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
- Performing active measurement without probe traffic
- Using data packets in a TCP connection as probe packets
- Adjusting the transmission intervals of some data 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: 1Gbps => 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)
  - Intel, "Interrupt moderation using Intel Gigabit Ethernet Controllers"
- Measuring the available bandwidth from arrival intervals of ACK packets

TCP

Network

TCP receiver

TCP sender

Interrupt generated
Packet arrival

Time

Kernel does not know the real inter-arrival intervals of packets
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 1 µs [2].

Proposed method: ICIM

- Objective: New inline measurement method
- Measuring available bandwidth when IC is enabled
- Not requiring packet pacing

Basic idea: Packet-burst pair is considered 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

At the TCP sender

- \( A \geq N \frac{P}{S} \)
- \( S \) is unchanged

At the TCP receiver

- \( A < N \frac{P}{S} \)
- \( NP/S \) is enlarged
- \( S \) is unchanged

Steps in a measurement

- Measurements are repeated continuously
- 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
- \( S = 4 \) 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
ICIM can deliver good measurement results in high-speed network.
- $Q = 2\text{RTTs}$ 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

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
- 0.25 results/s (when 200-packet stream is used)

Comparison in number of packets

<table>
<thead>
<tr>
<th>A - Gbps ICIM</th>
<th>IC-aware Pathload</th>
<th>Ratio ICIM:Pathload</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Gbps</td>
<td>116</td>
<td>200 - 12 - 8 - 19 200</td>
</tr>
<tr>
<td>3 Gbps</td>
<td>136</td>
<td>200 - 12 - 9 - 21 600</td>
</tr>
<tr>
<td>4 Gbps</td>
<td>154</td>
<td>200 - 12 - 10 - 24 000</td>
</tr>
</tbody>
</table>

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

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](http://www.anarg.jp/imtcp)

Supplement: Fairness with RenoTCP

<table>
<thead>
<tr>
<th>Connections</th>
<th>$Q = 1\text{RTT}$</th>
<th>$Q = 2\text{RTTs}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>446.4 : 400.6 (0.05:1)</td>
<td>483.7 : 475.6 (1.01:1)</td>
</tr>
<tr>
<td>8</td>
<td>451.1 : 544.4 (0.82:1)</td>
<td>505.1 : 496.2 (1.02:1)</td>
</tr>
<tr>
<td>12</td>
<td>418.7 : 572.7 (0.82:1)</td>
<td>503.5 : 491.2 (1.02:1)</td>
</tr>
</tbody>
</table>

A number of TCP with ICIM conflict with RenoTCP
When $Q = 2\text{RTTs}$, TCP with ICIM has almost the same performance with RenoTCP