N93-14773

Network Issues for Large Mass Storage Requirements

James "Newt" Perdue

Ultra Network Technologies, Inc.
San Jose, California USA

CLIENT/SERVER NETWORK & STORAGE NEEDS

The major performance demand on today's networks in the Science and Engineering environment by far derives from mass storage requirements. The need to move massive amounts of data between the different parts of the computing environment dictate the topology and performance requirements of the local area network. This paper will explore such requirements and provide some insights into solutions which address the increased need for network performance as a result of the explosive growth of data in the science and technology area.

Data plays a key role in determining the architecture and performance needs of a computing environment. Basically, mass storage is the repository for the data and information which drive the entire computing scenario. In fact, we can think of mass storage as holding the major assets of any institution or corporation. Here is stored the "kings jewels" of the organization. In today's society information is the king and rapid access to it is the king's road. If we look at Data as the center of any organization (See fig 1), Compute Servers and Clients surround it with paths for fast access between Servers, Clients and data.

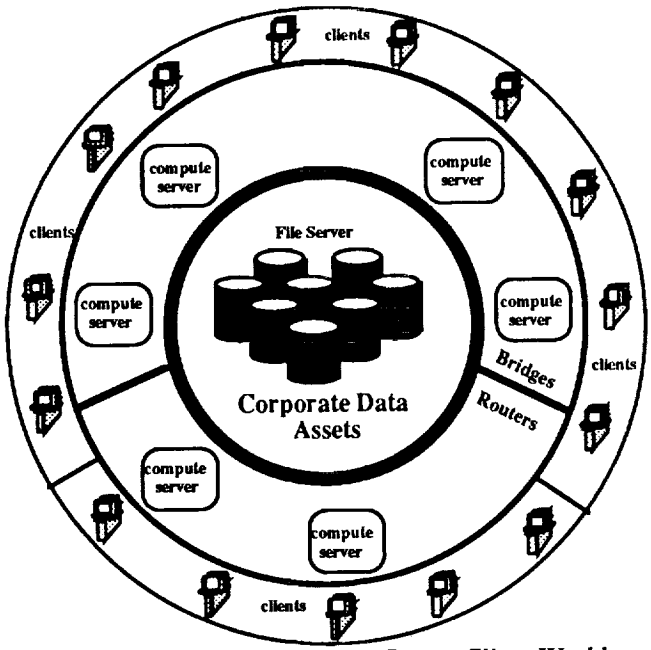


Figure 1 - File Server In The Server-Client World

These servers are usually linked to one or more data or File Servers (containing the corporate data) by a local area network with large throughput and bandwidth capabilities. Outside of a ring of such Compute Servers, lies the users of the information, the clients. They also require fast access to the data but usually in lesser amounts than the Compute Servers since they are looking at the pieces in small amounts, analyzing it in some greater detail. The clients in todays scenario are usually a variety of workstations and personal computers linked through a local area network with low to medium bandwidth and throughput capabilities. This describes the Client-Server scenario today: fast compute servers, such as supercomputers, near-supers, and high end workstations, sharing data between themselves and file servers, all working to create and manipulate the data into a form accessible by many clients, usually workstations and personal computers.

The central file server in this scenario provides for the:

- storage of corporate data in a single secure location;
- rapid access and movement of data between compute servers;
- creation of a hierarchy of storage devices for economic handling of the data;
- remote data access to workstation/pc clients in both file and record oriented (diskless operation) modes;
- temporary storage of massive data sets for distributed processing needs or common access requirements;
- storage and retrieval of large graphic images for later playback (digital VCR);
- cacheing data between Compute Servers and networks of differing speeds;
- long term reliability of the storage by implementation of archiving methods.

These are heavy demands in the supercomputing environment due to the very rapid growth of the amount of data required to feed the ever faster Compute Servers and the need to save data for both development and liability needs. In most supercomputing environments today, it is not uncommon to see existing on-line storage requirements greater than 250 Gigabytes of storage, and on-line tape storage (silo's) in the range of several terabytes of data. It's also clear that most users consider these capacities to be inadequate for the near future.

These demands create several issues for access to the storage and thus the local area networks. Because the individual file sizes can reach several gigabytes in size, many local area networks can't handle them. Often the mean time to failure of a network can be less than the time to transfer such a file making file transfer via a network not feasible. Such a single file can take over an hour to transfer between servers or servers and clients. Multiple files being transfered at once can literally stop

PRECEDING PAGE BLANK NOT FILMED

an ethernet from functioning due to the congestion of such transfers.

The usefulness of a local area network must be measured in terms of it's ability to provide EFFECTIVE performance for such files, not by the rated performance or speed of the bits on some part of the wire connecting hosts. Further, the need to transfer such large files places large burdens on the processing capabilities of the servers and clients involved in the transfer. On typical hosts today, protocol processing consumes between 40 to 100% of the CPU power to maintain an average of 8 megabits per second transfer capability (ref: SHIFT, CERN 2 Mar 91). Further, as the speed of the network "wires" increase the demand on the CPU increases if it is to maintain efficiency. Further, most host system structures are undersized for the efficient movement of large data files:

- buffers in both system and applications are small, causing many interupts to host I/O systems,
- disk "effective" transfer rates are not sized to the mammoth file sizes;
- copying between various system buffers creates a large amount of overhead which affects the transfer rate and the cpu utilization of the host systems;
- the "wire speeds" of the ethernet, and even FDDI systems don't measure up to the needs to move gigabytes of data.

Clearly, for the science and technology marketplace, if the file systems and networks are to maintain the pace set by the CPU industry, major improvements must occur in the integration of technologies.

Fortunately, the technologies needed to address these requirements are moving forward at an acceptable rate. But it is not enough for technology to be available, for there must also be the ability to integrate the elements of technology into products which address the specific needs. In figure 2, major technology developments affecting the computing and storage needs are listed in a relative timeline. From this diagram one can see trends that may come to our rescue. Developments such as International Standards for the definition of File Server interfaces such as the one in development by the IEEE Mass Storage Committee, standards for high-speed data connectivity such as HIPPI (ANSI X3T9.3), development of high-speed protocol processors such as UltraNet, and development of new disk architectures such as the RAID devices all contribute technology to the solution of the next generation high performance File Servers. Already such technologies are being combined to give us a taste of what's to come: DISCOS is supporting the IBM RAID product in a distributed environment with UniTree™, Cray Research has a Data Migration Facility which uses a dedicated Cray using HIPPI for I/O and striped disks for increased throughput, and NASA Ames has developed their own file system software using parallel channel connections (8) for both disk and network I/O from Amdahl systems to be able to achieve transfer rates to the Cray up to 20 MBytes/second.

The way we treat data in the supercomputing environment has certainly evolved over the last 20 years. As shown in Figure 3, file systems originally were thought as part of the host which they served. Each host owned the data it produced and networks permitted sharing by moving entire files across the network when required. Such sharing was not so important in this scenario due to the slow transfer rates possible. Next as network speeds increased, a concept of a centralized file store was introduced and is widely accepted today as the architecture most applicable to the networked environment. This permits access both at record level and file level to any host on the network.

Data produced by a host may still reside on that host, but the opportunity to move it to a central system for later retrieval by other systems or for archive of the data is now possible. This reduces the amount of disk space required on each host and has the advantage of permitting economies of scale to apply for storage purchased for the File Server. Of course, the File Server remains a point of failure for the entire system, and generally causes a large performance bottleneck for access to the data.

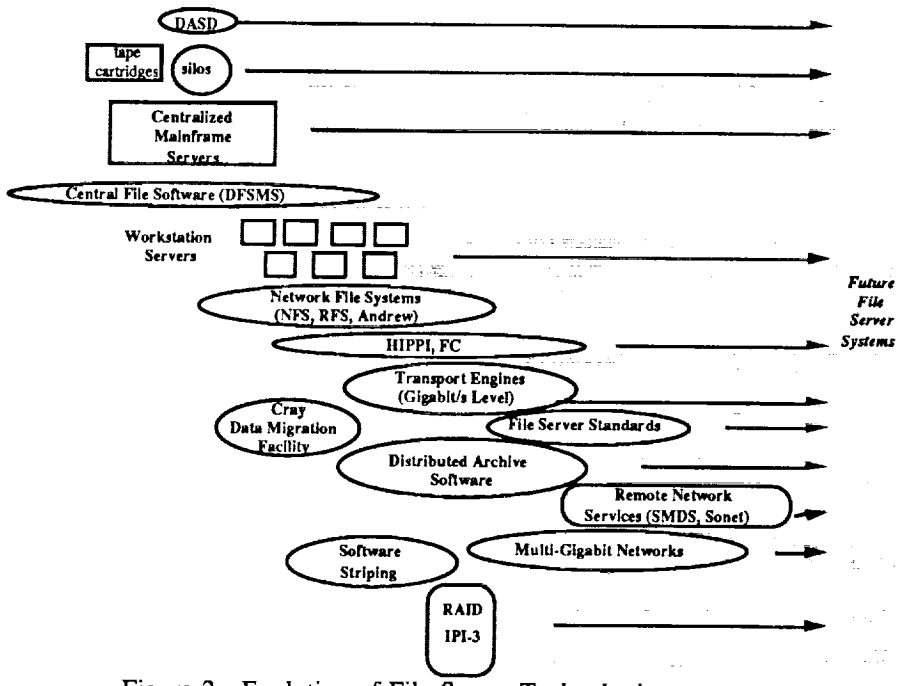


Figure 2 - Evolution of File Server Technologies

However, the most likely future scenario for File Servers is seen as the last diagram in figure 3. In the Distributed File Server scenario, data may or may not be associated directly with a specific host or even a specific File Server CPU. Software maintains knowledge of the location of the files throughout the network and manages its migration over the network from source to requestor as a third party manager rather than directly manipulating the storage itself. This scenario presents the possibility of connecting storage directly to the network itself, without a host to manage it due to the evolution of two technologies: intelligent controllers and standardized high-level command languages such as Intelligent Peripheral Interface Level 3 (IPI-3). The intelligent controllers can play the role of the low-level disk driver and the network interface. Further, with the availability of network interfaces with the effective performance of high-speed I/O channels AND the ability to run network protocols at the TRANSPORT level (such as the UltraNet Network Processors), these stand-alone disk servers are made even more practical. This scenario may provide the supercomputer owners freedom of choice in INDEPENDENTLY selecting the peripherals, the CPU's, the file software and the network. Each element can evolve it's own competitive marketplace which is sure to drive the prices for mass storage, computer systems and networks to more advantageous levels for the users.

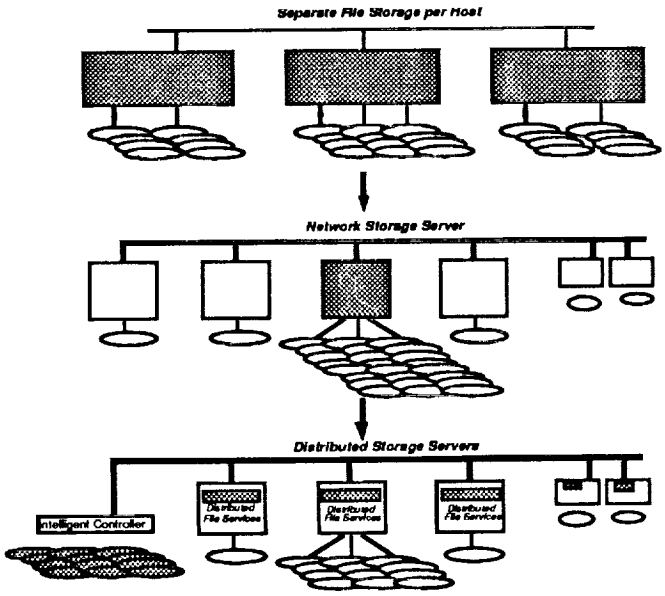


Figure 3 - Trends in File Server Architecture

Of course, performance is still a major consideration when designing file systems for the supercomputer environment. Figure 4 diagrams four different approaches to increasing performance to disk systems. Today, several vendors (including Cray Research) have implemented software to stripe the data from several disk drives in parallel to achieve raw disk transfer rates in the range of 32 to 120 MBytes per second.

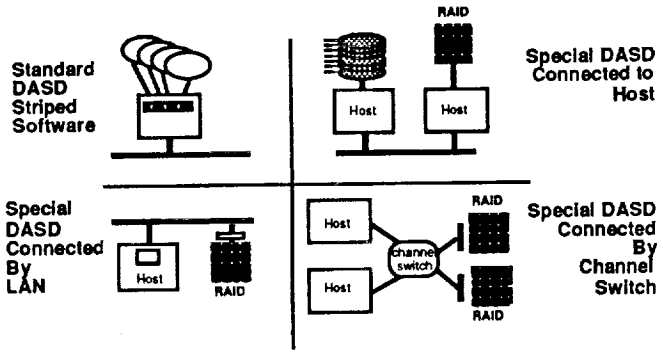


Figure 4 - Methods To Achieve High Performance Disk Throughput

Generally the drawback of this approach has been the problem of "all the eggs in one basket". If a disk fails, it is possible to lose the entire (VERY LARGE) file store. Another approach includes the use of parallelism in the disk devices themselves. Parallel heads on a single platter can increase transfer rates today up to about 20 MBytes per second, and a new area of development called RAID (Redundant Arrays of Inexpensive Disks) parallelize the data streams from a number of inexpensive disks. The advantage of this approach includes the ability to offer redundant paths to protect against most loss of data. For network access, several possibilities arise. The CPU can manage the RAID or parallel head disk devices directly and pass the data across the network. The main problem with this approach is that most networks today cannot maintain the transfer rates required. Another approach now possible with the introduction of IPI-3 and HIPPI interfaces is the connection of the RAID devices directly to the network. Although this has not been done in any operational implementation yet, it is possible and developments are in progress. Finally, with the introduction of HIPPI channel switches, these devices can be connected between hosts (which have HIPPI channels) much in the same way that multi-channel controllers permit access by more than one host.

The most promising approach for increased performance with economic rewards may prove to be the distributed file server concept with network attached disk devices and peripherals. A major advantage of this approach is the ability to move data directly from the disk device to the requestor without going through a host mainframe, which only adds to the performance overhead and the cost of the system. The data management software, if centralized can reside on a much smaller host, such as a workstation, with dramatic savings possible in both initial capitalization, maintenance costs and in-house system personnel costs. In this scenario, the network must be fast enough to maintain effective rates higher than a single host to a variety of hosts with a variety of connection capabilities (BMC, HSX, HIPPI, LSC, VME, Microchannel, etc.). A standard disk I/O command language, such as IPI-3 makes it possible to ask for block data and the intelligent controllers execute the low-level

commands required for disk I/O. Large blocks then get sent to the destination directly over the network.

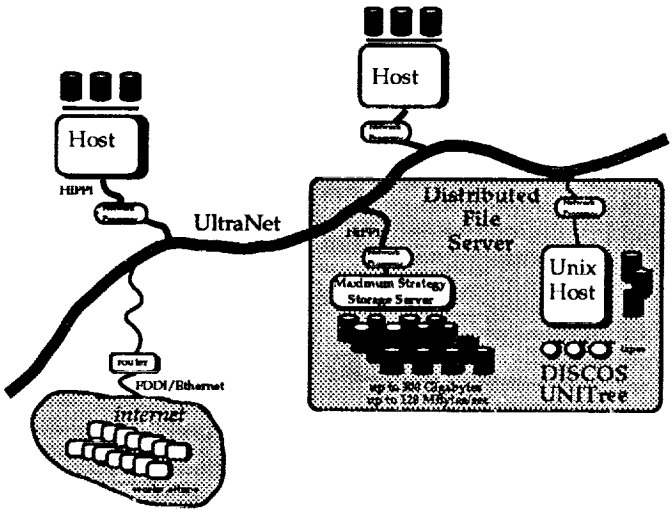


Figure 5 - Concept for Distributed File Server

Figure 5 diagrams a concept for a Distributed File Server using network attached peripherals. In this scenario, a RAID disk with HIPPI interfaces permit transfer to hosts at transfer rates between 20 to 40 MBytes/second. In this specific scenario, UltraNet is proposed for it's high performance and it's ability to do the network protocol processing, Maximum Strategies HIPPI RAID devices for maximum transfer performance and redundancy (using HIPPI channels and IPI-3 command languages), and DISCOS UniTree for it's distributed hierarchical storage management software. Although this combination must still be proven, it is an example of a system that could be constructed to provide a completely distributed file server environment with very high performance and a variety of connectivity (must greater than allowed by HIPPI only devices). Figure 6 (at the end of this paper) presents details of this implementation.

SERVER NETWORKS

From a network perspective two major requirements exist for performance-based file access. First, Server-to-Server traffic must be managed. Data from a single users job may exist on several servers, and may take several days to accumulate. Data is transfered between the servers to accomplish the integration of the task. Gigabytes of data flow between File Server and Computer Server each day. Server-to-Server traffic accounts for a large amount of the traffic in a local area network.

Second, once the data is computed, client systems such as workstations need access for data analysis, visualization and presentation. Usually transaction-oriented access, such as that provided by NFS dominates Client-Server communications. The large databases are generally not transfered to the client, but only accessed in pieces as needed. Further, this Server-Client access serves as the path for software development, not requiring major amounts of traffic but frequent access.

Therefore, in view of the Client-Server model discussed earlier, VOLUME data is required between Servers but TRANSACTION oriented traffic dominates between Client and Servers. An ideal network model for this would segment the network in such a way that permits use of multiple network technologies. For the Server-to-Server traffic utilize the highest speed network technology and for the Server to Client (and Client-Client) utilize the technology with the lowest cost, highest connectivity and most standardization.

Figure 7 contrasts two approaches using technology available today. One connects both Servers and Clients using a single network.

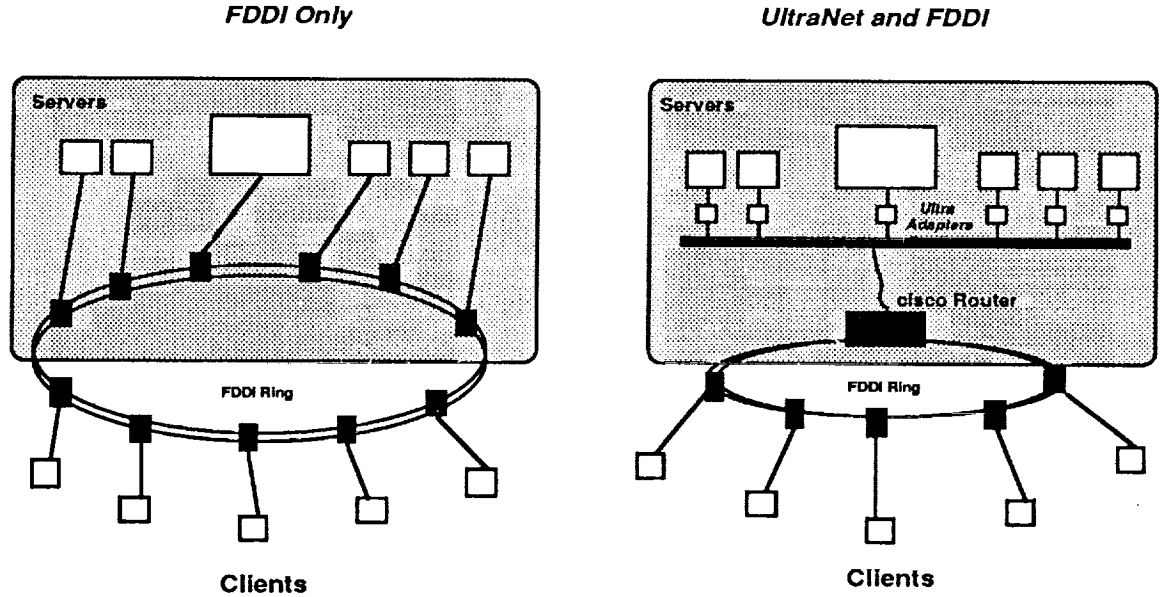


Figure 7 - Alternatives for Networking Servers

The other connects two networks via a high-speed router. In the later instance, one network excels in large aggregate performance (for backboning) and also in high task-to-task transfer rates for Server-to-Server communication. The other network would be more transaction oriented, standard, and have a lower cost point.

By measure of published EFFECTIVE performance results, the only network available today which can maintain the task-to-task data rates and the large aggregate rates needed in the supercomputing scenario is UltraNet. In figure 8, UltraNet is used as a FRONT-END network for Servers bridging (routing) to Clients connected by FDDI. One (or more) high-speed routers are needed to connect to FDDI. No Server needs a direct connection to FDDI or ethernet. This saves the cost of multiple network interfaces for the Servers. As a Server Network, UltraNet can sustain at least a gigabit/second aggregate transfer rate (for the Backbone function) and can sustain task-to-task transfer rates up to 50 MBytes/second. Connectivity to multiple hosts are available using interfaces such as HIPPI, HSX, LSC, BMC, VME and Microchannel.

As a way to explore the merits of using UltraNet as a Server Network instead of using only FDDI, a simple network model was built and then tested to confirm it's results. Figure 9 shows the results of the network model. For the FDDI (or ethernet) only solution (on the left), the same maximum transfer rate is achieved for either Server-Server traffic or Server-Client traffic. The maximum transfer rate from the host, in the demonstrated case is limited to an effective rate of ~1.5 MBytes/second for FDDI and is evenly shared between Server traffic and Client traffic. If 50% of the traffic is between Servers, the maximum available bandwidth is .7 MBytes/sec, and the other .7 MBytes/sec is available for Server-Client traffic. In the UltraNet scenario, the Server-Client traffic is still limited to the .7 MBytes/sec (due to bottlenecks in the Client systems), but between Servers, UltraNet can now

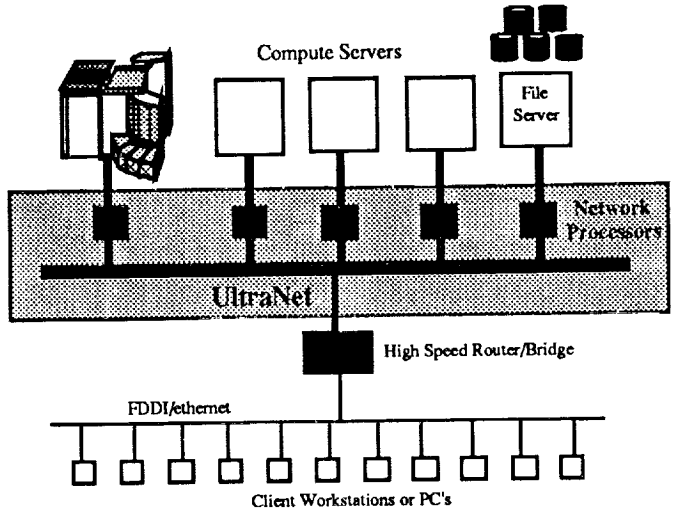


Figure 8 - UltraNet as a Server Network Solution

provide up to 4 MBytes per second while still reserving 50% of the host transaction capabilities for Server-Client traffic. When all traffic is between Servers UltraNet can provide up to 9 MBytes/second. This example features a typical low end near-supercomputer as Compute Server. For Cray based systems, the UltraNet can provide over 40 MBytes/second effective performance. This model was based on the number of transactions per second possible by the host for I/O, and the amount of data that could be transfered per transaction. In this case, I/O to Clients is limited to FDDI packet sizes of 4.5 Kbytes. For each transaction, only 4.5 Kbytes is transfered. Between Servers connected through the UltraNet, 32 Kbytes of data can be transfered, over 7 times the amount of data for a single FDDI transaction.

Figure 10 shows the results of an actual test done to demonstrate this point. Near-supercomputers (Convex C-1 and Alliant FX-80) were used as the main Compute Servers together with several workstations. Although FDDI was not actually used, it was simulated by limiting the transactions to 4.5 Kbytes. An actual test using 6 computers was run using the TSOCK test program and the

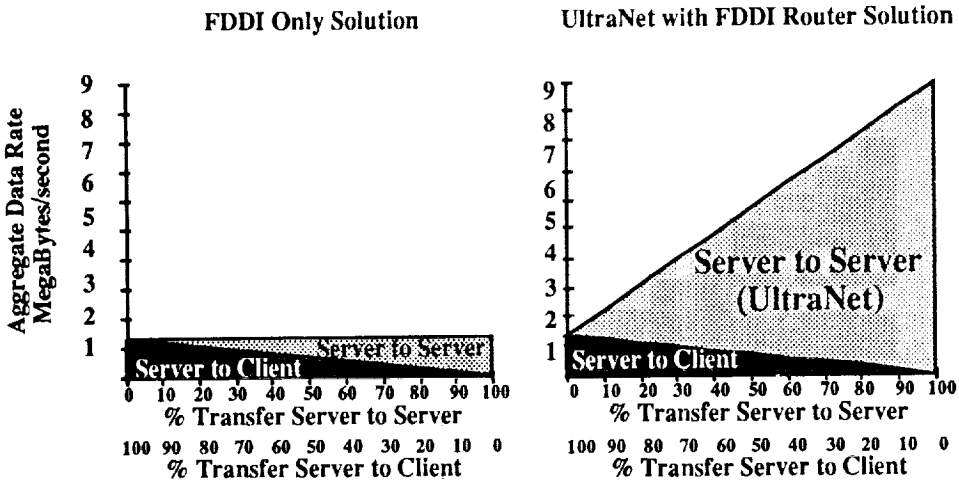


Figure 9 - Results of Modeling Server Network Scenarios - FDDI only vs UltraNet/FDDI

results show that the model in figure 9 is very closely approximated. When 50% of the traffic is to Clients (using 4.5 Kbyte transactions), about half of the FDDI sized traffic is possible (about .5 MBytes/s in this test). But at the same point, UltraNet provides over 5 MBytes per second to the other Servers.

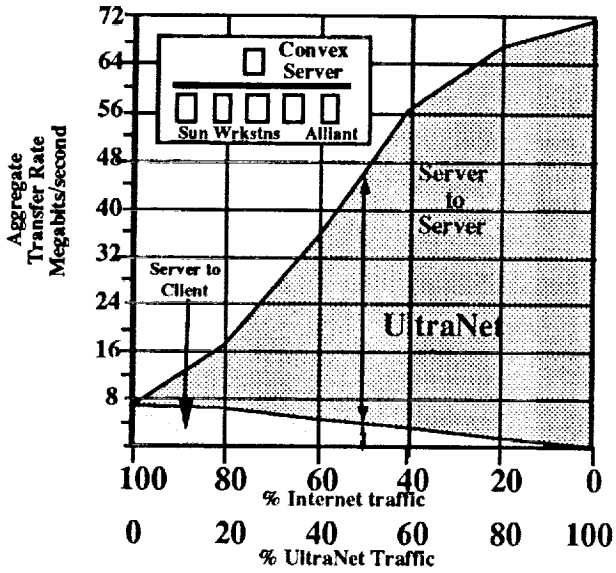


Figure 10 - UltraNet as Server Network (Actual Results)

The point of this is to demonstrate that if a network architecture can be selected which maximizes the transfer rates for Server-to-Server traffic and minimizes the cost of the Server to Client traffic then the best result is achieved for implementing high performance file systems.

Another factor which is important in evaluating networks to use for file systems is the total aggregate data rate possible. Task-to-task transfers are extremely important

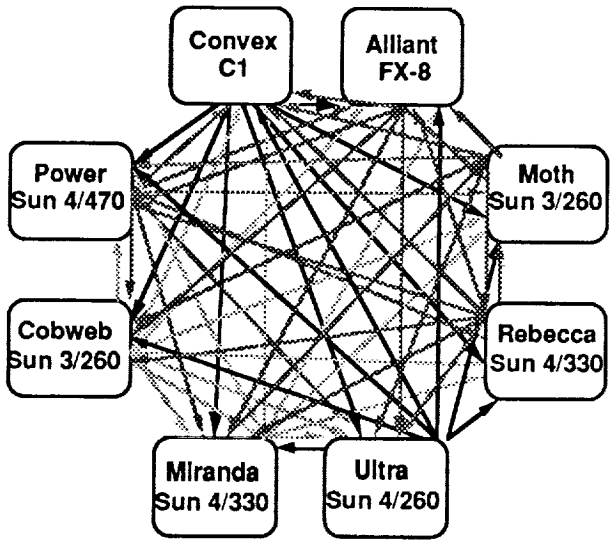


Figure 11 - Test Environment for Bandwidth Test

but more important in an busy network environment is that the aggregate transfer rate does not drop off dramatically when additional conversations are added. Figure 11 shows the basis for an experiment using UltraNet and ethernet to demonstrate this point. Eight computers, (including Convex, Alliant, and Sun Servers) were used to demonstrate this point.

Each computer transmitted and received to every other computer 114 MBytes of data. Each computer established 7 vitual circuits to each of the other computers (a total of 56 virtual circuits for the test). Each computer began the transfers within about 5 seconds of the other. A total of 3,600 MBytes of data was transfered between the computers using the TSOCK test program. Each computer had both an UltraNet connection and an ethernet connection. Two ethernet segments were utilized to increase the aggregate transfer rate on the slower network.

Figure 12 demonstrates that UltraNet delivered the entire 3.6 Gigabytes of data in less than 6 minutes. Ethernet, on two circuits complete the transfer in 26 minutes. If one segment had been used, it would have take over 52 minutes. In the UltraNet case, each computer sustained over 3 MBytes/second, generally limited by the Sun workstation transfer rates.

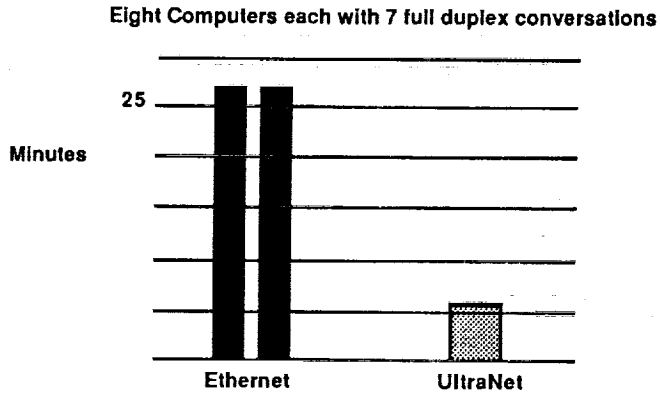


Figure 12 - Results of Aggregate Bandwidth Test

Ethernet performed well in the test from the viewpoint that each segment sustained over .83 Megabytes/second or almost 65% of the ethernet bandwidth. UltraNet sustained over 11.8 MBytes per second. However, with UltraNet, only 10% of the total bandwidth of 125 MBytes/second was utilized, leaving another 110 MBytes/second of bandwidth for additional conversations.

Although this test shows the large aggregate capability of UltraNet as a Server Network, probably more instructive is how it performs when supporting actual file applications. Disk-to-disk transfer rates are most instructive in evaluating any network to be used for a file server. Figure 13 summarizes the results of a test performed by Cray Research on the ability of FTP to transfer between two Supercomputers over UltraNet. The

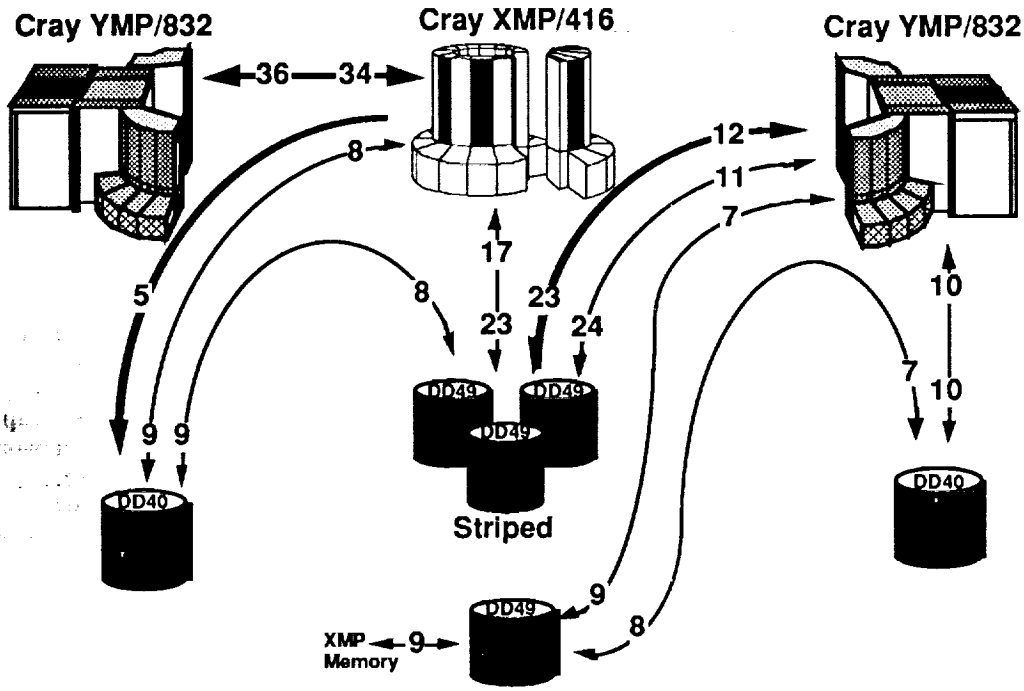


Figure 13 - FTP Test Results Using UltraNet with 2 Cray's

Cray XMP had both striped (3X) and non-striped disks available. The Cray YMP had only non-striped DD40 disks available. In this test, FTP was modified to have buffers of up to 1 MByte in size to take advantage of the transfer capabilities of the UltraNet.

Clearly, this test shows that generally speaking the UltraNet was not a bottleneck in maintaining the high disk transfer rates available on the Cray. Over the network, the Cray YMP could write the striped disks on the Cray XMP as fast as a user sitting directly on the Cray XMP (23 MBytes/sec). Disk-to-disk rates between the DD40 on the YMP and the DD49 on the XMP was very close to the non-network rates (8 MBytes per second).

Finally, it should be instructive to examine the network performance of an actual installation using a high-speed Compute Server (Cray 2 and YMP), file servers (Amdahl 5880) and workstation computer servers (SGI). At NASA Ames Research Center, UltraNet is installed as a gigabit/second backbone across several buildings connecting the supercomputers, file servers and more than 40 SGI workstations (all equipped with the Powerchannel I/O option). Figure 14 diagrams the configuration at NASA Ames Research Center.

Tests were run on a variety of these hosts to demonstrate actual performance achieved in a variety of scenarios. Each test was run in a heavily loaded system with over 100 users logged in and competing for resources. Therefore, each run was repeated several times (variation noted in the results) to give an idea of the range of results possible. Dedicated testing should prove higher effective data rates.

Figure 15 summarizes this data. For memory-to-memory tests, using the TSOCK application, it is observed that the maximum transfer rates possible approach up to 92 MBytes per second for a single graphics application. (Over 32 MBytes per second still left for other traffic.) For transfer between Computer Server (Cray YMP) and File Server (Amdahl 5880), UltraNet can sustain memory-to-memory transfer rates approaching 22 MBytes/second using striped BMC channels on the Amdahl (UTS) and HSX channels on the Cray (UNICOS). Although testing has not been performed as yet on the disk to disk rates on the Amdahl, it is expected that near 20 MBytes/sec can be achieved due to the software striping possible at the NASA site on the UTS based Amdahl 5880 system.

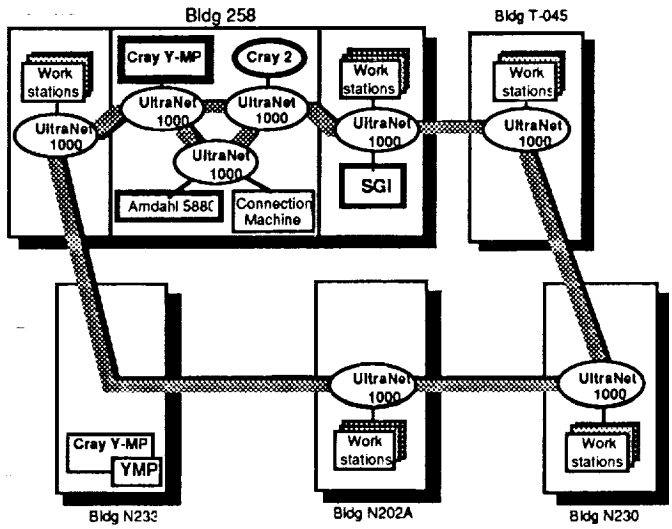


Figure 14 - NASA AmesGigabit/s Network Configuration

Perhaps notable also is the transfer rates between the disk on the Cray YMP and the memory of the SGI workstation. A user running an interactive application on the workstation can sustain data transfer at rates of over 4.5 MBytes per second from the Cray YMP disk. This would permit a user to run a large simulation on the Cray and access the results as it progresses interactively from the SGI without interrupting the Cray simulation.

For disk to disk tests, FTP was used between the various computers. FTP between the Cray and SGI disks maintained the maximum data rate possible for the SGI disks (about 1.5 MBytes/second) as demonstrated by timing the SGI disk rates (using the dd command in UNIX). Between the Cray computers, FTP transfered at somewhat lower rates than possible from a single Cray to it's own disk, in the range of 2.5 to 3.5 MBytes/ second over the network. However, it was shown by TSOCK disk to disk tests that if the FTP buffer sizes are increased to 1 or 2 MBytes the transfer rates approach that of the dd rates on a single disk (without the network).

Transaction oriented file access provided by NFS was also measured. Only NFS reads are possible at NASA Ames. Users can create the files only on the same computer they are using, but can read disks attached to the other Cray using NFS over UltraNet. Transfer rates were significantly better than what might be achieved using Ethernet, but were considerably lower than those achieved using FTP. NFS is a very transaction oriented protocol, uses the host stack instead of the network-based protocol processing provided by an UltraNet processor and therefore sees much less performance than other applications over the UltraNet. However, the Ultra was still able to provide from 1 to 2.6 MBytes/ second transfer rates, more than available with ethernet. However, UltraNet can support a much large number of NFS transactions than ethernet (not shown here).

It is expected that future modifications to NFS by Sun Microsystems and improvements in UltraNet's ability to handle the small packets of NFS transfers will improve this NFS rate.

In summary, File Servers and Supercomputing environments need high performance networks to balance the I/O requirements seen in today's demanding computing scenarios. UltraNet is one solution which permits both high aggregate transfer rates and high task-to-task transfer rates as demonstrated in actual tests.

UltraNet provides this capability as both a Server-to-Server and Server to Client access network giving the supercomputing center the following advantages:

- highest performance Transport Level connections (to 40 MBytes/sec effective rates)
- matches the throughput of the emerging high performance disk technologies, such as RAID, parallel head transfer devices and software striping;

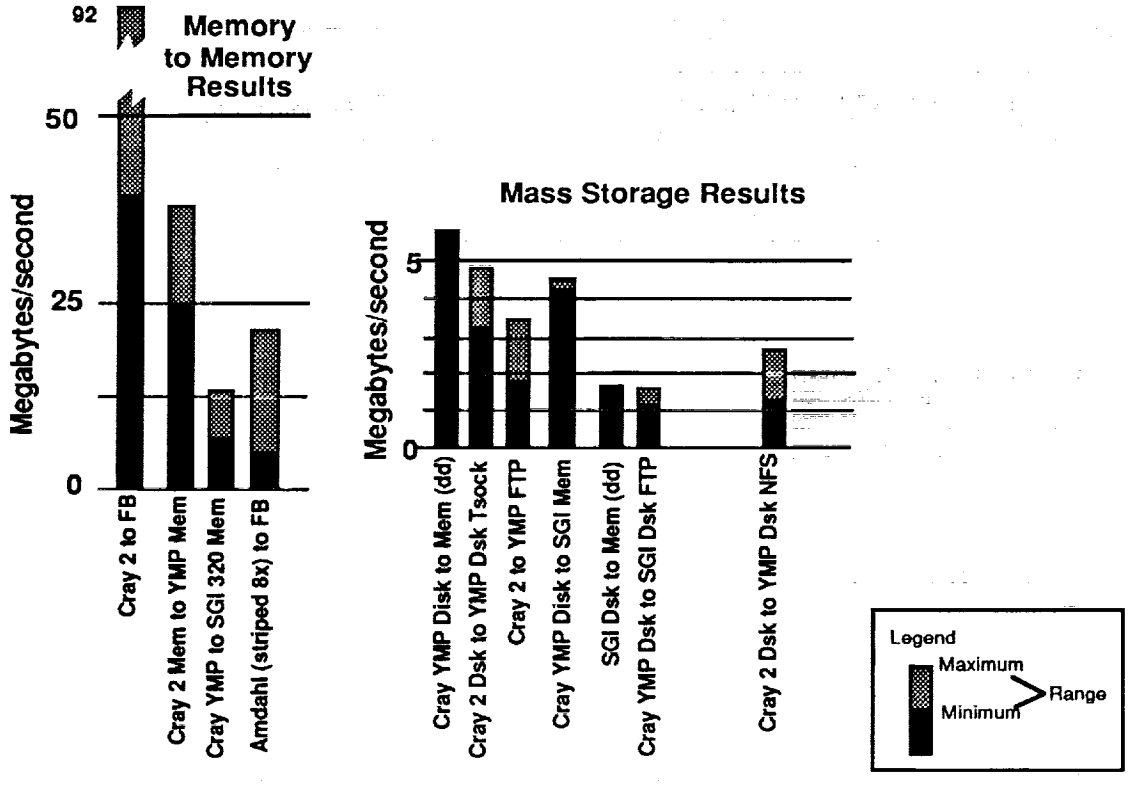


Figure 15 - Typical Application Test Results At NASA Ames Research Center (UltraNet)

- supports standard network and file system applications using SOCKET's based application program interface such as FTP, rcp, rdump, etc.
- Supports access to NFS and LARGE aggregate bandwidth for large NFS usage;
- provides access to a distributed, hierarchical data server capability using DISCOS UniTree product;
- Supports file server solutions available from multiple vendors, including Cray, Convex, Alliant, FPS, IBM, and others.

This paper appeared in the Cray User Group Spring 91 Proceedings (London, England).

UltraNet® is a registered trademark of Ultra Network Technologies, San Jose, California, USA. UTS is a trademark of Amdahl Corporation. UniTree is a trademark of DISCOS, GA Technologies, San Diego, California, and Cray YMP, Cray XMP and UNICOS are trademarks of Cray Research, Inc., Minn, Minn.

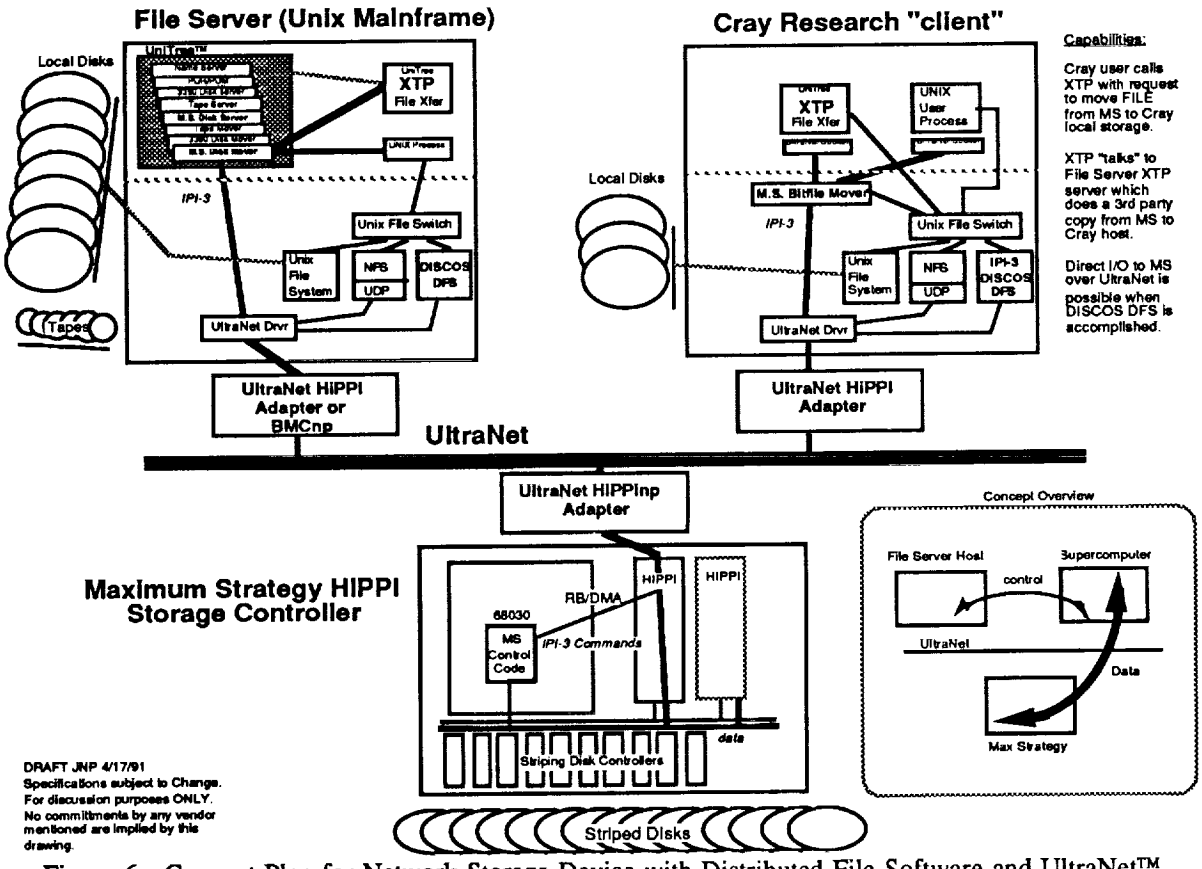


Figure 6 - Concept Plan for Network Storage Device with Distributed File Software and UltraNet™

Network Issues for Large Mass Storage Requirements

Presented to the NSSDC Conference on Mass Storage Systems & Technologies for Space & Earth Science Applications

By

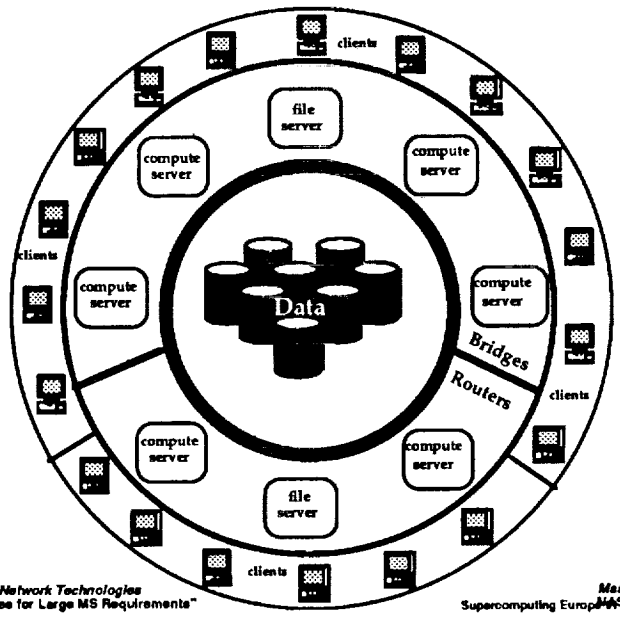
Newt Perdue
Vice President
Ultra Network Technologies
San Jose, California
U.S.A.

July 24, 1991

Overview

- **File Server Network Requirements**
- **File Server Performance Trends**
- **UltraNet as a Performance Solution**
- **UltraNet File Performance Data**

Emerging Client-Server Model



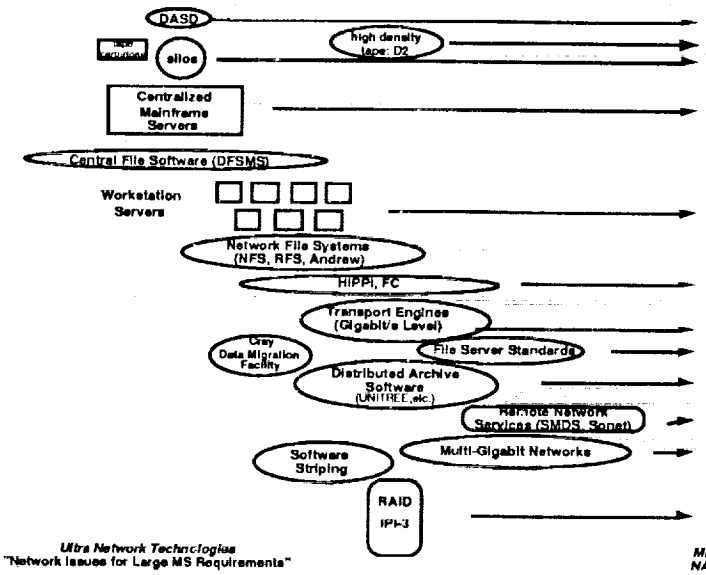
File Server Uses for Sci & Eng

- **Central File Storage**
  - Move files between Storage Hierarchies & Compute Servers
  - Support DISKLESS nodes (remote record reads)
- **Temporary or Workspace Storage**
  - Distributed Processing (provide common buffers for large data sets)
  - Image/Graphics display (digital VCR)
  - Network cache to match high speed systems to lower speed
  - Wide Area communication buffering (similar to cache)
- **Archiving to Reliable Storage Medium**
  - Very large but frequent used files
  - All files for reliable long term storage

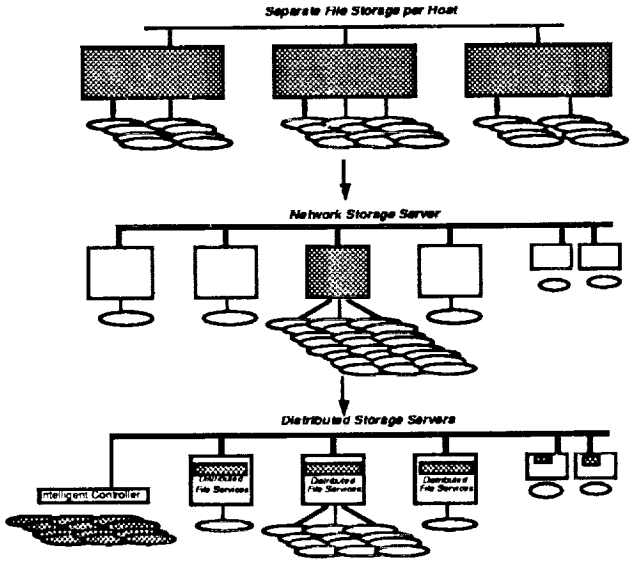
Networked File Server Requirements

- **File Management**
  - archive management
  - spooling between hierarchies
  - catalog management/file scheduling
- **Disk Performance Technologies**
  - parallel disk head technology
  - striped disks (software)
  - striped disk controllers (hardware)
  - striped file servers
- **Network Performance**
  - high effective Throughput (pt to pt)
  - low latency for transaction oriented applications
  - connectivity to highest performance channels/busses
  - standard protocols for heterogenous systems
  - high aggregate bandwidth

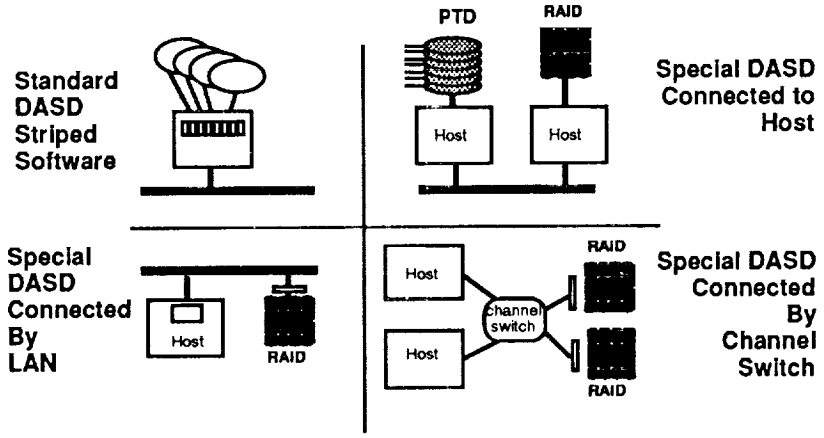
File Server Systems - Evolution not Revolution



Trends in File Servers



High Performance Disk Network Connection Strategies



Distributed File Server Trends

Factors Which Will Affect Cost/Performance Trends

- Smaller CPU's with medium I/O capabilities can control Distributed File Systems
- Transport Based Protocol Engines can provide reliable transport for network storage devices
- Standards (ala HIPPI, FC, !PI-3) create more competitive marketplace for devices
- Standards (ala IEEE MS) create more competitive marketplace for software
- Technology advancements continue in improving cost/performance of devices

Major Issues for File Transfer

- Files are getting substantially larger
  - files today range from small to several GIGAbytes in size
  - slow LAN performance limits feasibility of some file transfers
  - LAN is congested during transfers
  - MTBF of hosts/disks/LAN can be less than file xfer time
- LAN utility is determined by EFFECTIVE performance
  - EFFECTIVE Performance 7 - 100 times slower than "wire" speed
- File transfer impacts valuable host resources
  - As network "wires" get faster, it takes more CPU to be efficient
- Current system structures sized for small transfers
  - slow LAN's
  - slow effective disk transfer rates
  - application I/O buffers small
  - system buffers small & require copies

File Server Performance

Clients
- OS latencies
- Access Methods
- Network Software
- Application
- CPU Speeds
- I/O Speeds
- Device Speeds

Networks
- End to End Performance
- Aggregate Performance
- Connectivity Choices
- Interoperability
- Distance Latencies
- Host Integration

Servers
- Application
- Server Algorithms
- System Algorithms
- Network Software
- CPU Speeds
- I/O Speeds

Devices
- channel speeds
- device speeds
- device latencies
- allocation methods
- block sizes
- controller intelligence
- controller memory
- controller speeds
- controller protocols

The Performance Chain

The User Sees One Level of Performance and Reliability:

Influenced by the slowest and most unreliable in the chain

Network Performance Myths

You'd Have A Much Higher Performance Network If You Only Had:

- **Fiber Optics**
- **Switches Instead Of Busses**
- **A Lighter Weight Network Protocol**
- **A Faster Channel, ala HIPPI or ESCON**
- **Multiple Simultaneous Data Paths**
- **A New Computer**
- **Wave Division Multiplexing**
- **A Faster Disk System**

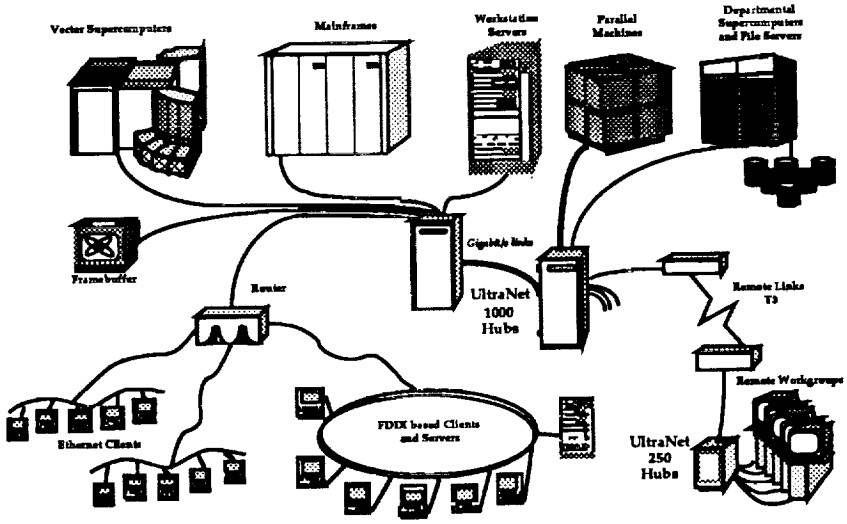
Gigabit/s Network Issues

- System Issues Dominate Performance Not Fabric
- Network Problems Dominated By Large Speed Range
- Applications Determine Realized Performance
- Higher Speeds Uncover Many Vendor/System "limits"
- Integration With Existing Network Technologies

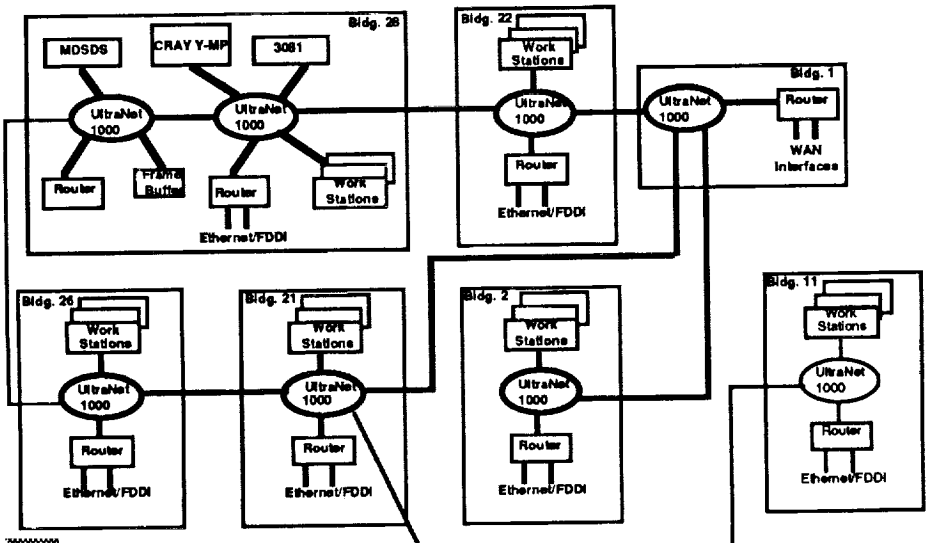
UltraNet as a System Solution

- UltraNet Is Transport Level Service To Host
- Data Delivered Directly to User Buffer From Channel
- Protocol Processing In Adapter - Reduces Host CPU cycles
- Decouples Host Transaction Sizes From Network Packet Sizes
- Uses Large Packet Sizes When Between UltraNet Connected Servers
- Uses Standard Packet Sizes To Other Networks
- Fully Participates In Internetting Environment (RIP, ARP, SNMP)

UltraNet Topology



Server Network Concept
NASA Goddard



UltraNet as Server Network

- Front End

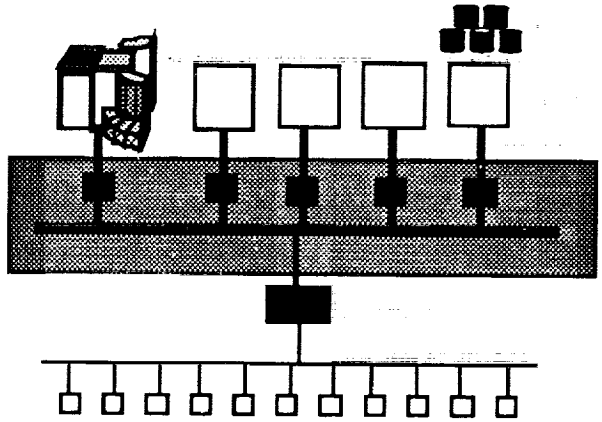
Eliminates need for FDDI Adapters directly on Servers

- Server Network

Large Aggregate and pt-pt xfer rates for direct connected servers

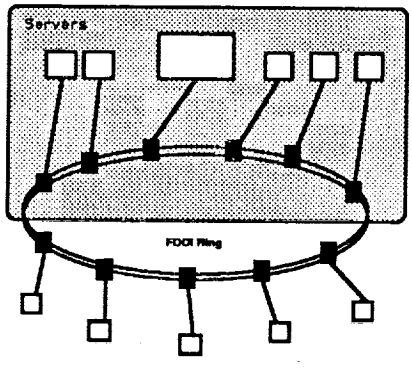
- Backbone

Connect multiple buildings & other networks at gigabit/s rates

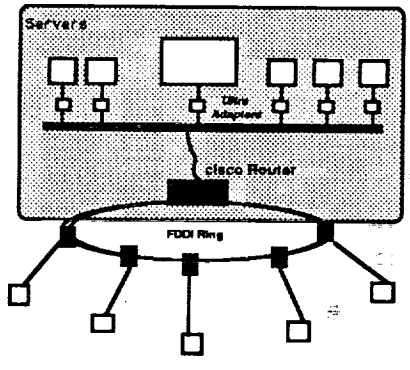


Server Network Alternatives

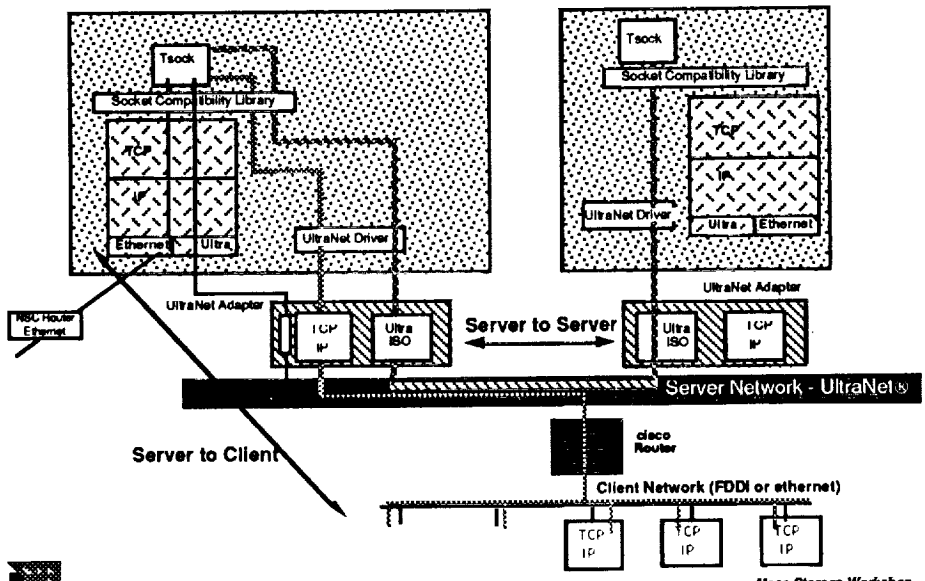
FDDI Only



UltraNet and FDDI

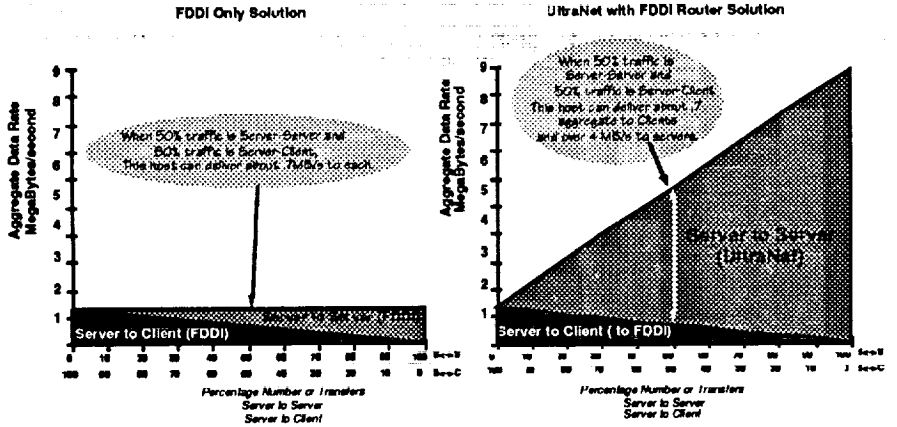


UltraNet as Server Network- the Path



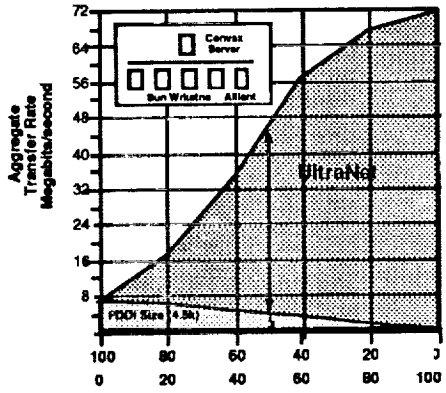
UltraNet Server Model

Application to Application Performance Simply Modeled
From One Server to Multiple Server/Clients



Above data computed based on typical host transaction rates, network transfer rates, and network packet sizes.

UltraNet Server Test Results



Test Configuration

Tests Performance at Ultra Factory using TSOCK application.
Memory to Memory Xfer rates shown.

Computers used:

Sun 4, Sun 3
Sun OS 4.1.1

Convex C1
Convex OS 8.0.0.1

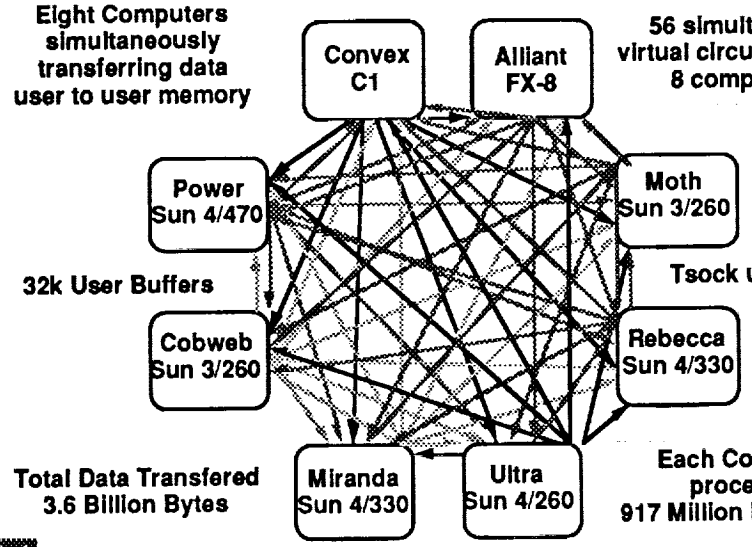
Alliant FX-8 Concentrix

UltraNet 3.50-03 software

FDDI was not tested - FDDI host transaction size used to simulate the UltraNet performance from an Internet source.

UltraNet: Bandwidth Test

Eight Computers simultaneously transferring data user to user memory



56 simultaneous virtual circuits among 8 computers

32k User Buffers

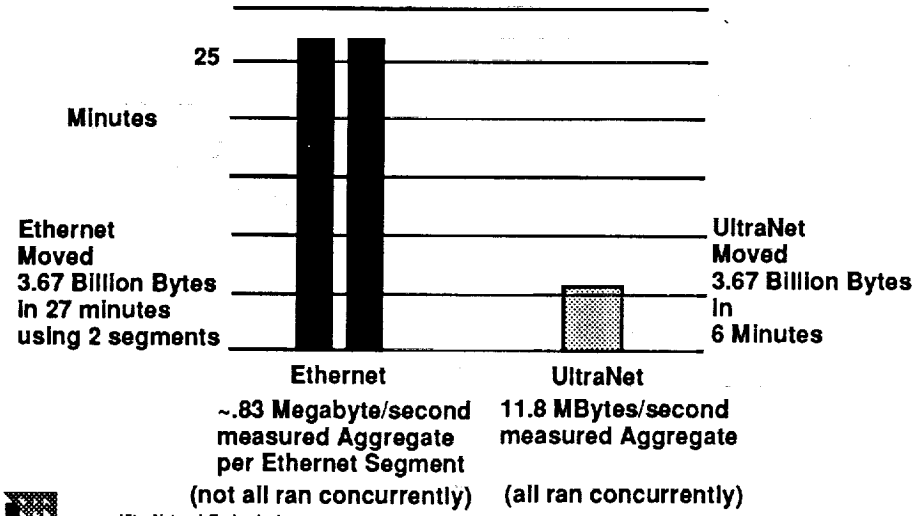
Tsock used for test.

Total Data Transfered 3.6 Billion Bytes

Each Computer processed 917 Million Bytes each.

Bandwidth Stress Test Comparison

Eight Computers each with 7 full duplex conversations



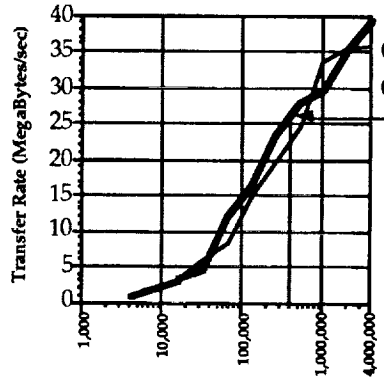
Cray File Server
Networking using HIPPI Interface

Data taken
at
University of
Stuttgart

Unicos 6.0

July 17, 1991

Data acquired
during production
time



Cray 2 to YMP2E
Cray YMP2E to Cray 2

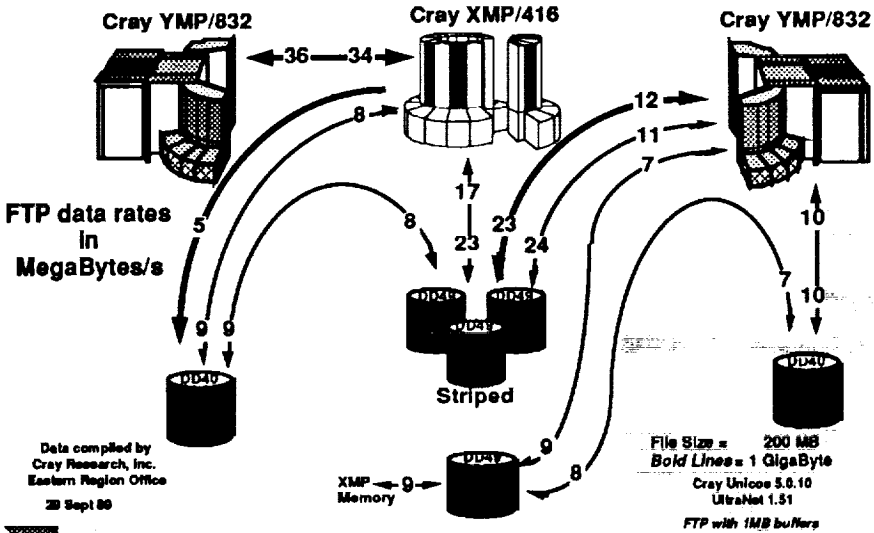
Data xfer
rate
up to
25 MBytes/s
possible

with UltraNet
using
128k buffers
&
3 way
Striped Disks

Cray 2 connected to Ultra HSXnp
Cray YMP2E connected to Ultra HIPPInp

TSOCK Test program
user buffer to user buffer

UltraNet Performance - FTP Rates Between Two Crays

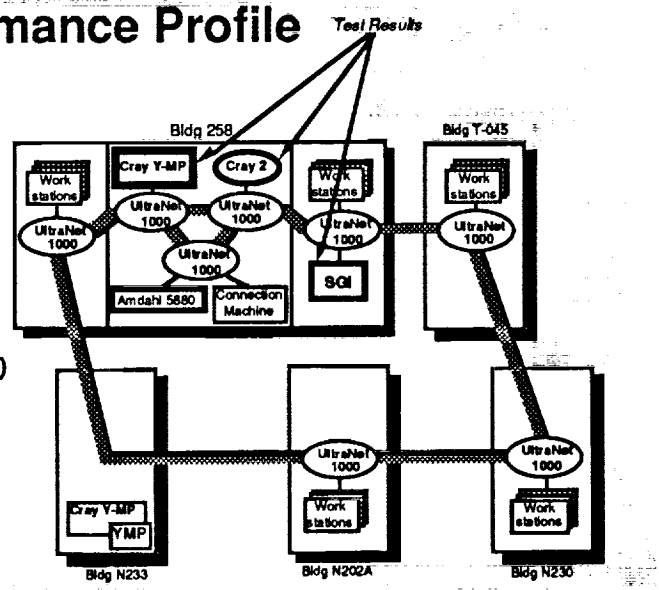


NASA Ames Research Center Performance Profile

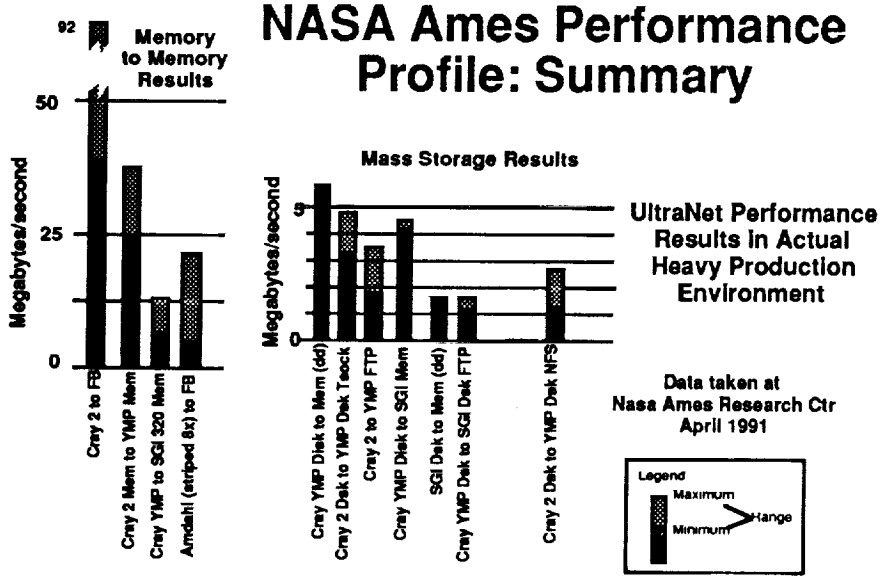
Cray 2
Unicos 5.1.11

Cray YMP
Unicos 5.1.11

SGI 4D/320 VGX
(with Powerchannel)
Irix 3.3.1



NASA Ames Performance Profile: Summary



File Server FTP Performance

	Typical FTP results Ethernet		UltraNet		*BENEFIT/Server* Single	Aggregate
Super <--> Wks	.25		.75		3.0 X	
aggregate		.35		30.0		85 X
Super <--> Mainf	.26		.60		2.5 X	
aggregate		.40		4 - 20		10-50 X
Mainf <--> Wks	.25		.60		2.5 X	
aggregate		.40		4 - 20		10-50 X
Wks <--> Wks	.25		.75		3.0 X	
aggregate		.35		4.0		11 X

Significantly More Users Can Be Supported with the Same Computing Resources for File Transfer Operations Using a Faster Network

UltraNet as File Server Transport

- Provides Highest Performance TRANSPORT LEVEL connection available 2 - 40 MBytes/second range for host to host transfers;
- Matches throughput of high performance emerging disk devices, i.e. RAID, vendor striped disks
- Supports standard SOCKET based Applications at increased speeds for FTP, rcp, rdump, user written applications
- Supports host based NFS access - improves network wide bandwidth for large NFS internets
- UNITREE application supports UltraNet for Distributed File and Archive Server Applications
- Other Applications in Test for Network Backup over UltraNet
- Supports several vendor based File Server Solutions:

  Cray Superserver; Convex, Alliant, IBM HMS, FPS