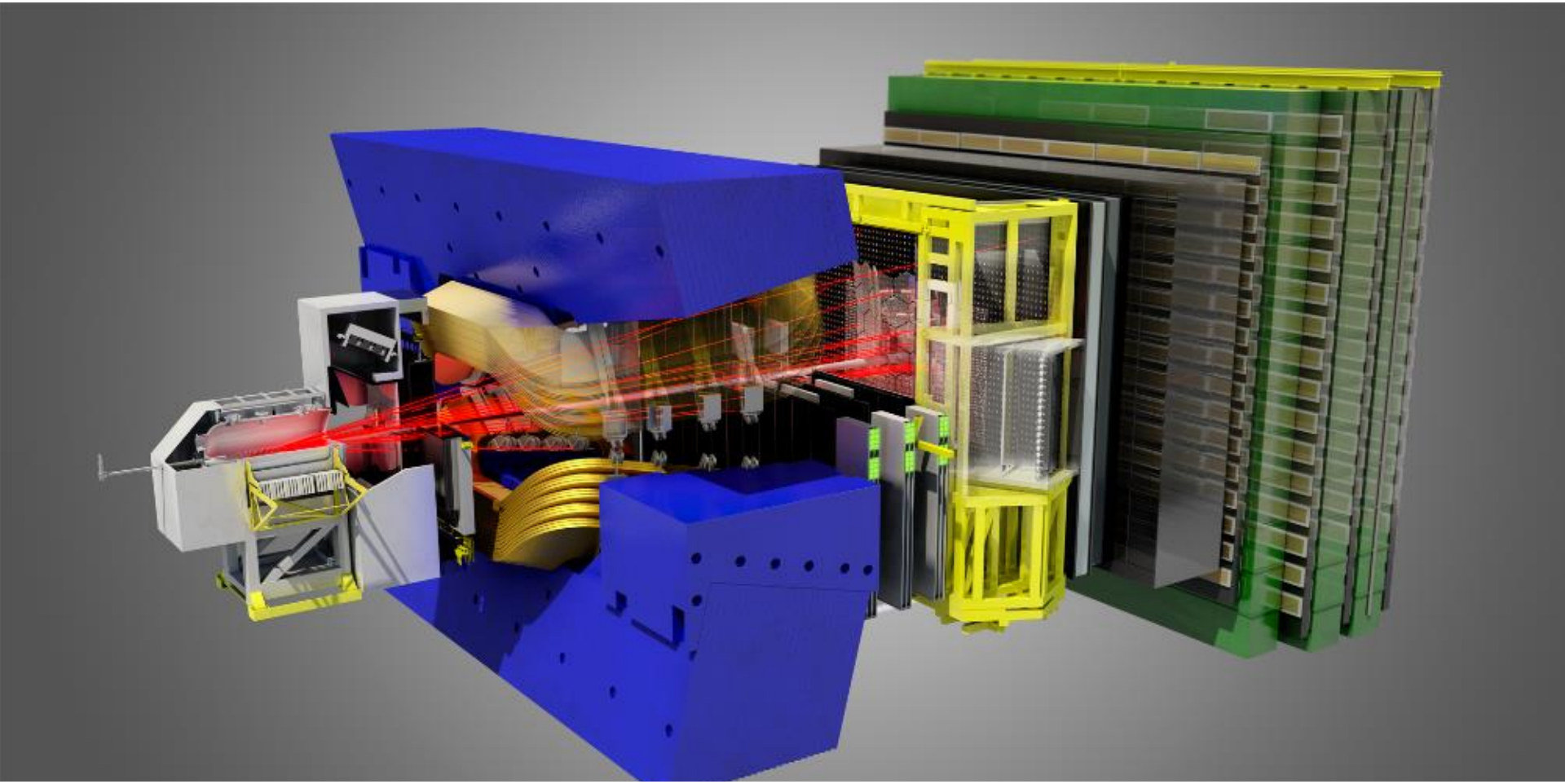
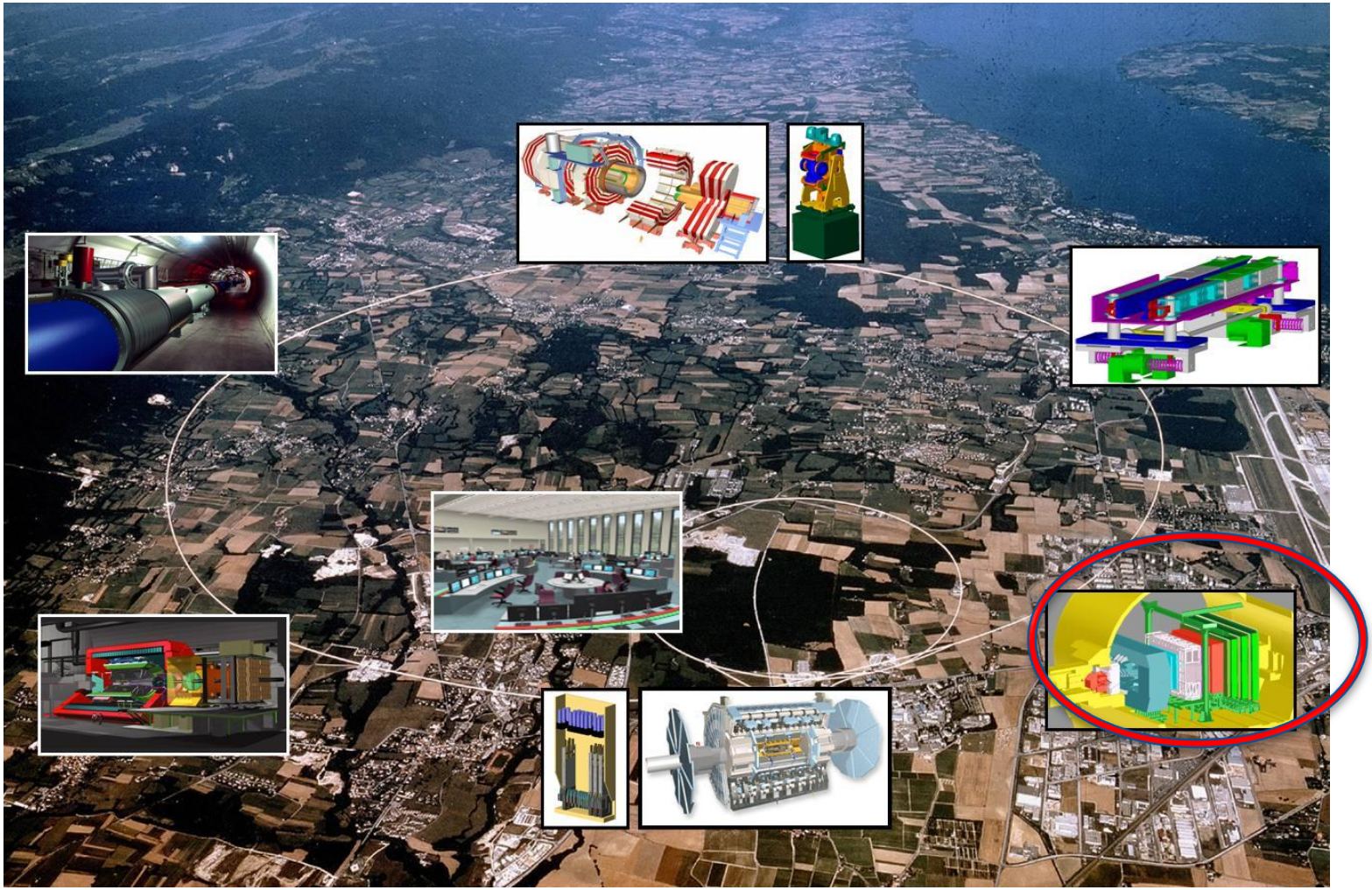


LHCb Upgrade Online Computing Challenges

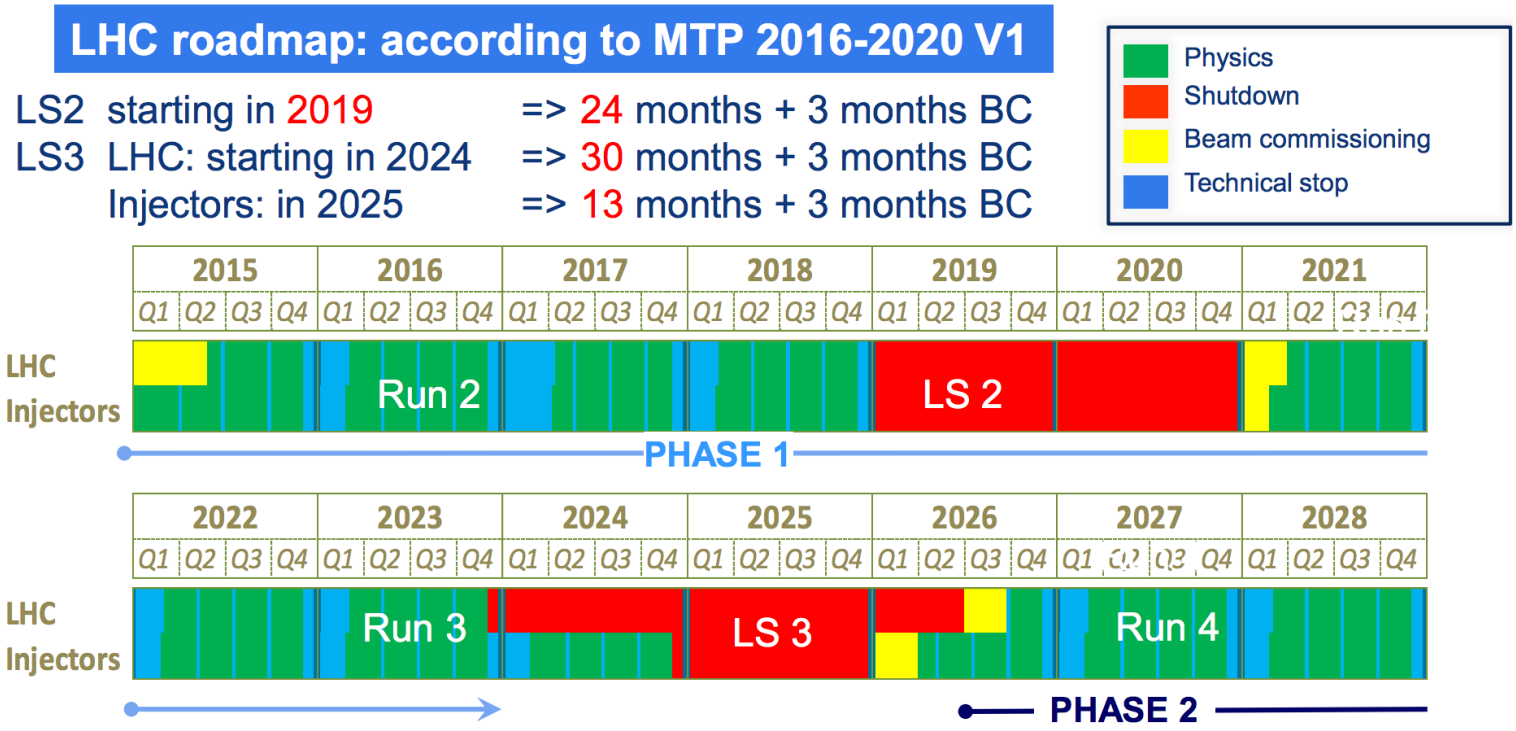
CERN openlab Workshop on Data Center Technologies and Infrastructures, Mar 2017

Niko Neufeld
niko.neufeld@cern.ch

The Large Hadron Collider

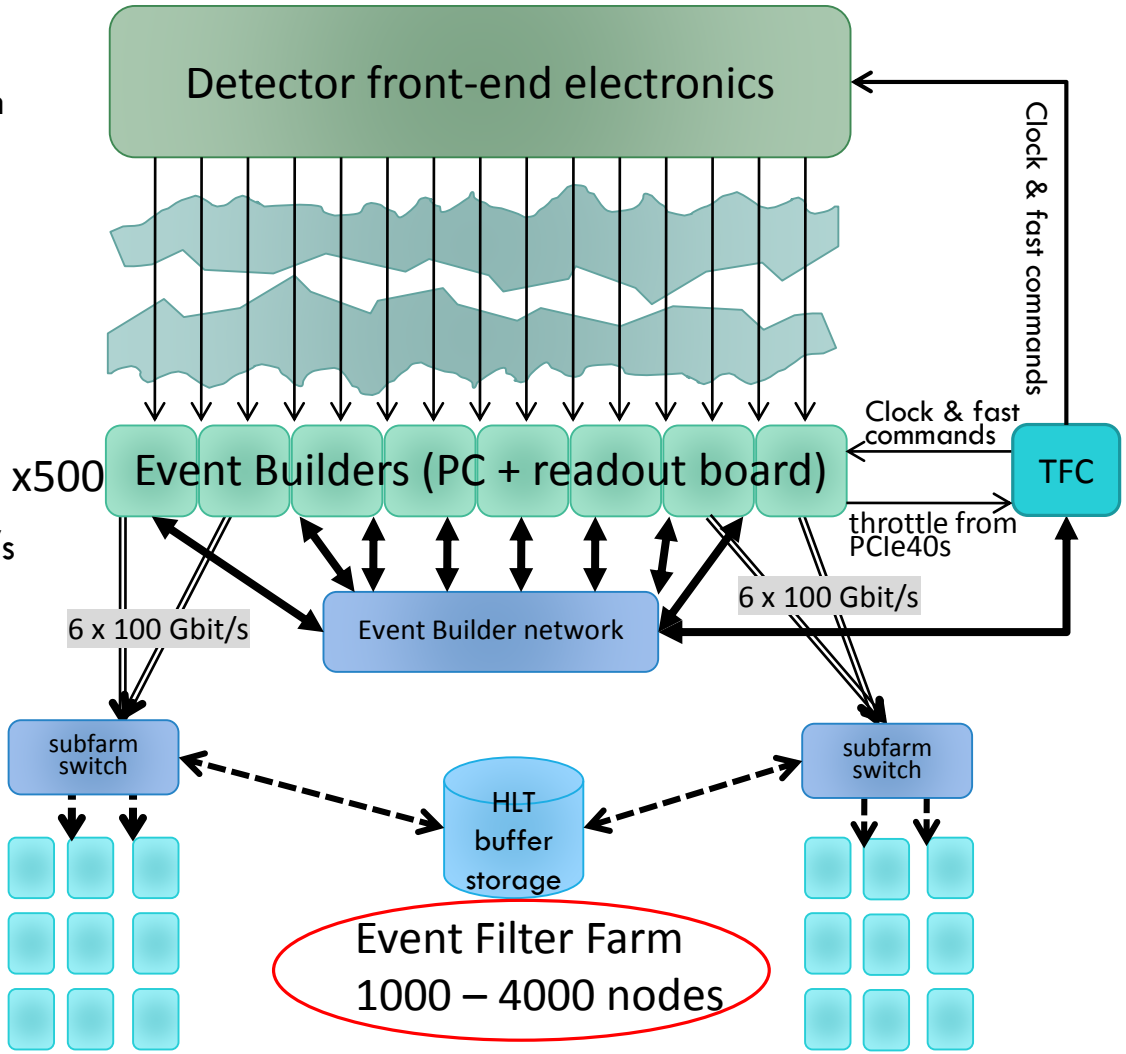


LHC long-term planning



Run3 Online System

- Dimensioning the system:
  - **~10000** versatile links ~ 300 m
  - **~500** readout nodes
  - **~40 MHz** event-building rate
  - **~130 kB** event size
- High bisection bandwidth in event builder network
  - ~40 Tb/s aggregate bandwidth
  - Use industry leading 100 Gbit/s LAN technologies
- Global configuration and control via ECS subsystem
- Global synchronization via TFC subsystem
- 100 PB buffer storage



Required components for a TByte/s scientific instrument at the LHC

"Zero-suppression" on front-end in real-time, 10000 radiation hard links over 350 m, etc… (not covered here)

A 100 Gbit/s acquisition card (1 slide!)

"Event-builder" PCs handling 400 Gbit/s

A high-throughput / high-link load network with 1000 x 100 Gbit/s ports. Candidates: OmniPath, InfiniBand, Ethernet

A ~ 2 MW datacentre for about 2000 to 4000 servers (with accelerators: FPGAs? GPGPUs?)

O(100) PB @ O(100) GB/s read and write for intermediate storage, ~ 4000 streams

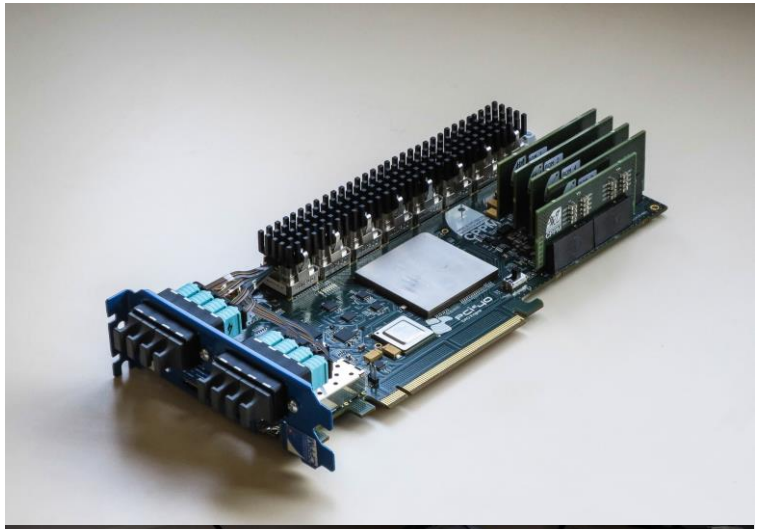
Processing capacity to process all those data and reduce them to an amount we can (afford to) store, which is about O(10) GB/s

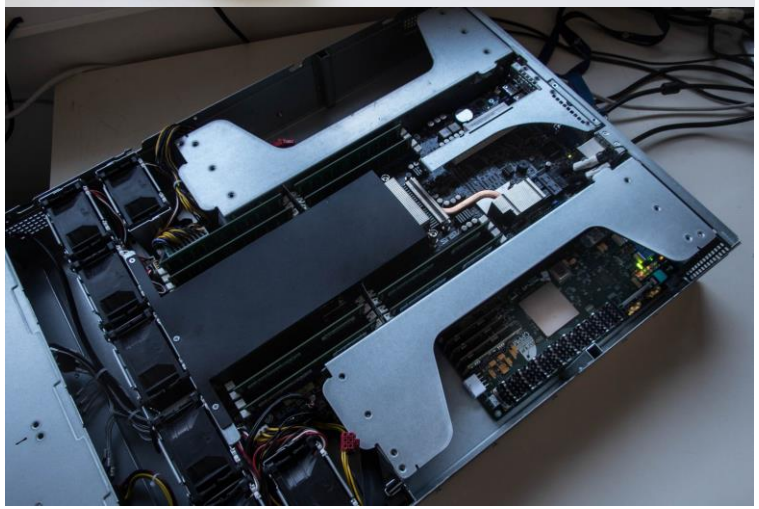
Design principles

- Architecture-centered, not tied to a specific technology
- Guiding principle is the overall cost-efficiency of the entire system, where lowest initial capital expenses are favored
- Open for any technological development which can help / cut cost as long as it is (mostly) COTS
- Focus on firmware and software

Custom hardware and I/O

- PCI Express add-in card
  - Altera Arria10 FPGA
  - 100 Gbps DMA engine to event-builder memory
- High-density optical IO
  - Up to 48 transceivers (Avago MiniPODs)
  - Reuse same HW for timing distribution system
- Decouple FPGA from network
  - Maximum flexibility in network technology
- Exploit commercial technologies
  - PCI Express Gen3 interconnect
  - COTS servers designed for GPU acceleration
- With sufficient on-FPGA / on-card memory this could be re-purposed as a very powerful OpenCL accelerator





Online Cost optimisation: it's all about TCO

- Control and Monitoring cost mostly fixed – determined by number of "devices"
- Main cost drivers:
  - CPU (+ accelerators)
  - Storage
  - Number and length of fast interconnects
- Detector links are there and optical anyhow →
- Bring all data to single (new) data-centre

NETWORK

DAQ Network challenges

- Transport multiple Terabit/s reliably and cost-effectively
- **500 port full duplex, full bi-sectional bandwidth network, aiming at 80% sustained link-load @ >= 100 Gbit/s / link** (industry average is normally 50% on fully bisectional network)
- Integrate the network closely and efficiently with compute resources (be they classical CPU, "many-core" or accelerator-assisted (FPGA))
- Multiple network technologies should seamlessly co-exist in the same integrated fabric ("the right link for the right task")

Long-distance data-transport option

The CPU cluster can be hosted elsewhere. In that case the 32 Tbit/s raw data need to be transported off-site

Distance: about 10 km

Amount of SMF available dictates the use of DWDM

Using 25 Ghz wavelength about 1200 lambdas are needed in the maximum configuration.

The solution should be compact, does not need redundancy (non-critical data), it should be scalable (starting smaller to be ramped up to max later)

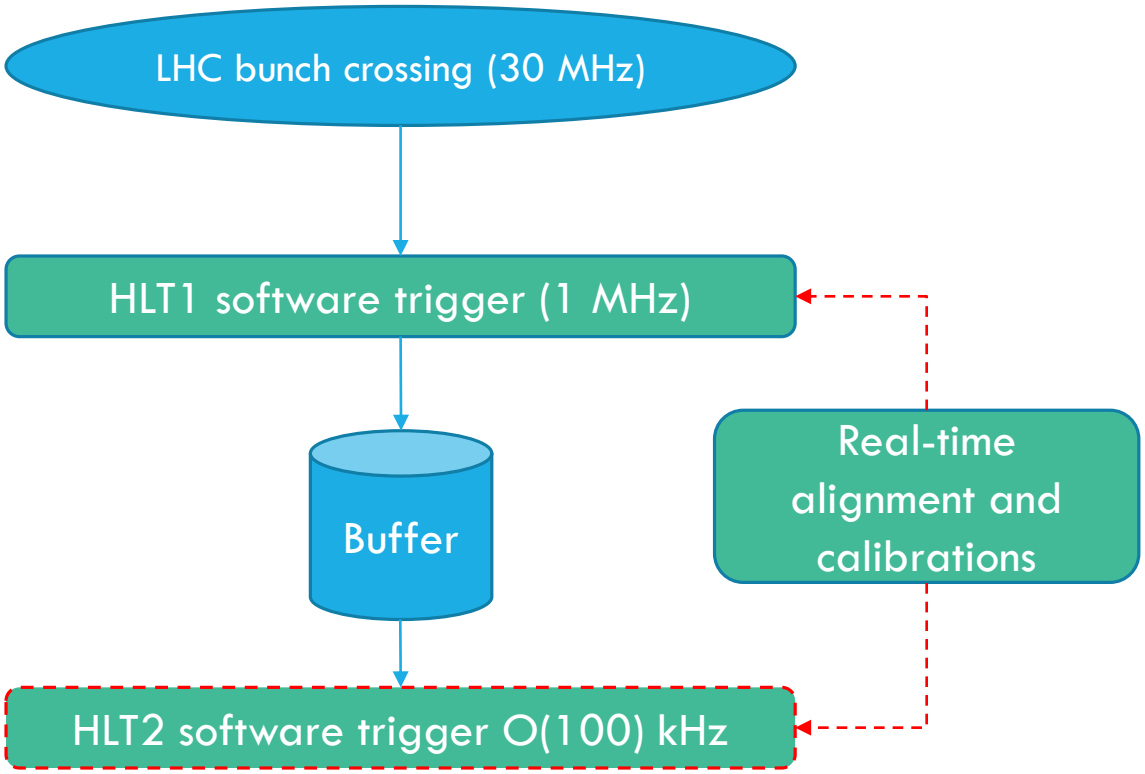
Traffic is essentially uni-directional → Could this be exploited?

Would prefer a Layer 1 option (protocol agnostic), but ok to use Layer 2 or Layer 3, if lower cost

In any case data transport cost / Tbit/s is significant, compression of data is attractive, if cost-efficient

STORAGE

Buffer-needs for Run3

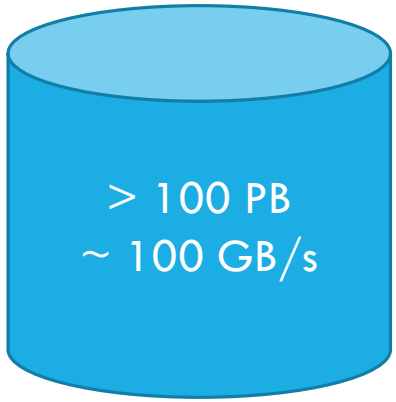


- Same concept but now very challenging numbers
- 1 MHz * 130 kB → **4.5 PB / day** @ a Hübner(*)-factor of 0.4
- To cover a good LHC week might need > 100 PB
- Very benign I/O (streams, large files, single write / single read), no need for POSIX

(*) Hübner-factor: availability of the accelerator: average ~ 0.3 over a year
However can be up to 0.8 over good days, typically 0.4 – 0.5 during physics production runs

Implementation possibilities

- **Completely decentral in local disks of servers**
  ☺ most likely cheapest
  ☹ not well adaptable to different node speeds / histories, limits node choice, operationally most complicated, no redundancy
- **Fully central**
  ☺ maximum flexibility, easiest to operate, many implementation options (tiered storage)
  ☹ expensive, single point of failure → require high level of redundancy
- **Partially centralized**
  ☺ several hardware options, redundancy relatively cheap, not a single point of failure depending on granularity, covers node differences well
  ☹ probably more costly than local disks, more complicated to operate



Example numbers,
Not binding ☺

More storage facts

- Asynchronous 2nd phase of processing allows maximising utilization of compute resources → exploit non-availability of accelerator → up to a factor 3(!) at equal cost, if sufficient buffering
- Practically no assumptions about storage in the application: can be solid-state, spinning, could be even tape ☺, can be shared with other users/experiments as long as performance is guaranteed (IMHO a prime example of a synergy potential between IT and experiments and industry)

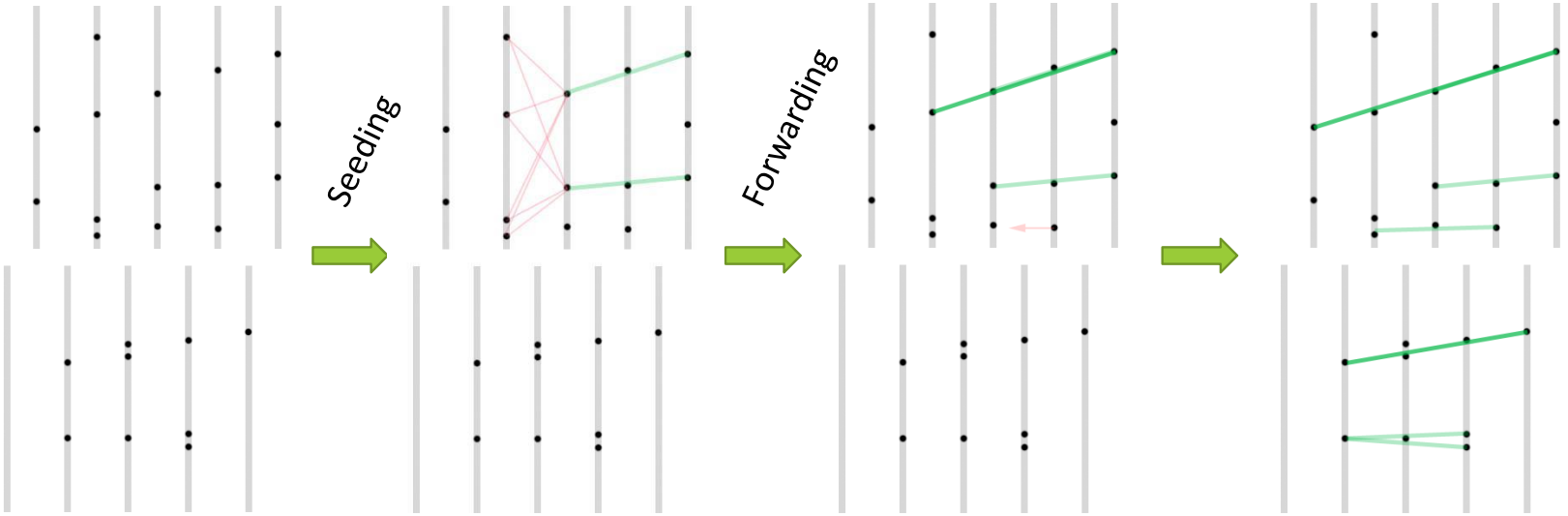
COMPUTE

Data-processing: example track-finding in LHCb

Iterative algorithm that finds straight lines in collision event data in VeloPixel sub-detector

Triplets of hits with best criterion are searched (seeding)

Triplets are extended to tracks if a fitting hit can be found



Flexible infrastructure

Ideally could seamlessly switch between or run side-by-side

- batch-type (simulation, analysis "offline") and
- (near) real-time workloads

Require easy access to disk-storage

High speed network between (some) servers

Housing of custom I/O cards (PCIe)

Flexible amount of accelerators (FPGA, GPGPU, Xeon/Phi), should also be easily assigned to different work-loads

→ and all this ideally in the same infrastructure, easily reconfigurable

→rack-level and data-centre oriented design

A building-block for Online Compute

Required hardware for efficient online processing profits from locality and sharing of some resources

Currently we organize entire racks into “pods”

Ideal granularity could be a bit smaller:

- 8 – 16 servers
  - 25/50 Gbit/s network, dual or single socket CPU
- Internal network over back-plane with very high-speed uplinks (200G/400G → reduce cabling
- Shared storage
  - Cost effective redundancy (local drives need mirroring)
  - Spinning drives (capacity)
  - Deployed as NAS (cost)
- Flexible amount of accelerators (GPGPUS, FPGAs, etc…)
  - Not 1:1 ratio with servers
- Optional: possibility to plug custom PCIe card (and link to server) (for DAQ)