Second Large Scale Cluster Computing Workshop

21st-22nd October 2002

Fermilab National Accelerator Laboratory

Proceedings

Alan Silverman
First Published 10 March 2003
Revised 30 June 2003

Introduction

This was the second in this series, the last having taken place also in Fermilab in May 2001[1]. For this second meeting, two themes had been set – experiences with real clusters and technology needed in building and running clusters. There were some 100 attendees. The meeting lasted 2 full days with no parallel sessions and was summed up on the last day by two *rapporteurs.*

HEP[2] Computing and the Grid[3] - Matthias Käsemann (FNAL)

Considering only one LHC experiment as an example, very physicist in ATLAS must have transparent and efficient access to the data irrespective of his or her location with respect to that of the data. Computing resources are available world-wide and physicists should be able to benefit from this spread by making use of Grid projects.

An early example of such distributed computing is the BaBar experiment at SLAC with its tiers of centres across the globe. Similar world-wide processing chains exist already for other current experiments and for the early testing for future generations of experiments. But such schemes are not easy to setup and you need to include the costs not only for investment but also for the human resources for setup and operation of these centres and this latter must include the cost of developing the techniques needed to make such distributed schemes interoperate.

Given the accepted advantages of Grid computing, why now? Since recently, network performance has made distributed computing feasible and so has offered us an opportunity not previously available. Today's networks make interconnection possible at affordable prices.

Grid computing is not new (I.Foster and C.Kesselman published "The Grid – Blueprint for a New Computing Infrastructure" already in 1999) but nowadays tools start to exist which make it possible to create realisable world-wide virtual organisations. There are many examples of grid projects now in development or even in some cases starting production. There is considerable consensus on grid standards and methods. Significant research and development (R & D) funding is being applied to produce grid middleware but it is important to coordinate these efforts to avoid diversity and to promote inter-operability. HEP is not the only, or even the main, driver but we need to participate to benefit from this new environment.

Turning to the LHC Computing Grid (LCG) project at CERN, to cope with the needs of the LHC experiments, no single HEP centre could satisfy the demands where computing is 10-20% of the cost of a modern experiment. CERN is not even the largest centre today associated with at least one LHC experiment (CMS). Grid technologies

[1] The programme of the first workshop, most of the overheads and the full proceedings are available from http://conferences.fnal.gov/lccws/.
[2] HEP – High Energy Physics
[3] Grid – see for example the Grid Computing Info Centre

are needed to build and share such resources on a world-wide scale but it comes back to the provision of basic clusters at individual centres and workshops like this one are necessary in order that cluster managers get together to share experiences and learn from each other how to collaborate and build and offer to the end-users the computing facilities demanded.

BaBar Computing – Stephen Gowdy (SLAC)

The BaBar Collaboration spans the globe with 560 people in 76 institutes in 9 countries arranged in multi-tiers of 3 levels. In order to permit members of the collaboration around the world to participate fully in software development, they have adopted tools such as AFS for the file system and CVS for controlling remote developments. In collaboration with Fermilab, they use tools such as SoftRelTools and SiteConfig for the same purpose.

There are 4 Tier A centres; no Tier Bs in use in practice. And Tier Cs are small sites (about 20 of these) or personal computers or workstations. Tier As centres include SLAC itself plus RAL in the UK, IN2P3 at Lyon and INFN at Padua. There is a proposal for a new Tier A centre in Germany which would initially be for Monte Carlo but later also for analysis.

To perform BaBar analysis, the SLAC centre has about 900 SUN systems plus 2 Redhat Linux PC clusters, one of 512 PCs from VA Linux (no longer in the hardware business) and 256 newer PCs. All servers are SUNs, including 50 for data plus another 100 for various tasks. SLAC stores a copy of all raw data and extracts can be found at the other Tier A centres as required.

CCIN2P3 (Lyon) was the first Tier A centre, the centre is shared among some 30 experiments, not only HEP. BaBar has some 11 SUN data servers with 22TB of disc. This means they rely heavily on HPSS to access this data because of the need to stage more data from near-local storage. The worker nodes are mostly IBM eServers.

The INFN Tier A centre at Padova has an LTO[4] robot which will be used to backup a copy of all data in case of series seismic activity at SLAC! The fourth remote Tier A centre is at RAL where analysis-level data is available via ROOT. It is currently suffering from severe disc problems and needing to replace all the discs in the farm.

Data reconstruction tasks include prompt reconstruction as the data comes out of the detector and offline reprocessing with the latest calibration data. The primary format is Objectivity, while remote sites tend to use a ROOT-based scheme known as Kanga (Kind ANd Gentle data Analysis). The speaker showed how the different analysis steps are being performed across the various sites and how the next data run will be handled. As luminosity rises, they will have a choice as to where to add capacity, Padua or SLAC. And under-used cycles in any of the sites can be made available on the analysis and data production farms to generate Monte Carlo simulation statistics. They found it was necessary to add more memory for this, going up from 1GB to 2GB per node. Much of the Monte Carlo simulation data comes from the RAL centre.

Three “Grid-type” applications are in development. Simulation is the first where centrally-managing the different Monte Carlo production sites could reduce the human resources needed to manage each of them independently. Analysing the data via the Grid could lead to better use of world-wide resources and allow users to collaborate better. The third potential Grid application where BaBar sees an interest is for data distribution.

All farms are currently Pentium PIIIs or SUN SPARCs. They found that moving from a P3 Pentium at 1GHz to a 2GHz P4 showed only a 30% increase but someone in the audience claimed that can be a chip set effect. There are currently no plans to move away from Objectivity. They rely on tools such AFS, CVS and SoftRelTools.

[4] LTO – Linear Tape Open

Computing at CDF - F.Wurthwein (MIT/FNAL)

CDF Run II collaboration comprises some 525 scientists in 55 institutes across 11 countries. Their computing environment allows more than 200 collaborators to perform physics every day. These facilities include a reconstruction farm for data reconstruction and validation, as well as Monte Carlo event generation. It has 154 dual Pentium P3s, uses Fermilab's home-grown FBSng for job management, running a single executable program. Data is handled by ENSTORE (described briefly as network-attached tape store) using STK9940 drives in a robot. Oracle DB is used for metadata and ROOT for the data itself. The data storage has been controlled by a dual SUN system but they recently added a Linux-based quad-processor because of bottlenecks. They are evaluating MySQL for the metadata where MySQL would avoid the need for licensing, especially beneficial for offsite users.

ROOT I/O is used for raw and processed data. Analysis jobs run on 1GHz PCs quite well. CDF purchase mostly AMD Athlon CPUs currently because of disappointing Intel Pentium P4 results. The experiment has 176TB on tape as of 20th October but expect to grow to 700TB of disc space by end-2005 and they are looking at IDE discs.

The past CDF Analysis Farm (CAF) architecture was based on a mainframe-like model with a large 128 processor SGI but it has become too expensive to maintain and operate. They began to look at a new computing model where users would develop and debug their jobs interactively from desktop. They would then build execution jobs and dispatch them to a central farm. But there should be no user accounts on the farm, only a scheme to give users access to temporary storage space with quotas. Code management would be needed for the desktops, a Kerberos gatekeeper for job submission and FBSNG and ENSTORE on the central cluster.

CDF created a design team to build this architecture including experts from outside labs. The farm consists of 132 CPUs worker nodes plus some servers to look after 35TB of disc space. They have achieved 200MB/s for local reads and 70 MB/s NFS reads. They have had some performance issues (resource overload, software bottlenecks, and memory hardware errors) which have required workarounds.

Statistics show that current cluster capacity is being used at about 80% but they don't expect much more because the jobs are not very long, typically 30 minutes and job startup and rundown reduces the overall efficiency. Nevertheless, an upgrade is pending. For Stage 2 they will upgrade by 238 dual Athlon PCs and 76 file servers. They transfer some 4-8 TB of data per day and they plan to integrate SAM[5] for global data distribution. One problem worth mentioning was with their discs where expansion of the initial purchase was complicated because the original vendor went out of business.

Access to the temporary staging area on the farm in stage 1 used NFS to access data but they are now moving to dCache[6] after tests showed this is now reliable and stable. The advantages expected include more automation of data replication but not necessarily better performance. They have noted resource overloads in dCache usually related to simultaneous access to single servers or single data stores where the solution often implies distributing the target data over several nodes.

From the user's view he or she needs not only to submit and kill a job but also to monitor its progress. CDF built an interface in Python to a server process which can perform simple UNIX commands to manipulate files and processes in the context of the executing job.

Unlike BaBar, CDF is based on a single farm, at FNAL. But many users across the world need access so a. remote replicated models of CAF (DCAF) are planned.

[5] SAM stands for Sequential data Access via Metadata.

[6] dCache was jointly developed by FNAL and DESY in Hamburg.

D0 and SAM – Lee Lueking (FNAL)

The D0 experiment involves 500 physicists, from 76 institutes in 18 countries. They expect to collect 600TB of data over the coming 5 years. All Monte Carlo simulated data is produced offsite at 6 centres. SAM[7] is a data-handling system used in many clusters in D0 and they are now working with CDF to integrate it into the latter's environment.

SAM's design goals include optimisation of data storage and data delivery. In this respect it uses caches and meta-data to define the data. Access may be from individual consumers, projects or user groups; it makes use of fair shares and data and compute co-allocation. A SAM station is a place where data is stored and from which it is accessed. Data can be passed between SAM stations and to or from mass store. The control flow is based on CORBA. It has adaptors to various batch systems, mass storage systems and different transfer protocols. SAM is now considered to be stable, flexible, fault tolerant and has become ubiquitous to D0 users. It has interfaces to several batch systems, several mass storage systems and various file transfer methods.

SAM is used today in many scenarios: on a central large 176 node SGI SMP[8] system at FNAL; on a number of PC-based farms of various sizes up to 1-200 nodes, for example on the central compute farm at FNAL and on clusters at other D0 sites, including a shared 160 node cluster in Karlsruhe where the SAM station is installed on a single gateway node.

By September this year D0 were serving some 450,000 files per month on the main analysis farm. Statistics show that after regular use, some 80% of the served files were in the cache. Other examples where SAM is used were presented and are described in the overheads. At remote sites, D0 often shares worker nodes on a private network with local applications so the SAM station is installed on a gateway which has its own cache and copy of the D0 name database. Data access may be via staging (for example at Nijmegen) or via a local file server such as a RAID server (e.g. Karlsruhe).

SAM-GRID is an extension of SAM which should extend it to include job and information management (JIM) along with the basic SAM data management layer. As part of the PPDG collaboration, they are working with the Condor team and using Globus middleware. A demonstration has been put in place recently between 3 sites, FNAL, Imperial College in London and the University of Texas at Austin. As a consequence of this extension, they must understand how to make use of user certificates for authentication and authorisation and they now have to deal with other security issues.

Other work planned will investigate integrating dCache[9] and some more generalised HSM[10] features. SAM must be made less centralised in order to make full use of the Grid. They are looking at NAS[11] files to enable users to access only that part of a file which is of interest without copying the entire file.

Managing White Box Clusters – Tim Smith (CERN)

CERN buys standard boxed PCs in batches as needed and currently has just over 1000 dual processor systems installed. This farm processes 140K jobs per week and hosts 2400 interactive users daily. They can perform 50 parallel re-installs at a time if needed using parallel command engines. They estimate that they are about 7th in the Top 500 clusters in terms of number of boxes installed or 38th in terms of installed power.

The acquisition policy breeds a certain complexity – the farm is comprised of 12 separate acquisitions over the past 6 years, leading to 38 different configurations of CPU, disc and memory! Until recently, they were using in

[7] SAM initially stood for Sequential data Access via Metadata but is now known as Sequential Access Model

[8] SMP – Symmetric MultiProcessor

[9] dCache is referenced in the previous talk and described in brief detail in the CMS talk later in this conference.

[10] HSM – Hierarchical Storage Management scheme

[11] NAS – Network Attached Storage

parallel four different major Redhat releases from V4.x to V7.y. There were 37 identifiable clusters serving 30 experiments, each with its own requirements, a total of some 12,000 registered users.

They experience a certain number of issues related to the "dynamics" of the installation.

- They have noted a definite hardware "drift" due to hardware changes during lifetime of the PC's.
- With such a large user base, the password file needs to be regenerated and distributed every 2 hours.
- With over 1000 nodes operational, there is always hardware failure; a peak of 4% of the farm was registered recently as being instantaneously unavailable ("on holiday" as the speaker put it).
- Hardware replacement after some failures may lead to new hardware configurations.
- Failure management is manpower-intensive as one needs to perform some first line analysis before dispatch to the vendor.

Because of CERN's acquisition policy, some one-third of the installed base is out of warranty at any one moment and thus prime candidates for replacement, even apart from the fact that such nodes are much less powerful than more recent systems and more prone to aging problems.

To address the challenge, they think to

- replace older nodes in the interactive farm with a largely uniform configuration
- merge the multiple batch clusters into a large single shared resource whose redundancy offers more flexibility to handle the failures of individual nodes and permits to alter the resources allocated to any individual experiment at a given time
- create a separate farm dedicated to LHC experiment data challenges.

With such a large number of installed systems, all system administration tasks must be controlled via workflows in a problem tracking package; CERN uses the Remedy product.

On the software side, there is a legacy scheme based on 12 year's experience. The base operating system (Redhat Linux) is (usually) installed by Kickstart. A local CERN procedure (SUE[12]) then tailors this for the CERN environment, common applications come from CERN's public domain software repository (ASIS) and finally some scripts configure the target nodes for the particular task in hand. But there are risks that different tools in this set may conflict. The various objects for each of these may come from an ORACLE database, the AFS file base or from local files. The result is difficult to define and administer when neighbouring nodes ought to be as standard as possible in a cluster of several hundred nodes.

They have performed a redesign in a campaign to adopt tools coming from various current projects such as the European DataGrid (EDG) and the LHC Computing Grid (LCG). They started with a defined target configuration and linked this to an installation engine, a monitoring system and a fault monitoring scheme which between them are responsible for installing and maintaining a node in the target configuration. The fault management engine is responsible for comparing the actual node configuration to the target configuration at any moment and informing the installation engine what software changes may be needed to move to the target configuration.

They start from a clean initial state using the Linux Standard Base via RPMs and CVS is used for software versioning. To avoid software drift, no unregistered application provider can trigger software updates.

Their conclusion is that maturity and scaling have their own dangers and there is a need for ongoing strong management automation.

University Multidiscipline Scientific Computing – Alan Tackett (Vanderbilt University)

[12] SUE stands for Standard Unix Environment

This was a talk on VAMPIRE – VAnderbilt Multi-Processing Integrated Resource Engine. Conceived and built by a combination of biologists and physicists as a research tool for their needs, it currently consists of some 55 dual Pentium P3 PCs and a 700GB disc server. There are more than 70 users from various departments and it is also used to teach High Performance Computing to students by applying it to help solve problems in multiple disciplines. The philosophy is to try not to "reinvent the wheel" by using or adapting existing products and tools to their needs.

Among the user communities, HENP[13] users submit lots and lots of long serial jobs while other users need to run small or medium-sized parallel jobs requiring up to 20 CPUs, for example for problems in human genetics. The latter application requires a high performance network where the different threads or programs communicate between themselves. A third class of jobs commonly seen involves large ASCI[14]-class jobs requiring up to 512 CPUs; problems in condensed physics are typically of this class. These jobs often run offsite at ASCI sites but are debugged at VAMPIRE. Again a high performance network is needed. Here the use of Myrinet, although much more expensive that Fast Ethernet, shows the cost benefit of fast networking.

Because of the need to support a diverse user population, a range of products, including compilers, is required to be available. This often includes multiple versions of the same compiler. SystemImager is used to propagate system updates and using it, they can wipe and re-install the entire cluster in 30 minutes and this is thought to be scalable to thousands of nodes. System administration tools include

- Nagios (formerly Netsaint) and Ganglia are used for health monitoring and automatic service restart with system administrator paging as the last resort.
- Job execution is controlled by a combination of openPBS (for resource scheduling) and Maui (for batch scheduling).
- Parallel execution is controlled by a locally- written tool (pexec).
- GPFS is being evaluated for the parallel file system.

The features of Maui of interest here include fair shares, advanced reservation for both serial and parallel jobs, configurable job priorities and options for interactive debugging. Advanced reservation is important because of the job mix: if the cluster is nearly full and a job needs many CPUs, it needs to be scheduled only when enough CPUs are available.

A recent proposal is to extend VAMPIRE into a thousand node Scientific Computing Centre (SCC). The question of which CPU – Intel P4 or AMD Athlon (or Intel P3) – remains open at this time. The most recent (October 2002) Athlon chip appears to be 10-20% faster than the most recent P4 for some applications but the P4 is twice as fast for double precision BLAS[15]. The chip set choice is also important, as demonstrated in a table in the overheads. And the recent Intel compilers, although slow to compile, produce the fastest code. They believe that XFS gives the best performance for accessing large files from a shared file base.

VAMPIRE is currently being upgraded with another 200 compute nodes and 110 TB tape backup and there are plans for further significant upgrades in following years.

PASTA – Michael Ernst (FNAL)

PASTA is a series of technology reviews started in 1996 in relation to the LHC computing needs. The PASTA III review should revise the information and project it to 2005 and beyond. What are the business drivers, taking into account the state of the market and how the market has changed recently? Seven technology areas have been studied. The reports were now in the final draft stages but generally available online.

[13] HENP stands for High Energy and Nuclear Physics
[14] ASCI - Advanced Simulation and Computing Program
[15] BLAS - Basic Linear Algebra Subprograms

For semiconductors, the first PASTA started from the SIA[16] 1997 forecast whose predictions have been conservative to say the least. In 1999 PASTA II followed the still-pessimistic view of SIA. A more recent SIA forecast shows a major change from the point of view of the much denser chips now available. (60% more transistors on the chip) and PASTA III now needs to review the forecast but perhaps not as radically as SIA. Some negative aspects include the unexpectedly relatively poor performance jump in the Intel P4 chip (only a 20% performance improvement) and the slowness of improvement in compiler performance. Limits are also foreseeable in the CMOS technology. Another concern is the continuously-rising power needs. On the other hand, they note that chip size remains constant.

Nevertheless, the cost performance continues to improve. The comparison of the Intel P4 and the AMD Athlon chip is interesting where the latter is rather interesting for HEP but it currently runs rather hot.

In the interconnect area, PCI-X[17] is slow to arrive but the PCI-X2 standard looks interesting, as does PCI Express (also called 3GIO). Various other new interconnects are being studied and evaluated by different vendors.

The summary is that there are no major surprises at the component level but they need to watch how the market changes. There are approaching MOS physical limitations and the slight risk of needing more than air cooling for the densest chips. The question arises of the need to create our own HEP reference application in order to measure the performance they can expect for our applications. The existing CERN benchmarks are now considered old and possibly out of date.

Turning to disc technology, there are only a handful of vendors left from around 10-12 in 1999. And the existing ones (principally Seagate and Quantum) dominate different market segments (Quantum for desktop, Seagate for enterprise). Disc capacity still doubles every 18 months and 500GB drives seem feasible in the near future but rotational speeds and seek times are only improving enough to match rising disc capacities. SCSI is alive and well and still being developed (support for 320 MB/s is announced but not released). For commodity discs, IDE is adopting serial connections and moving to up to 600 MB/s. Fibre channel products are still expensive but the market for them is still active. DVD devices have not kept their promise, they are still too expensive and not shown to be in use in large libraries.

Tape technology will still be needed for LHC-era experiments and various products are or will be available. The LTO (Linear Tape Open) roadmap predicts up to 4 generations of LTO devices where each generation doubles both capacity and access speeds. This indicates yet again that they will need to deal with media changes in the future as much as in the past. Capacity-oriented devices must be traded against models intended to offer the best access speeds. One summary is that tape drives should no longer be treated as random access. A large persistent disc cache will be essential for LHC experiments. Media costs will dominate, not acquisition costs, so they need to continue moving to higher density devices. But there are no major challenges for LHC.

As mentioned earlier, networking is more and more important in the present and future generations of HEP experiments. The projected bandwidth to CERN by 2006 should be 20 Gb/sec. Experiments are already taking advantage of this unexpected bandwidth. The largest problem is that created by the comings and (mainly) goings in the market place. The choice of protocol plays more of a role as network bandwidth rises and processing is needed to optimise this. The DataTAG project is already investigating some networking aspects as are other research projects.

Storage offers probably the largest challenge for LHC. The future of storage architecture is hard to predict. The field is very vendor-oriented with not much inter-working between vendors. SAN versus NAS is still an open debate but object storage technologies are emerging where individual discs manage their own stored data. HPSS is still a good HSM[18] tool for large scale systems but there are few commercial alternatives for our market place

[16] SIA - Semiconductor Industry Association

[17] PCI stands For **P**eripheral **C**omponent **I**nterconnect.

[18] HSM stands for Hierarchical Storage Management

and so still a scope for in-house developments such as ENSTORE at FNAL, Castor at CERN, JASmine at Jefferson Lab. Cluster file systems are being developed, such as IBM's Storage Tank and the LUSTRE[19] scheme of which Sandia will install a large implementation shortly. Three possible interconnects are available – fibre channel which continues to grow, iSCSI or its equivalent over Gigabit Ethernet and the emerging Infiniband from HP/Compaq and others which uses IPv6 as the protocol. Similarly to faster and faster network bandwidth, work is needed to develop a light-weight protocol for file system access. Cost-wise, management costs still dominate.

Some Overall Conclusions:-

- Tape and Network trends match or exceed our initial needs
- CPU trends need to be carefully interpreted
- Disc trends continue to make a large (multi PB) disc cache technically feasible
- More architectural work is needed in the next 2 years for the processing and handling of LHC data.

Building a Computer Centre – Tony Cass (CERN)

CERN is currently refurbishing a new computer room and this talk will cover the needs for power, cooling, space, and fire security.

Power: discs, tape drives and robots need relatively low power; discs in the 10s of watts per device, tape drives and robots in the 100s. Be careful however not to start all discs at once because of the inrush current they need at spin-up. But CPU power consumption continues to rise, at a rating of approximately 1 watt per SpecInt95. Power supplies exist of varying quality levels and at corresponding varying price levels for a given power rating. Given the criticality of power supplies, the more expensive units are probably justified if amortised over the planned lifetime of the centre. Care also needs to be taken to properly plan the current expected to be drawn in order to optimise costs while reducing fire risks.

How to deal with primary power failures? Power for a long shutdown (measured in hours or more) is probably very expensive, so a more reasonable question is - what is needed to permit a controlled shutdown when primary power is interrupted, at least for critical systems (especially disc and database servers)? Battery UPSs are probably too expensive for the power needed; diesels would be better but are also expensive. You could consider a UPS to cover at least 5 minutes to smooth out power dips and microcuts. A middle way could be to install UPS or diesels for designated critical systems only.

You need to be careful of so-called differential circuit breakers on multi-socket strips. Experience leads us to believe that they are more likely to cause problems when such strips are used to power racks of PCs in HEP environments.

Cooling: which vendor will be the first to admit defeat and include water cooling in their latest offering? Our belief is that they are much more likely to stay with air cooling for some time yet. The various drawbacks of water cooling make it unlikely for the foreseeable future. Air cooling is still popular but it needs more recycling to be efficient. This implies frequent air changes and this has an effect on the working conditions if people need to work in proximity to the equipment for long periods. The use of fresh air, as opposed to closed circuits, usually helps, at least in winter, to keep costs down. The air feed can be from the top or the bottom and both have advantages; but top-feed requires a high ceiling room. The talk overheads contain some air flow numbers for the CERN installation.

Space: it turns out that at least in our environment, box size not an issue for the moment, clearance (accessibility) is the only real concern.

[19] Lustre is an association of Linux and Clusters. It is a novel storage and file system architecture and implementation suitable for very large clusters.

Fire precaution: they recommend a laser-based detection scheme although these are very sensitive and should not be the only trigger device for fire suppression systems. You need to localise the problem when it occurs if the air flow is very high. What to do if it happens? Fire suppression systems are popular if you want to keep running at all costs (banks, airlines) but these are expensive, bulky and need a hermetic computer room. Water-based systems are also possible. In our environment, fire prevention is a better option, with power being cut as soon as fire is detected.

CPU Technology Overview – John Gordon (RAL)

Not only are chip speeds important, cache sizes also affect performance, and how many caches, bus speeds, etc also play a part. Despite predictions, there is still no real movement yet to IA64 (Itanium), partly because of their current cost and the availability of smart compilers able to extract the best of the inherent power of the EPIC[20] architecture. The potential is there but it is still a guess when it will take off in the HEP market.

AMD chips are fast becoming popular in many large sites - good chip performance (especially with bigger caches) and faster internal buses. And AMD's 64 bit chip should also offer good 32 bit performance. A true comparison to Intel chips is non-trivial and depends on the application but in general you get more Gflops[21] per GHz and for a better price.

Looking at the roadmap of Intel chips, the Celeron chip in its P4 implementation seems to the base of Intel's single CPU systems. P3 systems have probably come to an end. For the high end and multi-CPU systems, Xeon seems to be the name of the game for now with Prescott (3 or 4 GHz and beyond) for the next generation. As noted above, IA64 adoption for HEP will need someone to write a really good compiler for Linux. The McKinley second generation IA64 chip is still too expensive, perhaps Madison will be more affordable.

And don't forget other chip vendors such as SUN (SPARC), IBM (PowerPC), SGI (MIPS), HP (PA and Alpha).

Dual-CPU motherboards are common in HEP but at increasing costs compared to those for single-CPU systems since the CPU cost is rising faster and there is a price for these boards. Blade systems[22] appear to offer higher density.

For parallel CPUs, IA64's shared bus to memory could be a bottleneck as opposed to the planned AMD design.

Summary:

- Don't look only at clock speed.
- Is it time to consider AMD?
- What is the optimal number of CPUs per box?
- Watch blades but not until they've solved the cooling problem.

Tier 1 Storage – John Gordon (RAL)

Tier 1[23] sites need to offer at least 40TB of disc capacity at reasonable prices on commodity servers. They need to restrict the disc per server ratio in order to offer good network bandwidth. RAL[24] were recently offered bids consisting of both IDE[25] and SCSI discs plus Fibre Channel and network-attached solutions. They finally decided on internal SCSI-based RAID controllers with IDE discs. In choosing the discs, no vendor offered evidence that SCSI discs were worth the premium over IDE. RAL found no objective review of individual vendors. Some

[20] EPIC - Explicitly Parallel Instruction Computing is the internal architecture of the Itanium® processor family
[21] Gflops – 1,000,000,000 floating point operations per second.
[22] See for example the description of the IBM blade offering.
[23] Tier 1 in LHC terms, Tier A in BaBar terms
[24] RAL - Rutherford Appleton Laboratory, UK
[25] See Pasta report above for links for these disc technologies.

vendors were unable or unwilling to provide good benchmarks despite a relatively large order in prospect so RAL performed benchmarks on a final list using the benchmark tools IOZONE with Bonnie as a sanity check. They saw 40 MB/s on NFS reads and 30 MB/s on writes; other results are given in the overheads.

After 6 months actual experience with the results of the tender, RAL discovered that 25% of discs came from a bad batch and the supplier is in the middle of replacing the full batch. More worrying is that the RAID controllers have a limited remap area and with a really bad batch of discs, the remap area overflows and the file system itself is lost! A firmware fix is expected from the supplier.

In looking at the next tender, RAL note that most IDE vendors only offer 1 year's warranty as opposed to 3 years last time. Also disc capacity is rising and at least one vendor offers "enterprise class" IDE discs, presumably at a premium.

Conclusion: SCSI/IDE has good characteristics but it is too early to see if this is the correct technology choice.

NSF TeraGrid – Remy Evard (ANL)

TeraGrid is funded by the NSF[26] to provide cyber-infrastructure facilities for many scientific disciplines. It could be considered a follow-on to NSF-funded supercomputer centres. Several different proposals have been funded or planned. The subject of this talk is a pre-production update on the Distributed Terascale Facility (DTF) which will consist of clusters at each of 4 sites (ANL, Caltech, SCSC and NCSA), each with a different target application. Most clusters will be a mix of Intel IA64 and Pentium P4 systems.

The objectives are to deploy a significant capacity enhancement balanced across multiple sites with an open and extensible infrastructure. They are trying to roll out a real production grid between 4 sites spread across the US and hope to start production by the middle of 2003.

The first challenge is to support high bandwidth between the sites, including possible extension beyond the initial 4 sites. They have set a target of 40Gbps. Partnering with Qwest and using advanced networks techniques. To maximise connectivity they will install two 40 Gbps connection points, one in Chicago and one on the west coast. Individual sites connect to these at 30 Gbps which is therefore the maximum speed between 2 end-points. The result, called I-Wire, will be a closed network for TeraGrid but how to define a TeraGrid site? Installation of the network is well advanced and on schedule.

However, they are still not 100% certain of the production hardware which will be installed, partly because of the "youth" of the IA64 architecture. They need to evaluate the target application load using a small Pentium P3-based testbed.

Although the TeraGrid is a multiple site system, it should have a coherent environment across the sites but this is very difficult to achieve. They have defined standard services and a minimum environment for run time and compile time coherence to which each site can add its own tailoring. This seems to be acceptable but the file system choices need to be verified.

They have met a number of social issues: 4 sites tend to have 4 directions. Opinions often differ and may be strongly held. Then there are historical rivalries to overcome. Participation must be equal and fair between sites. Decisions affect all sites. The organisation chart reflects this in its complexity. Lots of things get discussed and debated in working groups but these have mixed effectiveness. Experience of previous collaboration helps, for example networking people are used to working together, cluster managers less so. It is too early to know how these issues will develop.

[26] NSF - **National Science Foundation, a US Government funding agency**

Grid/Fabric Relationship - Bernd Panzer-Steindal (CERN)

This was an interactive discussion of some topics related to the implementation of Grids on existing fabrics. In principle, there exists a clear separation of Grid middleware from the fabric with defined interfaces between them but current implementations have not yet achieved this goal.

- How do grid choices interact with existing fabrics?
- Is this only a concern of current middleware status?
- What is an acceptable interface level?

Don Petravick FNAL) believes that Grid middleware should be limited to protocols and not implementations; FNAL does not like imposed implementations. John Gordon (RAL) warns of the risk of considering the reality (interim solutions) of today instead of the planned model of the future. There is a need to make remote sites feel they can trust "incoming" packages. Tony Cass (CERN) noted that nevertheless they have a problem now, how do they solve it? In order to deploy the LCG[27] soon, it should have minimal impact. Middleware writers should take this message on board.

The CMS experiment suggest constrain the problem in the short term and not mandate which version of the operating system, the compiler or other tools should be installed. According to Markus Schulz (CERN), EDG[28] software writers have not accepted this principle, for example they try to insist on certain daemons being present on target nodes. They should define good clear protocols first. Another complication is between small sites (one or only a few user groups to support) and large sites with many user groups with different needs.

Olof Barring (CERN) asks if you can create middleware to separate Grid from non-Grid applications. The OGSA[29] initiative even goes in a different direction, calling for more control of grid nodes.

Matthias Käsemann (FNAL) believes that if grids are to take over from webs, they must have very open protocols and get away from any constraints on local fabrics. They are clearly a very long way from this and it does not solve today's problem.

Should they distinguish individual clusters or deal with only a total grid? Surely some level of intrusiveness is required? Bernd Panzer-Steindal (CERN) does not consider this realistic today when they deal with virtual organisations but each site is independent. Tim Smith (CERN) believes that you should not intrude on another site, you should not know what is at the other site. It was suggested that some level of negotiation between sites is reasonable. The problem is that there are well-established clusters being used today whose managers are reluctant to re-negotiate working environments. John Gordon said there is a need to accept that some level of intrusiveness today; one can dedicate some part of a local fabric for grid work.

There is a conflict between all existing local and mature resources and the promise of the advantages of grids. Remy Evard (ANL) contended that clusters as a class are not mature either. And plugging in grids to clusters only complicates matters and problems must be expected. He said that such discussions are healthy, how can we feed them back into the grid developers?

Olof: Barring noted that HEP users face similar problems today in distributing their applications across huge collaborations, even without considering Grids.

[27] LCG – LHC Computing Grid

[28] EDG – European DataGrid

[29] OGSA - Open Grid Services Architecture

Workshop Proceedings Day 2

European DataGrid Fabric Management – Olof Barring (CERN)

The EDG is a 3 year project with 6 main and 15 assistant contractors. There are 150 FTEs split among 12 work packages (WP). WP4 is related to fabric management. It aims to deliver a computing fabric comprised of all the necessary tools to manage a computing centre providing grid services on clusters of up to thousands of nodes. Such a centre needs tools for managing user jobs and the fabric nodes on which they should run. WP4 has around 14 FTEs split over 30-40 people within 6 partners.

In May 2001 there was a report to the first LCCWS workshop which described using LCFG[30] for the initial installation tool; this was indeed used in the first prototype but will be replaced later this year. The monitoring plans at the previous meeting were based on a mix of PEM[31] and WP4's own ideas. This has been developed and recently deployed. The design of the overall architecture has been refined over the first year and now generally adopted.

The architecture model is based on keeping the configuration of the node (the desired target state) distinct from the monitoring (the actual state). Configurations should be hierarchical using templates. There should be node autonomy where possible, fixing problems locally if possible. Intrusive actions need to be scheduled, taking account of the job load which would be affected by such actions. They permit different security policies at the sites by plugging in authorisation as needed.

He showed a layered diagram of the various modules of the DataGrid and how jobs are scheduled (by WP1, the work scheduling package) and processed (by WP4). Then he showed how the automated management of the fabric nodes is performed by the various WP4 modules, in particular how node repair actions need to be scheduled. These are triggered by monitoring alarms and must take account of the load on the node and the level of intrusion of the repair, balanced against the severity of the alarm. Can the repair wait until the current jobs complete?

The Configuration Management task uses a high level definition language to define node configurations. These are based on templates, adding specific features for the individual nodes as needed. They are stored in a configuration database on the target node after compilation into an XML format. A new language was defined for this because no good match was found for the needs of this task.

Installation Management takes care of deploying actual node configuration, including the installation of the basic system and subsequent installation and updates of packages. The configuration is defined in terms of components and they configure themselves. The base installation should use native tools where possible, for example Kickstart for Redhat. Software Package Management (SPM) should be done also with pluggable packages. SPM is also a component like any other.

For both installation and configuration, LCFG is currently used and although much has been learned using it, some needed features are missing such as a lack of compile-time validation of the high level language definition of a node. So a new development has been started and should be deployed around April 2003.

Fabric monitoring is a framework for collecting monitoring information from sensors on the nodes, storing it locally and transporting it to a central repository. A first version is deployed but depends on ORACLE so public domain databases are being studied as alternatives. The fault tolerance module (another sub-task) uses this information to decide on repairs. If accessing the data on a local node it reads the monitoring cache directly but it

[30] LCFG is a system for automatically installing and managing the configuration of large numbers of Unix systems

[31] PEM (Performance and Exception Monitoring) was a CERN-written project some time ago, now replaced by a newer project using the PVSS product from ETM in Austria as its basis.

can run remotely and access its required data using the SOAP protocol. There are rules for correlating monitoring information. The fault tolerance module has suffered a serious delay but should be deployed soon.

The resource manager has also suffered delays due to problems with the Globus packages.

The Gridification sub-task handles user authentication, resource authorisation and credential mapping. This sub-task was deployed in an initial version in May this year but will be further refined.

In the past year, a lot has been learned, in particular by using LCFG to get started quickly. Not only did it help in getting feedback on what was going to be needed in the long-term, it taught middleware providers to follow certain rules for delivering software. Inter-work package coordination involves considerable overhead and social issues are a constant challenge.

MOSIX at CDF – Andreas Korn (MIT)

The MIT[32] group in the CDF experiment at Fermilab needed a group computing facility which would accommodate interactive users. They decided to adopt MOSIX from the Hebrew University in Jerusalem. MOSIX is a form of dynamic load balancing, effectively a kernel patch to Linux by which processes may transparently migrate across nodes, running jobs migrating to free nodes.

MOSIX includes Direct File System Access (DFSA): the discs may be mounted on any node but can be accessed from all nodes. Jobs may be migrated to the node on which their data resides if possible but if not, the inter-node data transfer speed is close to that of NFS. The measured overheads of DFSA are within tolerant levels.

MOSIX comes with its own monitoring and scheduling tools as well as tools to modify the behaviour of running jobs.

The configuration in this instance consists of 10 dual Pentium P3 PCs with 6 gateway nodes. The actual cluster nodes are on a private network – needed because there is no user authentication on individual MOSIX nodes in order to permit transparent job migration! The kernel is the standard Fermilab Redhat 6 version with the 2.2.19 kernel and MOSIX 0.98.0. Jobs are not allowed to migrate back to the gateways. The worker nodes have no user accounts, only root, and no special software.

The cluster has been used for some 18 months, initially for Monte Carlo data production but it is now also used for code development and some data analysis by some up to15 users.

The problems met include:

- the “home node” concept is an intrinsic problem where some system calls can only be issued from the home node, leading to unnecessary migrations.
- it is very dependent on reliable network connections since there is no check-pointing if a target node for migration, such as the home node, cannot be accessed.
- there are a few internal MOSIX bugs which are being worked on by the MOSIX developers.
- maintenance is problematic: the developers seem to be more interested in ongoing development and there is little or no support for older versions and many bugs can only be fixed by upgrading to the latest version, often in conflict with operating stable production clusters.

The group is currently looking at the latest release of MOSIX (1.8) on the latest Redhat version (7.3). Some bugs appear to be fixed but testing is in its early stages. The home node concept however has not changed and remains a limiting factor. They are also looking at a centralised network boot scheme since all the cluster nodes are identical. There are several new developments such as single system imaging and the openMOSIX split-off.

[32] Massachusetts Institute of Technology

Their conclusion is that MOSIX is ok for small clusters but probably not yet ready for large scale deployment. There are hints that Java-based applications don't migrate and some applications suffer from random migrations for no apparent reason and with no apparent benefit. Could Scyld be an alternative?

CMS Tier 1 Computing Centre at Fermilab – Hans Wenzel (FNAL)

Fermilab is the home of the CMS Tier 1 centre for the US, based on the MONARC model adopted for the LHC experiments. A number of US Tier 2 centres have already been identified. The definition of a Tier 1 centre is that it should produce its share of the CMS Monte Carlo data, hosts a major data store and offers code development facilities. It should also evaluate new H/W and S/W solutions. Being housed at Fermilab, the CMS Tier 1 centre benefits from local experience both in the main computing centre and running experiments.

For monitoring they use Ganglia for low-level monitoring, FBSNG from Fermilab for batch scheduling (see next talk) and dCache from DESY for data caching. They use the standard Fermilab process for selecting new hardware with formal tests with the target applications. The latest selection was for 65 PCs based on AMD dual Athlon CPUs. Nodes must then pass a 4 week burn-in. Currently there are only 3 PC generations on the floor. They are also upgrading the 7 dCache nodes and adding 3TB of disc storage. They have tried MOSIX (see previous talk) but rejected it because the CMS applications with their database dependencies do not take advantage of the migration features.

dCache was developed at DESY and it is used as the front end to the ENSTORE mass storage (effectively making it look like network-attached tape). dCache interfaces to this for end-users and should make the multi-TB tape system look like a homogeneous storage system. It should smooth out tape and disc access rates, it is fault-tolerant and it tries to make optimal use of the tape robot. User should not have to care where the data actually sits at any one moment – no explicit staging necessary. The nodes use IDE discs but need a 2.4.18 Redhat kernel to avoid memory management problems. They found that PCs used as servers benefit from using a server-specific kernel. The overheads show some results of measuring dCache performance and demonstrate that individual processes can achieve good access speeds even with many concurrent processes performing I/O.

They are looking into how best to configure a farm for interactive use, possibly using CERN's interactive farm (LXPLUS) as a model. In parallel, they would like to make batch more user-friendly and they are also evaluating NPACI ROCKS for farm configuration and various disc management systems such as zambeel.

They feel it is essential that service providers, users and middleware providers must talk together in order to design and implement something which allows the end-users to run their applications in the most efficient manner within local constraints but benefiting from the projected Grid environment.

FBSNG and Disc Farm – Igor Mandrichenko (FNAL)

FBSNG is the second generation of a farm batch processing scheme developed at Fermilab. It adds support for file parallelism. FBS V1 used LFS to actually schedule the jobs but this was removed in V2. Rather than being based on load measurement and load sharing, FBSNG is based on resource counting. It should recognise available farm resources and know what resources are required by running and queued jobs and hence where to schedule new jobs.

The unit of operation is known as a job section; it could be a single job or an array of job processes. Resources are abstracted and counted as semaphores. Local resources may be CPU power, disc space and local tape drives. Node resources include operating system flavour, installed software and so-called logical attributes (for partitioning a farm). There are also global resources such as network throughput between nodes, NFS-exported discs and some global semaphores for job coordination.

This scheme permits great flexibility, offers options for prioritising jobs, fair-share scheduling, guaranteed scheduling and resource utilisation quotas. Since failures happen, robustness has been planned for. The farm can be reconfigured dynamically and the overall service is stable with respect to failures of individual nodes. As elsewhere in Fermilab, the farm is fully Kerberised. The software, largely written in Python, is portable and runs on Linux, SUN/Solaris, SGI/Irix and Compaq/Tru64.

There are both command line and graphical user interfaces (GUIs) as well as a Python API[33]. There is also a web interface. There is a light-weight client which can be installed remotely and can submit jobs over the network.

It currently runs on several large managed farms at Fermilab, including those for the D0, CDF and fixed-target experiments and CMS plan to adopt it for their local farms also. It is also used at remote sites (e.g. NIKHEF) and at least one known corporate site.

A typical computing farm will have 5-10 TB of disc space with aggregated throughputs of the order of magnitude of 1 GB/s. However, because of the distributed nature of the storage, sometimes unreliable components and the need to coordinate access and allocation, the disc space on such farms is difficult to use and the Disc Farm tool has been developed to help use this unused disc space. Disc Farm is effectively distributed data storage: you view a farm as an array of discs with attached computers. It organises the unused space into a global virtual name space. A given file has both a physical and a virtual path name. User interfaces have been created to access files in this space. Data replication is supported for redundancy. There is load management and load balancing. It scales with farm size. WAN[34] access is available via a Kerberised ftp server and a grid ftp interface is in development.

Data Challenges and Fabric Architecture – Bernd Panzer-Steindal (CERN)

Within the general CERN fabric nodes can be moved between different usages as load on different "fabrics" changes. The current architecture is based on commodity components, PCs with dual CPU, NAS[35] disc servers with EIDE disc arrays, Redhat Linux and open software packages.

Any architecture must be validated against reliability, performance, and functionality and it must match the computing model of the experiments. The batch farm runs on some 700 nodes with 65% CPU utilisation currently. Across the farm there are some 7 reboots per day and 0.7 vendor interruptions per day (mostly for disc problems). This is considered to be stable production. The comment was made that, while the exact number may be debated, one node with problems is more resource-expensive than 200 without, because hardware problems can be hard to diagnose and fix and software failures can absorb vast amounts of time.

Network performance is also very stable given the number of devices installed (29 interventions in 6 months). CERN has started looking at 10 GB routers and switches but problems were quickly seen and vendor support must fix these before CERN will continue with the evaluation. Network bottlenecks seen today are acceptable for production and for the Grid testbeds. This is not true for data challenges which should be bleeding edge performance. They see average network performance running at 400MB/s with peaks of 600. With special tuning for data challenges they have achieved 1.4 to 1.6 GB/s.

Disc performance and stress tests have been performed on Pentium III systems with 500GB of mirrored discs and 45MBps aggregate throughput. Total rates up to 500 MB/s write + 500 MB/s read have been achieved, limited by network setup and load balancing. The stability was about 1 reboot per week (out of ~200 disc servers in production) and ~one disk error per week (out of ~3000 discs in production). They measured translate to an MTBF[36] of 160,000 hours. Scaling tests were performed and they saw linear performance with only network

[33] API – Application Programming Interface
[34] WAN – Wide Area Network
[35] NAS - Network Attached Storage
[36] MTBF – Mean Time Between Failures

limitations. File system multi-stream I/O disc to network transfers (or vice versa) runs at 45 MB/s aggregate. With new P4 Xeon nodes they achieve 80 MB/s.

Tapes run at 12 MB/s uncompressed but only 8 MB/s including overhead as measured with real user programs. They mount some 45,000 tapes per week. The error and intervention rate has improved from last year. New drives are being installed very soon for the LCG[37] prototype tests. It has not been easy to match the different tape and disc server speeds.

CASTOR couples a lot of these components together, from online data to data storage. CASTOR HSM[38] currently stores some 1.3 PB of data in 7 million files. In the recent ALICE data challenge they achieved 85 MB/s on to tape. The target for November 2002 is 200MB with new tape drives.

They believe that they have a viable model for data storage and access. It is stable, it seems to scale and it performs well. It is dependent on the market (risk of paradigm changes) and the eventual analysis model of the LHC experiments is crucial.

Data Centre Experience – Don Petravick (FNAL)

Experimenters and computing professionals work together and bring their respective insights into play when solving any given problem. The key philosophy is intellectual collaboration between both parties. We as computer professionals must avoid thin interfaces to the experimenters. We must do the mundane things well and earn credibility. Modern and, especially, next-generation experiments loom large because of their size and steady long-term ramp up. But we must not forget the smaller ones, many of whom may be good customers for grid and cluster technology.

Important technical entities include:

- Connectivity: today exclusively standard Ethernet within the centre; fibre channel did not live up to some people's expectations, nor did large MTU[39] Ethernet.
- Computational nodes: at Fermilab there is still a significant SGI component but the main focus, as it has been now for some time, is on PC Linux white boxes. Existing personnel and their skill base are important in making good use of these masses of systems.
- Storage: data architecture is vital in matching expectations of modern experiments. FNAL uses a combination of locally-written ENSTORE for handling mass storage and dCache from DESY for buffering and caching the data. Access methods now include grid interfaces.

The main direction for disc storage is to make use of installed space on farm nodes (see previous talk on FBSNG and the Disk Farm) and to build a Linux-driven TeraByte store. The transport will be via Ethernet. For tape storage, STK 9940 and IBM LTO are the production devices; they are moving to STK T9940B which are faster (30 MB/s) and higher capacity (200 GB).

ENSTORE handles tape staging; it is very scalable and used both in the computer centre and in the data acquisition chain. It currently operates at about 10 TB per day, over 100 MB/s in sustained mode, spread over 10 tape drives.

dCache performs rate adaptation and takes care of disc failures. The interface is more abstract than ENSTORE and so it is more flexible. Protocols are Grid/WAN and LAN; access modes are staging, POSIX-like direct file access and management access. Congestion control is essential where each dCache pool only serves a finite number of users and can only have a finite number of streams to ENSTORE.

[37] LCG – LHC Computing Grid

[38] HSM – Hierarchical Storage Management

[39] MTU – Maximum Transmission Unit

File transfer (ftp) is available in several modes: with weak authentication, protected by Kerberised authentication and via GridFTP.

Storage Resource Management (SRM) provides management functions such as pinning of a file or pre-staging or space reservation. They are working on inter-operation with other labs such as Jefferson Lab and LBL. SRM is integrated into CDF's framework and the local Grid environment.

dcap is the LAN protocol for dCache, used for direct file access as well as staging. It has been integrated with both ROOT and Objectivity and DESY are currently adding Kerberos support.

There are masses of farm middleware, some of which have been described in earlier talks. Many of these were designed and implemented in collaboration with the users. Administration and operations support are very important and a lot of energy has gone into developing the necessary tools.

Summary: the FNAL facility provides a comprehensive suite of tools and services to the large and varied FNAL scientific community. There has required a substantial intellectual investment.

EDG[40] Testbed Experience – Markus Schulz (CERN)

This talk will concentrate on problems and issues raised but the speaker emphasised that experience so far has been generally good and has already been covered in earlier talks. EDG itself was summarised earlier. Although a development project, it must demonstrate production quality software. Some important features include:

- like many Grid projects it is based on Globus
- the current testbed comprises 21 sites, some 400 nodes, some shared with local batch farms. Services include site-independent authentication using GSI (Grid Security Infrastructure, based on PKI[41])
- Globus Information Services (GIS) based on MDS[42]
- storage management including file replication and GSI-enabled FTP. Catalogue replication is based on LDAP
- resource management – covering resource brokering, job manager, job submission, accounting, monitoring.

Services are inter-dependent and failure in one may affect several others. Services may be composite and may need to be interfaced (e.g. Condor, MySQL) among themselves.

Services are mapped to logical machines and imply constraints on local node setup as the DataGrid is currently structured. The speaker showed a map of which services must run on which nodes as well as a minimal configuration of a testbed.

The current CERN EDG testbed consists of 90 nodes split into 3 distinct functions. The Production Testbed (now called the Application Testbed) runs the stale release and is updated only rarely. It is used for test production codes and demonstrations but it still requires frequent node restarts.. The Development Testbed has constantly-changing releases and hence is quite unstable. The third testbed is used to integrate new major releases (including changes to Redhat or Globus software as well as the middleware itself). There are also a small number of systems which may be individually allocated to developers for testing.

The CERN testbed is based on CERN-standard Redhat Linux. There are 2 NFS file servers with 1 TB of disc. NIS account servers to manage user accounts (since many Grid users do not have CERN accounts). LCFG is used

[40] EDG – European DataGrid
[41] PKI – Public Key Infrastructure
[42] MDS – Meta-computing Directory Services

for installation and configuration. They run a Certificate Authority (CA) to provide CERN users with X509 user certification.

To LCFG has been added support for PXE boot sequences and DHCP to allow completely network-based installations. This works well for well-tested configurations and for identical nodes. But the number of machine types must not be too large and it has certain limitations, such as a maximum of 4 disc partitions. However LCFG has certain drawbacks in a rapidly-changing environment:

- configurations are always changing and re-installing a node from scratch which has been first installed and later updated to a new state does not always work correctly
- there is limited feedback on the success or failure of an LCFG update.
- some LCFG objects never existed in a working state or were not invoked correctly
- the chosen philosophy of replacing installed systems by RPMs does not match the environment needed by a developer
- impractical user management functions.

They decided to separate releases of the tools and the middleware. LCFG is turned off for developers after installation. They wrote tools to check on success or failure. But they are still missing some LCFG objects. And they have gone away from using LCFG for managing all accounts and are now use NIS for everything except the root and few other service accounts. Despite these issues, they have realised that using a tool such as LCFG is mandatory for such a large number of nodes and that code developers must get used to delivering appropriate installation and/or configuring objects with their code where appropriate implies a match to the chosen project tool, LCFG or other.

In constructing EDG Middleware they have noted that many services are fragile with very complex fault patterns and they still need to decide the correct way to do some things. Better error codes would be appreciated. Scalability is not so good in some places. Some software modules are in conflict, requiring different operating system or library releases. All these points have led to creating ad-hoc tools for monitoring and repair and sometimes there needs to be multiple instances of services. Some grid specific problems include an unfamiliar model for user authentication and authorisation; there is no central grid administration and propagation of required changes is slow.

Different testbeds have different usage patterns and this can cause resource problems. There are some high visibility demonstrations and tutorials which must have high priority support. Some users have effectively performed production work and need production-level support. All users expect fast response and even casual users create overhead because they assume the testbeds should offer production level service.

With the coming introduction of Redhat Linux 7.3 and EDG software release 2, will they need to run all combinations, where is the hardware? Where are the people to manage and support all of this? Running large testbeds remains a complex task requiring tools, administrators who understand the needs and are able to react quickly to changing environments, and dedicated user support.

RHIC Facility Evolution – Shigeki Misawa (BNL)

There are many factors driving the evolution of a computer centre, including stability, security policy, increasing user needs, technology changes, budgets, etc. These are often in conflict and a choice has to be made. At the BNL lab, RHIC has sacrificed reliability and power. They have been forced to add security gateways for user login (ssh) and file transfer (bbftp) and a facility firewall as well as a site firewall which was installed subsequently. They use HPSS for managing their tape archive, 19 SUN NFS servers, a Gigabit Ethernet backbone, 2 AFS cells (the IBM flavour) and 1000 dual CPU Linux worker nodes with local scratch space, a lot of it in many cases.

The NFS and local scratch space is managed with low overhead, mostly delegated to users themselves. There are a variety of data processing models, at least for analysis if not for data reconstruction which is more similar between the experiments. This diversity makes it difficult to configure or tune the systems for optimal use.

Security-wise, exposure is thought to be improving and management and monitoring is getting better. They had to implement rigorous US Department of Energy-mandated user life cycle management. They would like to move to Kerberos version 5 and LDAP for user authentication and authorisation but being tied-in to AFS complicates this (as AFS uses Kerberos version 4); tests have been successful but they are wary about introducing it into production.

HPSS is fairly stable now and is expected to remain so. Workarounds are in place where limitations or failure modes have been uncovered and understood. All this results in a heavy investment of resources which makes them reluctant to consider switching to another scheme.

Gigabit Ethernet not an issue, there is plenty of performance and plenty of headroom for now. The only issue is vendor shake-out, disappearing suppliers.

The farms have adequate processing power for now and the nodes have more disc space than initially planned, benefiting from changes in the price/performance of disc capacity. This fits the processing model adopted by some experiments. But now individual nodes may become mission-critical and maintenance is now an issue since the original VA Linux[43] boxes were not easy to fix or swap in place. More recent acquisitions cause fewer problems and are easier to fix. And different user groups have different disc configurations which restricts node re-allocation or swapping.

NFS servers provide some 100 TB of storage; availability and reliability are getting better. Many problems have been seen in almost every component of the disc configuration. It has become more stable some since months although the servers are currently overloaded. NFS logging was switched on (on Solaris) to see where the problems may be but this was inconclusive (and creates huge logs which are virtually impossible to analyse) and has now been disabled. Not enough meaningful statistics were gathered to determine the cause. The result is that both the servers and the NFS space are poorly utilised. The answer for now is to roll out more of the same and the various alternatives are being looked at.

In the medium to long-term they would like to go to OpenAFS and Kerberos version 5 based on Linux servers. Who has done that? What about Grid integration implications?

They feel that they have pushed their installed systems to their limit and it may be time to force more discipline on their user base.

GridKa – Holger Marten (Karlsruhe)

The Karlsruhe Forschungszentrum is one of 15 research centres in Germany, it has 40 institutes and divisions covering many disciplines. The central computing and communications department provides services to the campus and has recently added some research activities, one of which is working on GridKa. The mission of GridKa is to provide Tier 1 facilities in Germany for the 4 LHC experiments as well as Tier A centres for today's generation of experiments. This makes a total of some 41 user groups from 19 German institutes.

In order to support multiple experiments, which versions of which Linux (Redhat or SuSE) should be installed? Should they split the cluster into different parts? Should they reconfigure it each time for a new experiment? In the end, they decided to create a shared resource running Redhat 7.2 plus each experiment gets one dedicated

[43] VA Linux was the supplier of the hardware; they have since transformed into a software-only supplier

software server which can be used as a gateway. Named user group representatives may request permission to install any software on his/her node, even a different flavour of Linux.

There are 124 dual Pentium P3 compute nodes with a total of 5 TB of disc space. They use OpenPBS for batch scheduling and NPACI ROCKS for assisting with system administration tasks.

They require a plan for scalability, heterogeneity, consistency and all this with limited manpower. For scalability, they build on a hierarchical management architecture. They believe that installation using RPMs instead of disc cloning should help cope with heterogeneity. They have chosen a very prosaic naming scheme to designate allocated task of a node, namely its cabinet address and a sequence number. This is intended to help identify individual nodes when required.

NPACI ROCKS is used for node installations and upgrades. System monitoring is done by Ganglia with a web interface which can be accessed (in read-only mode) by users but there are doubts about how well it scales. Cluster management is performed with Nagios which is better able to handle "events"; it also has a web interface. They are combining Nagios with their Tivoli Enterprise Console from IBM.

They have developed with a local firm a closed rack-based cooling system for their cluster. It has some 10kW of cooling power and they estimate to make a 70% saving on air conditioning charges.

Currently they have 45 TB net capacity on a mixture of disc architectures and various file architectures. This includes a successful fibre-channel SAN[44] with Linux servers and IDE discs. But there are concerns about the scaling of disc storage and disc management and for future file architectures, they intend to evaluate GFS and GPFS for Linux.

Various tools exist for tape storage management but most seem to be specific to vendors so they are looking at linking the IBM Tivoli Storage Manager to both CASTOR as used in the EDG [45]and dCache/SAM.

Modular Screen Saver – Frederic Hemmer (CERN)

The normal PC Windows desktop lives for typically 3 years but is only used for less than 25% of this. CERN has a very CPU intensive application for LHC particle tracking called SIXTRACK which is "embarrassingly parallel". It is currently running on a dedicated Linux cluster which must be expanded to test more particle tracking models as real data about the magnets is obtained. A typical run takes 2 CPU hours on a Pentium P3 with low I/O, or at least low read I/O.

CERN has over 5000 installed desktops, most of which of course are unused overnight and at weekends. Seti@home is a good model for using this unused power and UD[46] is another example, used for example for medical research.

CERN has designed a client screen saver to control the execution of an application and pass back the results to a central server using HTTP and SOAP as the transfer protocols. This makes the client platform independent of the server platform and removes any intranet/internet. It also promises scalability.

The prototype consists of a VB script to register that the node is available, to cycle the job and to return results. The server side consists of standard web-like actions and includes limited management features such as updating the client.

[44] SAN – Storage Area Network
[45] EDG – European DataGrid
[46] UD is United Devices Inc

A first prototype runs on a few PCs and there are plans for a wider deployment to a limited number of nodes in November (2002) if all goes well.

Clearly this is not a new idea but this may be time to apply it to HEP applications. The scheme could be extended beyond desktop screen savers. It could be applied as a background task for servers and even beyond CERN to home PCs. And why stop at only Windows desktops?

Summaries

Technology Stream Summary – John Gordon (RAL)

VAMPIRE – there's a world beyond HEP and they have different needs and different characteristics; maybe we also should look more at Maui.

TeraGrid – clusters are not only Linux.

CERN's PASTA review shows that tape and network trends match our initial needs but CPU trends need to be interpreted. In the disc area, should we aim for NAS or SAN or what? Disc size may not be an issue, although we will need a lot of spindles, but they are not getting much faster. And there are still many file systems in use. Tape is still cheaper than disc and speed and capacity are still increasing increasing; STK still dominates our market segment but LTO is coming in.

For CPU, Intel is still widely used but AMD is becoming more common for floating point applications. White boxes are still the cheapest capital cost for PC power but racking may change the equation for some cases.

Network: techniques exist for very high throughput and capacity is matching rising demands although not for general purposes. Given that available and planned capacity, should we re-look at the MONARC model again?

O/S: Linux is a given but agreeing the particular release to use is not trivial. MOSIX is interesting but has been thus for some time without getting any closer. [The current version is 0.98.0 – not yet ready for version 1!]

Infrastructure: many large clusters exist, some by natural evolution, others by major acquisitions. But the larger you get, the more professional you must become at all levels from the ground up (literally).[47] We need to be more professional when building a really large Computer Centre. There is a rich selection of management tools; not a problem in itself, but which to choose? An unsolved problem is that of application installation and that of configuration management is not fully understood.

Items not covered include managing large user communities, software certification (control over incoming software), security (discussed in many other places).

Do we always agree on the taxonomy? For example, do we all understand the same thing between authentication and authorisation?

What is the overall cost of building and running a white box farm? For many users, only up-front costs matter, at least matter most; lifecycle costs matter less. Should we buy in solutions? Are they available at affordable prices?

Experience Stream Summary – Ruth Pordes (FNAL)

In the large majority, today's experiments are using commodity technology and shared usage seems to be increasingly supporting multiple user communities. This translates to more emphasis on dynamic configuration management.

The basic model and use of a cluster seems not to have changed much since last workshop but there has been a steady scaling and increasing of cluster size although not by orders of magnitude. As a consequence there needs to be lots of management, lots of effort in finding ways to automate this.

[47] Comment by Chuck Boeheim of SLAC: "10 nodes is Arts and Crafts, 100 nodes is Carpentry, 1000 nodes is Engineering and the design challenges take different skills at each level."

Users are starting to expect production level grids and not research projects and users expect the same environment across sites.
.

It may be time to include desktops in the “grid” but there is no real experience yet, maybe last talk of the conference (on CERN’s Modular Screen Saver) will provoke more of this.

There was very little said about commercial software, but it is clear that there is no silver bullet.

NFS serving does not scale but maybe XFS does.