

WHITE PAPER

Collaborative SRB[®] Data Federations

*A Unified View for Heterogeneous
High-Performance Computing*



Nirvana
A DIVISION OF GENERAL ATOMICS



INTRODUCTION

This paper describes Storage Resource Broker (SRB): its architecture and capabilities as a base for **data grid implementations**. SRB unifies heterogeneous HPC environments, providing easy discovery, sharing and management of high-value data in R&D federations. With SRB, advanced collaborative R&D grids can be implemented today.

SRB presents applications and clients with a uniform interface to heterogeneous distributed resources, **including file systems, databases, and archival storage**. A Metadata Catalog (MCAT) is the vehicle for abstraction and provides accredited end users an enterprise-wide, "collection"-oriented view of all data, wherever located. SRB enables productive **collaboration and efficient data exchange within a secure, scalable, easily administered environment**.

Under development since 1995, SRB has become the most widely adopted data federation software in U.S. research communities and is the vehicle for managing massive digital libraries and image archives at over 150 sites including NASA, Stanford and Harvard. These comprise the world's most advanced image repositories, digital libraries, archives and data grids, in the most demanding HPC work streams.

Nirvana developed SRB in cooperation with the San Diego Supercomputer Center at UCSD. The product addresses the needs of collaborating research institutions operating compute and data-intensive **projects predicated on simplified, easy access to shared, high-value data in complex ad hoc working groups**. In these projects, SRB forms the foundation for the world's most advanced and demanding data grids conducted at leading research institutions in the U.S. and spanning efforts in earth sciences, neuroscience, particle physics and cosmology.

View other whitepapers:
www.ga.com/nirvana/solutions/whitepapers.htm.

FEDERATING HPC ENVIRONMENTS

An immense challenge in today's HPC communities is the management of disparate storage systems and HSM architectures in distributed data centers. Even within a single organization it is not unusual to see such obstacles to collaboration. Data centers are often designed separately, managed independently, and each typically implements a unique infrastructure without designing for cross-enterprise compatibility. It is rare that all operating groups in a complex enterprise agree on a common infrastructure.

This is where SRB brings value. SRB accommodates existing, legacy storage and HSM infrastructures, while providing a global view of all data across the enterprise – or even across an entire HPC community.

SRB introduces an MCAT abstraction layer and creates and maintains a global namespace organized in a logical hierarchy independent of the physical implementation of the local data storage infrastructure. Logical data collections, containing files from several physical locations, can be created through SRB. Data can be migrated to new storage systems or different locations without changing the logical "collection"-oriented view of the data.

Where the objective is achieving a unified view of high-value data for collaborating R&D workers - and the legacy compute environment is heterogeneous, distributed, and complex - SRB is the correct choice.

SRB is client/server middleware that connects applications with diverse data resources including file systems on Storage Area Networks (SANs) or Network Attached Storage devices (NAS), and online Content Addressed Storage (CAS) disk archives such as EMC Centera. Layers of data from disparate sources and in various formats can be assembled and grouped into logical data collections, for transparent access by authorized users.

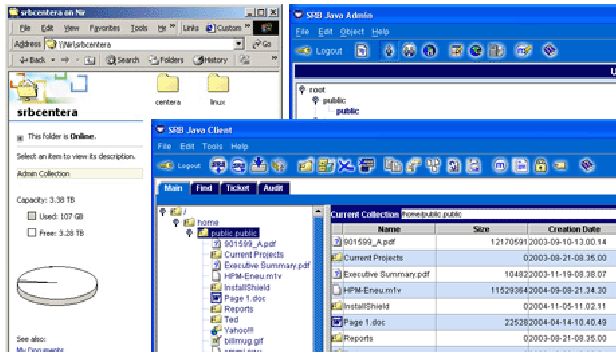


Figure 1: Selection of SRB User Interfaces

The result is an effective mechanism for managing massive and rapidly-expanding data stores. SRB scales easily to very large enterprise-wide and even cross-enterprise federations, and provides multiple user interfaces, including Windows gateways and simple point-and-click Java applications.

FEDERATING EOS R&D

NASA's Earth Observing System (EOS) Program implemented an ambitious Remote Data Store (RDS) project in 2003, similar to solutions architected for NOAA's R&D HPCS, including the heterogeneous legacy environment, geographic distribution of centers, and high-value content. Since 2003 the RDS has been architected with SRB at its center. NASA is now collecting unprecedented volumes of data to aid the U.S. government in its understanding of the earth's near and long-term climate processes. These data sets are a national resource that must be carefully preserved to maximize the return on the EOS Program. To address this need, the Earth Science Data & Information System (ESDIS) established a remote data storage backup and recovery capability that operates independent of, but closely allied to, the current EOS mission Data Active Archive Centers (DAACs). The RDS was designated for the persistent storage of portions of the DAAC data holdings for backup purposes, and eventually for some degree of load balancing across the DAACs.

With this infrastructure in place at the RDS, the same architecture could be expanded in the future to span NASA's DAACs and to create a global namespace across all data centers ("One NASA Vision"). Nirvana is helping NASA realize this vision by providing the following capabilities:

- Global namespace for a seamless, unified view of all data
- Single sign-on to global namespace
- Fail-safe architecture with all components clustered (MCAT, SRB Agents, database)
- Transient and persistent data stores
- High performance (30-45MB/s for persistent, 60-78MB/s for transient; limited by hardware)
- Vertical and horizontal scalability
- Policy-based automated data management
- FTP access
- GridFTP gateway for SRB connectivity into compute grids (Virtual Data Product)

The files are ingested into the RDS from the Goddard Space Flight Center (GSFC) using a custom application. This Java application can query the GSFC databases and extract only data from the AMASS storage system that matches specified criteria. The application then transfers the extracted data from GSFC to the RDS in West Virginia using SRB protocol, parallel I/O transfer mechanisms and bulk operations involving data packing and data streaming.

At the RDS, the SRB Policy Daemon automatically migrates files from various transient systems (currently 150TB EMC CLARiiON CX700 and 50TB SGI TP9500) to the persistent data store (10TB EMC Centera) and vice versa depending on criteria specified in its policies. Connectivity to StorageTek tape silos is supported and pending implementation.

NASA scientists use the RDS for data recovery into GSFC or for direct data access via FTP protocol. They can also use the SRB Web Client via any Internet browser to query the RDS data store using custom



metadata attributes. This eases data discovery in large collections with thousands of data objects.

NOAA HPCS

The NOAA R&D HPCS RFP maps well to SRB's capabilities. The following illustrates areas where SRB addresses NOAA's challenges.

Performance and Scalability

NOAA specified performance requirements and benchmark tests to validate such requirements. In addition to supporting a variety of scientific user activities at different sites, the storage component of the HPCS must also perform common tasks such as backup and data migrations in the background.

SRB's architecture is modular and scalable. Every component can be clustered for load balancing. More capacity can be easily added by bringing new SRB Servers online, offering nearly linear scalability. The heart of SRB, the MCAT, is built on robust relational database technology and is highly scalable in the number of transactions that can be supported and the number of files that the system can manage. The database can be replicated across multiple sites, so that each site has real-time access to all metadata.

There are also numerous performance-enhancing features implemented in SRB: parallel I/O for maximum bandwidth utilization, bulk operations for efficient handling of large batch jobs, and intelligent resource selection for transparent access to the fastest and closest data stores.

High Availability and Reliability

NOAA has a requirement for all scientific data to be available 99% of the time, which translates into 3 days, 15 hours and 40 minutes of downtime per year. This can only be achieved if all components of the backend storage system are redundant and can failover in the case of unscheduled down-time.

SRB components can all be clustered and configured to automatically fail-over to secondary or tertiary

systems. Both server-side SRB components – MCAT servers and SRB Agents – are redundant. Multiple MCAT servers can provide an automated fail-over mechanism for the metadata database communication. SRB Agents can host replicas of data in distributed locations, failing-over if one of the locations goes offline. Data replicas can be kept in synch either synchronously or asynchronously as scheduled. Data checksum operations can be scheduled and data corruption can be detected and eliminated through clean replicas.

System upgrades are performed on redundant system components so that operations can continue during the upgrade process though with degraded performance.

"One NOAA" Vision

NOAA's vision, that their users can (a) access any computational platform, (b) access all data within the organization independent of the location or underlying storage system, and (c) migrate data from existing to future archives transparent for the scientific users, is already implemented today in dozens of SRB compute grid federations.

SRB overlays a global namespace on all storage systems and can share-out this global namespace according to the client's interface requirements: NFS mounts, CIFS shares, Web folders through WebDAV, command-line interfaces, Web Services, Java applications, or APIs in both C and Java. A single file system will span the entire organization, including all home file systems and all existing HSM systems.

The global namespace, a logical representation of all data enterprise-wide, is independent of the organization of the underlying physical storage. This has advantages for both end-users and administrators: end-users can retain a logical, hierarchical, or even ontological organization of all their data without being affected by system upgrades or data migrations. Administrators are free to migrate data to new archives and maintain replicas and data backups without disturbing users.



Automated Data Management and HSM

NOAA sites are currently running at least three different HSM systems – ADIC StorNext, IBM HPSS, and SGI DMF. Each system manages its own archive and has its own policies. Some of the systems will reach full capacity soon and must be upgraded or replaced. It would be an enormous undertaking to replace just one of those systems without disrupting operations significantly. SRB can provide a global namespace overlay of each HSM system and either expand them incrementally or roll up their function over time.

SRB contains very powerful integrated HSM and ILM capabilities. They are carried out by the ILM Daemon, a policy-based automation module of SRB.

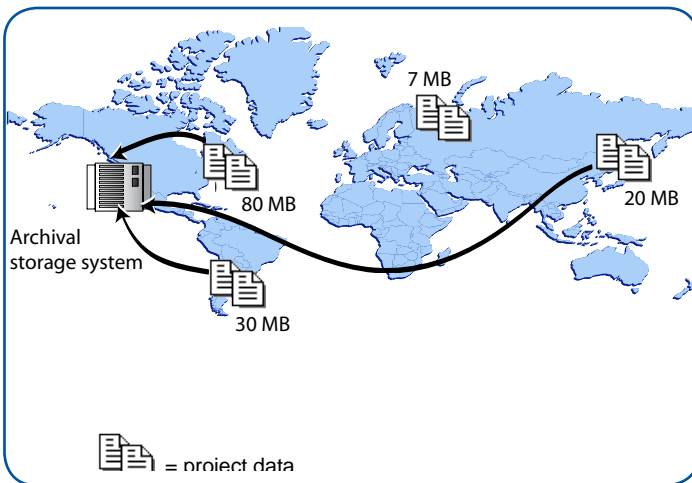


Figure 2: ILM Daemon

Within an SRB federation, administrators can deploy any number of ILM Daemons, while controlling them from a central location. ILM Daemons routinely query the MCAT servers and execute actions – such as migration, replication, backup, expiration, or synchronization – on a specified schedule.

The policies are extremely flexible and can contain system-level metadata attributes (watermarks, object sizes, object age, average access count, etc.) or user-level metadata attributes (e.g. longitude,

latitude, data source, satellite, product name, model name, etc.). Policy execution is based on administrator-defined schedules and occurs behind the scenes, transparent to end-users or applications.

Metadata

Without indices and a descriptive catalog, any file archive with more than a million files becomes a write-once-read-never (WORN) system. It is common in this scenario for individuals to ingest large amounts of high-value data into the archive and subsequently leave the organization, stranding this data and losing its value forever. The MCAT server at the heart of an SRB federation tracks several metadata attributes, prevents the genesis of a WORN archive, and supports effective data management. The attributes are grouped into system and user-level metadata:

System-level attributes are automatically maintained by MCAT and are mostly queried, displayed, and used for data management. System-level attributes are maintained on all SRB objects. Examples are file path, object size, resource type, etc.

User-level attributes are created and maintained by applications or users and can be attached to both files and directories. This can be either a manual process or automatic if electronic metadata repositories already exist. User-level attributes too are useful for query, display, and data management, and are access-controlled on a per-attribute-basis so that administrators can create access policies such as “Engineering can modify the revision attribute, analysts can read it, and the public can not even see it.” Other examples for user-level attributes might be satellite, projects, etc.

Containers

The container technology referenced in NOAA’s RFP is tightly integrated within SRB and patented by General Atomics. SRB Container addresses three major problems which typically occur with traditional HSM systems, and which will almost certainly impact the projected operating environment.

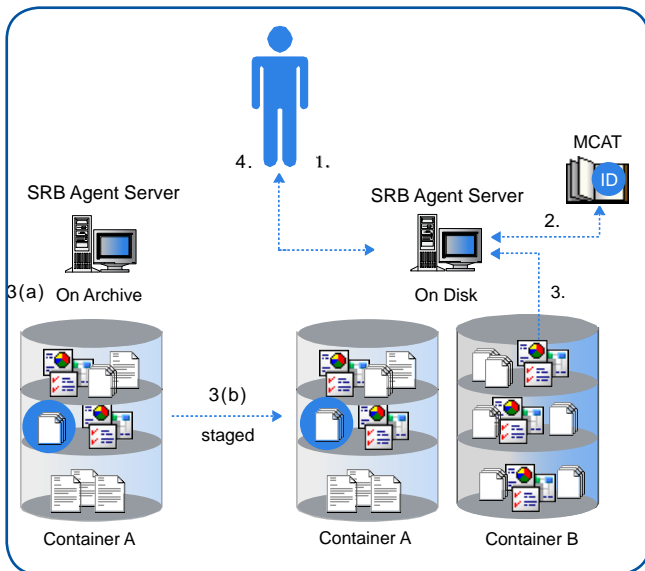


Figure 3: SRB Container Design

The first is high-latency when retrieving data from tape. The second is file storage limitations. Archives are designed to store small numbers of very large files whereas data collections are more likely to contain large numbers of small files. Finally, if the data is distributed over a WAN, serious latencies can ensue when transferring a large number of files. Through its patented Container technology, SRB addresses all these problems.

SRB manages containers so that archives can be transparently integrated with all other enterprise-wide data. These containers are transparent to end-users or applications by presenting in-container objects just like any other SRB Object in the global namespace.

Data objects are physically packed into a container before being stored into an archive. This assures that associated data objects will reside on the same tape while decreasing the number of data objects managed by the archive. When later retrieving these data objects, the correct container is automatically staged to disk and all read access occurs against the staged copy. As related data objects are retrieved together, access time can be reduced dramatically.

Latency Minimization across WANs

The NOAA RFP encouraged innovation in the WAN solution in order to optimize total system performance. The main enemy of WAN communications is of course latency. If one were to transfer 100,000 files across a WAN with just 10 hops (10ms latency each), one would incur a total time loss of almost 3 hours just for sending the instructions (1 packet per instruction) for the data transfer.

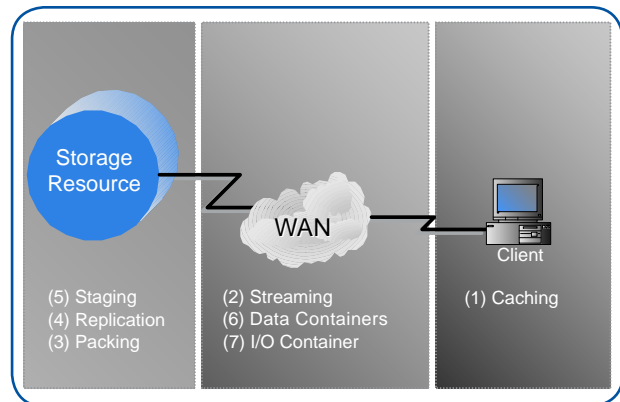


Figure 4: Latency Management in SRB

SRB features multiple mechanisms to reduce the latencies common to such data transfers:

- (1) Caching keeps local copies of frequently accessed files on fast storage or memory.
- (2) Streaming is continuous transfer and receipt of data without noticeable lag time.
- (3) Packing aggregates multiple files into a single buffer before sending the files over the network reducing latencies associated with multiple requests.
- (4) Replication keeps multiple synchronized copies of files at different sites.
- (5) Staging writes files stored on archival storage media to disk for faster access.
- (6) Data Containers transparently aggregate multiple objects into one large file that can be stored and transferred more efficiently than multiple smaller files.
- (7) I/O containers allow for a remote execution of



Nirvana
A DIVISION OF GENERAL ATOMICS

Nirvana Storage, a Division of General Atomics

3550 General Atomics Court

San Diego, CA 92121

P: (858) 455-2500

F: (858) 455-2529

Email: srb@nirvanastorage.com

www.ga.com/nirvana