Daniele De Sensi[1], Salvatore Di Girolamo[1], Kim McMahon[2], Duncan Roweth[2], Torsten Hoefler[1]

[1]ETH Zurich, Switzerland
[2]HPE Cray

# An In-Depth Analysis of the Slingshot Interconnect

SPCL
ETH zürich
Hewlett Packard Enterprise

All HPC traffic layered over **RoCEv2**

Efficient **software stack**

High-Radix **Switches**

Low-Diameter **Topology**

**Congestion Control**

**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**

All HPC traffic layered over **RoCEv2**

Efficient **software stack**

High-Radix **Switches**

Low-Diameter **Topology**

**Congestion Control**

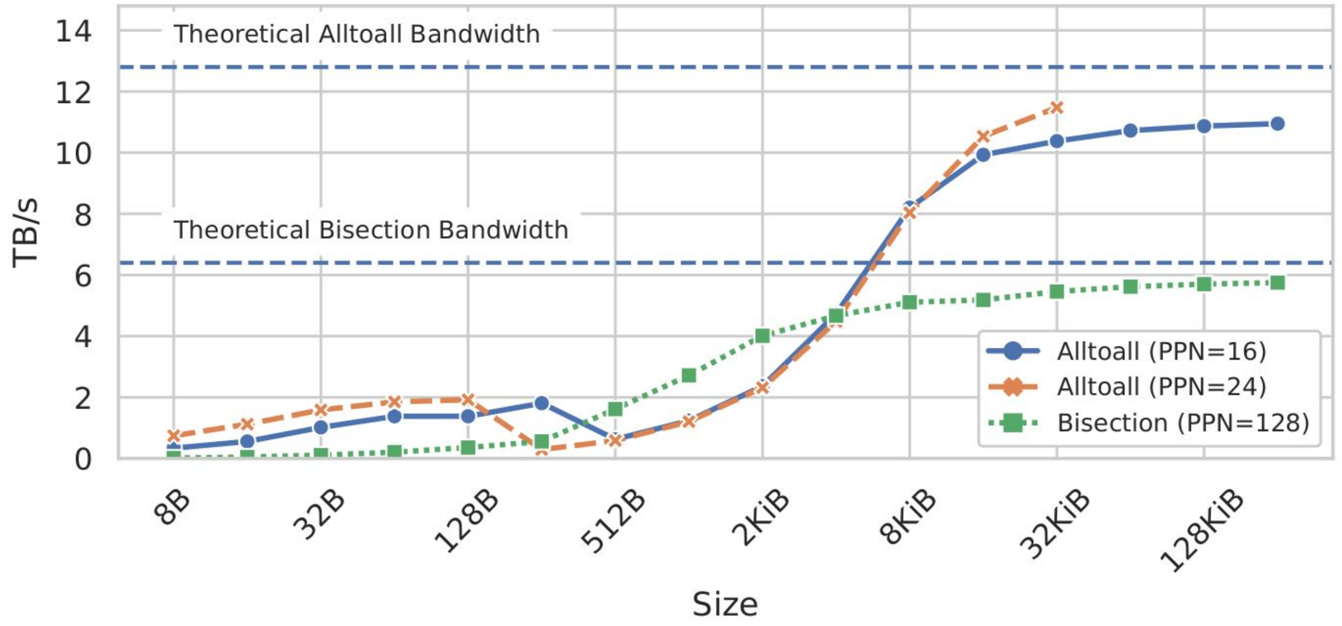
**Adaptive Routing**

**Quality of Service**

# ETHERNET ENHANCEMENTS

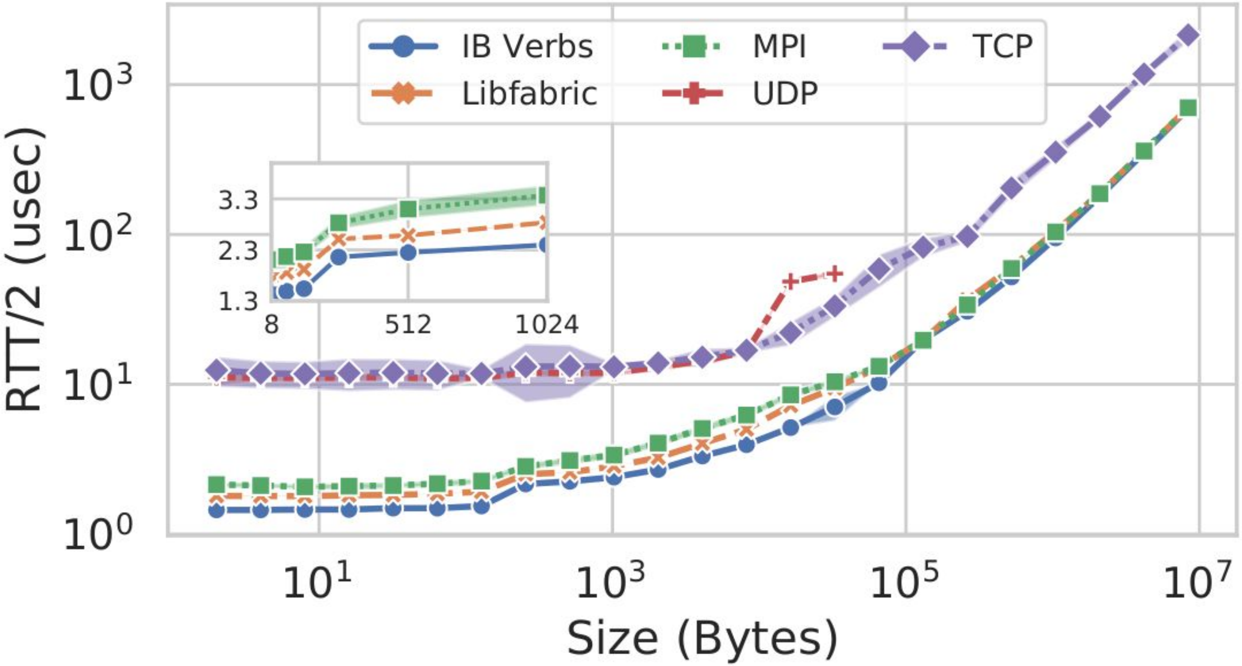
Can process both **standard** and **enhanced** Ethernet packets

1024 nodes



# SOFTWARE STACK

Standard **TCP/IP** stack or **libfabric**



All HPC traffic layered over **RoCEv2**

Efficient **software stack**

High-Radix **Switches**

Low-Diameter **Topology**

**Congestion Control**

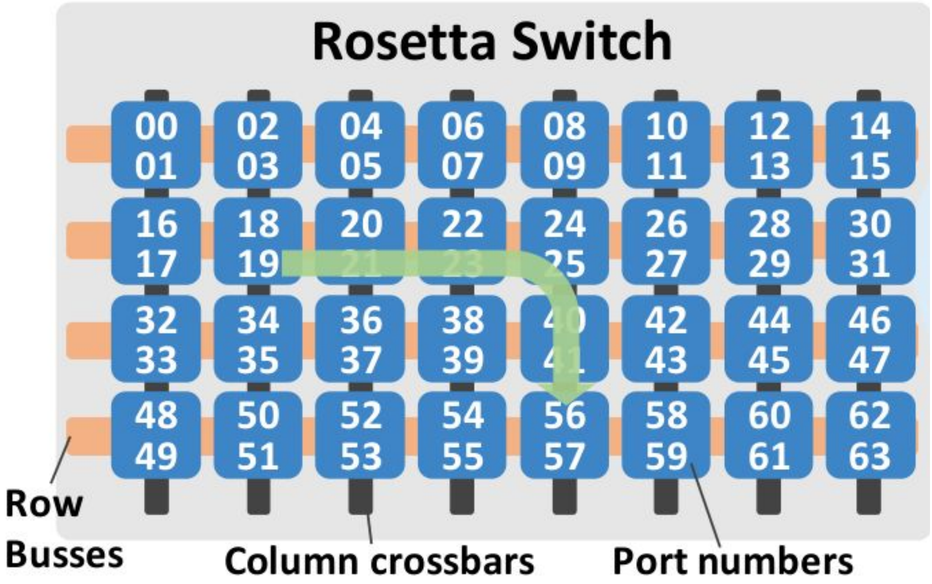
**Adaptive Routing**

**Quality of Service**

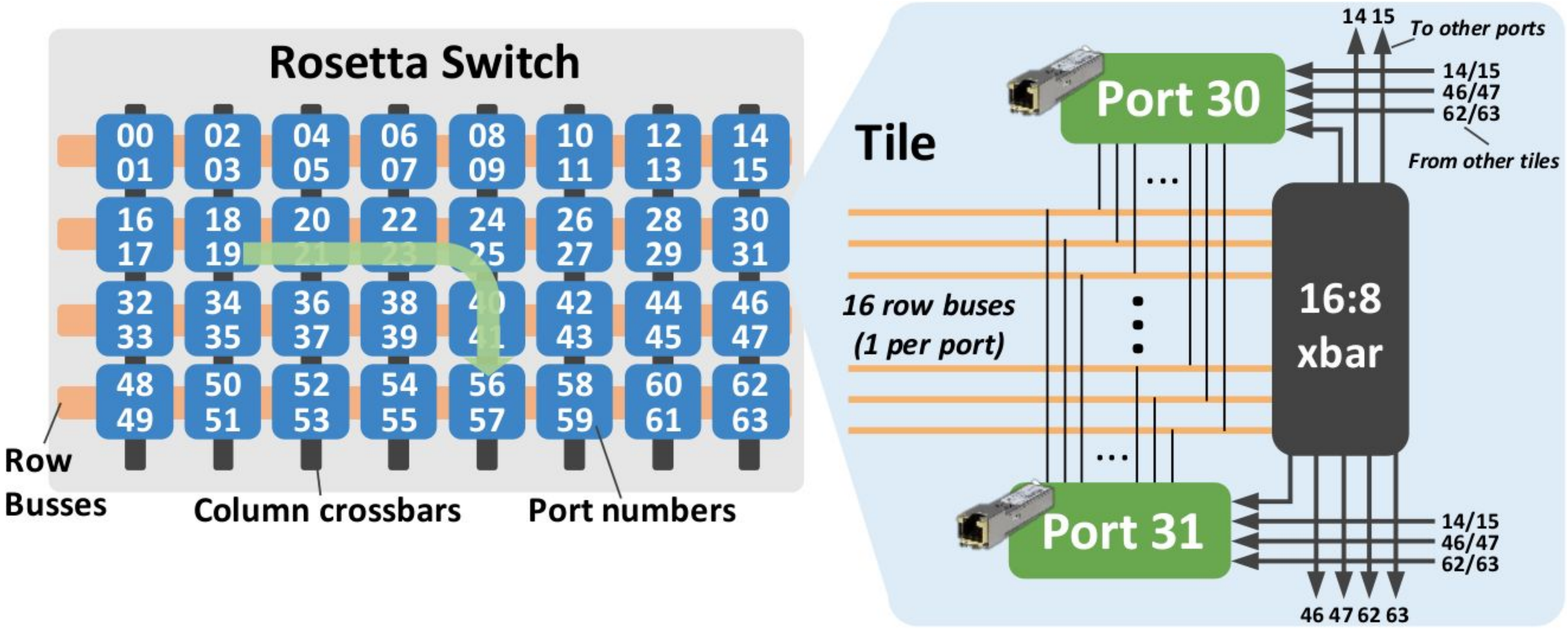
# SWITCH - ROSETTA

64 x 200Gb/s ports

32 tiles



# SWITCH - ROSETTA



All HPC traffic layered over **RoCEv2**

Efficient **software stack**

High-Radix **Switches**

Low-Diameter **Topology**

**Congestion Control**

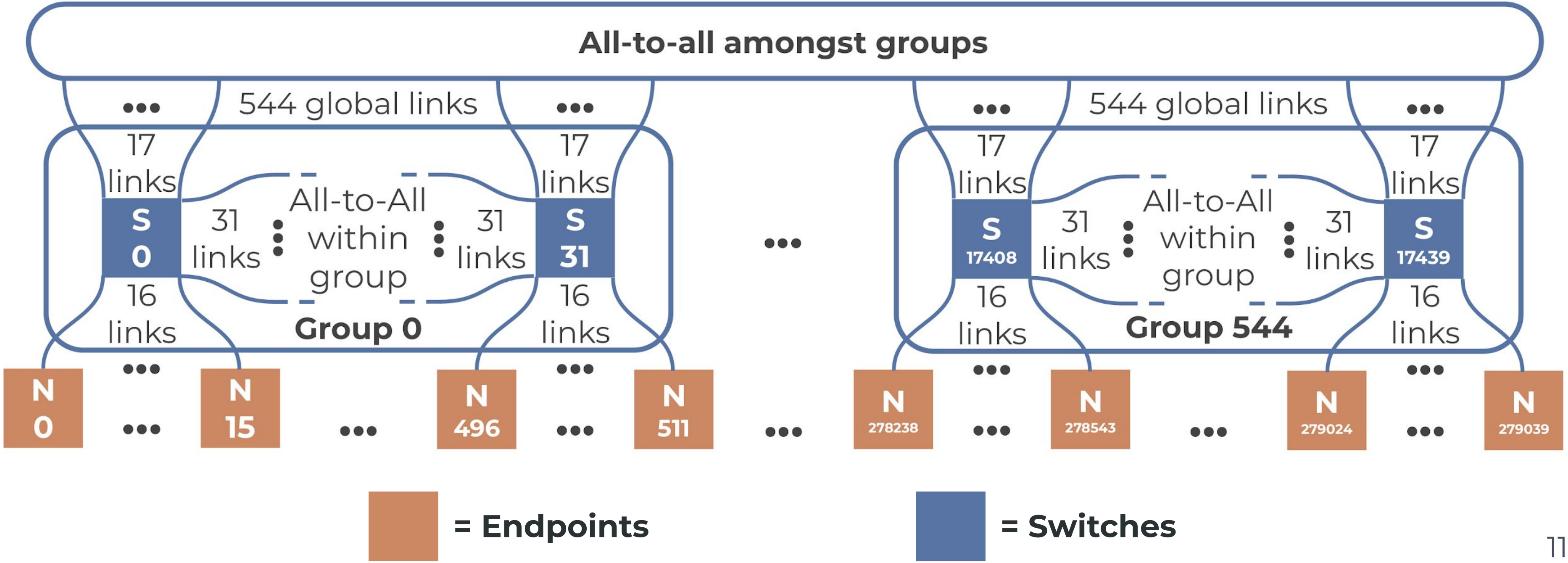
**Adaptive Routing**

**Quality of Service**

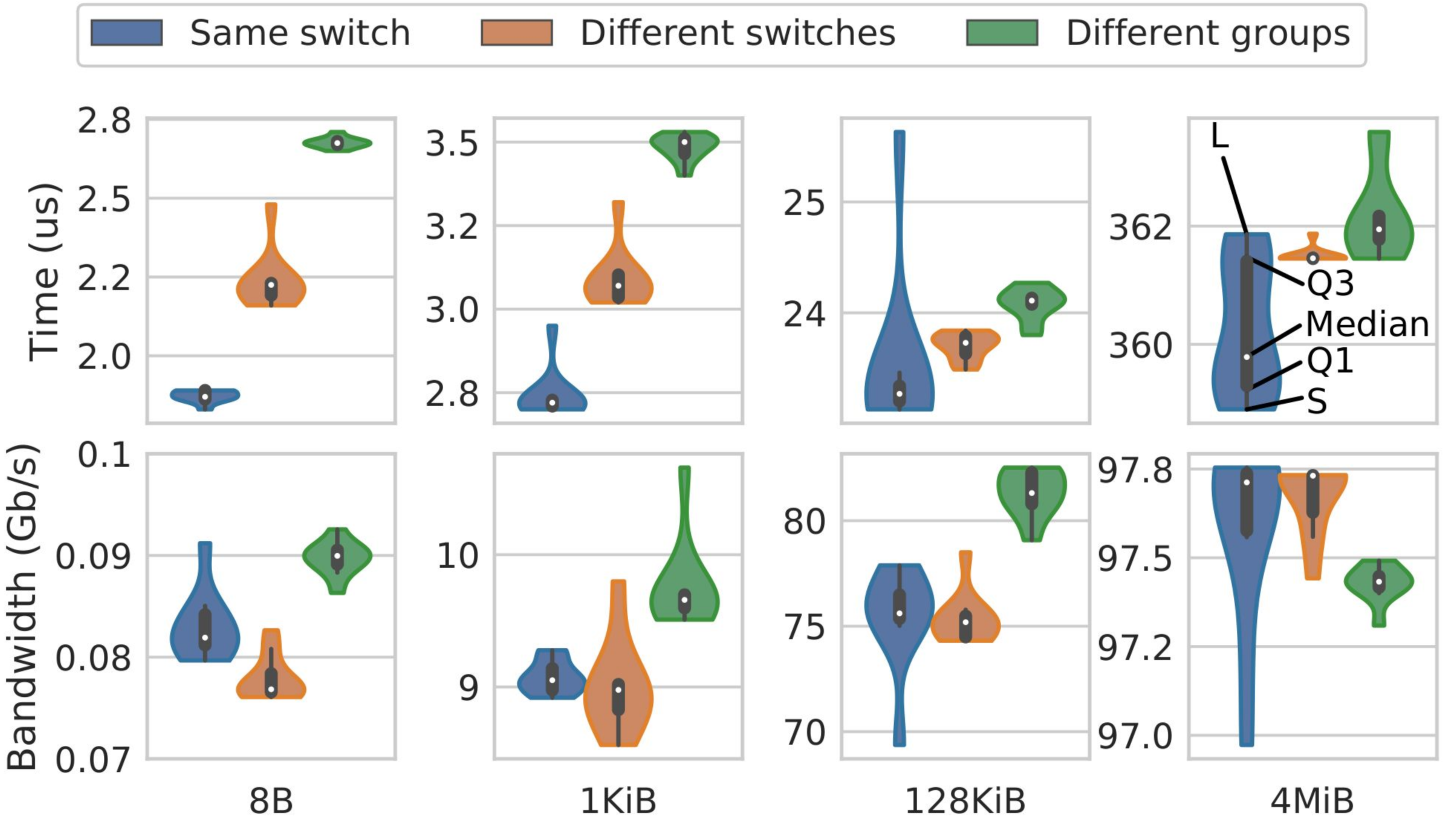
# SLINGSHOT TOPOLOGY

Switches can be connected into **arbitrary topologies**

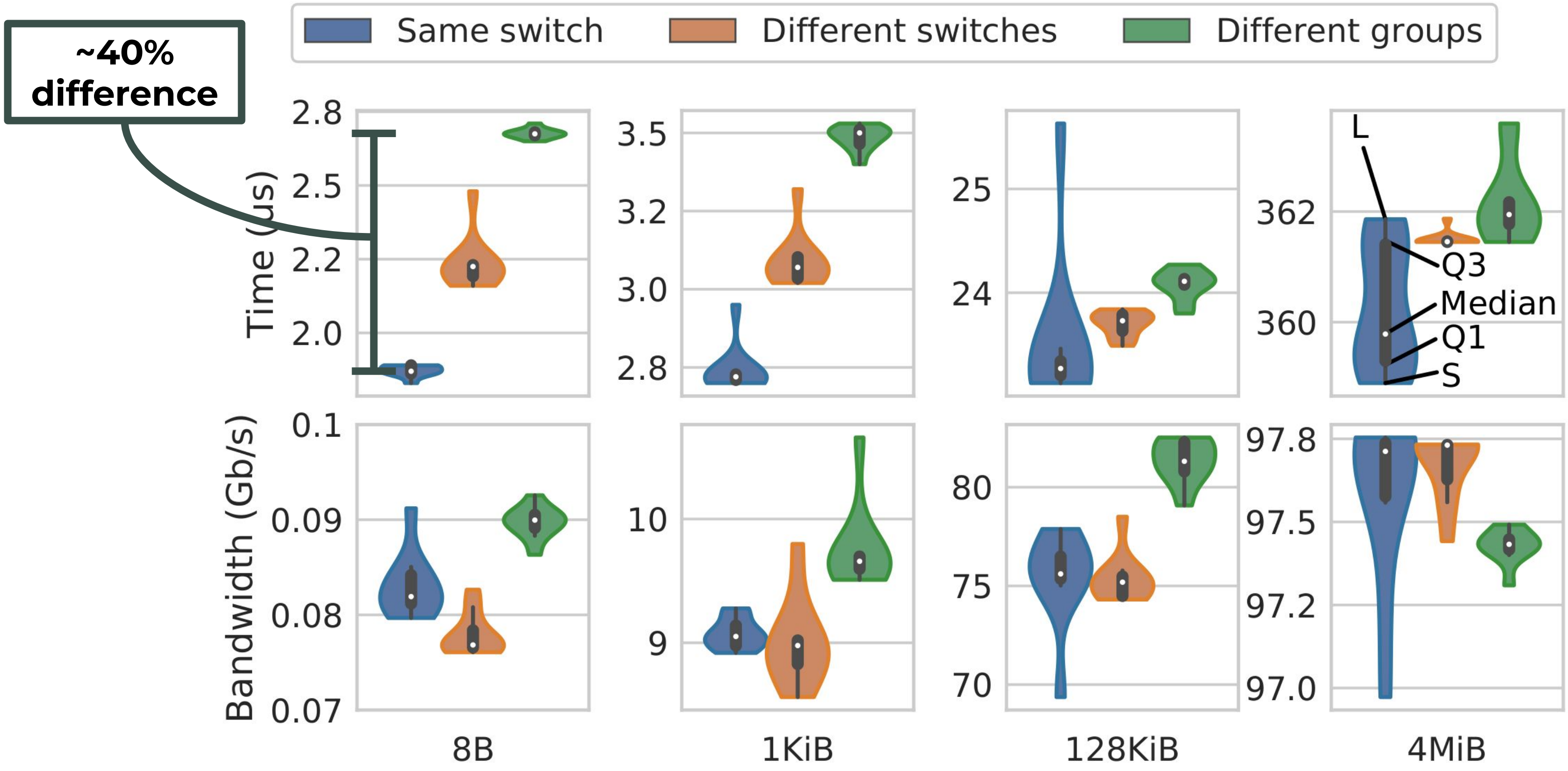
**Dragonfly** is the default topology



# SLINGSHOT TOPOLOGY - LATENCY & BANDWIDTH



# SLINGSHOT TOPOLOGY - LATENCY & BANDWIDTH



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**
t congestion)

**all-to-all**
(intermediate congestion)

# CONGESTION IMPACT - 512 NODES

Color scales: 1 – 45 (applications); 1 – 92 (microbenchmarks)

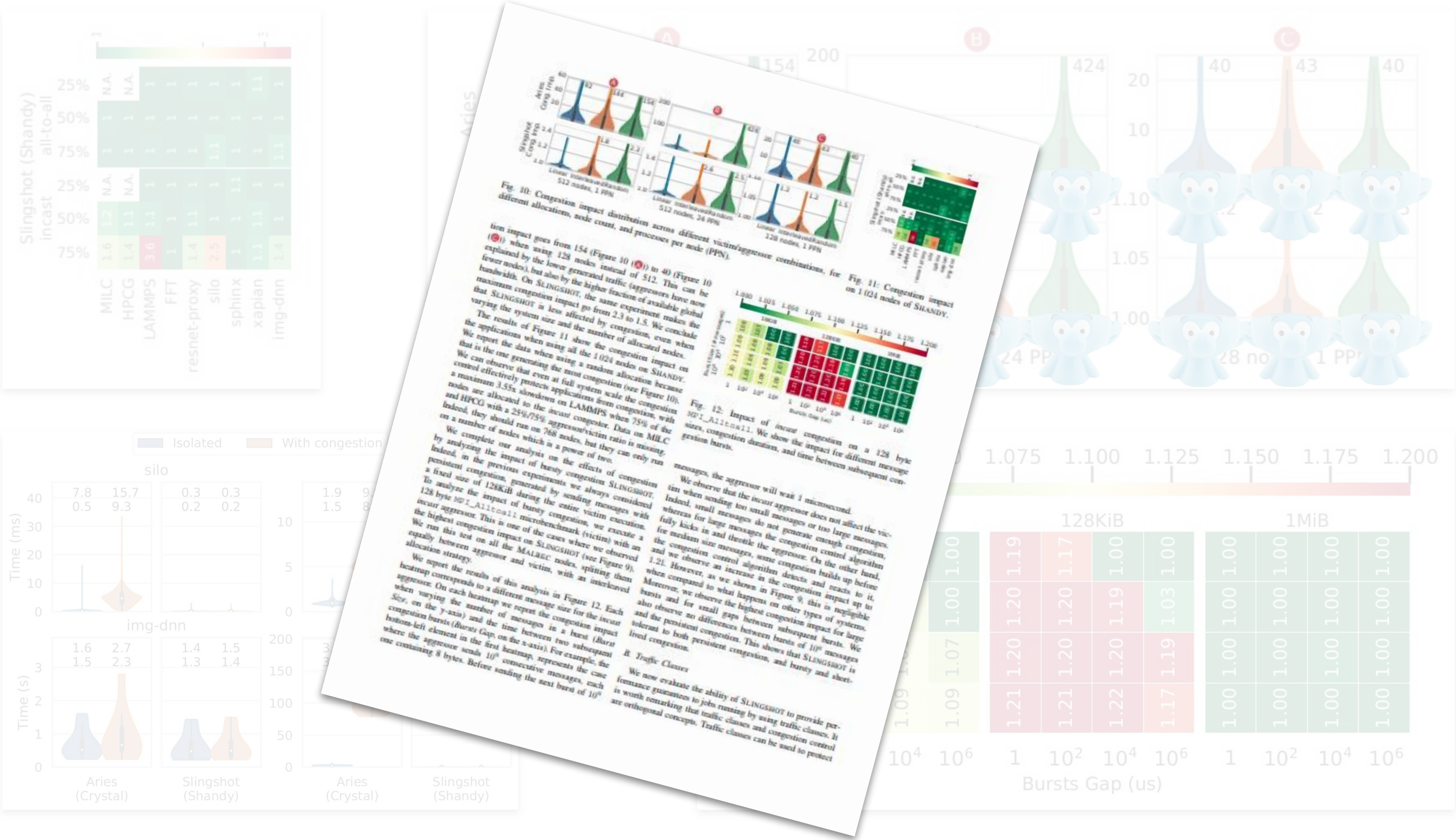
| System | Pattern | Load | MILC | HPCG | LAMMPS | FFT | resnet-proxy | silo | sphinx | xapian | img-dnn |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Aries (Crystal) | all-to-all | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Aries (Crystal) | all-to-all | 50% | 1 | 1 | 1 | 1 | 1 | 1.3 | 1 | 1.1 | 1 |
| Aries (Crystal) | all-to-all | 90% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.2 | 1 |
| Aries (Crystal) | incast | 10% | 1.6 | 1.5 | 5.7 | 1.8 | 1.3 | 1.1 | 1 | 5.2 | 1.2 |
| Aries (Crystal) | incast | 50% | 4.6 | 3.9 | 17 | 6.9 | 6.5 | 12 | 2.2 | 24 | 4.2 |
| Aries (Crystal) | incast | 90% | 2.6 | 2.3 | 7.3 | 2.2 | 7.3 | 11 | 1.4 | 46 | 5.6 |
| Slingshot (Shandy) | all-to-all | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1.1 | 1.2 | 1 |
| Slingshot (Shandy) | all-to-all | 50% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.1 | 1 |
| Slingshot (Shandy) | all-to-all | 90% | 1 | 1 | 1 | 1 | 1 | 1 | 1.1 | 1.1 | 1.1 |
| Slingshot (Shandy) | incast | 10% | 1 | 1 | 1 | 1.1 | 1 | 1 | 1.1 | 1.1 | 1 |
| Slingshot (Shandy) | incast | 50% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.2 | 1 |
| Slingshot (Shandy) | incast | 90% | 1 | 1.2 | 1 | 1 | 1 | 1 | 1.1 | 1 | 1 |

| System | Pattern | Load | pingpong 8B | pingpong 128B | pingpong 1KiB | pingpong 16KiB | pingpong 128KiB | pingpong 1MiB | pingpong 4MiB | pingpong 16MiB | allreduce 8B | allreduce 128B | allreduce 1KiB | allreduce 16KiB | allreduce 128KiB | allreduce 1MiB | allreduce 4MiB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aries (Crystal) | all-to-all | 10% | 1.3 | 1.3 | 1.3 | 1.2 | 1.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Aries (Crystal) | all-to-all | 50% | 1.1 | 1.1 | 1.1 | 1.1 | 1 | 1 | 1 | 1 | 1.2 | 1.2 | 1.1 | 1.1 | 1 | 1 | 1 |
| Aries (Crystal) | all-to-all | 90% | 1.5 | 1.6 | 2 | 1.6 | 1.2 | 1 | 1 | 1 | 1.9 | 2 | 1.8 | 1.4 | 1.1 | 1 | 1 |
| Aries (Crystal) | incast | 10% | 13 | 13 | 16 | 15 | 13 | 2.9 | 2.4 | 2.2 | 14 | 10 | 13 | 9.4 | 6 | 4.4 | 4.6 |
| Aries (Crystal) | incast | 50% | 44 | 47 | 54 | 35 | 44 | 20 | 9.9 | 4.4 | 59 | 66 | 53 | 43 | 28 | 13 | 14 |
| Aries (Crystal) | incast | 90% | 80 | 87 | 93 | 70 | 67 | 30 | 12 | 6.9 | 62 | 71 | 62 | 32 | 23 | 14 | 14 |
| Slingshot (Shandy) | all-to-all | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | all-to-all | 50% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | all-to-all | 90% | 1 | 1.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 10% | 1.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 50% | 1 | 1 | 1 | 1.1 | 1.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.2 | 1.2 | 1.2 | 1.1 |
| Slingshot (Shandy) | incast | 90% | 1.3 | 1.5 | 1.2 | 1.4 | 1.2 | 1.1 | 1 | 1 | 1.1 | 1 | 1.1 | 1.1 | 1.1 | 1 | 1 |

| System | Pattern | Load | alltoall 8B | alltoall 128B | alltoall 1KiB | alltoall 16KiB | alltoall 128KiB | alltoall 1MiB | alltoall 4MiB | barrier |
|---|---|---|---|---|---|---|---|---|---|---|
| Aries (Crystal) | all-to-all | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Aries (Crystal) | all-to-all | 50% | 1.1 | 1 | 1 | 1.1 | 1 | 1 | 1 | 1.2 |
| Aries (Crystal) | all-to-all | 90% | 1.9 | 1.5 | 1.4 | 1.2 | 1.1 | 1.1 | 1.1 | 2 |
| Aries (Crystal) | incast | 10% | 8.4 | 4.4 | 5.7 | 4 | 3.9 | 3.9 | 4 | 11 |
| Aries (Crystal) | incast | 50% | 56 | 34 | 28 | 19 | 15 | 15 | 15 | 59 |
| Aries (Crystal) | incast | 90% | 70 | 69 | 63 | 36 | 24 | 23 | 23 | 77 |
| Slingshot (Shandy) | all-to-all | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | all-to-all | 50% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | all-to-all | 90% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 50% | 1 | 1.3 | 1 | 1 | 1.1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 90% | 1 | 1 | 1 | 1 | 1.1 | 1 | 1 | 1 |

| System | Pattern | Load | broadcast 8B | broadcast 128B | broadcast 1KiB | broadcast 16KiB | broadcast 128KiB | broadcast 1MiB | broadcast 4MiB | broadcast 16MiB | hal 128B | hal 1KiB | swp 128B | swp 512B | inc 8B | inc 128B | inc 1KiB | inc 16KiB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aries (Crystal) | all-to-all | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Aries (Crystal) | all-to-all | 50% | 1.1 | 1 | 1.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.1 | 1 |
| Aries (Crystal) | all-to-all | 90% | 1.5 | 1.6 | 1.6 | 1.2 | 1 | 1 | 1.1 | 1.1 | 1 | 1.1 | 1.1 | 1.1 | 1.9 | 1 | 1.2 | 1 |
| Aries (Crystal) | incast | 10% | 10 | 9.8 | 11 | 9.5 | 3.8 | 1.5 | 1.2 | 1.1 | 1.3 | 1.1 | 6.8 | 7.4 | 1.5 | 1.3 | 1.4 | 2.4 |
| Aries (Crystal) | incast | 50% | 39 | 36 | 44 | 35 | 24 | 9 | 5.5 | 4.1 | 5.9 | 7.4 | 35 | 41 | 16 | 13 | 9.6 | 9.7 |
| Aries (Crystal) | incast | 90% | 50 | 46 | 48 | 67 | 15 | 12 | 4.9 | 2.2 | 14 | 12 | 48 | 58 | 55 | 25 | 47 | 20 |
| Slingshot (Shandy) | all-to-all | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | all-to-all | 50% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | all-to-all | 90% | 1.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 10% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 50% | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.1 | 1 | 1.1 | 1 | 1 | 1 | 1 | 1 |
| Slingshot (Shandy) | incast | 90% | 1 | 1 | 1.2 | 1.3 | 1.1 | 1 | 1 | 1 | 1.2 | 1 | 1.2 | 1.1 | 1 | 1 | 1.1 | 1.1 |

# 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

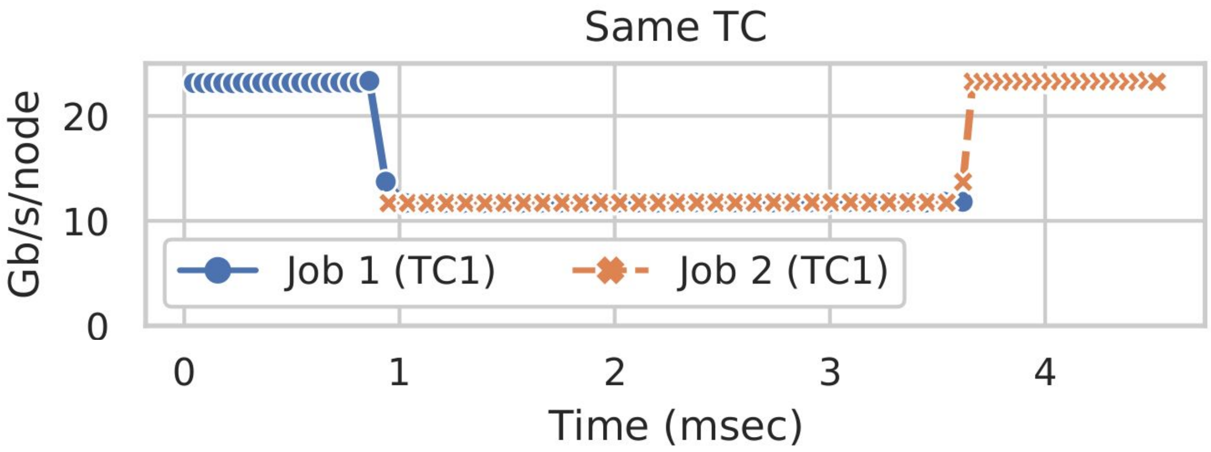
# QOS TESTS

**25%** bandwidth **tapering**

2 jobs running **bisection bandwidth** tests

**TC1**: 80% minimum bandwidth

**TC2**: 10% minimum bandwidth



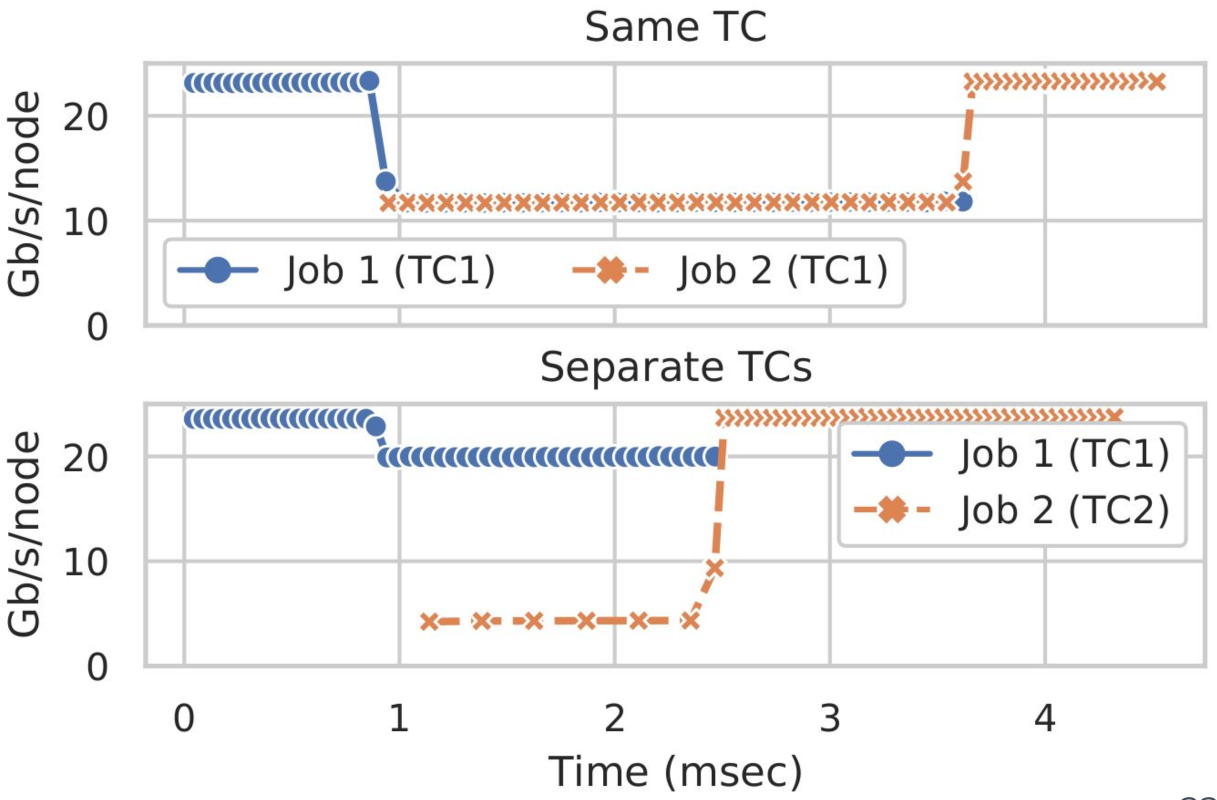
# QOS TESTS

**25%** bandwidth **tapering**

2 jobs running **bisection bandwidth** tests

**TC1**: 80% minimum bandwidth

**TC2**: 10% minimum bandwidth



# CONCLUSIONS

# CONCLUSIONS