XSEDE
Extreme Science and Engineering Discovery Environment

BRIDGES

A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE

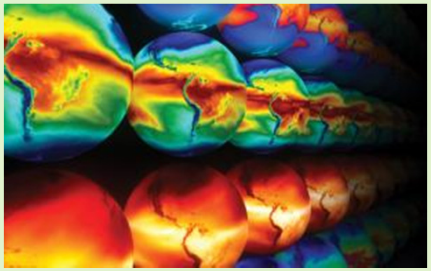
Bridges and the *Data Exacell*: Database-Enabling Technologies for the National Research Community

Nick Nystrom · PSC · nystrom@psc.edu
May 20, 2015

The Shift in Open Research to Big Data

Pan-STARRS telescope
http://pan-starrs.ifa.hawaii.edu/public/

Genome sequencers
(Wikipedia Commons)



NOAA climate modeling
http://www.ornl.gov/info/ornlreview/v42_3_09/article02.shtml

New Emphases

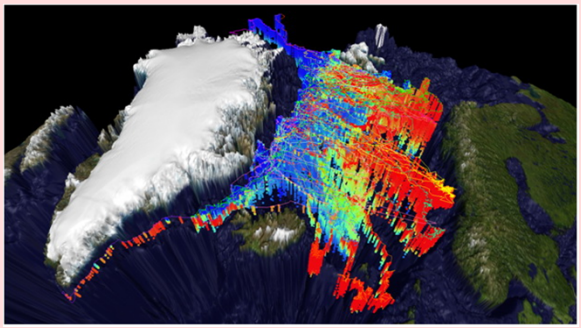
Social networks and the Internet

Video
Wikipedia Commons

Library of Congress stacks
https://www.flickr.com/photos/danlem2001/6922113091/

Collections
Horniman museum: http://www.horniman.ac.uk/get_involved/blog/bioblitz-insects-reviewed

Legacy documents
Wikipedia Commons



Environmental sensors: Water temperature profiles from tagged hooded seals
http://www.arctic.noaa.gov/report11/biodiv_whales_walrus.html

Algorithms and Applications Have Also Changed

Structured Data

Statistics

Optimization (numerical)

Calculations on Data

Scientific Visualization

Unstructured Data

Machine Learning

Optimization (decision-making)

Natural Language Processing

Video

Image Analysis

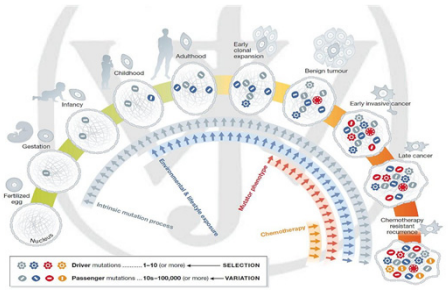
Sound

Graph Analytics

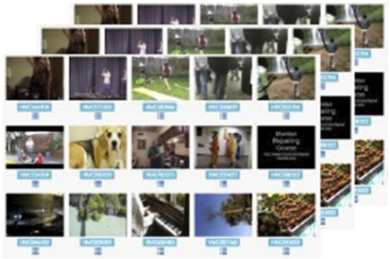
Information Visualization

Objectives and Approach

- Bring HPC to nontraditional users and research communities.
- Allow high-performance computing to be applied effectively to big data.
- Support diverse databases and high-productivity programming environments to allow the creation and execution of data-driven and data-intensive applications and workflows.



EMBO Mol Med (2013) DOI: 10.1002/emmm.201202388: *Proliferation of cancer-causing mutations throughout life*



Alex Hauptmann et. al.: *Efficient large-scale content-based multimedia event detection*

What's Currently Available

- Data Exacell (DXC)
  - Couples analytic resources with innovative storage
  - PSC experts available to help develop applications
  - Research pilot; limited number research projects
- Bridges
  - National-scale (*large*), production resource (XSEDE)
  - DXC users can become early users on Bridges
  - Software developed for DXC will transition to Bridges
- *For open research, these resources are available at no charge.*

The Data Exacell (DXC)

- NSF Data Infrastructure Building Blocks (DIBBs) award #ACI-1261721
- *A pilot project to create, deploy, and test software and hardware building blocks to enable data analytics in scientific research*
- Application-driven: Pilot applications having diverse data analytic requirements motivate, test, and demonstrate the DXC's capabilities
- Components:
  - Crucible: innovative, disk-based near-line storage system featuring low latency, high bandwidth, and high reliability for large-scale datasets
  - Blacklight: the world's largest shared-memory supercomputer, capable of running Java and applications of 1-2048 threads using up to 16TB
  - Sherlock: a unique system for hardware- and software-optimized graph analytics, using either RDF/SPARQL for productivity or threaded C++ for very broad applicability
  - Hadoop and Spark: a cluster of nodes for the Hadoop ecosystem
  - Application, Database, and Web Server Nodes: cutting-edge technologies enable the development of powerful new application architectures

Bridges: Overview

Bridges will be a uniquely capable data-intensive HPC resource designed to empower new research communities, bring desktop convenience to supercomputing, expand campus access, and help researchers needing to tackle vast data to work more intuitively.

The $9.65M acquisition is made possible by National Science Foundation (NSF) award #ACI-1445606:
Bridges: From Communities and Data to Workflows and Insight

HP is delivering *Bridges*

Target schedule:

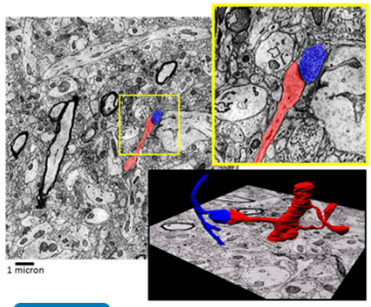
- Oct–Dec 2015: Installation, testing, early user period, and acceptance
- January 2016: Begin production (allocated usage)

An Important Addition to the National Advanced Cyberinfrastructure Ecosystem

Bridges will be a new resource on XSEDE and will interoperate with other XSEDE resources, Advanced Cyberinfrastructure (ACI) projects, campuses, and instruments nationwide.

Examples:

High-throughput genome sequencers



Reconstructing brain circuits from high-resolution electron microscopy

Carnegie Mellon University's Gates Center for Computer Science

Temple University's new Science, Education, and Research Center

Social networks and the Internet

Data Infrastructure Building Blocks (DIBBs)
- Data Exacell (DXC)
- Other DIBBs projects

Other ACI projects

BRIDGES
A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE

PITTSBURGH SUPERCOMPUTING CENTER

Bridging to Nontraditional HPC Users and Enabling HPC + Big Data Workflows

Leveraging PSC's expertise with shared memory, *Bridges* will feature 3 tiers of large, coherent shared-memory nodes – 12TB, 3TB, and 128GB – to support a uniquely flexible and user-friendly environment:

- Interactivity is the feature most frequently requested by nontraditional HPC communities and for doing data analytics and testing hypotheses.
- Gateways and tools for gateway building will provide easy-to-use access to *Bridges'* HPC and data resources, reaching large numbers of users who aren't programmers.
- Database and web server nodes will provide persistent NoSQL and relational databases to enable data management, workflows, and distributed applications.
- High-productivity programming languages & environments (R, Python, MATLAB, Java, Hadoop, etc.) will let users scale familiar applications and workflows.
- Virtualization will allow users to bring their particular environments for portability, reproducibility, and security and provide interoperability with clouds.
- Campus bridging will streamline interoperation with campus resources and enable burst offload capability through a pilot project with Temple University.

Interest from new communities is already very high: examples include the digital humanities, machine learning, statistics, genomics, and radio astronomy.

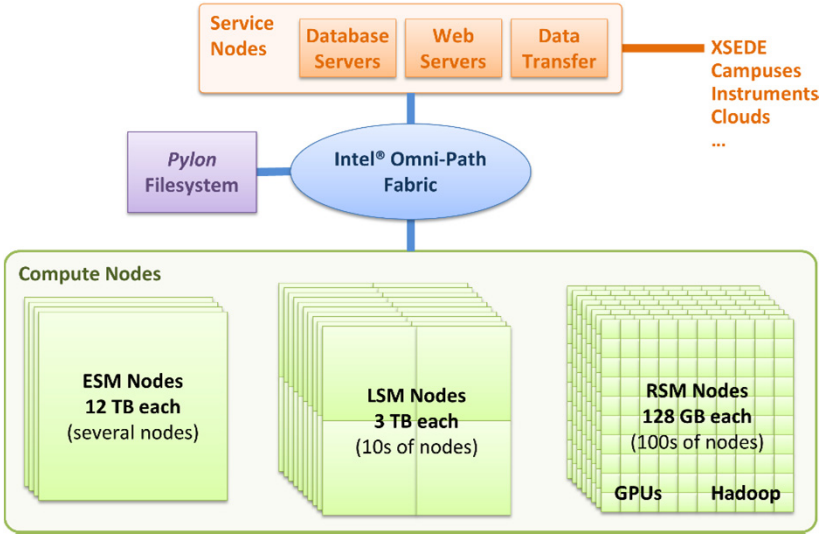
Bridges' Architecture for High-Performance, Data-Intensive Computing (a *very* brief summary)

- 3 tiers of large, coherent shared memory nodes featuring the latest Intel® Xeon® CPUs and NVIDIA® Tesla® dual-GPU accelerators

Node type	Example applications
12 TB	Genomics, machine learning, graph analytics, other extreme-memory applications
3 TB	Virtualization & interactivity including large-scale vis and analytics; mid-range memory-intensive jobs
128 GB	Execution of most components of workflows, interactivity, Hadoop, and capacity computing

- Database nodes and Web server nodes to enable persistent relational and NoSQL databases and enable distributed, service-oriented architectures
- Intel® Omni-Path Architecture Fabric configured in a unique, PSC-developed topology to provide optimal communication and I/O performance
- Data management:
  - Pylon: A large, central, parallel, high-performance filesystem
  - Node-local storage for portability, performance, and performance consistency

High-Level Architecture of Bridges



BRIDGES
A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE
PITTSBURGH SUPERCOMPUTING CENTER

Database and Web Server Nodes

- Dedicated database nodes will power persistent relational and NoSQL databases
  - Support data management and data-driven workflows
  - SSDs for high IOPs; RAIDed HDDs for high capacity

(examples)

- Dedicated web server nodes
  - Enable distributed, service-oriented architectures
  - High-bandwidth connections to XSEDE and the Internet

High-Productivity Programming

Bridges will feature high-productivity programming languages and tools.

Hadoop Ecosystem

- *Bridges* will provide acceleration for Hadoop applications running on its 128GB nodes.
- Large memory will be great for Spark.

BRIDGES
A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE

PITTSBURGH SUPERCOMPUTING CENTER

Meeting Research Needs – *Examples*

Scaling research questions beyond the laptop

From individual researchers to teams & collaborations

Cross-domain analyses

Shared data collections & related analysis tools

Workflows – HPC power without the programming

Large-memory applications & in-memory databases

Optimization & parameter sweeps

Powerful collections of application & tools

Modern, widely-used software environments

Examples of Early Applications

- Finding causal relationships underlying cancer, chronic lung disease, and brain disorders (autism, schizophrenia)
- Assembling large genomes and metagenomes
- Analysis of financial markets and policies
- Improving the effectiveness of organ donation networks
- Recognizing events and enabling search for videos
- Understanding how the brain is connected from EM data
- Addressing public health issues from social media data
- Analyzing large bodies of work in the digital humanities
- Data integration for history, political science & cultural studies
- Cross-observational analyses in astronomy & other sciences

For Additional Information

Project website: www.psc.edu/bridges
@psc_live

Bridges PI: Nick Nystrom
nystrom@psc.edu

Co-PIs: Michael J. Levine
Ralph Roskies
J Ray Scott

Project Manager: Janet Brown

BRIDGES
A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE
PITTSBURGH SUPERCOMPUTING CENTER