CSE 599c
Scientific Data Management

Magdalena Balazinska and Bill Howe

Spring 2010

Lecture 3 – Science in the Cloud

References

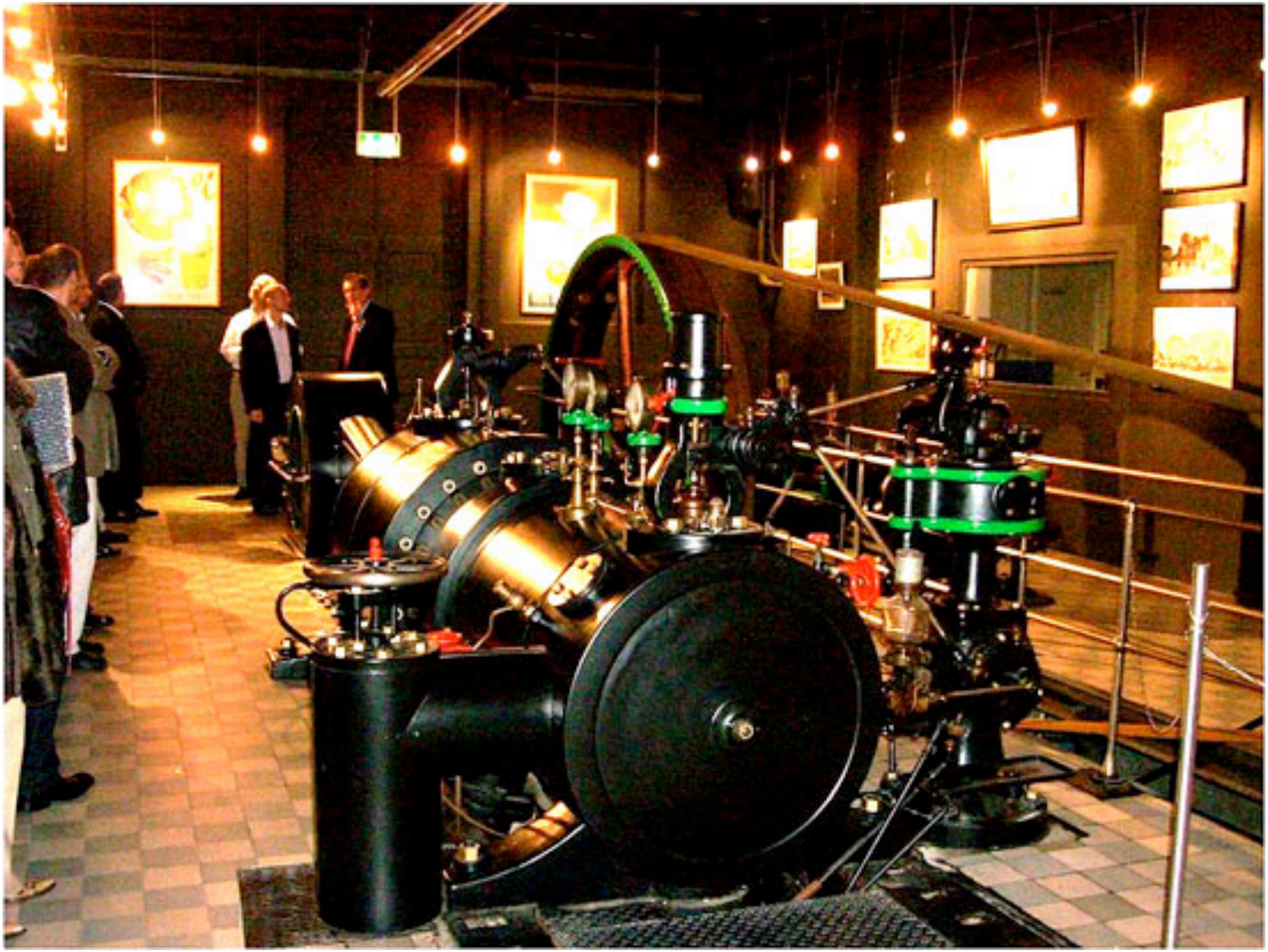
- Existing Clouds
  - Amazon Web services, Google App Engine, & Windows Azure
  - And Science Clouds (listed later in the talk)

- Microsoft Cloud Futures Workshop
  http://research.microsoft.com/en-us/events/cloudfutures2010/default.aspx

- Today's readings
  - E-Science Central: eScience on the Web, powered by Clouds. P. Watson et. al. Under submission
  - Integating marine Observatories into a System-of-Systems: Messaging in the US Ocean Observatories Initiative. M. Arrot et. al. In Proc. of IEEE/ MTS Oceans 2009

Outline

- Review of cloud computing
- Why push science into the cloud?
- Highlights of Cloud Futures workshop
- Two case studies
  - eScience Central
  - Ocean Observatories Initiative



Cloud Computing

- A definition [Wikipedia]
  - "Style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet"

- Basic idea
  - Developer focuses on application logic
  - Infrastructure and data hosted by someone else in their "cloud"
  - Hence all operations tasks handled by cloud service provider

- A few history points
  - "computation may someday be organized as a public utility" (J. McCarthy – 1960)
  - 1990's: Grid computing, Hotmail
  - Early 2000s: Web services, ASPs, Salesforce.com
  - 2005, Google docs
  - 2006, Amazon Web Services
  - And now it's a craze!

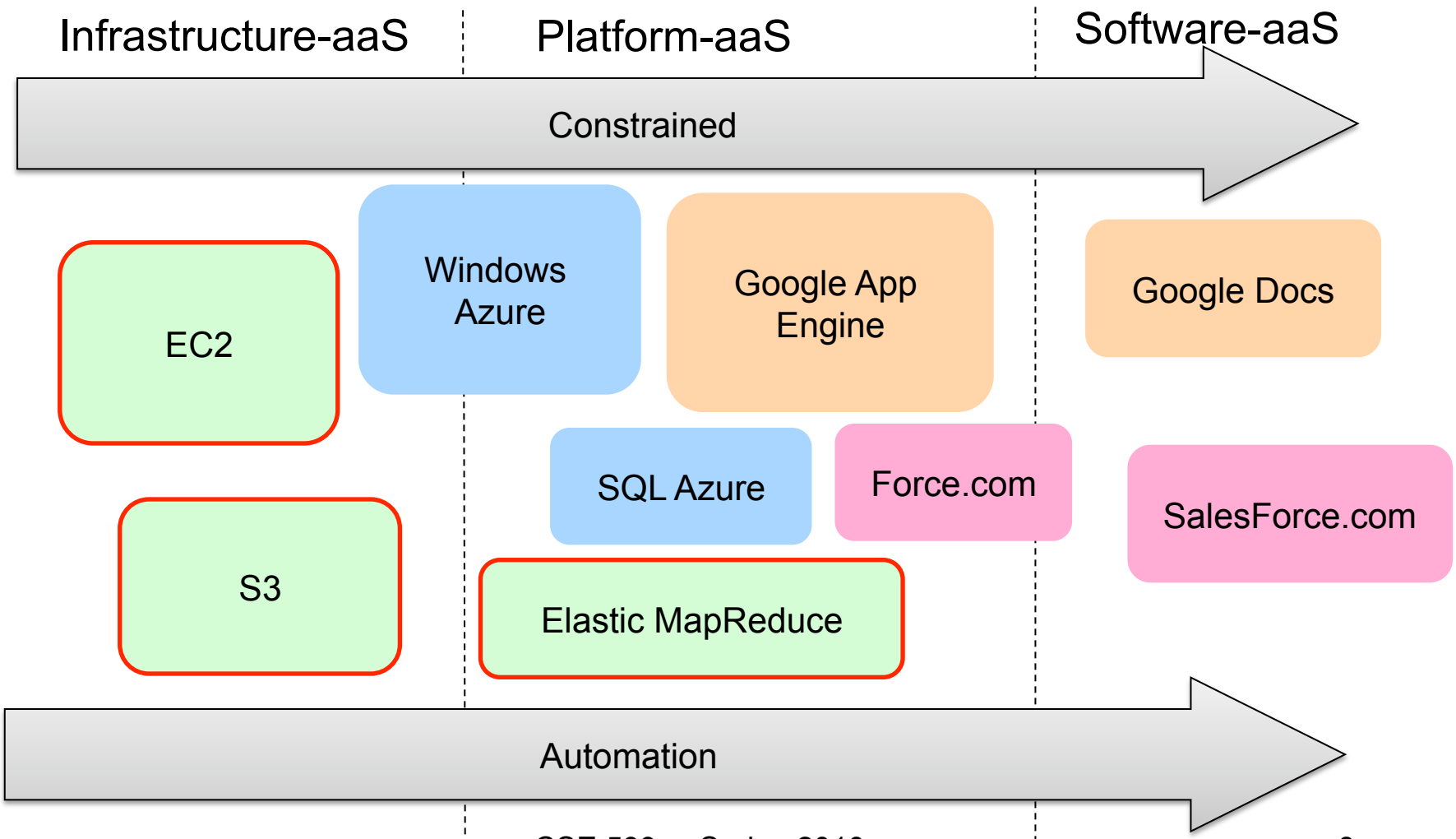
Key Features

- Elastic scalability
- Pay-as-you-go
- Accessible from anywhere
- Accessible as a service
- Multi-tenant (= cheaper but also performance can vary)
- Many admin/operations issues handled by provider…

Levels of Service

- Infrastructure as a Service (IaaS)
  - Example Amazon EC2
- Platform as a Service (PaaS)
  - Example Microsoft Azure, Google App Engine, Force.com
- Software as a Service (SaaS)
  - Example Google Docs

Levels of Service



How About Data Management as a Service?

- Running a DBMS is challenging
  - Need to hire a skilled database administrator (DBA)
  - Need to provision machines (hardware, software, configuration)
  - Problems:
    - If business picks up, may need to scale quickly
    - Workload varies over time
- Solution: Use a DBMS service
  - All machines are hosted in service provider's data centers
  - Data resides in those data centers
  - Pay-per-use policy
  - Elastic scalability
  - No administration!

Basic Features for Data Management as a Service

- Data storage and query capabilities
- Operations and administration tasks handled by provider
  - Include high availability, upgrades, etc.
  - **Elastic scalability**: Clients pay exactly for the resources they consume; consumption can grow/shrink dynamically
    - No capital expenditures and fast provisioning
- Three different types exist at the moment
  - Simplified data management systems (e.g., Amazon SimpleDB)
  - Standard relational data management systems
  - Analysis services such as Amazon Elastic MapReduce

Why push science into the cloud?

- Scientists have great data management needs
  - Collect, store, archive, and analyze growing datasets
- Processing power needs are bursty (e.g., paper deadline)
- Scientists need to share their data and analyses
  - Simply co-locating data is already a huge help here!
- Scientists do not have computer science expertise
  - They should not have to be DBAs!

Why is this Hard?

- Cloud platforms can still be hard to use
- Clouds do not necessarily provide services needed by scientists

Lots of Ongoing Efforts

- **DataOne** (earth science observatory data, nothing deployed yet)
- **HubZero** (platform for science collaboration, helps create dynamic websites)
- **CrowdLabs** (social visualization repository, based on VisTrails)
- **myGrid** (creating, running, and sharing Taverna workflows)
- **Nimbus** (turnkey clusters for science)
- **CAMERA** (ocean microbiology)
- **caBig** (biology)
- Galaxy, BioMart, BioMobi, InnateImmunityPortal, LANL HIV, Pathway Commons, GEO, ArrayExpress (bio, varying success, usually focused on data quality and simple data retrieval)

Cloud Futures Workshop

- Keynotes
  - Ed Lazowska
  - David Patterson
  - Dan Reed
- Technical talks
  - Several example uses of clouds, including for science
  - Several technical talks about building cloud infrastructures
    - Virtualization, failures, consistency, replication, elasticity, etc.
  - One track on existing clouds: Amazon, Google, Microsoft
- All talks were filmed, so they should be available online

Two Case Studies

- E-Science Central
- Ocean Observatories Initiatives

Discussion Questions

- What are the key goals/requirements of the system?
- What are the key features of the system?
  - Any good ideas?
  - Any terrible ideas?
  - Anything missing?
- Is it likely to succeed? Why or why not?