Parallel Session - Data Management, Access and Organization / Data Streaming - Summary

Tanu Malik and Oliver Gutsche

2nd S2I2 HEP/CS Workshop, Princeton

03. May 2017

The session

- Mixed audience
  - Good HEP representation, from Atlas, IF, CMS
  - Good CS representation, special thanks to
    - Michela Taufer
    - Carlos Maltzahn
    - Anton Burtsev
    - **for very fruitful and interesting discussion and asking very good questions!**
- We didn't have any lightning talks. We used all the time for discussion.
- Tanu introduced the CS side of data management, OLI talked about the HEP process
- What follows is what we wrote down as the conclusion of the parallel session
  - And I added some lesson's learned which are mostly personal but hopefully help the overall process

Question 1

- Are there places where the HEP language to describe a problem or system doesn't match how a CS person would describe the same problem?
  - Yes
  - Describing the HEP science process, the general scientific workflow, the structure of the data, the workflows, what simulation means, the growth of the scale for HL-LHC time scales ➜ this is very important for CS to understand the overall HEP software and computing problem
  - CS asks for more information about the science use case. Which science would not be possible with today's technologies, what science would be enabled by future technologies? HPC's miniapps and micro-benchmarks mentioned as good example for the HPC community addressing this questions, maybe there is something similar for HEP to describe the data management challenges.
  - For CS it is difficult to see where the bottlenecks are. Where will the system break down. It is ok if this will change over time.

Question 2

- Which things does the HEP community want to do, does not know how, and believes that Computer Scientists may be able to figure out?
  - Is the current approach still the right optimization for HL-LHC?
  - What can we learn from other communities, structural biology, genomics, … ?
  - What are the CS research directions that HEP community can track and leverage?

Question 3

- How do these problems map to CS research questions?

  - Can we built a taxonomy of existing technology and applications that use this technology? (this could be a concrete research topic for CS)

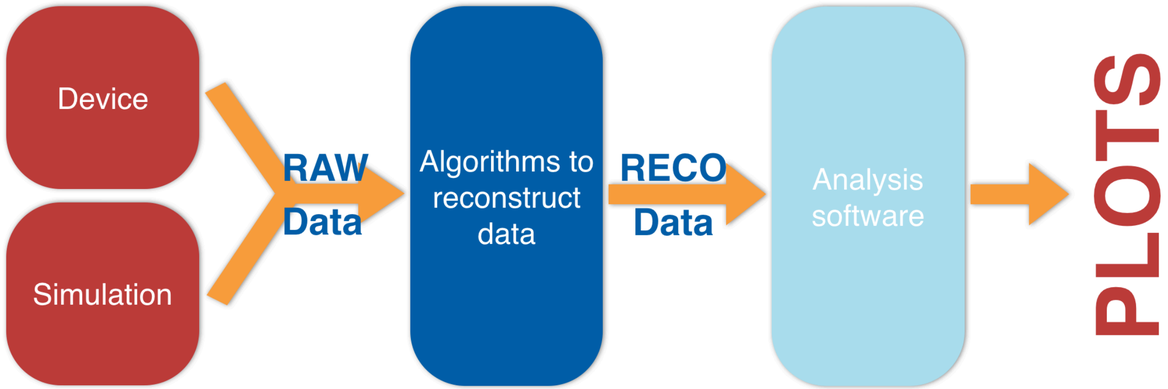
    - Different open source technologies that might be applicable to the HEP data problem (examples: BigTable implementations: Impala, Kudu, Drill), CS is interested in exploring the flexibility of these and other systems using the HEP use cases

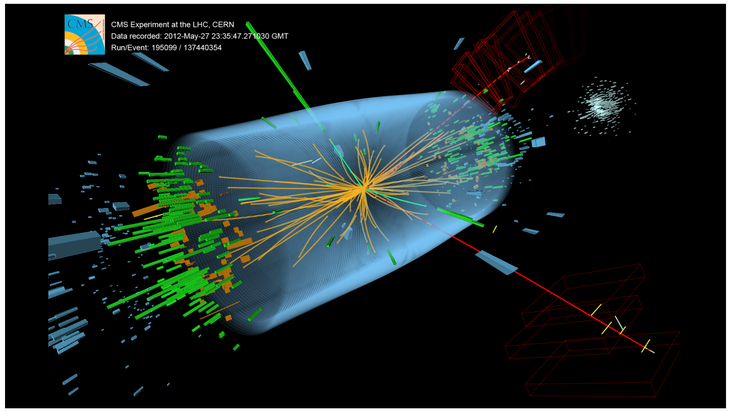
    - If there are similar workflows in for example in genomics, we should investigate what they do and if this is applicable, and perform concrete tests of our workflows with their implementations

    - Building taxonomy means:
      - Developing metrics for the current system
      - Investigate new technologies and their impact on metrics ➜ concrete tests are needed

Lesson number 1

- We are “still” not talking the same language!
  - We spend most of the time to talk about what high energy physics is all about, how we extract the results.
  - For example that we (HEP) treat an event (all detector information and all derived information for either a single recorded particle collision or a simulated particle collision) as an atomic unit was not entirely clear
    - “Oh, but then you are embarrassingly parallel!”





Lesson number 2

- Iteration
  - To really enable CS to combine research directions with helping HEP, we have to iterate much more about HEP's problems and challenges
  - Ideally, when introducing the HEP computing challenges, a CS person would give the talk!

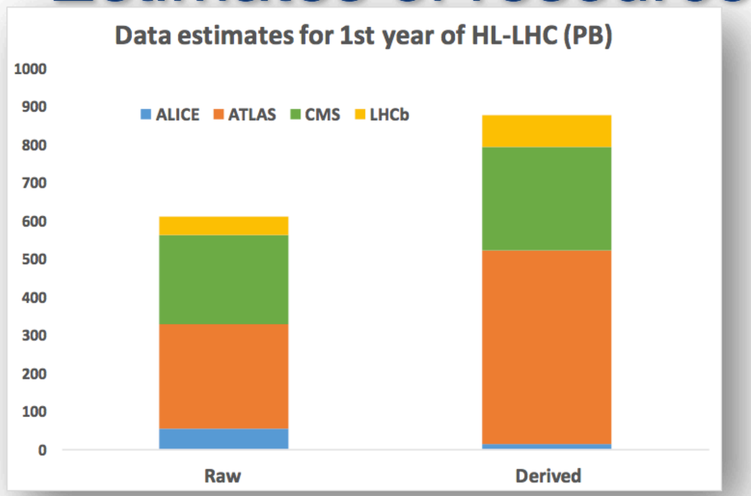
- On the way to concrete projects
  - Possibilities for concrete things emerge
  - More iteration is needed

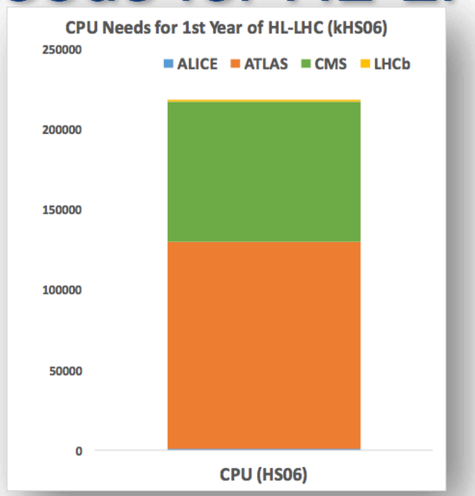
Lesson number 3

- CS is asking good questions ➜ HEP does not necessarily have good answers ready
  - Our computing problem is not trivial, we have many and complicated interdependencies
  - Questions like: “What is your biggest bottleneck” or “What are your top 5 problems” are very important input to CS to start helping us
    - Everyone in HEP has an opinion about this (because we have very detailed knowledge of our current systems and ways of doing business)
    - This is more confusing than helping for our CS friends
  - More and more, the question of trade-offs comes up
    - Trade-offs in the sense of physics trade-offs
    - “How much more physics would you be able to do if you would get X”
- HEP has to be much more focussed and consistent in answering these questions!

No talk in the next 10 years can end without the drinking plot!

Estimates of resource needs for HL-LHC





Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:
- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

- Simple model based on today's computing models, but with expected HL-LHC operating parameters (pile-up, trigger rates, etc.)
- At least x10 above what is realistic to expect from technology with reasonably constant cost


Fermilab
U.S. DEPARTMENT OF ENERGY
Office of Science