Best-Effort versus Reservations: A Simple Comparative Analysis*

Lee Breslau and Scott Shenker
{breslau,shenker}@parc.xerox.com
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

Abstract

Using a simple analytical model, this paper addresses the following question: Should the Internet retain its best-effort-only architecture, or should it adopt one that is reservation-capable? We characterize the differences between reservation-capable and best-effort-only networks in terms of application performance and total welfare. Our analysis does not yield a definitive answer to the question we pose, since it would necessarily depend on unknowable factors such as the future cost of network bandwidth and the nature of the future traffic load. However, our model does reveal some interesting phenomena. First, in some circumstances, the amount of incremental bandwidth needed to make a best-effort-only network perform as well as a reservation capable one diverges as capacity increases. Second, in some circumstances reservation-capable networks retain significant advantages over best-effort-only networks, no matter how cheap bandwidth becomes. Lastly, we find bounds on the maximum performance advantage a reservation-capable network can achieve over best-effort architectures.

1 Introduction

The current Internet offers a single class of best-effort service. That is, the Internet offers no guarantees about when (or even if) packets will be delivered, and clients need not ask permission before transmitting packets. This architecture has been tremendously successful in supporting data applications, as most recently demonstrated by the astonishing growth of Internet usage and the dramatic emergence of the World-Wide-Web. While the Internet architecture has been an undeniable success for data applications, there are many who do not think the present Internet architecture provides sufficient support for audio, video, and other so-called *real-time* applications. The Internet research community has devoted much effort to designing an integrated services Internet architecture, which is an architecture capable of supporting real-time applications as well as data applications (see, for example, [4, 6, 7, 8, 10, 15, 18] and references therein for a small sampling of the research in this area). In a culmination of these efforts, the Internet Engineering Task Force (IETF) recently promoted to Proposed Standard extensions to the Internet architecture that will enable it to support *reservations*, in which resources (*e.g.*, bandwidth) are set aside for a particular flow; see [2, 12, 13, 16, 17] for the relevant RFCs and for additional supporting material. In this architecture, clients can still send best-effort packets, but in addition clients have the option of requesting a reservation for their flow.[1] To obtain a reservation, a client requests a certain amount (characterized by a traffic specification) and quality (specified by a service specification) of service; the network then decides whether or not it can satisfy this request. While there are many mechanistic differences between the various integrated services proposals, they all share the two fundamental aspects that (1) applications have the ability to reserve bandwidth, and (2) the network exercises control – known as admission control – over these reservation requests so it can ensure the level of service given to reserved traffic. These are the two most fundamental conceptual changes brought to the Internet by an integrated services architecture.

During the past few years there has been a substantial research focus on the design details of this integrated services architecture investigating, for instance, the nature of the reservation protocol, the behavior of measurement-based admission control algorithms, and the appropriate service model. The vigor and extent of this research activity should not be interpreted as a sign of consensus about the wisdom of this endeavor. Simmering in the background has been a rather intense debate over the more fundamental question: are reservations necessary, or would the Internet be better off

*This research was supported in part by the Advanced Research Projects Agency, monitored by Fort Huachuca under contract DABT63-94-C-0073. The views expressed here do not reflect the position or policy of the U.S. government. We would like to acknowledge the crucial contributions made in the early stages of this project by Deborah Estrin and Subramaniam Vincent.

[1] A flow, for the purposes of this paper, is the traffic stream generated by a particular application.

retaining its best-effort-only architecture? Advocates of reservations claim that high fidelity interactive audio and video applications need higher quality, and more predictable, network service than that delivered by the best-effort-only Internet. Opponents of reservations, on the other hand, contend that this is a simple matter of provisioning; a reservation-capable network will not deliver satisfactory service unless its blocking rate (the rate at which it denies reservation requests) is low, and at such provisioning levels best-effort networks will provide completely adequate service – service that is nearly as good as that of the reservation-capable network. Moreover, opponents maintain, any differential in quality can be offset by adding a modest amount of additional capacity to the best-effort-only network which, when bandwidth becomes inexpensive, should be far cheaper than the added complexity of the proposed integrated services architectural extension. In addition, the most ardent opponents claim that the adaptability of modern network applications renders reservations unnecessary, since applications can adapt to whatever service the network offers.[2]

This research was initiated in an attempt to formalize some of these claims and provide a more solid footing for this debate. This paper introduces a simple analytical model in which we can more concretely pose the question of whether the Internet should adopt a reservation-capable architecture or retain its best-effort-only architecture. This model is not intended to be a complete representation of reality, but is instead intended merely to illustrate, in an accessible and tractable fashion, some of the essential issues. While we hope to *inform* the debate, by providing an intellectual framework in which the debate can constructively continue, we do not in any way expect that this work will *settle* the debate, since the relative merits of the two architectures depend on many practical concerns – such as the future cost of bandwidth and the burden of architectural complexity – that are inputs to, not outputs of, our model. In the spirit of full disclosure, we admit that we (the authors) are biased in favor of reservations; we strived to keep our analysis as neutral as possible, but we obviously aren't the best judges of our success in that regard.

To evaluate a network architecture, one must ask how well it meets user needs. For a user employing a given network application, the *utility* – or value – the user derives from that application will depend on the application's performance (*e.g.*, the picture quality for video, the sound quality for audio applications, etc.); the application's performance, in turn, depends on the nature of the network service the application receives. Network architectures are intended to provide a high degree of total utility (the sum of utility over all network users). For the simple case where a single link has a fixed load of k identical applications, it has been shown in [14] that for certain classes of utility functions the reservation-capable architecture provides a higher level of total utility. We review that derivation here, in Section 2, since our subsequent work will build on these results.

[2] These statements are simplifications of the actual debate. We include them only to give a flavor of the contending arguments.

However, optimizing utility is not the only design goal; competing with it is the goal of keeping the network architecture simple. The complexity of the network architecture is hard to quantify, and we do not attempt to do so here, but it is clear that any integrated services architecture is significantly more complex than a best-effort-only architecture. The key question, then, is whether or not the performance advantage of reservation-capable networks alluded to above is significant. If this utility difference is quite small then there is little reason to incorporate reservations into the Internet architecture, since the burden of adding significant additional complexity to the Internet architecture would far outweigh the small increase in utility it would bring. We initially address this question in Sections 3 and 4 through the use of a variable load model that extends the fixed load model used in [14]. Our results for this variable load model show that the answer to the question we posed depends critically on characteristics of the network load and on the nature of application utility. We find that, under many conditions in our variable load model, the arguments of the opponents of reservations are largely correct. However, we also find that, under certain conditions, the incremental bandwidth needed to equalize performance in the two kinds of networks increases as the capacity increases. We also show that in a subset of these cases in our variable load model, reservation-capable networks retain a significant, but bounded, performance advantage over best-effort networks, no matter how cheap bandwidth becomes. In Section 5 we consider two extensions to the variable load model that increase the performance advantage of reservation-capable networks. We conclude in Section 6 with a brief discussion of our results.

2 Fixed Load Model

The key difference between a best-effort-only architecture and a reservation-capable one is that, in the former, flows are never denied access to the network – they can send packets whenever they want – whereas in a reservation-capable architecture the network can deny reservation requests. The question we address here, reviewing the material in [14] to provide necessary context, is: does denying reservation requests result in an increase in utility under some conditions, or does allowing all flows access to the network always maximize utility?

We do this by considering a very simple fixed load model. We consider a single link of capacity C and we assume that the load on this link is, at any one time, comprised of k identical flows. We further assume that the bandwidth is allocated evenly among these k flows, so that each flow receives the same bandwidth share $\frac{C}{k}$.

Each flow represents an application whose performance or utility π as a function of the bandwidth b allotted it is given by the function $\pi(b)$. We assume that $\pi(0) = 0$ (*i.e.*, when the application receives zero bandwidth it provides no value) and that $\pi(\infty) = 1$ (*i.e.*, when the application receives as much bandwidth as it wants, it provides a value of 1). At bandwidths $0 < b < 1$, different applications have different levels of performance, but in all cases $\pi(b)$ is a nondecreasing function. We model a flow that requested a reservation but was denied service as receiving zero bandwidth ($b = 0$) and so has zero utility ($\pi = 0$).

If all flows are given service, the total utility of the network is given by $V(k) \equiv k\pi(\frac{C}{k})$. If the function $V(k)$ is increasing then utility is maximized by always allowing all flows access to the network. In this case, the best-effort-only architecture, which admits all flows, provides the higher total utility.

However, if the function $V(k)$ is maximized at some finite value k_{max}, then if $k > k_{max}$ the utility would be maximized by denying service to the additional flows $k_{max}+1, k_{max}+2, \ldots, k$. Such denial of service requires an integrated services architecture.

Thus, the nature of the function $V(k)$ determines which architecture produces the higher utility. In turn, the nature of the function $V(k)$ depends on the character of the utility function $\pi(b)$. There are two general results of interest. First, if there exists some neighborhood of the origin in which the function π is convex but not concave (*i.e.*, is convex, but not linear in the whole region), then there exists some k_{max} such that $V(k_{max}) > V(k)$ for all $k > k_{max}$. For such functions π, admission control should keep the number of users at or below k_{max}. Second, if the function π is everywhere strictly concave, then $V(k)$ is a strictly monotonically increasing function of k. In this case, access should never be denied and so admission control is not needed. The next question, then, is: what do real application utility functions look like?

The traditional data applications, like electronic mail and file transfer, are somewhat *elastic*, in that they are not particularly sensitive to individual packet delays, and typically do not have hard real-time constraints. This suggests that while giving such applications additional bandwidth certainly aids performance, the marginal improvement for additional bandwidth decreases in b and so $\pi(b)$ is strictly concave everywhere. Thus, $V(k)$ is always maximized when no users are denied access; the current best-effort-only architecture is ideal for such elastic applications.

At the opposite end of the spectrum are *rigid* applications, which need their data to arrive within a given delay bound (and performance does not improve if data arrives earlier than this bound). Traditional telephony is an example of such a rigid application, as are other applications that rely on circuit-switched service. For such applications needing $\bar{b}$ units of bandwidth

$$\pi(b) = 0 \; for \; all \; b < \bar{b} \;\; and \;\; \pi(b) = 1 \; for \; all \; b \geq \bar{b} \quad (1)$$

and so the function $V(k)$ is given by

$$V(k) = 0 \quad k > \frac{C}{\bar{b}} \qquad and \qquad V(k) = k \quad k \leq \frac{C}{\bar{b}}$$

and thus admission control is clearly necessary here to constrain usage at or under $k_{max}(C) = \lfloor \frac{C}{\bar{b}} \rfloor$. Applications whose utility curves give rise to finite $k_{max}(C)$ are deemed to be *inelastic*, and function better with a reservation-capable architecture.

These two extreme cases – elastic and rigid applications – illustrate the fact that the telephone and Internet network architectures were both designed to meet the needs of their original class of applications; rigid applications perform better with reservations, and data application fare better without them.

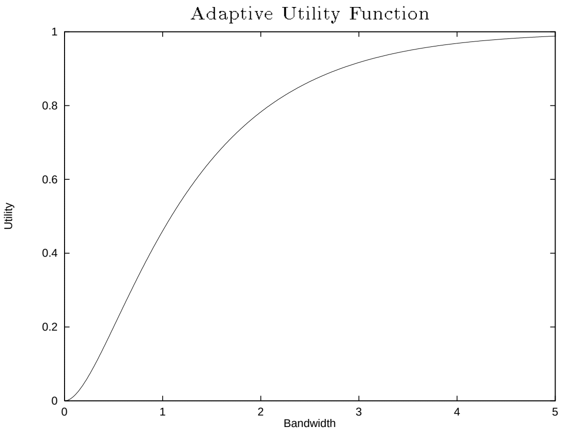

Figure 1: The performance curve $\pi(b)$ for a rate and delay adaptive application.

However, video and voice applications are becoming much more common on the Internet, and these Internet voice/video applications are not built to expect circuit-switched service. Instead, they are designed to adapt to the currently available bandwidth and to variations in packet delay.[3] It appears, due to human perceptual factors, that minimal levels of bandwidth are not very useful, so that at low bandwidths the marginal utility of additional bandwidth is fairly small. Similarly, at high bandwidths the signal quality is quite good and so the marginal utility of additional bandwidth at high bandwidths is also small. At intermediate levels, when the signal quality first starts to be viable, the marginal utility of extra bandwidth is significant.

One such utility function modeling adaptive but inelastic applications, that we will use in our later analysis, is given by:

$$\pi(b) = 1 - e^{-\frac{b^2}{\kappa + b}} \qquad (2)$$

where $\kappa = .62086$.[4] This function is depicted in Figure 1; note that for small b, $\pi(b) \approx \frac{b^2}{\kappa}$ and that for large b $\pi(b) \approx 1 - e^{-b}$. While the function $V(k)$ for this utility function has a peak at some finite $k_{max}(C)$ (because of the convex neighborhood around the origin), the adaptive nature of the application means that the decrease in $V(k)$ for $k > k_{max}(C)$ is much more gentle than the abrupt drop from full utility $V(k) = k$ to zero utility $V(k) = 0$ that rigid applications have as k passes through $k_{max}(C)$. Thus, while it is true that an integrated services architecture produces superior performance for these adaptive audio and video applications, it is not at all clear that the performance

[3]Most current Internet audio and video applications are delay adaptive but not rate-adaptive, in that they do not adjust their sending rate, but do adjust to varying packet delays. See [14] for an elucidation of the differing utility curves between these two styles of adaptation. For our treatment here, we assume that the applications are both rate and delay adaptive. This assumption makes the case for best-effort-only service stronger, by considering only those applications most suited to best-effort.

[4]This value of κ yields $k_{max}(C) = C$, facilitating comparisons with the rigid case, which also has $k_{max}(C) = C$.

advantage is of a significant magnitude. We address this issue in the next section.

3 Variable Load Model

The model from [14] reviewed in the previous section used a fixed load k of flows. If the applications are inelastic, the total utility is higher in a reservation-capable network when the offered load k is higher than $k_{max}(C)$. This does not tell us the likelihood of such overload conditions, and thus we cannot evaluate the extent of the performance advantage of the reservation-capable architecture. To quantify this performance advantage, we extend the fixed load model from [14] to include variable loads. The load on the network is described not by a fixed number of inelastic flows, but by a probability distribution of the number of inelastic flows on the link. To keep the level of complexity manageable, we do not model the dynamics of flows arriving and departing the network, but rather only consider a probability distribution of possible static loads.

In what follows we use a mixture of numerical computations on a more realistic discrete model and analytical calculations on a more tractable continuum model. The two models are quite similar in spirit, if different in detail (the probability distributions are slightly different and the adaptive utility functions are substantially different). The results of the two models are, at least in the asymptotic case of large C, completely equivalent. We first present the discrete version of our variable load model.

3.1 Discrete Model

Let $P(k)$ denote the probability that there are k flows requesting service, and let $\bar{k}$ denote the average number of flows requesting service: $\bar{k} = \sum_{k=0}^{\infty} P(k)k$. We assume that a user's utility is the average of her utility at these various load levels. In the best-effort-only architecture, each flow receives bandwidth $\frac{C}{k}$ so the total utility of the system, $\bar{V}_B(C)$, is given by:

$$\bar{V}_B(C) = \sum_{k=1}^{\infty} P(k)k\pi(\frac{C}{k})$$

We will use the notation $B(C)$ to refer to the normalized utility:

$$B(C) = \frac{\bar{V}_B(C)}{\bar{k}}$$

In the reservation-capable architecture, when k flows request service each of the $\min[k, k_{max}(C)]$ admitted flow receives bandwidth $\frac{C}{\min[k,k_{max}(C)]}$, and each of the $k - \min[k, k_{max}(C)]$ rejected flows gets zero bandwidth. The total utility of the system, $\bar{V}_R(C)$, is given by:

$$\bar{V}_R(C) = \sum_{k=1}^{k_{max}(C)} P(k)k\pi(\frac{C}{k})$$
$$+ \sum_{k=k_{max}(C)+1}^{\infty} P(k)k_{max}(C)\pi(\frac{C}{k_{max}(C)})$$

We use the notation $R(C)$ to refer to the normalized utility:

$$R(C) = \frac{\bar{V}_R(C)}{\bar{k}}$$

Clearly we have $R(C) \geq B(C)$, with the inequality strict if $k_{max}(C)\pi(\frac{C}{k_{max}(C)}) > (k_{max}(C)+1)\pi(\frac{C}{k_{max}(C)+1})$ and $P(k_{max}(C)+1) > 0$; these conditions always hold in the cases we consider, and so $R(C) > B(C)$ in what follows. The key question, though, is whether this difference is significant.

One way to answer that question would be to compare the numerical values of these two quantities, looking at the *performance gap*:

$$\delta(C) = R(C) - B(C)$$

Since the units of utility are somewhat arbitrary, this approach may be of limited value. Perhaps a better way of assessing significance is to determine how much additional bandwidth is needed to make a best-effort-only network have the same performance as the reservation-capable one. This is an important quantity given that arguments against the need for reservations suggest that the same performance can be achieved by adding a modest amount of additional capacity to a best-effort-only network. We can define the incremental bandwidth requirements $\Delta(C)$, called the *bandwidth gap*, via the relation:

$$R(C) = B(C + \Delta(C))$$

The reservation-capable architecture imposes the burden of additional complexity. The best-effort-only architecture allows one to avoid extra complexity, but it requires additional bandwidth in order to match the performance of the reservation-capable architecture. The bandwidth gap $\Delta(C)$ quantifies this bandwidth versus complexity tradeoff, and depends on the functions $P(k)$ and $\pi(b)$.

In modeling $\pi(b)$, we consider two separate cases, representing the two classes of inelastic applications discussed above: rigid and adaptive. Rigid applications have utility functions given by Equation 1 (with $\bar{b} = 1$) and the adaptive applications have utility functions as given by Equation 2. Our choice of the particular form of the adaptive utility function is arbitrary, and represents an extremely adaptive function in that it provides non-negligible marginal utility over a wide range of bandwidths.

In modeling $P(k)$, we claim no special wisdom about the nature of future network loads. To cover a broad spectrum of possibilities, we consider three quite different load distributions $P(k)$:

Poisson $P(k) = \frac{\nu^k e^{-\nu}}{k!}$

Exponential $P(k) = (1 - e^{-\beta})e^{-\beta k}$

Algebraic $P(k) = \frac{\nu}{\lambda + k^z}$

Note that $\bar{k} = \nu$ in the Poisson distribution and $\bar{k} = (e^{\beta} - 1)^{-1}$ in the exponential distribution. The power z in the algebraic distribution controls the asymptotic rate of decrease in $P(k)$; we only consider cases where $z > 2$ so that $\bar{k}$ is well defined. The constants ν and

λ are chosen so that the probability function is normalized, $1 = \sum_{k=0}^{\infty} P(k)$. We introduce two parameters (rather than just setting λ to zero and taking a simple power law) so that we can vary the average of the distribution while holding the asymptotic power law z fixed. In all of our numerical calculations, we set $\bar{k} = 100$. Note again that we do not justify these load distributions by any detailed arrival and departure processes. There are too many unknown aspects of what these might be in reality, especially in terms of correlations and diurnal rhythms, and so instead we just model their resulting stationary distributions.

We now have six cases to investigate: two choices for the utility function $\pi(b)$, combined with three choices for the load distribution $P(k)$. Since solving the various quantities of interest analytically is difficult in this discrete variable load model, we instead numerically evaluate these quantities. This allows us to do our modeling without regard for tractability. However, since our numerical calculations are necessarily done over a finite range of C values, it is impossible to make definitive conclusions about the asymptotic (large C) behavior of the various quantities based on these computations; for that we need analytical calculations, to which we now turn.[5]

3.2 Continuum Model

We augment our treatment of the variable load model by introducing a continuum version, where the variable k varies continuously from 0 to ∞. This model, while requiring some additional simplifications, is more analytically tractable than the discrete model. However, these simplifications do not affect the asymptotic behavior of the quantities we examine. In the continuum model, the formulae for $\bar{V}_R(C)$ and $\bar{V}_B(C)$ become:

$$\bar{V}_R(C) = \int_0^{k_{max}(C)} dk P(k) k \pi(\frac{C}{k})$$

$$+ \int_{k_{max}(C)}^{\infty} dk P(k) k_{max}(C) \pi(\frac{C}{k_{max}(C)})$$

and

$$\bar{V}_B(C) = \int_0^{\infty} dk P(k) k \pi(\frac{C}{k})$$

In our continuum model, we only consider the exponential and algebraic load distributions, as they are most easily computable.[6] In addition, to make the algebraic distribution more tractable, we consider the form $P(k) = (z-1)k^{-z}$ for $k \geq 1$ and $P(k) = 0$ for $k < 1$. Moreover, since the calculations are no longer tractable when we consider adaptive applications with utility functions given by Equation 2, we use a modified form of adaptive utility function in the continuum model. These utility function is parametrized by a constant a, $a \in (0, 1)$, and is given by the following:

$$\pi(b) = 0 \quad b < a$$

$$\pi(b) = \frac{b-a}{1-a} \quad a \leq b < 1$$

$$\pi(b) = 1 \quad b \geq 1$$

Note that when $a = 1$ this reduces to the rigid case. Decreasing a represents increasing levels of adaptivity of the application. For all $a > 0$, $k_{max}(C) = C$, so the calculations of $\bar{V}_R(C)$ are identical to the rigid case, but the best-effort results are significantly altered. When $a = 0$, the utility function is no longer inelastic and $\bar{V}_R(C) = \bar{V}_B(C)$.

3.3 Results

We now address the six cases. For each of the three different load distributions, we first consider rigid applications, and then adaptive ones. In each case, we begin with the relevant results of the numerical calculations using the discrete model and then, where appropriate, augment the discussion with results from the continuum model.

The Poisson load distribution describes a situation where the load is fairly tightly controlled within a region around the average, and excursions to large (or small) loads are extremely rare. This describes the behavior of a Poisson arrival process with uncorrelated and independent departure processes. The results here are the closest (of our three load distributions) to the fixed load model. Figure 2a shows the performance functions for reservations and best-effort, $B(C)$ and $R(C)$, and Figure 2b shows $\Delta(C)$ for Poisson load and rigid applications. Note that for small C (by which we mean $C < \bar{k}$) $R(C)$ is close to linear in C (with slope $\frac{1}{\bar{k}}$, and recall that $\bar{k} = 100$) while $B(C)$ is almost zero throughout most of this region. The difference in performance $\delta(C) = R(C) - B(C)$ reaches a peak of 0.8 and the bandwidth gap $\Delta(C)$ reaches a peak of 80, and so both gaps are significant in this region. However, as soon as C is slightly greater than $\bar{k}$, both $R(C)$ and $B(C)$ are very close to unity, and even closer together, and so both $\delta(C)$ and $\Delta(C)$ vanish extremely quickly (faster than exponentially).

Figures 2d and 2e show the functions $B(C)$, $R(C)$, and $\Delta(C)$ for Poisson load and adaptive applications. In contrast to the rigid case, the $R(C)$ and $B(C)$ curves are quite close for all but the smallest C; this reflects the fact that adaptive applications tolerate overload conditions reasonably well, and so the performance under best effort does not degrade so severely. Note that the $R(C)$ curve continues to increase well past $C = \bar{k}$. This is because the utility $\pi(b)$ continues to increase for $b > \frac{C}{k_{max}(C)}$, reflecting that adaptive applications not only tolerate overload conditions better, they also take advantage of underload conditions more effectively than rigid applications. As for the performance and bandwidth gaps, the maximum values of $\delta(C)$ and $\Delta(C)$ are substantially lower than in the rigid case. As before, in the region $C > \bar{k}$ both of these difference curves vanish superexponentially.

The exponential load distribution describes a situation where the load is not peaked around the average, but instead decays over the whole range at an exponential rate. Figures 3a and 3b show the functions $B(C)$, $R(C)$, and $\Delta(C)$ for exponential load and rigid applications. The performance curves, $R(C)$ and $B(C)$,

[5] Conversely, the analytical calculations are often *only* tractable in the asymptotic (large C) limit, so we need the numerical calculations to illustrate the nonasymptotic regime.

[6] The asymptotic behavior of our numerical computations for Poisson distribution are completely unambiguous. Therefore, failure to treat this case analytically does not impact our conclusions.

increase more gradually than in the Poisson case, reflecting the greater variability in load levels. Additional differences from the Poisson case are that $B(C)$ is not vanishingly small when $C < \bar{k}$ nor is $R(C)$ linear, and so the performance gap $\delta(C)$ peaks at less than half the value of the Poisson peak. However, in the region $C > \bar{k}$, $\delta(C)$ decays more slowly in the exponential case. At capacities of $2\bar{k}$ and $4\bar{k}$ with rigid applications, $\delta(C)$ is approximately .27 and .07, respectively. For the Poisson distribution, $\delta(C)$ is less than 10^{-15} at the same capacities. The biggest contrast with the Poisson case, though, is that the bandwidth gap $\Delta(C)$ is monotonically increasing throughout the entire domain. As capacity increases, the incremental bandwidth needed to make best-effort performance equivalent to reservations increases.

We can articulate this behavior more precisely in the continuum model. For rigid applications, the continuum equations become (recalling that $k_{max}(C) = C$):

$$\bar{V}_R(C) = \int_0^C dk P(k)k + C\int_C^\infty dk P(k)$$

and

$$\bar{V}_B(C) = \int_0^C dk P(k)k$$

For the exponential distribution $P(k) = \beta e^{-\beta k}$, we find that $\bar{V}_B(C) = \frac{1}{\beta}(1 - e^{-\beta C}(1+\beta C))$ and $\bar{V}_R(C) = \frac{1}{\beta}(1 - e^{-\beta C})$. In addition, $\delta(C) = Ce^{-\beta C}$ and $\Delta(C)$ is the solution to $\beta\Delta(C) = \ln(1 + \beta(C + \Delta(C)))$ which grows asymptotically like $\frac{\ln C}{\beta}$ for large C. Thus, the bandwidth gap grows logarithmically for exponential loads and rigid applications. This is a somewhat surprising result, in that it says that as you increase the overprovisioning in the limit $C \gg \bar{k}$ it takes increasingly more bandwidth to render the performance of the two architectures the same. At first it might seem puzzling that the performance gap $\delta(C)$ is decreasing while the bandwidth gap $\Delta(C)$ is increasing; the phenomena is most easily understood by noting that $\Delta(C)$ can be approximated by $\frac{\delta(C)}{B'C}$ so if the derivative $B'(C)$ is decreasing faster than the gap $\delta(C)$ then $\Delta(C)$ is an increasing function. Thus, even though the performance gap is shrinking, the increase in utility per unit of bandwidth is shrinking faster, so it takes increasingly more bandwidth to make up this performance difference.

This behavior disappears in the adaptive case. Figures 3d and 3e show the functions $B(C)$, $R(C)$, and $\Delta(C)$ for exponential load and adaptive applications. The $R(C)$ and $B(C)$ curves are much closer together. The peak of the performance gap $\delta(C)$ is reduced by a factor of 10 and it has a value less than .01 when capacity equals $2\bar{k}$, and less than .001 when capacity equals $4\bar{k}$. Note that after hitting a peak of 9, the bandwidth gap $\Delta(C)$ decreases for $C > \bar{k}$.

In the continuum model, (recalling that a is a parameter in the continuum adaptive utility function) we find $\bar{V}_B(C) = \frac{1}{\beta}(1 - \frac{e^{-\beta C}}{1-a} + \frac{ae^{-\frac{\beta C}{a}}}{1-a})$. In this case, $\delta(C) = \frac{ae^{-\beta C}}{1-a}(1 - e^{-\frac{a\beta C}{1-a}})$. $\Delta(C)$ is the solution to $\beta\Delta(C) = -\ln(1-a) + \ln(1 - ae^{-\frac{a\beta(C+\Delta(C))}{1-a}})$. For large C, $\Delta \approx \frac{-\ln(1-a)}{\beta}$, so the bandwidth performance gap goes to a constant (as opposed to the logarithmic growth we found in the rigid case). Thus, the exponential distribution illustrates the profound difference between rigid and adaptive applications, and the qualitative, not just quantitative, impact adaptivity has on the tradeoffs between reservation-capable and best-effort-only architectures.

The algebraic distribution is like the exponential distribution in that it decreases over the whole range, but here the decrease is much slower. Figures 4a and 4b show the functions $B(C)$, $R(C)$, and $\Delta(C)$ for algebraic load and rigid applications, with $z = 3.0$ (recall that z is the power of the algebraic distribution). The gap between the $R(C)$ and $B(C)$ remains substantial over a wide range of C's (for instance taking on the values of .20 at $C = 2\bar{k}$ and .10 at $C = 4\bar{k}$), and so, while the performance gap $\delta(C)$ peaks at a fairly low value it decays quite slowly. In contrast, the bandwidth gap $\Delta(C)$ increases linearly throughout the entire domain.

Figures 4d and 4e show the functions $B(C)$, $R(C)$, and $\Delta(C)$ for algebraic load and adaptive applications. The performance gap between the $R(C)$ and $B(C)$ is much less than with rigid applications, but the increasing nature of $\Delta(C)$ remains unchanged, although the slope is much less (decreased by a factor of over 20).

The continuum calculations shed some light on the behavior of $\Delta(C)$. Recall that the algebraic distribution for the continuum model is given by $P(k) = (z-1)k^{-z}$ for $k \geq 1$ and $P(k) = 0$ for $k < 1$. For rigid applications, $\bar{V}_B(C) = \frac{z-1}{z-2}(1 - C^{2-z})$ and $\bar{V}_R(C) = \frac{z-1}{z-2}(1 - \frac{C^{2-z}}{z-1})$. The gaps are given by $\delta(C) = C^{2-z}$ and $\Delta(C) = C((z-1)^{\frac{1}{z-2}} - 1)$. Thus, the linear increase in $\Delta(C)$ applies for all powers z. For $z = 3.0$, the constant of proportionality is 1, similar to what we saw in our numerical calculations in the discrete case.

Note that in the limit as $z \rightarrow 2^+$, $\Delta(C) = (e-1)C$. We conjecture that this limit represents the greatest asymptotic advantage reservations can have over best-effort-only in our basic variable load model.[7] If so, this means that in the worst asymptotic case in this variable load model, best-effort-only networks require e times more bandwidth than reservation networks to match their performance. Hence, while a reservation-capable network always has some performance advantage over a best-effort one, the ratio of additional bandwidth needed to make up this difference is bounded.

In the adaptive case, $\bar{V}_B(C) = \frac{z-1}{z-2}(1 - C^{2-z}\frac{1-a^{z-1}}{(1-a)(z-1)})$. $\delta(C) = \frac{C^{2-z}}{z-2}\frac{a(1-a^{z-2})}{1-a}$ and $\Delta(C) = C((\frac{1-a^{z-1}}{1-a})^{\frac{-1}{z-2}} - 1)$. Note that in the limit as $z \rightarrow 2^+$, $\Delta(C) = Ce^{\frac{-a\ln a}{1-a}}$. In this limit, the constant of proportionality can vary from 1 (for $a \rightarrow 0^+$) to e (for $a \rightarrow 1^-$), depending on the nature of the adaptive utility function.

Notice that while the asymptotic behaviors of the discrete and continuum models agree for the algebraic distribution, they disagree for smaller C. In particular, while the continuum model has $\frac{\Delta(C)}{C}$ constant for

[7] For this asymptotic advantage we care only about the nature of $\frac{\Delta(C)}{C}$ for large C; there may be cases where the ratio $\frac{\Delta(C)}{C} > (e-1)$ for smaller C.

all C, the discrete model has some some structure in the $\Delta(C)$ curve at lower values. This is due to the difference in the algebraic distribution. To be able to vary the average $\bar{k}$ without changing the power law z, we inserted a constant in the discrete version that perturbed the distribution for lower values of C: *i.e.*, k^{-z} versus $\frac{1}{\lambda+k^z}$. We think the latter is likely to be a more realistic distribution than the continuum one, but that is pure speculation and our point here is merely to explain the impact on the results for $C \leq \bar{k}$. In any event, the asymptotic behavior is unaffected.

The adaptive utility function used in our numerical computations behaves as $\pi(b) \approx 1-e^{-b}$ for large b (see Equation 2). While we think this exponential approach to the asymptotic value is the most realistic choice, one can also consider utility functions that approach their asymptotic values more slowly. For instance, the family of functions $\pi(b) = \frac{b^r}{1+b^r}$ approach algebraically: $\pi(b) \approx 1 - b^{-r}$ for large b. This difference turns out to be important for the algebraic load distributions. To focus on the important aspect (the large b behavior) and to make the calculation tractable,[8] we use instead the form:

$$\pi(b) = 0 \quad b \leq 1$$

$$\pi(b) = 1 - b^{-r} \quad b > 1$$

This captures the behavior at high b but ignores the behavior at low b. For this form of $\pi(b)$, we find $k_{max}(C) = C(r+1)^{\frac{-1}{r}}$. For algebraic loads, the total utilities take the form:

$$\bar{V}_B(C) = \omega_1 + \omega_2 C^{-r} + \omega_3 C^{2-z}$$

and

$$\bar{V}_R(C) = \omega_1 + \omega_2 C^{-r} + \omega_4 C^{2-z}$$

with the ω_i being constants and $\omega_4 > \omega_3$. Note that the asymptotic behavior of $\Delta(C)$ for large C depends on whether $r > z-2$ or not. If $r > z-2$ then $\Delta(C) \sim C$ for large C, but if $r < z-2$ then $\Delta(C) \sim C^{r+3-z}$; thus, if $z-2 > r > z-3$ then $\Delta(C)$ asymptotically increases with C, but not linearly, and if $r < z-3$ then $\Delta(C)$ asymptotically decreases with C. We have observed similar behavior in our calculations.

While the variable load model introduced in this section illustrates the performance and bandwidth gaps between the two architectures as a function of C, it yields no insight as to what value of C is likely to be relevant. This is crucial, since the behavior for $C \approx \bar{k}$ can be radically different than the behavior for $C \gg \bar{k}$. In the next section, we try to gain some insight into the choice of C.

4 Variable Capacity Model

What capacity level C is likely to be present in the network? This is clearly an impossible question to answer in general, since so much will depend on market factors like the future cost of bandwidth and the level of network usage. However, in an attempt to clarify the situation, we present a very simplified analysis of the economic tradeoff between the cost of additional bandwidth and the utility it provides.

We assume that a network service provider making the provisioning decision can provide additional bandwidth at a cost p per unit bandwidth. Moreover, we assume that the service provider sets the provisioning level so as to maximize the total welfare $\bar{V}(C) - pC$. This is based on the assumption that the provider can recover the utility $\bar{V}(C)$ from charging customers, and so the quantity $\bar{V}(C) - pC$ represents the profits of the network provider. Maximizing this welfare gives rise to a function $C(p)$ describing the capacity as a function of price. Then the welfare provided by the network can be computed via:

$$W(p) = \bar{V}(C(p)) - pC(p)$$

This is the total utility derived from the network minus its cost. We will denote the quantities for the two architectures as $C_R(p)$ and $C_B(p)$, and similarly $W_R(p)$ and $W_B(p)$. With this model we now compare the quantities $W_R(p)$ and $W_B(p)$, rather than comparing the quantities $\bar{V}_R(C)$ and $\bar{V}_B(C)$ as we did in the previous section. Before we compared utilities at a given capacity level, we now compare welfare values at a given price for bandwidth. The comparison of welfares $W_R(p)$ and $W_B(p)$ recognizes the fact that one might make capacity decisions based in part on the choice of architecture.

We must always have $W_R(p) \geq W_B(p)$ (with a strict inequality holding as long as $W_R(p) > 0$ in all of the cases we consider). The welfare difference $W_R(p) - W_B(p)$ must be compared to the additional complexity needed. However, as before, comparing absolute (or relative) welfare values may not be very informative. A better measure is to ask, given a bandwidth price p, at what bandwidth price $\tilde{p}$ are the two welfares equal: $W_R(\tilde{p}) = W_B(p)$. Thus, the ratio $\gamma(p) = \frac{\tilde{p}}{p}$ indicates how much more expensive bandwidth in the integrated services architecture would have to be (assuming that its cost is linear in bandwidth) in order for the best-effort-only network to be the more cost-effective one. Thus, if we quantify the cost of additional complexity as how much extra per-unit-bandwidth it takes to build such a network (which is probably not a good approximation of reality, but it may be sufficient to illustrate our point), we can then compare the additional utility provided by reservations to their additional cost of complexity. Figures 2c, 2f, 3c, 3f, 4c and 4f display the equalizing price ratio $\gamma(p)$ for our six cases.

For the Poisson load distribution with rigid applications, the provisioning levels (not shown) remain quite moderate (below $1.4\bar{k}$) for all but the very smallest pricing levels. The price ratio that makes two architectures equivalent varies, for most values of p, between 1.1 and 1.2 (see Figure 2c). Thus, if adding reservations added less than 10% to the cost of bandwidth, then over most of the domain of prices the reservation-capable network is the preferable choice. If bandwidth is exceedingly cheap, then this no longer holds.

When we switch to adaptive applications with the Poisson load distribution, the capacity levels are sig-

[8] We also investigated the form:

$$\pi(b) = b^r \quad b \leq 1$$

$$\pi(b) = 1 \quad b > 1$$

which captures the low b behavior but not the high b behavior and obtained asymptotic results that resembled those of the original utility function.

nificantly higher, reflecting the fact that adaptive applications can take better advantage of underloaded situations. The capacity levels $C_R(p)$ and $C_B(p)$, and the welfare levels $W_R(p)$ and $W_B(p)$, are nearly the same at all price levels. As shown in Figure 2f, the equalizing price ratio $\gamma(p)$ is effectively 1 for all but the higher values of p, so that if adding reservation capability to the network incurred any significant per-unit-bandwidth cost then the best-effort-only architecture is the preferable choice.

The results for the exponential load distribution are fairly similar to those for the Poisson distribution. However, we can treat the exponential case analytically. With rigid applications, the overall welfare is maximized for best-effort-only when $p = \beta C e^{-\beta C}$ and so $W_B(p) = \frac{1}{\beta}(1 - p - \frac{p}{h(p)} - ph(p))$ where the function h is defined implicitly as the largest solution to $p = h(p)e^{-h(p)}$. For the reservation case, the maximizing capacity is $C = \frac{-\ln p}{\beta}$ and so $W_R(p) = \frac{1}{\beta}(1 - p + p\ln p)$. The ratio of prices $\gamma(p)$ that gives equal welfares – $W_B(p) = W_R(\gamma(p)p)$ – is given by the solution to the equation $\gamma(p)(1 - \ln\gamma(p) - \ln p) = 1 + \frac{1}{h(p)} + h(p)$. Note that when $p \rightarrow 0^+$ this ratio is converges to one as $\gamma(p) \approx 1 + \frac{\ln(-\ln p)}{-\ln p}$.

With adaptive applications, the overall welfare is maximized for best-effort-only when $p = \frac{1}{1-a}(e^{-\beta C} - e^{-\frac{\beta C}{a}})$. For small p and $a < 1$ the first term dominates, so we have, approximately, $p \approx \frac{1}{1-a}e^{-\beta C}$ and so $W_B(p) \approx \frac{1}{\beta}(1 - p + a(1-a)^{\frac{1}{a}-1}p^{\frac{1}{a}} + p\ln(p(1-a)))$. Recall that for the reservation case, $W_R(p) = \frac{1}{\beta}(1 - p(1 - \ln p))$. The ratio of prices $\gamma(p)$ that gives equal welfares for $a < 1$ is given approximately by $\gamma(p) = 1 + \frac{\ln(1-a)}{\ln p}$ which approaches 1 logarithmically. In going from rigid to adaptive in this case, the rate at which $\gamma(p)$ converges to 1 differed by a factor of $\ln(-\ln p)$.

The key feature of the exponential and Poisson distributions is that the equalizing ratio $\gamma(p)$ converges to one as the price of bandwidth approaches zero. This implies that as bandwidth becomes cheaper, a best-effort-only network is preferable if the complexity of reservations imposes any significant cost on building or managing a network. In contrast, in the algebraic case $\gamma(p)$ does not converge to one. This is shown in Figures 4c and 4f and confirmed by the analysis. That is, if a reservation capable network only imposes a small but nonvanishing additional per-unit bandwidth, then no matter how inexpensive bandwidth becomes the reservation-capable architecture is the preferable choice.

For rigid applications, the overall welfare is maximized for best-effort-only when $C = (\frac{p}{z-1})^{\frac{1}{1-z}}$ and so $W_B(p) = \frac{z-1}{z-2}(1 - (z-1)^{\frac{1}{z-1}}p^{\frac{z-2}{z-1}})$. For the reservation case, the maximizing capacity is $C = p^{\frac{1}{1-z}}$ and so $W_R(p) = \frac{z-1}{z-2}(1 - p^{\frac{z-2}{z-1}})$. The ratio of prices $\gamma(p)$ that gives equal welfares is $\gamma(p) = (z-1)^{\frac{1}{z-2}}$. This is consistent with our numerical computations, where the $\gamma(p)$ takes on values approaching 2 when p approaches zero (recall that $z = 3$). Note that when $z \rightarrow 2^+$ this ratio is $\gamma(p) = e$. As before, we conjecture that the limit $z \rightarrow 2^+$ represents that greatest asymptotic advantage reservations can have over best-effort-only. Thus, we conjecture that the asymptotic ratios $\lim_{p\rightarrow 0^+}\gamma(p) = e$ and $\lim_{C\rightarrow\infty}\frac{\Delta(C)}{C} = (e-1)$ are the maximal cases. In the worst asymptotic case, we conjecture, best-effort-only networks require e times more bandwidth than reservation networks, and if the price of constructing reservation-capable networks is more than e times more expensive (than constructing best-effort-only networks) then best-effort-only networks are always more advantageous (no matter what the load distribution).

For adaptive applications, the overall welfare is maximized for best-effort-only when $C = (p\frac{1-a}{1-a^{z-1}})^{\frac{-1}{z-1}}$ and so $W_B(p) = \frac{z-1}{z-2}(1 - p^{\frac{z-2}{z-1}}(\frac{1-a}{1-a^{z-1}})^{\frac{-1}{z-1}})$. The ratio of prices $\gamma(p)$ that gives equal welfares is $\gamma(p) = (\frac{1-a^{z-1}}{1-a})^{\frac{1}{z-2}}$. Note that when $z \rightarrow 2^+$ this ratio becomes $e^{\frac{-a\ln a}{1-a}}$. This ratio varies from 1 (for $a = 0$) to e (for $a = 1$). In this case, direct comparisons with our numerical calculations are not possible because we use different adaptive utility functions in the discrete and continuum models. In the discrete case, $\gamma(p)$ is approximately 1.02 as p approaches zero.

5 Extensions to the Model

We have considered several extensions to our model that capture potentially relevant elements not included in the basic model. Many of these extensions – such as having heterogeneous flows (both in size and in utility), risk-averse utility functions (where the utility is not the average performance experienced, but something less), and nonstationary loads (where the probability distribution is not fixed) – did not change the basic nature of our asymptotic (large C) results (although some of them substantially perturbed the results in the $C \approx \bar{k}$ region). Below we report briefly on two extensions that did alter the asymptotic results somewhat more significantly. See [3] for a more detailed description of these two extensions and their results.

5.1 Sampling

In our basic model, we evaluate the utility of a flow at a single load level; that is, we assume that a flow shares the link with $k-1$ other flows with probability $Q(k) = \frac{kP(k)}{\bar{k}}$, and that the load level is constant for the entire duration of the flow. In reality, during the lifetime of a given flow other flows might arrive and/or depart, so a flow could, and usually will, experience a fluctuating load level rather than a constant one. This fluctuating load level, in turn, creates fluctuations in the instantaneous application performance; *e.g.* the picture quality of a video stream will vary over time in a teleconferencing application. Given this varying quality, a user's utility may not merely be the average performance experienced, but may instead be closer to the minimal performance experienced.

To understand what impact this might have on our results, we examined an extension to our model where a flow samples its performance S times, with $S > 1$. For each sample, the number of flows sharing the link is picked independently from the distribution $Q(k) = \frac{kP(k)}{\bar{k}}$, and the performance is a function of the maximal value k from those S samples. For the reservation case we have to also stipulate that the acceptance/rejection decision is based on the first sample (*i.e.*, if $k > k_{max}(C)$ the flow is admitted with probability $\frac{k_{max}(C)}{k}$) and then the effective load for the subsequent samples is taken to be $\min[k_{max}(C), k]$ (*i.e.*, once the flow is admitted, it never faces a total load greater than $k_{max}(C)$).

The resulting formulae for the normalized average utilities $R(C)$ and $B(C)$ are:

$$B(C) = \sum_{k=1}^{\infty} \tilde{Q}_S(k)\pi(\frac{C}{k})$$

and

$$R(C) = \sum_{k=1}^{k_{max}(C)} Q(k) \left(\sum_{j=1}^{k_{max}(C)} \tilde{Q}_{S-1}(j)\pi(\frac{C}{\max[k,j]}) + \sum_{j=k_{max}(C)+1}^{\infty} \tilde{Q}_{S-1}(j)\pi(\frac{C}{k_{max}(C)}) \right) + \sum_{k=k_{max}(C)+1}^{\infty} Q(k)\frac{k_{max}(C)}{k}\pi(\frac{C}{k_{max}(C)})$$

with $\tilde{Q}_S(k)$ the probability that k is the maximal value obtained in S independent samples.

Multiple samplings has little effect on the the Poisson case since this distribution results in very little variance in load. The exponential and algebraic cases, on the other hand, reveal significant changes. With both adaptive and rigid applications, the performance and bandwidth gaps between best-effort and reservations increase relative to the original model. For example, the performance gap $\delta(C)$ in the exponential distribution with adaptive applications has a value of .21 at capacity $2\bar{k}$ in the sampling model; the corresponding value in the original model was less than .01. This difference is also reflected in the bandwidth gap, $\Delta(C)$. In the basic model, $\Delta(C)$ had a peak of less than $.1\bar{k}$ of the load occurring for $C \approx .5\bar{k}$. With multiple samplings, the peak in the bandwidth gap $\Delta(C)$ occurs for $C \approx 1.5\bar{k}$ and has a value of roughly $2\bar{k}$. However, asymptotically $\Delta(C)$ in this case still converges to zero. Similar changes are evident with the algebraic distribution.

Corresponding versions of the above equations in the continuum model allow us to understand the asymptotic behavior with multiple samples. Looking at the exponential distribution with rigid applications, we find that $\delta(C) \approx e^{-\beta C}(S(1+\beta C)-1)$ and $\Delta(C) \approx \frac{\ln C}{\beta}$, so the sampling extension does not significantly alter the asymptotic nature of our results. Similarly, the asymptotic results for $\gamma(p)$ in the limit of small p are not altered by the sampling extension.

We next consider the algebraic distribution with rigid applications. For large C and fixed S, we have $\delta(C) \approx C^{2-z}(S - \frac{1}{z-1})$ and $\lim_{C\to\infty} \frac{C+\Delta(C)}{C} = \lim_{p\to 0^+} \gamma(p) = (S(z-1))^{\frac{1}{z-2}}$. Note that here in the limit as $z \to 2^+$, the asymptotic ratio $\frac{\Delta(C)}{C}$ for large C and the asymptotic price ratio $\gamma(p)$ for small p diverges for any $S > 1$. Thus, we no longer have the apparent bounds of $e-1$ on the asymptotic ratio $\frac{\Delta(C)}{C}$ and of e on the asymptotic ratio $\gamma(p)$ that we had in the basic model.

Moreover, when one computes the analogous quantities for adaptive utilities, the asymptotic ratios are given by:

$$\lim_{C\to\infty} \frac{C+\Delta(C)}{C} = \lim_{p\to 0^+} \gamma(p) = \left(\left(S + \frac{a(1-a^{z-2})}{1-a} + S(1-a^{z-1})\frac{z-2}{z-1} \right) (z-1) \right)^{\frac{1}{z-2}}$$

Thus, even with adaptive applications, these limits still diverge in the limit $z \to 2^+$. If the load distribution is algebraic with z close to 2, then the amount of extra bandwidth needed to close the performance gap is exceedingly large, and that unless the cost penalty for reservation-capable networks is extremely high, reservation-capable networks provide higher levels of welfare.[9]

5.2 Retrying

In our basic model of a reservation-capable network, a rejected flow is modeled as having zero utility. In reality, however, rejected flows may try again at some later time. If a previously rejected flow is admitted at some later point, then it receives its full performance utility, but there is likely to be some user dissatisfaction due to the delay incurred. Thus, reservation networks trade off assured levels of performance at the cost of delay in getting access to the network.

We can model this by assuming that there is a utility penalty for having to retry, call it α. To avoid having to model the actual retry process, we assume that the retries of these rejected flows obey the same basic distribution as the original probability distribution. This is best expressed by introducing the notation $P_L(k)$ denoting the distribution with average $\bar{k} = L$.[10] Then, if the original model is described by some parameter L, the total offered load including retries is then given by $P_{\tilde{L}}(k)$ for some $\tilde{L} > L$.

The average utility $\tilde{R}_L(C)$ once we incorporate retries is given by:

$$\tilde{R}_L(C) = \frac{1}{L}\left(\tilde{L}R_{\tilde{L}}(C) - \alpha DL\right) = \frac{\tilde{L}}{L}R_{\tilde{L}}(C) - \alpha D$$

where $R(C)$ represents the average per-flow utility in our basic model without retries and D denotes the average number of retries each flow makes.

[9] An interesting aspect of this extension is that even with elastic applications (*e.g.*, $\pi(b) = 1 - e^{-b}$) the reservation-capable network can provide higher utility. However, in this case we need to discard the standard value of $k_{max}(C)$ as the maximizer of $k\pi(\frac{C}{k})$, which is infinite for elastic applications, and use some finite value.

[10] We use L rather than $\bar{k}$ because we are treating the load level as a variable here.

With $\alpha = .1$ (a flow suffers a performance penalty of 0.1 each time it retries), the Poisson and exponential cases show minimal effects of retrying, but the algebraic cases exhibit significant changes. Interestingly, these effects are more apparent in the region $C \gg \bar{k}$. For instance, with adaptive utility, the performance gap $\delta(C)$ has a value of .027 at capacity $4\bar{k}$ with retries, in contrast to a value of .0025 without retries. More significantly, perhaps, the price ratio curve $\gamma(p)$, which in all previous cases was monotonically increasing, now decreases for very small p. This means that as bandwidth gets cheaper, the advantage of reservation-capable networks increases! The theory suggests (see below) that this curve does not increase without bound as p decreases, but instead converges to some finite maximum value.

The continuum formulation does not yield a closed form solution. However, we can analyze the large C limit where the blocking rate is small. Let θ_L denote the blocking rate at load L; for large C, θ_L and $\tilde{L} - L$ are small, and so θ_L is a good approximation to $\theta_{\tilde{L}}$. Here, to first order in θ_L, we have $\tilde{L} \approx \frac{L}{1-\theta_{\tilde{L}}} \approx L(1 + \theta_L)$ and $D \approx \theta_L$. Note that $R_{\tilde{L}}(C) = R_L(C\frac{L}{\tilde{L}})$ so:

$$\tilde{R}_L(C) \approx (1 + \theta_L) R_L(C(1 - \theta_L)) - \alpha\theta_L$$

But in the large C limit $R_L(C) \approx 1 - \theta_L$ (the average utility per flow is just unity minus the blocking rate) so, to first order in θ_L, the expression becomes:

$$\tilde{R}_L(C) \approx 1 - \alpha\theta_L$$

This just expresses the obvious result that for large C, the only disutility is the penalty for retrying.

For exponential loads, $\theta = e^{-\beta C}$ so $\tilde{R}_L(C) \approx 1 - \alpha e^{-\beta C}$ for large C. For rigid applications, $\Delta(C)$ remains logarithmic in C and for adaptive applications $\Delta(C) \approx \frac{-\ln(\alpha(1-a))}{\beta}$. So retrying changes little in the exponential case except the asymptotic constant for adaptive applications.

For algebraic loads, $\theta = \frac{C^{2-z}}{z-1}$. For rigid applications:

$$\lim_{C\to\infty} \frac{C + \Delta(C)}{C} = \lim_{p\to 0^+} \gamma(p) = (\frac{\alpha}{z-1})^{\frac{-1}{z-2}}$$

For adaptive applications:

$$\lim_{C\to\infty} \frac{C + \Delta(C)}{C} = \lim_{p\to 0^+} \gamma(p) = (\frac{\alpha(1 - a^{z-1})}{1-a})^{\frac{-1}{z-2}}$$

Note that in both the rigid and adaptive cases, the asymptotic ratios $\frac{\Delta(C)}{C}$ and $\gamma(p)$ diverge in the limit $z \to 2^+$. Thus, extending our basic model to incorporate retrying blocked flow requests leaves most of the qualitative results unchanged, except that now in the algebraic case the asymptotic ratios $\frac{\Delta(C)}{C}$ and $\gamma(p)$ are unbounded in the $z \to 2^+$ limit.

Note that both of the extensions presented here retained the property that in the algebraic case $\Delta(C)$ was proportional to C in the large C limit, and $\gamma(p)$ was finite in the small p limit, and for the other two distributions these quantities were smaller. This appears to be quite a generic property of such models; to see this, consider rigid applications where the analysis is easier. Clearly, for large C the disutility for reservation-capable networks, $1 - R(C)$, must be proportional to the fraction of blocked flows θ. The disutility for best-effort-only networks, $1 - B(C)$, is proportional to the number of flows that are present during overload periods; call this ζ. As long as the ratio $\frac{\zeta}{\theta}$ is finite, the above generic results hold. In our class of models, the only time this ratio diverges is when the average number of flows is infinite, which occurs for $z \leq 2$. One other way the ratio might diverge is if flows are very long lived, and so each flow will eventually experience an overload condition. This would correspond to the case of S diverging in our sampling extension.

6 Discussion

We now review our results and discuss the implications of this analysis on whether or not the Internet should adopt a reservation-capable architecture. Our analysis addressed both rigid and adaptive applications. Before beginning our review we should note that questions about the extent of application adaptivity remain. Certainly the rigid utility function embodies an extreme that should not be representative of any future Internet application. However, the adaptive utility function we used embodies fairly large changes in utility across a wide range of bandwidths, both above and below the bandwidth level at which a reservation-capable architecture would admit such an application. Thus, it too may represent an extreme case by overstating the extent to which applications can adapt. Hence, we caution that the rigid and adaptive utility functions we used may in fact represent two extremes on a continuum, and that reality may lie somewhere in between.

Our initial model showed significant performance and bandwidth gaps between best-effort-only and reservation-capable networks with rigid applications. This was true even with the Poisson distribution, which is the load model that exhibits the least variance among the three distributions we considered. For example, across a wide range of bandwidth prices, reservations were superior to best-effort even if the complexity of the reservation architecture adds 10% to the total cost of the network.

Considering adaptive applications changed the picture dramatically. The basic model does not make a case for a reservation-capable network with exponential or Poisson load models and adaptive applications. With the Poisson model, at all but the highest price levels the two architectures perform the same. With the exponential model, differences between the two architectures are still very small. Only with the algebraic distribution does there appear to be doubt. In this case, the bandwidth gap grows linearly as a function of capacity and the price ratio at which welfare is equalized does not converge to one as bandwidth becomes cheap. Here the answer depends on how much cost the increased complexity of reservations adds to the network.

Two extensions to our basic model – using sampling to account for variation in performance over time, and including retries in a reservation-capable architecture – increased gaps between the two architectures for both rigid and adaptive applications. These changes could be seen in two different price regimes. For example, if bandwidth is relatively expensive and $C \approx \bar{k}$, sampling opened up a large gap with the exponential distribution and adaptive applications. When bandwidth is relatively cheap and $C \gg \bar{k}$, the extensions changed the asymptotic behavior for the algebraic distribution. For example, in our basic model, the asymptotic extra benefit of reservation-capable networks are bounded in the $z \to 2^+$ limit, with $\lim_{p \to 0^+} \gamma(p) \leq e$ and $\lim_{C \to \infty} \frac{\Delta(C)}{C} \leq (e-1)$. However, in the modified model with retries and/or sampling, these limits are removed and the performance advantages become unbounded in the $z \to 2^+$ limit.

Our results make clear that the answer to our original question depends in large part on the load patterns in the future Internet. In general, there is not a strong case for reservations with Poisson and exponential distributions. The tail of these distributions is such that a reasonable amount of provisioning likely makes the differences between the two architectures insignificant. With the algebraic distribution, particularly with a low z value, reservations yielded significant benefits. In this case, best effort performance degrades under the wider variance in load. It is not at all clear how likely it is that network loads will be described by such distributions. However, recent results on self-similar behavior in a variety of contexts [1, 5, 9, 11] make algebraic distributions less far-fetched than they might have been a few years ago. Nonetheless, there is still no definitive evidence for them that we are aware of. Thus, while our results are frustratingly ambiguous on the fundamental question of which architecture is best, they do unambiguously point to the need to more fully understand the load distributions future networks are likely to face.

References

[1] Jan Beran, Robert Sherman, Murad S. Taqqu, and Walter Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43(2):1566–1579, February 1995.

[2] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, and S. Jamin. Resource ReSerVation protocol (RSVP) – version 1 functional specification. Technical Report RFC 2205, Internet Engineering Task Force, September 1997.

[3] Lee Breslau and Scott Shenker. Best-effort versus reservations: A simple comparative analysis. Submitted to ACM Transactions on Networking, June 1998.

[4] David D. Clark, Scott Shenker, and Lixia Zhang. Supporting real-time applications in an integrated services packet network: Architecture and mechanism. In *Proceedings of ACM Sigcomm*, pages 14–26, August 1992.

[5] Mark Crovella and Azer Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. In *Proceedings of SIGMETRICS '96*, 1996.

[6] Domenico Ferrari, Anindo Banerjea, and Hui Zhang. Network support for multimedia: A discussion of the Tenet approach. *Computer Networks and ISDN Systems*, 10:1267–1280, July 1994.

[7] Sally Floyd. Comments on measurement-based admissions control for controlled-load services. submitted to CCR, July 1996.

[8] Sugih Jamin, Peter B. Danzig, Scott J. Shenker, and Lixia Zhang. A measurement-based admission control algorithm for integrated services packet networks. *IEEE/ACM Transactions on Networking*, 5(1):56–70, February 1997.

[9] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[10] Abhay K. Parekh and Robert G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, June 1993.

[11] Vern Paxson and Sally Floyd. Wide-area traffic: the failure of Poisson modeling. In *Proceedings of ACM Sigcomm*, pages 257–268, London, United Kingdom, August 1994. ACM.

[12] S. Shenker, C. Partridge, and R. Guerin. Specification of guaranteed quality of service. RFC 2212, Internet Engineering Task Force, September 1997.

[13] S. Shenker and J. Wroclawski. Network element service specification template. Technical Report RFC 2216, Internet Engineering Task Force, September 1997.

[14] Scott Shenker. Fundamental design issues for the future internet. *IEEE Journal on Selected Areas in Communications*, 13(7), September 1995.

[15] C. Topolcic. Experimental internet stream protocol, version 2 (ST-II). RFC 1190, SRI Network Information Center, October 1990.

[16] J. Wroclawski. Specification of the controlled-load network element service. RFC 2211, Internet Engineering Task Force, September 1997.

[17] J. Wroclawski. The use of RSVP with IETF integrated services. Technical Report RFC 2210, Internet Engineering Task Force, September 1997.

[18] Lixia Zhang, Steve Deering, Deborah Estrin, Scott Shenker, and Daniel Zappala. RSVP: A new resource reservation protocol. *IEEE Network Magazine*, 7(5):8–18, September 1993.

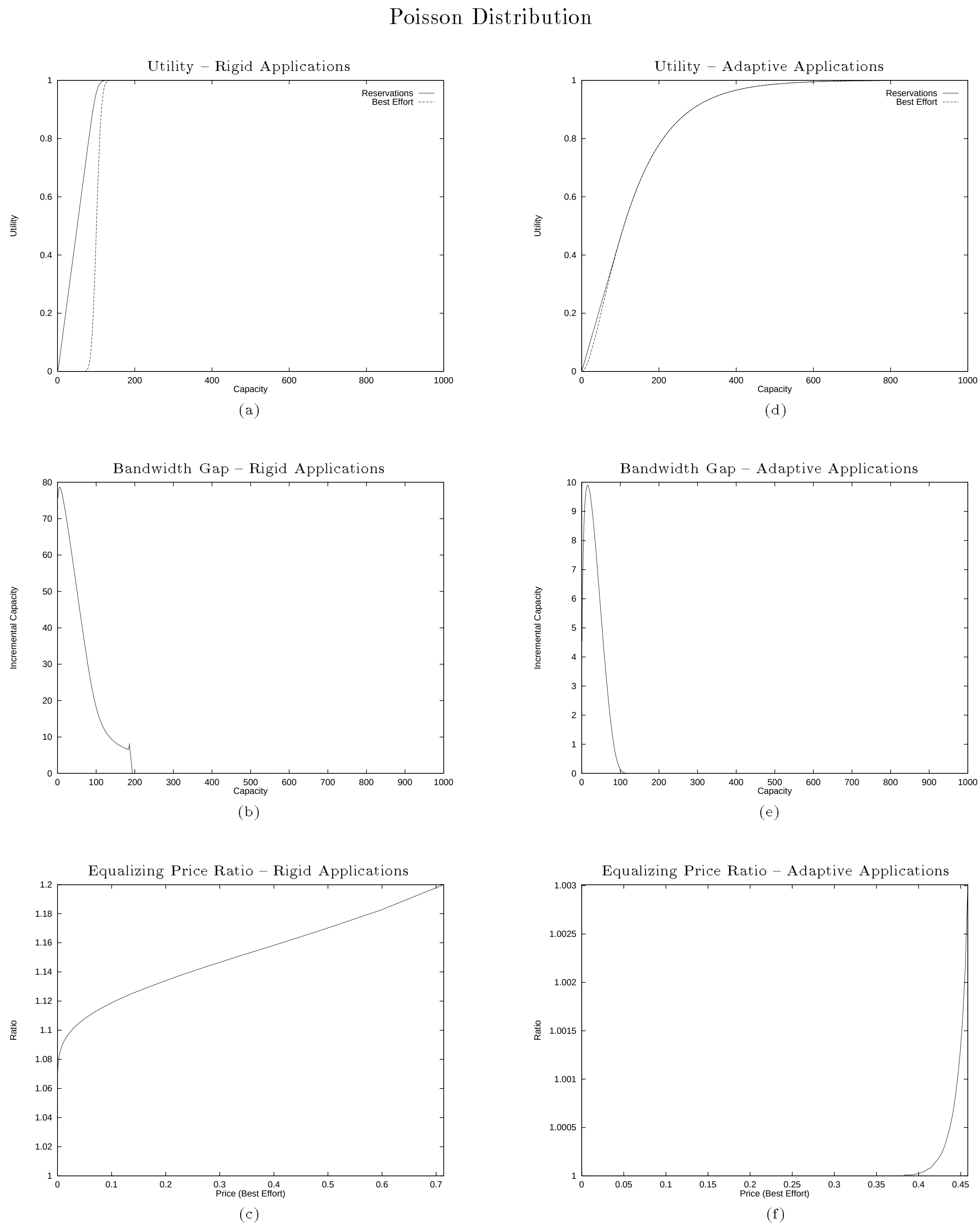

Figure 2: Poisson Distribution – Utility, Bandwidth Gap, and Price Ratio to Equalize Welfare for Rigid and Adaptive Applications.

Exponential Distribution

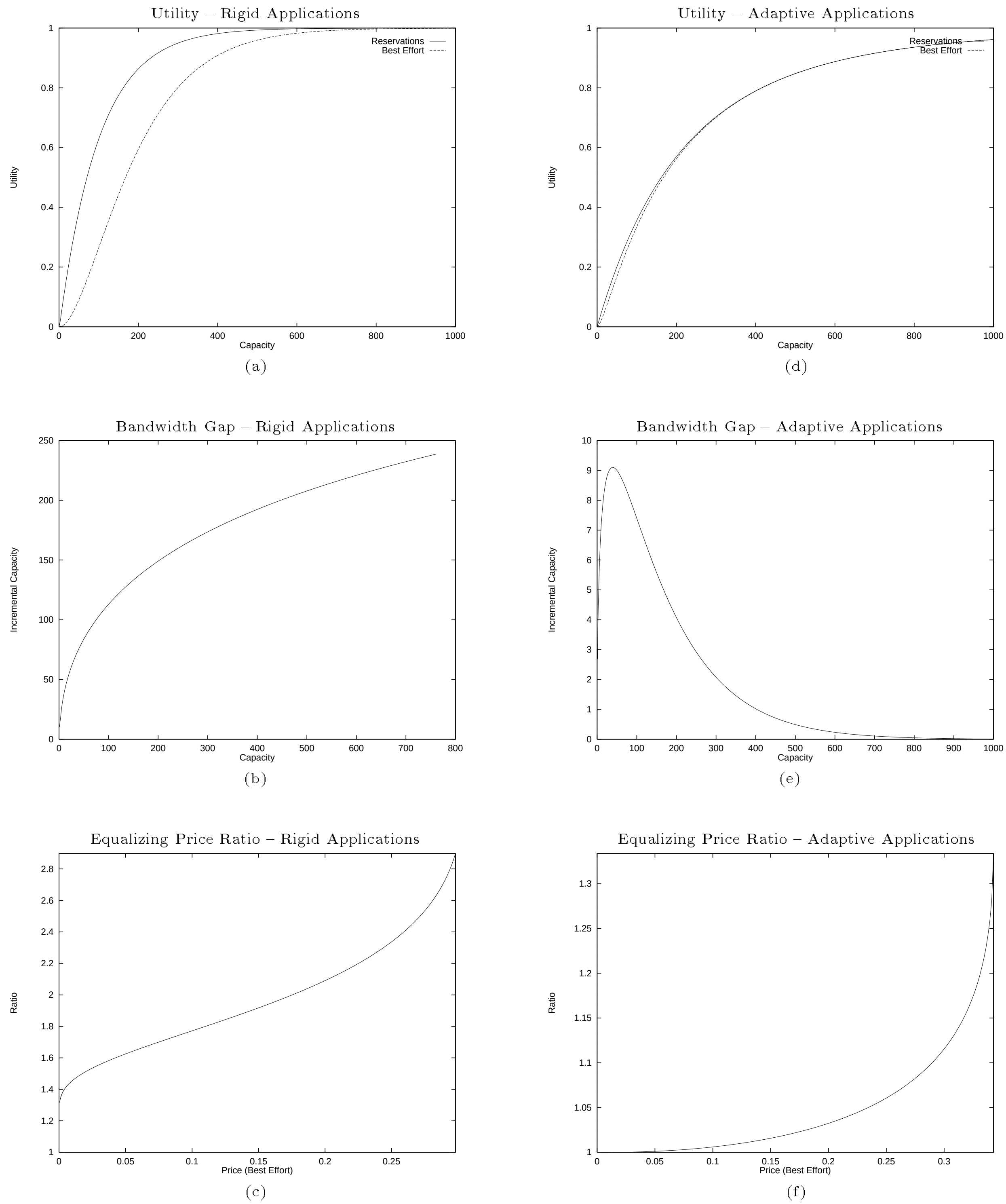

Figure 3: Exponential Distribution – Utility, Bandwidth Gap, and Price Ratio to Equalize Welfare for Rigid and Adaptive Applications.

Algebraic Distribution

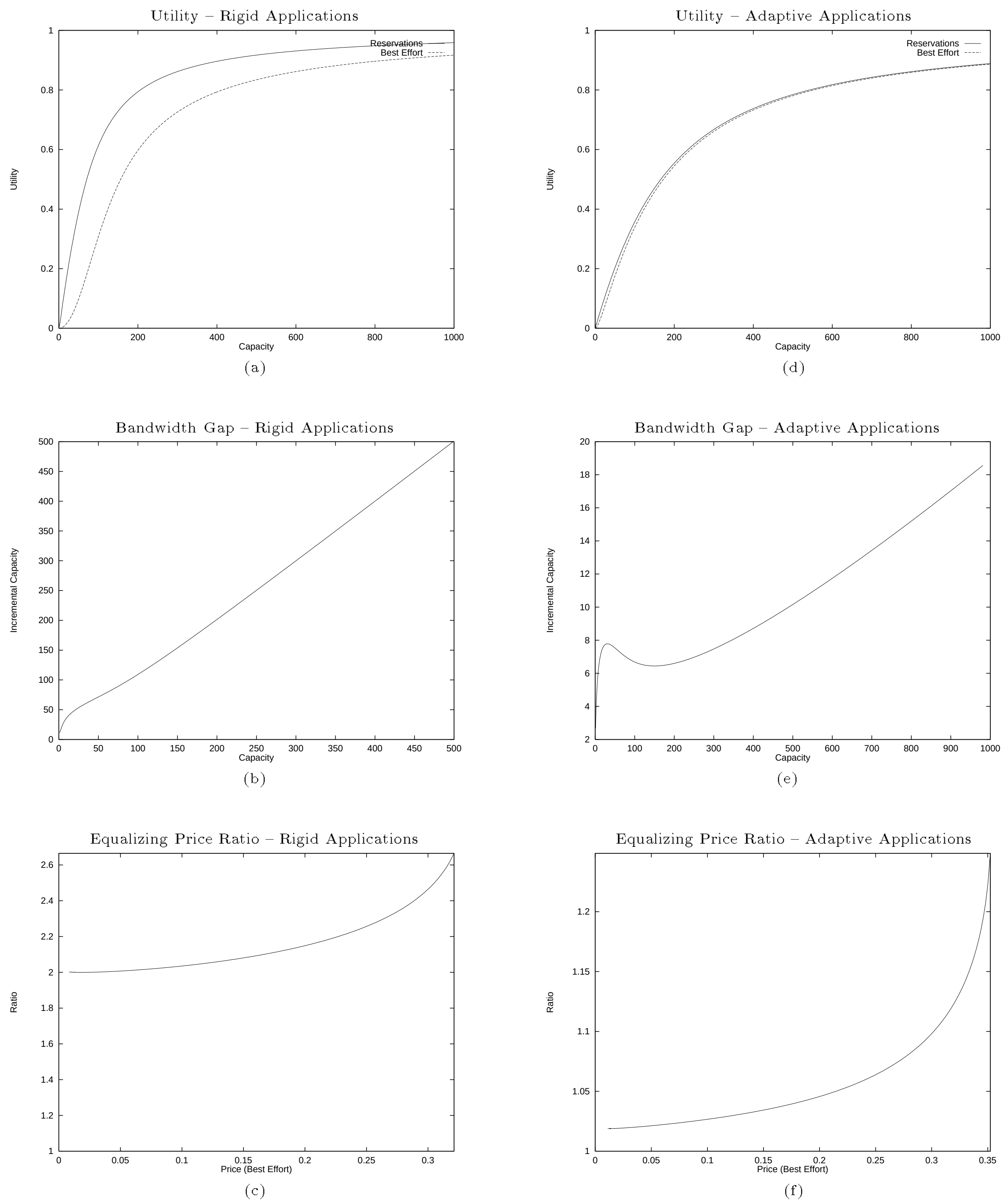

Figure 4: Algebraic Distribution – Utility, Bandwidth Gap, and Price Ratio to Equalize Welfare for Rigid and Adaptive Applications.