# Collaborative Access to Large Data



# SLASH2 - Data Exacell

J. Ray Scott
Pittsburgh Supercomputing Center

(For more info contact: <a href="mailto:scott@psc.edu">scott@psc.edu</a>)
[Presented, <a href="mailto:very">very</a> quickly, by Mike Levine]

Support NSF, NARA, Commonwealth of PA SLASH2 Availability
Search "slash2" on github

## Overview & Glossary

- Covering 3 different topics (very quickly!)
   All of potential relevance to a National Data Service.
  - File systems (software)
  - Physical systems w/data storage & other things
  - Use or purpose
- File system software
  - SLASH
  - SLASH2
- Physical systems
  - DSC aka Data Supercell (Storage)
  - DXC aka Data Exacell (Storage & analysis)
  - Bridges (Analysis & storage)
- Purpose
  - Regional data service
  - DIBBs pilot project for data intensive research
  - Production facility for data intensive research

(DSC)

(DXC)

(Bridges)





#### SLASH2 background & features (file system)

- SLASH: a file system designed to
  - Provide storage shared between multiple HPC systems
  - Serve as a user interface and cache between disk storage systems and a tape-based "mass store" (Cray/SGI DMF)
  - Production support for the 1<sup>st</sup> NSF Terascale system: Compaq's Quadrics-based AlphaServer
- SLASH2: an elaboration of SLASH designed from the ground up to be:
  - Portable
  - Scalable
  - Interoperable: in both computing platforms served and underlying file systems
  - Serviceable over wide-area networks (issues of latency and consistency)
- SLASH2 is an encapsulating file system (think Lustre)
  - Overall metadata services manage files as chunks on possibly heterogeneous and WAN distributed underlying storage systems
    - Can, and did, incorporate a tape-based mass store.
- Features
  - Multiple file residencies
  - System managed file replication and migration
  - Multiple error checking capabilities
  - Support for striping across underlying storage systems
  - Open source



Production implementations: **D**ata **S**uper**C**ell (5PB raw), **D**ata e**X**a**C**ell (variable), Bridges (14PB raw)

#### SLASH2 Architecture Overview\* (file system)

- Three software components
  - usually run on separate hardware but can all run on one server if performance is not an issue
- MetaData Server (MDS)
  - Provides file attribute and object management
  - Orchestrates data replication
  - Extensive control utility for MDS management msctl
- SLASH2 I/O Daemons (sliod)
  - File servers that store the file content as objects
  - Objects are stored in a local, native file system on the server
    - e.g. EXT3, ZFS, Lustre, NFS, tape-based DMF
  - There are usually many of these in a production system
  - Can utilize space on existing storage systems with SLASH2 as a "user"
  - They are orchestrated by the MDS
- Clients (mount-slash)
  - highly portable FUSE library
  - SLASH2 appears as a mounted file system
  - Data movement is third party

(In production; available via github)





## Data SuperCell (DSC): multi-PB regional data service

- Low cost
  - Replaced tape-based archive at same or lower price point
  - Low cost to operate
  - Open source software; commodity hardware
  - Modest foot-print
- Reliability
  - Redundancy
  - Multiple layers of RAID and checksums
  - Remote management reduces cost, repair time & probability of data loss
- Scalability
  - SLASH2 based
  - Allow use of tape or any other technology for underlying storage systems
- Performance: "Faster than tape"
  - 25x transfer rate(/\$)
  - 1/10,000 data access time (100s  $\rightarrow$  10 ms)
  - (Think data-intensive work!)
- Usage: ~3.2PB, ~500M files





# DXC: an NSF DIBBs pilot project.

- Data service w/data-analytics & architectural issues
- SLASH2 based + large memory analysis engine(s)
- Improved performance (cf DSC)
  - Next generation hardware
  - IOPS considerations
- System implementation & management additions
  - Database
  - Web
  - Virtual Machines
- Multiple user-partners
  - Provide goals and tests
  - Geographically separated
  - Multiple administrative domains
    - Functional support for workflows

- Users see an improving production environment.
- Example collaborator: <u>P</u>ittsburgh
   <u>G</u>enome <u>R</u>esource <u>R</u>epository
  - Collaborative effort dealing with <u>T</u>he
     <u>C</u>ancer <u>G</u>enome <u>A</u>tlas
  - Using SLASH2 to collect data and support 2-data center access.
  - University of Pittsburgh: Institute for Personalized Medicine (IPM), U. Pitt. Cancer Institute (UPCI), Department of Biomedical Informatics (DBMI), Center for Simulation and Modeling (SaM)
  - University of Pittsburgh Medical Center (UPMC)
  - Pittsburgh Supercomputing Center (PSC)



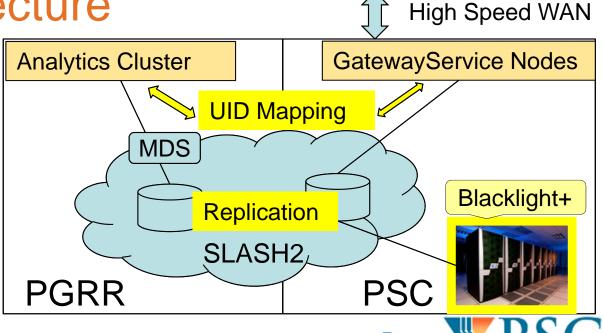
## SLASH2 Features Enabling PGRR

- Wide-area network
  - Resilience (keeps on going!)
  - Robustness (maximize performance)
  - mountable filesystem allowing access to custom TCGA client : genetorrent

- Selective data availability
  - cache data at PSC
  - active data at Pitt
- Local user credentials
  - id mapping

#### **DXC PGRR Architecture**

- Features relevant to NDS data access:
  - Managed
  - Protected
  - Active (mountable)
  - Shared



Data Source

