200Gbps Challenge – Next Steps Discussion

IRIS-HEP SB21

Brian Bockelman

MORGRIDGE
INSTITUTE FOR RESEARCH

What were the original goals?

- Use the “Grand Challenge” format to show progress toward HL-LHC analysis.
  - We define the “Grand Challenges” to be a series of incremental, increasingly-realistic exercises toward a common goal.
  - What makes them so useful?
    - Focuses effort
    - Helps the community find “common truths”.
    - Can include both scale and technology readiness.
- The “200Gbps Challenge” was an **integrative**, **deadline-driven**, and **quantitative** exercise showing progress toward HL-LHC scale.
  - 200Gbps was selected as a target as, arguably, it’s 25% of HL-LHC scale.

What’s next?

Dimensions to consider

- "Line speed of IRIS-HEP Data Analysis Pipeline (IDAP)" as was done for the 200Gbps challenge is a single dimension.
  - What are other relevant dimensions for HL-LHC?
  - Ideas:
    - More community coordination (additional AFs, additional services/pipelines).
    - Multiple concurrent workloads running.
- Data Challenge parallels: DC24's goal was 25%; DC26's goal should be 50%. What is 50% of analysis?
  - And what's a reasonable milestone to consider for winter 2024 or 2025?
  - Let's avoid running DC26 and future Analysis Challenges in the same quarter…
- Not everything needs to be a "grand challenge": projects are expected to independently progress, fix bugs, test their standalone scaling.
  - The 200Gbps work showed shortcomings in some of the Python tooling, need for new hardware deployments in testbeds. These should be done regardless.

Some ideas for discussion

- My idea: Smaller IRIS-HEP-driven “complexity challenge” in winter 2024 (**post-CHEP!**); community AF challenge in spring/summer 2025.
- IDAP “Complexity challenge”:
  - Stay with the 200Gbps scale target, same pipelines.
    - Target far enough out to provide testbeds time to acquire hardware to address bottlenecks discovered in the 200Gbps activity.
  - Chance to show issues identified in 200Gbps challenge have been fixed.
    - E.g., use Coffea 2024, ServiceX Client V3, PHYSLITE branch reading fixes.
  - **Increased realism**: Multiple pipeline instances running simultaneously; add in more realistic physics to the pipeline (histogram aggregation, cuts).
  - **Schedule**: Are there any “good months” between November 2024 and March 2025 for such an exercise?

Some ideas for discussion, part 2

- Community AF challenge, spring/summer 2025 (later?):
  - **Idea**: Execute a common HL-LHC (proxy) analysis using multiple pipelines and facilities.
    - M pipelines and N facilities does not imply M*N workloads – encourage groups to hew to their local priorities and interests.
    - IDAP team would build on top of the “complexity challenge”, perhaps modest scale increase & better histogram management (or add in systematics).
  - **Much harder!** We need to identify community leaders who can spearhead the organization.
    - Consider DC24 as a blueprint. Coordination is not a full-time job but probably several FTE-months over the course of a year.
    - External deadlines help! Is there a clear community deadline/workshop?
    - DC24 also had several pre-planning workshops that benefitted from only having to “tweak” goals set in 2022.
      - This would be one way to precipitate the idea of having an “What is an HL-LHC analysis?” workshop.