HANDS ON
- submitting jobs on bash script

Paul Veen

ALICE

cea irfu

universit
PARIS-SACL

General idea and motivation

- We need more statistics…

- Let's assume we need around 100 million events (of biased simulations, i.e. HardQCD)
  - Impossible to do with your computer (you can try..)

- On Stoomboot we can submit batch jobs to the cluster

Structure

First simulate pythia job (hardQCD) with X number of events

Submit this Y amount of times

Result is Y groups of X number of events (‘batched’)

This batching is nice because we can use it to determine the uncertainties (with **subsampling** = repeating the analysis on the separate datasets and comparing the outputs)

Submit again the batched jobs, but now with the analysis script

Proceed to merging (and keep the batched outputs as well)

- I implemented the subsampling in the plotting macro (link) so that it can be done at run-time. It needs some more testing, however..

Condor cluster

https://kb.nikhef.nl/ct/Stoomboot_cluster.html

https://kb.nikhef.nl/ct/Batch_jobs.html

Uses condor cluster (see also https://htcondor.readthedocs.io/en/latest/users-manual/submitting-a-job.html)

Also important: storage: https://kb.nikhef.nl/ct/Storage_overview.html

sub.sh

```
executable              = test.sh
log                     = test.log
output                  = outfile.txt
error                   = errors.txt
## Can use "el7", "el8", or "el9" for UseOS or you can specify your own
## SingularityImage but an OS must be specified and in string quotations.
+UseOS                  = "el9"
## This job can run up to 4 hours. Can choose "express", "short", "medium", or "lo
+JobCategory            = "short"
queue
```

Exercise

1. Try submitting a simple test job that prints “I love quarks !” by using the cluster.
   - Make sure to use the short queue and be sure to submit only 1 job.. we don’t want to overload the system with our small test.
2. If everything works ok, try to submit the `qqbarcorrelations_status.cpp` script, or the one you have been using so far. Do it with ~1000 events and using the short queue. Think carefully about the directory to put this information. The size of the output files will increase by a lot when you simulate more events (`/data` is a good choice for now, but be careful when you really start simulating a lot of files)
3. Try using the `status_analysis.C` macro on the output files. This can be done interactively if the files are not too large, but should be done in another batch job if the outputs start getting bigger.

Note: be careful

Some tips from my own experience:

- Always submit 1 job first, then scale-up. It won't be the first time that a syntax error makes your 1000 jobs fail.
- In general, 1-100 jobs are ok, more can also be done for more statistics, but it's on the larger side. I would do that only after some checks and other analysis has already been done.
- The number of events to simulate can be picked in such a way that you stay under the short queue limit (4 hours). This is much faster than staying in a longer queue.
- Be careful with bash scripts. Never use * unless it's in a full path. Do checks if your directories and paths actually exist, otherwise exit. Bash scripts have the nice feature that they can go to your home directory and execute commands from there if the path you ask cannot be found/doesn't exist. It can have consequences..