USCMS Analysis Facility Discussion

Matteo Cremonesi - June 13, 2024
What is an “Analysis Facility”

- Common, jointly developed, software-defined infrastructure
- Tuned to improve ease and turnaround time of end-user analysis
- Deployed on hardware sufficiently optimized to realize those improvements
What is an “Analysis Facility”

- Ultimate goal: deliver high performance analysis capabilities to users
  - 200 Gbps + 200 MHz / user
- Mostly a **software concept**
  - Collection of services that make it easier and faster to do analysis
- For hardware, define “minimum system requirements”
  - One of this year’s goals
The USCMS R&D Analysis Facility Effort
This Year Goals and Milestones

- **Benchmarking** emerged as the highest priority
  - Summary [document](#)

- **Recommendation:** run Analysis Grand Challenge
  - Test capabilities to run real analysis
  - Helps comparing AFs

- **Moving forward:** consider 200 Gbps challenge
  - Helps identify bottlenecks
    - i.e. compression of inputs has large effect
  - Can inform on hardware specs
What is AGC?

• “The AGC is developing a series of increasingly **realistic** benchmarks for workflows designed to analyze the data collected at the HL-LHC”
• IRIS-HEP-led effort
  => analysis workflow designed using the IRIS-HEP stack of tools
• Analysis workflow includes:
  - columnar data extraction from large datasets
  - data processing (event filtering, construction of observables, evaluation of systematic uncertainties) into histograms
  - statistical model construction and statistical inference
  - data visualization

  **well aligned with the data analysis workflow supported by the USCMS AFs**
### AGC and Scaling

**getting ready for HL-LHC**

<table>
<thead>
<tr>
<th>Timeline</th>
<th>Fraction of HL-LHC dataset processed in 1h</th>
</tr>
</thead>
<tbody>
<tr>
<td>2025</td>
<td>20% (40 TB)</td>
</tr>
<tr>
<td>2026</td>
<td>50% (100 TB)</td>
</tr>
<tr>
<td>2027</td>
<td>75% (150 TB)</td>
</tr>
<tr>
<td>2028</td>
<td>100% (200 TB)</td>
</tr>
</tbody>
</table>

With 30 simultaneous users
AGC Metrics

• Quantitative metrics
  - Data volume processed (per time and core)
  - Event processing rate per core
  - Scheduling efficiency

• Additionally
  - data pipeline comparisons to test speedups provided by caching data
  - measurements of the effect of concurrent users
  - evaluations of the user experience
• **Nebraska** (coffea-casa) provides the backbone to AGC
  • Large contribution to AGC from *Fermilab* EAF
    - UW student serving as the contact person for AGC at EAF
    - EAF client implemented in the AGC settings
200 Gbps Challenge

- 8-week exercise
- [https://github.com/iris-hep/idap-200gbps](https://github.com/iris-hep/idap-200gbps)
  - Used to run the challenge at Nebraska
- Simplified analysis setup compared to what is done in AGC
  => project focused on data throughput