

The Social Grid

Leveraging the Power of the Web
Focus on Development Simplicity

Roger S. Barga, PhD
Architect, Technical Computing at Microsoft



Technical Computing at Microsoft

<http://www.microsoft.com/science>

Earth
Sciences



Life
Sciences



Social
Sciences



Collaborative
Research

Computer &
Information
Sciences



Accelerating Discovery



New Materials,
Technologies
& Processes

$$E=MC^2$$

Math and
Physical Science



Emergence of a New Science 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**

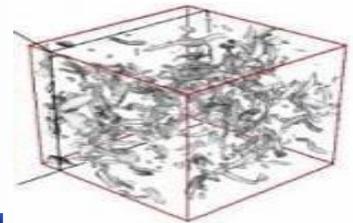
- Simulation of complex phenomena

Today – **Data-Centric Science or eScience**

- Unify theory, experiment, and simulation
- Using data exploration and data mining
 - Data captured by instruments
 - Data generated by simulations
 - Data generated by sensor networks



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



Slide thanks to Jim Gray



A New Set of Challenges

- Grid middleware and cyberinfrastructure is fast becoming too complex and difficult to use
- Support the transition to data-centric eScience
 - *Computation is no longer the bottleneck*
- New dynamic in science investigations
 - *Collaborative, distributed, cross disciplinary*
 - *Science is becoming highly social (Research 2.0)*
- Need for simplicity and ease of development
- Can we learn lessons from the Web



The Web as a Platform for Research?

- Has tremendous momentum
- It is **the** channel for result dissemination
- The browser is the **universal canvas** for the delivery of information and functionality
- Web protocols, technologies, and middleware are well supported by the IT industry
- Today it is the contemporary **platform** for distributed, internet-scale applications
- Emergence of “*software-as-a-service*”
- Collection of ‘Web 2.0’ technologies is maturing



Scientific Data Servers for Hydrology

Work with Berkeley Water Center to use modern (relational) database technology

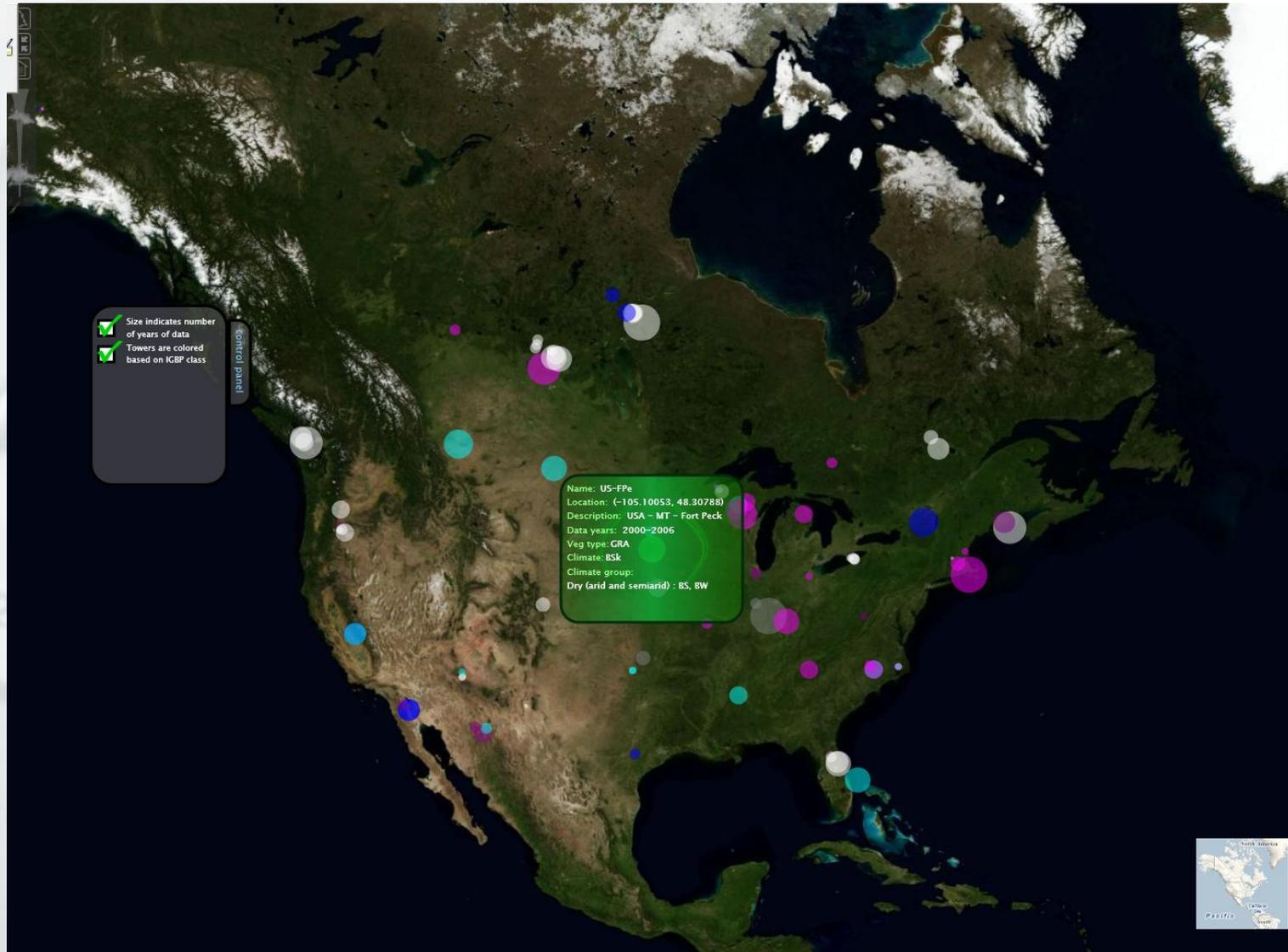
- 149 Ameriflux sites across the Americas reporting minimum of 22 common measurements
- Carbon-Climate Data published to and archived at Oak Ridge
- Total data reported to date on the order of 192M half-hourly measurements since 1994

<http://public.ornl.gov/ameriflux/>

Microsoft Project Lead: Catharine van Ingen



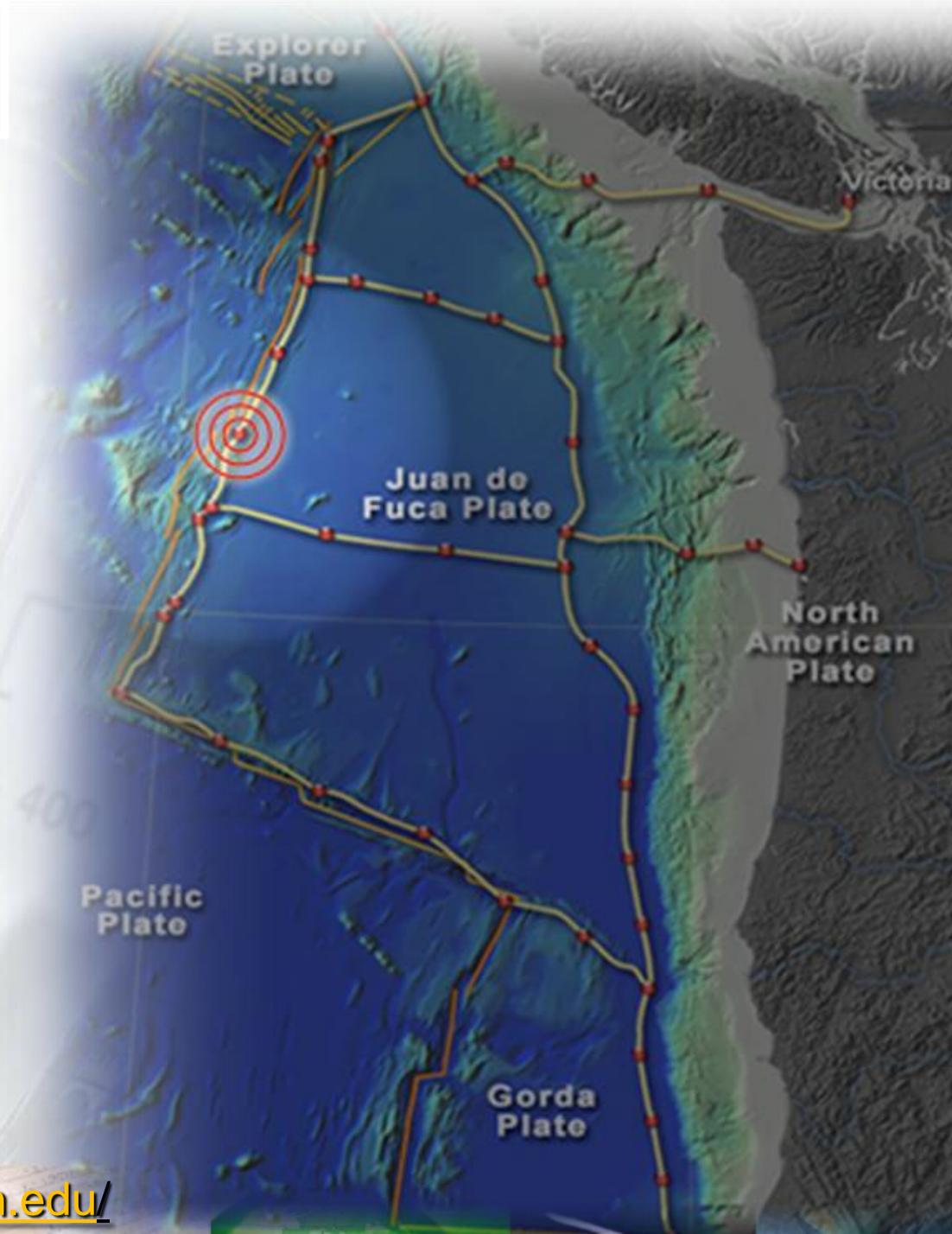
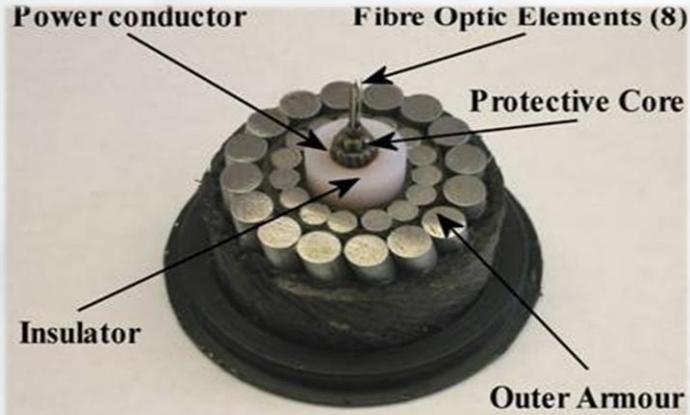
Mashup of Ameriflux Sites



Work of Savas Parastatidis



Project Neptune



Programmable Sensors & Remote Instruments

Undersea Sensor Network

The screenshot shows the NEPTUNE web interface in a Microsoft Internet Explorer browser. The address bar displays <http://www.neptune.washington.edu/>. The page features navigation tabs for "Scientists", "Teachers & Students", and "General Public", along with a "Log Out" button. Below the navigation is a menu with "Highlights", "News & Events", "Manage Feeds", "Sensor Controls", "Collaborations", and "Plan Experiment".

The main content area is divided into two panels:

- Interactive Map:** Displays a 3D bathymetric map of the seafloor with a network of red nodes connected by black lines. A specific node is highlighted with a red target symbol. The map is labeled "Axial".
- Node Sensors:** A list of sensor data for three nodes: Node D-433, Node D-436, and Node D-437. Each node has a "SUBMIT" button and a list of sensors with checkboxes indicating their status.

Node D-433 Sensors:

- Thermal (floor, always on)
- Thermal (10m)
- Thermal (50m)
- Seismometer (always on)
- Salinity
- Current field vector (offline)
- Microbial concentration
- Oxygen
- Doppler current profiler
- Microbial concentration
- Video
- Hydrophone
- Sample floats (20 remaining)
- AUV

Node D-436 Sensors:

- Thermal (floor, always on)
- Thermal (10m)
- Thermal (50m)
- Seismometer (always on)
- Salinity

Node D-437 Sensors:

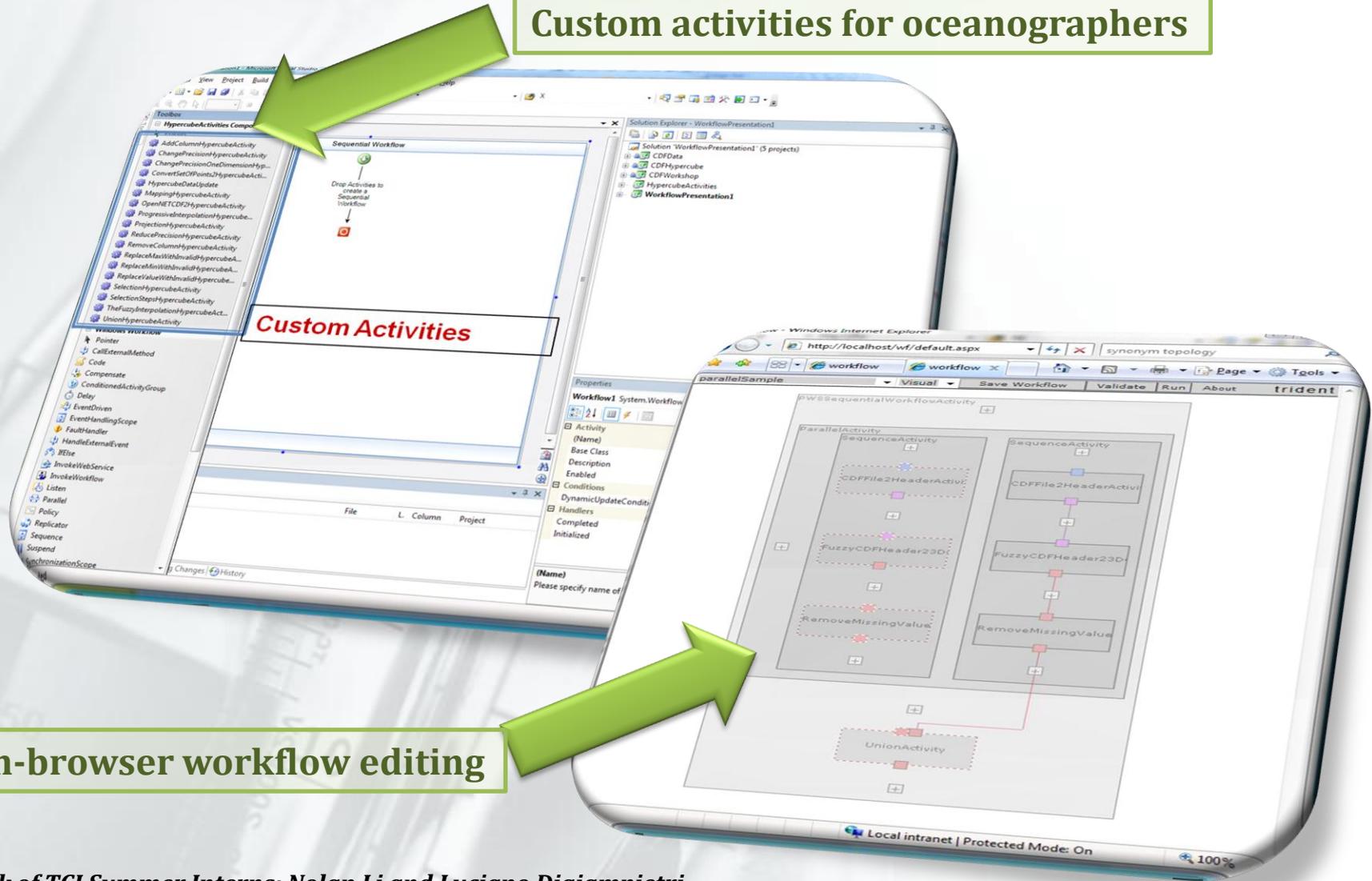
- Thermal (floor, always on)
- Thermal (10m)
- Thermal (50m)
- Seismometer (always on)
- Salinity

Connected & Controllable Over the Internet



Trident – Scientific Workflow for Neptune

Custom activities for oceanographers



In-browser workflow editing

Work of TCI Summer Interns: Nolan Li and Luciano Digiampietri

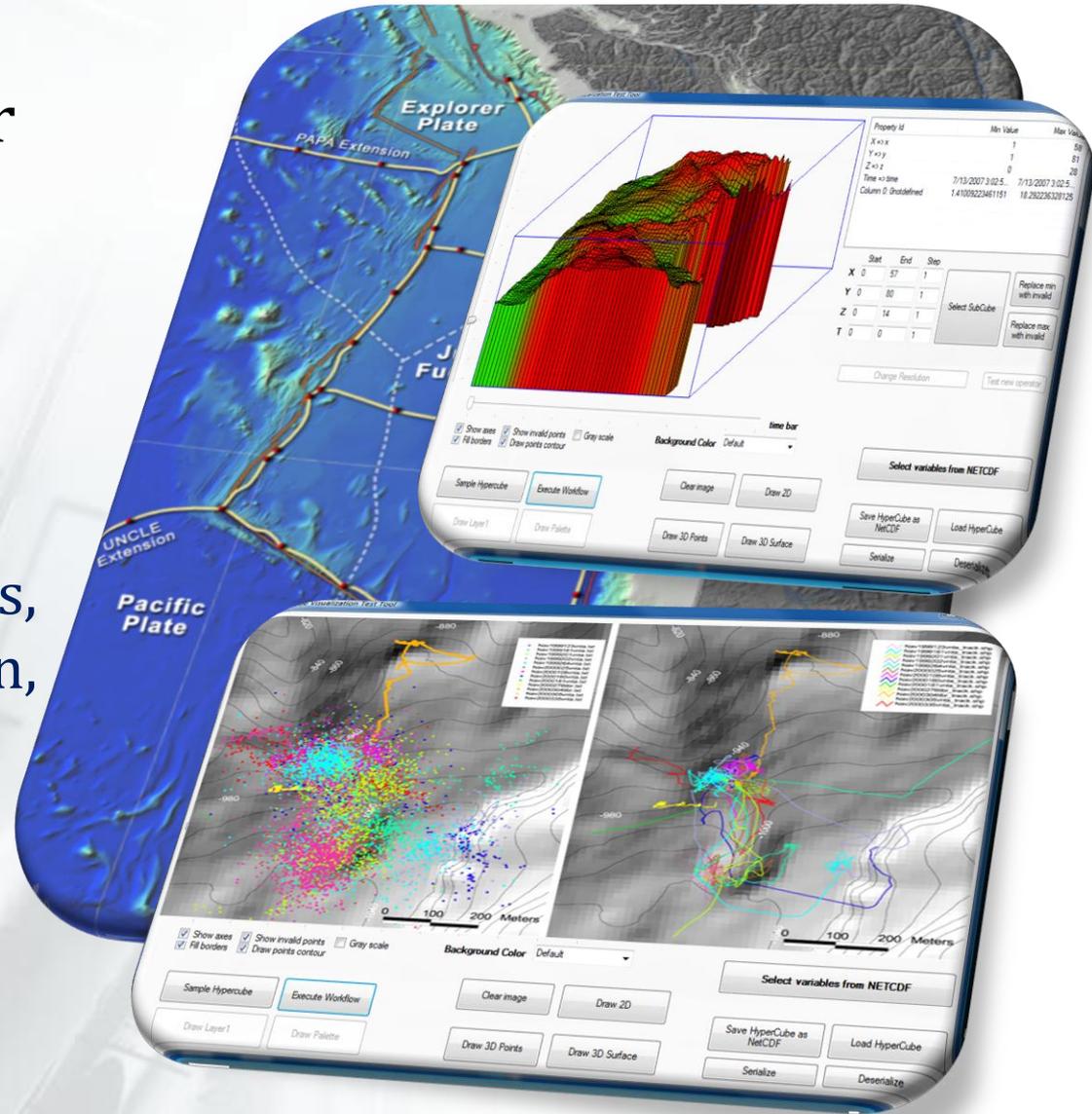
Trident for Neptune

Workflow workbench for oceanography

From raw sensor data to useable data products

Automatic data cleaning,
Integrates multiple models,
Regridding & interpolation,
Analysis

Real time, on-demand
visualization from the
Neptune sensor array



myGrid and the Taverna Workflow System

- Independent third party world-wide service providers of applications, tools and data sets – *in the cloud*.
 - 850 databases, 166 web servers: Nucleic Acids Research Jan 2006.
- My local applications, tools and datasets. In the Enterprise. In the laboratory.
- Easily incorporate new service without coding. So even more services from the cloud and enterprise.



- 3500+ service operations
- All major providers
- Integration application for service providers like BioMOBY and BioMART

Slide thanks to Carole Goble and David DeRoure

eScientists in the Cloud



Individual life scientists, in under-resourced labs, who use other people's applications, with little systems support.

- Exploratory workflows
- Developers (often) the users.
- Consumers are providers.

- A distributed, disconnected community of scientists.
- Decoupled suppliers and consumers of services and workflows.
- Scientists in an enterprise and in large projects
- **Scientists out of the enterprise, in small projects or sole traders.**

Slide thanks to Carole Goble and David DeRoure

200+ projects and sites, ~1000 individual users.
Users throughout UK, USA, Europe, and SE Asia



On the Web

- Users generate content on the Web
 - Blogs, wikis, photographs, videos, etc.
 - They do not have to know HTML
- They form communities
 - Social networks, virtual worlds
- They interact, collaborate, share
 - Instant messaging, web forums, content sites
- They consume information and services
 - Search, annotate, syndicate



And Scientists Today...

- Annotate, share, discover data
- Collaborate, exchange ideas over the Web
- Create communities, social networks
- Use workflow tools to compose services



Example - Connotea (Nature Publishing)

The screenshot shows the Connotea website interface. At the top, the URL is <http://www.connotea.org/search?q=social+networking>. The search bar contains "social networking" and a "Find results" button. The user is logged in as "savas". The page title is "Connotea Organize. Share. Discover." and the navigation menu includes "Home", "Latest News", "About Connotea", "Site Guide", and "Community pages".

The main content area displays search results for "social networking". A blue callout box labeled "search" points to the search bar. Another blue callout box labeled "find resources bookmarked by other users" points to the search results. The results list includes:

- Social network - Wikipedia, the free encyclopedia** (en.wikipedia.org) - Section 4, as of 5/1/2007, is about "Social Networking, Internet Social Networks." Posted by **ascoppin** to **social computing Web 2.0** on **Tue May 01 2007** at 19:26 UTC | [info](#)
- Social Networking Leaves Confines of the Computer - New York Times** (www.nytimes.com) - Posted by **library_mistress** to **networking Social social software** on **Mon Apr 30 2007** at 15:34 UTC | [info](#)
- Social-networking sites link Hispanic youth** (www.cnn.com) - MIAMI, Florida (AP) -- Indie rocker Eric Monterrosa checks his ElHood.com Web page at least three times a day, answering fans, surfing for other new Latin artists and keeping in touch with friends from his native Colombia. Posted by **msgbeeph** to **news** on **Sun Apr 29 2007** at 16:47 UTC | [info](#)
- Social networking in the health context** (www.ingentaconnect.com) - Software and services for creating online social networks. Posted by **Spiky** to **social apps** on **Tue Apr 24 2007** at 15:58 UTC | [info](#)

On the right side, there is a "Hospital Staffing Revenue Cycle Staffing including PFS, Pt. Access, Med. Records" (www.hrgpros.com) and a "Report a problem" link. Below the results, there is a "Related tags:" section with links to [social bookmarking](#), [folksonomy](#), [bookmarking](#), [collaborative - tagging](#), [tagging](#), [collaborative tagging](#), and [csdl-picasso-folks](#).

At the bottom, the status bar shows "Internet | Protected Mode: On".

Mashups: Composing Data and Functionality



SensorMap

Functionality: Map navigation

Data: sensor-generated temperature, video camera feed, traffic feeds, etc.

The Web as a Platform for eResearch

Services not middleware

- *No need to install many thousands of lines of middleware*

Core Services in the Cloud

- Identity
- Blogging, Messaging
- Search, Discovery
- Data processing/visualization
- Content upload, sharing, discovery
- Computation and Storage

University of Southampton Crystal Structure Report Archive

Home About Browse Search Register User Area Help

6,7,9,10,12,13,15,16-Octahydro-benzo-1,4,7,10,13-pentaoxacyclopentadecin

Simon J Coles, Michael B Hursthouse, Jeremy G Frey and Esther Rousay, University of Southampton

$C_{14}H_{20}O_5$

InChI=1C14H20O5/c1-2-4-14-13(3-1)18-11-9-16-7-5-15-6-8-17-10-12-19-14/h1-4H,5-12H2

DOI: 10.594/ecrystals.chem.soton.ac.uk/145

Compound Class: Organic

Keywords: crown ether

Creation Date: 07/07/2004

Deposited By: A. J. Coles

Deposited On: 20/07/2004

Final Result

Structure already known, but accurately redetermined for a local research project.

Data collection parameters

Chemical formula	C14 H20 O5
Crystallisation Solvent	
Crystal morphology	Plate
Crystal system	Orthorhombic
Space group symbol	Pbca
Cell length a	16.4963(18)
Cell length b	8.325(3)
Cell length c	20.061(6)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00

Refinement results

Solution figure of merit	0.0409
R Factor (Obs)	0.0487
R Factor (All)	0.0977
Weighted R Factor (Obs)	0.1008
Weighted R Factor (All)	0.1192

Citation: Coles, S.J., Hursthouse, M.B., Frey, J.G. and Rousay, E. (2004), Southampton, UK, University of Southampton, Crystal Structure Report Archive (doi:10.594/ecrystals.chem.soton.ac.uk/145)

04sjc0831.cif 13k

04sjc0831.cml 6k

Validation

04sjc0831_checkcif.htm 7k

Refinement

04sjc0831.res 6k

04sjc0831_xl.lst 34k

Solution

04sjc0831_prp 6k

04sjc0831_xs.lst 39k

04sjc0831.hkl 702k

04sjc0831.htm 10k

04sjc0831_0k.jpg 57k

04sjc0831_h0l.jpg 85k

04sjc0831_hk0.jpg 88k

Data Collection

04sjc0831_crystal.jpg 17k

Other Files

04sjc0831.doc 78k

04sjc0831.fcf.bt 155k

<http://ecrystals.chem.soton.ac.uk>

Thanks to Jeremy Frey

Services Expose Functionality

BLAST service delivered through a Web browser

The image is a collage illustrating the BLAST service. It features three overlapping windows:

- Top-left window:** The main BLAST web interface. It includes a navigation bar with 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. Below the header, there's a section for 'BLAST Assembled Genomes' with a list of species: Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. A 'Basic BLAST' section is also visible.
- Top-right window:** A 'BLAST' service interface with a form for 'Enter Query Sequence'. It includes fields for 'Enter accession number, gi, or FASTA sequence', 'Query subrange', 'From', 'To', 'Or, upload file', 'Job Title', and 'Choose Search Set' (with options for Human genomic + transcript, Mouse genomic + transcript, and Human genomic plus transcript).
- Bottom-right window:** A window displaying XML code, likely a WSDL document. The code includes definitions for 'WSNCBIblast' and 'EBIParams', and lists various SOAP actions like 'program', 'matrix', 'exp', 'filter', 'numal', 'scores', 'align', 'gapalign', 'opengap', 'dropoff', and 'async'.

A blue box with the text 'BLAST' is overlaid on the main interface window.

BLAST service (WSDL) that can be integrated into an application

Services can be Composed

Taverna
Workflow

The screenshot displays the Taverna Workbench interface with several key components:

- Workflow diagram (A):** A flowchart showing the composition of services. It includes nodes like 'GetUniqueHomolog', 'GetMouseGenes', 'CreateFasta', 'hsSeq', 'mmSeq', 'rnSeq', 'fasta', 'sequence_export', 'FlatImageList', 'outputPlot', and 'MMusIDs'. The diagram shows how these services are interconnected to process data.
- Available services (C):** A list of services available for use, categorized by provider. Services include 'GetDomainsFromGAMthEvalue', 'GetAccFromRetredGI', 'ProteinReportSetDescription', 'GetFastaForKeyedList', 'RedundantGroupForKeyedList', 'GetFastaFromRedundantGroupForKeyedList', 'Biomart_ensembl_mart_22_1@martub.ebi.ac.uk', 'frutripes_gene_ensembl', 'hsapiens_gene_est', 'cbriggiae_gene_est', 'morvegicus_gene_est', 'drosio_gene_ensembl', 'ggallus_gene_ensembl', 'colegans_gene_ensembl', 'morvegicus_gene_ensembl', 'drosio_gene_est', 'ggallus_gene_est', and 'cbriggiae_gene_ensembl'.
- Resource usage report (D):** A report showing the usage of external resources. It lists resources on 'martub.ebi.ac.uk' (4 instances) and 'industry.ebi.ac.uk' (3 instances). The report includes a table with columns for 'Biomart', 'Dataset Name', and 'Proc'.
- Configuring query for GetHSGenes (D):** A configuration window for the 'GetHSGenes' service. It shows options for 'Type of sequence to export' (REGION, GENE, PROTEIN), 'Sequence export options', 'Type of sequence to fetch' (Genes), 'Desired sequence options' (5' upstream only), 'Extents' (5' flank: 200, 3' flank: 1000), and 'Sequence glyph'.
- Advanced model explorer (B):** A table showing the workflow object and its metadata. It includes columns for 'Retries', 'Delay', 'Backoff', 'Threads', and 'Critical'.
- Enactor invocation (E):** A window showing the status and results of the workflow execution. It includes a 'Process report' and a 'Results' section displaying sequence data for 'MMusIDs', 'HScapIDs', and 'RnorIDs'.

Data is Easily Shareable

The image displays three overlapping browser windows from the Sloan Digital Sky Survey website, illustrating the ease of sharing and accessing astronomical data.

Background Window: Sloan Digital Sky Survey / SkyServer
The main website interface, featuring navigation menus (Home, Tools, Schema, Projects, Astronomy, SDSS, Contact Us, Download, Site Search, Help) and a sidebar with 'SkyServer Tools' (Famous places, Get images, Visual Tools, Explore, Search, Object upload, CasJobs) and 'Info Links' (About Astronomy, About the SDSS, SDSS Data Release 5, SDSS Project Website, Open SkyQuery, Images of RC3 galaxies).

Middle Window: SDSS DR5 Image List Tool
A tool for searching and viewing astronomical images. It includes a search form with fields for name, ra, dec, and a 'Get Image' button. Below the search form is a grid of image thumbnails with their respective coordinates.

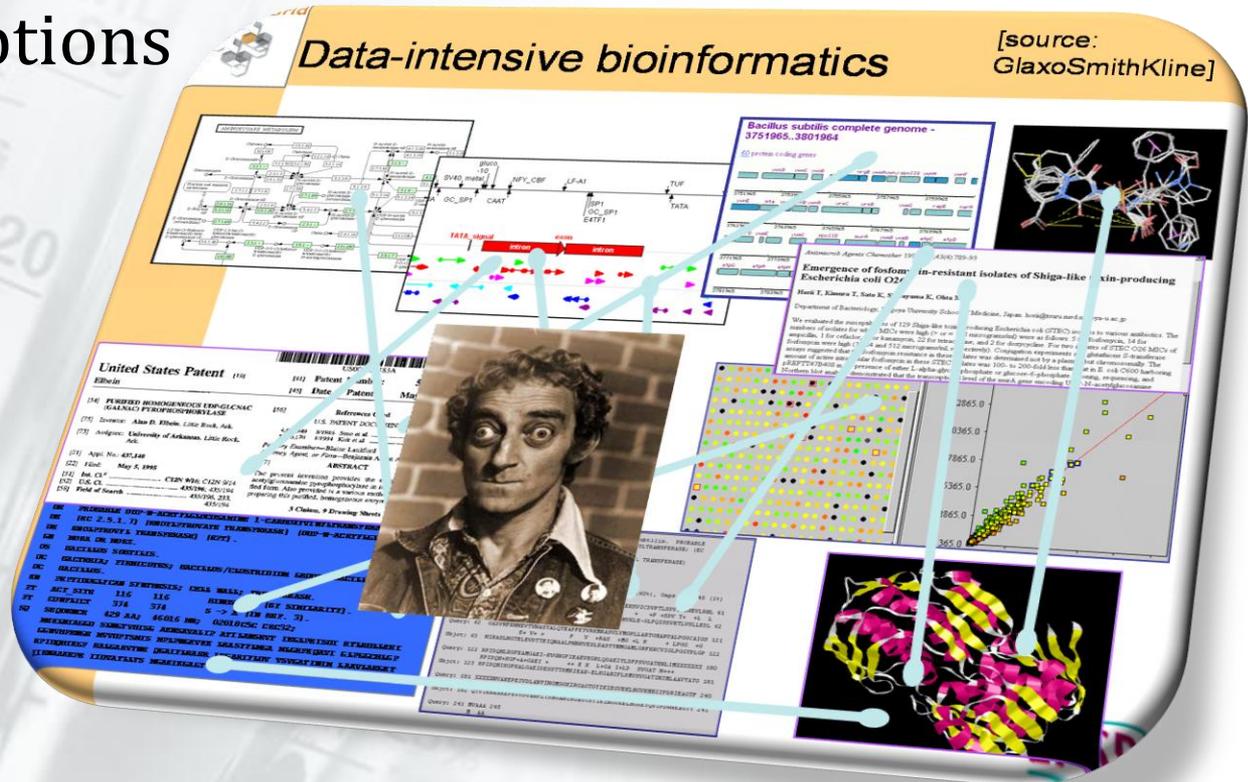
Foreground Window: SkyServer Object Explorer
A detailed view of a specific object, identified as 'SDSS J113459.47+002509.1'. It displays the object's name, coordinates (ra=173.747818, dec=0.419213), and object ID (ObjId = 588848900446814264). The window shows a 'Summary' section with various parameters (mode, ra dec, status, flags, photoObj, specObj) and a 'SpecObjID = 79597814924967936' section with a spectrum plot and associated parameters (plate, imgid, fiberid, z, zErr, zConf, specClass, ra, dec, fiberMag_r, objid).

Sloan Digital Sky Server/SkyServer
<http://cas.sdss.org/dr5/en/>

Knowledge can be created/published/archived/discovered

myGrid

- Semantic relationships between different data
- Semantic descriptions of services
- Annotations
- Provenance
- Repositories
- Ontologies
- Folksonomies



Grids in Industry

- Google, Amazon, Yahoo, eBay and Microsoft are the major 'Cloud Platform' providers
 - All have infrastructures of hundreds of thousands of servers
 - Many large data centers, distributed across multiple continents
 - Have developed proprietary technologies for job scheduling, data sharing and management
 - Care about power consumption, fault tolerance, scalability, operational costs, performance, etc.

They are living the "Grid dream" on a daily basis

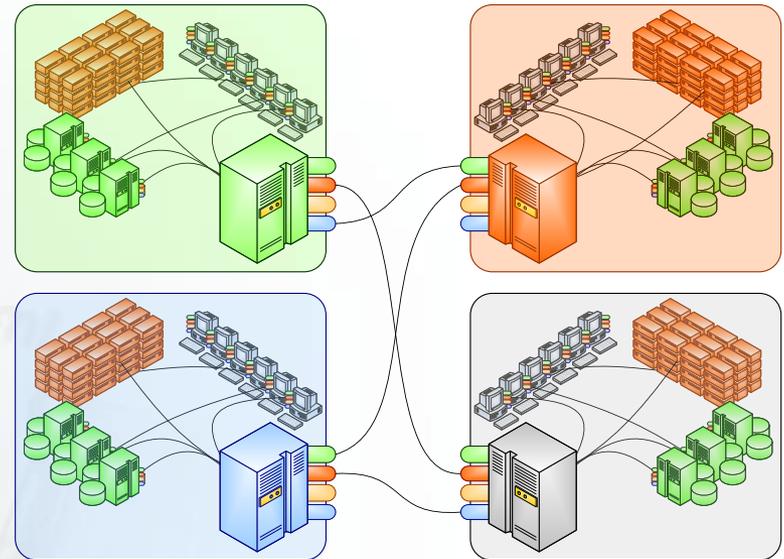


Google as an Example

- Estimated 450,000* servers distributed around the world

*source: Wikipedia

- Google File System - highly distributed, resilient to failures, parallel, etc.
- Schedulers and load balancers for the distribution of work
 - Use their 'Map-Reduce' middleware as parallel computational model



Amazon web services: simple storage service (s3)

- S3 is storage for the Internet
 - Designed to make web-scale computing easier for developers
- Provides a simple Web Services interface to store and retrieve any amount of data from anywhere on the Web
 - ‘CRUD’ philosophy – Create, Read, Update and Delete operations
- Uses simple standards-based REST and SOAP Web Service interfaces
 - Built to be flexible so that protocol or functional layers can easily be added



Amazon s3 Functionality

- Intentionally built with a minimal feature set
 - Write, read, and delete objects containing from 1 byte to 5 gigabytes of data each
- Can store unlimited number of objects
 - Each object is stored and retrieved via a unique, developer-assigned key
- Authentication mechanisms provided
 - Objects can be made private or public, and rights can be granted to specific users
- Default download protocol is HTTP
 - BitTorrent protocol interface is provided to lower costs for high-scale distribution



Amazon web services: elastic compute cloud (ec2)

- Compute on demand service that works seamlessly with their S3 storage service
- Create Amazon Machine Image (AMI) containing application, libraries and data
- Use EC2 Web Service to configure security and network access
- Use EC2 to start, terminate and monitor as many instances of your AMI as you want

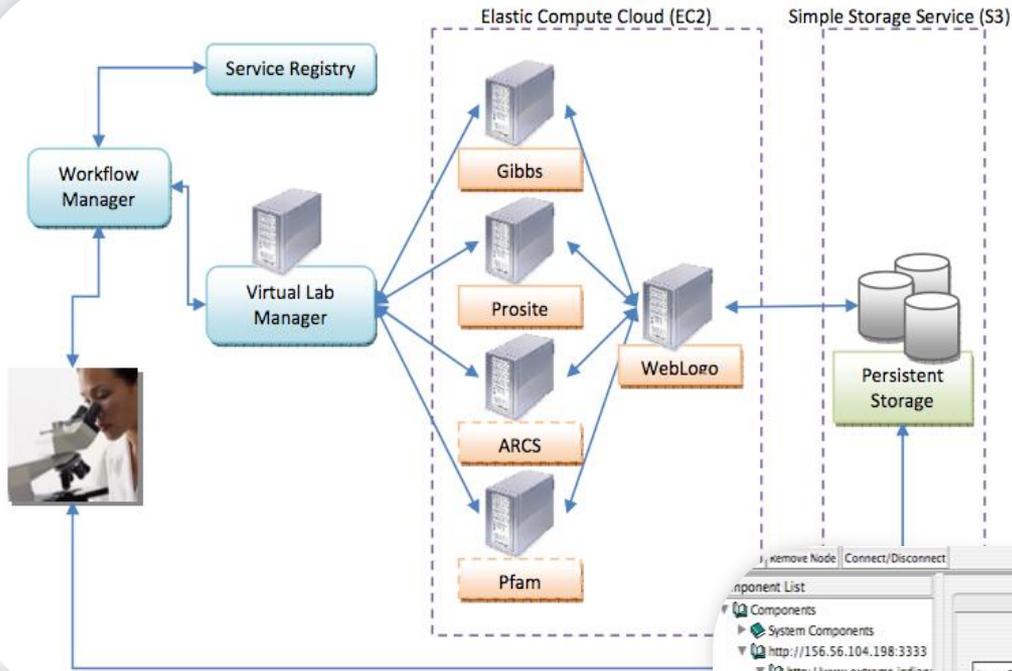
Each instance has:

- 1.7 GHz x86 Processor
- 1.75 GB RAM
- 160 GB local disk
- 250 MB/s network bandwidth

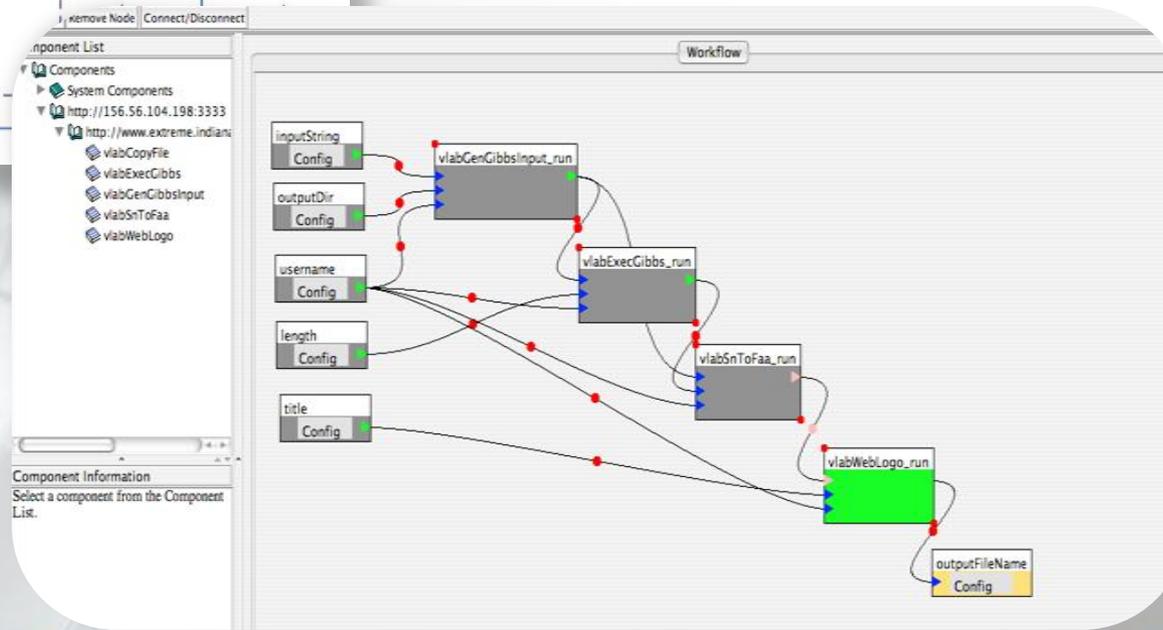
*Used by Catlett and Beckman as capacity computing alternative to TeraGrid
'SPRUCE' capability computing for emergency urgent response*



A Grad Student Project Using S3 and EC2



Gene Analysis Virtual
Lab Experiment
by Jong Youl Choi
at Indiana
(For Beth Plale and Sun Kim)



Data-Intensive High-Performance Computing

- A new generation of facilities to support eResearch on the cloud
- Data-intensive
 - Large storage capacity
 - Functional-style programming for data filtering, searching (e.g. MapReduce)
 - Storage-as-a-service
- Compute-intensive
 - State-of-the-art clusters
 - No need to be the fastest in the world; few top100 ones
 - Scientific applications-as-services



Social Grids and the Web

A Call to Action

- Focus on **solutions** for scientific/technical computing and **not just on infrastructure**
- Focus on “data-centric eScience”
 - Help domain experts define formats for representing and annotating domain-specific data
- Keep it simple, build on known Web technologies
 - Solutions that “just work” without the need for complicated middleware platforms
 - Leverage only existing, Web infrastructure (HTTP, XML, simple Web Services, services in the cloud)





Microsoft[®]

Your potential. Our passion.[™]

