CPUs, GPUs and accelerators

x86

Intel server micro-architectures (1/2)

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)

SNEAK PEEK INTO THE FUTURE

Intel server micro-architectures (2/2)

Microarchitecture	Technology	Launch year	Highlights	CPU codename
Sunny Cove	10nm+	2019	Single threaded performance New instructions Improved scalability Larger L1, L2, µop caches and 2nd level TLB More execution ports AVX-512	Ice Lake Scalable Tiger Lake?
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	?

Source: Anandtech

Other Intel CPU architectures

- 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
- Hybrid CPUs
  - Will be enabled by Foveros, the 3D chip stacking technology recently demonstrated
- Itanium
  - It will be finally discontinued in 2021 (the only remaining customer is HP)

Other Intel-related news

- Record Q3 2018 results
  - Data-centric revenue rose 22%
  - PC revenue rose 16%
- Could not keep up with demand for the latest Xeon chips in 2018
- Serious issues with 10nm process as years behind scheduled
  - Pushing 14nm process to its limits
  - Claims that volume delivers on track for late 2019 and later
  - Being superseded by 7nm sooner than intended, which will be based on EUV lithography
    - Hopes that it will put Intel on track with Moore’s Law

AMD News

- Next gen desktop Matisse CPU (7nm) using Zen2 core achieves IPC parity with Intel, consumes less power and supports PCIe 4.0
  - Improved branch predictor unit and prefetcher, better micro-op cache management, larger micro-op cache, increased dispatch bandwidth, increased retire bandwidth, native support for 256-bit floating point math, double size FMA units, double size load-store units
- CSC announced an upcoming supercomputer using 3125 64-core EPYC “Rome” CPUs in 2020
- Market trend
  - Revenues increased by 23% over 2018 and profitability at its highest since 2011

AMD EPYC Naples (since Q2 ‘17)

AMD EPYC processors target the datacenter and specifically (not limited to) mono-processor servers. EPYC Naples (Zen architecture) is a single chip made up of 4 separate dies (multi-chip module), interconnected with Infinity Fabric links.

Main specs:

- 4 dies per chip (14nm), each die embedding IO and memory controllers, no chipset, SP3 sockets
- range of frequencies : 2.0-2.4 GHz, turbo up to 3.2 GHz
- 8 DDR4 memory channels with hardware, on the fly, encryption, up to 2600 MHz
- up to 32 cores (64 threads)
- up to 128 PCI gen3 lanes per processor (64 in dual)
- TDP range: 120W-180W

EPYC Naples processors have similar computing power compared to Intel Skylake processors (HS06 benchmarks on close frequencies and core count CPUs) with cutoff prices up to 49% (AMD claim).

Mostly compatible with Intel based x86, sparing for user code modifications.

AMD EPYC Rome (starting Q2 ‘19)

Next AMD EPYC generation (Zen2 based), embeds 9 dies (8 CPU 7nm chiplets for 1 I/O 14 nm die). All I/O and memory access is concentrated into a single die.

Main specs:

- 9 dies per chip : a 7nm single IO/memory die and 8 CPU 7nm chiplets
- 8 DDR4 memory channels, up to 3200 MHz
- up to 64 cores (128 threads) per processor
- up to 128 PCI gen3/4 lanes per processor
- SP3 / LGA-4094 sockets
- TDP range: 120W-225W (max 190W for SP3 compatibility)

Claimed **+20% performance per zen2 core** (over zen), **+75% through the whole chip** with similar TDP over Naples.

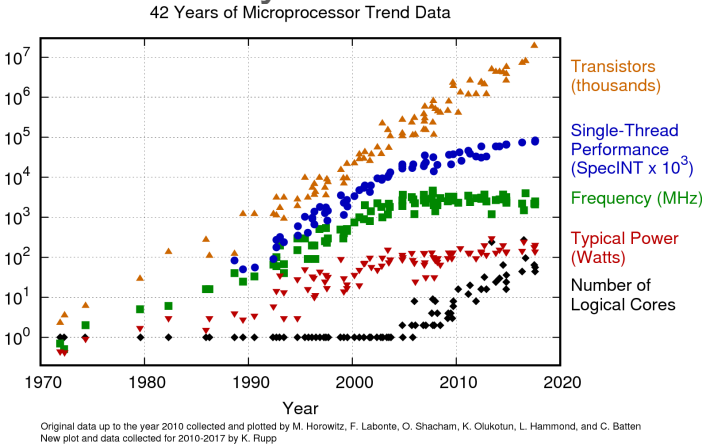
Available on DELL C6525 chassis starting from october.

Manufacturing technologies

- 10nm
  - Intel will ramp up in 2019, late by several years
  - Relies on DUVL (deep ultraviolet lithography) (193nm wavelength laser) requiring heavy use of multipatterning, which is problematic
- 7nm
  - Uses EUVL (extreme ultraviolet lithography) (13.5nm wavelength laser) reducing use of multipatterning and reducing costs
  - Intel on track and will start at the end of 2019
  - TSMC already making or will make chips for AMD, Apple, Nvidia and Qualcomm
  - Samsung Foundry started production and will make POWER CPUs for IBM from 2020
  - GlobalFoundries put it on hold indefinitely
- 5, 3, ? nm
  - Design costs increase exponentially
  - [to be expanded]

GPUs - NVIDIA Architecture

- For what concerns raw power GPUs are following the exponential trends wrt number of transistors and cores
- New features appear unexpectedly, driven by market



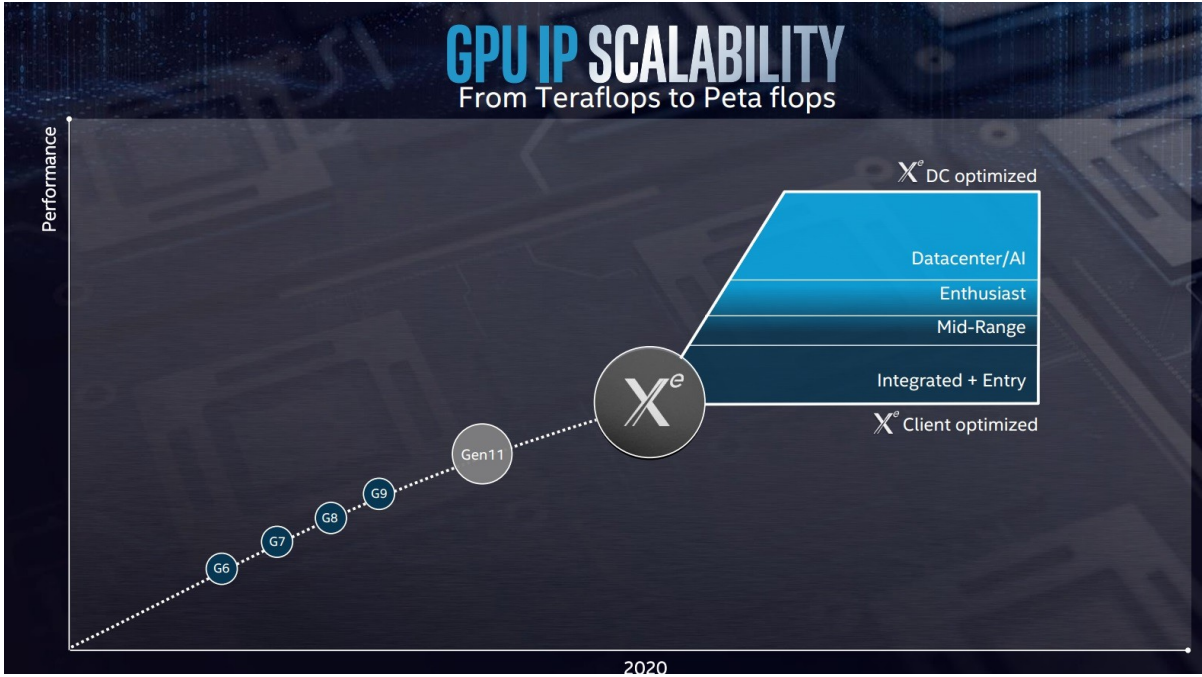
AMD Vega 20

- 7nm process allows to shrink the die and have more space for HBM2 memory, up to 32GB
- 2x bandwidth per ROP, texture unit, and ALU wrt Vega 10
- added support for INT8 and INT4 data types
  - useful for low-precision inference
- PCI Express 4 on AMD Radeon Instinct MI60

Intel

Entering the discrete GPU market in 2020 “Xe”

ATM just rumors

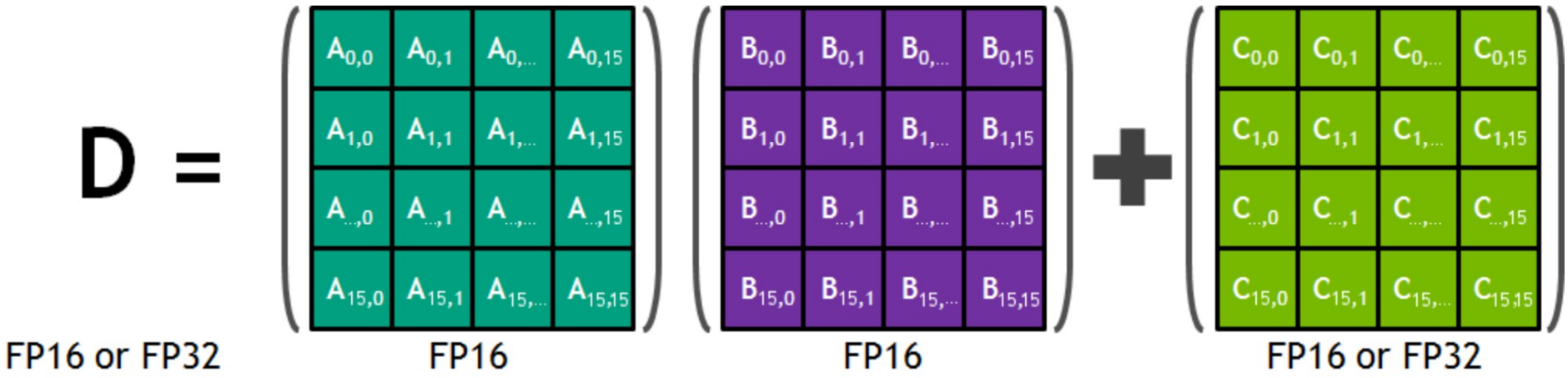


Tensor cores on NVIDIA Volta and AMD Vega 20

Tensor cores integrated on the GPU

Fast half precision multiplication and reduction in full precision

Useful for accelerating NN training/inference



GPUs - Programmability

- NVIDIA CUDA:
  - C++ based (supports C++14)
  - Many external projects
  - New hardware features available with no delay in the API
- OpenCL:
  - Not supported by NVIDIA
  - Can execute on CPU/iGPU/NVIDIA/AMD 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:
  - Similar to CUDA, still supports only a subset of the features
- GPU-enabled frameworks to hide complexity (Tensorflow)

GPUs - Programmability

Issue is performance portability and code duplication

At the moment, only possible solutions are based on trade-offs and DSL for very simple codes

- might work very well for analysis/ML, less for reconstruction

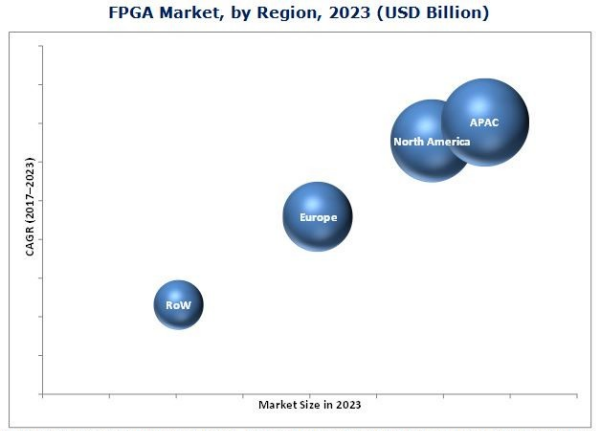
Hemi, Kokkos, RAJA...

GPUs in LHC experiments software frameworks

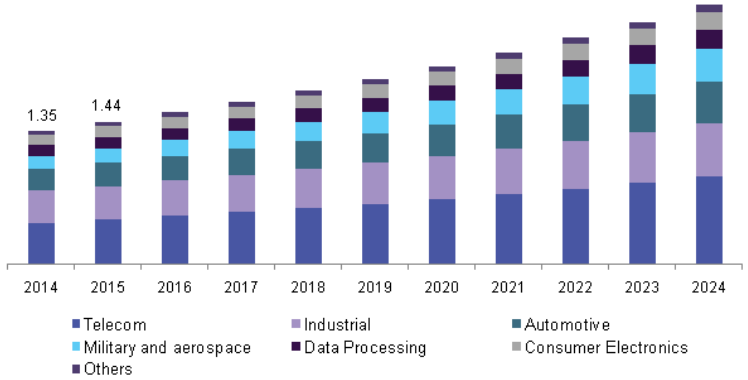
- Alice, O2
  - Tracking in TPC and ITS
  - Modern GPU replaces 40 CPU cores
- CMS, CMSSW
  - Demonstrated advantage of heterogeneous reconstruction from RAW to Pixel Vertices at the CMS HLT
  - 1 order of magnitude both in speed-up and energy efficiency wrt full Xeon socket
- LHCb (online) Allen: HLT-1 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

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 was valued at USD 5.34 Billion in 2016 and is expected to be valued at 9.50 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
- Telecommunications held the largest size of the FPGA market in 2016



Source: Investor Relation Presentations, Annual Reports, Expert Interviews, and MarketsandMarkets Analysis



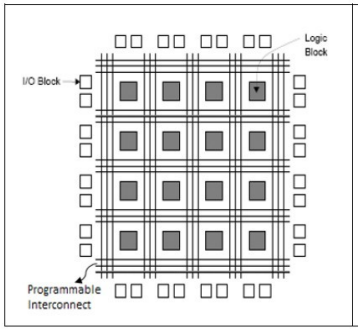
Sour
ce:

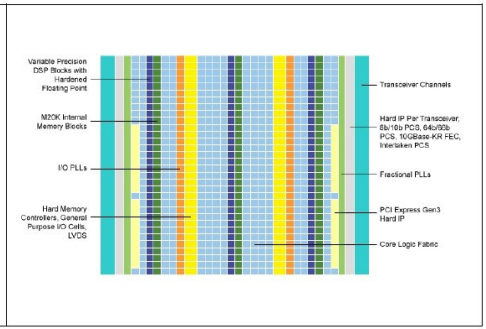
https://www.marketsandmarkets.com/Market-Reports/fpga-market-194123367.html

https
://ww
w.gra

FPGAs for Application Acceleration

- Programmability without sacrificing efficiency
- Highly suited for low latency applications
- Accelerates edge and streaming applications





Source: https://www.nextplatform.com/2018/10/15/where-the-fpga-hits-the-server-road-for-inference-acceleration/

Process Technology	20 nm		16 nm		14 nm	
	Intel®	Xilinx®	Intel®	Xilinx®	Intel®	Xilinx®
Best Performance Or Fastest, Most Powerful		Virtex® UltraScale®		Virtex® UltraScale+® Zynq® UltraScale+®	Intel® Stratix® 10	
Best Price/performance/watt Or Balance of cost, power, performance	Intel® Arria® 10	Kintex UltraScale®				
Cost-Optimized Or Low system cost plus performance	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

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

Sour
ce:
https
://ww
w.eet

FPGAs 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/