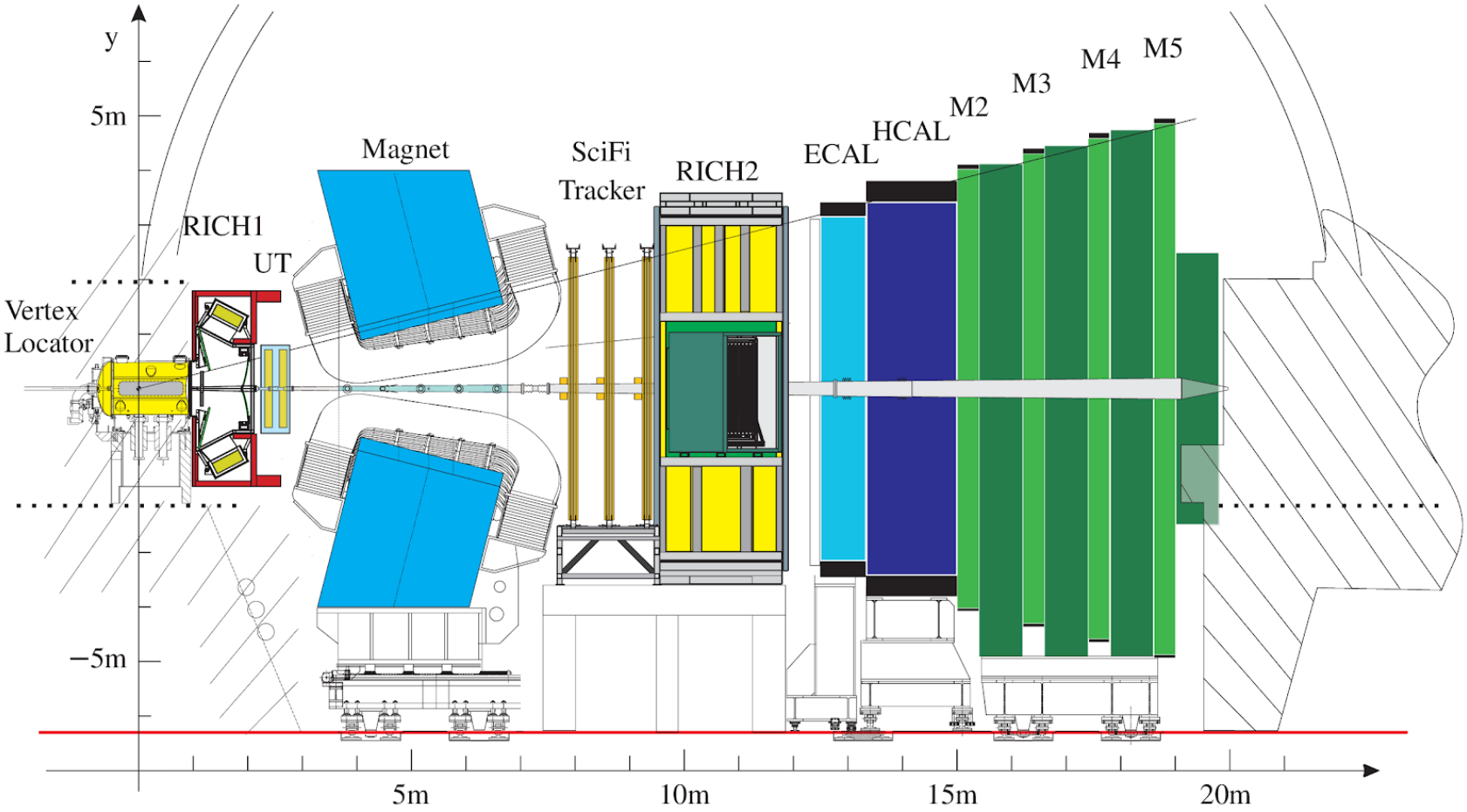
LHCb Data Acquisition Network

Tommaso Colombo
Niko Neufeld

CERN, EP

4th SIG-NGN Meeting
Jan. 16th 2020
CERN, Geneva, Switzerland

The LHCb Experiment



Some facts about LHCb

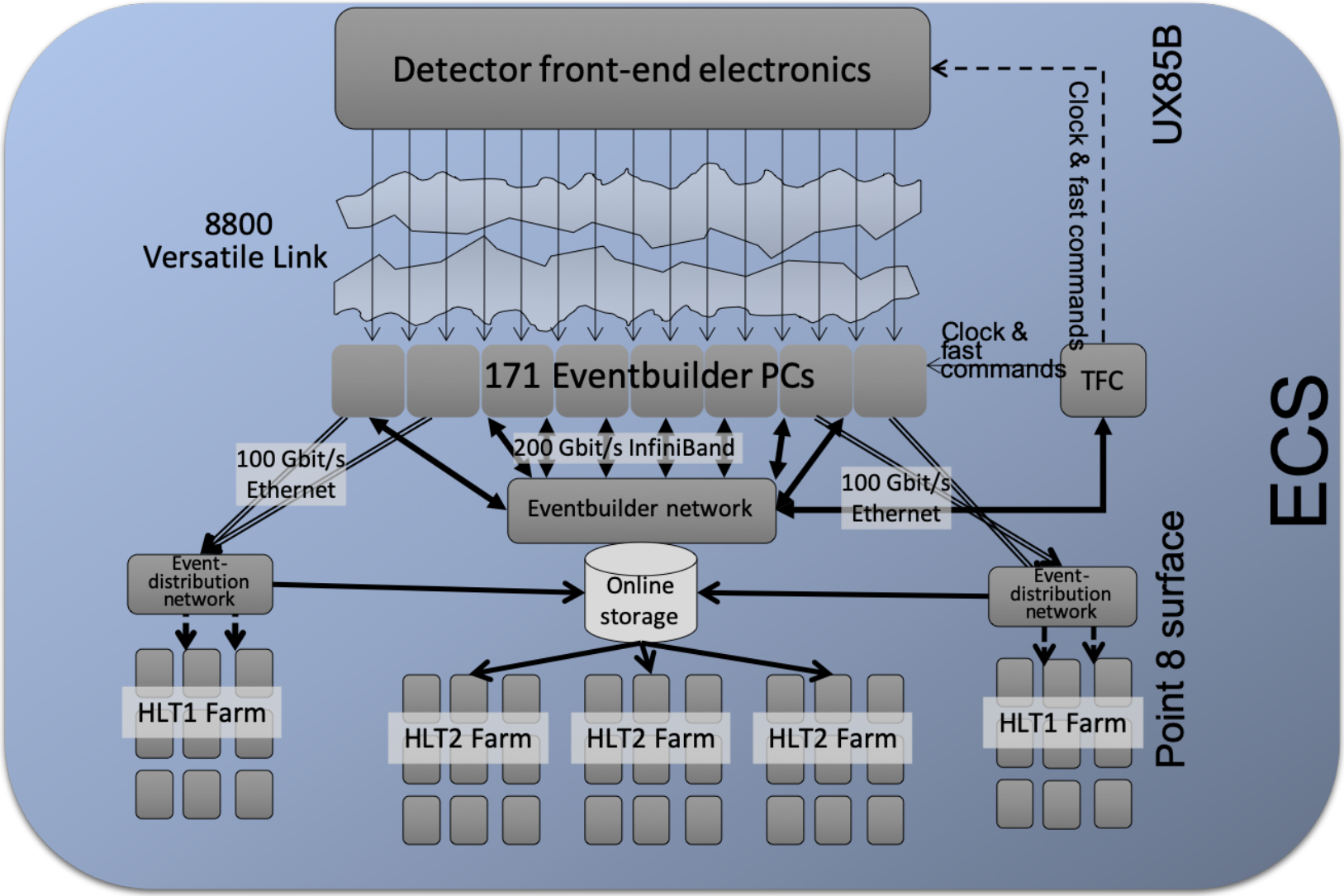
- In the LHC proton-bunches collide every 25 ns
- LHCb will read out the *entire* detector for every collision
- Aggregated data from one collision are approximately 100 kB in size

- The data arrive on $\sim$ 10000 optical fibres
- Thus each fibre on average contains 8 to 10 bytes of data for every collision
- They are collected into 478 FPGA receiver cards (called "TELL40")

Data collection and combination - "event-building"

- Needs to collect data from 478 TELL40 FPGA boards into a single "location"
- And hand them over to compute units for further processing
- **The rest of this talk is about how we combine the data before the compute**

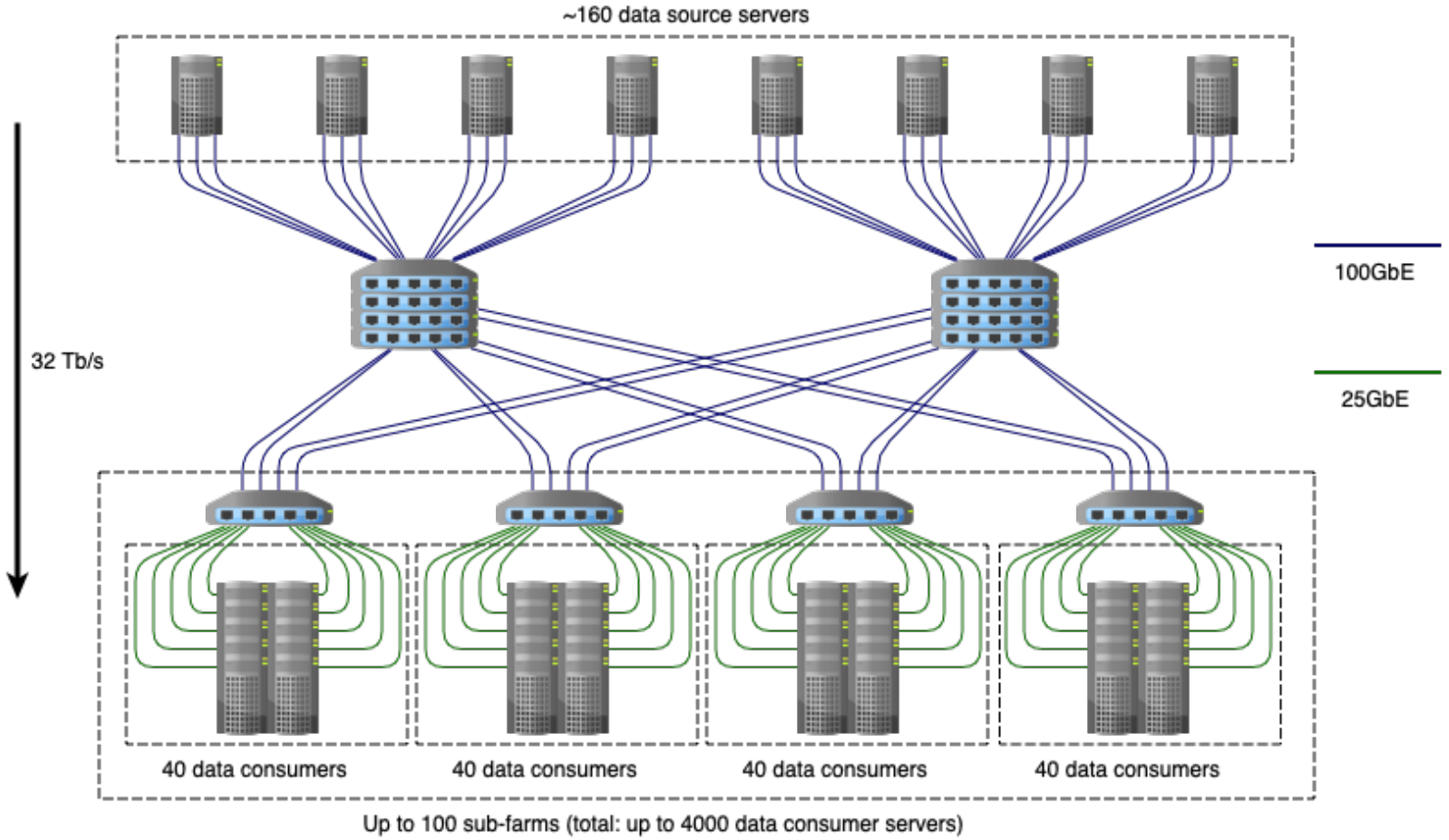
Schematic vie of the "Event-builder"



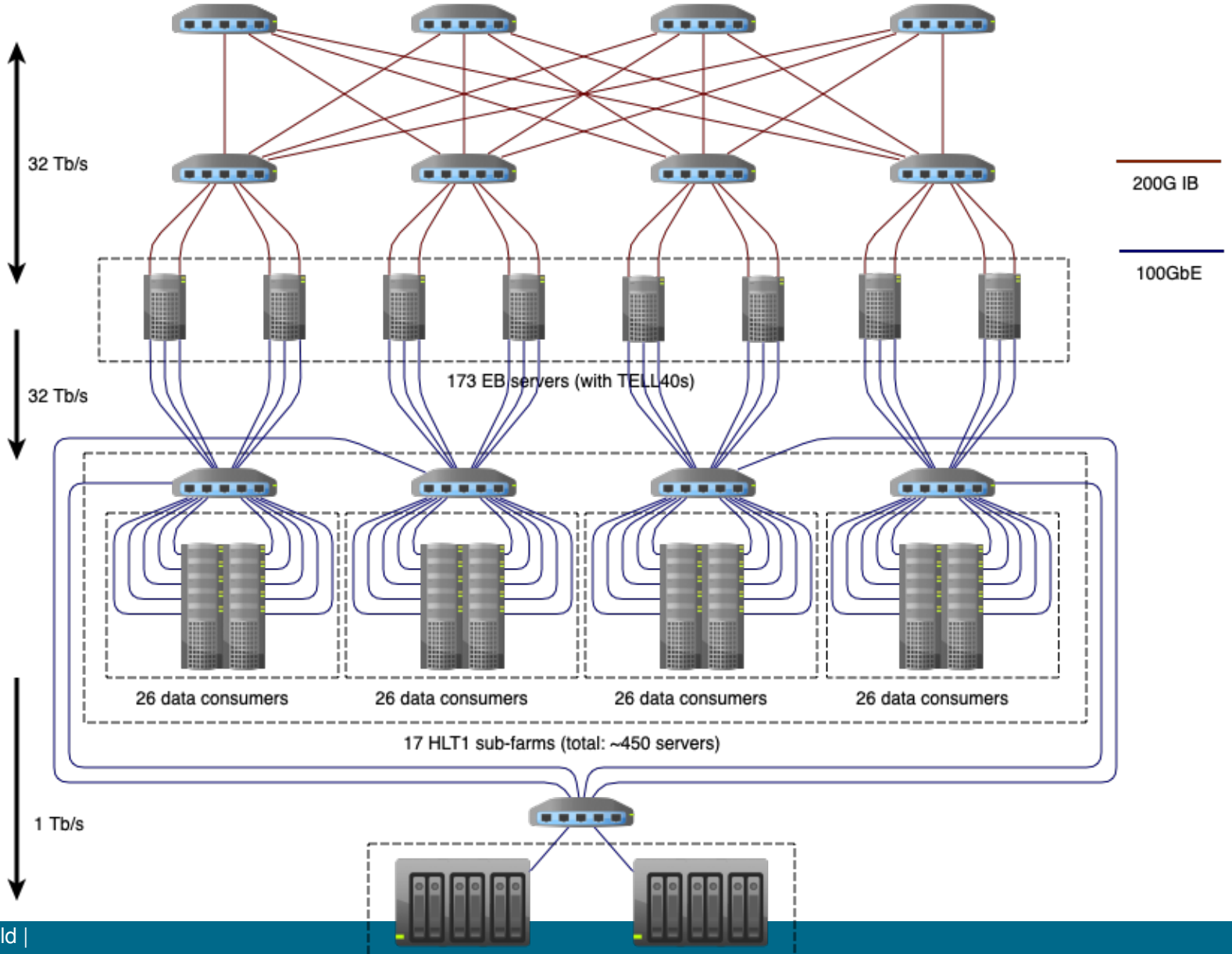
Technical Challenges

1. I/O in the server hosting the TELL40s (event-builder server / EB-server)
2. Scalability of the network, which is composed of several individual network switches
3. Limit the costs by pushing for a compact system at relatively high link-load (which increases I/O and makes scaling more difficult)

Single direction large Ethernet network



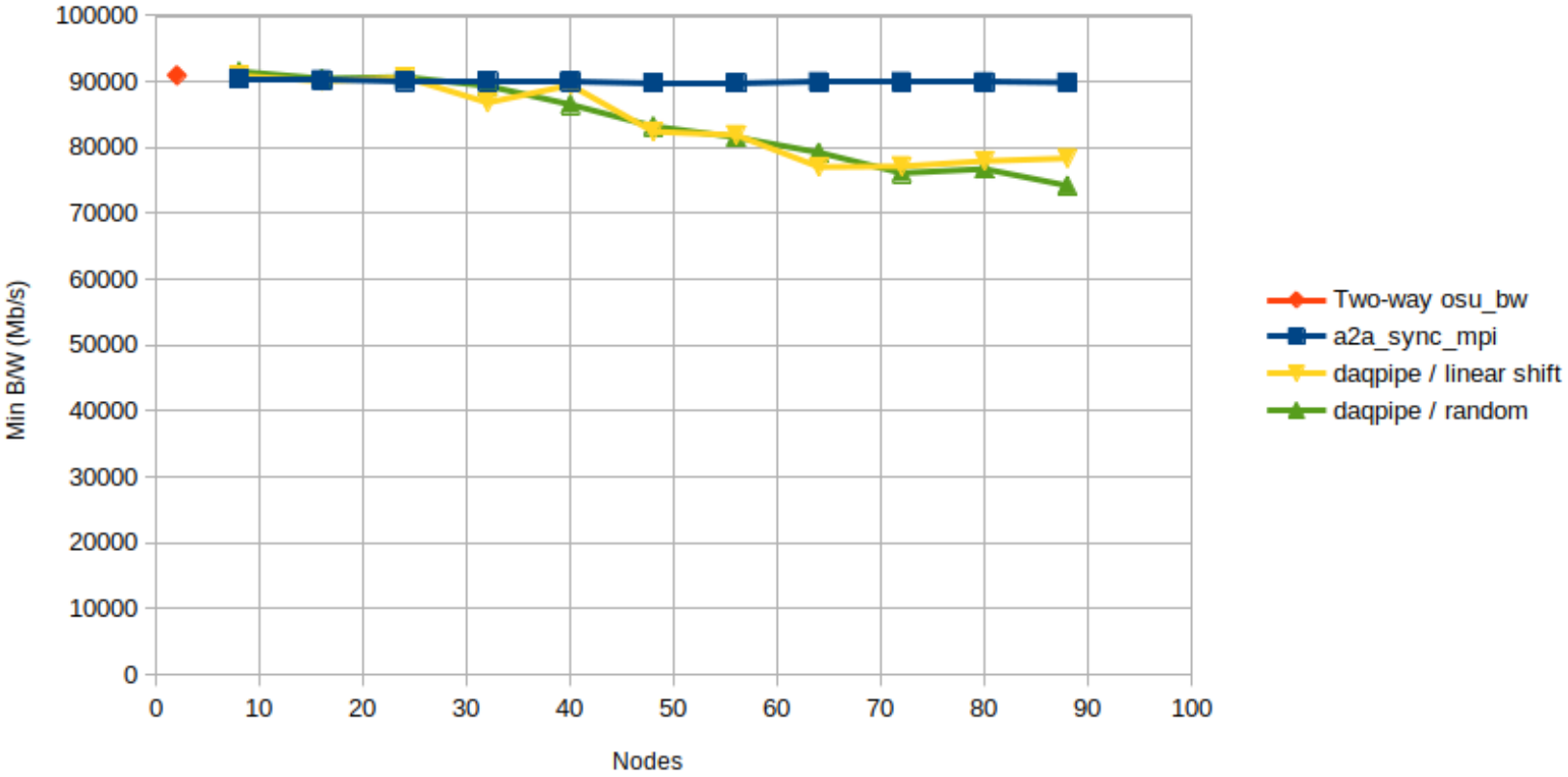
All-to-all InfiniBand network with Ethernet distribution



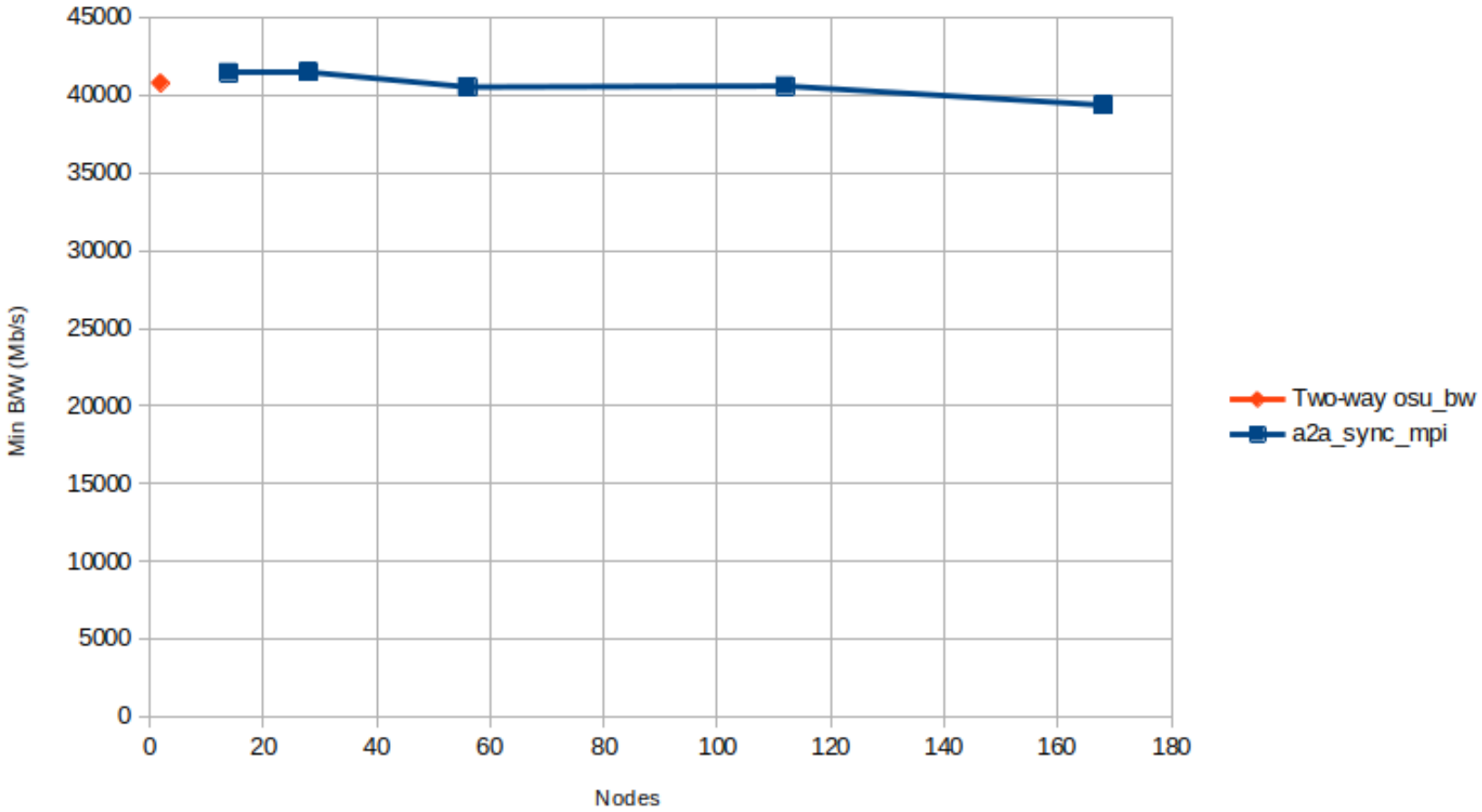
Challenges for the network

- Want high link-load (cost)
- Traffic is inherently bursty
- Want to use some kind of remote DMA to reduce server-load

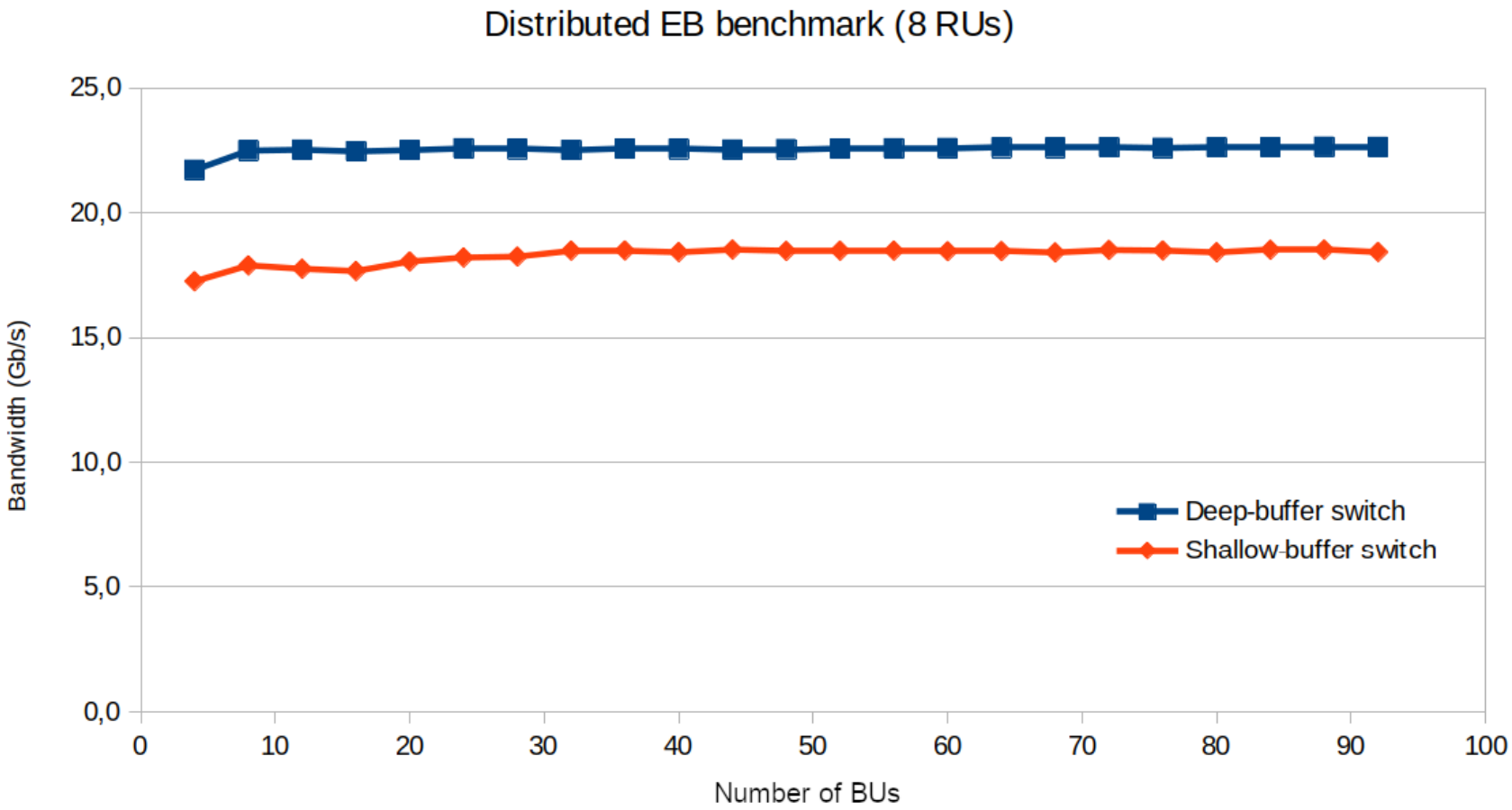
Scalability InfiniBand



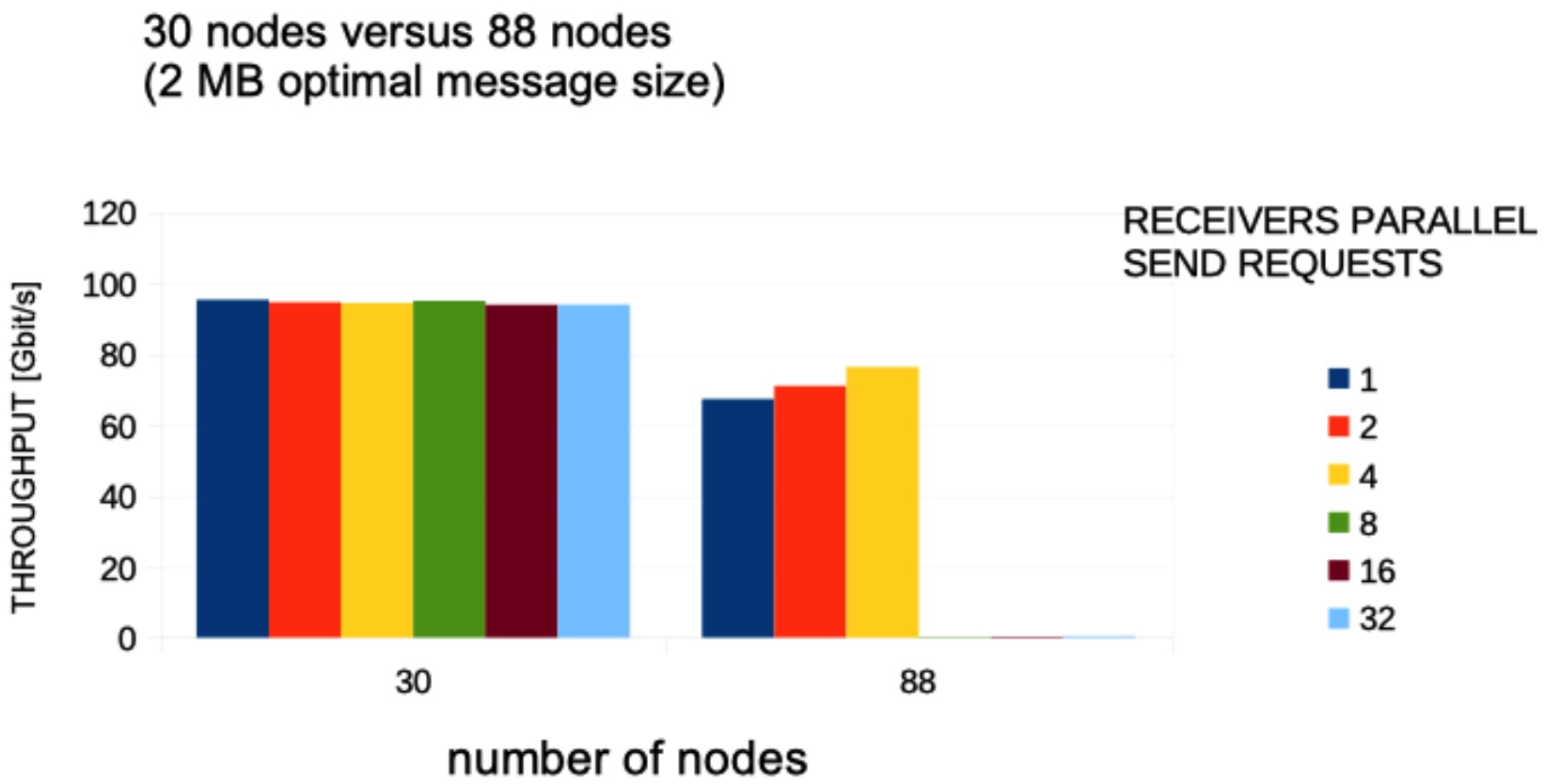
Scalability InfiniBand



Scalability Ethernet (shallow vs deep buffers)



Scalability Ethernet (deep buffers)



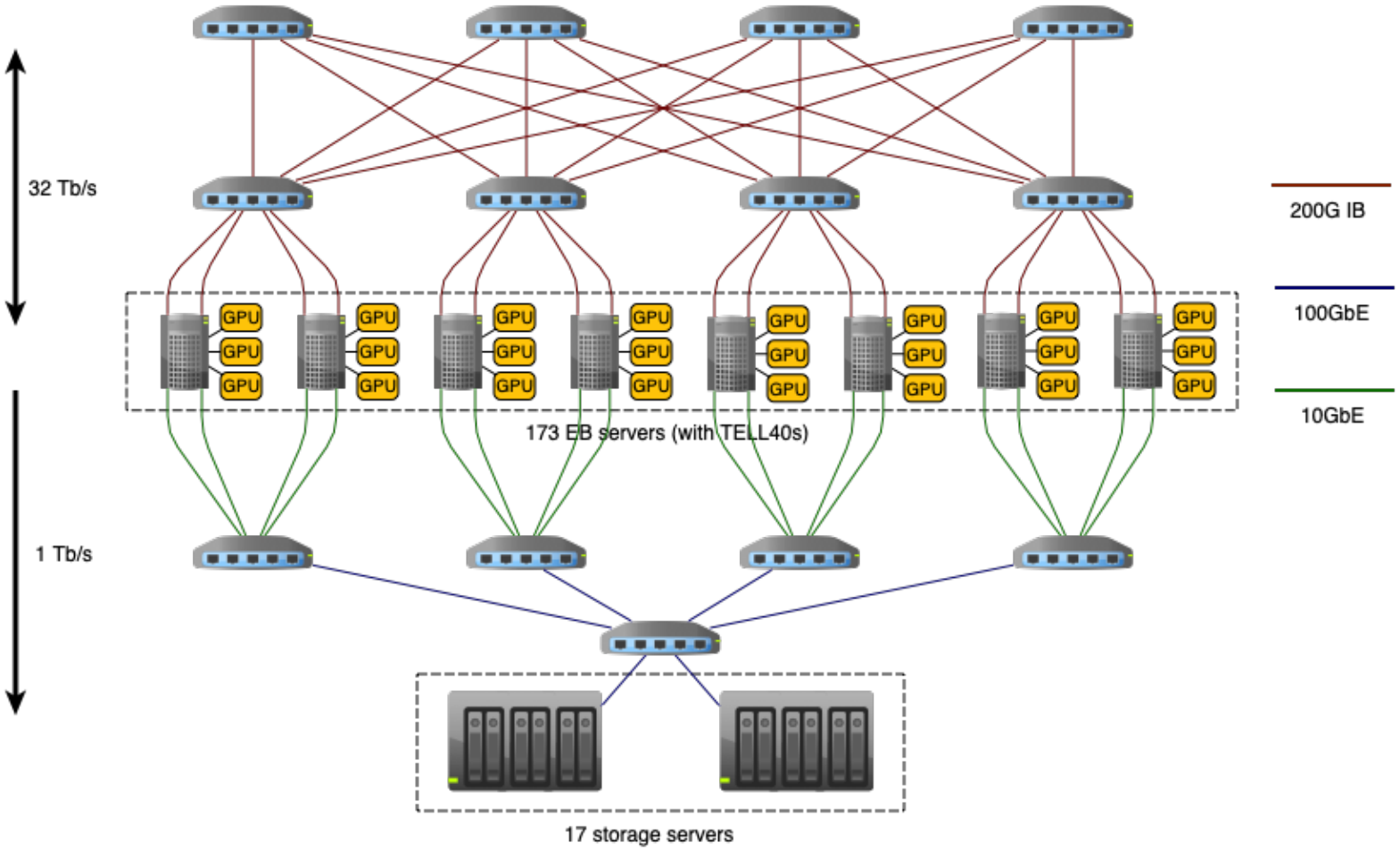
Reasons for going with InfiniBand

- PCIe Gen4 allows using 200 Gbit/s connections which save cost and help with scalability. However 200 Gbit/s so far only effectively exists for InfiniBand!
- Ethernet flow-control could not be made to work properly on available reference platforms
- Ethernet remains - for us - affected by worrying / irritating scaling issues
- Probably most important: could never get access to a really big Ethernet test-system: need the full event-builder for testing. For InfiniBand can and have used super-computer sites and the CMS DAQ (based on InfiniBand)

...ergo

Lowest risk solution at equal cost is the InfiniBand solution

Alternative with compute contained in building network



Conclusions

- Up to everyone to draw their own conclusions
- Personally (Niko): It's certainly possible to do this all with shallow buffer Ethernet, but we need a sizeable test-system, more time to test and tune, and probably settle for a lower average link-load
- To be continued until LHC Run4