

440-TP-009-001

# Network Attached Storage Concepts & Industry Survey for the ECS Project

Technical Paper

Technical Paper—Not intended for  
formal review or government approval.

February 1995

Prepared Under Contract NAS5-60000

## RESPONSIBLE ENGINEER

Alla Lake /s/ 2/23/95  
Alla Lake, DADS Engineering Specialist      Date  
EOSDIS Core System Project

## SUBMITTED BY

Stephen Fox /s/ 3/18/95  
Steven Fox, SDPS Segment Manager      Date  
EOSDIS Core System Project

Hughes Applied Information Systems  
Landover, Maryland

This page intentionally left blank.

# Summary

---

This paper is to be updated periodically as the Earth Observing System Data and Information System Core System (ECS) Data Archival and Distribution System (DADS) performance requirements become more concrete and newer developments arise in the networking and network attached storage market. The initial purpose of this study is to identify candidate methods and technologies of network attached storage for a distributed archive and to examine the associated implementations issues and risk factors.

This page intentionally left blank.

# Contents

---

## Summary

### 1. Introduction

### 2. Review of the I/O Interface Suitability

2.1	Fiber Distributed Data Interface (FDDI), FDDI-II & FDDI Follow-On LAN (FFOL) Standards .....	2-1
2.1.1	FDDI, FDDI-II & FFOL DADS Application Potential.....	2-2
2.2	High Performance Parallel Interface (HIPPI) Standard.....	2-2
2.2.1	High Performance Parallel Interface DADS Application Potential .....	2-3
2.3	Fibre Channel .....	2-4
2.3.1	Fibre Channel DADS Application Potential.....	2-4
2.4	SCSI.....	2-5
2.4.1	SCSI DADS Application Potential.....	2-5
2.5	Ethernet .....	2-6
2.5.1	Ethernet Application Potential .....	2-6

### 3. Architectural Considerations

3.1	Network Attached Storage Architecture .....	3-1
3.2	Existing Distributed Archives .....	3-2
3.3	Existing ATL Interfaces .....	3-2
3.4	FSMS Suitable for a Network Attached Architecture.....	3-3
3.5	Risks Summary.....	3-3

## 4. Selected Architectural Approaches

4.1	Direct Network Attachment.....	4-1
4.1.1	Configuration.....	4-1
4.1.2	Advantages.....	4-2
4.1.3	Disadvantages.....	4-2
4.1.4	Feasibility .....	4-2
4.2	Concentrator Based Configuration.....	4-3
4.2.1	Configuration.....	4-3
4.2.2	Advantages.....	4-3
4.2.3	Disadvantages.....	4-4
4.2.4	Feasibility .....	4-4
4.3	I/O Computers.....	4-4
4.3.1	General Purpose Computer .....	4-5
4.3.2	Custom Solution Using COTS Components .....	4-5
4.3.3	Dedicated Function I/O Computer.....	4-6

## 5. Conclusion

## 6. References

### Figures

4-1.	Direct Network Attachment Architecture .....	4-2
4-2.	Concentrator Based Architecture .....	4-3
4-3.	I/O Computers.....	4-4

### Tables

3.1	Automatic Tape Libraries Interface Devices.....	3-3
-----	---	-----

## Abbreviations and Acronyms

# 1. Introduction

---

While the upper limits of the required Earth Observing System Data and Information System Core System (ECS) Data Archival and Distribution System (DADS) data throughput performance have not been established at this time, it is clear, that very high cumulative data rates, in excess of a hundred Mega bits per second (Mbps), will have to be supported. Support of the high data rates necessitates a very careful choice of the architecture in order to facilitate high throughput, rather than to hamper it.

Two alternative architectures are best known in large data archival systems. The first is a front-end approach, where all data flow, as well as most control functions, are handled by a central front end machine. The sheer size of the on-line ECS archive and the required data throughput rates demand a very large computer with a high data input and output (I/O) processing capacity. This architecture is simple in both implementation and maintenance. However, it is not readily expandable or cost effective. The file storage management function for an archive of the ECS magnitude is a Central Processing Unit (CPU) intensive operation. The file storage management would compete for CPU time with the I/O function, if the same computer is shared. The front end CPU may easily become a bottleneck for a large archive.

This paper examines the second approach, a distributed archive implementation with separate control and data paths. Staging-to-archive device I/O traffic is off loaded from the central host CPUs to dedicated I/O controllers in order to maximize the system performance. An optimal network attached configuration must be scalable and should accommodate the heterogeneous tape and disk storage devices that are likely to be used by ECS DADS.

The discriminating factors for the I/O interfaces are as follows. First is the suitability of the interface for staging-to-archive use. The other factors are: the bandwidth, transfer rate, command set possibilities offered, the number of host and storage devices making use of the interface, and the existing direct transfer and third party transfer software support.

Section 2, Brief Overview of the Current I/O Interface Standards, reviews the I/O interface standards in place currently or being developed in the near future. Section 3, Architectural Considerations, examines the availability of the interface technologies discussed in Section 2, their suitability, and risk factors associated with existing or future implementations.

This page intentionally left blank.

## 2. Review of the I/O Interface Suitability

---

Let us first state the definitions of terms *network* and *channel* as commonly used in the industry. A *network* is defined as a general purpose data communication path interconnecting computers (or devices). Networks are, as a rule, characterized by serial transmission, longer device to device distances, and more complex message oriented protocols.[1,2] Ethernet and Fiber Distributed Data Interface (FDDI) are examples of networks.

A *channel* is a special purpose communications path which connects a computer (I/O controller) to a peripheral, such as a storage device. It is usually stretched over limited distances and runs over a number of parallel signal connections. Data integrity is commonly assured by a parity check. Small Computer System Interface (SCSI) and High Performance Parallel Interface (HIPPI) are two examples of communication channels.[1,2]

Fibre Channel, while classified as a channel by its very name, could alternatively be called a network. Like HIPPI, it supports high-speed dedicated communications path (although, like a network, it is bit-serial as opposed to the 32- or 64-bit HIPPI bus).

In the ECS DADS architecture, like in other large scale distributed archives, such as those mentioned in Section 3, a channel interconnection is most suitable for use as a data path and a network interconnection can be better used as a control path. Given that various types of interface are likely to be used, Fibre Channel fabric capable of accommodating most of them can be of great advantage. (A communication fabric is an interconnection of multiple communication segments of the same type.) The remainder of Section 2 examines the suitability of the existing interface standards to DADS applications in greater detail.

### 2.1 Fiber Distributed Data Interface (FDDI), FDDI-II & FDDI Follow-On LAN (FFOL) Standards

The Fiber Distributed Data Interface standard set was developed in American National Standards Institute (ANSI) Committee X3T9.5. FDDI and FDDI-II offer services suitable for lower bandwidth and long distance communications networks. The concept of a fiber-optic-based communication ring was introduced for standardization in 1982. FDDI topology is a dual ring of trees. The protocol for FDDI was initially fully distributed with no master station.

Originally intended for fiber media only, FDDI standards today cover non-fiber media, such as copper wires, as well as a centralized ring master station for FDDI-II. FDDI is a packet switching network. FDDI-II adds a circuit switched isochronous service compatible with the circuit switched carriers of public networks.[6,8]

The following set of definitions, all taken from [8], must accompany the previous paragraph: 1. "*Circuit* is a bi-directional communications capability provided over a continuous isochronous channel(s) between two or more Circuit Switching-Multiplexer

(CS-MUX) level entities." 2. "*Circuit switching* is a service that provides and manages a set of circuits." 3. The term "*isochronous*" indicates the essential characteristic of a time-scale or a signal such that the time intervals between consecutive significant instants either have the same duration or durations that are integral multiples of the shortest duration."

Both FDDI and FDDI-II support 100 Mbps token ring network in large configurations: over 100 stations (1000 physical connections), up to 100 km of cable (200 km total fiber optic path length). The FDDI token ring protocol provides for: 1) synchronous service with preallocated bandwidth and strictly bounded access delays; 2) asynchronous service with no bandwidth preallocation, but weak access guarantees (unused synchronous bandwidth is available for asynchronous use); 3) restricted token service for dedicated real time control applications. Time delays taking a toll on performance are associated with the token ring protocols. FDDI technology is stable, and the products are available from a wide pool of vendors.

The FDDI Follow-On LAN (FFOL) standard is being defined to serve as a backbone for multiple FDDI and FDDI-II networks. FFOL can carry signals at the rates in excess of a Gigabit (Gb) per second over distances ranging from 100 to 200 meters.[6,12]

### **2.1.1 FDDI, FDDI-II & FFOL DADS Application Potential**

Due to the token ring architecture of FDDI, its definition as a network rather than a channel, and the lack of Automatic Tape Library (ATL) vendor implementations, the FDDI, in any of its incarnations, is more appropriate for use as a control network, rather than for data channels to the ATL. One vendor is said to be implementing a maximal-rate, 17 Mega Bytes per second (MB/sec) Enterprise Systems Connection (ESCON) channel (IBM variety of FDDI) interface for a helical scan tape drive. Aside from that one vendor, there do not seem to be any other ATL vendors actively implementing or considering a FDDI tape drive interface.

FDDI is applicable primarily as an Ethernet alternative or an Ethernet enhancement in the implementation of DADS control and request paths, as well as for processor to processor communications in a distributed archive. The second application falls more under the purview of Communications and System Management Segment (CSMS) than DADS.

3Com is one of the more notable FDDI vendors on the market. DEC, with its cluster architectures, has long been associated with FDDI technology. Practically no FDDI vendor research has been done for this paper since this is one of the lower risk technology areas.

## **2.2 High Performance Parallel Interface (HIPPI) Standard**

The High Performance Parallel Interface standard encompasses a number of documents, such as "High Performance Parallel Interface - Mechanical, Electrical, and Signaling Protocol Specification (HIPPI-PH)" [X3.183-1991], HIPPI-FP (Framing protocol), HIPPI-LE (Link Encapsulation), etc. Some of the documents, such as HIPPI-PH, have been completed, others are in a draft form in ANSI committee Task Group X3T9.3 The intent of the standard is to provide for high-performance point-to-point channel connection

between high-end systems.[2, 9] HIPPI provides an interface layer to support Upper Layer Protocols (ULPs), such as mapping Intelligent Peripheral Interface (IPI) commands to HIPPI, HIPPI-IPI.

The physical level of the IPI standard (IPI-0/1) defines a 16, 32, or 64-bit parallel, master/slave electrical bus, with transfer rates up to 200, 400 or 800 Mbps. The maximum bus length is 150 meters. Device Specific level (IPI-2) defines the low level controller to device command sets for peripherals dependent on the features of the IPI-1 bus. The IPI-3 standard is defining a device generic command set.

HIPPI is a simplex point-to-point connection (no multi-drop). A full-duplex implementation requires two HIPPI channels. In the 800 Mega bits per second (Mbps) version of HIPPI, the word is transferred over 32 data lines. The 1600 Mbps, double wide version, is 64 data lines wide. The maximum twisted copper cable length of a HIPPI segment is specified at 25 meters. Serial HIPPI (an implementer's agreement) specifies the use of fiber optic cabling that extends the maximum allowed cable length up to 10 km distances.

### **2.2.1 High Performance Parallel Interface DADS Application Potential**

The high throughput rate of HIPPI best lends it to serving as an ATL data channel interface. HIPPI channels can be connected via crossbar switches. Each switch allows for multiple simultaneous connections, so that concurrent transfers can occur, each at its peak data rate.[10] The switches are available from a number of vendors and do not present an integration problem. HIPPI is a comparatively older standard. Therefore, HIPPI-based attachments are preceding newer technology, such as Fibre Channel attachments to the market.

The HIPPI attachment to the DADS devices will require considerable integration work. The development of HIPPI-IPI, the specification defining the IPI-3 Disk and Tape commands over HIPPI may result in product development that would facilitate such integration. Although the ATL vendors are moving towards a HIPPI-based interface implementation with the helical scan-based tape drives, few of the tape drive manufacturers are ready to offer it in product form. A HIPPI implementation is under consideration for a helical scan drive of at least one manufacturer. Although, in that case HIPPI adoption is not favoured, because it is perceived as a "stop-gap" measure in light of Fibre Channel developments.

Ampex DD2 drive provides an IPI-3 interface, not a HIPPI interface per say. The National Storage Lab's (NSL) attempt to use MaxStrategy HIPPI controller for a direct Redundant Array of Inexpensive Disks (RAID) to DD2 tape drive connection was only partially successful. Instead, an RS6000 serves as a controller in the current configuration at the NSL.

All in all, the emerging Fibre Channel standard seems to be superseding the current HIPPI. A number of vendors may decide to forego HIPPI development and "sit on the fence" waiting for the Fibre Channel technology to mature. Vendor offerings in the area of HIPPI attached disk arrays must be investigated more thoroughly, but it is already clear, that the

market is sparse. IBM RS6000 and Cray are two of the better known vendors with the appropriate products.

The following is a brief list of vendors offering HIPPI products: Network Systems, NetStar, Input/Output Systems, Maximum Strategy (HIPPI attached RAID controller tested in the NSL configuration). Further vendor and product survey needs to be done.

## **2.3 Fibre Channel**

The Fibre Channel standard is under development in the ANSI X3T9.3 working group (X3.230-199x). The standard specifies bit-serial communications over optical fiber. The spelling "Fibre" is deliberate. The standard documents are international, which is why the internationally accepted spelling of the name is used.

Fibre Channel provides a transport vehicle for IPI-3 (applied over HIPPI) and packetized SCSI command sets, as well as HIPPI data framing, and IBM S/370 Block Multiplexer commands. The command sets and protocols are encapsulated into a new framing-type protocol.[13] It defines an interface which is used with a switching matrix for the interconnection of heterogeneous ports (multiple channel and networking protocols). Direct peripheral attachment is going to be one of the number of services to be provided.[11,15] Dedicated connections, multiplexed connections, and datagrams are all going to be supported by the Fibre Channel. Of those three levels of network service to be supported, the dedicated connection resembles most closely the service provided by HIPPI.

The data rates supported by Fibre Channel are: 100 MB/sec, 50 MB/sec, 25 MB/sec, 12.5 MB/sec. In summary, Fibre Channel will probably supersede HIPPI on the market. It will accommodate both HIPPI and other existing interfaces in a switching fabric. Fiber Channel is more complex to implement, as compared to HIPPI, due to its increased capability.

### **2.3.1 Fibre Channel DADS Application Potential**

Fibre Channel based switching fabric products are becoming more. However, because it is a relatively new standard, the number of available host and peripheral interfaces at this date is very few. On the other hand, the very attractiveness of the new standard is causing the vendors not to proceed with the HIPPI development they would have otherwise undertaken and wait for the Fibre Channel products to become available. Both the scarcity of Fibre Channel products and vendor hesitation in making interface design choices present a problem to DADS.

The use of Fibre Channel technology within a DADS network attached storage architecture would initially be limited to the switching fabric supporting a variety of other interconnection technologies.

Preliminary vendor list: Canstar (Canada), Ancor Communications, Inc. (High Performance Storage System [HPSS] participant)

## **2.4 SCSI**

The original Small Computer System Interface (SCSI) is specified in ANSI standard X3.131-1986. The standard defines a peripheral interface that distributes data among peripherals independent of the host. The SCSI standard interface is capable of supporting 8 physically addressable devices.

SCSI provides two handshaking protocols: asynchronous and synchronous. Asynchronous mode requires a handshake for every byte transferred, and in synchronous mode, a series of bytes are transferred before the handshake occurs increasing the data transfer rate.

There are multiple SCSI standards. The older SCSI-1 has an 8 bit data bus with an added parity bit. The SCSI-1 standard supports 2 MB/sec in asynchronous mode and 5 MB/sec in synchronous mode. This standard has been around since 1986, and most ATL vendors have at least a SCSI-1 data path interface. Migration to the SCSI-2 standard is in progress in the marketplace.

The SCSI-2 standard supports 8 bit data bus with an optional extension to a 16 or 32 bit data path at a faster rate yielding an effective maximum transfer rate of 20 MB/sec in asynchronous mode and 40 MB/s in synchronous mode. SCSI-2 also provides for queuing of commands by their logical unit number (LUN), removing this responsibility from the host. Single-ended SCSI-2 extends to a length of 6 meters, differential SCSI-2 can be extended up to 25 meters.

Three gradations of SCSI-2 implementation exist. SCSI-2 command set implementation without either the speed or the width increase results in the same data rates as SCSI-1. SCSI-2 "fast" implementation does not increase the data bus width, but in addition to SCSI-2 command set support is capable of data rates of up to 10 MB/sec. Finally, the complete "fast" and "wide" - 32 bit implementation provides the SCSI-2 level data rates as well as command set. Since the width extension is optional, the SCSI-2 vendor specification does not necessarily imply the 20 MB/sec and 40 MB/sec data rates that can be achieved with a "fast and wide" SCSI-2.

The SCSI-3 standard is under development. It represents further evolution of the diverse I/O interface standards toward interconnection. SCSI-3 subdivides the SCSI standard into three layers: Command Set, Protocol, and Interface. It extends peripheral device command sets. The SCSI-3 interface will be capable of supporting up to 16 directly addressable devices, as well as dual port operation, and optional 16-bit transfers on a single cable. A "packetized" stack will be available for a bit serial interface to Fiber Channel, HIPPI or the SCSI-3 Parallel Interface.[16] No SCSI-3 compliant products exist at the moment.

### **2.4.1 SCSI DADS Application Potential**

The SCSI interface will have to be "reckoned with" in DADS integration since it is ubiquitous in the ATL market.

## **2.5 Ethernet**

In a generic sense, Ethernet (Trademark of the Xerox Corporation) refers to all CSMA/CD protocols. Ethernet was originally conceived in 1973 and later the Institute of Electrical and Electronics Engineers (IEEE) 802.3 (Ethernet) LAN standard based on Ethernet became the official standard. The IEEE 802.4 (token bus) and IEEE 802.5 (token ring) standards were established at the same time as IEEE 802.3 and are also derived from Ethernet.

Ethernet is a 1-persistent Carrier Sense Multiple Access (CSMA) Collision Detection (CD) protocol. CSMA/CD works by having all the machines listen to the cable. If the cable is idle, any machine may transmit. Collisions cause a termination of transmission and a retransmission after a random period of time. A 1 persistent transmission means that the probability of transmitting when the channel is empty is 100%.

The original standard uses a 50-ohm coaxial cable. Maximum cable length is 500m, but longer distances can be achieved by using repeaters (devices that receive a signal, perform amplification or regeneration, then retransmit the signal). The cable is tapped by use of a transceiver. Transceiver is defined in IEEE 802.3 as "the attachment hardware connecting the controller interface to the transmission cable. The transceiver contains the carrier sense logic, the transmit/receive logic, and the collision-detect logic" [18]. The cable extending from the transceiver may be 50 meters. Two Ethernet cables can be connected by use of a bridge (also called a selective repeater). A typical maximum speed of Ethernet is 10-Mbps.

### **2.5.1 Ethernet Application Potential**

Ethernet connections between DADS devices should be straight forward because of Ethernet's age and popularity. Most commercial systems have Ethernet accessories available today.

Preliminary vendor list: 3Com

## 3. Architectural Considerations

---

### 3.1 Network Attached Storage Architecture

There are several benefits associated with a distributed network-attached architecture:

- 1) performance optimization due to the off-loading of the data flow from the machine performing the file storage management task;
- 2) decreased data transfer latency due to the direct transfer routes between storage and the requesting clients;
- 3) inherent evolvability - potential for the capacity increase and for the addition of newer technologies as network nodes;
- 4) cost efficiency in foregoing a large expensive single computer configuration;
- 5) the ability to accommodate multiple clients that are running disparate File Storage Management Systems (FSMS).

While a number of vendors offer HIPPI and Fibre Channel fabric solutions as network products, limited integration experience exists with ATL products. As should be expected, most of the work to date has been done by the storage manufacturers whose tape drives are based on the helical scan technology (due to the I/O throughput of the helical scan drives). Among the large computers that would be suitable for a dedicated I/O front end function, Cray is on the forefront of HIPPI connectivity. That again is not surprising, since HIPPI arose from the work done at the Los Alamos National Laboratory (LANL) in trying to interface peripheral devices directly to Cray's HSX data channel.[16] Cray has performed a significant amount of work with HIPPI and has reputedly achieved very high Transmission Control Protocol/Internet Protocol (TCP/IP) performance over HIPPI channels.

One aspect of a network attached configuration, attractive in terms of performance, but risky in terms of implementation, is a channel data path connection between the data staging (disk) and the data archiving (ATL) portions of the archive. To date, the use of such a data path has been tried at the NSL using Maximum Strategy Corporation HIPPI RAID controller. According to some reports it was unsuccessful due to buffering problems on the controller. It does not, however, preclude DADS from pursuing a prototype development in this area in cooperation with the Maximum Strategy Corporation or another vendor.

With Commercial Off The Shelf (COTS) hardware products being used to implement network attached storage, the bulk of the DADS integration work will be in software: device drivers, buffering mechanisms, priority schemes, and lower level volume serving interfaces to COTS FSMS products.

## 3.2 Existing Distributed Archives

One good example of an existing distributed archive architecture is the High Performance Data System (HPDS) at LANL [5]. Another example is the High Performance Storage System (HPSS) at the National Storage Laboratory (NSL) (U.S. Department of Energy Lawrence Livermore National Laboratory (LLNL) in Livermore, California) [14]. HPDS makes use of a LANL-developed HIPPI Data Transfer (HDT) network. TCP socket connections on Ethernet are used for control communications. HIPPI is used as a separate data path.

Another implementation of a network-attached system is the MaSSIVE™ storage system at the National Center for Atmospheric Research (NCAR) [1]. All three of the above systems are termed "fourth generation" to indicate that the data transfers occur directly between storage devices and client machines. Workstation-class machines are used for control and management.[5] ECS needs to take advantage of the body of experience generated by the above projects.

Yet another large scale archive where interesting work with HIPPI attached helical scan tape (DD2) storage is being done currently is at the Fort Meade, Maryland NSA facility. Due to the nature of the agency and the recency of the experience, less publicly available information exists. However, ECS DADS is periodically updated on the progress of this effort.

## 3.3 Existing ATL Interfaces

SCSI interface, whether SCSI I, II, or the newly defined SCSI III is the most readily available ATL interface. HIPPI interface is available with DD1 and DD2 helical scan drives. STK will support an upgraded ESCON channel interface for their DD3 drive.

While nominally these interfaces conform to their respective standards, there is enough design latitude to make their implementation slightly different for each vendor. Network integration with other devices compliant with the appropriate interface standards is potentially a non-trivial effort.

Table 3.1 below lists the interfaces known to be supported by the ATL and tape drive manufacturers in the market.

**Table 3.1 Automatic Tape Libraries Interface Devices**

ATL Equipment	Data interface(S)	Control interface(s)
Bosch Robotics	-	DASD, LU 6.2/APPC Token Ring LAN, VTAM LU2, EXCP/3270 (NON-SNA 3274 control unit), V24 (RS-232)
Data Tape DD1 drive [DCTR-LP]	HIPPI (50 MB/s)	RS-422
DEC DLT drive	SCSI-2 (10 MB/s synchronous)	
EMASS AMPEX IBM (DD2 drive)	enhanced IPI physical interface (IPI-3 logical constructs) Future: HIPPI, Fibre Channel	Drive: RS-232C or RS-422 Tower & Library: Ethernet
Exabyte	SCSI-2 command set. Single ended or differential	
IBM 3495	IPI protocol support	
Metrum RSS48 or RSS600	8-bit single ended SCSI	RS232
STK	Ethernet (?), FDDI (?), SCSI FIPS-60 (Block Mux), HIPPI (?), Fibre Channel (?)	

### 3.4 FSMS Suitable for a Network Attached Architecture

The only FSMS that supports network-attached storage as of 2/94 is NSL Unitree from IBM. While, arguably, the most palatable of the UniTree products currently on the market, it is prone to the deficiencies that are inherent in the UniTree product family.[7]

Several other vendors have expressed their intention to implement network attached storage management capability with their products. It is the assessment of the DADS personnel, that some FSMS products lend themselves to such enhancement, while other do not. However, to date, no other vendor has begun the work in this area.

### 3.5 Risks Summary

**RISKS:**

- 1) concurrent access control
- 2) hardware implementation
- 3) lack of appropriate FSMS
- 4) susceptibility to network or server failure
- 5) network performance

This page intentionally left blank.

## 4. Selected Architectural Approaches

---

A number of network layouts are possible. The three most characteristic ones, explored below, span a range from a completely off the shelf configuration to a configuration requiring sizable customization work. The advantages of all of the outlined configurations lie in the maximum efficiency of direct peer to peer communications. The data does not flow through the host CPU, which minimizes both the data latency and the CPU loading.

On one end of the spectrum is a flat, fully network attached, architecture. Clusters of client devices are attached to their specific FSMS servers and to other clusters via fabric switches. In this configuration data flow is direct from device to device. This configuration requires the least amount of custom work, consequently it is most reliant on the existence of the appropriate COTS hardware and software products. The configuration is discussed in Section 4.1, Direct Network Attachment. On the other end of the spectrum data flow is routed via I/O computers, which would require a sizable integration effort in software, hardware, or both. A concentrator based configuration takes the middle ground. The required buffering and signal and protocol conversions are performed by a concentrator switch.

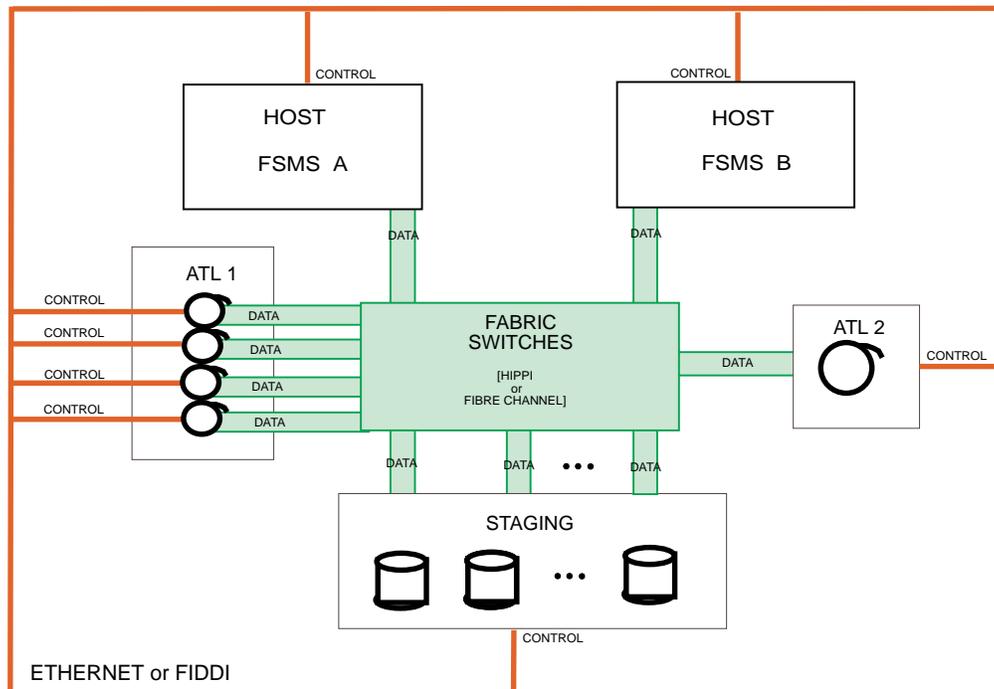
In all cases a control network is implemented separately from the data channel fabric. As discussed previously in this paper, a lower performance network is suitable for the control interconnection, but a higher throughput, parallel channel fabric will be required for the data path implementation. Therefore, in the following discussion the term *network* is used in reference to the control path interconnections. The term *channel* is applied in reference to the data paths. Such channels may be direct peer-to-peer connections or switched fabric interconnections as appropriate.

In the multi-FSMS environment, that is likely to be adopted for ECS, each FSMS will control its own set of peripheral devices (physical or logical). Each FSMS host with its associated peripherals and/or clients can, for the purpose of our discussion, be treated as an isolated network segment. Whenever *network* or *channel fabrics* are discussed below, it can be understood as a "single FSMS-centered" segment of an overall multi-FSMS controlled system.

### 4.1 Direct Network Attachment

#### 4.1.1 Configuration

In this configuration the FSMS is hosted on a dedicated FSMS server computer. Data switching is routed via fabric switches (HIPPI or Fibre Channel). The control signals use a network (Ethernet, FDDI). Each FSMS server controls all peripherals scheduling and initiates all data transfers for the associated peripherals. The data flow between the peripherals is peer-to-peer. Only such buffering as allowed by the individual devices or the device network adapters is performed. Figure 4-1 illustrates the architecture. Data flow to data requester is not shown.



**Figure 4-1. Direct Network Attachment Architecture**

### 4.1.2 Advantages

The network access control and the concurrent access control mechanisms are simple. The file system integrity is guaranteed by the central FSMS controller (not shown in Figure 4.1). Should the appropriate hardware and software be commercially available (see Section 4.1.4, Feasibility), this configuration is the least expensive, the easiest to implement and maintain, and the least prone to technology aging.

### 4.1.3 Disadvantages

The fabric switch is a single point of failure. Lack of sufficient buffering may become a performance problem. The transfer rate between two components will match the transfer rate capacity of the slower device.

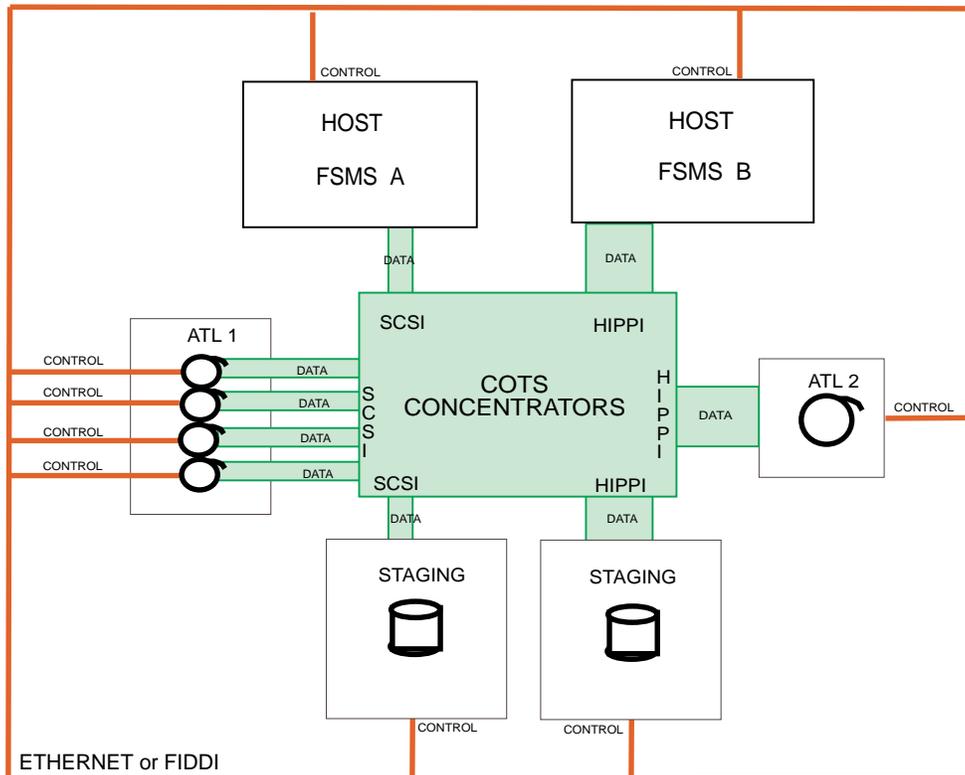
### 4.1.4 Feasibility

All network attached devices must either be capable of supporting a single communication interface protocol or have appropriate adapters installed at the point of network (channel) attachment. The existence of such COTS products is a significant risk factor. Among all the configurations reviewed here, this one is the most attractive, but also the least feasible at this time.

## 4.2 Concentrator Based Configuration

### 4.2.1 Configuration

As in the previous configuration, FSMS resides on the servers. An FSMS server controls all associated peripherals and initiates all data transfers. The control network is Ethernet or FDDI. It can be assumed, that most COTS devices would have control interfaces complying with a common communication interface standard. The data flow between the peripherals is routed through COTS concentrator adapters. Moderate amount of buffering is provided at the concentrators. Figure 4-2 illustrates the configuration.



**Figure 4-2. Concentrator Based Architecture**

### 4.2.2 Advantages

This is a more feasible configuration as opposed to direct network attachment via crossbar switches, because there is less of a demand placed on the availability of homogeneous network interfaces on the hosts and peripherals. This configuration is both less expensive and less complex to implement as compared to the dedicated I/O computers solution outlined in Section 4.3.

### 4.2.3 Disadvantages

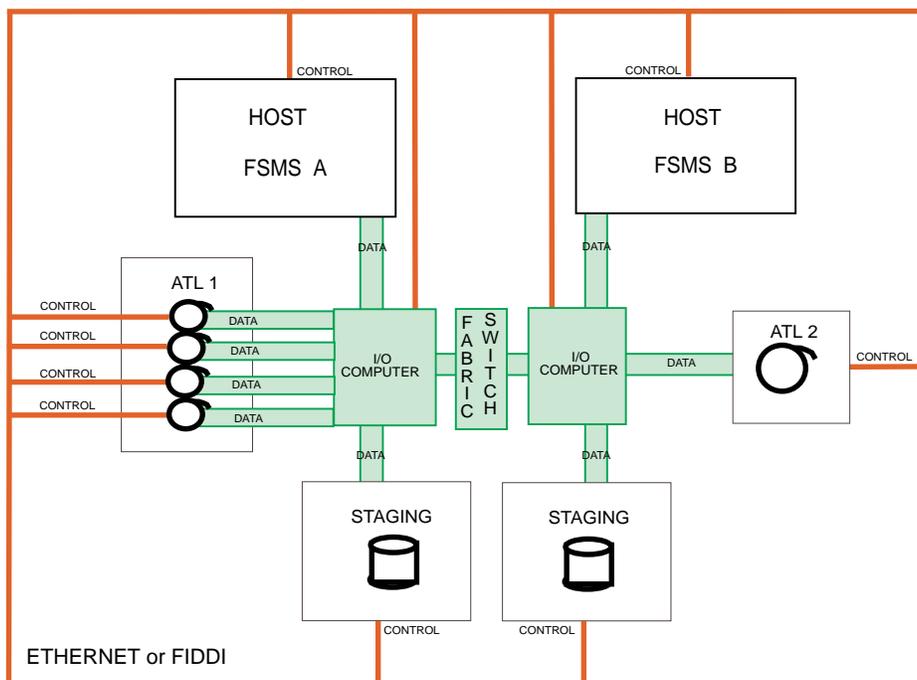
The foremost risk is in the availability of suitable concentrator switches on the market. There is a risk of the concentrator not being able to keep up with the data load. Adequate buffering is an important performance factor. The complexity of network is increased.

### 4.2.4 Feasibility

This configuration is more feasible than the previous one, but its performance is a risk factor to a point, where it could be so poor, as to invalidate the feasibility. The availability of suitable COTS products must be evaluated. It is also a risk factor. Careful study of DADS loading and throughput capacities of available COTS concentrators needs to be performed.

## 4.3 I/O Computers

Several options exist in implementing this configuration. 1) Use a general purpose computer to perform I/O function. 2) Use a custom ECS-developed configuration of COTS components to perform I/O function. 3) Use a special purpose COTS machine for I/O control. The three options are explored below. Figure 4-3 can be used to illustrate all three.



**Figure 4-3. I/O Computers**

## **4.3.1 General Purpose Computer**

### **4.3.1.1 Configuration**

A general purpose workstation class machine can be used to perform I/O buffering and protocol conversion. The FSMS host will be responsible for access and traffic control.

### **4.3.1.2 Advantages**

The benefits of a general purpose computer used as an I/O processor are the ease of availability and the relative simplicity of design and integration. Most of the custom work will be performed in software. Software changes can be introduced easier, than hardware changes, although not necessarily less expensively.

This solution is less vendor dependent than the one utilizing a custom, dedicated function I/O computer. At the cost of software adjustments a general purpose I/O computer can be replaced by another general purpose I/O computer. Therefore the design represents less danger of hardware aging influence on the network configuration overall.

### **4.3.1.3 Disadvantages**

The cost associated with the use of general purpose workstations and with the custom software development presents the most significant disadvantage.

### **4.3.1.4 Feasibility**

This configuration is the least dependent on the market availability of the appropriate hardware products. It is also the most dependent on custom code development.

## **4.3.2 Custom Solution Using COTS Components**

### **4.3.2.1 Configuration**

This solution to DADS I/O interconnect consists of custom ECS configured rack-mounted COTS hardware, such as microprocessor-based Single Board Computer (SBC) boards.

### **4.3.2.2 Advantages**

It is the most flexible configuration that is geared to the function it performs. If designed correctly, it is also the best performing configuration. There is no dependency on the market availability of the specific communication hardware. The configuration is much less prone to aging, since keeping it "in stride" with technology advances implies changing SBC components rather than the entire configuration. The cost of the hardware is the lowest in the I/O computer category.

#### **4.3.2.3 Disadvantages**

The hardware integration and software development efforts involved are the most significant among all of the outlined configurations. If this path is chosen, its impact on the delivery schedule must be carefully evaluated.

ECS DADS would be fully responsible for the maintenance of the I/O computer hardware.

#### **4.3.2.4 Feasibility**

There are no implementation obstacles, aside from the potential schedule impacts mentioned above.

### **4.3.3 Dedicated Function I/O Computer**

#### **4.3.3.1 Configuration**

Small-scale commercially available dedicated computers are used for the data channel I/O control. Software development will be required.

#### **4.3.3.2 Advantages**

This configuration has the potential of being just as flexible and well performing as the one custom configured by ECS.

#### **4.3.3.3 Disadvantages**

This configuration is most prone to technology aging.

#### **4.3.3.4 Feasibility**

No obstacles to implementation aside from the hardware availability on the market.

## 5. Conclusion

---

Several large distributed archives, such as the ones mentioned in Section 3 exist at this time. The issues of network attached storage are being pursued by a number of vendors. The recent advances in disk and tape drive technology as well as continuing standardization efforts are going to provide DADS with attractive viable design options. Network fabric products are available from a growing number of vendors and should be the area of the least risk.

However, the very recency of the product and standards development, and the EOS charter of evolving EOS DADS with the evolving technology, make every other aspect of the network attached storage non-trivial to integrate. If ECS DADS is to do the integration successfully, proof of concept prototyping must be done for the most risky portions of the storage configuration: I/O control to storage device and staging to storage device attachments. As experience at the other large archives has shown, performance and even feasibility surprises spring up in both in the hardware and in the micro code.

Another area of high risk, perhaps overshadowing the hardware concerns in importance, is the timely availability of a choice of FSMS products capable of performing the task. Aside from the NSL UniTree no other commercial FSMS can support network attached storage. Both LANL and NCAR, for example, use FSMS of their own design.

This page intentionally left blank.

## 6. References

---

1. "The MaSSIVE<sup>TM</sup> Project at NCAR", J. Sloan, et al., Twelfth IEEE Symposium on Mass Storage Systems, 1993
2. "High-Performance Network and Channel Based Storage", R. Katz, Proceedings of the IEEE, Vol. 80, No. 8, August 1992
3. "An Introduction to X3T9 and I/O Interface Standards", W. E. Burr, 13 February 1992.
4. "HIPPI World -- The Switch is the Network", K. Hardwick, Network Systems Corporation, 1992
5. "Los Alamos HPDS: High-Speed Data Transfer", W. Collins, et al., Twelfth IEEE Symposium on Mass Storage Systems, 1993
6. "FDDI: Current Issues and Future Plans", R. Jain, IEEE Communications Magazine, September 1993
7. "FSMS White Paper", T. Smith, ECS DADS, 1993
8. "FDDI-II Operation and Architectures", M. Teener, R. Gvozdanovic, IEEE
9. "HIPPI / Serial-HIPPI", D. Tolmie, M. Halvorson, IEEE, 1992
10. "HIPPI: Simplicity Yields Success", D. Tolmie, J. Renwick, IEEE Network, January 1993
11. "High Performance Networks" manual, NS1364, Network Systems Corporation, 1993
12. "Following the Fiber Distributed Data Interface", R. Fink, F. Ross, IEEE Network, March 1992
13. "High Performance Switching with Fibre Channel", T. Anderson, R. Cornelius, Ancor Communications, Inc., IEEE 1992
14. "The High Performance Storage System", R. A. Coyne, H. Hullen, R. Watson.
15. "Distributed Computing with Fibre Channel Fabric", K. Malavalli, B. Stovhase, Canstar, IEEE, 1992
16. "HIPPI", J. Hughes, IEEE, 1992
17. "An Introduction to X3T9 and I/O Interface Standards", W. Burr, NIST, 1992
18. *Data Communications and Networking Dictionary*, T. Pardoe, R. Wenig, Professional Press Books, 1992

This page intentionally left blank.

# Abbreviations and Acronyms

---

ANSI	American National Standards Institute
ATL	Automatic Tape Library
CD	Collision Detection
COTS	Commercial Off The Shelf
CPU	Central Processing Unit
CSMA	Carrier Sense Multiple Access
CSMS	Communications and System Management Segment
DADS	Earth Observing System (EOS) Data and Information System (EOSDIS) Core System (ECS) Data Archive and Distribution System
ECS	Earth Observing System Data and Information System Core System
ESCON	Enterprise Systems Connection
FDDI	Fiber Distributed Data Interface
FFOL	FDDI Follow-On LAN
FSMS	File Storage Management Systems
Gb	Gigabit
HDT	HIPPI Data Transfer
HIPPI	High Performance Parallel Interface
HIPPI-PH	High Performance Parallel Interface - Mechanical, Electrical, and Signaling Protocol Specification
HIPPI-FP	High Performance Parallel Interface - Framing protocol
HPDS	High Performance Data System
HIPPI-LE	High Performance Parallel Interface - Link Encapsulation
HPSS	High Performance Storage System
IEEE	Institute of Electrical and Electronics Engineers
I/O	input and output
LAN	Local Area Network
LANL	Los Alamos National Laboratory
LLNL	Lawrence Livermore National Laboratory
Mbps	Mega bits per second

MB	Mega Bytes
NCAR	National Center for Atmospheric Research
NSL	National Storage Lab
RAID	Redundant Array of Inexpensive Disks
SBC	Single Board Computer
SCSI	Small Computer System Interface
TCP/IP	Transmission Control Protocol/Internet Protocol
ULP	Upper Layer Protocols
IPI	Intelligent Peripheral Interface