LAWRENCE BERKELEY NATIONAL LABORATORY

COMPUTATIONAL RESEARCH DIVISION

Storage Resource Management

Alex Sim, Arie Shoshani, Junmin Gu, Vijaya Natarajan

Scientific Data Management Group

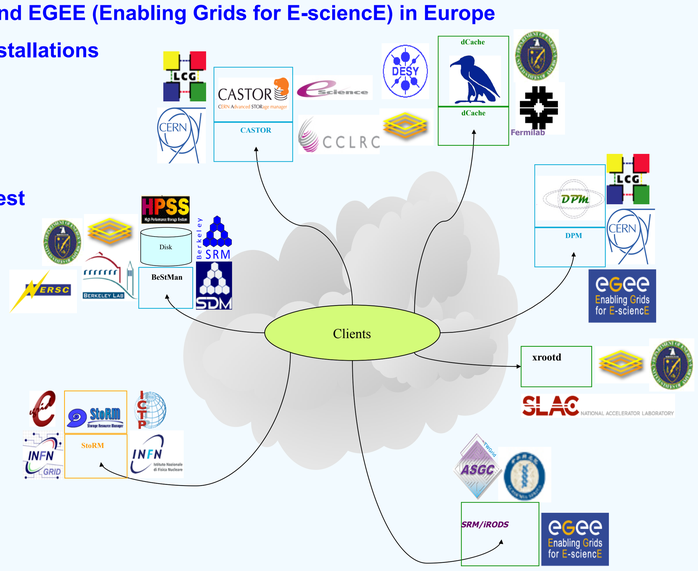
Storage Resource Managers: Essential Components for Distributed Applications

SRMs are developed independently and inter-operate in multiple sites

- The SRM concept was initiated at LBNL by the Scientific Data Management Group
- The SRM standard provides a common interface to various storage and file systems
- Interoperation of multiple Mass Storage Systems (MSSs) on the wide-area-network was achieved
- Multiple implementations of SRMs by institutions exist, yet they all interoperate
- Endorsed as a standard by Open Grid Forum (OGF) - effort led by LBNL
- Standard was adopted by large communities, such as the LHC (Large Hadron Collider) experiments, OGF (Open Science Grid) in the US, and EGEE (Enabling Grids for E-sciencE) in Europe
- 7 implementations world-wide, 300 installations

SRM functionality

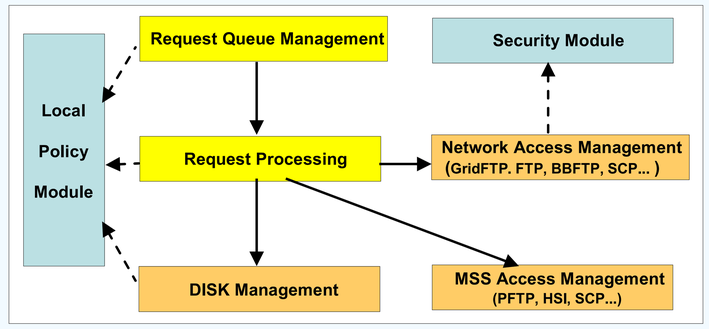
- Accepts multi-file requests, manages a request queue, allocates space per request
- Invokes GridFTP or other protocols to get/put files from/to remote sites
- Queues staging and archiving to MSSs
- Pins files for a limited lifetime, accepts early release of files
- Provides automatic garbage collection
- Monitors and recovers from failures of file staging, archiving, and transfer
- SRMs can be invoked by clients, other middleware, and *other SRMs*



BeStMan: the Berkeley Storage Manager

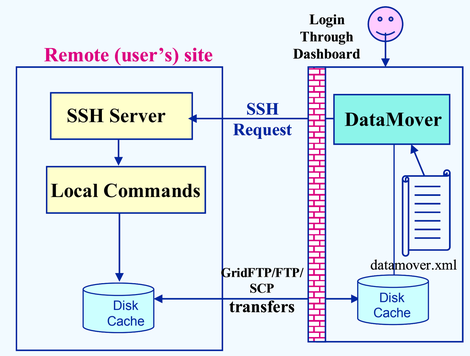
Main features

- Modular design to adapt to different storage and file systems
- Designed to work with disk systems, …
- as well as MSSs to stage/archive from/to its own disk (e.g. HPSS)
- Uses in-memory database (BerkeleyDB)
- Supports multiple transfer protocols
- Java implementation for portability



BeStMan-related products

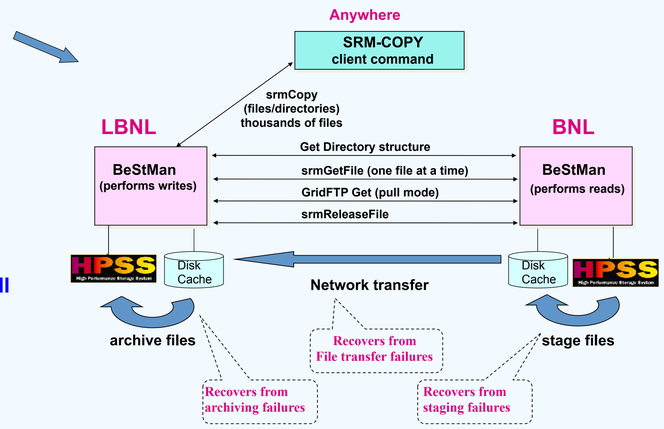
- BeStMan Full Mode
  - Supports queue management for multi-file requests
  - Supports quota space management and space reservation
- BeStMan Gateway
  - Designed as thin layer on top of file systems
  - No queue management or space management
- BeStMan as-a-client (called DataMover)
  - Can work to pull files to client site (see ESG below)
  - Can work to push files to client behind a firewall (used in SDM center framework through Dashboard)



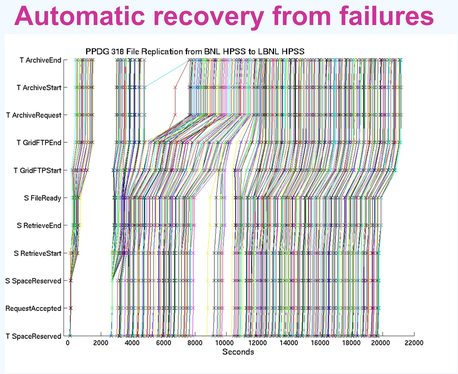
Robust File Replication Using BeStMan

Robust multi-file replication in STAR

- Replicating thousands of files is a lengthy error-prone task – needs automation
- Error recovery needed for MSS and WAN transient failures or maintenance interruptions
- Entire directory structure gets replicated
- Concurrent file transfers performed to take advantage of available bandwidth
- Used in Earth System Grid (ESG) project as well



Automatic recovery from failures

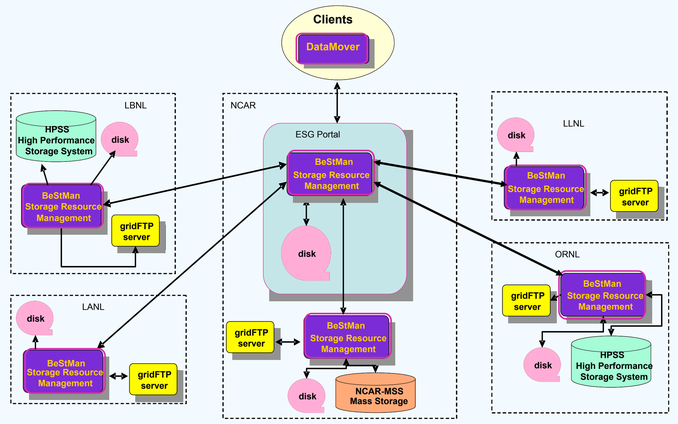


- Events of file transfer recorded (space allocated, stage file, move file, archive file, release file, …)
- Graph shows progress of events over time: each vertically connected line represent history of one file
- Gaps in graph show downtime (or failure) periods in the source site, network, or target site
- Graph shows full recovery and completion of request

BeStMan Deployments

Earth System Grid (ESG)

- Serving about 14000 users
- Over a million files and 170TB of data
- Files pulled into Portal disk cache by BeStMan from 5 storage sites, some with HPSS (LANL, LBNL, LLNL, NCAR, ORNL)
- A special adaptation of BeStMan was developed for NCAR's own MSS



Open System Grid (OSG)

- ~ 30 deployments of BeStMan in the US and Europe
- US CMS (an LHC experiment)
  - BeStMan Gateway used as an SRM frontend for Hadoop at UNL, Caltech, UCSD
  - Compatible with CMS through EGEE in Europe – passed all tests
- US ATLAS (an LHC experiment)
  - BeStMan on NFS
  - BeStMan Gateway on Xrootd/FS, GPFS, IBRIX
- STAR experiment
  - Robust file replication (see left panel) between BNL and LBNL
    - in production for over 4 years
    - HPSS access at BNL and NERSC
  - Used in analysis scenario to move job-generated data files from LBNL to remote BNL storage

CRD computational research division

U.S. DEPARTMENT OF ENERGY Office of Science