Hewlett Packard
Enterprise

The HPE Slingshot 400 Expedition

Slingshot Product Team, HPE

May 4, 2025

HPE Slingshot Interconnect

HPE Slingshot

The Best of Traditional HPC Interconnects

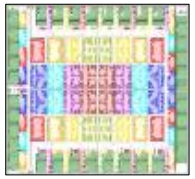
- Low latency protocols
- Efficient for small to large payloads
- Message acceleration offloads

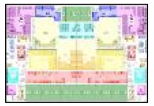
The Best of Ethernet Networks

- Ubiquity
- Interoperability
- Native IP protocol software

Best in Class for AI Workloads

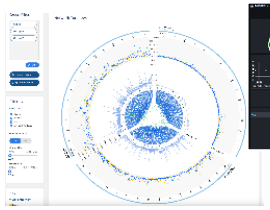
Switch and NIC Silicon and Networking Innovations





Switches and NICs

Systems, Software, and Solutions



HPC Leadership

OAK RIDGE National Laboratory

CSCS

Argonne NATIONAL LABORATORY

Lawrence Livermore National Laboratory

pawsey

EuroHPC Joint Undertaking

CSC

NREL NATIONAL RENEWABLE ENERGY LABORATORY

Los Alamos NATIONAL LABORATORY

KAUST

US Air Force Weather (ORNL)

NCAR

INES

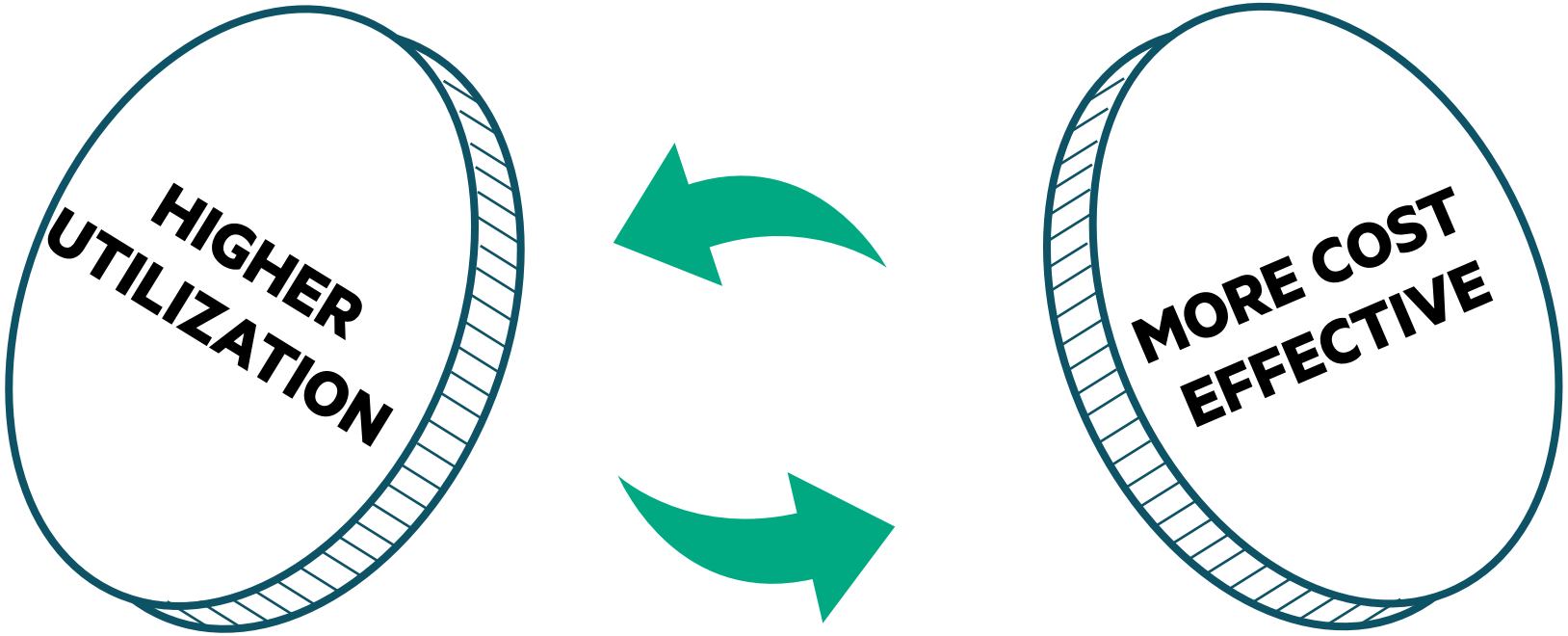
Worldwide adoption, large ecosystem of GPUs, CPUs and NICs, breadth of applications

Nov '24 Top500 w/HPE Slingshot

#1	LLNL El Capitan
#2	ORNL Frontier
#3	ANL Aurora
#5	ENI HPC6
#7	CSCS Alps
#8	CSC LUMI
#10	LANL Tuolumne
#13	LANL Venado
#19	LBNL Perlmutter
#20	Sandia El Dorado
#30	GENCI Adastra
#37	KAUST Shaheen III – CPU
#42	LANL Crossroads
#44	Pawsey Setonix – GPU
#45	Exxon Discovery 5
#46	ANL Polaris
#49	LLNL rzAdams
#62	EPSRC Archer2
#67	Aramco Ghawar-1
#68	ORNL Frontier TDS
#69	Cyfronet Helios GPU
#71	NREL Kestrel CPU
#105	ERDC Carpenter
#106	NREL Kestrel CPU
#114	NCAR Derecho CPU
#117	NOAA Cactus
#118	NOAA Dogwood
#121	KTH Dardel GPU
#138	Skoda C24

Plus 21 more!

Achieving “Efficiency” Through Better Technology



- Run fabric at higher utilization
- Better performance at scale and under load means more utility from expensive computes
- Reduced time/energy-to-solution increases capability

- Minimize cable cost and power with Dragonfly
- Avoid bandwidth overprovisioning to prevent congestion
- Scale and perform without expensive DPUs
- Converge networks
- Eliminate expensive (and unreliable) gateway nodes

And it’s Ethernet!

Fabric Innovations That Maximize RDMA Performance & Minimize Congestion

The HPE Slingshot network is engineered from the silicon-up to deliver high utilization and low bandwidth at any scale, even under heavy load, and with any network interface card.

Fine-Grained Adaptive Routing

Achieve higher utilization by dynamically load balancing both ordered and unordered traffic across many paths to avoid mid-fabric congestion

Flow-based Congestion Control

Achieve consistent and much **lower tail latency under load** by rapidly detecting and mitigating edge congestion, even on short-lived, HPC-sized messages

Ethernet with Flow-based Fairness

Enable **highest performance at scale not possible on typical Ethernet** with consistency regardless of proximity of NICs to switches

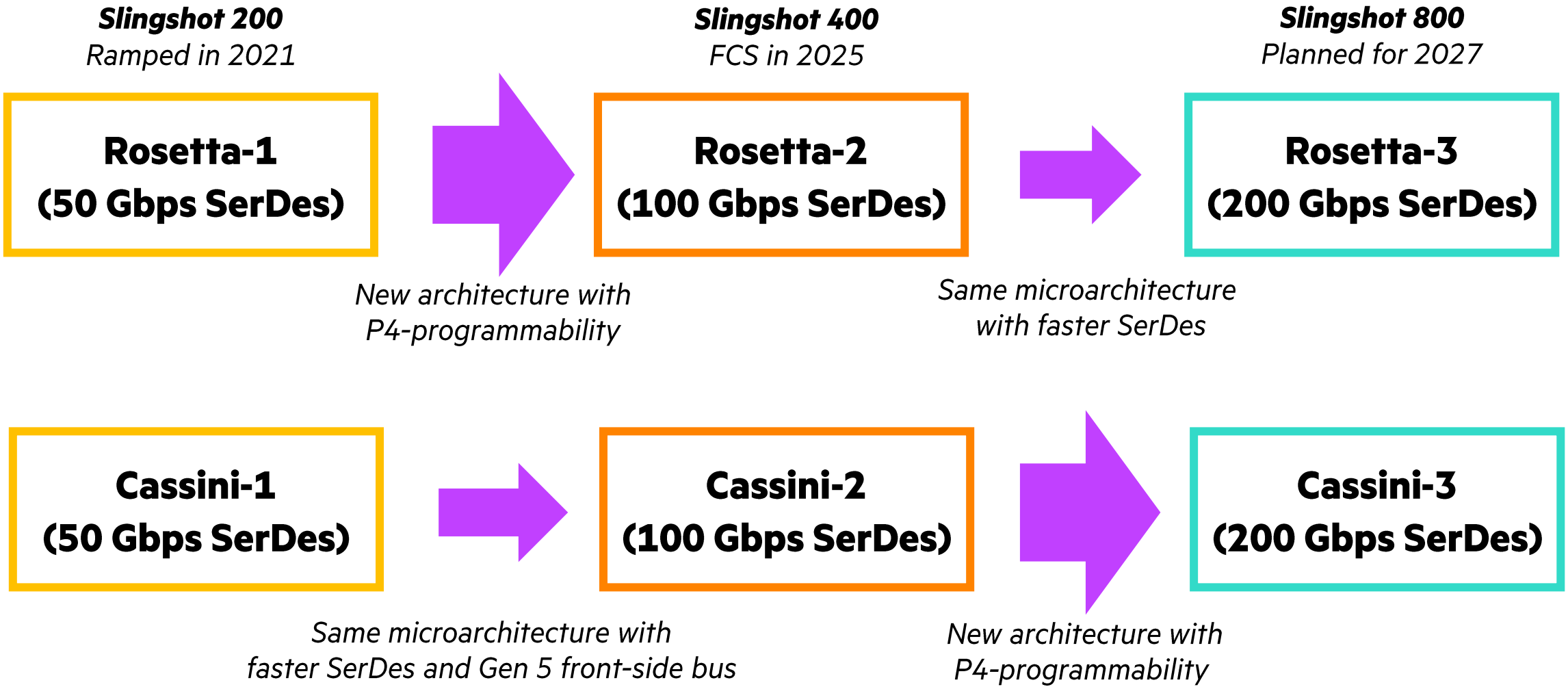
Enhanced Link-layer Reliability

Enables **high performance using Ethernet** by reducing the impact of physical layer signaling errors

Dragonfly Topology

No more than **three switch-to-switch hops** on the minimal path even on the worlds largest Exascale systems

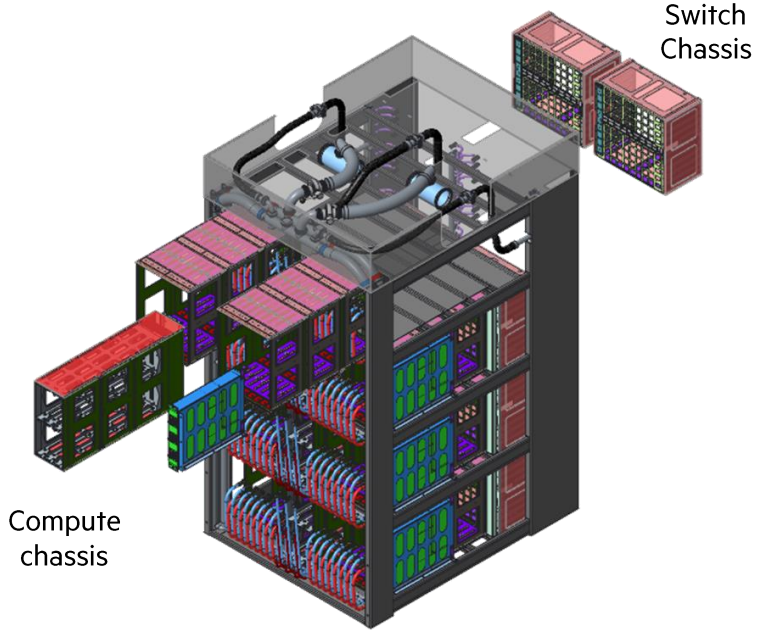
Slingshot Roadmap At-a-Glance



HPE Slingshot 400 Goals

- HPE Cray's 9th generation interconnect architecture
  - Proven leader in HPC for many decades
  - Ethernet 400 Gbps NICs and 64x 400 Gbps switching
  - New switching architecture and improved performance-cost ratio
- Expand beyond HPC and address mixed HPC/AI workloads
  - Classic Cray EX and next-gen Cray GX supercomputers
  - Slingshot for HPE Cray XD and ProLiant clusters
- Sustain leadership in low-latency, scale, and congestion management solutions for Ethernet Supercomputing
- Interoperate with HPE Slingshot installed base
- Introduce option for Fat Tree topology
- Enhance ability to "cloudify" HPC/AI clusters

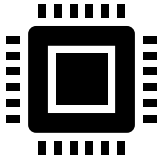
Cray-1 Supercomputer (1976)



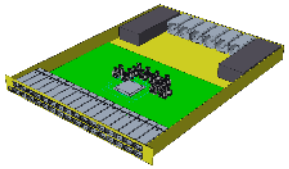
HPE Cray EX

HPE Slingshot 400 Portfolio

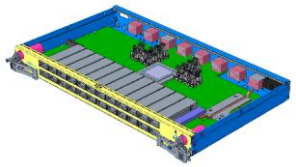
HPE Slingshot 400 Switches



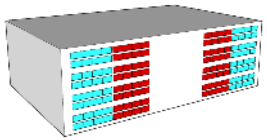
Rosetta-2 ASIC



Top-Of-Rack



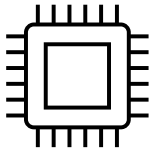
HPE Cray EX Switch Blade



HPE Cray GX Switch Blade and Wired Midplane

- HPE Developed Switching ASIC
- 64 ports @ 400 Gbps (25.6 Tbps switch BW, 51.2T bi-directional)
- 2x200G & 4x100G port breakout
- 100 Gbps PAM4 SerDes, supports 50 Gbps PAM4

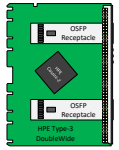
HPE Slingshot 400 NICs



Cassini-2 ASIC

PCIe NIC

HPE Cray EX Mezz



HPE Cray GX OCP Two-NIC Optimized Form Factor

- HPE Developed NIC ASIC
- 400 Gbps Per NIC (800G bidirectional)
- 2 NICs per chip. (PCIe card provides 1 NIC)
- Ethernet, MPI HW offload, Libfabric RDMA
- Software compatibility w/ 1st Generation

Industry NICs

- 400 Gbps NICs (CX7)
- 200 Gbps NICs (4x 50G, 2x 100G)
- 100 Gbps NICs (1x 100G, 2x 50G)

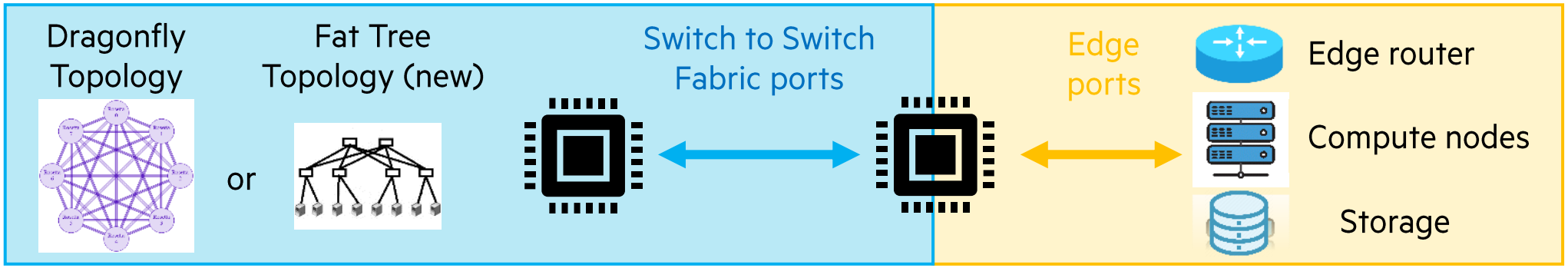
Software

- Next Gen Fabric Management
- Next Gen Advanced Analytics
- HPE Cray and Industry Open-source SW using Libfabric for AI/HPC (HPE Slingshot NIC)
- UCX support on 3rd party NICs

Cables

- QSFP-DD 800G cables:
  - Passive Electric Cables up to 1.6m
  - Active Electric Cables up to 5m
  - Active Optical Cables up to 100m
- Validated interconnection with 1st Generation
- New L1 wired midplane for HPE Cray GX
- OSFP 800G cables & transceivers for Cray GX

HPE Slingshot Rosetta Switch overview



Rosetta Benefits

- Edge and Fabric ports with a single chip
- Ethernet for edge plug-and-play
- Link-Layer enhancements for HPC Ethernet
- Low latency fabric (~300ns)
- Edge and mid-fabric congestion control
  - Fine-grained adaptive routing
  - Credit-based flow control
  - Lowest tail latency under load
- Versatility for use in standalone compute and supercomputer clusters

Rosetta-2 Enhancements

- Improved Forwarding and Scalability with up to 260K endpoints
- Improved end-to-end Congestion Management
- Multi Fabric RDMA
  - Slingshot Transport across 2 generations
- More Traffic Classes
  - Enables dedicated classes for classic Ethernet
- ACL support for HPC/Cloud isolation & security
- Fat Tree Topology support
- Programmable data path (P4)
  - Flexibility for hardening and enhancing match action rules

Rosetta-2 vs Rosetta-1

	Rosetta-1	Rosetta-2
Si Node	16nm	7nm
Ports (Bandwidth)	64x200G (12.8T, 25.6T bidirectional)	64x400G (25.6T, 51.2T bidirectional)
SerDes Speed	56G/28G	112G/56G/28G
Port Splitting	No, 64-radix	Yes (2x, 4x), up to 256-radix
Latency	300ns	300-350ns
Input Buffer	24MB	64MB
Port groups (fwd'ing granularity)	32 (2 ports each)	64 (1 full, 2 half-split, or 4 quarter-split ports each)
Exact Match Table	8K per port group	Up to 384K, 8K cache per port group
LPM Rules	2K per port group	2K per port group
ACL Rules	-	Ingress from hash/LPM, Egress 1K per port group
ARP Table	256 per port group	1K per port group
LAG Table	4K per port group	8K per port group
Traffic Classes	8 (4 bidirectional Slingshot Traffic)	12 (4 bidirectional Slingshot Traffic + 4 Ethernet)

HPE Slingshot 400 Cassini NIC

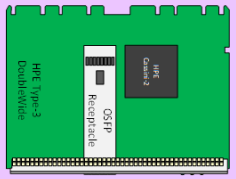
- **ASIC Enhancements beyond Cassini-1**
  - Double host bandwidth and link bandwidth
    - PCIe Gen5 x16 per core (dual core ASIC)
    - 100Gbps signaling (also 50G and 25G)
    - Single 400G network port per core
      (200G and 100G also supported, compatible with Rosetta-1)
  - Same NIC software stack as Cassini
  - Same MPI and PGAS features as Cassini
    - Nominal reductions in latency due to higher clock rate
    - Some improvements in matching rate from improved clock rate
    - Improved throughput where PCIe-Gen4 or network was the limit

Metric	Cassini-1	Cassini-2	Reason
PGAS, 8B	435 Mmsgs/s	**500 Mmsgs/s**	PCIe Gen4 => Gen5
PGAS, 64B	192 Mmsgs/s	**421 Mmsgs/s**	Gen4=>Gen5, 200=>400 Gb/s
MPI, 8B	100 Mmsgs/s	**110 Mmsgs/s**	Clock rate increase (~ 10%)
MPI, 1KB	24 GB/s	**47 GB/s**	200 => 400 Gb/s

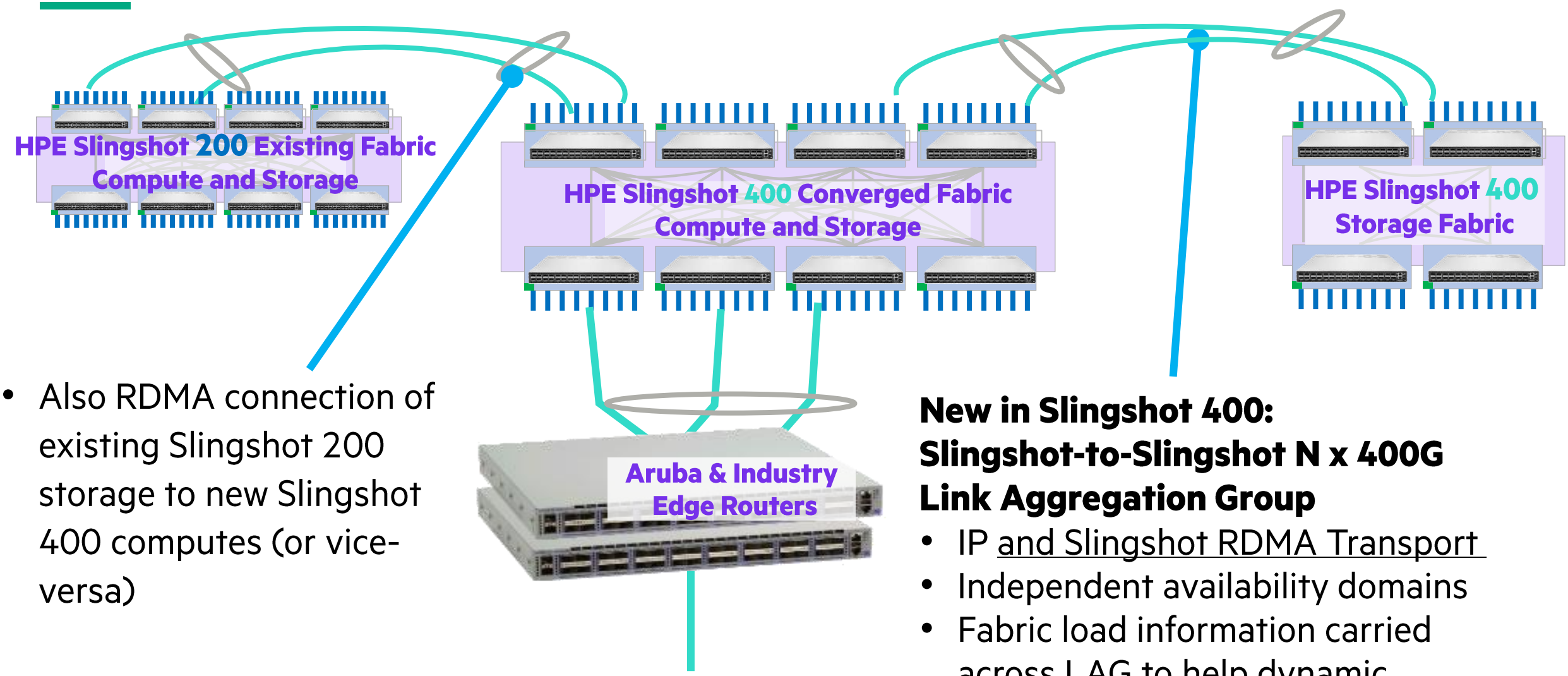
ASIC Performance Fixes

- 8K TRS entries (was 2K) to reduce NO_TRS
- Additional options for controlling setting of RO=1
  - All types of traffic on receives
  - "Get" response can now be set to RO=1 writing to host memory
- Fix ODP bugs (shouldn't need work-arounds)

- **Products**
  - EX Mezz. card (SA420M)
    - 2 x 400 Gbps
    - 2 x PCIeGen-5 to compute
  - OCP-W card (SA420OW)
    - 2 x 400Gbit OSFP
    - 2 x PCIe Gen5 to compute
  - PCIe-Gen5 NIC (SA410S)
    - 1 x 400Gbit QSFP
    - 1 x PCIe Gen5



New In HPE Slingshot 400: Slingshot Transport Across Fabric Edge



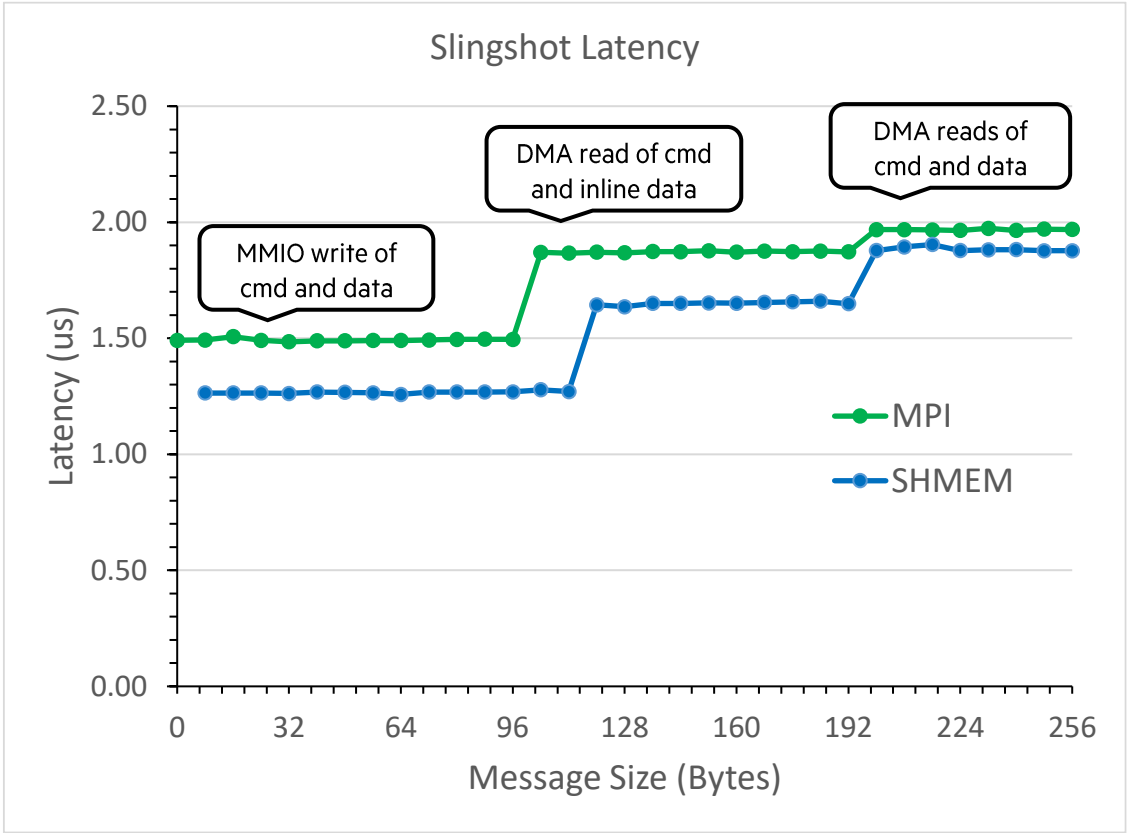
- Also RDMA connection of existing Slingshot 200 storage to new Slingshot 400 computes (or vice-versa)

New in Slingshot 400:
Slingshot-to-Slingshot N x 400G Link Aggregation Group

- IP and Slingshot RDMA Transport
- Independent availability domains
- Fabric load information carried across LAG to help dynamic

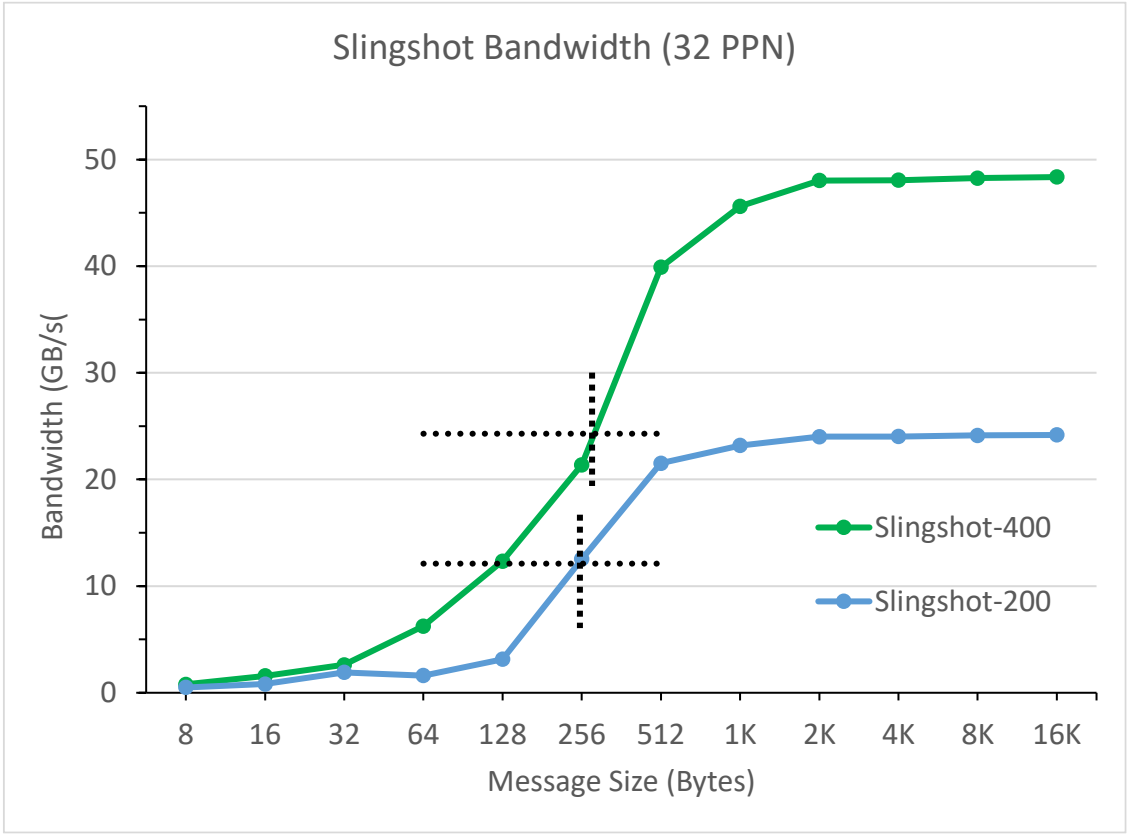
Quiet Network Latency

- Results are for NIC-switch-NIC path
- Performance is CPU type specific
  - Data is for Genoa, EPYC 9654, 2400MHz
- Steps illustrate protocol changes
  1. MMIO write of command and inline data (≤128 bytes)
  2. DMA read of command and inline data (≤ 256 bytes)
  3. DMA reads of command and data (> 256 bytes)
- Additional system costs
  - ~300ns for each additional switch hop
  - ~100ns for each active optical cable
- System impact (dragonfly)
  - Group of up to 512 endpoints + 300ns
  - System of any size + 1000ns



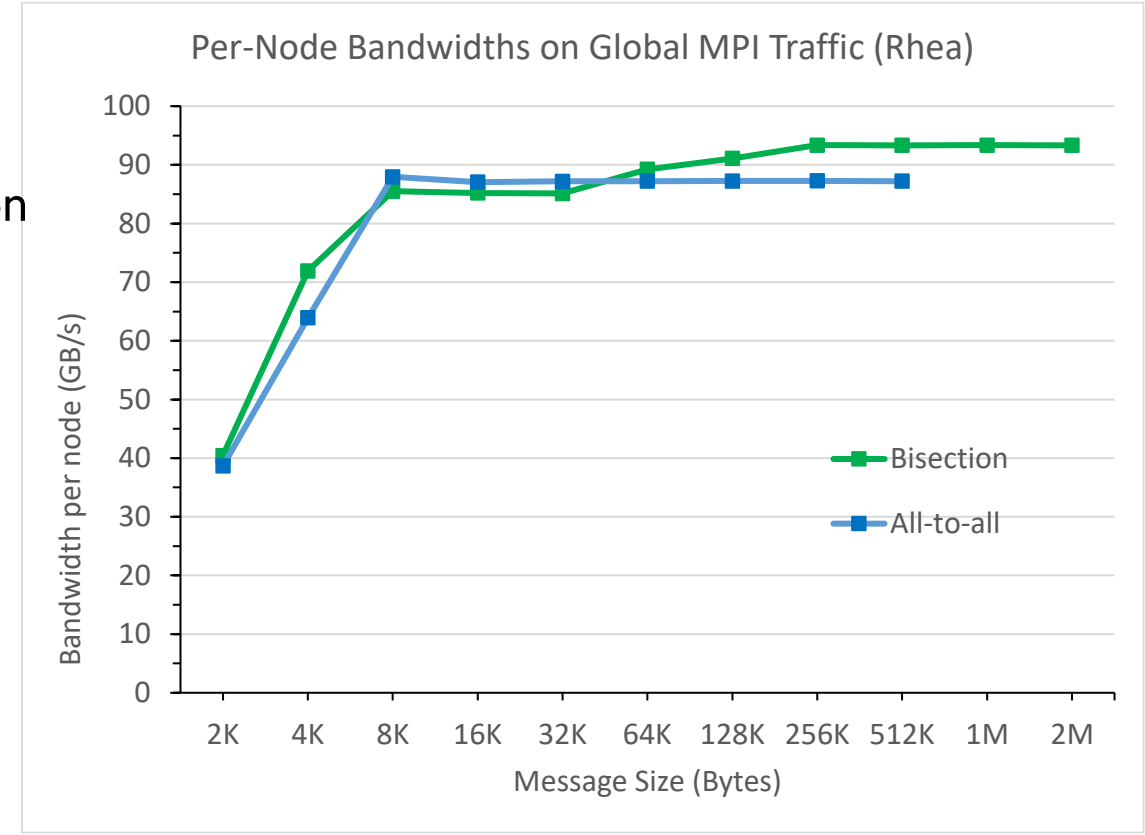
Point-to-Point Bandwidth

- Multiple processes per node, unidirectional traffic
  - MPI payload bandwidth after protocol overheads
- Good small message bandwidths
  - Half peak bandwidth at ~256 bytes
- Performance is largely independent of CPU type
- No large system effects
  - DMA engines issue enough transfers to cover the network round trip time



System Performance

- Tests were performed on Rhea
  - 64 dual NIC nodes with AMD EPYC 9654 (Genoa) CPUs
- Bisection traffic
  - Stresses non-minimal paths but doesn't cause congestion
- All-to-all traffic
  - Stresses minimal paths, causes congestion
- Sustained MPI payload bandwidth is ~90% of peak



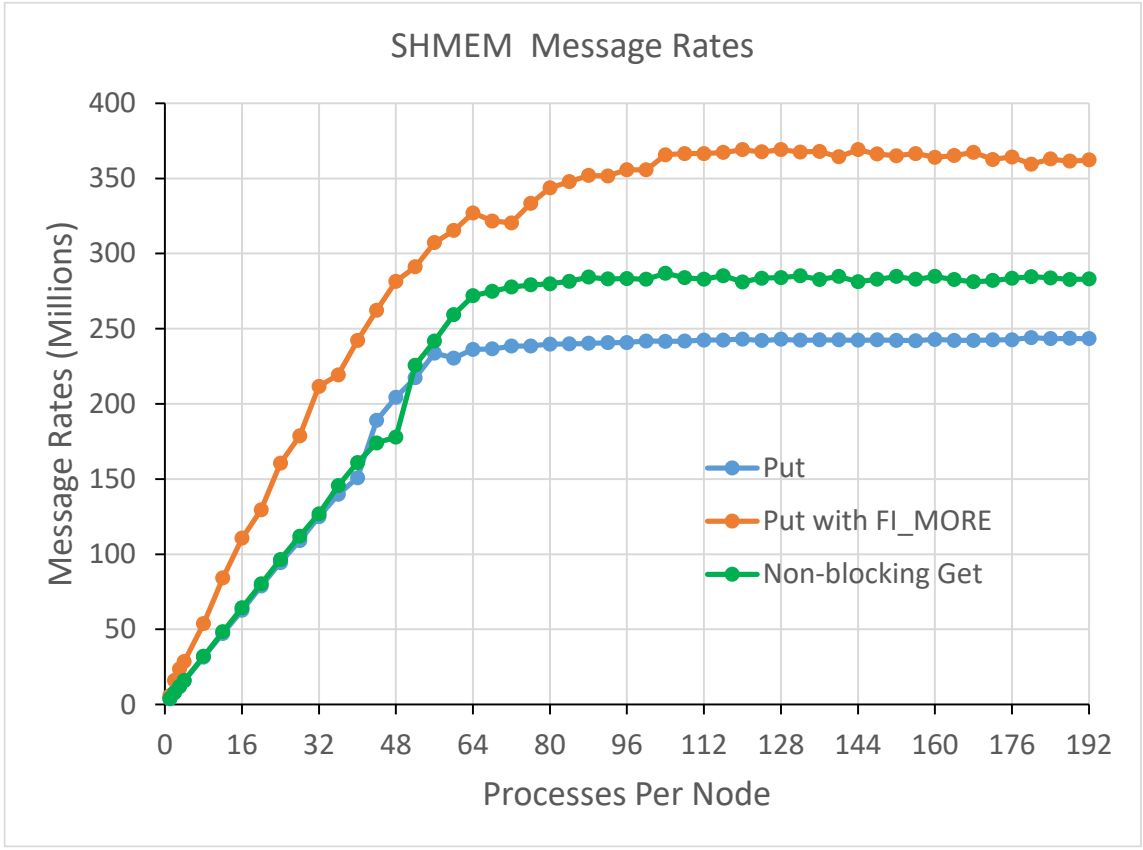
SHMEM Message Rates

- Performance is CPU type specific
  - Data is for Genoa, EPYC 9654, 2400MHz with Cassini-2

- Rates on 1-1 traffic are limited by PCIe write performance
- Read limited tests achieve higher performance
  - 465M Puts/sec on 1-to-4 traffic
  - 450M Gets/sec on 4-to-1 traffic

- Results show the value of Libfabric FI_MORE
  - Reduces software & CPU overheads

- Results show that Get rate related changes from Cassini-1 are effective in Cassini-2



Ethernet Performance

- Microsoft tcp & iperf2 tests
- Between 2 nodes in Rhea
  - AMD EPYC 7654 (Genoa) system

- Minimal tuning

```
ethtool -G hsn0 rx 8192
ethtool -G hsn0 tx 8192
ethtool -L hsn0 rx 16
ethtool -L hsn0 tx 16
ip link set dev hsn0 txqueuelen 2000
```

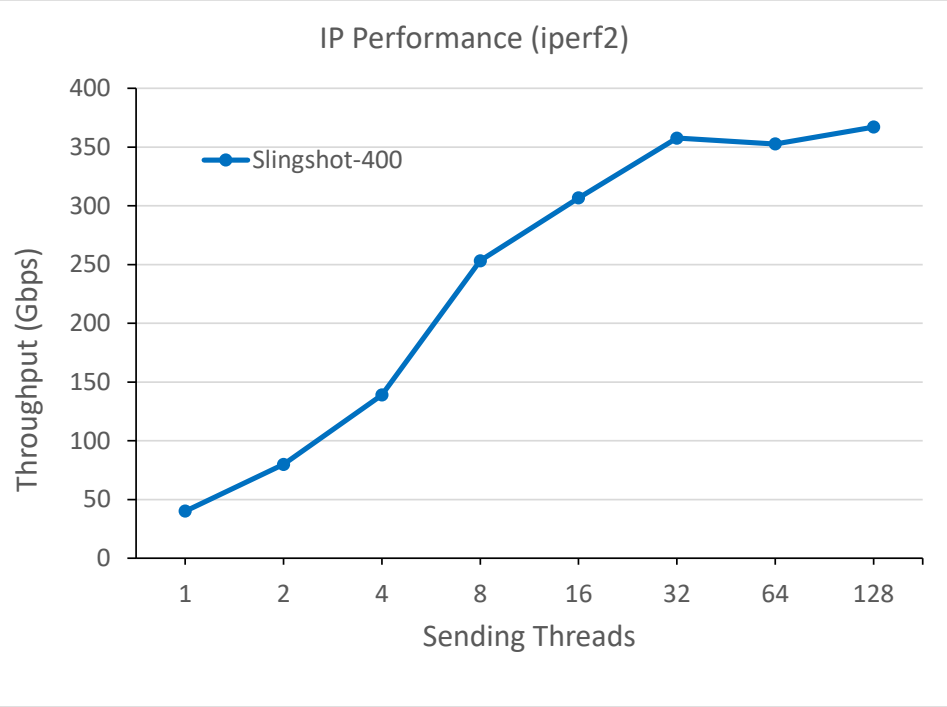
- Throughput of 364 Gbps

Cmd on sender:
$ numactl -N 3,2,1,4,5,6,7 ./ntttcp -s 10.150.0.59 -b 1048576 -W 10
Cmd on receiver:
$ numactl -N 3,2,1,4,5,6,7 ./ntttcp -r -W 10

NTTTCP for Linux 1.4.0

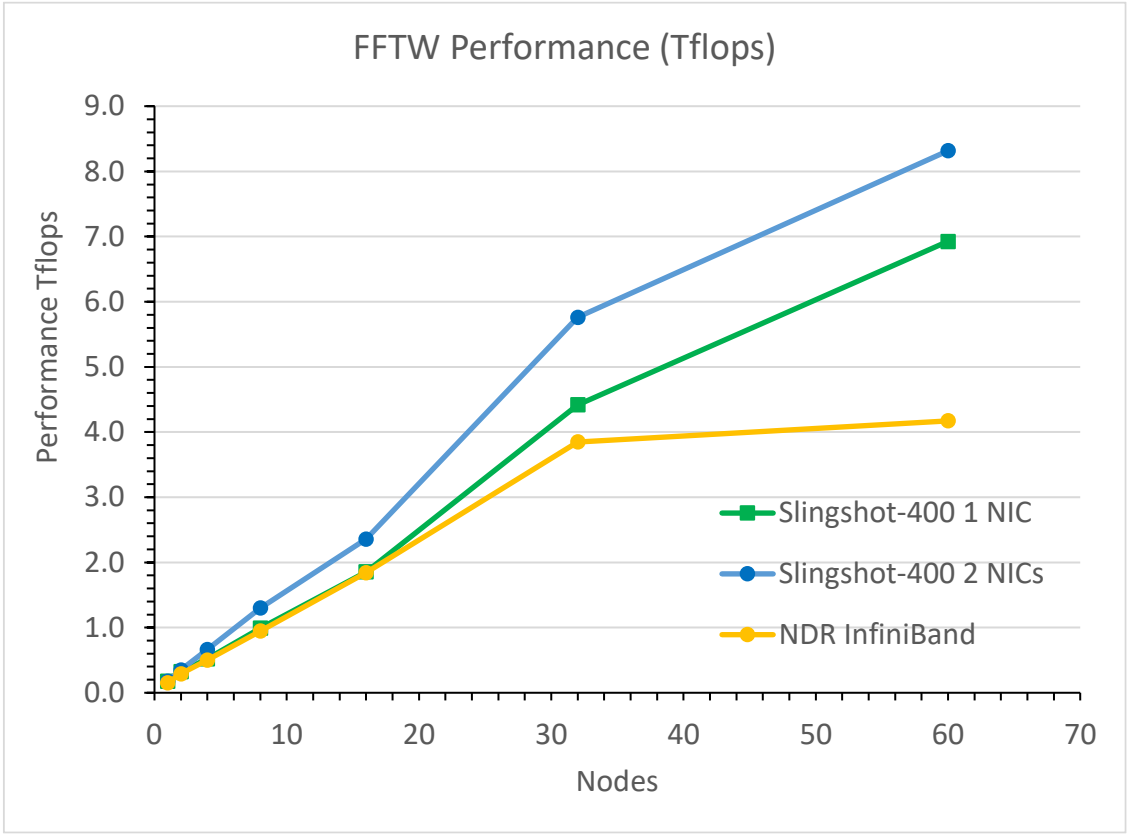
11:23:41 INFO: 17 threads created
11:23:50 INFO: Test cycle time negotiated is: 70 seconds
11:23:50 INFO: Network activity progressing...
11:24:00 INFO: Test warmup completed.
11:25:00 INFO: Test run completed.
11:25:00 INFO: Test cycle finished.
11:25:00 INFO: ##### Totals: #####
11:25:00 INFO: test duration : 60.00 seconds
11:25:00 INFO: total bytes : 2728715901064
11:25:00 INFO: throughput :363.83Gbps
11:25:00 INFO: retrans segs :0
11:25:00 INFO: cpu cores :384
11:25:00 INFO: cpu speed :3694.474MHz
11:25:00 INFO: user :0.05%
11:25:00 INFO: system :3.36%
11:25:00 INFO: idle :94.41%
11:25:00 INFO: iowait :0.00%
11:25:00 INFO: softirq :2.19%
11:25:00 INFO: cycles/byte :1.74
11:25:00 INFO: cpu busy (all) :1305.12%

- Iperf2
  - 40 Gbps with a single thread
  - 350 Gbps with 32 threads



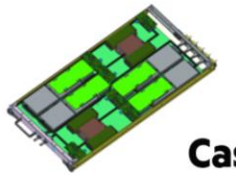
Comparing Slingshot 400 with NDR InfiniBand on FFTW

- Measurements on AMD EPYC 7654 (Genoa) systems
  - Same CPUs and memory
  - NDR InfiniBand, full bandwidth Fat-Tree
  - Slingshot 400, full bandwidth Dragonfly
- Realistic benchmarking
  - Cray MPI and Open MPI on both systems
  - Rendezvous protocol on or off on both systems
  - Local shared memory optimizations on
  - Select the highest performing option for each datapoint
- Slingshot 400 outperforms NDR at 32 nodes or above
- 30% gain in application performance from two Slingshot 400 NICs



The HPE Slingshot Heritage in UEC

Rosetta Switch



Cassini NIC

Hewlett Packard Enterprise

Slingshot Fabric Technology → Ultra Ethernet Consortium

HPC Ethernet (w BROADCOM)

>75% of UEC based on Slingshot (with HPC Ethernet)

Since the beginning, **HPE Slingshot's** focus is to address limitations of InfiniBand on scalability, reliability, and inter-operability **with an Ethernet capable solution**

Key foundational features, developed by the Slingshot team, are part of the UEC draft standard and are already proven and shipping in Slingshot:

- ☑ Ethernet as a connectionless protocol, with less state overhead and more memory available for applications
- ☑ Industry-adopted Libfabric north-bound API
- ☑ Slingshot Transport with Credit-based Flow Control and Link Layer Retry to achieve lossless/low-loss networks
- ☑ Advanced congestion management, flexible packet ordering, and lowest tail latency for MPI/RDMA messages
- ☑ Telemetry (Fabric Monitoring)

Thank you

jesse.treger@hpe.com