20 Years of AFS at CERN

Rainer Többicke, CERN/IT

HEPiX 2011, Vancouver

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Reminder - what is AFS

- A planetary File System
  - Client disc caching
  - Kerberos Security
  - Posix semantics
  - Global Name Space
  - Location Transparency
  - Support for lots of Unix and Windows Systems

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

1990s – LEP Times

CERN IT Department

- CERNVM
  - 1000 users including 200 batch jobs, on 6-way CPU
- VAX VMS cluster
- DECstations, SunOS 4.1 MIPS-based workstations
- Apollo cluster: Motorola 680x0 workstations
  - 30/40 MHz processors, 16 MB memory
- 10 Mbits/s Ethernet – coaxial “cheapernet” bus

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

AFS Beginnings

CERN IT Department

- 1991 Tests on HP, IBM and Sun hardware
- After a few false starts
  - Oldest volume: Wed May 20 13:24:17 1992
  - 2 x 300 MB file systems
  - 2 servers: HP & AIX
  - Jamie Shiers donated a 1GB disc
- HEPiX: rtb preaching AFS
  - NIKHEF(93), Saclay(94), Prague(95), Zeuthen(97), JLab(97)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Paleolithic gems

CERN IT Department

- ASIS – public domain software repository
  - Mon Jul 27 14:40:30 1992
- Cernlib
- Users
  - Jamie Shiers - Fri Sep 4 10:21:44 1992
  - Chris Onions - Wed Jan 13 15:40:44 1993
  - Helge Meinhard - Fri Nov 12 11:05:18 1993

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

1993 – Workgroup Services

CERN IT Department

- The first loosely coupled Unix cluster, files shared over AFS
- /afs/cern.ch/user/*initial*/username
  - Limit number of entries per directory
  - Scheme cast in a meeting, Tony Cass counting r, s, t peaks from phone book
- /afs/cern.ch/group/xxxx – group environment
- Guinea Pig – the CHORUS experiment

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Getting used to it

CERN IT Department

- Kerberos Authentication
  - Login hooks
  - Training users that tickets expire
  - Explain why “root” has no special powers
  - Batch job authentication – token renewal
  - Cron jobs – acron service → HEPiX '95 Prague
- Patience-stretching stability
  - Excitement when VM mainframe pushed out

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Early Performance Challenges

- Single threaded file server
  - During disc I/O, RPC protocol stack stops
    - UDP input buffer overflow, retries, short lock-ups
- Home-made fileserver running all I/Os in POSIX threads
- File server RPC call starving
  - A 100 MB file read in by hundreds of NA45 jobs
  - Other volumes on the server timing out
  - Had to stop batch for 2 hours to move volume

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Mid-term – AFS challenged

CERN IT Department

- DCE/DFS, the constantly delayed AFS successor, never really made it
- AFS's reputation: high maintenance, expensive hardware, unpopular OSes (AIX, Solaris)
- Venue of Windows 2000 created high expectations
- Stony path to Linux servers
  - Immature OS, immature AFS port

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Struggling for Survival

CERN IT Department

- Open investigation to abandon AFS
  - Closely involved the Physics community, lead by Marco Cattaneo (LHCb)
- Conclusion: no alternative, AFS is mission critical!
- Consequent investment in high quality hardware and manpower
- Increased capacity

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Firewalls

- UDP-based RX protocol
  - Firewalls do not understand “session” semantics
  - Firewall administrators...
    - Treat it like DNS. Well...
    - Assume a request-response protocol and allow for responses in a time window, hereby wrecking callbacks
  - NAT
    - Routers do not do any better

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Disc Overload – Thread Starving

CERN IT Department

- Highly popular AFS volumes
  - Access rates in the 2000-3000 per second
  - Disc I/O's pile up, thread starving sets in
    - Other volumes affected, even on other discs
- Isolation in a queue with fixed, dedicated threads
  - Fixed number limits load on discs
  - Request starts normally
  - Heuristics to determine “overloaded” volumes
  - Request re-queued

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Callbacks

- The client as a server: cache consistency
- Synchronous close()
  - Dirty data flushed to file server
  - Clients notified - N at a time (N ~ 10, ~10s timeout)
  - During this time the flush waits to return
- Now apply this to a file cached by 10000 clients, some of which block callbacks or idle behind a NAT
- … and to a volume with 500000 files...

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Callbacks - fixed

CERN IT Department

- Fixed callback collection
  - Collect all in one go, memory is cheap, avoids loops when callbacks are re-established while broken
- Revisited the client probing process
  - Now highly parallel, timeouts do no pile up
  - For o(10000) partly ill-behaved clients, down from 10-12 hours (!!) to < 2-3 minutes
  - Result: “dead/deaf” clients discovered early
  - “deaf” clients can be nagged

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

AFS Service Today

CERN IT Department

- Found it's niche
  - Low volume: 100-TB-range
  - High access rate:
    - 4-6 billion accesses/day → 70 kHz
    - 0.5 billion disc I/Os → 7 kHz
  - Low latency
- Manages itself 80% automatically
  - Effort spent elsewhere
  - Hardware investigations, SSD, Backup policy, Service integration

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

AFS Over the Years

CERN IT Department

- Sound design has set a reference point
  - Adapted to technological evolution from 1990 to now
    - Processors: 100 times faster, 1000 times more memory
    - Discs: 3000 times bigger, only ~5 times faster
    - Network:100 times faster
- Constantly seriously challenged
  - NFS, Windows DFS, GFS, IBM Storage Tank, GPFS, Lustre
  - NFS version 4.1...

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Outlook

- AFS maintains its place as an infrastructure file system
  - Small size, high access rate, low latency
  - Challenged again... (e.g. CVMFS)
- Orthogonal to Cloud and Mass Storage
- Appliance-style “YFS”

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Retrospectively

- AFS design defined state-of-the-art
- The implementation didn't
- Touches on a very rich set of different aspects and problems in data management
- Fun to work with
- HAPPY BIRTHDAY
- … and ...
- … thank you for listening!

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it