BIG DATA SYSTEM INFRASTRUCTURE AT EXTREME SCALES

Ioan Raicu
Illinois Institute of Technology
Argonne National Laboratory
Northwestern University

Northwestern University
May 1st, 2017

WHO AM I?

- **History**
  - 1997-2002: BS/MS in CS at **Wayne State University**; MS thesis in IPv6 Network Protocols under Sherali Zeadally
  - 2003-2009: PhD in CS at **University of Chicago** in Many-Task Computing under Ian Foster
  - 2009-2010: Postdoc at **Northwestern Univ.** with Alok Choudhary
- **Current Affiliations**
  - Visiting Scholar on sabbatical at EECS at **Northwestern University**
  - Associate Professor in CS at **Illinois Institute of Technology**
  - Guest Research Faculty in MCS at **Argonne National Laboratory**
  - Advisory Board Member at **Ocient LLC**

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.

DataSys
Data-Intensive Distributed Systems Laboratory

DATASYS:
DATA-INTENSIVE DISTRIBUTED SYSTEMS LABORATORY

Emphasize ystems,
protocols, data-
intensive stems,
from nd

FUNDING ACKNOWLEDGEMENTS

- Federal:
  - National Science Foundation (CAREER, REU, CCC)
  - Department of Energy (ANL, LANL, FNAL)
  - NASA (ARC)
- Industry:
  - NVIDIA, Mellanox, Intel
- Infrastructure:
  - Amazon AWS, Microsoft Azure, Chameleon, ANL ALCF, XSEDE, Blue Waters

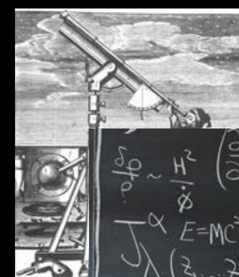
COMPUTER VISIONARIES

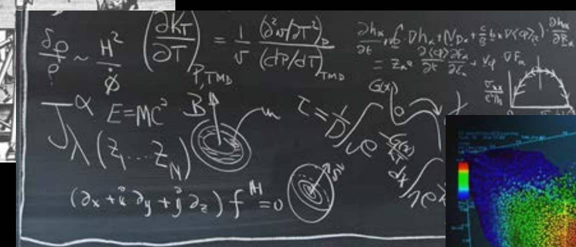
The advent of computation can be compared, in terms of the breadth and depth of its impact on research and scholarship, to the invention of writing and the development of modern mathematics.

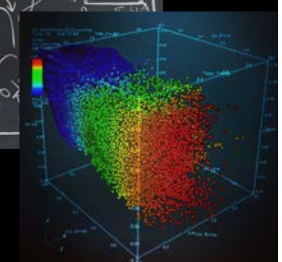
Ian Foster

SCIENCE PARADIGMS

- Thousands years ago: science was empirical ➔ described natural phenomena
- Last few hundred years: theoretical branch ➔ used models and generalizations
- Last few decades: computational branch ➔ simulated complex phenomena
- Today: data exploration (eScience) ➔ unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics





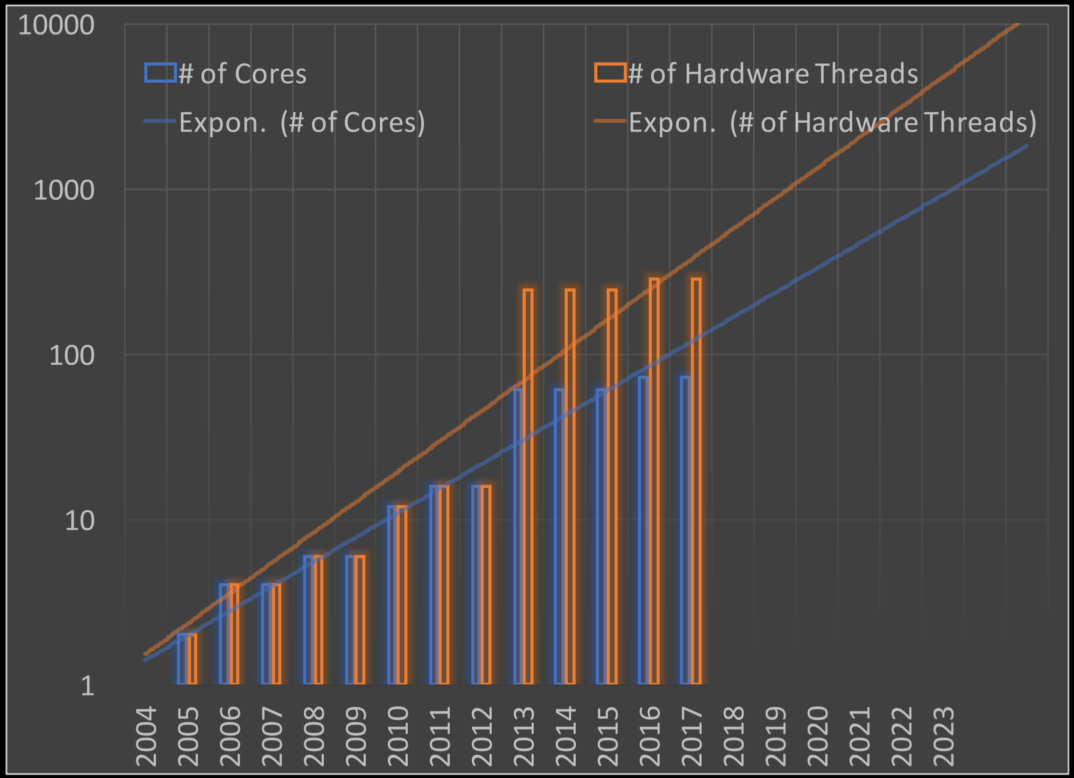


EVERYTHING IS CENTERED AROUND DATA

- Everything about science is changing because of the impact of information technology
- Huge increased in science productivity and advancement
- Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, "data-intensive" science paradigm has emerged (aka ***Big Data*** or ***Data Science***)
- Many new tools and systems need to be created to allow scientists to focus on their science

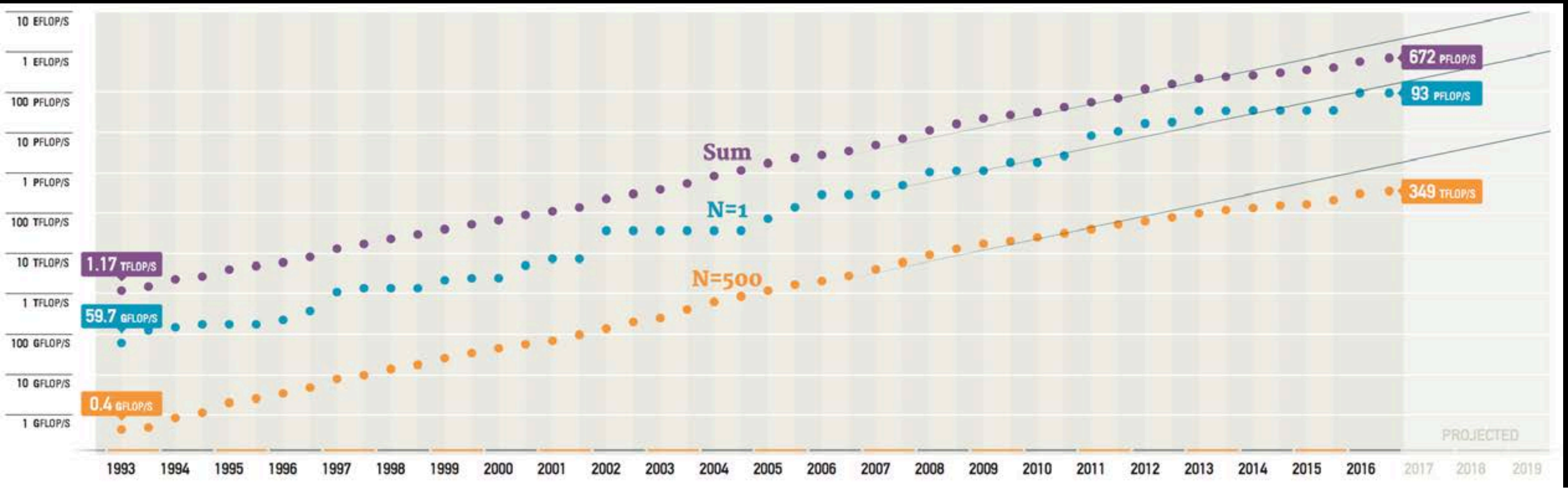
INCREASING COMPUTING INTRANODE

- Many-core processors with x86 architecture from Intel and AMD are increasing in capacity and concurrency
- In 13 years (2004 – 2017)
  - 1 core ➔ 72 cores and 288 HT
  - **288X increase in concurrency**
  - 5.6 GFLOPS ➔ 3456 GFLOPS
  - **617X increase in capacity**



INCREASING COMPUTING INTERNODE

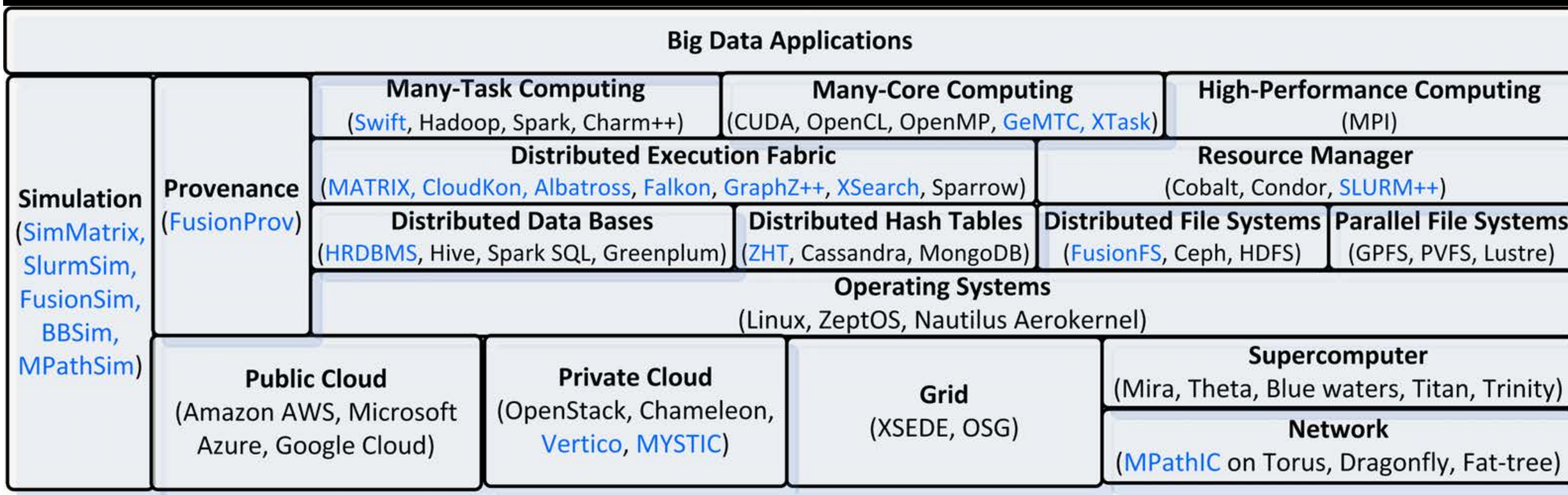
- High-performance Computing systems are increasing in capacity at an exponential rate
- In 13 years (2004 – 2017)
  - 5120 processors ➔ 10649600 cores (**2000X increase**)



ACTIVE RESEARCH PROJECTS TOWARDS SCALABLE RESOURCE MANAGEMENT

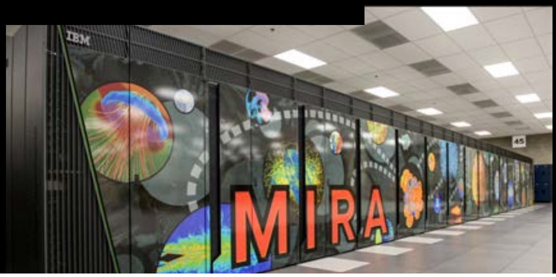
- **Active Projects**
  - **FusionFS:** Fusion distributed File System
  - **ZHT:** Zero-Hop Distributed Hash Table
  - **HRDBMS:** A Scalable Distributed Relational Database for Commodity Hardware
  - **Vertico:** VolunteER cloud compuTIng researCh frameworK
  - **GraphZ++:** Lightweight and Scalable Graph Processing System
  - **FastWorm:** Studying C.Elegans Behaviour through Big Data Computing
  - **XSearch:** Distributed Indexing and Search in Large-Scale Storage Systems
  - **BBSim:** Exploring Burst Buffer Storage Architectures through CODES/ROSS Simulations
  - **XTask:** eXTreme fine-grAined concurrent taSK invocation runtime
  - **MPathIC:** Multi-Path Network Support on Multi-Dimensional Network Architectures
  - **MYSTIC:** prograMmable sYstems reSearch Testbed to explore a stack-wIde adaptive system fabriC

SOFTWARE STACK TOWARDS SCALABLE RESOURCE MANAGEMENT



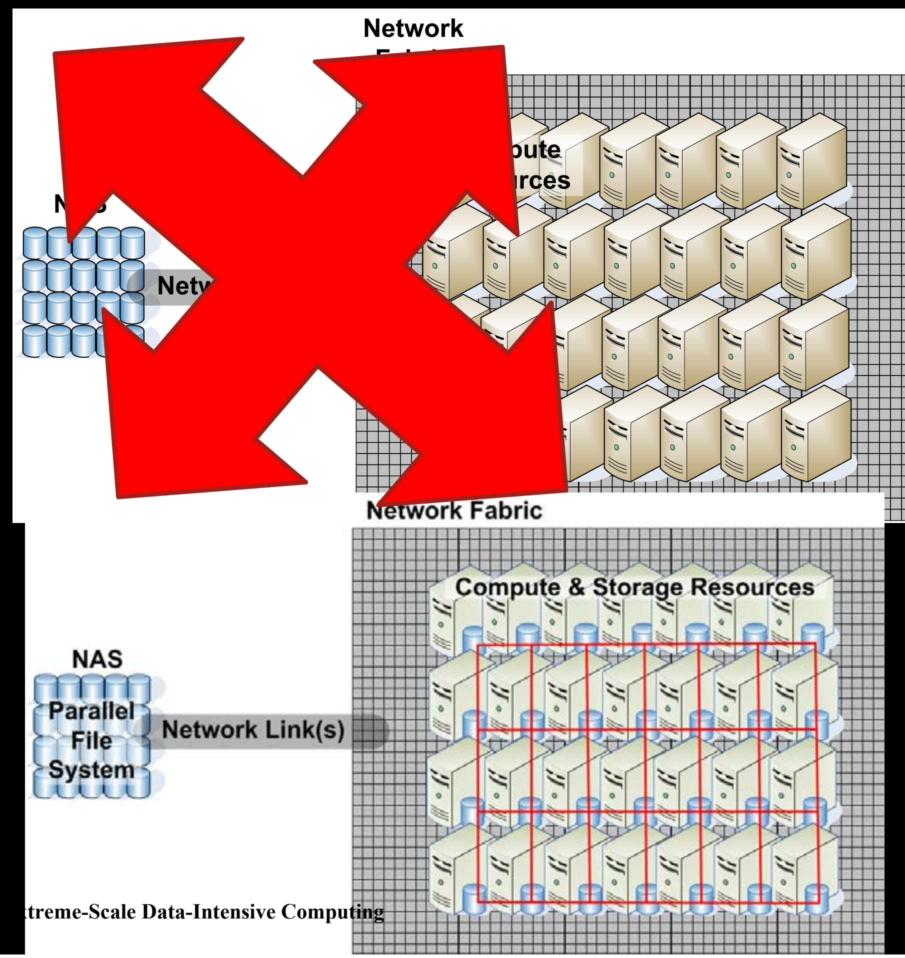
NSF CAREER 2011 – 2018: AVOIDING ACHILLES' HEEL IN EXASCALE COMPUTING WITH DISTRIBUTED FILE SYSTEMS

- Design & impl. of scalable storage systems
  - ➔ extreme scale distributed systems
  - Data management & Scheduling
    - Distributed metadata management
    - Information dispersal algorithms
    - Cooperative caching
    - Dynamic compression
    - Data-aware scheduling
  - **FusionFS: Fusion distributed File System**
  - **ZHT: Zero-Hop Distributed Hash Table**
- Long term goals
  - Support data-intensive science at extreme scales
  - Take current storage system prototypes to production
  - Tackle deep search challenges

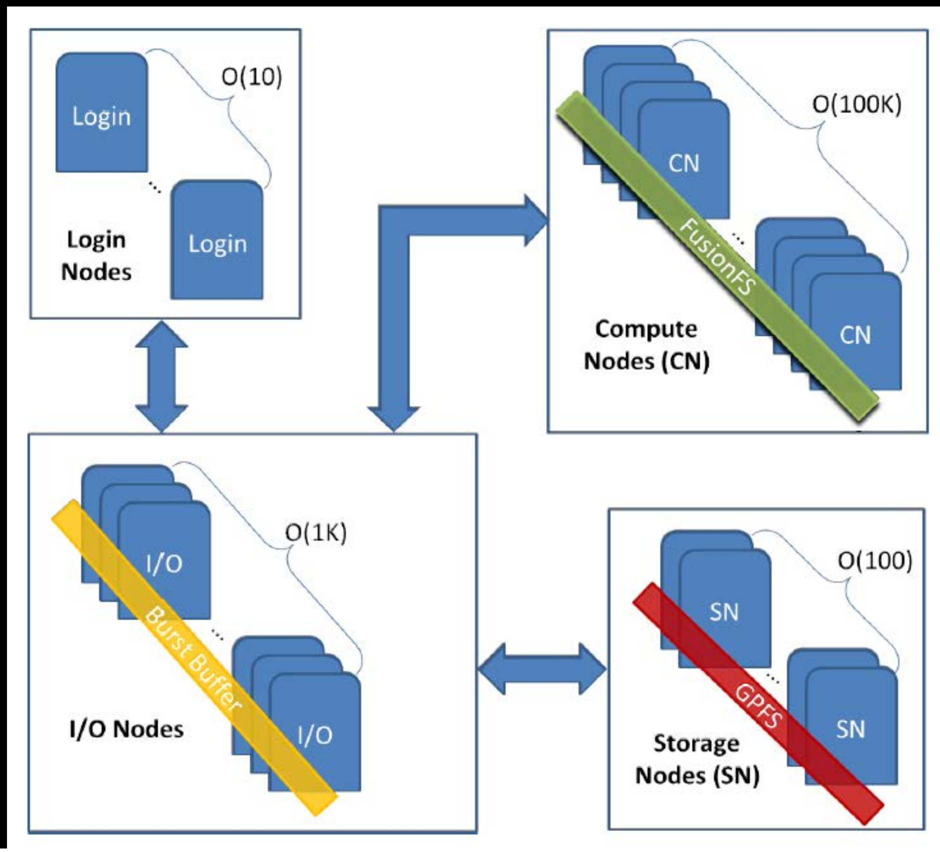


FUSIONFS DISTRIBUTED FILE SYSTEM

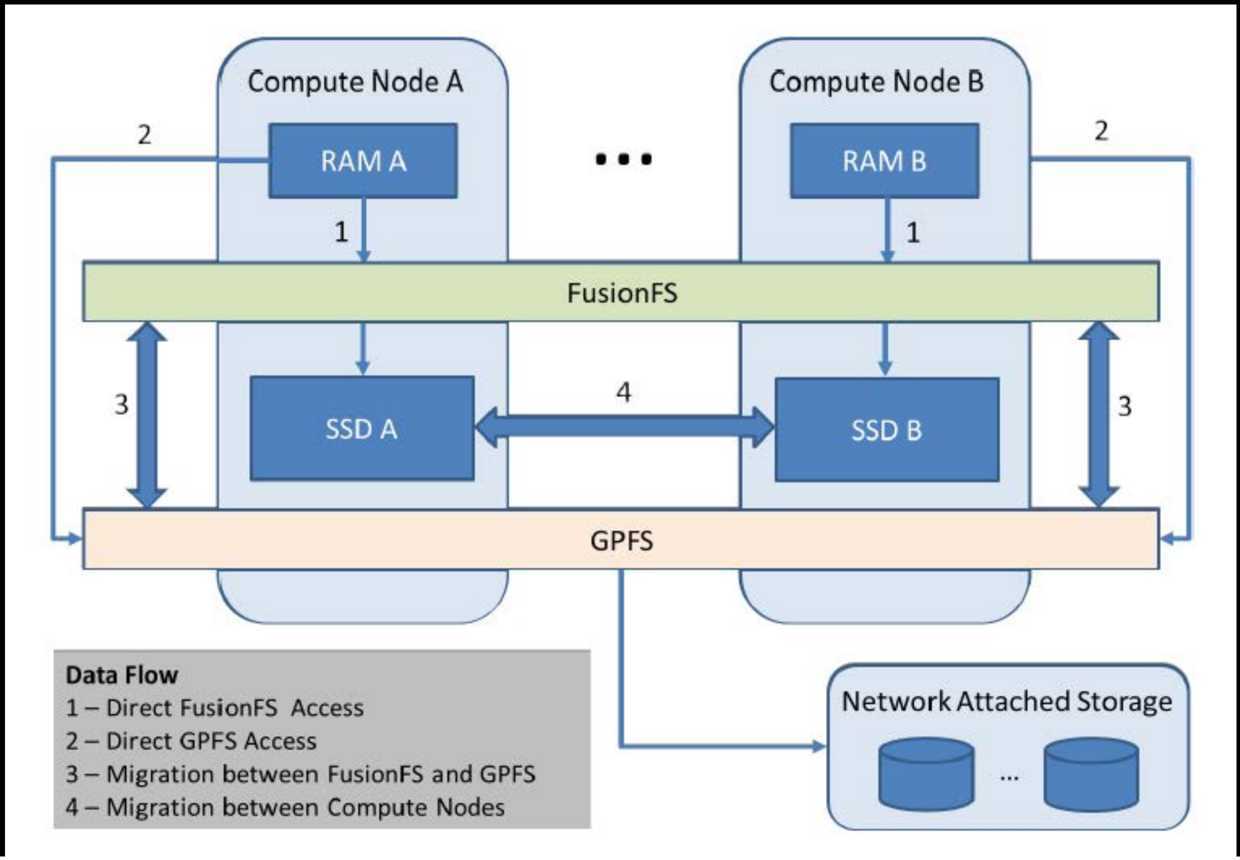
- A distributed file system co-locating storage and computations, while supporting POSIX
- Everything is decentralized and distributed
- Aims for millions of servers and clients scales
- Aims at orders of magnitude higher performance than current state of the art parallel file systems



FUSIONFS OVERVIEW



FUSIONFS OVERVIEW



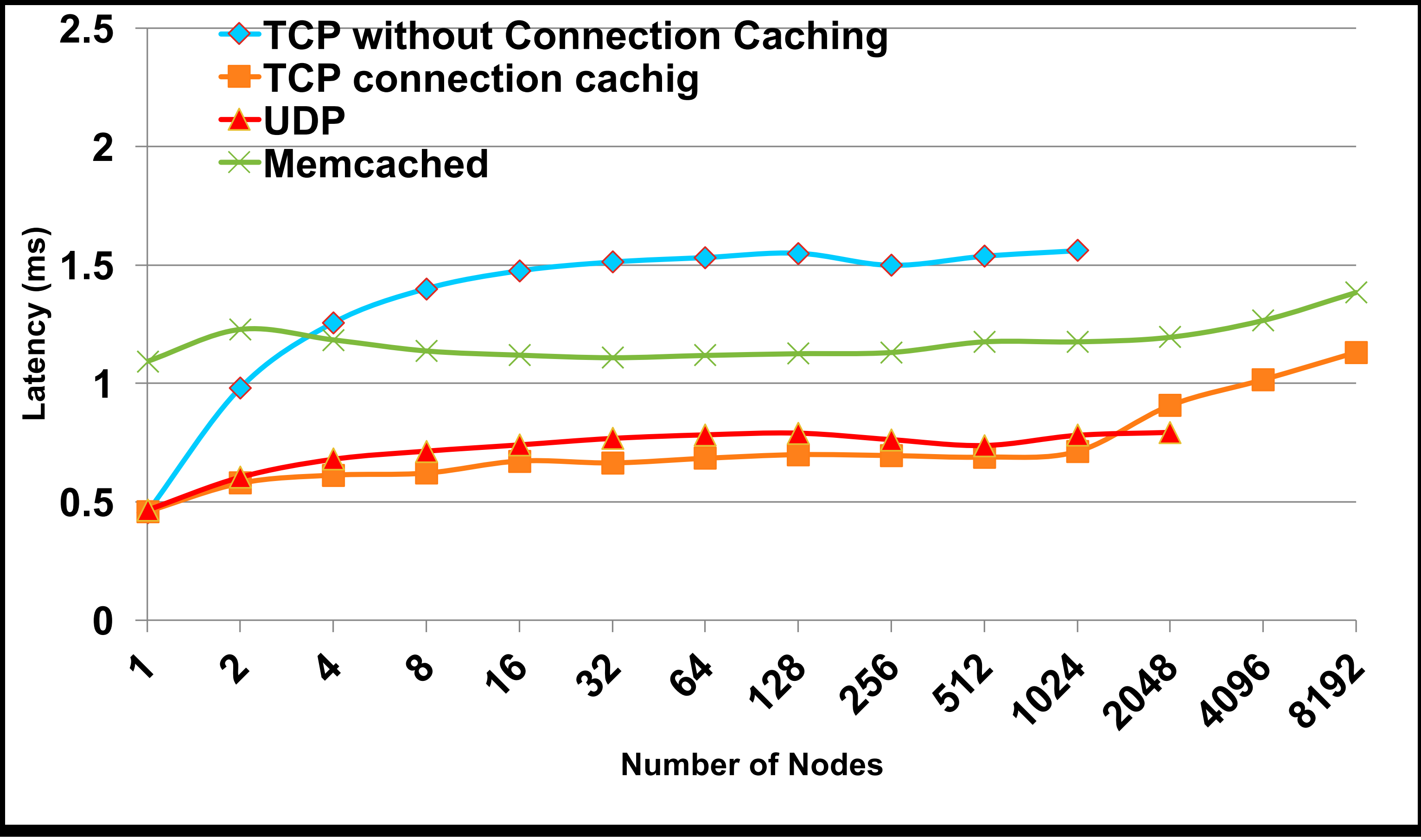
METADATA MANAGEMENT

- Metadata co-located to computation, i.e. metadata is stored directly on compute nodes
- Metadata is evenly distributed on compute nodes
  - Pro: good load balance
  - Con: bad metadata locality
- Approach: implementation on top of Zero-hop distributed Hash Tables (ZHT)

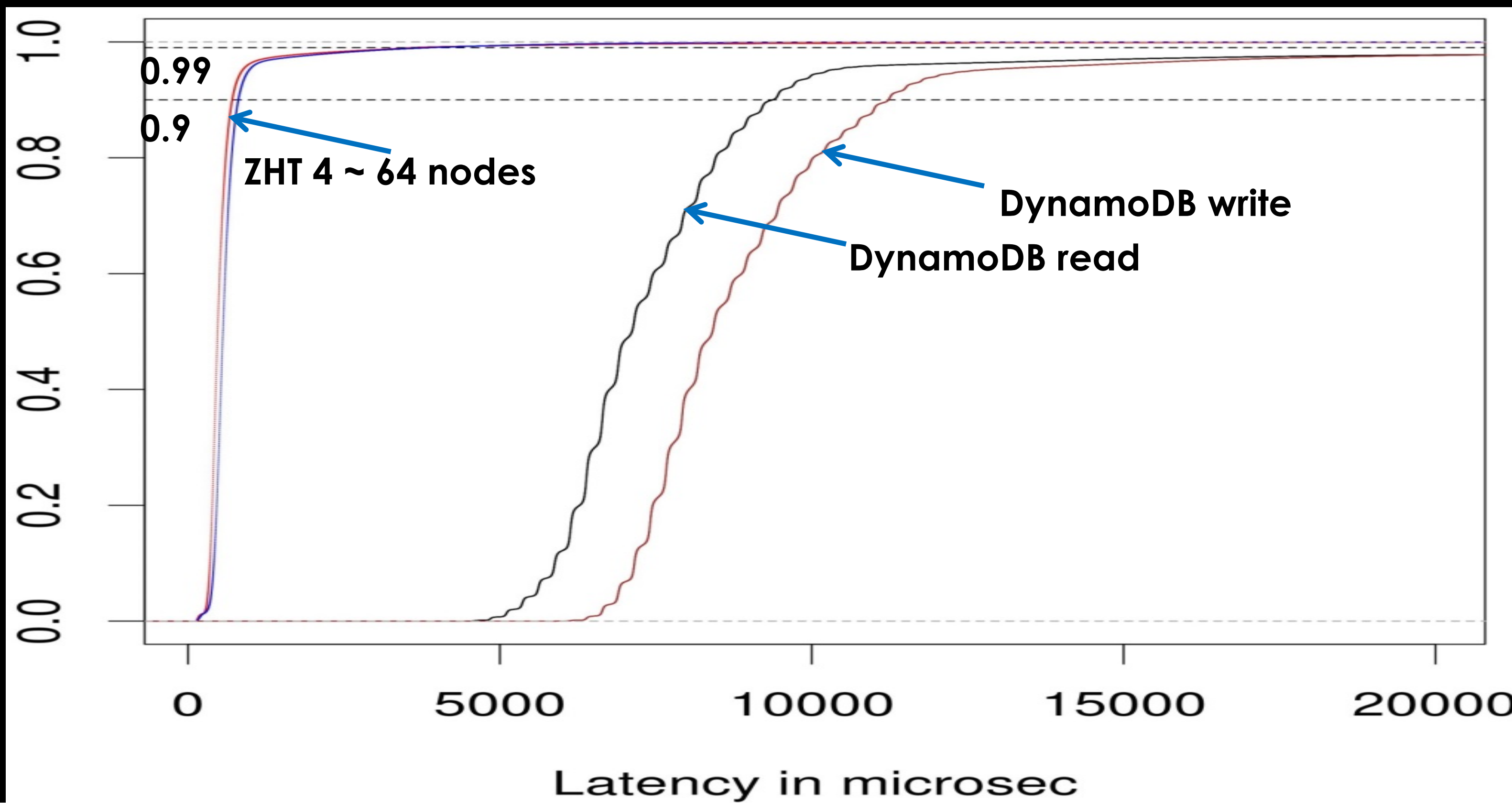
ZHT: ZERO-HOP DISTRIBUTED HASH TABLE

- Light-weighted
- High performance
- Scalable
- Dynamic
- Fault tolerant
- Strong Consistency
- Persistent
- Versatile: works from clusters, to clouds, to supercomputers

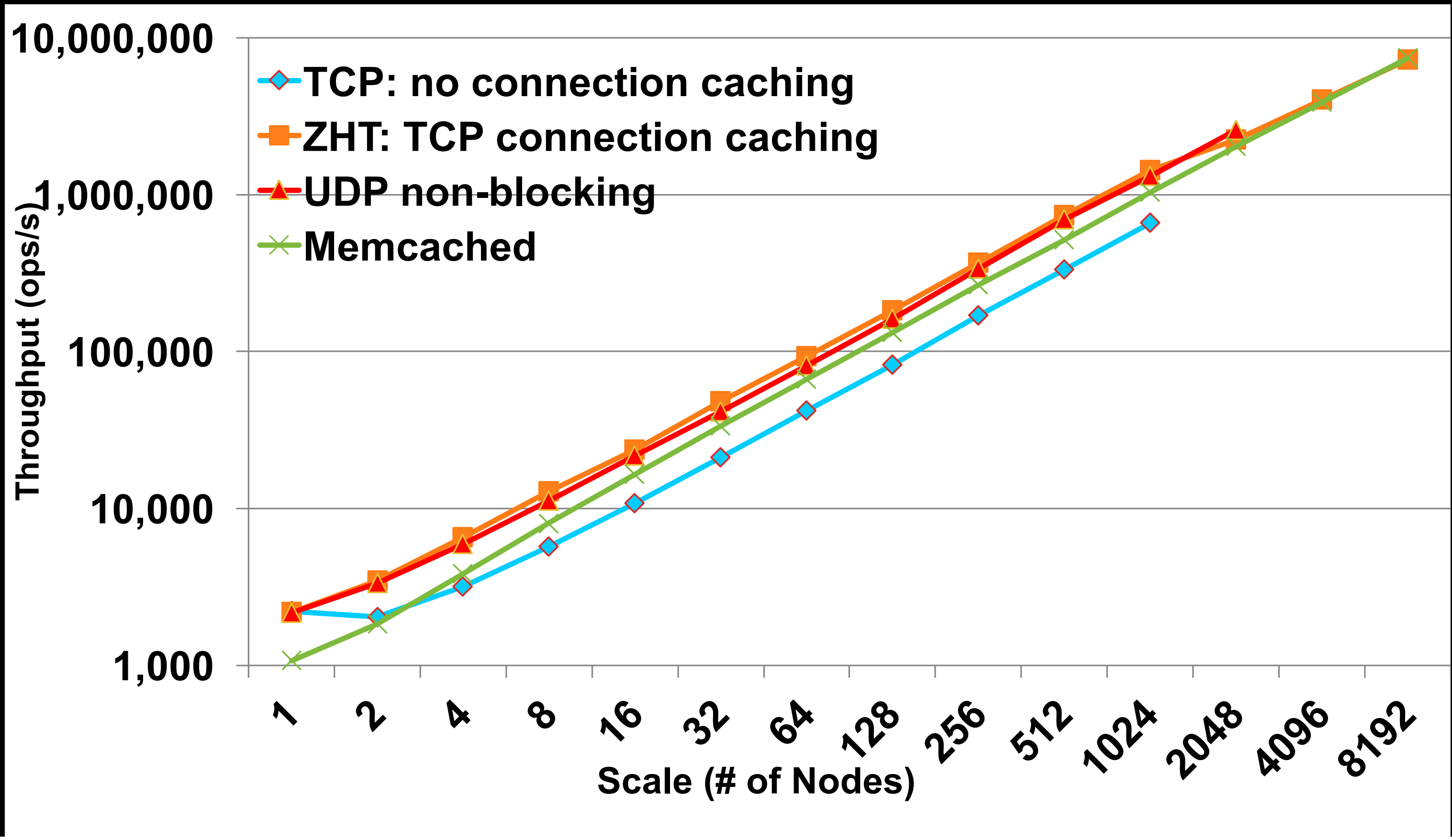
ZHT: ZERO-HOP DISTRIBUTED HASH TABLE



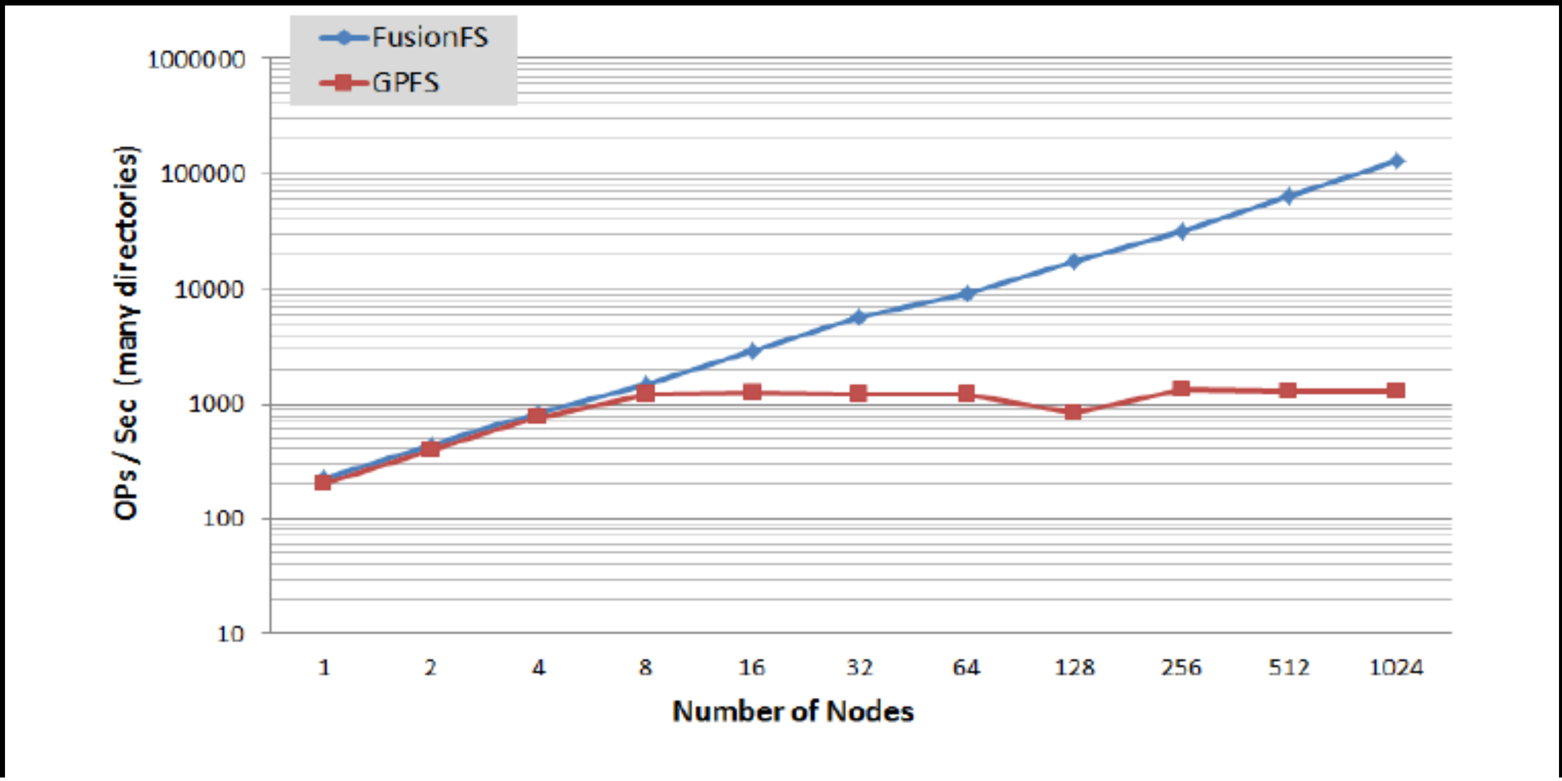
ZHT: ZERO-HOP DISTRIBUTED HASH TABLE



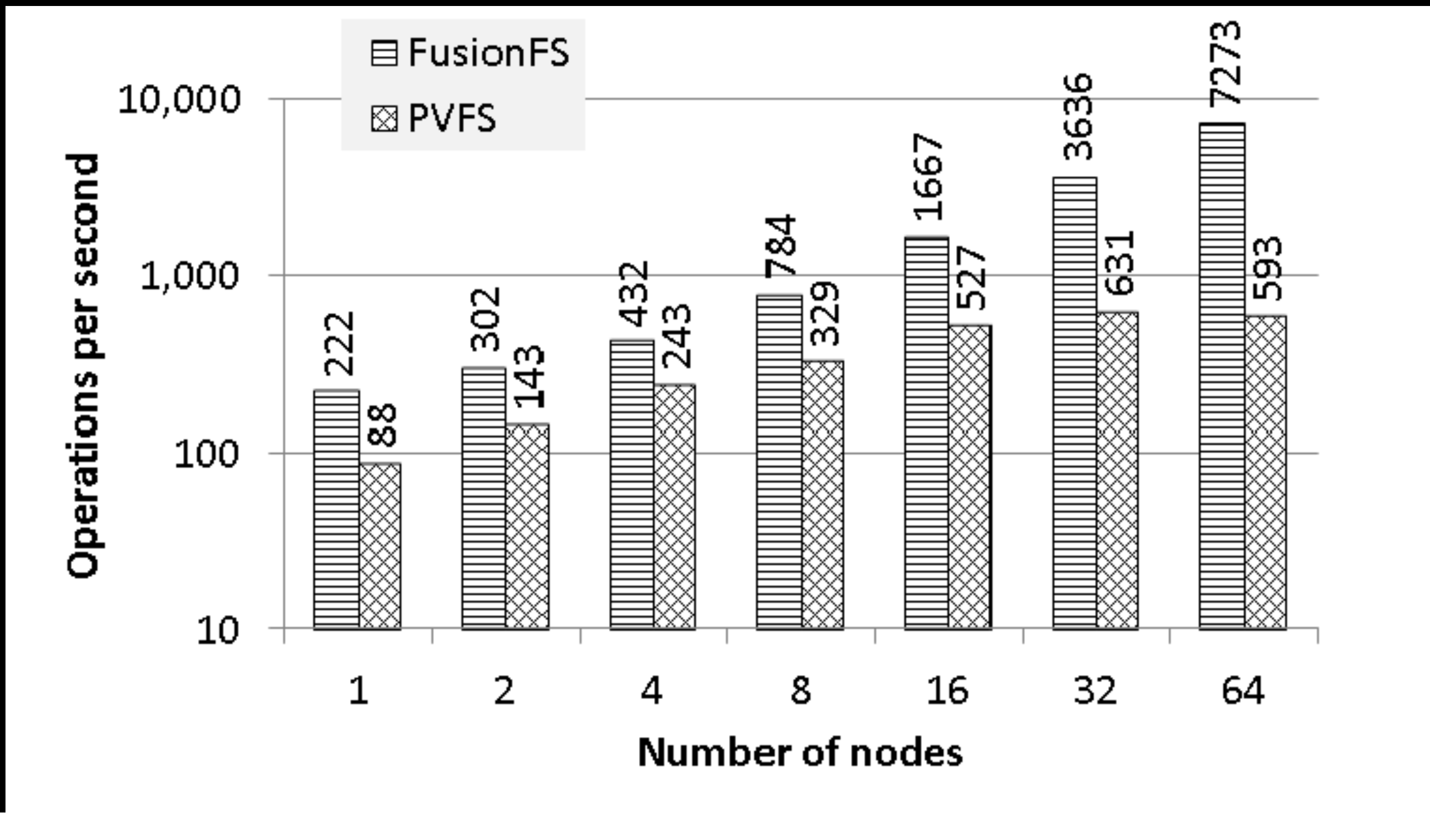
ZHT: ZERO-HOP DISTRIBUTED HASH TABLE



FUSIONFS METADATA MANAGEMENT



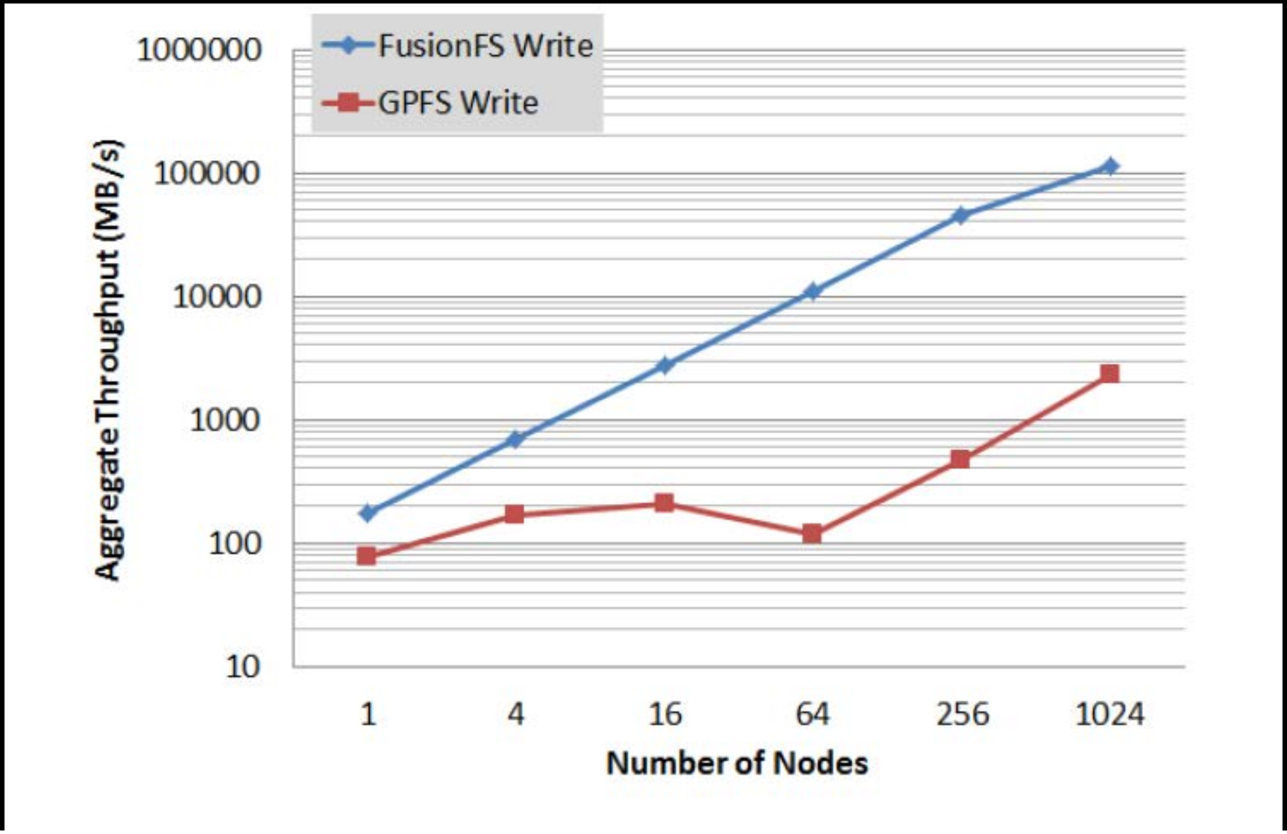
FUSIONFS METADATA MANAGEMENT



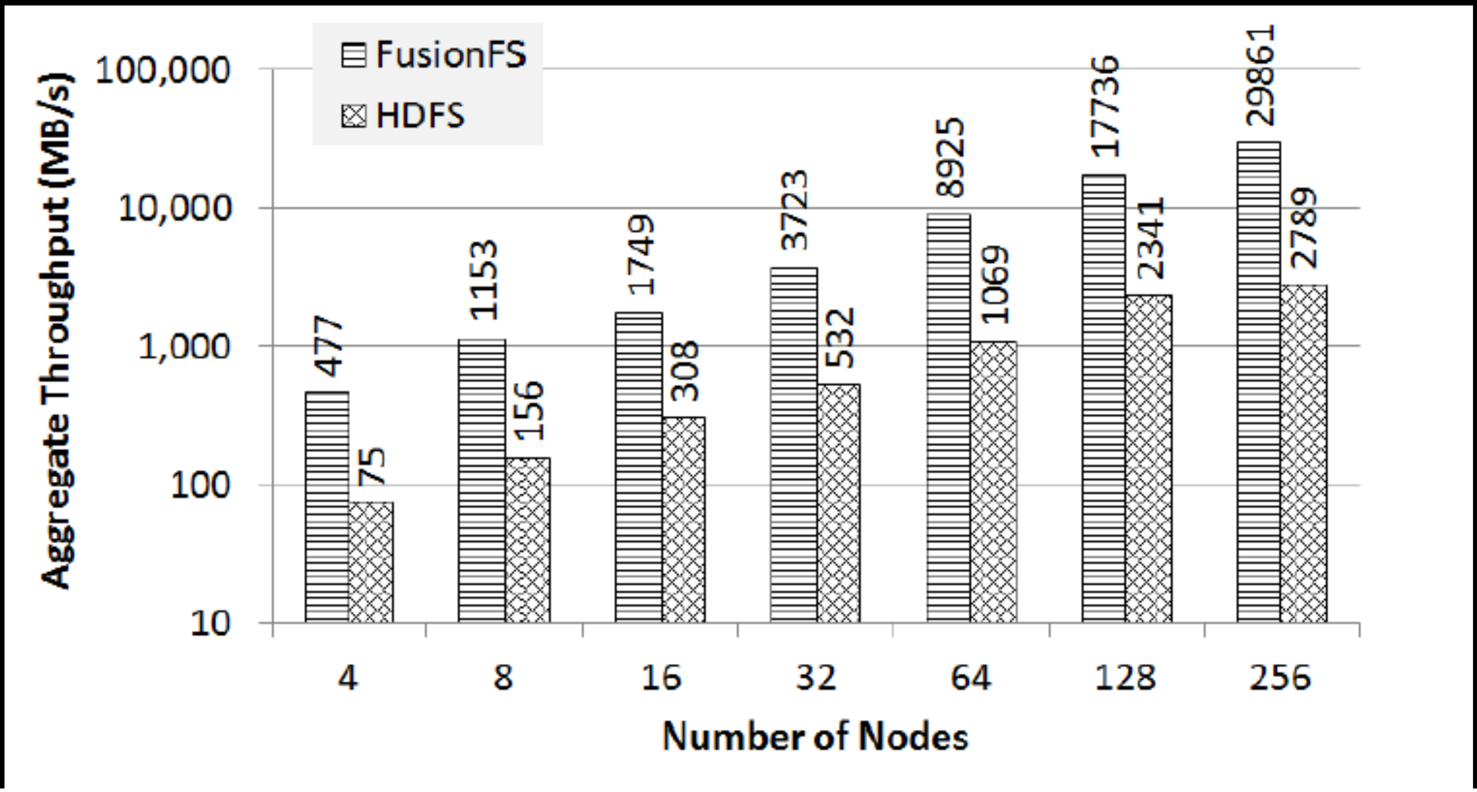
FUSIONFS I/O THROUGHPUT

- Always write to the local compute node
  - Pro: maximal aggregate throughput
  - Con: bad load balance
    - Solution: Asynchronous rebalancing
- No locality-awareness for data read
  - Transfer from the node that holds the requested file
    - Maybe it is just loopback if we are lucky
    - Can we schedule the job on the node with the required data?
  - Then read locally

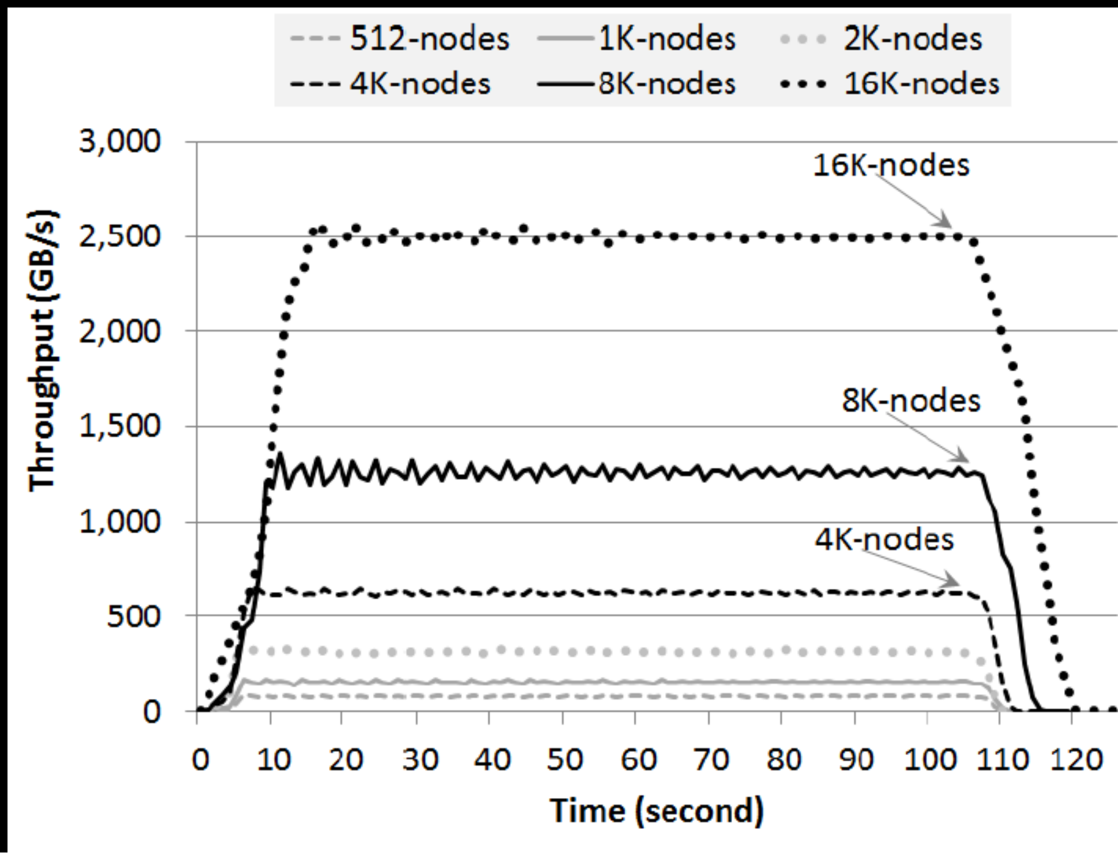
FUSIONFS I/O THROUGHPUT



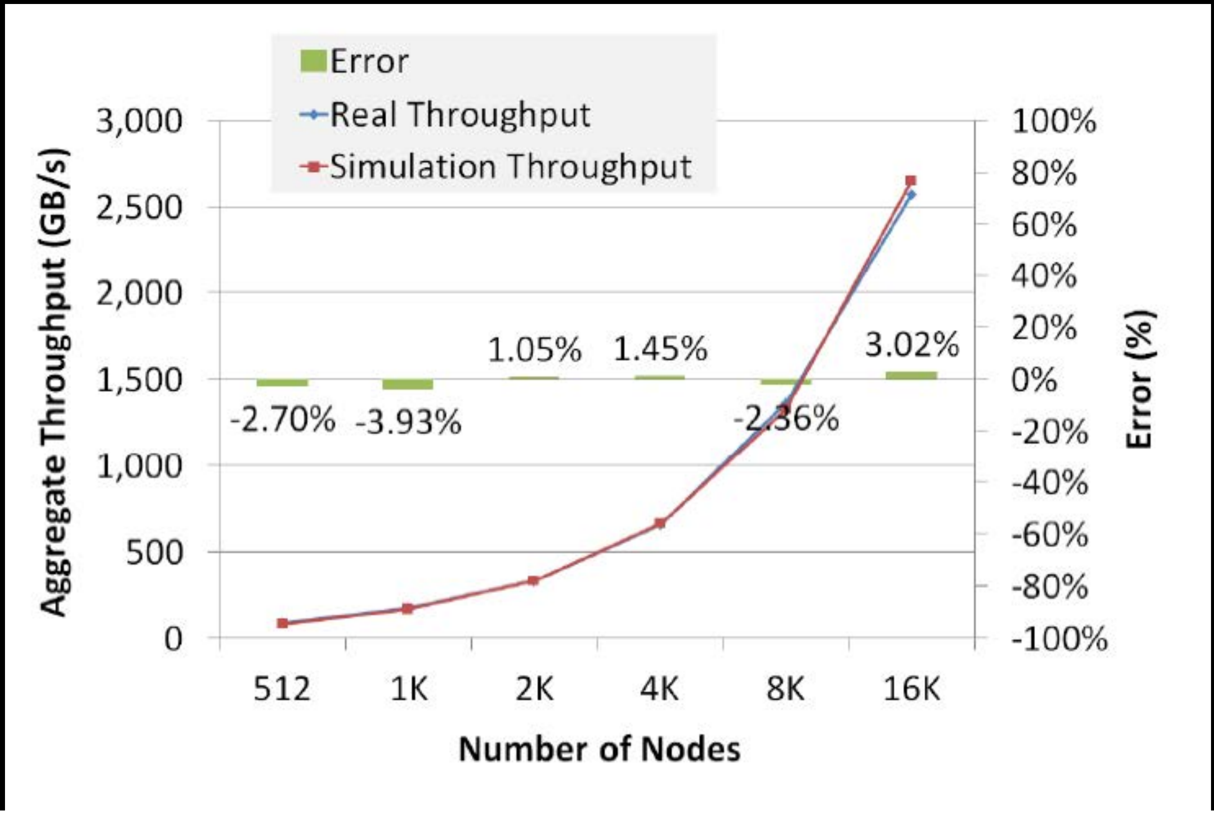
FUSIONFS I/O THROUGHPUT



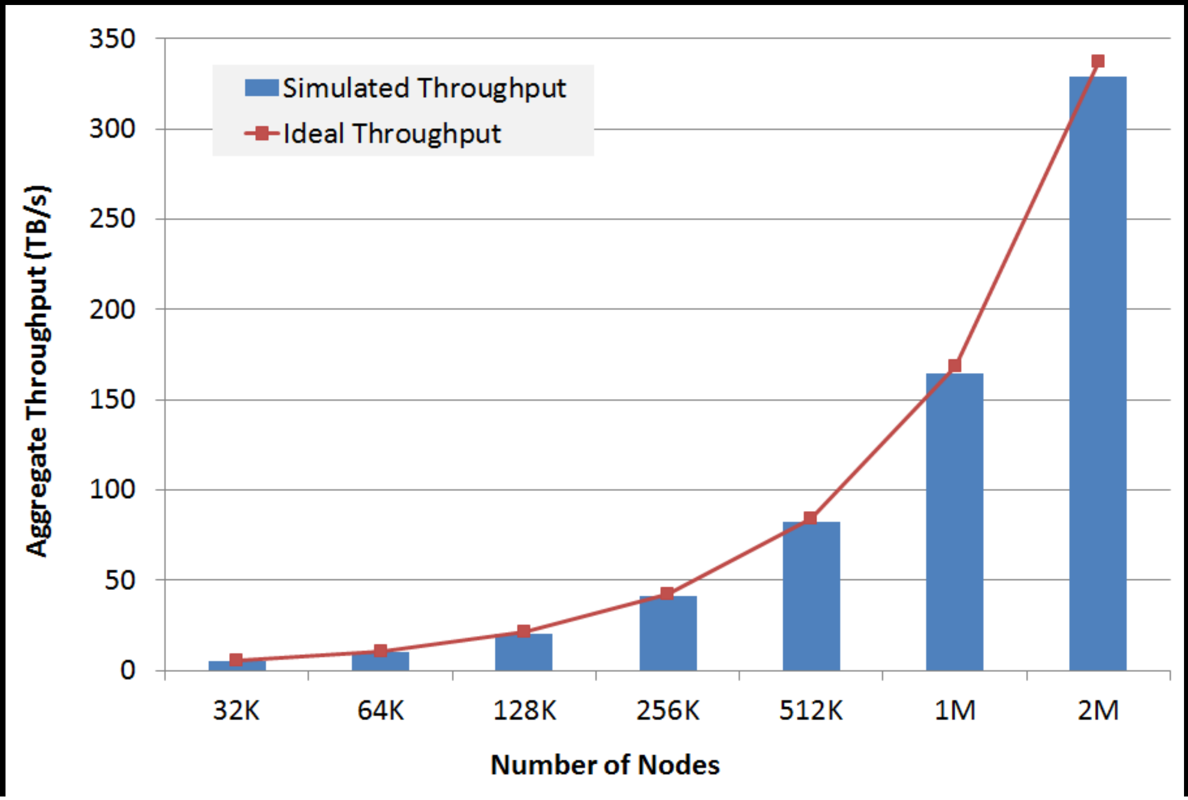
FUSIONFS I/O THROUGHPUT



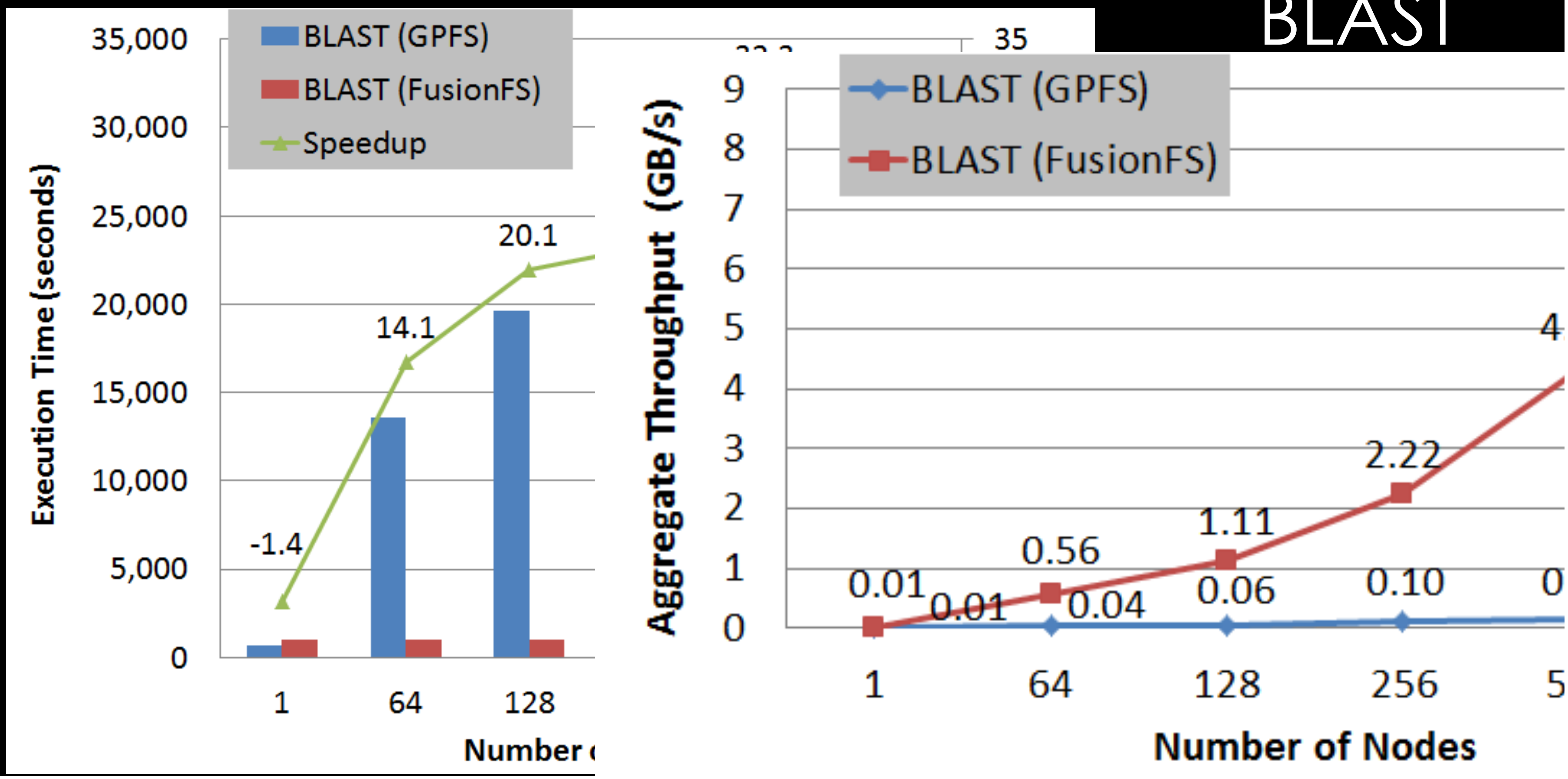
FUSIONFS I/O THROUGHPUT SIMULATION VALIDATION



FUSIONFS I/O THROUGHPUT SIMULATION

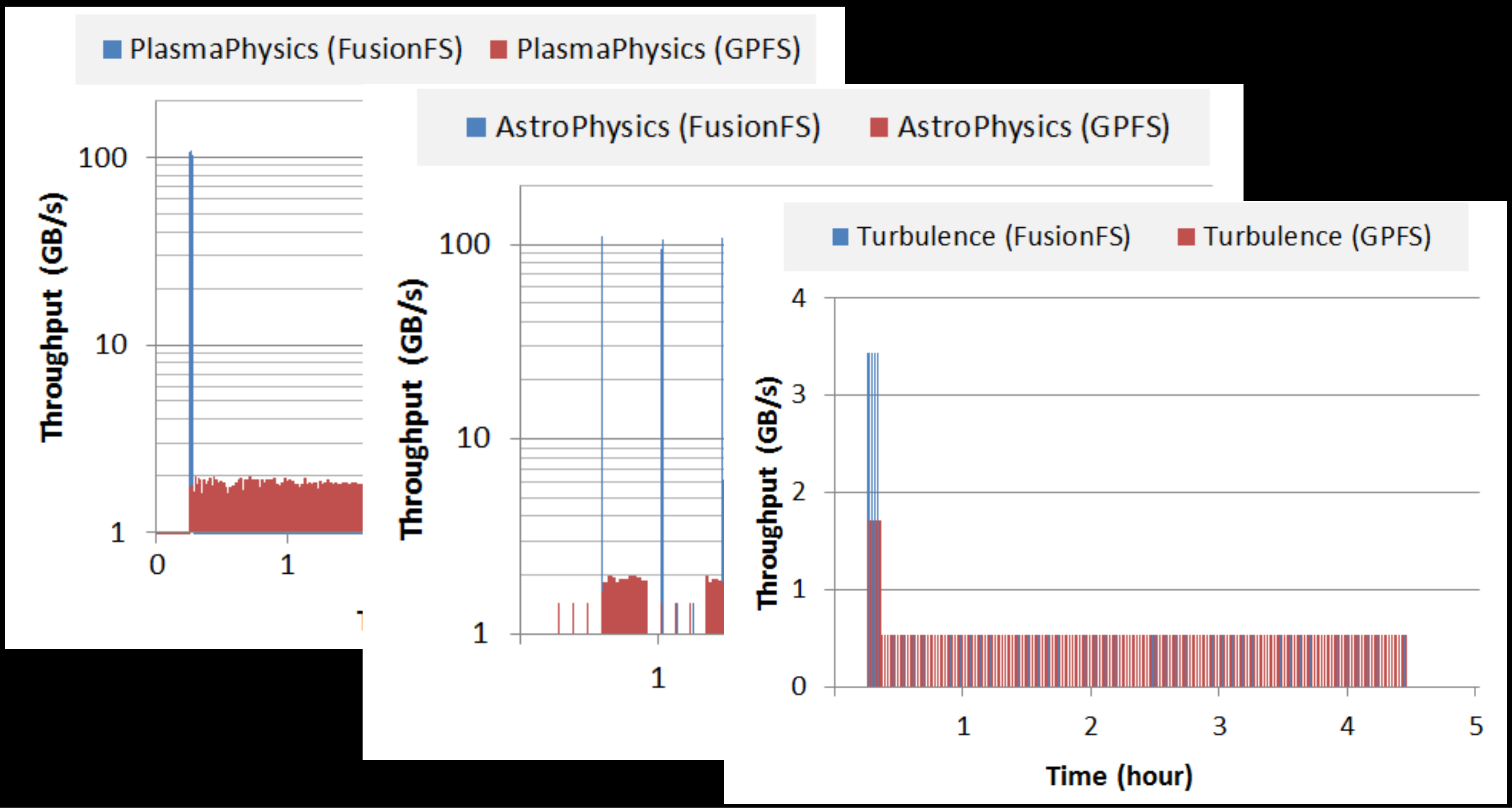


APPLICATIONS
BLAST



APPLICATIONS

TOP 3 DATA-INTENSIVE APPLICATIONS ON INTREPID BG/P



MAIN MESSAGE FROM DISTRIBUTED STORAGE RESEARCH

- ***Decentralization is critical***
  - Computational resource management
  - Storage systems
- ***Preserving locality is critical!***
  - POSIX I/O on shared/parallel file systems ignore locality
  - Data-aware scheduling coupled with distributed file systems that expose locality is the key to scalability over the next decade
- *Co-locating storage and compute is **GOOD***
  - Leverage the abundance of processing power, bisection bandwidth, and local I/O
- **2010 → 2017: stateless compute nodes → burst-buffer architecture**

SOME CHALLENGES AND OPPORTUNITIES IN THE COMING DECADE

- Scale **storage** in capacity, performance, and features
  - DFS, NoSQL, active storage, burst buffer, deep search
- **Programmability** at extreme concurrency and scale
  - MPI, PGAS, MTC, OpenMP
- Improve network performance and resilience through multi-path routing protocols on **multi-dimensional networks**
  - Fat-tree, Torus, Dragonfly
- Improve **power efficiency**
  - GPUs, ARM, FPGAs, approximate computing
- Simplify programming and improve resilience with **universal memory**
  - 3D Xpoint, PCM

THE BIG BANG OF COMPUTING

BS PRODUCTION IN COMPUTER SCIENCE

Similar level to 1985 when …

<1K Internet hosts ➔ 7B+

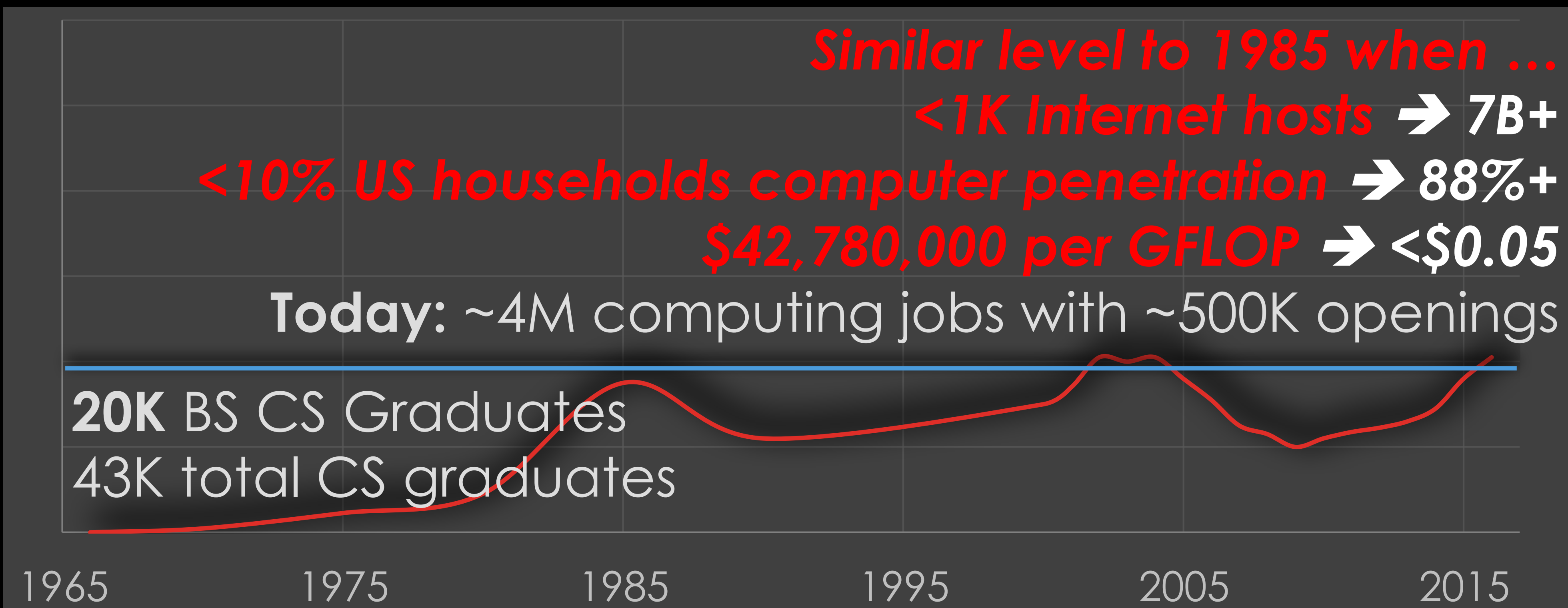
<10% US households computer penetration ➔ 88%+

$42,780,000 per GFLOP ➔ <$0.05

Today: ~4M computing jobs with ~500K openings

20K BS CS Graduates

43K total CS graduates

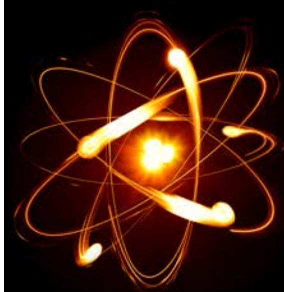


STUDENTS

- **Research Engagement** (145 students from 2010 – 2017)
  - **PhD** students: (8/13) through CAREER/ANL/LANL/IIT
  - **Master** students: 57 through independent studies
  - **Undergraduate** students: 60 through CAREER/REU/SCC
  - **High-school** students: 7 through CAREER/SCC
- **Teaching Engagement** (~1000 students from 2010 – 2016)
  - Intro. to Parallel and Distributed Computing (introduced in 2012)
  - Advanced Operating Systems
  - Cloud Computing (introduced in 2012)
  - Data-Intensive Computing (introduced in 2010)

BIGDATAX REU SUMMER PROGRAM

STUDENT CLUSTER COMPETITION AT IEEE/ACM SUPERCOMPUTING/SC



ACTIVE COLLABORATORS

- **Illinois Institute of Technology (CS)**
  - Kyle Hale
  - Xian-He Sun
  - Zhiling Lan
  - Sanjiv Kapoor
  - Gruia Calinescu
  - Boris Glavic
- **Argonne National Lab. (MCS/ALCF)**
  - William Scullin
  - Rob Ross
- **University of Chicago (CS/CI) & Argonne National Lab. (MCS)**
  - Ian Foster
  - Kyle Chard
  - Michael Mile
  - Justin Wozniak
- **Univ. of Chicago (Radiology)**
  - Samuel G. Armato III
- **Northwestern University (EECS)**
  - Peter Dinda
- **Los Alamos National Lab. (USRC)**
  - Mike Lang
- **Fermi Nat. Accelerator Lab. (FermiCloud)**
  - Steven Timm
- **Rosalind Franklin University of Medicine and Science (Biology)**
  - Hongkyun Kim
- **DePaul University (CS)**
  - Daniela Stan Raicu
  - Jocob Furst

QUESTIONS

- Contact:
  - ioan.raicu@northwestern.edu
  - iraicu@cs.iit.edu
- More information:
  - http://www.cs.iit.edu/~iraicu/
  - http://datasys.cs.iit.edu
  - https://www.linkedin.com/in/ioanraicu
  - https://scholar.google.com/citations?user=jE73HYAAAAAJ