Cloud Computing for Research

Fabrizio Gagliardi

Microsoft Research Europe
External Research, Director

Microsoft

Introduction

- Happy to be here at a HealthGrid event
- Still remember the start of this initiative in the early 2000's at the same time we started EU DataGrid and CrossGrid
- Interesting evolution of a self-sustained infrastructure, leveraging DCI in Europe, mostly EU supported
- Interesting critical mass of well organized scientific user communities with a common focus to Health and Well Being
- Important to understand and develop longer term sustainability
- Important to review technology evolution and current trends in DCI

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
  - Search, email, social networks
  - File storage

The Cloud is built on massive data centers

Range in size from "edge" facilities to megascale.

Economies of scale

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

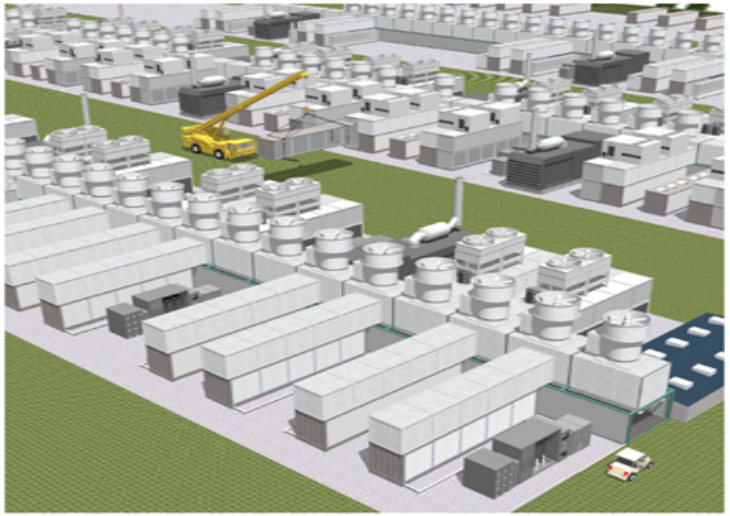
Technology	Cost in small-sized Data Center	Cost in Large Data Center	Ratio
Network	$95 per Mbps/ Month	$13 per Mbps/ month	7.1
Storage	$2.20 per GB/ Month	$0.40 per GB/ month	5.7
Administration	~140 servers/ Administrator	>1000 Servers/ Administrator	7.1

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

Microsoft Advances in DC Deployment

Conquering complexity.

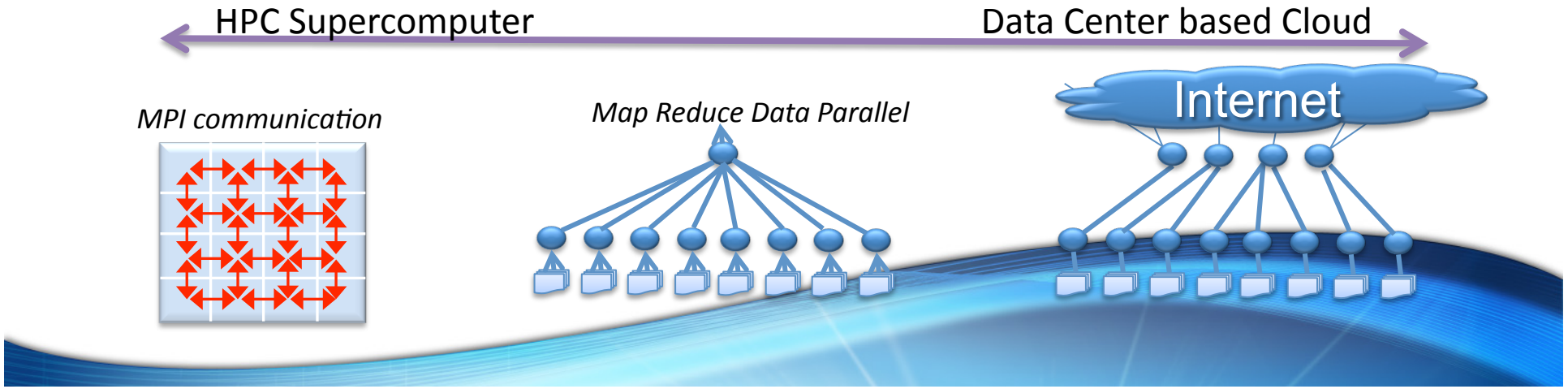
- Building racks of servers & complex cooling systems all separately is not efficient
- Package and deploy into bigger units



http://www.microsoft.com/showcase/en/us/details/36db4da6-8777-431e-aefb-316ccbb63e4e

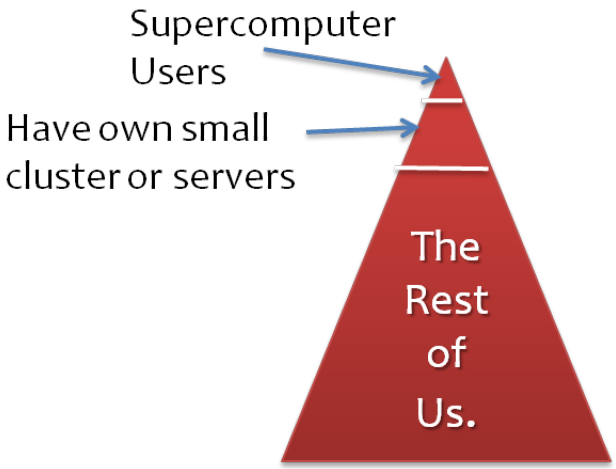
DCs vs Grids, Clusters and Supercomputer Apps

- Supercomputers
  - High parallel, tightly synchronized MPI simulations
- Clusters
  - Gross grain parallelism, single administrative domains
- Grids
  - Job parallelism, throughput computing, heterogeneous administrative domains
- Cloud
  - Scalable, parallel, resilient web services



Changing the way we do research

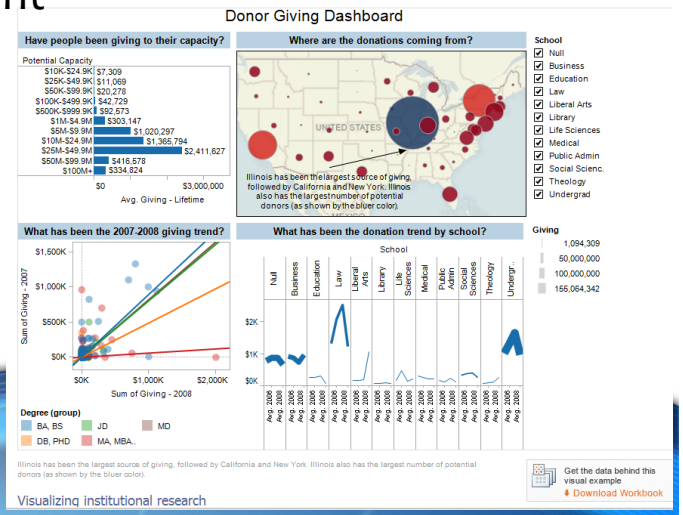
- The Rest of Us
  - Use laptops
  - Got data, now what?
  - And it is really is about data, not the FLOPS…
    - » Our data collections are not as big as we wished
    - » When data collection does grow large, not able to analyze
  - Tools are limited, must dedicate resources to build analysis tools
- Paradigm shifts for research
  - The ability to marshal needed resources on demand
    - » *Without caring or knowing how it gets done…*
  - Funding agencies can request grantees to archive research data in common public repositories
  - The cloud can support very large numbers of users or communities in a flexible way



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…

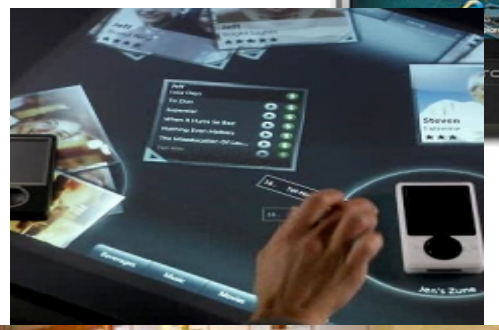


The Clients+Cloud Platform

- At one time the “client” was a PC + browser

  Now:
  - The Phone
  - The laptop/tablet
  - The TV/Surface/Media wall

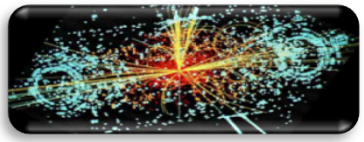
  And the future:
  - The instrumented room
  - Aware and active surfaces
  - Voice and gesture recognition
  - Knowledge of where we are
  - Knowledge of our health



The Future: an Explosion of Data

Experiments

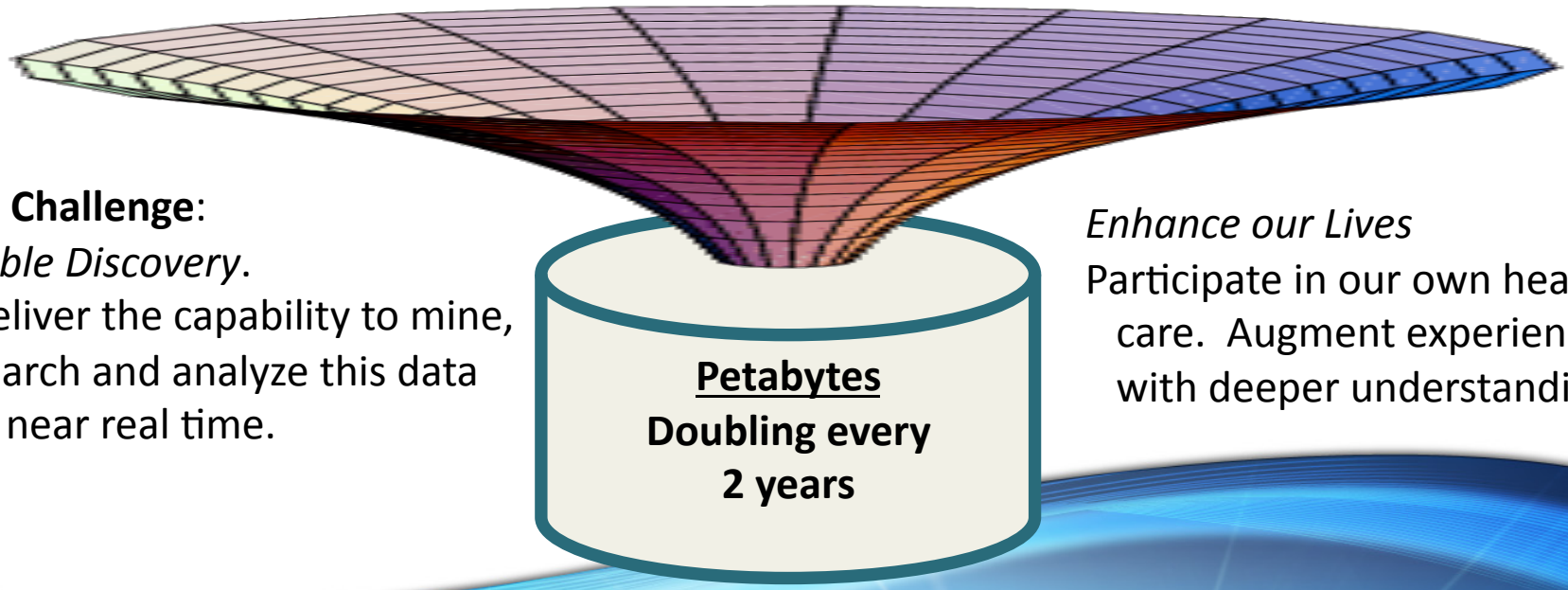
Simulations



Archives

Literature

Instruments



The Challenge:

Enable Discovery.

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

Enhance our Lives

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

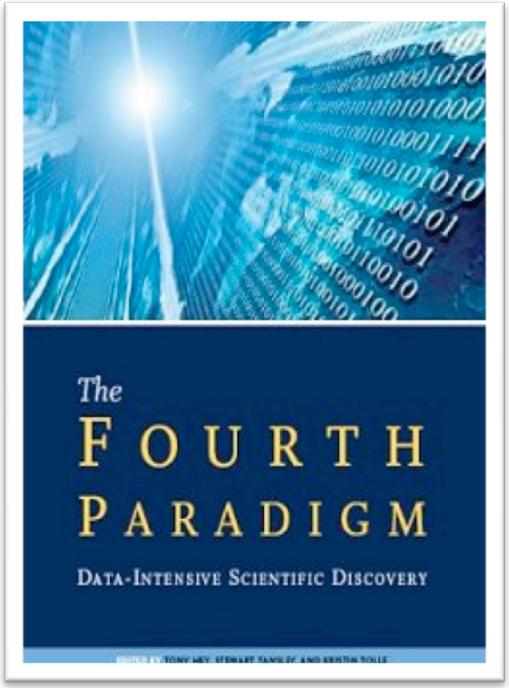
Changing Nature of Discovery

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

Diverse expectations
- Client-centric and infrastructure-centric



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

The Cloud Landscape

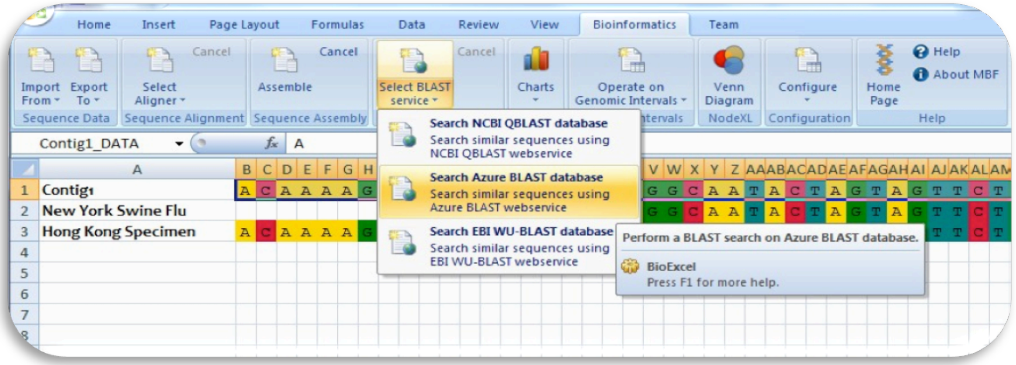
- Infrastructure as a Service (IaaS)
  - Provide a data center and a way to host client VMs and data.
- Platform as a Service (PaaS)
  - Provide a programming environment to build a cloud application
  - The cloud deploys and manages the app for the client
- Software as a Service (SaaS)
  - Delivery of software from the cloud to the desktop

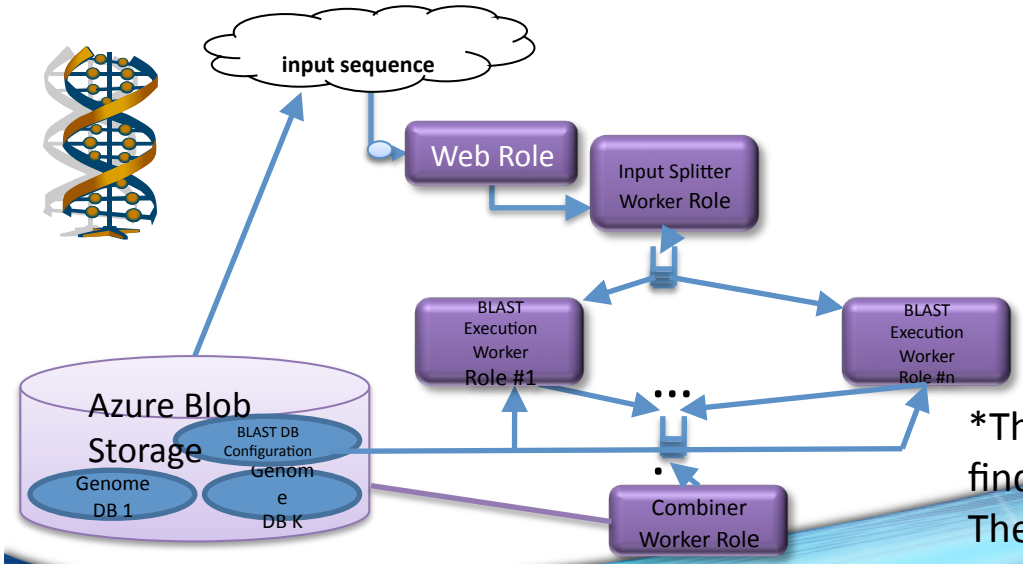
Some examples based on MSR ongoing Cloud Research Engagements

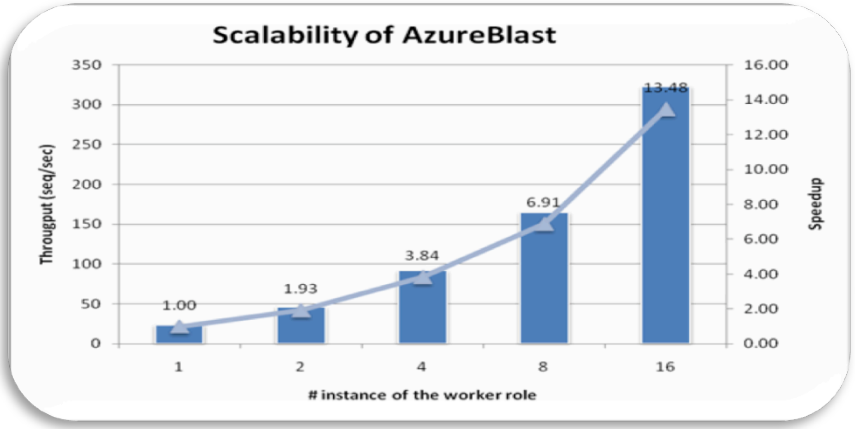
AzureBLAST*

Seamless Experience

- Evaluate data and invoke computational models from Excel.
- Computationally heavy analysis done close to large database of curated data.
- Scalable for large, surge computationally heavy analysis.
- Test local, run on the cloud.

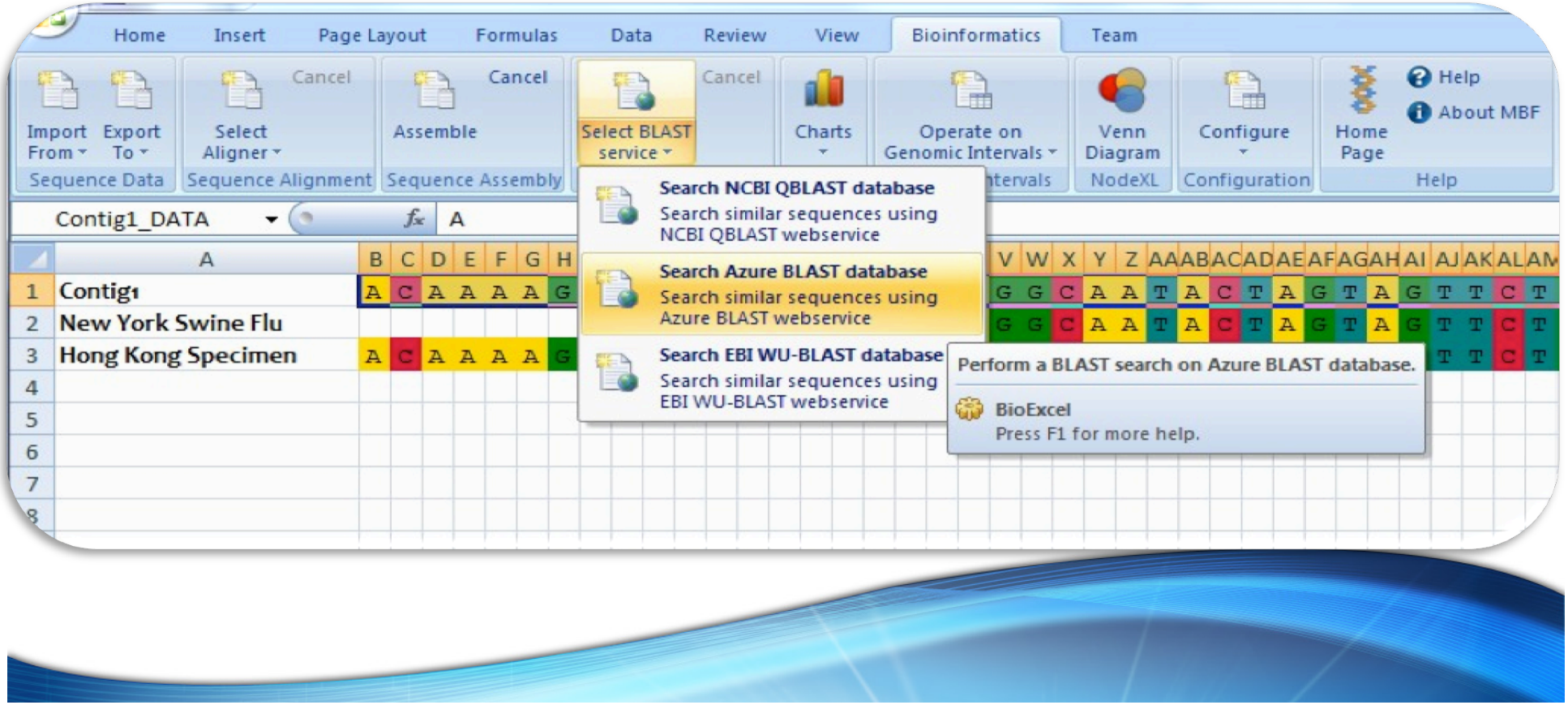




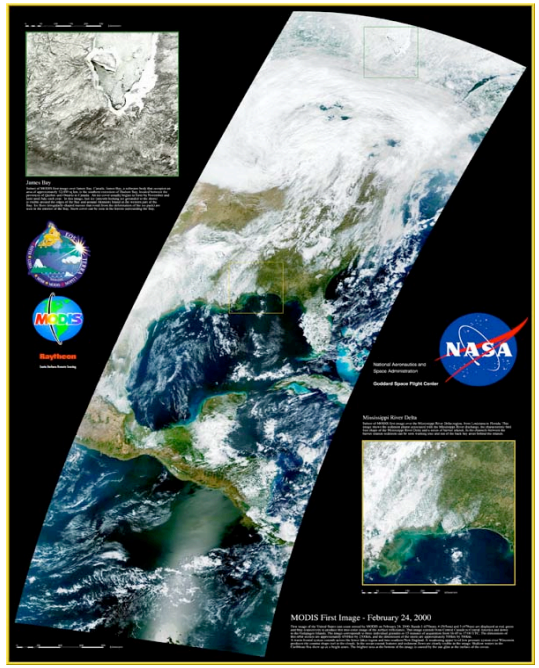


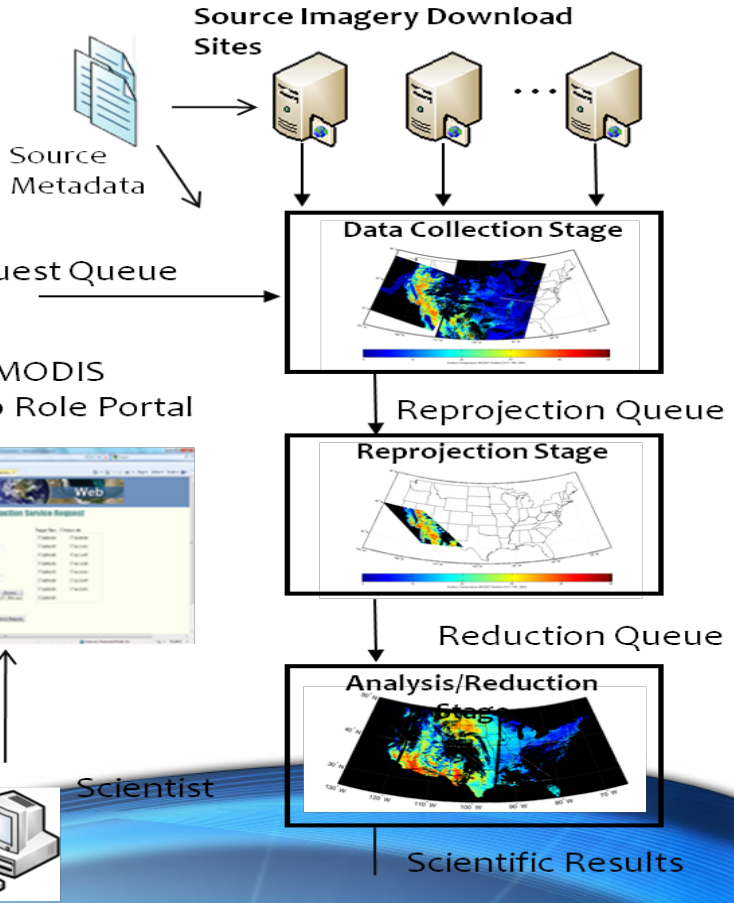
*The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases

Making Excel the user interface to the cloud



AzureMODIS – *Azure Service for Remote Sensing Geoscience*





- 5 TB (~600K files) upload of 9 different imagery products from 15 different locations (~6 days of download)
- 4 TB reprojected harmonized imagery ~35000 cpu hours
- 50 GB reduced science variable results ~18000 cpu hours (~14 hour download)
- 50 GB additional reduced science analysis results ~18000 cpu hours (~14 hour download)

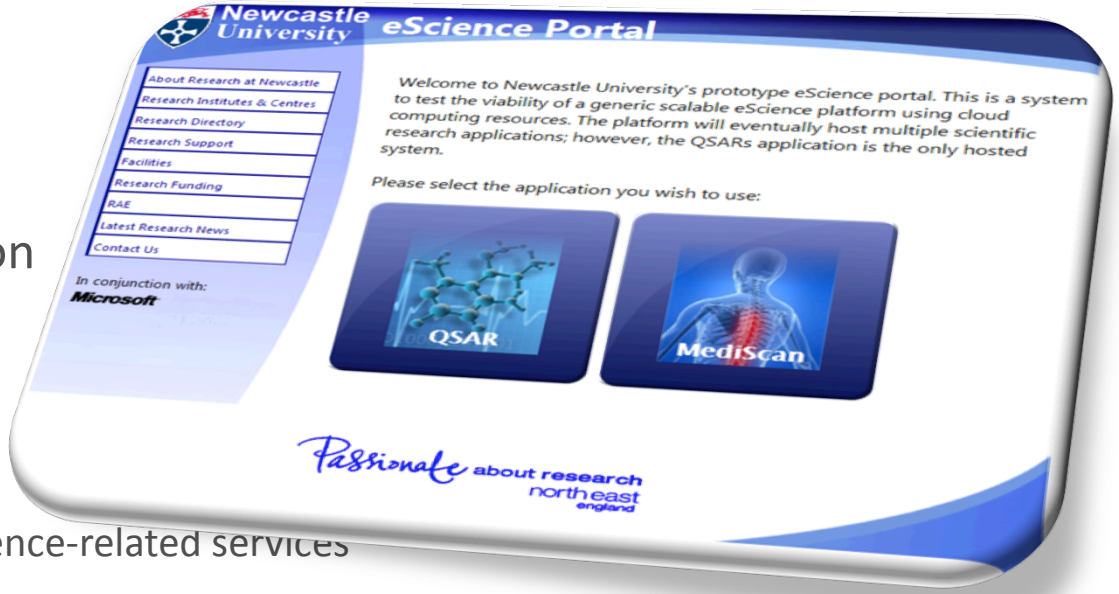
Project JUNIOR

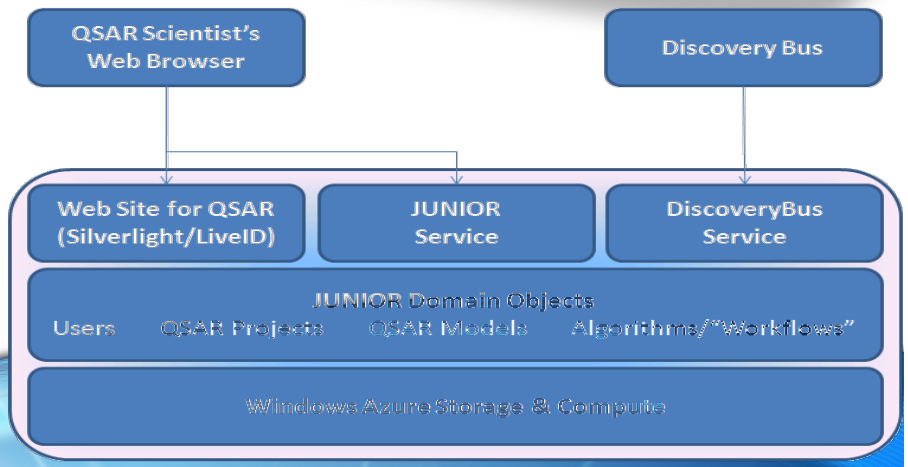
Newcastle University, UK -Paul Watson

- Investigating applicability of commercial clouds for scientific research
- Build a working prototype for use-cases in chemo-informatics
- Uses Microsoft technologies to build science-related services (Windows Azure, Silverlight…)
- Exploits Azure and Amazon Clouds

Built initial proof-of-concept

- Silverlight UI for basic Quantitative Structure-Analysis Relationship (QSAR) modeling
- Demonstrated ability to scale QSAR computations in Windows Azure





Dynamic shared state

- multiplayer games, virtual worlds, social networks, clinical tests
  - New modalities of collaboration

Reaching Out: Azure Research Engagement project

In the U.S.

Memorandum of Understanding with the National Science Foundation

- Provide a substantial Azure resource as a donation to NSF
- NSF will provide funding to researchers to use this resource

In Europe

- We interested in direct engagement with the thought leaders in the U.K., France and Germany
- EC engagements where possible (VENUS-C to start off)

In both we provide our engagement team

- We provide workshops, tutorials, best practices and shared services, learn from this community, shape policy…

In Asia

- We wish to explore possibilities.

VENUS-C

Virtual multidisciplinary EnviroNments USing Cloud infrastructures

Funding Scheme: *Combination of Collaborative Project and Coordination and Support Action: Integrated Infrastructure Initiative (I3)*

Program Topic: *INFRA-2010-2 1.2.1. Distributed Computing Infrastructures*

Consortium

1 (co)	Engineering Ingegneria Informatica S.p.a.	ENG	IT
2	European Microsoft Innovation Centre	EMIC	DE
3	European Charter of Open Grid Forum	OGF.eeig	UK
4	Barcelona Supercomputing Center – Centro Nacional de Supercomputación	BSC-CNS	ES
5	Universidad Politecnica de Valencia	UPV	ES
6	Kungliga Tekniska Hoegskolan	KTH	SE
7	University of the Aegean	AEG	GR
8	Technion	TECH	IL
9	Centre for Computational and Systems Biology	CoSBi	IT
10	University of Newcastle	NCL	UK
11	Consiglio Nazionale delle Ricerche	CNR	IT
12	Collaboratorio	COLB	IT

Goals

1. Create a platform that enables user applications to leverage cloud computing principles and benefits.
2. Leverage the state of the art to on-board early adopters quickly, incrementally enable interop with existing DCI and push the state of the art where needed to satisfy on-boarding and interop
3. Create a sustainable infrastructure that enables the cloud computing paradigms for the user communities inside the project, the one from the call for applications, as well as others.

Supporting multiple basic research disciplines

Biomedicine: Integrating widely used tools for Bioinformatics (UPV), System Biology (CosBI) and Drug Discovery (NCL) into the VENUS-C infrastructure

Civil Protection and Emergency: Early fire risk detection (AEG), through an application that will run models on the VENUS-C infrastructure, based on multiple data sources.

Civil Engineering: Support complex computing tasks on Building Information Management for green constructions (provided by COLB) and dynamic building structure analysis (provided by UPV).

D4Science: Integrating computing through VENUS-C on data repositories (CNR). In particular focus will be on Marine Biodiversity through Aquamaps.

Open Call for 20 e-Science Applications

20K€ funding each (in addition to Azure Compute , Storage and Network Resources)

-> porting applications to the cloud

-> education and traning

-> scalability tests

Thanks for your attention