Journal of Physics: Conference Series **513** (2014) 042046 doi:10.1088/1742-6596/513/4/042046

Toward a petabyte-scale AFS service at CERN

Daniel van der Ster, Jakub T. Mościcki, Arne Wiebalck
CERN, Geneva, Switzerland
E-mail: daniel.vanderster@cern.ch

Abstract. AFS is a mature and reliable storage service at CERN, having worked for more than 20 years as the provider of Unix home directories and project areas. Recently, the AFS service has grown at unprecedented rates (200% in the past year); this growth was unlocked thanks to innovations in both the hardware and software components of our file servers.

This work presents how AFS is used at CERN and how the service offering is evolving with the increasing storage needs of its local and remote user communities. In particular, we demonstrate the usage patterns for home directories, workspaces and project spaces, as well as show the daily work which is required to rebalance data and maintaining stability and performance. Finally, we highlight some recent changes and optimisations made to the AFS Service, thereby revealing how AFS can possibly operate at all while being subjected to frequent—almost DDOS-like—attacks from its users.

1. Introduction

OpenAFS [1] is a mature network filesystem with more than 20 years of history at CERN. AFS offers universal access to users' files from all relevant platforms—including desktop PCs (Windows, Mac OS, Linux), interactive workstations, and batch nodes—as well as globally across the Internet. Figure 1 shows the growth of AFS at CERN since it was first deployed in 1994. Since then the service has grown to serve more than 30.000 active users at CERN, and is relied upon for many core data use-cases, including home directories for personal development work, workspaces for large scale batch processing, and project spaces for experiment data and applications.

We present the status of AFS at CERN as of September 2013. We detail the usage of AFS and present some recent improvements to the service which are enabling its continued growth. Finally, we conclude and present some ideas for the future.

2. AFS Usage at CERN

During the past year, AFS has grown by around 200% in volume. Figure 2 shows the nearly linear growth since January 2012.

Table 1 presents the status of all AFS volumes at CERN as of 18 September 2013. With close to 250TB of stored data, AFS may be considered small relative to the data storage services CASTOR and EOS [2]; however, the total number of files ($\sim$ 2 billion) and access rates ($\sim$ 75kHz) are at least an order of magnitude larger than those other services. Regarding the breakdown of volume types, the majority of data volume is used by workspaces, which are offered to users as 100GB volumes which they can use for batch processing (which is confirmed by the relatively large file size in the workspace volumes). Project volumes are numerous and

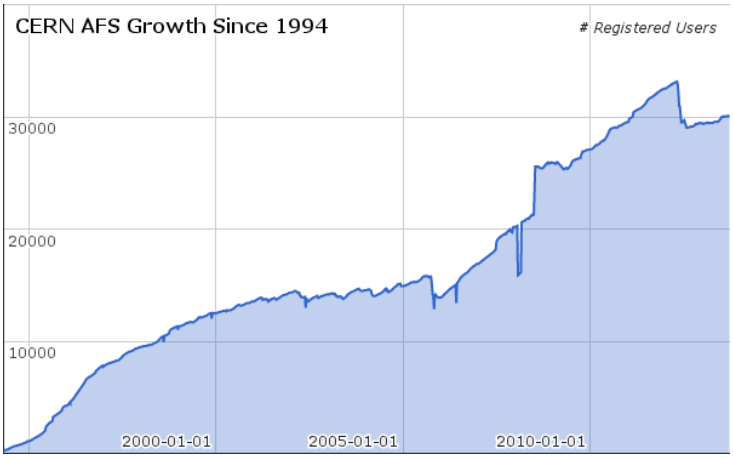

Figure 1. Growth of the AFS service at CERN since 1994. Large drops are the result of user archiving campaigns.

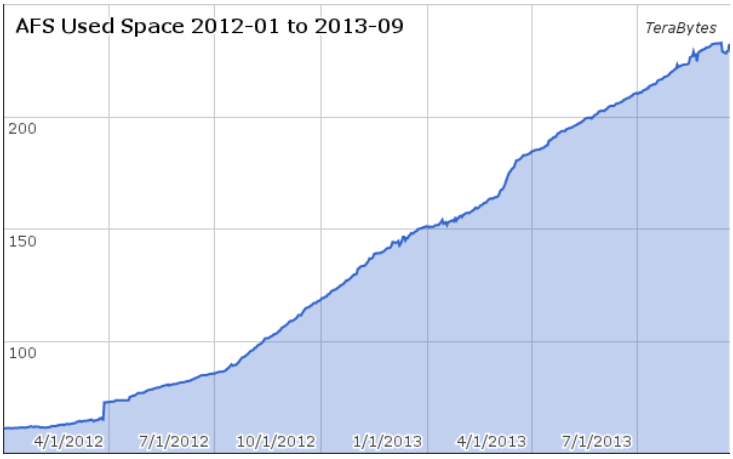

Figure 2. AFS used space since January 2012

Table 1. Status of the AFS Volumes at CERN, 18 September 2013

Type	# Volumes	Quota (TiB)	Used (TiB)	# Files ($*10^6$)	Daily Accesses ($*10^6$)
Home Directories	30156	47	14	199	832
Workspaces	4627	303	113	402	2504
Project Volumes	33773	182	99	1282	3108
Archived Homes	14872	3.6	0.88	21	0
TOTAL	83428	538	228	1904	6445

large – experiments rely on these for software and data delivery. Finally the home directories are relatively small due to the per-volume limit of 10GB – however the access rate of 4 IOs per file per day shows that these files are more active than the project volumes.

Table 2 presents the breakdown of local and remote clients in 2012 and 2013. The number

Table 2. Local and Remote Accesses to /afs/cern.ch/

Date	CERN Clients	Remote Clients	TOTAL
2012	16823	19008	35830
2013 (–Sept)	16344	16979	33322
Week of 19 Sept 2013	9673	4238	13910

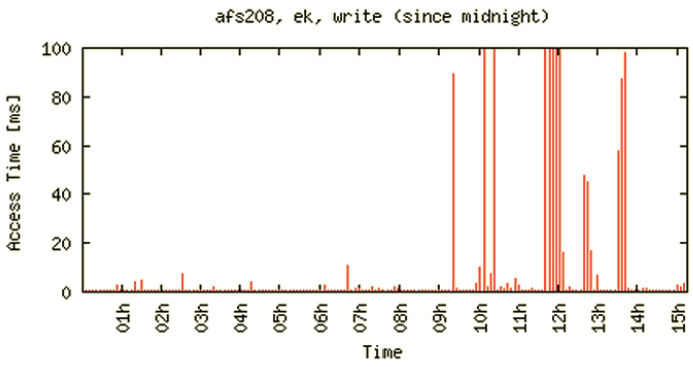

Figure 3. Example of a workspace server with an unusually high latency.

of total CERN and remote clients have been divided at close to 50/50 over the past two years, however for the current week at the time of writing, the more usual ratio of 70/30 local/remote clients was observed.

3. AFS Optimization

During 2012 and 2013, a number of improvements have been made to the AFS service at CERN, enabling more cost-effective servers to be deployed and thereby allowing further growth of the service.

3.1. Access Latencies

Because AFS is used most often for interactive work, users are very sensitive to the latency of the service. As the usage of workspace volumes increased throughout 2012, a number of high latency incidents were observed on a variety of disk servers, each of which causing a user-visible disruption. An example latency incident is shown in figure 3, where a server saw its latency to write and read 64 kilobytes increase from the usual 1-2 milliseconds to more than 100ms.

Early debugging work indicated that the latency events were being caused by unusually large traffic to one or two volumes on a file server; however, observations of the disks and CPUs did not reveal a culprit. In order to debug and solve this problem we undertook two parallel investigations: first, we sought to reproduce the problem by configuring a test server and using the LSF batch system to stress the server at load until the latency events were observed. In this test environment we found that 30 parallel clients were able to increase the latency dramatically.

Next, we used the *rxperf* tool (from the standard openafs package) to try to understand and optimise the underlying UDP-based protocol used by AFS, namely RX. With rxperf, we found that the default configuration led to an increased latency with only 5 clients.

After observing different limits for the two tests (30 vs. 5 clients), we looked for differences in the test configurations. Finally we noticed that the fileserver had a UDP buffer size of 2

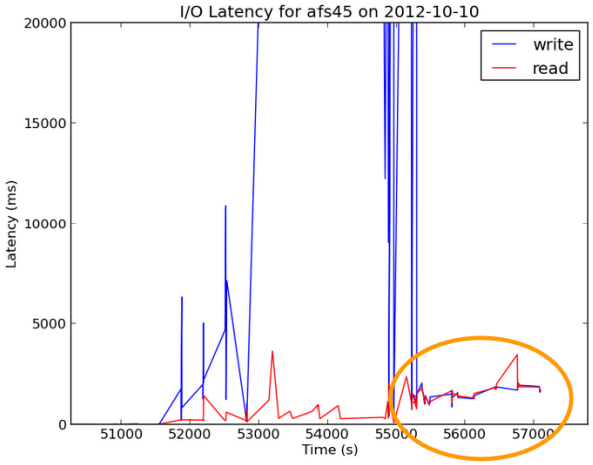

Figure 4. Applying the UDP buffer size fix in production on a highly loaded fileserver.

megabytes, whereas the default rxperf configuration uses 64 kilobytes. After increasing rxperf's UDP buffer to 2 megabytes, the performance became similar to the fileserver, which confirming that this variable was the key factor in the latency issue.

Further testing confirmed that a 16 megabyte UDP buffer would allow a single fileserver to host > 200 clients while maintaining an access latency of less than 300 milliseconds. This result convinced us to roll out the increased buffer size to all production fileservers.

Figure 4 shows the effect of the fix as it was applied to an in-production server which had an ongoing latency incident. At its peak, we observed > 300s latency on this server, however after increasing the buffer size the latency dropped to around 1s. Since deploying this fix in autumn 2012, the rate of user-reports related to AFS latency has dropped from around five per month to approximately zero.

The full details of the problem and solution to this latency problem is available online [3].

3.2. Other improvements

A variety of other improvements have enabled further growth of the AFS service at CERN:

Revisiting the Backup System The CERN AFS service level agreement guarantees users that all files stored in */afs/cern.ch* will be backed up daily and kept on tape for at least six months. As the total volume of data in AFS grew past 100TB, the scaling ability of CERN's AFS backup system (written more than 10 years ago) was being stressed. In order to better cope with this growth, the backup system was re-written to support increased archival backends, including the use of multiple TSM servers as well as backup to CASTOR.

Volume and Pool Management AFS supports transparent migration of volumes between servers; this feature is used frequently at CERN in order to balance load and disk usage across the fileservers. The tool which implements this automatic volume balancing—the pool monitor—was similarly unable to scale with the growth of AFS at CERN. A rewrite of that tool with increased parallelism now allows up to 250 volumes to be moved per hour.

Using Large Disk Servers Since the early days of AFS at CERN, the service has demanded specialised storage hardware with low-latency RAID arrays and redundant controllers. Recently,

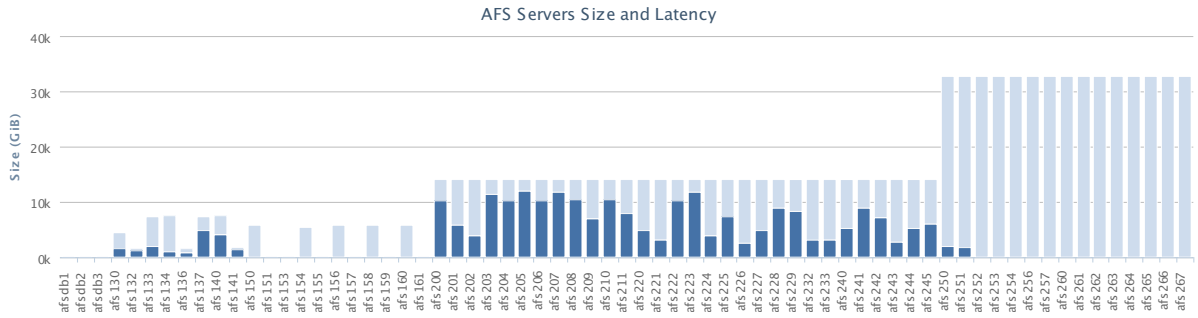

Figure 5. Rollout of new large servers resulted in a 1PB available space in the AFS service.

the operational and management overhead incurred by these special requirements has encouraged the team to investigate the usage of standard large capacity disk servers in future. These newest generation of these servers feature twenty four 3TB drives in JBOD configuration, without redundant storage controllers (and configured in software RAID1). Whereas in the past our hardware configuration could offer high availability (HA) even in the case of a server failure, the lack of redundant paths to the disks on the new servers precludes this level of HA. We have investigated techniques to rapidly recover from server failures (for example with standby virtual machines and ZFS) however these techniques were not yet found to be stable enough for production. As such, the new servers have been put into production without HA, relying on the statistics from the past that total server failures are highly infrequent (less than once per year).

4. Conclusions and Future Plans

With the aforementioned improvements in place, the AFS service has grown by more than 200% in the past year and used space is expected to continue to double yearly. Figure 5 shows the total available space in AFS at CERN – with the newly deployed 32TB servers we have achieved a total 1PB capacity.

In the near future we plan some significant updates which will further improve the service. First, by deploying OpenAFS 1.6 fileservers, we expect the performance and capacity to improve. Next, we plan to puppetize the service, and we expect that the efficiency benefits offered by puppet will allow us to increase the number of fileservers we operate. Finally, we hope to exploit the new cloud offerings at CERN [4] – virtual AFS fileservers with Ceph-backed storage volumes [5] a highly attractive solution to the operations of AFS at CERN.

References

[1] OpenAFS website, online at *http://www.openafs.org/*.
[2] X. Espinal et al. Disk storage at CERN: handling LHC data and beyond. *These proceedings.*
[3] J. Moscicki, D. van der Ster. Site Report - CERN. *European AFS and Kerberos Conference 2012.* Edinburgh. Online at *http://conferences.inf.ed.ac.uk/eakc2012/*.
[4] B. Moreira et al. Production Large Scale Cloud Infrastructure Experiences at CERN. *These proceedings.*
[5] D. van der Ster et al. Building an organic block storage service at CERN with Ceph. *These proceedings.*