CPUs, GPUs, accelerators and memory

Andrea Sciabà

On behalf of the Technology Watch WG

HOW Workshop
18-22 March 2019
Jefferson Lab, Newport News

Introduction

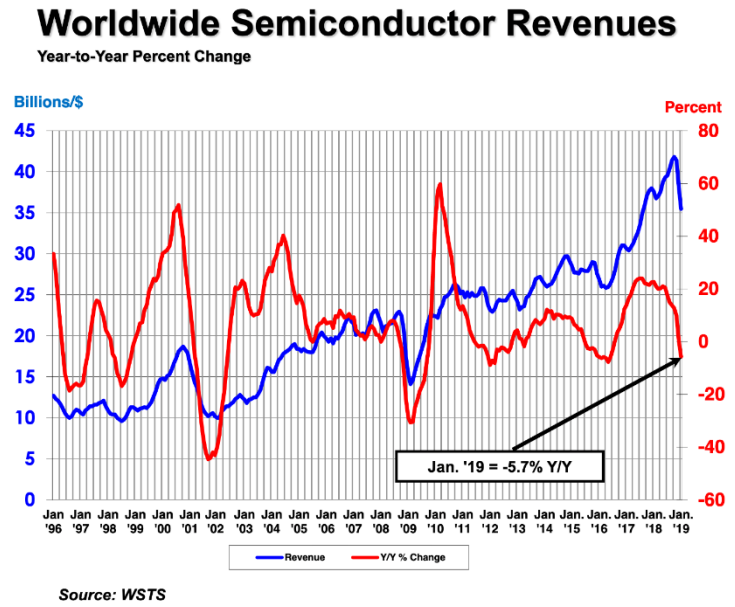
- The goal of the presentation is to give a broad overview of the status and prospects of compute technologies
  - Intentionally, with a HEP computing bias
- Focus on processors and accelerators and volatile memory
- The wider purpose of the working group is to provide information that can be used to optimize investments
  - Market trends, price evolution
- More detailed information is already available in a document
  - Soon to be added to the WG website

Outline

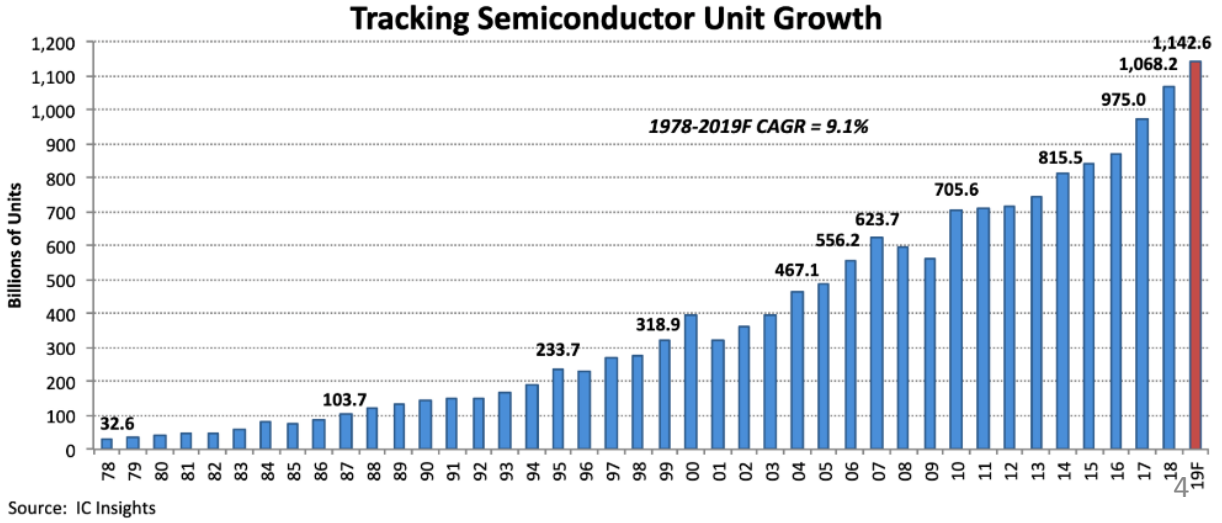
- General market trends
- CPUs
  - Intel, AMD
  - ARM
  - Other architectures
- GPUs
- FPGAs
- Supporting technologies
- Memory technologies

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

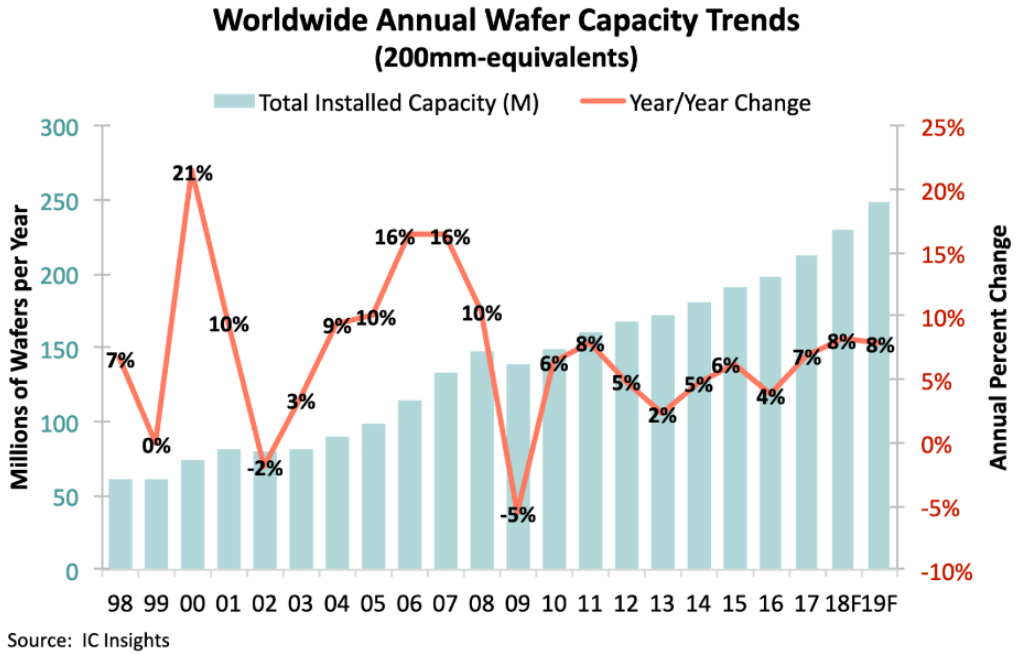


Source: WSTS

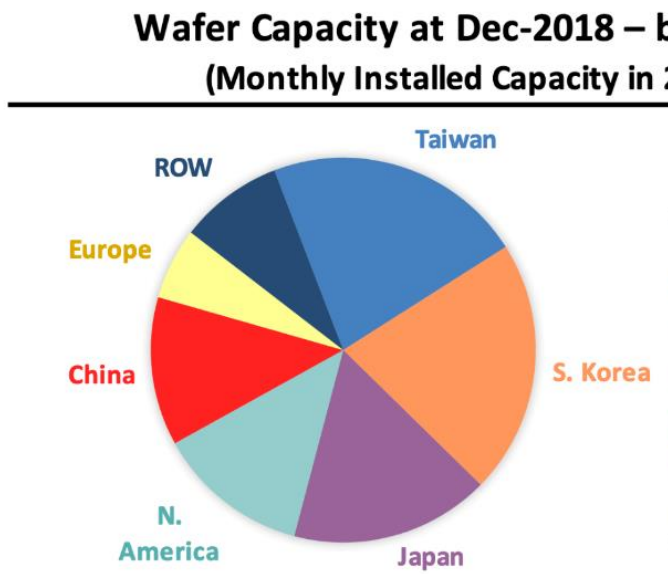


Source: IC Insights

Semiconductor fabrication



Source: IC Insights

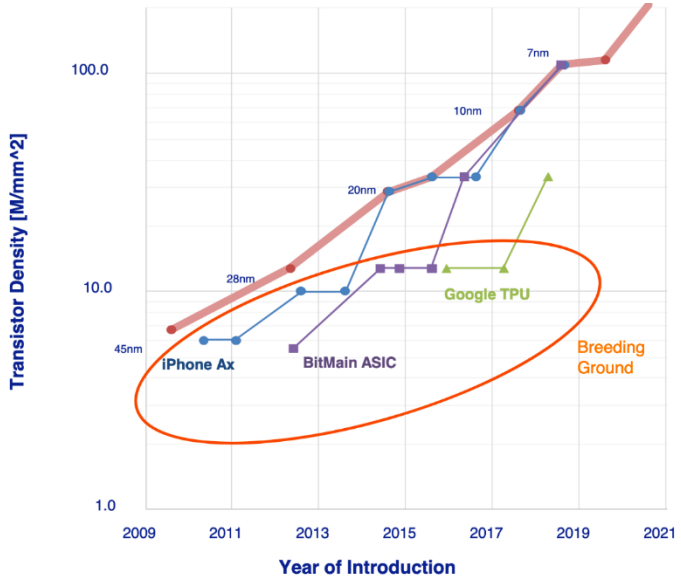


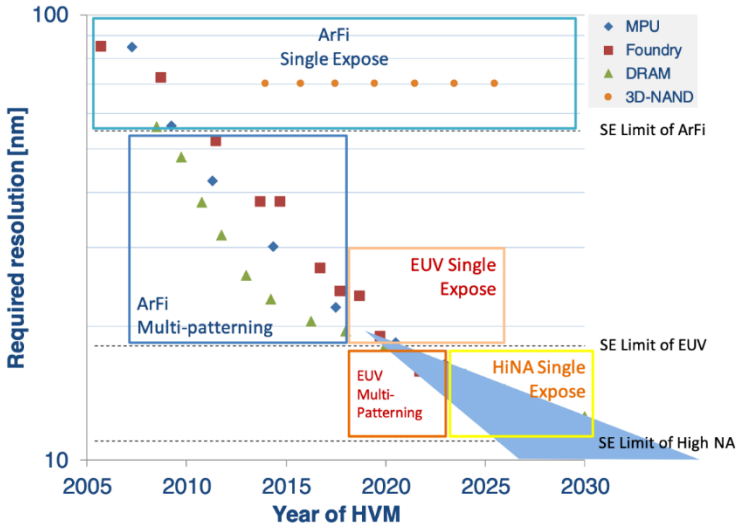
Region	Installed Capacity (K w/m)	% of Worldwide Total
Taiwan	4,126	21.8%
S. Korea	4,033	21.3%
Japan	3,168	16.8%
N. America	2,426	12.8%
China	2,361	12.5%
Europe	1,138	6.0%
ROW	1,646	8.7%
TOTAL	18,897	100%

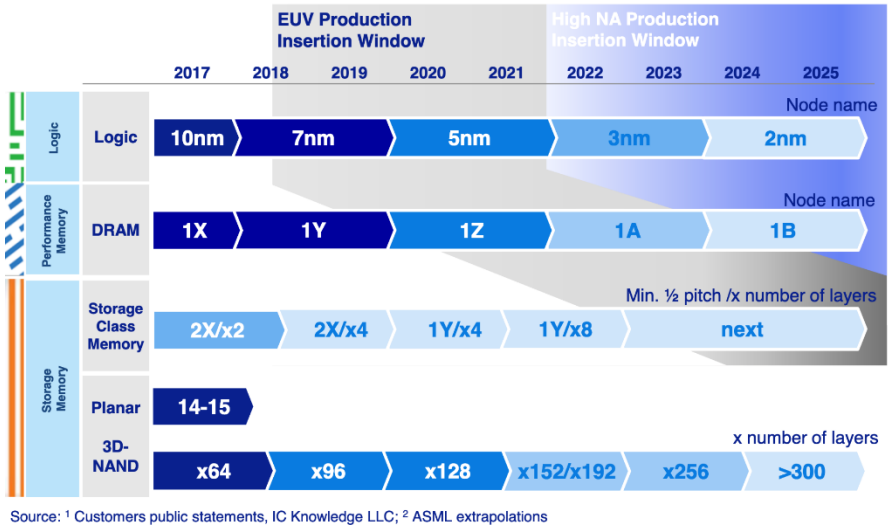
Source: IC Insights

- Taiwan leading all regions/countries in wafer capacity
- TSMC held 67% of Taiwan's capacity and is leading
- Samsung and SK Hynix represent 94% of the installed IC wafer capacity in South Korea
  - They are likely to influence memory prices (which are now very high)
- New manufacturing lines are expected to boost industry capacity by 8% in both 2018 and 2019

Process technology



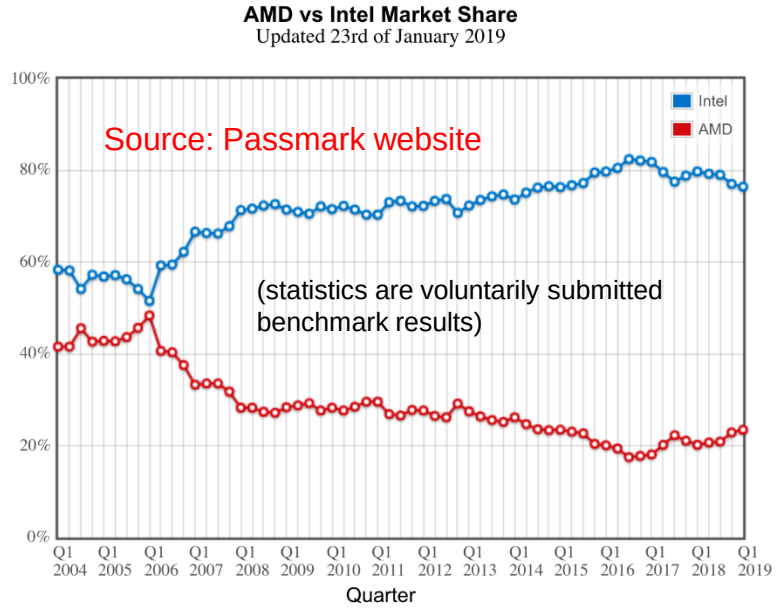


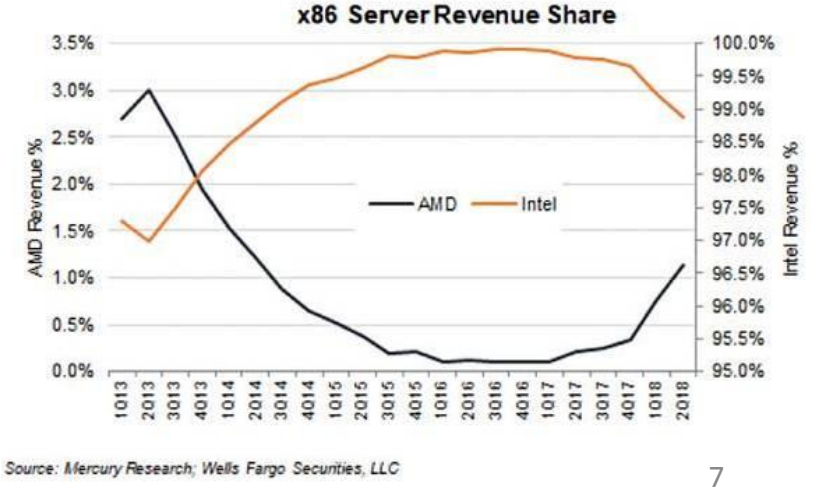


- Performance scaling in process technology continues to grow Moore's law prediction
- Embedded processors benefit the most from process manufacturing improvements
- EUV is forecast to be the dominant lithography technology in the coming years
  - Already used for 7nm by TSMC for AMD, Apple, Nvidia and Qualcomm

Intel and AMD market share

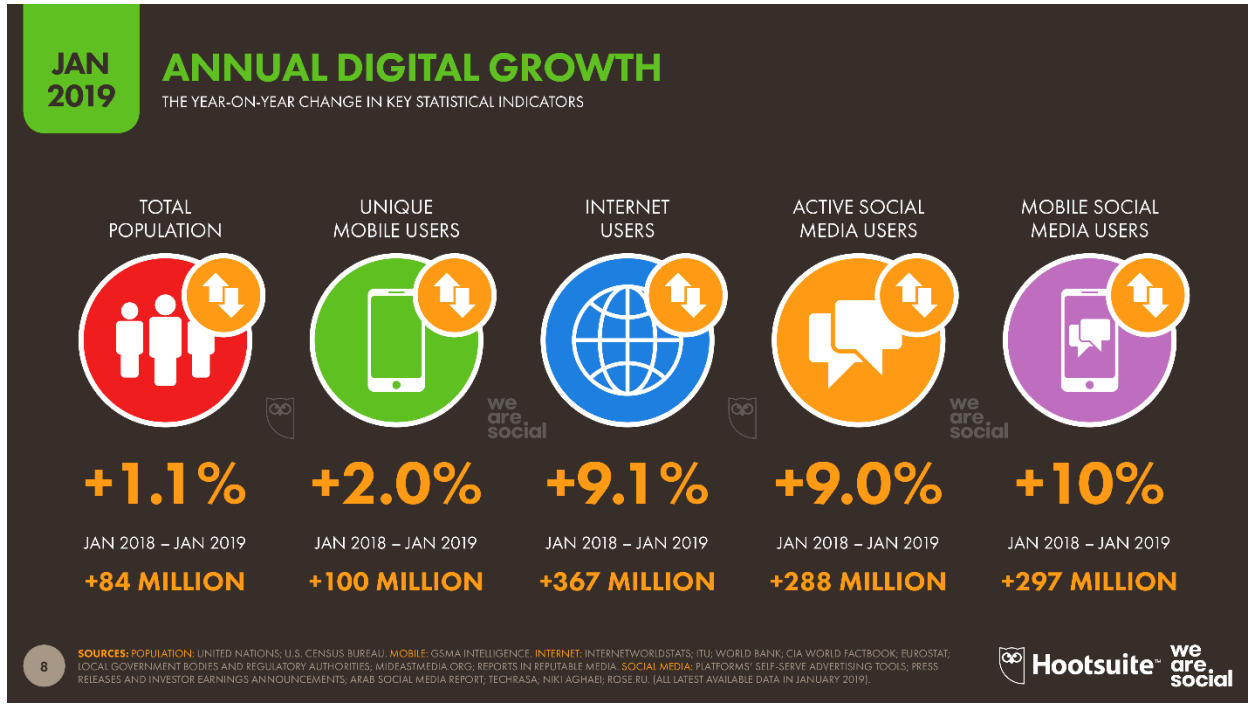
- AMD server market share is rapidly increasing since 2017, but from almost nothing
  - Zen architecture released in 2017
  - Achieved 5% of server shipment market share on Q4 2018, projected to 10% in one year
- AMD always had a reasonable (20-30%) share overall
- EPYC revenue was $58m in the second 2018 quarter vs $36m in the prior quarter

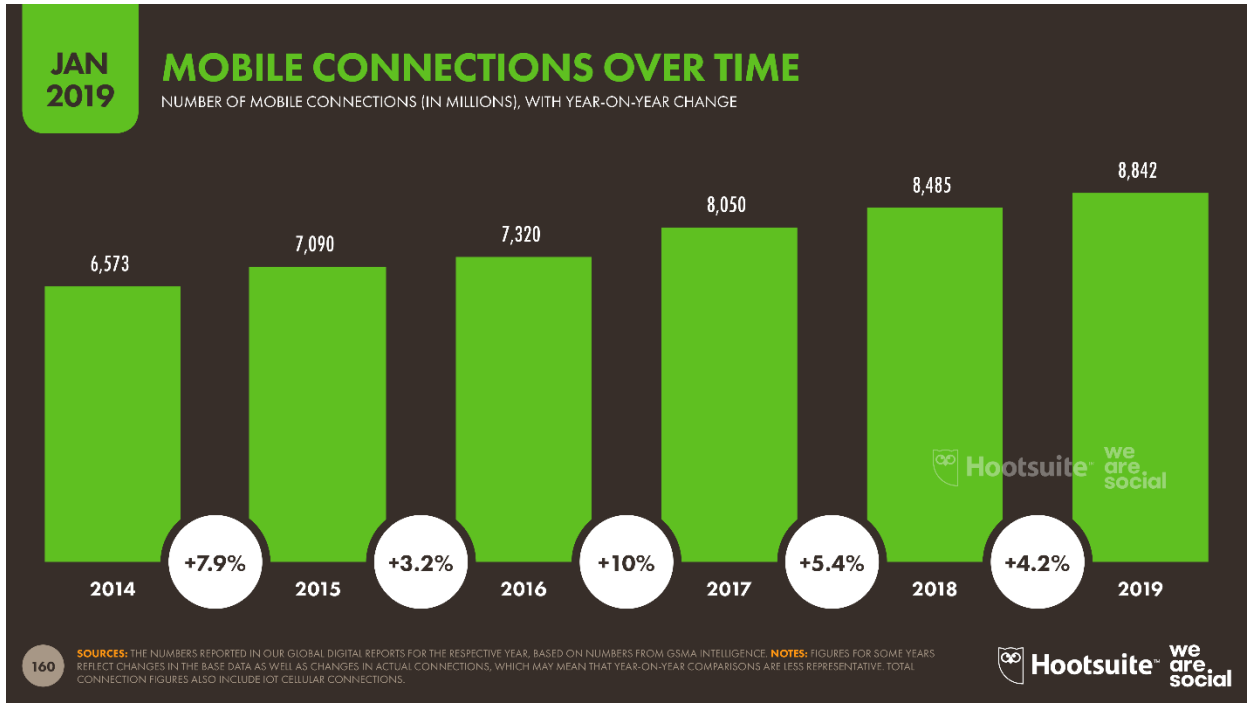




Source: Mercury Research; Wells Fargo Securities, LLC

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

CPUS AND ACCELERATORS

Intel server CPU line-up

- 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

Current and future Intel server architectures

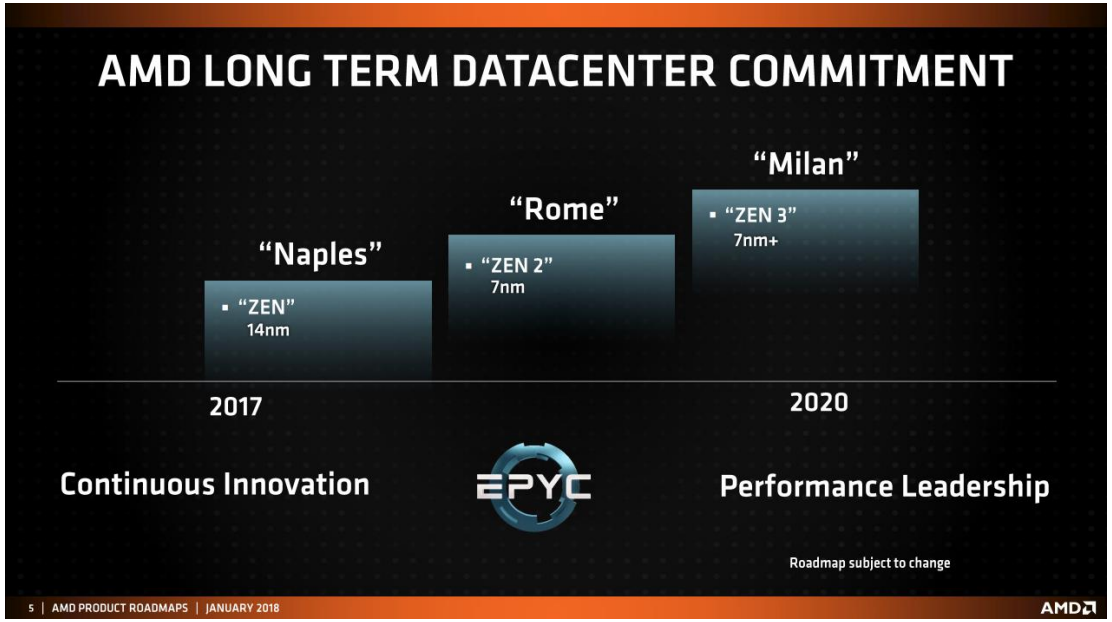
Microarchitecture	Technology	Launch year	Highlights
Skylake-SP	14nm	2017	Improved frontend and execution units More load/store bandwidth Improved hyperthreading AVX-512
Cascade Lake	14nm++	2019	Vector Neural Network Instructions (VNNI) to improve inference performance Support 3D XPoint-based memory modules 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

Other Intel x86 architectures

- Xeon Phi
  - Features 4-way hyperthreading and AVX-512 support
  - Elicited a lot of interest in the HEP community and for deep learning applications
  - Announced to be discontinued in summer 2018
- Networking processors (Xeon D)
  - SoC design
  - Used to accelerate networking functionality or to process encrypted data streams
  - Two families, D-500 for networking and D-100 for higher performance, based on Skylake-SP with on-package chipset
  - Hewitt Lake just announced, probably based on Cascade Lake
- Hybrid CPUs
  - Will be enabled by Foveros, the 3D chip stacking technology recently demonstrated

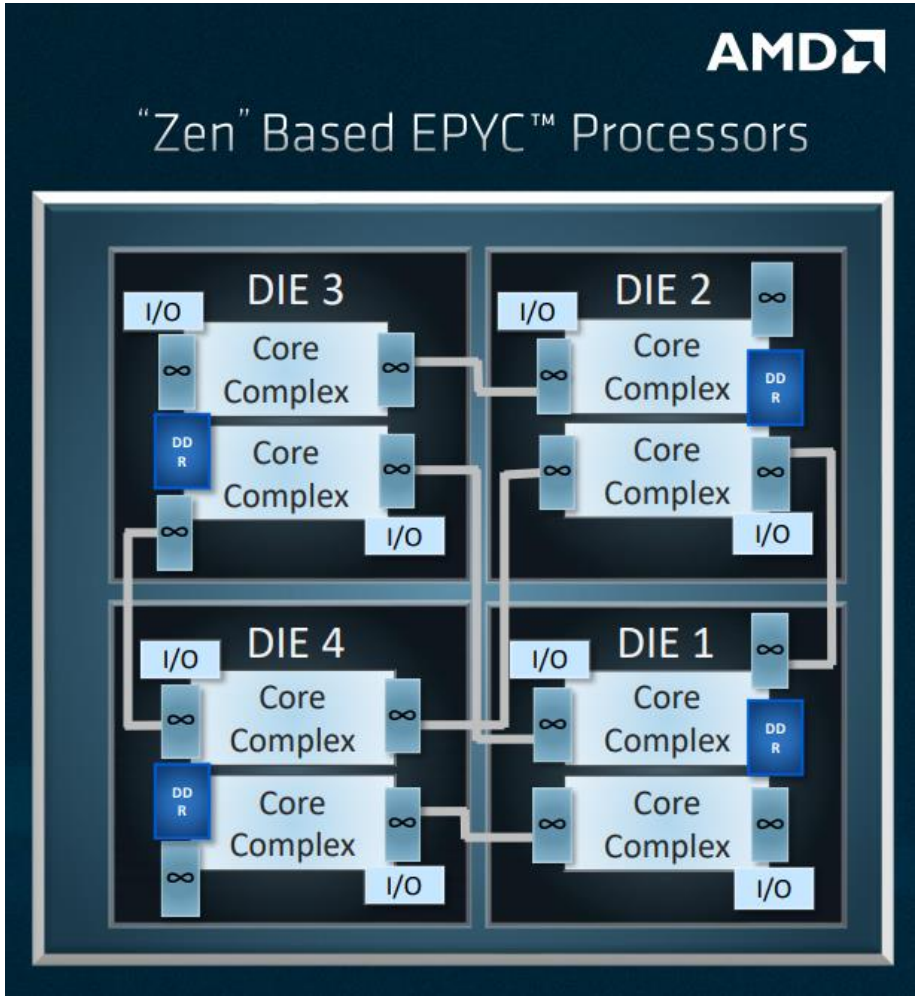
AMD server CPU line-up

- 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



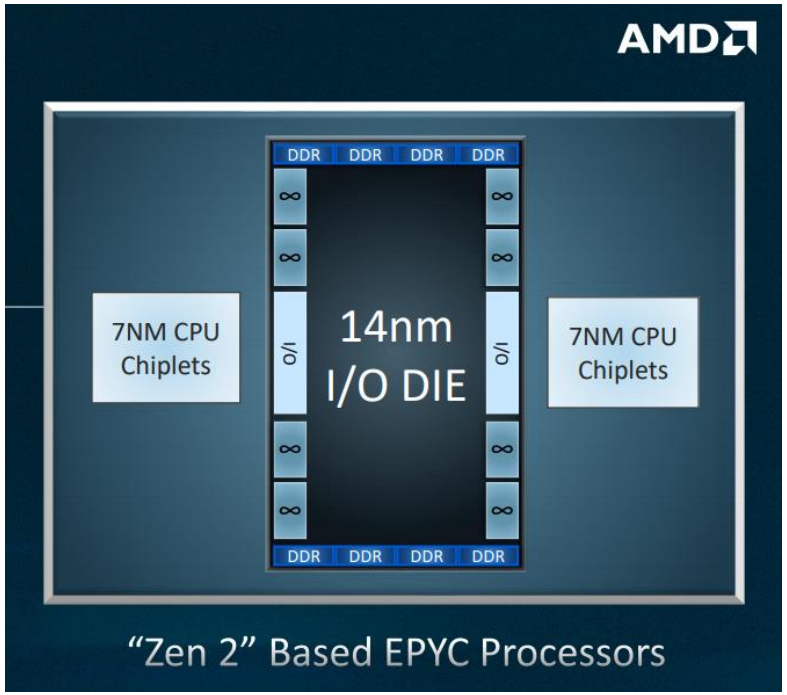
EPYC Naples

- 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



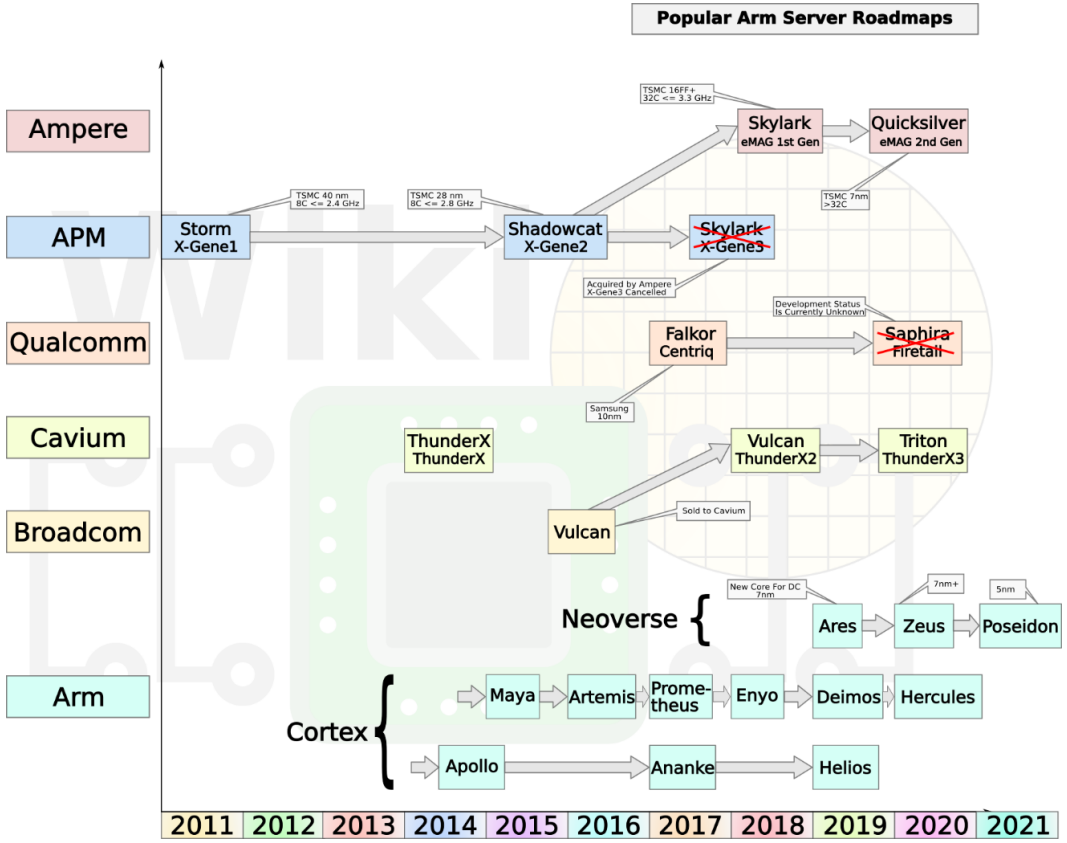
Recent experiences in WLCG

1. LHCb
   - Using some nodes with EPYC 7301 CPUs (16 cores)
   - Performance of LHCb trigger application almost equal to Xeon Silver 4114 (10 cores)
   - Need to populate all 8 DIMM slots for maximum performance
   - Testing it as potential hypervisor platform
   - Will competitively tender with Intel next year
2. NIKHEF
   - Have 93 single-socket 32 core EPYC 7551P nodes in production
   - A single EPYC 7371 node (single socket, 16 cores), available for tests
3. INFN
   - All WLCG sites have installed in 2018 a number of systems (40 in total) with EPYC 7351 (16 cores) in Twin Square configuration
   - Experience very positive
4. BNL
   - Extensive tests with several EPYC CPUs presented at HEPiX Fall 2018
   - Measured performance from mid/upper range EPYC similar to mid/upper range Xeon Gold

- Caltech
  - Two servers with EPYC 7551P (32 cores), soon available for benchmarking

ARM in the data center

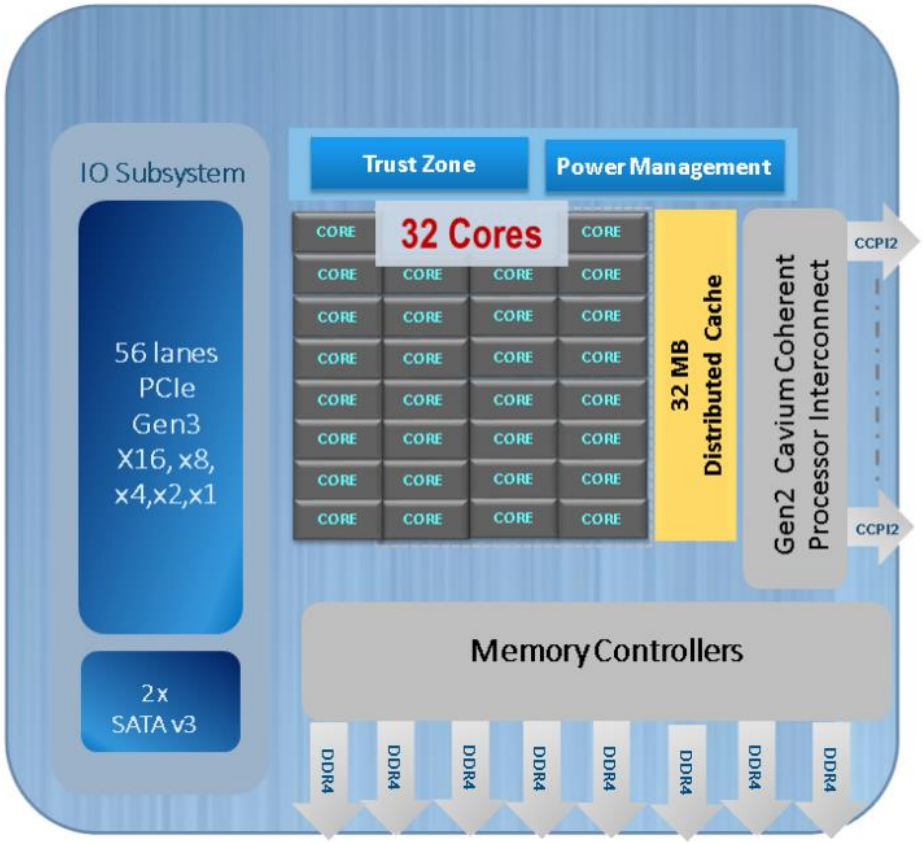
- ARM is ubiquitous in the mobile and embedded CPU word
- Data center implementations have been relatively unsuccessful so far
  - Performance/power and performance/$ not competitive with Intel and AMD
- LHC experiments are capable of using ARM CPUs if needed
  - Some do nightly builds on ARM since years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



Marvell ThunderX2 and Fujitsu A64FX

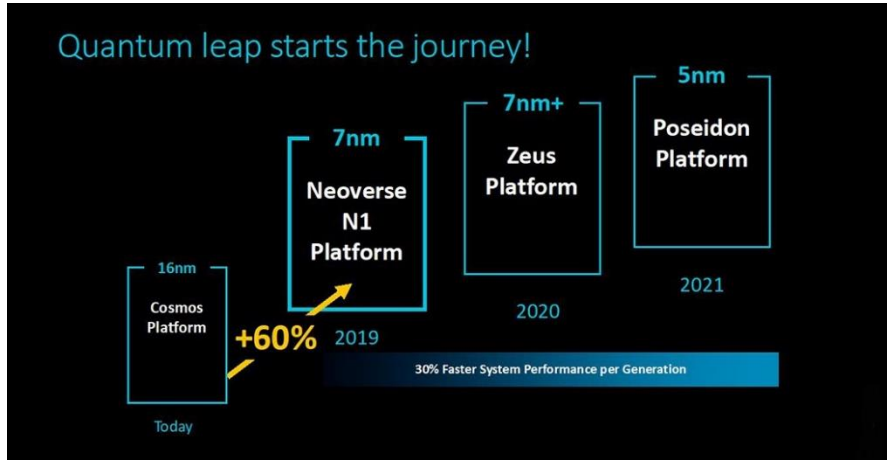
- ThunderX2 for mainstream cloud and HPC data centers, from 2018
  - Enjoys the greatest market visibility and reasonable performance/$
    - Used e.g. at CRAY XC-50 at Los Alamos and HPE Apollo 70 based Astra HPC system at Sandia National Laboratory
  - ARM V8.1 architecture
    - Up to 32 cores, 4-way SMT
    - Up to 8 DDR4 memory channels
    - Up to 56 PCIe Gen3 lanes

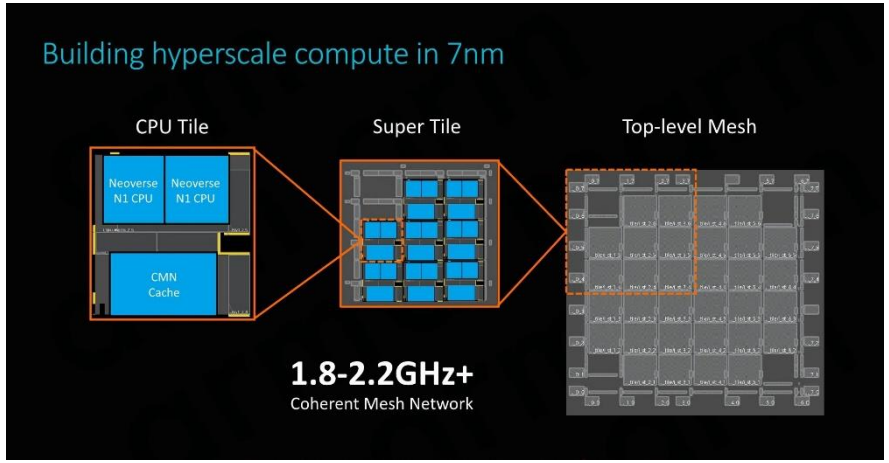
- Fujitsu A64FX to be used in supercomputer at RIKEN center
  - Based on the V8.2-A ISA architecture
    - First to deliver scalable vector extensions (SVE)
    - 48 cores
    - 32 GB of HBM2 high bandwidth memory
    - 7nm FinFET process
  - Interesting to see what performance will achieve as it may 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

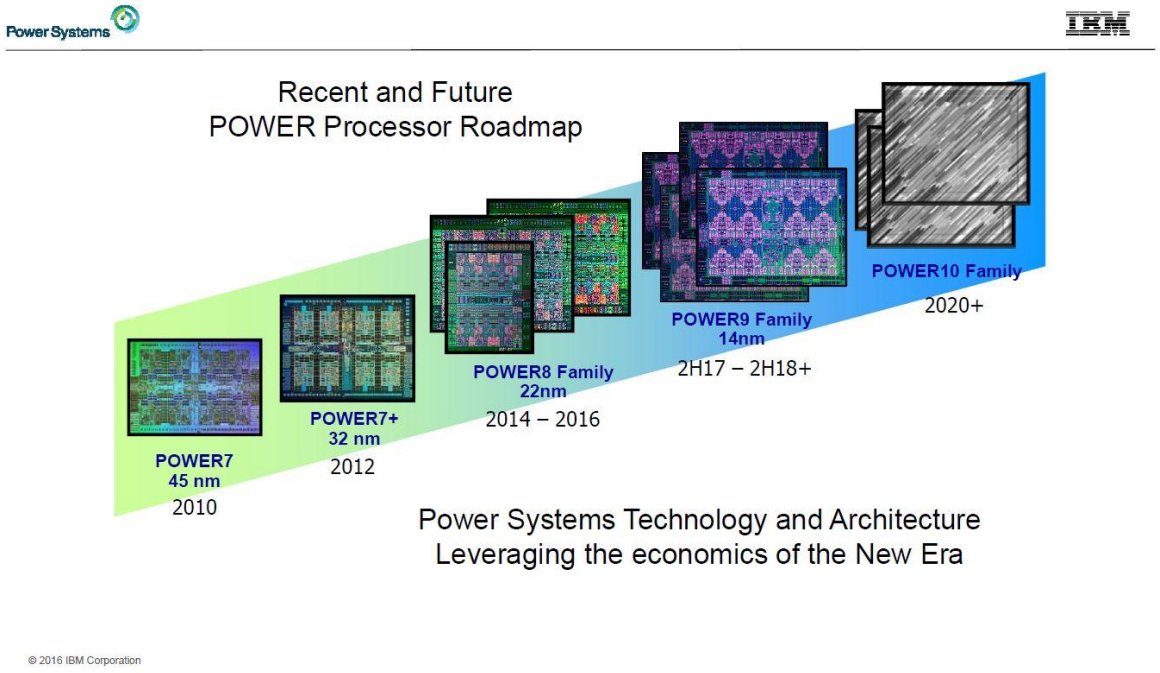




Source: Anandtech

IBM POWER

- POWER9
  - 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
  - 10nm process
  - Several feature enhancements
  - First to support PCIe Gen5

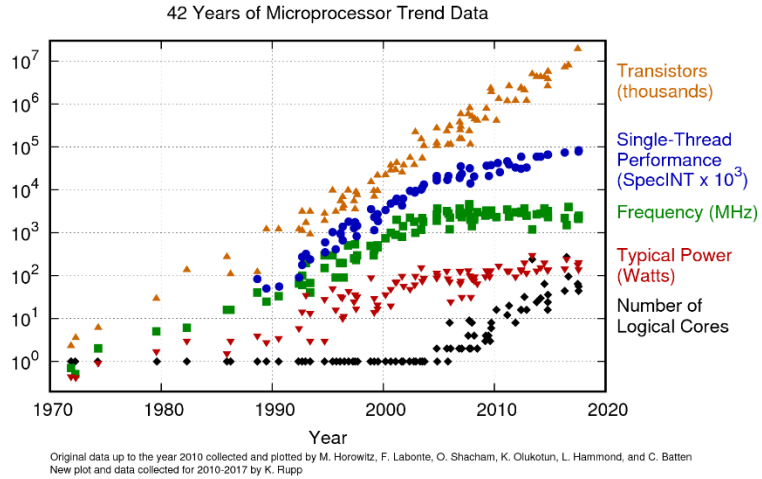


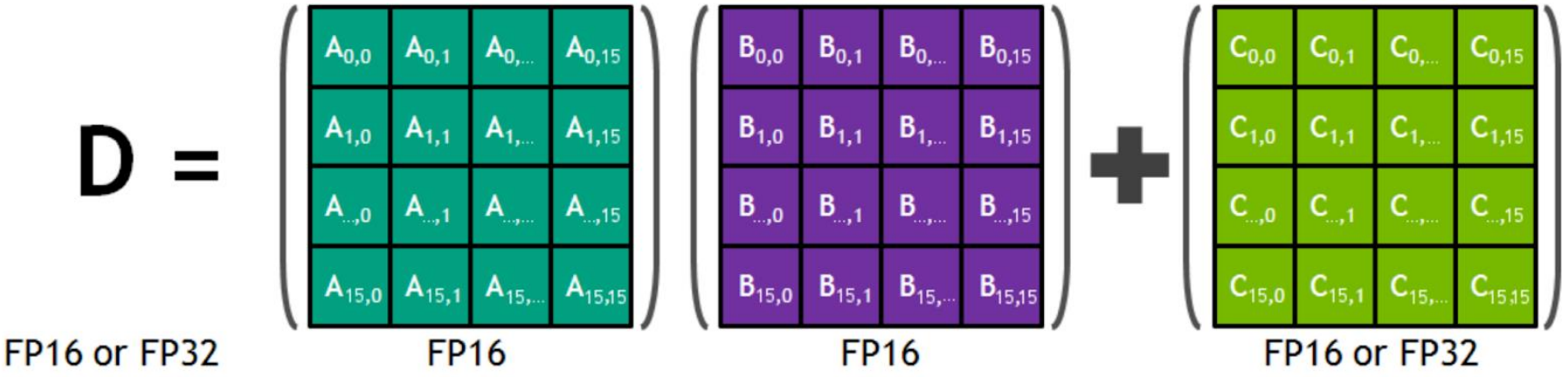
RISC-V and MIPS

- RISC-V is an open source ISA
  - To be used by some companies for controllers (Nvidia and WD), for FPGA (Microsemi), for fitness bands…
  - For the time being, not targeting the data center
  - Might compete with ARM in the mid term
  - Completely eclipsed MIPS
- MIPS
  - Considered dead

Discrete GPUs: current status

- GPU's 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/

Nvidia and AMD

- Volta addressing the server market, Turing the gaming market

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

- 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

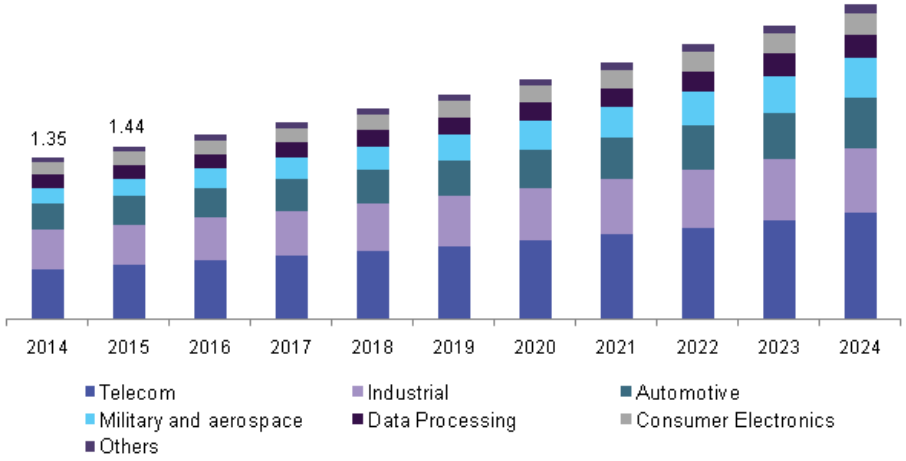
GPUs - Programmability

- 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 software frameworks

- 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

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#qom1512594527835__fn_soc_variab_avail_xlx

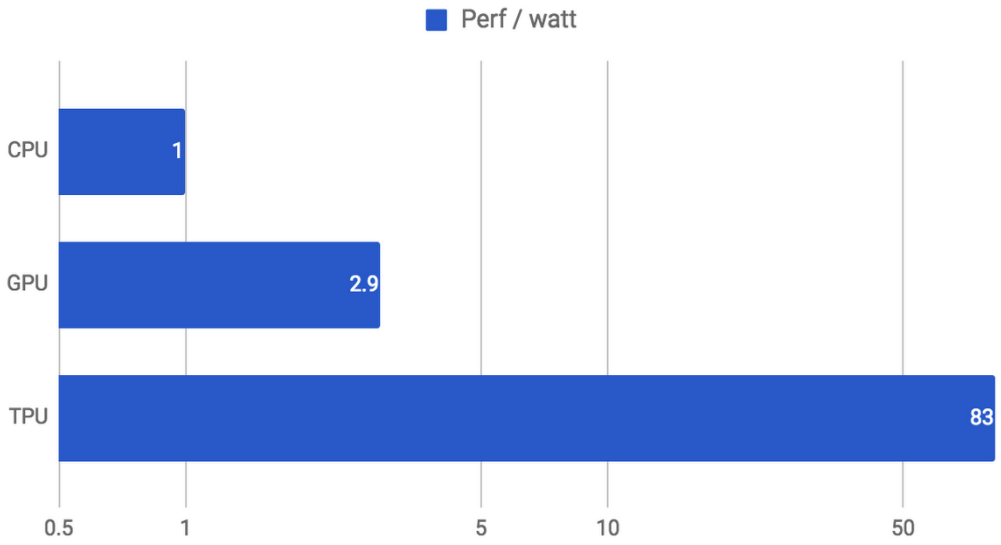
FPGA programming

- 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/

Other Machine Learning processors and accelerators

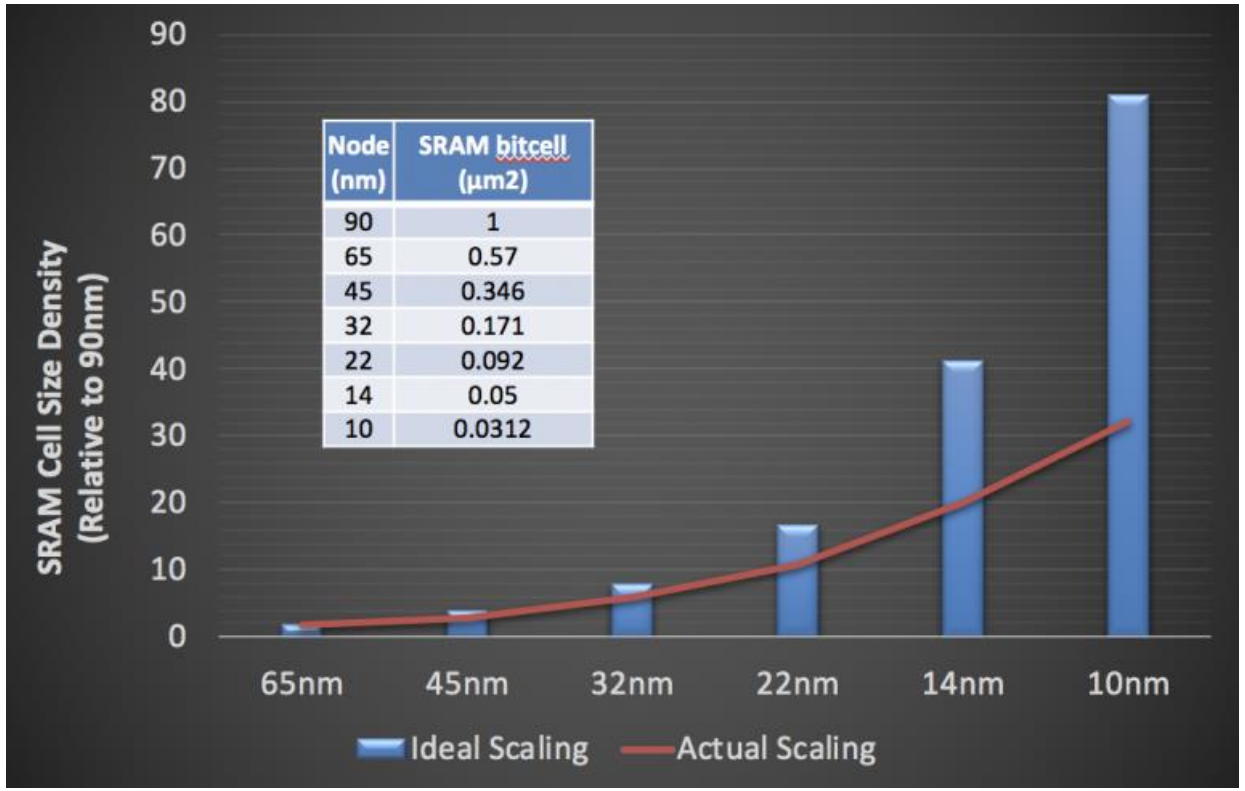
- Intel Nervana AI Processor NNP-L-1000 (H2 2019-)
  - Accelerates AI inference for companies with high workload demands
  - Optimized across memory, bandwidth, utilization and power
  - Spring Crest 3-4x faster training than Lake Crest, introduced in 2017
  - Supports bfloat16
- Google TPU
  - Huge increase in perf/watt for ML compared to CPUs and GPUs
- Intel Configurable Spatial Accelerator (CSA)
  - Dataflow engines that explicitly map the parallelism of the code onto an array of processing, storage and switching elements
  - Heavily customized for specific applications



MEMORY TECHNOLOGIES

Static RAM (SRAM)

- 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

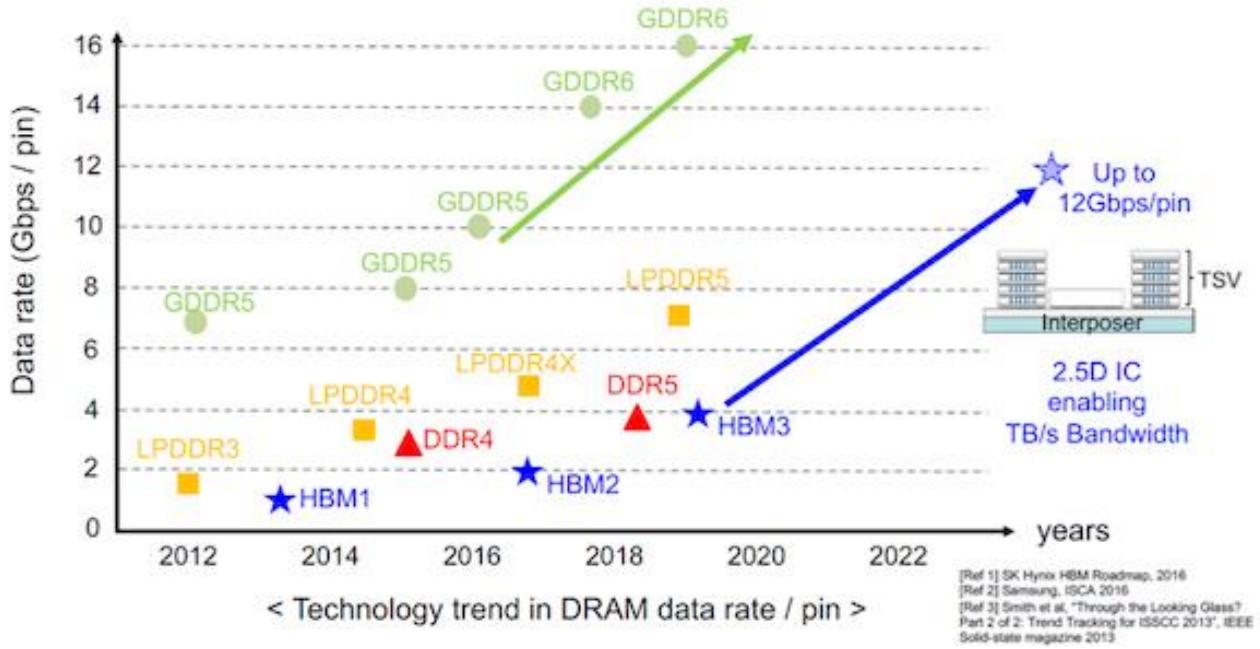


Node (nm)	SRAM bitcell (µm2)
90	1
65	0.57
45	0.346
32	0.171
22	0.092
14	0.05
10	0.0312

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

Dynamic RAM (DRAM)

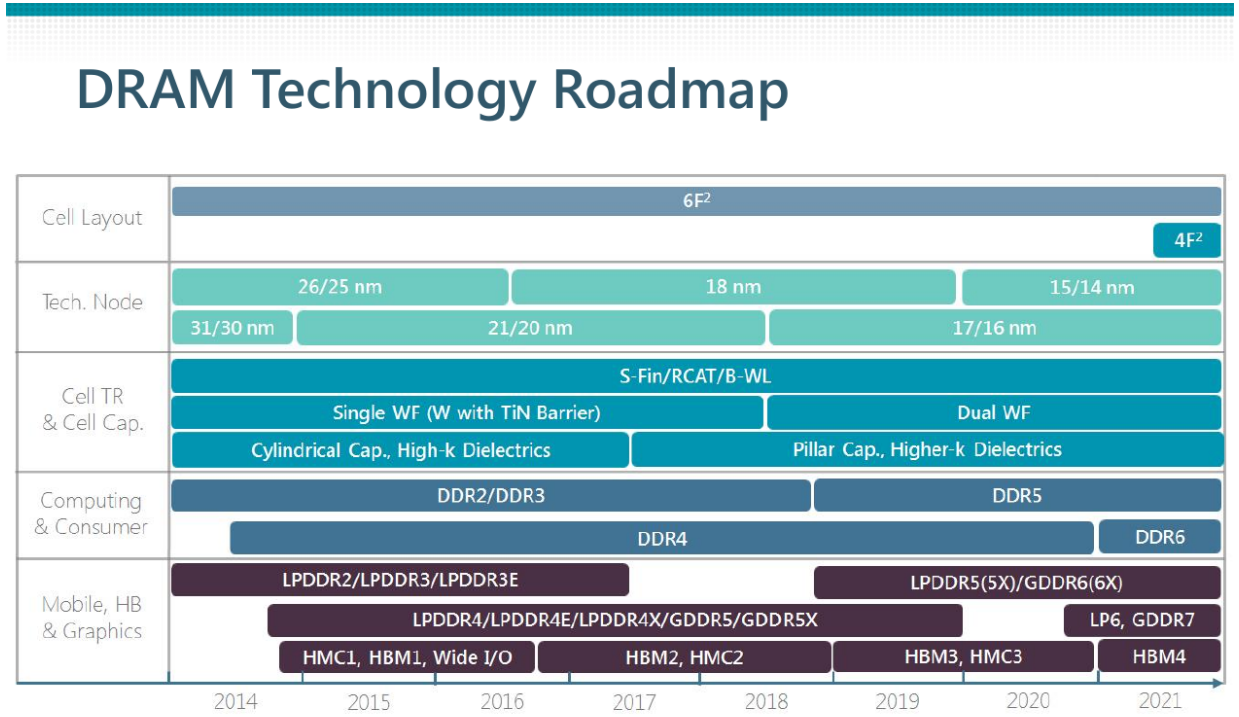
- Dominant standards continue to evolve
  - DDR4 -> DDR5
    - 3200MT/s -> 6400MT/s
    - 16Gb -> 32Gb chips
  - GDDR5 - > GDDR5X
    - 14 Gbps/pin -> 16Gbps/pin
    - 8Gb -> 16Gb chips
  - HBM -> HBM2
    - 1 Gbps/pin -> 2.4 Gbps/pin
    - 4 die stack -> 12 die stack
    - 2Gb die -> 8Gb die
- Note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

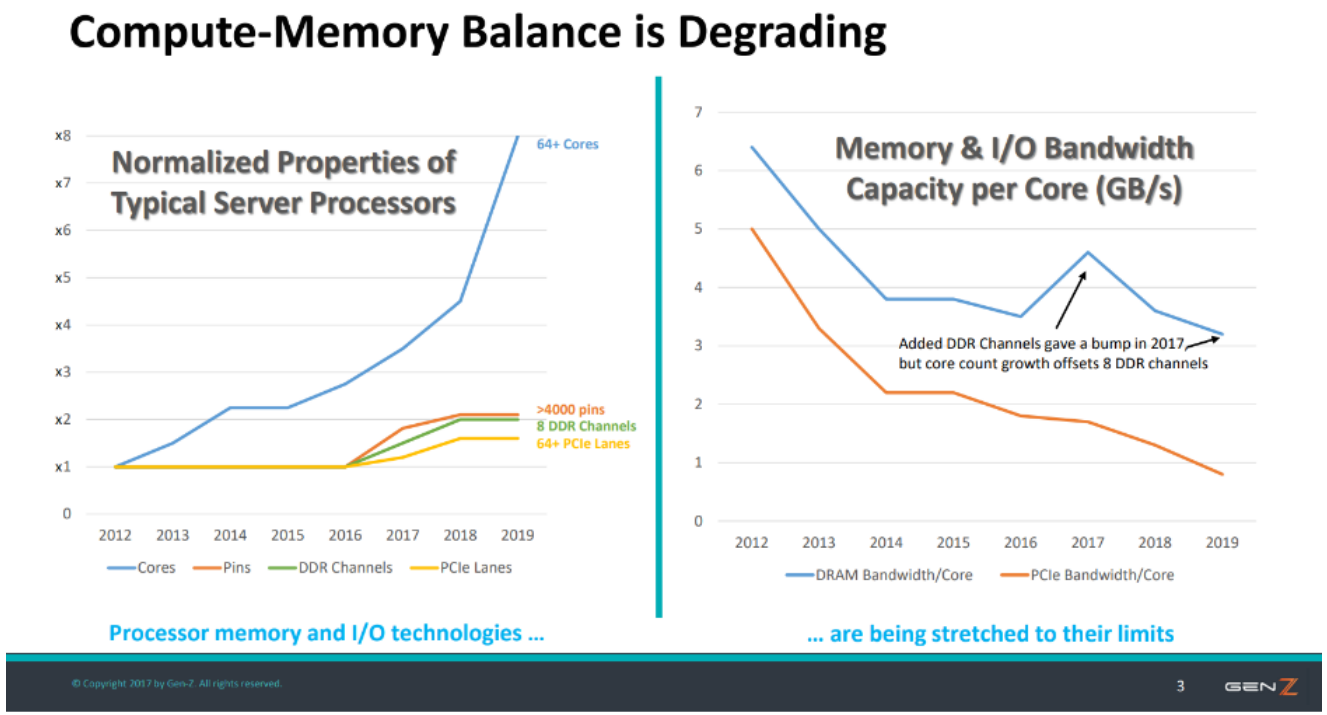
DRAM Outlook

- 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 thru 2019 due to additional production capacity coming online

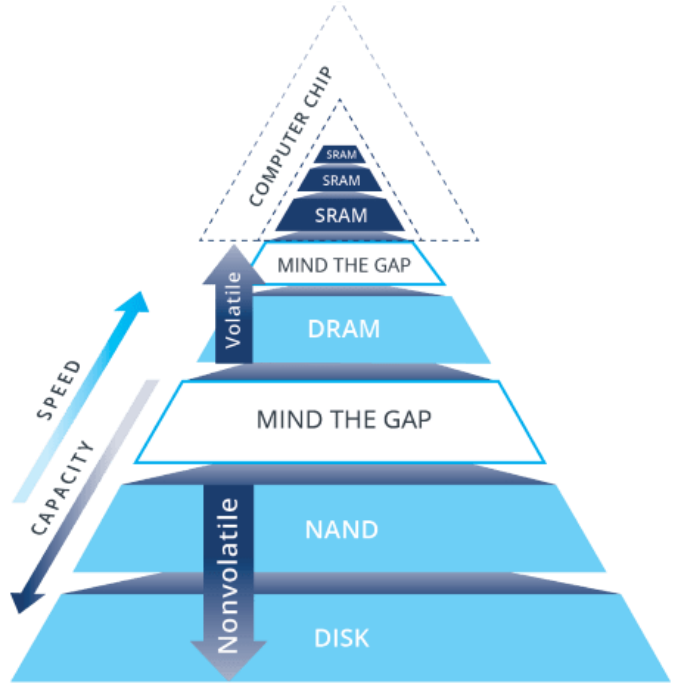


https://www.techinsights.com/technology-intelligence/overview/technology-roadmaps/

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 Comparison

SNIA PERSISTENT MEMORY PM SUMMIT
JANUARY 24, 2018 | SAN JOSE, CA

Technology	FeRAM	MRAM	ReRAM	PCM	DRAM	NAND Flash
Nonvolatile	Yes	Yes	Yes	Yes	No	Yes
Endurance	10^{12}	10^{12}	10^{6}	10^{8}	10^{15}	10^{3}
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

 14

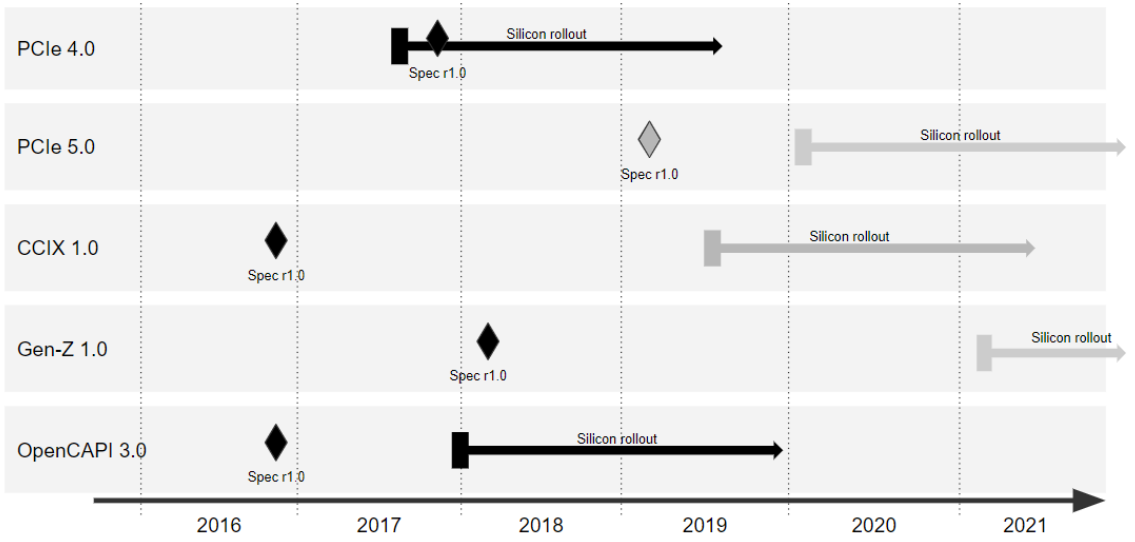
https://www.snia.org/sites/default/files/PM-Summit/2018/presentations/14_PM_Summit_18_Analysts_Session_Oros_Final_Post_UPDATED_R2.pdf

SUPPORTING TECHNOLOGIES

Interconnect technology

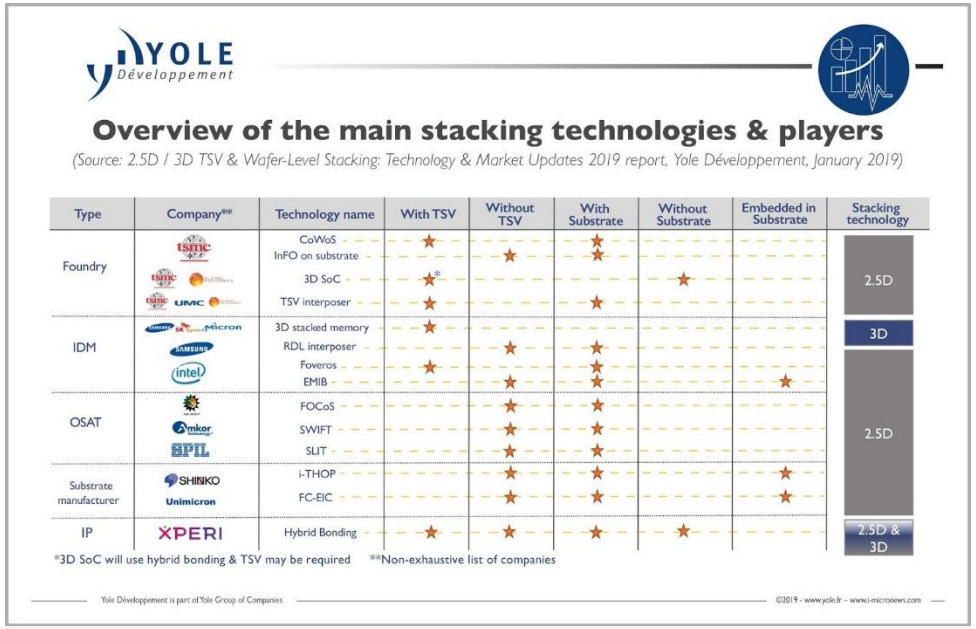
- Increasing requirements on bandwidth and latency driving the development
  - E.g. moving data between CPU and GPU is often a bottleneck
  - Several standards competing (PCIe Gen4/5, CCIX, Gen-Z, OpenCAPI, CXL…)
- Proprietary technologies
  - NVLink (GPU-to-GPU, GPU-to-POWER9)
  - Ultra Path (Intel),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

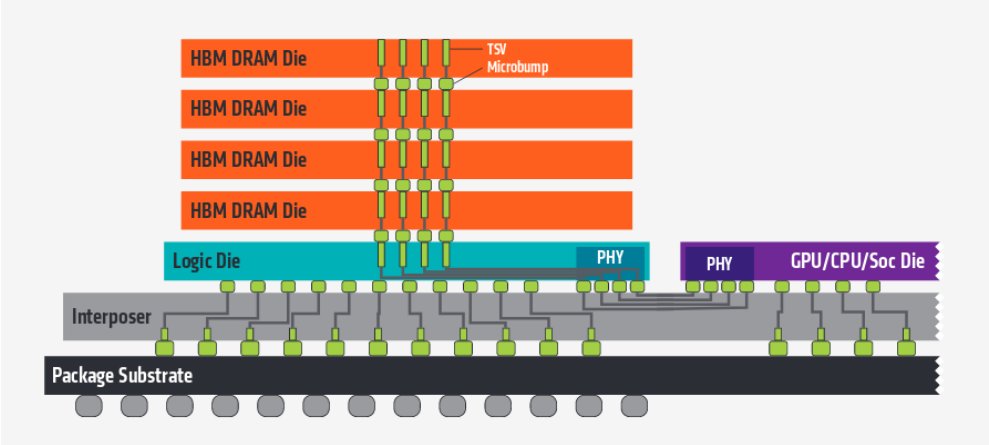
Standard	Physical Layer	Topology	Unidirectional Bandwidth	Mechanicals	Coherence
PCIe 4.0	PCIe PHY	p2p switched	16Gb/s/lane up to x16	PCIe	No
CCIX 1.0	PCIe PHY	p2p switched	25Gb/s/lane up to x16	PCIe	Full cache coherence between processors and accelerators
Gen-Z 1.0	IEEE 802.3 PCIe PHY	p2p switched meshed	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	SSF-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





What next?

- We do not really know what will be there in the HL-LHC era (2026-2037)
- Some “early indicators” of what might come next
  - Several nanoelectronics projects might help in
    - Increasing density of memory chips
    - Reducing size of transistors in IC
  - Nanocrystals, silicon nanophotonics, carbon nanotubes, single-atom thick graphene film, etc.
  - https://www.understandingnano.com/nanotechnology-electronics.html

Conclusions

- Market trends
  - Server market is increasing, AMD share as well
  - EUV lithography driving 7nm mass production
- CPU, GPUs and accelerators
  - AMD EPYC promising from a cost perspective
  - Nvidia GPUs still dominant due to the better software support
  - Recent developments for GPUs greatly favor inference workloads
  - FPGA market dominated by telecom, industry and automotive but there is also some HEP usage
- Memory technologies
  - SDRAM still the on-chip memory of choice, DRAM still for the main memory, no improvements in latency
  - NVDIMM – emerging memory packaging for memory between DRAM and NAND flash (see next talk)
  - Other non-volatile memory technologies in development

Additional resources

- All subgroups
  - https://gitlab.cern.ch/hepix-techwatch-wg
- CPUs, GPUs and accelerators
  - Document (link)
- Memory technologies
  - Document (link)

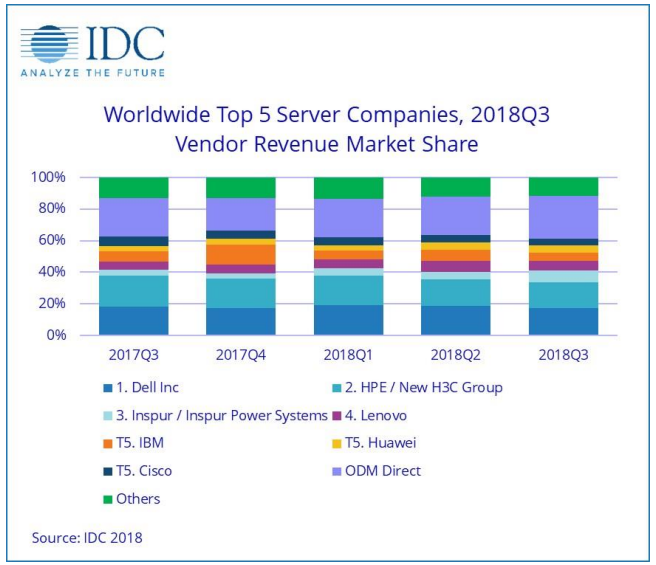
Acknowledgments

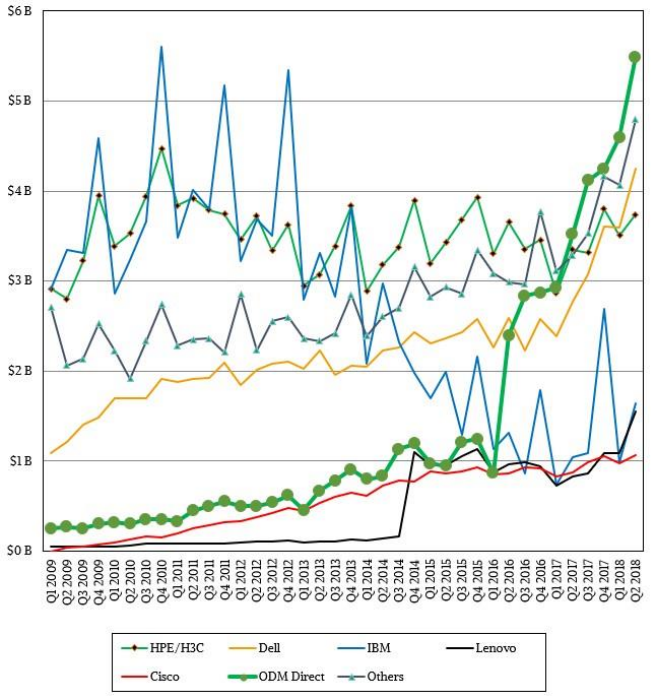
- Special thanks to Shigeki Misawa, Servesh Muralidharan, Peter Wegner, Eric Yen, Andrea Chierici, Chris Hollowell, Charles Leggett, Michele Michelotto, Niko Neufeld, Harvey Newman, Felice Pantaleo, Bernd Panzer-Steindel, Mattieu Puel and Tristan Suerink

BACKUP SLIDES

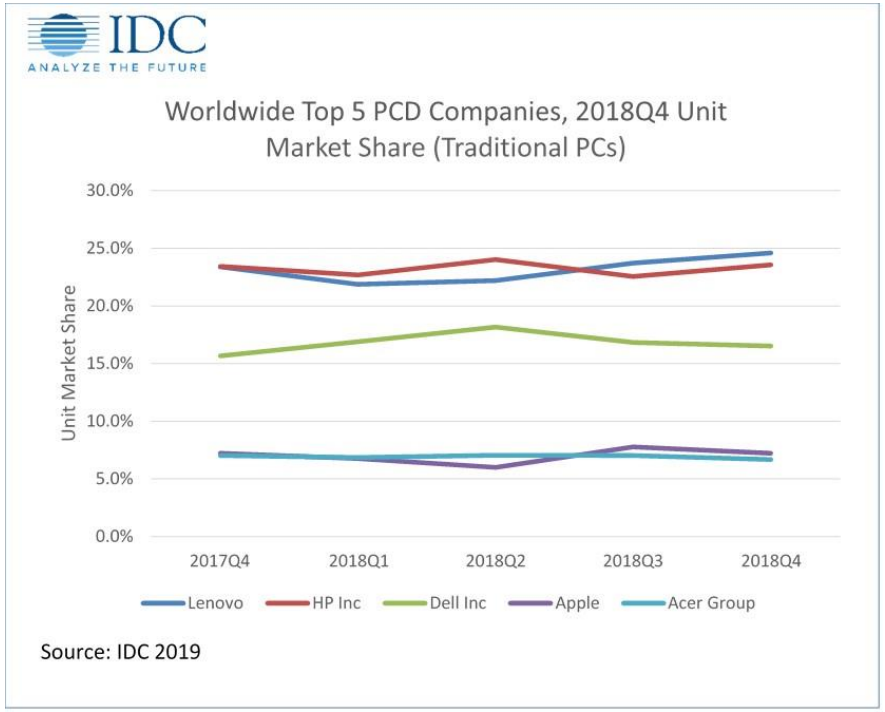
Market share of technology companies

Server companies





PC companies



- Worldwide server market increased 38%, year over year to $23 billion during the third quarter of 2018

eMAG, Graviton

- eMAG from Ampere is a V8 64 bit single socket SoC meant to compete with Xeon processors
  - Available in 16 and 32 cores
  - Eight DDR4 memory channels
  - 42 PCI-E Gen3 lanes
  - Using the TSMC 16nm FinFET+ process
- Graviton is available only via AWS
  - Could be the beginning of a new trend among hyperscalers, avoiding commercially available processors
  - not a good thing for HEP if it results in higher CPU prices due to drop of sales!