Science Impact of Sustained Cyberinfrastructure: The Pegasus Example

Ewa Deelman, Ph.D.
University of Southern California, Information Sciences Institute

NSF Office of Advanced CyberInfrastructure Webinar
May 17, 2018

Pegasus represents a long standing collaboration with Miron Livny, University of Wisconsin, Madison
October 16th 2017: “LIGO and Virgo make first detection of gravitational waves produced by colliding neutron stars”

And kick off a new era of multi-messenger astronomy

“The inspiral and merger of two neutron stars, as illustrated here, should produce a very specific gravitational wave signal, but the moment of the merger should also produce electromagnetic radiation that's unique and identifiable as such.”, credit LIGO

NASA’s Fermi space telescope had detected a burst of gamma rays at about the same time

Images credit: LIGO Scientific Collaboration
“aftermath of the BNS merger... On the left are six optical images taken between 10 and 12 hours after the merger by different telescopes. On the right are images constructed from x-ray and radio observations. The x-ray image was taken 9 days after the merger by NASA's Chandra X-ray Observatory. 16 days after the merger NRAO's Jansky Very Large Array (VLA) captured the radio image” from LIGO.org
Dependable Cyberinfrastructure

Takes time, the Pegasus and LIGO Partnership

- First Pegasus prototype
- Blind injection detection
- First detection of a GW from black hole collision
- Multi-messenger neutron star merger observation

First Pegasus prototype

Blind injection detection

First detection of a GW from black hole collision

Multi-messenger neutron star merger observation
First GW detection: Pegasus automated ~ 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

http://pegasus.isi.edu
What does it take to build and sustain Cyberinfrastructure?

The Pegasus lesson - One needs a holistic approach to build dependable CI!
Takes time to build a team and expertise

Back Row: Tu Mai Anh Do, Mats Rynge, Karan Vahi, George Papadimitriou
Front Row: Rosa Filgueira, Ewa Deelman, Rajiv Mayani
Missing: Rafael Ferreira da Silva, Ashwin Venkatesha
Takes Contributions from Many People

GRAs

PostDocs

Master Students and Visitors

Developers

Currently at Amazon, Google, NetApps, SpaceX, Samsung, startups

And others

Takes Collaboration with Many CS and Domain Scientists

http://pegasus.isi.edu
How did Pegasus Start?

Extend the concept of view materialization in DBs to distributed environments

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

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

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

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

Circa. 2001
Challenge: How Translate a Science Request to an Actionable Plan?

Explore AI planning techniques

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)
Benefits of Scientific Workflows (from the point of view of an application scientist)

• Conducts a series of computational tasks
• Chaining (outputs become inputs) replaces manual hand-offs
• Ease of use: gives non-developers access to sophisticated codes
• Provides framework to host or assemble community set of applications, can be multi-disciplinary
• Framework to define common formats or standards when useful
Typical local computational environment

Work Definition

Local Resource

Local Data Storage

http://pegasus.isi.edu
Typical local computational environment

Work Definition

Local Resource

Local Data Storage

Blue Waters
Campus Cluster
XSEDE
DOE Facilities
OSG
Chameleon
Amazon Cloud

http://pegasus.isi.edu
To run Hello World on TACC’s Wrangler

1. Login to TACC
   ```
   localhost$ ssh -l deelman wrangler.tacc.utexas.edu
   ```

2. Write submit script
   ```
   login1.wrangler$ emacs myjob.sub
   ```

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 Wrangler goes down/gets decommissioned? What if the job crashed? What about running on multiple platforms?
Our Approach: Submit locally, Compute globally

Local Resource

Workflow Management System

Work Definition

Local Data Storage

Blue Waters
Campus Cluster
XSEDE
DOE Facilities
OSG
Chameleon
Amazon Cloud

Multi-domain

HTCondor

http://pegasus.isi.edu
Pegasus Today

- Scientists describe their computational processes (workflows) at a logical level, without including details of the underlying CI
  - Operates at the level of files and individual codes
- Pegasus maps the abstract workflow to the available resources and infers the needed data transfers
- Pegasus generates an executional workflow, writes out submit files, and executes the workflow
- Underpins other user facing portals: NanoHub (Purdue)
- Provides workflow management for workflow composition tools: Wings (USC)
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
  - *In situ* workflows

---

New Direction: In-memory coupling of simulation and analytics
Collaboration with U Delaware, Cornell, UTEP

Image credit: Michela Taufer, U. of Delaware
Publications are important for dissemination, education, workforce development, career path, and funding.
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

• Allowed us to focus on other aspects of automation:
  • Workflow planning, and re-planning in case of failures
  • Automated data management
  • Specialized workflow execution engines for HPC systems
  • APIs for workflow composition: Python, R, Java, Perl, Jupyter Notebook
  • User-friendly monitoring and debugging tools
  • Provenance tracking
  • Data integrity
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
Need applications that push the boundaries of what you can do

SCEC’s CyberShake: What will peak earthquake shaking be over the next 50 years?

Useful information for:
- Building engineers
- Disaster planners
- Insurance agencies

2017: 21.6 million core hours, 777TB of data
On ORNL’s Titan and NCSA’s Blue Waters
Since 2007: CyberShake ran on 9 different HPC systems 100 million core-hours (11,416 years)

Pegasus Optimizations:
• Task clustering
• MPI-based workflow engine

Application Optimizations:
• Workflow restructuring
• MPI/code tuning
• Porting to GPUs

2010: World’s first physics-based probabilistic seismic hazard map,

2018: Incorporating earthquake simulator with a 1 million-year catalog of California seismicity

http://pegasus.isi.edu

Slide credit: Southern California Earthquake Center
Arming Individual Scientists with Pegasus on OSG

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

Ariella Gladstein, Ph.D. Student
University of Arizona

http://pegasus.isi.edu

Graph credit: Open Science Grid, Image credit: Gladstein
Cross-pollination between domains is highly beneficial

LIGO Driven
- Support for Replica Catalog: 7/03
- Data Cleanup: 12/04

Development started: 2001
- New data transfer and monitoring tools, 11/10
- Online monitoring dashboard: 1/13

2015
- LIGO’s first GW detection
- 0.2 Second before a pair of massive black holes collide

SCEC Driven
- Task clustering: 10/05
- New partitioning and clustering: 10/06
- Pegasus MPI-cluster: 8/12

Benefits the applications
Benefits the software

But, can make the software more complex

http://pegasus.isi.edu

Image credit: LIGO Scientific Collaboration
To sustain software, need many different funding sources and need to interleave research, software development, and user support.

### Pegasus-related funding

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Core funding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Application funding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CS Research</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Infrastructure support</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pegasus: NSF SDCI program</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Workflow Design: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LIGO subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SCEC subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NASA subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Monitoring effort: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Resource Provisioning effort: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SCEC subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Placement Research: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SCEC subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pegasus: NSF SSI Program</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Open Science Grid subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Resource Management: DOE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>XSEDE subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Elastic workflows: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LIGO/Pegasus grant: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Performance modelling: DOE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>New directions: Circuit design: DARPA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>XSEDE subcontract</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data integrity: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NHERI SimCenter: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pegasus: NSF SSI Program</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>In situ workflows: NSF</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Application to bioinformatics: NIH</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary of Observations

- Multi-domain Engagement
- Computer Science Research
- Strong team and collaborations
- Development and Re-Use of Existing CI
- Sustained Funding

Dependable CI

http://pegasus.isi.edu
Looking ahead: Application Trends

- More complex
- Faster time to solution: instrument steering
- More individual researchers in need of significant CI
  - Need intuitive workflow composition, better monitoring, error handling, assisted debugging

Outreach: How do you reach scientists that don’t know you are out-there?
Many scientists are going through the same pain
Leverage/enhance existing engagement: NSF’s Campus efforts? OSG/XSEDE outreach?
Education and outreach at instruments and experimental facilities?

Planned CyberShake for Northern California:
- 869 geographic sites
- 16,000 workflow jobs
- 70 million core-hours on Blue Waters and Titan
- 800 TB of data

Raw Data: up to $10^{12}$ events per second
Translated Data: Gigabytes to Terabytes
Reduced Data: e.g. Powder Diffraction Pattern
Analysis & Simulation: PDF, MD simulation, etc.
Feedback guiding changes to the experiment setup

Image credit: Vickie Lynch, ORNL
Looking ahead: Growing Demand for Automation

**HPC Systems**
- Complex
- Heterogeneous
- Specialized data storage
- Increasingly faulty

**Distributed Systems**
- Software Defined capabilities
- Specialized data storage

**Clouds**
- New platform for science
- Very heterogeneous
- Can be costly

Resource Management is Key
- Constraints: time to solution, budget
- Faulty environment: failure detection and attribution
- Heterogenous storage: memory, burst buffers, file systems, data xfer nodes

Need to keep track of big data technologies and machine learning solutions that are being developed at a rapid pace by industry
Thank you to the team, collaborators and funders.