Introduction to Trigger/DAQ challenges at CERN

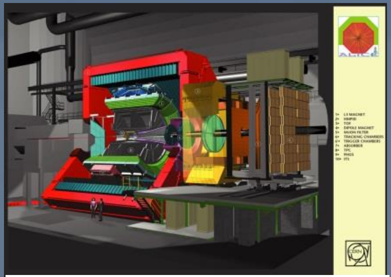
Niko Neufeld, CERN/PH

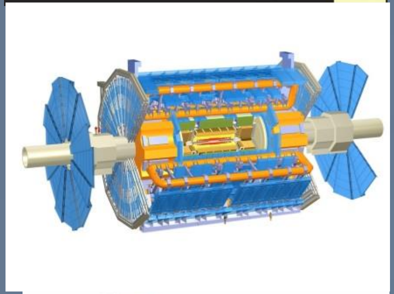
Many stimulating, fun discussions with my T-DAQ friends in ALICE, ATLAS, CMS and LHCb, and with a lot of smart people in CERN/IT (openlab) and industry are gratefully acknowledged

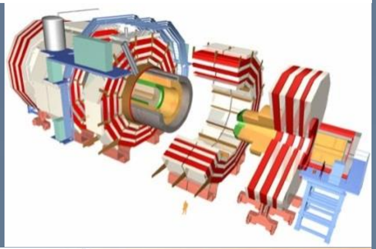
The LHC Experiments today

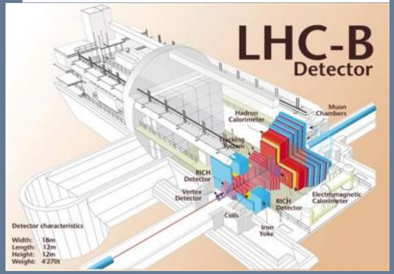
- ALICE – “A Large Ion Collider Experiment”
  - Size: 26 m long, 16 m wide, 16m high; weight: 10000 t
  - 35 countries, 118 Institutes
  - Material costs: 110 MCHF
- ATLAS – “A Toroidal LHC ApparatuS”
  - Size: 4 6m long, 25 m wide, 25 m high; weight: 7000 t
  - 38 countries, 174 institutes
  - Material costs: 540 MCHF
- CMS – “Compact Muon Solenoid”
  - Size: 22 m long, 15 m wide, 15 m high; weight: 12500 t
  - 40 countries, 172 institutes
  - Material costs: 500 MCHF
- LHCb – “LHC beauty” (b-quark is called “beauty” or “bottom” quark)
  - Size: 21 m long, 13 m wide, 10 m high; weight: 5600 t
  - 15 countries, 52 Institutes
  - Material costs: 75 MCHF
- Regular upgrades … first 2013/14 (Long Shutdown 1)

1 CHF ~ 1 USD

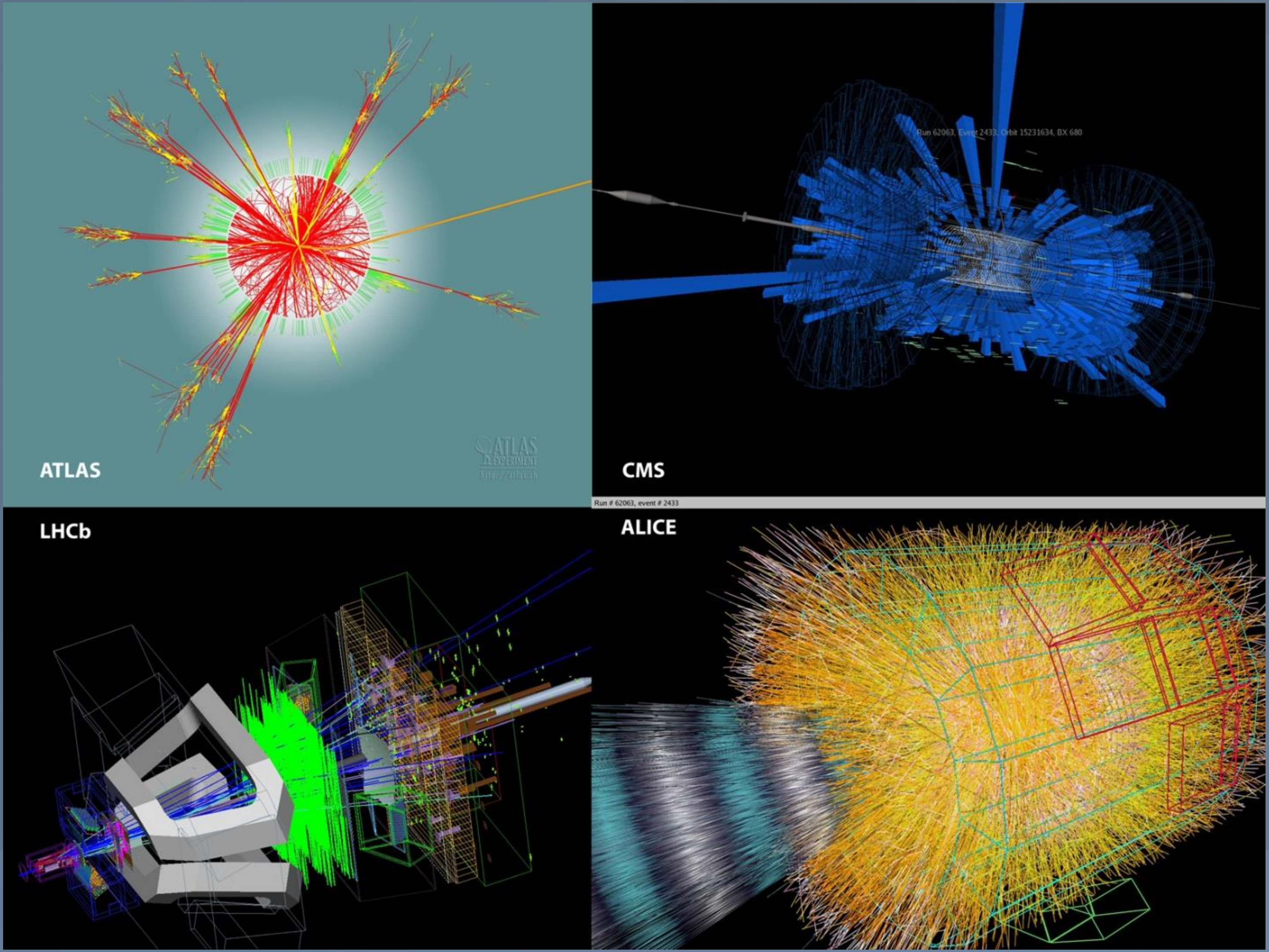






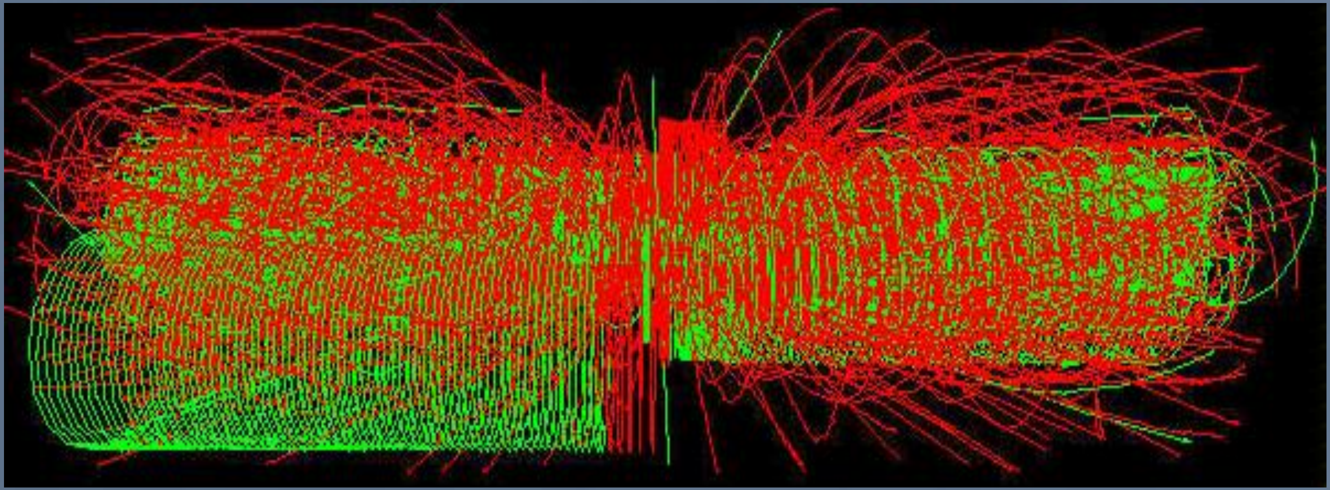


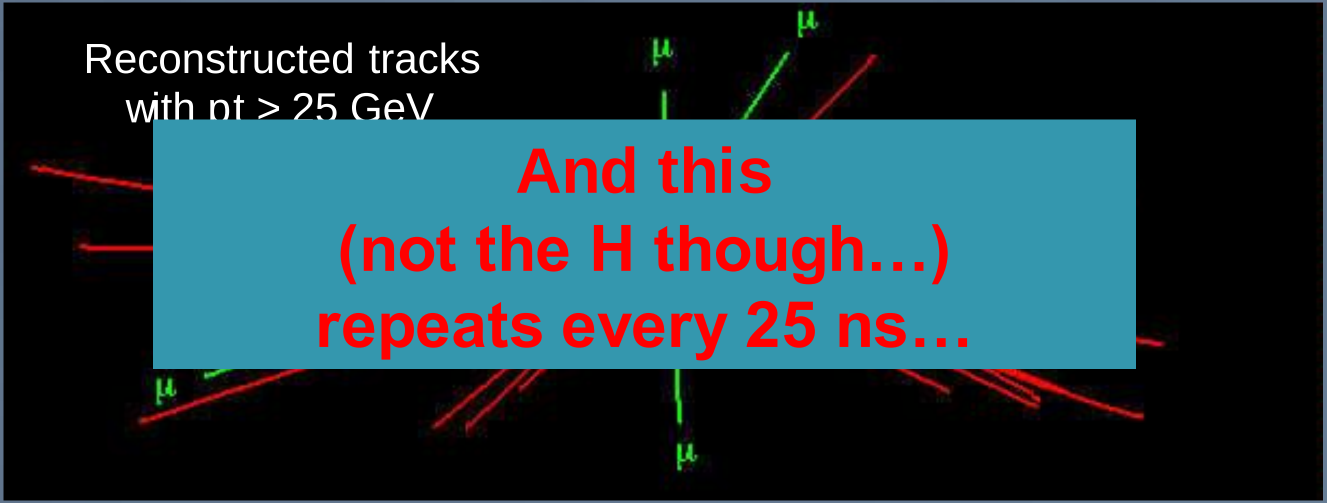
What do Events Look Like?


ATLAS
ATLAS EXPERIMENT
http://atlas.ch
Run 62063, Event 2433, Orbit 15231634, BX 680
CMS
Run # 62063, event # 2433
LHCb
ALICE

The needle in the hay-stack: pp Collisions at 14 TeV at 10^{34} cm^{-2}s^{-1}

- σ(pp) = 70 mb --> >7 x 10^{8}/s (!)
- In ATLAS and CMS* 20 – 30 min bias events overlap
- $H \rightarrow ZZ$
  $Z \rightarrow \mu\mu$
  $H \rightarrow$ 4 muons: the cleanest ("golden") signature





And this
(not the H though…)
repeats every 25 ns…

*)LHCb @4x10^{33} cm^{-2}-1 isn't much nicer and in Alice (PbPb) is even more busy

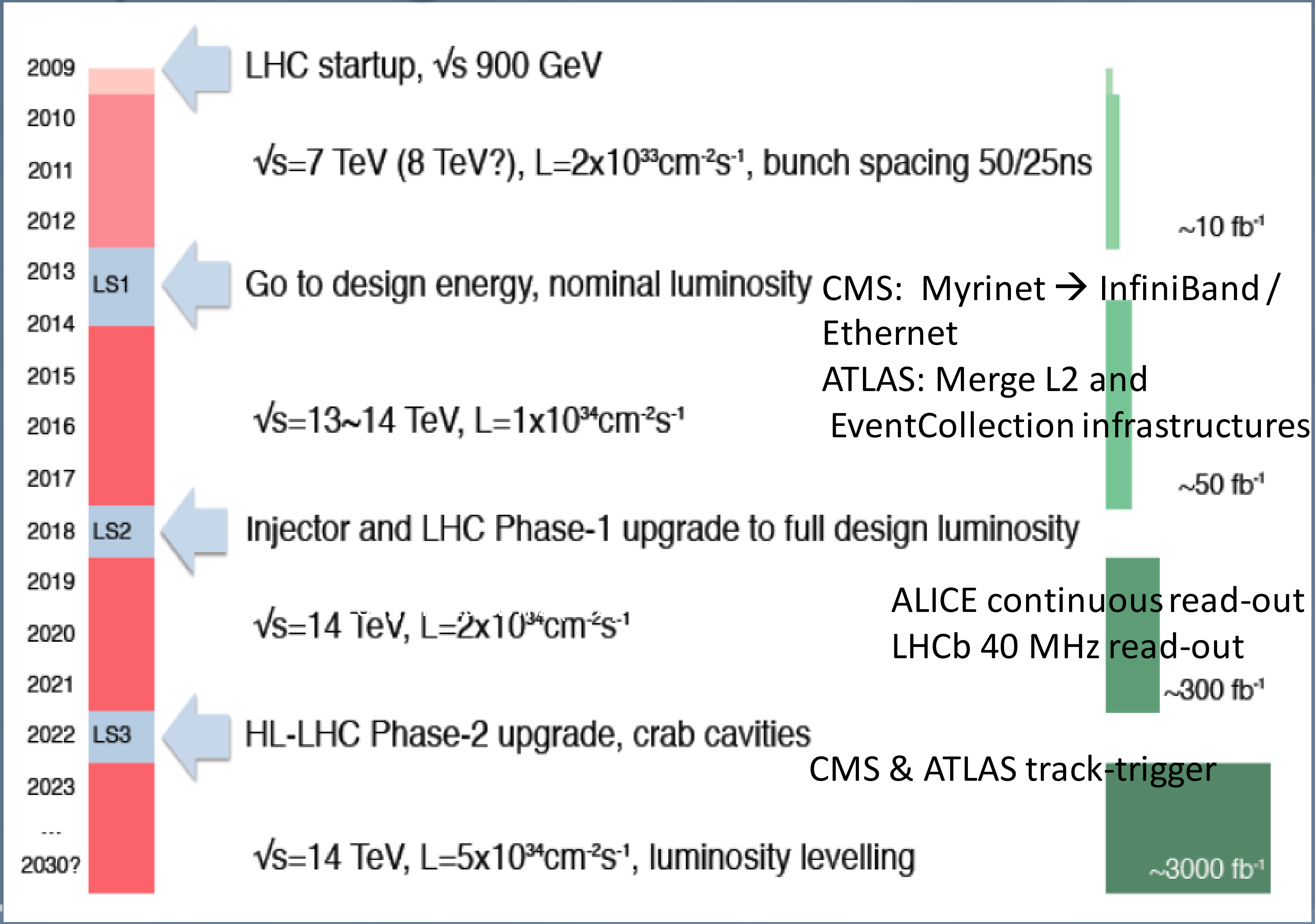
Data Rates

- Particle beams cross every 25 ns (40 MHz)
  - Up to 25 particle collisions per beam crossing
  - Up to 10^9 collisions per second
- Basically 2 event filter/trigger levels
  - Data processing starts at readout
  - Reducing 10^9 p-p collisions per second to O(1000)
- Raw data to be stored permanently: >15 PB/year

Physics Process	Events/s
Inelastic p-p scattering	10^8
b	10^6
$W \rightarrow e\upsilon ; W \rightarrow \mu\upsilon ; W \rightarrow \tau\upsilon$	20
$Z \rightarrow ee ; Z \rightarrow \mu\mu ; Z \rightarrow \tau\tau$	2
t	1
Higgs boson (all; m_H = 120GeV)	0.04
Higgs boson (simple signatures)	0.0003
Black Hole (certain properties)	0.0001

	Incoming data rate	Outgoing data rate	Reduction factor
Level1 Trigger (custom hardware)	40000000 s^{-1}	10^5 – 10^6 s^{-1}	400-10,000
High Level Trigger (software on server farms)	2000-1000000 s^{-1}	1000 -10000 s^{-1}	10-2000

LHC planning



LHC startup, √s 900 GeV

√s=7 TeV (8 TeV?), L=2x10^{33}cm^{-2}s^{-1}, bunch spacing 50/25ns

Go to design energy, nominal luminosity

√s=13~14 TeV, L=1x10^{34}cm^{-2}s^{-1}

Injector and LHC Phase-1 upgrade to full design luminosity

√s=14 TeV, L=2x10^{34}cm^{-2}s^{-1}

HL-LHC Phase-2 upgrade, crab cavities

√s=14 TeV, L=5x10^{34}cm^{-2}s^{-1}, luminosity levelling

CMS: Myrinet → InfiniBand / Ethernet

ATLAS: Merge L2 and EventCollection infrastructures

ALICE continuous read-out

LHCb 40 MHz read-out

CMS & ATLAS track-trigger

Future DAQs in numbers

	Event-size [kB]	Rate of events into HLT [kHz]	HLT bandwidth [Gb/s]	Year [CE]
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	2000	200	3200	2022
LHCb	100	40000	32000	2019

40000 kHz == collision rate
→ *LHCb abandons Level 1 for an all-software trigger*

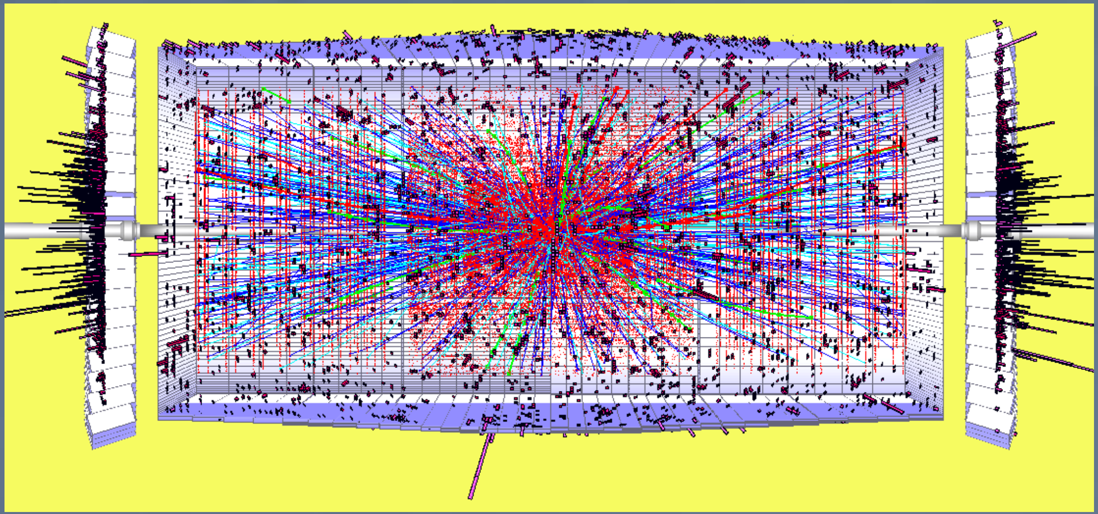
It’s a good time to do DAQ
CMS and ATLAS numbers are growing as we speak…

Challenge #1
The first level trigger

Level 1 Trigger

- The Level 1 Trigger is implemented in hardware: FPGAs and ASICs → difficult / expensive to upgrade or change, maintenance by experts only
- Decision time: ~ a small number of microseconds
- It uses simple, hardware-friendly signatures → looses interesting collisions
- Each sub-detector has its own solution, only the uplink is standardized →

A Track-Trigger at 40 MHz 2020++



- Goals:
  - Resolve up to 200÷250 collisions per bunch crossing
  - Maintain occupancy at the few % level
  - Maintain overall L1 rate within 100 KHz
  - Keep latency within ~ 6 µs (ECAL pipeline 256 samples = 6.4 µs)
    - The current limit is the Tracker
- L1 tracking trigger data combined with calorimeter & muon trigger data
  - With finer granularity than presently employed.
- Physics objects made from tracking, calorimeter & muon trigger data transmitted to Global Trigger.

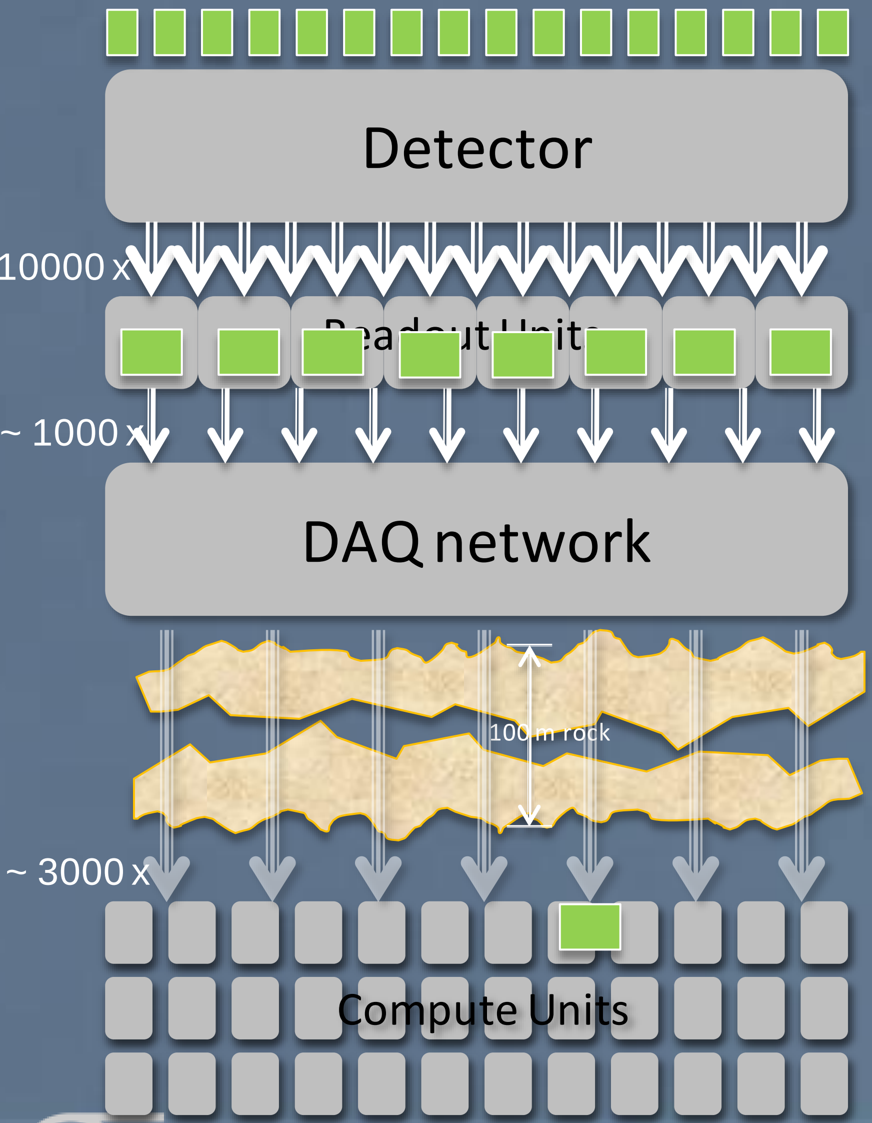
Intro to Trigger/DAQ challenges at CERN - N. Neufeld

Level 1 challenge

- Can we do this in software?
- Maybe in GPGPUs / XeonPhis → studies ongoing in the NA62 experiment
- We need low and – ideally – deterministic latency
- Need an efficient interface to detector-hardware: CPU/FPGA hybrid?
- Or forget about the whole L1 thing altogether and do everything in HLT → requires a lot of fast, low-power, low-cost links, did anybody say Si-photonics?

Challenge #2
Data Acquisition

Data Acquisition (generic example)



Every Readout Unit has a piece of the collision data
All pieces must be brought together into a single compute unit
The Compute Unit runs the software filtering (High Level Trigger – HLT)

↓ GBT: custom radiation- hard link from the detector 3.2 Gbit/s

↓ DAQ ("event-building") links – some LAN (10/40/100 GbE / InfiniBand)

↓ Links into compute-units: typically 10 Gbit/s (because filtering is currently compute-limited)

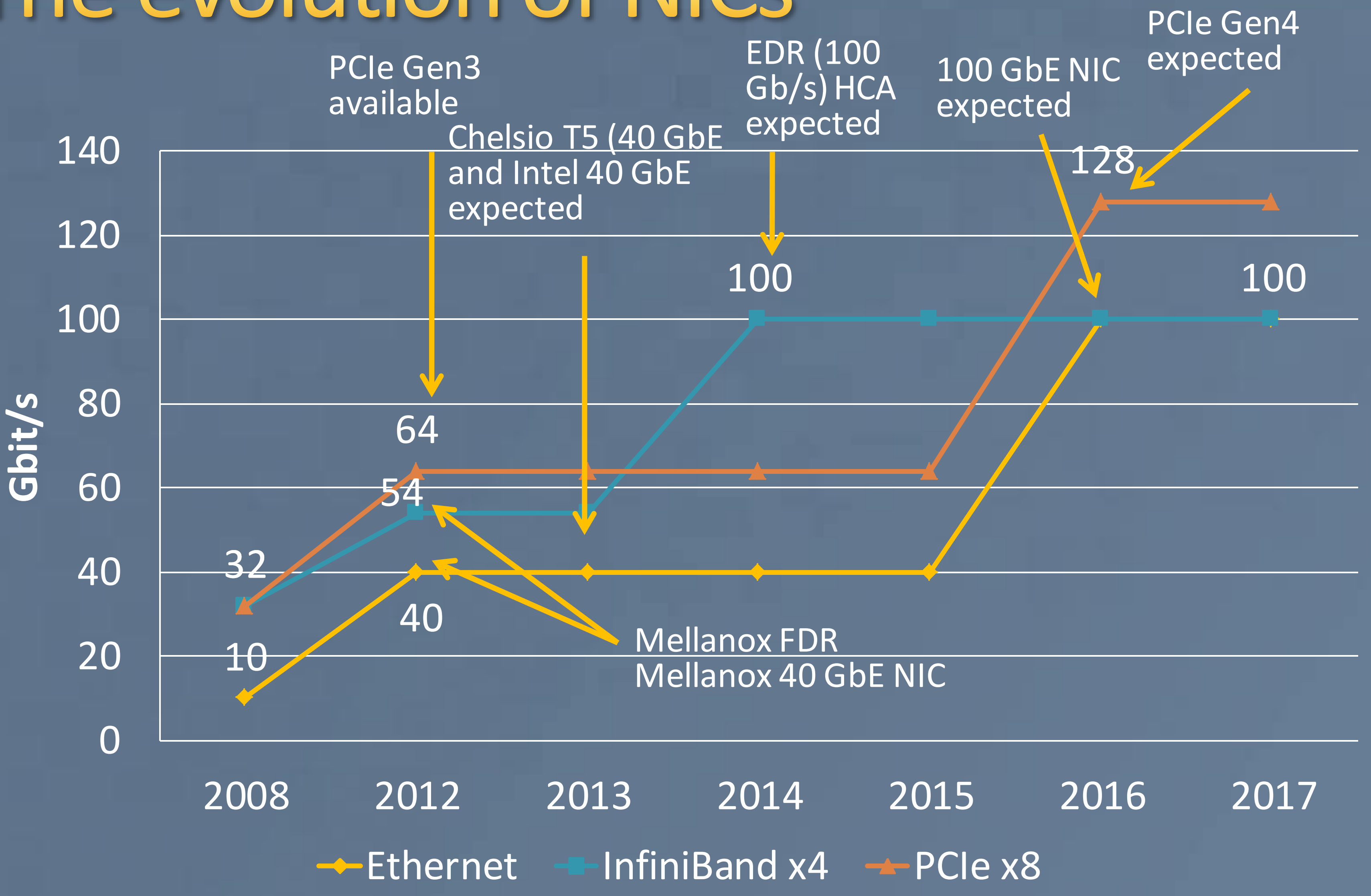
Key Figures – Example LHCb

- Minimum required bandwidth: > 32 TBit/s
- Number of 100 Gbit/s links: > 320
- Number of compute units: > 4000
- Event size: 100 kB
- Number of events per seconds: 10 – 40 Millions per second
- Number of events retained for permanent storage: 20k – 30k per second
    - storage allows to "defer" the decision at the cost of disks and tapes

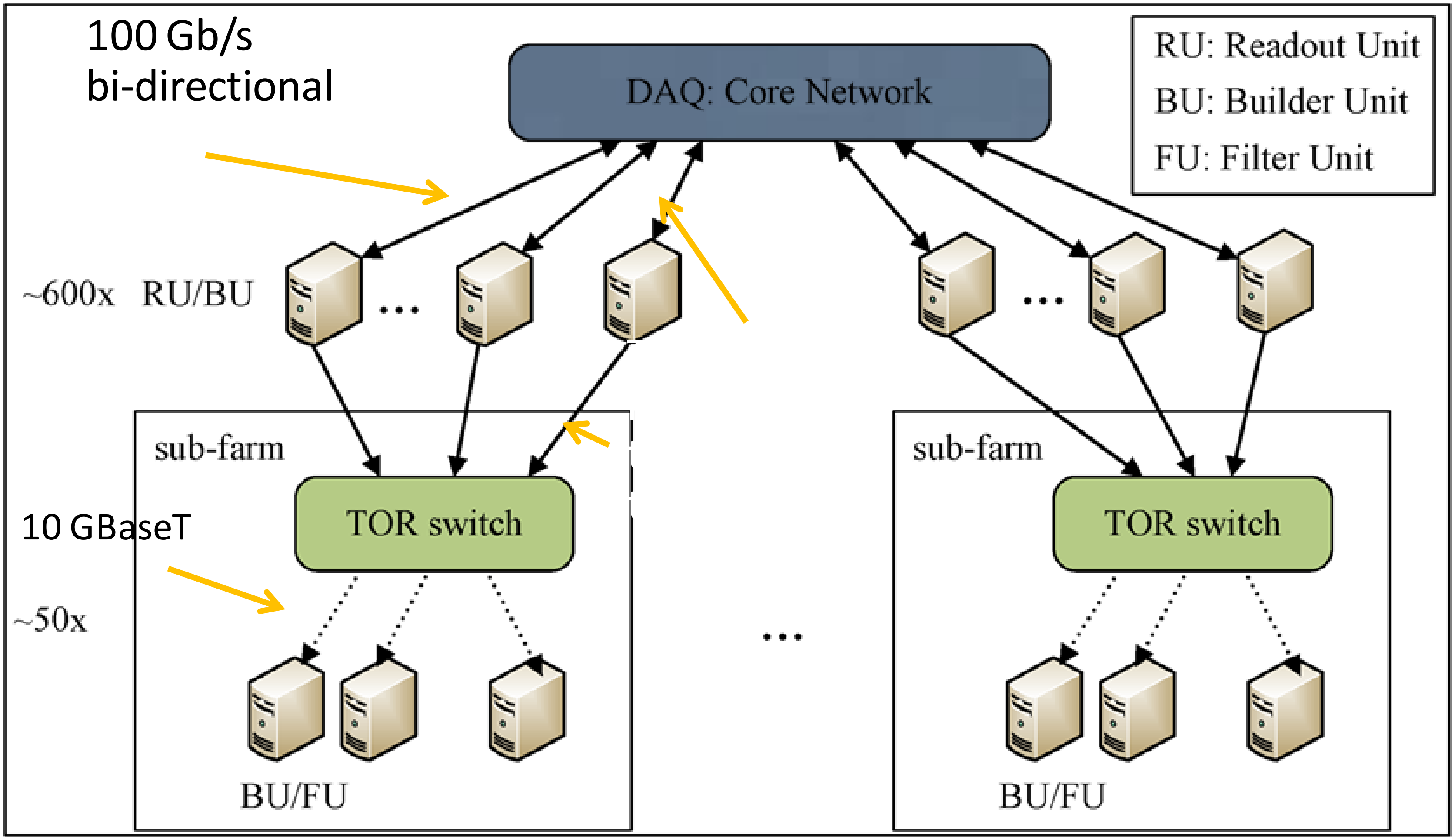
Design principles

- Minimize number of expensive "core" network ports
- Use the most efficient technology for a given connection
  - different technologies should be able to co-exist (e.g. fast for building, slow for end-node)
  - keep distances short
- Exploit the economy of scale → try to do what everybody does (but smarter ☺)

The evolution of NICs

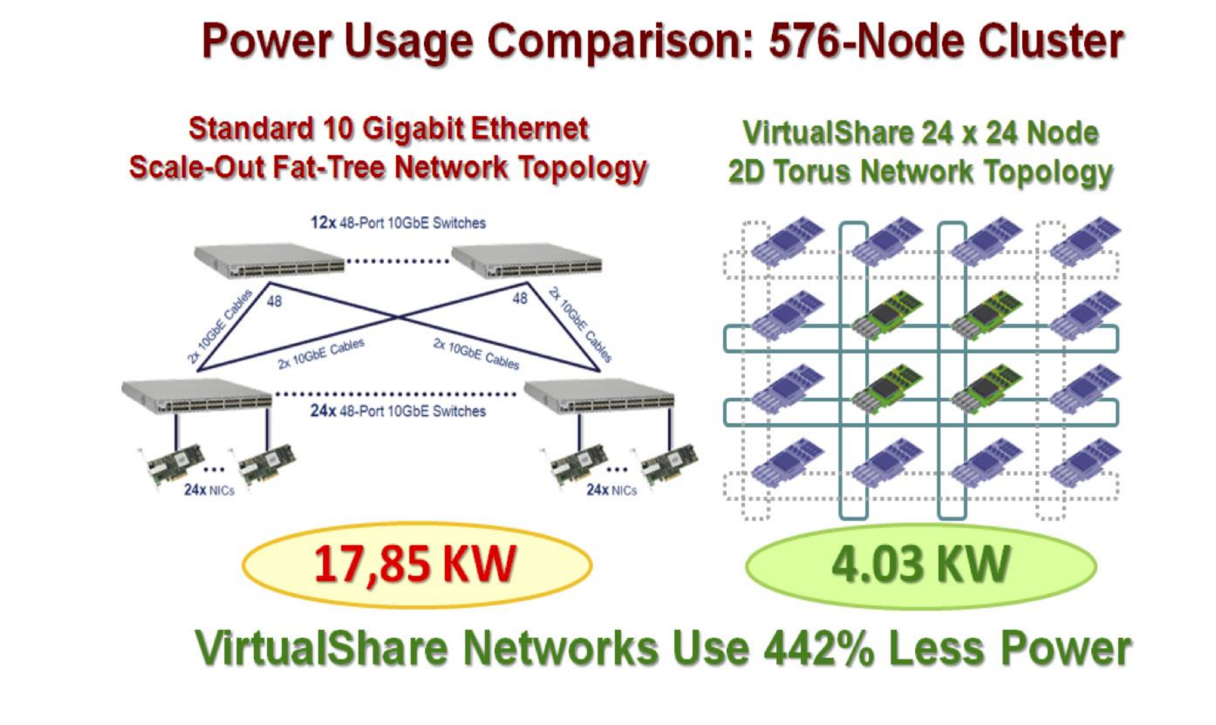


A realistic DAQ / HLT for LHC



Keep an eye on the fringe

- There is always the temptation to remove the switch altogether → merge fabric and network
- Modern versions of an old idea (token-ring, SCI)
  - PCIe based (for example VirtualShare Ronniee a 2D torus based on PCIe, creates a large 64 bit shared memory space over PCIe)
  - IBM blue-gene interconnect (11 x 16 Gb/s links integrated on chip – build a 5N torus)



Power Usage Comparison: 576-Node Cluster

Standard 10 Gigabit Ethernet Scale-Out Fat-Tree Network Topology

VirtualShare 24 x 24 Node 2D Torus Network Topology

17,85 KW

4.03 KW

VirtualShare Networks Use 442% Less Power

DAQ challenge

- Transport multiple Terabit/s reliably and cost-effectively
- Integrate the network closely and efficiently with compute resources (be they classical CPU or "many-core")
- Multiple network technologies should seamlessly co-exist in the same integrated fabric ("the right link for the right task")

Challenge #3
High Level Trigger

High Level Trigger: Key Figures

- Existing code base: 5 MLOC of mostly C++
- Almost all algorithms are single-threaded (only few exceptions)
- Currently processing time on a X5650 per event: several 10 ms / process (hyper-thread)
- Currently between 100k and 1 million events per second are filtered online in each of the 4 experiments

Online Trigger Farms at the end of Run 1

	ALICE	ATLAS	CMS	LHCb
# cores (+ hyperthreading)	2700	17000	13200	15500
# servers (mainboards)	~ 500	~ 2000	~ 1300	1574
total available cooling power [kW]	~ 500	~ 820	800	525
total available rack-space (Us)	~ 2000	2400	~ 3600	2200
CPU type(s)	AMD Opteron, Intel 54xx, Intel 56xx, Nvidia GPU	Intel 54xx, Intel 56xx	Intel 54xx, Intel 56xx Intel E5-2670	Intel 5450, Intel 5650, AMD 6220

Massive upgrades foreseen for Run 2

HLT needs for the future: 2018+

	Event-size [kB]	Rate of events into HLT [kHz]	HLT bandwidth [Gb/s]	Year [CE]
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	2000	200	3200	2022
LHCb	100	40000	32000	2019

Up a factor 10 to 40 from current rates - with much more complex events

Coprocessors and all that

- Many core co-processors (Xeon/Phi, GPGPUs) are currently very much in fashion in the HPC world
  - Lots of interest in HEP, but few successful applications so far: ALICE, NA62
- It might be that it will be most efficient to include them directly in the event-building network (i.e. receive data directly on the GPGPU/XeonPhi rather than passing through the main CPU)– this is supported today using IB by both Intel and Nvidial
- The "co-processor" could become an independent unit on he network → this will clearly require very high-speed network interfaces (>>100 Gb/s to make sense over PCIe Gen3)

High Level Trigger compared to HPC

- Like HPC:
  - full ownership of the entire installation → can choose architecture and hardware components
  - single “client” / “customer”
  - have a high-bandwidth interconnect

- Unlike HPC:
  - many independent small tasks which execute quickly
    → no need for check-pointing (fast storage)
    → no need for low latency
  - data driven, i.e. when the LHC is **not** running (70% of the time) the farm is idle → interesting ways around this (deferal, “offline usage)
  - facility is very long-lived, growing incrementally

High Level Trigger challenge

- Make the code-base ready for multi/many-core (this is not Online specific!)
- Optimize the High Level Trigger farms in terms of cost, power, cooling
- Find the best architecture integrating "standard servers", many-core systems and a high-bandwidth network

Summary

- LHC “online” computing needs to acquire move and process huge amounts of data in real-time
- Level 1 trigger challenge: replace custom by industry-standard hardware; move more data with less power
- Data acquisition challenge: very high bandwidth, low-overhead networks, tightly integrated with computing resources
- High Level Trigger challenge: make most out of the CPU power (parallelisation), find the most power- and cost-efficient way to provide as much computing power as possible for extracting the most interesting physics

More material

Challenge 4: I/O on x86 servers

The evolution of PCs

- PCs used to be relatively modest I/O performers (compared to FPGAs), this has radically changed with PCIe Gen3
- Xeon processor line has now 40 PCIe Gen3 lanes / socket
- Dual-socket system has a theoretical throughput of 1.2 Tbit/s(!)
  - Tests suggest that we can get quite close to the theoretical limit (using RDMA at least)
- This is driven by the need for fast interfaces for co-processors (GPGPUs, XeonPhi)
- For us (even in LHCb) CPU will be the bottle-neck in the server - not the LAN interconnect – 10 Gb/s by far sufficient

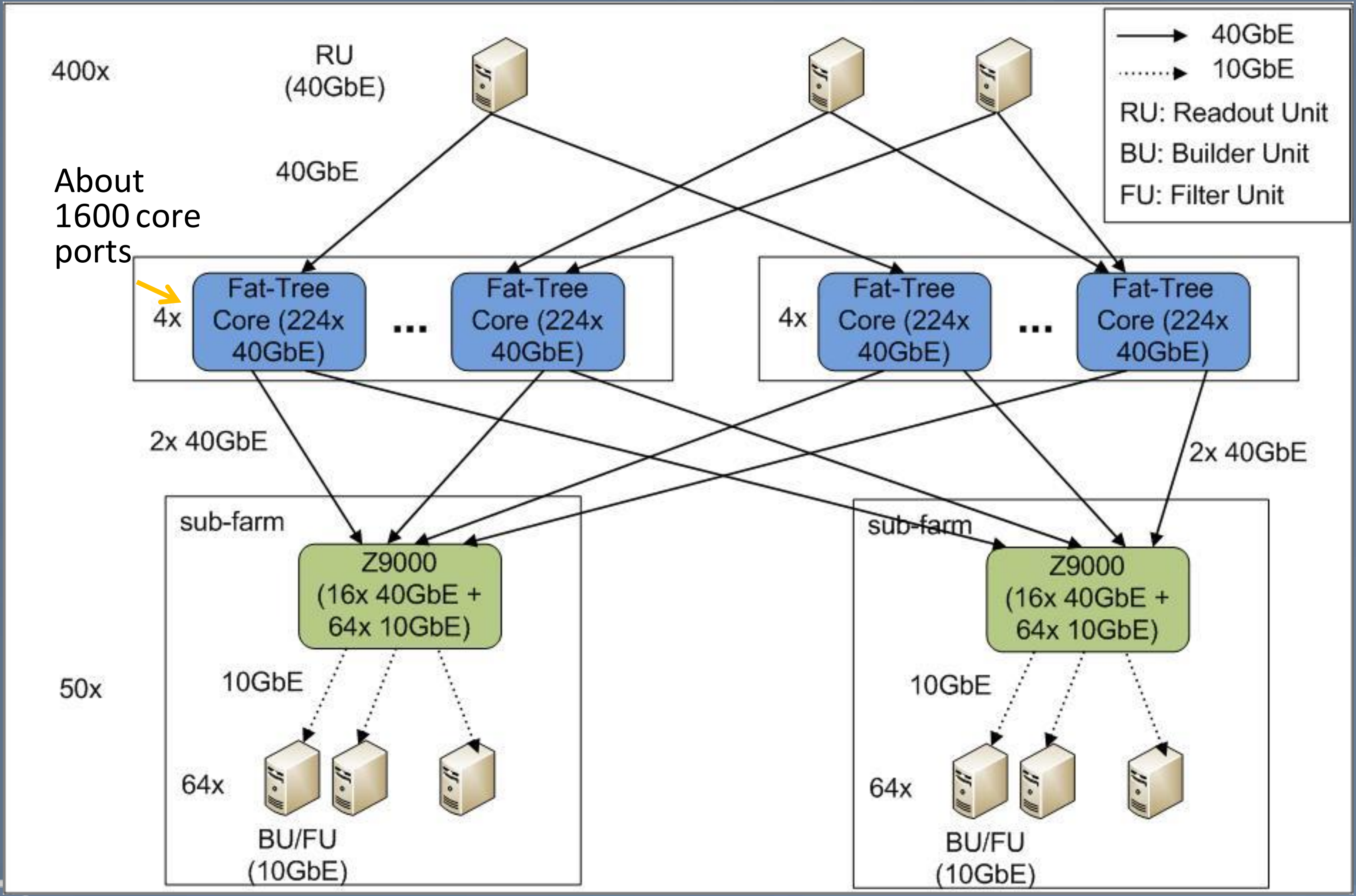
Keep distances short

- Multi-lane optics (Ethernet SR4, SR10, InfiniBand QDR) over multi-mode fibres are limited to 100 (OM3) to 150 (OM4) meters
- Cable assemblies (“direct-attach) cables are either
  - passive (“copper”, “twinax”), very cheap and rather short (max. 4 to 5 m), or
  - active – still cheaper than discreet optics , but as they use the same components internally they have similar range limitations
- For comparison: price-ratio of 40G QSFP+ copper cable assembly, 40G QSFP+ active cable, 2 x QSFP+ SR4 optics + fibre (30 m) = 1 : 8 : 10

The evolution of lane-speed

- All modern interconnects are multiple serial: (x something SR)
- Another aspect of “Moore’s” law is the increase of serialiser speed
- Higher speed reduces number of lanes (fibres)
- Cheaper interconnects also require availability of cheap optics (VCSEL, Silicon-Photonics)
- VCSEL currently runs better over MMF (OM3, OM4 for 10 Gb/s and above) → per meter these fibres are more expensive than SMF
- Current lane-speed 10 Gb/s (same as 8 Gb/s, 14 Gb/s)
- Next lane-speed (coming soon and already available on high-end FPGAs) is 25 Gb/s (same as 16 Gb/s) → should be safely established by 2017 (a hint for GBT v3 ☺?)

Classical fat-core event-builder


400x
RU
(40GbE)
40GbE
10GbE
RU: Readout Unit
BU: Builder Unit
FU: Filter Unit
40GbE
4x
Fat-Tree Core (224x 40GbE)
...
Fat-Tree Core (224x 40GbE)
4x
Fat-Tree Core (224x 40GbE)
...
Fat-Tree Core (224x 40GbE)
2x 40GbE
2x 40GbE
sub-farm
Z9000 (16x 40GbE + 64x 10GbE)
sub-farm
Z9000 (16x 40GbE + 64x 10GbE)
50x
10GbE
10GbE
64x
64x
BU/FU (10GbE)
BU/FU (10GbE)

About 1600 core ports