Enabling Grid features in dCache

Timur Perelmutov
Don Petravick
For dCache Team

dCache Team

Responsibility, dCache

Patrick Fuhrmann Rob Kennedy

Core Team (Desy and Fermi)

Jon Bakken

Mathias de Riese

Micheal Ernst

Alex Kulyavtsev

Birgit Lewendel

Dmitri Litvintsev

Tigran Mrktchyan

Martin Radicke

Neha Sharma

Vladimir Podstavkov

Responsibility, SRM

Timur Perelmutov

External Development

Nicolo Fioretti, BARI

Abhishek Singh Rana, SDSC

Support and Help

Maarten Lithmaath, CERN

Owen Synge, RAL

Grid Features in dCache

- Storage Resource Manager V1/V2
- GridFTP server
- Resilient Manager
- Interactive Web Monitoring

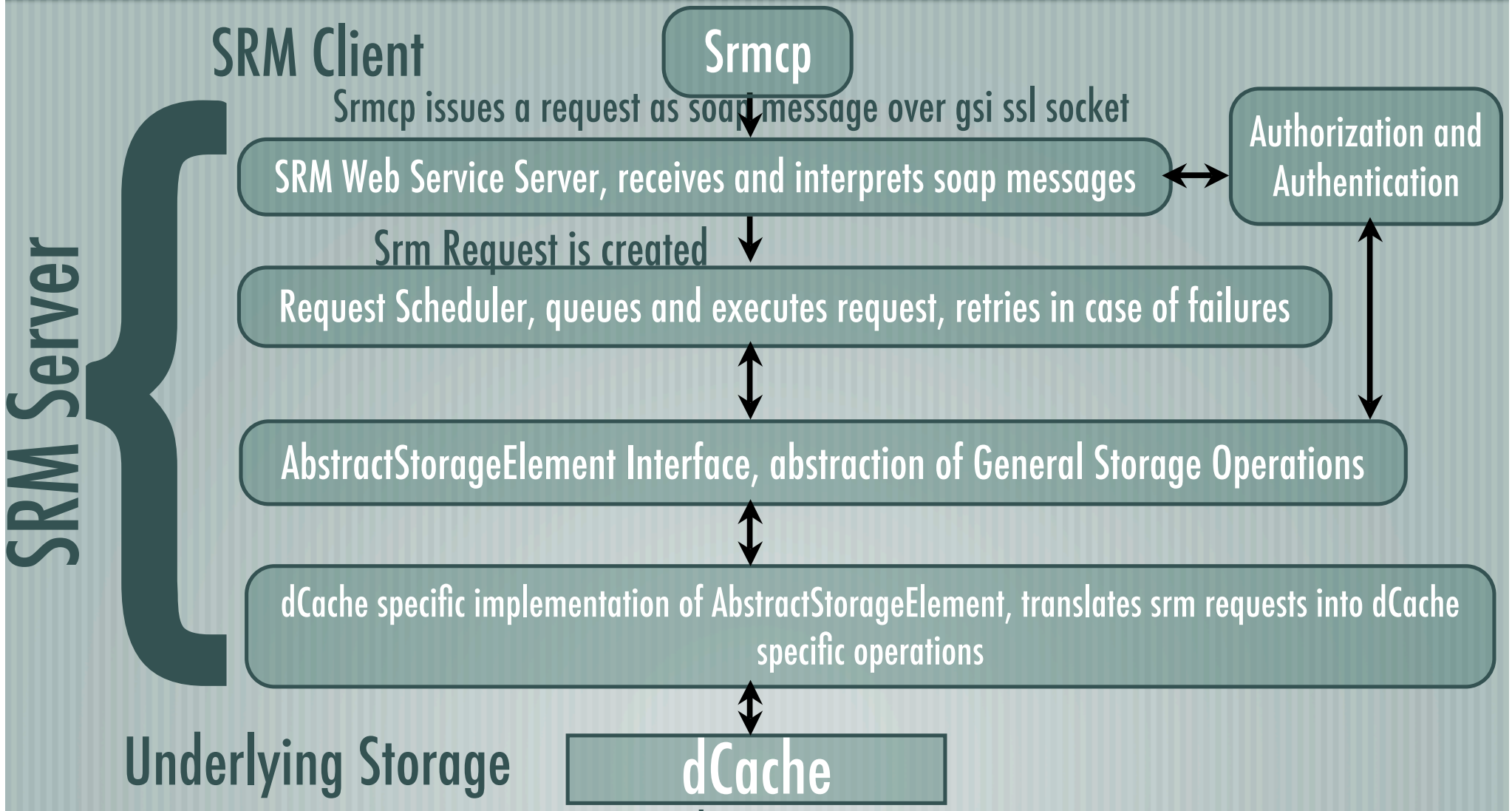
Storage Resource Managers

- SRMs are middleware components that manage shared storage resources on the Grid and provide:
  - Uniform access to heterogeneous storage
  - File Transfer Protocol negotiation
  - Dynamic Transfer URL allocation
  - Access to permanent and temporary types of storage
  - Advanced space and file reservation
  - Reliable transfer services

Storage Resource Manager versions

- Two SRM Interface specifications
  - SRM v1.1provides
    - Data access/transfer
    - Implicit space reservation
  - SRM v2.1 adds
    - Explicit space reservation
    - Namespace discovery and manipulation
    - Access permissions manipulation
- dCache SRM fully implements v1.1 specification and
- Just finished implementation of v2.1 data transfer and directory functions

SRM to dCache communication



dCache SRM Implementation Features

- Data Transfer Functions (get, put and copy)
- Data Transfer and directory functions from Version 2.1 protocol
  - srmPrepareToPut, srmPrepareToGet, srmCopy
  - srmLs, srmRm, srmMv, srmMkDir, srmRmDir
- Load balancing, throttling, fairness
  - Count number of trunsfer per door, select least loaded one
  - dCache poolManager selects “best” pool on basis of Space and CPU utilization
  - SRM bounds number of active transfers by limiting number of TURLs given to clients
- Scalable replication mechanism via gridftp
  - Pool movers are gridftp clients
  - Direct data node to data node connection in Mass Storage System (MSS) to MSS transfer
- Automatic directory creation
- Checksum verification
- Fault tolerance and reliability achieved by providing persistent storage for transfer requests and retries on failures
- SRM interface as a standalone product, adaptable to work on top of another storage system through a SRM-Storage interface
- A reference implementation of the SRM-Storage interface to a Unix File System
- Implicit Space Management

dCache SRM Implementation Plans

- Full implementation of SRM Version 2.1 interface
  - Explicit Space Management - April 2006
  - Support for at least Volatile and Permanent space types
  - Permission functions
- Research utilization of Lambda Station Interface by a Storage System.
  - Lambda Station gives optical path allocation and per flow routing
  - Utilize LS info in scheduling
- Open Science Grid Storage Element
- Monitoring, Administration and Accounting interfaces

dCache GridFTP Server

- GridFTP protocol v1 implementation
  - Most of standard and many advanced functions
  - Stream and Extended Block Transfer Modes
  - Scalable reads, data flows directly from data nodes
  - Write transfers go through Gridftp server - Scalable writes achieved by replication of doors on multiple nodes, and load balancing by SRM
    - Need Version 2 for achieving scalability for writes
  - Integrity verification on writes though checksum comparison

dCache Resilient Manager

- Typical problem of CMS Tier 1 installation
  - Computation farms
    - hundreds of nodes
    - data disks virtually unutilized
  - dCache PNFS can provide a global namespace
  - Worker node based pools are not reliable
- Tier 2 sites
  - Have
    - No tape backup
    - Limited resources - inexpensive (unreliable) disk storage
  - Want
    - Reliability
    - high throughput rates

Resilient dCache

- Resilient Manager
  - Top Level dCache service created to address above issues
  - Performs automatic replication to separate pools
  - Sets minimum and maximum number of replicas
  - PostgresSQL based local replica catalog
  - Reliable disk based storage
  - High IO rates

Resilient Manager in Action

1: Initial state, 2<= N <= 3

- All pools are online

	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Count
File A	A	A				2
File B	B		B			2
File C		C	C			2
File D			D	D		2
	online	online	online	online	online	

2: Pools 1 and 2 went down

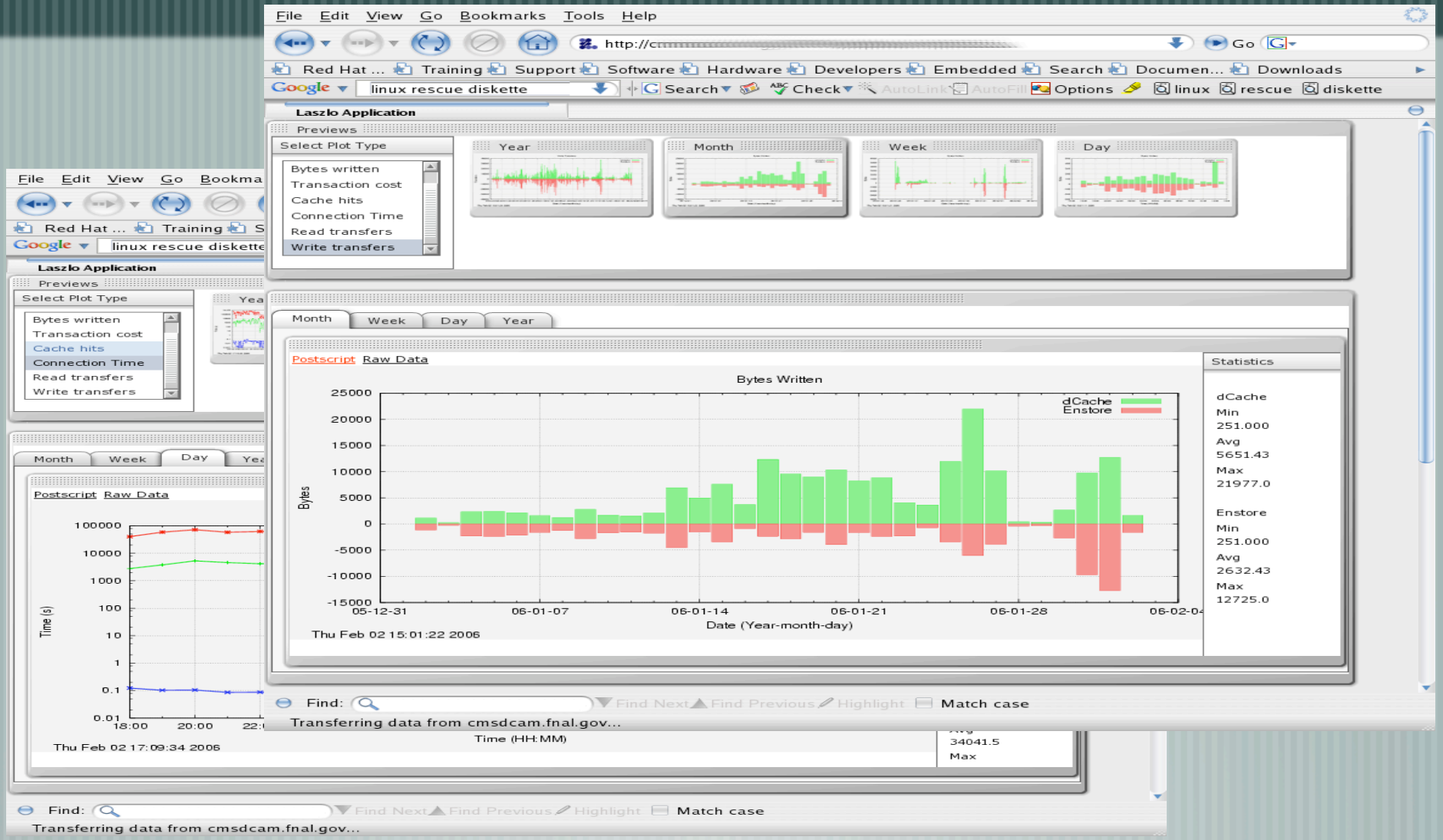
- Can't access File A; replicate B and C

	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Count
File A	A	A				0
File B	B		B → *B*			1
File C		C	C →		*C*	1
File D			D	D		2
	down	down	online	online	online	

dCache Monitoring Plots

- External application using dCache event logging database
  - Backend Database layer implemented Java/JDBC
  - OpenLaszlo based front end with rich user interface capabilities portable between browsers
- Builds plots based on various datasets and time periods
- Presentation parameters such as plot time period are controlled by xml, easily configurable by administrators
- Current system at Fermilab produces over 50 different plots

dCache Monitoring Plot examples



Resources

- DCache, Disk Cache Mass Storage System, http://www.dcache.org
- The Storage Resource Manager Collaboration, http://sdm.lbl.gov/srm-wg
- Fermilab SRM Project , http://srm.fnal.gov
- Resilient dCache Manual, http://cmsdcam.fnal.gov/dcache/resilient/Resilient_dCache_v1_0.html
- Dcache monitoring plots page https://plone4.fnal.gov/P1/DCache/dcache
- Lambda Station http://www.lambdastation.org/