CDF Analysis Hardware Review

2 December 1998

CDF Run II Analysis Hardware Overview

Elizabeth Buckley-Geer

- Introduction
- Design Goals
- Overall System design
  - Baseline design
  - Components of the system
    - Network File-server
    - CPU
    - Disk storage
    - Tape storage
  - Fall-back conservative design
  - The future
- Implications for the data handling software
- Connections to farms and online

Introduction - A little history

- The Run I analysis system was very successful

- We had a central VMS cluster (FNALD) with of order 20 nodes (4 multiprocessor machines) and a 28 CPU SGI Challenge. Each system had about 500 GBytes of directly attached disk space, access to the ~ 2 TB STK silo and about 20-30 operator serviced Exabyte 8mm tape drives. There was also a small IBM AIX system added towards the end.

- The focus of analysis moved from the VMS systems at the start of Run IA to the UNIX systems by the end of Run IB. The VMS systems were upgraded from the previous run and were stable and reliable systems. The SGI was a new system and took some time to shake down. Considerable effort from both CDF and Computing Division went into getting it into a stable state.

- The SGI system started life as a 4D/480 and was upgraded to a Challenge system to take advantage of new processors. It has been continually expanded since by replacing the CPU boards. We are always in the situation of having plugged in the last CPU board.

- Users tend to choose one or the other system for working

on, typically based on location of dataset, personal preference, other people working on the same analysis etc. This tended to limit the duplication of datasets on the two systems. One of the places where duplicate data is likely to occur is the staging space devoted to the STK silo which is separate between the two systems.

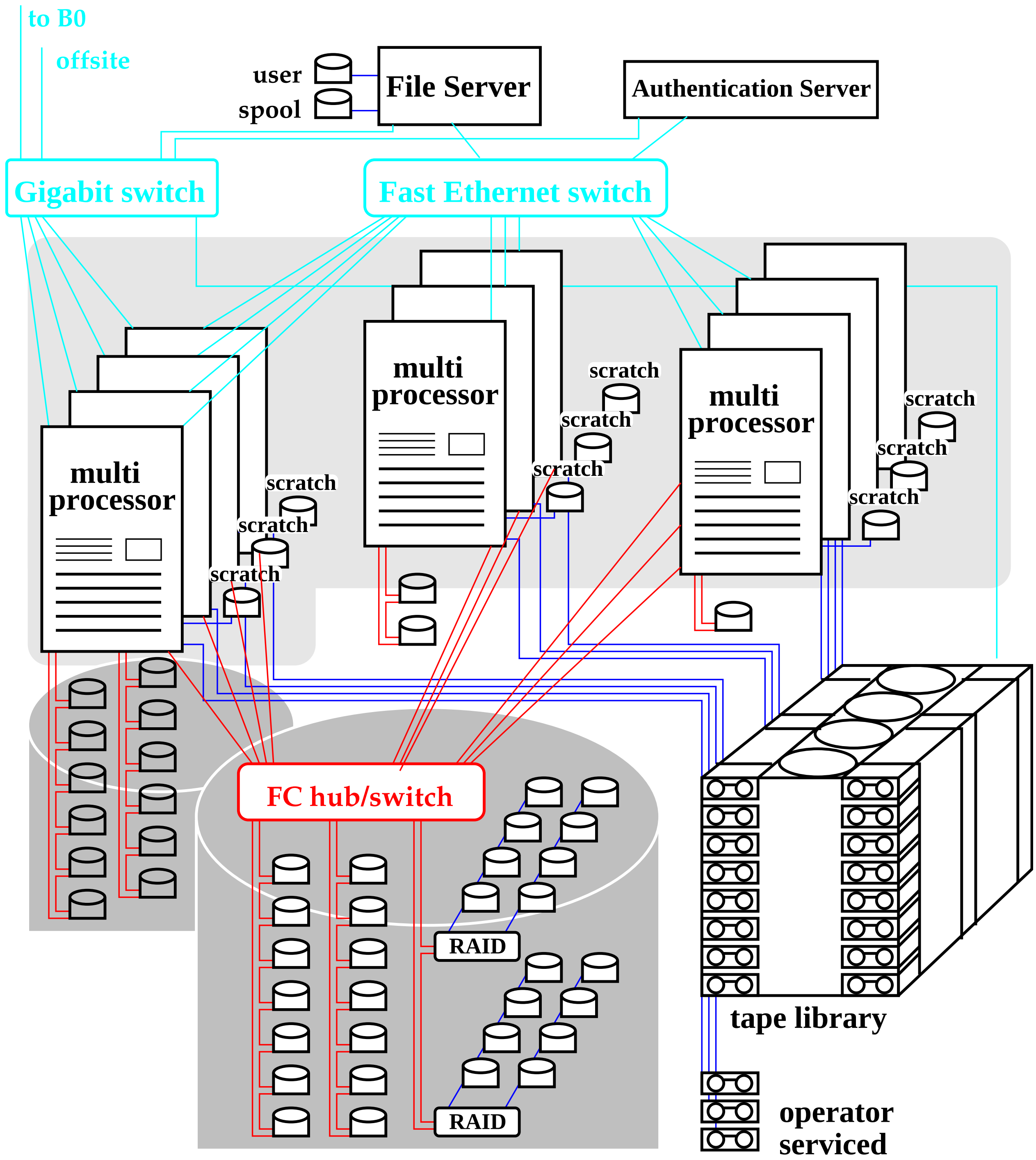
Design Goals

- There must be sufficient CPU available to do the physics we want to do but the system must be scalable beyond the current estimates for CPU at reasonable cost. We don't want to be in the situation where we are plugging in the last CPU board into a single large multiprocessor machine.
- There must be enough tape storage for all the data, raw data, reconstructed data, secondary, tertiary and user datasets. As much as possible should be in robotic storage.
- Buy as much data disk as possible (limited only by budget). We must store the data on tape but disk is what makes analysis possible. Means we have to filter out events of interest early in the chain so that the data fit on disk.
- Analysis jobs should only read and write data from disk and not deal with tape directly. Data should be spooled to disk first by an automated means which is transparent to the user.
- The system should be able to incorporate new technologies as they become available.
- The desktop systems in the CDF trailers should be better integrated than they have been in the past.
- Import/export of data should be easy.
- The system should be designed to minimize the operational support load.

Overall System design

- Commissioning the Run II analysis system will start in Fall 1999.
- We will present three scenarios
  - A baseline design which we believe is feasible by Summer 1999
  - A conservative fall-back position
  - The future that requires further developments in the area of Storage Area Networks (SANs), on the year 2000 timescale.

Baseline design



Components of the System

Network File Server and Authentication Server

An analysis system with standalone compute nodes is not practical from the user point of view. We have to be able to share files between systems. There are a number of ways we could imagine doing this

1. Use proprietary clustering software from the CPU vendor. Not desirable as we don't want to be tied to one vendor. AFS falls into this category also.

2. Use NFS.

   - Also in wide use at Fermilab.
   - Traditionally used one computer to NFS serve files to another. Performance not great either when used this way.
   - Also some reliability issues, especially with SGI
   - Can now buy dedicated NFS servers whose performance is typically a few MBytes/sec which is fine for interactive access. This is the direction we have chosen to evaluate. We purchased a small test system from Network Appliances. Stephan will present results from the measurements with this device. Such devices are in use at Jefferson Lab where they have the login areas ~ 1000 workstations served this way.

Authentication Server

We also need a way of serving passwords securely amongst the

machines in the cluster. This is a general problem at Fermilab that is being worked on and we will not discuss it further here.

CPU

- We estimate about 90,000 VUPS of CPU for Run II analysis (based on a scaling up from Run I - CDF4100). 1 VUP is the CPU power of 1 VAX-780 or about 1 MIP. 1 SpecInt95 is about x MIPs.
- This corresponds for example to a 128 processor SGI Origin 2000 machine. Such a system would provide for our needs but would suffer from the same problem of always plugging in the last CPU board - or replacing them to get higher capacity. It would also be expensive
- We expect moderate sized SMPs (workgroup server-class machines) to be more cost effective than large machines which are in the supercomputer league
- We would also expect Intel/NT based workgroup server to be at or beyond 8 CPUs by the year 2000 timeframe which will raise the price breakpoint for the other vendors from 4 to 8 CPUs machines.
- We will most likely start with of order a few larger machines as the price breakpoint is currently at 4 CPUS and that is too small a machine to consider. We would evolve towards something on the order of 10 machines over time as the price breakpoint changes.
- We expect to have a mixture of SGI/IRIX and SUN/Solaris (we proposed in January 1998 dropping IBM/AIX in favor of SUN). We would competitively bid these machines.

Given the experience of KTeV with DEC/Compaq UNIX we do not expect to include this OS in the system. Intel boxes would be bought as needed.

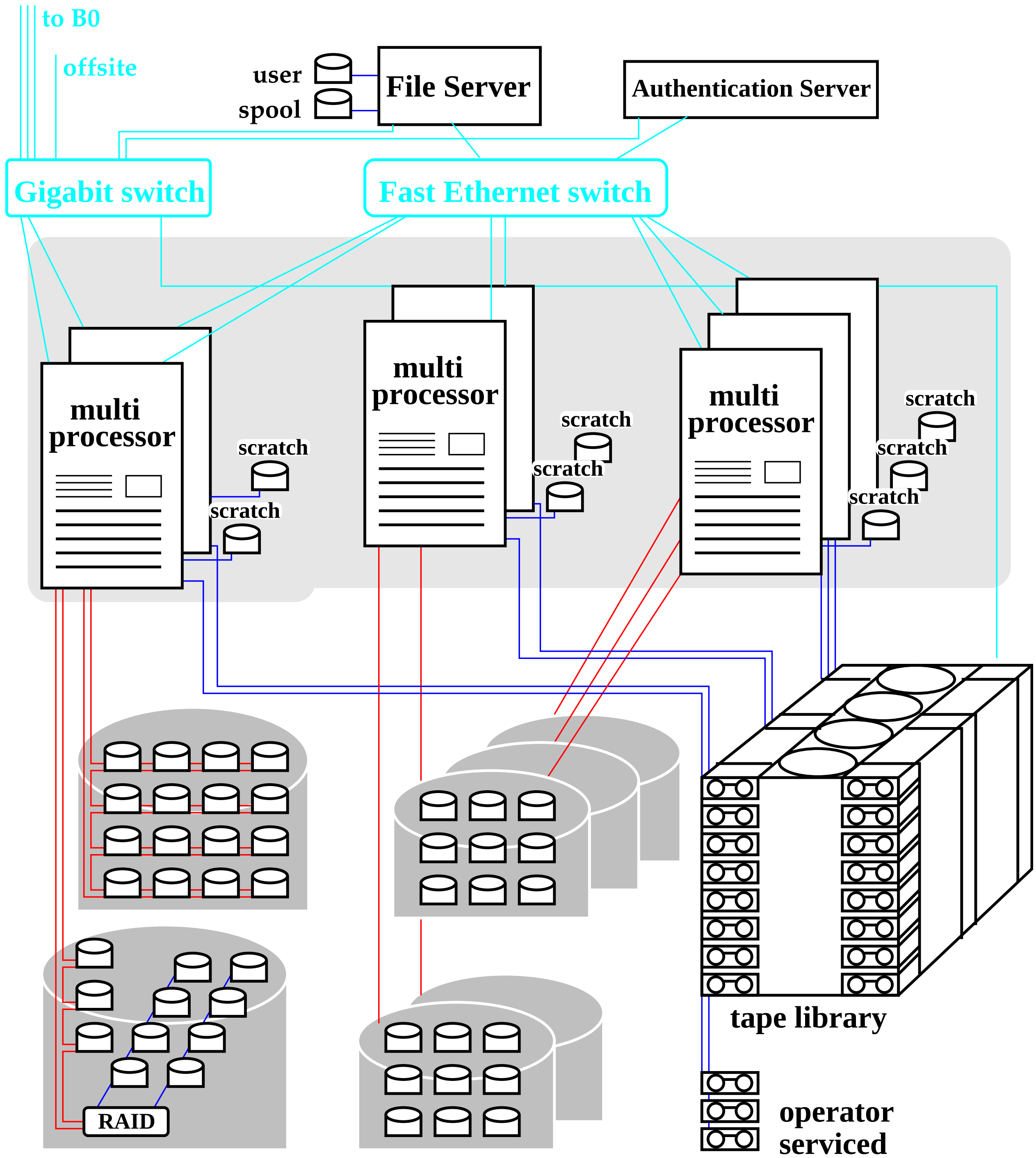
Disk Storage

- We estimated the amount of disk required to be 20-28 T-bytes (CDF4072/4100). The bandwidth to this disk was estimated at 1.5 GBytes/sec.
- Plan to use Fiber Channel to connect the disks to the computers to overcome the length limitations of SCSI (even a problem for low voltage differential).
- Fiber channel provides increased throughput and supports the SCSI protocol.
- SCSI and Fiber Channel disks are about the same price.
- Assuming a capacity of 75 GBytes in 2000 this means about 350 disks which means that the effect of disk failures must be minimized. RAID controllers are now available at moderate cost and we bought an Adaptec controller for tests. Stephan will talk more about this.
- If we want to share disks in a heterogeneous system then we need a filesystem that is understood by all computers. Stephan will talk about the options for this.
- In the baseline system the disk for tape staging in locally attached to each compute node. Similar analyses will congregate on certain machines. Will not be a problem in the initial system.

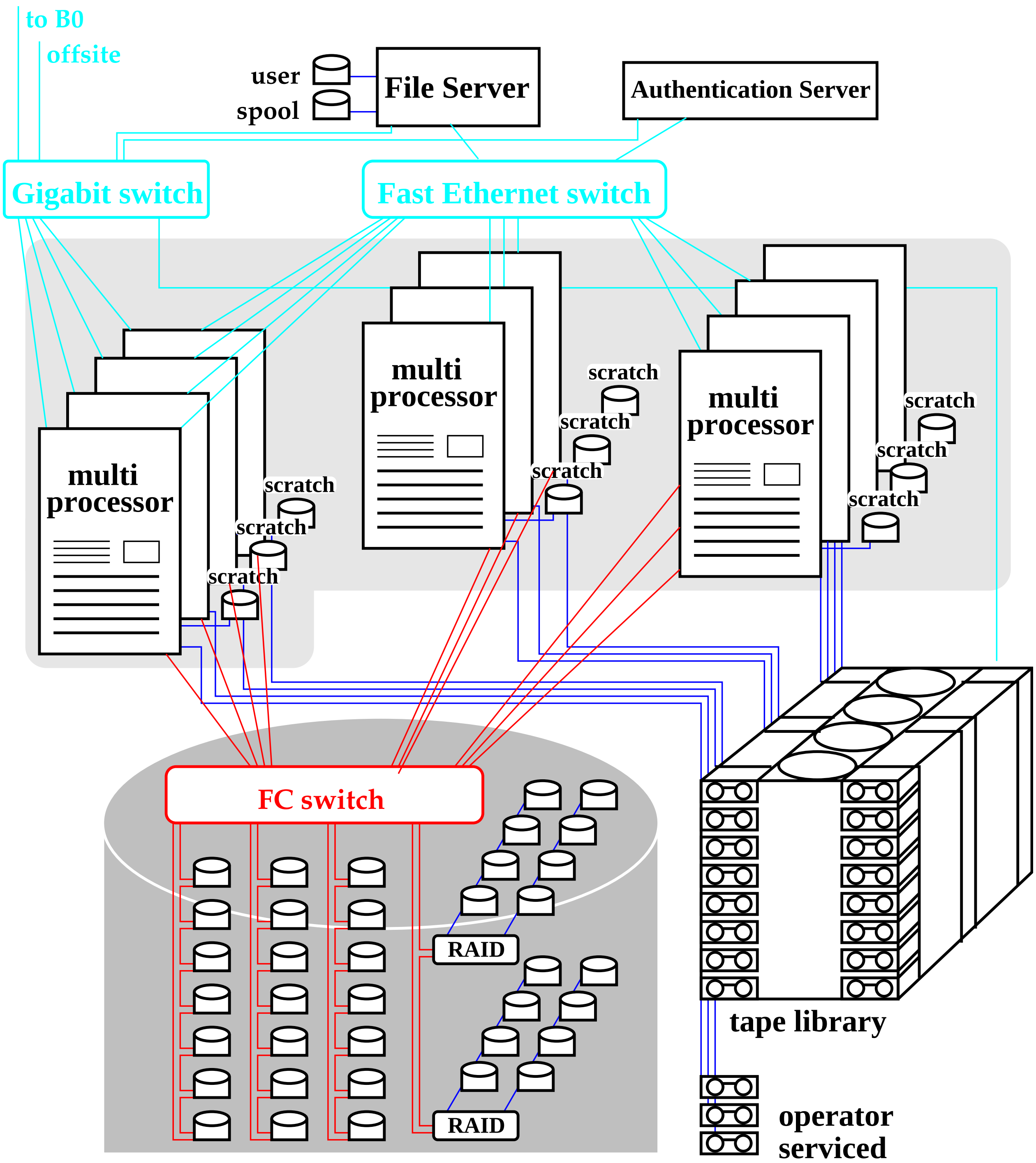
Tape Storage

- The recommendation for the Run II tape technology will be made by the Serial Media working group in March 1999.
- There are several promising commodity tape technologies that are under investigation which would minimize the cost of the tape subsystem.
- We will have a 4 tower EMASS tape library which can hold up to 256 drives. The library's tape capacity depends on the cartridge size, with 50 GBytes tapes it will hold 1 Petabyte.
- We hope to keep all of the data in the tape library and minimize the use of operator tape mounts.
- The tape drives will be directly connected via SCSI to the compute systems.

Fall-back conservative design



The future



to B0
offsite
user
spool
File Server
Authentication Server
Gigabit switch
Fast Ethernet switch
multi processor
multi processor
multi processor
scratch
scratch
scratch
scratch
scratch
scratch
scratch
scratch
scratch
FC switch
RAID
RAID
tape library
operator serviced

Implications for the data handling software

- The relevant parts of the software must be able to operate in a distributed environment, e.g The Disk Inventory manager must be a client server architecture, the CDF Data Catalog be accessible to all machines via a server.
- The batch system must be able to operate in a clustered environment - not a problem for currently available products under consideration.

Connections to farms and online