WLCG HEP-SCORE Deployment Task Force

Meeting on 03 February 2021 at 15:00 h UTC (teleconference)

Notes

Indico event page: https://indico.cern.ch/event/992804/

Welcome, note-taking, notes from previous meeting

Helge M. welcomes Andrew Melo who will represent US Tier2 sites, particularly ATLAS ones. Minutes of the previous meeting are approved

CMS workload(s): status and plans (Tommaso Boccali)

Tommaso B. presents an update on CMS workloads. 3 workflows are proposed: GEN+SIM, DIGI with Premix, and RECO+MINI. NANO has a negligible weight, so not included. RECO should have a higher weight in the eventual mix. 8 threads/cores is the most reasonable number, and it is the one used in production. Number of events should be of the order of few hundreds: 100 events on 8 cores currently take ~40 minutes on Lxplus, ~55min on Power8 and ~140min on AARCH64. GPUs are good to go, but no sizeable offline GPU resources are expected in Run3. For a given resource, the same GPU workload (HLT in this case) can be run twice: first pass forcing a CPU configuration, and second pass allowing for GPU discovery.

Domenico G. comments that images are rather large, and Tommaso replies that unneeded data sets can be removed, and containers can probably be re-used. Domenico further asks to expand on combined workflows; Tommaso explains that GEN+SIM is a combined number, where generation takes a much smaller fraction, while MINI is an additional module in the end of RECO, not being a significant contribution, but bringing it closer to the production scenario. Domenico also comments on ARM support: several HPC centers offer it, so a comparison of different benchmarks on different architectures is much welcomed.

Gonzalo M. asks whether all cores can be used (not just 8), and Tommaso replies that anything between 3 and 15 is good, but 8 is most common for CMS jobs. Helge M. adds that there's a study showing that a multiple of 4 is optimal.

Andrea V. asks whether 2x4core vs 1x8core jobs performance was compared, to which Tommaso replies that 2x4 performs much worse due to memory limitations. Andrea further asks whether generation always multithreaded, and Tommaso explains that two ways can be used, single- and multi-threaded, and even single-threaded is set to fill all available cores. Andrea also suggests to split GEN and SIM, because adding NNLO, many jets etc to GEN can have a significant impact on performance. He also notes that some events can take much longer that others, which can skew the results if the sample is not too big, so for benchmarking it might be useful to (re)run same events, but Tommaso disagrees.

Jeff T. warns that 8 cores may not be the most common number for every site, as it is 4 at NIKHEF. He further expects that GEN+SIM should have a higher weight than RECO, but Tommaso argues that RECO has a comparable weight in CMS.

Stefano P. agrees that 8 cores is optimal, and has a point on GPUs: HLT workflow may not be a suitable benchmark to be used for procurements. Tommaso agrees that this was more of a technical test, and GPUs are not

expected in Tier1 and Tier2 in Run3 anyway.

Helge M. adds that while finding workloads suitable for GPUs (and other architectures) is good, it is not a current priority. He also thinks that we should focus on workloads themselves and decide on their relative weights later, and supports splitting generation from simulation.

Manfred A. comments on 4 vs 8 threads, noting that there are machines with 12 or 20 cores, which is a not a multiple of 8, thus all cores can not be fully occupied by benchmark tests, and 4 is a preferable number.

LHCb workload(s): status and plans (Andrea Valassi)

Andrea V. presents the LHCb feedback on the HEP-SCORE21 benchmark. Simulation is the major consumer of LHCb resources, by far, SIM >> GEN, so most of CPU power is used by Geant4. Current software and compiler versions are rather old, and the benchmark is based on 5 reproducible events, taking around 20 min on a regular CPU. A more recent Gauss workload, still with a single thread, is planned to be used. In future, multi-threaded workload will be used (Gaussino), with a much better memory per core usage, and multi-threaded Geant4. It is not yet production-ready though, container will be eventually provided to the WG, but not in time for HEP-SCORE21. Reconstruction workloads may also become relevant in future: HLT1 data reconstruction will be done online on GPUs, but it is not really a Grid use case. Fast benchmarks are of a particular interest: LHCb may want to gauge individual node performance on the fly, to optimise job scheduling, and needs a reliable mechanism (like Machine-Job-Features, MJF) to know individual node performance. A fast benchmark should thus be computed in real time (like DB12 does). This won't be needed if benchmark value will be available per node.

Helge M. comments that the time scale is not fixed yet, and the WG wants to make sure that the experiments come with realistic workflows, such that a benchmark is good for next 5 years or so. Ideally the workloads should be complete before people leave for summer vacations. Andrea hopes that workflows will be ready, though the Gaussino one is less obvious. Helge further asks whether simulation is dominant for Tier0 as well, which Andrea confirms.

Jeff T. comments that Machine-Job-Features has numerous shortcomings, Andrea agrees, but such is the LHCb feedback.

Domenico G. also agrees that running conditions on nodes vary, and MJF may indeed not be representative of a node performance in general. Andrea agrees that another benchmark may need to be discussed, and input from other experiment is needed. Domenico notes that the two benchmarks should not be mixed, as for procurements we need a stable and reproducible benchmark that reflects CPU architecture.

Domenico G. also notes that most experiments don't mention analysis workload, and asks whether such should be only introduced for ALICE? Andrea confirms that for LHCb it is a really small fraction. Walter L. replies that ATLAS may think of including some derivations, but analysis itself is too diverse to quantify. Tomasso B. says that for CMS analysis takes a higher proportion of the resources, but he still sees no need to include it as a benchmark workload, partially because it will be more I/O- than CPU-bound. Stefano P. confirms that for ALICE analysis is indeed very significant, and agrees that it is very I/O-intensive. Jeff T. recalls that gunzipping a ROOT file used to be the heaviest analysis CPU load, but Stefano P. says that ALICE is is not doing it, and Walter L. adds that ATLAS uses a different compression.

Lastly, Andrea V. asks Tommaso B. about vectorisation: how much CMS relies on it, and whether it should be benchmarked; Tommaso can't estimate what fraction is vectorised, and says that nothing has changed recently anyway, so for all benchmarking purposes AVX2 should still be assumed.

Any other business

Helge M. announces agenda proposal for the next meeting: presentation of non-LHC workloads, such as Belle II, DUNE, and gravitational wave experimens. Randy S. agrees to present Belle II. Helge M. will contact the other experiments.

Next meeting

Wed 17 February at 15:00 h UTC (16:00 h in Geneva)

Annex: Attendance

Present:

Manfred Alef (KIT)
Tommaso Boccali (INFN Pisa)
Simone Campana (CERN)
Ian Collier (STFC)
Peter Couvares (Caltech)
Domenico Giordano (CERN)
Michel Jouvin (IJCLab)
Walter Lampl (U Arizona)
Andrew McNab (U Manchester)
Helge Meinhard (CERN, chair)
Andrew Melo (Vanderbilt U)
Gonzalo Merino (PIC)
Bernd Panzer-Steindel (CERN)
Stefano Piano (INFN Trieste)
Fazhi Qi (IHEP)
Oxana Smirnova (U Lund, notes)
Randall Sobie (U Victoria)
Jeff Templon (Nikhef)
Andrea Valassi (CERN)
Tony Wong (BNL)

Apologies: