Source-based $E^2 CM$:

Validation of the Orlando Proposal

Cyriel Minkenberg & Mitch Gusat
IBM Research GmbH, Zurich
March 22, 2007
Some Concerns re. E2CM Raised in Orlando

1. **Destination (DST)-based per-flow RX rate calculation (throughput accounting)**
   - Preferably, the source (SRC) should handle this job

2. **Global clock synchronization required for forward latency measurement**
   - Too costly
Modification addressing Concern #1

1. **SRC** measures throughput in between probes
   1. Generally this equals the configured mean probe interval (e.g. 75 KB)
   2. May vary due to imposed interval jitter and max-interval time limit (e.g. 10 ms)
   3. Byte count $B(P_n)$ is included in probe $P_n$
   4. Upon reception, **DST** returns probe $P_n$ including $B(P_n)$ and records probe arrival time $T_{dst}(P_n)$ in probe $P_n$
   5. Upon return, **SRC** stores $T_{dst}(P_n)$ for this particular flow
   6. **SRC** computes throughput as follows: $B(P_n) / (T_{dst}(P_n) - T_{dst}(P_{n-1}))$
   1. Clock synchro is not an issue: both time stamps are recorded at **DST**

Potential demerits:
1. Does not account for dropped frames
2. Less robust to lost/corrupted probes
Modification addressing Concern #2

- Use SRC clock to determine forward latency
  - Expedite probes on reverse path
    - Use top priority traffic class
    - Switches automatically preempt other traffic for probes
  - SRC includes time stamp $T_{src}(P_n)$ in probe $P_n$
  - Upon return, SRC computes round-trip latency $L(P_n) = \text{now} - T_{src}(P_n)$
  - SRC keeps track of minimum round-trip latency $L_0 = \min_n(P_n)$
  - SRC computes effective forward latency as $L(P_n) - L_0$
Output-Generated Single-Hop Hotspot

- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Node 1 service rate = 10%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected
Simulation Setup & Parameters

• Traffic
  - I.i.d. Bernoulli arrivals
  - Uniform destination distribution (to all nodes except self)
  - Fixed frame size = 1500 B

• Scenarios
  1. Single-hop output-generated hotspot

• Switch
  - M = 300 KB/port
  - Partitioned memory per input, shared among all outputs
  - No limit on per-output memory usage
  - PAUSE enabled or disabled
    - Applied on a per input basis based on local high/low watermarks
    - watermark\textsubscript{high} = 280 KB
    - watermark\textsubscript{low} = 260 KB
    - If disabled, frames dropped when input partition full

• Adapter
  - Per-node virtual output queuing
  - No limit on number of rate limiters
  - Unlimited ingress buffer size
  - Egress buffer size = 1500 KB
  - PAUSE enabled
    - watermark\textsubscript{high} = 1500 - rtt*bw KB
    - watermark\textsubscript{low} = watermark\textsubscript{high} - 10 KB

• ECM
  - W = 2.0
  - Q\textsubscript{eq} = 75 KB (= M/4)
  - G\textsubscript{d} = 0.5 / ((2*W+1)*Q\textsubscript{eq})
  - G\textsubscript{i0} = (R\textsubscript{link} / R\textsubscript{unit}) * ((2*W+1)*Q\textsubscript{eq})
  - G\textsubscript{i} = 0.005 * G\textsubscript{i0}
  - P\textsubscript{sample} = 2% (on average 1 sample every 75 KB)
  - R\textsubscript{unit} = R\textsubscript{min} = 1 Mb/s
  - BCN\_MAX enabled, threshold = 280 KB
  - No BCN(0,0), no self-increase

• E\textsuperscript{2}CM (per-flow)
  - W = 2.0
  - Q\textsubscript{eq} = 15 KB
  - G\textsubscript{d} = 2.5 / ((2*W+1)*Q\textsubscript{eq})
  - G\textsubscript{i} = 0.025 * G\textsubscript{i0}
  - P\textsubscript{sample} = 2% (on average 1 sample every 75 KB)
  - R\textsubscript{unit} = R\textsubscript{min} = 1 Mb/s
  - BCN\_MAX enabled, threshold = 56 KB
Results single-hop OG scenario ($N = 16$)

- Source- vs. destination-based (both mods 1 and 2)
- Switch PAUSE enabled/disabled
- No thresholding of OQ (unlimited within h/w boundaries)

<table>
<thead>
<tr>
<th>Switch frame drops</th>
<th>Dst-based</th>
<th>Src-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAUSE on</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PAUSE off</td>
<td>146,595</td>
<td>130,268</td>
</tr>
</tbody>
</table>
Results single-hop OG scenario - Impact of RTT

- Source-based (both mods 1 and 2)
- Switch PAUSE disabled
- Unlimited output queue length (hoggable)
- RTT = [0, 10, 100, 200, 500] μs

<table>
<thead>
<tr>
<th>RTT (μs)</th>
<th>Switch frame drops</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>134,879</td>
</tr>
<tr>
<td>10</td>
<td>148,816</td>
</tr>
<tr>
<td>100</td>
<td>135,874</td>
</tr>
<tr>
<td>200</td>
<td>144,239</td>
</tr>
<tr>
<td>500</td>
<td>189,371</td>
</tr>
</tbody>
</table>
Results single-hop OG scenario - OQ limit

- Source-based (both modifications)
- Switch PAUSE disabled
- 600 KB limit on output queue length
- RTT = [0, 10, 100, 200, 500] ms

<table>
<thead>
<tr>
<th>RTT (ms)</th>
<th>Switch frame drops</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16,083</td>
</tr>
<tr>
<td>10</td>
<td>14,230</td>
</tr>
<tr>
<td>100</td>
<td>11,116</td>
</tr>
<tr>
<td>200</td>
<td>7,171</td>
</tr>
<tr>
<td>500</td>
<td>11,300</td>
</tr>
</tbody>
</table>
Conclusions: Pat’s Orlando Proposal Works...

• Source-based and destination-based $E^2CM$ are practically indistinguishable in terms of SH-OG performance
  - consequential for h/w implementation...

• Stability is achieved even with RTTs up to 500 $\mu$s
  - However, mean queue level increases with RTT as consequence of additional transport lag

• In PAUSE-less mode frame drops* can be significantly (~10x) reduced by using per-OQ drop threshold
  - such, or more sophisticated, partitioning is recommended

* An arguable pursuit (reducing loss rate w/o LL-FC) ...