Data Science, Clouds and X-Informatics

May 8 2013

3rd International Conference on Cloud Computing and Services Science, CLOSER 2013

Eurogress Aachen

Geoffrey Fox

gcf@indiana.edu

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

School of Informatics and Computing

Digital Science Center

Indiana University Bloomington

https://portal.futuregrid.org

Abstract

- We explore the principle that much of "the future" will be characterized by "Using Clouds running Data Analytics processing Big Data to solve problems in X-Informatics". Applications (values of X) include explicitly already Astronomy, Biology, Biomedicine, Business, Chemistry, Crisis, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields defined implicitly. We discuss the implications of this concept for education and research. Education requires new curricula – generically called data science – which will be hugely popular due to the many millions of jobs opening up in both "core technology" and within applications where of course there are most opportunities. We discuss possibility of using MOOC's to jumpstart field. On research side, big data (i.e. large applications) require big (i.e. scalable) algorithms on big infrastructure running robust convenient programming environments. We discuss clustering and information visualization using dimension reduction as examples of scalable algorithms. We compare Message Passing Interface MPI and extensions of MapReduce as the core technology to execute data analytics.
- We mention FutureGrid and a software defined Computing Testbed as a Service

Big Data Ecosystem in One Sentence

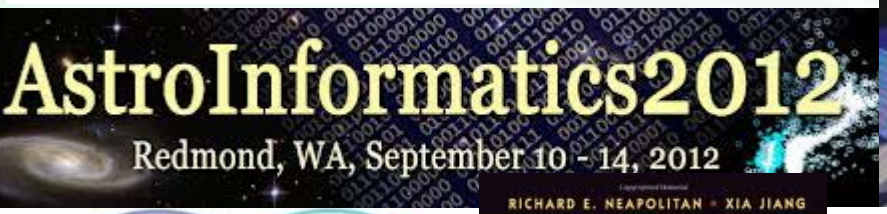
Use Clouds running Data Analytics processing Big Data to solve problems in X-Informatics (or e-X)

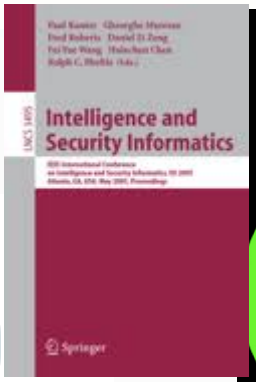
X = Astronomy, Biology, Biomedicine, Business, Chemistry, Crisis, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

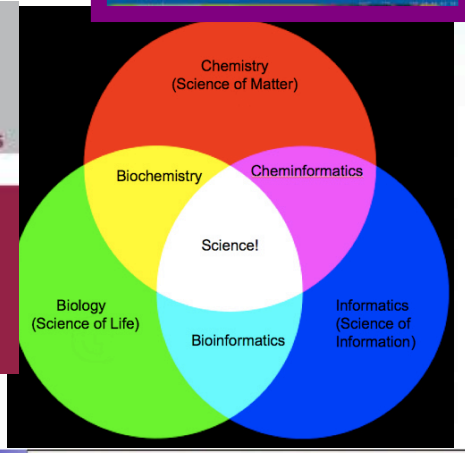
Spans Industry (AHEAD?) and Science (research)

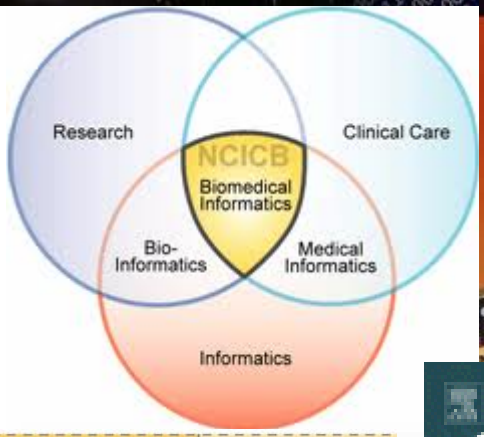
Education: Data Science see recent New York Times articles

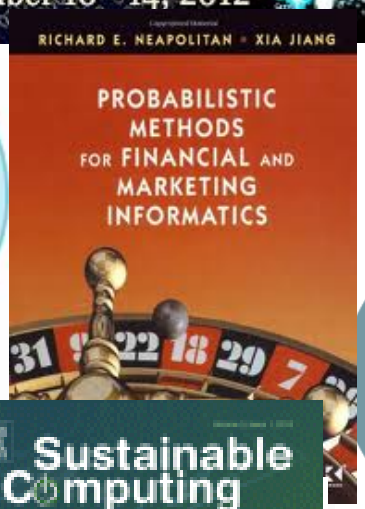
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

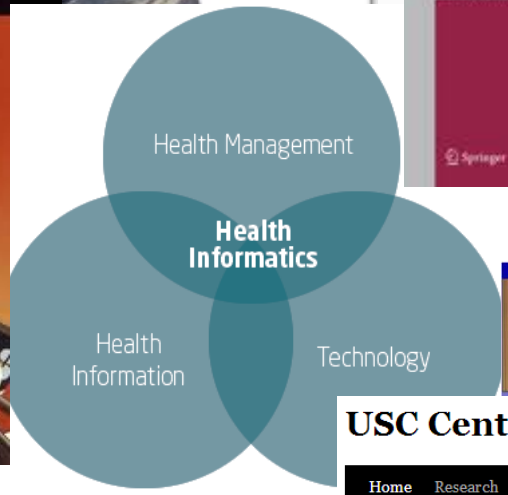


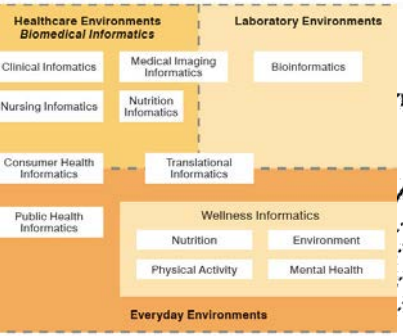


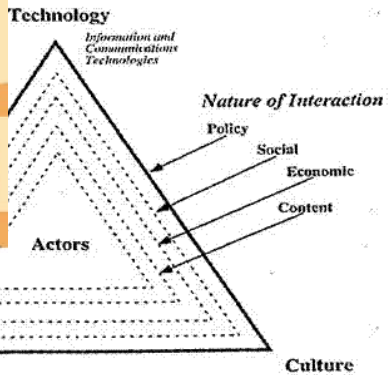




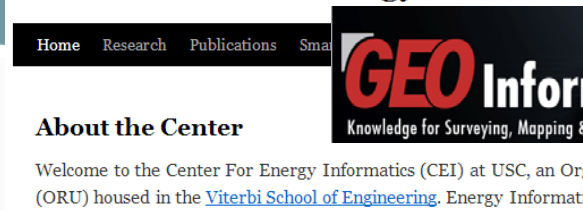


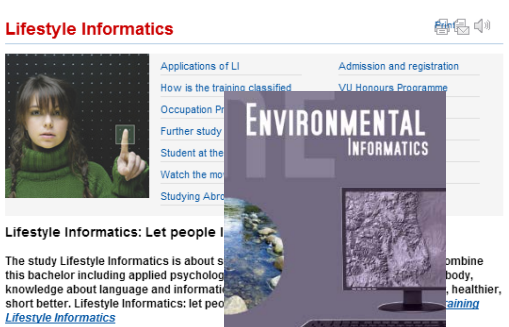






Social Informatics





Issues of Importance

- **Economic Imperative:** There are a lot of data and a lot of jobs
- **Computing Model:** Industry adopted clouds which are attractive for data analytics. Research has not adopted?
- **Research Model:** 4th Paradigm; From Theory to Data driven science?
- Progress in **Data Science Education**: opportunities at universities
- **Confusion in a new-old field:** lack of consensus academically in several aspects of data intensive computing from storage to algorithms, to processing and education
- Progress in **Data Intensive Programming Models**
- Progress in **Academic (open source) clouds**
- Progress in scalable robust **Algorithms**: new data need better algorithms?
- **FutureGrid:** Develop Experimental Systems

Economic Imperative
First Data

There are a lot of data and a lot of jobs

Some Trends

- **The Data Deluge** is clear trend from Commercial (Amazon, e-commerce) , Community (Facebook, Search) and Scientific applications
- **Light weight clients** from smartphones, tablets to sensors
- **Multicore** reawakening parallel computing
  - Compelling server side
- **Exascale initiatives** will continue drive to high end with a simulation orientation
- **Clouds with cheaper, greener, easier to use** IT for (some) applications
- **New jobs** associated with new curricula
  - **Clouds as (part of) a distributed system** (classic CS courses)
  - **Data Analytics** (Important theme in academia and industry)

Some Data sizes

- ~40 10^9 **Web pages** at ~300 kilobytes each = 10 Petabytes
- **Youtube** 48 hours video uploaded per minute;
  - in 2 months in 2010, uploaded more than total NBC ABC CBS
  - ~2.5 petabytes per year uploaded?
- **LHC** 15 petabytes per year
- **Radiology** 69 petabytes per year
- **Earth Observation** becoming ~4 petabytes per year
- **Earthquake Science** – few terabytes **total** today
- **PolarGrid** – 100's terabytes/year
- **Exascale simulation** data dumps – terabytes/second = 30 exabytes/year
- **Square Kilometer Array Telescope** will be 100 terabits/second = 400 exabytes/year


1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
1,700+ Firefox DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
20,000+ NEW POSTS ON tumblr.
13,000+ iPhone APPLICATIONS DOWNLOADED
New Craigslist Ads
THE LARGEST SOCIAL READING PUBLISHING COMPANY
IN 60 SECONDS
Google
Google Search
GO-Globe.com
Adobe Acrobat X Pro

Value of Data & Analytics

Monitor fleet of ~25,000* engines ... 3.6MM flight records/month

B777

Prognostics

- ✓ Dispatch reliability
- ✓ Preventive maintenance
- ✓ Asset utilization

Prevent failures = customer efficiency

+

GE90

Asset Productivity

- ✓ Enhanced service offerings
- ✓ Airline cost structure
- ✓ Fuel performance

Streamline operations = increased airline productivity

=

DATA

90,000 flight records analyzed

~200 parameters per flight record

~18MM parameters per month

System & Optimization

- ✓ Time & space management
- ✓ Fuel efficiency
- ✓ Airspace capacity

Integrated systems = value-added services

Drives strong alignment with customers

Creates productivity in long-term service agreements

Value-added services fuels growth

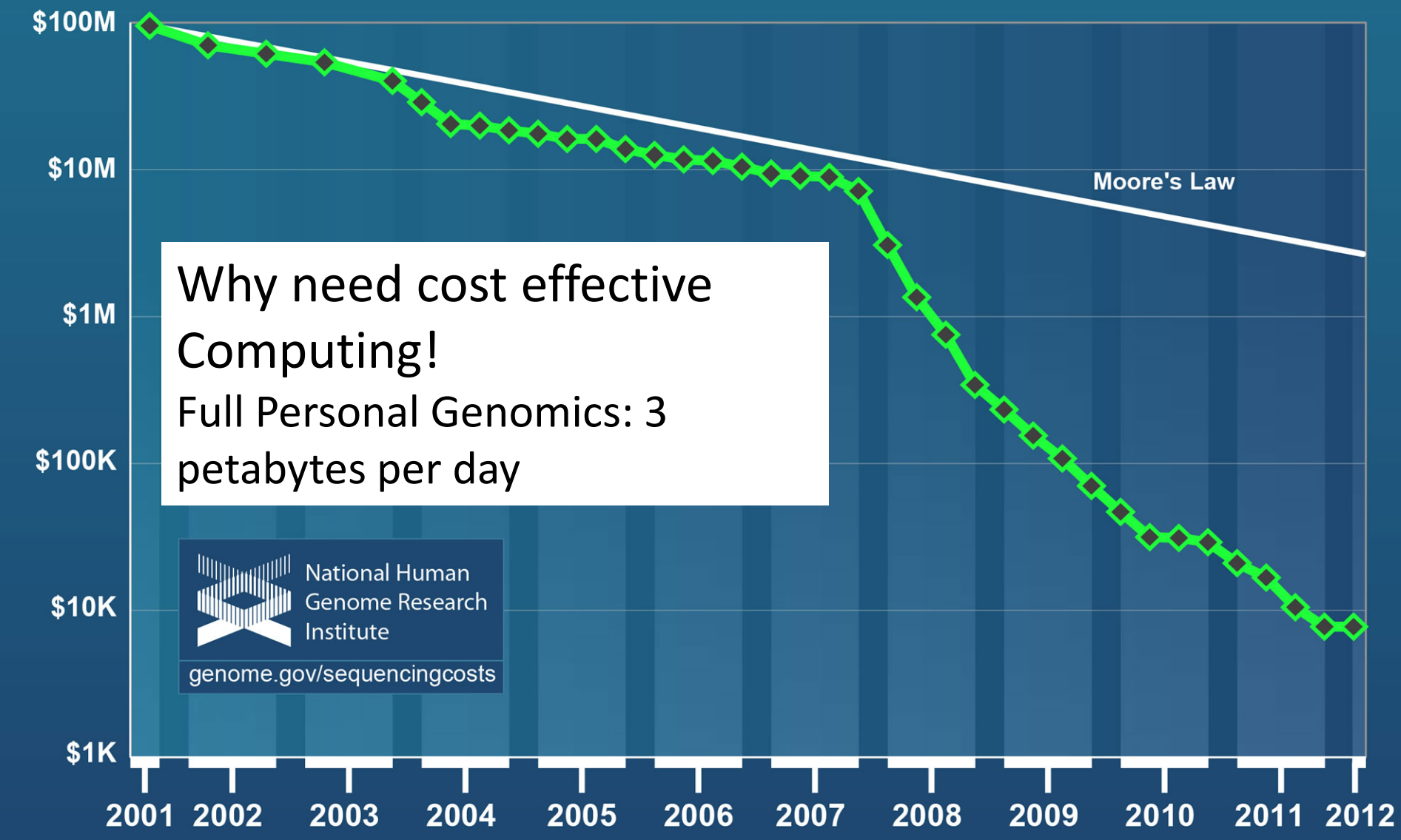
MM = Million

imagination at work

Ruh VP Software GE http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

is 50/50 JV with Pratt & Whitney

Cost per Genome



Why need cost effective Computing!

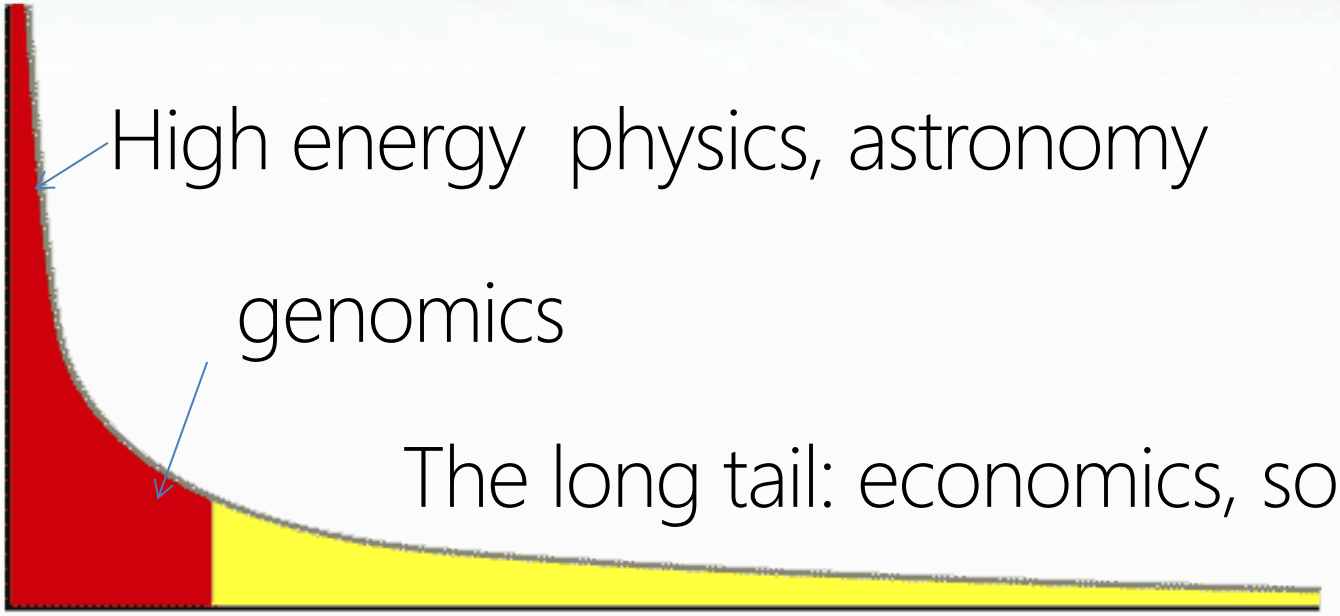
Full Personal Genomics: 3 petabytes per day

National Human Genome Research Institute

genome.gov/sequencingcosts

http://www.genome.gov/sequencingcosts/

The Long Tail of Science



Collectively "long tail" science is generating a lot of data
Estimated at over 1PB per year and it is growing fast.

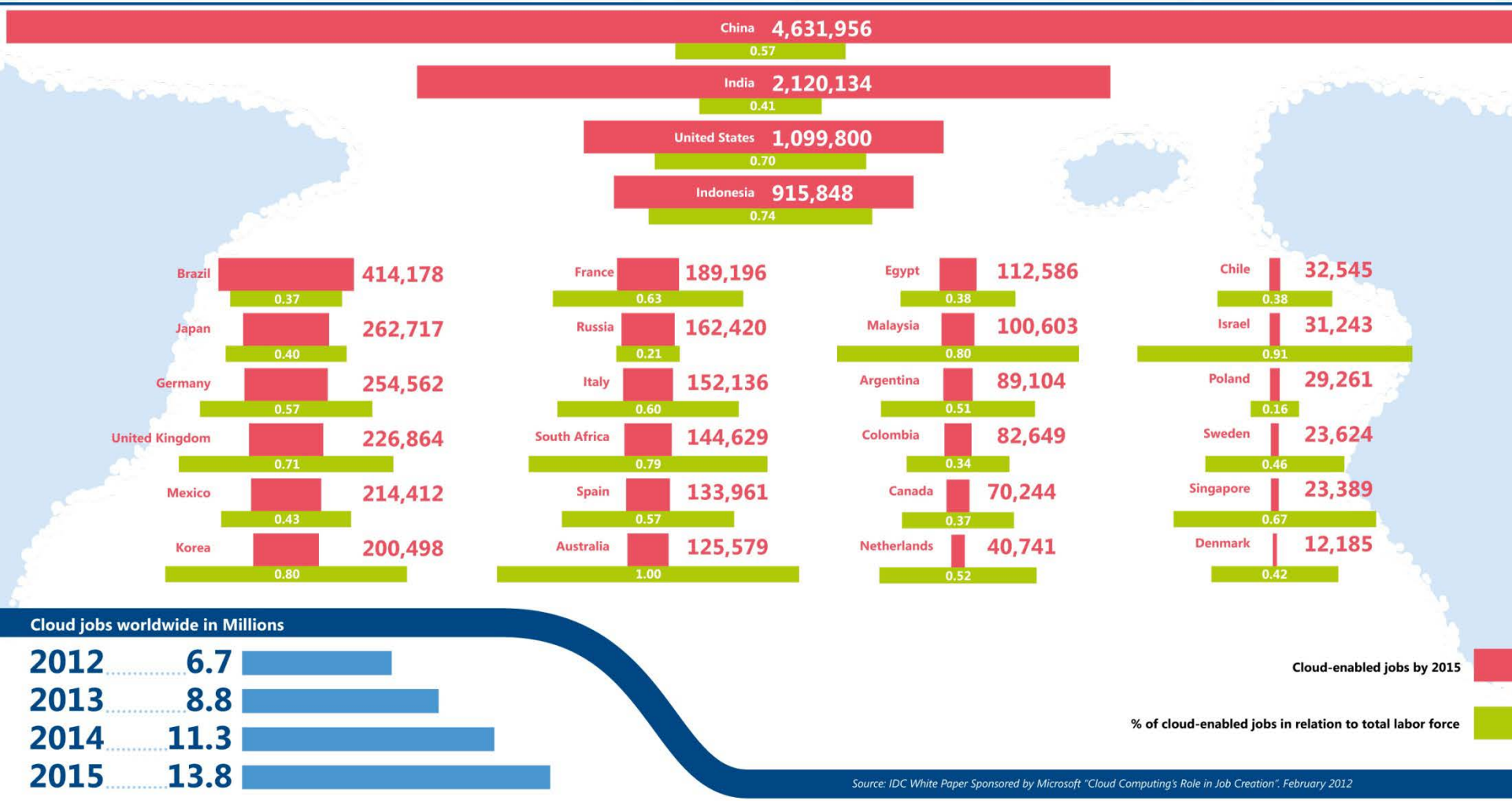
80-20 rule: 20% users generate 80% data but not necessarily 80% knowledge

From Dennis Gannon Talk

Economic Imperative
Now Jobs

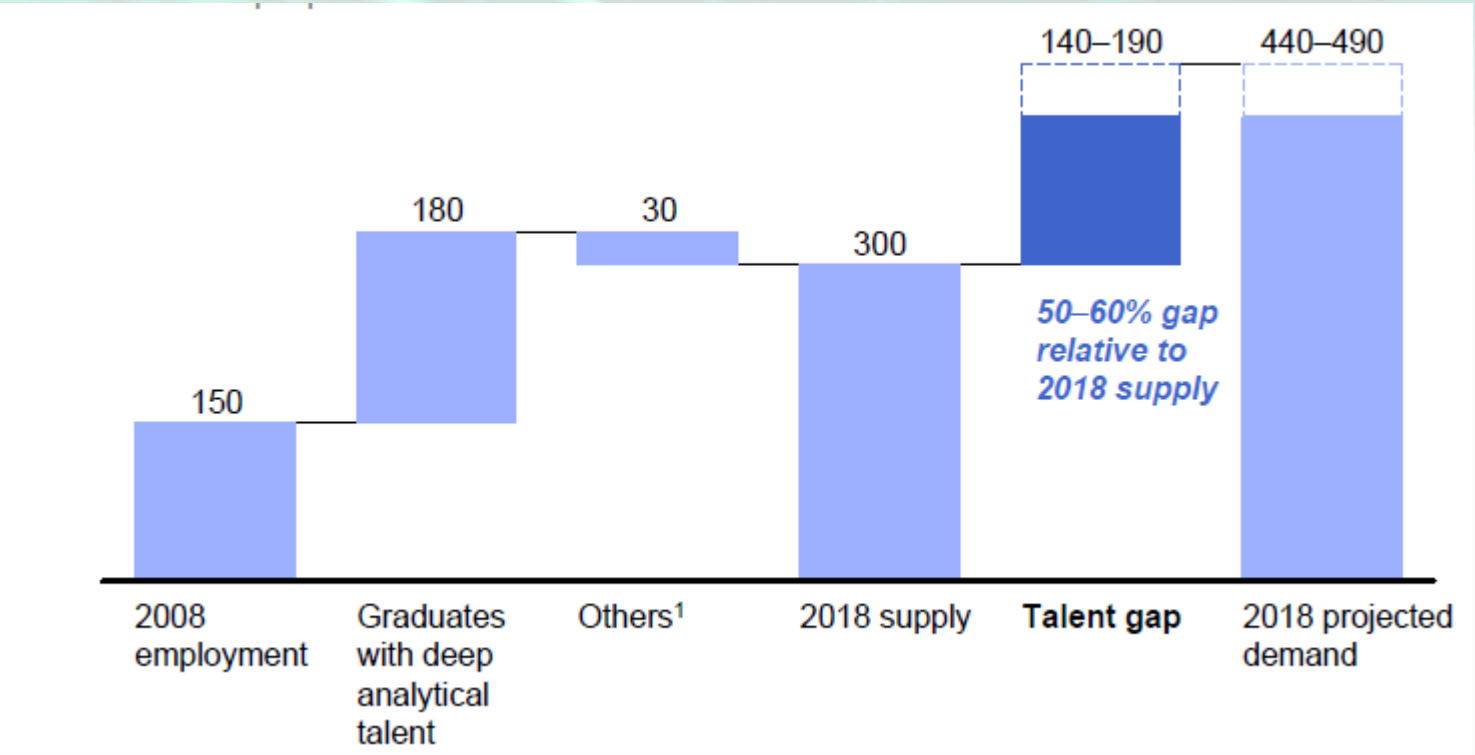
There are a lot of data and a lot of jobs

Jobs v. Countries



http://www.microsoft.com/en-us/news/features/2012/mar12/03-05CloudComputingJobs.aspx

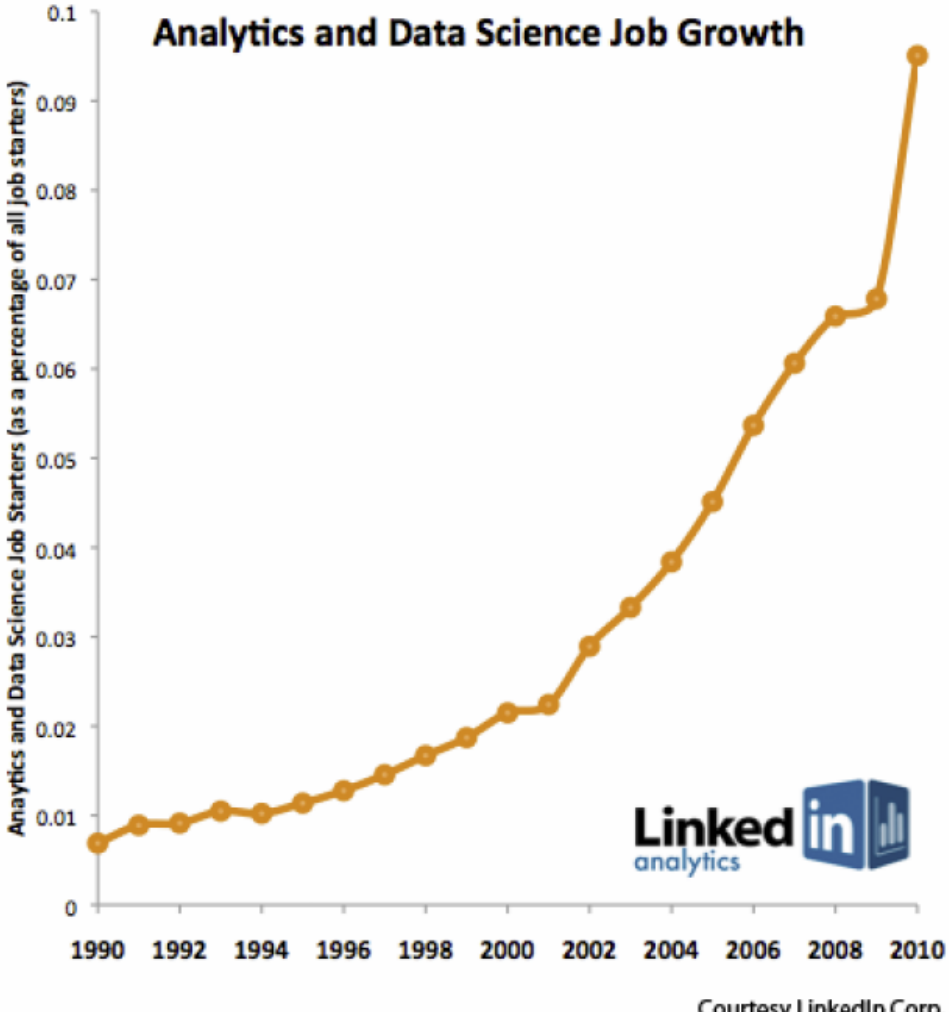
McKinsey Institute on Big Data Jobs



- There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.
- Informatics aimed at 1.5 million jobs. Computer Science covers the 140,000 to 190,000

http://www.mckinsey.com/mgi/publications/big_data/index.asp.

The Rise of Data Scientists and Analysts



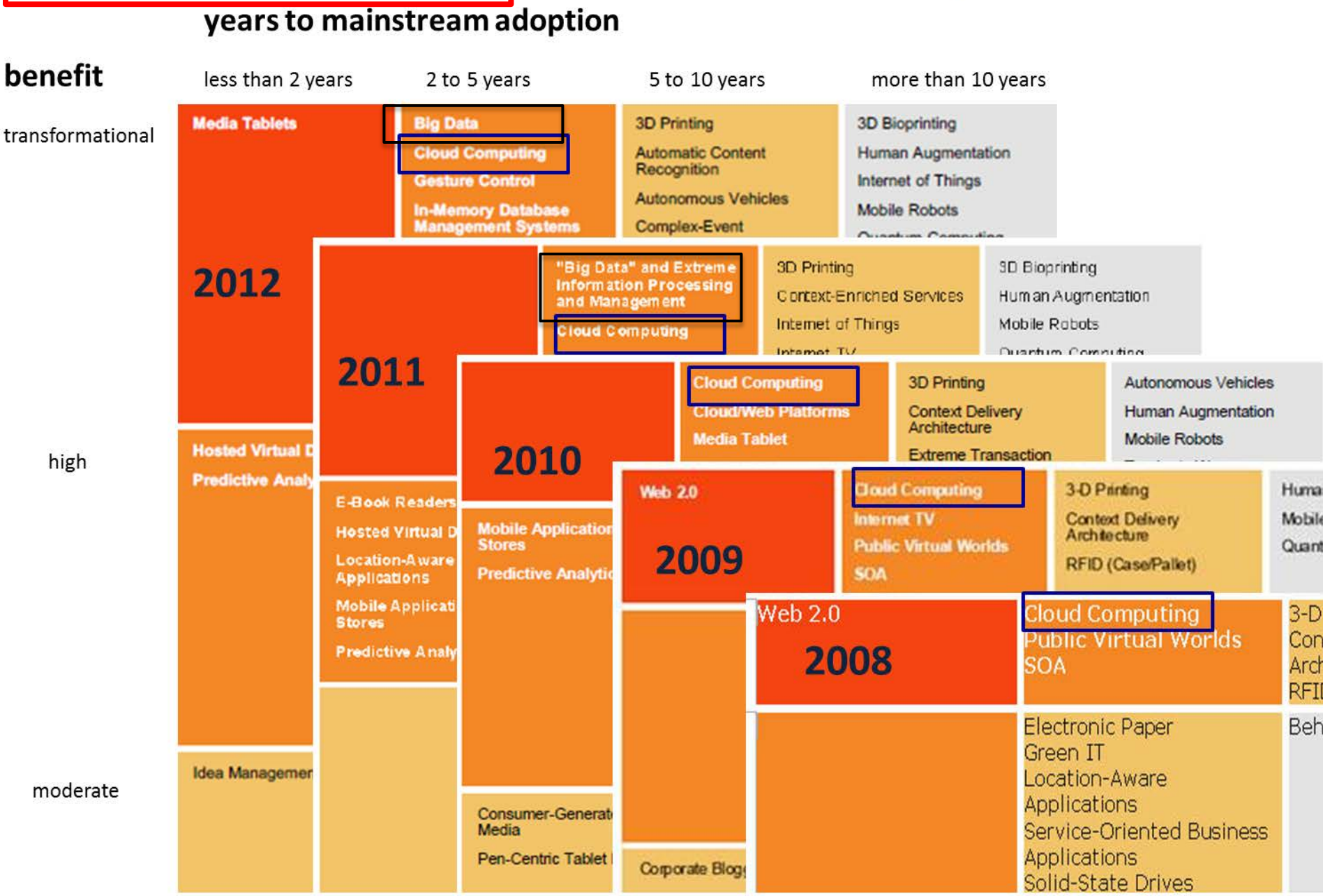
Tom Davenport Harvard Business School
http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html Nov 2012

Computing Model

Industry adopted clouds which are attractive for data analytics

5 years Cloud Computing
2 years Big Data Transformational

Gartner Priority Matrix



Amazon making money

- It took Amazon Web Services (AWS) eight years to hit $650 million in revenue, according to Citigroup in 2010.
- Just three years later, Macquarie Capital analyst Ben Schachter estimates that AWS will top $3.8 billion in 2013 revenue, up from $2.1 billion in 2012 (estimated), valuing the AWS business at $19 billion.

Physically Clouds are Clear

- A bunch of computers in an efficient data center with an excellent Internet connection
- They were produced to meet need of public-facing Web 2.0 e-Commerce/Social Networking sites
- They can be considered as “optimal giant data center” plus internet connection
- Note enterprises use private clouds that are giant data centers but not optimized for Internet access

Virtualization made several things more convenient

- Virtualization = abstraction; run a job – you know not where
- Virtualization = use hypervisor to support “images”
  - Allows you to define complete job as an “image” – OS + application
- Efficient packing of multiple applications into one server as they don’t interfere (much) with each other if in different virtual machines;
- They interfere if put as two jobs in same machine as for example must have same OS and same OS services
- Also security model between VM’s more robust than between processes

Clouds Offer From different points of view

- **Features from NIST:**
  - On-demand service (elastic);
  - Broad network access;
  - Resource pooling;
  - Flexible resource allocation;
  - Measured service
- **Economies of scale** in performance and electrical power **(Green IT)**
- Powerful new **software models**
  - **Platform as a Service** is not an alternative to **Infrastructure as a Service –** it is instead an incredible valued added
  - Amazon is as much PaaS as Azure
- They are **cheaper than classic clusters** unless latter 100% utilized

Research Model

4th Paradigm; From Theory to Data driven science?

http://www.wired.com/wired/issue/16-07 September 2008

The End of Science

The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.

The 4 paradigms of Scientific Research

1. **Theory**
2. **Experiment or Observation**
   - E.g. Newton observed apples falling to design his theory of mechanics
3. **Simulation of theory or model (computational Science)**
4. **Data-driven** (Big Data) or The Fourth Paradigm: Data-Intensive Scientific Discovery (aka Data Science)
   - http://research.microsoft.com/en-us/collaboration/fourthparadigm/ A free book
   - More data; less models

- Note Data → Information→ Wisdom →Knowledge→ Decisions pipeline

More data usually beats better algorithms

Here's how the competition works. Netflix has provided a large data set that tells you how nearly half a million people have rated about 18,000 movies. Based on these ratings, you are asked to predict the ratings of these users for movies in the set that they have not rated. The first team to beat the accuracy of Netflix's proprietary algorithm by a certain margin wins a prize of $1 million!

Different student teams in my class adopted different approaches to the problem, using both published algorithms and novel ideas. Of these, the results from two of the teams illustrate a broader point. Team A came up with a very sophisticated algorithm using the Netflix data. Team B used a very simple algorithm, but they added in additional data beyond the Netflix set: information about movie genres from the Internet Movie Database(IMDB). Guess which team did better?

Anand Rajaraman is Senior Vice President at Walmart Global eCommerce, where he heads up the newly created @WalmartLabs,

http://anand.typepad.com/datawocky/2008/03/more-data-usual.html

Data Science Education

Opportunities at universities

see recent New York Times articles

http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

Data Science Education

- Broad Range of Topics from Policy to curation to applications and algorithms, programming models, data systems, statistics, and broad range of CS subjects such as Clouds, Programming, HCI,
- Plenty of Jobs and broader range of possibilities than computational science but similar cosmic issues
  - What type of degree (Certificate, minor, track, “real” degree)
  - What implementation (department, interdisciplinary group supporting education and research program)

Computational Science

- Interdisciplinary field between computer science and applications with primary focus on simulation areas
- Very successful as a research area
  - XSEDE and Exascale systems enable
- Several academic programs but these have been less successful than computational science research as
  - No consensus as to curricula and jobs (don't appoint faculty in computational science; do appoint to DoE labs)
  - **Field relatively small**
- Started around 1990

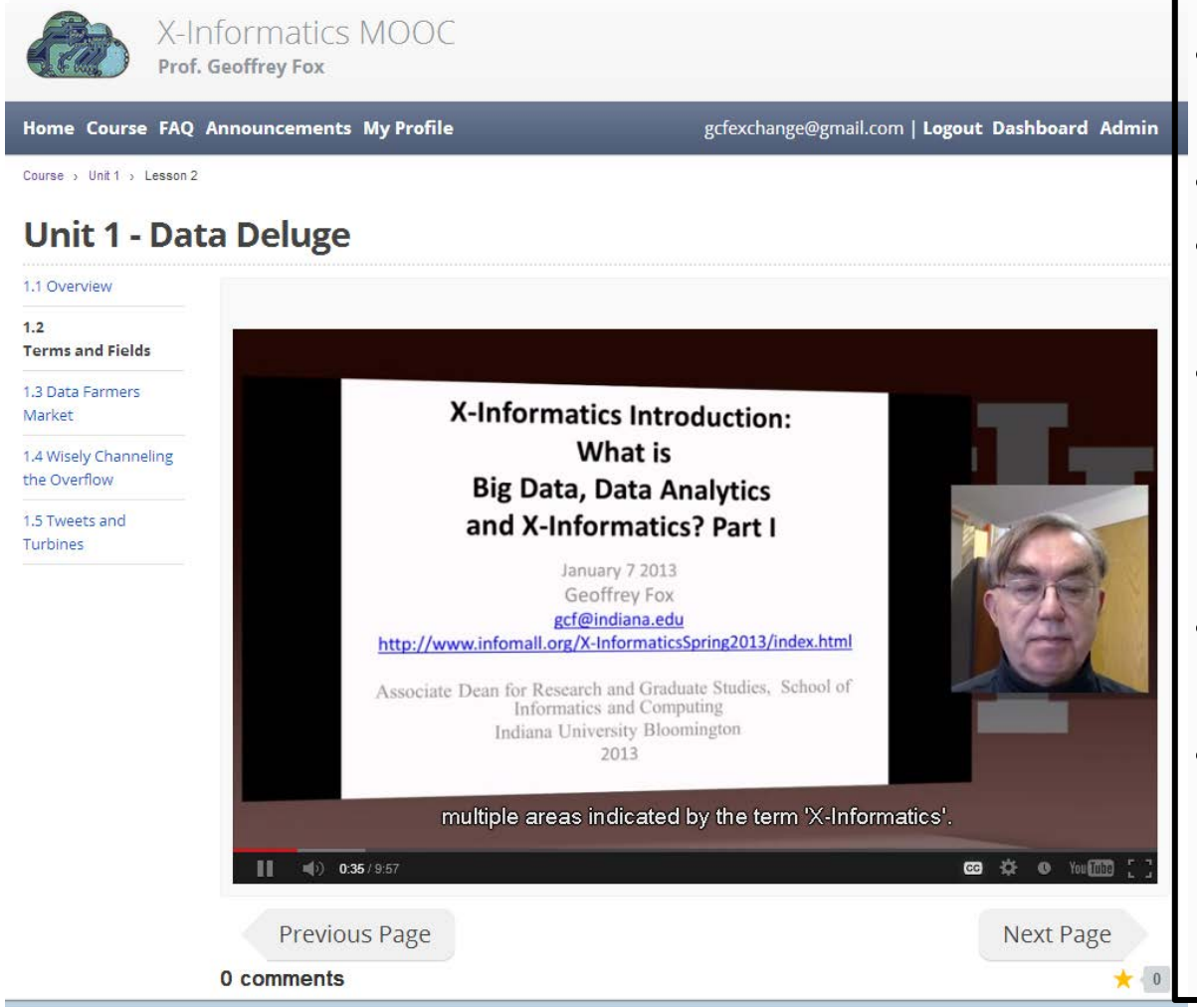
MOOC's

Massive Open Online Courses (MOOC)

- MOOC's are very "hot" these days with Udacity and Coursera as start-ups
- Over 100,000 participants but concept valid at smaller sizes
- Relevant to **Data Science as this is a new field with few courses** at most universities
- Technology to make MOOC's: Google Open Source **Course Builder** is lightweight LMS (learning management system)
- Supports MOOC model as a collection of short prerecorded segments (talking head over PowerPoint) termed **lessons** – typically 15 minutes long
- Compose playlists of lessons into sessions, modules, courses
  - Session is an "Album" and lessons are "songs" in an iTunes analogy

MOOC's for Traditional Lectures

- We can take MOOC lessons and view them as a "learning object" that we can share between different teachers



- i.e. as a way of teaching typical sized classes but with less effort as shared material
- Start with what's in repository;
- pick and choose;
- Add custom material of individual teachers
- The ~15 minute Video over PowerPoint of MOOC's much easier to re-use than PowerPoint
- Do not need special mentoring support
- Defining how to support computing labs with FutureGrid or appliances + Virtual Box

Confusion in the new-old data field

lack of consensus academically in several aspects from storage to algorithms, to processing and education

Data Communities Confused I?

- Industry seems to know what it is doing although it's secretive – Amazon's last paper on their recommender system was 2003
  - Industry runs the largest data analytics on clouds
  - But industry algorithms are rather different from science
- **Academia confused on repository model:** traditionally one stores data but one needs to support "running Data Analytics" and one is taught to bring computing to data as in Google/Hadoop file system
  - Either store data in compute cloud OR enable high performance networking between distributed data repositories and "analytics engines"
- **Academia confused on data storage model:** Files (traditional) v. Database (old industry) v. NOSQL (new cloud industry)
  - Hbase MongoDB Riak Cassandra are typical NOSQL systems
- **Academia confused on curation of data:** University Libraries, Projects, National repositories, Amazon/Google?

Data Communities Confused II?

- **Academia agrees on principles of Simulation Exascale Architecture:** HPC Cluster with accelerator plus parallel wide area file system
  - Industry doesn't make extensive use of high end simulation
- **Academia confused on architecture for data analysis:** Grid (as in LHC), Public Cloud, Private Cloud, re-use simulation architecture with database, object store, parallel file system, HDFS style data
- **Academia has not agreed on Programming/Execution model:** "Data Grid Software", MPI, MapReduce ..
- **Academia has not agreed on need for new algorithms:** Use natural extension of old algorithms, R or Matlab. Simulation successes built on great algorithm libraries;
- Academia has not agreed on **what algorithms are important?**
- **Academia could attract more students**: with data-oriented curricula that prepare for industry or research careers (as discussed)

Clouds in Research

Clouds have highlighted SaaS PaaS IaaS

Software (Application Or Usage) SaaS
- Education
- Applications
- CS Research Use e.g. test new compiler or storage model

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

Infra structure IaaS
- Software Defined Computing (virtual Clusters)
- Hypervisor, Bare Metal
- Operating System

Network NaaS
- Software Defined Networks
- OpenFlow GENI

But equally valid for classic clusters

- Software Services are building blocks of applications
- The middleware or computing environment including **HPC, Grids** …
- Nimbus, Eucalyptus, OpenStack, OpenNebula CloudStack plus **Bare-metal**
- OpenFlow – *likely to grow in importance*

Science Computing Environments

- **Large Scale Supercomputers** – Multicore nodes linked by high performance low latency network
  - Increasingly with GPU enhancement
  - Suitable for highly parallel simulations
- **High Throughput Systems** such as European Grid Initiative EGI or Open Science Grid OSG typically aimed at pleasingly parallel jobs
  - Can use "cycle stealing"
  - Classic example is **LHC data analysis**
- **Grids** federate resources as in EGI/OSG or enable convenient access to multiple backend systems including supercomputers
- Use **Services** (**SaaS**)
  - **Portals** make access convenient and
  - **Workflow** integrates multiple processes into a single job

Clouds HPC and Grids

- Synchronization/communication Performance
  Grids > Clouds > Classic HPC Systems
- **Clouds** naturally execute effectively **Grid** workloads but are less clear for closely coupled HPC applications
- **Classic HPC machines** as MPI engines offer highest possible performance on closely coupled problems
- The 4 forms of MapReduce/MPI
  1) **Map Only** – pleasingly parallel
  2) **Classic MapReduce** as in Hadoop; single Map followed by reduction with fault tolerant use of disk
  3) **Iterative MapReduce** use for data mining such as Expectation Maximization in clustering etc.; Cache data in memory between iterations and support the **large collective communication** (Reduce, Scatter, Gather, Multicast) use in data mining
  4) **Classic MPI!** Support small point to point messaging efficiently as used in partial differential equation solvers

What Applications work in Clouds

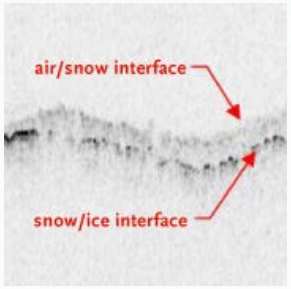
- **Pleasingly (moving to modestly) parallel** applications of all sorts (over **users** or **usages**) with roughly independent data or spawning independent simulations
  - **Long tail** of science and integration of distributed sensors
- **Commercial and Science Data analytics** that can use MapReduce **(**some of such apps) or its **iterative** variants (most other data analytics apps)
- **Which science applications are using clouds**?
  - **Venus-C** (Azure in Europe): 27 applications **not using** Scheduler, Workflow or MapReduce (except roll your own)
  - 50% of applications on **FutureGrid** are from Life Science
  - Locally **Lilly** corporation is commercial cloud user (for drug discovery) but not IU Biology
- **But overall very little science use of clouds**

Internet of Things and the Cloud

- It is projected that there will be **24 billion devices** on the Internet by 2020. Most will be small sensors that send streams of information into the cloud where it will be processed and integrated with other streams and turned into knowledge that will help our lives in a multitude of small and big ways.
- The **cloud** will become increasing important as a controller of and **resource provider for the Internet of Things.**
- As well as today's use for smart phone and gaming console support, "Intelligent River" "smart homes and grid" and "ubiquitous cities" build on this vision and we could expect a growth in cloud supported/controlled **robotics**.
- Some of these "things" will be supporting science
- Natural parallelism over "things"
- "Things" are distributed and so form a Grid

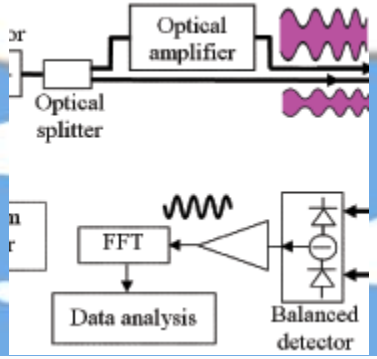
Sensors (Things) as a Service

Output Sensor



A larger sensor ……

Sensors as a Service



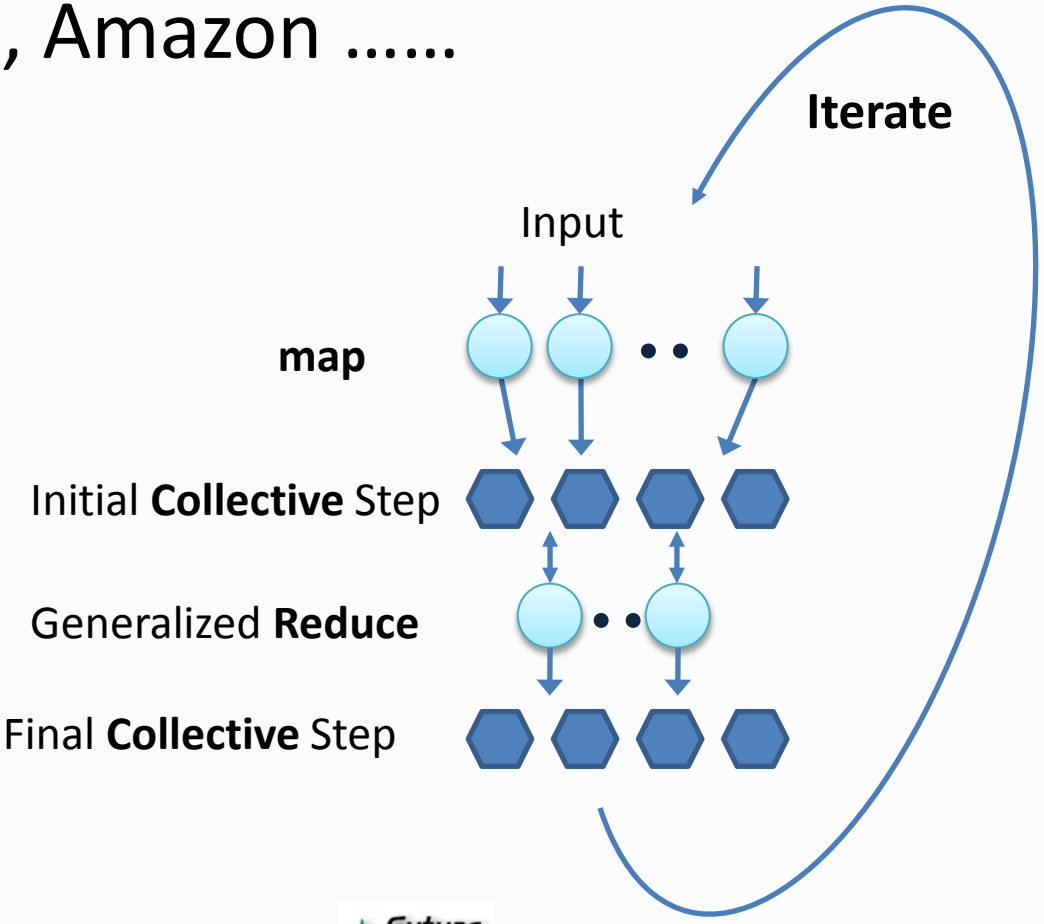
Sensor Processing as a Service (could use MapReduce)

https://sites.google.com/site/opensourceiotcloud/ Open Source Sensor (IoT) Cloud

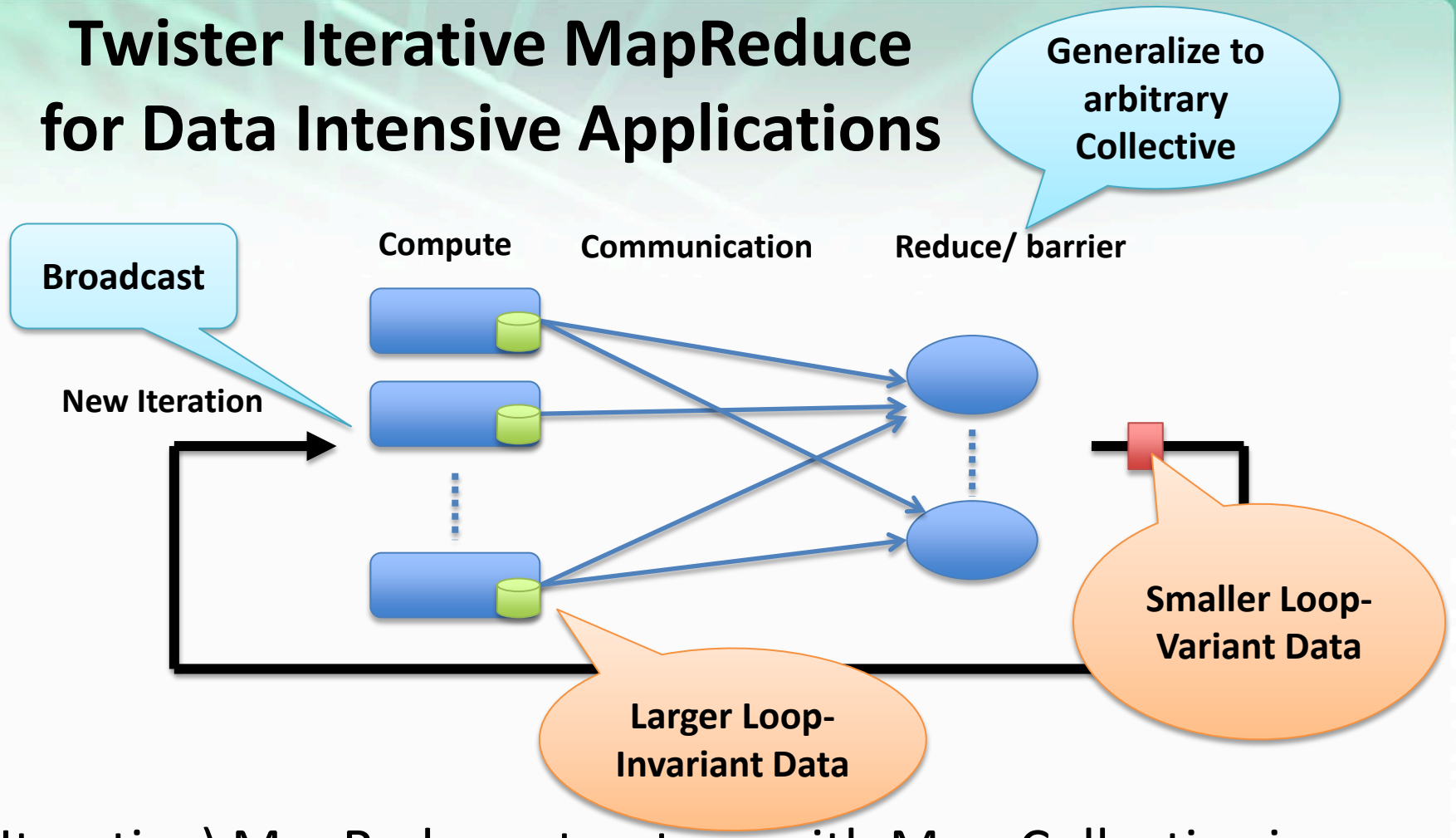
Data Intensive Programming Models

Map Collective Model (Judy Qiu)

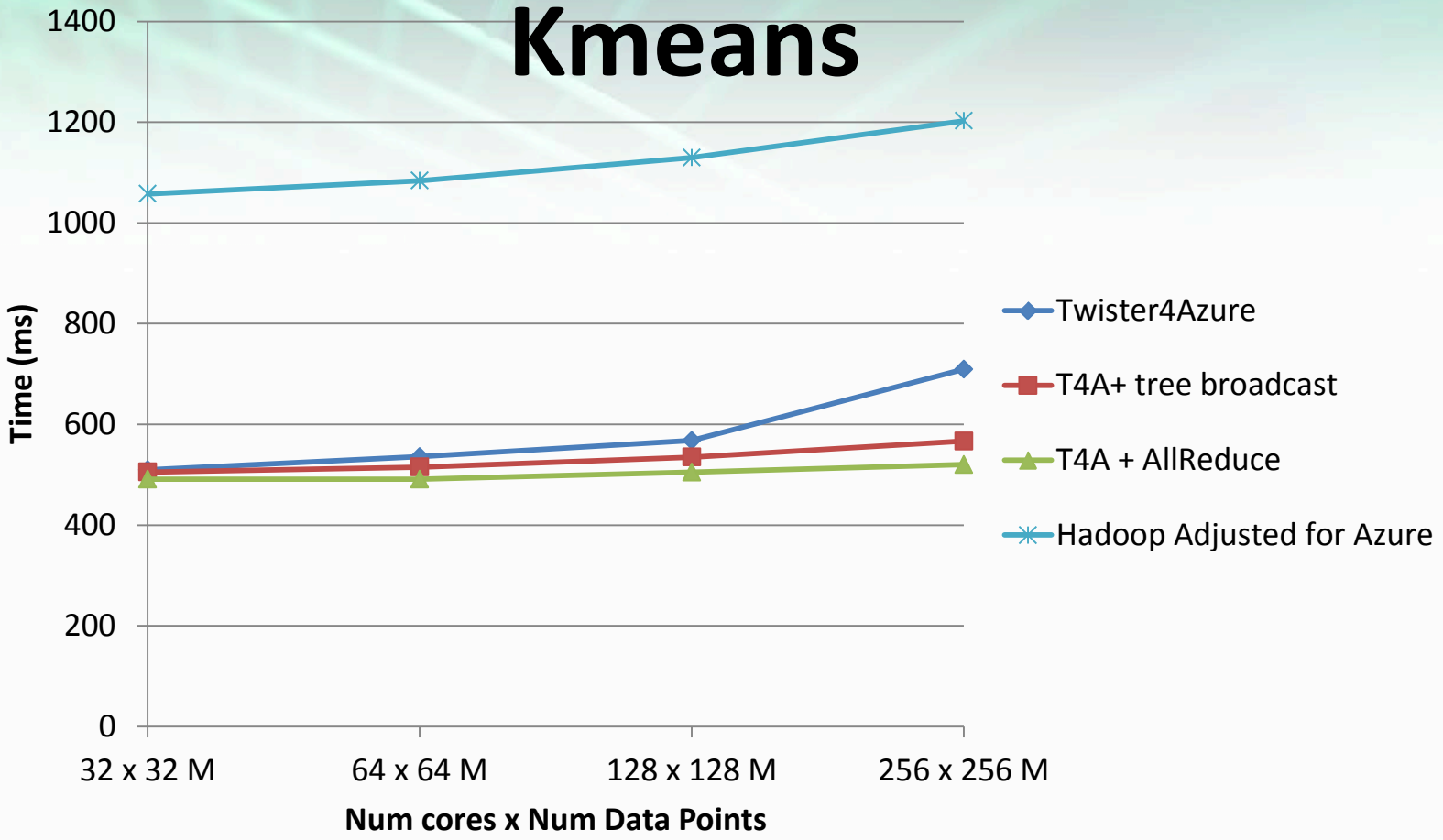
- Combine MPI and MapReduce ideas
- Implement collectives optimally on Infiniband, Azure, Amazon ……



Twister Iterative MapReduce for Data Intensive Applications



- (Iterative) MapReduce structure with Map-Collective is framework
- Twister runs on Linux or Azure
- Twister4Azure is built on top of Azure **tables**, **queues**, **storage**



Hadoop adjusted for Azure: Hadoop KMeans run time adjusted for the performance difference of iDataplex vs Azure

FutureGrid Technology addressing Poor Cloud Performance?

FutureGrid Testbed as a Service

- FutureGrid is part of XSEDE set up as a testbed with cloud focus
- Operational since Summer 2010 (i.e. now in third year of use)
- The FutureGrid testbed provides to its users:
  - Support of Computer Science and Computational Science research
  - 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; supports both classic HPC and Cloud storage

4 Use Types for FutureGrid TestbedaaS

- **292** approved projects (**1734 users**) **April 6 2013**
  - USA(79%), Puerto Rico(3%- Students in class), India, China, lots of European countries (Italy at 2% as class)
  - Industry, Government, Academia
- **Computer science and Middleware (55.6%)**
  - Core CS and Cyberinfrastructure; Interoperability (**3.6%**) for Grids and Clouds such as Open Grid Forum OGF Standards
- **New Domain Science applications (20.4%)**
  - Life science highlighted (10.5%), Non Life Science (9.9%)
- **Training Education and Outreach (14.9%)**
  - Long (27 full semester) and short events
- **Computer Systems Evaluation (9.1%)**
  - XSEDE (TIS, TAS), OSG, EGI; Campuses

Sample FutureGrid Projects I

- **FG18 Privacy preserving gene read mapping** developed hybrid MapReduce. Small private secure + large public with safe data. Won 2011 PET Award for Outstanding Research in Privacy Enhancing Technologies
- **FG132, Power Grid Sensor analytics on the cloud** with distributed Hadoop. Won the IEEE Scaling challenge at CCGrid2012.
- **FG156 Integrated System for End-to-end High Performance Networking** showed that the RDMA over Converged Ethernet (InfiniBand made to work over Ethernet network frames) protocol could be used over wide-area networks, making it viable in cloud computing environments.
- **FG172 Cloud-TM** on distributed concurrency control (software transactional memory): "When Scalability Meets Consistency: Genuine Multiversion Update Serializable Partial Data Replication," 32nd International Conference on Distributed Computing Systems (ICDCS'12) (good conference) used 40 nodes of FutureGrid

Sample FutureGrid Projects II

- **FG42,45 SAGA Pilot Job** P* abstraction and applications. XSEDE Cyberinfrastructure used on clouds
- **FG130 Optimizing Scientific Workflows** on Clouds. Scheduling Pegasus on distributed systems with overhead measured and reduced. Used Eucalyptus on FutureGrid
- **FG133 Supply Chain Network Simulator** Using Cloud Computing with dynamic virtual machines supporting Monte Carlo simulation with Grid Appliance and Nimbus
- **FG257 Particle Physics Data analysis for ATLAS LHC** experiment used FutureGrid + Canadian Cloud resources to study data analysis on Nimbus + OpenStack with up to 600 simultaneous jobs
- **FG254 Information Diffusion in Online Social Networks** is evaluating NoSQL databases (Hbase, MongoDB, Riak) to support analysis of Twitter feeds
- **FG323 SSD performance benchmarking** for HDFS on Lima

Education and Training Use of FutureGrid

- 27 Semester long classes: 563+ students
  - Cloud Computing, Distributed Systems, Scientific Computing and Data Analytics
- 3 one week summer schools: 390+ students
  - Big Data, Cloudy View of Computing (for HBCU's), Science Clouds
- 1 two day workshop: 28 students
- 5 one day tutorials: 173 students
- From 19 Institutions
- Developing 2 MOOC's (Google Course Builder) on Cloud Computing and use of FutureGrid supported by either FutureGrid or downloadable appliances (custom images)
  - See http://cgltestcloud1.appspot.com/preview
- FutureGrid appliances support Condor/MPI/Hadoop/Iterative MapReduce virtual clusters

Clouds have highlighted SaaS PaaS IaaS

Software (Application Or Usage) SaaS
- Education
- Applications
- CS Research Use e.g. test new compiler or storage model

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

Infra structure IaaS
- Software Defined Computing (virtual Clusters)
- Hypervisor, Bare Metal
- Operating System

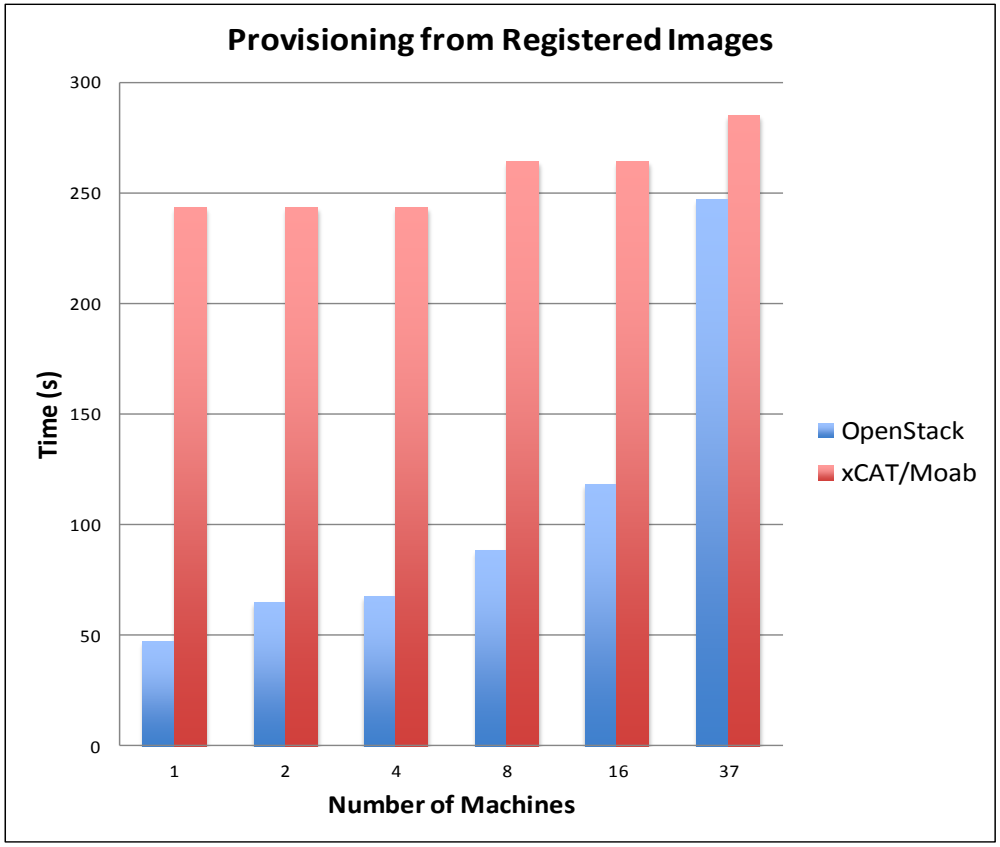
Network NaaS
- Software Defined Networks
- OpenFlow GENI

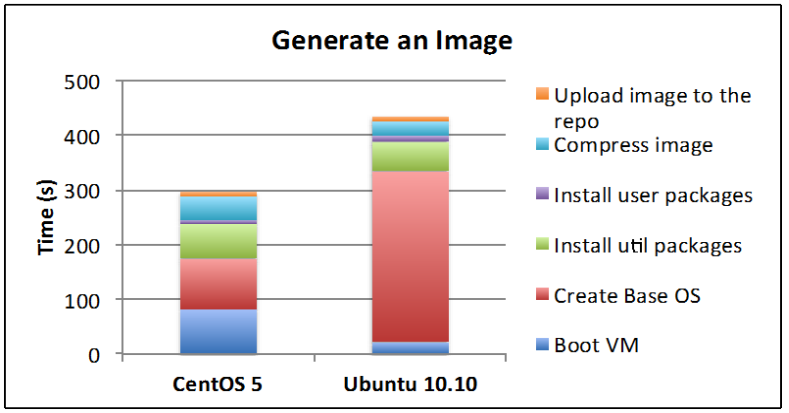
But equally valid for classic clusters

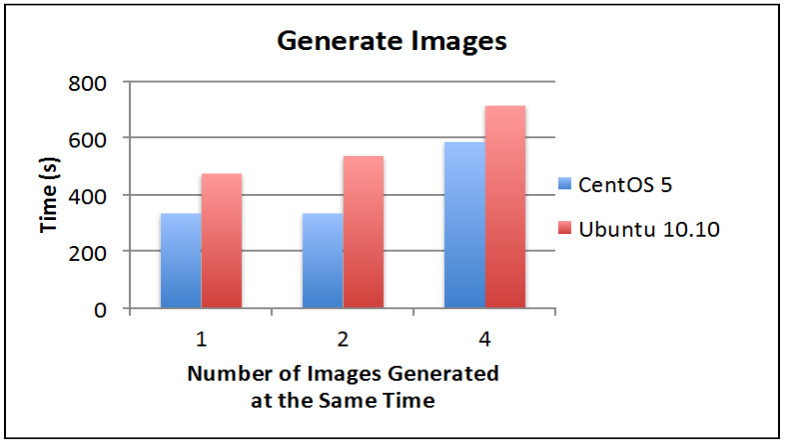
- Software Services are building blocks of applications
- The middleware or computing environment including **HPC, Grids** …
- Nimbus, Eucalyptus, OpenStack, OpenNebula CloudStack plus **Bare-metal**
- OpenFlow – *likely to grow in importance*

Performance of Dynamic Provisioning

- **4 Phases** a) Design and create image (security vet) b) Store in repository as template with components c) Register Image to VM Manager (cached ahead of time) d) Instantiate (Provision) image



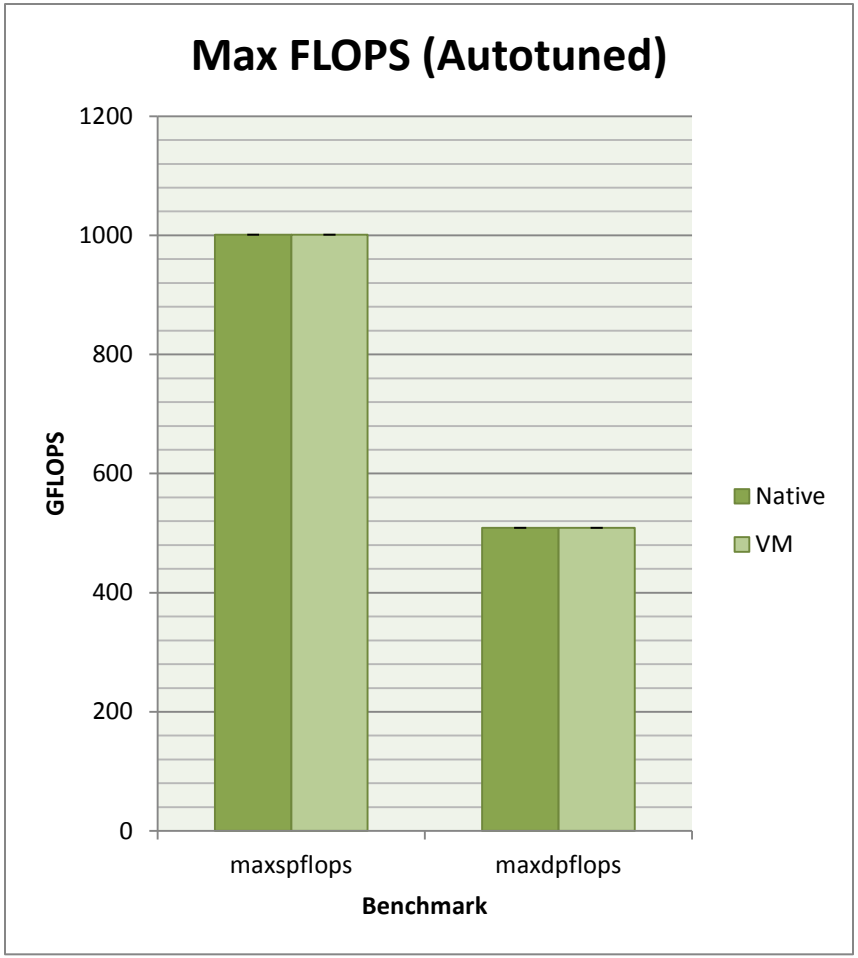


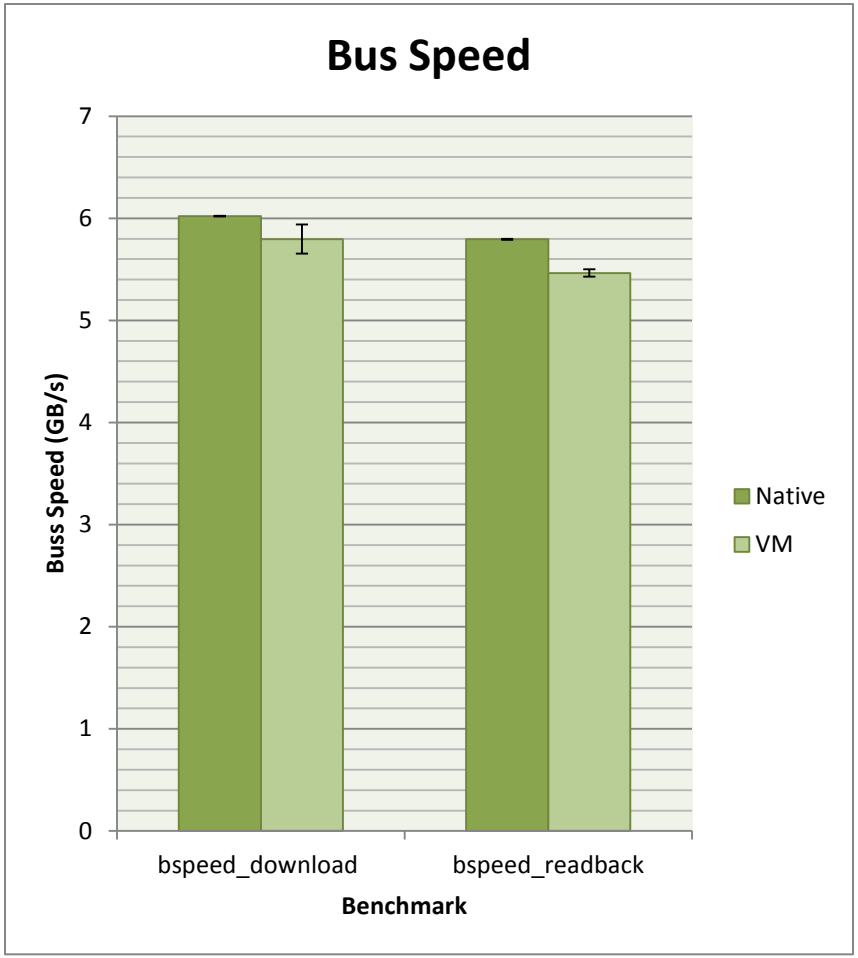


Direct GPU Virtualization

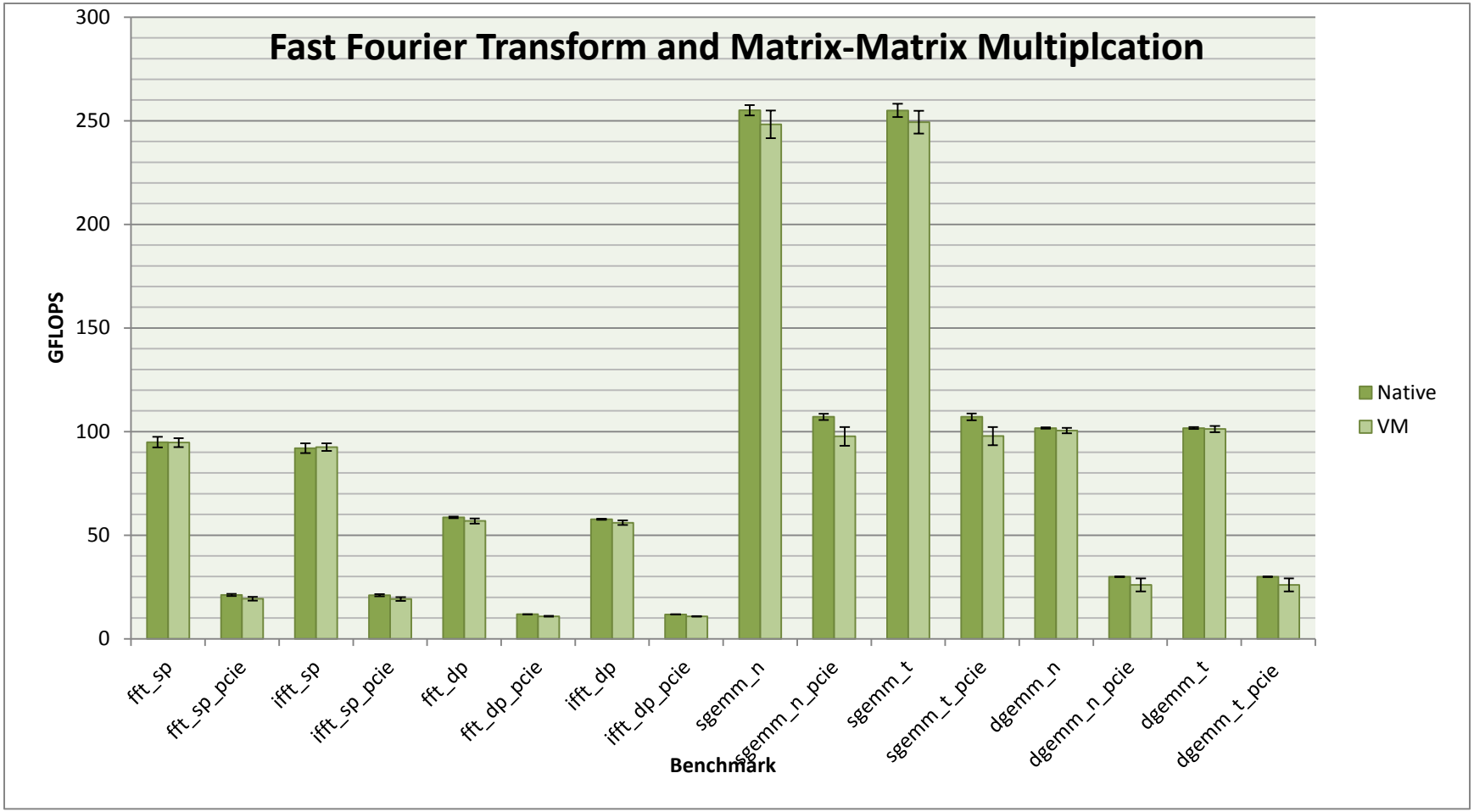
- Allow VMs to directly access GPU hardware
- Enables CUDA and OpenCL code – no need for custom APIs
- Utilizes PCI-passthrough of device to guest VM
  - Hardware directed I/O virt (VT-d or IOMMU)
  - Provides direct isolation and security of device from host or other VMs
  - Removes much of the Host <-> VM overhead
- Similar to what Amazon EC2 uses (proprietary)

Performance 1





Performance 2



Algorithms

Scalable Robust **Algorithms**: new data need better algorithms?

Algorithms for Data Analytics

- In simulation area, it is observed that equal contributions to improved performance come from increased computer power and better algorithms http://cra.org/ccc/docs/nitrdsymposium/pdfs/keyes.pdf
- In data intensive area, we haven't seen this effect so clearly
  - Information retrieval revolutionized but
  - Still using Blast in Bioinformatics (although Smith Waterman etc. better)
  - Still using R library which has many non optimal algorithms
  - Parallelism and use of GPU's often ignored

"Moore's Law" for fusion energy simulations

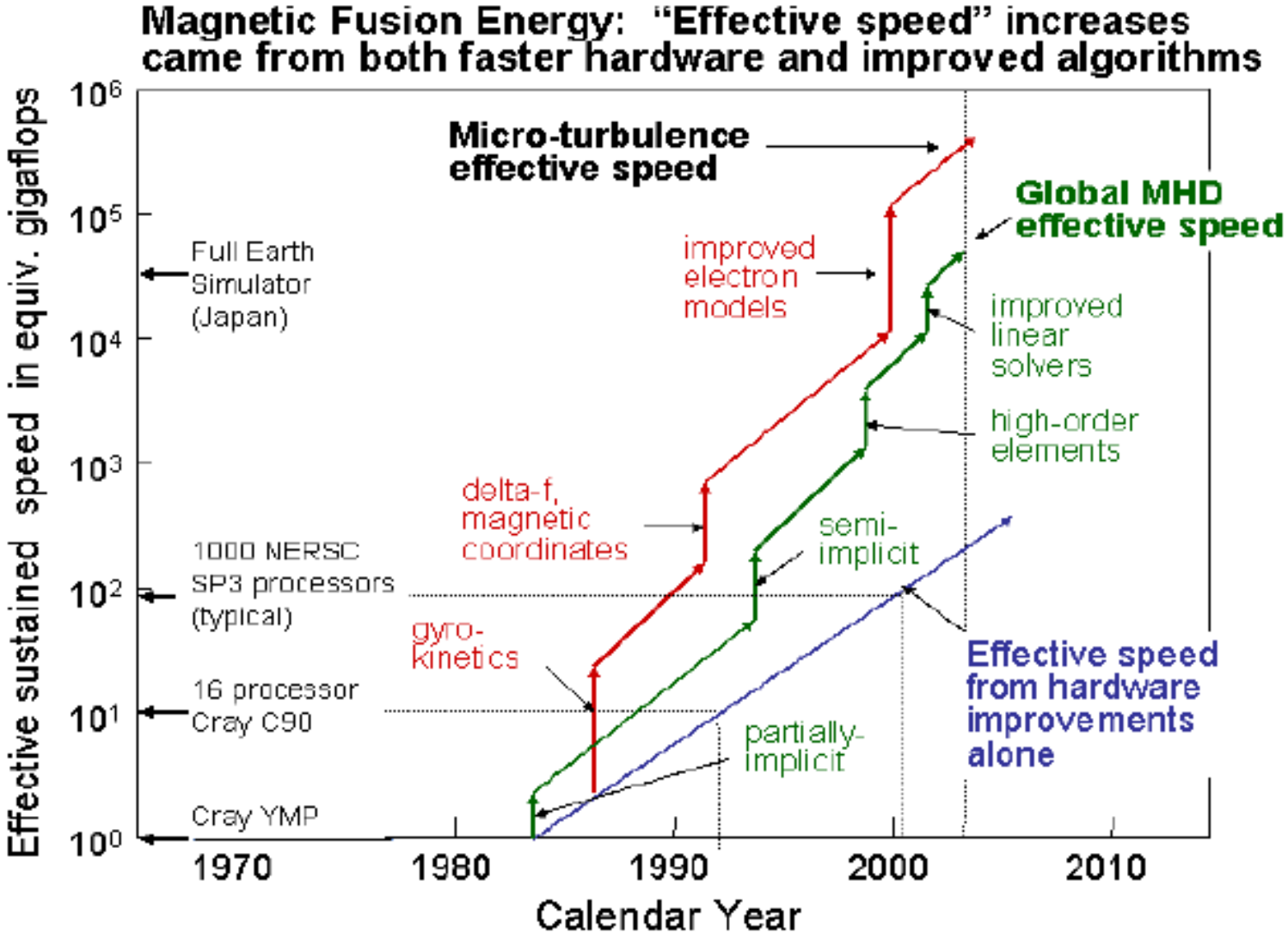


Figure from DOE "SCaLeS report" Volume 2 (Keyes et al., 2004)

Data Analytics Futures?

- **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**
  - Crosses Science, Business Network Science, Social Science
- Propose to build community to define & implement **SPIDAL** or **Scalable Parallel Interoperable Data Analytics Library**

Conclusions

Conclusions

- Clouds and HPC are here to stay and one should plan on using **both**
- **Data Intensive** programs are not like simulations as they have **large "reductions" ("collectives")** and do not have many small messages
  - Clouds suitable
- **Iterative MapReduce** an interesting approach; need to optimize collectives for new applications (Data analytics) and resources (clouds, GPU's …)
- Need an initiative to build **scalable high performance data analytics library** on top of **interoperable cloud-HPC platform**
- Many promising data analytics algorithms such as **deterministic annealing** not used as implementations not available in R/Matlab etc.
  - More sophisticated software and runs longer but can be **efficiently parallelized** so runtime not a big issue

Conclusions II

- **Software defined computing systems** linking NaaS, IaaS, PaaS, SaaS (Network, Infrastructure, Platform, Software) likely to be important
- More **employment opportunities** in clouds than HPC and Grids and in data than simulation; so cloud and data related activities popular with students
- Community activity to discuss **data science education**
  - Agree on curricula; is such a degree attractive?
- Role of **MOOC**'s as either
  - Disseminating new curricula
  - Managing course fragments that can be assembled into custom courses for particular interdisciplinary students