The Stability of Best Effort and Managed Services in the Internet, and the Role of Application Creation

Debasis Mitra
Department of Electrical Engineering
Columbia University
New York, NY 10027
email: debasismitra@columbia.edu

Qiong Wang
Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801
email: qwang04@illinois.edu

***Abstract*—Best Effort service has been the bedrock of the Internet. The simultaneous offering of Managed Service has been proposed as an efficient way to support quality of service, to make some applications feasible and many others more attractive, and give consumers greater choice. However, there is widespread concern that if a network provider is allowed to offer Managed Service and charge a per-use fee, then the network provider will find it in its short-term profit-maximizing interest to degrade the Best Effort service. Not only will consumers then face higher costs, but also fewer innovations and new applications will be created. To investigate these issues, we formulate a model in which the provider myopically maximizes its profit by setting the per-use fee and deciding on the bandwidth to provision for the Best Effort and Managed Services. We show that in the optimal solution the service offering is binary, i.e., either the Best Service or the Managed Service is offered, but not both simultaneously. Moreover, a necessary and sufficient condition gives the optimal service offering. Combining this condition with a birth-and-death model of applications, we identify a natural attractor around which the application creation process makes the service offering stable. Our finding alleviates the concern about the Managed Service. Still, the provider needs tow have foresight and the regulator needs to be vigilant, since we also show that there is a threshold number of applications belo which the stabilizing mechanism loses its effectiveness and myopic profit-maximization results in a shrinking network and application pool, which hurts both the provider and consumers.**

I. Introduction

There is general consensus that the low flat subscription fee for connections and the absence of a usage fee have been essential contributing factors to the Internet's explosive growth, as well as the immense innovations and creativity that the Internet has spawned. The mode of the offered service has been Best Effort, i.e., without guarantees on service features, such as jitter, latency, throughput, reliability and security. As the Internet evolves, new applications and needs have argued for the offering of Managed Services for applications, which offer higher assurances, if not guarantees, of previously missing service features, but at the cost of premium rates for usage of the managed applications [6]. These premium rates compensate the last-mile, broadband access provider for provisioning and reserving network resources, such as bandwidth, which is the proxy network resource that we will be considering here. All subscribers to the Best Effort service share the bandwidth allocated to the service, which makes the latency, for example, unpredictable, and also unavoidable if Best Effort is the only service offering. In contrast, subscribers to the Managed Service, by virtue of the network resource, i.e., bandwidth, dedicated to it, are unaffected by interference from usage by others. This makes some applications feasible; additionally, it gives users the option to trade-off lower delays for usage fees.

However, the very idea of broadband access providers offering Managed Services in conjunction with Best Effort has raised many red flags among regulators and the general public ([1],[5]). At a visceral level the concern is that disturbing the model for the Internet that has served society so well is fraught with danger. Specifically, a major concern is that the profit-maximizing service provider will take the "damaged goods" approach wherein it intentionally reduces the quality of Best Effort service (most easily, by reducing the bandwidth allocated to it) to force customers to choose the more expensive (and more profitable) Managed Service. The first reaction, also visceral, is to consider regulation. Yet we know from long experience that regulation comes at a high cost [2]. But even if the regulatory path is to be followed, the regulators will need guidance and a better understanding than is available today of the basic dynamics of interactions between customers and bandwidth access providers.

Our point of view is that quite basic issues have not yet been resolved. For instance, it has not been resolved satisfactorily if the access provider's profit-maximizing path is to squeeze the Best Effort service by starving it of bandwidth. After all, Best Effort multiplexes bandwidth, which is inherently efficient, whereas Managed Service uses dedicated bandwidth, which reflects a longstanding choice in networking with different end results at different eras. However, we believe that a little-understood, yet quite consequential, aspect is the interaction of the bandwidth allocation decision with the size of the pool of applications, the source of much of the utility of the Internet to the provider's customers. For instance, the possibility exists (and has not been ruled out) that even a myopic profit-maximizing broadband access provider will find in its interest to sustain the Best Effort service for the benefit of the spawning of applications from its usage. If fortuitously

978-1-4799-3088-3/14/$31.00 ©2014 IEEE

this turns out to be the case, no regulation would be necessary. Even otherwise, it may be possible to glean from the analysis just what minimal regulation is required. The central focus of our work has been on understanding this interaction between bandwidth allocation decisions, pricing, usage and spawning of applications, especially in its impact on the sustainability (or stability) of the Best Effort service in the Internet.

Recently we did an initial study of this subject. We considered the case where the provider is motivated by short-term profit maximization with no regard to the long-term implications of the creation of new applications. We showed that when allocating a fixed amount of bandwidth to the Best Effort and Managed services, this profit incentive drives the provider to favor the Best Effort service when there are very few applications. The increased allocation to the Best Effort service, and the consequent improvement in its quality of service, accelerates the spawning of new applications. When the number of active applications reaches a certain threshold, it becomes more profitable for the provider to shift its preference to the Managed Service to benefit from higher usage revenue. From these results, we concluded that even when the network provider does not explicitly take application innovations into account, it still shapes the provider's bandwidth allocation decision, rectifies its shortsighted near-term profit maximization, and sustains the Best Effort service.

This paper continues our prior work [4] by extending it to a new context. Instead of the network having a fixed amount of bandwidth, we consider the network's bandwidth to be part of the provider's decision. Specifically, the provider decides on the size of the networks bandwidth in return for payment of a proportional "rental fee" or "maintenance cost". Unchanged from the earlier model, the bandwidth provider decides on the allocation of the network bandwidth to the Best Effort and Managed Service and the per-use fee of the managed applications. Also as before, all applications are offered in two service offerings, Best Effort and Managed Services, and there exists a flat fee that users pay to the bandwidth provider to be connected to the Internet. This subscription fee is assumed to be regulated and outside the control of the broadband provider. Since this fee is revenue for the bandwidth provider, it is in its interest to have a large subscriber base, regardless of the subscribers' service choices.

This change in the model leads to new, qualitatively distinct consequences. When paying for bandwidth is an option, and not a sunk cost, the provider's preference becomes more polarized. In some cases, myopic profit maximization makes it desirable to provide bandwidth only for the Managed Service, while in all other cases exactly the opposite is true. The impact of new application spawning also involves more intricacy. While the aforementioned stabilization effect is still present when the network starts from a sufficiently high number of applications, when the latter number gets to be small, then, under myopic profit maximization, a loop of downward spiral may materialize between the bandwidth provisioning to the Best Effort service and the process of spawning new applications, driving the network towards extinction.

These new observations are largely a consequence of economy of scale, which is a salient feature of the Best Effort service. Given that the bandwidth of the Best Effort service is fully shared, its marginal cost is small if the service is operated at a large scale, but can be substantially higher if the service is used by a smaller population with fewer applications. This is not a concern when the provider operates a fixed-size network, in which case a dwindling number of new applications implies the usage revenue from the Managed Service is also diminishing. Hence allocating existing bandwidth to the Best Effort service becomes more attractive. More bandwidth means lower congestion and higher usage, which stimulates the growth of new applications. However, if the amount of the bandwidth is to be decided and paid for by the provider, then at least in the case of myopic profit maximization, there can be a point of no return once the usage of applications fall below a certain threshold. In this regime of low usage the economy of scale is lost, and with it the incentive to provision bandwidth to the Best Effort service, leading to an even smaller number of the applications and less usage, which forges a permanent decline.

These findings have strong managerial and policy implications. While illustrating the power of new application innovations in sustaining the Best Effort service, they also show the limits of its influence. The provider needs to be farsighted and the regulator needs to be vigilant not to slip into the regime of contracting bandwidth and usage for the Best Effort service, and the number of applications, which ultimately will be a loss to the business, consumers and society.

While much more can be said about our findings, the goal of this short paper is limited. We build a formal model to demonstrate the aforementioned points by a mix of analytic discussions and numerical examples. The consumer self-optimization model, with a different utility function from our previous study, will be presented in Section II, the provider's short-term profit maximization problem is discussed in Section III, and the interaction between profit maximization and application innovations will be analyzed in Section IV.

II. CONSUMERS: SUBSCRIPTION AND USAGE

We consider a fixed number of potential customers, some of them pay the subscription fee s at regular intervals, for access to a broadband network. The number of potential customers is normalized to unity. The network offers a set of applications, the number of which is denoted by A. For simplicity, assume that all consumers have the same exponential utility function from the use of each application,

$$u(\lambda) = \omega(1 - e^{-\gamma\lambda}), \tag{1}$$

where λ is the number of uses per unit of time (the usage rate). The maximum utility $\omega(> 0)$ is reached when $\lambda = \infty$ and the maximum marginal utility, $\omega\gamma$, is reached at $\lambda = 0$.

Let D be a QoS-related performance measure, such as the mean delay, that captures the (dis)quality of the Best Effort service. Subscribers differ in their tolerance to delay, and the difference is modeled by associating subscribers with index θ, which is assumed to be uniformly distributed over $[0, \bar{\theta}]$. Given

D, θD is the delay cost per use of subscriber θ. When using the Best Effort service, each subscriber chooses the usage rate to maximize her surplus

$$v^{BE}(\lambda;\theta) = u(\lambda) - \theta\lambda D, \quad \theta \in [0, \bar{\theta}]. \tag{2}$$

The optimal usage rate of each application is

$$\lambda^{BE*}(\theta) = \frac{1}{\gamma}\ln\frac{\omega\gamma}{\theta D}, \tag{3}$$

and the optimal surplus from using all A applications,

$$\begin{aligned} v^{BE*}(\theta) &= A[u(\lambda^{BE*}(\theta)) - \theta\lambda^{BE*}(\theta)D] \\ &= A\left[\omega - \frac{\theta D}{\gamma}(1 + \ln\frac{\omega\gamma}{\theta D})\right]. \end{aligned}$$

Both the optimal usage and surplus decrease with θ. The optimal usage rate of those with infinite tolerance to delay is infinity, i.e., $\lambda^{BE*}(0) = \infty$. These users achieve the maximum surplus $A\omega$. Since the distribution at $\theta = 0$ has zero mass, the total consumption of bandwidth by these users is negligible, which will be evident in the discussion below.

We next consider three cases of service offerings in the network: Best Effort service only, Managed Service only, and both services. When the network offers the Best Effort service only, consumers subscribe to the network if and only if they can derive an optimal surplus that exceeds the subscription fee. These customers θ are in the range $[0, \theta_B \wedge \bar{\theta}]$ where θ_B is determined by

$$s = A\left[\omega - \frac{\theta_B D}{\gamma} - \frac{\theta_B D}{\gamma}\ln\frac{\omega\gamma}{\theta_B D}\right]. \tag{4}$$

A profit-maximizing provider will always keep $\theta_B \le \bar{\theta}$, which is the situation we consider. In this case,

$$\Lambda^{BE} = \int_0^{\theta_B} \lambda^{BE*}(\theta)d\theta = \frac{A}{\gamma}\left(\ln\frac{\omega\gamma}{\theta_B D} + 1\right)\theta_B. \tag{5}$$

We use the mean sojourn time in the M/M/1 queueing model as the proxy of the (dis)quality. Let B^{BE} denote the amount of bandwidth provisioned to the Best Effort service, so that

$$D = \frac{\Lambda^{BE}/B^{BE}}{B^{BE} - \Lambda^{BE}} = \frac{1}{B^{BE}(B^{BE}/\Lambda^{BE} - 1)}. \tag{6}$$

In the parameter region $s \le A\omega$ and $B^{BE} > 0$, θ_B, Λ^{BE}, and D are uniquely determined by (4), (5), and (6). It is also easy to verify that the delay D and the total usage Λ^{BE} both increase in the number of applications A and decrease in the subscription fee s, as one would expect.

Next we consider the case where only the Managed Service is offered. Consumers choose the optimal usage rate to maximize their surplus and subscribe to the network only if the optimized surplus is not less than the subscription fee. Specifically, with quality of service guaranteed and the per-use fee of p, a user maximizes

$$u(\lambda) - p\lambda = \omega(1 - e^{-\gamma\lambda}) - p\lambda.$$

The optimal usage rate of each application is

$$\lambda^{MS*}(p) = \frac{1}{\gamma}\ln\frac{\omega\gamma}{p}, \tag{7}$$

and the optimal surplus from using all A applications is

$$\begin{aligned} v^{MS*}(p) &= A\left[u(\lambda^{MS*}(p)) - p\lambda^{MS*}(p)\right] \\ &= A\left[\omega - \frac{p}{\gamma}(1 + \ln\frac{\omega\gamma}{p})\right], \end{aligned}$$

which is the same for all consumers. Therefore if

$$v^{MS*}(p) \ge s, \tag{8}$$

then all consumers are willing to subscribe. To guarantee quality of service, bandwidth is provisioned on the per-use basis, and denoted by b^{MS}. Let $[0, \theta_s]$ be the segment of consumers who subscribe to the Managed Service. The amount of bandwidth needed to serve these customers, $b^{MS}A\lambda^{MS*}(p)\theta_s$, cannot exceed B^{MS}, the amount of bandwidth provisioned to the Managed Service. Thus

$$\theta_s = \min\left(B^{MS}/(b^{MS}A\lambda^{MS*}(p)), \bar{\theta}\right). \tag{9}$$

Finally, we consider the case where both services are offered. Consumers compare the surplus they can get from each service to determine whether to subscribe, and if so, to which service. At the same usage rate, a customer gets the same utility but pays different costs for different services, the per-use fee p for the Managed Service or a delay cost θD for the Best Effort service. The two costs are equal for

$$\theta_b = \frac{p}{D}, \tag{10}$$

with customer $\theta \le \theta_b$ preferring the Best Effort service to the Managed Service, and the preference is reversed for $\theta > \theta_b$.

For the Managed Service to attract any consumer,

$$s \le A\left[\omega - \frac{p}{\gamma}(1 + \ln\frac{\omega\gamma}{p})\right]. \tag{11}$$

Consumers in the range $[0, \theta_b]$ select the Best Effort service and those in the range $[\theta_b, \theta_s]$ select the Managed Service. Here θ_s is given by

$$\theta_s = \min\left(B^{MS}/(b^{MS}A\lambda^{MS*}(p)) + \theta_b, \bar{\theta}\right), \tag{12}$$

which includes (9) as a special case where $\theta_b = 0$. To justify subscription to the Best Effort service,

$$s \le A\left[\omega - \frac{\theta D}{\gamma}(1 + \ln\frac{\omega\gamma}{\theta D})\right], \ 0 \le \theta \le \theta_b, \tag{13}$$

which follows from (10) and (11).

III. SERVICE PROVIDER:MYOPIC PROFIT MAXIMIZATION

Under the current regulatory environment, a network provider is not free to change the subscription fee at will. This restriction is reflected in our model by the assumption that s is a fixed parameter. The provider makes three decisions. For the Managed Service, the provider sets p, the per-use fee, and B^{MS}, the amount of bandwidth provisioned. For the Best Effort service, the provider does not charge a usage fee and decides only the amount of provisioned bandwidth, B^{BE}.

The provider is assumed to be myopically maximizing its immediate profit, including the revenue from the Managed

Service, which is collected from both the subscription and usage charges, the revenue from the Best Effort service, which is from subscription only, and the costs of provisioning bandwidth to both types of services. Here we assume the cost per unit of bandwidth is constant and denote it by r.

We consider three cases of service offerings in the network provider's profit maximization: Best Effort service only, Managed Service only, and both services. Below we discuss each case separately, followed by a summary on which case applies under different conditions.

A. Best Effort Service Only

If the provider offers only the Best Effort service, then $B^{MS} = 0$ and the provider's profit

$$\Omega(B^{BE}) = s\theta_B - rB^{BE}, \tag{14}$$

where given B^{BE}, θ_B, together with D and Λ^{BE*}, are jointly determined by (4), (5), and (6).

To determine the outcome of maximizing $\Omega(B^{BE})$, solve (6) and use the solution

$$B^{BE} = \frac{\Lambda^{BE*}}{2}\left(1 + \sqrt{1 + 4/(\Lambda^{BE*}D)}\right). \tag{15}$$

Given D, θ_B can be obtained by solving (4) and Λ^{BE*} follows from (5), so the profit (14) can be maximized by choosing the optimal congestion level D instead of the bandwidth B^{BE}.

The maximum congestion cost from delay is paid by the marginal subscriber to the Best Effort service. This amount is

$$x = \theta_B D. \tag{16}$$

Let $g(x) = 1 + \ln\frac{\omega\gamma}{x}$, where $\omega\gamma/x$ is the marginal benefit-cost ratio of the marginal subscriber. From (4), (14), and (15),

$$\Omega(B^{BE}, \theta_B) = AG(x)\theta_B. \tag{17}$$

where

$$G(x) = \omega - \frac{xg(x)}{\gamma} - r\frac{g(x)(1 + \sqrt{1 + 4\gamma/(xAg(x))})}{2\gamma}. \tag{18}$$

Since $x = \theta_B D$ is uniquely determined by solving (4) and depends only on parameters ω, γ, and s, the solution that maximizes the profit given in (14) is either to serve all consumers or none, i.e.,

$$\theta_B^* = 0 \text{ if } G(x) \le 0, \quad \text{and otherwise} \quad \theta_B^* = \bar{\theta}. \tag{19}$$

This important result reflects the economy of scale in providing the Best Effort service. If the provider can break-even to serve one user, adding more subscribers drives down the marginal cost of bandwidth, making it economical to serve all consumers.

B. Managed Service Only

When the network offers only the Managed Service, $B^{BE} = 0$. The provider's profit is

$$\Omega(p, B^{MS}) = [A\lambda^{MS*}(p)p + s]\bar{\theta}_s - rB^{MS}, \tag{20}$$

where B^{MS} is given by (9). Also, to justify subscription, (8) needs to be satisfied.

From this profit function, it may be deduced that if the optimal per-use fee is p^*, then it must be the case that

$$p^* - rb^{MS} + s/(A\lambda^{MS*}(p^*)) \ge 0.$$

Hence (20) increases in θ_s, and it is optimal to set $\theta_s = \bar{\theta}$. Consequently, the profit function simplifies to

$$\Omega(p) = [A\lambda^{MS*}(p)(p - rb^{MS}) + s]\bar{\theta}. \tag{21}$$

Since $\Omega(p)$ is concave in p,

$$p^* = \min(p^o, p^s), \tag{22}$$

where p^s is the maximum value that satisfies (8), which is the unique value of p for which equality holds, i.e.,

$$s = v^{MS*}(p^s) = A\left[u(\lambda^{MS*}(p^s)) - p^s\lambda^{MS*}(p^s)\right], \tag{23}$$

and p^o optimizes $\Omega(p)$ without the constraint (8), i.e.,

$$p^o = \arg\max\left\{\lambda^{MS*}(p)(p - rb^{MS})\right\}. \tag{24}$$

Obviously, $p^o > rb^{MS}$.

C. Both Services

When both services are offered, the profit function becomes

$$\begin{aligned}\Omega(p, B^{MS}, B^{BE}) &= [s + pA\lambda^{MS*}(p)](\theta_s - \theta_b) \\ &\quad + s\theta_b - r(B^{MS} + B^{BE}),\end{aligned} \tag{25}$$

(recall that $[0, \theta_b]$ is the segment of consumers subscribing to the Best Effort service and $(\theta_b, \theta_s]$ is the segment subscribing to the Managed Service.)

Following the same discussion that establishes (21), either $\theta_s - \theta_b = 0$, in which case the network offers the Best Effort service only, or θ_s should be at its maximum value $\bar{\theta}$. In the latter case, using (15) to replace B^{BE}, $\theta_b = p/D$ to replace D, and (12) to replace B^{MS}, the profit function becomes

$$\begin{aligned}\Omega(p, \theta_b) &= A\lambda^{MS*}(p)(p - rb^{MS})(\bar{\theta} - \theta_b) \\ &\quad - r\frac{\Lambda^{BE}(\theta_b)}{2}\left(1 + \sqrt{1 + 4\frac{\theta_b}{p\Lambda^{BE}(\theta_b)}}\right) + S,\end{aligned} \tag{26}$$

where $S = s\bar{\theta}$ is the total revenue from the subscription fees. Here $\theta_b \in [0, \bar{\theta}]$ and the choice of p is subject to $0 \le p \le p^s$ where p^s is the unique value for which (11) holds at equality.

By observing from (3), (7), and $\theta_b D = p$ that

$$\begin{aligned}\Lambda^{BE}(\theta_b) &= A\int_0^{\theta_b}\lambda^{BE*}(\theta)d\theta = A\int_0^{\theta_b}\frac{1}{\gamma}\ln\frac{\omega\gamma}{\theta D}d\theta \\ &= \frac{A}{\gamma}[\ln\frac{\omega\gamma}{\theta_b D} + 1]\theta_b = \frac{A}{\gamma}[\ln\frac{\omega\gamma}{p} + 1]\theta_b \\ &= A\left(\lambda^{MS*}(p) + 1/\gamma\right)\theta_b,\end{aligned}$$

the profit function can be further simplified to

$$\Omega(p, \theta_b) = A\lambda^{MS*}(p)(p - rb^{MS})\bar{\theta} + S + AK(p)\theta_b, \quad (27)$$

where

$$K(p) = \lambda^{MS*}(p)(rb^{MS} - p) - \frac{r(\lambda^{MS*}(p) + 1/\gamma)}{2}\left(1 + \sqrt{1 + \frac{4}{pA(\lambda^{MS*}(p) + 1/\gamma)}}\right).$$

The above is a central result of this paper. It indicates an important property of the network provider's profit-maximizing solution, which is that the selection of the service offering is binary. When $K(p^*) \leq 0$, it is optimal to set $\theta_b^* = 0$, i. e., the provider should only offer the Managed Service. When $K(p^*) > 0$, the optimum is reached when $\theta_b^* = \bar{\theta}$, i.e., the provider should only offer the Best Effort service.

For instance, the former solution is the case when in (22), $p^* = p^o > rb^{MS}$: the Managed Service allows the provider to collect all subscription revenue and also make an extra profit from the usage, in which case there is no incentive to offer the Best Effort service. On the other hand, it is also possible that $p^* = p^s$, and the latter value causes $K(p^*)$ to be positive. In this case, the cost of providing dedicated bandwidth is too high, even after being compensated by the usage charge of the Managed Service, and cannot compete with the economy of scale from the shared use of bandwidth. Consequently the Best Effort service becomes the dominant solution.

To summarize, from the perspective of myopic profit maximization, it is not optimal for the service provider to offer both services simultaneously. To induce consumers to subscribe to the broadband network, the provider needs to deliver to them a certain amount of surplus. The provisioning of the service that allows the surplus to be generated at the minimum cost is the favored choice.

IV. Dynamic Process

In the previous section we showed that the optimal solution is binary. Nevertheless, when the optimal solution differs from the current offering, the provider can only migrate viscously, i.e., not instantly move to the "most-favored" service. During the migration process, parameters, in particular the number of applications, will change, which may alter the optimal solution and suggest to the provider a change in the direction. In this section, we examine this dynamic process.

We start by considering the dependence of the optimal per-user fee, p^*, on the number of applications A. Following (22) and (24), p^o is determined independently from A. If there is a sufficiently large number of applications such that

$$A \geq s\left(\omega - \frac{p^o}{\gamma}(1 + \ln\frac{\omega\gamma}{p^o})\right)^{-1}, \quad (28)$$

then (11) indicates that consumers get a positive surplus from subscribing to the Managed Service. Moreover, since $p^o > rb^{MS}$, $K(p^o) < 0$, and thus to myopically maximize its profit, the provider should offer the Managed Service only and charge the per-use fee p^o.

However, if the number of applications is smaller than the number given in (28), then the optimal per-use fee of the Managed Service is p^s, which is determined from (23). The latter equation indicates

$$\frac{dp^s}{dA} = \frac{\omega\gamma - p^s(\ln(\omega\gamma/p^s) + 1)}{A\ln(\omega\gamma/p^s)} = \frac{s}{A^2\ln(\omega\gamma/p^s)} > 0,$$

which means that the per-use fee in this parameter region increases with the number of applications available to the subscribers. It then becomes conceivable that by affecting p^s, the number of applications influences the value of $K(p^*)$ defined in the last section, and thereby the service offering. That is, for some number of applications, the solution is to offer the Best Effort service only, while in other cases, it is optimal to offer the Managed Service only.

To illustrate this influence, consider a numerical example with the following parameter values:

$$\omega = 10, \ \gamma = 0.5 \ , r = 1.5, \ b^{MS} = 2 \ \text{ and } s = 4.$$

We vary the number of applications and plot the resulting values of $K(p^*)$ in Figure 1. The key observation from the figure is that $K(p^*) > 0$ if and only if the number of applications is between two threshold values, $\underline{A}$ and $\bar{A}$. Recall from (21) that in the myopic profit maximization, it is optimal to offer the Managed Service only if $K(p^*) \leq 0$ and the Best Effort service only if $K(p^*) > 0$.

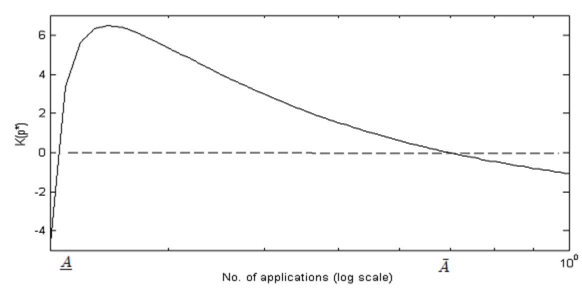

Fig. 1. Plot of $K(p^*)$ versus A, no. of applications, illustrating the change of the optimal service choice with respect to A

Here the Managed Service is the optimal choice if the number of applications is sufficiently large ($A \geq \bar{A}$). The conclusion is rather intuitive. In this case, the provider sets a high per-use fee. Subscribers will not make heavy use of each application, but even so, with the number of applications large and marginal utility high, enough surplus is generated to justify the subscription fee. As a consequence, the per-use fee is sufficient to cover the cost of bandwidth requirement for the Managed Service, and generates a new stream of profit to supplement revenue from the subscription.

As the number of applications gets smaller, the provider has to reduce the per-use fee. Subscribers derive their surplus from more frequent use of each application. The reduced usage fee may not be sufficient to recover all the cost of provisioning dedicated bandwidth for the Managed Service. In this case, the Best Effort service, which shares the use of bandwidth, can be more cost-effective and surpasses the Managed Service as the solution to profit maximization.

However, while shared usage produces multiplexing gains, it is associated with negative externality, i.e., the congestion cost that users impose on each other. When the number of applications is very small ($A \leq \underline{A}$), subscribers derive their surplus from extensive uses of each available application. The marginal utility is low, so for such uses to take place, the impact of negative externality needs to be substantially reduced, which is achieved by adding more bandwidth to mitigate the congestion. With reduced marginal utility and tightened tolerance to congestion, for the same amount of the surplus obtained, the amount of bandwidth needed increases sharply. As the number of applications reduces, at a certain point it becomes economical again to switch from shared free usage of bandwidth to controlled access, i.e., to replace the Best Effort service with the Managed Service.

We now take a broader perspective of the above discussions. Network applications are normally not imported from outside, but endogenously generated from the usage. As many have argued, the Best Effort service, with no per-use fee attached and allowing more flexible uses of bandwidth, gives subscribers more freedom to innovate, and hence is the major incubator of new network applications [3]. Network applications also have limited life times, so the pool of active applications shrinks in the absence of constant replenishment of new applications. As a first-order approximation, this process may be formulated by the following differential equation

$$\frac{dA}{dt} = \delta \Lambda^{BE*} - \mu A, \qquad (29)$$

where parameter $\delta > 0$ links the usage of the Best Effort service to the "birth" of new applications and $\mu > 0$ is the "death" rate of the existing applications.

Combining the dynamics described in (29) with myopic profit maximization leads to a feedback loop shown in Figure 2. In this scenario,the network starts from a reasonably large number of applications, e.g., greater than the lower threshold in the example of Figure 1, $\underline{A}$. Furthermore we assume that the network provider cannot switch service offerings instantaneously, and it uses the results of myopic profit maximization as signals to set direction to its provisioning decisions. For instance, when $K(p^*) > 0$, so the optimal solution is to offer the Best Effort service only, the provider increases the bandwidth provisioned for the Best Effort service and reduces the amount for the Managed Service. These amounts are subject to a maximum rate limit, so service replacements take place gradually, i.e., viscously. A similar process occurs when $K(p^*) \leq 0$,

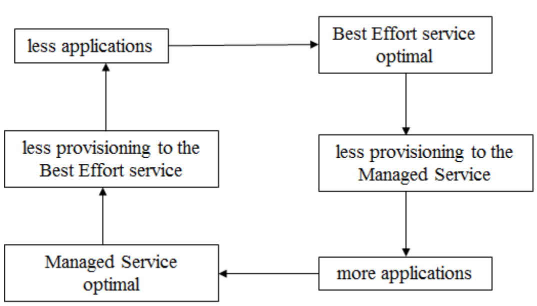

Fig. 2. Feedback Loop

The process has a natural attractor, i.e., stable point, $\bar{A}$ shown in Figure 1. The stabilization is achieved via the feedback mechanism in Figure 2. When the number of applications, $A > \bar{A}$, $K(p^*) < 0$, so the provider moves towards to the Managed Service. The reduction of the Best Effort Service causes the number of applications to drop. Once A falls below $\bar{A}$, $K(p^*) > 0$, in which case the provider begins to offer more Best Effort service, which makes A grow again.

The existence of this feedback mechanism should be welcomed by both the network provider and the regulator. For the provider, the mechanism aligns its incentive to maximize the short-term profit with it long-run interest. The regulator often faces a dilemma, i.e., while the Managed Service can be more efficient in both technology and economics than the Best Effort service, the fear is that allowing its introduction may force out the Best Effort service. This feedback mechanism can drive the provider towards a balanced offering of both services, even if the provider is oblivious to the role played by network application innovations.

It is important to observe that the feedback mechanism does not work unconditionally. In particular when the number of application falls below $\underline{A}$, then myopic profit maximization will require reduction of the Best Effort service, further reducing the number of applications. Unless the provider exercises some foresight to promote the Best Effort service for future benefits, the network will be stuck in the destabilizing cycle that leads to the demise of both the service and applications. Furthermore, our numerical experiments also revealed that the pattern in Figure 1 is not ubiquitous in all parameter regimes. When s, r, or b^{MS} is sufficiently small, $K(p^*) < 0$ in all cases, indicating the myopic profit-maximization always leads to the reduction of the Best Effort service. Understanding and interpreting the impacts of these parameter values is a subject that we are actively working on.

REFERENCES

[1] Economides, Nicholas, and Benjamin Hermalin. "The Economics of Network Neutrality," *RAND Journal of Economics*, Vol. 43, No. 4, Winter 2012, pp. 602629.
[2] Faulhaber, Gerald, and David Farber. "The Open Internet: A Customer-Centric Framework." *International Journal of Communication*, 4 (2010).
[3] Lee, Robin S., and Tim Wu. "Subsidizing creativity through network design: Zero-pricing and net neutrality." *The Journal of Economic Perspectives*, 23.3 (2009): 61-76.
[4] Mitra, Debasis, and Qiong Wang. "A model-based study of the impact of managed services and the spawning of applications in broadband networks." *Proceedings of the 24th International Teletraffic Congress*, Krakow Poland, September 2012 (http://dl.acm.org/citation.cfm?id=2414321&dl=ACM&coll=DL&CFID=273265459&CFTOKEN=84726096).
[5] Peha, Jon. "The benefits and risks of mandating network neutrality, and the quest for a balanced policy." *Proceedings of the 34th Telecommunications Policy Research Conference*, Arlington, VA, October 2006.
[6] Sen, Soumya, et al. "Pricing data: A look at past proposals, current plans, and future trends." arXiv preprint arXiv:1201.4197 (2012).