Efficient Ways to Run On Data

Carl Gwilliam

UNIVERSITY OF LIVERPOOL

16th June 2010

Liverpool Performance Meeting

Introduction

- What is the most efficient way to run over the data?
  - Especially as it gets larger
- Several possibilities:
  - Run directly on physics containers
  - Skim using TAG
    - Database
    - File-based
- Which is best will probably depend on if making histograms directly or ntuples
- How much can we rely on (sub)groups to skim the data?
  - Might need to further skim ourselves
- Don't know the answer but would like to start a discussion …

Physics Containers

- Physics Containers are a collection of coherent datasets
  - e.g data from may reprocessing
- This allows to run over a self consistent set of the data without having to specify the dataset for every single run
- If you're producing histograms directly it will become annoying to run over all the data every time
  - Probably OK if producing ntuples
  - Can possibly speed up by combining with TAG selection on-the-fly
    - Saves opening every single AOD/ESD
- Further information:
  - https://twiki.cern.ch/twiki/bin/view/AtlasProtected/PhysicsContainers

Skim using TAG

- TAG contains limited summary information on the event
  - e.g. triggers, multiplicities, four-vector info of 4 highest P_t electrons/ muons …
- Can use the TAG to skim the data and write out a much reduced set of events
  - Faster when running analysis but have to do the skim first
  - Can just update skim when new data comes in (until repro)
  - Are we going to be allowed to store (large) skimmed datasets
    - Can possibly get them "sponsored" by a group
- Database or file-based tags:
  - Can use central services e.g. ELSSI to run over database tag
  - Can perform more complicate queries using file-based tags
    - e.g. form four vectors, invariant masses …

ELSSI

https://voatlas18.cern.ch/tagservices/index.htm

- Relatively user-friendly web interface for querying TAG DB
  - Apply selection:
    - GRL
    - Trigger
    - Physics selection
      - "Njet>1 and abs(JetPt2)>10000 and abs(JetEta1)<2.5 and abs(JetEta2)<2.5"
  - Count number of selected events
  - Either produce TAG file for selected events or directly skim the events to produce skimmed AOD/ESD
- Experience:
  - Was able to use ELSSI to make a dijet skim of all April reprocessed data
  - Production ELSSI site not fully up-to-date so had to use nightly version
  - Took a relatively long time to skim - 3/4 days for April repro (long tail)
  - Had problems with too long DB query when try to repeat for May repro
- Can also use ganga directly with file-based or DB TAG
  - More control over jobs (e.g. number of subjobs)
  - https://twiki.cern.ch/twiki/bin/view/Atlas/PullGangaAtlasTutorial#8_6_Using_TAG_Files_with_Ganga