Managing Data Distributed Across Geographically Separated Storage Repositories

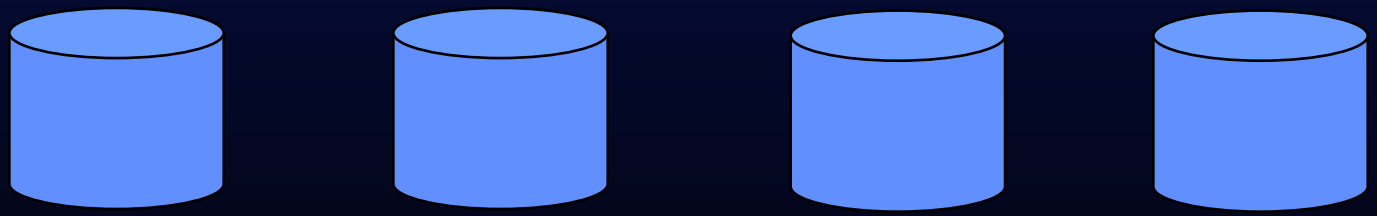
Reagan W. Moore - SDSC
Wayne Karpoff - YottaYotta
Kent Koeninger - Hewlett Packard
Michael Factor - IBM Research Lab in Haifa
Dave Berry - UK National e-Science Center
William Adamson - University of Michigan
Micah Beck - University of Tennessee

Who is in Control?

Data Grids

Global File Systems

Storage Virtualization



Object-based Storage Devices

Data Management Hierarchy

- **Data virtualization**
  - Span file systems, archives, ORBs
- **Global File Systems**
  - Span multiple sites
- **Storage virtualization**
  - Manage multiple disks
- **Object-based storage**
  - Manage local objects

Control versus State Information

- **Who manages assertions about data integrity, data authenticity, data access?**
  - Where does associated state information reside?
  - What level of assurance can be provided by each data management level?
  - What interactions between levels must be consistently managed?

Operations versus State Information

- **Who manages results of applying operations on data?**
  - RAID distribution
  - Replica creation
  - File locks
  - Checksum validation
  - File and user name spaces

Data Grids

- **Extensive state information**
  - File properties
    - Checksum, audit trail, owner, size, replica, version, backup, container, change data, ACLs
- **Descriptive information**
  - Authenticity
    - Provenance metadata
    - Descriptive metadata

Data Grid Operations

- **File access**
  - Open, close, read, write, seek, stat, synch, …
  - Audit, versions, pinning, checksums, synchronize, …
  - Parallel I/O and firewall interactions
  - Versions, backups, replicas
- **Latency management**
  - Bulk operations
    - Register, load, unload, delete, …
  - Remote procedures
    - HDFv5, data filtering, file parsing, replicate, aggregate
- **Metadata management**
  - SQL generation, schema extension, XML import and export, browsing, queries,
- **GGF, "Operations for Access, Management, and Transport at Remote Sites"**

Data Management Environments

- **Data grids**
  - Manage shared collections
- **Digital libraries**
  - Provide discovery, browsing, presentation services on top of collections
- **Persistent archives**
  - Manage technology evolution while the authenticity and integrity of the assembled collection is preserved
- **Real-time sensor networks**
  - Manage access to real-time data streams from thousands of sensors