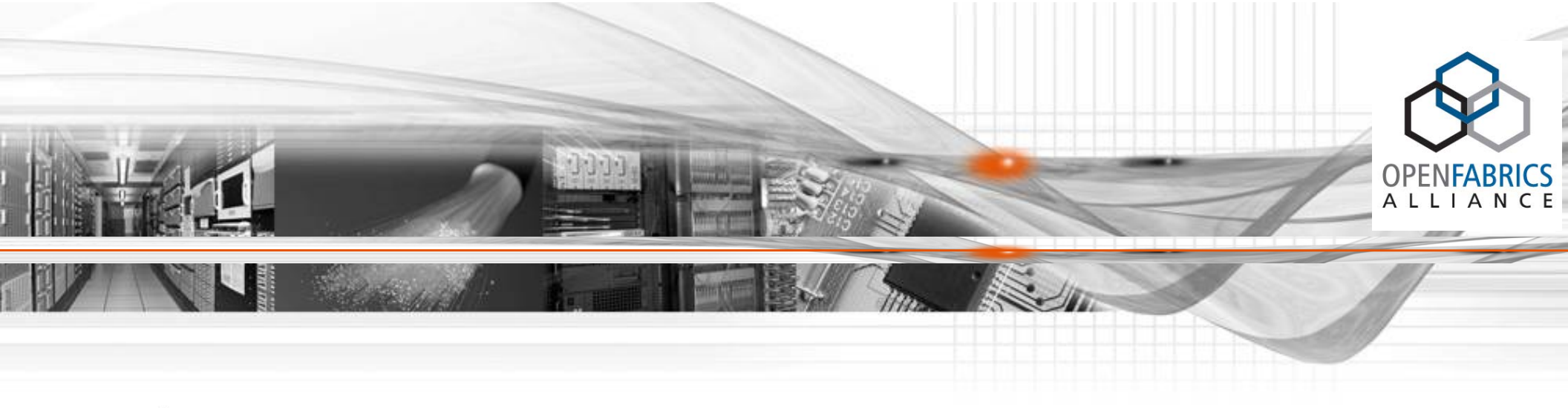
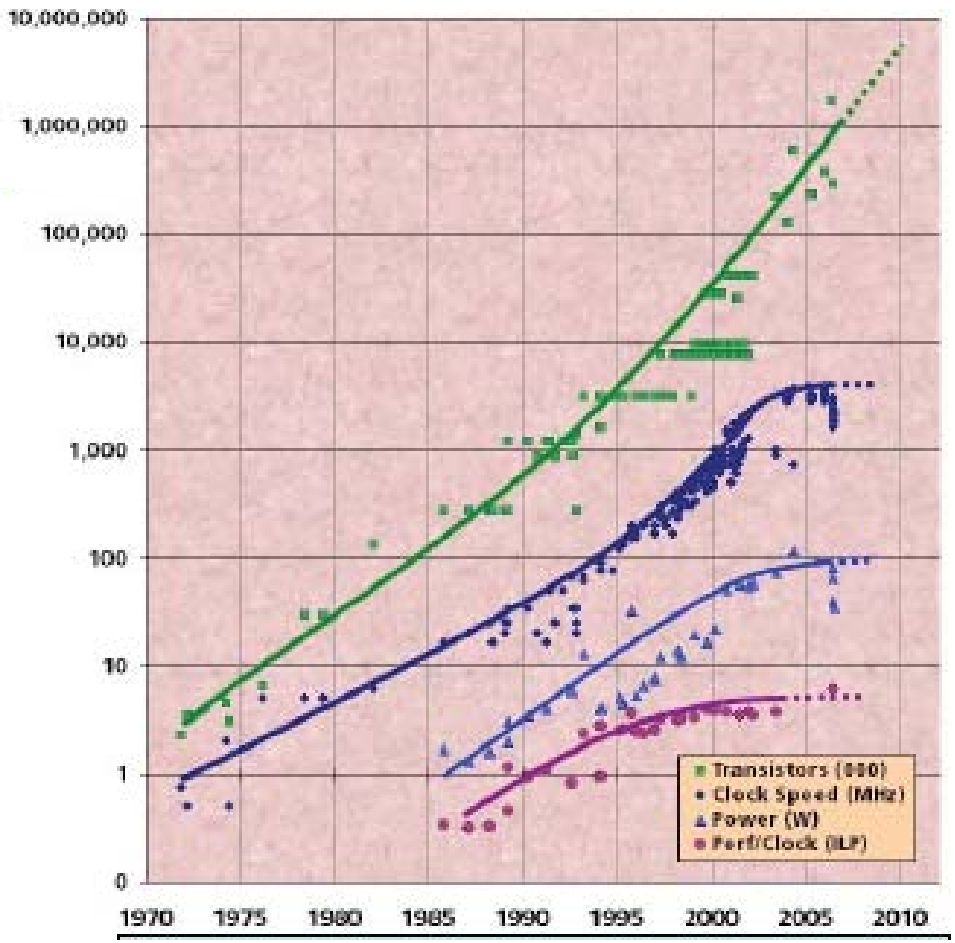


OFED RDMA on Cost-scalable Networks

Authors: Blake Fitch (IBM Watson Research),
Bernard Metzler (IBM Zurich Research)

Date: 04/04/2011

Processor Clock Frequency Scaling Ends



- Three decades of exponential clock rate (and electrical power!) growth has ended
- Yet Moore's Law continues in transistor count
- What do we do with all those transistors to keep performance increasing to meet demand?
- Industry response: Multi-core (i.e. double the number of cores every 18 months instead of the clock frequency (and power!)
- But, added transistors can be used for other functions such as memory/storage controllers, embedded networks, etc.

Source: "The Landscape of Computer Architecture," John Shalf, NERSC/LBNL, presented at ISC07, Dresden, June 25, 2007

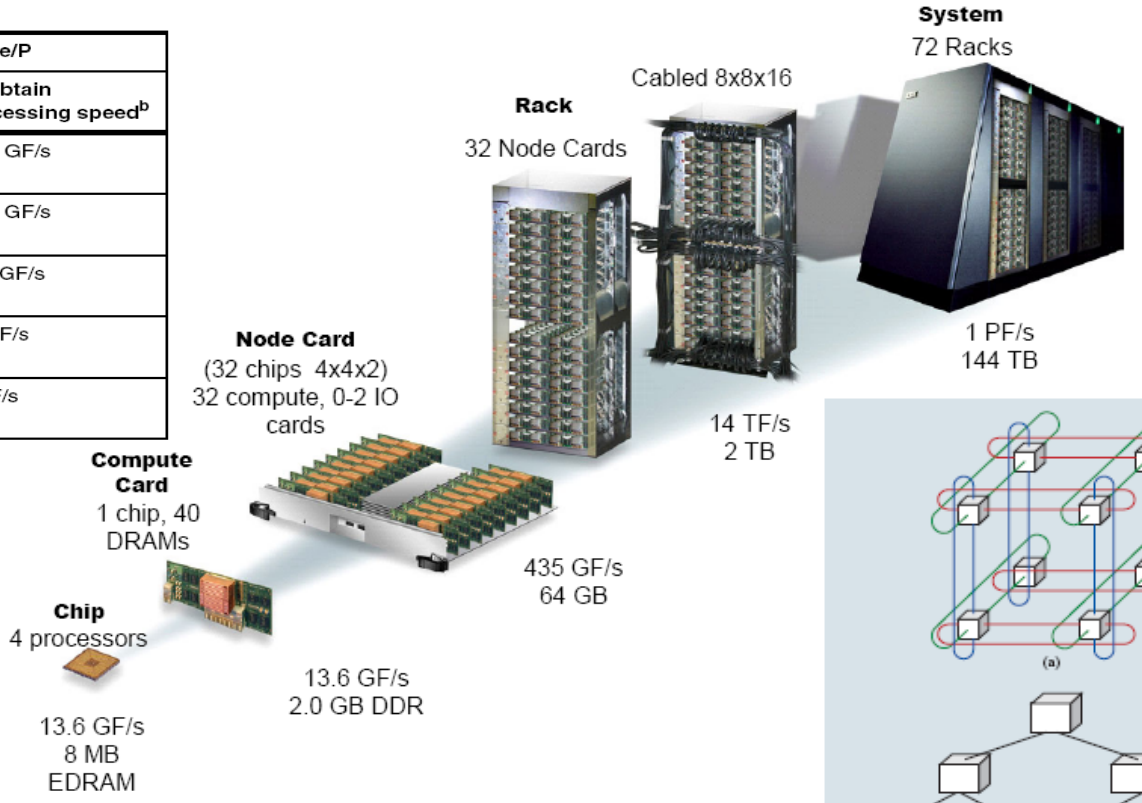
The Blue Gene Family BG/L, BG/P, next BG/Q: Scalable, System-on-a-Chip, Supercomputers.

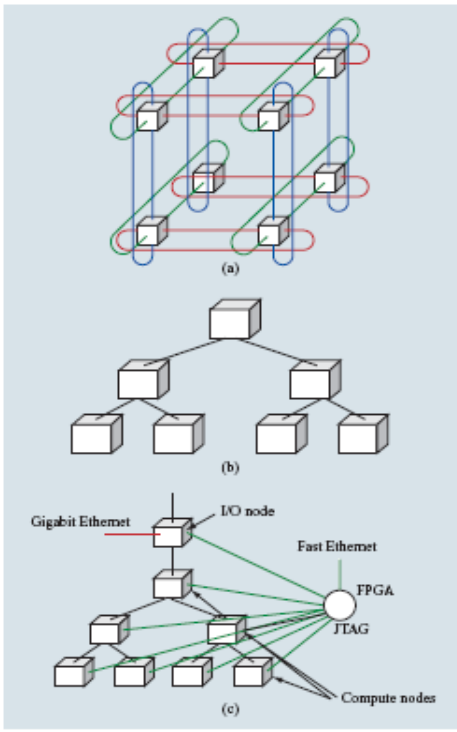
OPENFABRICS ALLIANCE

Table 1-1 Comparison of Blue Gene/L and Blue Gene/P packaging

	Blue Gene/L		Blue Gene/P	
	Quantity per component	To obtain processing speed[a]	Quantity per component	To obtain processing speed[b]
Chip	2 processors	2.8 GF/s 5.6 GF/s	4 processors	13.6 GF/s
Compute card	2 chips	5.6 GF/s 11.2 GF/s	1 chip	13.6 GF/s
Node card	32 chips; 16 per midplane	90 GF/s 180 GF/s	32 chips; 16 per midplane	435 GF/s
Rack	32 node cards	2.8 TF/s 5.6 TF/s	32 node cards	14 TF/s
System	64 racks	180 TF/s 360 TF/s	72 racks	1 PF/s

Feature	Blue Gene/L	Blue Gene/P
Network topologies		
Torus network		
Bandwidth	2.1 GB/s	5.1 GB/s
Hardware latency (nearest neighbor)	200 ns (32B packet) and 1.6 μs (256B packet)	100 ns (32B packet) and 800 ns (256B packet)
Global collective network		
Bandwidth	700 MB/s	1.7 GB/s
Hardware Latency (round trip worst case)	5.0 μs	3.0 μs
Full system (72 rack comparison)		
Peak performance	410 TFlop/s	~1 PFlop/s
Power	1.7 MW	~2.3 MW



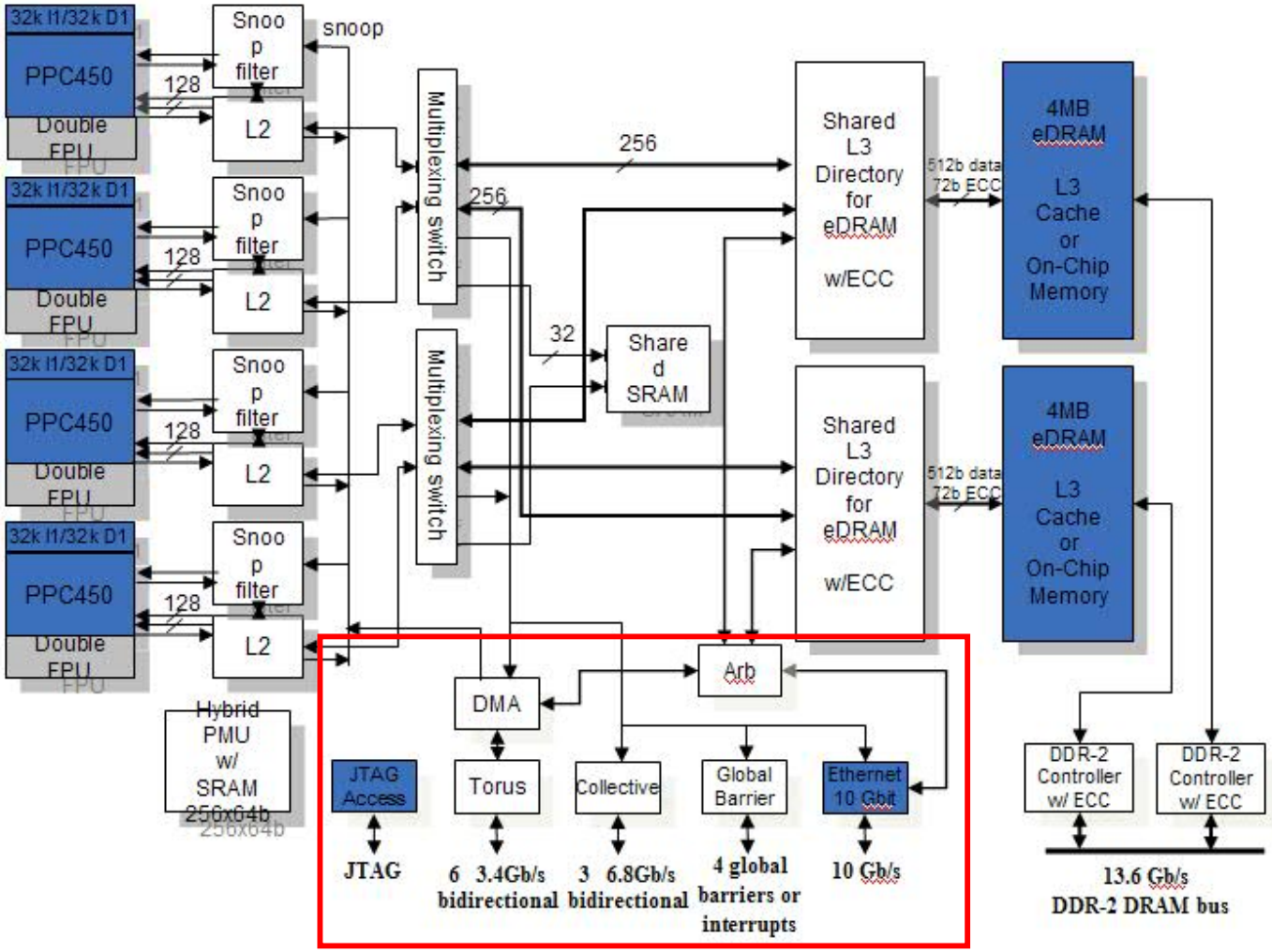


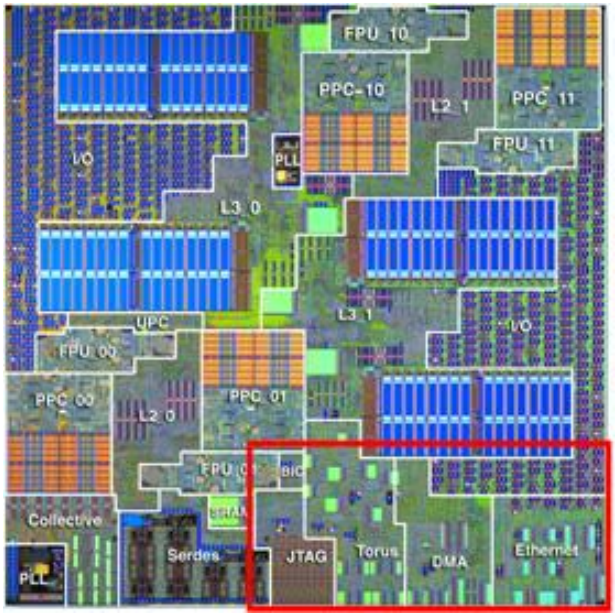
SoC architecture with focus on cost-scalable network, reliability, and power efficiency.

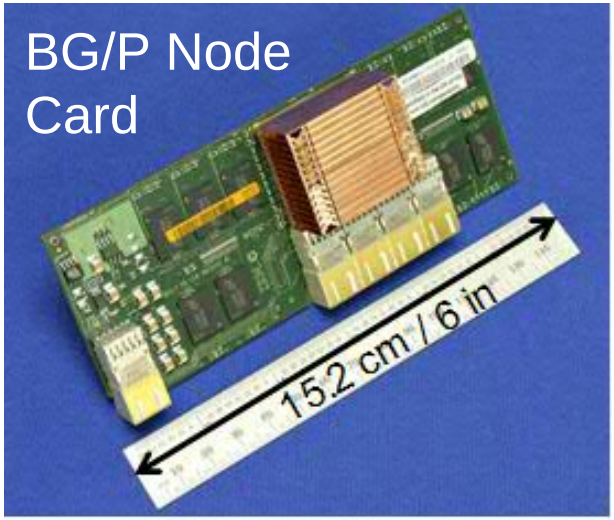
Minimize unutilized configurability cost.

BG/P System-on-a-Chip Compute Node

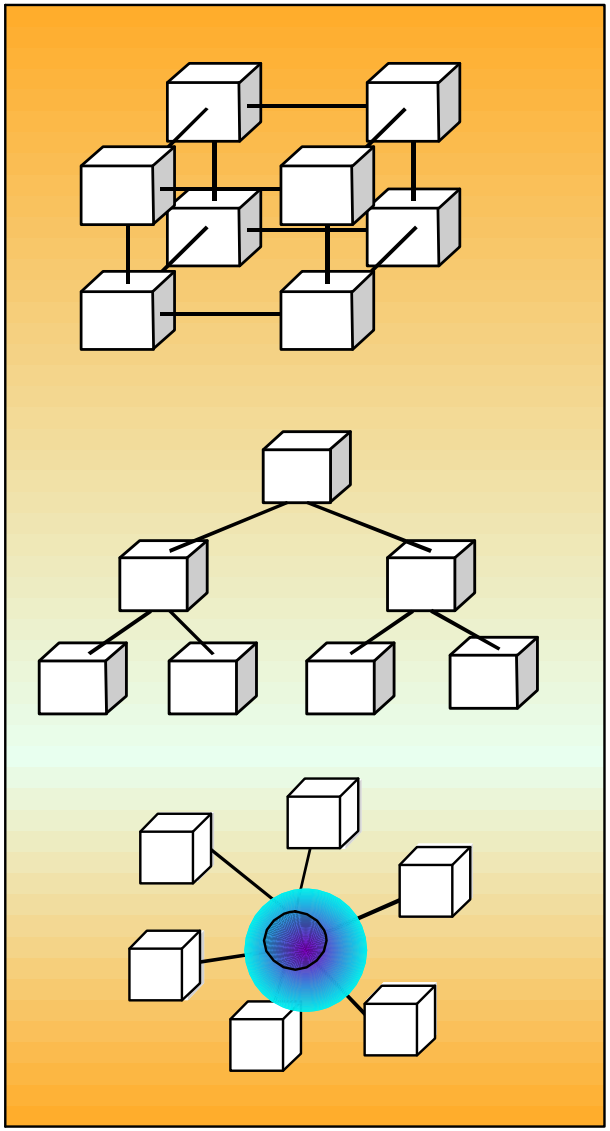
Network logic is a fraction of compute ASIC complexity/area.







Blue Gene/P Networks



3 Dimensional Torus

- Interconnects compute nodes
- Virtual cut-through hardware routing
- 425MB/s per link
- 5100 MB/s per node (6 links * 2 directions * 425MB/s)
- 0.5 µs latency to nearest neighbors, 5 µs to the farthest
- MPI: 3 µs latency for one hop, 10 µs to the farthest
- Communications backbone for computations

Collective Network

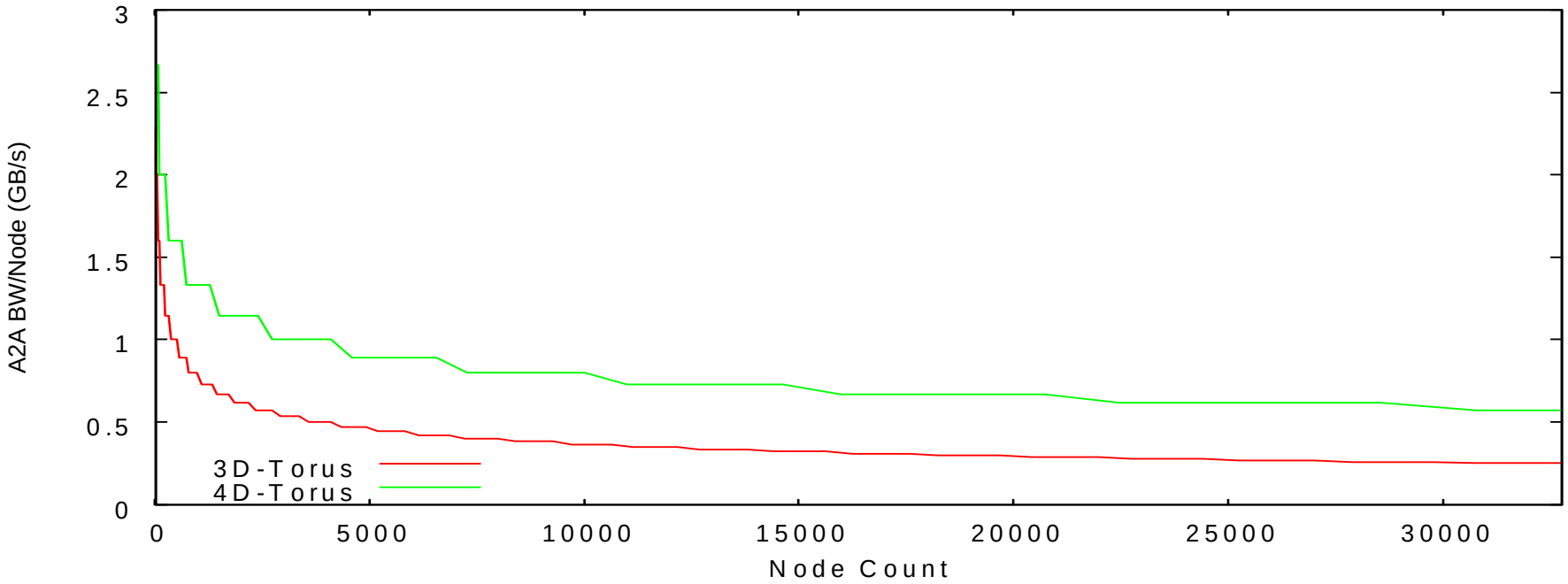
- Used to connect I/O nodes
- One-to-all broadcast functionality
- Reduction operations functionality
- 850MB/s of bandwidth per link
- Latency of one way tree traversal 1.3 µs, MPI 5 µs

Low Latency Global Barrier and Interrupt

- Latency of one way to reach all 72K nodes 0.65 µs, MPI 1.6 µs

SoC Torus enables Cost-effective Network

- Cost scales linearly with number of nodes
- Torus bisectional bandwidth does fall rapidly for small system sizes
- But, bisectional bandwidth continues to rise as system grows
- A 4D Torus with 1GB/s links yields peak all-to-all bandwidth of:
  - 1GB/s at 4k nodes (8x8x8x8)
  - Above 0.5GB/s out to 64k nodes (16x16x16x16)



Theoretical 3D and 4D Torus All-to-all throughput per node for 1GB/s Links

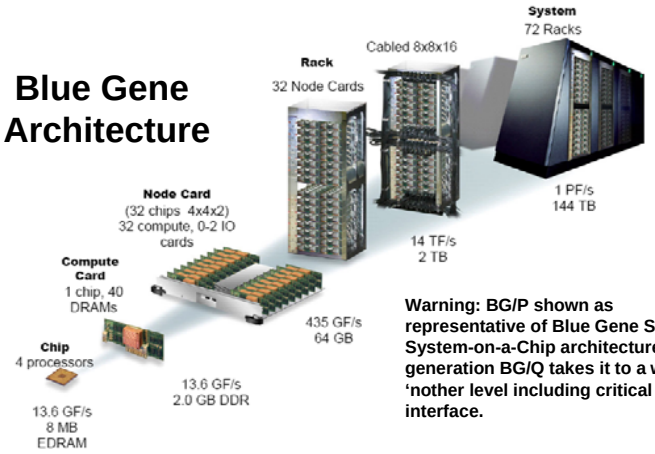
Exascale HPC and Big Data Analytics Share System Design Challenges

- Disk options increasingly infeasible
  - Exascale HPC I/O requirements: 60TB/s I/O Bandwidth
  - Big Data analytics requirements: high random I/O operations, high bandwidth
- Storage Class Memory (Flash, PCM, etc) offers a way forward
  - A single handful of SCM chips provides order 1GB/s I/O bandwidth
  - HPC still would require 60,000 port 10GbE equivalent full bisection network!
- Infrastructure costs drive integration of storage into scalable compute infrastructure
- **We see an emerging architectural class:**

  Scalable, System-on-a-chip, Storage Class Memory
  - **Reduces unutilized configurability (adaptors, connectors, ports)**
  - **Scalable RAS at each stage of the design process**
  - **Power efficiency enables very high density**
  - **Enables a cost-scalable network**

Though Experiment: Blue Gene Active Storage

Blue Gene Architecture



Warning: BG/P shown as representative of Blue Gene Scalable, System-on-a-Chip architecture. Third generation BG/Q takes it to a whole 'nother level including critical PCIe interface.

PCIe Flash Board Repackaged
(for example: Fusion-io)

ioDrive Duo SLC	Today
Flash Capacity	320 GB
I/O Bandwidth	1.5 GB/s
IOPS	207,000

Integrated scalable computation and storage

BGAS Rack

Nodes	512
Storage Cap	640 TB*
I/O Bandwidth	768 GB/s
Random IOPS	100 Million

Target Applications

- High performance shared file or object store
- Graph-based algorithms
- Join
- Sort
  - "order by" queries
- "group by" queries
- Map-Reduce (heavy reduce phase)
- Aggregation operations
  - count(), sum(), min(), max(), avg(), …
- Data analysis/OLAP
  - Aggregation with "group by"…
  - Real-time analytics

Key architectural balance point: All-to-all throughput roughly Equivalent to SCM bandwidth

* Assume a two fold added Flash capacity before scale up.

Blue Gene Active Storage Use Models

- Data Intensive Supercomputing (HPC)
  - Integrate with standard BG/Q system as I/O accelerators
  - Create/modify HPC applications to make direct use of new capabilities ex: neurosimulation
- Standard Middleware
  - BGAS utilized as a standard cluster with very high performance
  - Configure standard middleware such as GPFS, DB2, etc to run in BGAS environment
- New Frameworks
  - Restructured HPC applications and workflows to use new middleware to intercommunicate (non-posix FS, Hadoop etc)
  - Acceleration
    - Active File System to offload UNIX commands into BGAS
    - DB2 offload via Infosphere Federated Wrapper to offload and accelerate relational operators

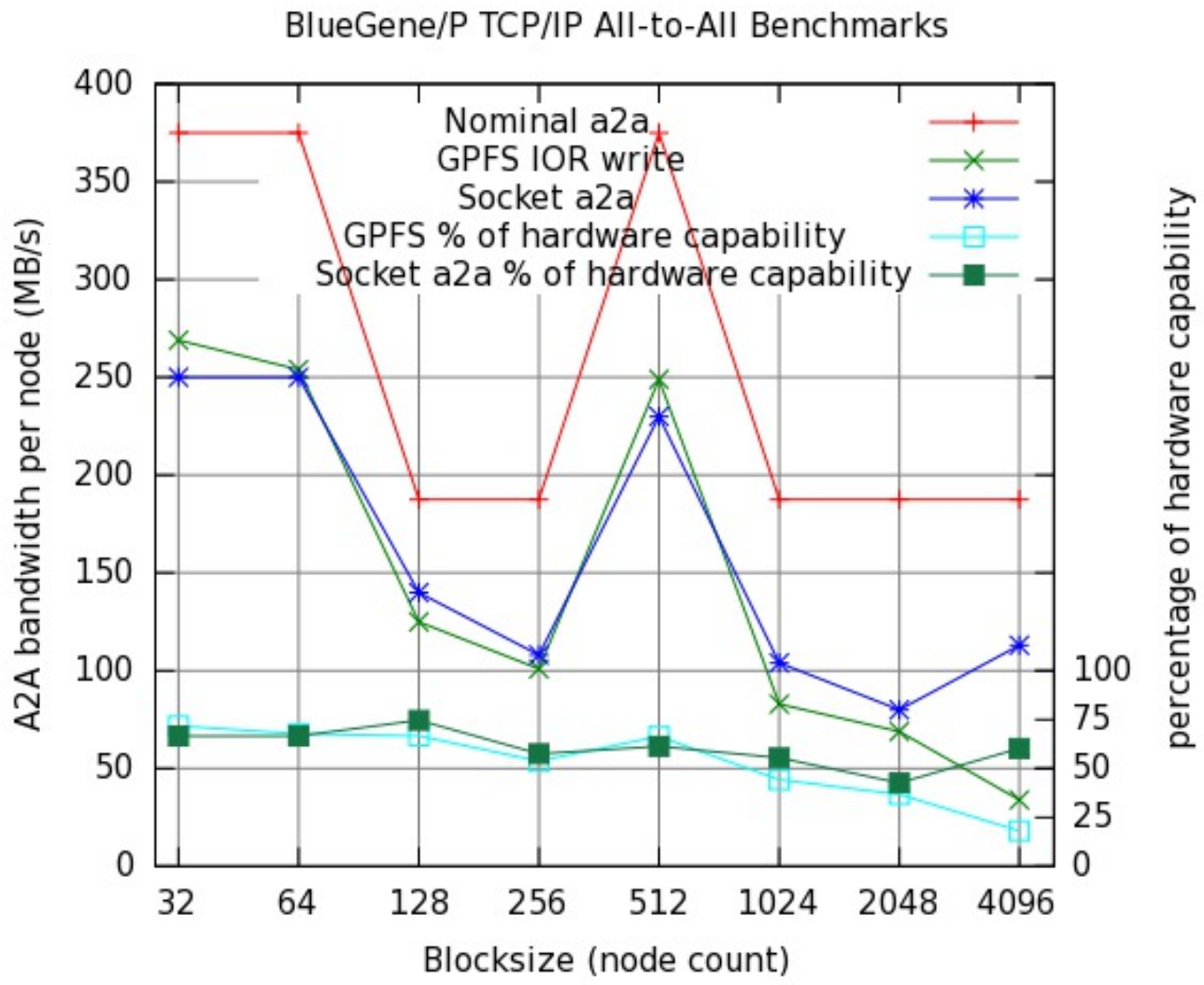
Blue Gene Active Storage as a Scalable Linux Cluster

- 2.6.29 kernel, 32-bit powerpc
- 4GB RAM, diskless, 4-way SMP, 1024 nodes/rack
- IP-over-torus to other compute nodes
- I/O nodes as IP routers for off-fabric traffic
- MPICH/TCP, OpenMPI/TCP, SLURM, TCP sockets
- 2 OFED verbs providers for Blue Gene Torus
  - SoftiWARP over TCP/IP
  - vrnic with direct torus access
- Experimental GPFS version for on-board Parallel File System
- Research platform for scalable data-centric computing and BG/Q prototyping

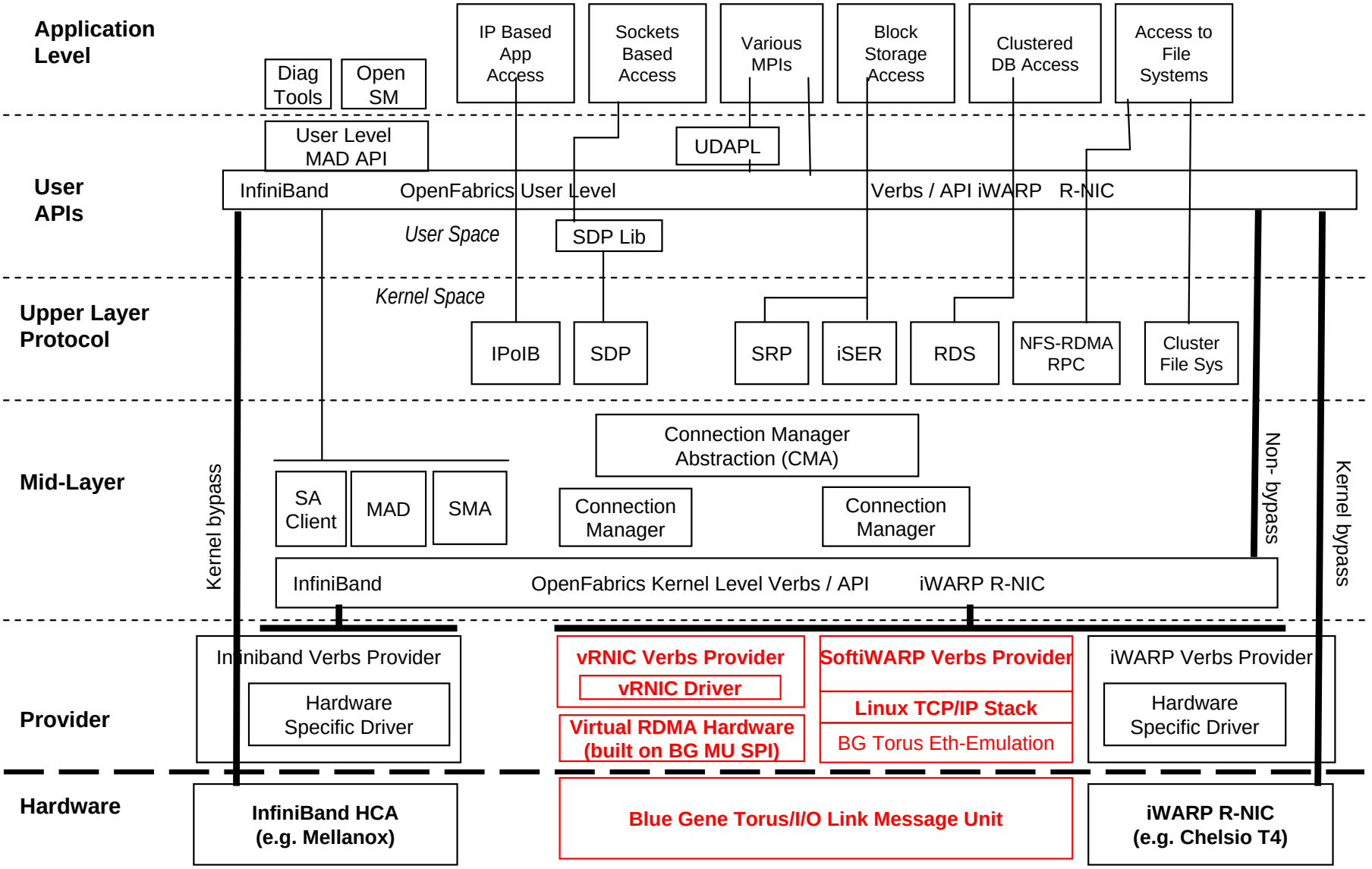
Blue Gene Scalable Linux Custer: Ethernet and TCP/IP

- **Emulate Ethernet over Blue Gene Torus and Collective Network**
- **Motivation**
  - High performance TCP/IP among compute nodes as well as outside of Blue Gene
  - I/O Nodes used as IP routers
- **Protocol**
  - Each skbuff handled independently
  - 'short message' protocol
    - Torus packets sent to reception FIFO, reassembled by software.
  - 'long message' protocol
    - 'Proposal' packet sent to reception FIFO, 'Acceptance' packet sent back to transmitter's reception FIFO, MU DMA 'put' delivers bulk data to skbuff and completes the interaction
- **Performance**
  - Torus bisection yields all-to-all throughput of 375MByte/s per node at 512 nodes
  - TCP sockets all-to-all test gets 60% of this
  - IBM GPFS gets 65% of this – driven by IOR benchmark
  - MPICH/TCP all-to-all gets 25% of this (sequoia/phloem) (This is an identified problem.)
  - Off-fabric 270 Gbit/s per rack to 10gE (netperf)

BG/P Linux Cluster Performance



OFED: Current BG Integration



Software OFED Verbs Providers

- Utilize Moore's Law anticipated transistors
  - A fraction of ASIC area dedicated to network/HCA
- Motivation
  - CPU Offload may not be as important as buffer management
  - Challenge to manage 1k to 10k, or 100k node network without CPU/memory off-load to additional hardware
- What is possible using a reasonable fraction of SoC ASIC area? (Cores, network interface, network logic)
- Several software OFED verbs providers
  - OSC software iWARP and derivatives
  - Soft RoCE
  - IBM Zurich SoftiWARP (siw)
  - Blue Gene/Q vRNIC

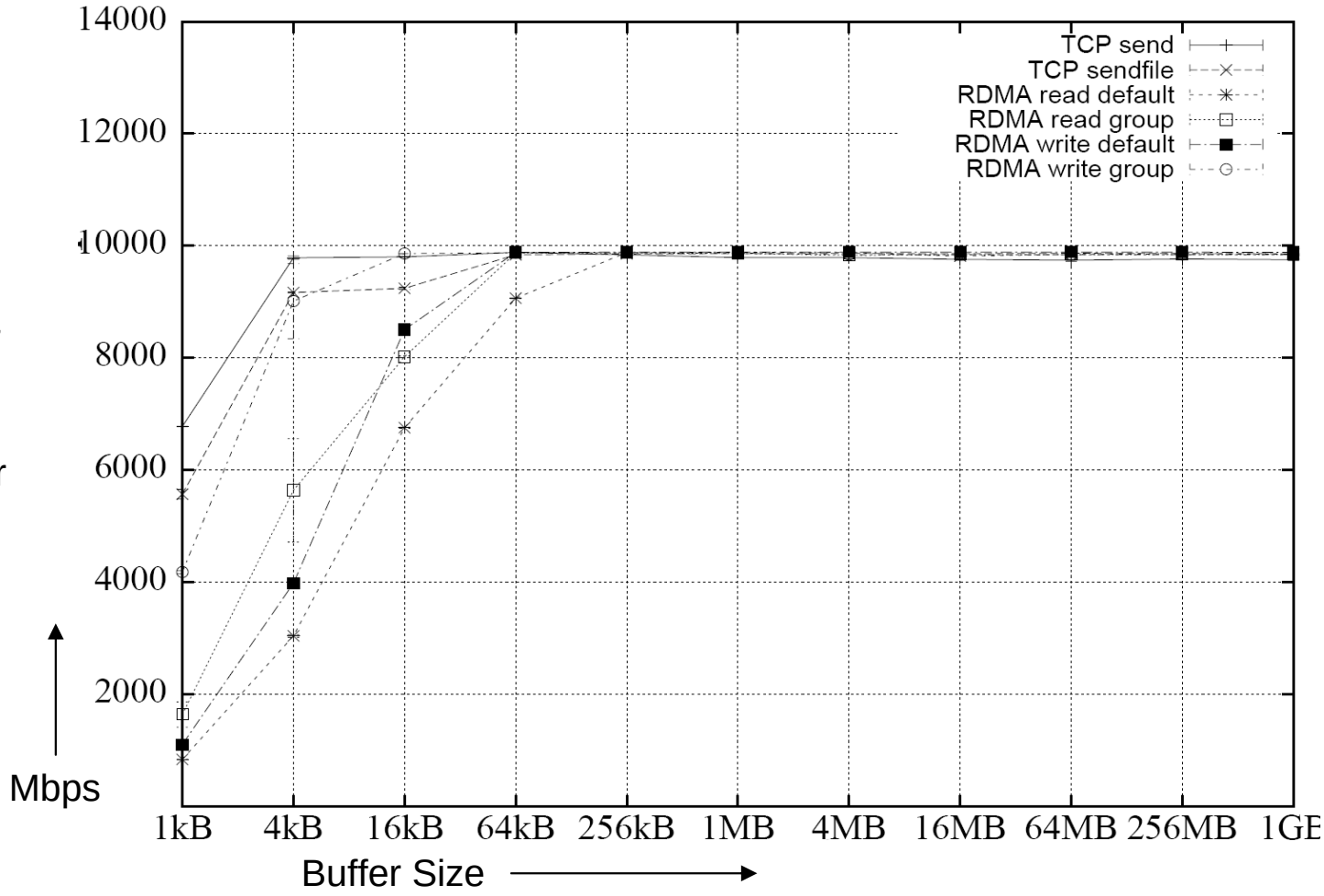
SoftiWARP (SIW) Update

- Kernel Client Support
  - Non-blocking Fast Path
  - DMA-Memory-Region Support
    - siw_get_dma_mr()
    - Private DMA address mapping functions, e.g.: siw_dma_map_single(), siw_dma_map_page() …
  - Fast memory registration (work in progress)
- Done all fixes as recommended on netdev/linux-rdma
  - Available at http://gitorious.org/softiwarp
- Applications tested
  - NFSv4
  - SDP (work in progress, needed FMR)
- GPFS tested on 2048 BG/P nodes
  - With SIW 15% better than using plain TCP sockets
    - More efficient communication buffer management
    - No skbuf buffering at receive socket

SIW Bandwidth Test

Simple point-to point Throughput

- Using netperf (extended with RDMA)
- Compare siw WRITE and READ throughput to plain TCP + TCP sendfile
- Zero copy transmit for siw and TCP sendfile (siw: non-signalled WRITE + extra 0-length READ)
- SoftiWARP on par for larger buffers
- Grouping work requests improves performance for small buffers
- SoftiWARP sensitive to SOCK_NODELAY for small buffers



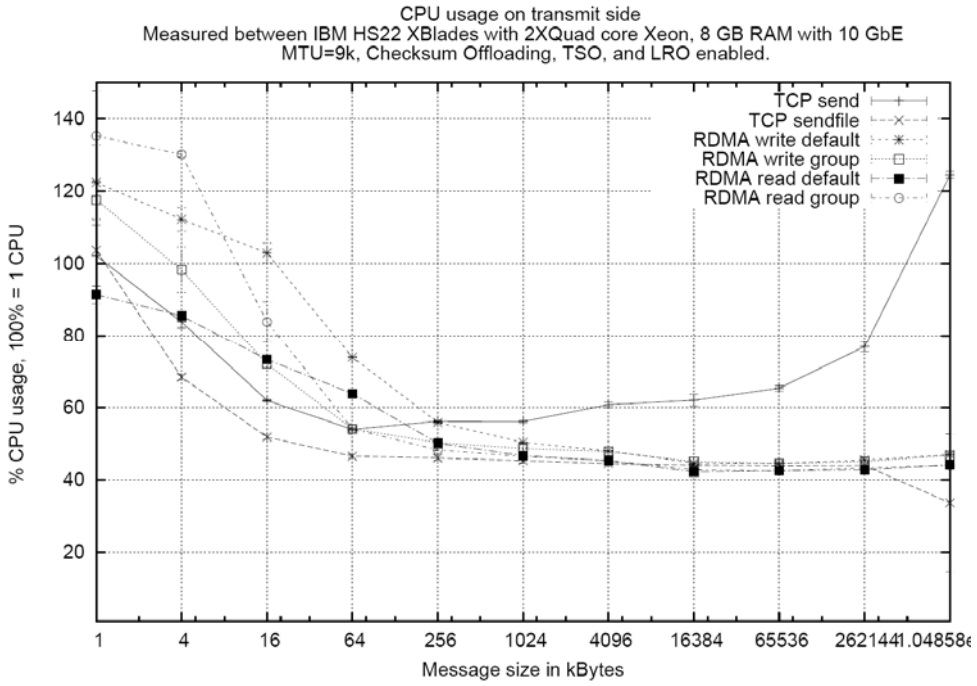
SIW CPU Usage

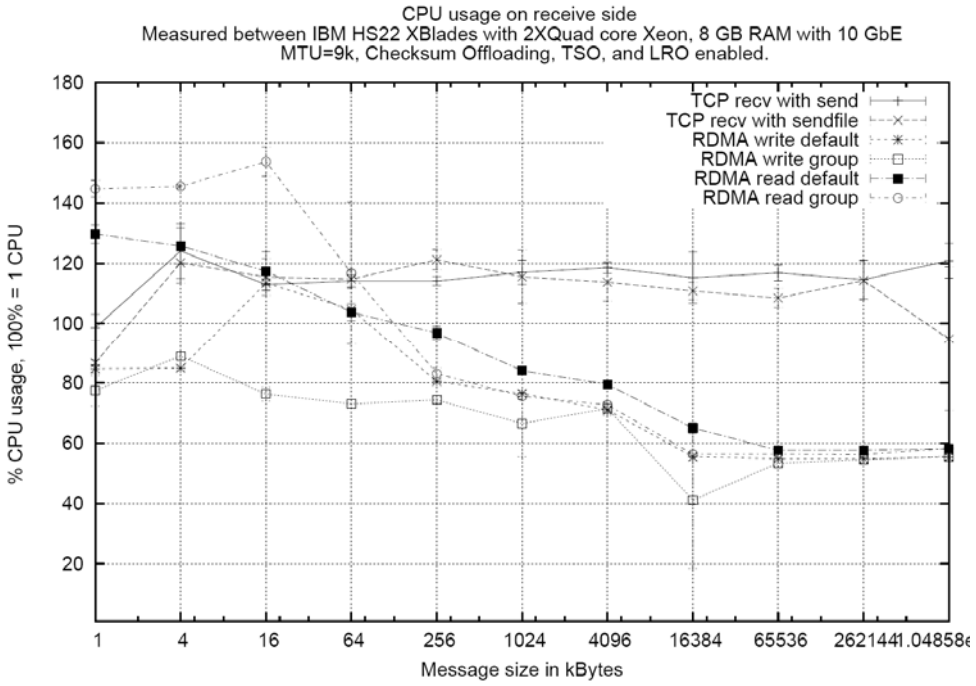
Send Side

- Zero copy transmit for SIW and TCP sendfile
- Comparable sender load for smaller messages
- TCP sender load problematic for large buffers

Receive Side

- SIW does data placement in softirq
- Using TCP, load stays constant over buffer size, always using more than one CPU
- With SIW, significantly less receiver load for larger buffers (about half)



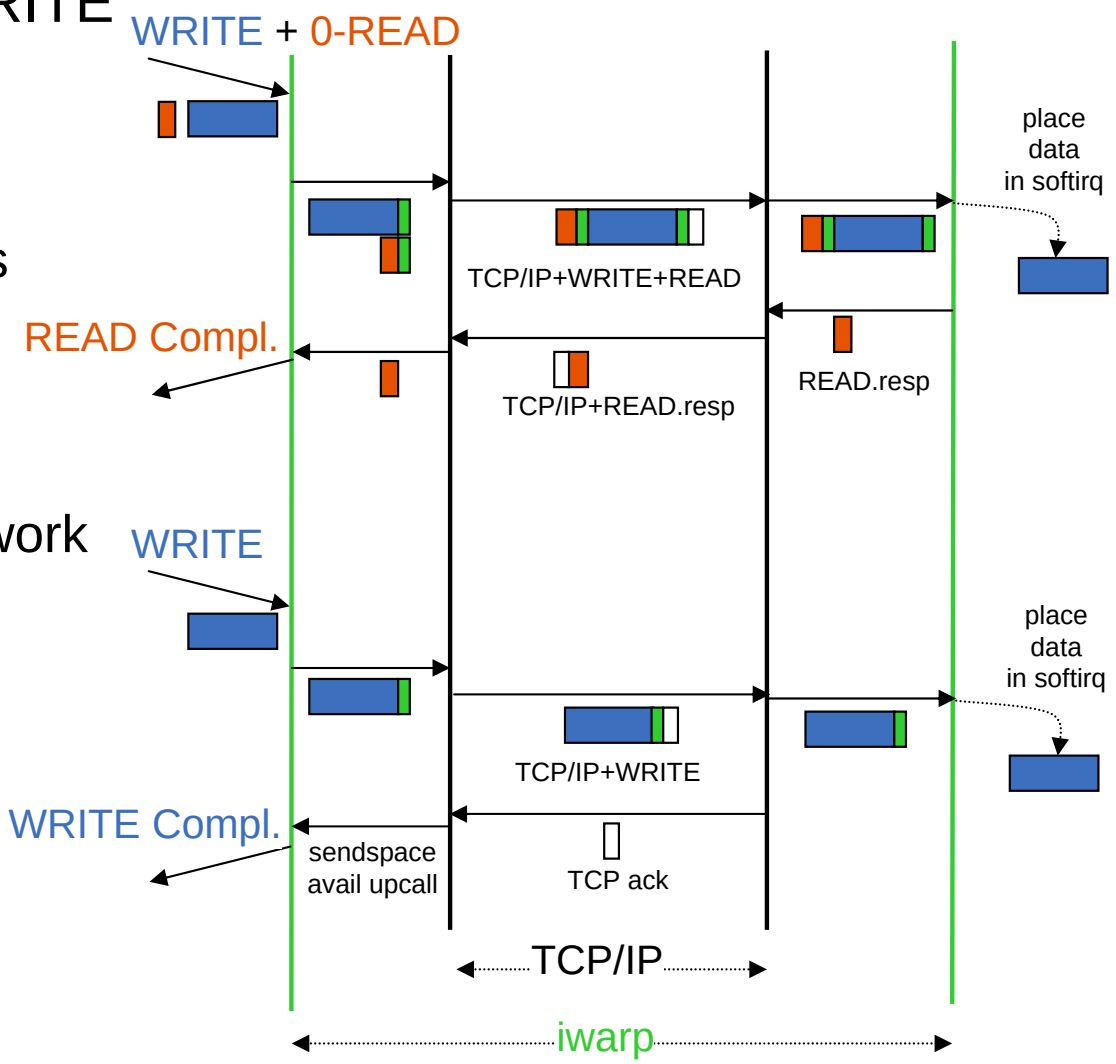


SIW Micro-benchmarks: Summary

- Bandwidth
  - Almost identical with using TCP on user socket
- CPU Usage
  - Significantly less CPU usage due to 0copy sending and softirq data placement
- Advantages of using iWARP including TCP
  - Benefit from stateless offloading
    - LRO/TSO available on most state-of-the-art adapters
  - Relatively good performance for small messages
  - Robustness and end2end of TCP/IP stack
  - RDMA semantics on anything and e2e
- Semantic Advantages
  - One-sided operations do not schedule peer application

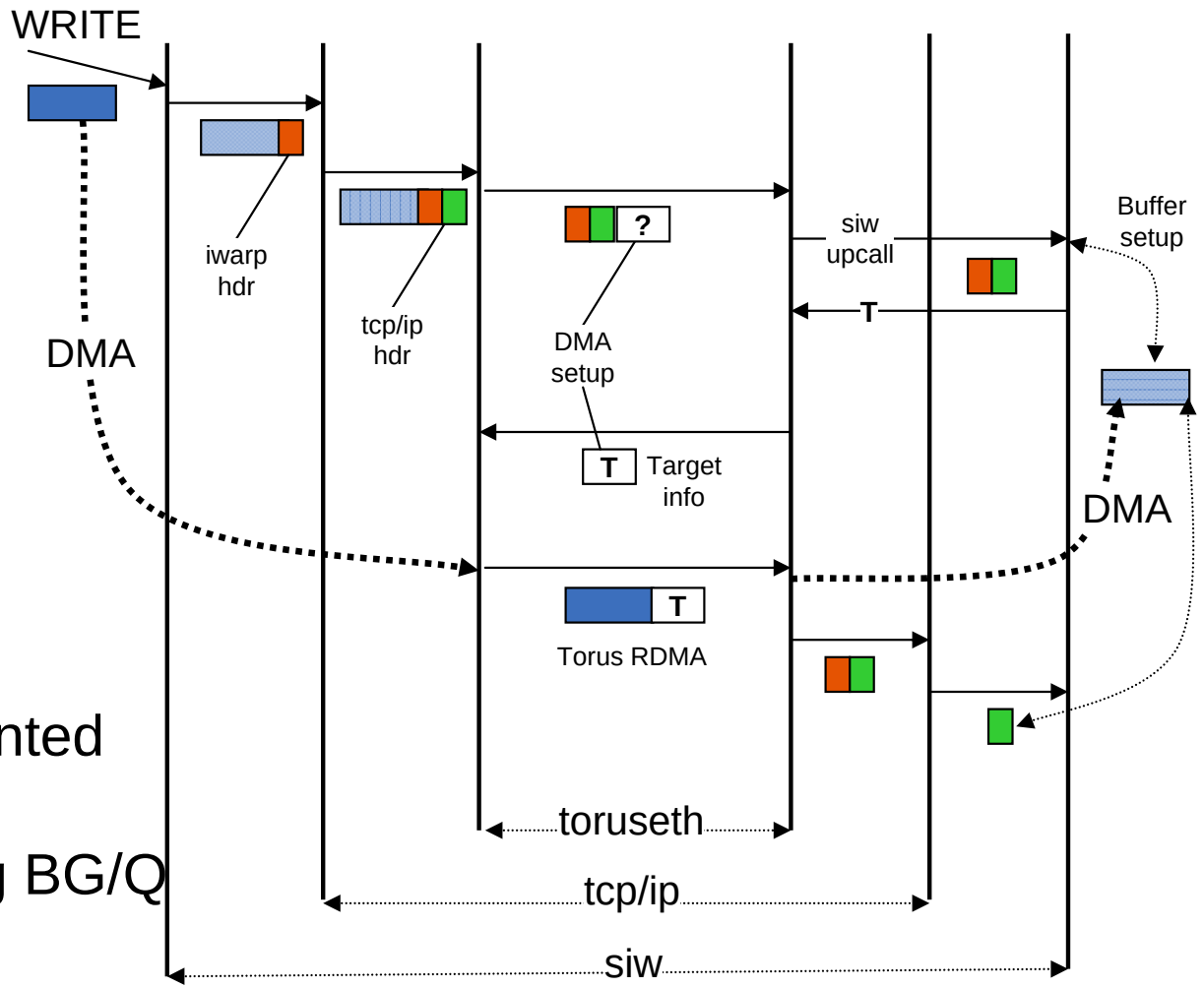
SIW: Current Work

- 0copy TX for signaled SEND/WRITE
  - Current solution: Non-signaled WR + zero length READ
  - Plan: Use local TCP send state info to determine when peer has received all data
- 0copy RX for all WR's
  - Restricted to HW support
  - Example: Blue Gene Torus network
- Memory Windows
- FMR
- Lazy Memory Registration
  - Don't pin application buffers, just bring in
- Applications, applications…

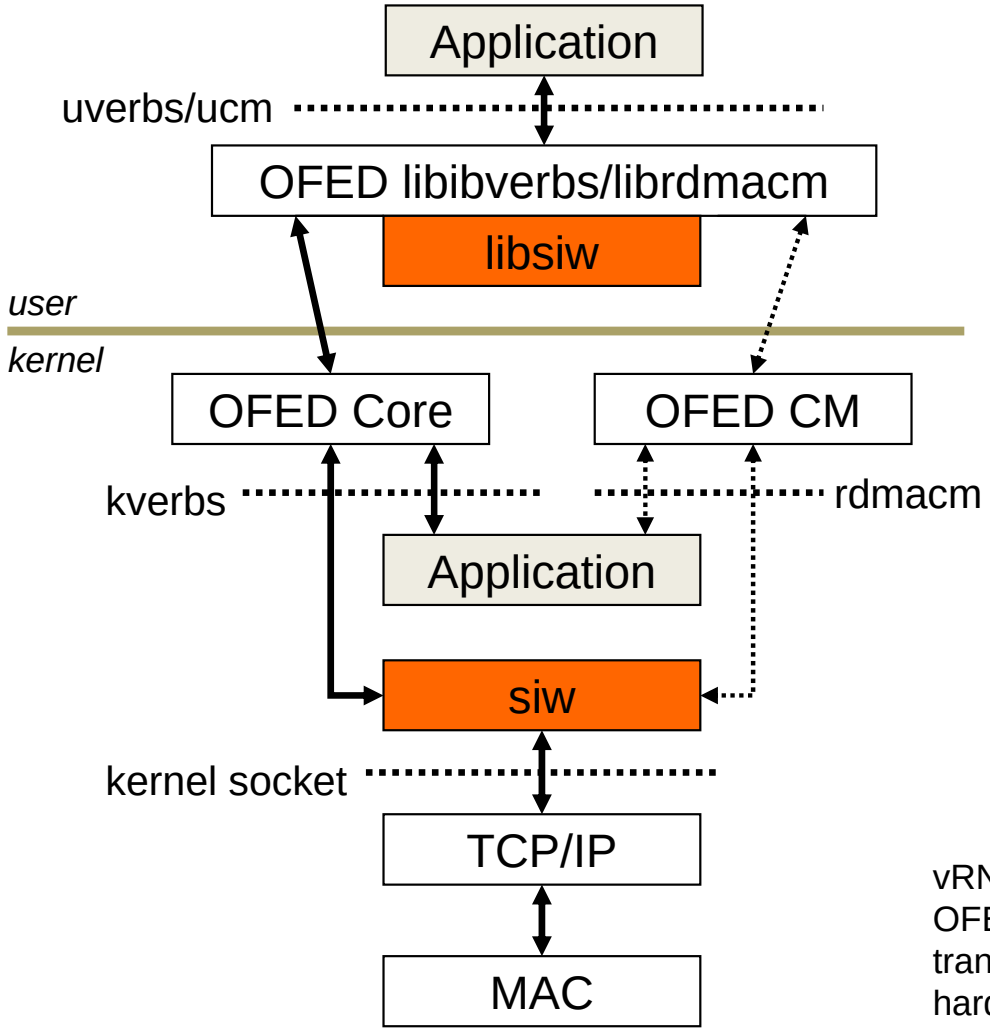


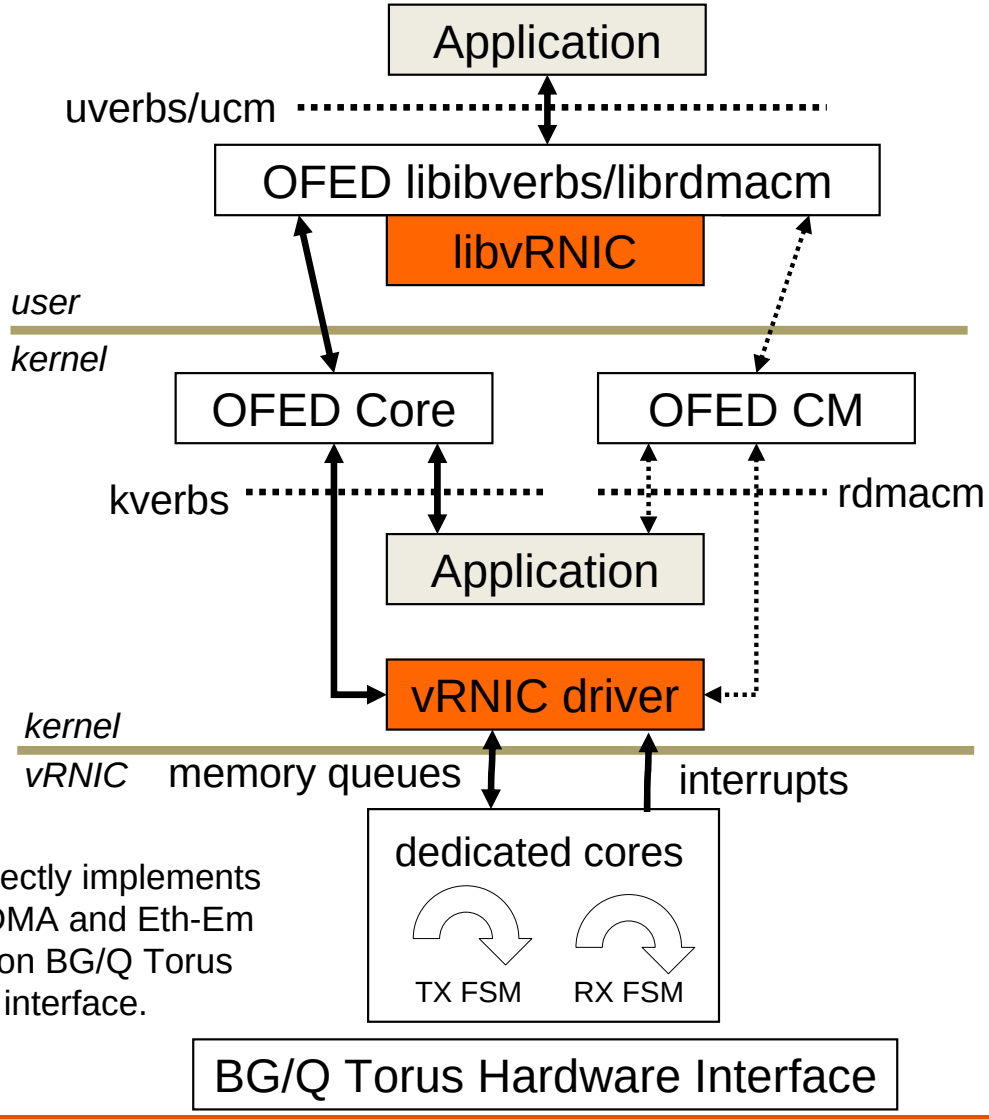
Zero Copy Receive on BG Torus

- Use DMA setup message to poll peer for target
  - Include transport headers in setup message
  - Directly inject headers in peer SIW control path
  - Feed back target buffer coordinates
  - Directly place into target
  - Inject TCP/IP/SIW envelop into stack after placement
- Basic functionality implemented for BG/P
- More flexibility on upcoming BG/Q



SoftiWARP vs vRNIC





vRNIC directly implements OFED RDMA and Eth-Em transport on BG/Q Torus hardware interface.

Conclusions and Further Research

- Cost-efficient network
  - High performance up to very high node counts
  - Avoids dedicated HW, keeps incremental transistor density busy
- Active Storage
  - Extends BG Architecture for Storage Class Memory integration
  - Environment for Exascale I/O and analytics applications
- OFED for BGAS
  - Industry standard environment
  - Industry standard (verbs) API
  - Better buffer management for middleware
  - Integrates HW assists
  - OFED can integrate more transports than InfiniBand and iWARP
- Two Software Verbs Providers discussed
  - SIW
    - 'RDMA for the people', aims for Linux mainline integration
    - Quick start for BGAS, can leverage HW assists
  - vRNIC
    - Better tailored for BGAS environment
    - Can seamlessly replace SIW (thanks to OFED!)