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

MPEG-Compressed Video Trace

Frame size

Frame number

Peak rate

Mean rate
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
Computing delays in such networks is notoriously hard ...
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)

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

Arrivals from flow \( j \)

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

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 \textbf{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**

\[
d_j \geq \frac{\sum_{k=1}^{j} b_k - \sum_{k=1}^{j-1} r_k d_k}{C - \sum_{k=1}^{j-1} r_k}
\]
C = 45 Mbps

MPEG 1 traces:

Lecture:
d = 30 msec

Movie
(Jurassic Park):
d = 50 msec

Numerical Result (Sigmetrics 1995)
Deterministic worst-case

Expected case

Deterministic worst-case

Probable worst-case
Statistical Multiplexing Gain

Worst-case arrivals

Arrivals

Flow 1
Flow 2
Flow 3

Backlog

Worst-case backlog

Time
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** \( G \):
  \[
  \Pr\{A(s, t) > G(t - s) + \sigma\} < \varepsilon(\sigma) \quad \forall s, t
  \]

- **Statistical sample path envelope** \( \mathcal{H} \):
  \[
  \Pr\{\sup_{s \leq t} \{A(s, t) - \mathcal{H}(t - s)\} > \sigma\} < \varepsilon(\sigma)
  \]

Statistical envelopes are non-random functions
Aggregating arrivals

Arrivals from group of flows:

with deterministic envelopes:

\[ A_1 \sim E_1 \]

\[ A_N \sim E_N \]

with statistical envelopes:

\[ A_C = \sum_{j} A_j \]

\[ E_C = \sum_{j} E_j \]

\[ G_C \ll \sum_{j} G_j \ll E_C \]
Statistical envelope for group of independent (shaped) flows

- Exploit independence and extract statistical multiplexing gain when calculating $G_C$

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

\[
G_C(t) = \inf_{s > 0} \frac{1}{s} \left( \sum_{j \in C} \log \bar{M}_j(s, t) - \log \varepsilon \right)
\]

\[
\bar{M}_j(s, t) = 1 + \frac{\rho_j t}{E_j(t)} \left( e^{sE_j(t)} - 1 \right)
\]

\[
\rho_j = \lim_{\tau \to \infty} \frac{E_j(\tau)}{\tau}
\]
Statistical vs. Deterministic Envelope

\[ E(t) = \min(P_t, \sigma + \rho t) \]

**Type 1 flows:**
- \( P = 1.5 \text{ Mbps} \)
- \( \rho = 0.15 \text{ Mbps} \)
- \( \sigma = 95400 \text{ bits} \)

**Type 2 flows:**
- \( P = 6 \text{ Mbps} \)
- \( \rho = 0.15 \text{ Mbps} \)
- \( \sigma = 10345 \text{ bits} \)

\[ \varepsilon = 10^{-6} \]
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:

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

Statistical Service

Delay bound violation with $\epsilon$ if:

$$\sup_s \left\{ \sum_p \mathcal{H}_{C_p}(s + \Delta_{qp}) - Cs \right\} \leq C d_q$$
Statistical Multiplexing vs. Scheduling (JSAC 2000)

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

- Thick lines: EDF Scheduling
- Dashed lines: SP scheduling

Statistical multiplexing makes a big difference
Scheduling has small impact

d_{\text{terminator}} = 100$ ms
\[ d_{\text{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
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

\[ G^\varepsilon(\tau) = \inf_{s > 0} \left\{ \tau \alpha(s, \tau) - \frac{\log \varepsilon}{s} \right\} \]
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

\( C = 100 \text{ Mbps}, \quad \varepsilon = 10^{-6} \)
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: \( D(t) \geq A \ast S(t) \)

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

- \((\min,+)\) algebra operators
  - Convolution:
    \[ f \ast g(t) = \inf_{0 \leq s \leq t} (f(s) + g(t-s)) \]
  - Deconvolution:
    \[ f \circ g(t) = \sup_{s \geq 0} (f(t+s) - g(s)) \]
Main result:

If $S^1, S^2, S^3$ describes the service at each node, then

$$S_{\text{net}} = S^1 \ast S^2 \ast S^3$$

describes the service given by the network as a whole.
Stochastic Network Calculus

- Probabilistic view on arrivals and service
  - Statistical Sample Path Envelope
    \[ Pr\{\sup_{s \leq t}(A(s, t) - H(t - s)) > \sigma\} \leq \varepsilon(\sigma) \]
  - Statistical Service Curve
    \[ Pr\{D(t) - A \ast S(t) > \sigma\} \leq \varepsilon(\sigma) \]
- Results on performance bounds carry over, e.g.:
  - Backlog Bound
    \[ Pr(B(t) > H \ominus S(0)) \leq \varepsilon \]
Stochastic Network Calculus

- **Hard problem:** Find $S^{net}$ so that $S^{net} = S^1 \ast S^2 \ast \ldots \ast S^H$

- **Technical difficulty:**

$$D^2(t) = \inf_{0 \leq s \leq t} \left( A^2(s) + S^2(t-s) \right)$$

$$= A^2(s_0) + S^2(t-s_0)$$

$s_0$ is a random variable!
• **Notation:** \( S_{-\delta}(t) = S(t) - \delta t \)

• **Theorem:** If \( S^1, S^2, \ldots, S^H \) are statistical service curves, then for any \( \delta > 0 \):

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

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

- Traffic with Exponentially Bounded Burstiness (EBB)

\[ P(A(s, t) - \rho(t - s) > \sigma) \leq M e^{-\alpha \sigma} \]

\[ G(t - s; \sigma) \]

\[ \varepsilon(\sigma) \]

- 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: $S_j = [Ct - H_c(t)]_+$
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
- $T = \frac{1}{\mu} + \frac{1}{\lambda} = 10$ 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 \(\text{}^{(\text{ToN 2011})}\)

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

**Theorem:** E2E delay \(W_H\) satisfies for all \(0 < z < 1\)

\[
Pr(W_H \leq \gamma_1 H \log(\gamma_2 H)) \leq z
\]

**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
\[ C = 100 \text{ Mbps} \]

Mean packet size
\[ \frac{1}{\mu} = 400 \text{ Bytes} \]

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**

![Diagram showing **H_c** and **S_j** formula: $S_j = [Ct - H_c(t)]_+$]

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

**Does link scheduling matter on long paths?**
How well can a service curve describe a scheduler?

For schedulers considered earlier, the following is ideal:

\[ S_j(t; \theta) = [Ct - H_c(t - \theta + \Delta_{j,k}(\theta))]_+ I(t > \theta) \]

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

- Peak rate: $P = 1.5$ Mbps
  - Average rate: $\tau = 0.15$ Mbps

- EBB traffic

- $C = 100$ Mbps
- $\epsilon = 10^{-9}$
- 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
  \[ Pr(X(t) > x) \sim K x^{-\alpha} \]
  with $1 < \alpha < 2$

- A self-similar process satisfies
  \[ X(t) \sim_{\text{dist}} a^{-H} X(at) \]
  \[ a > 0 \]
  \[ H \in (0, 1) \] Parameter
End-to-End Delays

Through traffic

Cross traffic

Node 1

Cross traffic

Node 2

... Cross traffic

Node H

End-to-end delay bound

Exponentially bounded traffic
\(\Theta(N \log N)\)
(Sigmetrics 2005, Infocom 2007)

Worst-case delays
\(\Theta(N)\)
(e.g., LeBoudec and Thiran 2000)

Number of nodes (N)
htss Traffic Envelope

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

\[ Pr(A(s, t) > r(t - s) + \sigma(t - s)^H) \leq K \sigma^{-\alpha} \]

- \( \mathcal{G}(t - s; \sigma) \)

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

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

- **Key contribution (not shown)**: Derive sample path bound for htss traffic
Example: Node with Pareto Traffic (Infocom 2010)

Traffic parameters:
\[ \alpha = 1.6 \]
\[ b = 150 \text{ Byte} \]
\[ \lambda = 75 \text{ Mbps} \]

Node:
- Capacity \( C = 100 \text{ Mbps} \) 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: \( \Theta(N^{\frac{\alpha+1}{\alpha-1}} \log N^{\frac{1}{\alpha-1}}) \)

Lower Bound: \( \Theta(N^{\frac{\alpha}{\alpha-1}}) \)
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:
  \[ Pr(X_i > x) = \left( \frac{x}{b} \right)^{-\alpha} \quad x \geq b \quad 1 < \alpha < 2 \]

- Arrivals are evenly spaced with gap \( \lambda \):
  \[ A(t) = \sum_{i=1}^{N(t)} X_i \]

- With Generalized Central Limit Theorem ...
  ... and tail bound

- ... we get htss envelope
  \[ G(t; \sigma) = \lambda E[X]t + \sigma t^{1/\alpha} \]
  \[ \varepsilon(\sigma) = \lambda \sigma^{-\alpha} \]
Example: Envelopes for Pareto Traffic (Infocom 2010)

Parameters:
\[ \alpha = 1.6 \]
\[ b = 150 \text{ Byte} \]
\[ \lambda = 75 \text{ Mbps} \]

Comparison of envelopes:
- htss GCLT envelope
- Average rate
- Trace-based
  - deterministic envelope
  - htss trace envelope
Single Node Delay Bound

- **htss envelope:**
  \[ G(t; \sigma) = r t + \sigma t^H \]
  \[ \varepsilon(\sigma) = K \sigma^{-\alpha} \]

- **ht service curve:**
  \[ S(t; \sigma) = [R t - \sigma]^+ \]
  \[ \varepsilon(\sigma) = L \sigma^{-\beta} \]

Delay bound:

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