Scaling Science to the Cloud

Tony Hey
Corporate Vice President
Microsoft Corporation
Huge opportunities for insight and innovation through ‘scaling’ our research capabilities
Scaling science

- Length scales: from infinitesimal to galactic
- Research teams: from individual to community
- Timescales: from instant to eon
- Complexity: from single source to webscale
- Data: from documents to digital libraries
from infinitesimal to galactic
Detailed neural circuitry of the brain - the infinitesimal

3D registration, segmentation, and complete neural circuit reconstruction

- Visualize and “fly through” 3D images of neural circuitry at least 100GBs

- Database design and management for storing and querying multi-TB datasets containing 3D microscopy images of neural circuits

**Harvard University:** Jeff Lichtman, Kenny Blum and Hanspeter Pfister

**MSR:** Michael Cohen
Seamless Rich Social Media Virtual Sky
Web application for science and education

• Science- Seamless integration of multi-wavelength, multiple telescope distributed image/data sets and one click contextual access to distributed web information/data sources
• Education- Easy as Powerpoint, rich social media authoring environment within the sky allowing astronomers, educators and kids to create and share rich narrated guided tours of the universe

www.worldwidetelescope.org

ID magazine International Design Annual
“Best in category; Interactive 2009”

TIME magazine “50 Best sites on the Internet 2009”

Harvard-Smithsonian: Alyssa Goodman
Johns Hopkins University: Alex Szalay
Microsoft Research: Curtis Wong, Jonathan Fay
from individual to community
Chasing HIV – to web scale analysis

Tracking the evolution of HIV inside an individual using advanced machine-learning algorithms
http://www.galaxyzoo.org
Hanny van Arkle’s Woorwerp
"Green Peas"
from single source to webscale
Carbo-Climate Synthesis

Understanding the global carbon cycle

- Measurements of CO2 in the atmosphere show 16-20% less than emissions estimates predict...the difference is either due to plants or ocean absorption.

- Cross site studies and integration with modeling increasingly important
  - 921 site-years of data
  - 240 sites around the world; 80+ site-years now being added
  - 60+ paper writing teams
  - American data subset is public and served more widely
  - Summary data products greatly simplify initial data discovery

www.fluxnet.org

UC Berkeley: Dennis Baldocchi
Microsoft Research: Catharine van Ingen
Sensor Networks in Brazil
Sensor Networks in Brazil

- Pilot project with hundreds of sensors
  - ‘Hostile environment’
  - Demonstrating good reliability
  - Key datasets for environmental science

- Now looking to scale this to thousands

- Microsoft has experience of dealing with webscale data challenges...
Web data has structure...

...and that counts

(e.g. Body, Title, Anchor)

Search engines report unigram body hits...

Rich context/meta-data ignored

Users form ‘query’
Multi-word Tag Cloud from Government Dataset Titles

Ref: Dr. Li Ding, Rensselaer Polytechnic Institute
from instant to eon
A moment in time – an instant
Our vision is to create an application that allows researchers to browse, overlay, and explore interdisciplinary data sources.
Given the history of the universe

See the demo live at www.chronozoomtimescale.org Walter Alvarez with the support of Bill Crow and the Live Labs Seadragon team.
Our vision is to create an application that allows researchers to browse, overlay, and explore research both inside and outside of their specific expertise. Applications for such a tool are massively interdisciplinary and touch not only earth sciences but physics, genomics, astronomy, economics, history, biology, marketing, and just about any field that produces or consumes data that is somehow related to time.
from documents to digital libraries
Project Tuva
Video hyperlinked to rich Web resources (e.g. Wikipedia)

Oddly, Cassini seems to have abandoned this reasoning, which Ramer adopted and set about buttressing in an irrefutable manner, using a selected number of observations performed by Picard and himself between 1671 and 1677. Ramer presented his results to the French Academy of Sciences, and it was summarised soon after by an anonymous reporter in a short paper, *Démonstration touchant le mouvement de la lumière trouvée par M. Roemer de l’Académie des sciences*, published 7 December 1676 in the *Journal des scéavans*. Unfortunately the paper bears the stamp of the reporter failing to understand Ramer’s presentation, and as the reporter resorted to cryptic phrasings to hide his lack of understanding, he obfuscated...
Full text search with contextual results linked to video
Linked Interactive Simulations and Visualizations
Richard Feynman: Law of Gravitation — An Example of Physical Law

planets and look at the planets to see if they attract each other?

This experiment—the direct test—was made by Cavendish on equipment which you see indicated on the next slide (if I get my slides right).

Well, I made a mistake. I was talking about the importance of gravitation, and I was overwhelmed by my dear remark about astrologers, and forgot to mention the important places where gravitation does have some real effect in the behavior of the universe.

One of the interesting ones is the formation of new stars.

In this picture, which is a gaseous nebula inside our own galaxy (and is not a lot of stars, but is gas), there are places where the gas has been compressed or attracted to itself here. It starts, perhaps, by some kind of shock waves to get collected, but the remainder of the phenomenon is gravitation pulls the cloud of gas closer and closer together.

So big mobs of gas and dust collect and form balls, which, as they fall still further, the heat generated by the falling, lights them up and they become stars—and we have in the next slide some evidence of the creation of new stars.

It is, unfortunately, harder to see than I thought it was when I looked at it before, but this is not exactly the same as this. This is more how it is seen outside.
• Novel ways to visualize and explore data, digital assets, relationships, meaning, sharing, archiving
Facilitating the move from static summaries to rich information vehicles

- Pace of science is picking up...rapidly
- The status quo is being challenged and researchers are demanding more
- Why can’t a research report offer more ...
Imagine...

- Live research reports that had multiple end-user ‘views’ and which could dynamically tailor their presentation to each user
- An authoring environment that absorbs and encapsulates research workflows and outputs from the lab experiments
- A report that can be dropped into an electronic lab workbench in order to reconstitute an entire experiment
- A researcher working with multiple reports on a Surface and having the ability to mash up data and workflows across experiments
- The ability to apply new analyses and visualizations and to perform new *in silico* experiments
Article Authoring Add-in for Word

Structure: Read, convert, and author NLM XML documents

Structure: Client-side XML validation

Services: repository deposit via SWORD

Relationships: ORE Resource Map creation

Relationships: Citation lookup and reference management

http://research.microsoft.com/authoring/
Ontology Plug-in for MS Word

Data Acquisition and Modeling

Collaboration and Visualization

Analysis and Data Mining

Disseminate and Share

Archiving and Preservation

Intent: Term recognition & disambiguation

Relationships: Ontology browser

Services: Ontology download web service

CC Science Commons: John Wilbanks
UCSD: Phil Bourne, Lynn Fink
Microsoft Research: Lee Dirks, Alex Waden
A “Smart” Cyberinfrastructure for Research

Viewpoint
A “Smart” Cyberinfrastructure for Research
A view of semantic computing and its role in research.

The Web has emerged as the largest distributed information repository on the planet. Human knowledge is encoded in various digital forms: Web pages, news articles, blog posts, digitized books, scanned paintings, videos, podcasts, blogs, speech transcriptions, and so forth. Over the years, services have emerged to aggregate, index, and retrieve the vast amount of data available online, but the full meaning of that data may only be perceivable by humans. In the common case, machines are far from understanding or recognizing about the vast amounts of data available online. However, the rise of artificial intelligence—machines capable of performing tasks in a way that was once thought to require human intelligence—has led to the development of new technologies that are fundamentally changing the way we think about and interact with data. These technologies include natural language processing, machine learning, and knowledge representation techniques, which allow machines to understand and reason about the world in ways that were previously impossible.

Data Mesh
At the center of our discussion is the concept of a “data mesh,” a term we use to refer to the modern information systems and knowledge representation techniques that allow us to draw meaningful insights from vast amounts of data. The data mesh is a fundamental building block of the modern information system, and it is defined as a network of interconnected data sources that provide a unified view of the world. The data mesh enables us to integrate information from various sources, including structured and unstructured data, and to extract meaningful insights from the data. This allows us to make sense of the vast amounts of data available online and to make informed decisions based on that information.

Software Infrastructure
Scalable languages, XML, Entity-Data Models, Microformats, RSS, Atom, RDF (see http://www.w3.org/RDF/), OWL (see http://www.w3.org/2002/07/owl/), and other technologies are being used to capture the meaning in data. In this context, machine learning, entity extraction, neural networks, clustering, and other techniques are used to extract information from data. The field is an active area of research and continues to be rapidly evolving. It includes the field of “Semantic Computing” or “Semantic Web.”

Figure 1. Information and Knowledge: Knowledge management at scale is a great opportunity for innovation.
Handling semantic relationships

A semantic computing platform to store and expose relationships between digital assets

Flexible data model enables many scenarios and can be easily extended over time

Zentity

Default web UI with CSS support and custom ASP.Net controls

Native support for RSS, OAI-PMH, OAI-ORE, AtomPub and SWORD

Data Acquisition and Modeling

Collaboration and Visualization

Analysis and Data Mining

Disseminate and Share

Archiving and Preservation

A semantic computing platform to store and expose relationships between digital assets
Richard Feynman: The Messenger Series

Data Acquisition and Modeling
Collaboration and Visualization
Analysis and Data Mining
Disseminate and Share
Archiving and Preservation
Scaling science

- **Length scales:** from *infinitesimal* to *galactic*
- **Research:** from *individual* to *community*
- **Timescales:** from *instant* to *eon*
- **Complexity:** from *single source* to *webscale*
- **Data:** from *documents* to *digital libraries*

Enabled/powered/accelerated by Cloud Computing
The Cloud

• A model of computation and data storage based on “pay as you go” access to “unlimited” remote data center capabilities
• Provides a framework to manage scalable, reliable, on-demand access to applications
• The “invisible” backend to many of our mobile applications
• Historical roots in today’s Internet apps
  - Search, email, social networks
  - File storage (Live Mesh, MobileMe, Flickr, ...)
A Cloud Service: www.smugmug.com

Devoted to priceless photos.

Most internet companies dream of selling to bigger ones and getting rich.
We don’t.

Living a dream.
We dream of an independent company devoted to nothing but your priceless photos.
A company that backs up your photos to three data centers across the U.S.
A profitable, debt-free company.
That earns your fanatical loyalty.
We’re living that dream.

Details, details.
36 employees. More than 300,000 paying customers. 372,720,004 photos and counting.

We’ll always be smaller than the photo-sharing divisions of giant companies.
Which is a very good thing.

Our story.
A Definition:
• Cloud Computing means using a remote data center to manage scalable, reliable, on-demand access to applications
• Providing Applications and Infrastructure over the Internet
• Scalable means possibly millions of simultaneous users of the application
• Reliable means on-demand; 5 “nines” available right now
Range in size from “edge” facilities to mega scale.

Unprecedented economies of scale
Approximate costs for a small size center (1K servers) and a larger, 50K server center.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cost in small-sized Data Center</th>
<th>Cost in Large Data Center</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>$95 per Mbps/month</td>
<td>$13 per Mbps/month</td>
<td>7.1</td>
</tr>
<tr>
<td>Storage</td>
<td>$2.20 per GB/month</td>
<td>$0.40 per GB/month</td>
<td>5.7</td>
</tr>
<tr>
<td>Admin</td>
<td>~140 servers/Admin</td>
<td>&gt;1000 Servers/Admin</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Each data center is **11.5 times** the size of a football field.
Advances in Data Center Deployment

Conquering complexity
• Building racks of servers & complex cooling systems all separately is not efficient.
• Package and deploy into bigger units
• 3 Sockets: Power, Cooling, Bandwidth
Programming the Cloud

**Infrastructure as a Service (IaaS)**
Provide a way to host virtual machines on demand

**Platform as a Service (PaaS)**
You write an Application to Cloud APIs and the platform manages and scales it for you.

**Software as a Service (SaaS)**
Delivery of software to the desktop from the Cloud
PhyloD as an Azure Service

- Statistical tool used to analyze DNA of HIV from large studies of infected patients
- PhyloD was developed by Microsoft Research and has been highly impactful
- Small but important group of researchers
  - 100’s of HIV and HepC researchers actively use it
  - 1000’s of research communities rely on results

Typical job: 10 – 20 CPU hours; Extreme jobs: 1K – 2K CPU hours
- Large number of test runs for a given job (1 – 10M tests)
- Highly compressed data per job ( ~100 KB per job)
Source Imagery Download Sites

Data Collection Stage

Download Queue

Source Metadata

Request Queue

AzureMODIS Service Web Role Portal

Reduction #1 Queue

Derivation Reduction Stage

Reduction #2 Queue

Analysis Reduction Stage

Reprojection Queue

Reprojection Stage

Catharine van Ingen (Microsoft Research), Jie Li, Marty Humphrey (UVA), Youngryel Ryu (UCB), Deb Agarwal (BWC/LBL), Keith Jackson (BL), Jay Borenstein (Stanford), Team SICT: Vlad Andrei, Klaus Ganser, Samir Selman, Nandita Prabhu (Stanford), Team Nimbus: David Li, Sudarshan Rangarajan, Shantanu Kurhekar, Riddhi Mittal (Stanford)
Scaling science

- **Length scales:** from *infinitesimal* to *galactic*
- **Research:** from *individual* to *community*
- **Timescales:** from *instant* to *eon*
- **Complexity:** from *single source* to *webscale*
- **Data:** from *documents* to *digital libraries*

Enabled/powered/accelerated by Cloud Computing
Today...

Computers are great **tools** for storing, computing, managing, and indexing huge amounts of **data**.

For example, Google and Microsoft both have copies of the entire Web for indexing purposes.
Semantic computing combines concepts and technologies that
- Enable data modeling
- Capture relationships
- Allow communities to define ontologies
- Exploit machine learning

Will empower computers to reason about the data

Data → Information → Knowledge

Current technologies → Possibilities for innovation
Tomorrow...

Computers will still be great **tools** for huge amounts of **data**

We would like computers to also help with the **automatic**

acquisition  discovery  aggregation

organization  correlation  analysis

interpretation  inference

storing  computing  managing  indexing
Moving to a world where all data is linked ...

• A knowledge ecosystem:
  • A richer authoring experience
  • An ecosystem of services
  • Semantic storage
  • Open, Collaborative, Interoperable, and Automatic

• Data/information is inter-connected through machine-interpretable information (e.g. paper X is about star Y)

• Social networks are a special case of ‘data meshes’

Attribution: Chris Bizer
Vision of Future Research: scaling cyber-infrastructure with Client + Cloud
PowerPoint Guidelines

- Font, size, and color for text have been formatted for you in the Slide Master
- This template uses Segoe UI, a standard Windows Vista and Office 2007 font
- Hyperlink color: www.microsoft.com
- Use the color palette shown below
Microsoft Research
Faculty Summit 2010
Guarujá, Brasil | May 12 - 14 | In collaboration with FAPESP

Video Title