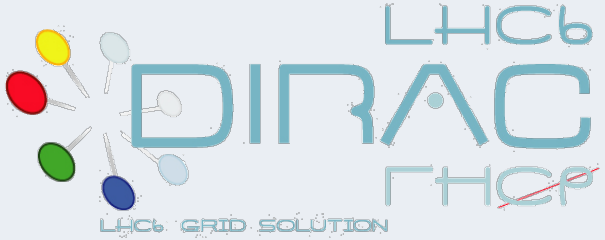


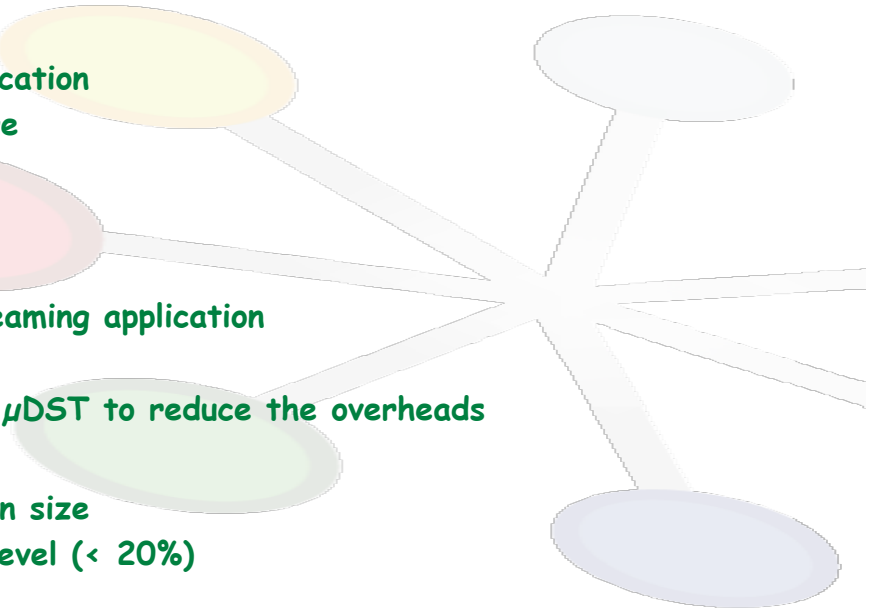
Distributed Computing for Run 3

Basic assumptions

- Most of the HLT output will be TURBO
  - Throughput? Let's assume 5 GB/s for the sake of this talk
    - This is already 4 times the "impossible" goal for ALICE in the early days of WLCG
  - One single online stream is sufficient
    - as discussed earlier this week
  - Run structure (1 hour max) as now for DQ and calibration purposes
- Will there still be some other streams?
  - Probably a TURCAL-like stream for calibration
  - Any "to-be-reconstructed" stream (à la FULL stream)?
- TESLA will just be a conversion and streaming application
  - Lossless, i.e. no need to "reprocess"
  - All events go to at least one stream
- A few numbers then
  - 18 TB/run : 1000 files of 18 GB
    - File size limited by what a job can download (don't expect much more)

Offline data processing

- Mostly streaming TURBO data
  - Keeping the run granularity as now
- A possible model
  - First level processing
    - Produce 10 streams with a TESLA application
    - Each file is then 1.8 to 2 GB on average
  - Second level processing
    - Merge and Stream all in one
    - Using 10 input files (18 GB)
    - Produce 10 streams with a specific streaming application
    - Each file again 1.8 to 2 GB
    - Optionally further merge the streamed μDST to reduce the overheads
  - Requirements
    - Streams should be as even as possible in size
    - Overlap should remain at a reasonable level (< 20%)
- Result for each run
  - 100 streams with 100 files of 2 GB each (or 20 files of 10 GB)



Next analysis steps

- Do like now, wild world, "chacun pour soi"
  - Each user chaotically runs jobs on its favourite stream
    - 1% of the whole dataset
    - Still 180 GB per hour of data taking, or 300 TB per stream for a whole year
      - How frequently can this be done?
  - Is this conceivable?
    - How long would it take to run each job?
    - What if each user is doing that?
- Produce Ntuples (or whatever) centrally for (sub-)WGs
  - Back to the story of WG-productions
  - Requires organisation, coordination, preparation
    - But can we avoid that?
      - I don't think so
    - Seems to be a unique opportunity
  - Caveats
    - Can Ntuples be larger than the initial dataset, just because they are inefficient?
    - How frequently should one run on the whole sample?

User analysis job submission tool still needed

- Why?
  - Necessary to run on a sensible fraction of the datasets
  - Even 10% is a few TB, hence few 1000's jobs!
    - Can't run interactively… Direct batch submission is disappearing
      - Who has run jobs on HTCondor at CERN? Who is willing?
- What is the future of ganga?
  - Manpower situation is critical
  - Functionality
    - Users seem to always complain but in the end they manage… and don't help improving…
- Should we develop a more LHCb/Dirac/Gaudi-centric job submission tool?
  - Revise user requirements
  - Clarify the use cases
    - For example avoid copying from EOS to EOS if not needed, run on remote files (this is easy already now)
- Analysis data format
  - What are the requirements?
  - Use (py)ROOT, use GaudiPython / Bender

Conclusion

- Too many unknowns still
- Run2 is a unique opportunity to try and learn on how to deal with larger datasets in Run3
  - Not much progress in WG-production adoption since last workshops
  - It is a step to go, but there may be large benefits
- Analysing like now with 10 times more data will not be possible
- More precise numbers would greatly help
  - Don't I say that at every workshop?