

Brainstorming for Order in which to proceed

**Frank Würthwein
Professor of Physics
UCSD/SDSC**

Proposal for Discussion

- With these slides, we'd like to propose an order for how to get started.
- Basic idea is to try and devote a set of consecutive meetings to a question, and invite presentations around that question.
- Move forward one question at a time.
- Try to plan ahead as much as possible in order to give people a chance to prepare to address the questions.

Questions (Round 1)

Q1: What deployments (production, prototypes, testing) do exist today ?

- Solicit presentations on these deployment and their experience.
- Q&A on what has already or can in the future be learned from these deployments.

Q2: What effort exists in the community in this area ?

- Solicit presentations on projects that have started, their goals, effort, and timelines. I.e. work through the list in google doc.

Q3: What are the cost drivers for ATLAS and CMS for storage ?

- Have the two large experiments provide information on what they think the cost drivers are for storage, and how the various things we have seen in response to Q1 and Q2 may be useful to reduce costs. I.e. what are the problems from ATLAS & CMS' perspective?
- Are there challenges other than cost ?

- After the initial round of educating each other, we switch gear and ask more detailed “technical” and/or “quantitative” questions, e.g.:
 - What is known about the Application IO patterns ?
 - How does this affect the ability to do remote reads ?
 - To what extent do measurements exist today?
 - Are more measurements needed?
 - Is caching needed/desirable to hide latencies due to RTT ?
 - If yes, then how large a cache and what network bandwidth is needed for a canonical small, medium, large site ?
 - What is known/planned regarding “data reuse” ?
 - How large is the “working set” of data experiment wide?
 - Has this been measured for Run2 ?
 - If yes, do we know how to use Run2 to project to HL-LHC?
 - What is the expected tape recall bandwidth needed, given disk space, and working set size ?
 - What is the expected network bandwidth given expected cache refresh rates and policies ?

These questions revolve around understanding the experiments data use models.

What Else?

- Are the previous two slides sufficient for an initial plan to proceed ?
- What is missing in terms of questions ?
- Do we need to plan further ahead, or can we proceed with this plan for the near term in this working group ?
- Or should we organize this completely differently ?