Jobsub_lite proposal: Tarball timestamps

Shreyas Bhat on behalf of the jobsub_lite group

FIFE Group Meeting

May 27, 2023

Background and Steps to Reproduce

- Background: Uploading tarballs to RCDS from wrappers that call jobsub_submit
- Specifically, the -f dropbox:// flag and the --tar-file-name tardir:// flag
- Steps to reproduce:
    1. Create a tarball
    2. Submit a job using -f dropbox://, pointing to that tarball
    3. jobsub_lite will upload that tarball to RCDS.
    4. Now recreate the SAME tarball, with the same filename.
    5. Submit a job using -f dropbox://, pointing to the second tarball
    6. jobsub_lite will see this second tarball as a different file, since even though the files are the same, the underlying tarball has different timestamps.

How this could happen, and what's the effect?

- Experiment wrapper scripts could easily do this if they're called twice, but tarring the same directory to pass to jobsub_submit
- Increases number of file uploads to RCDS (increasing server load unnecessarily)
- Slows down submission time

Proposal

- Any time jobsub_lite creates a tarball (the -f dropbox://, --tar-file-name tardir:// cases for RCDS uploads), coerce the timestamps on underlying files to be a sentinel value (one second past the epoch, for example), so that any hash calculation on two identical files yields the same hash.
- This action will NOT change the original files on disk - only the copies of the file that jobsub_lite adds to the tarball it uploads to RCDS
- Benefits:
  - We anticipate that implementing this proposal will decrease the number of duplicated files uploaded to RCDS, which lessens the load on the RCDS servers
  - Speed up submission time