WLCG HEP-SCORE Deployment Task Force

Meeting on 17 February 2021 at 15:00 h UTC (teleconference)

Notes

Indico event page: https://indico.cern.ch/event/992814/

Welcome, note-taking, notes from previous meeting

The minutes from the previous meeting are approved. Thanks to Oxana Smirnova for preparing them.

Belle 2 workload(s): status and plans (Randy Sobie)

Randy presents the slides attached to the agenda. Randy is a member of both Belle2 and ATLAS. The following additional points mentioned by Randy are noted during his presentation.

- Belle2 is installed at an asymmetric, low energy, high intensity, e+e- collider. It is similar to Belle and Babar but aims to collect a factor 50 higher integrated luminosity.

- The events have low multiplicity. There is no pileup. The primary background are beam interactions, this needs to be overlayed on MC events.

- There are many small sites contributing resources.

- Belle2 uses DIRAC. The DB12 benchmark is used for job scheduling in DIRAC, but also in CPU usage plots.

- A HEP-benchmark container has been prepared using B0B0bar production, a typical event. The software is single core, so the benchmark uses multiple copies of a single core job.

The following questions and comments are noted after the presentation.

- Question/Helge: what is the memory consumption for the workload you have packaged? Answer/Randie: we reserve 2 GB, but typically they are not needed and we only use 1.5 GB.

- Q/Helge: how do you make sure the workload is reproducible? A: we fix the random number seeds.

- Q/Tommaso: I understand that now the accelerator is running at low luminosity, but will you have more complex events with the final luminosity? A: no, we have no pileup, the events will not be more complex; we expect that initially the beam background will be higher, but eventually this will decrease as it gets better understood.

- Q/Tommaso: what is the total amount of disk/cpu that you request for a typical year? A: we are requesting approximately 15k-20k cores per year now. Comment/Tommaso: thanks, this is essentially 10% of CMS.
- Q/Andrea: in general would you say that you have a computing resource problem? A: not really; if anything, we do not use all resources available, we often have some idle periods.

- Q/Andrea: slide 4 mentions that “cvmfs is not required” in the workload, can you clarify? A: the workload container is built like all other HEP workloads, using cvmfs shrinkwrap to extract the relevant files from our software installation on cvmfs; once the container is prepared, at that point it is self contained and we no longer need access to cvmfs.

- Q: Helge: on slide 4 you mention that SIM+DIGI is 37% while RECO is 61%, while in the LHC experiments SIM is often a dominant consumer of CPU, can you explain this? A: I cannot explain now why this is so (but can

ask colleagues for a detailed explanation if needed); in any case I confirm that these are our current numbers, so RECO is almost a factor 2 heavier than SIM+DIGI.
- Comment/Helge: I encourage you to benchmark independently SIM and RECO, so that we can have the two figures separately and gain more insight. A: will do, it should be easy to do this.

- Comment/Helge: this looks in good shape! And unlike the LHC experiments that are concentrating on getting the new software for Run3, your software seems quite stable. A: yes it is pretty stable.

- Q/Tommaso: to ask resources from your funding agencies, do you use HS06 or numbers of cores? A: we use HS06 (for the resource board that is the equivalent of the RRB).

GW experiments' workload(s): status and plans (Josh Willis)

Josh presents the slides attached to the agenda, describing computing workloads in LIGO. The following additional points mentioned by Josh are noted during his presentation.

- Gravitational wave experiments are different from HEP. The data from each observatory is a time series. Unlike in astronomy, the instruments are not directional.

- The data is analysed in many different ways, including: searches (which look at all the data and can either use models or be model independent); detector characterization and data quality; parameter estimation (which is a pretty big consumer because GW experiments have been lucky with many observations).

- Data analysis can be parallelised along many dimensions, so it fits well on HTCs and has no need for HPCs. Previously the analyses were running on private clusters, now they are moving to OSG using Condor. Condor is also used now on own clusters and on computing resources contributed by VIRGO.

- A typical example of a modeled transient search is CBC, Compact Binary Coalescence (two objects, e.g. two black holes or two neutron stars). A relevant benchmark score is the number of "templates" computed per core. It is not templates per core per unit time, because time in the denominator cancels out with the duration of the time window you are analysing in the whole time series, in the numerator. Match filtering itself is not complex; removing detector noise (i.e. signal/background discrimination) is the more difficult area, where the software work is concentrating.

- One example of a pipeline is PyCBC (developed by Josh). This does the analysis in the frequency domain (the problem is that you do not know the moment in time when a signal arrived). One technique is FFTW ("Fastest Fourier Transform in the West"): this is fast but it needs the creation and sharing of 'plans' specific to each architecture. Sometimes it takes days to make one such plan. The Intel MKL library and CUDA do not have this issue. The highest CPU usage is in the match filtering operation. The workload is already containerized in singularity. Josh has not yet started to look at porting this to a HEP workload benchmark container, but it should be relatively easy to find a typical example of a job that lasts approximately one hour (or maybe less if needed).

- Parameter estimation is another important workload (for instance, a fit of the mass of the coalescing object). It is based on Markov Chain Monte Carlo (MCMC). The CPU usage here is dominated by the MC generation of the signal of the gravitational waveform. The software is not well vectorized now.

The following questions and comments are noted after the presentation.

- Question/Walter: can you confirm that you have no RECO (data processing shared by many analyses) and that you essentially go straight from data taking to analysis? Answer/Josh: there is an internal library that is shared by several analyses, but many people have moved away from it. Also, there is a calibration process that produces "conditioned string" data, these are produced by the labs and then archived, and the analyses run downstream of that: it is a critical step, but takes relatively few computing resources.

- Comment/Andrea: interesting that, in parameter estimation, generation is the largest CPU consumer.

- Q/Andrea: how much data would you need to encapsulate for a benchmark lasting 1 hour? A: at most a few 5-10 GBs, but probably hundreds of MB would be enough, and the workload would probably run with 2GB of RAM.

- Q/Andrew: what is the granularity of the architecture-specific plan that you need for FFTW, e.g. do you need one for Haswell and one for Broadwell? A: it is very granular, it is a bit like the ATLAS math library, the plan can enable different instruction sets and also determines if we run single threaded or multi threaded (e.g. on Sandy Bridge using 8 cores together is faster than running separately on 8 individual cores).
- Comment/Domenico: this fine graining is very interesting, it is different from what we have done so far in the WG, where we can move from one CPU model to another without reoptimizing for each architecture. Q/Domenico: how would you see this addressed in a benchmark container? A: we should set aside FFTW as an option, and instead use MKL (which essentially just "knows" that it is on a given architecture and uses the optimized version); what would be interesting for LIGO is to check if the resulting benchmark moves in the same way as other HEP becnhmarks from one machine to another, i.e. if they are correlated or not.

- Q/Domenico: you mentioned as another option that you could adapt the configuration: is this something that you can read in from an existing docker image, or would you need to rebuild a different container? A: the information is stored in internal format in a file, so one should build with the option to use largest range of options (e.g. avx512), and then choose at runtime what is going to be used.

- Q/Gonzalo: what is the ratio of GPUs to CPUs currently, and how will that change? A: we mostly use CPUs, while for instance at Caltech we use 600 consumer GPUs with 15-year old CPUs that we could not use to run the CPU code anyway; we do not have many (or any!) dedicated software engineers, it is scientists writing the code.

- Comment/Helge: if we need to add external libraries in containers, then we need to look at license, this must be understood carefully. A: FFTW is open source, but MKL is not, so indeed we should check that. C/Domenico: from CERN KT, I understood that as long as we yum install a library, but do not copy the source code, this is fine (it is not very different from distribution over cvmfs), but someone should cross-check.

Q/Helge: how many cores do you have as dedicated resources, beyond the lab's own resources that are used for the initial processing (like our HLTs)? A: about 20k cores or a bit more (the largest site, Caltech, has 11k, then there is a large site in Milwaukee for instance, and other resources at OSG and from VIRGO).

Q/Helge: do I understand correctly from your use of Fourier transforms that a lot of your processing is floating point? A: yes, now we are actually memory bound, but eventually we will be FP bound.

Comment/Helge: I encourage you to include some of these workloads as part of our “matrix”, so that we can see how they behave with respect to particle physics (i.e. if they are correlated or not to other HEP workloads).

Any other business

Comment/Helge: at the next meeting on March 3, there will be a presentation by Fazhi Qi about JUNO and other IHEP experiments, and a presentation by Ian Collier on accounting.

Q/Domenico: one of the next Task Force meetings is scheduled during the HEPIX workshop and there may be some overlap, is this meeting confirmed? A/Helge: I discussed this with Tony Wong who is the HEPIX chair, I will follow up with him to investigate several options.

Next meeting

Wed 03 March at 15:00 h UTC (16:00 h in Geneva)

Annex: Attendance

Present:

Manfred Alef (KIT)
Tommaso Boccali (INFN Pisa)
Domenico Giordano (CERN)
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)
Randall Sobie (U Victoria)
Andrea Valassi (CERN, notes)
Tony Wong (BNL)

Apologies:

Ian Collier (STFC)
Michel Jouvin (IJCLab)
Fazhi Qi (IHEP)
Oxana Smirnova (U Lund)
Jeff Templon (Nikhef)