The Data Exacell (DXC): *Data Infrastructure Building Blocks for Integrating Analytics with Data Management*

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

August 19, 2014

The Data Exacell (DXC)

- NSF Data Infrastructure Building Blocks (DIBBs) award #1261721
- *A pilot project to create, deploy and test software and hardware implementing functionalities specifically designed to support data-analytic capabilities for data intensive scientific research*
- Extends PSC's
  - Data Supercell (DSC): an 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
- Offers cutting-edge database technologies to enable development of powerful new application architectures
- Pilot applications having diverse data analytic requirements motivate, test, and demonstrate DXC's capabilities

Objectives

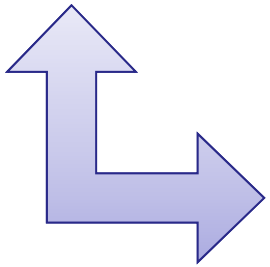
- *The Data Exacell will develop, test, and make available hardware and software building blocks for data-intensive science. Several anticipated examples include:*
  - SLASH2: file system building blocks for remote deployment for easy-to-use, high-performance, distributed data management
  - Software building blocks to leverage the tight coupling of powerful analytical engines with low-latency, high-bandwidth storage
  - Software building blocks to facilitate distributed workflows between the DXC and campus resources
  - Database building blocks for incorporation of relational and NoSQL database technologies into applications and workflows to allow cross-domain data integration, improve data management and provenance, ease use, and leverage distributed resources

High-Level Architecture

Data Exacell

Sherlock

Blacklight



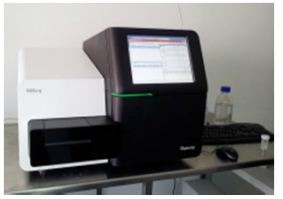
Campuses

XSEDE

Instruments

Enabling and Accelerating Data-Analytic Applications

- *Data analytics requires a system architecture that is very different from traditional HPC.*
- The following pilot applications were selected for the diverse demands they place on data-analytic and storage systems:
  - *Identifying changes in gene pathways that cause tumors*
    - Michael Becich, Rebecca Crowley, et al.,
      University 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
  - *Data integration and fusion for world history*
    - Vladimir Zadorozhny and Patrick Manning, Univ. of Pittsburgh
      School of Information Sciences and World History Data Center



SLASH2 Distributed Filesystem

- An open-source, WAN-friendly distributed file system featuring multi-residency at the file chunk level, inline checksum verification, ...
- Provides:
  - geographical replication for access locality
  - replicas for valuable data
  - the continuing emergence of cloud computing and the need for universal interfaces
  - research collaboration
  - Federates other filesystems: ZFS, Lustre, GPFS, NFS, etc.
- www.psc.edu/slash2

DXC Extensions

- Extends PSC's SLASH2 (file system software) and Data Supercell (hardware storage system) to enable:
  - *Collaborative data analytics across researchers' sites and datasets*
    - Cross-domain analytics
    - Distributed, web-based workflows
  - *Tightly-coupled computational resources for data analytics*
    - Uniquely large shared memory
    - Purpose-built graph capabilities
  - *Improved performance*
  - *Enhanced ease of use*

SLASH2: Internal, Distributed Architecture

- 4 types of component sub-systems, each with its own hardware and software:
  - Metadata Servers (MDS) maintain the namespace with POSIX and SLASH2 metadata including block maps, replication tables, data checksums, etc.
  - Gateway Service Nodes (GSN) provide user access
  - I/O Service Nodes (IOS) encapsulate different POSIX-compliant file systems as backing stores
    - Physically configured as Storage Building Blocks (SBBs)
    - Allow for vendor heterogeneity
  - Administrative Servers provide system administration and monitoring functions

SLASH2: I/O Service Nodes

- Stand-alone IOS is used for a single storage server which contributes storage to a SLASH2 file system
- Cluster No Share Service logically binds a set of stand-alone nodes for striping or file-wise load balancing and supports file replication therein
- Archival IOS is used for storage systems requiring an arbitrary retrieval period
- Parallel IOS is useful for accessing parallel file systems via multiple endpoints.
  - *In a typical configuration, the SLASH2 I/O service runs a parallel file system client. SLASH2 does not replace systems' native clients but operates within a parallel file system for parallel data staging.*
- Interfaces to cloud storage

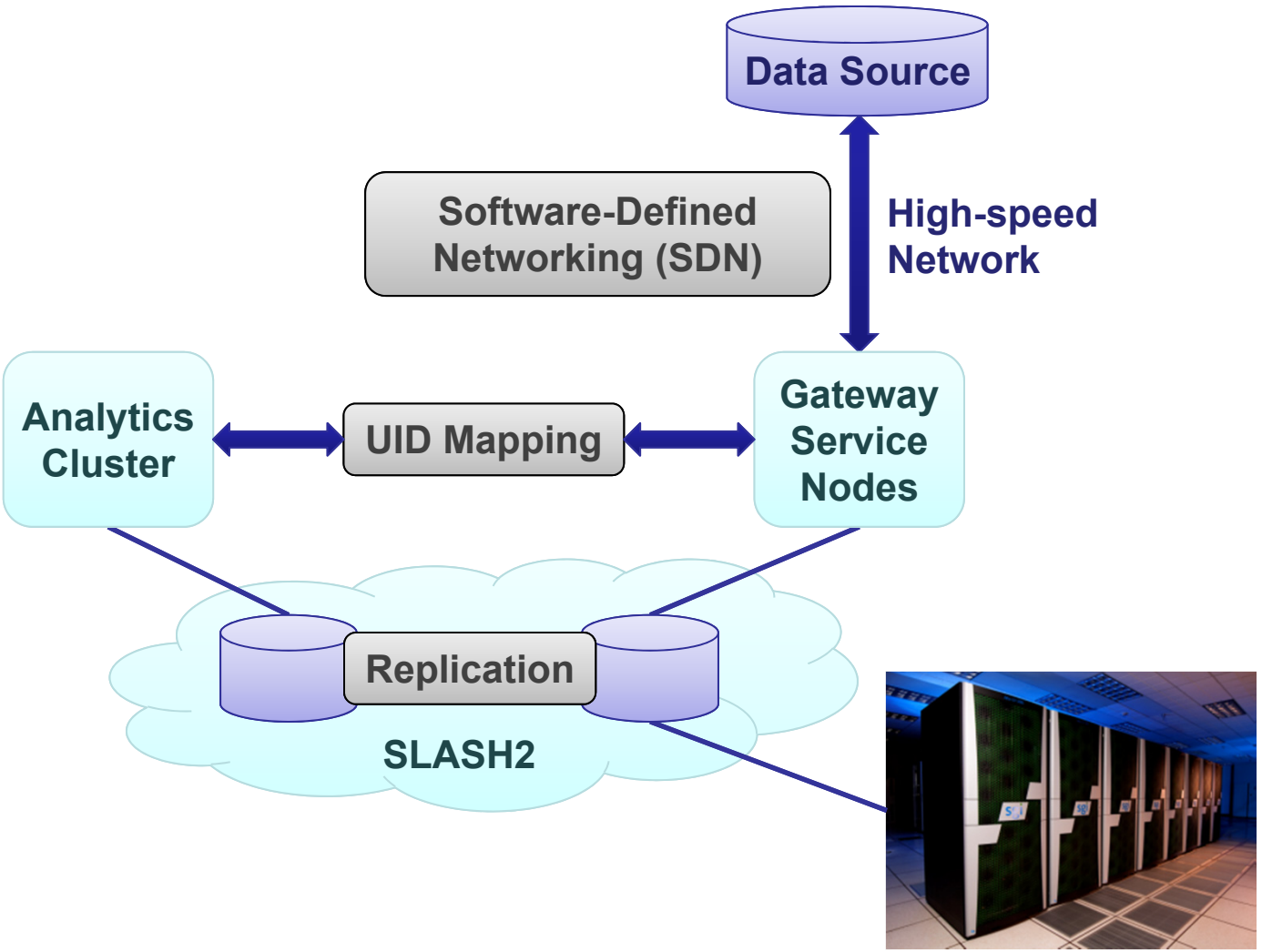
Background: The Data Supercell (DSC)

- A ***PSC-developed, groundbreaking, disk-based*** data management solution for ***low cost***, ***high bandwidth***, ***low latency***, ***high reliability***, and ***high capacity***
- ***SLASH2 managed, in production >2yrs, 4PB (usable)***
- As cost-effective as tape
  Bandwidth/$ ~ 24 × better than tape
  Latency ~10,000 × better than tape
  Scalable
- Highly secure
- Highest reliability for petascale storage
  - Enhancements beyond standard RAID
  - Options for geographical redundancy
  - Optimize data replication and data movement

PSC's Blacklight (SGI Altix® UV 1000 × 2)

- **2 × 16 TB of cache-coherent shared memory**
  - hardware coherency unit: 1 cache line (64B)
  - 16 TB exploits the processor's full 44-bit physical address space
  - *ideal for fine-grained shared memory applications, e.g. graph algorithms, sparse matrices*
- **32 TB addressable with PGAS languages (e.g. SGI UPC)**
  - low latency, high injection rate supports one-sided messaging
  - *also ideal for fine-grained shared memory applications*
- **NUMAlink® 5 interconnect**
  - fat tree topology spanning full UV system; low latency, high bisection bandwidth
  - *hardware acceleration for PGAS, MPI, gather/scatter, remote atomic memory operations, etc.*
- **Intel Nehalem-EX processors: 4096 cores (2048 cores per SSI)**
  - 8-cores per socket, 2 hardware threads per core, 4 flops/clock, 24MB L3, Turbo Boost, QPI
  - 4 memory channels per socket → *strong memory bandwidth*
  - x86 instruction set with SSE 4.2 → *excellent portability and ease of use*
- **SUSE Linux operating system**
  - supports OpenMP, p-threads, MPI, PGAS models → *high programmer productivity*
  - supports a huge number of ISV applications → *high end user productivity*

Example Architecture



Sherlock: a YarcData Urika™ Appliance with PSC Enhancements

- Graph Analytics Platform — *Massive multithreading and sophisticated memory handling for latency hiding*
  - *Next-generation Cray XMT*
- uRiKA application architecture
  - "Universal RDF Integration Knowledge Appliance"
- 32 Graph Analytics Platform nodes, each containing:
  - 2 Cray Threadstorm 4.0 processors, 128 threads/proc — *Remote Memory Access block with Extended Memory Semantics, providing a single, shared address space*
  - SeaStar 2 ASIC
- 1 TB globally shared memory
  - can accommodate graphs of up to ~10 billion edges
- General-purpose XT5 (AMD Opteron) nodes — *Enable additional, heterogeneous applications*

Urika™: Standards-based Graph Analytics

- Leverage emerging Web 3.0 standards
  - "*The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.*" – W3C[1]
- Resource Description Framework (RDF)
- SPARQL Protocol and RDF Query Language
- Jena framework for semantic web applications
- Application-specific GUIs for user interaction

1. "W3C Semantic Web Activity." World Wide Web Consortium (W3C). November 7, 2011.

Early Status: Overview

- Initial hardware deployed and configured; additional hardware on order
- Development of performance test tools and procedures
- Ongoing work to extend and expand filesystem functionality
- Collaborative research groups are engaged and developing pilot applications

Early Status: Pilot Applications (1)

- *Data integration and fusion for world history*
  (Vladimir Zadorozhny, University of Pittsburgh)
  - Remote SLASH2 instance at Pitt SIS
  - Distributed queries to Neo4j server on Blacklight for similarity joins (may move to Sherlock)
  - Background processing of similarity scores on Blacklight PSC
  - Data distributed between DXC and Pitt SIS
- *Semantic understanding of large, multimedia datasets*
  (Alex Hauptmann et al., Carnegie Mellon University)
  - Feature detection, speech, OCR, etc., and semantic indexing for a large corpus of video data
  - Data distributed between DSC and CMU SCS
  - Computation distributed between PSC's Blacklight and SCS's Rocks (mostly on Blacklight)

Early Status: Pilot Applications (2)

- *Identifying changes in gene pathways that cause tumors*
  (Michael Becich, Rebecca Crowley, et al., Univ. of Pittsburgh)
  - Remote SLASH2 instance at Pitt DBMI
  - Analytics running on Blacklight
  - Developing RDF representation of TCGA
  - Data distributed between DXC and Pitt DBMI
- *Exploring and understanding the universe*
  (David Halstead et al., National Radio Astronomy Observatory)
  - Planning a SLASH2 instance at NRAO
  - Will port applications to the DSC's analytic engines
  - Data to be distributed between NRAO and PSC
    - Automated, periodic transfers from NRAO to PSC of astronomy data
    - Automated transfers from PSC to NRAO of workflow flags placed in the filesystem by analytics applications running at PSC

Early Status: Pilot Applications (3)

- *Enabling bioinformatic workflows*
  (Anton Nekrutenko, Penn State University)
  - Working toward distributed execution between Galaxy Main at Penn State University and large-memory *de novo* genome assembly at PSC
  - Distributed workflow benefiting from efficient storage on DXC

Summary

- The Data Exacell tightly couples analytic engines with an innovative storage architecture to meet the requirements of diverse data-analytic applications.
- Pilot applications are being used to motivate, test, and harden DXC functionality.
- Through the DXC, we are creating hardware and software data infrastructure building blocks that will complement other XSEDE resources of today and tomorrow.