Fibre channel Storage Area Network Design for an Acoustic Camera System with 1.6 Gbits/s Bandwidth

Zhang Hong, Koay Teong Beng, Venugopalan Pallayil, Zhang Yilu, John R Potter

Acoustic Research Laboratory, Tropical Marine Science Institute, National University of Singapore, 10 Kent Ridge Crescent, Singapore, 119260

Lawrence Wong Wai Choong

Department of Electrical and Computer Engineering, National University of Singapore,

10 Kent Ridge Crescent, Singapore, 119260

Fax: (65)8748325

Email:engp9774@nus.edu.sg http://www.arl.nus.edu.sg

Topic area: information technology; computer networking

Abstract-- **Driven by Gigabit-rate Fibre Channel (FC) networking capability, storage area networks (SAN) have emerged as a high performance interface that can interconnect multiple servers and storage devices with high bandwidth, high availability and high fault tolerance. Most FC SAN designs employ fabric switches to retime and manage traffic. This is a relatively expensive approach but provides the robustness demanded by workstation network users. We present the application of FC-AL (Arbitrated Loop)-based SAN technology to a novel problem of constructing a data management system for a broadband acoustic camera demanding 1.6 Gbit/s data flow rate. The needs of such a dedicated SAN system are quite different from that of a generic computing network. Our SAN consists of Intel embedded processors running Windows NT4.0 and Seagate FC hard drives. In order to improve the end user throughput, the mechanisms of FC I/O systems are examined and major factors affecting FC-AL performance have been analysed. Furthermore, we have observed the application performance of FC-AL systems employing different types of host bus adapters (HBA) and CPUs. Finally, test results are presented and evaluated.**

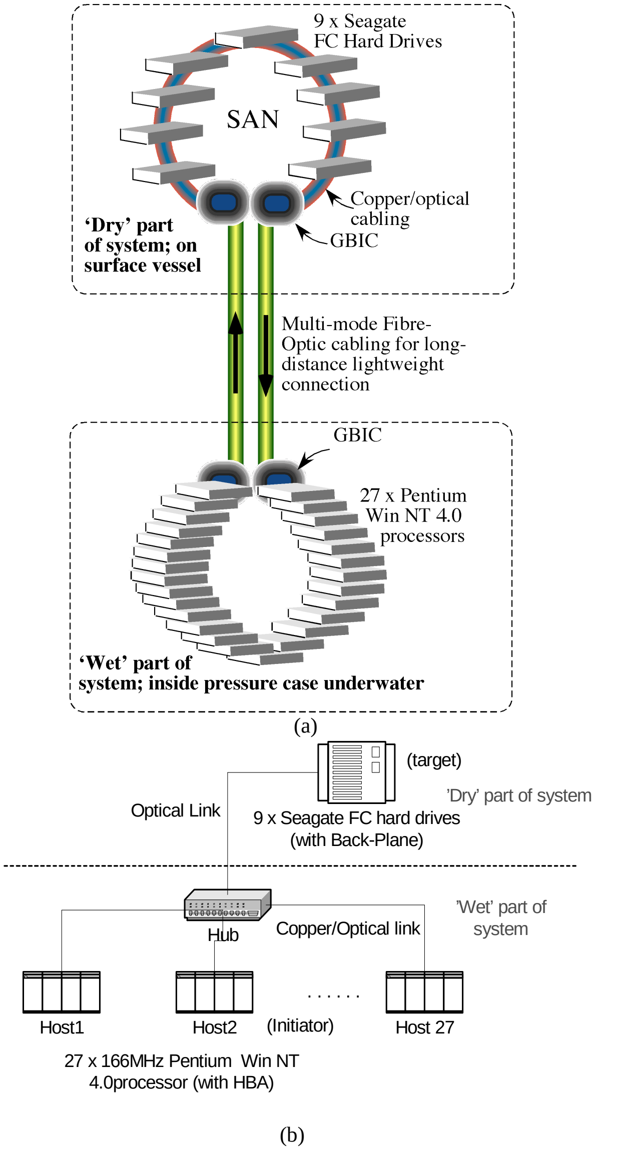
Keyword-- **Acoustic Imaging, Fibre Channel Arbitrated Loop, Storage Area Network, Device Driver.**

I. INTRODUCTION

As the topologies of computing environments evolve into a mix of self-contained workstations and distributed assets, they have led to the need for efficient, stable and fast Storage Area Networks (SAN). To support the high bandwidth required, a high-speed and flexible networking protocol is needed, and Fibre Channel (FC) is often chosen for its Gigabit/s (and increasing) capability. In addition, FC can be implemented on a variety of physical layer infrastructures and supports many higher-level protocols such as TCP/IP and SCSI. In this paper, we describe the application of FC SAN to a novel environment, a broadband acoustic camera.

Broadband acoustic cameras with a large number of individual sensors have not been constructed before due to the technical limitations on multiplexing and storing acquired data with sufficient bandwidth. Narrowband sonar systems demand only a small fraction of the bandwidth needed by broadband systems, and are in wide use. As the importance and power of broadband receivers has become clear in many acoustic applications, the computer networking capability necessary to handle the high bandwidth has also serendipitously emerged. Using FC to relax the bandwidth bottleneck, the Acoustic Research Laboratory (ARL) at the Tropical Marine Science Institute (TMSI) in the National University of Singapore (NUS) has designed and constructed the world's first broadband planar acoustic camera, called the Remotely Operated Mobile Ambient Noise Imaging System (ROMANIS)[1].

The traditional approach of building a data acquisition system is to acquire data from many sensors, multiplex, then pass this dense serial stream to a storage system. A RAID array is normally required to stripe the data across many drives to achieve the throughput required. To retrieve the data, the striped drives must be accessed and the striped data re-assembled, followed by a de-multiplexing stage. This is inelegant and costly. With a FC-AL SAN, the data can in principle be acquired by many embedded processors, and stored on many hard drives on a super data highway, without the need for multiplexing, RAID striping, RAID reassembly or demultiplexing. ROMANIS has 54 Pentium processors, each running Windows NT 4.0, 508 sensors, and an array of 54 Seagate FC drives, arranged in 2 x FC-ALs. In each loop the processors are locally networked by a copper physical layer, as are the cluster of FC drives. These two sub-assembles are then connected by a Fibre Optic physical layer to achieve the loop topology. Planned enhancements include moving over to a fully-optical fabric layer and 4 x FC loops instead of 2. Fig 1 shows the schematic of the present FC-AL-based SAN system.



(a)

(b)

Fig.1. (a) Schematic of one of the FC-AL arrangements.
(b) Actual loops are physically constructed using FC hubs to connect individual nodes at the Pentium and FC disk array ends of the loop.

Theoretically, an FC-AL-based SAN has the ability to provide a raw communication bandwidth of up to 106.25MByte/s. However, in practice the achievable transfer rate is quite limited. As FC-AL-based SAN is a network system consisting of many hardware and software components, the way each component is implemented has a substantial impact on the system communication performance finally obtained by end user. The major cause of performance degradation from raw bandwidth to user level throughput is the interaction required between the host computer and the network interface [2]. The network subsystem related to I/O operation, called the I/O subsystem, involves components of three main areas: the hardware architecture of the host, the host software system, and the network interface. We will be looking at some of these components more closely in section 2, especially DMA and the device driver. In section 2, we compare Fibre Channel design alternatives for protocol function partitioning between hardware and the host software. Since each of our FC-AL-based SAN subsystems consists of 27 embedded Pentium processors and 9 FC hard drives, this creates quite a congested loop. Future enhancements plan up to 14 Pentium processors and 14 FC disks on each FC-AL. The change in aggregate performance of an Arbitrated Loop as the number of devices and congestion increases is also discussed in section 2. Based on these design issues, section 3 covers the configuration of our FC-AL-based SAN test bed. An evaluation and analysis of the performance of FC systems employing different types of host bus adapters and CPUs are presented. We close with a summary in section 4.

II. FACTORS AFFECTING OUR AL-FC-BASED SAN COMMUNICATION PERFORMANCE

A. I/O subsystem components

Generally speaking, the I/O subsystem includes components in three major areas: the host hardware architecture, the host software system, and the network interface [3].

The host hardware architecture includes: memory, system bus, I/O adapter, and I/O bus. Two major steps to be performed to transmit data from the host computer to disk destination are DMA (direct memory access) operation across a PCI bus and data transmission across a FC communication link.

DMA transfers all data from the acquisition device to memory with minimal host intervention and handshaking. A typical DMA phase contains delays of physically locking the memory pages of the user buffer, preparing the address list and moving data between main memory and the interface. DMA devices must be programmed using physical addresses. This is a "hidden" feature of DMA that can seriously impact performance, since the system has to assure that a large data block actually resides in contiguous physical memory, which may cause copying of the buffer in the kernel. One solution to the problem of insufficient contiguous physical memory is a technique called Scatter/Gather[4]. A DMA device that supports this technique scatters incoming data to a collection of physically disjoint memory blocks or gathers outgoing data from a similar collection. Mechanisms for implementation of Scatter/Gather vary. Depending upon the design of the host bus adapter, the Scatter/Gather table may exist on the host as a buffer available to the HBA's driver software, or it may exist in specialized memory on the HBA itself. This is the difference between software and hardware Scatter/Gather. Maximum sustained transfer rates of a system employing a software Scatter/Gather technique will be about 80MB/s while maximum transfer rate of a system using hardware Scatter/Gather can be 120MB/s. NT 4.0's enhanced Scatter/Gather supports handling very large SCSI I/O transfers. It supports up to 256 Scatter/Gather segments of 4096 bytes each, allowing transfers of up to 1,048,576 bytes.

The host software consists of the operating system, the application programming interface (API), higher level protocol processes, and the device driver for the network interface.

To protect user applications from accessing and/or modifying critical operating system data, Windows NT uses two processor access modes: user mode and kernel mode. User application code runs in user mode, whereas operating system code (such as device drivers) runs in kernel mode. The device driver for an I/O subsystem acts as a system agent that converts user-level commands to the network interface. It is not even possible for user-mode code to access I/O hardware without the aid of a kernel mode driver. The software operating system (Windows NT 4.0) communicates with the SCSI HBA through a layer of device drivers. Fig 2 shows the SCSI driver architecture of the Windows NT operating system. An application request for an I/O operation with HBA is routed to a Disk Class Driver. The Disk Class Driver handles aspects of the requested operation generic to all disk devices. The SCSI Port Driver handles aspects of the SCSI protocol specific to disk operation. The Disk Class Driver is multi-threaded while the SCSI Port Driver is single-threaded, which is the main performance inefficiency that enhanced drivers eliminate. Both of the Disk Class and SCSI Port Drivers are part of the Microsoft supplied operating system. The Miniport Driver communicates with the firmware in HBA which, in turn, communicates with the physical disks. The Miniport Driver is device specific, which is generally supplied by the manufacturer as a companion to the physical device.

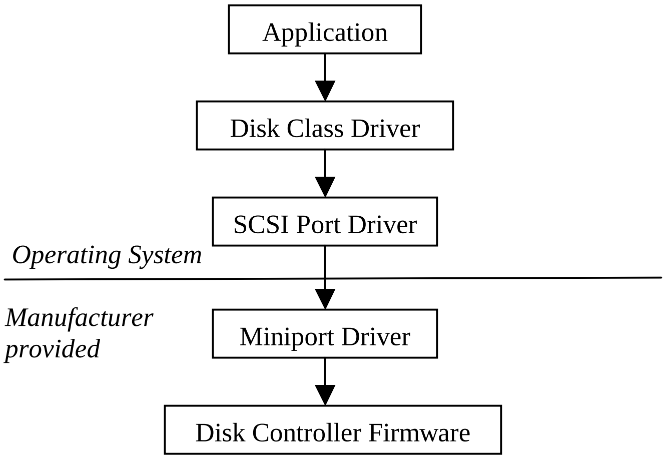


Fig 2: Typical SCSI driver architecture

A significant consideration in designing a Device Driver is how the application needs to access the device. High-level composite operations can minimise the user-to-kernel-to-user transactions. Optimized handling of interruptions can reduce context switching. Carefully chosen buffering strategy, which facilitates direct I/O to the user space instead of using buffered kernel, can minimise the copying.

B. Protocol function partitioning between hardware and host software

Fibre channel protocol is defined as a multi-layered stack of functional levels. The layers of the Fibre Channel standard define the physical media and transmission rates, encoding scheme, framing protocol and flow control, common service, and the upper-level applications interfaces. Many designs will delegate some of the protocol functions to a microprocessor. This microprocessor may be physically part of the adapter itself. Or the design may rely on the host system's processor (typically as part of a host device driver). For example, a Fibre Channel host bus adapter may contain an on-board processor where the control function is implemented, or the control function may be on the host system's device driver. Different implementations partition the protocol functions between hardware and firmware in different degrees. Generally speaking, incorporating hardware functionality can enhance performance while firmware reduces costs at the expense of system performance[5].

C. Large number of devices in an Arbitrated Loop

One characteristic of FC-AL is that it allows connecting a relatively large number (up to 127) of devices in a single loop as well as maintaining low latency and high throughput. As the number of devices on the loop increases, the aggregate performance of the system will increase to a limit imposed by the host bus adapter and/or the FC-AL itself. At the same time, the performance of each device will decline as contention for the loop increases due to the increasing number of devices arbitrating for access. Even if these devices are idling and just receiving and re-transmitting data on the loop, according to the FC-AL standard, there is a transmission delay of up to the time required to transmit six transmission words through each device. This delay is normally insignificant compared to the latencies imposed by the disk drives themselves (i.e rotational and seek latencies). But for a large number of devices, this delay can become significant. In order to avoid using a very populated Arbitrated Loop and to divide the total required bandwidth of 1.6 Gbits/s into two manageable rates of 800 Mbits/s per loop, we have divided our 54 processors and 18 hard drives into two independent loops. Future enhancements plan to further divide these into a total of 4 loops, and to increase the number of FC Hard Drives, approximately maintaining the number of nodes per loop.

III. PERFORMANCE ANALYSIS OF FC SAN

We conducted a case study to compare the performance of different Fibre Channel controllers and host system CPUs in terms of CPU utilisation and data throughput. The performance was measured using IOmeter, a vendor-neutral tool developed by Intel. IOmeter runs under Windows NT. The NT performance monitor is used to gather CPU performance results. IOmeter measures operation rate (IOps) and CPU utilisation directly. The performance metrics we used are data transfer rate and CPU utilisation, in which the data transfer rate was computed by multiplying IOps and transfer size.

A. Test bed configuration

An FC-AL test bed including one host computer and up to 8 hard drives was built. The test configuration is shown in Fig.1. All the nodes (host computer and hard drives) were connected through an FC hub. The test bed equipment consisted of the following:

1) Pentium III 550MHz MMX, 256MB RAM, Intel 440BX dual processor chipset motherboard (running single processor), S3 SVGA display card. Operating system: Windows NT4.0.

2) Pentium 166MHz MMX, 32MB RAM, Intel VX chipset motherboard, S3 SVGA display card. Operating system: Windows NT 4.0.

3) Vixel 2000 Manageable Hub in unmanaged configuration

4) 8 x Seagate Cheetah Fibre Channel hard disks (ST39103FC)

5) HP HHBA5100B evaluation host bus adapter card (Tachyon TL chip)

6) QLogic QLA2200 evaluation host bus adapter card (ISP 2200A chip)

7) Iometer version 1998.10.08 Copyright Intel Co.

The criteria for selecting the test set up components were to explore at least two types of HBA chipset and two processor speeds and RAM configurations that would span the range of hosts from the embedded Pentium used in ROMANIS to a typical desktop PC, as available to us in the laboratory.

We set the IOmeter workload according to the real data stream of our FC-AL-based SAN system as follows:

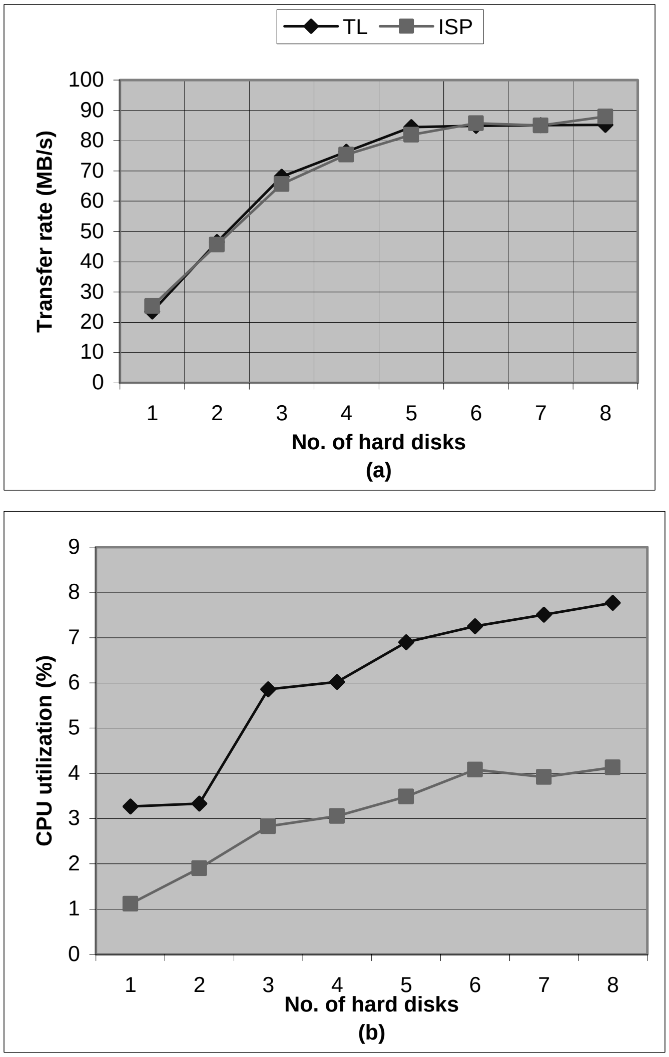
1) Data transfer size: 2MB
2) Number of transaction per connection: 1
3) 100% sequential write
4) Ramp up time 10 sec, run time 1 min.

During the design period of our SAN system, we tried two types of Fibre Channel HBA: HP HHBA5100B(TL) and Qlogic QLA2200(ISP). These two HBAs have different protocol function partitioning policies and device drivers. The testing results show the advantages and disadvantages of the two approaches.

Both TL and ISP support FC Class 2 and 3 services. The frame size of TL is up to 1024 byte, while ISP supports up to 2112-byte frame payload. A large frame payload indicates small overhead. TL has enhanced FCP hardware assists that support a single command from the host to fully execute all phases of an FCP transaction. TL has no on board processors. The design focus of ISP is on minimising CPU interaction, maximising I/O throughput and improving host and FC utilisation. ISP incorporates an onboard high-speed enhanced RISC processor; a FC protocol manager (FPM); integrated frame buffer memory; and a host bus, three-channel, and bus master DMA controller. The FPM and DMA controller operate independently and concurrently under the control of the onboard RISC processor for maximum system performance. ISP firmware provides two interfaces to the host system: the command interface and the Fibre Channel transport interface. The single-threaded command interface facilitates debugging, configuration, and recovering errors. The multi-threaded transport interface maximises use of the FC and host buses. Apart from the Miniport driver, ISP also provides an enhanced adjunct driver: QLdirect, which intercepts all communications from the disk class driver ordinarily handled by the SCSI port driver, then handles them directly in a multi-threaded fashion.

B. Test results analysis and conclusion

In our Arbitrated Loop test bed, we connected one host computer to up to 8 FC hard disks. We observed the performance using TL and ISP FC chips on 550MHz and 166MHz host computers.



(a)

(b)

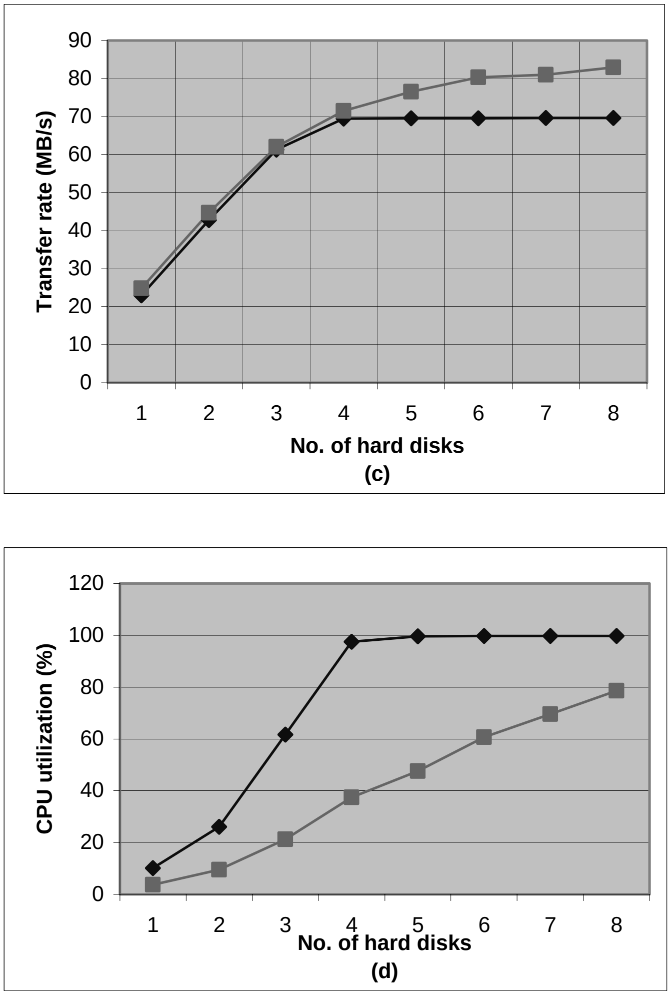


Fig. 3. Transfer rates and CPU usage for TL and ISP chips:
(a) P550 TL & ISP transfer rate vs. number of hard disks;
(b) P550 TL & ISP CPU utilisation vs. number of hard disks;
(c) P166 TL & ISP transfer rate vs. number of hard disks:
(d) P166 TL & ISP CPU utilisation vs. number of hard disks.

From the results in Fig.3 (a) we can see that the transfer rates of the P550 using TL and ISP are nearly equal, rising almost linearly with the number of disks until at least 5 hard disks are attached. We interpret this as an indication that the hard disk throughput is limiting the total bandwidth of the FC-AL until the load is spread over at least 5 disks.

The CPU utilisation is the fraction of the processing power of the CPU required to service the target I/O activity. Fig.3 (b) shows that the CPU utilisation for the P550 is consistently lower for the ISP than for the TL chip, because the TL chip relies more than the ISP on the host computer to control the I/O transfer.

Fig.3 (c) shows a similar throughput performance for the P166 as for the P550 for less than 5 hard disks, so that for 4 disks or less the transfer rate is essentially unaffected by different types of FC chips or host computer. This is consistent with the total throughput being limited by the individual disk sustained write rates. Fig.3 (c) also shows that using the slower P166 the transfer rate using ISP is about 20% higher than using TL when writing to more than 5 hard disks. If the host computer is not powerful enough, once a sufficient number of disks are attached to spread the load, the total system performance becomes limited by the host computer’s power. Using the ISP, the P166 achieves similar transfer rates to the P550 for any number of hard disks, since the ISP completes almost all the control functions by itself and needs minimum interaction with the host computer.

Fig.3 (d) shows that the CPU usage for the P166 is always much higher than for the P550 for both TL and ISP (note the very different scales between Fig.3 (b) and Fig.3 (d)), and that the limit to throughput for the P166 for more than 4 disks shown in Fig.3 (c) is reached because the CPU is 100% occupied handling the transfer protocols. From Fig.3 (b) and 3 (d), we can see that using the ISP, both the P550 and P166 have CPU much lower utilisation rates than for the TL chip. The low CPU utilisation of the ISP is due to its on board RISC taking over much of the control function from the host CPU.

The transfer rate is an important performance benchmark, but it cannot give a complete picture of system performance without considering CPU utilisation. A high transfer rate at the expense of high CPU utilisation means that overall system performance is degraded if the CPU is required to support other functions.

In our FC-AL-based SAN system, we employ 166MHz embedded Pentium processors. The host processor must also perform some pre-processing on received data, such as an FFT to reduce the bandwidth traveled on FC links as well as to off-load subsequent computational requirements at the beamformer. Based on the test results, the ISP FC controller, which provides both higher transfer rates and lower CPU utilisation, is certainly more suitable than using the TL chip because of its on-board RISC processor and enhanced device driver, particularly in our case where a low power embedded CPU is used.

IV. SUMMARY

Applying FC-AL-based SAN technology to a data communication link in a broadband acoustic camera is a novel application. We have presented our SAN system architecture and investigated the impact of CPU processing power, FC chip type and number of target disks on both total throughput and CPU utilization. Several factors impacting the user level throughput were analysed, such as DMA transfer technology, device driver, FC protocol function partitioning between hardware and software, and the large number of devices on a single Arbitrated Loop. The test results of the FC-AL system employing different FC HBAs and host CPUs have been evaluated. We have found that assigning more FC control functions to the HBA on board microprocessor to minimise host computer intervention can significantly improve data transfer rate as well as decreasing CPU utilisation, thus providing more efficient system performance for slower host CPU’s. Other design issues of I/O subsystems such as the bus architecture and the software system (data moving and copying policy) need to be studied further to take more efficient advantage of the FC high-speed network.

REFERENCES

[1] P.Venugopalan, P.Deshpande, S.Badiu, S.Constantin, B.Lu & J.R Potter. A 1.6 Gigabit/Second, 25-85 kHz Acoustic Imaging Array-novel Mechanical and Electronics Design Aspects, OCEANS '99 MTS/IEEE. Riding the Crest into the 21st Century , Volume: 1 , 1999 Page(s): 352 -358 vol.1

[2] Mengjou Lin, Jenwei Hsieh, and David H.D. Performance of High-Speed Network I/O Subsystems: Case Study of A Fibre Channel Network. Supercomputing '94 Proceedings, 174-183.

[3] K.K. Ramakrishnan. Performance Considerations in Designing Network Interfaces. IEEE Journal on Selected Areas in Communications, 11(2):203-219, February 1993.

[4] David A. Solomon. Inside Windows NT. Microsoft Press, 1998

[5] Robert W.Kembel and Horst L.Truestedt. Fibre Channel Arbitrated Loop, the Fibre Channel Consultant Series. Northwest Learning Associates, Inc. 2000

[6] Alan F. Benner. Fibre channel: Gigabit Communications and I/O for Computer Networks. McGraw-Hill, 1996

[7] Thomas M.Ruwart. Performance Characterisation of Large and Long Fibre Channel Arbitrated Loops. Mass Storage Systems, 1999, 16th IEEE Symposium on, 1999, 11-21.

[8] Edward N. Dekker and Joseph M. Newcomer. Developing Windows NT Device Drivers, a Programmer's Handbook. Addison Wesley Longman, Inc. 1999

[9] John R.Heath and Peter J.Yakutis. High-Speed Storage Area Networks Using a Fibre Channel Arbitrated Loop Interconnect. IEEE Network, March/April 2000, 51-56.