AFS at CERN: Growing with the users' needs

Arne Wiebalck
(for the CERN AFS team)

HEPiX 2013 Spring Meeting
Bologna, 16 April 2013

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

"If the AFS service cannot grow, that will be its end."

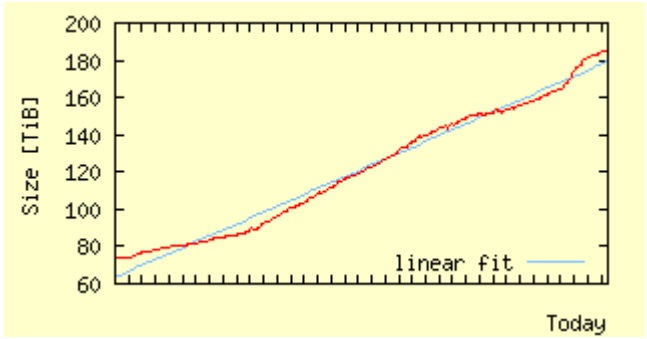
(Name withheld, 2011)

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

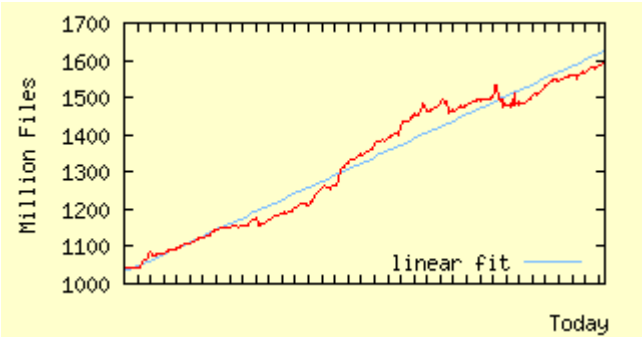
Introduction

- 180TB of data
  - growth from 50% to almost 300%
- 1.6 billion files
- 15k clients
- ~50k accesses/sec
- 74k volumes
- ~50 servers w/ 450 partitions
  - going down slightly

AFS used space in past 12 months



AFS #files in past 12 months



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

User-side: More clients

CERN IT Department

- Extend our client base
  - Only about 120 Windows based OpenAFS clients at CERN
- OpenAFS for Windows now installable via CMF
  - Installing OpenAFS on Windows is not straight-forward
  - collaboration with OIS colleagues, Win 7 (64bit)
  - even the Windows colleagues use it now :-)

Windows 7

- HOWTOs for Mac OS/X and Mobile Devices
  - Installing OpenAFS on Mac OS/X is easy
  - Native (commercial) client for iPad, iPhone, and iPod Touch
  - SFTP clients for Android (native client in preparation)

ANDROID

- Details available from cern.ch/afs !

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

CERN

User-side: More space

- Personal work spaces
  - Up to 100GB per user
  - SSD-enhanced servers
  - /afs/cern.ch/work/d/dwight
- Home directories
  - Up to 10GB
  - Critical power

User acceptance

3'800 work spaces created since service opening

work spaces occupy now more than 30% of the total space in AFS

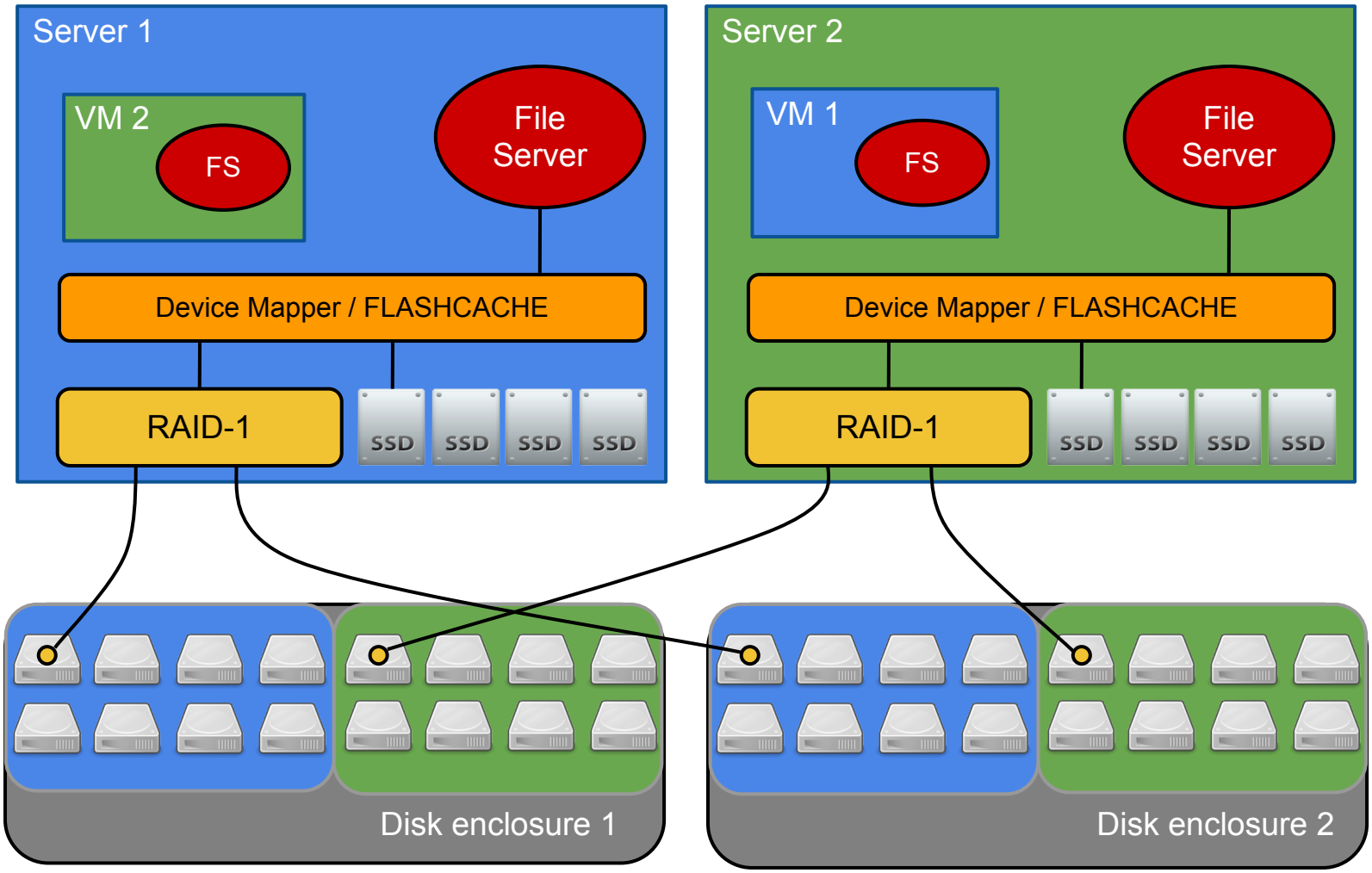
user feedback very positive

- Reminder: All space is backed up
  - Retention is 6 months
- Get more AFS space from cern.ch/account !

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

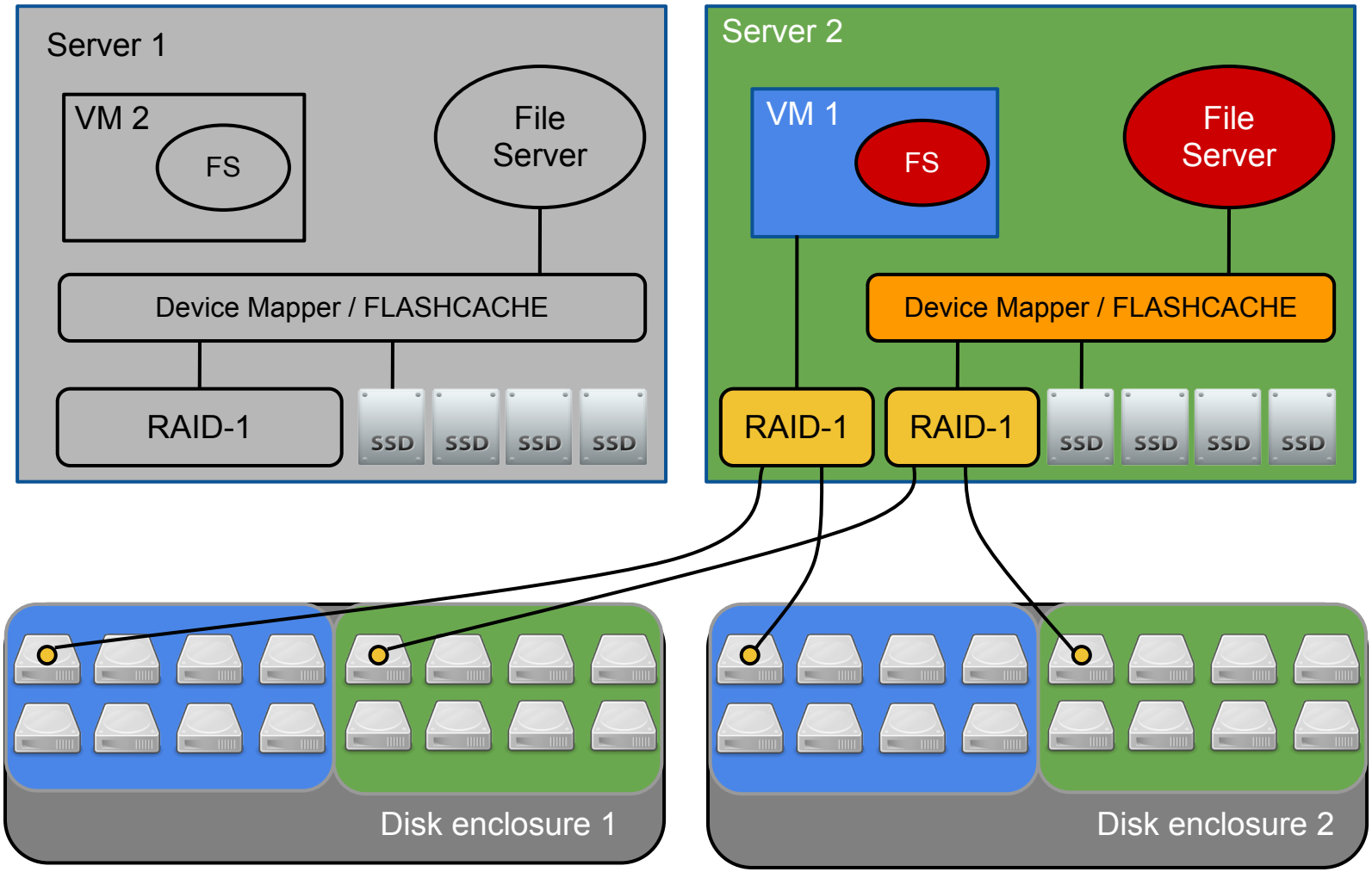
Server-side: New architecture

CERN IT Department



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

Server-side: New architecture



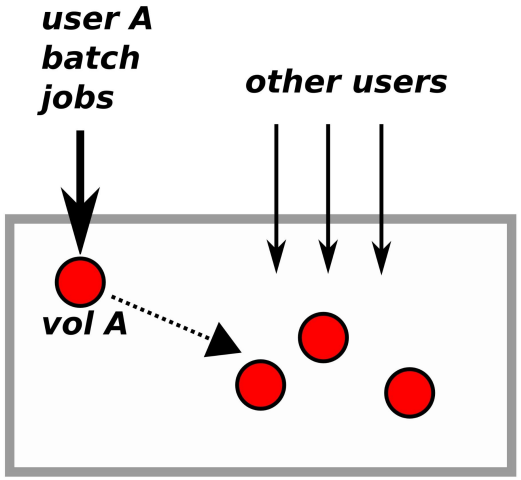
- VMs for volume separation
- Time to flip: 120secs (< client timeout)

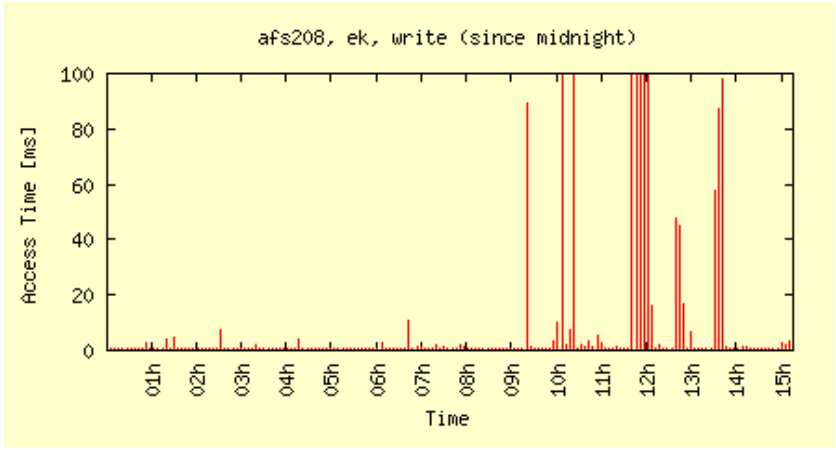
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Server-side: Latency issues (1)

- Users (usually) do not complain about throughput
  - but they do complain about **access latency** to read files or lists directories at interactive prompt
  - "ls of death" is an extreme example

one user / one volume impacts
all others on the same server

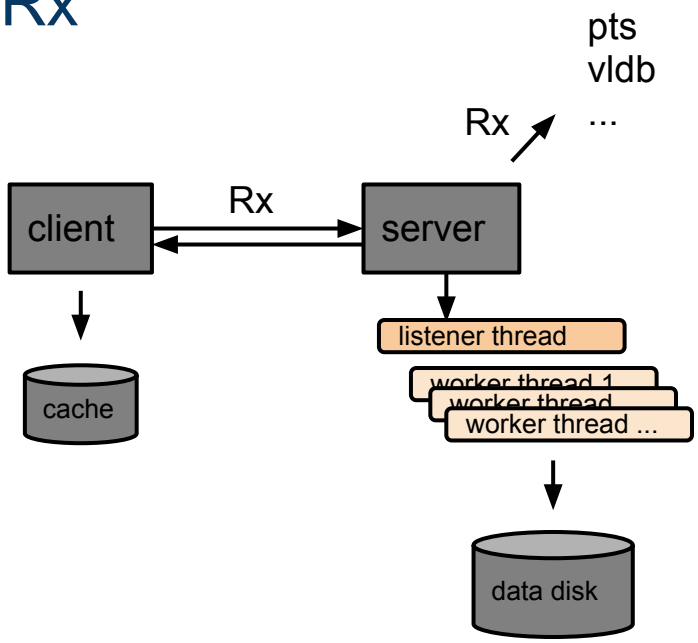




CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Server-side: Latency issues (2)

- Hardware limits excluded
  - Disks idle, network below peak values
  - CPU flat at 125% on a 4 core system
- AFS basics: file server and Rx
  - RPC over UDP
  - 1 dispatcher, 240 workers
- Two symptoms
  - Thread starvation
  - "Rx-limit"



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

Server-side: "Rx-limit" (1)

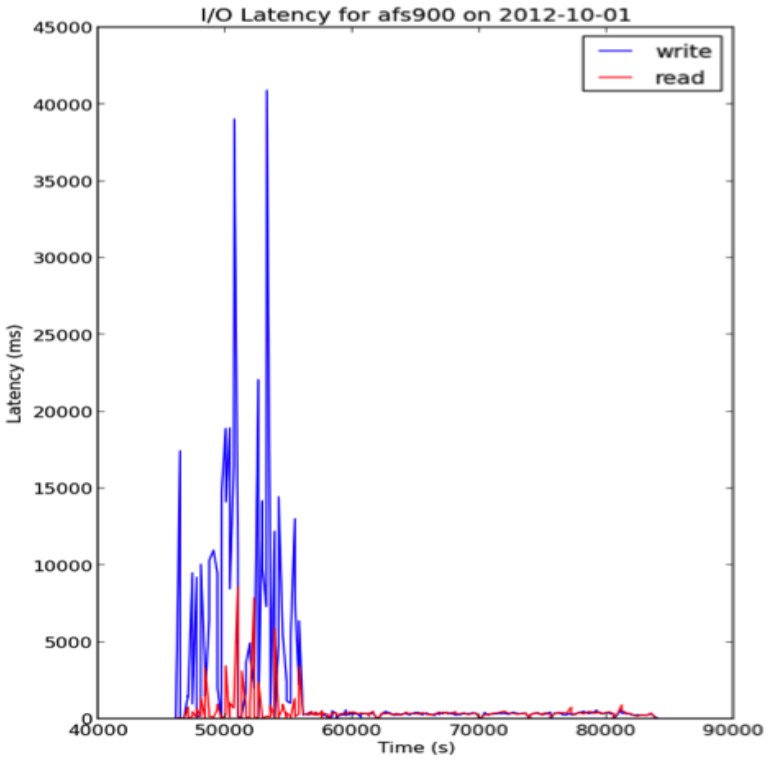
... studying code ... analysing dumps ... looking through logs ...

- Synthetic stress testing
  - file server accesses from a fast batch queue
    30 clients: OK, but 180 clients: server unresponsive
- `rxperf`
  - Rx testing tools that ships with OpenAFS
  - 5 clients: server unresponsive!
- Why two limits? UDP buffer sizes!
  - listener thread is bottleneck of handling incoming packets
  - packet queue length ~f(#clients)
  - UDP buffer overflow = lost packets = RX busy handling errors = not feeding worker threads = latency goes up
  - `/proc/net/snmp` show inError rate > 10%

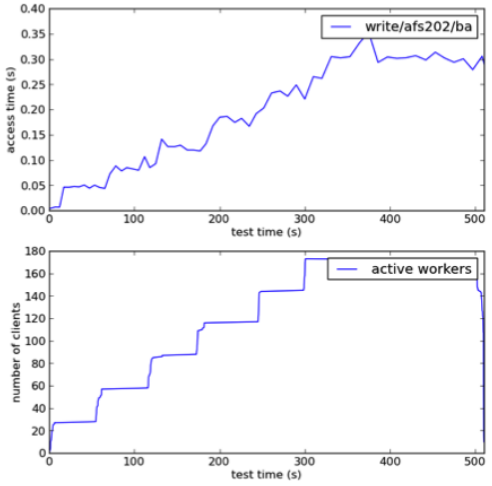
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Server-side: "Rx-limit" (2)

- UDP buffer size of 16 MB eliminates the problem
  - dramatically improves access latency under load
  - can sustain 8 heavy users w/o server slowdown



- applied during incident
- access time from 40s to 300ms
- server quite usable



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

Server-side: "Rx-limit" (3)

- Presented at EAKC 2012
  - positive feedback, "unknown" issue
  - general recommendation by the lead developers on ML to review these settings
  - several sites have deployed the fix since
- Deployed at CERN since autumn last year
  - rate of performance related tickets ("AFS is slow!") gone down dramatically since
  - mostly throttled users, hammering from batch while trying to work interactively in the same directory

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

Server-side: "ulimit crashes"

- Newly deployed servers crashed every few hours

```
#0  0x000000000044c81c in GetHandler (fdsetp=0x6fb920,
    maxfdp=0x7ffa1d2b3d1c) at ../vol/fssync.c:816 sdf

816         FD_SET(HandlerFD[i], fdsetp);
```

- Core analysis showed flipped bits
  - `-33554433` is all ones with a zero at bit 25, `-32769` is all ones ...
  - so we found the problem: bad memory!
  - however: extensive memory testing revealed nothing :(
- Back to the core ... `FD_SET` ...
  - will set random bits when `FD_SETSIZE > 1024` !
  - default value changed in RHEL kernel 2.6.32-279 to 4096
    (increased on purpose to expose "dangerous" apps :-)

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

What else?

- New backup system
  - currently being phased in
  - to overcome some of the current system's limitations
- 1.6.x
  - Triage of CERN patches
- Wigner Data Center
  - "shadow" cell for disaster recovery?
- AFS on the cloud
  - we have an S3 prototype that works with Huawei and Openstack/Swift

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

OpenAFS and IPv6: Status

The situation today:

- There is no support for IPv6 in OpenAFS
- There is no activity to add IPv6 support

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

OpenAFS and IPv6: Reasons

- Why no IPv6 support?
  - AFS is a complex distributed system
  - IPv4 is embedded everywhere
  - Backwards compatibility is high priority
- Why no IPv6 activity?
  - No serious request (read: with funding)
  - Don't assume OpenAFS is for free!
- Bottom line:

  IPv6 in OpenAFS won't come by itself.

OpenAFS and IPv6: Options

- Rely on dual-stack
  - comes with some limitations
- Fund the implementation
  - code would become mainline
  - timelines? price?
- Do it ourselves
  - accepted for mainline?
  - timelines? price?
- Look for AFS alternatives
  - "dropbox" for home, CVMFS for s/w, NFS for batch/dev, ...

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

OpenAFS and IPv6: Options

- Rely on dual-stack
  - comes with some limitations
- Fund the implementation
  - code would become mainline
  - timelines? price?
- Do it ourselves
  - accepted for mainline?
  - timelines? price?
- Look for AFS alternatives
  - "dropbox", CVMFS, NFS, ...

BoF session later today!

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

Summary

- AFS is able to grow
  - Combination of hardware and software changes
  - User have 50-100 times more space
- Performance has improved
  - The UDP fix was a major improvement
  - The demand increases and new bottlenecks show up
  - Continuous effort
- IPv6 support
  - See BoF session later today

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