Trends in Computing Technologies and Markets: The HEPiX TechWatch WG

Helge Meinhard / CERN

Contributions by Andrea Sciaba, Shigeki Misawa, Chris Holloway, Michele Michelotto, Edoardo Martelli, Rolf Seuster

EGI Conference, Amsterdam, 06-08 May 2019

Challenges

- Huge computing demands of Big Science projects
  - LHC Run 4 / phase II
  - DUNE
  - SKA
  - Other Big Sciences
- Understanding better where technology and markets go becomes indispensable
- Wide area of topics to be covered in a regular and systematic manner

HEPiX Techwatch Working Group

- Evolution previously tracked by individuals
- More systematic effort required
- A lot of expertise and interest in community

- Proposal in spring 2018: Form working group
- Main purposes:
  - Follow trends, better predict costs, optimise investments
  - Provide input to computing models and software development

Techwatch WG – Current Status (1)

- 59 persons subscribed

- Scope: Evolution of technology and markets
  - Excludes benchmarking, cost modelling, software techniques and products
- Target communities: HEP experiments, others as they collaborate and contribute to WG

Techwatch WG – Current Status (2)

- Chairs: Helge Meinhard, Bernd Panzer-Steindel
- Six sub-groups (with conveners):
  - General market trends, semiconductor markets, unit sales (Servesh Muralidharan, Peter Wegner)
  - Server markets (Chris Hollowell, Michele Michelotto)
  - CPUs and accelerators (Andrea Sciaba, Eric Yen)
  - Memories (Shigeki Misawa)
  - Storage (German Cancio, Martin Gasthuber)
  - Network (Edoardo Martelli)

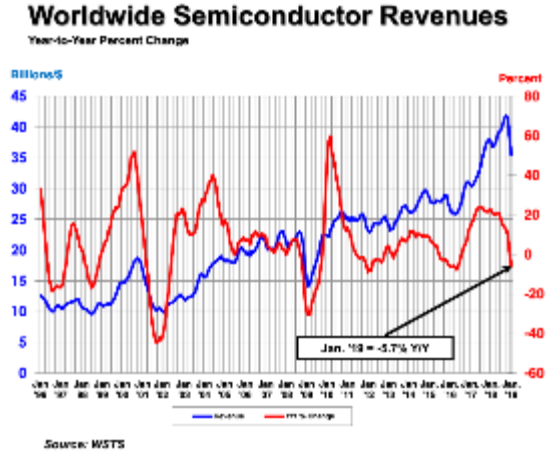
Techwatch WG – Deliverables

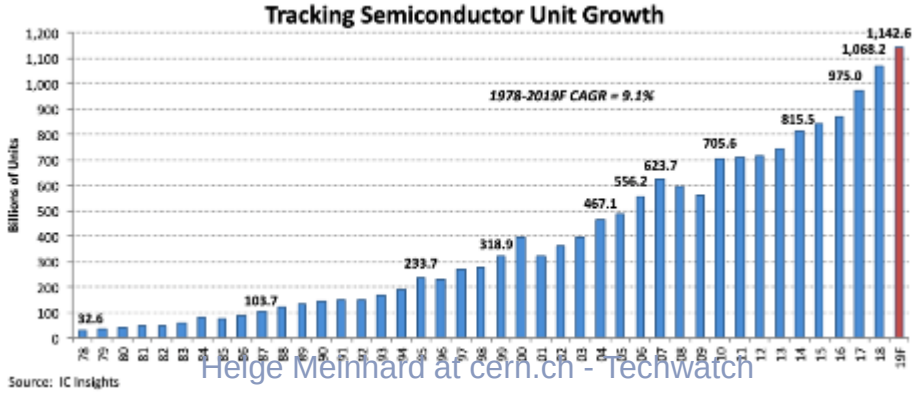
- Presentations at meetings / workshops / conferences
- Set of maintained Web pages
- Regular publications (e.g. in CSBS)

- Common follow-up with other WGs and stakeholders

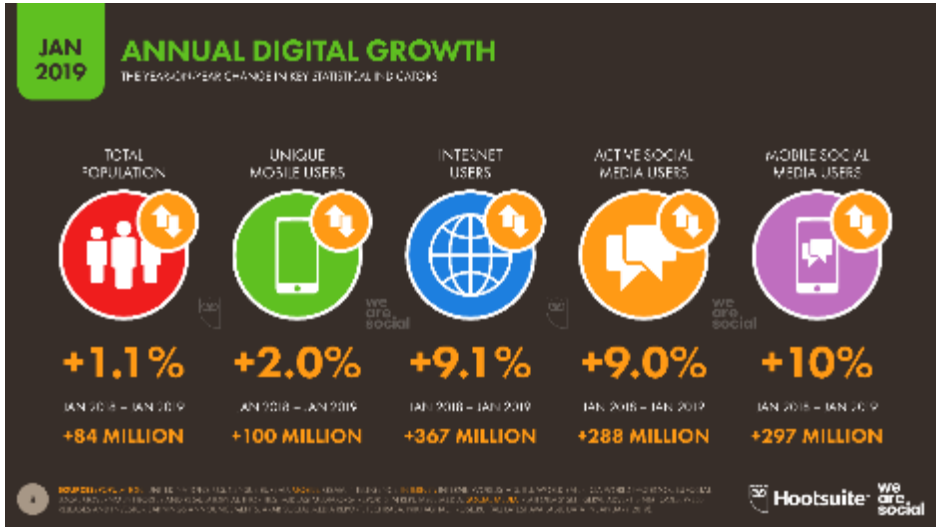
Semiconductor Device Market and Trends

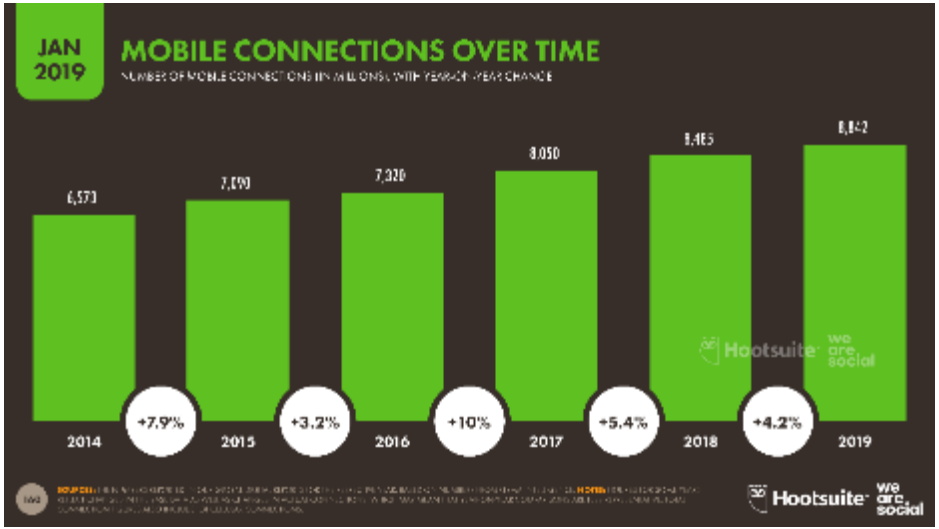
- Global demand for semiconductors topped 1 trillion units shipped for the first time
- Global semiconductor sales got off to a slow start in 2019, as year-to-year sales decreased
- Long-term outlook remains promising, due to the ever-increasing semiconductor content in a range of consumer products
- Strongest unit growth rates foreseen for components of
  - smartphones
  - automotive electronics systems
  - devices for deep learning applications





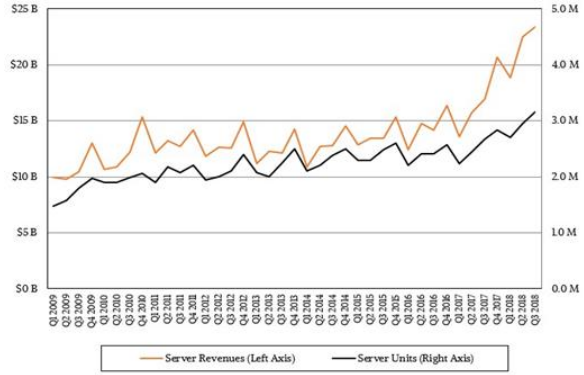
Internet and Smart Population Growth and Effects

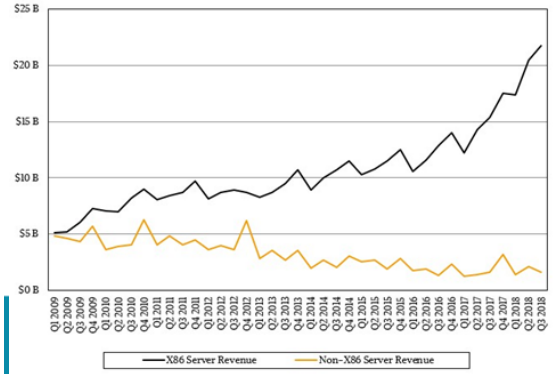




- Small changes in smart population trend from 2018
- Significant increase in mobile social media usage over the past year
  - Otherwise pretty stagnating

Server Market (1)





- Global server market revenue up substantially
  - $23.37 billion in Q3 2018 with 3.16 million servers shipped
  - Primarily due to hyperscale cloud provider purchases
  - May grow to $100 billion/year industry
- Revenue outpacing units sold
  - Increased cost of components including CPU, GPUs/FPGAs, memory and solid state storage
- Total server shipments per quarter now almost double what they were for an entire year ~20 years ago (1997)
  - Vast majority of revenue from x86 servers
  - Other server platforms only sold $1.6 billion in Q3 2018: ~6.8%
- New ARM Neoverse server CPU platform hopes to change the downward trend for non-x86 servers

Server Market (2)

- ODMs have larger market share than any single Tier-1 brand (Dell, HPE, Lenovo)
- 25% units purchased by hyperscale cloud providers
- AMD share still small
  - But was zero before EPYC Naples
  - Driven by hyperscale cloud providers
- Data centres responsible of 3% of total electricity consumption in 2017
  - Expected to raise to 20% by 2025
  - Efficiency (in terms of PUE, and in terms of compute per Watt) is key
  - Rack power typically 12…20 kW now, can be up to 40 kW
    - Sophisticated cooling required
- Open19 and Open Compute Platform are interesting approaches
  - Little uptake so far

CPUs: Intel for Servers (1)

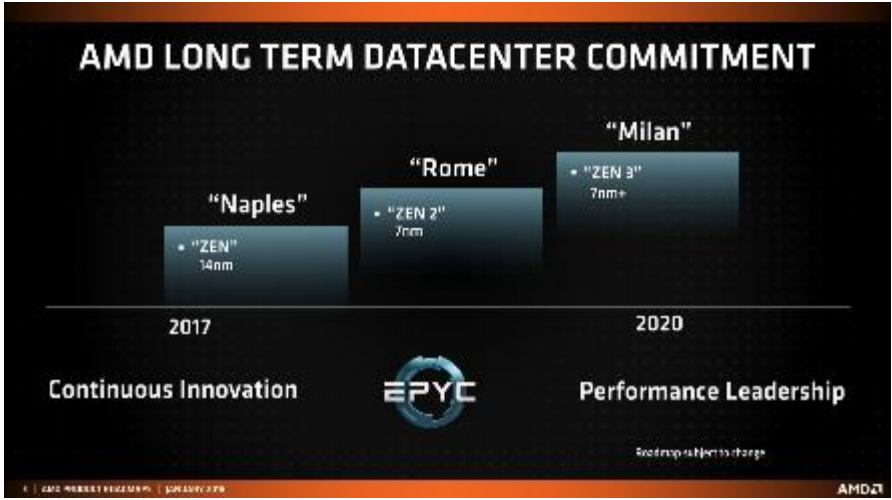
- Intel Xeon Scalable Processors
  - Currently based on Skylake-SP and coming in four flavours, up to 28 cores
- Only minor improvements foreseen for 2019
  - Adding support for Optane DC Persistent Memory and hardware security patches
- New microarchitecture (Sunny Cove) to become available late 2019
  - Several improvements benefiting both generic and specialised applications

CPUs: Intel for Servers (2)

Microarchitecture	Technology	Launch year	Highlights
Skylake-SP	14nm	2017	Improved frontend and execution units More load/store bandwidth Improved hyperthreading AVX-512
Cascade Lake	14nm++	2019	Vector Neural Network Instructions (VNNI) to improve inference performance Support 3D XPoint-based memory modules and Optane DC Security mitigations
Cooper Lake	14nm++	2020	bfloat16 (brain floating point format)
Sunny Cove (aka Ice Lake)	10nm+	2019	Single threaded performance New instructions Improved scalability Larger L1, L2, µop caches and 2nd level TLB More execution ports
Willow Cove	10nm	2020?	Cache redesign New transistor optimization Security Features
Golden Cove	7/10nm?	2021?	Single threaded performance AI Performance Networking/5G Performance Security Features

CPUs: AMD for Servers (1)

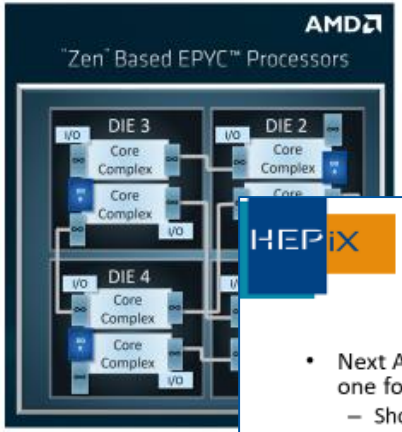
- EPYC 7000 line-up from 2017
  - Resurgence after many years of Bulldozer CPUs thanks to the Zen microarchitecture
    - +40% in IPC, almost on par with Intel
    - 2x power efficiency vs Piledriver
  - Up to 32 cores
- Already being tested and used at some WLCG sites



CPUs: AMD for Servers (2)

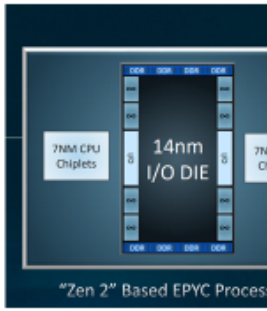
EPYC Naples

- EPYC Naples (Zen) consists of up to 4 separate dies, interconnected via Infinity Fabric
  - Chiplets allow a significant reduction in cost and higher yield
- Main specifications
  - up to 32 cores
  - 4 dies per chip (14nm), each die embedding IO and memory controllers
  - 2.0-3.1 GHz of base frequency
  - 8 DDR4 memory channels with hardware encryption
  - up to 128 PCI gen3 lanes per processor (64 in dual)
  - TDP range: 120W-200W
- Similar per-core and per-GHz HS06 performance to Xeon



EPYC Rome

- Next AMD EPYC generation (Zen 2), embeds 9 dies, including one for I/O and memory access
  - Should compete with Ice Lake
- Main specs:
  - 9 dies per chip : a 14nm single IO/memory die and 8 CPU 7nm chiplets
    - +300-400 MHz for low core count CPUs
  - 8 DDR4 memory channels, up to 3200 MHz
  - up to 64 cores
  - up to 128 PCI Gen3/4 lanes per processor
  - TDP range: 120W-225W (max 190W for SP3 compatibility)
  - Claimed +20% performance per-core over Zen, +75% through the whole chip with similar TDP over Naples
  - To be released during 2019



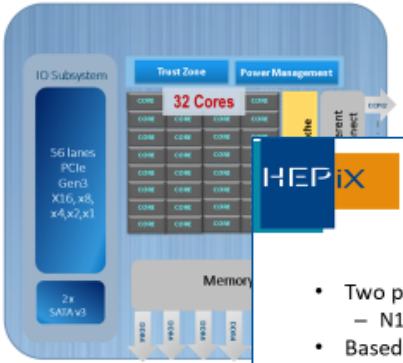
CPUs: ARM in Data Centre (1)

- ARM is ubiquitous in the mobile and embedded CPU world
- Data centre implementations have been relatively unsuccessful so far
  - Performance/power and performance/$ not competitive with Intel and AMD
- Only a few implementations (potentially) relevant to the data center
  - Cavium ThunderX2
  - Fujitsu A64FX
  - ARM Neoverse
  - Ampere eMAG, Graviton
- LHC experiments are capable of using ARM CPUs if needed
  - Some do nightly builds on ARM since years

CPUs: ARM for Data Centre (2)

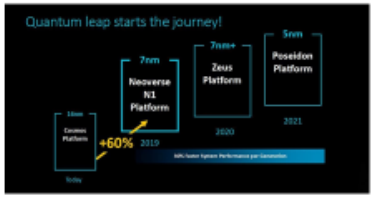
Marvell ThunderX2 and Fujitsu A64FX

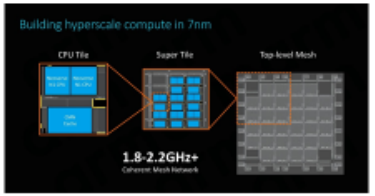
- ThunderX2 for mainstream cloud and HPC data centers, from 2018
  - Enjoys the greatest market visibility and reasonable performance/$
    - Used e.g. at CRAY XC-50 at Los Alamos and HPE Apollo 70 based Astra HPC system at Sandia National Laboratory
  - ARM V8.1 architecture
    - Up to 32 cores, 4-way SMT
    - Up to 8 DDR4 memory channels
    - Up to 56 PCIe Gen3 lanes
- Fujitsu A64FX to be used in supercomputer at RIKEN center
  - Based on the V8.2-A ISA architecture
    - First to deliver scalable vector extensions (SVE)
    - 48 cores
    - 32 GB of HBM2 high bandwidth memory
    - 7nm FinFET process
  - Interesting to see what performance will achieve as it may lead to a more competitive product



ARM Neoverse

- Two platforms for the data center
  - N1 for cloud, E1 for throughput
- Based on the Neoverse N1 CPU
  - Very similar architecture to Cortex A76 but optimized for high clock speeds (up to 3.1 GHz)
  - Two N1 cores each with L1 and L2 caches
  - To be combined by licensees with memory controller, interconnect and I/O IP
- Demonstrated the N1 Hyperscale Reference Design
  - 64-128 N1 CPUs each with 1 MB of private L2
  - 8x8 mesh interconnect with 64-128 MB of shared cache
  - 128x PCIe/CCIX lanes
  - 8x DDR4 memory channels
- Intended to strengthen ARM's server market share
  - Not expected to be available for another 1-1.5 years





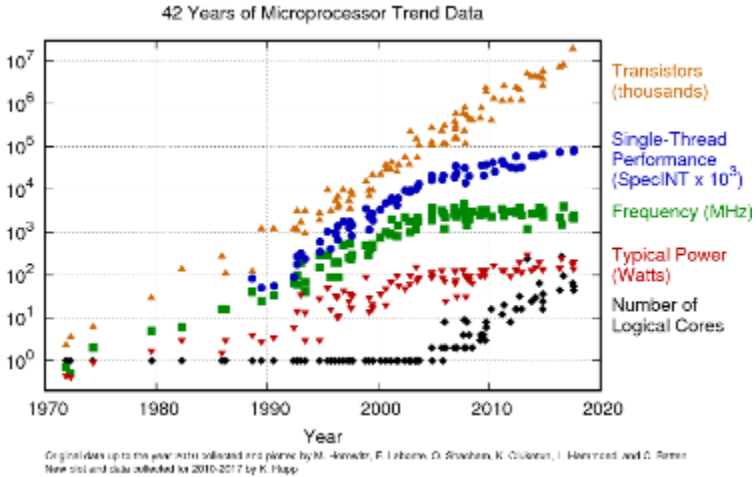
Source: Anandtech

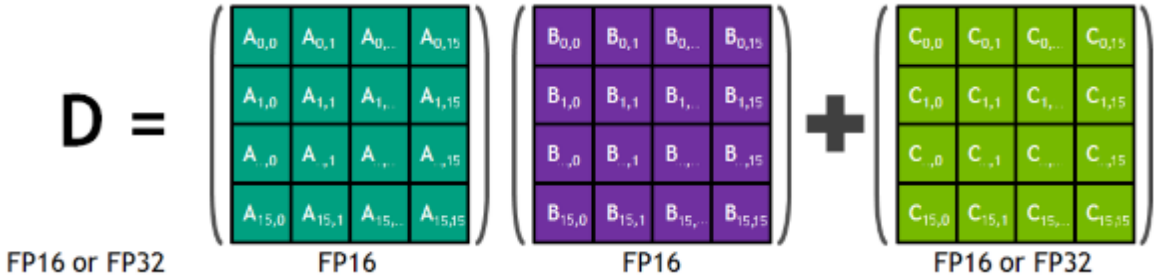
CPUs: IBM POWER, RISC-V, MIPS

- POWER9 (2017)
  - Used in Summit, the fastest supercomputer
  - 4 GHz
  - Available with 4-way (up to 24 cores)
  - First supporting PCIe-Gen4
  - CAPI 2.0 I/O to enable
    - Coherent user-level access to accelerators and I/O devices
    - Access to advanced memories
  - NVLink to increase bandwidth to Nvidia GPUs
  - 14nm FINFET process
  - Product line with full support for RHEL/CENTOS7
- POWER10 (2020)
  - 10nm process
  - Several feature enhancements
  - First to support PCIe Gen5
- RISC-V: Open-source ISA for controllers
  - Not currently targeting data centre
  - May evolve into competition with ARM
- MIPS: Dead horse

Discrete GPUs

- Raw power follows the exponential trend on numbers of transistors and cores
- New features appear unexpectedly, driven by market (e.g. tensor cores)
  - Tensor cores: programmable matrix-multiply-and-accumulate units
  - Fast half precision multiplication and reduction in full precision
  - Useful for accelerating deep learning training/inference





https://devblogs.nvidia.com/programming-tensor-cores-cuda-9/

GPUs: Nvidia and AMD

Nvidia: Volta addressing servers, Turing addressing gaming

Feature	Volta (V100)	Turing (2080 Ti)
Process	12nm	12nm
CUDA cores	yes	yes
Tensor cores	yes	yes
RT cores	NA	yes
FP performance	FP16: 28 TFLOPS FP32: 14 TFLOPS FP64: 7 TFLOPS Tensor: 112 TFLOPS	Same, but FP64: 1/32 of FP32
Memory	HBM2	GDDR6
Memory bandwidth	900 GB/sec	616 GB/sec
Multi-GPU	NVLink 2	NVLink 2/SLI
Applications	AI, datacenter, workstation	AI, workstation, gaming

- AMD: Vega 20
  - Directly aimed at the server world (Instinct MI50 and MI60)
- Evolution of Vega 10 using a 7nm process
  - more space for HBM2 memory, up to 32GB
  - 2x memory bandwidth
  - Massive FP64 gains
  - PCIe Gen4
- Some improvements relevant for inference scenarios
  - Support for INT8 and INT4 data types
  - Some new instructions

GPUs: Programming

- NVIDIA CUDA:
  - C++ based (supports C++14), de-facto standard
  - New hardware features available with no delay in the API
- OpenCL:
  - Can execute on CPUs, AMD GPUs and recently Intel FPGAs
  - Overpromised in the past, with scarce popularity
- Compiler directives: OpenMP/OpenACC
  - Latest GCC and LLVM include support for CUDA backend
- AMD HIP:
  - Interfaces to both CUDA and AMD MIOpen, still supports only a subset of the CUDA features
- GPU-enabled frameworks to hide complexity (Tensorflow)
- Issue is performance portability and code duplication

GPUs in LHC Experiments

- ALICE: O2
  - Tracking in TPC and ITS
  - Modern GPU can replace 40 CPU cores
- CMS, CMSSW
  - Demonstrated advantage of heterogeneous reconstruction from RAW to Pixel Vertices at the CMS HLT
  - ~10x both in speed-up and energy efficiency wrt full Xeon socket
  - Plans to run heterogeneous HLT during LHC Run3
- LHCb (online - standalone) Allen framework: HLT-1 reduces 5TB/s input to 130GB/s:
  - Track reconstruction, muon-id, two-tracks vertex/mass reconstruction
  - GPUs can be used to accelerate the entire HLT-1 from RAW data
  - Events too small, have to be batched: makes the integration in Gaudi difficult
- ATLAS
  - Prototype for HLT track seed-finding, calorimeter topological clustering and anti-kt jet reconstruction
  - No plans to deploy this in the trigger for Run 3

FPGAs, Others

FPGA

- Players: Xilinx (US), Intel (US), Lattice Semiconductor (US), Microsemi (US), and QuickLogic (US), TSMC (Taiwan), Microchip Technology (US), United Microelectronics (Taiwan), GLOBALFOUNDRIES (US), Achronix (US), and S2C Inc. (US)
- Market valued at USD 5 Billion in 2016 and expected to be valued at 10 Billion in 2023
- Growing demand for advanced driver-assistance systems (ADAS), developments in IoT and reduction in time-to-market are the key driving factors

Process Technology	20 nm		16 nm		14 nm	
	Intel®	Xilinx®	Intel®	Xilinx®	Intel®	Xilinx®
Top Performance Tier		Virtex® UltraScale		Virtex® UltraScale+® Zynq® UltraScale+®	Intel® Stratix® 10	
Mid Performance Tier	Intel® Arria® 10	Kintex UltraScale				
Low Performance Tier	Intel® Cyclone® 10 GX					

Source: https://www.intel.com/content/www/us/en/programmable/documentation/mtr1422491996806.html#gom1512594527835__fn_soc_variab_avail_xlx

26

FPGA programming

- Used as an application acceleration device
  - Targeted at specific use cases
    - Neural inference engine
    - MATLAB
    - LabVIEW FPGA
- OpenCL
  - Very high level abstraction
  - Optimized for data parallelism
- C / C++ / System C
  - High level synthesis (HLS)
  - Control with compiler switches and configurations
- VHDL / Verilog
  - Low level programming

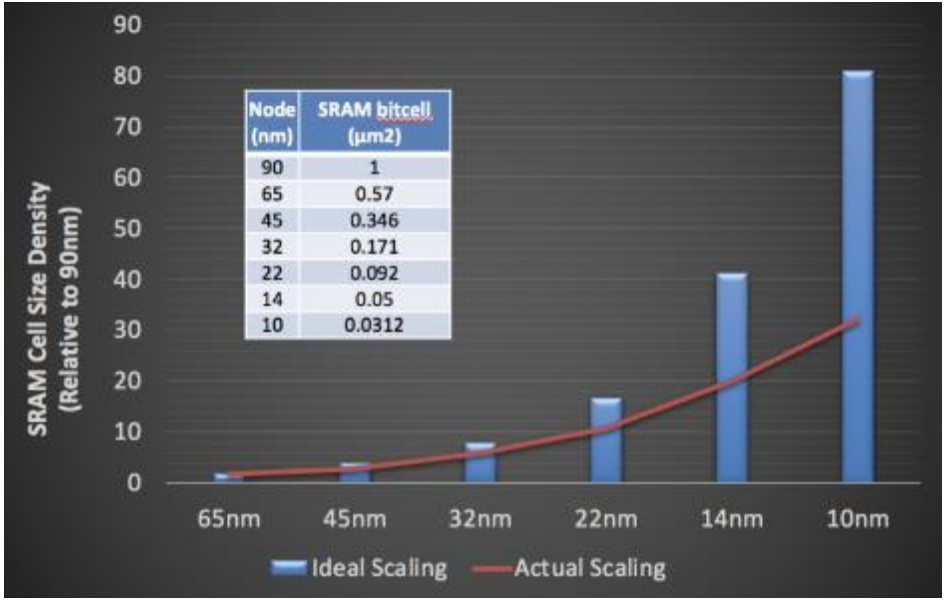
- In HEP
  - High Level Triggers
    - https://cds.cern.ch/record/2647951
  - Deep Neural Networks
    - https://arxiv.org/abs/1804.06913
    - https://indico.cern.ch/event/703881/
  - High Throughput Data Processing
    - https://indico.cern.ch/event/669298/

27

- Intel Nervana, Google TPU (huge boost in perf/W wrt CPU and GPU), Intel Configurable Spatial Accelerator

Memories: Static RAM

- On-die memory on the CPU used for L1/L2/L3 cache
  - SRAM cell size not scaling with node
    - SRAM cache constitutes large fraction of area on modern CPUs
- Power consumption is an issue
- Applications driving larger caches
- No direct replacement in sight for L1/L2
- Alternate L3 cache technologies
  - eDRAM - Used in IBM Power CPUs
  - STT-MRAM - proposed as possible replacement

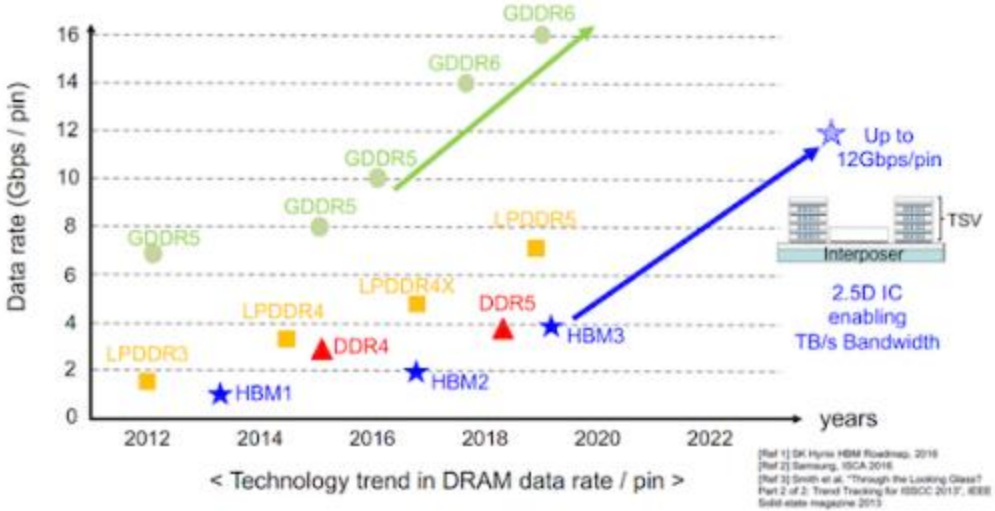


Node (nm)	SRAM bitcell (µm2)
90	1
65	0.57
45	0.346
32	0.171
22	0.092
14	0.05
10	0.0312

https://www.sigarch.org/whats-the-future-of-technology-scaling/

Memories: Dynamic RAM

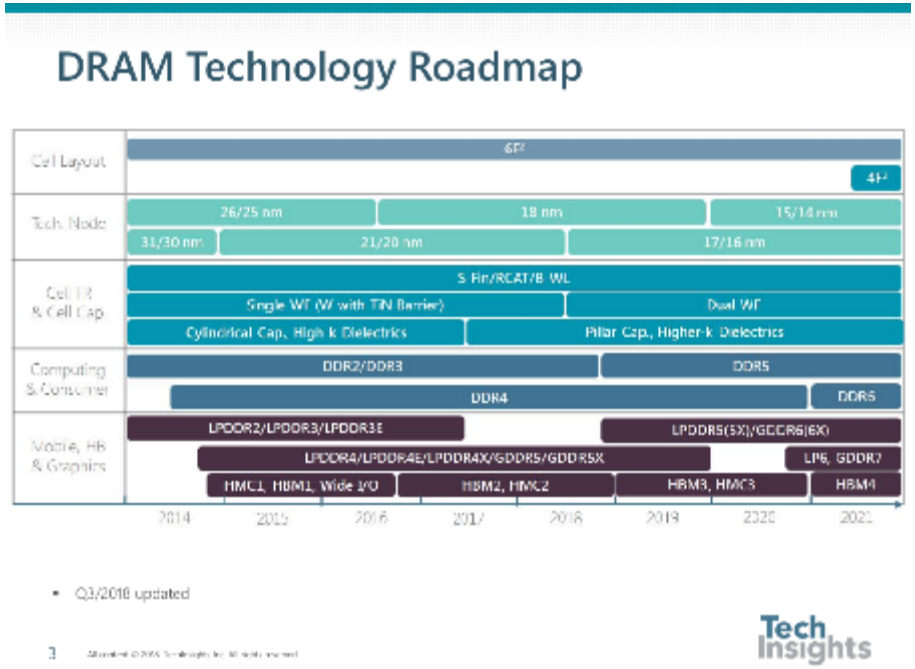
- Dominant standards continue to evolve
  - DDR4 -> DDR5
    - 3200MT/s -> 6400MT/s
    - 16Gb -> 32Gb chips
  - GDDR5 - > GDDR5X
    - 14 Gbps/pin -> 16Gbps/pin
    - 8Gb -> 16Gb chips
  - HBM -> HBM2
    - 1 Gbps/pin -> 2.4 Gbps/pin
    - 4 die stack -> 12 die stack
    - 2Gb die -> 8Gb die
- Memory latency remains mostly unchanged



(Youngwoo Kim, KAIST's Terabyte Labs)
https://www.3dincites.com/2019/02/designcon-2019-shows-board-and-system-designers-the-benefits-of-advanced-ic-packaging/

Memories: DRAM Outlook

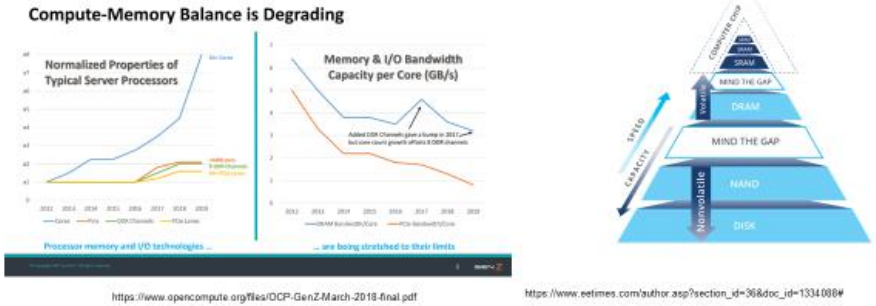
- Major vendors showing next generation chips (DDR5/GDDR6)
- Multiple technologies being investigated for future DRAM
- EUV lithography not needed for at least 3 more generations (Micron)
- Contract DRAM pricing fell ~30% in Q1 2019
- Pressure expected on DRAM prices through 2019 due to additional production capacity coming online



https://www.techinsights.com/technology-intelligence/overview/technology-roadmaps/

Memories: Issues, Emerging Technologies

Performance gaps in memory hierarchy

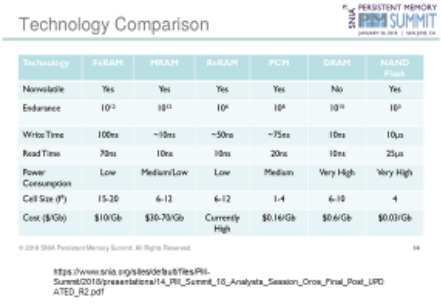


https://www.opencompute.org/files/OCP-GenZ-March-2018-final.pdf

https://www.eetimes.com/author.asp?section_id=36&doc_id=1334088#

Emerging technologies

- May eventually fill the gap
  - STT-MRAM between SRAM and DRAM (work in progress)
  - "Persistent Memory" in NVDIMM package for the DRAM/NAND gap
    - Low latency NAND (e.g. Z-NAND)
    - 3D XPoint (aka "Optane")
  - Technologies still in the lab
    - MRAM
    - NRAM
    - FeRAM
    - PCRAM
    - ReRAM



Technology	FeRAM	MRAM	ReRAM	PCM	DRAM	NAND Flash
Nonvolatile	Yes	Yes	Yes	Yes	No	Yes
Endurance	10^{12}	10^{12}	10^{8}	10^{8}	10^{16}	10^{5}
Write Time	100ns	~10ns	~50ns	~75ns	10ns	10µs
Read Time	70ns	10ns	10ns	20ns	10ns	25µs
Power Consumption	Low	Medium/Low	Low	Medium	Very High	Very High
Cell Size (F^2)	15-20	6-12	6-12	1-4	6-10	4
Cost ($/Gb)	$10/Gb	$30-70/Gb	Currently High	$0.16/Gb	$0.6/Gb	$0.03/Gb

https://www.snia.org/sites/default/files/PM-Summit/2018/presentations/14_PM_Summit_18_Analysts_Session_Cross_Final_Post_UPDATED_R2.pdf

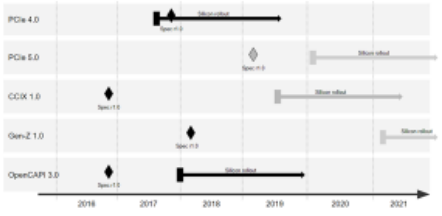
Any fab for new technologies requires multi-billion investment – not likely to happen

Interconnects and Packaging

Interconnect technology

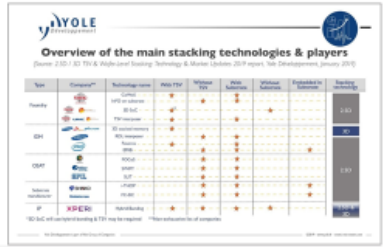
- Increasing requirements on bandwidth and latency driving the development
  - E.g. moving data between CPU and GPU is often a bottleneck
  - Several standards competing (PCIe Gen4/5, CCIX, Gen-Z, OpenCAPI, CXL…)
- Proprietary technologies
  - NVLink (GPU-to-GPU, GPU-to-POWER9)
  - Ultra Path (Intel), CPU-to-CPU
  - Infinity Fabric (AMD), chiplet-to-chiplet

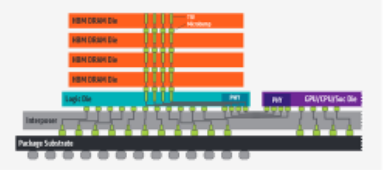
Standard	Physical Layer	Topology	Unidirectional Bandwidth	Mechanicals	Coherence
PCIe 4.0	PCIe PHY	p2p switched	16Gb/s/lane up to x16	PCIe	No
CCIX 1.0	PCIe PHY	p2p switched	25Gb/s/lane up to x16	PCIe	Full cache coherence between processors and accelerators
Gen-Z 1.0	IEEE 802.3 PCIe PHY	p2p switched meshed	16/25/56Gb/s/lane up to x256	SFF-TA	Full cache coherence
OpenCAPI 3.0	IEEE 802.3	p2p	25Gb/s/lane up to x8	In definition	Coherent access to system memory
PCIe 5.0	PCIe PHY	p2p switched	32Gb/s/lane up to x16	SFF-TA	No



Packaging technology

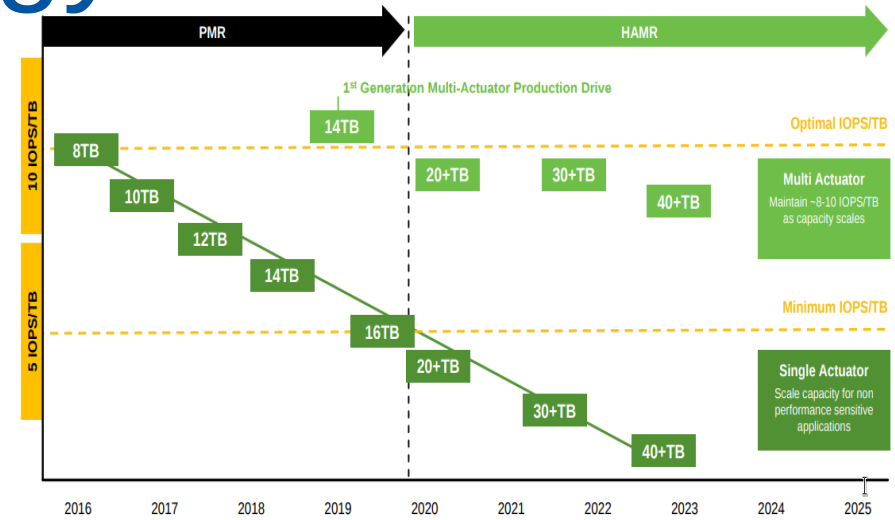
- Traditionally a silicon die is individually packaged, but more and more CPUs package together more (sometimes different) dies
- Classified according to how dies are arranged and connected
  - 2D packaging (e.g. AMD EPYC): multiple dies on a substrate
  - 2.5D packaging (e.g. Intel Kaby Lake-G, CPU+GPU): interposer between die and substrate for higher speed
  - Intel Foveros, a 2.5D with an interposer with active logic (Intel "Lake Field" hybrid CPU)
  - 3D packaging (e.g. stacked DRAM in HBM), for lower power, higher bandwidth and smaller footprint
- Can alleviate scaling issues with monolithic CPU dies but at a cost, both financial and in power and latency





HDDs: Technology

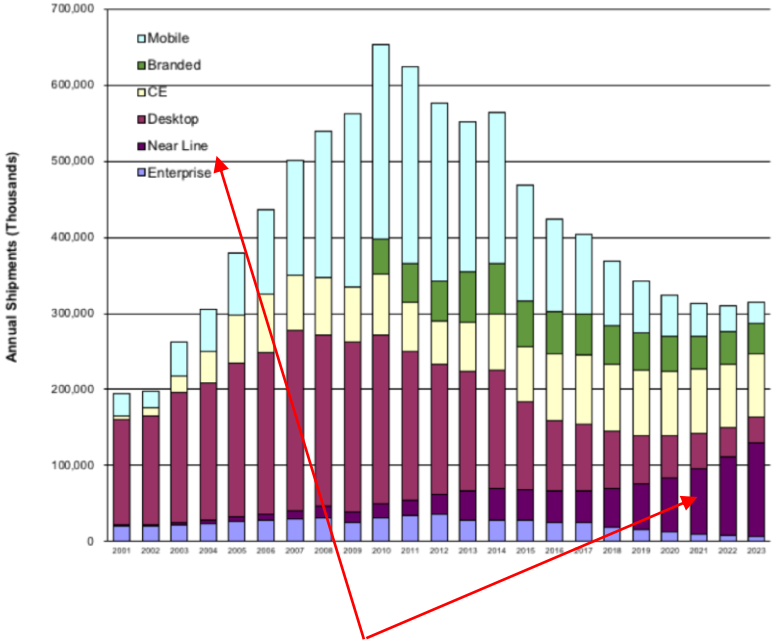
- Problems with existing HDD technology
  - Perpendicular magnetic recording at areal density limit
  - IOPS per TB continues to fall
  - Bandwidth not keeping up with drive capacity



- New technology expected soon
  - Energy Assisted Magnetic Recording (MAMR & HAMR) should allow 40TB drives by 2025
    - 2019 - 16TB MAMR drives from WD and HAMR drives from Seagate
  - Dual Actuator drives double disk IOPS and drive bandwidth
    - 2019 - 14TB Dual actuator drives from Seagate

HDDs: Market

- Total HDD Market is shrinking
- Sole growth market is in near-line (capacity) HDD used by Cloud and Big Science
- Shrinking market introduces risks
  - Higher costs -> Reduced economies of scale
  - Production risk -> Fewer factories
  - Technology risk -> Insufficient revenue to finance continued R&D
- Narrowing HDD/SSD price gap causing more HDD to be replaced by SSD (eg HSM buffers)



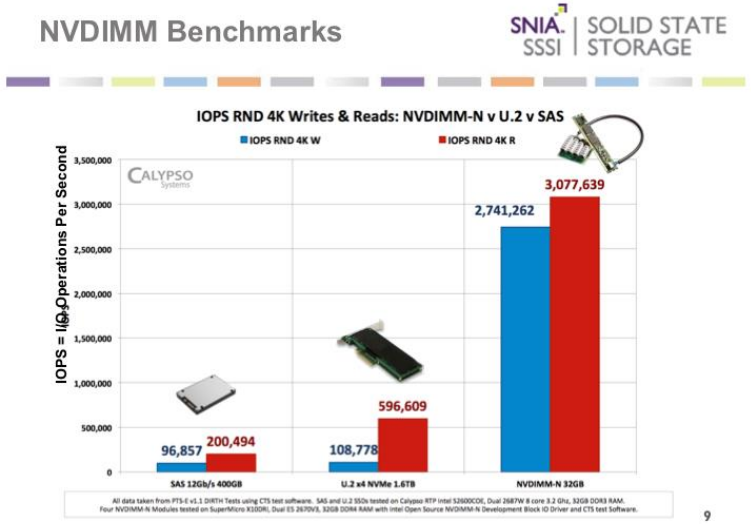
HDD type used for Cloud + HEP

Solid-State Storage: Current State

- 3D NAND Flash
  - Continues to be the underlying non-volatile memory technology of choice
  - Capacity increasing through additional layers (96 layers now, roadmaps out to a few hundred layers, >200 cost incr.)
  - Additional capacity also obtained by increasing bits per memory cell
  - SLC (1 bits) -> MLC (2 bits) -> TLC (3 bits) -> QLC (4 bits) > ???
  - Challenge - endurance and retention - increased ECC overhead

- Solid State Disks (SSD)
  - SAS/SATA software stack and hardware severely limits IOPS and rates (i.e. incl. Linux SCSI driver - sd)
  - NVMe + PCI-e new interconnect and protocol replacing SATA/SAS, significantly alleviating I/O bottleneck
  - NVDIMM (persistent memory) technology effectively eliminates all I/O bottlenecks (connected to memory bus)
  - New form factors EDSFF (“ruler”) and U.2 in addition to existing 2.5” HDD, M.2, and PCI-e card form factors

- Flash Systems
  - All Flash Arrays (AFA) common (analogous to HDD HW RAID arrays, but all flash instead of disk)
  - Mostly SAS attached, although other interconnects also available. NVMe-oF will come to replace SAS



Solid-State Storage: Evolution

- NVMe and NVMe-oF
  - Suppress multiple software layers in storage OS
  - Extend to other interconnects
  - Expand to enclosure and drive management, …
- NVDIMM
  - Non-volatile, on CPU memory bus
  - Requires CPU and OS support
  - Expect higher density, lower cost than DRAM; latency only factor 3 higher
  - Candidate technologies: 3D NAND, 3D XPoint

Solid-State Storage: Market

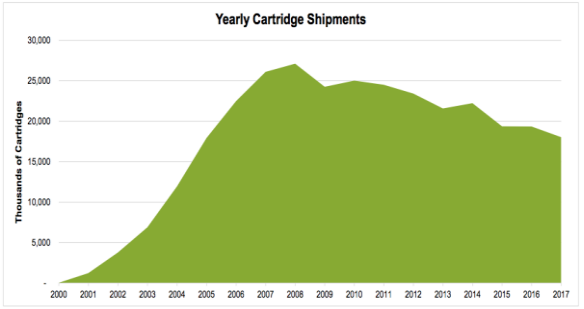
- 3D NAND dominant
- 3D XPoint: slow start, large investments, IP issues
- 2018: 40% more bits shipped, but revenues declining
- 2019: NAND flash market will remain in oversupply

- NVMe/NVMe-oF
  - NVMe SSDs: Expected to grow from ~$2B in 2017 to $9 billion in 2022
  - NVMe expected to become dominant over SSD in 2021/2022
  - NVMe-oF adapters: 1.75 million units sold by 2022, with the bulk of the adapters being smart (NVMe-oF offload)

Tape: Technology

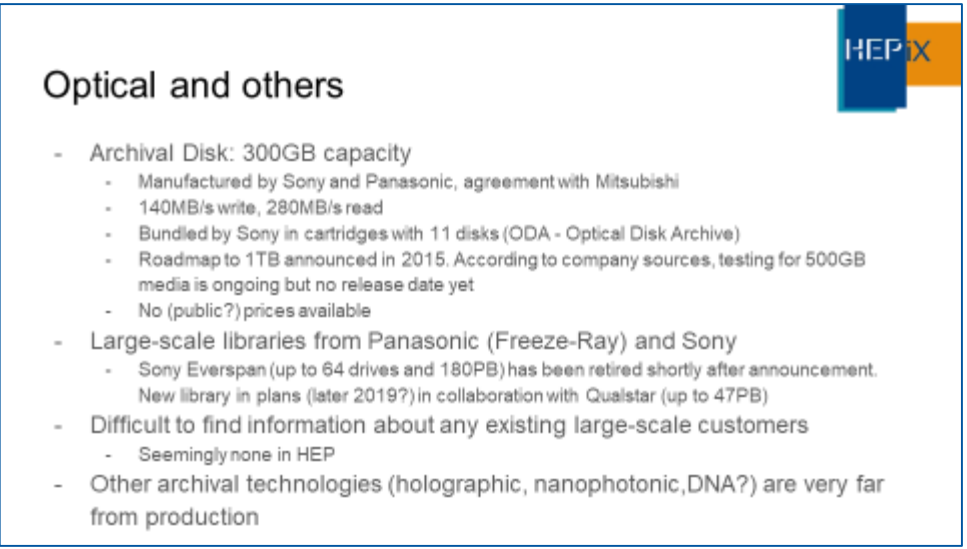
- Enterprise: TS1160 released in Q4 2018
  - 400MB/s (+11% over previous gen), 20TB on new JE media (+33% over previous gen)
- LTO: Gen-8 released in Q4 2017
  - 360MB/s (+20% over LTO-7), 12TB (+100% over previous gen).
  - Can use LTO-7 media @ 9TB instead of 6TB, but LTO-9 support for that format unclear
  - LTO-9 expected by EOY 2020. Roadmap claims 24TB but more realistically ~18-20TB
- Libraries (>10K slots): IBM, Oracle, Spectralogic and Quantum
  - Concerns about Quantum and Oracle future strategy (and support model/pricing for the latter)
- R&D
  - 123Gb/sqin on BaFe (~220TB tape) and 201Gb/sqin (330TB, CoPtCr media, but material/manufacturing costs expected to be high). Allows for 8-9 years of headroom at 30% density CAGR
  - Streaming read/write performance predicted to be 15-20% CAGR. But no significant improvements expected on seek times (random access) as limited by mechanics

Tape: Market



- Drives
  - Left only with LTO and IBM Enterprise after Oracle's retirement
  - LTO drive technology, R&D and manufacturing seemingly dominated by IBM. Will the two product lines be merged eventually?
- Media
  - Fujifilm entangled in ongoing patent war with Sony wrt LTO-8 media. LTO-7 media available at interesting prices (~5-6CHF/TB), not the case for LTO-8
- Market continues a ~10-year sustained contraction
- Revenues estimated 0.7B USD
- Significant risk factor for Big Science: limited competition, contracting market

Optical Storage and Others

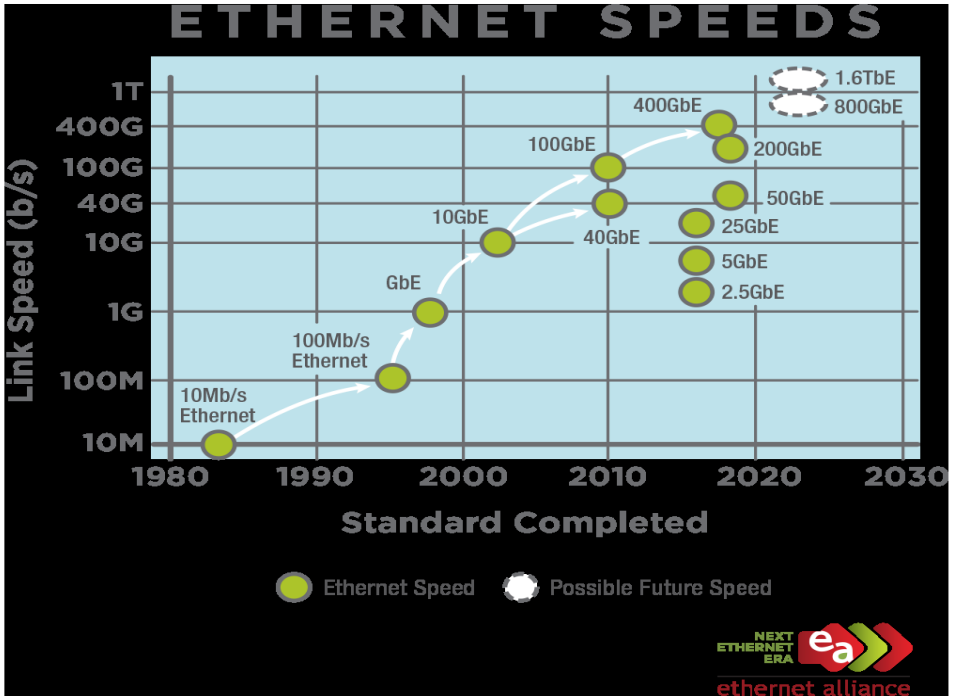


- More like technology studies rather than production systems with significant market penetration
- Others even more so
- No role in Big Science

Network: Introduction

- Expect > 25% CAGR of Internet traffic until 2022
  - Main driver is media streaming over 5G
- Evolution okay until now
  - But free lunch is over

Network: Ethernet

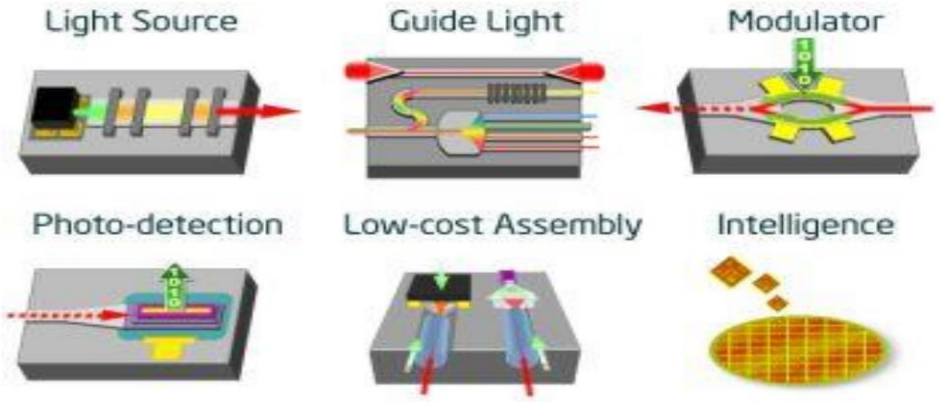


- 2019
  - Support for links at 50, 200, 400 Gbps over single-mode fibre up to 10 km, to extend to 40 km later
  - Bi-directional optical PHY for point-to-point at 10, 25, 50 Gbps
  - 100 and 400 Gbps over DWDM up to 80 km
- 100Gbps becoming commodity, but too many variants; 400 Gbps is out
- SC18 demonstrated maturity of 100Gbps servers

Forward Error Correction

- Inflates payload by N bits to protects M bits
  - E.g. 300 / 5140 [RS(544,514,10)] or less for other standards
- Can use lower quality optical specs to reduce cost significantly
- Requires less retransmits as long as all errors can be corrected → overall higher bandwidth
- Can show ‘cliff’ in bandwidth if corrections fail due to too many transmission errors
- Little additional latency: O(100ns)
- Algorithms tuned to transmission parameters

Silicon Photonics



- Almost all parts on same die
  - Cheaper to produce than traditional transceivers
  - Allows also for more integration testing
  - Manufacturing in existing CMOS production processes
  - Lower power consumption than conventional transceivers
  - Available since ~2 years from Cisco(Luxtera), Intel, GlobalFoundries, IBM (research only?), …
- See also e.g. https://community.mellanox.com/s/article/inside-the-silicon-photonics-transceiver

Areas of Network Evolution

- Improving efficiency of data transfers
  - TCP BBR - version 2 is in the works with promising improvements
  - Exploring alternative protocols for transfers (UDP)
- Caching
  - Data caches co-located with network hubs in a similar way as on commercial CDNs
- Federations/Clouds
  - Overlay networks spanning multiple domains
  - Multi-clouds - expanding DC networking via L3VPNs
- Technology
  - SDN/NFV approaches - currently looked at by HEPiX NFV WG
    - Compute - Agile service delivery on Cloud Infrastructures (OpenStack, Kubernetes)
    - Data Transfers - Network resource optimisation - dynamically optimising the network based on its load and state (more in Shawn/Ilija)
  - SD-WAN approaches - https://www.mode.net/

Network: Other things to look out

- What to expect from Mellanox/NVidia deal?
  - Nvidia paid $6.9b for Mellanox (~$1b revenue)
  - 3 CPU/GPU(FPGA) interconnects now:
    - NVlink / OpenCAPI (Coherent Accelerator Processor Interface)
      AMD, IBM, Google, Micron, Mellanox (Nvidia as ‘contributor’)
    - CCIX (Cache Coherent Interconnect for Accelerators)
      AMD, ARM, IBM, Qualcomm, Xilinx, Huawei, Mellanox
    - CXL (Compute Express Link)
      Cisco, Dell EMC, Facebook, Google, HPE, Huawei, Intel, Microsoft
- Remote DMA (RDMA) with all its flavours
  - RoCE RDMA over Converged Ethernet, iWARP, ...

Conclusions (1)

- Technology progress per se remains good, but obstacles ahead
- Key computing markets in the hand of very few companies
  - Intel still dominating server market
- Price/performance advances are slowing down
- New processors and architectures are mainly focused on Machine Learning, highly profitable markets expected – not necessarily very relevant for Big Science
- HDD still key storage for the foreseeable future, SSDs not cost effective, but concerns over markets
- Have to closely watch the tape development – big concerns

Conclusions (2)

- Meta-conclusions:
  - Some years ago: Concerns that evolution may stop due to physics or technology
  - Now: Clear that business, market, financial arguments are the real limit
    - Markets largely saturated
- Meta-meta-conclusions:
  - Tech watch is becoming essential for the Big Science communities
  - Complex and demanding
  - Interesting and rewarding

Final Remarks

- Thanks to all WG members and the conveners for their dedication and contributions so far
- Interested? Never too late to join – contact conveners or HM
- Ideas for opportunities (in particular presentations)? Contact conveners or HM

References

- Presentations:
  - HOW2019 workshop: https://indico.cern.ch/event/759388/sessions/295048/#20190319
  - HEPiX spring 2019 workshop: https://indico.cern.ch/event/765497/sessions/303981/#20190328
- Techwatch WG internal Web site: http://w3.hepix.org/techwatch/
- Subgroup documents:
  - CPUs, GPUs and accelerators: https://docs.google.com/document/d/1dUcuuSubLO4WeFzF0mlsA8mPgHV-46z6Rga7wCt0kFk/edit?usp=sharing
  - Memory: https://docs.google.com/document/d/1a8K_BA8ipy5l0NvcNjrOLqDsN_RjJXDnX4LHRrfrNE0/edit?usp=sharing
  - Storage: https://docs.google.com/document/d/1IS4_raw7PE0wVTNWDJmUGmneV1zpg9-29vz7XP4ChRA/edit
  - Networking: https://twiki.cern.ch/twiki/bin/view/HEPIX/TechwatchNetwork/WebHome