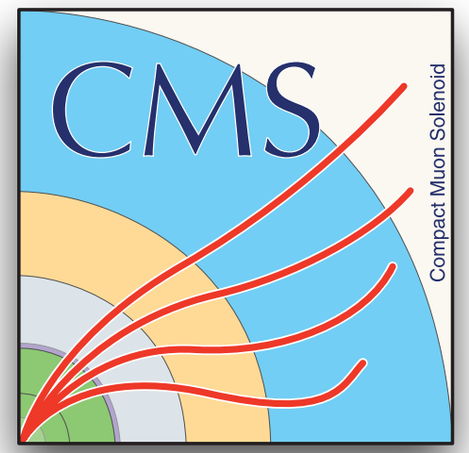
CMS Networking: Current and Future



Liz Sexton-Kennedy and Daniele Bonacorsi

Currently

- CMS is still in a regime where it **treats WAN bandwidth as an infinite** resource.
  - We can manually manage the cases where this isn't true.
- CMS uses a dynamic data management system that heavily relies on reliable, consistent network transfers.
- Biggest recent changes are in computing model: new reprocessing strategy for MC ("premixing") reduces storage&LAN load by >10x.
  - We can **effectively stream data to remote sites** without needing *any* local datasets. This was viewed as impossible 4 years ago.
  - High bandwidth is still needed at a few sites (FNAL, CERN) to create libraries of pileup data.

Evolution of the Data Federation (AAA)

- CMS increasingly relies on its data federation to increase effectiveness of production activities.
  - Our workflow management system can “overflow” jobs from sites whose CPU resources are fully occupied. Data placed at them is read over the WAN.
    - For premixing, both signal and background events are read via WAN.
  - In 2016, this became an everyday technique within “network regions” (locations with strong network connectivity, for some definition of strong).
  - In 2017, we’ll do this over the transatlantic links for select site pairs.
- Opportunistic computing is dependent on strong networking (may have large CPU capacity, little disk, and lots of WAN bandwidth).

Requirements of the Federation

- In terms of network activity, there is no special distinction between T0<->T1, T1<->T2, or T2<->T2 traffic *for CMS*.
    - The distinction between LHCOPN and LHCONE is no longer supported by the CMS Computing model (still - might be useful for other cases!).
- Current production use cases need sites to export to AAA at 10Gbps.
- Production now requires the federation. Sites should plan for production re-processing over WAN to go from the current 20% share to 50%.
    - This does not necessarily imply a doubling of the provisioned bandwidth!
    - Analysis is more chaotic and less predictable; we still view analysis use of AAA as an optimization which reduces failures.
- Priority of the data federation traffic should be as high as FTS3-based traffic at the infrastructure level.
    - We ask sites to no longer prioritize GridFTP (storage-to-storage) over AAA (storage-to-workers) traffic.

Evolution of Requirements

- How will bandwidth required grow?

  A. As the LHC delivers more data, the total amount of **data transferred per event will increase**.

  B. Due to the increased event complexity, CPU time increases, meaning **event processing rate decreases**.

- Does the decrease in (B) cancel out the in increase in (A)? Not clear at this point.

- We are sometimes blocked by bursts of staging from tape, rather than steady-state. Investigating how we mitigate this: this may result in a decrease in bursts.

ECOM2017

- CMS needs to re-evaluate it's computing model. The charge for the ECOM17 working group was just sent by the SP. It will help the S&C project address the issues raised in the CRSG.

- Many of the things we could decided to do in order to address funding shortfalls in 2018 would have an effect on data movement and storage.

  - For example we could decide that it is faster and cheaper to remake the simulated hit data than to archive it and later stage to a reprocessing site.

New Network Capabilities Needed?

- Can we increase network utilization (when needed)?
- Improved quality of service: make network performance more consistent.
  - Broaden use of monitoring; automate problem identification and alerting?
- Can we inform our workload management systems about expected network/transfer performance between sites (inc. worker nodes)?
  - Pairs of “good sites for AAA” are manually maintained.
- Better ways to describe our instantaneous data movement patterns to the underlying networks.
  - Could we inform the network of relative traffic priority for bulk transfers? CMS sees this as a potentially beneficial project between network layer and FTS3.

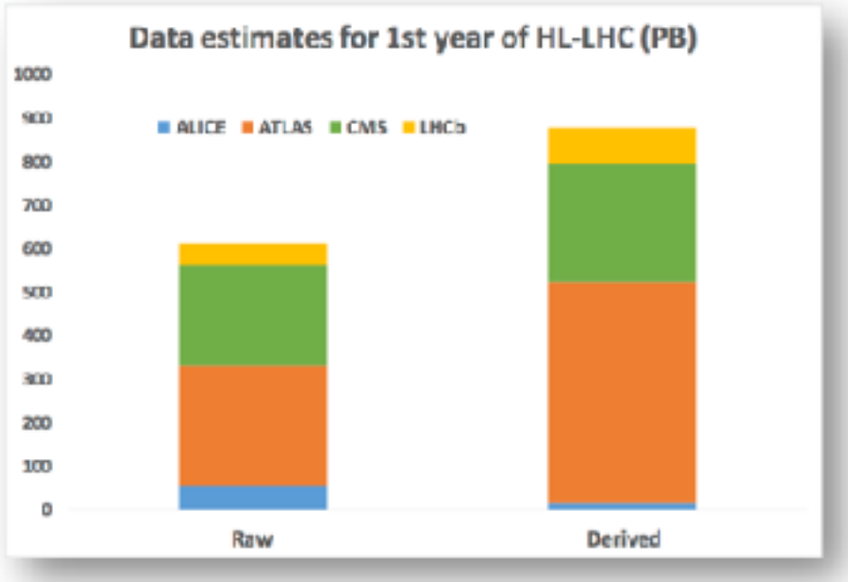
HL-LHC

- **Scarcity drives innovation**: struggle to forecast scarcity in the network, given solid track record of historical upgrades.
    - In some regions, 100Gbps upgrades are only recently completed. Scarcity feels a ways off, especially compared to better-understood CPU challenge; hard to create urgency.
- The jump in data storage between Run3 and Run4 is large. Whether this results in corresponding jumps in network needs depends on (unknown, future) changes in the computing model.
    - One thing is clear: flat hardware budgets and zero-sum reallocations of funds.
- We fully support investments into network research even if in the medium term it will be mostly “R” and less “D”.

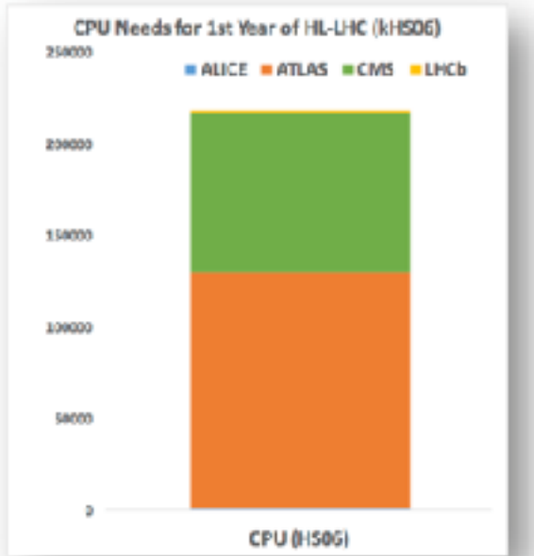
Backup

Beyond Run2

Estimates of resource needs for HL-LHC



Storage
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

=> x10 above what is realistic to expect from technology with constant cost

- This plot was shown many times at CHEP16