How HL-LHC Challenges inform WMS development for CMS

Frank Würthwein

SDSC/UCSD

HSF/WLCG Meeting in Naples

March 27th 2018

“Data Driven Design”

There’s a **desire to base design & development decisions on detailed measurements** of current practice and careful evaluations of change in scale.

As this evaluation is still ongoing, it is **too early to make strong statements about where we are going** with regard to design.

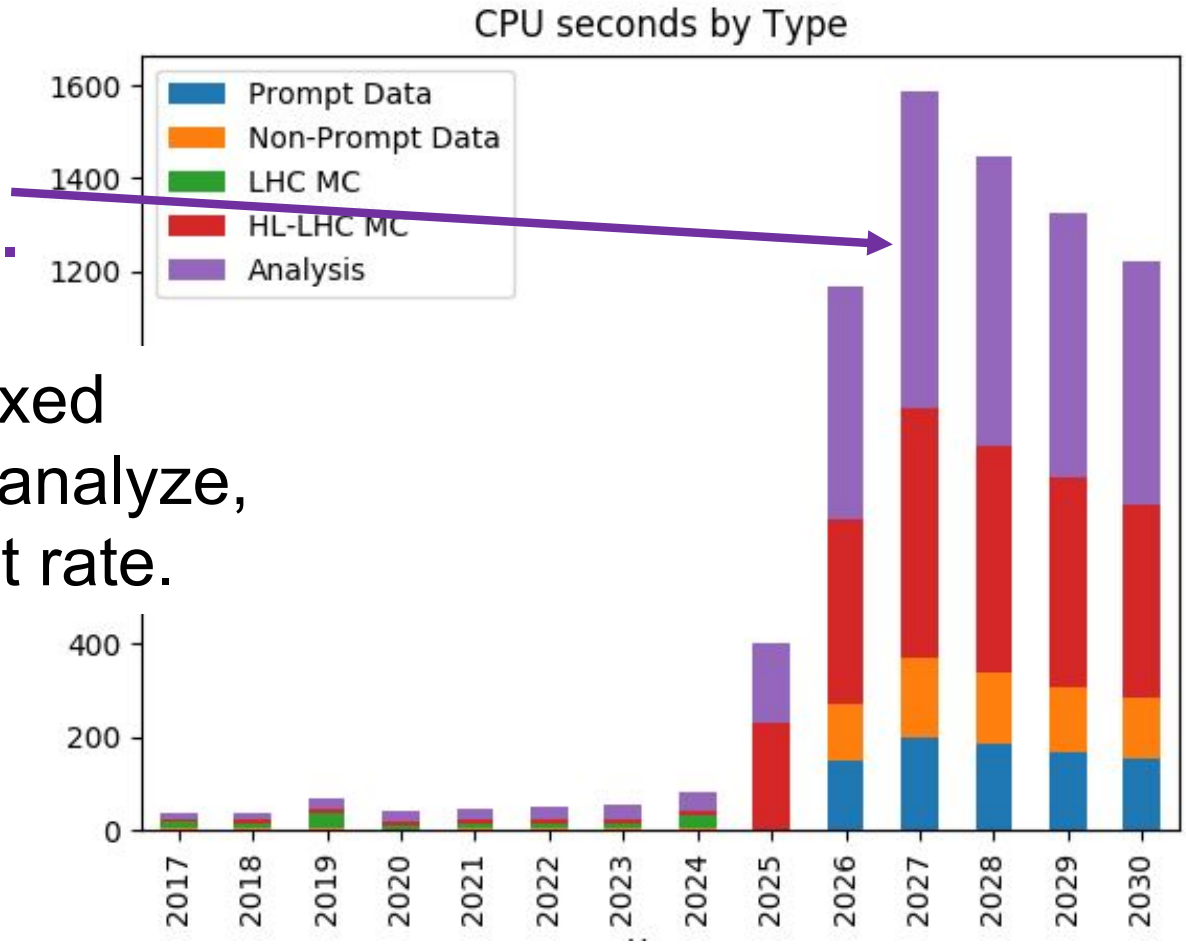
Two High Level Goals

- Understand what it takes to fit computing operations costs within budget
  - All of the detector components we design have a budget, and get reviewed against that budget.
  - Why should computing be allowed to blow the budget by x3?
- Understand what it takes to reduce the student & post-doc effort required to get physics out.
  - It's likely that the effort available shrinks during the HL-LHC time period as other programs in HEP start taking on more resources out of a fixed budget.

Thoughts on CPU Budget

Analysis is fixed relative to production.

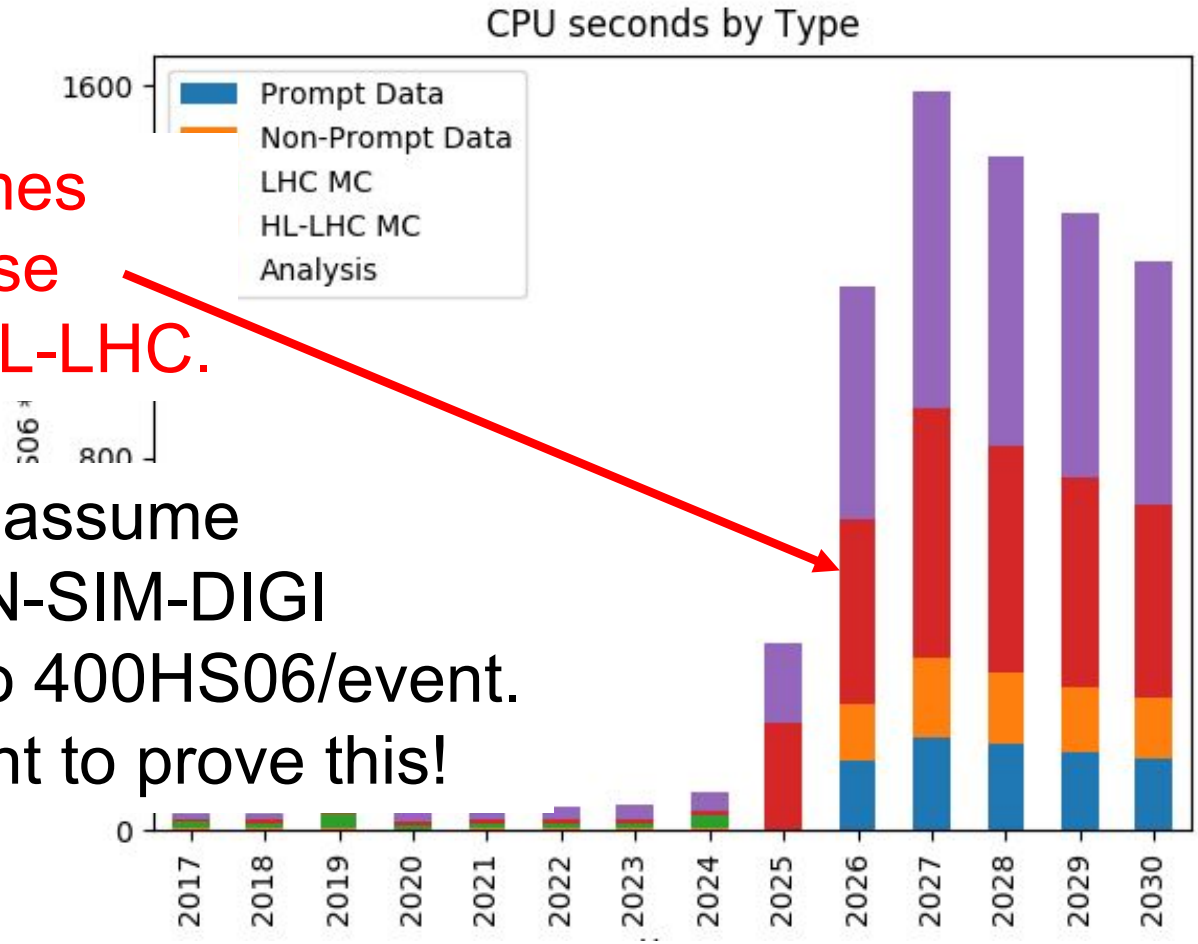
Should instead get fixed relative to # of events to analyze, with ~ fixed CPU/event rate.



A model like that will reduce analysis computing needs to <10% of the total.

Thoughts on CPU Budget

GEN-SIM-DIGI assumes
no pre-mixing because
that is not yet ready for HL-LHC.



It's not unreasonable to assume
pre-mixing reduces GEN-SIM-DIGI
from 4000HS06/event to 400HS06/event.
=> Ongoing development to prove this!

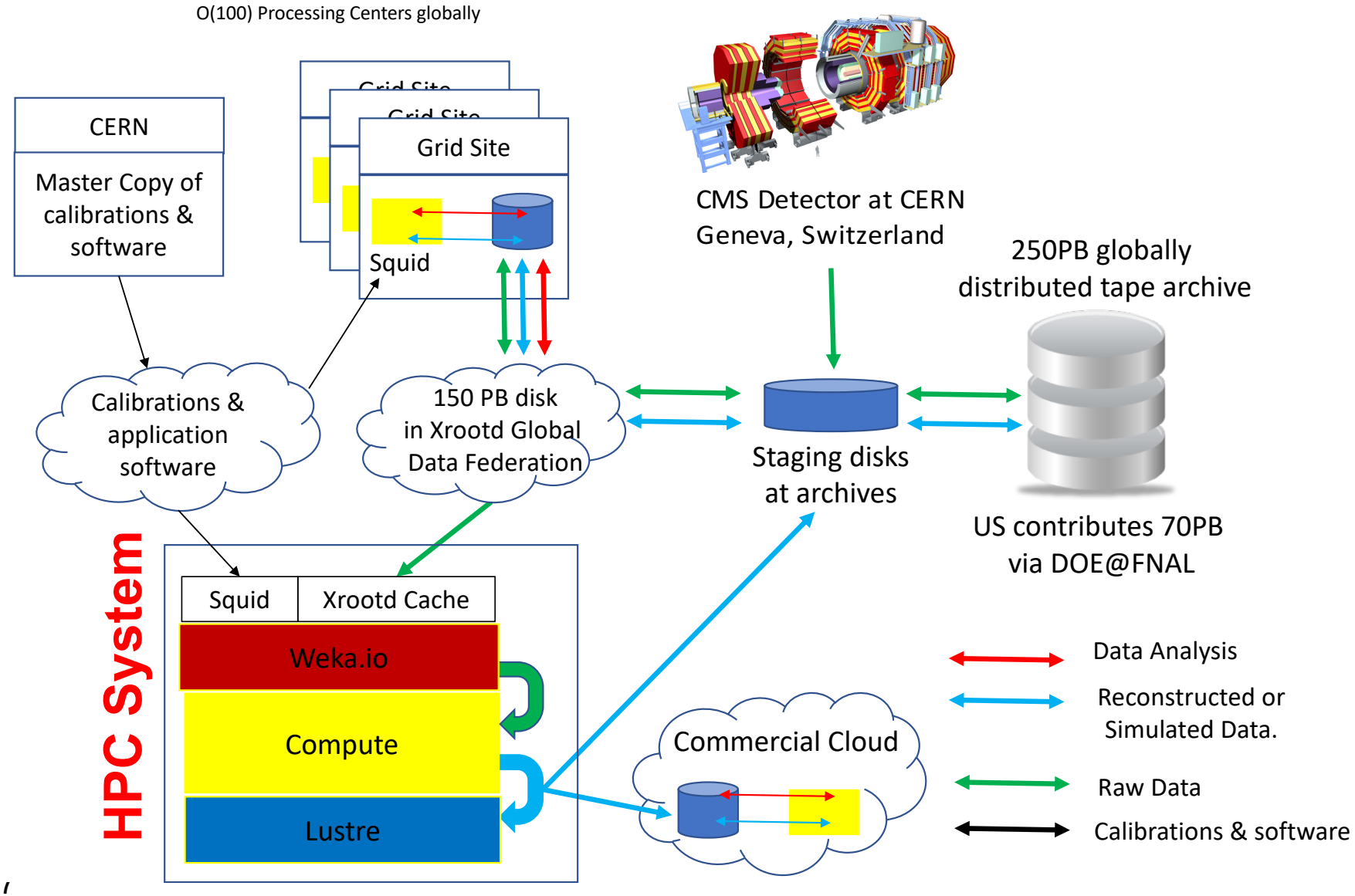
If this works out then HL-LHC MC is completely dominated by CPU time in reconstruction.

WMS Implications

- Pre-mixed PU library today ~ 600TB => expect O(10)PB size PU library for HL-LHC
  - Creation of this data sample is completely IO limited.
- Possible workflow for future HPC systems
  - Individual collisions stored in NVME or Flash based “global” filesystem (e.g. weka.io or alike)
  - Each CPU draws at random from this input sample, and writes out PU library sequentially into SATA storage.
    - A couple hundred random inputs get merged into one output.

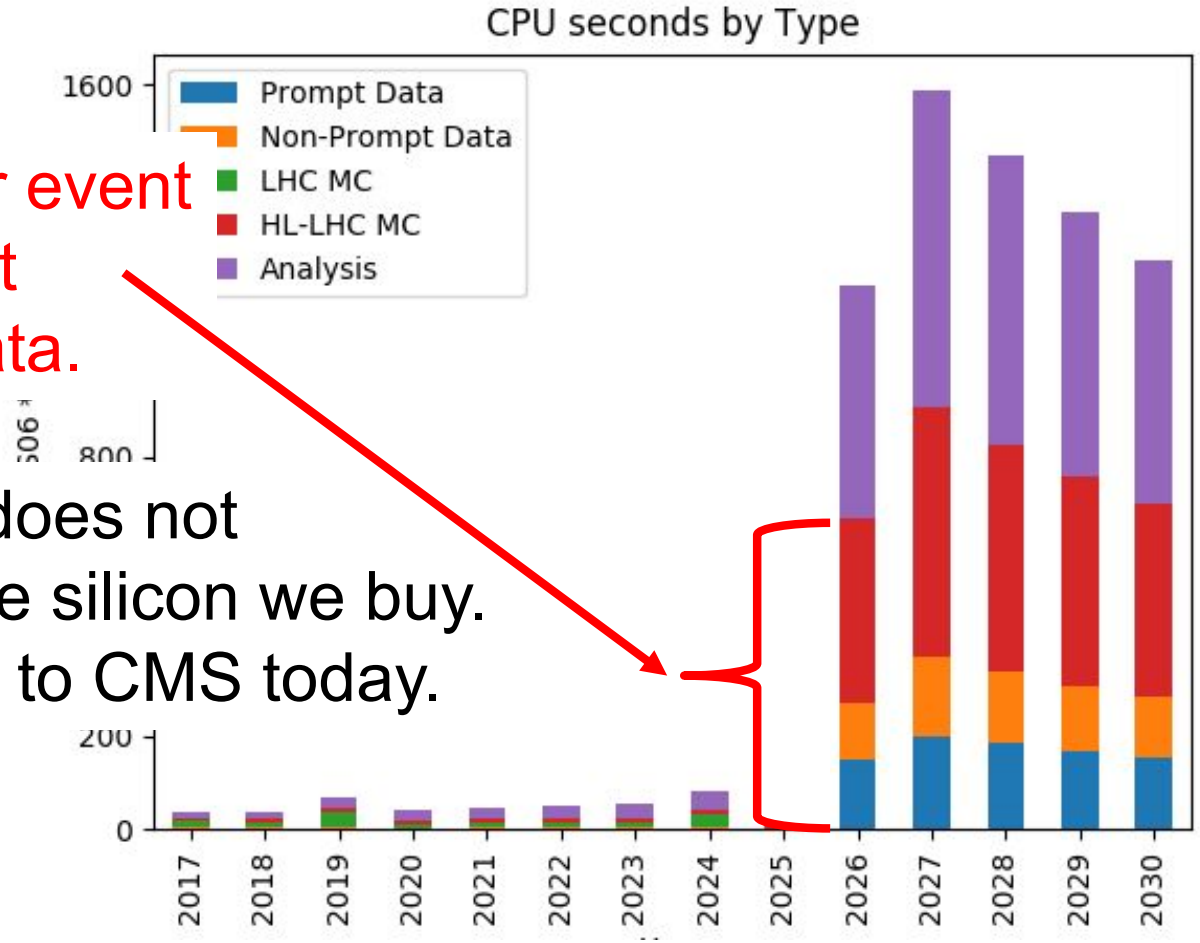
Unusual workflow with modest CPU needs but has to complete before the rest of the MC campaign can start.

Possible HPC Data Integration at today's scale



Thoughts on CPU Budget

Reconstruction costs per event
~ 4000 HS06/event
Same for MC and Data.



Present reconstruction does not make effective use of the silicon we buy.
E.g. AVX-512 is useless to CMS today.

A 25% improvement in reconstruction, 4k to 3k HS06/event, would likely be sufficient to have CPU within budget given the previous 2 improvements.

Thoughts on Storage Budget

- The tape budget is something we can not shrink significantly
  - "virtual data" makes no sense given the high CPU cost of reconstruction.
- Disk is ~x5 more expensive than tape.

=> The best bet for reducing storage budget is to rely more on tape and less on disk.

Thoughts on Storage Budget

- Size of different event formats covers x500
  - RAW ~ 5MB
  - AOD ~ 2MB
  - MiniAOD ~ 200kB
  - NanoAOD ~ 10kB
- DPG/TSG/POG needs might dominate the total disk budget because they need subsets of larger formats.
  - How large a subset is sufficient for algorithm & physics object development ? Which ones for what development ?
- This is a paradigm shift for how the content of the Nano(Mini)AOD is developed !!!

Implications for WMS are unclear !

Reducing effort to produce physics results from primary data

Today, we spent O(100) FTE on converting primary data into user data (=ntuples).

Can we reduce this dramatically ?

Two Strategies

- NanoAOD
  - A common, CMS wide, centrally produced event format that is smaller, faster, and more useful than most user level ntuples.
  - Can it eliminate the need for user ntuples ?
- A radical paradigm shift towards “declarative programming” at “data facilities”.
  - Some T2 centers would turn into data facilities.

Both of these may have implications on WMS that we don’t fully appreciate yet.

Common Project Ideas

- Data Analytics that underpins the measurements we need to drive the design.
  - E.g. we don't have a measurement of the "working set" of data for a time interval commensurate with recalling data from tape.
- Where does tape recall fit ?
  - Is it a facility issue hidden from the experiments ?
  - What do the experiments specify to guarantee performance?
    - Tape recall of an Exabyte for reprocessing !!!
- Is there a common interest in "Data Facility" issues to support "declarative programming" for analysis ?