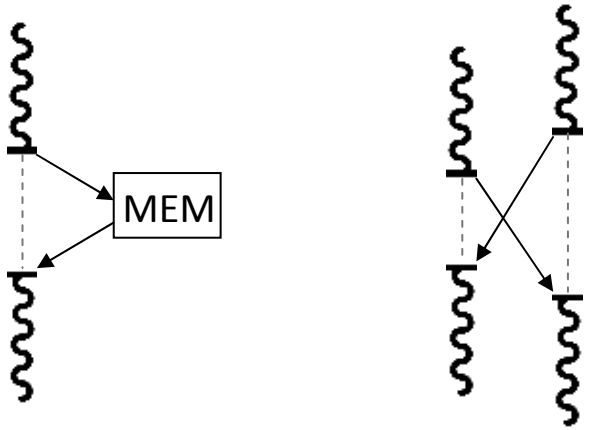
Multi-tier Interconnect and Mechanisms for Exascale Communication

Dr. Javier Navaridas

The University of Manchester

Implications of Communications on Performance

- Not having Data readily available severely limits performance
  - CPUs stall for Memory access
  - MPI tasks stall for IPC traffic
- This can have a huge impact (as seen in Top500 list)
  - **Linpack** (CPU intensive) *vs* **HPCG** (some Comm): 2 orders of magnitude lower processing throughput

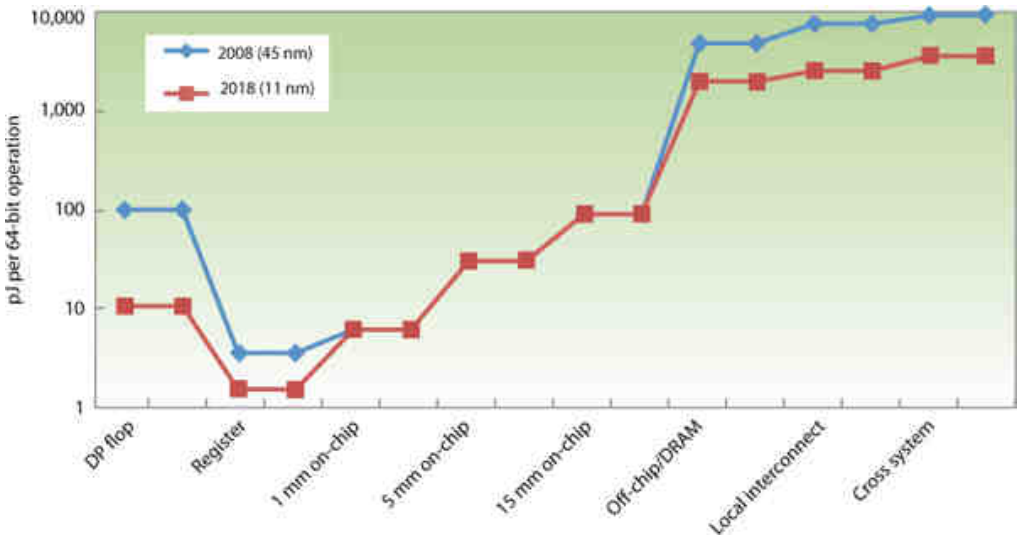


Computer	Country	HPCG (Pflop/s)	Rmax (Pflop/s)	HPCG Peak	HPCG HPL
K Computer SPARC64 VIIIfx 2.0GHz, Tofu Interconnect	Japan	0.6027	10.5	5.3%	5.7%
Tianhe-2 NUDT TH-IVB-FEP, Xeon 12C 2.2GHz, IntelXeon Phi	China	0.5801	33.9	1.1%	1.7%
Trinity Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries	USA	0.5461	14.1	1.2%	3.9%
Piz Daint Cray XC50, Xeon E5 12C 2.6GHz, Aries, NVIDIA Tesla P100	Switzerland	0.4864	19.6	1.9%	2.5%
Sunway TaihuLight NRCPC Sunway SW26010, 260C 1.45GHz	China	0.4808	93.0	0.4%	0.5%

- Many different forms of traffic
  - Inter-processor
  - Collectives
  - Storage
  - System-level & Control
- Traditionally traffic types are segregated over separate networks

Huge Impact of Data Transfers on Power/Energy

- Networks in large installations can consume a large proportion of the power budget
  - 10-50% according to [1]
  - 10-20% according to [2]
- The energy needed for moving data around is much higher than for performing computation
- Can Exascale's strict power budget really sustain several parallel networks?
  - We believe such design is inefficient



[1] Abts, D., et al: Energy proportional datacenter networks. In: Intl. Symposium on Computer Architecture. pp. 338{347. ISCA '10, ACM, New York, NY, USA

[2] Heller, B., et al.: Elastictree: Saving energy in data center networks. IN: NSDI'10 Proceedings of the 7th USENIX conference on Networked systems design and implementation

Existing Interconnection Technologies

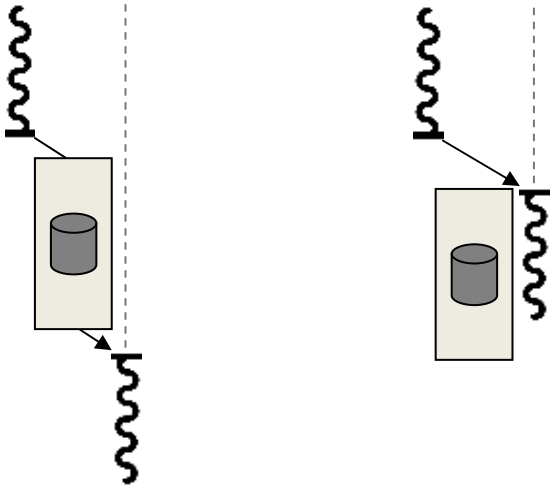
- **AXI**: Mandatory at the processor level (ARM subsystems - other architectures have their own)
  - Designed for high locality, not for scalability
    - Short messages, low latency, low number of concurrent transactions
    - Can we leverage for Exascale?

- **Ethernet**: De facto standard in most systems
  - Low performance: some HPC implementations exist, not enough for our purposes
  - Limited scalability, IP layer helps with scalability but severely degrades performance

- **Infiniband**: High Performance Interconnect BUT
  - Expensive and power-hungry (due to its excessive complexity)
  - No FPGA IP readily available
  - Can it really scale to Exascale? (millions of endpoints?)

ExaNeSt Interconnection Solution

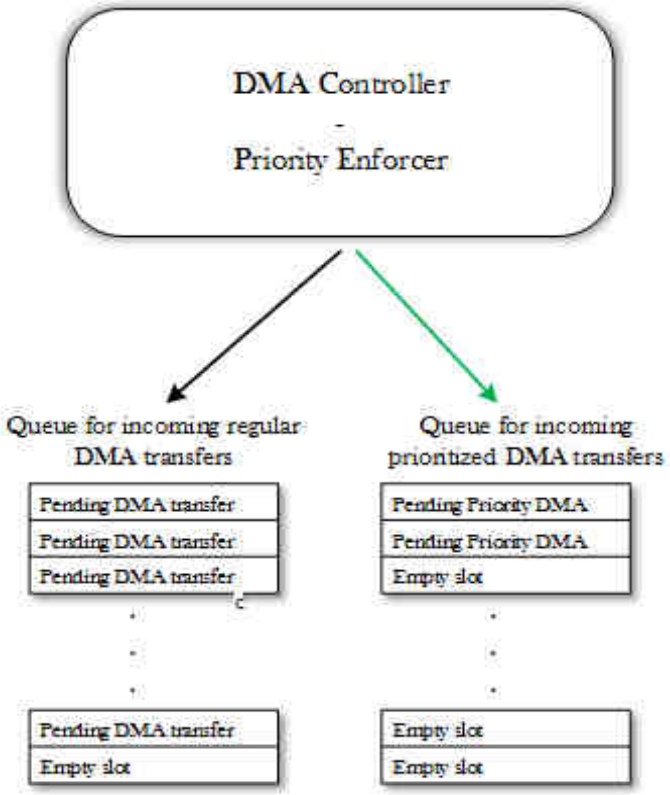
- Unified interconnect to reduce energy
  - Many new issues appear due to interferences between traffic types
    - Mechanisms for QoS
    - Congestion control
    - Locality Strategies



ExaNeSt Interconnection Solution (QoS)

- Quality of Service at the DMA level
  - Advanced scheduling that assigns higher priority to small transfers
  - Two transfer queues: high / low priority queue based on a threshold

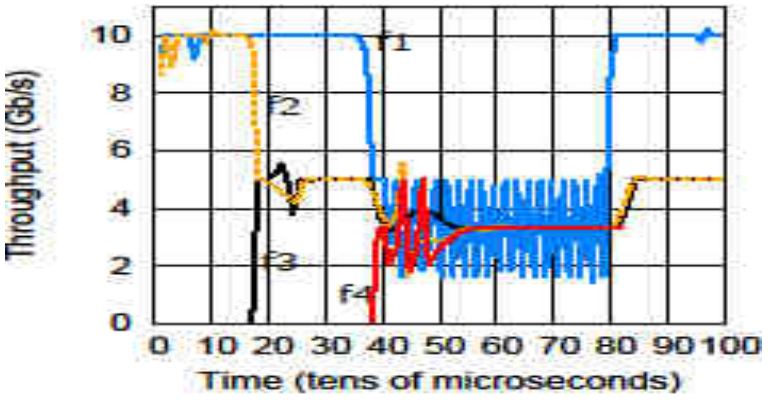
- Preliminary results show
  - Great improvement for critical small transfer transmissions
  - A small increase in the overall latency



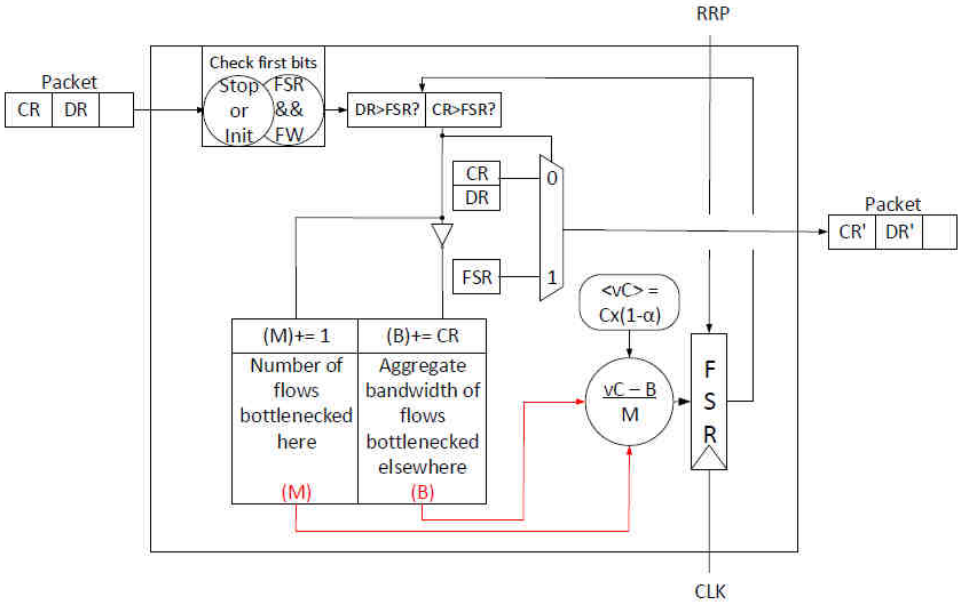
Transfer policy	Latency of the small transfer	Total latency
FIFO scheduler	40µs	42µs
Priority Scheduler	10µs	45µs

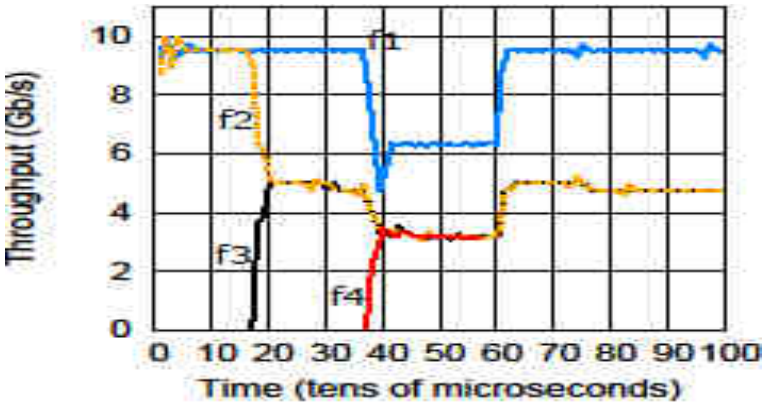
ExaNeSt Interconnection Solution (Congestion Control)

- A novel congestion control, DMMF
  - Contention points located at links
  - Reaction points placed at the sources (e.g. RDMA engines)
  - Multi-channel RDMA engine with per-channel rate throttling



Baseline

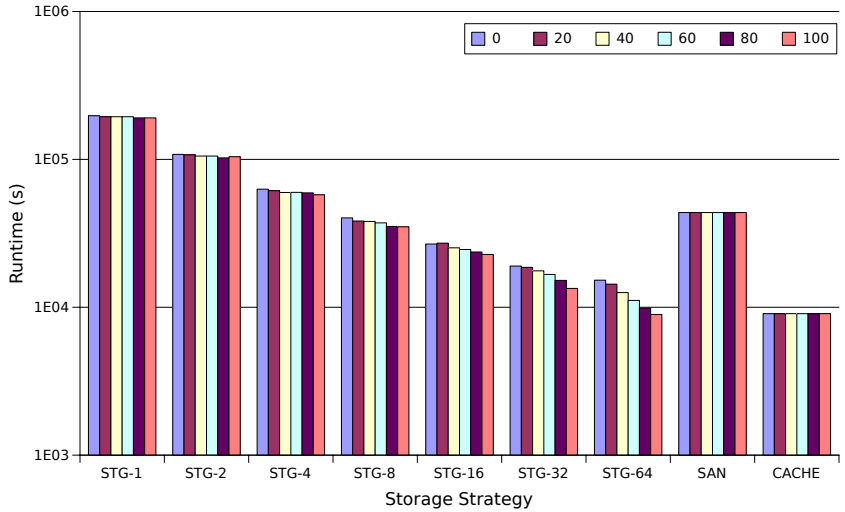




DMMF

ExaNeSt Interconnection Solution (Exploiting Locality)

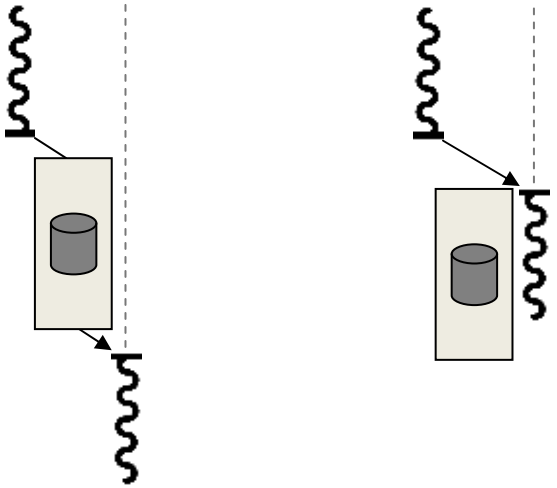
- We investigated several aspects of data-aware allocation [3]
  - Effects of spatial and temporal locality
  - Affinity of data to storage sources
  - Per-flow bandwidth allocation
- Many opportunities for the Scheduling system to exploit locality to improve performance
  - Temporal locality can reduce application runtime up to a 10%
  - Spatial locality can be more significant (one order of magnitude faster with perfect locality)
  - Traffic prioritization provides up to 17% reduction in runtime
  - Data-locality information can be essential for extreme-scale systems
- Distributed storage can outperform traditional SAN architectures



[3] JA Pascual, et al. “On the Effects of Allocation Strategies for Exascale Computing Systems with Distributed Storage and Unified Interconnects”. Invited Paper. CC-PE

ExaNeSt Interconnection Solution

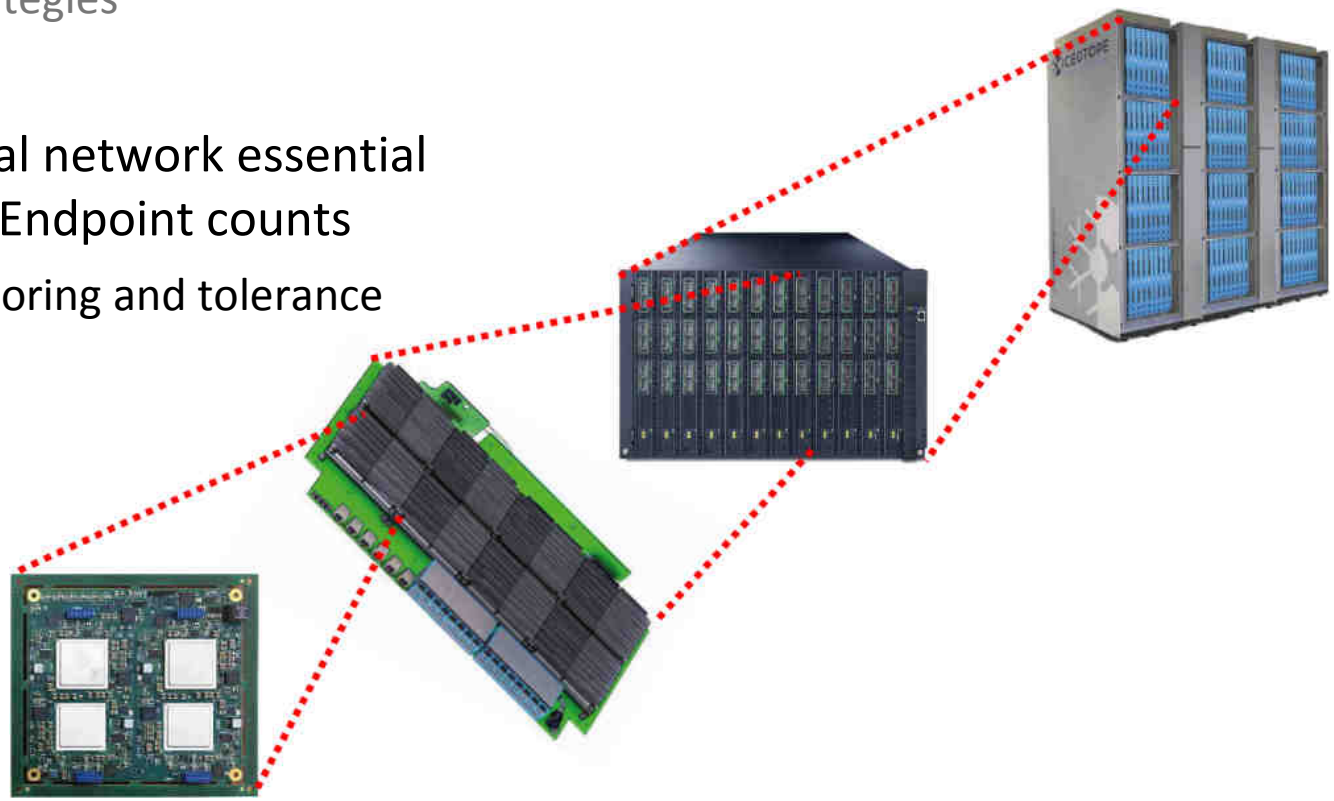
- Unified interconnect to reduce energy
  - Many new issues appear due to traffic interferences
    - Mechanisms for QoS
    - Congestion control
    - Locality Strategies



ExaNeSt Interconnection Solution

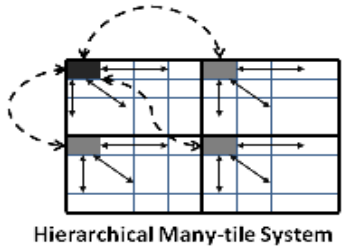
- Unified interconnect to reduce energy
  - Many new issues appear due to traffic interferences
    - Mechanisms for QoS
    - Congestion control
    - Locality Strategies

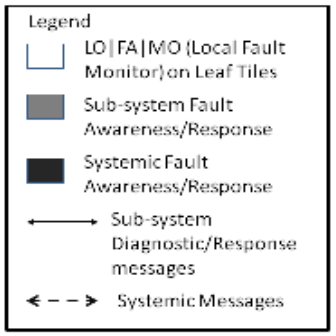
- Multitier hierarchical network essential to support massive Endpoint counts
  - Makes fault monitoring and tolerance more manageable

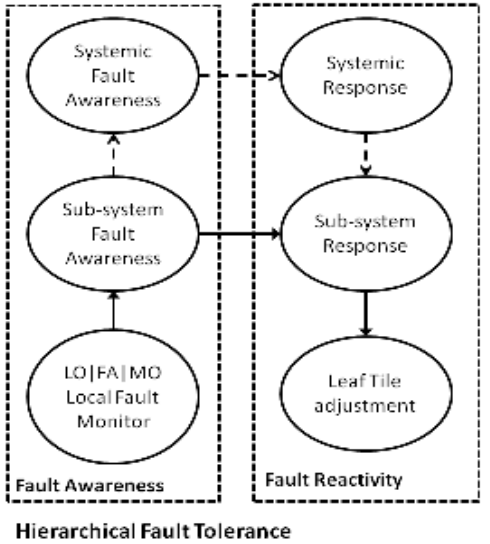


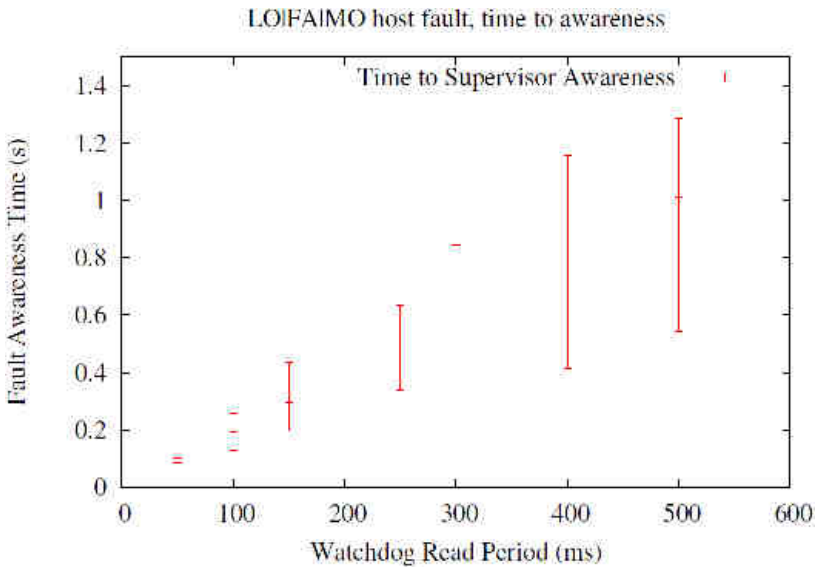
Scalable Fault-tolerance (LO|FA|MO)

- Fault-tolerance is another of the challenges of Exascale
  - Millions of Endpoints
    - Very low MTBF (hours?)
    - Lots of control traffic
- LO|FA|MO: a distributed, hierarchical mechanisms to enable systematic awareness for extreme-scale systems [4]
- Relies on hierarchical information
  - LO|FA|MO component runs in every node to detect faults and other critical events
  - Information is propagated upward the system hierarchy
  - Reactions can be autonomously initiated at every level based on that information







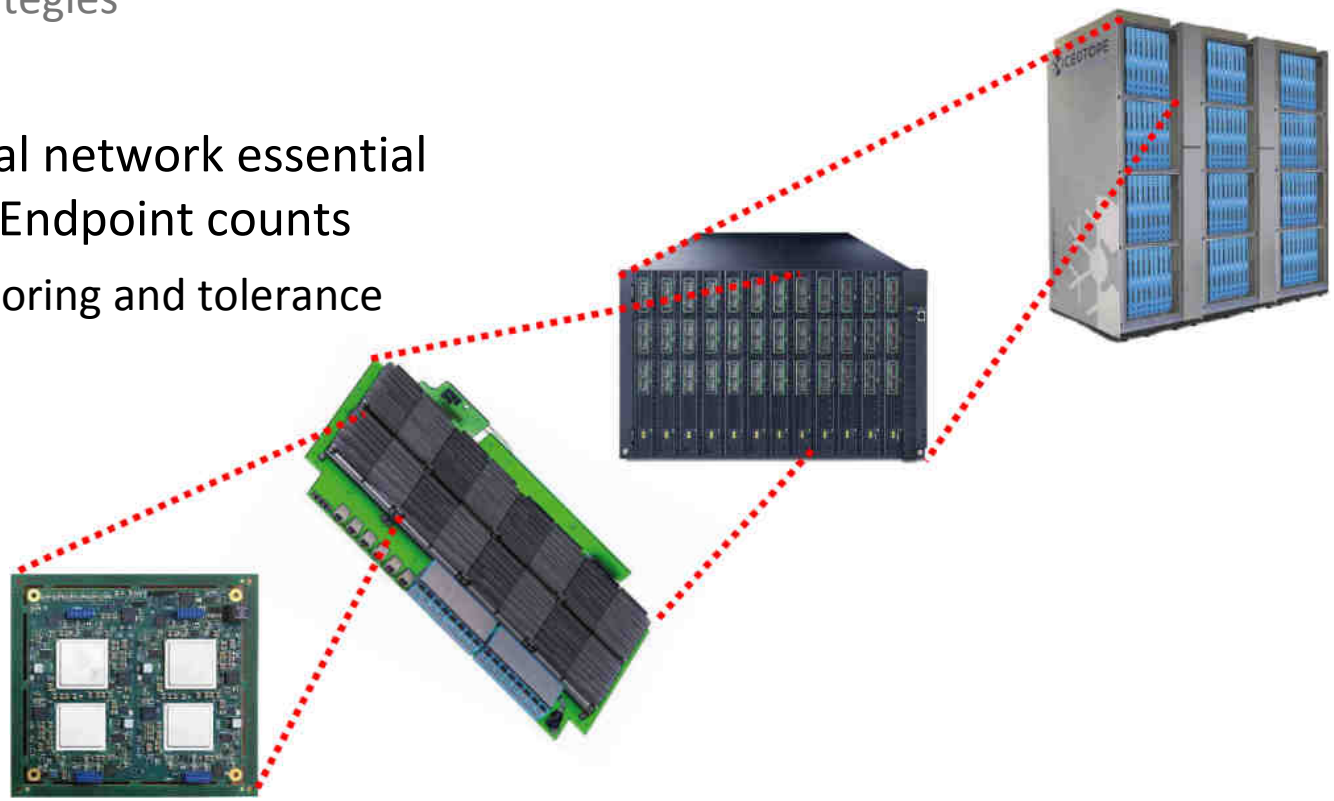


[4] R Ammendola, et al. "A hierarchical watchdog mechanism for systemic fault awareness on distributed systems." Future Generation Computer Systems, 53:90–99, 2015.

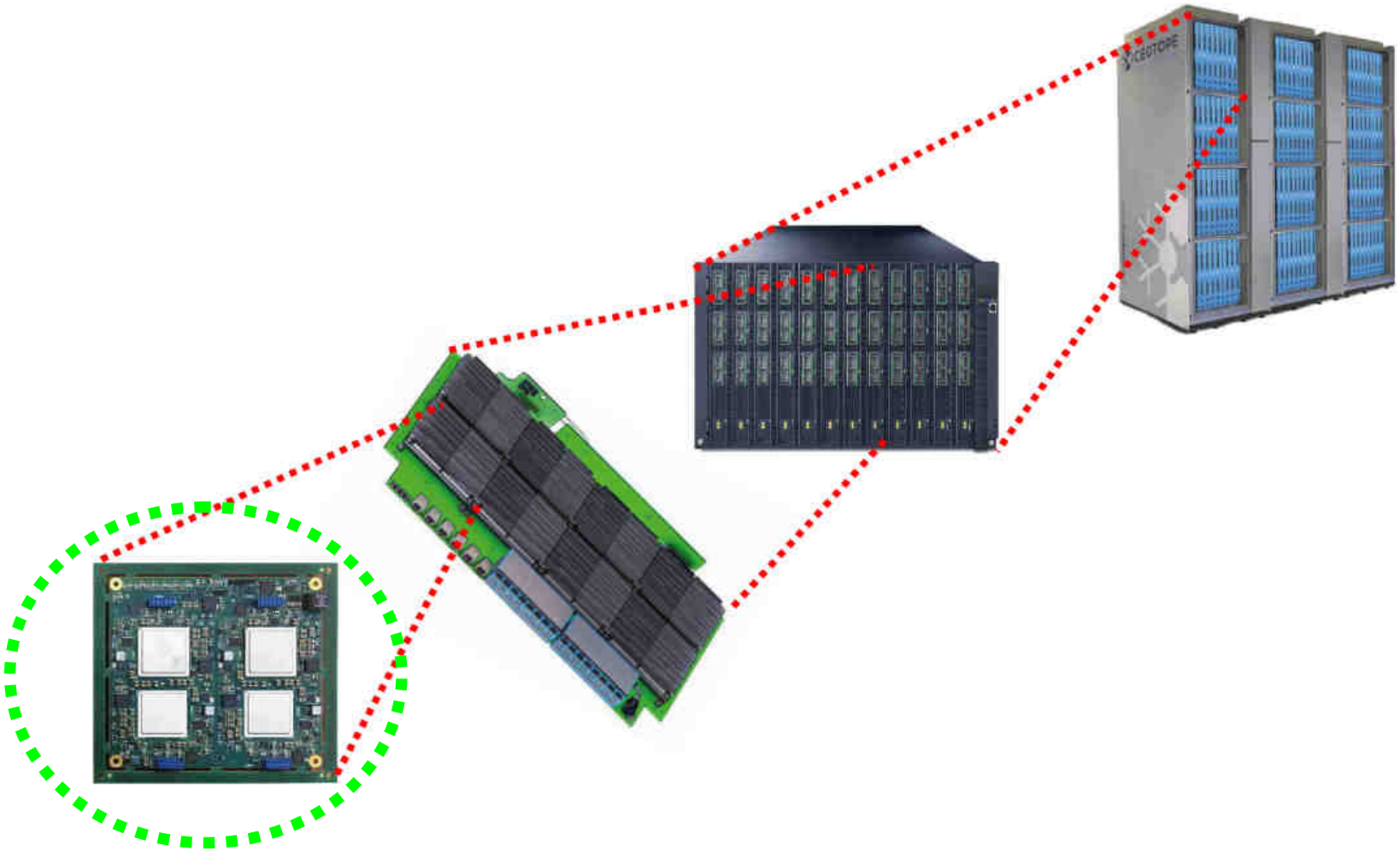
ExaNeSt Interconnection Solution

- Unified interconnect to reduce energy
  - Many new issues appear due to traffic interferences
    - Mechanisms for QoS
    - Congestion control
    - Locality Strategies

- Multitier hierarchical network essential to support massive Endpoint counts
  - Makes fault monitoring and tolerance more manageable
  - Simplifies routing
  - Tier 0: AXI
  - Tier 1-2: ExaNet
  - Tier 3+: ToR

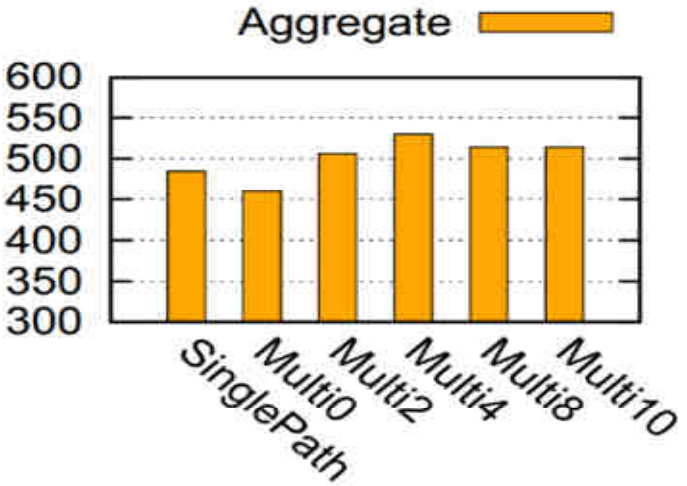


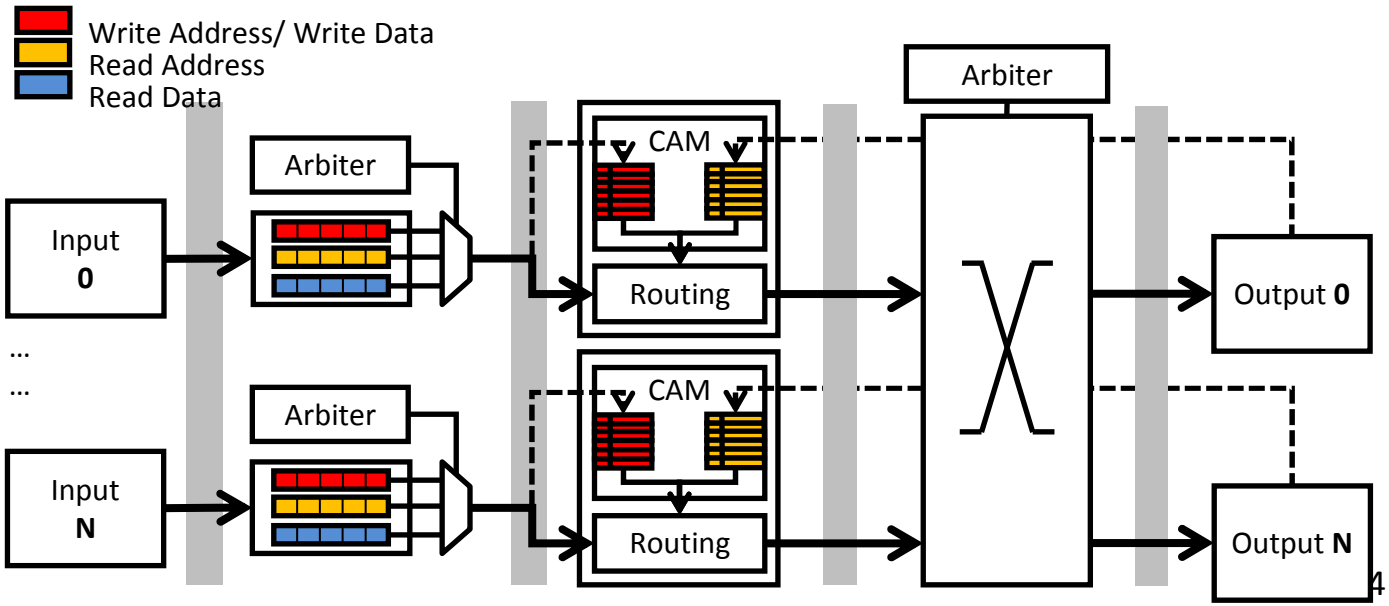
Tier 0: QFDB-level AXI-Crossbar



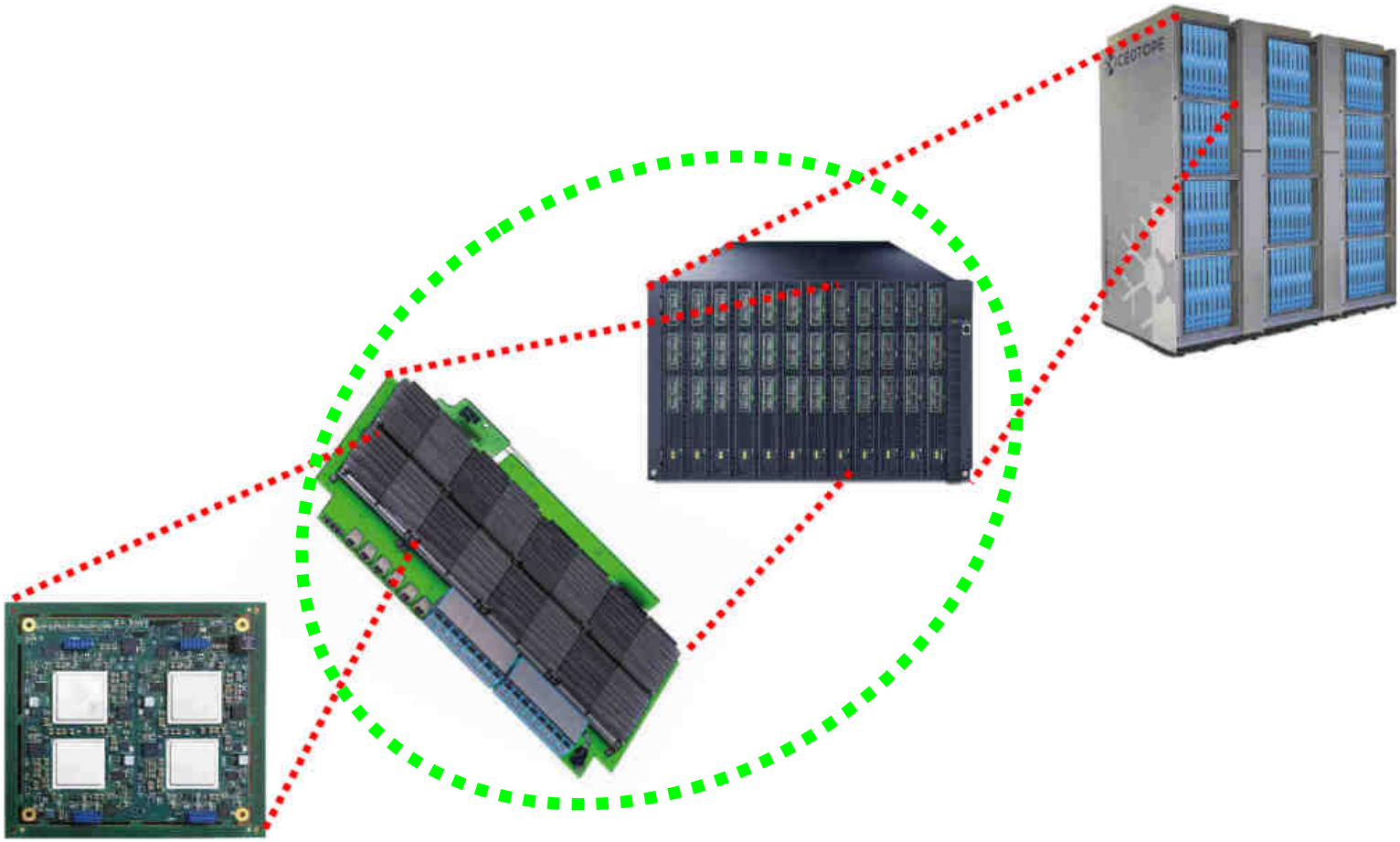
Tier 0: QFDB-level AXI-Crossbar

- Extend AXI protocol for intra-QFDB Routing
  - No need for protocol translation
- Implemented a Multipath routing scheme
  - Reduces in-node congestion when high congestion in a link occurs
  - Configurable multipath threshold values
    - 4 flits threshold seems the sweet spot
- Plans for leveraging it for fault-tolerance



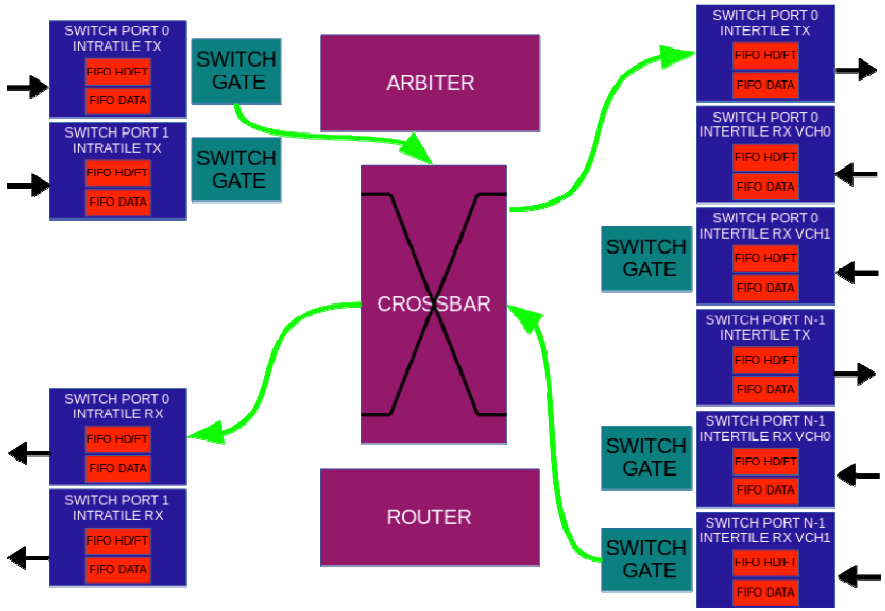


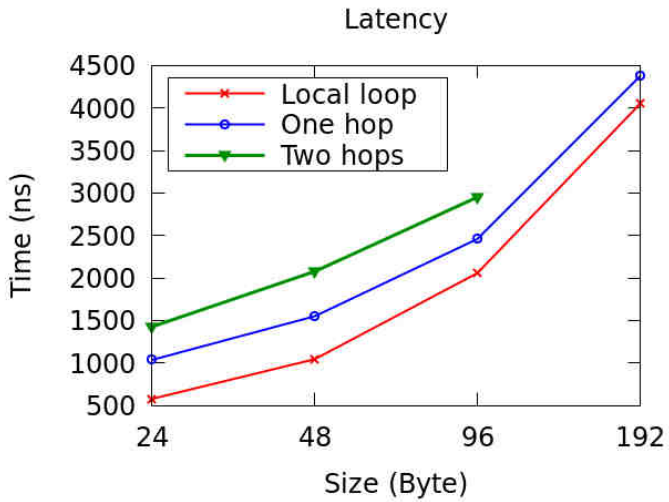
Tier 1-2: Blade/Chassis level



Tier 1-2: Blade/Chassis level

- Fully parametric (width, VCs, credit) [5]
  - Torus-like topologies (Dragonfly?)
  - Virtual cut-through
  - 2 VCs to avoid deadlocks
- Data Link Controller (APElink)
  - low latency, AXI compliant, valid/ready interface with Aurora IP
  - low latency credit management: 8 bit per VC, programmable threshold values
- byte enable management developed
  - Routing&Arbiter infrastructure allows to implement an enhanced DOR, VC select based priority
- Interfaces with the lower level by means of NI + RDMA engine

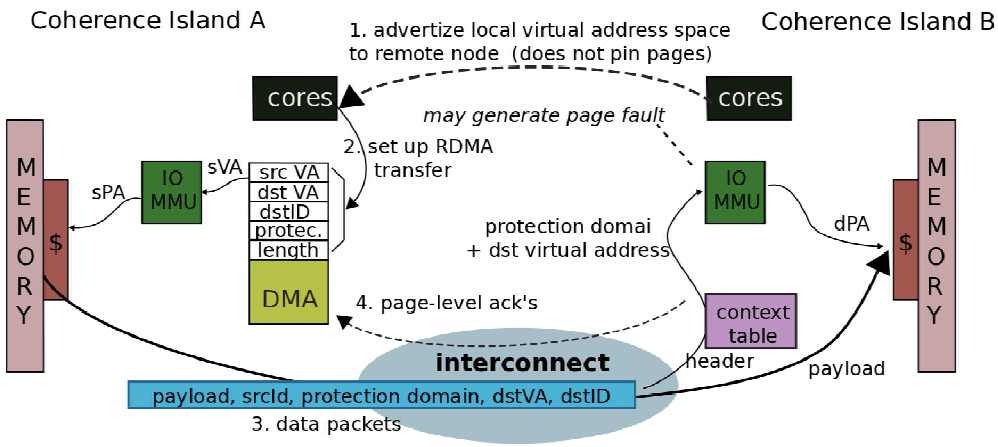




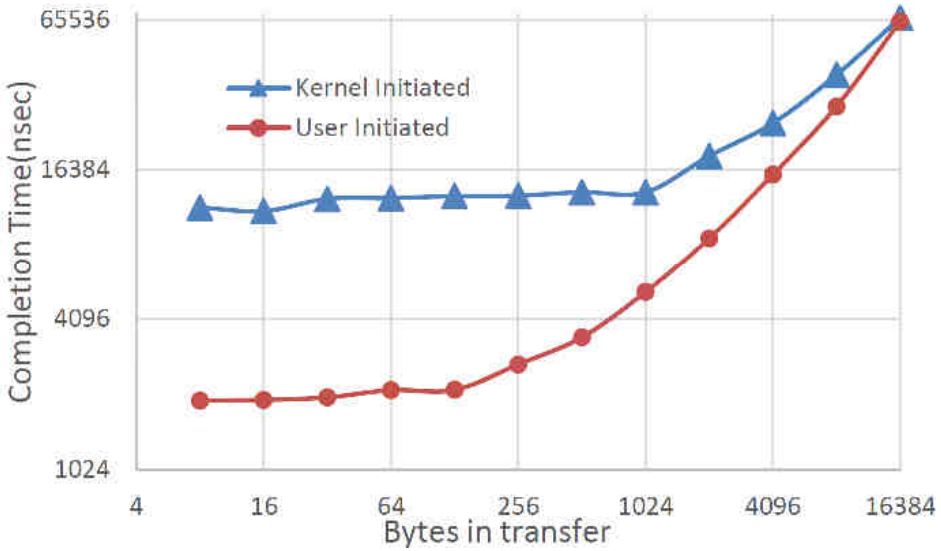
[5] R. Ammendola, et al. “Low latency network and distributed storage for next generation HPC systems: the ExaNeSt project” J. Phys.: Conf. Ser. 898 082045, 2017

Network Interface + RDMA Engine

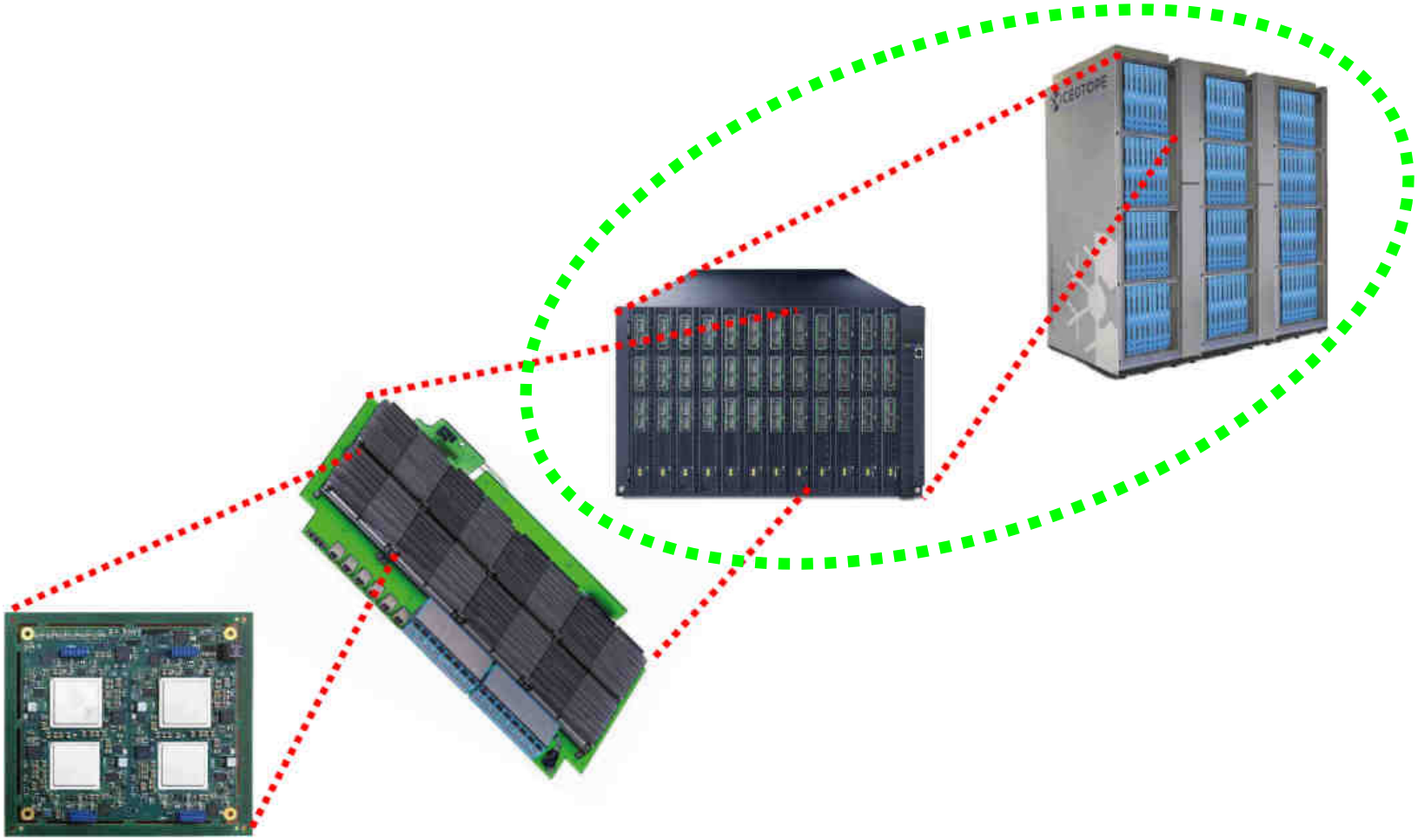
- Virtualized ExaNet MBOX and Packetizer
- SMMU middleware configuration
- ExaNet ←→ AXI adapters
- Virtualized 10G Ethernet NIC
- Advanced DMA Engine
  - Full 64 bit addresses + 16 bit PDID
  - 1024 source channels
  - CmpltNotification @ destw. 256 (~fully associative) contexts
  - resiliency: ACKs/ re-xmitper 16 KB block, time-outs
  - Multipath at block level
  - Packets payload aligned to dest. address: arbitrary addresses
  - Software can configure paths & define transfer dependencies



← Is this the same mechanisms as before DMMF?



Tier 3-: Chassis/Cabinet level



Top-of-Rack Switch Architecture

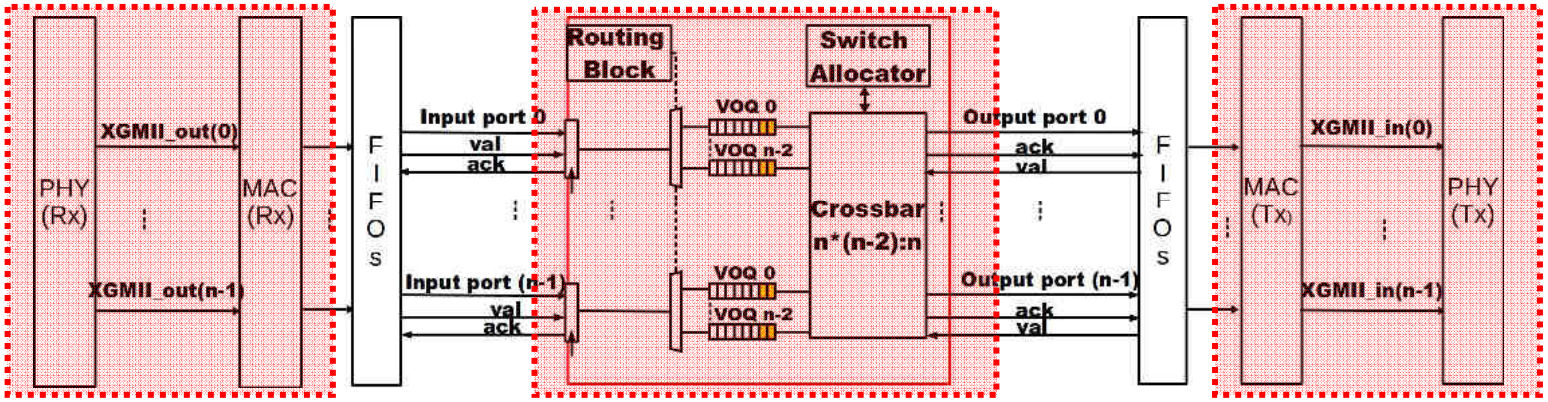
- 3-stage VCT/Wormhole architecture [6]
  - Routing | Allocation | Traversal
- Arithmetic and multipath routing
  - No need for power-hungry CAMs
  - Great flexibility for topologies
- Virtual Output Queues (VOQ)
  - Reduces contention for resources
- Currently interfaced through a simple packetizer [7]
  - Seamless sharing of memory and FPGA resources (Evaluated with some test applications)
- Looking into interfacing with lower layers

STREAM benchmark

Implementation	Function	Best Rate MB/s
PS DRAM	Copy:	3344.6
	Scale:	1825.9
	Add:	2033.2
	Triad:	1683.2
Local BRAM	Copy:	45.4
	Scale:	44.0
	Add:	44.2
	Triad:	44.6
Remote BRAM	Copy:	4.8
	Scale:	4.7
	Add:	4.7
	Triad:	4.7

Climate modelling kernel

Platform	Exec. Time (s)
CPU Local	7.52
CPU Remote	68.232
FPGA Local	2.533
FPGA Remote	24.186

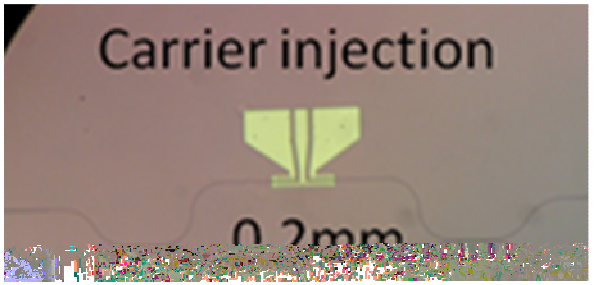


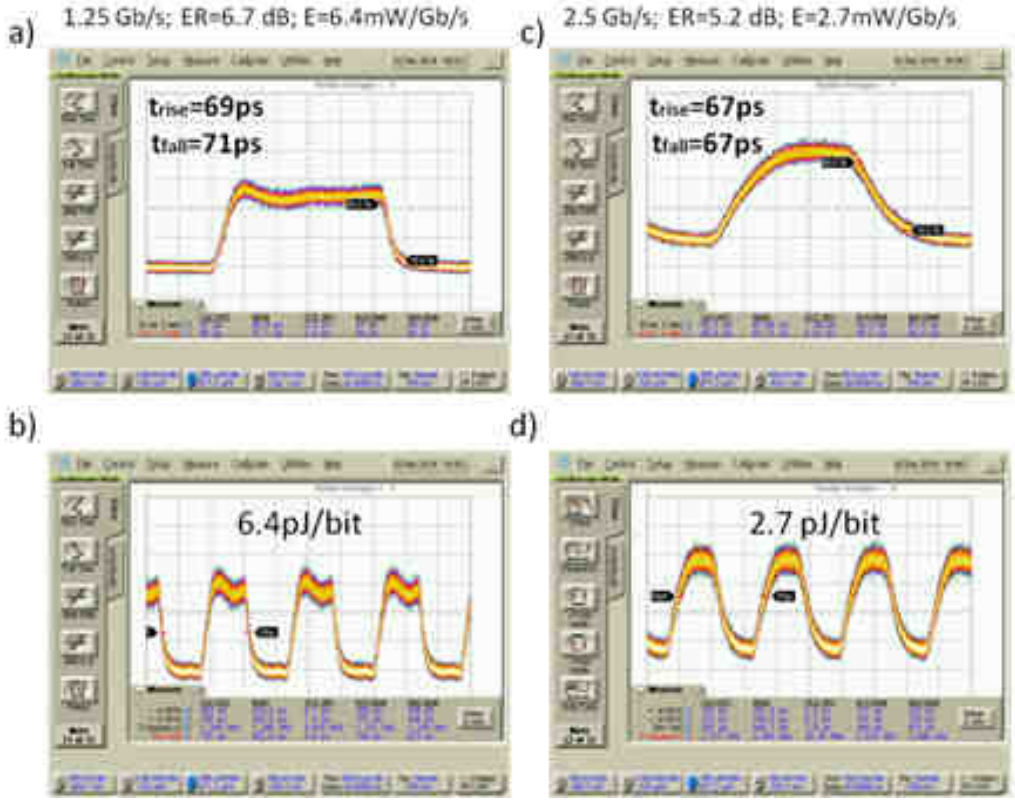
[6] C Concatto, et al. "A CAM-free Exascalable HPC Router for Low-energy communications". ARCS'18

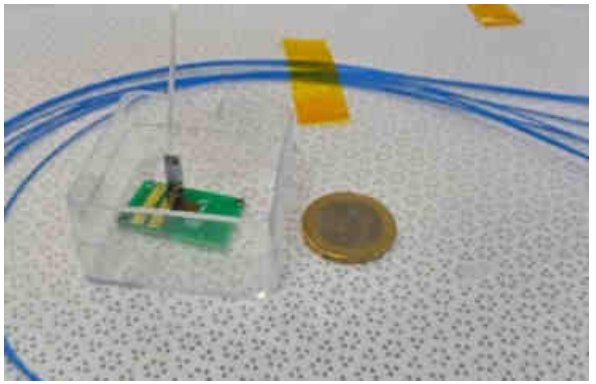
[7] J Lant, et al. "Shared Memory Communication in Networks of MPSoCs". Under review

Optical Switch Demonstrator

- Developed and fabricated a small 2x2 full optical switch prototype
- Can be composed into matrices of switches for larger NxN crossbars

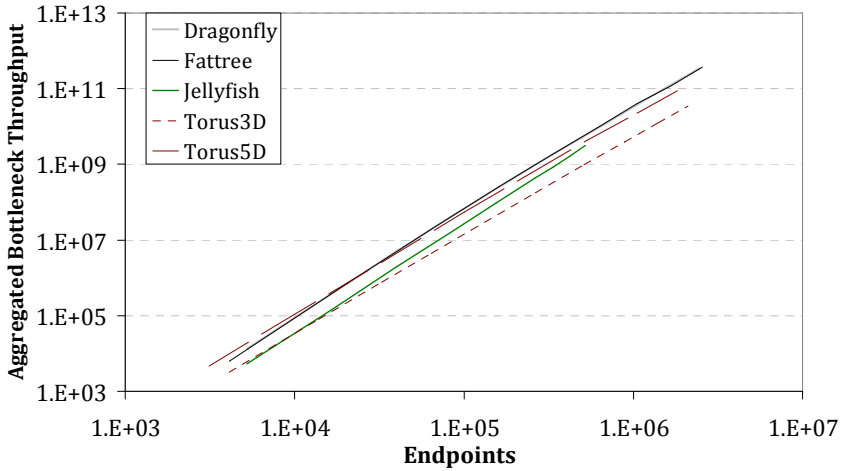


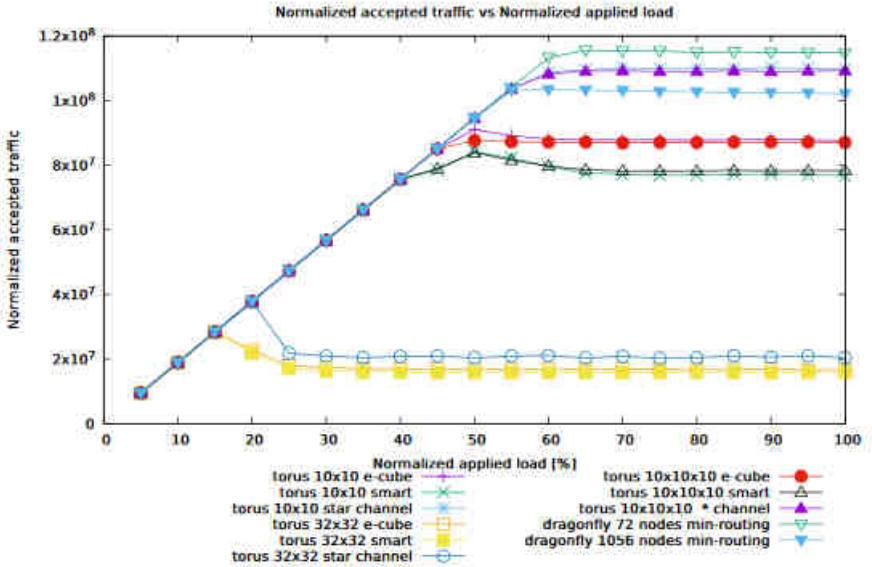


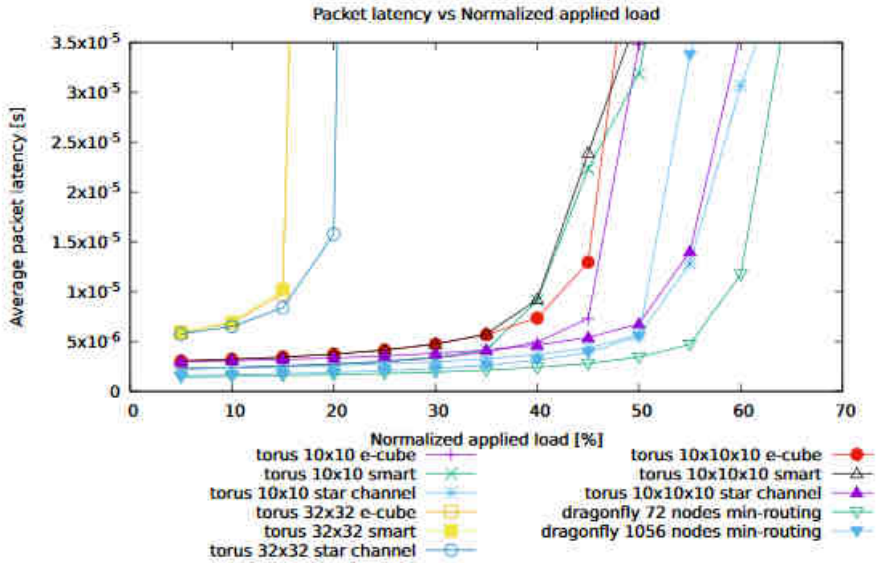


Considerations on Topologies

- Studied state-of-the-art HPC topologies
  - Fattree, dragonfly, tori
  - Graph-based topologies (Jellyfish, de Bruijn, Kautz)

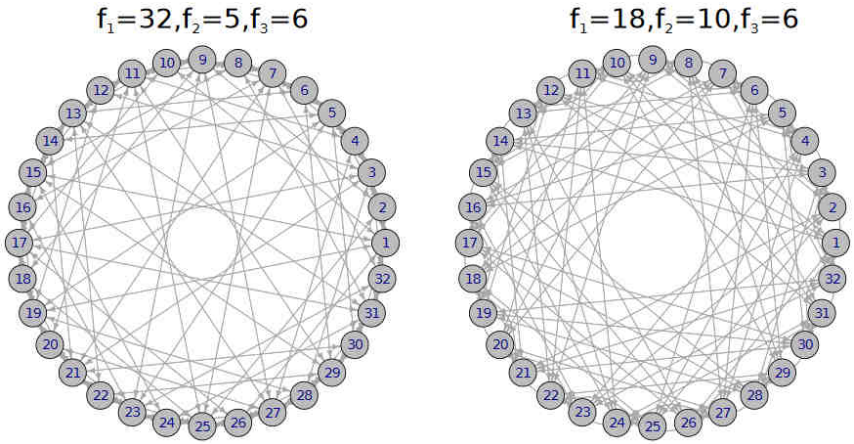


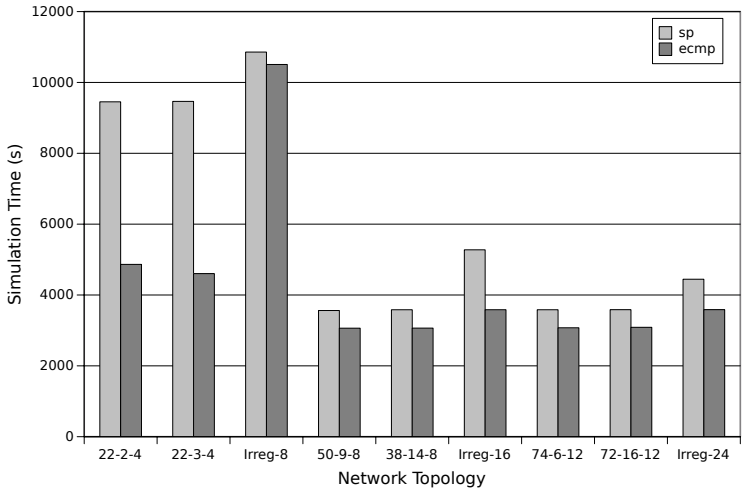




Considerations on Topologies

- Studied state-of-the-art HPC topologies
  - Fattree, dragonfly, tori
  - Graph-based topologies (Jellyfish, de Bruijn, Kautz)
- But also proposed a multi-objective optimization framework [8]
  - Objectives: Performance, Resilience, Cost
  - Metrics: Bisection width, Path diversity, Number of links
  - Algorithms: NSGA-II, SMS-EMOA
- Most of the above are Deadlock-prone
  - Dragonfly and Torus had their own deadlock-avoidance mechanisms; others do not
  - We propose a collection of novel deadlock mechanisms for arbitrary topologies and routing [9]





[8] JA Pascual, et al. “Designing an exascale interconnect using multi-objective optimization”. CEC 2017: 2209-2216

[9] JA Pascual, et al. “High-Performance, Low-Complexity Deadlock Avoidance for Arbitrary Topologies/Routings”. To be Submitted. ICS’18

Summary of ExaNeSt Interconnection

- **Fully-functional interconnect** supporting inter-node transfers
  - NI + ExaNet switches + high-speed links
  - user-level, zero-copy transfers through virtualized RDMA engines
  - microsecond application-side latency
- A prototype FPGA-based architecture for **ToR switches**
  - Supports resource sharing across devices through a (very) basic packetizer
- A new **AXI crossbar** architecture for intra-DB communications
- A small-scale physical demonstrator of an **optical Switch**
- Many mechanisms to handle the challenges of Exascale systems
  - QoS, congestion control, locality strategies, fault tolerance and monitoring
- Research around **topologies**
  - Studied most state-of-the art topologies
  - Proposed a multi-objective topology optimization framework
  - Proposed new deadlock avoidance mechanisms

Related papers / presentations

[3] JA Pascual, et al. “On the Effects of Allocation Strategies for Exascale Computing Systems with Distributed Storage and Unified Interconnects”. Invited Paper. CC-PE

[4] M. Katevenis, et al.“The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems”, DSD’16

[5] M. Katevenis and N. Chrysos, “Challenges and Opportunities in Exascale-Computing Interconnects”, Keynote, AISTEC’16.

[6] P Xirouchakis ea, “Low latency RDMA Engine for High-Performance Computing”, under preparation.

[7] D. Giannopoulos ea, “Fair Allocation for RDMA Transfers”, under preperation

[4] R Ammendola, et al. "A hierarchical watchdog mechanism for systemic fault awareness on distributed systems." Future Generation Computer Systems, 53:90–99, 2015

[5] R. Ammendola, et al. “Low latency network and distributed storage for next generation HPC systems: the ExaNeSt project” J. Phys.: Conf. Ser. 898 082045, 2017

[6] C Concatto, et al. “A CAM-free Exascalable HPC Router for Low-energy communications”. ARCS’18

[7] J Lant. et al. “Shared Memory Communication in Networks of MPSoCs”. Under