Data Mining Using High Performance Data Clouds: Experimental Studies using Sector & Sphere

Robert Grossman, University of Illinois at Chicago & Open Data Group
Yunhong Gu, University of Illinois at Chicago

What is a Cloud?

•Clouds provide on-demand resources or services over the Internet, usually at the scale and reliability of a data center
•No standard definition
•Cloud architectures are not new
• What is new is the scale – capacity is being added by the data center not by the rack
• Examples include:
  ○Google's Google File System, MapReduce & BigTable
  ○Amazon's S3, EC2 and SimpleDB
  ○ Open source Hadoop

Performance

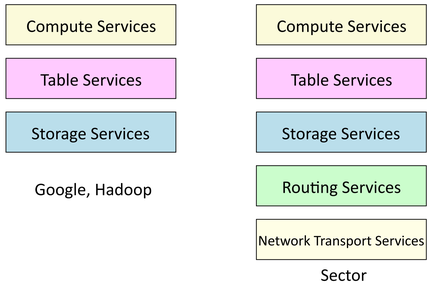
	Number of Nodes	Hadoop v0.1.17 (sec)	Sector v1.8 (sec)
WAN-2 (UIC, SL, UCSD, JHU)	118	3702	1526
WAN-1 (UIC, SL, UCSD)	88	3069	1430
MAN (UIC, SL)	58	2617	1301
LAN (UIC)	29	2252	1265

The table shows the time required in sec to complete the Terasort benchmark. The time required to generate the data is excluded. The test used 10 GB of data per node. The four clusters were connected with 10 GE networks. The tests were done on the Open Cloud Testbed using Dell 1435 computers with 4GB mem, 1TB disk, 2.0GHz dual-core AMD Opteron 2212, with 1 Gb/s network interface cards

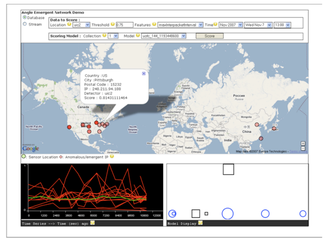
Sector

•Sector is fast
  ○Terasort over 118 nodes requires 1526 seconds with Sector and 3702 sec with Hadoop
•Easy to program
  ○Supports User-defined Functions over records
  ○Supports MapReduce style over (key, value) pairs
•Customizable
  ○Sector is layered and designed to support specialized functionality by customizing the layers
  ○Example: using specialized network protocols for high performance networks
•Open source and available from sector.sourceforge.net

Stacks for Clouds



Sector Applications



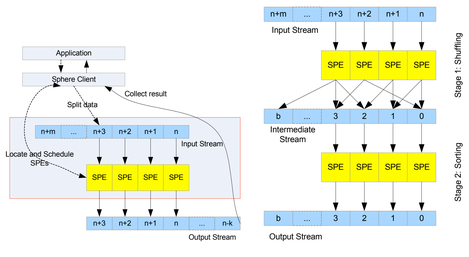
• Angle anomaly detection project uses Sector as its distributing computing platform.

Source	Destination	LLPR*	Link	Bandwidth
Chicago	Greenbelt	0.98	1 Gb/s	615 Mb/s
Chicago	Austin	0.83	10 Gb/s	8000 Mb/s

*LLPR = local / long distance performance
• Sector LLPR for SDSS varies between 0.61 and 0.98

•Distributing the Sloan Digital Sky Survey (SDSS)
• SDSS dataset is 14 TB in size.

Sector / Sphere Programming Model

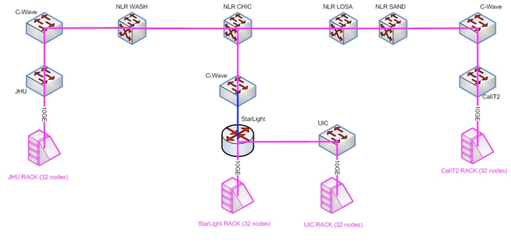


•Sector dataset consists of one or more physical files
•Sphere applies User Defined Functions over streams of data consisting of data segments
  ○ Example of UDFs: Map function, Reduce function, Split function for CART
•Data segments can be data records, collections of data records, or files
•Outputs of UDFs can be returned to originating node, written to local node, or shuffled to another node.

Open Cloud Consortium

•Supports the development of open source software for cloud based computing.
•Develops standards and standard based interfaces for interoperating different software supporting cloud based computing.
•Manages a testbed for cloud computing called the Open Cloud Testbed.
•Sponsors workshops and other events related to cloud computing.
•See www.opencloudconsortium.org

Open Cloud Testbed



•Distributed cloud testbed with clusters at UIC, StarLight, Calit2 and JHU
•All clusters connected via 10+ GE wide area network