LUXE SAS meeting, 2nd May 2022

Proposal for dataset naming and plan for grid storage

Yee Chinn Yap, Federico Meloni (DESY)

+ fruitful discussions with Sasha and DESY local members (Mikael, Jenny, Frank)

Grid

- We want to eventually move from using DUST for data storage to grid.
  - DUST is currently close to full and has no back-up.
- Dataset naming on grid should have a flat structure and not in the hierarchical directory structure as on DUST.
  - Dataset name should be concise and contains all important information at first glance and to trace its provenance.
  - Character limits in DDM.
- Project name: mc/data + 2-digit year
  - For MC, year refers to when the campaign started, does not necessarily coincide with the production year.
  - mc21 (old), mc22 (current)
  - In the future when the production changes significantly e.g. with a uniform software framework: mc23 (if this happens in 2023).
- Each project contains dataset containers which contain files from different jobs/bunch crossings.

Naming policy

- Format should work for signal and background MC as well as data.
- "." must only be used to separate fields of the dataset name and "_" may be used to separate parts of a field.

{Project}.{Process}.{Generator}.{Type}.{Tag}

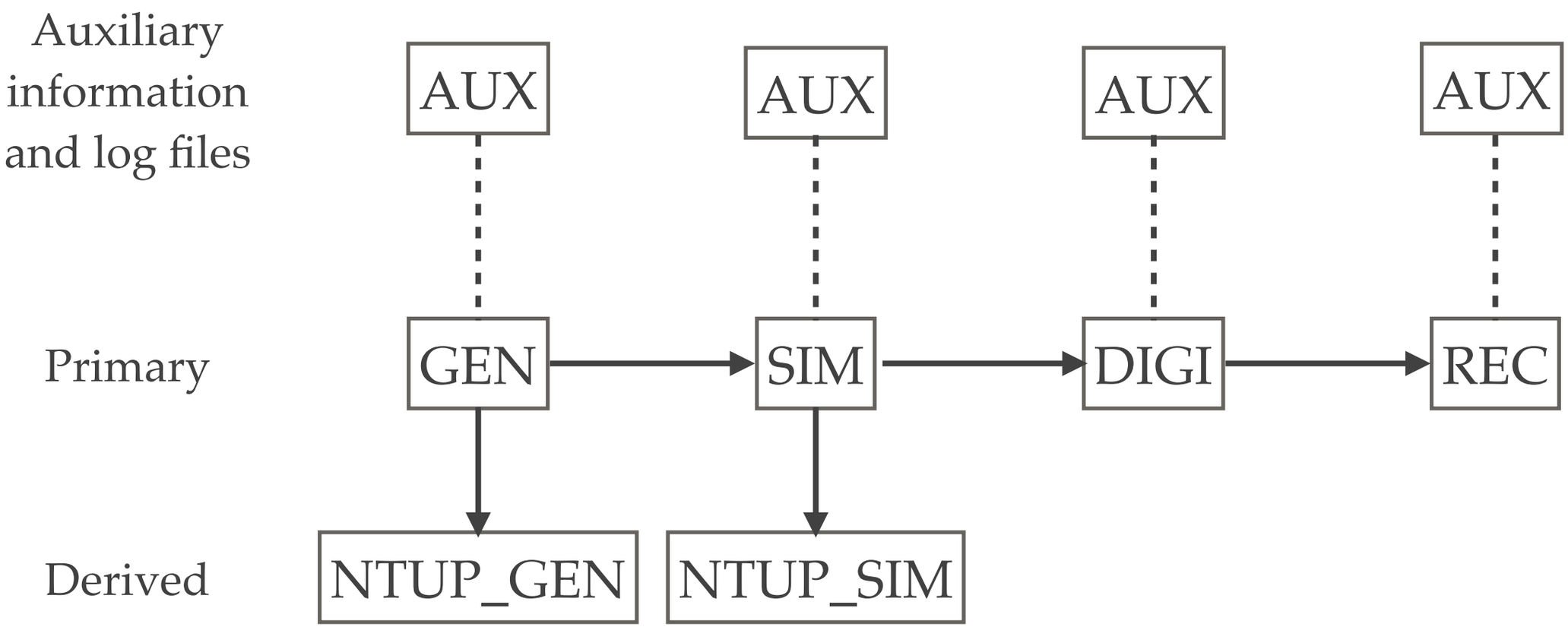
- Process:
  - For signal, process should also include xi value and laser power separated by "_" and "p" as decimal, e.g. e-laser_xi5p0_40TW, brem-laser_xi0p15_350TW, etc.
  - Background examples: ebeam, brem, ics.
  - Single particles: singlemuon, singlepositron, etc.
  - Others (BSM?): ALP_m10MeV

Naming policy

{Project}.{Process}.{Generator}.{Type}.{Tag}

- Generator:
  - Signal: ptarmigan, IPstrong (for archival)
  - Background: G4gun
  - Data: runXXXX, periodXXXX
- Type: GEN, SIM, DIGI, REC, NTUP_GEN, AUX, RAW etc. (See next slide)
- Tag: 4 digit tags for generation, simulation, reconstruction, etc. gXXXX, sXXXX, rXXXX, dXXXX. (See later slides)
- Files: {Project}.{Process}.{Generator}.{Type}.{Tag}.{Number}.{Extension}

Provenance



For examples, many don't exist at the moment

G-tag

- For Ptarmigan, this should contain the version used and a corresponding template yaml file.
  - The individual yaml files should still be stored in the AUX container so that the information is kept and samples are be reproduced if needed.
- For each sample production requiring yaml file that differs other than in the random seed and xi value, a different g-tag should be created.
  - One tag for all xi values in e-laser, another for berm-laser.
  - Different tag for different laser power.
- From a given g-tag (via the corresponding template yaml file), one would know e.g.:
  - beam spot size, polarisation, beam offset, bunch length, electron beam energy, etc.
  - which particle weight scheme (i.e. increase_pair_rate_by) and dt_multiplier.

S-tag

- S-tag should correspond to a given snapshot of the LUXE simulation (i.e. branch and commit hash)
- Each s-tag should have corresponding template Mac files (which contain the settings).
- From the s-tag, one would know e.g.:
  - G4 version
  - geometry version/commit
  - magnet strength, shielding, physics list
  - approximations in particle tracking
  - fast/full sim
  - parameters for the background generation

Tags

- Further tags can be added as necessary, e.g. r-tag is foreseen for reconstruction.
- Tag information can be managed in git during the production.
  - First create tag then launch production.
    - It's the responsibility of whoever launches an official production to do this systematically.
- For now, the tags and what they correspond to should just be documented on a confluence page.

Examples

- After each step of processing, the new file name will have different {Type} plus extra tag at the end.
- mc22.e-laser_xi0p5_40TW.ptarmigan.GEN.g0001 processed with two different geometries/G4 versions becomes
  - mc22.e-laser_xi0p5_40TW.ptarmigan.SIM.g0001_s0001
  - mc22.e-laser_xi0p5_40TW.ptarmigan.SIM.g0001_s0002
- After reconstruction
  - mc22.e-laser_xi0p5_40TW.ptarmigan.REC.g0001_s0001_r0001

Signal examples

- /nfs/dust/luxe/group/MCProduction/Signal/ptarmigan-v0.8.1/e-laser/phase0/gpc/0.5/e0gpc_0.5_*_particles.h5
  ☞ mc22.e-laser_xi0p5_40TW.ptarmigan.GEN.g0001

Ptarmigan

- /nfs/dust/luxe/group/MCProduction/Signal/ptarmigan-v0.8.1/e-laser/phase0/gpc/0.5/e0gpc_0.5_*.yml
  ☞ mc22.e-laser_xi0p5_40TW.ptarmigan.AUX.g0001

Ptarmigan yaml

- /nfs/dust/luxe/group/MCProduction/Signal/g4/ptarmigan-v0.8.1/e-laser/phase0/lp/e0lp_10_0_0_particles_g4.root
  ☞ mc22.e-laser_xi10p0_40TW.ptarmigan.SIM.g0002_s0001 (new g-tag because of different polarisation)

Geant4

- /nfs/dust/luxe/group/MCProduction/Signal/IPstrong_V1.1.00/JETI40/g_laser/16.5GeV/w0_3000nm/g_laser_JETI40_16.5GeV_0.8Jpulse_3000.0nm_w0_1000_zmsh_10000_nmp_IPstrong_V1.1.00_run_1000_events.stdhep
  ☞ mc21.brem-laser_w3_40TW.IPStrong.GEN.g0003 (based on waist instead of xi)

IPStrong

Background examples

- /nfs/dust/luxe/group/MCProduction/Background/elaser/10022022_e67ace8b/sim*_run0_9.root
  ☞ mc22.ebeam.G4gun.SIM.s0001 (no g-tag)

  E-laser bkg

- /nfs/dust/luxe/group/MCProduction/Background/elaser/10022022_e67ace8b/log/sim*_log.tar.gz
  ☞ mc22.ebeam.G4gun.AUX.s0001

  E-laser bkg log

- /nfs/dust/luxe/group/MCProduction/Background/gammalaser/18102021_55ae8938/ob/sim*_run0_9.root
  ☞ mc22.brem.G4gun.SIM.s0002

  brem-laser bkg

Currently on DUST

- In /nfs/dust/luxe/group/MCProduction:

```
hodnoam  Dec  6 14:37 BSM
lhelary  Apr 27 17:00 Background
fmeloni  Dec 29 11:15 Digitisation
lhelary  Mar  3 16:07 Signal
ychinn   Apr 21 08:31 SimplifiedTrackerSim
lhelary  Feb  8  2021 SinglePositron
ruth     Nov  8 14:04 Testbeam
hallforj Jan 18 20:56 e-laser-cherenkov-xi-scan
lhelary  Nov  3  2020 py.py
lhelary  Mar  1 15:29 tmp
santraar Feb  1  2021 user
```

→ By far the largest in terms of space usage
→ 5 datasets. Type: DIGI_PIX, tag: dXXXX?
→ IPStrong, ptarmigan and g4, many are old
→ Csv files of tracker hits using simplified sim
→ Code. Can this be deleted?
→ Fluka simulation + other things
→ Empty. Can this be deleted?

- Google spreadsheets with list of directories and plans for them (delete/keep/move to grid).
  - Please have a look and comment!
  - Proposed new names for those marked for moving.

Plan

- We'll have some storage space on grid courtesy of ILC.
- The plan is to start archiving old files there, i.e. those mc21 files.
- Rearrange the datasets currently on DUST with the new naming and move them to grid.
  - I'll take care of the moving but the renaming and rearranging should be done by the person with ownership of the files (you'll be contacted).
  - Google spreadsheets for organising.
- We'll stick with the ILC VO for the time being. The current space will need to be expanded by buying more disk (exact amount to be worked out).
  - When that's available, the rest of the files as well as new production should also go on grid.