An In-Depth Analysis of the Slingshot Interconnect
All HPC traffic layered over **RoCEv2**

Efficient **software stack**

High-Radix **Switches**

Low-Diameter **Topology**

Congestion Control

Adaptive Routing

Quality of Service
Adaptive Routing

Efficient software stack

High-Radix Switches

Low-Diameter Topology

Congestion Control

All HPC traffic layered over RoCEv2

Adaptive Routing

Quality of Service
All HPC traffic layered over RoCEv2

Efficient software stack

- High-Radix Switches
- Low-Diameter Topology
- Congestion Control
- Adaptive Routing
- Quality of Service
ETHERNET ENHANCEMENTS

Can process both **standard** and **enhanced** Ethernet packets

1024 nodes
SOFTWARE STACK

Standard TCP/IP stack or libfabric

![Graph showing RTT/2 (usec) vs Size (Bytes)]
All HPC traffic layered over RoCEv2

Efficient software stack

High-Radix Switches

Low-Diameter Topology

Congestion Control

Adaptive Routing

Quality of Service
SWITCH - ROSETTA

64 x 200Gb/s ports

32 tiles
SWITCH - ROSETTA

Rosetta Switch

<table>
<thead>
<tr>
<th>00</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
</tr>
<tr>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
<td>57</td>
<td>58</td>
<td>59</td>
<td>60</td>
<td>61</td>
<td>62</td>
<td>63</td>
</tr>
</tbody>
</table>

Row Busses

Column crossbars

Port numbers

Tile

Port 30

16 row buses (1 per port)

Port 31

16:8 xbar

To other ports

14/15
46/47
62/63

From other tiles

14/15
46/47
62/63

46 47 62 63

9
All HPC traffic layered over RoCEv2

Efficient software stack

High-Radix Switches

Low-Diameter Topology

Congestion Control

Adaptive Routing

Quality of Service
SLINGSHOT TOPOLOGY

Switches can be connected into arbitrary topologies

Dragonfly is the default topology

All-to-all amongst groups

= Endpoints

= Switches
SLINGSHOT TOPOLOGY - LATENCY & BANDWIDTH
SLINGSHOT TOPOLOGY - LATENCY & BANDWIDTH

~40% difference

- Same switch
- Different switches
- Different groups

Time (µs)
- 2.0
- 2.2
- 2.5
- 2.8

Bandwidth (Gb/s)
- 0.07
- 0.08
- 0.09
- 0.1

8B
- 9
1KiB
- 10
128KiB
- 70
4MiB
- 97.0

L
- Q3
- Median
- Q1
- S
All HPC traffic layered over RoCEv2

Efficient *software stack*

High-Radix *Switches*

Low-Diameter *Topology*

**Congestion Control**

Adaptive Routing

Quality of Service
CONGESTION CONTROL

- **ECN/QCN** hard to tune and slow to converge
- Tracks the traffic between **every pair of endpoints**
- Slows down **offending traffic** only
- Improves average and tail *latencies*
CONGESTION CONTROL TESTS

Run two concurrent jobs: **victim** and **aggressor**

- Microbenchs.
- MILC
- HPCG
- LAMMPS
- FFT
- silo
- sphinx
- xapian
- img-dnn
- resnet-proxy

- Tailbench
  - **incast** (endpoint congestion)
  - **all-to-all** (intermediate congestion)
CONGESTION IMPACT - 512 NODES
CONGESTION IMPACT - ADDITIONAL ANALYSIS
All HPC traffic layered over RoCEv2

Efficient software stack

High-Radix Switches

Low-Diameter Topology

Congestion Control

Adaptive Routing

Quality of Service
QUALITY OF SERVICE

Each traffic class occupies **hardware resources** in the switches.

**Tunable** priority, ordering, minimum/maximum bandwidth, ...

Jobs can be assigned to a small number **traffic classes**.

Traffic class can be changed on a **per-packet** basis.
QOS TESTS

25% bandwidth tapering

2 jobs running bisection bandwidth tests

TC1: 80% minimum bandwidth

TC2: 10% minimum bandwidth

Same TC

Gb/s/node

Time (msec)

Job 1 (TC1)  Job 2 (TC1)
QOS TESTS

25% bandwidth tapering

2 jobs running bisection bandwidth tests

TC1: 80% minimum bandwidth

TC2: 10% minimum bandwidth

Same TC

Separate TCs
CONCLUSIONS

All HPC traffic layered over RoCEv2
Efficient software stack
High-Radix Switches
Low-Diameter Topology
Congestion Control
Adaptive Routing
Quality of Service

CONGESTION IMPACT - 512 NODES

CONGESTION IMPACT - ADDITIONAL ANALYSIS