MANAGEMENT SCIENCE
Vol. 55, No. 1, January 2009, pp. 115–131
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5501 | 0115

informs®

DOI 10.1287/mnsc.1080.0918
© 2009 INFORMS

# The Value of Partial Resource Pooling: Should a Service Network Be Integrated or Product-Focused?

Barış Ata, Jan A. Van Mieghem
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208
{b-ata@kellogg.northwestern.edu, vanmieghem@kellogg.northwestern.edu}

We investigate how dynamic resource substitution in service systems impacts capacity requirements and responsiveness. Inspired by the contrasting network strategies of FedEx and United Parcel Service (UPS), we study when two service classes (e.g., express or regular) should be served by dedicated resources (e.g., air or ground) or by an integrated network (e.g., air also serves regular). Using call center terminology, the question is whether to operate two independent queues or one N-network. We present analytic expressions for the delay distributions and the value of network integration through partial resource pooling. These show how the value of network integration depends on service quality (speed and reliability of service) and demand characteristics (volume averages and covariance matrix). Our results suggest that network integration is of little value and operating dedicated networks is a fine strategy if the firm primarily serves express requests with high reliability and if the correlation with regular requests is not strongly negative. In contrast, network integration offers significant gains for firms serving primarily regular requests, almost independent of correlation. Our analysis provides the intuition behind these findings in terms of three main drivers of integration value: arrival pooling, the substitution effect, and the correlation effect.

*Key words*: network integration strategy; quality of service; flexible technology; N-network, skill-based routing; queueing
*History*: Accepted by Michael Fu, stochastic models and simulation; received January 26, 2006. This paper was with the authors 1 year and 2 months for 2 revisions. Published online in *Articles in Advance* October 7, 2008.

## 1. Introduction and Summary

In this paper, we study the value of dynamic resource substitution in service systems. Using a stylized analytic model of a firm that serves two separate markets, we investigate whether these two markets should be served by dedicated resources or by one integrated network. Despite Frederick Taylor's quest, there is no one best way to design every operation; rather, the appropriate network design depends on the strategy and market characteristics. Consequently, our main focus is to generate qualitative managerial insight by explaining how the value of integration depends on demand characteristics (including mean and covariances) and service guarantees. Although our model is too stylized to serve as a precise decision support tool, it still represents a notoriously hard problem in queueing networks that is intractable via exact analysis. We present an approximate analysis that yields closed-form expressions and an intuitive explanation of the impact of key parameters, including covariances. The approximation is appropriate for highly congested systems, which is exactly when resource substitution is useful.

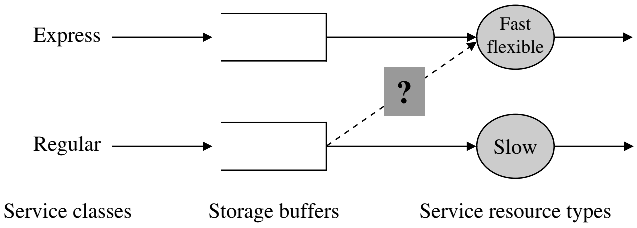
To provide the reader with a concrete example of our research question, we consider two world-class service firms—FedEx and United Parcel Service (UPS)—which operate in the same industry yet use contrasting operating systems, as indicated by the following quotations:

> ...We strongly believe that the optimal way to serve very distinct market segments, such as express and ground, is to operate highly efficient, independent networks with different facilities, different cut-off times and different delivery commitments. (FedEx Corporation 2000)

> Our integrated air and ground network enhances pickup and delivery density and provides us with the flexibility to transport packages using the most efficient mode or combination of modes. (UPS 1999, p. 4)

The essential question here is whether different service classes should be served by separate networks or by an integrated network. Serving different markets by separate (or dedicated) networks allows resource specialization and complexity reduction. In contrast, serving multiple markets with a single network enjoys

**Figure 1** We Study Whether Two Separate Markets Should Be Served by Separate Resources or by an Integrated Network Where Two Markets Dynamically Share the Flexible Resource



economies of scale and economies of scope derived from resource sharing.

To answer our research question, we study the impact of integration using the stylized processing network of Figure 1 that captures resource specialization, resource sharing, and (statistical) economies of scale. In a product-focused network (no dashed arrow in Figure 1) two separate markets, or *service classes*, are processed on demand by their specialized resource type. Because of finite resource capacity, service requests may have to wait in storage buffers before being served. Quality of service, as measured by responsiveness or delay, is a natural key performance indicator for a service operation, and we thus adopt a dynamic model to evaluate its dependence on network design. The express class is time sensitive and is served by a fast resource, whereas the regular class is served by a slow resource. (The fast resource will also be referred to as server 1, and the slow resource will be called server 2. Similarly, the express class is labeled class 1, and class 2 refers to the regular class.) In the integrated network, the slow resource can be dynamically substituted by the fast one, as indicated by the dashed arrow in Figure 1.

The option value of network integration is the incremental performance of the integrated network over the dedicated network. Given that the integrated network can emulate its dedicated counterpart by choosing not to process regular requests with the fast server, the option value is nonnegative. Our analysis seeks to identify the conditions that yield a sufficiently high option value (so that it exceeds the costs of network restructuring, integration, and other complexities, which we do not model). The option value here stems from using occasional excess capacity at the fast server to process regular requests whenever the express queue is empty, thereby reducing the load of the slow resource. (Given that it can serve two service classes, the fast server is also a flexible resource.) We will refer to this occasional and inherently asymmetric resource substitution as partial resource pooling. This associated efficiency gain can be translated either into better quality of service for the regular class if capacities of the resources are kept unchanged, or into investment savings from reduction in slow-server capacity if service quality is kept unchanged. We analyze both. In addition, our third analysis combines both effects by optimizing capacity.

The fundamental tension in the model is driven by two market segments with different service requirements where the resources of serving the more demanding segment are more than adequate for serving the other segment. This applies to many service settings, including call centers, entertainment parks, and professional services such as technical support, health care, and legal advice. Indeed, the central question of this paper can be stated using call center terminology as whether one should operate two independent queues or one N-network. Clearly, the N-network enjoys the benefits of skill-based routing. We investigate whether these benefits are high enough to compensate potential costs of restructuring. To illustrate our findings throughout the paper, we will use the motivating FedEx-UPS example because our model covers two important sources of integration value in their naturally heavily loaded systems. The first one is the dynamic substitution of fast transportation for a slow transportation mode for long-haul traffic originating from one node or hub in the physical network. The fast resources would be airplanes that can serve both markets while the slow trucks would serve only the regular market. The second source of integration value stems from dynamic transportation mode substitution for local traffic around a hub when express requests with nearby destinations could go by truck. That model with dynamic downward substitution is the mirror image of Figure 1, and our analysis can be modified to handle that case. A third source of integration savings in the FedEx-UPS example that is not covered by our model is the reduction in local transportation costs due to increased spatial density of pickups and deliveries. Valuing this source requires a more detailed spatial model of the distribution systems and is studied by Smilowitz and Daganzo (2007). In many ways, our model and that of Smilowitz and Daganzo (2007) are complementary, each focusing on different dimensions of a rich problem. Nevertheless, we will see that both models find similar qualitative conclusions.

To properly value the integrated service network, we must capture the crucial impact of demand correlation on capacity requirements and response times. This is a notoriously hard problem in queueing networks that is intractable via exact analysis. The methodological novelty in our paper is in the sequential application of two powerful analytic approximations: first we adopt a heavy-traffic approximation to derive a correlated Brownian model of the queueing network. Second, we apply a large deviations approximation to the (still intractable) correlated

Brownian model. The final results are closed-form analytical expressions for the tail delay probabilities under heavy, correlated loading, exactly the conditions where integration would derive value. We also illustrate the impact of nonbasic activities in Brownian models, cf. §4.

From a broader perspective, we address the strategic, high-level question of network integration with fairly sophisticated Brownian and large deviations approximations for queueing systems. Although these approximations are too crude to drive precise decisions, they are appropriate for preliminary analysis to generate insight and directionally correct results. Indeed, they yield novel analytic expressions that identify the key drivers and allow comparative static analysis. In addition, this methodology provides a graphical representation to explain nonobvious results in an intuitive manner, as the remainder of this section will summarize.

Integration through partial resource pooling obviously benefits the regular class, so one expects the value of integration to depend on the service mix, defined as the fraction of the regular demand to total demand. To analyze the mix dependence of integration value we consider so-called "express firms," which are firms similar to FedEx that primarily serve express requests, and "regular firms," which are similar to UPS and primarily serve regular requests. There are two conflicting forces that determine which type of firm would benefit most from integration: A regular firm serves many regular requests, so it has much to gain from integration. At the same time, it has relatively little fast-server capacity, so it has little to gain. Our results show that the first force dominates: network integration offers significant gains for regular firms, almost independent of the correlation between express and regular requests. In contrast, operating dedicated networks is a fine strategy for express firms with high service reliability unless express and regular requests are strongly negatively correlated. Our analysis thus supports the different network strategies of FedEx and UPS.

To provide intuition behind these findings, we identify three main drivers of value of integration: arrival pooling, the substitution effect, and the correlation effect. Arrival pooling means that the standard deviation $\sigma$ of the arrival process grows sublinearly with the mean arrival rate $\lambda$ because more opportunities exist for independent, individual fluctuations to partially offset each other. (For example, a square-root relationship holds for Poisson arrivals.) The fact that high-volume arrivals feature relatively less uncertainty reduces the relative amount of capacity $\mu$ needed for a given service level. To capture these statistical economies of scale, we introduce, inspired by inventory theory, the concept of standardized excess capacity $z_i = (\mu_i - \lambda_i)/\sigma_i$ for service class $i$ as the comparable measure of excess capacity. We show that the standardized excess capacity to support a given service level decreases in the arrival rate $\lambda_i$. This means that an express firm has $z_1 \ll z_2$ whereas the reverse relationship applies to a firm serving primarily regular requests. Thus, although a regular firm indeed has little fast-server capacity compared to slow-server capacity, the relevant comparison is in terms of standardized excess capacity, which has the opposite order: $z_1 \gg z_2$.

The substitution effect refers to the frequency of resource substitution, which is greatly impacted by the ratio of the standardized excess capacities $z_1/z_2$. Consider the two opposite cases of large and small ratio $z_1/z_2$, as illustrated in Figure 2. Our analysis will show that $z_1/z_2$ determines the slope of the drift vector of the Brownian motion that approximates the queue length vector $Q$ and impacts the stationary distribution of the queue length process. Intuitively, the queus will be more likely to change in the direction of the drift vector. This means that the queue vector is more likely to hit the vertical axis when the standardized excess capacity of the flexible resource

**Figure 2** **Standardized Excess Capacity Ratio $z_1/z_2$ Determines the Queue Drift Vector and the Likelihood of Hitting Boundary $Q_2 = 0$ When Resource Substitution Is in Effect**

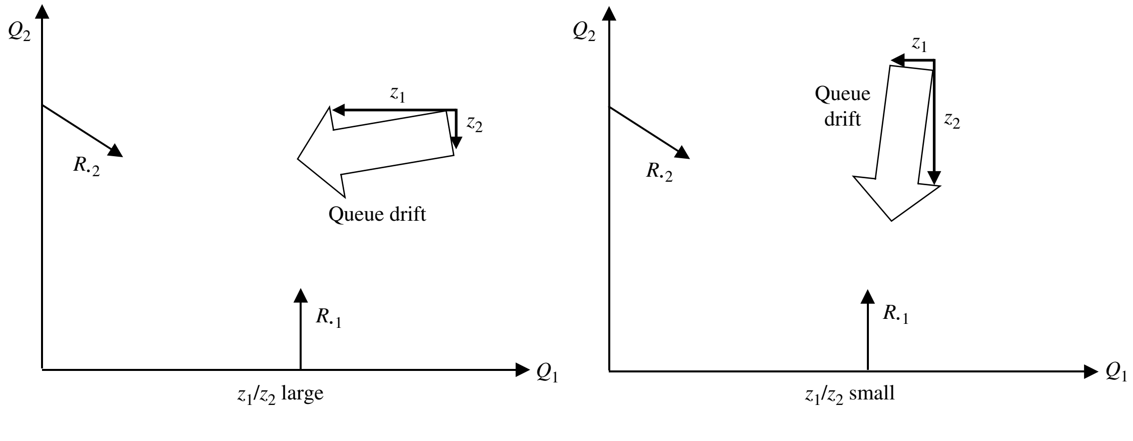
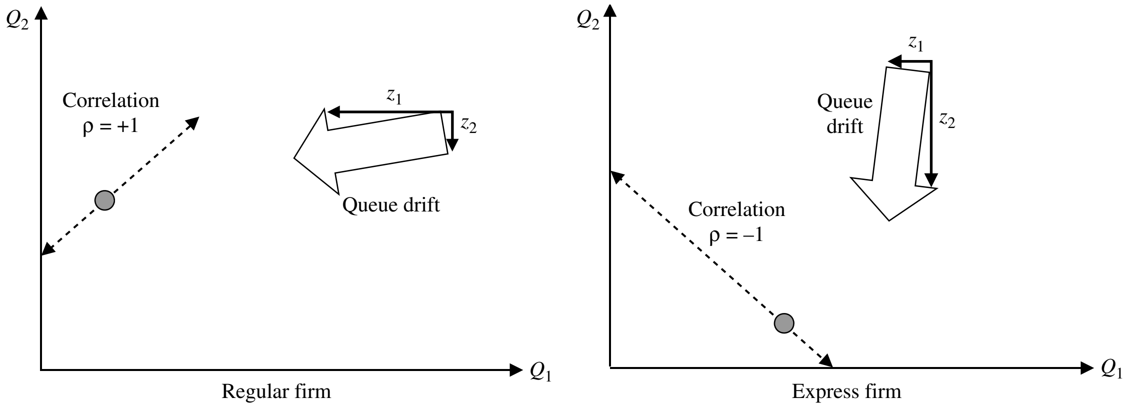


**Figure 3** With Highly Positive Correlation, Both Queues Tend to Move in Lockstep Along the Line with Positive Slope (Left) and Frequently Hit the Vertical Axis for Regular Firms. Express Firms Require Strongly Negative Correlation for the Queues to Hit the Vertical Axis and Enjoy Dynamic Substitution



is high relative to that of the slow resource (i.e., $z_1/z_2$ is large). Then the express queue will be empty quite often, resulting in frequent substitution of the fast resource for its slow counterpart.[1] In contrast, when $z_1/z_2$ is small, the horizontal axis will be visited more often, the regular queue will be empty more often, and resource substitution will be seldom. Combining this substitution effect with arrival pooling explains why the option value of resource substitution is high for a regular firm like UPS (which has large $z_1/z_2$) but low for an express firm like FedEx (small $z_1/z_2$).

It is intuitive that the value of dynamic resource substitution increases as the correlations between arrivals for the two classes becomes more negative. We show, however, that the impact of correlation on the value of integration is strongly mix-dependent, and our analysis provides an intuitive explanation. Consider a given queue length vector, as represented by the point in Figure 3. With highly positive correlation, the short-term deviations of both queues tend to move in lockstep along the dashed line with positive slope (left) and will still frequently hit the vertical axis for a regular firm. An express firm, however, requires strongly negative correlation for the queue length to hit the vertical axis frequently and enjoy dynamic substitution. Combining the intuition from the three effects explains the nonobvious result that correlation has a strong (first-order) impact on the value of integration for an express firm but is of secondary importance to a regular firm. In other words, dynamic resource substitution is always valuable for a regular firm but only with strong negative correlation for an express firm.

The outline of this paper is as follows. After reviewing related literature in the next section, we present the integrated service network model in §3. Section 4 derives the Brownian approximation and the large deviations approximation. Our key analytic results for the option value of network integration, the intuition, and managerial implications are provided in §§5–8. Finally, §9 concludes. All proofs as well as a minimal discussion on large deviations analysis are in the online appendix (provided in the e-companion).[2] The online appendix also presents a detailed discussion of the comparative statics of our results and a simulation study.

## 2. Literature Review

In operations management, economies of scope stemming from substitution have typically been studied in an inventory (goods) setting using newsvendor models focusing on transshipment, product substitution, commonality, and flexibility. Van Mieghem and Rudi (2002) provide a unifying model of a newsvendor network and apply it to our network of Figure 1, which features "discretionary substitution" as referred to by the authors. Similar to our results in §8, they confirm the intuition that integration allows the slow-server capacity to be decreased while the fast-server capacity is increased. The difference with our paper is that any newsvendor model takes stockout probabilities as the service criterion and focuses on pooling as the benefit. In contrast, in a service network it is natural to consider the dynamics of processing service requests (rather than of stocking goods) and their associated delay probabilities, which is naturally done using queueing theory.

In queueing theory, our paper contributes to the growing literature on resource pooling. The simplest

[1] The arrow displayed on the vertical axis corresponds to "idling" of the fast server when there are no express requests to process, that is, $Q_1 = 0$. The fast server can then process regular requests, which is represented by the direction of the downward arrow.

[2] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

implementation of resource pooling serves homogeneous markets with one pool of interchangeable resources. Basic textbooks show that this is better than serving each market with a dedicated resource; see, for example, p. 222 in Anupindi et al. (2005). In practice, markets and resources may be heterogeneous, which reduces the value of resource pooling. Whether to use specialized or flexible resources in real service systems such as call centers, retail banking, and health care has been studied by various authors; see, for example, van Dijk (2002) and Gans et al. (2003) and the references therein. The analytic study of heterogeneous resource pooling in queueing systems easily becomes analytically intractable so that approximate analysis becomes a natural resort. A class of approximate models, called Brownian models, was proposed by Harrison (1988), was generalized in Harrison (2000), and is applied here. Brownian models are powerful for studying various aspects of dynamic control in queueing networks, including dynamic resource substitution. Their validity is established through heavy-traffic limit theorems; Whitt (2002) provides an overview. One can study the value of resource pooling using Harrison's framework. In that setting, the best possible performance is achieved under the so-called complete resource pooling assumption, which amounts to assuming that the servers have sufficiently overlapping capabilities and work collectively to the extent that they act as a single super-server in the heavy-traffic limit. That is, processing capacities of *all* resources are exchangeable in the heavy-traffic limit, which naturally leads to excellent system performance; see, for example, Harrison (1988), Harrison and Lopez (1999), Bell and Williams (2001), Ata and Kumar (2005), and Tezcan and Dai (2008). Nevertheless, this assumption is unrealistic in our asymmetric setting and would assume away the problem studied in this paper; see §4.1 for further discussion.

To study the value of integration we need the delay tail probabilities in the integrated network whose exact computation is intractable. We thus resort to a large deviations analysis. Foley and McDonald (2005a, b) study a network similar to ours under Markovian assumptions and derive both rough and sharp asymptotics. Although their approach is more direct than ours, the Markovian assumptions do *not* allow a study of the covariance effects of demand (because Poisson processes are characterized by a single parameter and cannot be negatively correlated). In contrast, our Brownian model resulting from the heavy-traffic approximation captures the entire demand covariance matrix and offers the intuition summarized in §1. Moreover, the large deviations analysis of the resulting Brownian model leads to explicit, simple-to-analyze formulas for the various quantities of interest. Indeed, Avram et al. (2001) study a variational problem that arises in the large deviations analysis of such a Brownian model and characterize its solution explicitly. We specialize their results to our setting and characterize the tail behavior of steady-state delays in our model and closed-form expressions for various quantities of interest.

This paper also relates to network design, contributing to understanding when a process should follow a product layout with dedicated resources or a process layout where products share resources. (Refer to Anupindi et al. 2005 for a general overview and Kulkarni et al. 2004 and Lu and Van Mieghem 2008 for recent analysis.)

Finally, this paper can be applied to transportation systems, as illustrated by our FedEx-UPS example. Our focus is on substitution of transportation modes; see chapter 14 of Chopra and Meindl (2004) for a general overview. Another source of integration gains in transportation systems would derive from enhanced pickup and delivery density, which is studied by Smilowitz and Daganzo (2007). The authors study value of integration and provide a modeling framework for large-scale integrated networks that starts with a mathematical programming approach and then develops a continuous density approximation to minimize facility location and transportation costs. The authors conclude from a case study that the benefits of integration seems to be larger when the regular demand exceeds the express demand. In contrast, our model is much simpler (and hence more tractable), ignoring some aspects of the problem while focusing on response time and correlation between the two types of demand. Our model and that of Smilowitz and Daganzo (2007) are complementary, each focusing on different dimensions of a rich problem. In §§6–8, we will show that the qualitative conclusion of Smilowitz and Daganzo (2007) is supported by our model as well.

## 3. An Integrated Service Network Model

Consider the network illustrated in Figure 1 of two resources serving requests from two different customer classes that arrive randomly over time. Let $\lambda_i$ denote the average arrival rate of request of class $i$. Requests of the express class, suitably labeled first class or class 1, are time-sensitive, meaning that such requests should be served within a short time window. Class 2 is the *regular* class. This service differentiation is manifested by the network's quality-of-service (QoS) guarantees, which promise that a class $i$ request will be processed within a given time window $d_i$. The guarantee means that if the promise is broken, the customer is entitled to a compensation payment $p_i$. Such service guarantees and compensation schemes for service failures are well documented in practice.[3]

Service failures thus induce a service cost onto the network, whose long-run average rate is

$$C = \lambda_1 p_1 \mathbb{P}(D_1 > d_1) + \lambda_2 p_2 \mathbb{P}(D_2 > d_2), \tag{1}$$

where $D_i$ is the steady-state delay of class $i$ requests and $\mathbb{P}(D_i > d_i)$ denotes the steady-state probability of a class $i$ service failure or violation of the delay commitment. We have $d_1 < d_2$.

The network's processing resources are of two types: type 1 are "fast servers" (e.g., airplanes) that are flexible in that they can serve both classes, and type 2 are "slow servers" (e.g., trucks) that can serve only regular requests satisfactorily. The service capacity of resource type $j$ comprises two factors: (1) the time $m_j$ that resource $j$ needs to process a single request and (2) the number of requests $K_j$ that resource $j$ can process in parallel. In our motivating example, $m_1 < m_2$ represent the travel times of an airplane and truck, respectively, and $K_j$ represents their total cargo space. The aggregate service capacity of resource $j$ is then $\mu_j = K_j/m_j$, meaning that it can process up to $\mu_j$ requests per unit of time. This capacity model is easiest explained as a batch process with constant, deterministic capacity, but for the remaining that interpretation is indistinguishable from a conventional server[4] in queueing theory with processing rate $\mu_j$.

The operational decisions involve dynamic routing and sequencing: route regular requests to either the slow or the fast server, and sequence express and regular requests at the fast server. Equivalently, the control is one of dynamic resource allocation: allocate servers to requests. We assume that the system manager implements a simple greedy policy that (i) prioritizes first-class requests at the fast server; (ii) fills up any remaining capacity with regular requests; and (iii) processes remaining regular requests at the slow server, up to its capacity.[5] To describe this policy, it is easiest to assume that the system manager observes the state of the system at discrete points in time, yet the distinction between continuous and discrete review will be immaterial to our results.

At time $t = 1, 2, \ldots,$ the system manager observes the number $N_i(t)$ of class $i$ requests in queue and uses a greedy resource allocation policy that can be described in terms of an activity vector $x$ as follows. Let $x_1(t)$ be the number of express requests served by the flexible resource in period $t$; $x_3(t)$ be the number of regular requests served by the flexible resource; and $x_2(t)$ be the number of regular requests served by resource 2. Upon observing the backlog vector $(N_1(t), N_2(t))$ at time $t$, the system manager allocates resources for the upcoming period as follows:

$$x_1(t) = \min(N_1(t), \mu_1), \tag{2}$$

$$x_3(t) = \min([\mu_1 - N_1(t)]^+, N_2(t)), \tag{3}$$

$$x_2(t) = \min([N_2(t) - x_3(t)]^+, \mu_2). \tag{4}$$

To facilitate future analysis, define the cumulative allocation processes $T_k$ for $k = 1, 2, 3$ as

$$T_k(t) = \sum_{s=0}^{t-1} x_k(s),$$

which are nonnegative and nondecreasing. The capacity constraints can then be expressed as follows: For $0 \le s \le t$,

$$[T_1(t) + T_3(t)] - [T_1(s) + T_3(s)] \le \mu_1(t-s), \tag{5}$$

$$T_2(t) - T_2(s) \le \mu_2(t-s). \tag{6}$$

Let $\alpha_i(t)$ be the number of class $i$ requests that arrive in period $t$. We assume that the arrival process $\{(\alpha_1(t), \alpha_2(t))\}_{t=0}^{\infty}$ is a sequence of independent and identically distributed random vectors with mean vector $\lambda$ and covariance matrix

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

where $\sigma_i$ is the standard deviation of $\alpha_i(0)$ and $\rho \in [-1, 1]$ the correlation coefficient of $(\alpha_1(0), \alpha_2(0))$. Arrivals can be correlated within a time period but not across time. (The reader familiar with queueing theory should bear in mind that $\rho$ here denotes correlation; we shall need no symbol for utilization.) Let $A_i(t)$ denote the cumulative number of class $i$ requests arrived up to time $t$:

$$A_i(t) = \sum_{s=0}^{t-1} \alpha_i(s) \quad \text{for } t = 1, 2, \ldots$$

Assuming the system is empty initially, dynamics of the backlog process can be described as

$$N_1(t) = A_1(t) - T_1(t),$$

$$N_2(t) = A_2(t) - T_2(t) - T_3(t).$$

[3] For example, FedEx offers money-back guarantees for service failures: "…At our option, we will, upon request, either refund or credit your transportation charges in the event of a service failure (which means delivery of your package 60 seconds or more after the published delivery commitment time for the selected service and destination, except as otherwise described in these terms and conditions)" (p. 189 of FedEx Terms and Conditions/FedEx Express U.S. available at http://www.fedex.com/us/services/terms/index.html as of January 6, 2005).

[4] We assume that the travel times (measured as the time until the vehicle completes the delivery and is available for service again) are deterministic and therefore the resulting queueing system has deterministic service times. Our approach can be extended to stochastic service times at the expense of additional notational complexity.

[5] This reflects practice at UPS according to Wright (2006).

Ultimately, we seek to evaluate the performance of this service network in terms of the service cost $C$. This requires the specification of the violation or tail delay probabilities in this queueing network with correlated arrivals and dynamic resource substitution. Unfortunately, such a delay process is not amenable to exact analysis. Therefore, we adopt the more modest goal of deriving an analytic approximation for the service cost. The next section will adopt a heavy-traffic approximation to derive a Brownian system model that is amenable to large deviations analysis and will yield the delay probabilities in closed form.

Basic queueing theory tells us that quality of service is a function of excess capacity and variability. Indeed, the concept of standardized excess capacity introduced in §1 plays a prominent role in our analysis. Recall that for class $i$ the standardized excess capacity is defined as

$$z_i = \frac{\mu_i - \lambda_i}{\sigma_i}, \quad i = 1, 2. \tag{7}$$

It measures the excess capacity in units of the standard deviation of demand and is similar to the $z$-value associated with the safety stock in a single-period inventory model with normal demand.

## 4. The Approximating Brownian Model

The discrete nature of requests makes many queueing models intractable. Often, tractability is enhanced by adopting a heavy-traffic approximation. The essence of that approximation is in rescaling time and state to obtain a simpler model driven by an underlying Brownian motion. This section develops such a Brownian approximation to our service network. We focus on the basic intuition and refer to Harrison (1988, 2000) for an elaborate treatment.

### 4.1. Deriving the Brownian Model

We start by extending the discrete time arrival, queue count, and allocation processes $A$, $N$, and $T$ to continuous time by defining $A(t) = A(\lfloor t \rfloor)$, $N(t) = N(\lfloor t \rfloor)$, and $T(t) = T(\lfloor t \rfloor)$ for $t \geq 0$ and where $\lfloor t \rfloor$ denotes the largest integer not exceeding $t$. Then, we define the cumulative unused capacity $U(t)$ up to time $t \geq 0$ as follows:

$$U_1(t) = \mu_1 t - T_1(t) - T_3(t), \tag{8}$$

$$U_2(t) = \mu_2 t - T_2(t). \tag{9}$$

The Brownian approximation procedure then expresses all processes in terms of mean-centered arrival and server-time allocation processes $\hat{A}$ and $\widehat{T}$, defined as

$$\hat{A}_i(t) = A_i(t) - \lambda_i t,$$

$$\widehat{T}_i(t) = \mu_i t - T_i(t)$$

for $i = 1, 2$ and $t \geq 0$. Rearranging terms yields the following representation of the backlog process:

$$N_1(t) = \hat{A}_1(t) + (\lambda_1 - \mu_1)t + \widehat{T}_1(t), \tag{10}$$

$$N_2(t) = \hat{A}_2(t) + (\lambda_2 - \mu_2)t + \widehat{T}_2(t) - T_3(t). \tag{11}$$

Similarly, the cumulative unused capacity processes can be expressed as

$$U_1(t) = \widehat{T}_1(t) - T_3(t) \quad \text{and} \quad U_2(t) = \widehat{T}_2(t).$$

The Brownian approximation procedure then considers a sequence of closely related systems indexed by a parameter $n$. (A superscript $n$ will be attached to quantities associated with the $n$th system.[6]) The processes in the $n$th system are then scaled to give rise to the approximating Brownian model in the limit as $n \to \infty$. Now fix an arbitrary, large $n$ and define the scaled standardized excess capacity $\theta_i = -\sqrt{n} z_i^n$ for $i = 1, 2$ and define the following scaled[7] arrival and queue-count processes for $i = 1, 2$ and $t \geq 0$:

$$X_i(t) = \frac{1}{\sqrt{n}\sigma_i} \hat{A}_i^n(nt) \quad \text{and} \quad Q_i(t) = \frac{1}{\sqrt{n}\sigma_i} N_i^n(nt). \tag{12}$$

Similarly, define the scaled cumulative unused capacity and server allocation processes:

$$I_i(t) = \frac{1}{\sqrt{n}\sigma_i} U_i^n(nt), \quad i = 1, 2,$$

$$Y_i(t) = \frac{1}{\sqrt{n}\sigma_i} \widehat{T}_i^n(nt), \quad i = 1, 2, \quad \text{and}$$

$$Y_3(t) = \frac{1}{\sqrt{n}\sigma_1} T_3^n(nt).$$

The Brownian approximation is essentially obtained by letting the parameter $n \to \infty$. The scaled processes $Q$ and $I$ then represent the limiting scaled processes and will still be referred to as the queue length and cumulative idleness or unused capacity process, respectively.

Using (10)–(12) we can express the dynamics of the queue length process $Q$ as

$$Q_1(t) = X_1(t) + \theta_1 t + Y_1(t), \tag{13}$$

$$Q_2(t) = X_2(t) + \theta_2 t + Y_2(t) - \frac{\sigma_1}{\sigma_2} Y_3(t) \tag{14}$$

for $t \geq 0$. Similarly, the cumulative unused capacity or idleness process (8)–(9) gives rise to

$$\begin{aligned} I_1(t) &= Y_1(t) - Y_3(t), \quad t \geq 0, \\ I_2(t) &= Y_2(t), \quad t \geq 0. \end{aligned} \tag{15}$$

To justify the Brownian approximation we need some technical assumptions. First, the scaled standardized excess capacity $\theta_i$ must be of moderate value

[6] To be precise, the service rates $\mu_i^n$ vary with $n$ while we keep $\lambda_i$ and $\sigma_i$ constant across systems.

[7] The particular scaling relative to $\sigma_i$ is chosen to yield a standardized Brownian model in accordance with the setup of Avram et al. (2001).

for $i = 1, 2$, which means that $\lambda_i^n$ is close to the capacity $\mu_i^n$. In other words, the arrival rate is near capacity for each resource, which is the *heavy-traffic regime*. This assumption seems to hold in practice, cf. Leonhardt (2005). Second, we also assume that $X(t) + \theta t \approx B(t)$ for $t \geq 0$, where $B$ is a two-dimensional Brownian motion with drift vector $\theta$ and covariance matrix[8]

$$\Gamma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Finally, we require the queue length process $Q$ to be nonnegative and the allocation process $Y$ and the cumulative unused capacity process $I$ to be nonnegative and nondecreasing.

Let us now discuss *our control* in the Brownian model, which leads to further simplification. Recall that our greedy allocation policy (2)–(4) will prioritize class 1 and serve it at capacity $\mu_1$ unless queue 1 is empty. Therefore, in the limit under our greedy policy, $Y_1$ increases only when $Q_1 = 0$ and is proportional to the cumulative time that queue 1 is empty. Only at those times will the fast server process the other queue, also at full capacity unless queue 2 is empty, so as to minimize the unused capacity of the flexible resource in each period. In the heavy-traffic regime, the probability that both queues are empty is negligible, so it is very likely that the regular class backlog is sufficiently large to fill up any remaining flexible capacity (after serving the express backlog). In other words, the unused flexible capacity—and thus also the cumulative unused flexible capacity process $I_1$—would always be negligible. Under our greedy control rule, the limiting (or idealized) Brownian model simply sets $I_1(t) = 0$ for all $t \geq 0$. Such ideal system behavior is quite common in Brownian models as observed by Kelly and Laws (1993, p. 48), "…the important features of good control policies are displayed in sharpest relief."

Our greedy allocation policy thus yields an idealized Brownian model with $I_1(t) = 0$, and thus $Y_1(t) = Y_3(t)$ given Equation (15) for $t \geq 0$. We also know that $Y_2(t) = I_2(t)$ so that $Y_2$ is the cumulative idleness process of the slow server and can increase only when the regular backlog is zero. Eliminating $Y_3$ in (14) finally yields the approximating Brownian model[9] for our service network under the greedy policy: For $t \geq 0$,

$$Q(t) = B(t) + RY(t) \geq 0$$
$$\text{where } R = \begin{bmatrix} 1 & 0 \\ -\sigma_1/\sigma_2 & 1 \end{bmatrix}. \quad (16)$$

Recall that $Y$ represents our greedy control policy, and $Y_i$ is proportional to the cumulative time that server $i$ is not serving class $i$ (for $i = 1, 2$). Formally,

$$Y_i(\cdot) \text{ is nondecreasing and continuous with}$$
$$Y_i(0) = 0 \text{ and increases only when } Q_i = 0. \quad (17)$$

The Brownian model (16)–(17) can be interpreted graphically as follows. The backlog vector $Q$ is the linear combination of the Brownian motion $B$ and the control $Y$. By definition, $Q$ is nonnegative and lives in the nonnegative quadrant $\mathbb{R}_+^2$. Recall that $Y_i$ increases only when $Q_i = 0$. Thus, in the positive quadrant the queues $Q$ are nonempty and behave like the Brownian motion $B$. Only when queue $i$ is empty does control $Y_i$ increase: when $Q$ "hits" a boundary $Q_i = 0$, the control $Y_i$ "pushes" $Q_i$ in the direction specified by the $i$th column of $R$, denoted $R_{\cdot i}$ in Figure 2, to prevent $Q$ from leaving the positive quadrant. The result is that $Q$ behaves as Brownian motion in the interior of the positive quadrant but is reflected on the boundaries in the direction of the corresponding column of matrix $R$, which is aptly called the *reflection matrix*.

### 4.2. Heavy-Traffic Estimate of the Express Service Failure Probability

Estimating the service cost $C$ requires calculating the service failure probabilities $\mathbb{P}(D_i > d_i)$, which can be expressed in terms of the tail probabilities of $Q$ in the approximating Brownian model. First, the snapshot principle of Reiman (1984) states that the stationary delay distribution is approximately equal to the stationary distribution of the scaled queue count:

$$\mathbb{P}(D_i > x) \approx \mathbb{P}\left(\frac{N_i}{\lambda_i} > x\right).$$

[8] This can be justified by a straightforward application of a functional central limit theorem, cf. Whitt (2002).

[9] The Brownian model of a network closely related to ours was studied in Harrison (1998) under the so-called complete resource pooling assumption. That condition basically ensures that the two servers act as one "super-server" giving rise to one-dimensional dynamics. Using Harrison's terminology, the complete resource pooling assumption requires that all possible resource allocations in our model are basic activities. Intuitively, this means that a significant fraction of fast-server capacity is *always* used by and reserved for regular traffic. In other words, the slow-server capacity is not sufficient to handle the incoming regular demand on its own in the long-run, which is an unrealistic assumption in our setting. In our model, the regular class can be served by the fast server only occasionally, i.e., when there is some excess fast capacity. This means that the activity of serving the regular class by the fast server is a nonbasic activity in Harrison's terminology. Moreover, under the complete resource pooling assumption, one can keep all the backlog in one buffer at all times, which in essence assumes away the issues we study in this paper. Therefore, we do *not* assume complete resource pooling, and hence we have a two-dimensional Brownian model. In particular, both queue lengths vary stochastically over time allowing us to model delays associated with each class.

One can think of this approximation as a distributional heavy-traffic extension of Little's law. The second step is to approximate the steady-state distribution of the scaled queue-count process by that of the reflected Brownian motion process:

$$\mathbb{P}\left(\frac{N_i}{\sqrt{n}\sigma_i} > x\right) \approx \mathbb{P}(Q_i > x).$$

Similar approximations have been justified rigorously under heavy traffic by various authors; for example, see Gamarnik and Zeevi (2006) and the references therein. Combining both steps yields:[10]

$$\mathbb{P}(D_i > d_i) = \mathbb{P}\left(Q_i > \frac{\lambda_i d_i}{\sqrt{n}\sigma_i}\right). \tag{18}$$

The only remaining task is to evaluate the marginal stationary distributions $\mathbb{P}(Q_i > x)$ of the reflected Brownian motion $Q$. Given that class 1 receives priority service at server 1, we can easily find the marginal distribution of $Q_1$, cf. Harrison (1985), which yields:

Proposition 1. *The express service failure probability is*

$$\mathbb{P}(D_1 > d_1) = \exp\left\{-2\frac{\lambda_1}{\sigma_1} z_1 d_1\right\}. \tag{19}$$

Unfortunately, the stationary distribution of the two-dimensional reflected Brownian motion $Q$ does not admit a closed-form expression in general, nor does $\mathbb{P}(Q_2 > x)$. Therefore, the next subsection advances an estimate of $\mathbb{P}(Q_2 > x)$ using large deviations theory.

### 4.3. Large Deviations Estimate of the Regular Service Failure Probability

Large deviations theory typically approximates the tail distribution of a random variable by an exponential distribution (Dembo and Zeitouni 1998). For our service network, the sought-after estimate of the steady-state probability $\mathbb{P}(Q_2 > x)$ using a large deviation approximation is:[11]

$$\mathbb{P}(Q_2 > x) = e^{-r_q x} \quad \text{for large } x, \tag{20}$$

where $r_q$ is the (model-specific) large deviations rate.

Deriving a large deviation estimate is done in two steps. First, one must prove that a large deviation principle for $Q_2$ (i.e., an exponential approximation to the tail probability) holds. Majewski (1998b) did this for Brownian models with a large class of reflection matrices. These reflection matrices are referred to as $\mathscr{M}$-matrices in Avram et al. (2001), and our reflection matrix $R$ belongs to this class so that the first step is done. Second, one must solve a variational problem to specify the large deviations rate. Avram et al. (2001) explicitly solve the variational problem for a large class of two-dimensional Brownian models that include ours. The structure of the solution to that problem, however, is strongly parameter dependent, as discussed in Online Appendix A.

Applying the large deviations approximation (20) to (18) gives the sought-after expression

$$\mathbb{P}(D_2 > d_2) = \mathbb{P}\left(Q_2 > \frac{\lambda_2 d_2}{\sqrt{n}\sigma_2}\right) = \exp\left(-r_q \frac{\lambda_2}{\sqrt{n}\sigma_2} d_2\right)$$
$$= \exp\left(-r\frac{\lambda_2}{\sigma_2} d_2\right), \tag{21}$$

where $r = r_q/\sqrt{n}$.[12]

Online Appendix A tailors the general results of Avram et al. (2001) to our Brownian model and shows that the large deviation rate $r$ of (21) can take on four expressions for negative correlation and positive excess capacity.[13] Defining

$$r_1 = \frac{1}{1-\rho^2}\left[\sqrt{z_1^2 - 2\rho z_1 z_2 + z_2^2} + z_2 - \rho z_1\right],$$
$$r_2 = 2z_2 - 4z_1\rho, \qquad r_3 = \frac{2z_1^2}{z_2 - 2z_1\rho} + 2z_2,$$
$$r_4 = \frac{2(\sigma_1 z_1 + \sigma_2 z_2)}{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}\sigma_2$$

gives rise to the following proposition, which is proved in Online Appendix A.3.

Proposition 2. *Assume positive excess capacities and negative correlation (i.e., $\rho \leqslant 0$ and $z > 0$). Then the regular service failure probability $\mathbb{P}(D_2 > d_2) = \exp(-r(\lambda_2/\sigma_2)d_2)$, where $r$ depends on the correlation $\rho$,*

[10] The equality in (18) holds asymptotically as one approaches the heavy-traffic limit. For a precise statement of such an approximation see Gamarnik and Zeevi (2006). An alternative approach is to model service failure probability by (18) directly, which gives rise to appealing closed-form expressions and is supported by heavy-traffic limit theorems.

[11] The equality in (20) holds asymptotically as $x$ gets large. For a precise statement see Avram et al. (2001). Hereafter, whenever we use an equality sign in such statements it is implicit that the statement holds approximately and becomes accurate in the limit.

[12] Majewski (1998a) recently demonstrated that, roughly speaking, one may switch the heavy-traffic and large deviations limits in feedforward networks with deterministic service times, indicating that the rare event behavior of a semimartingale reflecting Brownian motion gives insight into the rare event behavior of the associated heavily loaded queueing network. This justifies the use of large deviations estimates of delay probabilities in the Brownian model to approximate those in the original queueing network.

[13] The appendix shows that similar expressions exist for positive correlation or for negative excess capacity ($z_2$ can conceivably be negative when $z_1$ is strongly positive).

*the variability ratio* $\sigma_1/\sigma_2$, *and the standardized excess capacity ratio* $z_1/z_2$ *as follows*:

$$r = \begin{cases} r_2 & \textit{if } (1-(\sigma_1/\sigma_2)^2)\dfrac{z_1}{z_2} \le 2(\sigma_1/\sigma_2)+2\rho \\ & \textit{and } \dfrac{z_1}{z_2}(4\rho^2-1) \ge 2\rho, \\ \min(r_1, r_3) & \textit{if } (1-(\sigma_1/\sigma_2)^2)\dfrac{z_1}{z_2} \le 2(\sigma_1/\sigma_2)+2\rho \\ & \textit{and } \dfrac{z_1}{z_2}(4\rho^2-1) < 2\rho, \\ \min(r_2, r_4) & \textit{if } (1-(\sigma_1/\sigma_2)^2)\dfrac{z_1}{z_2} > 2(\sigma_1/\sigma_2)+2\rho \\ & \textit{and } \dfrac{z_1}{z_2}(4\rho^2-1) \ge 2\rho, \\ \min(r_3, r_4) & \textit{if } (1-(\sigma_1/\sigma_2)^2)\dfrac{z_1}{z_2} > 2(\sigma_1/\sigma_2)+2\rho \\ & \textit{and } \dfrac{z_1}{z_2}(4\rho^2-1) < 2\rho. \end{cases}$$

## 5. Value of Integration Part A: Keeping Capacities Constant

Having expressed the service failure probabilities—and thus the service cost rate—in terms of model parameters we now proceed with analyzing the value of network integration $V$, which we define as the incremental value of the integrated network over that of the dedicated network.

### 5.1. Three Valuation Assessments

Assessing the value of integration requires a meaningful comparison between two networks. We will present three different comparisons that all confirm our main result. The first comparison keeps capacities in both networks constant, whereas the second keeps quality of service constant. Clearly, the integration value under constant capacities derives from providing better service to the regular class. In contrast, under constant service quality, the value of integration stems from reduced slow-server capacity requirements, as will be shown in §7. The first valuation is relevant when capacity exhibits high irreversibility, and the second is a proxy for a competitive industry with equilibrium service requirements. Relaxing either constraint and solving for a general competitive equilibrium is nonobvious and beyond the scope of this paper. However, our third valuation in §8 gives some insight by comparing two optimally designed networks in a rather restricted setting where a monopolist can optimize both capacity and service without impacting demand.[14] As expected, integration then reduces slow-server capacity but increases fast-server capacity. However, it also improves express service, a result that is less likely in a competitive industry.[15] Thus, although each of the three valuations has different strengths and weaknesses, they reinforce each other and collectively provide evidence of the robustness of our main result.

We start the first value assessment by discussing the resource substitution and arrival pooling effects, which give an intuitive explanation of the subsequent analytic results.

### 5.2. The Resource Substitution Effect

The graphical interpretation of the Brownian model (16)–(17) yields insight into resource substitution. Recall that the fast server can substitute for the slow server only when the express queue is empty. The effect of substitution on queue lengths is captured by $R_{\cdot 1}$ in Figure 2: When the queue length process $Q$ hits the vertical boundary $Q_1 = 0$, the control $Y_1$ pushes the queue length process $Q$ in the direction $R_{\cdot 1} = (1, -\sigma_1/\sigma_2)'$. This corresponds to an empty express queue and the flexible server processing regular requests thereby reducing the regular class queue length $Q_2$. The slope $-\sigma_1/\sigma_2$ represents the fact that one unit of (unused) standardized excess capacity can be substituted for $\sigma_1/\sigma_2$ units of standardized regular capacity.

In assessing the impact of resource substitution, the ratio of the standardized excess capacities $z_1/z_2$ plays an important role. Consider the two opposite cases of large and small ratio $z_1/z_2$, as illustrated in Figure 2. Recall that $z_1/z_2$ determines the slope of the drift vector of the Brownian motion $B$ and impacts the stationary distribution of the queue length process $Q$. Intuitively, the queues will be more likely to change in the direction of the drift vector. This means that the queue length process is more likely to hit the face $Q_1 = 0$ when the standardized excess capacity of the flexible resource is high relative to that of the slow resource (i.e., $z_1/z_2$ is large). Thus, the express queue will be empty quite often, resulting in frequent substitution of the fast resource for its slow counterpart. In contrast, when $z_1/z_2$ is small, the face $Q_1 = 0$ will be visited less often while the other face $Q_2 = 0$ will be visited more often. Then the regular class queue will be empty more often, and resource substitution will be seldom. In conclusion, we expect the option value of resource substitution to be high when $z_1/z_2$ is large.

### 5.3. The Arrival Pooling Effect

Understanding when the value of integration is significant requires comparative statics on the model

[14] Incorporating elastic demand requires a more comprehensive economic model of customer classes' willingness to pay to determine equilibrium customer arrival patterns and resulting quality of service; e.g., see Van Mieghem (2000) for a single-server setting.

[15] It is unlikely that UPS would increase its express service level above that of FedEx.

parameters: demand data in terms of means $\lambda$ and (co)variances, network data in terms of capacities $\mu$, and service guarantees $d$ and $\varepsilon$. To understand the role of variability, which is always important in option values, we must model its dependence on the scale of demand. We introduce the "arrival pooling parameter" $\gamma_i$ to express how arrival variability scales with mean arrival rate. To be more specific, we assume

$$\sigma_i = \lambda_i^{\gamma_i}, \quad \text{where } \tfrac{1}{2} \le \gamma_i < 1. \tag{22}$$

This relationship captures most typical effects, and $\gamma_i$ is a measure of statistical economies of scale or pooling in arrivals. At one extreme, statistical averaging in arrivals leads to variance growing linearly in scale, like with Poisson arrivals. This is captured by $\gamma_i = 1/2$ in our model and implies that the relative impact of variability (or coefficient of variation in arrivals) decreases in volume. Without strong arrival pooling benefits, variance would grow super-linearly as captured by $\gamma_i > 1/2$.

To study the value of integration in a meaningful way while keeping capacities constant, we suggest the following mechanism to set capacities. Consider the dedicated network and set capacities to the minimal level necessary to achieve a given quality of service. Specifically, let this given quality of service level be denoted by the "QoS-pair" $(d_i, \varepsilon_i)$ where $d_i$ specifies the promised delay guarantee or speed and $\varepsilon_i$ measures the reliability in terms of the service failure. That is, we require that

$$\mathbb{P}(D_i > d_i) \leqslant \varepsilon_i \quad \text{for } i = 1, 2. \tag{23}$$

It is natural to assume that the express market is served not only faster, but also with higher reliability: $d_1 \le d_2$ and $\varepsilon_1 \le \varepsilon_2$.

In the dedicated network each class–resource pair can be viewed in isolation so that the tail probability for each class is given by the same simple expression (19). The minimal standardized excess capacity $z_i$ to achieve the QoS pair $(d_i, \varepsilon_i)$ thus is

$$z_i = \frac{\sigma_i}{2d_i\lambda_i}\log\left(\frac{1}{\varepsilon_i}\right). \tag{24}$$

Incorporating the arrival pooling relationship (22) directly yields the following proposition.

Proposition 3. *The minimal standardized excess capacity $z_i$ to achieve the QoS pair $(d_i, \varepsilon_i)$ in a dedicated network with arrival pooling parameter $\gamma_i$ is given by*

$$z_i = \frac{1}{2d_i\lambda_i^{1-\gamma_i}}\log\left(\frac{1}{\varepsilon_i}\right) \quad \textit{for } i = 1, 2. \tag{25}$$

In principle, Proposition 2 provides all of the machinery to investigate the value of integration. However, this requires a case-dependent analysis. Instead, we will focus on two canonical cases of interest to gain insight: an "express firm" that primarily serves express requests (i.e., $\lambda_1 \gg \lambda_2$ and $\lambda_2 \approx 1$) and a "regular firm" that focuses on the regular market (i.e., $\lambda_1 \ll \lambda_2$ and $\lambda_1 \approx 1$). Arrival pooling suggests that $z_1/z_2$ is strongly impacted by the mix: Proposition 3 shows that $z_1/z_2$ is indeed proportional to $\lambda_2^{1-\gamma_2}/\lambda_1^{1-\gamma_1}$ and thus small for an express firm but large for a regular firm. We can now combine the arrival pooling effect with the resource substitution effect:

1. The standardized excess regular capacity for an express firm is much larger than that of its flexible resource. In other words, the ratio $z_1/z_2$ is small and the right panel of Figure 2 applies. Roughly speaking, the regular class queue will be empty more often than the express class queue so that the substitution frequency will be small.
2. The opposite applies to a regular firm. Its standardized excess regular capacity is much smaller than that of its flexible resource. In other words, the ratio $z_1/z_2$ is large and the left panel of Figure 2 applies. Thus, the express queue will be empty quite often, so the resource substitution frequency will be high.

### 5.4. A Simple Upper Bound on Integration Value

When both the dedicated and the integrated networks have the same capacity, and hence identical standardized excess capacity, the integration value $V$ derives from providing better service to the regular class, thereby reducing the service cost. Let $C_D$ and $C_I$ denote the service cost in the dedicated network and integrated network, respectively, given capacities $\mu$. Clearly,

$$C_D = \lambda_1 p_1 \varepsilon_1 + \lambda_2 p_2 \varepsilon_2, \tag{26}$$

$$C_I = \lambda_1 p_1 \varepsilon_1 + \lambda_2 p_2 \mathbb{P}(D_2 > d_2), \tag{27}$$

where the random variable $D_2$ denotes the stationary delay for class 2 in the integrated network given capacities $\mu$. Thus, the value of integration is

$$V = \lambda_2 p_2 \varepsilon_2 F, \quad \text{where } F = 1 - \frac{\mathbb{P}(D_2 > d_2)}{\varepsilon_2}, \tag{28}$$

and it "only" remains to investigate the regular service violation probability. Setting $\mathbb{P}(D_2 > d_2) = 0$ yields a simple upper bound $\overline{V}$ to the value of integration: $V \le \overline{V} = \lambda_2 p_2 \varepsilon_2$. The factor $F \in [0, 1]$ shows how tight the bound is and will be computed analytically. We also have the following bound on the relative value of integration:

$$\frac{V}{C_D} = \frac{\lambda_2 p_2 \varepsilon_2}{\lambda_1 p_1 \varepsilon_1 + \lambda_2 p_2 \varepsilon_2} F \le \frac{\lambda_2 p_2 \varepsilon_2}{\lambda_1 p_1 \varepsilon_1 + \lambda_2 p_2 \varepsilon_2}. \tag{29}$$

To reduce complexity, our analysis will focus on the more important case of negative correlation hereafter. (All calculations can be done for $\rho > 0$, but we will show those results only in the figures.)

Proposition 4. *The regular service violation probability for an integrated firm (keeping capacities constant) that serves primarily express requests (i.e., $\lambda_1 \gg \lambda_2$) with negative correlation ($\rho \le 0$) is*

$$\mathbb{P}(D_2 > d_2) = \exp\left\{-r_2 \frac{\lambda_2}{\sigma_2} d_2\right\} = \varepsilon_2 \varepsilon_1^{-2\rho(\lambda_2/\lambda_1)(\sigma_1/\sigma_2)(d_2/d_1)}.$$

These formulas allow the study of various comparative statics on the value of integration. Although most of that analysis follows immediately from those formulas, the following corollary facilitates our analysis further by providing a Taylor's expansion of the expressions in Proposition 4:

Corollary 1. *The value of integration (keeping capacities constant) for an express firm ($\lambda_1 \gg \lambda_2$ with $\rho \le 0$) is*

$$\begin{aligned} V &= \lambda_2 p_2 \varepsilon_2 \big(1 - \varepsilon_1^{-2\rho(\lambda_2/\lambda_1)(\sigma_1/\sigma_2)(d_2/d_1)}\big) \\ &= 2\rho p_2 \varepsilon_2 \log(\varepsilon_1) \frac{d_2}{d_1} \frac{\lambda_2^{2-\gamma_2}}{\lambda_1^{1-\gamma_1}} + O\left(\frac{\lambda_2^{3-2\gamma_2}}{\lambda_1^{2-2\gamma_1}}\right). \end{aligned} \tag{30}$$

The corollary highlights the effect of correlation, which will be discussed after first presenting the analogous results for a firm that serves primarily regular requests:

Proposition 5. *The regular service violation probability for an integrated firm (keeping capacities constant) that serves primarily regular requests (i.e., $\lambda_2 \gg \lambda_1$) with negative correlation ($\rho \le 0$) is*

$$\begin{aligned} \mathbb{P}(D_2 > d_2) &= \exp\left\{-r_4 \frac{\lambda_2}{\sigma_2} d_2\right\} \\ &= \exp\left\{-\frac{2(\mu_1 - \lambda_1 + \mu_2 - \lambda_2)}{\sigma_T^2} \lambda_2 d_2\right\} \\ &= \varepsilon_2^{\sigma_2^2/\sigma_T^2} \varepsilon_1^{(\sigma_1^2/\sigma_T^2)(\lambda_2/\lambda_1)(d_2/d_1)}, \end{aligned} \tag{31}$$

*where the total demand variance* $\sigma_T^2 = \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2 < \sigma_2^2$.

In stark contrast to an express firm, a regular firm enjoys "pooling benefits": Equation (31) shows that its regular service violation probability is determined by $\mu_1 - \lambda_1 + \mu_2 - \lambda_2$, which is the sum of the excess capacities of the two resources, and by the variance of the total demand $\sigma_T^2$, which is less than $\sigma_2^2$ for the regular firm (given the assumption of negative correlation and $\lambda_2 \gg \lambda_1$). Given that capacities are kept constant, this benefit does not accrue to the express requests. Proposition 5 thus is a mathematically precise statement of what we referred to as partial resource pooling in the introduction. The fact that only a "regular" firm enjoys partial resource pooling agrees with our intuitive explanation of the arrival pooling and the substitution effects. To highlight the impact of correlation on the integration value, consider the following corollary:

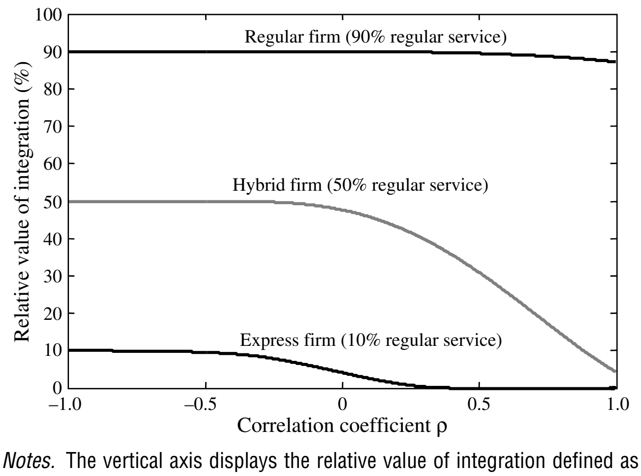
Corollary 2. *The value of integration (keeping capacities constant) for a regular firm ($\lambda_2 \gg \lambda_1$ with $\rho \le 0$) is*

$$\begin{aligned} V &= \lambda_2 p_2 \big(\varepsilon_2 - \varepsilon_2^{\sigma_2^2/\sigma_T^2} \varepsilon_1^{(\sigma_1^2/\sigma_T^2)(\lambda_2/\lambda_1)(d_2/d_1)}\big) \\ &= -p_2 \frac{d_2}{d_1} \log(\varepsilon_1) \lambda_1^{2\gamma_1 - 1} \lambda_2^{2-2\gamma_2} + 2p_2\rho \log(\varepsilon_2) \lambda_1^{\gamma_1} \lambda_2^{1-\gamma_2} \\ &\quad + O\left(\frac{\lambda_1^{3\gamma_1-1}}{\lambda_2^{3\gamma_2-2}}\right). \end{aligned}$$

## 5.5. The Correlation Effect

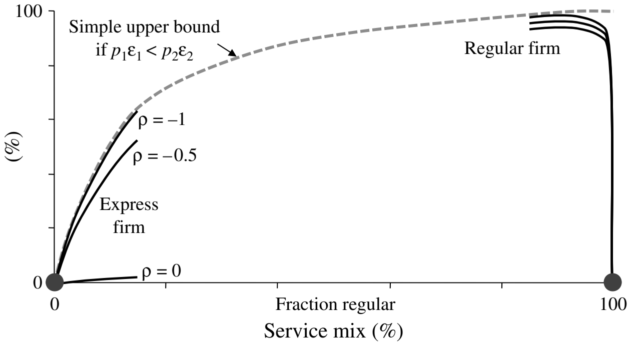
It is intuitive that the value of resource substitution increases as the correlation between arrivals becomes more negative: as regular requests are more negatively correlated with express requests, bursts of regular requests are more likely to find excess fast-server capacity. Our results also show that the option value of integration decreases in correlation, in agreement with many option values in other settings. The strength of this correlation effect, however, is strongly mix-dependent. The three examples in Figure 4 differ in the demand mix and show that the impact of correlation is unimportant for regular firms. That same insight derives from analyzing the quality of the simple upper bound $\overline{V}/C_D$ on the relative value of integration. Figure 5 shows the actual values for both types of firms relative to the simple upper bound (29) as a function of the service mix and parameterized by the correlation. It shows that the relative value is

**Figure 4** **Correlation Between Express and Regular Demand Is a Key Value Driver Unless the Regular Demand Dominates (≥90% of Total Mix)**



*Notes.* The vertical axis displays the relative value of integration defined as $V/C_D$. Data: $p_1 = 10$, $p_2 = 5$, $\varepsilon_1 = 0.01$, $\varepsilon_2 = 0.02$, $d_1 = 2$, $d_2 = 4$, $\lambda_1 + \lambda_2 = 100$, $\gamma = 0.5$.

**Figure 5** **Our Analysis Gives Analytic Expressions for the Relative Value of Integration, Which Strongly Depends on Correlation for an Express Firm**



*Notes.* In contrast, the quality of the simple bound is good for a regular firm for which the relative value of integration is insensitive to the value of correlation. The figure displays the relative value of integration on the vertical axis defined as $V/C_D$ as a function of the service mix (ratio of the regular demand volume to the total demand volume). Data: $p_1 = 10$, $p_2 = 5$, $\varepsilon_1 = 0.01$, $\varepsilon_2 = 0.02$, $d_1 = 2$, $d_2 = 4$, $\lambda_1 + \lambda_2 = 100$, $\gamma = 0.5$.

high for a regular firm, almost independent of correlation, so that correlation is a secondary effect. In contrast, the relative value for an express firm is strongly dependent on correlation. In other words, the simple upper bound is tight for a regular firm but can be far off for an express firm, depending on the value of correlation, in line with the intuitive explanation provided in the introduction. Corollary 1 confirms that correlation has a first-order impact on value for an express firm whereas it has only a second-order effect for a regular firm (Corollary 2).

## 6. The Main Result and Comparative Statics

In this section, we present the first main result on the value of integration and various comparative statics. One could conjecture that, given that a regular firm has relatively little fast-server capacity, the impact of integration is much less than that for an express firm. However, that conjecture is false as Theorem 1 will show. Let the superscript $e(r)$ denote a firm that primarily serves express (regular) requests. For a meaningful comparison, we assume that total demand rates for the two firms are comparable, that is, $\lambda_1^e + \lambda_2^e \approx \lambda_1^r + \lambda_2^r$.

Theorem 1. *The value of integration* (*keeping capacities constant*) *is higher, in both absolute and relative terms, for a regular firm than for an express firm* (*assuming* $\rho \le 0$).

To see the intuition behind Theorem 1, consider the three main drivers of value introduced earlier: arrival pooling, the substitution effect, and the correlation effect. Our discussion of arrival pooling suggests that the standardized excess regular capacity for an express firm is much larger than that of its flexible resource. In other words, the ratio $z_1/z_2$ is small and the right panel of Figure 2 applies. Consequently, the regular queue will be empty more often than the express class queue, so the substitution frequency will be small. Combining this with our discussion of the correlation effect we conclude that the value of integration will be small for an express firm unless the demand for the two classes is strongly negatively correlated. In contrast, the arrival pooling effect suggests that a *regular firm*'s excess regular capacity is much smaller than that of its flexible resource. That is, the ratio of $z_1/z_2$ is large and the left panel of Figure 2 applies. Thus, the express queue will be empty quite often, so the resource substitution frequency will be high, resulting in a high value of integration. Moreover, because there is a strong resource substitution effect, the impact of correlation is of second order.

**Table 1** **Impact of Key Drivers on the Option Value of Integration Given Equal Capacities in Both Networks**

| | Change in express class | Change in regular class |
|---|---|---|
| Guaranteed speed increase | ↑ | ↓ |
| Reliability increase | ↑ | ↓ (for express firm)<br>↑ (for regular firm) |
| Correlation increase | ↓ | ↓ |
| Variance increase | ↑ | ↓ |
| Volume increase | ↓ | ↑ |
| Combined volume-variance increase | ↓ (for express firm)<br>↑ (for regular firm) | ↑ |

*Notes.* Each row of the table corresponds to a parameter, and each column corresponds to a particular demand class so that each entry of the table is associated with a parameter of a particular class. In particular, each entry of the table displays whether the value of integration increases (denoted by the arrow ↑) or decreases (denoted by the arrow ↓) as the parameter (for the associated demand class) corresponding to that entry increases.

The theorem offers a possible explanation behind the differences in network strategy of FedEx and UPS, as discussed in the introduction. (The next sections will provide two additional explanations.) Theorem 1 results from the closed-form expressions provided in §5, which also allow us to study the comparative statics as summarized in Table 1. The discussion and intuition behind these comparative statics can be found in Online Appendix B.

## 7. Value of Integration Part B: Constant Service Quality

In this section, we assess the value of integration when the service quality is kept the same in both the integrated network and the dedicated network. Equal service failure probabilities imply equal flexible capacities but a regular capacity reduction in the integrated network. The value of integration is thus measured by the degree of reduction in the slow-server capacity. As in the previous section, we consider our two canonical firms and compute their capacity reduction in closed form:

Proposition 6. *The reduction in slow-server capacity for an integrated firm (keeping service quality constant) that serves primarily express requests (i.e., $\lambda_1 \gg \lambda_2$) with negative correlation ($\rho \le 0$) is*

$$\begin{aligned}\Delta\mu_2 &= \mu_2^{ded} - \mu_2^{int} = -2\rho\sigma_2 z_1\\ &= -\rho\sigma_2\frac{\sigma_1}{\lambda_1 d_1}\log\left(\frac{1}{\varepsilon_1}\right) = -\rho\frac{\lambda_2^{\gamma_2}}{\lambda_1^{1-\gamma_1}d_1}\log\left(\frac{1}{\varepsilon_1}\right).\end{aligned}$$

The proposition shows the key drivers of the value of integration when service quality is kept constant: capacity savings decrease in correlation $\rho$ but increase in the flexible excess standardized capacity $z_1$ and in the variability of regular arrivals. The effect of class 1 parameters is similar to that in the previous section because the value is primarily driven by the standardized excess capacity of the flexible server, which depends on the express's class QoS pair $(d_1, \varepsilon_1)$ and its arrival mean and variability (cf. Proposition 3). In contrast, the impact of class 2 is different, and only its variability $\sigma_2$ and correlation with class 1 matter because there is no strong substitution effect for the express firm and the value is driven primarily by the correlation effect. Next, consider a regular firm.

Proposition 7. *The reduction in slow-server capacity for an integrated firm (keeping service quality constant) that serves primarily regular requests (i.e., $\lambda_2 \gg \lambda_1$) with negative correlation ($\rho \le 0$) is*

$$\Delta\mu_2 = \mu_2^{ded} - \mu_2^{int} = (\mu_1 - \lambda_1) + \frac{\sigma_2^2 - \sigma_T^2}{\sigma_2^2}(\mu_2^d - \lambda_2),$$

*where the total demand variance* $\sigma_T^2 = \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2 < \sigma_2^2$.

The following corollary provides a more explicit representation of the capacity reduction for a regular firm, helping us identify the first-order drivers of value.

Corollary 3. *The reduction $\Delta\mu_2$ in the slow-server capacity for a regular firm is given by*

$$\begin{aligned}\Delta\mu_2 &= \frac{\sigma_1^2}{2d_1\lambda_1}\log\left(\frac{1}{\varepsilon_1}\right) - \frac{2\rho\sigma_1\sigma_2 + \sigma_1^2}{2\lambda_2 d_2}\log\left(\frac{1}{\varepsilon_2}\right)\\ &= \frac{\lambda_1^{2\gamma_1 - 1}}{2d_1}\log\left(\frac{1}{\varepsilon_1}\right) - \left[\frac{\rho\lambda_1^{\gamma_1}}{\lambda_2^{1-\gamma_2}d_2} + \frac{\lambda_1^{2\gamma_1}}{2\lambda_2 d_2}\right]\log\left(\frac{1}{\varepsilon_2}\right).\end{aligned}$$

Proposition 7 and Corollary 3 show that the value of integration for a regular firm again depends on the express class primarily; the impact of the regular class is only of second order. As in §5, correlation has a second-order impact whereas the substitution effect has a first-order impact on the value of integration for the regular firm. In contrast, correlation has a first-order impact for an express firm whereas substitution has negligible impact.

**Table 2** **Changes in Value of Integration in Terms of the Regular Capacity Savings When Service Quality in the Integrated and Dedicated Networks Is Equal**

| | Express firm | | Regular firm | |
|---|---|---|---|---|
| | Express class | Regular class | Express class | Regular class |
| Guaranteed speed increase | ↑ | — | ↑ | ↑ |
| Reliability increase | ↑ | — | ↑ | ↑ |
| Correlation increase | ↓ | ↓ | ↓ | ↓ |
| Variance ($\sigma^2$) increase | ↑ | ↑ | ↑ | ↑ |
| Volume ($\lambda$) increase | ↓ | — | ↓ | ↓ |
| Combined volume-variance increase | ↓ | ↑ | ↑ | ↓ |

*Notes.* Each row of the table corresponds to a parameter, and each column corresponds to a particular demand class, so each entry of the table is associated with a parameter of a particular class. In particular, each entry of the table displays whether the value of integration increases (denoted by the arrow ↑) or decreases (denoted by the arrow ↓) as the parameter (for the associated demand class) corresponding to that entry increases.

Interestingly, whether one keeps capacity or quality constant to measure the value of integration, the main result is unchanged: regular firms derive more value from integration than express firms, as Theorem 2 will show below. Of course, to assess the value of integration we need to evaluate the financial gain from a reduction in slow-server capacity. For simplicity, we assume that marginal cost $c_i$ of resource $i$ capacity is constant[16] so that the value of integration (keeping service quality constant) is $V = c_2\Delta\mu_2$. Then Theorem 2 follows from Propositions 6 and 7 immediately.

Theorem 2. *The capacity reduction and the value of integration (keeping service quality constant) is higher, in both absolute and relative terms, for a regular firm than for an express firm (assuming $\rho \le 0$).*

The comparative statics are summarized in Table 2. Comparing with Table 1 shows that the effects of changes in express class parameters are the same in both cases; so is the intuition. In contrast, it is striking at first to see that the effects of changes in regular class parameters (in Table 2) either do not exist or are the exact opposite of those in Table 1. To explain this consider the setting of the previous section where capacities were kept constant. Then, a change in regular class parameters that improves the quality of service implies a lower value of integration simply because there is less room for improvement. In contrast, in the setting of this section the better quality of service in the integrated network of the same change in a regular class parameter (before reducing resource 2 capacity) implies a higher reduction in

[16] Undoubtedly, Theorem 2 holds for more general cost structures.

resource 2 capacity. Therefore, the value of integration is higher.

## 8. Value of Integration Part C: Monopoly Capacity Optimization

Our third assessment considers the value of integration to a monopolist whose market demand is unaffected by changes in quality of service. The monopolist then optimizes resource capacities $\mu_1$ and $\mu_2$ to minimize the sum of service and capacity cost rates:

$$\underset{\mu_1,\,\mu_2\geq 0}{\text{Minimize}}\sum_{i=1}^{2}[\lambda_i p_i \mathbb{P}(D_i > d_i) + c_i\mu_i], \qquad (32)$$

where $c_i$ is the capacity cost rate for resource $i$ per unit of time. The value of integration in this case equals the difference between the optimal total cost for the dedicated and integrated systems. The solution to the capacity optimization problem (32) is straightforward for the dedicated network and is given by the following proposition, where the superscript *ded* or *int* denotes optimal quantities in the dedicated or integrated network, respectively.

Proposition 8. *The optimal standardized excess capacities and service failure probabilities for the dedicated network are*

$$z_i^{ded} = \frac{1}{2d_i}\frac{\sigma_i}{\lambda_i}\log\left(\frac{2p_i d_i}{c_i}\left(\frac{\lambda_i}{\sigma_i}\right)^2\right) \quad \textit{and}$$

$$\mathbb{P}^{ded}(D_i > d_i) = \frac{c_i}{2p_i d_i}\left(\frac{\sigma_i}{\lambda_i}\right)^2 \quad \textit{for } i = 1, 2.$$

For an integrated network, however, the optimization problem (32) is not analytically tractable in general. Luckily, the analysis simplifies for the canonical cases of the express and regular firms, which allow explicit solutions:

Proposition 9. *Consider a firm serving primarily express requests* (*i.e.*, $\lambda_1 \gg \lambda_2$) *with negative correlation* ($\rho \leq 0$). *Its optimal standardized excess capacities and service failure probabilities for the integrated network are*

$$z_1^{int} = \frac{1}{2d_1}\frac{\sigma_1}{\lambda_1}\log\left(\frac{2p_1 d_1}{c_1 + 2\rho c_2(\sigma_2/\sigma_1)}\left(\frac{\lambda_1}{\sigma_1}\right)^2\right) > 0,$$

$$z_2^{int} = \frac{1}{2d_2}\frac{\sigma_2}{\lambda_2}\log\left(\frac{2p_2 d_2}{c_2}\left(\frac{\lambda_2}{\sigma_2}\right)^2\right) + \frac{\rho}{d_1}\frac{\sigma_1}{\lambda_1}$$

$$\cdot\log\left(\frac{2p_1 d_1}{c_1 + 2\rho c_2(\sigma_2/\sigma_1)}\left(\frac{\lambda_1}{\sigma_1}\right)^2\right) > 0,$$

$$\mathbb{P}^{int}(D_1 > d_1) = \frac{c_1 + 2\rho c_2(\sigma_2/\sigma_1)}{2p_1 d_1}\left(\frac{\sigma_1}{\lambda_1}\right)^2,$$

$$\mathbb{P}^{int}(D_2 > d_2) = \frac{c_2}{2p_2 d_2}\left(\frac{\sigma_2}{\lambda_2}\right)^2.$$

Comparing Propositions 8 and 9 shows that integration induces a monopolist express firm to decrease the optimal slow-server capacity while increasing the optimal flexible server capacity. Van Mieghem and Rudi (2002) also observed this effect, but our explicit solutions show how the capacity substitution depends on the coefficient of variation and correlation. It is as if the effective cost of a fast server in the integrated network has decreased by $2\rho c_2(\sigma_2/\sigma_1)$. Moreover, the optimal service quality for the express class is higher under integration, whereas the optimal service quality for the regular class is unchanged. This changes for a regular firm:

Proposition 10. *Consider a firm serving primarily regular requests* (*i.e.*, $\lambda_2 \gg \lambda_1$) *with negative correlation* ($\rho \leq 0$). *Its optimal standardized excess capacities and service failure probabilities for the integrated network are*

$$z_1^{int} = \frac{1}{2d_1}\frac{\sigma_1}{\lambda_1}\log\left(\frac{2p_1 d_1}{c_1 - c_2}\left(\frac{\lambda_1}{\sigma_1}\right)^2\right) > 0,$$

$$z_2^{int} = \frac{1}{2d_2}\frac{\sigma_T^2}{\sigma_2\lambda_2}\log\left(\frac{2p_2 d_2}{c_2}\left(\frac{\lambda_2}{\sigma_T}\right)^2\right) - \frac{1}{2d_1}\frac{\sigma_T^2}{\sigma_2\lambda_1}$$

$$\cdot\log\left(\frac{2p_1 d_1}{c_1 - c_2}\left(\frac{\lambda_1}{\sigma_1}\right)^2\right) > 0,$$

$$\mathbb{P}^{int}(D_1 > d_1) = \frac{c_1 - c_2}{2p_1 d_1}\left(\frac{\sigma_1}{\lambda_1}\right)^2,$$

$$\mathbb{P}^{int}(D_2 > d_2) = \frac{c_2}{2p_2 d_2}\left(\frac{\sigma_T}{\lambda_2}\right)^2.$$

Although integration also induces a monopolist regular firm to decrease slow-server capacity while increasing fast-server capacity, there are three distinct differences. First, the effective cost of a fast server in the integrated network has decreased by $c_2$, independent of correlation or variances. Second, the regular class enjoys partial resource pooling: its service quality and the slow-server capacity are driven by the total demand variance $\sigma_T^2$, which is less than $\sigma_2^2$ for a regular firm with $\rho \leq 0$. This yields the third difference with the express firm: a regular firm improves the quality of service for both classes under integration.

Not only do we observe the same partial resource pooling and dependence on correlation as in the preceding two valuation assessments, the main result under monopoly capacity optimization agrees with the earlier results:

Theorem 3. *The value of integration under capacity optimization is higher for a regular firm than for an express firm, in both absolute and relative terms* (*assuming* $\rho \leq 0$).

## 9. Discussion and Limitations

In summary, our results suggest that operating dedicated networks is a fine strategy (or that network integration is of little value) if the firm primarily serves express requests with high reliability and if the correlation with regular requests is not strongly negative. In contrast, network integration offers significant gains if the firm primarily serves regular requests (almost regardless of correlation). Our analysis also reveals three main drivers of value of integration: arrival pooling, the substitution effect, and the correlation effect. Arrival pooling shows that resource substitution will be frequent for a regular firm yet infrequent for an express firm. In particular, a strong substitution effect exists when the standardized excess capacity of the fast server is (much) larger than that of the slow server; that is, $z_1/z_2$ is large. In that case, its magnitude is driven by $z_1\sigma_1/\sigma_2 = (\mu_1 - \lambda_1)/\sigma_2$. Moreover, the correlation effect is unimportant for a regular firm but plays a key role for an express firm, which requires strong negative correlation for integration to be valuable.

The main focus of this paper is to provide structural insights on value drivers. Although one can estimate the value of integration numerically for any given set of parameters by simulation, gaining structural insights through numerical studies becomes exceedingly more difficult as the number of model parameters increases. Our analytic approximations offer structural insights that would be hard to get by other means because our model has 13 parameters. While emphasizing different viewpoints, our three valuation methods yield the same main result, showing robustness of the insights. Nonetheless, we have performed a simulation study (see Online Appendix D) as a sanity check corroborating our analytical results.

To decide whether or not to integrate two networks, the present value of value of integration must be compared with the cost of integration. While our analysis supports the different network strategies pursued by FedEx and UPS, history provides further explanation. Initially, FedEx served only express requests. In 1998, it expanded into the regular class market by acquiring the ground transportation company RPS. In addition to the low value of integration as a (still primarily) express firm, the complexity and high cost of merging infrastructures and processes of two separate firms only reduce the net value of network integration for FedEx. Furthermore, the fact that FedEx Ground employees are contractors makes the integration harder (and perhaps less desirable from FedEx's benefits perspective). In contrast, UPS initially served regular requests only but started in 1982 building its air network organically, which has been integrated with its ground network since the beginning. Therefore, the integration cost seems to be low for UPS. Consequently, it would follow from our analysis and the historical development that the integration costs outweigh the value of integration for FedEx, whereas for UPS the value of integration is significant and probably well exceeds the (low) cost.

Like any stylized model, ours suffers from limitations. We have assumed a stationary regime where demand characteristics and capacity remain constant over time. In reality, UPS and FedEx modulate capacity over time according to predictable variability. A timescale argument suggests that our results may carry over, at least approximately, to the nonstationary case: The relevant transaction timescale of seconds suggests that a stationary analysis is appropriate over time spans of hours while holding capacity constant. It then seems a reasonable approximation to divide the week into "almost stationary" periods and apply our results in each of those periods. The use of overtime would be one such period with higher capacity (which UPS can predict fairly well according to Wright 2006)—note that overtime is an action relevant on the hours timescale, not on the seconds timescale. (This is similar to basic call center analysis, where the day is broken into half-hour segments, and a stationary analysis is performed for each.)

Our model also assumes constant transportation times and focuses on queueing delays at various hubs or sorting centers. Finally, our stylized model does not capture the geographical network structure of FedEx and UPS. Rather, it best represents a bottleneck hub. However, we believe that insights of our analysis carry over to the FedEx-UPS setting because bottleneck hubs are key in determining system performance.

## 10. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

**Acknowledgments**

The authors are indebted to Mustafa Hayri Tongarlak for his assistance with the simulation study. They also thank the editors and reviewers for their constructive feedback.

### References

Anupindi, R., S. Chopra, S. Deshmukh, J. A. Van Mieghem, E. Zemel. 2005. *Managing Business Process Flows*: *Principles of Operations Management*, 2nd ed. Prentice Hall, Upper Saddle River, NJ.

Ata, B., S. Kumar. 2005. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Ann. Appl. Probab.* **15**(1A) 331–391.

Avram, F., J. Dai, J. Hasenbain. 2001. Explicit solutions for variational problems in the quadrant. *Queueing Systems* **37** 261–291.

Bell, S. L., R. J. Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann Appl. Probab.* **11**(3) 608–649.

Chopra, S., P. Meindl. 2004. *Supply Chain Management: Strategy, Planning, and Operation*. Prentice Hall, Upper Saddle River, NJ.

Dembo, A., O. Zeitouni. 1998. *Large Deviations Techniques and Applications*, 2nd ed. Springer, New York.

FedEx Corporation. 2000. FedEx unleashes the power of its brand. (January 19), http://www.fedex.com/us/about/express/pressreleases/pressrelease011900.html?link=4.

Foley, R. D., D. R. McDonald. 2005a. Bridges and networks: Exact asymptotics. *Ann. Appl. Probab.* **15** 542–586.

Foley, R. D., D. R. McDonald. 2005b. Large deviations of a modified Jackson network: Stability and rough asymptotics. *Ann. Appl. Probab.* **15** 519–541.

Gamarnik, D., A. Zeevi. 2006. Validity of heavy traffic approximations in open queueing networks. *Ann. Appl. Probab.* **16** 56–90.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.

Harrison, J. M. 1988. Brownian models of queueing networks with heterogenous customer populations. W. Fleming, P.-L. Lions, eds. *Stochastic Differential Systems, Stochastic Control Theory and Their Applications*, Vol. 10. Springer-Verlag, New York.

Harrison, J. M. 1998. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies. *Ann. Appl. Probab.* 822–848.

Harrison, J. M. 2000. Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probab.* **10**(1) 75–103.

Harrison, J. M., M. J. Lopez. 1999. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems* **33** 339–368.

Kelly, F. P., C. N. Laws. 1993. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Systems* **13** 47–86.

Kulkarni, S. S., M. J. Magazine, A. S. Raturi. 2004. Risk pooling advantages of manufacturing network configuration. *Production Oper. Management* **13**(2) 186–199.

Leonhardt, D. 2005. Have recessions absolutely positively become less painful? *New York Times* (October 8), http://www.nytimes.com/2005/10/08/business/08fedex.html.

Lu, L. X., J. A. Van Mieghem. 2008. Multimarket facility network design with offshoring applications. *Manufacturing Service Oper. Management*, ePub ahead of print April 17, http://msom.journal.informs.org/cgi/content/abstract/msom.1070.0198v1.

Majewski, K. 1998a. Heavy traffic approximations of large deviations of feedforward queueing networks. *Queueing Systems* **28** 125–155.

Majewski, K. 1998b. Large deviations of the steady-state distribution of reflected processes with applications to queueing systems. *Queueing Systems* **29** 351–381.

Reiman, M. I. 1984. Some diffusion approximations with state space collapse. F. Bacelli, G. Fayolle, eds. *Proc. Internat. Seminar on Model. Performance Eval. Methodology, Lecture Notes Control Inform. Sci.* Springer, New York, 209–240.

Smilowitz, K. R., C. F. Daganzo. 2007. Cost modelling and design techniques for integrated package distribution systems. *Networks* **3** 183–196.

Tezcan, T., J. Dai. 2008. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.* Forthcoming.

UPS. 1999. Form S-1/A as filed with the Securities and Exchange Commission on November 5, 1999. Registration 333-83347. http://sec.edgar-online.com/1999/11/05/17/0000940180-99-001306/Section2.asp.

van Dijk, N. M. 2002. To pool or not to pool? The benefits of combining queueing and simulation. E. Yucesan, C.-H. Chen, J. L. Snowdon, J. M. Charnes, eds. *Proc. 2002 Winter Simulation Conf., San Diego*, 1469–1472.

Van Mieghem, J. A. 2000. Price and service discrimination in queuing systems: Incentive compatibility of G$c\mu$ scheduling. *Management Sci.* **46**(9) 1249–1267.

Van Mieghem, J. A., N. Rudi. 2002. Newsvendor networks: Inventory management and capacity investment with discretionary activities. *Manufacturing Service Oper. Management* **4**(4) 313–335.

Whitt, W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York.

Wright, A. 2006. Corporate engineering manager for UPS. Personal conversations on May 23, 2006.