Research Computing Building Blocks

INFRASTRUCTURE FOR DATA AT PURDUE
PRESTON SMITH, DIRECTOR OF RESEARCH SERVICES
PSMITH@PURDUE.EDU
Discussion

WHAT ARE THE GAPS BETWEEN THE BUILDING BLOCKS AND THE SCIENCE?

Data in IT

IT has always been about data! Computing and data are inextricably linked.

Purdue has had computing on campus for a very long time, since the days of the CDC 6500 in the 1960s.

We see both analysis and simulation!

https://upload.wikimedia.org/wikipedia/commons/e/e3/CDC_6600_introduced_in_1964.jpg
What is big data?

- Not just Facebook-style analytics!
- 3.5 PB of high-energy physics detector data
- 1 PB of climate model data
  - 90 TB in an active workflow!
- 200 TB of astrophysics simulations
- 150 TB of CFD model output
- 120 TB of audio files
- 100 TB of actively-used next-gen sequencing data
  - Millions of files used in an active workflow
- 10s of TB of video files
- 5 TB of electron microscope images generated per day
- ..to the 75% of users on Conte using less than 1TB
- ... and to the social science researcher with stacks of excel sheets

Big data: A data set that is larger/faster/more complex than one feels comfortable dealing with.
## Scope of Data problems at Purdue

<table>
<thead>
<tr>
<th></th>
<th>Domain 1</th>
<th>Domain 2</th>
<th>Domain 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform 4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Not just a matrix
Scope of Data problems at Purdue

A 3D cube of:

- Domain
- Technology/Methods
- Computing Platform

*Bioinformatics - using Bioconductor on the Snyder Supercomputer*
Discussion: How can we scope this challenge?
Can there be a one-stop place to go?
Research Computing Support of Data

A PLATFORM

VARIOUS DOMAINS AND APPLICATIONS

https://news.uns.purdue.edu/images/+2008/jiang-bacteriophage.jpg
# Our Domains

<table>
<thead>
<tr>
<th>DOMAINS</th>
<th>APPLICATION SPACES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>Molecular Dynamics</td>
</tr>
<tr>
<td>Physics</td>
<td>Image Processing</td>
</tr>
<tr>
<td>Astrophysics</td>
<td>Quantum Chemistry</td>
</tr>
<tr>
<td>Earth and Atmospheric Sciences</td>
<td>Weather Modeling</td>
</tr>
<tr>
<td>Computer Science</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>Chemical Engineering</td>
<td>Big Data</td>
</tr>
<tr>
<td>Electrical and Computer Engineering</td>
<td>Computer Architecture</td>
</tr>
<tr>
<td>Cell and Molecular Biology</td>
<td>Finite Element Analysis</td>
</tr>
<tr>
<td>Agriculture</td>
<td>Statistics</td>
</tr>
<tr>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td></td>
<td>Geospatial</td>
</tr>
<tr>
<td></td>
<td>Remote Sensing</td>
</tr>
<tr>
<td></td>
<td>Visualization</td>
</tr>
</tbody>
</table>
Community Cluster Program

2015 Systems:

*Rice* – Parallel Computing
*Snyder* – Data-Intensive Life Science
*Hammer* – High-Throughput Computing
Steele Cluster, 2008
Your Personal Supercomputer

Commonly-used software, toolkits, compilers, and libraries installed and maintained by ITaP computational scientists.

Easy-to-use graphical access available.
Data Storage

The Fortress archive is a large, long-term, multi-tiered file caching and storage system utilizing both online disk and robotic tape drives.

Ideal for permanent storage of your research data.

https://upload.wikimedia.org/wikipedia/commons/e/e7/Interior_of_Storagetek_tape_library_at_NERSC_(1).jpg
Explosions of Data

Fortress Archive Growth

<table>
<thead>
<tr>
<th>Month</th>
<th>TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar-97</td>
<td>110.91</td>
</tr>
<tr>
<td>Sep-97</td>
<td>550.15</td>
</tr>
<tr>
<td>Mar-98</td>
<td>1,002.30</td>
</tr>
<tr>
<td>Sep-98</td>
<td>2,109.91</td>
</tr>
<tr>
<td>Mar-99</td>
<td>3,273.25</td>
</tr>
<tr>
<td>Sep-99</td>
<td>3,932.26</td>
</tr>
</tbody>
</table>
Data Storage

The Research Data Depot is a high-capacity, high-performance, reliable and secure data storage service designed, configured and operated for a lab’s active research data.

220 research labs
.75 PB allocated

More than just file services!
Data Storage

Supercomputer systems are built with a 1 Petabyte+ scratch filesystem for running jobs.

Holding input data, writing results. Data copied to Fortress or Data Depot.

100T allocated per user.

Very high-speed, very scalable.
No data protection beyond RAID!
Discussion: What about the costs of data storage?

At large scale, costs add up quickly when borne by the researcher.
Data Analytics

- “hathi” Hadoop cluster for prototyping big data applications
- Spark, Hbase, Hive, Pig, Storm etc.

Spark Software fully supported on community clusters as well!

http://bit.ly/1QCennM
Research Networking

As science gets more data-intensive – researchers require increasing amounts of bandwidth

The last mile to the labs is key!

https://pmcdeadline2.files.wordpress.com/2014/05/greenacres132__140501163754.jpg

https://www.nwcouncil.org/media/24501/rural.jpg
Instruments are getting cheaper, more common, and generate more data.

*High-speed (10Gb+) connections for labs and instruments to move data into clusters, storage, and research WAN connections.*
Data Transfer and Sharing

Transfer and share large datasets....

.... With dropbox-like characteristics ....

.... Directly from your own storage system!
Networking
How to balance security, performance, and accessibility to have a high-speed, friction-free end-to-end experience between the lab and HPC?

Instruments
How can we reliably collect and move data?
Hubzero: Collaboration, Online Simulation, and Data

- Databases and digital publications
- Uploaded by researchers in the community
- Digital Object Identifiers and license options
- Data ↔ tools for analysis
Science Gateways

Web-based portals that enable a community to share data, tools, and collaborate.
Research Solutions

A staffing gap exists between the science and the expertise in advanced research technology, for creating new solutions.

- Applied technology and software developers
Computing Literacy

Our computational scientists are investing heavily in teaching faculty and students

- UNIX literacy
- Effective use of clusters
- Programming models (MPI)
- Visualization
- “Big Data” Tools
- Software carpentry

One-on-one instruction as well!
Computing Literacy

Is computing like a car?

As a driver going back and forth to campus, I could say “I don’t know how it does what it does, I just drive it”. It tells me when something goes wrong.

Should researchers be shielded from the details of how computing works for them?
Computing Literacy

Or..

Is the driver in Indianapolis a better analogy?

There are people who make sure the track is in good shape and the car is running fast, but you can bet that the driver understands his car.

- Downforce
- Wind
- Traffic
- Heat
- Tire Wear
- Aerodynamics
Education

How do we train our graduate students to use the computing and data resources they need to develop into computationally-literate scientists?