Exceptional service in the national interest

Sandia National Laboratories

An Evaluation of BitTorrent's Performance In HPC Environments

Matthew G. F. Dosanjh, Patrick G. Bridges,
Suzanne M. Kelly, James H. Laros III,
Courtenay T. Vaughan

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Motivation

Sandia National Laboratories

- Need scalable system services in large HPC systems
  - Example services: file systems, job launch, system monitoring
  - Communication bottlenecks; performance degradation
- Can we use recent distributed systems techniques here?
  - E.G. Peer-to-peer services developed for Internet systems
  - Must scale to large systems with complex topologies and bottlenecks
  - Not designed for HPC use-cases, networks, etc.

An Example Problem

Sandia National Laboratories

- Dynamic Shared Libraries distribution on HPC platforms
  - Read-only data-transfer
  - Many simultaneous downloaders
  - Looks a lot like a peer-to-peer problem
- Can we use BitTorrent?

Sandia National Laboratories

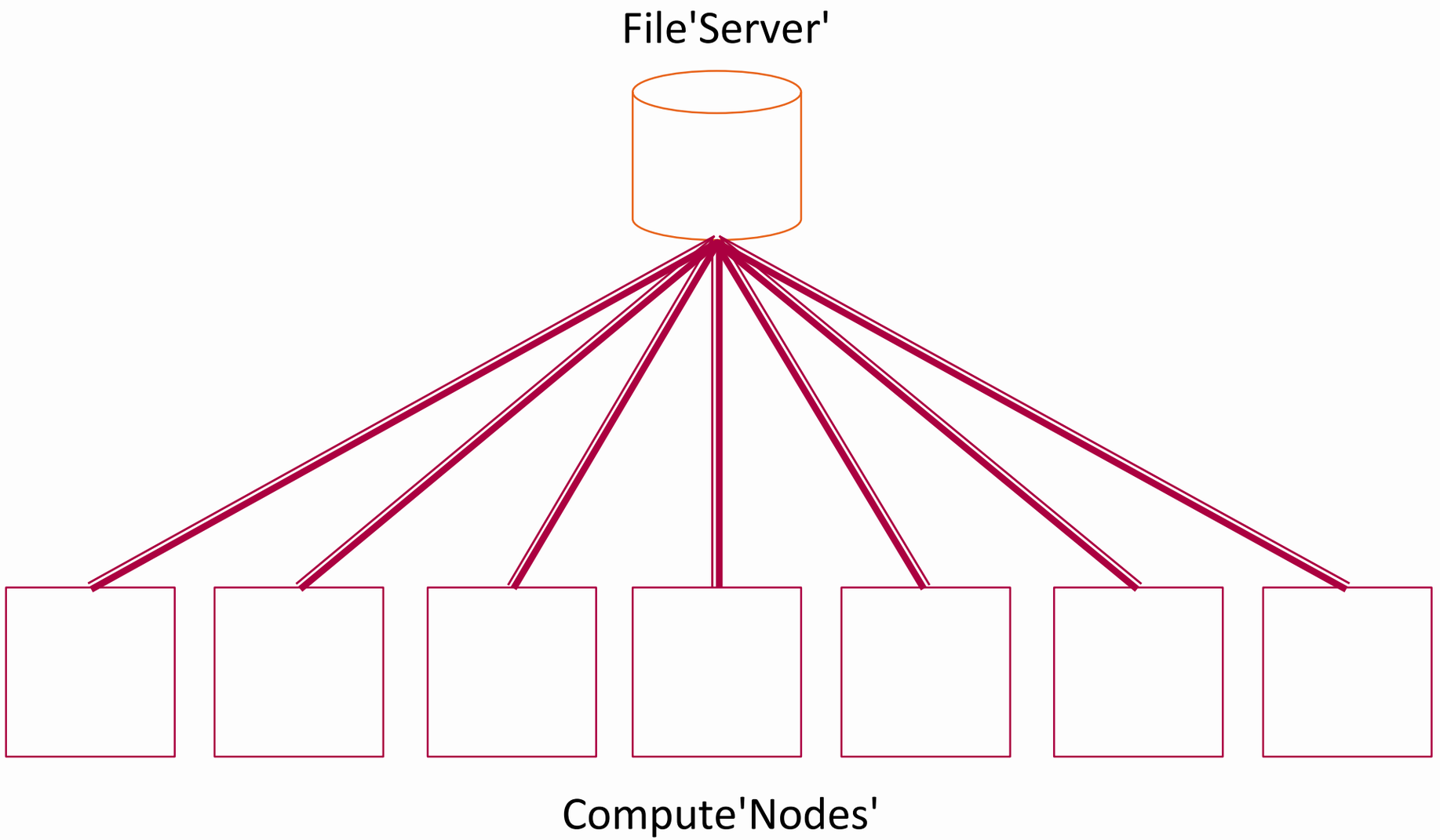
DYNAMIC SHARED LIBRARIES

The Problem

Sandia National Laboratories

- Demand for Dynamic Shared Library support is increasing
- Two methods of use:
  - Linked at program launch
  - dlopen() during runtime

The Naïve Approach



Current Solutions

Sandia National Laboratories

- There are a number of current solutions to this problem
  - Cache nodes
  - Bulk pre-distribution
  - Tree-based overlay networks
- These solutions look like peer-to-peer networks
  - Can we use an off-the-shelf peer-to-peer technology?

Sandia National Laboratories

BITTORRENT

Why BitTorrent?

Sandia National Laboratories

- One of the more popular peer-to-peer protocols
- Open-source implementations
- Private networks
- Extensible protocol
  - Peer Exchange
  - Magnet Links

Preparation

Sandia National Laboratories

- Tracker
  - Daemon listens for incoming peers
  - Maintains a list of peers
- Description File
  - User specified file or directory
  - Splits file(s) into smaller pieces
  - Lists tracker information
- Initial Seeder
  - Starts with a full set of data
  - Registers with tracker

Starting a Client

Sandia National Laboratories

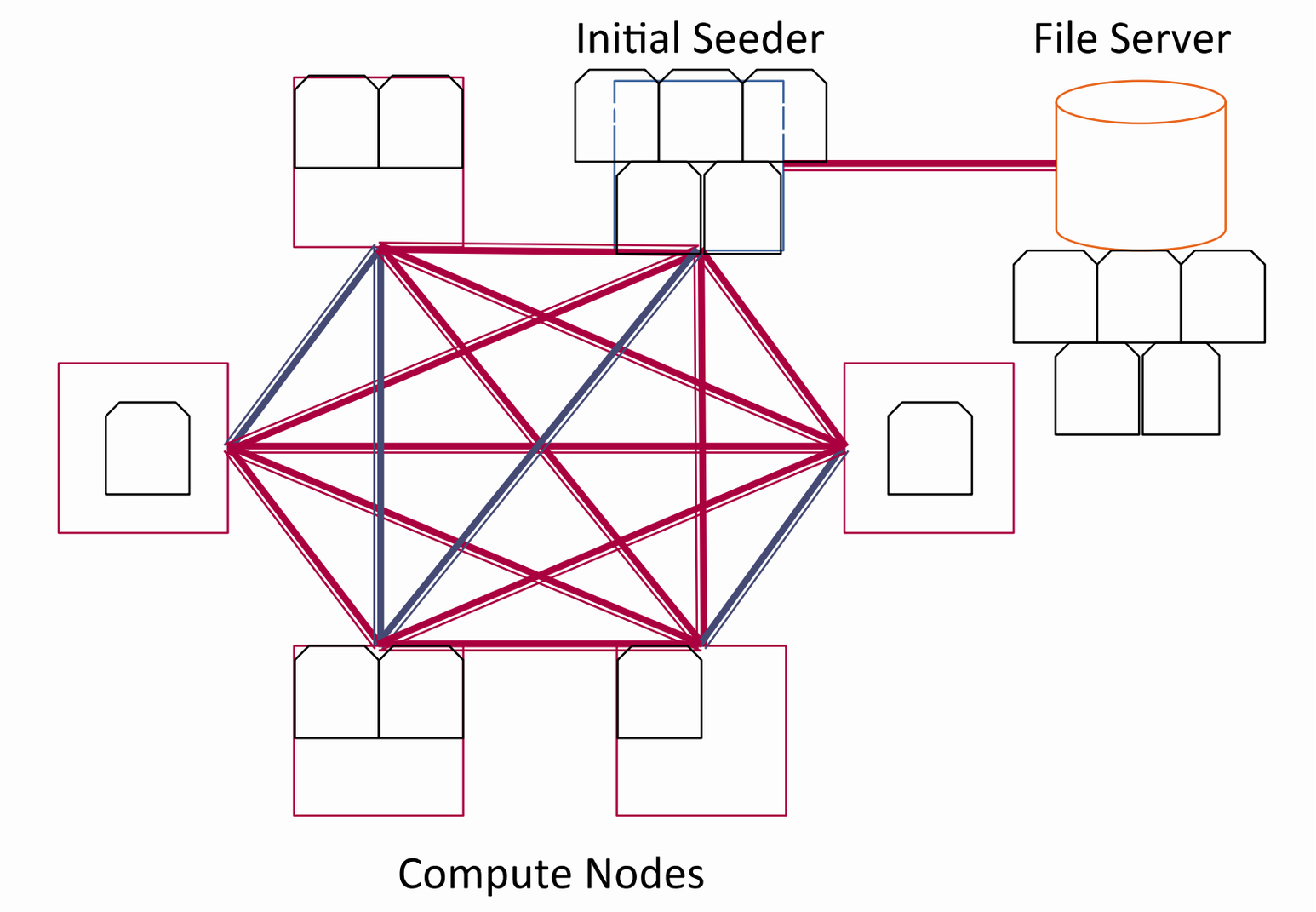
- Reads description file
- Connects to tracker
  - Registers as a peer
  - Gets a list of peers
- Connects to other peers

Data Distribution

Sandia National Laboratories

- Leechers
  - Keep a list of piece availability
  - Request pieces they don’t have
- Seeders and Leechers
  - Respond to requests
  - Distributes Data according to fairness algorithm
- Fairness algorithm
  - Leechers use Tit-for-tat
  - Seeders try to push complete copies

Data Distribution



Sandia National Laboratories

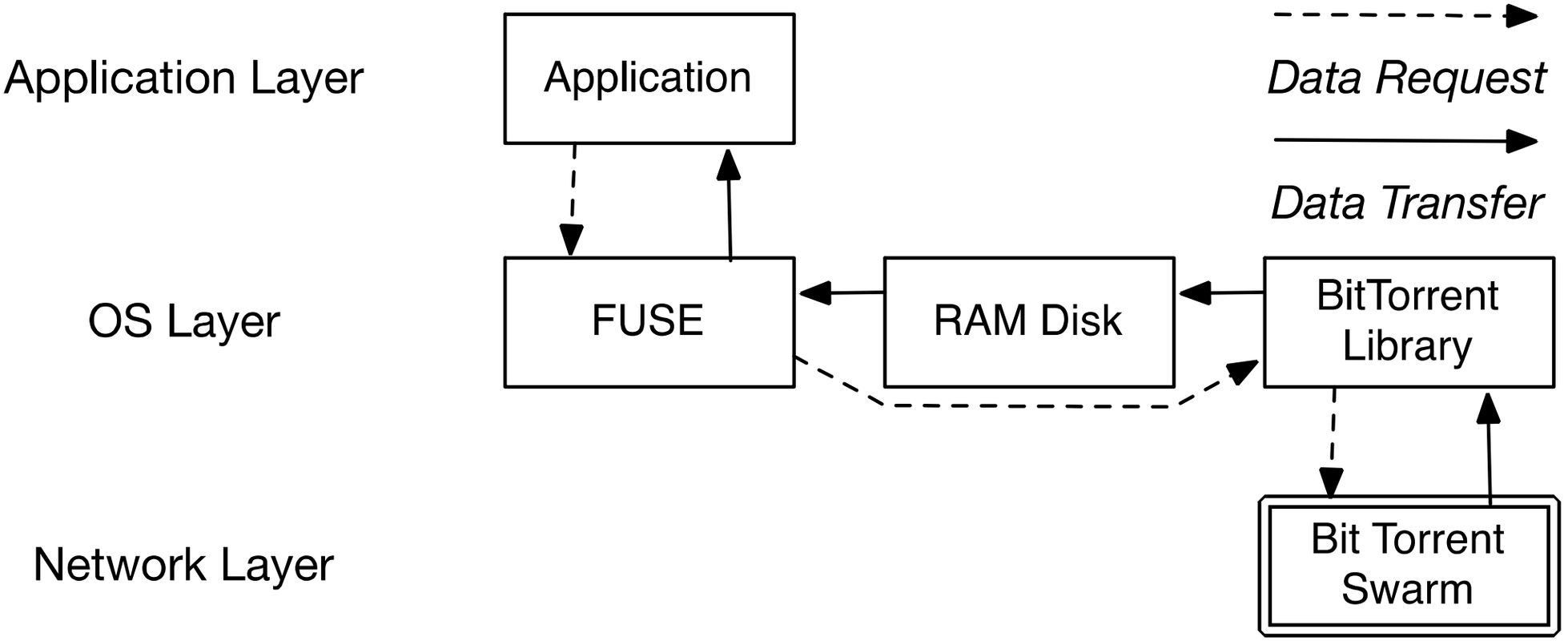
ARCHITECTURE

Added Actions

Sandia National Laboratories

- Before Job Launch
  - Create the description file
  - Start Tracker & Initial Seeder
- During Job Launch
  - Distribute the description file
  - On each node
    - Start FUSE client
    - Add directory to LD_LIBRARY_PATH
- During Runtime
  - DL requests first search the FUSE directory
  - If a file isn't downloaded yet, request it from the BitTorrent swarm
  - If a file is downloaded, pass through to a local RAMDisk

Data Request Path



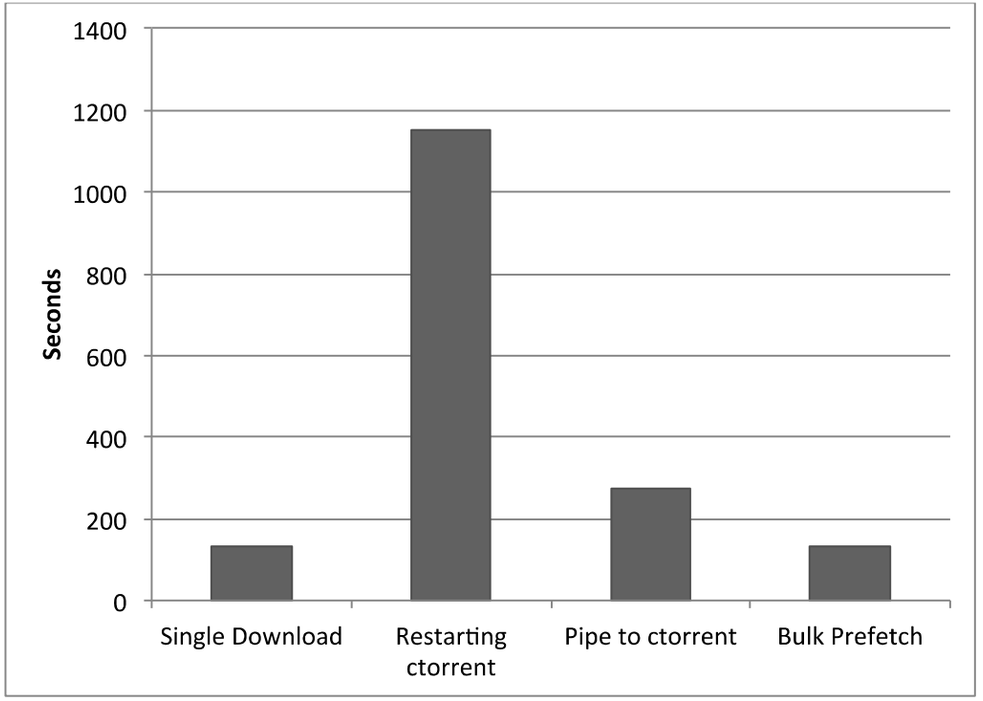
Sandia National Laboratories

MAKING BITTORRENT WORK FOR HPC

BitTorrent Assumptions Get in the Way

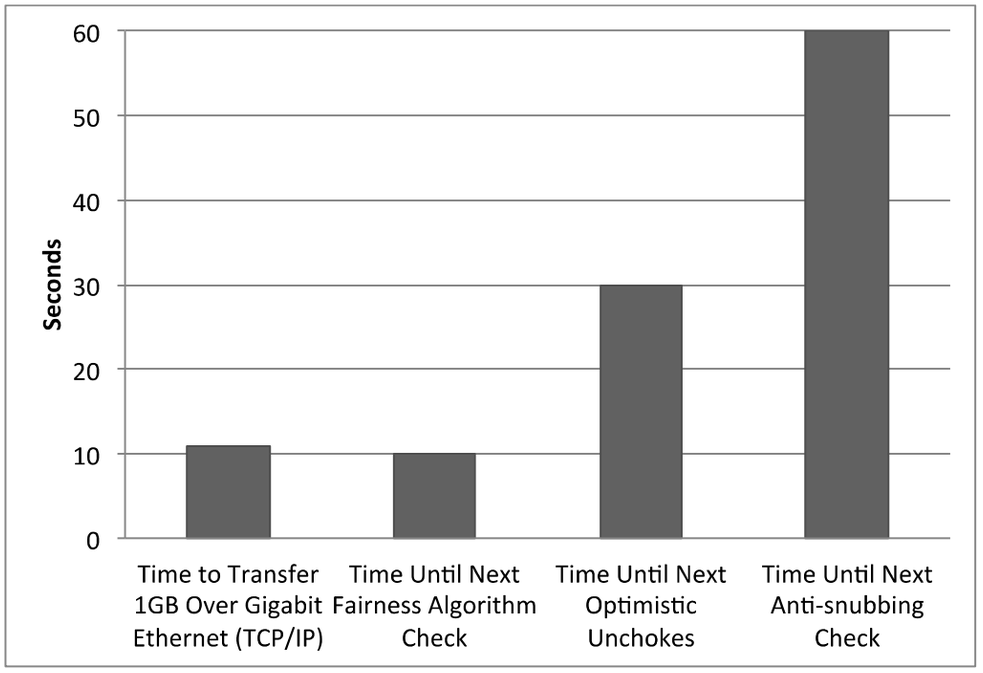
Sandia National Laboratories

- Basic approach: Downloads individual files on-demand
- Problem: Violates basic BitTorrent usage assumption



Fairness Algorithm Problems

- BitTorrent tries to tolerate adversarial peers
- Fairness algorithm tuned for Internet speeds and timescales
- Sequential requests interfere with tit-for-tat choking algorithm



Working Around Fairness

Sandia National Laboratories

- Solution: Bulk prefetching of launch loaded libraries
- Libraries can be loaded two ways
  - On launch
  - On demand (dlopen)
- List of launch-loaded libraries provided by ldd
- Downloading these together, we match BitTorrent's use case
- We still support on-demand requests

Sandia National Laboratories

RESULTS

Experimental Setup

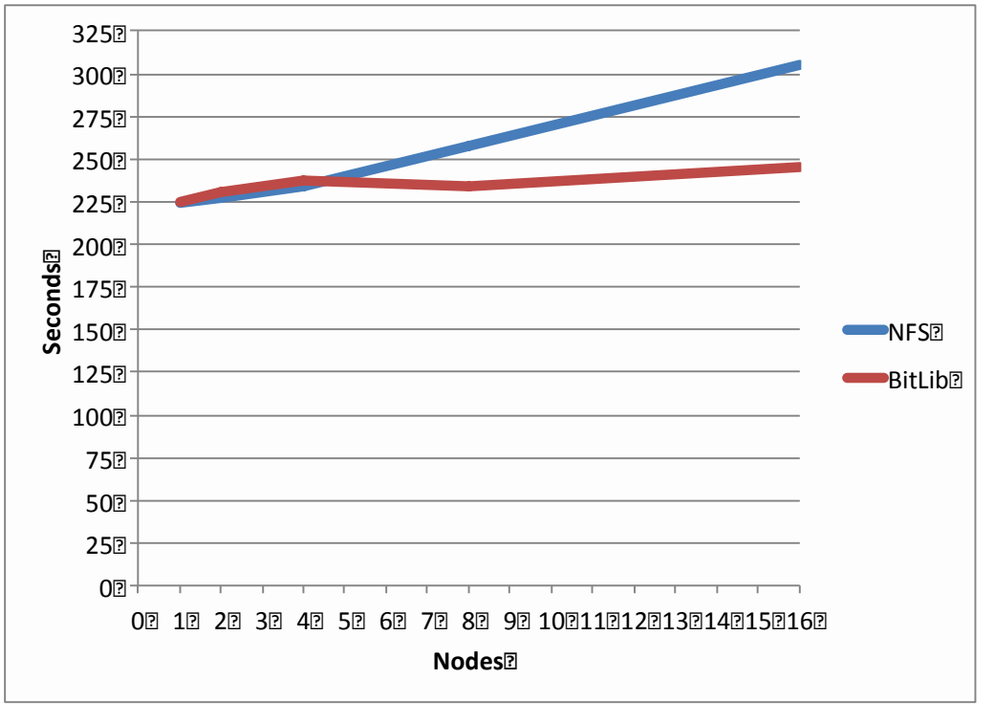
Sandia National Laboratories

- The Pynamic benchmark
  - Test case has 495 libraries totaling ~1.1GB
  - Test case based on an LLNL scientific code
- Three Machines
  - Teller – 100 node, AMD Fusion with Infiniband
  - Muzia – 20 node, Cray XE6
  - Cielo - 8,944 Node, Cray XE6

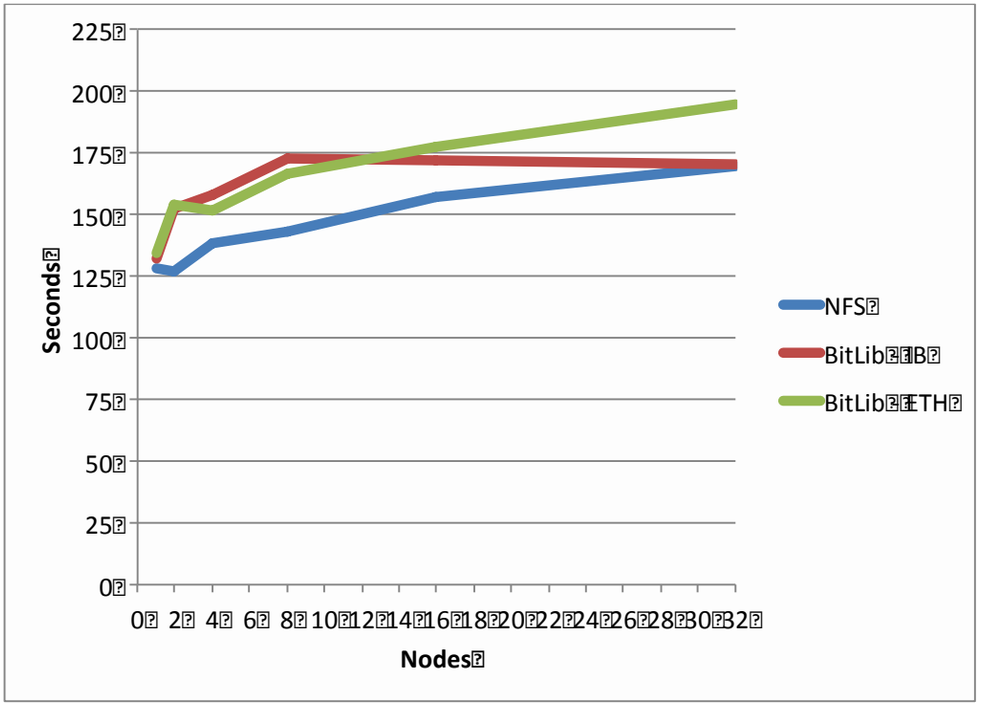
Small Scale

Sandia National Laboratories

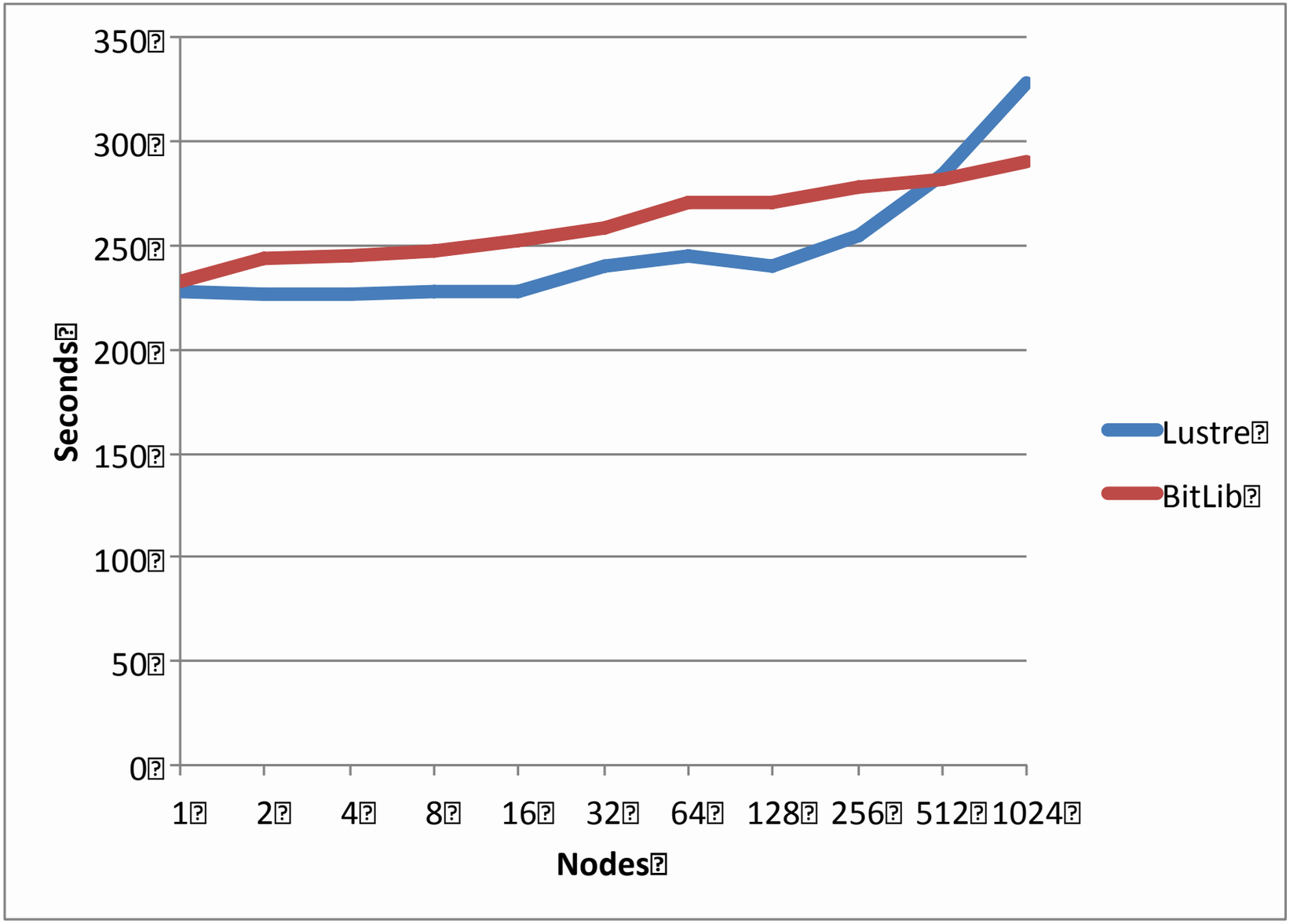
Muzia: a Gemini system



Teller: an Infiniband system



Cielo


350
300
250
200
150
100
50
0
Seconds
1
2
4
8
16
32
64
128
256
512
1024
Nodes
Lustre
BitLib

Sandia National Laboratories

ISSUES AND FUTURE WORK

Network Features

Sandia National Laboratories

- Can we use HPC network features to enhance and optimize BitTorrent?
- Multicast
- Topology-aware
- Multiple-paths

Acknowledgements

This work was funded by NNSA's Advanced Simulation and Computing (ASC) Program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE- AC04-94AL85000.

Sandia National Laboratories

QUESTIONS?