

Concatenation @ remote sites

- Need to start generating next round soon
- Risky schedule for using SAM concatenation tools
- Instead, develop simple home-brew solution
- Concatenation on-the-fly:
 1. Carefully split sample to generate into N “projects”, such that each project will produce X GB of output
 2. Cron job to monitor simulation output & submit concatenation jobs

cron job to submit concatenation jobs

Runs on head-node; can see durable cache.

1. Find any new small files in durable cache.
2. Read project list to determine the project number associated with each small file.
3. Move each small file to an intermediate directory called DATASET.PROJECT
4. For each directory DATASET.PROJECT, determine if all required small files exist (must again read project list).
5. If all small files exist, then submit a concatenation job.

Status

- Cron job mostly written.
- Testing with 5.3.3 wtop3i sample
 - Submit simulation jobs to DCAF(s)
 - Send small files to fcdpdata004
 - Cron job submits concatenation jobs to CAF
- Limited testing because no cdfmc queue on new CAF (should be resolved today)
- Concatenation job cannot delete small files.
- Need to add error recovery to simulation script
- Still no error recovery of concatenation jobs.

Error recovery

1. Simulation failure

- If the simulation script can catch it, then send a `ti022904.0001topu.ok` file to durable cache. This is a signal to the cron job to disregard this file in concatenation job.
- If the simulation job cannot catch error, then cron job must!
If `ti022904.0001topu.proj` file created > 24 hours ago, then assume simulation failure.

2. Concatenation failure: cron job can handle it. Needs more thought.