

Very large networks for future Event Builders

Niko Neufeld, CERN/PH

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

Disclaimer

- *Prediction is very difficult, especially about the future.*
(attributed to Niels Bohr)
- I use what I learned over the last 4 – 5 years in presentations, private talks and on the net
- I have tried to discard anything, which I remember to be under NDA. If you are interested in something I did not say/write, come see me after
If you think I spilt any beans don't tell anybody (but tell me!)

Introduction

- I am not going to explain event-building, in particular not the couple of 1000 lines of C-code, which are often referred to as the “event-builder”
- This is about technology, architecture and cost

Future DAQs in numbers

	Event-size [kB]	Rate [kHz]	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

It's a good time to do DAQ

Technology

Links, PCs and networks

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

The evolution of PCs

- More integration (Intel roadmap)
 - All your I/O are to belong to us 😊 → integrate memory controller ✓, PCIe controller ✓, GPU ✓, NIC (3 – 4 years?), physical LAN interface (4 – 5 years)?
 - **Advertisement**: we are going to study some of these in our common EU project ICE-DIP
- Aim for high density
 - short distances (!)
 - efficient cooling
 - less real-estate needed (rack-space)
- Same thing applies mutatis mutandis to ARM (and other CPU architectures)

Ethernet

- The champion of all classes in networking
- 10, 40, 100 out and 400 in preparation
- As speed goes up so does power-dissipation → high-density switches come with high integration
- Market seems to separate more and more in to two camps:
 - carrier-class, deep-buffer, high-density, flexible firmware (FPGA, network processors) (Cisco, Juniper, Brocade, Huawei), \$\$\$/port
 - data-centre, shallow buffer, ASIC based, ultra-high density, focused on layer 2 and simple layer 3 features, very low latency, \$/port (these are often also called Top Of the Rack TOR)

The “two” Ethernet’s

Speed	Core [USD / port]	TOR [USD / port]
10 Gb/s	400 – 1000	200 - 250
40 Gb/s	1000 - 4000	500 - 900

- If possible less core and more TOR ports
- buffering needs to be done elsewhere

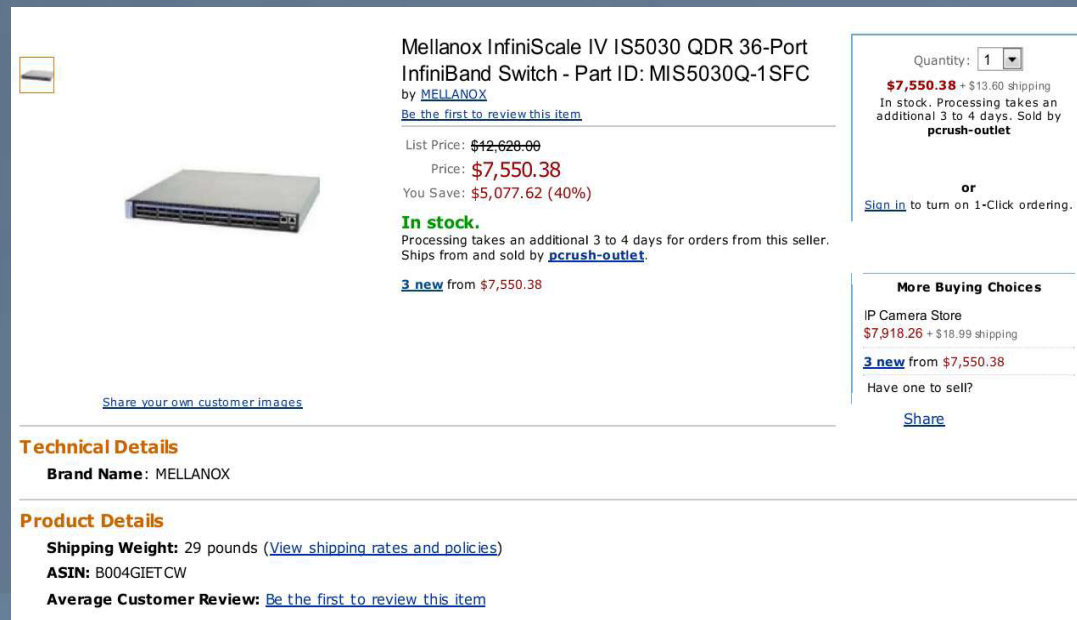
DCE/B: can we have the cake and eat it too?

- Loss-less Ethernet claims to give us what deep-buffer switches do, but without the price (why: because this is made for FCoE and this wants low latency)
- Basically try to avoid congestion and need for buffering → push buffer-needs into end-nodes (PCs)
- DCB is an umbrella-term for a zoo of IEEE standards (among them):
 - [IEEE 802.1Qb](#) PFC (Priority based Flow Control)
 - [IEEE 802.1Qaz](#) ETS (Enhanced Transmission Selection)
 - [IEEE 802.1Qau](#) CN (Congestion Notification)
- Currently only partially supported, in particular CN not really working in practice, converged NICs expensive → but this might change if FCoE takes off – keep an eye on this.
- Not clear that this is more effective than explicit traffic shaping on a dedicated network

InfiniBand

- Driven by a relatively small, agile company: Mellanox
- Essentially HPC + some DB and storage applications
- Only competitor: Intel (ex-Qlogic)
- Extremely cost-effective in terms of Gbit/s / \$
- Open standard, but almost single-vendor – unlikely for a small startup to enter
- Software stack (OFED including RDMA) also supported by Ethernet (NIC) vendors

- Many recent Mellanox products (as of FDR) compatible with Ethernet



Mellanox InfiniScale IV IS5030 QDR 36-Port InfiniBand Switch - Part ID: MIS5030Q-1SFC by MELLANOX

[Be the first to review this item.](#)

List Price: ~~\$12,620.00~~
Price: **\$7,550.38**
You Save: **\$5,077.62 (40%)**

In stock.
Processing takes an additional 3 to 4 days for orders from this seller. Ships from and sold by [pcrush-outlet](#).

[3 new](#) from \$7,550.38

[Share your own customer images](#)

Technical Details
Brand Name: MELLANOX

Product Details
Shipping Weight: 29 pounds ([View shipping rates and policies](#))
ASIN: B004GIETCW
Average Customer Review: [Be the first to review this item](#)

Quantity:

\$7,550.38 + \$13.80 shipping
In stock. Processing takes an additional 3 to 4 days. Sold by [pcrush-outlet](#)

or
[Sign in](#) to turn on 1-Click ordering.

More Buying Choices
IP Camera Store
\$7,918.26 + \$18.99 shipping
[3 new](#) from \$7,550.38
Have one to sell?
[Share](#)

InfiniBand prices

Speed	Core [USD / port]	TOR [USD / port]
52 Gb/s	850	250

Speed	Core [USD / port]	TOR [USD / port]
10 Gb/s	400 – 1000	200 - 250
40 Gb/s	1200 - 4000	500 - 900

→ Worth some R&D at least

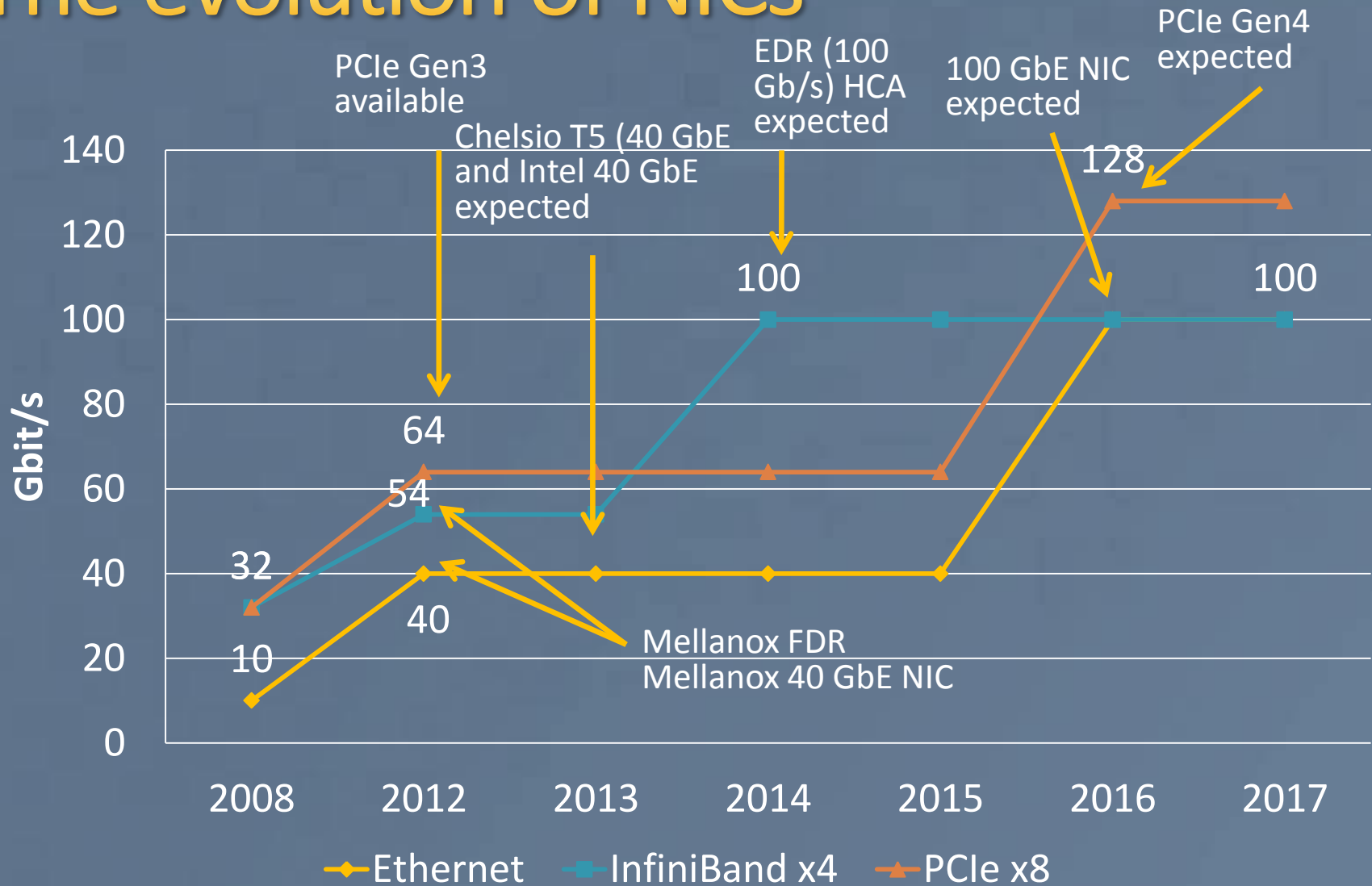
The devil's advocate: why **not** InfiniBand *?

- InfiniBand is (almost) a single-vendor technology
- Can the InfiniBand flow-control cope with the very bursty traffic typical for a DAQ system, while preserving high bandwidth utilization?
- InfiniBand is mostly used in HPC, where hardware is regularly and radically refreshed (unlike IT systems like ours which are refreshed and grown over time in an evolutionary way). Is there not a big risk that the HPC guys jump onto a new technology and InfiniBand is gone?
- The InfiniBand software stack seems awfully complicated compared to the Berkeley socket API. What about tools, documentation, tutorial material to train engineers, students, etc...?



* Enter the name of your favorite non-favorite here!

The evolution of NICs



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 😊?)

The evolution of switches – BBOT (Big Beasts Out There)

Date of release

- Brocade MLX: 768 10-GigE
- Juniper QFabric: up to 6144 10-GigE (not a single chassis solution)
- Mellanox SX6536: 648 x 56 Gb (IB) / 40 GbE ports
- Huawei CE12800: 288 x 40 GbE / 1152 x 10 GbE



Inside the beasts

- Many ports does not mean that the interior is similar
- Some are a high-density combination of TOR elements (Mellanox, Juniper Qfabric)
 - Usually just so priced that you can't do it cheaper by cabling up the pizza-boxes yourself
- Some are classical fat cross-bars (Brocade)
- Some are in-between (Huawei, CLOS but lots of buffering)

The eternal copper

- Surprisingly (for some) copper cables remain the cheapest option, iff distances are short
- A copper interconnect is even planned for InfiniBand EDR (100 Gbit/s)
- For 10 Gigabit Ethernet the verdict is not yet passed, but I am convinced that we **will have 10 GBaseT on main-board “for free”**
 - It is not yet clear if there will be (a need for) truly high-density 10 GBaseT line-cards (100 ports+)

Keep distances short

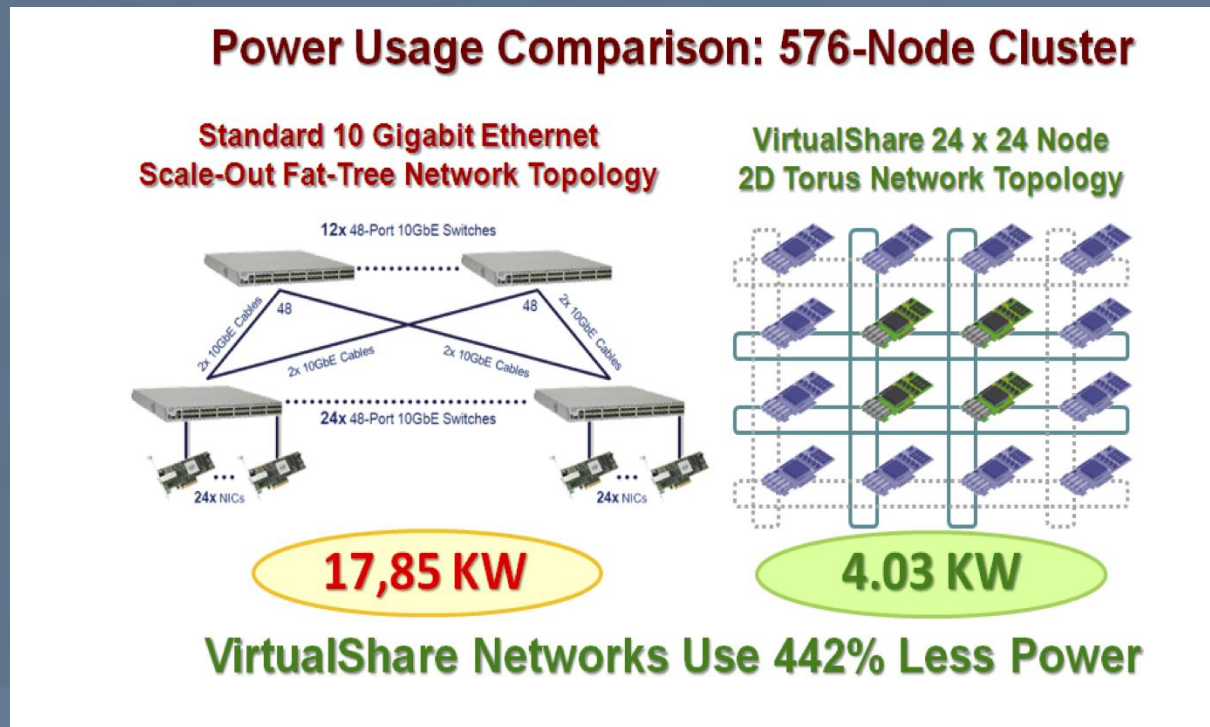
- Multi-lane optics (Ethernet SR4, SR10, InfiniBand QDR) over multi-mode fibres are limited to 100 (OM3) to 150 (OM4) meters
- Cable assemblies (“direct-attach) cables are either
 - passive (“copper”, “twinax”), very cheap and rather short (max. 4 to 5 m), or
 - active – still cheaper than discreet optics , 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

Coprocessors and all that

- After event-building all these events need to be processed by software triggers
- In HPC (“Exascale”) co-processors (Xeon/Phi, GPGPUs) are currently the key technology to achieve this
 - There is also ARM but this is for another talk 😊
- If they are used in the trigger, it is likely that it will be most efficient to include them into the event-building (i.e. receive data directly on the GPGPU rather than passing through the main CPU – this is supported today using InfiniBand by both Nvidia and Intel)
- Bottle-neck is the interface bus → main driver for more and faster PCIe lanes
- Interestingly Intel seems also to develop the concept to make the “co-processor” an independent unit on the network → this will clearly require very high-speed network interfaces (>>100 Gb/s to make sense over PCIe Gen3)

Exotica: toroi in N dimensions

- 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)



Exotica: sequel and finish

- If we had an infinite number of students, it would be at least cool check it in simulation
- In any case this **will not scale gracefully for event-building**, and to be economical would need to be combined with conventional LAN

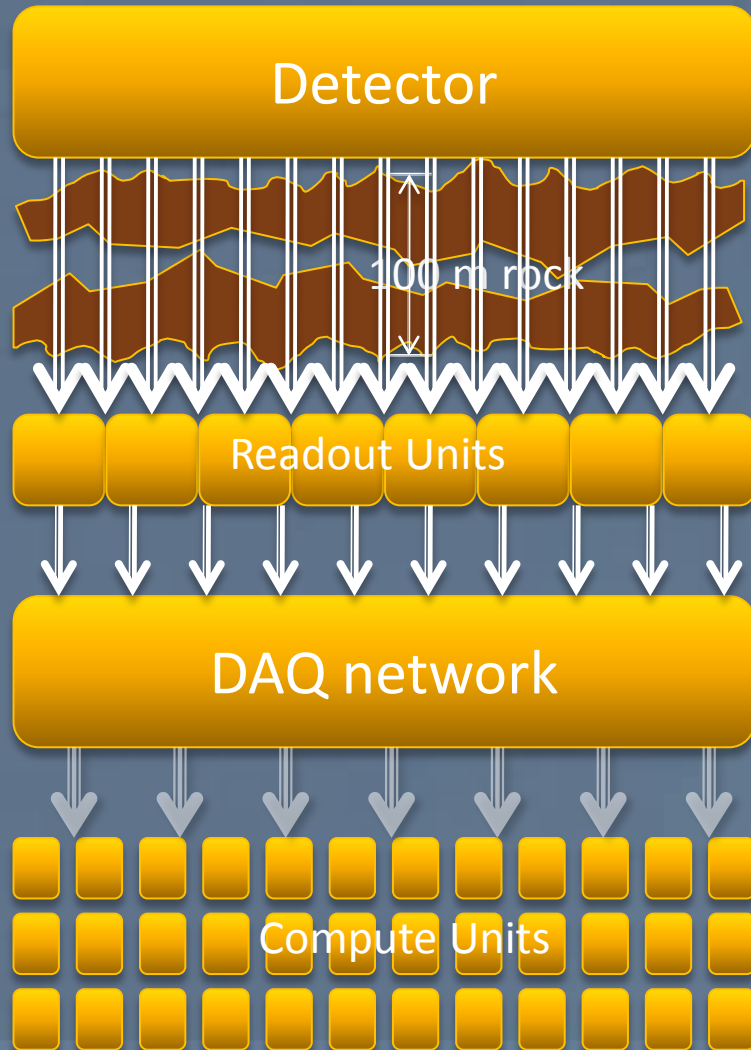
Architecture

Make the best of available technologies

Principles

- Minimize number of “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 😊)

Minimising distances for LHCb



Long distance covered by low-speed links from detector to Readout Units.

- ☺ Cheap and links required anyhow
- ☹ Many MMF required (low speed)



GBT: custom radiation-hard link over MMF, 3.2 Gbit/s (about 12000)

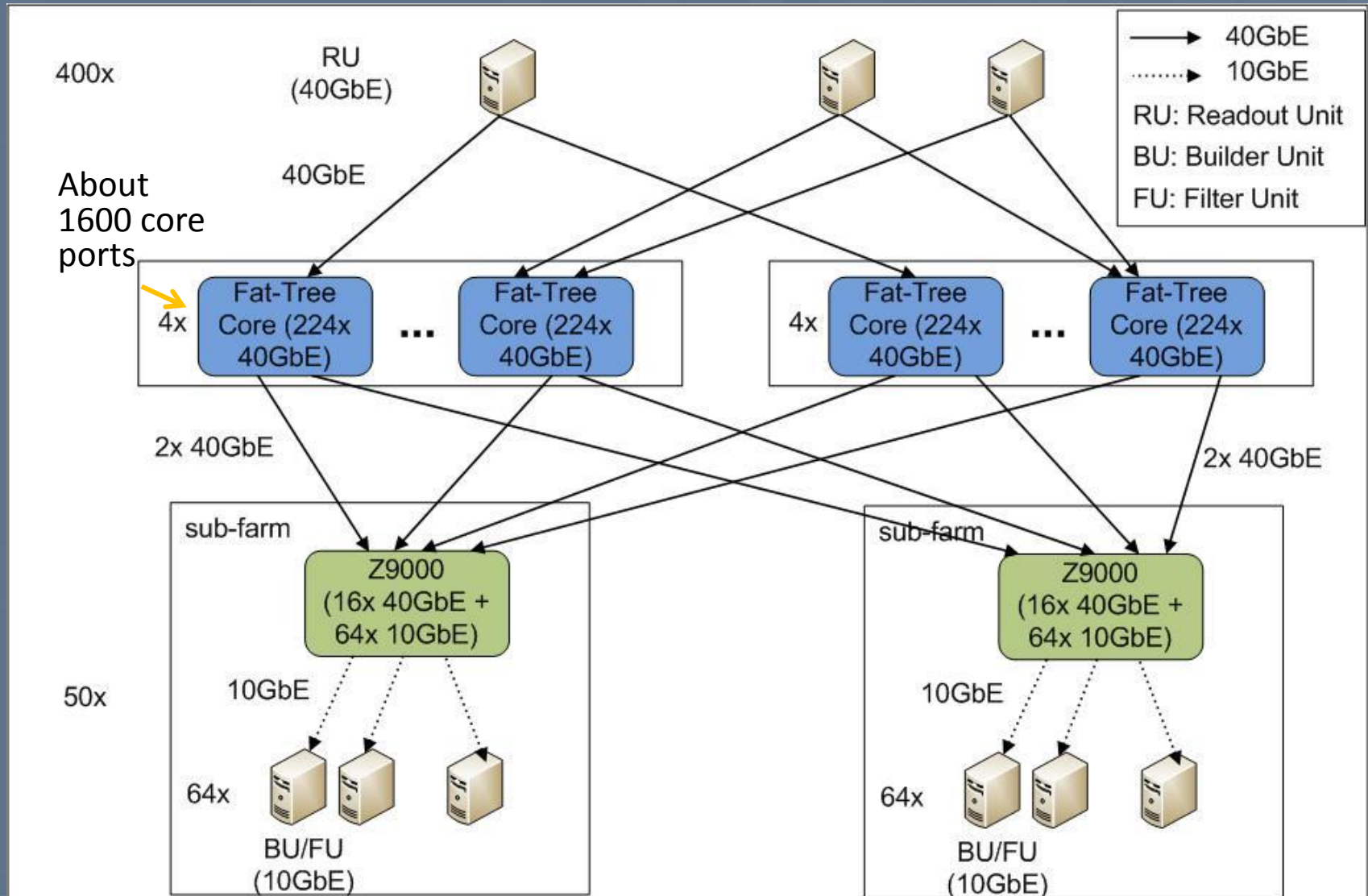


Input into DAQ network (10/40 Gigabit Ethernet or FDR IB) (1000 to 4000)



Output from DAQ network into compute unit clusters (100 Gbit Ethernet / EDR IB) (200 to 400 links)

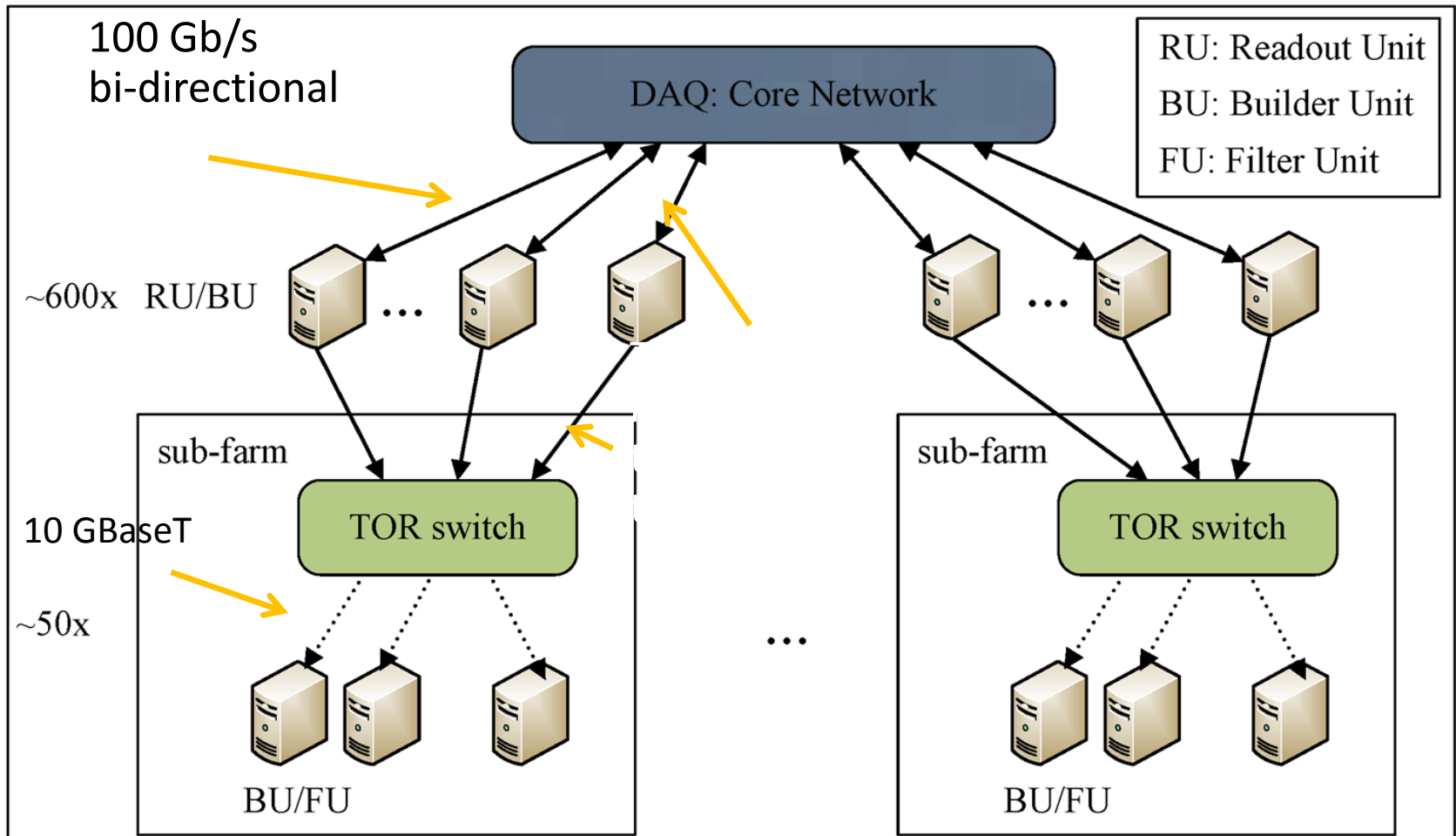
Classical fat-core event-builder



Protocols & topologies

- Pull, push, barrel-shifting can be done of course on any topology
- It is more the switch-hardware (and there in particular the amount of buffer-space) and the properties of the low-level (layer-2) protocol which will make the difference

Reduce number of core-ports



Conclusions

- Technology is on our side
- Truly large DAQ systems can be built and afforded
- Event-building is still an “unusual” problem – We need to watch out for the best architecture to match what’s “out there” to our needs
- Which should maximize our options
 - Not follow any crazy fashion
 - But be ready to exploit a new mainstream we cannot anticipate today