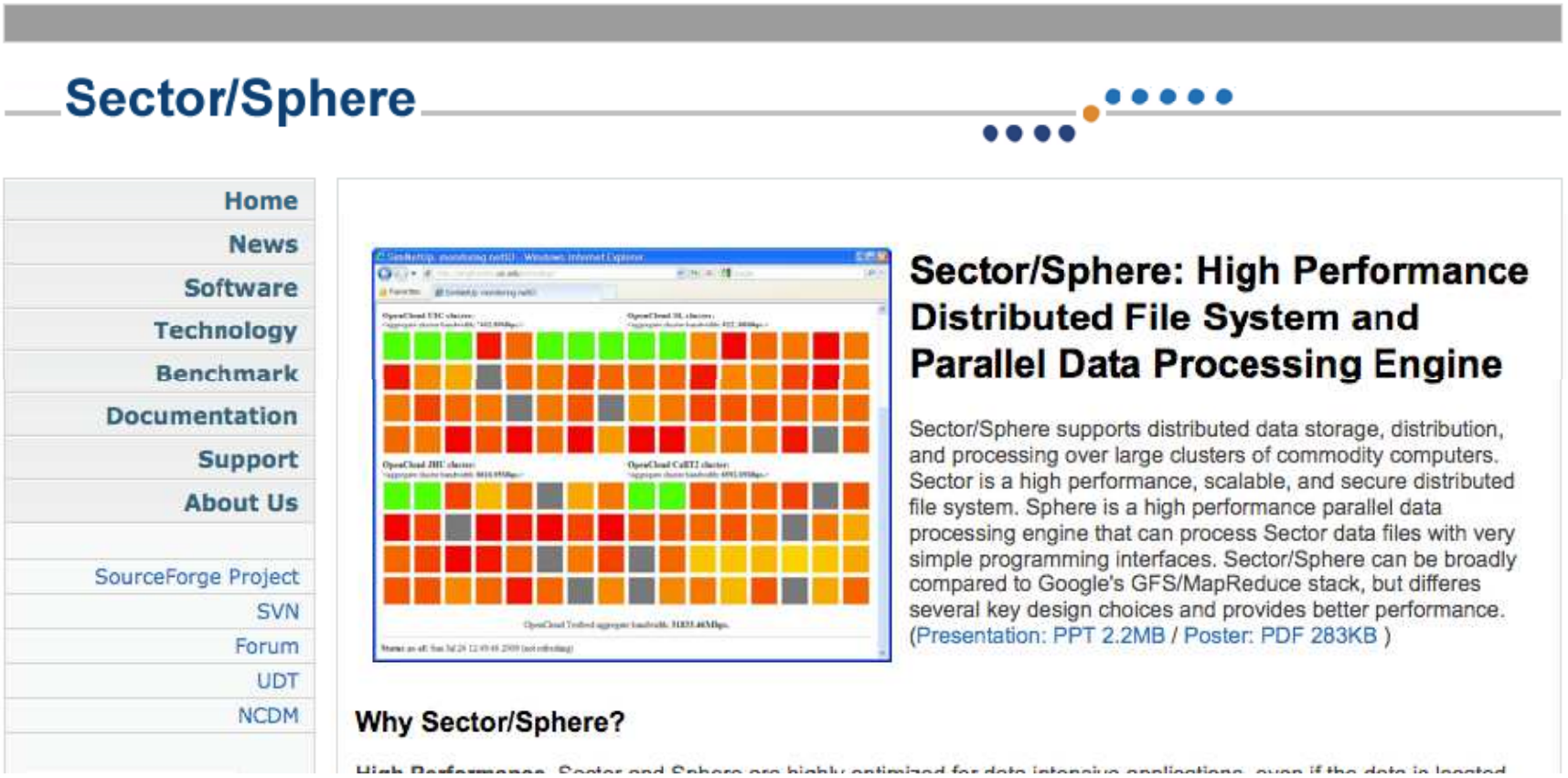
Lessons Learned From a Year's Worth of Benchmarks of Large Data Clouds

Yunhong Gu and Robert L Grossman

Laboratory for Advanced Computing
University of Illinois at Chicago

November 16, 2009

Part 1. Overview of Sector

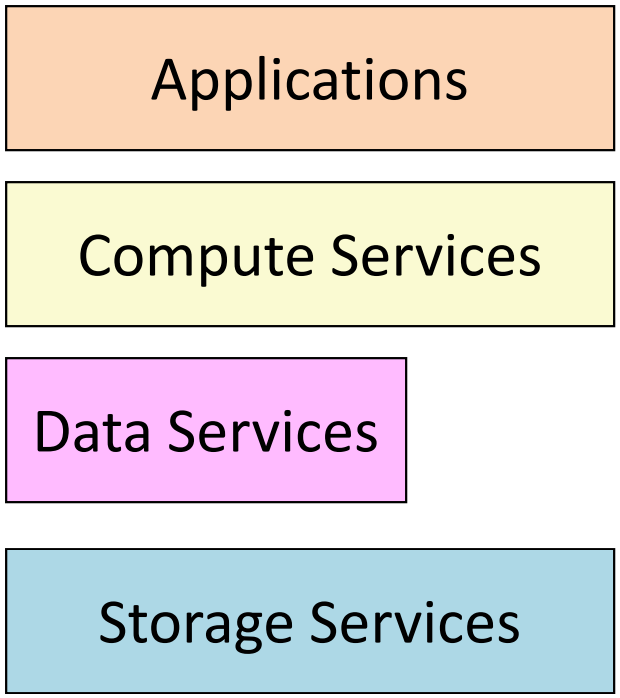


http://sector.sourceforge.net

Sector Overview

- Sector is fastest open source large data cloud
  - As measured by MalStone & Terasort
- Sector is easy to program
  - UDFs, MapReduce & Python over streams
- Sector is secure
  - A HIPAA compliant Sector cloud is being launched
- Sector is reliable
  - Sector v1.24 supports multiple active master node servers

Google's Large Data Cloud



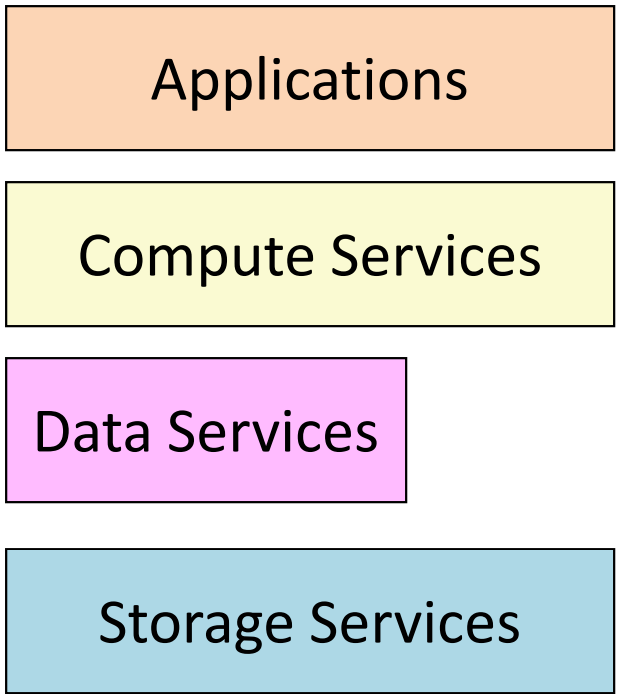
Google's MapReduce

Google's BigTable

Google File System (GFS)

Google's Stack

Hadoop's Large Data Cloud

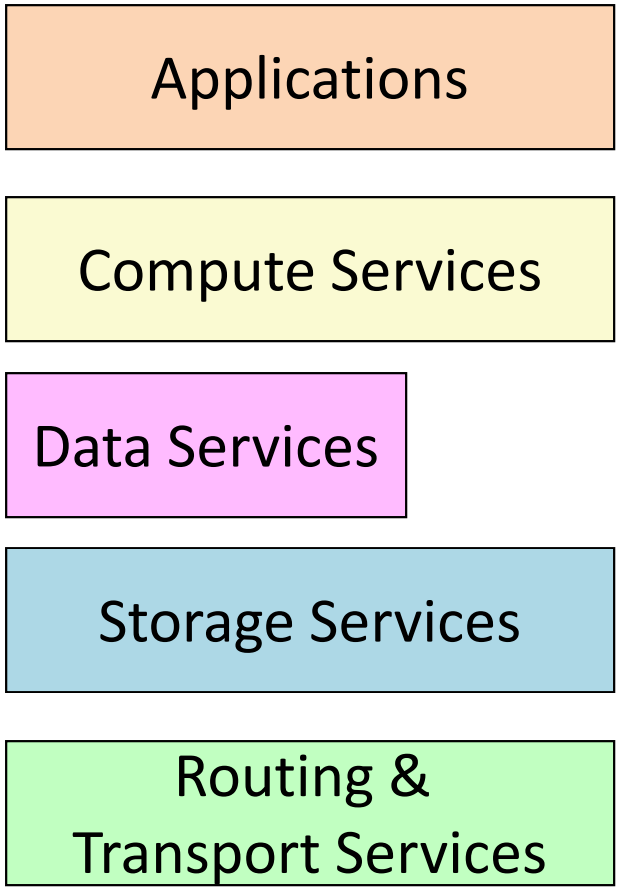


Hadoop's MapReduce

Hadoop Distributed File System (HDFS)

Hadoop's Stack

Sector's Large Data Cloud



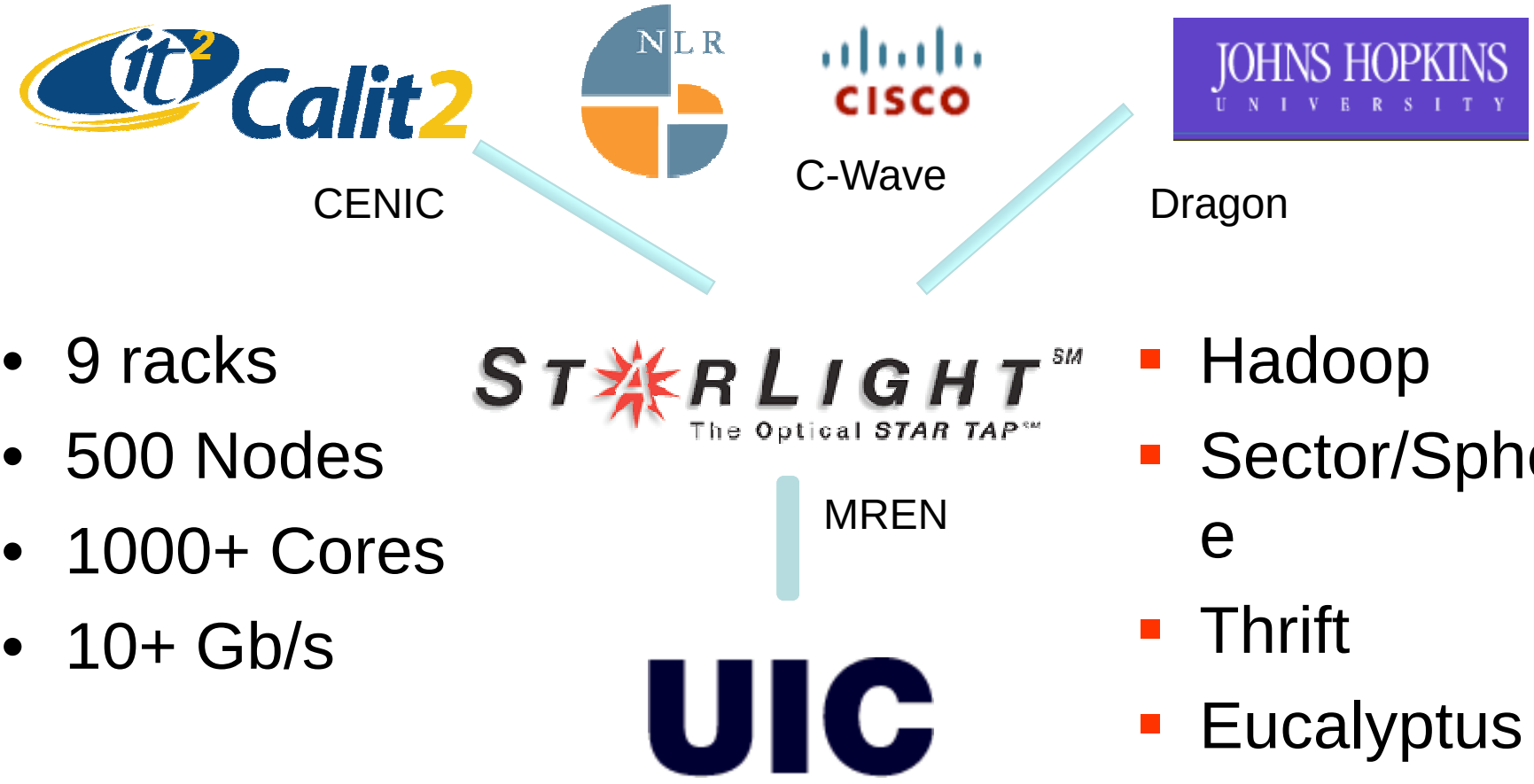
Sector's Stack

Sphere's UDFs

Sector's Distributed File System (SDFS)

UDP-based Data Transport Protocol (UDT)

Open Cloud Testbed (2009)



- 9 racks
- 500 Nodes
- 1000+ Cores
- 10+ Gb/s

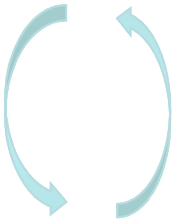
- Hadoop
- Sector/Sphere
- Thrift
- Eucalyptus

What I Learned

- Dead nodes are easy; tardy nodes are hard.
- Terasort is only a small part of the picture.
- There are many surprises in how best to program a data center.

Cloud Interoperability

OCC Large Data Cloud Interoperability Framework



- Platform as a Service
  - Cloud Compute Services
  - Data as a Service

Large Data Cloud Interoperability Framework

SNIA Cloud Data Management Interface (CDMI)

- Infrastructure as a Service
  - Virtual Data Centers (VDC)
  - Virtual Networks (VN)
  - Virtual Machines (VM)
  - Physical Resources

Open Virtualization Format (OVF)

Open Cloud Computing Interface (OCCI)

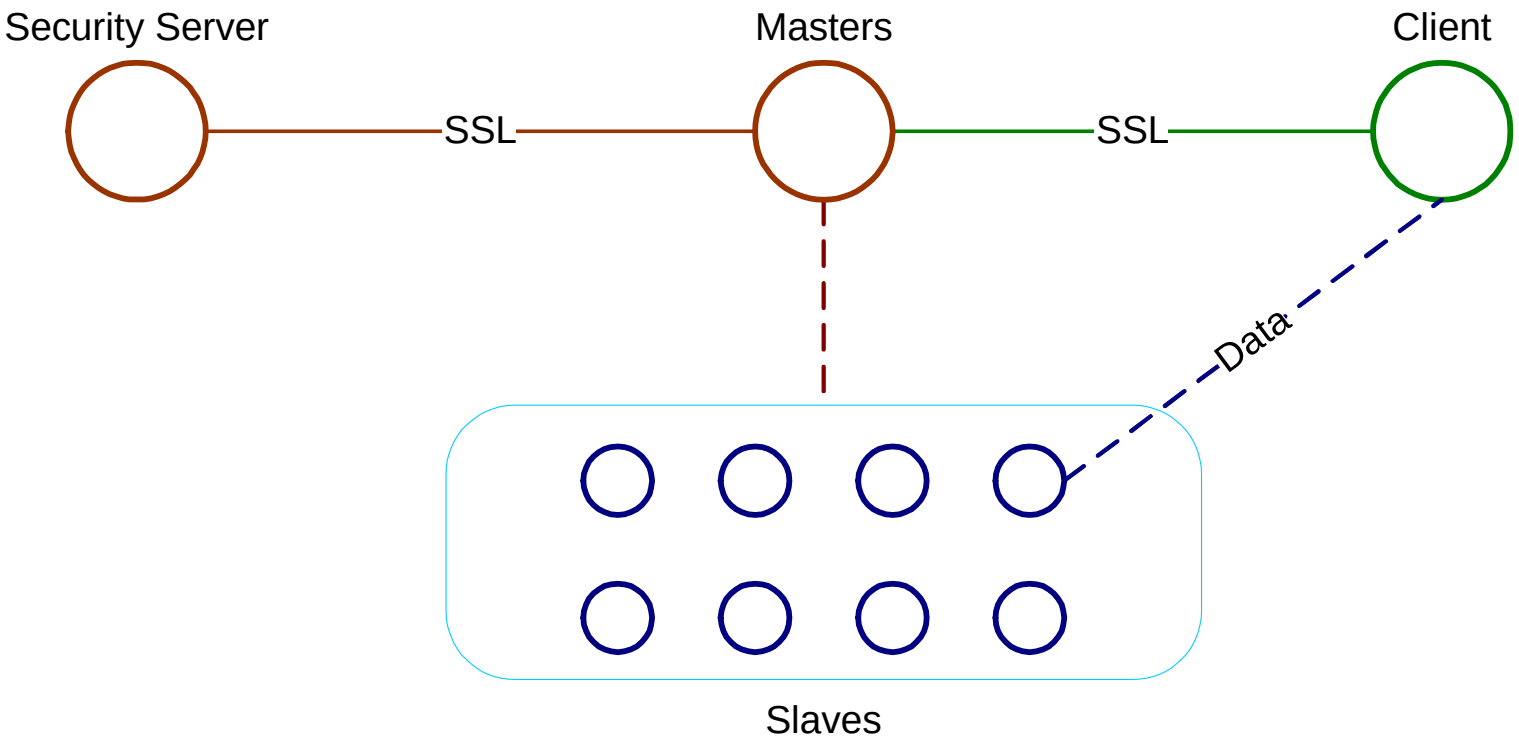
Raywulf Cluster

- 32 nodes: 31 compute/storage nodes and 1 head node
- Data switch (Ethernet): 96X1Gbs ports and 2X10Gb/s uplink
- Management switch
- Each compute/storage node:
  - Intel Xeon 5410 Quad Core CPU with 16GB of RAM
  - SATA RAID controller
  - Four (4) SATA 1TB hard drives in RAID-0 configuration
  - Two 1 Gbps NIC
  - IPMI management
  - $2200
- Total $85,000

Part 2. Sector Architecture

Data intensive computing that scales to a data center.

Sector Architecture



Sector DFS vs Hadoop DFS

- Block vs File
  - Sector stores files on the native file system of each node; Files are not split into blocks.
- Replication
  - Sector supports both on-write replication and periodical replication
- Data transfer
  - Sector uses UDT for data transfer
- Security, compression, etc.

Sphere UDF

```
for (int i = 0; i < total_segment_num; ++ i)
        UDF(segment[i]);

for (int i = 0; i < total_segment_num; ++ i)
        UDF(segment_A[i], segment_B[i]);
```

- The processing results from one node can be sent to multiple destination nodes.
- UDF is a superset of MapReduce
- Sphere provides a Map and Reduce UDFs

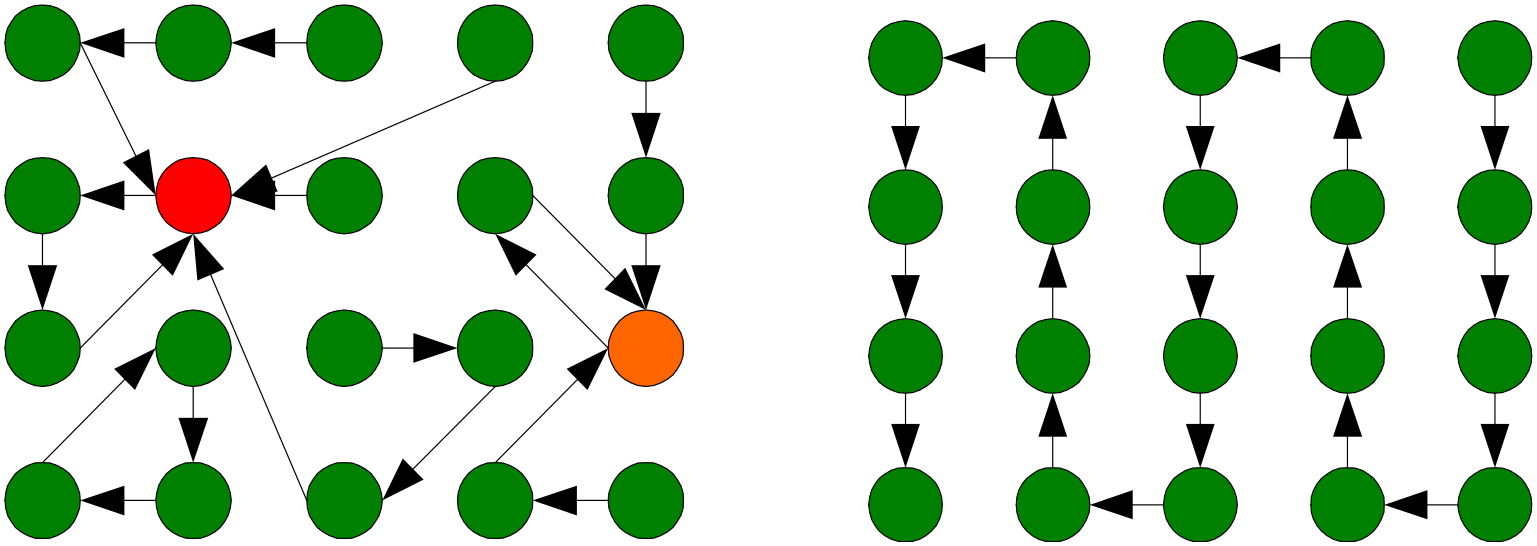
Sphere UDF vs. MapReduce

Sphere UDF	**MapReduce**
Record Offset Index	Parser / Input Reader
UDF	Map
Bucket	Partition
-	Compare
UDF	Reduce
-	Output Writer

Comparing Sector and Hadoop

	Hadoop	**Sector**
Storage Cloud	Block-based file system	File-based
Programming Model	MapReduce	UDF & MapReduce
Protocol	TCP	UDP-based protocol (UDT)
Replication	At write	At write or period.
Security	Not yet	HIPAA capable
Language	Java	C++

Balancing Data Flow



Terasort Benchmark

	1 Rack	2 Racks	3 Racks	4 Racks
Nodes	32	64	96	128
Cores	128	256	384	512
Hadoop	85m 49s	37m 0s	25m 14s	17m 45s
Sector	28m 25s	15m 20s	10m 19s	7m 56s
Speed up	3.0	2.4	2.4	2.2

Sector/Sphere 1.24a, Hadoop 0.20.1 with no replication on Phase 2 of Open Cloud Testbed with co-located racks.

MalStone (OCC Benchmark)

	MalStone A	MalStone B
Hadoop	455m 13s	840m 50s
Hadoop streaming with Python	87m 29s	142m 32s
Sector/Sphere	33m 40s	43m 44s
Speed up (Sector v Hadoop)	13.5x	19.2x

Sector/Sphere 1.20, Hadoop 0.18.3 with no replication on Phase 1 of Open Cloud Testbed in a single rack. Data consisted of 20 nodes with 500 million 100-byte records / node.

Lessons Learned

- Data Locality
- MapReduce is not the only parallel paradigm available
- Load balancing is critical to performance
- Fault tolerance comes with a price
- Balanced system
- Streaming

For More Information

- http://sector.sf.net

- Please come to booth 1309 (NCDM) and/or L6 (Disruptive Technologies).