

Fabric Convergence Implications on Systems Architecture

Kevin Leigh (HP)

Partha Ranganathan (HP Labs)

Jaspal Subhlok (U. of Houston)



Executive summary

- A converged fabric
 - can support data, storage and cluster networking
 - is desirable for cost and management advantages
- Fabric convergence is not new, but recent trends motivate revisiting
 - High-speed fabrics, physical layer similarities, blade servers, ...
- In this presentation, we will
 - describe various aspects of fabric convergence
 - present some case-study eval results of two fabric convergence methods
 - discuss challenges/opport. for future research, esp. implications to sys arch
- Our goal
 - to initiate examination of general issues in a broader academic community

Contents

- Part I: Introduction
 - Fabric convergence and related hardware infrastructure
- Part II: Case study
 - Network consolidation vs. I/O consolidation
- Part III: Future challenges and opportunities
- Closing remarks

All opinions in this positioning paper represent the views of the authors and do not represent official HP positions on these subjects.

Fabric convergence

- Fabric protocols can have
 - Memory semantics: PCIe (I/O); HT, QPI (processor links), IB
 - Network semantics: Ethernet, FC, SAS, IB, ...
- Traditional and emerging fabrics usage
 - Ethernet – Data communication networks
 - SAS, FC – Local/remote storage networks
 - Ethernet, IB – Cluster networks
 - PCIe – shared I/O [initially intended for local I/O]
- Fabric convergence means
 - A fabric for multiple usages (data, storage, cluster)
 - We define fabric convergence at three levels – nw, I/O, coherent

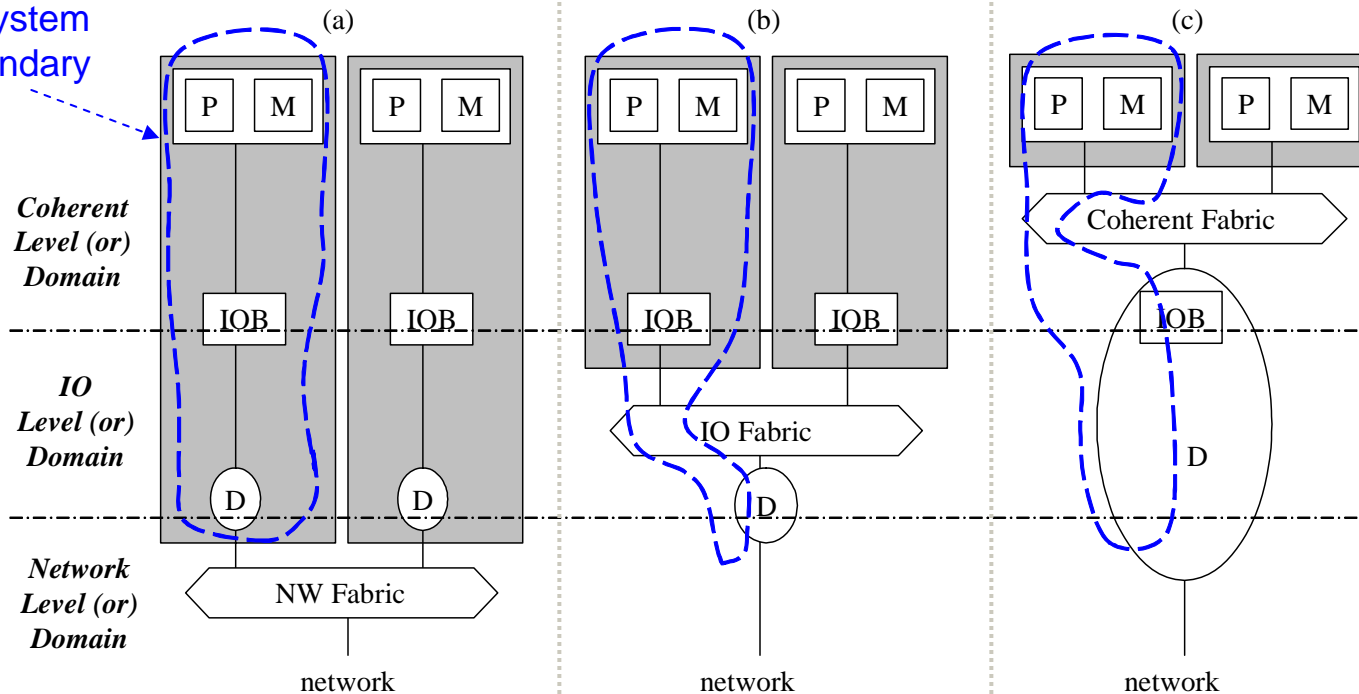
Three fabric convergence levels

Replacing multiple network protocols with one – network consolidation

Replacing multiple network protocols with I/O consolidation

Replacing multiple network protocols with I/O consolidation via coherent fabric.

A system boundary

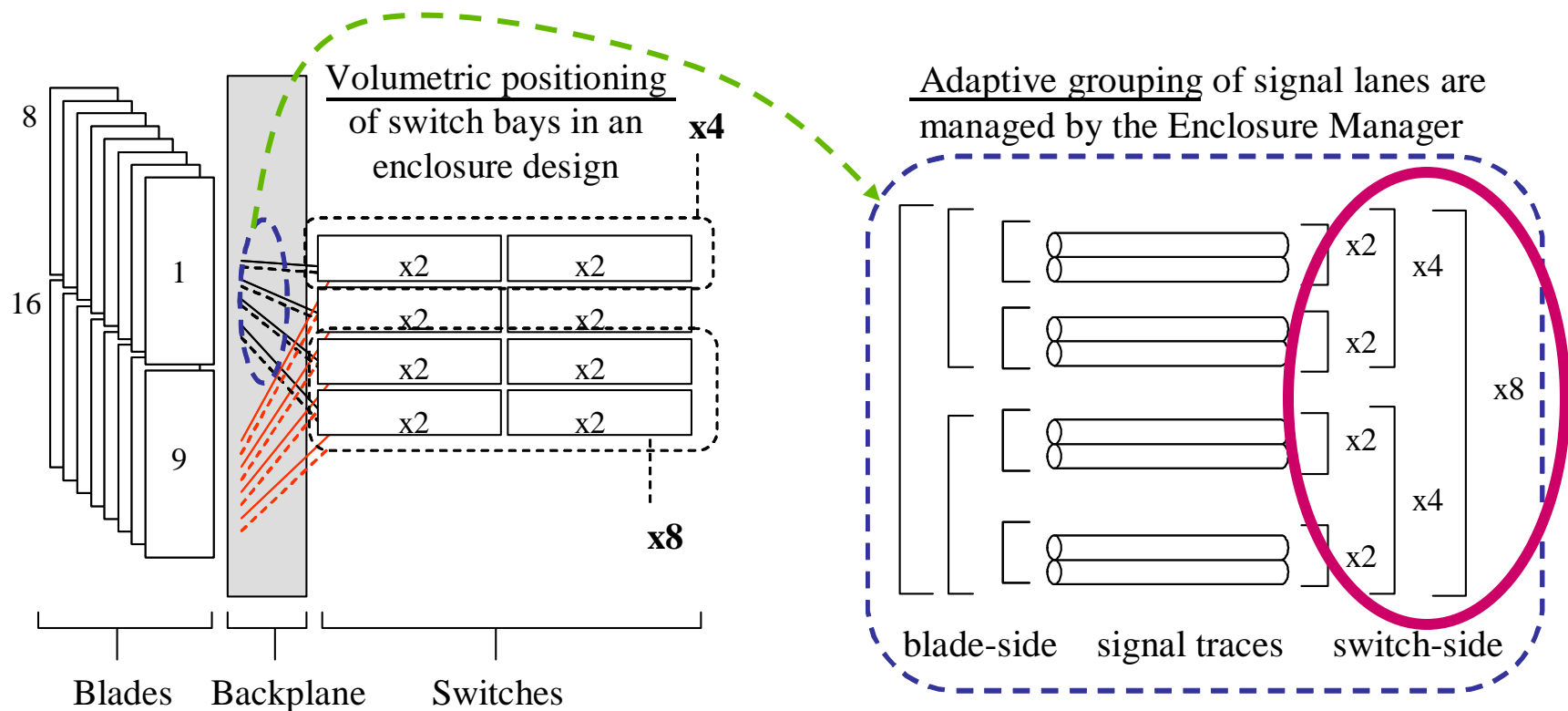


P: Proc
M: Mem
IOB: I/O Bridge
D: Device

- High-level system architecture implication
 - Flexible (logical) system boundaries in (b) and (c)
- Case-study on (a) vs. (b); (c) for future work

Relevant hardware infrastructure for our case study

- Enable fabric convergence at different levels with one phy infrastructure
- Achieved by using blade encl as a General-Purpose Infrastructure (GPI)



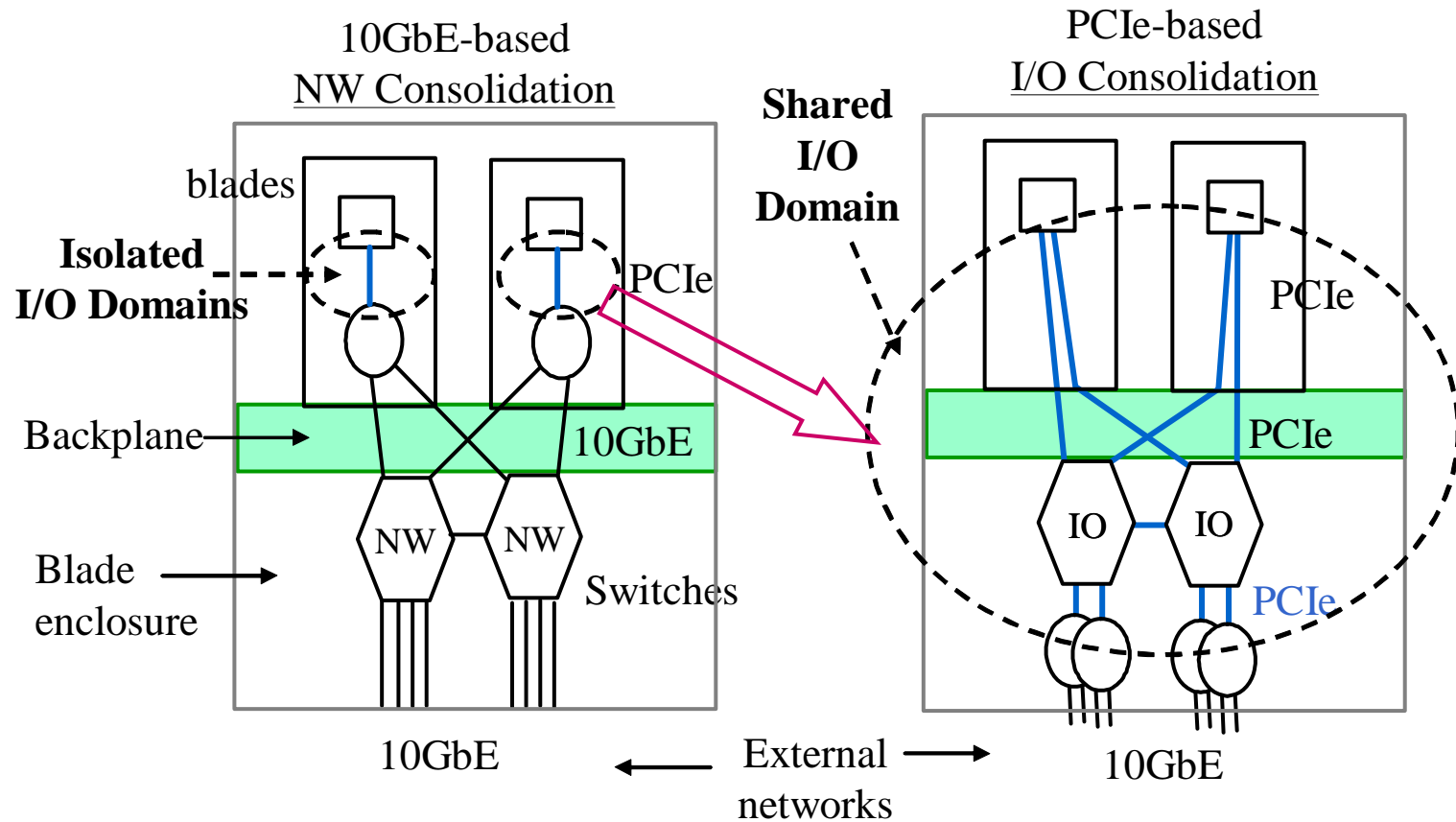
K. Leigh, P. Ranganathan, J. Subhlok, "General-Purpose Blade Infrastructure for Configurable System Architectures," Journal on Parallel and Distributed Databases, March 2007.

Contents

- Part I: Introduction
 - Fabric convergence and related infrastructure
- Part II: Case study
 - Network vs. I/O consolidation
- Part III: Future challenges and opportunities
- Closing remarks

Network vs. I/O consolidation

- I/O consolidation reduce components, but has new issues
 - Isolated local I/O domains in each system → a shared I/O domain



Challenges to evaluate NW vs. I/O consolidation

I/O consld. is new for mainstream servers

Metrics Challenges

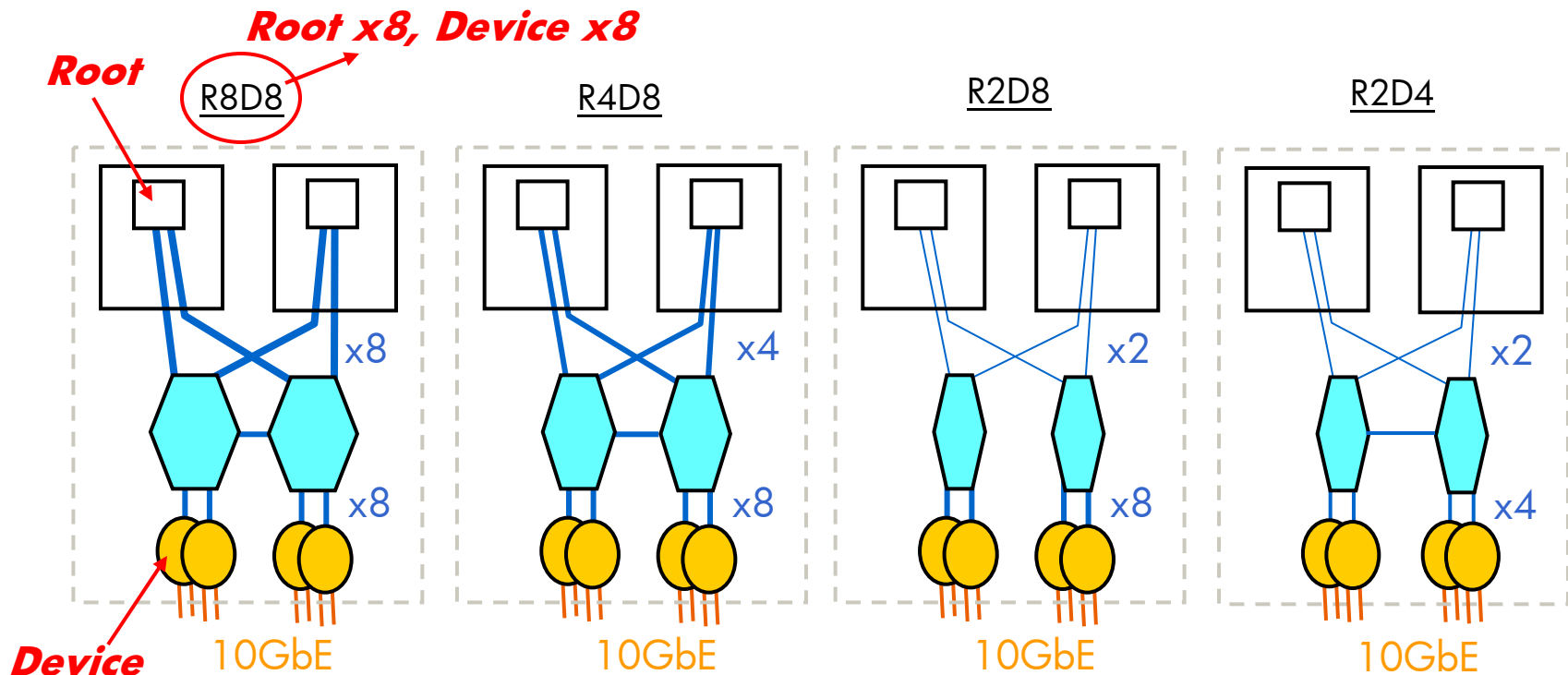
- What configurations and metrics to evaluate I/O consolidation?
 - For different applications, workloads, ...

Methods Challenges

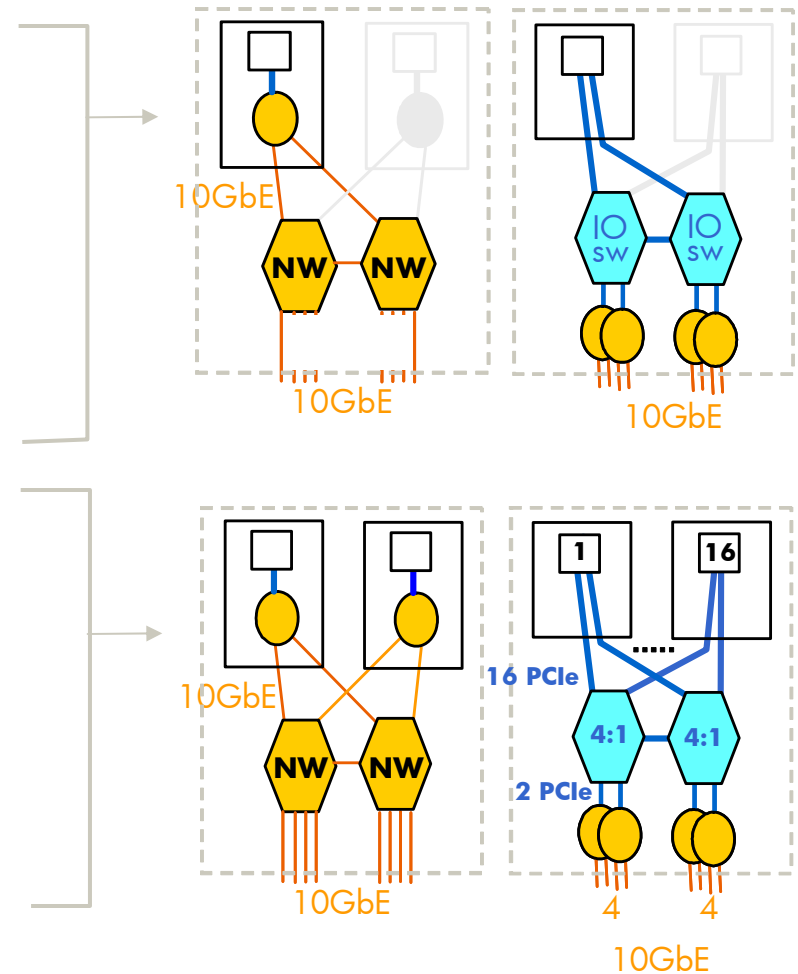
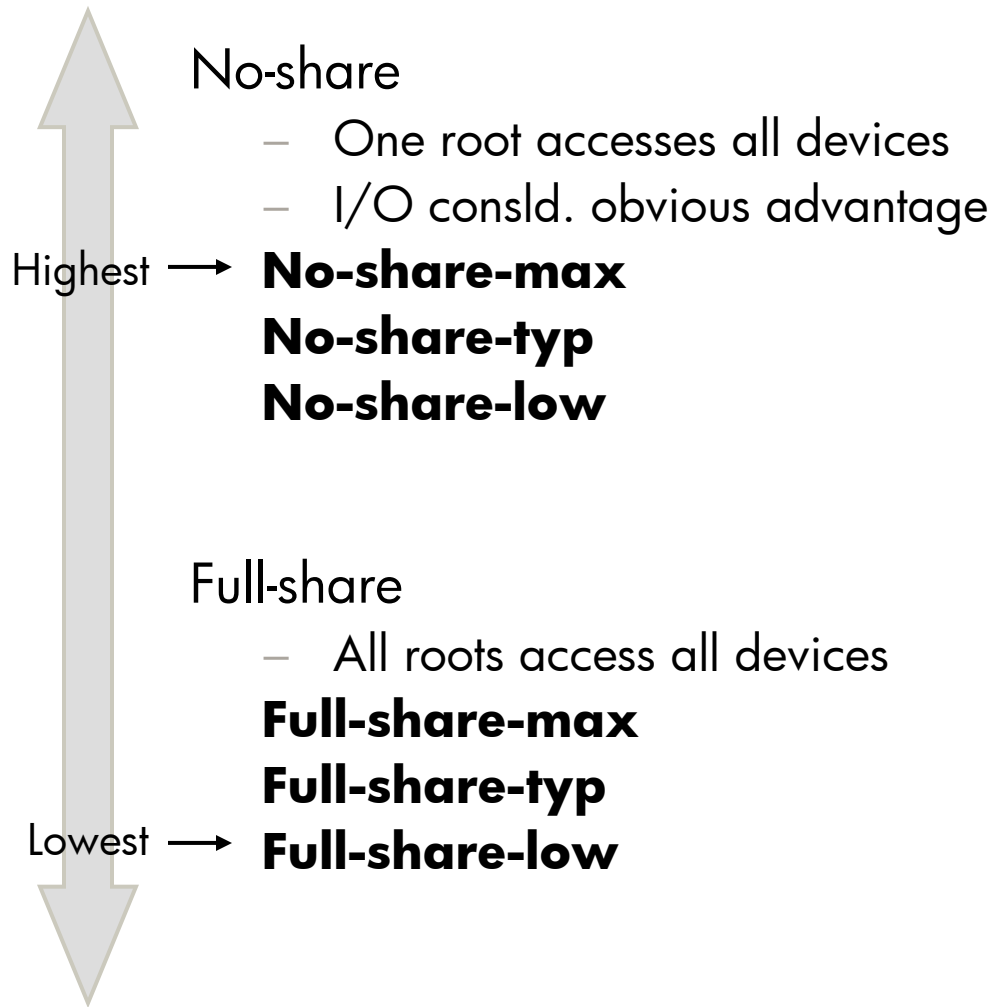
- No existing off-the-shelf full HW/SW systems and workloads
- Even so, inter-dependent param space is very large
- Simulating the entire I/O consld. env is impractical
- Building proposed arch in full HW systems is impractical
- Will describe a hybrid method – HW emulation + simulation + ...

Practical configs: I/O consolidation

- NW and I/O consolidation in same infra
- Practical configs ← BW over-prov at roots, and over-subs of switches
- I/O link-widths significant impact I/O switch
 - Cost, performance, availability (inter-switch cross-links, more switches)



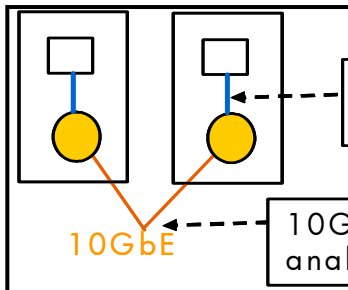
Bracketing Performance: Metric definitions



Max, Typ and Low conditions depend on parameters that can affect bandwidth

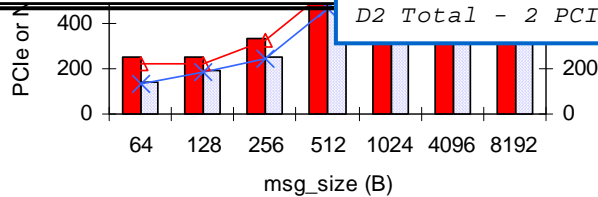
- Link-width ratio, switch latency, sharing penalty, ...

Methodol



per-root BW
Cut-thru PC

Root link width
Root BW max (Gbps)
Switch BW over-sub ratio
Optimal # of roots sharing a dev
Device link width
Rel Non-share max BW
Rel Full-share max BW
Rel Full-share BW (w/ Throughput-Overhead)
And so on...

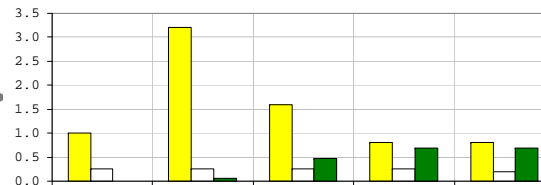


■ PCIe_up x8 sw0 ■ NW_Rx x8 sw0
—x PCIe_up x8 sw1 —x NW_Rx x8 sw1

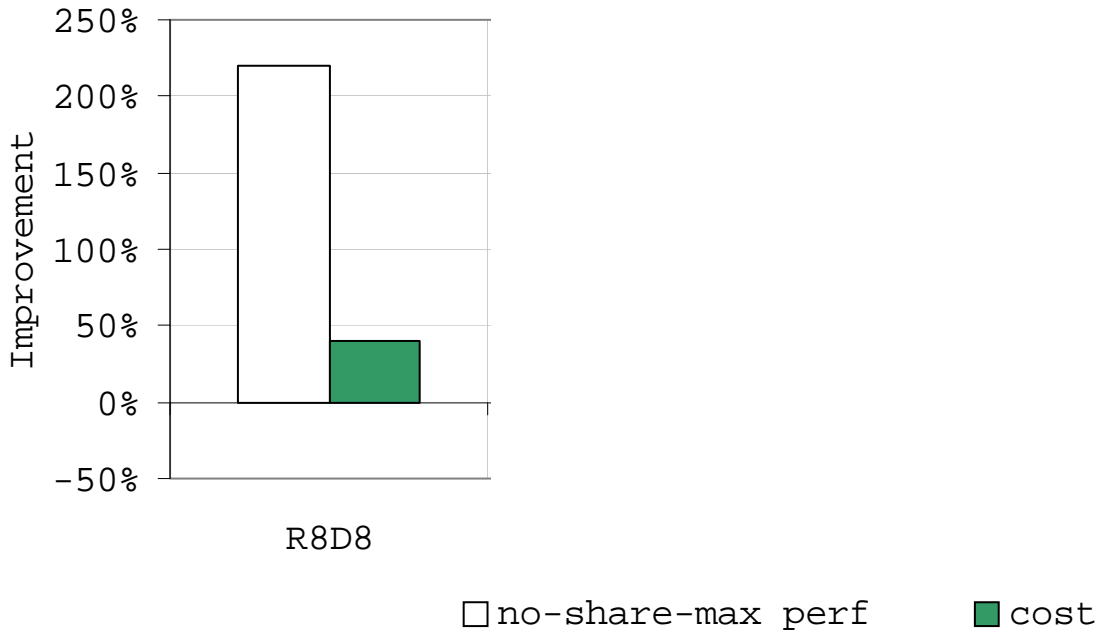
Normalized Costs	# of Blades			
	4	8	12	16
10-10 [10GbE NICs and switches w/4:1 over-sub ratio]				
10GbE dual-port NIC [16 total]	0.07	0.15	0.22	0.29
10GbE 24-port switch [2 total]	0.71	0.71	0.71	0.71
10GbE dual-port NIC [16 total]	0.07	0.15	0.22	0.29
10GbE 24-port switch [2 total]	0.71	0.71	0.71	0.71
<i>N Total with 2 10GbE sw (4 ext ports per sw)</i>	0.78	0.85	0.93	1.00
<i>N2 Total with 4 10GbE sw (8 ext ports per sw)</i>	1.56	1.71	1.85	2.00
R8D8 [x8 per blade; x8 for shared-IO dev]				
PCIe I/F (redrivers, etc.)	0.02	0.05	0.07	0.09
PCIe 96-lane 12-port switches [4 total]	0.72	0.72	0.72	0.72
10GbE dual-port shared-NIC (with premium) [4 total]	0.03	0.06	0.09	0.12
10GbE dual-port shared-NIC (with premium) [4 total]	0.03	0.06	0.09	0.12
<i>A Total - 2 PCIe sw (2 shared NICs, 4 ext ports per sw)</i>	0.77	0.83	0.88	0.93
<i>A2 Total - 2 PCIe sw (4 shared NICs, 8 ext ports per sw)</i>	0.80	0.88	0.97	1.05
R4D8 [x4 per blade; x8 shared-IO dev]				
PCIe I/F (redrivers, etc.)	0.01	0.02	0.04	0.05
PCIe 96-lane 20-port switches [2 total]	0.36	0.36	0.36	0.36
10GbE dual-port shared-NIC (with premium) [4 total]	0.03	0.06	0.09	0.12
10GbE dual-port shared-NIC (with premium) [4 total]	0.03	0.06	0.09	0.12
<i>B Total - 2 PCIe sw (2 shared NICs, 4 ext ports per sw)</i>	0.40	0.44	0.48	0.52
<i>B2 Total - 2 PCIe sw (4 shared NICs, 8 ext ports per sw)</i>	0.43	0.50	0.57	0.64
R2D8 [x2 per blade; x4 or x8 shared-IO dev]				
PCIe I/F (redrivers, etc.)	0.01	0.02	0.04	0.05
PCIe 48-lane 20-port switches [2 total]	0.14	0.14	0.14	0.14
10GbE dual-port shared-NIC (with premium) [4 total]	0.03	0.06	0.09	0.12
10GbE dual-port shared-NIC (with premium) and sw [4 total]	0.17	0.19	0.22	0.25
<i>C Total - 2 PCIe sw (2 shared NICs, 4 ext ports per sw)</i>	0.18	0.22	0.26	0.30
<i>C2 Total - 4 PCIe sw (2 shared NICs, 4 ext ports per sw)</i>	0.34	0.41	0.48	0.55
R2D4 [x2 per blade; x4 or x8 shared-IO dev]				
PCIe I/F (redrivers, etc.)	0.01	0.02	0.04	0.05
PCIe 48-lane 20-port switches [2 total]	0.14	0.14	0.14	0.14
10GbE dual-port shared-NIC (with premium) [4 total]	0.03	0.06	0.09	0.12
10GbE dual-port shared-NIC (with premium) [4 total]	0.06	0.12	0.18	0.24
<i>D Total - 2 PCIe sw (2 shared NICs, 4 ext ports per sw)</i>	0.18	0.22	0.26	0.30
<i>D2 Total - 2 PCIe sw (4 shared NICs, 8 ext ports per sw)</i>	0.24	0.34	0.44	0.54

NW vs. IO consolidation Perf/Cost trade-offs

Results



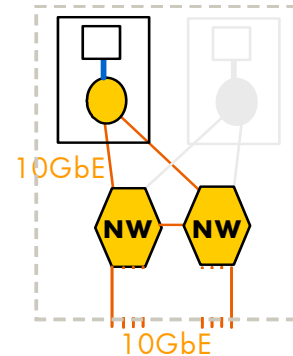
(No-share-max) Performance/cost results



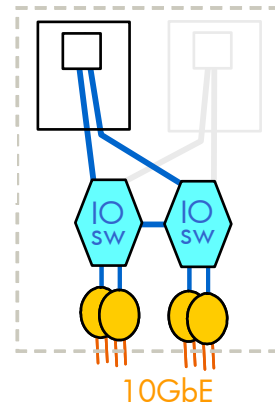
- R8D8: **203% perf gain @ 39% cost saving**
- R4D8: **52% perf gain @ 53% cost saving**
- R2D8: **24% perf loss @ 60% cost saving**
- R2D4: **24% perf loss @ 60% cost saving**

I/O consolidation can be more cost/perf effective

Compared to 10-10



I/O cons config



Some insights learned from our case-study

- Smaller (x2) I/O fabrics enable
 - Finer granular cost/performance scalability
 - Higher level of fabric isolations across shared devices
- Largest max_payload_size not always provide best perf eff
 - Larger than 512B provided no significant benefits
- Cut-thru can be slower than store-n-fwd mode in I/O switch
 - Link width differences can throttle on the smaller link

Contents

- Part I: Introduction
 - Fabric convergence and related infrastructure
- Part II: Case study
 - Network vs. I/O consolidation
- Part III: Future challenges and opportunities
- Closing remarks

Challenges and opportunities

- System architecture implications
 - **Disaggregated system resources** (proc, mem, HDD, I/O)
 - Independent scaling of system, I/O, infrastructure
 - New optimal design sizing for various applications
 - New issues due to overlapping domains (clk, perf, security, usage, ...)
 - **Level-merging of converged fabrics**
 - I/O consolidation eliminates network fabrics
 - Coherent fabrics eliminates I/O fabrics and network fabrics
 - **Hierarchical fabrics**
 - Opportunities to nest ensembles of systems to share resources
 - **Heterogeneous fabric consolidations**
 - Opportunities to mix different types of fabric consld in an architecture
 - **Address Translations**
 - Sizing, placement, algorithms, protocols

Challenges and opportunities (contd.)

- Resource management
 - **Multi-level resource allocations w/ heterogeneous envr.**
 - Deterministic bandwidths and latencies
 - Sizing of buffers, payloads, flow control credits
 - **Hardware/software layer interoperability**
 - Combinations of hardware (proc, chipsets, devices, firmware)
 - Combinations of software (apps, OS/stacks, dev drivers, option ROM)

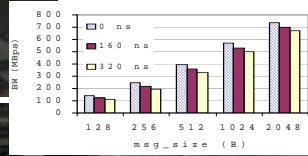
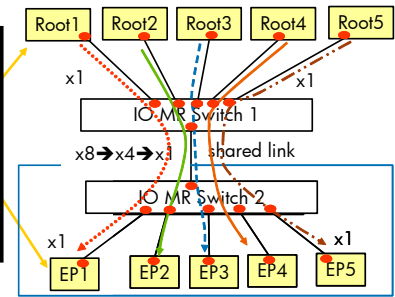
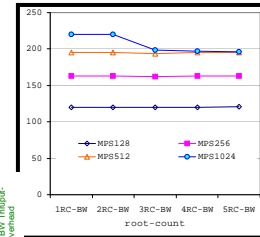
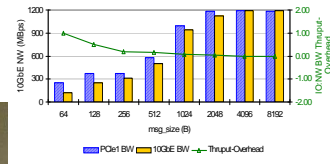
Challenges and opportunities (contd.)

- Evaluation methods
 - **Application-specific message-size profiling**
 - Sequence, distribution, etc.
 - Relationships to I/O- and network-level packet sizes
 - **Correlated workloads**
 - Synthetic workloads that can be correlated to actual workloads
 - **Network performance analysis tools**
 - Workload profile driven capabilities highly desired
 - **Robust instruments**
 - Single- or multi system domains traffic sources/sinks
 - Different-level time-synchronized traces
 - Lower level parameter adjustments (e.g., max_payload_size)
 - **TCO and industry ecosystem dynamics**
 - Capturing TCO (downstream costs, e.g., mgmt SW)
 - How economy of scale affecting costs

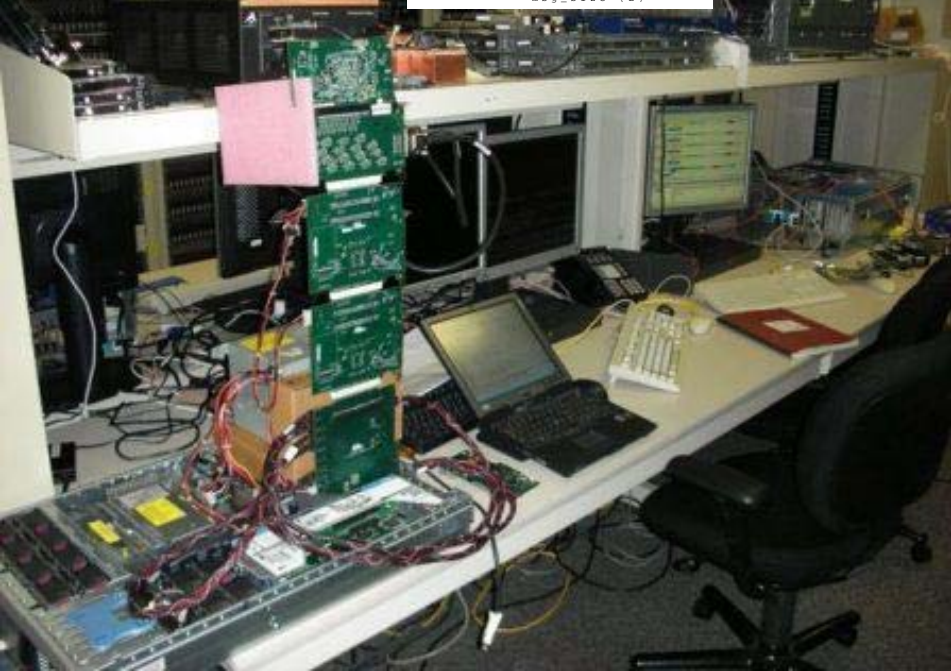
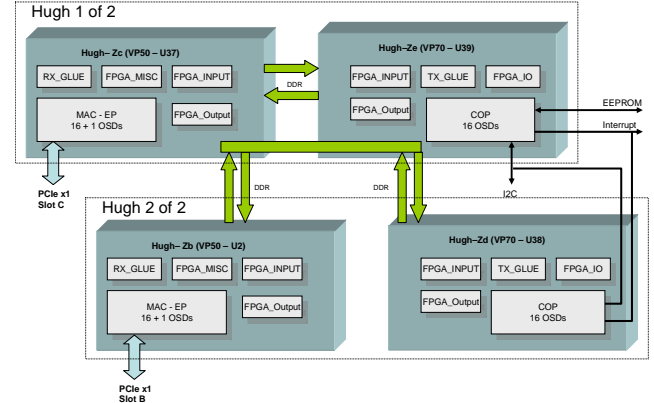
Closing remarks

- **Fabric convergence is imminent**
 - Potential benefits include lower solution costs, flexible resource allocations
 - Recent GPI, like HP BladeSystem c-Class, makes it easy to support & study
- **We discussed one case study**
 - first evaluation of I/O vs. NW consolidation
 - NW consolidation does not significantly affect system architecture
 - I/O consolidation does → flexible system partitions
 - I/O consolidation can be more cost/perf efficient, but more work needed
- **Lots of interesting open challenges still left**
 - System architecture implications, resource mgmt, evaluation methods, ...
- **I/O in computer arch is interesting and challenging ... and needs more research**

Q&A



HW emulation world



Acknowledgement:
NextIO's FPGA emulators & PCIe multi-root switch

