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

Ying-Ju Chen, Costis Maglaras, and Gustavo Vulcano

**Abstract** We study an aggregated marketplace where potential buyers arrive and submit requests-for-quotes (RFQs). There are $n$ independent suppliers modeled 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 as the demand and the supplier capacities grow large, and subsequently extract insights about the equilibrium behavior of the suppliers. We show that supplier competition results into a mixed-strategy equilibrium phenomenon that is significantly different from the centralized solution. In order to overcome the efficiency loss, 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.

Ying-Ju Chen
School of Business and Management & School of Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, e-mail: imchen@ust.hk

Costis Maglaras
Columbia Business School, 409 Uris Hall, 3022 Broadway, New York, NY 10027 e-mail: c.maglaras@gsb.columbia.edu

Gustavo Vulcano
School of Business at Universidad Torcuato di Tella, Buenos Aires, Argentina, e-mail: gvulcano@utdt.edu

# 1 Introduction

## *1.1 Background and motivation*

Electronic markets (e-markets) have proliferated in the last decade or so as means to efficiently aggregate supply and demand for services or goods, in an effort to reduce search and transaction costs, improve market outcomes, and benefit both participants that supply and demand services. We study a mathematical model motivated by such marketplaces for services or goods that are subject to congestion effects, manifested in terms of delays until the good is delivered. The focus is on analyzing the market dynamics and gaining insight regarding the competition across suppliers when services are produced in a make-to-order manner (that is modeled via a queueing facility) and where congestion signals are state-dependent.

As part of the dynamics of the e-market evolution, several large-scale, web-based service sites have recently emerged. An incomplete list of examples includes ODesk, Elance, Vworker, Freelancer.com and Guru, which are freelancing sites that facilitate and streamline the process of hiring virtual or remote workers. In these platforms, many small service providers (e.g., individual computer programmers) seek work orders. Customers (e.g., employers) post job descriptions in the form of RFQs, and have service providers bid on the work. Customers then look at previous ratings and work history of the different candidates before settling on either a contract rate, or a pay-per hour agreement. Generally, money is escrowed by each of the websites (intermediaries), which release the payment to the service provider when the work is completed, while skimming a commission –typically 5-15% of money that changes hands. In addition, sometimes intermediaries also charge a membership fee to the parties involved. The volume of registered freelancers and effective transactions conducted through these platforms has been growing exponentially over time. Just to illustrate, the numbers of work hours per week transacted through oDesk were 50,000 by August 2008, 100,000 by June 2009, and 450,000 by July 2011.[1] An important portion of the projects auctioned out via these markets are complex and could be better addressed via a team as opposed to an individual. To better serve and bid in such cases, many freelance workers are represented via agents that pool capacity as traditional agents would do, and also provide project management in executing the complex projects so as to best use the pooled resources. These agents aggregate the capacity of many individual freelance providers.

In these settings, customers usually require specific skill sets, quality, and timeliness from their providers, and account for these needs as well as for cost in their utility function. This multidimensional assessment can be captured by a scoring index that a customer assigns to each potential provider. The bids are ranked, and the order is then awarded to the most "desirable" service provider. In this way, the final allocation for each work order is decided based on a reverse auction. Even though multiple service attributes can be subsumed in the scoring function, two of the most relevant ones are expected delay and price. Customers visiting these marketplaces

[1] Data available from www.odesk.com.

usually seek quick solutions and are willing to trade prices with waiting times. In this regard, aggregated marketplaces like the aforementioned ones raise several interesting practical and theoretical questions. In general, the role of the intermediary is passive, neutral, and limits to pooling supply and demand in an exchange platform. In this case: How does the system dynamics evolve as service providers (suppliers) compete for each potential order by posting a price and a state-dependent delay estimate? How should these suppliers determine their bidding strategies? Is the market efficient under competitive behavior? If the market were inefficient, then the role of the intermediary may become more active, in the sense of proposing a coordination scheme to align the suppliers' incentives. Is it possible to achieve this centralized optimal solution? If so, how can it be implemented in practical terms?[2]

We make some initial progress in addressing these questions. Specifically, we introduce a stylized mathematical model to study the aggregated marketplace in settings characterized by high volume of transactions. The goal of our study is to understand the dynamics and performance of the system, and gain insight on the pricing and capacity game among suppliers. We assume that potential buyers arrive according to a Poisson process and submit order requests, and that the suppliers (modeled as $M/GI/1$ queues) compete for these requests. Initially, suppliers decide the capacity (i.e., service rate) to offer, which is a static, long term decision. While operating, each supplier processes orders in a first-in-first-out manner, and submits a bid that comprises a fixed price and a target leadtime that depends on his own queue status. In fact, a key distinction of our work is that the delay quotations are dynamic rather than based on a steady-state assessment of the queue size. When suppliers submit bids, they face 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 long delay quotations thus increasing the full cost of the respective bid. The arriving buyer then uses a scoring function to compute the net utility associated with her bid, and awards the order to the lowest-quote supplier in order to maximize her own surplus (provided that it is nonnegative).

[2] There are other applications that share the same salient feature of several firms competing in offering some type of substitutable service that is differentiated with respect to its price and delay. Perhaps one of the most pervasive comes from the US equities market, which comprises of many exchanges, such as the NYSE, NASDAQ, ARCA, BATS, etc. Exchanges typically function as electronic limit order books, operating under a "price-time" priority rule, and their high-frequency dynamics can be modeled as multi-class queueing systems. Exchanges offer a rebate to liquidity providers, i.e., traders that post limit orders that "make" markets when their orders get filled, and charge a fee to "takers" of liquidity that initiate trades using marketable orders that transact against posted limit orders. The magnitudes of these make-take fees vary across exchanges and are comparable to the spread plus a significant fraction of the overall trading costs. Exchanges often change their fees and rebates in an effort to attract liquidity. Market participants employ so called "smart order routers" that take into account real-time market data, including queue and trading rate information, and formulate an order routing problem to trade off between rebates and a notion of expected delay, fill probabilities, and/or expected adverse selection. Once again, prices are fixed but delays are state dependent.

## 1.2 Overview of results

This appears to be one of the first papers to study competition in queues with substitutable products or services and state dependent congestion information. The discrete choice among substitutable products of potential consumers, the state dependent nature of the congestion signals, and the decentralized control among suppliers complicate the analysis of this system, rendering brute force analysis to be essentially intractable.[3]

The first contribution is to propose a tractable way for studying the decentralized market. As a preliminary step towards solving the capacity and pricing game, we analyze the queueing performance of this stochastic dynamic system assuming that the price vector is given. The solution to 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, our first set of results characterizes the behavior of the marketplace using an asymptotic analysis 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 (e.g., see [7] and [19]). Specifically, if $\Lambda$ is the market size, then the above result states that the economically optimal price is of the form $p^* = \bar{p} + \pi/\sqrt{\Lambda}$, where $\bar{p}$ is the price that induces full resource utilization in the absence of any congestion, and $\pi$ is a constant.

With the above fact in mind, we formulate the performance analysis sub-problem in a novel way that becomes asymptotically tractable in settings with large capacities and large volume of transactions. Specifically, based on the above observation, the starting point of our analysis is to write the suppliers' price bids as perturbations around the price $\bar{p}$ of the form $p_i = \bar{p} + \pi_i/\sqrt{\Lambda}$, for a constant $\pi_i$, where $\Lambda$ is viewed as a natural proxy for system size. Letting $\Lambda$ grow large, we derive the corresponding fluid and diffusion approximations. The fluid model transient analysis is helpful in establishing an important state space collapse (SSC) result through a variation of an approach developed by [8]. The SSC property establishes that the suppliers' dynamics are asymptotically coupled and can be described as a function

[3] The distinction regarding consumer choice model is important. With partially substitutable products, one could model consumer choice through a multi-product demand function, where the demand for one product depends on the price and delay of all products in a continuous manner. This is not the case with perfectly substitutable products, where demand may switch from one product to another in a discontinuous manner. For example, in a setting with two equally priced products, all consumers will select the one with the lower delay. This would increase the delay estimate of the faster option, causing all consumers to choose the alternative option. That is, small differences in price and delay, may lead to dramatic differences in demand for one of suppliers.

of the aggregate (system-wide) workload process, and, moreover, SSC implies that a supplier is able to know his competitors' bids by simply observing awaiting orders in his own buffer. We prove a weak convergence result of the workload process to a one-dimensional reflected Ornstein-Uhlenbeck (O-U) process, where interestingly the reflection point may be away from zero depending on the suppliers' prices. The latter implies the nonintuitive property that the aggregate workload process can never drain even though some of the suppliers may be idling, and this happens if the suppliers differ in their pricing. The derivation of the diffusion model extends "standard" results to this setting with self-interested routing policies based on dynamic information, which is of independent interest.

Second, using the asymptotic characterization of system behavior at a fixed choice of prices and capacities, we characterize the revenue stream of each supplier using the steady state properties of the reflected O-U process. This is then used to study the resulting pricing game. We find that the pricing game does not admit a pure strategy equilibrium. We specify the structure of the supporting mixed strategy equilibria where suppliers randomize over their pricing decisions. We also prove that the second-order efficiency loss of the decentralized solution can be arbitrarily large. 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 identified and used in our analysis. The framework of studying the appropriate asymptotic formulation of the aggregated market in the context with self interested buyers and state dependent congestion information, and using the derived diffusion to study the supplier game is novel. Such problems had not been studied in the literature before, in part due to their inherent complexity, and their proposed roadmap advanced seems to be of broader interest.

Third, the discrepancy between the centralized solution and the decentralized equilibrium calls for the development of a mechanism to coordinate the marketplace, in the sense that all suppliers would self-select to price according to the centralized solution. Our proposal relies on a transfer pricing scheme that compensates suppliers during idle periods. The existence of the intermediary in the motivating examples described above provides the natural support to implement it.

### *1.3 Literature review*

Our work touches on three related bodies of literature: 1) Economics of queues, 2) Competitive models in queueing contexts, and 3) Approximation schemes to analyze complex queueing models.

The literature that studies pricing in the context of single-server queues dates back to [23]. The demand model that we consider here is inspired by [21]: There is a single class 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. [22] extends that model to multiple customer types, and [1] extends it to a revenue maximization setting. In the context of queueing models with pricing and service competition, starting from the early papers by [16] and [18], 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 [16] and [18], competition is modeled in a duopoly setting where firms operate as $M/M/1$ queueing systems. Relaxations of the early papers include [17], which studies a variant of [18] in which the providers are modeled as symmetric $M/GI/1$ systems. [15] generalizes [17] for arbitrary number of service providers. [3] treats the price and waiting time cost as separate firm attributes that can be traded off differently by each arriving customer. These papers focused on customers making decisions based on steady-state performance measures.

Our use of asymptotic approximations and heavy traffic analysis to study the supplier game is motivated by the results of [7] and [19], who showed that in large scale systems the heavy traffic regime is the one induced by the revenue maximizing price. Our work implicitly assumes the validity of the heavy traffic regime in deriving its asymptotic approximation (as opposed to proving it as in the two papers above). The equilibrium pricing behavior of the competing suppliers supports the rationale of this assumption in the sense that no supplier wishes to price in a way that would deviate from that operating regime. The derivation of our limit model makes heavy use of the work by [20] on queues with state dependent parameters, and of the framework developed by [8] for proving state space collapse results. We also use technical results from [5] and [26] in our analysis. However, the combination of all our model features does not fit neatly the technical requirements of the aforementioned papers, as shown in the proofs contained in the online appendix.

A queueing paper that studies a model that is similar to our in a heavy traffic asymptotic regime is [24]. The key differences are the following: a) [24] assumes strictly convex delay cost functions as opposed to linear, b) it does not consider pricing (or some term that could account for its effect in the routing rule), and c) it does not allow for admission control decisions that can turn away users when the system is congested. The latter is captured by the behavior of self-interested users that differ in their valuations, and as a result will choose not join the market if the full cost exceeds their value. Taken together, these three elements necessitate a new analysis that leads to some insights that differ than what was observed in [24]. Perhaps the most notable difference is the fact that as a result of the pricing game, the workload process will not reflect at the origin, but instead it will reflect at some strictly positive quantity.

Recent papers accounting for congestion pricing include [11], which studies two types of contractual agreements in oligopolistic service industries. [4] appeals 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. More recently, [2] studies a large-scale marketplace with a moderating firm and numerous service providers. [2] also uses a static measure of the waiting time standard (usually the expected value or some percentile of the steady state distribution).

The remainder is organized as follows. In §2, we describe the model details. Next, §3 derives the asymptotic characterization of the marketplace behavior, and §4 characterizes the equilibrium behavior of the supplier pricing game. Finally, §5 includes our concluding remarks. All the proofs and more details of the analysis can be found in the technical report [10].

# 2 Model

## 2.1 Description of the market

We consider an aggregated marketplace where a homogeneous product (e.g., computer programming hours) is exchanged. The market functions as follows:
*Order arrivals:* Potential buyers arrive to the marketplace according to a Poisson arrival process with intensity $\Lambda$, and submit requests-for-quotes (RFQs). Each RFQ corresponds to the procurement of one unit of the product. 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 a RFQ process to procure one unit of the product.
*Suppliers:* The market is served by a set of suppliers $\mathscr{N} = \{1,\ldots,n\}$. 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 mean $1/\mu_i^{'}$ and standard deviation $\sigma_i$. Let $\hat{\mu} := \sum_{i\in\mathscr{N}} \mu_i^{'}/\Lambda$ be the (normalized) aggregated service rate of the market. We assume that the capacity vector $\mu^{'} \equiv \{\mu_i^{'}\}$'s is common knowledge. This fact can also be sustained by information provided by intermediary entities like the ones discussed in §1.
*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. (6) later on. We are assuming here that the supplier always submits a truthful estimate of the expected delay $d_i(t)$. In fact, in the presence of a market intermediary, the misreport of expected delays is discouraged through the display of past experiences of buyers with a given supplier, e.g. through publicly available ratings and reviews. This revealed information acts as a threatening device to favor the honest disclosure of suppliers' availabilities.

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., $v \geq \min_{i\in\mathscr{N}}\{p_i + c\,d_i(t)\}$; otherwise, the buyer leaves without submitting any order. Whenever a tie occurs, the order is awarded by randomizing uniformly among the cheapest suppliers.[4]

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

$$\lambda(p, d(t)) = \Lambda \bar{F}(\min_i\{p_i + c\,d_i(t)\}). \tag{1}$$

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.[5] 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 $\varepsilon(x) = -\frac{d\lambda(x)}{dx}\frac{x}{\lambda(x)}$. The expression $\varepsilon(x)$ can be regarded as the price elasticity of the demand rate, as it measures the proportional change of demand rate in response to the price change. We will make the following intuitive economic assumption:

**Assumption 1** $\lambda(p,0)$ *is elastic in the sense that* $\varepsilon(x) > 1$ *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 the minimum price would result in an increase in the market-wide aggregated revenue rate $p \cdot \lambda(p,0)$.[6] Of course, this would increase the utilization levels of the suppliers and lead to increased congestion and delays, thereby moderating 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 $t$,

$$A(t) = N\left(\Lambda \int_0^t \bar{F}(\min_{i\in\mathscr{N}}\{p_i + c\,d_i(s)\})ds\right), \tag{2}$$

[4] 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.

[5] 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 $v \sim \mathrm{Exp}(\alpha)$, then the demand is exponential, with $\lambda(x) = \Lambda\, e^{-\alpha x}$.

[6] This implication follows directly from the economics literature. When $\varepsilon(x) > 1$, the proportional increase of demand rate is larger than the proportional decrease of price. As the revenue is the product of price and demand rate, the aggregated revenue rate ends up being higher. Suppose further that there are no congestion effects, 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 $\Lambda$, 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.

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

$$\mathscr{J}(t) \equiv \{i \in \mathscr{N} : p_i + cd_i(t) \leq p_j + cd_j(t), \forall j \in \mathscr{N}\} \tag{3}$$

as the set of cheapest suppliers at time $t$. Further, define $\Xi_{\mathscr{J}(t)}$ as the random variable that assigns the orders uniformly amongst the cheapest suppliers. That is, $\Xi_{\mathscr{J}(t)} = i$ with probability $\frac{1}{|\mathscr{J}(t)|}$ if $i \in \mathscr{J}(t)$, where $|\mathscr{J}(t)| > 0$ is the cardinality of $\mathscr{J}(t)$, and $\Xi_{\mathscr{J}(t)} = i$ with zero probability otherwise. For the ease of the exposition, we will assume that $\mathscr{J}(t)$ and $\Xi_{\mathscr{J}(t)}$ 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

$$A_i(t) = \int_0^t 1\!\!1\{\Xi_{\mathscr{J}(s)} = i\} dA(s), \tag{4}$$

where $1\!\!1\{\cdot\}$ is the indicator function. Note also that $A(t) = \sum_{i \in \mathscr{N}} A_i(t)$.
*Supplier dynamics:* Let $Q_i(t)$ denote supplier $i$'s 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 a time $t$ when 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:

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t). \tag{5}$$

Given this notation, then

$$d_i(t) = \frac{Q_i(t) + 1}{\mu_i'}, \tag{6}$$

is the expected sojourn time of the new incoming order, given that it gets awarded to supplier $i$ and the current queue length is $Q_i(t)$. Under our modeling assumptions, supplier $i$ will therefore bid $(p_i, d_i(t))$, where $d_i(t)$ is given by (6). Each supplier knows his own system queue length $Q_i(t)$, but is not informed about his competitors' queue lengths.

## 2.2 Problems to address

We study three problems related to the market model described above:

1. *Performance analysis for a given $p$ and $\mu'$*: 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 average revenue is

$$\Omega_i(p_i, p_{-i}) \equiv p_i \cdot \lim_{t\to\infty} \frac{S_i(T(t))}{t}, \tag{7}$$

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 study the competitive equilibrium that characterizes the supplier capacity and pricing games, both of which are one-shot games where each supplier selects his service rate and static price, successively. We further show that the capacity selections constitute the first-order effects on the suppliers' payoffs and the pricing decisions are of second order; thus, we can conveniently decouple the equilibrium analysis into two separate stages. For the first-stage, capacity game, since the capacities (service rates) are assumed to be publicly observable, we adopt the Nash equilibrium as our solution concept. For the second-stage, pricing game, the suppliers may be uninformed about the queue lengths of the competitors, which may potentially lead to information incompleteness. However, as we will show, the suppliers' competitive behavior is insensitive to this knowledge; consequently, we adopt again the standard Nash equilibrium (under complete information) as our solution concept. Given the revenue specified in (7), a Nash equilibrium $\{p_i^*\}$ requires that $p_i^* = \arg\max_{p_i} \Omega_i(p_i, p_{-i}^*)$, $\forall i \in \mathscr{N}$.
3. *Market efficiency and market coordination:* Our objective here is to characterize the efficiency loss:

$$\max_p \left\{ \sum_{i\in\mathscr{N}} \Omega_i(p_i, p_{-i}) \right\} - \sum_{i\in\mathscr{N}} \Omega_i(p_i^*, p_{-i}^*), \tag{8}$$

i.e., the difference between the revenue of a system where a central planner would control the pricing decision of each supplier (i.e., the *first best solution*), and the sum of the revenues collected in the competitive framework. 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 allocation of orders. Our approach is to develop an approximate model for the market dynamics that is rigorously validated in settings where the demand volume 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 provides 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 supplier*

As a motivation for our subsequent analysis, this subsection 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 $\Lambda$ and the processing capacity $\mu$ grow large. Specifically, we will consider a sequence of problem instances indexed by $r$, where $\Lambda^r = r$ and $\mu^r = r\mu$; 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. That is, neglecting congestion effects, $\hat{p}$ is the price that maximizes the revenue rate and $\bar{p}$ is the price that induces full resource utilization, and both of these quantities are independent of $r$. Assumption 1 implies that $\hat{p} < \bar{p}$ (or equivalently, $\Lambda^r\bar{F}(\hat{p}) > \mu^r$) thus accentuating the tension between revenue maximization and the resulting congestion effects. [7] showed that the revenue maximizing price, denoted by $p^{*,r}$, is of the form

$$p^{*,r} = \bar{p} + \pi^*/\sqrt{r} + o(1/\sqrt{r}), \tag{9}$$

where $\pi^*$ is a constant independent of $r$. Moreover, the resulting queue lengths $Q^r(t)$ 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 $1/\sqrt{r}$. Following [19], the delays are moderate in absolute terms (of order $1/\sqrt{r}$) but significant when compared to the actual service time (of order $1/r$). If the supplier can select the price and capacity $\mu$, the latter assuming a linear capacity cost, then the optimal capacity choice is indeed such that $\hat{p} < \bar{p}$ (where $\bar{p}$ is determined by $\mu$), 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 ([7]).

### 3.2 Setup for asymptotic analysis

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

1. Define the normalized parameters $\mu_i = \mu_i'/\Lambda$ for every supplier $i$.
2. Define $\bar{p}$ to be the price such that $\Lambda \bar{F}(\bar{p}) = \sum_{i \in \mathscr{N}} \mu_i'$. 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:
$$\Lambda^r = r, \;\; \mu_i^r = r\mu_i, \; \forall i \in \mathscr{N}, \;\; c^r = c, \;\; v \sim F(\cdot), \tag{10}$$
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 takes the form assumed in item 3) above. This would be true if the market were managed by a central planner that could coordinate the supplier pricing and capacity decisions. The approach we pursue is to embed the system we wish to study in the sequence of systems indexed by $r$ and described in (10), 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 (10), where $\Lambda$ 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 $\Lambda$ 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 $(p, \mu')$ that are embedded in the sequence (10).

### 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

$$\bar{Q}_i(t) = \bar{Q}_i(0) + \bar{A}_i(t) - \bar{D}_i(t), \forall i \in \mathscr{N}, \tag{11}$$

$$\bar{A}_i(t) = \int_0^t \frac{1}{|\mathscr{I}(t)|} 1\!\!1\{i \in \mathscr{I}(t)\} \Lambda \bar{F}(\bar{p}) ds, \forall i \in \mathscr{N}, \tag{12}$$

$$\bar{D}_i(t) = \mu_i \bar{T}_i(t), \forall i \in \mathscr{N}, \tag{13}$$

$$\int_0^t \bar{Q}_i(s) d\bar{Y}_i(s) = 0, \forall i \in \mathscr{N}, \tag{14}$$

$$\bar{T}_i(t) + \bar{Y}_i(t) = t, \forall i \in \mathscr{N}, \tag{15}$$

$$\bar{W}(t) = \sum_{i \in \mathscr{N}} \frac{\bar{Q}_i(t)}{\mu_i}. \tag{16}$$

captures the market's transient dynamics over short periods of length $1/\sqrt{r}$. This subsection studies the transient evolution of (11)-(16) starting from arbitrary initial conditions.

In (11)-(16), the processes $\bar{Q}, \bar{A}, \bar{D}, \bar{T}, \bar{Y}$ are the fluid analogues of $Q, A, D, T, Y$ defined in §2, and $\bar{W}$ is the fluid analog of the system workload $W$. Equation (11) keeps track of the queue sizes. Equation (12) 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 §2. From Equation (12), the orders get awarded to the various suppliers at a rate $\Lambda \bar{F}(\bar{p}) = \sum_{i \in \mathscr{N}} \mu_i'$, i.e., $\bar{F}(\bar{p}) = \sum_{i \in \mathscr{N}} \mu_i$ (as indicated by the aggregate counting process $N(\bar{F}(\bar{p})t)$). Equation (14) demonstrates the non-idling property: $\bar{Y}_i(t)$ cannot increase unless $\bar{Q}_i(t) = 0$. Equation (15) is a time-balance constraint. Finally, Equation (16) establishes the connection between the total workload and the queue lengths.

The next proposition establishes that starting from any arbitrary initial condition, the transient evolution of the market (as captured through (11)-(16)) converges to a state configuration where all suppliers are equally costly in terms of the full cost of the bids given by (price + $c$ × 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 onwards, 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 share the same full cost.

**Proposition 1** *Let $\bar{Q}, \bar{A}, \bar{D}, \bar{T}, \bar{Y}, \bar{W}$ be the solution to (11)-(16) 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*

$$\max_{i,j \in \mathscr{N}} \left| \left( \pi_i + c\frac{\bar{Q}_i(s)}{\mu_i} \right) - \left( \pi_j + c\frac{\bar{Q}_j(s)}{\mu_j} \right) \right| < \delta, \ \forall s > s(\delta, M_0). \tag{17}$$

### 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 evolve 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 $W^r(t) = \sum_{i\in\mathscr{N}} \dfrac{Q_i^r(t)}{\mu_i^r}$.

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 the re-scaled queue length processes for all suppliers according to

$$\tilde{Q}_i^r(t) = \frac{Q_i^r(t)}{\sqrt{r}}. \tag{18}$$

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

Define $\tilde{Z}^r(t) = \bar{\pi} + \bar{c}\tilde{W}^r(t)$, where $\bar{\pi} = \sum_{i\in\mathscr{N}} \pi_i/n$ and $\bar{c} = c/n$. $\tilde{Z}^r(t)$ can be regarded as a proxy for the average of the second-order terms of suppliers' bids since $\tilde{Z}^r(t) = \dfrac{1}{n}\sum_{i\in\mathscr{N}}\left(\pi_i + c\dfrac{\tilde{Q}_i^r(t)}{\mu_i}\right)$. 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* $\pi_i + c\dfrac{\tilde{Q}_i^r(0)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(0)$ *in probability,* $\forall i \in \mathscr{N}$*. Then, for all* $\tau > 0$*, for all* $\varepsilon > 0$*, as* $r \longrightarrow \infty$*,*

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

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

The proof applies the "hydrodynamic scaling" framework of [8], which is 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 [8], but as we show in the online 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 those of $\tilde{W}^r(t)$ and $\tilde{Q}^r(t)$.

**Theorem 1.** (WEAK CONVERGENCE) *Suppose* $\pi_i + c\dfrac{\tilde{Q}_i^r(0)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(0)$ *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*

$$\tilde{Z}(t) = \tilde{Z}(0) - \gamma c \int_0^t \tilde{Z}(s)ds + \tilde{U}(t) + \frac{c\sqrt{\sigma^2 + \hat{\mu}}}{\hat{\mu}} B(t), \tag{19}$$

*where* $B(t)$ *is a standard Brownian motion,* $\tilde{U}(0) = 0$, $\tilde{U}(t)$ *is continuous and nondecreasing, and* $\tilde{U}(t)$ *increases only when* $\tilde{Z}(t) = \hat{\pi}$. *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,* $\tilde{W}^r(t) \Rightarrow \dfrac{1}{\bar{c}}(\tilde{Z}(t) - \bar{\pi})$, *and* $\tilde{Q}_i^r(t) \Rightarrow \dfrac{\mu_i}{c}(\tilde{Z}(t) - \pi_i)$, $\forall i \in \mathcal{N}$.

The process $\tilde{U}(t)$ is the limiting process of $\tilde{U}^r(t) \equiv \dfrac{c}{\hat{\mu}} \sum_{i\in\mathcal{N}} \mu_i \tilde{Y}_i^r(t)$ (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 $\tilde{Z}(t) \geq \hat{\pi}$, 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 at a given time $t$, 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, 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. $\gamma$, that controls the speed of the reversion of the aggregate workload process, is extracted from the customer valuation distribution. $\gamma$ measures the sensitivity of the demand function to changes to the full price "$\pi + cd(t)$", and it is proportional to the demand elasticity at $\bar{p}$.

To summarize, in the limit model, the suppliers' queue length processes follow from $\tilde{Q}_i(t) = \dfrac{\mu_i}{c}(\tilde{Z}(t) - \pi_i)$, $\forall i \in \mathcal{N}$, where $\tilde{Z}(t)$ is defined through (19). The next section will use this result as an input to study the suppliers' pricing game.

### 3.6 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 0.1. 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 $\bar{\pi} = \frac{\pi_1 + \pi_2}{2} = -1.5$.

We run simulations using Arena (a discrete-event simulation software) and at places supplement it with Matlab to compute the relevant parameters. The Arena model is illustrated in the online appendix.

In Figures 1 and 2, we illustrate the workload trajectories for $r = 30$ and $r = 80$, respectively, for one replication. Given these parameters, the respective boundaries are 0.365 and 0.224 for $r = 30$ and $r = 80$. Note that our mathematical statement is established on the steady-state workload process. We can either perform a very long run (say with 600,000 arrivals in expectation) and break each output record from the (single) run into a few large batches. Alternatively, we can run many replications and identify appropriate warm-up and run-length times (for example 3,000 replications, each of which generates roughly 200 arrivals). We find that both approaches lead to very similar statistical outcomes, and hence we report only the former. When $r = 80$ and the simulation time is 600,000/80 units, we find that 92.93% of time the system workload is above the boundary 0.224. When $r = 30$ (and the corresponding simulation time is 600,000/30 units), however, this proportion goes down to 71.22%. This suggests that our mathematical result of asymptotically negligible time below boundary is more applicable when the scaling factor is large.

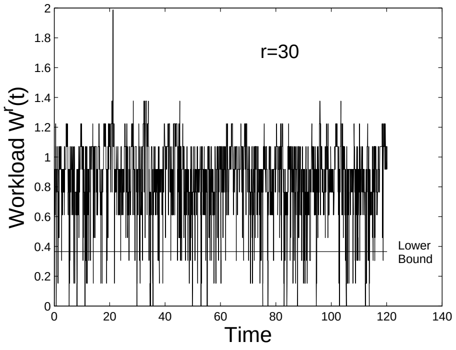

**Fig. 1** An instance of workload process when $r = 30$.

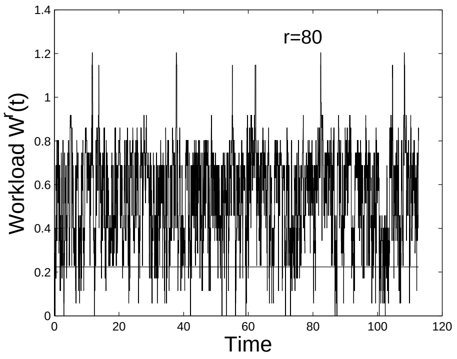

**Fig. 2** An instance of workload process when $r = 80$.

## 4 Competitive behavior and market efficiency

In this section, we discuss the equilibrium behavior of suppliers in the capacity and pricing games described in §2. We use the performance characterization of §3 and cast the suppliers' prices as "small" deviations around the market clearing price $\bar{p}$. This distinction results in a first-stage capacity game that affects the first-order revenues, and a second-stage, pricing game, that adjusts prices around the first-order price. Then, we 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' first-order payoffs and the capacity game*

Let $R_i^r(t)$ denote supplier $i$'s cumulative revenue. Since the pricing is static, supplier $i$ earns $R_i^r(t) = (\bar{p} + \dfrac{\pi_i}{\sqrt{r}})S_i^r(t - Y_i^r(t))$, 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* $\quad r \to \infty, \forall i \in \mathscr{N}.$

Lemma 1 demonstrate that the capacity choice ($\mu_i$) has a first-order effect on the suppliers' revenues. If we attach an appropriate capacity cost $c_i(\mu_i)$ to the suppliers, the capacity game can be explicitly posted. In the centralized system, a central planner decides the service rates (capacities) to maximize the net revenue:

$$\max_{\{\mu_i\}} \left\{ \sum_i \mu_i \bar{p}(\sum_i \mu_i) - \sum_i c_i(\mu_i) \right\}. \tag{20}$$

This centralized solution can be obtained via a two-stage problem in which we first find the optimal allocation for each individual to minimize the aggregate cost:

$$C(\hat{\mu}) = \min_{\{\mu_i\}} \left\{ \sum_i c_i(\mu_i), s.t. \ \sum_i \mu_i = \hat{\mu}, \ \ \mu_i \geq 0, \ \forall i \right\}, \tag{21}$$

and then optimize over the aggregate service rate via $\max_{\hat{\mu}} \{\hat{\mu}\bar{p}(\hat{\mu}) - C(\hat{\mu}) | \ \hat{\mu} \geq 0\}$.

In a Nash equilibrium $\{\mu_i^*\}$'s, supplier $i$ chooses a capacity such that

$$\mu_i^* = \arg\max_{\mu_i} \left\{ \mu_i \bar{p}(\mu_i + \sum_{j \neq i} \mu_j^*) - c_i(\mu_i) \right\}, \tag{22}$$

where $\bar{p} = (\bar{F})^{-1}(\sum_i \mu_i/\lambda)$ is the price that induces the full resource utilization. Furthermore, if $\hat{\mu}\bar{p}(\hat{\mu})$ is concave in $\hat{\mu}$ and $c_i(\mu_i)$ is convex in $\mu_i$, $\forall i$,[7] there exists a pure-strategy Nash equilibrium in the capacity game. Based on this existence result, we can characterize the pure-stra tegy equilibrium from the best responses of suppliers against others' strategies. Specifically, a Nash equilibrium$\{\mu_i^*\}_{i=1}^n$ satisfies

$$\mu_i^* \bar{p}'(\sum_i \mu_i^*) + \bar{p}(\sum_i \mu_i^*) - c_i'(\mu_i^*) = 0, \ \forall i. \tag{23}$$

It can be verified that in the decentralized (Nash) equilibrium, each supplier intends to build a capacity $\mu_i^*$ higher than the centralized solution. This over-investment result follows from the ignorance of the negative externality a supplier brings to the entire system, as a supplier may benefit from over-investment since this allows him to capture a higher market share. This is reminiscent of the demand-stealing effect in the classical Cournot competition.

## 4.2 Suppliers' second-order payoffs and the pricing game

To study the suppliers' pricing game we will focus on the second order correction around $R_i^r(t)$ defined as $r_i^r(t) \equiv \frac{1}{\sqrt{r}}(R_i^r(t) - r\bar{p}\mu_i t), \forall i \in \mathscr{N}$. The limiting processes of these corrected terms are characterized in the following lemma.

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

Instead of using the revenue functions $\{\Omega_i(p_i, p_{-i})\}$'s defined in (7), we will study the suppliers' pricing game based on their (second-order) revenues given by

$$\Psi_i(\pi_i, \pi_{-i}) \equiv \lim_{t\to\infty} \frac{r_i(t)}{t} = \mu_i(\pi_i - \bar{p}E[\tilde{Y}_i(\infty)]), \tag{24}$$

where $\pi_{-i} \equiv (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$, and (with some abuse of notation) $E[\tilde{Y}_i(\infty)] := \lim_{t\to\infty} \frac{\tilde{Y}_i(t)}{t}$. Define $h_i$ as the (steady-state) proportion of the market idleness incurred by supplier $i$, i.e.,

$$E[\tilde{Y}_i(\infty)] = h_i E[\tilde{U}(\infty)], \tag{25}$$

where we again denote by $E[\tilde{U}(\infty)] := \lim_{t\to\infty} \frac{\tilde{U}(t)}{t}$ the long-run average of the aggregate idleness $\tilde{U}(t)$ specified in Theorem 1.

Dividing (19) by $t$ and letting $t \to \infty$, we obtain that

$$E[\tilde{U}(\infty)] = \lim_{t\to\infty} \frac{\tilde{U}(t)}{t} = \gamma c E[\tilde{Z}(\infty)] = \gamma c \beta \frac{\phi(\hat{\pi}/\beta)}{1 - \Phi(\hat{\pi}/\beta)}, \tag{26}$$

[7] These assumptions are commonly adopted in revenue management, although in some context the arrival rate is used instead of the service rate, which makes no difference in heavy traffic regime, see e.g. [13].

where

$$\beta = \sqrt{\frac{c(\hat{\mu}+\sigma^2)}{2\gamma\hat{\mu}^2}}, \tag{27}$$

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 ([9, Proposition 1]).[8] We do not derive the closed-form expressions of $\{h_i\}$'s since 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). From Theorem 1 we have that $\tilde{Z}(t) \geq \hat{\pi}$ for all $t \geq 0$ and that $\tilde{Q}^r_j(t) \Rightarrow \frac{\mu_j}{c}(\tilde{Z}(t) - \pi_j) > 0$, for all $t \geq 0, \forall j \notin J$. It follows that $\tilde{Y}_i(t) = 0$ for all $t \geq 0$, and therefore $h_j = 0$, $\forall j \notin J$. Given (24) and (25), we obtain the suppliers' second-order long-run average revenue functions as follows:

$$\Psi_i(\pi_i, \pi_{-i}) = \begin{cases} \mu_i\pi_i - \mu_i\bar{p}h_i\gamma c\beta\frac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)}, & if\ i \in J, \\ \mu_i\pi_i, & otherwise. \end{cases} \tag{28}$$

### 4.3 Centralized system performance

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

$$\max_{\pi} \left\{ \sum_{i\in\mathcal{N}} \mu_i\pi_i - \bar{p}\gamma\hat{\mu}\beta\frac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)} \,:\, \pi_i \leq \hat{\pi} \right\}, \tag{29}$$

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 $\mathscr{L}(\hat{\pi}) \equiv \bar{p}\gamma\hat{\mu}\beta\dfrac{\phi(\hat{\pi}/\beta)}{1-\Phi(\hat{\pi}/\beta)}$ 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 $\pi_i$'s are equal. The optimal static price is $\pi^C := \arg\max_{\pi}[\hat{\mu}\pi - \mathscr{L}(\pi)]$.*

[8] 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 $\hat{\pi}$.

## 4.4 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 $\mathscr{L}(\hat{\pi})$, we can rewrite the supplier's payoff in (28) as

$$\Psi_i(\pi_i, \pi_{-i}) = \mu_i \pi_i - \mu_i h_i \frac{c}{\hat{\mu}} \mathscr{L}(\hat{\pi}) 1\!\!1\{i \in J\}] \tag{30}$$

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

### 4.4.1 Homogeneous service rate case

We first consider the case where the service rates are the same across suppliers, i.e., $\mu_i = \mu_j \equiv \mu, \ \forall i, j \in \mathscr{N}$ , and focus on symmetric equilibria. Define $\pi^* = \arg\max_\pi [\mu \pi - \mathscr{L}(\pi)]$ and $\Psi^* \equiv \mu \pi^* - \mathscr{L}(\pi^*)$. Note that we are charging all the idling penalty to a single supplier. In this way, a price $\pi^*$ guarantees a lower bound for the payoff $\Psi_i(\pi_i, \pi_{-i})$. Hence, $\Psi^*$ is the payoff that any supplier can guarantee for himself, i.e., his *minmax* level. We further let $\underline{\pi} := \Psi^*/\mu = \pi^* - \mathscr{L}(\pi^*)/\mu < \pi^*$ 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.

Although a standard approach is to look for a pure-strategy Nash equilibrium, in the next proposition we show that none exists. Instead, we shall focus on the mixed-strategy competitive equilibrium. Let $G(\pi)$ denote the mixing cumulative probability distribution of a supplier's pricing strategy $\pi$. The next proposition characterizes the structure of these mixing probabilities.

**Proposition 3** *With homogeneous rates, there exists a unique symmetric equilibrium in which all suppliers randomize continuously over $[\underline{\pi}, \pi^*]$, and every supplier gets $\Psi^*$. The randomizing distribution is $G(\pi) = [\frac{\mu(\pi - \underline{\pi})}{\mathscr{L}(\pi)}]^{1/(n-1)}, \forall \pi \in [\underline{\pi}, \pi^*]$.*

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. 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 additional revenues.

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

**Proposition 4** *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 $|\Pi^C - \Pi^*| > M$, $\forall n > N_M$.*

Proposition 4 shows that as the number of suppliers 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 a coordination scheme, as we investigate in §4.5. Note that this statement is asymptotic in the sense of the number of suppliers, which is different from the case in §3, and it is particularly relevant in the context of the large-scale systems discussed in §1. By restricting ourselves to the case of fixed aggregate service rate, we can then illustrate that the efficiency loss that results from the market idleness term also plays a pivotal role.

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 §3.

### 4.4.2 Heterogeneous service rate case

Now we consider the scenario where suppliers are endowed with different service rates. We again first define a global maximizers $\pi_1^*$, $\pi_2^*$,..., $\pi_n^*$, if supplier $i$ ($1 \leq i \leq n$) is the one who proposes the highest price solely; i.e., we define $\pi_i^* = \arg\max_{\pi_i} [\mu_i \pi_i - \mathscr{L}(\pi_i)]$ and $\Psi_i^* \equiv \mu_i \pi_i^* - \mathscr{L}(\pi_i^*)$ 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 5** *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, $\underline{s} \geq \max_{i \in \mathscr{N}} \underline{\pi}_i$.*
- *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*

$$G_2(\pi) = \frac{\mu_1(\pi - \underline{\pi}_1)}{\mathscr{L}(\pi)}, \forall \pi \in [\underline{\pi}_1, \pi_1^*], \; G_1(\pi) = \frac{\mu_2(\pi - \underline{\pi}_1)}{\mathscr{L}(\pi)}, \forall \pi \in [\underline{\pi}_1, \pi_1^*), \quad (31)$$

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

The first result on left endpoints is not surprising. This comes directly from an analogous argument for Proposition 3. 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 for 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 a capacity advantage, he can afford to price lower to capture more customers.

### 4.4.3 Numerical results

In this section, 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 0.1, and $\bar{p}$ is set such that the effective arrival rate $P(v \geq \bar{p})$ 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 $\Lambda = r$ and $\hat{\mu} = \hat{\mu}^r$, $\bar{p}$ 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 $a \equiv \dfrac{\mu_1}{\hat{\mu}} \in (0.5, 1)$ 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 the price for the mixing distributions. Note that the mixing distribution may have a point mass at the upper bound $\pi_1^*$, in which case the mixing distribution jumps to 1 at $\pi_1^*$ (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 capacities. This might induce higher prices. Second, the increase of $a$ also increases the variance of the service time (since $\sigma = ((\frac{1}{a\hat{\mu}})^2 + (\frac{1}{(1-a)\hat{\mu}})^2)^{1/2}$ is increasing in $a \in (0.5, 1)$). This increases the magnitude of the second-order price through the parameter $\beta$. 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 not monotonic (in the same direction) as the degree of heterogeneity increases (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 $\pi^C$ is strictly positive, the prices in the competitive equilibrium are significantly lower than the price under the centralized control.

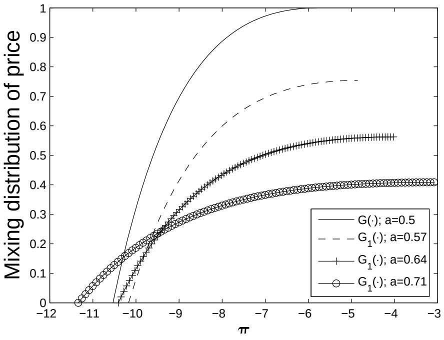

**Fig. 3** The mixing distribution of prices versus the heterogeneity of service rates.

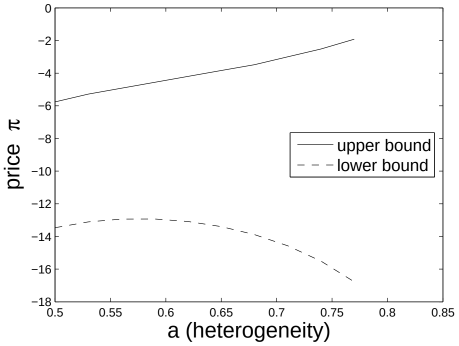

**Fig. 4** The bounds of prices versus the heterogeneity of service rates.

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 homogeneous 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. The individual service rate is $\mu_i = \hat{\mu}/n, \forall i \in \mathscr{N}$. As demonstrated in Figure 5, the range of prices becomes more negative when more suppliers participate in the market, due to a more severe competition among suppliers. In Figure 6, we draw the expected aggregate (second-order) revenue of the market, $\sum_{i \in \mathscr{N}} \Psi_i(\pi_i, \pi_{-i})$, 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.

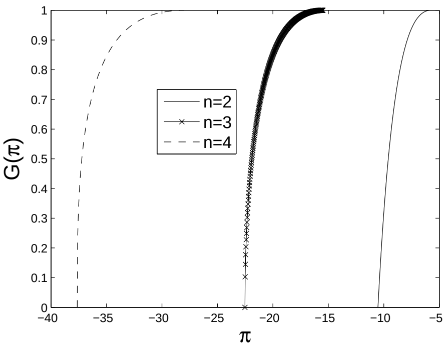

**Fig. 5** The mixing distribution of prices versus the number of suppliers.

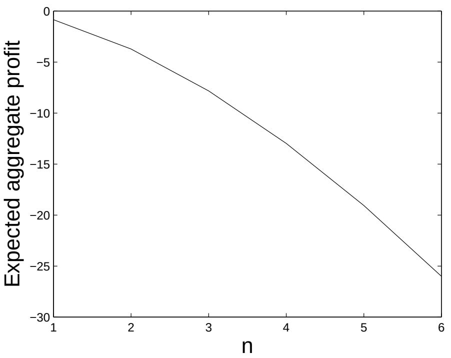

**Fig. 6** The expected aggregate revenue versus the number of suppliers.

### 4.4.4 A remark on the suppliers' participation

In characterizing the equilibrium behavior of the suppliers' 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 to maximize

$$\max_{\mu_i \geq 0} \bar{p}\mu_i - c_i(\mu_i), \tag{32}$$

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 [25, Corollary 1], this capacity game has a unique equilibrium if the following conditions are satisfied: 1) $\mu\bar{F}^{-1}(\mu)$ is concave in $\mu$; 2) $c_i(\mu_i)$ is weakly convex in $\mu_i$; and 3) there exists a sufficiently large $\mu^*$ such that $\mu\bar{F}^{-1}(\mu) - c_i(\mu)$ is decreasing in $\mu$ 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 [25] 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 $c_i < \bar{p}$.

## 4.5 Coordination scheme

The above competitive equilibrium analysis reveals that 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. In addition, as the market size grows, the competitive behavior among suppliers may result in an unbounded efficiency loss. This demonstrates a significant inefficiency due to the market mechanism and motivates the search for a coordination scheme.[9]

### 4.5.1 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

$$\Psi_i^{PS}(\pi_i, \pi_{-i}) \equiv \mu_i \pi_i - \frac{\mu_i}{\hat{\mu}} \mathscr{L}(\max_{j \in \mathscr{N}} \pi_j), \tag{33}$$

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 $\frac{\mu_i}{\hat{\mu}} \mathscr{L}(\max_{j \in \mathscr{N}} \pi_j)$ corresponds to his penalty share that is proportional to his service rate $\mu_i$. Note that this scheme is budget-balanced, i.e., no financing from outside parties is required. Let $\{\pi_i^{PS}\}' s$ denote the equilibrium prices under this sharing scheme. Under this sharing scheme, the centralized solution can be achieved.

**Proposition 6** *Under the penalty sharing schemes that satisfy (33), $\{\pi_i^{PS} = \pi^C, \forall i \in \mathscr{N}\}$ is the unique equilibrium.*

[9] It is worth mentioning that the mixed-strategy equilibrium is studied mainly to demonstrate the discrepancy between the centralized system and the decentralized market equilibrium. It is conceivable that the implementation or identification of such a mixed-strategy equilibrium requires fairly sophisticated communication and consensus among the suppliers. Nevertheless, the Pareto dominance result justifies why such a coordination scheme is required and desired irrespective of the implementation issue of the mixed-strategy equilibrium.

Under the *PS* scheme, a supplier's objective is in fact an affine function of the aggregate revenue (29). Hence, this coordination mechanism eliminates the wrong incentives of suppliers, regardless of the number of suppliers and their service rates. Since in both the competitive and the coordinated equilibria the suppliers' expected payoffs are proportional to their service rates, all suppliers have a natural incentive to join.

### 4.5.2 "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 $\eta_{ij}$ 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

$$\tilde{r}_i^{PS}(t) = \tilde{r}_i(t) + \tilde{\delta}_i(t), \tag{34}$$

where

$$\tilde{r}_i(t) \equiv \mu_i \pi_i t + \bar{p}\sigma_i B_{s,i}(t) - \mu_i \bar{p}\tilde{Y}_i(t), \tag{35}$$

$$\tilde{\delta}_i(t) \equiv \sum_{j\in\mathscr{N}, j\neq i} \eta_{ji}\tilde{Y}_i(t) - \sum_{j\in\mathscr{N}, j\neq i} \eta_{ij}\tilde{Y}_j(t). \tag{36}$$

According to (35), the three terms correspond to his revenue from serving customers. In (36), $\tilde{\delta}_i(t)$ corresponds to the net transfers for supplier $i$ : $\sum_{j\in\mathscr{N}, j\neq i} \eta_{ji}\tilde{Y}_i(t)$ is the compensation he receives from other suppliers during the idle period, and $\sum_{j\in\mathscr{N}, j\neq i} \eta_{ij}\tilde{Y}_j(t)$ is the cash outflow to other suppliers while compensating their idleness.

Given (34), supplier $i$'s long-run average revenue can be expressed as

$$\begin{aligned}\tilde{\Psi}_i(\pi_i, \pi_{-i}) &\equiv \lim_{t\to\infty} \frac{1}{t}[\tilde{r}_i(t) + \tilde{\delta}_i(t)] \qquad (37)\\ &= \mu_i\pi_i - \mu_i\bar{p}E[\tilde{Y}_i(\infty)] + \sum_{j\in\mathscr{N}, j\neq i} \eta_{ji}E[\tilde{Y}_i(\infty)] - \sum_{j\in\mathscr{N}, j\neq i} \eta_{ij}E[\tilde{Y}_j(\infty)].\end{aligned}$$

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

**Proposition 7** *The transfer prices*

$$\eta_{ij} = \frac{\mu_i\mu_j}{\hat{\mu}}\bar{p}, \forall i \neq j, i, j \in \mathscr{N}, \text{ and } \eta_{ii} = 0, \forall i \in \mathscr{N}, \tag{38}$$

*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 7 essentially eliminate the imbalance between the current share of the market idleness incurred by an individual supplier and his required share $(\frac{\mu_i}{\hat{\mu}}\mathscr{L}(\max_{j\in\mathscr{N}}\pi_j))$. 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 (33)), and thus all suppliers are induced to set prices equal to $\pi^C$.

Proposition 7 shows that we are able to achieve coordination since given any chosen prices, we can align the suppliers' objectives with the planer's objective. To implement this compensation scheme in the original system, we can simply request each supplier make transfers according to (38). This mechanism can be implemented and monitored by the intermediary in the market.

Note that the coordination scheme is independent of the static prices $\{\pi_i\}$'s; it only requires the information of the service rates $\{\mu_i : 1 \le i \le n\}$, which are publicly available in our model. In fact, to facilitate the coordination scheme, the market intermediary needs to have access to the current queue lengths, and should be able to perfectly observe the idleness of suppliers.

## 4.6 Simulation results

To close the loop, we shall return to the original system described in §3.2 and see how the competitive equilibrium and coordination scheme fare. To this end, we again run simulations using the Arena model illustrated in Figure A1 of §3.6. The parameters are the same as those in §4.4.3: $c = 0.5$, $\Lambda = 1$, $\hat{\mu} = e^{-1.3}$, and the valuation $v$ is exponentially distributed with mean 0.1.

**Mixing distributions of prices**. Compared with §3.6, a new challenge arises: we cannot arbitrarily assign prices, because now the suppliers determine their competitive prices as equilibrium outcomes. We shall use the results from §4.4.3 as inputs to our Arena model for both homogeneous and heterogeneous cases of suppliers. We note that the equilibrium pricing strategy is described by a continuous distribution without simple expressions (see Propositions 3 and 5). Thus, the usual inverse-transform method fails to apply (because the inverse of distribution function is not known). Furthermore, the acceptance-rejection method is also not suitable for this problem, because it requires an explicit expression of density function that is not available. Our treatment follows from a similar idea to Figure 3. We first discretize them and record the cumulative distribution at discrete points. We choose the mesh sufficiently small and make linear interpolation to replicate approximately the original continuous distribution.

**Second-order revenues**. In terms of suppliers' profits, we focus exclusively on the pricing game in which the second-order correction around $R_i^r(t)$ defined as $r_i^r(t) \equiv \frac{1}{\sqrt{r}}(R_i^r(t) - r\bar{p}\mu_i t), \forall i \in \mathscr{N}$. Note that given $\hat{\mu} = e^{-1.3}$, $\bar{p} \approx 0.13$. When there are two suppliers ($n = 2$), we let $a \equiv \frac{\mu_1}{\hat{\mu}} \in (0.5, 1)$ denote the heterogeneity of service rates between these two suppliers. We examine two scenarios: in the homogeneous case $a = 0.5$, these two suppliers are endowed with the same service rate (capacity). In the heterogeneous case, we choose $a = 0.71$.

Regarding the tie-breaking rule, here we examine two rules: the smallest index first rule by which supplier 1 gets the priority, and the random priority rule by which customers choose between tied suppliers with equal probabilities. For the random priority rule, we can revise the conditions inside the ”Decide Suppliers” module in Figure A1. Specifically, we add a ”two-way by chance” module to rout the customers randomly with 50-50 chances when there is a tie. For all the following simulations, we conduct 1,000 replications, each of which takes the warm up of 120 arrivals and regular simulation of 600 arrivals in expectation. The scaling factor $r$ is fixed at 1,000. The confidence level is set at 5% when we make the statistical statements of hypothesis testing. We use two-sample-t two-tailed tests when we compare across different scenarios and paired-t tests when comparing between the two suppliers within each scenario.

**Revenue comparison: Symmetric case**. First, we consider two symmetric suppliers ($a = 0.5$) and compare the suppliers' second-order revenues in the competitive equilibrium and under the coordination scheme. We find that the average difference of suppliers 1's and 2's revenues in these two scenarios are statistically significant. For supplier 1, the estimated revenue difference is 0.758 whereas the 95% confidence interval is 0±0.00419. Similarly, supplier 2's estimated revenue difference 0.76 and 0.76 falls outside ±0.00421. Therefore, the coordination scheme indeed leads to higher expected revenues for both suppliers.

**Revenue comparison: Asymmetric case**. Second, we consider two asymmetric suppliers ($a = 0.71$). In this case, the coordination scheme again yields higher expected revenues for both suppliers that are statistically significant. The estimated revenue improvements for suppliers 1 and 2 are 0.992 and 0.351 respectively, and they fall outside the 95% confidence intervals ±0.0026 and ±0.00284. We can also compare the two suppliers' revenues. Naturally, their (second-order) revenues are different due to heterogeneous service rates. Using paired-t tests, we observe that the revenue differences are statistically significant in both the competitive and coordinated scenarios (-0.435 and 0.206 on average, and their corresponding confidence intervals are ±0.00295 and ±0.00488.

**Tie-breaking rule**. Third, we can also examine the impact of tie-breaking rule. For this matter, we use the symmetric supplier case as illustration. We first start with the competitive equilibrium and compare the two tie-breaking rules: smallest index first and random rules. For the competitive equilibrium we fail to reject the null hypothesis, i.e., the suppliers' revenues are statistically indistinguishable under the two rules. In contrast, under the coordination scheme the tie-breaking rule matters substantially. The estimated revenue difference is 0.0857 and it falls outside the 95% confidence interval ±0.00323.

The above discrepancy can be explained intuitively. In the competitive equilibrium, both suppliers randomize their prices. Thus, the chance of seeing an actual tie is infinitesimal (zero probability in theory). Therefore, the tie-breaking rule rarely comes in action. However, under the coordination scheme, both suppliers are induced to set prices at $\pi^C$. Because their service rates are identical, often times ties actually happen and the priority rule goes in favor of supplier 1. In this case, random tie-breaking ensures the fair routing between suppliers and it leads to statistically significant consequences. Further to the above observation, we run additional comparisons between the two suppliers. Under the smallest index first rule, supplier 1 earns on average 0.176 more than supplier 2, which is outside the 95% confidence interval ±0.00308. Under the random priority rule, this difference is negligible (-0.000792 on average).

To summarize, our simulations suggest that (1) the coordination scheme is effective in both homogeneous and heterogeneous scenarios, and this benefit applies to all suppliers; (2) tie-breaking rules are inconsequential when suppliers adopt randomized pricing, but they do matter when instead deterministic prices are chosen.

## 5 Conclusions

We study an oligopolistic model in which suppliers compete for buyers that are both price and delay sensitive. 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 suppliers' competition results in a price randomization over bounded 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 in intermediary platforms such as online exchanges.

## References

1. Afeche, P.: Incentive-compatible revenue management in queueing systems: optimal strategic delay. Manufacturing & Service Operations Management **15**(3), 423–443 (2013)
2. Allon, G., Bassamboo, A., Cil, E.B.: Large-scale service marketplaces: The role of the moderating firm. Management Science **58**(10), 1854–1872 (2012)

3. Allon, G., Federgruen, A.: Competition in service industries. Operations Research **55**(1), 37–55 (2007)
4. Allon, G., Gurvich, I.: Pricing and dimensioning competing large-scale service providers. Manufacturing & Service Operations Management **12**(3), 449–469 (2010)
5. Ata, B., Kumar, S.: 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)
6. Baye, M., Kovenock, D., Vries, C.D.: It takes two to tango: Equilibria in a model of sales. Games and Economic Behavior **4**, 493–510 (1992)
7. Besbes, O.: Revenue maximization for a queue that announces real-time delay information. Working paper, Graduate School of Business, Columbia University (2006)
8. Bramson, M.: State space collapse with applications to heavy-traffic limits for multiclass queueing networks. Queueing Systems **30**, 89–148 (1998)
9. Browne, S., Whitt, W.: Piecewise-linear diffusion processes. In: Advances in Queueing, pp. 463–480. CRC Press (2003)
10. Chen, Y.J., Maglaras, C., Vulcano, G.: Design of an aggregated marketplace under congestion effects: Asymptotic analysis and equilibrium characterization. Technical Report, Columbia University (2008)
11. DiPalantino, D., Johari, R., Weintraub, G.Y.: Competition and contracting in service industries. Operations Research Letters **39**(5), 390–396 (2011)
12. Ethier, S., Kurtz, T.: Markov processes: Characterization and convergence. Wiley Series in Probability and Mathematical Statistics, New York (1986)
13. Gallego, G., van Ryzin, G.: Optimal dynamic pricing of inventories with stochastic demand over finite horizons. Management Science **40**, 999–1020 (1994)
14. Glynn, P.: Diffusion approximations. In: D.P. Heyman, M.J. Sobel (eds.) Handbooks in OR & MS, vol. 2. North-Holland (1990)
15. Lederer, P., Li, L.: Pricing, production, scheduling and delivery -time competition. Operations Research **45**, 407–420 (1997)
16. Levhari, D., Luski, I.: Duopoly prcing and waiting lines. European Economic Review **11**, 17–35 (1978)
17. Loch, C.: Pricing in markets sensitive to delay. Ph.D. dissertation, Stanford University, Stanford, CA (1991)
18. Luski, I.: On partial equilibrium in a queueing system with two servers. The Review of Economic Studies **43**, 519–525 (1976)
19. Maglaras, C., Zeevi, A.: Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. Management Science **49**, 1018–1038 (2003)
20. Mandelbaum, A., Pats, G.: State-dependent queues: approximations and applications. In: F. Kelly, R. Williams (eds.) Stochastic Networks, *Proceedings of the IMA*, vol. 71, pp. 239–282. North-Holland (1995)
21. Mendelson, H.: Pricing computer services: Queueing effects. Communications of ACM **28**, 312–321 (1985)
22. Mendelson, H., Whang, S.: Optimal incentive-compatible priority pricing for the M/M/1 queue. Operations Research **38**, 870–883 (1990)
23. Naor, P.: On the regulation of queue size by levying tolls. Econometrica **37**, 15–24 (1969)
24. Stolyar, A.L.: Optimal routing in output-queued flexible server systems. Probability in the Engineering and Informational Science **19**, 141–189 (2005)
25. Watts, A.: On the uniqueness of equilibrium in Cournot oligopoly and other games. Games and Economic Behavior **13**(2), 269–285 (1996)
26. Williams, R.J.: An invariance principle for semimartingale reflecting Brownian motions in an orthant. Queueing Systems **30**, 5–25 (1998)
27. Zeltyn, S., Mandelbaum, A.: Internet supplement to “Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue”. Queueing Systems **51**, 361–402 (2005)

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

APPENDIX

Ying-Ju Chen* Costis Maglaras † Gustavo Vulcano‡

**Notation.** We use the following notation throughout the paper. We say that $f(x) = o(g(x))$ when $f(x)/g(x) \to 0$ when $x \to \infty$, and that $f(x) = g(x)$ to state 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 $y \in R^k$, we define $|y| := \max\{|y_j|, j = 1,\dots,k\}$ as its maximum norm. For any function $f : R_+ \to R^k$ and constant $L > 0$, we define $||f(\cdot)||_L := \sup_{0\le t\le L}|f(t)|$. For any sequence $\{a_r, r \in N\}$, $a_r \to \alpha$ if for any $\delta > 0$, there exists $r_\delta$ such that $|\alpha - a_r| < \delta$, whenever $r > r_\delta$. The function $\phi(\cdot)$ denotes the standard normal density, and $\Phi(\cdot)$ its corresponding cumulative distribution function (c.d.f.).

## A1 Schematic descriptions of the simulation model

We run simulations using Arena and the model is illustrated in Figure A1, where we assign the (randomly generated) valuation as an ”attribute” to each customer. In the module ”Decide Suppliers”, we compare the customer’s valuation and the full price (i.e., the price plus the expected waiting time) of each supplier. This determines whether a customer should balk (when the valuation is too low), and when not balking, which supplier to go to. Regarding the tie-breaking rule, we use the smallest index first rule: when a customer feels indifferent between suppliers 1 and 2, she joins supplier 1 by default. Other modules are self-explanatory. Arena software allows us to record the frequencies our specified variable (such as system workload) falls into distinct intervals. Thus, we can record the proportion of time the system workload falls below the boundary and above.

* School of Business and Management & School of Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; tel: +852-2358-7758; fax: +852-2358-2421; email: imchen@ust.hk

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

‡ School of Business at Universidad Torcuato di Tella, Buenos Aires, Argentina, gvulcano@utdt.edu

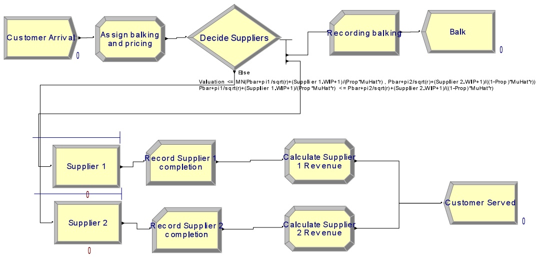
Customer Arrival
0
Assign balking and pricing
Decide Suppliers
Recording balking
Balk
0
Else
Valuation <= MN(Pbar+pi1/sqrt(r)+(Supplier 1.WIP+1)/(Prop*MuHat*r) , Pbar+pi2/sqrt(r)+(Supplier 2.WIP+1)/((1-Prop)*MuHat*r))
Pbar+pi1/sqrt(r)+(Supplier 1.WIP+1)/(Prop*MuHat*r) <= Pbar+pi2/sqrt(r)+(Supplier 2.WIP+1)/((1-Prop)*MuHat*r)
Supplier 1
0
Supplier 2
0
Record Supplier 1 completion
Record Supplier 2 completion
Calculate Supplier 1 Revenue
Calculate Supplier 2 Revenue
Customer Served
0

**Fig. A1** The simulation model in Arena.

## A2 Proofs of main results

### *Proof of Proposition 1*

Define $\kappa_i(t) = \pi_i + c\dfrac{\bar{Q}_i(t)}{\mu_i}, \forall i \in \mathscr{N}$ as the total cost supplier $i$ submits for the buyer at epoch $t$, and $\kappa(t) = [\kappa_1(t), ..., \kappa_n(t)]$. Let $\bar{J}(t) \subseteq \mathscr{N}$ be the set of suppliers tied as the cheapest at time $t$ and $\bar{J}^c(t) \equiv \mathscr{N} \setminus \bar{J}(t)$ be its complement. In this continuous fluid limit model, we can imagine that buyers arrive in a continuous manner and are routed to the cheapest supplier(s) accordingly. Given that this routing process is continuous, oscillations, that would be a core feature of the discrete buyer setting, will never occur. Critically, the buyers in our fluid limit model are not counted one by one; rather, they are infinitesimally small.

We hereby formalize the above statement. Define the function

$$g(\kappa(t)) = \max_{i,j \in \mathscr{N}} (\kappa_i(t) - \kappa_j(t)) = \max_{i \in \mathscr{N}} \kappa_i(t) - \min_{j \in \mathscr{N}} \kappa_j(t). \tag{A1}$$

Note that $g(\kappa(t))$ is nonnegative, and $g(\kappa(t)) = 0$ if and only if $\kappa_i(t) = \kappa_j(t)$, $\forall i, j \in \mathscr{N}$. 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.

Note that as long as not all suppliers receive new orders, the aggregate arrival rate, that equals $\sum_{i \in \mathscr{N}} \mu_i$, is always higher than the aggregate service rate of servers that tie as the cheapest ones. That is, $\sum_{i \in \mathscr{N}} \mu_i > \sum_{j \in \bar{J}(t)} \mu_j$ for all $t$ before the resource pooling regime (if it exists). Let $k \in \bar{J}(t)$ be an arbitrary cheapest supplier. We now claim that for any time epoch, the following equality holds:

$$\dot{\kappa}_k(t) = \frac{c}{\mu_k} \frac{\mu_k}{\sum_{i \in \bar{J}(t)} \mu_i} \left[ \sum_{i \in \mathscr{N}} \mu_i - \sum_{i \in \bar{J}(t)} \mu_j \right], \tag{A2}$$

where $\sum_{i\in\mathscr{N}}\mu_i - \sum_{i\in\bar{J}(t)}\mu_j$ is the imbalance between the aggregate arrival rate and aggregate service rate for those cheapest suppliers, and $\frac{\mu_k}{\sum_{i\in\bar{J}(t)}\mu_i}$ is the proportion of customers routed to supplier $k$; the term $\frac{c}{\mu_k}$ simply accounts for the sensitivity of $\kappa_k(t)$ on the queue length increase.
We prove Equation (A2) by contradiction. Suppose, on the contrary, that (A2) is violated at some time for some supplier in the cheapest suppliers' set. Let

$$t_1 := \inf\{t|t, \dot{\kappa}_k(t) \neq \frac{c}{\mu_k}\frac{\mu_k}{\sum_{i\in\bar{J}(t)}\mu_i}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{i\in\bar{J}(t)}\mu_j\right],\ for\ some\ k \in \bar{J}(t)\} \tag{A3}$$

be the time right after which violation occurs and (with abuse of notation) let $k$ be such a supplier. There are two cases for this violation: Case 1) $\dot{\kappa}_k(t_1) < \frac{c}{\mu_k}\frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$ and Case 2) $\dot{\kappa}_k(t_1) > \frac{c}{\mu_k}\frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$.
Consider the first case. Recall $\bar{J}(t_1)$ is the set of suppliers tied as the cheapest at time $t_1$ and define $\varepsilon_1 := \min_{i\in\bar{J}^c(t_1)}\kappa_i(t_1) - \min_{j\in\bar{J}(t_1)}\kappa_j(t_1)$ as the mimimum difference between those $\{\kappa_j(t_1)\}$'s inside and outside $\bar{J}(t_1)$ at time $t_1$. If $\varepsilon_1 = 0$, the problem is non-existent and we have already reached the resource pooling regime. Thus, below we consider the case $\varepsilon_1 > 0$.
Because $\min_{i\in\bar{J}^c(t_1)}\kappa_i(t_1) - \min_{j\in\bar{J}(t_1)}\kappa_j(t_1) > 0$, the inequality $\dot{\kappa}_k(t_1) < \frac{c}{\mu_k}\frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$ is strict, and $\{\kappa_k(t)\}$'s are all continuous functions, we can find a sufficiently small $\delta < \frac{\varepsilon_1}{c(1+\sum_{i\in\mathscr{N}}\mu_i/\min_{j\in\mathscr{N}}\mu_j)}$ such that 1) the sets $\bar{J}(t) \subseteq \bar{J}(t_1)$, for all $t \in [t_1, t_1+\delta)$, and 2) $\dot{\kappa}_k(t) < \frac{c}{\mu_k}\frac{\mu_k}{\sum_{i\in\bar{J}(t)}\mu_i}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{i\in\bar{J}(t)}\mu_j\right]$, for all $k \in \bar{J}(t)$, for all $t \in [t_1, t_1+\delta)$.
Consider any $i_1 \in \bar{J}^c(t_1)$ and $j_1 \in \bar{J}(t_1)$. For $t \in [t_1, t_1+\delta)$, we note that

$$\begin{aligned}
&\kappa_{i_1}(t) - \kappa_{j_1}(t)\\
&= \pi_{i_1} + c\frac{\bar{Q}_{i_1}(t)}{\mu_{i_1}} - \left\{\pi_{j_1} + c\frac{\bar{Q}_{j_1}(t)}{\mu_{j_1}}\right\}\\
&\geq \pi_{i_1} + c\frac{\bar{Q}_{i_1}(t) - \delta\mu_{i_1}}{\mu_{i_1}} - \left\{\pi_{j_1} + c\frac{\bar{Q}_{j_1}(t) + \delta\sum_{i\in\mathscr{N}}\mu_i}{\mu_{j_1}}\right\}\\
&\geq \pi_{i_1} + c\frac{\bar{Q}_{i_1}(t)}{\mu_{i_1}} - \left\{\pi_{j_1} + c\frac{\bar{Q}_{j_1}(t)}{\mu_{j_1}}\right\} - c\delta\left\{1 + \frac{\sum_{i\in\mathscr{N}}\mu_i}{\mu_{j_1}}\right\}\\
&\geq \varepsilon_1 - c\frac{\varepsilon_1}{c(1+\sum_{i\in\mathscr{N}}\mu_i/\min_{j\in\mathscr{N}}\mu_j)}\left\{1 + \frac{\sum_{i\in\mathscr{N}}\mu_i}{\mu_{j_1}}\right\}\\
&> 0.
\end{aligned}$$

Therefore, within the time window $[t_1, t_1+\delta)$, no supplier in $\bar{J}^c(t_1)$ will join the set of cheapest suppliers. This suggests that all the arrivals will be routed to $\bar{J}(t_1)$ within the time window $[t_1, t_1+\delta)$ and $\bar{J}(t) \subseteq \bar{J}(t_1)$ for $t \in [t_1, t_1+\delta)$.

Now we focus on the set $\bar{J}(t_1)$. Because $\dot{\kappa}_k(t_1) < \frac{c}{\mu_k}\frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{\mathscr{J}}(t_1)}\mu_j\right]$, there must exist another supplier $l$ in $\bar{J}(t_1)$ such that $\dot{\kappa}_l(t) > \frac{c}{\mu_l}\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$. To see this, recall that the aggregate arrival rate is $\sum_{i\in\mathscr{N}}\mu_i$ during $[t_1,t_1+\delta)$, which is always higher than $\sum_{j\in\bar{J}(t_1)}\mu_j$, the aggregate service rate of servers that tie as the cheapest ones in $\bar{J}(t)$. Since no supplier in $\bar{J}^c(t_1)$ receives any arrival during $[t_1,t_1+\delta)$, the queue length accumulation rate is at least $\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j$ (if $\bar{J}(t)$ is strictly smaller than $\bar{J}(t_1)$, then the queue length accumulation rate can only go higher). If for all $l\in\bar{J}(t_1)\setminus\{k\}$, $\dot{\kappa}_l(t) \le \frac{c}{\mu_l}\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$, then the queue length accumulation of each $l$ is

$$\bar{Q}_l(t) \le \frac{\mu_l}{c}\left\{\frac{c}{\mu_l}\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]\right\} = \frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right].$$

Likewise, for supplier $k$ we have $\bar{Q}_k(t) < \frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$. Thus, the aggregate queue length accumulation among all suppliers is

$$\begin{aligned}
&\bar{Q}_k(t) + \sum_{l\in\bar{J}(t_1)\setminus\{k\}}\bar{Q}_l(t) + \sum_{i\in\bar{J}^c(t_1)}\bar{Q}_i(t)\\
&< \frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right] + \sum_{l\in\bar{J}(t_1)\setminus\{k\}}\left\{\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]\right\} + 0\\
&< \sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j,
\end{aligned}$$

which leads to a contradiction.

Given that $\dot{\kappa}_l(t) > \frac{c}{\mu_l}\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$, we derive the corresponding $\kappa_l(t)$ as follows ($l\in\bar{J}(t_1)\setminus\{k\}$):

$$\begin{aligned}
\kappa_l(t) &> \kappa_l(t_1) + \frac{c}{\mu_l}\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right](t-t_1)\\
&= \pi_l + c\frac{\bar{Q}_l(t_1)}{\mu_l} + \frac{c}{\mu_l}\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right](t-t_1)\\
&= \pi_k + c\frac{\bar{Q}_k(t_1)}{\mu_k} + \frac{c}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right](t-t_1),
\end{aligned}$$

where the last equality comes from the fact that both suppliers $k$ and $l$ belong to $\bar{J}(t_1)$. On the other hand, for supplier $k$ we have:

$$\begin{aligned}\kappa_k(t) &< \kappa_k(t_1) + \frac{c}{\mu_k}\frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right](t-t_1)\\ &= \pi_k + c\frac{\bar{Q}_k(t_1)}{\mu_k} + \frac{c}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right](t-t_1).\end{aligned}$$

Thus, $\kappa_l(t) > \kappa_k(t)$, for all $t \in [t_1,t_1+\delta)$. However, this implies that supplier $l$ shall never receive any arrival during $[t_1,t_1+\delta)$, and its queue length can only drop during $[t_1,t_1+\delta)$. This leads to a contradiction.

The second case, $\dot{\kappa}_k(t) > \frac{c}{\mu_k}\frac{\mu_k}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$, is the mirror image of the above. In this case, we must be able to find another supplier $l \in \bar{J}(t_1)$ such that $\dot{\kappa}_l(t) < \frac{c}{\mu_l}\frac{\mu_l}{\sum_{j\in\bar{J}(t_1)}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{j\in\bar{J}(t_1)}\mu_j\right]$. Otherwise, the queue length accumulations will be violated within the set $\bar{J}(t_1)$. However, we can then observe that $\kappa_k(t) > \kappa_l(t)$ in a small time window and supplier $k$ shall never receive any arrival. This implies that $\dot{\kappa}_k$ in that small time window is strictly negative and leads to a contradiction. Collectively, Equation (A2) must hold.

Next, we return to the proposed Lyapunov function $g(\kappa(t))$. We shall prove that there exists a lower bound for the depreciation rate of $g(\kappa(t))$. We observe that $\frac{d}{dt}\left(\max_{i\in\mathscr{N}}\kappa_i(t)\right)$ is non-positive, because those suppliers in $\bar{J}^c(t)$ do not receive arrivals and their queue lengths can only depreciate. On the other hand, Equation (A2) shows that $\frac{d}{dt}\left(\min_{j\in\mathscr{N}}\kappa_j(t)\right) \geq 0$, because $\sum_{i\in\mathscr{N}}\mu_i - \sum_{i\in\bar{J}(t)}\mu_j > 0$ and therefore the derivative is always non-negative. Thus, $\dot{g}(\kappa(t)) = \frac{d}{dt}\left(\max_{i\in\mathscr{N}}\kappa_i(t)\right) - \frac{d}{dt}\left(\min_{j\in\mathscr{N}}\kappa_j(t)\right)$ is strictly negative and $|\dot{g}(\kappa(t))| > |\dot{\kappa}_k(t)| = \dot{\kappa}_k(t)$, where $k \in \bar{J}(t)$. We have

$$|\dot{g}(\kappa(t))| \geq \dot{\kappa}_k(t) = \frac{c}{\mu_k}\frac{\mu_k}{\sum_{i\in\bar{J}(t)}\mu_i}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{i\in\bar{J}(t)}\mu_j\right] \geq \frac{c}{\sum_{i\in\mathscr{N}}\mu_i - \min_{j\in\mathscr{N}}\mu_j}\left[\sum_{i\in\mathscr{N}}\mu_i - \sum_{i\in\bar{J}(t)}\mu_j\right]. \tag{A4}$$

From $\sum_{i\in\mathscr{N}}\mu_i - \sum_{i\in\bar{J}(t)}\mu_j \geq \min_{j\in\mathscr{N}}\mu_j$ as long as $\bar{J}(t) \neq \mathscr{N}$, we then have $|\dot{g}(\kappa(t))| \geq \dfrac{c\min_{j\in\mathscr{N}}\mu_j}{\sum_{i\in\mathscr{N}}\mu_i - \min_{j\in\mathscr{N}}\mu_j}$. 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. To wit, our argument to establish Equation (A2) does not hinge on any specific tie-breaking rule. Thus, the above result applies to all tie-breaking rules. When $\bar{J}(t) = \mathscr{N}$, the system has reached the resource pooling regime.

Let $\triangle\kappa$ be such that $\pi_l + c\dfrac{\bar{Q}_l(0)}{\mu_l} + \triangle\kappa = \max_{i\in\mathscr{N}}\ \{\pi_i + c\dfrac{\bar{Q}_i(0)}{\mu_i}\}$. 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 $s^*(\bar{Q}(0),\bar{W}(0)) =$

$\frac{\triangle\kappa(\sum_{i\in\mathscr{N}}\mu_i-\min_{j\in\mathscr{N}}\mu_j)}{c\min_{j\in\mathscr{N}}\mu_j}$. 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 $\delta>0$, $s^*(\delta,M_0)\leq s^*(M_0)=\frac{\triangle(M_0)(\sum_{i\in\mathscr{N}}\mu_i-\min_{j\in\mathscr{N}}\mu_j)}{c\min_{j\in\mathscr{N}}\mu_j}$. This completes the proof for the convergence of fluid model. □

## *Proof of Proposition 2*

The proof builds on the framework advanced by [8]. To facilitate exposition we will follow [8] 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 A1-A5 will establish that appropriately scaled processes that focus on the system behavior over order $1/\sqrt{r}$ time intervals satisfy the fluid equations (11)-(16). 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. [8, 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 $\dfrac{1}{\sqrt{r}}t+\dfrac{1}{\sqrt{r}}m$. 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 [8, Proposition 5.1]. For ease of notation, let us define $Q^{r,m}(t) = \{Q_i^{r,m}(t)\}$. The routing indicator is $\mathbf{I}_i(Q^{r,m}(t)) = 1$ if $\Xi_{\bar{J}(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)} = i$, and 0 otherwise. Note that $\sqrt{r}Q_i^{r,m}(t) = Q_i^r(\dfrac{1}{\sqrt{r}}t+\dfrac{1}{\sqrt{r}}m)$ is the queue length of supplier $i$ at epoch $\dfrac{1}{\sqrt{r}}t+\dfrac{1}{\sqrt{r}}m$. Therefore,

$$\bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} = \bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{Q_i^r(\frac{1}{\sqrt{r}}t+\frac{1}{\sqrt{r}}m)+1}{\mu_i^r} \tag{A5}$$

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

$$\mathbf{J}_i^r(Q_i^{r,m}(t)) = P\left\{v \geq \bar{p}+\frac{\pi_i}{\sqrt{r}}+c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i}\right\}, \forall i \in \mathscr{N} \tag{A6}$$

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

**Lemma A1** *Fix* $\varepsilon > 0$, $L > 0$, *and* $\tau > 0$. *For r large enough,*

$$P\left\{\max_{m<\sqrt{r}\tau}\max_{i\in\mathscr{N}}||A_i^{r,m}(t)-\int_0^t \Lambda^r\mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L > \varepsilon\right\} \leq \varepsilon, \tag{A7}$$

$$P\left\{\max_{m<\sqrt{r}\tau}\max_{i\in\mathscr{N}}||D_i^{r,m}(t)-\mu_i T_i^{r,m}(t)||_L > \varepsilon\right\} \leq \varepsilon, \tag{A8}$$

$$P\left\{\max_{m<\sqrt{r}\tau}\max_{i,j\in\mathscr{N}}||\left(\pi_i+c\frac{Q_i^{r,m}(t)}{\mu_i}\right)-\left(\pi_j+c\frac{Q_j^{r,m}(t)}{\mu_j}\right)||_L > \varepsilon\right\} \leq \varepsilon. \tag{A9}$$

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

**Lemma A2** *Fix* $\varepsilon > 0$, $L > 0$, *and* $\tau > 0$. *There exists a constant* $N_0$ *such that for r large enough,*

$$P\left\{\max_{m<\sqrt{r}\tau}\sup_{t_1,t_2\in[0,L]}|X^{r,m}(t_2)-X^{r,m}(t_1)| > N_0|t_2-t_1|+\varepsilon\right\} \leq \varepsilon. \tag{A10}$$

The above two lemmas imply that the measure of "ill-behaved" events is negligible for the hydrodynamic scaled processes. Note that (A9) only shows the convergence in probability for one particular instance. Nevertheless, the ultimate state space collapse requires an aggregation/ integration over all instances, and continuous integrations over measure-zero events could lead to non-trivial consequences.
Let $N_0$ denote the constant required in Lemma A2 and focus on the complement of these ill-behaved events. We can choose a sequence $\varepsilon(r)$ decreasing to 0 sufficiently slowly such that the inequalities (A9), (A8), and (A10) still hold. For such a sequence $\varepsilon(r)$, we define

$$K_0^r = \left\{ \max_{m<\sqrt{r}\tau} \sup_{t_1,t_2\in[0,L]} |X^{r,m}(t_2) - X^{r,m}(t_1)| \le N_0|t_2 - t_1| + \varepsilon(r) \right\},$$

$$K_1^r = \left\{ \begin{array}{c} \max_{m<\sqrt{r}\tau} \max_{i\in\mathscr{N}} ||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L \le \varepsilon(r); \\ \max_{m<\sqrt{r}\tau} \max_{i\in\mathscr{N}} ||D_i^{r,m}(t) - \mu_i T_i^{r,m}(t)||_L \le \varepsilon(r) \\ \max_{m<\sqrt{r}\tau} \max_{i,j\in\mathscr{N}} || \left(\pi_i + c\frac{Q_i^{r,m}(t)}{\mu_i}\right) - \left(\pi_j + c\frac{Q_j^{r,m}(t)}{\mu_j}\right) ||_L \le \varepsilon(r); \end{array} \right\},$$

$$K^r = K_0^r \cap K_1^r.$$

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 [8, Proposition 5.2, Corollary 5.1]. and follows from Lemmas A1 and A2.

**Lemma A3** $P(K^r) \to 1$, *as* $r \to \infty$.

Let $E$ be the space of functions $x : [0,L] \to R^5$ that are right continuous and have left limits. Define $E^{'} = \{x \in E : |x(0)| < M_0, |x(t_2) - x(t_1)| < N_0|t_2 - t_1|, \forall t_1, t_2 \in [0,L]\}$, where $M_0$ is a fixed constant. Let $E_0^r = \{X^{r,m}(\cdot,\omega), m < \sqrt{r}\tau, \omega \in K^r\}$ denote the sets of well-behaved events and $E_0 = \{E_0^r, r \in R_+\}$ 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 $E_0$. A function $\widehat{X}$ is said to be a cluster point of the functional space $E_0$ if for all $\delta > 0$, there exists a $X_\delta \in E_0$ such that $||\widehat{X} - X_\delta||_L < \delta$.

**Lemma A4** *Fix* $\varepsilon > 0, L > 0$, *and* $\tau > 0$. *There exists a sufficiently large* $r(\varepsilon)$ *such that for all* $r > r(\varepsilon)$, *for all* $\omega \in K^r$ *and for all* $m = 1,2,...,\lfloor\sqrt{r}\tau\rfloor$, $||X^{r,m}(\cdot,\omega) - \widehat{X}(\cdot)||_L < \varepsilon$, *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 [8, 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 (11)-(16) (cf. [8, Proposition 6.2]).

**Lemma A5** *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 (11)-(15).*

*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 $\eta, \xi, \varepsilon > 0$. By Lemma A3, there exists a sufficiently large $r(\xi) > 0$ such that $P(K^r) > 1 - \xi$ for all $r > r(\xi)$.
Take $L = s(\varepsilon, M_0) + 1$ 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 $m_r(\varsigma) = \min\{m \in N_+ : m < \sqrt{r}\varsigma < m + L\} = \max\{\lceil \sqrt{r}\varsigma - L \rceil, 0\}$ 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

$$\tau_r^{'}(\varsigma) = \sqrt{r}\varsigma - m_r(\varsigma) = \sqrt{r}\varsigma - \max\{\lceil \sqrt{r}\varsigma - L \rceil, 0\} \geq \sqrt{r}\varsigma - (\sqrt{r}\varsigma - L - 1) = L - 1 = s(\varepsilon, M_0). \tag{A11}$$

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 $\pi_j + c\frac{\tilde{Q}_j^r(\varsigma)}{\mu_j}$ is upper bounded (note that the additional terms $\{1/\mu_i^r\}$'s in the suppliers' bids have been taken into account in Lemma A1 through (A9)):

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

From Lemmas A1 and A4, for fixed $L, \tau, \xi,\ \eta$, and $\varepsilon > 0$, there exists a sufficiently large $r(L, \tau, \xi,\ \eta, \varepsilon) > \max\{r(\xi), r(\eta)\}$ such that for all $r > r(L, \tau, \xi,\ \eta, \varepsilon), \omega \in K^r$, for all $\varsigma \in [\frac{L}{\sqrt{r}}, \tau]$, we have

$$\max_{i \in \mathscr{N}} | \left( \pi_i + c \frac{Q_i^{r,m_r(\varsigma)}(\tau_r'(\varsigma))}{\mu_i} \right) - \left( \pi_i + c \frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i} \right) | \leq \frac{\varepsilon}{3}, \tag{A12}$$

and $\max_{i,j \in \mathscr{N}} | \left( \pi_i + c \frac{\widehat{Q}_i(\tau_r'(\varsigma))}{\mu_i} \right) - \left( \pi_j + c \frac{\widehat{Q}_j(\tau_r'(\varsigma))}{\mu_j} \right) | \leq \frac{\varepsilon}{3}$. This implies that $\max_{i,j \in \mathscr{N}} | \left( \pi_i + c \frac{\tilde{Q}_i^r(\varsigma)}{\mu_i} \right) - \left( \pi_j + c \frac{\tilde{Q}_j^r(\varsigma)}{\mu_j} \right) | \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$. Note that $P(K^r) \to 1$ as $r \to \infty$, i.e., the state space collapse holds in probability (the additional term $1/\mu_i^r$ is of order $1/r$ and therefore does not affect the result as demonstrated in Lemma A1). We can then choose appropriate $L, \tau, \xi,\ \eta$, and $\varepsilon$ 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 $W^r(t) = \sum_{i \in \mathscr{N}} \dfrac{Q_i^r(t)}{\mu_i^r}$, 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 ([14, 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 $A^r(t)$. Note that potential buyers that find it costly to purchase and choose not to purchase are automatically left out from the effective arrival process.

*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.,$\forall i \in \mathscr{N}$, and $\dfrac{\Lambda^r(t)}{r} \ \to (\sum_{i \in \mathscr{N}} \mu_i)t,\ u.o.c.$

*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 $(\varphi, \psi)$ (similar to the framework in [26]) and apply Lemma 7 in [5] 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 [20] 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 $\pi_i + c\dfrac{\tilde{Q}_i^r(t)}{\mu_i} = \bar{\pi} + \bar{c}\tilde{W}^r(t) + o_p(1),\ \forall i \in \mathscr{N}$. 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,

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

Rearranging the above equation, we see that

$$\frac{A_i^r(t)}{\sqrt{r}} = \frac{\mu_i}{c}[\bar{\pi} + \bar{c}\tilde{W}^r(t)] + \frac{S_i^r(t - Y_i^r(t))}{\sqrt{r}} - \frac{\mu_i \pi_i}{c} - \tilde{Q}_i^r(0) + o_p(1), \tag{A13}$$

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

$$\frac{A^r(t)}{\sqrt{r}} = \frac{\sum_{i\in\mathscr{N}} \mu_i}{c}[\bar{\pi} + \bar{c}\tilde{W}^r(t)] + \sum_{i\in\mathscr{N}} \frac{S_i^r(t - Y_i^r(t))}{\sqrt{r}} - \frac{\sum_{i\in\mathscr{N}} \mu_i \pi_i}{c} - \sum_{i\in\mathscr{N}} \tilde{Q}_i^r(0) + o_p(1). \tag{A14}$$

Next we focus on the demand and service time processes. A strong approximation ([14, Theorem 5]) for the Poisson process $N(\cdot)$ allows us to rewrite $A^r(t) = N(\Lambda^r(t))$ as

$$A^r(t) = \Lambda^r(t) + B_a(\Lambda^r(t)) + O(\log rt), \tag{A15}$$

where $\Lambda^r(t) = \int_0^t \lambda^r(\min_{i\in\mathscr{N}}[\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}\tilde{Q}_i^r(s)+1}{r\mu_i}])ds$ 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, according to [12], we can apply strong approximation to service completion process $t - Y_i^r(t)$ (which is a random time) and represent the renewal process $S_i^r(\cdot)$ as

$$S_i^r(t - Y_i^r(t)) = \mu_i^r t - \mu_i^r Y_i^r(t) + \sqrt{r}\sigma_i B_{s,i}(t - Y_i^r(t)) + O(\log r[t - Y_i^r(t)]), \tag{A16}$$

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.”[1] Note that $B_a(t)$ and $\{B_{s,i}(t)\}$'s describe distinct Brownian motion processes and they are mutually independent. The service rate is of order $r$, and therefore the Brownian motion is scaled by $\sqrt{r}$.

From Proposition 2 we have that $\min_{i\in\mathcal{N}}\left\{\pi_i + c\dfrac{\tilde{Q}_i^r(t)}{\mu_i} + c\dfrac{1}{\mu_i^r}\right\} = \bar{\pi} + \bar{c}\tilde{W}^r(t) + o_p(1)$, $\forall t$, and $P(\dfrac{\tilde{W}^r(t)}{\sqrt{r}} > 0) \to 0$, 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

$$\Lambda^r(t) = (\sum_{i\in\mathcal{N}}\mu_i^r)t - \sqrt{r}(\sum_{i\in\mathcal{N}}\mu_i)\frac{f(\bar{p})}{\bar{F}(\bar{p})}\int_0^t[\bar{\pi}+\bar{c}\tilde{W}^r(s)]ds + o_p(\sqrt{r}t). \qquad \text{(A18)}$$

Combining (A14), (A15), (A16), and (A18), and for $\gamma \equiv \dfrac{f(\bar{p})}{\bar{F}(\bar{p})}$, 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:

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

[1] More specifically, Corollary 5.5 of Chapter 7 in [12] ensures the existence of a probability space in which a Poisson process $N$ and a Brownian motion $B$ exist such that

$$\sup_{t\geq 0}\frac{|N(t)-t-B(t)|}{\log(2\vee t)} < \infty, \;\; a.s. \qquad \text{(A17)}$$

See also [20, §7.2, p. 268] for a similar treatment.

$$+\frac{B_a(\Lambda^r(t))}{\sqrt{r}} - \sum_{i\in\mathscr{N}} \sigma_i B_{s,i}(t - Y_i^r(t)) + o_p(1).$$

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

Using these substitutions, we get that

$$\frac{\hat{\mu}}{c}\tilde{Z}^r(t) = -\hat{\mu}\gamma\int_0^t \tilde{Z}^r(s)ds + \sum_{i\in\mathscr{N}} \mu_i \tilde{Y}_i^r(t) + \tilde{Z}(0) + [\frac{B_a(\Lambda^r(t))}{\sqrt{r}} - \sum_{i\in\mathscr{N}} \sigma_i B_{s,i}(t - Y_i^r(t))] + o_p(1). \tag{A19}$$

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

$$\tilde{Z}^r(t) = \tilde{Z}(0) - \gamma c\int_0^t \tilde{Z}^r(s)ds + \tilde{U}^r(t) + \frac{c}{\hat{\mu}}\left[\frac{B_a(\Lambda^r(t))}{\sqrt{r}} - \sum_{i\in\mathscr{N}} \sigma_i B_{s,i}(t - Y_i^r(t))\right] + o_p(1). \tag{A20}$$

*Step 2:*

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

$$\frac{\Lambda^r(t)}{r} \rightarrow \hat{\mu}t,\ inprobability,\ u.o.c. \tag{A21}$$

$$Y_i^r(t) \rightarrow 0, inprobability,\ u.o.c., \forall i \in \mathscr{N}. \tag{A22}$$

First we prove (A21). From (A18), we have $\dfrac{\Lambda^r(t)}{r} = \hat{\mu}t - \dfrac{\hat{\mu}\gamma}{\sqrt{r}}\displaystyle\int_0^t \tilde{Z}^r(s)ds + o_p(\frac{1}{\sqrt{r}})$, where the last term vanishes in the limit. From Proposition 2, we know that for any fixed constant $C$, $P(\sup_{t\leq T} \tilde{W}^r(t) > C) \rightarrow 0$, and hence $P(\sup_{t\leq T} \dfrac{\tilde{W}^r(t)}{\sqrt{r}} > \varepsilon) \rightarrow 0$, $\forall\varepsilon > 0$. Moreover, $\dfrac{\hat{\mu}\gamma}{\sqrt{r}}\displaystyle\int_0^t \tilde{Z}^r(s)ds = \dfrac{\hat{\mu}\gamma}{\sqrt{r}}[\bar{\pi}t + c\displaystyle\int_0^t \tilde{W}^r(s)ds] \rightarrow 0$, in probability, u.o.c., as $r \rightarrow \infty$, because the first term inside the brackets is a constant and the integral is uniformly bounded by $\varepsilon\sqrt{r}$. Therefore, $\dfrac{\Lambda^r(t)}{r} \rightarrow \hat{\mu}t,\ inprobability,\ u.o.c.$, which completes the proof of (A21).

Next we show (A22) by contradiction. Suppose that there exists a supplier $i$ such that $Y_i^r(t)0$ u.o.c., then we can find a constant $\varepsilon_1 > 0$ and a sequence $(r_k, t_k)$ such that $r_k \rightarrow \infty$ as $k \rightarrow \infty$ and $Y_i^{r_k}(t_k) > \varepsilon_1$, $\forall k$. 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 $\tilde{U}^{r_k}(t_k) > \sqrt{r_k}(c\mu_i\varepsilon_1/\hat{\mu})$. Thus, as $k \rightarrow \infty$, $\tilde{U}^{r_k}(t_k) \rightarrow \infty$. From (A20), this implies that $\tilde{Z}^{r_k}(t_k) \rightarrow \infty$, and therefore that $\tilde{W}^{r_k}(t_k) \rightarrow \infty$. This, however, contradicts the fact that for all $T > 0$, $\varepsilon > 0$, there exists a constant $C$ such that $P(\sup_{t\leq T}\tilde{W^r}(t) > C) < \varepsilon$. This completes the proof of (A22).

*Step 3:*

We can rewrite (A20) 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 $\frac{\Lambda^r(t)}{r} \to \hat{\mu}t$ a.e. and $Y_i^r(t) \to 0$ in probability, we can apply the invariance principle of Brownian motion and convergence together lemma to obtain

$$\frac{B_a(\Lambda^r(t))}{\sqrt{r}} D{=} B_a(\frac{\Lambda^r(t)}{r}) \Rightarrow B_a(\hat{\mu}t) D{=} \sqrt{\hat{\mu}} B_a(t), \sum_{i\in\mathcal{N}} \sigma_i B_{s,i}(t - Y_i^r(t)) \Rightarrow \sum_{i\in\mathcal{N}} \sigma_i B_{s,i}(t) D{=} \sigma B_s(t), \tag{A23}$$

where $\sigma \equiv \sqrt{\sum_{i\in\mathcal{N}} \sigma_i^2}$ and $B_a, B_s$ are independent standard Brownian motions. Thus, $\tilde{V}^r(t) \Rightarrow \frac{c}{\hat{\mu}}[\sqrt{\hat{\mu}}B_a(t) - \sigma B_s(t)] D{=} \frac{c}{\hat{\mu}}\sqrt{\sigma^2 + \hat{\mu}}B(t)$, 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 $\tilde{U}^r(t) \geq 0$.

Now we quote a technical lemma:

**Lemma A6** *Let $\varepsilon^r$ be a real-valued sequence such that $\varepsilon^r \to 0$ as $r \to \infty$, and $\zeta \equiv \frac{1}{\bar{c}} \max_i (\pi_i - \bar{\pi})$. Then, $\int_0^t 1\!\!1\{|\tilde{W}^r(s) - \zeta| > \varepsilon^r\} d\tilde{U}^r(s) \to 0$ in probability, u.o.c.*

Motivated by the result in Lemma A6, we will rewrite $\tilde{U}^r(t)$ in two parts as follows:

$$\tilde{U}^r(t) = \underbrace{\int_0^t 1\!\!1\{|\tilde{W}^r(s) - \zeta| > \varepsilon^r\} d\tilde{U}^r(s)}_{\tilde{U}^r_\varepsilon(t)} + \underbrace{\int_0^t 1\!\!1\{|\tilde{W}^r(s) - \zeta| \leq \varepsilon^r\} d\tilde{U}^r(s)}_{\tilde{U}^r_\zeta(t)}. \tag{A24}$$

Note that $\tilde{U}^r_\zeta(\cdot), \tilde{U}^r_\varepsilon(\cdot)$ are nonnegative, continuous, and nondecreasing. Moreover, $\tilde{U}^r_\varepsilon(t) \to 0$, 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 $V(t) = \frac{c}{\hat{\mu}}\sqrt{\sigma^2 + \hat{\mu}}B(t)$, 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 $\zeta$, 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,

$$\tilde{X}(t) = \tilde{X}(0) - \gamma c \int_0^t [\tilde{X}(s) + \hat{\pi}]ds + V(t) + U(t) \equiv \varphi(\hat{X}(t)), \tag{A25}$$

where $\varphi(\cdot)$ is the reflection operator ([20]) 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 $\tilde{Z}^r(t) = \tilde{Z}^r(0) - \gamma c \int_0^t \tilde{Z}^r(s)ds + \tilde{V}^r(t) + \tilde{U}^r_\zeta(t) + \tilde{U}^r_\varepsilon(t), \tilde{Z}^r(t) = \tilde{Z}^r(0) - \gamma c \int_0^t \tilde{Z}^r(s)ds + \tilde{V}^r(t) + \tilde{U}^r_\zeta(t) + \tilde{U}^r_\varepsilon(t)$, 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 $\tilde{H}^r(s) \Rightarrow \tilde{X}(t)$ in distribution, which will imply the weak convergence of $\tilde{Z}^r(t)$ to $\tilde{R}(t)$. From Lemma A6 and Proposition 2, for any sequence $\varepsilon^r > 0$, s.t. $\varepsilon^r \to 0$ 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 $P(\inf_{s \leq t} \tilde{W}^r(s) \geq \zeta - \varepsilon^r) = 1 - \delta^r$. Let $\Omega(\varepsilon^r)$ be the set of sample paths for which $\inf_{s \leq t} \tilde{\tilde{W}}^r(s) \geq \zeta - \varepsilon^r$. Note that $P(\Omega(\varepsilon^r)) = 1 - \delta^r \to 1$ as $r \to \infty$.
In the sequel we will use Lemma 7 of [5]. We first observe that $\forall \omega \in \Omega(\varepsilon^r), \tilde{W}^r(t) \geq \zeta - \varepsilon^r \Leftrightarrow \tilde{Z}^r(t) \geq \hat{\pi} - \bar{c}\varepsilon^r \Leftrightarrow \tilde{H}^r(t) \geq -\bar{c}\varepsilon^r$, and, similarly, $\tilde{W}^r(s) > \zeta + \varepsilon^r$ implies $\tilde{H}^r(s) > \bar{c}\varepsilon^r$. If we define $H^r_1(t) = \tilde{H}^r(t) + \bar{c}\varepsilon^r$, $H^r_2(t) = \hat{H}^r(t) + \bar{c}\varepsilon^r$, and focus on the event $\omega \in \Omega(\varepsilon^r)$, then $H^r_1, H^r_2, \tilde{U}^r$ satisfy the following conditions:

$$H^r_1(t) = H^r_2(t) + \tilde{U}^r(t), H^r_1(t) \geq 0, \tilde{U}^r(\cdot) \textit{ nondecreasing}; \; \tilde{U}^r(0) = 0, \int_0^t 1\!\!1\{H^r_1(s) > 2\bar{c}\varepsilon^r\} d\tilde{U}^r(s) = 0, \tag{A26}$$

where the last equation follows from a simple transformation $\tilde{H}^r(s) > \bar{c}\varepsilon^r \Leftrightarrow H^r_1(s) > 2\bar{c}\varepsilon^r$. 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 $H^r_1, H^r_2, \tilde{U}^r$ fit the conditions of Lemma 7 in [5]. 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 [5] that $\psi(H^r_2(t)) \leq \tilde{U}^r(t) \leq \psi(H^r_2(t)) + 2\bar{c}\varepsilon^r$.
Recall that $\psi(\cdot)$ is nonincreasing and Lipschitz continuous with unity constant, and $H^r_2(t) = \hat{H}^r(t) + \bar{c}\varepsilon^r$. It follows that

$$\psi(\hat{H}^r(t)) - \bar{c}\varepsilon^r \leq \tilde{U}^r(t) \leq \psi(\hat{H}^r(t)) + 2\bar{c}\varepsilon^r, \tag{A27}$$

which in turn implies $\varphi(\hat{H}^r(t)) - 2\bar{c}\varepsilon^r \leq \tilde{H}^r(t) \leq \varphi(\hat{H}^r(t)) + \bar{c}\varepsilon^r$.
*Step 4:*
Subtracting $\hat{H}^r(t)$ by the auxiliary process $\hat{X}(t)$, we obtain

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

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 (A27), this inequality can be rewritten as $|\hat{H}^r(t) - \hat{X}(t)| \leq \gamma c K \int_0^t |\hat{H}^r(s) - \hat{X}(s)|ds + |V^r(t) - V(t) + o_p(1)| + 2\gamma c^2 \varepsilon^r t$.
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

$$\sup_{t \le T} |V^r(t) - V(t)| \le v^r, \;\; w.p.\; 1 - \theta^r, \tag{A28}$$

where $\theta^r \downarrow 0$ as $r \to \infty$. In the sequel we concentrate on sample paths in $\Omega(\varepsilon^r)$ for which (A28) holds, i.e., we consider only $\omega \in \Omega(\varepsilon^r, v^r) \equiv \Omega(\varepsilon^r) \cap \Omega(v^r)$, 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 $\Omega(\varepsilon^r, v^r)$ is more than $1 - \delta^r - \theta^r$. Then,

$$|\hat{H}^r(t) - \hat{X}(t)| \le \gamma c K \int_0^t |\hat{H}^r(s) - \hat{X}(s)| ds + v^r + 2\gamma c^2 \varepsilon^r t \le (v^r + 2\gamma c^2 \varepsilon^r T) e^{\gamma c K T}, \tag{A29}$$

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

From the Lipschitz continuity of $\varphi(\cdot)$, (A25), and (A29), we get that $|\tilde{H}^r(t) - \tilde{X}(t)| \le K(v^r + 2\gamma c^2 \varepsilon^r T) e^{\gamma c K T} + 2\bar{c}\varepsilon^r$. Let $r \to \infty$, sequences $\varepsilon^r, v^r, \delta^r, \theta^r$ all vanish and therefore we conclude that

$$\sup_{t \le T} |\tilde{H}^r(t) - \tilde{X}(t)| \to 0 \;\; inprobability. \tag{A30}$$

From (A30), 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*

We will follow [6] 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, $G(\pi) = 0, \forall \pi < \underline{s}, 0 \le G(\pi) < 1$, $\forall \pi \in (\underline{s}, \bar{s})$, and $G(\pi) = 1, \forall \pi \ge \bar{s}$. 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 $\pi$.

First we observe that $\underline{s} \ge \underline{\pi}$, $\forall i \in \mathscr{N}$. To see this, if a supplier chooses $\pi^*$, his expected payoff is at least $\mu \pi^* - \mathscr{L}(\pi^*) = \Psi^* = \mu \underline{\pi}$. 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

$$\mu \underline{s} - \frac{1}{n} \mathscr{L}(\underline{s}) [G(\underline{s})]^{n-1} < \mu \underline{s}. \tag{A31}$$

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 $\underline{s}-\varepsilon$ while other suppliers set prices at $\underline{s}$. Thus, his expected payoff is $\mu(\underline{s}-\varepsilon)$ (without any penalty incurred). When $\varepsilon$ is small enough, $\mu(\underline{s}-\varepsilon)$ is strictly higher than the LHS of (A31). This implies that the supplier will always intend to undercut the price by a sufficiently small $\varepsilon$.

We now define $\Psi_i(\pi)$ as the expected payoff of supplier $i$ if he places price $\pi$, and $\Psi_i^e$ his equilibrium payoff. We claim that the suppliers get equal payoffs in equilibrium, i.e., $\Psi_i^e = \Psi_j^e = \mu\underline{s},\ \forall i,j \in \mathscr{N}$. 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 $\Psi_i^e = \Psi_i(\underline{s}) = \mu\underline{s},\ \forall i \in \mathscr{N}$.

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

$$\Psi_i(\bar{s}) = \mu\bar{s} - \mathscr{L}(\bar{s}) < \mu\pi^* - \mathscr{L}(\pi^*) = \mu\underline{\pi} \leq \mu\underline{s}, \tag{A32}$$

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

$$\mu\underline{s} = \Psi_i^e = \Psi_i(\pi^*) = \mu\pi^* - \mathscr{L}(\pi^*) = \mu\underline{\pi}, \tag{A33}$$

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 $\pi$. Following the argument in Lemma 10 of [6], it pays for a supplier to transfer a $\varepsilon$-neighborhood mass above $\pi$ to a $\delta$-neighborhood below $\pi$, 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 $(\pi_1, \pi_2)$ such that $G(\pi_1) = G(\pi_2)$. Since $[G(\pi)]^{n-1}$ (i.e., the probability that a supplier quoting $\pi$ becomes the most expensive supplier) does not change in $(\pi_1, \pi_2)$, by moving a small mass from slightly below $\pi_1$ to $\pi_2$, a supplier is strictly better off. That is, the supplier charges a higher price $\pi$, 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, $\pi$ is involved in supplier $i$'s equilibrium strategy. Therefore a supplier should get the same payoff for all such $\pi$'s (otherwise he would not be willing to play the equilibrium strategy). Thus, For all $\pi \in [\underline{\pi}, \pi^*]$,

$$\Psi_i^e = \mu\underline{s} = \Psi_i(\pi) = \mu\pi - \mathscr{L}(\pi)[G(\pi)]^{n-1}, \forall \pi \in [\underline{\pi}, \bar{\pi}^*], \tag{A34}$$

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 $G(\pi)$.

Finally, the above derivations also imply that no pure-strategy equilibrium exists, because the formulation actually allows for probability mass in $G(\pi)$. As we verify here, there always exist profitable deviations in such a scenario. This completes the proof. □

## *Proof of Proposition 4*

Given $\hat{\mu}$ (and other relevant parameters such as c and $\sigma$), $\mathscr{L}(\pi)$ is a known function and therefore $\Pi^C$ is a fixed constant independent of $n$, the number of suppliers. From Proposition 3, $\Pi^* = n\Psi^*$, where $\Psi^* = \max_\pi[\mu\pi - \mathscr{L}(\pi)] = \max_\pi[\frac{\hat{\mu}}{n}\pi - \mathscr{L}(\pi)]$. The function $\mathscr{L}(\pi)$ is strictly convex and positive, and therefore $\Psi^*$ can be obtained via the first-order condition. When $n$ is sufficiently large, the maximizer $\pi^* = \arg\max_\pi[\frac{\hat{\mu}}{n}\pi - \mathscr{L}(\pi)]$ 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

$$|\Pi^C - \Pi^*| = \Pi^C - \Pi^* = \Pi^C - n\max_\pi[\frac{\hat{\mu}}{n}\pi - \mathscr{L}(\pi)] > \Pi^C - n\frac{\hat{\mu}}{n}\pi^* = -\hat{\mu}\pi^* + \Pi^C > M. \tag{A35}$$

This completes the proof. □

## *Proof of Proposition 5*

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 $\pi$. Let $\Psi_i(\pi, G_{-i})$ be the expected payoff of supplier $i$ when he plays $\pi$ and other suppliers adopt the mixing probabilities $G_{-i}$. Let $\Psi_i^e$ 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 $\max_{j\in\mathscr{N}} \underline{s}_j$. Combining the above, we know that $\underline{s}_i := \underline{s} \geq \max_{j\in\mathscr{N}} \underline{\pi}_j, \forall i \in \mathscr{N}$.

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

$$\Psi_i(\underline{s}, G_{-i}) = \mu_i \underline{s} - \mathscr{L}(\underline{s}) \times 0 = \mu_i \underline{s}, \tag{A36}$$

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

We now focus on the two-supplier case. The following lemma provides some relations of $\pi_1^*, \pi_2^*$ and $\underline{\pi}_1, \underline{\pi}_2$, which are needed for characterizing the equilibrium mixing probabilities.

**Lemma A7** *If $\mu_1 > \mu_2$, then $\pi_1^* > \pi_2^*$ and $\underline{\pi}_1 > \underline{\pi}_2$.*

We now return to the proof of Proposition 5. 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 $\pi_1^*$. 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

$$\Psi_2(\bar{s}, G_{-2}) = \mu_2 \bar{s} - \mathscr{L}(\bar{s}) \leq \Psi_2^* = \mu_2 \underline{\pi}_2 < \mu_2 \underline{\pi}_1 \leq \mu_2 \underline{s}, \tag{A37}$$

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,

$$\Psi_1(\bar{s}, G_{-1}) = \mu_1 \bar{s} - \mathscr{L}(\bar{s}) \leq \mu_1 \pi_1^* - \mathscr{L}(\pi_1^*) = \mu_1 \underline{\pi}_1 \leq \mu_i \underline{s} = \Psi_1(\underline{s}, G_{-1}), \tag{A38}$$

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

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

Similarly,

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

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

## *Proof of Proposition 6*

From the *PS* rule, we can define $\Pi_i^{PS}(\pi_i, \pi_{-i}) = \mu_i\pi_i - \frac{\mu_i}{\hat{\mu}}\mathscr{L}(\max_{j\in\mathscr{N}}\pi_j) = \mu_i[\pi_i - \dfrac{1}{\hat{\mu}}\mathscr{L}(\max_{j\in\mathscr{N}}\pi_j)]$ as the revenue for supplier $i$ when other suppliers select $\pi_{-i}$. 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 $\Pi_i^{PS}$ with respect to $\pi_i$ becomes $\dfrac{\partial \Pi_i^{PS}(\pi_i, \pi_{-i})}{\partial \pi_i} = \mu_i[1 - \dfrac{1}{\hat{\mu}}\mathscr{L}'(\max_{j\in\mathscr{N}}\pi_j)\mathbb{1}\{\pi_i = \max_{j\in\mathscr{N}}\pi_j\}]$.

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

## *Proof of Proposition 7*

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

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

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

### *Proof of Lemma 3*

Since the second term in (29) is independent of the lower static prices $\pi_i, i \notin J$, at optimality these $\{\pi_i\}'$s should all be equal, i.e., $\pi_i = \hat{\pi}$, $\forall i \in \mathscr{N}$. 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 ([27, §5]).[2] Thus, $\pi^C$ is well-defined and unique. □

## A3 Proofs of auxiliary lemmas in Sections 3 and 4

### *Proof of Lemma 1*

Recall that using the strong approximation theorem ([14, Theorem 7]), the service time process of supplier $i$ can be expressed as follows:

$$S_i^r(t) = \mu_i^r(t - Y_i^r(t)) + \sqrt{r}\sigma_i B_{s,i}(t - Y_i^r(t)) + O(\log rt), \tag{A39}$$

where $B_{s,i}$ is the associated Brownian motion with standard deviation $\sigma_i$, 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 $r \to \infty$,

$$\frac{R_i^r(t)}{r} = \frac{1}{r}(\bar{p} + \frac{\pi_i}{\sqrt{r}})\left[r\mu_i(t - Y_i^r(t)) + \sqrt{r}\sigma_i B_{s,i}(t - Y_i^r(t)) + O(\log rt)\right] \to \bar{p}\mu_i t, \tag{A40}$$

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

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

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

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 A1

Our first goal is to show that the scaled arrival process $A_i^{r,m}(t)$ has an upward jump if and only if $\mathbf{I}_i(Q^{r,m}(t)) = 1$ and $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq v$, where $v$ is the customer's willingness to pay, where the routing indicator clearly specifies which supplier should get a new buyer if her valuation is higher than the full price.
To this end, we shall consider two events

$$\{\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{\sqrt{r}Q_j^{r,m}(t)+1}{r\mu_j}\} \qquad \text{(A41)}$$

and

$$\{\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq v\}. \qquad \text{(A42)}$$

We recall the definition of $Q^{r,m}(t)$ and establish the following equivalence:

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

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, (A43) implies that the total cost submitted by supplier $i$ is less than that by supplier $j$.

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

$$\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{Q_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)+1}{\mu_i^r} = \bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq v. \tag{A43}$$

Therefore, if $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{\sqrt{r}Q_j^{r,m}(t)+1}{r\mu_i}$, $\forall j \neq i$, and $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{\sqrt{r}Q_j^{r,m}(t)+1}{r\mu_i}$, $\forall j < i$, and $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq v$, then $A_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m)$, the arrival process routed to server $i$ at time epoch $\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m$, has an upward jump. 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 $A_i^{r,m}(t)$ has an upward jump if and only if $\mathbf{I}_i(Q^{r,m}(t)) = 1$ and $\bar{p} + \frac{\pi_i}{\sqrt{r}} + c\frac{\sqrt{r}Q_i^{r,m}(t)+1}{r\mu_i} \leq v$. Thus, the variation associated with $A_i^{r,m}(t)$ for a specific supplier $i$ is upper bounded by the variation associated with the counting process of the aggregate arrivals, and this bound applies uniformly to all the suppliers $i \in \mathcal{N}$. Consequently, we obtain that

$$r\max_{i\in\mathcal{N}} ||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L \leq ||N(t) - t||_{Lr}, \tag{A44}$$

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

$$0 \leq r\int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du \leq Lr, \forall t \in [0,L]. \tag{A45}$$

Note that in the above discussions, we have applied the tie-breaking rule that the buyer chooses the supplier with the smallest index. This is inconsequential as it is also the buyer's best response and thus can be sustained as an equilibrium.

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

$$P\left(\max_{m<\sqrt{r}\tau}\max_{i\in\mathcal{N}} ||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L > \varepsilon\right)$$
$$\leq \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} P\left\{||A_i^{r,m}(t) - \int_0^t \Lambda^r \mathbf{I}_i(Q^{r,m}(u))\mathbf{J}_i^r(Q_i^{r,m}(u))du||_L > \varepsilon\right\}$$

$$\begin{aligned}
&\leq \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} P\left\{\frac{1}{r}||N(t)-t||_{Lr} > \varepsilon\right\} \\
&\leq \sum_{m=1}^{\lfloor\sqrt{r}\tau\rfloor} \frac{\varepsilon}{L^2 r} \\
&\leq \lfloor\sqrt{r}\tau\rfloor \frac{1}{L^2 r}\varepsilon,
\end{aligned}$$

where in the third second inequality we have applied [8, 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

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

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

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

$$= \pi_i + \sqrt{r}c\frac{Q_i^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) + 1}{r\mu_i} - \left(\pi_j + \sqrt{r}c\frac{Q_j^r(\frac{1}{\sqrt{r}}t + \frac{1}{\sqrt{r}}m) + 1}{r\mu_j}\right) - \frac{c}{\sqrt{r}}(\frac{1}{\mu_i} - \frac{1}{\mu_j})$$

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

Therefore,

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

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:

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

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 $\varepsilon$ is small. This completes the proof. □

## Proof of Lemma A2

The proof is similar to that of [8, 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

$$A_i^{r,m}(t_2) - A_i^{r,m}(t_1) = \frac{1}{\sqrt{r}} N\left( \int_{\frac{1}{\sqrt{r}}t_1 + \frac{1}{\sqrt{r}}m}^{\frac{1}{\sqrt{r}}t_2 + \frac{1}{\sqrt{r}}m} \Lambda^r \mathbf{I}_i(Q^r(t)) P(v \geq \min_{j \in \mathcal{N}} \left\{ \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(t+m)+1}{\mu_j^r} \right\}) dt \right), \tag{A46}$$

where $N(\cdot)$ is the counting process of a unit-rate Poisson. Note that the arrival rate is bounded $\Lambda^r \mathbf{I}_i(Q^r(t)) P(v \geq \min_{j \in \mathcal{N}} \left\{ \bar{p} + \frac{\pi_j}{\sqrt{r}} + c\frac{Q_j^r(t+m)+1}{\mu_j^r} \right\}) \leq r$. 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 [8, 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 A4

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 A3 and the definition of $K^r$ lead to this lemma immediately according to [8, Proposition 4.1]. □

## Proof of Lemma A5

This follows from [8, 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:

$$\widehat{Q}_i(t) = \widehat{Q}_i(0) + \int_0^t \widehat{\Lambda}\mathbf{I}_i(\widehat{Q}(u))du - \widehat{D}_i(t), \forall i \in \mathscr{N}. \tag{A47}$$

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 A6

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

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

$$\sum_{j \neq m} \frac{\tilde{Q}_j^r(s)}{\mu_j} < \zeta - \varepsilon^r - \frac{\tilde{Q}_m^r(s)}{\mu_m} \leq \zeta - \varepsilon^r. \tag{A48}$$

Now since $\varepsilon^r$ is given, we can define $\eta^r = \dfrac{c}{2(n-2)}\varepsilon^r > 0$ and $\eta^r \downarrow 0$ because $\varepsilon^r \downarrow 0$.

If for all $r$ we can find a pair $i, j \in \mathscr{N}$ such that $\pi_j + \dfrac{\tilde{Q}_j^r(s)}{\mu_j} \leq \pi_i + \dfrac{\tilde{Q}_i^r(s)}{\mu_i} - \eta^r$, then letting $r \to \infty$ we have found a sequence that contradicts Proposition 2. Thus from now on we assume that

$$\pi_j + \frac{\tilde{Q}_j^r(s)}{\mu_j} > \pi_i + \frac{\tilde{Q}_i^r(s)}{\mu_i} - \eta^r, \ \forall j \neq i, m; i, j \in \mathscr{N}. \tag{A49}$$

Combining Equations (A48) and (A49), we obtain

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

The proposed cost by supplier $i$ is upper bounded by:

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

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

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

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

$$\pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i} \leq \pi_m + c\frac{\tilde{Q}_m^r(s)}{\mu_m} - \frac{c}{2(n-1)}\varepsilon^r, \tag{A50}$$

which contradicts Proposition 2 if we let $r \to \infty$. since costs proposed by supplier $i$ and $m$ are different. Thus, $\tilde{W}^r(s) \geq \zeta - \varepsilon^r$, $\forall r \geq r_0$.
The next thing is to show that $\tilde{Q}^r(s) > 0$ in probability if $\tilde{W}^r(s) \geq \zeta + \varepsilon^r$. We again prove this by contradiction. Suppose that there exists $i$ such that $\tilde{Q}_i^r(s) = 0$. If $i \notin \arg\max_{j \in \mathscr{N}} \pi_j$, then

$$\pi_m + c\frac{\tilde{Q}_m^r(s)}{\mu_m} \geq \pi_m > \pi_i = \pi_i + c\frac{\tilde{Q}_i^r(s)}{\mu_i}, \tag{A51}$$

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 $\eta^r$ and apply state space collapse, we have

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

Note that $\tilde{Q}^r_m(s) = 0$ implies $\zeta + \varepsilon^r \leq \tilde{W}^r(s) = \sum_{j \neq m} \dfrac{\tilde{Q}^r_j(s)}{\mu_j}$, and therefore

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

The cost proposed by supplier $i$ is lower bounded by

$$\pi_i + c\frac{\tilde{Q}^r_i(s)}{\mu_i} \geq \pi_i + \frac{c}{n-1}[\frac{1}{c/n}\pi_m - \frac{1}{c/n}\frac{1}{n}\sum_{k \in \mathscr{N}} \pi_k - \frac{n-2}{c}\pi_i + \frac{1}{c}\sum_{j \neq i,m} \pi_j] + c(\varepsilon^r - \frac{(n-2)}{c}\eta^r). \tag{A52}$$

After some manipulation, the above inequality can be rewritten as

$$\pi_i + c\frac{\tilde{Q}^r_i(s)}{\mu_i} \geq \pi_m + \frac{c}{2}\varepsilon^r = \pi_m + c\frac{\tilde{Q}^r_m(s)}{\mu_m} + \frac{c}{2}\varepsilon^r, \tag{A53}$$

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

## *Proof of Lemma A7*

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

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

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 $\mu_1, \mu_2$, and hence $\Psi^*(\mu)$ is convex in $\mu$.

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