CDF/DOC/COMP_UPG/PUBLIC/5802
B. Ashmanskas, S. Belforte, I. Bird, T.A. Davis, F. Donno,
R. Hughes, P.T. Keener, A. Kreymer, P. Murat, A. Reichold,
M. Shimojima, D. Waters, S. Wolbers, F. Würthwein, Ch. Young
December 7, 2001

Final Report CDF CAF Review Fall 2001

Abstract

This is the final report of the CDF central analysis facility (CAF) review from August 15th to November 15th, 2001. This report draws on two additional documents that describe CDF software benchmarking [1] and CDF physics needs expectation [2], as well as a workshop at FNAL on October 18th/19th. This includes the discussion of a straw proposal of how to augment the CAF to meet the expected computing needs.

1 Executive Summary

A committee that broadly represents the CDF collaboration, advised by a set of four external reviewers from JLAB (I. Bird), NERSC (T. Davis), SLAC (Ch. Young), and INFN (F. Donno) took a careful look at the CDF central analysis computing facilities (CAF).

The detailed charge may be found in Appendix A. We evaluated the CAF starting with detailed benchmarking [1] of a variety of CDF applications on a variety of platforms, the results of which were presented to the collaboration at the September 2001 collaboration meeting.

Based on this benchmarking and an initial estimate of the physics needs using the original CAF specs [3] as input we recommended against buying of an additional Sun SMP and focused the efforts of the committee for Charge 2 on evaluating possible PC cluster based CAF implementations. Any architecture that uses cheap commodity PC's as main computing resource is referred to here as a PC cluster. The detailed recommendation in response to Charge 1 may be found in Appendix B.

Based on a detailed physics needs assessment with significant input from the physics groups [7] in conjunction with our benchmarking effort we arrive at the following conclusions:

- We estimate secondary dataset creation in Summer 2002 to take all of an upgraded fcdfsgi2 for one full week if done in an organized fashion. Justification of this number may be found in Appendix E which summarizes Ref. [2].

- If 200 users wanted to analyze one nominal secondary dataset each on an upgraded fcdfsgi2 for Summer 2002 then this would take 6-12 weeks. The higher number corresponds to 50% use of fcdfsgi2 for other purposes than user analysis of secondary datasets. Some of this "other usage", like code management and user code development is unavoidable. Justification of this number may be found in Appendix E.

- It is thus quite clear that even an upgraded fcdfsgi2 (128 CPUs) will not provide sufficient CPU power for the collaboration to produce physics results at the level expected for summer conferences 2002.

- If there were roughly 100 Dual PIII 1GHz nodes available for users to analyze the secondary datasets next summer then the 6-12 weeks above would shrink by a factor 2 (for Duals) $\times$ 4 (for performance) = 8 to a much more reasonable 1 week period. If such resources are not available then most groups within CDF will need to copy full secondary datasets to their home institutions for analysis in order to make data analysis possible for the rest of us. For Summer 2002, a "typical" secondary dataset is expected to comprise 1 million events. There will be 100 such chunks of data.

- For the full Run 2 dataset everything scales by a factor of ten in our simplistic model. I.e. O(1000) Dual PIII 1GHz nodes are needed to maintain a reasonable 1 week turn around for the analysis of secondary datasets and a typical dataset will comprise 10^7 events or 1TB at an event size of 100kB.

Given these findings we arrive at the following recommendations:

1. **We recommend against buying a second Sun SMP.**
2. **We recommend a shift away from large SMP's and towards a large cluster of commodity PC's. For this to be available in time for Summer 2002 analysis, we recommend immediate prototyping of batch, storage, and user interface. In particular, we urge offline management to:**
   (a) **Procede as quickly as possible with acquisition of hardware required for a prototype server plus roughly 32 compute nodes.**
   (b) **Work out a timeline with milestones for implementing a stage 1 PC cluster for summer 2002, including 100 Dual CPU compute nodes, tens of TB disk space, DH connection, batch queue, and user interface.**
3. **We recommend that offline management identify a person that will carefully evaluate the CPU consumption for reading in whatever standard data format the collaboration arrives at. This includes identifying which storable objects consume the lion share of the CPU power and why. The results of this investigation needs to be taken into account when deciding on a final data format that optimizes both disk space and CPU consumption without compromising the physics needs.**

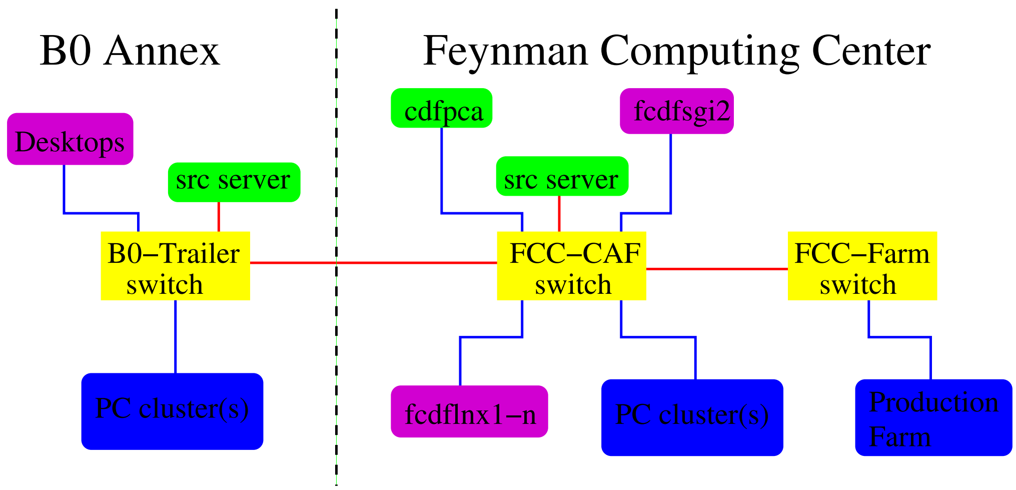

Figure 1: Schematics of potential CAF arrangement. See text for details.

4. **We recommend that offline management work out a detailed plan for how to organize re-processing of data on the farm for summer 2002, and beyond. This is particularly important given that re-procesing of a dataset can not be done as part of a user analysis job due to shortage of available CPU.**

The remainder of this document as well as the accompagnying documents [1] [2] provide more detailed information. In particular, Section 2 describes a straw proposal for a future PC based CAF.

2 Straw Proposal for a future CAF

Figure 1 depicts a possible scenario. Magenta indicates nodes with interactive user access while blue refers to nodes that may be restricted to batch access only. Green boxes indicate restricted access for code management. The yellow squares refer to the three main switches in the current CDF LAN. Blue connections refer to FastEthernet while red ones refer to Gigabit Ethernet. Multiple Gigabit Ethernet conections may be needed on some of these links, especially the one between FCC-CAF and B0-Trailer. Data servers are not shown in this figure. They will require Gigabit Ethernet connections as discussed in Section 2.1.

Conceptually, desktops, PC clusters, and fcdflnx1-n could be integrated by a common batch system augmented with a user front end that would allow specification of a user application (e.g. user exe & tcl file) as well as a destination directory for the job output. The latter may be on the user desktop. The user could thus compile, link, and even "execute" their jobs from their desktop. The user application would be copied to the "remote" PC cluster for "local" execution in a temporary directory. This complete directory could then be copied to the user specified destination directory after the job terminates.

Initially "remote" would probably remain within FNAL and simply mean "not NFS mounted". However, the basic concept described is easily extendable to a remote institution's desktops and PC cluster. The user front end could initially be quite limited, and thus require only modest human resources to implement. E.g. the batch queue name might signify a certain dataset that can be served on a given PC cluster.

Not all data needs to be accessible via each and every such batch queue. In particular, as part of the physics need assessment we identified 5-10nb, i.e. 1-2TB for the full Run2a dataset, as being the generic dataset size. This is a rather convenient size for a single data server. We discuss what this might mean for the client/server ratio in Section 2.1.

Over the next several years additional functionality like the data file catalogue might be integrated into the user front end, and grid tools might be used to implement additional functionality like resource discovery, etc.

One option for a batch system might be FBSNG [6], the batch system already in use on the production farm. FBSNG supports the idea of staged execution including pre- as well as post-processing steps, and is supported from within the FNAL CD. As additional feature FBSNG supports the concept of an "interactive batch process". Using this mechanism a user may submit a request for an interactive shell on a PC in the cluster that executes the batch job. One could imagine providing interactive user access on the PC cluster using this feature. If one envisioned "blurring the boundary" of production farm and PC cluster in the future, e.g. within the context of re-processing data on the production farm then FBSNG might be the most natural choice for a batch system. However, a careful evaluation of the merits of LSF, FBSNG, as well as a variety of other batch systems goes beyond the scope of this report.

2.1 Estimating the client/server ratio

To estimate the number of clients a single server may need to feed we consider the following:

- A Fast Ethernet link provides < 12MB/sec bandwidth, or <120Hz event rate at 100kB event sizes. A Gigabit Ethernet link then provides <1200Hz event I/O rate.
- Typical disk read bandwidth for currently available IDE RAID systems is 50-80 MB/sec [8] per RAID array. A typical server might have two such arrays, thus probably exhausting the maximum sustainable bandwidth of a 32bit PCI bus.
- For multi-branch I/O one might guess that on average only 1/5 of the event will be

accessed. The event rates then increase to <600Hz and <6000Hz respectively.

- We expect user analysis of secondary datasets to be CPU power limited to roughly 4Hz on a PIII 1GHz CPU. For convenience, let's call a Dual PIII 1GHz compute node a 10Hz data analyzer.

We thus conclude that a FastEthernet connection at the compute node provides an order of magnitude more bandwidth than a Dual PIII 1GHz provides CPU power. This doesn't really change significantly even when we consider more optimistic CPU need scenarios discussed in Ref. [2] as those generally tend to assume smaller event sizes (i.e. different data formats) as well.

A Gigabit Ethernet connection on the data server is thus likely to sustain the bandwidth needs of up to a few hundred clients. We thus conclude that neither client nor server bandwidth is a major concern in architecting a client-server based CAF. Instead, there is probably going to be a limit to the number of simultaneous clients that is more related to harddrive head movement.

Alternatively, one might assess the needs based on expected usage patterns. If each server was dedicated to a single dataset, and we have 200 simultaneous users accessing 50-100 datasets then the average client/server ratio would be more like 2-4. It is thus clear that the actual client/server ratio ought to be controllable well below the yet unknown hardware limitations as long as data management isn't done too mindlessly.

Aggregate LAN bandwith for a CAF of O(100) servers plus O(1000) clients is clearly an issue that needs to be dealt with. Partitioning of the LAN according to groups of datasets may be desirable.

3 Data Handling Requirements of a future CAF

Apart from issues regarding computing power that are the focus of this review there are also a number of demands on the data handling system that we believe are essential. The list below is more or less prioritized from the top down:

- handling of multiple disk staging areas.
- more general "data movement tools":
  - tape to networked disk
  - networked disk to tape
  - disk to disk staging as well as buffering where buffering is simply staging from a large to a small disk volume such that one file at a time is moved, for example.

- fast access to raw data
- fast access to production output without the need for intermediate saving to tape.
- handling of non-Edm data: stage in, catalogue, stage out also by data set name.
- data replication/striping across disk servers, i.e. tools for populating "static disk sets" (fully, partial, partial after disk failure). E.g. there may be a need to be able to respond quickly to high demand for a certain dataset by replicating it on a set of "fallback" servers.
- path towards grid for Run2b.

4 Acknowledgements

We greatfully acknowledge the substantial input we have received from many people within the CDF collaboration as well as the Fermilab computing division, D0, CMS, and SDSS. In particular, we want to thank all speakers at the CAF workshop, as well as Andy Beretvas, Ray Culbertson, Rob Kennedy, Pierre Savard, Willis Sakumoto, the CDF task force and Paul Hubbard, as well as many others we've probably forgotten to mention here. In addition, we'd like to thank CDF offline management for many fruitful discussions and courageous as well as speedy adoption of many of our detailed recommendations.

A Charge of the Review

A.1 Charge 1:

```
CDF has purchased a Sun Fire 6800 computing system with 24 750 MHz SMP nodes
capable of being seen as a single system image. The nodes will be
upgraded to 900MHz in September. CDF's plan is to purchase a second identical
Sun Fire 6800 system in FY01 or FY02, if the performance of the first system
with the CDF software is adequate.

The most immediate charge to this committee is to review the performance of
this Sun system, fcdfsun2, and recommend to the offline whether or not
they should proceed with the purchase of a second sun system. This should
be done in the context of specific performance requirements driven by
expected Run II datasets that will be available (a) for summer 2002 analysis
(300pb-1) and (b) for completed Run IIa analyses (2fb-1).

fcdfsun2 should be available to users near the end of August. The committee
```

should begin by evaluating the performance of running the offline software on fcdfsun1, a four processor 400 MHZ machine, and extrapolate to the performance of fcdfsun2. Before the completion of the report, this extrapolation should be explicitly verified on fcdfsun2.

If the committee decides to recommend against the purchase of a second Sun machine, the committee should suggest alternative uses for the funds that were earmarked for the second Sun, approximately $$400K, within the central systems.

A.2 Charge 2:

The committee will review the CDF central analysis systems and recommend directions for FY02 and FY03. The committee is invited to review both the current technical choices within the CDF centralised computing model, and the model itself, and recommend directions either inside our outside the current model. The committee may consider some or all of the following:

1. The existing and planned central analysis systems: fcdfsgi2, fcdfsun1, fcdfsun2, fcdflnx1.
2. The storage and CPU resources available to these and future systems.
3. Recommended UNIX flavors in the central systems in the future, restricted to the three possibilities IRIX, Linux and Solaris.
4. Possible future farm-like systems for batch driven central analysis.
5. Possible future distributed computing environments.

The report on this "long term" charge will provide important input for planning for RunIIb, which will commence late this fall.

B Recommendation regarding Charge 1

Based on detailed performance measurements (CDF 5743) on fcdfsun2, in particular in comparison to a variety of Intel/Linux systems the committee recommends not to purchase a second Sun SMP.

Moreover, due to significant set-up problems and overheads associated with maintaining SunOS, the return of fcdfsun2 to Sun is being actively pursued. It should be noted that the approximately $$400k that this would result in is EXTRA money, NOT the $$400k referred to in Charge 1 above. It is therefore strictly outside the remit of this committee to recommend how to spend this additional money.

Nevertheless, the committee has discussed with the offline management how these additional funds could be spent. The key requirement is that the money be spent in such a way that it helps to meet CDF's central analysis facility computing requirements in time for data analysis prior to the 2002 summer conferences. The option with the smallest risk in this regard is the upgrade of fcdfsgi2 with extra processors. In particular, doubling the number of processors with another 64 300MHz to the tune of $$408,000 gives the best value for money measured in MHz per dollar for an SGI solution.

Several members of the committee feel that, given the deployment of sufficient manpower, an Intel/Linux solution could be devised on a similar timescale with a high probability of success and providing much greater benefits in terms of computing power. However there are serious issues to be addressed relating to the management of very large disk volumes and the use of the data handling system in the context of an Intel/Linux central analysis facility.

The committee does not currently have a recommendation for how to spend the $$400k earmarked for the second Sun and referred to in Charge 1. Work is underway to arrive at this recommendation, which is naturally focussed on how Intel/Linux systems can be best employed to meet our computing needs. We consider this an integral part of our second charge and will explicitly comment on it in the forthcoming recommendations with respect to charge 2.

In detail, we arrived at the following conclusions in CDF note 5743:

* There are still significant setup issues for fcdfsun2 to be resolved.
* We consider the porting of CDF software to the sun platform at best incomplete and inefficient.
  If we were to believe that the port is efficient then we would have to conclude that SunOS itself is rather inefficient. Whether or not this is an issue related to CDF's use of KAI instead of a native compiler is impossible for us to assess.
* Performance improvements with maxopt are presently more at the level of roughly 30% than a factor of 2. It is conceivable that this might improve in the future when the focus shifts from algorithm development to performance tuning.
* Reading of storable objects in CDF is severely CPU limited,

```
   and differs between different objects by at least a factor
   about 3. CdfTrack was shown to have particularly large
   CPU needs.
*  Run2 analysis modules (i.e. ignoring I/O issues mentioned
   above) require roughly an order of magnitude more CPU
   power per event than similar modules in Run1. We do not
   understand the origin of this discrepancy.
*  We recommend that the ongoing design of the PAD format
   take CPU consumption as well as multi-branch I/O into account.
   In particular, a PAD format that shrinks event size at the
   cost of CPU power, or combines essential with non-essential
   information into the same storable object is unlikely to
   be optimal.
*  In light of the apparent difference in performance of run1 and
   run2 user analysis skim software we feel compelled to point
   out the cost differential in $$/MHz of processor power in
   an SMP (roughly $$20/MHz) versus a dual processor PC (< $$1/MHz).
   We expect this difference to increase in the near
   future as dual P4 become as common as dual P3 are at present.
```

C CAF workshop

To allow for detailed discussions of the second charge a two day workshop was organized on October 18th/19th the agenda of which can be found in Appendix D. The workshop focused on answering the following questions:

1. **Q: What level physics analysis computing will be possible in summer 2002 if all we have available is a 128 processor fcdfsgi2 to access the data?**
   A: If an average analysis job could process 100 events/sec on fcdfsgi2 then a 128 processor fcdfsgi2 might be sufficient to satisfy the physics needs for summer 2002. At present this would require ntuples as secondary data format and restrict analysis to root scripts that aren't too sophisticated. More details can be found in Appendix E. People would not be able to go back to PADS and redo tracking in their analysis jobs. At least not on the CAF.

2. **Q: How is user analysis computing handled at JLAB, PDFS, and SLAC? What can we learn from their experience?**
   A: All three places have data served by a small number of nodes to a large number of nodes. The largest such cluster is at SLAC where data is served to close to 2000 CPU's. Roughly 1/2 of these are housed in dual-CPU PC's the other half in single CPU Sun's. All of these are pure batch worker nodes. Some of them are specialized for building executables in batch. Interactive access is provided by two clusters, one

consisting of 10 quad-CPU Sun the other of 10 dual-CPU Linux PC's. Those are used for debugging, running small samples, and other interactive use. Submission to the large cluster is mostly done from the interactive nodes using a set of LSF batch queues. CPU power is "partitioned" such that never more than $\sim$ 20% of the clients can mount disks from any given server. Clusters at JLAB and PDFS comprise a few hundred CPU's and a few tens of data servers. A single data server is typically a dual-CPU PC with up to 15 IDE disks attached. Both JLAB and PDFS use IDE disks rather than SCSI or Fibrechannel for cost reasons. A detailed cost comparison can be found in Appendix F. However, both point out that the disk controllers they used are no longer available as the relevant companies, Raidzone and 3ware, are specializing on selling complete systems rather than components. **Note added: 3ware appears to have changed its mind. They continue to sell their RAID controller cards, at least for the time being.** *Detailed testing of disk controllers from the one remaining company, Promise Technology [5], that sells appropriate controller cards would be beneficial to more than just CDF. The clusters at JLAB and PDFS are maintained by 0.3 and 2 FTE respectively. For PDFS this includes installation and purchasing expertise & support, as well as general system administration like account maintenance, security, etc. . For JLAB it is the FTE needed to keep the cluster running. The JLAB number would thus be applicable if one needs to account for only the additional hardware maintenance, while the PDFS number is relevant if one were to start a new subproject within CDF offline operations and had that subgroup take care of all aspects of the PC cluster. See below for further discussion of human resources.*

3. **Q: How do other experiments at FNAL (e.g. D0, CMS) satisfy their user analysis computing needs? What can we learn from their experience?**
   A: The most comprehensive design is the one from D0. It is based on the concept of a "SAM station" which could for example be a set of PC's forming a small cluster. Files are transferred to such a station using a global data file catalogue to identify the location of a file. At present, analysis requires movement of files rather than movement of jobs. In the future both is envisioned. The D0 SAM concept is expected to incorporate grid tools as they become available. The SAM development team comprises roughly 5.5 FTE's. Regarding system administration D0 has also found an interesting modus operandi. Desktop computing is maintained by a collaboration between Dave Fagan from CD and a number of University volunteers who co-ordinate maintenance of systems on site.

4. **Q: Are there any infrastructure or resource limitations in the way CDF computing is organized right now that will make a PC cluster based CAF difficult, or impossible? At what price can they be overcome?**
   A: Apart from issues with respect to data handling mentioned below, we can identify the following potential bottlenecks:

   - *The CDF software server ncdf09 insufficient as it is only a 300MHz CPU, 1/3 GB RAM, and FE LAN. Compile & link time for cdfSim is 1.5 times longer using ncdf09 than a software distribution on a local disk. We understand that CDF* *compiling and linking is in principle CPU limited on Linux. A central software server in the B0 annex with sufficient CPU power and a Gigabit ethernet link to serve releases to all desktops in B0 annex is thus likely to be popular among desktop users.*

- *fcdflnx1 is insufficient as general user Linux platform, and is lacking data access.*
- *One or more CDF software servers in FCC to guarantee a uniform code platform on fcdflnx1 and any future PC cluster are likely to be needed.*
- *While the Gigabit link from B0 annex to FCC is presently not a bottleneck one should not wait for it to become one. An upgrade with an additional Gigabit link is likely to be needed in the future.*
- *The issue of how to re-process data in general, and data on a future PC cluster in particular needs to be thought through more carefully. If one were to decide that data on the PC cluster is best re-processed on the production farm then the FastEthernet crossover between the CAF and Farm switches in FCC needs to be upgraded to Gigabit ethernet in order to allow for data movement from the PC cluster to the production farm.*

We think that all of these limitations can be overcome rather easily. We thus see no fundamental obstacles other than human resources discussed below that would make a PC cluster based CAF impossible.

5. **Q: What resources and synergies are there within the FNAL CD that would facilitate implementation of a PC cluster based CAF?**
A: Steve Timm in OSS is presently testing two categories of PCs, desktops and farm nodes, on a regular basis. It seems that adding a category "data server" might be beneficial to a large number of groups at FNAL. Such a data server category would essentially be a farm node with a minimum of 4 IDE disks attached. To give an extreme example, a Dual Athlon 1800MHz system with 2 SuperTrak SX6000 cards from Promise, 2×6 Maxtor D540X 160GB drives, SuperSWAP hot swappable chassis from Promise, and arranged as RAID 5 would make a very interesting system to test if a chassis could be found with sufficient space for all the drives. The batch system in use on the production farm is another example for potential synergies. It was developed and is maintained by ISD in the FNAL CD. In addition to FNAL CD, there is significant experience with Fibrechannel and GFS in SDSS, and a shared interest in O(TB) PC fileservers and Mosix in CMS.

6. **Q: To what extend is it conceivable to have the full 200TB of run2a PAD data on disk, and served via NFS?**
A: The most important shortcoming of the present computing model is its lack of flexibility. I.e. the model is built around a very specific hardware choice, and its success or failure depends to a large degree on the assumption that CDF physics analysis is I/O rather than CPU limited. Being wrong once ought to make it clear that limiting one's flexibility in the future is a bad idea. Any computing model whose success depends on data being purely disk resident should thus be avoided. Some fraction of the total disk space available via a PC cluster needs to be accessible via the data handling

system. However, it is reasonable to expect that not all of the available disk space is equally accessible. E.g. serving a secondary dataset from largely static disk space in B0 annex should not be ruled out. Apart from these strategic considerations there are also significant technical as well as maintenance challenges to serving 200TB of disk space. E.g. at a mean time between failures of ~ 15 *years and 100GB capacity for a single disk, one ought to expect a disk failure every 2-3 days. And given recent experience with a particular disk vendor it is clear that the actual mean time between failures can be significantly smaller despite claims to the contrary. From the perspective of management and maintenance one would almost certainly want large servers with hotswappable disks and probably at least RAID 5 to allow for recovery without repopulation from tape. The architecture would thus be more that of a SAN (storage area network, possibly using fibrechannel disk arrays) behind every server rather than vast arrays of commodity IDE disks. See also Appendix F for a related discussion. The most significant technical problems are a result of 2TB filesystem limit in Linux.*

7. **Q: What impact does a PC cluster based CAF have on human resources within the CDF task force?**
   A: The CDF task force was busy even before the ongoing restructuring of the data handling system. As part of this restructuring the task force has taken on additional responsibility within data handling operations. It is thus obvious that the CDF task force can not play a major role in an initial development and implementation of a PC cluster based CAF. We thus urge the CDF collaboration to identify collaborating institutions who are willing and capable of taking on the development and initial implementation as an institutional responsibility. Once a PC cluster based CAF is up and running we would expect that routine maintenance can be picked up by the task force. We understand that the CD intends to hire an additional Linux system administrator to be added to the CDF task force. Based on the experience at JLAB and PDFS we expect that this one additional person should be sufficient to guarantee long term routine maintenance of a future PC cluster.

Based on the answers to these questions we identified the following general features that any PC cluster based CAF redesign should strive for. It is reasonably obvious that some of these goals may not be achievable for some time.

- Make maximal use of cost effective commodity components.
- Given the uncertainties in the expected physics motivated computing needs any design has to be easily scalable.
- Any design needs to be flexible enough to adjust to new hardware. The conceptual design should not be self-limiting by being focused around a particular technology.
- Make maximal use of the desktop. E.g. it would be desirable if the user could compile, link, and debug their analysis executable on their desktop, then run it on a large dataset on a PC cluster, and analyze the resulting ntuple again on their

desktop. An ideal solution would include data access from the desktop. However, there are significant security, stability, and maintenance concerns with NFS mounting data disks to the desktop. In addition, once one gives up on mounting data disks to desktops all desktops anywhere in the world are conceptually the same. A solution that works for those in B0 ought to be easily extended to one located in Italy, for example.

- A central linux platform with interactive login is essential. This may be thought of as a "central desktop" facility rather than a "central analysis" facility. I.e. this interactive resource is in addition to the principal batch driven computing resources.
- To simplify maintenance a central CDF software release server for the PC cluster is desirable.
- To allow for efficient release building a dedicated build node like cdfpca is desirable. We do not see any reason to have more than one such node. Releases can be copied to more than one release server after they are built on the build node.
- The design should be flexible enough to easily accomodate more than one PC cluster in more than one location. In particular, we forsee that some University groups may want to integrate (some of) their own computing resources in the B0 annex into the overall CAF scheme. A conceptual example of this sort of arrangement is the concept of a "SAM station" in the D0 user analysis scheme. An actual implementation may be quite different, though. In particular, for CDF the focus ought to be on moving jobs rather than data.
- We recommend that the collaboration adopt a model where CDF computing and data handling is slowly "gridified". We expect that this would best be done in some evolutionary fashion, and largely transparent to the user. It is clearly mandatory on the timescales of run2b in order to exploit computing resources in Europe and elsewhere.
- The LAN links between the farm switch and the CAF switch in FCC, as well as the Gbit link between FCC and the B0 annex are likely to require updates to support significantly larger bandwidths. For the CAF - Farm link this is needed to be able to send data from the PC cluster to the production farm for re-processing. For the B0 Annex - FCC link it is needed especially if some PC cluster(s) are located in the B0 annex.

D Workshop Agenda

```
***************** THURSDAY, OCTOBER 18TH **********************

Time:           Topic:                  Person:
8:30am     Intro & where we stand        fkw
```

	(this incl. benchmarking) what we want from this workshop	
8:45am	Budget & Human Resources	Rob Harris 10&15min disc.
9:10am	DH review recom. charge 1	Pekka Sinervo (via video) 15&15min disc.
9:40am	JLAB analysis computing	Ian Bird 15&15min disc.
10:10am	NERSC/PDFS analysis computing	Tom Davis 15&15min
10:40am	Coffee break	
11:00am	Physics needs report	Savard/Belforte 30&30min
12:00am	Discussion time	30min
12:30pm	Lunch	
1:30pm	CDF computing & networking infrastructure	Steve Wolbers 20&10min disc.
2:00pm	CDF options considered	Glenn Cooper 15&15min
2:30pm	PC Farm testing & maintenance	Steve Timm 15&15min
3:00pm	Activities in ISD	Igor Mandrichenko 15&15min
3:10pm	Discussion	
5pm	End of our time in Theory room	
7pm	Discussions until we're tired.	

***************** FRIDAY, OCTOBER 19TH ***********************

Time:	Topic:	Person:
8:30am	Serving 200TB via NFS What are the issues?	Tom Davis 15&15min disc.
9:00am	Mosix: moving jobs to data instead of data to jobs.	A.Korn & fkw 30&20min disc.
9:50am	Fibrechannel & GFS @ SDSS	Jim Annis 15&15min disc.
10:20am	Coffee	
10:40am	Discussion Session I.e. wrap up Run2a in this session.	
12:00 Noon	************ Lunch *******	
1:30pm	CMS: What's on the floor & near term plans	Hans Wenzel 15min+15min
2:00pm	UK-grid for run2b	David Waters 15min+15min
2:30pm	D0 SAM	Vicky White 15min+15min
3:00pm	Discussion of Run2b related issues.	

E Physics Needs Summary

The present section summarizes the estimated computing needs based on the expected dataset sizes, # of physicists doing analysis, and the CDF software performance as determined in Ref. [1].

Let us start with a list of useful numbers and assumptions.

- A 75Hz L3 accept rate at a luminosity of $10^{32} cm^{-2} sec^{-1}$, i.e. 1/2 the Run2a design luminosity, and an order of magnitude larger than the highest luminosity achieved so far, translates into a 750nb L3 accept crossection. Out of these 750nb the TDWG document CDF 4718 has currently aloted 556nb specifying 46 different datasets.

- 750nb $\times 2fb^{-1} = 1.5 \times 10^9$ events. We thus expect the Run2a primary physics datasets to comprise roughly 2×10^9 events total.

- 5nb means 10^7 events for Run2a occupying 1TB of disk space at an expected event size of 100kB. We find that this is roughly the average crossection for typical datasets in CDF 4718. Taking this and the 556nb face value we will work under the assumption of 100 datasets of 1TB each, and assume that this applies to both primary as well as secondary datasets. We consider this to be at the low end of expectations. I.e. this could easily be underestimated by a factor of two.

- For simplicity we take a day to have 10^5 seconds.

- We differentiate two types of analysis activity. Creation of secondary datasets is known/expected/estimated to proceed on fcdfsgi2 at a rate of 2-3Hz. User analysis of secondary datasets is expected/estimated to proceed on fcdfsgi2 roughly a factor of 3 slower. In comparison, ProductionExe 3.18.0 takes 7 seconds per event on fcdfsgi2. I.e. the analysis examples we consider are clearly not allowed to rerun significant parts of the reconstruction software!!! Creation of a secondary dataset of 10^6 events will thus take roughly 1 CPU-week on fcdfsgi2. User analysis of such a secondary dataset is estimated to take roughly 3 CPU-weeks.

- We take as nominal CPU a 1GHz Pentium III which was measured to outperform the CPU's on fcdfsgi2 by roughly a factor 4 on a variety of CDF applications. We use this unit whenever we estimate CPU needs beyond fcdfsgi2.

- One might estimate the data sample size for summer 2002 ($200pb^{-1}$) as 1/10 the full sample based on the integrated luminosity or 1/4 the full sample based on running 6 months at 30Hz L3 accept rate. We base our needs assessment for summer 2002 on 1/10 the data sample.

- We assume 200 active simultaneous users of the CAF. We are fully aware that this number depends on the amount of frustration the users will have to endure. More frustration, fewer users, more users who just copy complete secondary datasets to their home institutions.

Using these "facts" we estimate secondary dataset creation in Summer 2002 to take all of the upgraded fcdfsgi2 (128 CPUs) for one full week. If each user wanted to analyze one nominal secondary dataset on fcdfsgi2 then this would take 6-12 weeks. The higher number corresponds to 50% use of fcdfsgi2 for other purposes than user analysis of secondary datasets. Some of this "other usage", like code management and user code development is unavoidable. It is thus quite clear that even an upgraded fcdfsgi2 (128 CPUs) will not provide sufficient CPU power for the collaboration to produce physics results at the level expected for summer conferences 2002.

If we are to assume that there were roughly 100 Dual PIII 1GHz nodes available for users to analyze the secondary datasets next summer then the 6-12 weeks above would shrink by a factor 2 (for Duals) $\times$ 4 (for performance) = 8 to a much more reasonable 1 week period. If such resources are not available than most groups within CDF will need to copy full secondary datasets to their home institutions for analysis.

For the full Run 2 dataset everything scales by a factor of ten in our simplistic model. I.e. O(1000) Dual PIII 1GHz nodes are needed to maintain a reasonable 1 week turn around for the analysis of secondary datasets.

F Disk space "architecture"

At the workshop we discussed a variety of architecture models for how to distribute disk space such that a cluster of PC's has access to a large disk volume. The variants can be grouped into three basic categories:

1. NFS client-server
2. Fibrechannel SAN
3. Mosix

In the present section we contrast these three approaches in the hope that their respective strengths and weaknesses are clarified. We do so by describing the prototype "building blocks" and estimating their cost.

F.1 NFS client-server

A few servers serve large disk volumes to many worker nodes. The worker nodes automount the disks from the servers. The total pool of worker nodes may be partitioned such that only a limited number can access any given server. This partitioning might be controlled by the batch queue arrangement.

Typical worker node: Dual-CPU PC, limited local disk space, FE, $\sim$ \$1200 per node.

Typical server node: Gigabit Ethernet (\$400), Dual-CPU CP (\$ 1200), 2$\times$ SuperTrak SX6000 hardware RAID 5 card from Promise Technologies (2$\times$ \$ 470, incl. 3 hotswap chassis), 4 additional SuperSWAP chassis (4$\times$ \$ 90), 10 IDE disks (\$ 2500). This all adds up to close to \$ 6000 per 1 TB.

F.2 Fibrechannel SAN

Storage as well as compute nodes are connected to a Fibrechannel switch. The storage units attached to the switch are likely to be already Fibrechannel arrays in themselves while the nodes are each with their own link such that all have equal access to the SAN.

Switch ports: \$ 1200-1600 for 8-64 ports total. More ports means higher per port price. Each fibrechannel card per node \$1000 . Fibrechannel disk comes to roughly \$ 12,000 per 1TB.

Server equivalent: $\sim$ \$ 13,400 per 1 TB.

Typical worker node: $\sim$ \$ 1200 + \$ 2400 = \$ 3600

F.3 Mosix

All nodes have roughly equal amount of disk space. No distinction is made between server and worker node. The price overall is the same as for the NFS case except that the disks are distributed across all nodes. One of the key strengths of a Mosix cluster is that jobs migrate to the data rather than serving data to the jobs. This is the less important the larger the ratio of CPU need per I/O .

F.4 Discussion

Looking at the numbers above it seems clear that a Fibrechannel SAN is cost effective only if the number of compute nodes per I/O is small. It is thus an obvious solution if the CAF is based on a few large SMP's but probably inappropriate if it is based on a few hundred dual-CPU PCs. The cost per MHz compute power is roughly an order of magnitude larger for 8way Linux/Intel than 2way Linux/Intel. Fibrechannel thus makes sense only if disk space needs per MHz are large.

The limit for which fibrechannel makes the most sense is exactly the limit for which an NFS client-server arrangement is neither cost effective nor practical. After all, it makes little sense to have one \$ 6000 server for every \$ 1200 worker node.

A Mosix farm reduces effectively to a client-server arrangement for practical reasons if disk space per CPU is very low. It has its biggest advantage over NFS client-server if disk space per CPU is large.

From this we conclude that a combination of NFS client-server and Mosix covers the full range of disk space vs CPU and both architectures may be implemented using essentially the same hardware.

References

[1] Report on benchmarking, CDF CAF Review Fall 2001; CDF note 5743.

[2] Physics Analysis Computing needs assessment, CDF CAF Review Fall 2001; CDF note 5787.

[3] CDF note 4100, March 1997. This note specifies the CAF design requirements. This note assumes: 1 MIPS = 1 Tiny = 1 SPECint92 = 1/40 SPECint95

[4] http://www-cdf.fnal.gov/upgrades/computing/projects/central/workshop/

[5] http://www.promise.com

[6] http://www-isd.fnal.gov/fbs/FBS2/

[7] Special thanks to Willis Sakumoto, Pierre Savard, and Ray Culbertson for their input.

[8] Thanks to Patrick Schemitz for providing us with performance measurements on RAID5 arrays using 3ware Escalade 7580 controllers. For details see cdf_comp_reps listserv archive.