The effect of layer-2 store-and-forward devices on per-hop capacity estimation

Ravi S. Prasad[*], Constantinos Dovrolis[*]

and

Bruce A. Mah[†]

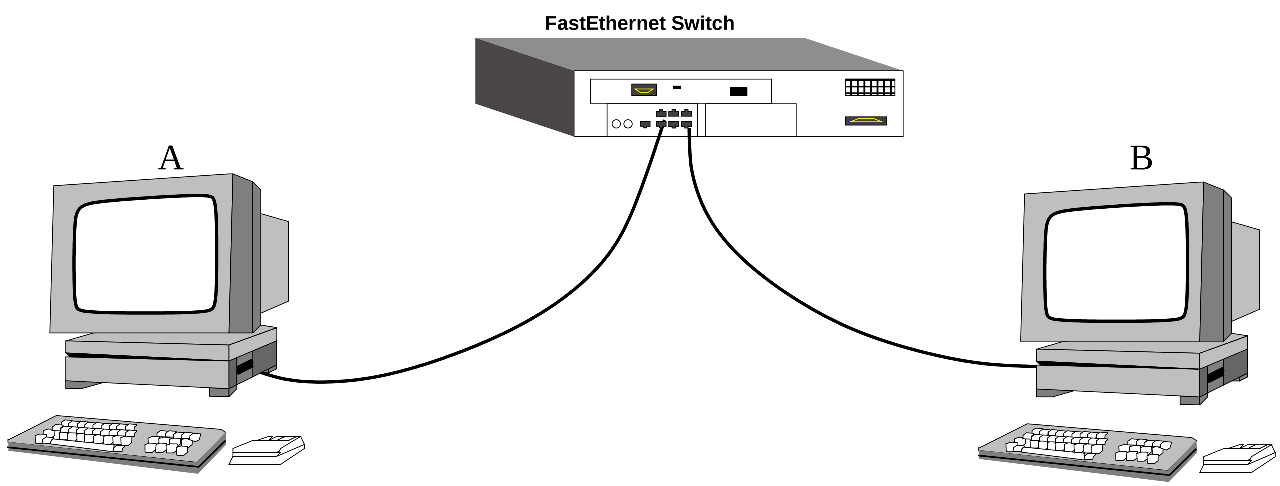
[*]University of Delaware

[†]Packet Design

Capacity estimation tools

Tool	Author	Measurement objective	Methodology
bprobe	Carter	End-to-End Capacity	Packet Pairs
nettimer	Lai	End-to-End Capacity	Packet Pairs
pathrate	Dovrolis	End-to-End Capacity	Packet Pairs & Trains
pathchar	Jacobson	Per-Hop Capacity	Variable Packet Size
clink	Downey	Per-Hop Capacity	Variable Packet Size
pchar	Mah	Per-Hop Capacity	Variable Packet Size
pipechar	Guojun	End-to-End Bottleneck	Packet Trains
cprobe	Carter	End-to-End Avail-BW	Packet Trains
pathload	Jain	End-to-End Avail-BW	Self-Loading Periodic Streams
cat	Allman	Bulk-Transfer-Capacity	Standardized TCP throughput
IPerf	NLANR-DAST	Maximum TCP throughput	Parallel TCP streams

A single-hop path



- A and B both have Fast Ethernet cards.
- What is the capacity from A to B ?

And we get...

Tool	Capacity estimate
pathchar	49.0$\pm$1.5Mbps
clink	47.5$\pm$1.0Mbps
pchar	47.0$\pm$1.0Mbps
pipechar	93.5$\pm$3.0Mbps
pathrate	97.5$\pm$0.5Mbps
bprobe	95.5$\pm$2.0Mbps

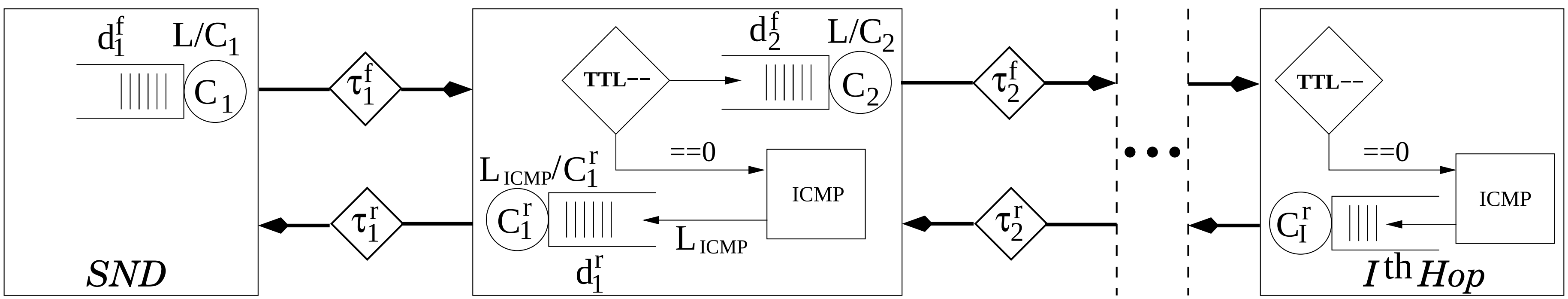
Questions

- What is wrong with pathchar, pchar and clink?
  - The tool?
  - The methodology used?

Overview

- *Variable Packet Size* (VPS) methodology
- Effect of L2 store-and-forward devices
- Some experimental results
- Other sources of error
- Conclusions

VPS methodology



$$\frac{L}{C} : \text{Serialization delay} \tag{1}$$

Components of RTT

VPS tools assume : Minimum RTT for each packet size L, doesn't include any queuing delays.

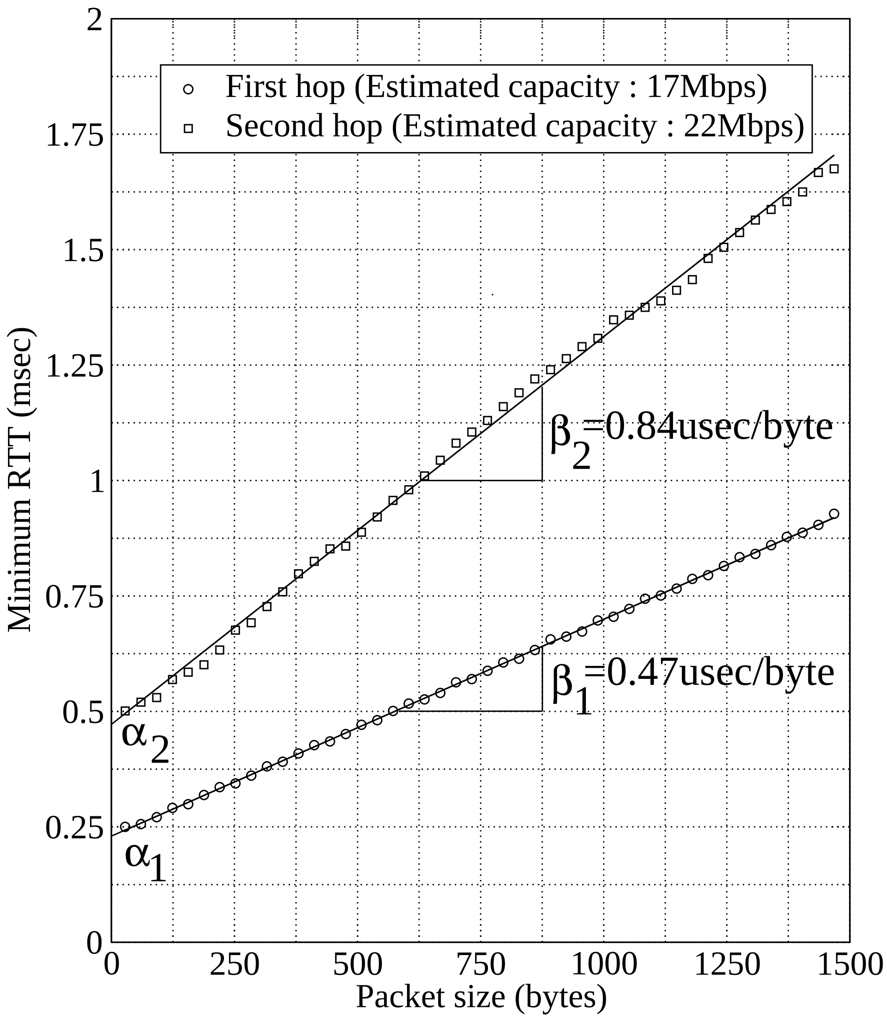
$$T_I(L) = \sum_{i=1}^{I} \left(\frac{L}{C_i} + \tau_i^f + \frac{L_{ICMP}}{C_i^r} + \tau_i^r \right) \tag{2}$$

- Linear fit to the minimum RTT obtained experimentally,

$$T_I(L) = \alpha_I + \beta_I L \tag{3}$$

- $\alpha_I = \sum_{i=1}^{I} \left(\tau_i^f + \frac{L_{ICMP}}{C_i^r} + \tau_i^r \right)$ and $\beta_I = \sum_{i=1}^{I} \frac{1}{C_i}$

An example for 2-hop path

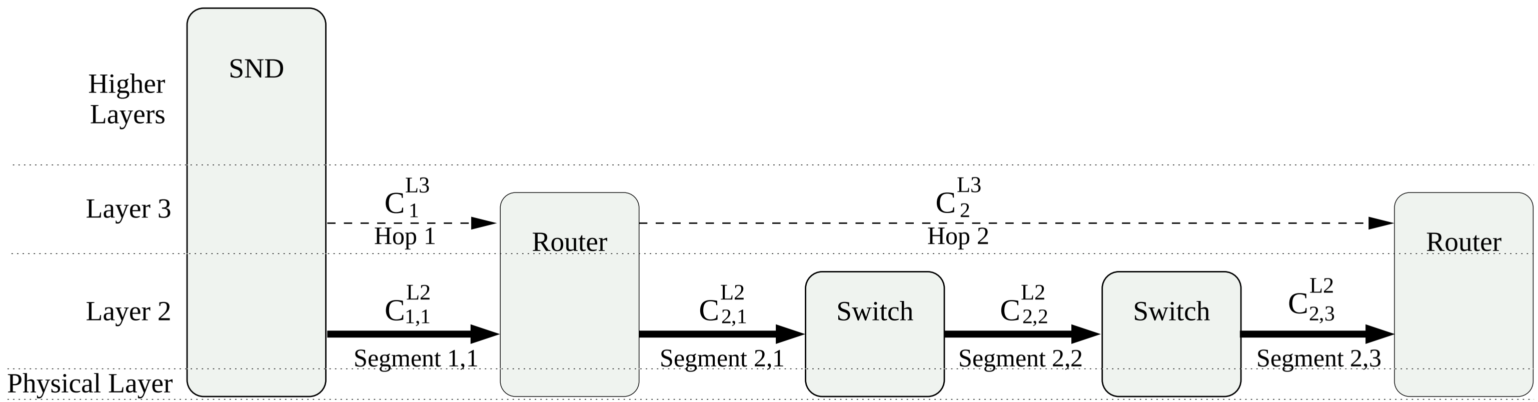


$$C_1 = \frac{1}{\beta_1}$$

$$\beta_2 = \frac{1}{C_2} + \frac{1}{C_1}$$

$$C_2 = \frac{1}{\beta_2 - \beta_1}$$

Links : Layer3 (L3) vs Layer2 (L2)



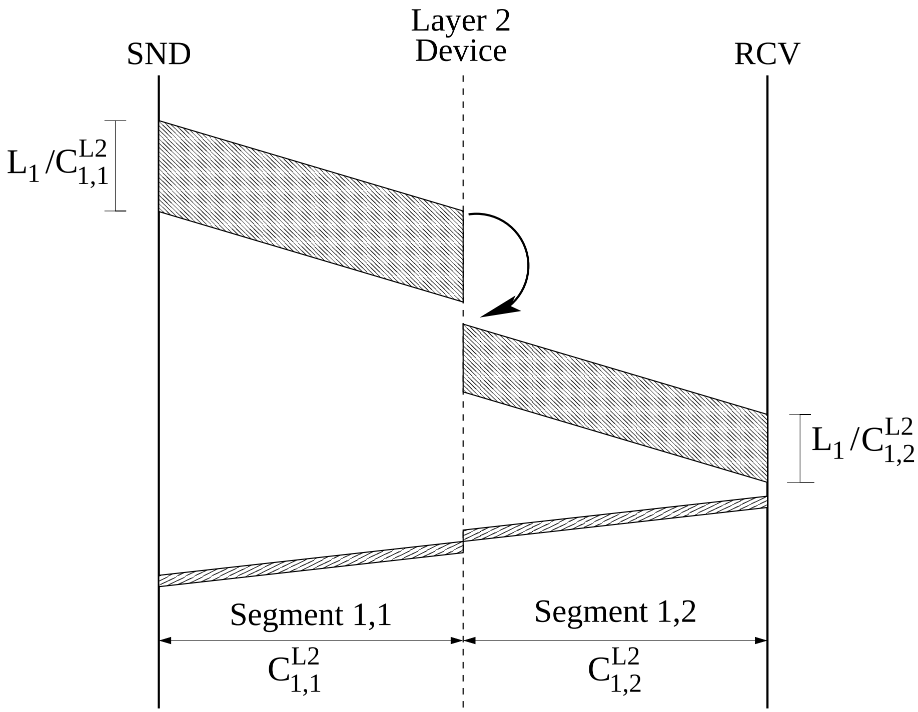
- Each L3 device has an L2 interface
  - Each L3 hop has at-least one L2 segment
- However, If an L3 hop has intermediate L2 devices,
  - May have more than one L2 segment
  - Different L2 segments may have different capacities
  - The capacity of i^{th} L3 hop consisting M_i L2 segments

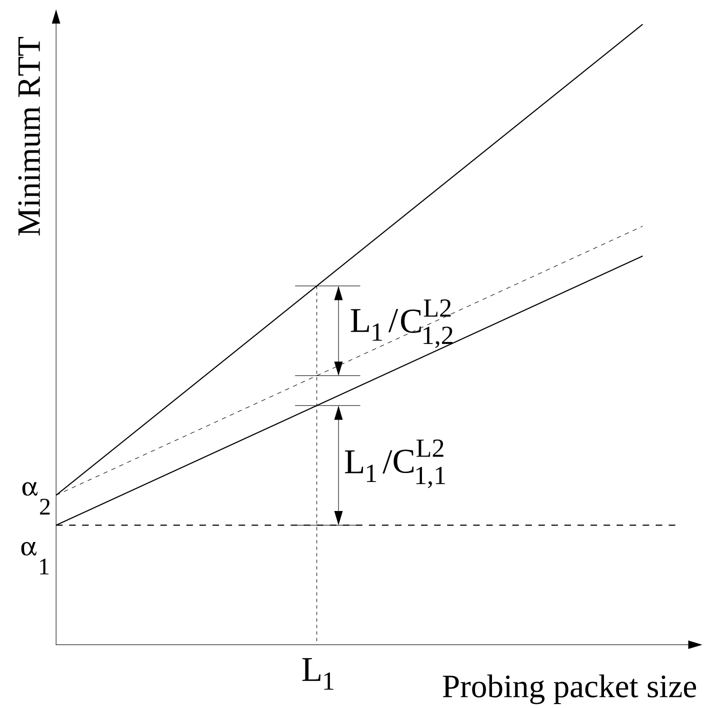
$$C_i^{L3} = \min_{j=1 \ldots M_i} \{C_{i,j}^{L2}\} \tag{4}$$

L2 store-and-forward devices

- Can not be detected by upper layers
  - do not decrease TTL field
  - do not generate ICMP packets
- Affect capacity estimated with VPS tools
  - increase RTT proportional to the packet size
  - change relation between β and capacity

L2 store-and-forward devices & serialization delay





$$\beta_1 = \frac{1}{C^{L2}_{1,1}} + \frac{1}{C^{L2}_{1,2}} \tag{5}$$

$$\hat{C}^{L3}_1 = \frac{1}{\frac{1}{C^{L2}_{1,1}} + \frac{1}{C^{L2}_{1,2}}} \leq C^{L3}_1 \tag{6}$$

Does this error propagate?

- No intermediate L2 store-and-forward devices in the I^{th} hop

$$\beta_I = \frac{1}{C_I^{L3}} + \beta_{I-1} \tag{7}$$

  - Path up to $(I-1)^{th}$ hop may include L2 devices.

- The estimated capacity of the I^{th} hop

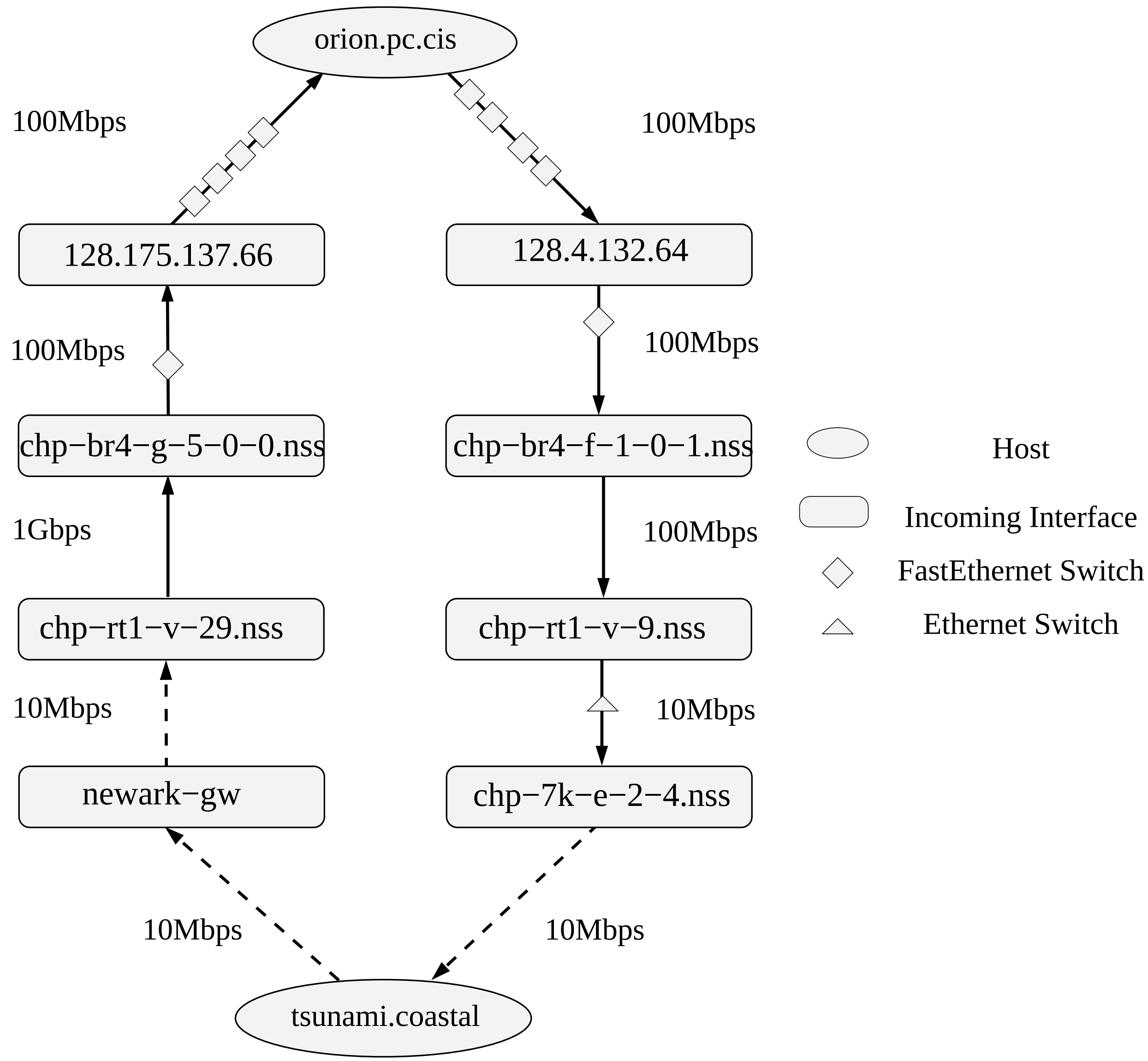
$$\hat{C}_I^{L3} = \frac{1}{\beta_I - \beta_{I-1}} = C_I^{L3} \tag{8}$$

Experimental results : LAN path

- Single-hop path with two L2 segments
- $C_{1,1}^{L2} = C_{1,1}^{L2}$ = 100Mbps
- $\hat{C}_1 = \frac{1}{\frac{1}{100}+\frac{1}{100}}$ = 50 Mbps

Tool	Capacity estimate
pathchar	49.0±1.5Mbps
clink	47.5±1.0Mbps
pchar	47.0±1.0Mbps
Nominal capacity	100.0 Mbps

Campus paths



Experimental results : Campus path 1

L3 hop	capacity	*pathchar*	*clink*	*pchar*
from *orion.ps.cis* to *128.4.132.64*	100Mbps	17.0±0.0	17.0±0.0	17.0±0.4
from *128.4.132.64* to *chp-br4-f-1-0-1.nss*	100Mbps	62.2±7.2	64.7±9.3	62.3±9.1
from *chp-br4-f-1-0-1.nss* to *chp-rt1-v-9.nss*	100Mbps	100.5±15.0	100.3±22.0	101.9±26.0
from *chp-rt1-v-9.nss* to *chp-7k-e-2-4.nss*	10Mbps	5.75±0.15	5.6±0.1	5.7±0.1
from *chp-7k-e-2-4.nss* to *tsunami.coastal*	10Mbps	4.5±0.1	3.7±0.1	6.5±0.6

Table 1: Capacity estimates for the path from *orion.pc.cis* to *tsunami.coastal*.

Experimental results : Campus path 2

L3 hop	capacity	*pathchar*	*clink*	*pchar*
from *tsunami.coastal* to *newark-gw*	10Mbps	4.05 ± 0.05	4.0 ± 0.0	4.0 ± 1.2
from *newark-gw* to *chp-rt1-v-29.nss*	10Mbps	10.5 ± 0.5	10.8 ± 0.4	11.1 ± 0.9
from *chp-rt1-v-29.nss* to *chp-br4-g-5-0-0.nss*	1Gbps	613.33 ± 150.0	414.70 ± 580.0	450.2 ± 110.0
from *chp-br4-g-5-0-0.nss* to *128.175.137.66*	100Mbps	38.3 ± 1.7	39.9 ± 6.0	35.6 ± 8.8
from *128.175.137.66* to *orion.pc.cis*	100Mbps	6.95 ± 0.5	6.1 ± 0.2	21.5 ± 7.8

Table 2: Capacity estimates for the path from *tsunami.coastal* to *orion.pc.cis*.

Experimental results : WAN path 2

L3 hop	capacity	*pathchar*	*clink*	*pchar*
from *abilene-wash-gsr.nss.udel.edu* to *atla-wash.abilene.ucaid.edu*	2480Mbps	$460\pm^{800}_{200}$	$520\pm^{680}_{410}$	$1031\pm^{12600}_{800}$

Table 3: Capacity estimates for an Abilene OC-48 core link.

Other sources of errors in VPS tools

- Traffic load
- Non-zero queuing delays
- Limited clock resolution
- Error propagation from the previous hop
- ICMP generation latency?

Traffic load

- High network traffic $\Longrightarrow$ High probability of observing queuing delays.
- The probability of not observing any queuing for a packet in i^{th} link

$$P_i = (1 - \rho_i) \tag{9}$$

where ρ_i is the utilization of the i^{th} link.

- The probability of not observing any queuing in I hops by at least 1 out of K packets

$$P(I, K) = 1 - \left[1 - \prod_{i=1}^{I} P_i\right]^K \tag{10}$$

Traffic load (contd.)

Path length I	ρ=0.2	ρ=0.4	ρ=0.6	ρ=0.8
1	2	3	5	11
2	3	6	14	57
4	5	17	89	1438
6	8	49	562	35977
8	13	136	3515	899447
10	21	380	21959	22486182

Table 4: minimum number of packets K so that $P(I, K) \geq 0.9$.

- VPS tools use same number of probes (default 32) for each hop
  - too few for remote hops under heavy load

Limited clock resolution

- If clock resolution is 2σ,

$$T_1 = \alpha + L_1\beta \pm \sigma \tag{11}$$

$$T_2 = \alpha + L_2\beta \pm \sigma \tag{12}$$

- The estimated capacity would be

$$\hat{C} = \frac{C}{1 \pm \frac{2\sigma C}{\Delta L}} \tag{13}$$

- For OC-48, 1μsec resolution and $\Delta L = 1500B$ can result in 25% error.

Error propagation from previous hop

- Any probabilistic error will propagate to next hop.
  - if measured RTT slopes are

$$\hat{\beta}_1 = \beta_1(1+\epsilon_1),\ \hat{\beta}_2 = \beta_1(1+\epsilon_2) + \beta_2 \tag{14}$$

estimated capacity of second hop

$$\hat{C}_2 = \frac{1}{\hat{\beta}_2 - \hat{\beta}_1} = \frac{C_2}{1 + (\epsilon_2 - \epsilon_1)\frac{C_2}{C_1}} \tag{15}$$

- Error in a Gigabit hop after an Ethernet hop gets magnified by a factor of 100

ICMP generation latency

- Latency of ICMP generation
  - not related to probing packet size
  - doesn't affect RTT slope measurement
- Variation of these latencies may affect RTT slope
- Minimum ICMP generation latency in high traffic load
  - large number of probes required to catch this
  - effect is similar to that of non-zero queuing delays

Conclusions

- Methodology used by VPS tools can introduce large errors
- Errors due to L2 store-and forward devices
  - consistent and hard to identify
- Probabilistic errors
  - can be detected by repetitive run of the tools

Thank you!

Non-zero queuing delays

- Minimum RTT measurement for packet sizes L_1 and L_2

$$T_1 = \alpha + L_1\beta + \frac{q_1}{C} \tag{16}$$

$$T_2 = \alpha + L_2\beta + \frac{q_2}{C} \tag{17}$$

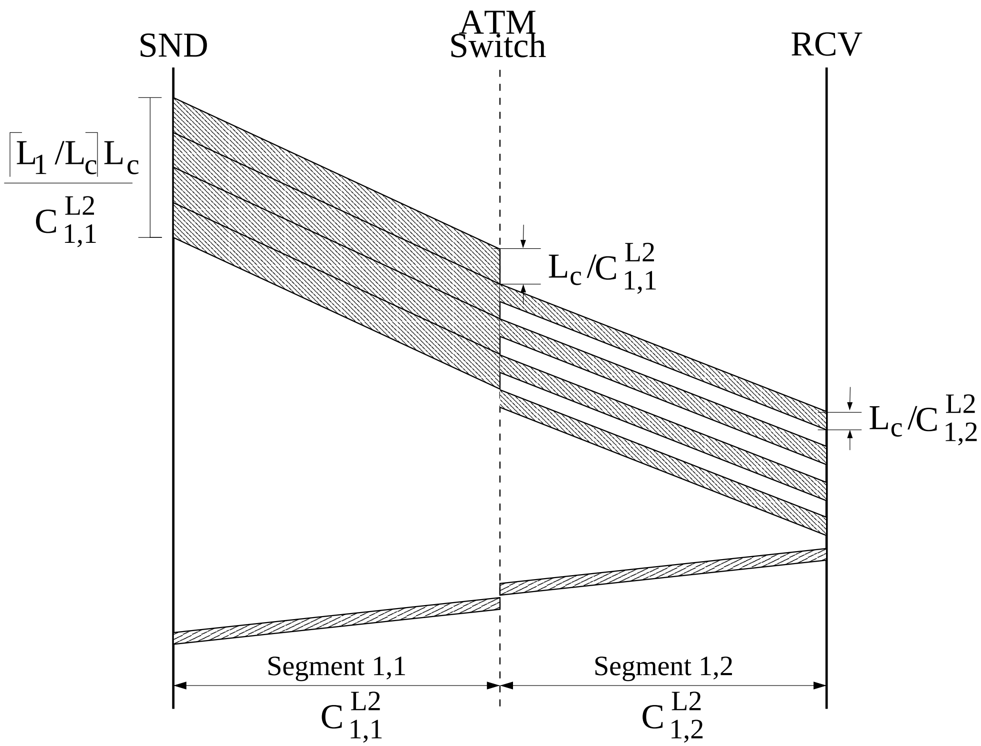
where q_1 and q_2 are minimum queue sizes

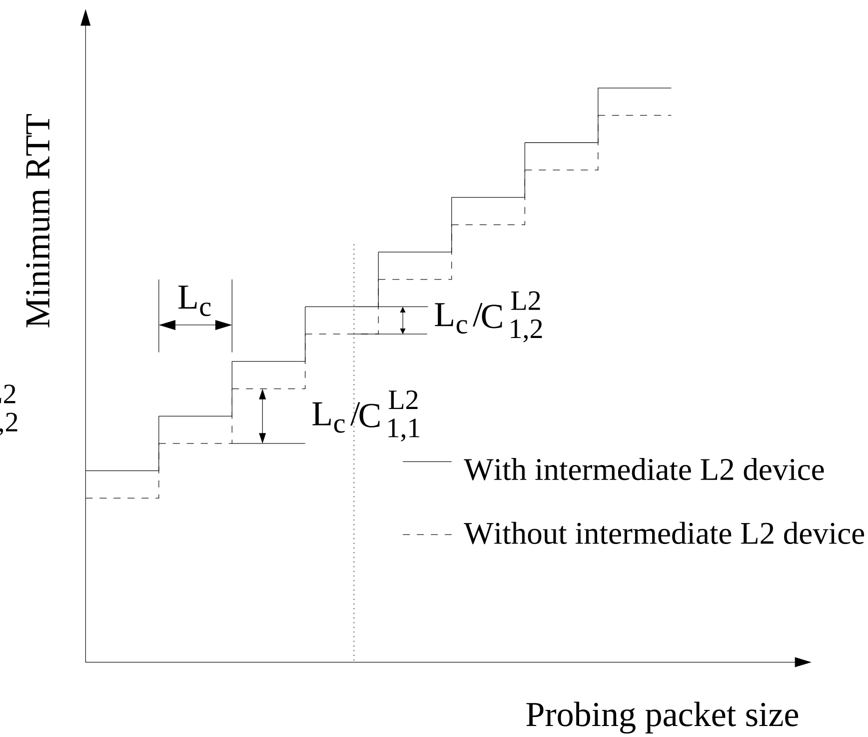
- Estimated capacity will be

$$\hat{C} = \frac{\Delta L}{\Delta T} = \frac{C}{\left(1 + \frac{\Delta q}{\Delta L}\right)} \tag{18}$$

- Non-zero queuing delays cause a multiplicative error in capacity estimate

Effect of ATM switches on RTT





Experimental results : WAN path 1

L3 hop	capacity	*pathchar*	*clink*	*pchar*
from *chp-br4-f-1-0-1.nss.udel.edu* to *delaware-gw-f2-0.voicenet.net*	45Mbps	30.5$\pm$3.5	30.3$\pm$5.6	28.3$\pm$5.6
from *delaware-gw-f2-0.voicenet.net* to *delaware2-gw-H2-0-T3.voicenet.net*	45Mbps	44.6$\pm$20.0	48.0$\pm$1.6	45.2$\pm$10.0

Table 5: Capacity estimates for the Univ-Delaware access link to VoiceNet, and for a VoiceNet edge link.