ICIM:
An Inline Network Measurement Mechanism for Highspeed Networks

Osaka University
Cao Le Thanh Man, Go Hasegawa and Masayuki Murata
mlt-cao@ist.osaka-u.ac.jp

Outline

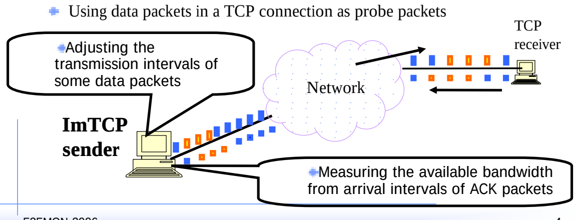
- Background
  - Available bandwidth & bandwidth measurement
  - Inline network measurement
- Measurement in high-speed networks
  - Problems of existing tools
- Proposed method: ICIM
  - Interrupt Coalescence –aware inline measurement
  - Works well in high-speed networks
- Simulation results

Available bandwidth & bandwidth measurement

- Available bandwidth of an E2E network path
  - Available bandwidth = Capacity – used bandwidth
  - Important key in network adaptive control: server selection, routing in overlay networks…
- 2 measurement approaches: Active & passive
  - Active approach
    - Fast
    - High accuracy
    - Extra load to the network
      - Extra traffic for probing
        - Pathload: 2.5 ~ 10MB of probe traffic for one measurement
        - IGI: 130 KB, Spruce: 300 KB

Our approach

- **ImTCP** [E2EMON 2004] *Inline network measurement TCP*
- Inline network measurement
  - **Performing active measurement without probe traffic**
  - Using data packets in a TCP connection as probe packets

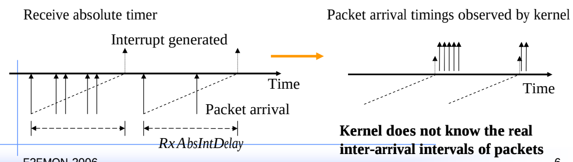


Focus of this research

- **Measurement in high-speed network**
  - 1Gbps or higher network
  - Popular nowadays: the need for measuring, observing them are emerging
- **Why existing tools (including ImTCP) can not work?**
  - **Limitation in packet pacing**
    - High-speed network measurement requires small packet sending/receiving intervals: 1Gbp => 0.012ms (packet size 1500B)
    - For a general-purpose machine, such small intervals cause high CPU overhead
  - **Effects of Interrupt Coalescence**
    - Inter-arrival intervals of packets are changed and is not visible to the measurement tools

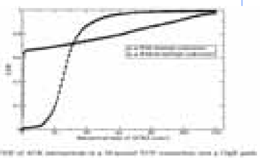
Interrupt Coalescence (IC)

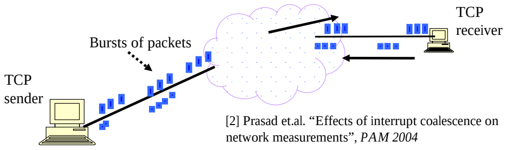
- Deployed in most high-bandwidth Network Interface Cards (NICs)
- Multiple packets are grouped and passed to the kernel in a single interrupt
  - Absolute timer (effective in high speed data transmission)
    [1]: Intel, "Interrupt moderation using Intel Gigabit Ethernet Controllers" http://www.intel.com/design/network/applnots/ap450.pdf (2003)



Effect of Interrupt Coalescence on TCP

- **Bursty transmission of TCP packets**
  - The bursty arrival of packets at the receiver causes bursty transmission of ACKs, and subsequently bursty transmission of more data packets from the sender.
  - With IC, 65% of ACKs arrive with intervals of less than 1µs [2]





[2] Prasad et.al. "Effects of interrupt coalescence on network measurements", *PAM 2004*

Effect of Interrupt Coalescence on TCP

- **Bursty transmission of TCP packets**
  - The bursty arrival of packets at the receiver causes bursty transmission of ACKs, and subsequently

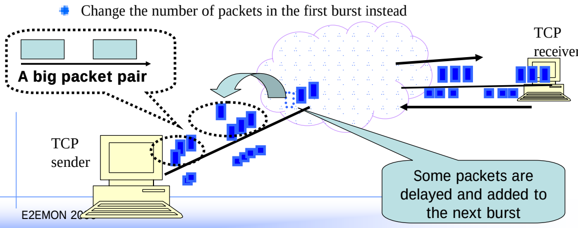
Busty transmission of TCP is the important key to our new inline measurement method

[2] Prasad et.al. "Effects of interrupt coalescence on network measurements", *PAM 2004*

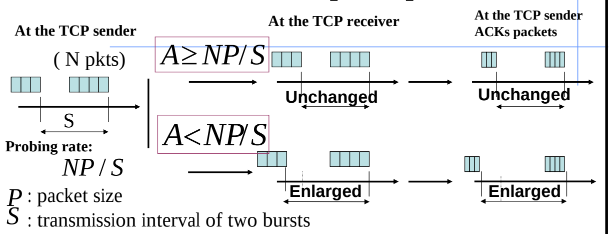
Proposed method: ICIM

Interrupt Coalescence-aware Inline Measurement

- Objective: New inline measurement method
  - Measuring available bandwidth when IC is enabled
  - Not requiring packet pacing
- Basic idea: **Packet-burst pair** is consider as **(big) packet pair**
  - IC automatically forms the bursts of packets in a TCP connection
    - No need for pacing
  - How to set the burst transmission intervals?
    - Change the number of packets in the first burst instead



Measurement principle



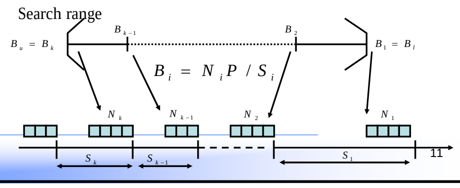
P : packet size

S : transmission interval of two bursts

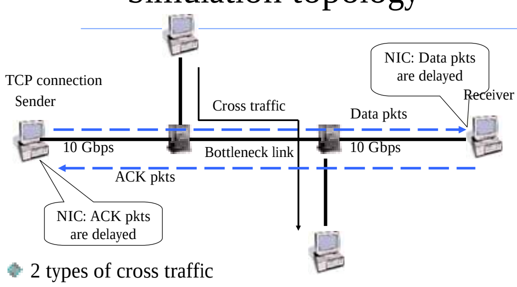
- Probing rate NP/S is set to various values
  - S is estimated by the size of the burst and the average throughput of TCP
  - N must be adjusted appropriately
- If the arrival interval of the two ACK bursts is enlarged
  - the available bandwidth (A) will be larger than the probing rate.
  - otherwise, the available bandwidth will be smaller

Steps in a measurement

- Measurements are repeated continuously
- Steps in a measurement
  - **Decide a search range** from the statistical information of previous results
    - High probability of including the available bandwidth
    - Faster measurement
  - Adjust packets in K bursts to **probe K points in the search range**
    - $K = 4$ in the simulations
  - **Infer the available bandwidth** from the probing results
  - **Wait** for a while before starting the next measurement
    - Avoid affecting the next measurement

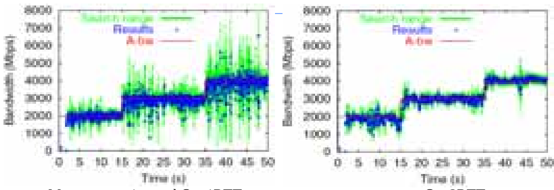


Simulation topology



- 2 types of cross traffic
  - Udp flows: Packet size according to the monitored results of the Internet traffic reported in NLANR
  - Web traffic: Numerous TCP connections

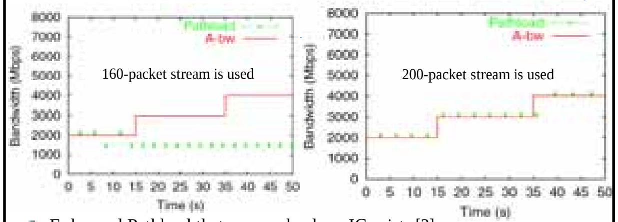
Measurement results for ICIM



Measurement interval Q= 1RTT   Q= 2RTTs

- **ICIM can deliver good measurement results in high-speed network**
- *Q*=2RTTs has better accuracy
  - A measurement is not affected by the previous one
  - Measurement frequency is only a half: 16.7 results/s vs. 34.2 results/s

Measurement results of Pathload



- Enhanced Pathload that can work where IC exists [2]
  - Only the last packet in each burst is used for original Pathload algorithm
  - The last packet is supposed to be delayed shortly at NIC
- Long packet streams are required
  - Results are yielded in rather long intervals
  - 0.28 results/s (when 200-packet stream is used)

[2] Prasad et.al. "Effects of interrupt coalescence on Network measurements", *PAM 2004*

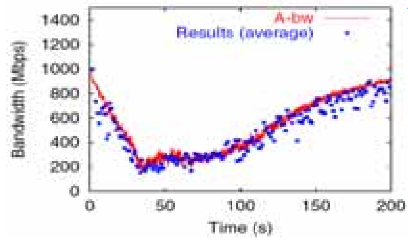
Comparison in number of packets

Number of packets required for a measurement

$A-bw$	ICIM	IC-aware Pathload	Ratio ICIM:Pathload
2 Gbps	110	200 · 12 · 8 = 19 200	0.006
3 Gbps	130	200 · 12 · 9 = 21 600	0.006
4 Gbps	154	200 · 12 · 10 =24 000	0.006

- Pathload probes 8,9,10 times for one result
  - For one probe, 12 streams are sent
- ICIM uses far fewer packets than Pathload
  - ICIM is suitable for using in TCP

Web traffic environment



- Web traffic: Numerous long/short TCP connections
  - Cross traffic is bursty
- The measurement results deviate only a little from the correct values
  - In general, the results follow the changes of available bandwidth.

Summary

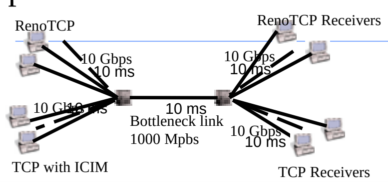
- We introduce ICIM
  - A packet burst-based measurement method
  - Deployed in a TCP connection
  - Effective in highspeed networks

Future studies

- Implementation of ICIM in the real network environment
  - Early results are reported at

http://www.anarg.jp/imtcp

Supplement: Fairness with RenoTCP



#connections	*Q*=1 RTT	*Q*=2 RTTs
4	466.4 : 490.6 (0.95:1)	483.7 : 475.6 (1.01:1)
8	451.1 : 544.4 (0.82:1)	505.1 : 490.5 (1.02:1)
12	418.7 : 577.7(0.72:1)	503.5 : 493.2 (1.02:1)

- A number of TCP with ICIM conflict with RenoTCP
- When Q=2 RTTs, TCP with ICIM has almost the same performance with RenoTCP