Network neutrality, usage-based pricing and service differentiation

George Kesidis, Penn State
kesidis@engr.psu.edu

G. de Veciana, U.T. Austin, and A. Das

Outline

- Present-day issues
  - Network neutrality
  - Managed service migration to commodity Internet, e.g., VoIP
- Motivating usage-based pricing and differentiated services
- Resource savings via differentiated services
- Comparison of flat-rate and usage-based pricing

BitTorrent issues

- The data network is not designed to handle even a small number of “persistent elephants”, i.e., large-volume users over an extended period of time: 5% of users account for 80% of traffic load.
- That is, the Internet, the access portion in particular, is designed around the concept of *statistical multiplexing*.
- If blocking BitTorrent is allowed, then why not also block/throttle every other big provider of content, e.g., Google?
- Carriers/ISPs use the network for their *own* “managed” services which constitute significant volume.
- So, American FCC ruled in favor of network “neutrality” for public ISPs.
- Comcast now introducing partially usage-based pricing to deter load, $F+C(\rho - R)^{+}$ where
  - F \$/month is the flat rate price depending on maximum access bandwidth
  - C \$/byte is the usage-based charge for net throughput

  ρ-$R>0$ for a threshold R byte/month (i.e., after which overage charges are applied)

Why not just expand the access network?

- Despite the need for statistical multiplexing, the principal of over-engineering (i.e., *no* traffic engineering in) the Internet is generally simple but economically inefficient.
- Note: issues of long-term economic efficiencies were hardly considered during the dot-com boom when the Internet was built out.
- BitTorrent users unlikely willing to pay, so present-day elephants will not finance infrastructure expansion.
- Expanded access may accelerate the migration of profitable managed services (telephony PSTN/POTS or broadcast TV) to third-party providers over "commodity" Internet access.
- So, as it is, expanding commodity Internet access infrastructure may result in *less* revenue for carriers.
- Need to improve broadband penetration (at least in the US), and generally make the Internet more profitable to those who operate it, by exploring other revenue paradigms besides those involving monopolies of network infrastructure and content.
- So, given tiered pricing, why not tiered *services* ?

Possible solution

- Invest in build-out of the Internet for “premium” service classes to generate significant additional revenues from services that use them.
- That is, develop support for differentiated/tiered services for
  - certain high-volume interactive real-time services,
  - generic data networking applications dynamically specified by the end-user, or
  - traffic aggregates (e.g., leased lines).
- Network neutrality not relevant to premium services.
- Service classes obviously need to be differentially priced to provide end-user incentives to, e.g., not designate every packet high priority or use “promotion” scheduling (later).

Service differentiation & resource savings: A two-class M/GI/1 FIFO queue

- Delay-sensitive flow requires $E[W] \leq \delta$ for some $\delta > 0$.
- Mean arrival rates of two traffic classes are λ_d (delay-sensitive) and λ_t (throughput-sensitive).
- Arrivals are packets of mean size s and variance σ^2 so that load $\rho = \lambda s$.
- Using Pollaczek-Khintchine on a FIFO queue without service differentiation, the delay constraint requires $\alpha_d(\delta) + \alpha_t(\delta) \leq c$, where c is the total service capacity and α the flow "effective bandwidth".
- Let μ_1 be the minimal c such that this inequality holds.
- Closed form expressions in [CISS'08]

Service differentiation & resource savings: A two-class M/GI/1 priority-service queue

- A queue where the delay-sensitive traffic gets pre-emptive priority requires
  - $\alpha_d(\delta) \leq c$ (delay constraint) and
  - $\alpha_d(\infty) + \alpha_t(\infty) \leq c$ (throughput constraint, i.e., stability).
- Let μ_2 be the minimal c such that these inequalities hold under service differentiation [CISS'08].
- Clearly, we found $\mu_2 < \mu_1$
  - but also the savings from introducing service differentiation reduces with total traffic load,
  - i.e., service differentiation may have a greater impact in the network edge than in the core.
- This argument can be generalized to other traffic models via large-buffer asymptotic effective bandwidths and virtual-delay tail constraints.

Flat rate pricing

- Flat rate pricing does promote growth in traffic [Odlyzko'01].
- Moreover, flat rate pricing is preferred by end-users.
- Again, raising flat prices difficult because of competition and regulations against price fixing.
- And additional traffic may not "proportionately" yield additional revenue to ISPs.
- Finally, end-users may hate a "ticking clock" but they expect additional costs for a premium service, e.g., history of long-distance telephony.
- Using tiered flat rates for premium classes of service motivates end-user "promotion" scheduling (e.g., [Haddad'07]) resulting in much higher utilization without additional revenue.

Motivating usage priced differentiated services

- Consistent with an only temporary need for dedicated bandwidth.
- Will reduce the volume of premium traffic in play as users will only assign applications to premium CoSs as needed.
- Premium applications receiving premium CoS could be statically designated (by dest port, src IP, etc.) or they could be dynamically specified by the user.
- Authentication and metering overhead, but potential generic security benefits.

Tiered flat rate versus usage-based pricing: overload conditions

- Consider a queue nominally handling only delay-sensitive traffic with a fixed service rate μ and N users.
- n^{th} user has total offered load $\rho_n = \lambda_n s_n$ (subscripts now indicate the user) and utility
  - $U_n(\underline{\rho}) = \rho_n$ if $\Sigma_k \rho_k \leq \Lambda$ and
  - $U_n(\underline{\rho}) = \rho_n \exp(- \Sigma_k \rho_k / \Lambda \beta_n)$ else;
  - where $\underline{\rho}$ is the N-vector of traffic loads, Λ is a limit to the total traffic load, and
  - β_n captures the extent to which the n^{th} user can tolerate an overload condition:
    - If $\beta_n = 0$, then the n^{th} user is intolerant of overload
    - If $\beta_n = \infty$, then the n^{th} user is actually transmitting best-effort traffic, as might be the case with automated promotion scheduling under flat-rate pricing.
- We want to compare flat versus usage-based pricing under excess demand: $\Sigma_k \rho_k > \Lambda$

Flat-rate pricing under overload conditions

- n th user has maximal load $\rho_n^{\max}$ ($C=0$ and flat rate charge F depends on $\rho^{\max}$).
- If $\Sigma_k \rho_k > \Lambda$, then $\partial U_n(\underline{\rho}) / \partial \rho_n = 0$ when $\Lambda\beta_n \leq \rho_n^{\max}$ and the arrival rate is chosen to be $\rho^f_n = \Lambda\beta_n$.
- This requires $\Sigma_k \beta_k > 1$ so that the overload condition holds under $\underline{\rho}^f$.
- Assuming $\rho^f_n \leq \rho_n^{\max}$ (feasibility) for all users n, then we need to overbook allocations in a flat rate system to utilize resources efficiently, i.e., $\Sigma_k \rho_k^{\max} > \Lambda$.

Usage-based pricing under overload conditions

- n th user chooses load ρ^u_n that maximizes net utility $U_n(\underline{\rho})-C\rho_n$ where C is the usage-based price.
- Using Jensen's inequality, we can show that $1-r \geq C\, e^{Nr}$ where $r= \Sigma_k \rho^u_k /(\Lambda\, \Sigma_k \beta_k)$.
- A necessary condition for a solution r is $C<1$ in which case:

$$\Lambda < \Sigma_k \rho^u_k < \Lambda\, \Sigma_k \beta_k /N = \Sigma_k \rho^f_k /N$$

- This requires $\Sigma_k \beta_k > N$.
- Weaker Jensen's gives just $\Sigma_k \rho^u_k < \Sigma_k \rho^f_k$ when $C<1$ and $\Sigma_k \beta_k > 1$ (as before).

Comparison between usage-based and flat-rate pricing

- So, under excess overall demand $\Sigma_k \rho_k > \Lambda$ wherein every user does not achieve their peak demand $\rho^{\max}$, overall demand is lower under usage-based pricing, recalling that we also assumed
  - a sufficiently low usage-based charge $C<1$
  - $\rho^{\max}$ not a factor (i.e., a sufficiently low flat rate cost F), specifically, feasible $\rho^f_n = \Lambda\beta_n < \rho_n^{\max}$
- This continues to hold if the β parameters are higher for flat rate pricing (again, as would be the case under promotion scheduling of best-effort traffic).
- Also, clearly $\rho^f_n > \rho^u_n$ for all users n.
- Again, the context is overload conditions.
- But we do not account for sensitivities to flat rate costs (even component costs) in this analysis, including the effects of time-of-day variations in flat rates (which are not application discriminative but may help to balance load over the course of a day).

Summary

- We used present-day issues of network neutrality and migration of managed services to motivate tiered services with usage-based pricing.
- Resource savings under differentiated services, especially in the access network, was explained using a simple queuing system.
- Finally, we developed a simple model of access under overload conditions to assess resource management with flat-rate pricing compared to usage-based pricing.