Best-Effort versus Reservations: A Simple Comparative Analysis

Lee Breslau
Scott Shenker

September 2, 1998

Context

Question: How best to support real-time applications in the Internet?

One answer: Extend Internet architecture to support resource reservations

- applications explicitly request enhanced quality of service from the network
- network says *yes* or *no*

Status:

- lots of research, standardization activity and product development
- however, widespread disagreement about the wisdom of resource reservations remains

Basic Argument: 1991

Deering The best-effort Internet works just fine as it is! Why mess with success?

Shenker Sure it works great for data applications, but some audio and video applications need reservations.

Deering Modern audio and video applications are *adaptive* and therefore don't need reservations.

Shenker Yes, but even some adaptive audio and video applications need reservations to perform adequately.

Deering No, they don't.

Shenker Yes, they do.

Deering No, they don't.

Shenker Yes, they do.

. . .

Basic Argument: 1998

...

Deering No, they don't.

Shenker Yes, they do.

Deering No, they don't.

Shenker Yes, they do.

...

Goals:

- Develop a *simple* model that captures key issues
- Increase our understanding of the essential features
- Inform the debate

Non-goals:

- A model that completely reflects reality
- Characterization of costs of resource reservations
- Settle the debate

Basic Model

Link of capacity C shared by k flows

Per flow utility, π is a function of a flow's bandwidth share b

- $\pi(0) = 0$
- $\pi(\infty) = 1$
- non-decreasing

If k flows each receive equal bandwidth total utility equals:

- $V = k\pi(\frac{C}{k})$

Variable load represented by $P(k)$

Basic Model (cont.)

Best Effort

- $V_B(C) = \sum_{k=1}^{\infty} P(k)k\pi(\frac{C}{k})$

Reservations

- For a certain class of utility functions, utility is maximized by limiting number of flows to $kmax$
- $V_R(C) = \sum_{k=1}^{k_{max}} P(k)k\pi(\frac{C}{k}) +$

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

Discrete model allows direct computation; continuum version enables examination of asymptotic behavior as C increases

$V_R(C) \geq V_B(C)$, but by how much?

Performance Measures

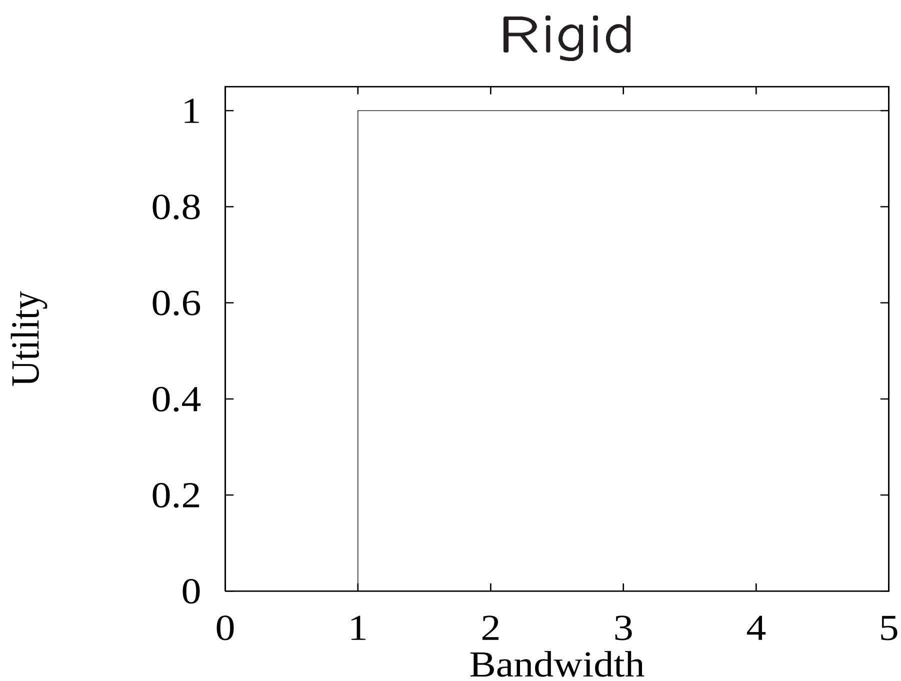
Performance gap, δ

- $\delta(C) = V_R(C) - V_B(C)$

Bandwidth gap, Δ

- How much additional bandwidth must be added to a best-effort network to achieve the same utility as a reservation network?
- $V_R(C) = V_B(C + \Delta(C))$

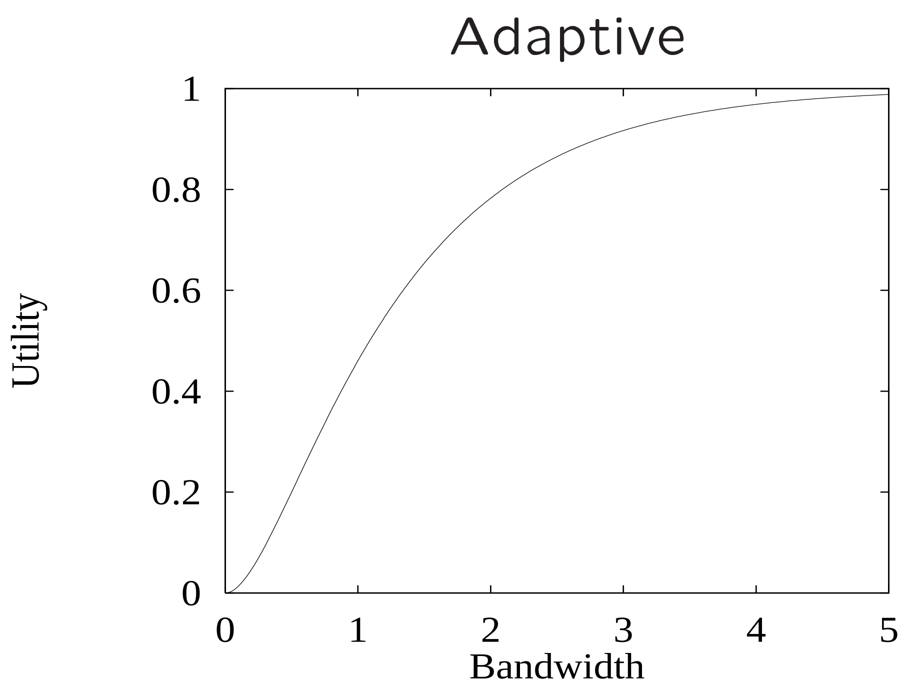
Utility Functions – $\pi(b)$



$\pi(b) = 0$ for $b < \bar{b}$

$\pi(b) = 1$ for $b \geq \bar{b}$

Utility Functions – $\pi(b)$ (cont.)



Minimum bandwidth requirement

Significant marginal utility over a wide range of b

- able to adjust to different levels of network service
- still benefit from reservations

Load Models – $P(k)$

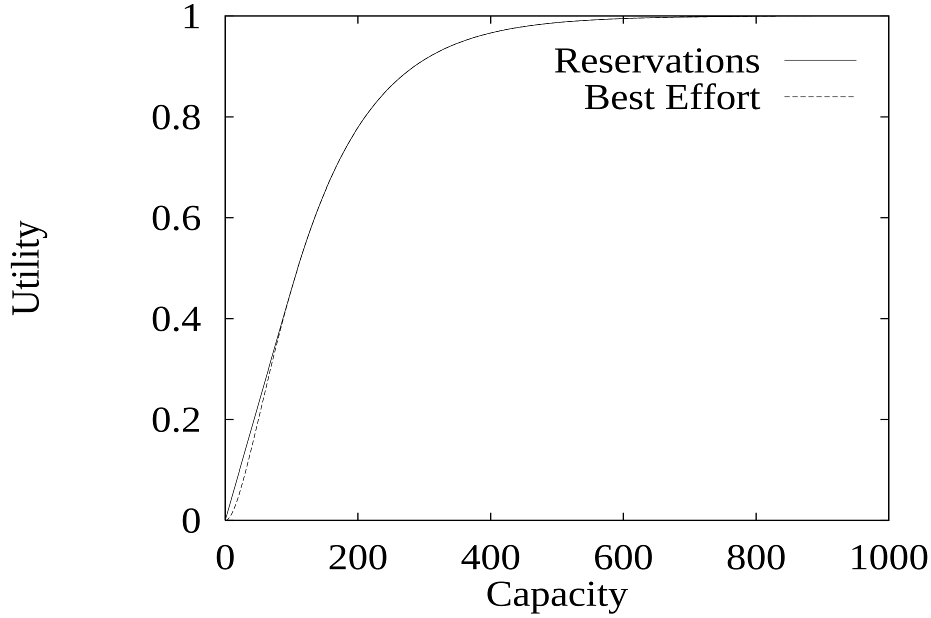
3 distributions

- Poisson: $P(k) = \frac{\nu^k e^{-\nu}}{k!}$
- Exponential: $P(k) = (1 - e^{-\beta})e^{-\beta k}$
- Algebraic: $P(k) = \frac{\nu}{\lambda + k^z}$

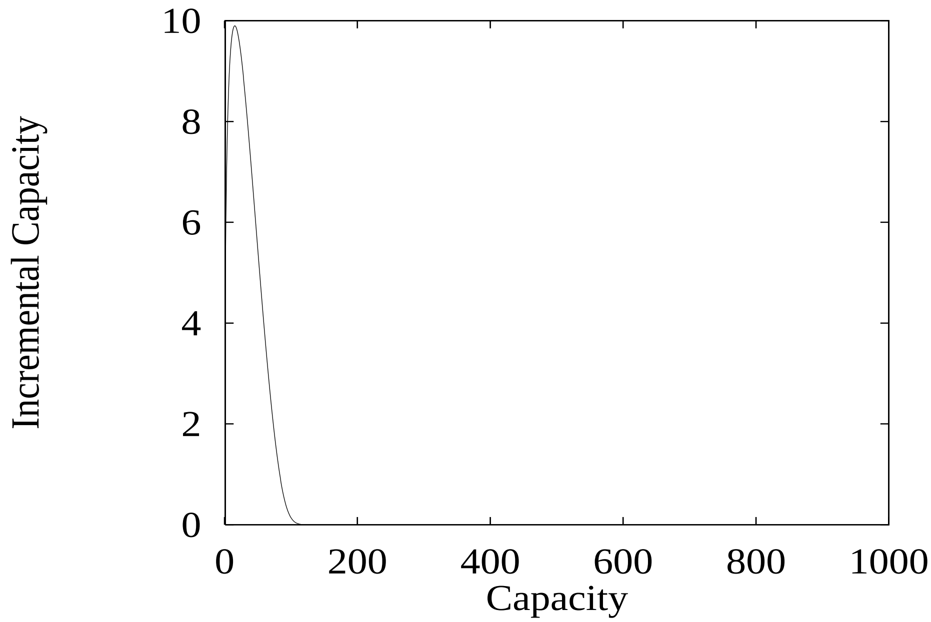
Represent a range of load models, no claim about their validity

Results – Poisson Adaptive

Performance

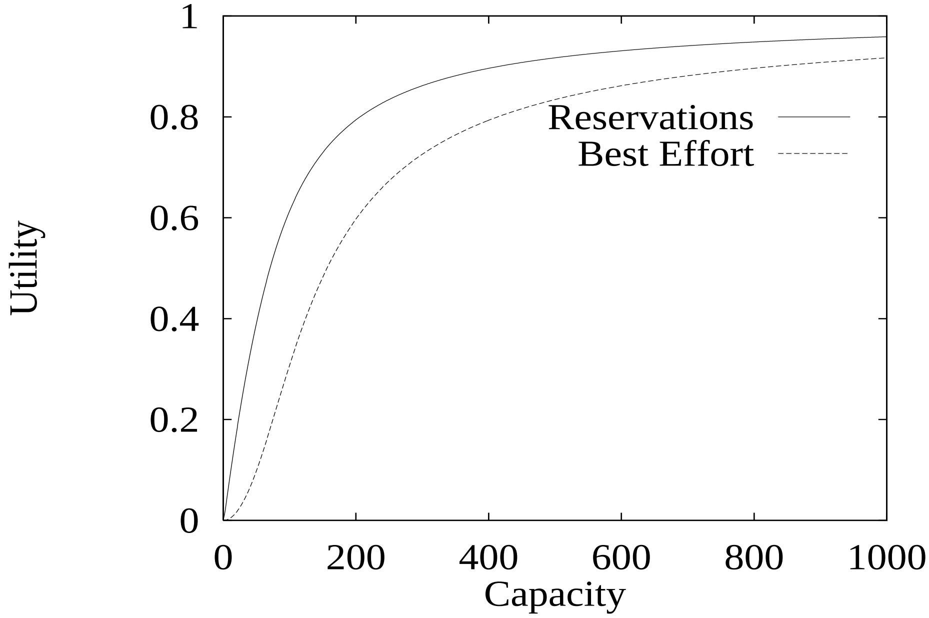


Bandwidth Gap

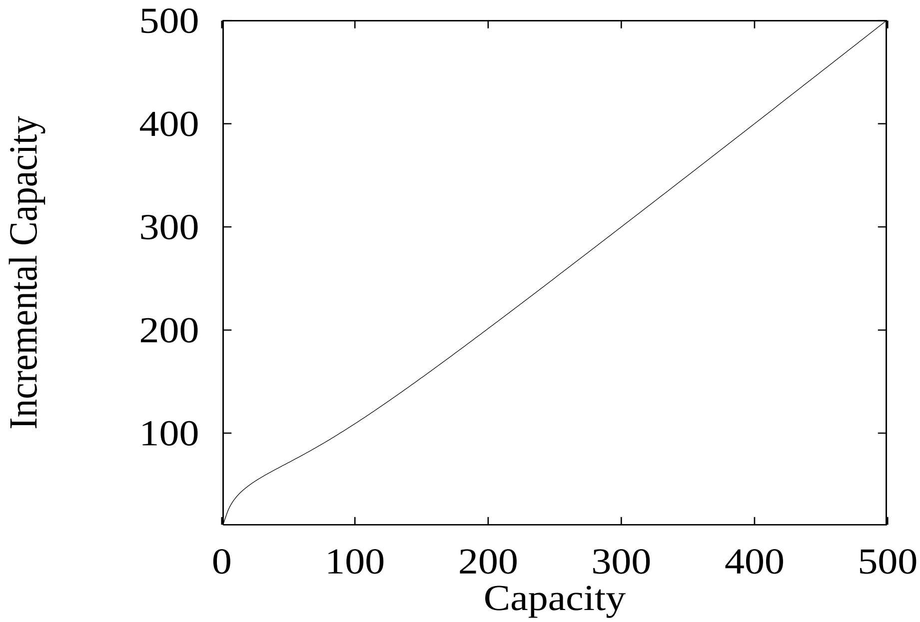


Results – Algebraic Rigid

Performance



Bandwidth Gap



Summary of Results

Performance Gap, δ

- Significant for small C (i.e., $C < L$) but quickly converges to zero (except in the algebraic case)

Bandwidth Gap, Δ

- Poisson: $\Delta \rightarrow 0$
- Exponential/Adaptive: $\Delta \rightarrow 0$
- Exponential/Rigid: $\Delta \approx \ln C$
- Algebraic: $\Delta \propto C$

Conjecture:

- $\Delta(C) = (e - 1)C$ is maximum bandwidth gap

Variable Capacity

Given a price per unit bandwidth p, provision network to maximize total welfare: $V(C) - pC$

Compute capacity as a function of price: $C(p)$

Total welfare:

- $W_B(p) = V_B(C_B(p)) - pC_B(p)$
- $W_R(p) = V_R(C_R(p)) - pC_R(p)$

Price ratio to equalize welfare:

- $\gamma(p) = \frac{\tilde{p}}{p}$ where $W_R(\tilde{p}) = W_B(p)$

Variable Capacity – Results

As $p \to 0$:

- Poisson: $\gamma(p) \to 1$
- Exponential: $\gamma(p) \to 1$
- Algebraic: $\gamma(p) \to \alpha$, with $\alpha > 1$

For algebraic distribution, no matter how cheap bandwidth becomes, reservation-based network retains an advantage over best-effort

Conjecture: $\lim_{p \to 0+} \gamma(p) \leq e$ for all distributions

Extensions

Sampling

- Performance varies over time
- Utility may be a function of the maximum load experienced
- For each flow, assume utility is the minimum value taken over S samples

Retry

- Rejected flows can request service later and receive non-zero utility
- But some penalty for delay
- Model rejected flows retrying as additional load

Extensions – Results

Poisson – no effect

Exponential – little effect, except with $C \approx L$ in sampling extension

Algebraic – significant change both with $C \approx L$ and in asymptotic behavior

- $\frac{\Delta(C)}{C}$ and $\gamma(p)$ no longer bounded

Conclusions

No simple answer to our original question

Over-provisioning appears sufficient with Poisson and Exponential load models

Reservations are useful with Algebraic distribution

What is the nature of future Internet load?