

Lawrence Livermore National Laboratory

Installation of LLNL's Sequoia File System



Marc Stearman

Parallel File Systems Operations Lead

marc@llnl.gov

April 2012

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

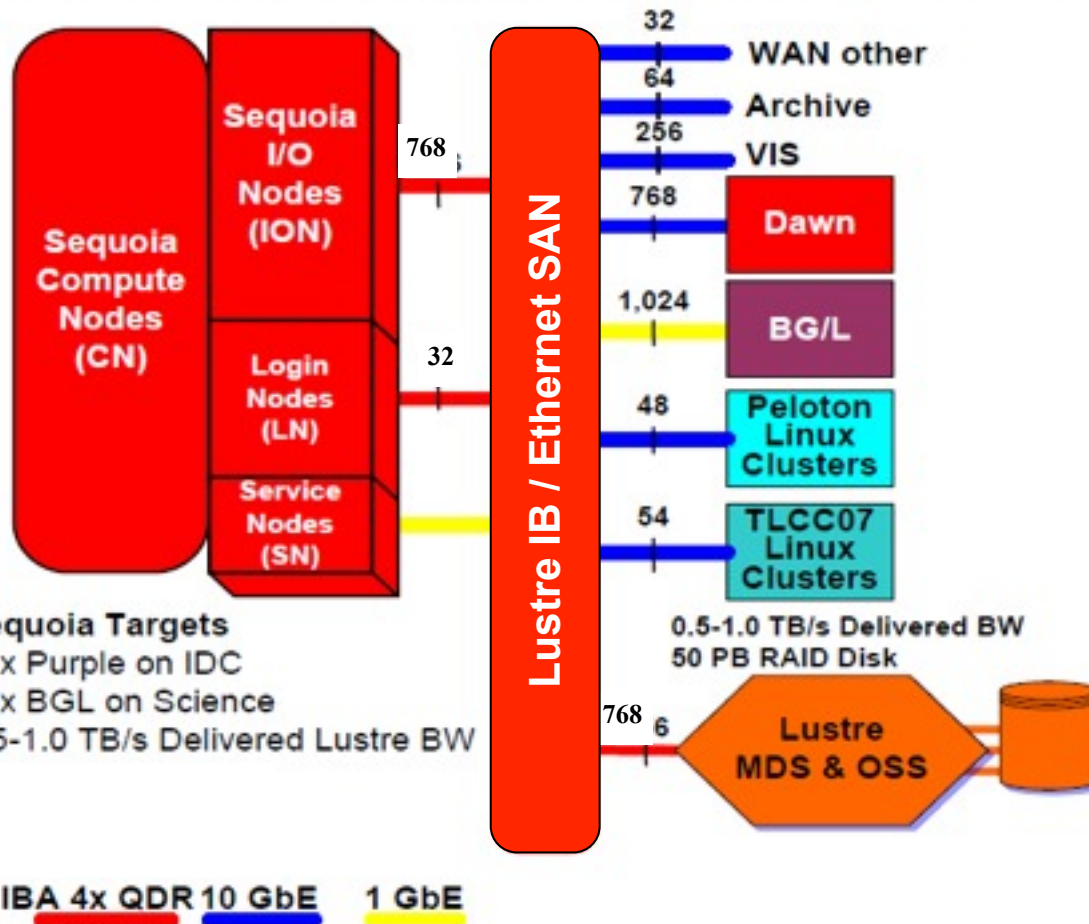
LLNL-PRES-551092

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Sequoia Compute Platform

ASC Sequoia Simulation Environment Lawrence Livermore National Laboratory 2011/12



Sequoia Stats:

- 20PF Compute Platform
- 96 Racks
- 1.6 Million Cores
- 1.6 PB Memory
- 768 I/O Nodes - QDR IB
- Liquid Cooled



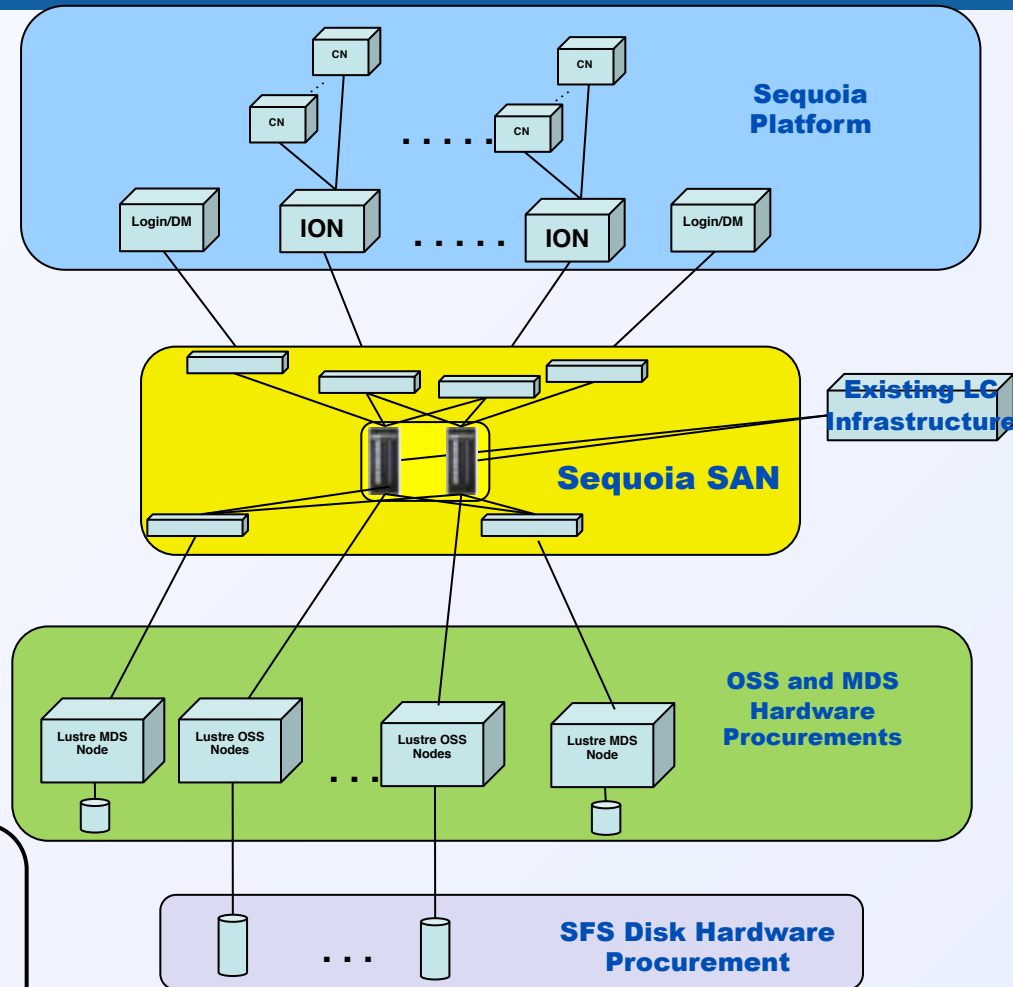
Sequoia I/O Infrastructure

Requirements

- 50PB file system
- 500GB/s minimum, 1TB/s stretch goal
- QDR InfiniBand SAN connection to Sequoia
- Must integrate with existing Ethernet infrastructure

>\$20M Budget

- Across five procurements dominated by RAID file system and IB SAN hardware procurements



Phased Bandwidth Delivery

- Phase 1: 10% Oct 2011
- Phase 2: 50% Dec 2011
- Phase 3: 100% Feb 2012



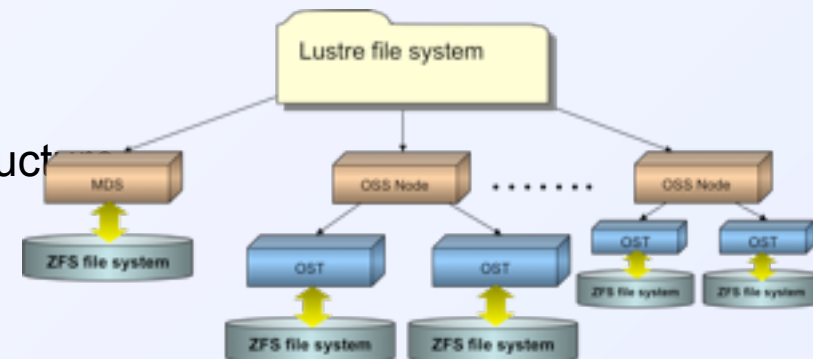
Sequoia I/O Challenges Aboard

- **ZFS-based Lustre**

- Has never been done before
- Is dependent on ongoing local development and D&E contract investments
- Is the pioneering implementation of new backend fs for Lustre community
- Will be buggy, does not meet 1TB/s stretch goal without performance improvements

- **InfiniBand SAN**

- Lack of tools and experience as HPC SAN
- Need to bridge to existing Ethernet infrastructure



- **Sequoia IONs**

- ION/CN ratio is a factor of two less than Dawn and BG/L

D&E Efforts

■ Lustre OSD work

- Abstracts all backend storage access into a single portable API
- Enables Lustre to use any backend file system with a minimum amount of “glue” logic

 – ldiskfs

– ZFS

– btrfs (future work) 

■ SMP Checksum Performance

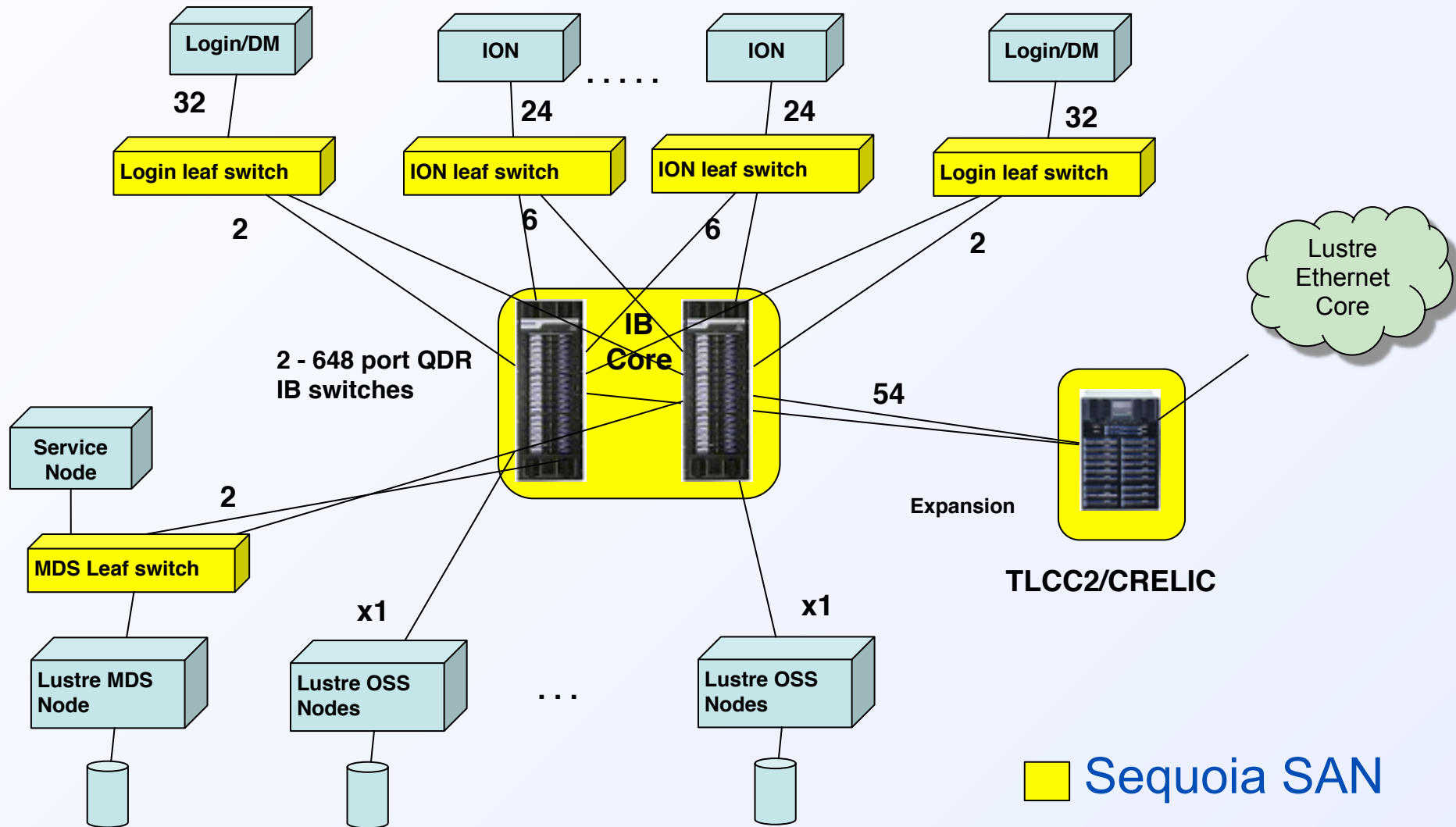
- Parallelize checksums across multiple threads

■ ZFS Optimization

■ Quotas



Sequoia SAN Architecture



Procurement Status

- **RAID Hardware**
 - Contract Awarded to IAS/NetApp

- **SAN Infiniband Hardware**
 - Contract Awarded to Advanced HPC/Mellanox

- **OSS**
 - OSS Contract Awarded to Appro

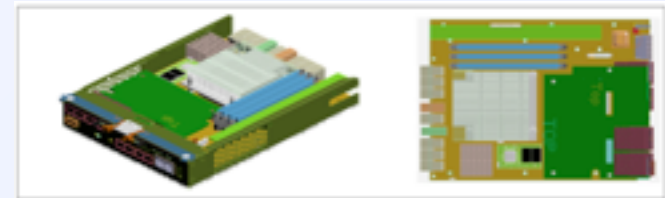
- **MDS**
 - Supermicro Westmere nodes, with RAID Inc. JBOD and OCZ Talos2 SSDs

- **Sequoia Platform**
 - All racks at LLNL, integration in progress



RAID Hardware Details

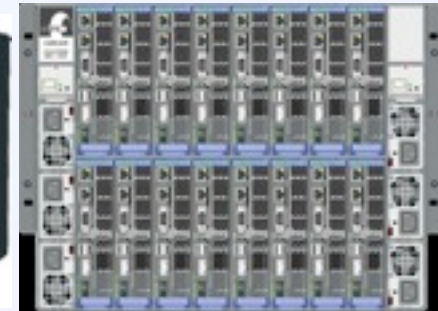
- Contract Awarded to NetApp (LSI Engenio equipment)
- NetApp E5400
 - 60-bay 4U Enclosure with 2 RAID controllers
 - 3TB SAS drives
 - 180TB RAW capacity
 - 130TB RAID6 capacity
 - 6Gb SAS lanes
 - FC or IB Host interfaces (We chose QDR IB)



Lustre Server Architecture

- **OSS uses TLCC2 design**

- Appro GreenBlade
- Intel Xeon E5-2670 @ 2.60GHz
- Dual Socket, 8 core
- 64GB RAM
- QDR Mellanox ConnectX-3 IB down (LNET)
- Dual Port QDR ConnectX-2 HCA (SRP to Disk)



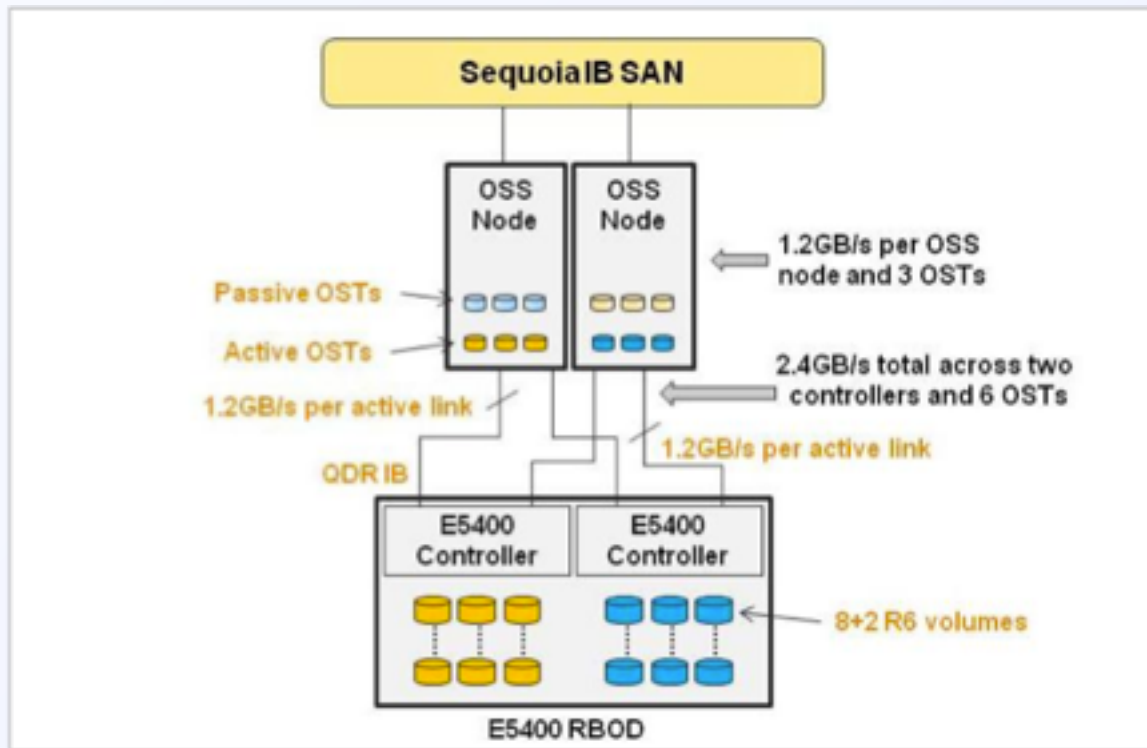
- **MDS**

- Supermicro X8DTH
- Intel Xeon X5690 @ 3.47GHz
- Dual Socket, 6 core (24 cpus with Hyperthreading)
- 192GB RAM
- JBODS with OCZ Talos2 SDDs (40 Drives, SAS connected using ZFS RAID10)
- Configure as a failover pair (active/passive) for reliability



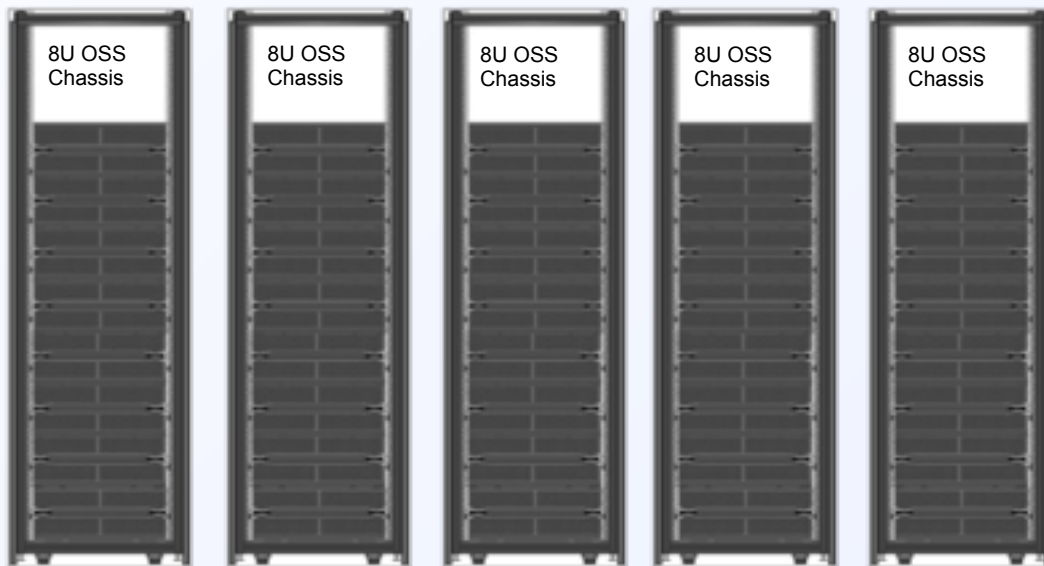
OSS Connectivity

- 2 OSS nodes per E5400 as a failover pair
- Using RHEL6 multipath rdac drivers



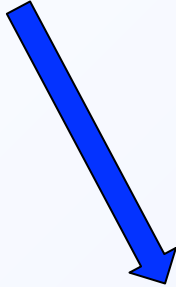
Rack Layout

- 8 E5400s per RSSU (Rack Storage Scalable Unit)
- 16 OSS nodes at the top of the rack
- 48 Racks total: 384 E5400s, 768 OSS nodes
- 55PB Capacity, aiming for 1TB/s

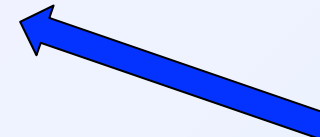


Performance – XDD

Seagate Drives



Hitachi Drives



- Aggregate 463 GB/s
- Average Node: 615 MB/s



Performance – XDD, Post Drive Swap

Performance – XDD

Seagate Drives



Hitachi Drives



- Aggregate 1.2 TB/s (752 nodes)
- Average Node: 1.5 GB/s



Performance – XDD, Post Drive Swap



Performance – ZPIOS

- Aggregate: Write 693 GB/s, Read 603 GB/s (616 nodes)
- Projected: Write 864 GB/s, Read 752 GB/s (768 nodes)
- Average Node: Write 1,125 MB/s, Read 979 MB/s

Performance – ZPIOS, Post Drive Swap

- Aggregate: Write 747 GB/s, Read 689 GB/s (624 nodes)
- Projected: Write 919 GB/s, Read 848 GB/s (768 nodes)
- Average Node: Write 1197 MB/s, Read 1104 MB/s



Time Lapse Build



Time Lapse Build



Questions?

