

AFS at CERN, Status, Plans, Benefits and Problems

EP Forum 18 June 2001
H.Renshall PDP/IT

Current Status

- Over the last two years the AFS service has been thoroughly modernised in servers, disks and software levels at a cost of about 600KCHF. A standard home directory and project space configuration was chosen (Sun E220R with two A1000 SCSI Raid arrays) and 11 are now in place. The top level volume location database servers were replaced by 3 dedicated Sun Netra servers and all file servers and clients run a modern release of AFS (not necessarily the latest).
- The scratch volumes were moved to a recuperated Sun E450 server equipped with cheap JBOD scsi disks. These are now at the end of their life and following price reductions will now be replaced by A1000 Raid arrays with larger disks but will still not be backed up.
- The staffing has been reduced to 1.5 FTE spread over 4 people with tape reloads outsourced.
- Reliability has increased and user problems have decreased.

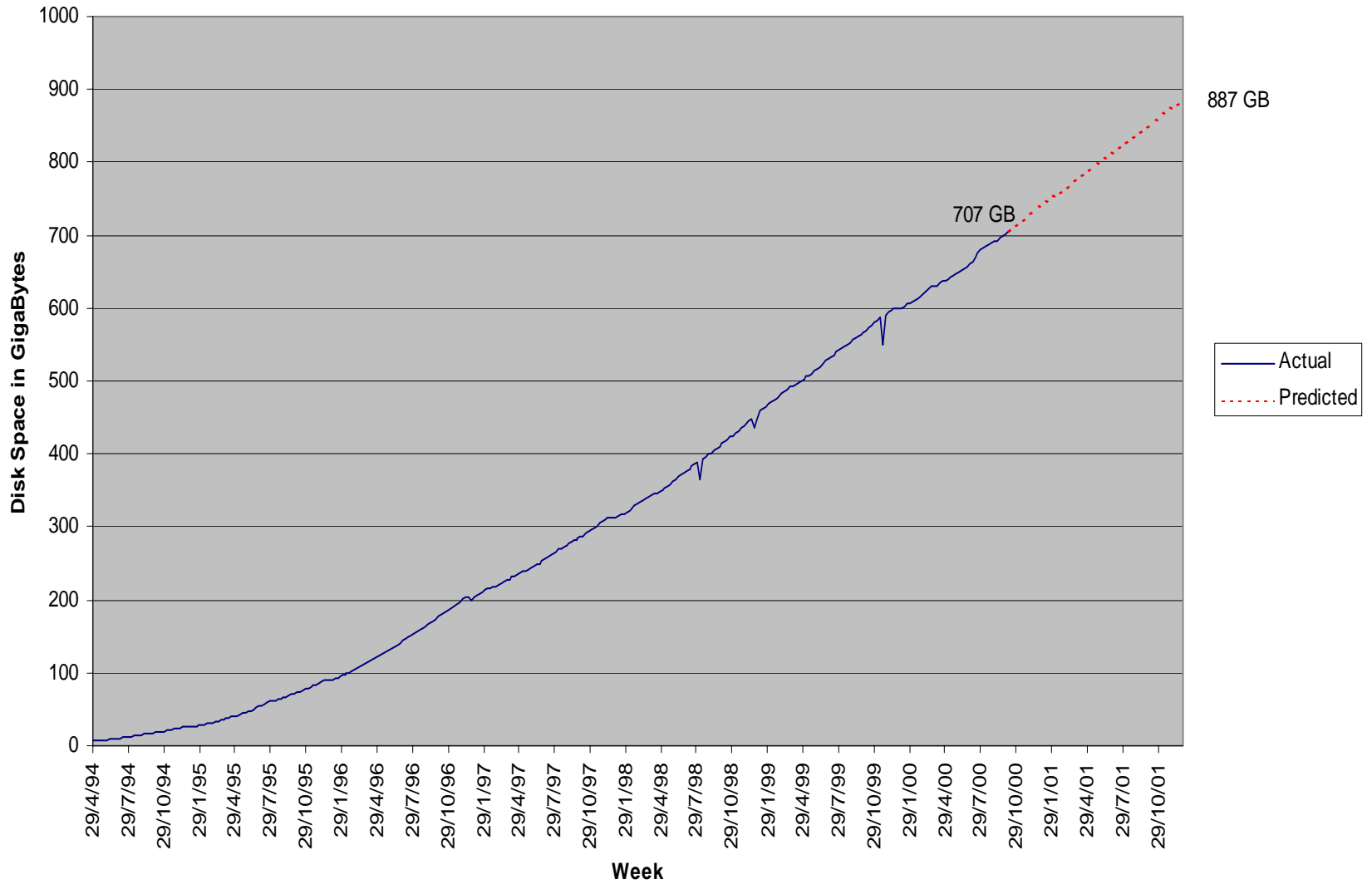
VLDB Server and a Standard File Server + 2 Raid boxes



AFS File, Volume Location data base and miscellaneous servers



Evolution of AFS Home Directory Usage since 1994 with Extrapolation to end-2001



See <http://consult.cern.ch/service/afs/stats/space.html> and <http://consult.cern.ch/service/afs/stats/users.html>

Project and scratch space mid 2001

- Project space mid 2001 1078GB used 1824GB allocated
 – (not including online backup copy)
- Scratch space mid 2001 368GB used 604GB allocated

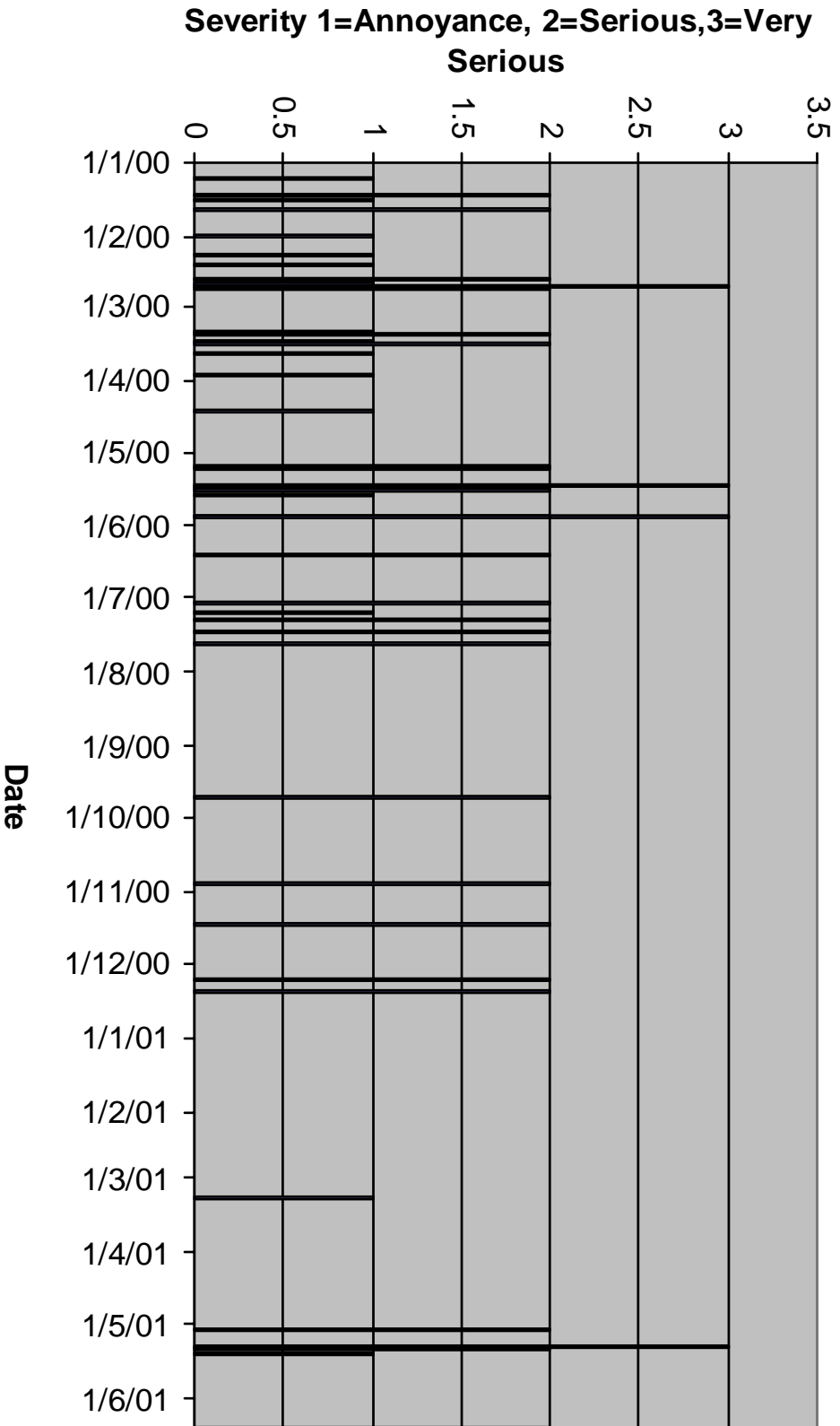
- Home dir space mid 2001 794GB used 1422GB allocated

- Project+scratch increase in 2000 333 GB used

- New project space requested 2001 430 GB
- New scratch space requested 2001 310 GB

- Total available is 4300GB (includes this years growth) with 2290GB used (53% of available) and 3850GB allocated (90% of available).

AFS Interrupts 2000/mid-2001 by Severity



Service Problems in last 6 Months

(see <http://consult.cern.ch/service/afs/afsproblems.html>)

- Wed 13 Jun 14.15 afs40 fileserver response very slow. No obvious cause so recycled it. This cleared problem and back in production 14.30
- Sun 13 May 10.00 afs48 shutdown to replace cpu. Back in production at 10.20
- Fri 11 May 13:55 afs48 crashed again with hardware memory error (Sun had said this should not happen) and partition c needed an fsck a reboot. Service was back at 15.20. 1700 users were affected.
- Thu 10 May 13.40 main AFS network switch power cable was loosened inadvertently while adding a new server. Nearly all afs services were lost until reconnection at 14.05.
- Thu 3 May 23.56 afs48 fileserver crash with hardware memory error. Automatic reboot and fileserver was back in production at 00:23. About 1700 user home directories were inaccessible during this 30 minutes.
- Fri 9 Mar 08.20 all fileserver processes were recycled before 09.00 causing logins to hang for a few minutes and many 'waiting for busy volume' messages but no failures. This was a precaution after heavy network pollution the previous day.

Last Weeks User Problems/Questions to AFS support

- UK university asks advice on starting an OpenAFS service
- User asks how to set same acl for all directories in a tree
- 3 Atlas users ask for scratch space replacement for their nfs scratch
- User directory inaccessible from one lxplus, ok on others
- Large scratch volume reports negative size to fs lq command
- Cern Atrans (write bypassing local cache) requested for Solaris 7
- AFS client requested for Solaris 8 by Cern certification team
- IT Controls Group requests new project creation
- LSF jobs get wrong AFS token in Solaris 7
- Scratch volume gone offline
- Two home directory reload requests

- An average week

Current Service Problems

- Global client afs hang. Seen most on busy linux machines (lxplus, lxbatch) with 1 machine per day out of 500. We are trying with new kernels+clients. Current fix – reboot client.
- Cheap Linux server better but still very sensitive to EIDE disk/bus problems. We will try internal Raid cards and use for personal scratch.
- Client process rapidly accessing hundreds of small files (eg build of Geant4) performance is poor and is likely to hang. Current fix – build locally. We will try different cache parameters.
- Client loses connection to a volume and either never recovers or takes many hours. User sees ‘file not found’ or ‘connection timed out’. Current fix – move volume to another server.
- Busy large (2GB and up) volumes go offline (1 per 2 weeks)
- Little or no knowledge of Cern extensions (arc, acron, batch tokens ...) after staff leave makes support for new environments difficult.
- We do not (yet) know the scalability limits of AFS.

Information from IBM Transarc Lab

(see <http://www.transarc.ibm.com>)

- IBM Transarc lab support AFS, DFS and DCE.
- Standard IBM support pricing using local support (IBM Suisse) as first level. Local support will be trained by IBM Transarc.
- Currently AFS 3.6 (supporting RH Linux kernels 2.2 and 2.4) and no plans for 3.7. Admit AFS does not generate a lot of revenue. Interested in Kerberos 5 (Cern would like this for Grid 'Globus' software) support but think this should be done by Open-AFS in which case may take it back into product.
- Solaris 8 and W2K clients released recently. W2K is MS tolerant only.
- Plan future NAS/SAN Enterprise-wide file system to replace AFS and DCE-2/3 years and driven by IBM San Jose.
- AFS development team same size but partly moved to India.
- Open AFS strategy announced. Source tree maintained by OpenAFS organisation.
- Official end of 3.6 support predicted for 2004. Will support new OS releases of existing architectures meanwhile and guarantee compatibility with OpenAFS.
- HEPIX concludes AFS is now in maintenance mode from IBM who will want their customers to either go to Open AFS or Enterprise system in a few years.

Status of Open AFS

(see <http://oss.software.ibm.com/developerworks/opensource/afs/>)

- IBM announced in August 2000 Linux world their open source contribution of the AFS Enterprise File system saying they will actively work with the open source community to extend AFS.
- Source was released end October at IBM site above after extensive code cleanup. Has been built at HEP sites e.g. Caspur (Rome) and works.
- An advisory council has been formed composed of::
 - Derrick Brashear Carnegie Mellon
 - Travis Broughton Intel
 - Craig Everhart IBM
 - Peter Honeyman Univ of Michigan
 - Ted McCabe MIT
 - Phil Moore Morgan Stanley Dean Witter
 - Bob Oesterlin IBM
 - Laura Stentz IBM
- First meeting was October 26
- CERN now runs the OpenAFS client on most central Linux machines.

Potential Work Items from Open AFS Council

(see bulletin board at openafs-info@openafs.org)

- Kerberos 5
- AFSDDB Resource Record Identification (using DNS)
- Disconnected Operations (done by UMich, but needs work)
- Win NT mount – so you don't need /machinename/afs
- Re-implementation Windows client (not the SMB trick)
- Mac OS 10 (now released)
- Backup clean up of Backup and BUTC (MIT, has some design changes could suggest design)
- Groups in Groups (done by UMich, but needs work)
- Large files/volumes
- Performance
- Backup Performance
- W2K
- W95/98 – (done by Shyh Wei Luan at IBM Almaden)
- Porting and Testing Efforts
- Incorporating changes across ports

Comparison with current User requirements (General)

- Transparent file access from any node on site:
 - yes from all unix, semi-transparent from Windows/NT and 2000
- Native access to files:
 - yes from all Unix, some windows applications
- Customisable protections
 - yes, at directory granularity. Multiple/overlapping groups supported.
- Authentication
 - yes but remote execution and batch need afs aware application
- Fast and reliable
 - Not fast: average speed 1-3 MB/sec but low latency and use of local cache probably satisfy the $\ll 1$ sec requirement for small files most of the time. Reliability now acceptable (only one total down of 35 minutes in last 12 months). Daily backup and instant restore from previous day, tape mount otherwise within a few hours during working day.
- Unique site-wide login script
 - hepix uses afs to implement this feature
- Source control and versioning
 - not built into the product

Comparison with current User requirements (Other)

- Home from Cern, Cern from Home
 - Yes if each site runs an AFS server and clients. User can klog into multiple cells at the same time and get full functionality. Currently 26 HEP cells worldwide and increasing.
- Maximum once/day authentication at remote site
 - afs token lifetime is 25 hours
- Installable at home labs
 - cheap (free but needs manpower) if OpenAFS is taken. Mirroring of directories would need a simple application to be written (as for ftp).
- Authentication and file-sharing software on laptop
 - yes, the afs client
- Automatic synchronisation of laptop and offline working
 - no. I know of no current plans for this desirable feature in OpenAFS. Could be simulated by running own cell on laptop and writing a synchronisation tool.
- Software development support
 - nothing built in. Must be managed by application layer and build performance is poor.
- Web access
 - nothing built in. Must be done by Web server application layer.

IT Analysis and strategy 2001 (Evolving)

- 1. Data sharing requirements
 - 1.1. Users need to share data between different nodes and platforms, both on the CERN LAN and remotely over WAN.
 - 1.2. The AFS distributed file system provides a single tool with maximum functionality that satisfies essentially all these needs.
 - 1.3. There is (was) no recent detailed description and analysis of the requirements. The users simply state they need AFS.
 - 1.4. AFS was expected to evolve to DFS in the framework of the Open Software Foundation. This evolution never happened.
 - 1.5. AFS is only available from one vendor. (Can now use OpenAFS)
 - 1.6. Recent announcements by the vendor and its parent company cast substantial doubt on the future of AFS as a viable commercial product. (IBM since extended support for 2 more years till 2004)

IT strategy (continued)

- 1.7. AFS is a very complex product that needs substantial effort to operate (1.5 FTE for 14000 users ?). Because AFS is a niche-market product, this effort must be provided by staff rather than outsourcing.
- 1.8. AFS is not an ideal solution for high performance applications with high data volumes, such as those in massive farms for physics computing.
- 1.9. Alternative tools based on client-server, such as the *hsm* command, have recently been introduced to handle high volume applications such as scratch and physics data. These tools will very likely be able to satisfy the “Large File System” requirement from the physics and engineering communities.
- 1.10. A lot of developments have taken place since AFS was chosen, in particular the Web and Linux and the consolidation of the Network Attached Storage market.
- More recently the extension of support has taken the pressure off and we will continue with AFS, including following OpenAFS, for now.

IT Strategy (last)

- Recommendations

A three person team, composed of representatives from the physics user community, PDP and IS groups should produce a concise document consolidating the existing user requirements for Data Sharing by April 2001. The already existing input from the engineering community should be taken into account. (This was the first item at this meeting).

These requirements should be matched to possible solutions, and implementation and long-term maintenance plans should be presented to after-C5 and the GLM by October 2001. Emphasis should be put on modern solutions based on widely accepted industrial, open-source or de-facto standards. It is recognized that multiple products may be needed to meet all the requirements (see the next two presentations. This meeting is input to this process).