On Growth of Parallelism within Routers and Its Impact on Packet Reordering

A. A. Bare\textsuperscript{1}, A. P. Jayasumana \textit{and} N. M. Piratla\textsuperscript{2}

Electrical & Computer Engineering, Colorado State University

\textsuperscript{1}Agilent Technologies, Inc., Loveland, CO
\textsuperscript{2}Deutsche Telekom Laboratories, Berlin, Germany
Outline

- Growth trends
- Impact on router architectures
- Metrics for reordering
- Simulation results
- Conclusions
Growth Trends

“Internet traffic continues to grow vigorously, approximately doubling each year, as it has done every year since 1997.” (annual growth between 70 and 150%.)

Growth Trends

Progress in transmission technology appears sufficient to double network capacity each year for about the next decade.

Figure 1. Progress in lightwave transmission capacity.
Growth Trends

“For the first time in history, performance improvements are required at a rate faster than 18-month doubling of semiconductor performance that Moore’s Law predicted in 1975.”

<table>
<thead>
<tr>
<th>System</th>
<th>Fiber Capacity</th>
<th>Wide Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>8x2.5 Gbps</td>
<td>20 Gbps</td>
<td>1996</td>
</tr>
<tr>
<td>16x2.5 Gbps</td>
<td>40 Gbps</td>
<td>1996</td>
</tr>
<tr>
<td>32x2.5 Gbps</td>
<td>80 Gbps</td>
<td>1996</td>
</tr>
<tr>
<td>80x2.5 Gbps</td>
<td>200 Gbps</td>
<td>2000</td>
</tr>
<tr>
<td>40x10 Gbps</td>
<td>400 Gbps</td>
<td>2000</td>
</tr>
<tr>
<td>160x2.5 Gbps</td>
<td>400 Gbps</td>
<td>Late 2000</td>
</tr>
<tr>
<td>80x10 Gbps</td>
<td>800 Gbps</td>
<td>2002</td>
</tr>
<tr>
<td>160x10 Gbps</td>
<td>1.6 Tbps</td>
<td>2003</td>
</tr>
<tr>
<td>40x40 Gbps</td>
<td>1.6 Tbps</td>
<td>2003</td>
</tr>
<tr>
<td>80x40 Gbps</td>
<td>3.2 Tbps</td>
<td>2003/4</td>
</tr>
<tr>
<td>100x40 Gbps</td>
<td>4 Tbps</td>
<td>2005</td>
</tr>
<tr>
<td>160x40 Gbps</td>
<td>6.4 Tbps</td>
<td>2007</td>
</tr>
</tbody>
</table>

A Generic Router

- Line Card
- Routing Software
- Routing Table
- Packet Processing - Network Processors
- Switching Fabric
- Buffer Space
- Line Card
BGP Table Size

\[ S = 7.804 \times 10^{-13} \times T_u^2 - 0.00076 \times T_u + 9.603 \times 10^4 \]

Source: http://bgp.potaroo.net, AS1221 (Telstra) router
A Generic Router

- Line Card
- Routing Software
- Routing Table
- Packet Processing - Network Processors
- Switching Fabric
- Buffer Space
- Line Card
Growth Factors

- $\alpha$ - increase in network link speed
- $\beta$ - increase in processing speed
- $\gamma$ - increase in routing table size

Computations for packet processing:

$$\log_2 (\gamma)$$

Number of NPUs:

$$\alpha \log_2 (\gamma) / \beta$$
Parallelism within routers

<table>
<thead>
<tr>
<th>Link change</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$\omega$</th>
<th>$n$</th>
<th>Mean Packet Processing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC-12</td>
<td>4</td>
<td>2.00</td>
<td>1.60</td>
<td>2.71</td>
<td>3</td>
<td>$0.00722 \times L + 8.476$</td>
</tr>
<tr>
<td>OC-24</td>
<td>8</td>
<td>2.83</td>
<td>1.87</td>
<td>7.22</td>
<td>5</td>
<td>$0.00680 \times L + 7.977$</td>
</tr>
<tr>
<td>OC-48</td>
<td>16</td>
<td>4.00</td>
<td>2.16</td>
<td>17.77</td>
<td>9</td>
<td>$0.00592 \times L + 6.944$</td>
</tr>
<tr>
<td>OC-96</td>
<td>32</td>
<td>5.66</td>
<td>2.46</td>
<td>41.56</td>
<td>15</td>
<td>$0.00489 \times L + 5.736$</td>
</tr>
<tr>
<td>OC-192</td>
<td>64</td>
<td>8.00</td>
<td>2.77</td>
<td>94.07</td>
<td>24</td>
<td>$0.00391 \times L + 4.593$</td>
</tr>
<tr>
<td>OC-384</td>
<td>128</td>
<td>11.3</td>
<td>3.10</td>
<td>208.9</td>
<td>37</td>
<td>$0.00307 \times L + 3.608$</td>
</tr>
<tr>
<td>OC-768</td>
<td>256</td>
<td>16.0</td>
<td>3.44</td>
<td>456.3</td>
<td>57</td>
<td>$0.00237 \times L + 2.785$</td>
</tr>
</tbody>
</table>

Table i. parameters used in simulations for different link speeds
Parallelism within Routers => Reordering?

Ex: Juniper M160
Impact of Reordering

Ex: Degradation of performance in TCP
- No of unnecessary transmissions increase (drop in throughput)
- Congestion window becomes small
- RTT estimate degrades
- Receiver performance degrades
- Detection of lost packets delayed
- Forward/reverse path causes loss of self-clocking

[Laor & Gendel, IEEE Network 2002]
Parallelism within Routers => Reordering?

Ex: Juniper M160

Counter Measures: input flow tracking, output buffering
Reorder Density (RD)

EX: Received sequence (1, 3, 4, 2, 5, 6, 8, 7, 9, 10)

<table>
<thead>
<tr>
<th>Sequence number</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>2</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>7</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receive index</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Displacement</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Each arrival is assigned a receive index
- Displacement of a packet = (receive index – sequence number)

RD is computed as

<table>
<thead>
<tr>
<th>Displacement</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Normalized Frequency</td>
<td>0.3</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Reorder Buffer Density RBD

- Ex: Received sequence: 1,3,4,2,5,7,6,8
  - Compute the buffer occupancy after each arrival

<table>
<thead>
<tr>
<th>Arrived sequence number (S)</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>2</th>
<th>5</th>
<th>7</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence number expected (E)</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Buffer contents after arrival</td>
<td>-</td>
<td>3</td>
<td>3,4</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Buffer occupancy (B)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- RBD is the normalized frequency of each buffer occupancy

<table>
<thead>
<tr>
<th>Buffer occupancy</th>
<th>Frequency</th>
<th>Normalized frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>5 / 8</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2 / 8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1 / 8</td>
</tr>
</tbody>
</table>

![Graph showing normalized frequency of buffer occupancy]
Router Simulation
Router Parameters

- Time taken for processing a packet arriving over OC-3 link
  \[ d(L) = (0.0213 \times L + 25) \text{ in } \mu s \]
  where L is length of a packet in bytes

- For simulation, packet processing time is
  \[ P(L) = d(L) \times \log_2 (\gamma) / \beta \]

- Standard deviation in packet processing time: 5% used unless specified

Simulation Parameters

- Multiple streams combined per input to provide the required utilization
  Ex: 4000 streams
- Reordering measured on one of the streams
  Ex: 1000 packets
- Self-similar traffic arrivals in each stream
- Constant packet size in each stream, aggregate follows given distribution
- Round-robin scheduling
  (Conservative choice)
- No buffer overflow

RD vs. Link Speed

(at 50% utilization)
RBD vs. Link Speed

(at 50% utilization)
RBD vs. Link Utilization

(For OC768, packet size 1500 bytes)
RBD vs. Packet Size

(with stream of interest occupying 10Mbps on a 50% utilized OC768)
Internet Measurements

(a) Armenia
2. MA, USA
3. Lazio, Italy
4. Wellington, NZ
5. India
6. ACT, Australia
### Internet Measurements

<table>
<thead>
<tr>
<th>Net</th>
<th>IR%</th>
<th>OR%</th>
<th>ER%</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>24</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>38</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>86</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>84</td>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Basic Patterns form significant amount of reordering patterns.

- **IR** – Independent Reordering
- **OR** – Overlap Reordering
- **ER** – Embedded Reordering
Solutions?

- Deal with it at end nodes
- Input tracking and output buffering
- Creative designs – pipelined solutions
- Switch bursts of packets, use longer packets, ...
Increasing gap between link speeds and processing speeds demands increased parallelism within routers resulting in increased packet reordering.

Need to deal with packet reordering proactively to prevent it from negating some of the performance gain resulting from link and processing speeds.

Throughput and delay characteristics are no longer sufficient to characterize network performance. Need to pay attention to secondary measures such as packet reordering.