Shared or Dedicated Infrastructure?

On the Impact of Reprovisioning

Abstract

Over the last decade, the information technology (IT) sector has witnessed convergence in the deployment of voice, video, and data services. The trend to integrate operational technology (OT) with existing IT infrastructure is yet another move in that direction. Similarly, as areas such as health-care, facilities management, surveillance, etc., become network-enabled, the potential for convergence on the existing communication infrastructure arises as well. Although convergence, i.e., sharing of a common infrastructure across services, can benefit firms, particularly through cost savings and greater efficiency, combining heterogeneous services on the same infrastructure need not always be the right answer. Sharing can produce complex interactions between services, and the resulting diseconomies of scope can more than offset any of the benefits it affords. This work proposes a model to analyze the trade-off between shared and dedicated infrastructures; thus providing a framework to facilitate managerial decisions. The model accounts for key factors such as potential demand-side synergies between services deployed on the same network, demand uncertainty, (dis)economies of scope in costs, and the ability to dynamically reprovision resources in response to excess demand. The model helps reveal that the extent to which reprovisioning is feasible can affect if and when convergence is beneficial. In particular, it singles out two operational metrics, gross profit margin and return on capacity, that play an important role in the impact of reprovisioning on this decision. The main contribution of this study is in developing a framework that can help decision makers evaluate the potential benefits of convergence.

1. Introduction

One of the major developments in telecommunications and Information technology (IT) in the past decade has been the emergence of convergence, a term used to refer to the ability to carry voice, data and video traffic and multiple IT-enabled services on a single network or IT platform. For example, the Internet has evolved from a simple data network to a global communication infrastructure that carries a multiplicity of services. Similarly, cable operators that previously offered only TV services have expanded their offerings to include telephony and broadband services on the same infrastructure. Other examples include the integration of Operational Technology (OT) consisting of "devices, sensors and software used to control or monitor physical assets and processes" with existing IT infrastructure in enterprises (Roberts and Steenstrup, 2010). Such IT/OT integration can be seen in green buildings with networked sensors and actuators that run their facilities management infrastructure on the existing IT infrastructure (Brandel, 2007). Gartner highlights the management of converged services as one of the critical changes in CIO roles as it "moves from leading the IT delivery organization to leading the systematic, coordinated exploitation of the business assets ... across all technologies" (Mahoney and Steenstrup, 2009).

Convergence can help firms realize several benefits, ranging from cost savings, tighter integration of services and greater efficiency (Cisco Report, 2003). Despite its many obvious advantages, combining services with disparate requirements onto a shared network can also have a cost. It often calls for the entire network to be "upgraded" with features required by only a handful of services, and at a cost that is borne by all of them. Resource sharing can also introduce complex interactions among services and call for tracking and trouble-shooting problems of previously little consequence, e.g., minor routing instabilities don't affect most data services but can severely degrade voice or video quality. Hence, while sharing a network across many services is often advantageous, it need not always be, and it is of interest to determine when and why this is the case or not.

Recent instances of service deployments point to a complex decision process when it comes to deploying multiple services on a single network or platform. For example, in deploying its new U-verse TV service, AT&T chose to create a dedicated network (Yager, 2009) This was in part to ensure it could be managed more easily for better reliability and for delivering higher quality video. In contrast, one of its competitors, Verizon, chose to share a common fiber optic network (Crosby, 2008) for its voice, video, and data services. Similarly, the facilities management infrastructure for green buildings can be set up by piggybacking on existing IT infrastructure of a building (Brandel, 2007), or by creating a dedicated facilities management network (Koebbe, 2007). Brandel (2007) provides an example of the New York public school system that is using a shared IT and facilities management infrastructure to reduce peak energy usage. He also cites Eddie Bauer, a sportswear retailer, which created a dedicated Ethernet backbone for its facilities management traffic because of concerns over costs, throughput and security. Thus, neither shared nor dedicated network choices emerge as an obvious winner in this scenario.

As networking and communication technologies continue to improve and more services become network enabled, e.g., health-care, infrastructure monitoring, surveillance, etc., the question of whether to offer this access over shared or dedicated networks looms large. The question has become even more relevant with the advent of new technologies such as virtualization (Peterson et al., 2005; Touch et al., 2003) and software defined networks (e.g., see http://www.openflow.org), which can further facilitate the deployment of new network "slices" dedicated to an individual new service. Answering the question calls for a framework that systematically examines the trade-off between shared and dedicated network infrastructures. The main motivation for this study is to develop such a framework and to provide managerial guidance on when shared/dedicated infrastructure are desirable for deployment of multiple services.

In this paper, we propose a model for offering two network services, an existing service with a known demand and a new one with uncertain demand that can either be deployed on the same network as the existing service or on its own dedicated network. The model allows for demand-side synergies when the two services are deployed on the same infrastructure, economies/diseconomies of scope from

deploying a new service on an existing network, and the ability to adjust network resources (reprovision) in response to higher than anticipated demand. The model establishes that the extent to which reprovisioning is feasible can affect which infrastructure, shared or dedicated, is more effective. In particular, two operational metrics, the gross profit margin and the return on capacity, play a major role in determining which infrastructure benefits more from reprovisioning. We note that although the problem is cast in the context of networks, the model is equally applicable to other types of infrastructures, e.g., computing, on top of which multiple services can be deployed.[1] The main contribution of this study is in offering a framework for service providers to evaluate infrastructure options, and in particular to decide whether it is profitable to deploy a new service on an existing infrastructure. Given the importance of IT infrastructure in supporting a large number of services in the modern economy, and the relative lack of knowledge about when to deploy on dedicated versus shared infrastructure, our paper helps answer key managerial questions relevant to the providers of IT-enabled services. A recent paper by Tilson et al. (2010) calls for the IS community to put the study of digital infrastructures at the centre of its research endeavor. Our research efforts are consistent with that goal.

The rest of the paper is structured as follows. Section 2 reviews prior works, particularly in the operations management and manufacturing flexibility literature, and highlights their relevance to our research question. Section 3 introduces the model and its parameters. Section 4 presents the analysis. Section 5 summarizes the paper's findings and concludes the study.

2. Literature Review

A number of papers in the Information Systems and Management Science communities have studied operational issues surrounding deployment of networked services. For example, Mendelson and Whang (1990) study pricing in a queuing system with multiple user classes. Gupta et al. (1997a, 1997b) and Zhang et al. (2007) study pricing related to the prioritized transmission of data on the Internet. Keon and

[1] The terms network and infrastructure are used interchangeably throughout the paper.

Anandalingam (2005) study deployment of multiple service classes in telecom networks and explore the use of price discounts as a congestion avoidance scheme. Gupta et al. (2011) analyze the investment incentives for infrastructure owners under the prevalent flat-rate pricing and an alternative congestion-based pricing scheme. Hosanagar et al. (2005, 2008) and Du et al. (2008) study operational aspects of deploying multiple service classes for distributed caches on the Internet. Tawarmalani et al. (2009), Tan et al. (2006), and Hosanagar and Tan (2011) extend that stream of work and focus on cache coordination in distributed caches. Several of these studies consider the problem of supporting multiple service classes on an integrated network. This is analogous to our notion of a shared network in which two different services run on a common network substrate. These studies implicitly assume that the benefits from sharing outweigh the costs of sharing and hence do not analyze the option of dedicated infrastructure. More generally, the topic of "integration" has been both a major motivation and a source of considerable debate (e.g., Steinberg, 1996, Hankins, 1999) in the development of broadband network standards, with Asynchronous Transfer Mode (ATM) and the Internet Protocol (IP) presenting two differing views on how to best achieve such integration. At the core of this debate is the very same question that this paper seeks to address, namely, the extent to which the benefits of integration justify its cost[2]. For example, Triden et al. (2006) investigate this question in the context of safety networks that have stringent quality requirements, and explore the cost trade-offs that integration involves. The model identifies how the outcome could easily be changed based on the weights assigned to different cost components. Fishburn and Odlyzko (1998) offer a slightly different perspective in that it explores both separate and integrated networks but under different pricing configurations that can be used to influence demand. The conclusions are, however, similar in that neither network option is found to be consistently better, and that the outcome is sensitive to assumptions about demand and costs. Our study contributes to this stream of work on networked services operations by formally modeling the cost of operating a dedicated network

[2] The latter clearly depends on the approach used for integration, which is where IP and ATM differ the most.

and providing insights to help managers decide whether to deploy new IT-enabled services on dedicated or shared infrastructure.

The decision of whether to use a shared or dedicated infrastructure for network services involves two main trade-offs. First, for a given network infrastructure (shared or dedicated), capacity sizing under uncertain demand involves trading off the cost of capacity with the cost of loss of some demand. The second trade-off relates to the benefits and costs of sharing. Positive externalities between services deployed on a shared network as well as economies/diseconomies of scope in cost components may serve to favor or oppose sharing. Below, we discuss two streams of research that inform our understanding of these trade-offs and serve a foundational role for our modeling exercise.

Capacity Planning under Uncertain Demand: Capacity sizing for a given infrastructure choice within our setting is in some ways analogous to that of the classical news-vendor problem, which has been studied in a number of papers in Operations Management, e.g., Khouja (1999); Lau (1980). The classical single-period single-product news-vendor problem is to select an inventory/order level for a product under uncertain demand so as to maximize the expected profit in a single period. Both over-provisioning and under-provisioning have associated costs and the inventory level cannot be readjusted if demand exceeds capacity. A rich literature has extended the study of the classical news-vendor to allow for multiple periods (Kogan and Portougal, 2006; Petruzzi and Dada, 1999), multiple products (Abdel-Malek et al., 2004; Erlebacher, 2000; Lau and Lau, 1996; Zhang and Du, 2010), and multi-product multi-period decision problems (Mileff and Nehez, 2007). Alp and Tan (2008) consider capacity sizing with volume flexibility in which firms choose an upfront capacity level but can upgrade total capacity in response to high demand, albeit with a penalty. Similarly, Tomlin (2006) considers contracting in environments in which a firm can source from a supplier with volume flexibility. The notion of volume flexibility in these papers is analogous to reprovisioning in our setting. Although computing optimal capacities under costly reprovisioning is part of our modeling exercise, our main focus is instead to decide between shared and dedicated infrastructures. Traditional newsvendor problems have only one product and the issue of sharing infrastructure does not even arise in that context. Although some papers have

studied multi-product news-vendors, they focus on finding the optimal production quantities of each product under capacity or budget constraints and do not delve into the trade-off associated with servicing the demands for the two products on dedicated versus shared infrastructures.

Manufacturing Flexibility: The manufacturing flexibility literature investigates the trade-off between using flexible resources to manufacture multiple products versus using dedicated resources for each product. Flexible plants capable of producing different types of products are more expensive to build, but have benefits in dealing with uncertain demand. There is, therefore, a trade-off that needs to be investigated to determine how much capacity to build into flexible and dedicated plants. In all these models, investment decisions in manufacturing plants have to be made before the actual demands for products are realized. Fine and Freund (1990) develop a two-stage model to analyze this trade-off. Plant capacity decisions are made in the first stage, when demand is still uncertain. Production decisions are implemented in the second stage after demand is realized. The authors set up an optimization problem to establish the firm's optimal investments in flexible and/or dedicated resources and the optimal production levels using these resources. A similar setting is considered by Van Mieghem (1998), with an emphasis on the role of price margin and cost mix differentials. The author shows that an investment in flexible resources can be beneficial even with perfectly positively correlated product demands because a flexible plant can shift production towards the product with a higher profit margin.

Our decision problem shares some basic properties with these works. Choosing between shared and dedicated networks parallels selecting flexible or dedicated manufacturing plants, as does the need to decide how to provision the network in the face of demand uncertainty. There are, however, several differences between our setup and these earlier works. Most of the manufacturing flexibility papers do not model reprovisioning because production in manufacturing plants usually cannot be rapidly ramped-up in response to higher than expected demand. Goyal and Netessine (2011) is a notable exception in the literature that allows for volume flexibility through capacity upgrades. However, the extent of reprovisioning allowed by technology is not a variable of interest in their analysis. In contrast, "upgrading" network capacity on a relatively short time-scale is becoming increasingly feasible and therefore relevant

to a study of network convergence[3]. We show that it affects not only the optimal capacity levels chosen by providers, but can also impact the decision to go with a shared or dedicated infrastructure. In addition, the manufacturing flexibility literature focuses on the benefits from pooling uncertain demands for two or more products but does not consider the impact of economies and diseconomies of scope in the underlying cost parameters, which is a key aspect of our investigation. Finally, network services deployed on the same infrastructure may demonstrate positive demand-side externalities, e.g. the deployment of IPTV solutions on integrated networks together with the availability of Internet-enabled TV sets is likely to increase demand for both TV and Internet services. While this is a key feature of networked systems, flexible manufacturing resources do not deliver any demand-side externalities and thus the literature does not incorporate such synergies.

The above discussion reveals two themes. First, a number of recent papers in IS have studied operational issues surrounding networked services but none of them have addressed the question of when to deploy services on shared versus dedicated infrastructure. Second, works in Operations Management point towards a useful modeling framework but the models do not directly apply to networked services. Specifically, the literature on news-vendor problems investigates mainly how to size capacity when demand for a product is uncertain. Recent extensions consider multi-product problems but do not delve into the benefits of infrastructure sharing between these products. The manufacturing flexibility literature considers the benefits of resource sharing but focuses primarily on manufacturing settings in which reprovisioning of capacity is often infeasible and flexible infrastructure offer no demand-side externalities. This is a major limitation when considering network services where the reprovisioning of network resources can often be done in short order. Our model builds on modeling frameworks from these streams of work to study the deployment of network services in shared versus dedicated infrastructures, while allowing for demand-side synergies and for the reprovisioning of network resources in response to realized demand.

[3] As mentioned earlier, the advent of virtualization technology and software defined networks will contribute further to this ability.

3. Model Formulation

We consider the most basic setting in which to explore whether to share a network across services, or instead deploy them on dedicated networks. Specifically, one service has already been deployed and has a predictable demand, and the service provider is introducing a second one. There is uncertainty in the demand for the second service, and possible economies or diseconomies of scope when adding it to the same network as the existing service. When the two services are integrated on the same network, there are also positive externalities to the demand of each service that can arise from the value-add of integration. Our goal is to develop a simple model that accounts for these factors in determining the optimal infrastructure choice. For analytical tractability, we ignore any economies of scale that may arise when combining services on a shared infrastructure. The magnitude of such economies of scale are typically limited in networks (e.g., Laoutaris et al., 2009 (ref. Fig. 5), Cisco's Featured Routers Products, 2010). In Appendix B, we show that our results remain qualitatively similar even when we relax this assumption.

The provider's objective is to maximize its total profit from the two services. This decision problem can be modeled as a three stage sequential process, as shown in Figure 1. In the first stage, the provider makes an infrastructure choice, namely a shared or a dedicated network. At this stage, the provider does not know the profit from Service 2 since its demand is uncertain. Given an infrastructure choice, in the second stage the provider provisions capacity. If Service 2 is deployed in a dedicated network, the provider provisions capacity on this network for the yet unknown demand for Service 2. If Service 2 shares the same network with Service 1, the provider provisions flexible capacity in addition to the existing capacity for Service 1 that can be used to support both the unknown demand for Service 2 and any additional demand for Service 1 that results from the externalities of service integration. Demand for Service 2 (and any resulting additional demand for Service 1 in a shared network) is realized in the third stage, where the provider now has the opportunity to reprovision the network if the demand exceeds the capacity provisioned up-front. A penalty for under-provisioning is incurred, and only a fraction of the excess demand can be captured through reprovisioning. Conversely, when the realized demand is lower

than the existing capacity, the provider takes no further action[4]. The three stages of the decision process are referred to as Infrastructure Decision Stage, Capacity Allocation Stage and Reprovisioning Stage, respectively.

[Insert Figure 1 Here]

The above sequential decision problem is solved in the reverse order. We first solve for the provider's decision in the Reprovisioning Stage, i.e., we evaluate whether the provider must reprovision resources after demand is realized, conditional on both the capacity provisioned up-front and the infrastructure choice. Next, we evaluate the provider's expected profit as a function of its capacity sizing decision in stage 2 when demand is uncertain. This is used to compute the optimal capacity to be provisioned up-front. Based on these results for Capacity Allocation Stage and Reprovisioning Stage, we finally evaluate the provider's total expected profit for each infrastructure choice, and select the one that yields the higher expected profit. These three steps are discussed in greater details in Section 3.2 after introducing the model parameters.

3.1 Model Parameters

Given that Service 1 is a mature service with a predictable demand, we assume for simplicity that prior to the introduction of Service 2 it operates at full capacity, i.e., its provisioned capacity matches its realized demand, X_1. The new service, Service 2, has uncertainty in its demand that is denoted by a random variable x_2 with known probability density function, g_{x_2}. We use the notation X_2 to indicate a realization of this demand. If Service 2 shares the same network with Service 1, this service integration can bring additional value or add-on services to the users using both services, which is not available when the two services are on two dedicated networks. As a result of this externality, for the shared network, there will

[4] Contractual obligations are assumed to preclude downward adjustment of resources. This captures the cost of over-provisioning.

be an additional demand of fX_1 for Service 2 and an additional demand of rX_2 for Service 1.[5] In other words, the demands for Services 1 and 2 in a dedicated network are X_1 and X_2, respectively, while the demand for Service 2 in a shared network is $(X_2 + fX_1)$ and the additional demand for Service 1 is (rX_2). The provisioned capacity is a decision variable denoted by K_s and K_{d2} for shared and dedicated networks, respectively[6]. K_s is flexible capacity and can support either service whereas K_{d2} is reserved for service 2. We assume throughout the paper that each user consumes one unit of capacity and, as a result, *per unit* and *per user* are the same. This assumption is not critical and has been made mainly to reduce the number of parameters in the model. If demand exceeds the provisioned level (e.g., $X_2 > K_{d2}$ for the dedicated network), network resources can be adjusted to accommodate a fraction α of the excess demand (i.e., resources are increased to $K_{d2} + \alpha(X_2 - K_{d2})$). The parameter α, or reprovisioning coefficient, reflects the fact that reprovisioning can take time and, as a result, some of the excess demand may be lost. When $\alpha = 0$, reprovisioning is unable, e.g., too slow, to capture any excess demand, while $\alpha = 1$ corresponds to a scenario where reprovisioning succeeds in accommodating the entire excess demand. In other words, when $\alpha = 1$, a "provisioning phase," is unnecessary as resources can be secured on-the-fly. Different levels of provisioning flexibility, e.g., as afforded by different types of virtualization technology, can be accounted for by varying α. This fraction is assumed to be independent of the magnitude of the reprovisioning required. In other words, the latency in securing additional capacity is the same regardless of the amount of capacity requested (at least within some bounds). Note that α is not a decision variable for the provider; it is an exogenous system parameter whose value depends on the reprovisioning technology available to the service provider. Of interest, as discussed in Section 4, is the fact that a change in α can also change the outcome of the provider's decision process, i.e., which infrastructure yields the higher profit.

[5] We assume *f* and *r* satisfy: $fX_1 \leq X_1 - X_1 \cap X_2$ and $rX_2 \leq X_2 - X_1 \cap X_2$. This assumption ensures that the users who would purchase both services in dedicated networks are not double counted when we calculate demand for a shared network.

[6] A shared network is, therefore, provisioned to handle a demand of up to $X_1 + K_s$.

Next, we describe the revenue and cost components of the model. To simplify notation, we consider only the present value of all future revenues and costs. This is similar to the approach in Fine and Freund (1990) and Van Mieghem (1998).

Services generate revenues from subscription fees paid by users. These fees are assumed set based on exogenous market factors. Offering a service also incurs a per user connection cost, e.g., cost of enabling last-mile connectivity, installing end-user access equipment, operational costs of billing, etc. We denote by p_{s1} and p_{s2} the per user contribution margins - price less the variable costs - for Services 1 and 2 respectively in a shared network. Similarly, p_{d1} and p_{d2} denote contribution margins for the two services in dedicated networks. We note that p_{s1} and p_{d1} can differ from each other due to cost implications of sharing. For example, support for voice service in a FiOS network[7] (a shared network used to carry voice, data and video) calls for network termination equipment that is significantly more complex than that used in a traditional voice network, e.g., the FiOS equipment needs to come with a battery pack to handle power outages. This then may translate into $p_{s1} < p_{d1}$. We also note that the model assumes that contribution margins are independent across services, i.e., there is no bundling discount from the same user subscribing to both services, and per user connection costs are also additive across services.

In addition to per-user connection costs, offering network services also involves fixed and capacity costs. Up-front fixed costs are independent of demand and capacity levels, e.g., they include facility rent, research & development expenses. These costs are denoted by c_s for a shared network, and by c_{d1} and c_{d2} when each service is deployed on a dedicated network. Capacity costs grow with network resources; they are incurred up-front because of provisioning and may also be incurred subsequently during reprovisioning. Unit capacity costs in the dedicated networks are denoted by a_{d1} and a_{d2} for Services 1 and 2, respectively, and by a_{s1} and a_s for the existing capacity for Service 1 and the new flexible capacity that can serve both Service 1 and 2 in the shared network. The cost of integrating the

[7] See http://en.wikipedia.org/wiki/Verizon FiOS for an informal description.

existing capacity and the new flexible capacity is included in c_s, a_{s1} and a_s. We use the term return on capacity to refer to the ratio of contribution margin to capacity cost, $\frac{p_i}{a_i}, i = \{d1, s1\}$ for the existing capacity for Service 1, with similar definitions for the dedicated capacity for Service 2 ($\frac{p_{d2}}{a_{d2}}$) and the flexible capacity in a shared network ($\frac{p_{s1}}{a_s}$ and $\frac{p_{s2}}{a_s}$). This metric represents the return from every unit of used capacity.

The values that the above parameters take in shared and dedicated networks are obviously related to each other. These relationships can exhibit different levels of economies and diseconomies of scope. We illustrate this through the example of overlay and integrated networks, which represent two possible options for realizing a shared network.

An overlay involves limited use of an existing infrastructure to deploy a new network service. For example, early versions of the Internet were deployed as an overlay on the existing phone network. End-systems connected using modems to transmit data over existing phone lines, and early routers were interconnected using available telephony transmission facilities such as T1 and T3 links. Control functions of the nascent Internet were, however, kept separate from those of the phone network, e.g., the Internet relied on its own routing protocols and did not use the phone network signaling system (SS7). In general, when a new service is deployed by way of an overlay, the networks of the two services share a common infrastructure (that of Service 1), but remain largely decoupled from each other. This limits the diseconomies of scope that could arise from complex interactions between them, but it also precludes significant economies of scope.

In contrast, an integrated network solution will operate both services on a truly common network infrastructure. For example, many cable providers with a “triple-play” offering have upgraded their infrastructure (backbone network and cable access network) so that it can carry the voice, data, and video traffic from those three services. This required upgrading backbone and access routers to allow differentiation (and prioritization) of different traffic types, but allowed reuse of the same router platforms and transmission facilities for all three services. In other words, an integrated network solution offers

opportunities for greater economies of scope, but often mandates more expensive equipment to handle the individual requirements of each service. This in turn can translate into higher diseconomies of scope in other cost components. Our model can be configured to reflect any combination of economies and diseconomies of scope between shared and dedicated networks.

3.2 Model Setup and Solution

We describe next solving the Three Stage model of Figure 1 to obtain the expected profits associated with shared and dedicated networks. As alluded to earlier, the solution proceeds in the reverse order of the decision process of Figure 1, i.e., the reprovisioning stage is solved first, followed by capacity allocation stage, and finally the infrastructure choice stage. Because the solution method is similar for shared and dedicated networks, we present it for the former and then simply provide final expressions for the latter.

3.2.1 Reprovisioning Stage

As mentioned earlier, reprovisioning takes place after the demand for Service 2 has been realized. In the presence of excess demand, i.e., when the realized demand exceeds the originally provisioned capacity, the provider secures additional capacity to capture a fraction α of the excess demand. In the absence of excess demand, no reprovisioning takes place.

We present next an expression for the gross profit after the reprovisioning phase from deploying Service 2 in a shared network. As defined in Sub-section 3.1, the contribution margins for Services 1 and 2 are p_{s1} and p_{s2}, respectively. The variable cost is a_{s1} for the existing capacity for Service 1, and a_s for the new provisioned flexible capacity (K_s). The original demand for Service 1 (X_1) is exactly met by the existing capacity. The flexible capacity (K_s) can serve both the demand for Service 2 ($X_2 + fX_1$) and the new additional demand for Service 1 (rX_2). If the realized total demand ($X_2 + fX_1 + rX_2$) exceeds the provisioned flexible capacity (K_s), the capacity is adjusted to accommodate a fraction α of the excess demand. i.e., capacity is increased to $K_s + \alpha(X_2 + fX_1 + rX_2 - K_s)$. How the capacity is allocated to the

demand of the two services depends on the contribution margin of the two services. Because Service 2 is new and is the primary service that the provider focuses on in the current setting, we assume $p_{s2} > p_{s1}$. Thus, the gross profit is given by

$$R_s(X_2 + fX_1 + rX_2 > K_s > X_2 + fX_1) = (p_{s2} - a_s)(X_2 + fX_1) + (p_{s1} - a_s)\big(K_s + \alpha(X_2 + fX_1 + rX_2 - K_s) - (X_2 + fX_1)\big) + (p_{s1} - a_{s1})X_1. \quad (1)$$

$$R_s(K_s \leq X_2 + fX_1) = (p_{s2} - a_s)\big(K_s + \alpha(X_2 + fX_1 - K_s)\big) + (p_{s1} - a_s)\alpha r X_2 + (p_{s1} - a_{s1})X_1. \quad (2)$$

Conversely, when the realized demand is less than or equal to the provisioned capacity, the gross profit is

$$R_s(X_2 + fX_1 + rX_2 \leq K_s) = p_{s2}(X_2 + fX_1) + p_{s1} r X_2 - a_s K_s + (p_{s1} - a_{s1})X_1 \quad (3)$$

Similar expressions can be obtained in the case of a dedicated network.

$$R_d(X_2 > K_{d2}) = (p_{d2} - a_{d2})\big(K_{d2} + \alpha(X_2 - K_{d2})\big) + (p_{d1} - a_{d1})X_1 \quad (4)$$

$$R_d(X_2 \leq K_{d2}) = p_{d2} X_2 - a_{d2} K_{d2} + (p_{d1} - a_{d1})X_1 \quad (5)$$

Next, we use these expressions to compute the optimal up-front capacity in the capacity allocation stage.

3.2.2 Capacity Allocation Stage

Assuming a known probability density function g_{x_2} for the demand of Service 2, the expected gross profit R_s given the capacity provisioned up-front K_s in a shared network can be expressed as

$$E(R_s)_{[K_s]} = \int_0^{\frac{K_s - fX_1}{1+r}} R_s(x_2 + fX_1 + rx_2 \leq K_s) g_{x_2}\, dx_2 + \int_{\frac{K_s - fX_1}{1+r}}^{K_s - fX_1} R_s(x_2 + fX_1 + rx_2 > K_s > X_2 + fX_1) g_{x_2}\, dx_2 + \int_{K_s - fX_1}^{X_2^{max}} R_s(K_s \leq x_2 + fX_1) g_{x_2}\, dx_2, \quad (6)$$

where the different $R_s(.)$ functions are given in Equations (1) - (3). Here, we assume that $K_s < X_2^{max} + fX_1$. That is, we focus on the scenarios where the positive demand externalities are not so big that the provider will optimally invest more than the maximum total demand of Service 2.[8]

[8] If we substitute the optimal capacity from Equation (7) in this assumption, it is equivalent to $r\left(\frac{p_{s1}}{a_s} - 1\right) < 1$.

For analytical tractability, we assume that x_2 is uniformly distributed in $[0, X_2^{max}]$. In Appendix B, we show that our results are qualitatively similar under non-Uniform distributions. Under the above assumptions, we can then compute the optimal capacity K_s^* such that $\frac{\partial E(R_s)_{[K_s^*]}}{\partial K_s^*} = 0$:

$$K_s^* = \frac{(1+r)(1-\alpha)(\frac{p_{s2}}{a_s}-1)}{\alpha+(1-\alpha)\frac{p_{s2}}{a_s}+(1-\alpha)(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s})r} X_2^{max} + fX_1 \tag{7}$$

The optimal capacity K_s^* is the one at which the cost incurred from a unit of over-provisioning is balanced against the loss from a unit of under-provisioning. As expected, Equation (7) yields $K_s^* = fX_1$ when $\alpha = 1$, i.e., the ability to reprovision without penalty obviates the need for provisioning up-front beyond the stable demand (fX_1).

Substituting the expression for K_s^* from Equation (7) in Equation (6), we get

$$E(R_s)_{[K_s^*]} = (p_{s2} - a_s)\left(\frac{(1+r)(1-\alpha)^2(\frac{p_{s2}}{a_s}-1)}{\alpha+(1-\alpha)\frac{p_{s2}}{a_s}+(1-\alpha)(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s})r}\frac{X_2^{max}}{2} + fX_1 + \alpha\frac{X_2^{max}}{2}\right) + (p_{s1} - a_s)\alpha r\frac{X_2^{max}}{2} + (p_{s1} - a_{s1})X_1 \tag{8}$$

Similar expressions can be obtained if Service 2 is deployed on a separate dedicated network as shown below

$$K_{d2}^* = \frac{(1-\alpha)(\frac{p_{d2}}{a_{d2}}-1)}{(1-\alpha)\frac{p_{d2}}{a_{d2}}+\alpha} X_2^{max} \tag{9}$$

$$E(R_d)_{[K_{d2}^*]} = (p_{d2} - a_{d2})\left(1 - \frac{(1-\alpha)}{\alpha+(1-\alpha)\frac{p_{d2}}{a_{d2}}}\right)\frac{X_2^{max}}{2} + (p_{d1} - a_{d1})X_1 \tag{10}$$

Next we proceed to use the results of Equations (8) and (10) to compute profits from shared and dedicated networks and finalize a choice of infrastructure.

3.2.3 Infrastructure Choice Stage

In this last stage, the overall profit of the two network options, shared or dedicated, are evaluated to select the one with the higher profit.

In a shared network, the expected profit Π_s is given by subtracting the up-front fixed cost from Equation (8):

$$\Pi_s = (p_{s2} - a_s)\left(\frac{(1+r)(1-\alpha)^2(\frac{p_{s2}}{a_s}-1)}{\alpha+(1-\alpha)\frac{p_{s2}}{a_s}+(1-\alpha)(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s})r}\frac{X_2^{max}}{2} + fX_1 + \alpha\frac{X_2^{max}}{2}\right) + (p_{s1} - a_s)\alpha r\frac{X_2^{max}}{2} + (p_{s1} - a_{s1})X_1 - c_s \quad (11)$$

In a dedicated network, the expected profit Π_d is given by subtracting the up-front fixed costs from Equation (10):

$$\Pi_d = (p_{d2} - a_{d2})\left(1 - \frac{(1-\alpha)}{\alpha+(1-\alpha)\frac{p_{d2}}{a_{d2}}}\right)\frac{X_2^{max}}{2} + (p_{d1} - a_{d1})X_1 - c_{d1} - c_{d2} \quad (12)$$

The optimal network infrastructure choice is the one yielding the higher overall profit. In the next section, we explore how this choice is affected by the model parameters. Before proceeding, we first derive in Lemma 1 a number of basic properties on optimal capacity that are used later in the analysis (the proof is available in Appendix A).

Equations (7) and (9) reveal that the optimal capacity critically depends on the return on capacity metrics ($\frac{p_{d2}}{a_{d2}}$, $\frac{p_{s2}}{a_s}$ and $\frac{p_{s1}}{a_s}$), the demand externalities parameters (r and f), and the reprovisioning coefficient (α).

Lemma 1 *For both shared and dedicated infrastructures, the optimal capacity (weakly) increases with return on capacity ($\frac{p_{d2}}{a_{d2}}$ for dedicated and $\frac{p_{s1}}{a_s}$ and $\frac{p_{s2}}{a_s}$ for shared) and decreases with α. The optimal capacity in the shared infrastructure increases with f and (weakly) increases with r.*

It is very intuitive that a higher return on capacity induces the provider to invest in greater up-front capacity. The optimal capacity is chosen to balance the cost of over-provisioning against the loss from under-provisioning. An increase in α allows a provider to recover more of the excess demand and therefore reduces the provider's cost of under-provisioning resources, resulting in lower up-front capacity.

These apply to both dedicated and shared networks. In the shared network, r and f reflect the increase in demand as a result of the positive externalities of service integration on the same infrastructure. Higher r and f suggest higher demand and therefore require higher capacity.

4. Analysis

In this section, we use the results of Section 3 to study the impact of various system parameters on the choice of network infrastructure. This is done in two phases. First, in Section 4.1, we consider the impact of different cost and revenue parameters. This is relatively straightforward. Second, in Section 4.2, we focus on the impact of the reprovisioning coefficient (α) on network choice, and show that it can produce more subtle and interesting behaviors.

4.1 Impact of Cost/Revenue parameters

The preferred infrastructure is found by comparing Π_s and Π_d (Equations (11) and (12)) of the shared and dedicated networks, respectively, and choosing the one that yields a higher profit. These profits are affected by the cost and revenue parameters in a similar fashion. For example, it can be shown that $\frac{\partial(\Pi_s-\Pi_d)}{\partial p_{s2}} > 0$. As Service 2's contribution margin in the shared network increases, the shared network becomes more profitable and thus is more likely to be the preferred choice. Similarly, we can show that $\Pi_s - \Pi_d$ increases with p_{s1}, a_{d1}, a_{d2}, c_{d1}, and c_{d2}, but decreases with p_{d1}, p_{d2}, a_{s1}, a_s and c_s. These results are all very intuitive and suggest a similar effect, i.e., economics of scope in costs (which leads to higher contribution margin or lower cost in a shared network) favor the shared network while diseconomies of scope favor dedicated networks.

Additionally, it can be shown that $\frac{\partial(\Pi_s-\Pi_d)}{\partial r} > 0$ and $\frac{\partial(\Pi_s-\Pi_d)}{\partial f} > 0$. This suggests that, as expected, a greater increase in demand driven by the positive externalities of service integration favors the shared network over dedicated networks.

4.2 Impact of Reprovisioning

To study the impact of reprovisioning coefficient (α), we substitute the expressions for K_s^* and K_{d2}^* from Equations (7) and (9) into the condition $\Pi_s > \Pi_d$ to obtain:

$$\begin{aligned}(1-\alpha)(p_{d2}-a_{d2})(X_2^{max}-K_{d2}^*)-(1-\alpha)(p_{s1}-a_s)rX_2^{max}-(1-\alpha)(p_{s2}-\\ a_{s2})(fX_1+X_2^{max}-K_s^*)\ >2\gamma,\end{aligned} \tag{13}$$

where γ is independent of α and is given by

$$\begin{aligned}\gamma=(p_{d2}-a_{d2})\tfrac{X_2^{max}}{2}-(p_{s2}-a_s)\left(\tfrac{X_2^{max}}{2}+fX_1\right)-(p_{s1}-a_s)\tfrac{X_2^{max}}{2}r+(p_{d1}-\\ a_{d1})X_1-(p_{s1}-a_{s1})X_1-(c_{d1}+c_{d2}-c_s),\end{aligned} \tag{14}$$

As seen in Equation (14), γ captures the difference in expected profits between the dedicated and shared networks conditioned on capacity exactly meeting the realized demand (as would for example be the case when $\alpha = 1$). The left hand side of Equation (13) captures the difference in the maximum loss from under-provisioning between the dedicated and shared network infrastructures as a function of α. For ease of exposition, we introduce the following notations to denote this difference.

$$h(\alpha) = h_d(\alpha) - h_s(\alpha) \tag{15}$$

$$h_d(\alpha) = (1-\alpha)(p_{d2}-a_{d2})(X_2^{max}-K_{d2}^*) \tag{16}$$

$$h_s(\alpha) = (1-\alpha)(p_{s1}-a_s)rX_2^{max} + (1-\alpha)(p_{s2}-a_s)(fX_1+X_2^{max}-K_s^*) \tag{17}$$

Note that $h(1) = 0$ as under-provisioning can be fully compensated by reprovisioning for both shared and dedicated infrastructures when $\alpha = 1$. In contrast, $h(0)$ can be positive or negative depending on whether the dedicated or shared network incurs a higher loss in the absence of reprovisioning.

As specified in Equation (13), the network infrastructure choice at any value of α depends on the value of $h(\alpha)$ relative to the constant baseline of 2γ. At each value of α where $h(\alpha)$ intersects with 2γ, a switch occurs from preferring one network choice to another. Understanding how reprovisioning affects network choice therefore calls for understanding how the loss difference, $h(\alpha)$, varies with α. This is the topic of Sub-section 4.2.1. In Sub-section 4.2.2, we enumerate the possible intersection(s) of $h(\alpha)$ with 2γ and their implications on network choice.

4.2.1 Analyzing the effect of α on the loss difference $h(\alpha)$

Before we analyze the effect of α on the loss difference between the two networks ($h(\alpha)$), we first examine how the maximum loss from under-provisioning in each network, $h_i(\alpha), i = \{s, d\}$, is affected by α. This is shown in Lemma 2 (the proof is in Appendix A).

Lemma 2 *For both shared and dedicated infrastructures, the maximum loss from under-provisioning* ($h_i(\alpha), i = \{s, d\}$) *decreases with* α.

An increase in α allows a provider to recover more of the excess demand and therefore reduces the provider's cost of under-provisioning resources. On the other hand, according to Lemma 1, an increase in α also induces the provider to reduce the capacity it provisions up-front, thus increasing the excess demand. Lemma 2 suggests that the former effect dominates the latter, so that an increase in reprovisioning (α) benefits both networks by reducing their maximum loss from under-provisioning. In particular, when $\alpha = 1$, i.e., the entire excess demand is captured, there is no loss from under-provisioning, i.e., $h_d(1) = h_s(1) = 0$.

Although Lemma 2 shows that the maximum loss from under-provisioning ($h_i(\alpha), i = \{s, d\}$) decreases with α for both shared and dedicated networks, the difference in these losses, as captured by $h(\alpha)$, may increase or decrease as α varies in $[0, 1]$. Proposition 1 specifies the conditions under which $h(\alpha)$ is increasing (shared benefits more) or decreasing (dedicated benefits more). Proposition 1 (and all the subsequent results) is derived under the condition that $h'(\alpha)$ satisfies the single crossing property (i.e., $h'(\alpha)$ intercepts 0 at most once for $\alpha \in [0, 1]$). This property is violated only in a relatively small parameter region where our main results also hold qualitatively. Therefore, we focus on the primary

parameter region where the single crossing property is satisfied and defer the details for the scenario where this property is not satisfied to Appendix A. [9]

Proposition 1 *Increasing reprovisioning capability* (α) *benefits both shared and dedicated networks. Additionally,*

(*i*) *if* $h'(0) \geq 0$ *and* $h'(1) \geq 0$, *an increase in* α *benefits a shared network more than a dedicated network* $\forall \alpha \in [0, 1]$.

(*ii*) *if* $h'(0) < 0$ *and* $h'(1) < 0$, *an increase in* α *benefits a dedicated network more than a shared network* $\forall \alpha \in [0, 1]$.

(*iii*) *if* $h'(0) \geq 0$ *and* $h'(1) < 0$, *an increase in* α *benefits a shared network more at low* α *and a dedicated network more at high* α.

(*iv*) *if* $h'(0) < 0$ *and* $h'(1) \geq 0$, *an increase in* α *benefits a dedicated network more at low* α *and a shared network more at high* α.

Proposition 1 establishes that the signs of $h'(0)$ and $h'(1)$ characterize the behavior of $h(\alpha)$. Proposition 1 is useful for two reasons. First, it helps identify key operational metrics that determine whether a dedicated or a shared network benefits more from improvements in reprovisioning. Second, it provides a useful graphical aid to understand the factors driving the optimal network choice. We elaborate on both these points below.

Since the sign of $h'(\alpha)$ at $\alpha = 0$ and $\alpha = 1$ determines which network benefits more from reprovisioning, we focus on the relations $h'(0) = 0$ and $h'(1) = 0$:

[9] The exact condition when $h'(\alpha)$ satisfies the single crossing property is given in Appendix A. When the single crossing property is not satisfied, $h'(0) > 0$, $h'(1) > 0$, and an increase in α benefits a shared network more at both low and high values of α and a dedicated network more at intermediate values of α. This only affects result (i) in Proposition 1 and takes place in a relatively small parameter region. It does not affect the other propositions and corollaries as its impact on $h'(\alpha)$ manifests itself only for intermediate values of α in (i). In addition, as shown in Appendix A, our main result, namely, that changes in α affect infrastructure choice and that the optimal choice can switch multiple times as α varies in the range $[0, 1]$ still holds.

$$h'(0) = 0 : \frac{(p_{d2}-a_{d2})}{\left(\frac{p_{d2}}{a_{d2}}\right)^2} = \frac{p_{s2}-a_s+r(p_{s1}-a_s)\left(1+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\left(1+\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)\right)}{\left(\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)^2} \quad (18)$$

$$h'(1) = 0 : p_{d2} - a_{d2} = (p_{s2} - a_s) + r(p_{s1} - a_s) \quad (19)$$

From Equations (18) and (19), we observe that two operational metrics, the return on capacity ($\frac{p_{d2}}{a_{d2}}, \frac{p_{s2}}{a_s}, \frac{p_{s1}}{a_s}$) and the gross profit margin for each unit of used capacity ($p_{d2} - a_{d2}, p_{s2} - a_s, p_{s1} - a_s$), determine which network choice benefits more from increases in α. In Figure 2, we identify the regions in the $\frac{p_{d2}}{a_{d2}}$ (y-axis) and $p_{d2} - a_{d2}$ (x-axis) plane associated with the four conditions from Proposition 1. Note that the y-axis, $\frac{p_{d2}}{a_{d2}}$, only takes values greater than 1 since Service 2 should generate positive profit margin.

[Insert Figure 2 Here]

We observe from Figure 2 that the line $p_{d2} - a_{d2} = (p_{s2} - a_s) + r(p_{s1} - a_s)$, i.e., $h'(1) = 0$, partitions the plane into two regions such that at high α a dedicated network always benefits more on one side and a shared network on the other. This observation is formalized in Proposition 2.

Proposition 2 *A dedicated network benefits more from better reprovisioning at high* α *(i.e.,* $\alpha \to 1$*) if* $p_{d2} - a_{d2} > (p_{s2} - a_{s2}) + r(p_{s1} - a_{s2})$*, and a shared network benefits more at high* α *if* $p_{d2} - a_{d2} < (p_{s2} - a_{s2}) + r(p_{s1} - a_{s2})$*,*

For α close to 1, the difference between the maximum losses from under-provisioning $h(\alpha)$ is mainly determined by the profit margins[10]. When the gross profit margin in a dedicated network is sufficiently high (i.e., $(p_{d2} - a_{d2} > (p_{s2} - a_{s2}) + r(p_{s1} - a_{s2}))$, it has a higher maximum loss from under-provisioning. When $\alpha = 1$, the losses become zero for both network choices (i.e., $h_d(1) = h_s(1) = 0$). Therefore, the network that starts with a higher loss, i.e., the dedicated network, experiences

[10] When $\alpha \approx 1$, $K_{d2}^* \approx 0$, $K_s^* \approx fX_1$, and so $h(\alpha) \approx (1-\alpha)X_2^{max}\left((p_{d2} - a_{d2}) - (p_{s1} - a_s)r - (p_{s2} - a_s)\right)$.

a more significant decrease in its under-provisioning loss as α approaches 1. This explains why the dedicated network benefits more from better reprovisioning when its gross profit margin is high.

The condition in Proposition 2 is less likely to hold when r increases. This follows straightforwardly from Proposition 1, because a higher r increases the threshold for $p_{d2} - a_{d2}$. This suggests that when service integration brings higher additional demand for Service 1 (i.e., when r increases), a shared network is more likely to benefit more from better reprovisioning at high α. This result is summarized in Corollary 1.

Corollary 1 *When r increases, a shared network is more likely to benefit more from better reprovisioning at high α (i.e., $\alpha \to 1$).*

Proposition 2 and Corollary 1 focus on scenarios with high α, and we now turn to scenarios with low α, i.e., very limited reprovisioning. We observe from Figure 2 that $h'(0) = 0$ partitions the plane into two regions. More formally,

Proposition 3 *A dedicated network benefits more from better reprovisioning at low α (i.e., $\alpha \sim 0$) if* $\frac{p_{d2}}{a_{d2}} < \frac{\sqrt{p_{d2}-a_{d2}}\left(\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)}{\sqrt{p_{s2}-a_s+(p_{s1}-a_s)r\left(1+r\left(\frac{p_{s2}}{a_s}-\frac{ps1}{as2}\right)\left(1+\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)\right)}}$, *and a shared network benefits more at low α if* $\frac{p_{d2}}{a_{d2}} > \frac{\sqrt{p_{d2}-a_{d2}}\left(\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)}{\sqrt{p_{s2}-a_s+(p_{s1}-a_s)r\left(1+r\left(\frac{p_{s2}}{a_s}-\frac{ps1}{as2}\right)\left(1+\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)\right)}}$.[11]

[11] $\frac{p_{d2}}{a_{d2}} = \frac{\sqrt{p_{d2}-a_{d2}}\left(\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)}{\sqrt{p_{s2}-a_s+(p_{s1}-a_s)r\left(1+r\left(\frac{p_{s2}}{a_s}-\frac{ps1}{as2}\right)\left(1+\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)\right)}}$ is the solution to Equation (18).

Proposition 3 indicates that in addition to the gross profit margin ($p_{d2}-a_{d2}, p_{s2}-a_s, p_{s1}-a_s$), another metric, return on capacity ($\frac{p_{d2}}{a_{d2}}, \frac{p_{s2}}{a_s}, \frac{p_{s1}}{a_s}$), also affects which network choice benefits more from better reprovisioning at low α. For α close to 0, the difference between the maximum losses from under-provisioning $h(\alpha)$ is determined by both gross profit margins and the difference between maximum demand and optimal capacity. According to Lemma 1, optimal capacity is mainly determined by the return on capacity metric. This explains why both gross profit margin and return on capacity play a role when α is low.

The condition in Proposition 3 is less likely to hold as r increases as long as p_{s1} is not too small. This suggests that when service integration brings higher additional demand to Service 1 (i.e., when r increases), a shared network is more likely to benefit more from an improved reprovisioning capability at low α only if p_{s1} is not too small. The intuition behind this behavior is as follows. For α close to 0, the maximum loss from under-provisioning in a shared network $h_s(\alpha)$ is mainly determined by $(p_{s1}-a_s)rX_2^{max}+(p_{s2}-a_s)(fX_1+X_2^{max}-K_s^*)$. The first term increases with r, but the second term decreases with r since K_s^* increases with r (given by Lemma 1). When p_{s1} is large, the first effect dominates, implying that $h_s(\alpha)$ increases with r, whereas the contrary is true if p_{s1} is small. Since $h_d(\alpha)$ is not affected by r, the shared network is more likely to benefit more from better reprovisioning when r increases, only if p_{s1} is sufficiently large. This result is summarized in Corollary 2.

Corollary 2 *When r increases, a shared network is more likely to benefit more from better reprovisioning at low α (i.e., $\alpha\sim 0$) as long as p_{s1} is not too small.* [12]

Our analysis thus far identifies for given values of the metrics, gross profit margin and return on capacity, which region of Figure 2 we operate in, which network benefits more from better reprovisioning in each region, and how this behavior is affected by demand externalities to Service 1 (i.e., r). The

[12] The proof and the cutoff value for p_{s1} are given in Appendix A.

additional demand that service integration brings to Service 2 does not matter here because fX_1 is based on the known, stable demand for Service 1 so that its capacity is always provisioned up front.

4.2.2 Optimal Network Choice

The analysis in Section 4.2.1 characterizes which network choice benefits more from reprovisioning. However, the provider's optimal network choice depends on how these relative benefits compare to the other cost and revenue parameters. As specified in Equation (13), this choice depends on the value of $h(\alpha)$ with respect to the baseline of 2γ, with each intersection between $h(\alpha)$ and 2γ marking a switch in network choice. In this section, we characterize the provider's optimal network choice.

As specified in Proposition 1 (see also Figure 2) there are four possible behaviors associated with an increase in reprovisioning coefficient α.

First, consider the region in which a shared network always benefits more from increases in α. This corresponds to the upper left region in Figure 2. In this region, if the shared network is already the preferred choice at $\alpha = 0$, then it obviously remains the provider's optimal network choice irrespective of reprovisioning ability. This requires $\gamma < 0$ (because $h(\alpha) > 2\gamma, \forall\alpha$ and $h(1) = 0$), which can arise if the shared network enjoys significantly lower fixed costs (i.e., $c_s \ll c_{d1} + c_{d2}$) or variable costs (($p_{d2} - a_{d2} \ll (p_{s2} - a_s) + r(p_{s1} - a_s)$ and/or $p_{d1} - a_{d1} \ll p_{s1} - a_{s1}$). A numerical example is shown in Figure 3(C). On the other hand, if a dedicated network is initially preferred and if the benefits that the shared network receives from reprovisioning are never sufficient to overcome the impact of other parameters (i.e., $h(\alpha) < 2\gamma, \forall\alpha \in [0, 1]$), then a dedicated network remains preferred irrespective of α, as shown in Figure 3(A) (where a low up-front fixed cost favors a dedicated network). A more interesting outcome arises when a dedicated network is the preferred choice for $\alpha = 0$, but as α increases, the benefits that the shared network receives are sufficiently high to overcome the impact of diseconomies of scope in other costs (i.e., $h(\alpha)$ and 2γ intersect). As a result, the optimal network choice switches to a

shared network at high α. A numerical example is shown in Figure 3(B), in which a dedicated network is preferred for $\alpha \precsim 0.75$ and shared for higher values.

[Insert Figure 3 Here]

Second, consider the region in which a shared network benefits more from increases in α at low α and a dedicated network at high α. This corresponds to the shaded region in the upper right hand side of Figure 2. A numerical example for this scenario is shown in Figure 4 for the case of $h(0) > 0$, which shows that there are four possible network choice outcomes depending on 2γ: (i) a dedicated network is preferred irrespective of α, (ii) a shared network is preferred irrespective of α, (iii) a shared network is preferred at low α and a dedicated network at high α, (iv) a dedicated network is preferred at both low and high α, and a shared network for intermediate values. A dedicated (shared) network is chosen irrespective of α if there are significant diseconomies (economies) of scope as shown in Figure 4(A) (Figure 4(D)). In both cases, the impact of reprovisioning is negligible relative to the impact of other cost and revenue parameters. For values of cost parameters such that $h(\alpha)$ and 2γ intersect, the optimal network choice switches. Moreover, as shown in Figure 4(B) and 4(C), there can be one or two such switches in the optimal network choice. In the case of $h(0) < 0$, there are also four possible network choice outcomes similar to Figure 4. The only difference is that when $h(\alpha)$ and 2γ intersect only once, a dedicated network is preferred at low α and a shared network at high α, opposite to what's shown in Figure 4(C). If $h(0) = 0$, $h(\alpha)$ and 2γ either do not intersect as in Figure 4(A) and 4(D), or intersect twice as in Figure 4(B), i.e., there is no equivalence of Figure 4(C).

[Insert Figure 4 Here]

Lastly, for the other two regions of Figure 2, the analyses are analogous to the previous ones. In the lower right region of Figure 2 where a dedicated network always benefits more than a shared network $\forall \alpha \in [0, 1]$, if the diseconomies (economies) of scope in the costs are very large, a dedicated (shared) network is preferred irrespective of α, otherwise the network choice switches from shared to dedicated as α increases. In the shaded region in the lower left hand side of Figure 2 where a dedicated network

benefits more from increase in α at low α and a shared network at high α, there can be three or four possible outcomes depending on the sign of $h(\alpha)$.

4.3 Discussion

In this section, we illustrate the paper's findings through two examples that help contrast the different possible outcomes the model predicts when the ability to reprovision resources improves (α increases).

Recall that shared infrastructures benefit from reusing equipment across services, but that these benefits can all but disappear when the sharing is poorly controlled and produces diseconomies rather than economies of scope. Technologies that control sharing have, therefore, played an important role in the emergence of shared solutions, e.g., witness the impact of "virtualization" on the growing popularity of both cloud computing and virtual networks. At the same time, technologies that enable better control of shared resources, often also facilitate more dynamic provisioning of those same resources. As seen in the previous section, better reprovisioning capabilities and greater economies of scope, as measured by improvements in gross profit margin and return on capacity, need not always combine to favor shared solutions. We illustrate this next through two examples.

Consider the task of providing computing services in the late eighties, early nineties. There were two major competing options for delivering such services. Systems such as IBM mainframes were representative of shared solutions that would support multiple services (and users). In contrast, DEC mini-computers and later on a wide range of "workstations" were the pillars of dedicated solutions, with individual machines assigned to specific tasks or users. The cost of equipment and therefore computational capacity was substantially lower for dedicated solutions than it was for IBM mainframe shared solutions, i.e., $a_d < a_s$. As a result, and even if services based on IBM mainframes often carried a premium ($p_s \gtrsim p_d$), this scenario maps to the upper right quadrant of Figure 2, i.e., dedicated solutions have high gross profit margins and return on capacity. In this area, better reprovisioning abilities (α) can favor either shared or dedicated solutions. In particular, improving α in the low-α region benefits shared

solutions more than dedicated ones. Small improvements in reprovisioning that technology advances afforded, e.g., through processor and memory upgrades or even additional processor cards, benefited both mainframes and mini-computers, but would have favored (shared) IBM mainframes more than (dedicated) mini-computer-based solutions. This factor, and obviously many others, may have enabled mainframes to survive in spite of the emergence of cheaper distributed solutions.

Contrast the previous situation with the current environment for computing services, where dedicated and shared solutions both rely on the same type of equipment, i.e., a stand-alone blade server fulfills the needs of an individual computation service, while racks of the same blade servers can be shared across services. In addition, previously mentioned technologies such as virtualization offer tight control of resources sharing across services, which enables shared solutions to take full advantage of the economies of scope they afford. The similar equipment costs $a_d \approx a_s$ and the ability to fully leverage the economies of scope of shared solutions ($p_s \geq p_d$) suggest that we might now be operating in the lower left quadrant of Figure 2, i.e., dedicated solutions display both lower gross profit margins and return on capacity. Furthermore, the same technology that is behind stackable blade servers and virtualization makes highly dynamic reprovisioning a reality, i.e., idle CPUs can be rapidly allocated to individual services, and adding new blades to an existing system can be done with little turn-around time. In other words, we are now in a high α environment. Hence, throughout the lower left quadrant of Figure 2, improvements in re- provisioning abilities would take place in a high α environment, and therefore only further the advantage of shared solutions. In other words, unlike the "mainframe vs. workstation" scenario where improving reprovisioning tilted the balance back towards the less competitive mainframe solution, it now further strengthens the solution of choice, shared systems, which augurs well for the continued growth of large-scale cloud computing systems.

Of course, in the absence of detailed estimates for model parameters, it is impossible to demonstrate the exact impact that changes in reprovisioning ability and economies of scope may have on the ultimate success of shared computing solutions, or conversely had on the survival of mainframe based

solutions. Nonetheless, the discussion is meant to illustrate the effects and interactions of these two parameters, and the kinds of analyses that are feasible to understand their impact.

5. Conclusions

This work introduces an analytical framework to investigate which of shared or dedicated infrastructures offer a more cost-effective solution in the deployment of new services. The choice of an infrastructure is influenced by many factors such as fixed and variable costs, capacity costs, demand synergies, and the ability to dynamically reprovision resources. The results demonstrate that although strong economies or diseconomies of scope in the cost components can, as one would expect, favor a shared or a dedicated solution, the ability to dynamically provision resources also has an important effect. Reprovisioning improves the profits of both shared and dedicated solutions, but can do so differently as a function of their respective gross profit margins and returns on capacity. The selection of a preferred infrastructure is, therefore, influenced not only by economies and diseconomies of scope, but also by how the infrastructure is affected by reprovisioning. Additionally, changes in synergies between services that affect demand in a shared infrastructure were also found to have a different impact as a function of the infrastructure's reprovisioning ability.

Although the model demonstrates the impact of reprovisioning and identifies operational metrics that influence infrastructure choice, it relied on a number of assumptions that we briefly review. First, the model focuses on economies of scope and ignores economies of scale. A natural extension would be to allow economies of scale in cost components. Numerical investigations incorporating economies of scale demonstrated qualitatively similar results, and suggest that the main findings are likely to hold (see Appendix B). Second, for analytical tractability the model assumes a uniform distribution for the demand of Service 2. As shown again in Appendix B, the results remain qualitatively similar under other non-uniform demand distributions. The model also assumes that reprovisioning is equally available in shared and dedicated networks. The use of different reprovisioning parameters for shared and dedicated solutions,

i.e., α_s and α_d, would be a relevant extension. Also, the model considers that reprovisioning is invoked only in the presence of excess demand, i.e., provisioned capacity could not be relinquished when demand was insufficient. Allowing (a)symmetric reprovisioning in both directions represents another interesting extension. Finally, modeling bundled pricing is a natural extension of the investigation now that we have developed a clearer understanding of the impact of core supply-side parameters such as economies and diseconomies of scope across dedicated and shared infrastructures and the ability to dynamically reprovision capacity. These issues, and in particular pricing strategies to best leverage demand externalities arising from service integration in a shared network, are topics we plan to address in future work.

References

Abdel-Malek, L., R. Montanari, L. C. Morales. 2004. Exact, approximate, and generic iterative models for the multi-product newsboy problem with budget constraint. *International Journal of Production Economics*. 91(2) 189–198.

Alp, O., T. Tan. 2008. Tactical capacity management under capacity flexibility. *IIE Transactions* 40(3) 221-237.

Brandel, M. 2007. Facilities management goes green. www.computerworld.com/s/article/295379/ Green_Buildings_ Smarter_More_Energy_Efficient.

Cisco Report. 2003. How does network convergence support a business strategy? www.cisco.com/global /EMEA /sidewide assets/pdfs/tdm/iptel/roi benefits paper.pdf.

Cisco's Featured Routers Products. 2010. www.cisco.com/en/US/products/hw/routers/index.html# products -menu.

Crosby, T. 2008. How FiOS works. http://communication.howstuffworks.com/fiber-optic-communications/fios.htm/printable.

Du, A. Y., X. Geng, R. Gopal, R. Ramesh, A. B. Whinston. 2008. Capacity provision networks: Foundations of markets for sharable resources in distributed computational economies. *Information System Research*.19 144–160.

Erlebacher, S. J. 2000. Optimal and heuristic solutions for the multi-item newsvendor problem with a single capacity constraint. *Production and Operations Management*. 9(3) 303–318.

Fine, C. H., R. M. Freund. 1990. Optimal investment in product-flexible manufacturing capacity. *Management Science*. 36(4) 449–465.

Fishburn, P. C., A. M. Odlyzko. 1998. Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet. Proc. First International Conference on Information and Computation Economics (ICE 98), Charleston, SC, October 1998.

Goyal, M. and S. Netessine. 2011. Volume flexibility, product flexibility or both: the role of demand correlation and product substitution. *Manufacturing & Service Operations Management*, 13(2) 180-193.

Gupta, A., D. O. Stahl, A. B. Whinston. 1997a. *Priority pricing of integrated services networks*. MIT Press, Cambridge, MA, USA, 323–352.

Gupta, A., D. O. Stahl, and A. B. Whinston. 1997b. A Stochastic Equilibrium Model of Internet Pricing. *Journal of Economic Dynamics and Control*, 21 697-722.

Gupta, A., B. Jukic, D. O. Stahl, and A. B. Whinston. 2011. An Analysis of Incentives for Network Infrastructure Investment Under Different Pricing Strategies. *Information Systems Research*, 22 (2) 215-232.

Hankins, M. L. 1999. Protocol Spawns Debate Over Future of Voice, Data, Video Communications. Signal Online, May 1999.
www.afcea.org/signal/articles/templates/SIGNAL_Article_Template.asp?articleid=878.

Hosanagar, K., J. Chuang, R. Krishnan, M. D. Smith. 2008. Service adoption and pricing of content delivery network (CDN) services. *Management Science*. 54 1579–1593.

Hosanagar, K., R. Krishnan, J. Chuang, V. Choudhary. 2005. Pricing and resource allocation in caching services with multiple levels of quality-of-service. *Management Science*. 51 1844–1859.

Hosanagar, K., Y. Tan. 2011. Optimal duplication in cooperative web caching. *Information Systems Research* Forthcoming.

Keon, N., G. Anandalingam. 2005. A new pricing model for competitive telecommunication services using congestion discounts. *INFORMS Journal on Computing*. 17 248–262.

Khouja, M. 1999. The single-period (news-vendor) problem: literature review and suggestions for future research. *Omega*. 27(5) 537–553.

Koebbe, P. H. 2007. A case for separate networks. Communications News. http://www.comnews.com/ features/2007_ november/1107a case.aspx.

Kogan, K., V. Portougal. 2006. Multi-period aggregate production planning in a news-vendor framework. *Journal of the Operational Research Society*. 57 423–433.

Laoutaris, N., G. Smaragdakis, P. Rodriguez, R. Sundaram. 2009. Delay tolerant bulk data transfers on the Internet. SIGMETRICS '09: Proc. *11th international joint conference on Measurement and modeling of computer systems*. ACM, 229–238.

Lau, H. S. 1980. The newsboy problem under alternative optimization objectives. *The Journal of the Operational Research Society*. 31(6) 525–535.

Lau, H. S., A. H. L. Lau. 1996. The newsstand problem: A capacitated multiple-product single-period inventory problem. *European Journal of Operational Research*. 94(1) 29 – 42.

Mahoney, J., K. Steenstrup. 2009. How the intersection of IT and OT is changing the CIO's role, and what to do about it. Tech. rep., Gartner Research. ID Number=G00172115.

Mendelson, H., S. Whang, 1990. Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue. Operations Research, 38 (5) 870-88.

Mileff, P., K. Nehez. 2007. Solving capacity constraint problems in a multi-item, multi-period newsvendor model. Proc. of microCAD 2007. 169–176.

Peterson, L., S. Shenker, J. Turner. 2005. Overcoming the Internet impasse through virtualization. *Computer*.38 34–41.

Petruzzi, N. C., M. Dada. 1999. Pricing and the newsvendor problem: A review with extensions. *Operations Research*. 47(2) 183–194.

Roberts, J. P., K. Steenstrup. 2010. The management implications of IT/OT convergence. Tech. rep., Gartner Research. ID Number=G00174016.

Tan, Y., V. S. Mookerjee, Y. Ji. 2006. Analyzing document- duplication effects on policies for browser and proxy caching. *INFORMS Journal on Computing* 18 506–522.

Tawarmalani, M., K. Kannan, P. De, C. Kumar. 2009. Allocating objects in a network of caches: Social welfare and incentive compatibility. *Management Science* 55 132– 147.

Tilson, D., K. Lyytinen, C. Sørensen. 2010. Digital infrastructures: The missing IS research agenda. *Information Systems Research* 21 748–759.

Tomlin, B. 2006. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science* 52(5) 639-657.

Touch, J., Y. Wang, L. Eggert, G. Finn. 2003. Virtual Internet architecture. *Proc. of ACM FDNA*. Karlsruhe, Germany.

Triden, B., S. Mantri, K. Schroeder, A. Thomas, J. Moyne, D. Tilbury. 2006 Dedicated vs. Shared Networks for Safety and Controls: An analysis of the trade-offs involved. Proc. IEEE Conference on Emerging Technologies and Factory Automation (ETFA'06), September 2006.

Van Mieghem, J. A. 1998. Investment strategies for flexible resources. Management Science. 44(8) 1071–1078.

Yager, T. 2009. Putting AT&T U-verse to the test. *PC World Business Center*. http://www.pcworld.com/businesscenter/article/160761/putting_atandt_uverse_to_the_test.html.

Zhang, B., S. Du. 2010. Multi-product newsboy problem with limited capacity and outsourcing. *European Journal of Operational Research*. 202(1) 107 – 113.

Zhang, Z., D. Dey, Y. Tong. 2007. Pricing communication services with delay guarantee. *INFORMS Journal on Computing*. 19 248–260.

6. Figures

Figure 1. The three-stage sequential decision process

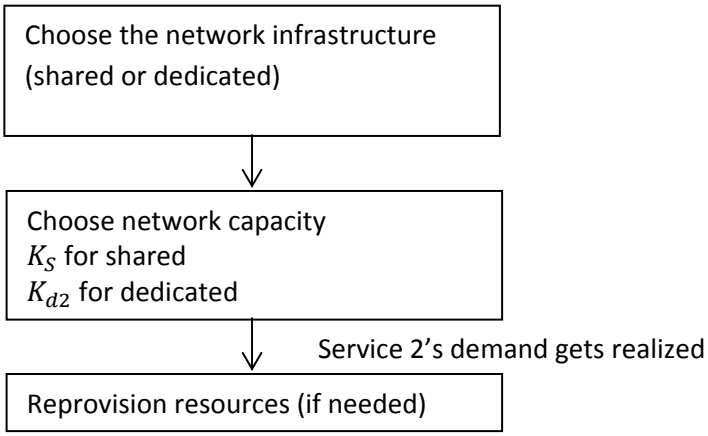
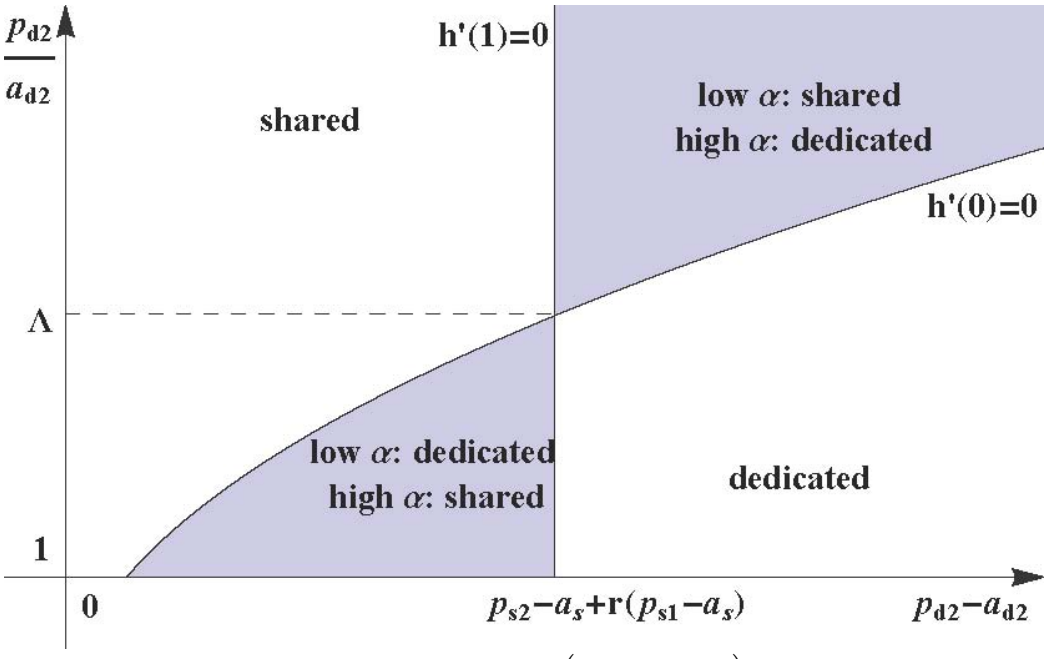
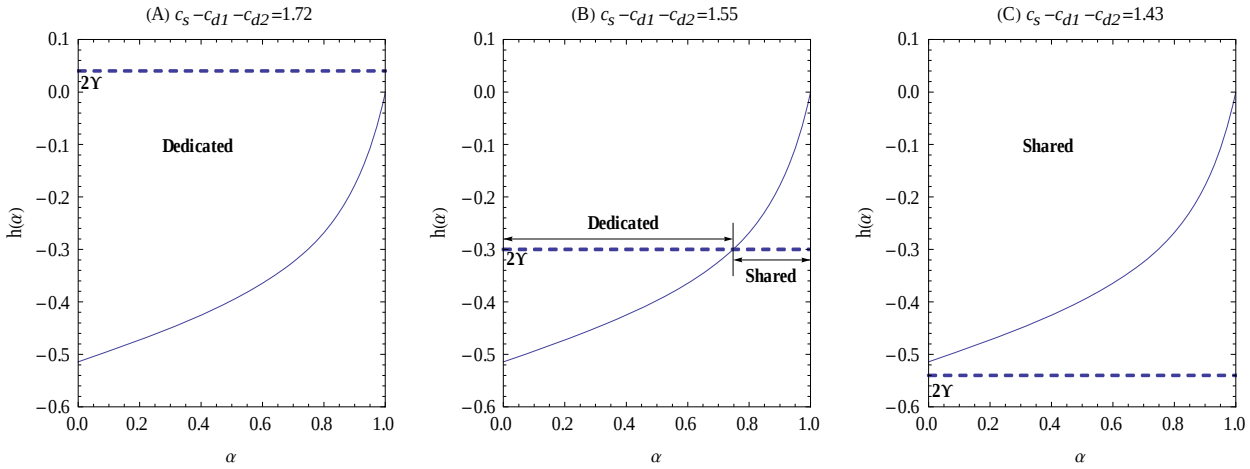


Figure 2. Partition of parameter space into regions corresponding to cases of Proposition 1 for the ($p_{d2} - a_{d2}$, $\frac{p_{d2}}{a_{d2}}$) plane



$$(\Lambda = \frac{\sqrt{p_{S2}-a_S+r(p_{S1}-a_S)}\left(\frac{p_{S2}}{a_S}+r\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)\right)}{\sqrt{p_{S2}-a_S+r(p_{S1}-a_S)\left(1+r\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_{S2}}\right)\left(1+\frac{p_{S2}}{a_S}+r\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)\right)\right)}})$$

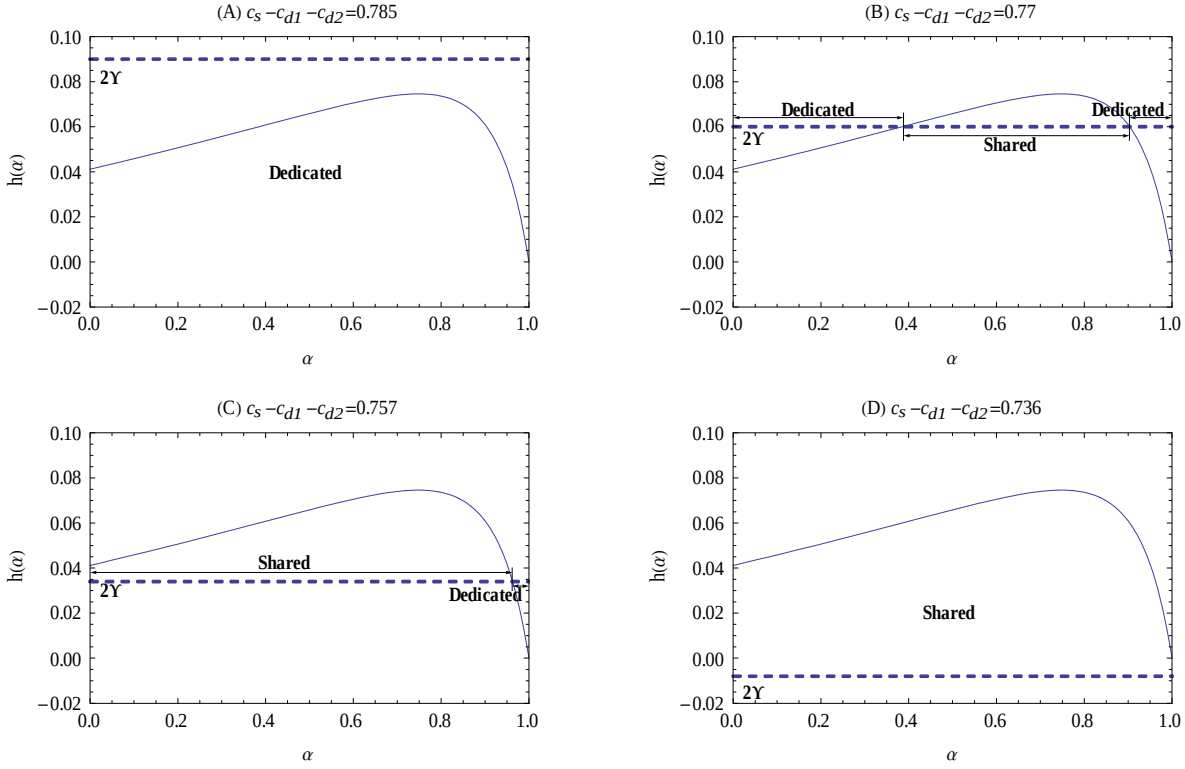
Figure 3: Impact of α on infrastructure choice when a shared network always benefits more from increases in α.



$$(p_{d2} - a_{d2} < (p_{s2} - a_s) + r(p_{s1} - a_s) \text{ and } \frac{p_{d2}}{a_{d2}} > \frac{\sqrt{p_{d2} - a_{d2}}\left(\frac{p_{s2}}{a_s} + r\left(\frac{p_{s2}}{a_s} - \frac{p_{s1}}{a_s}\right)\right)}{\sqrt{p_{s2} - a_s + (p_{s1} - a_s)r\left(1 + r\left(\frac{p_{s2}}{a_s} - \frac{ps1}{as2}\right)\left(1 + \frac{p_{s2}}{a_s} + r\left(\frac{p_{s2}}{a_s} - \frac{p_{s1}}{a_s}\right)\right)\right)}})$$

(Parameters: $p_{s2} = 4.5$, $p_{s1} = 2.4$, $a_s = 1$, $a_{s1} = 1$, $p_{d2} = 2$, $a_{d2} = 0.5$, $r = 0.5$, $f = 0.1$, $p_{d1} = 2.4$, $a_{d1} = 1$, $X_1 = 1$, $X_2^{max} = 1$)

Figure 4: Impact of α on infrastructure choice when a shared network benefits more from increases in α at low α and a dedicated network at high α.



$$\left(p_{d2} - a_{d2} > (p_{s2} - a_s) + r(p_{s1} - a_s) \text{ and } \frac{p_{d2}}{a_{d2}} > \frac{\sqrt{p_{d2} - a_{d2}}\left(\frac{p_{s2}}{a_s} + r\left(\frac{p_{s2}}{a_s} - \frac{p_{s1}}{a_s}\right)\right)}{\sqrt{p_{s2} - a_s + (p_{s1} - a_s)r\left(1 + r\left(\frac{p_{s2}}{a_s} - \frac{ps1}{as2}\right)\left(1 + \frac{p_{s2}}{a_s} + r\left(\frac{p_{s2}}{a_s} - \frac{p_{s1}}{a_s}\right)\right)\right)}}\right)$$

(Parameters: $p_{s2} = 4.5$, $p_{s1} = 2.4$, $a_s = 1$, $a_{s1} = 1$, $p_{d2} = 6$, $a_{d2} = 1$, $r = 0.2$, $f = 0.1$, $p_{d1} = 1.4$, $a_{d1} = 1$, $X_1 = 1$, $X_2^{max} = 1$)

7. Appendix

A. Proofs of Lemmas and Proposition 1

Proof of Lemma 1: It can be derived that $\frac{\partial K_{d2}^*}{\partial \frac{p_{d2}}{a_{d2}}} = \frac{(1-\alpha)X_2^{max}}{\left((1-\alpha)\frac{p_{d2}}{a_{d2}}+\alpha\right)^2}$, $\frac{\partial K_S^*}{\partial \frac{p_{S1}}{a_S}} = \frac{(1-\alpha)^2\left(\frac{p_{S2}}{a_S}-1\right)r(1+r)X_2^{max}}{\left(\alpha+(1-\alpha)\frac{p_{S2}}{a_S}+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)^2}$,

$\frac{\partial K_S^*}{\partial \frac{p_{S2}}{a_S}} = \frac{(1-\alpha)(1+r)\left(1-(1-\alpha)\left(\frac{p_{S1}}{a_S}-1\right)r\right)X_2^{max}}{\left(\alpha+(1-\alpha)\frac{p_{S2}}{a_S}+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)^2}$, $\frac{\partial K_{d2}^*}{\partial \alpha} = \frac{-\left(\frac{p_{d2}}{a_{d2}}-1\right)X_2^{max}}{\left((1-\alpha)\frac{p_{d2}}{a_{d2}}+\alpha\right)^2}$, $\frac{\partial K_S^*}{\partial \alpha} = \frac{-(1+r)\left(\frac{p_{d2}}{a_{d2}}-1\right)X_2^{max}}{\left(\alpha+(1-\alpha)\frac{p_{S2}}{a_S}+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)^2}$,

$\frac{\partial K_S^*}{\partial r} = \frac{(1-\alpha)\left(\frac{p_{S2}}{a_S}-1\right)\left(\alpha+(1-\alpha)\frac{p_{S1}}{a_S}\right)X_2^{max}}{\left(\alpha+(1-\alpha)\frac{p_{S2}}{a_S}+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)^2}$ and $\frac{\partial K_S^*}{\partial f} = X_1$. Given $r\left(\frac{p_{S1}}{a_S}-1\right) < 1$, it can be easily checked that $\frac{\partial K_{d2}^*}{\partial \alpha} < 0$, $\frac{\partial K_S^*}{\partial \alpha} < 0$, $\frac{\partial K_S^*}{\partial f} > 0$, and in addition, that $\frac{\partial K_{d2}^*}{\partial \frac{p_{d2}}{a_{d2}}}$, $\frac{\partial K_S^*}{\partial \frac{p_{S1}}{a_S}}$, $\frac{\partial K_S^*}{\partial \frac{p_{S2}}{a_S}}$ and $\frac{\partial K_S^*}{\partial r}$ are all greater than or equal to zero, and that they are zero only if $\alpha = 1$..

Proof of Lemma 2: It can be derived that $\frac{\partial h_d(\alpha)}{\partial \alpha} = -\frac{(p_{d2}-a_{d2})}{\left((1-\alpha)\frac{p_{d2}}{a_{d2}}+\alpha\right)^2}$ and $\frac{\partial h_s(\alpha)}{\partial \alpha} = -\frac{(p_{S2}-a_S)+(p_{S1}-a_S)r+(1-\alpha)(p_{S1}-a_S)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)\left(1+\alpha+(1-\alpha)\frac{p_{S2}}{a_S}\right)r^2+(1-\alpha)^2(p_{S1}-a_S)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)^2r^3}{\left(\alpha+(1-\alpha)\frac{p_{S2}}{a_S}+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)^2}$. It can be easily checked that both are lower than zero.

Proof related to Proposition 1: $h'(\alpha)$ satisfies the single-crossing property (i.e., changes its sign at most once) for $\alpha \in [0,1]$ if the equation $h'(\alpha) = 0$ has at most one solution for $\alpha \in [0,1]$. To examine the behavior of $h'(\alpha) = 0$, we derive:

$$h'(\alpha) = \frac{a_S\left(\frac{p_{S2}}{a_S}-1\right)+a_S\left(\frac{p_{S1}}{a_S}-1\right)r\left(1+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\left(\left(1+\alpha+(1-\alpha)\frac{p_{S2}}{a_S}\right)+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)\right)}{\left(\alpha+(1-\alpha)\frac{p_{S2}}{a_S}+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)^2} - \frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)}{\left((1-\alpha)\frac{p_{d2}}{a_{d2}}+\alpha\right)^2}$$

$$h''(\alpha) = \frac{2a_S\left(\frac{p_{S2}}{a_S}-1\right)^2(1+r)}{\left(\alpha+(1-\alpha)\frac{p_{S2}}{a_S}+(1-\alpha)\left(\frac{p_{S2}}{a_S}-\frac{p_{S1}}{a_S}\right)r\right)^3} - \frac{2a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2}{\left(\alpha+(1-\alpha)\frac{p_{d2}}{a_{d2}}\right)^3}.$$

$h''(\alpha) = 0$ can be rewritten as:

$$\alpha + (1 - \alpha)\frac{p_{s2}}{a_s} + (1 - \alpha)\left(\frac{p_{s2}}{a_s} - \frac{p_{s1}}{a_s}\right)r = \left(\frac{a_s\left(\frac{p_{s2}}{a_s}-1\right)^2(1+r)}{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2}\right)^{\frac{1}{3}}\left(\alpha + (1 - \alpha)\frac{p_{d2}}{a_{d2}}\right).$$

It can be easily seen that $h''(\alpha) = 0$ has at most one solution for $\alpha \in [0, 1]$. If it doesn't have a solution for $\alpha \in [0, 1]$, $h'(0) = 0$ has at most one solution for $\alpha \in [0, 1]$. If the solution exists, it is

$$\alpha = \frac{\frac{p_{d2}}{a_{d2}}\Phi - \Omega\Psi}{\left(\frac{p_{d2}}{a_{d2}}-1\right)\Phi - (\Omega-1)\Psi},$$

where $\Phi = \left(a_s\left(\frac{p_{s2}}{a_s} - 1\right)^2(1 + r)\right)^{\frac{1}{3}}$, $\Psi = \left(a_{d2}\left(\frac{p_{d2}}{a_{d2}} - 1\right)^2\right)^{\frac{1}{3}}$, and $\Omega = \frac{p_{s2}}{a_s} + \left(\frac{p_{s2}}{a_s} - \frac{p_{s1}}{a_s}\right)r$. Let $\bar{\alpha} = \frac{\frac{p_{d2}}{a_{d2}}\Phi - \Omega\Psi}{\left(\frac{p_{d2}}{a_{d2}}-1\right)\Phi - (\Omega-1)\Psi}$. $\bar{\alpha} \in [0, 1]$ if either of the following conditions holds:

(1) $\frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2\Omega^3}{(1+r)\left(\frac{p_{d2}}{a_{d2}}\right)^3\left(\frac{p_{s2}}{a_s}-1\right)^2} < a_s < \frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2}{(1+r)\left(\frac{p_{s2}}{a_s}-1\right)^2}$ and $\frac{p_{d2}}{a_{d2}} > \Omega$,

(2) $\frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2}{(1+r)\left(\frac{p_{s2}}{a_s}-1\right)^2} < a_s < \frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2\Omega^3}{(1+r)\left(\frac{p_{d2}}{a_{d2}}\right)^3\left(\frac{p_{s2}}{a_s}-1\right)^2}$ and $\frac{p_{d2}}{a_{d2}} < \Omega$.

Under Condition (1), we can confirm that $h''(\alpha) > 0$ for $0 < \alpha < \bar{\alpha}$, $h''(\alpha) < 0$ for $\bar{\alpha} < \alpha < 1$, and $h'(0) > 0$. Therefore, $h'(0) = 0$ has at most one solution for $\alpha \in [0, 1]$.

Under Condition (2), we can confirm that $h''(\alpha) < 0$ for $0 < \alpha < \bar{\alpha}$, $h''(\alpha) > 0$ for $\bar{\alpha} < \alpha < 1$. In this case, $h'(0) = 0$ has at most one solution for $\alpha \in [0, 1]$ except when $h'(0) > 0$, $h'(1) > 0$ and $h'(\bar{\alpha}) < 0$, for which $h'(0) = 0$ has two solutions. This region appears if

$$\text{Max}\left[\frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)\Omega^2}{\left(\frac{p_{d2}}{a_{d2}}\right)^2\left(\frac{p_{s2}}{a_s}-1+\left(\frac{p_{s1}}{a_s}-1\right)r\left(1+\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)r(1+\Omega)\right)\right)}, \frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)}{\frac{p_{s2}}{a_s}-1+\left(\frac{p_{s1}}{a_s}-1\right)r}\right] < a_s <$$

$$\frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)(\Omega-1)^2 - 3\Psi\Phi\left((\Omega-1)\Psi - \left(\frac{p_{d2}}{a_{d2}}-1\right)\Phi\right)}{\left(\left(\frac{p_{d2}}{a_{d2}}-1\right)^2\left(\frac{p_{s2}}{a_s}-1+\left(\frac{p_{s1}}{a_s}-1\right)r\right)+\left(\frac{p_{s1}}{a_s}-1\right)\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)r^2\left(1-2\frac{p_{d2}}{a_{d2}}+\Omega\right)\right)}$$

In summary, $h'(\alpha)$ satisfies the single-crossing property except when the following two conditions are both satisfied:

(C1) $\frac{p_{d2}}{a_{d2}} < \Omega$

(C2) $\text{Max}\left[\frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)\Omega^2}{\left(\frac{p_{d2}}{a_{d2}}\right)^2\left(\frac{p_{s2}}{a_s}-1+\left(\frac{p_{s1}}{a_s}-1\right)r\left(1+\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)r(1+\Omega)\right)\right)}, \frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)}{\frac{p_{s2}}{a_s}-1+\left(\frac{p_{s1}}{a_s}-1\right)r}, \frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2}{(1+r)\left(\frac{p_{s2}}{a_s}-1\right)^2}\right] < a_s <$

$$\text{Min}\left[\frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)^2\Omega^3}{(1+r)\left(\frac{p_{d2}}{a_{d2}}\right)^3\left(\frac{p_{s2}}{a_s}-1\right)^2}, \frac{a_{d2}\left(\frac{p_{d2}}{a_{d2}}-1\right)(\Omega-1)^2-3\Psi\Phi\left((\Omega-1)\Psi-\left(\frac{p_{d2}}{a_{d2}}-1\right)\Phi\right)}{\left(\left(\frac{p_{d2}}{a_{d2}}-1\right)^2\left(\frac{p_{s2}}{a_s}-1+\left(\frac{p_{s1}}{a_s}-1\right)r\right)+\left(\frac{p_{s1}}{a_s}-1\right)\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)r^2\left(1-2\frac{p_{d2}}{a_{d2}}+\Omega\right)\right)}\right]$$

When (C1) and (C2) are satisfied, $h'(0) = 0$ has two solutions for $\alpha \in [0,1]$, $h'(0) > 0$, and $h'(1) > 0$. In this case, a shared network benefits more from increases in α at low and high values of α and a dedicated network at intermediate values of α. A numerical example for this scenario is shown in Figure 5, which shows that there are five possible network choice outcomes depending on 2γ. A shared (dedicated) network is chosen irrespective of α if there are significant economies (diseconomies) of scope as shown in Figure 5(A) (Figure 5(E)). When $h(\alpha)$ intersects 2γ, there can be up to three switches in the optimal network choice as shown in Figure 5(B), 5(C) and 5(D). Thus our main result that changes in α affect infrastructure choice and that the optimal choice can switch multiple times as α varies in the range $[0,1]$ still holds.

Proof of Corollary 2: According to Proposition 3, a shared network benefits more from better reprovisioning at low α if

$$\frac{p_{d2}}{a_{d2}} > \frac{\sqrt{p_{d2}-a_{d2}}\left(\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)}{\sqrt{p_{s2}-a_s+(p_{s1}-a_s)r\left(1+r\left(\frac{p_{s2}}{a_s}-\frac{ps1}{as2}\right)\left(1+\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)\right)}}.$$

This condition is equivalent to

$$\frac{(p_{d2}-a_{d2})}{\left(\frac{p_{d2}}{a_{d2}}\right)^2} < \frac{p_{s2}-a_s+r(p_{s1}-a_s)\left(1+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\left(1+\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)\right)}{\left(\frac{p_{s2}}{a_s}+r\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}\right)\right)^2}.$$

The left hand side of the above inequality is independent of α, and the first derivative of the right hand side on α equals $a_s\left(\frac{p_{s1}}{a_s}-1-\frac{2\frac{p_{s1}}{a_s}\left(\frac{p_{s2}}{a_s}-1\right)^2}{\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}r+\frac{p_{s2}}{a_s}r\right)^3}+\frac{\left(\frac{p_{s1}}{a_s}-1\right)\left(\frac{p_{s2}}{a_s}-1\right)^2}{\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}r+\frac{p_{s2}}{a_s}r\right)^2}\right)$, which is greater than zero if p_{s1} is less than the solution to the following equation that falls in the range of $(a_s,\ p_{s2})$[13]:

$$\frac{p_{s1}}{a_s}\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}r+\frac{p_{s2}}{a_s}r\right)^3-\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}r+\frac{p_{s2}}{a_s}r\right)^3-\left(\frac{p_{s1}}{a_s}-1\right)\left(\frac{p_{s2}}{a_s}-1\right)^2\left(\frac{p_{s2}}{a_s}-\frac{p_{s1}}{a_s}r+\frac{p_{s2}}{a_s}r\right)-$$

$$2\frac{p_{s1}}{a_s}\left(\frac{p_{s2}}{a_s}-1\right)^2=0.$$

B. Robustness To Model Changes

In this section, we demonstrate that the results are robust to several changes in the model. In particular, we show that the behaviors and outcomes that the model helps elucidate are still present when economies of scale are included or when using a non-uniform demand distribution. The investigation is carried out by numerically computing optimal provisioning decisions for shared and dedicated networks under these new conditions. It reveals that changes in the reprovisioning factor α still affect which network choice yields a higher profit. Furthermore, scenarios where multiple such changes arise as α varies in the range [0, 1] remain present as well.

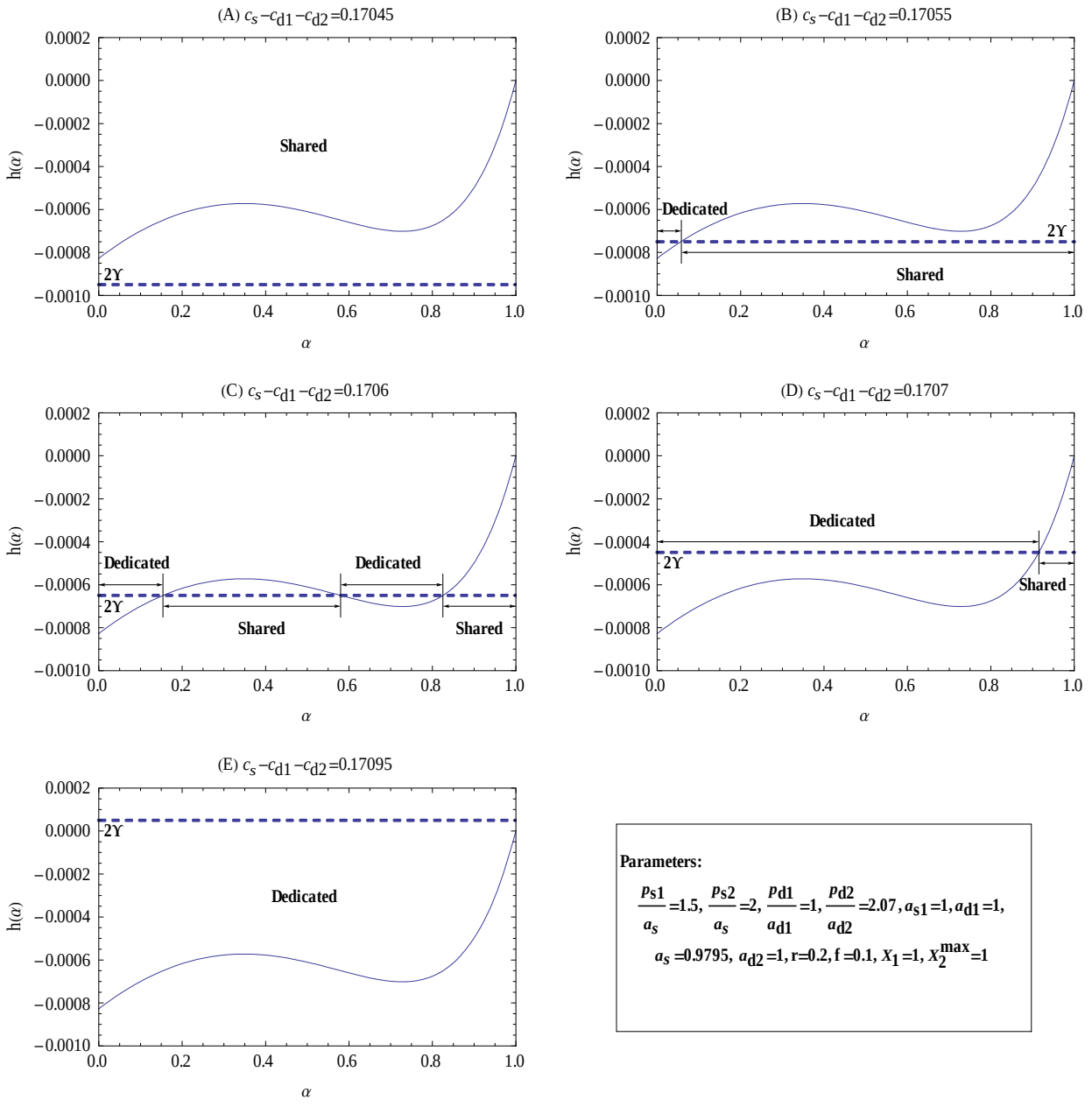
The inclusion of economies of scale is a natural extension, as they represent a common benefit associated with shared solutions. It is, therefore, of interest to verify that the presence of such a benefit (for shared solutions) does not eliminate the impact that the coefficient α can have on determining the solution of choice. Similarly, validating that changes in demand distribution do not significantly affect the outcome is another standard test of the robustness of the results.

In Figure 6(A), we use $a_{d2}K_{d2}^{0.8}$ and $a_{d1}X_1^{0.8}$ to capture economies of scale in capacity costs for Services 1 and 2 respectively in the dedicated network, $a_sK_s^{0.8}$ for the flexible capacity in the shared network, and $a_{s1}X_1^{0.8}$ for the existing capacity for Service 1 in the shared network. The example shows an

[13] There is a closed-form solution but it is too complex to show.

instance of infrastructure choice where dedicated networks are preferred at both high and low α, while a shared network is preferred at intermediate values of α. In Figure 6(B), Service 2's demand distribution follows a beta distribution with parameters (1.5, 1), which is negatively skewed. In this scenario, a shared network is preferred at both high and low α, while dedicated networks are preferred at intermediate values of α. Figure 6(C) displays a similar example with the demand distribution of Service 2 now following an Erlang distribution with parameters (2, 5), which is positively skewed. In this scenario, dedicated networks are preferred at both high and low values of α, while a shared network is preferred at intermediate values of α.

Figure 5. Impact of α on infrastructure choice when $h(\alpha)$ does not satisfy the single-crossing property.

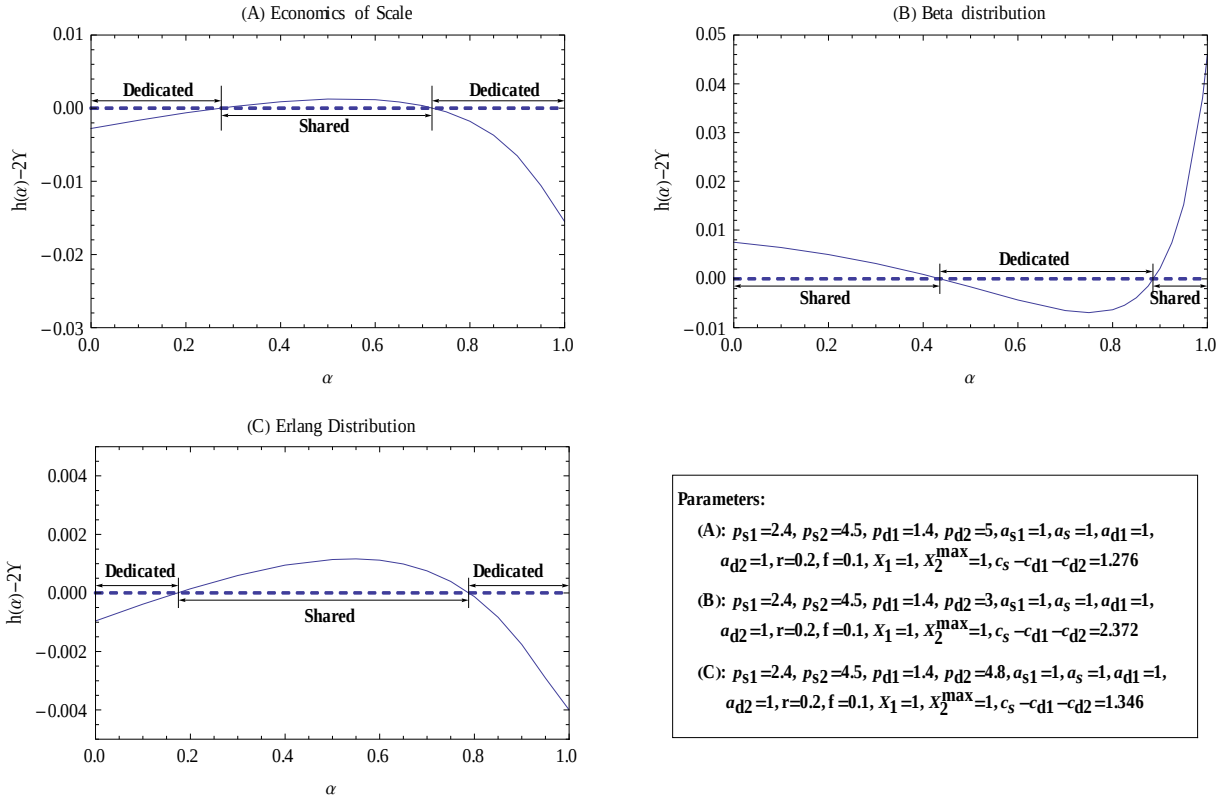


Parameters:

$\frac{p_{s1}}{a_s} = 1.5,\ \frac{p_{s2}}{a_s} = 2,\ \frac{p_{d1}}{a_{d1}} = 1,\ \frac{p_{d2}}{a_{d2}} = 2.07,\ a_{s1} = 1,\ a_{d1} = 1,$

$a_s = 0.9795,\ a_{d2} = 1,\ r = 0.2,\ f = 0.1,\ X_1 = 1,\ X_2^{max} = 1$

Figure 6. Impact of α on infrastructure choice when economics of scale or different form of demand distribution is assumed



Parameters:

(A): $p_{s1}=2.4$, $p_{s2}=4.5$, $p_{d1}=1.4$, $p_{d2}=5$, $a_{s1}=1$, $a_s=1$, $a_{d1}=1$, $a_{d2}=1$, $r=0.2$, $f=0.1$, $X_1=1$, $X_2^{max}=1$, $c_s-c_{d1}-c_{d2}=1.276$

(B): $p_{s1}=2.4$, $p_{s2}=4.5$, $p_{d1}=1.4$, $p_{d2}=3$, $a_{s1}=1$, $a_s=1$, $a_{d1}=1$, $a_{d2}=1$, $r=0.2$, $f=0.1$, $X_1=1$, $X_2^{max}=1$, $c_s-c_{d1}-c_{d2}=2.372$

(C): $p_{s1}=2.4$, $p_{s2}=4.5$, $p_{d1}=1.4$, $p_{d2}=4.8$, $a_{s1}=1$, $a_s=1$, $a_{d1}=1$, $a_{d2}=1$, $r=0.2$, $f=0.1$, $X_1=1$, $X_2^{max}=1$, $c_s-c_{d1}-c_{d2}=1.346$