Building Blocks for Scalable Distributed Storage Systems

Ioan Raicu
Computer Science Department, Illinois Institute of Technology
Math and Computer Science Division, Argonne National Laboratory

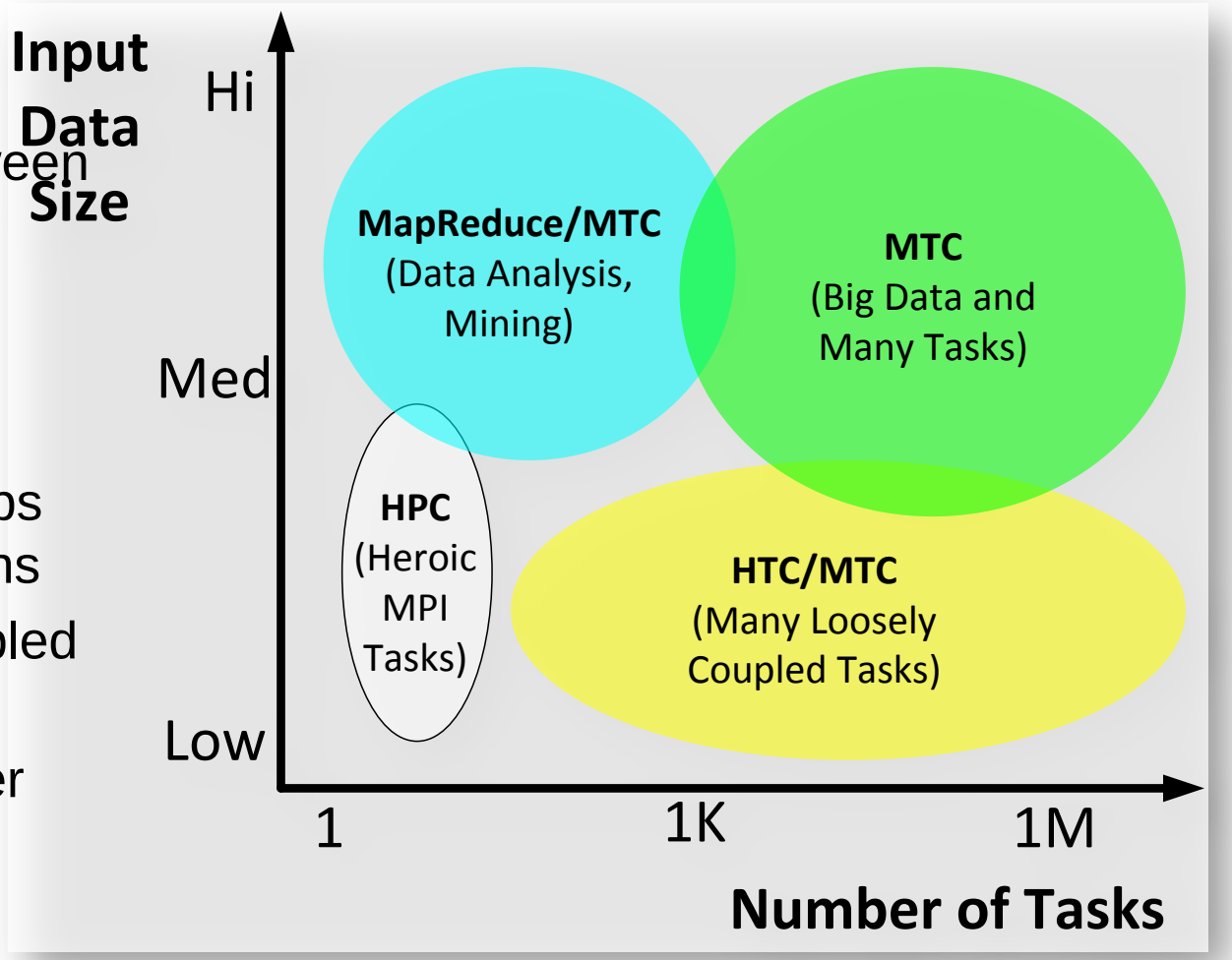
Greater Chicago Area System Research Workshop 2012
May 22nd, 2012

Who am I?

- **Current position:**
  - Assistant Professor at Illinois Institute of Technology (CS)
  - Guest Research Faculty, Argonne National Laboratory (MCS)
- **Education:** PhD, University of Chicago, March 2009
- **Funding/Awards:**
  - NSF CAREER, 2011 – 2015
  - NSF/CRA CIFellows, 2009 – 2010
  - NASA GSRP, 2006 – 2009
- **Over 70+ Collaborators (many here in this room):**
  - **Ian Foster** (UC/ANL), **Rick Stevens** (UC/ANL), Rob Ross (ANL), Marc Snir (UIUC), Arthur Barney Maccabe (ORNL), **Alex Szalay** (JHU), Pete Beckman (ANL), Kamil Iskra (ANL), Mike Wilde (UC/ANL), Douglas Thain (ND), Yong Zhao (UEST), Matei Ripeanu (UBC), Alok Choudhary (NU), Tevfik Kosar (SUNY), Yogesh Simhan (USC), Ewa Deelman (USC), Roger Barga (MSR), Chris Gladwin (Cleversafe), Mike Lang (LANL), Teresa Tung (Accenture), and many more…
- **More info:** http://www.cs.iit.edu/~iraicu/index.html

Best Known For

- **MTC: Many-Task Computing**
  - Bridge the gap between HPC and HTC
  - Applied in clusters, grids, and supercomputers
  - Loosely coupled apps with HPC orientations
  - Many activities coupled by file system ops
  - Many resources over short time periods



[MTAGS08] "Many-Task Computing for Grids and Supercomputers"

Best Known For

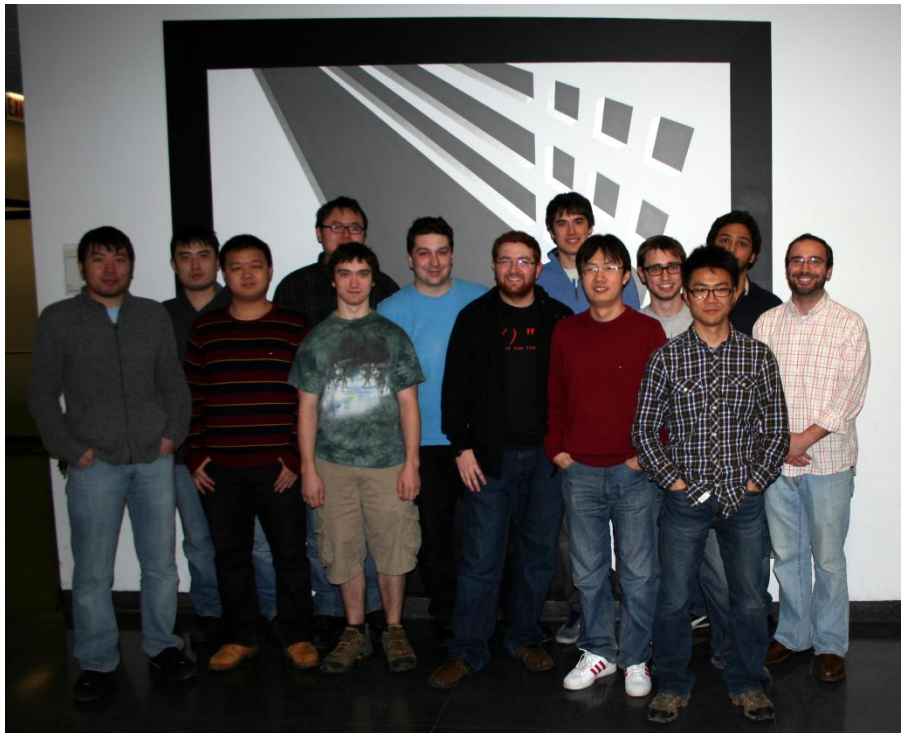
- **Falkon**
  - **Fast and Lightweight T Execution Framework**
  - http://dev.globus.or ubator/Falkon
- **Swift**
  - **Parallel Programming System**
  - http://www.ci.uchica wift/index.php

Field	Description	Characteristics	Status
Astronomy	Creation of montages from many digital images	Many 1-core tasks, much communication, complex dependencies	Experimental
Astronomy	Stacking of cutouts from digital sky surveys	Many 1-core tasks, much communication	Experimental
Biochemistry*	Analysis of mass-spectrometer data for post-translational protein modifications	10,000-100 million jobs for proteomic searches using custom serial codes	In development
Biochemistry*	Protein structure prediction using iterative fixing algorithm; exploring other biomolecular interactions	Hundreds to thousands of 1- to 1,000-core simulations and data analysis	Operational
Biochemistry*	Identification of drug targets via computational docking/screening	Up to 1 million 1-core docking operations	Operational
Bioinformatics*	Metagenome modeling	Thousands of 1-core integer programming problems	In development
Business economics	Mining of large text corpora to study media bias	Analysis and comparison of over 70 million text files of news articles	In development
Climate science	Ensemble climate model runs and analysis of output data	Tens to hundreds of 100- to 1,000-core simulations	Experimental
Economics*	Generation of response surfaces for various eco-nomic models	1,000 to 1 million 1-core runs (10,000 typical), then data analysis	Operational
Neuroscience*	Analysis of functional MRI datasets	Comparison of images; connectivity analysis with structural equation modeling, 100,000+ tasks	Operational
Radiology	Training of computer-aided diagnosis algorithms	Comparison of images; many tasks, much communication	In development
Radiology	Image processing and brain mapping for neuro-surgical planning research	Execution of MPI application in parallel	In development

Note: Asterisks indicate applications being run on Argonne National Laboratory's Blue Gene/P (Intrepid) and/or the TeraGrid Sun Constellation at the University of Texas at Austin (Ranger).

DataSys: Data-Intensive Distributed Systems Laboratory

- **Research Focus**
  - Emphasize designing, implementing, and evaluating systems, protocols, and middleware with the goal of supporting ***data-intensive applications on extreme scale distributed systems***, from **many-core systems, clusters, grids, clouds, and supercomputers**
- **People**
  - 1 Faculty Member
  - 5 PhD Students
  - 6 MS Students
  - 2 UG Students
  - 2 HS Students (over the summer)
  - Alumni: 5 MS, 1 UG
- **Contact**
  - http://datasys.cs.iit.edu/

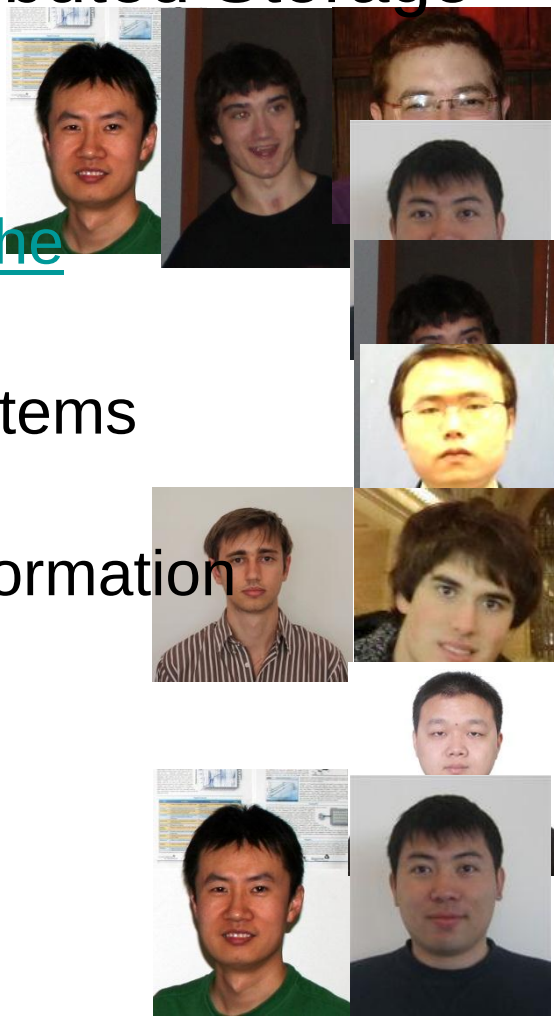


Other Faculty at IIT in Distributed Systems

- Xian-He Sun
  - HPC/architecture
- Zhiling Lan
  - HPC/Reliability
- Shangping Ren
  - Real-time Systems
- Ioan Raicu
  - MTC/HPC/Clouds

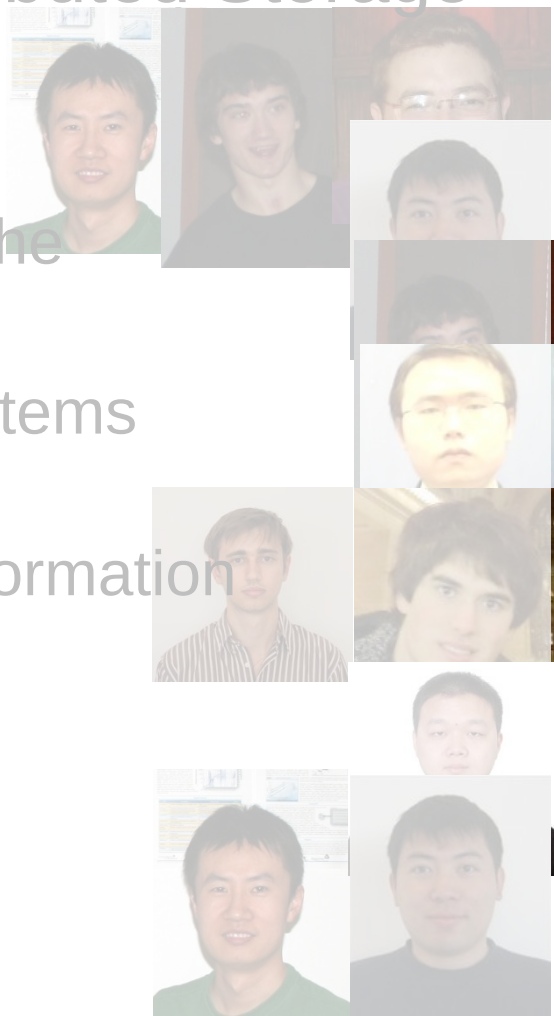
What this talk IS about

- Building Blocks for Large-Scale Distributed Storage Systems
  - Distributed Hash Tables ➔ ZHT
  - Hybrid SSD+HHD file systems ➔ HyCache
  - Persistent Key/Value Stores ➔ NoVoHT
  - Provenance Enabled Distributed File Systems ➔ PAFS
  - Increasing Storage Efficiency through Information Dispersal Algorithms
  - Reliability/Checkpointing ➔ SimHEC
- Long Term Goal:
  - Distributed File Systems ➔ FusionFS



What this talk IS about

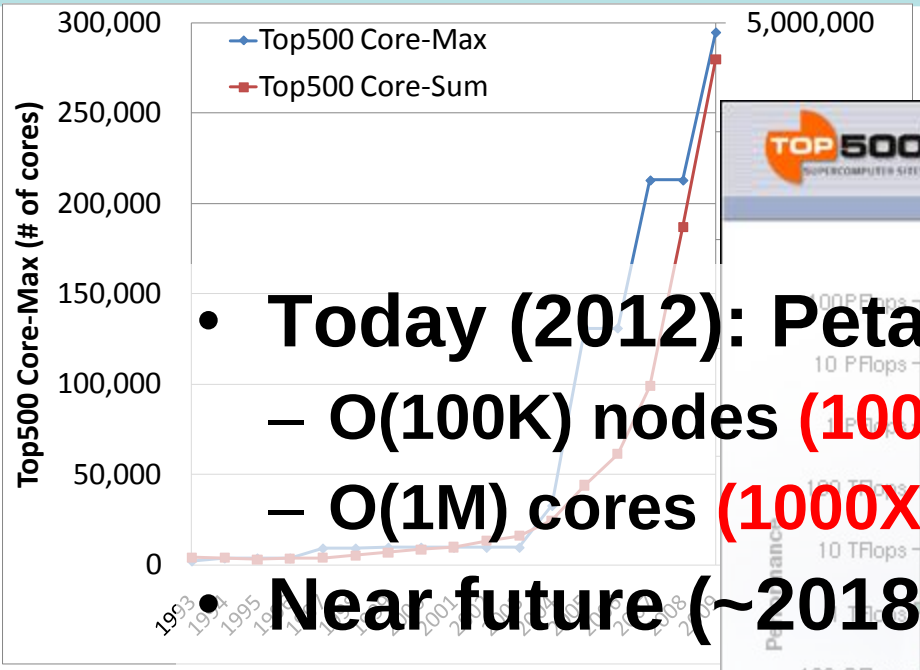
- Building Blocks for Large-Scale Distributed Storage Systems
  - Distributed Hash Tables ➔ ZHT
  - Hybrid SSD+HHD file systems ➔ HyCache
  - Persistent Key/Value Stores ➔ NoVoHT
  - Provenance Enabled Distributed File Systems ➔ PAFS
  - Increasing Storage Efficiency through Information Dispersal Algorithms
  - Reliability/Checkpointing ➔ SimHEC
- Long Term Goal:
  - Distributed File Systems ➔ FusionFS

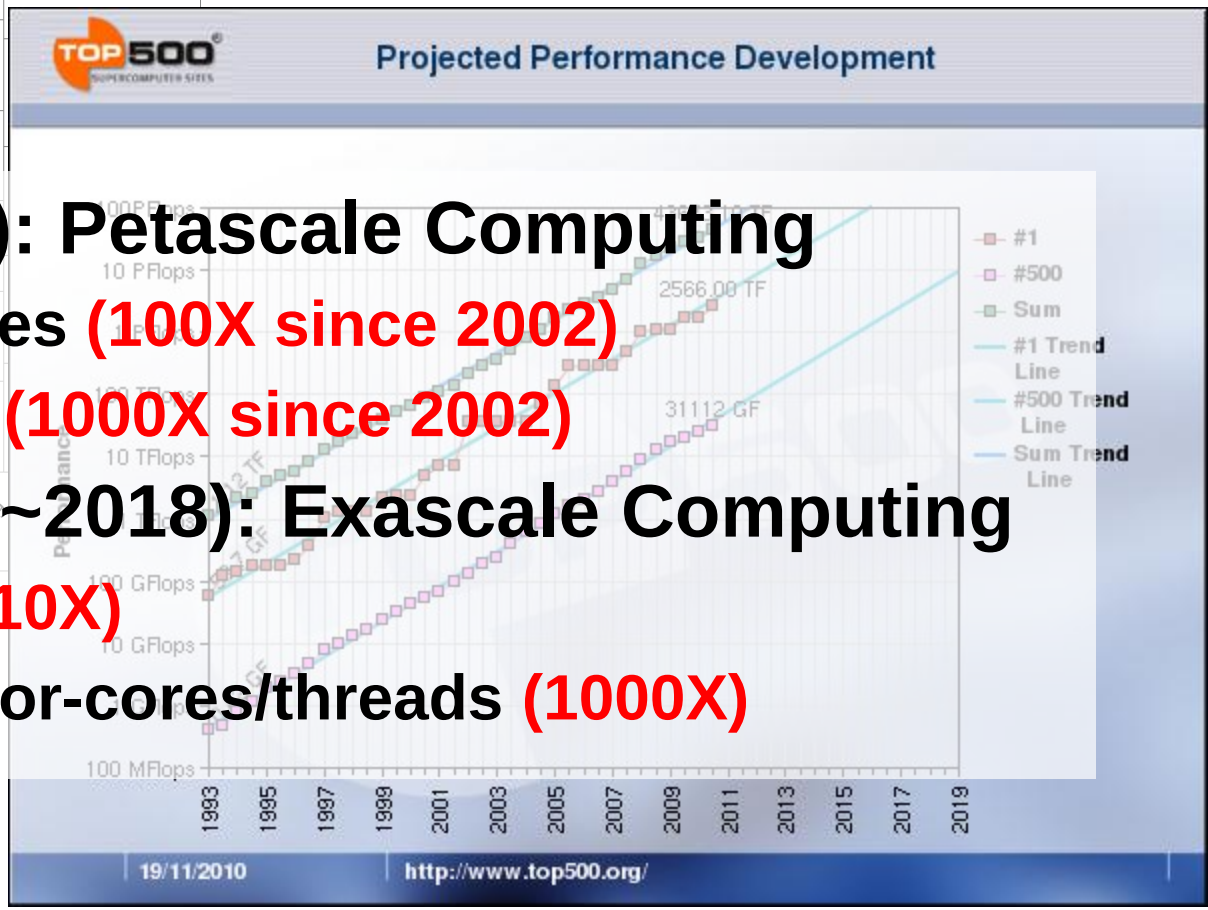


More Projects in the DataSys Lab NOT covered in this talk

- Compute Resource Management Systems SimMatrix & MATRIX
- Scheduling Algorithms ➔ Work Stealing, DAG Scheduling
- GPGPU Computing ➔ vGPGPU
- Cloud Computing ➔ Understanding the Cost of Clouds
- Mobile Computing ➔ CiteSearcher

Motivation
Exascale Computing



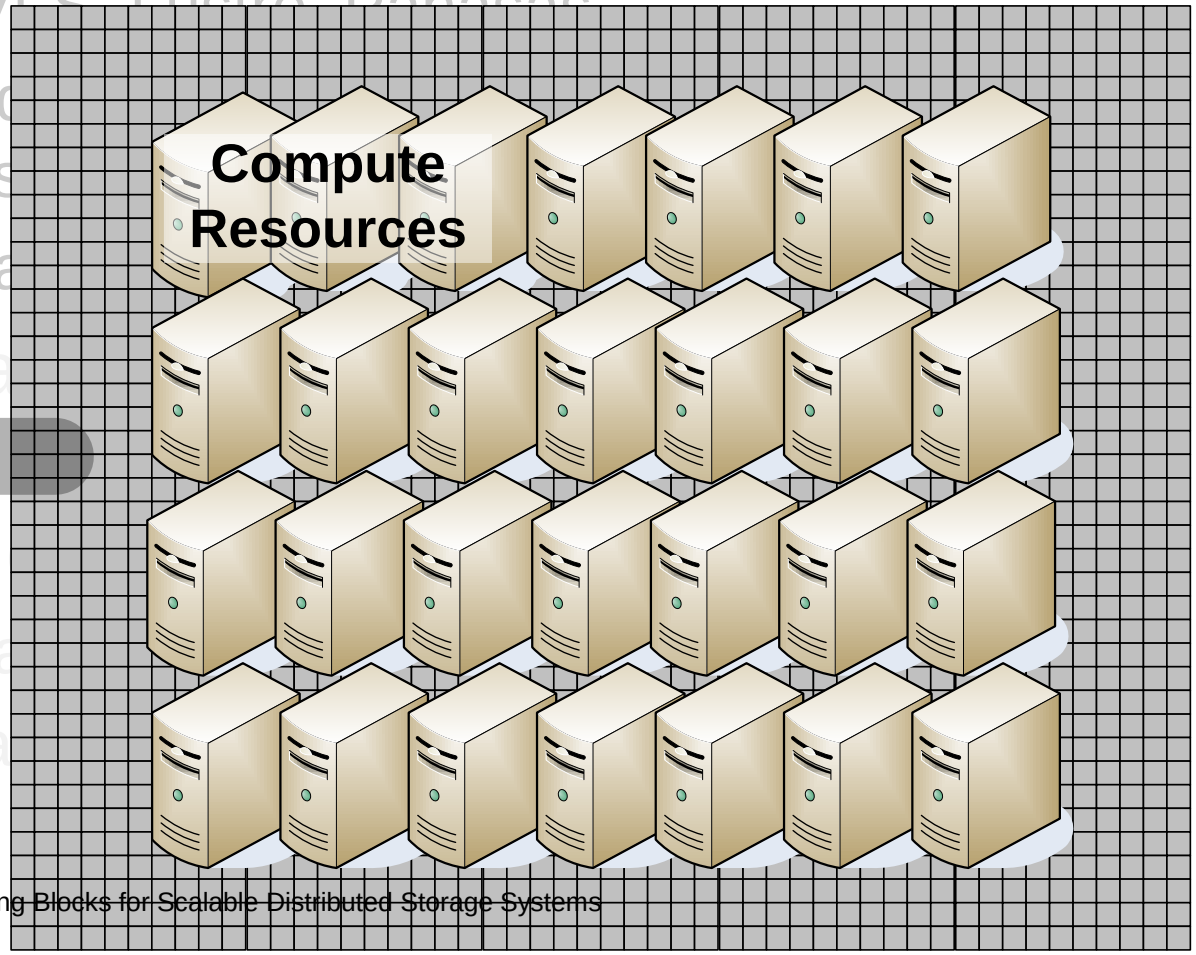


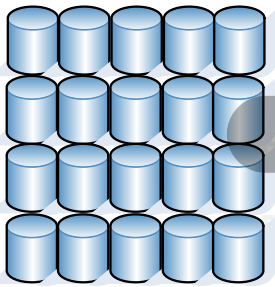
- **Today (2012): Petascale Computing**
  - **O(100K) nodes (100X since 2002)**
  - **O(1M) cores (1000X since 2002)**
- **Near future (~2018): Exascale Computing**
  - **~1M nodes (10X)**
  - **~1B processor-cores/threads (1000X)**

Top500 Projected Development,
http://www.top500.org/lists/2010/11/performance_development

Motivation
Parallel File Systems

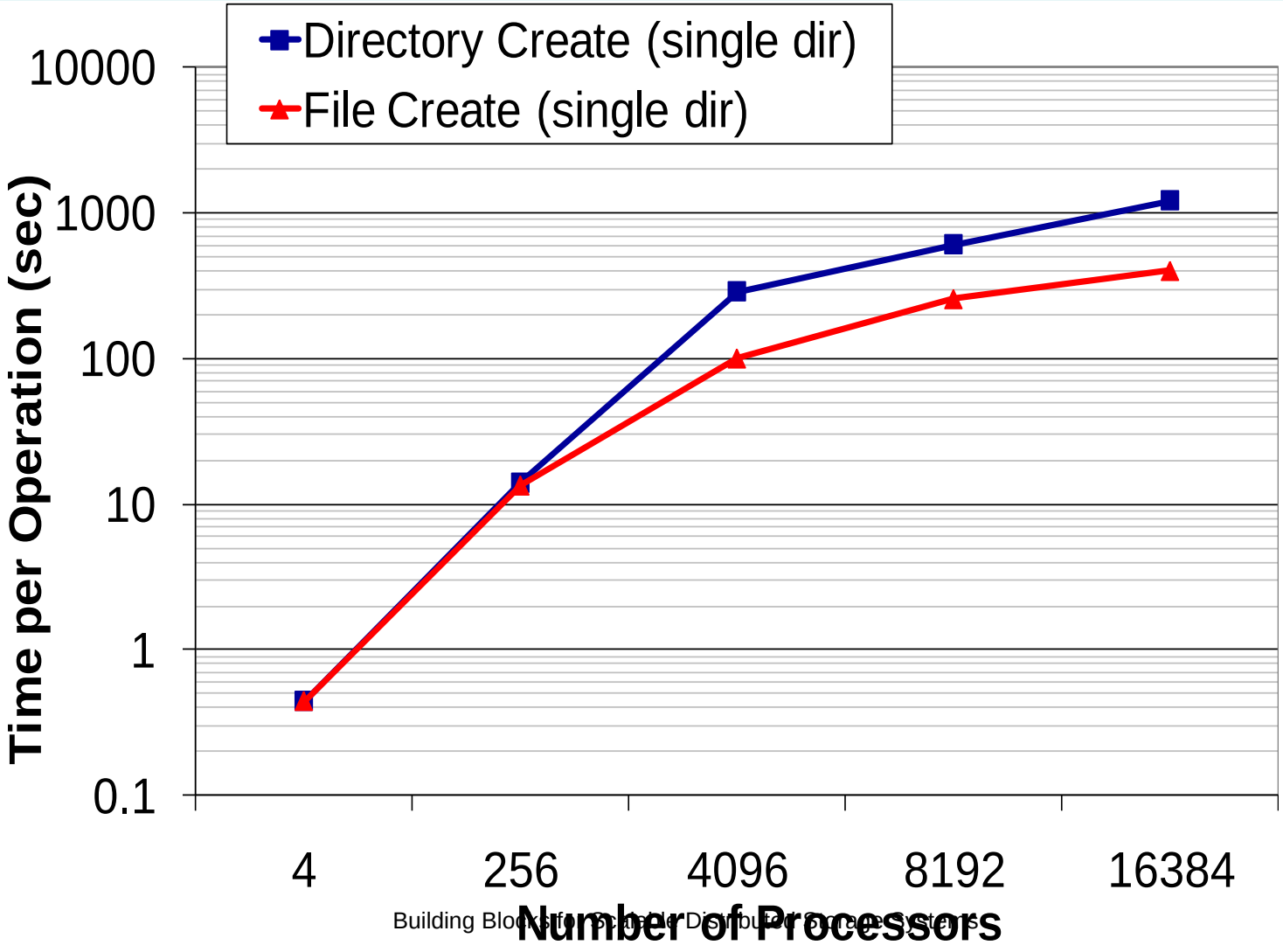
- Segregated storage and compute
  - NFS, GPFS, PVFS, Lustre, Panasas
  - Batch-scheduled
    Supercomputers
  - Programming pa
- located stora
  ata centers at
  - Programming pa
  - Others from aca





Network Link(s)

Motivation
Poor Meta-data Scalability on GPFS



Distributed Meta-data Management

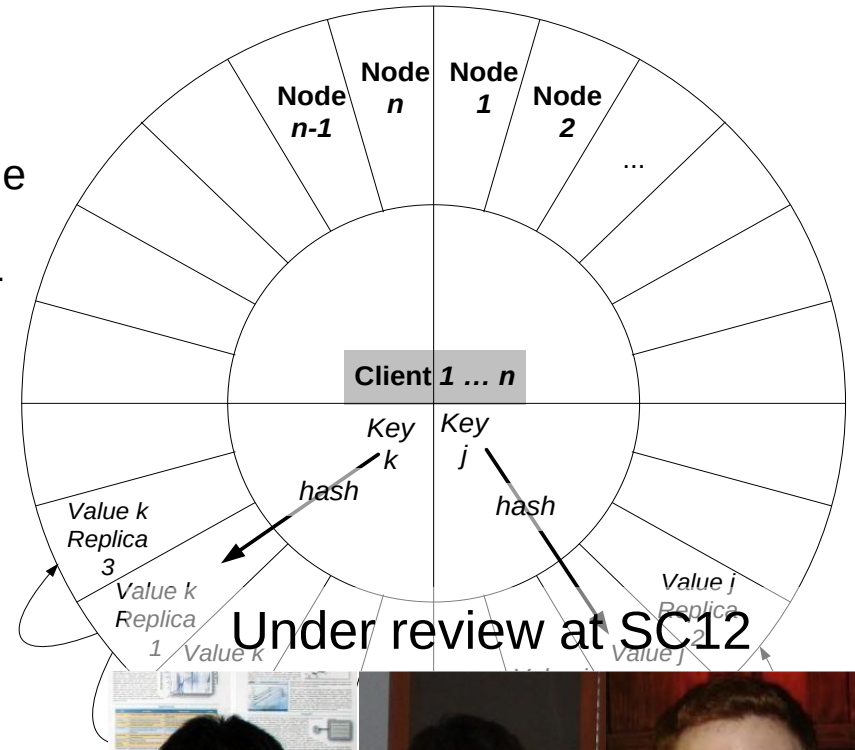
- Leverage distributed hash tables (DHT) to implement distributed meta-data management
- Existing DHT: old, slow, multi-hop
  - Chord, Kademlia, Pastry, Tapestry
- Amazon Dynamo: commercial use only, not open.

Assumptions of High-End Computing System

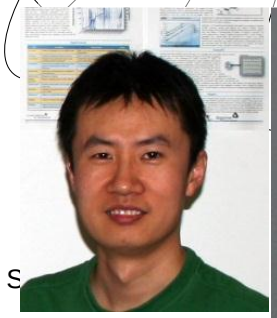
- Reliable hardware
- Fast network interconnects
- Non-existent node “churn”
- Batch oriented: steady amount of resource

ZHT:
Zero Hop Distributed Hash Table

- Simplified distributed hash table tuned for the specific requirements of HEC
- Emphasized key features of HEC are:
  - Trustworthy/reliable hardware, fast network interconnects, non-existent node "churn", the requirement for low latencies, and scientific computing data-access patterns
- Primary goals:
  - Excellent availability and fault tolerance, with low latencies
- ZHT details:
  - Static/Dynamic membership function
  - Network topology aware node ID space
  - Replication and Caching
  - Efficient 1-to-all communication through spanning trees
  - Persistence



Under review at SC12



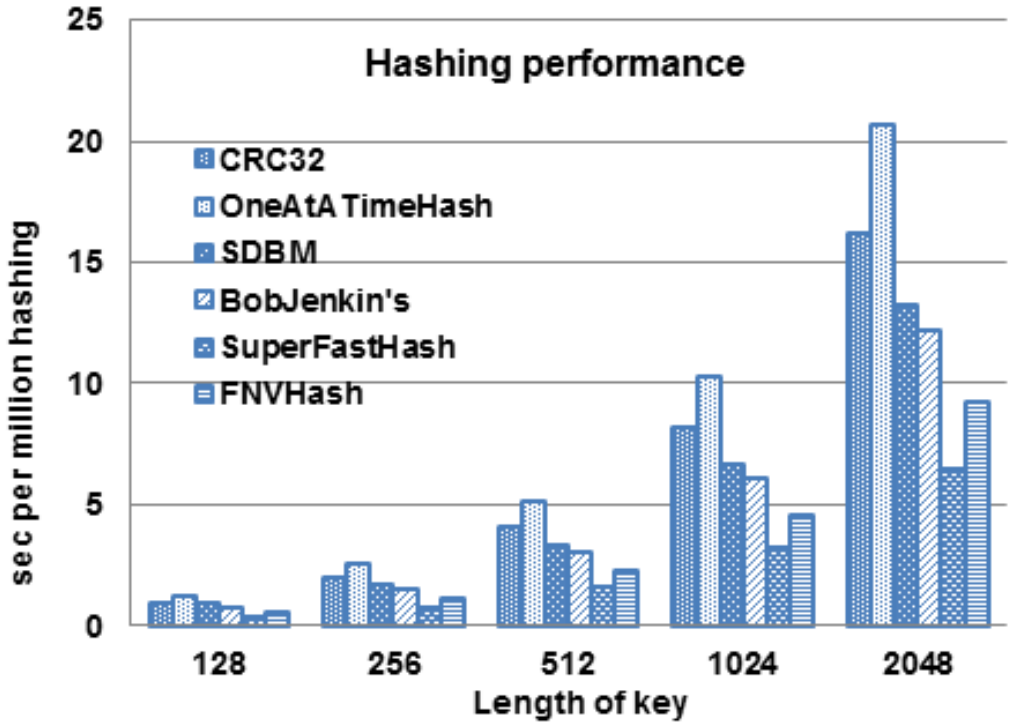
ZHT Prototype Implementation

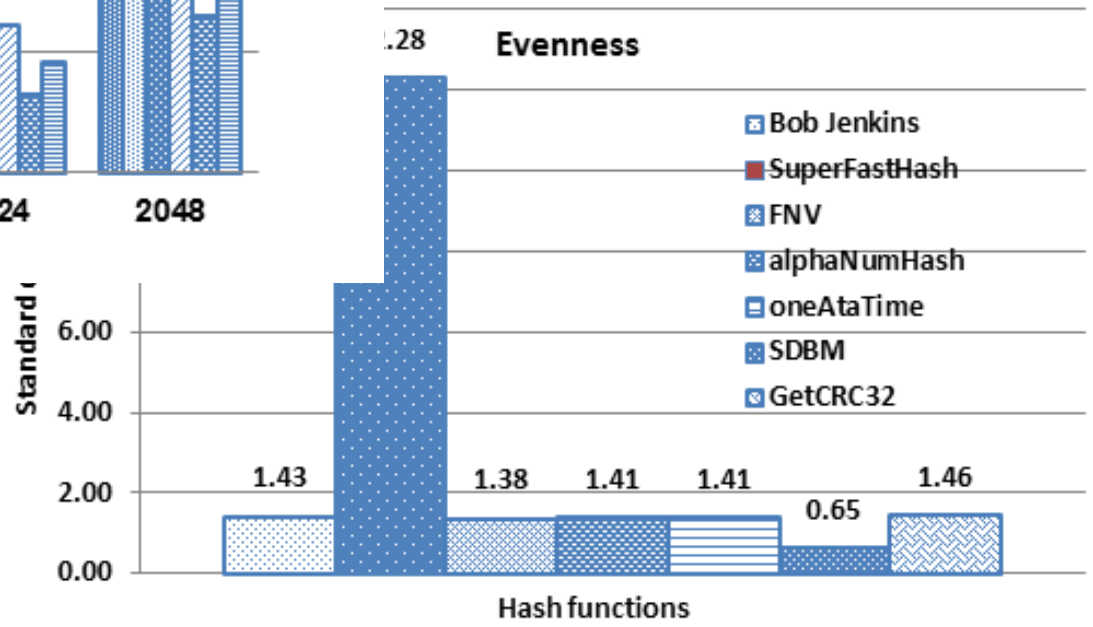
- C++/Linux
- Simple API
  - Insert, Find, Remove
- Communication
  - TCP & UDP, connection caching
  - Evaluating MPI & BMI
- Hashing functions
  - SuperFastHash, FNVHash, alphaNumHash, BobJenkins, SDBM, CRC32, OneAtATimeHash
- Architecture
  - Multi-threading, epoll
- Persistence
  - NoVoHT
- Leverages other work
  - Google Buffer

Testbeds

- Majority of experiments: IBM BlueGene/P
  - 1024 nodes
  - 2GB RAM/node
  - 4096 cores
  - OS: ZeptOS
  - Batch execution system: Cobalt
- Other testbeds
  - 64-node Linux cluster
  - 972-node SiCortex SC5832
  - 300-instances on Amazon EC2

Hash functions



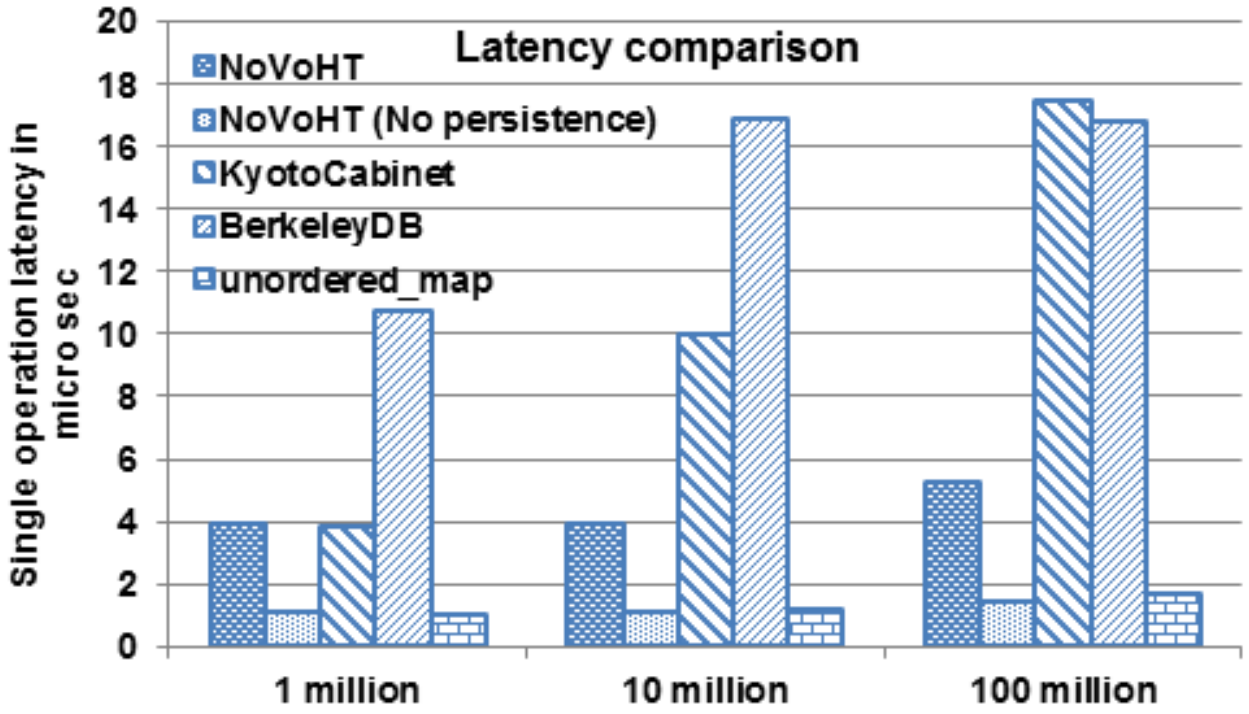


Data object

- Store complex data object
- Key-Value store accept plain string only
- Serialize/deserialize data structure
  - Boost: big, more dependencies
  - ACE: full-featured, but huge, heavy, complicated, failed to install on SiCortex
  - **Google protocol buffers: simple, light-weighted, easy to install**

Persistency

- Kyotocabinet: poor garbage collection
- BerkeleyDB: slow
- MongoDB: slow, complex
- MySQL: re
- Conclusior
  - **NoVoHT:**

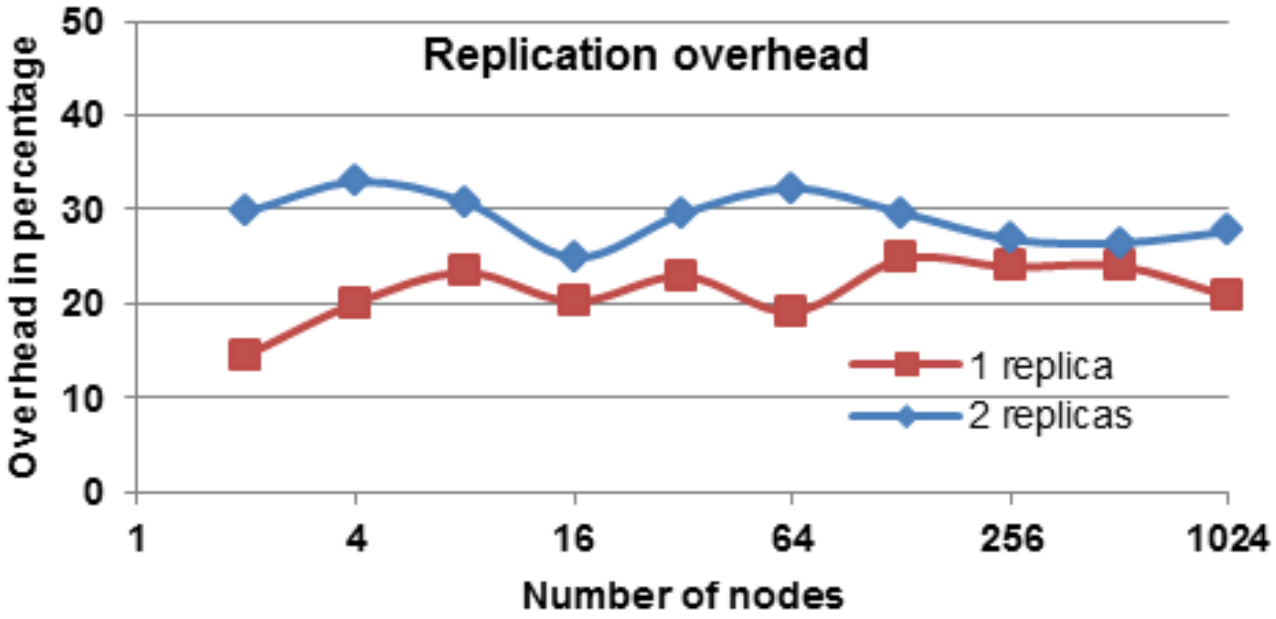


Failure handling

- Insert
  - If one try failed: send it to next replica
  - Mark this record as primary copy
- Lookup
  - If one try fail: try next one, until go through all replicas
- Remove
  - Lazy - mark record removed

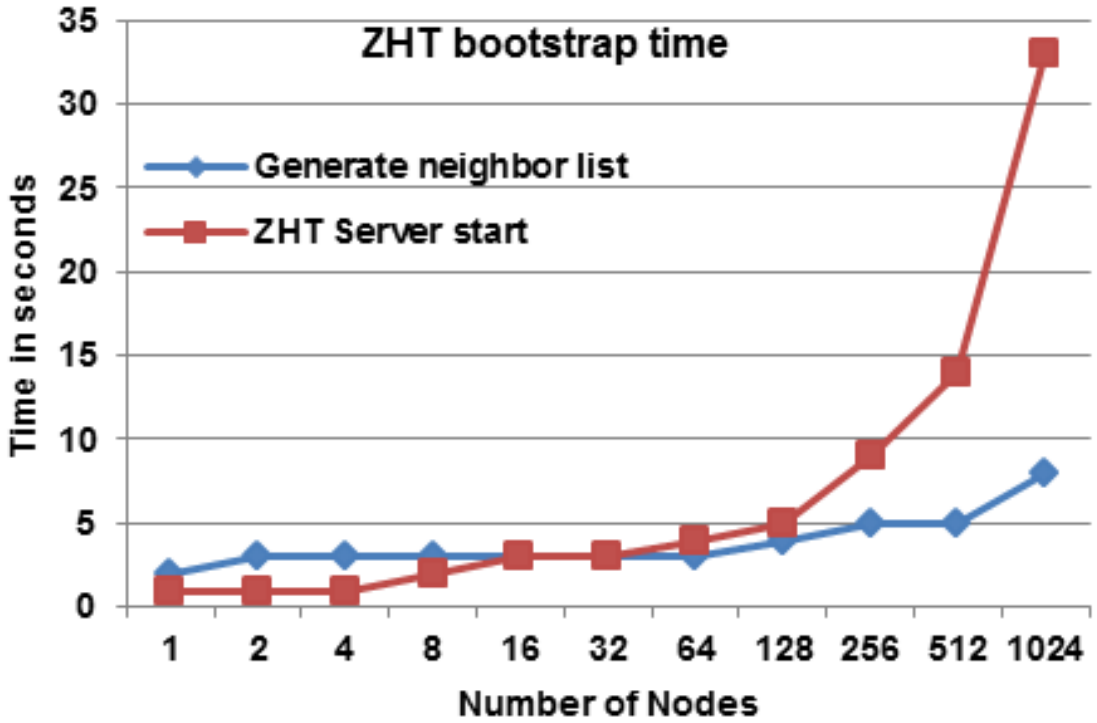
Replication

- Allow key/value pair replication
- Client side
  - Another thread from client deal with replication operation asynchronously
- Server si
  - After se
    replicati

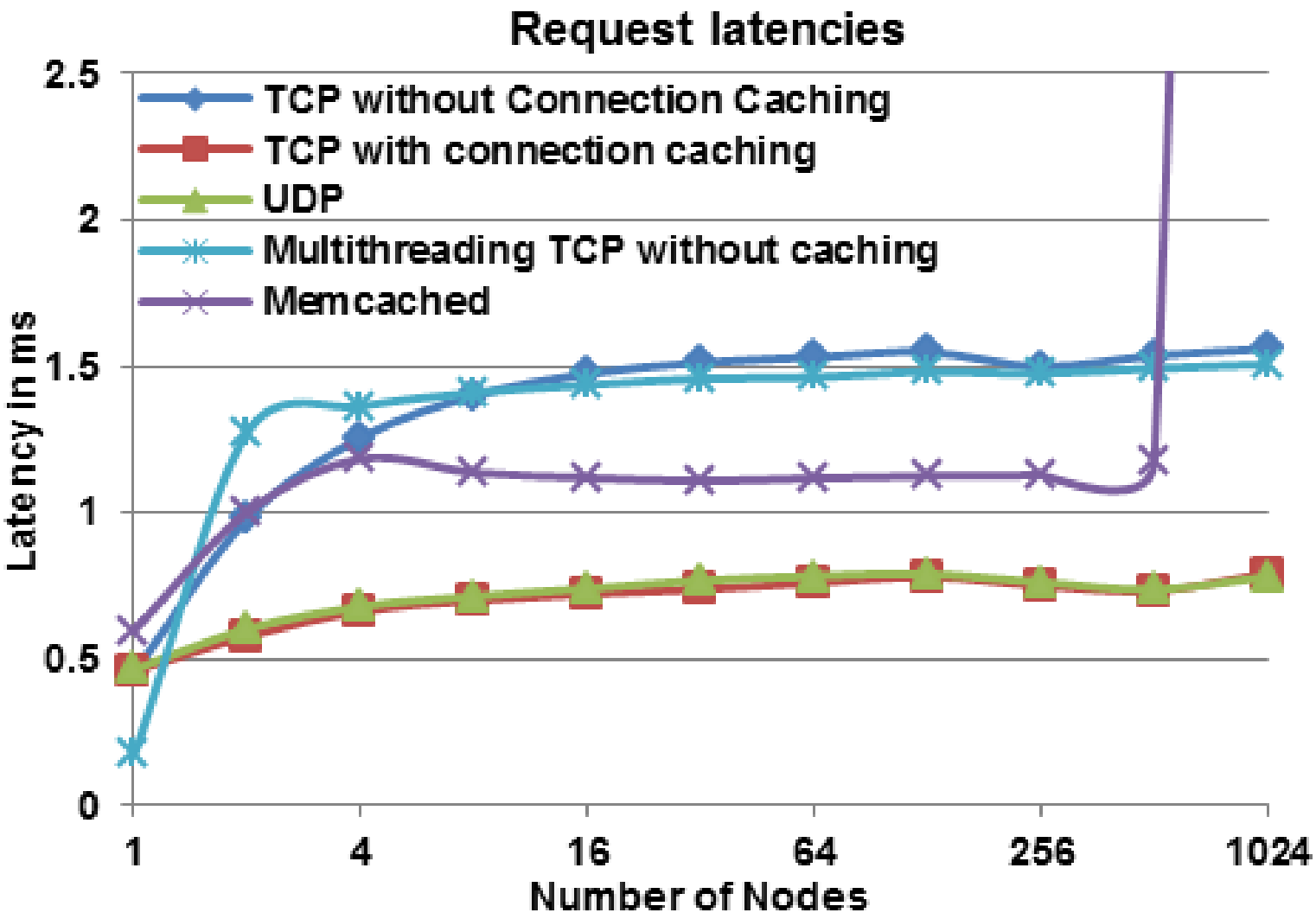


Membership management

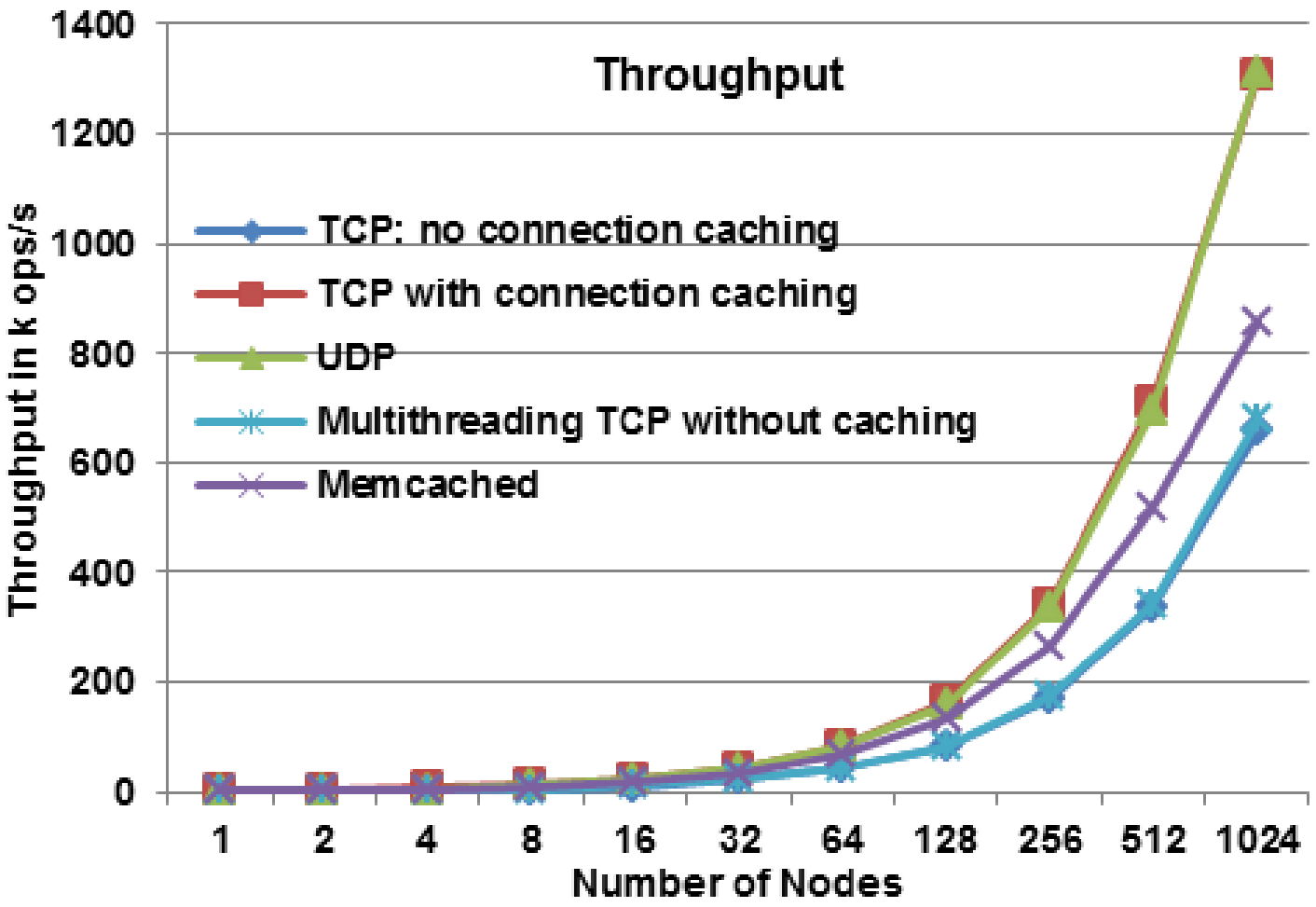
- Static member list
  - reliable hardware
  - non-existent node “churn”
  - Membership es
- If a node fails, it
recover
  - Remove failed



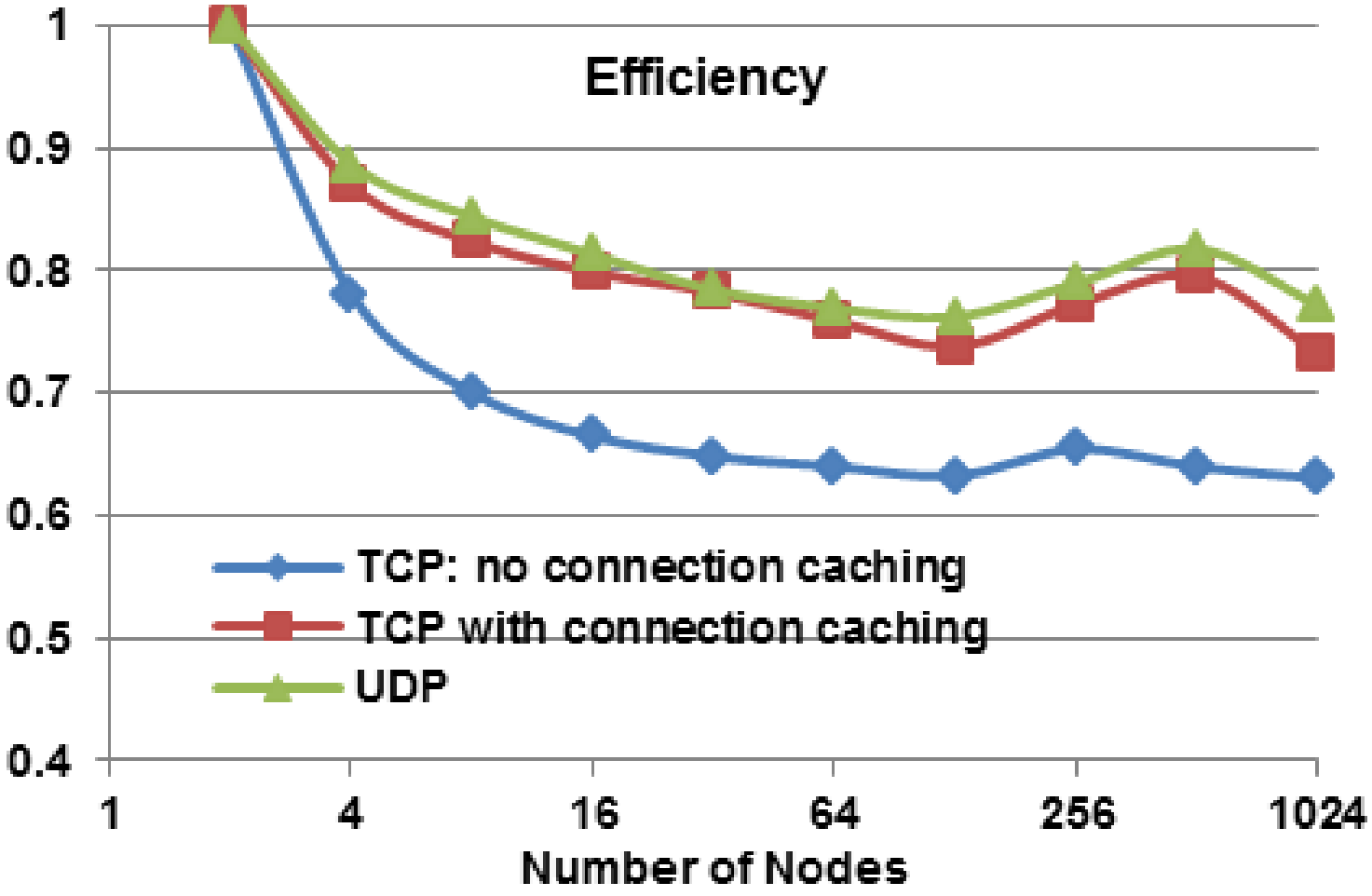
Latency


Request latencies
TCP without Connection Caching
TCP with connection caching
UDP
Multithreading TCP without caching
Memcached
Latency in ms
2.5
2
1.5
1
0.5
0
1
4
16
64
256
1024
Number of Nodes

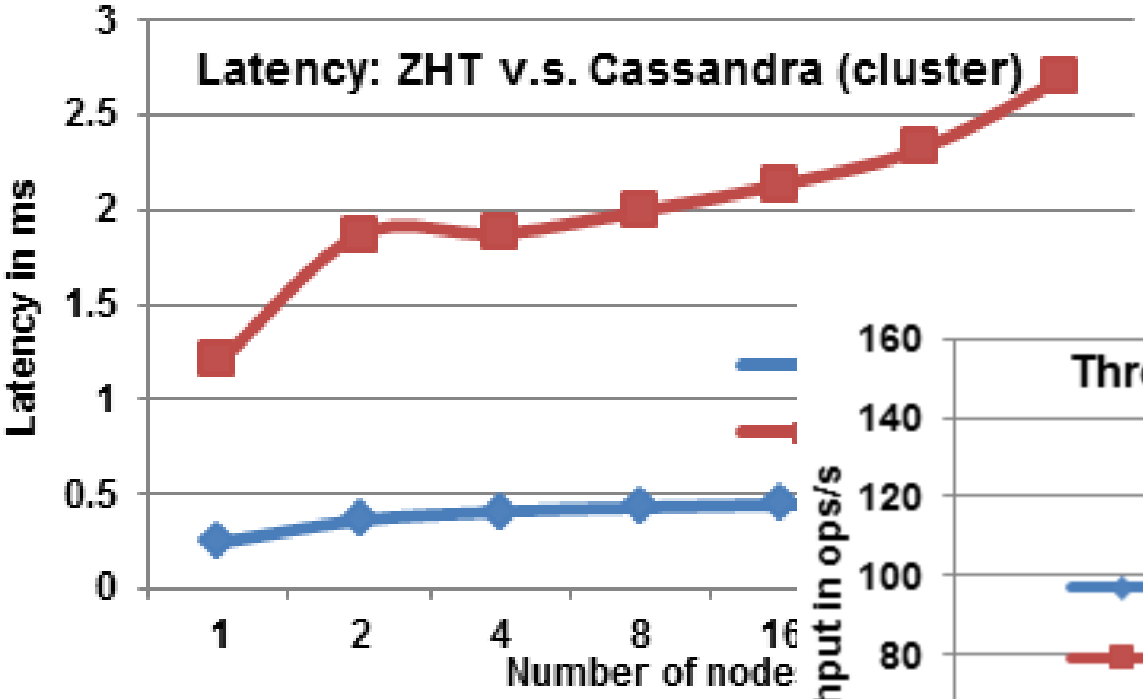
Throughput

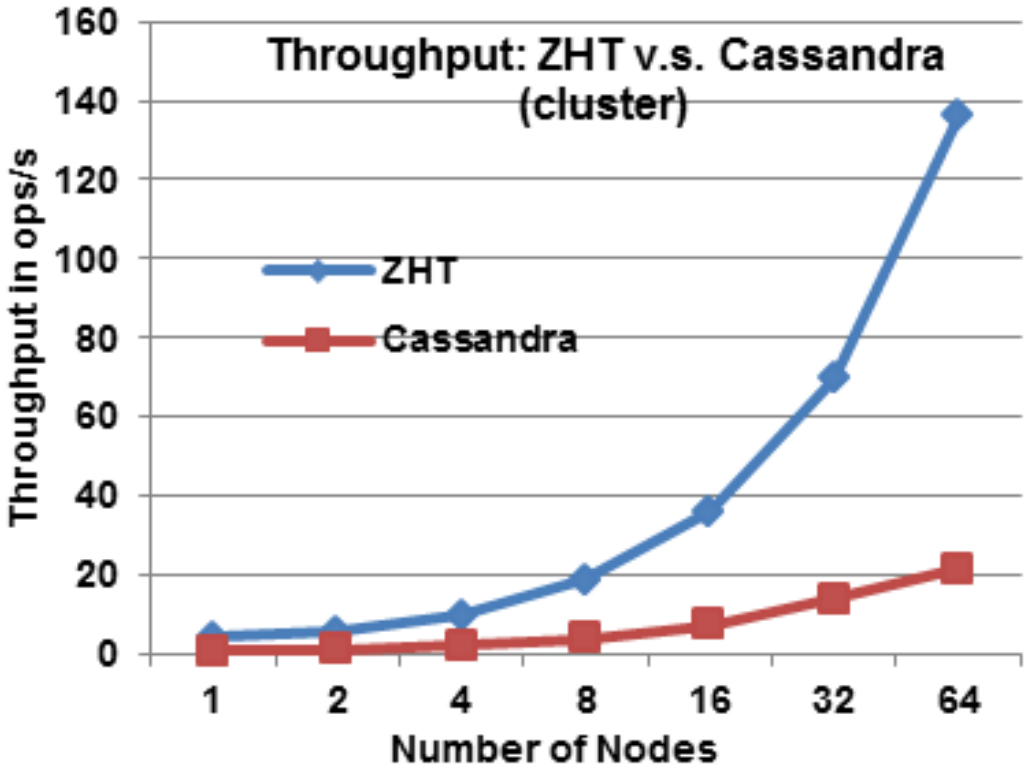

Throughput
1400
1200
1000
800
600
400
200
0
Throughput in k ops/s
TCP: no connection caching
TCP with connection caching
UDP
Multithreading TCP without caching
Memcached
1
2
4
8
16
32
64
128
256
512
1024
Number of Nodes

Efficiency


Efficiency
1
0.9
0.8
0.7
0.6
0.5
0.4
TCP: no connection caching
TCP with connection caching
UDP
1
4
16
64
256
1024
Number of Nodes

ZHT V.S Cassandra



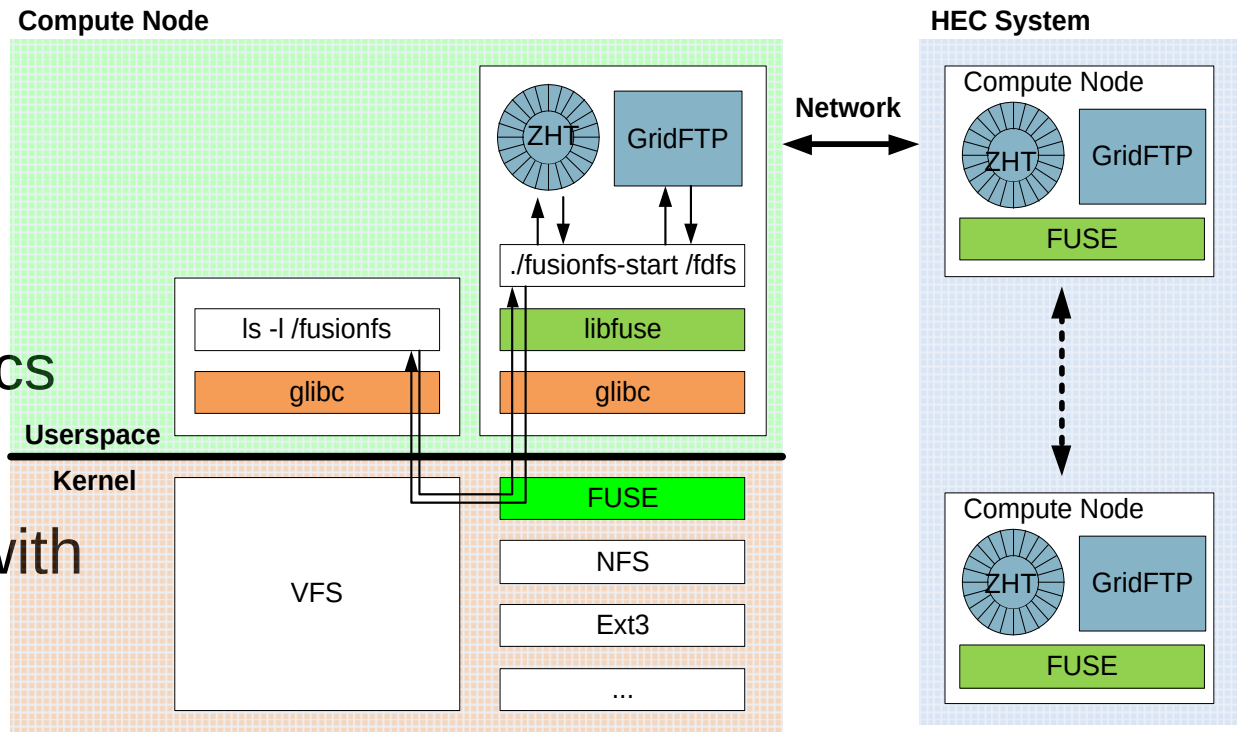


Conclusion

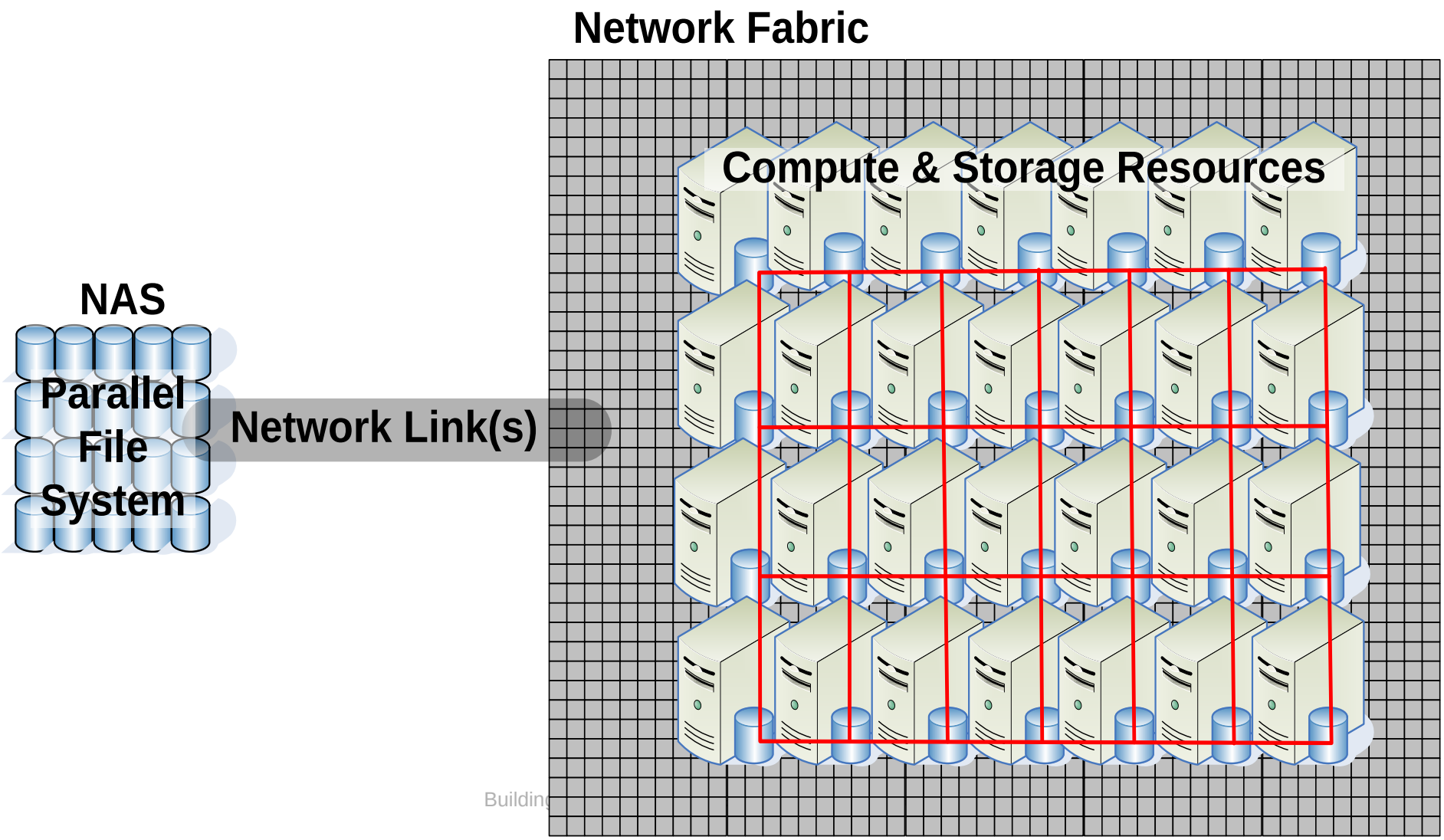
- ZHT is a distributed Key-Value store
  - Light-weighted
  - Scalable
  - High performance
  - Low latency
  - Few dependency
  - Wide range of use

FusionFS Details

- Distributed Metadata Management
- Distributed Data Management
- Data Indexing
- Relaxed Semantics
- Data Locality
- Overlapping I/O with Computations
- POSIX



Storage System Architecture



Main Message

- ***Preserving locality is critical!***
- *Segregating storage from compute resources is **BAD***
- *Parallel file systems + distributed file systems + distributed hash tables + nonvolatile memory*
  ➔ new storage architecture for extreme-scale HEC
- *Co-locating storage and compute is **GOOD***
  - *Leverage the abundance of processing power, bisection bandwidth, and local I/O*

Teaching

- Master Of Computer Science With a Specialization in Distributed and Cloud Computing
- Bachelor of Science in Computer Science with a Specialization in Distributed and Cloud Computing
- Courses
  - Introduction to Distributed Systems (CS495)
  - Advanced Operating Systems (CS550)
  - Cloud Computing (CS553)
  - Data-Intensive Computing (CS594)

Recent Workshops and Journals

- **IEEE MTAGS 2011:** 3rd IEEE Workshop on Many-Task Computing on Grids and Supercomputers, co-located with IEEE/ACM Supercomputing 2010, November 15th, 2010
  - http://datasys.cs.iit.edu/events/MTAGS10/
- **IEEE DataCloud 2011**: 1st Workshop on Data Intensive Computing in the Clouds, co-located with IEEE IPDPS 2011, May 16th, 2011
  - http://www.cse.buffalo.edu/faculty/tkosar/datacloud2011/
- **ACM ScienceCloud 2011:** 2nd Workshop on Scientific Cloud Computing, co-located with ACM HPDC 2011, June 8th, 2011
  - http://datasys.cs.iit.edu/events/ScienceCloud2011/
- **Scientific Programming Journal**, Special Issue on Science-driven Cloud Computing, Volume 19, Number 2-3 / 2011
  - SI: http://datasys.cs.iit.edu/events/SPJ_ScienceCloud_2011/
  - Table of Contents: http://iospress.metapress.com/content/n561462255r3/
  - Editorial: http://iospress.metapress.com/content/d421756381083576/fulltext.pdf
- **IEEE Transactions on Parallel and Distributed Systems**, Special Issue on Many-Task Computing, June 2011; vol. 22 no. 6
  - SI: http://datasys.cs.iit.edu/events/TPDS_MTC/
  - Table of Contents: http://www.computer.org/portal/web/csdl/abs/trans/td/2011/06/ttd201106toc.htm
  - Editorial: http://www.computer.org/portal/web/csdl/abs/html/trans/td/2011/06/ttd2011060897.htm
- **Springer Journal of Grid Computing**, Special Issue on Data Intensive Computing in the Clouds, April 2012
  - SI: http://datasys.cs.iit.edu/events/JGC-DataCloud-2012/index.html

Future Events

- **ACM MTAGS 2012:** ACM Workshop on Many-Task Computing on Grids and Supercomputers (co-located with SC12 -- pending)
- **IEEE DataCloud 2012:** 3rd IEEE Workshop on Data Intensive Computing in the Clouds (co-located with SC12 -- pending)
- **ACM ScienceCloud 2012:** 3rd ACM Workshop on Scientific Cloud Computing (will submit to HPDC12)
- **IEEE/ACM SC 2012:** in Salt Lake City, Utah
- **ACM HPDC 2012:** in Delft Netherlands
- **IEEE eScience 2012:** in **Chicago** IL (General Chair: Ian Foster)
- **IEEE/ACM CCGrid 2012:** in Ottawa Canada
- **IEEE/ACM CCGrid 2014:** in **Chicago** IL (General Chairs Xian-He Sun & Ian Foster)

More Information

- More information:
  - http://www.cs.iit.edu/~iraicu/
  - http://datasys.cs.iit.edu/
- Contact:
  - iraicu@cs.iit.edu
- Questions?