IRIS-HEP retreat

Current plans for AGC (30’)

Alexander Held (University of Wisconsin–Madison)
Oksana Shadura (University Nebraska–Lincoln)

Introduction

• We will summarize the AGC plans laid out in the strategic plan

• Our views of the best way forward for the AGC project have evolved in parts
  • will outline those here to move towards an updated plan
  • discussion session at 13:30 today
  • important: focus on realistic high-impact goals given available personpower
Challenges motivating the project
Scaling & turnaround

- **Scaling to HL-LHC data volumes** with available computing resources
  - Need for new methods for efficient data scaling, caching at AFs to handle more data-intensive analysis pipeline => DOMA, SSL

- **Analysis turnaround time**
  - Reaching interactive analysis turnaround times requires efficient analysis facility (AF) usage => SSL, AS, DOMA, user
UX & sustainability

- **User experience (UX)** for complex analyses: *increase scientific reach* of result => AS
  - User improvement experiences allow physicists to *focus on the physics*
  - Need *expanded Machine Learning (ML) tooling* with good user experience + performance
  - Leverage *ML technology* to automatically optimize analysis sensitivity
  
- **Sustainability** => ?
  - Limited person power to develop & maintain full stack => rely on industry solutions & external developments (e.g. tokens)
  - Limited of *analysis reproducibility & reusability*
  - Need for *central gathering point* for community to discuss & develop analysis approaches
Work plan from strategic report
Strategic report plan

• Proposed: expand AGC with **new flagship analyses** (high complexity, high volume)

• ATLAS-/CMS-specific analyses were meant to help **bridge the gap to production**
  • e.g. nanoAOD & PHYSLITE formats, systematic uncertainty handling
  • **different focus per experiment**: on-the-fly systematics in ATLAS, ServiceX & column joining in CMS

• Designing new analyses is a **significant amount of effort** that requires experienced physicist personpower

• We believe that the **relevant R&D can efficiently happen within the experiments**, with IRIS-HEP members participating and interfacing -&gt; **do not develop additional analyses within AGC**
  • some relevant details are too specific for a broader inter-experimental forum like AGC
  • instead: maximize impact by focusing on dedicated areas
What we already have right now

- **We have:**
  - Open Data-based analysis of modest complexity capturing all generic workflow aspects
  - Setup with a lot of configuration options to emulate different types of analyses (including different processing pipelines)
  - Many combinations of configuration settings need to be benchmarked & understood
Proposed evolution and new focus of AGC
Focus areas: exploiting AGC setup (1)

• **Benchmarking with existing setup** *(as already mentioned in strategic plan)*
  - we can use the AGC setup as a tool to study facility, library and implementation performance (line profiling, comparing measurements to hardware, …)
  - facility improvements (including ML/MLops), important: stability of distributed execution
  - collaborating with DGC

• **Deliverables**
  - performance reports at internal meetings, workshops, conferences
  - evolve existing AGC task as needed to capture new functionality
    - estimated 1–2 years of UX & facility improvement work possible from lessons learnt with AGC
Focus areas: exploiting AGC setup (2)

- **Analysis preservation and reinterpretation** *(as already mentioned in strategic plan)*
  - close collaboration with REANA team
  - demonstrate preserved AGC analysis
  - propose plan for relevant services: do we assume those still exist? package them up too?

- **Deliverables:** AGC running in REANA, reinterpretation example (?)
Focus areas: autodiff

- **Gradient-based analysis optimization demonstrator** *(as already mentioned in strategic plan)*
  - completely new standalone analysis example
  - connects many projects & people: AGC ideal home for this type of effort

- **Deliverable**: demonstrator project showcased in public meeting

**Today**: information flows forward through pipeline

**Required**: gradient information passed backwards through pipeline

**Automatic analysis optimization**

*figure source: arXiv:1805.04829 [cs.AI]*
Focus areas: support & interfacing

• **Support** for experiment-internal demonstrator projects
  • interface, forum to discuss inter-experimental aspects
  • column joining workflow

• **Community engagement / reaching new audiences**
  • bridge gaps to new user groups in experiments

• **Deliverable: ?**
An idea of a timeline

- **Year 1**
  - 25% faster analysis via improvements following benchmarking, stable execution at scale

- **Year 2**
  - partially complete fully differentiable analysis example

- **Year 3**
  - ?

- **Year 4**
  - showcase substantially complete fully differentiable analysis example

- **Year 5**
  - 1h turnaround time for analysis at HL-LHC scale (lots of variables to be determined here…)
Backup
Overall view on AGC timeline

- Benchmark analysis on dataset 20% HL-LHC scale to be completed in 1 hour
- Benchmark analysis on dataset 100% HL-LHC scale completed in 1 hour
- Execute ATLAS AGC analysis
- Execute fully differentiable analysis
- Execute CMS AGC analysis with column extraction feature
- ATLAS and CMS Coffea-casa facilities are ready to be used in production

Year 1 Year 2 Year 3 Year 4 Year 5

HL-LHC
Metrics and targets

- Number of deployed analysis facilities in production operation: [at least 2]
- Number of AS components fully supporting distributed analysis: [at least 5]
- Fraction of components fully preserved in AGC pipeline: [100%]
- Fraction of benchmark analysis at HL-LHC scale executed in 1 hour: [20% / 50% / 100%]
  - Compare efficiency through time & data rate metric
- Number of AS tools that support integration with automatic optimization: [at least 3]
  - Tracks progress towards fully-differentiable analyses, which the institute believes is a powerful and critical tool for HL-LHC analysis
Timeline: year 1

- **Setting up** facilities, services, and analysis task for the next generation of the AGC

<table>
<thead>
<tr>
<th>Department</th>
<th>Task Description</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOMA</td>
<td>ATLAS and CMS Coffea-casa facilities are ready to be used in production</td>
<td>Year 1</td>
</tr>
<tr>
<td>SSL / DOMA</td>
<td>ServiceX deployed inside Fabric at CERN</td>
<td>Year 1</td>
</tr>
<tr>
<td>AS</td>
<td>All components of AS pipeline are fully supporting distributed analysis</td>
<td>Year 1</td>
</tr>
<tr>
<td>AS</td>
<td>Define analysis tasks for the top quark mass and di-Higgs measurement, create implementations</td>
<td>Year 1</td>
</tr>
<tr>
<td>AS</td>
<td>New version of AGC analysis with incorporated ML techniques</td>
<td>Year 1</td>
</tr>
</tbody>
</table>
### Timeline: year 2

- **Checkpoint for performance** and throughput & **functionality** demonstration

<table>
<thead>
<tr>
<th>AS</th>
<th>Benchmark analysis on dataset 20% HL-LHC scale to be completed in 1 hour</th>
<th>Year 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOMA</td>
<td>Demonstrate analysis running at 200 Gbps as a part of AGC</td>
<td>Year 2</td>
</tr>
<tr>
<td>AS</td>
<td>Demonstration of running full analysis able to use statistical models defined in unified HS3 serialization format</td>
<td>Year 2</td>
</tr>
<tr>
<td>AS</td>
<td>All core components of Analysis System pipeline support integration of differentiable operations and passing of gradients</td>
<td>Year 2</td>
</tr>
<tr>
<td>AS</td>
<td>Demonstrate AOD extraction prototype</td>
<td>Year 2</td>
</tr>
</tbody>
</table>
Timeline: years 3–5

- **New functionality**: column joining, reinterpretation, differentiable analysis
- **Scaling** to full HL-LHC requirements

<table>
<thead>
<tr>
<th>Year</th>
<th>Project</th>
<th>Task Description</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 3</td>
<td>AS</td>
<td>Demonstrate AOD extraction with column joining workflow</td>
<td>Year 3</td>
</tr>
<tr>
<td>Year 3</td>
<td>DOMA</td>
<td>Demonstration of an AGC analysis used in reinterpretation platform</td>
<td>Year 3</td>
</tr>
<tr>
<td>Year 4</td>
<td>AS</td>
<td>Demonstrate fully differentiable analysis</td>
<td>Year 4</td>
</tr>
<tr>
<td>Year 5</td>
<td>DOMA</td>
<td>Benchmark analysis on dataset 100% HL-HL scale completed in 1 hour</td>
<td>Year 5</td>
</tr>
</tbody>
</table>
Primary risks for AGC

Analyzers not adopting proposed workflows
- **Impact**: Activity not useful to community
- **Probability**: Low to Medium
- **Mitigation**: Deep investment in community engagement

Unable to achieve intended throughput rates
- **Impact**: Slow analysis turnaround time for analyzers
- **Probability**: Low to Medium
- **Mitigation**: Give guidance on patterns and workflows to avoid

Facility evolution diverging from patterns used in AGC implementations
- **Impact**: Need to adopt technologies
- **Probability**: Low
- **Mitigation**: Invest in partnership with AFs and plan for technology nimbleness

Unable to put together a substantially complete differentiable pipeline
- **Impact**: Cannot study and benefit from gradient-based optimization, physics reach limited
- **Probability**: Low to Medium
- **Mitigation**: Rigorous integration testing between all components