\mathcal{L}(\pi)}, \forall \pi \in [\underline{\pi}_1, \pi_1^*]. \end{aligned}$$

Similarly,

$$\begin{aligned} \Psi_2(\pi, G_{-2}) &= \mu_2\pi - G_1(\pi)\mathcal{L}(\pi) = \mu_2\underline{\pi}_1 = \Psi_2(\underline{\pi}_1, G_{-2}), \forall \pi \in [\underline{\pi}_1, \pi_1^*), \\ &\Rightarrow \quad G_1(\pi) = \frac{\mu_2(\pi - \underline{\pi}_1)}{\mathcal{L}(\pi)}, \forall \pi \in [\underline{\pi}_1, \pi_1^*), \end{aligned}$$

and $\alpha_1(\pi_1^*) = 1 - \frac{\mu_2(\pi_1^* - \underline{\pi}_1)}{\mathcal{L}(\pi)}$. This completes the proof. □

The first result on left endpoints is not surprising. This comes directly from an analogous argument for Proposition 5. The second result captures the ex ante difference between suppliers' payoff function: higher service rate brings higher equilibrium payoff. When we restrict to the duopoly setting, we know perfectly the range over which suppliers randomize their prices, and we can obtain closed-form expressions of their expected payoffs. They randomize the prices over the same range, and the supplier with a higher service rate tends to set a lower price: his mixing distribution stochastically dominates the other's in the usual, first-order sense. This implies that when a supplier has capacity advantage, he can afford to price lower to capture more customers.

Numerical results

We in the following give some numerical examples. Our goal is to compare the performance between the centralized solution and the competitive equilibrium. We consider a system with n $M/M/1$ servers, delay sensitivity parameter $c = 0.5$, and arrival rate of buyers $\Lambda = 1$. The valuation v of each customer is assumed to follow an exponential distribution with mean $\frac{1}{10}$, and $\bar{p}$ is set such that the effective arrival rate $\mathbb{P}(v \geq \bar{p})$ matches the total service rate $\hat{\mu}$. As an example, if we let $\hat{\mu} = e^{-1.3}$, then $\bar{p}$ can be obtained as follows: $\Lambda e^{-10\bar{p}} = \hat{\mu} \Leftrightarrow \bar{p} \approx 0.13$. The other relevant parameter is $\gamma = f(\bar{p})/(1 - F(\bar{p})) = 10$. Note that as we scale according to $\Lambda = r$, $\hat{\mu} = \hat{\mu}^r$, $\bar{p}$ stays unchanged.

The next two figures compare the centralized solution and the competitive equilibrium. Take $n = 2$ and assume $\hat{\mu} = \mu_1 + \mu_2 = e^{-1.3}$. Without loss of generality, we assume that supplier 1 has a higher capacity and let $a \equiv \frac{\mu_1}{\hat{\mu}} \in (0.5, 1)$ denote the heterogeneity of service rates between these two suppliers. Figure 3 demonstrates the mixing distributions of supplier 1 under a competitive equilibrium with different values of a. Figure 4 presents the upper and lower bounds of price of the mixing distributions. Note that the mixing distribution may have a point mass at the upper bound π_1^*, in which case the mixing distribution jumps to 1 at π_1^* (e.g., $a = 0.57, 0.64, 0.71$ in Figure 3). Although the mixing distributions of supplier 2 have no point mass, the comparison of the mixing distributions across different degrees of heterogeneity is qualitatively similar and therefore is omitted. Combining Figure 3 and Figure 4, there is no unambiguous prediction for the suppliers' pricing decisions when the capacity heterogeneity increases. The increase of the heterogeneity, a, has two effects. First, it mitigates the competition between the suppliers due to the difference of the capacity. This might induce higher prices. Second, the increase of a also increases the variance of the service time (since $\sigma = \sqrt{(\frac{1}{a\hat{\mu}})^2 + (\frac{1}{(1-a)\hat{\mu}})^2}$ is increasing in a in $(0.5, 1)$). This increases the magnitude of the price through the parameter β. Since the second-order price is negative, it implies that the suppliers would set a lower price when the variance is higher. Because of these two conflicting forces, no clear ranking of the mixing distribution can be obtained (as seen in Figure 3), and the bounds are non-monotonic in the degree of heterogeneity (see Figure 4). Note also that in the centralized solution, only one price is set: $\pi^C \equiv \arg\max_\pi \left\{\hat{\mu}\pi - \bar{p}\gamma\hat{\mu}\beta\frac{\phi(\pi/\beta)}{1-\Phi(\pi/\beta)}\right\} \in [4.5, 6.0]$ when $a \in [0.5, 0.71]$. Since π^C is strictly positive, the prices in the competitive equilibrium are significantly lower than that under the centralized control.

Finally, we investigate how the number of suppliers affects the efficiency gap between the centralized solution and the competitive equilibrium. To this end, we focus on the case with

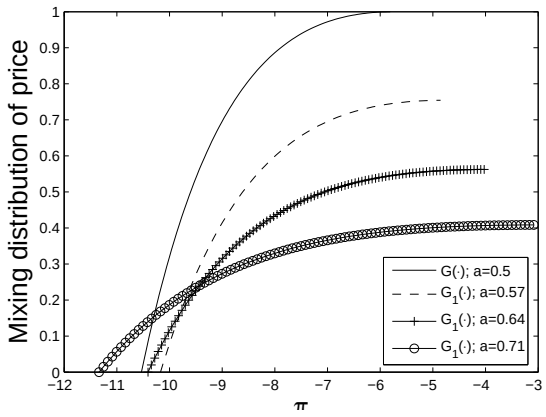

Figure 3: The mixing distribution of prices versus the heterogeneity of service rates.

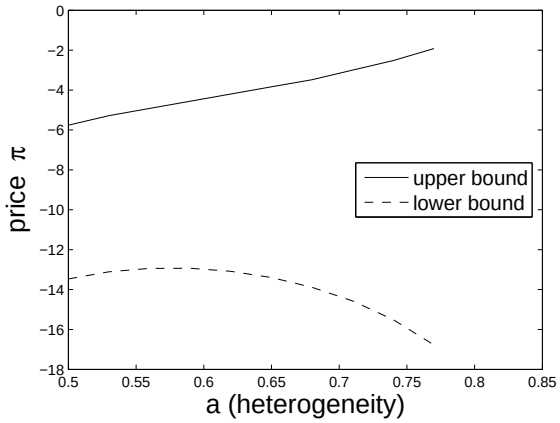

Figure 4: The bounds of prices versus the heterogeneity of service rates.

symmetric suppliers. This allows us to fully characterize the equilibrium pricing strategies and the suppliers' expected (second-order) revenues. We first assume $\hat{\mu} = e^{-1.3}$ and increase n, the number of suppliers. Given the symmetry, the individual service rate is $\mu_i = \frac{\hat{\mu}}{n}$, $\forall i \in \mathcal{N}$. As demonstrated in Figure 5, the range of price is shifted downwards when more suppliers participate in the market, which implies that the increase of n results in a more severe competition among suppliers. In Figure 6, we draw the expected aggregate (second-order) revenue of the market, $\sum_{i\in\mathcal{N}} \Psi_i(\pi_i, \pi_{-i})$, and vary the number of suppliers. We find that the expected aggregate revenue decreases when there are more suppliers due to the increasing price competition (as presented in Figure 6). Thus, the mis-coordination problem becomes more serious when more suppliers participate in the market.

A remark on the suppliers' participation

In characterizing the equilibrium behavior of the supplier pricing game, we have neglected the suppliers' participation decisions. Note that the pricing decisions $\{\pi_j\}$'s only affect the suppliers' second-order revenues, which are simply small perturbations around the first-order revenues $\bar{p}\mu_i$. (as seen in Lemma 1). Thus, a supplier is willing to participate if and only if his first-order revenue $\bar{p}\mu_i$ is positive, which depends on the capacity (service rate) decisions rather than the pricing decisions.

To study the capacity game, we can assume that each supplier incurs a cost of capacity, $C_i(\mu_i)$. Since the pricing decisions do not affect the first-order term, the suppliers choose their capacities

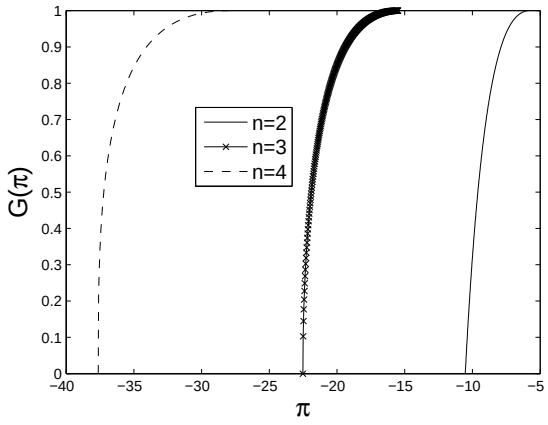

Figure 5: The mixing distribution of prices versus the number of suppliers.

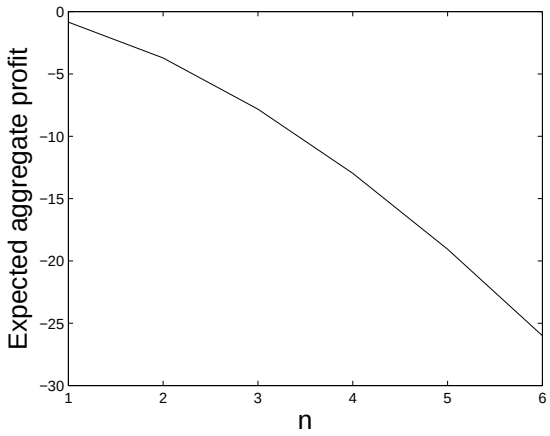

Figure 6: The expected aggregate revenue versus the number of suppliers.

to maximize

$$\max_{\mu_i \geq 0} \bar{p}\mu_i - C_i(\mu_i),$$

where $\bar{p}$ is endogenously determined through $\bar{F}(\bar{p}) = \sum_{j\in\mathcal{N}} \mu_j$, i.e., $\bar{p} = \bar{F}^{-1}(\sum_{j\in\mathcal{N}} \mu_j)$. The function $\bar{F}^{-1}(\sum_{j\in\mathcal{N}} \mu_j)$ can be interpreted as the inverse demand function, since it represents the customers' effective arrival rate given the aggregate capacity $\sum_{j\in\mathcal{N}} \mu_j$. Notably, the above capacity game does not involve any stochastic term.

Moreover, according to [Watts, 1996, Corollary 1], this capacity game has a unique equilibrium if the following conditions are satisfied: 1) $\mu\bar{F}^{-1}(\mu)$ is concave in μ; 2) $C_i(\mu_i)$ is weakly convex in μ_i; and 3) there exists a sufficiently large μ^* such that $\mu\bar{F}^{-1}(\mu) - C_i(\mu)$ is decreasing in when $\mu > \mu^*$. The first condition is related to the price elasticity of the demand, the second condition implies a diseconomy of scale for the capacity investment, and the third condition simply ensures that the aggregate market payoff never explodes. These conditions are widely adopted in many surplus sharing games, which contains the celebrated Cournot competition as a special case, to ensure that the competitive equilibrium is well-behaved (see Watts [1996] and the references therein).

Finally, a supplier is willing to participate in the market whenever in equilibrium $\max_{\mu_i\geq 0} \bar{p}\mu_i - C_i(\mu_i) \geq 0$. If we consider a special case in which the marginal cost of capacity is constant, i.e., $C_i(\mu_i) = c_i\mu_i, \forall\mu_i$, it is verifiable that only the suppliers that are more cost efficient will participate, ie., the ones for whom $c_i < \bar{p}$. The cost efficiency therefore ties in the suppliers' participation decisions.