Spectroscopy of Traceroute Delays

Andre Broido, Young Hyun and kc claffy

C A I D A

CAIDA / SDSC / UCSD

http://www.caida.org

PAM

Passive and Active Measurement

Boston

2005-04-01

Plan

Introduction
Router ICMP generation
A glimpse of the results
Details
Conclusion

This version is updated w.r.t. proceedings

Common-sense assumptions

about traceroute generation delays
at the routers

- Min delay is linear in size, d = d0 + x/C
- The constant C is the inbound link capacity
- Delay over linear fn. is due to cross-traffic
  (in the absense of CT, delay = min = linear fn)
- Delays are i.i.d. – independent identically distributed random variables
- Delay is independent of payload content

  all published work assumes these properties

We found all of them violated

Traceroute RTTs could be used to:

- construct router and PoP-level Internet maps (instead of IP address level maps)
- find latencies/capacities of remote links for realistic models/simulations
- user-level path diagnosis (Mahajan e.a.)
- fingerprint routers

See ”Reverse engineering the Internet”, other papers by Spring e.a. for more inspiration

History

- 1997: Skitter collects topology and RTT by running traceroutes to 30k destinations
- 2000: Skitter IPv4 list – 10x coverage
- 2003: Intermediate RTTs – 20x more data.

Cannot make sense of all this RTT data

To understand RTT we need:

Precision timestamping

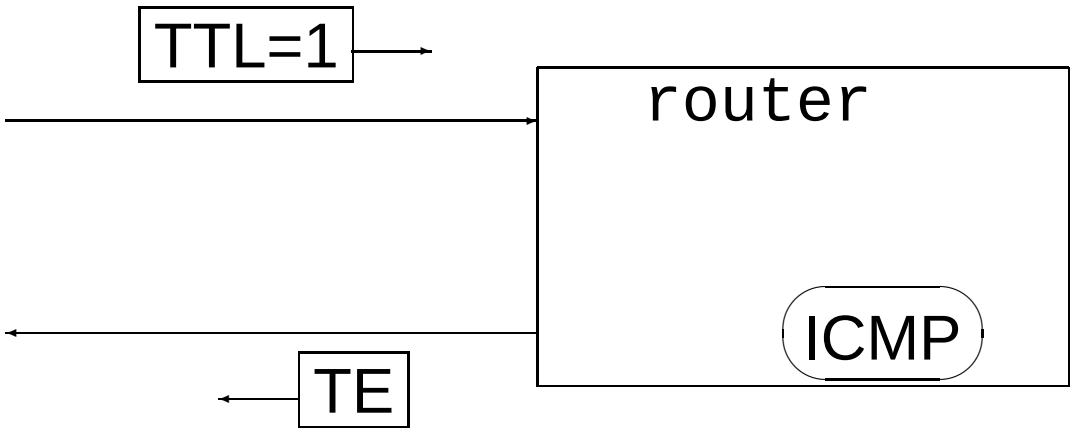
Delay summands

We study delay of one packet at one router:

ICMP TimeExceeded generation delay

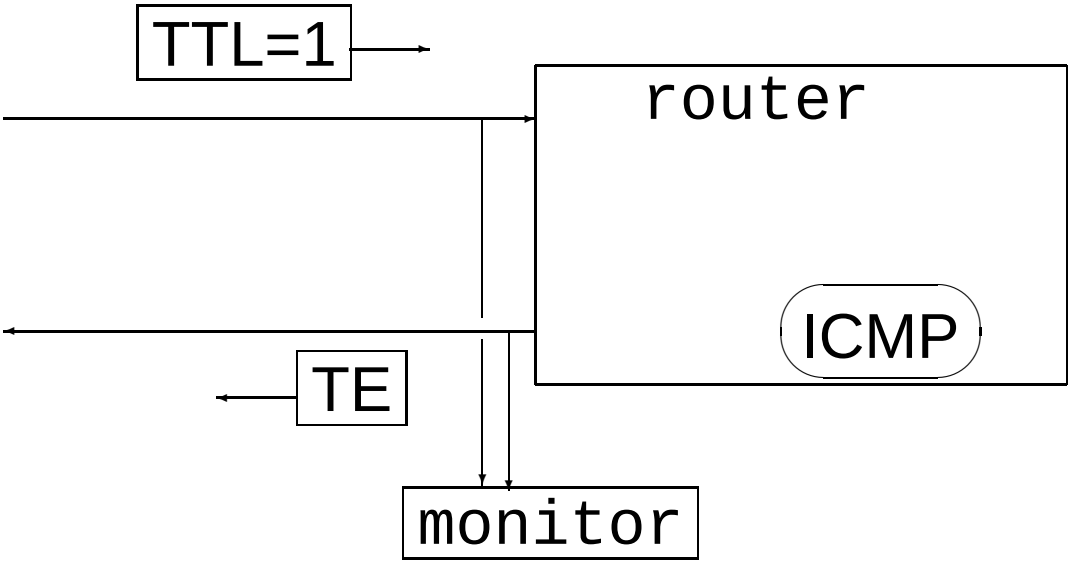
Isolate single router contribution
for future synthesis of the whole path delay

Router ICMP generation



1. IP packet with TTL=1 enters the router
2. TTL-1 = 0
3. ICMP Time Exceeded (TE) generated
4. TE message leaves the router

We want to measure



1. IP packet with TTL=1 enters the router
2. TTL-1 = 0
3. ICMP Time Exceeded (TE) generated
4. TE message leaves the router

how long does it take?
how packet size affects ICMP delay?

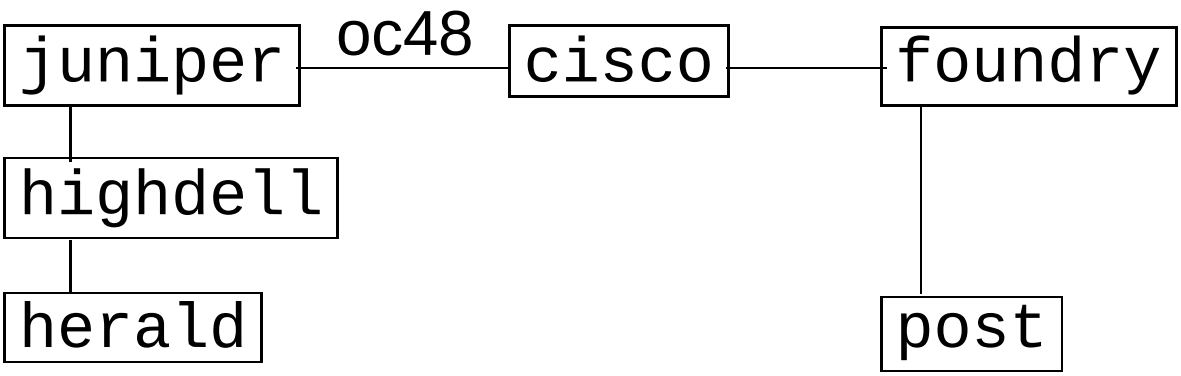
Method

- Traceroute between end hosts
- Make sure there is no cross traffic
- Send one packet at a time
- Capture probes/responses at each router as they enter and exit router's interfaces (line cards)
- Measure the ICMP Time Exceeded delay by timestamp difference

Advantages: Fully controlled setup
Many causes and effects are observable

Caveat: need to know how timestamping is done (talk to Stephen Donnelly if you use Dag cards)

Lab diagram



Equipment (clockwise):
Juniper M20 router
Cisco 12008 router
Foundry BigIron 8000 router/switch
Links: oc48 (Juniper to Cisco)
GigabitEthernet (all other links)

Clarification

One packet at a time means:

- We wait for a packet to come back to sender
- At any given time, there is only one packet in the whole test network
- The router works on one packet or idles

No one made this experiment,
everyone ”knew” the result

Published work:

operational traffic (Papagiannaki e.a, Hohn e.a.)
100% utilization (LightReading)
forwarding delays (Bovy e.a.)
remote routers (Govindan, Paxson)

Variable packet size method (Van J)

reasoning:

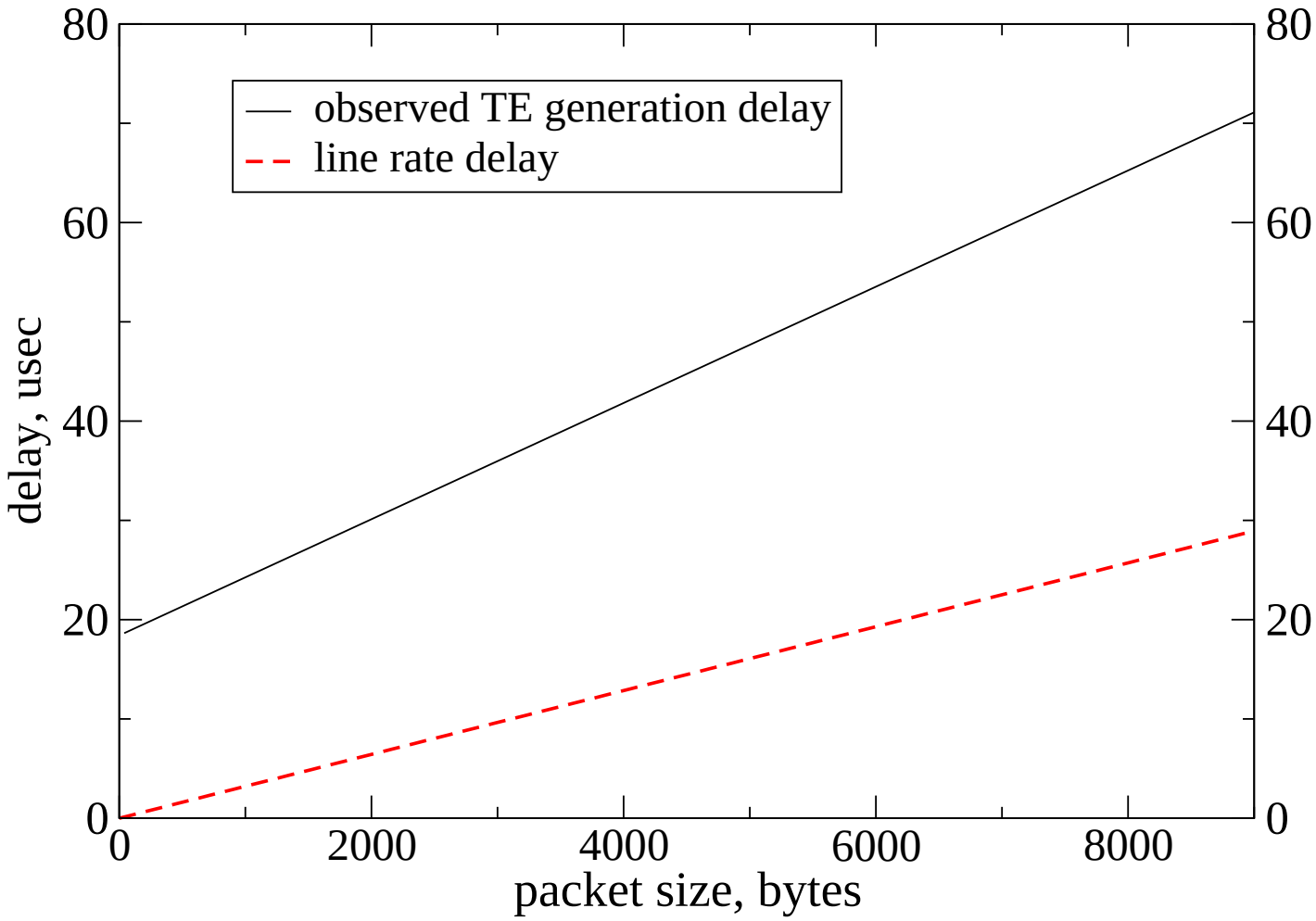
- The only component of delay dependent on size is packet input (deserialization)
- it takes constant time to generate small, fixed size ICMP packet
- The router ICMP delay must grow at link rate i.e. packet size divided by link capacity (x/C)
- e.g. as 1 ns/bit for gigE

this is how pathchar and related tools estimate link capacities

The Controversy

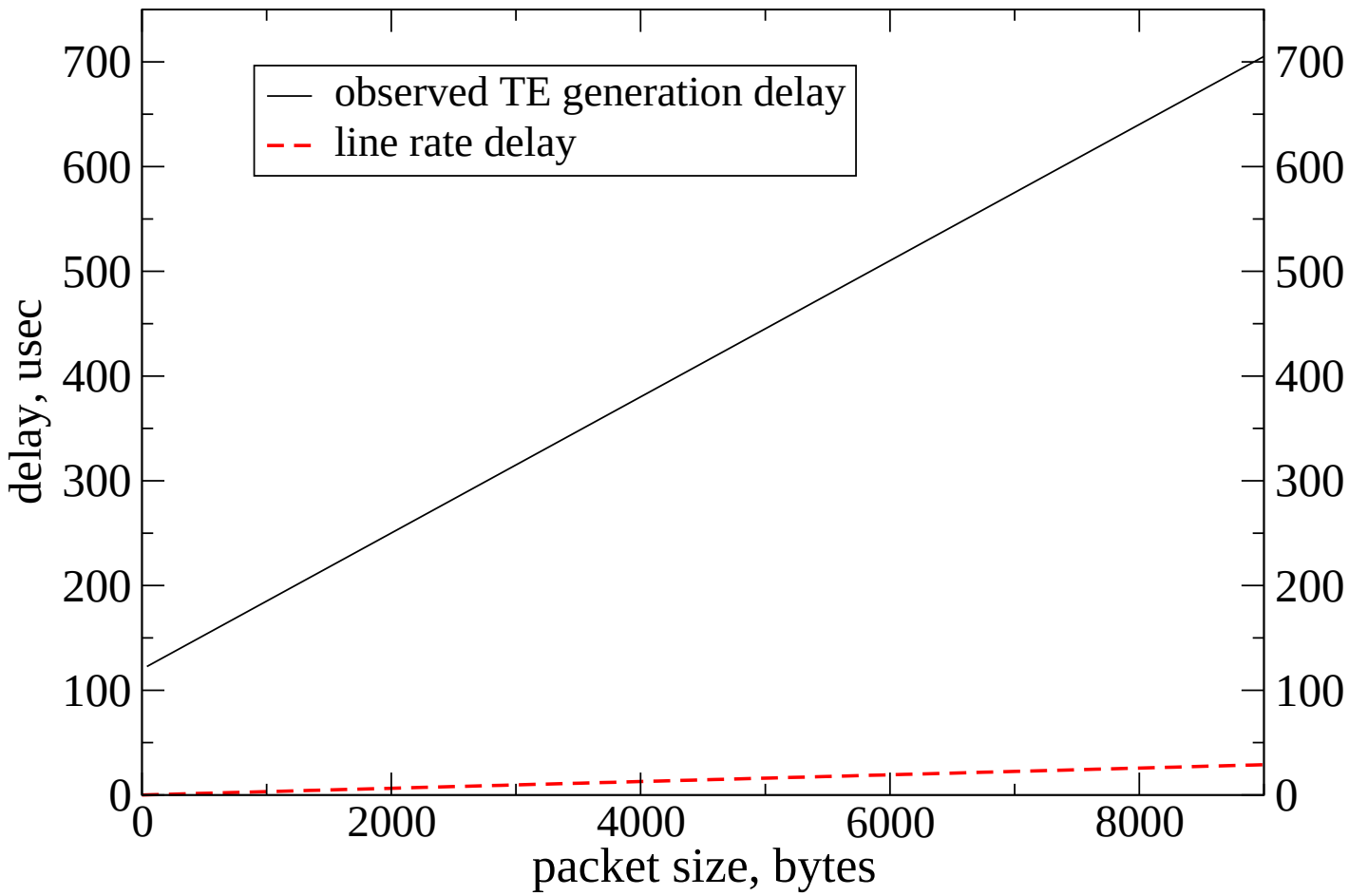
- Van's argument sounds reasonable
- However, pathchar measures 114 Mbps on Juniper's 2.5 Gbps link, an error of 20x
- Challenge: to understand why it's wrong

(Prasad e.a.: Layer 2 switches)



X axis: packet size, 40-9000 bytes
Y axis: min.Time Exceeded delay

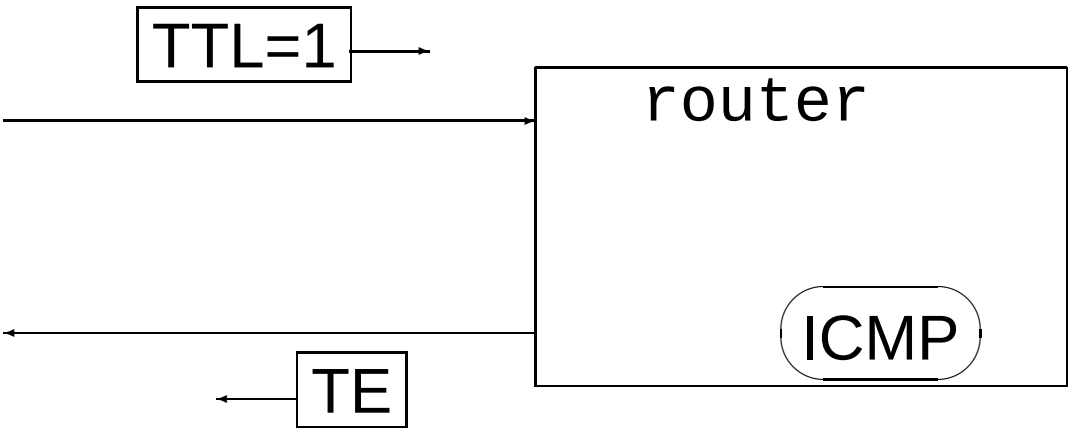
Cisco 12000: TE generation is 80% slower than link rate



X axis: packet size, 40-9000 bytes
Y axis: min.Time Exceeded delay

Juniper M20: TE generation is 20 times slower than link rate

Recall ICMP generation



1. IP packet with TTL=1 enters the router
2. TTL-1 = 0
3. ICMP Time Exceeded (TE) generated
4. TE message leaves the router

Proposed explanation

- Packet needs to move inside the router before ICMP generation can occur
- ICMP data path can be provisioned at lower-than-link rate
- Deliberate rate limiting (e.g. leaky buckets) can be part of the design

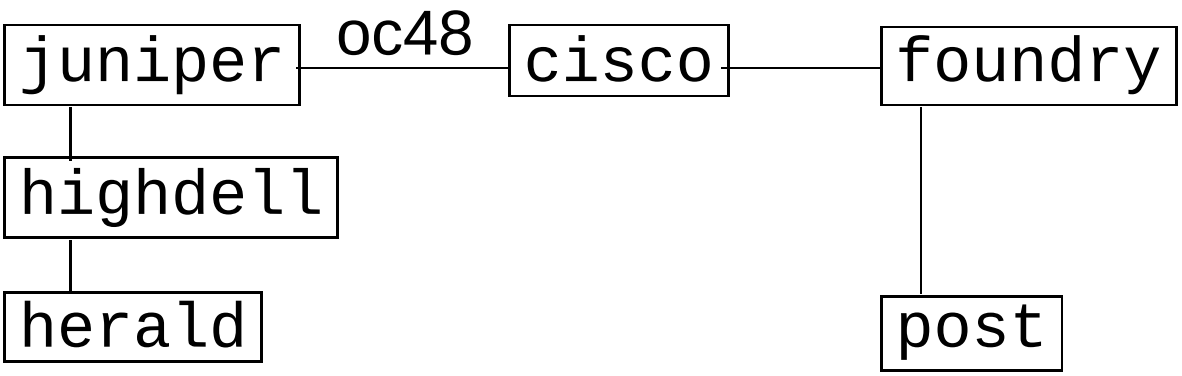
We measure ICMP box speed, not link speed

Details

Data analysis

- Study size dependence:
- assumed model: for packet size x,
- t = ax + b + (positive residual delay)
- (residual is *not* queueing)
- Under these assumptions
- ax+b = lower bound for t that can be found by linear programming
- (R.Graham’s convex hull algorithm, 1972, see Moon e.a.)

Experiment



Equipment (clockwise):
IBM eServer `herald`, FreeBSD 4.8
Dell PowerConnect 5212 switch
Juniper M20 router
Cisco 12008 router
Foundry BigIron 8000 router/switch
IBM eServer `post`
Links: oc48 (Juniper to Cisco)
GigabitEthernet (all other links)
more FreeBSD and Linux boxes

Varied parameters

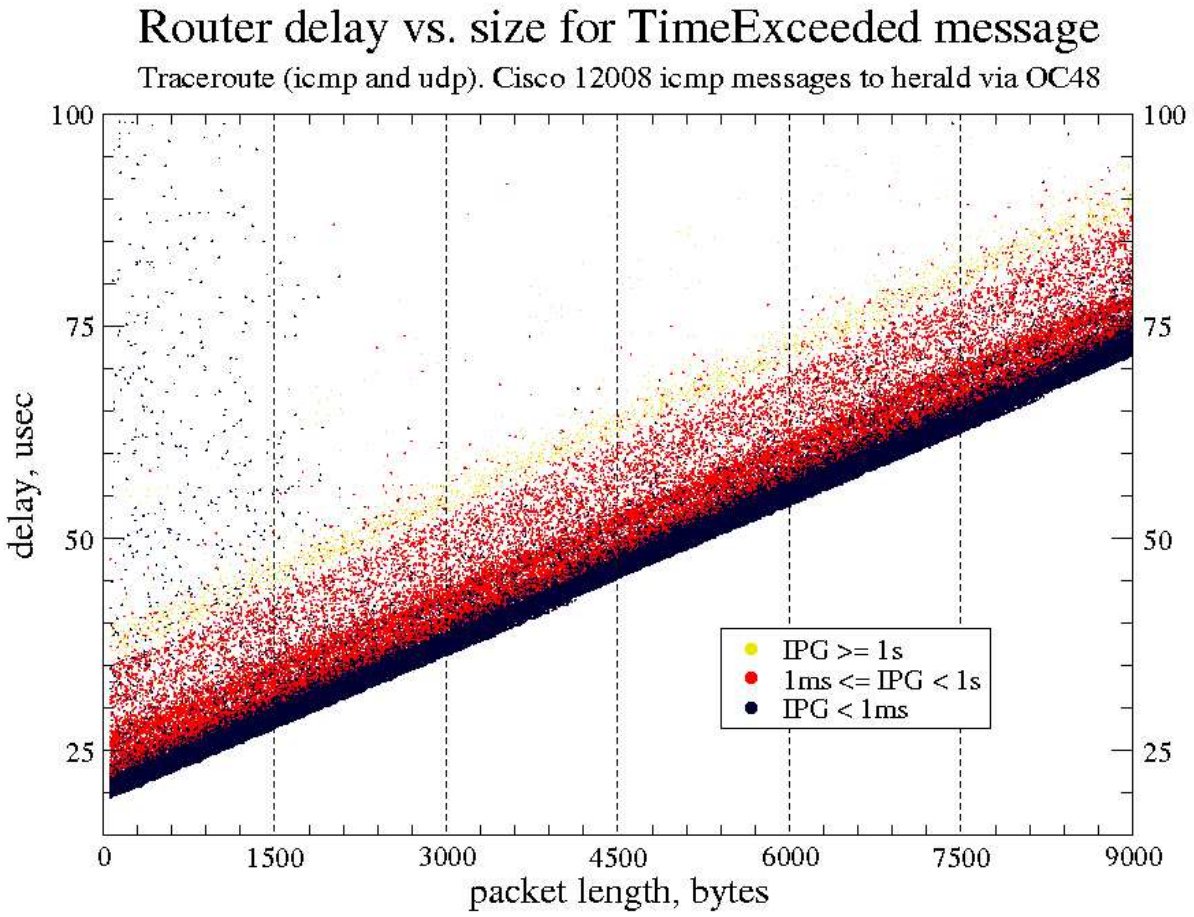
- Three router vendors
- OC48 vs. GigE line cards
- Packet sizes (full range up to 9000)
- Interprobe gap (micro/milli/whole seconds)

Observed

- Extra latency caused by inter-probe gap perhaps cache flushing/warm-up, 20-30 usec (observed for all routers for 2 sec spacing)
- Stepwise growth (64 byte cells) for Juniper
- Juniper delays some closely spaced packets by 9-10 ms (enforces 100 pps)
- Non-queueing residual delays always present
  - 95 percentile at 3-6 usec - same order as 1500b of cross traffic (5 usec)
  - 99% under 20-30 usec
  - max at 0.3 ms (Cisco),
  - 1.5 ms (Foundry),
  - 11 ms (Juniper)

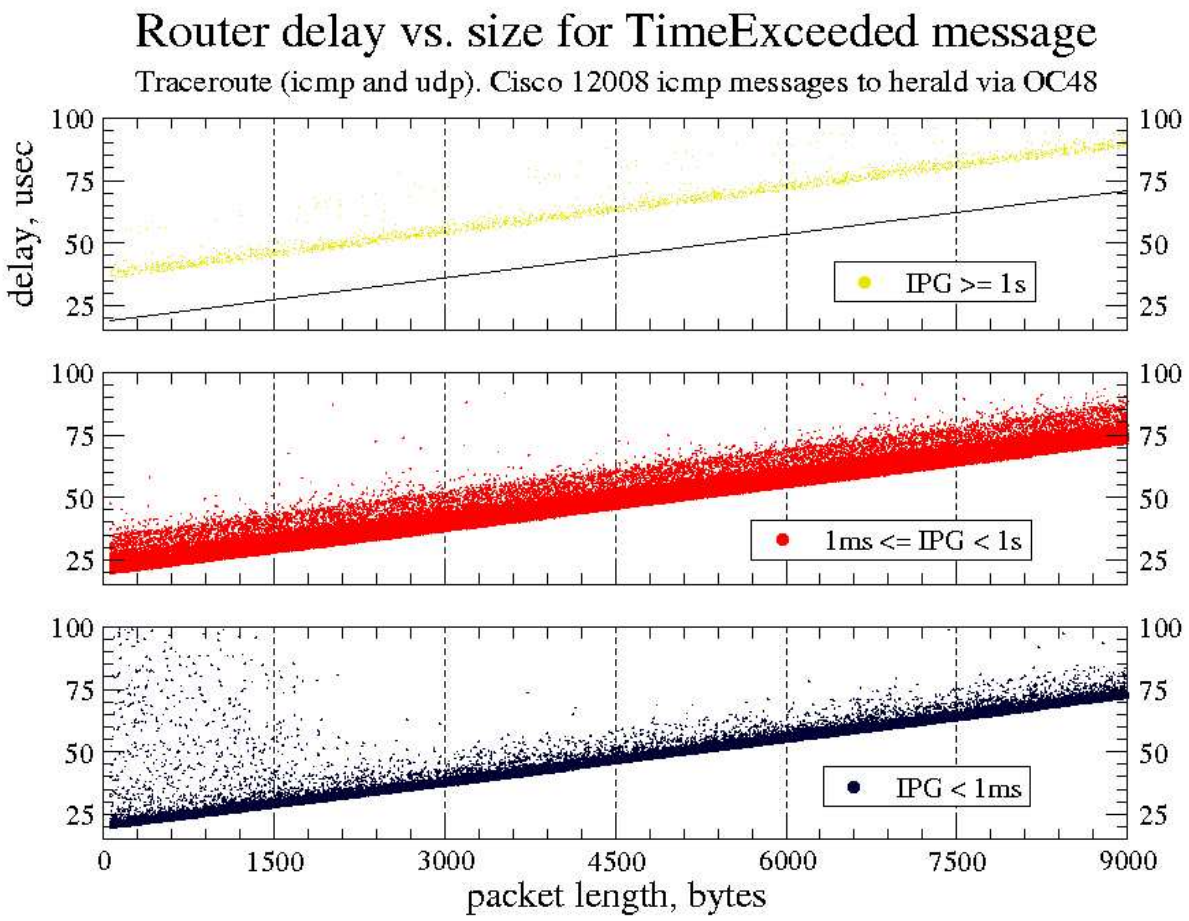
ICMP gen.rate not equal to input link capacity

Cisco OC48 TE generation delay



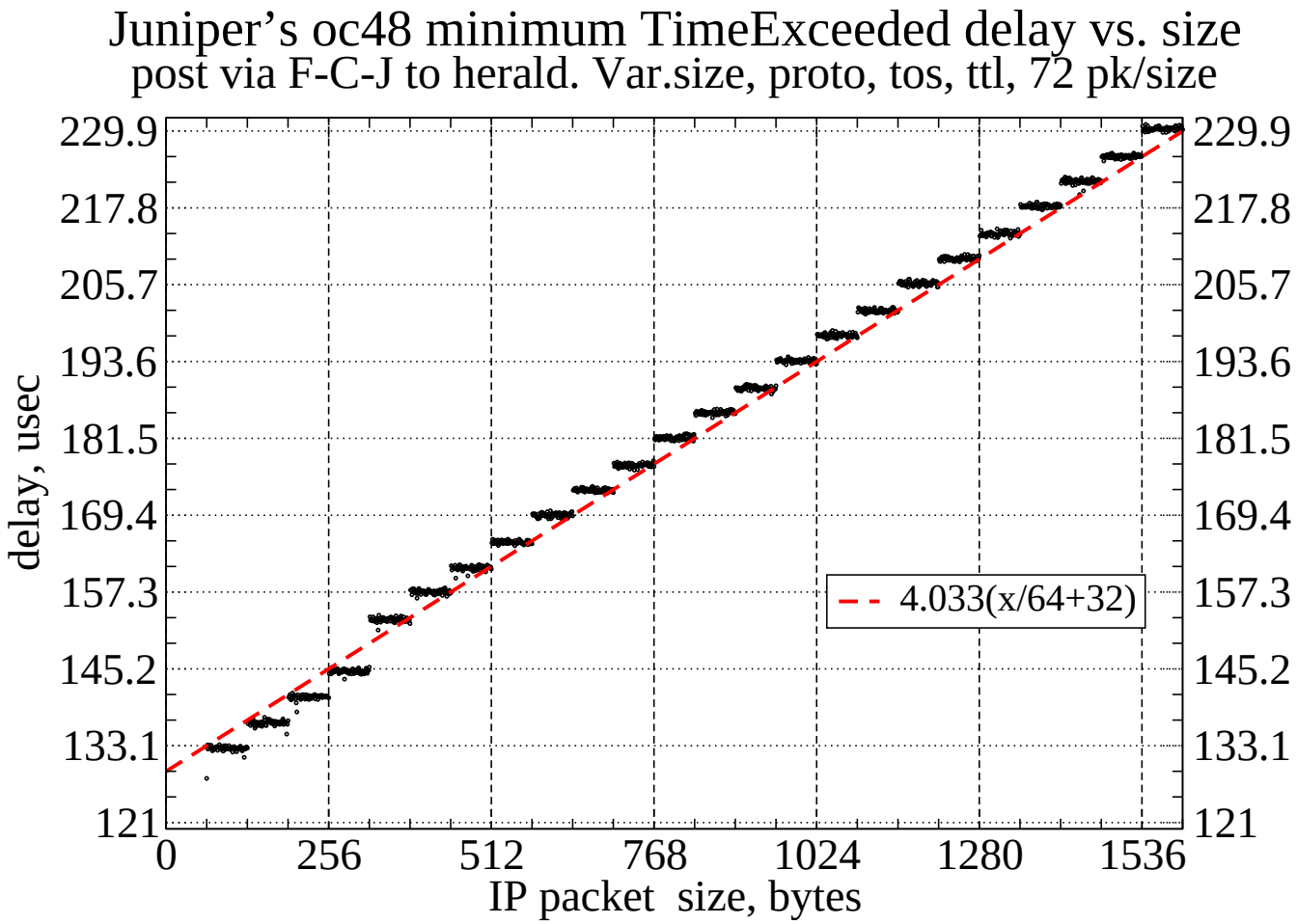
X axis: packet size, 40-9000 bytes
Y axis: Time Exceeded delay
Interprobe gap: 2 sec; 10-20 ms; under 1 ms

Cisco OC48 TE delay by interprobe gap



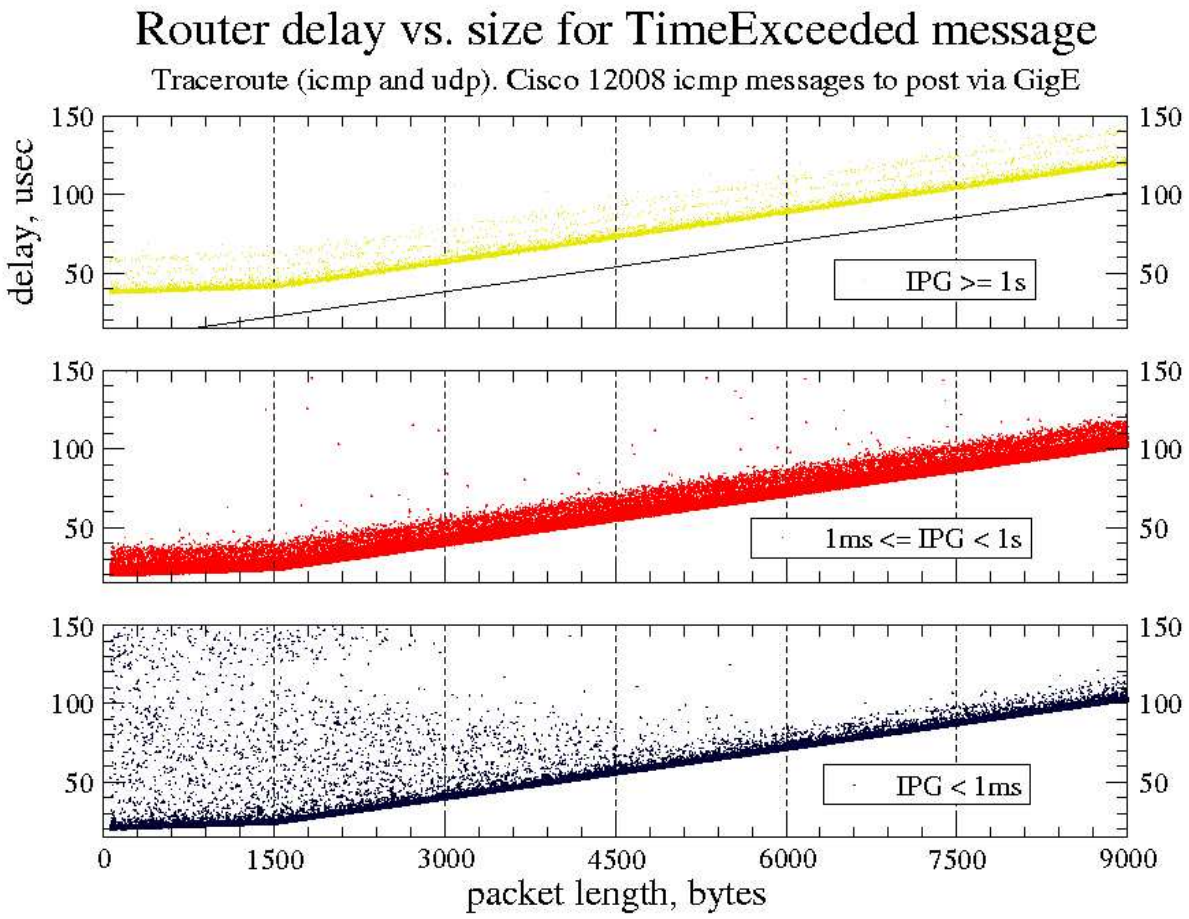
X axis: packet size, 40-9000 bytes
Y axis: Time Exceeded delay
Interprobe gap: 2 sec; 10-20 ms; under 1 ms

Min TE delay, Juniper OC48



X axis: packet size, 40-9000 bytes
Y axis: min.Time Exceeded delay
(64 byte cells)

Cisco Gigabit Ethernet TE delay



X axis: packet size, 40-9000 bytes
Y axis: Time Exceeded delay, Cisco gigE
Interprobe gap: 2 sec; 10-20 ms; under 1 ms

Can slope change be Dag card related?

- Slope changes occur on Cisco and Foundry gigE interfaces at around 1500 bytes
- Stephen Donnelly (Endace):
  Timestamps are at byte min(x, 1540)
  in firmware verisions released before
  October 2004

  In 2.5.2 and subsequent releases,
  Dag GE card timestamps first 4 bytes

Conclusions

- Routers don’t generate ICMP at line rate
- Observed non-queueing residual delay in 10-100 usec range
- Residual delays are large enough to upset some spectroscopy tools
- Residual delay clusters in bands
- Some bands caused by inter-probe gaps
- Know thy capture cards

Future work

- EchoReply
- PortUnreachable
- Forwarding delay
- Loaded routers (with cross-traffic)
- Continuous IAT range, 200 usec-2 sec

Our related work on spectroscopy

- Radon tranform for ATM rate evaluation
- DSL and cable modems' rates
- OS fingerprinting by DNS updates
- Remote device fingerprinting

Acknowledgements:

- Stephen Donnelly
- Dan Andersen
- UCLA IPAM
- Ryan King
- Yoshi Kohno
- Margaret Murray
- Evi Nemeth
- Robert Nowak
- David Moore