Big Data and Clouds: Challenges and Opportunities

NIST
January 15 2013

Geoffrey Fox

gcf@indiana.edu


School of Informatics and Computing
Digital Science Center
Indiana University Bloomington

https://portal.futuregrid.org
Charge to Presenters

• Discuss opportunities and challenges presented by the intersection of cloud and big data.

• For example, on the opportunity side we have been thinking about the ability of cloud to make big data approaches feasible and cost-effective for small and medium enterprises and for the combination to enable new, data-as-a-service business models.

• On the challenge side, we have been thinking about how “bring-the-computation-to-the-data-rather-than-the-data-to-the-computation” approaches could work in cloud environments and what quality metrics and measurement methods could work across heterogeneous data types of uncertain provenance, including methods for quality discovery.

• These are just examples and we are very interested in hearing your take on the intersection of cloud and big data.
Some Topics

• Curricula
• Consensus on Architecture and value of clouds
• High Performance Library
• FutureGrid
Education and Training

- Microsoft says there will be **14 million cloud jobs** around the world by 2015.

- McKinsey says that there will up to **190,000 nerds** and **1.5 million extra managers** needed in Data Science by 2018 in USA.

- Many more jobs than simulation (third paradigm) where **computational science** not very successful as curriculum.

- Need curricula to educate people to use/design **Clouds** running **Data Analytics** processing **Big Data** to solve problems in **X-Informatics** (X= Bio...LifeStyle...Policy...Wealth)

- Cover Data curation/management, Analytics (algorithms), run-time (MapReduce, Workflow, NOSQL), Applications

- Not many courses aimed at any one aspect of this; let alone everything and their integration

- Look at Massive Open Online Courses (MOOCs)

[https://portal.futuregrid.org](https://portal.futuregrid.org)
Clouds for Scientific Data Analysis

• There has been plenty of trials and several successes from particle physics (LHC) data analysis to genome sequencing

• **MapReduce/NOSQL** with Iterative extensions good for data intensive problems which have very different communication requirements from large scale simulations
  – Large collective communication v. smallish local messages

• However no agreement on good data architecture or even requirements for this either in cloud or on conventional HPC style systems

• No agreement on value of commercial clouds as cost effective solution

• **Need to generate a consensus on data architectures as exists for simulations**
  – Exascale discussion builds on agreed principles
Data Analytics Futures?

• Better algorithms contribute as much as better hardware in HPC

• PETSc and ScaLAPACK and similar libraries very important in supporting parallel simulations

• Need equivalent Data Analytics libraries

• Include datamining (Clustering, SVM, HMM, Bayesian Nets ...), image processing, information retrieval including hidden factor analysis (LDA), global inference, dimension reduction
  – Many libraries/toolkits (R, Matlab) and web sites (BLAST) but typically not aimed at scalable high performance algorithms

• Should support clouds and HPC; MPI and MapReduce
  – Iterative MapReduce an interesting runtime; Hadoop has many limitations

• Need a coordinated Academic Business Government Collaboration to build robust algorithms that scale well

• Propose to build community to define & implement SPIDAL or Scalable Parallel Interoperable Data Analytics Library

https://portal.futuregrid.org
FutureGrid offers Computing Testbed as a Service

**Software (Application Or Usage)**
- CS Research Use e.g. test new compiler or storage model
- Class Usages e.g. run GPU & multicore
- Applications

**Platform**
- Cloud e.g. MapReduce
- HPC e.g. PETSc, SAGA
- Computer Science e.g. Compiler tools, Sensor nets, Monitors

**Infrastructure**
- Software Defined Computing (virtual Clusters)
- Hypervisor, Bare Metal
- Operating System

**IaaS**
- Provisioning
- Image Management
- IaaS Interoperability
- NaaS, IaaS tools
- Expt management
- Dynamic IaaS NaaS
- Devops

**FutureGrid Uses Testbed-aaS Tools**
- Provisioning
- Image Management
- IaaS Interoperability
- NaaS, IaaS tools
- Expt management
- Dynamic IaaS NaaS
- Devops

**FutureGrid Usages**
- Computer Science
- Applications and understanding Science Clouds
- Technology Evaluation including XSEDE testing
- Education & Training
FutureGrid key Concepts

- FutureGrid is an international testbed modeled on Grid5000
- Supporting international Computer Science and Computational Science research in cloud, grid and parallel computing (HPC)
- The FutureGrid testbed provides to its users:
  - A flexible development and testing platform for middleware and application users looking at interoperability, functionality, performance or evaluation
  - FutureGrid is user-customizable, accessed interactively and supports Grid, Cloud and HPC software with and without VM’s
  - A rich education and teaching platform for classes
- Offers OpenStack, Eucalyptus, Nimbus, OpenNebula, HPC (MPI) on same hardware moving to software defined systems; classic HPC and Cloud storage

https://portal.futuregrid.org
4 Use Types for FutureGrid TestbedaaS

• 285 approved projects (1500 users) January 13 2013
  – USA(80%), China, India, Pakistan, lots of European countries
  – Industry, Government, Academia

• Training Education and Outreach (14.7%)
  – Semester and short events; interesting outreach to HBCU

• Computer science and Middleware (56%)
  – Core CS and Cyberinfrastructure; Interoperability (3.3%) for Grids and Clouds; Open Grid Forum OGF Standards

• Computer Systems Evaluation (8.8%)
  – XSEDE (TIS, TAS), OSG, EGI; Campuses

• New Domain Science applications (20.5%)
  – Life science highlighted (10.6%), Non Life Science (9.9%)

• Could emphasize Data Science and more experimentation by Government and Industry

https://portal.futuregrid.org