The Data Exacell (DXC):
Data Infrastructure Building Blocks for Coupling Analytics and Data Management

Michael J. Levine, Nick Nystrom, J Ray Scott, and Ralph Roskies
Pittsburgh Supercomputing Center
{levine,nystrom,scott,roskies}@psc.edu

NITRD FASTER Briefing
December 15, 2014

Outline

1. Motivation & Architecture	Michael J. Levine
2. Pilot Applications	Nick Nystrom
3. SLASH2 and Systems Software	J Ray Scott
4. Introduction to *Bridges*	Nick Nystrom
5. Q&A	

Abstract

Working in collaboration with, and striving to serve the needs of a growing, diverse set of emerging and existing applications requiring computing and other IT capabilities beyond those typically available to individual research groups yet not well served by existing HPC systems the Pittsburgh Supercomputing Center (PSC) is carrying out an accelerated, development pilot project to create, deploy and test software building blocks and hardware implementing functionalities specifically designed to support data-analytic capabilities for data intensive scientific research. This Data Exacell (DXC) project, building on the PSC's successful Data Supercell (DSC) technology which replaced a conventional tape-based archive with a disk-based system to economically provide the much lower latency and higher bandwidth data success necessary for data-intensive activities, will implement and bring to production quality additional functionalities important to such work. These include improved local performance, additional abilities for remote data access and storage, enhanced data integrity, data tagging and improved manageability. In support of data-analytics and data-intensive processing, we are acquiring hardware appropriate for running a broad collection of databases and interacting with them directly (e.g. over the Web) or as part of data-intensive workflows thereby increasing DXC's effectiveness and applicability for the full range of data-analytic research. In the technological ecosystem, DXC fills the void between computationally-intensive systems (now driving towards exascale) and shared-nothing clusters (including commercial clouds).

We will discuss DXC's current application complement, its architecture, computational engines (including small and large database systems, large cache-coherent memory system, graph analytics and data-analytic systems) and storage systems (including shared SSDs, local and multi-petabyte, shared, high-performance file systems) interconnected by a high-performance fabric in a 'data-centric' architecture.

In addition we will sketch the extension of the DXC pilot project to the recently announced NSF award to the Bridges, large-scale production facility scheduled to go on-line in 2016.

Introduction

- The Pittsburgh Supercomputing Center:
  - Joint effort of Carnegie Mellon University and the University of Pittsburgh
  - 28 years national leadership in:
    - High-performance and data-intensive computing
    - Data management technologies
    - Software architecture, implementation, and optimization
    - Enabling researchers nationwide
    - Networking and network optimization
    - Ground-breaking science and engineering
  - Supported by: NSF, NIH, the Commonwealth of Pennsylvania, DOE, DoD, foundations, and industry
- The presenters…

Introduction

- Few words on scope;
  - We will be describing the Data Exacell (DXC) accelerated, development pilot project, a part of the NSF Data Infrastructure Building Blocks program (DIBBs), which is creating an advanced environment for data intensive computing.
  - This is being created on the current PSC infrastructure
  - "DXC" is used to refer both to the overall DIBBs project and to the large data store which is but a part of that project.
  - *Bridges* award: a production environment scheduled for late 2015
    - DXC is now the de facto testbed for many parts of Bridges
    - Today, Nick will give a brief summary of Bridges. [⟶Nick]
  - The many DXC hardware and software components are being constructed as *building blocks* useable in other environments.

Presentation Plan

- Given the breadth and complexity of the DXC project, we are not able to go into full detail in this presentation.
- Instead, we present first a general description followed by more details as part of the historical background and motivation for DXC.
- We then go into greater depth on 2 subjects + a *preview*:
  - The wide range of pilot, collaborating, data-intensive application groups which are providing both guidance on the course of development and testing of DXC capabilities. (Nick Nystrom)
  - The SLASH2 software supporting the DXC large, shared file & archival system. (JRay Scott)
  - Brief introduction to Bridges (Nick Nystrom)
- Depending on time we can certainly go into greater depth where desired.

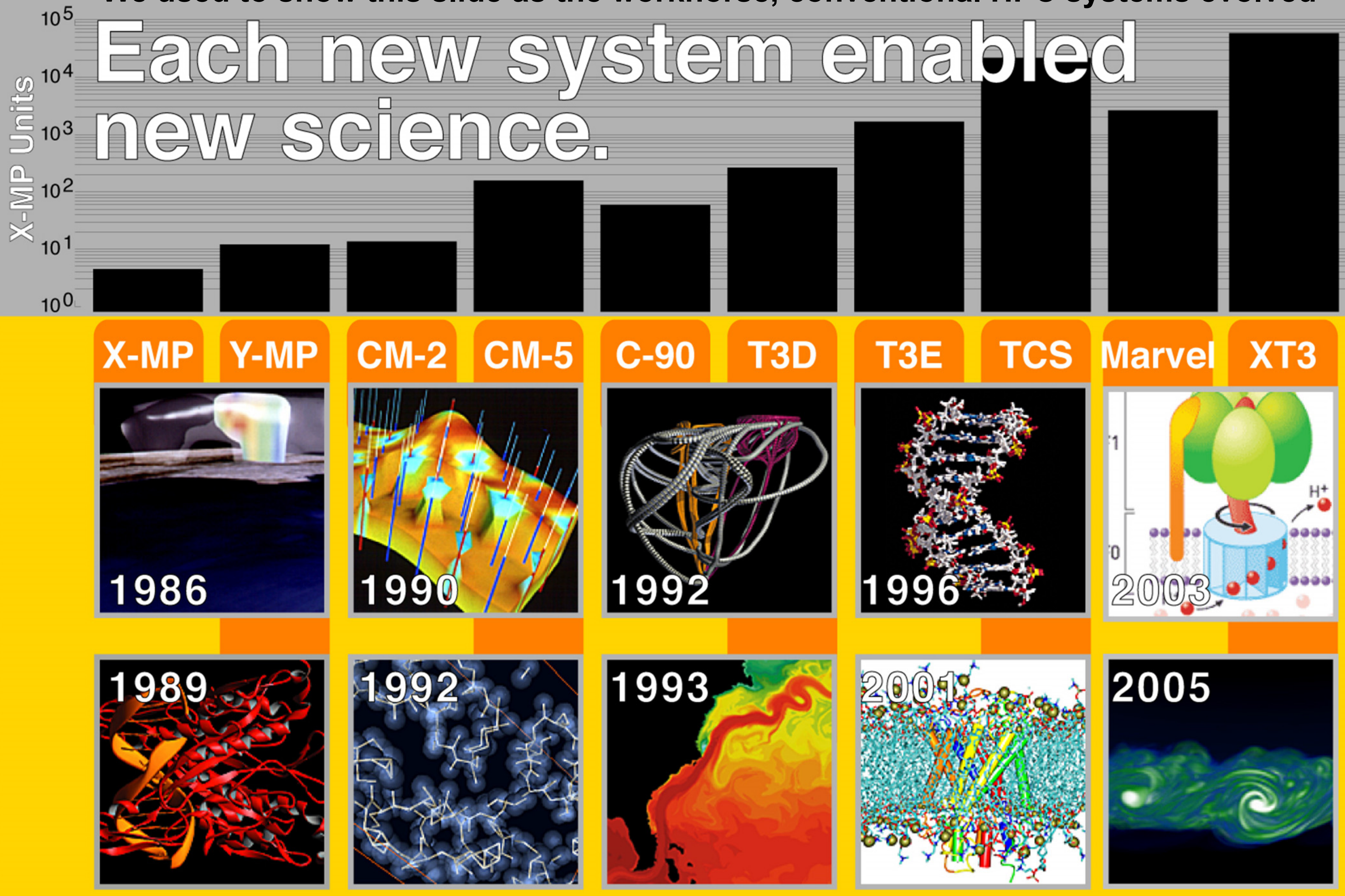
DXC: General Description

- Components (at PSC & some collaborators)
  - PSC's existing *data-analytics relevant* systems: Blacklight (very large memory) & Sherlock (graph analytics) to be followed by upgrades.
  - Additional servers , of varying memory sizes, for data intensive analysis & database work
  - A multi PB, fast, shared, disk-based file system – an enhancement of an existing, disk-based, production mass-store system – coupled to a variety of computers' local data stores.
  - A data-centric, high-performance interconnect (with off-site extensions)
  - Specialized file system software and analysis software
- Collaborating, data intensive research groups
  - Guiding the design of the overall system as a powerful and effective environment for a wide range of data-intensive, data-analytic work.
  - Providing testing of that system as it develops.

Now, on to the **Historical Background and Motivation**

We used to show this slide as the workhorse, conventional HPC systems evolved

Each new system enabled new science.



History of first or early systems

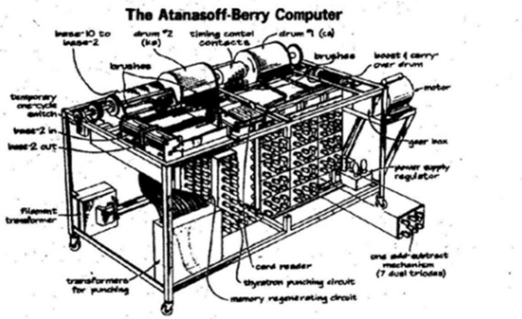
But, with the growth in importance of data-intensive research, things have changed.

New Systems ←→ New Science

- *Historical* HPC: Systems → Science
- Data Intensive: Systems ← Science
  - New fields are driving the selection of different types of systems
- System characteristics (which we will treat in turn)
  1. Computers
  2. Storage systems
  3. System architectures
  4. Interconnection fabrics
  5. Services

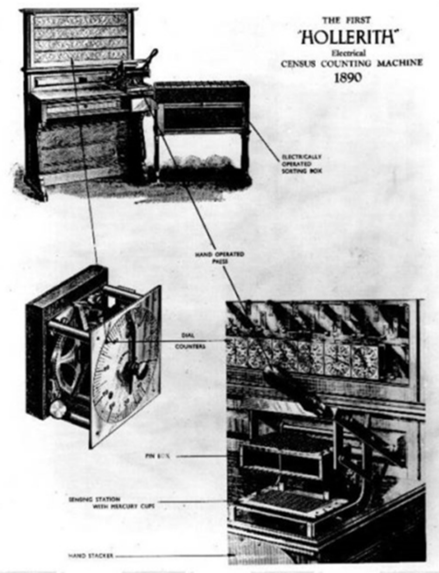
Characteristic: Computers

- *Historical* HPC
  - More powerful versions but of similar design
    - PSC: Cray vector MP, Cray MPP, …
    - Innovations: hard to retain edge (TMC, XMT)
      - Small research market limits engineering input
    - Few systems; mostly large, homogeneous ensembles.
    - (Evolved from Atanasoff & Berry; solving coupled linear equations)
- Data Intensive
  - Added HW capabilities & programming modes
  - Multiple system types able to share data access.
    - Range of memory sizes
    - Provision for GPGPUs
    - Ensembles of many diverse systems
    - (Evolving from Hollerith & 1890 US Census)



Binary Breakthrough

Few know of John Atanasoff, creator of the first digital computer.



Hollerith Tabulator and Sorter
Showing details of the mechanical counter and the tabulator press.

Characteristic: Computers

- Data Intensive
  - Valuing technologies not commonly used in HPC.
    - Cloud-type services for capacity computing problems.
    - New, very large memory systems being created for the Enterprise database & business analytics communities.
    - (Similar to the needs of the intelligence community.)
- PSC/DXC:
  - *Sherlock* (YarcData Urika)
    - Multi-threaded architecture
    - Proprietary processors & memory interfaces
    - Optimized for specific graph-analytics methodologies

Characteristic: Computers

- PSC/DXC:
  - *Blacklight* (SGI UV1000) [above, right]
    - Hardware cache-coherent memories
    - 2 16TB systems
    - Running as (16 + 8 + 8)TB to accommodate jobs
  - DXC: planned upgrade of existing large, shared memory system.
  - Hoping to acquire 2x HP Superdome X [at right]
    - Hardware cache-coherent memory
    - 12 TB, 240 cores each
    - Stronger
      - IO bandwidth,
      - inter-processor bandwidth,
      - memory bandwidth,
      - processor strength.
    - (Code named *DragonHawk* likely to stick*)*

Characteristic: Storage Systems

- *Historical* HPC
  - Dominated by reuse of expensive fast storage & system snapshots
  - Most data: write once, read never, latency tolerant.
  - Form: computer ↔ file system ↔ tape archive (HSM)
- Data Intensive
  - Demanding, complex data motion patterns
  - Latency & low bandwidth intolerant
  - Multiple modes: memory, SSD, disks
  - Both computer-local and shared
  - Form: varied and application dependent

Characteristic: Storage Systems

- PSC/DXC
  - HSM (tape) replaced by DSC (disk)
  - Disk advantages
    - Disk latency = Tape latency / 10,000
    - Much lower cost/bandwidth
    - Lower overall cost (advantage is time varying)
      - High volume & commodity vs low volume & proprietary
  - 5.5 PB (raw)
  - In production for ~3 years
  - PSC developed *SLASH2* file system **[→JRay]**
    - Development supported by National Archive.

Characteristic: Storage Systems

- PSC/DXC (the storage system) [Data Exacell]
  - Improved performance & flexibility
  - New HW technologies for performance
    - SAS3, IB, Haswell
  - New software: manageability
    - Linux
  - New configuration for performance
    - Higher ratio of servers to disks
    - Higher peak bandwidths
    - Storage Building Block
      - 176 3TB disks into
      - 38 GB/s aggregate SAS3 bandwidth into
      - More powerful server into
      - ~15GB/s IB fabric
  - More complex topologies and added functionality

Characteristic: System Architectures

- Three types (see next slide)
  1. A computer + (disk file system encapsulating tape HSM)
     - File system combines fast storage for active jobs with
     - Staging of snapshots and less frequently used files to tape archive
  2. A few computers each with local store +
     - Shared disk store with HSM tape backing +
     - Weak connections to remote users.
  3. Multiple (10’s -100’s) computers
     - Multiple types
     - Computer-local and shared storage (disk & SSD)
     - Multiple interconnects
     - Extensions “off-site”

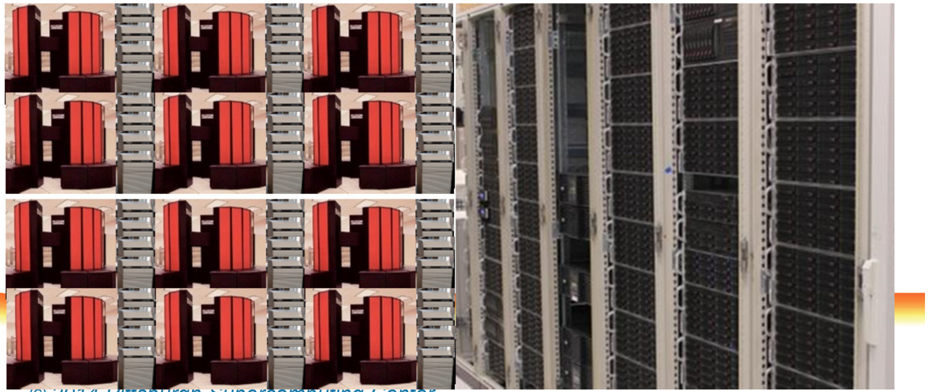
Elaboration of Compute/Storage Systems

First type

- Computer (XMP for history buffs)
- Local file system (DD49s)
- Tape archive (STC, later)

Second type

- A few computers (not XMP!) w/local disks (not DD49)
- Shared file system (SLASH)
- Tape archive, multi-PB



Third type (DSC ⟶ DXC)

- 10’s of diverse computers (not XMP!!!!) w/local disks & SSDs, some remote
- Shared, multi-PB, shared file system & mass-store (SLASH2) **[⟶JRay]**

Characteristic: System Architectures

- *Historical* HPC
  - Type 1: Only IO traffic outside of main computer(s)
  - Type 2: Modification to allow sharing of the tape archive by 2-3 computers (SLASH).
- Data intensive
  - Type 3: Integration of computation & data storage
  - Need to support data traffic between computers and multiple instances and levels of data stores.
  - Performance requirements drive mass-store transition from tape to disk
- PSC/DXC
  - Blacklight, Sherlock & non-DXC systems each with local storage
  - Coupled into the shared, 5.5 TB DSC/DXC storage systems
  - SLASH2 file systems spanning multiple sites

Characteristic: Interconnect Fabrics

- *Historical* HPC
  - Within a computer:
    - Very high bandwidth, low latency, uniformity
    - Topologies: largely fat tree & mesh
  - Within a facility (between computers & storage)
    - Medium bandwidth, latency tolerant (excluding Exascale *burst buffers*)
    - Simple patterns
- Data intensive
  - Within small clusters: as above (limited to single switch)
  - High bandwidth between computers & storage:
- PSC/DXC
  - Higher host IO bandwidth (Blacklight & Sherlock inadequate)
  - Topology: vetting alternative routing algorithms (w/Intel)
  - Extending large data transfers to remote user groups' systems
  - Multiple protocols

Characteristic: Services

- *Historical* HPC
  - Software & minimal: compilers, libraries & application packages.
- Data Intensive
  - Software
    - Data base and analysis tools
    - Workflow & data movement support
  - Operations
    - Semi-stable dedicated functionality (available database systems with data resident for extended periods to avoid startup latencies)
  - Superior Hadoop capabilities
- PSC/DXC
  - Substantial: as requested by DXC's (invaluable) collaborating, new community, research groups! **[→Nick]**

Outline

1. Motivation & Architecture	Michael J. Levine
2. Pilot Applications	Nick Nystrom
3. SLASH2 and Systems Software	J Ray Scott
4. Introduction to *Bridges*	Nick Nystrom
5. Q&A	

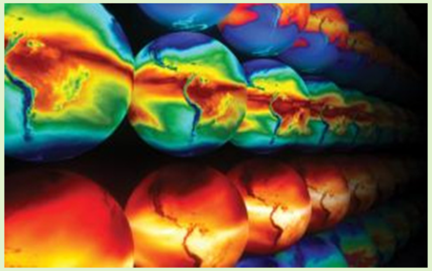
Coming Up…

- Nick Nystrom: Pilot Applications
- J Ray Scott: SLASH2 File System
- Nick Nystrom: Bridges *Preview*

Shifting to Nontraditional Data-Intensive HPC

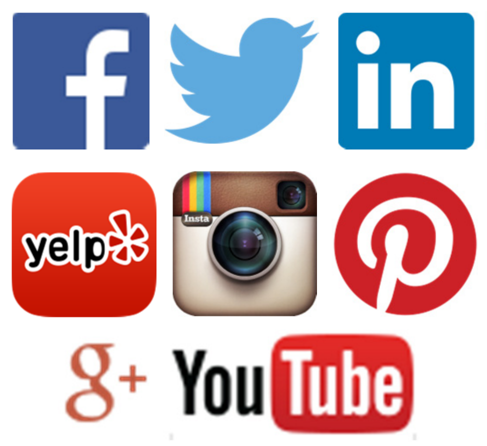
Pan-STARRS telescope
http://pan-starrs.ifa.hawaii.edu/public/

Genome sequencers
(Wikipedia Commons)



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

Unstructured; High Variety, Volume, and Velocity



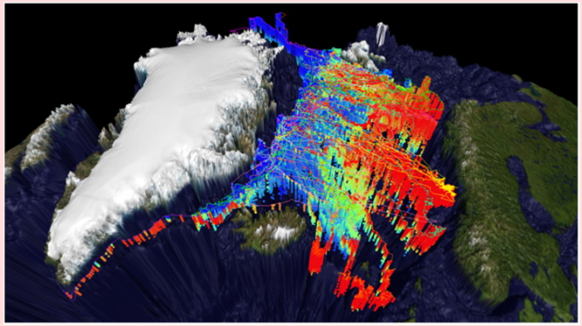
Social networks and the Internet

Video
Wikipedia Commons

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

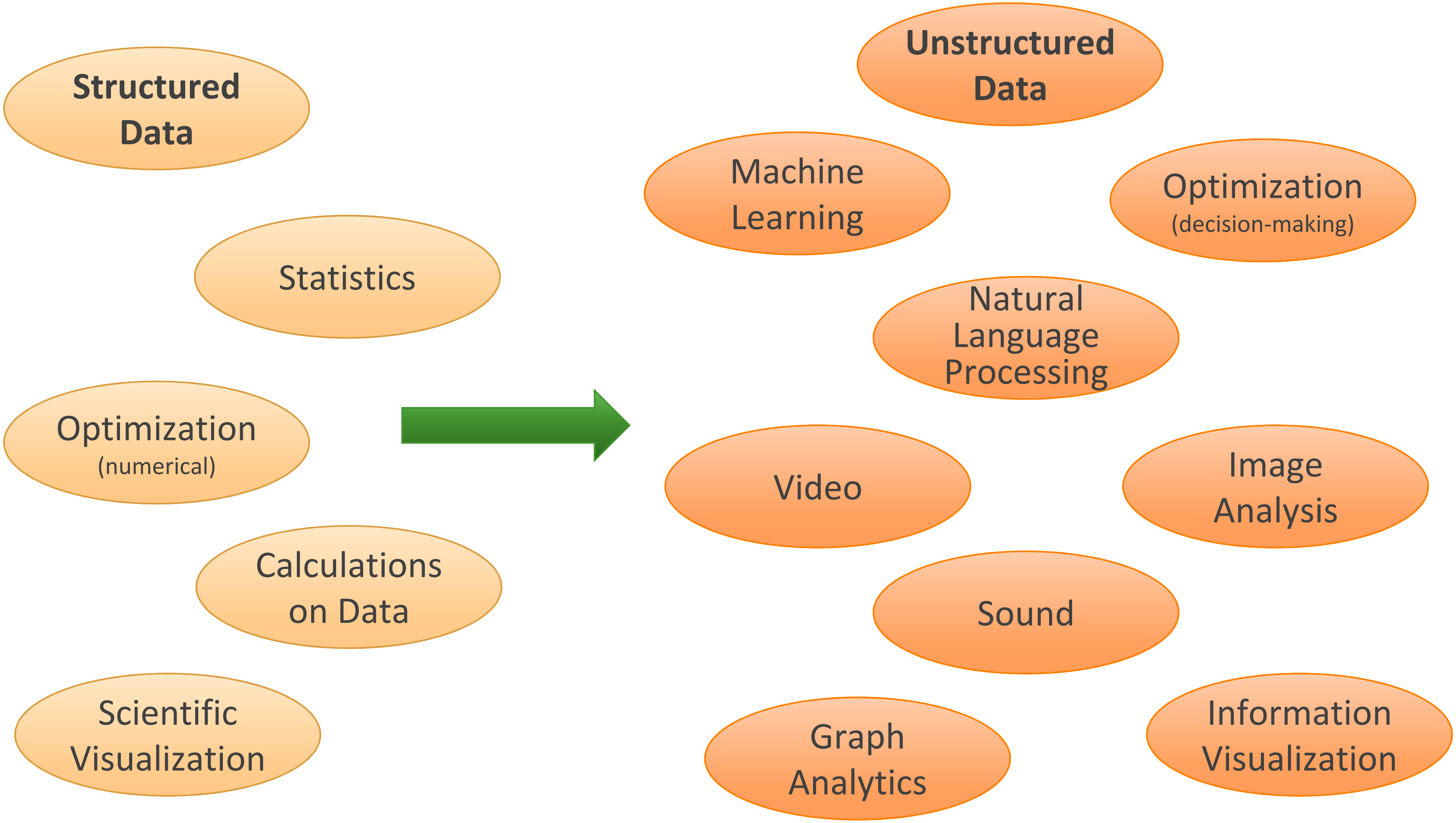
Collections
Horniman museum: http://www.horniman.ac.uk/get_involved/blog/bioblitz-insects-reviewed

Legacy documents
Wikipedia Commons



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

New Areas of Emphasis



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

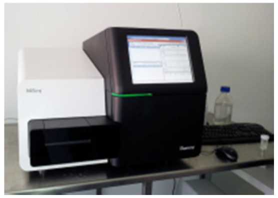
Information Visualization

Pilot Applications

- Pilot applications in data-intensive research areas are being used to motivate, test, demonstrate, and improve building blocks within the DXC project.
- Pilot applications are selected according to their ability to advance research through some combination of:
  - High data variety, velocity, and/or volume
  - Novel approaches to data management or organization
  - Novel approaches to data integration or fusion
  - Integration of data analytic components into workflows
  - Uniqueness with respect to existing DXC pilot applications
- *For each, a PSC specialist works closely with the research group to formulate and implement an effective solution.*

Initial Pilot Applications

- *Pittsburgh Genome Resource Repository / Identifying changes in gene pathways that cause tumors*
  - Michael Becich, Rebecca Crowley, et al., Univ. of Pittsburgh Department of Biomedical Informatics



- *Semantic understanding of large, multimedia datasets*
  - Alex Hauptmann et al., CMU School of Computer Science
- *Exploring and understanding the universe*
  - David Halstead et al., National Radio Astronomy Observatory

- *Enabling bioinformatic workflows*
  - Anton Nekrutenko, Penn State University, and James Taylor, Johns Hopkins University
- *Data integration and fusion for world history*
  - Vladimir Zadorozhny and Patrick Manning, University of Pittsburgh School of Information Sciences and History Department / World History Data Center

Additional Pilot Applications

- *Center for Causal Discovery (BD2K)*
  - Greg Cooper, Univ. of Pittsburgh Department of Biomedical Informatics
- *Recognizing textual entailment*
  - José Fernando Vega, University of Puerto Rico at Mayagüez ECE
- *Constructing the connectome (detailed structure of the brain)*
  - Art Wetzel, PSC / NRBSC
- *Optimization for systems with incomplete knowledge*
  - Tuomas Sandholm, CMU Computer Science and Machine Learning
- *High-performance semantic databases*
  - Nick Nystrom, PSC
- *Analysis of financial markets*
  - Mao Ye, University of Illinois College of Business
- *Aligning medieval manuscripts*
  - David Birnbaum, Univ. of Pittsburgh Slavic Languages & Literatures

Pittsburgh Genome Resource Repository

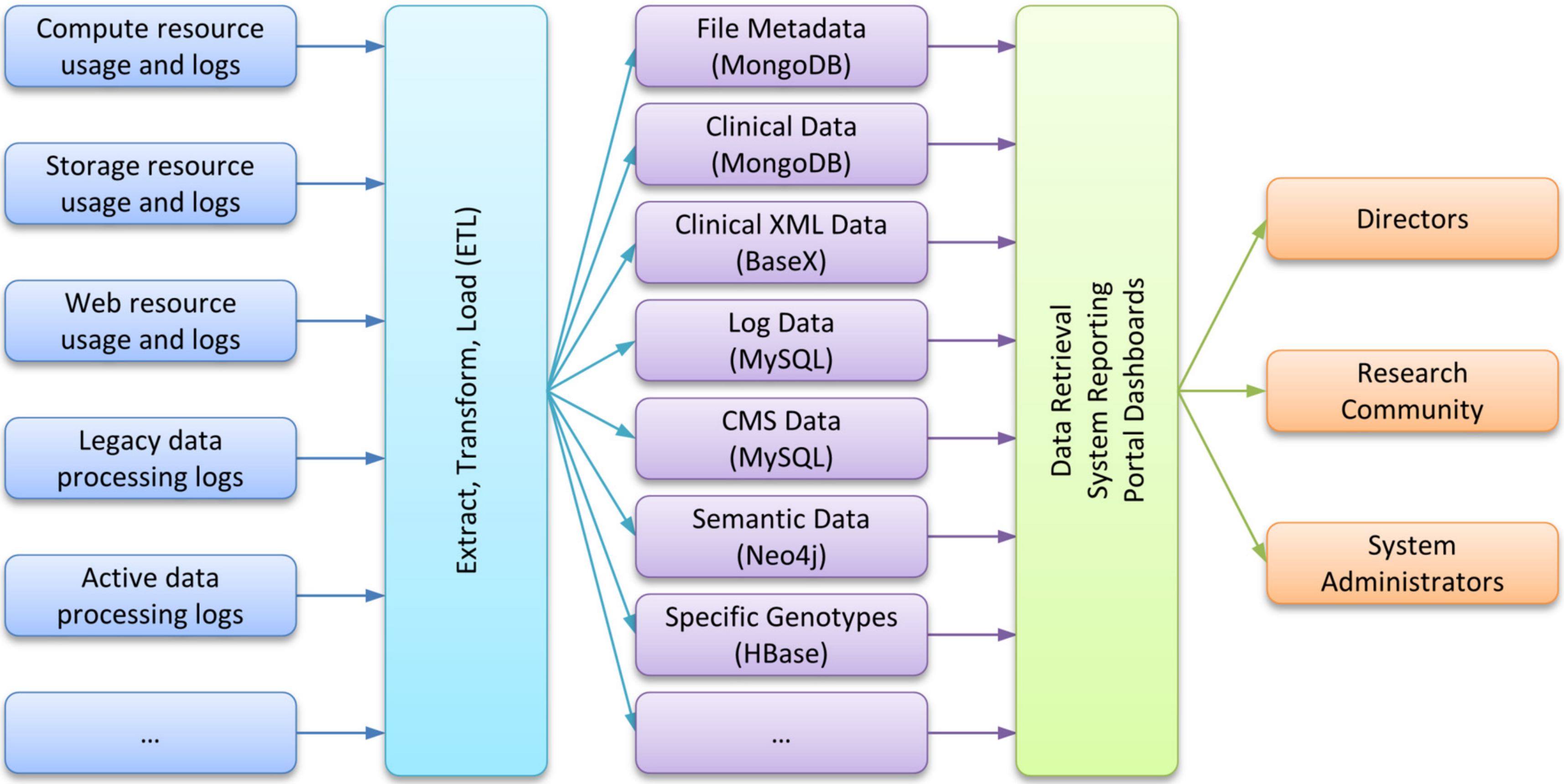
Identifying Changes in Gene Pathways that Cause Tumors

Michael Becich, Rebecca Crowley, Mike Barmada, Xinghua Lu, et al.;
Univ. of Pittsburgh Department of Biomedical Informatics

- Transferring the Cancer Genome Atlas (TCGA) from the Central Cancer Genome Hub (CGHub) at San Diego to DXC
  - PSC helped DBMI to optimize data transfers from CGHub
  - Currently ~1 PB
- Remote SLASH2 instance at Pitt DBMI
  - Data distributed between DXC and Pitt DBMI
- Analytics running on Blacklight
- Developing RDF representation of TCGA
- Now analyzing mesothelioma data obtained from high-throughput gene sequencing

A Database Approach to Genomics

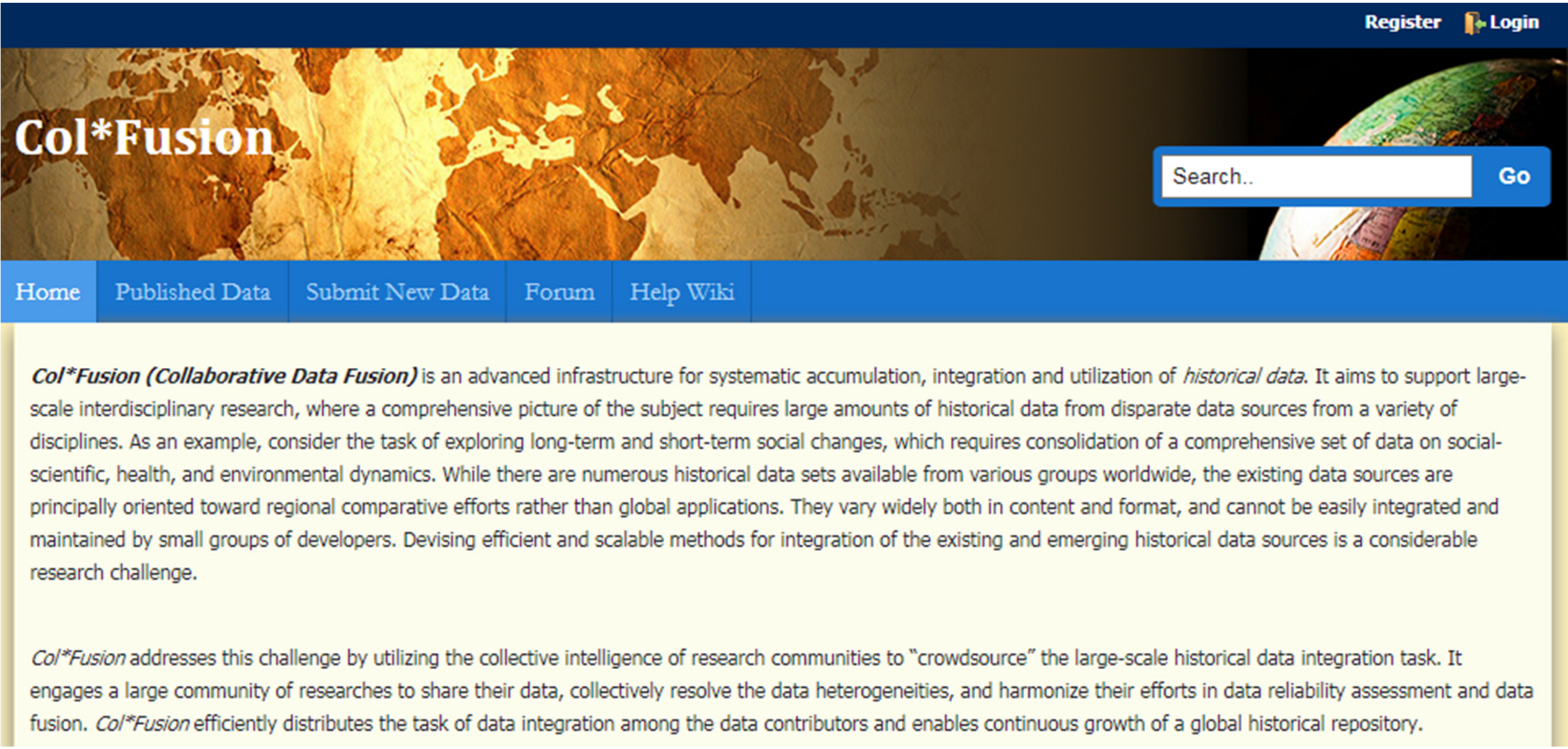
"Polyglot Persistence"



Compute resource usage and logs
Storage resource usage and logs
Web resource usage and logs
Legacy data processing logs
Active data processing logs
…
Extract, Transform, Load (ETL)
File Metadata (MongoDB)
Clinical Data (MongoDB)
Clinical XML Data (BaseX)
Log Data (MySQL)
CMS Data (MySQL)
Semantic Data (Neo4j)
Specific Genotypes (HBase)
…
Data Retrieval
System Reporting
Portal Dashboards
Directors
Research Community
System Administrators

Data integration and fusion for world history

Vladimir Zadorozhny, Patrick Manning, and Evgeny Karataev, Univ. of Pittsburgh



Col*Fusion (Collaborative Data Fusion) is an advanced infrastructure for systematic accumulation, integration and utilization of *historical data*. It aims to support large-scale interdisciplinary research, where a comprehensive picture of the subject requires large amounts of historical data from disparate data sources from a variety of disciplines. As an example, consider the task of exploring long-term and short-term social changes, which requires consolidation of a comprehensive set of data on social-scientific, health, and environmental dynamics. While there are numerous historical data sets available from various groups worldwide, the existing data sources are principally oriented toward regional comparative efforts rather than global applications. They vary widely both in content and format, and cannot be easily integrated and maintained by small groups of developers. Devising efficient and scalable methods for integration of the existing and emerging historical data sources is a considerable research challenge.

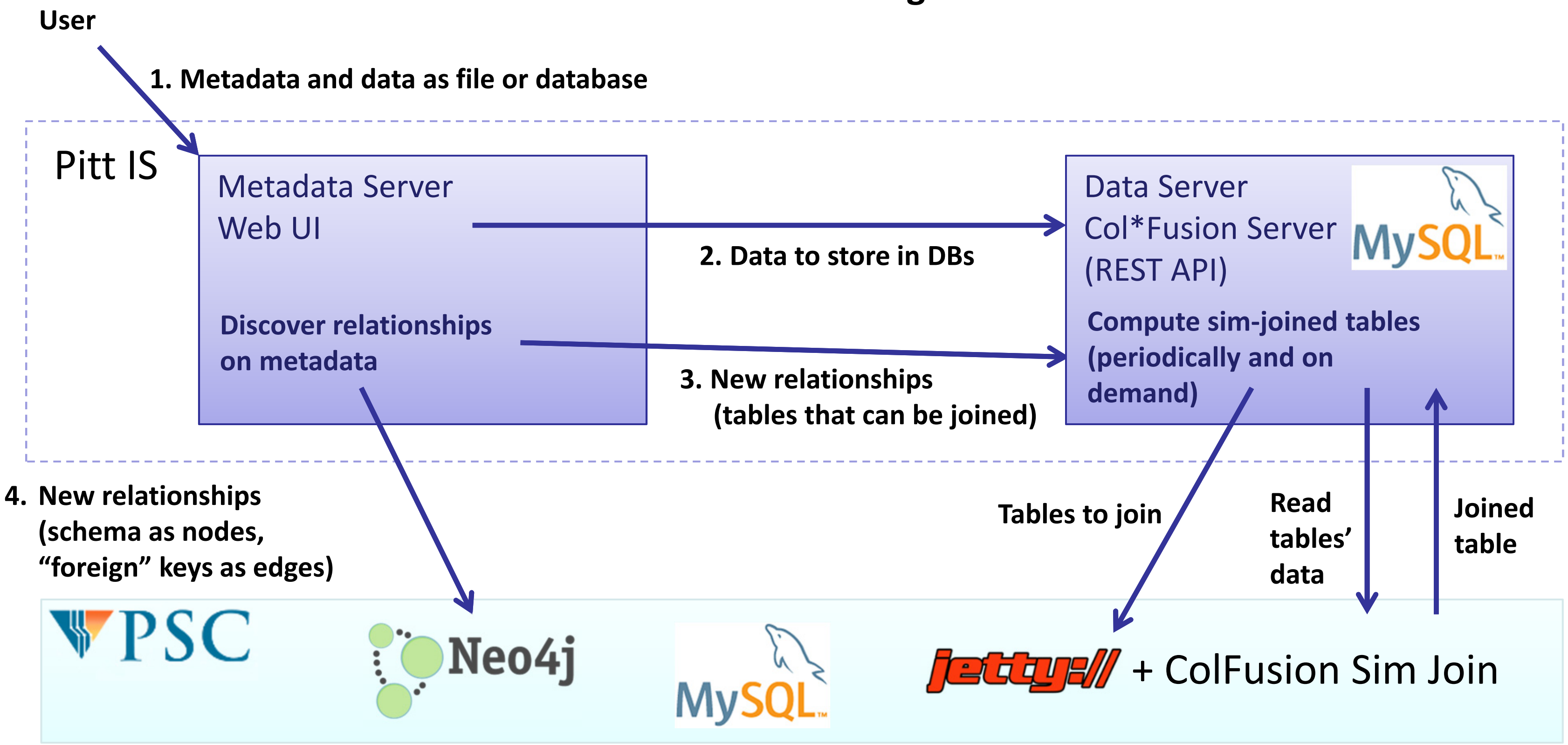
*Col*Fusion* addresses this challenge by utilizing the collective intelligence of research communities to "crowdsource" the large-scale historical data integration task. It engages a large community of researches to share their data, collectively resolve the data heterogeneities, and harmonize their efforts in data reliability assessment and data fusion. *Col*Fusion* efficiently distributes the task of data integration among the data contributors and enables continuous growth of a global historical repository.

Data Integration and Fusion for World History

A Distributed, Crowdsourcing Workflow

Vladimir Zadorozhny, Patrick Manning, and Evgeny Karataev, Univ. of Pittsburgh

Use Case: Data Ingest


User
1. Metadata and data as file or database
Pitt IS
Metadata Server
Web UI
Discover relationships on metadata
2. Data to store in DBs
3. New relationships (tables that can be joined)
Data Server
Col*Fusion Server
(REST API)
MySQL
Compute sim-joined tables (periodically and on demand)
4. New relationships (schema as nodes, "foreign" keys as edges)
Tables to join
Read tables' data
Joined table
PSC
Neo4j
MySQL
jetty:// + ColFusion Sim Join

Database and Web Server Nodes

- Dedicated database nodes will power persistent relational and NoSQL databases to support sophisticated data management and data-driven workflows.
  - High-performance local storage (SSDs)
  - High-capacity (HDDs and distributed databases)
  - Hadoop (HBase, Cassandra, Spark, ...)
- Dedicated web server nodes will enable distributed, service-oriented architectures.
  - High-bandwidth connections to XSEDE and the Internet

DXC Database Servers Address A Range of Needs

- High-performance storage
  - Common trade-offs: capacity vs. IOPs, performance vs. replication vs. cost
- Computational capacity and capability
  - Capacity: for largely independent operations (Hadoop, etc.)
  - Capability: for operations that aren't readily partitioned
    - Examples: graph analytics, in-memory databases, applications not written for distributed memory
- Effective data organization and manipulation
  - Files
  - Databases
    - Relational, column, document, graph, key/value, XML, hybrids
    - Disk-based vs. in-memory
      - Higher IOPs of other types of persistent storage
        - » Write vs. read performance, durability, capacity
  - Hybrid approaches
    - Databases for metadata, some data, analysis products, and to support querying
    - Files to support bulk storage of large data

Adding strong database support allows HPC to be applied to large-scale data analytics.

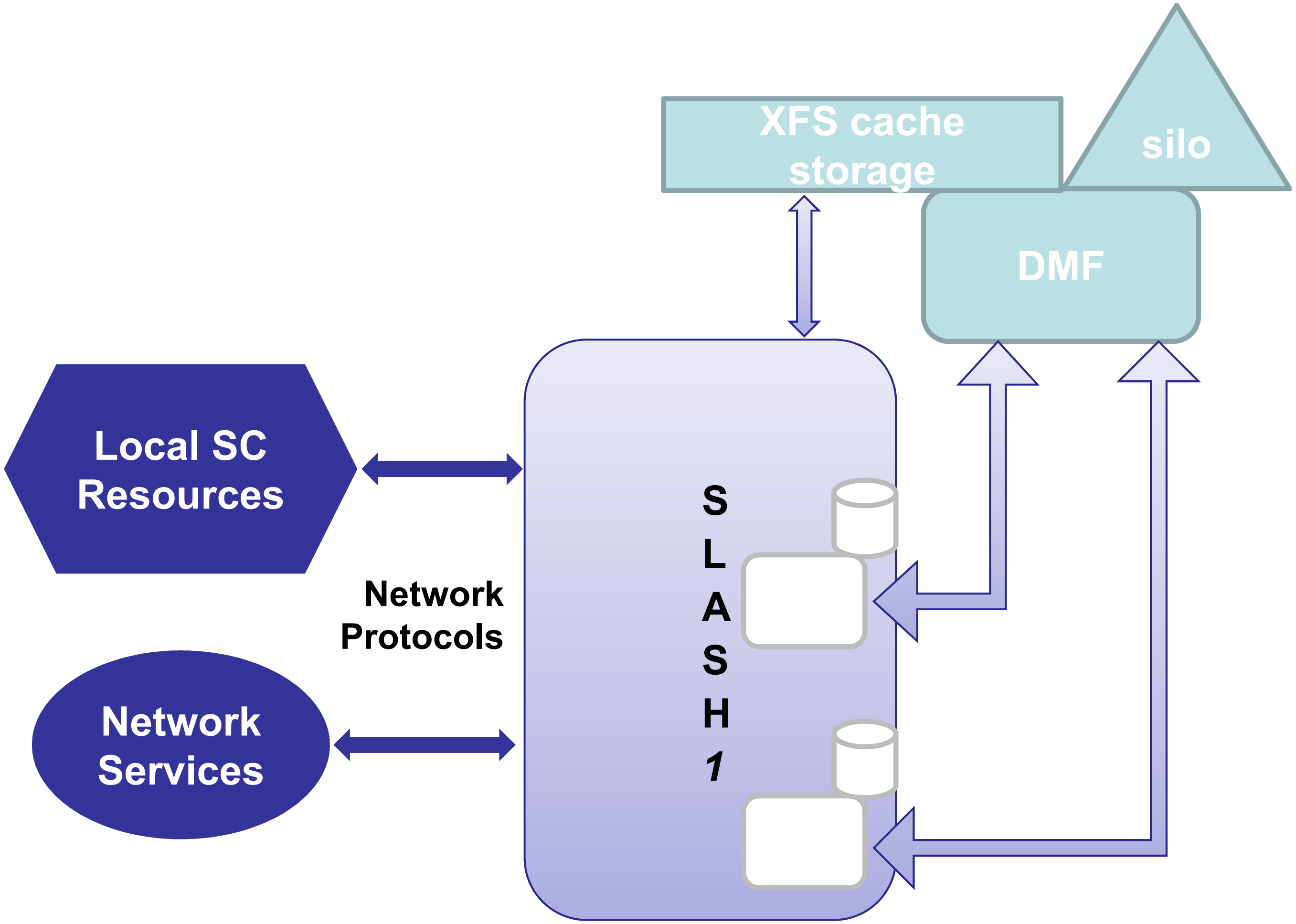
Applications Drive the Building Blocks

- For example:
  - Software architectures for data-intensive applications
  - Database frameworks, especially for polyglot persistence
  - Software and database frameworks for data integration and fusion; cross-dataset analyses
  - Database-driven workflow execution across heterogeneous resources
  - Hardware configurations for optimal performance and flexibility

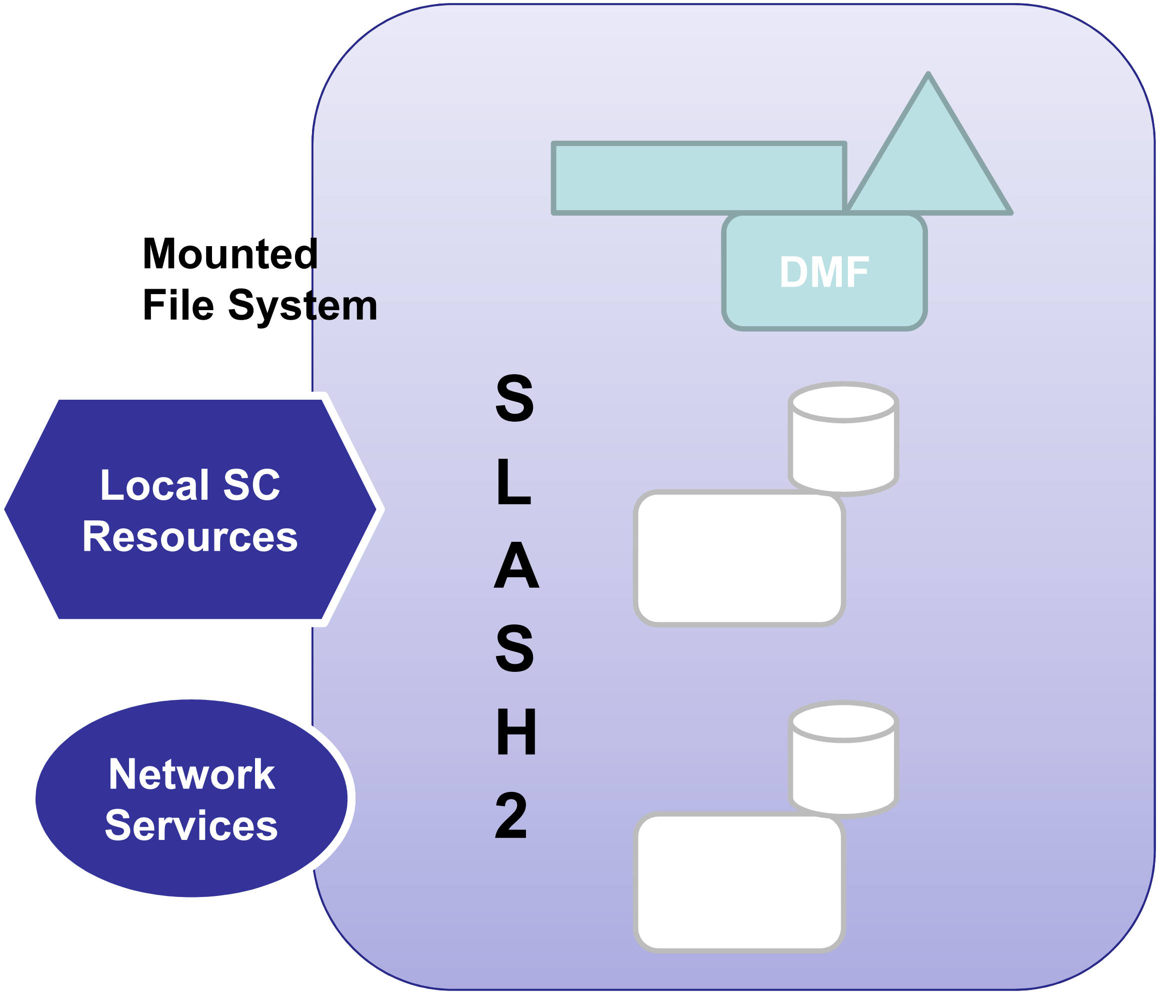
Outline

1. Motivation & Architecture	Michael J. Levine
2. Pilot Applications	Nick Nystrom
3. SLASH2 and Systems Software	J Ray Scott
4. Introduction to *Bridges*	Nick Nystrom
5. Q&A	

PSC Archiving Architecture v1

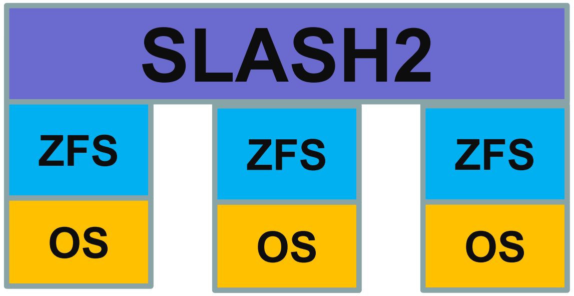


PSC Archiving Architecture v2



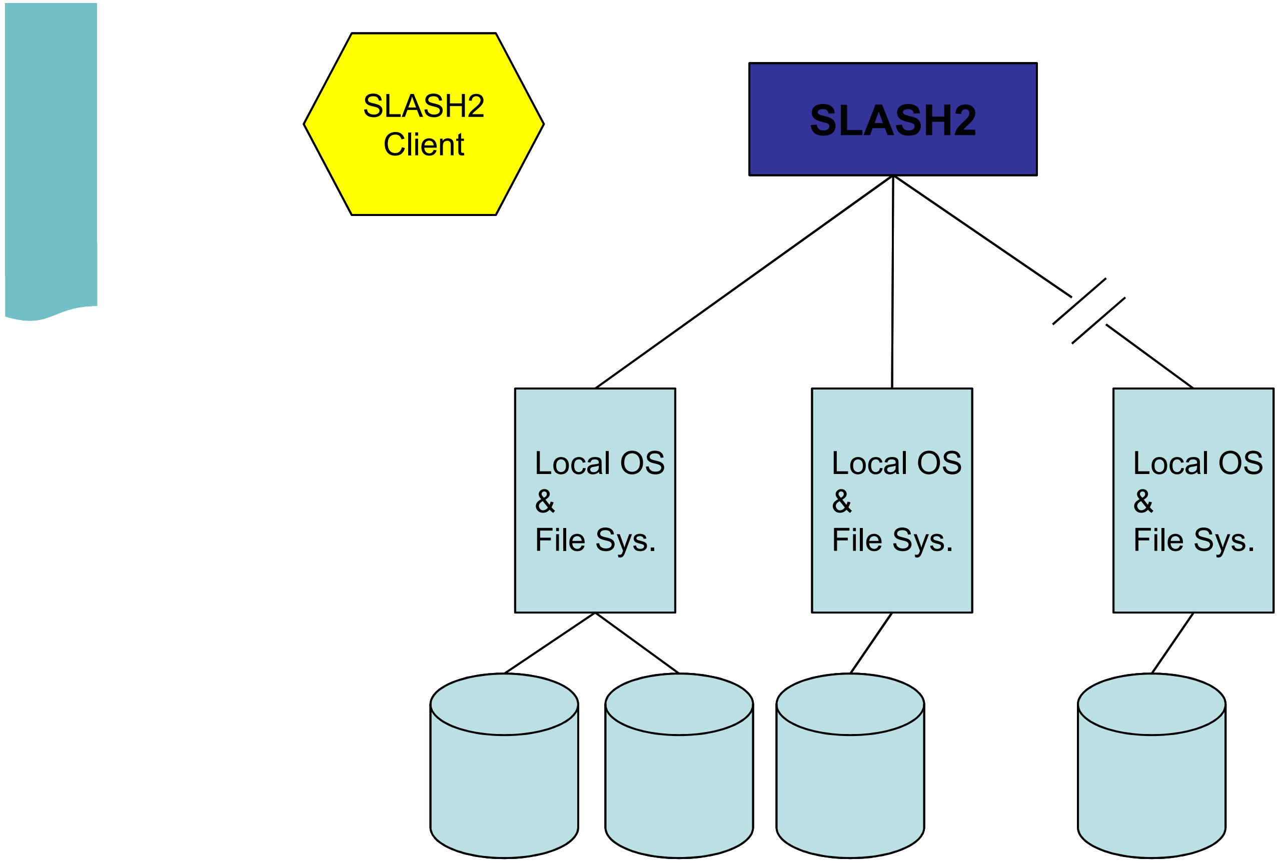
Data Supercell Software Architecture

- FreeBSD Operating System
- ZFS File System
- = One File System Namespace per Building Block
- SLASH2
  - PSC Developed
  - Wide Area Network File System
  - Fuse Mount Client
  - Replication
  - Error Checking

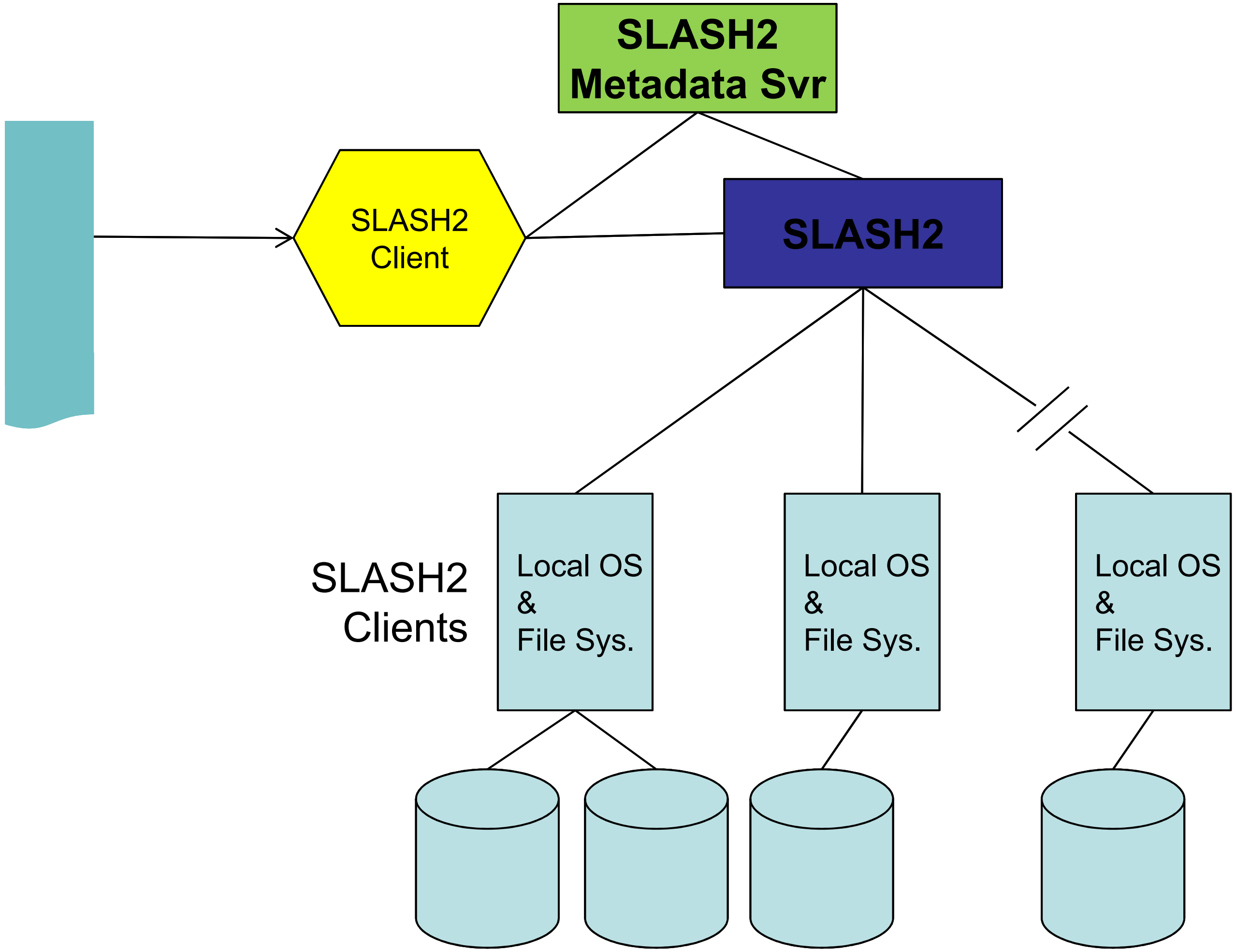


Server
Building
Block

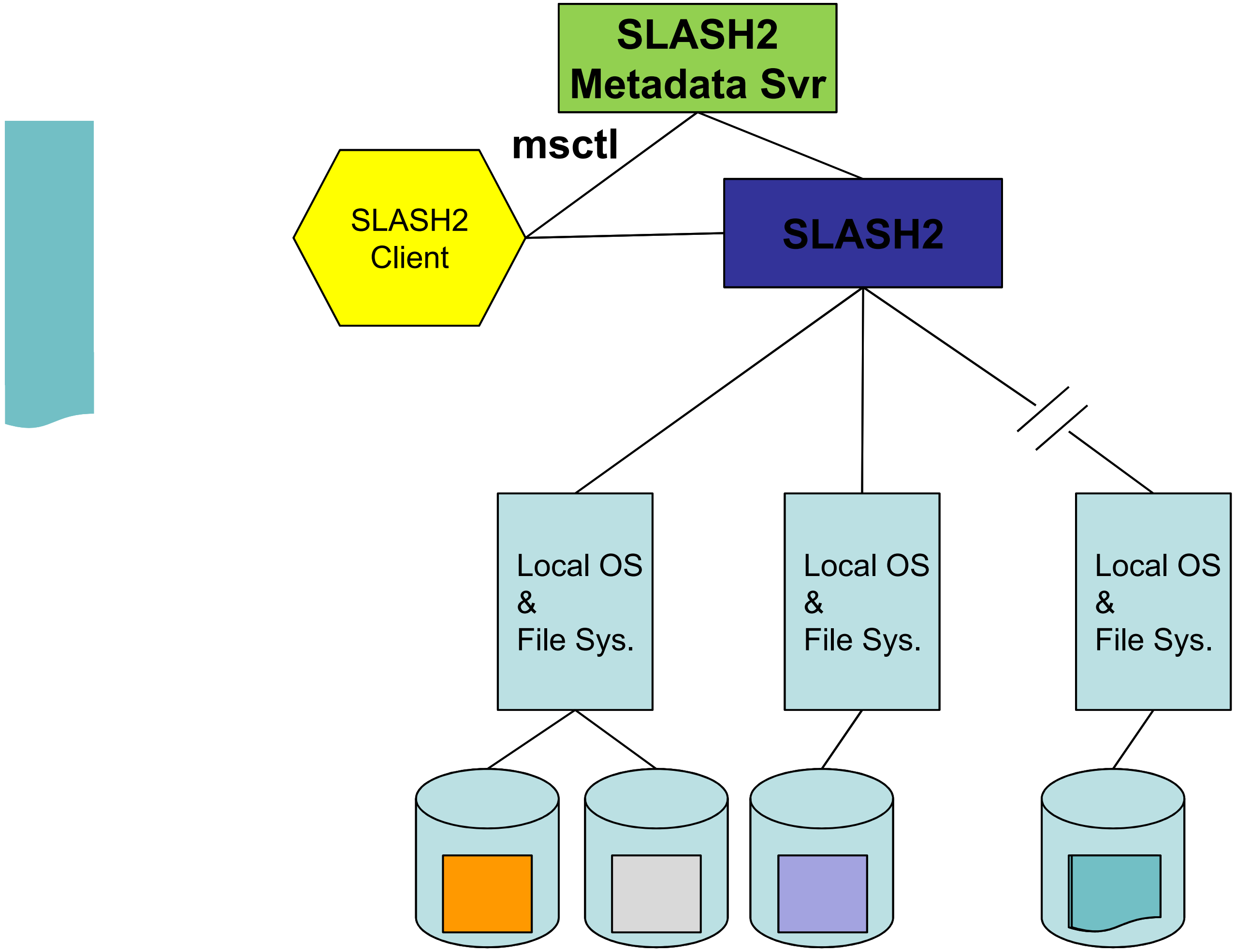
SLASH2 File Residency and Replication



SLASH2 File Residency and Replication



SLASH2 File Residency and Replication



DSC Design Goals

- Cost
  - Capital
  - Operations
- Reliability
  - Redundancy
  - Multiple layers of RAID and checksums
  - Remote Management
- Scalability
  - Leave open the possibility of tape or any other technology
- Performance – “Faster than tape”
- Space

DXC is a more complex environment

- File system performance
- Hardware testing
- Application requirements
- Multi-Site support
- Embedded execution
- Data tagging tools
- Administrative tools

SLASH2 Performance Improvements

- Improve performance based on use cases that fall outside the DSC archiver use case
  - GridFTP
  - GeneTorrent
  - Database
- Read-ahead code.
  - 20% improvement in performance.
  - We can document some of this since we've seen the numbers reported by the speedpage improve
- Enhanced queued activity support
  - SDN
  - workload management

Storage Hardware
Performance Testing and Tuning

- tests and numerous graphs showing performance of different combinations of disks, disk shelves, and server speeds and configurations.
- SSD wear support
  - . TRIM
  - . Modify SLASH2 MDS to be more SSD friendly
- Building block: Storage performance tools and framework

Complex Access Patterns

- Looking at the random i/o patterns of client tools like the bioinformatics tool “GeneTorrent” and updating SLASH2 to better handle these access patterns
- Building Block: Parallel file transport tool
  - out of this work has grown an interesting parallel file copy utility
  - Drop in replacement for rsync
  - 10x transfer improvement

Multi-site Support

- Multiple metadata servers
- SLASH2 <-> local file multi-site file import/export
- Workflow integration
- Cross site UID mapping
  - Security
  - Federated Authentication
- Enhanced access controls
  - E.g. read/delete only
- Public cloud support
- Building block : SCAMPI filesystem

Embedded Execution

- Enable the ability for executables to run on the file system servers
- Initial investigation with graphic processing

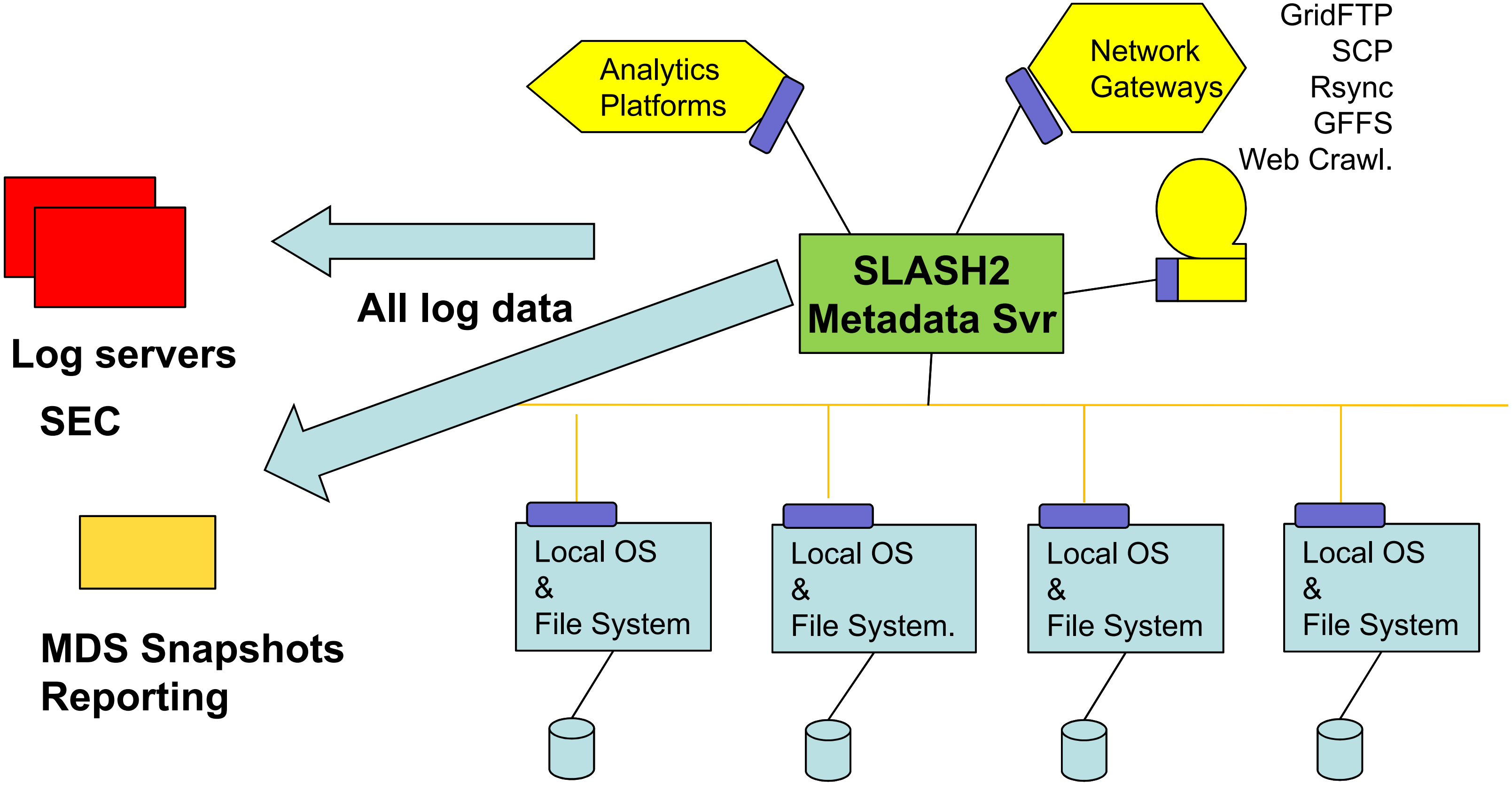
Data Tagging Tools

- Integrate Metadata Management for content into file system
- Application driven

Administrative tools

- Extensive metering built into SLASH2
- Activity logging
  - Files
  - Transfer tools
- Event management
- Performance monitoring
- Inventory management
- Reporting
- System Administration
- Private cloud integration
  - Openstack

Administration



Outline

1. Motivation & Architecture	Michael J. Levine
2. Pilot Applications	Nick Nystrom
3. SLASH2 and Systems Software	J Ray Scott
4. Introduction to *Bridges*	Nick Nystrom
5. Q&A	

BRIDGES

A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE

A Brief Introduction

Pittsburgh is a city of bridges: from its history in steel to its leadership in computer science and biotechnology, between diverse neighborhoods housing its many universities, and at PSC, from science-inspired national cyberinfrastructure to researchers' breakthroughs.

Bridges will be a new XSEDE resource that will integrate advanced memory technologies to empower new communities, bring desktop convenience to HPC, connect to campuses, and intuitively express data-intensive workflows.

The $9.65M *Bridges* 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*

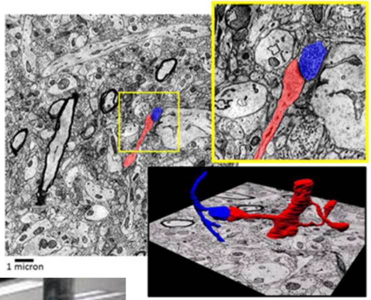
Integrating with the National Advanced Cyberinfrastructure Ecosystem

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

Examples:

Green Bank Telescope, National Radio Astronomy Observatory

High-throughput genome sequencers



Reconstructing brain circuits from high-resolution electron microscopy

Data Infrastructure Building Blocks (DIBBs)

- Data Exacell (DXC)
- Other DIBBs projects

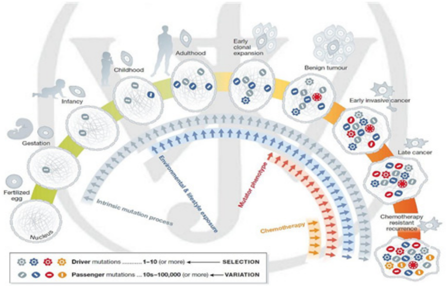
Other ACI projects

Carnegie Mellon University's Gates Center for Computer Science

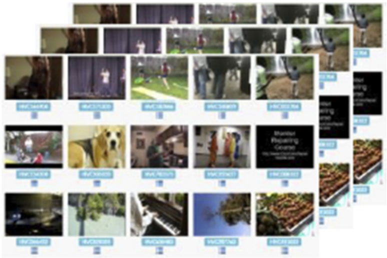
Temple University's new Science, Education, and Research Center

Objectives: Bridging Computational Needs

- Bring HPC to nontraditional users and research communities.
- Allow high-performance computing to be applied effectively to big data.
- Bridge to campuses to ease access and provide burst capability.
- To realize these goals, *Bridges* will be built on 3 tiers of large, coherent shared-memory nodes.


Fertilized egg
Nucleus
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late cancer
Chemotherapy resistant recurrence
Intrinsic mutation process
Environmental & lifestyle exposure
Mutator phenotype
Chemotherapy
Driver mutations .... 1–10 (or more) ← SELECTION
Passenger mutations ... 10s–100,000 (or more) ← VARIATION

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*

- *Bridges* will leverage its large memory for interactivity and to seamlessly support applications through virtualization, gateways, familiar and productive programming environments, and data-driven workflows.

Examples of Potential Applications

- Enabling complex workflows for genome assembly
- Topic modeling from text databases for sociology
- Machine learning and remote data analytics for astronomy
- Event detection in multimedia
- Understanding the brain (fMRI, electron microscopy)
- Agent-based modeling for epidemiology and other fields
- Understanding structure and evolution of materials (APS)
- Decision-making in situations with incomplete information
- Optimizing organ (e.g. kidney) exchange networks
- Causal modeling to interpret big data (esp. biomedical)
- Analysis of financial markets
- Data integration and fusion for world history

Bridging to New Research Communities

- 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.
- Database and web server nodes will provide persistent databases to enable data management, workflows, and distributed applications.
- High-productivity programming languages & environments will let users scale familiar applications and workflows.
- Virtualization will allow users to bring their particular environments and provide interoperability with clouds.

Application-Driven Hardware Architecture

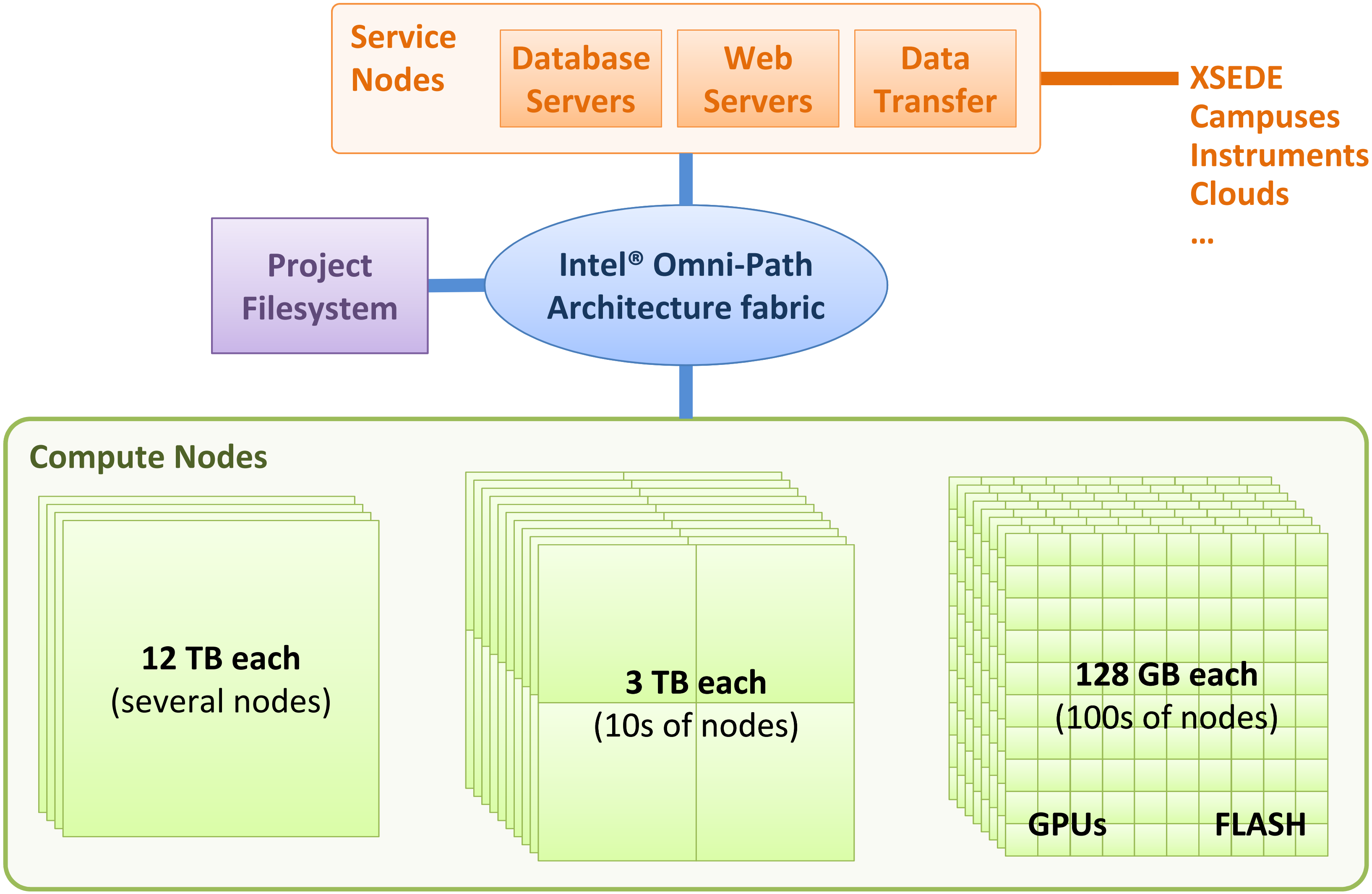
- Data-intensive computation
  - 3 tiers of node types, with each node having large, coherent shared memory: 12TB, 3TB, and 128GB
  - The latest Intel® Xeon® CPUs
  - NVIDIA® Tesla® K80 dual-GPU accelerator and next- generation NVIDIA Tesla GPUs
- Data management and movement
  - Database nodes with fast local storage
  - Web and data transfer nodes with fast networking
  - A shared, parallel, high-performance Project File System
  - Distributed storage for convenience and performance
  - Intel Omni-Path® Architecture fabric

A Hierarchy of Large-Memory Compute Nodes Will Serve a Spectrum of Memory Requirements

- Nodes with 12TB, 3TB, and 128GB each will support data-intensive applications:

Memory per node	Number of nodes	Example applications
12 TB HP Superdome X	Several	Genomics, machine learning, graph analytics, other extreme-memory apps
3 TB HP ProLiant DL580	Tens	Virtualization & interactivity including large-scale visualization & analytics; mid-spectrum memory-intensive jobs
128 GB HP Apollo 6000	Hundreds	Execution of most components of workflows, interactivity, Hadoop, and capacity computing

High-Level Architecture



Service Nodes

Database Servers

Web Servers

Data Transfer

XSEDE
Campuses
Instruments
Clouds
…

Project Filesystem

Intel® Omni-Path Architecture fabric

Compute Nodes

12 TB each
(several nodes)

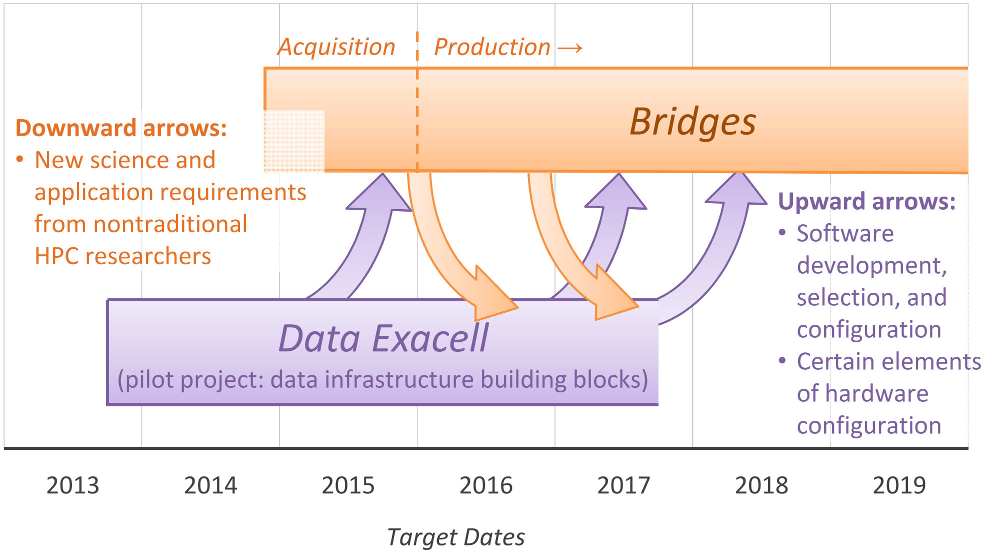
3 TB each
(10s of nodes)

128 GB each
(100s of nodes)

GPUs **FLASH**

Bridges and the *Data Exacell* : A Valuable Engineering Lifecycle

- Hardware and software “building blocks” developed at PSC through its *Data Exacell* (*DXC*) pilot project, funded by NSF’s DIBBs program, will enable new application architectures on *Bridges* and convenient, high-performance data movement between *Bridges* and users, campuses, and instruments.
- *Bridges* and *DXC* will provide complementary roles for production and application prototyping.



Acquisition
Production →
Bridges
Downward arrows:
• New science and application requirements from nontraditional HPC researchers
Upward arrows:
• Software development, selection, and configuration
• Certain elements of hardware configuration
Data Exacell
(pilot project: data infrastructure building blocks)
2013
2014
2015
2016
2017
2018
2019
Target Dates

Bridges Target Schedule

- Acquisition
  - Begins December 2014
  - Construction to begin October 2015
    - Will allow for including important new technologies
  - Early user period in late 2015
- Production
  - January 2016

For Additional Information

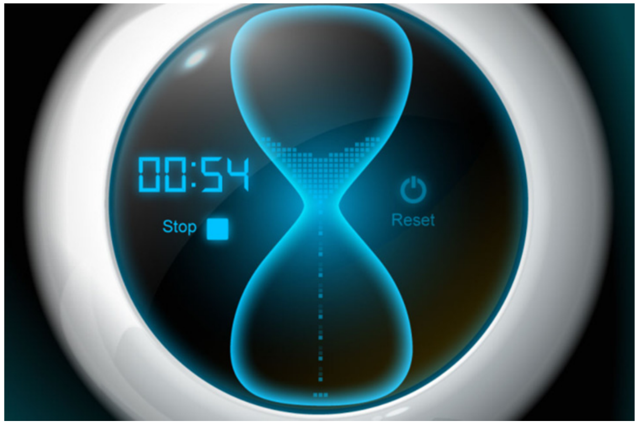
Project website: www.psc.edu/bridges

Bridges PI: Nick Nystrom, nystrom@psc.edu

Additional Content

Interactivity

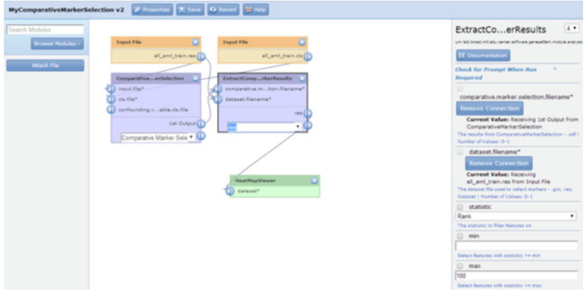
- *Interactivity is the feature most frequently requested by nontraditional HPC communities.*



- Interactivity provides immediate feedback for doing exploratory data analytics and testing hypotheses.
- *Bridges* will offer interactivity through a combination of virtualization for lighter-weight applications and dedicated nodes for more demanding ones.

Gateways and Tools for Building Them

Gateways will provide easy-to-use access to *Bridges'* HPC and data resources, allowing users to launch jobs, orchestrate complex workflows and manage data from their web browsers.



comparative.marker.selection.filename*

Current Value: Receiving 1st Output from ComparativeMarkerSelection

The results from ComparativeMarkerSelection - .odf | Number of Values: 0-1

dataset.filename*

Current Value: Receiving all_aml_train.res from Input File

The dataset file used to select markers - .gct, .res, Dataset | Number of Values: 0-1

statistic: Rank

The statistic to filter features on

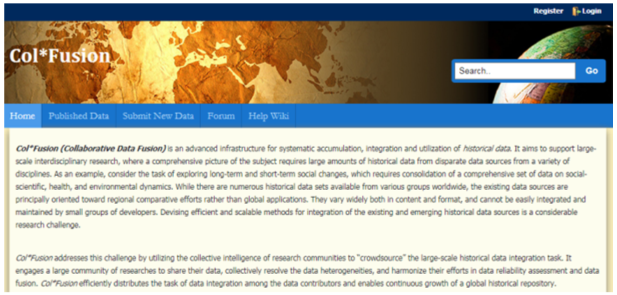
min:

Select features with statistic >= min

max: 100

Select features with statistic <= max

Interactive pipeline creation in GenePattern (Broad Institute)



Col*Fusion (Collaborative Data Fusion) is an advanced infrastructure for systematic accumulation, integration and utilization of *historical data*. It aims to support large-scale interdisciplinary research, where a comprehensive picture of the subject requires large amounts of historical data from disparate data sources from a variety of disciplines. As an example, consider the task of exploring long-term and short-term social changes, which requires consolidation of a comprehensive set of data on social-scientific, health, and environmental dynamics. While there are numerous historical data sets available from various groups worldwide, the existing data sources are principally oriented toward regional comparative efforts rather than global applications. They vary widely both in content and format, and cannot be easily integrated and maintained by small groups of developers. Devising efficient and scalable methods for integration of the existing and emerging historical data sources is a considerable research challenge.

*Col*Fusion* addresses this challenge by utilizing the collective intelligence of research communities to "crowdsource" the large-scale historical data integration task. It engages a large community of researches to share their data, collectively resolve the data heterogeneities, and harmonize their efforts in data reliability assessment and data fusion. *Col*Fusion* efficiently distributes the task of data integration among the data contributors and enables continuous growth of a global historical repository.

Col*Fusion portal for the systematic accumulation, integration, and utilization of historical data, from http://colfusion.exp.sis.pitt.edu/colfusion/

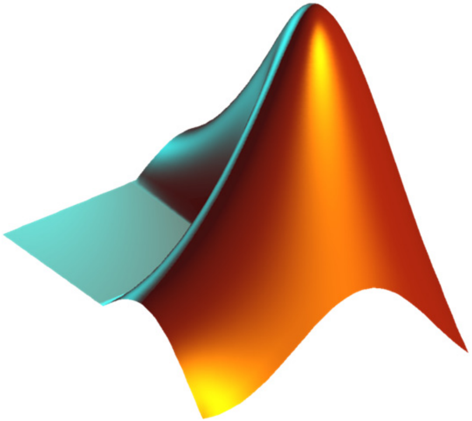
Download sites for MEGA-6 (Molecular Evolutionary Genetic Analysis), from www.megasoftware.net

Virtualization and Containers

Users will be able to import their entire computational environments into *Bridges*, conferring ease of use, reproducibility, and interoperability with cloud services.

High-Productivity Programming

Bridges will feature high-productivity programming languages and tools.



Supporting the Hadoop Ecosystem

- A shared, flash-based array will accelerate Hadoop applications.
- 128GB nodes can be configured to accommodate additional demand.

Cassandra
HIVE
PROJECT PEGASUS
APACHE
HBASE
mahout
Spark

Database and Web Server Nodes

- Dedicated database nodes will power persistent relational and NoSQL databases to support sophisticated data management and data-driven workflows.
  - High-performance local storage (SSDs)
- Dedicated web server nodes will enable distributed, service-oriented architectures.
  - High-bandwidth connections to XSEDE and the Internet

GPUs

- NVIDIA® Tesla® K80 dual-GPU accelerator and next-generation NVIDIA Tesla GPUs will accelerate a wide range of research through:
  - A large portfolio of existing, accelerated applications
  - Drop-in libraries
  - Easy-to-use OpenACC directives
  - The CUDA programming platform

Hundreds of supported applications and even more research applications will benefit from *Bridges*' state-of-the-art GPUs.

Parallel and Distributed Storage for High, Consistent I/O Performance

- *Bridges'* Project File System, built from technologies developed for PSC's *Data Supercell* and *Data Exacell*, will provide a central, high-performance, parallel filesystem.
- Distributed, node-local storage will provide maximum flexibility and performance for data-intensive applications.
- A shared, flash-based array will accelerate Hadoop-based applications and databases.

Interconnect

- The Intel® Omni-Path Architecture fabric will provide *Bridges* and its users with:
  - The highest-bandwidth internal network
  - Valuable optimizations for MPI and other communications
  - Early access to this important, forward-looking new technology

How Does *Bridges* Differ from *Comet*? (1)

- Interactivity
  To make HPC familiar and productive for nontraditional HPC communities, Bridges emphasizes interactivity, which we will implement through shared access to resources (which will be fine for most uses) and dedicated access to nodes when required either for memory or performance.
- Persistent Databases and Distributed Applications
  Bridges will have dedicated nodes to support persistent databases (relational and NoSQL) for effective data management and integration, enabling community data repositories, and supporting data-driven workflows. Related to this will be Bridges' dedicated nodes for supporting gateways and service-oriented architectures, which will interoperate closely with Bridges' database nodes to address requirements we're already seeing from communities in, for example, genomics, history, and imaging.

How Does *Bridges* Differ from *Comet*? (2)

- Larger Memory
  Comet will have four 1.5TB nodes; Bridges will have around that many 12TB nodes, plus around ten times as many 3TB nodes. From our experience, we are aware of applications in genome sequence assembly, causal modeling and cybersecurity (and other applications of graph analytics), and other fields that we expect to make good use of Bridges' 12TB nodes.

- Campus Bridging
  Bridges includes a pilot project (with Temple University, in eastern PA) to develop effective ways to handle burst processing from campuses.

How Does *Bridges* Differ from *Comet*? (3)

- Hadoop
  Bridges and Comet approach Hadoop differently. Bridges will offer a flash array to accelerate Hadoop applications, together with larger storage per node to scale data-intensive Hadoop applications. Comet, on the other hand, has no hardware specifically for accelerating Hadoop, and only 2×200GB per node can be limiting for Hadoop applications.

- Technology
  Bridges will introduce important technologies beyond those to appear in Comet. Qualitatively, the most prominent will be the Intel® Omni-Path Architecture fabric, which will couple all of Bridges' compute, data, database, and web resources at exceptionally high performance. Bridges will also introduce a new generation of Intel CPUs and NVIDIA GPUs, including the forthcoming NVLink GPU interconnect.