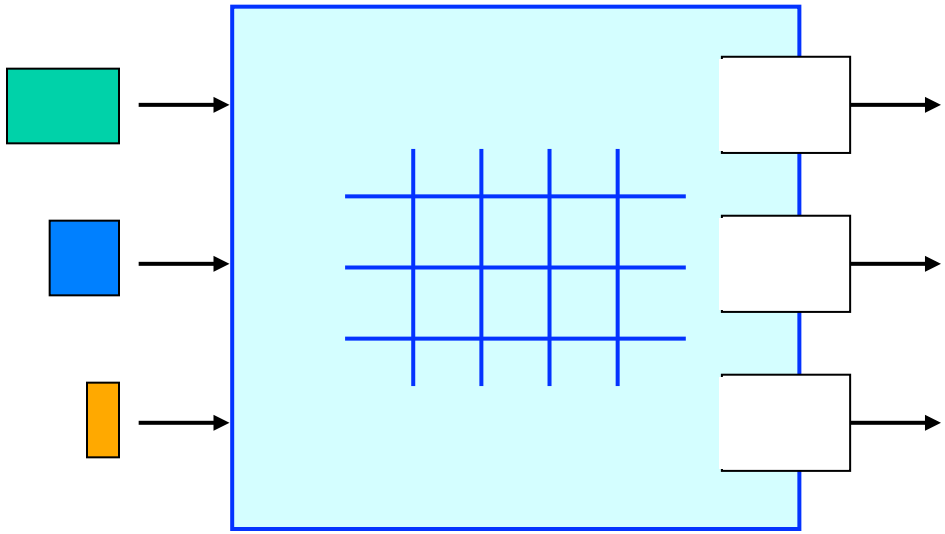
# Getting a Grip on Delays in Packet Networks

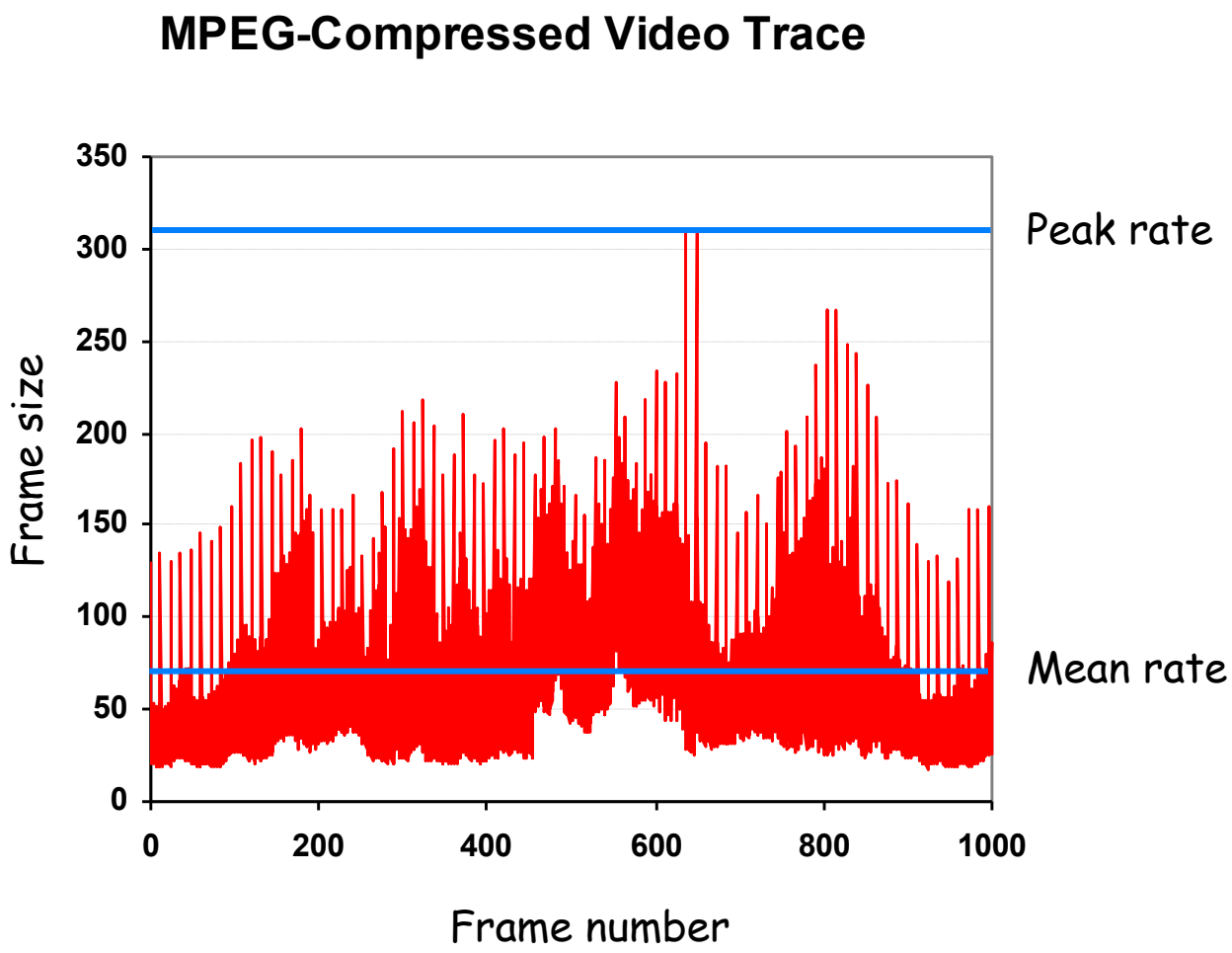
Jorg Liebeherr
Dept. of ECE
University of Toronto

# Packet Switch

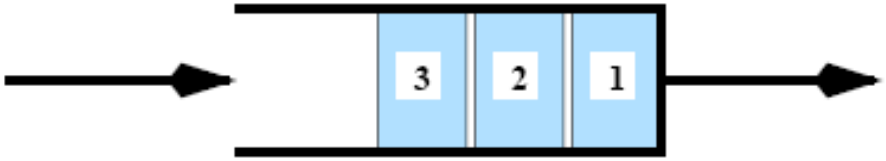


- Fixed-capacity links
- Variable delay due to waiting time in buffers
- Delay depends on
  1. Traffic
  2. Scheduling

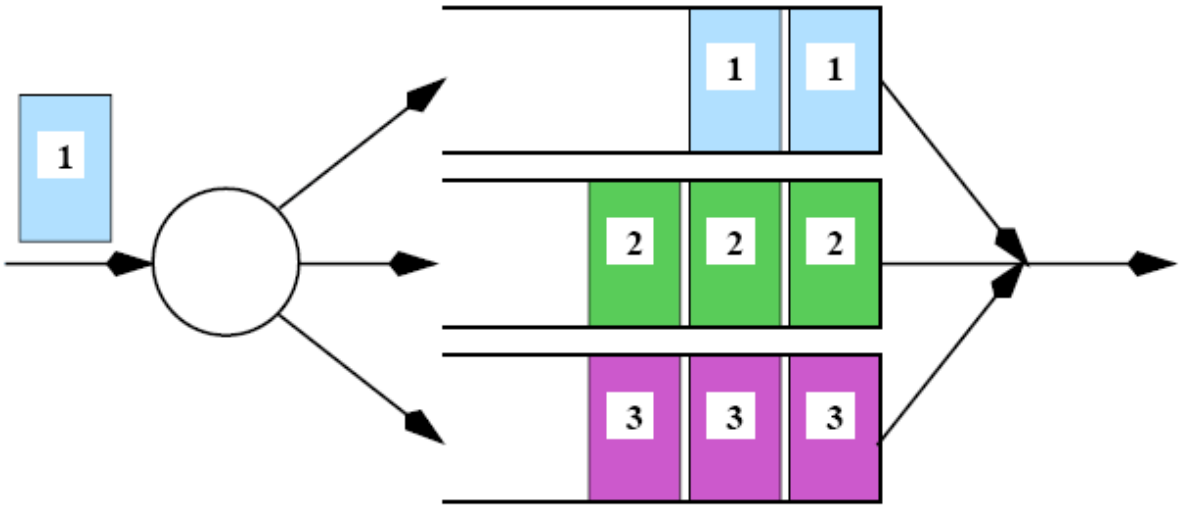
# Traffic Arrivals



# First-In-First-Out

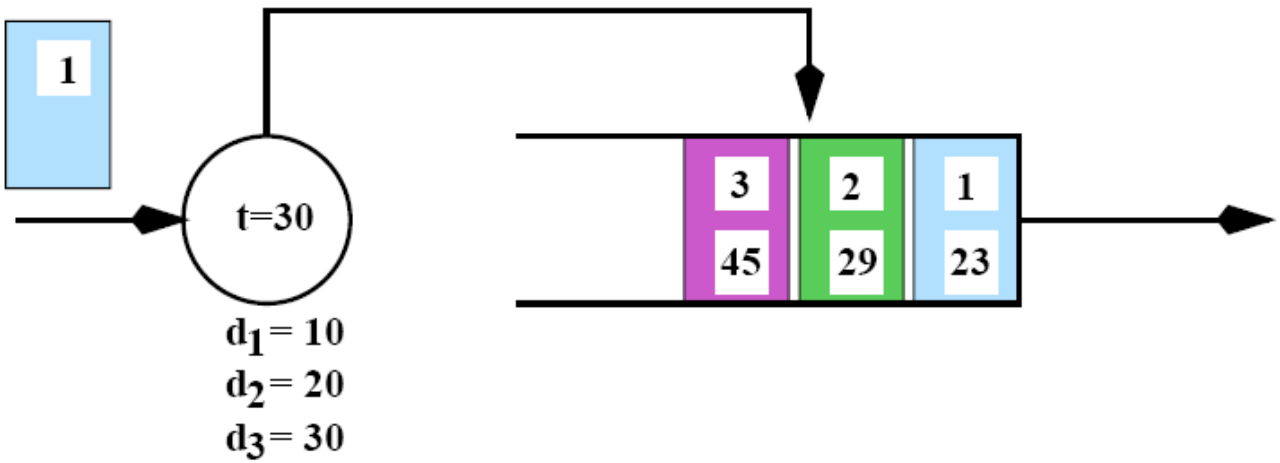


# Static Priority (SP)



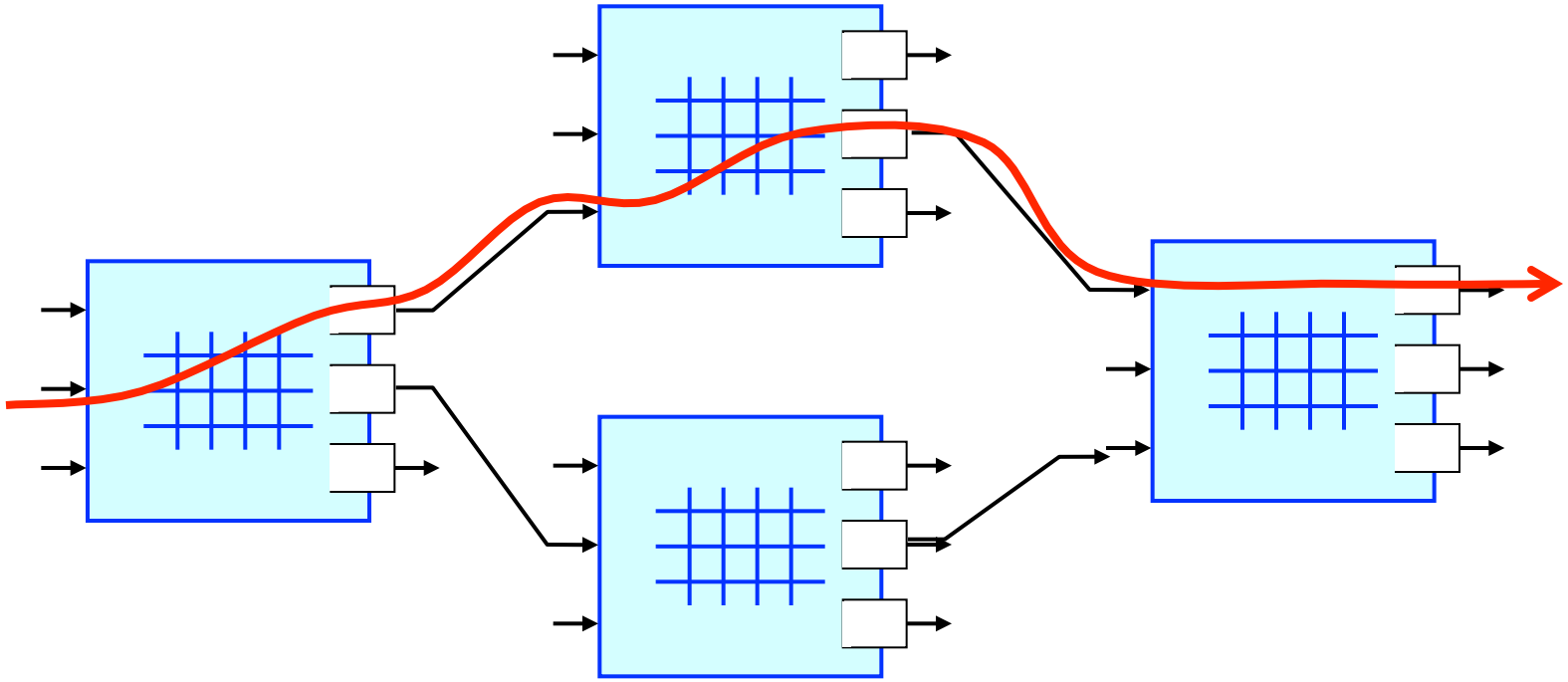

- **Blind Multiplexing (BMux):**
  All "other traffic" has higher priority

# Earliest Deadline First (EDF)

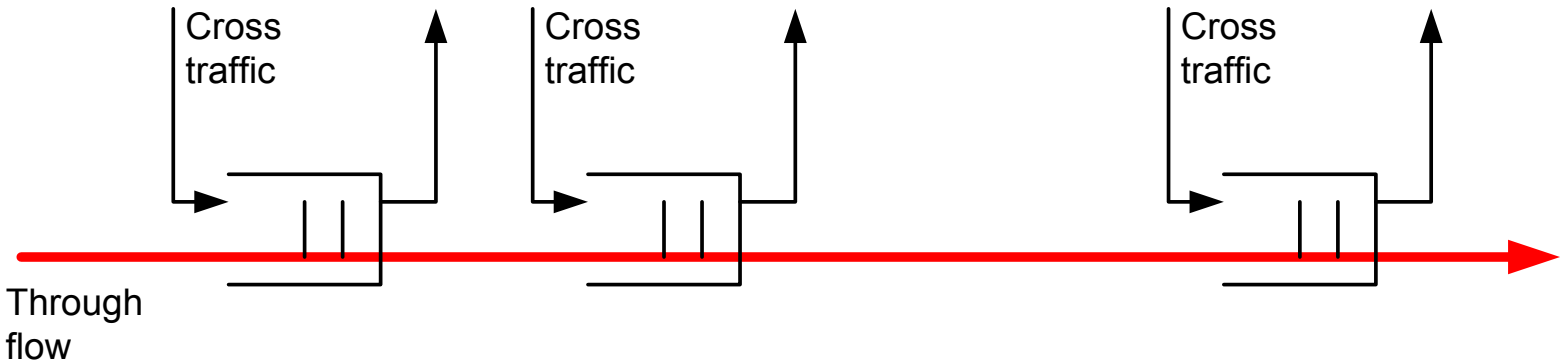



Benchmark scheduling algorithm for meeting delay requirements

# Network



# Simplified Network




Computing delays in such networks is notoriously hard …

# … but tempting

Over the last 20+ years, I have worked on problems relating to network delays:

- Worst-case delays
- Scheduling vs. statistical multiplexing
- Statistical bounds on end-to-end delays
- Difficult traffic types
- Scaling laws

# Collaborators

- Domenico Ferrari
- Dallas Wrege
- Hui Zhang
- Ed Knightly
- Almut Burchard
- Robert Boorstyn
- Chaiwat Oottamakorn
- Stephen Patek
- Chengzhi Li
- Florin Ciucu
- Yashar Ghiassi-Farrokhfal

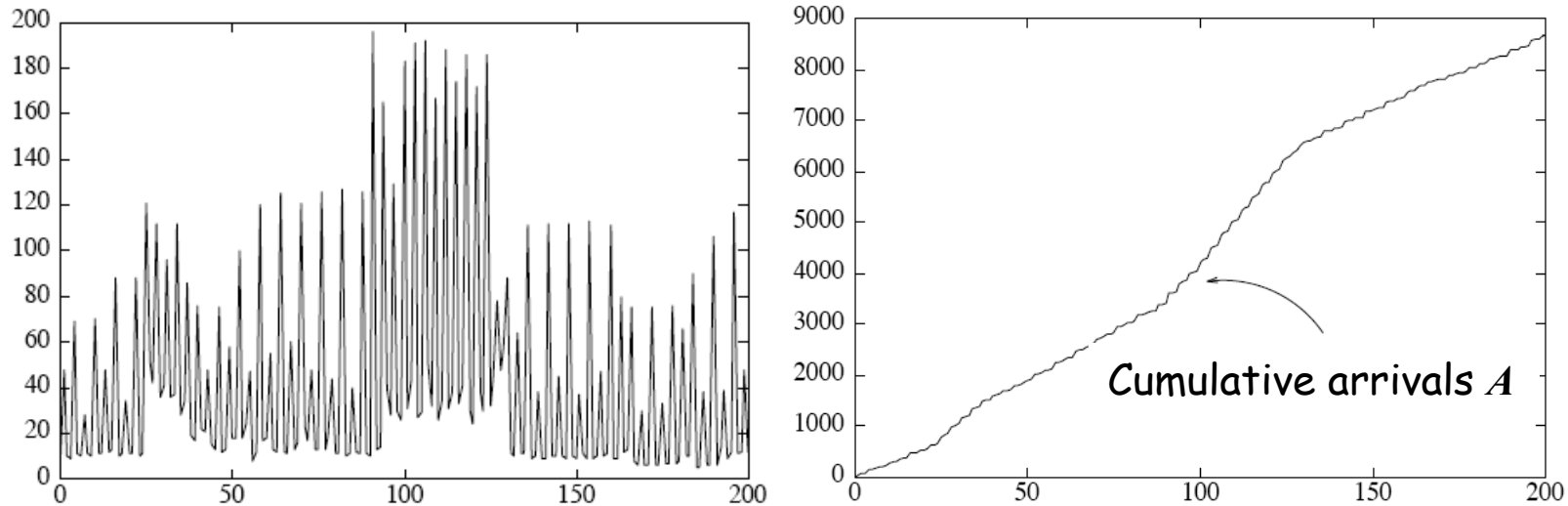
# Papers *(relevant to this talk)*


- J.Liebeherr, D. E. Wrege, D. Ferrari, "Exact admission control for networks with a bounded delay service," ACM/IEEE Trans. Netw. 4(6), 1996.
- E. W. Knightly, D. E. Wrege, H. Zhang, J. Liebeherr, "Fundamental Limits and Tradeoffs of Providing Deterministic Guarantees to VBR Video Traffic," ACM Sigmetrics, 1995.
- R. Boorstyn, A. Burchard, J. Liebeherr, C. Oottamakorn. "Statistical Service Assurances for Packet Scheduling Algorithms", IEEE JSAC, Dec. 2000.
- A. Burchard, J. Liebeherr, S. D. Patek, "A Min-Plus Calculus for End-to-end Statistical Service Guarantees," IEEE Trans. on IT, Sep. 2006.
- F. Ciucu, A. Burchard, J. Liebeherr, "A Network Service Curve Approach for the Stochastic Analysis of Networks", ACM Sigmetrics 2005.
- C. Li, A. Burchard, J. Liebeherr, "A Network Calculus with Effective Bandwidth," ACM/IEEE Trans. on Networking, Dec. 2007.
- J. Liebeherr, Y. Ghiassi-Farrokhfal, A. Burchard, "On the Impact of Link Scheduling on End-to-End Delays in Large Networks," IEEE JSAC, May 2011.
- J. Liebeherr, Y. Ghiassi-Farrokhfal, A. Burchard, "The Impact of Link Scheduling on Long Paths: Statistical Analysis and Optimal Bounds", INFOCOM 2011.
- A. Burchard, J. Liebeherr, F. Ciucu, "On Superlinear Scaling of Network Delays, ACM/IEEE Trans. Netw., August 2011
- J. Liebeherr, A. Burchard, F. Ciucu, "Delay Bounds in Communication Networks with Heavy-Tailed and Self-Similar Traffic ," IEEE Trans. on IT, Feb. 2012.

# Disclaimer

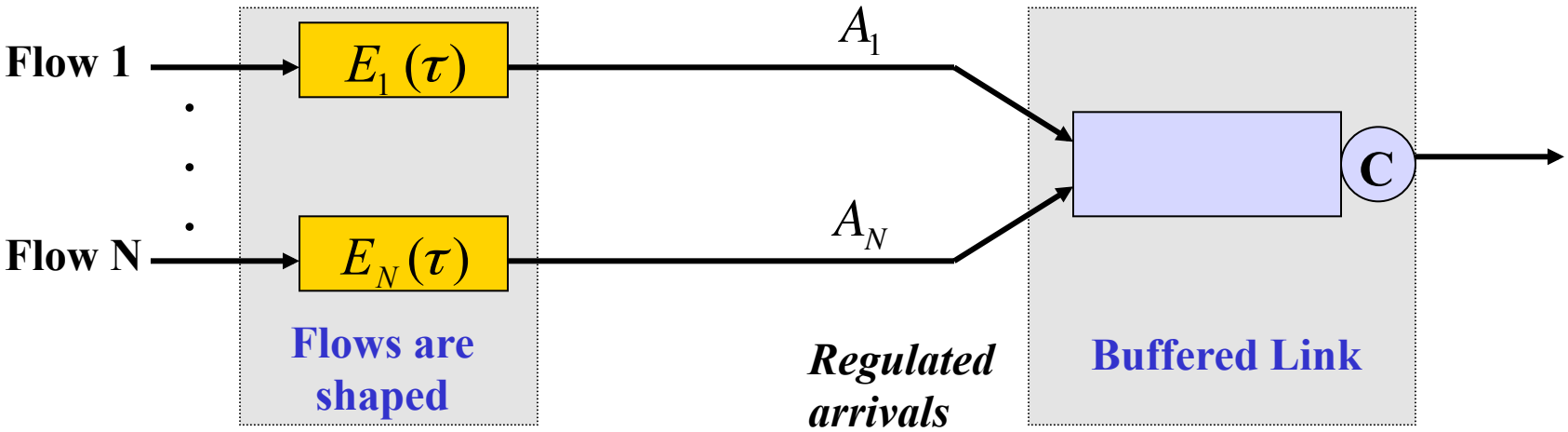
- This talk makes a few simplifications
- Please see papers for complete details

# Traffic Description




- Traffic arrivals in time interval [s,t) is $A(s,t)$
- Burstiness can be reduced by "shaping" traffic

# Shaped Arrivals

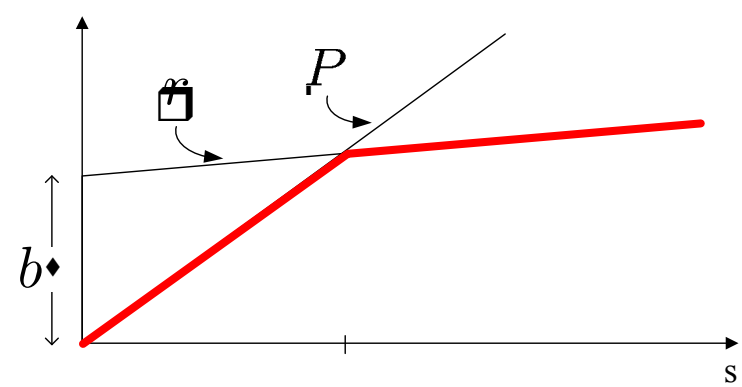



Traffic $A_j$ is shaped by an envelope $E_j$ such that:

$$E(t-s) \geq \sup_{s \leq t} \{A(s,t)\}$$

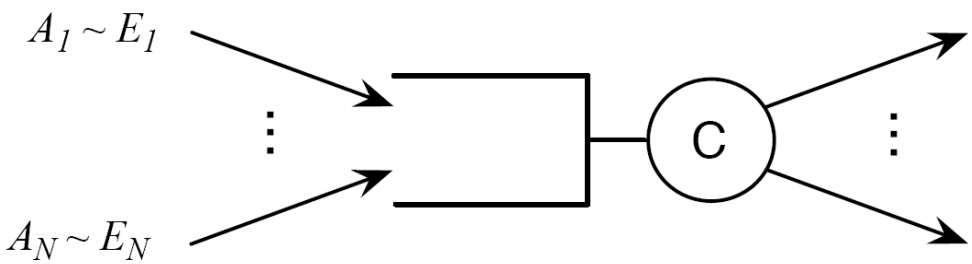
Popular envelope: "token bucket"

$$E(s) = \min(Ps, b + rt)$$



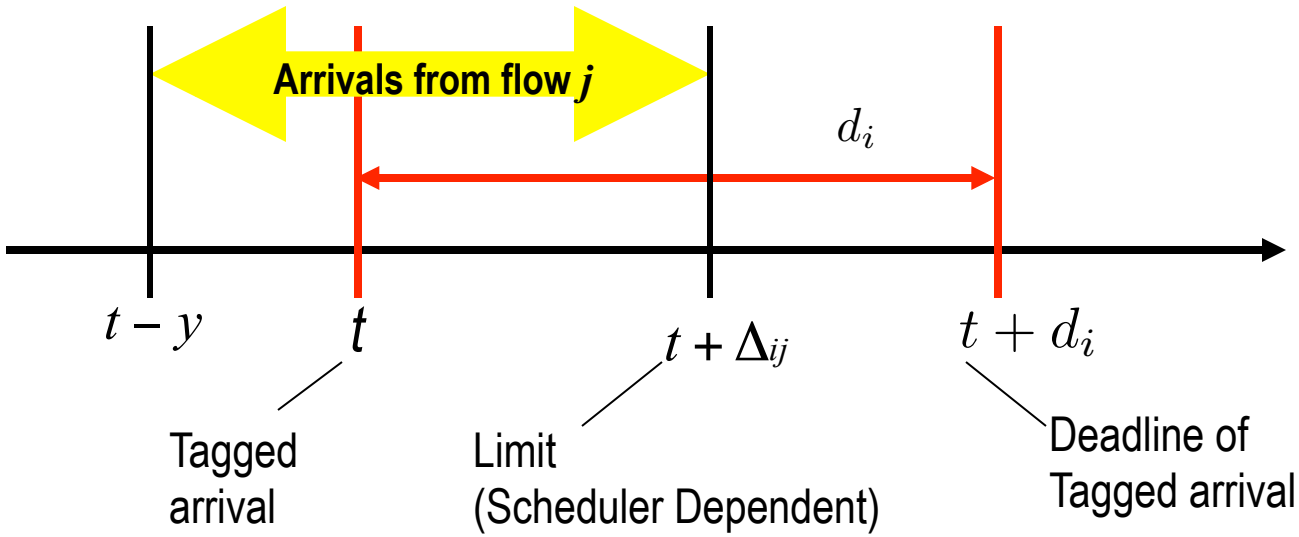
What is the maximum number of shaped flows with delay requirements that can be put on a single buffered link?

- Link capacity $C$
- Each flows j has
  - arrival function $A_j$
  - envelope $E_j$
  - delay requirement $d_j$



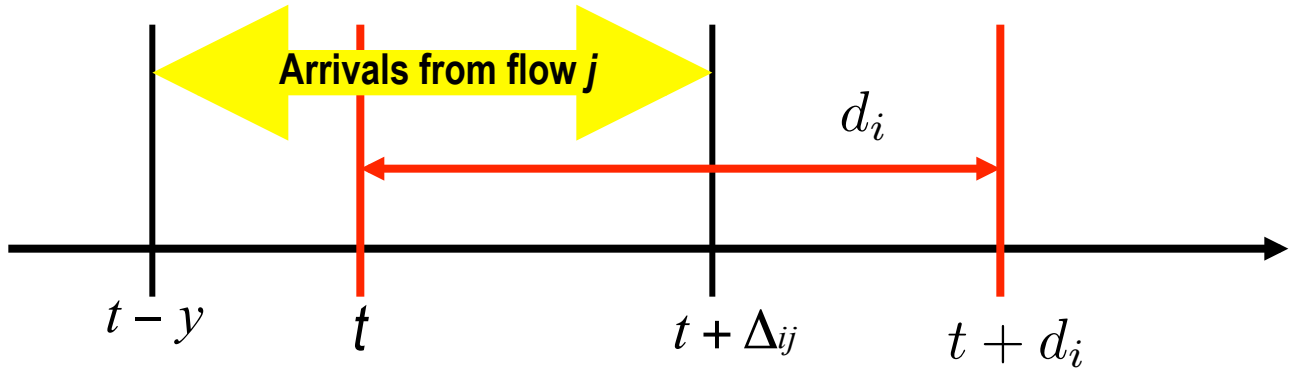
# Delay Analysis of Schedulers

- Consider a link scheduler with rate C
- Consider arrival from flow $i$ at $t$ with $t+d_i$:




$$\Delta_{ij}(x) := \min\{\Delta_{ij}, x\}$$

# Delay Analysis of Schedulers




$$d_i \geq \sup_{s \geq 0} \frac{1}{C} \left\{ \sum_j A_k(t - s, t + \Delta_{ij}(d_i)) - Cs \right\}$$

with

**FIFO**: $\Delta_{ij} = 0.$

**Static Priority**: $\Delta_{ij} = -\infty$ (lower) , $0$ (same) , $d_i$ (higher).

**EDF**: $\Delta_{ij} = d_i - d_j$

# Schedulability Condition

We have: $E_j(t-s) \geq A_j(s,t) \quad \forall s \leq t$

Therefore:

An arrival from class i **never** has a delay bound violation if

$$d_i \geq \sup_{s \geq 0} \frac{1}{C} \left\{ \sum_j E_j(s + \Delta_{ij}(d_i) - Cs \right\}$$

Condition is tight, when $E_j$ is concave

# Plugging in ...

Let: $E_j(t) = b_j + r_j t$

FIFO

$$d_j \geq \frac{1}{C} \sum_j b_j .$$

SP

$$d_p \geq \frac{\sum_{q=p}^{P} b_p}{C - \sum_{q=p+1}^{P} r_q}$$

EDF

$$d_j \geq \frac{\sum_{k=1}^{j} b_k - \sum_{k=1}^{j-1} r_k d_k}{C - \sum_{k=1}^{j-1} r_k}$$

# Numerical Result

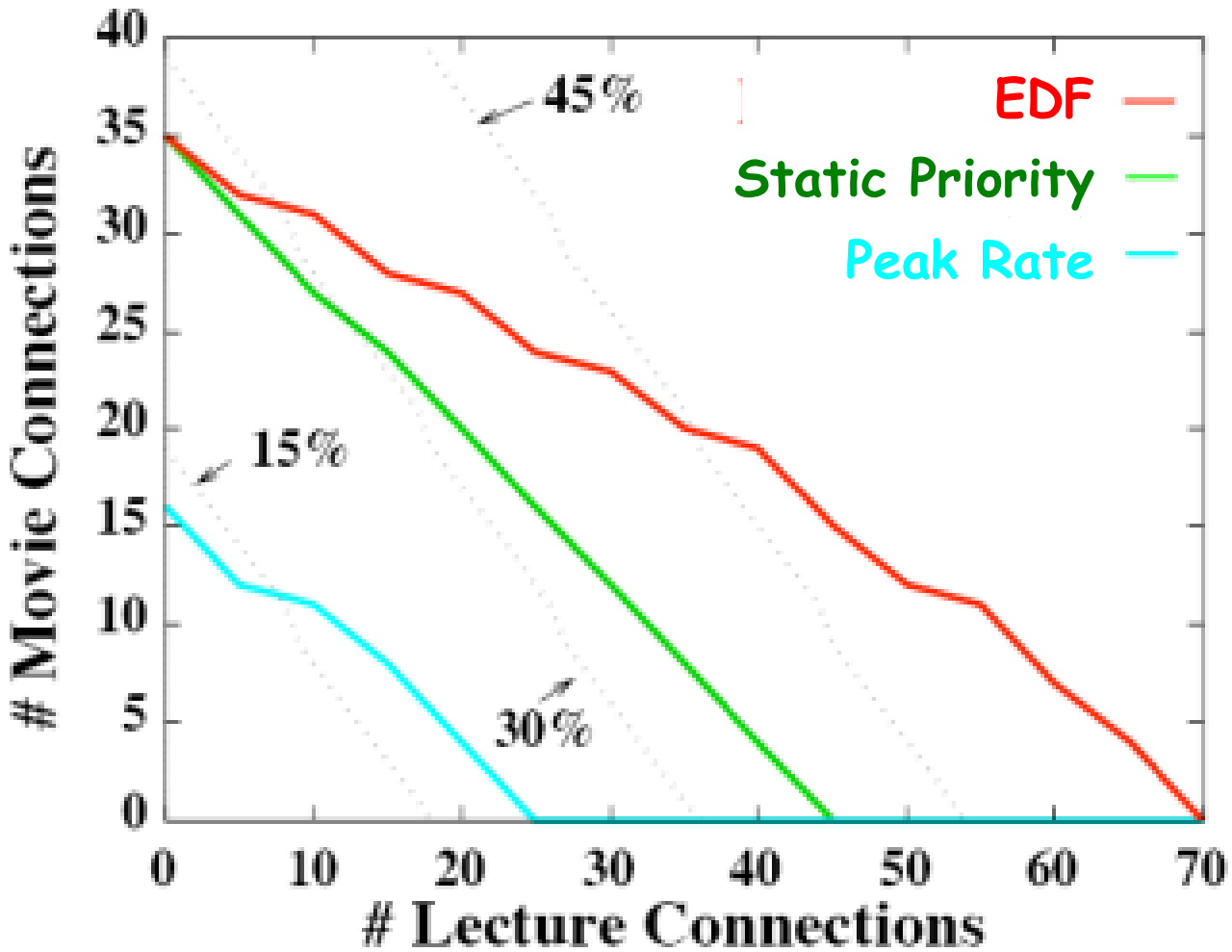
*(Sigmetrics 1995)*

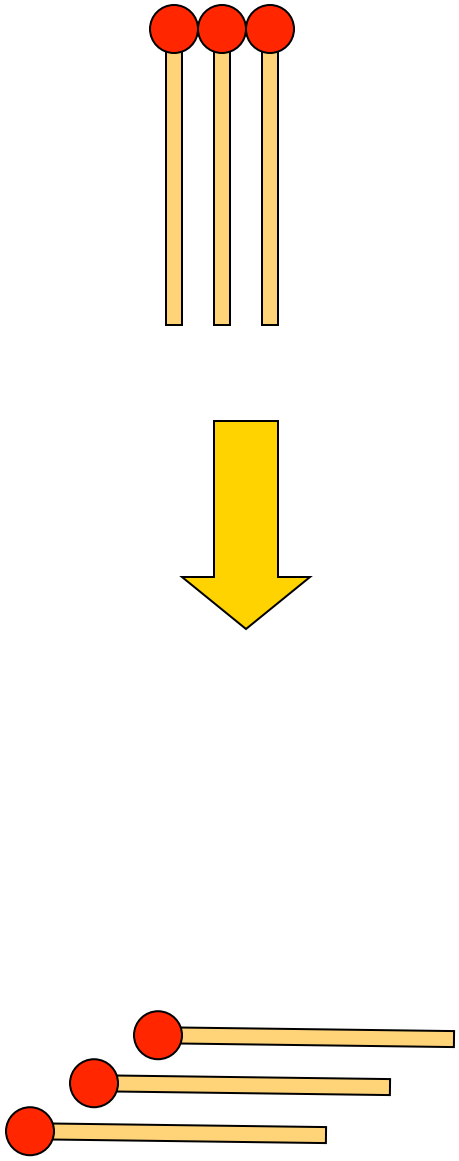
**C = 45 Mbps**

**MPEG 1 traces:**

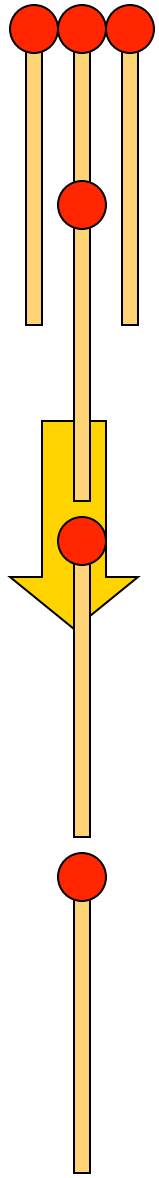
Lecture:
d = 30 msec

Movie
(Jurassic Park):
d = 50 msec

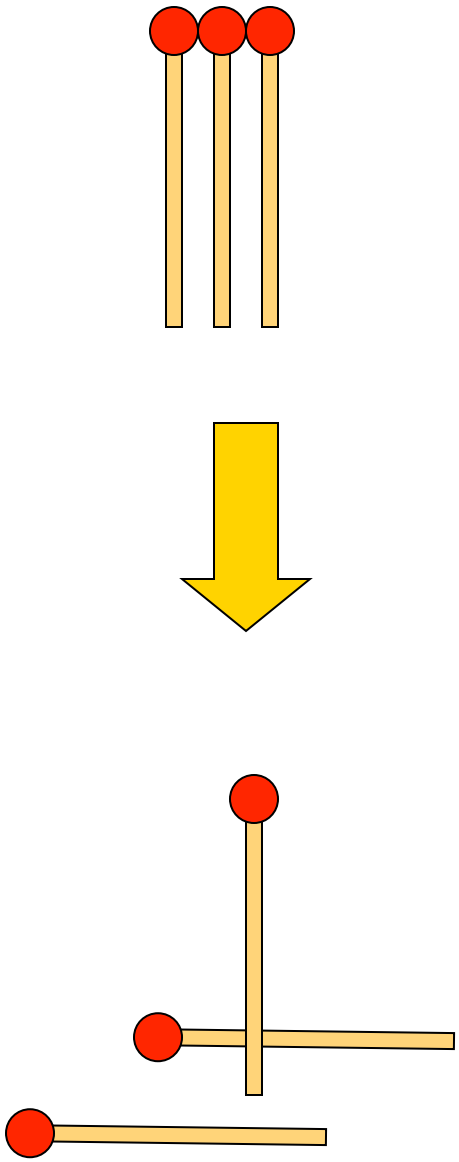




Expected case



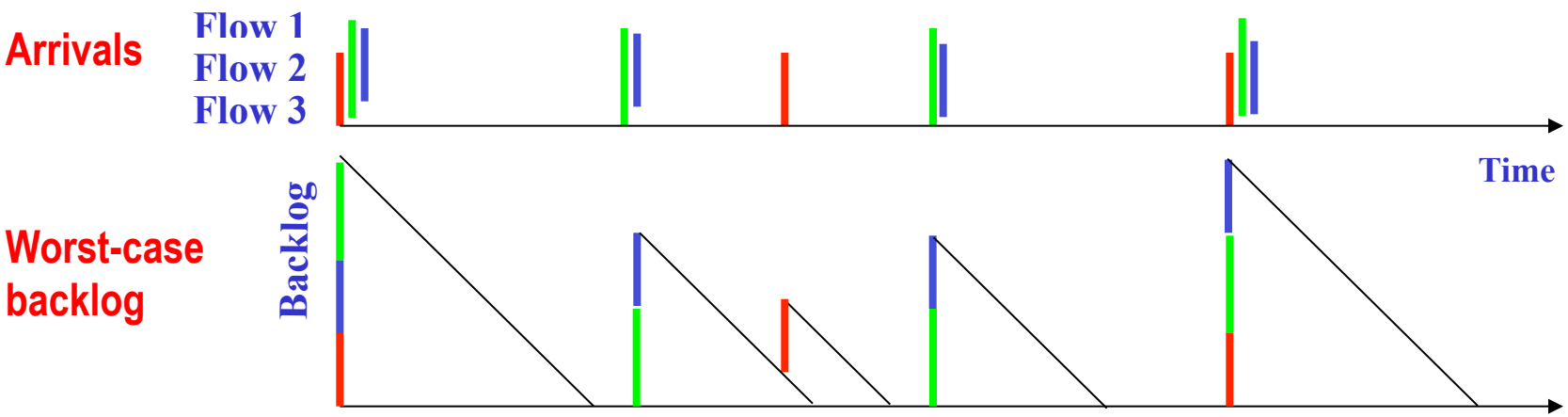
Deterministic worst-case



Probable worst-case

# Statistical Multiplexing Gain

Worst-case arrivals



# Statistical Multiplexing Gain

$$\begin{pmatrix}\text{Resources needed}\\ \text{to support}\\ \text{guarantees}\\ \text{for N flows}\end{pmatrix} << N \cdot \begin{pmatrix}\text{Resources needed}\\ \text{to support}\\ \text{guarantees}\\ \text{for 1 flow}\end{pmatrix}$$

Statistical multiplexing gain is the raison d'être for packet networks.

**What is the maximum number of flows with delay requirements that can be put on a buffered link and considering statistical multiplexing?**

Arrivals $A_j(s,t)$ are random processes

- Stationarity: $A_j$ is stationary random processes
- Independence: Any two flows $A_i$ and $A_j (i \neq j)$ are stochastically independent

# Envelopes for random arrivals

Statistical envelope bounds arrival from flow j with high certainty

- **Statistical envelope $\mathcal{G}$ :**

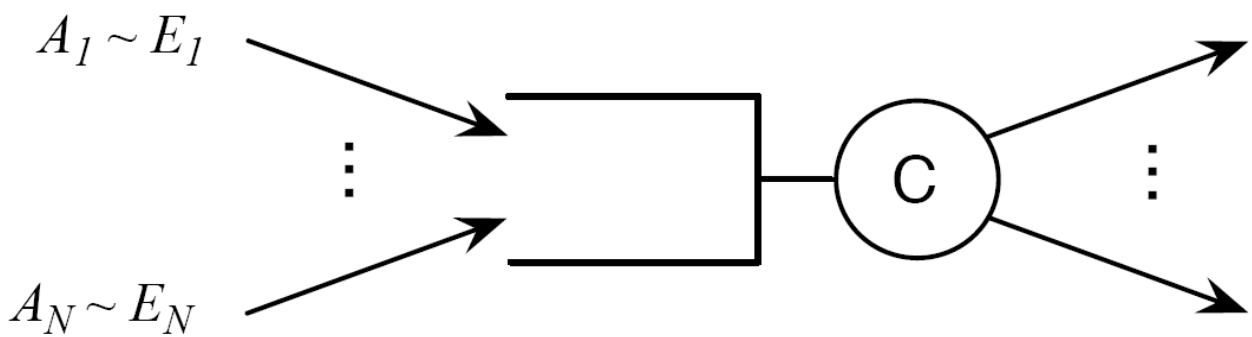
$$Pr\{A(s,t) > \mathcal{G}(t-s) + \sigma\} < \varepsilon(\sigma) \quad \forall s,t$$

- **Statistical sample path envelope $\mathcal{H}$ :**

$$Pr\{\sup_{s \le t} \{A(s,t) - \mathcal{H}(t-s)\} > \sigma\} < \varepsilon(\sigma)$$

Statistical envelopes are non-random functions

# Aggregating arrivals




Arrivals from group of flows: $A_{\mathcal{C}} = \sum_j A_j$

**with deterministic envelopes:** $E_{\mathcal{C}} = \sum_j E_j$

**with statistical envelopes:** $\mathcal{G}_{\mathcal{C}} \ll \sum_j \mathcal{G}_j \ll E_{\mathcal{C}}$

# Statistical envelope for group of indepenent (shaped) flows

- Exploit independence and extract statistical multiplexing gain when calculating $\mathcal{G}_{\mathcal{C}}$

- For example, using the Chernoff Bound, we can obtain

$$\boxed{\mathcal{G}_{\mathcal{C}}(t) \;=\; \inf_{s>0} \frac{1}{s}\Big(\sum_{j\in\mathcal{C}} \log \overline{M}_j(s,t) - \log \varepsilon\Big)}$$

$$\overline{M}_j(s,t) \;=\; 1 + \frac{\rho_j\, t}{E_j(t)}\big(e^{sE_j(t)} - 1\big)$$

$$\rho_j = \lim_{\tau\to\infty} E_j(\tau)/\tau$$

# Statistical Envelope vs. Deterministic Envelopes

*(JSAC 2000)*

$$E(t) = \min(Pt, \sigma + \rho t)$$

**Type 1 flows:**

P =1.5 Mbps

ρ = .15 Mbps

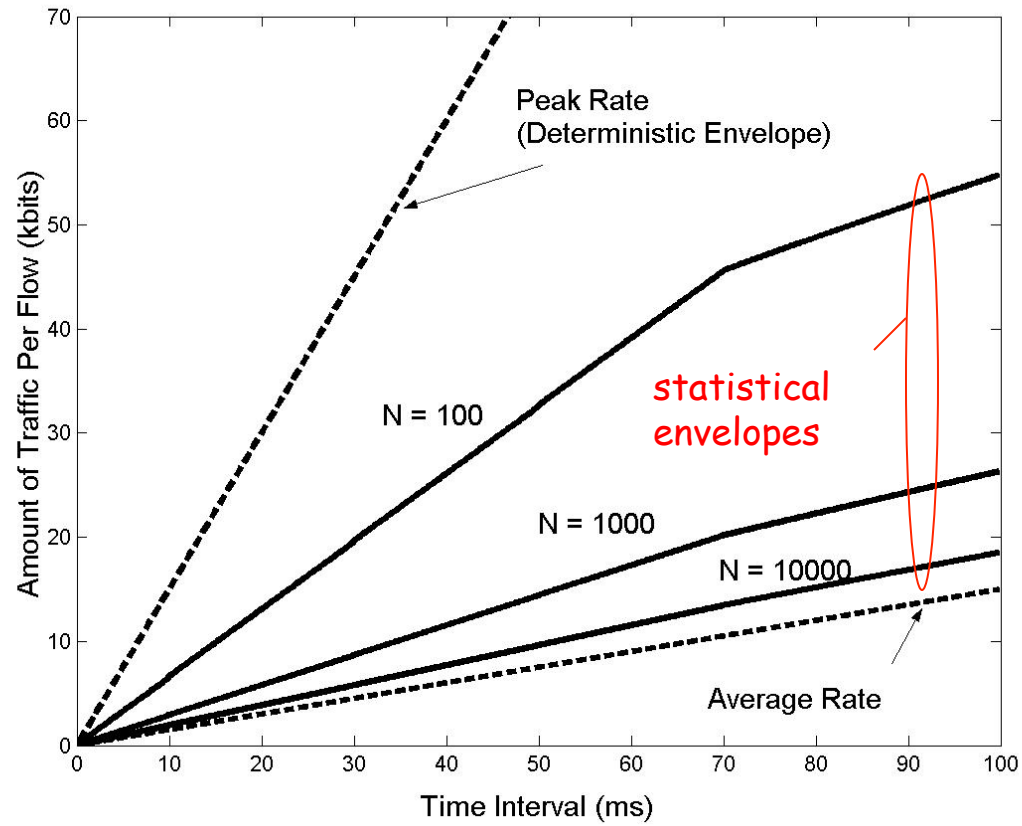
σ =95400 bits

**Type 2 flows:**

P = 6 Mbps

ρ = .15 Mbps

σ = 10345 bits



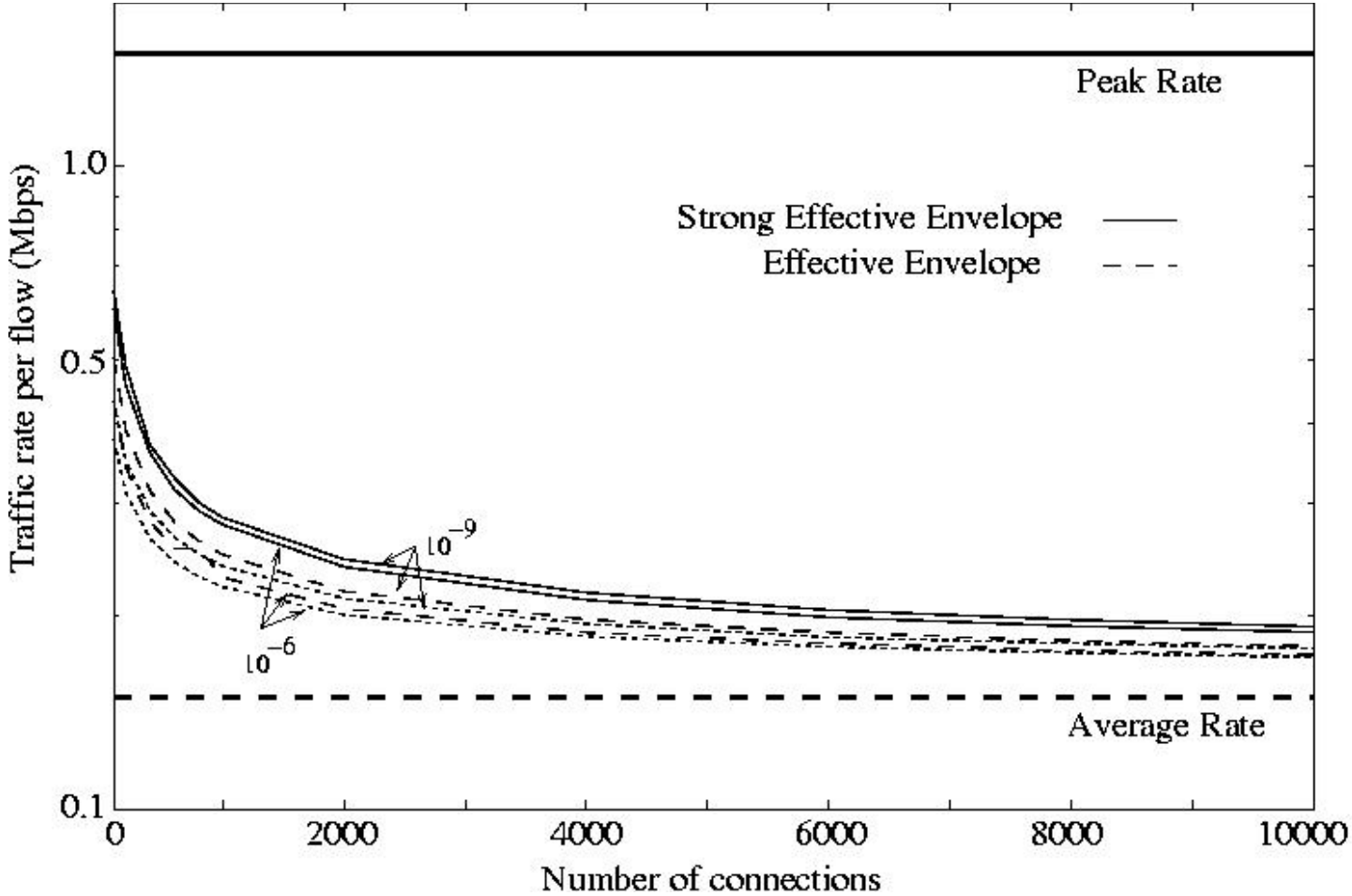

**Type 1 flows**

$$\varepsilon = 10^{-6}$$

# Statistical vs. Deterministic Envelope Envelopes

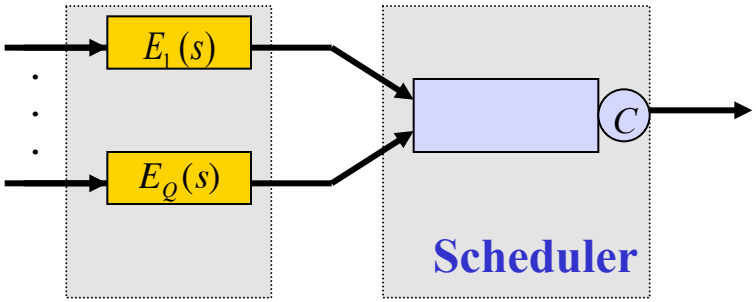
*(JSAC 2000)*

**Traffic rate at t = 50 ms**
**Type 1 flows**



# Scheduling Algorithms

- Work-conserving scheduler that serves Q classes
- Class-q has delay bound $d_q$
- $\Delta$-scheduling algorithm




**Deterministic Service**

Never a delay bound violation if:

$$\sup_s \left\{ \sum_p E_{\mathcal{C}_p}(s + \Delta_{qp}) - Cs \right\} \leq C d_q$$

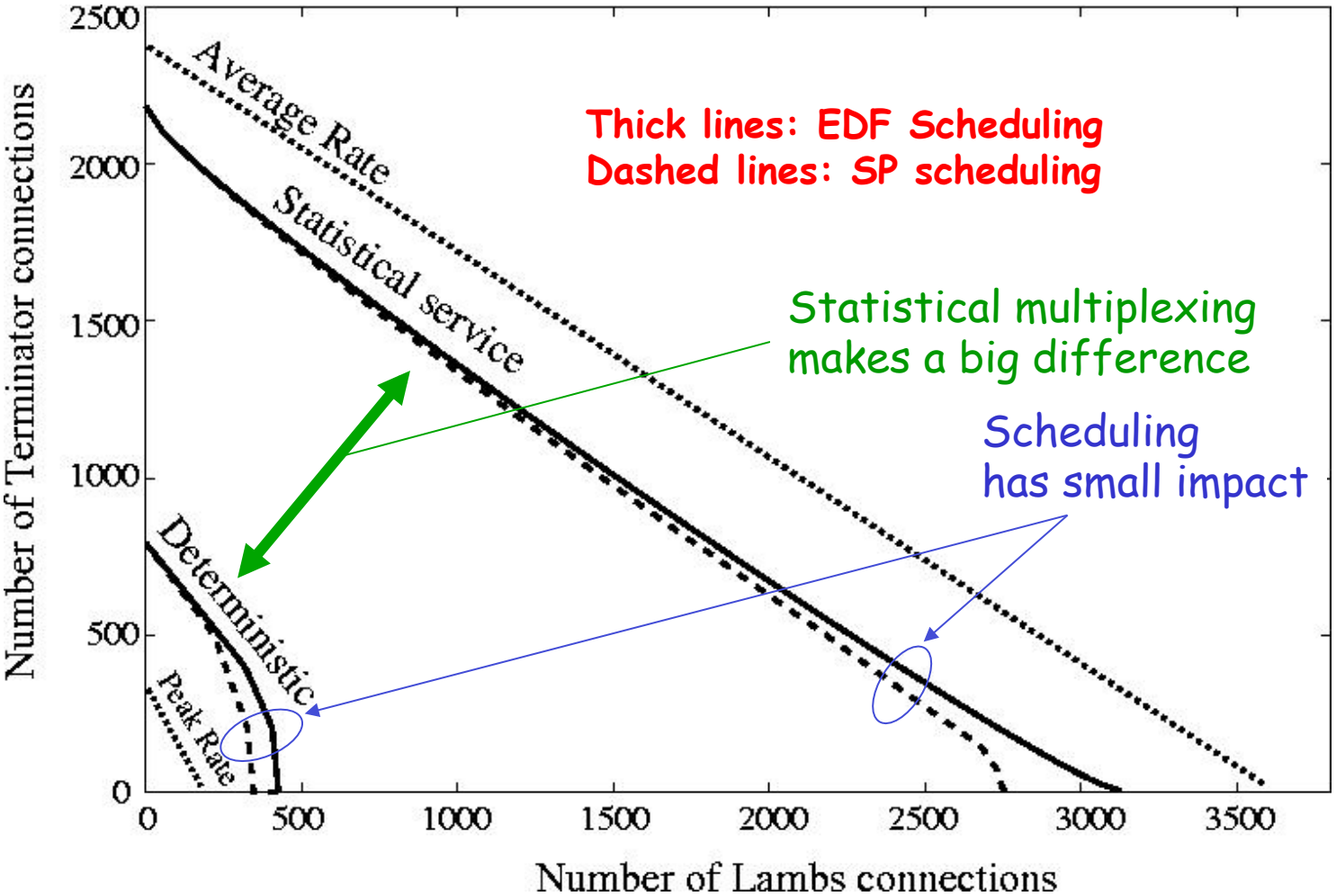
**Statistical Service**

Delay bound violation with $\varepsilon$ if:

$$\sup_s \left\{ \sum_p \mathcal{H}_{\mathcal{C}_p}(s + \Delta_{qp}) - Cs \right\} \leq C d_q$$

# Statistical Multiplexing vs. Scheduling *(JSAC 2000)*

**Example: MPEG videos with delay constraints at C= 622 Mbps**
**Deterministic service vs. statistical service ($\varepsilon = 10^{-6}$)**



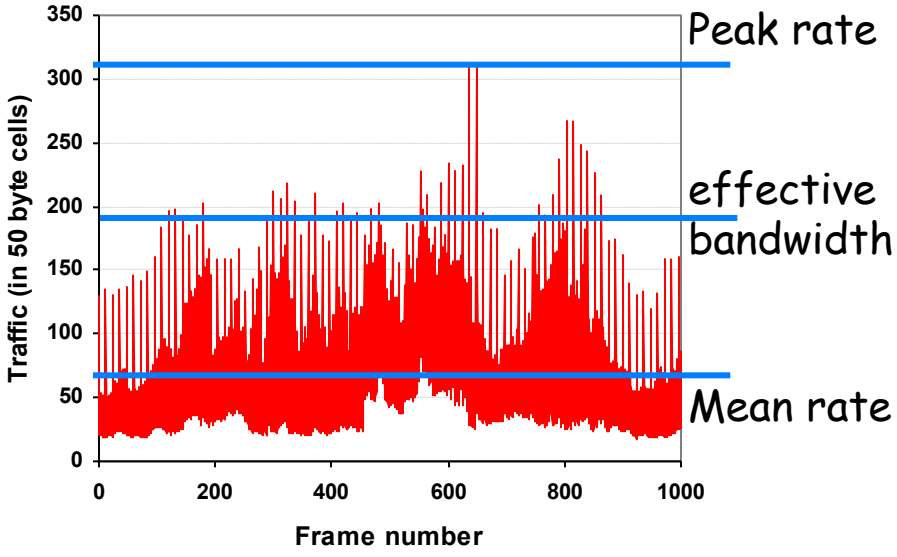

$d_{terminator}$=100 ms
$d_{lamb}$=10 ms

# More interesting traffic types

- **So far:** Traffic of each flow was shaped
- **Next:**
  - On-Off traffic
  - Fraction Brownian Motion (FBM) traffic

**Approach:**

- Exploit literature on Effective Bandwidth
- Derived for many traffic types



# Statistical Envelopes and Effective Bandwidth

Effective Bandwidth (Kelly 1996)

$$\alpha(s,\tau) = \sup_{t \geq 0} \left\{ \frac{1}{s\tau} \log E[e^{s(A(t+\tau)-A(t))}] \right\}$$

$$s, \tau \in (0, \infty)$$

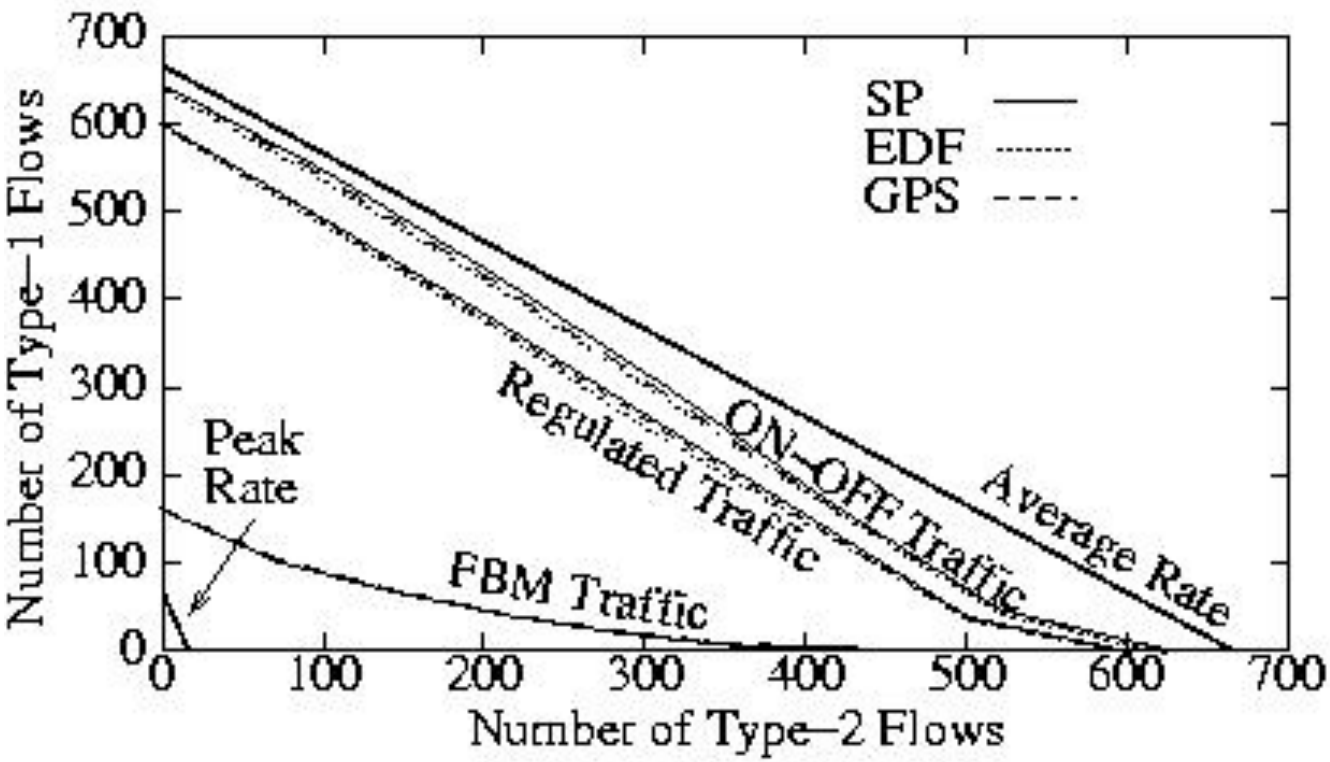
Given $\alpha(s,\tau)$, an effective envelope is given by

$$\mathcal{G}^{\varepsilon}(\tau) = \inf_{s>0} \{ \tau\alpha(s,\tau) - \frac{\log \varepsilon}{s} \}$$

# Different Traffic Types

*(ToN 2007)*

Comparisons of statistical service guarantees for different schedulers and traffic types




**Schedulers:**

**SP-** Static Priority
**EDF** – Earliest Deadline First
**GPS** – Generalized Processor Sharing

**Traffic:**

**Regulated** – leaky bucket
**On-Off** – On-off source
**FBM** – Fractional Brownian Motion

**$C$= 100 Mbps, $\varepsilon = 10^{-6}$**

# Delays on a path with multiple nodes:

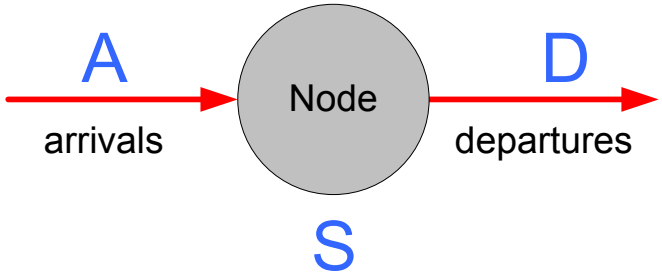
- Impact of Statistical Multiplexing
- Role of Scheduling

- How do delays scale with path length?
- Does scheduling still matter in a large network?

# Deterministic Network Calculus (1/3)

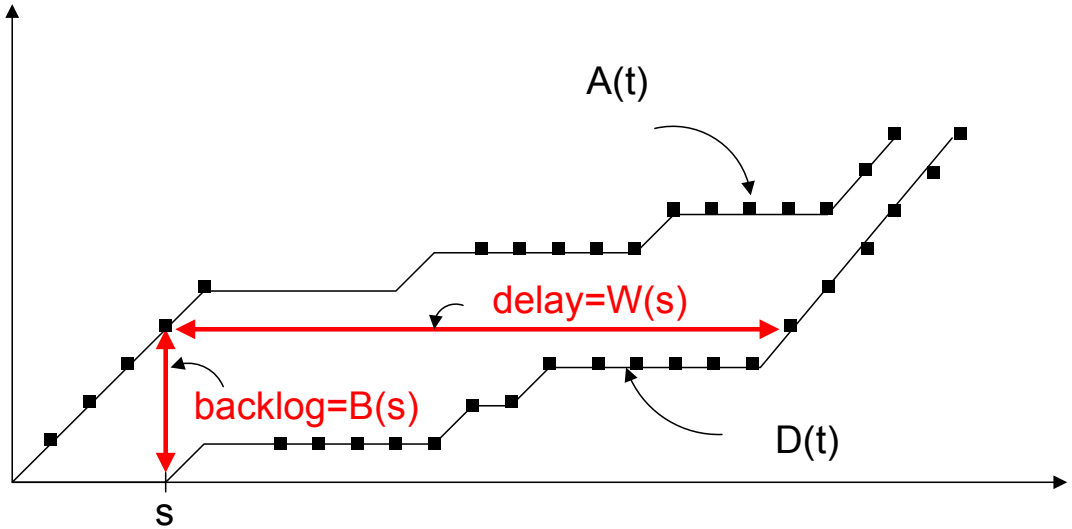
- Systems theory for networks in (min,+) algebra

  *developed by*
  *Rene Cruz, C. S. Chang, JY LeBoudec (1990's)*



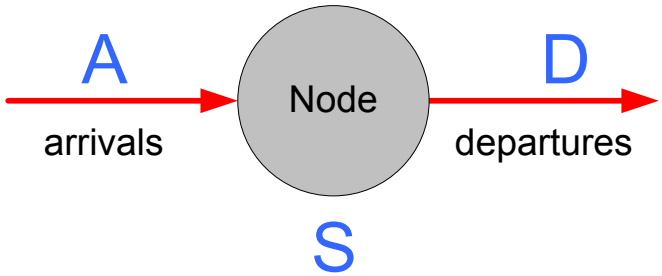

- Service curve *S* characterizes node

- Used to obtain worst-case bounds on delay and backlog

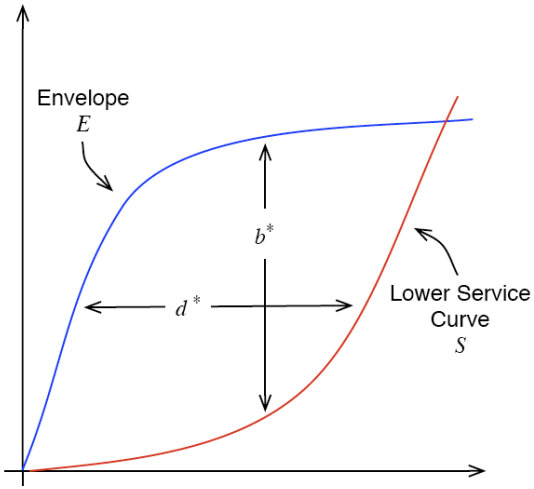


# Deterministic Network Calculus (2/3)

- Worst-case view of
  - arrivals: $A(s,t) \leq E(t-s)$
  - service : $D(t) \geq A * S(t)$




- Implies worst-case bounds
  - backlog: $B(t) \leq E \oslash S(0)$
  - delay : $W(t) \leq \inf\{d | E(s) \leq S(s +$




- (min,+) algebra operators
  - Convolution:

$$f * g(t) = \inf_{0 \leq s \leq t} (f(s) + g(t-s))$$

  - Deconvolution:

$$f \oslash g(t) = \sup_{s \geq 0} (f(t+s) - g(s))$$

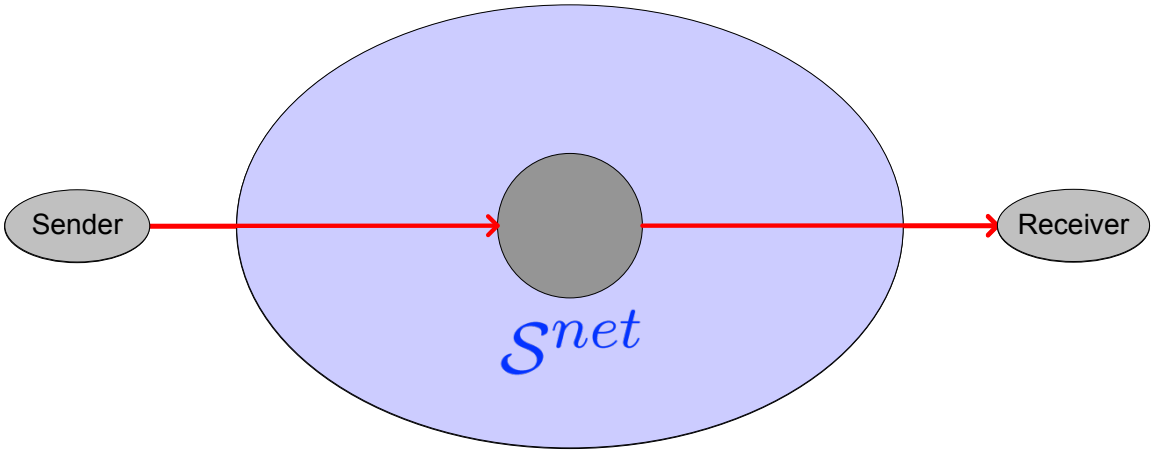
# Deterministic Network Calculus (3/3)

- Main result:

  If $\mathcal{S}^1, \mathcal{S}^2, \mathcal{S}^3$ describes the service at each node, then

$$\mathcal{S}^{net} = \mathcal{S}^1 * \mathcal{S}^2 * \mathcal{S}^3$$

  describes the service given by the network as a whole.



# Stochastic Network Calculus

- Probabilistic view on arrivals and service
  - Statistical Sample Path Envelope

$$Pr\{\sup_{s \leq t}(A(s,t) - \mathcal{H}(t-s)) > \sigma\} \leq \varepsilon(\sigma)$$

  - Statistical Service Curve

$$Pr\{D(t) - A * \mathcal{S}(t) > \sigma\} \leq \varepsilon(\sigma)$$

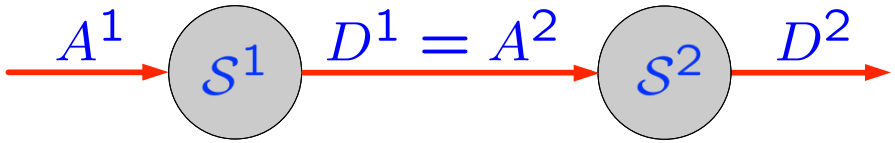
- Results on performance bounds carry over, e.g.:
  - Backlog Bound

$$Pr(B(t) > \mathcal{H} \oslash \mathcal{S}(0)) \leq \varepsilon$$

# Stochastic Network Calculus

- <u>Hard problem</u>: Find $\mathcal{S}^{net}$ so that $\mathcal{S}^{net} = \mathcal{S}^1 * \mathcal{S}^2 * \ldots * \mathcal{S}^H$

- Technical difficulty:




$$
\begin{aligned}
D^2(t) &= \inf_{0 \le s \le t} \left( A^2(s) + \mathcal{S}^2(t-s) \right) \\
&= A^2(s_0) + \mathcal{S}^2(t-s_0) \\
&\neq A^1 * \mathcal{S}^1(s_0) + \mathcal{S}^2(t-s_0) \\
&\neq A^1 * \mathcal{S}^1 * \mathcal{S}^2(t)
\end{aligned}
$$

$s_0$ is a random variable!

# Statistical Network Service Curve (Sigmetrics 2005)

- Notation: $\mathcal{S}_{-\delta}(t) = \mathcal{S}(t) - \delta t$

- **Theorem**: If $\mathcal{S}^1, \mathcal{S}^2, \ldots, \mathcal{S}^H$ are statistical service curves, then for any $\delta > 0$:

$$\mathcal{S}^{net} = \mathcal{S}^1 * \mathcal{S}^2_{-\delta} * \cdots * \mathcal{S}^H_{-(H-1)\delta}$$

is a statistical network service curve with some finite violation probability.

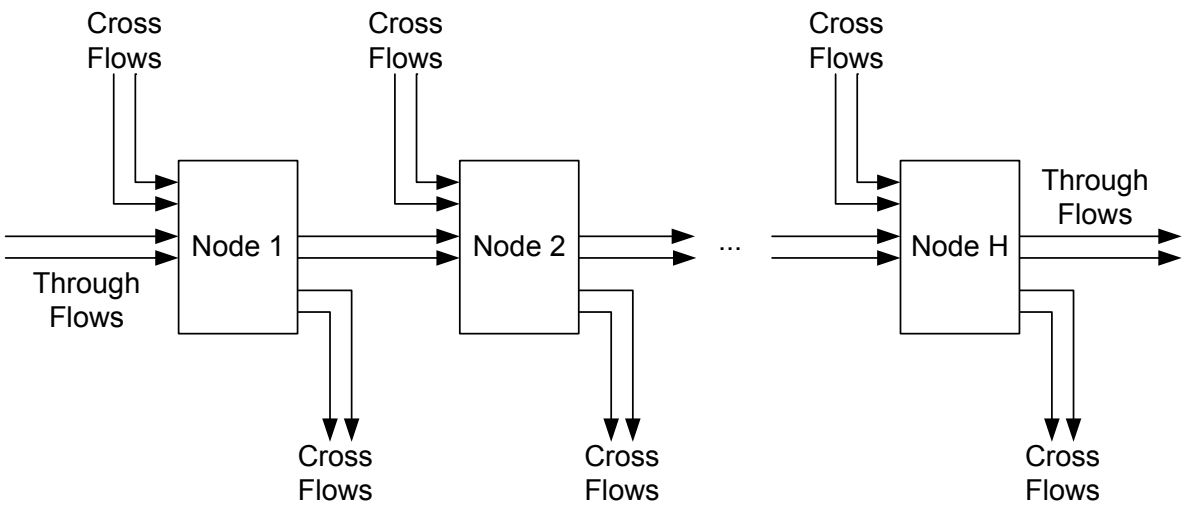
# EBB model

- Traffic with **Exponentially Bounded Burstiness (EBB)**

$$P\big(A(s,t) - \underbrace{\rho(t-s)}_{\mathcal{G}(t-s;\sigma)} > \sigma\big) \leq \underbrace{Me^{-\alpha\sigma}}_{\varepsilon(\sigma)}$$

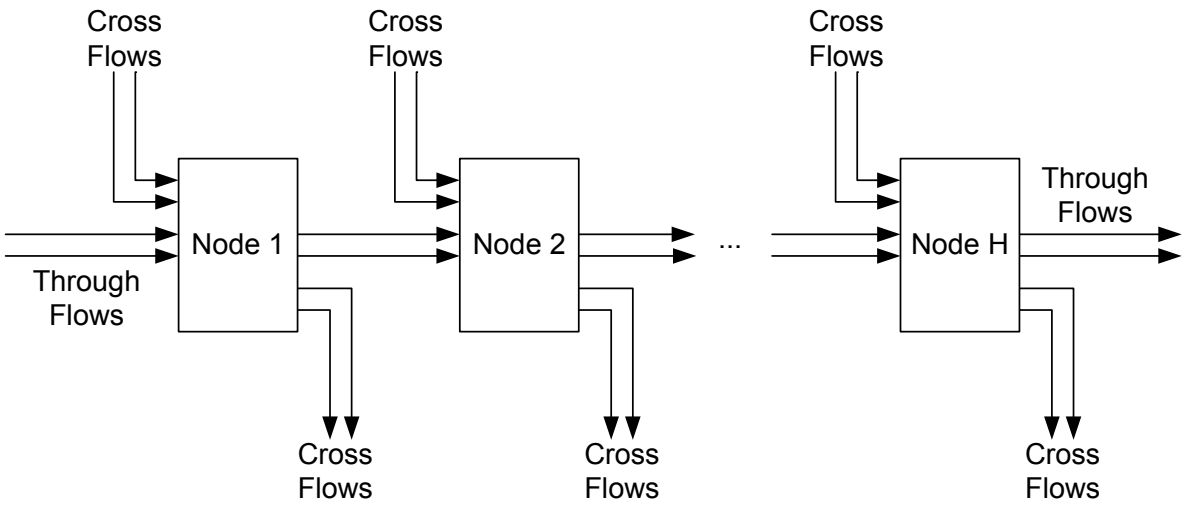
- Sample path statistical envelope obtained via union bound

# Example: Scaling of Delay Bounds




- Traffic is Markov Modulated On-Off Traffic (EBB model)
- All links have capacity C
- Same cross-traffic **(not independent!)** at each node
- Through flow has lower priority: $\mathcal{S}_j = [Ct - \mathcal{H}_c(t)]_+$

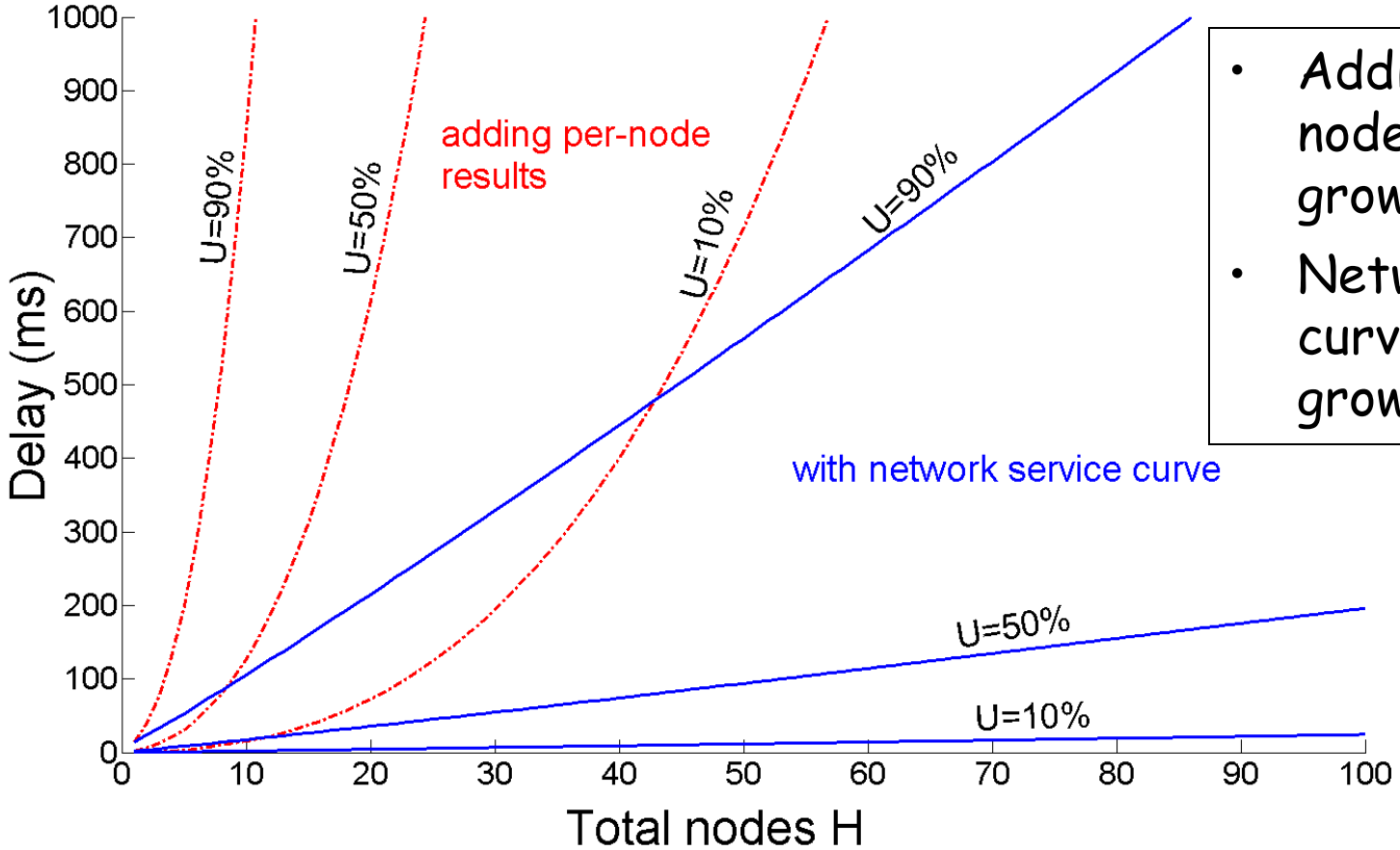
# Example: Scaling of Delay Bounds




- Two methods to compute delay bounds:
  1. **Add per-node bounds:**
     Compute delay bounds at each node and sum up
  2. **Network service curve:**
     Compute single-node delay bound with statistical network service curve

# Example: Scaling of Delay Bounds *(Sigmetrics 2005)*

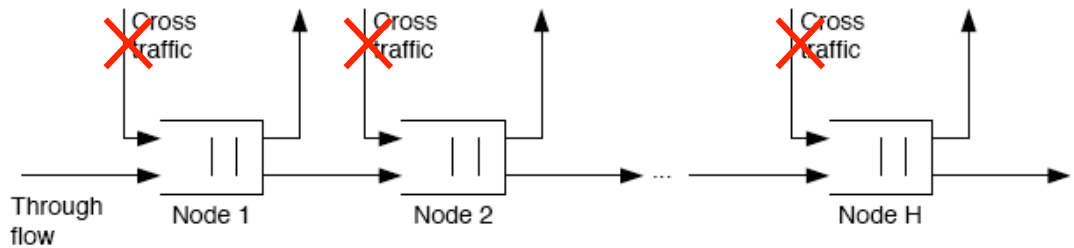
- Peak rate: P = 1.5 Mbps
  Average rate: $\rho$ = 0.15 Mbps
- T= $1/\mu + 1/\lambda$ = 10 msec
- $C$ = 100 Mbps
- Cross traffic = through traffic
- $\varepsilon = 10^{-9}$




- Addition of per-node bounds grows $O(H^3)$
- Network service curve bounds grow **$O(H \log H)$**

# Result: Lower Bound on E2E Delay *(ToN 2011)*

- M/M/1 queues with identical exponential service at each node



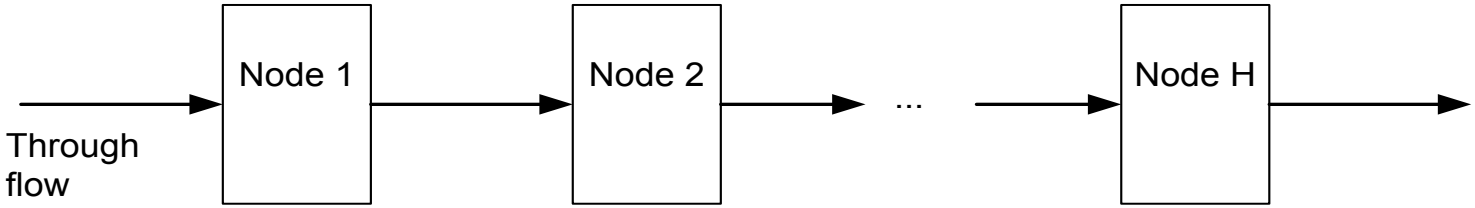

**Theorem**: E2E delay $W_H$ satisfies for all $0 < z < 1$

$$Pr\Big(W_H \leq \gamma_1 H \log(\gamma_2 H)\Big) \leq z$$

**Corollary**: $z$-quantile $w_H(z)$ of $W_H$ satisfies

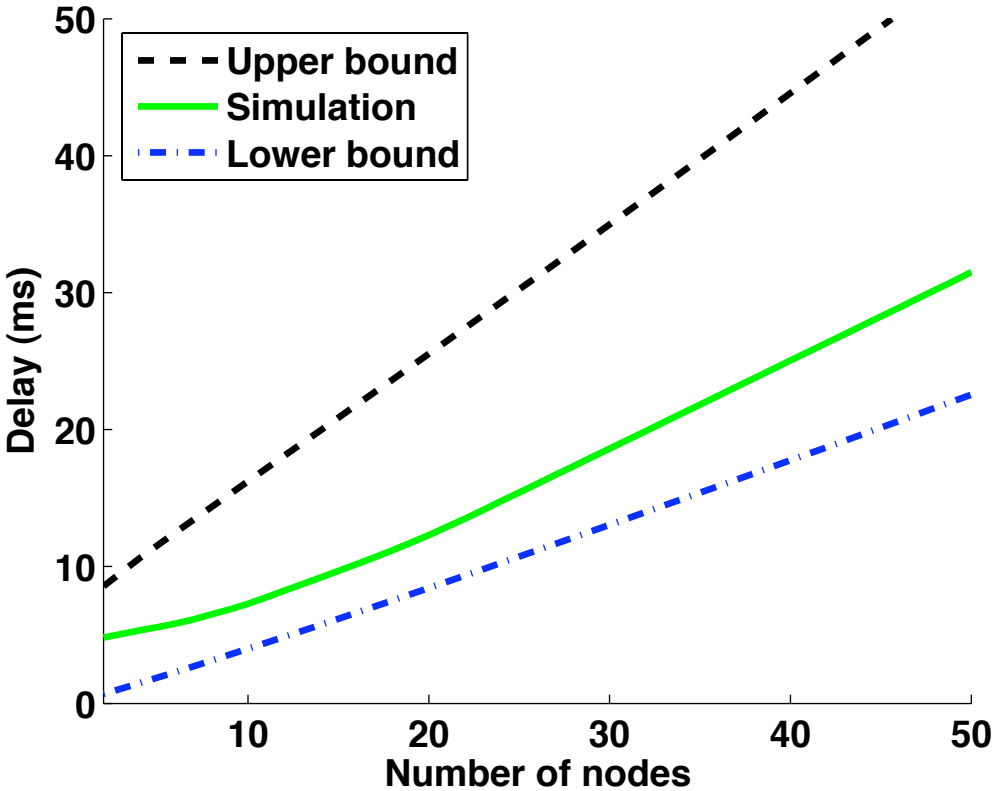
$$w_H(z) = \Omega(H \log H)$$

# Numerical examples




- Tandem network without cross traffic
- Node capacity: $C$
- Arrivals are compound Poisson process
  - Packet arrival rate: $\lambda$
  - Packet size: $Y_i \sim exp(\mu)$
- Utilization: $\rho = \lambda/(\mu C)$

# Upper and Lower Bounds on E2E Delays *(ToN 2011)*




Capacity

$C = 100\ Mbps$

Mean packet size

$\frac{1}{\mu} = 400$ Bytes

Load factor

$\rho = 90\%$

Violation probability

$\varepsilon = 10^{-6}$

# Superlinear Scaling of Network Delays

- For traffic satisfying "Exponential Bounded Burstiness", E2E delays follow a scaling law of $\Theta(H \log H)$

- This is different than predicted by
  - … worst-case analysis
  - … networks satisfying "Kleinrock's independence assumption"

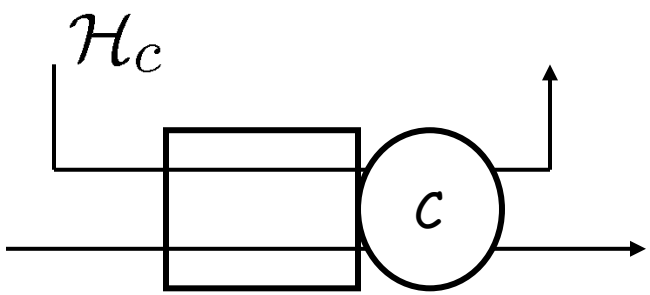
# Back to scheduling …

**So far:**

Through traffic has lowest priority and gets leftover capacity

→ **Leftover Service**

or **Blind Multiplexing**




$$\mathcal{S}_j = [Ct - \mathcal{H}_c(t)]_+$$

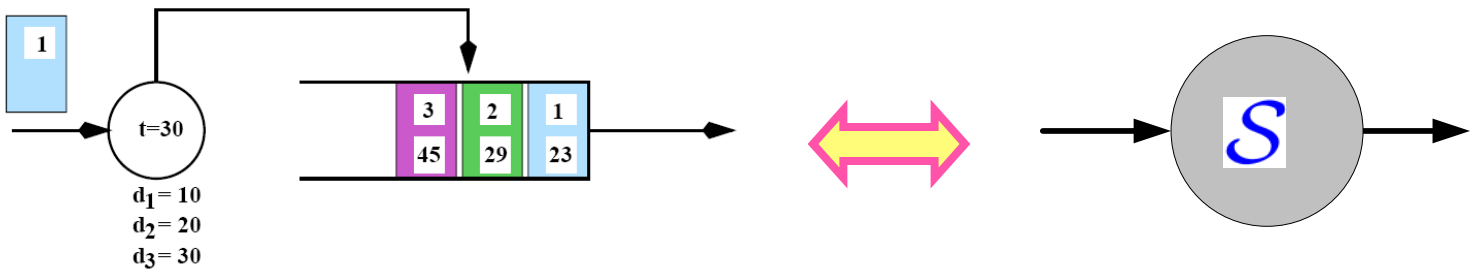
How do end-to-end delay bounds look like for different schedulers?

Does link scheduling matter on long paths?

# Service curves vs. schedulers

*(JSAC 2011)*

- How well can a service curve describe a scheduler?

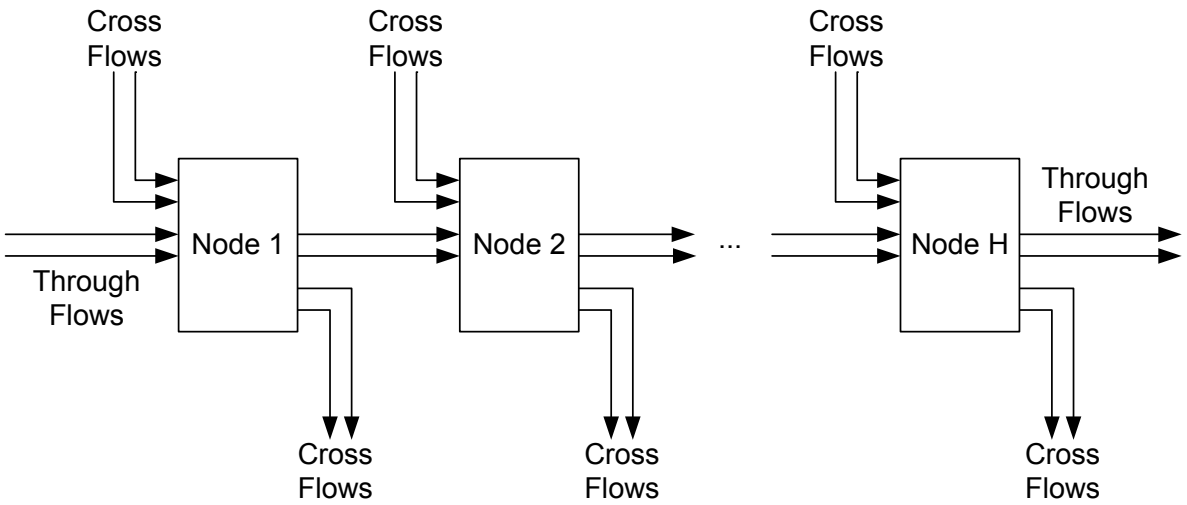



- For schedulers considered earlier, the following is ideal:

$$\mathcal{S}_j(t;\theta) = [Ct - \mathcal{H}_c(t - \theta + \Delta_{j,k}(\theta))]_+ I(t > \theta)$$

with indicator function $I(\text{expr})$ and parameter $\theta \geq 0$

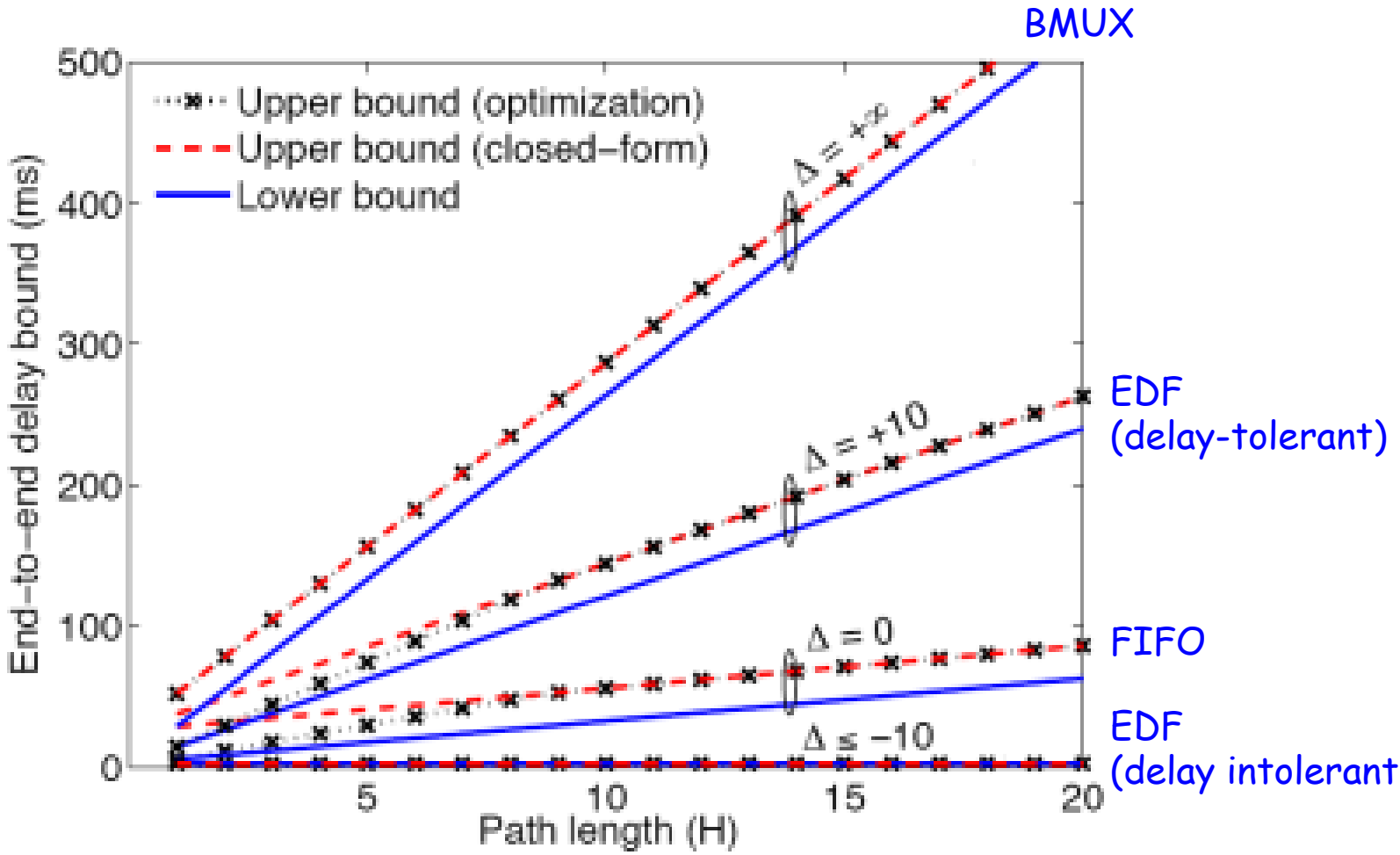
# Example: End-to-End Bounds




- Traffic is Markov Modulated On-Off Traffic (EBB model)
- Fixed capacity link

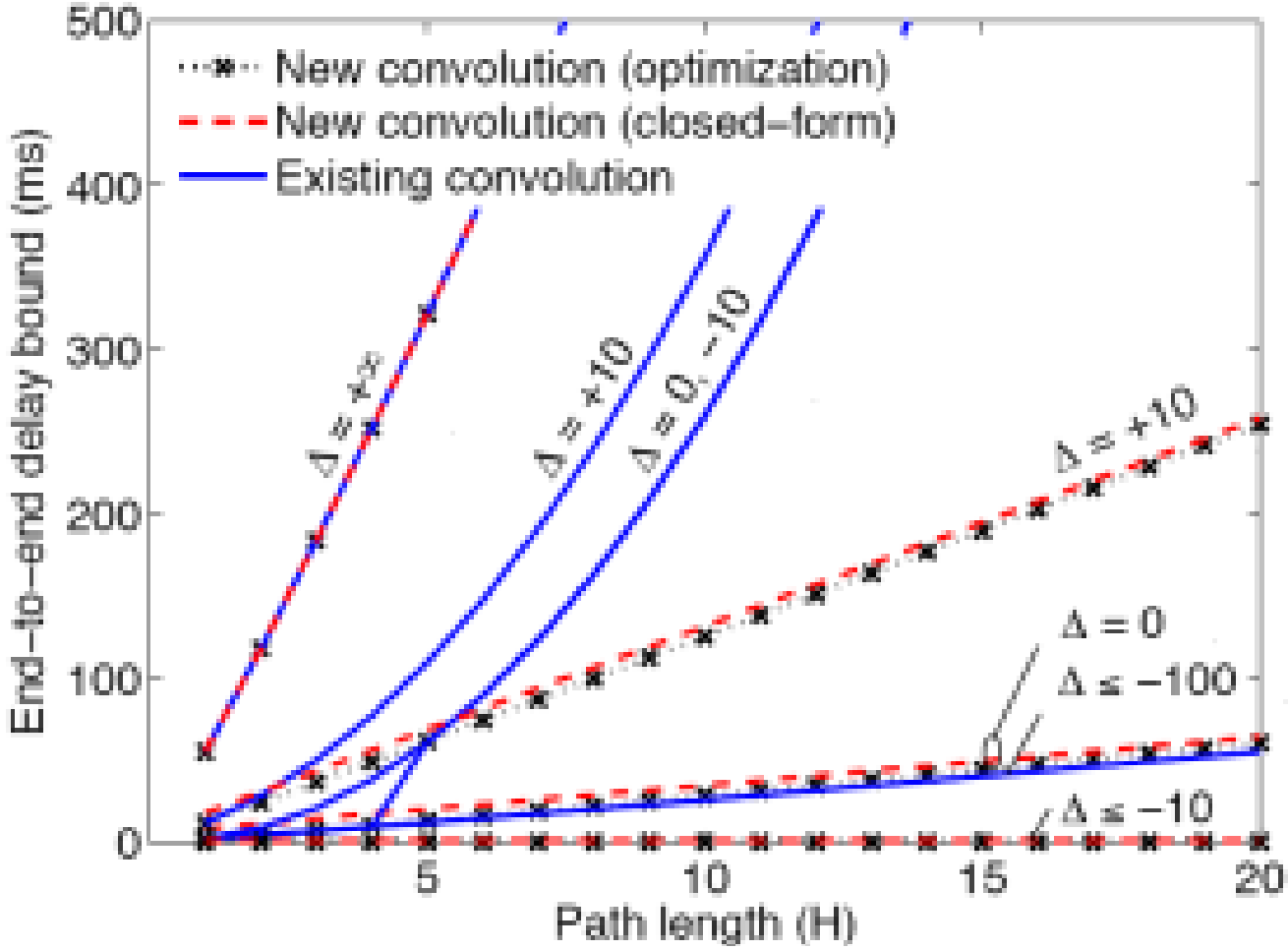
# Example: Deterministic E2E Delays *(Infocom '11)*

- Peak rate: E(t) = b+rt
  Average rate: r = 0.15 Mbps
- *C* = 100 Mbps
- Link utilization: 90% (through: 1.5%)



# Example: Statistical E2E Delays *(Infocom`11)*

- Peak rate: P = 1.5 Mbps
  Average rate: $\tau$ = 0.15 Mbps
  EBB traffic
- $C$ = 100 Mbps
- $\varepsilon = 10^{-9}$
- Link utilization: 90% (through: 1.5%)



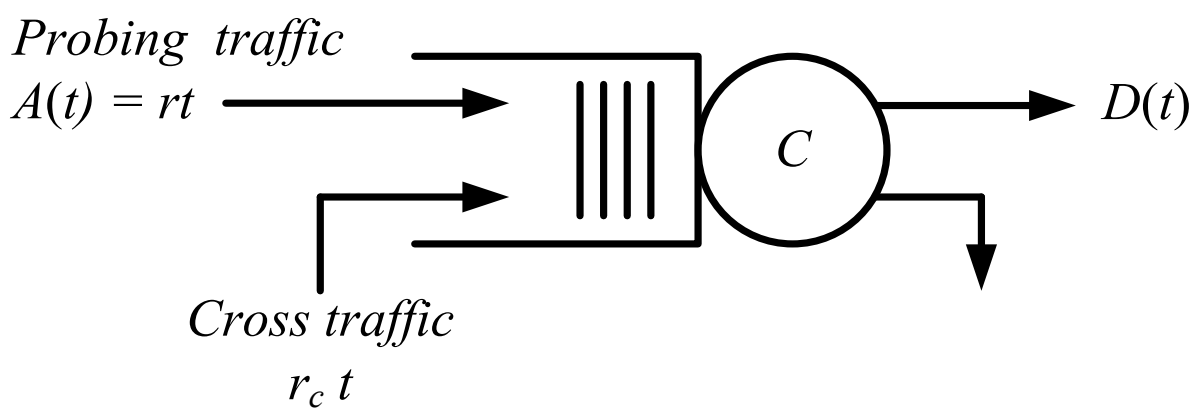
# How about an overloaded scheduler ?

- Delays are of course unbounded?
- But how about throughput?

# CBR traffic at a FIFO scheduler

- Problem appeared in probing method for bandwidth estimation

- FIFO system




- Output:

$$D(t) = \begin{cases} rt \ , & \text{if } r \leq C - r_c \ , \\ \frac{r}{r+r_c} Ct \ , & \text{if } r > C - r_c \ . \end{cases}$$

- Service curve:

$$S(t) = [Ct - r_c]^+ t$$

# Overloaded systems

- FIFO shares bandwidth proportional to input
- Service curve becomes BMUX

- The same holds
  - for any $\Delta$-scheduler with finite $\Delta$s
  - for any traffic type with an average traffic rate

Can we compute scaling of delays for nasty traffic ?

# Heavy-Tailed Self-Similar Traffic

- A **heavy-tailed** process $X$ satisfies

$$\boxed{Pr(X(t) > x) \sim Kx^{-\alpha}}$$

with $1 < \alpha < 2$

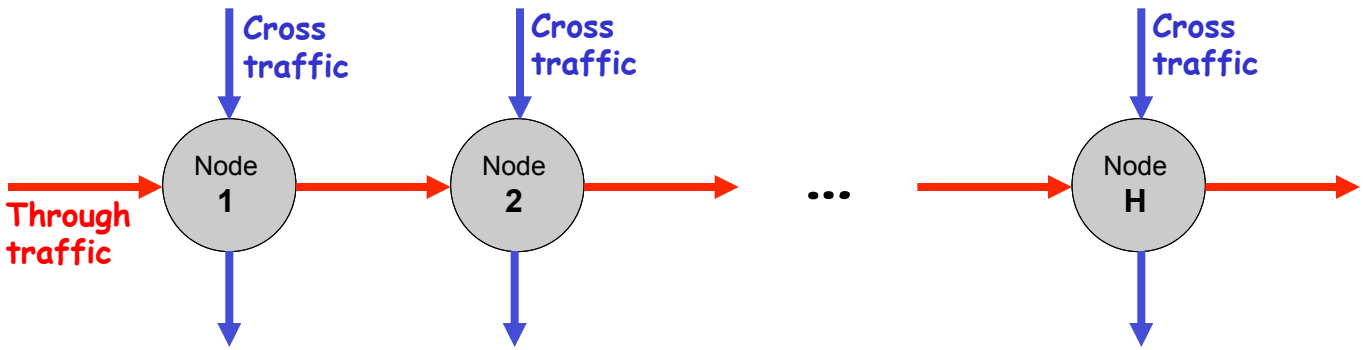
- A **self-similar** process satisfies

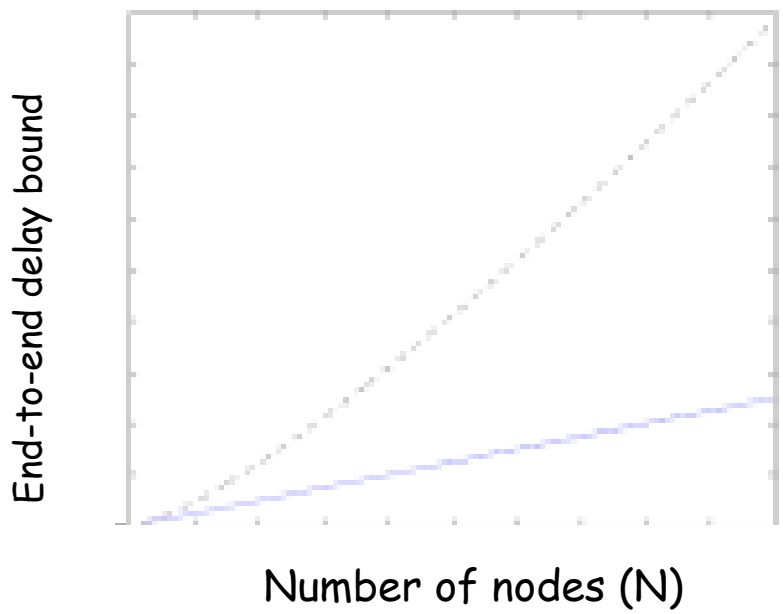
$$\boxed{X(t) \sim_{dist} a^{-H} X(at)}$$

$a > 0$

$H \in (0, 1)$ Parameter

# End-to-End Delays




Exponentially bounded traffic
**Θ (N log N)**
(Sigmetrics 2005, Infocom 2007)

Worst-case delays
**Θ (N)**
(e.g., LeBoudec and Thiran 2000)

# htts Traffic Envelope

- Heavy-tailed self-similar (**htss**) envelope:

$$Pr(A(s,t) > \underbrace{r(t-s) + \sigma(t-s)^H}_{\mathcal{G}(t-s;\sigma)}) \leq \underbrace{K\sigma^{-\alpha}}_{\varepsilon(\sigma)}$$

- **Main difficulty:** Backlog and delay bounds require sample path envelopes of the form

$$Pr(\sup_{s \leq t} \{A(s,t) - \overline{\mathcal{G}}(t-s;\sigma)\} > 0) \leq \varepsilon(\sigma)$$

- **Key contribution (not shown):**
  Derive sample path bound for htss traffic

# Example: Node with Pareto Traffic

(Infocom 2010)

Traffic parameters:

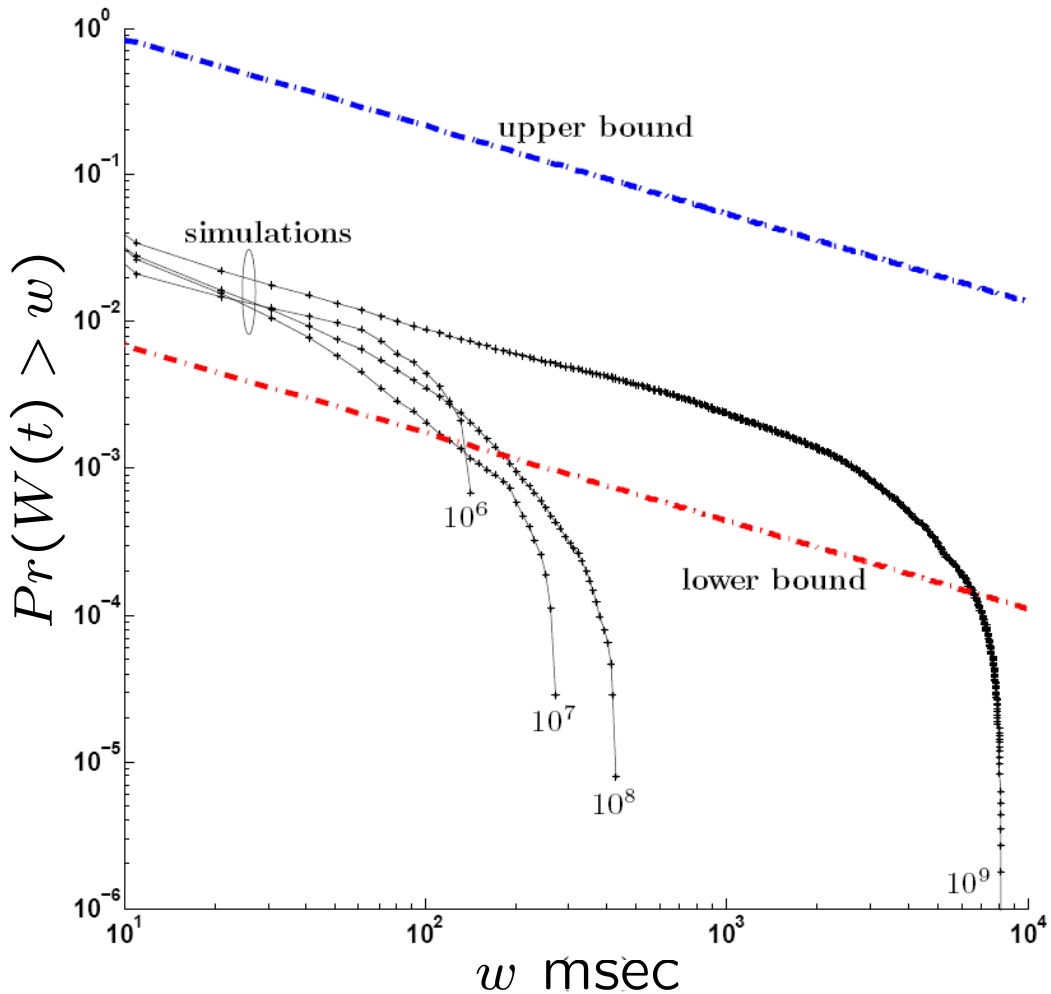
$$\alpha = 1.6$$
$$b = 150 \ Byte$$
$$\lambda = 75 \ Mbps$$

Node:

- Capacity C=100 Mbps with packetizer
- No cross traffic

Compared with:

- Lower bound from ToN 2011 paper
- Simulations



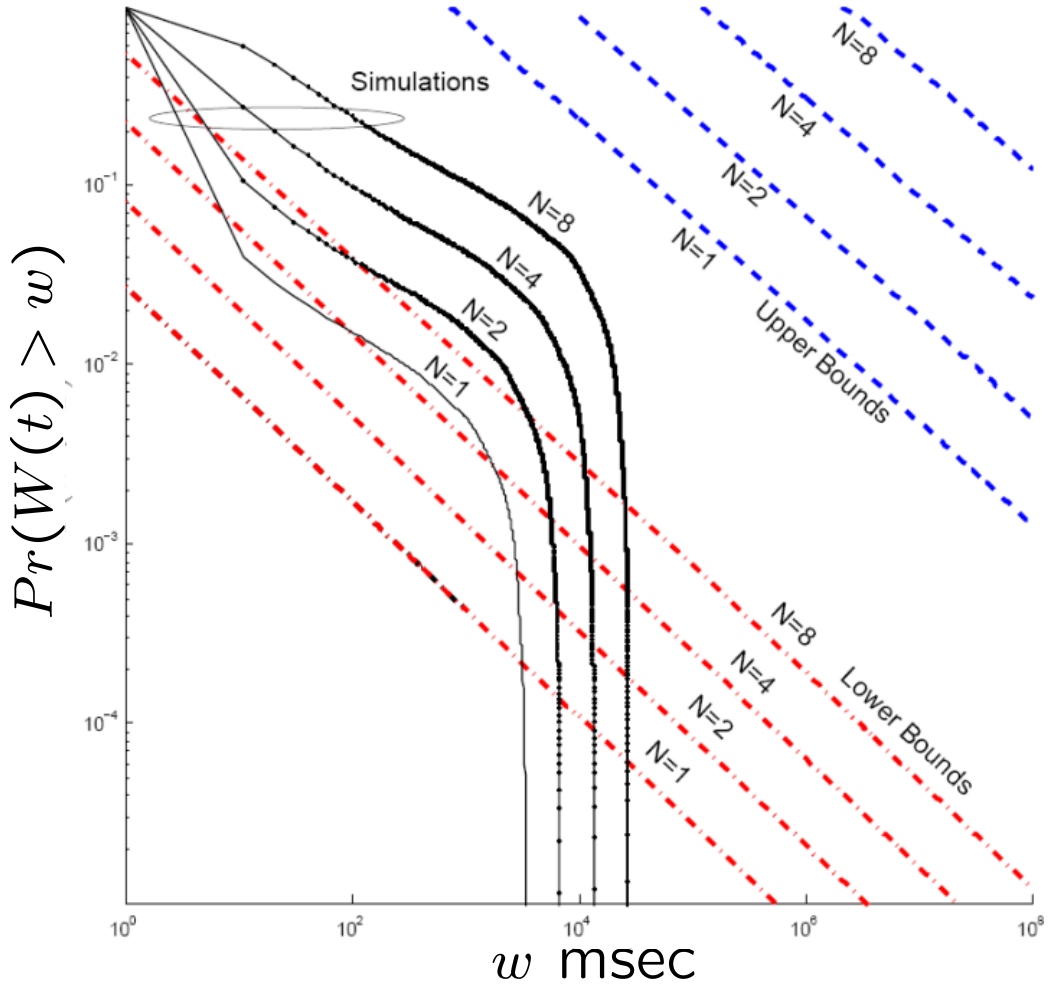
# Example: Nodes with Pareto Traffic (End-to-end)

Parameters:

$$N = 1, 2, 4, 8$$

Compared with:

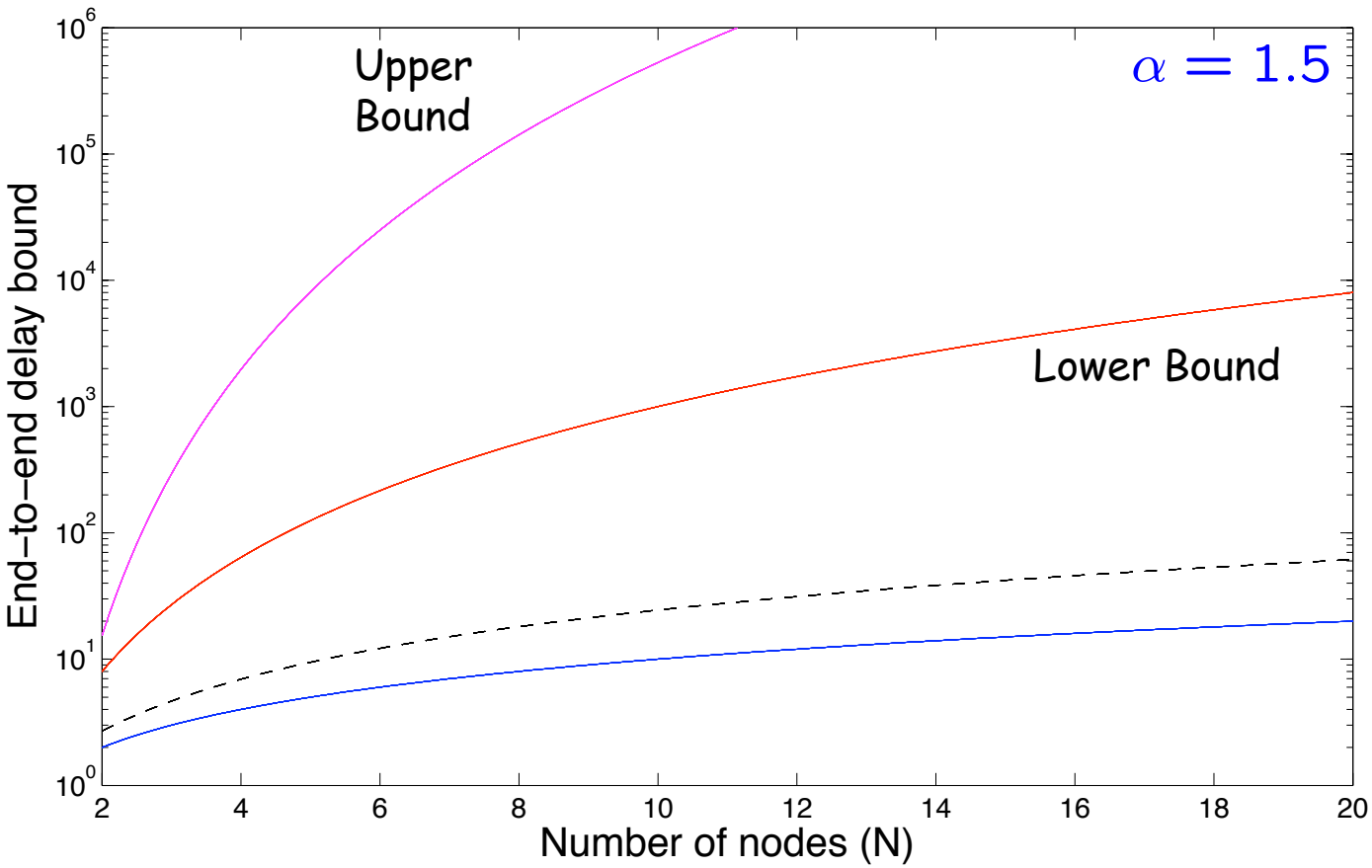
- Lower bound from ToN 2011 paper
- Simulation traces of $10^8$ packets



# Illustration of scaling bounds *(Infocom 2010)*

Upper Bound: $\mathcal{O}(N^{\frac{\alpha+1}{\alpha-1}}(\log N)^{\frac{1}{\alpha-1}})$

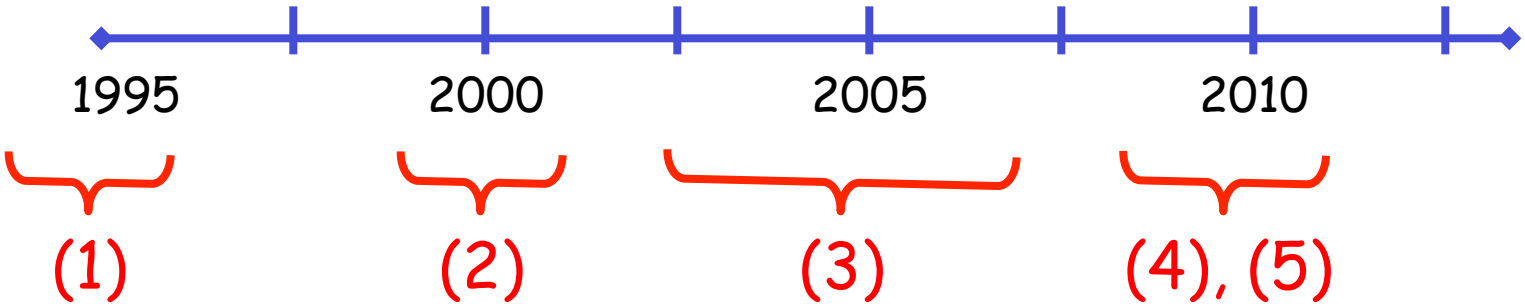
Lower Bound: $\Theta(N^{\frac{\alpha}{\alpha-1}})$




**Θ (N log N)**

**Θ (N)**

# Summary of insights




① Satisfying delay bounds does not require peak rate allocation for complex traffic

② Statistical multiplexing gain dominates gain due to link scheduling

③ $\Theta(H \log H)$ scaling law of end-to-end delays

④ *New laws for heavy-tailed traffic*

⑤ Link scheduling plays a role on long path

# Example: Pareto Traffic

- Size of i-th arrival: $$Pr(X_i > x) = \left(\frac{x}{b}\right)^{-\alpha}$$ $x \geq b$ $1 < \alpha < 2$
- Arrivals are evenly spaced with gap $\lambda$: $$A(t) = \sum_{i=1}^{N(t)} X_i$$
- With Generalized Central Limit Theorem … $$A(t) \approx \lambda t E[X] + c_\alpha (\lambda t)^{1/\alpha} S_\alpha$$
  … and tail bound $$Pr(S_\alpha > \sigma) \sim (c_\alpha \sigma)^{-\alpha}$$

  $\alpha$-stable distribution
- … we get htss envelope

$$\begin{aligned} \mathcal{G}(t;\sigma) &= \lambda E[X] t + \sigma t^{1/\alpha} \\ \varepsilon(\sigma) &= \lambda \sigma^{-\alpha} \end{aligned}$$

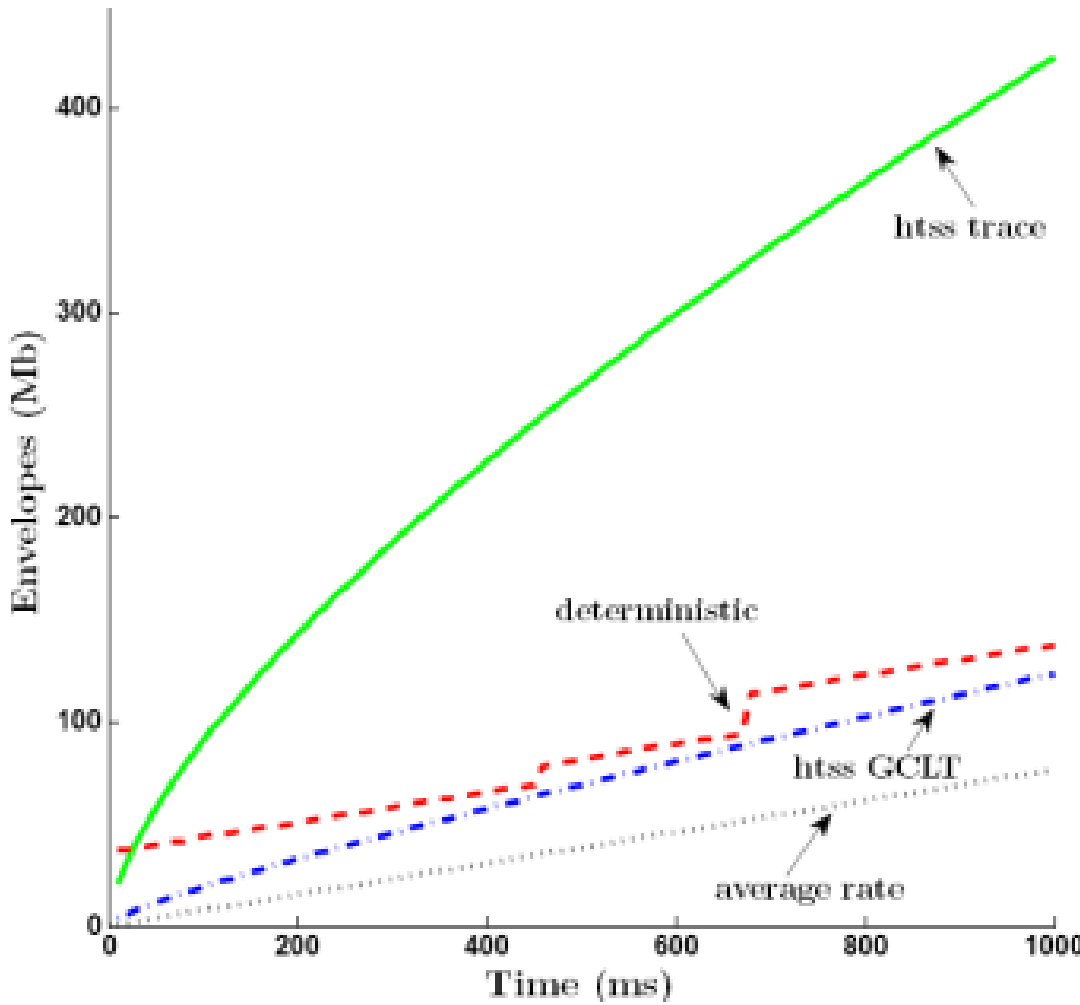
# Example: Envelopes for Pareto Traffic (Infocom 2010)

Parameters:

$$\alpha = 1.6$$
$$b = 150\ Byte$$
$$\lambda = 75\ Mbps$$

Comparison of envelopes:

- htss GCLT envelope
- Average rate
- Trace-based
  - deterministic envelope
  - htts trace envelope



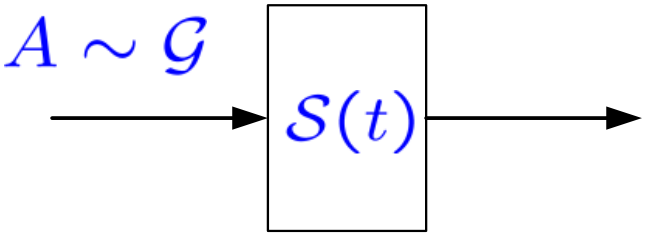
# Single Node Delay Bound

- htss envelope:

$$\begin{aligned} \mathcal{G}(t;\sigma) &= rt + \sigma t^H \\ \varepsilon(\sigma) &= K\sigma^{-\alpha} \end{aligned}$$

- ht service curve:

$$\begin{aligned} \mathcal{S}(t;\sigma) &= [Rt - \sigma]_+ \\ \varepsilon(\sigma) &= L\sigma^{-\beta} \end{aligned}$$




- Delay bound:

$$Pr(W(t) > w) \leq M(Rw)^{-\min\{\alpha(1-H),\beta\}}$$