Supporting Data-Intensive Distributed Computing in an Exascale Era

Ioan Raicu

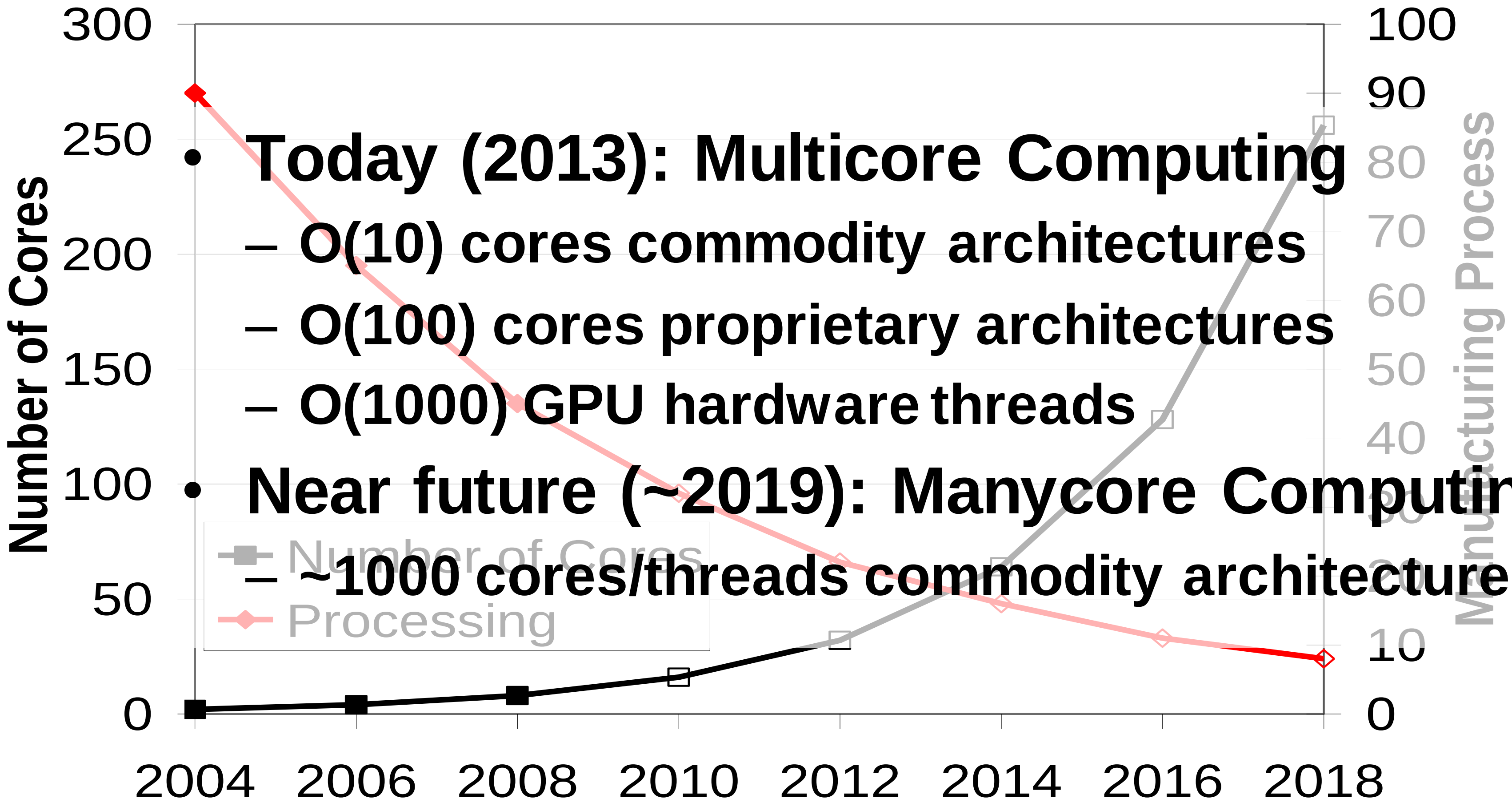
Computer Science Department, Illinois Institute of Technology

Math and Computer Science Division, Argonne National Laboratory

August 7th, 2013

MAGIC Meeting: 2020-2025 Scientific Computing Environments

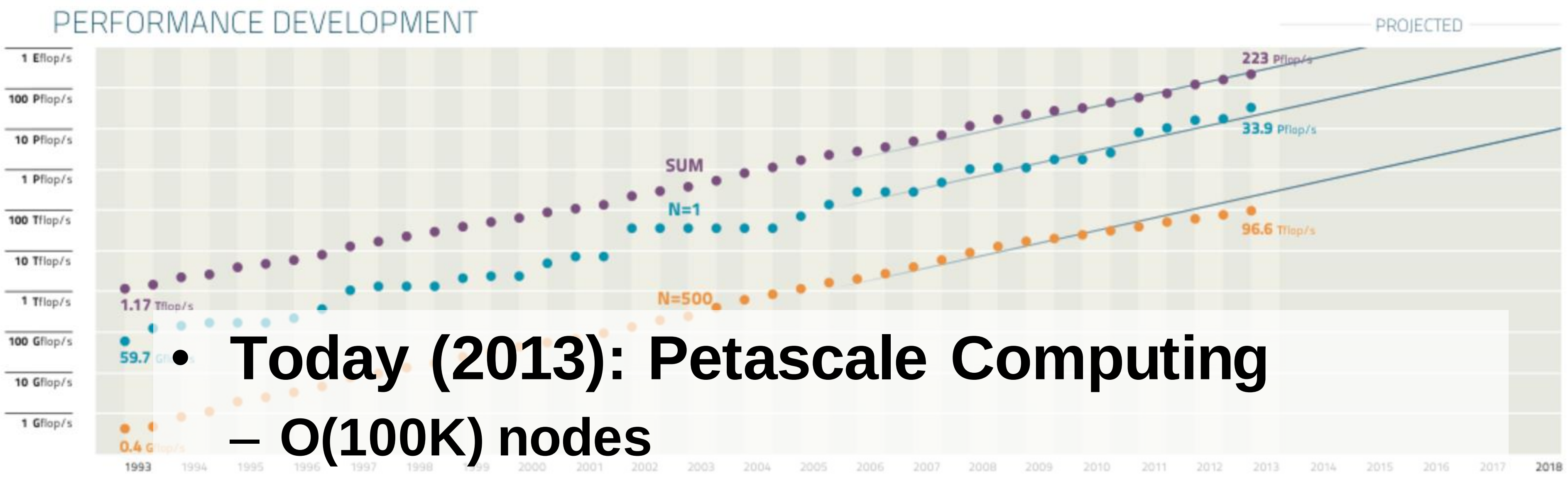
Manycore Computing



- **Today (2013): Multicore Computing**
  - **O(10) cores commodity architectures**
  - **O(100) cores proprietary architectures**
  - **O(1000) GPU hardware threads**
- **Near future (~2019): Manycore Computing**
  - **~1000 cores/threads commodity architectures**

Pat Helland, Microsoft, The Irresistible Forces Meet the Movable Objects, November 9th, 2007

Exascale Computing



- **Today (2013): Petascale Computing**
  - **O(100K) nodes**
  - **O(1M) cores**

http://www.top500.org/

- **Near future (~2018): Exascale Computing**
  - **~1M nodes (10X)**
  - **~1B processor-cores/threads (1000X)**

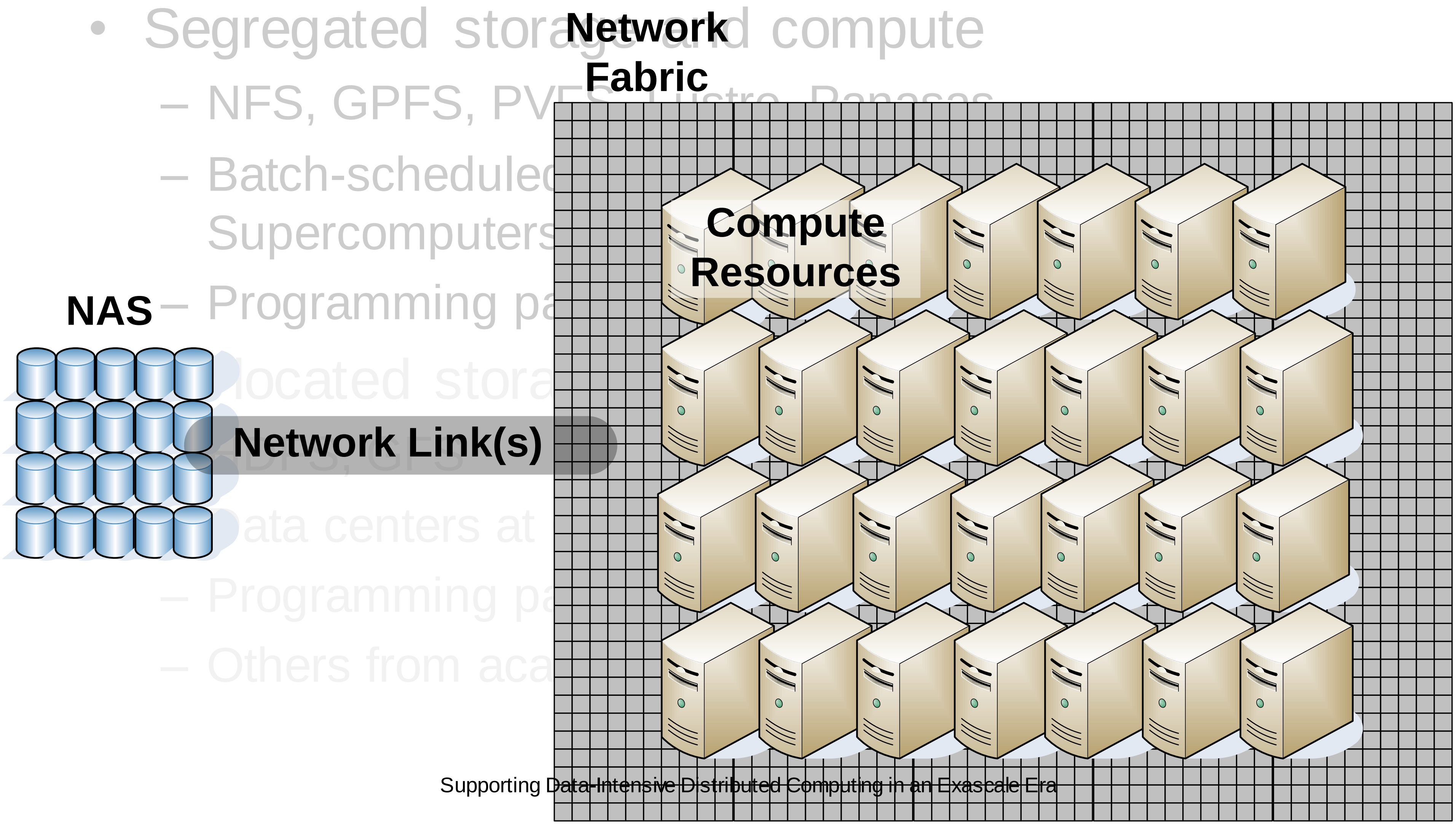
http://s.top500.org/static/lists/2013/06/TOP500_201306_Poster.png

Exascale Computing Architecture

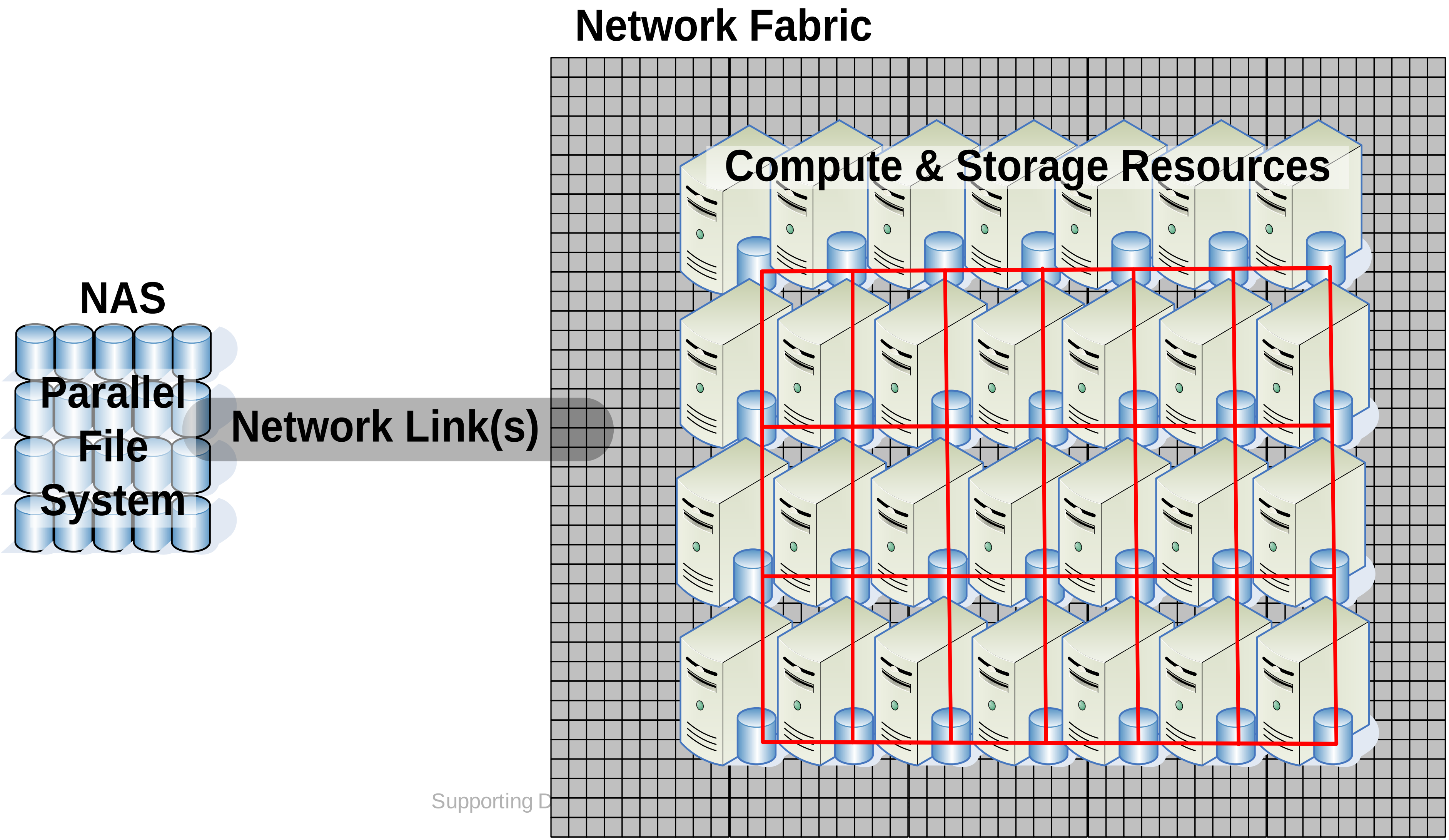
- Compute
  - 1M nodes, with ~1K threads/cores per node
- Networking
  - N-dimensional torus
  - Meshes
- Storage
  - SANs with spinning disks will replace today's tape
  - SANs with SSDs might exist, replacing today's spinning disk SANs
  - SSDs might exist at every node

State-of-the-Art Storage Systems in HEC
Parallel File Systems

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



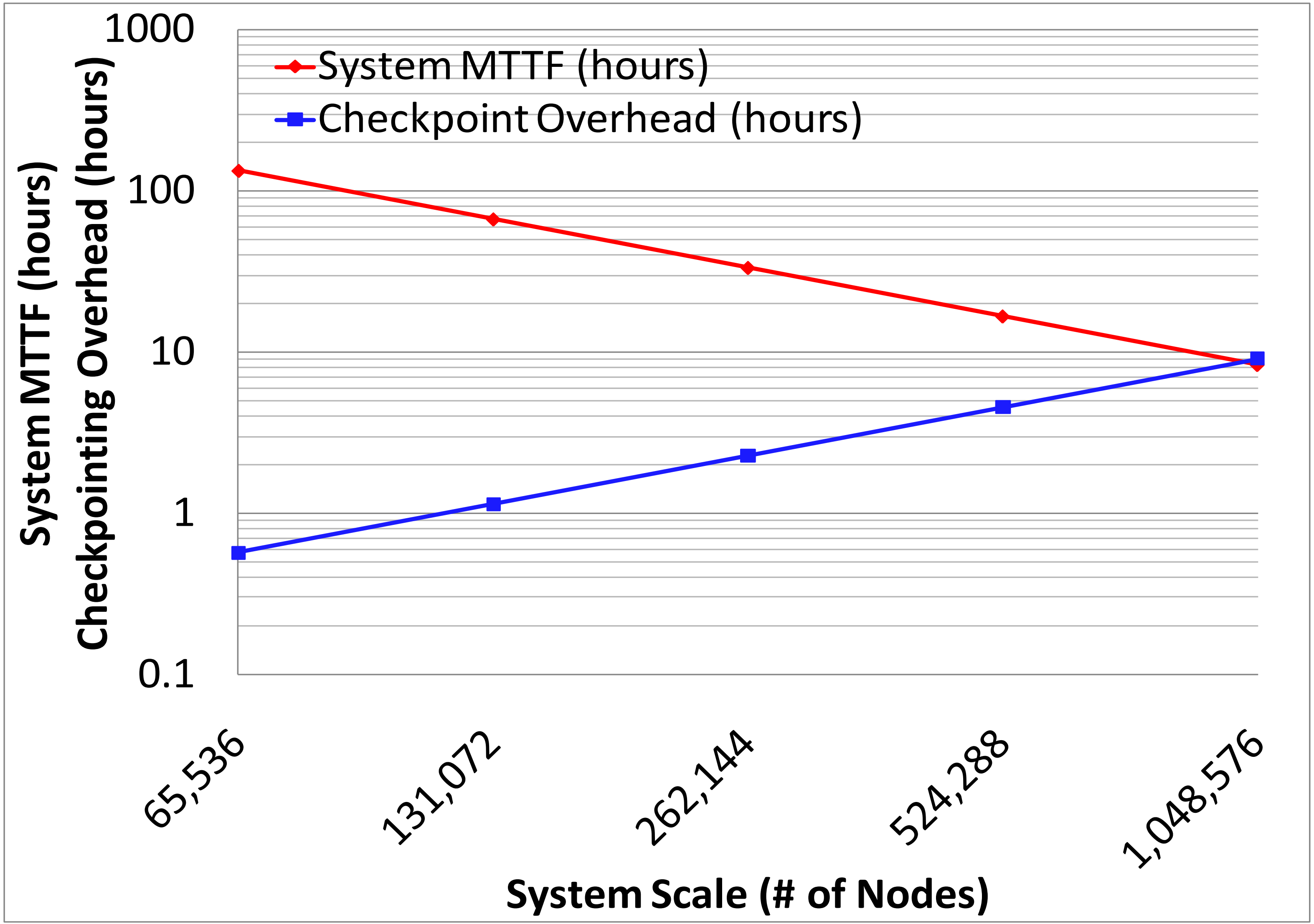
Future Storage System Architecture for Extreme Scale HEC



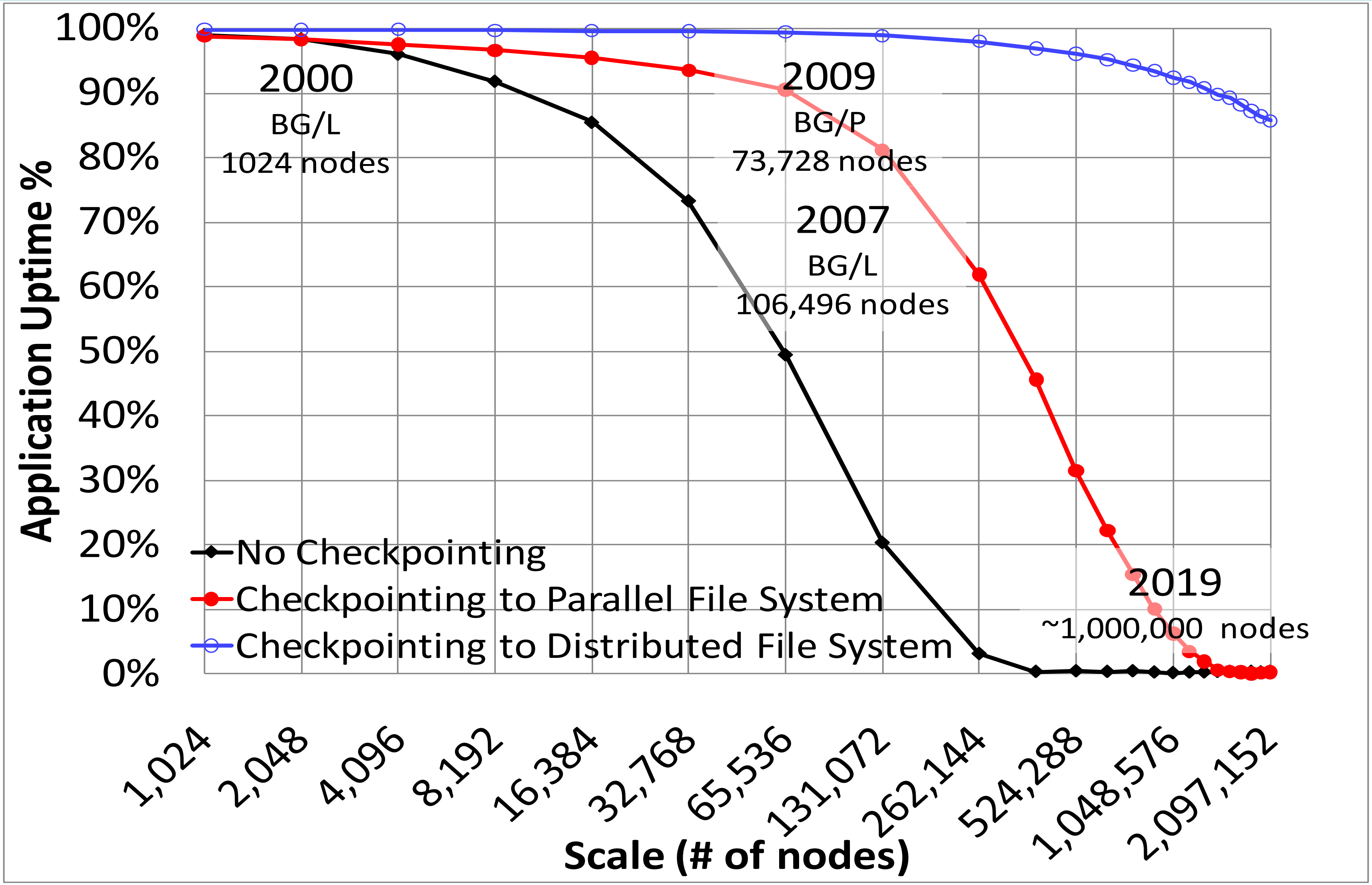
Some Challenges to Overcome at Exascale Computing

- Programming paradigms
  - HPC is dominated by MPI today
  - Will MPI scale another 3 orders of magnitude?
  - Other paradigms (including loosely coupled ones) might emerge to be more flexible, resilient, and scalable
- Storage systems will need to become more distributed to scale ➔ Critical for resilience of HPC
- Network topology must be used in job management, data management, compilers, etc
- Power efficient compilers and run-time systems

Expected checkpointing cost and MTTF towards exascale



Simulation application uptime towards exascale



Main Message

- ***Decentralization is critical***
  - Computational resource management (e.g. LRMs)
  - Storage systems (e.g. parallel file 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

Critical Technologies Needed to achieve Extreme Scales

- Fundamental Building Blocks (with a variety of resilience and consistency models)
  - Distributed hash tables (aka NoSQL data stores)
  - Distributed Message Queues
- Deliver future generation distributed systems
  - Global File Systems, Metadata, and Storage
  - Job Management Systems
  - Workflow Systems
  - Monitoring Systems
  - Provenance Systems
  - Data Indexing

More Information

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