Deploying Hadoop on Lustre Storage:

Lessons learned and best practices.

J. Mario Gallegos- Dell, Zhiqi Tao- Intel and Quy Ta- Dell

Apr 15 2015

Integration of HPC workloads and Big Data workloads on Tulane's new Cypress HPC cluster

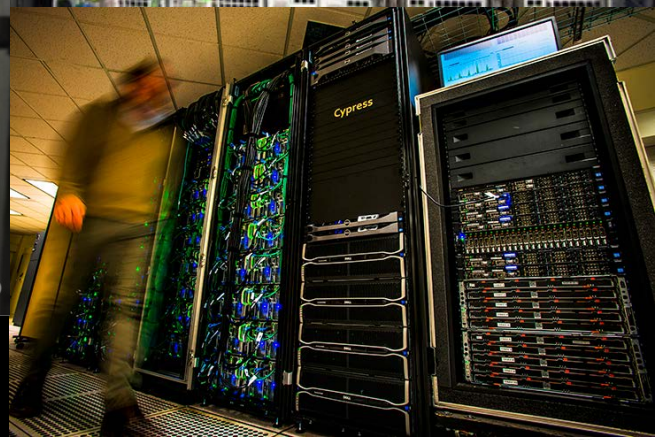
- Big Data problems are large, HPC resources can help
  - HPC Brings more compute power
  - HPC storage can provide more capacity and data throughput
- **Requisites/constraints**
  - HPC Compute nodes local storage space only for OS.
  - Integration with non-Hadoop applications and storage
    - Need something that is compatible with POSIX
  - Map-Reduce should access storage efficiently
  - Fast storage access(40GbE) and fully supported
  - Solution had to be designed, integrated, deployed and tested in 7 weekdays to be ready for SC14

Tulane University Cypress HPC hybrid cluster

RANKING

Tulane University

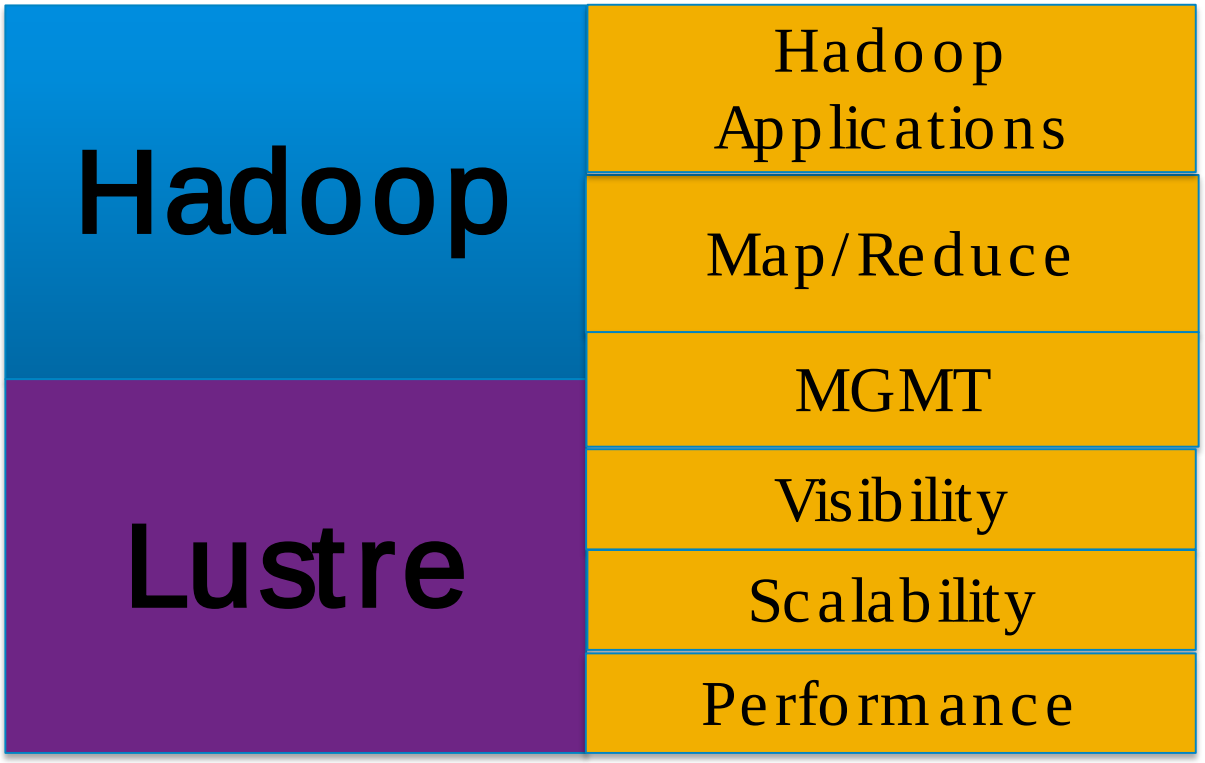
LIST	RANK	SYSTEM	VENDOR	TOTAL CORES	RMAX (TFLOPS)	RPEAK (TFLOPS)	POWER (KW)
11/2014	271	Dell C8220X Cluster, Intel Xeon E5-2680v2 10C 2.8GHz, 40G Ethernet, Intel Xeon Phi 7120P	Dell	17,608	221.0	355.1	87.00



Why Hadoop Map/Reduce on Lustre?

Effective Data Processing

High Performance Storage



Contrasting HDFS vs. Lustre

Hadoop = MapReduce + HDFS

LUSTRE

- Computations share the data
- Uses LNET to access data
- No data replication (uses RAID)
- Centralized storage
- POSIX compliant
- Widely used for HPC applications
- All data available at all time.
- Allows backup/recovery using existing infrastructure (HSM).

HDFS

- Data moves to the computation
- Uses HTTP for moving data
- Data replication (3X typical)
- Local storage
- Non-POSIX Compliant
- Widely used for MR applications
- Used during shuffle MR phase

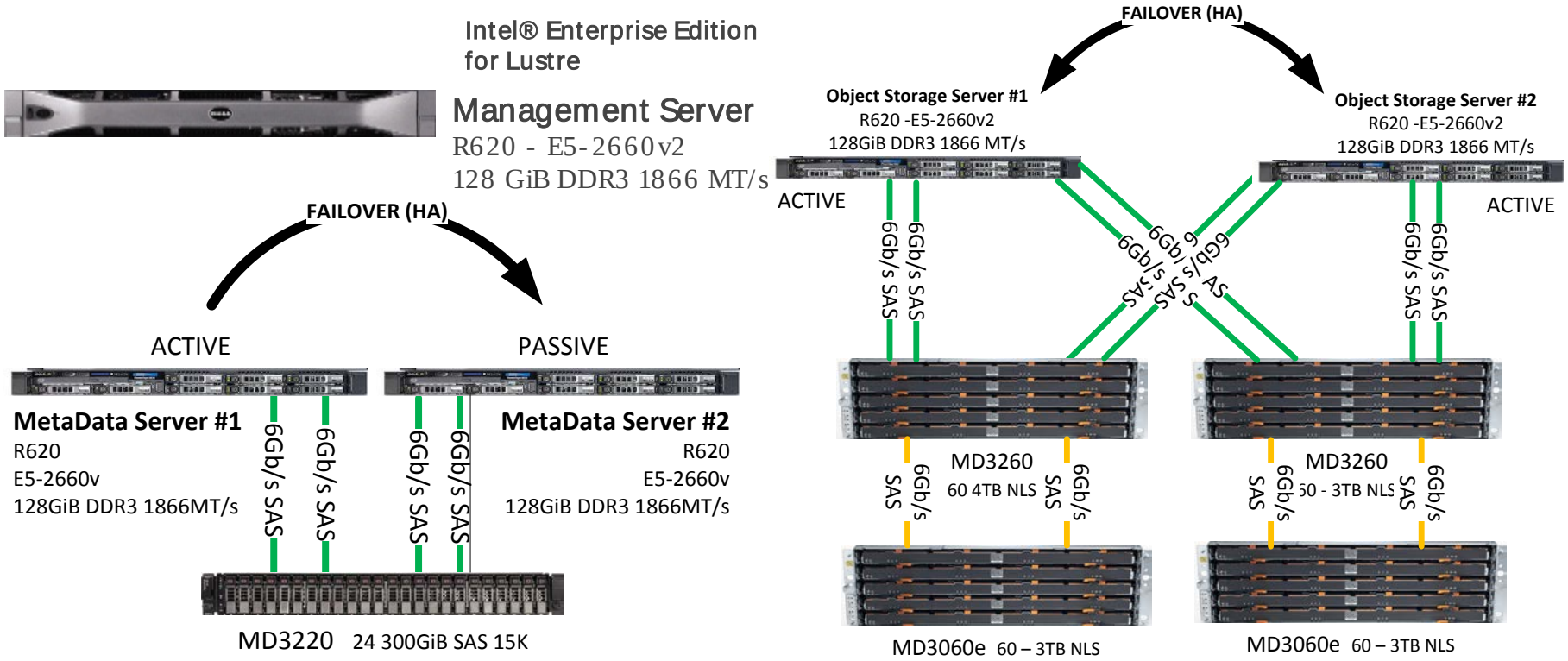
Integration of HPC workloads and Big Data workloads on Tulane's new HPC cluster

- MapReduce workloads added to HPC
  - Unknown mix of Hadoop & HPC applications from multiple fields, changing over time
  - Hadoop workloads run differently than typical HPC applications
    - Hadoop and HPC applications use different schedulers
    - Need to run MapReduce workloads just like HPC workloads
- **More requisites/constraints**
  - Train or hire admins to manage Hadoop
  - Lustre also require training or hiring admins
  - OR simplify/automate/assist deployment, administration and monitoring to avoid/minimize training/hiring.

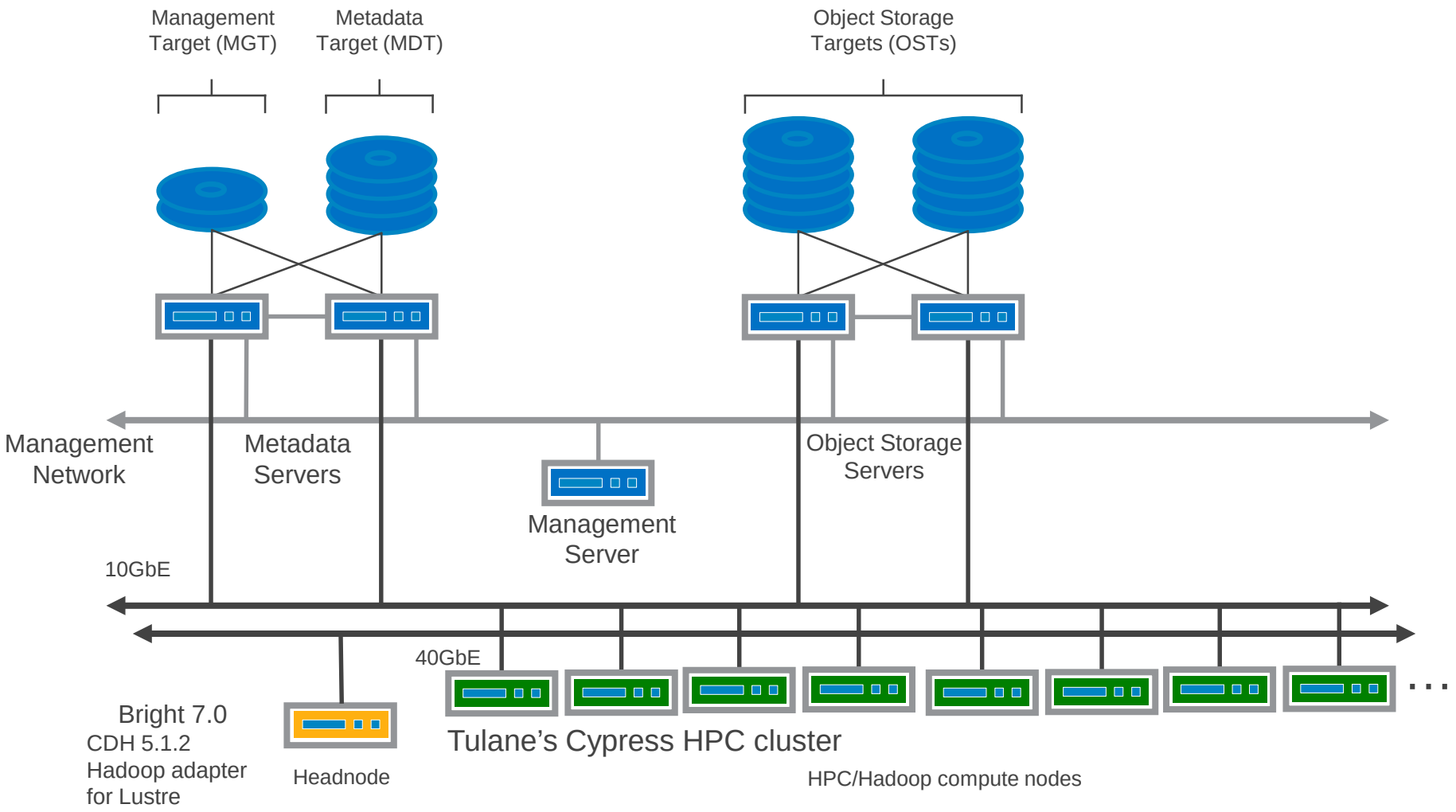
Design and initial implementation

- Rely on partners and existing products to meet deadline
- Reuse existing Dell Lustre based storage solutions
  - A Lustre training system was appropriated for the POC
  - Storage System fully racked/deployed/tested and shipped
  - Plan was: connect power and 40 GbE and add Hadoop onsite
  - A box with all types cables, spare adapters, other adapters and tools was included
- Intel was chosen as the partner for Lustre
  - Intel EE for Lustre was already used in a Lustre solution
  - Included the components needed to replace HDFS
  - A replacement for YARN was in the making
- Bright was chosen for cluster/storage management
  - Already used to deploy/manage/monitor Cypress HPC cluster
  - Supports Cloudera distribution of Hadoop
  - Supports deployment of Lustre clients on Dell's clusters

Proof Of Concept components



Proof of Concept Setup



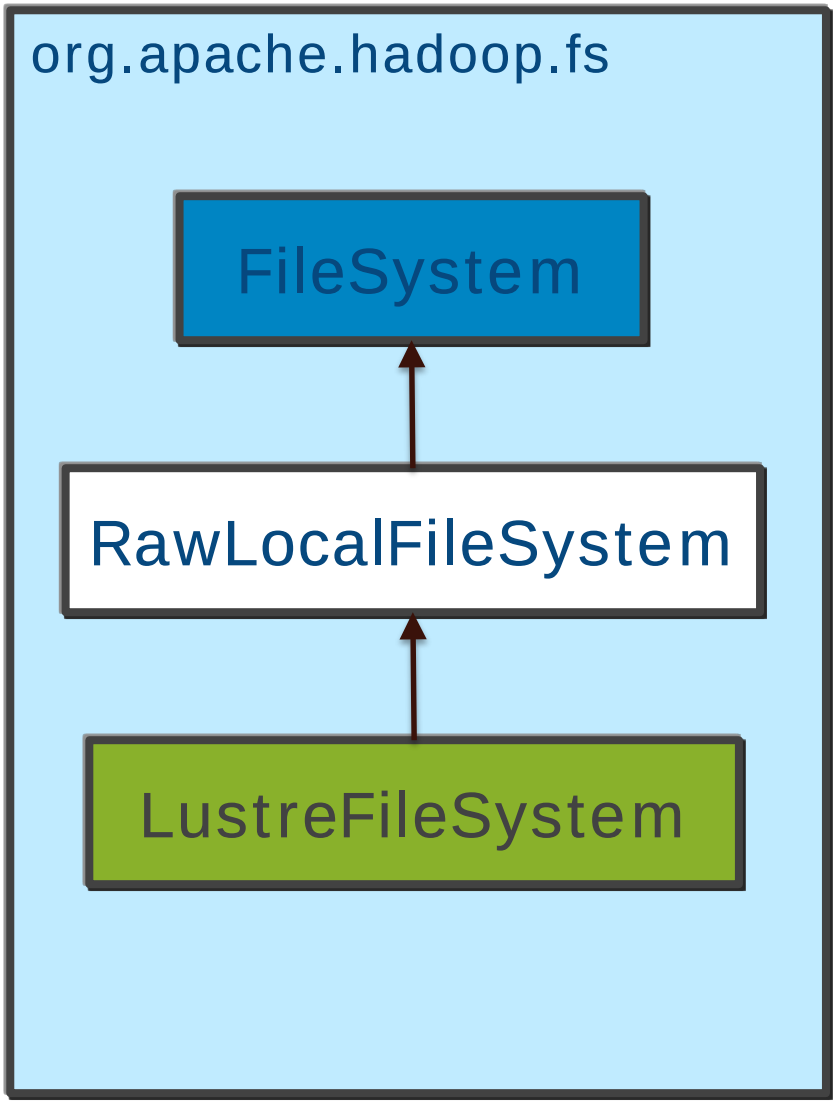
Hadoop Adapter for Lustre - HAL

- Replaces HDFS
- Based on the Hadoop architecture
- Packaged as a single Java library (JAR)
  - Classes for accessing data on Lustre in a Hadoop compliant manner. Users can configure Lustre Striping.
  - Classes for “Null Shuffle”, i.e., shuffle with zero-copy
- Easily deployable with minimal changes in Hadoop configuration
- No change in the way jobs are submitted
- Part of Intel Enterprise Edition for Lustre

HPC Adapter for MapReduce - HAM

- Replaces YARN (Yet Another Resource Negotiator)
- SLURM based (Simple Linux Utility for Resource Management)
  - Widely used open source resource manager
- Objectives
  - No modifications to Hadoop or its APIs
  - Enable all Hadoop applications to execute without modification
  - Maintain license separation
  - Fully and transparently share HPC resources
  - Improve performance
- New to Intel Enterprise Edition for Lustre

Hadoop MapReduce on Lustre. How?



- Use Hadoop's built-in LocalFileSystem class to add the Lustre file system support
  - Native file system support in Java
- Extend and override default behavior: LustreFileSystem
- Defined new URL scheme for Lustre, i.e. `lustre:///`
- Controls Lustre striping info
- Resolves absolute paths to user-defined directories
- Optimize the shuffle phase
- Performance improvement

Hadoop Nodes Management

- Bright 7.0 was used to provision the 124 Dell C8220X compute nodes from bare metal
- Used to deploy CDH 5.1.2 onto eight Hadoop nodes, not running HDFS
- The Lustre clients and IEEL plug-in for Hadoop were deployed by Bright on the Hadoop nodes
- The Hadoop nodes could access Lustre storage
- Different mix of Hadoop or HPC nodes can be deployed on demand using Bright

On Site work and challenges

- Dell, Intel and Bright send people onsite for the last 5 days and allocated resources to support remotely
- Rack with POC system took longer to arrive than planned
- Location assigned to POC system was too far from HPC cluster than anticipated, 40 GbE cables were too short
  - New requirement: Relocation of POC system closer to the cluster
- Z9500 40 GbE switch only had 1 open port. POC system needed four
  - A breakout cable and four 10 GbE cables were used
  - Z9500 port had to be configured for breakout cabling
  - IB FDR/40 GbE Mellanox adapters were replaced by 10 GbE
- After connecting the POC system, one of the servers had an iDRAC failure (everything was working during the initial testing in our lab
  - iDRAC is required for HA, to enforce node eviction
  - A replacement system was expedited, but ETA was one day
  - We decided to swap the management server and affected system, since all servers were R620s, same CPU/RAM/local disk
  - POC system was redeployed from scratch (OS and Intel software)

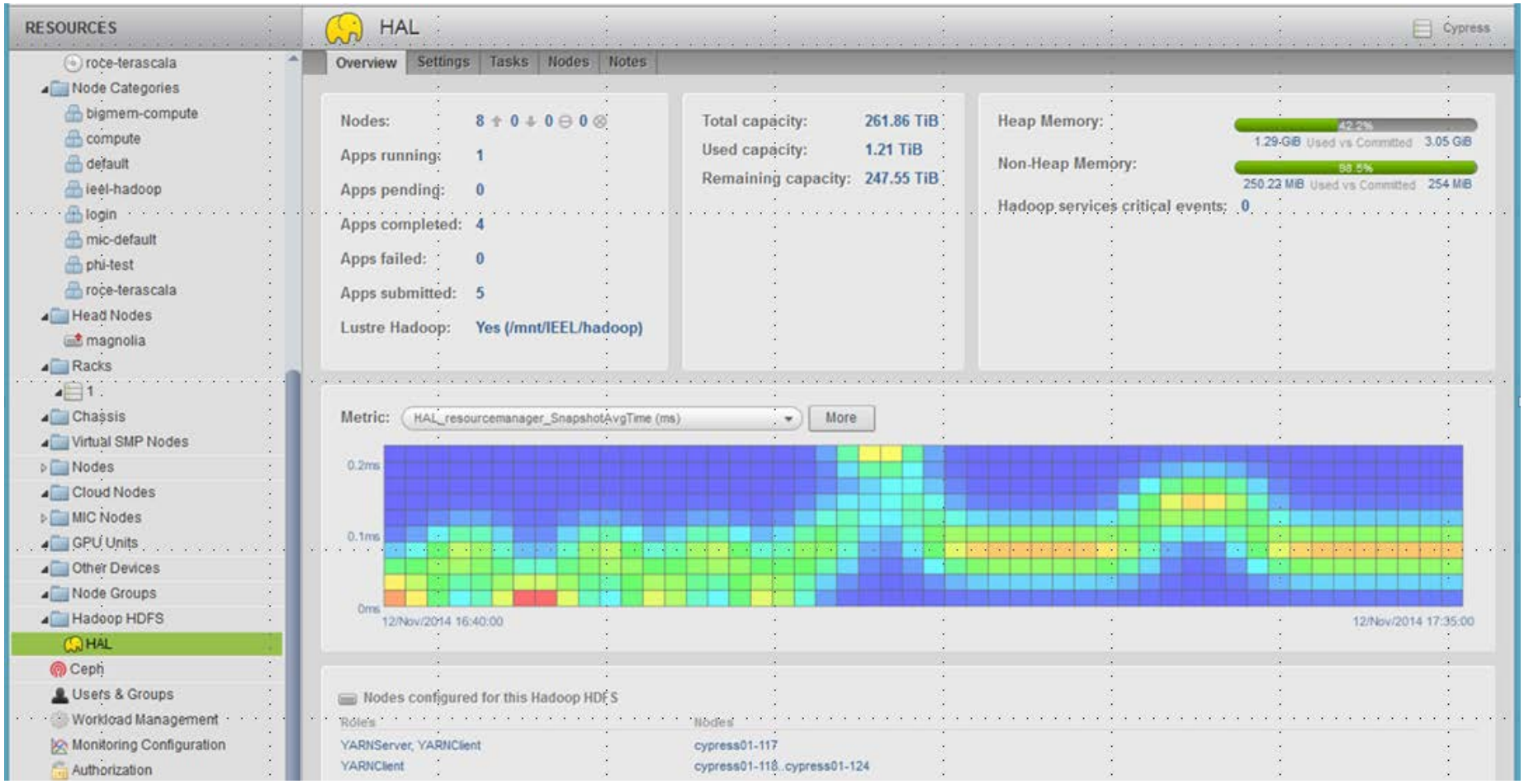
On Site work and challenges

- Local and remote resources started working immediately on Hadoop integration and client deployment, testing Lustre concurrently.
- Since POC had to use only 10 GbE, performance was limited to a maximum of 2 GB/s (2 OSS links).
- Focus changed to demonstrating feasibility, capabilities and GUI monitoring, but performance was also reported.
- All on site work was completed in four days, with last day to spare.
- Tulane University Execs and Staff support was invaluable to adjust to initial problems and expedite their solution.

POC Performance

- Performance was reported without tuning; there was plenty of room for improvement. Starting by using 40 GbE as originally planned.
- DFSio reported an average of 134.91 MB/sec (12 files, 122880 MB).
- Lustre reported max aggregated read/write throughput: 1.6 GB/s.
- TeraSort on 100 Million records:
  - Total time spent by all maps in occupied slots (ms)=341384
  - Total time spent by all reduces in occupied slots (ms)=183595
  - Total time spent by all map tasks (ms)=341384
  - Total time spent by all reduce tasks (ms)=183595
  - Total vcore-seconds taken by all map tasks=341384
  - Total vcore-seconds taken by all reduce tasks=183595
  - Total megabyte-seconds taken by all map tasks=349577216
  - Total megabyte-seconds taken by all reduce tasks=188001280

Hadoop on Lustre Monitoring



Lessons learned and best practices

Keeping in mind the extremely aggressive schedule:

- Allocating in advance the right and enough resources was key under time constraints.
- Without our partners full support, this POC could not happened.
- Reuse of existing technology (solution components) and equipment was crucial.
- Gather in advance all possible information about existing equipment, power and networks on site, as detailed as possible.
- Get detailed agreement in advance about location for new systems and their requirements. Use figures and photographs to clarify everything.
- Murphy never takes a brake and hardware may fail at any time, even if it was just tested.

Lessons learned and best practices

- Get in advance user accounts, lab access, remote access, etc. and if possible, get Lab decision-makers aware of work.
- Conference calls with remote access to show and sync all parties involved is a necessity.
- Under large time zones disparity, flexibility is gold.
- Have enough spare parts: adapters, hard disks, cables and tools… but you may need more.
- Using interchangeable servers proved to be an excellent choice.
- Division of assignments by competence and remote support: priceless.
- Single chain of command for assignments and frequent status updates were very valuable.
- Communicate, communicate, communicate!

Acknowledgements

Dell:

Joe Stanfield
Onur Celebioglu
Nishanth Dandapanthu
Jim Valdes
Brent Strickland.

Intel:

Karl Cain

Bright Computing:

Robert Stober
Michele Lamarca
Martijn de Vries

Tulane:

Charlie McMahon
Leo Tran
Tim Deeves
Olivia Mitchell
Tim Riley
Michael Lambert

Questions?

Thank you!

www.dellhpcsolutions.com
www.hpcatdell.com