56-32
83188

Petabyte Class Storage at Jefferson Lab (CEBAF)

Rita Chambers, Mark Davis
Jefferson Lab Computer Center
12000 Jefferson Ave.
Newport News VA 23606
chambers@cebaf.gov
davis@cebaf.gov
Tel: 757-269-7514
Fax: 757-269-7053

Abstract

By 1997, the Thomas Jefferson National Accelerator Facility will collect over one Terabyte of raw information per day of Accelerator operation from three concurrently operating Experimental Halls. When post-processing is included, roughly 250 TB of raw and formatted experimental data will be generated each year. By the year 2000, a total of one Petabyte will be stored on-line.

Critical to the experimental program at Jefferson Lab (JLab) is the networking and computational capability to collect, store, retrieve, and reconstruct data on this scale. The design criteria include support of a raw data stream of 10-12 MB/second from Experimental Hall B, which will operate the CEBAF Large Acceptance Spectrometer (CLAS). Keeping up with this data stream implies design strategies that provide storage guarantees during accelerator operation, minimize the number of times data is buffered, allow seamless access to specific data sets for the researcher, synchronize data retrievals with the scheduling of postprocessing calculations on the data reconstruction CPU farms, as well as support the site capability to perform data reconstruction and reduction at the same overall rate at which new data is being collected.

The current implementation employs state-of-the-art StorageTek Redwood tape drives and robotics library integrated with the Open Storage Manager (OSM) Hierarchical Storage Management software (Computer Associates, International), the use of Fibre Channel RAID disks dual-ported between Sun Microsystems SMP servers, and a network-based interface to a 10,000 SPECint92 data processing CPU farm. Issues of efficiency, scalability, and manageability will become critical to meet the year 2000 requirements for a Petabyte of near-line storage interfaced to over 30,000 SPECint92 of data processing power.

Introduction

The Thomas Jefferson National Accelerator Facility (formerly CEBAF, the Continuous Electron Beam Accelerator Facility, and now known as Jefferson Lab), located in Newport News, Virginia, operates a 4 GeV continuous wave electron beam accelerator, with the capability to drive fixed target experiments in nuclear physics simultaneously in three Experimental Halls. By 1997, when all three halls are under production operation, the data generation capability of the experiments, including both raw and reconstructed data, is expected to approach 250 TB per year. By the year 2000, a total of one Petabyte of data will be stored on-line for access to users on both Local and Wide Area Networks.

In this paper, we outline some of the major design decisions and strategies employed in the development of an automated facility which can collect raw experimental results from three separate data acquisition operations, plus serve this information to a 10K+ SPECint92 batch data reconstruction farm. The central mass storage system must also store output from the data reconstruction and analysis process, provide intuitive access to files associated with specific data runs and phases of the analysis, plus provide a data export capability for transport of the summary information back to the researcher's home institution where final analysis steps will be performed.

Some of the most critical decision points in the process require coordination in the design of both the on-line and off-line phases of the operation. The size of the individual data run, which represents a specific running period for each spectrometer, naming conventions for raw data files and associated calibration, target mapping, and other auxiliary files, and the methods used by each experimental hall to funnel data to the central mass storage system must be anticipated in the design of the off-line data handling capability. In some cases, particularly in the size of the raw data file, limitations and optimizations in the off-line process will influence operations during the data acquisition stages. This paper will summarize the basic assumptions in the development of the data handling operation, including considerations in designing the data path for both on-line and off-line operations. We provide a description of the current evolution of the design, status of the current production operation serving one Experimental Hall, plus anticipate the challenges ahead as we scale the operation to support a Petabyte-class data storage requirement.

Data Handling Requirements

Inherent in the design of the data handling operation at Jefferson Lab is the requirement for an automated, "hands-off" operation. Physicists historically have run experiments with their hands "on the wheel" -- actively managing and monitoring the experiment itself while manually loading multiple small tape units to store the generated raw data. In this mode, the volume of the output and the success of the operation are immediately apparent. The researcher is responsible both to develop effective tracking and logging systems as well as to determine and resolve problems encountered in the experimental and data storage facets of the operation. When designing for the collection of 1 TB per day of raw data, it is immediately apparent that this classic mode of operation will not scale: people time is expensive; the task of recording and tracking large numbers of potentially large files is daunting. A 2 GB raw data file, for instance, represents less than 3 minutes of operation of the CEBAF Large Acceptance Spectrometer (CLAS) in Experimental Hall B. Manual logging methods developed when data rates were on the order of Kilobytes per second become unmanageable when data is being generated in Megabytes per second. Part of the design consequently is to meet the human requirements for visual verification of the success of the data storage operation, to develop intuitive methods to locate specific data runs and associated files, and to implement robust strategies to withstand interruptions in the central storage capability without affecting the on-line data acquisition (DAQ) process.

Due to the large scale of the operation, efficiency is tantamount. Critical to the design is the effort to minimize the number of times that raw data must be copied on its path to the central mass storage facility. The analysis of early designs, in fact, revealed up to four separate copies of the raw data file on its way to a robotics silo: DAQ to disk; disk to disk via network to the tape storage server; a re-copy to disk buffers required by some tape management applications; and a final copy to tape in the robotics library. Recopying 1 TB of data incurs large costs in both time and hardware and could significantly increase the resources required. Furthermore, data reconstruction processes in nuclear physics on the average make two to three passes through the original raw data. While a true data reduction in this phase of the analysis is desirable, "reduced" data may, in fact, equal or even exceed the size of the raw input in some instances. There may be many reduction stages in the final creation of a Data Summary Tape (DST) sufficiently small to be transported to the home institution for final analysis. Consequently, efficient algorithms to coordinate the use of tape transports and disk pool areas, to optimize network and batch node performance with central and local disk buffers, to manage on-line storage of output information anticipated for near term re-access, and to vault experimental results to be maintained 10 years or more in off-line storage, are essential in maximizing the use of expensive resources (tape, disk, CPU, network).

The overall design must also meet the requirements of three separate experimental operations, and, in fact, arbitrate resources between the halls, isolating them from each other. JLab's three experimental halls each impose a different set of requirements and timelines for computational support, and in many cases make use of a variety of procedures and standards in the operation of their experiments. The data transport, storage, and post-processing requirements for Experimental Hall B, due to begin production during FY97 (10/1/96-9/30/97), significantly surpass the standard operational requirements for Halls A and C. While planning has focused on meeting the technical challenges posed by the collection and processing of approximately 1 Terabyte per day of raw data (after the compression phase in the data acquisition process) from the CEBAF Large Acceptance Spectrometer (CLAS) in Hall B, a data stream equivalent to approximately 10-12 Megabytes per second, the plans must also provide viable solutions for the lower data rates projected for Halls A and C. The other two halls, which begin operation in an earlier time frame than Hall B, generally incur lower data rates (1-2 MB/second) and while eventually requiring a separate data path from Hall B, can in fact be used to test, tune, and refine the solution for Hall B.

A summary of the basic data storage requirements and timelines for the three halls illustrates that the data handling requirements for Hall B are an order of magnitude greater than for the other two halls:

Hall	Test Runs	Production	Event Size	Events/ Sec	MB/Sec	Data/Day
C	Complete	Current	1 KB	200 - 2000	0.2 - 2	1 - 100 GB*
A	2Q96	4Q96	1 KB	200 - 2000	0.2 - 2	1 - 100 GB *
B	4Q96	1Q97	10 KB	1000	10	1 TB

* Approximately 1-25 GB/day under normal operation. Exceptions to this rate are the Hall A Helium parity experiment which will run at 10 KHz (10 MB/sec or about 1 TB/day) for a few months, and Visual Compton Scattering and other experiments which are expected to collect data at 2 KHz (2 MB/sec) with peaks up to 10 KHz.

The data reconstruction requirements for Halls A and C are estimated as 1/10 those of Hall B. Hall B estimates that each Byte of data will require on the order of 1000 instructions to reconstruct. At 10 KB/event and 1000 events/second, this is roughly equivalent to 10K MIPS (or 10K SPECint92). Using this projection plus anticipated increases in data rates, CPU and data resources must be implemented in the following scale:

FY	CPU (MIPS)	Disk (GB)	Near-Line Tape (TB)
96	2 K	100	5
97	10 K	500	150
98	20 K	1000	300
99	30 K	2000	1200

Meeting the usability, efficiency, and flexibility requirements outlined above imposes a special set of challenges for the modest budget and staff dedicated to this operation. The design described below consequently makes heavy use of commercial software applications and standard off the shelf hardware. The selection of hardware and software components has stressed cross-platform capabilities so that components can be replaced and/or upgraded as needed without major redesign of the facility. Project management has stressed the close involvement of users from all three Experimental Halls in addition to Computer Center personnel, who will implement and manage the central mass store, in order to insure that the design meets both the technical and human aspects of the overall requirement.

Factors in the Design of the Data Path

Several factors were evaluated in the design of the data path for both the on-line and off-line operations. In almost all cases, the final decisions represent trade-offs in terms of cost, efficiency, and robustness. With a small staff, plus some input and assistance from physics users, and limited budgets, simplicity is key.

Factors in Designing the On-Line Data Path

All three Experimental Halls use some combination of VME and Fastbus technology to collect raw data from one, and in some cases, two spectrometers. Hall B has developed an event building capability which employs the Asynchronous Transfer Mode (ATM) network technology to collect, sort, format, and compress the data points from each physics event. Details of their algorithm are provided in reference [1]. VME single board computers serve as readout controllers (ROCs) to collect data from the electronic crates; control and data messages are passed in the 53-byte ATM cells over OC-3 connections (155 Mbps) between the ROCs and an on-line SMP (Symmetric Multiprocessing) farm processor (OLFP), a UNIX server. Formatted event data must then be transported to a mass storage library for eventual replay to the off-line batch farms. Some local backup tape capability is desired both for convenience and redundancy.

Design decisions specifically related to the data path of the on-line physics events include (a) the location and number of robotics libraries required to collect the raw data, (b) the network implementation over which to transmit the information, (c) the number, speed, and capacity of the tape transports to employ, and (d) the number of copies of the raw data to be stored.

Robotics Libraries: Considerations such as redundancy and the need for visual feedback from storage operations led to the evaluation of implementing dual robotics libraries to support the on-line operation. In this model, a smaller robotics library would be located in the experimental area (the "Counting House" where the DAQ systems reside) so that tape storage of on-line information could continue even in the event that network connections to the central site were interrupted. With modern Hierarchical Storage Management (HSM) software, data loaded to the Counting House silo could be migrated in background to a higher capacity central silo used to feed the off-line batch farms. This arrangement, while providing good redundancy and failover capabilities, plus fulfilling the human need to keep the raw data of the running experiment "local", in fact results in one extra copy step, greater complexity in managing the location of the data and in freeing sufficient storage space for real time operations, plus most importantly doubles the cost of the operation. Locating all tapes (including duplicate copies if required) within one central silo (or silo-complex) insures that the data is where it is needed when it is needed and provides a central single point to expand when capacity requires. After evaluation of the options, simpler, cheaper solutions for redundancy and feedback can be implemented with graphical monitoring utilities to provide visual feedback to the researcher, and local disks and lower cost tape drives on the DAQ systems for buffering and emergency archives. This solution does require that some mechanism provide for the uncontested use of central tape drives for on-line operations. This is particularly critical for the high data rates for Hall B, where local buffers could quickly overflow if the real time operation waited on lower priority off-line use of the central drives. The decision to employ an HSM file management approach posed a problem in that most HSM applications do not provide tape allocation capabilities. Consequently the design of a local customized tape staging application must incorporate the capability to insure that a tape transport is immediately available for selected real time processes.

Network Medium: Viable network transports for the raw data include FDDI (Fiber Distributed Data Interface), ATM, HiPPI (High-Performance Parallel Interface), and Fast Ethernet (100BaseT). While the costs of ATM may prove to be lower than the more mature FDDI and HiPPI standards, many vendor offerings are unproven and still groping for a standard. Fast Ethernet (100BaseT) provides both higher capacity and cost effectiveness but may not meet the high speed throughput requirements of Hall B. Decisions regarding switching versus routing must insure that signals from the three halls do not interfere with each other, yet are not degraded by latency overheads.

Tape Transports: A major design decision was whether to use multiple lower speed/capacity tape transports (possibly DLT) or a fewer number of high end drives (Redwood--11.1 MB/sec., 50 GB cartridges; Ampex--15 MB/sec., 165 GB cassettes, etc.). The IBM Magstar Drive (9 MB/sec, 10 GB cartridges) offered a midrange choice in terms of cost and capacity, with relatively high end throughput (at least approaching the 10-12 MB/second data stream expected from Hall B). Employing multiple, lower cost drives has the advantage that losing one or even several drives has minimal impact on the overall operation, plus increases the possibility that the researcher can in fact read raw data tapes at the home institution. The disadvantage to this model is the increase in complexity in terms of fanning the data stream out to multiple drives as well as the significant increase in sheer floor space required to store tapes (both in near-line and off-line locations). The use of a higher throughput, higher density tape solution reduces the complexity of the algorithm, reduces the cost of tapes as well as storage, plus provides the throughput capacity required to "catch up" after scheduled and unscheduled interruptions.

Data Copies: It is interesting to note that the cost of generating the 300 TB possible to store in the laboratory's STK 4410 robotics silo, given site estimates of running costs, is many tens of millions of dollars (more depending on the volume of data collected from the lower intensity halls, Halls A and C, as well as the amount of processing required to produce any reconstructed and/or analyzed results). Consequently, the issue of whether duplicate copies of the raw data should be kept for all experimental runs is truly both a cost and research critical decision. Assuming Hall B produces approximately 50 GB per hour of operation running 125-150 days per year, the annual cost to save one copy of the raw data stream is on the order of $300K in tape costs alone. The cost, per copy, then is less than 1% of the overall generation cost. On the other hand, $300K, let alone $600K (assuming two copies), plus of course the original investment in additional $100K+ tape units, is a significant impact on tight experimental budgets. A survey of other energy research laboratories indicates that keeping duplicate copies of raw data is by no means universal even in far lower data rate environments, that total loss of a raw data tape is rare, and that the loss of some small percentage of an experiment's data would be unlikely to affect the overall results. This decision is still under consideration at Jefferson Lab and may be affected as much by budget restrictions as risk analysis.

Factors in Designing the Off-Line Data Path

Using the Hall B estimate of 10K SPECint92 to "reduce" (in many cases, just "reconstruct") the data from the CLAS spectrometer, the laboratory will require an off-line batch-mode CPU farm consisting of on the order of 50 CPUs ready for production operation during 1Q97. The final configuration of the farm and the supporting software will depend on several factors, including relative costs and performance of a range of processors, size/speed/cost of local node disks and central RAID subsystems, size of input raw data files as well as output files, and the complexity of the software algorithm to coordinate pre- and post-staging of data files with the data reconstruction job. The basic assumption in the design is that the first pass reconstructed data is approximately the same size as the input data. An actual reduction of the output data, to 10% for example, would drastically reduce the overall cost of the implementation, plus have significant impact on the overall throughput of the facility.

The data reconstruction operations on JLab's data involve a model of "trivial parallelism." One executable designed for an experiment can be used repeatedly against event after event either in sequence or in parallel to generate reconstructed events. Consequently the design decisions involve at what granularity to fan out events to a series of CPUs, making the basic assumption that a "pizza-box" style of post-processing will most likely cost less than the use of one, very high end multi-processing system. A PVM-approach (Parallel Virtual Machine), for instance, would use a "master" server to fan out single events to a series of CPUs each running the same code, collecting output back on the master node. Alternatively, blocks of events can be handled in a series of automatically generated batch jobs, with the naming of the output files used as a method to "collect" the results back into sequential order.

During off-line post-processing, raw data files must be retrieved from the central mass store via an automated multi-job generation process that loads required files "just in time" for batch processing and returns output to required locations (tape silo and/or on-line storage). Design decisions specifically related to the data path of the off-line processing include (a) how the researcher will access required files; (b) the algorithm and path used to pre- and post-stage files for the running batch job; and (c) the algorithm and implementation to allocate resources according to laboratory planning.

Data Access: One primary goal in the design is that access to the files associated with specific data runs should be reasonably intuitive to the end user. One method to implement this is of course to use the concept of the UNIX file system itself as a way to catalog files. The use of meaningful directory and file naming conventions then allows reasonable access, even without metadata, to specific file sets. Commercial HSM and other volume management applications support this access mode by implementing virtual file systems where only a portion of the files actually reside on-line. The use of standard HSM-style "file migration", where on-line water marks and recent use heuristics define which files are maintained on-line, provided one possibility to support the file system for JLab's experimental data. A disadvantage of this approach, however, is that files to be retrieved must first be "migrated" from tape to the local file system before they can be used. In the case of feeding an off-line post-processing CPU farm, the required location, for performance reasons, may very well be on dedicated central staging areas and/or local batch node disks, as opposed to the "cataloging file system," thus necessitating at least one extra file copy operation to locate the file where it is needed. A variant of file migration is the use of a file "stub", or marker to the actual tape location of the file, provided by some file management utilities. In some implementations, restrictions in the relocatability of stubs can pose a problem for expanding, dynamic file systems. In addition to intuitive access to the raw data and related output files, researchers must have the capability to store additional metadata related to both runs and data reduction phases. This requirement, however, calls for a database-oriented capability above and beyond the management of the virtual file system alone.

Staging Algorithms: Probably the most critical decisions for the overall design of the off-line batch farm revolve around how to make the input file, either a raw data file or the output from an earlier phase of data reduction, accessible to the batch job that eventually uses it. The question involves not only decisions regarding synchronization and job priorities, but from a design perspective even the anticipated input and output file sizes and how they may perform in either local or central staging models. Although data files could in theory be directly loaded from the tape silo to local disks on individual farm nodes, two limitations argue for an initial central staging area: (a) limitations in the I/O performance of the individual farm nodes. Although CPU alternatives exist with the required I/O performance, this will mandate higher performance and hence higher cost systems for the farm; (b) processor/tape utilization. Ideally, CPUs should not be idle while the next data file is loading, and the use of high cost, high performance tape drives should be optimized around efficient staging algorithms. De-coupling the two phases by means of central buffering best accomplishes each goal without compromising the other.

A variety of staging models can be considered. A real argument for smaller input files exists in both the current limitations in many UNIX operating systems for files less than 2 GB as well as the possibility to use inexpensive disks local to the batch node for actual input/output file storage. Considering that a 2 GB raw data file represents less than 3 minutes of beam operation in Hall B, such a limitation involves some level of inefficiency in terms of opening and closing files during the data acquisition process plus dramatically increases the number of associated raw data files. Substituting either a 25 GB (~30 minutes of operation) or 50 GB (~ one hour of operation) size for the raw data file may be more efficient from a DAQ perspective but effectively rules out truly local disk storage from a cost consideration and incurs the performance penalties of NFS or other network file access. Just to complicate the formula are considerations such as the time to "cold start" the farm, the time to "warm start" the farm after a brief interruption (e.g. take advantage of the files already pre-staged), plus considerations of the researcher's intent for longer term on-line storage of the associated output files. A PVM approach can solve a large file requirement by fanning out events from a large input file to the individual batch node, but network performance must be considered as well as the increased coding complexity for the researcher. Moreover, the entire model changes drastically if data reduction actually accomplishes a significant reduction in output file size during early processing cycles.

Resource Allocation: All resources required during the experimental process are allocated to an experimental collaboration according to laboratory planning, from the hours of scheduled beam time to the staging of input files for allocated use of the off-line batch farm. The bottom line for researchers, however, is how much CPU time they are getting to post-process the data collected during their on-line operation. The design of the algorithms to "feed the farm" must provide the best overall site throughput as far as quantity of data processed, yet accomplish this within the guidelines of allocation strategies mandated by the laboratory. In this central silo model, the design of tape drive allocation and staging algorithms must first of all meet the requirements to insure the uncontested use of storage mechanisms by real time operations. Beyond this, the focus must be on a fair share allocation of the farm resources; tape staging serves only the purpose of fueling the correct mix of jobs. The challenge in a fully automated system is at what point in the process to implement a fair share algorithm to achieve the overall allocation strategies of the laboratory--if during the tape loading stage, how can we determine in advance which files should be loaded to achieve the desired mix in the set of running jobs; if during the job submission stage, how can we insure that the required files will be available at the time to run? And what percent utilization can we hope to achieve with high performance, high cost tape transports? The prototyping and testing of these various models will be essential in selecting the optimal design given the specific CPU, I/O, and network parameters at hand.

Current Implementation and Status

The On-Line Model

The implementation in progress to handle the data flow from Jefferson Lab's experimental program includes the selection of the Open Storage Manager (OSM, Computer Associates, International) HSM software integrated with a StorageTek robotics library and Redwood helical scan tape drives. Fast Ethernet provides the base for the experimental network, carrying all data for the 1-2 MB/sec data streams of Halls A and C (See **Figure 1**, "High Speed LAN"). The use of a Fast Ethernet switched architecture provides enhanced performance by protecting raw data streams from outside interference. A developmental ATM switch will be used to prototype the potential use of ATM for the farm network fabric. The higher intensity data collection of Experimental Hall B (as well as potentially some experiments in Hall A) will use the network for control signals only, moving the raw data via dual-ported Fibre Channel RAID connected to both the Experimental Hall event-building CPU and the Computer Center data server. The StorageTek Redwood drives were selected for both their high performance (current benchmarks indicate an 11.1 MB/sec. throughput) and high density (50 GB cartridges are available). Due to the large volume of data anticipated, effectively ruling out the option of performing off-site data reduction, the advantages of these high performance/high density options for Jefferson Lab outweighed the disadvantage that the home institutions will most likely not be able to build similar environments. Furthermore, there was some advantage to the tape cartridge design of the StorageTek (STK) Redwood transports which contain no takeup reel, cutting the storage space required in half. One factor in the selection of the OSM software was the existence of a customized extension called *OSMcp*, developed at the Deutsches Elektronen-Synchrotron (DESY) which solved the relocation issue with stubs as well as provided exactly the data flow model required: the direct copy to a designated location (either disk-to-tape or tape-to-disk) using a stub file as a reference only. Furthermore, by use of UNIX protections on the stub files alone, the integrity of master copies of the raw data files, which should never be modified and rarely deleted, can be protected.

Figure 2, "High End Data Flow", illustrates the path to be taken by Hall B experimental data. Raw data collected during the Fastbus/VME-based data acquisition process will be channeled through an ATM switch to the Hall B SMP system (Data Acquisition Symmetric Multi-Processing, DA-SMP). This system will use the multiple input data streams to build events, locating the raw data files created on the dual-ported Fibre Channel RAID subsystem. The DA-SMP system will toggle between two separate RAID partitions and as it fills one, will signal the Computer Center data server (CC-SMP, a Sun

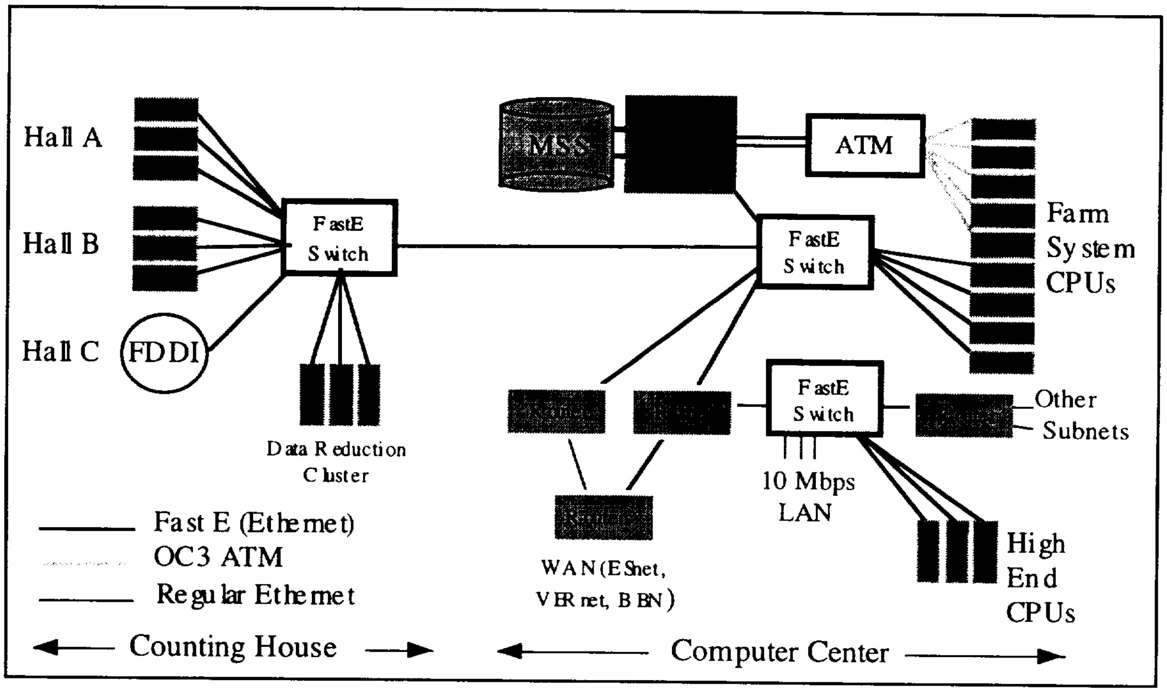

Figure 1: High Speed LAN

Enterprise 4000, 1 km away) via a network call, to begin moving the files on that partition to the STK silo. The DA-SMP will then resume data collection on the alternate RAID partition. This alternate writing then reading of the RAID subsystem, isolates the data transport from the network, and effectively enhances the throughput of the two data storage servers in that data does not have to pass twice through each system as required for a network-based transfer. Current testing of this model using a lower performance Sun 1000, including the use of the OSMcp utility, has resulted in transfers in the range of 9 MB/second, closing in on the performance goal of the maximum rating for the Redwood drives, 11.1 MB/sec. Configuring high performance components of the architecture (Redwood drives, RAID, network interface) with a separate SBus is expected to yield the remaining required throughput. While initial testing has involved only one RAID subsystem divided into two partitions and one Computer Center storage CPU, the solution can scale with the addition of multiple separate RAID units and storage servers. The CC-SMP will use the OSMcp utility to move the raw data directly to tape, leaving "stubs" in predefined directories corresponding to the specific Hall/Spectrometer. and experiment. The stubs, which appear to the user as regular UNIX files, are essentially pointers to the actual file locations on tape, as stored in the OSM database. The OSM software currently interfaces to two Redwood tape drives retrieving tapes from an STK 4410 robotics library (6000 tape maximum capacity, 140 robotic exchanges per hour). With the 50 GB cartridges available for these drives, the current maximum capacity of the silo is 300 TB. An aggregate of 30 MB/second of data throughput must be supported for post-processing to keep pace with the rate of new data collection (e.g. 10 MB/sec each for: new data in; raw data out; processed data in). Depending on the final efficiency factors realized, 6-8 Redwood tape transports will be required to support this level of throughput.

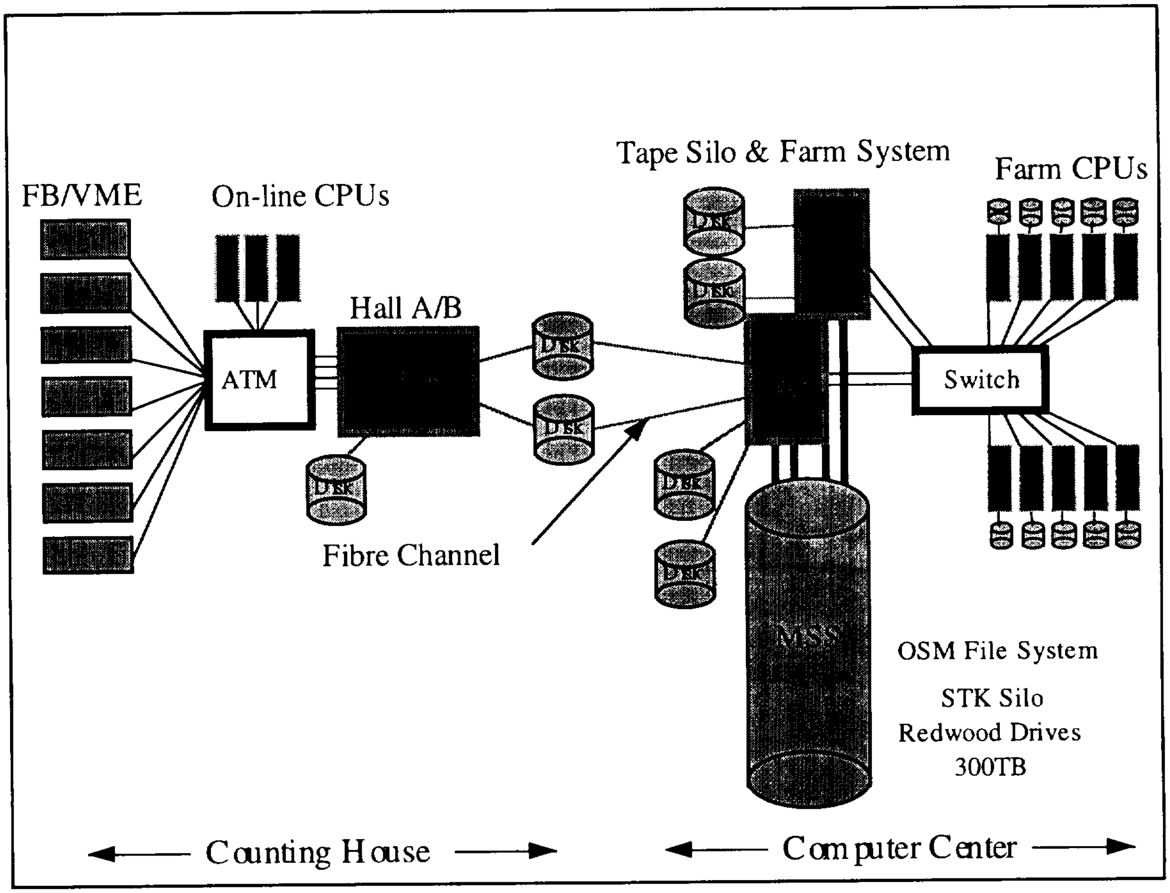

Figure 2: High End Data Flow

Halls C and A will most likely use a network-based data path, also making use of the OSMcp capability. A simple, automated process, parallel to the existing local Exabyte tape copy procedure, has been implemented for current Hall C production. The procedure transfers the data (via remote copies as opposed to NFS) from the Hall C Data cluster to a staging area directly on the Computer Center storage server, and then subsequently OSMcp's the files, leaving stubs in a separate file system. A user utility provides the reverse capability to retrieve data files for post-processing on the existing Data Reduction cluster (3 HP 9000/735s) located in the Counting House. Exabyte and 4mm DAT autoloaders available on this cluster currently provide an export capability for the experimental user.

The Off-Line Model

The Load Sharing Facility (LSF, Platform Computing) has been selected to provide the batch management software base for the off-line CPU farms. The initial design work has begun with the concept of a simple round-robin approach, using LSF to channel jobs to an array of "pizza-box" CPUs (low end, low cost, headless UNIX systems), each with a local disk to hold executable and output files, most likely reading input files from a central tape staging area. While the use of PVM is not ruled out entirely, a coarse grained parallelism with each individual job (submitted, however, in batches) processing one complete run is attractive due to the simplified coding and testing required on the part of the experimental user. A locally developed customized application to track runs and associated files, to pre- and post-stage files in coordination with OSM, and to generate data reduction batch jobs via LSF is under design and will most likely make use of the OpenIngres (Computer Associates) relational database software.

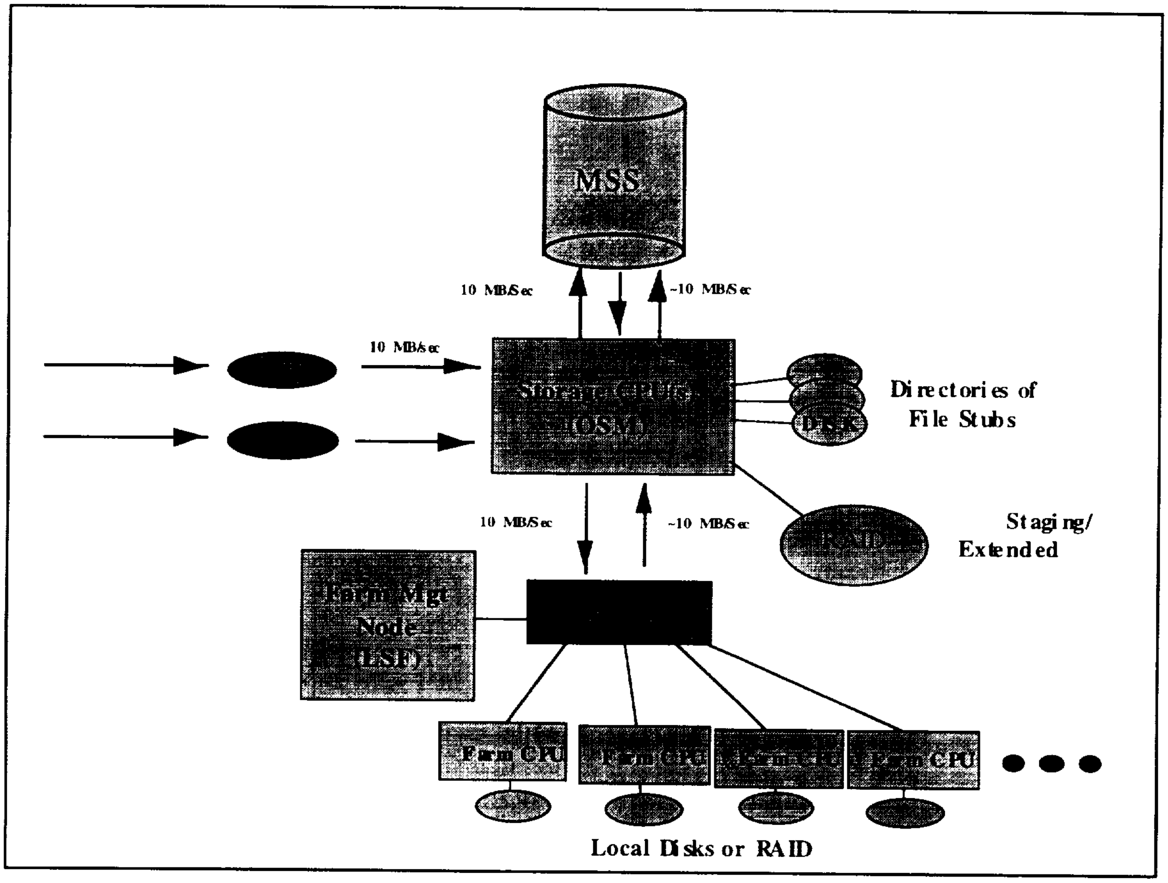

Figure 3: Batch Farm Data Flow

Figure 3, "Batch Farm Data Flow", illustrates the general path of data to and from the data reconstruction CPU farm. Batch jobs, managed by LSF, must coordinate both the retrieval (reverse OSMcp) of raw data and other auxiliary files (calibration files, trigger maps, executables, etc.) as well as the storage of generated output files. OSM is currently limited to a maximum file size of 2 GB. This, plus the economics of local disk storage mandates a maximum raw data file size of 2 GB at the current time. Work is in progress to simulate a variety of alternatives of central and/or local file staging to determine the optimal model considering both data throughput as well as cost. Critical calculations include the number of high end tape transports that will be required to provide both the necessary redundancy for on-line operations as well as sufficient aggregate throughput to pre- and post-stage data for the off-line farms. The right mix and volume of central and local RAID and/or striped disks must address cost, performance, and throughput, as well as simplicity in the algorithm to manage the farm.

Challenges Ahead

Several major challenges face the development team in the near future; the first is to develop the customized software extensions to the file and batch management layers provided by the commercial applications, OSM and LSF. A task force composed of Computer Center staff, Hall, and User representatives is nearing the end of the requirements phase for such a suite of applications. The functionality envisioned will replace manual and even electronic methods to manage the progress of an experiment through the pipeline required from raw data file, through reconstruction phases, to a final Data Summary Tape or even Micro-DST, and provide Jefferson Lab's collaborations with web-accessible, graphical tools to manage and track the entire reconstruction process. Results of prototyping the various data staging models will be critical to the final design of both the disk pool areas and staging algorithms. The developmental milestones include completion of both requirements and design reviews by October, 1996, and the release of a Beta version in early 1997. The procurement of required farm hardware will be concurrent to the coding and implementation phases of the software development. Production experimentation for Halls C and A, and early calibration runs for Experimental Hall B are expected to remain within the data handling capabilities of the existing script-based methods to store and retrieve sets of files from the central mass storage library.

Efficiency and scalability are challenges for every massive data handling operation. Experiences from other similar data collection and replay environments suggests that planning should anticipate overall efficiency rates of no more than 50%, building in sufficient tape, disk, and CPU resources to avoid bottlenecks at any point in the operation. The hard reality in today's research environment is that this approach costs real money--money that from the experimentalist's perspective reduces the amount of research that can be done. In JLab's setting, the hundreds of thousands of dollars that can be saved by tight management of high end disk and tape devices translates into new spectrometer equipment and additional beam time -- more "physics"! The goal, consequently, must be to develop finely tuned, smart systems that can anticipate scheduled requirements, manage resources closely, recover quickly from interruptions, and scale by modular upgrades. Right now, Hall B anticipates a data rate of 10-12 MB/second. Given history, that no doubt will ramp up not down. With the potential of two or three Experimental Halls running simultaneous, high intensity experiments, the initial infrastructure must anticipate an eventual doubling or tripling of the original design goals. Modularization of both hardware and software implementations must allow the upgrade and/or replacement of any one component, from the addition of multiple storage servers and farm nodes, or expansion of on-line storage pools, to the possibility of replacing magnetic-based near-line solutions with other future technologies. The useful lifetime of Jefferson Lab's raw data sets may be 10 years or more. In today's technology, that can represent two or even three implementation life cycles. Today's solutions must prepare not only for tomorrow's requirements but also lay the framework to build with future tools.

Conclusions

Designing for large data handling projects in today's computational environments involves the coordination of network design, tape and disk pool modeling, simulation of processing flows, as well as the detailed consideration of end user requirements and interfaces. The goal of Jefferson Lab's design to support the experimental data handling requirements of the laboratory is to employ modular hardware and software solutions that will scale to meet anticipated future requirements. We have chosen to employ commercial software foundations extended by locally developed applications to coordinate different components of the system. While the current design will provide the immediate capability to handle all facets of collecting and post-processing a new data stream of 10-12 MB/second, our objective must include the scalability to survive not only considerable expansion of the anticipated load, but significant changes in the technological alternatives available.

Acknowledgments

We would like to acknowledge the many individuals at Jefferson Lab whose combined contributions have resulted in the implementation in progress to provide the data handling required for the experimental program; in particular: Dave Doughty (Christopher Newport University), Lan Fan (College of William & Mary), Larry Dennis (Florida State University), David Hibler (Christopher Newport University), Andy Kowalski (JLab), Sandy Philpott (JLab), Dave Rackley (JLab), Roy Whitney (JLab), and Elliott Wolin (College of William & Mary). This work was supported by Department of Energy contract DE-AC05-84ER40150.

References

1 D. Doughty et al, “Event Building in Future DAQ Architectures Using ATM Networks”, Proceedings of the Computing for High Energy Physics Conference (CHEP95), Rio de Janeiro, Brazil, 1995.