eSCIENCE IN A CMU CLOUD CLUSTER

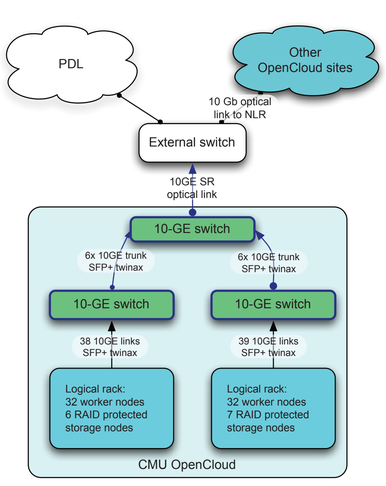
Julio López, Michael Stroucken, Mitch Franzos, Jason Sommerfield, Wittawat Tantisiriroj, Milo Polte, Lin Xiao,
Soila Pertet, Kai Ren, Greg Ganger, Garth Gibson (CMU)

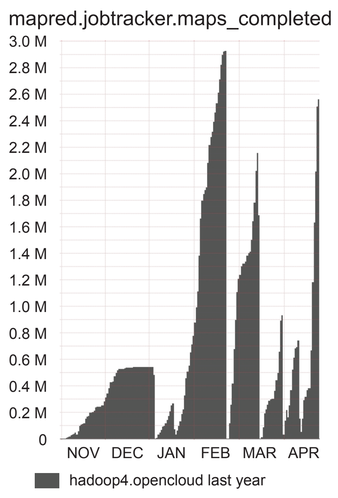
MOTIVATION

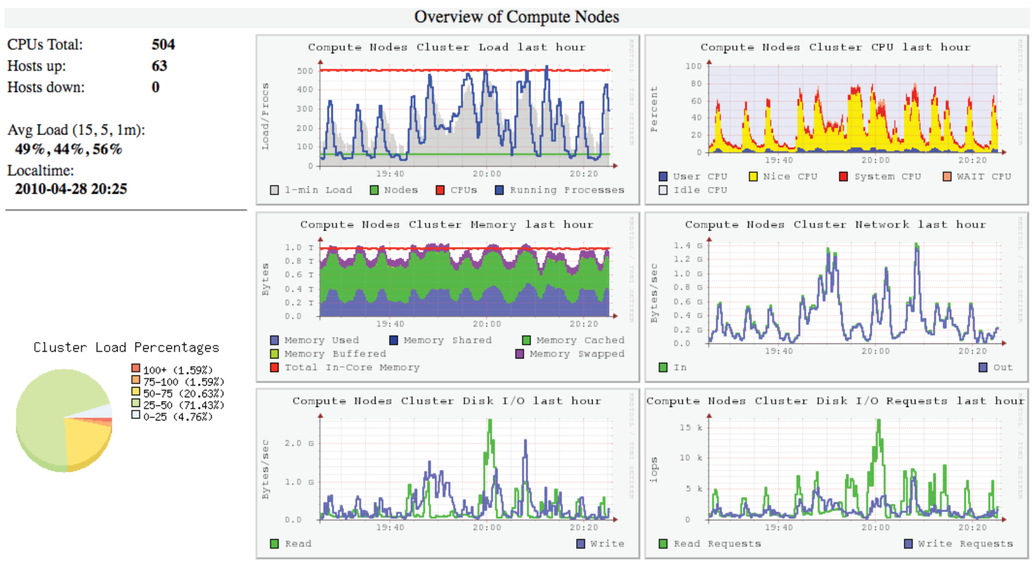
- Move scientists into a data-intensive scalable computing environment
- Focus on data-intensive applications and research
- Provide scientists with new tools for data analytics
- Capture application workload patterns
- Learn patterns from application structure
- Build better programming abstractions and tools for eScience and data analytics at scale

PLATFORM

- Data-intensive software stack: HDFS, Hadoop, HBase, Pig, Hive, Mahout
- 64 nodes, 512 CPU cores, 1/2 PB storage, 1 TB RAM, 10GE net
- Network: 10 GE to the host
  - 3x 10-GE switches, 60 Gbps bi-section bandwidth
  - 10 GE to the host
- Storage:
  - 250 TB Co-located with compute nodes
  - 282 TB RAID-protected external (13 PVFS servers)
- Node configuration:
  - CPUs: 2x quad-core Intel Xeon E5440 @2.83GHz
  - Storage: 4x 1TB SATA
  - RAM: 16 GB
  - Network: 10-GE NIC, SPF+ Twinax
- 6 map and 4 reduce slots per node
- 750 k map and 81 k reduce hours since 4/9/10







DATASETS SO FAR

- Large-scale cosmology simulations (75 TB)
- Eartquake simulation wavefields (10 TB)
- Black hole evolution (0.5 TB)
- Astronomy surveys digital images (26 TB)
- Tweeter: Social networking short communications (5 TB)
- ClueWeb09: Targeted web crawl (2 TB)
- Blogosphere snapshots (3 TB)
- Cluster job logs (1 TB)
- Fine-resolution cluster resource usage metrics (1.5 TB)
- Protein images from immunohistochemical staining (30 GB)

PROJECTS SO FAR

Astro-DISC: (Croft, Di-Matteo, López, Fink, Gibson):
- Analysis of massive cosmology datasets. Galaxy clustering, black hole merger tree construction, correlation functions, quasar classification

Computational Biology (Murphy, Langmead)
- Active learning of drug, protein association data, cellular simulations

Seismology (Bielak, Taborda, López):
- Large-scale seismic wavefield analysis

Twitter analysis (Von Ahn, Meeder)
- Algorithms on social network graphs

Distributed Malware Clustering (Brumley)

Large Scale Graph Mining (Faloutsos)
- Scalable pattern finding algorithms for finding large graphs: diameter, radius, PageRank, connected components, eigenvalues

Worldly Knowledge & Read the Web: (Mitchell, Cohen, Callan, Smith, Gibson)
- Large-scale information extraction from web documents

Applied statistical natural language processing (Smith, Callan)
- Forecasting economic and political phenomena by mining text streams

Informedia digital video Analysis (Hauptmann)
- Feature extraction machine understanding of video media.

Cluster performance diagnosis (Lopez, Narasimhan, Ganger, Gibson)

Mining the blogosphere (Guestrin)
- Parallelization of machine learning algorithms for large-scale classification using HBase

Intel Science & Technology Center for Cloud Computing

GORDON AND BETTY MOORE FOUNDATION

Carnegie Mellon University · Georgia Tech · intel · PRINCETON UNIVERSITY · UC Berkeley