9-20-2010

The Impact of Reprovisioning on the Choice of Shared versus Dedicated Networks

Soumya Sen
University of Pennsylvania, soumyas@princeton.edu

Kristin Yamauchi
University of Pennsylvania, kristin.yamauchi@gmail.com

Roch Guerin
University of Pennsylvania, guerin@acm.org

Kartik Hosanagar
University of Pennsylvania, kartikh@wharton.upenn.edu

Suggested Citation:
Sen, S., K. Yamauchi., R. Guerin and K. Hosanagar. "The Impact of Reprovisioning on the Choice of Shared versus Dedicated Networks." *Accepted for WEB 2010, Ninth Workshop on E-Business.* December 11, 2010. Saint Louis, Missouri, USA.

The Impact of Reprovisioning on the Choice of Shared versus Dedicated Networks

Abstract

As new network services emerge, questions about service deployment and network choices arise. Although shared networks, such as the Internet, offer many advantages, combining heterogeneous services on the same network need not be the right answer as it comes at the cost of increased complexity. Moreover, deploying new services on dedicated networks is becoming increasingly viable, thanks to virtualization technologies. In this work, we introduce an analytical framework that gives Internet Service Providers the ability to explore the trade-offs between shared and dedicated network infrastructures. The framework accounts for factors such as the presence of demand uncertainty for new services, (dis)economies of scope in deployment and operational costs, and the extent to which new technologies allow dynamic (re)provisioning of resources in response to excess demands. The main contribution is the identification and quantification of dynamic (re)provisioning as a key factor in determining the preferred network infrastructure, i.e. shared or dedicated.

Keywords

Network services, resource allocation, virtualization

Disciplines

Digital Communications and Networking | Other Operations Research, Systems Engineering and Industrial Engineering | Systems and Communications

Comments

Suggested Citation:

Sen, S., K. Yamauchi., R. Guerin and K. Hosanagar. "The Impact of Reprovisioning on the Choice of Shared versus Dedicated Networks." *Accepted for WEB 2010, Ninth Workshop on E-Business.* December 11, 2010. Saint Louis, Missouri, USA.

The Impact of Reprovisioning on the Choice of Shared versus Dedicated Networks[1]

Soumya Sen
ESE, University of Pennsylvania
ssoumya@seas.upenn.edu

Kristin Yamauchi
ESE, University of Pennsylvania
kristin.yamauchi@gmail.com

Roch Guérin
ESE, University of Pennsylvania
guerin@ee.upenn.edu

Kartik Hosanagar
Wharton, University of Pennsylvania
kartikh@wharton.upenn.edu

Abstract

As new network services emerge, questions about service deployment and network choice arise. Although shared networks, e.g., the Internet, offer many advantages, combining heterogeneous services on the same network need not be the right answer as it comes at the cost of increased complexity. Moreover, deploying new services on dedicated networks is becoming increasingly viable, thanks to virtualization technologies. In this work, we introduce an analytical framework to explore the trade-offs between shared and dedicated network infrastructures. We account for factors like the presence of demand uncertainty for new services, (dis)economies of scope in various costs, and the extent to which new technologies allow dynamic resource (re)provisioning in response to excess demands. The main contribution is the identification and quantification of how dynamic (re)provisioning impacts the preferred network choice, i.e., shared or dedicated.

Keywords: Networks, Resource Allocation, Virtualization, Capacity Planning

1. Introduction

Advances in network technologies have resulted in the Internet evolving from a simple data network to a global communication infrastructure that carries a multiplicity of services. Although such integration has advantages, combining services with disparate requirements onto a shared network comes at a cost. It often calls for the entire network to be "upgraded" with features required by only a handful of services, but at a cost that is borne by all of them. It can also introduce complex interactions among the services. Therefore, it is of interest to determine if, when, and why sharing a network across multiple services is beneficial or not. The question is becoming more relevant with the advent of new technologies, *e.g.* virtualization (Anderson et al., 2005; Touch et al., 2003), which facilitate the deployment of network "slices" dedicated to individual services. Conversely, even in the absence of new technologies recent instances of service deployments point to complex decisions. For example, in deploying its new U-verse TV service, AT&T chose to create a dedicated network to ensure better manageability and reliability in delivering HQ video (AT&T, 2009). In contrast, Verizon created FiOS (Crosby, 2008), a shared network for voice, video, and data services. As more new areas become network-enabled (*e.g.,* health-care, surveillance), the question of choosing shared or dedicated networks for such services become important. For instance, the emergence of green buildings results in a facilities management infrastructure that relies on networked sensors and actuators to monitor and control building operations. This can be realized either by piggybacking the existing IT infrastructure of a building (Brandel, 2007), or by creating a dedicated facilities management network (Koebbe,

[1] This work was supported by NSF grant CNS-0721610

2007), and neither option is an obvious winner. Making these decisions call for a framework that systematically examines the trade-offs between shared and dedicated infrastructures. Although this issue has got some attention in the business press, there is limited formal analysis. Creating a framework to evaluate the tradeoffs is the primary motivation for this work.

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 captures any economies or diseconomies of scope in resource sharing and accounts for the ability to adjust (reprovision) network resources in response to a higher than anticipated demand of the new service. The main contribution of this study is in offering a framework for making network infrastructure decisions and in establishing that the extent to which reprovisioning is feasible can by itself affect which infrastructure, shared or dedicated, is more profitable. In particular, two operational metrics, the gross profit margin and the return on capacity, influence which network benefits more from reprovisioning. Note that we will use the terms ‘network’ and ‘infrastructure’ interchangeably. The paper is structured as follows: Section 2 discusses the related works, Section 3 presents the model, Section 4 provides the analysis, and Section 5 concludes the work.

2. Related Literature

Our model shares some basic structural properties with works in the manufacturing systems literature on optimal resource planning in the presence of demand uncertainty (Fine and Freund, 1990; Van Mieghem, 1998). In these works, investment decisions in manufacturing plants with flexible (but more expensive) and/or dedicated resources have to be made prior to the realization of product demand. Plants with flexibility to produce different types of products are more expensive to build, but can deal with uncertain demand. This creates a trade-off regarding how much capacity to build into flexible and dedicated plants – a feature that parallels the decision of choosing between shared and dedicated networks in our work. However, there are several differences with these earlier works. First, rather than explore the benefits of a flexible (shared) plant (network) in dealing with uncertain demand, our focus is on investigating the impact of economies and diseconomies of scope in the underlying costs. A second and more significant difference is that building a new manufacturing plant typically involves a significant time-lag so that capacity decisions, once made, cannot be readily revisited even in the presence of large realized excess demand. This implies that realized demand in excess of the plant’s capacity is usually lost. In contrast, networks are becoming more akin to services, and adjusting capacity in response to an unexpected increase in demand can be realized relatively fast. Till recently the provisioning of a T1 connection (1.544 Mbps) required several months, while “dialing-up” an additional Gbps of bandwidth is now commonly available. The advent of technologies such as virtualization is making it increasingly feasible to “upgrade” (reprovision) network capacity in relatively short time-scales. As a result, even if some excess demand is lost (*i.e.,* reprovisioning incurs a penalty) networks can recover from insufficient provisioning decisions. As we shall see, this reprovisioning ability affects not only the optimal provisioning, but also the network choice.

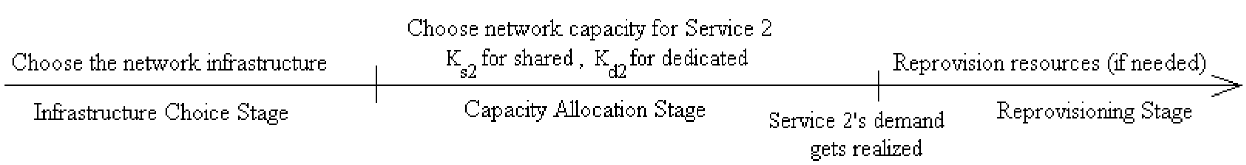

Fig. 1. The three-stage sequential decision process

3. Model Formulation

We consider the most basic setting to explore the trade-off between shared and dedicated networks. Specifically, one service has already been deployed and has a stable demand, and its 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 network of the existing service. For analytical simplicity, we ignore possible economies of scale, whose magnitude is typically limited in networks, *e.g.,* (Laoutaris et al., 2009) shows that bandwidth costs exhibit nearly linear growth at both access and backbone speeds. Our numerical investigations verify that economies of scale do not qualitatively affect our results (see Section 2 of (Sen et al., 2010)).

The network provider's objective is to maximize its total profit from the two services. We model the provider's decision problem as a 3-stage sequential process, as shown in Fig. 1. In the first stage, the provider makes the infrastructure choice, *i.e.* a shared or a dedicated network. At this stage, the provider does not know the profit from Service 2 as its demand is uncertain. Given a network choice, in the second stage, the provider provisions capacity for the yet unknown demand of Service 2. Service 2's demand is realized in the third stage, where the provider has the opportunity to reprovision if this demand exceeds the capacity provisioned upfront. But a penalty for under-provisioning is incurred as 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. We assume that contractual obligations preclude downward adjustment of resources. These 3 stages are referred to as *Infrastructure Choice Stage*, *Capacity Allocation Stage* and *Reprovisioning Stage*, respectively.

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 upfront and network 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 in Stage 2. Based on these results for Capacity Allocation and Reprovisioning Stages, we finally evaluate the provider's total expected profit for each network choice, and select the one that yields the higher expected profit. Section 3.2 discusses these steps.

3.1 Model Parameters

Service 1 is an existing service with a mature demand, and is assumed to operate at full capacity, *i.e.,* its provisioned capacity equals realized demand, X_1. Service 2, being new, has an uncertain demand, denoted by a random variable x_2 with known probability density function, f_{x_2}. We use X_2 to indicate a realization of this demand. The provisioned capacity (number of users the network can handle) for Service 2 is a decision variable denoted by K_{s2} and K_{d2} for shared and dedicated networks, respectively. If Service 2's demand exceeds the provisioned level ($X_2 > K_i$, $i = \{s2, d2\}$), network resources can be adjusted to accommodate a fraction α of the excess demand, *i.e.,* resources are increased to $K_i + \alpha(X_2 - K_i)$. The parameter α or *reprovisioning coefficient*, represents the fraction of excess demand that reprovisioning can capture. This fraction is assumed to be independent of the magnitude of the reprovisioning effort. In other words, the feasibility of and latency in securing additional capacity are the same regardless of the amount of capacity requested (at least within some bounds), with the latter possibly affecting the network's ability to retain all the excess demand that was present. This is consistent with the provisioning process of most computing and communications facilities. Therefore, when $\alpha=0$, reprovisioning is unable, *e.g.,* too slow, to capture any excess demand, while $\alpha=1$ means that reprovisioning succeeds in accommodating the entire excess demand, *i.e.,* a "provisioning phase" is unnecessary as resources can be secured on-the-fly. Different levels of provisioning flexibility (as afforded by different types of virtualization technology) can be accounted for by varying α.

Next, we introduce the revenue and cost components of the network choices. We follow standard practice (Fine and Freund, 1990; Van Mieghem, 1998) and consider only the present value of all future investments, revenues and costs. Services generate revenues from user paid subscription fees (determined by exogenous market factors). Offering a service incurs a per user connection cost, *e.g.,* cost of last-mile connectivity, installing end-user access equipments. We denote the per-user *contribution margins* for Services 1 and 2 respectively in a shared (dedicated) network by p_{s1} and p_{s2} (p_{d1} and p_{d2}). Implicit in their definition is the assumption that there are no bundling discount for subscribing to both services, and that the per user connection costs are additive across services, as common in settings like FiOS where the bulk of the connection costs lie in termination equipments specific to each service. Offering services also involve *fixed* and *capacity* costs. Fixed costs are independent of demand and capacity levels, *e.g.,* facility rent, R&D, and are denoted by c_s for shared, and by c_{d1} and c_{d2} for dedicated networks of Services 1 and 2. Capacity costs grow with network resources; they are incurred upfront because of provisioning and may also be incurred later during reprovisioning to accommodate some excess demand. Unit capacity costs for Service 1 are denoted by a_{s1} and a_{d1} in shared and dedicated networks respectively, and by a_{s2} and a_{d2} for Service 2. The term *return on capacity* refers to the ratio of contribution margin to capacity cost, *i.e.* (p_i/a_i), $i = \{s2, d2\}$ for Service 2. All these costs may exhibit different levels of (dis)economies of scope between shared and dedicated networks.

3.2 Model Setup and Solutions

Next we solve the 3-Stage model of Fig. 1 in the reverse order of the decision process to obtain the expected profits for shared and dedicated networks. Note that because Service 1 is mature and has a known demand, its revenue contribution is relevant only in the infrastructure choice stage. For Service 2, the solution method is essentially identical for both the shared and dedicated networks, and we use an index $i = \{s2, d2\}$ to denote the two respective cases.

3.2.1 Reprovisioning Stage

As mentioned earlier, reprovisioning takes place after the demand for Service 2 has been realized. If 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. Next, we present an expression for the gross profit from Service 2 after the reprovisioning phase for the two network choices. As defined in subsection 3.1, the contribution margin for Service 2 is p_i, and the variable cost of provisioning capacity is a_i (where $i=\{s2,d2\}$for shared and dedicated networks resp.). If the realized demand X_2 exceeds provisioned capacity, K_i, the capacity is adjusted to accommodate a fraction α of the excess demand, *i.e.,* capacity increases to $K_i + \alpha(X_2 - K_i)$. The gross profit for Service 2 is then $R_i(X_2 > K_i) = (p_i - a_i)(K_i + \alpha(X_2 - K_i))$. Conversely, when the realized demand is less than the provisioned capacity K_i, the gross profit for Service 2 is $R_i(X_2 \leq K_i) = p_i X_2 - a_i K_i$. Using these expressions, the optimal upfront capacity for Service 2 is found in the capacity allocation stage.

3.2.2 Capacity Allocation Stage

For a known demand density f_{X_2}, the expected gross profit R_i *given* the capacity provisioned upfront K_i (where $i = \{s2, d2\}$ for Service 2 in shared and dedicated network respectively) is:

$$\mathbf{E}(R_i)_{[K_i]} = \int_0^{K_i} R_i\,(X_2 \leq K_i) f_{x_2}\,\mathrm{d}(\mathrm{X}_2) + \int_{K_i}^{X_2^{max}} R_i(X_2 > K_i) f_{x_2}\,\mathrm{d}(\mathrm{X}_2) \qquad (1)$$

where $R_i\,(X_2 \leq K_i)$ and $R_i(X_2 > K_i)$ are given in Subsection 3.2.1. We assume f_{x_2} is uniformly distributed in [0, X_2^{max}]. Using $\mathbf{E}(R_i)_{[K_i]}$, the optimal capacity is

$$K_i^* = \frac{(1-\alpha)(p_i - a_i)X_2^{max}}{(1-\alpha)p_i + \alpha a_i} \qquad (2)$$

Substituting the expression for K_i^* from Eq. (2) in $\mathbf{E}(R_i)_{[K_i]}$, we get

$$\mathbf{E}(R_i)_{[K_i]} = \frac{(p_i - a_i)X_2^{max}}{2}\left(1 - \frac{(1-\alpha)a_i}{(1-\alpha)p_i + \alpha a_i}\right) \qquad (3)$$

3.2.3 Infrastructure Choice Stage

In this last stage, the overall profits of the two network choices are evaluated to select the one with the higher profit. We consider dedicated and shared networks in turn.

Dedicated Networks: The gross profit for Service 1 in a dedicated network is $R_{d1} = X_1(p_{d1} - a_{d1})$. Thus, the profit Π_{d1} from deploying Service 1 on a dedicated network is $\Pi_{d1} = X_1\,(p_{d1} - a_{d1}) - c_{d1}$. The expected profit under optimal provisioning for Service 2 is given by subtracting the fixed cost from $\mathbf{E}(R_{d2})_{[K_{d2}^*]}$ (ref. Eq. (3)).

$$\Pi_{d2} = \frac{(p_{d2} - a_{d2})X_2^{max}}{2}\left(1 - \frac{(1-\alpha)a_{d2}}{(1-\alpha)p_{d2} + \alpha a_{d2}}\right) - c_{d2} \qquad (4)$$

The total profit from both services, $\Pi_d = \Pi_{d1} + \Pi_{d2}$, is therefore

$$\Pi_d = \frac{(p_{d2} - a_{d2})X_2^{max}}{2}\left(1 - \frac{(1-\alpha)a_{d2}}{(1-\alpha)p_{d2} + \alpha a_{d2}}\right) - c_{d2} + X_1(p_{d1} - a_{d1}) - c_{d1} \qquad (5)$$

Shared Networks: The gross profit R_{s1} for Service 1 is $R_{s1} = X_1\,(p_{s1} - a_{s1})$, while that of Service 2 is given by $\mathbf{E}(R_{s2})_{[K_{s2}^*]}$ (ref. Eq. (3)). Hence, the total profit from both services is

$$\Pi_s = \frac{(p_{s2} - a_{s2})X_2^{max}}{2}\left(1 - \frac{(1-\alpha)a_{s2}}{(1-\alpha)p_{s2} + \alpha a_{s2}}\right) + X_1(p_{s1} - a_{s1}) - c_s \qquad (6)$$

The optimal network choice is the one yielding the highest overall profit. In the next section, we explore how this choice is affected by the model parameters, and in particular, α. Before proceeding, we derive some basic lemmas for our analysis whose proofs are in (Sen et al*,* 2010).

Lemma1: For both shared and dedicated networks, optimal capacity decrease as α increase.
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. With the cost of under-provisioning going down while the cost of over-provisioning remaining the same, the provider reduces the capacity it provisions upfront. In particular, when α = 1, no provisioning is needed, *i.e.* $K_{d2}^* = K_{s2}^* = 0$.
Lemma 2: If the return on capacity ($p_i/a_i, i = \{d2, s2\}$) is high, the optimal capacity decreases slowly with α when α is small and decreases very fast with α as α→1.

4. Analysis

Using the results of Section 3, we study the impact of various system parameters on the network choice. [Sen et al, 2010] shows that all cost and revenue parameters (*i.e.,* c_{d1}, c_{d2}, c_s, a_i, p_i; $i = \{d1, s1, d2, s2\}$) have a similar qualitative impact, *i.e.,* economies (diseconomies) of scope in costs favor a shared (dedicated) network. However, reprovisioning coefficient, α, can have a more interesting and subtle impact on network choice as discussed next.

4.1. Impact of Reprovisioning

To study the impact of reprovisioning coefficient (α), we substitute the expressions for K_{d2}^* and K_{s2}^* from Eq. (2) ($i = \{s2,d2\}$) into the condition $\Pi_s > \Pi_d$ to obtain:

$$a_{d2}K_{d2}^*(\alpha) - a_{s2}K_{s2}^*(\alpha) > 2\gamma \qquad (7)$$

where $\gamma = \left(p_{d2} - a_{d2} - (p_{s2} - a_{s2})\right)\frac{X_2^{max}}{2} + R_{d1} - R_{s1} - (c_{d1} + c_{d2} - c_s)$ is independent of α. γ captures the difference in expected profits between the dedicated and shared networks conditioned on capacity exactly meeting the realized demand (as would be the case when α = 1). The left hand side of Eq. (7) captures the difference in capacity costs under optimal provisioning between the dedicated and shared networks as a function of α. For ease of exposition, we introduce $h(\alpha) = a_{d2}K_{d2}^*(\alpha) - a_{s2}K_{s2}^*(\alpha)$ to denote this difference. $h(1) = 0$ as the upfront capacity provisioned is zero for both networks for α = 1, while $h(0)$ can positive or negative depending on which network incurs a higher capacity cost in the absence of reprovisioning.

As specified in Eq. (7), the network choice at any α depends on the value of $h(\alpha)$ for α in [0,1] and its position with respect to the constant baseline of 2γ. At each value of α where $h(\alpha)$ intersects 2γ, a switch occurs from preferring one network to another. Understanding network choices therefore calls for analyzing how the capacity cost difference, $h(\alpha)$, varies with α. This is the topic of Subsection 4.1.1, whose results are leveraged in Subsection 4.1.2 to enumerate possible intersection(s) of $h(\alpha)$ and 2γ and discuss their implication on network choice.

4.1.1. Analyzing the effect of α *on capacity cost difference*

The cost of capacity that needs to be provisioned upfront is decreasing with α for both shared and dedicated networks. This follows from Lemma 1 which shows that an increase in α benefits both shared and dedicated networks by helping them reduce their upfront capacity requirements, and hence their corresponding capacity ($a_iK_i^*$, $i = \{d2,s2\}$). But the difference in these costs, as captured by $h(\alpha)$, may increase or decrease as α varies in [0,1] depending on which network benefits most from an increase in α. Proposition 1 specifies the conditions under which $h(\alpha)$ is increasing (a shared network benefits more) or decreasing (a dedicated network benefits more).

Proposition 1: Increasing α benefits both shared and dedicated networks by reducing their optimal capacity costs. Additionally,

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

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

(iii) if $h'(0) \geq 0, h'(1) < 0$, 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 $h'(1) \geq 0$, 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 $h'(1)$ fully characterize the entire range of behaviors of $h(\alpha)$. Implicit in this characterization is that $h'(\alpha)$ can change its sign at most once for α in [0,1]. The proof that $h'(\alpha) = 0$ at most once for $\alpha \in [0,1]$ is given in (Sen et al., 2010).

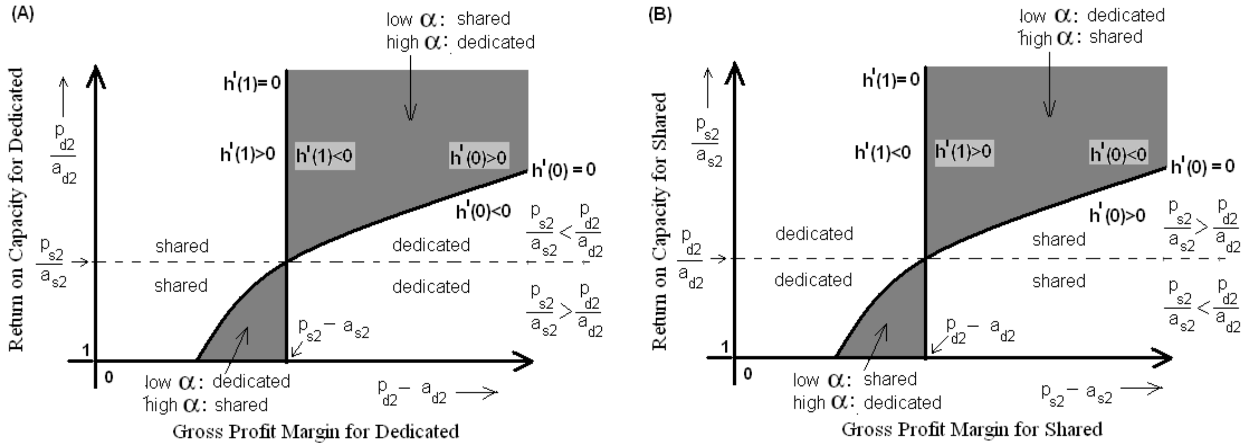

Fig. 2. Partitioning of the parameter space into cases of Prop.1: (A) ($p_{d2}-a_{d2}$, p_{d2}/a_{d2}) plane, (B) ($p_{s2}-a_{s2}$, p_{s2}/a_{s2}) plane

Proposition 1 is useful for two reasons. First, it helps clearly identify the key operational metrics that determine whether dedicated or 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 points below. The sign of $h'(\alpha)$ at α = 0 and α =1, determines which network benefits more from reprovisioning. Substituting, we get:

$$h'(0):\ \frac{p_{d2}-a_{d2}}{(p_{d2}/a_{d2})^2} = \frac{p_{s2}-a_{s2}}{(p_{s2}/a_{s2})^2} \qquad (8)$$

$$h'(1):\ p_{d2}-a_{d2} = p_{s2}-a_{s2} \qquad (9)$$

From Eqs. (8) and (9), we observe that two operational metrics, the *return on capacity* (p_i/a_i) and the *gross profit margin* for each unit of used capacity (p_i-a_i; i={s2, d2}), play key roles in determining which network choice benefits more from increases in α. In Fig. 2(A), we identify the regions in the p_{d2}/a_{d2} (y-axis) and $p_{d2}-a_{d2}$ (x-axis) plane associated with the four conditions identified in Proposition 1 (p_{s2}/a_{s2} and $p_{s2}-a_{s2}$ are held constants). Note that the y-axis only takes values greater than 1 since $\Pi_2 > 0$. A similar plot can be obtained for the plane p_{s2}/a_{s2} (y-axis) and $p_{s2}-a_{s2}$ (x-axis) as shown in Fig. 2(B). We observe from Figs.2(A) and 2(B) that the line $p_{d2} - a_{d2} = p_{s2}-a_{s2}$ partitions the plane into two regions such that dedicated network always benefits more on one side and shared network on the other. Proposition 2 formalizes this observation.

Proposition 2: A dedicated (shared) network benefits more at high α (*i.e.,* α →1) if $p_{d2} - a_{d2} > p_{s2}-a_{s2}$ ($p_{d2} - a_{d2} < p_{s2}-a_{s2}$), *i.e.,* if it has a higher (lower) gross profit margin.

For a≈1, the network choice with a higher gross profit margin will typically provision more upfront capacity and incur a higher capacity cost. When α = 1, there is no reprovisioning cost since $K_i^* = 0, i = \{s2,\ d2\}$. As a result, when α increases and approaches 1 and the upfront capacity cost drops to 0, the network choice with the greater upfront capacity (*i.e.,* with the higher gross profit margin) benefits more from reprovisioning. Next, we focus on scenarios with low α, *i.e.,* very limited reprovisioning is feasible. We observe from Figs. 2(A) and 2(B) that $h'(0) = 0$ partitions the plane into two regions.

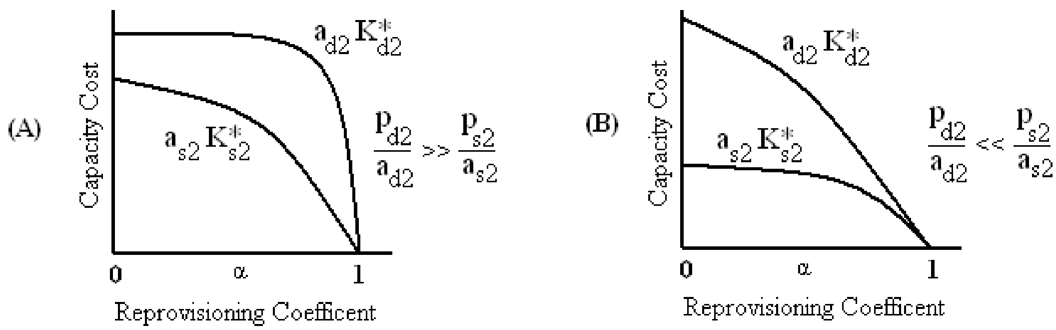

Fig. 3. (A) Shared (dedicated) benefits more at low (high) α, (B) Dedicated benefits more ∀α.

Proposition 3: A dedicated network benefits more at low α (*i.e.,* α →0) if $\frac{p_{d2}-a_{d2}}{(p_{d2}/a_{d2})^2} > \frac{p_{s2}-a_{s2}}{(p_{s2}/a_{s2})^2}$, and a shared network benefits more at low α if $\frac{p_{d2}-a_{d2}}{(p_{d2}/a_{d2})^2} < \frac{p_{s2}-a_{s2}}{(p_{s2}/a_{s2})^2}$

This indicates that besides p_i-a_i, another metric, return on capacity (p_i/a_i, $i = \{d2, s2\}$), affects which network choice benefits more from increases in reprovisioning at low α. Our analysis so far identifies which region of Fig. 2 we operate in for given values p_i/a_i and p_i-a_i. Next, we examine how these metrics influence which network choice benefits more from reprovisioning.

Consider the case in which the dedicated network enjoys a higher gross profit margin than the shared, *i.e.,* $p_{d2} - a_{d2} > p_{s2} - a_{s2}$. Suppose $p_{d2}/a_{d2} >> p_{s2}/a_{s2}$, then a scenario similar to Fig. 3(A) may arise. Since p_{d2}/a_{d2} is high, we know from Lemma 3.2 that the upfront capacity remains almost unaffected for the dedicated network at low α. Consequently, as shown in Fig. 3(A), the drop in the capacity provisioning cost for the dedicated network is not very significant at low α. On the other hand, since the shared network has a much lower return on capacity, its capacity and the associated provisioning cost decreases faster with increases in α at low α. But when α is high, the dedicated network starts to benefit more because the capacity requirements drop from a relatively high value to zero. Hence, shared network benefits more from increases in α at low α while the dedicated does so at high α. This observation is consistent with Fig. 2(A), where $p_{d2}-a_{d2} > p_{s2}-a_{s2}$ and $p_{d2}/a_{d2} >> p_{s2}/a_{s2}$ results in a point in the shaded upper right hand side region.

Next consider the case $p_{d2}/a_{d2} << p_{s2}/a_{s2}$. In Fig. 3(B), since p_{s2}/a_{s2} is high, the capacity cost of the shared network remains unaffected at low α in accordance with Lemma 3.2. But the dedicated network, which has a lower return on capacity, drops its capacity faster at low α. Moreover, in accordance with Proposition 2, the dedicated network continues to benefit more from increases in α even as α →1 since $p_{d2}-a_{d2} > p_{s2}-a_{s2}$. Thus, in this scenario, the dedicated network benefits ∀α ∈[0,1]. Once again, this is consistent with Fig. 2(A), where $p_{d2}-a_{d2} > p_{s2}-a_{s2}$ and $p_{d2}/a_{d2} << p_{s2}/a_{s2}$ correspond to a point in the bottom right hand side region. Similar explanations apply for other conditions, *e.g.,* $p_{d2}-a_{d2} < p_{s2}-a_{s2}$, and how they map to the regions of Fig. 2.

4.1.2 Implications for Network Choice

Section 4.1.1 characterized which network choice benefits more from improvements in reprovisioning. However, the provider's optimal network choice depends on how these relative benefits compare to other cost and revenue parameters. From Eq. (7), this choice depends on the value of $h(\alpha)$ with respect to the baseline of 2γ, and each of their intersections mark a switch in the network choice. As specified in Proposition 1 (also see Fig. 2(A)), there are four possible behaviors associated with an increase in reprovisioning coefficient, α.

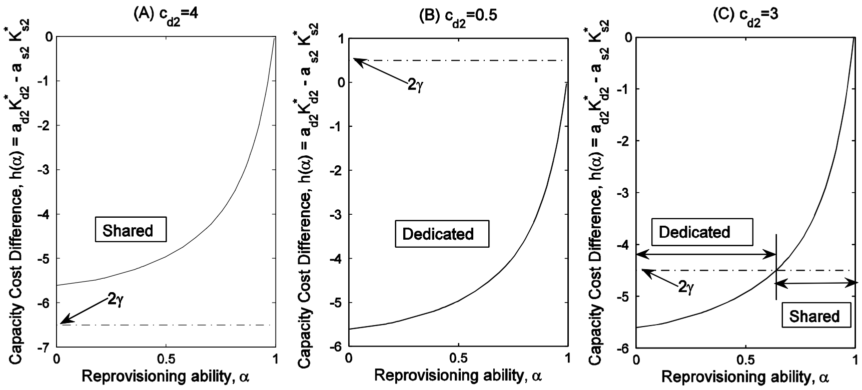

Fig. 4. (A) Shared ∀α ∈[0,1], (B) Dedicated ∀α ∈[0,1], (C) Dedicated (shared) at low (high) α.

First, consider the region in which a shared network always benefits more from increases in α. Now if a shared network is already the preferred choice at α = 0, then it obviously remains the optimal network choice irrespective of reprovisioning ability. This requires γ<0 (because $h(\alpha) > 2\gamma$, ∀α and h(1)=0) which can arise if the shared network enjoys significantly lower fixed costs ($c_s << c_{d1} + c_{d2}$) or variable costs ($p_{s2}-a_{s2} >> p_{d2}-a_{d2}$ and/or $R_{s1} >> R_{d1}$). A numerical example is shown in Fig. 4(A) (where $p_{d2}-a_{d2}$=9<16.7= $p_{s2}-a_{s2}$, p_{d2}/a_{d2}= 10 > 8.2609 = p_{s2}/a_{s2}, R_{d1} = 30, R_{s1} =25, c_s =20, c_{d1} =5, c_{d2}=4). 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.,* $h(\alpha)<2\gamma$,∀α ∈[0,1]), then a dedicated network remains preferred irrespective of α, as in Fig. 4(B) (where a low fixed cost, c_{d2} = 0.5 favors a dedicated network). A more interesting outcome arises when a dedicated network is the preferred choice for α = 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). Consequently, the network choice switches to a shared network at high α. An example is shown in Fig. 4(C), in which dedicated network is preferred for α<0.6 and shared for higher values.

Second, consider the region in which a shared network benefits more from increases in α at low α and dedicated at high α. An example of this is shown in Fig. 5 (where $p_{d2}-a_{d2}$=18> 11 =$p_{s2}-a_{s2}$, p_{d2}/a_{d2}=10>6.5=p_{s2}/a_{s2}, R_{d1} =25, R_{s1} =25, c_s =14, c_{d1} =11), which shows that there are four possible outcomes depending on 2γ: (i) dedicated network is preferred irrespective of α, (ii) shared network is preferred irrespective of α, (iii) shared is preferred at low α and dedicated at high α, (iv) dedicated is preferred both at low and high α, and shared for intermediate values. A dedicated (shared) network is chosen irrespective of α if there are significant diseconomies (economies) of scope as shown in Fig. 5(A) (Fig. 5(B)). 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. Fig. 5(C) and (D) show that there can be one or two such switches in the optimal network choice. Next, we consider the region in which a dedicated network always benefits more than a shared network ∀α ∈[0,1]. In this scenario, if the diseconomies (economies) of scope in the costs are very large, a dedicated (shared) network is preferred irrespective of α, else the network choice switches from shared to dedicated as α increases. Lastly, when a dedicated network benefits more from increases in α at low α and a shared network at high α, there can be four possible outcomes. The analyses for these last two cases are analogous to the previous ones in which the shared network was always benefiting more.

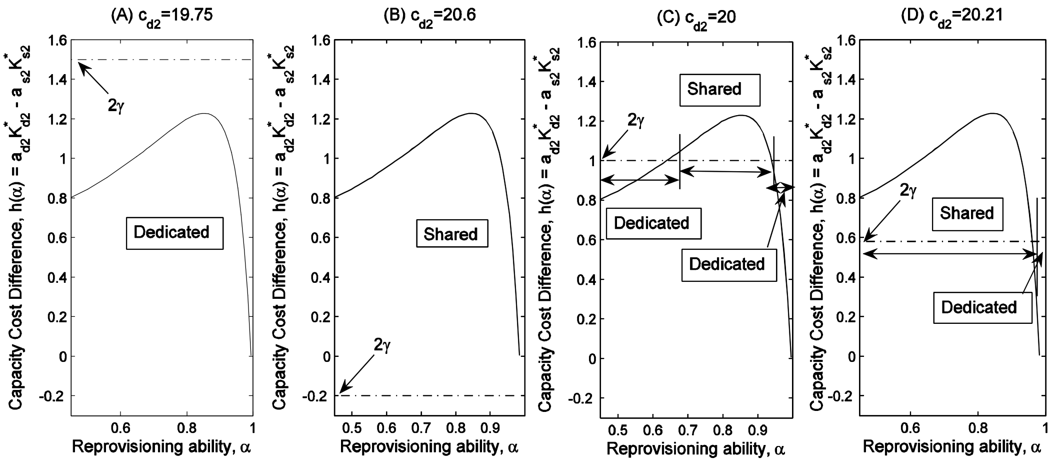

Fig. 5. (A) Dedicated network ∀α ∈[0,1], (B) Shared ∀α ∈[0,1], (C) Dedicated at both high and low α, shared for intermediate α, (D) Shared (dedicated) at low (high) α.

5. Conclusions

This work introduces an analytical framework to investigate the trade-offs between deploying new services on shared versus dedicated infrastructures. Our analysis clearly shows that new technological factors like the ability to dynamically reprovision resources, which was absent in earlier works, can have significant impact on network choices. We show that reprovisioning can lead to situations where one network choice is preferred when this ability is limited, while the other is preferred when it is easier to reprovision. We show that even more complicated decisions may arise due to reprovisioning. We find that the impact of reprovisioning is mediated by two key operational metrics, *gross profit margin* and *return on capacity* associated with each of the networks. In our future work, we will consider the provider's ability for downward adjustments of resources and allow the shared and dedicated networks to have different reprovisioning capabilities. We will also study the implications of our results in choosing between shared or dedicated infrastructures in the context of latest network advances like cloud computing etc.

References

1. Anderson, T., Peterson, L., Shenker, S., Turner, J., "Overcoming the Internet impasse through virtualization," *Computer*, vol.38, no.4, April 2005, pp. 34- 41.
2. AT&T, "What is IPTV?", att.com/Common/merger/files/pdf/IPTV_background.pdf, 2009.
3. Brandel, M.. "Facilities management goes green," *computerworld.com,* 25 June 2007.
4. Crosby, T., "How FiOS works," *HowStuffWorks.com*, 7 August 2008.
5. Fine, C. H., and Freund, R. M., "Optimal investment in product-flexible manufacturing capacity," *Management Science*, 36(4), 1990, pp. 449–465.
6. Koebbe, P. H., "A case for separate networks," *Communications News*, 1 November 2007.
7. Laoutaris, N., Smaragdakis, G., Rodriguez, P., and Sundaram, R., "*Delay tolerant bulk data transfers on the Internet," SIGMETRICS '09*, 2009, pp. 229–238.
8. Sen, S., Yamauchi, K., Guerin, R. and Hosanagar, K. "Supplementary Document: Proofs and numerical investigations," www.seas.upenn.edu/~ssoumya/WEB10_additional.pdf, 2010.
9. Touch, J., Wang, Y., Eggert, L., and Finn, G., "A virtual Internet architecture," *Proc. ACM FDNA'03*, Germany, 2003.
10. Van Mieghem, J. A., "Investment strategies for flexible resources," *Management Science*, 44(8), 1998, pp. 1071–1078.