Cloud Computing for Science

August 2009

CoreGrid 2009 Workshop

Kate Keahey

keahey@mcs.anl.gov

Nimbus project lead

University of Chicago

Argonne National Laboratory

Cloud Computing is in the news…

iSGTW INTERNATIONAL SCIENCE GRID THIS WEEK

About iSGTW | Contact iSGTW | Subscribe | Archive | Resources

Home > iSGTW - 20 May 2009 > Feature - A side of cloud with your grid, ma'am?

Feature - A side of cloud with your grid, ma'am?

Newsweek

TECHNOLOGY

Living in the Clouds

Is computer software becoming obsolete?

The Obama Team's Cloudy Ambitions

May 13th, 2009 : Rich Miller

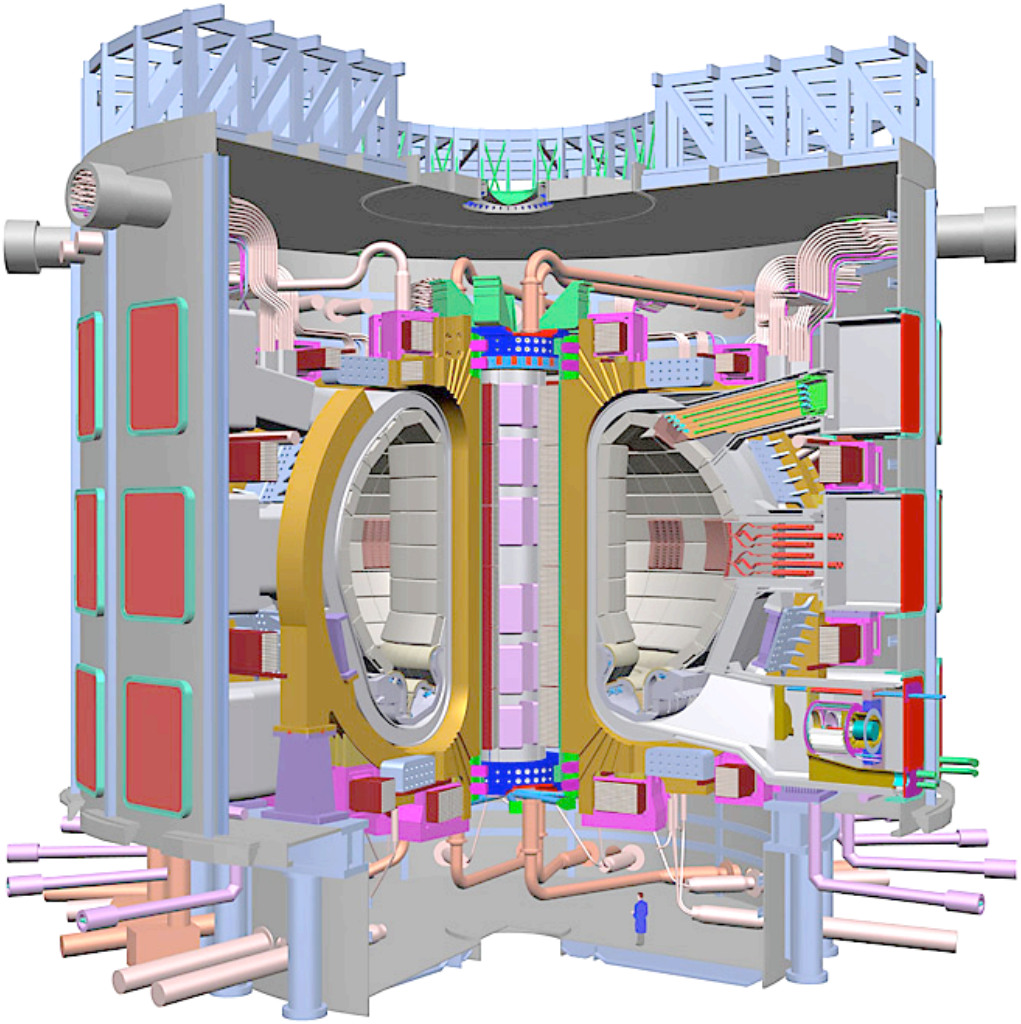
A new White House document reinforces the Obama administration's intent to shift a substanti

NEBULA: NASA's Cloud Computing Platform

Finally, a way to manage research-class computing capacity, with the ease and efficiency of the Enterprise Cloud.

…is it good news for Science?

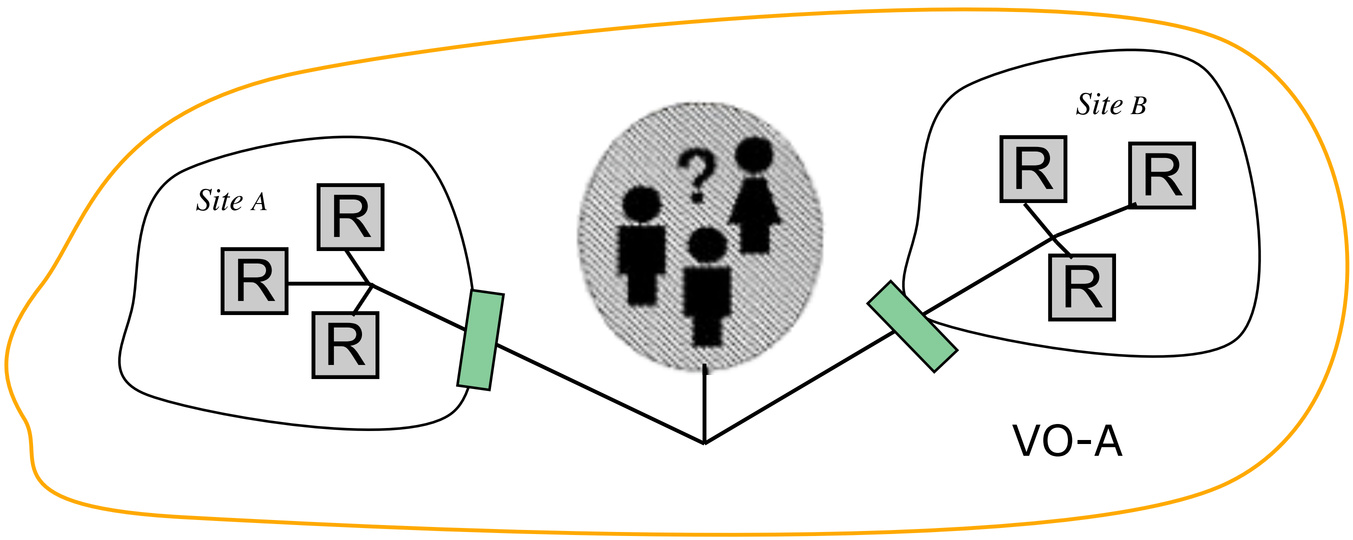
Cloud Computing for Science



- Complex codes
- Need for control

Grid Computing

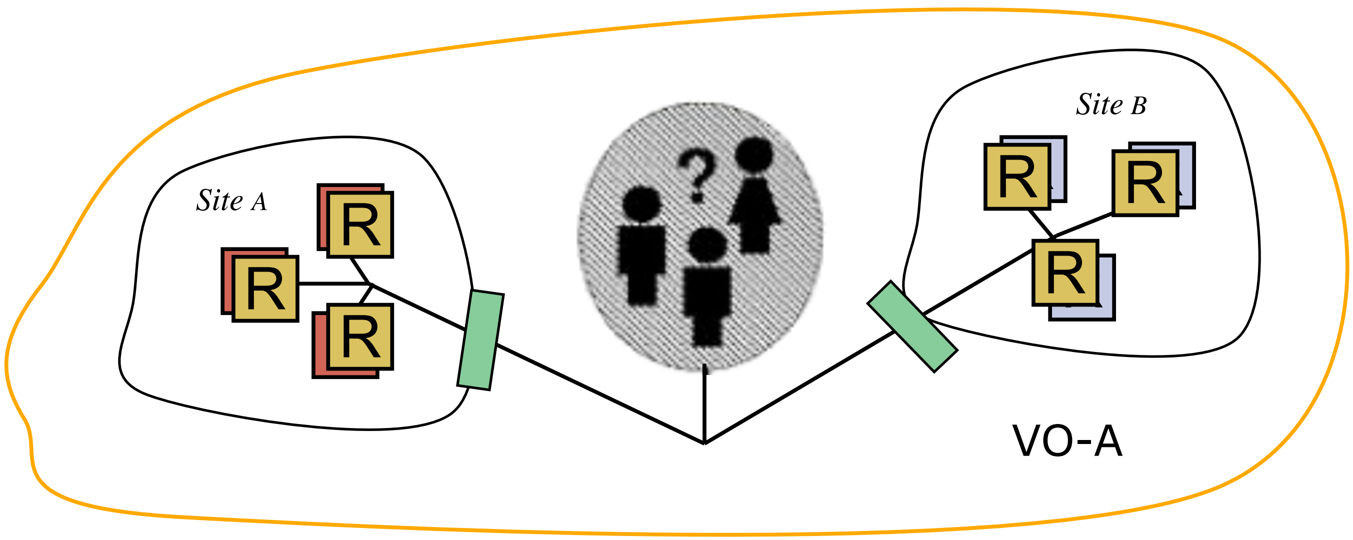
Assumption: control over the manner in which resources are used stays with the site



- Site-specific environment and mode of access
- Site-driven prioritization
- But: site control -> rapid adoption

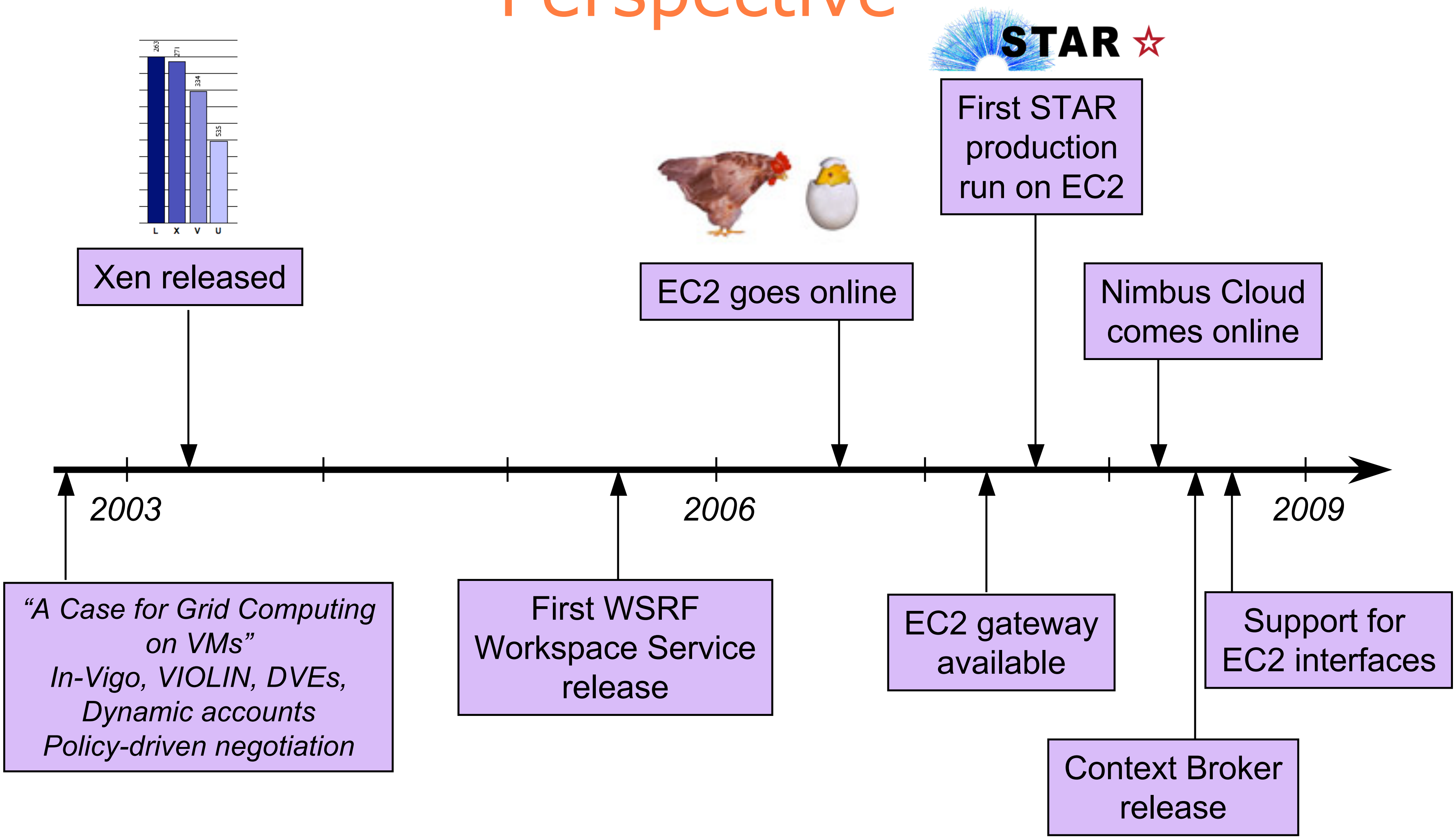
Cloud Computing

Change of assumption: control over the resource is turned over to the user



- Enabling factors: virtualization and isolation
- Challenges our notion of a site
- Lends itself to more explicit service level negotiation
- But: slow adoption

Grids to Clouds: a Personal Perspective



Benefits to Consumers

Eliminate expense and headaches of acquiring, managing and operating hardware

Elastic computing
Pay-as-you-go model

capital expense → **operational expense**

Benefits to Providers

Economies of scale to amortize the costs of buying and operating resources

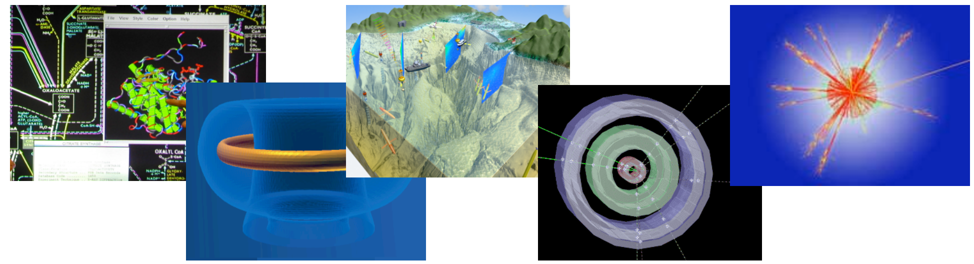
Avoid cost and complexity of managing multiple customer-specific environments and applications

Streamline and specialize

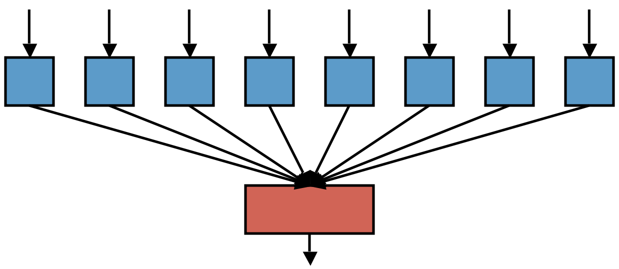
Unclouding the Cloud

Software-as-a-Service (SaaS)

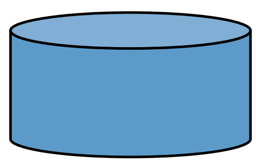
Community-specific applications and portals



Platform-as-a-Service (PaaS)



Infrastructure-as-a-Service (IaaS)

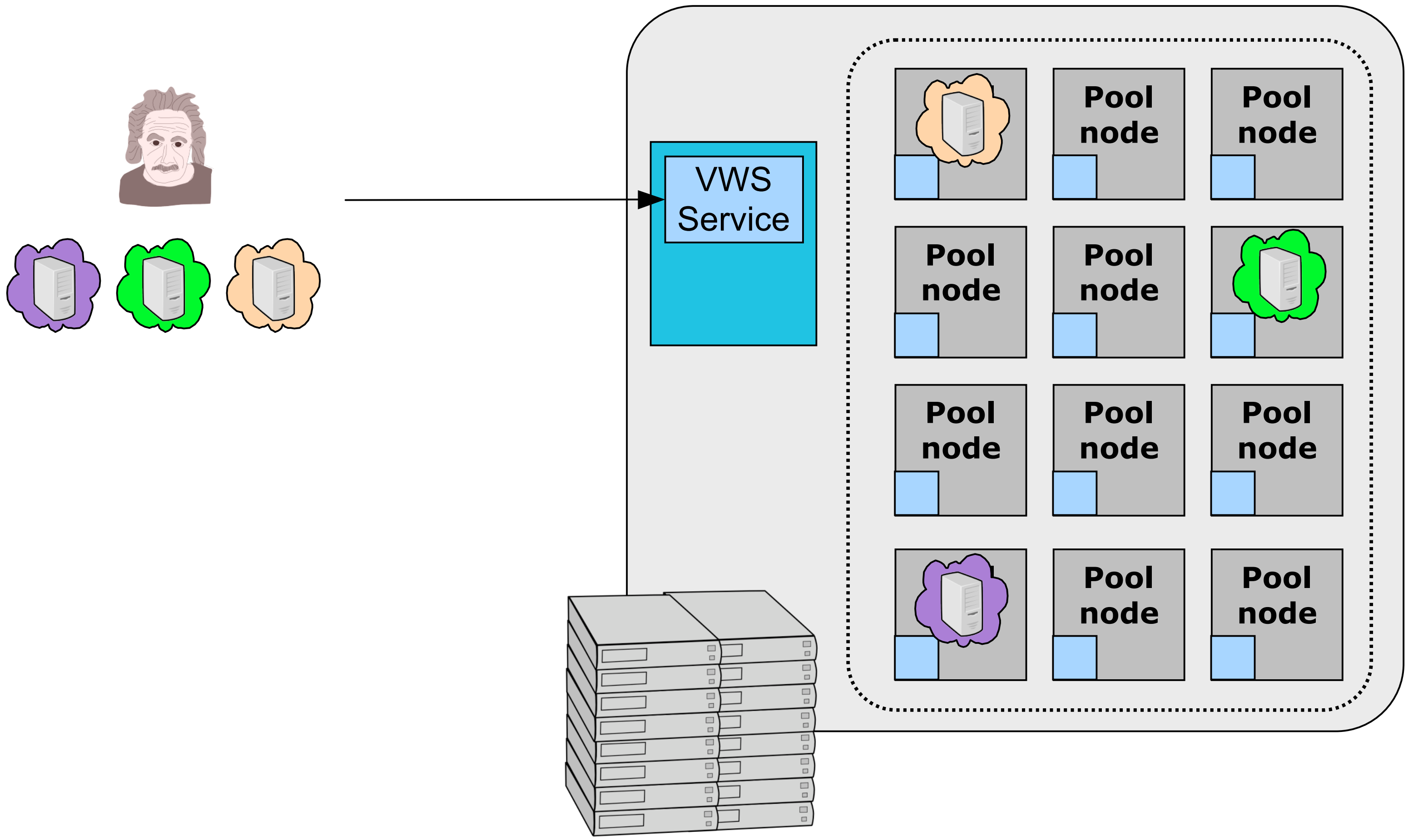


The Nimbus Toolkit: an Example Infrastructure-as-a-Service Implementation

Nimbus: Cloud Computing Software

- Allow providers to build clouds
  - Private&shared (privacy, expense considerations)
  - Workspace Service: open source EC2 implementation
- Allow users to use cloud computing
  - Do whatever it takes to enable scientists to use IaaS
  - Context Broker: turnkey virtual clusters,
  - Also: protocol adapters, account managers, scaling tools…
- Allow developers to experiment with Nimbus
  - For research or usability/performance improvements
  - Community extensions and contributions: UVIC (monitoring), IU (EBS), Technical University of Vienna (privacy, research)
- Nimbus: http://workspace.globus.org

The Workspace Service

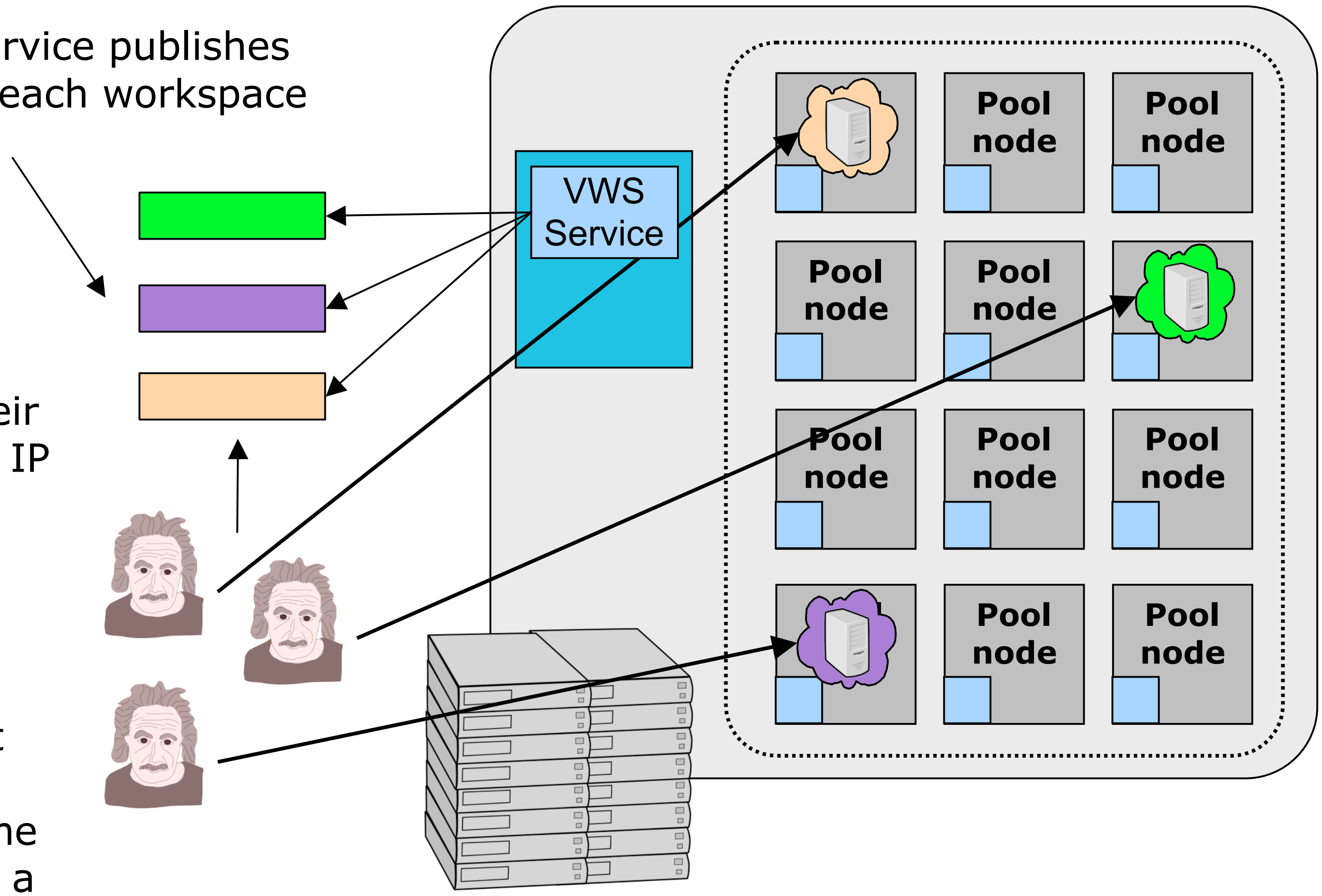


The Workspace Service

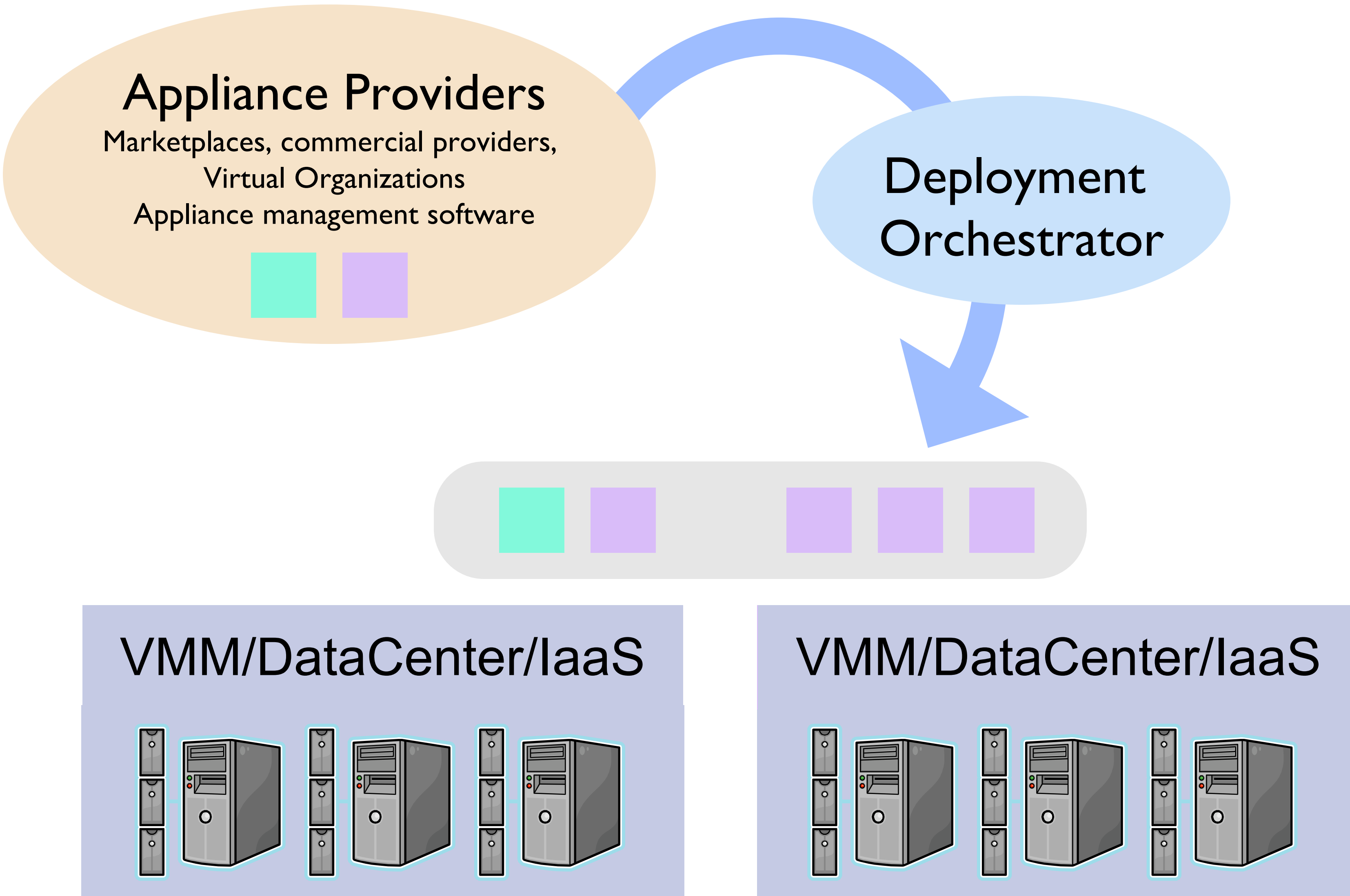
The workspace service publishes information about each workspace

Users can find out information about their workspace (e.g. what IP the workspace was bound to)

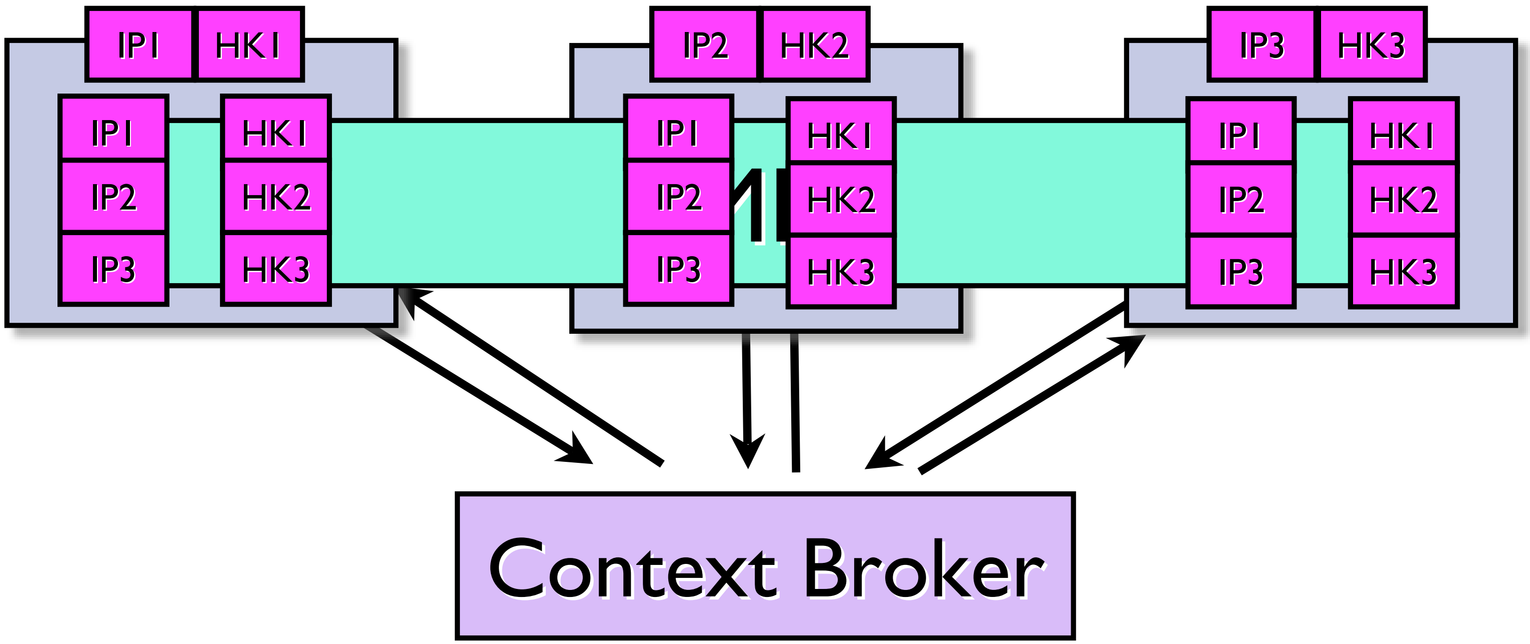
Users can interact directly with their workspaces the same way the would with a physical machine.



Cloud Computing Ecosystem



Turnkey Virtual Clusters



- Turnkey, tightly-coupled cluster
  - Shared trust/security context
  - Shared configuration/context information

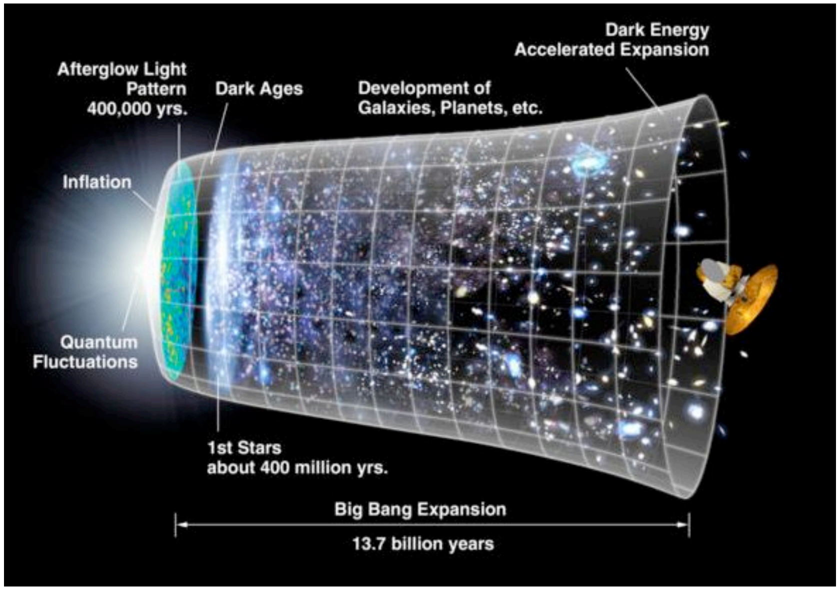
Scientific Cloud Resources and Applications

Science Clouds

- Goals
  - Enable experimentation with IaaS
  - Evolve software in response to user needs
  - Exploration of cloud interoperability issues
- Participants
  - University of Chicago (since 03/08), University of Florida (05/08, access via VPN), Masaryk University, Brno, Czech Republic (08/08), Wispy @ Purdue (09/08)
  - Using EC2 for large runs
- Science Clouds Marketplace: OSG cluster, Hadoop, etc.
- 100s of users, many diverse projects ranging across science, CS research, build&test, education, etc.
- Come and run: http://workspace.globus.org/clouds

STAR experiment

- STAR: a nuclear physics experiment at Brookhaven National Laboratory
- Studies fundamental properties of nuclear matter
- Problem: computations require complex and consistently configured environments that are hard to find in existing grids



STAR Virtual Clusters

Work by Jerome Lauret, Leve Hajdu, Lidia Didenko (BNL), Doug Olson (LBNL)

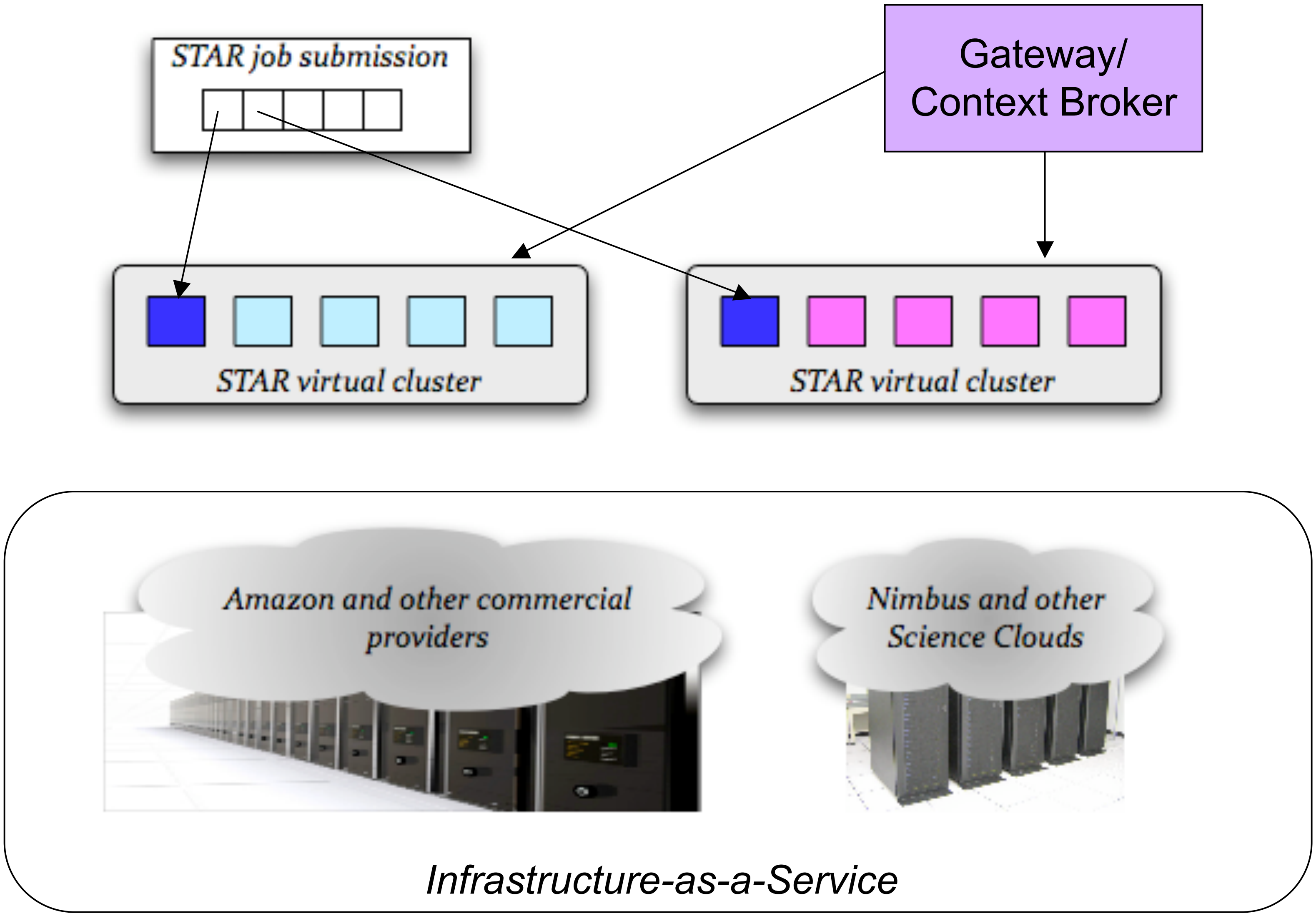
- Virtual resources
  - A virtual OSG STAR cluster: OSG headnode (gridmapfiles, host certificates, NFS, Torque), worker nodes: SL4 + STAR
  - One-click virtual cluster deployment via Nimbus Context Broker
- From Science Clouds to EC2 runs
- Running production codes since 2007
- The Quark Matter run: producing just-in-time results for a conference: http://www.isgtw.org/?pid=1001735

Newsweek

TECHTONIC SHIFTS

Number Crunching Made Easy

STAR Quark Matter Run



Priceless?

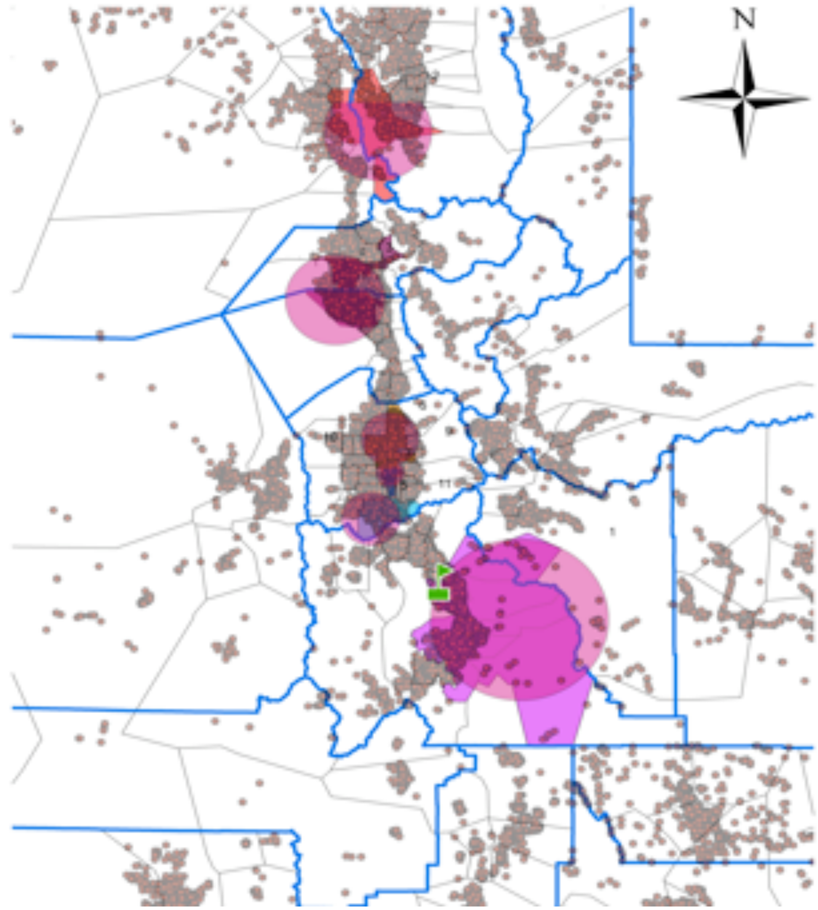
- Compute costs: $ 5,630.30
  - 300+ nodes over ~10 days,
  - Instances, 32-bit, 1.7 GB memory:
    - EC2 default: 1 EC2 CPU unit
    - High-CPU Medium Instances: 5 EC2 CPU units (2 cores)
  - ~36,000 compute hours total
- Data transfer costs: $ 136.38
  - Small I/O needs : moved <1TB of data over duration
- Storage costs: $ 4.69
  - Images only, all data transferred at run-time
- Producing the result before the deadline…

…$ 5,771.37

Modeling the Progression of Epidemics

Work by Ron Price and others, Public Health Informatics, University of Utah

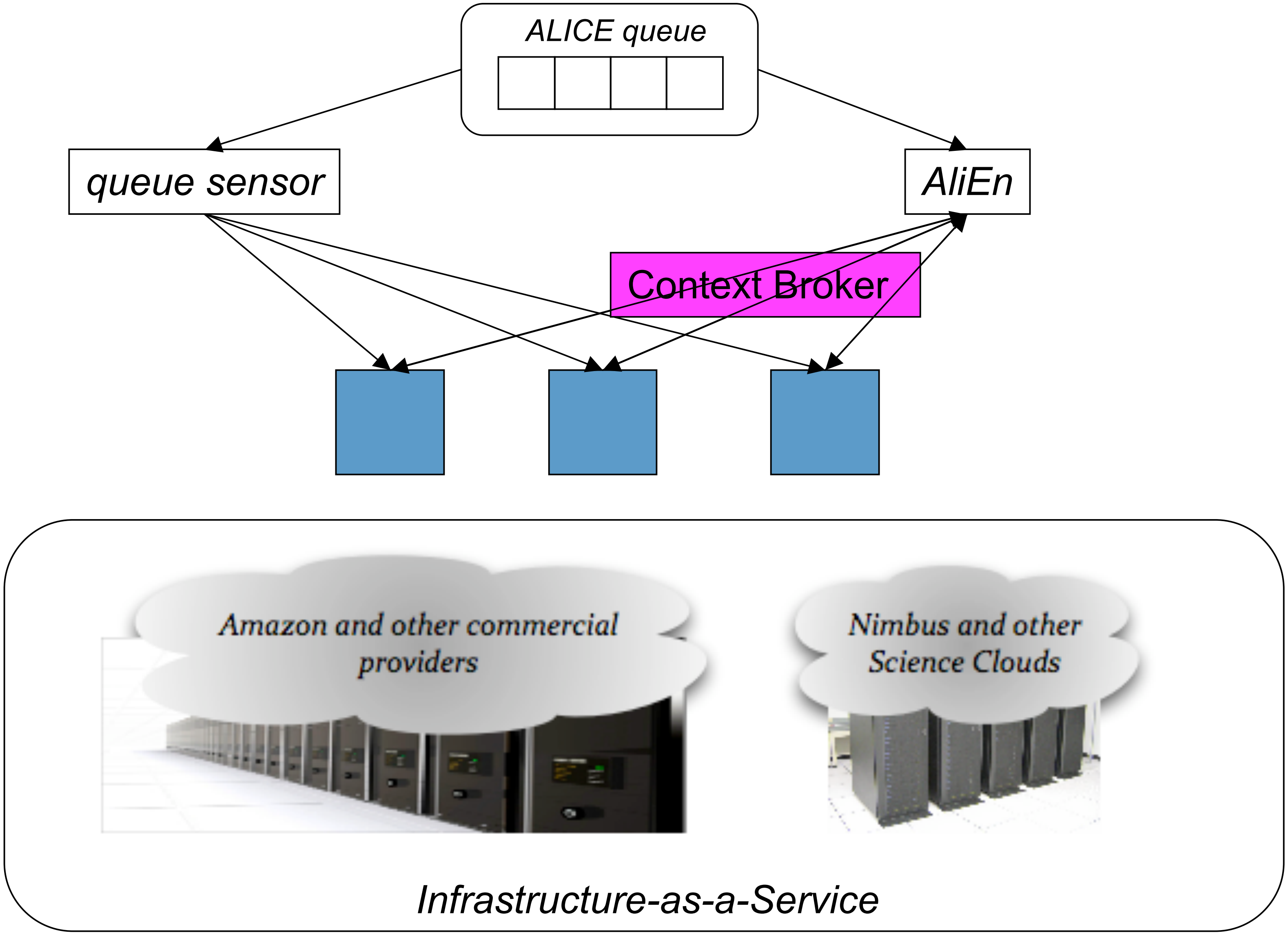
- Can we use clouds to acquire on-demand resources for modeling the progression of epidemics?
  - Monte-Carlo simulations
- What is the efficiency of simulations in the cloud?
  - Compare execution on:
    - a physical machine
    - 10 VMs on the cloud
    - The Nimbus cloud only
  - 2.5 hrs versus 17 minutes
  - Speedup = 8.81
  - 9 times faster



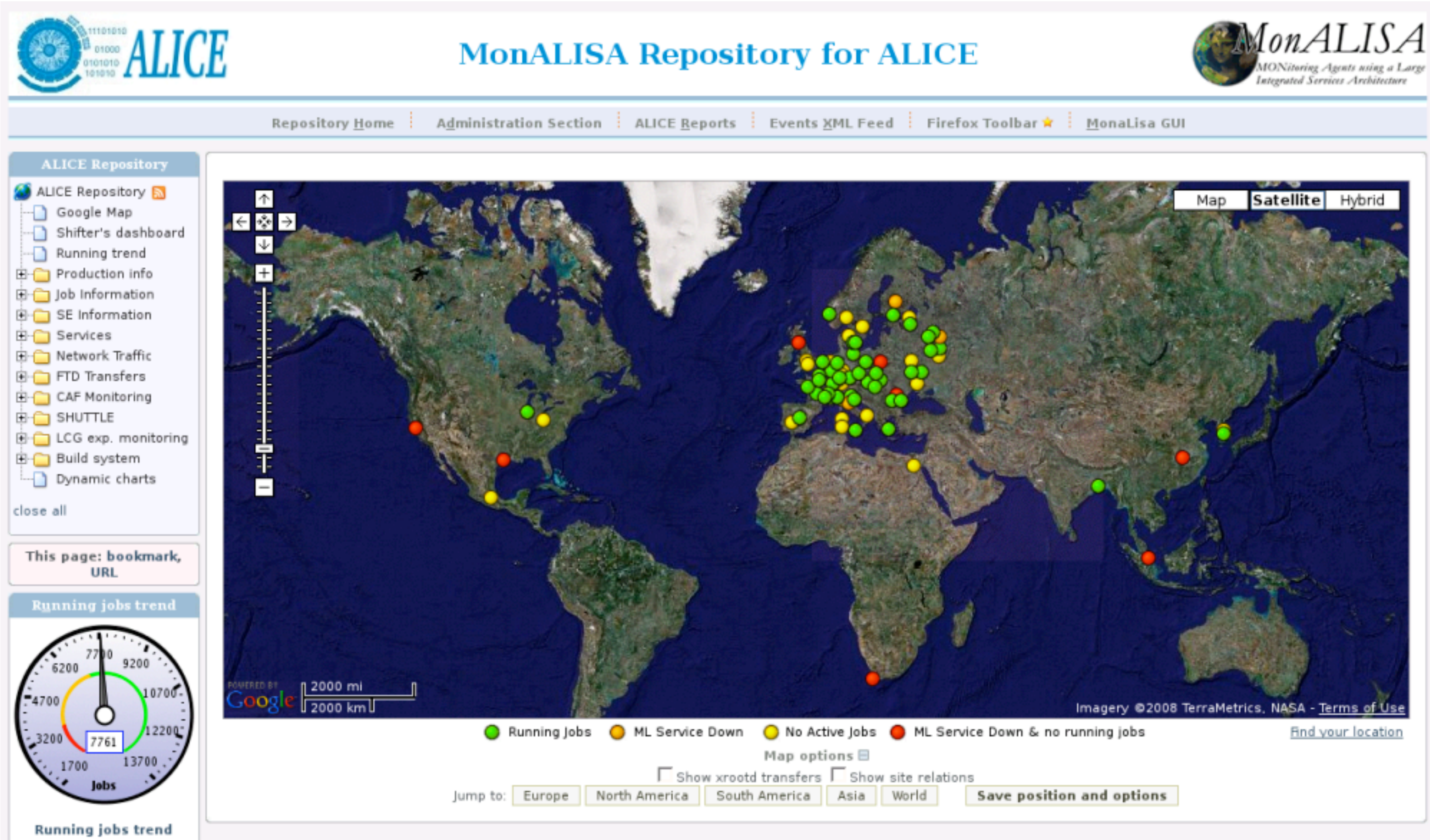
A Large Ion Collider Experiment (ALICE)

- Heavy ion simulations at CERN
- Problem: integrate elastic computing into current infrastructure
- Collaboration with CernVM project
- With Artem Harutyunyan and Predrag Buncic

Elastic Provisioning for ALICE HEP



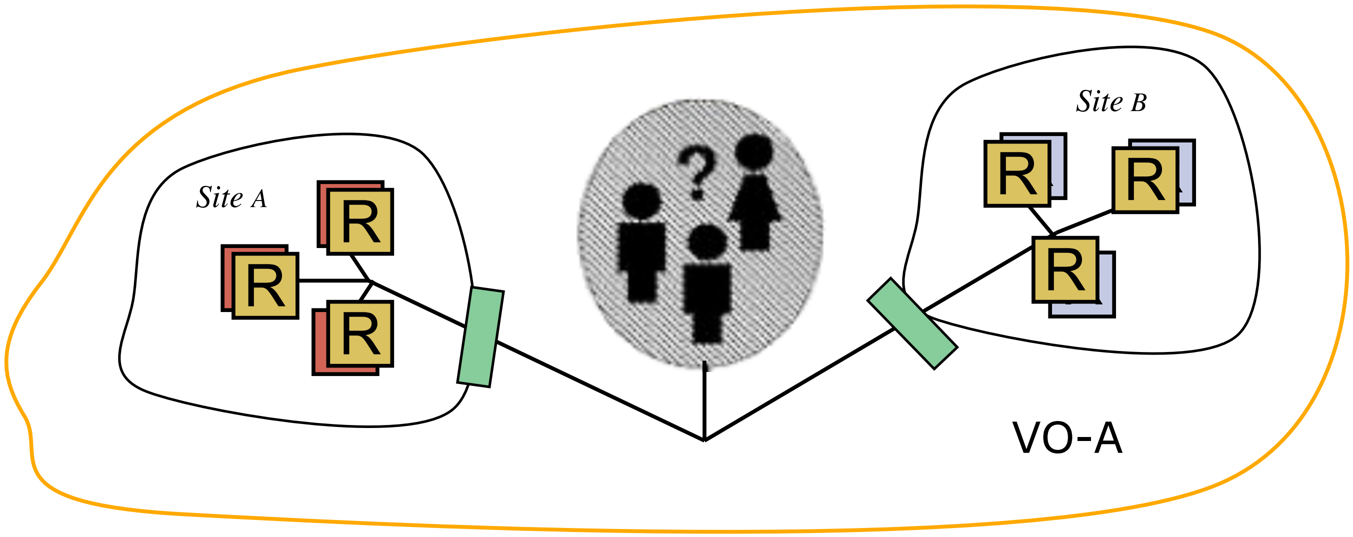
Elastically Provisioned Resources



- *CHEP09 paper, Harutyunyan et al.*
- *Elastic resource base: ElasticSite, ATLAS, and others*

Sky Computing

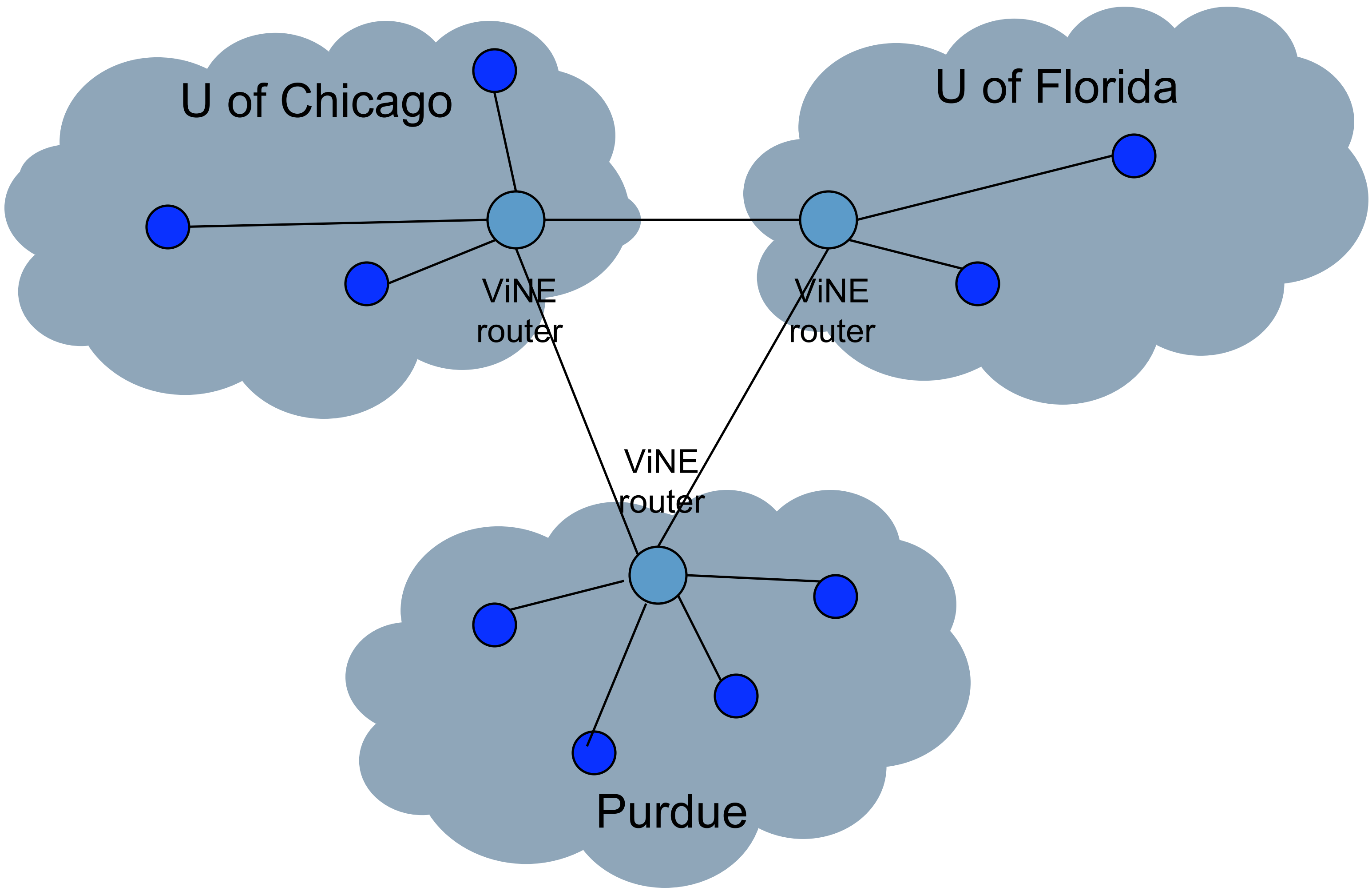
Change of assumption: we can now trust remote resources



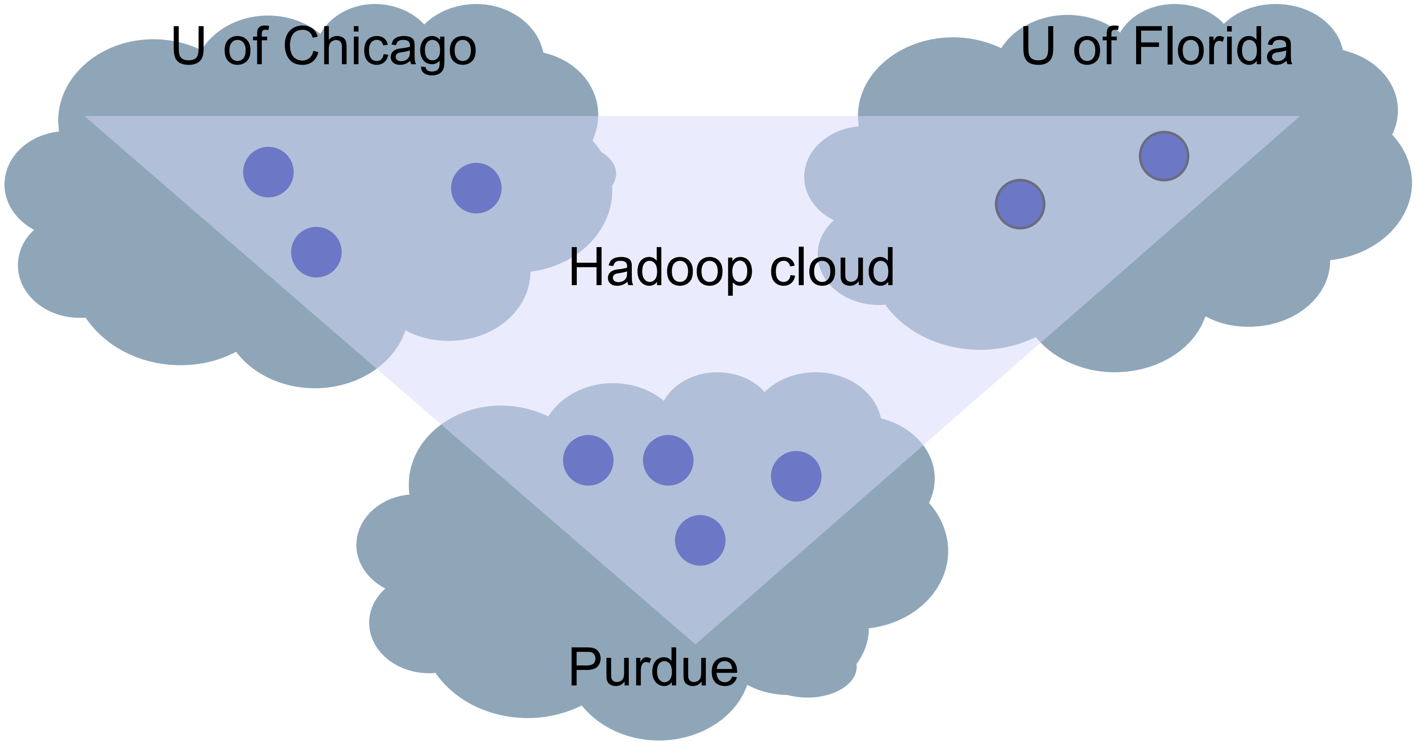
- Enabling factors: cloud computing and virtual networks
- Instead of a bunch of disconnected domains, one domain overlapping the Internet
- Network leases for a fully controlled environment

Sky Computing Environment

Work by A. Matsunaga, M. Tsugawa, University of Florida



Hadoop in the Science Clouds



- *Papers:*
  - *"CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications" by A. Matsunaga, M. Tsugawa and J. Fortes. eScience 2008.*
  - *"Sky Computing", by K. Keahey, A. Matsunaga, M. Tsugawa, J. Fortes, to appear in IEEE Internet Computing, September 2009*

Cloud Computing for Science: Issues and Challenges

Building the Ecosystem

- Configuring and maintaining appliances
  - Not just VMs, a variety of formats
  - CernVM, rBuilder (rPath)
- Licenses
  - Still vendor-specific approaches
- Getting used to dynamic sites
  - Host certificates and keys, community visibility, failure processing, etc.
- Infrastructure and leveraging

Security and Privacy Issues

- Security: new technology = new attacks
  - VMM issues: VM escape, drivers for smart NICs
  - Cloud infrastructure: IP spoofing?
  - Usage: is your VM up-to-date? are there any secrets on it? are there incentives to protect against attacks? Accepted “security” practices…
  - Attacks happen: e.g., VAServ
- Lack of features
  - Fine-grained authorization
  - *Paper: Palankar et al., Amazon S3 for Science Grids: a Viable Solution?*
- Data privacy
  - *Paper: Descher et al., Retaining Data Control in Infrastructure Clouds, ARES (the International Dependability Conference), 2009.*

Performance

- Difficult to track in a virtualized environment
  - I/O can be an issue
  - Tradeoffs between CPU power and throughput
  - Paravirtualized drivers
- Studies of cloud performance
  - *E.g., Walker, Benchmarking Amazon EC2 for high-performance scientific computing*
  - Low bandwidth from existing providers:
    - On the order of: 2-5 MB/sec, 17/21 MB/sec, 30MB/sec
  - Generally speaking, the existing cloud providers do not offer a very high-end computer… yet

Price

- Price for what?
  - Experimenting with business models
  - Estimating the cost is hard
- Price of Base Services for AWS:
  - Computation / EC2
    - On-demand: starting at $0.1 per hour
    - Reserved: starting at $227.50 per year for $0.03 per hour
  - Data / S3
    - Storage: $0.15 per GB/month,
    - Transfer: $0.17 per GB
    - AWS import/export for bulk
- Hosting Scientific datasets for free
  - Free on AWS for frequently used datasets

Service Levels

- Service levels
  - Computation: immediate, advance reservations, best-effort, periodic
  - Data: durability, high/low availability, access performance
  - Cross-cutting concern: security and privacy
- Different price points for different availability

Parting Thoughts

- IaaS cloud computing is science-driven
  - Scientific applications are successfully using the existing infrastructure for production runs
  - Promising new model for the future
- We are just at the very beginning of the "cloud revolution"
  - Significant challenges in building ecosystem, security, usage, price-performance, etc.
- Lots of work to do!