PERVASIVE TECHNOLOGY INSTITUTE

INDIANA UNIVERSITY

Big Data Workshop Summary

Virtual School for Computational Science and Engineering
July 30 2010

Geoffrey Fox

gcf@indiana.edu

http://www.infomall.org http://www.futuregrid.org

Director, Digital Science Center, Pervasive Technology Institute

Associate Dean for Research and Graduate Studies, School of Informatics and Computing

Indiana University Bloomington

Many Thanks!

- Speakers
- Collaboration (A-V) Specialists
- FutureGrid support staff
- Lunch and other support staff!
- Tom Scavo and Virtual School leadership (Thom Dunning and Sharon Glotzer)

- And Students!

Confusion in a Nutshell!

- Time of Change so
  - It's Exciting
  - But its hard to tell a coherent story
- In particular TeraGrid and Cloud approaches are not aligned
- Partly as TeraGrid mainly simulation and partly as technologies are changing

Applications

- Astronomy, Computational Fluid Dynamics, Sensors – Szalay, Budavari
- Life Sciences: (Biology/Bioinformatics/Cheminformatics) Qiu Poulain Wild Sun Grossman Figueiredo
- Particle Physics: Qiu Poulain

Data: Files Transport and I/O

- Latham: I/O Architecture
- Application – System Mapping
- MPI IO
- netCDF files for structured data (arrays)
- HDF5 for datasets
- Other application specific packages
- Darshan performance monitoring
- Examples: Flash and Volume Rendering
- Miller: Data Capacitor at IU and TeraGrid; application examples
  - Will be mounted on FutureGrid
- Lustre in action on Data Capacitor
  - Use for big files
  - Metadata
- Tatineni: Lustre File System in detail
  - Architecture
  - Parallel I/O performance with many cores; choice of stripe size
- Wide area File systems: GPFS DC Lustre(future)
- Data Transfer: ssh bbcp bbftp GridFTP
- Data Management and archivin

Cloud Technologies

- Focus as many data processing problems are suitable for either clouds (IaaS) or Cloud technologies (MapReduce, Data Parallel File Systems, Tables)
- Hadoop, Twister, Dryad, Sector(Sphere)
- NOSQL (Grossman) versus SQL (Szalay)
- Amazon EC2, Azure
- FutureGrid supports Clouds, Grids, Parallel computing
- Open Science Data Cloud supports data intensive clouds
- Appliances make deployment easier and especially attractive for classes; Virtual Clusters and Networks; on VM's or bare; GroupVPN supports VO sharing resources
  - You can deploy anywhere (maybe except TeraGrid!)

Some issues

- Szalay defined the Big Science Data problem
  - Disk I/O performance issues – Amdahl numbers, Graywulf
  - 100 TB practical limit today in science but Google/Microsoft are order(s) of magnitude larger
  - Data-compute collocation -- Cyberbricks
- Budavari noted role of GPUs, SQL, communities and standards
- Grossman posed SQL v NOSQL ; ACID v BASE; and Size issues (SQL doesn’t scale across datacenter)