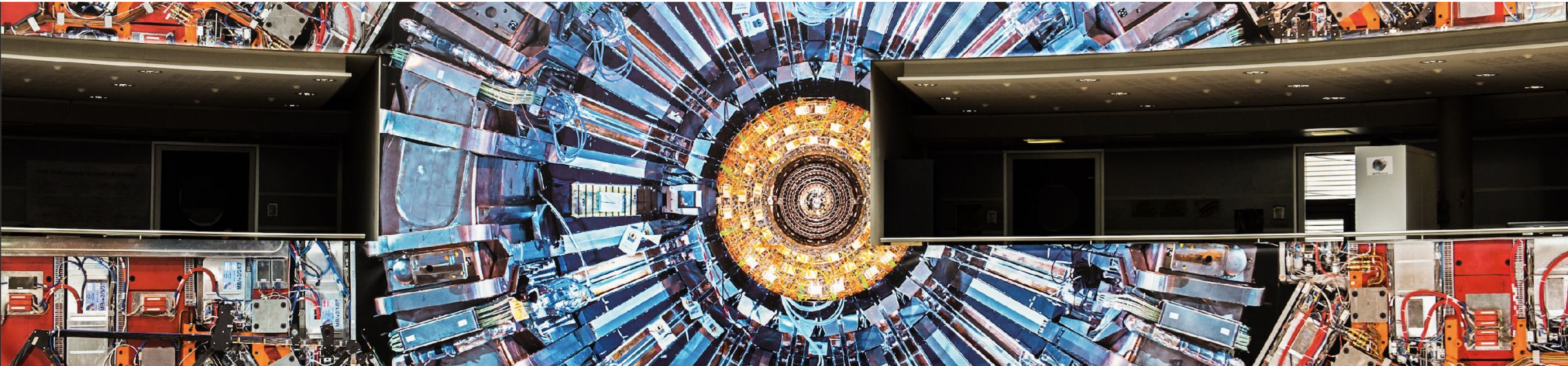
CMS Update

M. Klute (MIT), D. Piparo (CERN) for CMS Offline & Computing - LHCOPN/LHCONE Workshop - 13-1-2020



This Talk

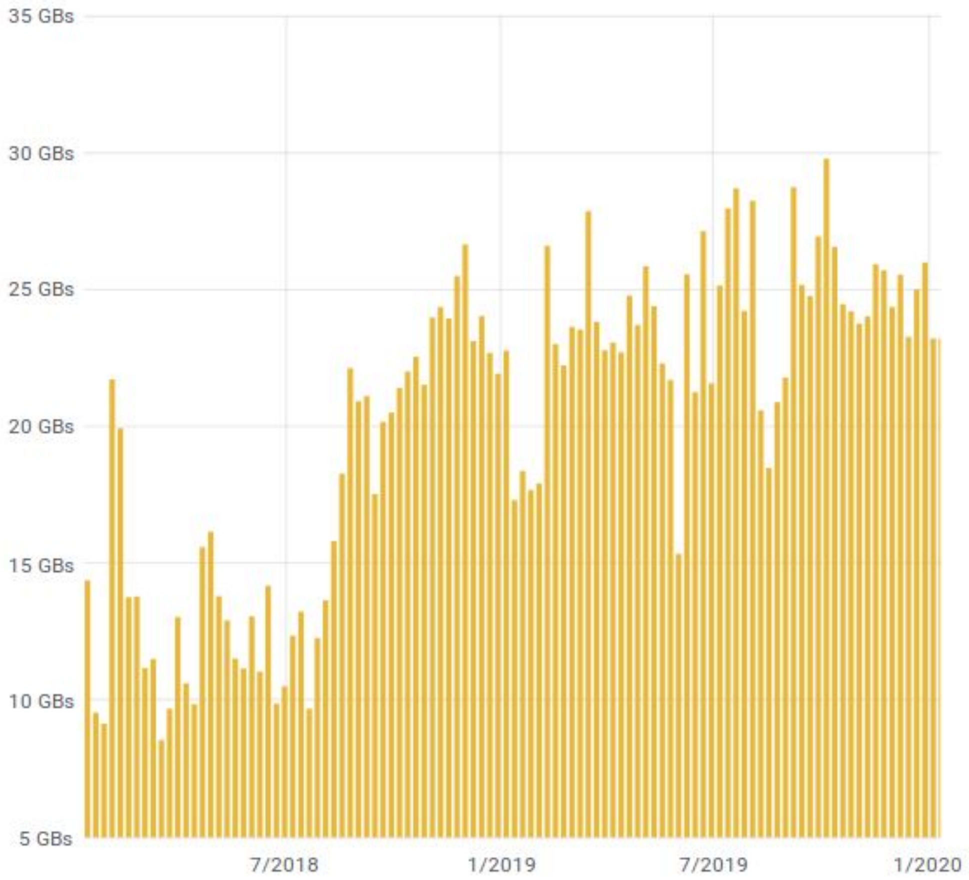
- Current usage of network
- Cost and availability of Network for HEP: a transition in our computing models?
- Interplay between compute, storage and network in the HL-LHC (and exascale) era
- Ideas for future research activities

Current Status

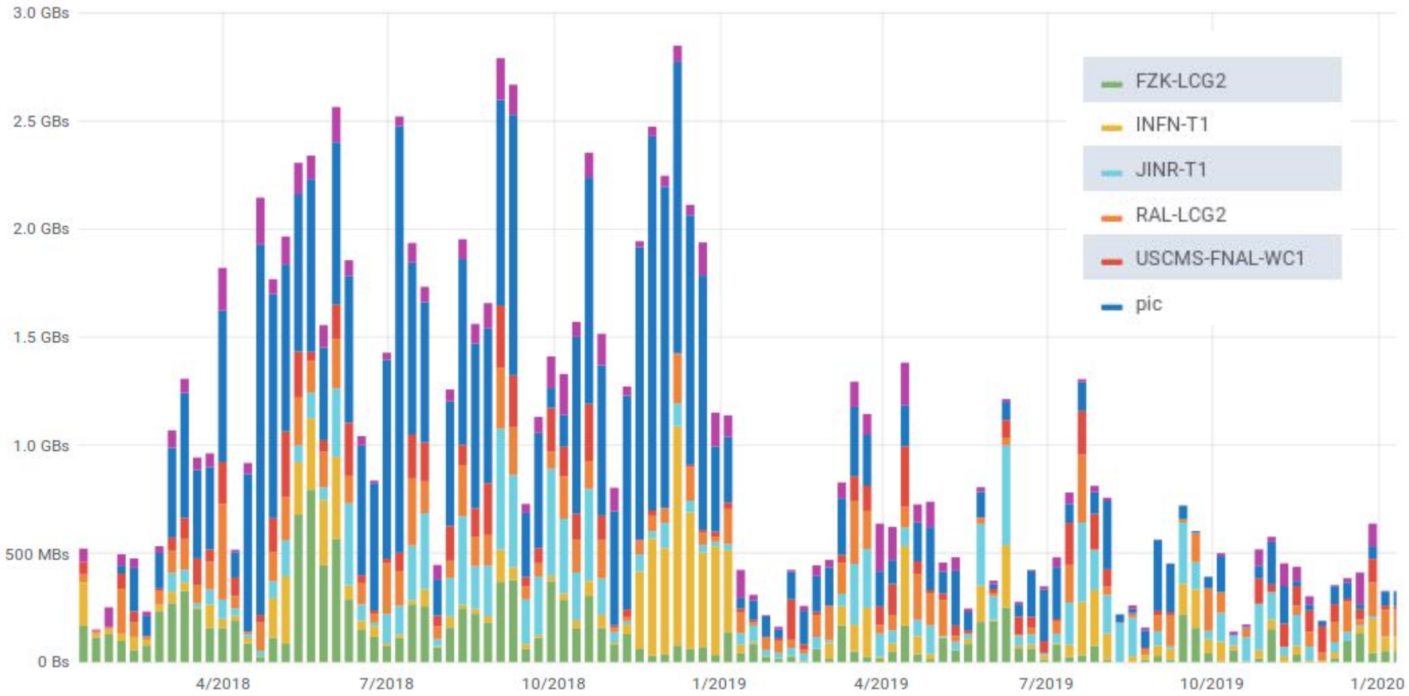
- An essentially **free, unlimited resource, global and regional**. Counter examples:
  - Not optimally connected areas, e.g. TIFR, or places where pay-per-use model was partially adopted
- CMS relies on a **dynamic data management system depending on reliable and consistent transfers**
- **Network cost (up to now) not considered** when planning data movement, e.g.
  - (Intercontinental) replicas created just with expectations of improving CPU usage
  - Establishing 2nd copy 10 or 10.000 km away has the same cost
  - Second copy established to accelerate finalisation of a dataset whose processing is almost completed
- **Network was a crucial ingredient for the success of CMS in RunI and RunII**
  - **CMS counts on the same quality for RunIII and HL-LHC**
  - At the cost of **no additional complexity for the experiment and sites**: ideally transparent evolutions

Transfer Throughput

Total Transfer Throughput



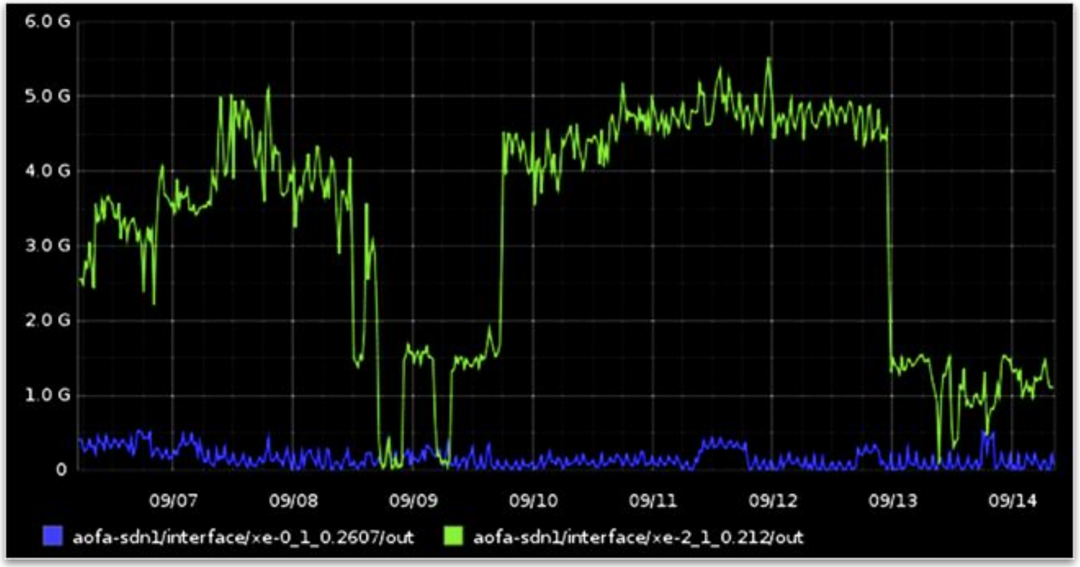
Transfer Throughput: T0 to T1s



Not only RAW data, see Simone's slides

Network in the Future

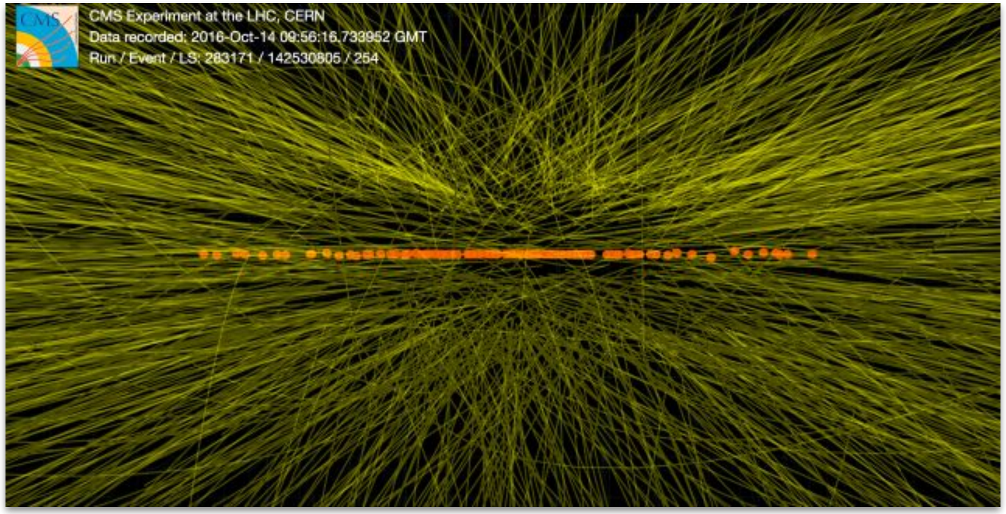
- Changes ahead of us in the next few years
- **Network evolution still in good shape**
  - But evidence that **for links needs are exceeding potential technology evolution rate**
- Changes are motivated both by **technological and non-technological reasons**, such as:
  - Limitations of the **optical interfaces and switching equipment**
  - Other **non-HEP large data volume** sciences coming online
  - Cloud native / commercial networking and their **impact on usage policies**



Courtesy of H. Newman

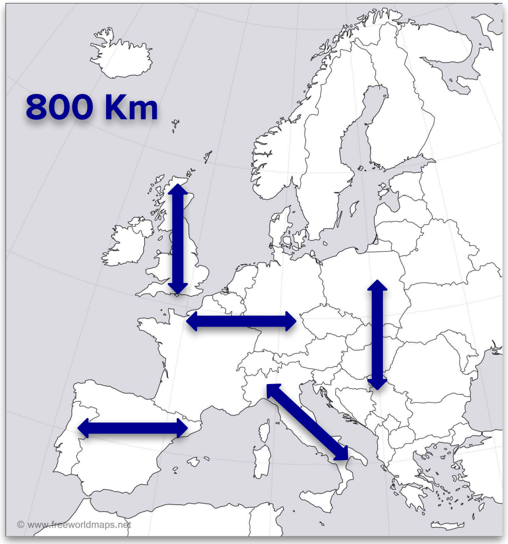
Example Use case: Pile-Up Simulation

- CMS approach **simulate hard scatter events and "overlay" pile up at a later stage**
  - I.e. during "digitisation" - the so-called "premixing"
- Secondary pp collisions modelled by **"Pile up only events", stored in a big dataset**
  - The "pile up library", two copies of it one in Europe and one in the US
  - Dataset at the PB scale: cannot afford more!
- During digitisation, **pick "pile up only events" from the library** and superimpose those
  - Only two copies, **remote reads from T1s, T2s and other sites**
- Premixing: **reduce overall I/O by ~100x wrt previous CMS mixing method**
  - Move reads from local to remote
- By far, not the only case of "remote read": strategy used for analysis and also for speeding up processing completion

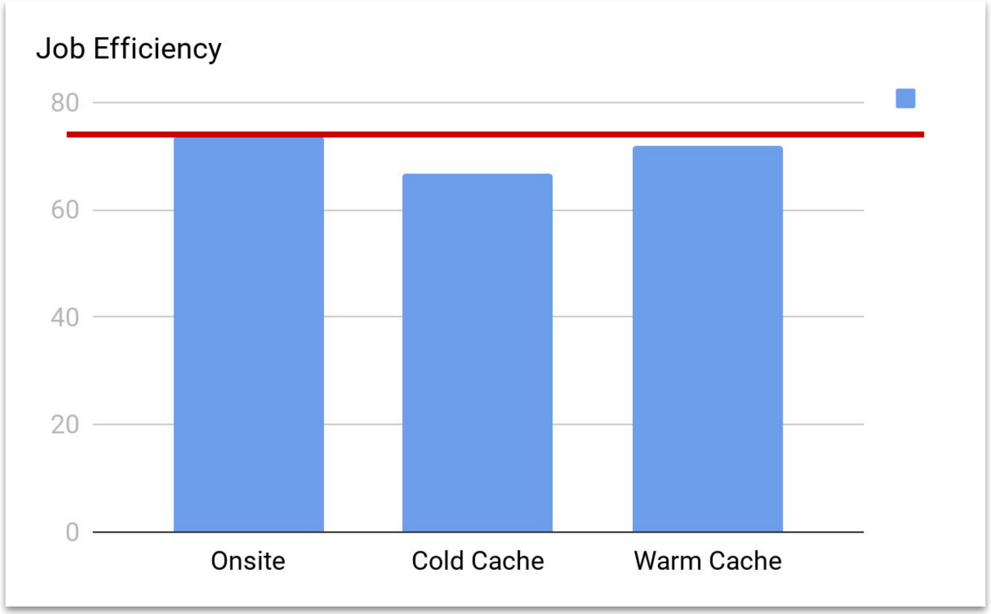

CMS
CMS Experiment at the LHC, CERN
Data recorded: 2016-Oct-14 09:56:16.733952 GMT
Run / Event / LS: 283171 / 142530805 / 254

Network as a Mean to Reduce Storage Needs?

- Activity in collaboration with DOMA
- Objective: **optimise storage amount and its operational costs**
- Concept: **geographically distributed caching layer**, mainly targeting analysis
  - Experiment agnostic
- Examples: XCache in production in California - SoCal
  - Access by Caltech and USCD
  - Other examples, e.g. in Italy
- A **“turn-key storage cluster” of JBODs**
  - No compute power, no RAID/HDFS replicas
  - Same storage across sites
  - Embedded data popularity
  - Good network is crucial for data delivery!



CMS Analysis formats: NanoAOD 1-2 kB/ev, MiniAOD 50 kB/ev



See this and this talk at CHEP

Caches: a promising way to reduce storage costs, heavily relies on good and cheap network?

Storage and Opportunistic Resources

- CMS cannot replace WLCG sites with HPCs/Commercial Clouds
  - But is committed to make the best use possible of allocations of that type
  - (Pre-)Exascale era: a single, big allocation can influence the processing plans of a full year

The example:

- Hypothetical opportunistic allocation: **2 months**
- Use for reprocessing one “nominal” HL-LHC year, i.e. about **300/400 PB of RAW** data
  - Many more events than Run2, bigger RAW events due to detector upgrades

The implications:

- **60 GB/s in input** to the data centre from a Lake
- **Output 2 - 16 GB/s** (depending on the tier, AOD or MiniAOD)

Network to connect foreign storage to the CPU at sites. Increased CPU sharing (especially HPCs) requires more network usage.

Make sure Tbit/s level connectivity is available between the lake and major computing power providers in a region

Encouraged Future Activities

- **CMS supports investments in network research** contributing to the HL-LHC program
- Before anything: **it is mandatory to make our network usage visible and carefully monitored**
  - **Capital mistake: theorize before one has data** - Risk to twist facts to suit theories!
- **Network usage monitoring, traffic labelling (e.g. experiment, workflow type), precision telemetry**
  - Identify patterns degrading network performance and tasks limited by network in non-obvious ways
  - Evolve computing operations practices and CMS-SW to use network more wisely
  - Compare between experiments and improve if possible
- **Shaping data flows, for example through packet pacing**
  - Needs to be achieved as transparently as possible for the experiments' ops teams and software
  - Interplay with streaming network telemetry, e.g. adapt pace according to packet loss per flow
  - Monitoring stability of the transfers, predictability of the system
- **Test of virtualised GPUs in the Cloud through novel protocols**
  - CMS could provide CPU-GPU heterogeneous workflows in the future

Summary

- Network treated as an **infinite commodity resource during RunI and RunII**
- **CMS built a computing model also based on copious and reliable network**
    - **Need plenty of this resource also for HL-LHC**
    - Ideally with **minimal complexity to deal explicitly with at experiment and sites level**
- **CMS supports investments in network research** contributing to the HL-LHC program
    - New experiments, technological evolution not keeping up with needs
    - **Need innovation**: sw defined networks? Intelligent pacing? Something else?
- **Detailed monitoring needed** (labelling, precision telemetry) to improve
    - operations and data management strategies
    - effectiveness of debugging, e.g. clear ownership of problems