Technology Developments for Trigger, Data Acquisition, Software and Computing

ECFA HL LHC Workshop: ECFA High Luminosity LHC Experiments Workshop

For the TDOC preparatory group
Niko Neufeld CERN/PH

TDOC Membership

- ALICE: Pierre Vande Vyvre, Thorsten Kollegger, Predrag Buncic
- ATLAS: David Rousseau, Benedetto Gorini, Nikos Konstantinidis
- CMS: Wesley Smith, Christoph Schwick, Ian Fisk, Peter Elmer
- LHCb: Renaud Legac, Niko Neufeld

Documentation

Relevant documents used as inputs by the preparatory group

- ALICE Upgrade LoI: LHCC-2012-012
- ATLAS Phase 2 Upgrade LoI: LHCC-2012-022
- CMS Draft Phase 2 Upgrade Document available over the summer 2013
- LHCb Framework Upgrade TDR: LHCC-2012-007.
- I. Bird *et al*. "Update of the Computing Models of the WLCG and the LHC Experiments", forthcoming
- For technology forecasting: document by Bernd Panzer (CERN-IT) http://cern.ch/go/DFG7 , with updates in I. Bird et al.

Intro & disclaimer

- Focus on Commercial Of the Shelf (COTS) relevant for Trigger, DAQ and computing in three broad topics:
- Processing
  - FPGA
  - Co-processors (many-cores)
  - CPUs (x86 and others)
- Interconnects
  - links & serializers
  - networks (LAN & WAN)
- Storage
  - tapes
  - disks

- Not covered:
  - ASICs, general semi-conductor development → electronics talks
  - Software → covered by David
  - Architecture →covered by Wesley
- Not covered ≠ not interesting

"Prediction is very difficult, especially about the future."

(attributed to Niels Bohr)

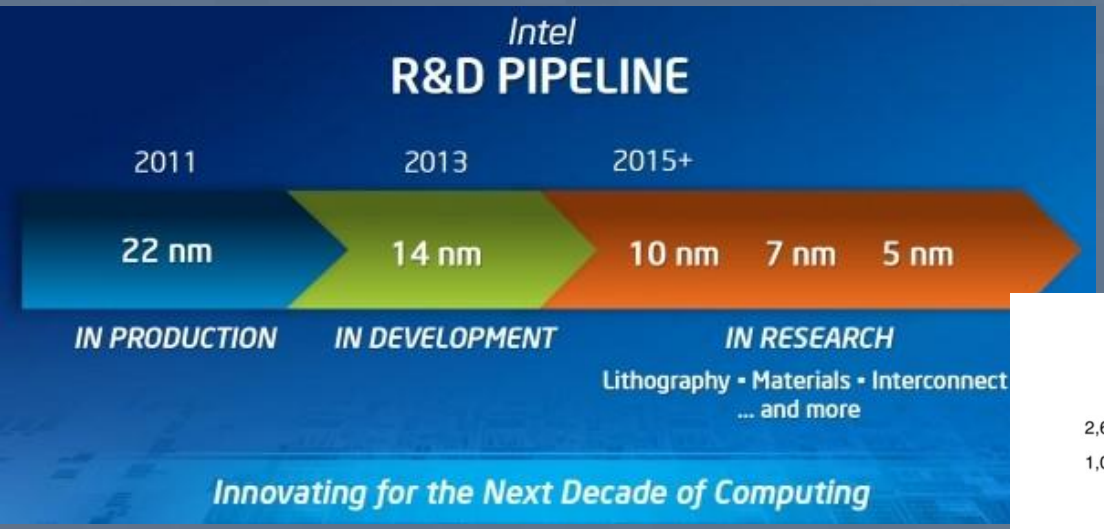
Due diligence…

- Technology talks are the more interesting the more “new stuff” they contain
- In summarizing our ideas about the future we are also relying on information we got from industry, often under non-disclosure agreement (NDA)
- Great care has been taken only to present information which has been cleared for public information
- Should you find anything on these slides, which you think is under NDA please contact the author (niko.neufeld@cern.ch)

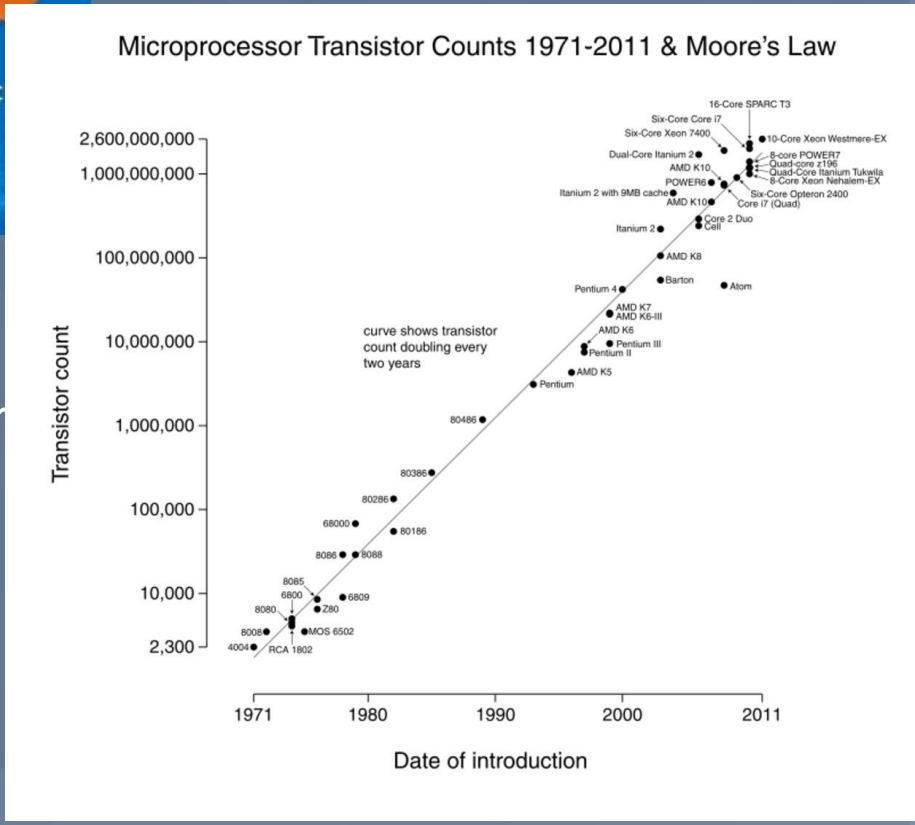
Processing

Moore's law

Source: http://www.techspot.com/



Source: wikipedia

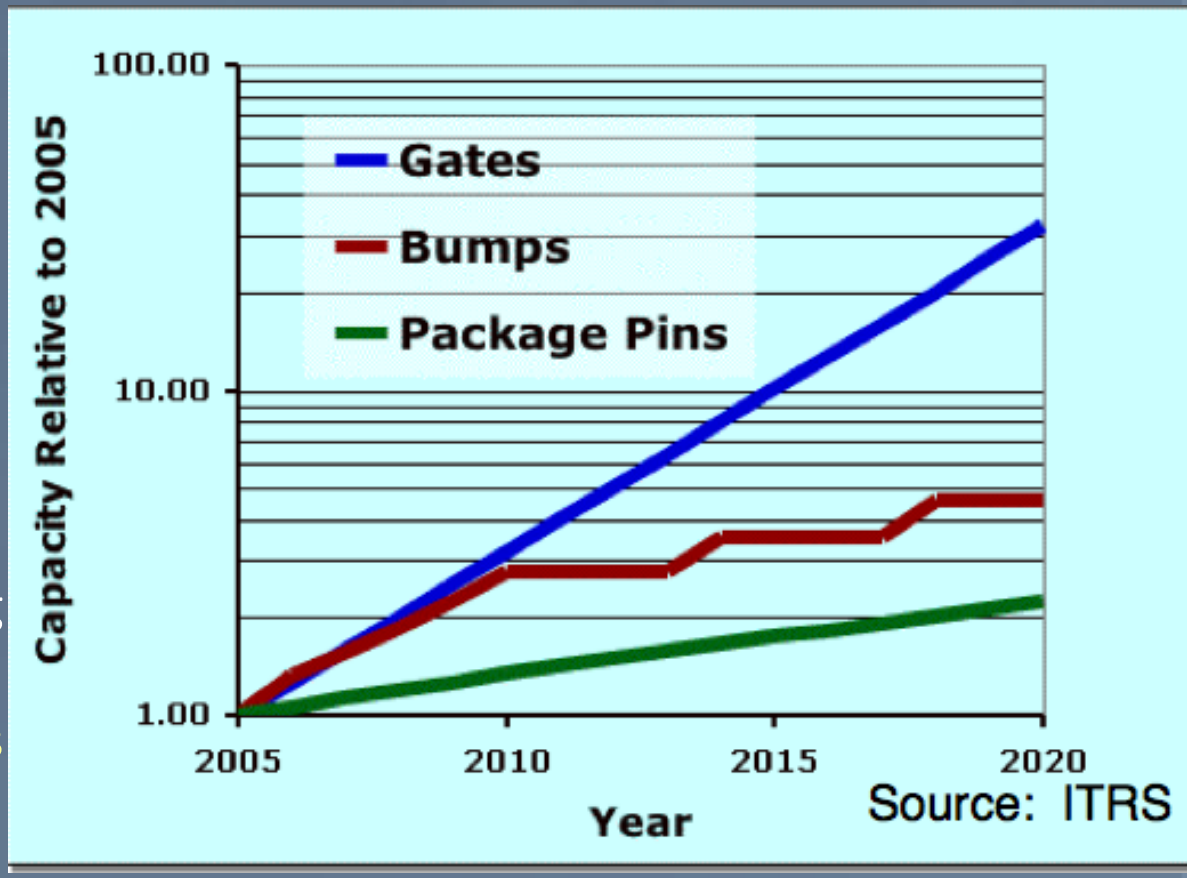


- 14 nm CPUs/FPGAs in 2014
  - (simpler NAND Flash already at 10 nm!)
- Intel strongly committed to new processor generation every 12 to 18 months: TSMC, Samsung, ST also still in this game
- expect 10 nm in 2015/16
- Moore's law will hold at least until 2020

FPGAs

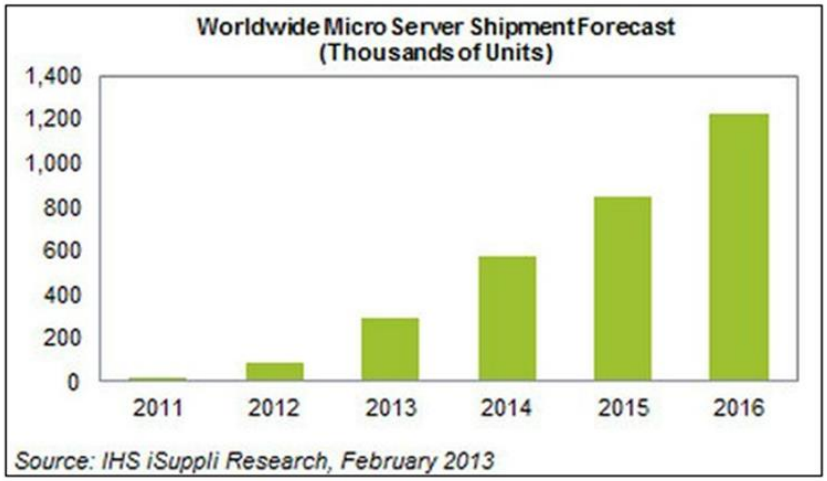
- Moore's law holds for FPGAs as well
  - Next generation Altera Stratix 10 in 14 nm
  - Smaller feature size means higher-speed and/or less power consumption
- No problem in logic density
- Challenges
  - programmability
  - long-term maintenance of design FPGA software which is rapidly changing → conserve tools in virtual machines, but this will need continuous support

Source: I. Bolsen (Xilinx) 2012



Architecture trends in micro-processors

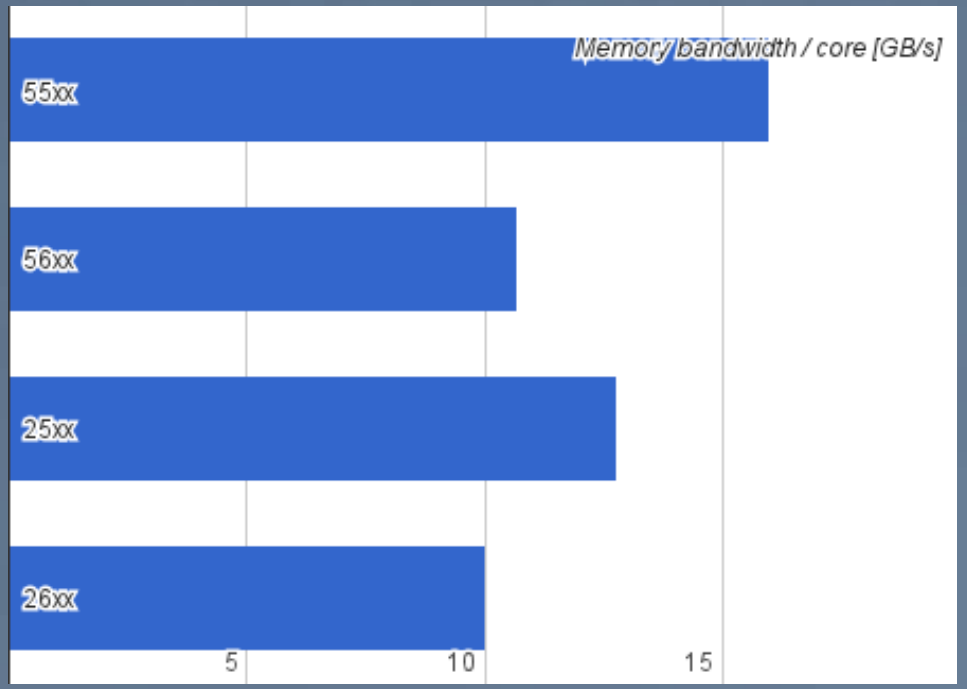
- Key challenges are power and memory band-width
- System-on-a-Chip
  - Memory control, voltage regulators integrated (helps with fine-grained power-management)
  - Integration of CPU and GPU (and DSPs) (these simpler elements tend to use less power for the same chip area than traditional cores)
  - Unified memory architecture, 3-D DRAM memory on chip
  - → micro-servers based on SOC
- Market focus on cost effective components for Smartphones, Phablets, Tablets, Ultrabooks, Notebooks
- Servers 'less' innovative
  - More memory channels & bandwidth
  - Larger caches, more cores
- Focus on vector units
  - HPC, integration of accelerator cards
- problem of "dark silicon" in processors



Microserver ships are expected to grow by a factor of 50 between 2011 and 2016

Memory band-width and latency

- Memory bandwidth / core is rather decreasing[1]
- DDR4 is expected to be the last parallel, pluggable memory as we know it
- This problem is obviously much more acute for many-core architectures

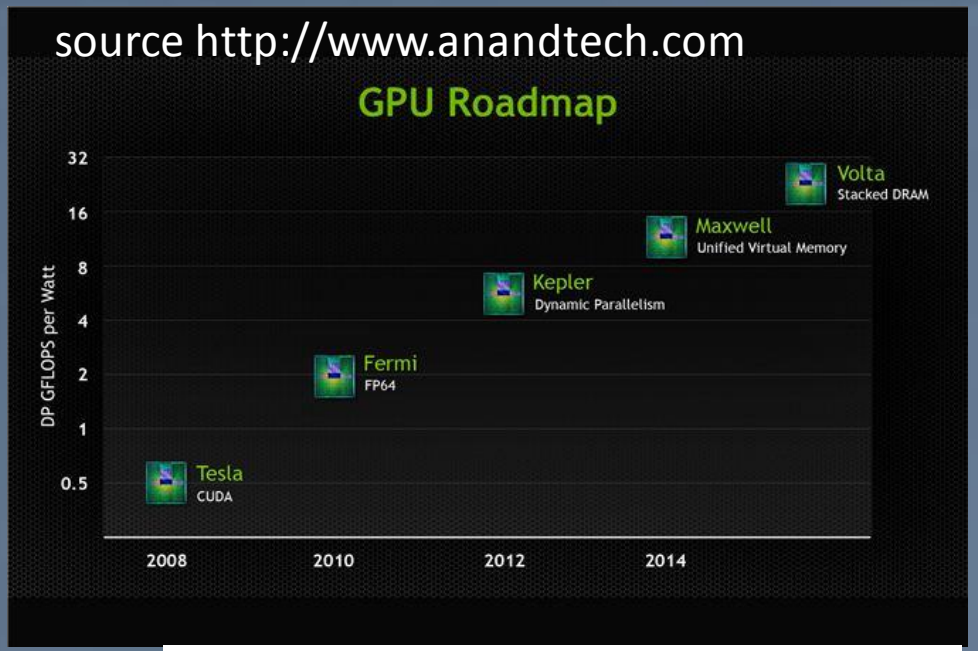


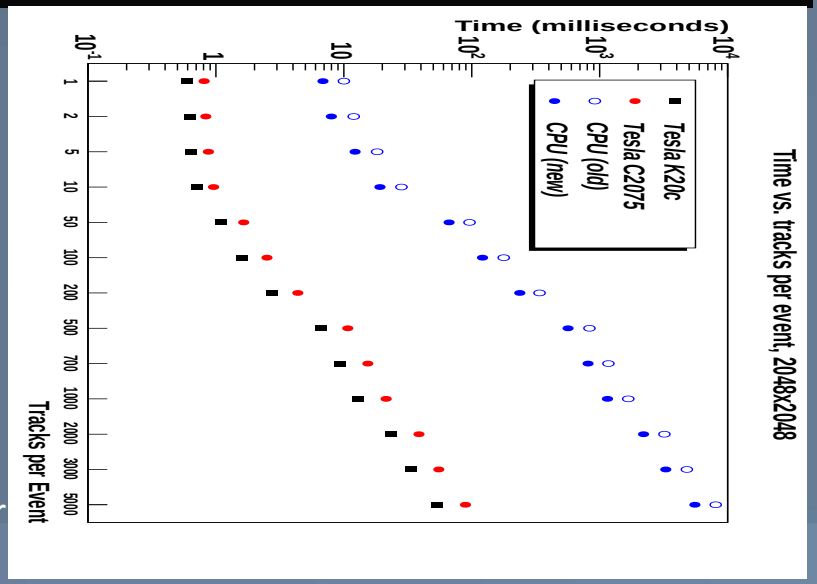
Source of raw data: wikipedia
Picture shows memory bandwidth /core for 4 recent Intel architectures using the fastest possible DDR3 memory in optimal configuration and the maximum number of cores available for dual-socket server processors

[1] Latency and bandwidth are strongly coupled. Actual application performance depends more often on the latency of memory access which in turn depend on the organization of the and size of the caches

Co-processors

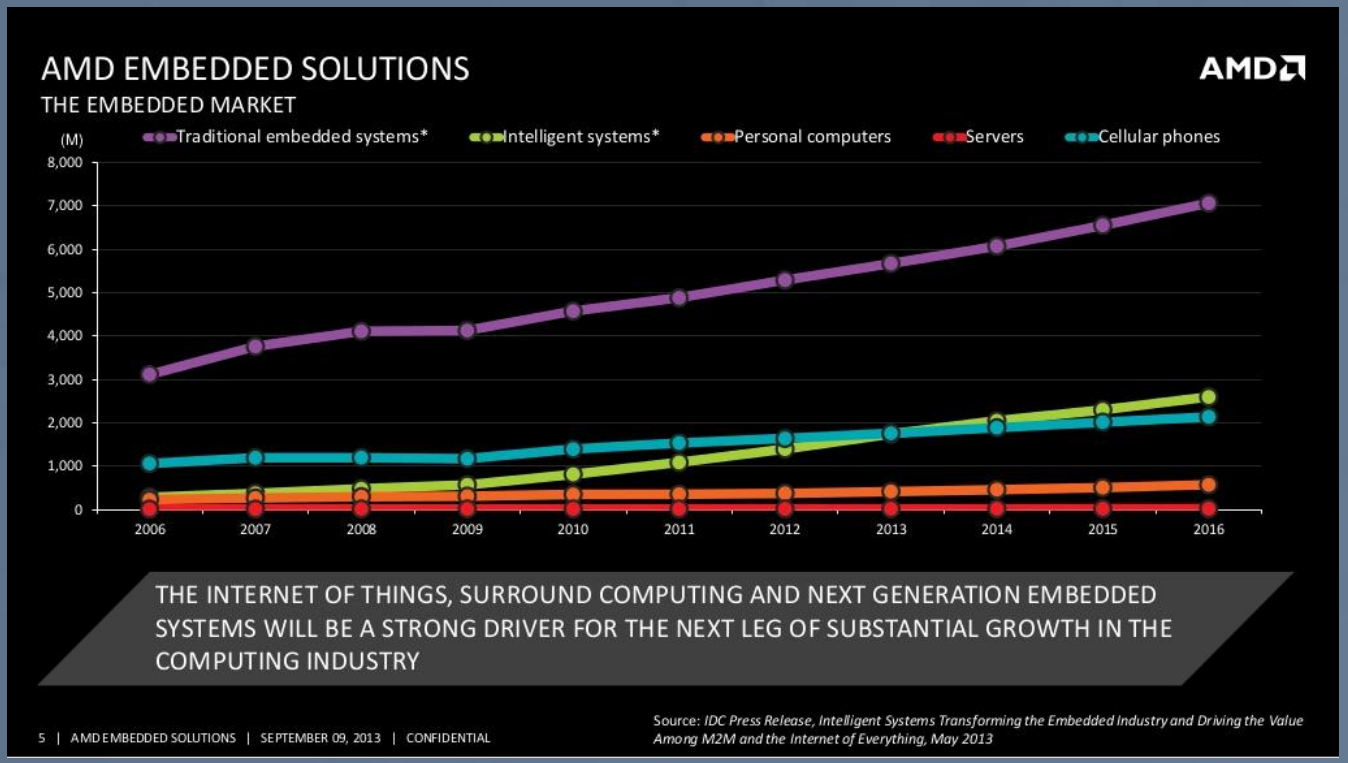
- Moore's law again
- Biggest challenges:
  - Programming & Vectorization (David's talk)
  - Memory bandwidth → stacked DRAM will come soon, should help with memory bottle-neck
- For scientific computing this domain is a match between two companies: Intel (Xeon/Phi) and Nvidia (CUDA) – but surprises (AMD) are always possible.





ATLAS study on triggering on displaced jets – up to 60 x faster than x86 version

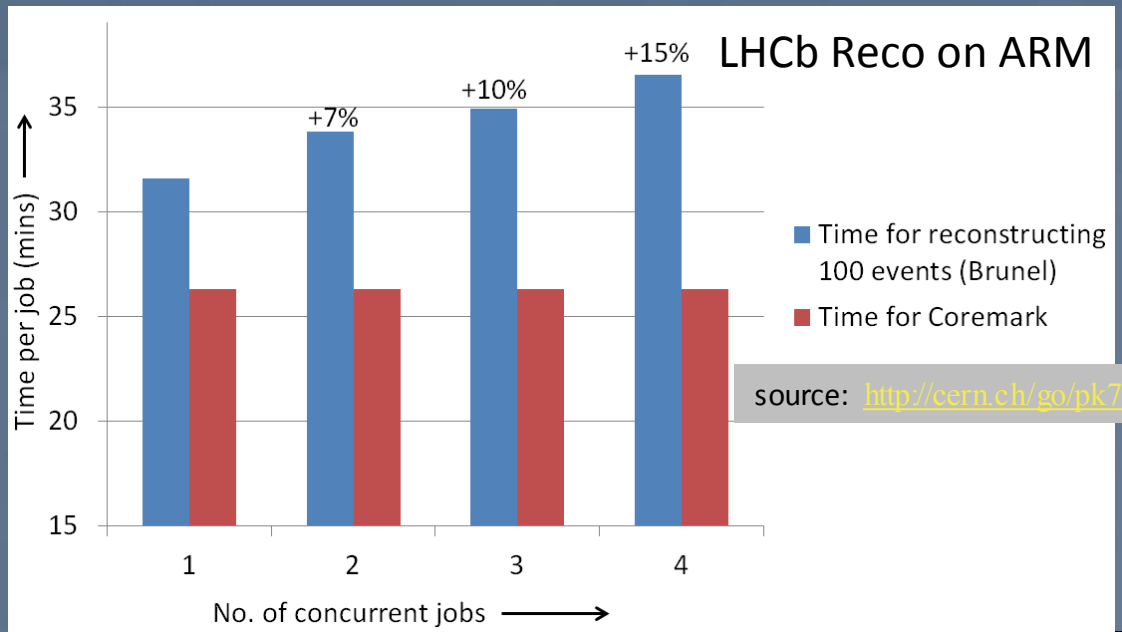
Non-x86 == ARM

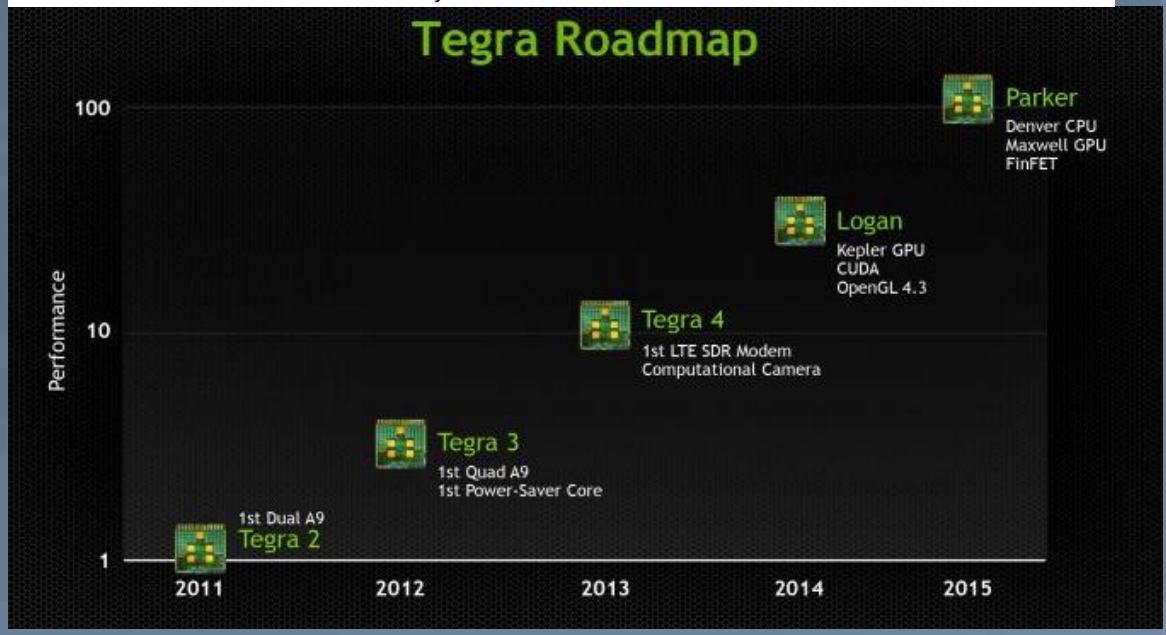


- Very strong in fast growing mobile market
- Interest in the HPC community ("Project Mont Blanc")
- Very power-efficient
- Very "European"

ARM in HEP

- Modern ARM SoC are competitive micro-servers
- HEP applications being ported
- Current challenges:
  - memory amount and bandwidth (scaling)
  - cost / performance ratio (when disregarding power)
- Moore's law holds for ARM as well → and GPUs are added to ARMs too



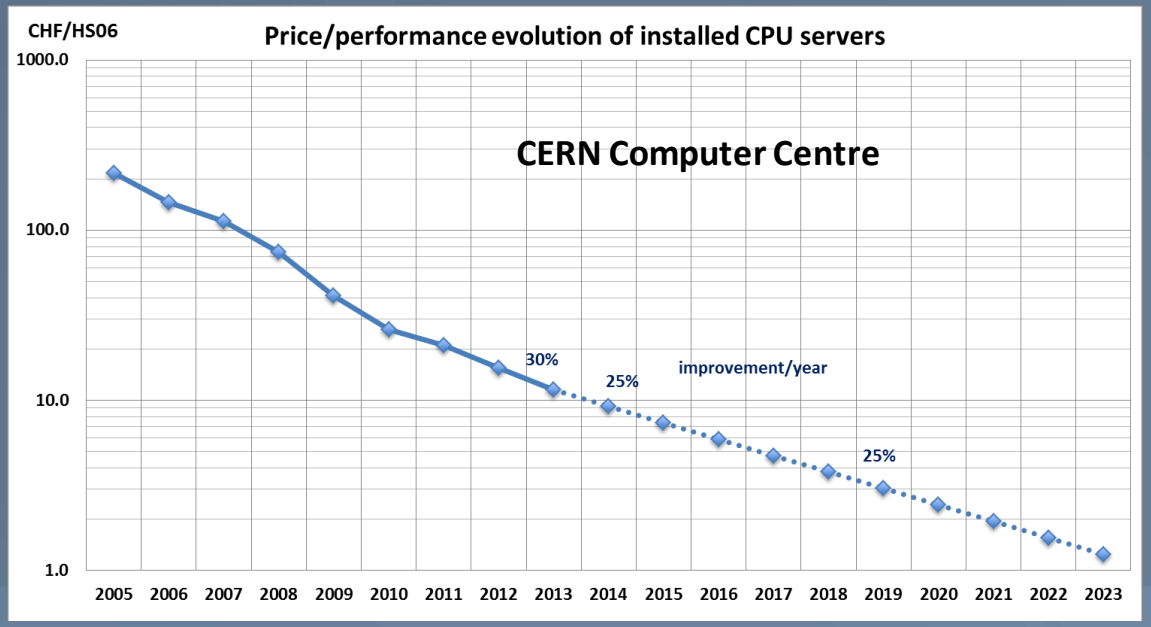


Summary (Processors)

- Principle technology roadmaps are very challenging, but not completely unrealistic
  - Clear focus on lower end consumer markets, server market at best stable
  - Very few companies can afford fabrication lines
- Price/performance improvements are achieved via:
  - Technology and fabrication (smaller structure sizes), or
  - longer amortization periods of fabs for the same chip generation
- Disadvantage
  - smaller structure sizes mean better power/performance ratio TCO, electricity costs increase
  - total number of cores increases ➔ I/O, streams per disk spindle go up

Better ratio in 2013 than expected
Assume now a 25% improvement,
IFF constant efficiency of the use of the processors by our software (not obvious! c.f. David's talk)

Reminder: 5% difference leads to a factor 1.63 over 10 years



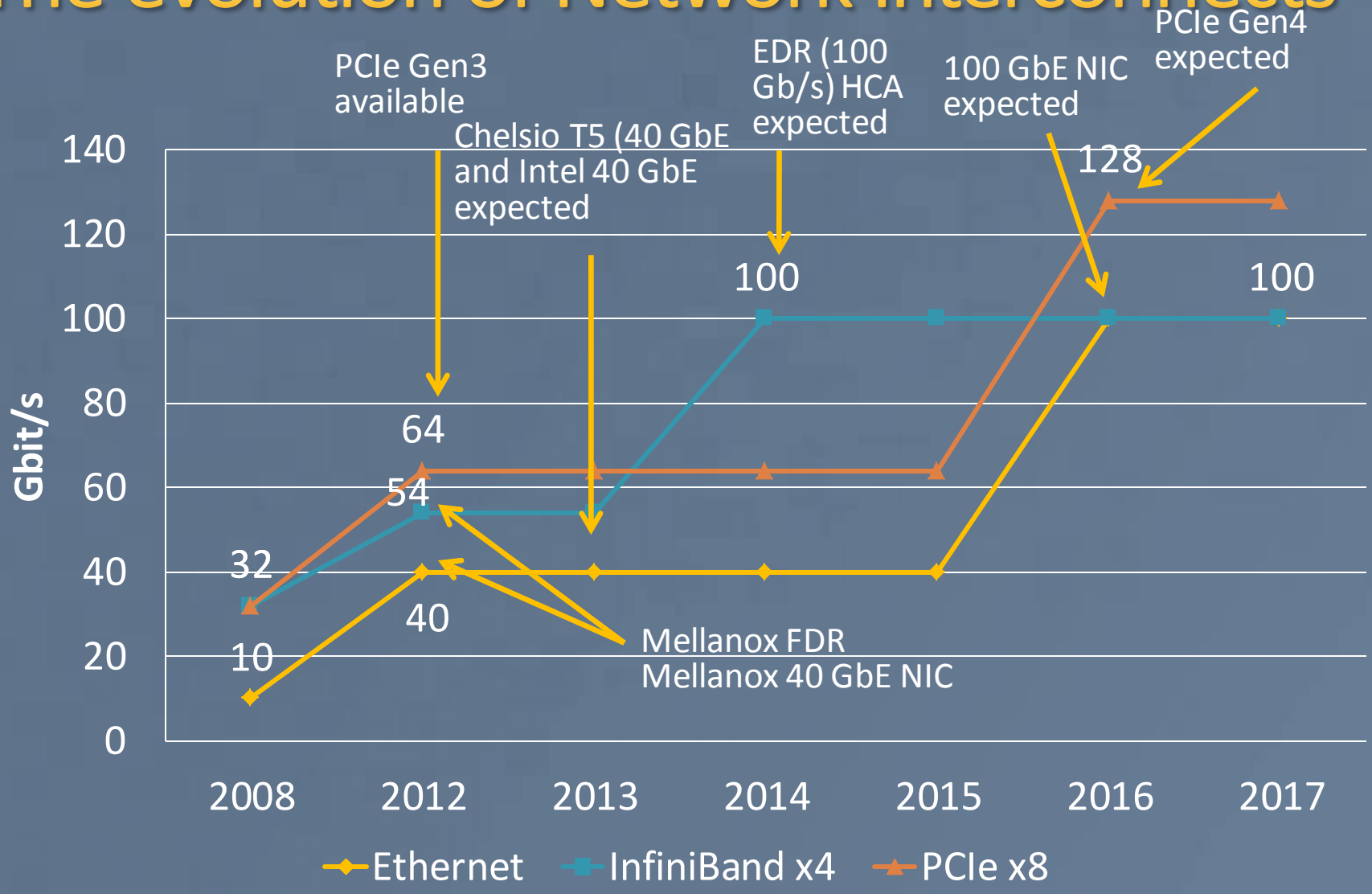
Serial interconnects & networks

Link technologies

- Commercially driven by data-center market
- For > 10 Gbit/s interfaces use 4 fibre-pairs in parallel. E.g.: 40/100 GigE, FDR/EDR InfiniBand
- Up to 150 m multi-mode optics is still cheaper than single-mode, but distances get shorter and expensive high-quality fibres (OM4) are required
- The next years will see higher integration
  - Network adapter in the CPU
  - Physical layer in the CPU (silicon photonics)

Speed [Gb/s] on *one* pair	Target Technology	Year
10	10/40 Ethernet	2000
14	FDR InfiniBand	2010
25	100G Ethernet	2011
36	??	2014
100	400G Ethernet	20??

The evolution of Network Interconnects



The evolution of switches

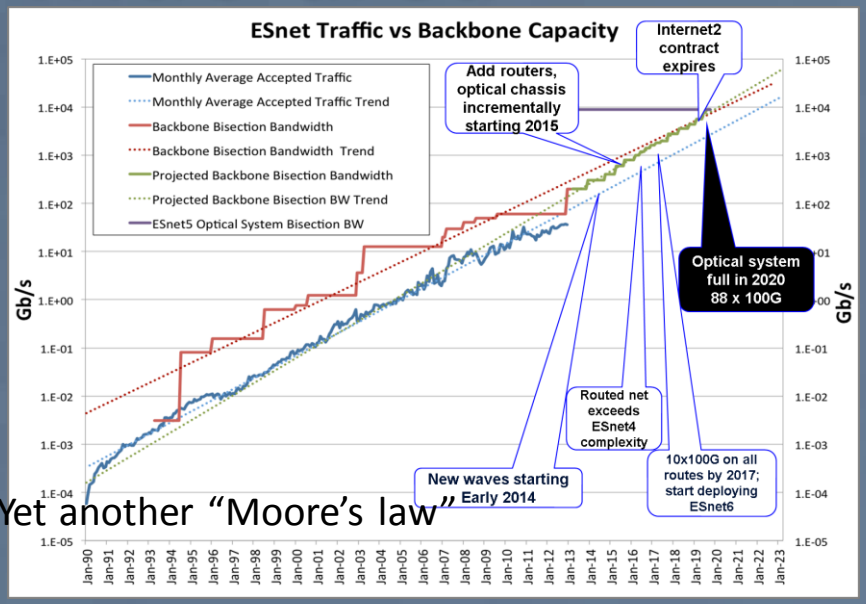
Date of release

- Brocade MLX: 768 10-GigE
- Juniper QFabric: up to 6144 10-GigE
- Mellanox SX6536: 648 x 56 Gb (IB) / 40 GbE ports
- Huawei CE12800: 288 x 40 GbE / 1152 x 10 GbE
- Each with sufficient bandwidth to run the entire Run #2 DAQ of all LHC experiments together ☺

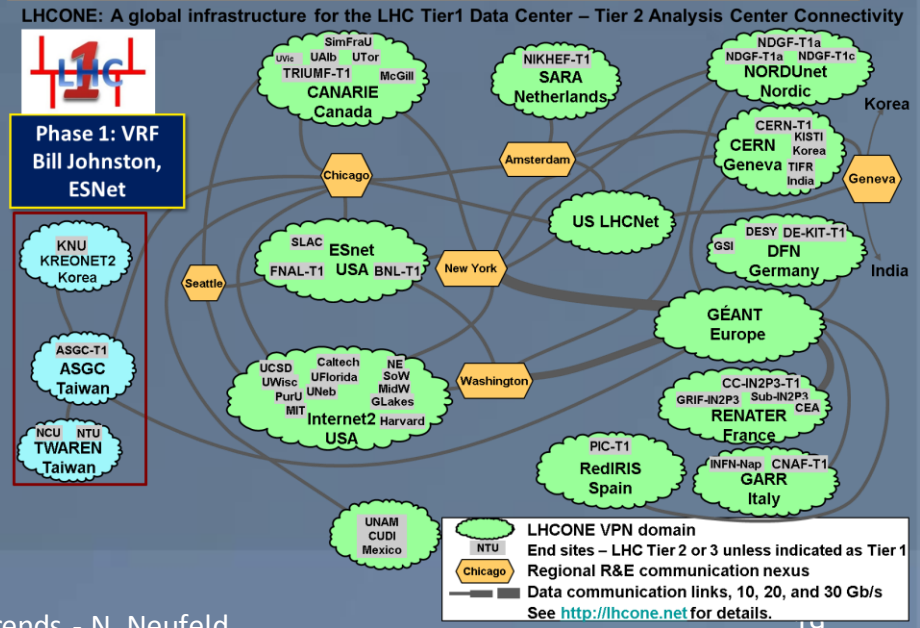



The other end: the wide-area

- ATLAS transferred 171 PB over the wide-area network (“internet”) in 2012
- Moving quickly into the 100 Gbps era, 400 Gbps not far behind
- Transatlantic 100 Gbps transmission link has been demonstrated (ANA-100 G)
- LHCONE project for future Tier1 – Tier 2 connectivity
- Technology moving forward steadily but funding needed to keep up with hardware development



Yet another “Moore’s law”



Ack. to A. Barczyk (CALTECH) for help with this slide

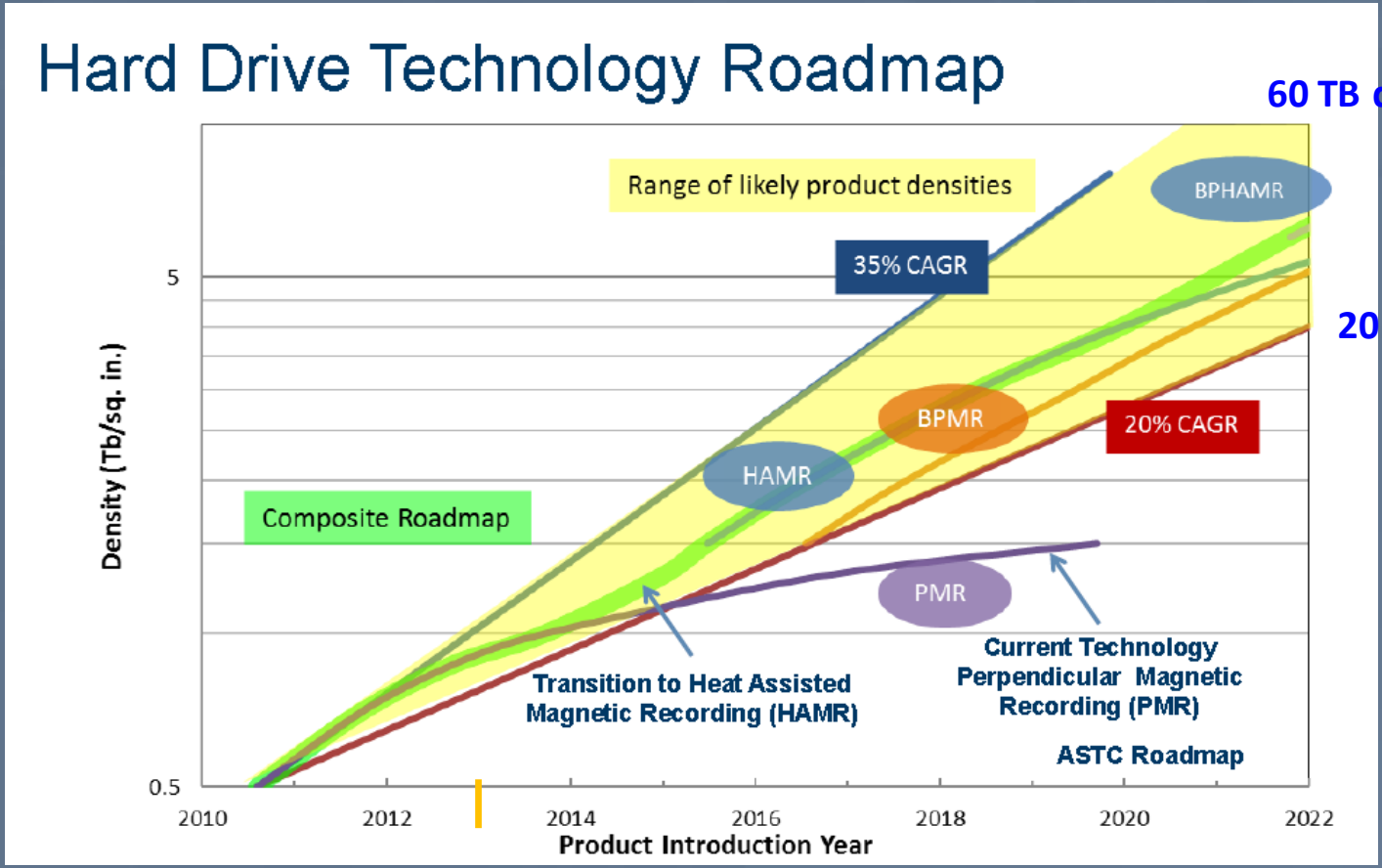
Summary interconnects and networks

- By the end of LS2 links with > 100 Gbit/s will be readily available
- Prices for networking in the local area are dropping steadily
  - The entry of silicon-photonics should reduce the price of optical links even more
- In non-radiation environments all network and link needs of LHC experiments will be satisfied by industry

	Event-size [kB]	Rate [kHz]	Bandwidth [Gb/s]	Year [CE]
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	4000	1000	32000	2022
LHCb	100	40000	32000	2019

Storage

Technology evolution: disks



Today 3- 4 TB disk

SSD are not replacing HDD any time soon - Same for non-volatile memory

Storage not a problem for Online applications (80 GB/s ALICE after LS2)

But: while capacity is increasing, parallel I/O capability is not → more applications sharing fewer disks → big challenge for offline applications

Tape

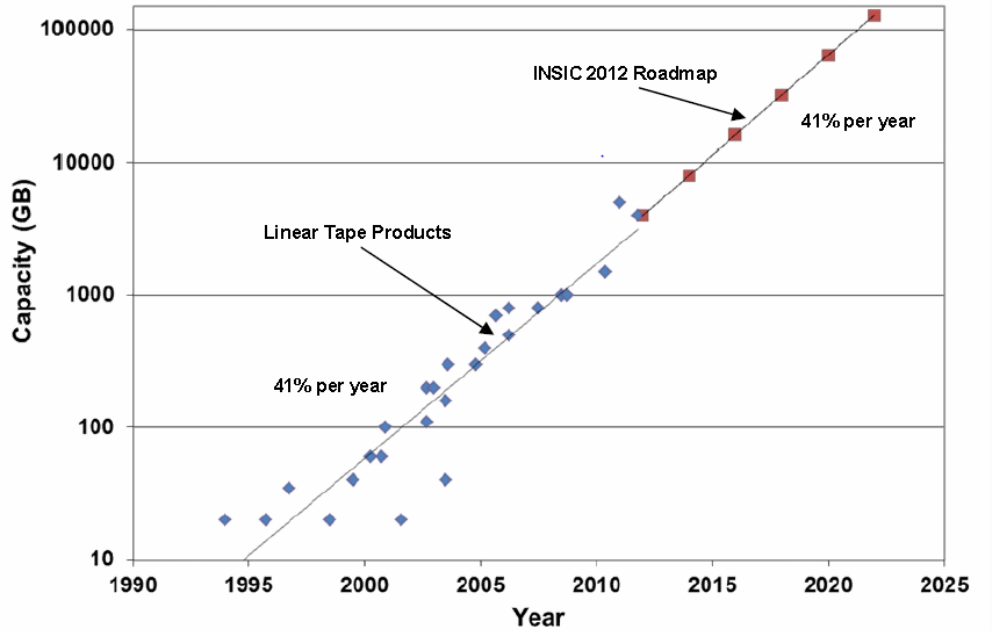


Figure 4: Tape Cartridge Capacity Trend.

International Magnetic Tape Storage Roadmap
May 2012

- Price continues to fall
- Still a price advantage over disk, but economic threshold moving out: i.e. need more and more data to justify tape infrastructure cost
- Overall a declining market

Summary: storage

- Technology roadmap for HDD and Tape is reasonable
- New HDD technologies complex and very expensive
- Market is shrinking for tapes & market for HDDs under pressure
- Price/performance improvements are achieved via:
  - Technology and fabrication (smaller structure sizes)
  - Longer amortization periods of production lines for the same HDD generation

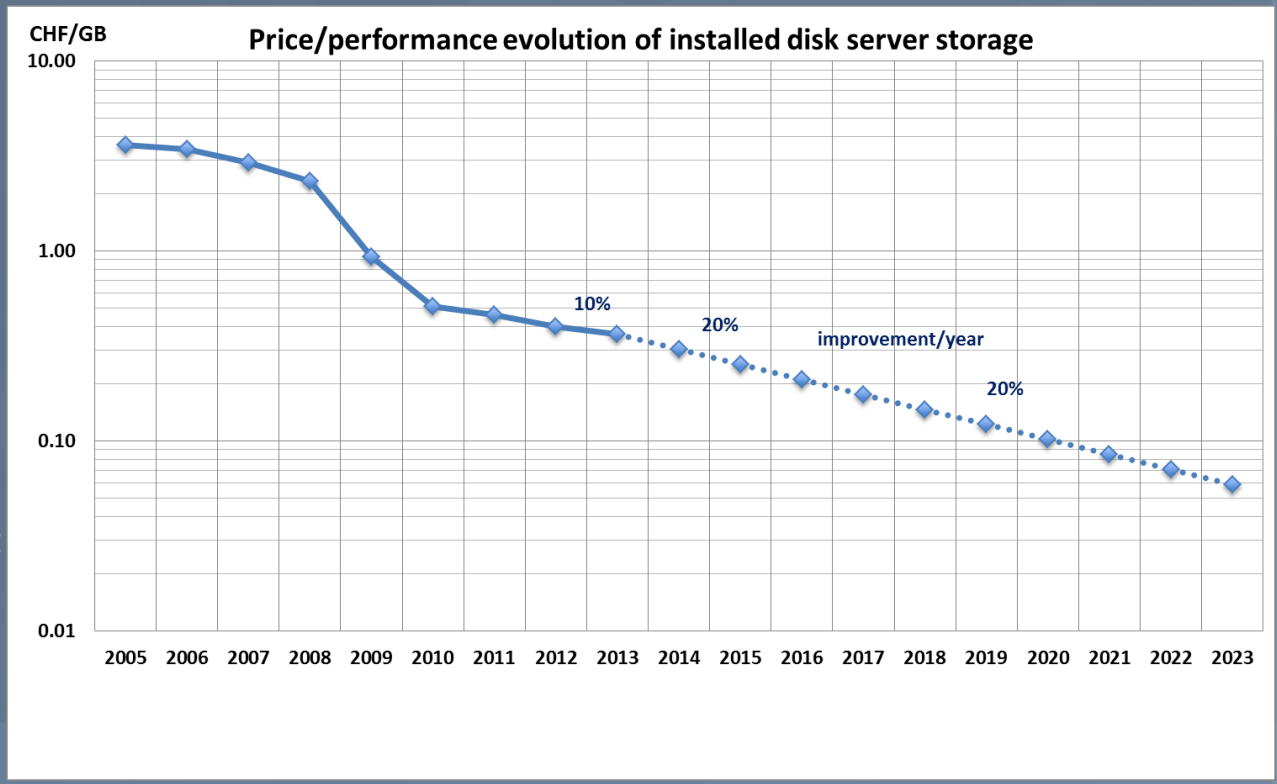
→ higher power consumption per GB, disk size not increasing (actually a good point)

Assume a factor 3? less costs for tape storage (CHF/GB)
Merge tape+disk costs?
Large site dependencies

Current focus is on space and not I/O performance, Problem ?!

2013 ratio as expected
Assume now a 20% improvement

Reminder: 5% difference leads to a factor 1.63 over 10 years



Infrastructure challenges

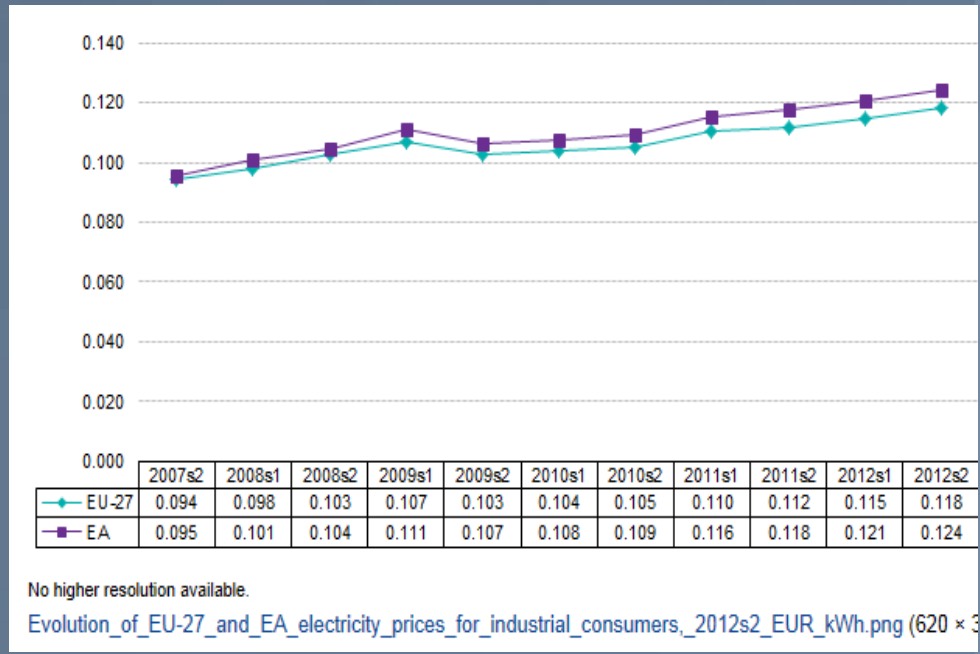
Cost structure for a site:

- CPU processing
- Disk storage
- Tape storage
- Networking (LAN and WAN)
- Services (Databases, home directory, backup, exp. Services, etc.)
- Electricity

Job characteristics:
MC and processing versus analysis I/O
10 Gbit necessity, local worker node SSDs,
Storage hierarchies, etc.

Boundary conditions:

- Constant or decreasing budget
- Limited space
- Limited cooling capacity
- Limited electricity capacity
- Lifecycle of equipment (4-5 years)



	2007s2	2008s1	2008s2	2009s1	2009s2	2010s1	2010s2	2011s1	2011s2	2012s1	2012s2
EU-27	0.094	0.098	0.103	0.107	0.103	0.104	0.105	0.110	0.112	0.115	0.118
EA	0.095	0.101	0.104	0.111	0.107	0.108	0.109	0.116	0.118	0.121	0.124

No higher resolution available.
Evolution_of_EU-27_and_EA_electricity_prices_for_industrial_consumers,_2012s2_EUR_kWh.png (620 × 3

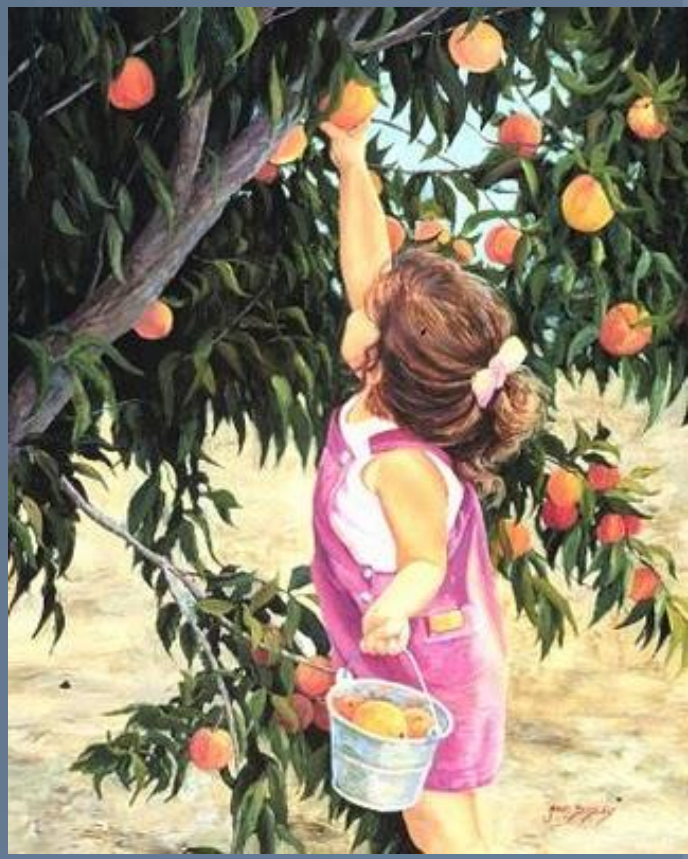
Power efficiency versus performance
Disk to tape ratio,
availability of tape technology

Cost for electricity is increasing at 4-5% per year (average euro-zone)
→ 60% in 10 years

Summary

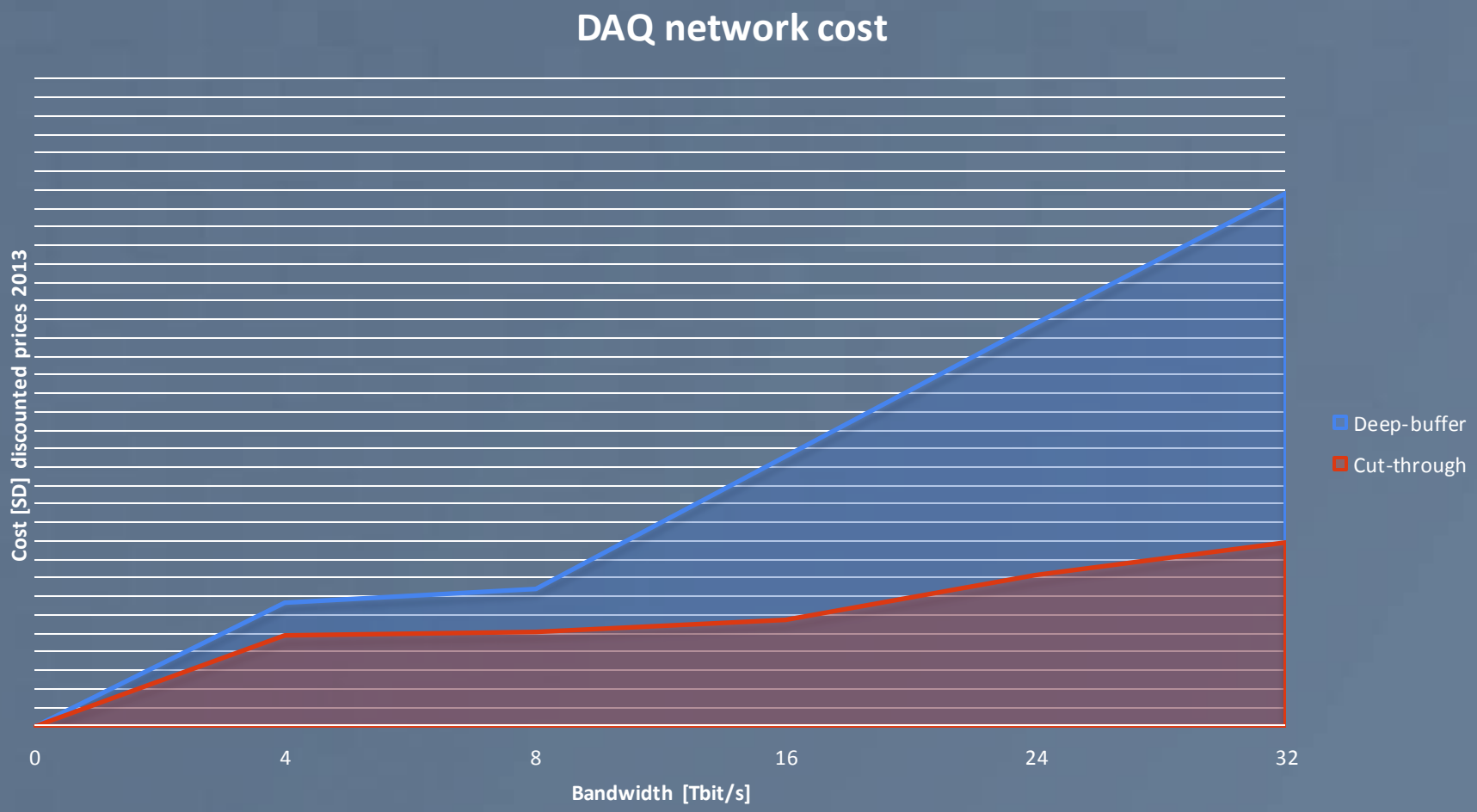
- Big-data and cloud-computing drive market for Commercial Off The Shelf IT equipment → HEP profits from that, but depends on economy at large
- We expect current performance rates (and price performance improvements) to grow at historic rates until at least LS2
  - 25% performance improvement per year in computing at constant cost
  - local area network and link technology sufficient for all HL-LHC needs also beyond LS2
  - wide area network growth sufficient for LHC needs, provided sufficient funding
  - 20% price-drop at constant capacity expected for disk-storage,
- Far beyond LS2 the technical challenges for further evolution seem daunting. Nevertheless the proven ingenuity and creativity of the IT justify cautious optimism
- The fruits of technology are there, but hard work is needed to make the best of it

source http://paintingmax.blogspot.com



More material

Cost of multi Tbit/s DAQ networks



based on a realistic model for the LHCb upgrade