High Performance interactive machines for CMS analysis

Josh Bendavid (EP-CMG)
with input from
M. Cipriani, M. Dünser, K. Long (EP-CMG)

June 24, 2021
IT R&D Advisory Group
Reminder: Discussion about possible high performance interactive machine for CMG group

Motivated by some of the needs of the W mass measurement, a high profile Run 2 physics deliverable

- $\sim 800M$ data + 1.5B signal MC events with little scope for skimming
- Complex with extreme precision requirements for both theory and experiment $\rightarrow$ many iterations needed
- For this type of analysis HL-LHC is now

Previous presentation in IT R&D Advisory meeting in February: https://indico.cern.ch/event/1005976/contributions/4243748/attachments/2195605/3712149/cmgInteractive-Feb24-2021.pdf
Introduction

So far have been mainly working on the machine provided to us in January (through OpenStack bare metal):

- Dual Xeon with total of 32 physical cores (64 threads)
- 192GB memory
- 16x1.92TB SATA SSDs in Raid0 (8Gbytes/sec theoretical throughput, 30.72 TB total capacity)
- 25gbps ethernet

Previously have shown a number of technical benchmarks

More recently focused on the full analysis framework/analysis
High Level Analysis Technical Overview

- **High level analysis step:** Compact CMS analysis data-format (NANOAOD, \(O(1\text{kB/event})\)) → histograms (for input to likelihood fit, diagnostics, and presentation)
- Older TTree-based analysis framework has been fully ported to RDataFrame (modern, parallelized, etc analysis library in ROOT)
- Significant technical optimizations have been made in close coordination with ROOT team to improve thread scaling (see presentation in ROOT Parallelism, Performance... meeting in March: https://indico.cern.ch/event/1018696/)
- Data volume: \(\sim 800\text{M data events}, 1.4\text{B signal, 1B background, 4TB total for one version}\)
- Currently filling \(18 \times 2\text{D and 136} \times 3\text{D histograms including all systematic variations and physics processes}\)
- Additional histogram axes are used for optimized treatment of systematic variations
- Full W analysis currently takes \(\sim 2\) hours interactively but requires two event loops
- “W-like” analysis of Z events used for validation etc is a bit smaller → \(\sim 40\) minutes runtime
Aside from pure technical gains expected from more CPU cores given previous scaling tests, a number of other optimizations are in progress.

- Eliminate second event loop and further consolidate histograms to a few 4D or 5D histograms
- Evaluating Boost histograms which our colleagues in Pisa have used extensively with RDataFrame (and I have already contributed some interface improvements to ROOT to make this easier)
- WIP Improvements to ROOT to allow native ND histogram support
Aside from pure technical gains expected from more CPU cores given previous scaling tests, a number of other optimizations are in progress

- Assess and eliminate impact of dynamic memory allocation inside event loop
  - Migrate to statically sized arrays to compute systematics variations where possible
  - WIP optimization by ROOT team to RDataFrame vector classes (small static buffer)
  - Possible further optimization in collaboration with ROOT team to vector classes (optimized memory allocation and/or expression templates)

- Carefully testing WIP developments from ROOT team to improve code optimization level for JIT’ed code, which becomes more important with complex multi-dimensional histogram filling and makes the use of statically sized arrays feasible in more places
Aside from high level analysis, a number of auxiliary measurements of corrections and calibration constants

Muon momentum scale calibrations are based on a high granularity correction for B-field, material and alignment residuals, approaching complexity of full tracker alignment

In large debugging version which was being used for R&D, \( \sim 130k \) parameters, 7TB of flat trees

Taking the same workflow and running with tight selection of tracks on top of the large trees \( \rightarrow \) reduced CPU load \( \rightarrow \) workflow becomes largely IO limited
<table>
<thead>
<tr>
<th>CPU</th>
<th>Storage</th>
<th>Avg. Rate (GBytes/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x Xeon (32C/64T)</td>
<td>eos/xrootd (eoscms) (25gbps)</td>
<td>1.64</td>
</tr>
<tr>
<td>2 x Xeon (32C/64T)</td>
<td>eos/xrootd (test inst.) (25gbps)</td>
<td>2.62</td>
</tr>
<tr>
<td>2 x Xeon (32C/64T)</td>
<td>CephFS HDD (25gbps)</td>
<td>2.60</td>
</tr>
<tr>
<td>2 x Xeon (32C/64T)</td>
<td>CephFS SSD (25gbps)</td>
<td>2.64</td>
</tr>
<tr>
<td>2 x Xeon (32C/64T)</td>
<td>Local SSD (16xSATA)</td>
<td>5.21</td>
</tr>
</tbody>
</table>

- Need to set e.g. XRD_PARALLELEVTL=16 to get good eos performance
- EOS+xrootd standard production instance not quite scaling up to network limits (possible xrootd client/ROOT bottlenecks?)
- Extremely good performance of EOS test instance, and CephFS (CentOS 8 kernel client), approaching limits of ethernet connection
- Reach 5.2GBytes/sec from local SSDs, approaching limits of disk array

thanks to IT-ST group for help setting up some of these tests
Next Steps

- A second ~ identical machine has been provided
- For $m_W$ analysis current focus is on optimization and/or scaling up to more cores in a single process → second machine will be used for other analysis and technical activities in the EP-CMG group
  - Several other NANOAOD-based physics analyses
  - Commissioning of HLT Scouting analysis for Run 3
  - Large maximum likelihood fits for statistical analysis (Higgs combination, etc.)
- To facilitate monitoring and ease some practical considerations (e.g. condor submission), in the process of setting up an appropriate puppet configuration for this one (plus future machines)
- Existing machine can be migrated to puppet when appropriate
- A future larger machine will likely be shared across more people/activities in the CMG group
Other Issues

- Higher performance solution for home directories desired than afs
  - Currently using local home directories
  - Some intermediate solution probably makes sense (CephFS?)
- How to best deal with mixed workflows which include a GPU component?
  - Some of our likelihood fitting workflows can already be GPU accelerated (tensorflow/jax implementations)
  - Requirements in this specific case are a bit awkward (double precision, modest compute requirements, but lots of VRAM)
  - Just running on CPU is ok for the moment
- Bigger computers on users desks in some cases?
  - Today possible to buy a 25gbps residential connection in Winterthur and a few other places in Switzerland
Next Steps

- Anticipate at least a factor of 2 speedup to the high level analysis from near-future software/workflow optimization, with an additional 3-4 possible with a larger machine.
- Having achieved excellent/sufficient performance on a single box, feasibility of disaggregation of storage and/or CPU to be assessed:
  - CephFS works out of the box with no additional tuning up to the wire limit in all tests so far (even HDD-backed behaves extremely well) → is this feasible for 10-100 TB scale high performance analysis? Will we overload the system with a few 100gbps clients?
  - Can the same level of performance be achieved with optimized eos instances?
  - Tests ongoing using XCache in front of the existing eoscms instance
  - Low-latency distribution of tasks over several hundred cores with Spark and/or Dask?
- Both performance and efficiency/convenience of resource use to be considered.
- Can the “bulk” resources even eventually be adapted to this use case and performance with the right software + modest use of SSD caches and/or storage instances?
If specialized hardware is desirable beyond this R&D phase, “minimal” possibility is to count this against CMG’s dedicated condor and eos allocations (currently 8kHS06 + 2PB of disk) in the absence of something more general.