CDF Run II Analysis System
Fibre Channel

Stephan Lammel

- **Fibre Channel Introduction**
- **Channel versus Networks**
- **Fibre Channel**
- **Fibre Channel Topologies**
- **Mainstream Computing**
- **What do we want from FC?**
- **Filesystems**
- **Prototype System**
- **Work done so far**
- **Future Work**
- **Conclusions**

Fibre Channel Introduction

Fibre Channel is a general name for an integrated set of standards developed by the American National Standards Institute which defines new protocols for information exchange.

Fibre Channel development begun in 1988 (as extension of work on the Intelligent Peripheral Interface (IPI)) mainly to address limitations in existing interfaces.

It was approved as ANSI standard X3.230-1994 in October of 1994.

Channels versus Networks

There are two basic peripheral protocols for device communication: channels and networks.

Traditionally, the term channel indicates a peripheral I/O interface to a host computer that transports large amount of data between the host and the peripheral device. The processing overhead is kept to a minimum by handling data transfer in hardware, with little or no software involved (once an I/O operation begins).

Channels usually operate in a closed, structured, and predictable environment. The host systems retain all the knowledge of the devices attached.

On the other hand, the term network indicates an I/O interface that usually supports many small transactions with higher overheads, usually the result of software involvement in the flow of information.

Networks operate in an open, unstructured, and essentially unpredictable environment. The unpredictable nature requires more software support for verifying access permission, setting up sessions, and routing transactions to the correct software service.

Fibre Channel

Fibre Channel attempts to combine the best of the two opposing methods.

It supports both network protocols like
IP, IEEE-802, or ATM
and channel protocols like
SCSI, HIPPI, IPI-3, or SBCCS.

Fibre Channel uses serial media (to be easy and cheap to install).

Fibre Channel is divided into functional levels so that physical layers can be upgraded/replaced without changes to higher protocols, scalability.

Fibre Channel is asynchronous to allow maximum throughput even under heavy load.

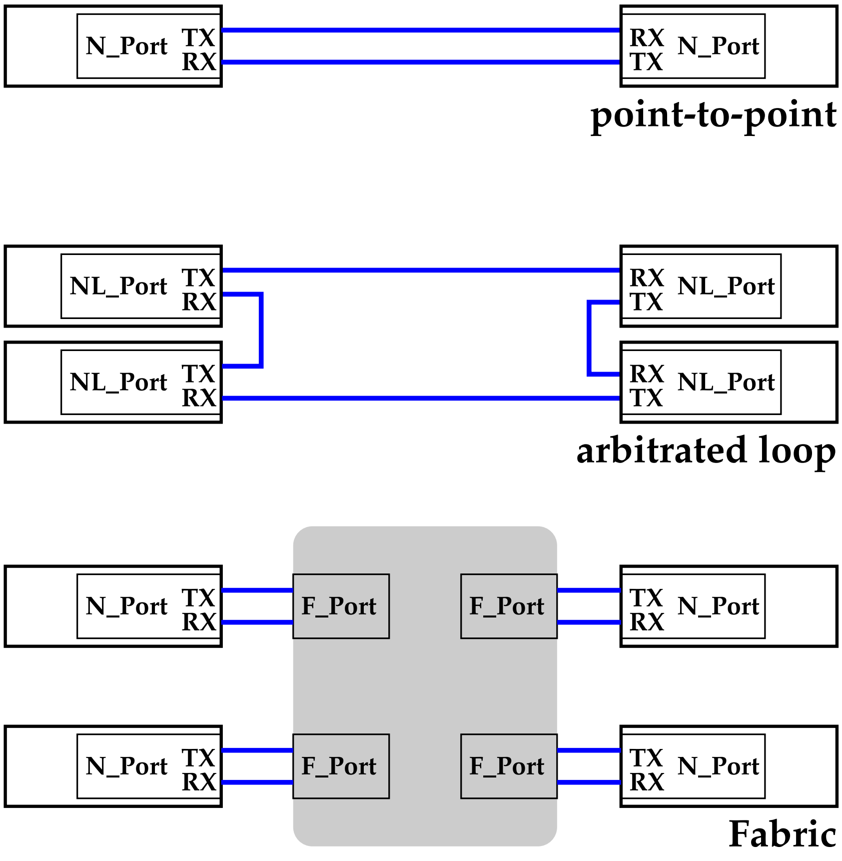
Fibre Channel can be switched (to get away from shared media).

Fibre Channel is not restricted to optical fiber but has standards for various media and speeds (twisted pair, coax, single-, multi-mode).

Fibre Channel Topologies

There are three different topologies:

- point-to-point,
- arbitrated loop,
- and fabrics.



Fibre Channel loops are limited to 125 devices, fabrics are limited by the address limit of 2^{24} (over 16 million).

Mainstream Computing

Since middle of 1998 the cost of a SCSI disk with a dual-ported Fibre Channel interface is the same as for a disk with 16-bit parallel SCSI interface.

On the other hand only a small fraction of all magnetic hard disks are SCSI disks.

In the past Seagate was the only disk manufacturer supporting Fibre Channel. This has changed and the last announcements of both Quantum and IBM include Fibre Channel interfaces.

An example: The latest filer model of Network Appliance is currently shipping with only Fibre Channel interface and disk. (Parallel SCSI is expected in a few weeks, though.)

Fibre Channel has no chance competing with Gigabit Ethernet in the network arena. However, it will most likely become the technology of choice for medium to high end storage connection.

What do we want from Fibre Channel?

Parallel SCSI is limited to 15 (or even 5/3) devices.

$\longrightarrow$ For 20 TBytes of disk space and 50 GByte disks this translates to 400 devices.

The length of a SCSI bus is very limited (25 to 1.5 meters). Every time the speed on the bus is increased, the maximum length is cut in half. LVD bumped us back to 25 meters in 1997 but with Ultra-2 we are already down to half of that length in 1998.

SCSI is a low overhead, efficient, flexible channel protocol. It works well and has broad support.

We want SCSI with a less limiting physical layer !

What do we get from Fibre Channel?

SCSI over Fibre Channel gives us exactly this:

SCSI with a less limiting physical layer !

Since we now get the network character for free, can we make use of it?

- to easier reconfigure storage space (devices) between systems
- to share storage in a homogeneous environment
- to share read-only storage in a heterogeneous environment
- to share read-write storage in a heterogeneous environment

Filesystems

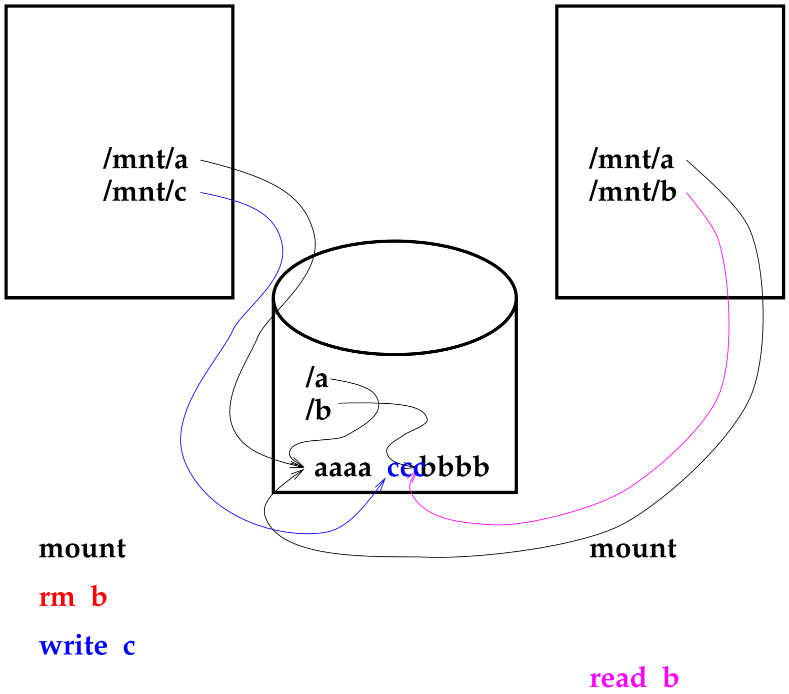
Sharing storage devices requires coordination beyond multiple initiator support on the SCSI interface. (This allows multiple hosts to communicate with the device at the same time. The hosts still arbitrate for the bus but can have outstanding commands.)

The disk space is used through a filesystem interface and rarely directly. Each UNIX flavour has its own filesystem format, efs and xfs for IRIX, ext2 for Linux, ufs for Solaris.

The OSes cache part of the filesystems in memory and combine individual read/writes into larger chunks for efficient disk access.

All of the four native UNIX filesystems above assume that they are the only user of the disk (or partition/slice), i.e. don't have a concept that their cache could expire or become invalid.

Filesystems



Another issue is that each UNIX flavour does not support the filesystem format of the other UNIX flavours. However, there are filesystem formats understood by all flavours. Those are normally connected with removable media:

- floppy DOS,
- CD-Rom ISO-9660/High-Sierra, and soon
- DVD-RAM (Panasonic) or DVD+RW (Sony, et al.) UDF.

Prototype System

We have acquired some small Fibre Channel equipment to learn about and understand the technology. (There is almost no FC expertise or even experience in Fermilab. HPSS used IP over quarter speed FC for a while.)

We have currently

- a few host bus adapters from Emulex, LP7000E,
- a dual host bus adapter from SGI,
- 2 Seagate Cheetah-18 disks in
- an 8-bay chassis from Trim,
- one FC/SCSI RAID controller from Adaptec, AEC-7412B,
- one 10-port hub from Emulex, LP5000,
- multimode and copper cables.

The equipment arrived during the last three month. The host bus adaper for the Sun workstation was installed October 20th. The SGI O2000 with the FC XIO board was setup by SGI November 10th and IRIX fully installed on November 22nd (the weekend before Thanksgiving).

Work done so far

We started with a simple one host and one device loop on Sun/Solaris. The device driver software is from Emulex and was downloaded from their ftp side. Installation was quick and straight forward. Default setting are good and additional configuration is easy.

The loop worked right away (we did not even switch transmit and receive). The disk (the RAID controller) was visible and showed up as standard SCSI disk.

We then did partitioning/formating and stress/performance tests. Things worked very well, no problems were encountered.

The Adaptec AEC-7412B runs a little slow, 10 MB/s write and 17 MB/s read, fare below our expectation. We contacted Adaptec and got new firmware for the controller. It took us a while to succeed with the download (poor support from Adaptec). The upgraded controller performs a bit better at small block sizes but levels out at the same 10 and 17 MB/s.

Work done so far

Next we added the chassis with the two Seagate disk onto the loop. It took us a while to setup the chassis correctly. Keep in mind that each FC chassis contains a little hub. We only got every other side of the manual and did not realize this right away.

We then added the hub into the loop and put each device on a separate port. Things continued to work flawlessly.

When the Sun MP came back from FCC (was used there for the EMASS acceptance test) we put a FC host bus adapter into the machine and connected it to the hub, i.e. same loop as the workstation.

We stress tested the the loop with both systems accessing different devices and same device. Formating partition one from one machine and partition two on the same physical disk from the other machine. No problems were encountered.

Work done so far

The Thanksgiving week we added the SGI to the loop. The SGI reported a lot of ”Notices” about ”unsolicited TCBs” during the boot. All disks showed up fine and neither of the Suns complained.

In discussions with MTI we learned that the SGI FC host bus adapter is not considered to be the greatest. MTI recommended using FC cards from Prisa on SGI. We are in the process of acquiring one. (For the time being we have put the SGI onto the second loop.)

In the meantime Walsh had prepared a utility to write ISO-9660 filesystems and a small test filesystem was written to one of the Seagate disks.

The Sun requires a Solaris-style label on the disk before it allows operation on a magnetic hard disk (it took us half an hour to realize this). After labeling the Sun mounts the ISO-9660 filesystem fine and the data in the filesystem are accessible (as on any other filesystem).

```
Wed Nov 25 21:39:14 CST 1998 0 successfull mount/umounts
 :
Fri Nov 27 11:58:24 CST 1998 1716600 successfull mount/umount
```

Work done so far

When we attempted to mount the ISO-9660 filesystem on the SGI we run into a problem: IRIX aparently makes a SCSI Inquiry and checks the product id against a database of known CD-Rom devices. Of course, the Seagate hard disks are not in the CD-Rom list. We discovered this the afternoon before Thanksgiving.

We have the IRIX source code to investigate the cause and look up how to overcome/add the missing entry. Unfortunately Randy (our IRIX guru) is in Java class this week.

So we bounced the issue to the local SGI system engineering specialist yesterday morning. He contacted the appropriate person at SGI headquater and told us in the afternoon that the check was put in to handle CD-Roms of different speed. He plans to get back to us today regarding what would need to be done and the RFE, etc.

Future Work

We need to resolve the ISO-9660 mount problem on IRIX. I assume that we will have at least a patch on a week timescale.

We will contact Sun and ask for hard disk access without having a valid Solaris disk label. (This is in case some other OS requires to have its label on a hard disk in the future.)

We have ask SGI to “ellaborate” on the ”unsolicited TCB” notices. Once we have the FC host bus adapter from Prisa we want to do some more compatibility stress tests. Right now, SGI and Sun, can coexist on the same loop without loop errors but those annoying messages.

The RAID controller from Adaptec did not perform satisfactory. We have researched and found two other FC/SCSI RAID controllers (in same price range). Both have vendor measured throughput of 50 MBytes/sec. (We did no succeed to get a sustained throughput number out of Adaptec.) One controller is on order, for the other we are waiting for a quote.

We would like to expand the prototype setup a bit more.

Operational experience is always a plus. The prototype analysis system, once it has moved to FCC, is an ideal place.

A fabric topology would be the next step. All the three disk vendors are adding support for class-2 right now. The disks with the new QLogic chip set will be available in Q1.

First propritary heterogeneous read-write storage will be available in December. We have to decide how aggressive we want to be here. I would base the decision on the price...

Conclusions

Where do we stand with Fibre Channel:

- **as parallel SCSI replacement**
  Sun $\surd$
  SGI $\surd$

- **easier storage management**
  Sun $\surd$
  SGI $\neg$ (wrong FC host bus adapter)

- **shared homogeneous storage (read-only)**
  Sun/UFS $\surd$
  SGI/XFS $\neg$ (to do)

- **shared heterogeneous storage (read-only)**
  Sun/ISO-9660 $\surd$
  SGI/ISO-9660 $\neg$ (mount problems)

- **shared heterogeneous storage (read-write)**
  awaiting quote from STK for Data Direct Networks evaluation unit