Cloud Innovation: Math & Science

Sylvia Spengler : Panel Moderator
National Science Foundation
NIST Cloud Computing Forum & Workshop III
April 8, 2011
Today’s Panelists

- **Vint Cerf** Vice President and Chief Internet Evangelist, Google
- **Dennis Gannon** Director of Cloud Research Engagements for the eXtreme Computing Group in Microsoft Research
- **Kate Keahey** Scientist, Argonne National Laboratory, Fellow, Computation Institute, University of Chicago
- **Joe Klimavicz** Chief Information Officer and Director, High Performance Computing and Communications, National Oceanic and Atmospheric Administration (NOAA)
- **Yelena Yesha** Professor, Computer Science and Electrical Engineering, University of Maryland at Baltimore
Democratizing Access to the Data Explosion with Cloud Computing

Dennis Gannon
Extreme Computing Group
Microsoft Research

dennis.gannon@microsoft.com
Microsoft’s Cloud Research Engagement Project

- **Basic premise**
  - 4th Paradigm: The data deluge is transforming all science.
  - Supercomputing is great ... for 2% of the research community

- **Our challenge**
  - Use the cloud as a platform to enable anybody to access data and analytics services anywhere

- **Our approach**
  - Build partnerships with government-sponsored research agencies and university consortia and provide researchers with free access to try new ideas
  - Build and release cloud data analytics and collaboration tools to the community
  - Let scientists focus on science. The cloud is in the background
Engagement Projects So Far (60)

• **NSF Computing in the Cloud**
  • 30 projects. Areas include genomics, phylogeny, algorithms for big data, sociology, semantic web, molecular dynamics, string theory, geo ....

• **European Commission and INRIA**
  • Venus-C - Demonstrate Cloud Interop
  • Applications: civil protection and emergencies, biomedicine, civil engineering, drug discovery, systems biology (20)

• **Japan National Institute for Informatics**
  • InfoPlosion : 7 projects around natural language and web data analytics

• **Australia – NICTA, National Comp Infrastructure, CSIRO**
  • 6 projects: Social network analysis, realtime data, computational chemistry, magnetotellurics, coal industry logistics
  • Future – More Asia Engagements
  • Additional – Massive scale protein folding, genomics, n-gram services, image analysis, great Kinect hacks in the community.
NCBI Blast on Windows Azure

- National Center for Biotechnology Information algorithm for comparing biological sequences
- Using the cloud, University of Washington reduced compute time on a project on protein interactions from 6 years to 1 week.
- Accelerating scientific discovery
- Released as open source
Next Step: Release Excel DataScope

• A framework of cloud services for data analytics with Excel as a user interface
  • Extensible collection of data analytics and machine learning algorithms running in the cloud
    • Filtering, anomaly detection, clustering, Bayesian nets, ...
  • Customizable data analytics ribbon.
  • Invoke models, perform analytics and visualization to gain insight from data.
  • Sample and manipulate extremely large data collections in the cloud;
  • Share data and results in the cloud, w/ annotations to facilitate data discovery and reuse

• We want to democratize access to advanced capabilities around big data for all scientists
Cloud Innovation for Science: How is Cloud Computing like a Horseless Carriage?

Kate Keahey
keahey@mcs.anl.gov
Mathematics and Computer Science Division,
Argonne National Laboratory
Computation Institute, University of Chicago
Horseless Carriage

Technical Innovations

Ecosystem

New Patterns
Cloud Computing as Horseless Carriage

Cloud Computing Today

- Isolation/virtualization
-aaSes of all kinds

Cloud Computing Tomorrow

- Technical Innovations
  - Control over environment
  - Availability
  - Performance & Reliability
  - Security and privacy
  - Cost models

- New Patterns
  - Elastic computing
  - Pay-as-you-go
  - Multi-tennancy

- Ecosystem
  - Appliance management
  - Platform tools
  - SLAs and legal underpinnings
  - Standards
  - Cloud markets
Parting Thoughts

• Outsourcing computing for science
  – Benefits
    • Economy of scale, access to different resources, no operation overhead, more flexible use
  – Criteria
    • Does it provide the right offering? Is it scalable? Easy to use? Easy to outsource? Cost-effective?
  – Outsourcing by degrees

• A virtuous spiral
  – Mohammad and the Mountain: will clouds change or will scientific computing?
  – See www.scienceclouds.org
Barriers in porting HPC to the cloud

• Performance implications of running the scientific applications on cloud vis-a-vis dedicated infrastructure.
  • Clouds are not optimized for HPC and also may suffer during high speed interconnections in HPC.
• Non-standardized data formats used in HPC not supported by cloud applications
• Security concerns – Multi-tenancy of Clouds
• Data privacy concerns – specially for public / third party provided Clouds
• Lack of intuitive interfaces for scientists – “problem solving environments”
Data sharing

Present:
- Setting up large data clearinghouses in domain specific verticals: Example: Genetics (Genbank)
- Sharing data still has privacy concerns - No incentive to share data
- Scientists working on mid-range clusters (non-HPC) are ideal cases for migrating to clouds that can do efficient resource utilization through virtualization

Future:
- Develop new approaches for data privacy and information sharing
- Incentivize data sharing
- Demonstrate low-hanging “wins” by sharing scientific data on the cloud
Legacy Scientific applications

- Legacy applications will continue to exist for some decades
- Hybrid clouds will be needed to manage legacy systems
  - Legacy code will remain on internal infrastructure
  - Data can be shared on public/community clouds
- Once applications are decommissioned, the legacy data, if needed, can be migrated to a public/community cloud or archived
Educational Opportunities

• Universities have begun offering certificate courses in Cloud Computing
• Very few universities have production cloud environments
  • Most begin by migrated their Internet/email functions to the cloud
• Even fewer have clouds in a laboratory setting to teach courses
• Grad students can work on Key Open issues in cloud research
  • Cloud Services discovery, acquisition and composition; governance issues
  • Virtualization: New architectures, OS designed for cloud environment
  • Power Management of cloud datacenters: Task balance, migration between Nodes
Cloud research @UMBC

- MC2 CHMPR lab focused on Cloud research
  - Internal Cloud infrastructure available for pilot studies
  - Pilot for Distributed Cloud with USD, Georgia Tech
- Cloud/Services related courses taught
- Some areas that we are working on
  - Cloud based Virtualized services: Automated lifecycle, policies and governance
  - Scientific workflows on Hadoop
  - Privacy on Cloud