Decentralized Service Supply Chains with Multiple Time-Sensitive Customer Segments: Pricing, Capacity, and Coordination

Philipp Afèche

Rotman School of Management, University of Toronto, Toronto, ON M5S3E6, Canada
afeche@rotman.utoronto.ca

February 2013

We study decentralized pricing and capacity decisions in a congestion-prone service supply chain that serves heterogeneous price- and time-sensitive customers. Independent, profit-maximizing providers control the network and offer complementary services. Some customers require the service bundle (cross-traffic), others only one of its components (local-traffic). The model is motivated by the provision of data transport services over the Internet. Our main findings are: 1. We identify two opposite performance effects under decentralized control without coordination, double-marginalization for cross-traffic prices and local-to-cross-traffic substitution. In the absence of local-traffic, decentralized control yields an undercongested network, due to double-marginalization and delay costs. However, in the presence of time-sensitive local-traffic the substitution effect may improve performance by offsetting the detrimental impact of double-marginalization. The benefit of this substitution effect critically depends on the relative delay costs and demand elasticities of the local- and cross-traffic flows. 2. We show for providers with balanced traffic flows that peering contracts, a common Internet interconnection agreement, overcongest the network, which lends support for the use of transfer prices. 3. We identify the optimal linear transfer price and show that a range of transfer prices yield a Pareto improvement compared to decentralized control and peering equilibria.

1 Introduction

We study decentralized pricing and capacity decisions in a congestion-prone service supply chain that serves heterogeneous price- and time-sensitive customers. Independent, profit-maximizing providers control the network and offer complementary services. Some customers require the service bundle (cross-traffic), others only one of its components (local-traffic). The model we study captures important features of the provision of data transport services over the Internet, where multiple time-sensitive traffic flows travel across or within interconnected subnetworks that are controlled by independent service providers such as AT&T, Cogent Communications, and Verizon. Similar service supply chains may also arise in sectors such as healthcare, transportation, financial services, and make-to-order manufacturing, whenever capacity-constrained complementary firms offer standalone and bundled products or services to time-sensitive customers.

The pricing and capacity decisions of decentralized service network providers jointly determine the prices and service qualities offered to customers. This raises fundamental questions on the relationship among demand structure, network control, and system performance: How does a decentralized service supply chain perform, compared to the centralized optimum, if providers make

their decisions independently, without coordination mechanisms? Which coordination mechanisms are effective in mitigating potential performance losses that arise under decentralized control, and how do these mechanisms compare to those that are prevalent in practice?

This paper contributes answers to these questions. We study the benchmark case of *decentralized control* without coordination, compare it to the overall optimum under *centralized control*, and consider the performance of two coordination mechanisms, *peering* contracts and *transfer prices*.

The interactions among decentralized service network providers depend on whether their resources and services are complements or substitutes. This paper focuses on the complementary case. We model the service supply chain as a two-node tandem queue. Each node has a single server with finite processing capacity. Under decentralized control, two independent providers each own one of the nodes, whereas under centralized control there is a single provider. Customer requests arrive in Poisson fashion to the network and are served FIFO at each node along their route. We consider customer segments which differ in their market sizes, routing requirements, service value distributions, and delay costs. Motivated by the data communication setting, we refer to segments that require service from both nodes as *cross-traffic*, and to those that require service from a single node as *local-traffic*. (More generally, one may interpret local-traffic as a standalone service offered by a specialized resource and cross-traffic as a bundle of complementary services that requires multiple resources with distinct capabilities.) Providers make pricing and capacity decisions, and customers base their purchase decisions on their service values and delay costs, and on the prices and expected delays. The analysis generates the following results.

Under decentralized control without coordination, providers independently set their prices and capacity levels, and customers pay the sum of prices along their route. We characterize the Nash equilibrium and identify two opposite performance effects, double-marginalization for cross-traffic prices and substitution between local- and cross-traffic. The equilibrium price expressions reflect how the double-marginalization and substitution effects depend on customer routes, demand elasticities, and delay costs. In the absence of local-traffic, decentralized control yields an *undercongested* system, a lower capacity and a higher price than at the overall optimum. These performance losses are due to the standard double-marginalization effect, and they are significantly amplified by the delay costs considered here. However, in the presence of *time-sensitive local-traffic* the substitution effect may improve performance by partially or completely offsetting *the detrimental double-marginalization* effect. The benefit of this substitution effect critically depends on the demand characteristics of local- and cross-traffic flows, i.e., their delay costs and demand elasticities. Specifically, if the local- and cross-traffic are equally elastic, then the performance loss under decentralized control is smaller the more time-sensitive the local- relative to the cross-traffic and the

larger the demand elasticity. However, if local- and cross-traffic differ in their elasticities, then the performance loss under decentralized control is minimized if local-traffic is infinitely elastic, and the loss is maximized if local-traffic is somewhat less elastic than cross-traffic.

We study two coordination mechanisms and discuss their effectiveness in eliminating the potential performance losses of decentralized control. *Peering* contracts, a common Internet interconnection agreement, whereby networks with comparable size and traffic flow balance agree to exchange traffic for free, and *transfer prices,* whereby networks pay each other for forwarding traffic. We show for providers with balanced traffic flows that *peering* contracts *overcongest* the network if customers are time-sensitive; peering only performs well in the absence of congestion effects. Peering eliminates the double-marginalization effect, but it introduces a qualitatively different externality: Since providers only generate revenue from a subset of customers, they ignore the delay cost of their pricing decisions on each others' customers. This result suggests that, contrary to the common Internet connection practice, providers should *not* peer for free if their customers are time-sensitive and network delays are significant, even if their traffic flows are balanced. This result also lends some support for the use of transfer prices. We study equilibria with linear transfer prices, identify the optimal transfer price, and show that a range of transfer prices yield a Pareto improvement compared to the decentralized control and peering equilibria.

Related Literature. This paper is at the intersection of three research streams: pricing for queues in the operations research/management science literature; network control in the electrical engineering/computer science literature; and network interconnection in the economics literature.

See Hassin and Haviv (2003) for a survey of pricing for queues. Virtually all papers in this literature study a *single* provider, e.g., Naor (1969), Mendelson (1985), Masuda and Whang (1999); or competing providers who offer *substitutable* services, e.g., Li and Lee (1994), MacKie-Mason and Varian (1995), Lederer and Li (1997), Shneorson and Mendelson (2003), Allon and Federgruen (2007). The case of *complementary* services considered here seems to have hardly been considered so far. Hassin and Veltman (2005) consider two complementary providers; however, only one of them is capacity-constrained and customers are homogeneous, in contrast to our model.

In the network control literature there is a growing number of studies that consider economic issues. Among these, papers on the efficiency implications of decentralized network control are somewhat close to this paper. See Acemoglu and Ozdaglar (2007) and references therein. That literature restricts attention to *homogeneous* users (in terms of their utilities and delay costs) and focuses on establishing worst-case bounds on efficiency losses under decentralized control. Papers in this stream that consider complementary providers also restrict attention to a single traffic flow; see

He and Walrand (2006) and Acemoglu and Ozdaglar (2007). In contrast, this paper provides results on how customer *heterogeneity* (in terms of utilities, delay costs, and routing) affects decentralized system performance, and it characterizes and compares several decentralized pricing schemes. The model of He and Walrand 2006 further differs from ours in that it ignores queueing effects and delay costs, which yields significantly different results. (Specifically, the capacity constraint can be binding in their model, in which case the provider with the bottleneck capacity charges the higher equilibrium price. In contrast, our results in §§3.1-3.2 for the model with only cross-traffic, prescribe equal equilibrium prices for both providers.)

In economics, studies of interconnection for traditional telecommunication networks, e.g., Cave and Donnelly (1996), Laffont et al. (1998a), Laffont et al. (1998a), or for Internet service providers, e.g., Milgrom et al. (1999), Cremer et al. (2000), Laffont et al. (2003), ignore operational characteristics, i.e., congestion and the resulting quality and substitution effects.

Finally, there are loose connections between this paper and further streams of the operations management literature. In service operations Farzan and Zhou (2012) study staffing and effort decisions in two-stage tandem M/M/s queues to affect waiting time and service quality. Interactions among complementary service providers with capacity constraints are also important in code sharing agreements among airline alliance partners (e.g., Netessine and Shumsky 2005; Hu et al. 2012).

There is an extensive literature on decentralized inventory supply chains; see Cachon (2003) for a survey. However, the results in that literature are not immediately applicable to service supply chains, because of the inherent differences between the two types of systems. Nevertheless, there are some interesting connections. E.g., Netessine and Zhang (2005) study for inventory supply chains how externalities due to retailer competition interact with double-marginalization. We identify for service networks without competition how externalities due to queueing and substitution effects interact with double-marginalization.

Plan of the Paper. In §2 we describe the model. In §3 we study the equilibrium prices and capacities under decentralized control without coordination. In §4 we analyze the performance of peering contracts and transfer prices. In §5 we discuss the robustness of our results. Our concluding remarks are in §6. Proofs are in the Online Supplement.

2 The Model

We specify the network and control structure in §2.1 and the demand structure in §2.2.

2.1 Network and Control Structure

We consider a tandem queueing network that consists of two single-server nodes (or resources), indexed by $n = 1, 2$. Customer requests for service arrive in Poisson fashion to the network and are served FIFO at each node along their route. Let μ_n be the average service rate of node n and $\boldsymbol{\mu} = (\mu_1, \mu_2)$ denote the capacity vector. Service times at node n are i.i.d. exponential random variables with mean μ_n^{-1} and are independent of the arrival process. The queueing process at each node is therefore quasi-reversible; the queue length vector has a product form stationary distribution with the marginal probabilities given by the standard M/M/1 formula, see Kelly (1979).

In the Internet context, a node corresponds to a subnetwork, μ_n is the bottleneck capacity of subnetwork n, and a request is for the transmission of a file such as an e-mail or a web page.

We assume that the variable cost of using capacity is negligible relative to that of building capacity. Specifically, once capacity is in place, the direct cost of serving a customer request is zero. This is appropriate for data communications where the marginal cost of sending traffic over an existing network is zero. (The capacity and congestion costs are discussed below.)

We study the performance under two control structures: *centralized control* by a monopoly and *decentralized control* whereby two independent providers each own one of the nodes.

2.2 Demand Structure

The network faces a population of small price- and time-sensitive potential customers whose individual arrival rates are infinitesimal relative to that of the market. Each arrival is for one unit of service. We consider three customer segments or types, which differ in their market sizes, routing requirements, service value distributions, and delay costs. The first segment requires service from node 1 only, the second from node 2 only, and the third segment requires the services of both nodes. Motivated by the data network setting, we index these segments by $i \in \{L1, L2, C\}$, where $L1$ ($L2$) refers to node-1 (node-2) *local-traffic* and C to *cross-traffic*. More generically, local-traffic corresponds to a standalone service offered by a single resource, and cross-traffic to a service bundle that requires complementary resources. Potential type-i customer requests arrive to the network according to an exogenous Poisson process with rate Λ_i. Their *actual* demand rate, denoted by λ_i, depends on their service values and delay costs, and on the prices and delays as discussed below.

Figure 1 portrays the network under decentralized control. Provider n controls node n, charges P_{Ln} for type-Ln service, and p_{Cn} for its part of the type-C service. The total type-C price is $P_C = p_{C1} + p_{C2}$. Each provider has sole control of local-traffic pricing, but the local- and cross-traffic flows are interdependent since they compete for capacity. This structure represents the simplest model of decentralized control for a network with standalone and bundled services.

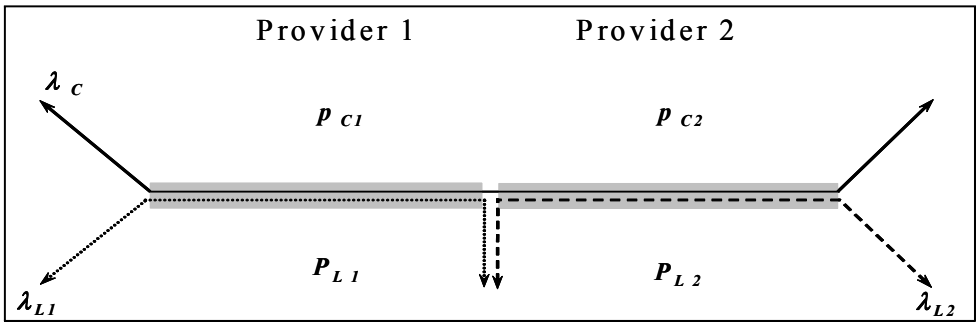

Figure 1: Decentralized network with local- and cross-traffic. Provider n charges P_{Ln} and p_{Cn} for local- and cross-traffic, respectively. The demand rates depend on service values, delay costs, prices, and delays.

Next, we specify the model with cross-traffic only. (We describe the local-traffic flows in §3.3.)

Service Values. Cross-traffic customers may be heterogeneous in their service values, i.e., their willingness to pay for delay-free service. Service values are i.i.d. draws from a continuous probability distribution with positive support and c.d.f. F, which is independent of the arrival and service process. Define the marginal value function

$$V_C'(\lambda_C) := \overline{F}^{-1}\left(\frac{\lambda_C}{\Lambda_C}\right), \quad \lambda_C \in [0, \Lambda_C], \tag{1}$$

where $\overline{F} = 1 - F$. That is, the value of the marginal type-C request corresponding to λ_C is $V_C'(\lambda_C)$. We consider the family of isoelastic marginal value functions

$$V_C'(\lambda_C) = A_C \cdot \lambda_C^{-\alpha}, \quad \lambda_C \in [0, \Lambda_C], \quad \alpha \in [0, 1), \tag{2}$$

where $A_C > 0$ is the demand intensity parameter and $\epsilon(\lambda) := -V_C'(\lambda)/(\lambda V_C''(\lambda)) = 1/\alpha$ is the elasticity. The aggregate gross value rate as a function of λ_C is $V_C(\lambda_C) = A_C \cdot (\lambda_C)^{1-\alpha}/(1-\alpha)$.

This service value model accommodates a range of scenarios. It yields an identical value A_C for all customers if $\alpha = 0$, and a Pareto value distribution with shape parameter $1/\alpha$ and minimum value $A_C \cdot \Lambda_C^{-\alpha}$ if $\alpha > 0$. In §5 we discuss the sensitivity of our results to the value distribution.

Delay Costs. Customers are time-sensitive. Their utility decreases in their service delay, defined as the time interval between a service request and its completion. We consider a *multiplicative* delay cost structure. A customer with service value v who experiences a delay w has a *net* value of $v \cdot D_C(w)$ from service. The *delay discount function* $D_C(w)$ is non-increasing in w and satisfies $D_C(0) = 1$ and $D_C(w) < 1$ for some $w > 0$. This multiplicative structure is well-suited to settings where delay reduces the value from service. (See Afèche and Mendelson Afèche and Mendelson (2004) for a detailed analysis of this delay cost structure for a single-provider system.) A variety of important phenomena lead to delay-driven value losses: (i) *Delayed information*: A delay in the

receipt or use of information adversely affects its value. For example, a delay in the execution of an order to trade a security deflates the trader's expected profit if part or all of the anticipated price change occurs prior to execution; see Dewan and Mendelson (1998). The rise of electronic markets and dynamic pricing extend the relevance of this scenario to many industrial markets where transactions involve the communication of valuable and time-sensitive information. (*ii*) *Audio or video signal distortions* caused by delays in real-time transmissions over the Internet. (*iii*) *Physical decay*, e.g., during transportation delays for perishable goods. (*iv*) *Technological obsolescence* of short life cycle products such as computer chips. While our model applies to a variety of D_C functions, we focus on the linear case $D_C(w) = 1 - d_C \cdot w$, where $d_C > 0$ is the linear *delay sensitivity parameter.*

Information Structure. The providers do not observe individual customers' service values, but they know the capacity levels and the aggregate demand characteristics, i.e., the value distribution F, the delay sensitivity parameter d_C, the rate Λ_C, and the statistical properties of the arrival and service processes. A customer knows her service value and delay sensitivity parameter but only her *expected* service time when making her purchase decision. Actual service times become known only once processing is completed. Customers do not observe the dynamically-changing network state. They base their purchase decisions on prices and the *expected* delays posted by the providers as detailed below. While the actual delays of individual customers deviate from the posted averages, we require that the posted averages equal the realized *average* steady-state delays, given the capacities $\boldsymbol{\mu}$ and the demand rate λ_C. This requirement captures the notion that reputation effects and third party auditors commit the providers to perform in line with their announcements.

Quality of Service and Expected Net Values. Customers evaluate their expected net value from service based on the *expected* delay discount factor $Q_C(\lambda_C, \boldsymbol{\mu}) := E[D_C(W_C(\lambda_C, \boldsymbol{\mu}))]$, where the random variable W_C is the steady-state end-to-end delay for cross-traffic. We also refer to Q_C as the Quality of Service or QoS. It satisfies

$$Q_C(\lambda_C, \boldsymbol{\mu}) = 1 - d_C \cdot E[W_C(\lambda_C, \boldsymbol{\mu})] = 1 - d_C \sum_{n=1}^{2} \frac{1}{\mu_n - \lambda_C}, \tag{3}$$

where $\lambda_C < \min(\mu_1, \mu_2)$ for stability. Note that $\partial Q_C/\partial \lambda_C < 0$, $\partial^2 Q_C/\partial \lambda_C^2 < 0$, $\partial Q_C/\partial \mu_n > 0$, and $\partial^2 Q_C/\partial \mu_n^2 < 0$. The expected net value of a request with service value v equals $v \cdot Q_C(\lambda_C, \boldsymbol{\mu})$, and the expected aggregate net value rate per unit time is $V_C(\lambda_C) Q_C(\lambda_C, \boldsymbol{\mu})$. For simplicity we suppress the argument $\boldsymbol{\mu}$ and write $Q_C(\lambda_C)$ when considering settings with fixed capacity.

Technical Assumptions. The analysis focuses on capacity levels $\boldsymbol{\mu}$ that satisfy the following.

A1. $Q_C(0, \boldsymbol{\mu}) > 0$.

A2. $Q_C(\lambda_C, \boldsymbol{\mu}) < 0$ as $\lambda_C \to \min(\Lambda_C, \mu_1, \mu_2)$.

Assumption *A1* rules out the trivial case where it is unprofitable to serve any customers, and *A2* ensures that it is not optimal to serve all customers.

3 Decentralized Control without Coordination

This section analyzes the benchmark case of decentralized control without coordinating contracts: Each provider independently chooses its own profit-maximizing prices and capacity. These independent decisions *jointly* determine the total cross-traffic price and the QoS offered to customers. To develop intuition for the problem, in §§3.1-3.2 we first consider the base case in which there is only cross-traffic. In §§3.3-3.4 we then consider the effect of adding local-traffic on each resource.

3.1 Cross-Traffic Service: Pricing for Fixed Capacity

Consider the following price game for a given capacity. Provider n charges a price p_{Cn} for its resource. The total charge for the service bundle, denoted by P_C, equals the sum of the resource prices: $P_C = p_{C1}+p_{C2}$. Providers set their prices simultaneously and independently. Customers are self-interested and seek to maximize their own expected utility. For given price $P_C < V_C'(0)Q_C(0)$ there is a unique Nash equilibrium demand rate $\lambda_C(P_C)$. The corresponding marginal customer with value $V_C'(\lambda_C(P_C))$ has zero expected utility, i.e., P_C satisfies the inverse demand function

$$P_C(\lambda_C) := V_C'(\lambda_C)Q_C(\lambda_C). \tag{4}$$

Decentralized Control. We consider Nash equilibria in pure undominated price strategies. Given firm 2 charges p_{C2}, provider 1 determines her best response $p_{C1}(p_{C2})$. It is convenient to perform the analysis in terms of the demand rate λ_C. Provider 1 solves:

$$\begin{aligned} \max_{\lambda_C} \Pi_1(\lambda_C; p_{C2}) &= \lambda_C \cdot p_{C1}(\lambda_C; p_{C2}) = \lambda_C \cdot (P_C(\lambda_C) - p_{C2}) \\ s.t. &\quad 0 \le \lambda_C < \mu_n, \quad n = 1, 2. \end{aligned}$$

The providers' first-order conditions are

$$\lambda_C \cdot P_C'(\lambda_C) + P_C(\lambda_C) = p_{Cn}, \quad n = 1, 2. \tag{5}$$

Proposition 1 characterizes the decentralized control equilibrium, denoted by the superscript D.

Proposition 1 *The price game has the following Nash equilibrium properties.*

1. *For $\alpha \in [0, \frac{1}{2})$, there is a unique interior equilibrium where $\lambda_C^D > 0$,*

$$P_C^D = -2\frac{1-\alpha}{1-2\alpha}V_C(\lambda_C^D) \cdot Q_C'(\lambda_C^D) = 2\frac{1-\alpha}{1-2\alpha}V_C(\lambda_C^D) \cdot d_C \sum_{n=1}^{2} \frac{1}{\left(\mu_n - \lambda_C^D\right)^2}, \tag{6}$$

and both firms set the same price: $p_{C1}^D = p_{C2}^D$.

2. For $\alpha \in [\frac{1}{2}, 1)$, there is no traffic in equilibrium: $\lambda_C^D = 0$, corresponding to infinite prices.

Performance vs. Centralized Control. If the network is under central control of a monopoly provider, the profit-maximizing demand rate λ_C^M and the corresponding price P_C^M satisfy

$$P_C^M = -V_C(\lambda_C^M) \cdot Q_C'(\lambda_C^M) = V_C(\lambda_C^M) \cdot d_C \sum_{n=1}^{2} \frac{1}{\left(\mu_n - \lambda_C^M\right)^2}.$$

The monopoly price equals the marginal *net value externality*, i.e., the expected delay-induced net value loss inflicted on the system by an infinitesimal demand rate increase. The monopoly price and demand are socially optimal (Afèche and Mendelson 2004, Prop. 1, p. 873), i.e., they maximize the expected net value rate $NV(\lambda_C) := V_C(\lambda_C)Q_C(\lambda_C)$. Let $\Pi(\lambda_C) := \Pi_1(\lambda_C) + \Pi_2(\lambda_C) = \lambda_C \cdot P_C(\lambda_C)$ be the expected total provider profit rate and $CS(\lambda_C) := NV(\lambda_C) - \Pi(\lambda_C)$ the consumer surplus rate. Proposition 2 compares the performance under decentralized and centralized control.

Proposition 2 *The equilibria under decentralized control (D), centralized monopoly control (M), and net value maximization ($*$) compare as follows:*

1. *Total price: $P_C^* = P_C^M < P_C^D$.*
2. *Demand rate: $\lambda_C^* = \lambda_C^M > \lambda_C^D$.*
3. *Expected system net value rate: $NV^* = NV^M > NV^D$.*
4. *Expected total provider profit rate: $\Pi^* = \Pi^M > \Pi^D$.*
5. *Expected consumer surplus rate: $CS^* = CS^M > CS^D$.*

Proposition 2 shows that decentralized control results in suboptimal performance and an *under-congested system,* compared to the overall optimum. The higher price under decentralized control is due to the double-marginalization effect which occurs in complementary monopolies (Spengler Spengler (1950)), and it is compounded by delay costs. In choosing its price each firm ignores the incremental profit that a lower price and higher demand rate would generate for the other provider.

As shown in Corollary 1 the performance loss under decentralized control is quite sensitive to the service value distribution (the parameter α) and to the delay cost (the parameter d_C).

Corollary 1 *The suboptimality of decentralized control specified in Proposition 2 depends as follows on the value distribution and time sensitivity. If $\alpha \in [0, \frac{1}{2})$, the decentralized equilibrium is suboptimal if and only if customers are time-sensitive ($d_C > 0$), and the loss under decentralized control increases in d_C. If $\alpha \in (\frac{1}{2}, 1)$, the decentralized equilibrium is suboptimal for all $d_C \geq 0$.*

Intuitively, the higher the delay sensitivity parameter d_C and/or the smaller the elasticity (the higher α), the more lucrative a high price. For sufficiently low elasticity, i.e., for $\alpha \in (\frac{1}{2}, 1)$, the providers "outbid" each others' prices to the point where no customer gets served.

The above results extend to a tandem network with N nodes and providers. Since each provider only gets a fraction $1/N$ of the total price, the larger the number of providers, the higher the total price, and the lower total profits and the consumer surplus. A natural measure of the price distortion under N providers is the ratio of equilibrium price to marginal congestion cost, which satisfies

$$\frac{P_C^{DN}}{-V_C(\lambda_C^{DN})Q'(\lambda_C^{DN})} = N\frac{1-\alpha}{1-\alpha N}.$$

The superscript DN refers to decentralized control by N providers. For $N \geq 2$, the price exceeds the marginal congestion cost. The price increases and the marginal congestion cost decreases in N.

3.2 Cross-Traffic Service: Price-Capacity Game

Next consider joint price-capacity decisions. We assume linear capacity costs. The node-n capacity cost per unit time is $b_n\mu_n$ where $b_n > 0$. Providers make price and capacity decisions simultaneously and independently. We consider pure strategy Nash equilibria for this price-capacity game.

Decentralized Control. Given provider 2's price p_{C2} and capacity μ_2, let $\Pi_1(\lambda_C, \mu_1; p_{C2}, \mu_2)$ and $P_C(\lambda_C, \mu_1; \mu_2)$ denote, respectively, the profit of provider 1 and the total price. Provider 1 determines her price $p_{C1}(p_{C2}, \mu_2)$ and capacity $\mu_1(p_{C2}, \mu_2)$ by solving:

$$\begin{aligned} \max_{\lambda_C, \mu_1} \Pi_1(\lambda_C, \mu_1; p_{C2}, \mu_2) &= \lambda_C \cdot [P_C(\lambda_C, \mu_1; \mu_2) - p_{C2}] - b_1\mu_1 \\ s.t. \qquad & 0 \leq \lambda_C < \mu_n, \quad n = 1, 2. \end{aligned}$$

For simplicity we suppress the arguments of the functions V_C, P_C, and Q_C. Provider 1's first-order conditions for an interior solution ($\mu_1 > 0$) are

$$\frac{\partial \Pi_1}{\partial \lambda_C} = \lambda_C \frac{\partial P_C}{\partial \lambda_C} + P_C - p_{C2} = \lambda_C \left[V_C'' Q_C + V_C' \frac{\partial Q_C}{\partial \lambda_C} \right] + P_C - p_{C2} = 0, \tag{7}$$

$$\frac{\partial \Pi_1}{\partial \mu_1} = \lambda_C \frac{\partial P_C}{\partial \mu_1} - b_1 = \lambda_C V_C' \frac{\partial Q_C}{\partial \mu_1} - b_1 = 0. \tag{8}$$

Proposition 3 *The price game has the following Nash equilibrium properties.*

1. *For $\alpha \in (0, \frac{1}{2})$, there are at most two interior equilibria. For an interior equilibrium, the prices of both firms are unique, given by*

$$p_{C1}^D = p_{C2}^D = \frac{P_C^D}{2} = \frac{b_1 + b_2}{1 - 2\alpha}, \tag{9}$$

and the marginal rate of QoS substitution between the node capacities satisfies:

$$\frac{\frac{\partial Q_C}{\partial \mu_1}}{\frac{\partial Q_C}{\partial \mu_2}} = \left(\frac{\mu_2^D - \lambda_C^D}{\mu_1^D - \lambda_C^D}\right)^2 = \frac{b_1}{b_2}. \tag{10}$$

The lower-cost firm has a higher capacity and profit: if $b_1 < b_2$*, then* $\mu_1^D > \mu_2^D$ *and* $\Pi_1^D > \Pi_2^D$*.*

2. *For* $\alpha \in [\frac{1}{2}, 1)$*, neither firm invests in capacity:* $\boldsymbol{\mu}^D = \lambda_C^D = 0$*.*

By (9), in equilibrium each firm's marginal revenue under constant QoS, $(\frac{1}{2} - \alpha) \cdot P_C^D$, equals the total marginal capacity cost $b_1 + b_2$. The intuition follows from firm 1's optimality condition in response to firm 2 capacity μ_2 and price $p_{C2}^D = \frac{P_C^D}{2}$: From (7)-(8),

$$(\frac{1}{2} - \alpha) \cdot P_C^D = b_1 \left(1 + \frac{\frac{\partial Q_C}{\partial \mu_2}}{\frac{\partial Q_C}{\partial \mu_1}}\right) = b_1 \cdot \left.\frac{d\mu_1}{d\lambda_C}\right|_{Q_C=\text{constant}}. \tag{11}$$

The RHS is the *induced marginal capacity cost* of provider 1, i.e., the marginal capacity cost of keeping the QoS constant (for fixed μ_2) as λ_C increases. By (11), keeping the QoS constant for fixed μ_2 requires provider 1 to increase μ_1 by more than λ_C, by an amount that equals the marginal rate of QoS substitution of node 1 for node 2 capacity, $\frac{\partial Q_C}{\partial \mu_2} \Big/ \frac{\partial Q_C}{\partial \mu_1}$. In equilibrium, this marginal substitution rate equals the capacity cost ratio b_2/b_1 by (10), so the induced marginal capacity cost of provider 1 equals the sum of marginal capacity cost $b_1 + b_2$.

Proposition 3 shows that the lower-cost provider invests in more capacity and earns a higher profit. Since both providers earn the same revenue in equilibrium, this means that the lower-cost provider has the lower total capacity cost, which follows because the marginal return of capacity (through a higher QoS) decreases in capacity.

Since the providers offer complementary services, either both invest in capacity or neither does. For $\alpha \in [\frac{1}{2}, 1)$ firms do not invest since they earn no revenue at any capacity level (Proposition 1).

While the equilibrium price P_C^D is unique, the equilibrium demand rate λ_C^D and capacities $\boldsymbol{\mu}^D$ need not be unique. There are at most two equilibria. Consider the symmetric case $b_1 = b_2 = b$. For small enough b, there is a unique interior equilibrium: If one provider restricts capacity, the other has an incentive to increase capacity since doing so is cheap. For sufficiently large b, there is no interior equilibrium since capacity is too costly for both firms. For intermediate values of b there may be two interior equilibria. In that case, b is high enough for a provider to refrain from expanding capacity if the other provider restricts hers, yielding a "small network"; and low enough to entice a provider to add capacity if the other firm also invests more, yielding a "large network" with higher demand rate, capacities and profits.

Performance vs. Centralized Control. Proposition 4 summarizes this comparison.

Proposition 4 *If the equilibria under decentralized control (D), centralized monopoly control (M), and centralized net value maximization (∗) are interior, then:*

1. *Total price:* $P_C^* = b_1 + b_2 < P_C^M = \frac{b_1+b_2}{1-\alpha} < P_C^D = 2\frac{b_1+b_2}{1-2\alpha}$.
2. *Capacities:* $\boldsymbol{\mu}^* > \boldsymbol{\mu}^M > \boldsymbol{\mu}^D$.
3. *Demand rate:* $\lambda_C^* > \lambda_C^M > \lambda_C^D$.
4. *Quality of Service:* $Q_C^* > Q_C^M > Q_C^D$.
5. *Expected system net value rate:* $NV^* > NV^M > NV^D$.
6. *Expected total provider profit rate:* $\Pi^* < 0 < \Pi^D < \Pi^D$.
7. *Expected consumer surplus rate:* $CS^* > CS^M > CS^D$.

Decentralized control results in a higher price and a lower capacity, demand rate, QoS, total profits, and consumer surplus, compared to the case where the network is controlled by a monopoly. These performance losses follow from double-marginalization, and they are increasing in delay costs.

3.3 Local- and Cross-Traffic Service: Pricing for Fixed Capacity

We now turn to the case where the network offers a mix of standalone and bundled services. We study how the delivery of this service mix affects the decentralized control equilibrium and the system inefficiency observed in §§3.1-3.2. Specifically, we introduce in addition to the cross-traffic also *local-traffic* service for each node. This yields three customer segments with distinct routes. The segments may also differ in their delay costs and value distributions. Demand for type-Ln local-traffic service ($n = 1, 2$) only requires resource n. Let Λ_{Ln} denote its market size. Let $\boldsymbol{\lambda} = (\lambda_{L1}, \lambda_{L2}, \lambda_C)$ be the demand rate vector and $\gamma_n := \lambda_{Ln} + \lambda_C$ the aggregate flow rate through node n. The type-Ln marginal value function is isoelastic, given by $V'_{Ln}(\lambda_{Ln}) = A_{Ln} \cdot \lambda_{Ln}^{-\alpha}$ for $\lambda_{Ln} \in [0, \Lambda_{Ln}]$. In §5 we discuss the effect of different elasticities for local- vs. cross-traffic on our results. Its delay sensitivity parameter is d_{Ln}. The expected QoS factors satisfy

$$Q_{Ln}(\boldsymbol{\lambda},\mu_n) = 1 - d_{Ln}\frac{1}{\mu_n - \gamma_n}, \quad n = 1, 2, \tag{12}$$

$$Q_C(\boldsymbol{\lambda}, \boldsymbol{\mu}) = 1 - d_C \sum_{n=1}^{2} \frac{1}{\mu_n - \gamma_n}. \tag{13}$$

The analysis focuses on capacity levels $\boldsymbol{\mu}$ that satisfy the following technical assumptions.

A3. $Q_C(\mathbf{0}, \boldsymbol{\mu}) > 0$, and $Q_C(\boldsymbol{\lambda}, \boldsymbol{\mu}) < 0$ as $\lambda_C \to \min(\Lambda_C, \mu_1 - \lambda_{L1}, \mu_2 - \lambda_{L2})$.

A4. $Q_{Ln}(\mathbf{0}, \mu_n) > 0$, and $Q_{Ln}(\boldsymbol{\lambda},\mu_n) < 0$ as $\lambda_{Ln} \to \min(\Lambda_{Ln}, \mu_n - \lambda_C)$ for $n = 1, 2$.

Assumptions *A3-A4* generalize *A1-A2* of §2. They ensure that each segment is profitable when no other segment is served, and that it is not optimal to serve all customers of any segment.

Each provider controls one node. Provider n charges P_{Ln} for type-Ln service and p_{Cn} for its part of the type-C service bundle. The total type-C price is $P_C = p_{C1} + p_{C2}$. (See Figure 1 for illustration.) Each provider has sole control of local-traffic pricing, but the QoS of local- and cross-traffic flows are interdependent since they compete for capacity. The demand equations are

$$P_{Ln}(\boldsymbol{\lambda}) = V'_{Ln}(\lambda_{Ln})Q_{Ln}(\boldsymbol{\lambda}) , \qquad n = 1, 2, \tag{14}$$

$$p_{C1} + p_{C2} = P_C(\boldsymbol{\lambda}) = V'_C(\lambda_C)Q_C(\boldsymbol{\lambda}). \tag{15}$$

First consider the price game for fixed capacity where providers set their prices simultaneously and independently. We focus on pure strategy Nash price equilibria. Given provider 2's prices P_{L2} and p_{C2}, provider 1 determines her prices $P_{L1}(P_{L2}, p_{C2})$ and $p_{C1}(P_{L2}, p_{C2})$ by solving:

$$\max_{\boldsymbol{\lambda}} \Pi_1(\boldsymbol{\lambda}; P_{L2}, p_{C2}) = \lambda_C \left[P_C(\boldsymbol{\lambda}) - p_{C2}\right] + \lambda_{L1} P_{L1}(\boldsymbol{\lambda}) \tag{16}$$

$$s.t.$$

$$P_{L2} = V'_{L2}(\lambda_{L2})Q_{L2}(\boldsymbol{\lambda}), \tag{17}$$

$$0 \leq \boldsymbol{\lambda}, \tag{18}$$

$$\lambda_{Ln} + \lambda_C < \mu_n, \quad n = 1, 2. \tag{19}$$

The constraint (17) requires that the demand rates be consistent with provider 2's local price. The constraints (19) do not bind at the maximum by assumptions *A3-A4*.

Decentralized Control. Let δ_1 and $\boldsymbol{\theta}_1 := (\theta_{11}, \theta_{12}, \theta_{1C})$ be the Lagrange multipliers of the price constraint (17) and the non-negativity constraints (18), respectively. Let

$$\mathcal{L}_1(\boldsymbol{\lambda}; P_{L2}, p_{C2}) := \Pi_1(\boldsymbol{\lambda}; P_{L2}, p_{C2}) + \delta_1 \cdot [V'_{L2}(\lambda_{L2})Q_{L2}(\boldsymbol{\lambda}) - P_{L2}] + \boldsymbol{\theta}'_1 \boldsymbol{\lambda} \tag{20}$$

be firm 1's Lagrangian. For simplicity we occasionally suppress the arguments of the price, marginal value and QoS functions. The Kuhn-Tucker conditions for a solution $\boldsymbol{\lambda}(P_{L2}, p_{C2})$ are (17)-(19) and

$$\frac{\partial \mathcal{L}_1}{\partial \lambda_{L1}} = \lambda_C \frac{\partial P_C}{\partial \lambda_{L1}} + P_{L1} + \lambda_{L1} \frac{\partial P_{L1}}{\partial \lambda_{L1}} + \theta_{11} = 0, \tag{21}$$

$$\frac{\partial \mathcal{L}_1}{\partial \lambda_{L2}} = \lambda_C \frac{\partial P_C}{\partial \lambda_{L2}} + \delta_1 \frac{\partial P_{L2}}{\partial \lambda_{L2}} + \theta_{12} = 0, \tag{22}$$

$$\frac{\partial \mathcal{L}_1}{\partial \lambda_C} = \lambda_C \frac{\partial P_C}{\partial \lambda_C} + P_C - p_{C2} + \lambda_{L1} \frac{\partial P_{L1}}{\partial \lambda_C} + \delta_1 \frac{\partial P_{L2}}{\partial \lambda_C} + \theta_{1C} = 0, \tag{23}$$

$$\boldsymbol{\theta}_1 \geq 0; \quad \boldsymbol{\theta}'_1 \boldsymbol{\lambda} = 0. \tag{24}$$

To economize on notation, let $MC_{Ln}(\boldsymbol{\lambda}) := -\lambda_{Ln} V'_{Ln}(\lambda_{Ln}) \frac{\partial Q_{Ln}}{\partial \gamma_n}$ denote the *marginal local-traffic node-n revenue externality*. It measures the delay-induced local-traffic revenue loss resulting from an infinitesimal flow rate increase at node-n. Similarly, let $MC_{Cn}(\boldsymbol{\lambda}) := -\lambda_C V'_C(\lambda_C) \frac{\partial Q_C}{\partial \gamma_n}$ be the marginal cross-traffic node-n revenue externality. Write $V_i^D = V_i(\lambda_i^D)$, $i \in \{L1, L2, C\}$, for the aggregate value rates in equilibrium and let $\Phi(\cdot)$ denote the indicator function.

Proposition 5 *The price game has the following Nash equilibrium properties.*

1. *For $\alpha \in [\frac{1}{2}, 1)$, there is no cross-traffic, $\lambda_C^D = 0$, corresponding to infinite prices p_{C1}^D and p_{C2}^D. The local-traffic demand rates and prices are unique and satisfy $\lambda_{Ln}^D > 0$ and*

$$P_{Ln}^D = \frac{1}{1-\alpha} MC_{Ln}(\boldsymbol{\lambda}^D) = \frac{V_{Ln}^D d_{Ln}}{\left(\mu_n - \lambda_{Ln}^D\right)^2}, \quad n = 1, 2. \tag{25}$$

2. *For $\alpha \in [0, \frac{1}{2})$, the prices and demand rates for an interior equilibrium $\boldsymbol{\lambda}^D > 0$ satisfy*

$$P_{Ln}^D = \frac{1}{1-\alpha} \left[MC_{Ln}(\boldsymbol{\lambda}^D) + MC_{Cn}(\boldsymbol{\lambda}^D) \right] = \frac{V_{Ln}^D d_{Ln} + V_C^D d_C}{\left(\mu_n - \gamma_n^D\right)^2}, \quad n = 1, 2. \tag{26}$$

$$P_C^D = \frac{1-\alpha}{1-2\alpha} \sum_{n=1}^{2} P_{Ln}^D \frac{2\alpha V_C^D d_C + V_{Ln}^D d_{Ln}}{\alpha V_C^D d_C + V_{Ln}^D d_{Ln}}. \tag{27}$$

$$p_{Cn}^D = \frac{1-\alpha}{1-2\alpha} \sum_{m=1}^{2} P_{Lm}^D \frac{\alpha V_C^D d_C + \left[\alpha + \Phi\{m=n\}(1-2\alpha)\right] V_{Lm}^D d_{Lm}}{\alpha V_C^D d_C + V_{Lm}^D d_{Lm}}, \quad n = 1, 2. \tag{28}$$

The drivers of these equilibrium prices are as follows. First, as in the monopoly case, by (25)-(26) the prices for local-traffic equal its marginal net value externalities, since a single provider sets each local-traffic price. However, the net value externalities are evaluated at the decentralized equilibrium demand rates $\boldsymbol{\lambda}^D$ which may differ from the monopoly demand rates $\boldsymbol{\lambda}^M$.

Second, consider the intuition for equilibrium cross-traffic prices p_{C1}^D, p_{C2}^D, and P_C^D. By (22) for an interior solution the Lagrange multiplier of the price constraint (17) satisfies

$$-\delta_1 = \frac{\partial \Pi_1}{\partial P_{L2}} = \lambda_C \frac{\frac{\partial P_c}{\partial \lambda_{L2}}}{\frac{\partial P_{L2}}{\partial \lambda_{L2}}} > 0, \tag{29}$$

indicating that provider 1's maximum profit increases in P_{L2}. The higher the other firm's local-traffic price, the lower the corresponding demand rate λ_{L2} at any fixed cross-traffic rate λ_C, and the higher the node-2 QoS, yielding a higher cross-traffic price P_C. Substituting for δ_1 into (23) and noting that $\partial P_c(\boldsymbol{\lambda}) / \partial \lambda_{L2} = MC_{C2}(\boldsymbol{\lambda})$ yields the cross-traffic price equilibrium equation

$$(1-\alpha) P_C - p_{C2} = MC_{L1}(\boldsymbol{\lambda}) + MC_{C1}(\boldsymbol{\lambda}) + MC_{C2}(\boldsymbol{\lambda}) \left(1 + \frac{\partial \lambda_{L2}}{\partial \lambda_C}\right). \tag{30}$$

Provider 1 equates her revenue from the marginal cross-traffic customer under constant QoS, the LHS of (30), to the *marginal revenue externality* it inflicts, the RHS of (30), which sums the components of this externality by node and customer type: $MC_{L1}(\boldsymbol{\lambda})$ and $MC_{C1}(\boldsymbol{\lambda})$ are the delay-induced revenue losses inflicted on local- and cross-traffic at node 1, and $MC_{C2}(\boldsymbol{\lambda})\left(1+\frac{\partial\lambda_{L2}}{\partial\lambda_C}\right)$ is the corresponding cross-traffic revenue loss at node 2. The multiplier $\left(1+\frac{\partial\lambda_{L2}}{\partial\lambda_C}\right)$ measures the *net* change in the aggregate node-2 demand rate as λ_C increases, where

$$\frac{\partial\lambda_{Ln}}{\partial\lambda_C}=-\frac{\frac{\partial P_{Ln}}{\partial\lambda_C}}{\frac{\partial P_{Ln}}{\partial\lambda_{Ln}}}=-\frac{(1-\alpha)d_{Ln}V_{Ln}(\lambda_{Ln})}{d_{Ln}V_{Ln}(\lambda_{Ln})+\alpha d_C V_C(\lambda_C)}\in[-1,0]\,,\quad n=1,2, \tag{31}$$

is the node-n *local-to-cross-traffic substitution rate* derived from the price constraint (17). This substitution effect measures the reduction in node-n local-traffic induced by a cross-traffic increase for fixed price P_{Ln}. By (30) provider 1's cross-traffic price response satisfies

$$p_{C1}=\frac{\alpha}{1-\alpha}p_{C2}+\frac{1}{1-\alpha}\left[MC_{L1}(\boldsymbol{\lambda})+MC_{C1}(\boldsymbol{\lambda})+MC_{C2}(\boldsymbol{\lambda})\left(1+\frac{\partial\lambda_{L2}}{\partial\lambda_C}\right)\right]. \tag{32}$$

For sufficiently inelastic marginal value function, i.e., $\epsilon\leq 2$ or $\alpha\in[\frac{1}{2},1)$, p_{C1} *exceeds* p_{C2} (the bracketed term is strictly positive), so there is no cross-traffic in equilibrium. For sufficiently elastic marginal value function, i.e., $\epsilon>2$ or $\alpha\in[0,\frac{1}{2})$, summing (30) and its counterpart for provider 2 yields the following cross-traffic price equilibrium equation, which is equivalent to (27):

$$2(1-\alpha)P_C-(p_{C1}+p_{C2})=\sum_{n=1}^{2}MC_{Ln}(\boldsymbol{\lambda})+MC_{Cn}(\boldsymbol{\lambda})\left(2+\frac{\partial\lambda_{Ln}}{\partial\lambda_C}\right). \tag{33}$$

Next, consider the providers' individual cross-traffic prices p^D_{C1} and p^D_{C2}. In equilibrium each provider's price must equal her marginal cross-traffic profit. It follows from (30) that

$$p_{Cm}=(1-\alpha)P_C-\left[MC_{Ln}(\boldsymbol{\lambda})+MC_{Cn}(\boldsymbol{\lambda})+MC_{Cm}(\boldsymbol{\lambda})\left(1+\frac{\partial\lambda_{Lm}}{\partial\lambda_C}\right)\right],\quad m\neq n. \tag{34}$$

The providers' equilibrium cross-traffic prices are therefore equal if, and only if, their marginal revenue externalities are equal. In general their difference is

$$p^D_{C1}-p^D_{C2}=MC_{L1}(\boldsymbol{\lambda}^D)\,\frac{V^D_{L1}d_{L1}+V^D_C d_C}{V^D_{L1}d_{L1}+\alpha V^D_C d_C}-MC_{L2}(\boldsymbol{\lambda}^D)\,\frac{V^D_{L2}d_{L2}+V^D_C d_C}{V^D_{L2}d_{L2}+\alpha V^D_C d_C}. \tag{35}$$

Loosely speaking, the price of cross-traffic is higher at the node where it inflicts a higher congestion-induced revenue loss. For example, in the absence of local-traffic, as in §§3.1-3.2, or if the local-traffic markets are symmetric, both providers charge the same cross-traffic price. However, if only one of the local-traffic flows is time-sensitive, e.g., $d_{L1}>0=d_{L2}$, then the equilibrium cross-traffic price is higher at that node, i.e., $p^D_{C1}>p^D_{C2}$, because $MC_{L1}(\boldsymbol{\lambda}^D)>MC_{L2}(\boldsymbol{\lambda}^D)=0$.

Performance vs. Centralized Control. Consider the monopoly price equations:

$$(1-\alpha)P_{Ln}^M = MC_{Ln}(\boldsymbol{\lambda}^M) + MC_{Cn}(\boldsymbol{\lambda}^M),\ n=1,2, \tag{36}$$

$$(1-\alpha)P_C^M = (1-\alpha)\left[P_{L1}^M + P_{L2}^M\right] = \sum_{n=1}^{2} MC_{Ln}(\boldsymbol{\lambda}^M) + MC_{Cn}(\boldsymbol{\lambda}^M). \tag{37}$$

The monopoly prices have two properties: (1) Each price equals its marginal net value externality, evaluated at the monopoly rates $\boldsymbol{\lambda}^M$. (2) The cross-traffic price P_C^M equals the sum of local charges $P_{L1}^M + P_{L2}^M$, since the marginal congestion costs are additive in network nodes. Inspection of (25)-(28) shows that the prices under decentralized control *may* but need not share these properties: (1) The prices of local-traffic flows equal their marginal net value externalities, evaluated at the equilibrium rates $\boldsymbol{\lambda}^D$, but the total cross-traffic price P_C^D typically does not. (2) The cross-traffic price P_C^D may exceed or equal the sum of local charges $P_{L1}^D + P_{L2}^D$.

To pinpoint how the control structure affects the cross-traffic prices, compare the RHS of (33) and (37). They capture for decentralized and monopoly control, respectively, the marginal revenue loss as cross-traffic increases. For given $\boldsymbol{\lambda}$ the marginal local-traffic revenue externalities $MC_{Ln}(\boldsymbol{\lambda})$ are equal under both control structures, since local-traffic is priced by a single provider. However, the marginal cross-traffic revenue externalities need not be equal. For given $\boldsymbol{\lambda}$ the marginal cross-traffic revenue externality equals $\sum_{n=1}^{2} MC_{Cn}(\boldsymbol{\lambda})$ for a monopoly and $\sum_{n=1}^{2} MC_{Cn}(\boldsymbol{\lambda})\left(2+\frac{\partial\lambda_{Ln}}{\partial\lambda_C}\right)$ under decentralized control. This difference is due to two opposite externality effects under decentralized control. First, the positive *double-marginalization* effect discussed in §§3.1-3.2 (captured by the factor 2) pushes decentralized providers to charge more and reduce cross-traffic, compared to centralized control. Second, a negative *substitution effect between one provider's cross-traffic revenue and the other's local-traffic revenue* (captured by $\frac{\partial\lambda_{Ln}}{\partial\lambda_C}, n=1,2$) gives each provider the incentive to lower its cross-traffic price, which partially or completely *offsets the double-marginalization effect*. Given provider 2's local-traffic price P_{L2}, consider provider 1's incentive to increase the cross-traffic rate λ_C by lowering her cross-traffic price p_{C1}. Increasing λ_C causes λ_{L2} to drop to maintain the price equilibrium $P_{L2} = V_{L2}'(\lambda_{L2})Q_{L2}(\boldsymbol{\lambda})$. This counterbalancing reduction in λ_{L2} increases the cross-traffic QoS, $Q_C(\boldsymbol{\lambda})$, making cross-traffic more profitable and enticing provider 1 to attract more cross-traffic than would be profitable without such substitution. Specifically, if there is no time-sensitive local-traffic (i.e., $d_{Ln}V_{Ln}(\lambda_{Ln}) = 0$) then $\frac{\partial\lambda_{Ln}}{\partial\lambda_C} = 0$ by (31) and the marginal cross-traffic revenue externalities under decentralized control are exactly *twice* as large as under a monopoly. Time-sensitive local-traffic reduces the difference since $\frac{\partial\lambda_{Ln}}{\partial\lambda_C} < 0$ if $d_{Ln}V_{Ln}(\lambda_{Ln}) > 0$.

The net effect of double-marginalization and substitution on the performance under decentralized control depends on the delay sensitivity parameters and the elasticity parameter. Since the

monopoly cross-traffic price equals the sum of local-traffic prices, the corresponding price ratio for decentralized control is a natural measure of system inefficiency. By (27)

$$\frac{P_C^M}{P_{L1}^M + P_{L2}^M} = 1 \leq \frac{P_C^D}{P_{L1}^D + P_{L2}^D} = \frac{1-\alpha}{1-2\alpha} \frac{\sum_{n=1}^{2} P_{Ln}^D \frac{2\alpha V_C^D d_C + V_{Ln}^D d_{Ln}}{\alpha V_C^D d_C + V_{Ln}^D d_{Ln}}}{P_{L1}^D + P_{L2}^D} \leq 2\frac{1-\alpha}{1-2\alpha}. \tag{38}$$

Corollary 2 summarizes the effect of time sensitivity on this price ratio.

Corollary 2 *For fixed $\alpha \in [0,1)$ the price ratio and substitution effects are bounded as follows:*

1. *If only the local-traffic is time-sensitive ($d_{Ln} > 0 = d_C, n = 1,2$):*

   (a) *The local-to-cross-traffic substitution is maximized: $\frac{\partial \lambda_{Ln}}{\partial \lambda_C} = \alpha - 1 < 0$, $n = 1,2$.*

   (b) *The cross-to-local-traffic price ratio attains its lower bound:*

$$\frac{P_C^D}{P_{L1}^D + P_{L2}^D} = \frac{1-\alpha}{1-2\alpha}. \tag{39}$$

2. *If only the cross-traffic is time-sensitive ($d_C > 0 = d_{L1} = d_{L2}$):*

   (a) *There is no local-to-cross-traffic substitution: $\frac{\partial \lambda_{Ln}}{\partial \lambda_C} = 0$, $n = 1,2$.*

   (b) *The cross-to-local-traffic price ratio attains its upper bound:*

$$\frac{P_C^D}{P_{L1}^D + P_{L2}^D} = 2\frac{1-\alpha}{1-2\alpha}. \tag{40}$$

By Corollary 2 the performance loss under decentralized control is minimized (maximized) when only the local-traffic (cross-traffic) is time-sensitive. In general the local-to-cross-traffic substitution effect benefits the overall system performance. Specifically, the local-traffic benefits from the lower than optimal cross-traffic demand under decentralized control since this frees up network capacity and in turn yields lower prices and higher demand rates, revenues and consumer surplus for local-traffic. In this sense, the cross-traffic effectively "subsidizes" the local-traffic.

In the case of infinitely elastic marginal value functions (i.e., $\alpha = 0$) the presence of time-sensitive local-traffic (i.e., $A_{Ln} d_{Ln} > 0$ for $n = 1,2$) has a particularly favorable effect on the performance of decentralized control: The local-to-cross-traffic substitution is perfect, i.e., $\frac{\partial \lambda_{Ln}}{\partial \lambda_C} = -1$, and cancels out the double-marginalization effect, so the equilibria under decentralized and monopoly control are identical. By (39) the total cross-traffic price equals the sum of local-traffic prices. In the absence of time-sensitive local-traffic the equilibrium under decentralized is suboptimal; by (40) the total cross-traffic price equals double the sum of local-traffic prices.

Example 1. We illustrate these performance comparisons numerically for a symmetric network with capacities $\mu_1 = \mu_2 = 100$. The marginal value functions have value intensities $(A_{L1}, A_{L2}, A_C) =$

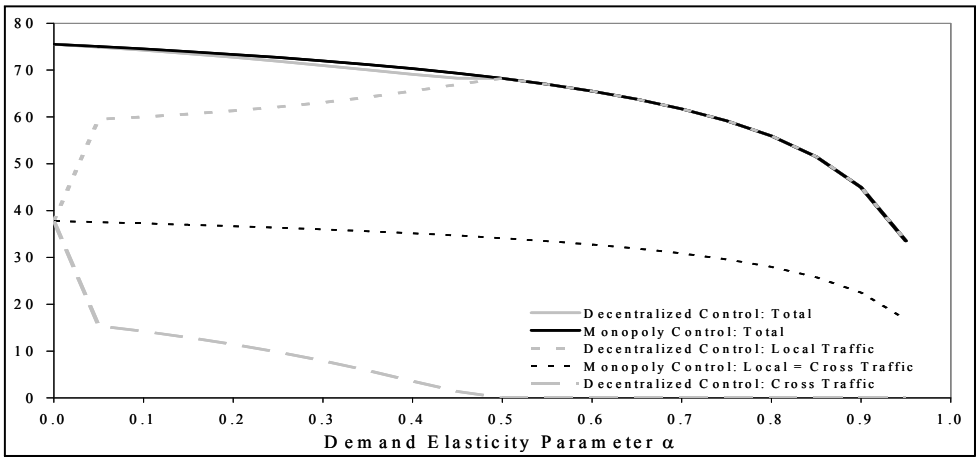

Figure 2: Symmetric example network: $\mu_1 = \mu_2 = 100$, $(A_{L1}, A_{L2}, A_C) = (5, 5, 10)$ and $(d_{L1}, d_{L2}, d_C) = (6, 6, 3)$. Equilibrium demand rates at each node under decentralized and monopoly control.

$(5, 5, 10)$, and the delay sensitivity parameters are $(d_{L1}, d_{L2}, d_C) = (6, 6, 3)$. The QoS is the same for all segments: cross-traffic is half as time-sensitive as local-traffic but expects double the delay. Figure 2 shows the equilibrium demand rates under decentralized and monopoly control for $\alpha \in [0, 0.95]$. For $\alpha = 0$, the decentralized equilibrium is efficient and the rates equal those of a monopoly. As α increases the double-marginalization effect intensifies and the substitution effect weakens, leading to a significant drop in cross-traffic and increase in local-traffic under decentralized control, compared to optimal levels. For $\alpha \geq 0.5$, there is no cross-traffic in equilibrium (Proposition 5). Figure 3 shows the ratios of total revenues under decentralized control and monopoly for the cases with and without local-traffic. For $\alpha = 0$ there is no revenue loss under decentralized control if, and only if, there is local-traffic on the network. For $\alpha > 0$, the revenue loss under decentralization is typically lower in the presence of local-traffic since the revenue loss on cross-traffic is partially offset by gains on local-traffic. By Figure 3 the beneficial effect of local-traffic is relatively significant, except for α-values around $\alpha \in [0.25, 0.4]$. In this range, both the double-marginalization effect and the substitution effect are relatively small, whereas for lower (higher) α the effect of substitution (double-marginalization) is significant. The performance gains due to local-traffic can be much larger than in this example, whenever local-traffic is more delay-sensitive relative to cross-traffic.

To summarize, the presence of time-sensitive local-traffic typically benefits performance: 1) The resulting substitution effect offsets or even cancels the detrimental effect of double-marginalization on cross-traffic. 2) If the cross-traffic price is distorted, the reduced delay benefits the local-traffic.

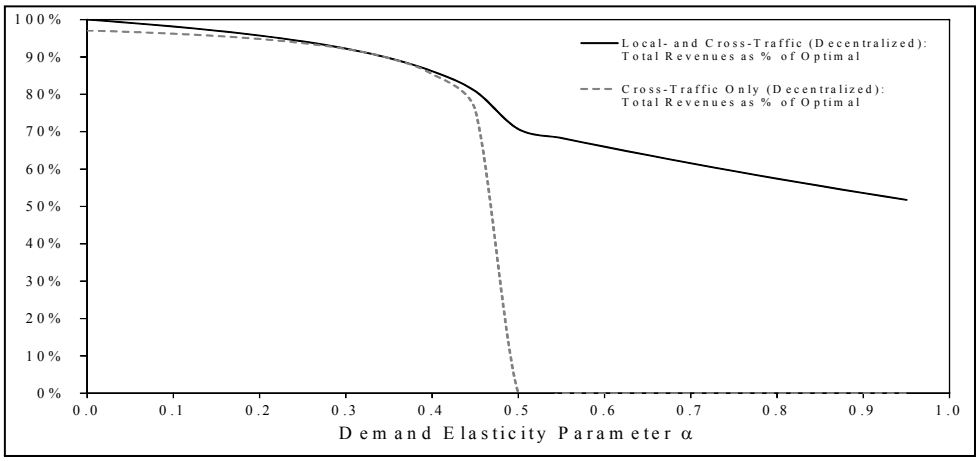

Figure 3: Symmetric example network: $\mu_1 = \mu_2 = 100$, $(A_{L1}, A_{L2}, A_C) = (5, 5, 10)$ and $(d_{L1}, d_{L2}, d_C) = (6, 6, 3)$. Impact of local-traffic on total revenues under decentralized control as % of optimal (monopoly).

3.4 Local- and Cross-Traffic Service: Price-Capacity Game

Now consider joint price and capacity decisions. Providers set their prices for local- and cross-traffic and their capacities simultaneously and independently. As in §3.2 capacity costs are linear. Given provider 2's prices $P_{L2},\ p_{C2}$ and capacity $\mu_2 > 0$, provider 1 determines her prices $P_{L1}(P_{L2}, p_{C2}, \mu_2)$, $p_{C1}(P_{L2}, p_{C2}, \mu_2)$ and capacity $\mu_1(P_{L2}, p_{C2}, \mu_2)$ by solving:

$$\max_{\boldsymbol{\lambda},\ \mu_1} \Pi_1(\boldsymbol{\lambda}, \mu_1; P_{L2}, p_{C2}, \mu_2) = \lambda_C \left[P_C(\boldsymbol{\lambda}, \mu_1; \mu_2) - p_{C2}\right] + \lambda_{L1} \cdot P_{L1}(\boldsymbol{\lambda}, \mu_1) - b_1\mu_1, \tag{41}$$

$$s.t.$$

$$P_{L2} = V'_{L2}(\lambda_{L2}) Q_{L2}(\boldsymbol{\lambda}, \mu_2), \tag{42}$$

$$0 \leq \boldsymbol{\lambda}, \tag{43}$$

$$\lambda_{Ln} + \lambda_C < \mu_n, \quad n = 1, 2. \tag{44}$$

Decentralized Control. We focus on equilibria with positive capacities, so the constraints (44) are slack by *A3-A4*. Let $\mathcal{L}_1(\boldsymbol{\lambda}, \mu_1; P_{L2}, p_{C2}, \mu_2)$ be provider 1's Lagrangian. It is the same as (20), except that μ_1 is a decision variable. A solution $\{\boldsymbol{\lambda}(P_{L2}, p_{C2}, \mu_2), \mu_1(P_{L2}, p_{C2}, \mu_2)\}$ with $\boldsymbol{\mu} > 0$ satisfies the Kuhn-Tucker conditions (21)-(24), (42)-(44), and

$$\frac{\partial \mathcal{L}_1}{\partial \mu_1} = \lambda_C \frac{\partial P_C}{\partial \mu_1} + \lambda_{L1} \frac{\partial P_{L1}}{\partial \mu_1} - b_1 = 0. \tag{45}$$

Proposition 6 *The price-capacity game has the following Nash equilibrium properties.*

1. *For $\alpha \in [\frac{1}{2}, 1)$, there is no cross-traffic ($\lambda_C^D = 0$), corresponding to infinite prices p_{C1}^D and p_{C2}^D. If $\lambda_{Ln}^D > 0$, the local-traffic prices are unique and satisfy*

$$P_{Ln}^D = \frac{b_n}{1 - \alpha}, \quad n = 1, 2. \tag{46}$$

2. For $\alpha \in (0, \frac{1}{2})$, the local-traffic prices for an interior equilibrium $\left(\boldsymbol{\lambda}^D, \boldsymbol{\mu}^D\right) > 0$ are unique:

$$P_{Ln}^D = \frac{b_n}{1-\alpha}, \quad n = 1, 2, \tag{47}$$

and the cross-traffic prices and demand rates satisfy

$$P_C^D = \frac{1}{1-2\alpha} \sum_{n=1}^{2} b_n \frac{2\alpha V_C^D d_C + V_{Ln}^D d_{Ln}}{\alpha V_C^D d_C + V_{Ln}^D d_{Ln}} \tag{48}$$

$$p_{Cn}^D = \frac{1}{1-2\alpha} \sum_{m=1}^{2} b_m \frac{\alpha V_C^D d_C + \left[\alpha + \Phi\{m=n\}\left(1-2\alpha\right)\right] V_{Lm}^D d_{Lm}}{\alpha V_C^D d_C + V_{Lm}^D d_{Lm}}, \quad n = 1, 2. \tag{49}$$

At the equilibrium prices and capacities, the marginal revenue local-traffic externality at a node equal that node's marginal capacity cost. This follows from two facts. First, as shown by (25)-(26), at any fixed capacity level the type-Ln local-traffic equilibrium price is such that its marginal revenue under constant QoS, $(1-\alpha)P_{Ln}^D$, equals the marginal revenue externality it inflicts on node-n traffic, $MC_{Ln}(\boldsymbol{\lambda}^D) + MC_{Cn}(\boldsymbol{\lambda}^D)$. Second, by inspection of (12)-(13), the expected QoS of services that require node-n remain constant under equal increases in the capacity μ_n and in the aggregate flow rate γ_n. The marginal revenue of node-n capacity, $\left[\lambda_{Ln} V_{Ln}'\left(\lambda_{Ln}\right) + \lambda_C V_C'\left(\lambda_C\right)\right] \frac{\partial Q_{Ln}}{\partial \mu_n}$, therefore exactly offsets the marginal revenue externality inflicted by type-Ln traffic, and in equilibrium both must equal the marginal cost b_n as shown by (46)-(47). The cross-traffic equilibrium prices (48)-(49) are immediate by substituting $b_n = P_{Ln}^D\left(1-\alpha\right)$ into the equilibrium price equations (27)-(28).

Performance vs. Centralized Control. The monopoly prices satisfy

$$P_{Ln}^M = \frac{b_n}{1-\alpha}, \quad n = 1, 2, \tag{50}$$

$$P_C^M = P_{L1}^M + P_{L2}^M. \tag{51}$$

Inspection of Proposition 6 shows that the local-traffic prices under decentralized and monopoly control are the *same*, since local-traffic is served by a single resource and firm. As discussed in §3.3 the monopoly prices are additive in the resources while the ratio of cross to local-traffic prices under decentralized control depends on double-marginalization and substitution effects. The price ratios, the counterparts of (38) for fixed capacity, satisfy

$$\frac{P_C^M}{P_{L1}^M + P_{L2}^M} = 1 \leq \frac{P_C^D}{P_{L1}^D + P_{L2}^D} = \frac{1-\alpha}{1-2\alpha} \frac{\sum_{n=1}^{2} b_n \frac{2\alpha V_C^D d_C + V_{Ln}^D d_{Ln}}{\alpha V_C^D d_C + V_{Ln}^D d_{Ln}}}{b_2 + b_2} \leq 2\frac{1-\alpha}{1-2\alpha}. \tag{52}$$

Corollary 3 *The prices under decentralized and monopoly control compare as follows.*

1. The local-traffic prices are equal.

$$P_{Ln}^D = P_{Ln}^M = \frac{b_n}{1-\alpha}, \quad n = 1, 2.$$

2. *If only the local-traffic is time-sensitive ($d_{Ln} > 0 = d_C, n = 1, 2$), then the cross-to-local-traffic price ratio under decentralized control attains its lower bound and satisfies:*

$$\frac{P_C^D}{P_C^M} = \frac{P_C^D}{P_{L1}^D + P_{L2}^D} = \frac{1-\alpha}{1-2\alpha}.$$

3. *If only the cross-traffic is time-sensitive ($d_C > 0 = d_{L1} = d_{L2}$), then the cross-to-local-traffic price ratio under decentralized control attains its upper bound and satisfies:*

$$\frac{P_C^D}{P_C^M} = \frac{P_C^D}{P_{L1}^D + P_{L2}^D} = 2\frac{1-\alpha}{1-2\alpha}.$$

4. *If there is time-sensitive local-traffic on the network ($A_{Ln} \cdot d_{Ln} > 0$ for $n = 1, 2$), then the equilibria under decentralized and monopoly control are identical as $\alpha \to 0$.*

Corollary 3 is immediate from Proposition 6 and (50)-(51). As discussed in §3.3, it shows that the presence of time-sensitive local-traffic benefits system performance under decentralized control.

4 Coordination Mechanisms

To summarize the results under decentralized control without coordination: (1) double-marginalization in setting the cross-traffic prices tends to reduce performance and yields an *undercongested* system, and (2) the local-to-cross-traffic substitution effect counteracts or completely cancels out the double-marginalization effect, thereby improving or restoring the optimal performance.

These insights raise a simple question: Which coordination mechanisms are effective in reducing/eliminating the potential performance loss under decentralized control? We study two mechanisms that are motivated by the interconnection practice among Internet service providers, and we discuss how their effectiveness depends on customers' time sensitivity: *Peering* contracts, the prevalent direct interconnection agreement, whereby networks of comparable size (with approximately balanced traffic flows) agree to exchange traffic for free; and *transfer prices,* whereby networks carry each other's traffic for a fee.[1] While transfer prices are typically only used among providers with traffic imbalance, we show that their use also increases the performance of a symmetric network.

We consider a network with two cross-traffic flows, indexed by $C1$ and $C2$. In the data network context, this captures the situation where type-Cn cross-traffic originates on provider n's network and terminates on that of provider $m \neq n$. (For simplicity we do not model local-traffic;

[1] Traditional peering contracts involve no transfer prices and are also referred to as "bill and keep" or "settlement-free". Transfer prices have traditionally only been part of *transit* agreements. More recently, they may also be part of "paid" or "settlement-based" peering contracts that are typically entered between networks with imbalanced traffic flows. In contrast to transit agreements, peering contracts obligate partners to carry traffic only to their own customers, not to those of third party networks. This distinction between transit and peering is immaterial in the context with two networks.

adding local-traffic complicates the analysis without substantively altering the main insights.) After providers agree on the terms for carrying each other's traffic, provider n independently determines the total cross-traffic price P_{Cn}. This model with two resources, two providers, and two cross-traffic flows captures the simplest structure for studying the coordination issues that are of primary interest here. First, by omitting local-traffic we eliminate the possibility of performance-improving substitution effects that may be realized in the absence of coordination mechanisms, as shown in §§3.3-3.4. Second, this model with two cross-traffic flows preserves the element of decentralized control in the study of inter-provider contracts such as peering and transfer prices, which eliminate double-marginalization by assigning each provider exclusive pricing control over a cross-traffic flow.

For simplicity we study a symmetric network with fixed capacity. For $n = 1, 2$, the marginal value functions, delay sensitivity parameters, and capacities satisfy $V'_{Cn}(\lambda_{Cn}) = A_C \cdot \lambda_{Cn}^{-\alpha}$, $\alpha \in [0, 1)$; $d_{Cn} = d_C > 0$; and $\mu_n = \mu$, respectively. Each provider carries both flows. Their QoS are the same:

$$Q_{C1}(\boldsymbol{\lambda}) = Q_{C2}(\boldsymbol{\lambda}) = 1 - 2d_C \frac{1}{\mu - \lambda_{C1} - \lambda_{C2}},$$

where $\boldsymbol{\lambda} = (\lambda_{C1}, \lambda_{C2})$. We further assume that $d_C < \mu/2$, so $Q_{Cn}(\mathbf{0}) > 0$ for $n = 1, 2$.

In §4.1 we show that peering agreements are suboptimal in equilibrium, except when customers are not time-sensitive. This inefficiency fundamentally differs in cause and effect from that under decentralized control: It is not due to double-marginalization and leads to *over- vs. undercongestion.* In §4.2 we study transfer pricing equilibria, identify the optimal transfer price that yields the centralized solution, and show that a range of transfers in proximity of the optimal transfer price yield better performance than either peering contracts or uncoordinated decentralized control.

4.1 Peering

Peering contracts are common among major Internet service providers. Providers who agree to peer charge their own subscribers directly and agree to carry or "terminate" each others' traffic for free. Providers typically only enter such an agreement if their traffic flows are balanced and their networks are "comparable" in terms of capacity and other measures of scale and reach. Providers are keen to avoid peering with networks that have less capacity or otherwise inferior capabilities, since the benefits from such partnerships are typically unfavorably one-sided. Instead they charge smaller partners for carrying their traffic. However, we show that when customers are time-sensitive, peering even "among equals" with balanced network traffic flows leads to suboptimal performance.

Peering Equilibrium. Provider n independently sets the type-Cn total price P_{Cn}, where

$$P_{Cn}(\boldsymbol{\lambda}) := V'_{Cn}(\lambda_{Cn})\, Q_{Cn}(\boldsymbol{\lambda}),\ n = 1, 2.$$

It is convenient to perform the analysis in terms of the demand rates. Providers make their decisions simultaneously and independently. Given λ_{Cm}, provider $n \neq m$ solves:

$$\begin{aligned} \max_{\lambda_{Cn} \geq 0} \Pi_n(\lambda_{Cn}; \lambda_{Cm}) &= \lambda_{Cn} \cdot P_{Cn}(\boldsymbol{\lambda}), \\ s.t. & \quad \lambda_{C1} + \lambda_{C2} < \mu. \end{aligned}$$

Define the *marginal* type-Cn *revenue externality* as

$$MC_n(\boldsymbol{\lambda}) := -\lambda_{Cn} V'_{Cn}(\lambda_{Cn}) \frac{\partial Q_{Cn}}{\partial \lambda_n} = -\lambda_{Cn} V'_{Cn}(\lambda_{Cn}) \frac{\partial Q_{Cn}}{\partial \lambda_m}, \quad n \neq m.$$

It measures the delay-induced revenue loss on type-Cn traffic that results from an infinitesimal increase in the rate of either cross-traffic flow. (Since both flows have the same route their impact is the same.) The first-order conditions for an interior equilibrium are

$$(1-\alpha)\, P_{Cn}(\boldsymbol{\lambda}) = MC_n(\boldsymbol{\lambda}) = \lambda_{Cn} V'_{Cn}(\lambda_{Cn}) \frac{2 d_C}{(\mu - \lambda_{C1} - \lambda_{C2})^2}, \quad n = 1, 2. \tag{53}$$

Provider n equates her marginal type-Cn cross-traffic revenue under constant QoS, the LHS of (53), *only* to the marginal revenue externality it inflicts *on its own revenue.* Ignoring the adverse effect of more congestion on the other provider's revenue leads to overcongestion as discussed below.

Proposition 7 *For $\alpha \in [0, 1)$, the symmetric peering game has a unique symmetric Nash equilibrium. Providers charge equal prices that satisfy*

$$P^P_{Cn} = P^P_C := \frac{MC_n(\boldsymbol{\lambda}^P)}{1-\alpha} = \frac{A_C \left(\lambda^P_C\right)^{1-\alpha}}{1-\alpha} \frac{2 d_C}{\left(\mu - 2\lambda^P_C\right)^2}, \quad n = 1, 2, \tag{54}$$

and the equilibrium cross-traffic demand rates $\boldsymbol{\lambda}^P$ are

$$\lambda^P_{C1} = \lambda^P_{C2} = \lambda^P_C := \frac{\mu}{2} + \frac{d_C}{4} \frac{2\alpha - 1}{1-\alpha} - \sqrt{\left(\frac{d_C}{4} \frac{2\alpha - 1}{1-\alpha}\right)^2 + \frac{d_C}{4} \frac{\mu}{1-\alpha}}. \tag{55}$$

Performance vs. Decentralized and Centralized Control. The results in §3.1 for a single cross-traffic flow imply the monopoly and decentralized equilibria for two cross-traffic flows.

Lemma 1 *For a symmetric network with two cross-traffic flows:*

1. *For $\alpha \in [0, 1)$, the monopoly prices are unique and satisfy*

$$P^M_{Cn} = P^M_C := \frac{\sum_{n=1}^2 MC_n(\boldsymbol{\lambda}^M)}{1-\alpha} = 2 \cdot \frac{A_C \left(\lambda^M_C\right)^{1-\alpha}}{1-\alpha} \frac{2 d_C}{\left(\mu - 2\lambda^M_C\right)^2}, \quad n = 1, 2, \tag{56}$$

where the cross-traffic demand rates $\boldsymbol{\lambda}^M$ are

$$\lambda^M_{C1} = \lambda^M_{C2} = \lambda^M_C := \frac{\mu}{2} + \frac{d_C}{2} \frac{\alpha}{1-\alpha} - \sqrt{\left(\frac{d_C}{2} \frac{\alpha}{1-\alpha}\right)^2 + \frac{d_C}{2} \frac{\mu}{1-\alpha}}. \tag{57}$$

2. There is a unique decentralized Nash price equilibrium. If $\alpha \in [0, \frac{1}{2})$, prices satisfy

$$P^D_{Cn} = P^D_C := 2 \cdot \frac{\sum_{n=1}^{2} MC_n(\boldsymbol{\lambda}^D)}{1-2\alpha} = 2\frac{1-\alpha}{1-2\alpha} \cdot 2 \cdot \frac{A_C \left(\lambda^D_C\right)^{1-\alpha}}{1-\alpha} \frac{2d_C}{\left(\mu - 2\lambda^D_C\right)^2}, \quad n = 1, 2, \tag{58}$$

where the cross-traffic demand rates $\boldsymbol{\lambda}^D$ are

$$\lambda^D_{C1} = \lambda^D_{C2} = \lambda^D_C := \frac{\mu}{2} + \frac{d_C}{2}\frac{1+2\alpha}{1-2\alpha} - \sqrt{\left(\frac{d_C}{2}\frac{1+2\alpha}{1-2\alpha}\right)^2 + d_C \frac{\mu}{1-2\alpha}}. \tag{59}$$

For $\alpha \in [\frac{1}{2}, 1)$, there is no traffic in equilibrium, corresponding to infinite prices.

Proposition 7 and Lemma 1 imply that *peering* yields an *overcongested network* with a higher demand rate and a lower price than under monopoly control, in contrast to decentralized control which yields an *undercongested and overpriced network*. The suboptimal performance of peering and decentralized control are the result of externalities that the providers inflict on each other. The externalities in these cases are fundamentally different: Under peering the externality is *negative* and takes effect *across* customer segments that are charged by different providers. Under decentralized control the externality due to double-marginalization is *positive* and takes effect *within* customer segments that are charged by both providers. Under peering, a single provider controls the price for each segment, but the congestion caused by that segment also increases delays for the other segment, which lowers the price that the other firm can charge. Since firms only generate revenue from one of the cross-traffic flows, they ignore the negative revenue externalities they inflict on their peers' customers; their equilibrium prices are hence lower than overall optimal levels. The equilibrium prices (54) and (56) capture this discrepancy between peering and monopoly decisions: The peering price for type-Cn depends only on its own revenue externality MC_n, whereas the respective monopoly price accounts for the resulting externalities on both segments, $\sum_{n=1}^{2} MC_n$. In contrast, under decentralized control each provider ignores the positive externality of a lower cross-traffic price on the other firm's profit. Therefore, the equilibrium prices (58) equal a *multiple* $2/(1-2\alpha)$ of the total marginal revenue externalities $\sum_{n=1}^{2} MC_n$.

It is worth highlighting that peering agreements may be effective for coordination if congestions costs are insignificant, specifically, if customers are *not* time-sensitive ($d_C = 0$). In this case, increasing the demand rate of one cross-traffic flow does not affect the QoS of the other. However, this does *not* hold under decentralized control: Double-marginalization may result in under-utilization even in the absence of time sensitivity (Corollary 1). Corollary 4 summarizes these comparisons.

Corollary 4 *For a symmetric network with two cross-traffic flows:*

1. If $d_C > 0$, peering yields an overcongested and underpriced network:

$$\lambda_C^D < \lambda_C^M < \lambda_C^P, \tag{60}$$

$$P_C^D > P_C^M > P_C^P. \tag{61}$$

2. If $d_C = 0$, there is a symmetric peering equilibrium that attains the centralized solution.

This analysis suggests that, contrary to common Internet connection practice, providers should not peer for free if their customers are time-sensitive, even if their traffic flows are balanced. Providers may instead consider the use of transfer prices.

4.2 Transfer Prices

The analysis so far shows that decentralized control typically undercongests whereas peering overcongests the network. Vertical integration through merger or acquisition would solve this problem, but data communication services are often jointly delivered by independent providers which must coordinate their decisions through subcontracting agreements. We consider subcontracts that specify linear transfer prices, i.e., a certain fee per unit of traffic. (The analysis of Section 4.1 implies that lump-sum transfers would not eliminate the performance losses that occur under peering.) Providers bargain over transfer prices since their incentives are misaligned: each would ideally like to forward traffic for free while charging the other for carrying its traffic. The game has three stages. 1. Providers bargain over transfer prices; 2. providers set their own cross-traffic prices simultaneously and independently; 3. customers decide whether to purchase service.

Without explicit analysis, for a symmetric network we can justify as the bargaining outcome of the first stage the single common transfer price that maximizes the total provider profit and gives each provider an equal share. Intuitively, providers will bargain for equal equilibrium payoffs since they are "equally powerful" and make equal profits in the absence of an agreement, i.e., under decentralized control (Lemma 1). Since both providers' profits increase in the total profit, it is in their interest to maximize it. More formally, the outcome where each provider makes half the total profit attained under centralized control is the unique Nash bargaining solution for the symmetric network with threat point given by the equilibrium profits under decentralized control, see Muthoo (1999). The Nash bargaining solution only specifies a profit distribution, but not the procedure that yields it. In our setting, the following simple procedure gives rise to the single common transfer price that yields the Nash bargaining solution: Either provider offers two unit transfer prices, one for each cross-traffic flow, and the other determines which of the two provider

gets to charge which price. (More generally, it is well known that under appropriate assumptions, a sequence of alternating offers produces the Nash bargaining solution, see Binmore et al. (1986).)

To illustrate the impact of the transfer price on system performance, let $T \geq 0$ be the common unit transfer price set in the first stage and consider the subsequent decisions. Of course, $T = 0$ corresponds to the peering agreement discussed in §4.1. Given $T \geq 0$, providers determine the price and demand rate for their own cross-traffic customers. Given λ_{Cm}, provider $n \neq m$ solves:

$$\begin{aligned} \max_{\lambda_{Cn} \geq 0} \Pi_n(\lambda_{Cn}; \lambda_{Cm}) &= \lambda_{Cn} \cdot [P_{Cn}(\boldsymbol{\lambda}) - T] + \lambda_{Cm} \cdot T \\ s.t. &\quad \lambda_{C1} + \lambda_{C2} < \mu. \end{aligned}$$

The first-order conditions for an interior equilibrium are

$$(1-\alpha)\, P_{Cn}(\boldsymbol{\lambda}) = \lambda_{Cn} V'_{Cn}(\lambda_{Cn}) \frac{2d_C}{(\mu - \lambda_{C1} - \lambda_{C2})^2} + T, \quad n = 1, 2. \tag{62}$$

Provider n equates her marginal type-Cn cross-traffic revenue under constant QoS, the LHS of (62), to the sum of marginal revenue externality inflicted on its own revenue *plus* the transfer price.

Proposition 8 *For the symmetric network with two cross-traffic flows and fixed transfer price* $T \in [0, V'_{Cn}(0)\, Q_{Cn}(\mathbf{0}))$*:*

1. *There is a unique symmetric Nash equilibrium with* $\lambda_{C1}(T) = \lambda_{C2}(T) = \lambda_C(T) > 0$.
2. *The optimal transfer price that yields the centralized solution (56)-(57) satisfies*
$$T^M = \frac{P_C^M}{2}(1-\alpha) = A_C \left(\lambda_C^M\right)^{(1-\alpha)} \frac{2d_C}{\left(\mu - 2\lambda_C^M\right)^2}. \tag{63}$$
3. *For* $\alpha \in [0, \frac{1}{2})$*, the transfer price that yields the decentralized equilibrium (58)-(59) satisfies*
$$T^D = \frac{P_C^D}{2}\left(\frac{3}{2} - \alpha\right) = \frac{3-2\alpha}{1-2\alpha} \cdot A_C \left(\lambda_C^D\right)^{(1-\alpha)} \frac{2d_C}{\left(\mu - 2\lambda_C^D\right)^2}. \tag{64}$$

By Proposition 8, the equilibrium profits are the same for both providers. The optimal transfer price T^M that yields the centralized solution equals the marginal revenue externality of *one* cross-traffic flow at the optimal arrival rates $\boldsymbol{\lambda}^M$. While the net payment between the providers is zero in equilibrium, the transfer price gives them the incentive to consider the impact of their pricing decisions on overall performance. This suggests that even providers with balanced traffic flows between their networks should *not* enter peering agreements with a zero transfer price. As (64) shows, the transfer price T^D that yields the decentralized solution is a multiple of each cross-traffic flow's marginal revenue externality, which reflects the double-marginalization effect.

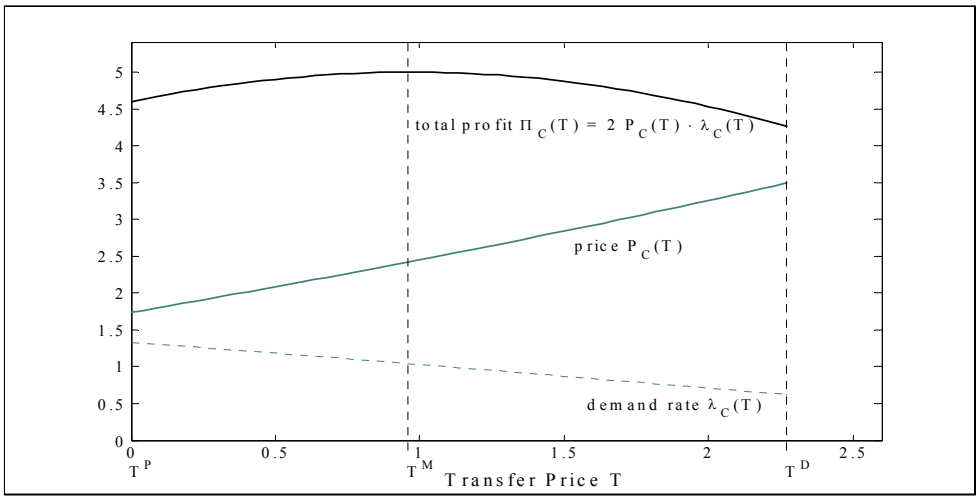

Figure 4: Symmetric example network: $\mu = 10$, $A_C = 10$, $\alpha = 0.2$, $d_C = 3$. Equilibrium demand rates, prices, and profits for transfer prices $T \in [0, T^D]$. Peering corresponds to $T^P = 0$; the centralized solution to T^M, defined in (63); and the decentralized equilibrium to T^D, defined in (64).

Example 2. The following example shows that a range of transfer prices yield a Pareto improvement compared to the decentralized control and peering equilibria. Each node has capacity $\mu = 10$, the demand parameters are $A_C = 10$, $\alpha = 0.2$, and the delay sensitivity parameter is $d_C = 3$. Figure 4 shows the equilibrium demand rates $\lambda_C(T)$, price $P_C(T)$ and total profit $\Pi_C(T)$ as the transfer price increases from $T^P = 0$ under peering to $T^D = 2.27$ which yields the decentralized control equilibrium, and the optimal transfer price is $T^M = 0.97$. The optimal total profit is $\Pi_C^M = 5.0$, roughly 10% and 17% higher than the profits under peering ($\Pi_C^P = 4.6$) and decentralized control ($\Pi_C^D = 4.3$), respectively. In the neighborhood of the optimal transfer price T^M, the total profit is relatively insensitive to the transfer price: $\Pi_C(T)|_{T=T^M-0.5} = \Pi_C(T)|_{T=T^M+0.5} = 4.89$. This suggests that even suboptimal transfer prices may yield a significant performance improvement.

5 Robustness of Results

In this section we discuss the robustness of our results to three assumptions: 1. Isoelastic marginal value functions; 2. identical demand elasticities for local- and cross-traffic in §§3.3-3.4; and 3. exponential service time distributions.

Isoelastic Marginal Value Functions. The family of isoelastic marginal value functions is appealing for analytical tractability, and it captures a range of demand scenarios. Our main results are not sensitive to the form of the marginal value functions. Three effects drive these results: double-marginalization and substitution under decentralized control, and the negative externality that yields overcongestion under peering. These effects are present under *any* value distribution.

Of course, the magnitudes of these effects and the equilibrium prices do depend on the marginal value functions. Two differences for non-isoelastic marginal value functions are worth noting.

First, contrary to the isoelastic case, for marginal value functions with increasing elasticity functions $\epsilon(\lambda)$, the decentralized price equilibrium may yield a *higher* system net value, relative to a monopoly. This follows from two facts. (i) If $\epsilon(\lambda)$ is increasing, the monopoly demand rate (price) in the absence of local-traffic is higher (lower) than socially optimal (Afèche and Mendelson 2004, Prop. 1, p. 873). (ii) By Proposition 2, decentralized control yields a lower (higher) equilibrium demand rate (price) compared to a monopoly (this holds for any marginal value function).

Second, the result that there is no cross-traffic in a decentralized equilibrium under an isoelastic marginal value function with elasticity $\epsilon \leq 2$ (i.e., $\alpha \geq 1/2$; see Propositions 1 and 5) generalizes as follows to the non-isoelastic case. In any decentralized equilibrium with a positive cross-traffic flow $\lambda_C^D > 0$, the cross-traffic elasticity must satisfy $\epsilon(\lambda^D) > 2$. For example, suppose the marginal value function for cross-traffic is linear, i.e., $V_C'(\lambda_C) = A_C\,(1 - \lambda_C/\Lambda_C)$. In this case, $\epsilon_C(\lambda_C) = \Lambda_C/\lambda_C - 1$, and $\lambda_C^D < \Lambda_C/3$ under any decentralized equilibrium with a positive cross-traffic flow.

Identical Demand Elasticities for Local- and Cross-Traffic. In §§3.3-3.4 we show that local-cross-traffic substitution in the presence of time-sensitive local traffic improves the revenue performance under decentralized control. These results assume identical demand elasticities for all flows. If the flows differ in their elasticities, the net effect of local-traffic on performance depends on the elasticity of the local-traffic vs. that of cross-traffic: Based on a numerical study we find that a decentralized network with local- and cross-traffic captures a higher fraction of optimal (monopoly) revenues than a decentralized network with only cross-traffic, if the *local-traffic is more elastic than cross-traffic.* We obtain this result by computing and comparing the price equilibria under decentralized and monopoly control for the following parameter values. As in Example 1, capacities are $\mu_1 = \mu_2 = 100$, the marginal value functions have $(A_{L1}, A_{L2}, A_C) = (5, 5, 10)$, and the delay sensitivity parameters are $(d_{L1}, d_{L2}, d_C) = (6, 6, 3)$. However, we choose different elasticity parameters for local- vs. cross-traffic, denoted by α_L and α_C, respectively: We set symmetric local-traffic elasticities ($\alpha_{L1} = \alpha_{L2} = \alpha_L$) and compute the equilibria for $(\alpha_C, \alpha_L) \in [0, 0.49] \times [0, 0.99]$. (For $\alpha_C \geq 0.5$ there is no cross-traffic in equilibrium.)

Example 3. Figure 5 shows the results for $\alpha_C = 0.35$, which are representative of our findings for $\alpha_C \in [0, 0.49]$. If the local-traffic is either more elastic than cross-traffic (i.e., $\alpha_L \leq 0.35$) or quite inelastic (i.e., $\alpha_L > 0.9$), the decentralized network captures a larger fraction of optimal revenues, compared to a decentralized network with only cross-traffic. In these ranges of α_L the local-traffic revenue gain exceeds the loss on cross-traffic, relative to a cross-traffic-only network. For low α_L,

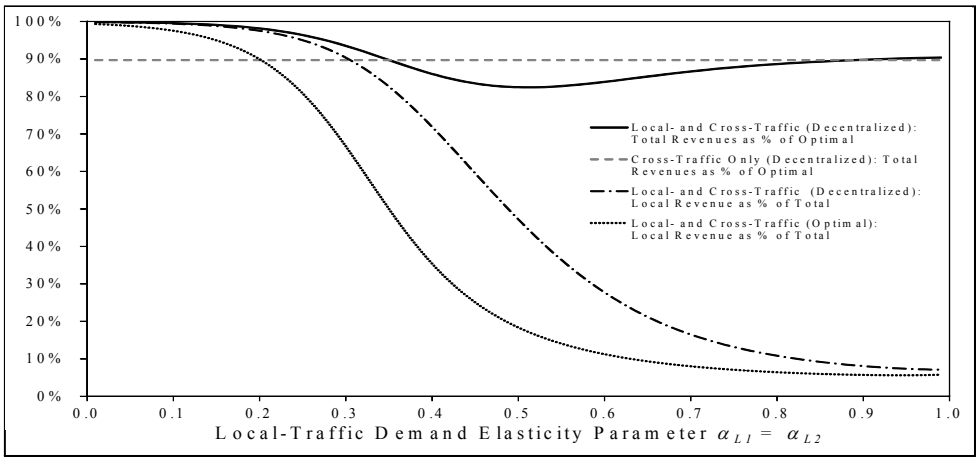

Figure 5: Symmetric example network: $\mu_1 = \mu_2 = 100$, $(A_{L1}, A_{L2}, A_C) = (5, 5, 10)$ and $(d_{L1}, d_{L2}, d_C) = (6, 6, 3)$. Cross-traffic elasticity parameter $\alpha_C = 0.35$. Impact of local-traffic elasticity on total revenues under decentralized control as % of optimal (monopoly).

local-traffic accounts for a large share of total revenues, and the substitution effect yields a large share of monopoly revenues for the decentralized network. For high α_L, cross-traffic accounts for most of the revenues; the decentralized networks, with or without local-traffic, have almost identical cross-traffic revenues, and the network with local-traffic generates additional revenue. The percentage revenue loss of the decentralized network vs. the optimal solution is largest when the local-traffic is somewhat less elastic than the cross-traffic (i.e., for $\alpha_L \in [0.4, 0.6]$): In this range, the local-traffic accounts for a modest share of total revenues and the substitution effect is less significant; as result, the local-traffic revenue gain is dominated by the loss on cross-traffic.

Exponential Service Time Distributions. Our main results hold for non-exponential service time distributions, indeed, for any tandem queues with well-defined mean steady-state delays. The key requirements are that the expected delays are increasing (decreasing) and convex in arrival rates (capacity). In particular, all of the equilibrium results that are not in closed-form hold for general tandem queues. Only the results that involve closed-form equations depend on the M/M/1 assumption, specifically, the price-capacity equilibria (Propositions 3 and 6) and the equilibrium prices/demand rates under peering and transfer pricing (Propositions 7 and 8, respectively).

It is worth noting that explicit mean steady-state formulae are also available for tandem queues with non-exponential service time distributions, specifically, if each node operates as a *symmetric queue*. For example, this holds under a processor-sharing discipline (see Chapter 3.3 of Kelly 1979).

6 Concluding Remarks

This paper analyzes the performance of a decentralized congestion-prone service supply chain that is controlled by independent profit-maximizing providers who offer complementary services to heterogeneous time-sensitive customers that require either the service bundle (cross-traffic) or only one of its components (local-traffic). We study decentralized control without coordination as a benchmark, compare it to the overall optimum under centralized control, and consider the performance impact of two practically relevant coordination mechanisms, peering contracts and transfer prices. Refer to §1 for a summary of results.

The model and results in this paper raise a number of fruitful questions for further research. In this context it is important to emphasize that the theory on complementary providers in the presence of queueing effects is relatively sparse. We outline three directions. First, the current model assumes uncorrelated traffic flows. An interesting and relevant issue is to consider pricing and capacity decisions that account for traffic flow correlation. E.g., signing up customers on one network is likely to spur both local- and cross-traffic. Second, a challenging yet important direction is to develop results on network performance for more general network topologies and provider control structures. Third, it is important to understand the effects of selling differentiated service, e.g., through priority queueing, in settings with complementary providers. For example, if one provider decides to sell priority service, how does this affect the incentives of others to follow suit?

References

Acemoglu, D., A. Ozdaglar. 2007. Competition in parallel-serial networks. *IEEE J. Sel. Areas in Comm.* **25**(6) 1180–1192.

Afèche, P., H. Mendelson. 2004. Pricing and Priority Auctions in Queueing Systems with a Generalized Delay Cost Structure. *Management Science* 50 869-882.

Allon, G., A. Federgruen. 2007. Competition in service industries. *Oper. Res.* **55**(1) 37-55.

Binmore, K., A. Rubinstein, A. Wolinsky. 1986. The Nash bargaining solution in economic modeling. *Rand Journal of Economics* **17**(2) 176-188.

Cachon, G. 2003. Supply chain coordination with contracts, in: S. Graves and T. de Kok (eds.) *Handbooks in Operations Research and Management Science: Supply Chain Management.* North Holland.

Cave, M. and Donelly, M. P. 1996. The Pricing of International Telecommunications Services by Monopoly Operators. *Information Economics and Policy* **8** 107-123.

Crémer, J., P. Rey, and J. Tirole. 2000. Connectivity in the Commercial Internet. *J. Industrial Economics* **48**(4) 433-472.

Dewan, S. and H. Mendelson. 1998. Information Technology and Time-based Competition in Financial Markets. *Man. Sci.* **44**(5) 595-609.

Farzan, A., Y.-P. Zhou. 2012. Analysis of a two-stage service process: coordination of staffing and effort. Working paper, University of Washington, Seattle, WA.

Hassin R., M. Haviv. 2003. *To Queue or not to queue: Equilibrium behavior in queueing systems.* Kluwer, Boston.

He, L., J. Walrand. 2006. Pricing and revenue sharing strategies for Internet service providers. *IEEE J. Sel. Areas in Comm.* **24**(5) 1180–1192.

Hu, X., R. Caldentey, G. Vulcano. 2012. Revenue sharing in airline alliances. In press, *Man. Sci.*

Kelly, F.P. 1979. *Reversibility and Stochastic Networks*, Wiley, New York.

Laffont, J.J., P. Rey and J.Tirole. 1998. Network Competition: I. Overview and nondiscriminatory pricing. *RAND J. of Economics* **29**(1) 1-37.

Laffont, J.J., P. Rey and J.Tirole. 1998. Network Competition: II. Price Discrimination. *RAND Journal of Economics* **29**(1) 38-56.

Laffont, J.J., P. Rey and J.Tirole. 2003. Internet Interconnection and the Off-Net-Cost-Principle. *RAND Journal of Economics* **34**(2) 370-390.

Lederer, P.J., L. Li. 1997. Pricing, Production, Scheduling and Delivery-Time Competition. *Oper. Res.* **45**(3) 407-420.

Li, L., Y.S. Lee. 1994. Pricing and delivery-time performance in a competitive environment. *Man. Sci.* **40**(5) 633-646.

MacKie-Mason, J.K. and H.R. Varian. 1995. Pricing Congestible Network Resources. *IEEE J. Sel. Areas in Comm.* **13**(7) 1141-1149.

Masuda, Y. and S.Whang. 1999. Dynamic Pricing for Network Service. *Management Science* **45**(6) 857-870.

Mendelson, H. 1985. Pricing Computer Services: Queueing Effects. *Communications of the ACM* **28**(3) 312-321.

Mendelson, H., S.Whang. 1990. Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue. *Oper. Res.* **38**(5) 870-883.

Milgrom, P., Mitchell, B. and Srinagesh, P. 2000. Competitive Effects of Internet Peering Policies, in: I. Vogelsang and B. Compaine (eds.) *The Internet Upheaval*, MIT Press, Cambridge, MA, 175-195.

Muthoo, A. 1999. *Bargaining Theory with Applications*, Cambridge Univ. Press, Cambridge, UK.

Naor, P. 1969. On the Regulation of Queue Size by Levying Tolls. *Econometrica* **37**(1) 15-24.

Netessine, S., R. Shumsky. 2005. Revenue management games: horizontal and vertical competition. *Man. Sci.* **51**(5) 813–831.

Netessine, S., F. Zhang. 2005. Positive vs. negative externalities in inventory management: implications for supply chain design. *M&SOM* **7**(1) 58-73.

Shneorson, S. and H. Mendelson. 2003. Internet Peering, Capacity and Pricing. Working paper, Stanford University.

Spengler, J. 1950. Vertical Integration and Anti-Trust Policy. *J. of Political Economy* **58**(3) 347-352.

Veltman, A., R. Hassin. 2005. Equilibrium in queueing systems with complementary products. *Queue. Syst.* **50** 325–342.

Online Supplement: Proofs

Proof of Proposition 1. Let $\Omega := \{\lambda_C : 0 \leq \lambda_C \leq \min(\Lambda_C, \mu_1, \mu_2)\}$ be the set of feasible rates. An interior equilibrium $\left(\lambda_C^D, p_{C1}^D, p_{C2}^D\right)$ with $\lambda_C^D > 0$ must satisfy $\lambda_C^D = \arg\max_{\lambda_C \in \Omega} \Pi_1(\lambda_C; p_{C2}^D) = \arg\max_{\lambda_C \in \Omega} \Pi_2(\lambda_C; p_{C1}^D)$ and $P_C(\lambda_C^D) = p_{C1}^D + p_{C2}^D$. The proof proceeds in three steps.

1. Necessary condition for interior equilibrium. Adding the first-order conditions (5) for provider $n = 1, 2$ and noting that $\lambda_C \cdot P_C'(\lambda_C) = -\alpha V_C'(\lambda_C) Q_C(\lambda_C) + (1-\alpha) V_C(\lambda_C) Q_C'(\lambda_C)$ yields the following necessary condition for an interior equilibrium:

$$f(\lambda_C) := (1-2\alpha) V_C'(\lambda_C) Q_C(\lambda_C) + 2(1-\alpha) V_C(\lambda_C) Q_C'(\lambda_C) = 0. \tag{65}$$

Note that $f(\lambda_C)$ has a unique root in the interior of Ω if $\alpha \in [0, \frac{1}{2})$: Assumption *A1* implies $f(0) > 0$, and *A2* implies $f(\lambda_C) < 0$ as $\lambda_C \to \min(\Lambda_C, \mu_1, \mu_2)$. The claim follows since $f'(\lambda_C) < 0$ on Ω. If $\alpha \in [\frac{1}{2}, 1)$ then $f(\lambda_C)$ has no strictly positive root: $f(0) \leq 0$ and $f'(\lambda_C) < 0$ on Ω. Hence there is at most one interior equilibrium if $\alpha \in [0, \frac{1}{2})$ and none if $\alpha \in [\frac{1}{2}, 1)$.

2. Existence of unique interior equilibrium for $\alpha \in [0, \frac{1}{2})$. Let $\lambda_C^\circ > 0$ be the unique root of $f(\lambda_C)$. Note that $P_C(\lambda_C^\circ) > 0$. By symmetry of the first-order conditions (5), the equilibrium prices must be equal. Let $p_{C1}^\circ = p_{C2}^\circ = P_C(\lambda_C^\circ)/2 > 0$. We will show that $(\lambda_C^\circ, p_{C1}^\circ, p_{C2}^\circ)$ is an equilibrium. By symmetry, it suffices to show that $\lambda_C^\circ = \arg\max_{\lambda_C \in \Omega} \Pi_1(\lambda_C; p_{C2}^\circ)$. By construction it follows that p_{C2}° and λ_C° satisfy the first-order condition (5) for provider 1:

$$\Pi_1'(\lambda_C; p_{C2}^\circ) = 0 \Leftrightarrow (1-\alpha) V_C'(\lambda_C) Q_C(\lambda_C) + (1-\alpha) V_C(\lambda_C) Q_C'(\lambda_C) = \frac{P_C(\lambda_C^\circ)}{2}. \tag{66}$$

Since $\Pi_1(\lambda_C; p_{C2}^\circ)$ is strictly concave on Ω:

$$\Pi_1''(\lambda_C; p_{C2}^\circ) = (1-\alpha)[V_C''(\lambda_C) Q_C(\lambda_C) + V_C(\lambda_C) Q_C''(\lambda_C) + 2V_C'(\lambda_C) Q_C'(\lambda_C)] < 0, \tag{67}$$

it follows that $\lambda_C^\circ = \arg\max_{\lambda_C \in \Omega} \Pi_1(\lambda_C; p_{C2}^\circ)$, so $(\lambda_C^\circ, p_{C1}^\circ, p_{C2}^\circ)$ is an interior equilibrium.

3. Infinite prices for $\alpha \in [\frac{1}{2}, 1)$. Step 1 establishes that there is no interior equilibrium. We show that equilibrium prices must be infinite. For given $p_{C2} < \infty$, we have $\Pi_1'(0; p_{C2}) = \infty$ and $\Pi_1'(\lambda_C; p_{C2}) < 0$ as $\lambda_C \to \min(\Lambda_C, \mu_1, \mu_2)$, so provider 1's price response $p_{C1}(p_{C2})$ must satisfy the first-order condition (5) for an interior solution:

$$(1-\alpha) V_C'(\lambda_C) Q_C(\lambda_C) + (1-\alpha) V_C(\lambda_C) Q_C'(\lambda_C) = p_{C2}. \tag{68}$$

Substituting $V_C'(\lambda_C) Q_C(\lambda_C) = P_C(\lambda_C) = p_{C1}(p_{C2}) + p_{C2}$ and rearranging yields

$$p_{C1}(p_{C2}) = p_{C2} \frac{\alpha}{1-\alpha} - V_C(\lambda_C) Q_C'(\lambda_C) > p_{C2}, \tag{69}$$

so firm 1 "outbids" firm 2. Symmetry implies that prices are infinite and $\lambda_C^D = 0$.■

Proof of Proposition 2. For the result $\lambda_C^M = \lambda_C^*$, see Afèche and Mendelson 2004, Prop. 1, p. 873. Define:

$$g(\lambda_C, N) := (1 - N\alpha) V_C'(\lambda_C) Q_C(\lambda_C) + N(1-\alpha) V_C'(\lambda_C) Q_C'(\lambda_C). \tag{70}$$

From (65), note that $g(\lambda_C, 2) \equiv f(\lambda_C)$. By Proposition 1, $\lambda_C^D > 0$ is the unique root of $g(\lambda_C, 2)$ for $\alpha \in [0, \frac{1}{2})$ and $\lambda_C^D = 0$ for $\alpha \in [\frac{1}{2}, 1)$. Similarly, it is straightforward to verify for $\alpha \in [0, 1)$ that the centralized rates satisfy $\lambda_C^M = \lambda_C^* > 0$ and are the unique root of $g(\lambda_C, 1)$.

We show that $\lambda_C^M = \lambda_C^* > \lambda_C^D$: this holds trivially for $\alpha \in (\frac{1}{2}, 1)$, so fix $\alpha \in [0, \frac{1}{2})$. Denote by g_{λ_C} and g_N the partial derivatives of g. Since $g_{\lambda_C}(\lambda_C; N) < 0$ for $\lambda_C > 0$, there is an implicit function $\lambda_C(N)$ defined on the interval $[1, \frac{1}{\alpha}]$ which satisfies $\lambda_C(1) = \lambda_C^M = \lambda_C^*$, $\lambda_C(2) = \lambda_C^D$ and

$$\lambda_C'(N) = -\frac{g_N(\lambda_C(N), N)}{g_{\lambda_C}(\lambda_C(N), N)} = -\frac{-\alpha V_C'(\lambda_C) Q_C(\lambda_C) + (1-\alpha) V_C(\lambda_C) Q_C'(\lambda_C)}{g_{\lambda_C}(\lambda_C(N), N)} < 0. \tag{71}$$

Hence $\lambda_C^M = \lambda_C^* > \lambda_C^D$. Since $P_C'(\lambda_C) < 0$ it follows that $P_C^D > P_C^M = P_C^*$. Since $NV'(0) > 0$, $NV'(\lambda_C^*) = 0$ and $NV_C''(\lambda_C) = V_C''(\lambda_C) Q_C(\lambda_C) + V_C(\lambda_C) Q_C''(\lambda_C) + 2V_C'(\lambda_C) Q_C'(\lambda_C) < 0$ implies that $NV(\lambda_C^D) < NV(\lambda_C^*)$. Since $CS(\lambda_C) = \alpha \cdot NV(\lambda_C)$ and $\Pi(\lambda_C) = (1-\alpha) \cdot NV(\lambda_C)$ the same ordering holds for the expected total provider profit Π and consumer surplus CS.■

Proof of Proposition 3. The result is trivial for $\alpha \in [\frac{1}{2}, 1)$: by Proposition 1, $\lambda_C^D(\boldsymbol{\mu}) = 0$ for any capacity vector $\boldsymbol{\mu}$, so neither provider invests in capacity and $\lambda_C^D = \boldsymbol{\mu}^D = 0$.

Fix $\alpha \in (0, \frac{1}{2})$: An interior equilibrium $(\lambda_C^D, \boldsymbol{\mu}^D)$ to the price-capacity game must satisfy the first-order conditions (7)-(8). Summing these conditions for both providers yields

$$(1-2\alpha) P_C(\lambda_C^D, \boldsymbol{\mu}^D) = -2\lambda_C^D V_C'(\lambda_C^D) \frac{\partial Q_C(\lambda_C^D, \boldsymbol{\mu}^D)}{\partial \lambda_C} \tag{72}$$

$$\lambda_C^D V_C'(\lambda_C^D) \left[\frac{\partial Q_C(\lambda_C^D, \boldsymbol{\mu}^D)}{\partial \mu_1} + \frac{\partial Q_C(\lambda_C^D, \boldsymbol{\mu}^D)}{\partial \mu_2} \right] = b_1 + b_2. \tag{73}$$

Note that $Q_{\lambda_C} = -\left[\frac{\partial Q_C}{\partial \mu_1} + \frac{\partial Q_C}{\partial \mu_2} \right]$ to obtain the total cross-traffic price equation (9). Proposition 1 implies that $p_{Cn}^D = P_C^D/2$, $n = 1, 2$. Equation (10) is immediate from (3) and (8).

To show that there exist at most two interior equilibria, substitute $\mu_n^D(\lambda_C) = \sqrt{\frac{\lambda_C V_C'(\lambda_C) d_C}{b_n}} + \lambda_C$ from (8) for the equilibrium capacities into (9):

$$P_C^D = V_C'(\lambda_C^D) \left(1 - \sqrt{\frac{d_C}{\lambda_C^D V_C'(\lambda_C^D)}} \left(\sqrt{b_1} + \sqrt{b_2} \right) \right) = 2 \frac{b_1 + b_2}{1 - 2\alpha}. \tag{74}$$

Define the function

$$g(\lambda_C) = V_C'(\lambda_C) \left(1 - \sqrt{\frac{d_C}{\lambda_C V_C'(\lambda_C)}} \left(\sqrt{b_1} + \sqrt{b_2} \right) \right) \tag{75}$$

Inspection of its derivative

$$g'(\lambda_C) = V_C''(\lambda_C) \left(1 - \frac{1+\alpha}{2\alpha} \sqrt{\frac{d_C}{\lambda_C V_C'(\lambda_C)}} \left(\sqrt{b_1} + \sqrt{b_2} \right) \right) \tag{76}$$

shows that there is a cutoff demand rate $\overline{\lambda}_C$ such that $g'(\overline{\lambda}_C) = 0$ with $g'(\lambda_C) > 0$ for $\lambda_C < \overline{\lambda}_C$ and $g'(\lambda_C) < 0$ for $\lambda_C > \overline{\lambda}_C$. Since $g(0) = -\infty$ and $g(\lambda_C) \to 0$ as $\lambda_C \to \infty$, it follows that (74)

has exactly zero, one or two solutions, as $g(\overline{\lambda}_C) < 2\frac{b_1+b_2}{1-2\alpha}$, $g(\overline{\lambda}_C) = 2\frac{b_1+b_2}{1-2\alpha}$ or $g(\overline{\lambda}_C) > 2\frac{b_1+b_2}{1-2\alpha}$, respectively. Therefore, the price-capacity game has at most two interior equilibria.

To show that the lower cost provider has a higher profit: note that $b_1 < b_2$ and (10) imply

$$1 > \left(\frac{\mu_2^D - \lambda_C^D}{\mu_1^D - \lambda_C^D}\right)^2 = \frac{b_1}{b_2} > \left(\frac{b_1}{b_2}\right)^2 \Rightarrow \mu_1^D > \mu_2^D \Rightarrow \left(\frac{\mu_2^D}{\mu_1^D}\right)^2 > \left(\frac{\mu_2^D - \lambda_C^D}{\mu_1^D - \lambda_C^D}\right)^2, \tag{77}$$

and so $\mu_1^D b_1 < \mu_2^D b_2$. The result follows since each provider makes the same revenue.■

Proof of Proposition 4. The equilibrium prices and their ordering follow immediately from Proposition 3 for decentralized control and the first-order conditions under centralized control. The conditions for an interior solution to the profit maximization problem yield (we suppress the arguments of P_C and Q_C)

$$(1-\alpha)P_C = -\lambda_C^M V_C'(\lambda_C^M)\frac{\partial Q_C}{\partial \lambda_C} = \lambda_C^M V_C'(\lambda_C^M)\left[\frac{\partial Q_C}{\partial \mu_1} + \frac{\partial Q_C}{\partial \mu_2}\right] = b_1 + b_2 \tag{78}$$

and the conditions for social optimization are similar.

We first establish the ordering of equilibrium demand rates $\lambda_C^* > \lambda_C^M > \lambda_C^D$. Define the function $f(\lambda_C, \boldsymbol{\mu}; \theta) = V_C(\lambda_C)Q_C(\lambda_C, \boldsymbol{\mu}) - \theta[b_1\mu_1 + b_2\mu_2]$, where $\theta \geq 0$. Thus

$$NV(\lambda_C, \boldsymbol{\mu}) = f(\lambda_C, \boldsymbol{\mu}; 1) \tag{79}$$

$$\Pi(\lambda_C, \boldsymbol{\mu}) = (1-\alpha)f(\lambda_C, \boldsymbol{\mu}; \frac{1}{1-\alpha}) \tag{80}$$

are the net value and profit functions, respectively. The function f is strictly concave in $\boldsymbol{\mu}$ for fixed demand rate $\lambda_C > 0$ (f is not jointly concave in $(\lambda_C, \boldsymbol{\mu})$), so the optimal capacity vector $\boldsymbol{\mu}(\lambda_C, \theta) := \arg\max_{\boldsymbol{\mu}} f(\lambda_C, \boldsymbol{\mu}; \theta)$ is the unique solution of $\nabla_{\boldsymbol{\mu}} f(\lambda_C, \boldsymbol{\mu}; \theta) = 0$ and satisfies

$$\mu_n(\lambda_C, \theta) = \sqrt{\frac{V_C(\lambda_C)d_C}{\theta \cdot b_n}} + \lambda_C, \quad n = 1, 2. \tag{81}$$

Substitute into f to obtain

$$f(\lambda_C, \boldsymbol{\mu}(\lambda_C, \theta); \theta) = V_C(\lambda_C)\left[1 - 2\sqrt{\frac{\theta \cdot d_C}{V_C(\lambda_C)}}\left(\sqrt{b_1} + \sqrt{b_2}\right)\right] - \lambda_C\theta(b_1 + b_2). \tag{82}$$

To show that $\lambda_C(\theta) := \arg\max_{\lambda_C} f(\lambda_C, \boldsymbol{\mu}(\lambda_C, \theta); \theta)$ is decreasing in θ, note that:

$$f_{\lambda_C\theta}(\lambda_C, \boldsymbol{\mu}(\lambda_C, \theta); \theta) = -\left(\frac{V_C'(\lambda_C)}{2}\sqrt{\frac{d_C}{\theta V_C(\lambda_C)}}\left(\sqrt{b_1} + \sqrt{b_2}\right) + b_1 + b_2\right) < 0, \tag{83}$$

so f has non-increasing differences in $(\lambda_C; \theta)$. Since it is maximized over $\mathbb{R}^+$ for any fixed θ, it follows that $\lambda_C'(\theta) < 0$ (**?**, Theorem 1) which implies that $\lambda_C^M = \lambda_C(\frac{1}{1-\alpha}) < \lambda_C^* = \lambda_C(1)$. Next consider the two-provider case. Substitute $\mu_n^D(\lambda_C) = \sqrt{\frac{\lambda_C V_C'(\lambda_C)d_C}{b_n}} + \lambda_C$ from (8) for the equilibrium capacities into the equilibrium price equation (9) to obtain

$$P_C^D = V_C'(\lambda_C^D)\left(1 - \sqrt{\frac{d_C}{\lambda_C^D V_C'(\lambda_C^D)}}\left(\sqrt{b_1} + \sqrt{b_2}\right)\right) = 2\frac{b_1 + b_2}{1 - 2\alpha}. \tag{84}$$

With the function $g(\lambda_C)$ as defined by (75), the price equations (78) and (84) imply that

$$g(\lambda_C^M) = \frac{b_1+b_2}{1-\alpha} \tag{85}$$

$$g(\lambda_C^D) = 2\frac{b_1+b_2}{1-2\alpha}. \tag{86}$$

Claim: If λ_C^M is interior it is unique and the largest solution of equation (85): Note that

$$\Pi(\lambda_C, \boldsymbol{\mu}(\lambda_C, \frac{1}{1-\alpha})) = \int_0^{\lambda_C} \left(g(x) - \frac{b_1+b_2}{1-\alpha} \right) \, dx. \tag{87}$$

It $\lambda_C^M = \arg\max_{\lambda_C \geq 0} \Pi(\lambda_C, \boldsymbol{\mu}(\lambda_C, \frac{1}{1-\alpha})) > 0$, it must satisfy (85) and $\Pi''(\lambda_C^M, \boldsymbol{\mu}(\lambda_C^M, \frac{1}{1-\alpha})) = g'(\lambda_C^M) \leq 0$. Recall from the proof of Proposition 3 that there is a cutoff demand rate $\overline{\lambda}_C$ such that $g'(\lambda_C) > 0$ for $\lambda_C < \overline{\lambda}_C$ and $g'(\lambda_C) < 0$ for $\lambda_C > \overline{\lambda}_C$. It follows that $\lambda_C^M > \overline{\lambda}_C$ and so $\Pi'(\lambda_C, \boldsymbol{\mu}(\lambda_C, \frac{1}{1-\alpha})) = g(\lambda_C) - \frac{b_1+b_2}{1-\alpha} < 0$ for $\lambda_C > \lambda_C^M$, which establishes the claim.

Noting that $g(\lambda_C^D) - \frac{b_1+b_2}{1-\alpha} = 2\frac{b_1+b_2}{1-2\alpha} - \frac{b_1+b_2}{1-\alpha} > 0$ implies that $\lambda_C^D < \lambda_C^M$, which establishes that $\lambda_C^* > \lambda_C^M > \lambda_C^D$. We use this ordering to prove the remaining inequalities of the proposition.
Capacities: From (81) it follows that $\boldsymbol{\mu}_{\lambda_C}(\lambda_C, \theta) > 0$ and $\boldsymbol{\mu}_{\theta}(\lambda_C, \theta) < 0$, which implies that

$$\boldsymbol{\mu}^D = \boldsymbol{\mu}(\lambda_C^D, \frac{1}{1-\alpha}) < \boldsymbol{\mu}^M = \boldsymbol{\mu}(\lambda_C^M, \frac{1}{1-\alpha}) < \boldsymbol{\mu}^* = \boldsymbol{\mu}(\lambda_C^*, 1). \tag{88}$$

Qualities of Service: from equation (81), $Q_C(\lambda_C, \boldsymbol{\mu}(\lambda_C, \theta)) = 1 - \sqrt{\frac{\theta d_C}{V_C(\lambda_C)}} \left(\sqrt{b_1} + \sqrt{b_2} \right)$, implying that $\frac{\partial Q_C(\lambda_C, \boldsymbol{\mu}(\lambda_C, \theta))}{\partial \lambda_C} > 0$ and $\frac{\partial Q_C(\lambda_C, \boldsymbol{\mu}(\lambda_C, \theta))}{\partial \theta} < 0$. Hence

$$Q_C(\lambda_C^D, \boldsymbol{\mu}(\lambda_C^D, \frac{1}{1-\alpha})) < Q_C(\lambda_C^M, \boldsymbol{\mu}(\lambda_C^M, \frac{1}{1-\alpha})) < Q_C(\lambda_C^*, \boldsymbol{\mu}(\lambda_C^*, 1)). \tag{89}$$

Expected total profits: from $P_C^* = b_1 + b_2$ and equation (81) it follows that

$$\Pi^* = \Pi(\lambda_C^*, \boldsymbol{\mu}(\lambda_C^*, 1)) = \lambda_C^*[b_1+b_2] - b_1\mu_1(\lambda_C^*, 1) - b\mu_2(\lambda_C^*, 1) = -\sum_{n=1}^{2} \sqrt{b_n}\sqrt{V_C(\lambda_C^*)d_C} < 0. \tag{90}$$

The monopoly profit for an interior solution clearly exceeds that under decentralized control, so

$$\Pi^M > \Pi^D > 0 > \Pi^*. \tag{91}$$

Expected consumer surplus rate: note that $CS(\lambda_C, \boldsymbol{\mu}) = \alpha \cdot V_C(\lambda_C) Q_C(\lambda_C, \boldsymbol{\mu})$, so

$$CS^* := \alpha V_C(\lambda_C^*) Q_C^*(\lambda_C^*) > CS^M := \alpha V_C(\lambda_C^M) Q_C^M(\lambda_C^M) > CS^D := \alpha V_C(\lambda_C^D) Q_C^D(\lambda_C^D). \tag{92}$$

Expected system net value rate: the result follows since $NV(\lambda_C, \boldsymbol{\mu}) \equiv CS(\lambda_C, \boldsymbol{\mu}) + \Pi(\lambda_C, \boldsymbol{\mu})$.■

Proof of Proposition 5. The equilibrium demand vector $\boldsymbol{\lambda}^D$, prices $(P_{L1}^D, P_{L2}^D, p_{C1}^D, p_{C2}^D)$ and Lagrange multipliers $\boldsymbol{\theta}_n^D$ and δ_n $(n = 1, 2)$ must satisfy (17)-(19) and:

$$\frac{\partial \mathcal{L}_n}{\partial \lambda_{Ln}} = 0, \quad n = 1, 2 \tag{93}$$

$$\frac{\partial \mathcal{L}_1}{\partial \lambda_C} + \frac{\partial \mathcal{L}_2}{\partial \lambda_C} = (1-2\alpha)P_C + 2\lambda_C V_C' \frac{\partial Q_C}{\partial \lambda_C} + \sum_{n=1}^{2} \left[(\lambda_{Ln} + \delta_{3-n}) \frac{\partial P_{Ln}}{\partial \lambda_C} + \theta_{nC} \right] = 0, \tag{94}$$

$$\boldsymbol{\theta}_n \geq 0; \quad \boldsymbol{\theta}_n' \boldsymbol{\lambda} = 0, \quad n = 1, 2. \tag{95}$$

From equation (22) the Lagrange multiplier δ_1 is

$$\delta_1 = \frac{\lambda_C \frac{\partial P_C}{\partial \lambda_{L2}} + \theta_{12}}{-\frac{\partial P_{L2}}{\partial \lambda_{L2}}} = \frac{\lambda_C V_C' \frac{\partial Q_C}{\partial \lambda_{L2}} + \theta_{12}}{-V_{L2}'' Q_{L2} - V_{L2}' \frac{\partial Q_{L2}}{\partial \lambda_{L2}}} = \lambda_{L2} \frac{(1-\alpha) V_C \frac{\partial Q_C}{\partial \lambda_{L2}} + \theta_{12}}{-\left[\alpha V_C \frac{\partial Q_C}{\partial \lambda_{L2}} + V_{L2} \frac{\partial Q_{L2}}{\partial \lambda_{L2}} + \frac{\alpha \cdot \theta_{22}}{1-\alpha}\right]}. \tag{96}$$

The last equality follows by substituting for P_{L2} from provider 2's first-order condition $\frac{\partial \mathcal{L}_2}{\partial \lambda_{L2}} = 0$:

$$P_{L2} = V_{L2}' Q_{L2} = -\left[V_C \frac{\partial Q_C}{\partial \lambda_{L2}} + V_{L2} \frac{\partial Q_{L2}}{\partial \lambda_{L2}} + \frac{\theta_{22}}{1-\alpha}\right]. \tag{97}$$

Substitute for δ_1 and δ_2 into (94) and define $f(\boldsymbol{\lambda}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \frac{\partial \mathcal{L}_1}{\partial \lambda_C} + \frac{\partial \mathcal{L}_2}{\partial \lambda_C} - [\theta_{1C} + \theta_{2C}]$ to obtain

$$f(\boldsymbol{\lambda}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (1-2\alpha) V_C' Q_C - \left[\sum_{n=1}^{2} 2MC_{Cn} + MC_{Ln} \frac{\frac{\alpha \theta_{n,n}}{1-\alpha} - \theta_{n,3-n} - (1-2\alpha) V_C \frac{\partial Q_C}{\partial \lambda_{Ln}} + V_{Ln} \frac{\partial Q_{Ln}}{\partial \lambda_{Ln}}}{\alpha V_C \frac{\partial Q_C}{\partial \lambda_{Ln}} + V_{Ln} \frac{\partial Q_{Ln}}{\partial \lambda_{Ln}} + \frac{\alpha \theta_{n,n}}{1-\alpha}}\right].$$

Now consider the two cases $\alpha \in [\frac{1}{2}, 1)$ and $\alpha \in (0, \frac{1}{2})$:

For $\alpha \in [0, \frac{1}{2})$: If the equilibrium is interior, then $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \mathbf{0}$. The local-traffic price equations (26) follow from the first-order conditions (93), and the cross-traffic equilibrium condition (94) reduces after some algebra to the price equation (27). Substitute for P_C, δ_1, δ_2 into (23), and set $\theta_{1C} = \theta_{2C} = 0$, to obtain the equations (28) for the cross-traffic prices p_{C1}^D and p_{C2}^D.

For $\alpha \in [\frac{1}{2}, 1)$: If $\lambda_C^D > 0$, then $Q_C(\boldsymbol{\lambda}^D) > 0$; otherwise, both providers' cross-traffic revenues is nonpositive and either can increase profits by reducing the cross-traffic demand rate and increasing its local-traffic . Hence $(1-2\alpha) V_C'(\lambda_C^D) Q_C(\boldsymbol{\lambda}^D) \leq 0$. The second summand of f (the bracketed term) is strictly positive, hence $f < 0$. Therefore, equation (94) is only satisfied if $\lambda_C^D = 0$ and $\theta_{nC} > 0$, $n = 1, 2$. Each provider independently maximizes her monopoly profit from local-traffic , solving $\max_{0 \leq \lambda_{Ln} \leq \mu_n} \Pi_{Ln}(\lambda_{Ln}; \lambda_C^D = 0) := \lambda_{Ln} V_{Ln}'(\lambda_{Ln}) Q_{Ln}(\lambda_{Ln}; \lambda_C^D = 0)$. Since $Q_{Ln}(0; \lambda_C^D = 0) > 0$, it follows that $\Pi_{Ln}'(0; \lambda_C^D = 0) = \infty$. For $\lambda_{Ln} = \overline{\lambda}_{Ln} := \mu_{Ln} - d_{Ln}$, $Q_{Ln}(\overline{\lambda}_{Ln}; \lambda_C^D = 0) = 0 \quad \Rightarrow \Pi_{Ln}'(\overline{\lambda}_{Ln}; \lambda_C^D = 0) < 0$. Since $\Pi_{Ln}'' < 0$, there is a unique demand rate $\lambda_{Ln}^D > 0$ which maximizes $\Pi_{Ln}(\lambda_{Ln}; \lambda_C^D = 0)$ and satisfies $\Pi_{Ln}'(\lambda_{Ln}^D; \lambda_C^D = 0) = 0$. Since for $\lambda_{Ln} > 0$ and $\theta_{nn} = 0$, $\Pi_{Ln}'(\lambda_{Ln}; \lambda_C^D = 0) = \frac{\partial \mathcal{L}_n(\lambda_{Ln}; \lambda_C^D = \theta_{nn} = 0)}{\partial \lambda_{Ln}}$, λ_{Ln}^D satisfies the equilibrium condition (93). The complementary slackness conditions (95) are also satisfied. Noting that $\Pi_{Ln}'(\lambda_{Ln}^D; \lambda_C^D = 0) = (1-\alpha) P_{Ln}^D - MC_{Ln}(\lambda_{Ln}^D; \lambda_C^D = 0)$ establishes the price equation (25).■

Proof of Proposition 6. By Proposition 5, for any fixed capacity level the equilibrium prices must satisfy (25) for $\alpha \in [\frac{1}{2}, 1)$ if there is local-traffic , and (26)-(28) for $\alpha \in [0, \frac{1}{2})$ if $\boldsymbol{\lambda}^D > 0$. Noting that (45), the optimality condition for the node-n capacity, satisfies

$$\frac{\lambda_C V_C'(\lambda_C) d_C + \lambda_{Ln} V_{Ln}'(\lambda_{Ln}) d_{Ln}}{(\mu_n - \lambda_C - \lambda_{Ln})^2} = MC_{Cn}(\boldsymbol{\lambda}) + MC_{Ln}(\boldsymbol{\lambda}) = b_n \tag{98}$$

and substituting into (25)-(28) yields the price equations (46)-(49).■

Proof of Proposition 7. First observe that any equilibrium must be interior. For $n = 1, 2$:

$$\Pi_n'(\lambda_{Cn}; \lambda_{Cm})\big|_{\lambda_{Cn}=0} = (1-\alpha) \cdot V_{Cn}'(0)\, Q_{Cn}(\boldsymbol{\lambda})|_{\lambda_{Cn}=0} > 0 \Leftrightarrow Q_{Cn}(\boldsymbol{\lambda})|_{\lambda_{Cn}=0} > 0. \tag{99}$$

Since the network is symmetric, $Q_{C1}(0) \equiv Q_{C2}(0)$, hence neither provider m has an incentive to choose her arrival rate λ_{Cm} such that $Q_{Cm}(0)|_{\lambda_{Cn}=0} \leq 0$. This and the fact that $Q_{Cn}(0) > 0$ for $n = 1, 2$, implies that $Q_{Cn}(0)|_{\lambda_{Cn}=0} > 0$ in equilibrium, and provider n chooses a positive arrival rate. Further note that $\Pi_n(\lambda_{Cn}; \lambda_{Cm})$ is strictly concave in λ_{Cn} for fixed λ_{Cm} $(m \neq n)$.

Since the equilibrium is interior, it follows from (53) that the arrival rates must satisfy

$$(1-\alpha)\, Q_{Cn}(\boldsymbol{\lambda}) = \lambda_{Cn} \frac{2d_C}{(\mu - \lambda_{C1} - \lambda_{C2})^2}, \quad n = 1, 2. \tag{100}$$

Hence, both cross-traffic arrival rates must be identical. Substituting $\lambda_{C1} = \lambda_{C2} = \lambda_C$ in (100) yields a quadratic equation in λ_C. If $2d_C < \mu$, then it has a unique real root λ_C^P that satisfies the capacity constraint $\lambda_C^P < \mu/2$. It is given by (55) and the equilibrium prices (54) follow from the demand equations.■

Proof of Corollary 4. First consider $d_C > 0$. Define the quadratic equation

$$g(\lambda_C, K) := \lambda_C^2 + \lambda_C \left[d_C (1-K) - \mu\right] + \frac{\mu}{4}(\mu - 2d_C) = 0, \tag{101}$$

where K is some constant. It is straightforward to verify that the equilibrium demand rates under peering, monopoly, and decentralized control, which are given by (55), (57), and (59), respectively, are the unique solutions of (101), where $K = K^P := \frac{1}{2(1-\alpha)}$ for peering, $K = K^M := \frac{1}{1-\alpha}$ in the monopoly case, and $K = K^D := \frac{2}{1-2\alpha}$ under decentralized control. Note that $0 < K^P < K^M$ for fixed $\alpha \in [0, 1)$ and $K^M < K^D$ for $\alpha \in [0, 1/2)$. Since

$$Q_{Cn}(\boldsymbol{\lambda})|_{\lambda_{Cn}=\lambda_C} > 0 \Longrightarrow \frac{\partial g(\lambda_C, K)}{\partial \lambda_C} = -(\mu - 2\lambda_C - d_C) - d_C K < 0 \tag{102}$$

for $K > 0$, there is for fixed α an implicit function $\lambda_C(K)$, defined for $K \in \left[K^P, K^D\right]$ if $\alpha \in [0, 1/2)$ and for $K \in \left[K^P, K^M\right]$ if $\alpha \in [1/2, 1)$. It satisfies

$$\lambda_C'(K) = -\frac{\partial g(\lambda_C, K)/\partial K}{\partial g(\lambda_C, K)/\partial \lambda_C} = \frac{-d_C K}{\mu - 2\lambda_C - d_C + d_C K} < 0, \tag{103}$$

which establishes the inequalities (60) and (61).
Next consider $d_C = 0$. In this case we have for peering:

$$\Pi_n'(\lambda_{Cn}; \lambda_{Cm}) = (1-\alpha) \cdot V_{Cn}'(\lambda_{Cn}) > 0, \quad n = 1, 2, \tag{104}$$

and it is optimal for each provider to fully utilize the capacity that is not used by the other provider's customers. As a result, any nonnegative arrival rates with $\lambda_{C1} + \lambda_{C2} = \mu$ are an equilibrium. The centralized solution requires in addition $V_{C1}'(\lambda_{C1}) = V_{C2}'(\lambda_{C2})$, i.e., equal marginal values for both cross-traffic flows, which implies $\lambda_{C1} = \lambda_{C2}$ since the network is symmetric. The centralized solution $\lambda_{C1} = \lambda_{C2} = \frac{\mu}{2}$ therefore is a peering equilibrium.■

Proof of Proposition 8. First observe that any equilibrium must be interior. For $n = 1, 2$:

$$\Pi_n'(\lambda_{Cn}; \lambda_{Cm})\big|_{\lambda_{Cn}=0} = (1-\alpha) \cdot V_{Cn}'(0)\, Q_{Cn}(\boldsymbol{\lambda})|_{\lambda_{Cn}=0} - T. \tag{105}$$

In equilibrium it must be that $\Pi_n'(\lambda_{Cn};\lambda_{Cm})|_{\lambda_{Cn}=0} > 0$. For $\lambda_{Cm}=0$ this holds since $V_{Cn}'(0)\,Q_{Cn}(0) > T$ by hypothesis. For $\lambda_{Cm}>0$, this follows since the network is symmetric:

$$\Pi_n'(\lambda_{Cn};\lambda_{Cm})\big|_{\lambda_{Cn}=0,\lambda_{Cm}>0} > \Pi_m'(\lambda_{Cn};\lambda_{Cm})\big|_{\lambda_{Cn}=0,\lambda_{Cm}>0}. \tag{106}$$

Hence, the equilibrium demand rates satisfy $\boldsymbol{\lambda}>0$ and $\lambda_{C1}+\lambda_{C2}<\mu$.

We next show that there is a unique symmetric Nash equilibrium $\lambda_{C1}(T)=\lambda_{C2}(T)=\lambda_C(T)>0$. Let $\lambda_{C1}(\lambda_{C2}) := \arg\max_{\lambda_{C1}\in[0,\mu-\lambda_{C2}]} \Pi_1(\lambda_{C1};\lambda_{C2})$ be provider 1's best response to λ_{C2}. First note that $\lambda_{C1}(\lambda_{C2})$ is unique since $\Pi_1(\lambda_{C1};\lambda_{C2})$ is strictly concave in λ_{C1}. Next note that $\lambda_{C1}'(\lambda_{C2})<0$: since $\Pi_1'(\lambda_{C1};\lambda_{C2})|_{\lambda_{C1}=\lambda_{C2}=0}>0$, $\Pi_1'(\lambda_{C1};\lambda_{C2})|_{\lambda_{C1}=0,\lambda_{C2}=\mu-2d_C}<0$ and $\partial\Pi_1'(\lambda_{C1};\lambda_{C2})/\partial\lambda_{C2}<0$, which implies there is a unique $\overline{\lambda}_{C2}\in(0,\mu-2d_C)$ such that $\Pi_1'(\lambda_{C1};\overline{\lambda}_{C2})\big|_{\lambda_{C1}=0}=0$. Therefore, $\lambda_{C1}(\lambda_{C2})>0$ for $\lambda_{C2}\in[0,\overline{\lambda}_{C2})$ and $\lambda_{C1}(\lambda_{C2})=0$ for $\lambda_{C2}\geq\overline{\lambda}_{C2}$. For $\lambda_{C2}\in[0,\overline{\lambda}_{C2}]$, $\lambda_{C1}(\lambda_{C2})$ is the unique solution of $\Pi_1'(\lambda_{C1}(\lambda_{C2});\lambda_{C2})=0$. By the implicit function theorem and since $\Pi_1(\lambda_{C1};\lambda_{C2})$ is strictly concave, the function $\lambda_{C1}(\lambda_{C2})$ is well defined for $\lambda_{C2}\in[0,\overline{\lambda}_{C2}]$ and

$$\lambda_{C1}'(\lambda_{C2}) = -\frac{\partial\Pi_1'(\lambda_{C1};\lambda_{C2})/\partial\lambda_{C2}}{\partial\Pi_1'(\lambda_{C1};\lambda_{C2})/\partial\lambda_{C1}} \tag{107}$$

$$= -\frac{f(\boldsymbol{\lambda})}{f(\boldsymbol{\lambda})+(1-\alpha)\,V_{C1}''(\lambda_{C1})Q_{C1}(\boldsymbol{\lambda})-(1-\alpha)\,V_{C1}'(\lambda_{C1})\frac{2d_C}{(\mu-\lambda_{C1}-\lambda_{C2})^2}}, \tag{108}$$

where

$$f(\boldsymbol{\lambda}) = -(1-\alpha)\,V_{C1}'(\lambda_{C1})\frac{2d_C}{(\mu-\lambda_{C1}-\lambda_{C2})^2} - \lambda_{C1}V_{C1}'(\lambda_{C1})\frac{4d_C}{(\mu-\lambda_{C1}-\lambda_{C2})^3} < 0.$$

Noting that $\left|\lambda_{C1}'(\lambda_{C2})\right|<1$ and by symmetry $\left|\lambda_{C2}'(\lambda_{C1})\right|<1$, it follows that the mapping $g(\boldsymbol{\lambda}) := (\lambda_{C1}(\lambda_{C2}),\lambda_{C2}(\lambda_{C1}))$ is a contraction and the equilibrium is unique and symmetric.

The fact that the optimal transfer price T^M satisfies (63) follows from substituting the monopoly price P_C^M from (56) into the first order conditions (62), setting $\lambda_{Cn}=\lambda_C^M$ and solving for T. Similarly, the transfer price T^D, given by (64), that corresponds to the decentralized control equilibrium is obtained by substituting the price P_C^D from (58) into the first order conditions (62), setting $\lambda_{Cn}=\lambda_C^D$ and solving for T.■

References

Topkis, D.M. 1978. Minimizing a Submodular Function on a Lattice. *Operations Research* 26 (2) 305-321.