Effective storage usage at DESY, Zeuthen

A guide to success ;-)

Andreas Haupt – DV –
Data Science Seminar, Zeuthen

Motivation

… what you should take home from this talk

- The Computing Centre hosts lots of (expensive) compute/storage resources
  - Aim to get the most output and value out of them
- Disclaimer: this talk is mainly a repetition of another talk presented 3 years ago
  - Nevertheless, the basic message still applies
  - Incorrect usage of central resources can badly affect performance & availability
    - not just only for you, but for every other user!
    - so it helps all of us to avoid bad usage patterns ;-)
- Another disclaimer …
  - Amount of information is quite huge and was shortened at some points
  - In case of missing details just ask :-)

Overview

Numbers, numbers, numbers ...

- Central storage resources

Storage product	Use case	capacity / servers	Notes
Lustre	mass storage with fast parallel data access (scratch space)	~3.5PB / 36	
dCache	data archive, mass storage with fast parallel data access	~8.5PB / 44	
AFS	$HOME, software, scratch	~260TB / 22	
Sync&Share	Document share		Provided by DESY-HH

Central storage systems

… some general remarks

- Network-attached storage is supposed to provide:
  - Data access on whichever client system you are running your stuff
  - Fast, aggregated data throughput by accumulating the power of many storage servers

- However, this flexibility has some drawbacks
  - Latency matters!
    - metadata i/o operations are comparable slow due to complex operations in background
    - open/close ops, recursive find, compiling software!
  - File locks are usually only enforced on the same client
  - Your notebook ssd will perform significantly better for some usage patterns
    - Keep this in mind whenever you are migrating software/workflows to our central compute environment

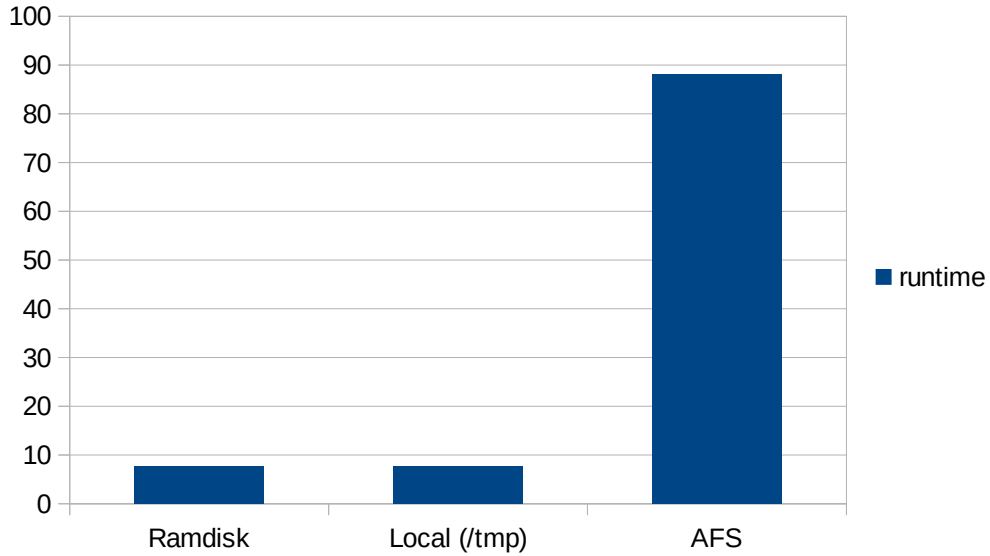
Central storage systems

… even more general advises / rules of thumb

- Avoid i/o as much as possible
  - example: avoid reading the same files again and again
- Avoid too many clients writing into the same directory
  - … and avoid too many files (>10k) per directory
- Avoid concurrent writes from multiple clients to the same file
  - in many cases the result will not be the expected one ...
- Try to move i/o operations which are affected by latency to local scratch (/tmp, $TMPDIR)
  - Example: untar software in /tmp, compile in /tmp but install to AFS
- In extreme cases: consider usage of ramdisks (/dev/shm)
  - but do not forget to clean up afterwards

Extract Linux kernel (>70k files)
Runtime in seconds:

```
tar -xJf linux-5.6.15.tar.xz
```



Inodes & blocks

… the undividable file-system units

- An inode (index node) is a data structure in a file system that describes a file-system object
  - … like files, directories, symlinks, etc.
  - Smallest unit to store data: block
- A file system has a fixed maximum amount of inodes and blocks usually (quota!)
  - Usual (data) block size on local file systems: 2kB – 4kB
  - and on network-attached cluster file systems: 64kB – 1MB (!)
  - An average file size is assumed during file-system creation: avg size = sum(blocks) / sum(inodes)
- Files bigger than the block size are split over multiple blocks
  - Files smaller than that allocate one block, anyway
- 3 ways how to occupy 10GB on a file system with 1MB block size
  - 1 file á 10GB
  - 100 files á 100MB
  - 10000 files á 1B (!!!)
- Homework: try to find out why it is a bad idea to store a Python venv on a file system with 1MB block size ;-)

In case of problems ...

when things go wrong ...

- In case of data access issues (slow ops, errors, etc.) contact uco-zn@desy.de
- Please be as much precise as possible in describing your problem!
  - It really helps us identifying the possible source of the problem
- A typical complaint: “My access to cluster is slow!”
  - … but what does that mean?
  - Therefore here some details we always would like to here from you:
    - Which operations do not perform as desired?
    - Which files/directories are affected?
    - Which client(s) are you using when hitting the problem?
    - In case you use many clients: Are all of them affected? Or just some of them? Which ones?

Getting storage resources

I need more space!

- Please keep in mind that we are not allowed to purchase large amounts of storage “on spec”
  - mainly: financial restrictions ...

- Unfortunately, especially for Lustre storage, we are not as flexible in providing new space as we want
  - Lustre is not designed to get easily extended once is has been established
    - “just put in another machine” does not really work :-(
  - We usually establish **one Lustre file system every 2nd year**
    - This should cover all the year's demands
  - So, we need to know yearly storage demands **well in advance**!

- Therefore: the first person to contact should always be your group leader!
  - He (alternatively: your group's computing contact) should have an overview of your group's resources
  - … and should announce further demands directly to us or via the Computing Board

Storage solutions

AFS

… still the work horse for personal data

- Organized in volumes with quotas
  - Your $HOME directory on WGS is one volume
  - Volumes have mountpoints (the place they are accessible for the client)
  - Partly self-service management via group admins
- Access restrictions via ACLs (access control lists) per directory
- Centrally managed backup
  - … and last day's snapshot of $HOME available in ~/.OldFiles !
- Authentication based on a so-called „AFS token”
  - Will be generated automatically during the login process
  - … but can and will finally expire if not renewed in time (lifetime 25 hours)
  - Expiring AFS tokens are the main source of access problems!
- Unfortunately AFS not designed to work flawlessly on mobile devices (like notebooks)
  - use sshfs or your client's “connect to server” (connection type: ssh) instead

AFS

… basic management

- Group‘s AFS space managed by groups themselves:
  - First person to contact: your group’s computing contact
  - Create, increase, … volumes beyond **/afs/ifh.de/group/<group>**
- Please do not mess with ACLs directly in $HOME
  - Use subdirectories and adapt ACLs there
  - Classic Unix access rights are mainly ignored, so something like this is useless:
  - `[wgs34] ~ % chmod 0600 my-private-file`
- Some subdirectories in AFS $HOME are automatically created during account initialisation:
  - ~/private : your private space, ACLs adjusted so that only you can access files stored there
  - ~/public : the opposite – any other user with AFS access can read files
  - ~/public/www : content visible on our public webserver (https://www.zeuthen.desy.de/~<user>)

AFS ACLs explained

- AFS rights:

l	lookup - read entries in this directory
r	read – read files in this directory
w	write – write files in this directory
i	insert – add new files to this directory
d	delete – delete files from this directory
a	administer – modify ACLs in this directory

- Some pre-defined groups are often set:

system:anyuser	any user in the world
system:administrators	an administrator
ifh-hosts	all hosts at DESY Zeuthen
desy-hosts	all hosts at DESY (Hamburg + Zeuthen)
group:<group name>	members of a DESY group

example ACL:

```
[wgs1d] /project/singularity % fs la .
Access list for . is
Normal rights:
  desy-hosts rl
  ifh-hosts rl
  system:administrators rlidwka
  group:usg_zn rlidwka
  system:anyuser l
  znasw rlidwka
```

AFS usage examples

… yes, you can do this at home :-)

- Show quota:

```
[wgs34] ~ % fs lq ~
Volume Name                    Quota       Used %Used   Partition
user.ahaupt                  2000000    1549680   77%         16%
```

- List ACLs:

```
[wgs34] ~ % fs la ~
Access list for /afs/ifh.de/user/a/ahaupt is
Normal rights:
  system:administrators rlidwka
  system:anyuser l
  ahaupt rlidwka
```

- Add an ACL to a directory:

```
[wgs34] ~ % fs sa ~/friends group:cta rl
[wgs34] ~ % fs la ~/friends
Access list for /afs/ifh.de/user/a/ahaupt/friends is
Normal rights:
  group:cta rl
  system:administrators rlidwka
  system:anyuser l
  ahaupt rlidwka
```

AFS usage examples

… continued

- Instant (world-wide) data sharing:

```
[wgs34] ~ % echo 'very cool content' > ~/public/www/my-share.txt
[wgs34] ~ % curl https://www.zeuthen.desy.de/~ahaupt/my-share.txt
very cool content
```

- Check if your AFS token is still valid:

```
[wgs34] ~ % tokens

Tokens held by the Cache Manager:

Tokens for afs@ifh.de [Expires May 29 09:20]
   --End of list--
```

- Fetch a new AFS token (including a new Kerberos ticket):

```
[wgs34] ~ % kinit
ahaupt@IFH.DE's Password:
```

CEPHFS

… future AFS replacement?

- Not available yet for public usage
- Planned as replacement for AFS $HOME in future
  - … with much more initial quota!
- Identical $HOME directories on Linux (WGS, batch) nodes and Windows clients
  - Export to Windows / MacOS / Linux desktop clients via Samba

Lustre

… put your huge datasets here

- Large and fast “scratch space”
  - Suitable for any bulk data (in large files preferably!)
  - NOT designed to host source trees, executables, etc.
    - minimum internal block size is 1MB – so if you think you read a 5kByte file, you read 1MB instead!
  - There is NO BACKUP!
- Currently five independent instances active:
  - /lustre/fs{20..24}
- Every Lustre file system lives for 5-6 years and gets replaced by a new one after that
  - Group's task to migrate data still needed
  - A bit of an effort, but also an easy way to clean up – by just doing nothing ;-)

dCache

… the data archive

- Biggest data provider at DESY (not just only in Zeuthen)
- Storage solution to share huge datasets with collaborations anywhere in the world
- Transparent tape backend possible
- Many access protocols
  - dCap (deprecated)
  - NFS-4.1 (makes it feel just like a „normal“ file system)
  - GSIFTP
  - WebDAV (https)
- Many authentication methods exist
  - GSI (Grid security infrastructure with X.509 certificates)
  - Kerberos
  - Macaroons (bearer tokens) – only for WebDAV access
  - OIDC (only tests atm.)

dCache usage

- dCache only partly behaves like a normal Posix file system
  - no file modifications possible – modification means: deletion / re-creation
- Two mountpoints exist:
  - /acs : Old, legacy entry point
    - Please do not use it any longer!
  - /pnfs/ifh.de/acs : nowadays preferred NFS-4.1 based access
    - Handle files just like on any other file system
- Beware of bulk data operations on areas with a tape backend
  - Unfortunately these areas are not easily identifiable from user's perspective
    - … but had been agreed with group’s contact “some time ago”
  - Nevertheless, usually only group's computing admin (or your collaboration) will modify content
  - Massive staging from tape (i.e. >1000 files) should be pre-announced to dCache admins

Sync&Share

… aka „DESYCloud”

- Provided by colleagues at DESY Hamburg
- Based on the file hosting service “NextCloud”
  - … with services like OnlyOffice directly integrated
- In theory without any quota!
- Data accessible via web browser, mobile app as well as sync client on any client device
  - Unfortunately data access on centrally managed nodes (wgs, farm, etc.) not available, yet!
    - Still unclear, when this will change :-(
- Some (un-)official documentation on how to access it & sync data from command line:
  - https://confluence.desy.de/display/MXW/rclone+and+desycloud
  - You will need to create an “app password” for this access
  - See FAQ here: https://it.desy.de/services/storage_services/desy_sync__share/index_eng.html
- Anyway, the perfect place to store and share your documents!

Storage summary

- Unfortunately a “silver bullet” that perfectly fits all storage use cases does not exist
  - It's also questionable whether it will ever exist …
  - Even if it exists, it will be very expensive most probably
- Some rules of thumb where to put your data:

Kind of data	Where to put
Your thesis	AFS $HOME / Sync&Share
Collaboration's, group's or private documents	AFS $HOME / Sync&Share
Collaboration or group (raw) data	dCache
private or group's simulated / derived data	Lustre
Software, histograms, ...	AFS group space

That's it folks ...

- Please also read our docs:
  - Storage: https://dv-zeuthen.desy.de/services/storage_recources/
  - Batch: https://dv-zeuthen.desy.de/services/batch/
- … as well as some hints to set up your notebook for easier remote access:
  - https://dvinfo.zeuthen.desy.de/BYOD/User-Info