HEPiX-HEPNT - TRIUMF, Vancouver

October 20-24th, 2003

Trip Report

Alan Silverman (with help from Helge Meinhard and others)

Introduction

This was the second HEPiX meeting held in TRIUMF. It attracted some 73 attendees, a high number for a meeting in North America. The first 3 days were devoted to general sessions, the fourth to security topics, organised by the HEPiX Large System SIG, and the last morning to a brain-storming on mass storage systems organised by Olof Barring in view of an expanded mass storage system session at the next meeting.

To explain the length and the different styles of this report (bullets for the first day, narrative for the rest), I missed the first day, on CHEP duty in SLAC. So Helge has very kindly sent me his notes from Day 1 which are included verbatim. Helge also contributed the notes on the Redhat support policy discussion (see start of day 3). As usual, almost all the overheads are present on the conference web site under the Programme tab.

Organisation

The meeting was extremely well organised by Corrie Kost of TRIUMF with a (small) band of helpers. He coped with late agenda changes, including the addition of 2 vendor demos within 2 weeks of the meeting; he added bus transport between the hotel and TRIUMF and organised web-casting of the sessions. He had even scouted out a number of restaurants within walking distance of the hotel which would be able to cope with large groups arriving unannounced during the week of the conference. He also appointed himself official photographer and the results can be viewed at http://www.triumf.ca/hepix2003/Pictures/

The next meeting will be May 24th to 28th in Edinburgh, Scotland and the meeting next Autumn is likely to be in BNL, the first which would be held there.

Highlights

- High attendance for a meeting in North America – the attraction of Vancouver as a site and security as a subject were cited as possible reasons plus the increasing visibility of HEPiX
- Excellent organisation
- Redhat discussions: although the Redhat speaker did not tell us anything new (and several attendees expressed disappointment that he did not announce the solution to our current problem), he did impress some of us with his sense of commitment to finding a solution for the current support situation. It seems clear that Redhat themselves are still working through the ramifications of their new release policy. [And discussions continue by e-mail since the meeting.]
- The internal discussion on Redhat the following day seemed to confirm that the negotiations being undertaken by SLAC, FNAL and CERN as representatives of the wider HEP community are supported (although not all sites were represented by decision makers)
- First (?) appearance of vendor exhibits at HEPiX plus talks from commercial vendors (especially Redhat and Microsoft).

- Quality of the presentations – I heard more than one comment on the quality of the overheads submitted by many speakers, several from CERN being noted in particular.
- Sharing of code among HEP sites (CERN's print manager in use in LAL; SLAC's console management tool being adopted by CERN via a collaboration; SLAC's monitoring tool used in DESY Zeuthen; etc)
- A very interactive Large Security SIG all-day session on the topic of Security. The security officers of all the major sites were present and the atmosphere often was more of an interactive workshop.
- The forming of a forum for Mass Storage Systems' interoperability.

Vendor Exhibits

Ibrix – sells a "global file system" aimed at offering a file system solution for high performance computing sites. From the viewpoint of 10,000 feet, it uses software implemented via a kernel mod to speed up file access from disc – keeping track of active files via metadata and careful file placement according to use patterns rather than file hierarchies. Fermlab evaluated this and similar products and chose this one, possibly because it is a software-only solution and therefore less risky that one combining hardware and software. First delivery to FNAL, in fact Ibrix's first production delivery anywhere, should happen in January. Jan Iven and I spoke to them and proposed to host a visit in spring of next year after they have some practical production experience to report. It can be expected that Fermilab will report on progress at future HEPiX meetings.

Panasas – another contender to solve cluster storage problems, this one based on "an object-based architecture". This product however includes the hardware, based on blade servers. The CTO of Panasas also gave a tutorial on his product (end of Day 3).

Day 1

Welcome (Jean-Michel Poutissou, TRIUMF)

- Triumf: national lab for sub-atomic physics; National funding: 200 CAD over five years. Organised as consortium of 11 universities
- Mission: cyclotron base program, ISAC for nuclear physics, support for activities abroad (CERN, SLAC, DESY). Main facility: 500 MeV cyclotron of the 70s. Using muons for condensed matter and chemistry research. Accelerating radioactive beams now. Isotope facility mostly used for medical research
- ISAC scientific programme: nuclear astrophysics, fundamental symmetries, nuclear structure, condensed matter, life sciences. Nuclear astrophysics: looking at isotopes with unusually high or low number of neutrons
- Particle physics and other fields: polarized muon decay, muon spin resonance, beta-NMR, muonium chemistry
- Medical imaging, isotope therapy, proton irradiation
- Infrastructure support: ATLAS at LHC (HEC LAR calorimetry, Tier 1.5 centre); Rare K decay programme at BNL (KOPIO); Long baseline neutrino oscillation experiment at KEK (J-PARC); R&D for experiments at linear collider
- Accelerators: LHC: warm dual quadrupole magnet, kicker magnets, …. Total 45 MCAD; AGS upgrades; …
- A little history: last HEPiX meeting at Triumf: April 1996. CERN had 5 TB of disk space, X stations dominated, CERNVM being de-commissioned (still 2300 users!) Desire to move to DFS due to AFS problems. Struggling with how to handle UNIX releases and deal with patches. Big

disks were 9 GB Seagates. Many sites moving from VMS to UNIX. Linux being rolled out. Windows 3.11 migrating to Windows NT. Mail services to move from POP to IMAP. Many emerging choices on handling batch queues and providing user interfaces

Site Reports

Oxford University Particle Physics (Pete Gronbech)

- One of the larges particle physics department in UK
- Using central physics support facilities: E-mail hubs (2.7 M messages relayed, 2.8 M rejected as spam), Windows terminal servers (on 2000 and 2003 on an 8-way CPU server), Web / Database, central Exchange servers (moving to Exchange 2003 now).
- Exchange 2003 comes with a very nice Web interface
- Desktops: about 200 Windows 2000 PCs with Exceed (migration from NT almost completed, XP Professional is default for new machines)
- Migration to Linux: central Unix systems running Linux, mainly RedHat 7.3. Linux on desktop being discouraged. DUX and VMS services closed in 2002. Batch farm growing (now 8 dual-processor machines). Using SCSI disks for file servers, had problems with IDE disks
- Monitoring of the farms: using Ganglia
- CDF Linux system: Morpheus, an 8-way 700 MHz Xeon system with 8 GB RAM and 1 TB Fibre Channel disks (IBM x370), plus 10 dual 2.4 GHz Xeons (Dell). Approx. 7.5 TB SCSI disks attached to master node as well
- Providing test bed for EDG as well
- Backup: Netvault running on Sony AIT3 drives (100 GB per cartridge)
- Network connectivity: Super Janet 4 (2.4 Gb/s). Bottleneck was 100 Mb/s between university computer centre and particle physics. Upgraded in 2002
- Network security: Firewall based on stateful inspection introduced. Rejecting 26000 connections per hour. Special protection (VPN) e.g. for Atlas construction area. Better control needed over how laptops access networks – considering NAT in order to insulate laptops
- Sophos site licence, update via Web pages
- Keeping OSs patched is a challenge !
- Plans: continue network security, reduce number of OSs. Problems: laptops, RedHat choice. Looking into single sign-on (just starting…)
- Q&A: IDE disk problem: referring to RAL, problems seen with 12-port arrays using bad IDE disks (running out of map space for bad blocks)

IN2P3 (Thomas Kachelhoffer)

- Batch computing: 70 processors under RH6.1, ~830 processors under RH 7.2, processors are mixture of PIII and P4. Also running Solaris 8 and AIX 5.1. Next: try blades, AMD Opterons; doubts about RedHat. Scheduler: BQS, serving 40 groups and 2700 users on 3 platforms, 6 OS, 6 hardware variants. All batch nodes accessible to everybody, no dedicated CPUs. About 1200 jobs running in parallel, up to 15'000 jobs per day.
- Storage: 6 STK Powderhorns, using STK 9940 and few 9940B, drives connected via SCSI and/or FC. DLT for export/import
- Disk space: ~60 TB on SCSI and/or FC (for AFS, HPSS cache, Objectivity, Oracle/Postgrsql/MySQL, Xstage, NFS
- Network: Direct access (1 Gbps) to French national academic network, from there to GEANT. Dedicated line to CERN (1 Gbps)
- Services: 2700 accounts (90% French, 10% international). 40 Experiments (HEP, nuclear and astrophysics, biology). Regional centre for Babar, D0, LHC, …

- HPSS through rfio reaching 10 TB/day transfer rate. Bbftp being used for data transfer in particular with US sites
- Batch scheduler: BQS (home-made)
- Grid involvements, database services (Objectivity, relational ODBMS), SRB, SAM
- General services: mail, news, listserv, DNS, EDMS, Web, LDAP, Webcast and streaming, videoconferencing (MCU IP & ISDN); Virus filtering, looking into Anti-spam. 6 gcc versions on Linux
- Specifics: no on-site user, no accelerator/experiment. All resources are shared and common (just very small dedicated grid testbed for software development). Limits being reached on UPS and cooling system.
- Coming next: 400 processors (Opteron?), 30 TB disks, more 9940B drives and tapes, servers for Objectivity and HPSS, HPSS supporting SAN, Solaris 9, Linux???, Kerberos 5, load balancing for Objy and xrootd with dynamic staging
- Q&A: Why SCSI/FC disks, not IDE or SATA? Want to buy 20 or 30 TB in one single unit. Found SCSI / FC disks still the best compromise
- What are your Objectivity servers? Solaris, no plan to migrate to Linux
- What have you done in terms of testing Opterons? Nothing on site, but relying on Babar tests. Not expecting major problems, running tests on 2-way machine now

SLAC (Chuck Boeheim)

- Experiment status: BaBar: resumed data taking in September, good startup at high rates. Converting to new computing model (root-based mini instead of Objectivity, but does not entirely get rid of Objy). Crunch on disk space because both systems being kept in parallel for some time. Glast: Simulations and pre-flight data
- Storage: expanded from 150 TB to 250 TB in 2002 in terms of disk space. HPSS migrated from AIX to Solaris, upgrading from 9940 to 9940B at the same time, HPSS upgrade to latest version – all remarkably smooth. Recently, some problems of tape thrashing (disk cache exceeded)
- OS: Babar on Redhat 7.2, some RH9. RH6.2 is gone. Kernel 2.4.20-20 ("kernel du jour"), investigating RHEL 3.0. Problem with NFS client mounts seen (both Linux and Solaris servers). Solaris 2.6 gone, 2.7 being phased out, 9 is production. Looking at Solaris-x86
- AFS: OpenAFS 1.2.9 on all file and database servers. Key changing broken, fixed in 1.2.10, otherwise very few problems with OpenAFS. AFS Perl module enhancements (bos and vos interfaces) contributed back to maintainer, soon on CPAN. AFS Best Practices workshop February 04 – 06 2004, see http://www.slac.stanford.edu/~alfw/OpenAFS_Best.pdf, complementing the LISA and Usenix workshops. Topics: scalability, backup, monitoring, management, …
- Kerberos: Unix: Heimdal K5 in test, sync passwords from kaserver to K5. Providing K5 tickets, AFS tokens, K524 service. Local tool conversion still to be completed. Windows: Active directory KDC. Project: synchronise passwords between KDC (appearance of single password for all services) – working on synchronization using native tools e.g. with Microsoft
- Windows status: XP migration scheduled to complete by end Dec 03 (400 machines still left of 1700 total). Moving from Exchange 5.5 to 2003 (skipping 2000), from Citrix Metaframe 1.8 to Xpe. Monitoring: NetIQ AppManager, investigating Application servers/J2EE. Discussions about level of service and management for desktops
- Q&A: Babar event store in Root, conditions and some metadata still on Objy

Triumf (Corrie Cost)

- Lean – not many staff involved
- WestGrid: Triumf/UBC first HEP site to install large blade cluster – 504 dual 3.06 GHz Xeon blades, RH7.3, PBS with Maui, 10 TB disk, 70TB tape, direct Gigabit connection between sites, upgradable to 10 Gbps. Subatomic physics allocation: 1/3 of total resources. Some delays

experienced. Blade power consumption exceeded expectations – hence using only 12 of 14 blades in a chassis, power supplies will be changed. Some teething problems of GPFS. Reduced cabling very welcome (wireless would still be better…). Very little infant mortality – not a single blade failed! 14 blades in a 7U chassis, 4 Gigabit connection
- First look at dual 1.6 GHz Opterons with SATA IDE drives. Found performance about equal to 2.8 GHz Xeons
- Misc.: 1 TB Raid 5 disks attached to Linux for Linux and Windows (via Samba) users. To be covered later: Mosix & PBS, bandwidth management, Colubris Radius server, SPAM fighting. Document management system: Docushare. Windows 2003 terminal server starting to be deployed
- Involved in 10 Gbps link to CERN, achieved 5.5 Gbps sustained

LAL (Michel Jouvin)
- Main storage system: Hitachi disk array (installed after last meeting) – 1.5 TB added, running very well. 10 dual Xeon on order (1U). Budget uncertainty still… No OS upgrades since last time
- Mail: virus and spam filtering (Spam Assassin + MimeDefang) tagging, mails still processed by clients. Less efficient since September… Delay with filtering at message delivery (SIEVE), upgrade of IMAP server to cyrus v2, authenticated SMTP
- Windows infrastructure: IN2P3 forest in production with 7 labs, 4 labs to join before end 03. No problem so far, even if master domain not reachable. Migration of LAL domain delayed, planned in November. Will go back to NT4 domain, and re-upgrade to ActiveDirectory (direct migration being more difficult)
- Viruses: Blaster/Lovesan didn't affect LAL too much. 7 PCs found infected, infected outside LAL, but continuing. Main tool: SMS for rapid deployment of fixes – 90% of running PCs patched in 2 hours. Infected PCs detected by inventory, can take 3 days to discover… They are investigating automatic isolation of weak/infected PCs.
- VPN server: W2000 server based on solution presented by CERN at Hepix@Fermilab. Tunnelling: PPTP+CHAP/MS-CHAP, encryption, works with Macintosh and supposedly Linux. Plan to investigate RADIUS authentication. Performance impact unclear
- Misc.: Unattended Linux installation server: currently based on PXE and kickstart, plan to investigate WP4 Quattor. Web content management: abandoned Zope, gone for home-grown solution (J. Charbonnel). Strong involvement in EDG/LCG/EGEE
- Q&A: What version of SpamAssassin is LAL using? Latest one – but spammers know the weaknesses…
- Q&A: Why use RADIUS for VPN authentication? RADIUS being looked at for potential authentication of wireless users
- Q&A: What about home machines connecting to the Lab? Seems to be a big problem, clear instructions missing. LAL providing instructions for home machines, but indeed not an easy problem

CERN (Helge Meinhard)
- See overheads

RAL (Martin Bly)
- Grid activities: LCG, Sam, EDG, UK QCD, SRB, …
- EDG status: 2.0.x deployed on production testbed since early September, 2.1 deployed on development testbed, EDG 2.0 gatekeeper as gateway into main CSF production farm, providing access for BaBar and Atlas. Further grid integration into production farms via LCG, not EDG
- LCG integration: LCG-0 mini testbed deployed in march, LCG-1 in July, upgraded in August/September. Worker nodes: 2 * 1 GHz, 1 GB, 540 GB. Will soon need to make decisions

about how much hardware to deploy in LCG, driven by experiments. Difficulties: installation and configuration, documentation, support, reliability
- SRB services (Storage Resource Broker) for CMS, based on Oracle, some issues that have been attacked with very good support of UC San Diego – considerable learning experience
- P4 Xeon experience: Disappointing performance, hoping for 1.5 performance over P3@1.4 GHz, getting 1.2…1.3. Hyperthreading not necessarily beneficial… (Linux CPU scheduling?) Performance better with Intel compiler, get 1.5 there. Efforts to run O(1) scheduler unsuccessful. Another issue: CPU accounting now depends on number of jobs running. Beginning to look closely at Opteron solutions…
- Datastore upgrade: STK 9310 robot with 6000 slots, phasing out IBM 3590 drives in favour of 9940B drives
- OS: RH6.2 closed end August, 7.2 upgraded to 7.3 during October for BaBar, residual 7.2 service closing soon. 7.3 main work horse now, bulk service opening soon. What to do next? Solaris 2.6…2.9 boxes being run as well, RH Advanced Server is around, too.
- Next procurement: 250 dual processors, …
- New helpdesk, replacing Remedy by Request Tracker
- YUMIT: RPM monitoring based on UYM
- Exception monitoring: Nagios as interim solution, waiting for EDG
- Outstanding issues: Many new developments and services. P4 poorly performing. RH support policy
- Q&A: NFS client problem seems to be fixed in new kernel releases

GSI (Walter Schoen)
- GSI: Member of Helmholtz-Gesellschaft, heavy ion lab – Darmstadtium element 110, cancer therapy project, condensed baryonic matter, external engagement (ALICE)
- Linux: central services (new Web, new mail), desktop, scientific computing (number crunching, analysis, file serving for experiments). Now about 400 Linux computers, 10 file servers, 10 group servers, 10 compute servers. All running Debian
- AIX and Windows servers being migrated to Linux (mail, DHCP, DNS)
- Mail: postfix on Linux, spamassassin, additional virus filters (complement Windows filters) [In conversation later with Alberto, it emerged that they use Exchange as their e-mail server.]
- Batch farms: 224 CPUs (96 CPUs PIII@600 MHz, ATX midi-towers, 128 CPUs Xeon 2.4 GHz 19" 2U). 100 more CPUs to come in spring
- Windows: Active directory with 2 domain controllers, 2 print servers, 2 file servers, 2 application servers, 2 Web servers. Windows 2003 server being tested
- Security: Security group established, developed concept for a GSI firewall with stateful routing, DMZ (being tested)
- Q&A: What are you using as anti-virus on mail server? Windows client: McAfee, Linux server: Sophos
- Q&A: What size is the security group? 4 people part-time (Walter himself is 50% security, 50% IT stuff), plus some people at IT Division for maintenance and support
- Q&A: What about DNS and Active Directory? Active Directory DNS will continue to be provide by a Windows Service
- Q&A: What services will you put into DMZ? Mail proxy, Web proxy, some more
- Q&A: What about Oracle and Debian? It's a disaster…

BNL (Ofer Rind)
- RHIC: DOE funded lab, research on site, support for Atlas

- Mass storage: 4 STK silos managed by HPSS, upgraded to 37 9940B drives. Aggregate bandwidth: 700 MB/s, expect 300 MB/s from experiments in next runs. 9 data movers with 9 TB of disk space
- Disk storage: Large SAN served via NFS for processed data store and user home directories, 16 brocade switches, 150 TB of Fibre Channel Raid5 managed by Veritas
- Linux farm: 1100 dual Intel CPU machines, totalling 918 kSpecInt. Allocated by experiment and further divided for reconstruction and analysis. 1 GB mem, 1.5 GB swap, Combination of local disks totals to > 120 TB available to users – plan to make use of it systematically. Most machines now at RH8, Atlas still at 7.3. Installation of customized image via kickstart. Reconstruction farms managed by python-based custom frontend to Condor. Analysis managed through lsf 5.1, condor is being deployed and tested as possible replacement
- Security and authentication: two layers of firewall, conversion to Kerberos5-based single sign-on paradigm. For RHIC, AFS/K5 fully integrated. Atlas: K4/K5 parallel authentication paths. Ongoing work: Integrate K5/AFS with LSF
- US Atlas Grid Testbed
- Monitoring: using a cornucopia of vendor-provided, open-source and home-grown software. Ganglia deployed on entire farm. Alert scripts re-implemented using direct TCP/IP pushes. Cluster management software is a requirement for Linux farm purchases (CACM, xCAT) – console access, power cycling, …
- Future plans: Linux farm to be expanded (100 nodes – 2U servers with local disk), more scalable solutions for file serving (Panasys, dCache); grid services
- Q&A: What version is Kerberos? How is Windows integrated? Kerberos version is MIT

DESY (Peter van der Reest)

- Site wide virus checks on E-mail, spam tagging (well accepted by user community, but continuous monitoring important)
- Heimdal/AFS environment in Zeuthen very stable, Hamburg preparing for transition
- Linux: DESY Linux will be based on SuSE (mainly on user demand – no changes now), new setup routines being prepared, new central software distribution concept (no longer AFS based). For future releases, considering Debian as well
- Windows migration to XP: Test phase for XP domain with selected users in August 2003. Workstations installed entirely via RIS. Still a lot of work to make applications available (Office, Exceed, SSH, OpenAFS, Acrobat). Production planned for January 2004
- User registry: So far using 3…4 different data bases, consolidating in a project involving 3…4 FTEs. Test environment: Jboss, STRUTS. Implementation underway now, Unix services to follow by end 2003
- Use of F5 switches: recent mail migration used them in high-load production. Transparent for the user, except a new certificate needs to be accepted by client
- dCache: now deployed in San Diego, Wisconsin, UWisc, Karlsruhe, Geneva, BNL. Now ~50 TB of cache space DESY-wide. Servers: Sun hardware@Linux prices
- Grid activities: centrally supported pool (EDG 1.4) for all Hera groups, based on SuSE (all tools ported). Cooperation with QM London. dCache has been gridified, in use at FNAL. DESY is a cooperating member of the EGEE, and founding member of dGrid (with Karlsruhe, Darmstadt, Juelich, …)

Jefferson Lab (Kelvin Edwards)

- Some personnel changes
- Sun: upgrade to 8 almost completed; HP: all upgraded to 11i, moving away for central services; Linux: still at RH 7.2, evaluating Fedora 1. Windows 2000 domain upgrade finished, working on group policy issues

- 2 file servers recently upgraded to FAS940 (~16 k NFS operations/s), about 4.5 TB online disk space. Linux file server using 3Ware SATA system, 2 TB scratch area (16 160 GB Seagate SATA drives). Backups: QuickRestore, Seagate LTOs, Overland tape library
- Scientific computing: JASMine & Auger (Bath farm management and monitoring); typical: 2…4 TB of input data through the farm, 2000…5000 jobs per day. Certificates used for all user authentification. Tape drives moving to 9940B. Linux file servers: 16 data movers (all SCSI drives, using Mylex or Adaptec RAID controllers), 32 cache/work file servers (mixture of Mylex and 3Ware cards). Batch farm: 178 RH 7.2 Linux dual processors
- Noteworthy: Kswapd failures solved (automount timeouts set to 60 seconds, not minutes). Adaptec RAID cards: not quite as fast as Mylar, but acceptable, good manageability. Adaptec TCP Offload Engine: Problems with RedHat 7.2, custom kernel, and their driver
- Projects: Standard windows builds (server, IIS server, desktop, laptop). Backup migrating to Reliaty. Studying ssh v2 internally. Gigabit connection to border router now available. Looking at VLans for use on site. JASMine: rewriting disk cache, supporting farm output caches. PPDG: looking at Storage Resource Manager, replication, remote job submission
- Q&A: What is the performance of a RAID array with 3Ware and SATA? Have seen 60…80 MB/s

Nikhef (Paul Kuipers)

- Unix : Run 25 Suns (Solaris 2.8), 180 Linux (RH7.3/RH9), 190 Windows (everything…). phased out: HP, SGI, SuSE. Unix servers: storage: 2 Supermicro systems with Infortrend 2 TB (RAID 5). Worries about RedHat
- Mail: Now using SpamAssassin (have phased out njabl.org free blacklist). Virus checking with Sophos, also on outgoing mail. Squirrel Webmail much appreciated by users. Because of increased load, three new mail servers (3 dual Xeon 2.4 GHz, round robin via MX records)
- Windows: SUS server for upgrades of desktops; installation via Ghost image and/or RIS server); domain migration: hardware prepared, still planned for this year. Windows terminal server running on 2 dual Xeons; logging out after 2.5 hours idle because of a licence problem. The latter does not exist with RDP 4, but restricted to 8-bit colour with Windows 2000 server. Blaster, Welchia virus infections: Most infections were Windows 2000 SP1, upgraded to XP SP1 with automatic updates
- Network: some parts of the infrastructure in CC migrated to Gigabit Ethernet
- Grid: Development and application testbed for EDG (140 nodes, 5 TB), contributing to data challenges of D0 and LHC experiments. Will be involved in LCG and EGEE

Castor evolution (Jean-Damien Durand)

- Usage at CERN : 1.6 PB data in 12 million files, mostly on 9940A and 9940B tapes
- Limit of 9999 files per tape has gone
- Vision: clusters of 100s of disk and tape servers, hence facing the same problems as with CPU clusters; expect to make use of configuration and monitoring from ELFms; Castor project to provide storage resource management, resource sharing, request scheduling. Vision is resource sharing facility. Main components: Request handler using a thread pool, scheduling – request handler must be much more performant
- Today's situation: file names don't really hide the location of the files, name of disk servers is encoded in path name. New version will only allow referring to files by their Castor name. Will also use consistently owner and group issue of the staged files (no longer the need to propagate password file to disk and tape servers)
- Detailed plan available in proposal document http://cern.ch/castor/DOCUMENTATION/ARCHITECTURE/NEW. Milestones: Oct 2003: demonstrate concept of pluggable scheduler (LSF, Maui…) and high rate request handling; February 2004: Integrated prototype; April 2004: production system ready for deployment.

Progress monitoring: via Savannah portal provided by LCG (http://savannah.cern.ch/projects/castor). Castor team prepared to have progress reviewed, if experiments can put the effort
- Aims of new stager: pluggable framework for policy controlled file access scheduling; evolvable storage resource sharing facility framework. Basically on time

First experience with Windows Terminal Services at CERN (Alberto Pace)

- AKA remote desktop
- Requires Windows 2000 or 2003 server
- Following successful lxplus model
- Reduces need for VMWare, virtual PCs and windows emulators, dual boot etc.
- Pilot starting in June 2003, 3 standard computers 2.4 GHz, 1 GB RAM, 40 GB mirrored disks. Clients: Linux Redhat, Mac OS X, all recent Windows versions (IE 4 or higher, built-in client on XP)
- Options investigated, but dropped: platform-independent clients (HOBLink JWT Java), Citrix ICA
- All documented on http://cern.ch/wts
- Linux clients: rdesktop (www.rdesktop.org), tsclient
- Clipboard can be shared between remote and local desktops, local files and printers can be made accessible on remote session depending on client software
- Modifying the bandwidth depending on connection is possible
- Core set of applications installed to start with, constant requests for new applications – does not scale. Also, some stability and/or functionality issues
- Pilot had 220 users for various motivations (Linux/Mac users, users abroad requiring CERN environment, application not installed locally, want faster machine)
- Large majority of user representatives requested continuation of the service – but need to concentrate on standard set of core application. Standard node could be cloned for user groups that could add their own stuff, e.g. LHCb build server, AB/CO controls applications, … Core applications: Office, Acrobat, SSH, phonebook, Zephyr, Symantec Antivirus. To be discussed: Perl, OpenAFS, Visual Studio, Project 2002, …
- Cost: server licence, client licence for Mac and Solaris. 3 nodes should be enough to get started. 0.5…1 FTE is the current estimate for manpower
- Expect go/nogo decision in November 2003
- Q&A: How is the load sharing done? With the load-sharing built into Windows Server 2003
- Q&A: How many users do you expect to be able to support? Not really known
- Q&A: Is the pilot still open? Yes, but unattended.
- Q&A: Is the pilot accessible through the firewall? Yes. But hence subject to man-in-the-middle attack

New fabric management tools in production at CERN (Thorsten Kleinwort)

- Node is the manageable unit – node is autonomous (local configuration files, programs working locally, no external dependencies, no remote management scripts), adhere to Linux Standard Base (startup scripts in /etc/init.d/, logfiles in /var/log, configs in /etc, programs in /(s)bin, /usr/(s)bin)
- Several nodes can form a cluster (same functionality, not necessarily same hardware), mgmt tools enforce uniform setup. Critical servers replaced by service clusters with redundant nodes
- Principles: Software installs/updates via RPM, configuration via one single tool (currently SUE, but being replaced), configuration information through one single interface and stored centrally. Node installation and configuration reproducible

- Framework: Configuration manager, talking to config agent on the node; SW manager, talks to SW cache and SW agent on node; Monitoring manager(!) talking to monitoring agent; State manager; hardware manager
- Configuration (CDB/CCM): Based on EDG WP4 work. All relevant node information stored there. 1500 nodes in 15 clusters in there, ~3200 templates. Creates one XML profile per node that includes all information required to install and run the nodes. Currently two Linux versions supported: RH 7.3 and ES 2.1. Additional information: state information, monitoring information, vendor & contract information. Local cache now working in test mode, deployed on a few nodes. In addition to CDB, SQL interface for queries available
- Software: SPMA (Software Package Management Agent), again based on WP4 tools. Supports RPM and Solaris PKG. Based on RPMT (transactional RPM). Currently installing 700…1000 RPMs per machine. Runs on every node on demand; manages either subset or full RPM list. Package list created from CDB. Rollback is possible. Using a software repository; no scalability issue for current size of clusters at CERN (uses http as sw distribution protocol), allowing to pre-cache RPMs on the node)
- Node configuration manager: Will replace SUE as local configuration tool. Installed on some test nodes
- Monitoring: again based on WP4 development. Client samples ~100 metrics, deployed on > 1500 nodes, configuration (to be) put into CDB. OraMon server uses Oracle database as backend, stores current values as well as history, user API in test phase
- State, hardware management: HMS tracks installation, move and reinstall, retirement, node repairs; SMS allows to set node state, validating state transitions. Will handle new machine arrivals (November). Prototype working already
- Success stories: LSF upgrade from 4.2 to 5.1 on > 1000 nodes in 15 minutes, without stopping batch (with pre-caching); kernel upgrade (possible to disentangle installation and re-boot of node); security updates (ssh, KDE – 400 MB per node, the latter not even using pre-caching)
- Working all very well for CERN, even though some tools still in test phase. Most if not all of this should work outside CERN as well, but not readily packaged yet
- Q&A: Can the system feed back hardware information into CDB? No, that is against the logic; changes would be spotted by the monitoring system
- Q&A: How many files did you really need to create? Most of the machine-specific stuff is created automatically
- Q&A: Where would you see the minimum cluster size for which these tools make sense? Probably about 100 machines
- Q&A: What about node-specific configuration files? Well, that's what CDB/NCM are for…

Mail service at GSI (Karin Miers)

- Started from AIX server with sendmail – many security updates, complicated configuration. Hardware not powerful enough, no spam filtering, no central virus scan, no internal failover system, only external relay (TU Darmstadt)
- Hardware: two identical servers (2 2.4 GHz dual Xeons with double disks in mirrored configuration)
- 1600 mail boxes, 1350 personal mail addresses, 90 mailing lists. Mail data provided by oracle data base, alias tables are re-created automatically
- Although not explicitly stated during the talk (!), in conversation later with Alberto, it emerged that they use Exchange as their e-mail server.
- Using postfix 1.1.11-0.woody2 and Amavis as the filter program, Sophos and Clamav (both concurrently) as virus scanner, spam assassin as spam filter. Postfix found secure (each of the

modules runs with lowest possible privilege in chrooted environment), and easy to configure, well performant, robust. Amavis written in perl, assuring high reliability, portability
- Virus scanner signatures updated once per day. If virus is found, it is put in quarantine, and forwarded to recipient without the attachment. In order to flag virus, one positive from Sophos and Clamav is sufficient
- Spamassassin: perl based filter, resulting in rating. Some false positives unavoidable
- 1 week on one of the servers: 114'000 mails, 73000 incoming mails, 60000 delivered, 6000 outgoing mails (old AIX server still working for outgoing). 2480 viruses detected!
- Future plans: Spam detection using Bayes system (Spam assassin learning what is spam), external blacklists, accept only mails sent to official mail addresses, move listserv@AIX to mailman@linux?
- Q&A: How sure are you about the black lists? Not at all, perhaps will not do it at all
- Q&A: Why two virus scanners? Sometimes only one finds a virus
- Q&A: How is the load balancing done? Via DNS, this provides failover in case one machine goes bad automatically
- Q&A: Do you block any attachments in general with known file extensions? No, only if a virus is found. (At RAL, all *.exe … attachments are blocked!)
- Q&A: What do you do about viruses in the main body of the mail? It gets forwarded only once disinfected
- Q&A: Does the user receive all spam? Yes – the MTA adds "SPAM:" to the subject such that all MUA can react accordingly
- Q&A: What about false negatives and false positives? There is infrastructure in place in order to report them, is being used to teach spam assassin
- Q&A: Why do you use Oracle for this little bit of information? Matter of a policy decision – Oracle is around anyway

CVS status and tools (Sebastian Lopienski)

- Two services: CERN central CVS service, LCG service
- Central service: hosting CERN-related software projects, has gone through user requirement collection, architectural design, implementation based on assigned resources. Service in production since August 2003. 45 projects, 3 GB
- Secure and robust service with data integrity (daily archiving to tape, mirroring every hour), K4, ssh, pserver access, cvs lock monitoring and reporting, Web interfaces (CVSweb and ViewCVS) – but not a project management tool
- Architecture: server machines via DNS alias, load balancing via ISS, repository stored in AFS. Monitoring the service every 10 minutes. Service availability: well above 99.5%. Hardware problems did not affect the service level thanks to redundant architecture. Total down time in 2003: less than 12 hours, due to CC power cut and several short network interruptions
- Service provided by 4 servers currently, recently two new machines added (2x2.4 GHz Xeons). Newest server version (1.12.1) installed, patches applied whenever necessary. Using Quattor and Lemon for installation, configuration, monitoring
- Interactions with users: some 200 received so far (via Remedy or via E-mail). Full documentation on the Web (http://cern.ch/cvs) – user documentation, manuals, Howtos, list of CVS books, technical documentation for administrators
- Tools: maintaining AFS volumes, CVS lock detection (and notification to librarian), …
- Then LCG came along… and insisted on a cvs service not dependent on AFS! Architecture: N+1 cluster (N active server nodes with repositories on local file system, one slave server with a copy of all repositories on its local disk). Project to server mapping via DNS aliases. Pro: fastest possible access, independence of AFS, no special hardware (like High Availability stuff). Con:

Constant mirroring affects performance, load balancing at repository level only (not at request level), if slave server is down, all redundancy is lost, fail-over requires human intervention (for now, automatic fail-over planned for the future). Data replication done via CVSup, access to repositories and home directories on file-system level via NFS, instant DNS update. Web page: http://cern.ch/lcgcvs

- Q&A: Is the slave, once failover has happened, a read-only copy? No, because bringing the original node back on-line is a manually controlled operation
- Q&A: Why did LCG people refuse this?

Day 2

NERSC Site Report

LBNL manages the NERSC contract for DoE and PDSF is one of its many production systems. PDSF has new hardware – 96 dual Athlons and 18TB of storage, all Gbit-attached. They are evaluating Sun Grid Engine as a possible replacement of LSF, first results look good. Zambeel's cluster solution was being tested but the firm folded, has caused some disturbance. ALICE has joined the user community.

The NERSC IBM SP has been upgraded with 208 new nodes, 16–way Nighthawks, putting it into 5th or 6th in the top 500 Supercomputer list with a 10 Tflops/s peak capacity, 7.8 TB of memory and 44TB of GPFS disc space. GridFTP is up and running and gatekeeper is deployed on all production systems. Account management system is being integrated with grid certificates. They have worked on a web interface to HPSS, running HPSS 4.3.

CERN's Solaris Service Update

Sebastien Lopienski, a Fellow in PS/UI section, presented the current status of Solaris in CERN, the work porting EDG WP4 Quattor to Solaris and the status of SUNDEV. Quattor should eventually be used to manage all Solaris systems and this will be implemented gradually, starting with the forthcoming introduction of Solaris 9, replacing SUE and ASIS. He described the technology refresh of SUNDEV along with the installation of the SUN Blade server which was part of the SUNDEV purchase agreement.

APT for RPM – Simplified Package Management

APT (Advanced Packaging Tool) is a tool used by INFN Napoli for installing their Redhat systems. They had examined a great number of packaging tools from different sources but decided APT, originally from the Debian release, looked best. APT has been ported for many releases and is now available from sourceforge although some features found on Debian are still missing for other releases. The speaker described its features and advantages in great detail, see the slides for details.

PDSF Host Database Project

This is a project for inventory management and tracking. The database is based on MySQL with Perl scripts to gather and store the data and for querying the database, which includes purchase information as

well as technical details. The tracking part uses rudimentary tickets and creates event logs. There is also some simple interface to their Nagios scheme for cluster management. There are web and command line interfaces. It is heavily used for daily operations on PDSF.

CERN's Console Management Infrastructure

Helge Meinhard described the problem faced in installing huge numbers of nodes and managing them at the physical, level. The main requirements are remote console access, preferably to the BIOS level, and remote reset, including remote power cycle. He described various possible solutions. It was decided to deploy an infrastructure based on serial cards to permit remote access but leave the remote reset for now, depending instead on onsite operators and the fact that in large farms there is some level of redundancy. Specs for future purchases with require support for BIOS redirection to a serial line and more control of the power cycle, ie. to permit stay-off after a power cut. It was decided to install dedicated head nodes for the actual control functions, one head node per 48 worker nodes. On the software side, after hearing of a similar scheme being developed by Chuck Boeheim at SLAC at the last HEPiX meeting, they have been working with Chuck to see how to share the tool, adding some features which CERN needs but maintaining a common code base. He described the current status and the work in progress.

Debian at GSI

GSI seems to be virtually the only major HEP site which has chosen this distribution, although for laptops they recommend SuSE. He started with a list of Debian features and a very personal comparison of Debian with Redhat and SuSE releases. For the GSI farm, Debian was chosen because of its perceived stability and ease of upgrade but for laptops, hardware detection and self-user admin is easier with SuSE and for enterprise applications such as Oracle, SuSE is preferred since few major commercial vendors formally support their software on Debian, although some packages, eg Tivoli's TSM, actually run fine.

The GSI Linux farm has over 400 nodes. He described the farm configuration and how Debian is used to install, upgrade and manage these nodes. He offered a list of strong and weak points of Debian and closed by agreeing that it was not that Debian is better or worse than other releases but rather which made sense for individual sites. Statically-linked binaries inter-work between Debian and other releases without problems. Dynamically-linked binaries have the same library dependencies as between different versions of Redhat.

Web-Based File Systems and Webdav

Alberto Pace presented Webdav, a standardised extension to the HTTP protocol using XML for file access via the web. It has document locking features since it was initially aimed at distributed web developers. The presentation was highly interactive, showing live access to his CERN file base. He also showed the various clients available, how these access the server, whether by http or https and how to use this to access one's files at CERN via the WebDAV-DFS gateway, and including access to Exchange.

Delegating NIS Group Administration

Alf Wachsmann, one of SLAC's system admins, described a tool he uses to managing a large farm with a large number of users, a combination which creates a constant stream of group change requests. He has created a ypgroup command based round Rainer Toebbicke's arc command for client-server communication, written many years ago for AFS administrators. Front-end Perl scripts modelled after the AFS pts command communicate with the server where he has used Perl's database interface to manipulate the NIS files; this should allow the use of a full-blown database if desired. He has updated arc to use

Kerberos 5 and he believes that the scheme is modular enough to permit the use of SOAP or an RPC scheme instead of arc, if preferred.

Exchange Deployment and Spam Fighting at CERN

Alberto Pace described how the Exchange deployment is proceeding and also recent activities in spam fighting.

Spam Fighting at TRIUMF

All mail transits via a single server and mail to be delivered on-site, whether sent from inside or outside the lab, is filtered through an anti-virus and an anti-spam system before delivery. He explained the flow of the checks and the types of checks performed. They are looking at adding Bayesian learning techniques, trusted relay lists, SMTP authorisation and digital signatures.

Redhat Linux

Don Langley, Redhat sales manager covering California including SLAC, was invited to discuss the situation around Redhat. In the new model, enterprise subscription includes access to both binaries and sources, access to any upgrades during the subscription period and all maintenance fixes, remedial calls with guaranteed response times (different options). Enterprise release 3 is due for official release tomorrow, October 22nd. RHEL 4 is due about this time next year, based on the Linux 2.6 kernel. Enterprise releases will have a 12-18 month release cycle but each will have a guaranteed 5 year support life.

He explained why this new model had been introduced. The support of the frequently-released (every 6 months) open source releases was becoming a drain on resources and there is a desire for these releases so these will continue – take the open source and build it – this is the Fedora project being offered to "the community". But the Enterprise version will offer a service to sites needing more stability and willing to pay for it.

The Enterprise release has a single code base for all platforms, including Itanium. There will be several packages ranging from a low-end retail version, through a Work Station version (suitable for our batch and interactive services) up to an Advanced Server as needed on a service needing 24x7 support. Of interest to CERN is the mid-range ES version, for medium sized database services for example (we already 25 of these licences for the Oracle physics services). The basic differences between WS, ES and AS are support level and a few packages shipped or not (it is assumed that AS systems, will not need Open Office but WS probably will for example). Also WS and ES are only supported on 1-2 node systems; they may work on larger systems but are not formally supported. Supported commercial s/w includes Oracle, Cadence, Tivoli and others.

Fedora will not be shipped but available for download. It should be community-driven and community-supported. Redhat will not offer support for it. It should be a proving ground for new technology, the best of which may eventually make the Enterprise release.

Support is normally channelled via a named Technical Account Manager and is available with different service levels. There is access to the Redhat Network, a web-based management platform built for distributed systems to help in updates and various management tasks. Support does not need to be linked to the number of subscriptions. For example in our case we could take a small number of WS full support subscriptions and a very large number of WS-basic subscriptions which come with no support.

He then described the technical details of the Enterprise 3 release, see slides for details.

AFS Cross-Cell Authentication Using Kerberos 5

INFN runs multiple cells across the country and cross-cell authentication is frequently needed. But a Kerberos 5 implementation is felt to be necessary for security reasons (a Kerberos 4 security alert in March of this year). And anyway OpenAFS is moving in this direction and MIT Kerberos 5 provides support for AFS authentication. He described in some detail their particular setup based on Redhat 9 and OpenAFS 1.2.10. But they only started working on it in detail last week with first results the morning of his talk! In the future they hope to spread the Kerberos 5 cell across the whole of INFN to allow it to be used by all local AFS cells for cross-cell authentication.

ADC Tests

Jan Iven described various "research activities" going on inside IT/ADC group. The main motivations of such research were to make the most of what we'll get and what we have in view of the approaching LHC. He described the data challenge process and the benchmarks performed. They have developed a benchmark framework aiming at repeatable results built around an open source regression test framework from IBM.

He described the Openlab initiative and the Opencluster. He listed the criteria and candidates for the choice of processors, interconnects and storage. He highlighted the recent successful high performance network tests with both Infiniband and 10Gbit. He ended with some real results from recent data challenges, namely the (IT-internal) tape data challenge and the recent ALICE data challenge.

Day 3

Redhat Linux Support Policy

We held an interactive session on reactions to and consequences of the new Redhat support policy. Jan started with some background than SLAC, FNAL and CERN (Alan) described recent discussions with their respective Redhat sales representatives, without obviously declaring exact terms, conditions and prices. There was then a long open discussion. Since I was in the role of animator of the session, the notes below were kindly taken and contributed (again) by Helge.

Although some speakers advised to be careful and suggested not to forget alternatives like SuSE and/or Debian (which were considered realistic, albeit at a very significant migration cost), the vast majority of labs present preferred to stay with RedHat, following their Enterprise Linux (WS is enough for most, if not all, nodes HEP is concerned with – dual processor, not an extraordinary amount of memory), provided an acceptable solution with RedHat can be found. A solution that can be applied HEP-wide is vastly preferred, although some labs need to support non-HEP communities as well.

There was general consensus that most HEP labs have provided internally a high-level support for Linux and are prepared to continue doing so. Hence HEP will not require a very significant support from RedHat. Given this fact and the large number of machines HEP will need to run under Linux, the existing

offerings for RHEL (RedHat Enterprise Linux) were not considered appropriate; the general feeling was that the price points per subscription (one node, one year) are way too high to be acceptable. Also, site-wide or even HEP-wide subscriptions would be preferred over numerous subscriptions for individual nodes. (Several participants mentioned that Microsoft licences are providing significantly better conditions than what is currently proposed by RedHat.) If running a large outfit, HEP may also need to run a RedHat Network satellite for updates, which is expected to cause significant additional licensing or subscription cost.

The situation gets complicated further because we usually need to modify the kernel (e.g. in order to support the OpenAFS file system); strictly speaking, this already voids being eligible for RHEL support.

It was suggested that as major hardware vendors offer RHEL subscriptions with their products, we might contact them and ask whether they could offer subscriptions for existing machines.

We could consider re-building a distribution from the sources of RHEL (other users are reportedly doing this already, although some lab reps said they would not trust these builds from outside our community). These are public (they need to be due to GPL), hence this is legal – CERN have even received an oral statement from a RedHat representative. However, if HEP go this way, a written statement from RH should be sought.

Concluding a long, intense and broad discussion, it was agreed that major labs (CERN, FNAL, SLAC) should contact RedHat together, and try to negotiate some framework agreement for HEP which all interested other HEP labs could buy into. All labs potentially interested are invited to already now contact Alan Silverman, Chuck Boeheim, Jack Schmidt and Jan Iven, who jointly volunteered to establish the contacts with RedHat.

New HEPiX Scripts

Jan described a major update which is being implemented by a member of his section (Peter Kelemen) on these scripts which are used not only in CERN and DESY but several other sites since many years. A re-write seemed appropriate to remove unused multi-vendor UNIX support, to make less use of external tools from within the scripts, to aim at faster execution time and to produce a code base which is easier to understand and support. He invited all sites to check the new versions and give feedback. [Wolfgang Friebel got a copy some weeks ago and has already submitted some constructive comments.]

TiBS – AFS Backup at FNAL

Weekly AFS full backups were taking over 40 hours and required operators while the lab wanted to go lights-out. They looked at different packages supporting both IBM AFS and OpenAFS, ranging from those from IBM and Veritas to small specialised vendors. The notes contain an analysis of the various tools looked at. Finally chose TiBS. Currently backing up 1.6TB and backup is down to 5 hours. After the first full backup over the network subsequent full weekly backups are generated from previous full and incremental backups stored on the server using a large cache there and this is what saves the time. Daily incrementals are also merged in this cache and there is a one-pass restore scheme. The product comes from a small firm created by a student from Carnegie Mellon where AFS was born. Their experience of the product and the support has been good so far.

TRIUMF Computing Services

There has been no public cluster at TRIUMF since the demise of the VMS cluster in the mid 90s, only individual large nodes which became overloaded and/or obsolete quickly. Having studied their needs, and noting that the new Canadian Westgrid removed the need for a large batch service, they chose and installed a small IBM cluster based on 12 x330s with 1TB of SCSI-attached IDE RAID 5 discs, Redhat 7.3 release with the OpenMOSIX[1] kernel and OpenPBS with the Maui scheduler (as in Westgrid). The simplicity and scalability of this solution should assure a smooth upgrade/replacement path as and when this is needed.

TRIUMF benefits from large capacity access to the University WAN but at a price so they have built a home-made traffic shaper based on an IBM x305 Linux-based system. Other public domain tools are installed for monitoring, logging and flow metering. But this only works with outgoing traffic; to control traffic in both directions, you need to setup Class Based Queues on all interfaces and they have done this.

He then described TRIUMF's wireless network which almost everyone in the room was using. There are 13 hubs of 802.11b based on 2 models from Orinoco. But the demand for wireless is growing and this area is being reviewed. They have installed a Colubris wireless access controller to control access (we had to register or pre-register our MAC addresses); access from an unregistered interface will be redirected to a registration application web page. UBC (Uni British Columbia) as a whole has some 1200 access points across 150 buildings.

LCG-1 Status and Deployment

Ian Neilson described the current status and deployment of LCG-1. After briefly describing what LCG is and the structure of the Grid Deployment group, he listed the major components of LCG-1, expanding the many acronyms. He went on to outline the deployment process inside CERN and how a new site is added. An LCG distributed operations centre is being developed at RAL and a distributed user support model at FZK. He then compared the 2003 plans with what actually happened and why. He closed with some of the lessons learned and the next steps, especially getting the LHC experiments to make serious use of the project and preparing for the 2004 data challenges.

LCG Overview and Scaling Challenge

David Smith continued the LCG theme by covering the issues around scaling in particular. Two principle problems are access to a shared file system for job submission and the need for a separate job manager for each user job. Not everyone is happy with a shared file system between batch workers so a cache scheme has been devised to avoid the need for this, exporting files from the cache on the Gatekeeper to caches on the target batch workers. For the second problem, multiple job managers, since Condor-G is already used as the actual job submission service, a trick from the Condor team can be used to make the job manager exit after job start and use a central grid monitor task on the Gatekeeper to query the status of running jobs, restarting the individual job managers only after the user jobs have completed execution. Not all problems are solved, the system is still vulnerable to a large number of queries, but the ceiling in terms of numbers of jobs which can be catered for with the current Gatekeeper model, has been raised to the 1000s level.

Windows Server Hardware Management at DESY

Currently building their Windows domain on a number of HP servers and manage it with HP's in-built Insight Manager tool. All target nodes have a management agent installed plus a central server for console management. The features of this tool were listed, covering various aspects of managing the hardware and

[1] The features of MOSIX have been described in the past 2 HEPiX reports and will not be repeated here.

firmware of servers. Most of the talk consisted of examples of the various functions with screen shots of the results. He gave examples of where the tool gave early warning of problems before they became critical.

Windows and UNIX Interoperability

This was a talk from someone from Microsoft on tips and tricks to make Windows and UNIX work together. Unfortunately it started with marketing slides on .NET but quickly got better. He listed various tools to run UNIX code on Windows and Microsoft has just released a new version of its tool, Microsoft Services For Unix (SFU), currently costing $99 but negotiable, which sits above the Windows kernel. He listed its features, going into some details on calling Win32 from SFU programmes and *vice versa*. A number of common UNIX tools are missing but Microsoft has chartered a small firm (Interop Systems) to fill in such gaps (Python for example).

He also described a Microsoft Virtual Server to create a full UNIX environment on a Windows system which is due around the turn of the year. This is where a virtual server sits above a hardware abstraction layer and you can run one or more UNIX kernels and one or more Windows kernels side by side. There is clearly a performance price to pay but it is better than dual boot. He agreed this compares with Vmware but refused to speculate on the relative merits. Lastly he described how to download, as opposed to buy, a guide to migrating UNIX applications to Windows.

CERN Print Manager Abroad

Michel Jouvin of LAL/IN2P3 presented some work done on Ivan Deloose's Windows Print Manager (which he described in some detail) to make it available outside CERN. The main client task, PrntTray, has been modified to permit the selection of the site, for example CERN or LAL. Depending on which site has been selected, a different database of installed printers is selected and displayed by the Printer Wizard task itself. The code has been modified in such a way as to make it easily extensible to other sites on demand and he strongly encouraged other sites to make use of it.

Panasas's Object-Based Storage

Sub-title is Scalable Bandwidth for Clusters. A rather technical description of the product in a vendor-neutral way (?) presented by the CTO of the firm, also the man who invented the concept of RAID. From his academic background in CMU, he understands the problems of providing storage for bleeding edge HPC systems and the need for scalability in this realm. Also that capital costs are only a part of the total cost of ownership, perhaps only 20%. He outlined several solutions to the problem and their drawbacks. Ideally you should aim at shared managed storage with controlled scalable bandwidth.

He has developed an object-based architecture which moves the low level storage functions into the storage device itself. He described the access protocol in some detail, see the overheads.

Day 3 – Large System SIG Day
Theme - Security

Computer Security Update

Bob Cowles of SLAC started the Security sessions organised under the banner of the Large System SIG of HEPiX by reviewing recent security "events". He began by showing the effect on network traffic of the Slammer attack and how prompt remedial action had prevented a much worse problem. But it did show how little system patching goes on. As seen in CERN and elsewhere, he described how the infections arrived in SLAC (VPN, DHCP, etc) and the steps taken to "encourage" users to patch their systems.

Although most of this summer's well-publicised attacks were on Windows, Linux was not spared and he showed the (long) list of Redhat security advisories, followed by a similarly-impressive list of MacOS vulnerabilities, some inherited from the underlying UNIX base of MacOS X, and some for Solaris and Cisco software. Across all the platforms, one must remember that it is not only the kernel which must be made secure but also the applications above that, especially those applications performing some system-related function (sendmail, ssl, samba, etc).

Stanford had direct contact with Microsoft and had gaven feedback on the security patching process but the conclusions remain as before – poor administration is still the major problem, patches are essential and firewalls are not a replacement.

CERN's Reactions to Recent Attacks

Alberto described how CERN had reacted to the Blaster and Sobig.F worms. He explained in more detail how these worms operated and infected systems and the timeline concerning the announcement of the warning from Microsoft, the patch campaign launched across CERN, the first and subsequent infections and the steps taken to quarantine vulnerable and infected systems. This was the first time CERN has taken such action against vulnerable as opposed to already-infected systems.

CERN uses a variety of patch mechanisms – SMS, SUS and start-up scripts – because no one solution addresses all scenarios. A total of more than 18 FTE hours were spent by central services in the fight against the Blaster/Welchia worm. But as a result of this effort, the majority of CERN's Windows users running standard configurations, were unaffected by these attacks.

Opportunities for Collective Incident Response

Matt Crawford of FNAL outlined a proposition for a possible collaboration in fighting security attacks. The motivations for this are fairly obvious but who has the time and energy to contact others. By the nature of HEP, we have many users in common, even perhaps are users of each other's sites. However there are obstacles such as local conventions or techniques in place, different missions and sometimes just the "not invented here" syndrome. And there are costs – for information sharing, making tools or procedures more portable, the need for documentation. However, there are already signs of convergence, more use of non-local tools, etc. He then listed the 8 step Fermilab incident response procedure with 2 additional steps which (a) estimated the effect or implication of protective steps for other sites and (b) notified affected or other concerned parties.

Regarding actual projects, he suggested :-

- We must alert each other when attacks are in progress – should be an absolute minimum duty to each other
- Share data on background probes and attacks but only if we can make some use of such data (not obvious at this time)
- Produce and share a summary of sources of vulnerabilities; for example each person summarise the sources he or she uses to monitor security issues

- Do this via a dedicated security discussion mail list as opposed to the more restricted mail list used by the site security officers to share information about current or ongoing incidents
- Take steps to protect the portables of visitors and conference attendees by extending a site's controls to such systems – faces some technical challenges. For example we could screen systems before granting protected access
- Support the LCG security policies and procedures as they are defined

He outlined an imagined scenario of a Grid worm and the extent of the damage it could cause by using the Grid itself with stolen grid credentials. Finally he exposed the current Fermilab's internal security checking procedures. There was then a very interactive debate on how such sharing could happen – incident mail lists, discussion mail lists, shared post-mortems after incidents. The security experts present will presumably now discuss some practical steps [later scheduled for Friday].

LCG Security Update

Presented by David Kelsey. He showed the mandate and membership of the LCG Security Group. They have so far helped prepare some formal security policies and procedures for LCG and are working on more. Moving on to the security technology being used by LCG, there is X509 PKI authentication and LDAP is used to access a virtual organisation (VO), mapped through a Global grid-map file, for authorisation He then explained in some detail the current VO scheme, adopted from EDG 2.0, from registration to login to job submission. The Security Group is currently working on a Risk Analysis document.

Security Components on CERN Farm Nodes

Presented on behalf of Vlado Bahyl by Thorsten Kleinwort. One of the principle "components" is to keep all nodes up-to-date with respect to security and other patches. Extensive logging is done and the logs are scanned by filters for suspicious patterns of behaviour. Only secure access methods are permitted (ssh rather than telnet or rlogin for example). Accounting information is compressed and kept for up to 3 months for forensic analysis where that might be necessary.

Root Kit Detection Tools

In an impromptu session, Shane Canon of NERSC/PDSF presented some tools he uses to maintain security at his site. The main one is called St. Michael – found on sourceforge. It attempts to keep the kernel safe by means of checking critical memory regions, backing up kernel text and checking these for changes. It is not so well maintained which could be a drawback if a new root kit method appears but it does work. In operation, it cloaks itself so intruders will not easily detect it. In short it raises the bar against most currently-known root kit intrusions.

PKI Tutorial

Alberto Pace presented PKI in great depth. He started with an introduction to Cryptography together with some simple examples of encryption and the use of public and private keys. PKI or Public Key Infrastructure on the other hand uses certificates. These certificates provide a technology which permits practical distribution of public keys. He showed some trust models and examples of creating and verifying digital signatures. A certificate contains a person's public key and information about the entity being certified, all being digitally signed by a trusted authority. These were illustrated by showing how a certificate can be authenticated. Certificates are not particularly sensitive to protect; they cannot be modified without affecting the digital signature and they can only be used in conjunction with a private

key, which of course must be well protected. If a private key gets compromised, the certification must be revoked via a certificate revocation list. This also explains why all certificates have defined expiration date. The last part explained PKI deployment via certificate authorities and the various classes of certificates.

One slide contained a table of the strengths of different encryption methods and he gave some hints on how to choose a solution to a given problem, including possibly employing someone to try to break-in to the final implementation of the chosen solution. He ended with a short demonstration.

CERN's Computer Security Challenge

Denise Heagerty started by summarising recent incidents and trends in the realm of security. Although the total number of incidents in 2002 was lower than the previous year, this year the total has jumped dramatically even at only three-quarters through the year. And some of these have been more serious than seen on site before – Spam relays, large numbers of infected nodes and so on. She described the limitations on accessing into CERN via Internet, VPN or modem and DHCP for onsite portables. CERN, like Fermilab and other sites, performs regular security scans. Different tools are used for scanning for different vulnerabilities; scans should be as non-intrusive as possible. Firewall through-access is only permitted after a system has been fully scanned and the results assessed. And there is extensive intrusion detection, usually based on available software with local modifications.

New actions include hardware registration for portables by end of the year, off-site ftp closure next January and AFS password expiry with first tests starting soon. Network connection rules are being prepared. But the trend seems to be towards more serious attacks which spread faster. The effect of these is aggravated by the large number of poorly-secured systems, especially home or portable systems, subsequently used to access the site.

Security Update from KEK

This was a rare visit from KEK. Security measures there include

- The creation of a DMZ to which a number of sensitive systems have been moved
- Since over a year, all systems are forced to register MAC addresses and there are currently some 4600 registered
- There is a VPN service with 300 users, 2900 connections per month
- They make an antivirus tool available to home users.

Like all the other sites, they saw the familiar attacks and launched a patch campaign with the usual few recalcitrant system owners whose systems must eventually be blocked until they are patched.

Cluster Security at SLAC

Alf Wachsmann listed some tips to enhance cluster security. He proposes to make a lot of use of scripts for routine administrative tasks: this has the advantages of reducing the risk of human error, encourages more standardisation, more consistency and a more rigorous approach to the task in hand. And although the talk is aimed at clusters, a similar approach applies to desktop support. The first area is automatic system installation, post-installation tailoring and daily operational tasks. One tool he proposes for this is cfengine but of course there are many more. Avoid cluster-specific tools because such procedures can also be applied to servers and even desktops. Always have a light-weight emergency script which runs regularly (hourly?): normally it does nothing but if it detects some strange condition, it can call for an

action (from a patch to a re-install). [Some members of the audience were against this because of the risk of introducing a serious bug which would then be run everywhere.] Use a tool to perform patching; especially important to ensure consistency across a cluster; a candidate here is autorpm. On clusters, create a high priority, non-preemtable queue which can be used to perform patching without interfering with other jobs (on multi-CPU systems, the queue can be set to take exclusive control of the whole system when other running jobs end, although noone in the audience knew how to do this in LSF).

If starting from scratch put the cluster behind extra network security – a DMZ for example, or in a private non-routable subnet and run as few daemons and extra services as possible; if possible, try not to allow user login.

A Walk Through a Grid Security Incident

Dane Skow related a recent experience of a test case at Fermilab. A risk assessment of a grid incident must ask why such an attack should take place – self-gratification, mass-action, etc. But many attacks are automatic and could enter a grid by accident and then spread by the nature of the grid architecture itself. All the players in a grid have responsibilities – users to protect their private keys, CA operators to follow their formal rules, resource administrators to protect proxies, application software authors must accept to patch defective code, and everyone must perform due diligence on checks.

The talk continued with a list of the different vulnerabilities and the responsibilities of each player in each area. He laid particular emphasis to the protection of private keys. One danger they have noted is that it is very easy to leave private keys openly-readable in AFS home directories by not understanding how AFS security works. The exposure of these and the reluctance of the compromised users to get their certificates revoked, indicates that some form of forced grid-wide revocation is and will be necessary. It also raises the question – what constitutes a private key compromise? And there is going to have to be a great level of coordination between the resource providers associated with a given VO.

Day 5

The last day, actually morning, consisted of 3 parallel sessions. There was a Windows Birds of a Feather led by Michel Jouvin which presumably Alberto can report on. There was a discussion on how to further the security team collaboration ideas proposed by Matt Crawford which Denise can report on. And Olof organised a video-conference on mass storage with participation remotely from at least FNAL and RAL. The idea of this workshop last was to prepare a fuller programme of mass storage presentations as the theme of the next Large System SIG day at the Edinburgh HEPiX.

Mass Storage (as reported by Olof Barring)

Sites represented at the meeting included FNAL, RAL, Kalrsruhe and IN2P3 by video-link and several other people from various labs (including SLAC, CERN and JLAB) present the room. All sites represented supported the forming of a forum for Mass Storage Systems' interoperability. At least there were no statements against.

The following steps forward were taken:

- A mailing list has been setup (hep-forum-mss@cern.ch)
- Each site drafts a document summarizing their current capabilities and plans for what concerns

    - WAN networking interfaces
    - Security
    - Monitoring & monitoring protocols
    - File transfer protocol
    - Management protocols
    - Replica systems
- The mailing list will be used for coordinating the work
- A phone-conf or VRVS to be scheduled for the week of the 7th of December. Exact date to be agreed over the mailing list
- Don and Martin will act as editors collecting the material for a presentation on the different sites' capabilities for the next HEPiX
- The theme for next HEPiX in Edinburgh is "Storage". 1.5 days will be attached to the HEPiX for storage meetings and we should plan for a ~2 hours slot for the MSSF meeting.
- For the time being we use the mailing list and HEPiX for coordinating the MSSF. The newly created GGF HENP WG is an alternative. We can use the next HEPiX storage meetings as an occasion for study GGF as a summary group for going to a standards forum.

Windows BoF (as reported by Alberto Pace)

The Windows session discussed, yet another time, patch distribution on Windows (using SMS, SUS and group policies) and the SLAC password synchronization project between Unix/Windows.

The session was postponed to 11:00 am to allow people also participate in the security discussion and it lasted a little bit more than one hour only.

Security (as reported by Jan Iven)

This was a follow-up on the presentation on the Thursday by Matt Crawford on the possibilities of sharing. It was agreed to set up a mail discussion list. Technical solutions for automatically screening visitors' laptops were discussed and it was agreed that cooperation in this area would be useful.

Alan Silverman
28th October 2003