Providing End-to-End Statistical Performance Guarantees with Bounding Interval Dependent Stochastic Models

Edward W. Knightly

knightly@tenet.berkeley.edu

The Tenet Group
University of California at Berkeley
and International Computer Science Institute
and Sandia National Laboratories

16th Annual ILP Conference
March 10, 1994

Outline

- Problem
  - QoS guarantees
  - Statistical and deterministic service
- Shortcomings of traditional solutions
- Our solution
  - Stochastic model of source (BIND)
  - Multiplexing heterogeneous sources
  - End-to-end statistical guarantees
- Conclusion

Types of Service

- Deterministic - performance bounds are met for all packets, even in the worst case
  - model: deterministically bound the rate as in Tenet (Xmin, Xave, I, Smax) model or Rene Cruz (σ, ρ) model
  - *Can* result in low resource utilization
- Statistical - guarantees of the form $Prob(delay \leq D) \geq z$
  - Many applications (e.g. voice, video conferencing) can tolerate some loss
  - Network utilization can be higher with statistical multiplexing

Statistical Service and Traffic Model

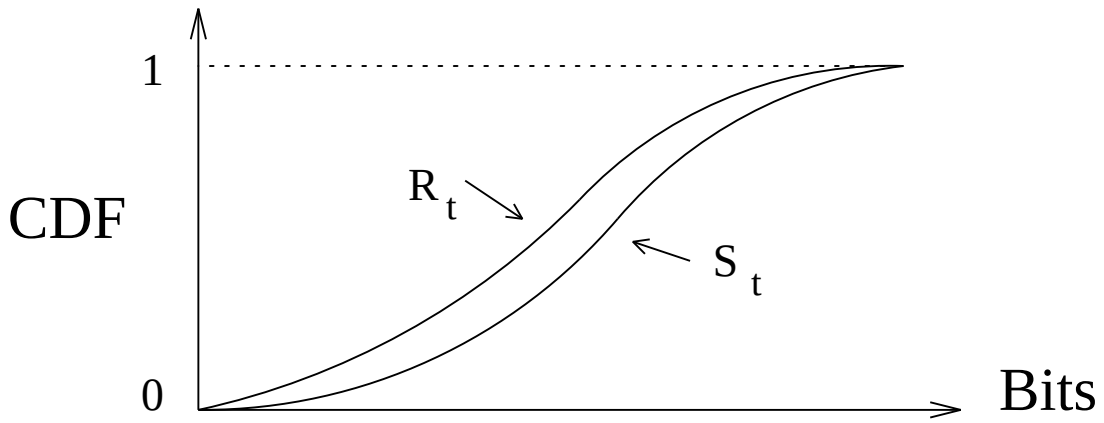
- Model: statistical service requires a stochastic model

- Traditional queueing approach: model traffic as stochastic process, approximate aggregate process and solve.

- Problems encountered with traditional queueing approaches:
  - Assume homogeneous traffic sources
  - Model fitting (as opposed to bounding)
  - Aggregate, average, or limiting results
  - Single switch analysis or no priorities

Our Model - BIND

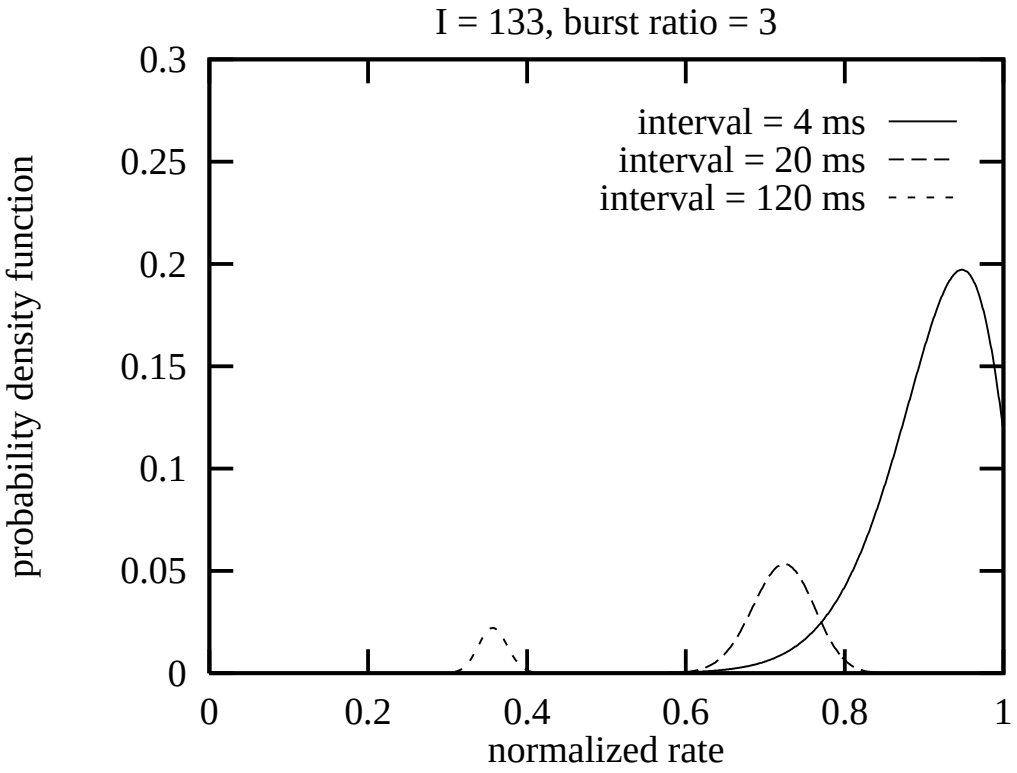
- Stochastically bound the sources rate over intervals of different length
  - $\{(\mathbf{R}_t, t) \mid t \geq 0\}$ stochastically bounds the source number of bits in an interval of length t as in [Kurose92]
  - $\mathbf{R}_t \succeq_{st} \mathbf{S}_t$ for all t



- The bounding distribution is an explicit function of the interval length

$$\frac{E(\mathbf{R}_t)}{t} \leq \frac{E(\mathbf{R}_s)}{s} \quad if \quad t > s$$

Example R_t/t distribution

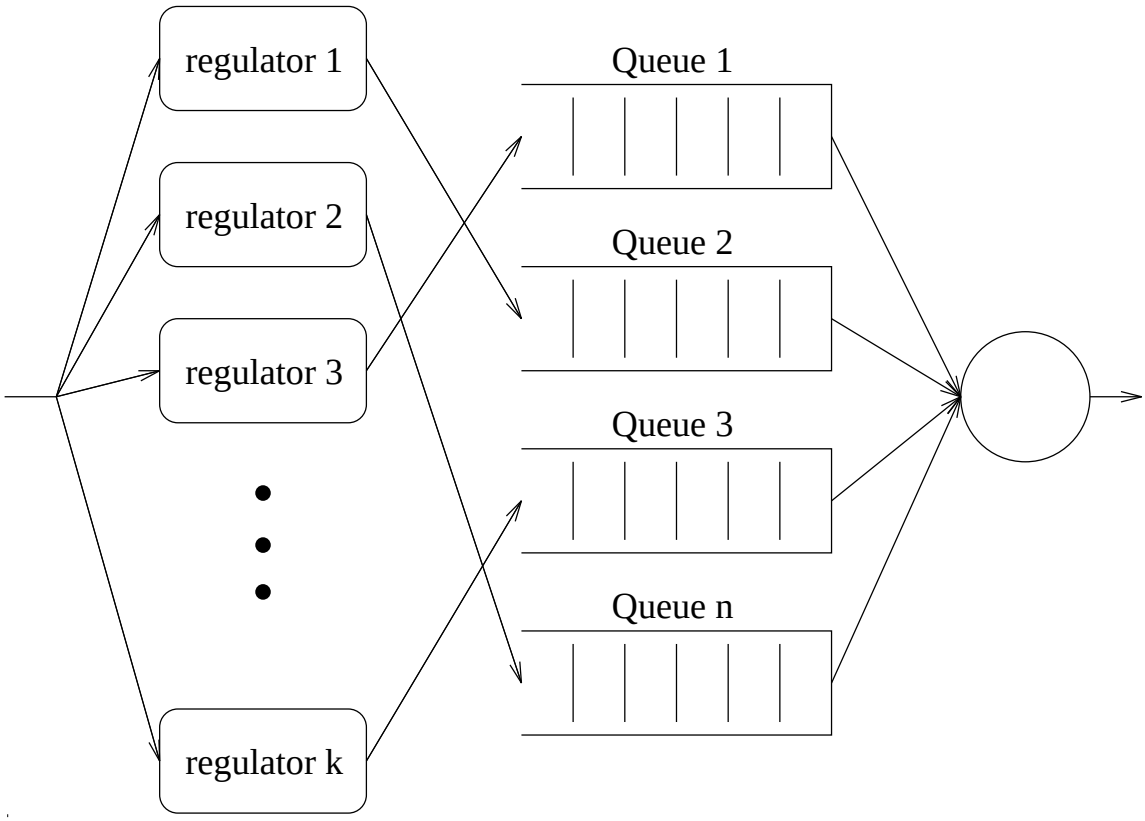


- As interval becomes large, PDF becomes concentrated at mean rate
- Sources will approach mean with different rates and variances

Example Distributions

- Discrete Stochastic BIND Model
  - Binomial (M_t, p_t)
  - $M_t = \lceil t/Xmin \rceil$
  - p_t decreases with t (with exp. rate γ)
- Continuous Stochastic BIND Model
  - Markov model with parameters $(\alpha_t,\ \beta_t,\ r_{1,t},\ r_{2,t})$
  - Different functions of t result in different families of curves

RCSP Scheduler



- Stat-mux gain depends on scheduler
- Each queue is FCFS and has a different priority based on delay; regulator enforces bandwidth and jitter constraints

Multiplexing Sources

- delay bound $\overline{d^m}$ associated with each priority level m, $(\overline{d^1} < \overline{d^2} < \cdots < \overline{d^n})$

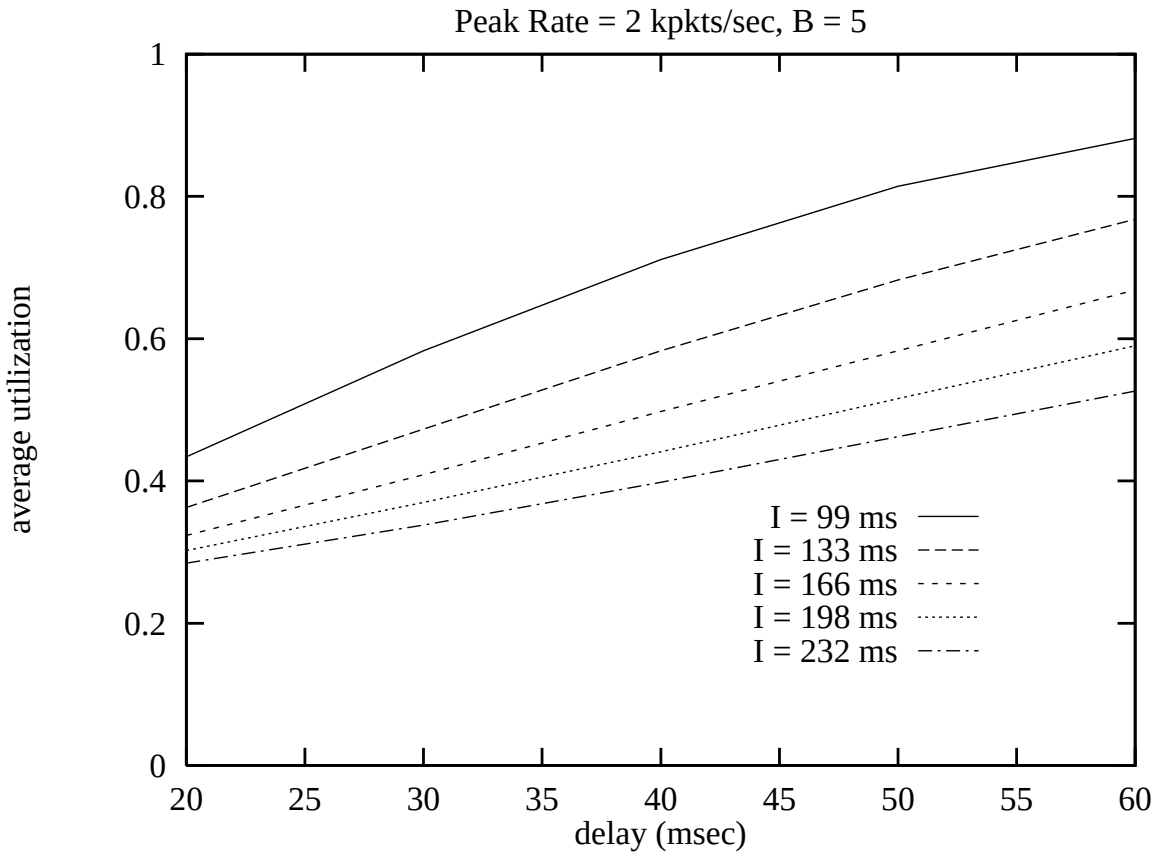
- delay-bound violation probability for level m is $Prob\{d^m > \overline{d^m}\}$

- At level q, the j^{th} connection has $\{(\mathbf{R}_{\overline{d^1},j}, \overline{d^1}), (\mathbf{R}_{\overline{d^2},j}, \overline{d^2}), \ldots, (\mathbf{R}_{\overline{d^n},j}, \overline{d^n})\}$

- If C_q is the set of connections at level q, link speed is l, and $\overline{Smax}$ is the max packet size:

$$Prob\{d^m < \overline{d^m}\} \geq Prob\{\sum_{q=1}^{m} \sum_{j \in C_q} \mathbf{R}_{\overline{d^m},j} + \overline{Smax} \geq \overline{d^m} l\}$$

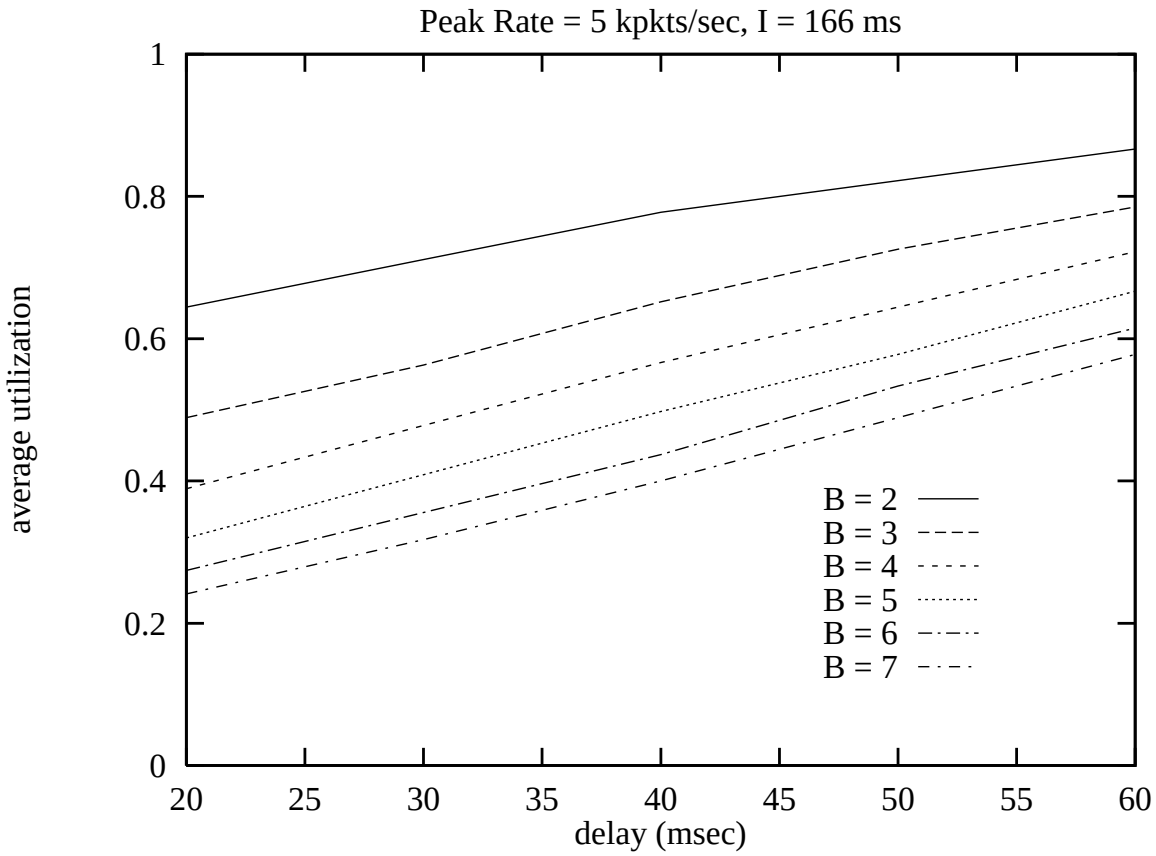
Effect of Averaging Interval



- Peak rate $= 800kbps$, $B = 5$, vary I and $\overline{d^1}$
  $l = 45Mbps$, $Prob\{d^1 > \overline{d^1}\} \leq 10^{-3}$

- Shorter interval and larger delay result in higher utilization

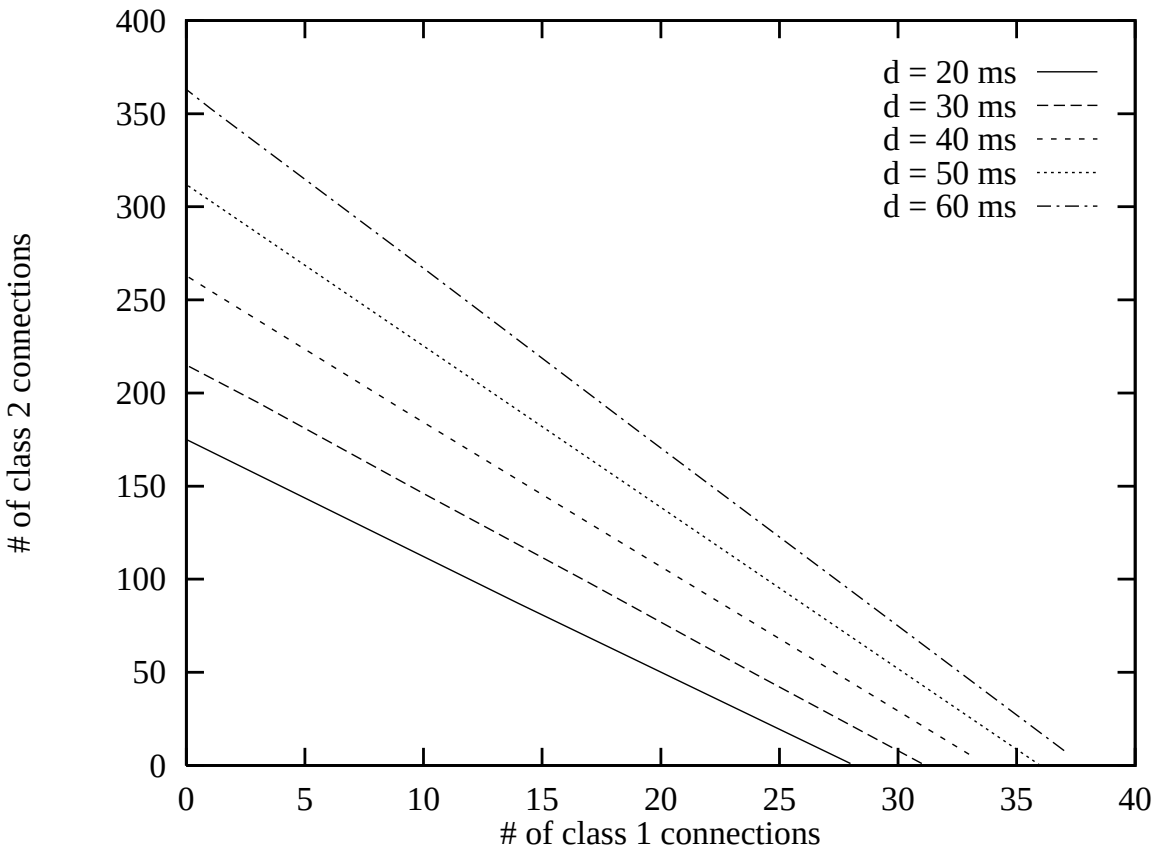
Effect of Burst Ratio



- Peak rate $= 2Mbps$, $I = 166$, vary B and $\overline{d^1}$
  $l = 45Mbps$, $Prob\{d^1 > \overline{d^1}\} \leq 10^{-3}$

- Burstier traffic results in lower average utilization but higher peak utilization

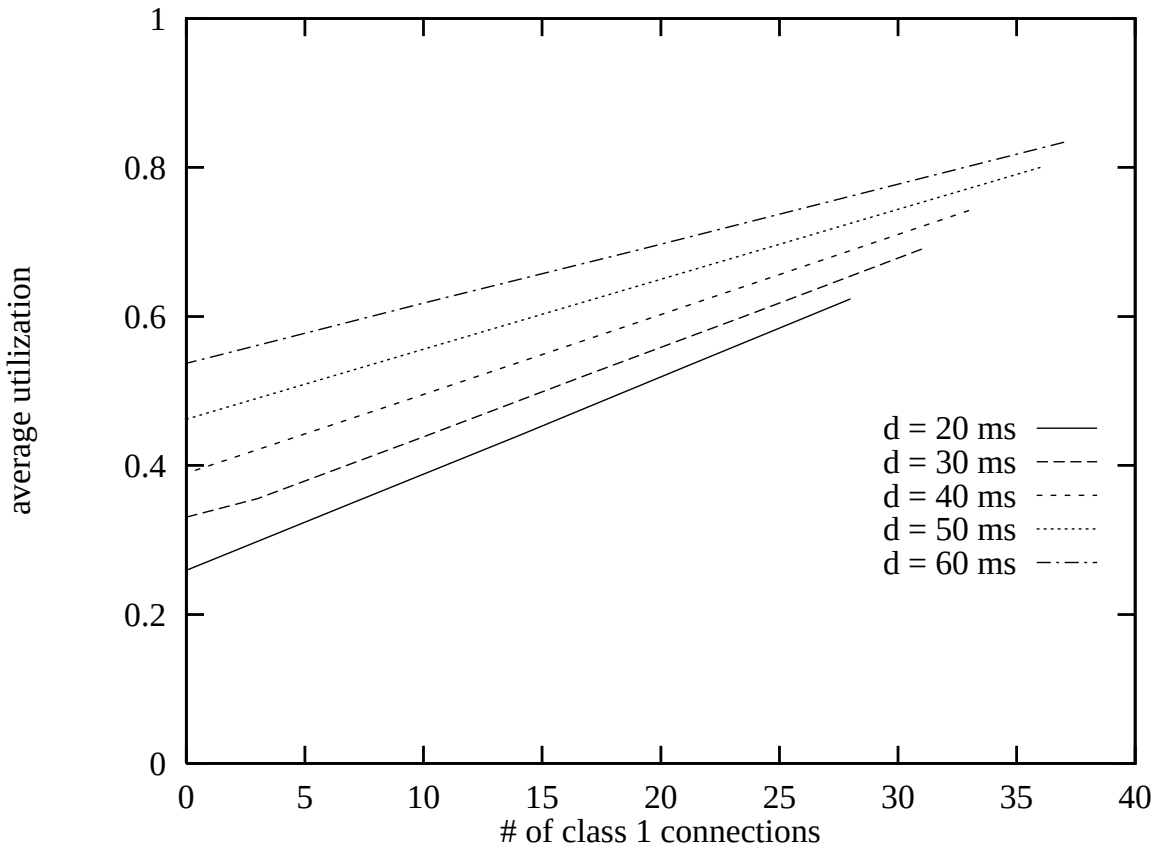
Effect of Delay with Heterogeneous Sources



- C1: peak rate = $2Mbps$, avg rate = $1Mbps$
  C2: peak = $400kbps$, avg = $66.7kbps$

- Generate bound on acceptable combinations of C1 and C2 connections

Average Utilization



- C1: peak rate = $2Mbps$, avg rate = $1Mbps$
  C2: peak = $400kbps$, avg = $66.7kbps$

- Less bursty sources result in higher utilization

Providing End-to-end Statistical Guarantees

- with distributed delay-jitter control [Ferrari91], network becomes a constant delay line
  - e.g., a MMPP at the source is a MMPP at the output queues of all switches
- we don't require:
  - restricted class of networks
  - calculation of transformation of process, busy period shift
  - unrealistic assumption of bottleneck node
- $D = \sum_{i=1}^{n} \overline{d}_i + \sum_{i=2}^{n} \overline{\pi}_i \ , \ Z = \prod_{i=1}^{n} z_i$

Conclusions

- BIND model - stochastic bounding interval-length dependent model
- Multiplexing - RCSP allows priority multiplexing of heterogeneous sources
- End-to-End guarantees - delay-jitter control efficiently extends the results to the network
- Current work - characterize real traffic with BIND model, investigate network utilizations for real sources, compact model