Data Storage Support for Grid Computing

18 September 2003 – v1.6

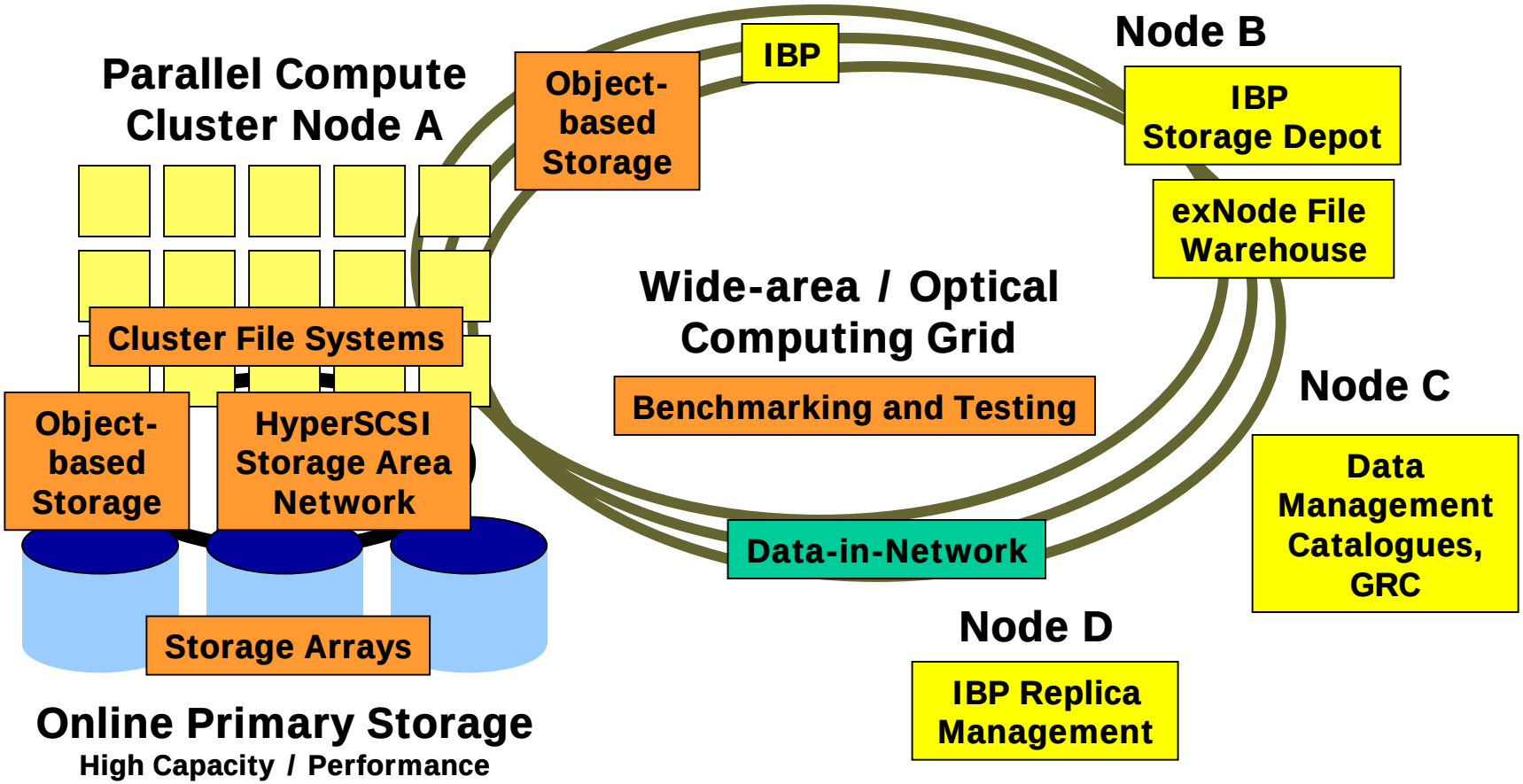
Storage as a Key Fabric Component

- Computing applications require input and produce output data which needs to be stored efficiently and reliably
- Two key aspects of this area are:
  - Intra-cluster Storage (local)
  - Distributed Storage (wide-area)
- This Presentation will cover:
  - Brief Overview of Current Efforts in Singapore
  - Overview of our Proposed Ideas in this Area
  - Conclusion

Current Local Storage Efforts in Grid

Intra-Cluster Storage

Distributed Storage



Leading organisation: DSI I2R NTU

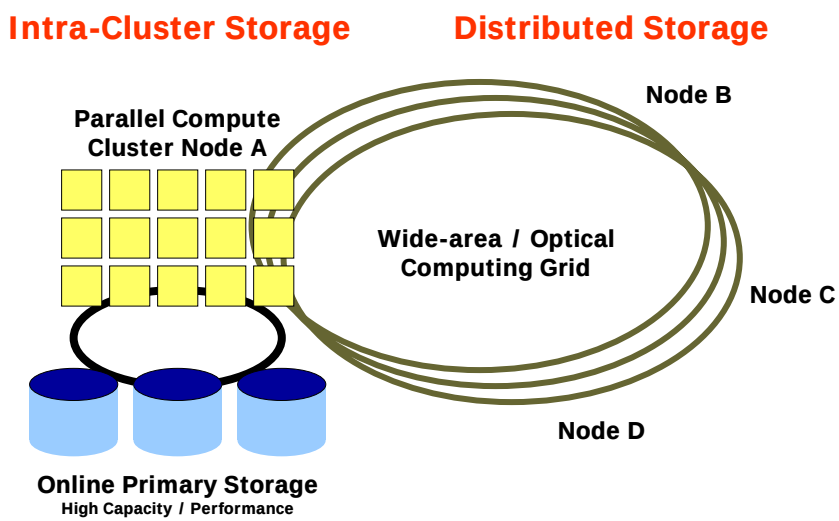
Proposed Storage Efforts in Grid

D-QoSS
Distributed Quality of Service Storage

Providing policy-based storage quality assurances through the management of storage and storage devices across the data grid.

Four key areas of effort:

- Management and Policies
- Inter-cluster Architecture
- QoSS Specifications
- Intra-cluster Provisioning



Management and Policies

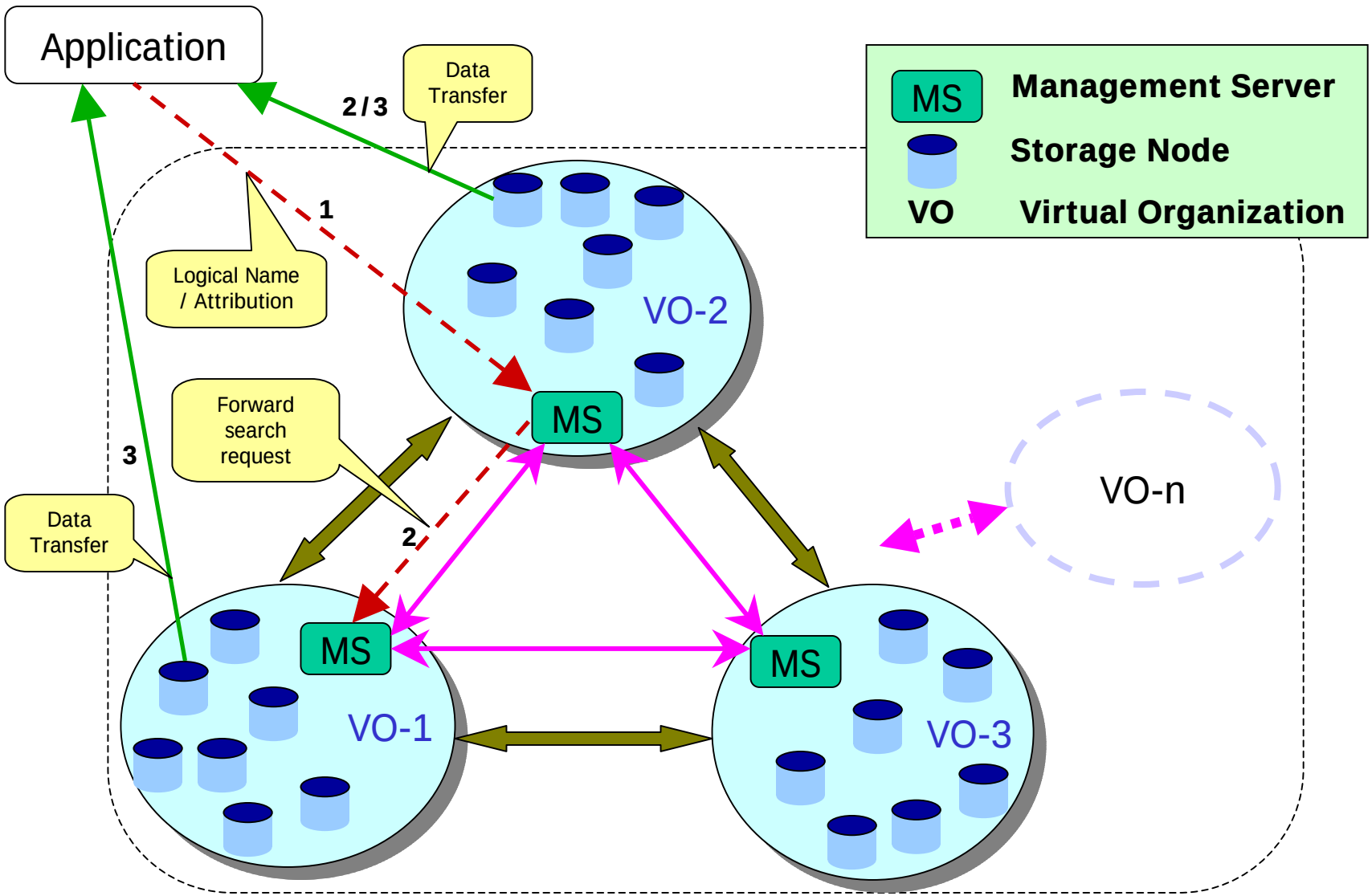
Key Tasks:

- Encapsulate the underlying systems and provide a uniform interface to user
- Universal name space for data
- Optimized wide-area data access/caching
- Management metadata (storage, network, data, user, ...)
- Interfaces to mass storage system
- Access control and authentication

Some Research Concerns:

- Authentication across VOs
- Storage and data sharing/access among VOs
- VOs can easily join or depart from the federation.

Management and Policies



Inter-cluster Architecture

Peer-to-Peer Object Oriented Storage Architecture

- A peer-to-peer object infrastructure forms the O-O overlay for future distributed storage systems
- A suitable overlay must allow for flexible storage object creation, modification and deletion at all distributed nodes. Other issues include scalability, fault-tolerant and self-organizing
- Algorithms for organizing, routing and accessing distributed objects have been proposed in DOIs such as Tapestry (UCB), CAN (UCB, AT&T), JavaSpace (Sun Microsystems), Tspace (IBM) and Chord (MIT), however, current works focus on object location and routing over standard best-effort Internet and routing algorithms

Inter-cluster Architecture

Peer-to-Peer Object Oriented Storage Architecture

- This project proposes a peer-to-peer storage architecture with object location and routing through data-in-network (DIN) and O-O file management techniques developed at I2R and DSI respectively
- The architecture will support storage object data communications for “Command” and “Data” channels for caching, synchronization, and data replication purposes
- Also proposed is a peer-to-peer metadata catalogue for efficient lookup and translation service

QoSS Specifications

QoSS Metrics

- Storage resource discovery
- Storage resource dynamic allocation / reconfiguration
- Capacity / performance monitoring
- Meta-data and object properties
- Resource utilization and billing
- Error monitoring and troubleshooting

QoSS Specifications

Common Information Model (CIM)

- Open framework to manage storage resource and systems
- Object oriented Models
- Support by SNIA and other big industry players

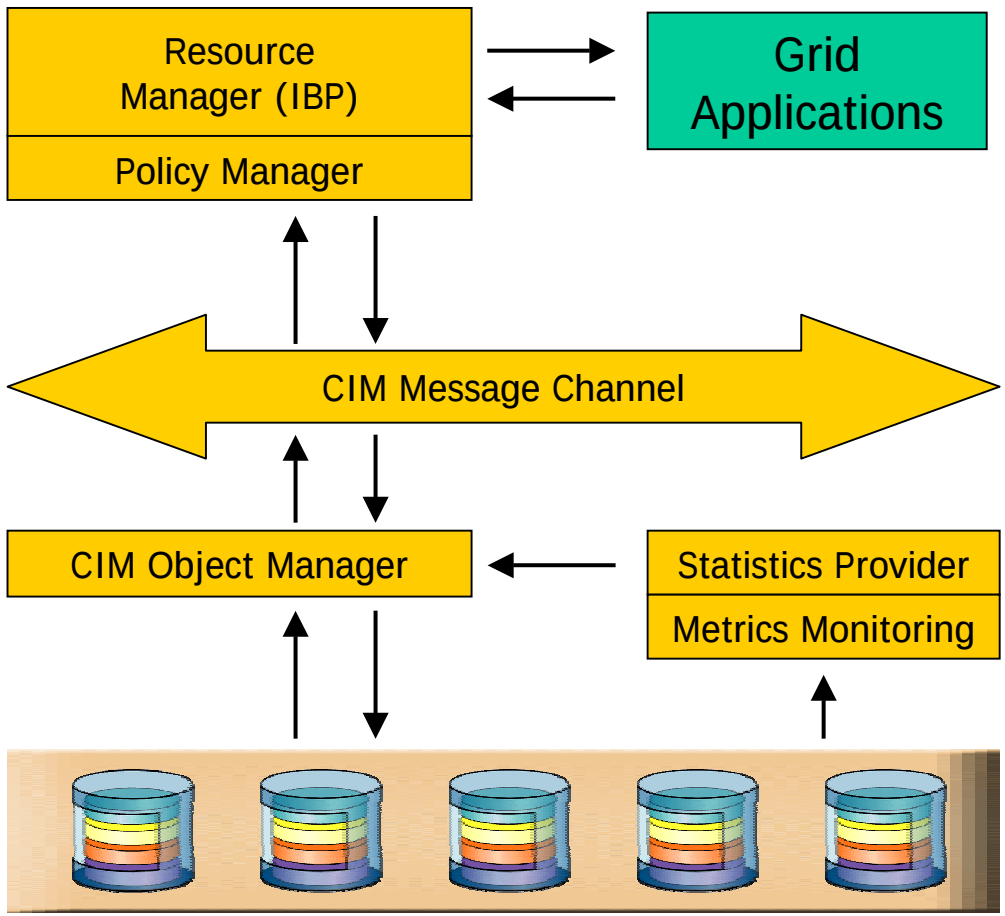
CIM Building Blocks:

- Core Model
- Common Models
- Extension Schema

Work Areas:

- Data/Replica placement
- Data/Replica Access
- Data Storage management

CIM-based QoSS for Grid



Intra-cluster Provisioning

Ethernet-based Network Storage for HPC Clusters

Protocols:

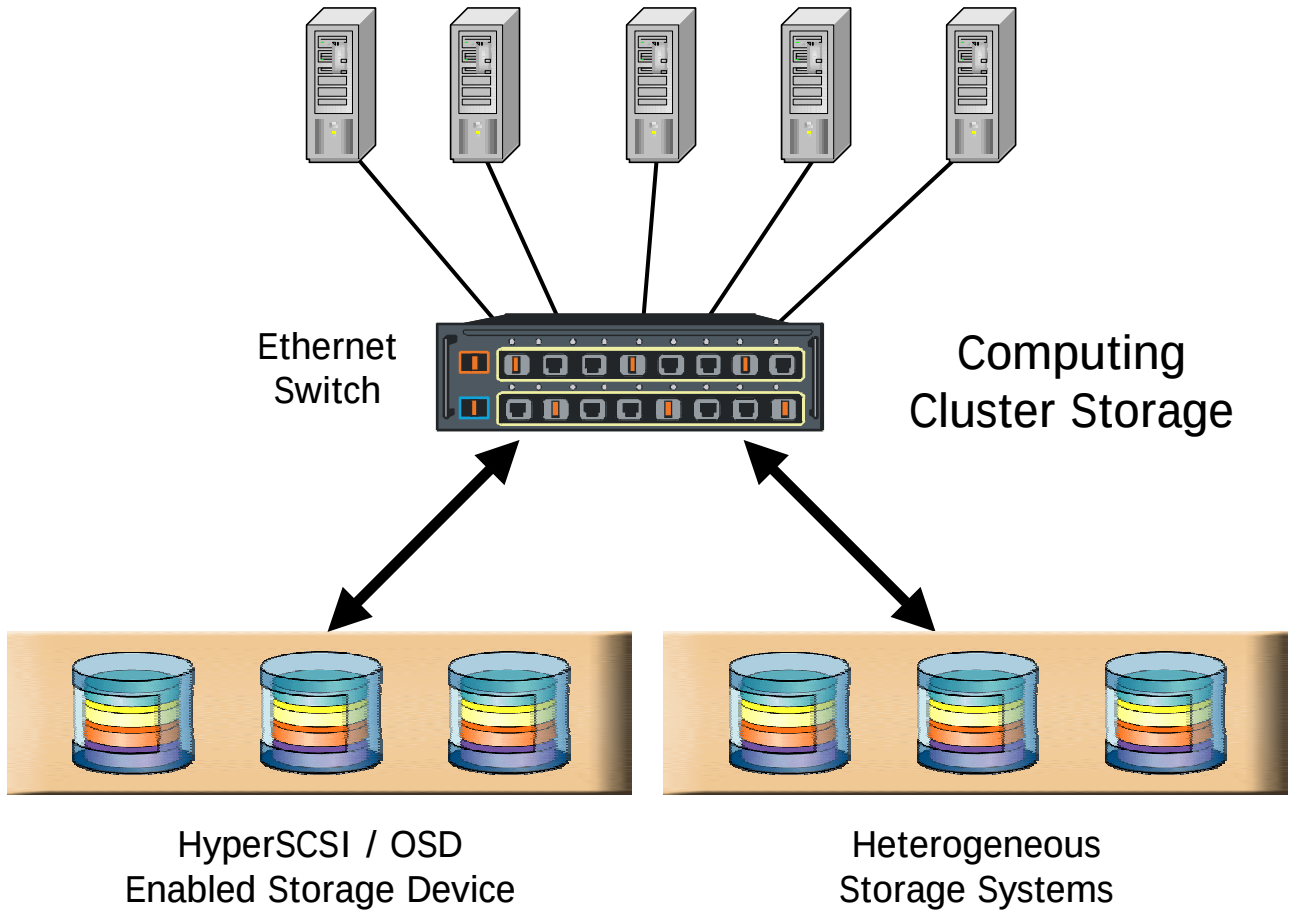
- iSCSI
- HyperSCSI

File Systems:

- DFS/CFS
- OSD

Management:

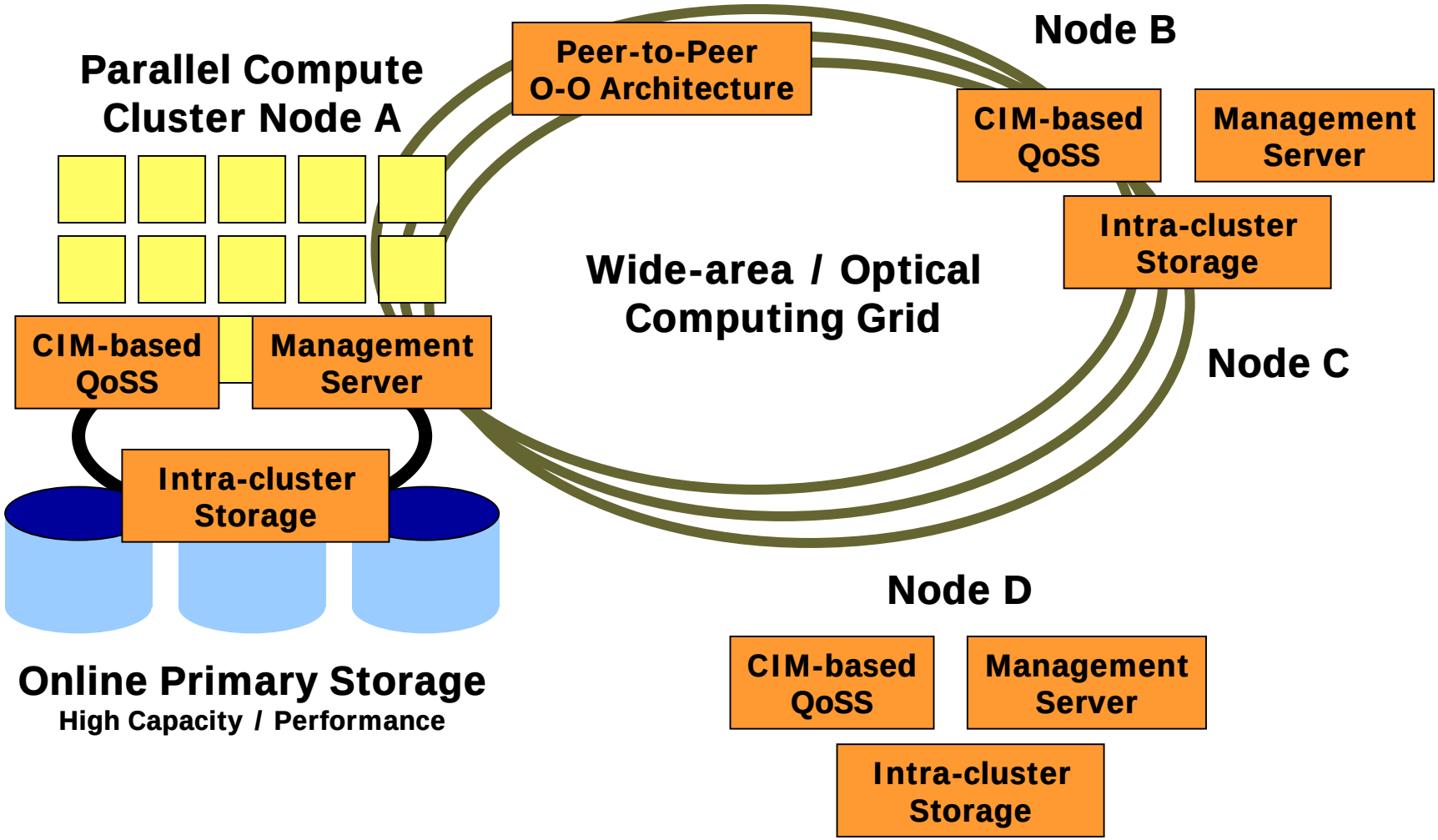
- Storage resource management
- CIM based QoSS



Summary of D-QoSS Efforts

Intra-Cluster Storage

Distributed Storage



Possible Collaborations and Synergies

- Computer Science, University of Tennessee
- CANARIE
- Canada Wavelength Disk Drive (WDD)
- Lawrence Berkeley National Labs
- Hitachi Data Systems
- IBM Research (Almaden)
- Veritas Research (Mountain View)
- Singapore National Grid Pilot Project
- NTU Campus Grid
- Singapore Advanced Research and Education Network
- A*STAR Optical Access Network Focused Interest Group
- NTU SCE, NTU EEE, NUS ECE, I2R, NST DSI, BII

What's Our Difference?

Distributed Network Storage Service with Quality-of-Service Guarantees

John Chung-I Chuang
Carnegie Mellon University/
University of California, Berkeley

Marvin A. Sirbu
Carnegie Mellon University

Abstract

This paper envisions a distributed network storage service with Quality-of-Service (QoS) guarantees, and describes its architecture and key mechanisms. When fully realized, this service architecture would be able to support, in one integrated framework, network storage services ranging from best-effort caching to replication with performance guarantees. Content owners could, through the use of standardized protocols, reserve network storage resources to satisfy their application-specific performance requirements. They would be able to specify either the number and/or placement of the replicas, or higher-level performance goals based on access latency, bandwidth usage or data availability. The network storage provider would then optimally allocate storage resources to meet the service commitments, using leftover capacity for best-effort caching. Content consumers would then retrieve the nearest copy of the data object, be it from a replica, cache, or the original source, in a completely transparent manner. Furthermore, a distributed network storage infrastructure should be integrated with the existing transmission-based QoS framework so that applications can select the optimal combination of storage and transmission resources to satisfy their performance requirements.

The SDSC Storage Resource Broker (SRB) is a client-server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets. SRB, in conjunction with the Metadata Catalog (MCAT), provides a way to access data sets and resources based on their attributes rather than their names or physical locations.

Notes:

- Focuses on data transmission QoS, replica and caching, or data locator services, not really storage management
- Does not take into account characteristics of the actual storage devices
- Does not provide interface to advanced features like on-demand capacity growth, or per-MB billing, meta-data and object properties, reliability, backup, content-type, etc

Conclusion

- Storage is a key fabric component of Grid Computing, and must be provisioned with adequate service quality levels and data and storage management
- Such provisioning cannot be done only at the grid-wide level, but also at the cluster, storage networks, storage devices and even meta-data levels
- Good integration is also required between the Distributed and Intra-cluster Storage areas
- Singapore, being very active in both Grid and Storage R&D, is therefore well positioned to undertake this collaborative effort

Thank You

Additional Information

Additional Information on Existing Projects

Institute for Infocomm Research

- Data-in-Network High-Speed Cache
  - Provide “persistent” data as a form of cache within conventional network infrastructure through the use of “data loops” created and partitioned using standard MPLS/GMPLS frameworks
  - Through the use of forward error correction schemes and multiple data loops, transparent data recovery is implemented thus providing cache data with high levels of reliability
  - Unique qualities includes shortening speed-of-light latency access to data and providing simultaneous access to data from multiple clients
  - Supports peer-to-peer read and write real-time dynamic data for Grid and Distributed Computing

School of Computer Engineering, NTU

- IBP Replica Management and exNode File Warehouse Systems
  - Integrates the Internet Backbone Protocol, which provides storage, and the Globus Replica Catalogue, for its replica management and cataloging, into a IBP Replica Management system
  - Provides management capabilities and functions to IBP-based storage depots already deployed globally, including relevant APIs for application integration
  - Includes required file management, expiration and mapping features between IBP exNodes and GRC catalogues through the exNode File Warehouse
  - Designed to provide globally managed storage resources to Grid and distributed computing clients

Network Storage Division, DSI

- HyperSCSI Data Transport Protocol
  - HyperSCSI is a open-source software-based data transport protocol for the development of Ethernet-based Storage Area Networks
  - Unlike conventional solutions, it does not run on TCP/IP at all and uses raw Ethernet instead
  - It matches the performance of Fibre Channel but at 10 times lower cost
  - This technology is designed as a drop-in replacement for Fibre Channel SANs for parallel compute clusters

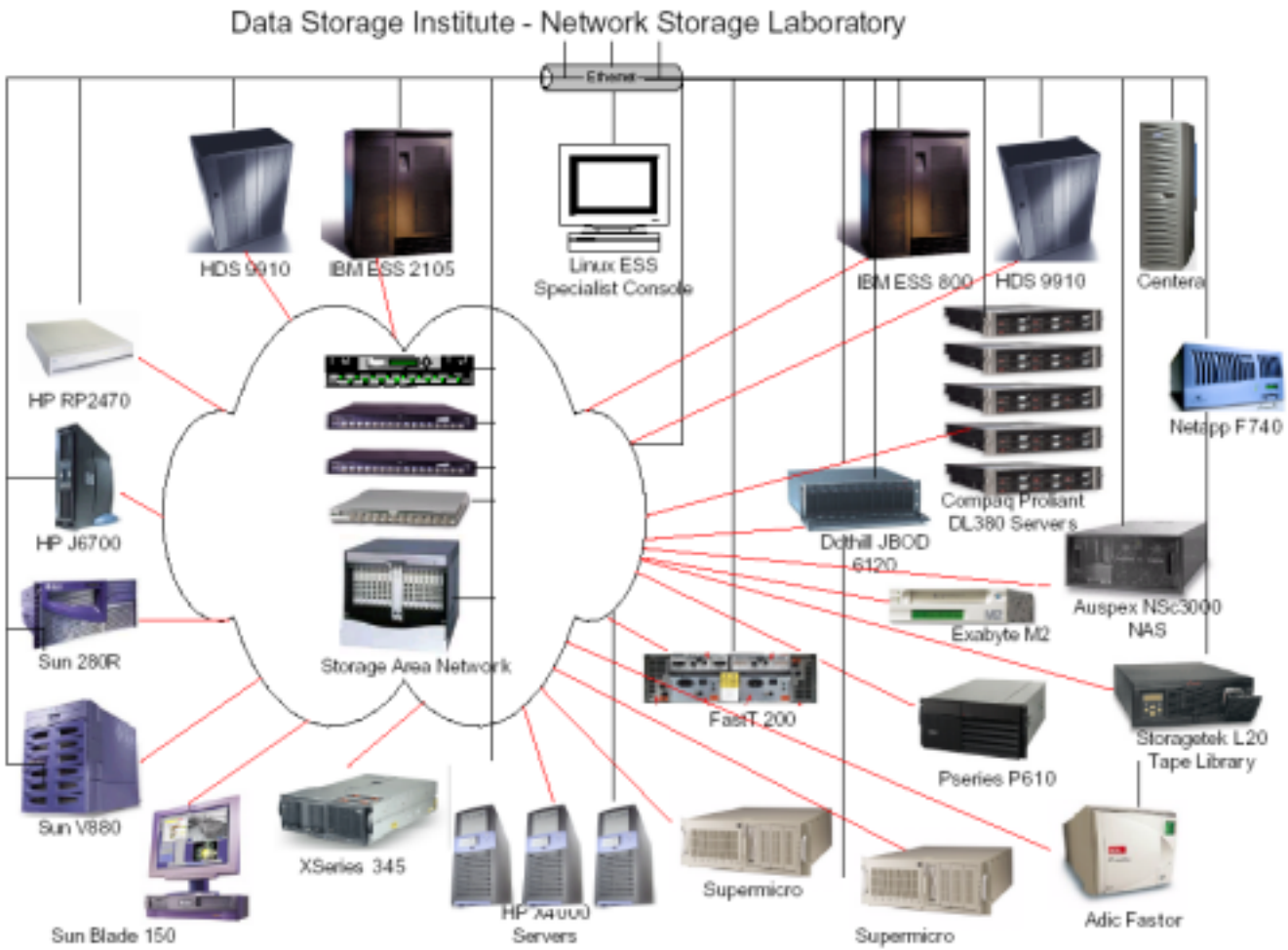
Network Storage Division, DSI

- Object-based Storage Solution
  - Implementation of new emerging SNIA OSD standards for data sharing in network storage environments with communications conducted over Fibre Channel or IP-based networks
  - Supports object storage in disk and cache with centralised management through an object manager or meta-data controller
  - Superior transaction processing performance (I/O per second and throughput) as well as high scalability
  - Designed to support large scalable storage (eg 3PB ASCI Purple 2005 and 6PB NASA RDS 2006)

Network Storage Division, DSI

- Cluster File Systems Work
  - In-depth performance analysis, comparison and characterisation of several commercial and open-source cluster and non-cluster file systems using Ethernet-based Storage Area Networks
  - Results included studies on multi-channel communications, starvation and loading in multi-client environments, performance analysis and storage configurations to support multiple clients in a parallel compute cluster
- Storage Performance Testing and Benchmarking
  - Diverse capabilities and experience in testing, benchmarking and analysing networks and network storage performance and reliability

DSI NST Lab Storage Resources



- Approximately 9TB of online heterogeneous primary disk space for development and testing
- Various Windows, Linux, Sun, HP and Compaq servers for specific services and functions
- 10-node dual PIII 1.1GHz CPU compute-cluster