2015 SPEC Distinguished Dissertation Award

Scalable End-to-End Data I/O over Enterprise and Data-Center Networks

Yufei Ren

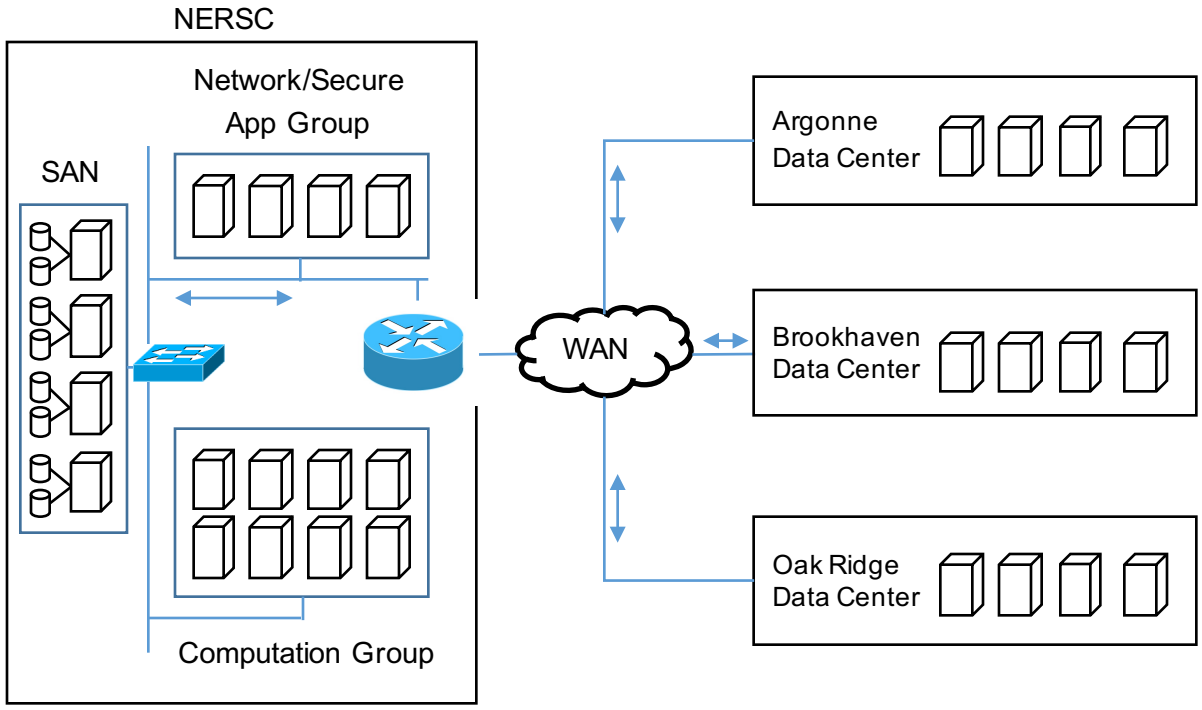
now with IBM T. J. Watson Research Center

graduated from Stony Brook University

Prof. Dantong Yu (*Advisor*), *BNL & Stony Brook University*

March 16th, 2016

Data I/O on Modern Hardware



- Bulk data movement
  - 100 Gbps high throughput
  - Across WANs with long latency
- Storage data caching
- Advanced hardware
  - Zero-copy network
  - Large-scale asymmetric memory layout
- Existing solutions
  - TCP/IP based
  - GridFTP, OpenSSH
  - iSCSI, NFS

GAMING

PRO VISUALIZATION

DATA CENTER

DEEP LEARNING AUTO

WATSON	CHAINER	THEANO	MATCONVNET
IBM	Preferred Networks	Université de Montréal	UNIVERSITY OF OXFORD

TENSORFLOW	CNTK	TORCH	CAFFE
Google	Microsoft	facebook	Berkeley UNIVERSITY OF CALIFORNIA

Research Challenges

- Achieve zero-copy in end-to-end data path
- Scale zero-copy based network transmission in WANs
- Improve data locality in large-scale NUMA systems

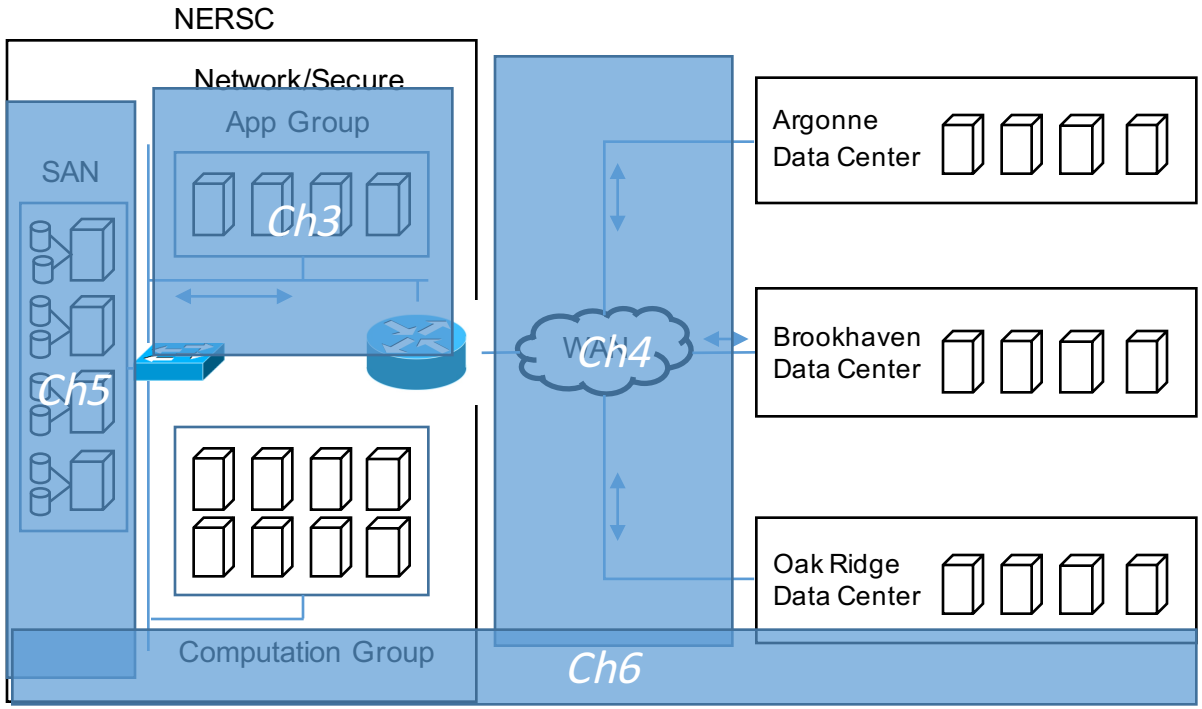
Our Goal: **Build** systems for scalable end-to-end data I/O and **Understand** system performance characteristics

Efficient bulk data movement in LAN and WAN

Effective cache design for large scale memory systems

Performance evaluation with advanced hardware

Thesis Outline



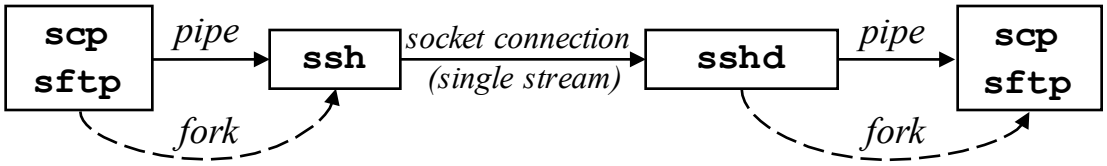
- Software design (Ch3)
- RDMA-based data transfer protocol (Ch4)
- Storage caching optimization (Ch5)
- End-to-end performance optimization (Ch6)

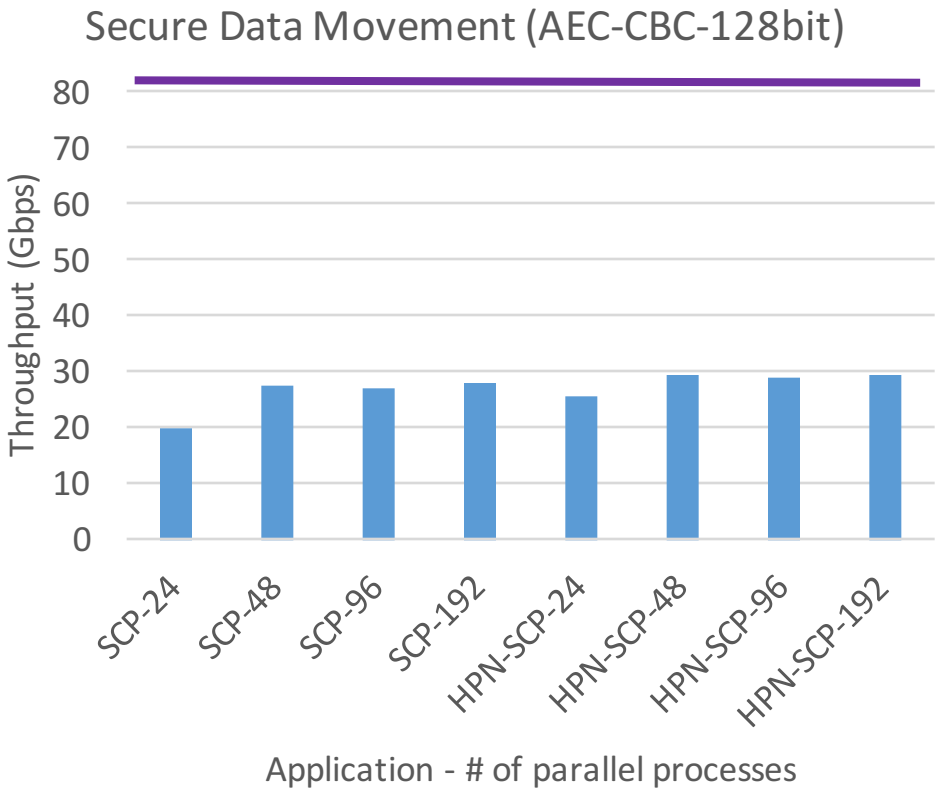
End-to-End Asynchronous I/O

Publications

- **Yufei Ren**, Tan Li, Dantong Yu, Shudong Jin, "**End-to-End Asynchronous Processing for High Throughput Computing**", *under revision*.
- [SC'13] **Yufei Ren**, Tan Li, Dantong Yu, Shudong Jin, Thomas Robertazzi, "**Design and Performance Evaluation of NUMA-Aware RDMA-Based End-to-End Data Transfer Systems**", In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (**SC '13**)*, Denver, Colorado, November 2013.
- [JSS'13] **Yufei Ren**, Tan Li, Dantong Yu, Shudong Jin, Thomas Robertazzi, "**Design and Testbed Evaluation of RDMA-based Middleware for High-performance Data Transfer Applications**", *Journal of Systems and Software (**JSS**)*, Volume 86, Issue 7, July 2013, Pages 1850-1863, ISSN 0164-1212, 10.1016/j.jss.2013.01.070.
- [SC'12] **Yufei Ren**, Tan Li, Dantong Yu, Shudong Jin, Thomas Robertazzi, Brian L. Tierney, Eric Pouyoul, "**Protocols for Wide-Area Data-intensive Applications: Design and Performance Issues**", In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (**SC '12**)*, Salt Lake City, Utah, November 2012.

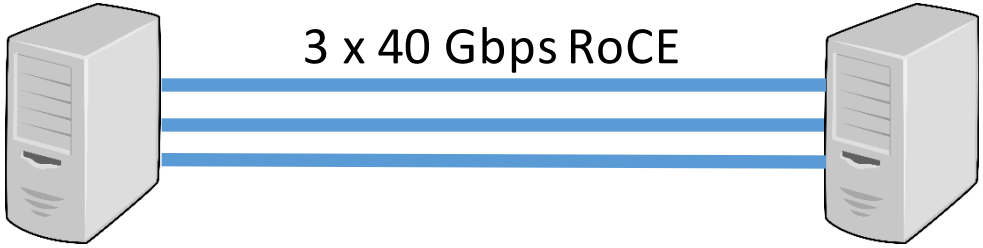
Classical Software Design

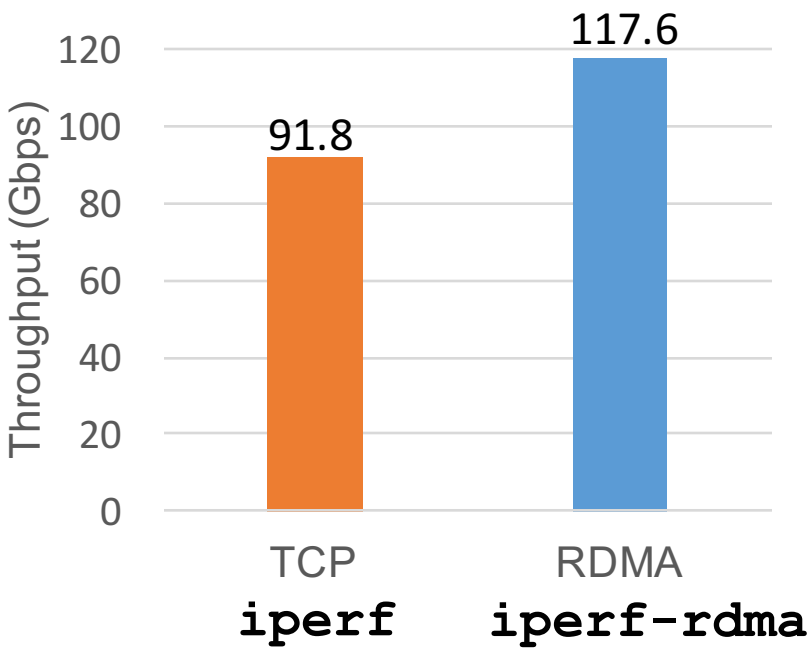




- OpenSSH & HPN-SCP
- Heavily rely on data copy based OS services
  - Inter-process communication
  - TCP-based data transfer
  - Massive interrupt handling
  - Synchronous and blocking I/O
- Hardware Performance
  - Network: **80**Gbps
  - Storage: **100**Gbps
  - CPU AES Encryption: **142**Gbps

Network Transmission Profile





- Memory Bandwidth: 400Gbps
- NUMA Random pick: 83.5Gbps
- NUMA-aware: 10% improvement
- TCP wasn't able to saturate this fat link
- TCP benchmark takes shortcut
  - **User space memory ->** CPU Cache -> Kernel space memory -> DMA to NIC
- TCP: 35% CPU is used for memory copy
  - `copy_user_generic_string()`
- RDMA achieved **98%** bare-metal throughput.

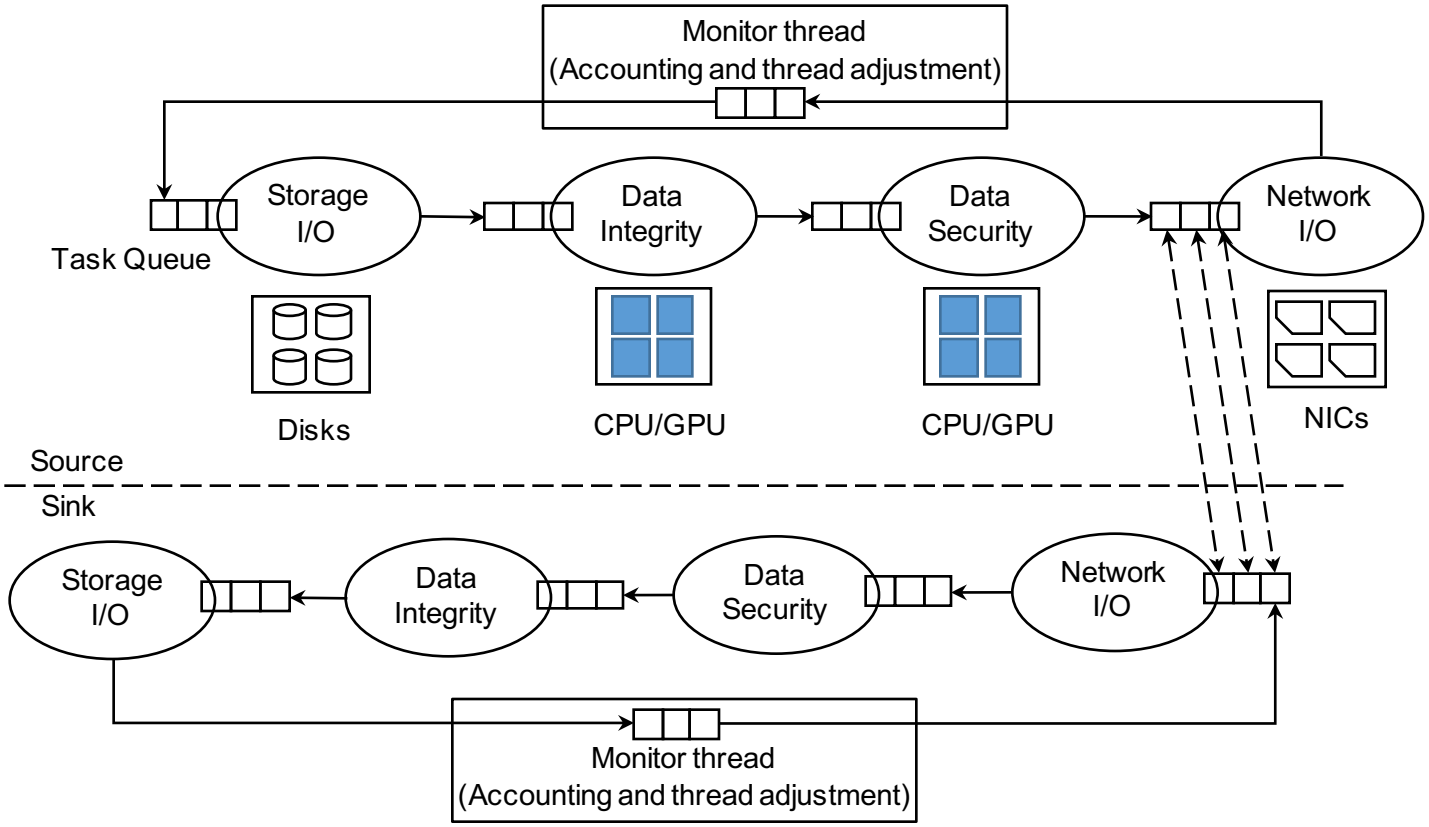
Related work

- High Performance TCP-based solutions
  - GridFTP
  - bbcp
- RDMA Send/Receive for bulk data movement
  - Lai et al. ICPP'09
- RDMA write based solution
  - Tian et al. Journal of Comp & Elec Eng '12
  - One RTT, not scalable to WAN
- Shared control channel and data channel
- Not studied in WAN environment

Overarching Goal

- Efficient secure data transfer
  - Storage I/O, Computing, network I/O
- Orchestrate hardware and retain zero-copy
- Observations on hardware trend
  - **Network** – **2** magnitude improvement (1 Gbps - 100 Gbps with a single port)
  - **Storage** – **1** magnitude improvement (200MB/s disk – 2GB/s flash)
  - **Multi-core** – 18 cores per die X 8 socket
  - **Memory** capacity improvement ≫ throughput (200Gbps on a NUMA node)
- Memory-centric design

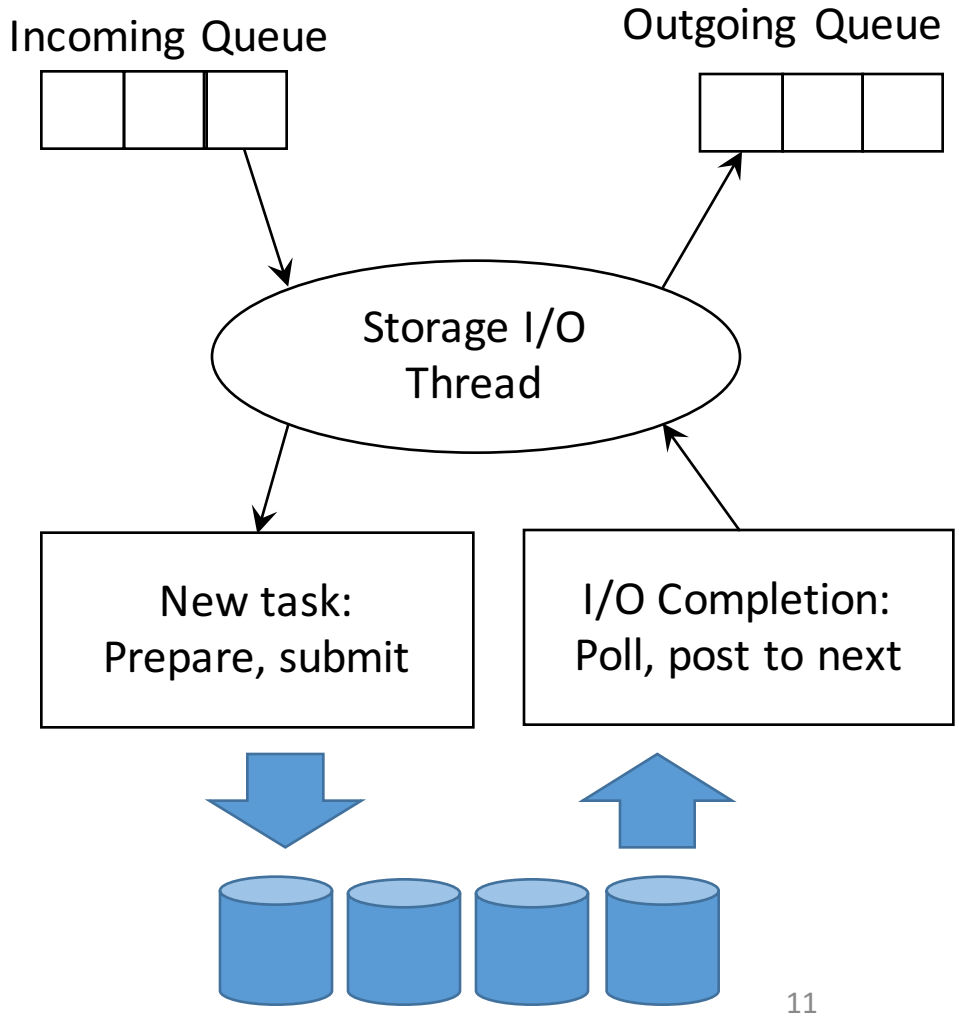
ACES Framework



- **A**synchronous **C**oncurrent **E**vent-driven **S**taged
- End-to-end zero-copy

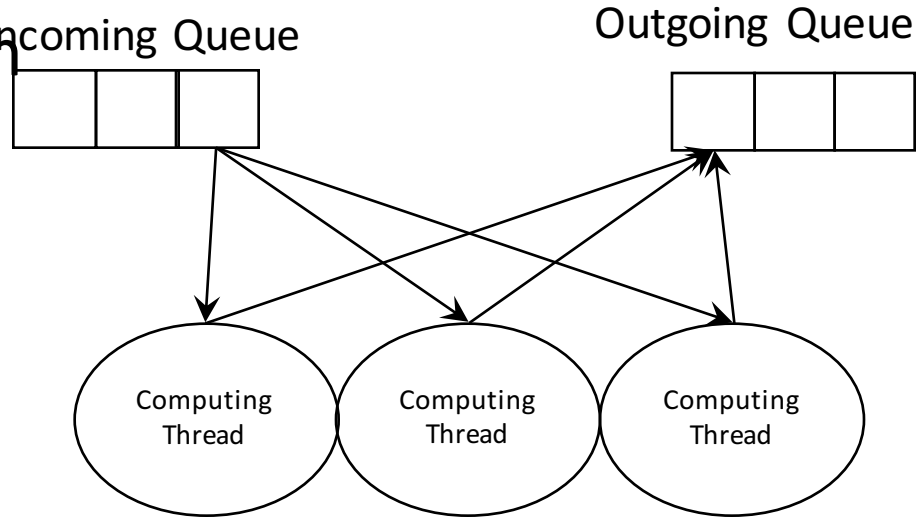
Asynchronous Storage I/O

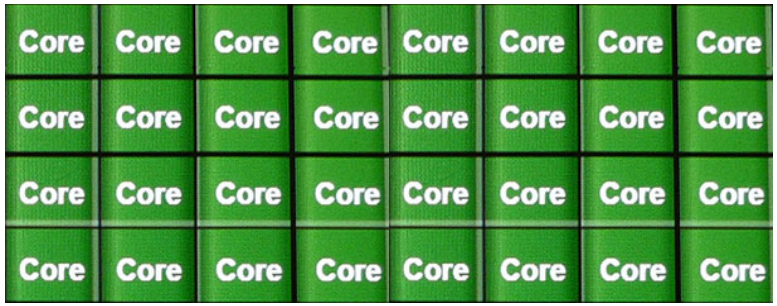
- Asynchronous I/O separates I/O submission and I/O completion
- `libaio` – Linux asynchronous I/O library
  - io_submit()
  - io_getevents()



Parallel Computing

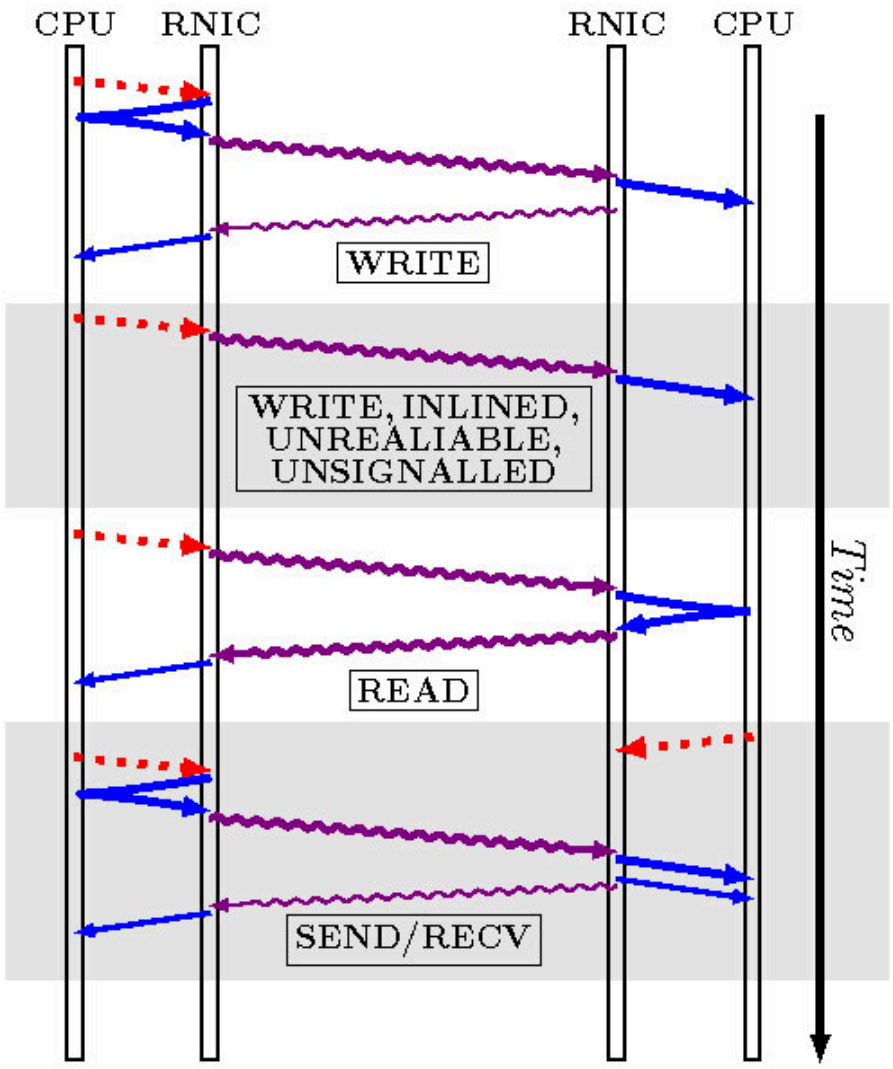
- Multi-threaded parallelism
- Encryption/Decryption
  - Advanced Encryption Standard (AES)
  - Session-based (i.e. single file)
- Data integrity
  - Block-based
  - crc32, Adler32
- Adjust # of threads dynamically





Zero-copy Network I/O

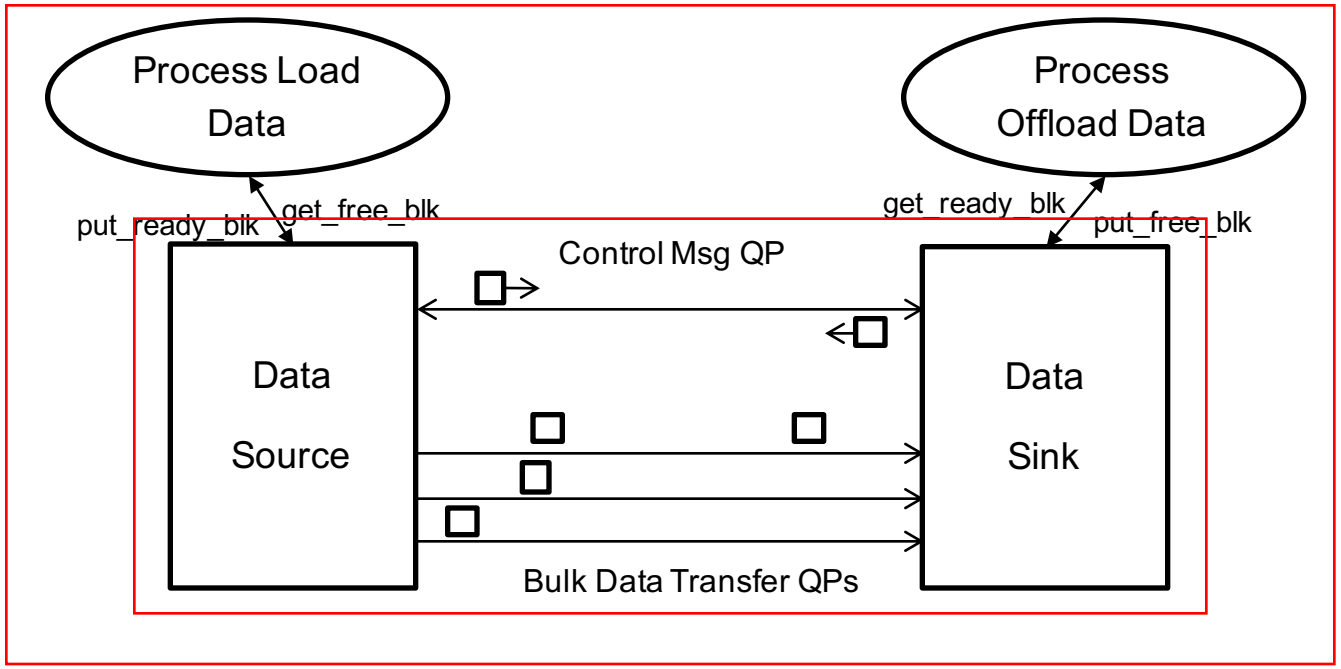
- High throughput
  - Achieve line speed
- Low cost
  - CPU utilization
  - Memory footprints
- Scalability
  - 100 Gbps and Beyond
  - Long latency WAN
- Network independent
  - InfiniBand: High I/O depth on a single QP
  - RoCE
  - iWARP: Multiple QPs
- Verbs
  - "Verbs Programming is like the assembly language version of network programming."



Network Protocol Overview

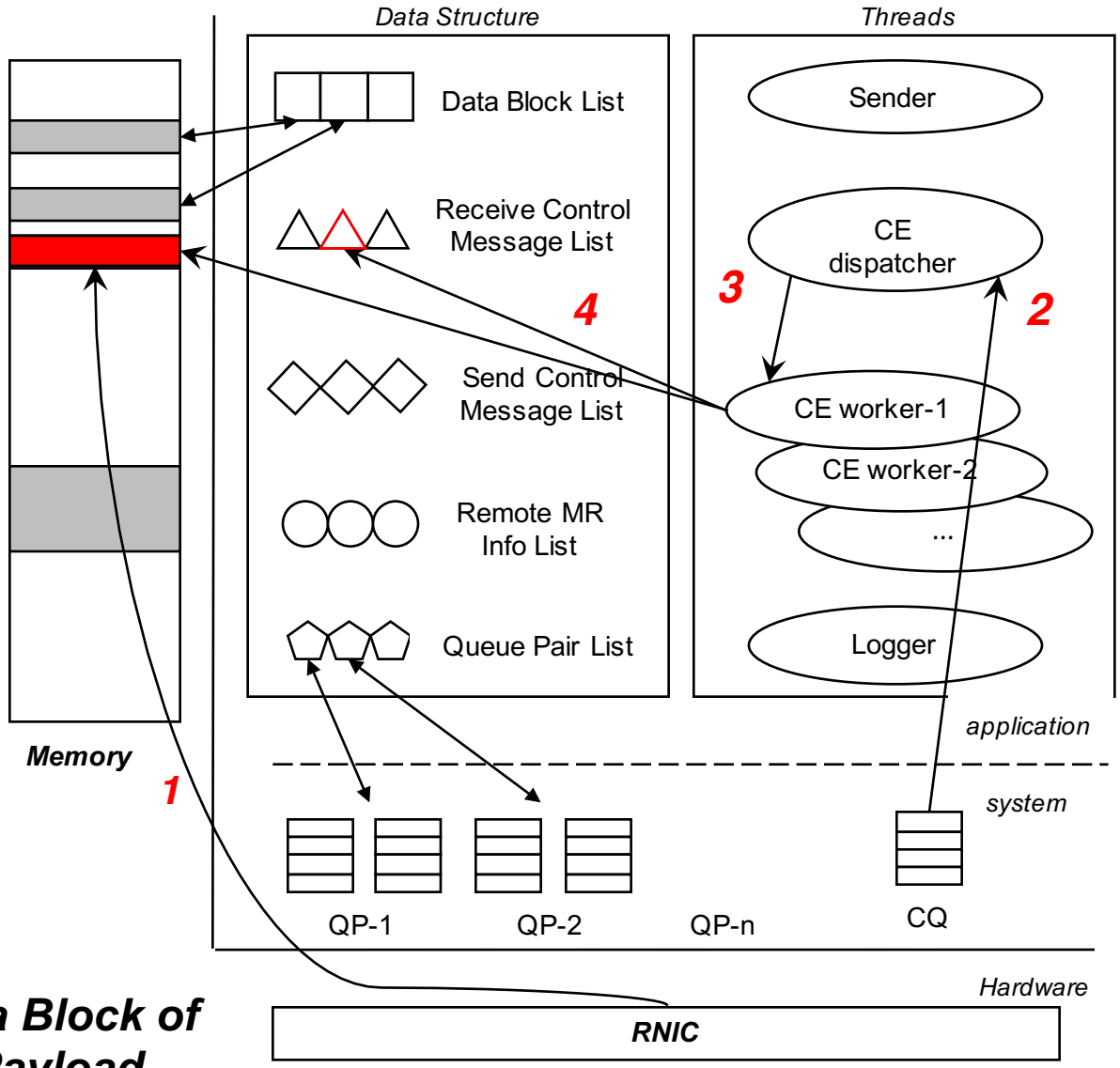
- Design choice
  - One-sided for bulk data movement (credit-based)
  - Two-sided for control messages (send with inline)
- Asynchronous protocol to overlap latency impact
- Multiple reliable queue pairs for data transfer
- Proactive feedback

RFTP: RDMA-enabled FTP Service



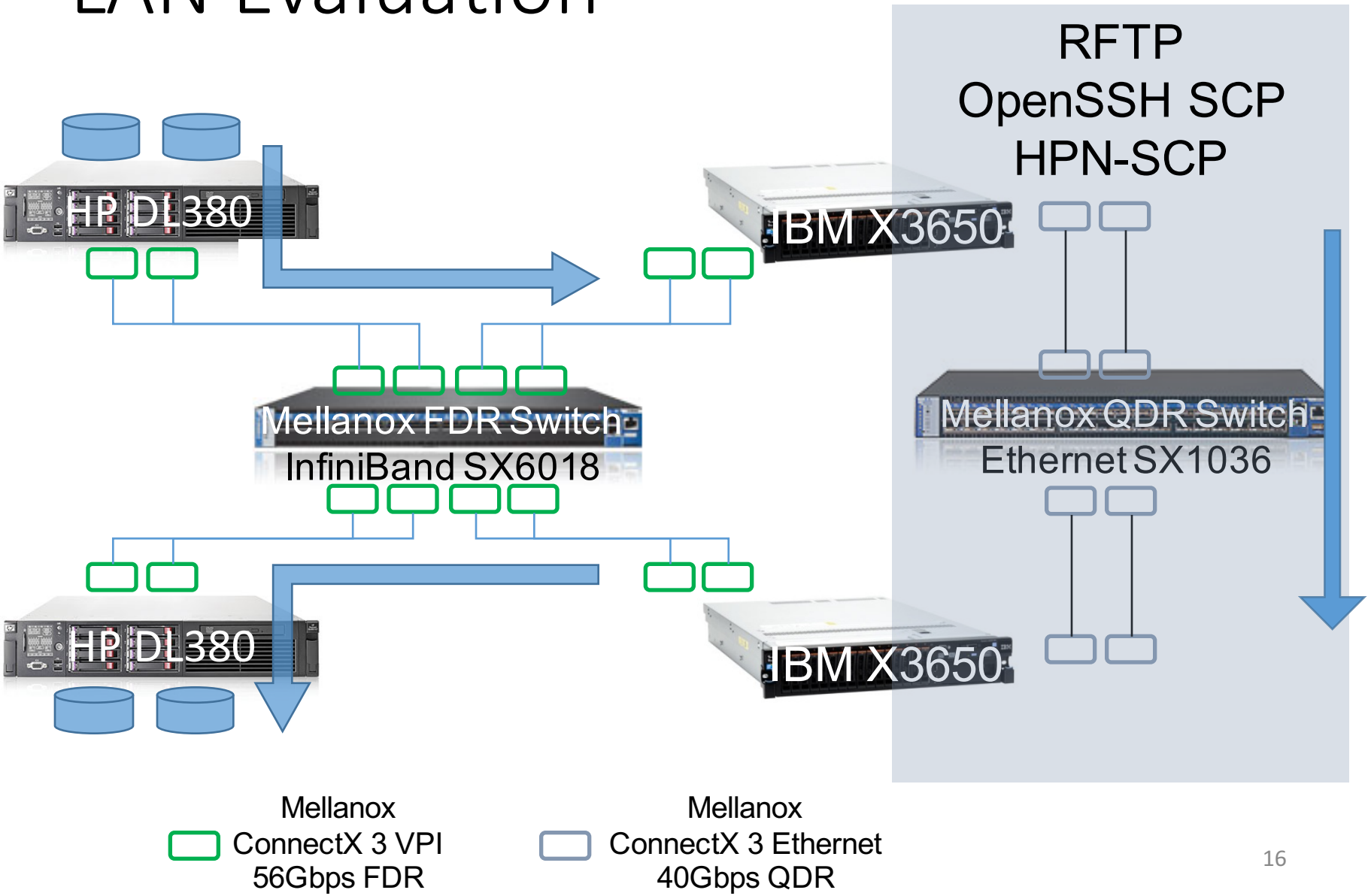
Implementation: Software Architecture

- Data Structure
  - Memory management
  - Control message management
  - Credit management
  - Connection management
- Threads
  - Multi-threaded
  - Master-worker
- Event-driven
  - Polling and dispatch

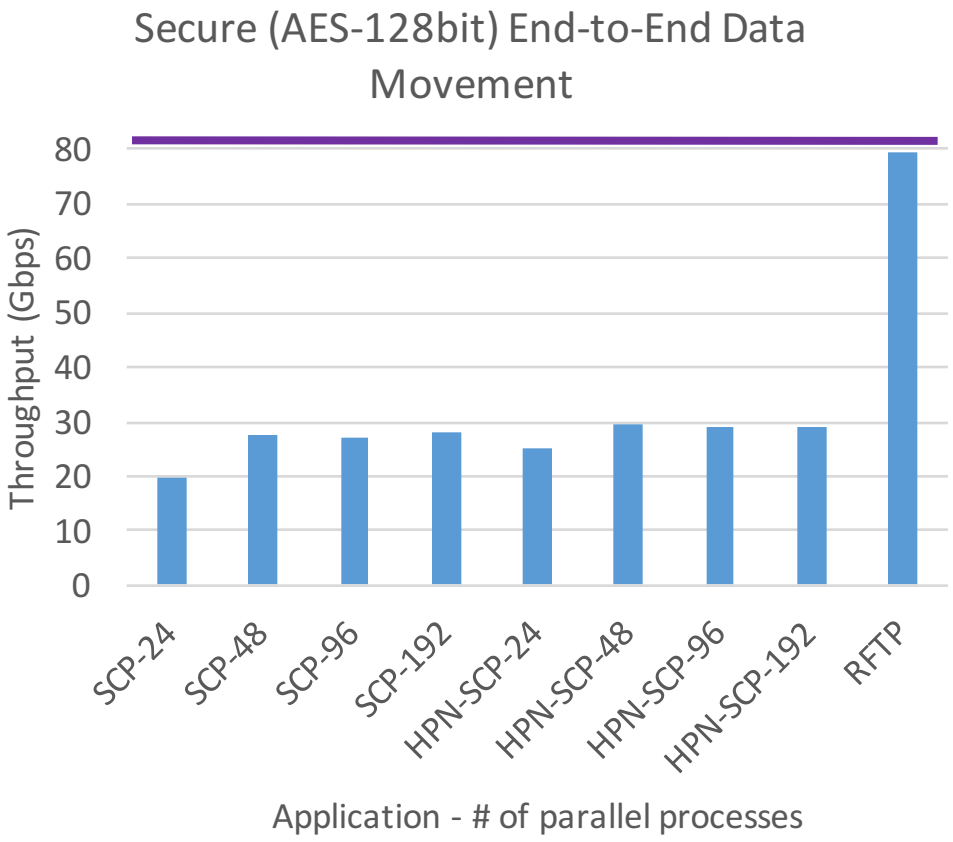


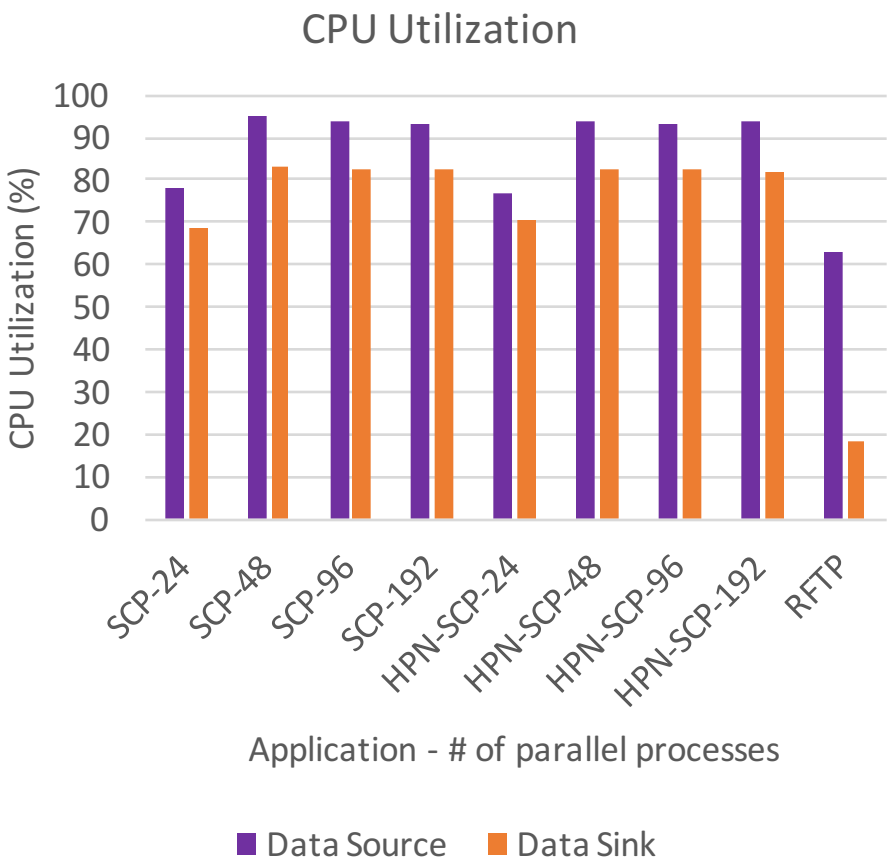
Receive a Block of User Payload

LAN Evaluation

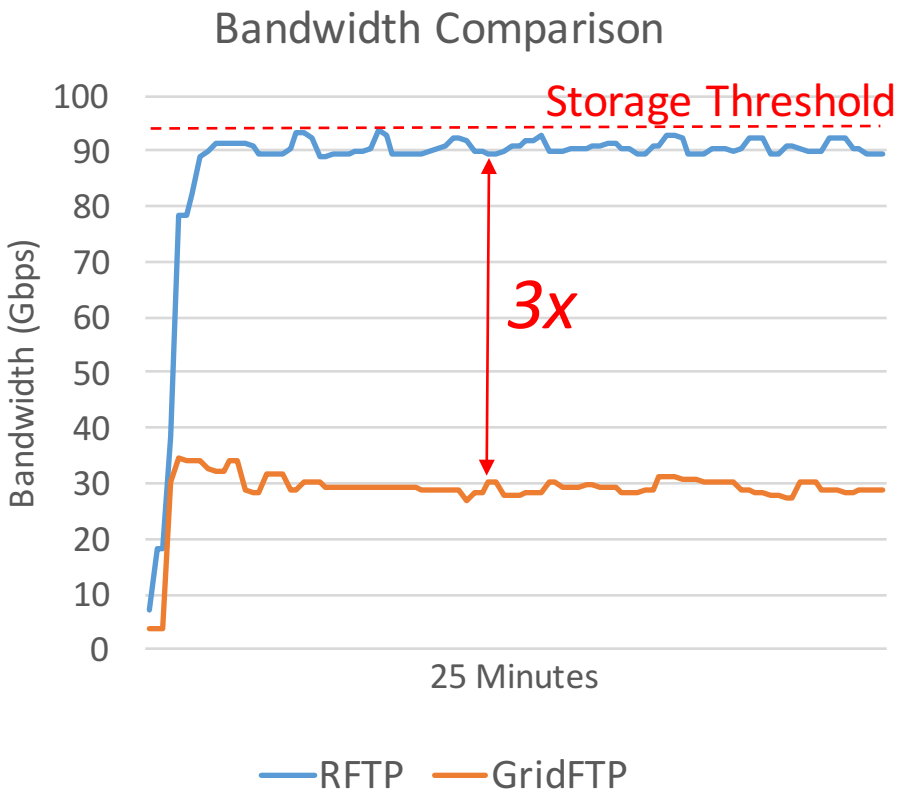


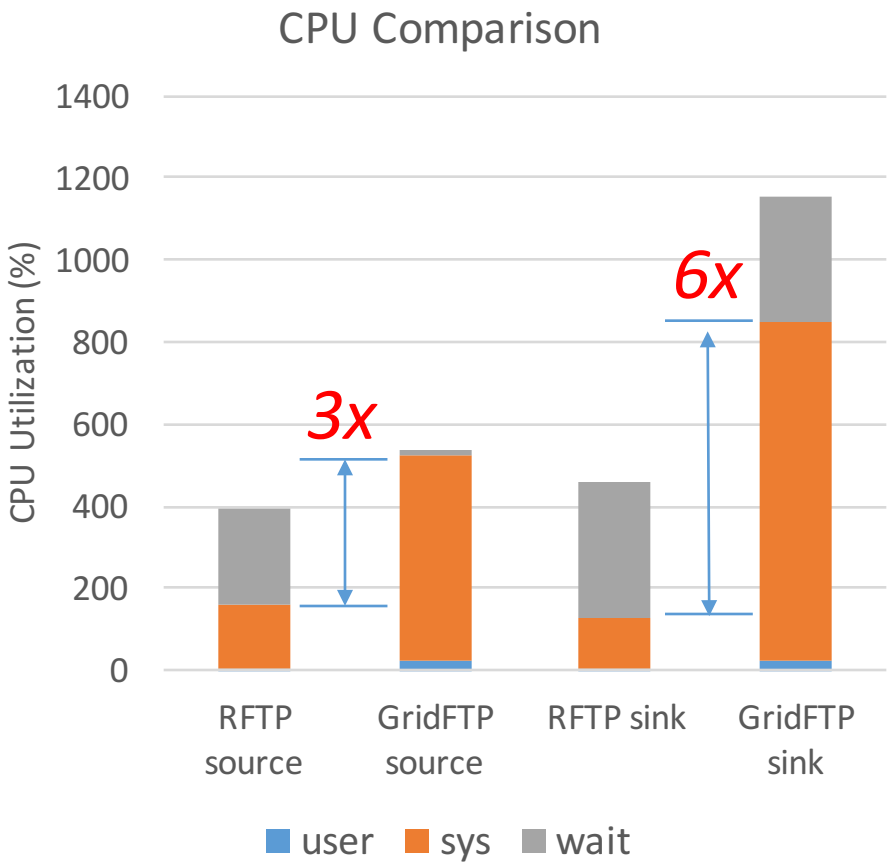
Secure Bulk Data Movement



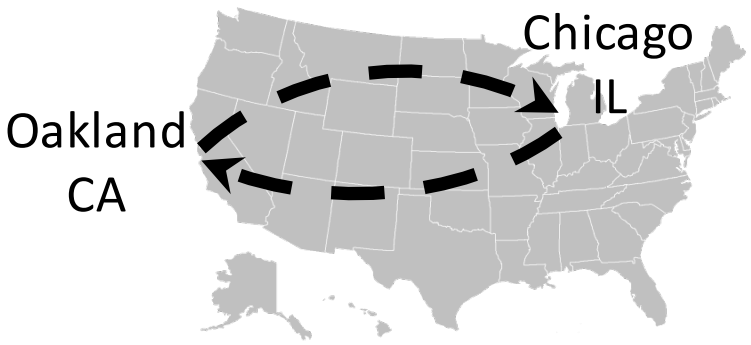


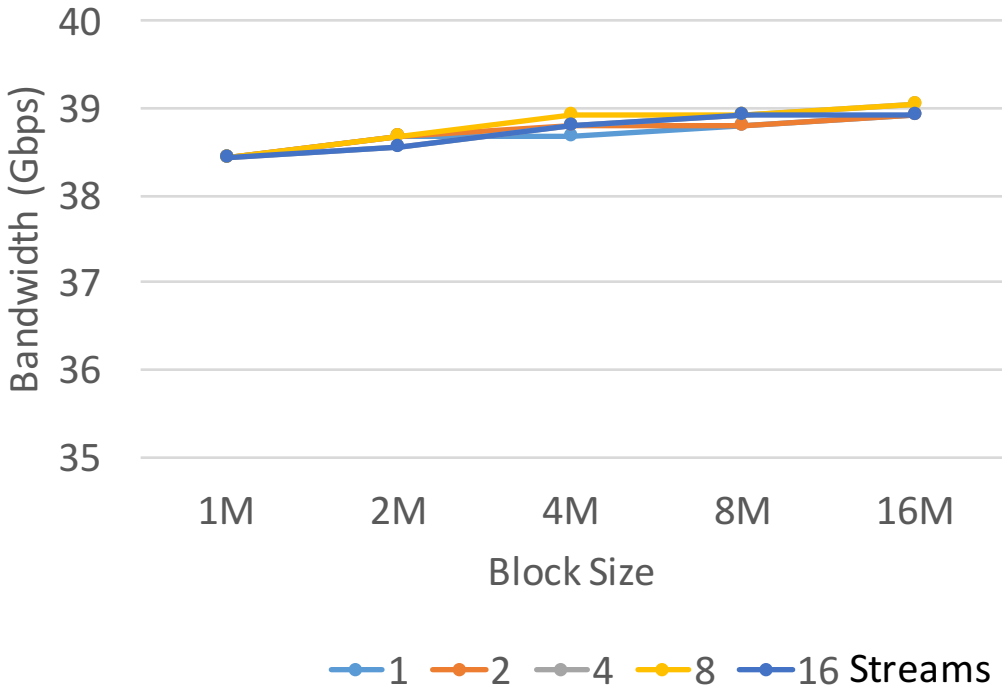
RFTP vs. GridFTP





40 Gbps Long Distance WAN Testbed





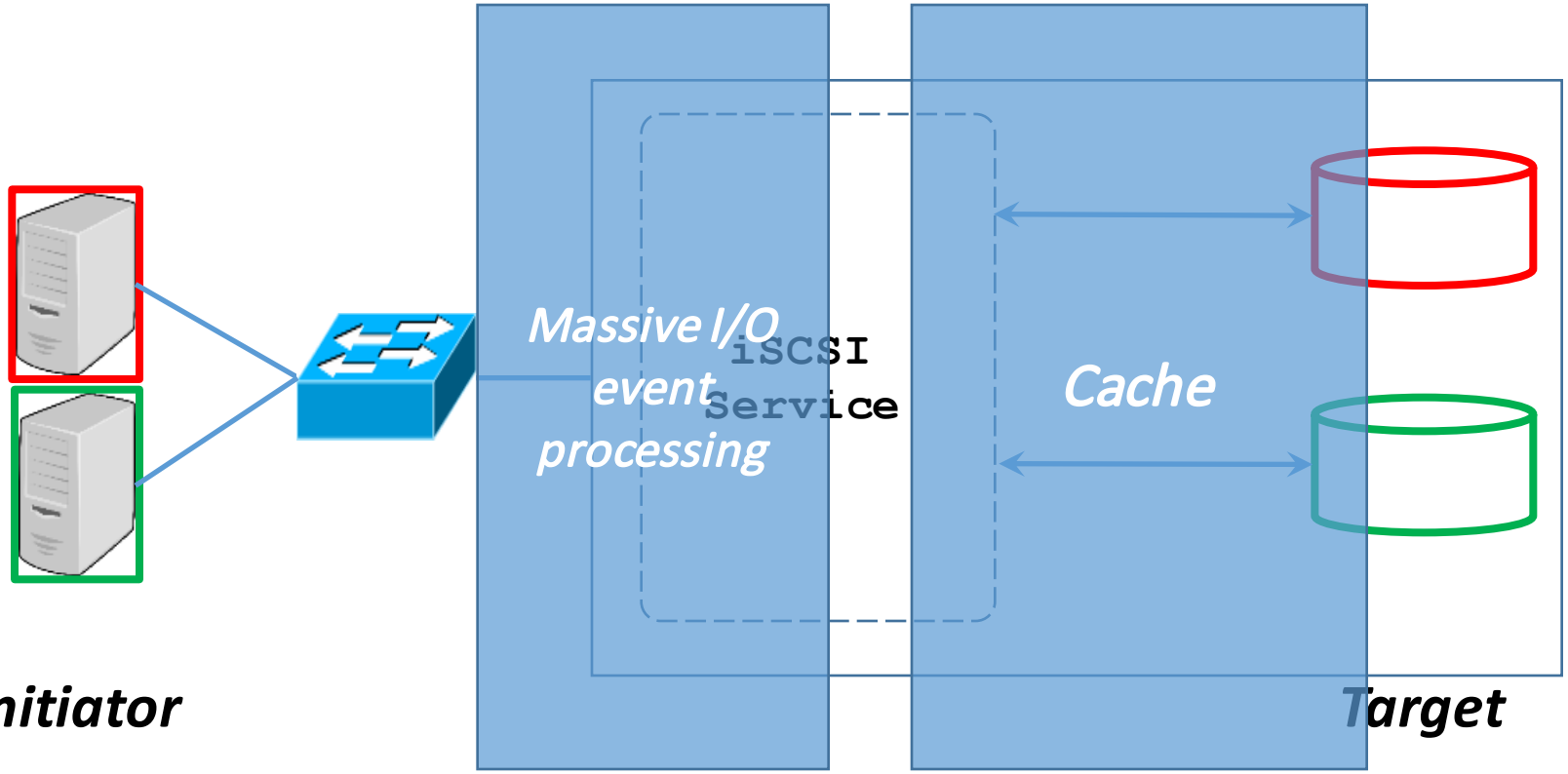
- 40 Gbps RoCE WAN
- 4,000 miles loopback
- RTT: 95 millisecond
- BDP: 500 MB
- *Will RFTP be scalable in WAN?*

NUMA-aware Cache for iSCSI Servers

Publications

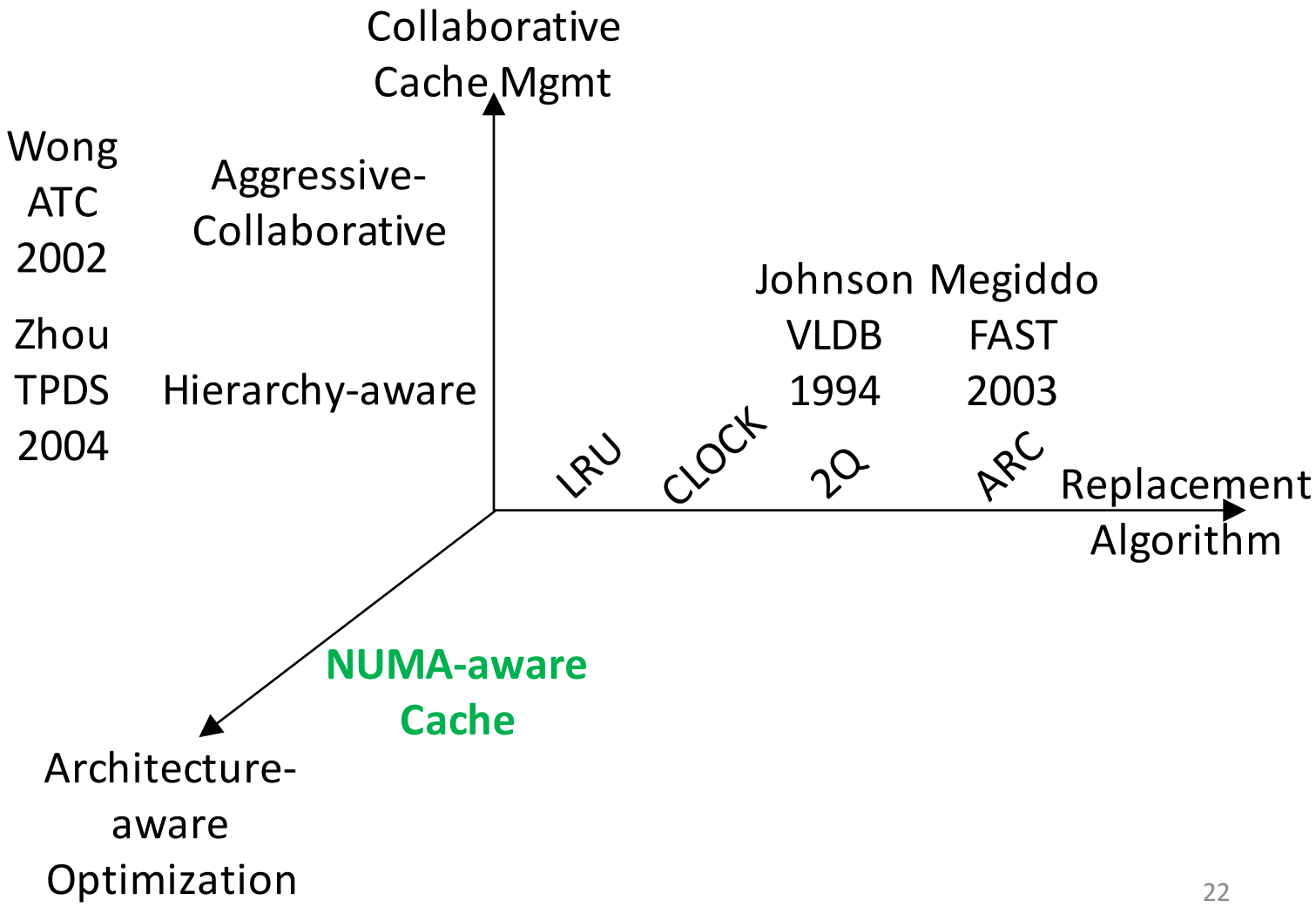
- [TPDS'15] **Yufei Ren**, Tan Li, Dantong Yu, Shudong Jin, Thomas Robertazzi, "Design, Implementation, and Evaluation of a NUMA-Aware Cache for iSCSI Storage Servers", IEEE Transactions on Parallel & Distributed Systems (TPDS), vol.26, no. 2, pp. 413-422, Feb. 2015

SAN Overview



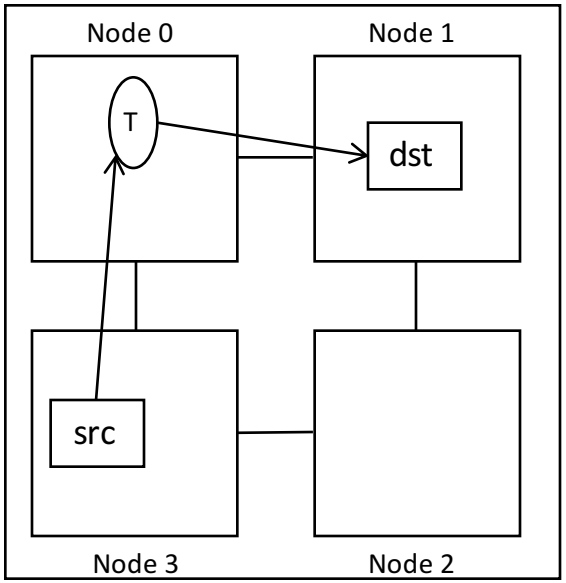
- Large-scale asymmetric memory caching system
- Massive I/O event processing
- iSCSI/iSER: SCSI cmd and data in TCP/IP and RDMA
- Cache: storage performance acceleration

Related work: Caching in SAN

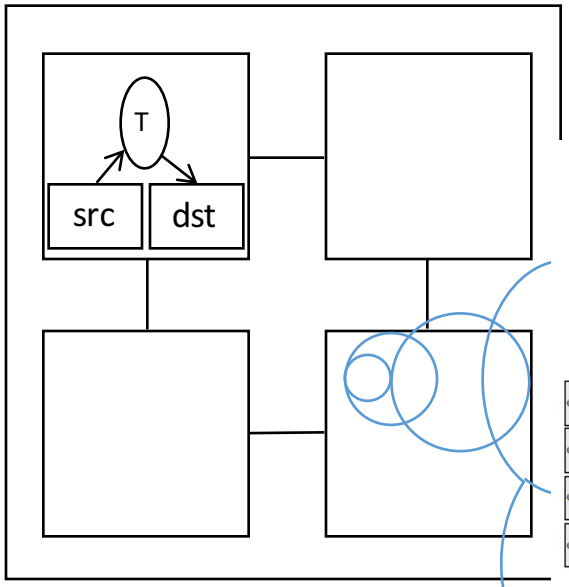


Caching in a NUMA host

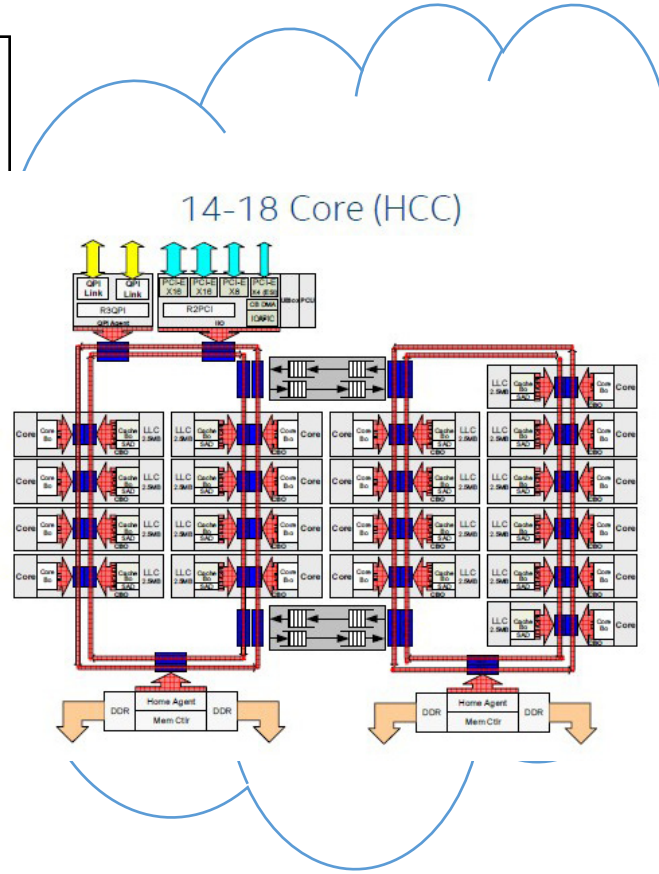
```
memcpy(dst, src, len);
```



NUMA-unaware



NUMA-aware



STREAM ***3.3 GB/s*** ***18.9 GB/s***

- [ICPP'13] Tan Li, **Yufei Ren**, Dantong Yu, Shudong Jin, Thomas Robertazzi, "**Characterization of Input/Output Bandwidth Performance Models in NUMA Architecture for Data Intensive Applications**", In *Proceedings of the International Conference on Parallel Processing ICPP '13, Lyon, France, October 2013*

iSCSI/iSER Performance Modeling for Cached Data

- Latency (from initiator perspective)

$$Latency = RTT + \frac{I/OSize}{MemBW} + \frac{I/OSize}{NetworkBW} + QueuingDelay$$

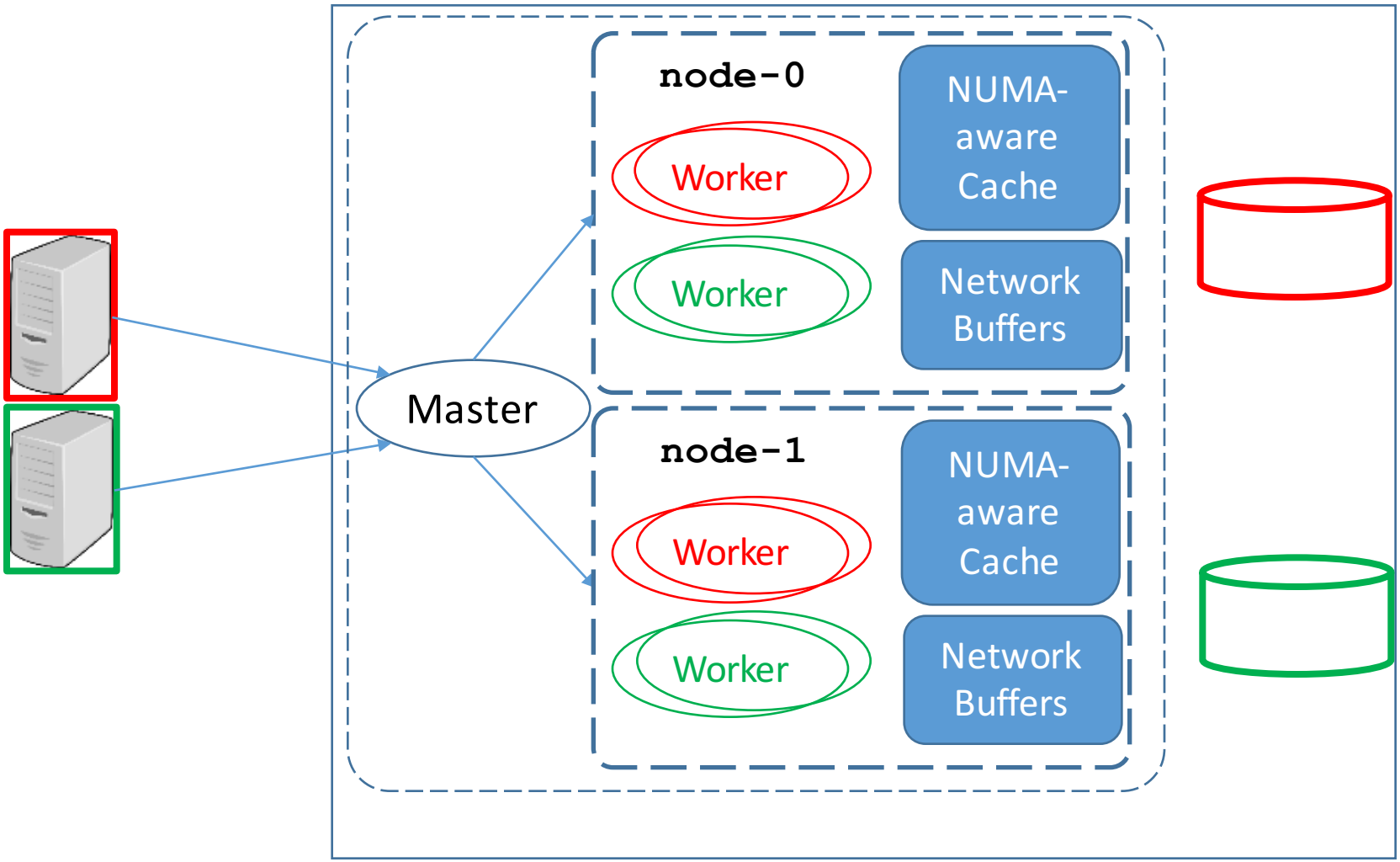
- Bandwidth

$$BW_{CachedData} = \min(BW_{memory}, BW_{network})$$

$$MemBW_{CachedData} = \sum_{i=0}^{n-1} Weight_i * MemBW_i$$

NetworkBW << MemoryBW —NUMA / RDMA→ NetworkBW < MemoryBW

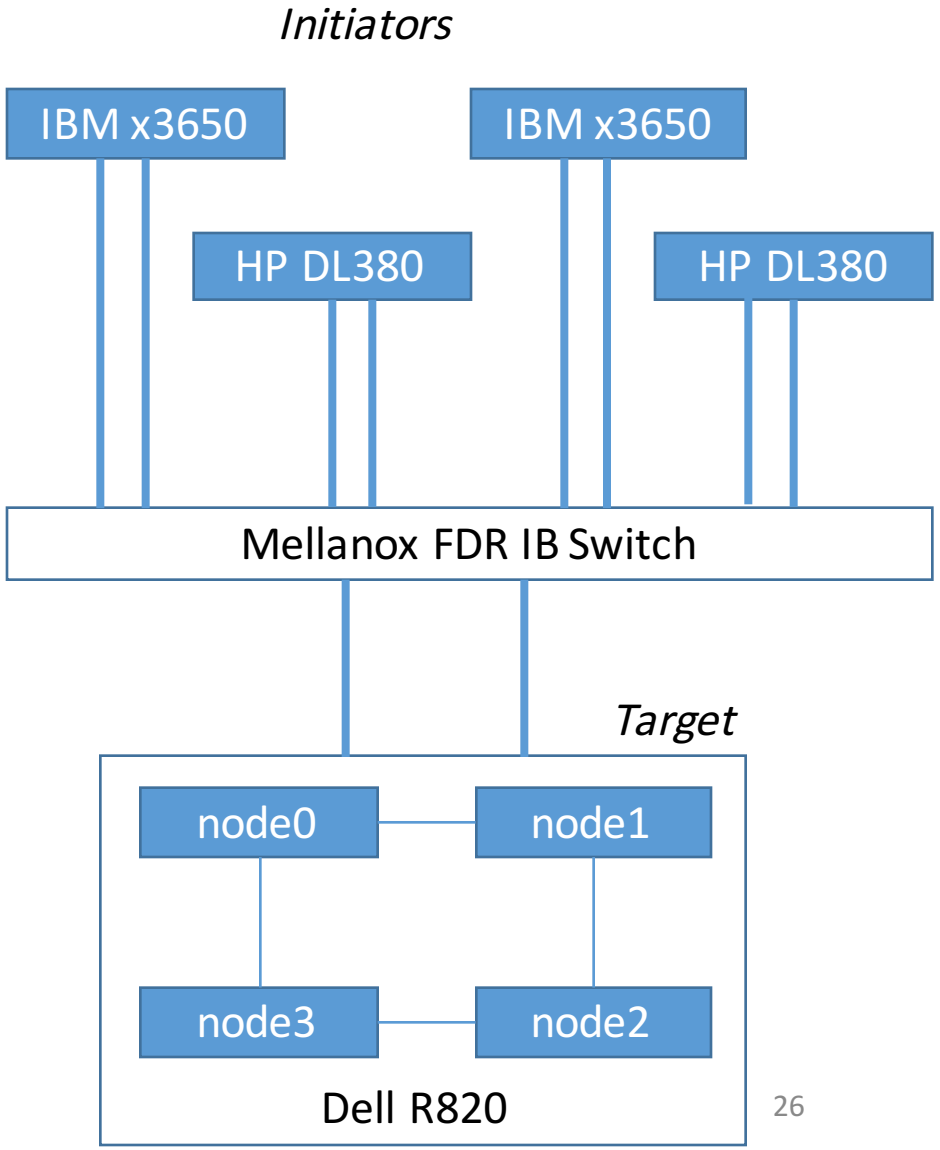
NUMA-aware Cache for SAN



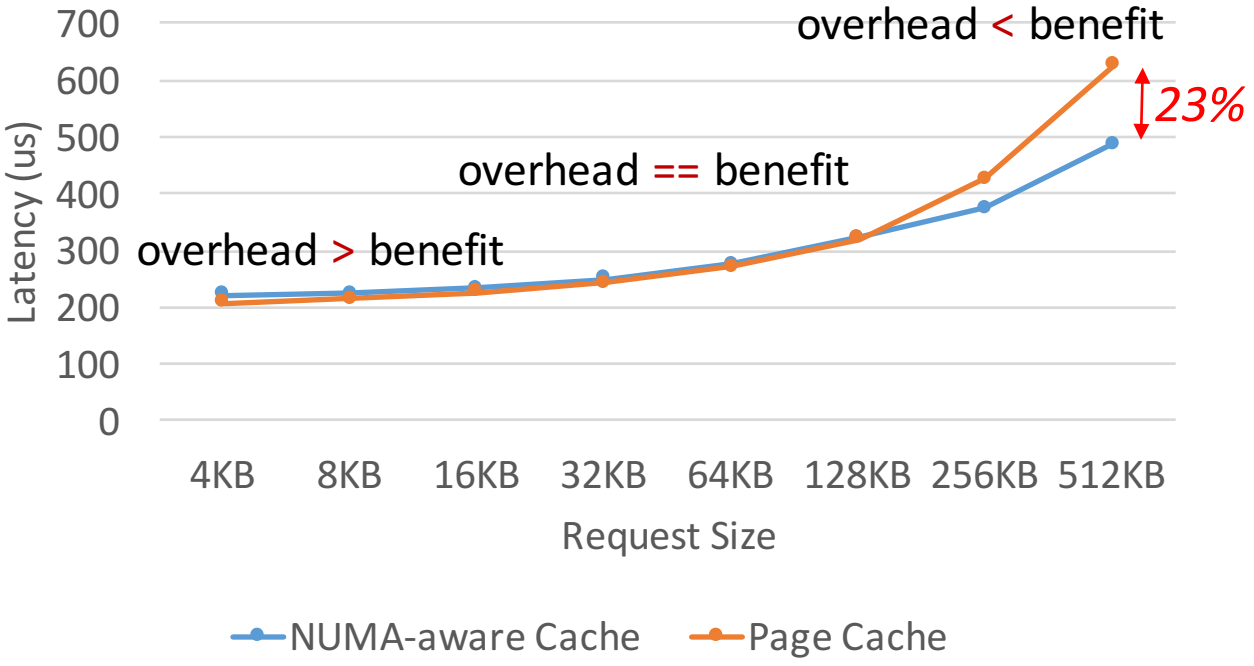
- Allocate thread, cache, and network buffer in each node

Evaluation

- 4-node NUMA host – Dell R820
  - Intel Xeon E5-4620 with QPI
  - **768 GB RAM**
  - Local memory bw: 18.71 GB/s
  - Remote memory bw: 3.26 GB/s
- 4 initiators
- Network adapters: 56Gbps InfiniBand FDR
- OS page cache vs. NUMA-aware Cache
- Fully cached data access
- Benchmark: `fio` with random read I/Os

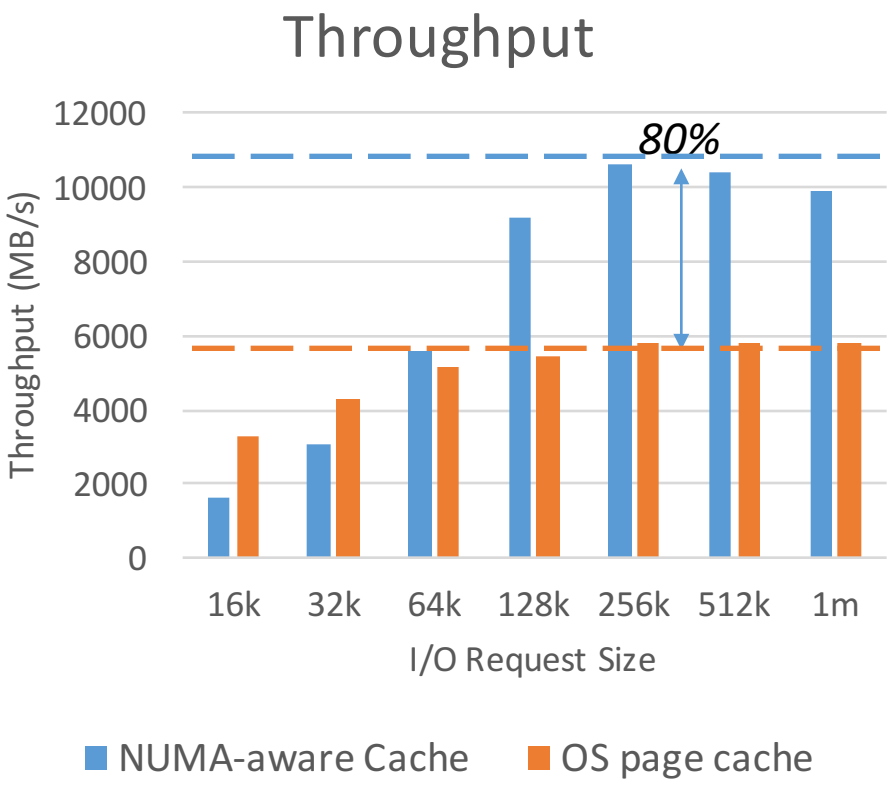


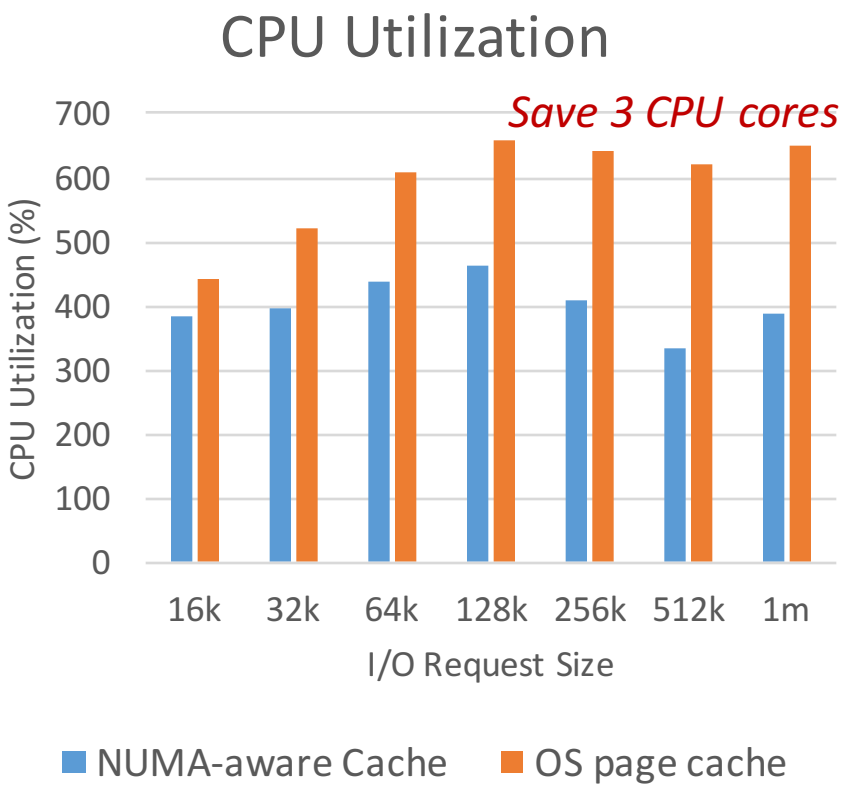
Latency (from Initiator perspective)



- Prompt response
- Overhead vs. NUMA-aware Benefit

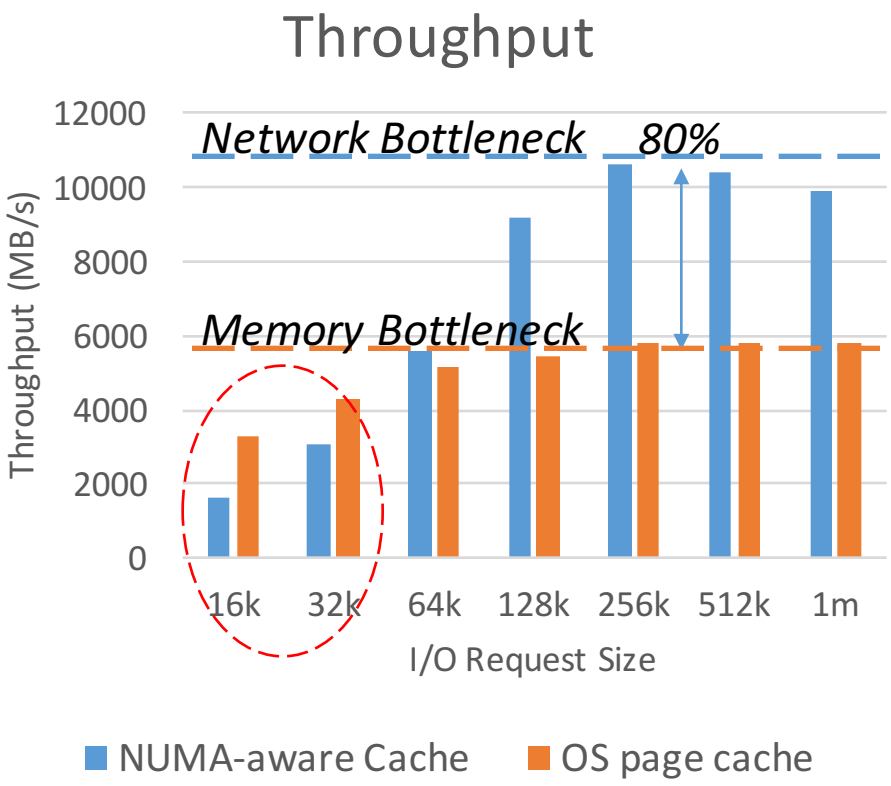
Throughput Comparison

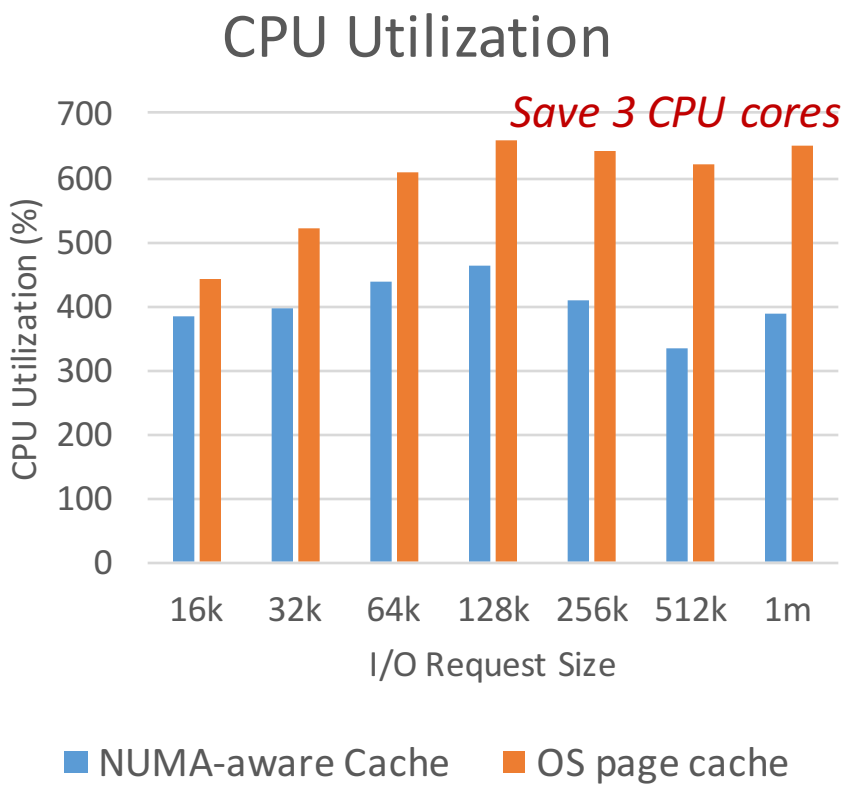




- Concurrent requests serving
- Aggregate throughput

Throughput Comparison

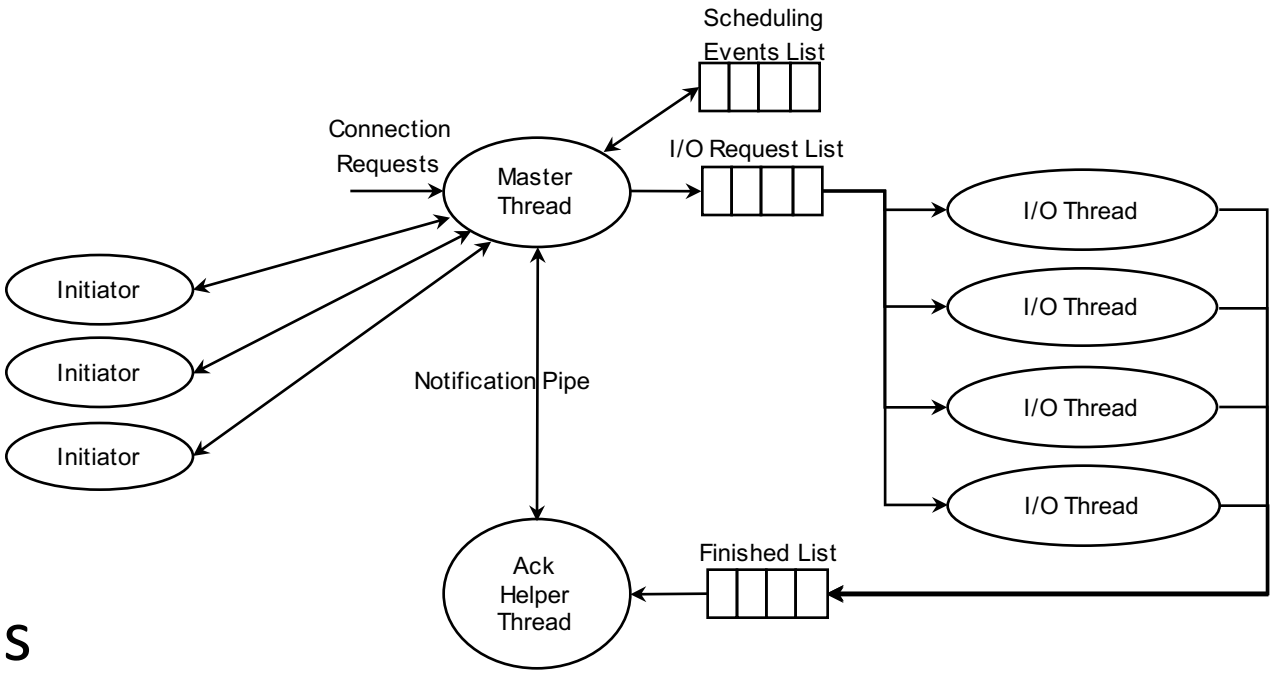




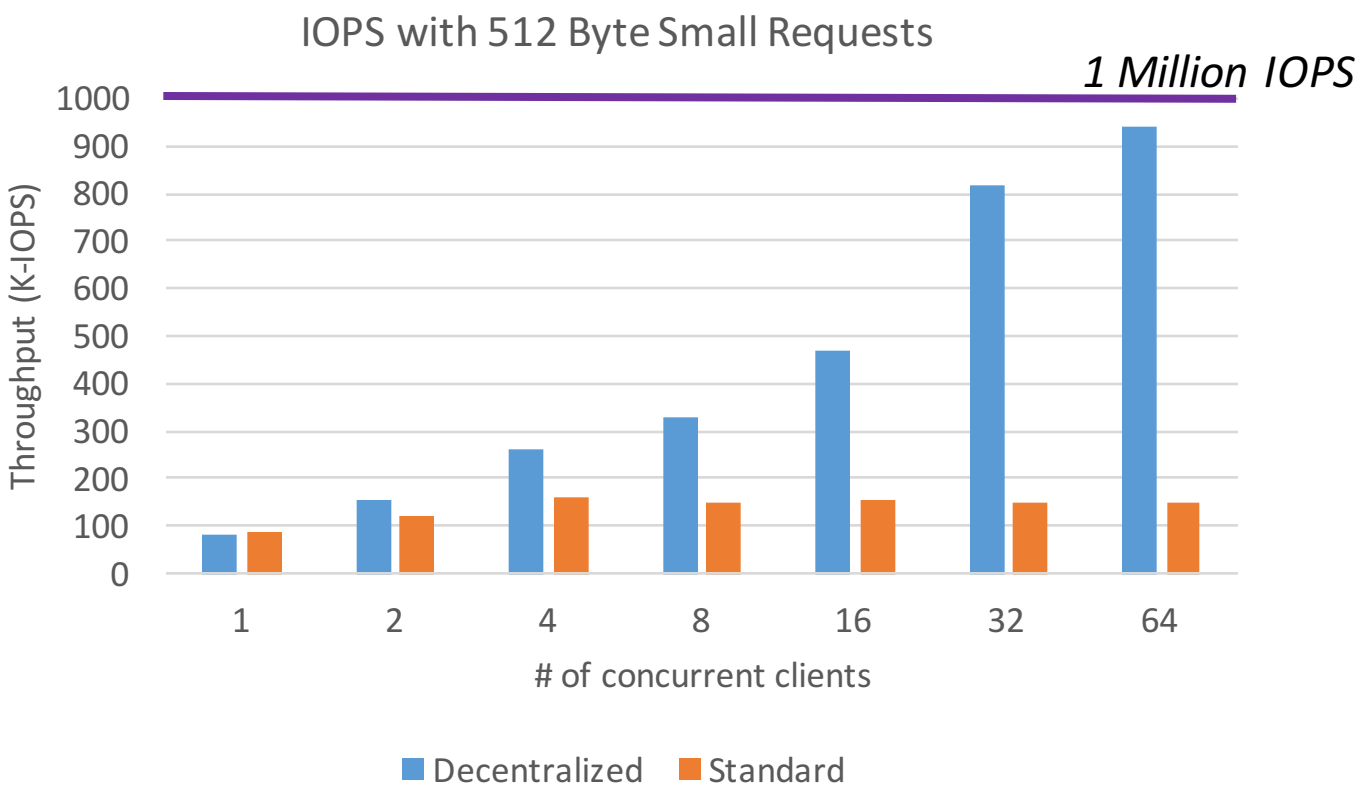
- Master thread becomes the bottleneck

Decentralized Event Processing

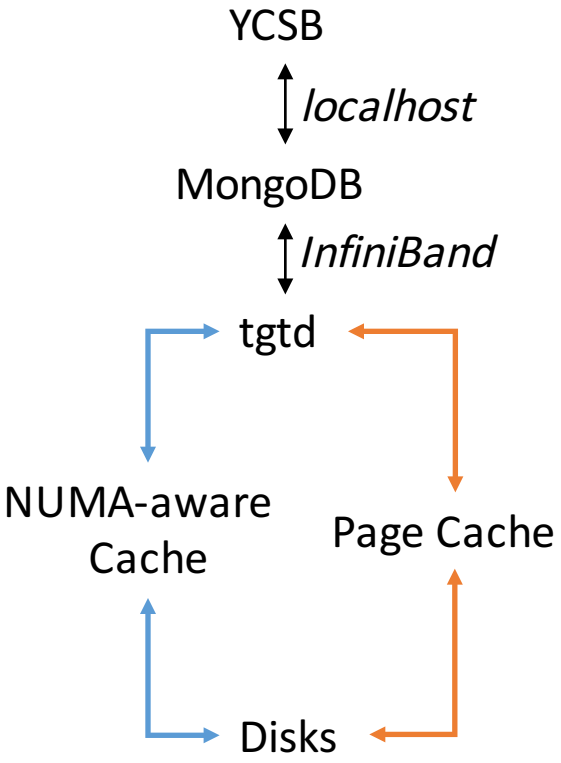
- Standard iSCSI: Centralized event processing with a single thread
- Create multiple threads and distribute event processing
- Processing events is 3X as the number of requests

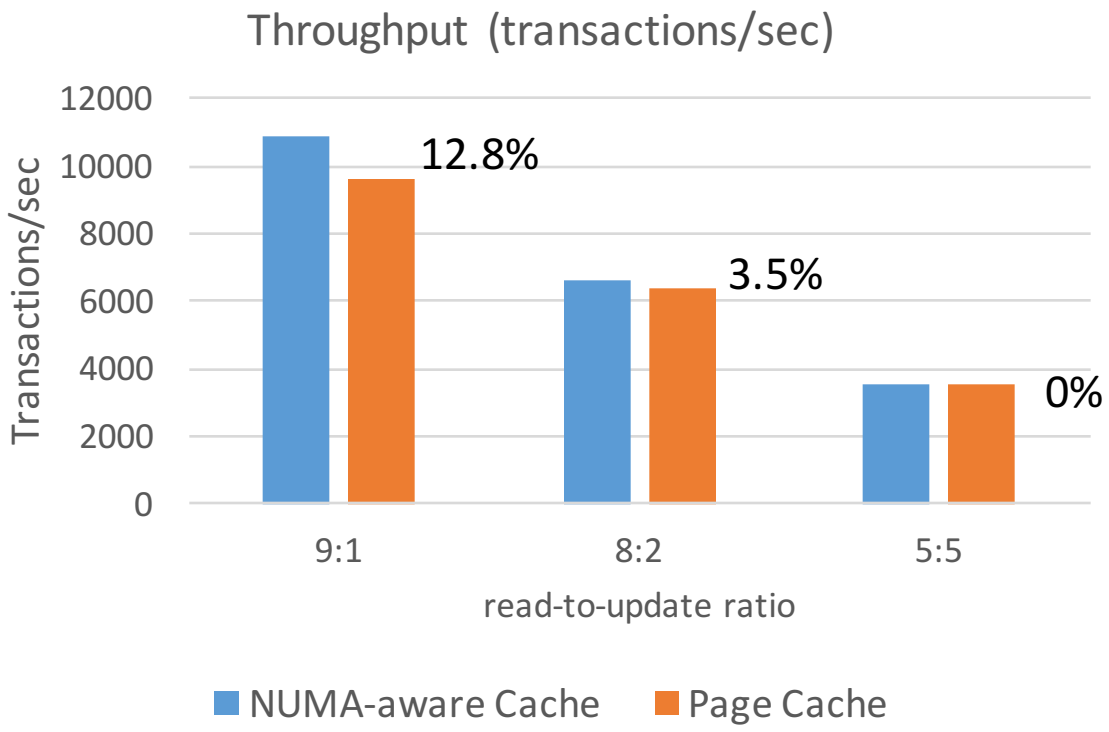


Event Processing Scalability



Real-life workloads



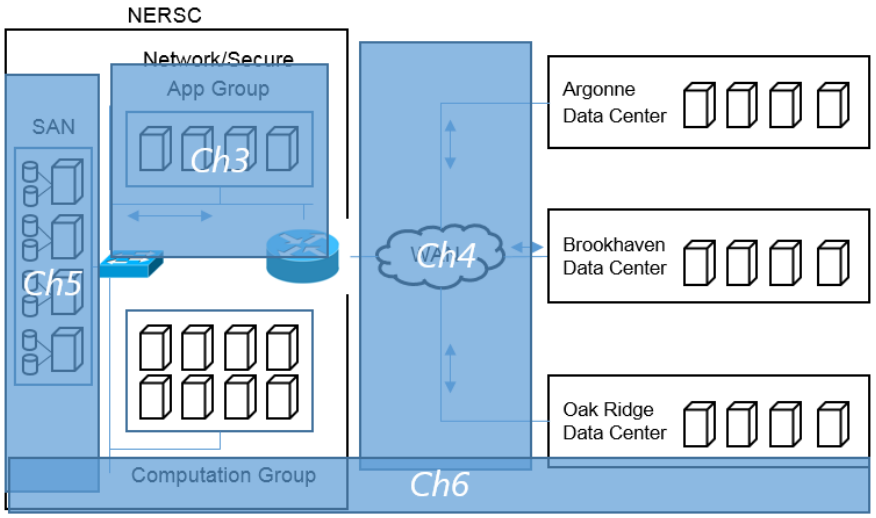


- YCSB – MongoDB – iSER Caching and Storage

Conclusions

Conclusions

Chapter	Contributions	Results
ACES (*ch3*)	Orchestrate resources on memory-centric architecture	Under revision
RFTP (*ch4*)	Scalable zero-copy based protocol for WAN	[SC12, JSS13, HPGC12]
NUMA-aware Cache (*ch5*)	Improve SAN caching data operation locality	[TPDS15]
End-to-End Optimization (*ch6*)	Extract line-speed I/O performance along the end-to-end data path	[SC13]



Software Contributions

- RFTP: RDMA-base FTP Software
  - http://ftp100.cewit.stonybrook.edu/rftp
  - The Three Clause BSD License
  - User: **High Energy Physics at Caltech**
- RDMA benchmark
  - fio: RDMA engine and NUMA module
    - **Chelsio**: http://www.chelsio.com/wp-content/uploads/resources/Lustre-Over-iWARP-vs-IB-FDR.pdf
    - **SanDisk**: http://events.linuxfoundation.org/sites/events/files/eeus13_shelton.pdf
  - iperf-rdma
    - https://github.com/yufeiren/iperf-rdma
    - **China MinSheng Bank**
- NUMA-aware Cache for iSCSI/iSER
  - Linux SCSI Target Framework tgtd
  - https://bitbucket.org/dtyu/tgt

Thank you Q & A