SCIENCE IMPACT OF SUSTAINED CYBERINFRASTRUCTURE:

The Pegasus Workflow Management System: Evolution and Impact

Ewa Deelman, Ph.D.

University of Southern California,

Information Sciences Institute

Exploring a Scientific Question

Scientific Problem

Earth Science, Astronomy, Neuroinformatics, Bioinformatics, etc.

Analytical Formulation

$$
\begin{aligned}
U\frac{\mathrm{d}^2}{\mathrm{d}x^2}U^{-1}g &= \left(U\frac{\mathrm{d}}{\mathrm{d}x}U^{-1}\right)\times\left(U\frac{\mathrm{d}}{\mathrm{d}x}U^{-1}\right)g \\
&= \frac{\mathrm{d}}{\mathrm{d}\psi}\left[g'\psi' + \frac{1}{2}g\frac{\psi''}{\psi'}\right]\cdot\psi' + \frac{1}{2}\left[g'\psi' + \frac{1}{2}g\frac{\psi''}{\psi'}\right]\times\frac{\psi''}{\psi'} \\
&= g''\psi'^2 + 2g'\psi'' + \frac{1}{2}g\times\left[\frac{\psi'''}{\psi'} + \frac{\psi''^2}{\psi'^2}\right]
\end{aligned}
$$

Computational Scripts

Shell scripts, Python, Matlab, etc.

Automation

Distributed Computing

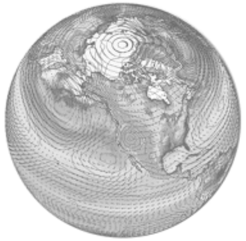
Clusters, HPC, Clouds, etc.

Monitoring and Debug

Fault-tolerance, Provenance, etc.

Scientific Result

Models, Quality Control, Image Analysis, etc.



Exploring a Scientific Question

Scientific Problem

Earth Science, Astronomy, Neuroinformatics, Bioinformatics, etc.

Analytical Formulation

$$
\begin{aligned}
U\frac{d^2}{dx^2}U^{-1}g &= \left(U\frac{d}{dx}U^{-1}\right)\times\left(U\frac{d}{dx}U^{-1}\right)g \\
&= \frac{d}{d\psi}\left[g'\psi' + \frac{1}{2}g\frac{\psi''}{\psi'}\right]\cdot\psi' + \frac{1}{2}\left[g'\psi' + \frac{1}{2}g\frac{\psi''}{\psi'}\right]\times\frac{\psi''}{\psi'} \\
&= g''\psi'^2 + 2g'\psi'' + \frac{1}{2}g\times\left[\frac{\psi'''}{\psi'} + \frac{\psi''^2}{\psi'^2}\right]
\end{aligned}
$$

Computational Scripts

Shell scripts, Python, Matlab, etc.

Distributed Computing

Clusters, HPC, Cloud, etc.

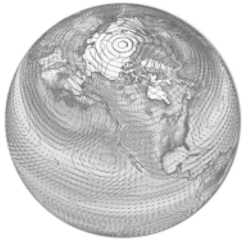
Automation

Monitoring and Debug

Fault-tolerance, Provenance, etc.

Scientific Result

Models, Quality Control, Image Analysis, etc.

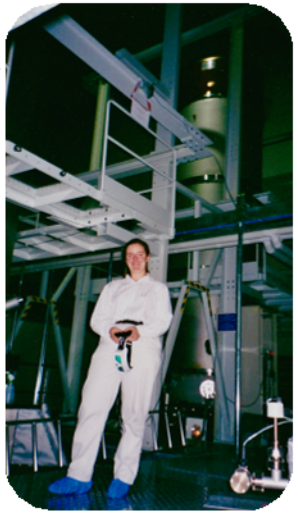


Pegasus: Grounding Research and Development

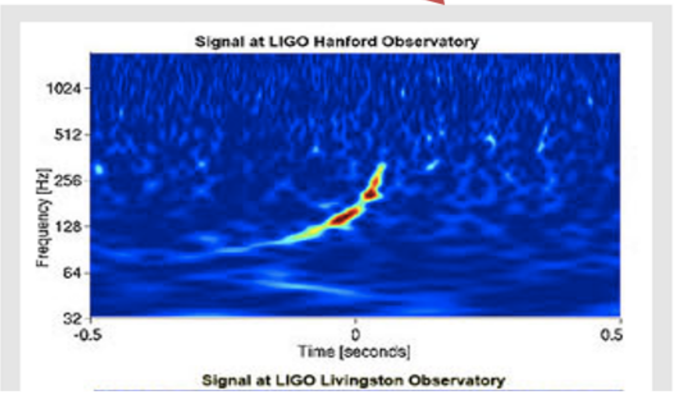
Nobel Prize

Working with LIGO (Laser-Interferometer Gravitational Wave Observatory)

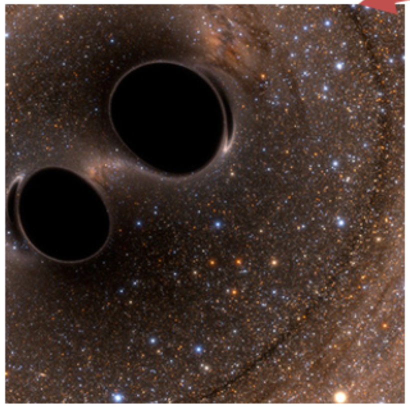
2001 → 2002 → 2003 → 2004 → 2005 → 2006 → 2007 → 2008 → 2009 → 2010 → 2011 → 2012 → 2013 → 2014 → 2015 → 2016 → 2017



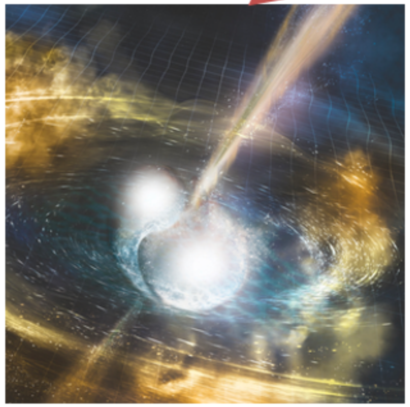
First Pegasus prototype



Blind injection detection



First detection of black hole collision



Multi-messenger neutron star merger observation

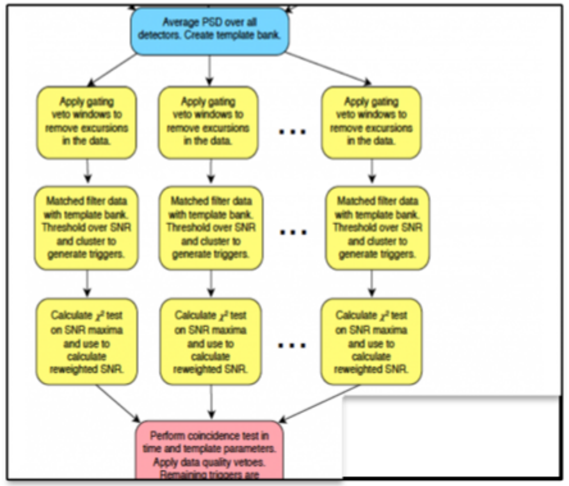
First GW detection: **~ 21K workflows with ~ 107M tasks**

Science workflow:
measure the statistical significance of data needed for discovery

Automated by Pegasus
execution of tasks and data access

Distributed Power
LIGO, Open Science Grid, XSEDE, Blue Waters

Science



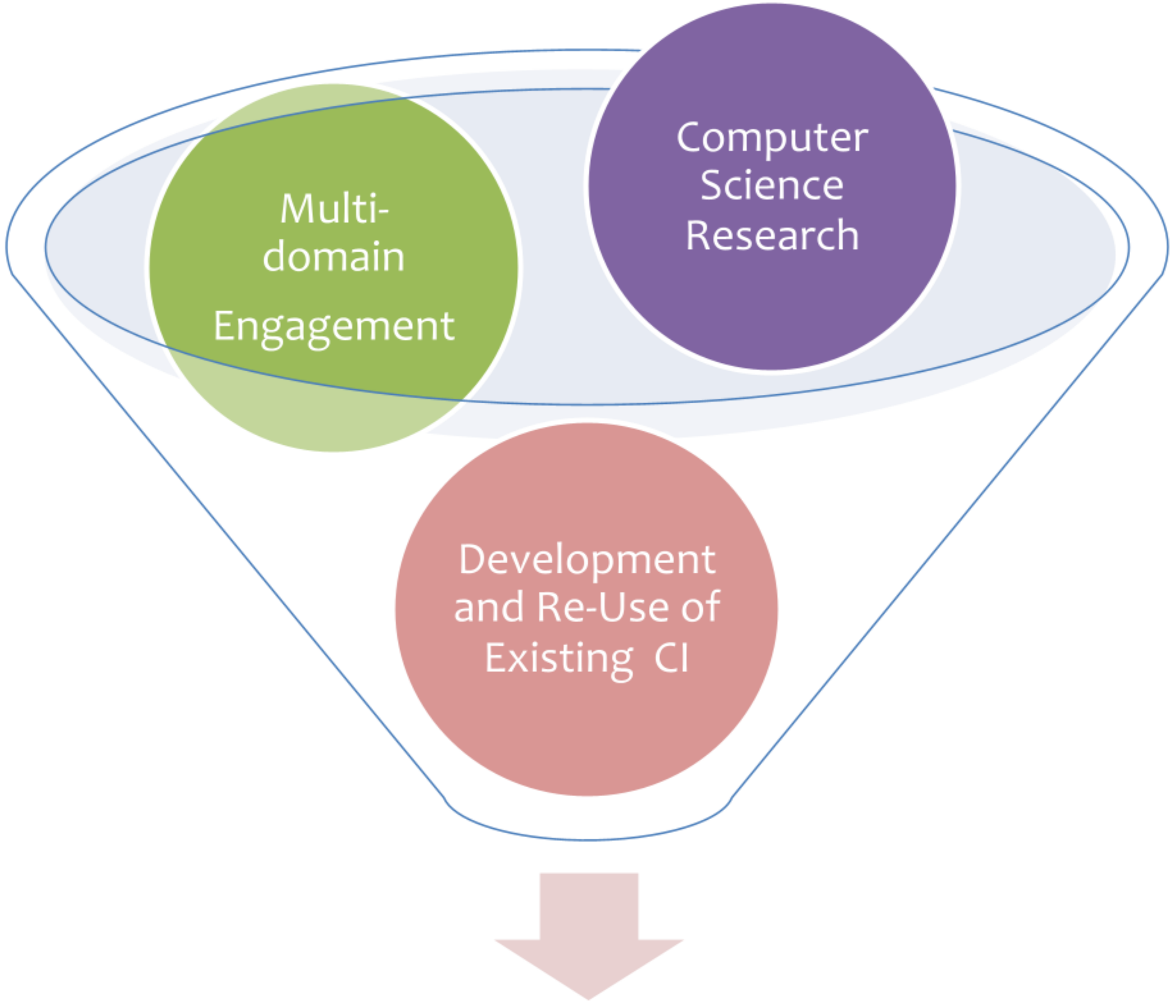
Cyberinfrastructure (CI) Middleware

pegasus

CI Platform

What does it take to build and sustain Pegasus?

Cyberinfrastructure (CI) =
computing systems
+ data storage systems
+ advanced instruments
+ data repositories
+ visualization environments
+ people
Connected by software
and high-performance networks
for research and breakthroughs not otherwise possible



Dependable Cyberinfrastructure

Takes time to build a team and expertise

2000 › 2001 › 2002 › 2003 › 2004 › 2005 › 2006 › 2007 › 2008 › 2009 › 2010 › 2011 › 2012 › 2013 › 2014 › 2015 › 2016 › 2017 › 2018 › 2019

Loïc Pottier

Ryan Tanaka

Patrycja Krawczuk

Front Row: Tu Mai Anh Do, Rajiv Mayani, Ryan Mitchell, Ragini Church, Ewa Deelman, Mukund Murrali
Back Row: Karan Vahi, Mats Rynge, George Papadimitriou, Rafael Ferreira da Silva

CS research

How did Pegasus Start?

Extend the concept of view materialization in DBs to distributed environments

GriPhyN
Data Intensive Science

The Virtual Data Grid (VDG) Model

- Data suppliers publish data to the Grid
- Users request raw or derived data from Grid, without needing to know
  - Where data is located
  - Whether data is stored or computed

NSF ITR: GriPhyN Project: Ian Foster (PI), Paul Avery, Carl Kesselman, Miron Livny, (co-Pis)

GriPhyN
Data Intensive Science

Virtual Data Scenario

- (LIGO) "Conduct a pulsar search on the data collected from Oct 16 2000 to Jan 1 2001"
- For each requested data value, need to
  - Understand the request
  - Determine if it is instantiated; if so, where; if not, how to compute it
  - Plan data movements and computations required to obtain all results
  - Execute this plan

How do you translate the Computer Science idea to the needs of science?

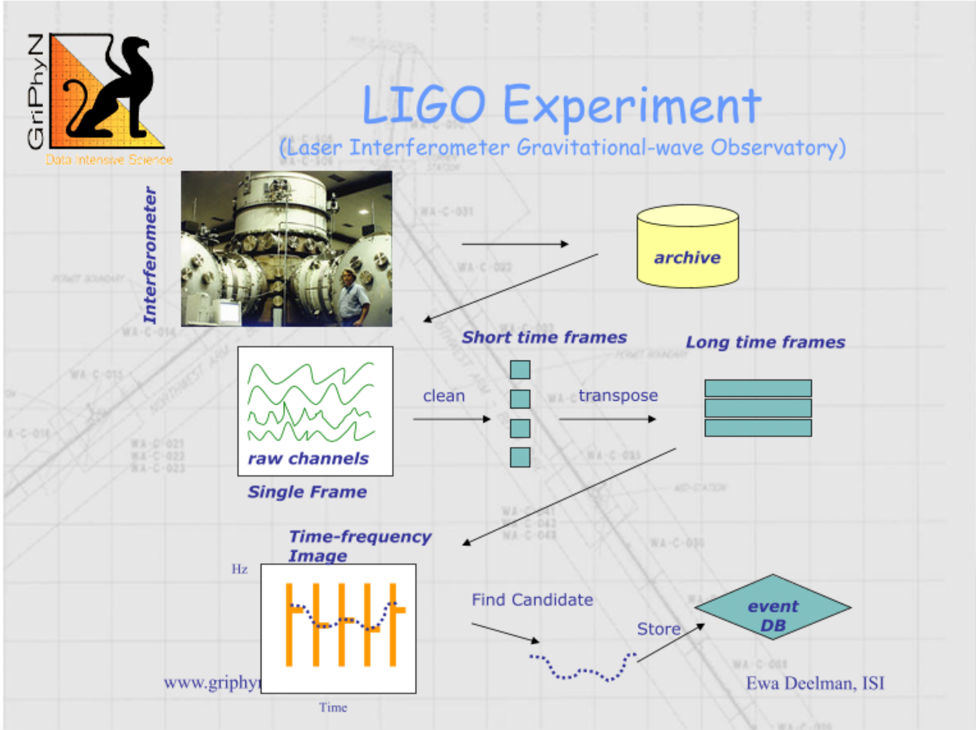
Circa. 2001

CS research

Challenge: How Translate a Science Request to an Actionable Plan?

Explore AI planning techniques

Work with Yolanda Gil and Jim Blythe



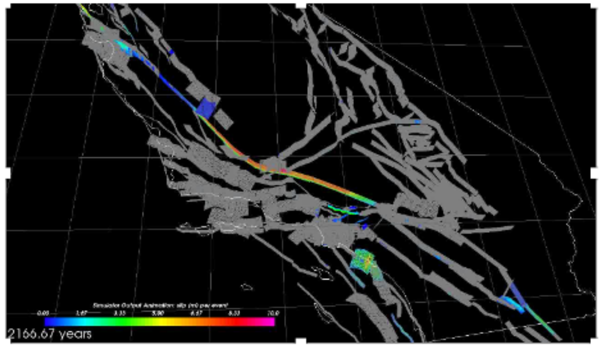
Lost in translation: high-level abstraction for this science domain

Found: new research direction: management of workflows in distributed environments

Challenges of Workflow Management

- Working with LIGO and other applications (astronomy, earthquake science), found common challenges:
  - Need to describe complex workflows in a simple way
  - Need to access distributed, heterogeneous data and resources
  - Need to deal with resources/software that change over time
- Our focus:
  - Separation between workflow description and workflow execution
  - Workflow planning and scheduling (scalability, performance)
  - Task execution (monitoring, fault tolerance, debugging)

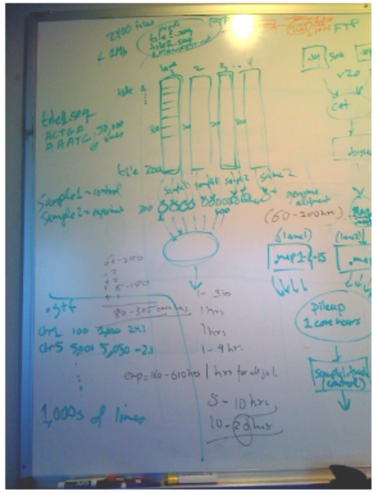
Sky mosaic, IPAC, Caltech



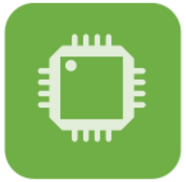
Earthquake simulation, SCEC, USC

Multi-domain

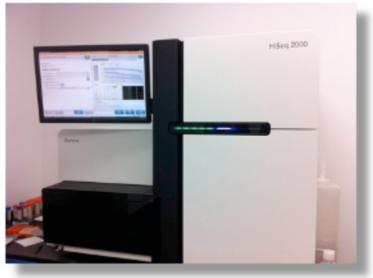
Typical local computational environment



Work Definition



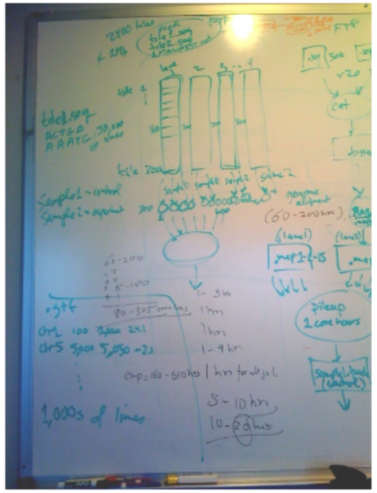
Local Resource



Local Data Storage

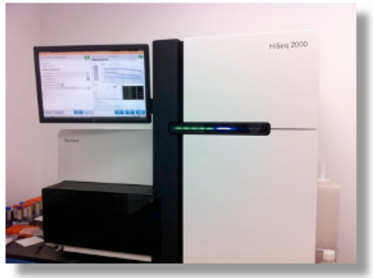
Multi-domain

Typical local computational environment



Work Definition

Local Resource



Local Data Storage

Blue Waters

Campus Cluster

XSEDE

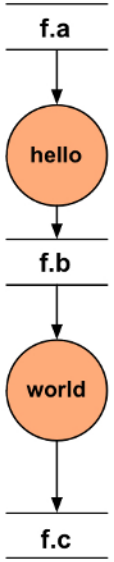
DOE Facilities

OSG

Chameleon

Amazon Cloud

To run Hello World on USC's HPC System



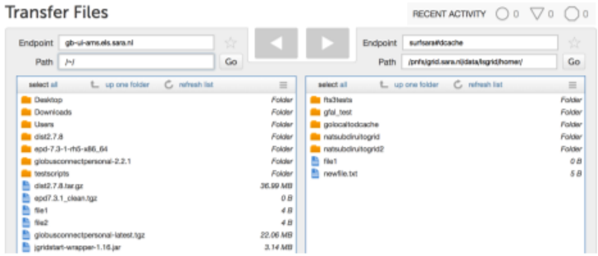
1. Login to System

```
localhost$ ssh -l deelman wrangler.tacc.utexas.edu
login1.wrangler$ emacs myjob.sub
```

2. Write submit script

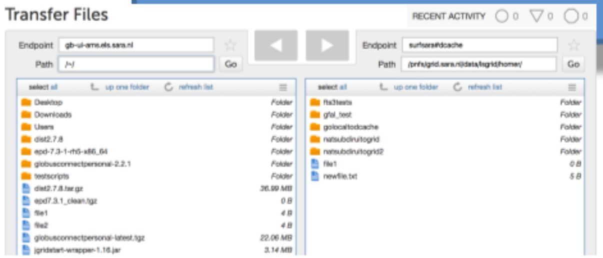
```
#/bin/bash
#SBATCH -J myjob
#SBATCH -o myjob.o%j
#SBATCH -e myjob.e%j
#SBATCH -p normal
#SBATCH -N 1
#SBATCH -n 1
#SBATCH -t 01:30:00
#SBATCH --mail-
user=deelman@gmail.com
#SBATCH --mail-type=all
#SBATCH -A myproject

mkdir $WORK/helloworld
cd $WORK/helloworld
cp $WORK/data/inputs/* .
~/hello
~/world
cp * $WORK/data/outputs/
~/my_output_files/
```

3. Find and bring in your input data



4. Submit script for execution

```
login1.wrangler$ squeue myjob.sub
```

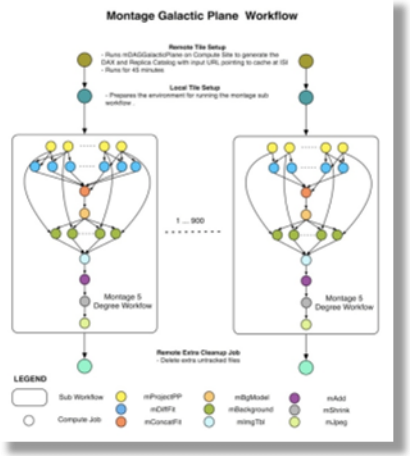


5. Stage out data for further analysis

What if the system goes down/gets decommissioned? What if the job crashed? What about running on multiple platforms?

Our Approach: Submit locally, Compute globally

Multi-domain



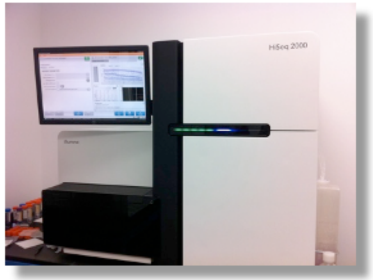
Work Definition

Local Resource

Workflow Management System

work

HTCondor
High Throughput Computing



Local Data Storage

data

Blue Waters

Campus Cluster

XSEDE

DOE Facilities

OSG

Chameleon

Amazon Cloud

Multi-domain

Pegasus Workflow Management System

- Operates at the level of files and individual applications
- Allows scientists to describe their computational processes (workflows) at a logical level
- Without including details of target heterogeneous CI (portability)
- Scalable to $O(10^6)$ tasks, TBs of data
- Captures provenance and supports reproducibility
- Includes monitoring and debugging tools

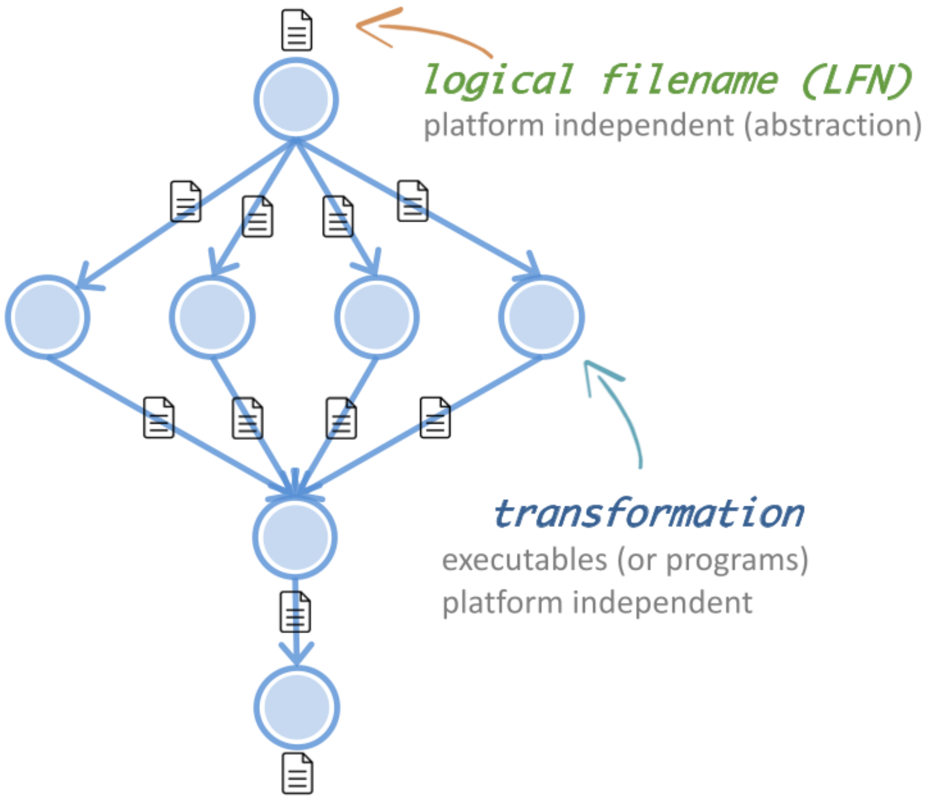
Composition in WINGS, Python, R, Java, Perl, Jupyter Notebook

hubzero

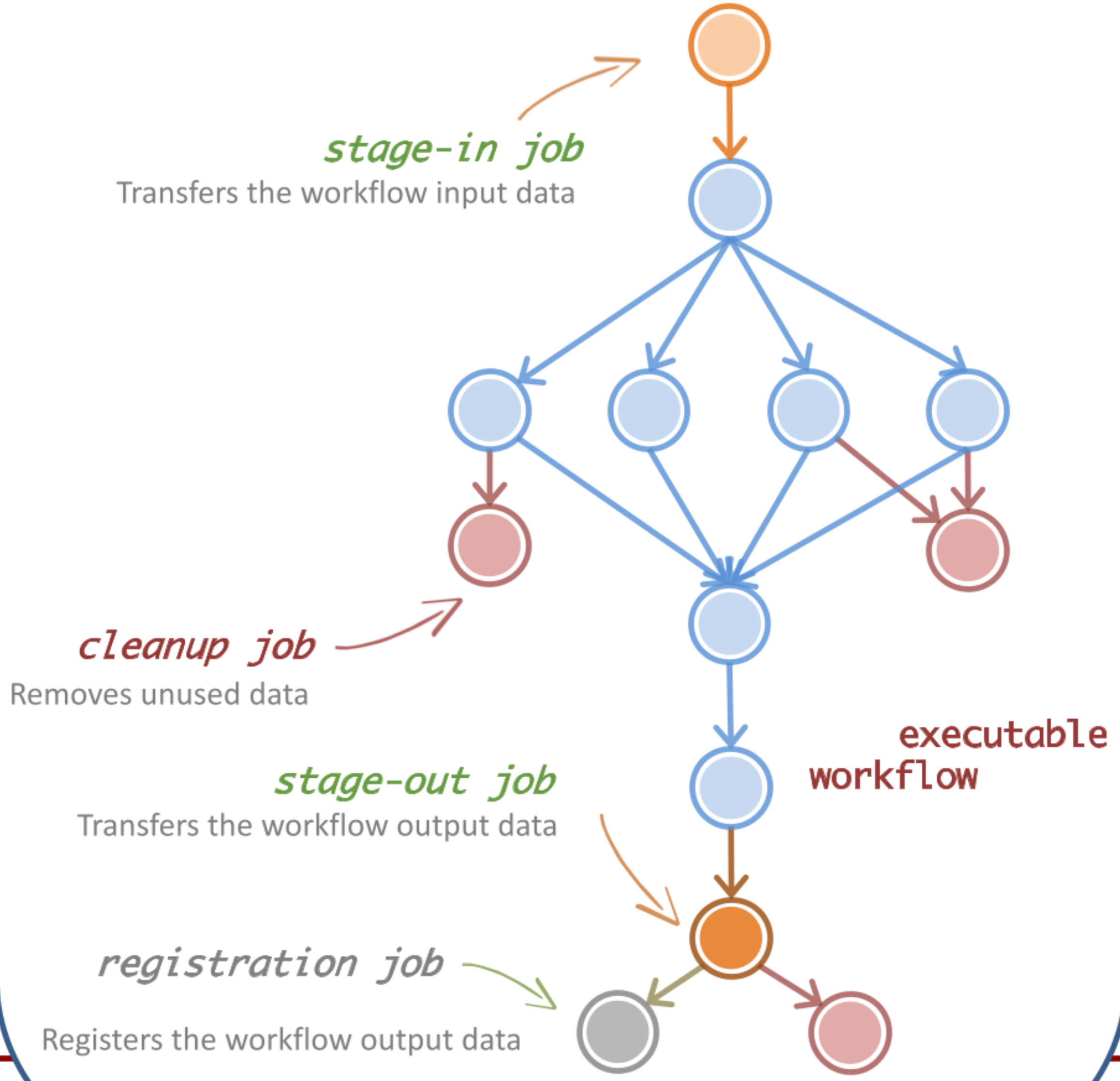
Abstract Workflow

Portable Description

Users do not worry about
low level execution details



Executable Workflow

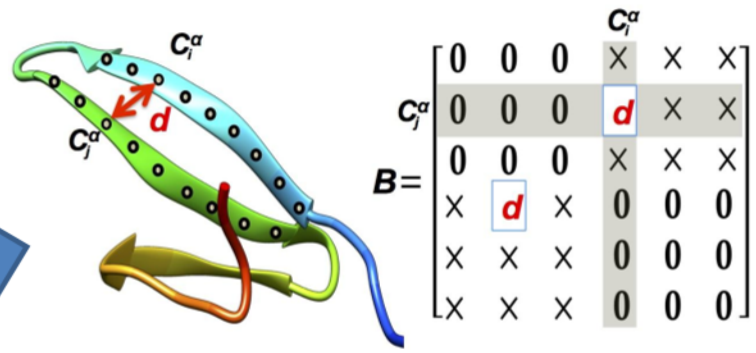


Http://cicoe-pilot.org

CS research

CS Principles Help in Cyberinfrastructure Development

- Structure workflows as **directed acyclic graphs (DAGs)**
  - Re-use of graph traversal algorithms, node clustering, pruning, other complex graph transformation
- Use hierarchical structures in DAGs
  - To achieve scalability, recursion, dynamic behavior
- Develop new algorithms:
  - Task clustering
  - Data placement
  - Data re-use
  - Resource usage estimation
  - Resource provisioning
  - **Insitu workflows**



New Direction:
In-memory coupling of simulation and analytics
Collaboration with U of Tennessee, Cornell, U. of New Mexico

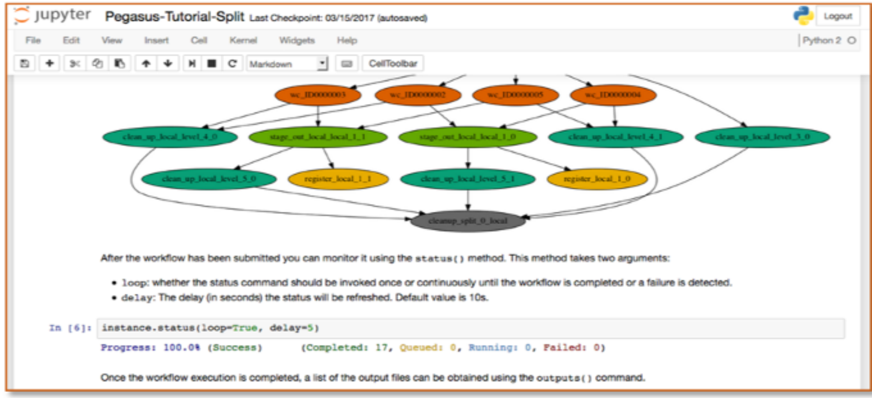
Image credit: Michela Taufer, U. of Tennessee

Re-use

Leveraging Proven Solutions Key to Innovation

- **Leveraged HTCondor's**
  - Job submission to heterogeneous, distributed resources
  - Managing job dependencies expressed as DAGs
  - Job retries and error recovery

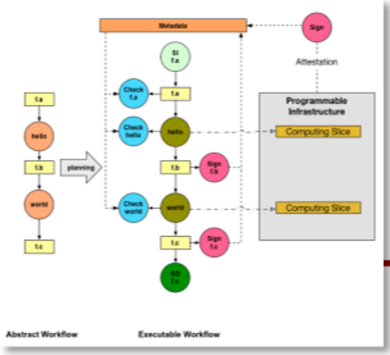
HTCondor
High Throughput Computing



- **Allowed us to focus on other aspects of automation:**
  - Workflow planning, and re-planning in case of failures
  - Automated data management
  - APIs for workflow composition in Python, R, Java, Perl, Jupyter Notebook
  - User-friendly monitoring and debugging tools
  - Specialized workflow execution engines for HPC systems
  - Provenance tracking
  - **Data integrity**

SWIP
Scientific Workflow Integrity with Pegasus

Indiana University
RENCI



CS research

CI Dev.

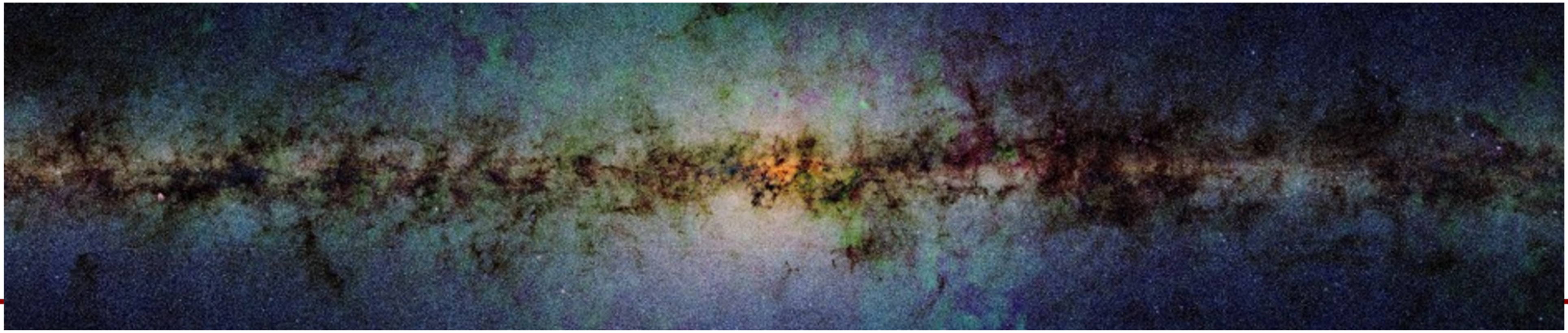
Using Real Applications Provides Realistic Testing and Evaluation

Montage: Important application for CS and CI

Open source, open data, scalable, robust

Helps advance CS and test CI: workflow scheduling, resource provisioning, provenance tracking

One of the workflows used in Pegasus' nightly build and test



Montage, an important Astronomy Application, collaboration with Caltech since 2002

Multi-domain

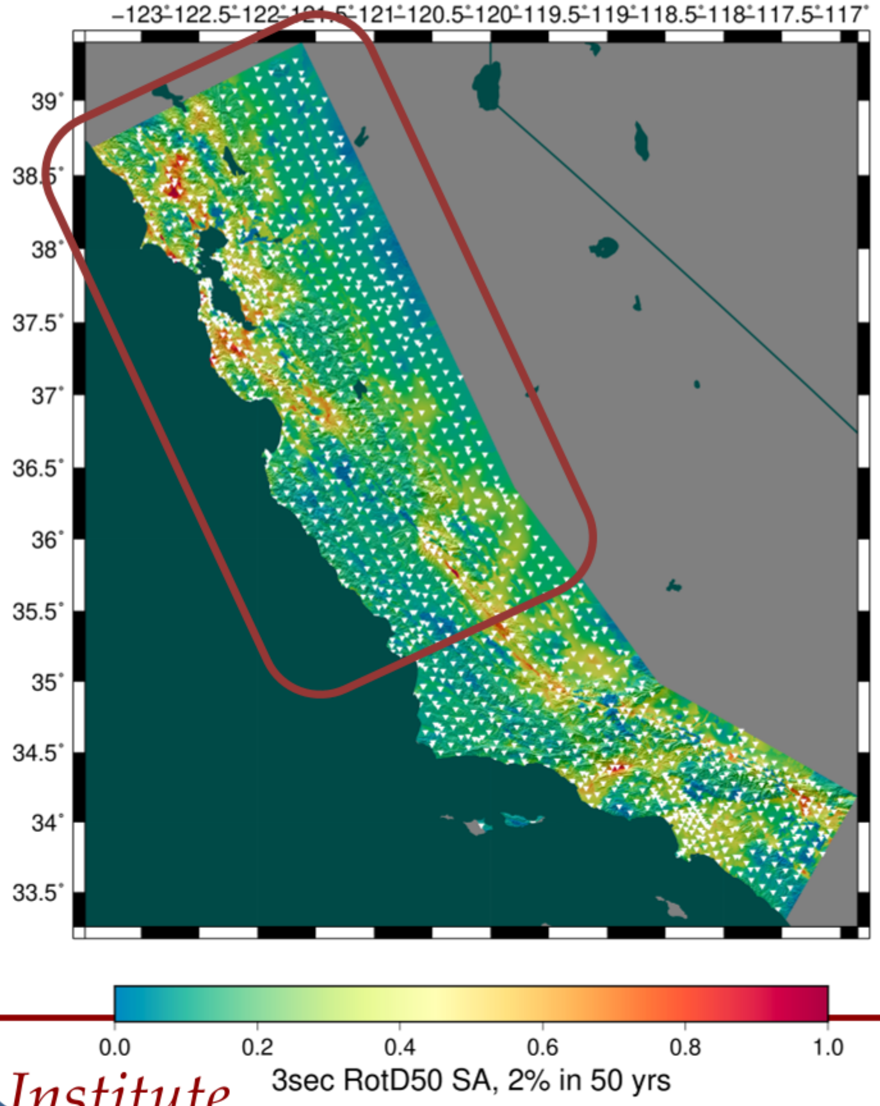
Supporting Large-Scale Applications

Slide credit: USC's Southern California Earthquake Center

SCEC's CyberShake: What will the peak earthquake motion be over the next 50 years?

Useful information for:

- Building engineers
- Disaster planners
- Insurance agencies

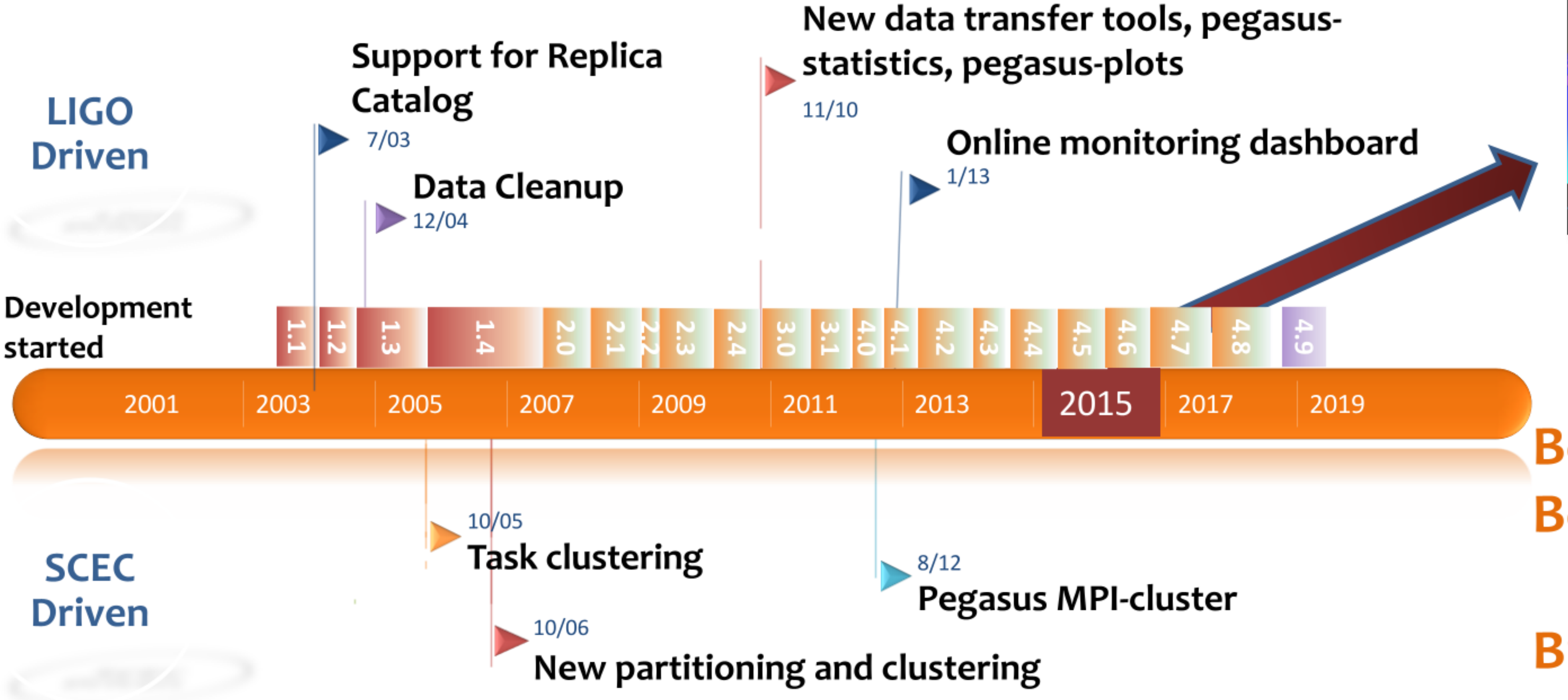


- 120 million core-hours
- 39,285 jobs
- 1.2 PB of data managed
- 157 TB of data automatically transferred
- 14.4 TB of output data archived

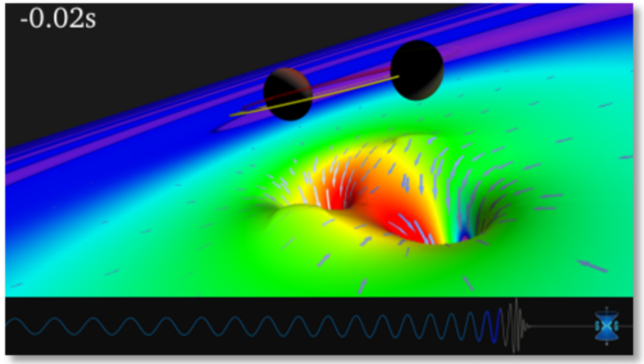
- NCSA *Blue Waters*
- OLCF *Titan*

Total map:
170 million core hours
> 19,407 core years

Cross-pollination between domains is highly beneficial



LIGO's first GW detection



0.2 Second before a pair of massive black holes collide

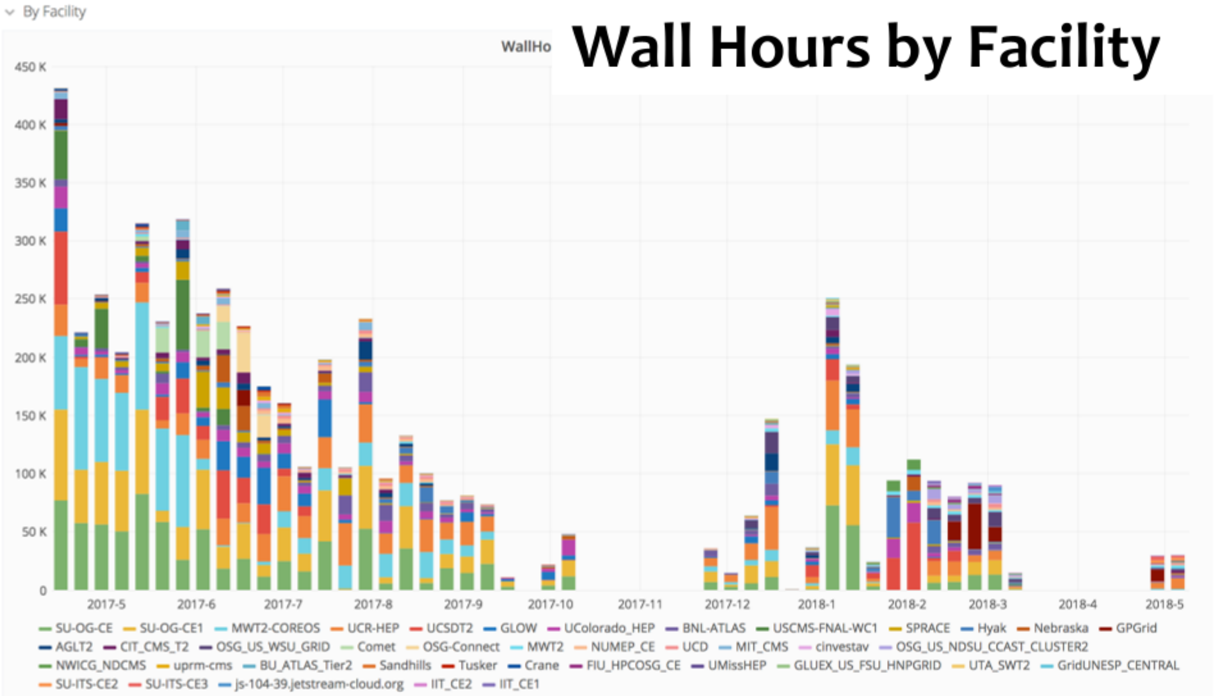
Benefits the applications
Benefits the software

But, can make the software more complex

Multi-domain

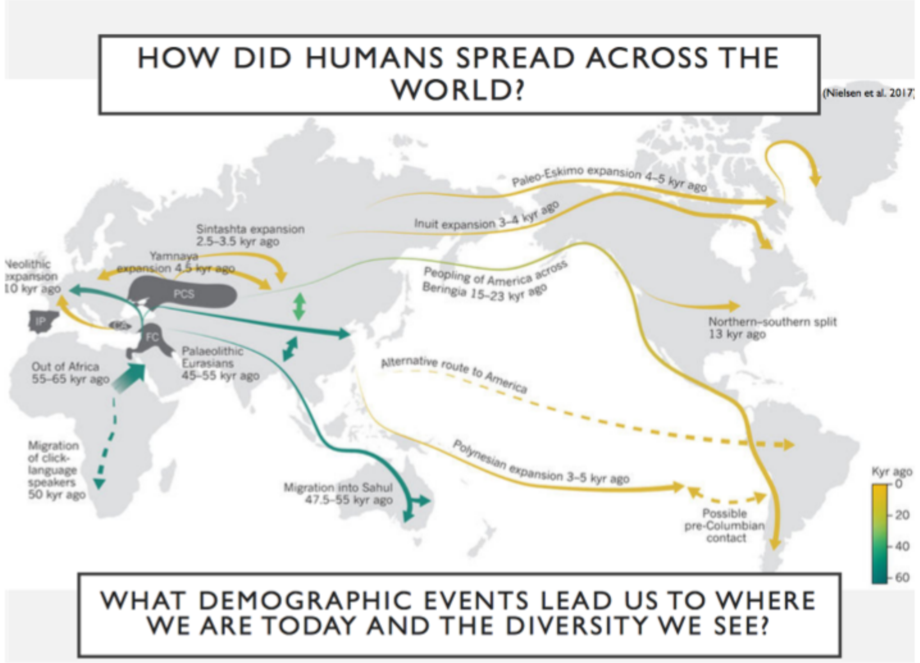
Arming Individual Scientists with Pegasus on OSG

Wall Hours by Facility

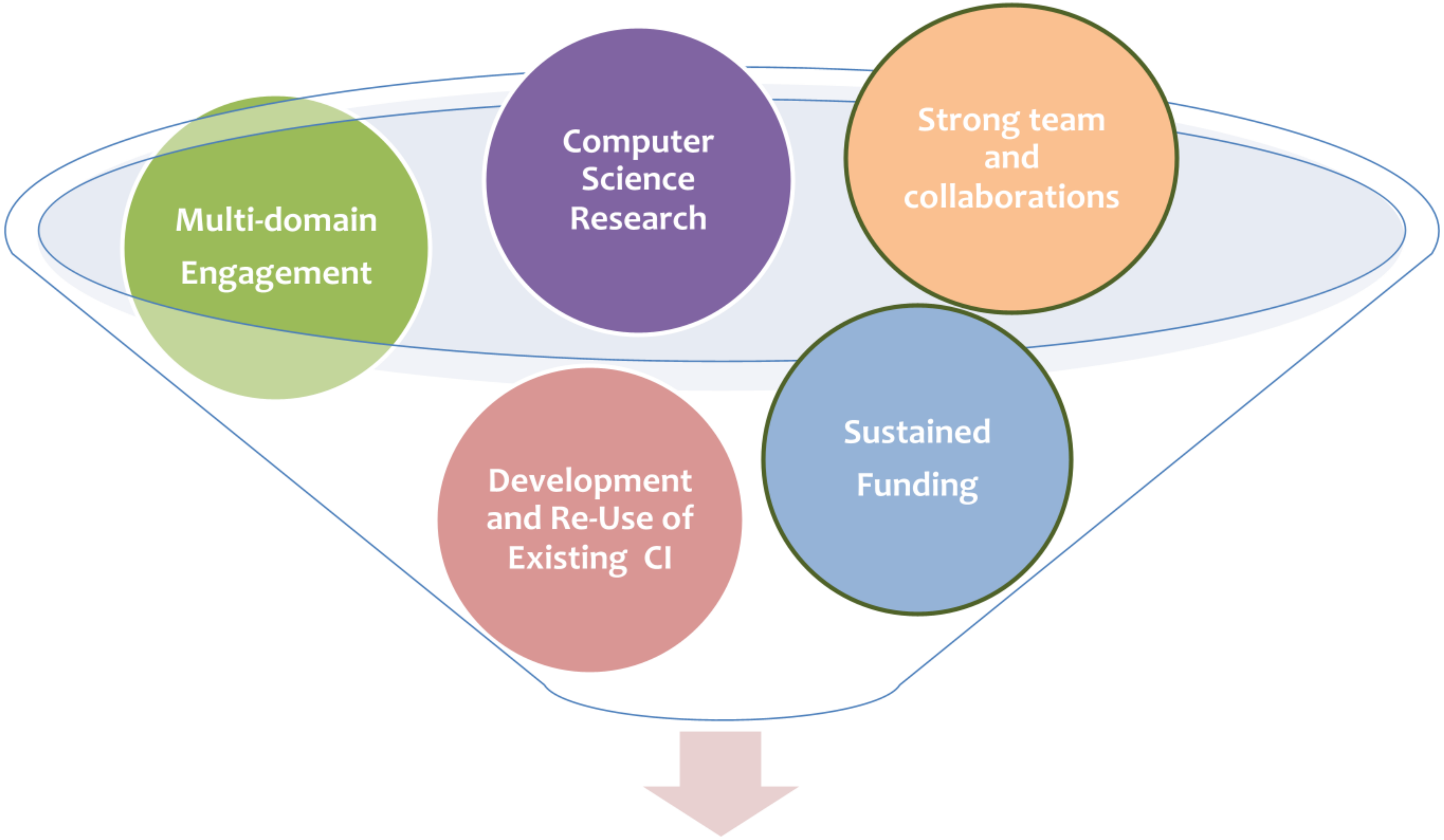


40 execution sites
12 million jobs across 342 workflows
~ 7.3 Million Core Hours

Ariella Gladstein, Ph.D. Student
University of Arizona



Summary of Observations



Dependable and Impactful Software

Looking ahead: Growing Demand for Automation

High Performance Computing Systems

- Complex
- Heterogeneous
- Specialized data storage
- Increasingly faulty

Distributed Systems

- Software Defined capabilities
- Specialized data storage

Clouds

- New platform for science
- Very heterogenous
- Can be costly

Resource Management is Key

Constraints: time, budget, resource capabilities

Faulty environment: detection and attribution

Heterogenous storage: memory, BB, FS, WAN

IoT devices

Programmable networks

Significant and applicable innovation in industry:

Need to keep track of big data technologies and machine learning solutions

Role of AI and Automation

Increased use of automation and ML presents a new set of challenges

- Trust: How do you know that what we observe is real?
- Transparency
- Understanding
- Reproducibility

Science Automation Changes the Workforce Landscape

How will the scientist of the future look like?
How will the human machine interfaces look like?

easy

http://pegasus.isi.edu

BIG Thanks
to the Pegasus Team
and amazing collaborators!

renci
University of Notre Dame
Oak Ridge National Laboratory
UMass Amherst
HTCondor High Throughput Computing
University at Buffalo The State University of New York
The University of Utah
Berkeley Lab
Missouri
Fermilab
Clemson University
Montage
LIGO
SCEC
neon
Rutgers
Trusted CI The NSF Cybersecurity Center of Excellence
TACC
Children's Hospital of Philadelphia
Open Science Grid
German Data Center for SDO DLR
XENON Dark Matter Project
XSEDE
The University of Arizona
Indiana University School of Engineering