Storage best practices @CC-IN2P3: Semi-Permanent Storage (SPS)

Loïc Tortay, 25 November 2019

This support is shared under the Creative Commons license:
Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)

https://creativecommons.org/licenses/by-nc-sa/4.0

EN: This license allows others to remix, tweak, and build upon this work non-commercially, as long as they credit the authors and license their new creations under identical terms
FR: Cette licence permet à d'autres personnes de remixer, d'ajuster et développer ce travail de manière non commerciale, à condition qu'elles créditent les auteurs et accordent une licence à leurs nouvelles créations aux mêmes conditions

- Per group shared (disk) space accessible from all computing and login nodes
- Storage for **active** data files accessed by jobs, not for long term (> 1 or 2 years) storage
- Filesystem: transparent access for programs using the standard UNIX API (POSIX), usual UNIX commands (`ls`, `cp`, etc. including `scp`, `sftp`, `rsync`, ...)
- Accessed through: `/sps/$GROUP` (or `/sps/hep/$GROUP` for some groups)
- Data transfer to/from long term storage (HPSS) using:
  - iRODS (`iput`/`iget`)
  - RFIO (`rfcp`)
  - XRootD tools (`xrdcp`, from HPSS only)

- Programs & sources, jobs log files OK
- **Not** a backup space for AFS, PBS, your laptop(s) or home machine(s)
- With very few exceptions (primary copy of unique experimental data only present at CC-IN2P3), **no** backup of the content: files removed are forever lost
- SPS relies on two distinct infrastructures/storage technologies: Spectrum Scale (GPFS, IBM) & Isilon (NFS access, Dell-EMC)
- From the user's point of view the infrastructure used is mostly unimportant (same access point & access interface)
- 107 groups, ~3.3 PiB used, ~1.2 billion files

- Active data files: last access (read) time **less** than 2 years ago
- Each users group is responsible for managing its allocated space
- New groups **must** define a data management plan (a.k.a. DMP)
- Automated cleanups **will return** for all groups:
  - cleanup trigger threshold (high watermark) is 95% space occupancy
  - low watermark (aim for cleanup) is 80% of space used
  - files not accessed for at least 2 years are removed

- low and high watermarks can be adapted for each group
- the last access time limit can also be adapted (to **less** than 2 years)
- however, the 2 years access time is a hard limit (but exceptions can be granted depending on the computing model)
- groups **can** define a single top-level directory which is excluded from cleanup

- When a cleanup occurs, an e-mail with a link to a report can be sent to a group provided address or set of adresses (*czars*, user's group mailing list, etc.)

- A quota management delegation tool is available: `spsquota`
- At the moment, `spsquota` is **only** available for groups using the **GPFS** infrastructure, the tool with support for the Isilon infrastructure will have the same interface and (almost) the same features
- `spsquota` allows:
  - *czars* to get & set individual users quotas for their group
  - *czars* to get & set default users quotas for their group
  - ordinary (*non-czar*) users to see their own quotas and current usage
  - ordinary (*non-czar*) users to see the default quotas

- `spsquota` documentation (GPFS focused):
  https://ccspsmon.in2p3.fr/spsquota
- With the GPFS infrastructure, `spsquota` can manage quotas for:
  - space (bytes used by files content)
  - files (number of files, including directories & symlinks)
- The Isilon infrastructure only supports space quotas, not files quotas, but files usage may be displayed by `spsquota` with Isilon support
- If quotas need to be defined or changed for groups using the Isilon infrastructure, please contact the CC-IN2P3 user support team

- Daily reports to help data management in SPS
- All reports are accessible to **all** users (not just the group)
- 3 main reports available:
  - Space used by user (how much space/how many files for each user in the group):
    - https://ccspsmon.in2p3.fr/users
    - 3 months history directly accessible in the reports
  - Space used by top-level directory (somewhat similar to "`cd /sps/$GROUP && du -sh *`"):
    - https://ccspsmon.in2p3.fr/dush
    - details can be adapted for each group

- 3 months history directly accessible in the reports
- Cleanups (simulated or not):
  - https://ccspsmon.in2p3.fr/cleanup
  - cleanup parameters can be adapted for each group
  - limited 2 weeks history

SPS: Criteria for space allocation/extension

- Yearly requests for space increments
- User support ticket to trigger space allocation
- Space allocation/extension is allowed **iff**:
  - less than 25% of space used has last access time older than 1-2 years
  - current space usage is at least 75%
  - requested allocation > free space

- Limit the number for files: more files ➪ more work managing data
- Avoid (lots of) extremely small files (< 512 B or 1 KiB), use a database when it makes sense
- Be careful with filenames: way too many files named `*`, `$`, `ESC`, `\n`, `\`, `:wq!`, etc.
- When transfering files to/from iRODS (`iput`/`iget`), please use `iput`/`iget -N0` on the login nodes (`cca.in2p3.fr`)
- Locate files with `find`, **not** `grep -R`
- Beware of programs setting the last access time (`tar xp`, `rsync -a`, ...) interactions with cleanup

- Use the default permissions we define when creating your space: they allow group read access
- Do no use `chmod -R 777`, `chmod 777` (or similar `chmod +arwx`, `chmod -R +arwx`, ...): giving *world* write permission serves no purpose
- Extended ACLs can be defined, but can rapidly become complicated and/or counter-intuitive
- Consider using `ln` (not `ln -s`), or even `lndir`, instead of `cp` when duplicating a directory tree (no space used), then use `cp` for the files you want to modify
- Put temporary files in `$TMPDIR`

- Put **temporary** files in `$TMPDIR`
- Access files from SPS directly, unless:
  - writes are temporary: copy to `$TMPDIR`, modify there, discard
  - hundreds or thousands of jobs would write concurrently to SPS: copy to `$TMPDIR`, modify there, copy back to SPS after writes are done
  - a single file (or set of files) is independantly or randomly **read** by many jobs (e.g. ROOT files), especially if the file is large (GiBs)
- Large ROOT files can be read efficiently from XRootD
- Jobs can cleanup `$TMPDIR` when they end, but the batch system will do it anyway

- Multiple jobs modifyfing the same file in SPS will probably not do what you might expect unless you take appropriate measures (*locks*), that includes log files
- Multiple jobs modifying the same file copied from SPS to `$TMPDIR`, will not work if the file is expected to be copied back to SPS
- Avoid compiling in every job, compile once (in SPS or PBS) and run the program from there
- Do not copy a directory from SPS to `$TMPDIR` and copy back the directory tree from `$TMPDIR` to SPS at the end of the job, especially with multiple jobs & when the directory tree contains log files: copy back **only** what was changed/created