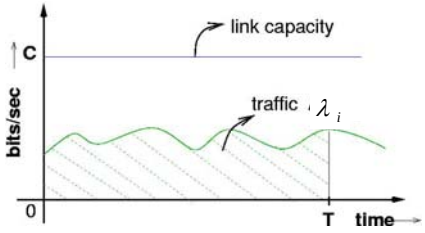
A Min-Plus System Interpretation of Bandwidth Estimation

J. Liebeherr
University of Toronto

Joint work with M. Fidler and S. Valaee.

Available Bandwidth

- Available bandwidth is the unused capacity along a path



- Available bandwidth of a link:

$$\alpha_i(t, t+\tau) = \frac{1}{\tau}\int_t^{t+\tau} C_i(x) - \lambda_i(x)dx$$

- Available bandwidth of a path:

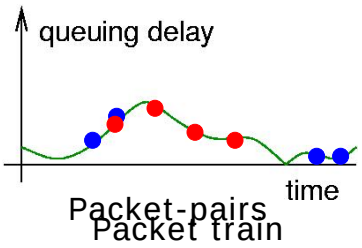
$$\alpha(t, t+\tau) = \min_{i=1,\ldots,H} \alpha_i(t, t+\tau)$$

- Goal:
  Use end-to-end probing to estimate available bandwidth

Edited slide from: V. Ribeiro, Rice. U, 2003

From packet pairs to packet trains

- Shortcoming: packet-pairs do not capture temporal queuing behavior essential for available bandwidth estimation

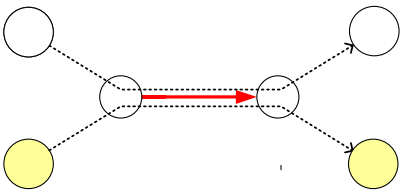


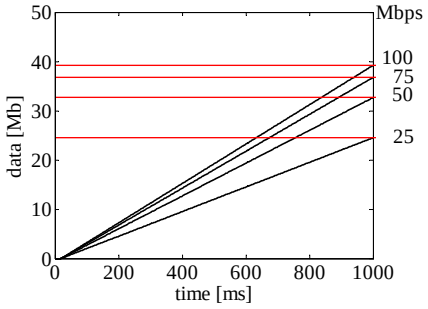
- Solution: Transmit multiple packet in a probe (→packet train)

Edited slide from: V. Ribeiro, Rice. U, 2003

How fast to send a packet train?

- Rate of packet train is determined by gap between packets



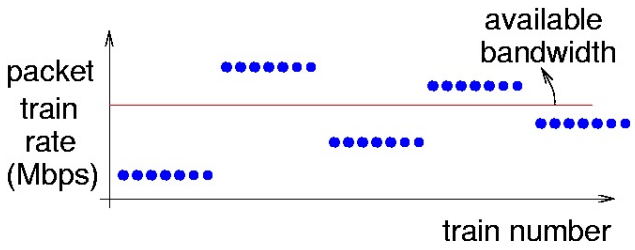


- So: rate at which the packet trains are sent is crucial:
  - Rate too high → probes preempt existing traffic
  - Rate too low → probes only measure the input rate

Bandwidth estimation methods

- Numerous techniques emerged in the last 5 years
- Two methods are relevant to this talk:
  - Pathload (PAM `02, Jain/Dovrolis)
  - Pathchirp (PAM '03, Ribeiro et. Al.)

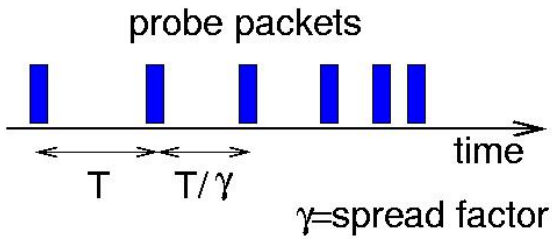
Pathload [Jain & Dovrolis, 2002]



- CBR packet trains
- Vary rate of successive trains
- Converges to available bandwidth

- We call this approach: **Rate scanning**

Packet Chirps [Ribeiro et.al., 2003]



- Uses a single packet train
- Successively decrease packet gap within packet train
  … thereby covering a range of probing rates

- We call this approach: **Rate chirps**

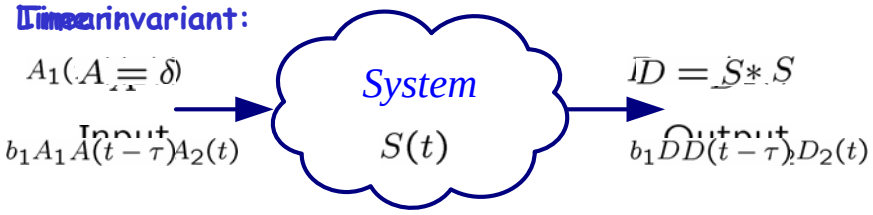
Our point of departure:

- There are dozens of probing schemes (with inventive acronyms), each offering their own heuristics and assumptions on the network
  - Can heuristics be made rigorous?
  - Is there a better way to view the problem?

IS THERE A (better) THEORY OF BANDWIDTH ESTIMATION ?

(Classical) System Theory

Linear Time Invariant (LTI) Systems



- If input is Dirac impulse, output is the system response S
- Output can be calculated from input and system response:

$$D(t) = \int_{-\infty}^{\infty} A(\tau) \cdot S(t-\tau) d\tau =: A * S(t)$$

"convolution"

System Theory for Networks

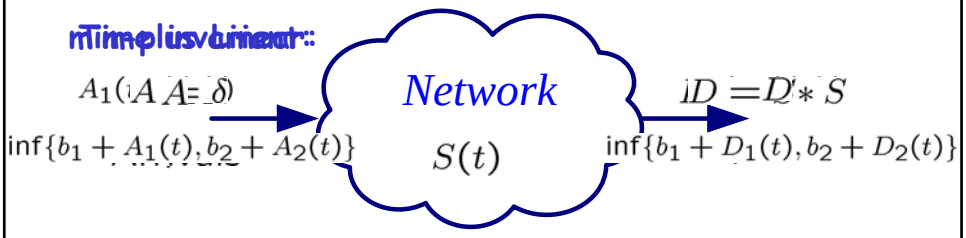
- Networks can be viewed as linear systems in a different algebra:
  - Addition (+) → Minimum (inf)
  - Multiplication (·) → Addition (+)

  Min-Plus Algebra

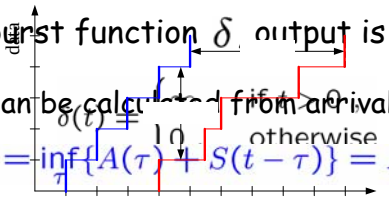
- Systems theory for networks is called **Network Calculus**
  - emerged in the 1990s

- Network service is described by a service curve S

Min-Plus Linear Systems



- If input is burst function δ, output is the service curve S
- Departures can be calculated from arrivals and service curve:



$$D(t) = \inf_{\tau}\{A(\tau) + S(t-\tau)\} = A * S(t)$$

"min-plus convolution"

Transforms

- Classical LTI systems

Time domain	Fourier transform	Frequency domain
$f(t)$	→	$\mathcal{F}_f(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$
$f * g(t)$		$\mathcal{F}_{f*g}(\omega) = \mathcal{F}_f(\omega) \cdot \mathcal{F}_g(\omega)$

- Min-plus linear systems

Time domain	Legendre transform	Rate domain
$f(t)$	→	$\mathcal{L}_f(r) = \sup_{\tau}\{r\tau - f(\tau)\}$
$f * g(t)$		$\mathcal{L}_{f*g} = \mathcal{L}_f + \mathcal{L}_g$

Properties: (1) $\mathcal{L}(\mathcal{L}(f)) \leq f$. If f is convex: $\mathcal{L}(\mathcal{L}(f)) = f$
(2) If g convex, then $f \geq g \Leftrightarrow \mathcal{L}_f \leq \mathcal{L}_g$

One more thing …

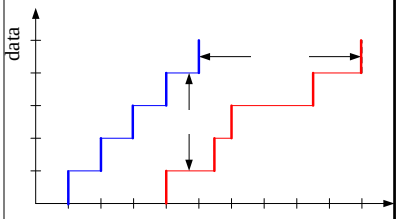
- Many networks are not min-plus linear
  - i.e., for some t: $D(t) \neq A * S(t)$
- … but can be described by a **lower service curve** $\underline{S}$
  - such that for all t: $D(t) \geq A * \underline{S}(t)$
- Having a lower service curve is often enough, since it provides a lower bound on the service !!

Bandwidth estimation in the network calculus

- View the network as a min-plus system that is either linear or nonlinear

Bandwidth estimation scheme:

1. Timestamp probes
   $A^p(t)$ - Send probes
   $D^p(t)$ - Receive probes
2. Use probes to find a $\underline{S}$ that satisfies $D(t) \geq A * \underline{S}(t)$ for all (A,D).
3. $\underline{S}$ is the estimate of the available bandwidth.



Bandwidth estimation in the network calculus

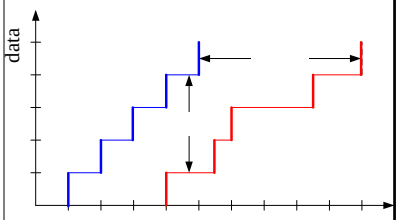
- View the network as a min-plus system that is either linear or nonlinear

Bandwidth estimation scheme:

1. Timestamp probes
   $A^p(t)$ - Send probes
   $D^p(t)$ - Receive probes

2. Use probes to find a $\underline{S}$ that satisfies $D(t) \geq A * \underline{S}(t)$ for all (A,D).

3. $\underline{S}$ is the estimate of the available bandwidth.



Why is bandwidth estimation hard ?

- Goal: Find $\underline{S}$ as large as is possible

- So: maximize $\underline{S}$
  subject to $D(t) \geq A * \underline{S}$
  $= \inf_\tau \{A(\tau) + \underline{S}(t-\tau)\},$

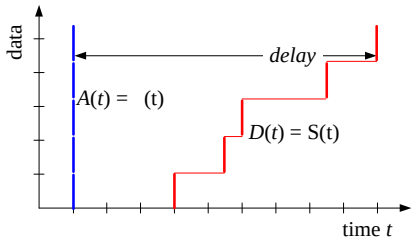
  $\forall t \geq 0,$ for all pairs (A, D).

- This is a max-min optimization (= very hard problem group)

Is there hope?

- If network is min-plus linear, we get $D = A * S$
- If we set $A = \delta$, then $D = \delta * S = S$

- **So**: We get an exact solution when the probe consist of a burst (of infinite size and sent with an infinite rate)



- However:
  - Large bursts consume a lot of bandwidth
  - Large bursts can make a network non-linear

→ Need to find better solutions

Three methods
(that work better than infinite sized bursts)

1. Passive measurements
2. Rate Scanning
3. Rate Chirps

For the time being, we will assume a linear network, i.e., $D = A * S$

Passive measurements

- **How much info can be extracted from passive measurements of a traffic flow? (without probes)**
- Define a min-plus deconvolution:

$$f \oslash g(t) = \sup_{\tau}\{f(t+\tau) - g(\tau)\}$$

- Deconvolution is not an inverse of the convolution, but:
  (a) "Duality": $f \leq g * h \quad \Leftrightarrow \quad h \geq f \oslash g.$

  (b) New result: $((h * g) \oslash g) * g = h * g$

Passive measurements (2): Algorithm

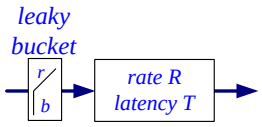
- **Step 1:** Collect measurements (A_{tr}, D_{tr})
  - We know that $D_{tr} = A_{tr} * S$, but we do not know S
- **Step 2:** Compute $\tilde{S} = D_{tr} \oslash A_{tr}$
  - Then:

$$\begin{aligned} D_{tr} &= S * A_{tr} \\ &= ((S * A_{tr}) \oslash A_{tr}) * A_{tr} \\ &= (D_{tr} \oslash A_{tr}) * A_{tr} \\ &= \tilde{S} * A_{tr} \end{aligned}$$

  - Applying "duality" to $D_{tr} = A_{tr} * S$ gives $S \geq D_{tr} \oslash A_{tr} = \tilde{S}$
- **Step 3:** So, we know that our $\tilde{S}$ is a service curve and that $S \geq \tilde{S}$
  In fact, $\tilde{S}$ is the best possible estimate based on the trace

Passive measurements (3): Example

Service: $S(t) = (b + rt) * (R[t - T]^+)$

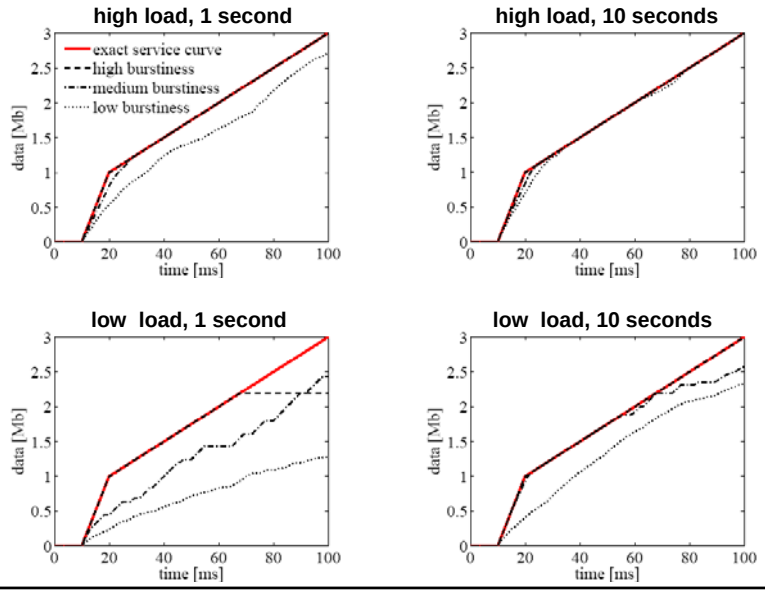


R = 100 Mbps T = 10 ms
r = 25 Mbps b = 750 kB

Traffic: On-Off Source

Scenario	high load			low load		
Burstiness	high	med	low	high	med	low
Number of sources	1	5	25	1	5	25
Source peak rate [Mbps]	200	40	8	200	40	8
Total average rate [Mbps]	20	20	20	10	10	10

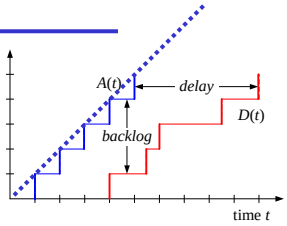
Passive measurements (4): Results



Rate Scanning (1): Theory

- **Backlog:** $B(t) = A(t) - D(t)$
- **Max. backlog:**

$$B_{max} = \sup_t \{A(t) - D(t)\}.$$



- **If** $A(t) = rt$**, we can write this as:**

$$\begin{aligned} B_{max}(r) &= \sup_t \{rt - \inf_\tau \{r\tau + S(t-\tau)\}\} \\ &= \sup_t \{\sup_\tau \{r(t-\tau) - S(t-\tau)\}\} \\ &= \sup_t \{rt - S(t)\} \\ &= \mathcal{L}_S(r) \end{aligned}$$

- **Inverse transform:** If S is convex we have

$$S(t) = \mathcal{L}(\mathcal{L}_S)(t) = \mathcal{L}_{B_{max}}(t) = \sup_r \{rt - B_{max}(r)\}$$

Rate Scanning (2): Algorithm

Step 1: Transmit a packet train at rate r,
compute $B_{max}(r)$
compute $S(t) = \mathcal{L}_{B_{max}}(t)$

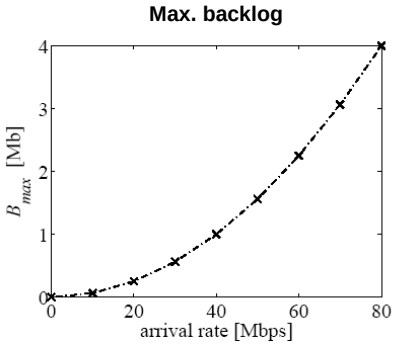
Step 2: If estimate of S has improved, increase r and go to Step 1.

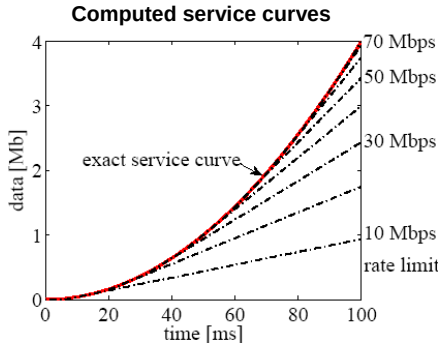
- This method is very close to **Pathload** !

Rate Scanning (3): Example

Service: $S(t) = 0.4t^2$

Rate scanning: 10, 20, …, 80 Mbps





Rate chirp (1): Theory

- In signal processing, chirp signal is a signal whose frequency changes in time
- In network probing, rate chirp is a packet train that is sent at different rates

- Suppose we have a service curve: $D = A * S$
- Take the Legendre transform: $\mathcal{L}_D = \mathcal{L}_{A*S} = \mathcal{L}_A + \mathcal{L}_S$
- Rearrange (Need: $\mathcal{L}_A(r) < \infty$) $\mathcal{L}_S = \mathcal{L}_D - \mathcal{L}_A$
- Taking the transform again: $\mathcal{L}(\mathcal{L}_S) = \mathcal{L}(\mathcal{L}_D - \mathcal{L}_A)$
- If S is convex we have $S = \mathcal{L}(\mathcal{L}(S))$

Rate chirp (2): Algorithm

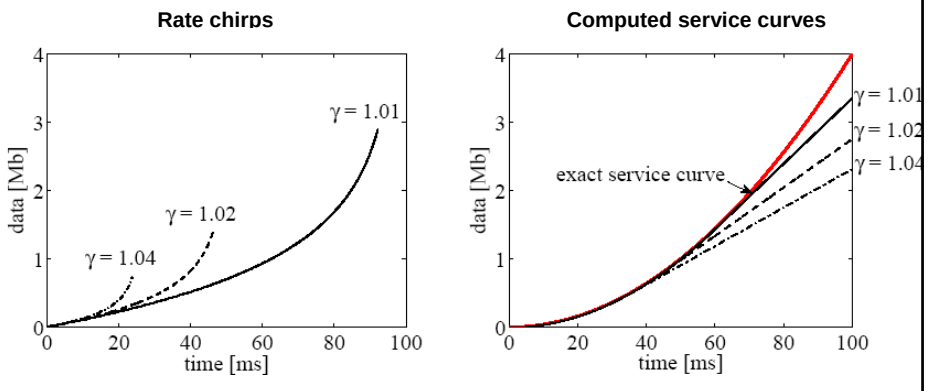
- **Note:** A packet chirp satisfies $\mathcal{L}_{A_{chrp}}(r) < \infty$
- **Step 1:** Collect measurements of a chirp: (A_{chrp}, D_{chrp})
- **Step 2:** Compute $\tilde{S}(t) = \mathcal{L}(\mathcal{L}_{D_{chrp}} - \mathcal{L}_{A_{chrp}})(t)$.
- **Then:** $\tilde{S} \leq S$, i.e., we have a lower service curve

  If S is convex, then $\tilde{S} = S$, and we have recovered the service curve completely.

Rate Scanning (3): Example

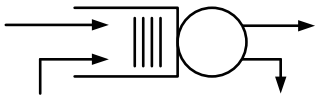
Service: $S(t) = 0.4t^2$

Rate chirp: Packet size: 1200 Bytes
spread factor: 1.01, 1.02, 1.04



Non-Linear Systems *(or: How about FIFO ?)*

- When we exploit $D(t) = A * S(t)$ we assume a min-plus linear system
- In a linear system, the system response S does not change with the input. But in FIFO, the system response changes with the input rate. So, FIFO is not min-plus linear.
- In a FIFO system with …



… we get *(this has been measured)*

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

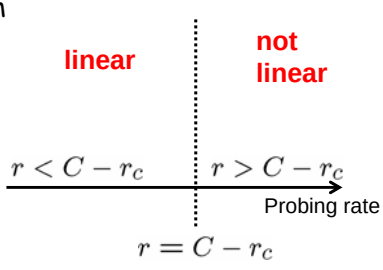
Non-Linear Systems *(or: How about FIFO ?)*

- If we set $S_{fifo}(t) = (C - r_c)t$

… we can describe a FIFO system as

$$D(t) = (rt) * S_{fifo}\ , \text{if } r \le C - r_c$$
$$D(t) \ge (rt) * S_{fifo}\ , \text{if } r > C - r_c$$

- **This means:** FIFO is a linear system if total traffic is below capacity, and non-linear otherwise.
- So, we should not increase rate of probe traffic beyond $r = C - r_c$



Detecting Non-linearity

Backlog convexity criterion

- Suppose that we probe at constant rates $A(t) = rt$
- Legendre transform is always convex
- In a linear system we have:

$$B_{max}(r) = \mathcal{L}_S(r)$$

- If we find that for some rate r

$$B_{max}(r) \neq \text{conv}_{B_{max}}(r)$$

we know that system is not linear

EmuLab Measurements

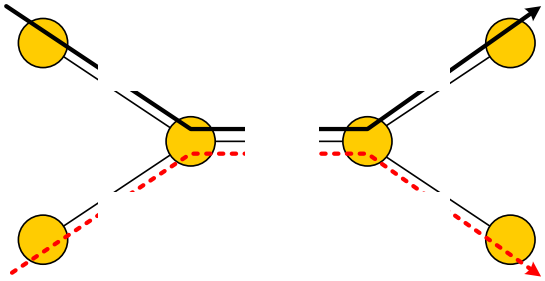
- Emulab is a network testbed at U. Utah
  - can allocate PCs and build a network
  - controlled rates and latencies

Some Questions:

- How well does our theory translate to real networks?
- Does representing available bandwidth by a function (as opposed to a number) have advantages?
- How robust are the methods to changes of the traffic distribution?

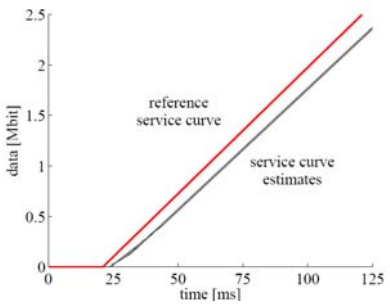
Dumbbell Network

- UDP packets with 1480 bytes (probes) and 800 bytes (cross)
- Cross traffic: 25 Mbps

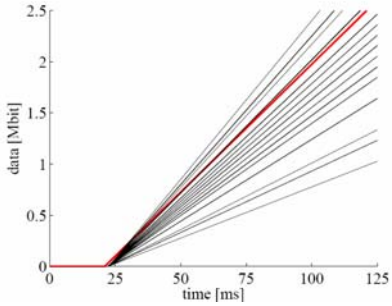


Constant Bit Rate (CBR) Cross Traffic

- Results of 100 repeated estimates of the service curve



Rate Scanning



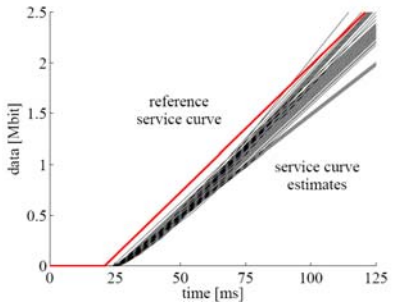
Rate Chirp

- Problems with detecting non-linear region with Ratechirp

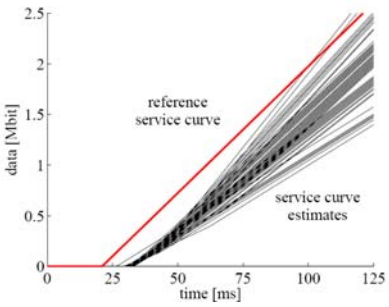
Cross traffic

Rate Scanning: Different Cross Traffic

- Exponential: random interarrivals, low variance
- Pareto: random interarrivals, very high variance



Exponential

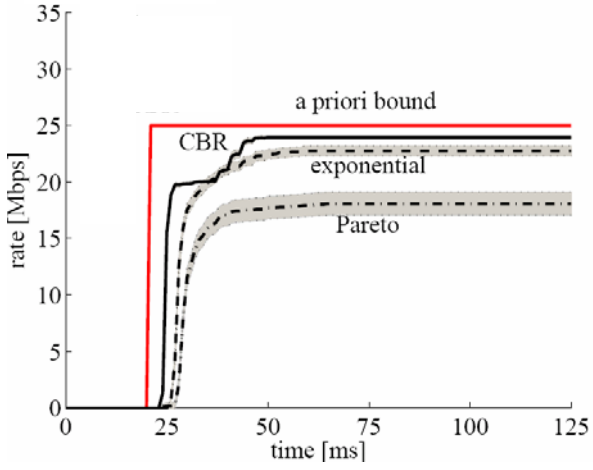


Pareto

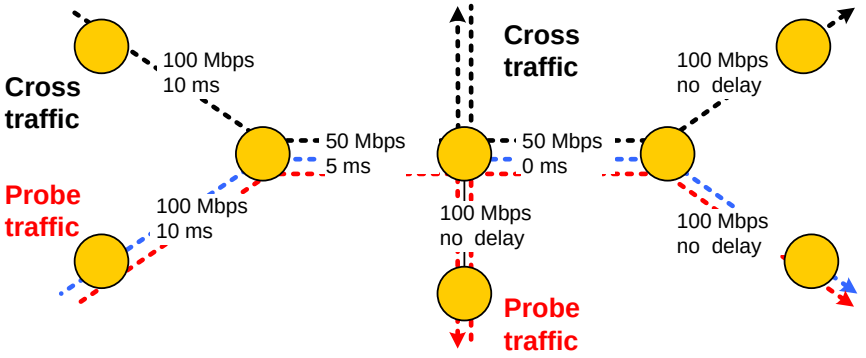
Summary: Rate of the service curve estimate

- Take average of 100 measurements

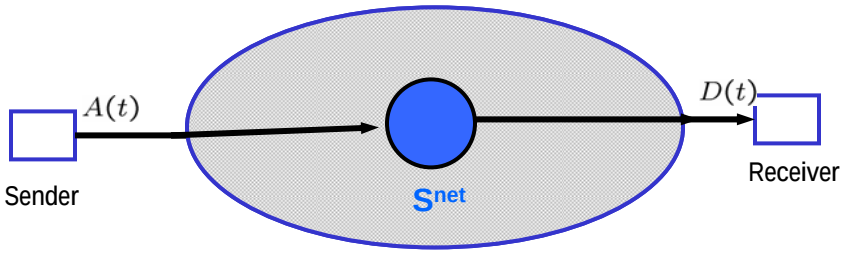
Pathload tool



Network with multiple bottlenecks



Network Service Curves



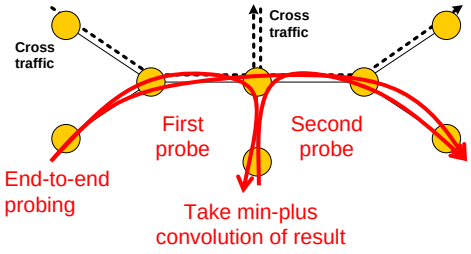
A network with two nodes

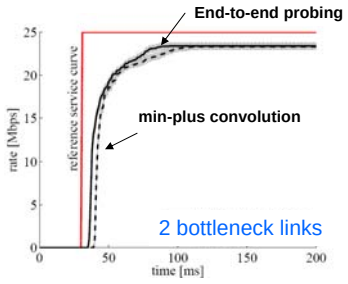
$$D = (A * S_1) * S_2 = A * (S_1 * S_2)$$

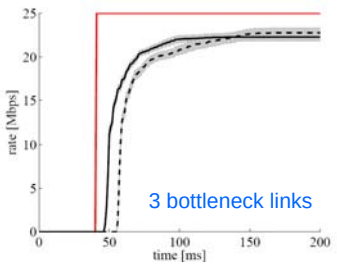
... can be simplified to a single node:

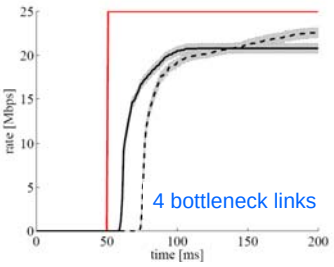
$$S^{net} = S_1 * S_2$$

Network multiple bottlenecks









Conclusions

- Posed available bandwidth estimation as a problem in min-plus linear systems
  - Available bandwidth is a service curve.
- The min-plus linear interpretation can provide a foundational justification of probing schemes:
  - Pathload (rate scan)
  - Path chirp (rate chirp)
- Difficulties with network probing can be related to non-linearities of the underlying system
- Can exploiting min-plus algebra for e2e measurements