Science Gateways – Leveraging Modeling and Simulations in HPC Infrastructures via Increased Usability

Sandra Gesing, Jens Krüger, Rion Dooley, Richard Grunzke, Marlon Pierce, Sonja Herres-Pawlis and Alexander Hoffmann
sandra.gesing@nd.edu

HPCS 2015
20 July 2015

UNIVERSITY OF NOTRE DAME
CRC CENTER FOR RESEARCH COMPUTING

Science Gateways

"A Science Gateway is a community-developed set of tools, applications, and data that is integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community."

TeraGrid/XSEDE

Science Gateways

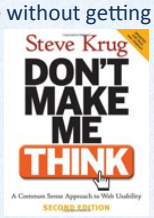
Science Gateways

Goals of Science Gateways

Usability

"After all, usability really just means that making sure that something works well: that a person … can use the thing - whether it's a Web site, a fighter jet, or a revolving door - for its intended purpose without getting hopelessly frustrated."

(Steve Krug in "Don't make me think!: A Common Sense Approach to Web Usability", 2005)

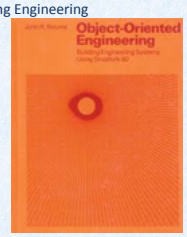


Goals of Science Gateways

Reusability

"The key to productivity is reusability. The easiest way to produce code is obviously to have it already!"

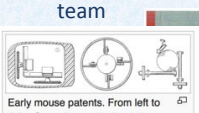
(John R. Bourne in "Object-oriented Engineering: Building Engineering Systems Using Smalltalk-80", 1992)



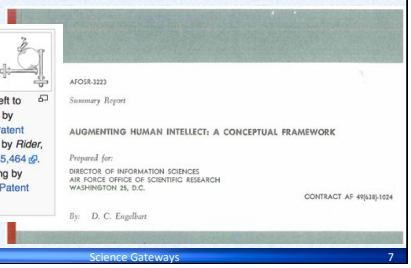
Usability - History

1946 - Invention of a trackball by Ralph Benjamin (patented 1947 but kept as military secret)

1962 - Prototype of a mouse by Douglas Engelbart's team



Early mouse patents. From left to right: Opposing track wheels by Engelbart, Nov. 1970, U.S. Patent 3,541,541. Ball and wheel by *Rider*, Sept. 1974, U.S. Patent 3,835,464. Ball and two rollers with spring by Opocensky, Oct. 1976, U.S. Patent 3,987,685



AFOSR-3223

Summary Report

AUGMENTING HUMAN INTELLECT: A CONCEPTUAL FRAMEWORK

Prepared for:

DIRECTOR OF INFORMATION SCIENCES
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
WASHINGTON 25, D.C.

CONTRACT AF 49(638)-1024

By: D. C. Engelbart

Usability - History

A. GENERAL

By "augmenting human intellect" we mean increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems. Increased capability in this respect is taken to mean a mixture of the following: more-rapid comprehension, better comprehension, the possibility of gaining a useful degree of comprehension in a situation that previously was too complex, speedier solutions, better solutions, and the possibility of finding solutions to problems that before seemed insoluble. And by "complex situations" we include the professional problems of diplomats, executives, social scientists, life scientists, physical scientists, attorneys, designers--whether the problem situation exists for twenty minutes or twenty years. We do not speak of isolated clever tricks that help in particular situations. We refer to a way of life in an integrated domain where hunches, cut-and-try, intangibles, and the human "feel for a situation" usefully co-exist with powerful concepts, streamlined terminology and notation, sophisticated methods, and high-powered electronic aids.

Usability - History

1969 - Alan Kay's Dynabook

1970 - Developments by Apple

Usability - History

1969 - Alan Kay's Dynabook

1970 - Developments by Apple

Usability - History

1980s - First science gateways

National Center for Biotechnology Information (NCBI) BLAST server, search results sent by email

1989 - World Wide Web developed by CERN

1990 - First web browser by Sir Tim Berners-Lee

WorldWideWeb, renamed to Nexus

MOSAIC evolved 1992

1992 - ISO 9241-2:1992 Ergonomic requirements for office work with visual display terminals (VDTs)

2004 - Announcement of TeraGrid's Science Gateway Program

TeraGrid – Scientific Profile

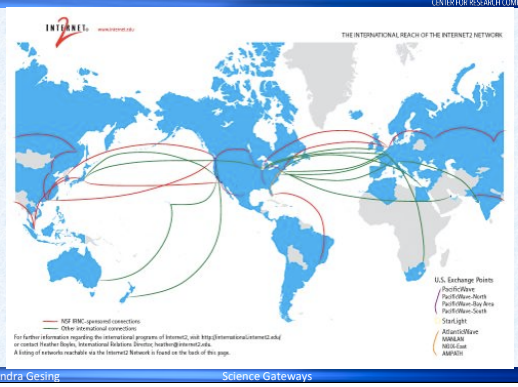
- Start 2001
- Operated 2004 – 2011
- roughly 4,000 users at over 200 universities
- XSEDE is the successor since 2011

Allocated (%)	Used (%)	Scientific Discipline
19	23	Molecular Biosciences
17	23	Physics
14	10	Astronomical Sciences
12	21	Chemistry
10	4	Materials Research
8	6	Chemical, Thermal Systems
7	7	Atmospheric Sciences
3	2	Advanced Scientific Computing
2	0.5	Earth Sciences
2	0.5	Biological and Critical Systems
1	0.5	Ocean Sciences
1	0.5	Cross-Disciplinary Activities
1	0.5	Computer and Computation Research
0.5	0.25	Integrative Biology and Neuroscience
0.5	0.25	Mechanical and Structural Systems
0.5	0.25	Mathematical Sciences
0.5	0.25	Electrical and Communication Systems
0.5	0.25	Design and Manufacturing Systems
0.5	0.25	Environmental Biology

User Profile

Generally not IT specialists…

Internet2 – Global International Reach



Distributed Infrastructures

- Definition Grid (Ian Foster, 1998)

„A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.“

- Definition Cloud (Sam Johnston, 2008)

„The Cloud is what The Grid could have been.“

  - Virtualization
  - Services
    - Infrastructure as a Service (IaaS)
    - Platform as a Service (PaaS)
    - Software as a Service (SaaS)

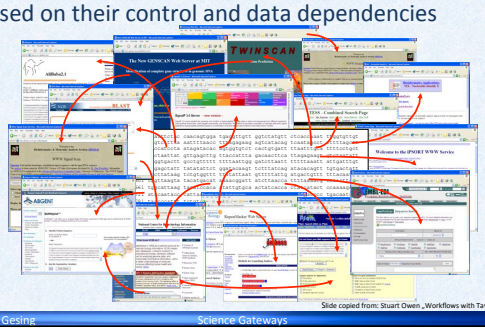
Distributed Infrastructures

- National and international infrastructures (EGI, PRACE, XSEDE)
- UNICORE, HTCondor, Globus Toolkit, gLite
- Docker, OpenStack, Amazon EC2, Windows Azure

- Mixed computing paradigms
  - Grid of federated clusters (NGI-DE)
  - Grid of federated clouds (EGI, CERN-openlab)
  - Grid over cloud (StratusLab)
  - Cloud over grid (WNoDeS (Worker Nodes on Demand Services))

Workflows

A sequence of connected steps in a defined order based on their control and data dependencies



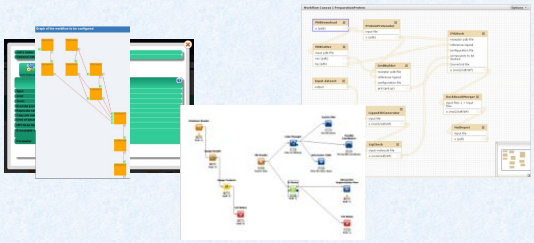
Workflow Systems

- Different workflow concepts
- Different workflow languages
- Different workflow constructs

OGSA-DAI, SUSE, TRIANA, Galaxy, Kepler, Taverna, UNICORE, Workflow 4Ever, KNIME

Workflow Editors

- Different technologies (workbenches, web-based)
- Different look-and-feel



Distributed Data Management

- Amazon Simple Storage Service (Amazon S3)
- Google File System (GFS)
- XtreemFS
- dCache
- iRODS

Figure copied from „iRODS Overview"

State-of-the-Art

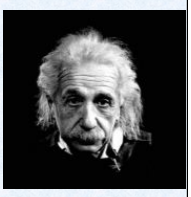
- Data-intensive and compute-intensive problems
- High-speed networks available
- Sophisticated tools and methods available
- DCIs (Distributed Computing Infrastructures) available
- Workflow systems available
- Distributed data management available

How do researchers use the tools and distributed environments on a large scale?

Science Gateways

Why Science Gateways?

- Modeling of data and simulations on a large scale with DCIs
- Usability of computational tools, applications, and DCIs is limited
- Scientists are generally not IT specialists



➔ Need for intuitive and self-explanatory GUIs

Solution Science Gateways

Characteristics of science gateway

- Convenient access to data and tools
- Single point of entry for multiple organizations
- No firewall issues
- Diverse frameworks and APIs

Frameworks and APIs

Re-inventing is not always necessary...

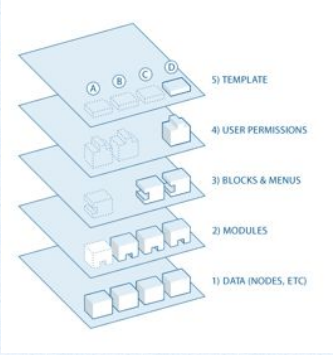
Frameworks and APIs

... but the model should fit to the demands of the community

Science Gateway Technologies

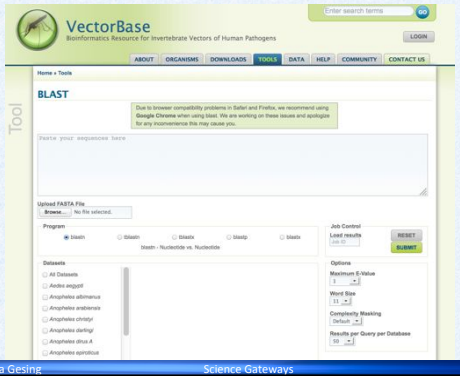
- Content management systems (Drupal)
- Libraries for implementation (Django)
- Portal frameworks (Liferay)
- Science gateway frameworks (WS-PGRADE, Galaxy)
  - Static layout
  - Layout extendable
  - Workflow-enabled
- APIs for implementation (Apache Airavata, Agave)

Drupal



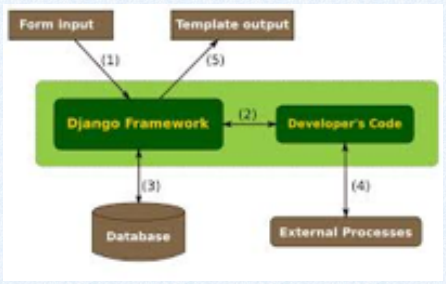
VectorBase - Example for Drupal

VectorBase

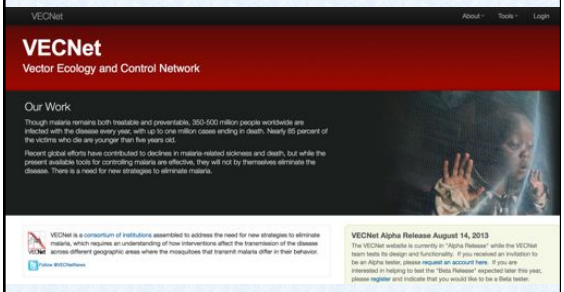


VectorBase

Django



VecNet – Example for Django



VecNet

VECNet enables national malaria control managers, researchers, product developers, funding bodies and policy makers to ask questions such as: *'What is the intensity and type of intervention/s required to achieve one malaria death per 100,000 in this particular population?'* and *'What is the impact on malaria transmission of a new tool that potentially kills a mosquito any time it takes a sugar meal, seeks a blood meal or lays eggs?'*

To find answers to these questions, VECNet is developing three resources: the **Digital Library**, the **Data Warehouse Browser** and a **Modeling Platform**. These tools provide users with access both to data on malaria transmission and to modeling software to create simulations of various scenarios. The simulations use geospatially specific data to analyze the potential of different combinations of control interventions to reduce malaria transmissions in a given area.

Digital Library
The Digital Library assembles all published and unpublished data on malaria vectors. This extensive database enables the analysis of transmission risk as a function of vector ecology and behavior via the Modeling Platform.

Data Warehouse Browser
The Data Warehouse Browser is an incentive-based platform for data sharing, and enables easy-to-use, secure access to data for use with any of the Modeling Platform tools. It presents research data in ways that allow model simulations with specific data in geographically defined areas. Users can access all existing data, including their own model input and output data.

VecNet

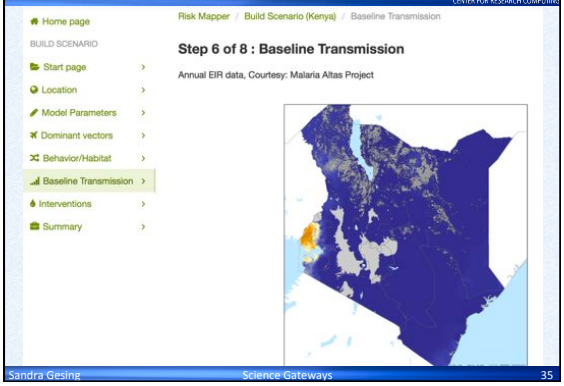
Transmission Simulator
Researchers use their data as input to model the sensitivity of transmission to changes in the behaviors of vectors resulting from responses to interventions or changes in the environment/ecology.

Risk Mapper
Risk Mapper analyzes the impact of particular interventions on malaria. National malaria control programs can use it to compare the distribution of vector control interventions against the distribution of specific vector species to determine the appropriateness of the intervention.

Product Impact Evaluator (PIE)
Investors use PIE to estimate the value of new control tools. With PIE, product developers estimate the effect of candidate tools on malaria transmission and can then develop and refine Target Product Profiles to achieve a desired level of impact.

Computational Intervention portFolio EvaluatoR (CIFER)
CIFER is an amalgamation of the outputs of Transmission Simulator, Risk Mapper, and PIE, combining the estimates of contributions from individual vector species, individual geographies, and individual interventions to overall malaria transmission. Policymakers can refine the suite of tools needed to achieve malaria eradication by analyzing how interventions affect the transmission of the disease and, as importantly, where interventions will not be able to achieve effective control and elimination.

VecNet



Liferay

Portal framework
- Authentication (e.g., OpenSSO, CAS)
- Authorization
- Standards compliant
  - JSR168/286
  - Web services
  - Web 2.0 websites
- Web Publishing and Shared Workspaces
- Collaboration
- Social Networking

WS-PGRADE

User Interface
WS-PGRADE
Liferay

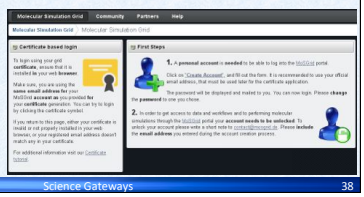
Workflow storage | Application repository | Information system
Workflow engine | Submitters | Logging

High-Level Middleware Service Layer
gUSE

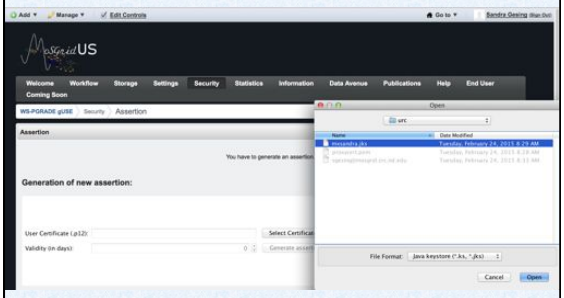
DCI Resources Middleware Layer

MoSGrid as WS-PGRADE Example

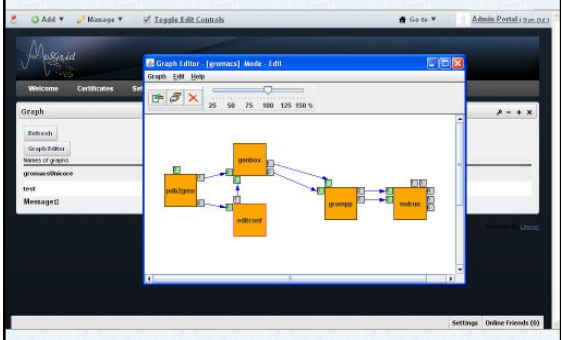
Molecular Simulation Grid
- Science gateway integrated with underlying compute and data management infrastructure
- Distributed workflow management
- Data repository
- Metadata management



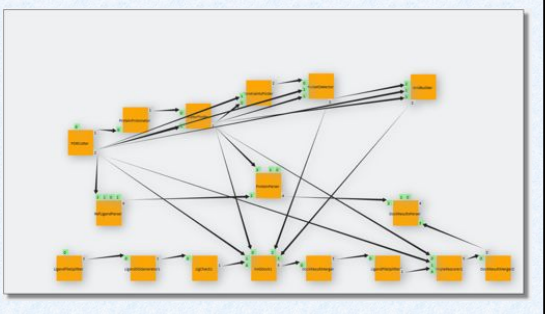
MoSGrid



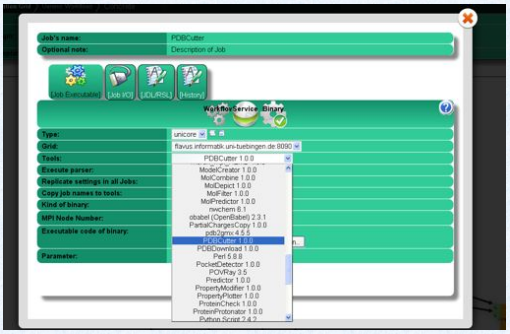
MoSGrid



MoSGrid



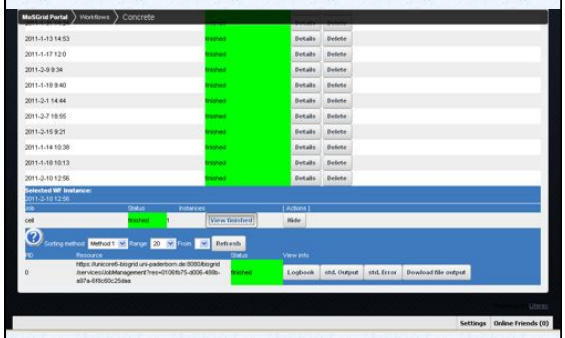
MoSGrid



MoSGrid

MoSGrid

MoSGrid



MoSGrid

Molecular Dynamics
- Study and simulation of molecular motion

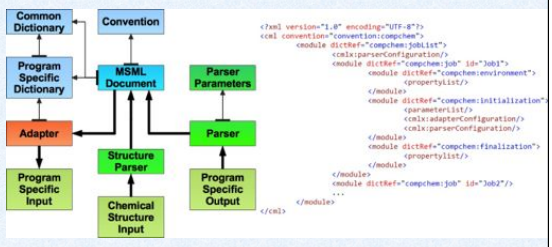
Quantum Chemistry
- Study and simulation of molecular electronic behavior relative to their chemical reactivity

Docking
- Main focus on evaluation of ligand-receptor interactions (e.g., for drug design)

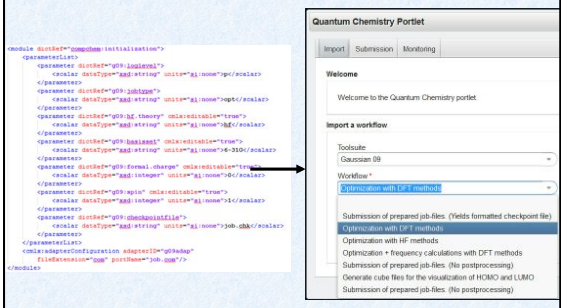
MoSGrid - Metadata

- Molecular Simulation Markup Language (MSML)
- CML compliant
- Template for each and every workflow
  - Molecular input
  - Domain specific tools
  - Job configuration
  - Optimized structures, trajectories, energies, …
- Semantic search (Apache Lucene)

MoSGrid - Metadata

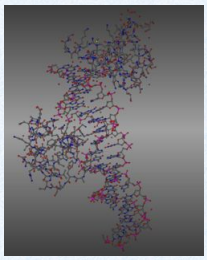


MoSGrid - Metadata



MoSGrid – Visualization

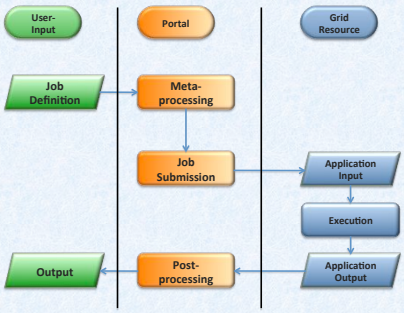
Testing of ChemDoodle and MolCAD



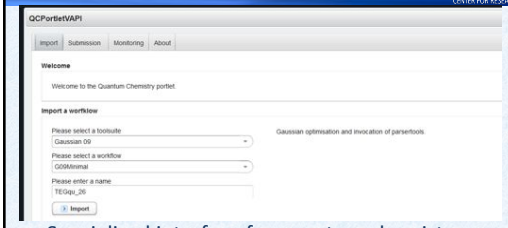
web.chemdoodle.com

molcad.de

MoSGrid – Basic Workflow



MoSGrid – QC Portlet

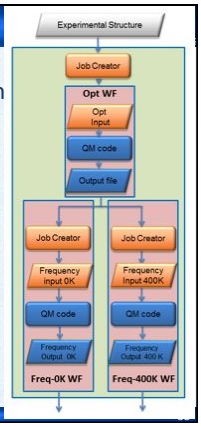


- Specialised interface for quantum chemistry software (Gaussian, NWChem, ORCA)
- Basic workflows
- Easy Generation or Uploading of Input Files
- Parsing of result files

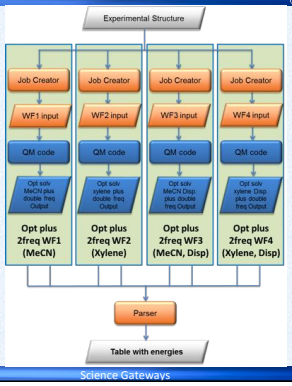
MoSGrid - Metaworkflows

- Fundamental step: optimisation
- Frequency calculation at 0 K and at 400 K (polymerisation temperature) for the same molecule

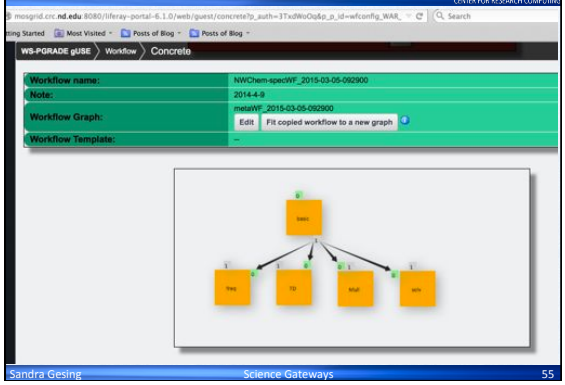
→ Small workflow with 3 atomic workflows (opt and freq)



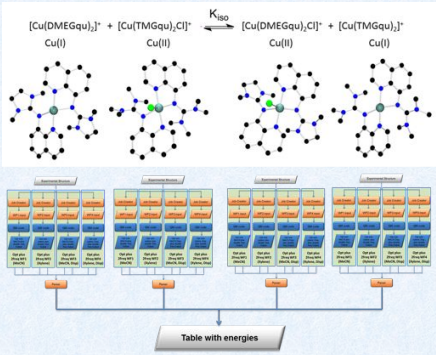
MoSGrid - Metaworkflows



MoSGrid - Metaworkflows

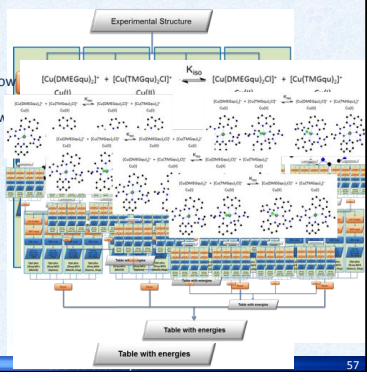


MoSGrid

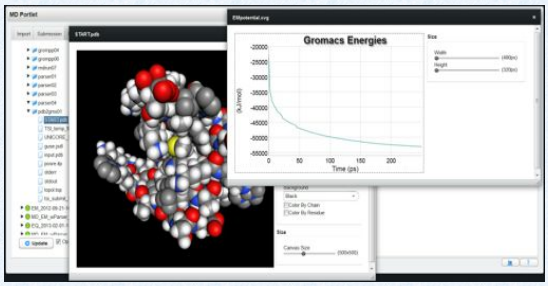


MoSGrid – Multiple Layers

First layer: atomic workflow
Second layer: metaworkflow
Third layer: meta-metaworkflow
Fourth layer: (meta)^3-workflow
Fifth layer: (meta)^4-workflow

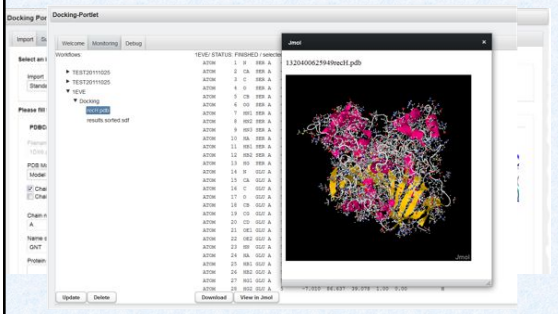


MoSGrid – MD Portlet



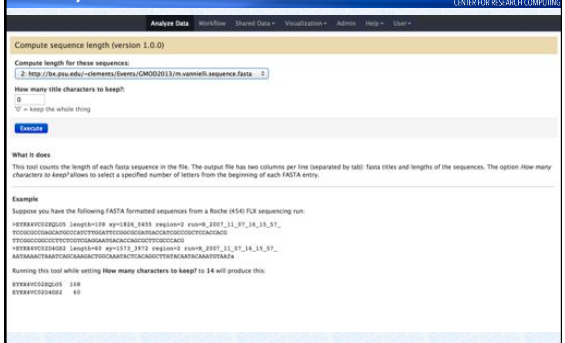
MoSGrid – Docking Portlet

MoSGrid – Docking Portlet



Galaxy

Galaxy

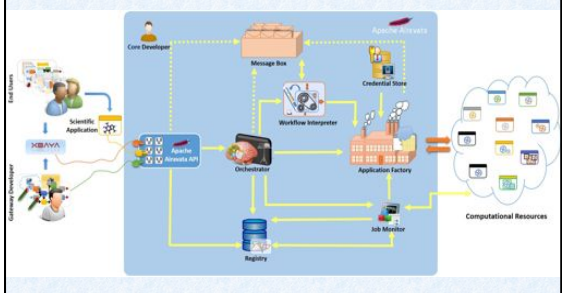


RNA-Seq Analysis

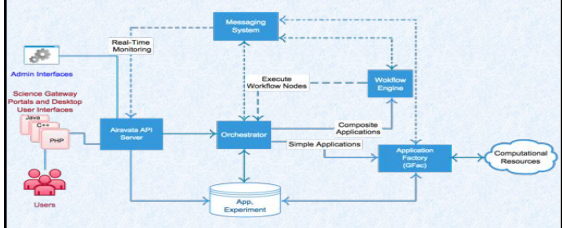
Apache Airavata

- Airavata is a general purpose distributed system software framework build on micro-service and component based architecture principles
- Airavata provides capabilities to compose, manage, execute and monitor large scale applications and workflows on distributed computing resources
- Airavata supports executions on local clusters, national grids, academic and commercial clouds
- Airavata is inherently multi-tenanted

Apache Airavata

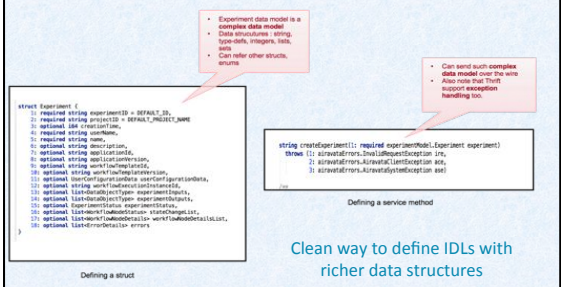


Apache Airavata



- External clients interact with Airavata API (based on Apache Thrift)
- Internally, components interact with each other through Component Programming Interfaces (thrift-based CPIs)

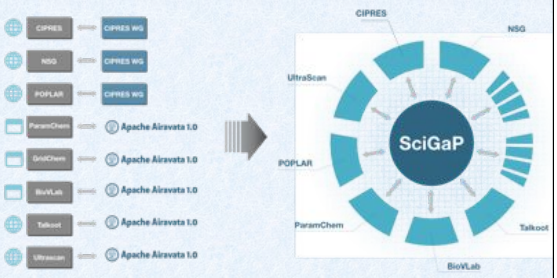
Apache Airavata



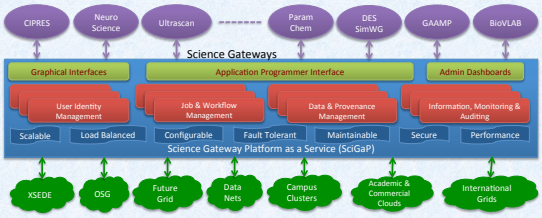
Clean way to define IDLs with richer data structures

SciGap – Example for Apache Airavata

Science Gateway Platform as a Service



Apache Airavata



Apache Airavata

Community Hangout

Mailing lists:

architecture@airavata.apache.org

dev@airavata.apache.org

users@airavata.apache.org

Extend Airavata from your project or extend your project from Airavata

Agave API

Agave is a *Science-as-a-Service* web API platform

Run scientific codes

- your own or community provided codes

...on HPC, HTC, or cloud resources

- your own, shared, or commercial systems

...and manage your data

- reliable, multi-protocol, async data movement

...from the web

- webhooks, rest, json, cors, oauth2

...and remember how you did it

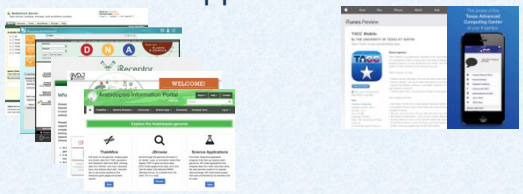
- deep provenance, history, and reproducibility built in

Agave API

- Multitenant
- Hosted identity management
- Supports multiple IdP
- OAuth2/OIDC server
- API Management
- Hosted or on premise
- Vertical SSO
- Analytics and reporting
- Developer resources
- Multiple SDK & CLI
- Reference gateway
- White labeled
- 100% open source

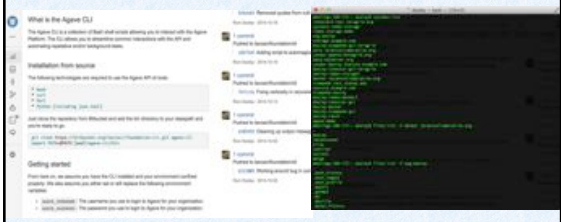
Agave API

Used to power web & mobile applications



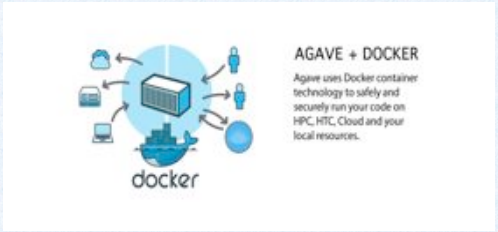
Agave API

Used to extend existing processes



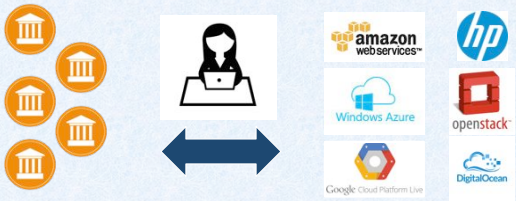
Agave API

(Re)Introducing the Micro App Paradigm

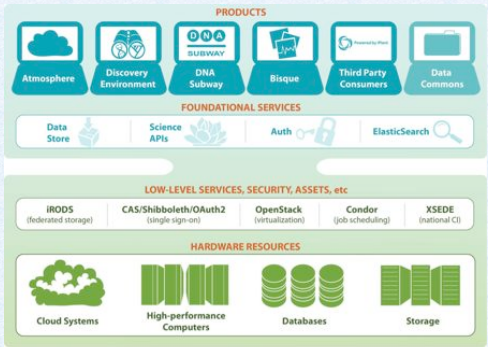


Agave API

Agave Delivers Process-as-a-Service

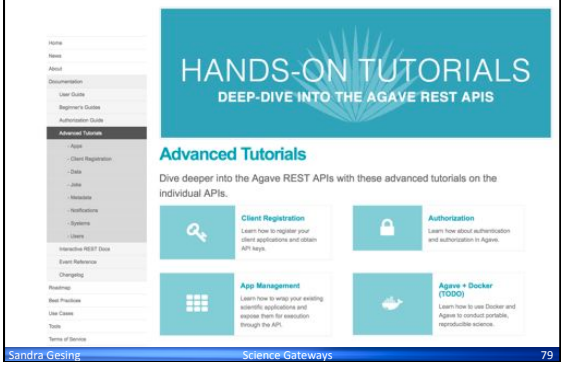


Agave API

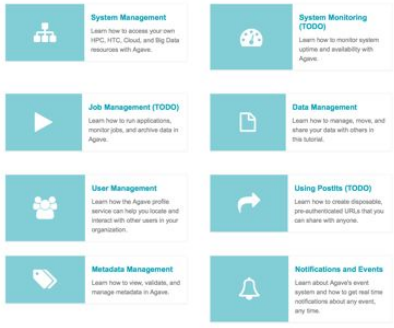


iPlant – Example for Agave API

Agave API - Tutorials



Agave API - Tutorials



Collaboration on Science Gateways

Crucial Topics
- Close collaboration with user communities
- Knowledge about available technical solutions

Sounds easy but…
- Requirements of user communities often not so clear
- Technologies sometimes still under development for certain building blocks

➔ Slow uptake of solutions

➔ Larger effort for creating science gateways

Science Gateways

A new era…
- Novel developments of web-based agile frameworks
- Infrastructure providers report that science gateways are more used than commandlines

http://www.iplantcollaborative.org

Science Gateways

A new era…
- Novel developments of web-based agile frameworks
- Infrastructure providers report that science gateways are more used than commandlines

But also always new challenges…
- Novel infrastructures
- Novel data sources like telescopes such as the Square Kilometre Array (SKA) (will create data rates in exa-scale size)

➔ Support of developers necessary

Science Gateway Institute

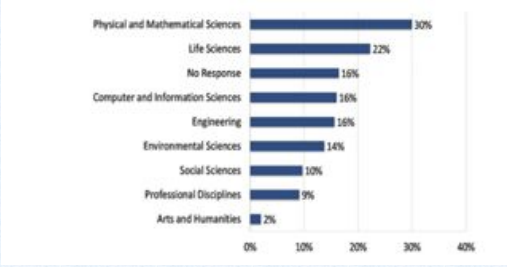
2012 NSF Software Institute conceptualization award
2015 NSF Software Institute implementation proposal ($15M)

Services
- Incubator
- Developer support team
- Gateway framework directory
- Workforce development

http://sciencegateways.org

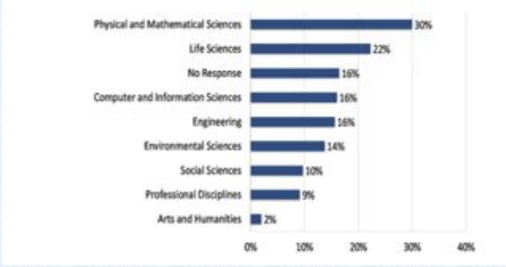
Science Gateway Survey 2014

- 29,000-person survey
- 4957 responses from across domains



Science Gateway Survey 2014

- 29,000-person survey
- 4957 responses from across domains



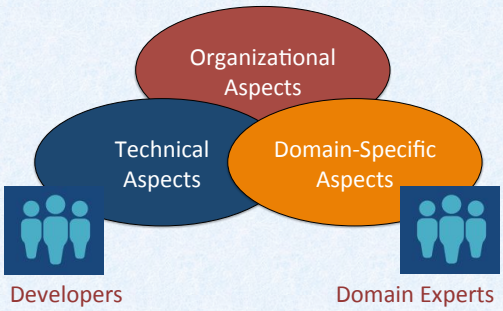
Science Gateway Survey 2014

What services would be helpful?

Proposed Service	% Interest
Evaluation, impact analysis, website analytics	72%
Adapting technologies	67%
Web/Visual/graphic design	67%
Choosing technologies	66%
Usability Services	66%
Visualization	65%
Developing open-source software	64%
Support for education	64%
Community engagement mechanisms	62%
Keeping your project running	62%
Legal perspectives	61%
Managing data	60%
Computational resources	59%
Mobile technology	59%
Database structure, optimization, and query expertise	59%
Data mining and analysis	58%
Cybersecurity consultation	57%
Website construction	57%
Software engineering process consultation	53%
Source code review and/or audit	51%
High-bandwidth networks	45%
Scientific instruments or data streams	44%
Management aspects of a project	38%

New Science Gateways - Checklist

DISCUSSION



New Science Gateways - Checklist

Domain-specific aspects:

- Goal, target area and target users
- Visions/demands on the layout
- Priorities of features and options, e.g., a list from must-have to great-to-have options
- Integration of existing applications or development of applications
- Technologies of the applications
- Visualization
- Security demands
- Workflows

New Science Gateways - Checklist

Organizational aspects:

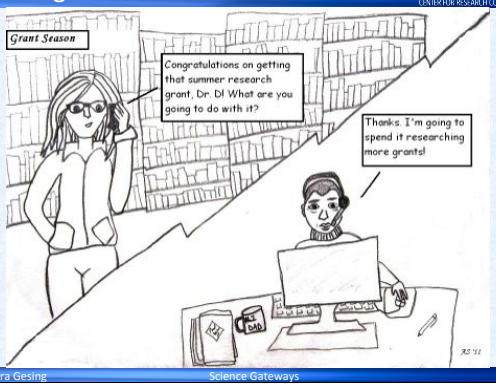
- Time constraints for the development, agreement on a (maybe even rough) project plan with milestones
- Agreement on alpha- or beta-tester
- Regular meetings

New Science Gateways - Checklist

Technical aspects:

- Experience with existing frameworks and programming languages
- Available infrastructure including security infrastructure and resources
- Available support of suitable technologies
- Scalability of suitable technologies
- Effort for extending existing technologies compared to novel developments
- Synergy effects with other science gateway projects

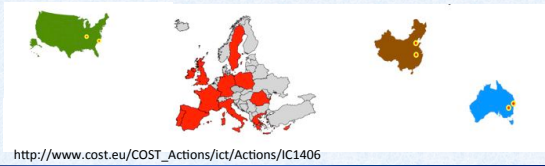
Funding



EU COST Action cHiPSet (IC1406)

cHiPSet – High Performance Modeling and Simulation for Big Data Applications

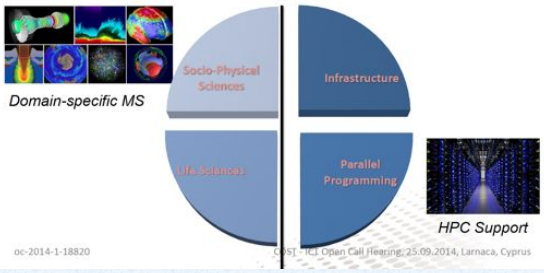
- April 2015 – April 2019
- 15 countries - 12 COST, 3 non-COST (US, China, Australia)
- 37 reseach organizations/companies (31 COST, 6 non-COST)



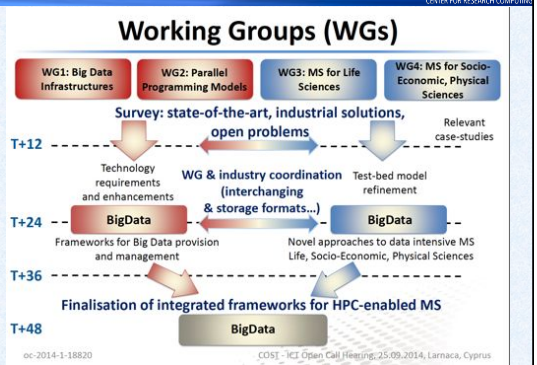
http://www.cost.eu/COST_Actions/ict/Actions/IC1406

EU COST Action cHiPSet

Structure & Coordinate Research on **HPC**-enabled **Modelling and Simulation (MS) for Big Data** in Europe.



EU COST Action cHiPSet



cHiPSet - Collaboration

Projects already declared interest for collaboration

- NESUS (Network for Sustainable Ultrascale Computing)
- KEYSTONE (Semantic keyword-based search on structured data sources)
- AAPELE (Algorithms, Architectures and Platforms for Enhanced Living Environment)

And maybe YOU?

Acknowledgements

Co-authors
Rion Dooley
Richard Grunzke
Sonja Herres-Pawlis
Alexander Hoffmann
Jens Krüger
Marlon Pierce

WS-PGRADE
Akos Balasko
Peter Kacsuk
Istvan Marton
Gabor Terstyanszki

XSEDE, especially
Tom Maiden
Suresh Marru
Marlon Pierce
Tabitha Samuel
Raminder Singh
Nancy Wilkins-Diehr

Information on Science Gateways

- IEEE Technical Area on Science Gateways
  http://ieeesciencegateways.org
- Science Gateway Institute
  http://sciencegateways.org
- XSEDE Science Gateways
  https://www.xsede.org/gateways-overview
- Science Gateway Workshops (partner workshops)
  Europe: IWSG - http://iwsg.info/
  USA: GCE - http://sciencegateways.org
  Australia: IWSG-A - http://iwsg.info

sandra.gesing@nd.edu