HEPiX/HEPNT, Fall 2001, NERSC, LBNL, Berkeley

Alan Silverman

Introduction

The meeting was held at the Oakland Facility of NERSC/LBNL where their main computer systems are located (described below). Despite the various problems associated with travel these days, some 40 people attended and all major sites were represented except BNL. Wolfgang Friebel and myself represented CERN.

Highlights

- Several small sites reported on the difficulties associated with having to serve multiple experiments.
- Several sites reported on problems, often severe problems, in using NFS for access to data. It seems in particular to be a problem with Linux unless care is taken in the set-up.
- SLAC has an interesting scheme to install large numbers of Linux nodes in a very short time.
- A scary talk from Bob Cowles on his attendance at the recent Defcon 9 hackers conference and the openness with which hackers share the tools of their trade.
- An interesting talk from Steve Lau on security measures at NERSC, which has installed the largest number of supercomputers that I have ever come across. He described how they handle the various risks and problems.
- Progress reports from FNAL on their Strong Authentication Project using Kerberos (limited to reducing the use of clear text passwords on login for the moment; no extension yet to mail or web authentication) and their NGOP monitoring project (where the first production release has recently taken place).
- Problems reported by many sites in using the Serverworks LE chipset on PCs
- Lack of progress of most sites in moving towards W2000 in any serious manner. FNAL is almost the only exception.
- Most of the overheads are available via the web site at http://pdsf.nersc.gov/hepix/

Future Directions in Scientific SuperComputing, Horst Simon (Director)

There is close co-operation between the University of California, Berkeley and NERSC in computer science and computational science. NERSC is an unclassified facility with more than 2000 users in DoE mission-relevant research. Examples include the BOOMERANG experiment to analyse cosmic microwave background data and many HEP experiments.

Simon identified several strategic components for te future – high throughput computing, supercomputing, Grid technology, international collaborations. NERSC has one of the very top systems in the Top500 Supercomputing, certainly the highest non-classified one. HPSS is another significant component. He noted that TFLOPS by themselves are not enough – you need lots of memory. NERSC is working increasingly in a Unified Science Environment where different sciences work in tandem, sharing Grid technology, compute power, storage, etc.

He cited the following trends – continued growth of computational and storage resources; increasing formation of large-scale multi-national collaborations; increasing use of Grid technology.

In more detail what matters is –

1. Continued rapid processor performance is still following Moore's law – an analysis of the Top500 systems shows growth of 1.82 per year recently. Will this continue? Known plans for coming systems until 2004/5 seem to indicate yes.
2. Open software model (Linux), notably the important contribution of Beowulf clusters which demonstrated the effectiveness of PC clusters to many classes of computation.
3. Network bandwidth will grow even faster than Moore's law
4. Aggregation, centralisation, collaboration
5. Use of commodity components everywhere.

NERSC has had a 3 TFLOPS system since December 2000. LANL will install a 30 TFLOPS system from Compaq soon.

IBM Blue Gene project is a 64000 32 GFLOPS system using PIM chips. It is a special purpose architecture for performing one task only, which, despite its name, has nothing to do with gene processing but this ensured good publicity! It will produce CMOS-based Petaflops.

The NERSC Oakland site is built for 2010-size clusters with 20,000 square feet of floor space with good power and infrastructure and room for expansion.

US DataPort won a contract to build a $1.2B centre in San Jose to run web servers for commercial customers (of the scale of Yahoo, E-bay, etc).

PCs are now saturating; non-PC devices and Internet devices are about to take over the growth trends. Low power, low cost, low consumption devices are coming.

He predicts by 2010 – Petaflops peak supercomputing performance using MPI and commodity parts although he refuses to make a prediction on which technology will be used for the chips; Grid will have happened, waiting for the killer application. Microsoft will have been split into 3. All this is liable to disruption due to some unrelated outside event, eg. Robotics.

Benchmarking CDF Code, Rochelle Lauer, Yale Physics Dept.

They are using Compaq TruCluster today with NFS file sharing but the CDF group require Linux. Another Yale group are members of BaBar. So she bought some Intel systems to experiment with Linux. She uses various FNAL products and LSF. She found really unacceptable NFS performance. Even applying known patches did not give Tru64 performance levels. Thus their farm model would not work for Linux – having no local disc does not work in this environment. But she was reluctant to build many independent systems, even clustered, with local discs that would all have to be kept current with the latest developed CDF codes. The chosen solution was a multiprocessor model (preferably a quad-CPU system) so she tried to create a benchmark to prove this. It gave favourable results with minimal overheads. More L2 cache helps a little. The downside is that a multi-processor system is less flexible than a farm – how to hot-swap a sick CPU? How to add one more CPU when it is needed?

Linux File Systems, Steve Chan, NERSC

Evaluated NFS, GFS (Uni Minnesota), AFS, GPFS (IBM for their SP systems), PVFS.

- GPFS got cut quickly - could not get it to run on their (IBM-built and managed) Alvarez cluster, even with IBM help!
- PVFS got cut – it was fast but not robust (single point of failure) and application codes need at least to be relinked if not recompiled to run properly (connected to use of libc library)
- AFS got cut next - heavy to deploy (needs a specialist administrator and a Kerberos installation)
- GFS got cut last – not fast, stability under doubt, really needs Fibrechannel; depends on a SCSI-level protocol supported by only a few vendors today
- This left NFS, the only realistic choice in his opinion.

How to configure NFS for maximum performance? Install the most recent Linux kernel for a start. Keep to a maximum of about 20-30 connections. NFSV3 should be chosen, better caching but still poor performance on writes. Details on his overheads.

A General Purpose High Performance Linux Installation Infrastructure (Alf Wachsmann - SLAC)

An installation scheme was needed for farm nodes and desktops. It should be fully unattended and able to install hundreds of nodes in parallel and in finite time.

The chosen solution uses a modified BIOS with the Pre-boot eXecution Environment (PXE) installed to offer network boot. This issues a bootp request to a DHCP server to get the IP address; then it uses tftp to load the bootloader programme which loads the kernel and its designated configuration files via NFS. Then Kickstart runs and completes the installation. All this is controlled by self-written scripts, including some post-installation tasks.

Since SLAC uses static IP (a local choice), how to load new IP information when new systems arrive? They use SNMP to read MAC addresses and by careful wiring (node 1 to cable 1, etc) they can match these to IP addresses using a simple mapping algorithm.

Benchmarks: they installed 256 nodes in a batch connected to a Gbit NFS server: 50 had a problem but the other 200 were installed in 30 minutes. Another procedure, re-installing 100 nodes connected to a 100Mb/s NFS server, took 45 minutes. The bottleneck is clearly the NFS server.

See his URL for details - http://www.slac.stanford.edu/~alfw/PXE-Kickstart/

Large Scale Distributed Computing (Nick Cardo - NERSC)

NERSC 3 is the second most powerful cluster in the world. It is basically a large (huge) IBM SP cluster running AIX 4.2. It is intended for parallel applications and the dominant programming model is MPI. It has 184 compute nodes, 3 interactive nodes from which users submit jobs and 16 server nodes. It uses the GPFS file system from IBM. Each “node” has 16 processors for a total of 2528 processors, 3.8 TFLOPS. A node has mirrored drives and memory ranging from 16GB to 64GB with a total of 4544GB of memory. They use IBM SSA drives with about 20TB formatted RAID 5 capacity in total. Each node has a Gbit

Ethernet connection to the “outside world” plus hidden network channels for storage and control access.

The use Loadleveler for batch. Different job classes may use different numbers of nodes from 8 up to 128 with different charging factors (using “monopoly money”) depending on class, wall time and number of nodes allocated. Jobs can be “premium” and have guaranteed service by up to 128 nodes (times 16 processors). Or the job can be “low” which can also have up to 128 nodes but has no guaranteed service, and costs less.

Nodes are not shared so there is incentive for programmers to use as many processors on the allocated nodes as possible and hence as few nodes as possible. Most programs are in Fortran with some in C. Totalview is used for debugging.

Installing and running Sun Grid Engine (Wolfgang Friebel, DESY Zeuthen)

DESY Zeuthen first installed Codine, as it was called when they chose it, in 1993/4. As SGE it is now available under open source from SUN. It runs on the DESY Zeuthen cluster, which contains some 120 systems. The original choice was influenced by the fact that it was a local product on which they felt they would have more influence than for example LSF, to which it is broadly similar, and it matched DESY’s preferred batch model.

He described the components of SGE including how jobs get scheduled and how administrators can affect this as the user community changes. SUN have developed the base product into a so-called Enterprise Edition (SGEEE) by adding more tools for administrators. In Zeuthen SGEEE is enhanced with AFS support, added for the HERA-B experiment in a previous version of Codine/SGE by an outside company. Zeuthen currently run the unlicensed open source version.

The system is highly configurable (for example it can have very complicated scheduling algorithms) and yet very stable. It is liked by the users who find it easy to use. It adapts dynamically to changing process priorities. Seems to be under active development.

Although there is a SUN licensed version and an open source version, new developments are fed back into the SUN official version, at least so far.

The Grid is Coming; Is Your Infrastructure Ready? (Bob Cowles - SLAC)

The Grid vision is that PKI-based authentication will allow pseudo-random jobs to be run at random sites. But different sites have vastly different authentication policies. There will be a need to provide trusted authentication and authorisation across security and trust domains and a need to determine the risk model.

One technique is grouping – group accounts, group resources, many users mapped to a single id – hence the proposed Community Authorisation Service – CAS. But the proponents of CAS appear to be not very far advanced on its full implications – te risks of many users accessing files via a single user id for example.

Are certificates issued by site, by country, by experiment (virtual organisation) or by Grid?

FBS (Farm Batch System) status report (Tanya Levshina - Fermilab)

FBSng (new generation) has all the usual batch scheduler features, including a fair-share scheme. It is used by FNAL experiments of course but also by D0 users at NIKHEF and D0 and CDF users at Northwestern University. The latest version (1.3), written largely in Python with a few C/C++ modules, was released in June and is available via Fermitools.

New features include optional Kerberos authentication support; it permits dynamic reconfiguration of a farm; job status change notification; new features in the GUI; and many others. There is a web interface built on the Python API.

The new version appears to be stable and of production quality. It is portable and easily configurable. It has an API, a GUI and a web interface. There is considerable documentation available.

Computer Security Update (Bob Cowles - SLAC)

The first part of this was a long list of recent security exposures and an exhortation to the audience to be sure that they have applied all the security patches associated with each of them.

Solaris has suffered various security incidents of late, including the ability for hackers to connect as local root – see overheads. Similarly, there have been a few security vulnerabilities in Cisco products. But his largest list of exposures was for Linux, including one giving remote root access via misuse of the lpd daemon. He also listed 3 (only) exposures in Microsoft's IIS as well as 2 in IE.

The big advance in the "black hats" world is "Combo Worms" – when a virus exploits an exposure in one system to get behind the firewall and then search for an exposure elsewhere. An example is a combination of a Solaris bug and, the eventual target, an IIS exposure. CodeRed was another example of a Combo Worm by exploiting e-mail and IIS; nimdA was another, exploiting e-mail, IIS and web servers. All these can be thwarted by patching the known bugs as quickly as possible, although nimdA was really ingenious.

He recently attended DefCon 9 – a conference for real hackers, held annually in Las Vegas, where the latest hacks and tricks are discussed by the hacking community. Various presentations were given on how to intrude on US government sites, how to snoop on the network, how to design small payloads on e-mails (to create buffer overflows), etc. And the talks are available on CDs! At least one speaker (a Russian, on breaking ebook security) was arrested by the FBI the day after his talk!

A particularly scary talk was on kis – kernel intrusion system. Dropped into a Linux kernel, for example after compromising local root, it is virtually undetectable and allows a remote user to execute commands, substitute commands requested by a local unsuspecting user, and so on. More information is trivially available via a Google search on kis. And "available" in this sense means available to potential hackers! It has been thus available since July.

His final message was that firewalls are no substitute for applying all the latest patches. Poor system administration is still a major problem.

NERSC network performance and tuning (Shane Canon - NERSC)

The approach is to monitor everything and react quickly. They try to stay nimble and make use of the latest technology. The Oakland site is connected directly to ESnet by an OC12 link. The backbone in the Centre is Gbit. They are looking at OC48 for the ESnet and perhaps also a 10Gbit uplink.

Common problems include traffic congestion, dropped packets, and latency issues in TCP transfers. They operate optical taps on all switches and plot all traffic patterns. Tools in use include Traceroute, Netperf, Tcpdump, TCPtrace, and Pathchar (although this last one is of debatable value).

He gave 2 examples where poor network performance between NERSC and other sites (Oak Ridge and BNL respectively) were traced to specific problems and solutions were applied. In one case, they were able to persuade Cisco to acknowledge and fix a problem in their software.

Site Reports

Fermilab

The strong authentication project rollout continues and is progressing fairly well; the deadline is Dec 31st for simple applications to be Kerberised. Significant resources have applied to the project and there have been some issues with LSF and Kerberos – now resolved. Meetings and seminars were held for users. The various tools are distributed via UPS/UPD. Documentation is in preparation. Access will be forced to be with a Kerberos token or using a Cryptocard. For example, access to the FNALU interactive cluster will need to be via ssh or with a Kerberos ticket from mid-October and ssh will be disabled from December. They still use LSF on the FNALU cluster, upgrading to 4.1 this week.

Run II – the accelerator is down right now for a 5-week shutdown. CDF has 89M events – 22TB. Hope to reconstruct 3 to 3.5M events per day on their farm. No information for D0. Farm systems growing – CDF has 154 worker nodes, D0 has 90 + 32 in burn-in. Both use SGI as I/O nodes. Issues with NIS/NFS – timeouts on I/O. Working to eliminate the need for either as an FY2002 project. FBS now supports Kerberos but now independent of L.SF. Will continue to use home-grown tools for cluster mgmt.

W2K migration proceeding – desktops being upgraded where possible – see separate talk.

HPSS decommissioning in progress; data migrated off to ENSTORE. Most experiments (not CDF, at least not yet) using ENSTORE as the path to mass storage.

An AFS upgrade to 3.6 is planned, no precise date. Capacity now 2 TB; using OpenAFS for Linux 7.1 clients, IBM version for other platforms.

Email gateways run Sophos virus-scanning software.

CERN

I presented the CERN Site Report as prepared by Maria. Among the reactions from the audience, several people claimed that the Open AFS client for W2000 works just fine, no problems, reasonable performance. Is CERN using the latest version?

Jefferson Lab

Using Solaris 2.6, HP-UX 10.2 and some 11; Redhat Linux 6.2 and 7.1 on some servers. Still using Windows NT, just starting to look at W2000. Completed migration from OSM for mass storage access to JASmine. Linux batch cluster uses a mix of SCSI and IDE discs; recently added 60 more dual-CPU PCs. Lot of work into Jpasswd in particular for UNIX/Windows/Calendar password synch; added password expiration and a web scheme for account mgmt with group/manager authorisation. Using the Corporate Time Calendar tool.

A Windows Terminal Server is in production and they are thinking of adding a second. Looking at various thin clients to replace X terms – Wyse seems easiest to set up at least, as well as being the cheapest ($300). Chose Workstations Solutions QuickRestore to replace Budtool.

Telnet turned off. Will use SafeTP for secured access only. CERN/JLab printing collaboration put on hold – resources assigned over the summer to install second STK silo. They use SilkyMail and the Trend Micro VirusWall scanning tool.

LAL

They have moved most central storage to a Fibrechannel SAN and they will attach a tape library later this year. Current controllers are Compaq but they hope to move to Solaris next year.

They use the CERN Printing Package on their growing W2000 domain but need Ivan Deloose's help to complete the port; one request is to remove some CERN dependencies (printer naming conventions for example).

Standardising on Dell PCs, $3K for a 512 MB 1.2GHz model. Testing VMware V3 with its support for integrated NAT (one IP address for the PC for Windows and Linux) and support for larger virtual discs (256GB instead of 2GB as now).

Using the OpenAFS client for W2000 and Linux. They find the W2000 client very stable, with no serious problems – this was echoed by other sites represented.

IN2P3

Still based on Redhat 6.2 but looking at Solaris 8 on the SUN side in response to a BaBar request. Will upgrade some AIX systems to 4.3.3. The last HP has been switched off. They have a variety of platforms controlling 35TB of disc (SUN, IBM and Hitachi) and are working on making disc access independent of controller platform.

They have interfaced HPSS with RFIO via C, C++ and Fortran APIs and with bbftp for secure parallel high performance ftp. A 64 bit RFIO is being developed and will be interfaced to HPSS at IN2P3 by Ph.Gaillerdon and to CASTOR at CERN. The current work on BQS is to add support for parallel jobs and a Java interface is in production.

For D0 they operate a SAM server and they use HPSS as SAM cache space. IN2P3 is officially a Tier A centre for BaBar and a regional centre for D0, EROS II and AUGER.

Current work includes the recent creation of a small informal working group studying the future evolution of mass storage, and another concentrating on cheap storage and integrating BQS with Globus. They are also participating in DataGrid WP6, testbed integration.

INFN

The trend across INFN is, like everywhere, in favour of Linux and Windows 2000. Using a 622 Mbps link from Milan to New York and will add another to connect to GEANT by December. Discussions underway to move AFS servers to OpenAFS running under Linux. Still 22% of users depend on VMS sendmail! Padua is implementing a new farm for BaBar data reprocessing; initially with 120 bi-processor Linux PCs, 15TB of disc space and 60TB of tape capacity.

SLAC

Solaris 8 (and 7) is now in production. The Redhat Linux production version is still 6.2 but 7.1 has started appearing on desks and BaBar is targeting 7.2. Their Linux farm now counts 512 dual-CPU VA Linux nodes. There have been some problems such as spontaneous reboots, random hangs and some translation look-aside buffer failure. All these have been very hard to find and are ongoing. VA Linux now do not sell hardware but they have been supportive in trying to keep the systems operational.

The Solaris farm has not changed much recently but disc space continues to expand and they are now up to 30 9940 tape drives and they work well. There are now plans to replace their E10000 64-node SUN server with smaller and separate disc and CPU servers. A second HPSS service on Solaris will be installed for general staging. No HPSS problems to report. Recently crossed the half-Petabyte threshold, mostly BaBar Objectivity data.

Desktops can be updated at any moment with the latest production software from a centrally maintained distribution site.

Difficulty with a Dell SAN used on the Windows side to serve Exchange and Windows home directories. Blamed for a major Exchange failure and now replaced by SUN T3 disc systems.

PDSF/NERSC

Now situated in downtown Oakland where they inhabit a huge, and hugely expandable, computer centre. Co-located with ESnet HQ. Also in the centre are several Cray supercomputers, the Alvarez cluster and the NERSC 3 IBM SP system (see earlier talk). Certainly the most powerful centre I have walked through.

PDSF focuses on commodity hardware, namely a Linux cluster based on 1U and 2U PIIIs with up to 2GB of memory. There started with 160 dual-CPU compute nodes (the 1U nodes), now 190, plus 34 2U I/O nodes for a total of 16,000 SPECints. They use cfengine for system consistency. Looking at Athlon chips as well as P4s and IA64 for future purchases. PDSF customers include many HEP experiments such as STAR, ALICE, ATLAS, CDF and others.

RAL

Expect to get £17M (42M CHF) from PPARC for the UK Grid project, although only a small portion actually approved for next year, to cover staff costs, some equipment and the UK contribution to the LHC Computing Grid project. They also got 12M CHF to set up an e-science Centre, taking staff from the RAL IT Department. Should enable collaboration between Grid projects.

Linux farm now 240 processors plus a SUN E4500 and 4 420s for BaBar. CDF have been installing a small PC/Linux simulation farm. Using the funds allocated, they are currently out to tender now for equipment for a BaBar Tier A centre combined with a Grid Tier 1 centre. It will probably end with about 200 CPU PIII Linux systems and about 50TB of disc space, either SCSI or IDE.

The official desktop is still W-NT with Office 97, mail via Outlook 98 and an Exchange 5.5 server. Supported notebooks now run W2000 and some servers now run W2000 but no date for a general migration to W2000.

Investigating under what conditions can home systems connect to RAL, in particular those who wish to connect "behind" the RAL firewall. Looking at funding virus protection and firewall software for home systems.

DAPNIA

User workstations are either W2000 or Linux – dual boot is forbidden for desktops but not laptops. There are plans to upgrade the Solaris service for their particle physics users. They plan to migrate from Euclid on Compaq to Catia on Windows. Working on plans to amalgamate the civil and military sides of CEA at one site and a small cluster of PCs has been installed to investigate this.

Microsoft and Unix Kerberos Interoperability (Chris Brew - Fermilab)

Work done as part of Fermilab's Strong Authentication Project which is largely based on Kerberos authentication and which targets single password, single sign-on. W-NT4 cannot authenticate via Kerberos but W2000 can so it should be part of the project although W2000 is not yet widespread at the lab. Several options for this range from separate UNIX and Windows domains (but then the user has to have 2 passwords) to integrating the domains with one or other as the master.

The various benefits and caveats with making either the UNIX domain or the Kerberos domain as master were listed. Microsoft's extension to Kerberos, where certain Windows information is included in their Kerberos ticket, does not make it easy to have the UNIX domain as the master and making the Windows domain as the master loses AFS and Cryptocard support.

Eventually decided to try to use the UNIX domain as master and a series of steps were taken to set this up (see overheads) including setting up the necessary trust relationships. Problems found include that users cannot change passwords from W2000 and they have been unable to get Cryptocard login to work via Windows because of delays in the notification although they believe this works elsewhere, e.g. MIT.

They have concerns about future support given the modifications that they would need to the Kerberos usage in W2000 and the fact that very few other sites have made an NT domain subservient to a UNIX domain. Plus the possible risks associated with not using Microsoft's standard login on other commercial software and future Windows versions or patches. All these risks, plus the estimation of support for such a complex solution have decided them to use the native Kerberos on each side and to establish a 2-way trust between the 2 domains and they will see what can be done to keep the passwords in synch between the 2 domains as a long term goal.

NGOP Status Report (Tanya Levshina - FERMILAB)

NGOP goals include monitoring service status as well as node status and performance analysis although this last is still under discussion. Since the rollout of the prototype (described in the report at the previous HEPiX in Paris) they have redesigned the configuration language to use XML. In Sep, they rolled out a production version. Now they monitor 524 nodes, 9000 objects and there are typically 5 instances of the GUI running at any time.

The monitor agent has some new features. Some system administrators refused to permit long-running daemons on their services and demanded a version of the agent running as a cron job. Agents can execute any sequence of UNIX commands. They provide a range of agents to monitor the most common UNIX services and conditions.

They intend to work later on a better GUI so the present GUI remains rather rudimentary for the moment. They have interfaced NGOP to Remedy to generate trouble tickets automatically although this is only done for a few situations at present (about 3-4 tickets per day). For this they made use of the Remedy API.

Next plans include a web interface, a monitoring client API, Correlation Agents, historical rules and escalations, adding W2000 monitoring agents, creating a performance analysis framework and scaling up to handle 1000 nodes. Their conclusions so far are that use of the prototype and now the production installation have increased system uptime and efficiency.

FERMI PC Vendor Qualification Update (Lisa Giacchetti - FERMILAB)

FNAL have a procedure to qualify a limited number of PC vendors for limited periods of time, described at previous HEPiX meetings. They found that many groups inside and outside the lab then used this list of vendors as their first choice. The vendor's ability to install the FNAL release of Linux was important. However, the fast-moving PC technology has been an ongoing concern and no vendor is immune from the effects. Despite this, the number of onsite Linux PCs has doubled in the past year.

FNAL have performed a new qualification this year according to a well documented procedure involving pre-determined specifications, a series of tests on software modules provided by FNAL (who also provided the preferred release of Linux) and defined criteria in delivery and support. All these were agreed in a meeting with all interested vendors. Vendors could bid for desktops or farm nodes or both.

Once test systems were delivered, with the tests having been performed by the vendors in-house, FNAL repeated the tests on site. I have copies of the handouts that were given out at the meeting with the tests and the results (and the prices!). Systems were graded

according to the results. Vendors were promised there would be at least 100-150 farm systems bought in the near future.

They contacted 18 vendors this time of which only 2 remained from the previous round of qualification. Three dropped out quickly, including Compaq. 3 sent only desktops (including SGI and IBM), 9 sent both farm nodes and desktops and Linux NetworX sent only a farm node.

Eventually decided to qualify 6 vendors for desktop purchases, including SGI; 5 of them use the Athlon/Thunderbird processor and RDRAM and DDR SDRAM memory systems performed well. 5 vendors were qualified for farm nodes; no major vendor in this second list (not even Dell as had been expected).

Following this, they purchased 10 systems from each of the top 2 desktop suppliers, negotiating fixed configurations, and several farm purchases were made for a total of 136 nodes. They have now realised they may also have to qualify a server supplier. There were a number of kinks in the farm purchase side mainly due to the decision to go for rack mounted units for the first time. Then there were many problems relating to installing Linux on the system disc drives; the vendor changed discs, cables and the rate of errors dropped enough to accept burn-in. But now errors are being seen on data drives. Eventually they traced most problems to the chosen motherboard with the Serverworks LE chipset and the to use of IDE discs with DMA turned on. [Two other sites present at the meeting had already reported this.] The problem has now been agreed by Serverworks and they are working together on a solution.

Despite, they believe the qualification is a success and it will be repeated every 18/24 months with mini-qualifications more frequently. They accept it will not avoid all hardware/OS issues and they are currently unsure what hardware for the next purchase should be. And they strongly believe a 30-day burn-in is essential.

NERSC Security Practices (Steve Lau - NERSC)

NERSC has no classified work, thus they are only constrained to implement the lowest level of security as defined by US Govt. but they have a diverse user community spread across the world and staff separated between 2 sites linked via ESnet which they consider as an external unsecured network. How to relate this to security, be an open site but secure from attack and disruption. Their primary tools are intrusion detection and platform hardening. They get scanned about 30-40 times per day and note that the scanning rate is increasing.

There are 3 layers of security –

1. They monitor links using the BRO software, a locally written package.
2. All latest security patches are applied.
3. And there are local firewalls available on demand for given systems or nodes.

VirusWall checks incoming mail but the primary detection system is BRO – it passively monitors a network link while it is open. It allows a site to write scripts to detect patterns that can be allowed or forbidden. It records all sessions and keystrokes. It is set to ignore some data, eg. the content of ftp transfers. It detects signatures of anti-social transfers and file sharing such as Gnutella and NAPSTAR. BRO permits the security staff to reconstruct an intrusion. It can be quickly adapted to detect signatures of new viruses, for example for the Code Red worm on July 19th. It can monitor up to OC12 link speeds. Freely accessible – see http://www.aciri.org/vern/bro-info.html.

Safe security –

- Maintain a good backup
- Keep up with the latest patches
- Use virus checkers and get the latest pattern file
- Eliminate clear text passwords, use ssh where possible
- Disable not-needed services
- Don't use IIS
- Disable open shares on Windows
- Don't run executable e-mail attachments
- Use good passwords
- Use host-based firewalls
- Scan your own workstation but DON'T SCAN OTHER PEOPLE'S. Use NESSUS or NBUS.

HEPNT

NT4 & W2K File Permission Incompatibilities; Is Microsoft Premier Support Needed? (Andrea Chan - SLAC)

The first part was a long list of bugs found by her group and how they were fixed. Details, if you really care, are on her overheads.

Is Microsoft premium support needed? SLAC now shares a Microsoft Technical Account Manager (TAM) with Stanford Uni, SLAC's share being 25%. This person acts as a central point into Microsoft to co-ordinate technical, consulting, sales and marketing support for Stanford and SLAC. Compared with Premium support used before, this is a major improvement giving faster responses to problems and advance information on fixes. The TAM is able to get SLAC in touch with the correct level of resource inside Microsoft. The annual cost is definitely considered better value than Premium support and recommended for mission-critical Microsoft services such as Exchange, Active Directory and others. And the cost? The TAM costs some $300K per year. SLAC pays 25% as their share.

Exchange and Email Anti-Virus at SLAC (Teresa Downey - SLAC)

[SLAC UNIX users still use a purely UNIX solution; this talk only concerns Windows users.] Last year they replaced a Uni Washington IMAP server and the Eudora client by Exchange 5 and adopted Outlook 2000 as the recommended Windows mail and calendar client. They have added web mail accessibility and since it uses IMAP, Netscape clients are possible. The UNIX users accessing via Windows can store mail on a UNIX mail spool via NFS.

Mail is subject to anti-virus measures namely MTA-PMDF on the Solaris mail gateway and CA/Sybari on the Exchange server. Various attachments (.exe, .com, etc) are stripped off and SPAM is blocked (some 6000 domains are currently blocked!). 75% of the Windows clients run Innoculan but not always with the most recent pattern file.

This year, the mail database was moved from a Dell SAN, now deemed "immature", to Sun StorEdge T3 RAID system. The single large Exchange server with 1500 (almost all Windows) users is being split into 4 smaller ones and they now feel they should have done this earlier before it grew so large.

No plans yet for Exchange 2000; waiting for SLAC to move towards Windows 2000 and they reckon "a year or two down the road".

The Fermilab Windows 2000 Domain Structure (Tim Doody - Fermilab)

Today there is a collection of largely independent NT4 workgroups (domains) on the site. Like CERN they have established a migration working group, led by Jack Schwarz, and including the main users, to discuss moving towards a centralised W2000 infrastructure. An additional incentive for moving towards more centralised operation is an official DoE instruction to implement better security, for example to have a single point where accounts can be authorised, or perhaps more importantly, blocked.

A root domain (win.fnal.gov) has been agreed and sub-domains are being negotiated with major user groups such as Business Systems, CDF and D0 Controls, etc. Local sub-domain controllers will have some flexibility to set policies for their domains but user accounts can only be created at the root domain level; the sub-domains, apart from the main Fermi sub-domain (Fermi.win.fnal.gov) where the desktops "reside", can only have machine accounts.

The target is to have the domain controllers for the root domain in place by Nov 1st along with those for the Fermi sub-domain. Then desktops can start to be moved. For the moment, they are actively encouraging users to migrate from NT4 to W2000 but to remain for the moment in the NT4 domain. Migration should in fact be a fresh installation, as recommended at CERN also, although a few users groups are insisting on upgrades. The goal is to have all Computer Department supported desktops migrated to W2000 but still in the NT4 domain by the third week of November.

A core system image has been defined with the most common tools and applications and other images with more specialised applications can be created for particular user groups as needed. Rollout of these images is very fast, 15 to 20 minutes per node is a typical installation time.

The current plan (2 weeks before their Nov 1st deadline) is to install the initial Domain Controller and establish the necessary trusts between the W2000 and NT4 domains. Then the W2000 desktops will be migrated into the Fermi.win.fnal.gov domain and then they will migrate the servers and the Wincenter servers. The overall goal is to complete this sequence by Jan 1st 2002.

Issues under discussion in the migration working group include

- User accounts, especially in relation with the Strong Authentication project. According to a policy decision, a Windows user must first have an account in the MIT Kerberos domain (UNIX) and thus the account name is subject to UNIX account restrictions (maximum 8 characters for example). Some accounts previously only used in Windows domains will now conflict with existing UNIX user accounts.
- Forcing naming conventions where none existed before
- Roaming profiles especially when moving back and forth between NT4 and W2000.

Windows 2000 Co-ordination Group status report (Michel Jouvin - LAL)

Michel, on behalf of Gian Pierro Siroli, reported on the W2000 co-ordination group, which met in CERN in July. I guess those most interested, namely IS Group, know it all and I refer others to read the overheads on the web. One point he raised concerned the Printing Package and whether effort should be found to make it more open source and applicable to other sites (LAL found some CERN dependencies, at least in the Windows part).

DESY Windows Report (Henner Bartels - DESY)

Today there are some 2400 users in their W-NT4 domain with only a few users on W2000. They install and update their stations with Netinstall that still has no support for W2000. In August a proposal for a W2000 deployment was agreed along with a re-organisation of the IT department and a new separation of responsibilities. This last has forced reconsideration of the detailed planning and there is still no clear time-scale for the migration and a great deal of uncertainty in the Windows team.

Alan Silverman
20 Oct 2001