Sustaining Distributed Workflow Management Research and Software in Support of Science

Ewa Deelman, Ph.D.

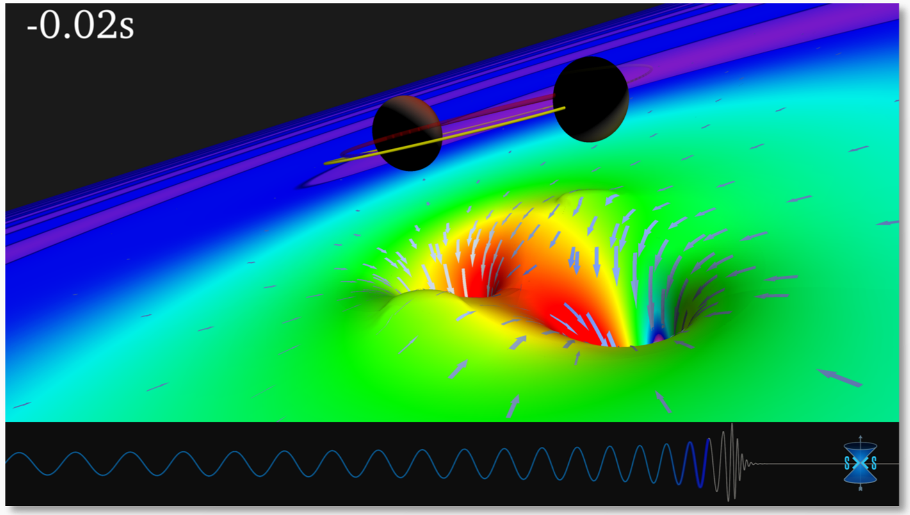
USC Viterbi
School of Engineering
Information Sciences Institute

Longstanding collaboration with Miron Livny

HTCondor
High Throughput Computing

http://pegasus.isi.edu

LIGO’s Gravitational Wave Detection



LIGO announced first ever detection of gravitational waves *Feb 2016*

Created as a result of coalescence of a pair of dense, massive black holes

Confirms major prediction of **Einstein Theory of Relativity**

Detection Event

Detected by both of the operational Advanced LIGO detectors
(4km long L shaped interferometers)

Event occurred at September 14, 2015 at 5:51 a.m. Eastern Daylight Time

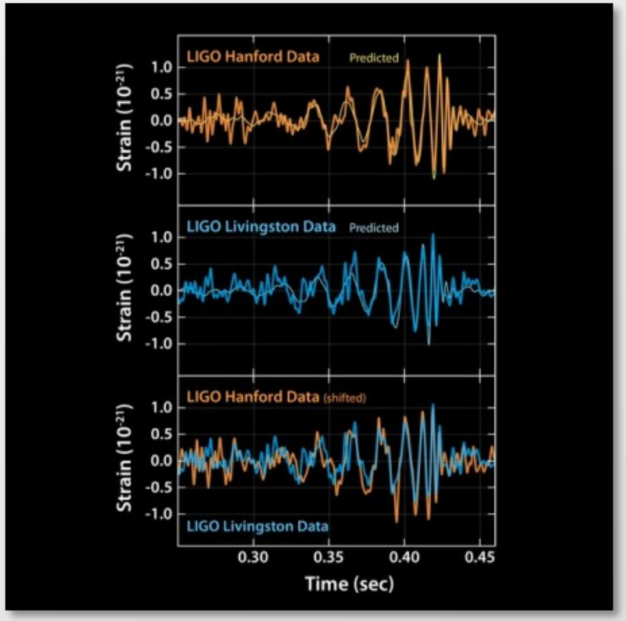


Image Credits: *0.2 Second before the black holes collide: SXS/LIGO*
Signals of Gravitational Waves Detected: Caltech/MIT/LIGO Lab

Advanced LIGO PyCBC Workflow

LIGO Laser Interferometer Gravitational-Wave Observatory
Supported by the National Science Foundation
Operated by Caltech and MIT

One of the main pipelines to measure the statistical significance of data needed for discovery

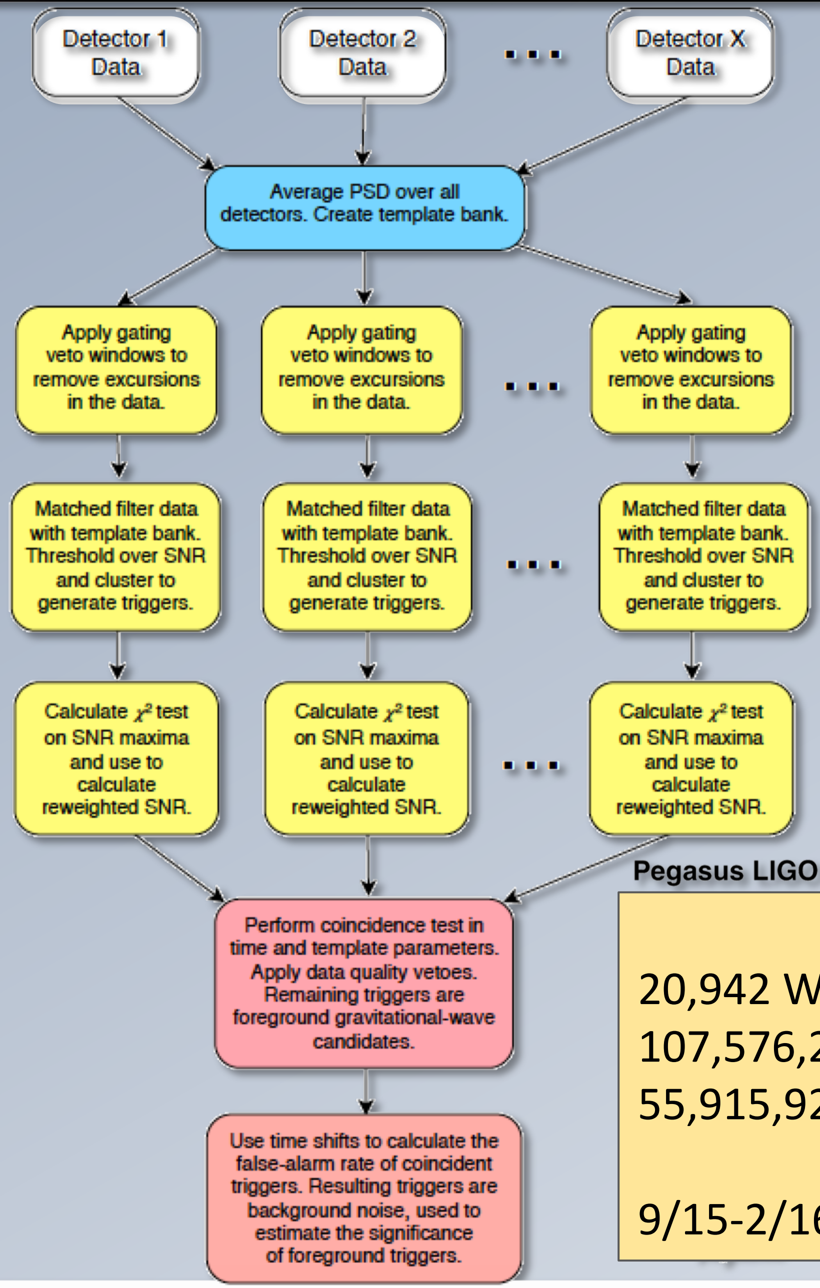
Contains 100's of thousands of jobs and accesses on order of terabytes of data

Uses data from multiple detectors

For the detection, the pipeline was executed on distributed resources in the US and Europe

Use our Pegasus software to automate the execution of tasks and data access

2015-2016



Pegasus LIGO PyCBC Workflow

20,942 Workflows
107,576,294 Tasks
55,915,928 Jobs

9/15-2/16

Pegasus

http://pegasus.isi.edu

Presentation to VIRGO Sept. 2001

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
- User can easily determine
  - What it will cost to obtain data
  - Quality of derived data
- VDG serves requests efficiently, subject to global and local policy constraints

www.griphyn.org

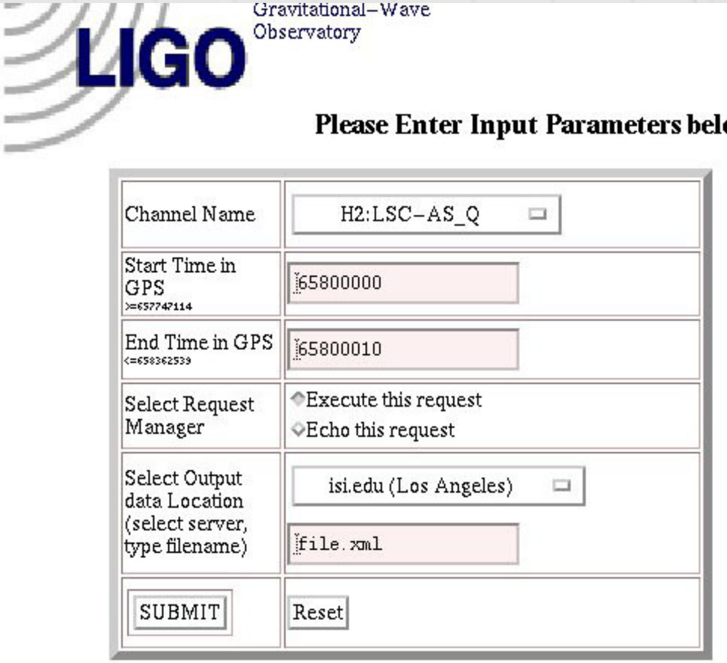
Ewa Deelman, ISI

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

http://pegasus.isi.

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


LIGO Gravitational-Wave Observatory

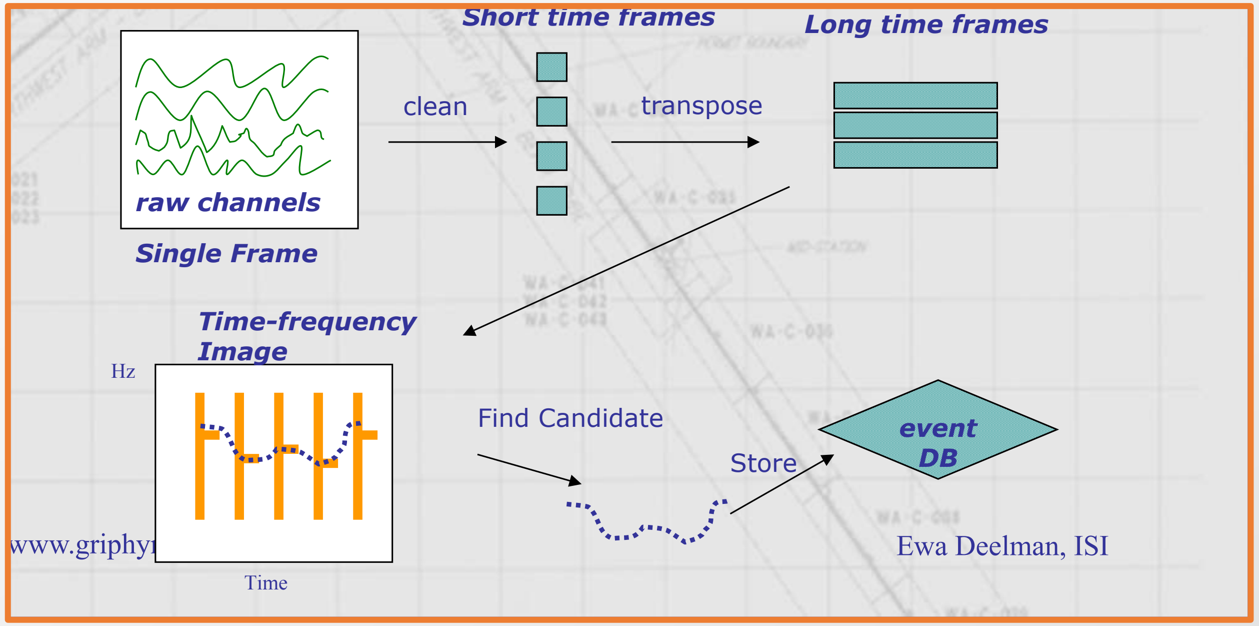
Please Enter Input Parameters below.

Channel Name	H2:LSC-AS_Q
Start Time in GPS >=657747114	65800000
End Time in GPS <=658362539	65800010
Select Request Manager	Execute this request Echo this request
Select Output data Location (select server, type filename)	isi.edu (Los Angeles) file.xml
SUBMIT	Reset

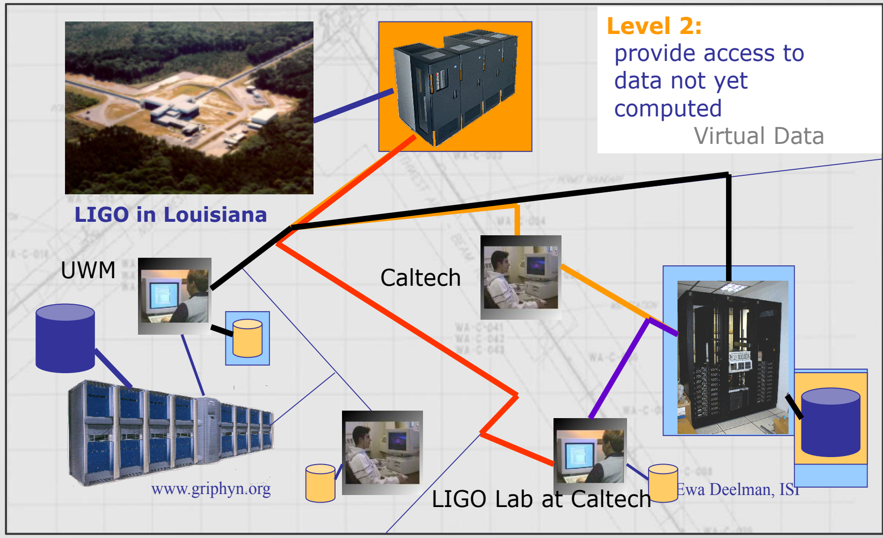
Completion Date November 2001

Lessons Learned

- Listen to the scientists needs – virtual data was a great concept, but too abstract



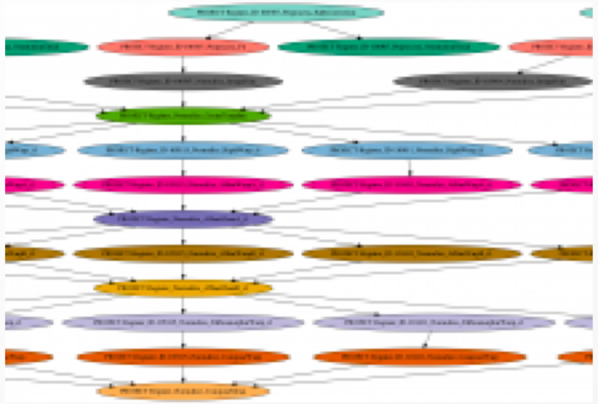
- Need to deal with distributed, heterogeneous data and compute resources
- Need to deal with changing resources/software over time
- Separation between work description and work execution
- Rewarding to work with real world problems



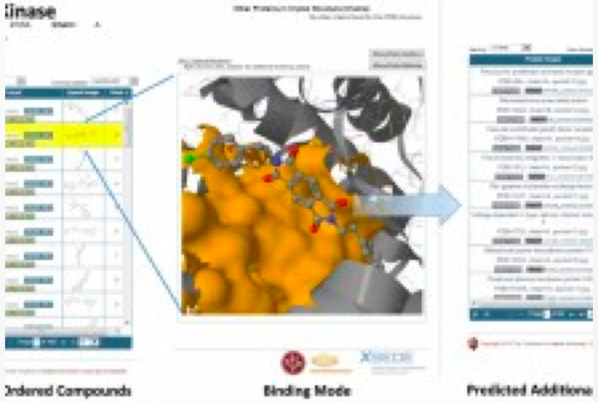
Focus:

- Workflow planning and scheduling (scalability, performance)
- Task execution (scalability, fault tolerance)

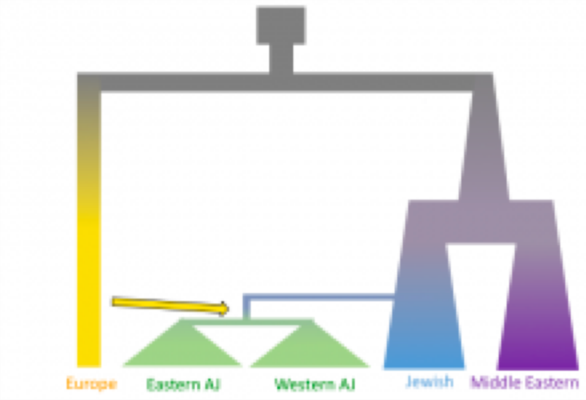
Example Pegasus Applications, varied domains, varied users expertise



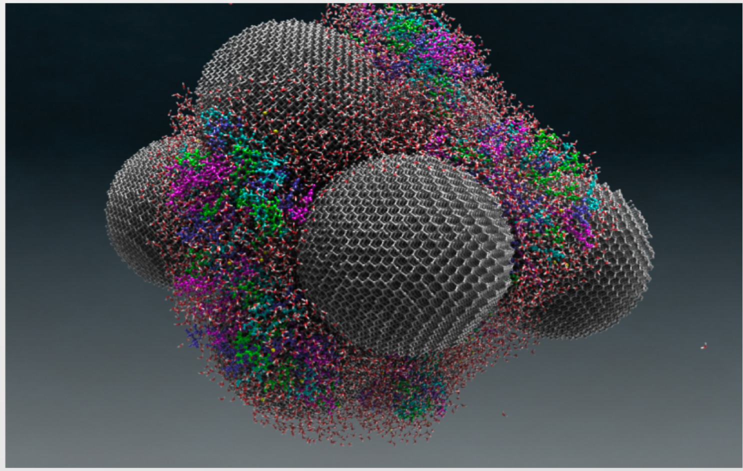
Processing of **MRI data for Alzheimer's** research



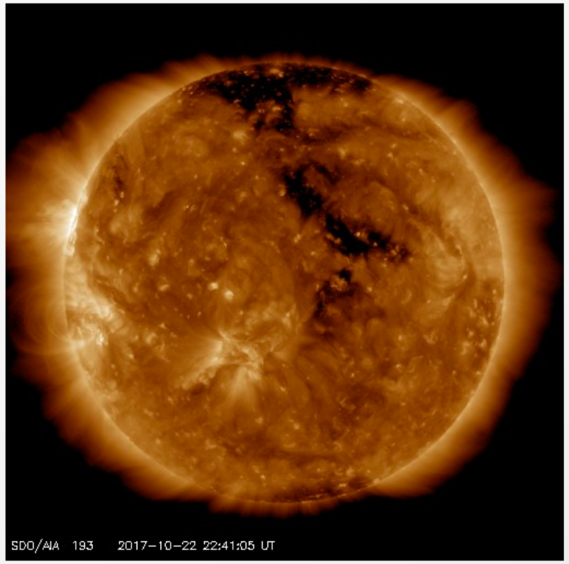
The **Structural Protein-Ligand Interactome** (SPLINTER) project predicts the interaction of thousands of small molecules with thousands of proteins.



Inference of Human Demographic History: Infer human demographic history, such as global migrations, population size changes, and mixing betwe populations through modeling.

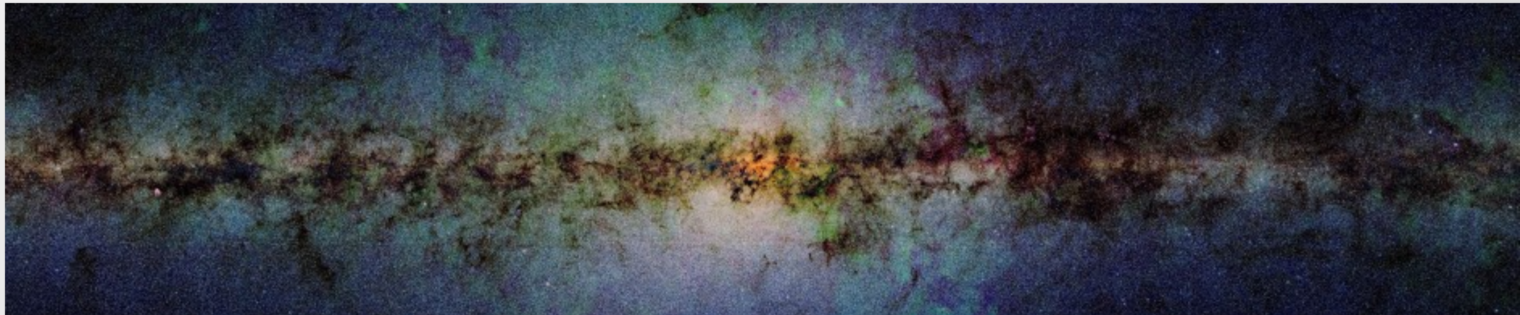


Molecular dynamics simulations for **drug delivery**



Helioseismology

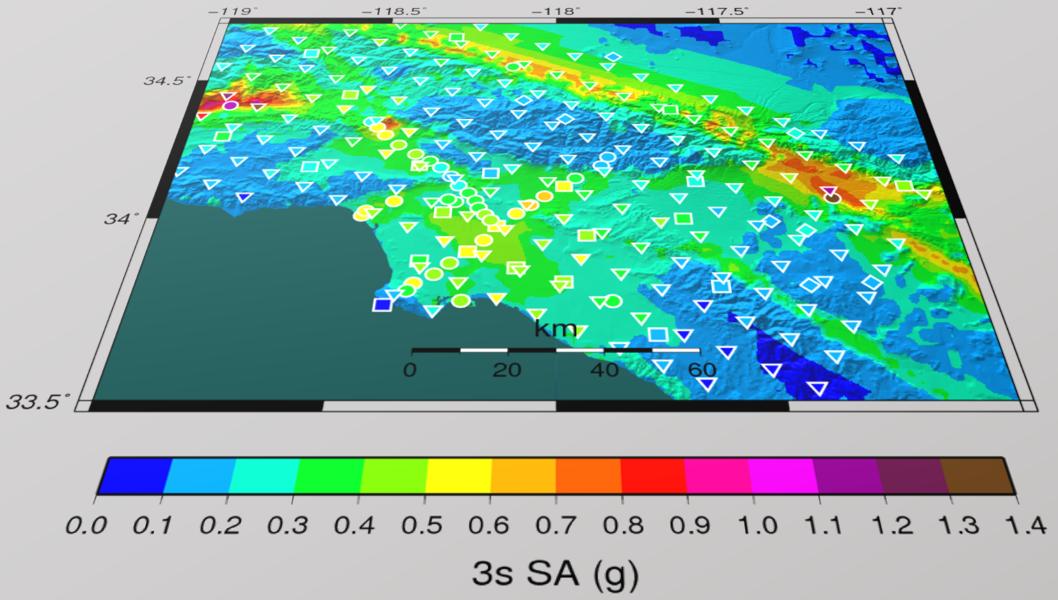
Astronomy/ Montage



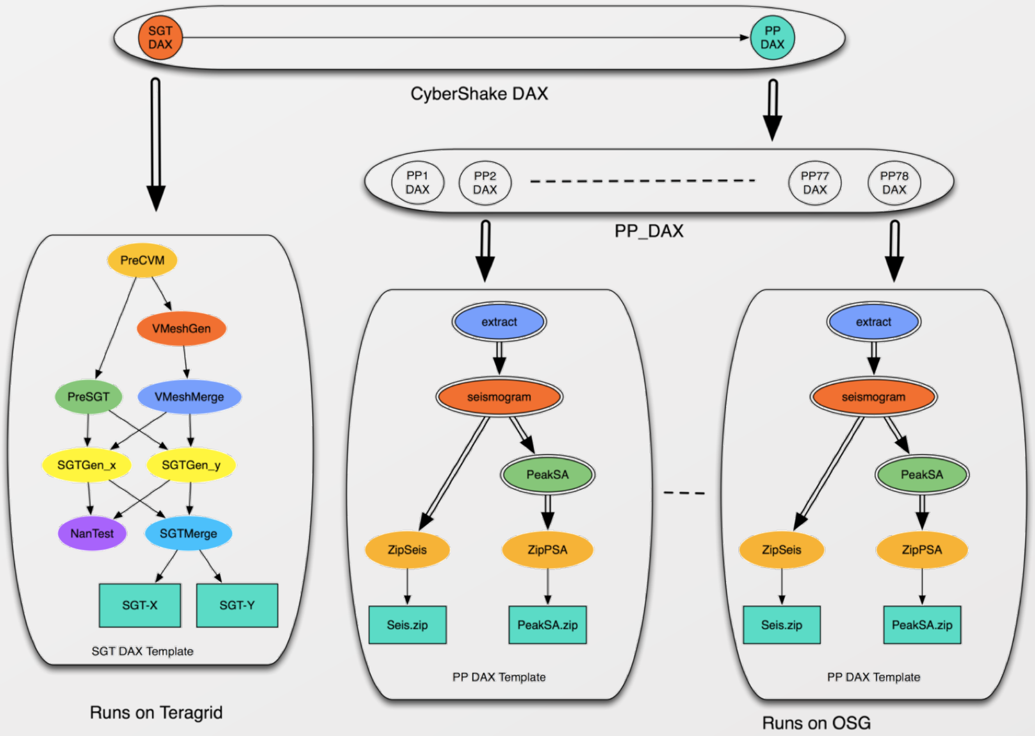
Southern California Earthquake Center's CyberShake PSHA Workflow

Civil engineers ask seismologists: What will the peak ground motion be at my new building in the next 50 years?

Seismologists answer this question using Probabilistic Seismic Hazard Analysis (PSHA)



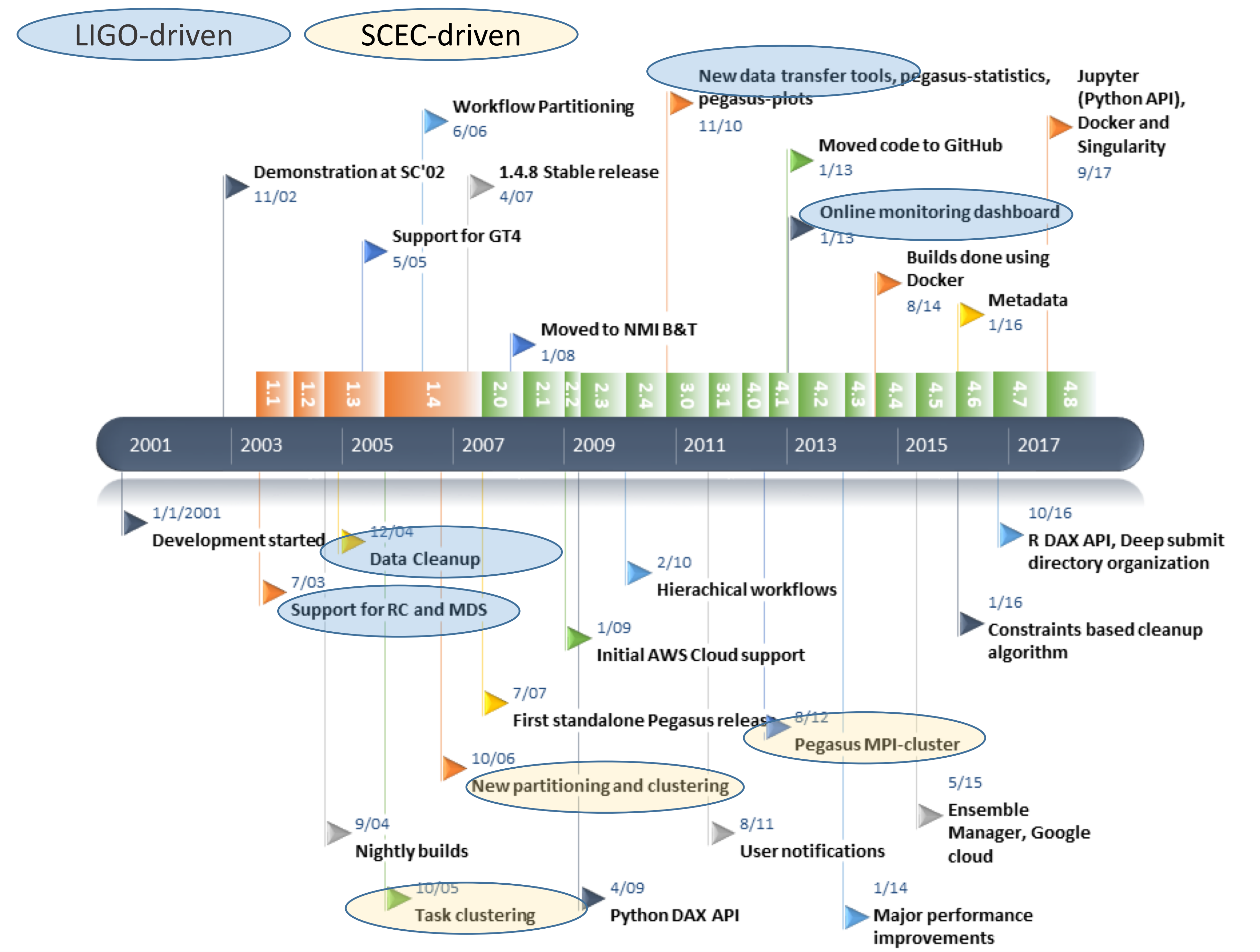
Workload does not match the infrastructure



293 workflows
each workflow has 820,000 tasks

Lessons learned

- Developing capabilities takes time
- Cross-pollination is highly beneficial
- Working with various applications makes software better but also more complex
- Need capable people and sound software engineering practices to make it work



Pegasus

Software Engineering

Small team – easy to communicate!

GitHub and public mailing lists

- Open Source
- Open development with community feedback

Atlassian tools

- Jira: feature/bug/task tracking
- Fisheye: a window into the code changes
- Bamboo: automatic builds and tests
- Confluence: wiki for roadmaps and
- HipChat: quick communication between team members

Test Driven Development

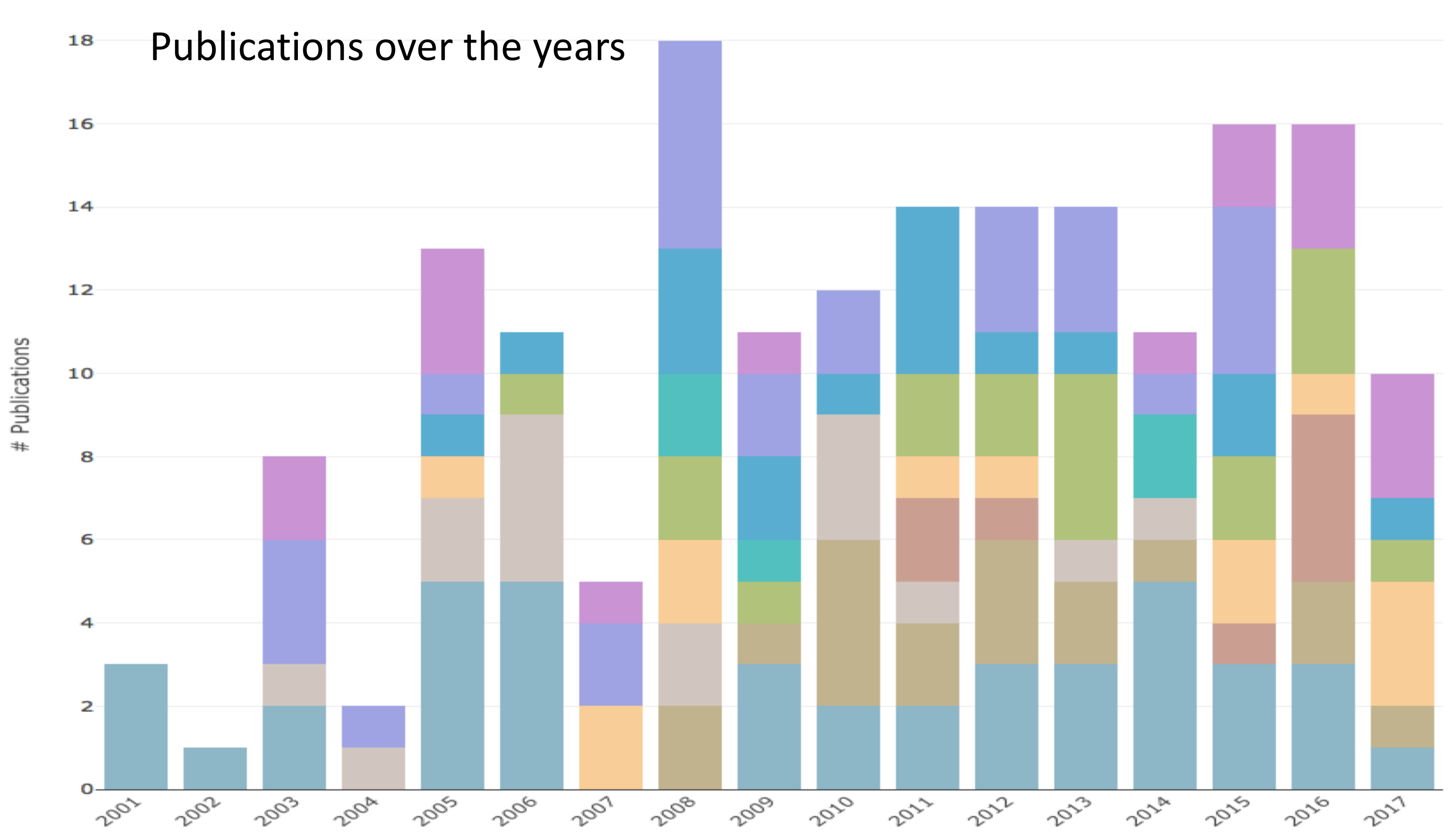
- Builds are run automatically for each code commit
- Unit tests are run as part of each build
- Large functional workflow tests are run every night. Many of the tested workflows are derived from production workflows from our users

Bamboo

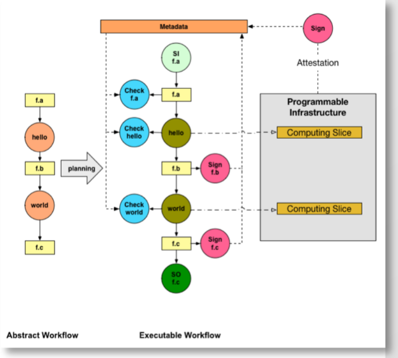
FishEye

JIRA

Lessons learned: It is important to interleave Research and Development, you are judged by your CS achievements, brings satisfaction



Scientific Workflow Integrity with Pegasus



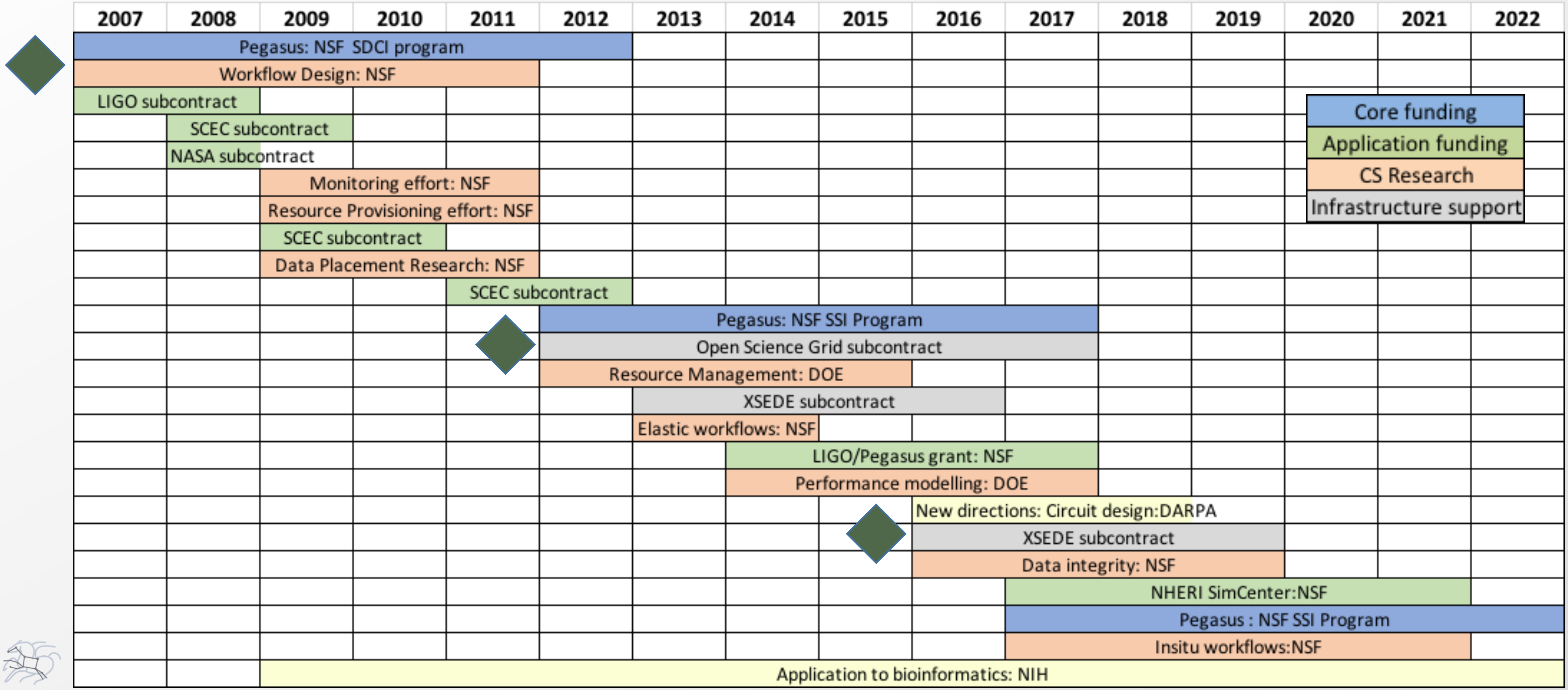
Data Integrity

Pegasus

https://pegasus.isi.edu/workflow_gallery/

To sustain software, need many different funding sources and interleave research, software development, and user support

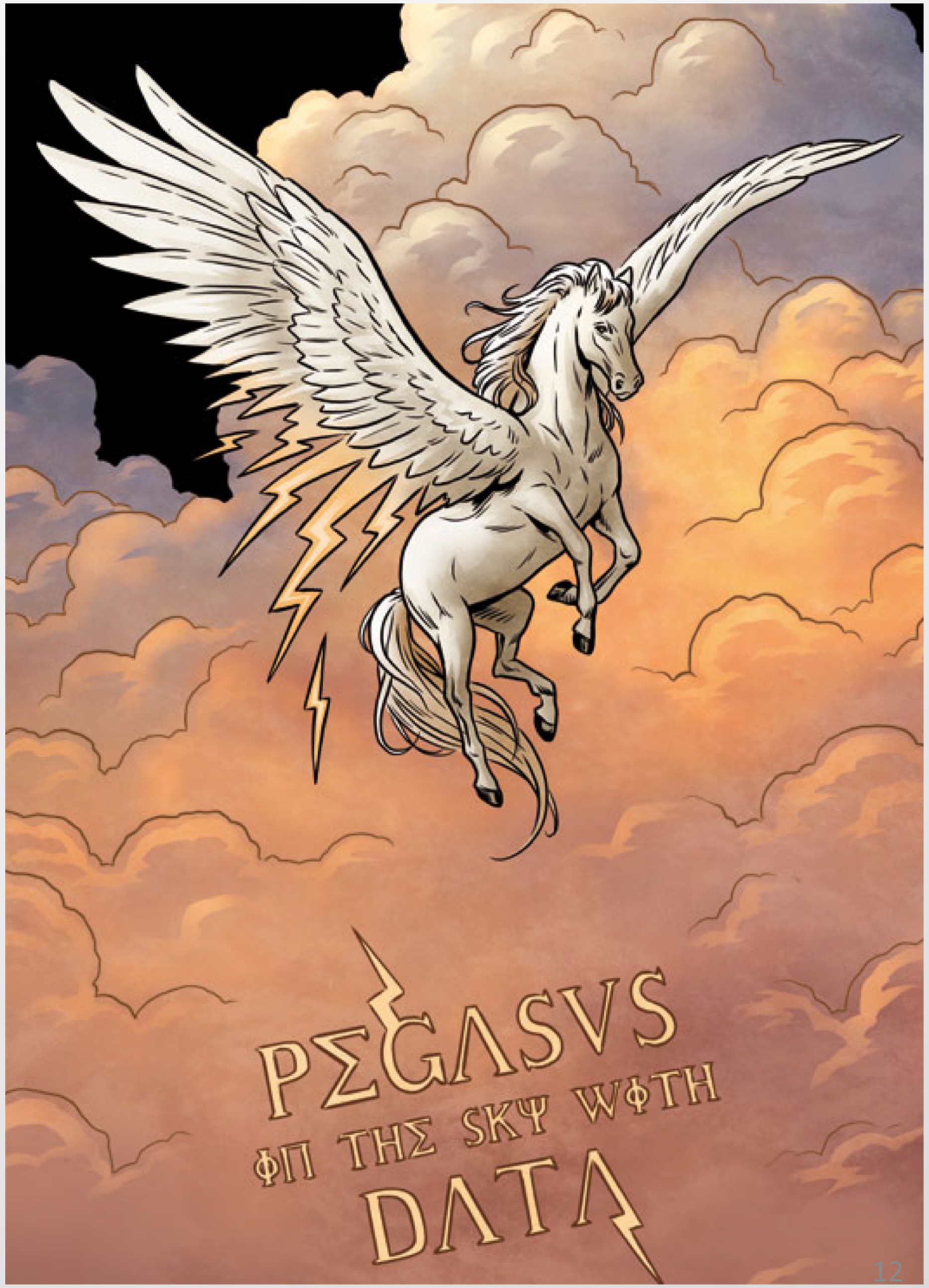
- Pegasus-specific funding



Lessons Learned Summary

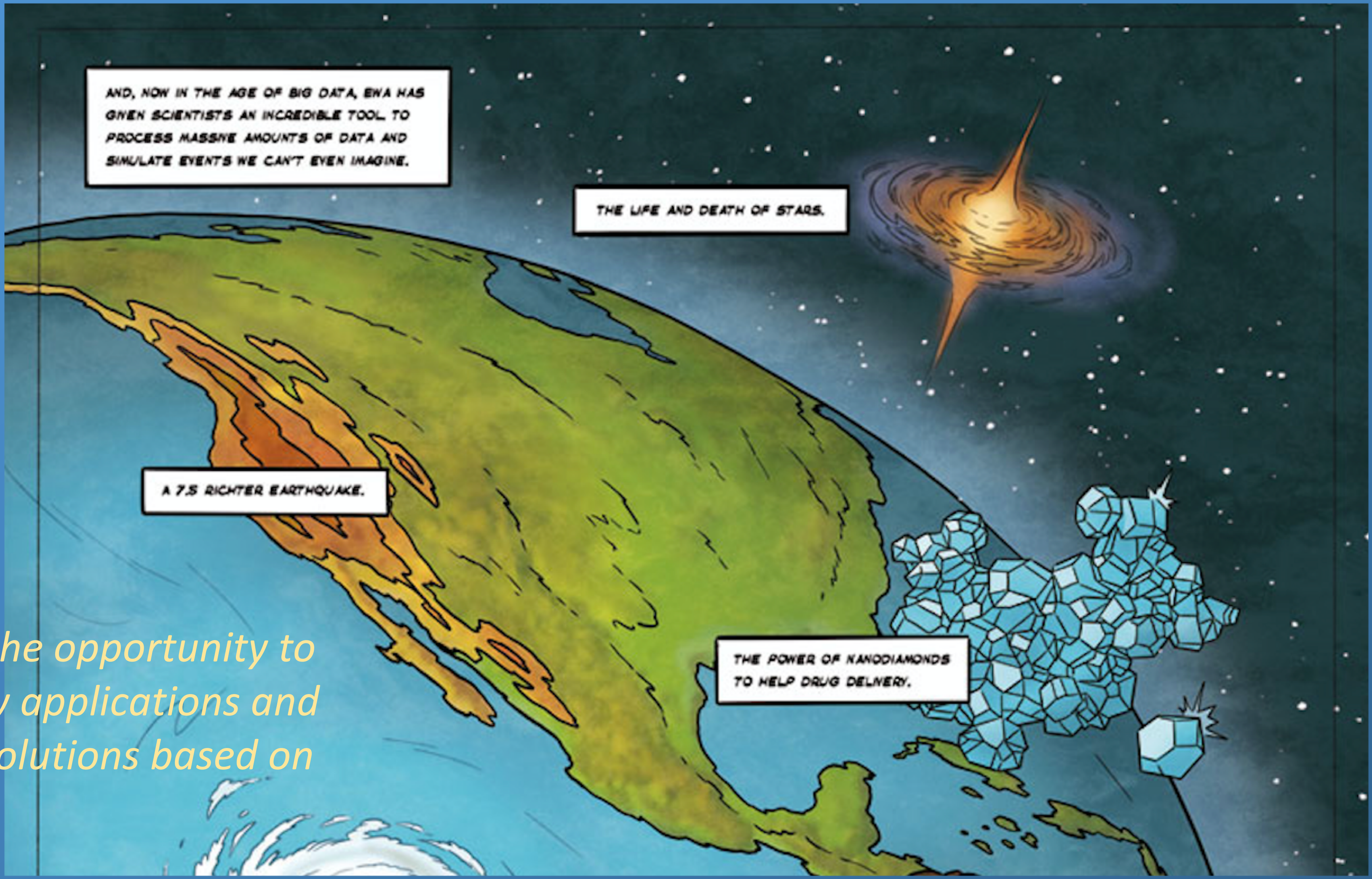
USC Viterbi Magazine, Spring 2018

- Developing production quality software targeting cutting edge science applications and heterogeneous cyberinfrastructure:
  - Involves algorithm development
  - Requires an experienced software development and research team (Karan Vahi, Mats Rynge, Rajiv Mayani, Rafael Ferreira da Silva) and employing good software engineering practices
  - Open source is important
  - Takes time and patience (not all collaborations are easy at times)
  - Needs sustained funding. (diversity is important)
- Need commitment to a vision, collaboration is key
- Collaboration between various CS expertise is critical to research and making software robust: cybersecurity, networking, data management, cloud computing, ..
- Collaboration between CS and domain scientists is critical to making the software relevant
  - Need to work with various applications and communities
- Need to listen carefully to scientists' needs, takes time to develop trust
- Need to abstract user's needs to general concepts applicable across domains
- Good to pick a catchy name and logo and stick to it



Pegasus est. 2001

Automate, recover, and debug scientific computations.



We welcome the opportunity to work with new applications and enhance our solutions based on user's needs.

Pegasus Website

http://pegasus.isi.edu

Users Mailing List

pegasus-users@isi.edu

Support

pegasus-support@isi.edu

HipChat