OCEANOGRAPHIC

Architecture, Implementation, and Deployment of a High Performance, High Capacity Resilient Mass Storage Server (RMSS)

Eighteenth IEEE Symposium on Mass Storage Systems
Ninth NASA Goddard Conference on Mass Storage Systems and Technologies
April 17-20, 2001
San Diego, CA

AUTHORS

Terry Jones

Beata Sarnowska

Logicon Information Systems and Services (LISS)

With

John Kothe, LISS

Frank Lovato, NAVOCEANO MSRC

David Magee, NAVOCEANO MSRC

ACKNOWLEDGEMENTS

USAF Aeronautical Systems Center (ASC) Major Shared Resource Center (MSRC)

Early Deployment of ASC MSRC RMSS Cluster Uncovered Issues Which, Through Information Sharing, Saved Us Many Valuable Hours

OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

Cluster Configurations

High Performance Computing Most Concerned with 3 Types of Clusters:

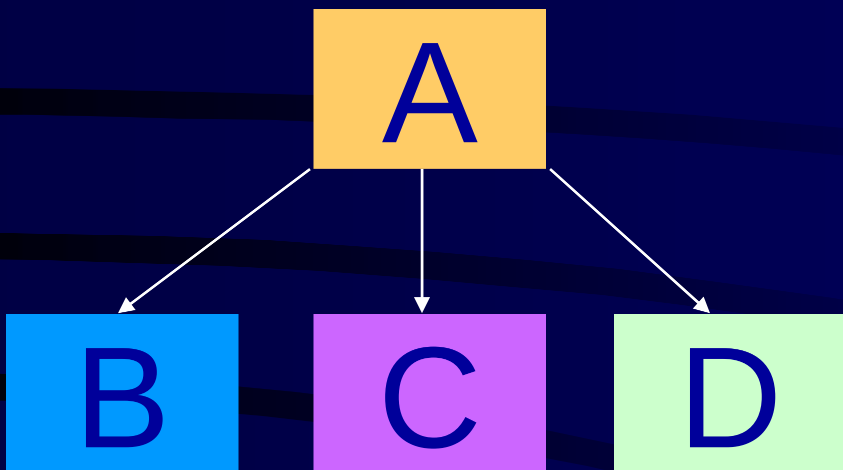
- **Capacity Cluster**
- **Capability Cluster**
- **High Availability**

Many Cluster Types Defined.

See IDC Report: *Clustering and High-Performance System Interconnect on Intel Architecture, 1995-2001*. D. Floyer, Report #14714, February 1998

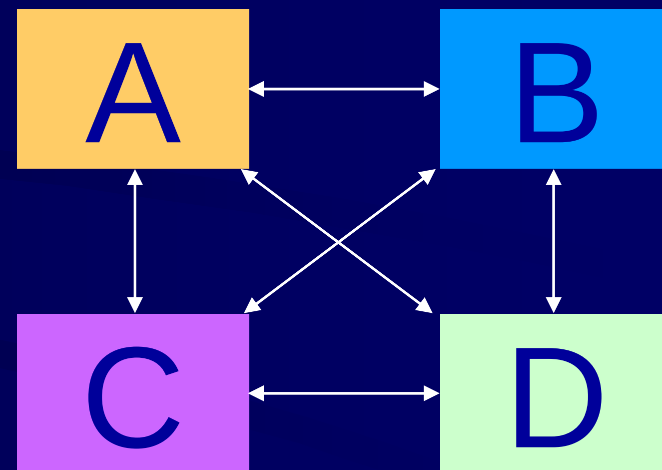
Cluster Configurations

Capacity



Application Runs on One Node; Runs Distributed Across All Nodes Based on Scheduler Policies

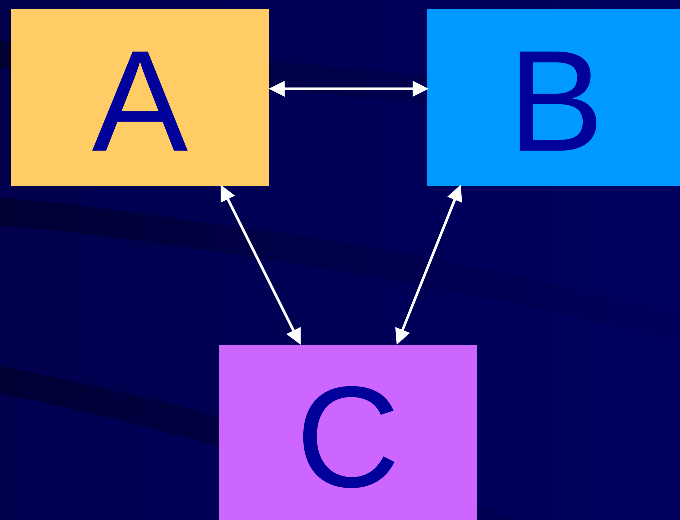
Capability



Application Runs Across As Many Nodes As Needed At the Same Time; Runs Scheduled By Available Nodes

Cluster Configurations

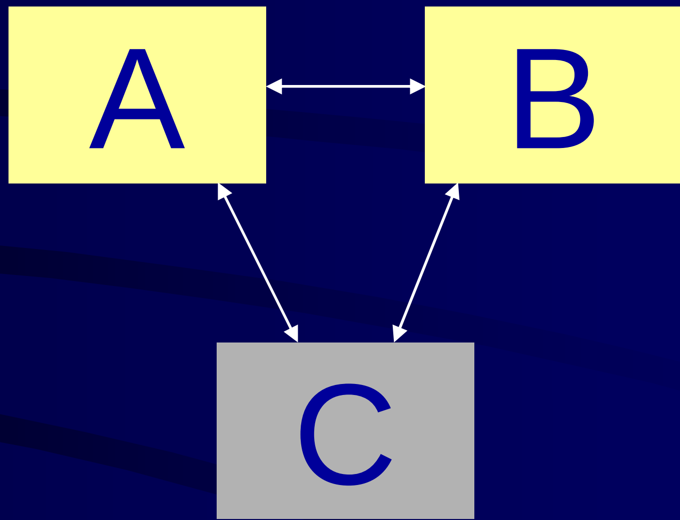
High Availability



Simplified View:
One Node Takes Over if Another Fails

HA-Cluster Configurations

Active-Passive

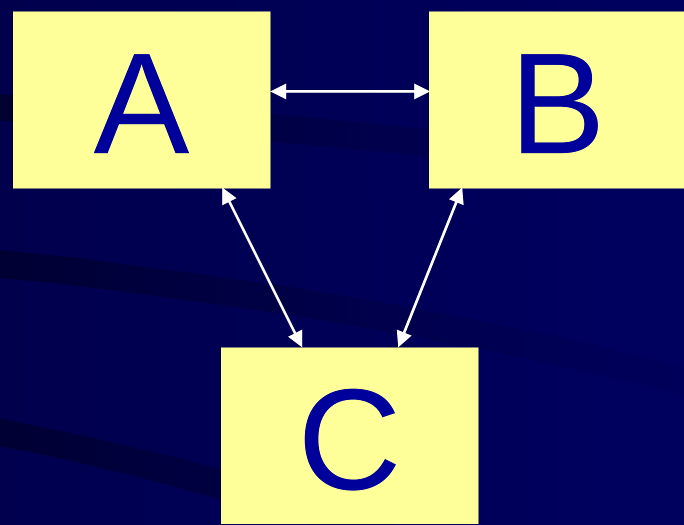


Node C Takes Over Applications on A or B Should Either A or B Fail

Suitable for Inexpensive Nodes

HA-Cluster Configurations

Active-Active



Nodes A, B and C Running Applications
Should One Node Fail, Remaining Nodes Take Over Failed Node's Applications
Cost-Efficient for Expensive Nodes

HA-Cluster Configurations

- Node Failure Does Not Necessarily Mean That Node Has Completely Crashed

- Partial Failure of Node, Such as Disk Controller, Network Adapter or Other Component Can Mean That Node Has "Failed"

- Node Failure Defined As:
  Any Hardware or Software Failure on a Node That Renders the Node Incapable of Supporting Its Application(s)

High Availability Technology

Machine Centric: Fault Tolerance
Focused on Increased Reliability to Improve Availability. Reduces MTTF with additional H/W Resources.

Application Centric: High Availability Clusters
Focused on Providing Resources to Applications from Pooled Devices in Clusters

High Availability Clusters

HPC Centers Require High-End Performance Mass Storage Servers

- **High Capacity:** Petabytes Stored Data
- **High Volume:** 100,000's File Requests/Day
- **High Data Traffic:** 2+ Terabytes/Day
- **High Availability:** 99.99% Desired

High Availability Clusters

NAVOCEANO MSRC RMSS DESIGN Combines Fault Tolerance and High Availability

Goals:

- Achieve High Reliability, High Capacity
- Position for Emerging Technologies
- Re-Engineer Existing Mass Storage Servers for Long-Term Sustainability and Growth Options

OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

Requirements Definition

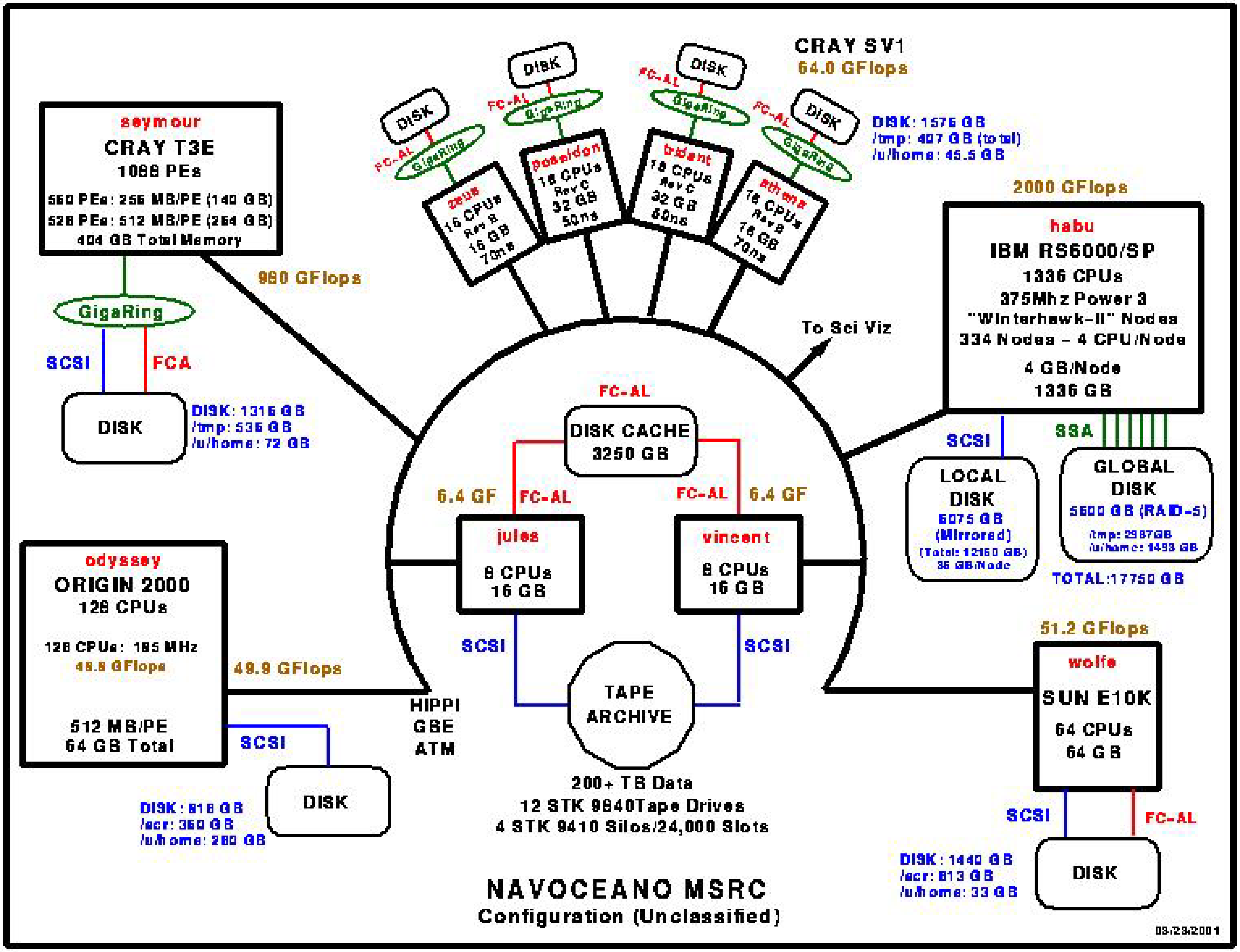
- Uncertain Support Future of Existing Mass Storage Solution (DMF/Unicos)
- Required Non-Proprietary Archive Data Format
- Disaster Recovery
- Scalable Solution Required
  - Scalable to 500 TB – 1,000 Total Capacity
  - Up to 3 TB+ Data Traffic Per Day
- High Data Availability to User Community
- System Internal Resiliency

NAVOCEANO MSRC

Total Capacities

TOTAL COMPUTE CAPACITY:	3158 GFLOPS
TOTAL NUMBER OF CPUs:	2696
TOTAL SYSTEM MEMORY:	1996 GB
TOTAL DISK STORAGE:	26,250 GB

MSRC CONFIGURATION



NAVOCEANO MSRC MASS STORAGE UTILIZATION GROWTH

Date	TOTAL TB	Growth
Dec 95	10.2	1X
Dec 96	20.1	2X
Dec 97	39.7	4X
Dec 98	78.4	8X
Dec 99	153.4	15X
Dec 00	223.9	22X

WORKLOAD PROJECTIONS

- **Storage Growth Modeled by Exponential Function From 1995 to July 2000**
- **Growth Rate Model After July 2000 Modeled By 10th Order Polynomial**

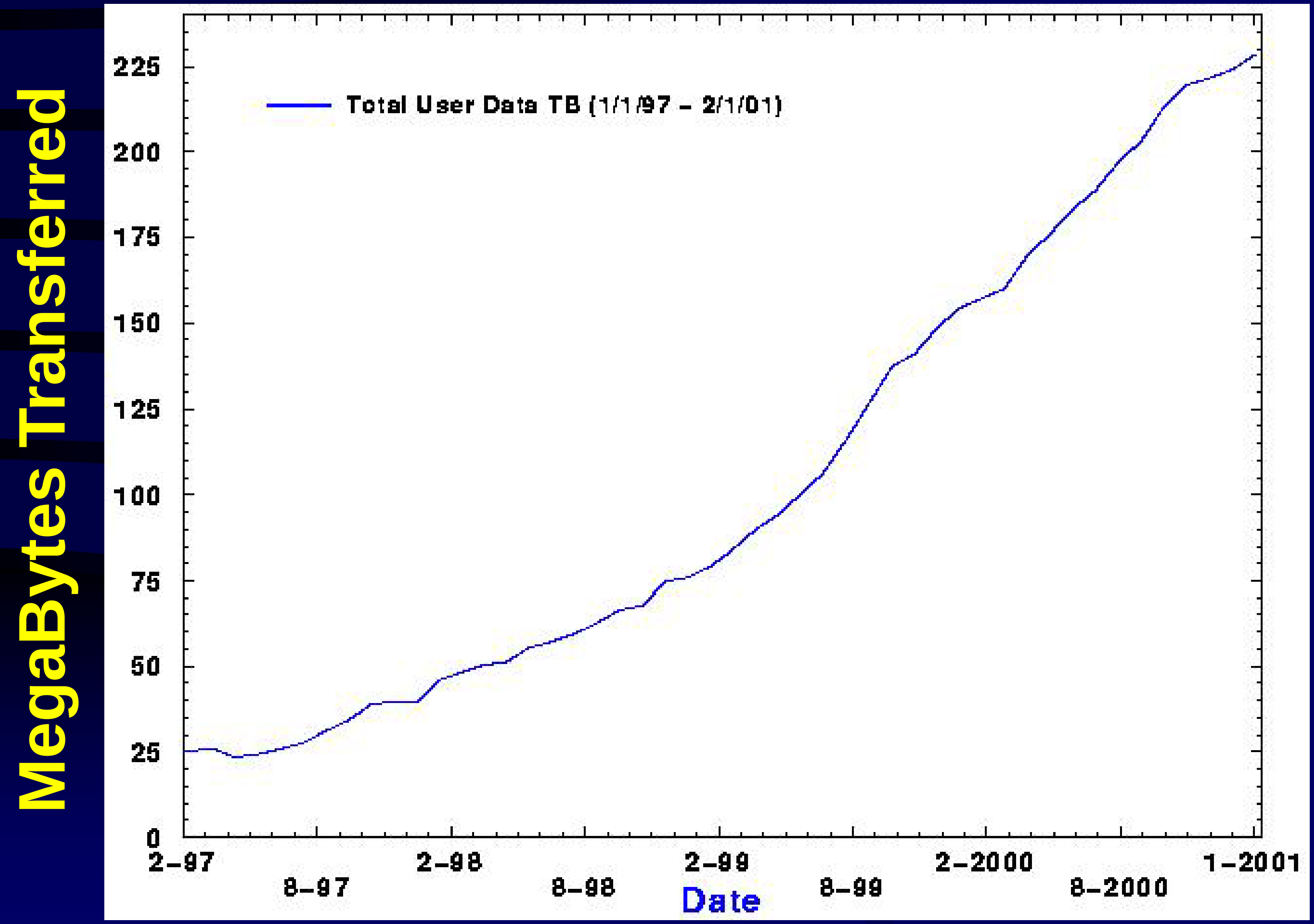
WORKLOAD PROJECTIONS

- Anticipated Advances in Key Model Resolution and Accuracy Factored into Growth Model

- Yielded Projected Six-Fold Increase in Storage Requirements

NAVO MSRC Mass Storage Utilization Trend

Total User Data Storage Trend Through February 1, 2001

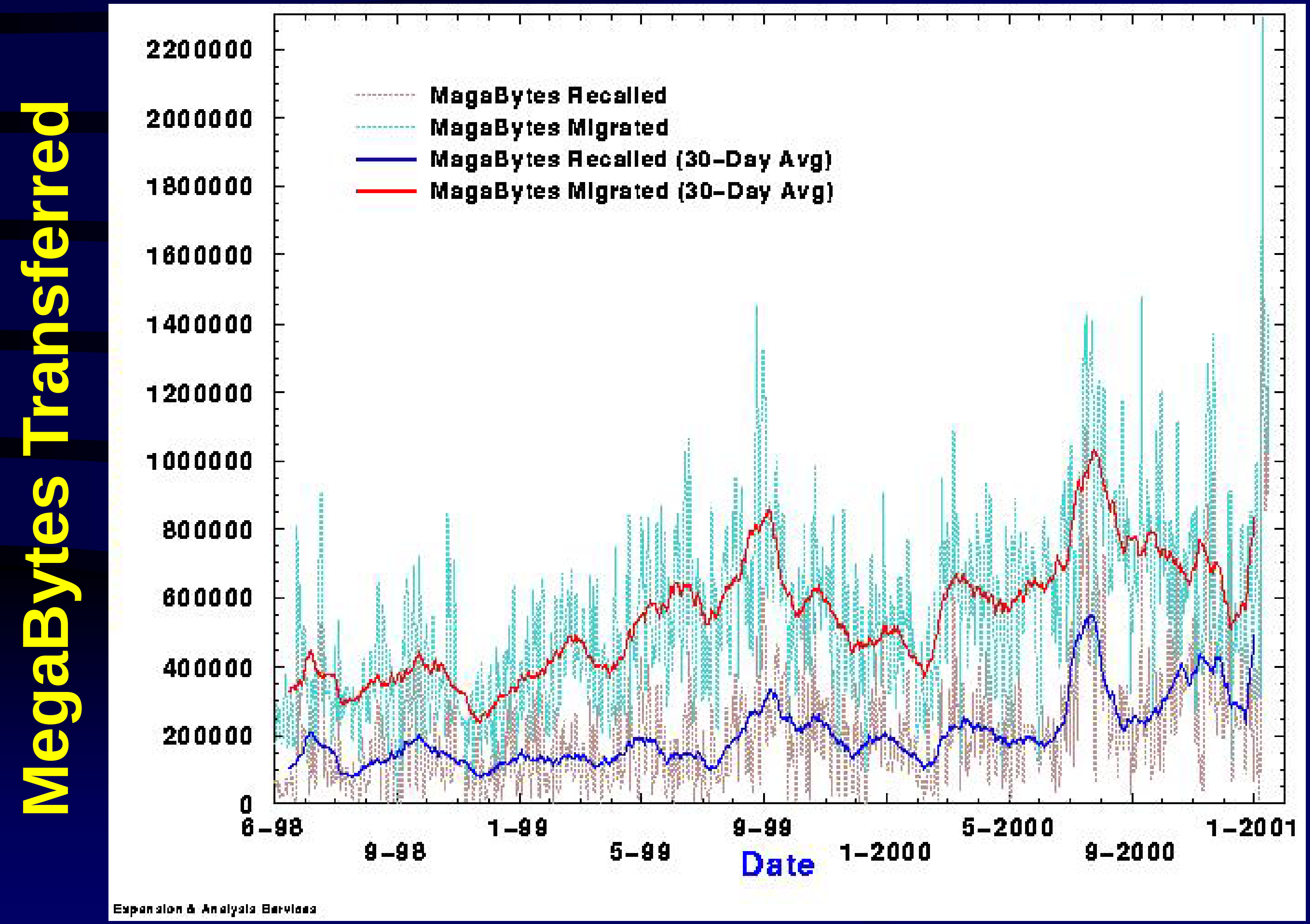


DATA DISTRIBUTION

- Follows the 90/10 Rule
  - 90% of Files/10% Storage Space
  - 10% of Files/90% Storage Space
- Single Project Has 60% of Data

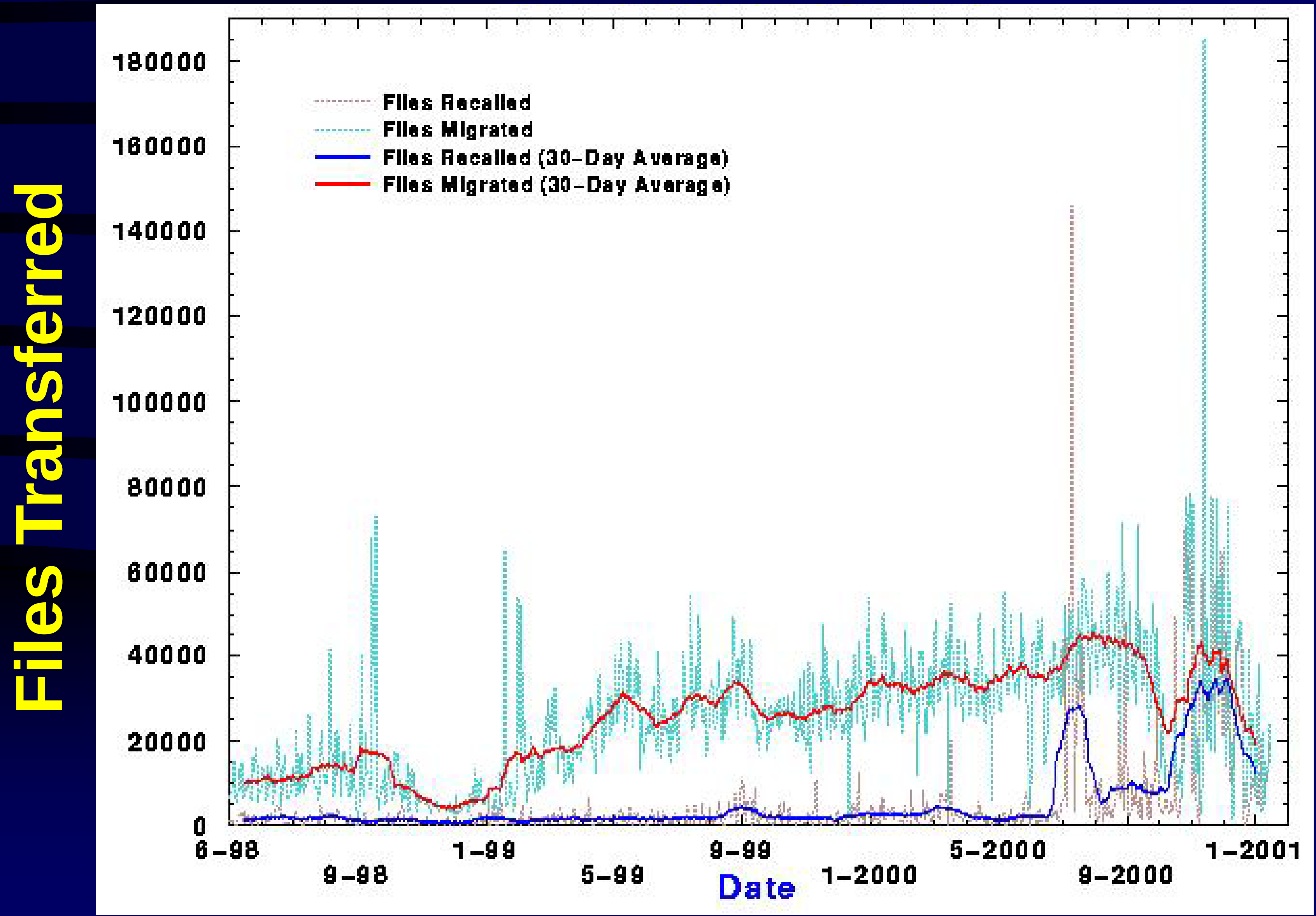
TRANSACTION ANALYSIS

DMF File Traffic – June 1998 to January 2001



TRANSACTION ANALYSIS

DMF Migrate/Recall Traffic – June 1998 to January 2001



DESIGN SPECIFICATIONS

Design Criteria	Node Design Limits	Cluster Design Limits
GB Data Network Traffic per Day	1024	2048
Data Archive/Recall Ratio	33%	33%
Target Disk Cache Retention Period	72 Hours	72 Hours
Largest File	25 GB	25 GB
Number of Files	25 Million	50 Million
Archived Filesystems	2	4
Scalability (3 Year Lifecycle)	6X	6X
Sustained Network Bandwidth	110 MB/Sec	220 MB/Sec
Sustained Tape Bandwidth	96 MB/Sec	192 MB/Sec
Peak Disk Bandwidth	110 MB/Sec	110 MB/Sec
I/O Memory Bandwidth	25.6 GB/Sec	51.2 GB/Sec
GB Memory	16	32
CPUs	12	24
System Boards	8	16

OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

Solution Development

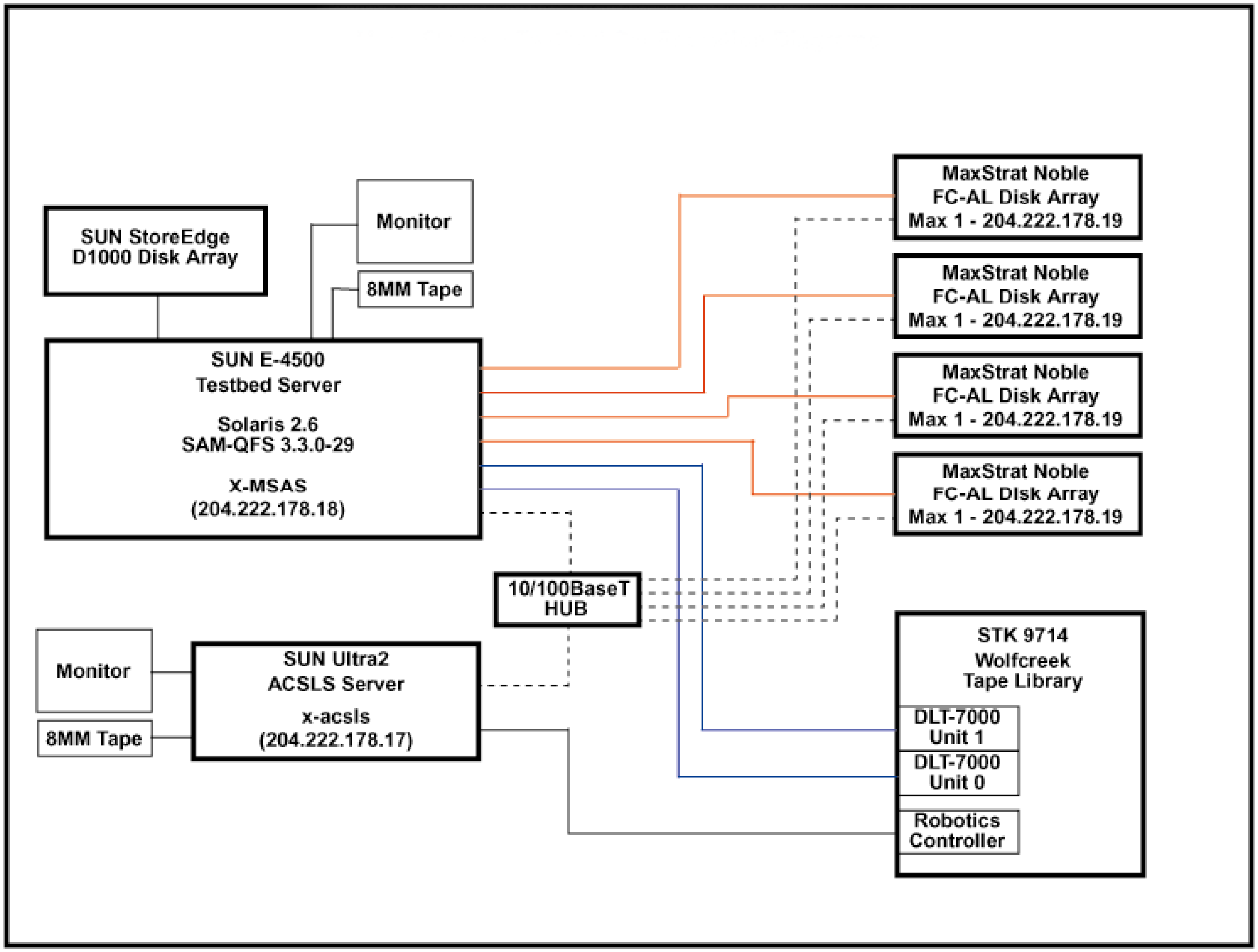
- Researched Architectural Models for Storage Solutions
  - Fibre Channel RAID Disk on Arbitrated Loops New and Stable with Good Potential
  - Storage Area Networks Still Evolving
  - High Availability Clustering Possible with Sun and SGI Architectures
- Two COTS HSM Solutions Identified
  - Data Migration Facility (DMF) Under IRIX [SGI, Inc.]
  - SAM-QFS Under Solaris [LSC, Inc., Now SUN, Inc.]

Mass Storage Testbed

- Testbed Constructed to Evaluate SAM-QFS
  - SUN E4500/Solaris 2.6
  - Fibre Channel Disk SUN T3
- Test Results Validated SAM-QFS and SUN Fibre Channel as Viable Solution
- Resulted in System Design and Specification

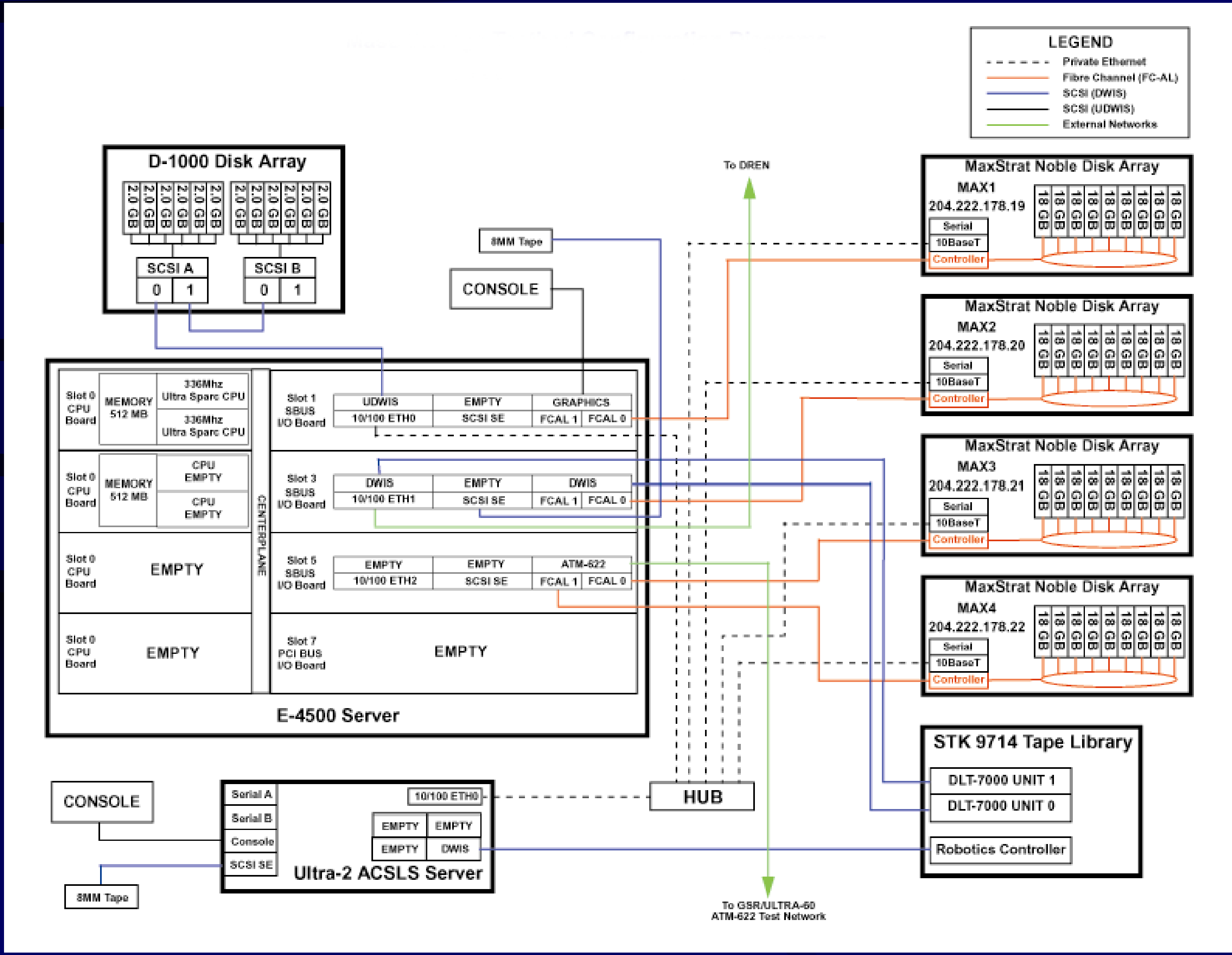
Mass Storage Testbed Configuration Diagrams

High-Level Configuration Diagram



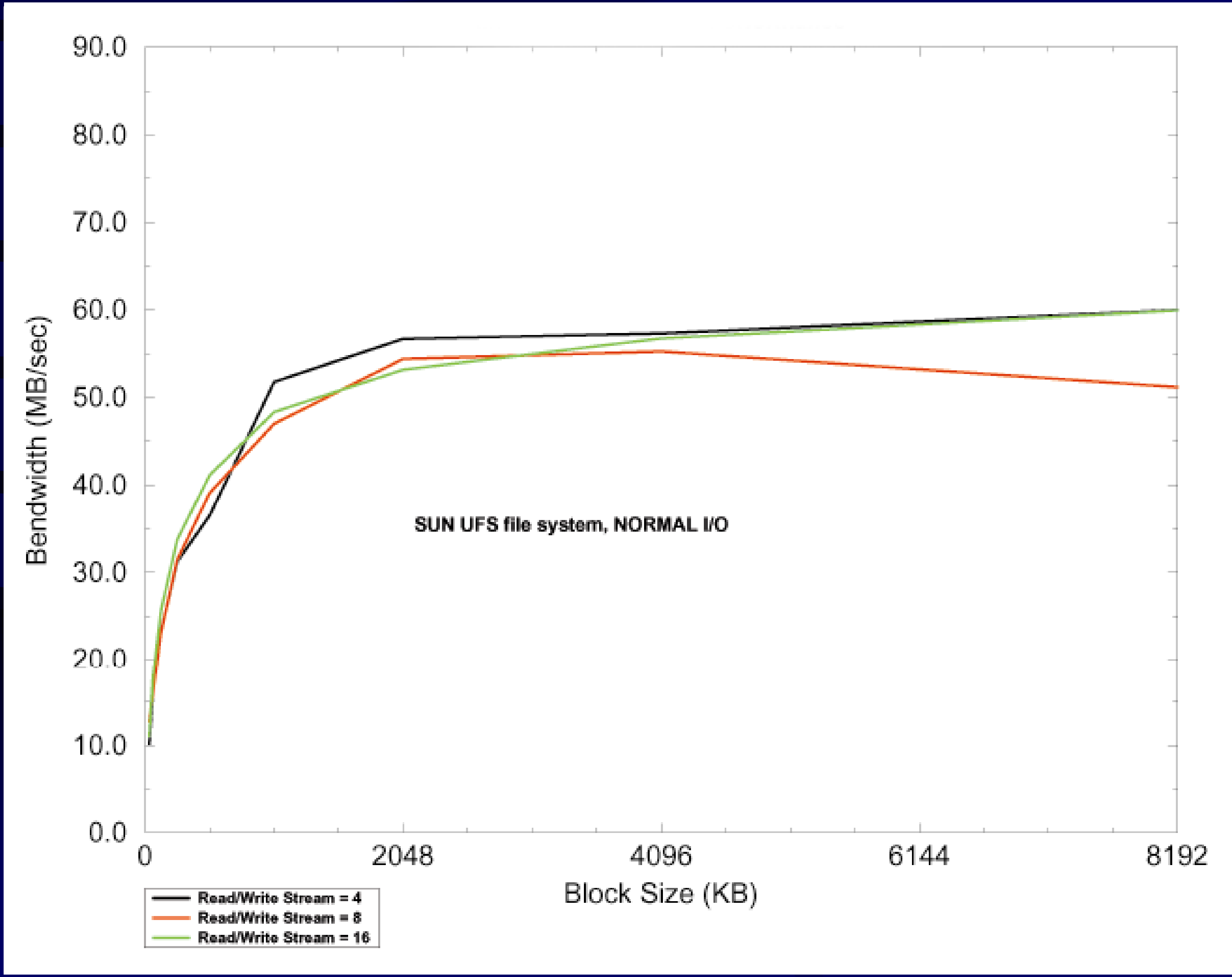
Mass Storage Testbed Configuration Diagrams

Low-Level Configuration Diagram



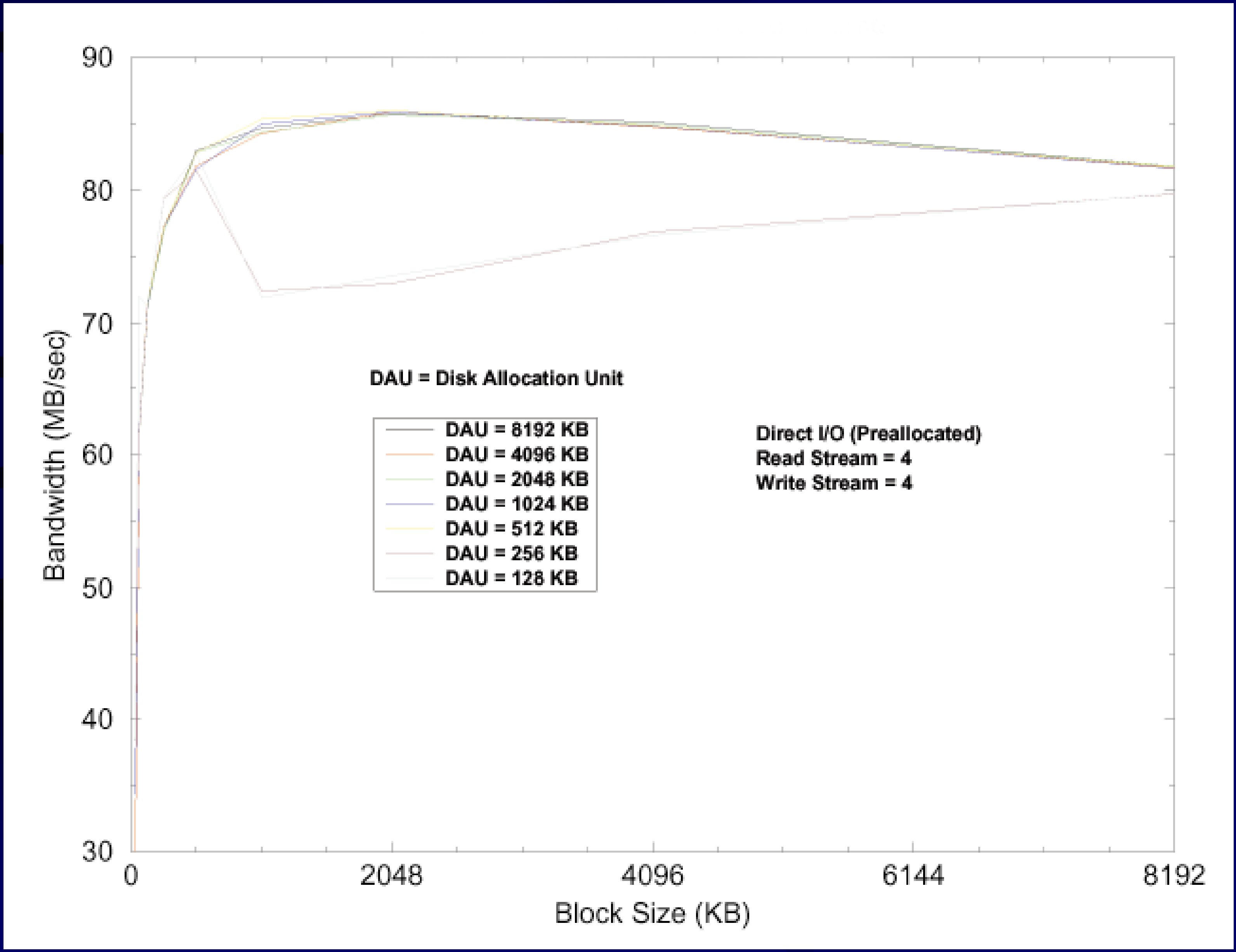
NAVO MSRC Mass Storage Testbed

MaxStrat Noble Write Performance



NAVO MSRC Mass Storage Testbed

MaxStrat Noble Write Performance

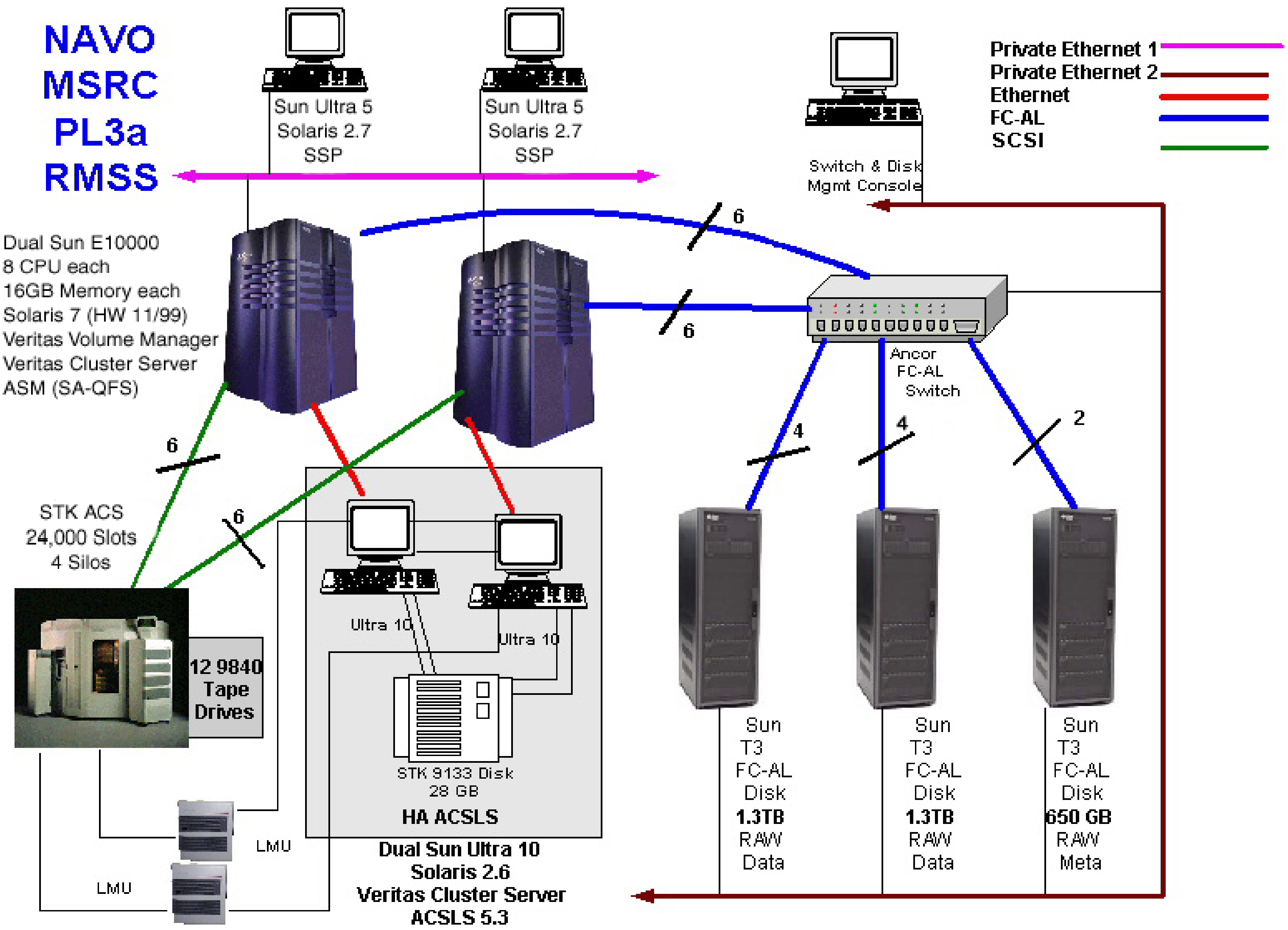


SYSTEM DESIGN

- **High Availability File Server Cluster**
  - Two Nodes: SUN E10K, 12 CPUs/8 System Boards, 16 GB
  - HIPPI, ATM (OC-3, OC-12c), 10/100 Ethernet, FC-AL, STK 9840 Tape (SCSI)
  - 3.2 TB SUN T3 FC-AL Disk (2.6 TB Data, 0.6 TB Metadata)
  - Veritas Cluster Server (VCS) Active-Active High-Availability Cluster Configuration
- **STK HA-ACSLS Tape Robot Controller**
  - Two Nodes: SUN Ultra10
  - Veritas Cluster Server Active-Passive Cluster Configuration

RMSS CLUSTER RESOURCES

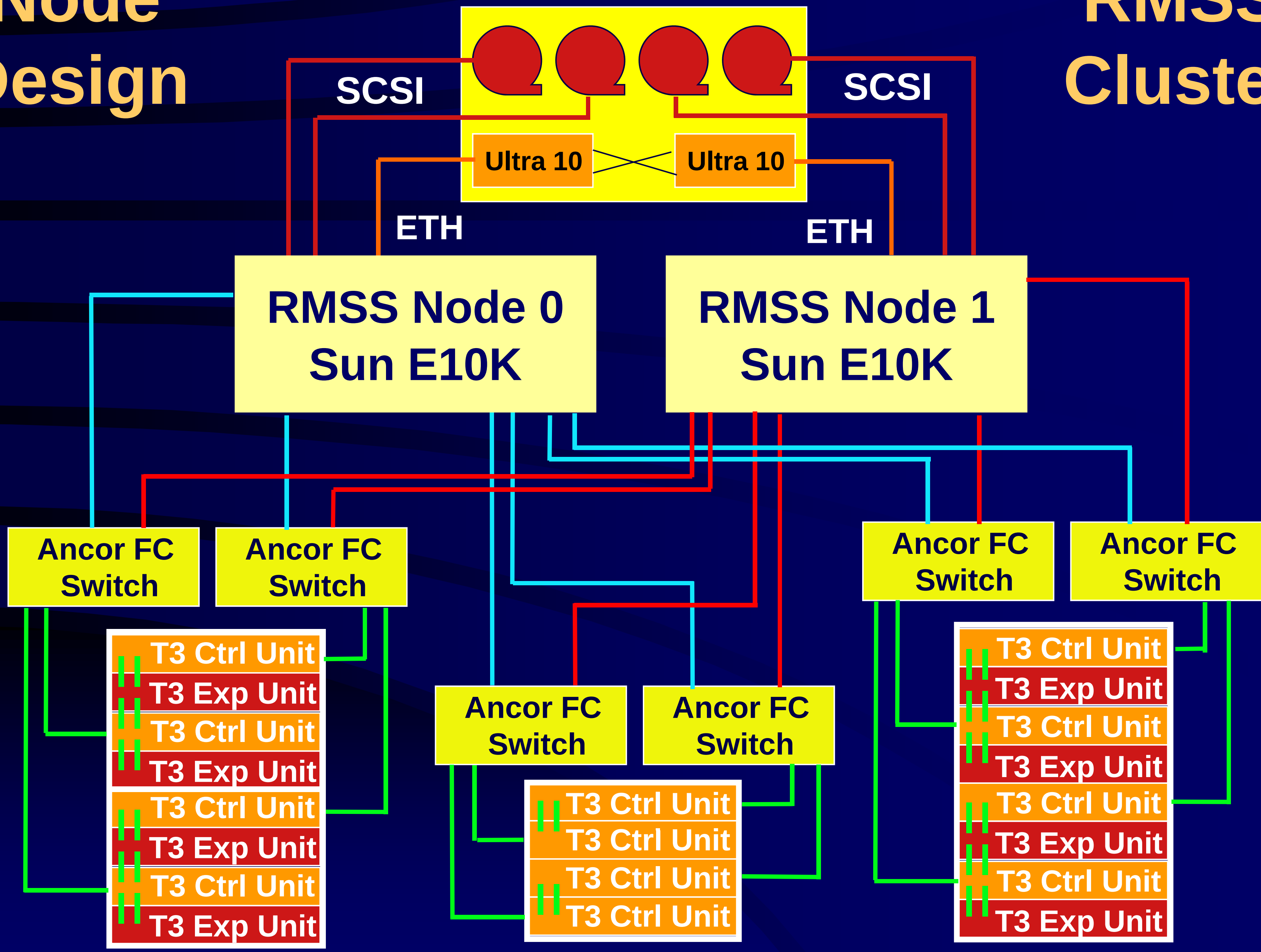
Resource	Node 0	Node 1	Total
System Boards	8	8	16
CPUs	12	12	24
Memory	16 GB	16 GB	32 GB
Fibre Channel Adapters	6	6	12
SCSI Adapter Slots	9	9	18
System Disk (D1000 18 GB JBOD)	10	10	20
Tape Drives	6	6	12
HIPPI NICs	2	2	4
ATM OC12c NICs	1	1	2
Quad-Fast Ethernet NICs	1	1	2



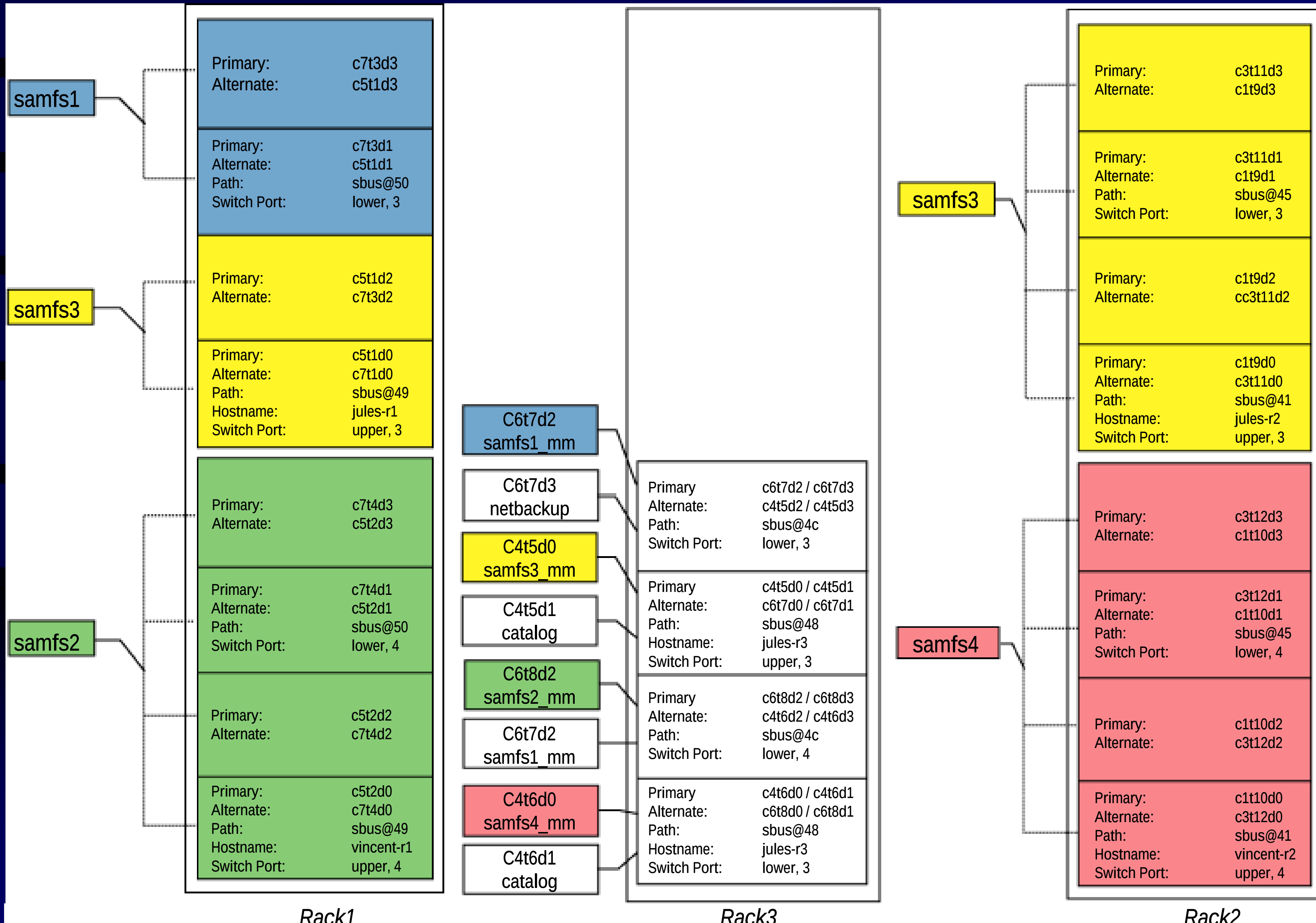
NAVO
MSRC
PL3a
RMSS
Sun Ultra 5
Solaris 2.7
SSP
Sun Ultra 5
Solaris 2.7
SSP
Switch & Disk
Mgmt Console
Private Ethernet 1
Private Ethernet 2
Ethernet
FC-AL
SCSI
Dual Sun E10000
8 CPU each
16GB Memory each
Solaris 7 (HW 11/99)
Veritas Volume Manager
Veritas Cluster Server
ASM (SA-QFS)
6
6
6
6
Ancor
FC-AL
Switch
4
4
2
STK ACS
24,000 Slots
4 Silos
12 9840
Tape
Drives
Ultra 10
Ultra 10
STK 9133 Disk
28 GB
HA ACSLS
Dual Sun Ultra 10
Solaris 2.6
Veritas Cluster Server
ACSLS 5.3
LMU
LMU
Sun
T3
FC-AL
Disk
1.3TB
RAW
Data
Sun
T3
FC-AL
Disk
1.3TB
RAW
Data
Sun
T3
FC-AL
Disk
650 GB
RAW
Meta

Node Design

RMSS Cluster



Naval Oceanographic Office - MSRC STOREdge T300 Layout




ENTERPRISE

SOFTWARE ELEMENTS

- **SAM-QFS (Storage Archive Manager/Quick File System)**
  - **OEM: LSC, Inc. (Acquired By SUN, Feb. 2001)**
  - **Hierarchical Storage Manager**
  - **Principal Application of RMSS**
- **Veritas Cluster Server (VCS)**
  - **Provides Resiliency Functions for Cluster**
  - **Detects Service Level Failures on Either Node**
  - **Initiates Recovery to Shift SAM-QFS Service from Failed Node to Remaining Node**
  - **Also used on HA-ACSLS Cluster**

SOFTWARE ELEMENTS

- **Veritas Volume Manager**
  - **Provides Dynamic Multipathing**
  - **Provides Control Over System Disks**
- **SOLARIS Operating System**
  - **Solaris 7**
  - **Nodes Run in 32-Bit Mode**
- **ACSLS Client (Bundled in SAM-QFS)**
  - **Interfaces Between SAM and STK Robotics**

SAM-QFS

- **Basic Functionality**
  - **Archive**
    - **Copies Files From Disk Cache to Permanent Storage on Tape. Copy Performed as Soon As Feasible After File is Created or Modified in Cache**
  - **Release**
    - **SAM-QFS Manages Disk Cache Free Space by Releasing Files When Site-Definable Thresholds are Reached. Thresholds Set by Filesystem**

SAM-QFS

- **Basic Functionality**
  - **Stage**
    - **Retrieval of Files from Tape Archive back to Disk Cache. A Stage Occurs for Released Files When Either an Explict Stage Command is Entered or Indirectly When File is Referenced in a Local Command, or by an FTP/RCP Command on a Remote System. Indirect References at Remote System for NFS Exported SAM-QFS Filesystems Also Causes Automatic Stage**

SAM-QFS

- **Basic Functionality**
  - **Recycle**
    - **Reclaims Tape Space as Archive Copies Become Obsolete When Replaced by Newer Archive Versions**

QFS

- Implemented Using Standard SOLARIS Virtual Filesystem (vfs/vnode) Interface
- Requires No Modifications to SOLARIS Kernel
- Supports Multiple Filesystems Up To 200 Partitions Each
- Allows Separation of Data and Metadata

QFS

- 64-Bit Filesystem (2^{64}-1 Files)
- Maximum File Size: (2^{64}-1 Bytes)
- Supports Striping (RAID-0) or Round-Robin Allocation
- QFS Supports a Fully Adjustable DAU from 16KB to 65,535KB Blocks
- Very Useful For Tuning Filesystem to Physical Storage Media

SAM-QFS

- **SAM-QFS and the SUN T3 Disk Array**
  - **SUN T3 Disk Supports 16KB, 32KB, and 64KB Physical Block Sizes**
    - **16 KB for Smaller Transactions**
    - **32 KB Better for Random Mix of Sequential**
    - **64 KB Best for Large I/O Requests**
  - **Block Size Choice Impacts How T3 Cache Memory Is Used to Avoid Read/Modify/Write Operations**

SAM-QFS

- **SAM-QFS and the SUN T3 Disk Array**
  - **Tune T3 Disk Array Physical Allocation Unit to Characteristics of Files within Filesystem**
  - **Then Tune QFS DAU to T3 Allocation Unit**
  - **Performance Improvement Significant**
  - **QFS Peak Theoretical Bandwidth: 1.5 GB/Sec**
  - **Bandwidth Independently Measured at Over 1 GB/Sec**

SAM-QFS

- **Metadata Disks**
  - **Small I/O Requests (512 Bytes Each)**
  - **Individual Requests, Low Rate**
  - **Little to No Locality of Reference**
  - **Use 16KB T3 Physical Allocation Size**

SAM-QFS DAU/T3 BLOCKSIZE TUNING

File System	Characteristics	QFS DAU	T3 Block Size
A	Smaller Files Mostly Sequential	32 KB	32 KB
B	Very Large Files Mainly Sequential Few Requests	64 KB	64 KB
C	Mixed File Sizes Many Requests	32 KB	32 KB
D	Many Small Files Mixed Sequential High Access Rates	32 KB	32 KB

VERITAS CLUSTER SERVER

- **Design HA-Cluster Actions**
  - **Define Anticipated Failures and Required Response to Each**
  - **Develop Failover Matrix**
  - **Identify System Component Responsible**
  - **Develop VCS Resource Tree**
  - **Define and Develop Agent Scripts Required**

RMSS CLUSTER FAILOVER MATRIX

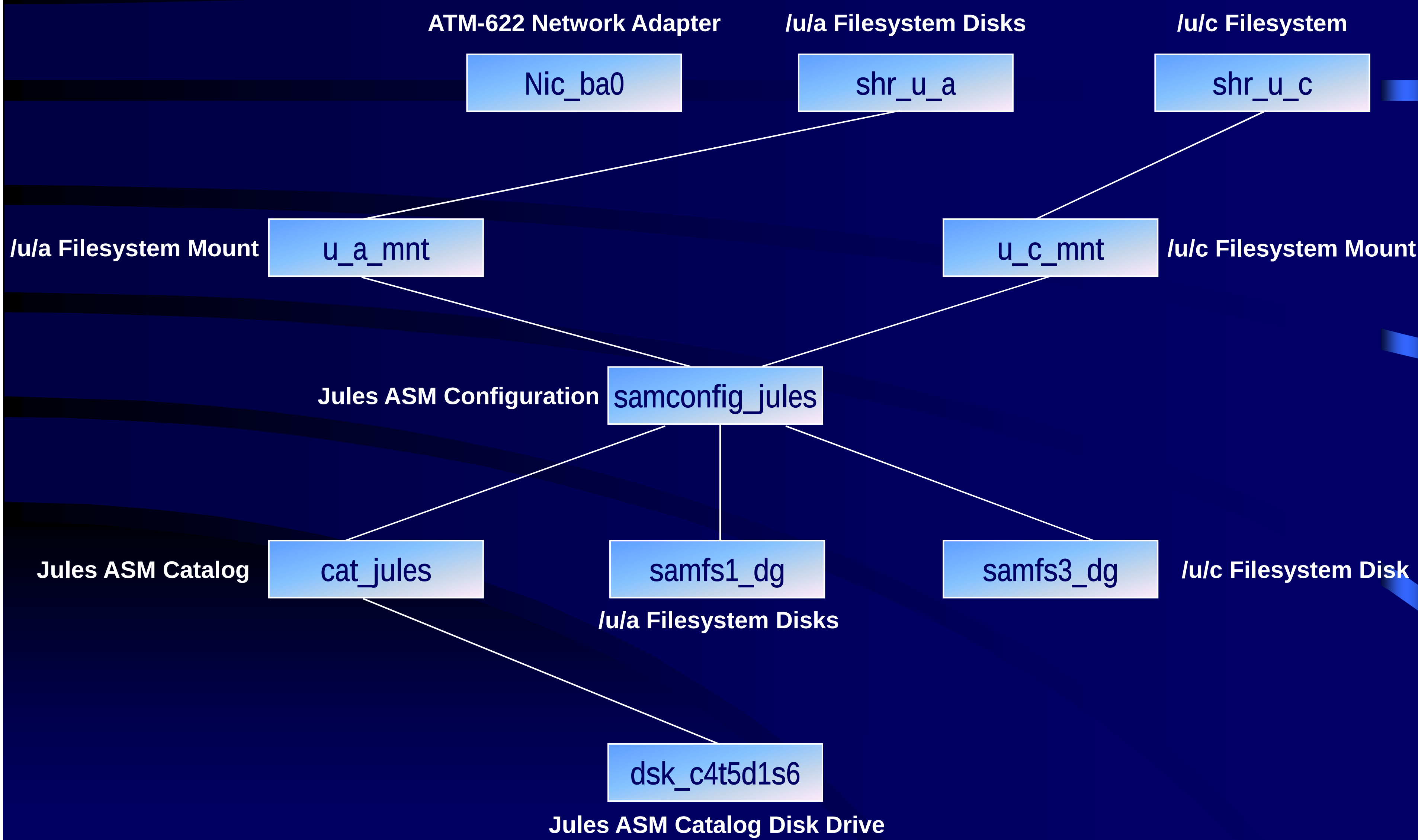
Failure	Software Element	Action Taken	Filesystem Resultant State
Disk Spindle Permanent Error	T-300 Firmware	Use RAID Parity & Mirror Data to Reconstruct Data	Remains on Node
User Filesystem Disk Controller/Path	VVM	Internal Switch to Alternate Path	Remains on Node
Tape Drive Error	ASM	Offlines Tape Drive	Remains on Node
SAM Catalog SCSI Disk	VCS	Failover QFS Filesystems	On other Node
T-3 Disk Controller Failure	VCS	Failover QFS Filesystems	On other Node

RMSS CLUSTER FAILOVER MATRIX

Failure	Software Element	Action Taken	Filesystem Resultant State
QFS Filesystem Unmount	VCS	Failover Filesystem	Other Node
Loss of QFS Catalog	VCS	Failover Filesystem	Other Node
System Crash	VCS	Failover Filesystems	Other Node
System Shutdown	Solaris	System Shuts Down	Unavailable
System Shutdown (with manual Failover)	VCS Solaris	Failover Filesystems System Shuts Down	Other Node
Manual Tape Drive Offline	ASM ASCLS	Offlines Tape Drive	Remains on Node

VCS CONFIGURATION

Resource Dependency Tree (Node 0)



OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

NETWORK CONNECTIVITY

- Default Network: ATM OC12c LAN
  - Peak Bandwidth: 77.75 MB/sec
  - Anticipated Peak Sustained: 62 MB/sec
  - Production Sustained Rates: 50 MB/sec
    - Network Buffer Shortages Identified
    - Tuning (In-Work) Will Improve to Near Anticipated Bandwidths

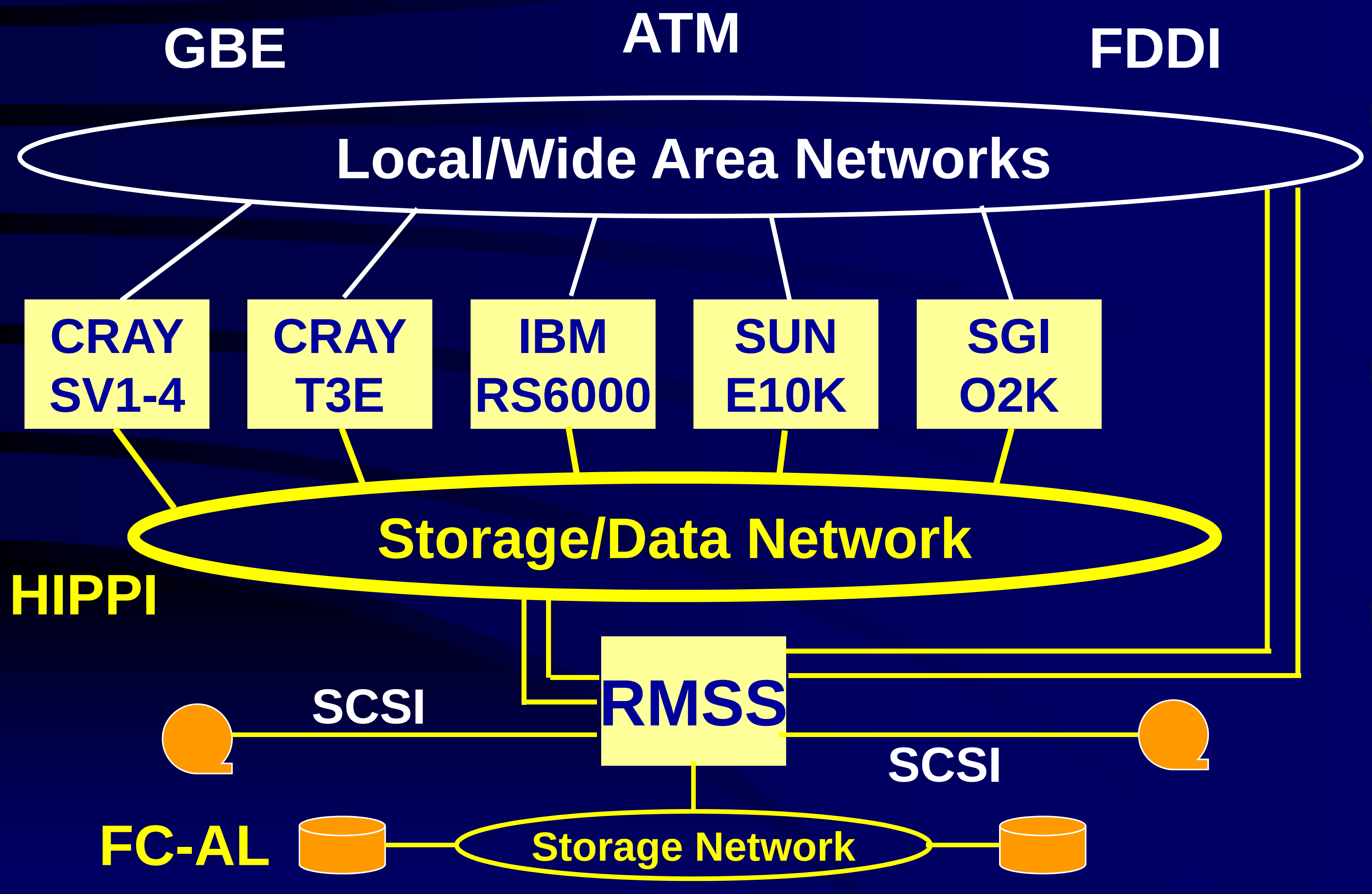
NETWORK CONNECTIVITY

- **Data Transfer Network**
  - **HIPPI**
    - **Peak Theoretical B/W: 100 MB/Sec**
    - **Anticipated Sustained B/W: 50 MB/Sec**
    - **Production B/W: 50 – 55 MB/Sec**
  - **Two HIPPI NICs Per Node**
    - **One for Connectivity to Data Network**
    - **One Dedicated to Transition**

NETWORK CONNECTIVITY

- Users Log In to ATM For File Management
- Users Transfer Files Between RMSS and HPC Platforms Via HIPPI Network
- SAN Concepts in Use For Over a Decade

STORAGE NETWORK CONFIGURATION



SECURITY

- Security Extremely Important for RMSS
  - Data Represent Years of Research
  - Substantial Investment by Users
  - Significant Use of Resources
- All Relevant DoD, HPCMO, NAVO and other Government Agency Regulations and Requirements Implemented

SECURITY

- Vendor and CERT Recommended Security Patches/Configuration Changes Applied
- System Fully Kerberized
- Local Security Team Worked with NRL to Implement Large-File Aware Kerberized FTP Special Version
- ktelnet, krcp, kftp, krlogin Available
- Secure Shell Available (ssh)

SECURITY

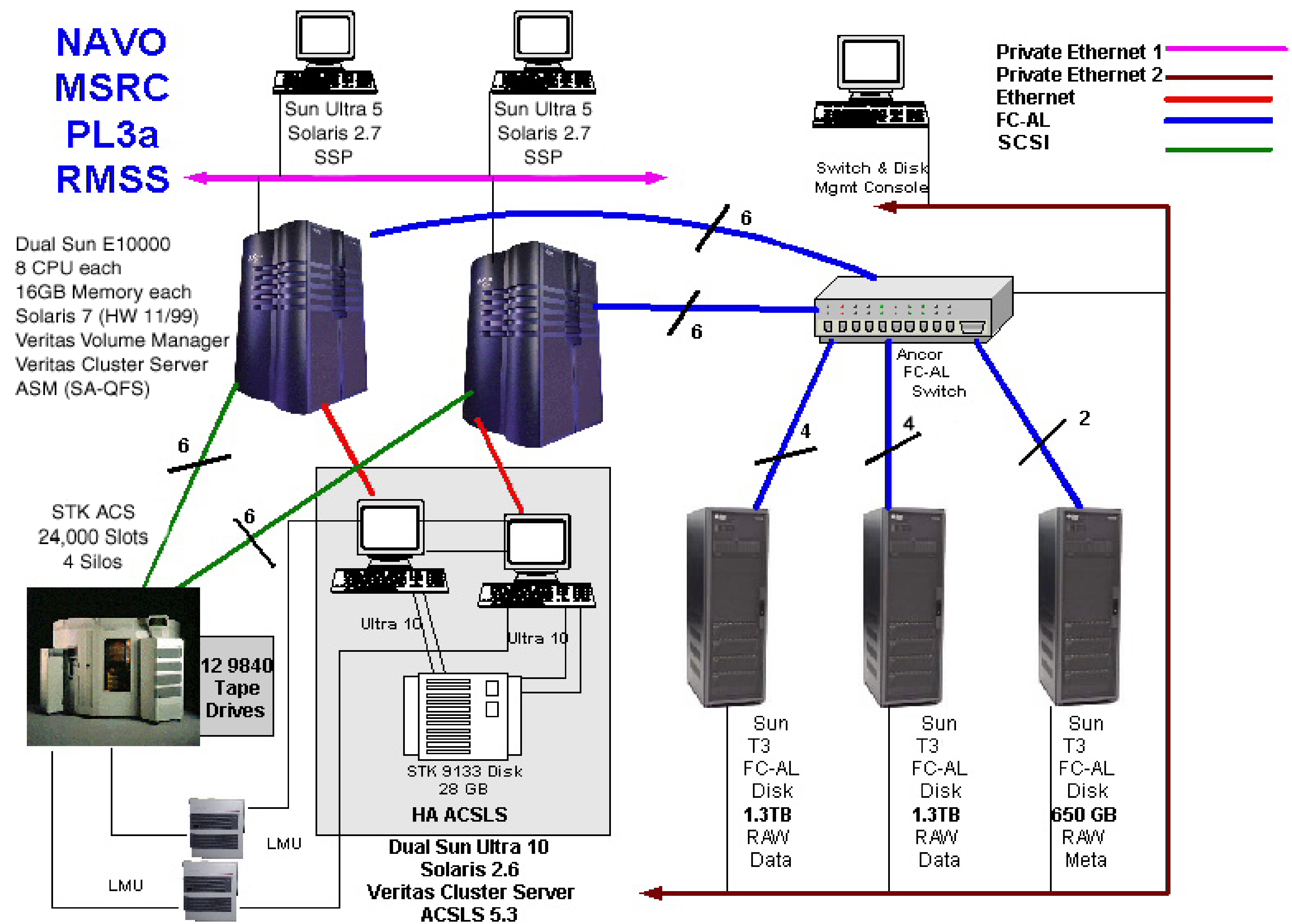
- Archive Data Protected
- No Direct User Access to Tape Devices Permitted

OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

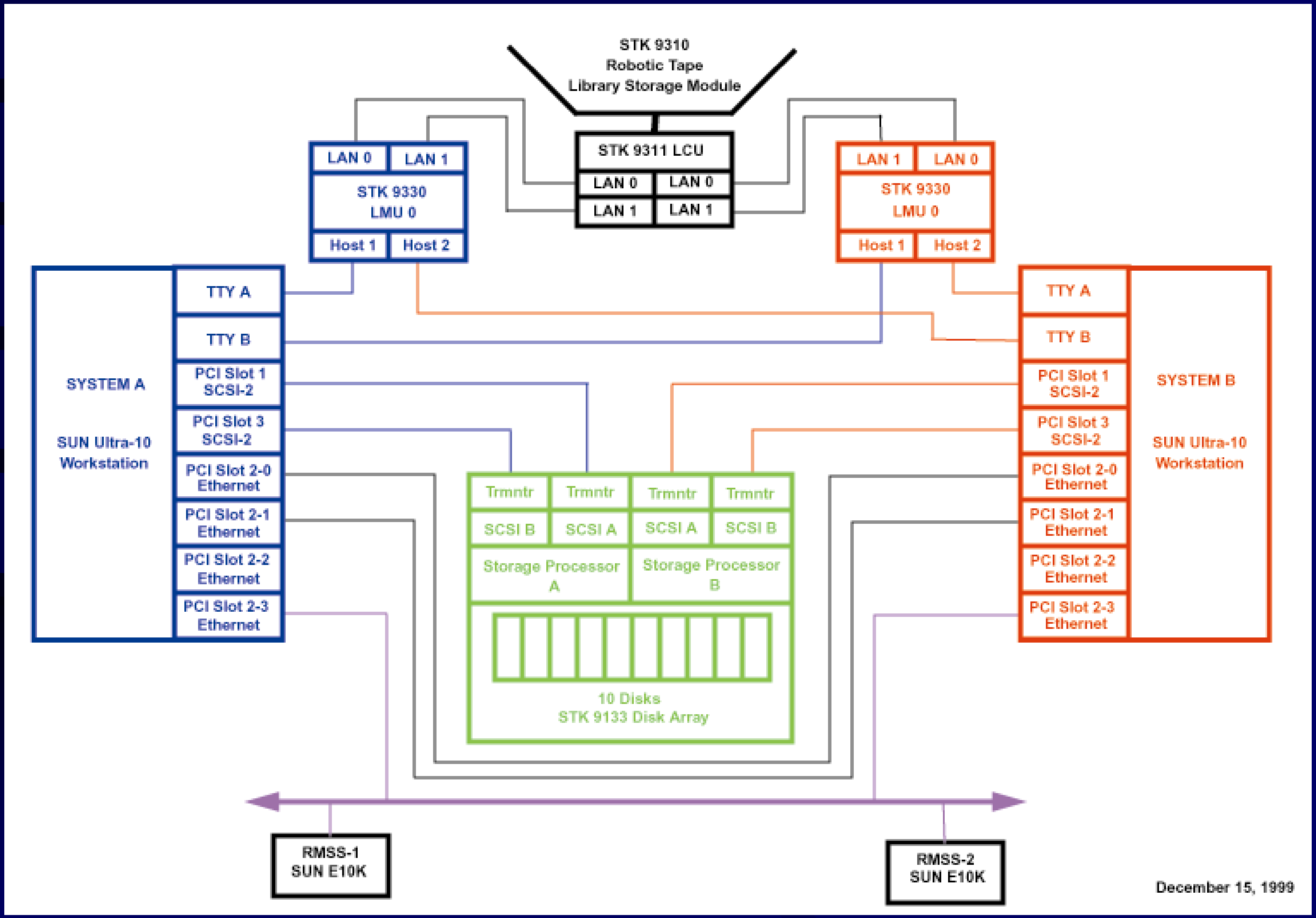
STK HA-ACSLS

- **STK High Availability Automated Cartridge System Library Software**
  - **SAM-QFS and HA-ACSLS Interoperate in Client/Server Relationship**
  - **ACSLS Application Runs on SUN Workstation (Ultra-10) And Controls STK Library Robotics (Server)**
  - **Client resides on Server (E10K) and Requests Tape Mounts as Directed by SAM**
  - **HA-ACSLS is Hardware and Software Package from STK. Provides HA for Library Server**


NAVO
MSRC
PL3a
RMSS
Sun Ultra 5
Solaris 2.7
SSP
Sun Ultra 5
Solaris 2.7
SSP
Switch & Disk
Mgmt Console
Private Ethernet 1
Private Ethernet 2
Ethernet
FC-AL
SCSI
Dual Sun E10000
8 CPU each
16GB Memory each
Solaris 7 (HW 11/99)
Veritas Volume Manager
Veritas Cluster Server
ASM (SA-QFS)
6
6
Ancor
FC-AL
Switch
4
4
2
6
6
STK ACS
24,000 Slots
4 Silos
12 9840
Tape
Drives
Ultra 10
Ultra 10
STK 9133 Disk
28 GB
HA ACSLS
Dual Sun Ultra 10
Solaris 2.6
Veritas Cluster Server
ACSLS 5.3
LMU
LMU
Sun
T3
FC-AL
Disk
1.3TB
RAW
Data
Sun
T3
FC-AL
Disk
1.3TB
RAW
Data
Sun
T3
FC-AL
Disk
650 GB
RAW
Meta

NAVOCEANO MSRC Resilient Mass Storage Server (RMSS)

High-Availability Automated Cartridge System Library Server Configuration



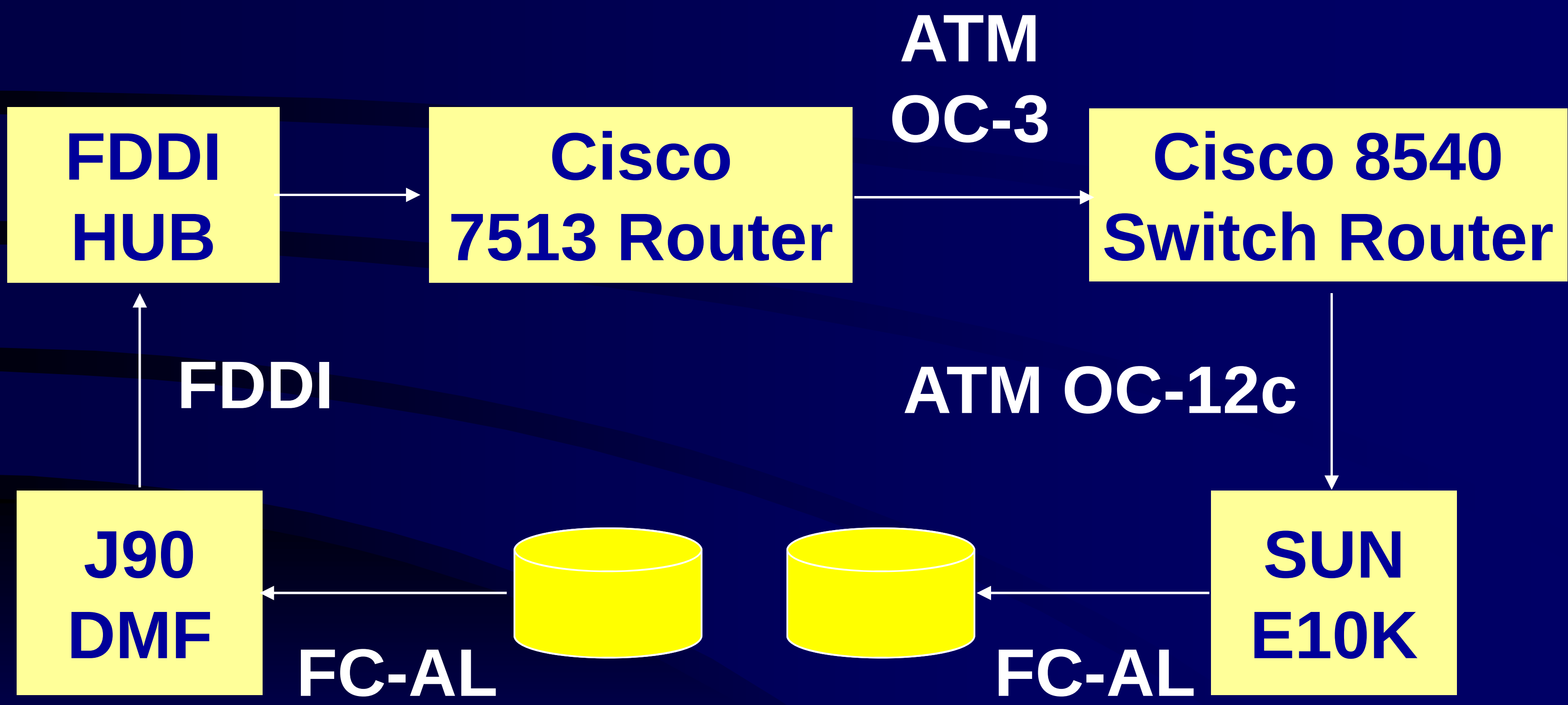
OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

PERFORMANCE
Simple File Transfer Test

- FTP Transfer of Compressible and Uncompressible Files From DMF Server to RMSS
- 1 MB Files Transferred Memory to Memory
- 1 and 5 GB Files Transferred Disk to Disk
- Introduced Disk and I/O Latencies

PERFORMANCE
Simple File Transfer Test



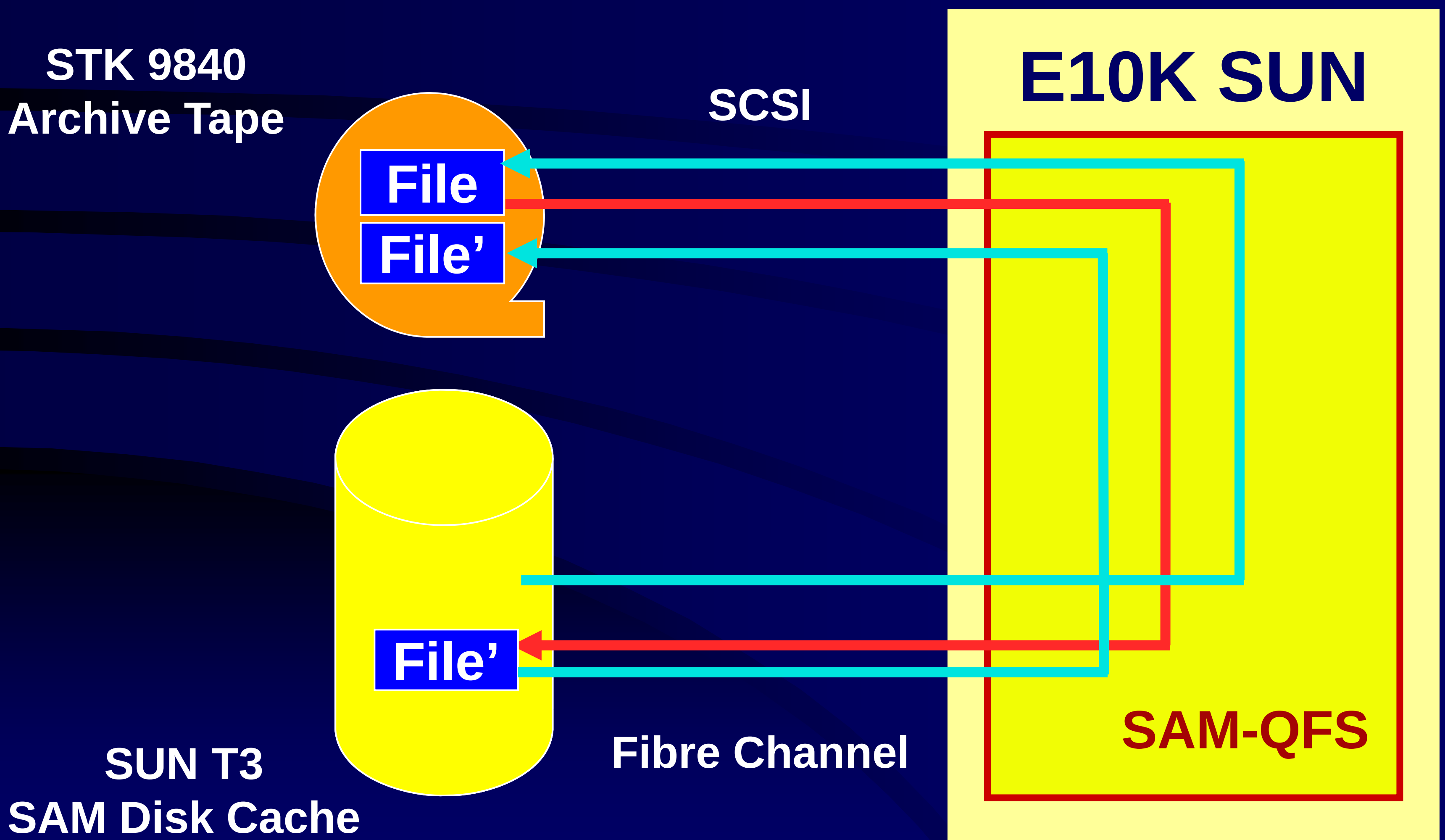
PERFORMANCE
Simple File Transfer Test

File Name	Transfer (Sec)	Attribute	Bandwidth (MB/Sec)	Size (Bytes)
1mb_ascii	0.22	Compr	4.55	1,048,576
1mb_binary	0.22	Uncompr	4.55	1,048,576
1gb_ascii	497	Compr	2.06	1,073,741,824
1gb_binary	261	Uncompr	3.92	1,073,741,824
5gb_binary	3901	Uncompr	1.31	5,368,709,120

PERFORMANCE
RMSS End to End Functional Test

- Test Measured Time Required to Write File from ASM Disk Cache to Tape (archive) and Then Recall File Disk Cache (stage).
- Tape Mounted at Start of Test, But Not Positioned
- Used Uncompressible 5 GB File
  - Gives Worst Case Attained Bandwidths

PERFORMANCE
Single File Functional Test



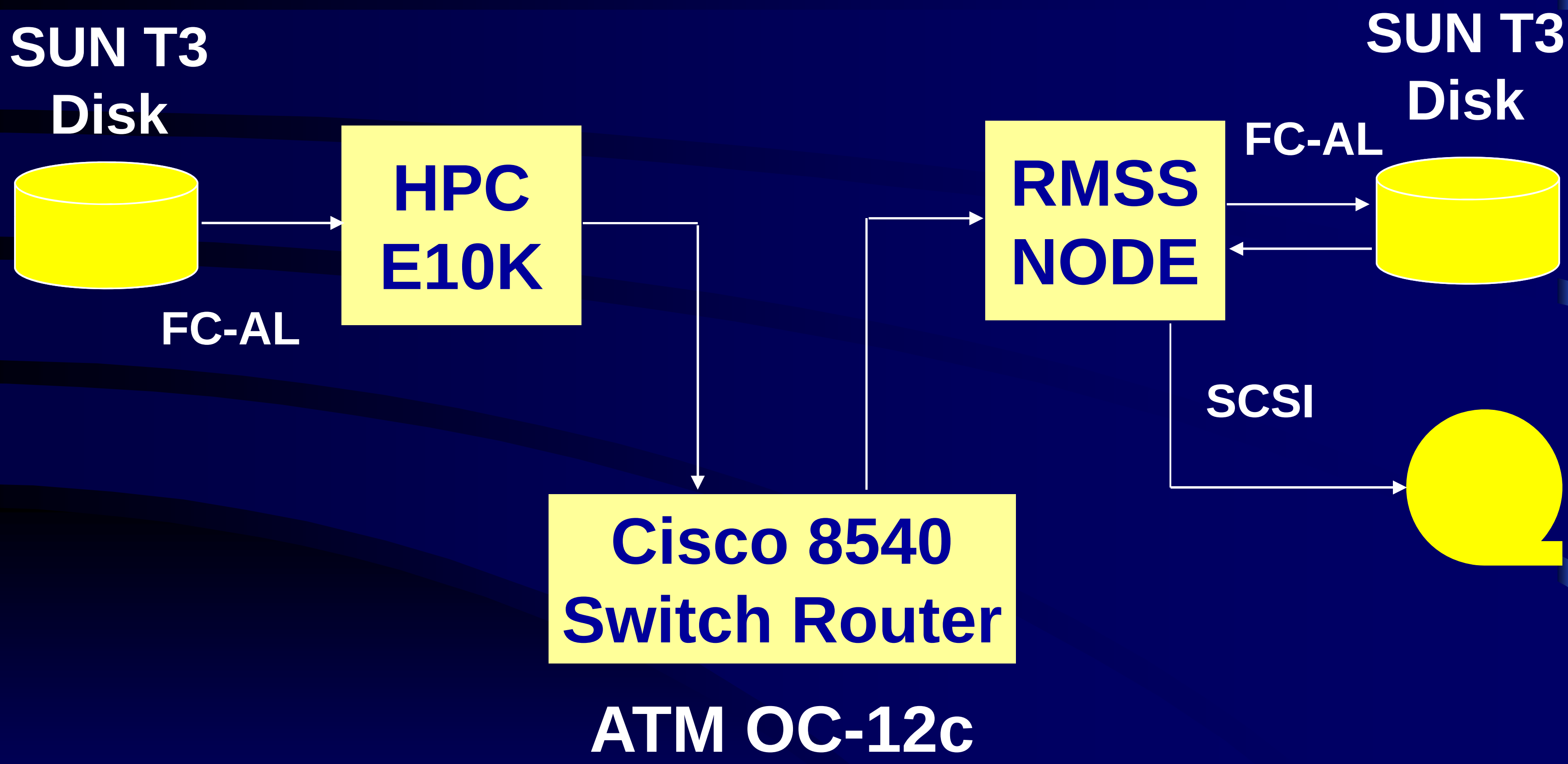
PERFORMANCE
Single File Functional Test

Operation	Data Transfer (Sec)	Tape Position (Sec)	Total Time (Sec)	Effective B/W (MB/Sec)
Archive	3334	8	3342	1.53
Stage (Recall)	602	10	612	8.51
Archive	3436	0	3436	1.49

PERFORMANCE
Profiled Test Data Set

File Size Range	/u/a	/u/b	/u/c	/u/d	Total
64 KB	20	28	84	44	176
512 KB	16	16	24	28	84
1 MB	8	4	4	16	32
10 MB	4	24	36	8	72
50 MB	4	4	12	0	20
100 MB	0	4	0	4	0
500 MB	0	4	0	0	4
1 GB	0	4	0	0	4
Total Files	52	88	160	100	400

PROFILED DATA SET TESTS
Test Configuration



PERFORMANCE
Profiled Data Set Tests

- Measures Typical Production Load
- Provides Analytical Data for Projecting Actual Performance Against Required Performance
- Two Sections to Tests
    - Network Transfer of File Stream
    - File Stream Archiving by SAM-QFS

PERFORMANCE
Profiled Data Set Tests

- **Used Production-Configured Filesystems**
- **Four Separate Streams to Both Nodes**
  - **One Stream Per Filesystem**
- **One STK 9840 Tape Drive Per Filesystem**
- **Data Transfers Across ATM OC-12c Using Non-Kerberized/Non Encrypted RCP**

PERFORMANCE
Profiled Test Data Set

File Stream Network Transfer Results

Stream	Elapsed Time (Sec)	MB/Sec	Avg Sec/File	Total MB Transferred
A	93	1.22	1.79	113.26
B	407	11.95	4.63	4862.19
C	378	1.89	2.36	713.40
D	291	1.34	2.50	389.29

PERFORMANCE
Profiled Test Data Set

File Stream Archiving Results

Stream	Elapsed Time (Sec)	MB/Sec	Avg Sec/File	Files Archived
A	510	0.22	9.8	52
B	493	9.86	5.6	88
C	640	1.11	4.0	160
D	290	1.34	2.5	116

PERFORMANCE
Profiled Test Data Set

End-to-End Stream Throughput Results

Stream	Elapsed Time (Sec)	MB/Sec	Avg Sec/File
A	709	0.16	13.6
B	611	7.96	6.9
C	776	0.92	4.9
D	396	0.98	3.4

PERFORMANCE
Profiled Test Data Set

Aggregate Performance

Stream	Elapsed Time (Sec)	MB/Sec	Avg Sec/File
Transfer Data	407	14.93	0.98
Archive Data	670	9.07	1.61
End-to-End	776	7.83	1.86

End-to-End Aggregate Performance =
7.83 MB/Sec = 28.19 GB/Hour = 676.56 GB/Day

PROFILED DATA SET TEST
Projection to Production System

- Production Cluster Has 15 Tape Drives
  - 7 on Node 0
  - 8 on Node 1
  - 150 MB/Sec Peak, 25% = 37.5 MB/Sec
- 72% Overlap Between Network Transfer Streams and File Archive Streams Attained During Test

PROFILED DATA SET TEST
Projection to Production System

- **Production ATM Bandwidth Measured at 50 MB/Sec Sustained**
- **Network Increase = 3.33**
- **Tape Drive Increase = 3.75**

PROFILED DATA SET TEST
Projection to Production System

- **Projected Network Transfer Time:**
  - 407 Sec X 0.3 = 122 Sec
- **Projected Archive Time:**
  - 670 Sec X 0.267 = 178 Sec
- **Projected Stream Time (72% Overlap)**
  - (122+178) X 0.72 = 216 Sec
- **Projected Speedup Factor: 776 / 216 = 3.6**

PROFILED DATA SET TEST
Projection to Production System

- **Projected Transfer Rate for Production System: 7.83 MB/Sec End-to-End X 3.6**
  - **28.2 MB/Sec**
  - **101.5 GB/Hour**
  - **2.43 TB/Day**
- **Anticipate Additional Increases with System Tuning**

OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- **Transition To Production**
- Future Work

USER DATA TRANSITION

- **Problem:** Transition all Data Managed by DMF/Unicos Server to SAM-QFS/Solaris Resilient Mass Storage Server (RMSS)
- **Problem Size:** 225+ TB

USER DATA TRANSITION

- **Constraints:**
  - Continued User Data Availability During Transition Process Required
  - Heterogeneous Archival Data Formats
  - Heterogeneous Platform and O/S Environments
  - DMF Archival Data Format Propriety

USER DATA TRANSITION

- **On-Demand File Transition!**
  - User-Centered
  - Allows Instantaneous Transition of Users with No Periods of Data Unavailability
  - User Accesses Files on RMSS Cluster
  - Files Retrieved from Old Mass Storage Server if Not on RMSS

USER DATA TRANSITION

- **On-Demand File Transition!**
  - **Underlying Tools Retrieve File from DMF/UNICOS Platform Transparently**
  - **Migrates Files Users Require Immediately**
  - **Uses SAM/QFS Migration Toolkit**

USER DATA TRANSITION

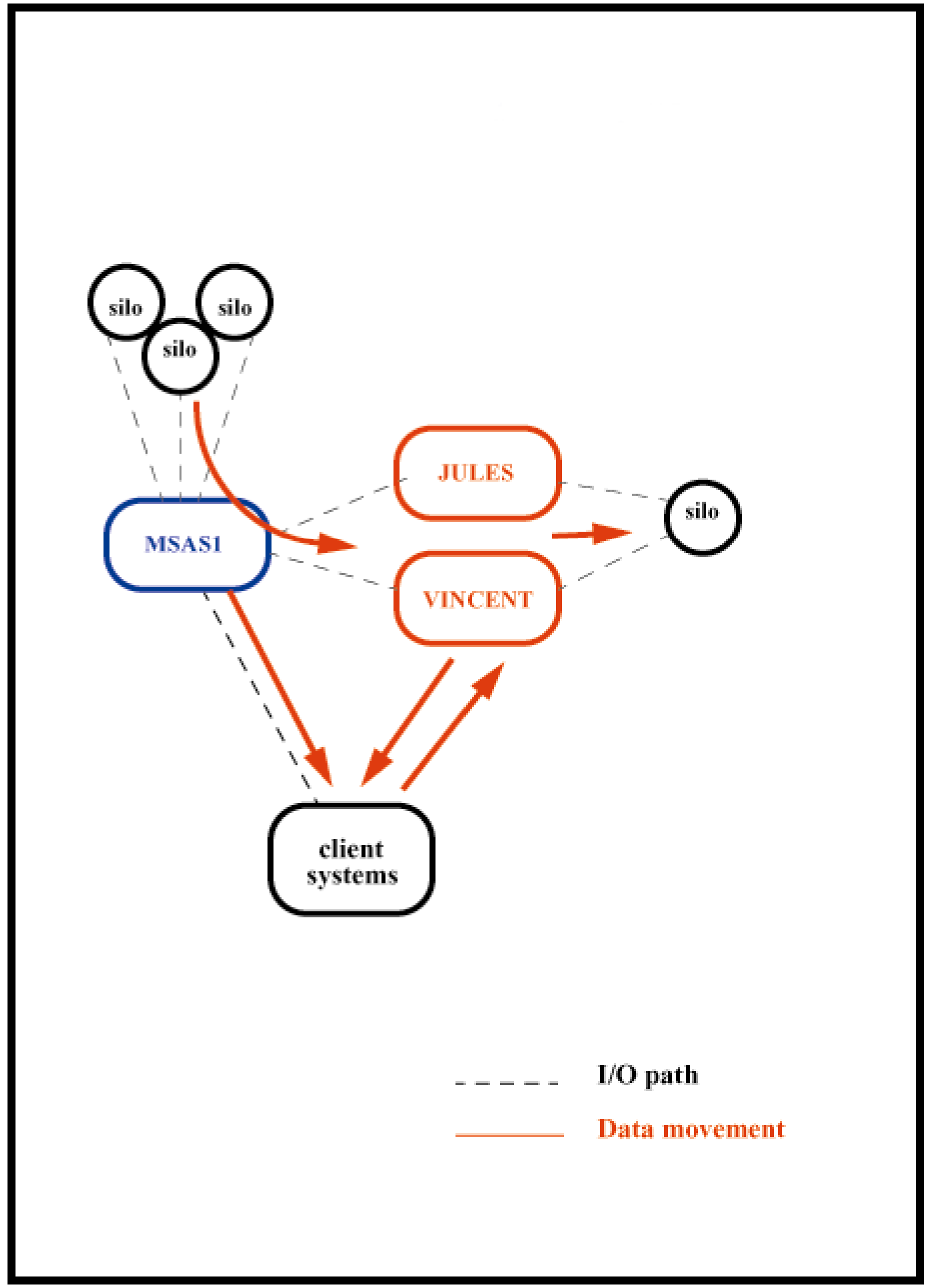
- Process
  - Application Accesses File
  - File Marked as Foreign: Not on RMSS
  - SAM/rsh Retrieves File From MSAS1 and Writes It To RMSS
  - File Delivered to User Application
  - File Flagged for Rearchive as Native SAM File on Local SAM Controlled Media
  - Next Access of File is Entirely on RMSS

Sample SLS File Listing Display

xxx/inwork_xmp:

mode:	-rwxr-xr-x	links:	1	owner: jkothe	group: NA0101
length:	142048	inode:	1680		
offline; archdone;					
copy 1: ----	Apr 24 16:38	28.0		za	jkothe
access:	Oct 12 1994	modification:	Oct 7 1994		
changed:	Nov 29 1998	attributes:	none		
creation:	none	residence:	none		

Migration Model



USER DATA TRANSITION

- **Background File Transition**
  - Bulk File Transfer
  - Moves Files Not Requested On-Demand
  - Filenames are sorted by DMF Tape VSN
    - Improves Throughput
  - Sustained Bulk Transfer Performance:
    - 1 TB/Day

USER DATA TRANSITION

- Before Transition
  - Establish user accounts on RMSS
  - Permit preliminary RMSS user logins
  - Review/Scan Filesystems

USER DATA TRANSITION

- **On the Transition Date**
  - RMSS Service outage for filesystem users
  - Restrict old filesystem to *read-only* access
  - Replicate old filesystem directories on RMSS
  - Mark all files as foreign Files (“za”)
  - Restore User Access to RMSS
    - +Read/Write Access to RMSS
    - +Read-Only Access MSAS1

USER DATA TRANSITION

- **After The Transition Date**
  - User archive via RMSS – supported by Migration Toolkit, kftp, krcp
  - Complete Facility activity to move *small* files
  - Process VSN/file list via Migration Toolkit rearchive (Background Bulk File Transition)

USER DATA TRANSITION

- Progress
  - Began July 2000, Internal Users,
    - 4.5 TB Moved in Two Months
  - Mid October 2000, First External User Filesystem
    - 20 TB Moved in Two Months
    - Transfer Rate Reached 660 GB/Day

USER DATA TRANSITION

- **Progress**
  - Mid-January 2001, Second User File System
    - Completed End of March
    - 60 TB Transitioned
    - Routines Tuned During Previous Filesystem Transfer
    - Transfer Rate Reached 1 TB/Day

USER DATA TRANSITION

- Progress
  - Final User Filesystem
    - Transition Began April 5, 2001
    - 160 TB To Transition
    - Anticipated Completion:
      - September, 2001
- Support Overhead
  - Administrative/Maintenance: 1.5 Analysts

OVERVIEW

- Cluster Types/High Availability Clusters
- Requirements Definitions/Workload Analysis
- System Design and Implementation
  - Node Description
  - Network Connectivity/Security
  - HA/ACSLS Subsystem
- Performance
- Transition To Production
- Future Work

FUTURE WORK

- **Refresh Disk Technology**
  - Replace Expansion Units with Controller Units
  - New Technology RAID Drives
- **Tape Drives on the Disk Cache SAN**
  - Increase Disk/Tape Transfer Performance
  - Maintains Availability of All Tape Drives in Event of Node Failure

FUTURE WORK

- Refresh Software Technology
  - Reengineered SAM-QFS Available (V3.5)
  - Solaris 8 Upgrade
  - Refresh VCS Cluster Manager S/W
- Refresh Hardware Technology
  - Updated Fibre-Channel HBAs
  - Refresh SAN Switch Technology
  - Native Fibre Channel Tape Drives
  - Large-File Optimized (Capacity) Tape Drives

FUTURE WORK

- **Refresh Disk Technology**
- **Improve Resiliency**
  - **IP Address Failover**
  - **Improved Disk Multipathing**

CONCLUSION

- RMSS Cluster Supporting 50% of NAVO MSRC Workload, 100% in Six Months
- Design Proven, Relatively Few Problems
- Scales to Meet Projected Requirements
- Remaining Two MSRCs Implementing
- Discussions to Develop Smaller Scale Version for Other Entities Now Expressing Interest

THANK YOU

Terry Jones

01.228.688.5344

jonestl@navo.hpc.mil

Beata Sarnowska

01.228.688.6334

sarnowsk@navo.hpc.mil