Clouds from FutureGrid's Perspective

April 4 2012

Geoffrey Fox

gcf@indiana.edu

Director, Digital Science Center, Pervasive Technology Institute

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

Indiana University Bloomington

Programming Paradigms for Technical Computing on Clouds and Supercomputers (Fox and Gannon)

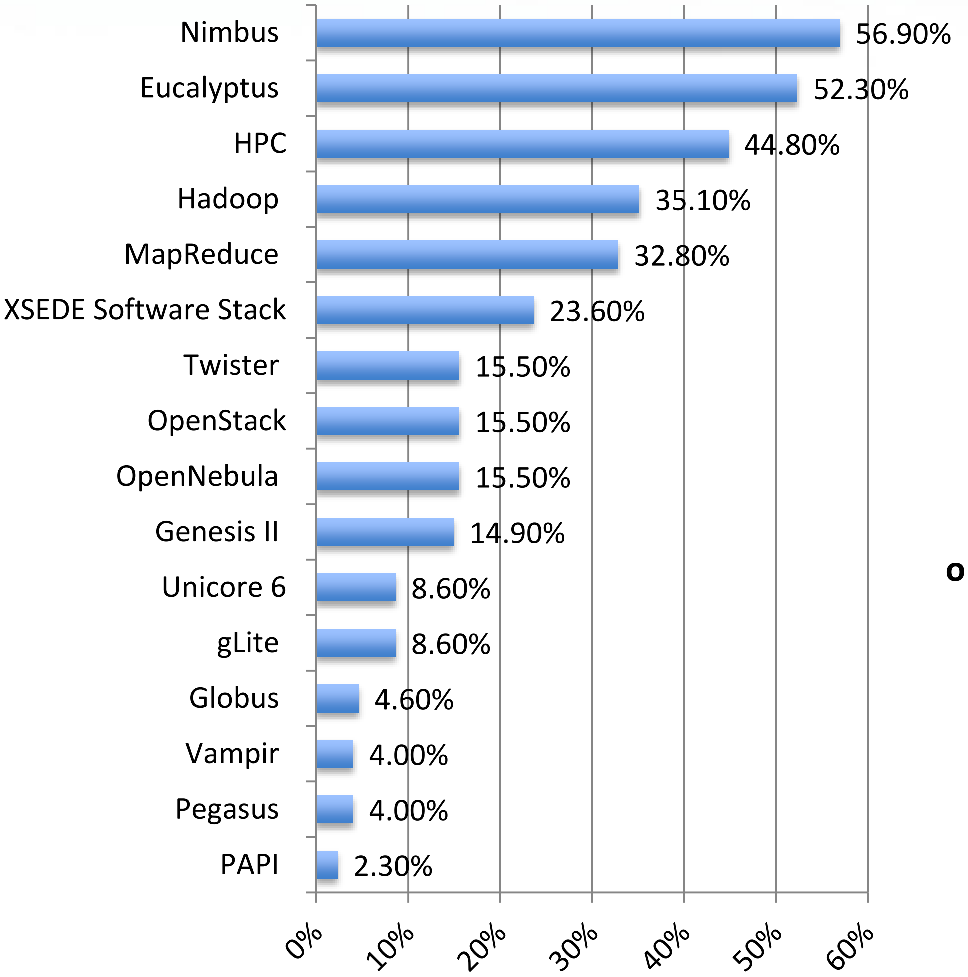
http://grids.ucs.indiana.edu/ptliupages/publications/Cloud%20Programming%20Paradigms_for__Futures.pdf

https://portal.futuregrid.org

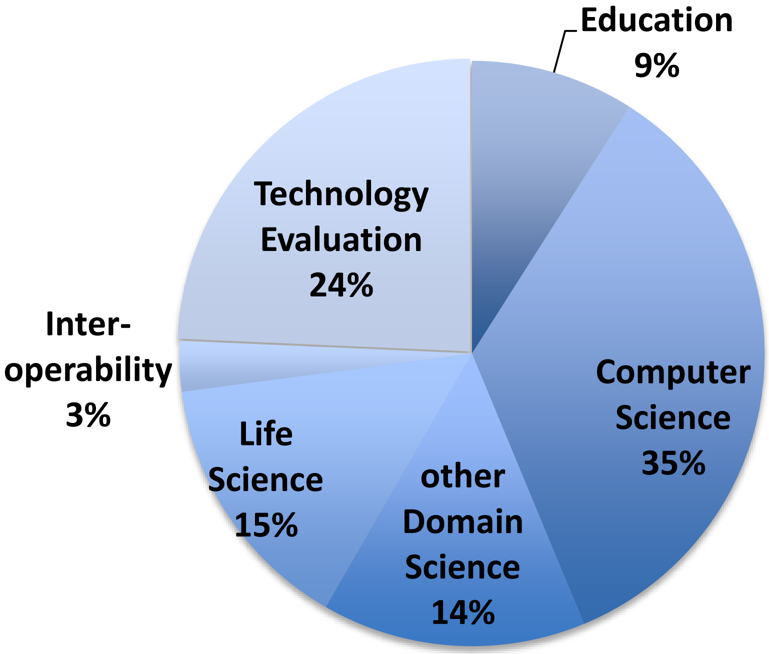
What is FutureGrid?

- The FutureGrid project mission is to enable experimental work that advances:
  a) Innovation and scientific understanding of distributed computing and parallel computing paradigms,
  b) The engineering science of middleware that enables these paradigms,
  c) The use and drivers of these paradigms by important applications, and,
  d) The education of a new generation of students and workforce on the use of these paradigms and their applications.
- The implementation of mission includes
  - Distributed flexible hardware with supported use
  - Identified IaaS and PaaS "core" software with supported use
  - Outreach
- ~4500 cores in 5 major sites

Distribution of FutureGrid Technologies and Areas



- 190 Projects

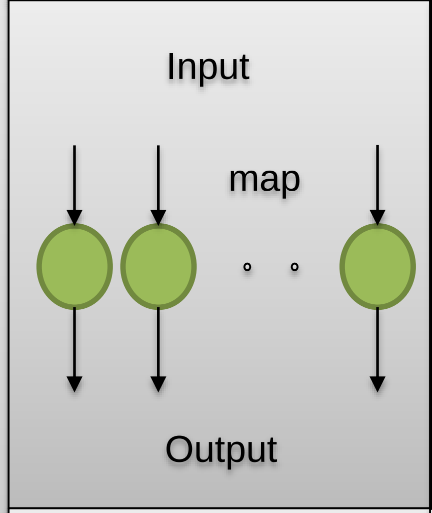
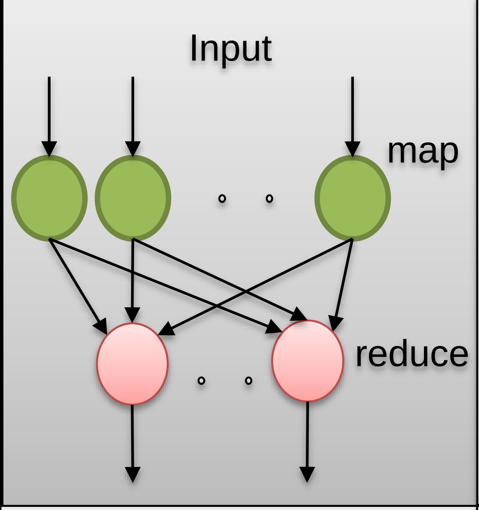
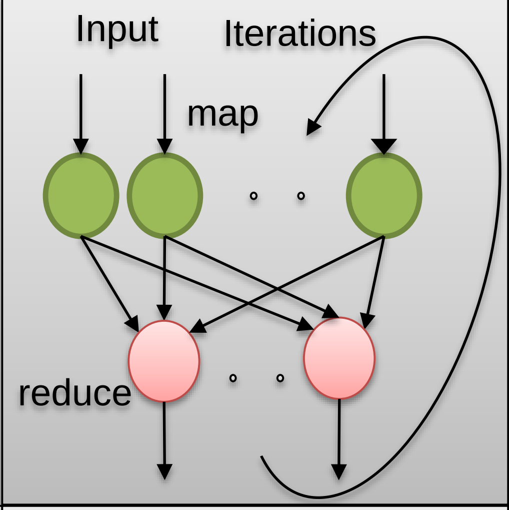
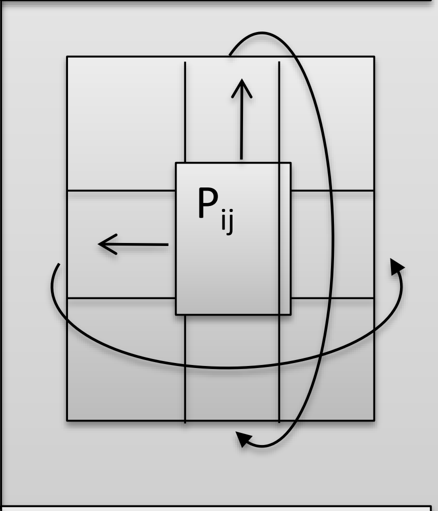


https://portal.futuregrid.org

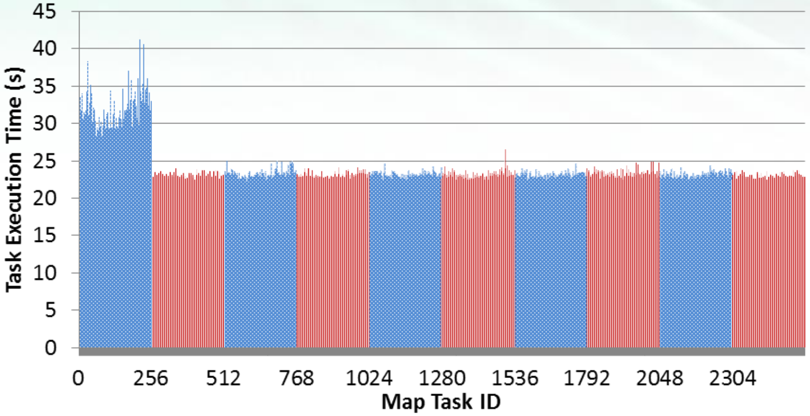
Using Clouds in a Nutshell

- High Throughput Computing; pleasingly parallel; grid applications
- Multiple users (long tail of science) and usages (parameter searches)
- Internet of Things (Sensor nets) as in cloud support of smart phones
- (Iterative) MapReduce including “most” data analysis
- Exploiting elasticity and platforms (HDFS, Queues ..)
- Use services, portals (gateways) and workflow
- Good Strategies:
  - Build the application as a service;
  - Build on existing cloud deployments such as Hadoop;
  - Use PaaS if possible;
  - Design for failure;
  - Use as a Service (e.g. SQLaaS) where possible;
  - Address Challenge of Moving Data

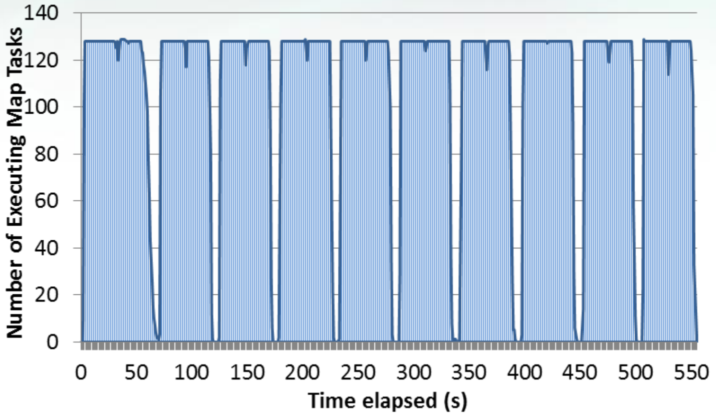
4 Forms of MapReduce

(a) Map Only	(b) Classic MapReduce	(c) Iterative MapReduce	(d) Loosely Synchronous
			
BLAST Analysis Parametric sweep Pleasingly Parallel	High Energy Physics (HEP) Histograms Distributed search	Expectation maximization Clustering e.g. Kmeans Linear Algebra, Page Rank	Classic MPI PDE Solvers and particle dynamics
← Domain of MapReduce and Iterative Extensions →			MPI

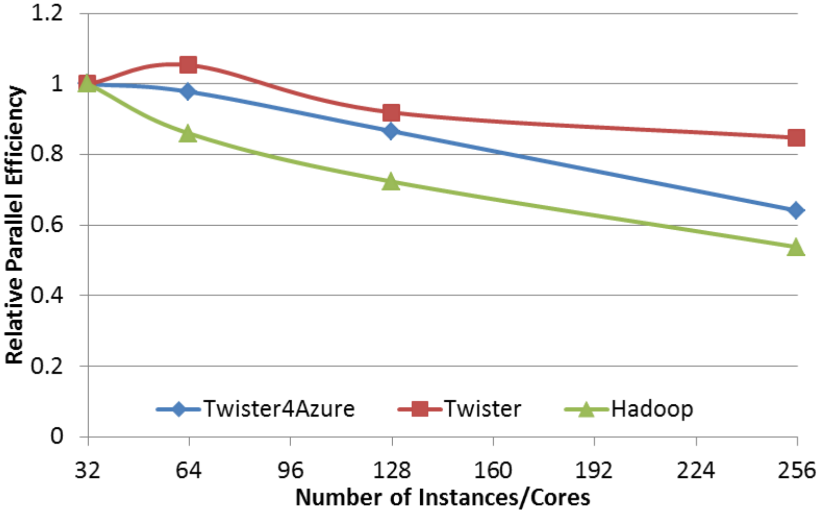
Iterative MapReduce Performance
Kmeans Clustering



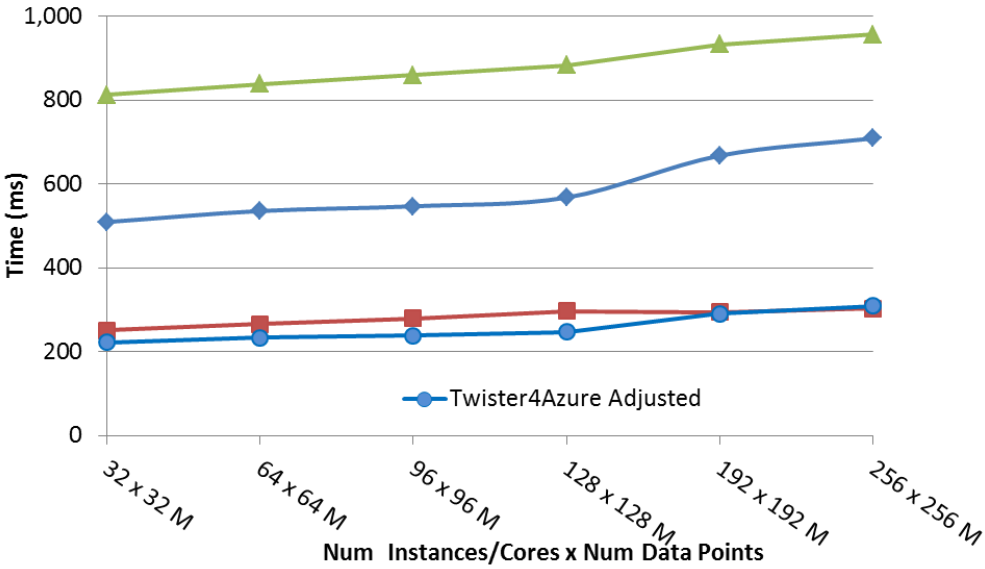
Task Execution Time Histogram



Number of Executing Map Task Histogram



Strong Scaling with 128M Data Points



Weak Scaling

https://portal.futuregrid.org

Some next Steps

- Clouds are suitable for several types of (but not all) applications
- Clouds can leverage major commercial software investment
- Current academic (open source) cloud software needs more investment both in core capabilities and in "Platform"
  - Hadoop not best MapReduce for science
  - HDFS and OpenStack storage don't have quality of Lustre and classic HPC storage
- 14 million cloud jobs worldwide by 2015 – Cloud curricula and experiences can help workforce development
- Science Cloud Summer School July 30-August 3
  - ~10 Faculty and Students from MSI's (sent by ADMI)
  - Part of virtual summer school in computational science and engineering and expect over 200 participants spread over 10 sites
- Science Cloud and MapReduce XSEDE Community groups