DATA-INTENSIVE COMPUTER CLUSTERS AND NETWORKS

Mitch Franzos, Michael Stroucken, Julio López, Garth Gibson (CMU)

OVERVIEW

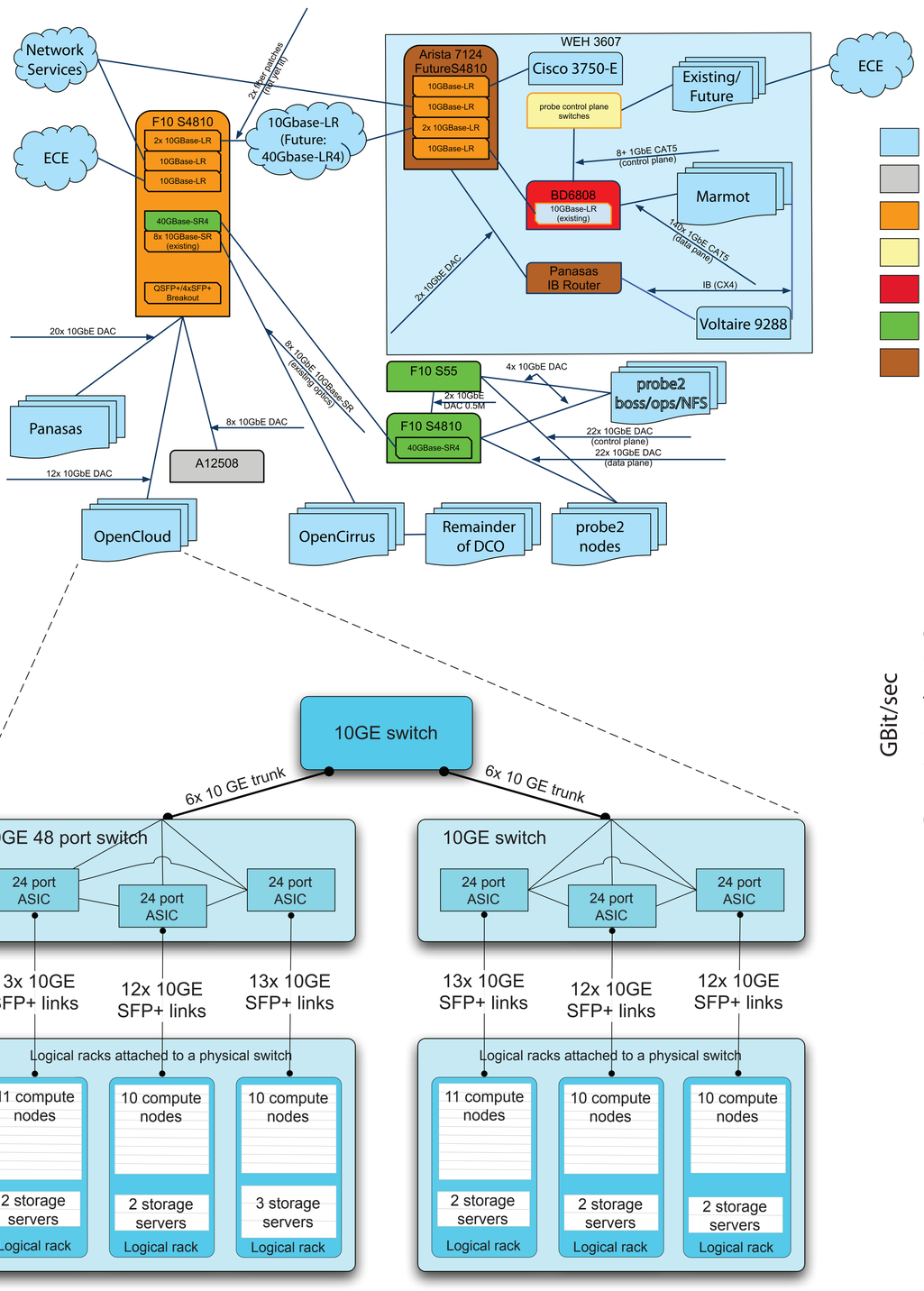
- New computer cluster for data-intensive research (DISC)
- Intra- and inter- data center experiments
- Emphasis on Table Software for large scale analytics
- Scalable Analytics for Astronomy (Astro-DISC)
- Part of the OpenCloud testbed
  http://www.opencloudconsortium.org/testbed.html
- Hosted in the PDL Data Center Observatory Zone 2
- Provided by NSF and Intel

SOFTWARE

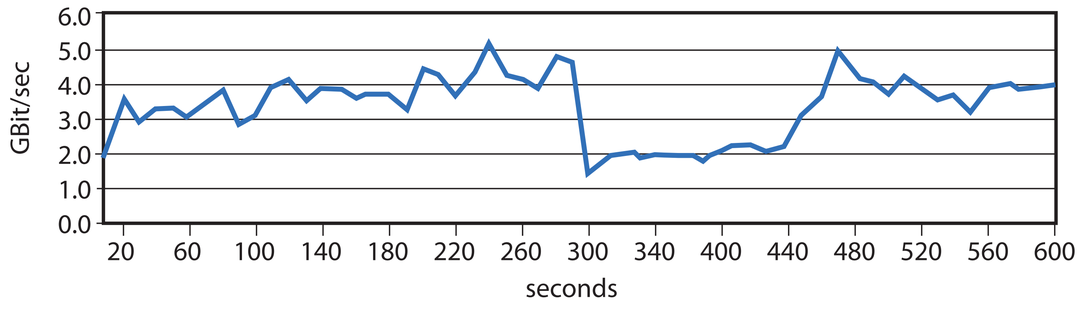
- Flexible configuration management
- OS: Linux distribution
- Data processing: Hadoop, HDFS
- Analytics support: Numerical and algorithm libraries
- Table Software: HBase, Hypertable, Cassandra

HARDWARE SPECIFICATIONS

- 64 nodes, 512 CPU cores, 1/2 PB storage, 1 TB RAM, 10GE net
- Network:
  - 3x 10-GE switches, 60 Gbps bi-section bandwidth
  - 10 GE to the host
  - 10-Gbps connection to the National Lambda Rail (NLR)
- 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



existing
HP-Samsung-VMWare vCloud project being built
has been ordered to expand OpenCloud/OpenCirrus
existing PRObE netgear switches
existing PRObE blackdiamond switches
proposed future GE network purchases for PRObE 2 (high core count) cluster
under consideration for increasing bandwidth between clusters



Bandwidth to the New Mexico Consortium (NMC) without jumbo frames.
Command: 'iperf -c perf.nmc-probe.org -P64 -i 10 -t 600 -f gG | grep SUM'.
The wide area network bandwidth to NMC has been improving of late: up to 7 Gbps sustained with jumbo frames.

OPENCLOUD CLUSTER ARCHITECTURE & NETWORKING

Intel Science & Technology Center for Cloud Computing

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