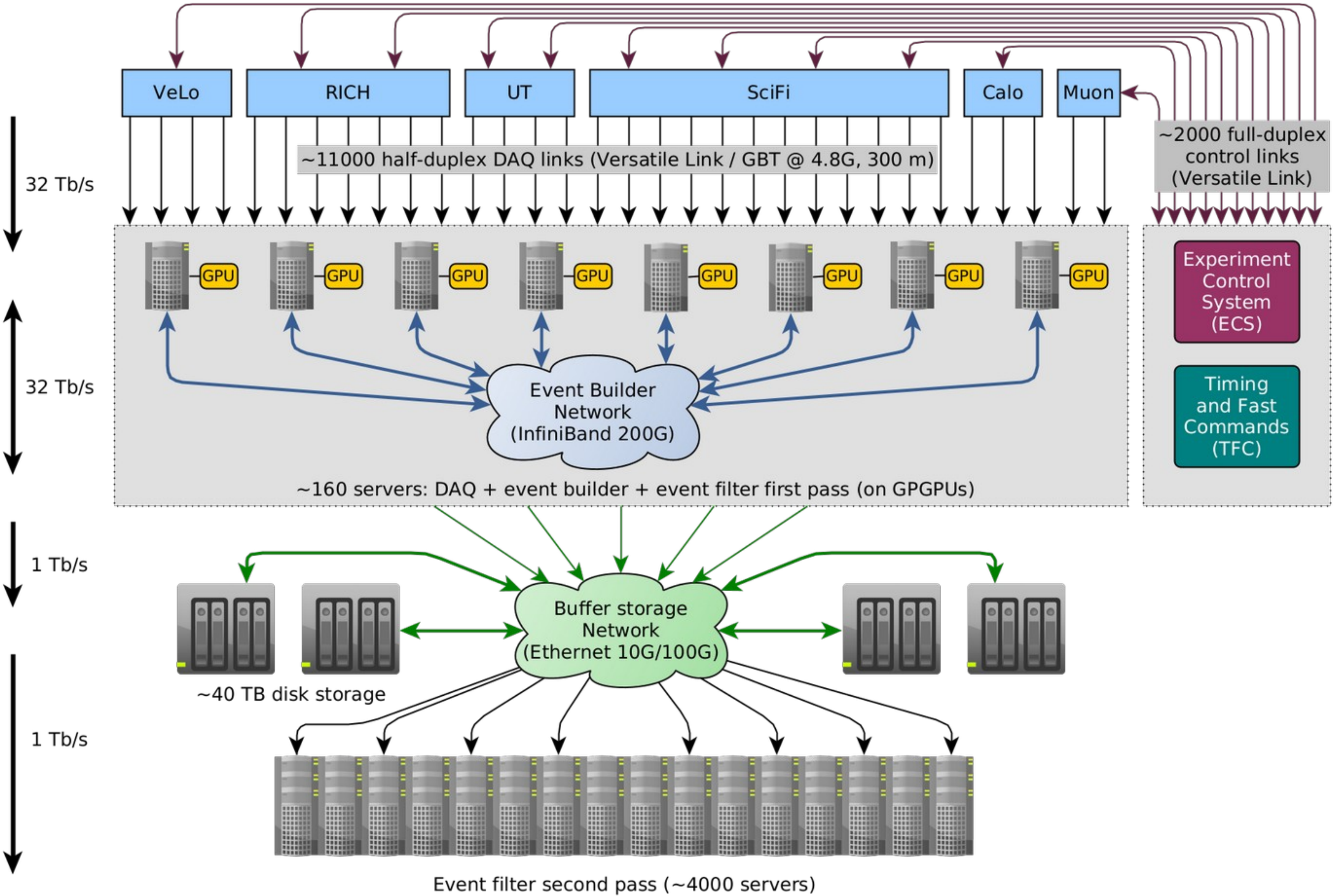
ONLINE SYSTEM
THE NEXT GENERATIONS

Tommaso Colombo
CERN

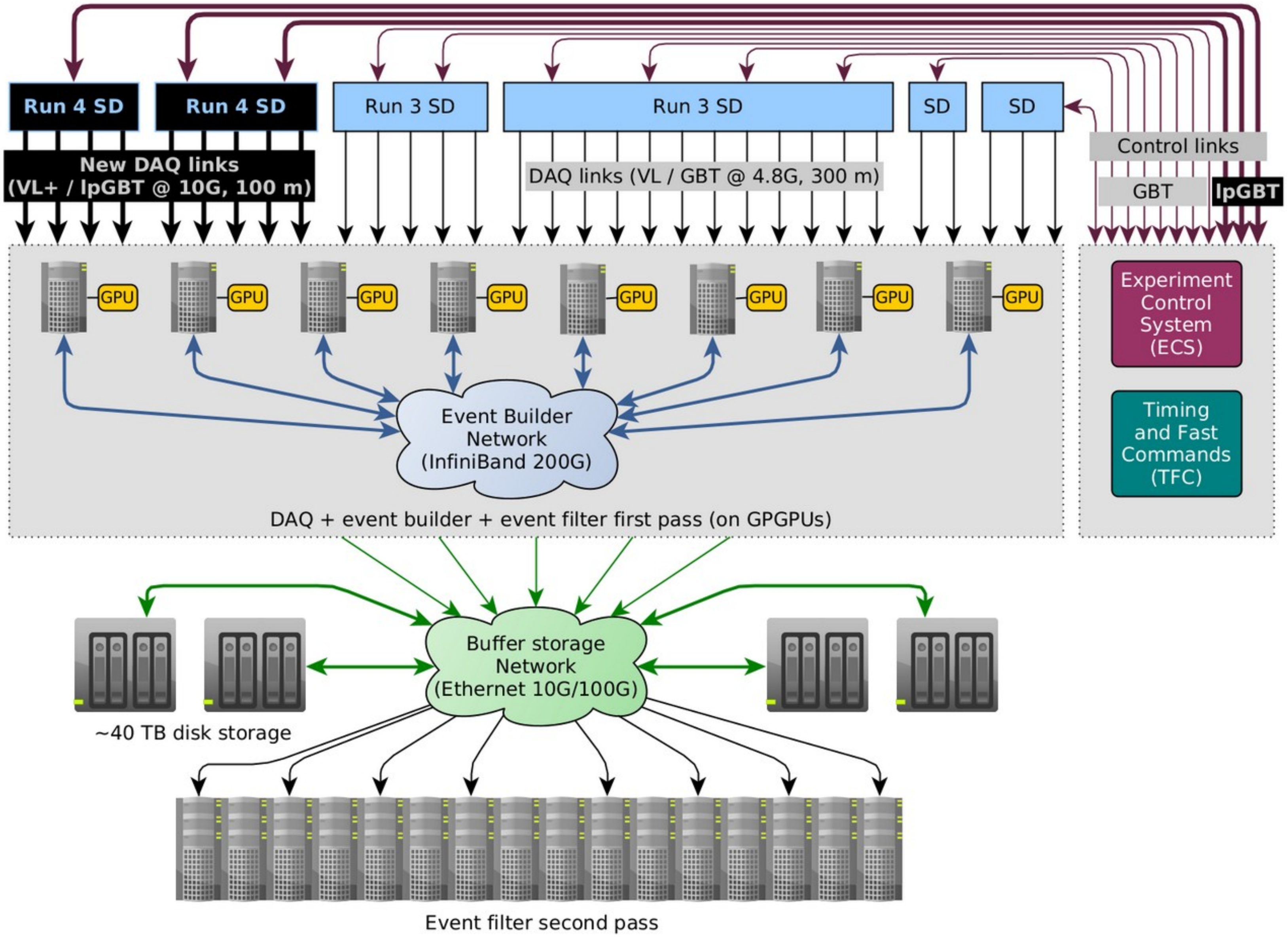
6th Workshop on LHCb Upgrade II
Barcelona, 30 March 2023

LS3 ENHANCEMENTS
OVERVIEW

Run 3 architecture


VeLo
RICH
UT
SciFi
Calo
Muon
~11000 half-duplex DAQ links (Versatile Link / GBT @ 4.8G, 300 m)
~2000 full-duplex control links (Versatile Link)
32 Tb/s
GPU
32 Tb/s
Event Builder Network (InfiniBand 200G)
Experiment Control System (ECS)
Timing and Fast Commands (TFC)
~160 servers: DAQ + event builder + event filter first pass (on GPGPUs)
1 Tb/s
Buffer storage Network (Ethernet 10G/100G)
~40 TB disk storage
1 Tb/s
Event filter second pass (~4000 servers)

Run 4 architecture

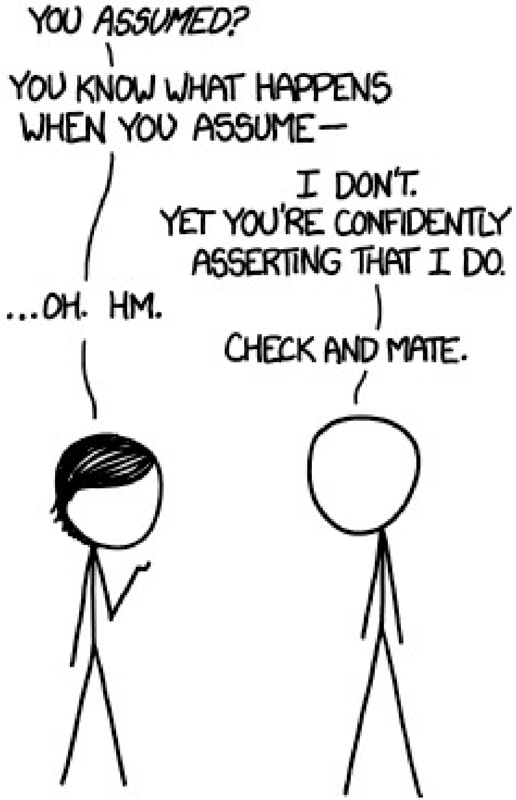


Assumptions

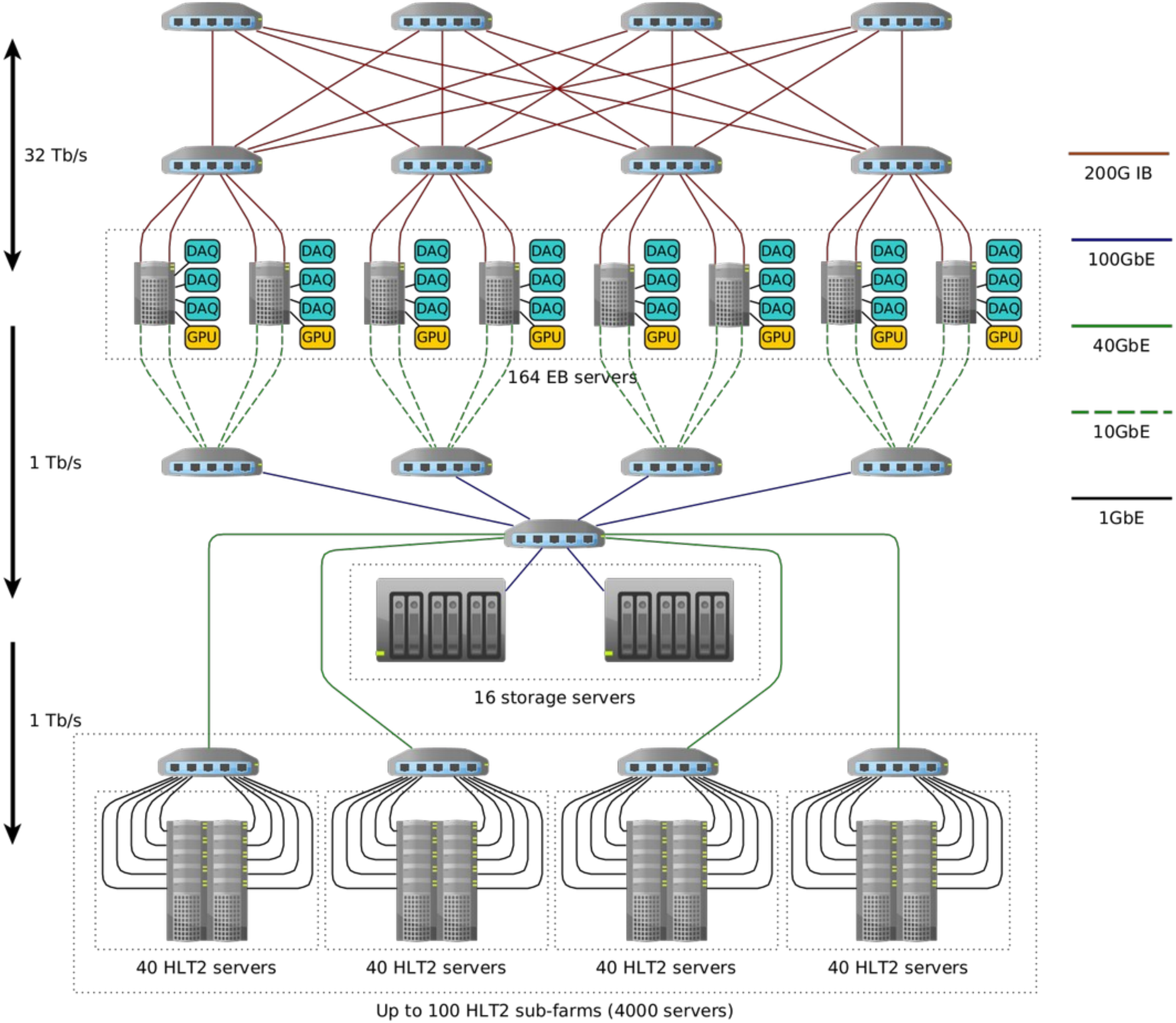
- “Enhanced” sub-detectors will standardize on lpGBT
- Versatile Link + / lpGBT cannot reach the surface data center

  → DAQ boards in cavern
- EB network links can reach cavern
- $\frac{\text{Run 4 EB throughput}}{\text{Run 3 EB throughput}} < 133\%$
- Rest of Online system handles higher throughput with no substantial changes

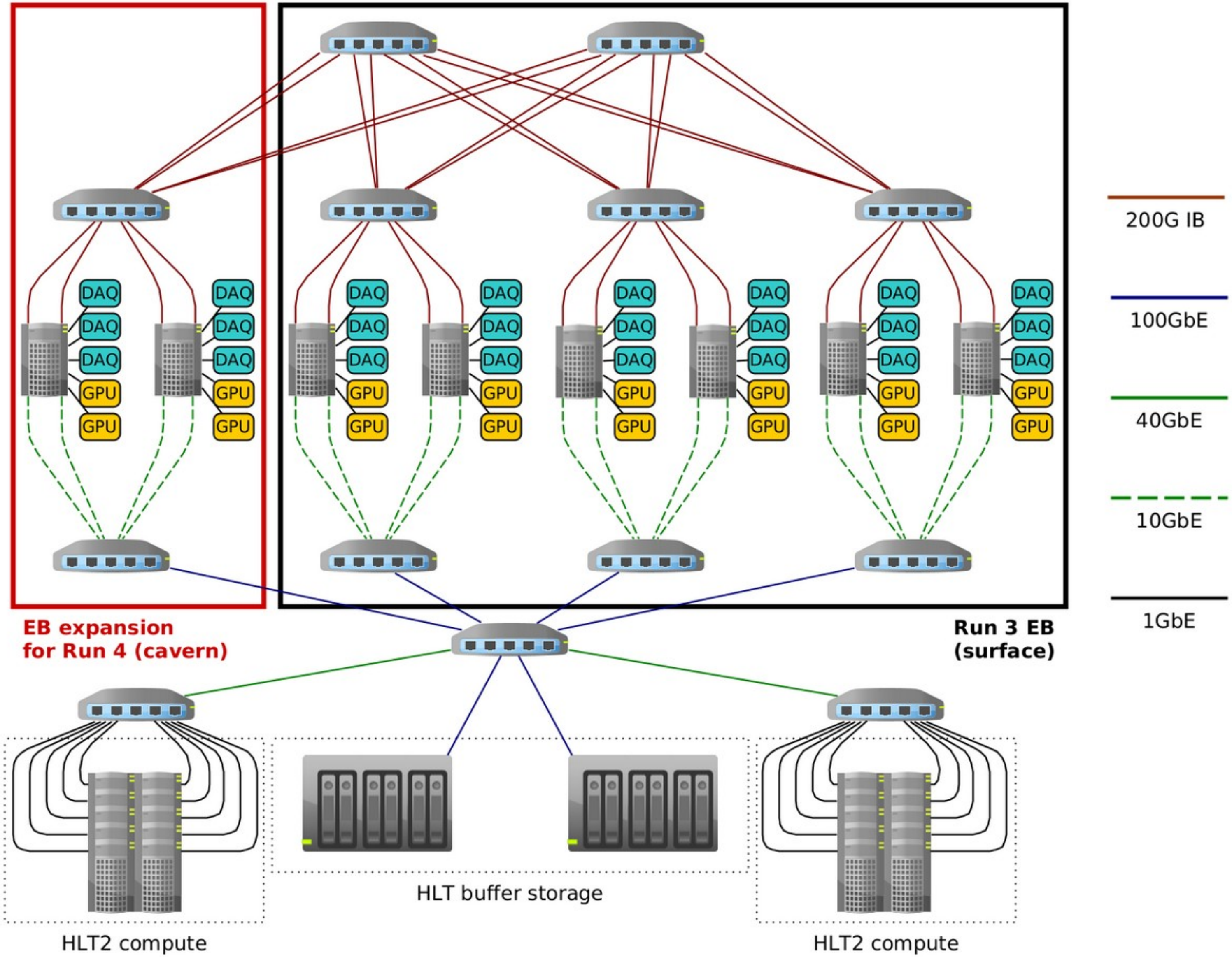
  → Extending the Run 3 EB is easier/cheaper than building a new one in the cavern


YOU ASSUMED?
YOU KNOW WHAT HAPPENS WHEN YOU ASSUME—
I DON'T. YET YOU'RE CONFIDENTLY ASSERTING THAT I DO.
...OH. HM.
CHECK AND MATE.

Run 3 system


32 Tb/s
1 Tb/s
1 Tb/s
DAQ
GPU
164 EB servers
16 storage servers
40 HLT2 servers
40 HLT2 servers
40 HLT2 servers
40 HLT2 servers
Up to 100 HLT2 sub-farms (4000 servers)
200G IB
100GbE
40GbE
10GbE
1GbE

Run 4 system


DAQ
DAQ
DAQ
GPU
GPU
200G IB
100GbE
40GbE
10GbE
1GbE
EB expansion for Run 4 (cavern)
Run 3 EB (surface)
HLT buffer storage
HLT2 compute
HLT2 compute

LS3 ENHANCEMENTS

~~BOXES ON A SLIDE~~

REAL SYSTEM

What we need from sub-detectors

- Amount of links (DAQ & ECS/TFC)
- Throughput
  (IN and OUT of the DAQ boards)
- Optical fibers plans

Contact: me

- DCS plans

Contact: Clara Gaspar, Luis Granado Cardoso

- Data formats
- Firmware plans

Contact:
Guillaume Vouters,
Paolo Durante

PLANNING AND PROCUREMENT
ALWAYS TAKE LONGER THAN EXPECTED.

2026 IS AROUND THE CORNER.

LS3 ENHANCEMENTS

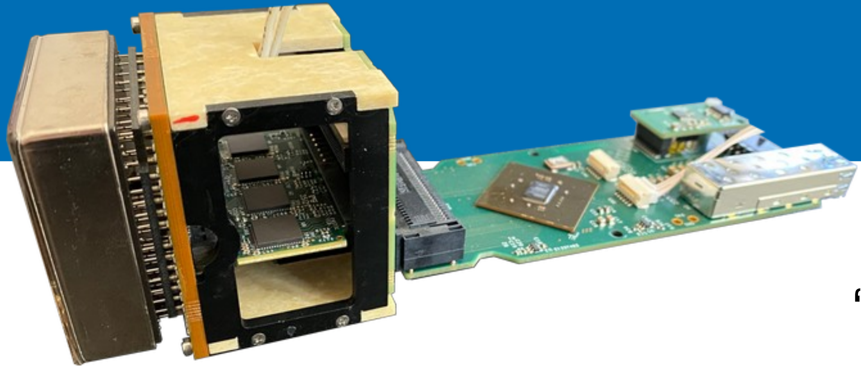
COST

Cost model

- Included:
  - Optical fibers
    from cavern to counting room
  - Additional EB
    servers & switches
    to cover throughput increase
  - Long-distance EB optics
    between cavern and surface

- **Not included:**
  - **ECS fibers, boards, servers**
  - **DAQ boards (PCIe400?s)**
  - **On-detector DAQ fibers**

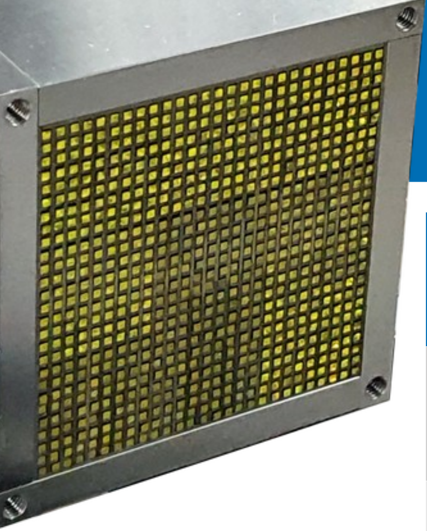
Warning: many optimizations / threshold effects make the relationship between system size and cost non-linear and non-intuitive



RICH

“Likely” scenarios (inspiration: last RICH review)

DAQ fibers	EB input	PCIe40s	PCIe400s	**Online costs**
2500	12.5 Tb/s		63	**450** kCHF
2500	12.5 Tb/s	125		**490** kCHF
2500	11.0 Tb/s		55	**400** kCHF
2500	11.0 Tb/s	105		**340** kCHF



CALO

DAQ fibers	EB input	PCIe40s	PCIe400s	**Online costs**
440	2 Tb/s		10	**150** kCHF
440	2 Tb/s	20		**170** kCHF

Caveat emptor (CALO & RICH)

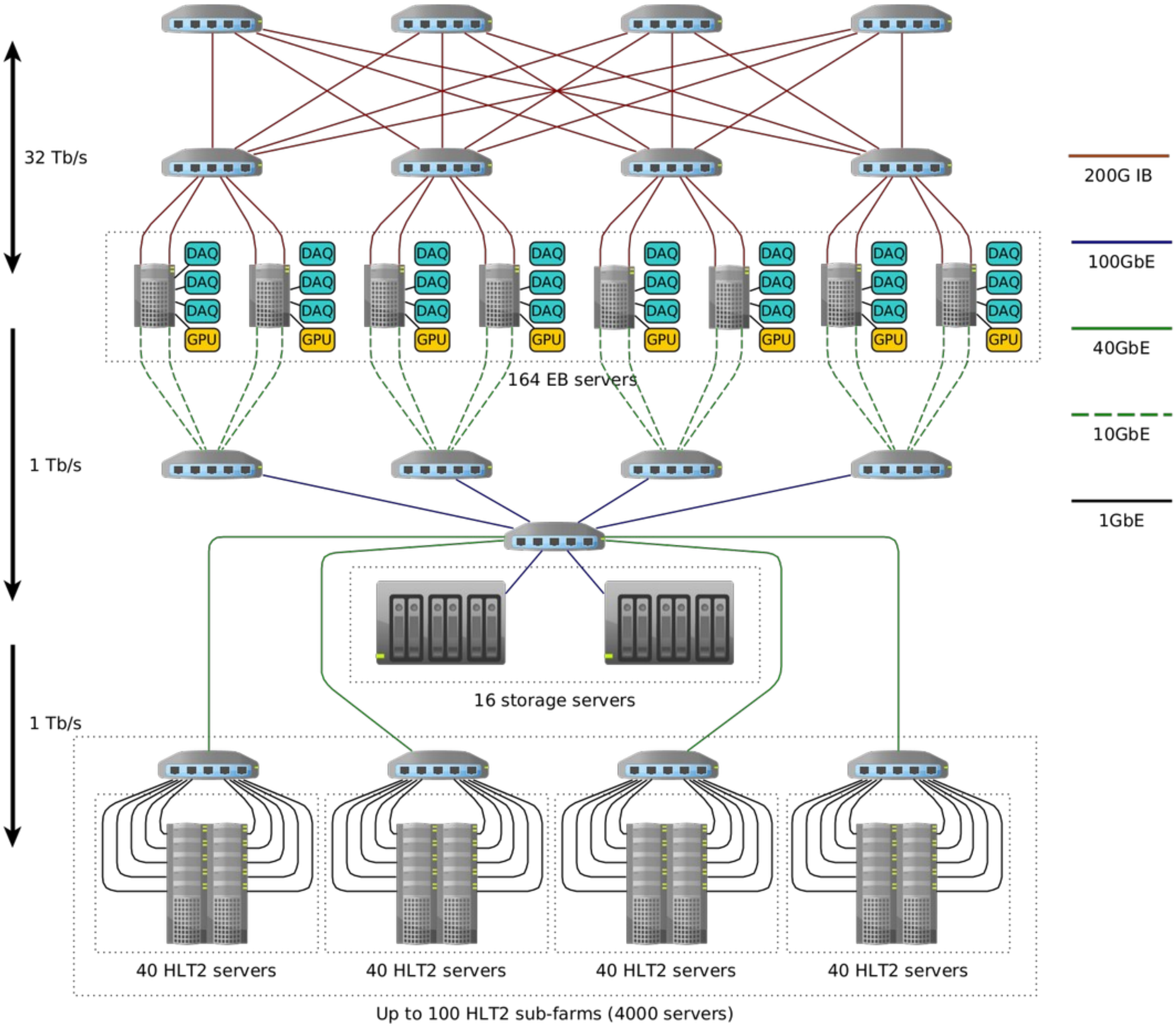
- All scenarios (RICH and CALO) assume PCIe400 output to the EB < 200 Gb/s
- If not, costs increase:
  - More expensive servers needed
  - Cannot reuse existing servers

- Total EB input > 16 Tb/s triggers a threshold effect in the EB network architecture

→ 110 kCHF added costs

UPGRADE 2
OVERVIEW

Current system



Current system: limitations

- It’s “hyper-converged”:
  - Sub-detector data processing (on DAQ FPGAs), readout, event building, and HLT1 compute (on GPUs) all together in the same server!
  - Result of a relentless effort to lower the Online system cost and free up resources to acquire more compute
- There is no such thing as a free lunch:
  - **Not-so flexible HLT1**:
    # of GPUs has to be a multiple of the # of EB servers
    - HLT1 compute power can only be doubled (happening now) or tripled
  - **Brutal DAQ firmware integration testing challenge**:
    - >10 different firmwares, hours to compile just one

Current system: limitations

- It’s “upstairs”:
  - Current 5 Gbps GBT/VL links can reach our surface datacentre
  - **Future 10 Gbps lpGBT/VL+ links will not**
    - Studies by EP/ESE show that 10 Gbps links to the surface only work:
      - For “low radiation” detectors (i.e. not VeLo, UT, or Mighty Tracker)
      - With highly-sensitive receivers (i.e. unobtanium)
- High EB performance is achieved through rigid scheduling of data transfers on the EB network:
  - **Low fault tolerance**:
    a few “well placed” link failures can severely lower the maximum EB rate

Requirements: what we know now

- Around **3×** data links: 11000 → 30000
- All lpGBT/VL+: 5 Gbps → 10 Gbps
- **5×** readout throughput: 32 → 160 Tbps
- **5×** output to tape: 80 → 400 Gbps

This presentation focuses on the components that will change substantially to handle the new requirements.

Other parts
(dataflow software, infrastructure services, ECS)
will follow a more natural evolution.
They are equally mission-critical,
so their continuous maintenance must be ensured.

Event builder for U2

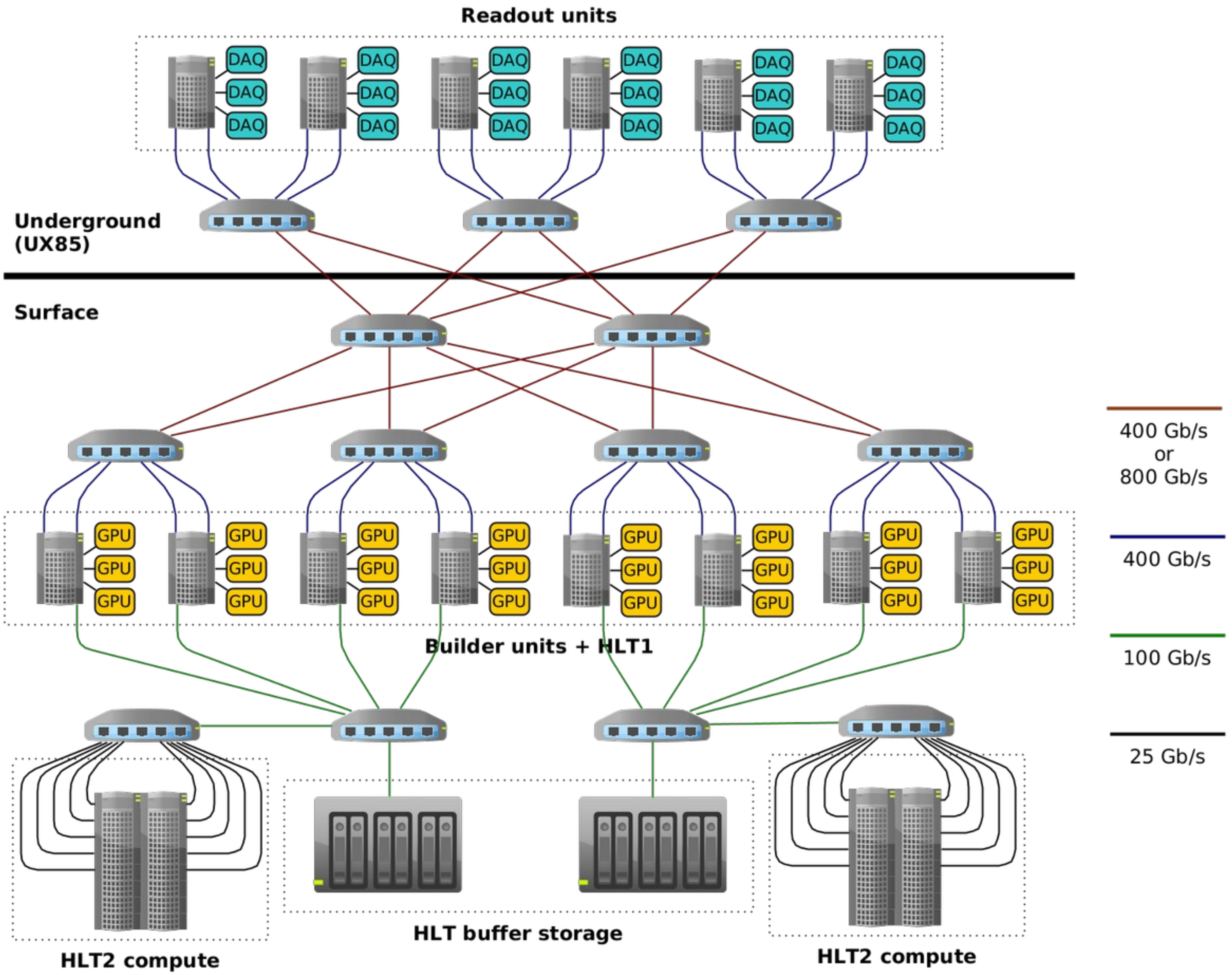
- OPTION:
  reproduce the current design, but in the cavern:
  - Cheaper (fewer servers, fewer network ports)
  - More challenging (very high server memory bandwidth use)
  - Not very flexible for HLT1 compute or HLT1 accelerators
- OPTION:
  readout servers in the cavern, *builder+filter* servers in the surface:
  - More expensive
  - Less risky
  - More flexible

Preferred option.

Can it be done at reasonable cost?

Must keep up with tech evolution and prototype early.

Upgrade 2: “split” EB

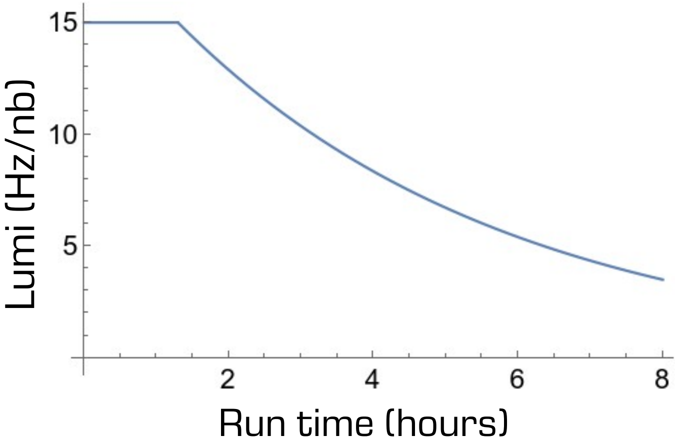


UPGRADE 2

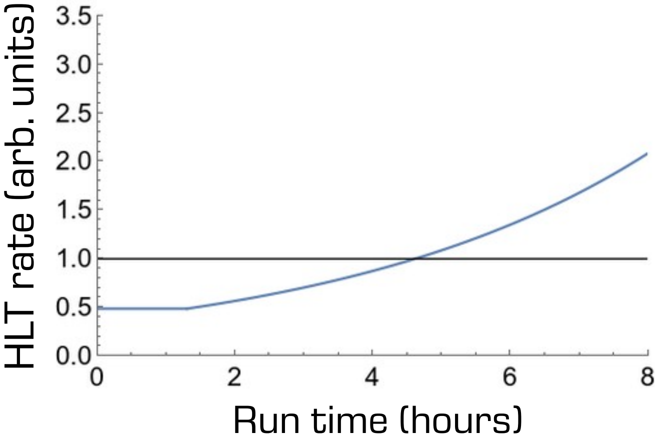
WHAT IFS

Luminosity curve & HLT throughput

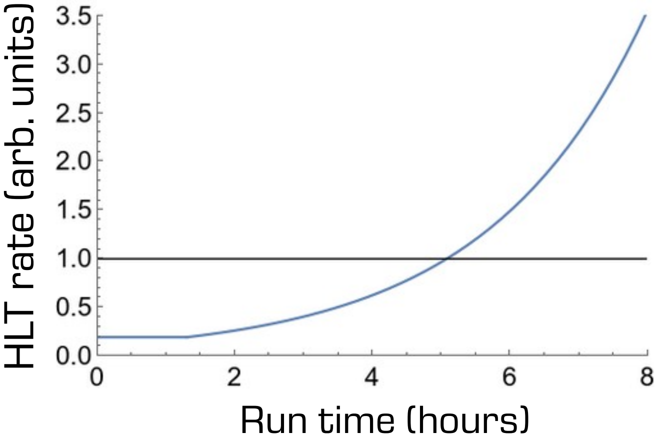
In Run 5, the luminosity provided to LHCb will look like this:



HLT throughput, assuming event processing time scales **linearly** with lumi:



HLT throughput, assuming event processing time scales **quadratically** with lumi:



Can storage help flattening it?

A TOY MODEL

- Run parameters [Framework TDR + guesswork] :
  - Lumi-levelled time = 1.3 hours
  - Total run time = 8 hours
  - Lumi = from 15 to 3.5 Hz/nb
  - EB throughput = 20 TB/s
- SSDs in 2030 [citation needed] :
  - Capacity: 100 TB
  - Write speed: 10 GB/s
  - Endurance: >2 DWPD

HLT vs. lumi	Min. space	Min. speed	SSDs needed
Linear	120 PB	10 TB/s	1200
Quadratic	210 PB	16 TB/s	2100

- Feasible? Maybe. Cheap? Likely not.
- Must evaluate cost/benefit in 2030

Accelerators

- For optimization reasons, the new system will still require all *builder+filter* servers to process an identical share of the events

  → Same amount of "acceleration" must be installed in each server

- Different approaches are possible, but will add significant costs

- Impact of different accelerators (IPUs, FPGAs, NPUs, GPUs, …) on total system cost and complexity is highly dependent on the specific accelerator requirements

  → **We can help modeling this**

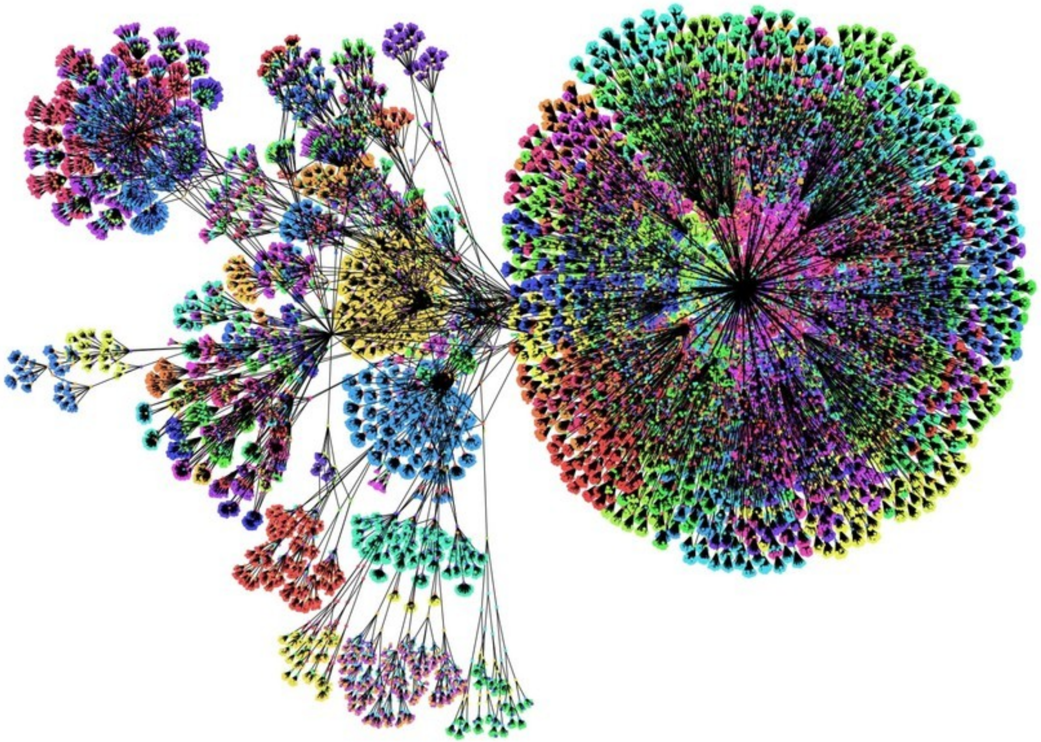
FIN.

LS3 ENHANCEMENTS

DETAILS

ECS

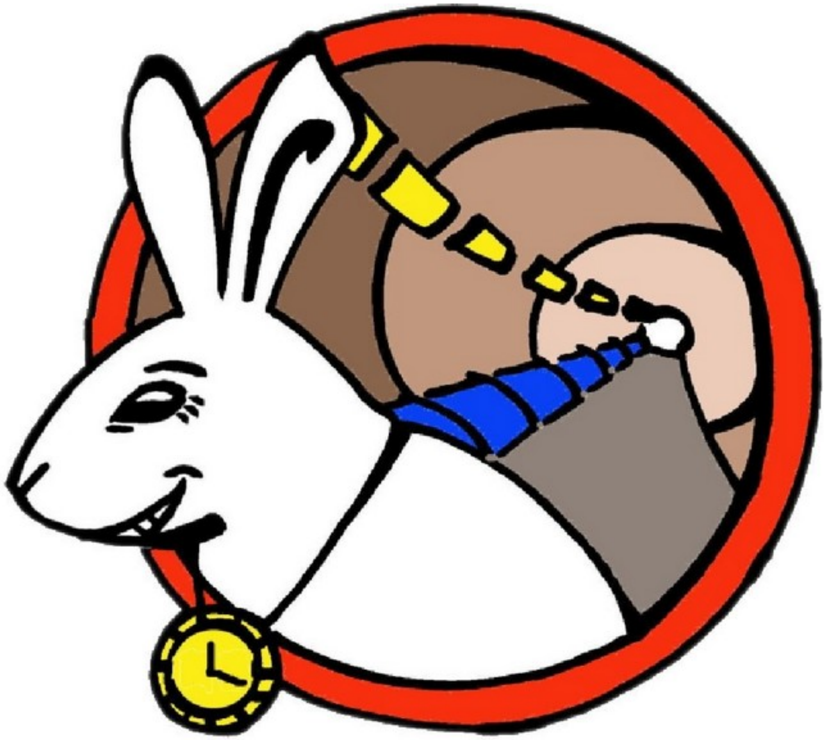
- System architecture and technology will stay the same
- **Please engage early with ECS experts!**
  - Use common interfaces and equipment for DCS
  - Integrate DCS, DAQ, Data Quality in the same global system



TFC

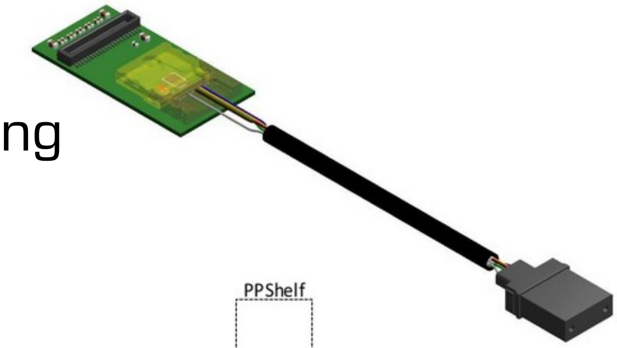
- New clock and timing distribution from the LHC:
  White Rabbit (Ethernet based)
  - New LHC clock Hardware interface
  - New Online clock distribution system
- **Sub-detectors mostly unaffected**

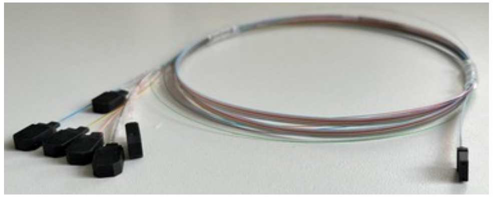
  → TFC interface (via control links) stays the same

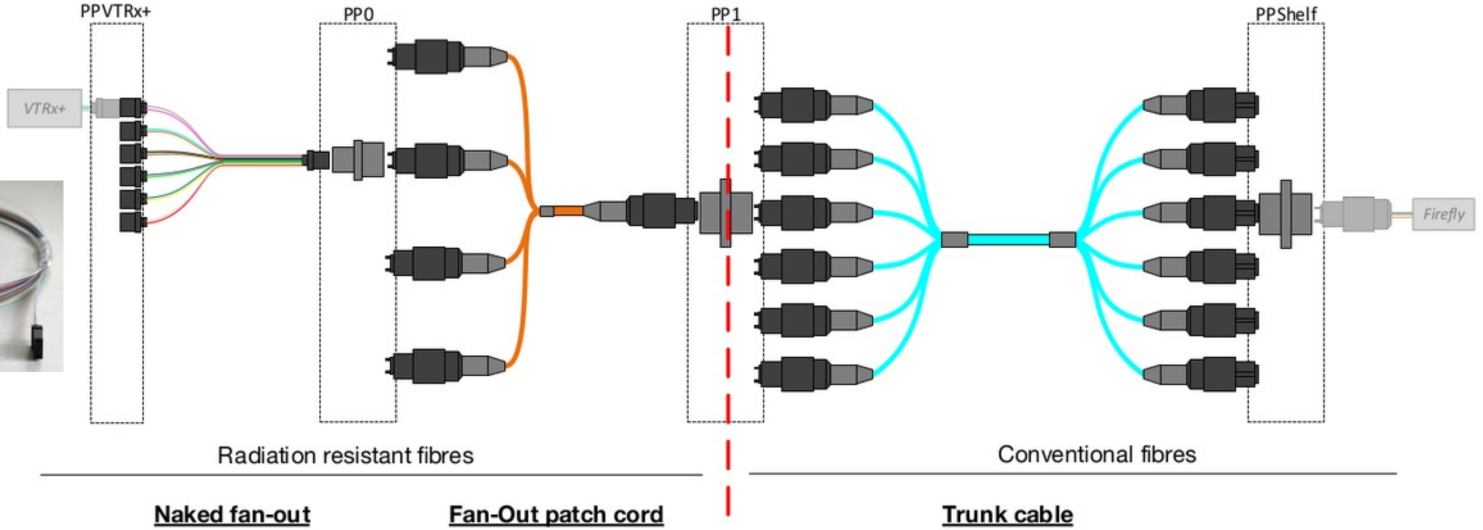


Optical fibers

- The VL+ cable plant is even more complicated than in Run 3
  - VTRx+ has 4 Tx + 1 Rx links on an MPO-12 pigtail
  - Needs custom fan-outs for concentrating/reshuffling
  - Procuring them will have a long lead time







Radiation resistant fibres

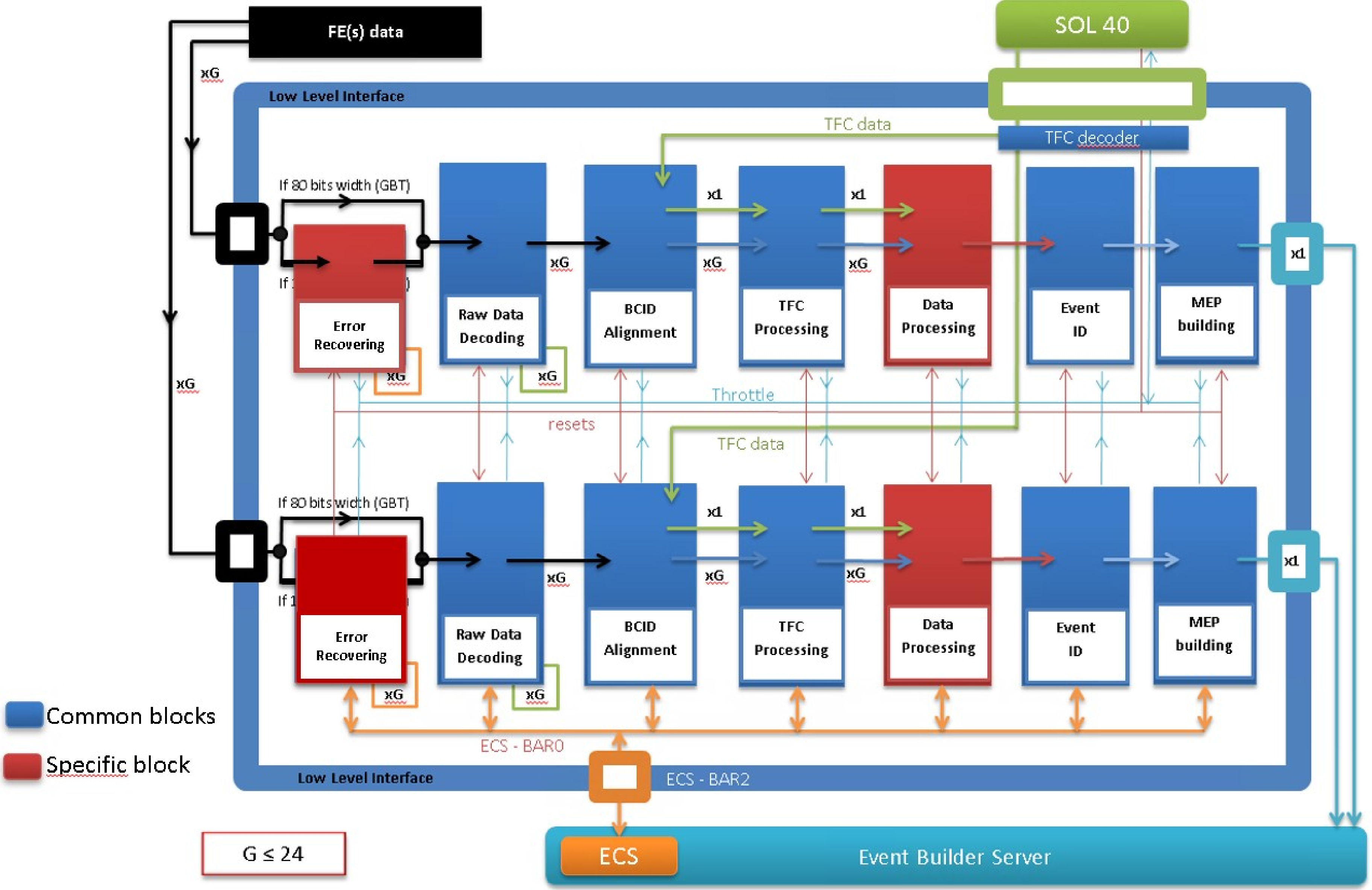
Conventional fibres

Naked fan-out

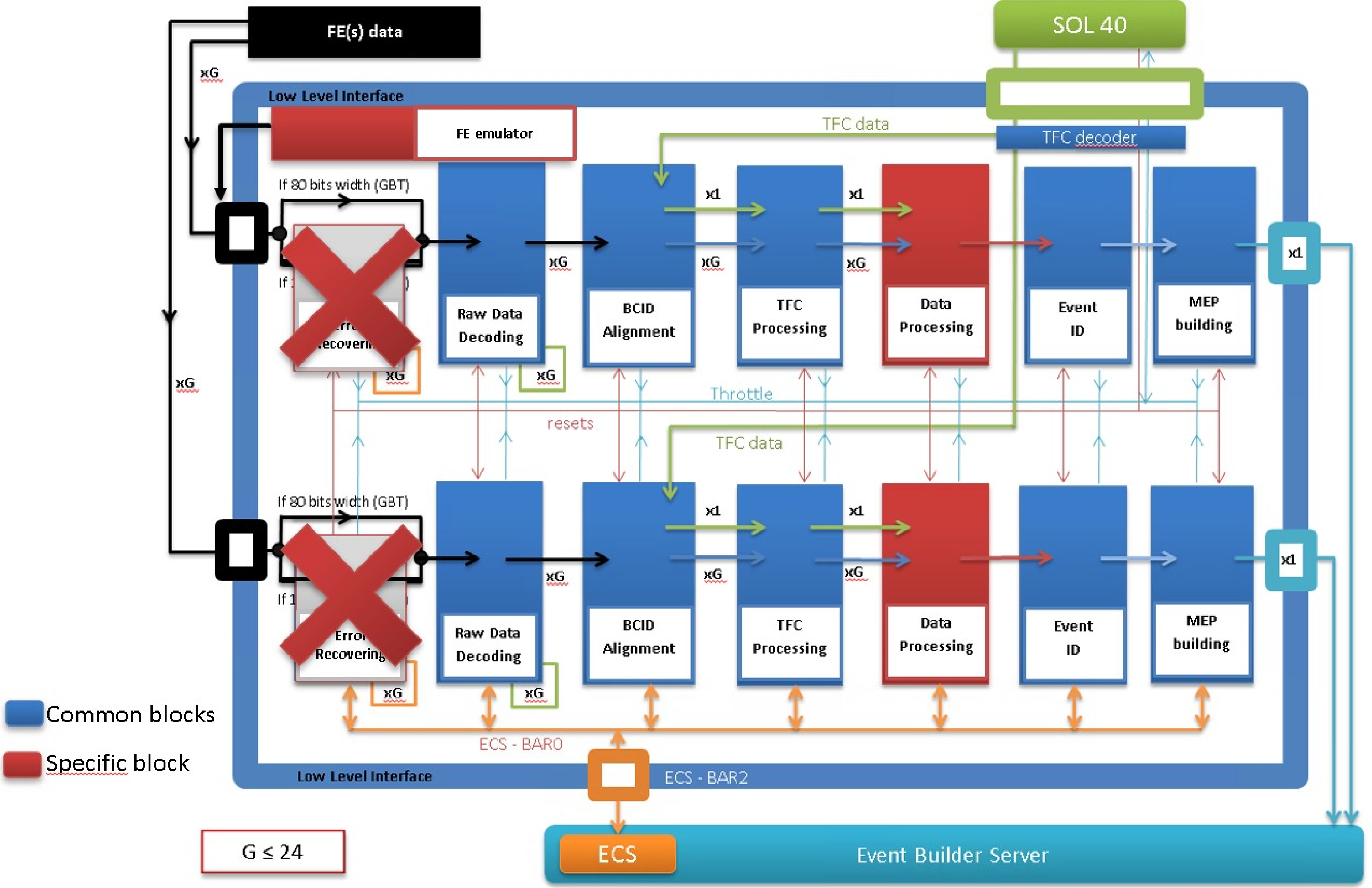
Fan-Out patch cord

Trunk cable

Run 3 DAQ firmware



Run 4 DAQ firmware



Data formats

FE data format impacts:

- FE ←→ BE bandwidth
- Number of links/boards
- Firmware encoding/decoding complexity
- Firmware data processing complexity
- BE data format complexity

- BE data format impacts:
  - BE output bandwidth
  - Number of boards
  - HLT decoding complexity
- **Engage with DAQ experts early and aim for simplicity!**

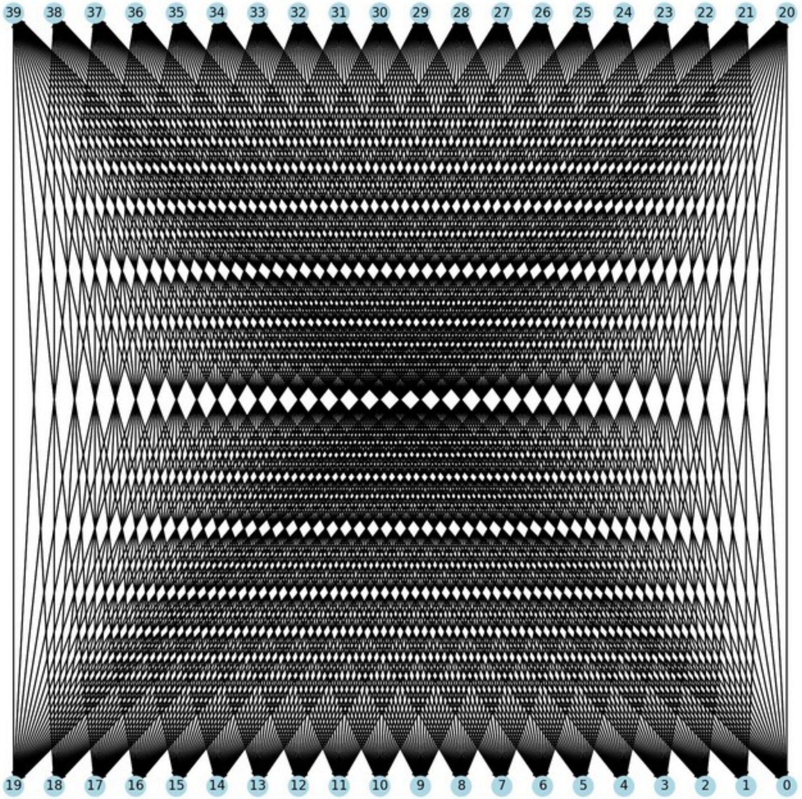
2 documents required per SD group:
¤ FE data format
¤ TELL40 data processing
Run 3 documents here:
https://edms.cern.ch/project/LHCB-0936

UPGRADE 2

R&D

Event builder fault tolerance

- Advanced EB algorithms:
  - Goal: **almost no loss of throughput in case of multiple link failures** on the EB network
  - Adaptive communication scheduling algorithm
  - Adaptive routing solution
- Now: focus on the current EB network topology
- Next: extend it to the more complex topologies of U1b and U2 networks



Idea: lpGBT aggregation

- Rad-hard front-end links (lpGBT) are too slow for modern FPGAs

  → A “mildly-rad-hard” aggregation board could help with more efficient use of back-end electronics (i.e. buying fewer expensive FPGAs)

- Collaboration with EP/ESE:
  - Fits in well with the nascent DRDT 7 collaboration (outcome of the ECFA roadmap process)
- Concept studies can be done with FPGA dev kits (no need to build custom hardware)
- Cost/benefit analysis after prototyping will tell us if it saves us more money than it costs