CDF/DOC/COMP_UPG/PUBLIC/4707

CDF Run II Data Handling Central Analysis System Hardware Architecture

Walsh Brown, Liz Buckley-Geer, Stephan Lammel, and Terry Watts

Jan. 23rd, 1999
version 2.7

This note describes the hardware architecture of the CDF central analysis system for Run II. The design builds upon successful strategies of Run I, eliminates shortcomings recognized in the Run I system, and integrates new computing technologies that have matured during the last years. The core of the analysis system is a pool of over 20 TBytes of data disks. Logically behind this disk pool is an automated tape library with a capacity of about 1 PByte. Shelf resident tapes space will be used in case of library overflow. The compute systems are located around this storage pool. A potentially heterogeneous cluster of moderate sized multi-processor systems will provide the necessary compute power of about 90,000 MIPS. Desktop systems in the B0 trailer complex can be part of the analysis cluster, sharing a common login area and user authentication. Remote sites will be able to replicate parts of the system to build simplified, easy to manage, local analysis systems. Risk-free and more aggressive variations on the basic design are discussed too.

1 Introduction

The next collider run of the Tevatron, Run II, is scheduled for spring of 2000. The CDF experiment plans to record about 1 PByte of data during the two year duration. About half of this data is raw data from the detector. The data presents an over 20 fold increase in data volume compared to the previous data recording, Run I. However, the data handling and analysis challenge is even larger: With the upgrade of the detector and trigger system a significantly higher good event purity is expected. And with the growth of the collaboration during the last years and of the physics potential of the detector, an increase of analysis activity with respect to Run I is anticipated.

The Run I analysis system was very successful. A DEC VMS cluster of over 20 nodes (with six being multi-processors, order 4 CPUs) and a large Silicon Graphics/IRIX multi-processor system (≥ 20 CPUs) formed the central analysis system. Each of the two systems had about 500 GBytes of disk space, access to a common Storage Tek silo with a 1.2 TBytes capacity and about 30 operator serviced Exabyte 8mm tape drives. A small IBM AIX system added extra analysis power towards the end of the run.

While the VMS system was the main analysis system at the beginning of Run I, the SGI system serviced the majority of the analysis activity towards the end of the run. The VMS system was an upgrade from the Run before and was very robust. The SGI was a new system. Significant effort was put into this system and its system management by CDF to make this machine reliable. It has been upgraded on a regular basis during the last years, with upgrades reaching and being limited by the maximum CPU capacity of the machine.

Most analyses are performed on either of the systems. Each physicist selects a system for her/his analysis on personal preferences, location of datasets, the system used by colleagues working on a similar analysis, and machines available at the home institution (to name some of the important criterias). This resulted in a rather small duplication of datasets on the various machines. Most of the data disk space is rather static. The order 100 GBytes of dedicated staging space in front of the silo on each VMS and SGI system is probably a good measure of the amount of disk space holding duplicate data.

While this note describes the hardware architecture, a second note [1] addresses the resource management of the central analysis system and data handling software.

2 Design Goals

Storing both raw and reconstructed data on permanent storage is a necessity. With the large amount of data (and our finite computing budget) the use of serial media (i.e. tape) is necessary. However, data on tape is hard to analyse. I/O bandwidth of tape is about an order of magnitude smaller than disk I/O bandwith (few MBytes/sec tape I/O versus tens of MBytes/sec disk I/O), and significantly increases the execution time (elapsed time) of an analysis job. With the use of Exabyte 8200 and 8500 technology in Run I there was an additional very noticable reliability difference between disk and tape resident data analysis.

Also to best explore the physics potential of the experiment we need to filter out unwanted data and get the data of interest quickly onto disk. Network connections from the datalogger and production to the storage pool will enable fast debugging of both detector and reconstruction programs and quasi realtime data analysis. Avoiding duplication of data on disk results in more effective disk space and the ability to accommodate more analyses or a faster analysis turn-around.

Sufficient CPU must be available to satisfy the ever increasing demands from analysis activities. The central system must be able to provide the necessary analysis power in order to support collaborative work and to keep analysis productivity high. It should not force physicists back onto their desktop or remote systems due to limited resources. However, making the use of both desktop and remote systems attractive is of high priority as it will offload analysis activity from the central system. With the steep drop in cost of commodity computing equipment the gap between commodity and supercomputing equipment is increasing more and more. The central analysis system must stay connected to the cost effectiveness of commodity desktop systems.

The analysis system is a tool, like any detector component, in the scientific research. What makes this tool special and a fitting design more important is the fact that it will be used directly by most physicists. As such it should be easily usable and support the physicist in her/his work. It should not enforce unnatural working methods but support the user as much as possible in her/his analysis approach.

The central analysis system for CDF Run II has been designed based on our Run I experience with the following primary goals in mind:

- The need to have enough tape storage for all data – raw data, reconstructed data, secondary, tertiary, and user datasets.

- To have as much data disk space as possible, limited only by budget.

- Analysis jobs should read and write data from/to disk and not deal directly with tape drives. Tape resident data should be copied to disk first, in a way that is automated and transparent to the user.

- The system needs to be scalable beyond current estimates and should be able to integrate new technologies as they become available.

- Analysis in the B0 trailers or at remote sites should be attractive and the import/export of data easy.

- The system should be robust and periodic tasks automated so that it requires as little operational support as possible.

3 The Central Analysis System

Commissioning of the central analysis system will start in early autumn of next year, 1999. The system will be extended and upgraded over four or more years and may even transition into our Run III analysis system. With such a long lifetime it is important to integrate new technologies as they become available and/or mature. Using the remaining time before the start of the first acquisition phase for extensive R&D will enable a progressive design that is both viable throughout Run II and can be realized with confidence before the beginning of Run II. On the other hand, one also needs to document one's ideas/plans from time to time, like now. In this document we present three scenarios: a design we believe to be feasible by summer of 1999, after a small development effort (called realistic design from here on); a "risk-free" scenario that could be implemented today; and a more aggressive design relying on storage area filesystems to mature on the timescale of 2000. The hardware architecture of all three designs allows a smooth transition all the way from the most conservative system to the most aggressive one.

The realistic design assumes a small amount of development work in the area of Fibre Channel (FC) shared storage. This development means testing that reading data off FC disks is robust in a heterogeneous environment with multiple hosts (SCSI initiators) per disk. According to information from storage vendors this should be no problem if the right, compatible hardware is selected. We need to convince ourselves here and understand the incompatibility issues. The design also assumes that we can read-share data on a filesystem among SGI/IRIX, Sun/Solaris, and Intel/Linux systems. We are currently verifying one of the filesystem options.

The design is based on a central storage pool. This pool consists of local and shared disks with a capacity of over 20 TBytes, an automated tape library (ATL) with a capacity of 1 PByte, and shelf resident tapes in case the capacity of the ATL is exceeded. The compute system consists of a small cluster (order 10 nodes) of moderate sized multi-processor (8 to 16 CPUs) machines. The machines are clustered loosely, sharing only the user login and disk spool area (disk scratch space is local on each machine) and authentication information (i.e. password, etc.). A sketch of the hardware architecture with the major components is shown in Figure 1. In the following subsections we discuss each component in more detail.

3.1 Network File Server and Authentication Server

An analysis system with more than two or three standalone compute nodes is not feasible from a user point of view. Even with work/physics group specific association, collaborative working will be degraded. Keeping files and passwords manually synchronized between multiple systems

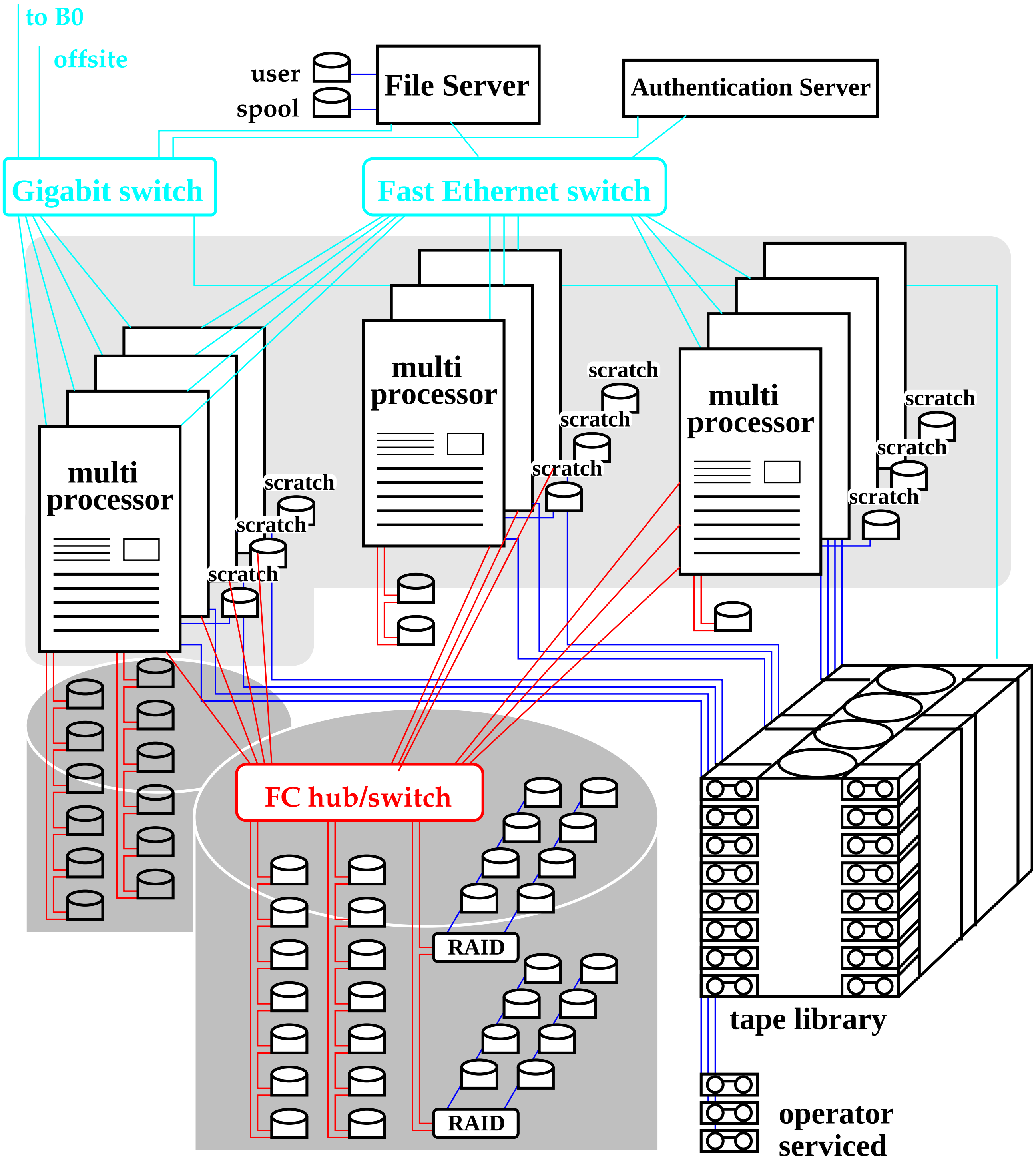

Figure 1: The CDF Run II central analysis system with locasl and shared data disks.

will be rather painful for the users. To enable the option of an analysis cluster (versus one big mainframe type machine) we need to share user files and passwords efficiently and reliably. Several vendors offer clustering software. Unfortunately most are restricted to homogeneous clusters of that vendor and are proprietary. If we want to have a cost competitive compute system and make use of Intel/Linux system in the central analysis cluster than such software is not an option.

Fermilab and CDF have a lot of experience with the Sun network filesystem (NFS). In the past we used computers to NFS-serve filesystems and used version 2 of NFS. Performance issues and reliability, especially with SGI servers, made clustering based on NFS worrysome. During the last years dedicated NFS fileservers have matured to a point where they claim performance similar to filesystems on locally attached disks. In addition, NFS version 3 has become available on most platforms. A dedicated NFS server will provide a clusterwide login and spool area for all users. We have acquired a small evaluation unit from Network Appliances. Setup and testing of the unit is in progress.

We send most of our passwords unencrypted over the network. Tightened security at Fermilab encourages us to use a more secure authentication system. Most advanced authentication systems are network based and thus support a cluster configuration naturally. A dedicated authentication server (most likely running Kerberos and providing SecurID tokens) will handle all authentication inside the analysis cluster. The Distributed Computing department has setup such a system and is in the process of testing it. We are following their activity with interest.

Both file and authentication server will have a dedicated network to talk to the compute systems or high priority access through quality of service. While the implementation details are of less importance and are currently discussed with the Data Com and Networks group, a reliable, low latency network connection between those two servers and the compute nodes could always be implemented with a dedicated network.

3.2 Compute Systems

The CPU requirement estimated for Run II analysis is about 90,000 VUPS, with a VUP being the CPU power of a VAX-780, roughly one MIPS. A fully populated SGI Origin 2000 should reach about 90,000 VUPS by the year 2000. Although such a system would satisfy our needs, it will not provide much upgrade capability (other than through technology advances from the vendor, i.e. CPU replacement) and will be rather expensive. Large scale multi-processor systems (SMP or NUMA) are supercomputing equipment and will, most likely, become less and less affordable for high energy physics. Moderate sized compute systems will be significantly more cost effective as they are commodity workgroup servers. By the year 2000 we expect Intel/Windows NT based workgroup servers to be at or past 8 CPUs. This will most likely be the price breakpoint for multi-processor UNIX servers as well. A cluster with order 10 nodes of moderate sized servers will also provide upgrade capability beyond pure CPU replacement. However, the clustering mechanism has to be well designed to archive a robust and user-friendly system.

We assume the flavour of the servers in the compute system to be a mixture of SGI/IRIX, Sun/Solaris, and Intel/Linux systems. The mixture of the first two would be based on competitive bids from the vendors, probably around a 1/3 to 2/3 fraction. For Intel based systems there is little benefit of a competitive bid and those would be acquired as needed. We believe we need SGI/IRIX or Sun/Solaris systems to serve the large interactive user load of CDF. Our past experience on SGI/IRIX with tape devices makes those machines not attractive for large tape I/O. In a computing representative meeting of CDF in early 1998 IBM/AIX was replaced in favour of Sun/Solaris as supported platform. DEC/UNIX is the fourth supported CDF Run II platform but at this time we do not expect significant DEC/UNIX presence in the central analysis system (based on KTeV's experience with DEC/UNIX as analysis platform).

3.3 Disk Storage

The mimimum amount of disk space needed for analysis activity was estimated in CDF-4072[2] and CDF-4100[3] to be between 20 and 28 TBytes. These estimates assumed a data duplication at the level of Run I and a similar purity of good events. The total I/O bandwith from/to disk was estimated based on Run I access patterns and measurements to about 1.5 GBytes/sec.

In Run I almost all data disk were SCSI disks. With transfer rates on the SCSI bus increasing, both cable length and number of devices per bus limitations become more and more severe. Low voltage differential (LVD) is addressing the length limitation but cable length is still reduced by a factor of two each time the transfer rate doubles. To overcome the limitations of SCSI and better accomodate the large number of disk drives required, we plan to use Fibre Channel (FC) as disk/computer interconnect. FC is not a new technology. ANSI chartered the FC working group in 1988 but only until recently have Gigabit implementations become widely available. FC is now supported by all major disk manufacurers. It allows the continued use of the SCSI protocol (but now over the FC transport). It also provides superior bandwidth and is becoming more and more mainstream computing, potentially surpassing SCSI. A nice side effect of FC is its network character (although being a channel technology). For CDF Run II this is a very interesting feature that allows us to share part of the disk space more easily among the nodes of the analysis cluster.

Both FC and SCSI disk are at the same price level, about a factor two more expensive than IDE disk of the same size. If we assume disk sizes of about 75 GBytes in 2000, then the required analysis disk space translates to over 350 disks. With such a large number of disks, failure rates at even the per mil level are significant without redundancy. RAID controllers have dropped in price to a level where they become attractive even for data disks. We have acquired a controller from Adaptec with SCSI device channels and FC host connection. The unit has just arrived and testing of the device is planned during the next weeks. We are interested in maximising the data disk space. With this goal in mind we cannot ignore the half price IDE disks. Several companies are currently selling RAID controllers with IDE device channels and SCSI host connection, and the first companies have started developing RAID controllers with IDE device channels and FC host connection. We need to follow up on this development and test devices as they become available and mature as this could potential be a big savings (or translated: a doubling of affordable disk space).

In order to share disk space in a heterogeneous environment, more than physical access to the disk is required. The filesystem on the disk must be understandable by all systems in the cluster. Two filesystem formats are understood by all three, SGI/IRIX, Sun/Solaris, and Intel/Linux: MSDOS (although only supported by SGI/IRIX on floppy diskette) and ISO-9660 used on CD-ROMs. A third filesystem, the Universal disk format (UDF), should be supported by all three operating systems within a few month as it is the filesystem used on DVDs. The ISO-9660 is an appropriate filesystem format for static write-once read-often disks like CD-ROMs. Most of CDF's Run I analysis data disk space was rather static. Well organized, static datasets could be written to disk into a ISO-9660 filesystem format and then accessed by all compute systems in the analysis cluster. We are currently verifying the feasibility of this option.

Disks used for selection output, to assemble new datasets, for tape staging, or for any temporary storage will be locally attached to the compute system via FC and have native filesystems. This will result in a clustering of similar analyses on the same compute system of the cluster, analogous to Run I. With the use of moderate sized multi-processor systems this should not provide a problem.

3.4 Tape Storage

The total data volume of CDF Run II is estimated at about 1 PByte. The media cost for this 1 PByte is significant and the use of commodity tape technology could result in large savings for this subsystem [5]. With a tape technology also affordable at remote sites, data export and import will be simplified. The serial media working group is thinking along the direction of commodity tape technology. The recommendation on Run II tape technology is expected for March of 1999. There are several promising commodity tape technologies. While we assume a commodity tape technology in the design of the central analysis system, there is little difference between commodity and data center quality tape technology from the hardware design point of view.

The tape library will be an EMASS AML/2 with four quadrotowers. Most tape drives will reside inside this automated tape library with only very few for shelf-resident tapes. The tape drives will be directly attached to the compute systems via SCSI; the plan is to use low voltage differential (LVD) SCSI. (All of the next generation commodity tape drive vendors plan to offer drives with LVD interface.) Device drivers for commodity tape drives should not be an issue for the operating systems under consideration. The required 24 hour average I/O bandwith to/from tape for CDF Run II was estimated to be about 150 MBytes/sec [4]. A moderate sized pool of commodity type tape drives (with about 5 MBytes/sec each) can easily provide this bandwidth. Each compute node will have access to a handful of tape drives and a few SCSI buses per node will suffice. Communication with the tape library controller (to request tape mounts) will be done over the general purpose network.

We hope to keep most of the data tapes in the ATL. Should the CDF Run II data exceed the capacity of the library, shelves will provide the extra capacity. Data tapes on shelf will be either read on operator serviced tape drives (for one time reads) or can be migrated back into the ATL. Although not a hardware issue, we like to remind the reader that user analysis jobs will never deal directly with tape drive. Any tape resident data will be spooled to disk for access.

The tape technology of Run II that will be use for buld data storage in the EMASS tape library is most likely based on helical scan technology. While such a technology is very cost effective for large data volumes it is less suited for small datasets. The raw data from the detector, the primary datasets from production, as well as most secondary datasets will have a size much larger than that of a single tape. The datasets will be written once with a large to infinite lifetime. When read, the entire dataset will be needed most of the time and the data is spooled to disk in large chunks. Tertiary datasets as well as user-defined data samples are much smaller and more dynamic. Such data are better kept disk resident or on tapes with linear recording technology. The total amount of disk space in the central analysis system is limited by budgetary constraints. Tertiary datasets and user-defined data samples are very important for a speedy analysis. As their size is rather small the desire to keep older versions for comparison and cross-checks rises. With the expected number of active analyses the amount of storage space for tertiary datasets and user-defined data samples is no longer small with respect to the available disk storage space. However, tenth of TBytes are still small from a tape storage point of view. Filesizes are expected to be significantly smaller than the raw and production target filesize of 1 GBytes. A hierarchical storage management (HSM) system would be ideal to extend the disk space used for tertiary datasets and user-defined data samples. Transparent migration to tape and restore of files from tape to disk could off-load the management of those data from the rest of the CDF data handling system. A tape technology based on linear recording handles small files, repeated loading and positioning of a tape, as well as start/stop operations intrinsically much better than tape technologies based on helical scan recording. With the migration and restore of individual (and mostly small) files tape mounting and dismounting becomes very significant. A fast robotic system will be required to satisfy the mount request from the HSM system. The EMASS library was bought with emphasize on capacity but not mount/dismount operations. The speed of the robotic system is well matched to the number of mount/dismount operations in a fully populated library when each drive reads or writes an entire tape worth of data. The Storage Tek silo that is currently used by CDF has a much faster robotic system. The system was upgraded last year in preparation for Run II. STK released end of 1998 a new family of tape technology based on linear recording. The STK 9840 has mid-point tape load and high speed tape search making it ideal for file pick operations. We plan to upgrade the 4490 tape technology in the STK silo to 9840 technology and use it to provide the much needed storage space for tertiary datasets and user-defined data samples.

3.5 Networks

We plan to use Fast Ethernet (or Gigabit Ethernet) to connect among any of the systems (network file server, authentication server, compute node, tape library control) and to a general purpose network. Each of those systems on this general purpose network will have its own port on a Gigabit switch which has an up & down link to B0 and offsite. We assume a similar connection (Fast Ethernet from each desktop to a Gigabit switch) in the B0 trailers. The file server and authentication network connection to the compute systems in the analysis cluster was discussed above.

3.6 B0 Trailer Desktop and Remote Systems

Workstations in the B0 trailers could be part of the analysis cluster and share authentication, login, and spool area with the central analysis system. We hope that this will result in a better integration and thus utilization of those resources, and allow to offload CPU intensive but data un-intensive tasks from the central analysis system.

Remote systems could be built according to the same architecture. None of the components in the central analysis systems are supercomputing style equipment. Moderate sized multi-processor systems should satisfy the analysis needs at most remote sites and thus eliminate the extra clustering effort. Both SGI and Sun have educational programs for universities. Fibre Channel equipment is becoming more common but could be replaced with SCSI devices for single system setups. The use of a commodity tape technology will allow to directly import/export data from/to Fermilab. It will also allow to integrate low cost tape stackers and small scale libraries into a remote setup. (The design of the tape I/O software is described in [1] and this software will be easy for remote sites to adopt.)

The various options for data access at remote sites need to be discussed further within the collaboration. We need to better understand the detailed iconstraints and needs at the collaborating home institutions so their prefered data access and analysis models are well supported.

4 Other Systems

Figure 2 shows the connection of the whole data handling hardware layout including online and the production system. The data logger of Run II will have two parts: A system in B0 sorts events into a handful of datastreams. Each datastream is written into a separate set of files. A system in the Feynman Computer Center, FCC, reads those files and writes them to the central storage pool. The production farms will have one or more dedicated I/O system. This machine will read data from tape, send individual files to the worker nodes for processing, collect the processed data, split and concatenate files as needed and write the output back to the central storage pool.

4.1 Dataloggers

The datalogger in B0 will receive events from the level 3 trigger system. The trigger system will be a farm of PCs running Linux. Events inside each farmlet are collected by an output node. The output nodes communicate with the online system and send events to the B0 datalogger machine. About 75 events per second are send to the datalogger at peak luminosity. Events are sorted by datastream and each datastream is written to its own set of files. To simplify bookkeeping and luminosity counting files are closed and opened at pre-defined boundaries. Events at the file boundary that are out-of-order (due to the parallelism in the trigger system) are sorted and written into either of the two files depending on their collision time. The online group currently plans this datalogger machine in B0 to be a Silicon Graphics O200. The data disks of this machine will have a dual Fibre Channel interface. (Dual interfaces are standard for most FC devices.) The disks will be read/write mounted on the B0 data logger machine and will appear like standard SCSI disks.

The datalogger in FCC will use the second FC port of those data disks. A single-mode fiber links the machine in FCC with the disks in B0. In order to understand the filesystem structure the FCC resident datalogger has to be of same OS flavour as the B0 machine. It will mount the data disks as needed in read-only mode. The machine in FCC has access to both levels in the central storage pool. A dedicated pool of tape drives in the tape library is connected via SCSI to the datalogger machine. The machine will read the data files from the disks in B0 and archive them on tape. Data needs to be archived at about the same speed at which they are recorded by the B0 datalogger. At peak luminosity this will be 20 MBytes/sec. It will also write some of the data files to the central analysis disk pool. The handshake between the two datalogger machines on which files to archive or delete will be done via the general purpose network.

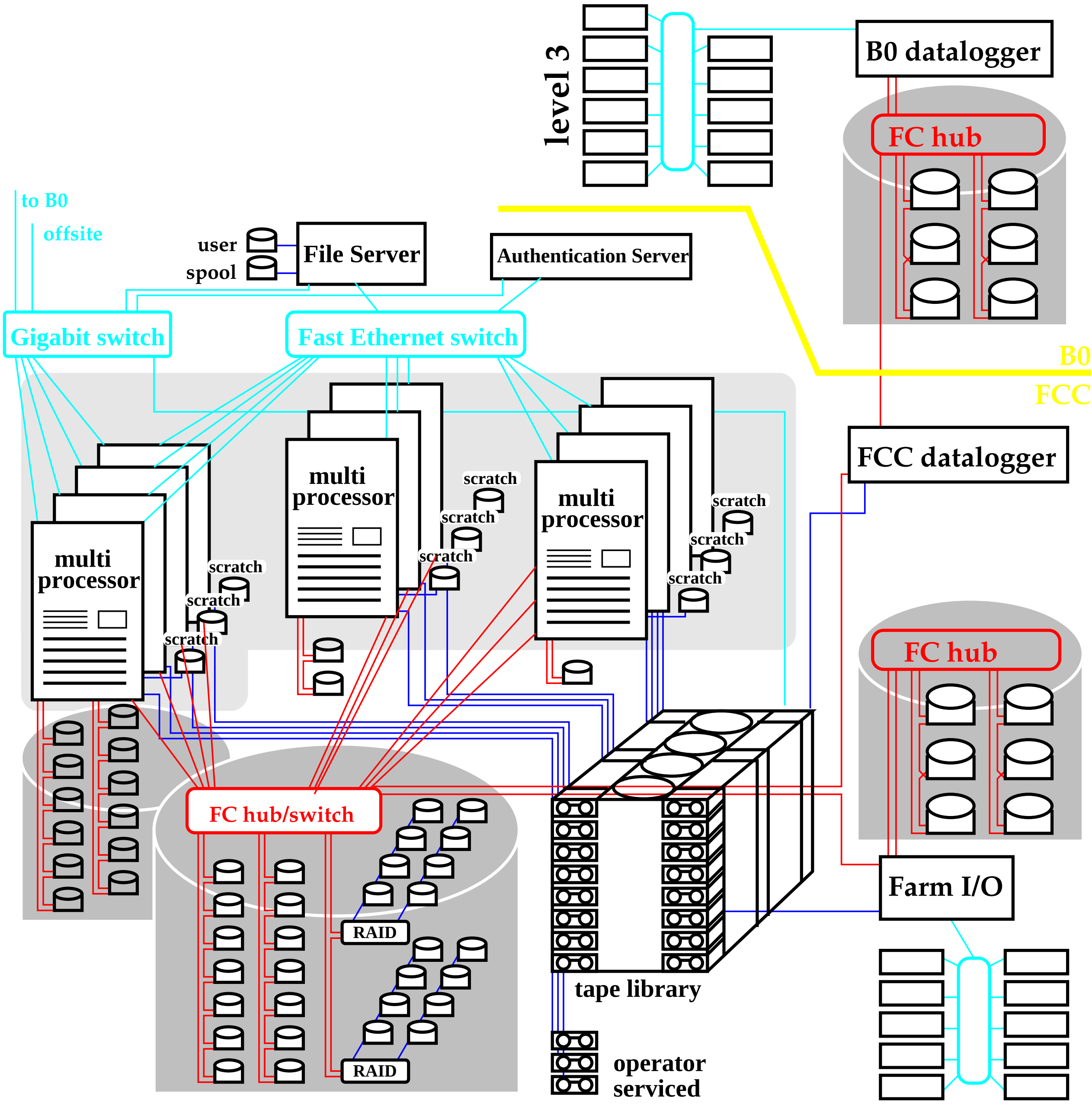

Figure 2: The CDF Run II data handling system.

4.2 Production System

The dataloggers records signal information from the detector. Those raw data have to be converted to physics quantities before the events can be analysed. The reconstruction of the events is very CPU time intensive. A farm of PCs running Linux is expected to be the most cost effective solution for this processing. One or more I/O nodes will have a connection to the central storage pool and fetch raw data files onto its local disks. The files will be distributed to the worker nodes and stored there on a local disk during processing. The worker nodes will write the reconstructed events to the local disk and will send the file to the I/O node when the processing of the raw data file is complete. The I/O node and worker nodes will be connected through a switch with Fast Ethernet connection to each worker node.

While the raw data are organized into a handful of datastreams, the production output should be organized along datasets. There will be around 8 datasets per datastream for a total of about 50 datasets. The dataset splitting will be done in the I/O node. While the size of a raw data event is approximately 250 kBytes, the reconstructed information per event is much smaller, around 65 kBytes. With the smaller event size and the dataset splitting, several files need to be concatenated in order to write files close to the target filesize of about 1 GBytes. It would be desirable to keep an inclusive, unsplit (but concatenated) set of files with processed data in addition to the split data. However, storage cost may prevent us from doing so. The I/O node will archive the files with processed data on tape. It will also write some of the data files to the central analysis disk pool. The disk space of the I/O node will be attached through Fibre Channel and could have a connection to the FC switch of the analysis disk pool to allow "spying" on production output from a machine in the analysis cluster.

5 Risk-Free Design

The design of the "risk-free" central analysis system is based on current technology and feasibility. The system contains no shared data disks and the number of multi-processor systems is reduced in favour of fewer, larger multi-processor systems. Figure 3 shows a sketch of the hardware architecture. CDF Run II software has not yet been ported to Sun/Solaris and there has not been any extensive tape drive usage on Intel/Linux systems that we know of. Sun/Solaris device reliability and robustness testing, however, is complete [6].

The design contains a network based file server for user login and spool area and a authentication server similar to the design described in section 3. The network file server is still unproven. Should such a file server not be possible we have to fall back to an unattractive, single large multi-processor design. The feasibility study of the network file server should be complete by middle of October. We'll ignore this "risk" for the next few days.

In the case where we cannot share any disk space (not even in a homogeneous environment, i.e. assuming two steps back from the realistic design) there should be as few compute systems

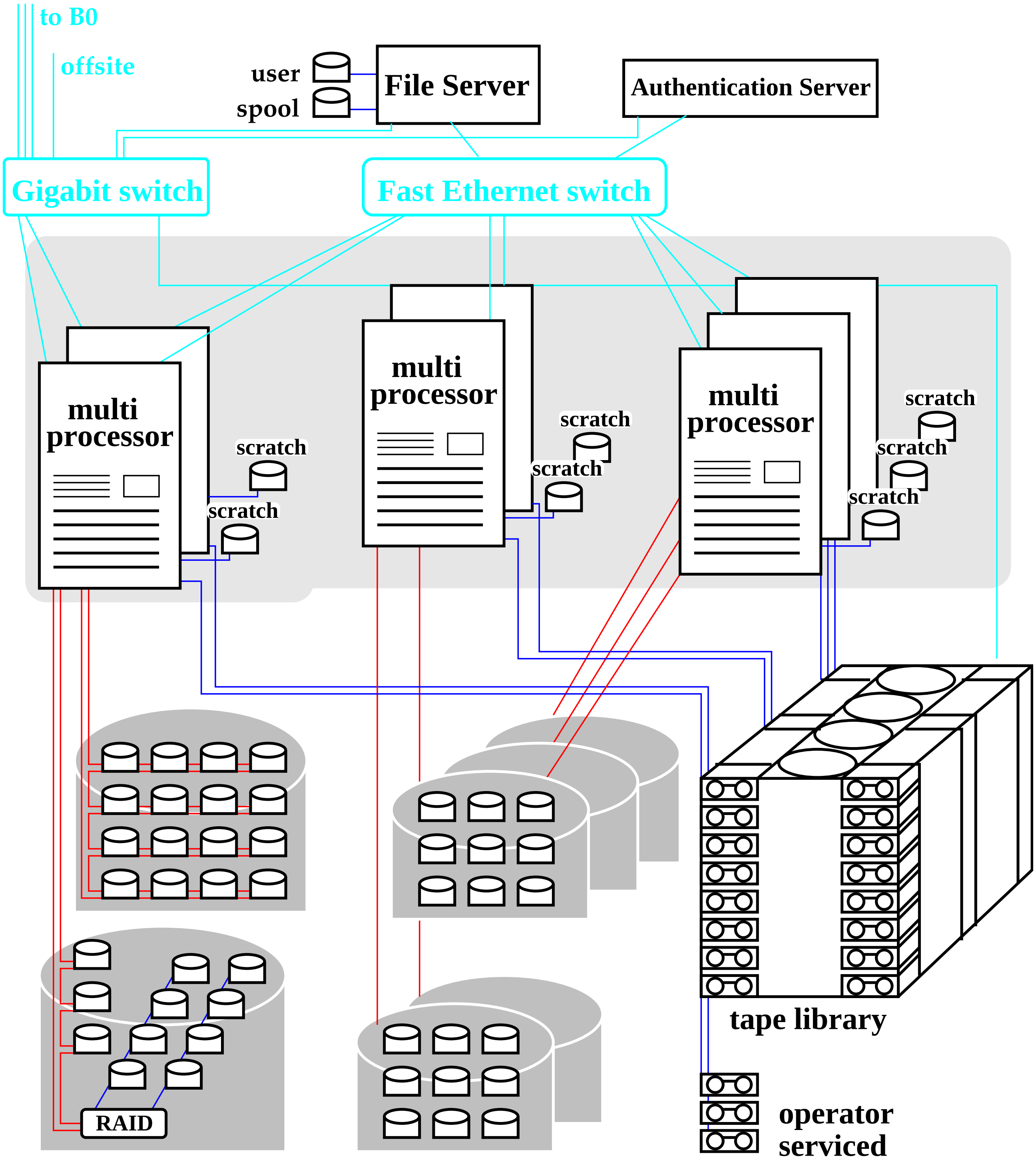

Figure 3: The CDF Run II central analysis system, risk-free design.

as possible so as to not partition the analysis data disk space too much. The exact size will be an optimization between efficient disk utilization and cost effective compute system. In general we expect larger multi-processor systems in this scenario.

All disk storage would then be locally attached (via SCSI although FC would be an option too) and all data disk would have native filesystems. Relying on only current, tested technology would mean the extensive use of SCSI expanders and converters to hook up the required disk space (we now have to assume 36 GByte capacity disks). Additional development work is required to integrate and manage such a large quantity of SCSI devices (over 700). This work could easily be of the same scale as the proposed FC testing in the realistic design.

Tape storage would remain as before. The amount of tape drives is still small and requires only a few but now precious SCSI channels. The EMASS automated tape library has already been acquired and it would still be our prefered choice since we still want to optimize for capacity. Luckily the selection of tape technology is in the hands of the serial media working group so we don't need to consider tape volumes of current capacity here.

The network technology currently available on all platforms is Fast Ethernet. To achieve the required bandwidth to the desktop systems in the B0 trailer complex today one would need to replace the Gigabit link with several Fast Ethernet links. To remote sites we assumed a low rate data transfer rate and shipping of large datasets via tape. With this assumption there is no change to data access at remote sites.

Of course no system is risk free. Our Run I experience has made clear the importance of good system management. We need a close collaborative approach involving the Operating System Support department, the Equipment Support department, the Distributed Computing department, and the Computing Services department with people from the experiment taking a leading role. Without this even the simplest system is doomed to fail.

6 Aggressive Design

The aggressive design differs from the realistic design in Section 3 by assuming one of the storage area filesystems (like the Global File System (GFS) or Cross Platform File System (CPFS) from Data Direct Networks/Storage Tek) has matured on the timescale of Run II. Such a filesystem would be based on FC attached disks, contain a global metadata server, and allow each compute system in the cluster to read and write from/to each disk. In recent discussions (Sep. 21st, 1998) with Data Direct Networks and Storage Tek the vendors were confident about marketing a product by December of 1998. Data Direct Networks has shipped the first test systems and plans general availability by December with three different metadata servers. Both price and performance are unknown at this time (except sales statements). Storage Tek is marketing the Data Direct Networks systems and will anounce pricing and performance measurements at the Storage Tek forum on October 18th to 24th. GFS [7] does not require a metadata server. Instead, device locks are used to synchronize writes. An implementation of GFS is available for SGI/IRIX. We don't know of implementation for other operating systems.

In the aggressive design there is almost no local disk space on the compute systems. A sketch of the architecture is shown in Figure 4. There is no need to change the clustering strategy from what was described in section 3. (Although one could now use the FC attached disk also for user login and spool space.) The compute systems are identical to what we described in setcion 3. Tape storage and network setup is also unchanged. There would still be no desire to use FC for Internet Protocol (IP) traffic as it is better handled by Ethernet (or Fast/Gigabit Ethernet).

Acknowledgements

We are grateful for useful discussions with Eric Wicklund.

References

[1] W. Brown *et al.*, CDF Run II Data Handling Resource Management and Software Components, CDF note 4759 (1998).

[2] L. Buckley-Geer, S. Lammel, and R. Plunket, CDF Run II Data Handling (Data Volume Estimates and Equipment Requirements), CDF note 4072 (1997).

[3] D. Baden *et al.*, Joint D0/CDF/CD Run II Data Management Needs Assessment, CDF note 4100 (1997).

[4] E. Berman *et al.*, Data Flow inside a Run II Light-Weight Event Database System, CDF note 4495 (1998).

[5] J. Bakken *et al.*, Serial Media Working Group First Run II Report, CDF note 4560 (1998).

[6] W. Brown *et al.*, Device Testing and Robustness of Sun/Solaris CDF note 4758 (1998).

[7] S. Soltis, T. Ruwart, and M. O'Keefe, The Global File System, Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, College Park, MD, (1996).

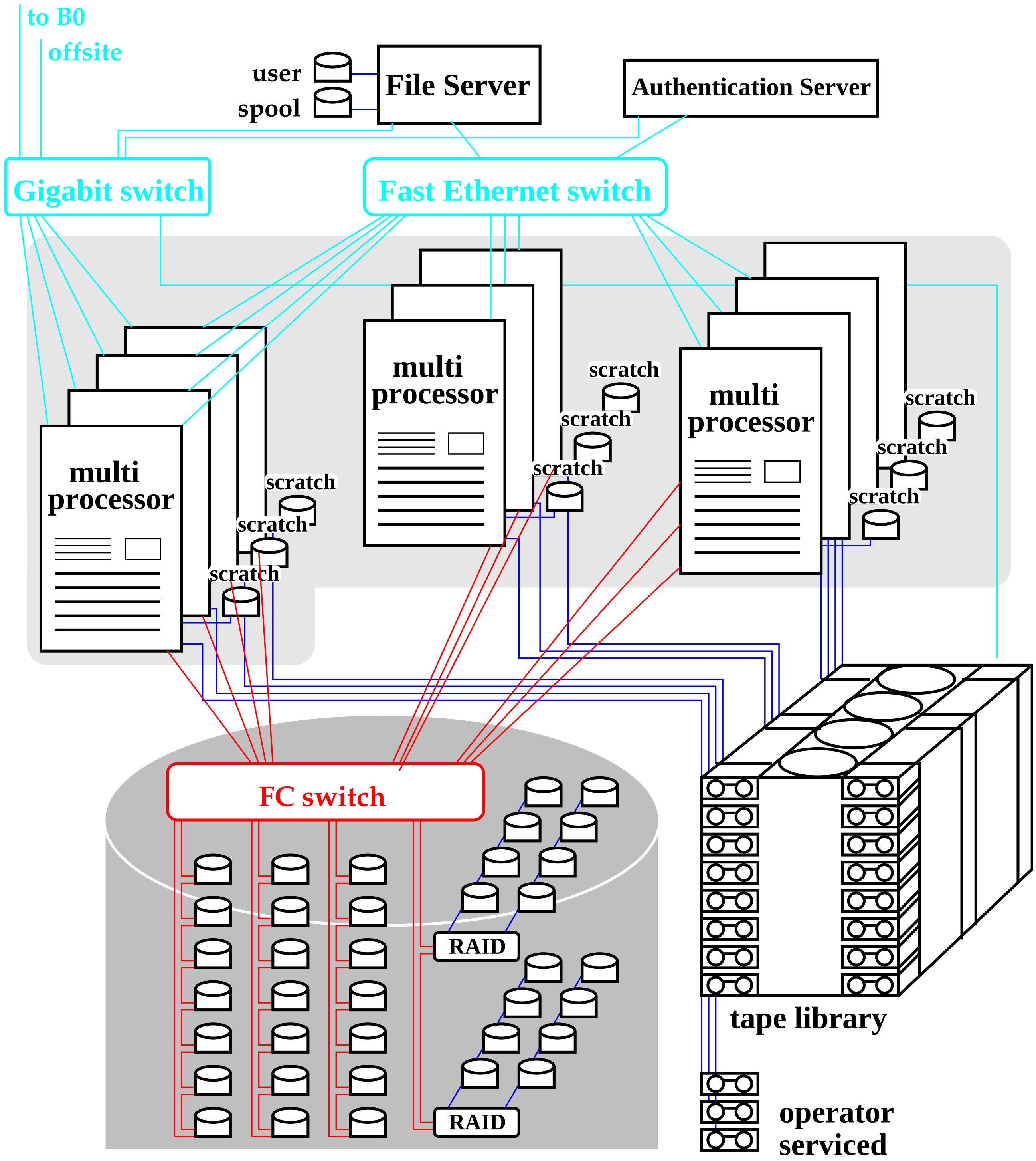

Figure 4: The CDF Run II central analysis system, aggressive design.