Microsoft Research Connections and Cloud Computing for Science

Dr. Fabrizio Gagliardi

Microsoft Research

Microsoft Research Connections

Introduction

From 1975 till 2005: Computing Science at CERN (www.cern.ch)

- Developing HPC distributed computing solutions for HEP
- Including EU-DataGrid and EGEE, foundation for the present LHC distributed computing Grid infrastructure (www.eu-egee.org)
- Extending support to other scientific communities in the EU European Research Area context
- Among them **OGF-Europe** and the follow on **SIENA** project active of Grid and Cloud computing standards
- Deploying Cloud Computing for Science and Technology with **VENUS-C** (www.venus-c.eu)

Microsoft Research Connections

Work with the worldwide academic research community to speed research, improve education, and foster innovation

Collaborations to pursue scientific breakthroughs

Inspire emerging computer and research scientists

Accelerate scientific exploration with computing

Microsoft Research

Microsoft

BSC-Microsoft Research Centre

Barcelona Supercomputing Centre: computer architecture, parallel programming models

MSRC expertise: programming language and operating system design & implementation

Research at the intersection of computer architecture, language implementation, and systems software

Transactional memory (TM)

- Abstraction for scalable shared-memory data structures
- Research on using TM in real applications; game servers, recognition-mining-synthesis
- Debugging and profiling
- Major publications include PPoPP 09, MICRO 09, PPoPP

Language runtime system

- Architecture support to accelerate synchronization and garbage collection
- "Dynamic filtering" support for GC read/write barriers (ASPLOS 10)
- H/W abstractions for fast and scalable locking

Low-power vector processors

- New vectorization techniques for cloud computing and mobile applications
- Fusion of Edge and E2 with vector techniques

More on: http://www.bscmsrc.eu/

The Microsoft Research-INRIA Joint Centre

MICROSOFT RESEARCH INRIA JOINT CENTRE

- Founded by INRIA (the French National Research Institute for Computer Science and Applied Mathematics), Microsoft Corporation, and the Microsoft Research Laboratory Cambridge

- The Centre's objective is to pursue fundamental, long-term research in formal methods, software security, and the application of Computer Science research to the Computational Science.

- The Joint Centre benefits from the collaboration of 35 researchers from INRIA and other French academic institutions, 25 post docs and PHD students and 15 researchers from Microsoft Research.
- More on: http://**www.msr-inria.inria.fr**/

The Microsoft Research - University of Trento Centre for Computational and Systems Biology (CoSBi)

- **Goals:** Perform computational system biology using latest HPC technology. Initially a MS HPC cluster was used to carry out simulations of complex biological systems modelled through the techniques developed at CoSBi.

- **Outcome-Impact:** Use of the cluster for time-consuming analyses, especially analyses and simulations that require multiple input data and with different parameters. It was impossible for CoSBi scientists to run this kind of programs before the introduction of the MS HPC cluster in their working environment.

- **Cloud Computing:** moving now to MS Azure cloud computing technology with EU FP7 Venus-C project.

- **More info on http://www.cosbi.eu/**

KU-MSR-BSC

- Focus areas
  - Software engineering, reliability, verification
  - Multicore and multiprocessor systems

- Teams collaborate during the design process:
  - architecture (BSC)
  - systems (MSR)
  - software engineering (KU)

- Software engineering tools for
  - novel multicore architectures
  - novel concurrent programming approaches

- Verification tools early in the design process
  - Not as a late-stage debugging tool only.

Collaborative Research in Computer Vision with MSU

Dr. Pushmeet Kohli,
MSR Cambridge

Dr. Anton Konushin,
MSU

Dr. Carsten Rother,
MSR Cambridge

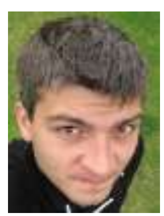
Dr. Olga Barinova.
MSU

Dr. Victor Lempitsky,
Yandex/MSU

Undergraduate and PhD students:

Mikhail Sindeev

Elena Tretiak



Sergey Milyaev

Roman Shapovalov

Tatiana Novikova

2011 Microsoft Computer Vision Summer School in Russia

Facts & figures:

- 520+ registrations
- 70+ cities
- 80 students selected

The school offered students a unique opportunity to learn about fundamental and state of the art on Computer Vision from top scientists , including Andrew Blake, Andrew Fitzgibbon, Carsten Rother (Microsoft Research, UK), Andrew Zisserman (University of Oxford, UK).

PhD Scholarship

- Goals
  - Encourage interdisciplinary research
  - Advance the state of the art
  - Create a community
  - Identify potential interns & employees
- Open & competitive
  - Application by research supervisors
  - Selection ratio 17%
  - Up to one year to find best possible students
- More than funding
  - Co-supervisions by MSR researchers
  - Internship
  - Summer School

MSR Summer Schools

- Networking
  - Other students, MSRC researchers, Cambridge academics
- ‘Transferable skills’
  - Write paper, give talk, becoming an entrepreneur, applying for funding
- Research talks
- Poster sessions
- Social activity

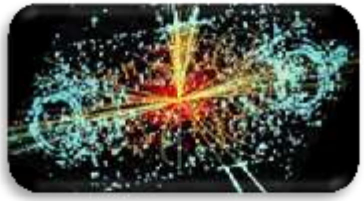




The Future: an Explosion of Data

Experiments

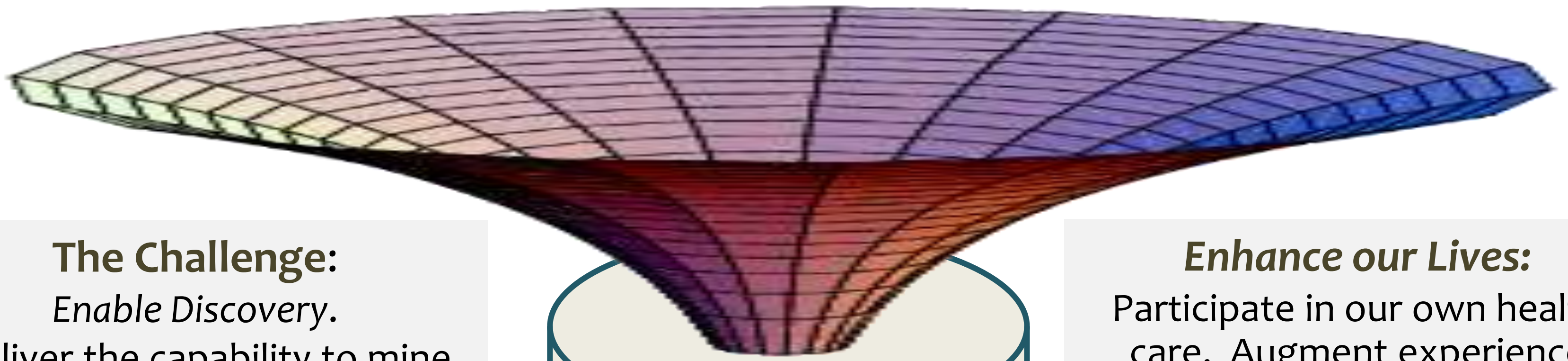
Simulations



Archives

Literature

Instruments



The Challenge:

Enable Discovery.
Deliver the capability to mine, search and analyze this data in near real time.

<u>Petabytes</u>

Digital information created annually will grow by a factor of 44 from 2009 to 2020

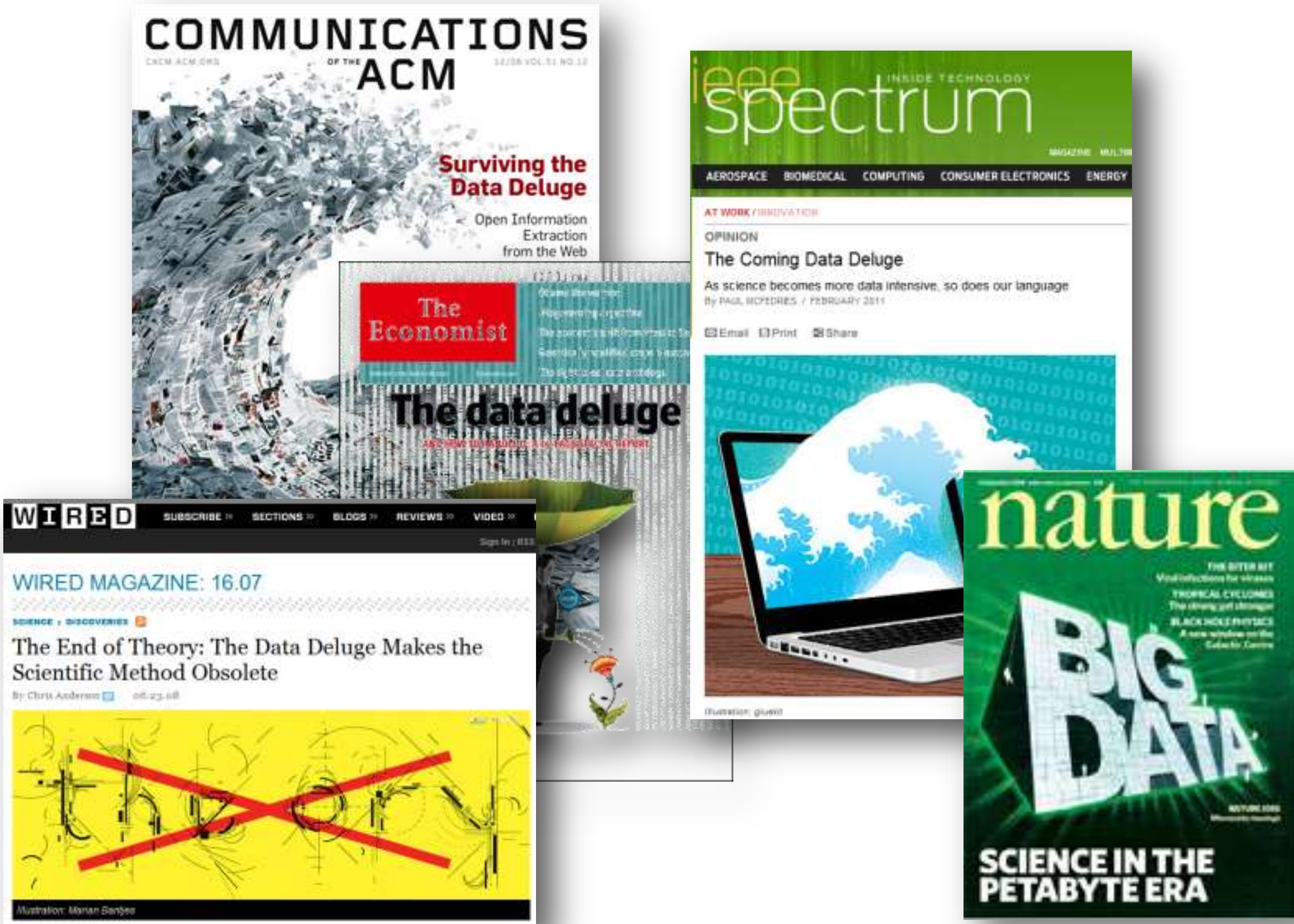
Enhance our Lives:

Participate in our own health care. Augment experience with deeper understanding.

By 2020, more than 1/3rd of all digital information created annually will either live in or pass through the cloud.

(**Source:** EMC-sponsored IDC study)

A Tidal Wave of Scientific Data



Emergence of a Fourth Research Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

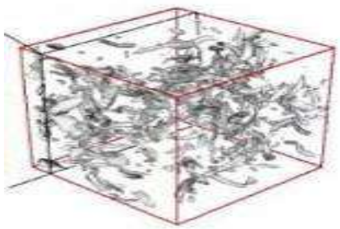
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Captured by instruments
  - Generated by simulations
  - Generated by sensor networks

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

(With thanks to Jim Gray)

Changing Nature of Discovery

Complex models

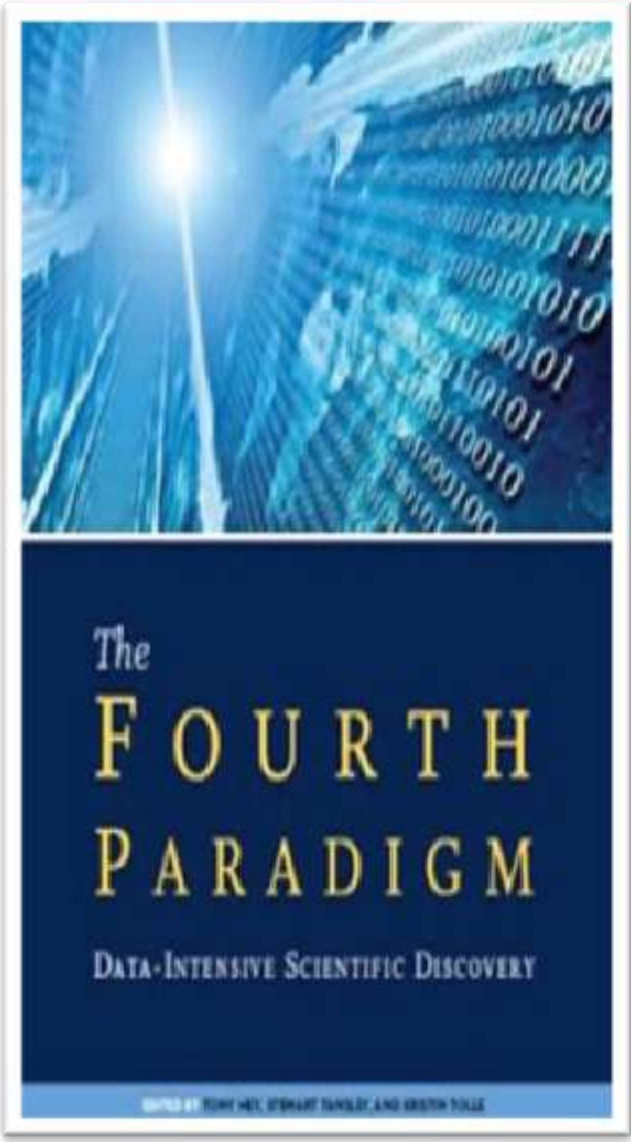
- Multidisciplinary interactions
- Wide temporal and spatial scales

Large multidisciplinary data

- Real-time steams
- Structured and unstructured

Distributed communities

- Virtual organizations
- Socialization and management

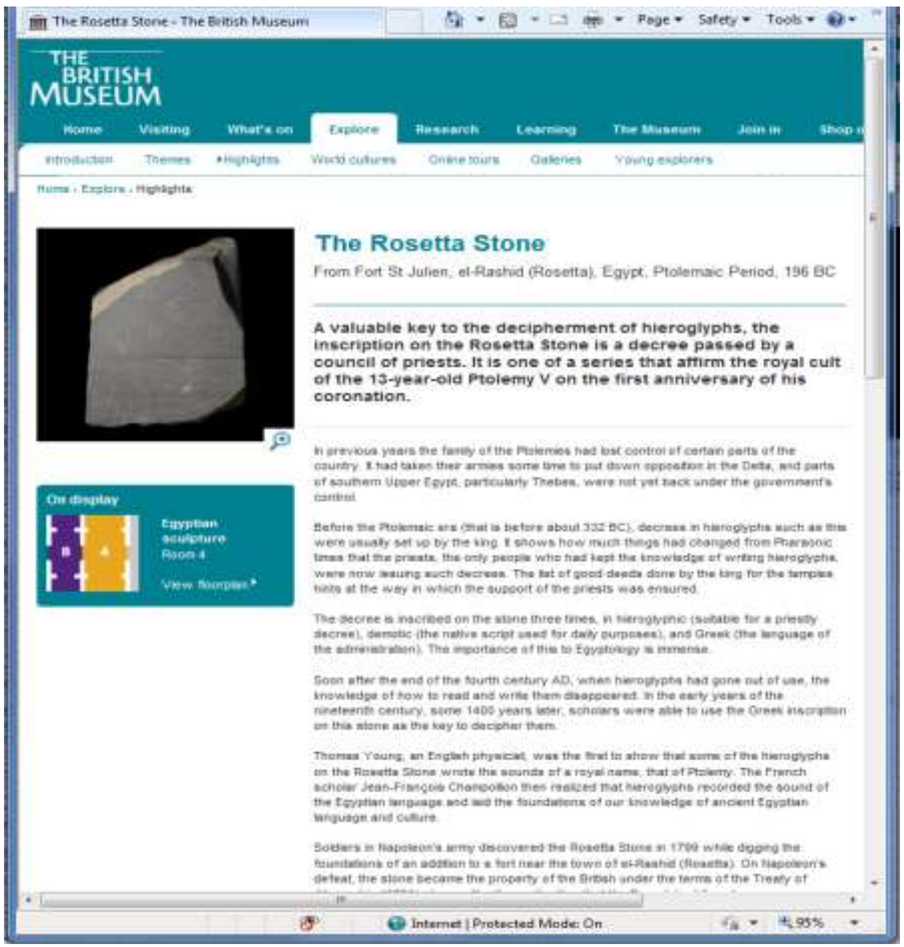


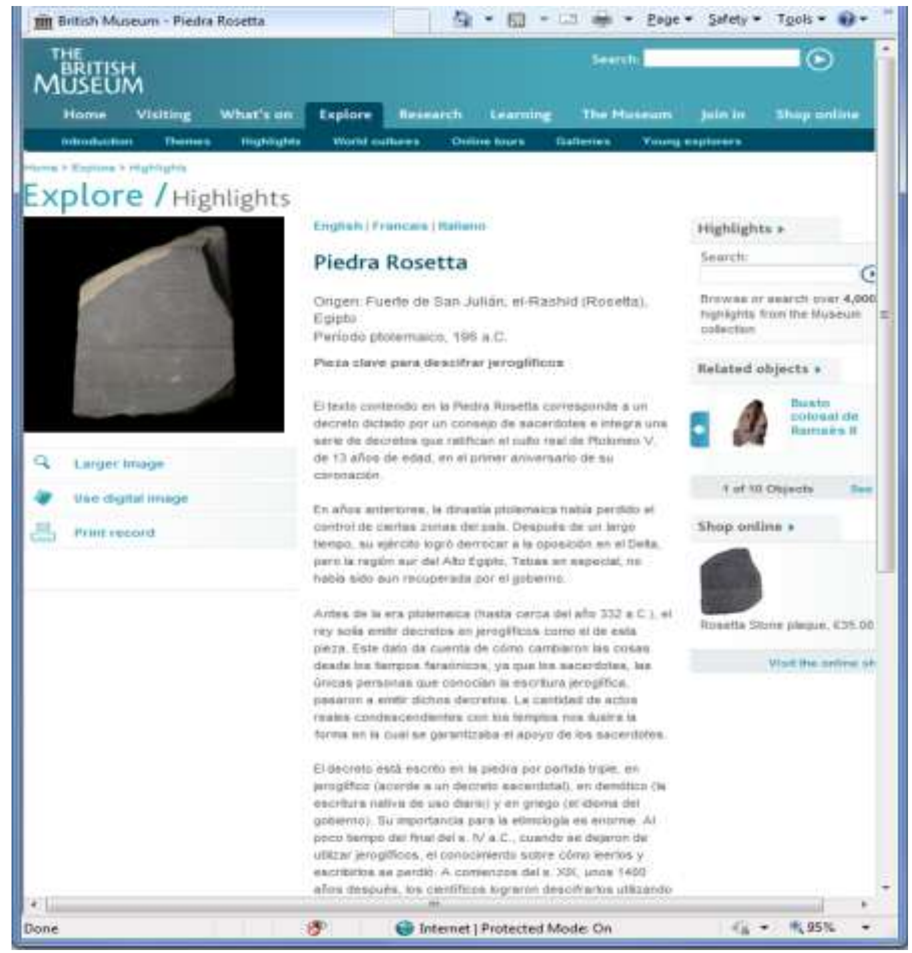
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Machine Translation: The Statistical Revolution

Instead of hand-coding rules

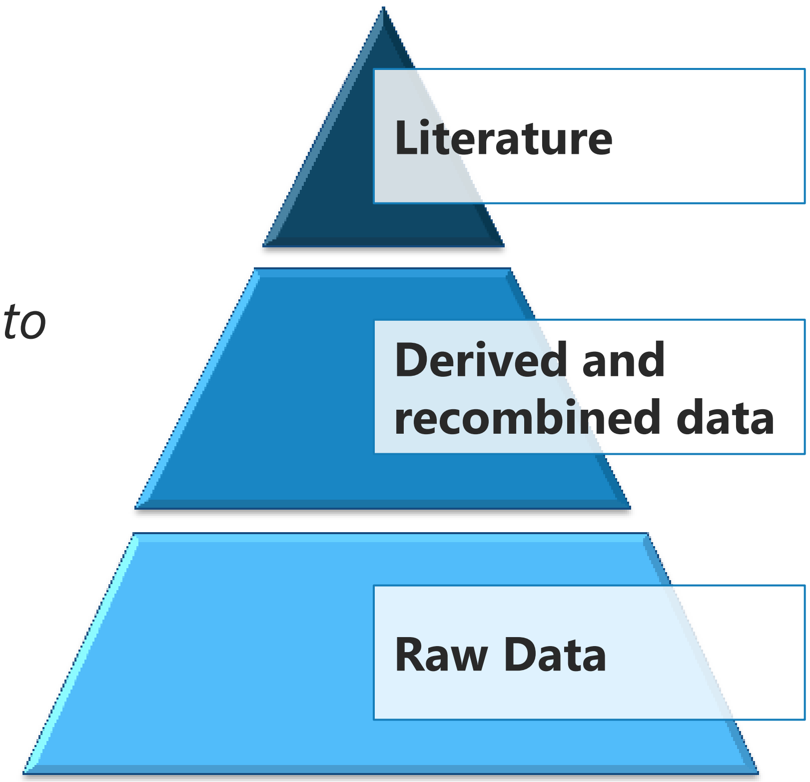
- Exploit large volumes of existing parallel text
- Learn how words, phrases, and structures translate in context





All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips – For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



(From Jim Gray's last talk)

The Cloud

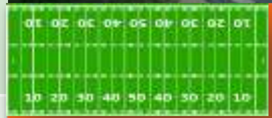
- A model of computation and data storage based on "pay as you go" access to "unlimited" remote data center capabilities
- A cloud infrastructure provides a framework to manage scalable, reliable, on-demand access to applications
- A cloud is the "invisible" backend to many of our mobile applications
- Historical roots in today's Internet apps and previous DCI computing (Cluster, Grid etc.)

The Cloud is built on massive data centers

Essentially driven by economies of scale

- Approximate costs for a small size center (1K servers) and a larger, 100K server center.

Technology	Cost in small-sized Data Center	Cost in Large Data Center	Ratio
Network	$95 per Mbps/ Month	$13 per Mbps/ month	7.1
Storage	$2.20 per GB/ Month	$0.40 per GB/ month	5.7
Administration	~140 servers/ Administrator	>1000 Servers/ Administrator	7.1

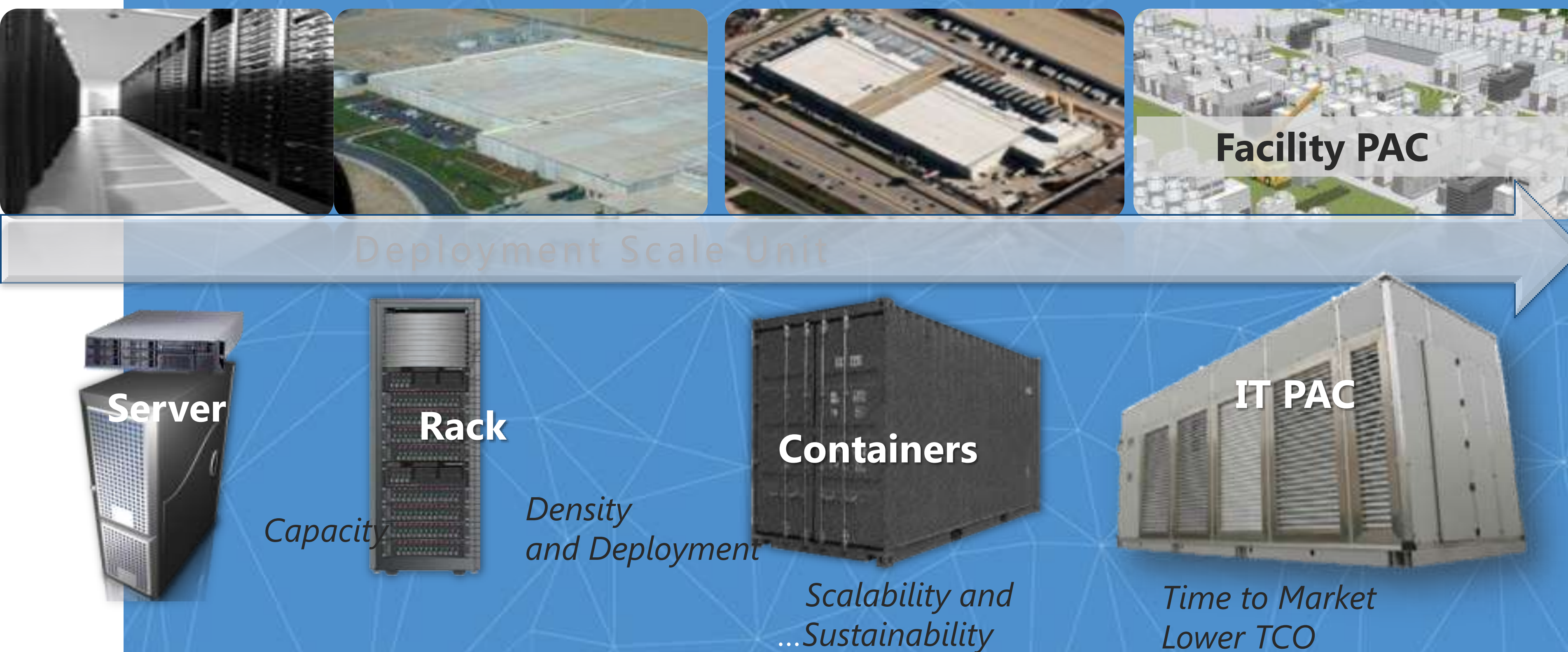


Each data center is
11.5 times
the size of a football field

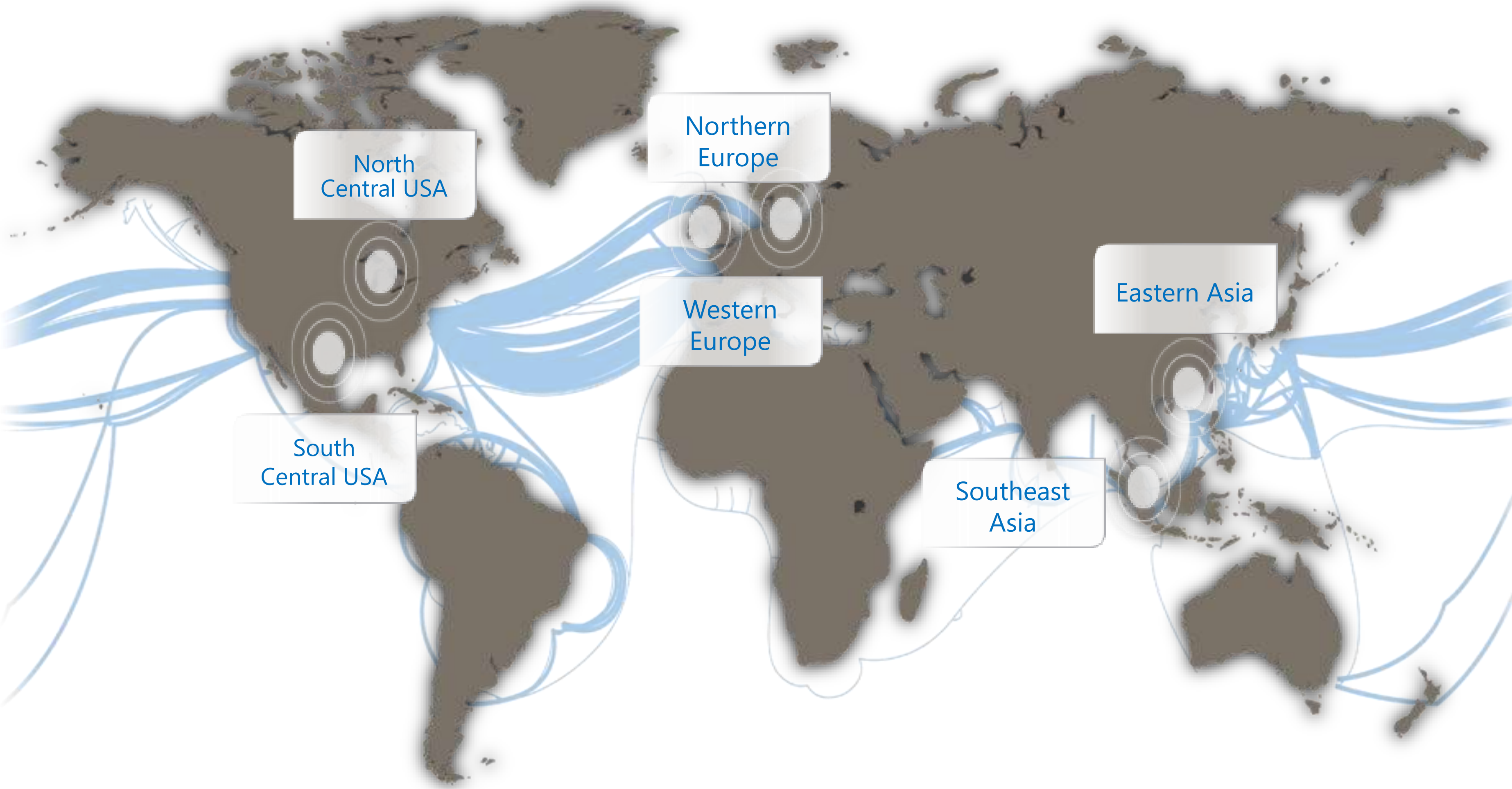
Microsoft's Datacenter Evolution

2005 — 2006 — 2007 — 2008 — 2009 — 2010

Datacenter Co-Location Generation 1	Quincy and San Antonio Generation 2	Chicago and Dublin Generation 3	Modular Datacenter Generation 4



Windows Azure Platform Availability



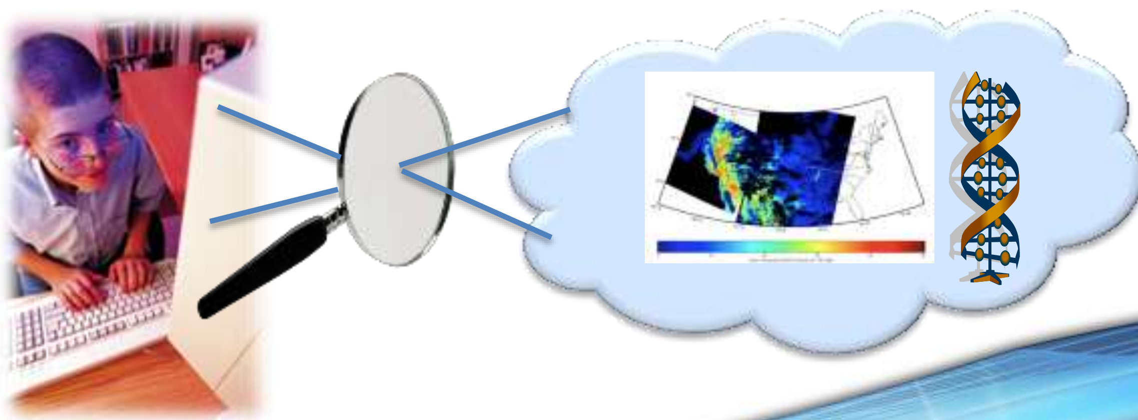
Major Motivations

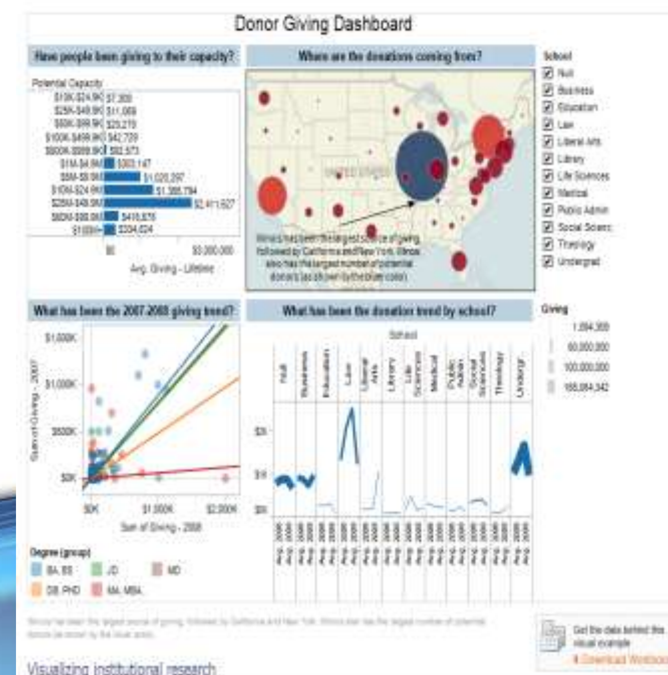
- Environmental responsibility
  - Managing energy efficiently
  - Adaptive systems management
- Provisioning 100,000 servers
  - Hardware: at most one week after delivery
  - Software: at most a few hours
- Resilience during a blackout/disaster
  - Service rollover for millions of customers
- Software and services
  - End-to-end communication
  - Security, reliability, performance, reliability

Focus Client + Cloud for Research

Seamless interaction

- Cloud is the lens that magnifies the power of desktop
- Persist and share data from client in the cloud
- Analyze data initially captured in client tools, such as Excel
  - Analysis as a service (think SQL, Map-Reduce, R/MatLab)
  - Data visualization generated in the cloud, display on client
  - Provenance, collaboration, other ‘core’ services…



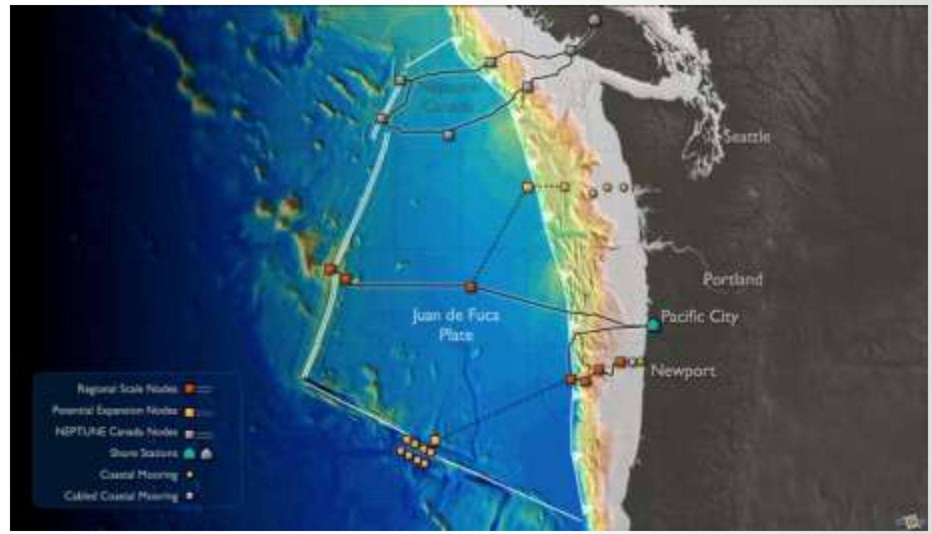


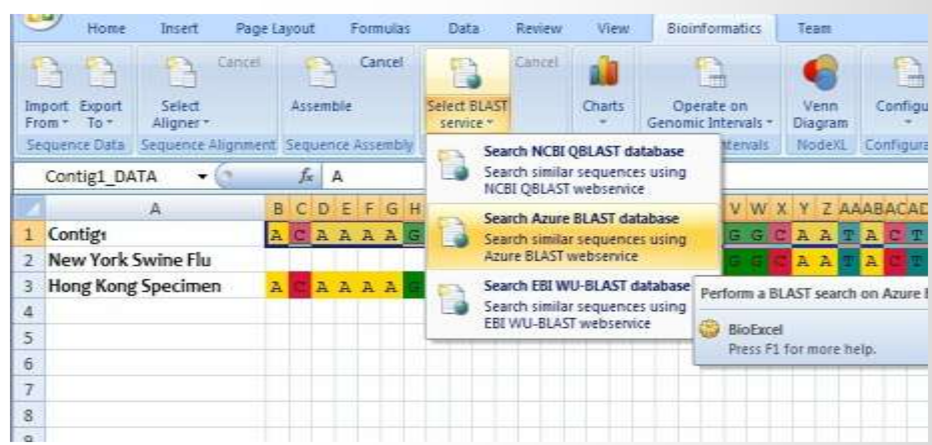
Simple Tools to Answer Complex Questions…

Imagine: the client plus the invisible backend for problem solving

- Give the standard science and engineering desktop tools a seamless extension
- Use a spreadsheet to invoke genomic analysis tools running on 600 servers
- Use a simple script to orchestrate data analytics and mining across 10000 MRI Images
- Pull data from remote instruments for visualization on the desktop

Create a revolution in scientific capability for everybody





Extend the research footprint

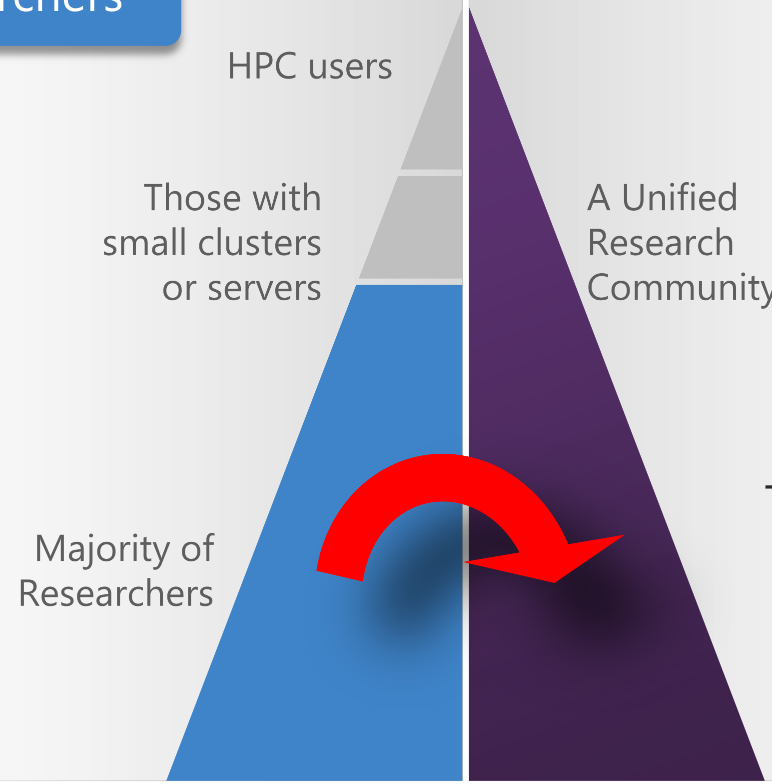
Today

Majority of Researchers

Use laptops and desktop computers

Overwhelmed by data

Finding analysis ever more difficult; sharing even harder



Tomorrow?

Paradigm Shift

Powerful tools

Data and analysis tools in the cloud
Cycles, storage, support

Building communities around research results

The ability to marshal needed resources on demand
Without caring or knowing how it gets done…

Accelerating discovery

VENUS-C

Venus-C

Virtual multidisciplinary EnviroNments USing Cloud infrastructures

European Cloud Computing Strategy

Vice-President Neelie Kroes, responsible for the Digital Agenda

Three Pillars for Cloud

- *Legal frameworks*
- *Technical and commercial fundamental elements*
- *Development of the cloud market by supporting pilot projects of cloud deployments*

Official opening of the Microsoft Cloud & Interoperability Center, March 2011

Cloud Power

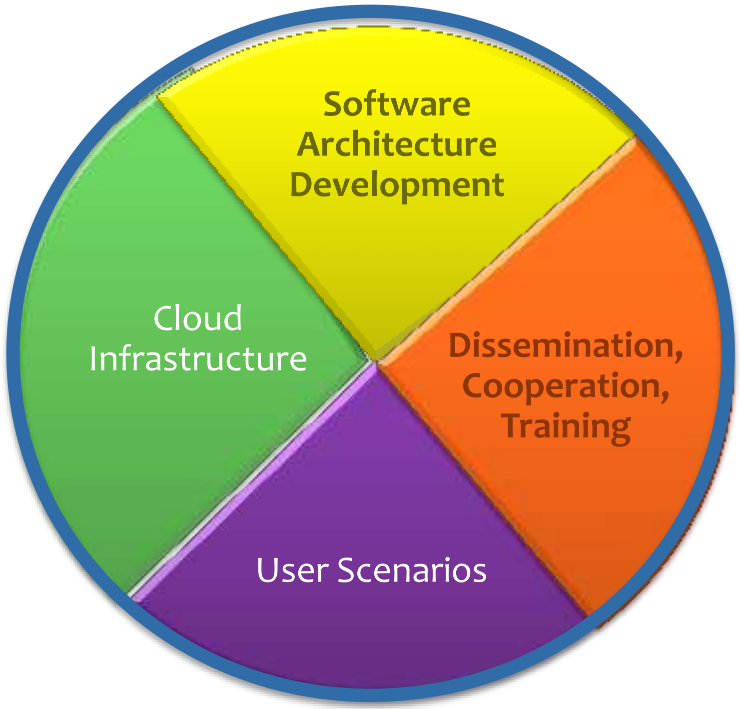
Neelie Kroes on international standardisation & open specifications

"I count here on the further support and commitment of Microsoft and all the other participants."

Industry contribution to the European Cloud Strategy

Building an industry-quality, highly scalable & flexibale Cloud infrastructure

Microsoft EMIC–MICGR-MRL



Coordinated by Engineering – Investment in infrastructure provision & software development.

Microsoft invests in Azure resources & manpower through Redmond & its European data centres

A user-centric Approach

Building a Cloud Infrastructure with user needs interwoven

Bringing about fundamental changes in scientific discovery & innovation

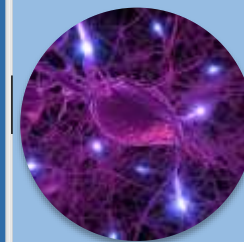
Structural Analysis of buildings

Building Information Management

AquaMaps – Marine Biodiversity data

Civil Protection and Emergencies

Bioinformatics



Systems Biology

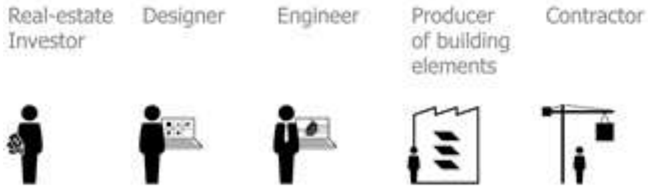
Drug Discovery

Some Success Stories

- Interactive computation of fire risk and fire propagation estimation
- Access to burst-scalable cloud compute and storage
- Web-based GIS based on Bing Map

[Wild Fire Demo]

collaboratorio
innovators in architecture

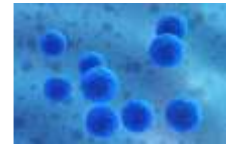


- Collaboratorio & its new start-up Green Prefab
- Collaborative platform for the design of ecofriendly & affordable buildings
- Selected by INTESA SAN PAOLO Start-up intitiative; expanding to US

"We feel like pioneers in the right direction to the still untouched gold mine," Furio Barzon

Extending Cloud Usage - Pilots & Experiments

Engineering & Science

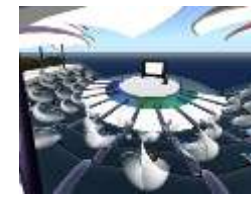



Architecture & Civil Engineering
Biology









NEW DISCIPLINES

Earth Sciences, Healthcare, Maths, Mechanical Engineering, Physics, Social Media, Education

Start-ups

Computer resources can be scaled as required without committing to large capital purchases, which is critical to the success of our small business. **Molplex UK**

DFRC is part of the EU Flagship project PERSEUS on maritime security. Scaling our platform with VENUS-C will enable us to support future growth in terms of vessels monitored in real time & usability by operators.

Value-add for eScience

- Distributing, managing and curating data is better served by a virtual, scalable and elastic infrastructure
- Economy of scale, energy costs and environmental impact are better addressed by Cloud computing
- Virtualisation of computing infrastructures can support funding agencies in developing new funding models:
  - Moving from CAPEX to OPEX
- Leading to more science per tax payer €
- Faster to deploy than conventional HPC in emerging scientific and business communities

Thank you

?

Resources

- Microsoft Research
  - http://research.microsoft.com
  - Microsoft Research downloads: http://research.microsoft.com/research/downloads
- Microsoft External Research
  - http://research.microsoft.com/en-us/collaboration/
- Science at Microsoft
  - http://www.microsoft.com/science
- Scholarly Communications
  - http://www.microsoft.com/scholarlycomm
- CodePlex
  - http://www.codeplex.com

References

- The Fourth Paradigm: Data-Intensive Scientific Discovery – Tony Hey (Editor), Stewart Tansley (Editor)
- How Will Astronomy Archives Survive the Data Tsunami? Communications of the ACM Vol. 54 No. 12
- Data-Intensive Science: A New Paradigm for Biodiversity Studies - Steve Kelling, Wesley M. Hochachka, Daniel FinkMirek Ried EwaldRich Caruana, Grant Ballard, Giles Hooker. Bioscience, 2009.
- Data-intensive e-science frontier research – HB Newman, MH Ellisman… - Communications of the ACM, 2003
- Data-intensive computing in the 21st century – I. Gorton, P Greenfield, A Szalay… - IEEE Computer, 2008