AI, FPGA and HPC

High Performance Computing is…

swarm of processors

TOP500: Benchmark in HPC

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,414,592	148,600.0	200,794.9	10,096
2	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM / NVIDIA / Mellanox	1,572,480	94,640.0	125,712.0	7,438
3	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	Texas Advanced Computing Center/Univ. of Texas United States	**Frontera** - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR Dell EMC	448,448	23,516.4	38,745.9	
	Industrial Science and Technology (AIST) Japan	**(ABCI)** - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR Fujitsu				
9	Leibniz Rechenzentrum Germany	**SuperMUC-NG** - ThinkSystem SD650, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path Lenovo	305,856	19,476.6	26,873.9	
10	DOE/NNSA/LLNL United States	**Lassen** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Tesla V100 IBM / NVIDIA / Mellanox	288,288	18,200.0	23,047.2	

"HPC" system is large as it consumes > MW.
Energy efficiency are even more critical.

TOP500 Nov 2019

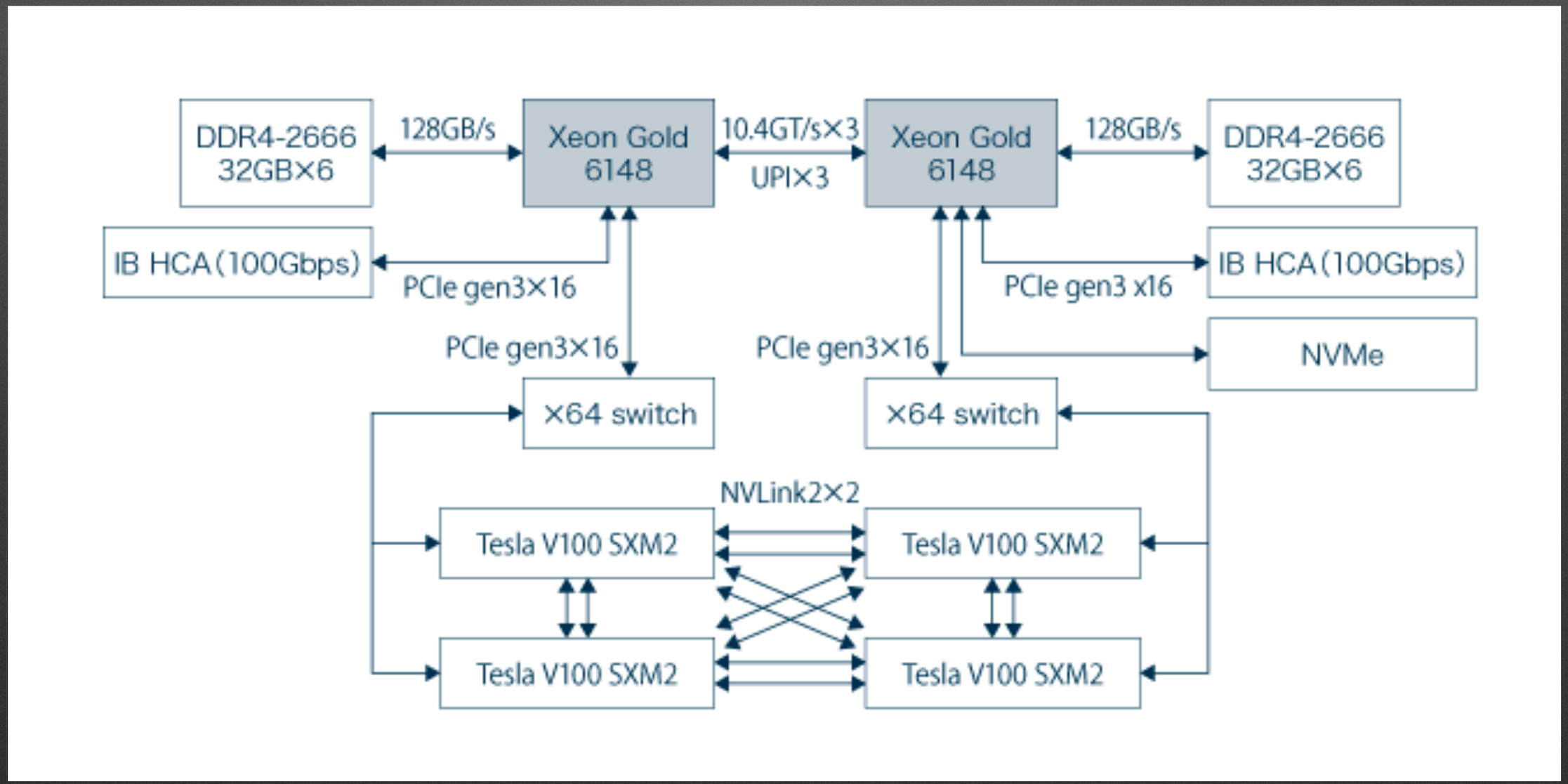
Energy Efficiency in HPC

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	159	**A64FX prototype** - Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D , Fujitsu Fujitsu Numazu Plant Japan	36,864	1,999.5	118	16.876
2	420	**NA-1** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	1,271,040	1,303.2	80	16.256
3	24	**AiMOS** - IBM Power System AC922, IBM POWER9 20C 3.45GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100 , IBM Rensselaer Polytechnic Institute Center for Computational Innovations (CCI) United States	130,000	8,045.0	510	15.771
4	373	**Satori** - IBM Power System AC922, IBM POWER9 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 SXM2 , IBM MIT/MGHPCC Holyoke, MA United States	23,040	1,464.0	94	15.574
5	1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	10,096	14.719
6	8	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	1,649	14.423
7	494	**MareNostrum P9 CTE** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Tesla V100 , IBM Barcelona Supercomputing Center Spain	18,360	1,145.0	81	14.131
8	23	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	135,828	8,125.0	792	13.704
9	11	**PANGEA III** - IBM Power System AC922, IBM POWER9 18C 3.45GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100 , IBM Total Exploration Production France	291,024	17,860.0	1,367	13.065
10	2	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	7,438	12.723

GREEN500 Nov 2019

AI bridging Cloud @ AIST

- ***Each node is Server CPU(Xeon) + GPU(Volta)***
- ***Mainly targeting for Machine Learning(ML)***



HPC(super computers) & Cloud

Accelerate machine learning and high performance computing applications with powerful GPUs

GPU is effective for ML & GPU are ubiquitous in the cloud

“HPC” system and “Cloud” system are somehow converging

Technology for better EE

- More use of natural energy
- More advanced silicon technology
- Faster processors
- Faster communication technology
- Better cooling system
- More energy efficient architecture
  - More “Parallel” !
  - Integration of domain specific units

Now “Computer” is…

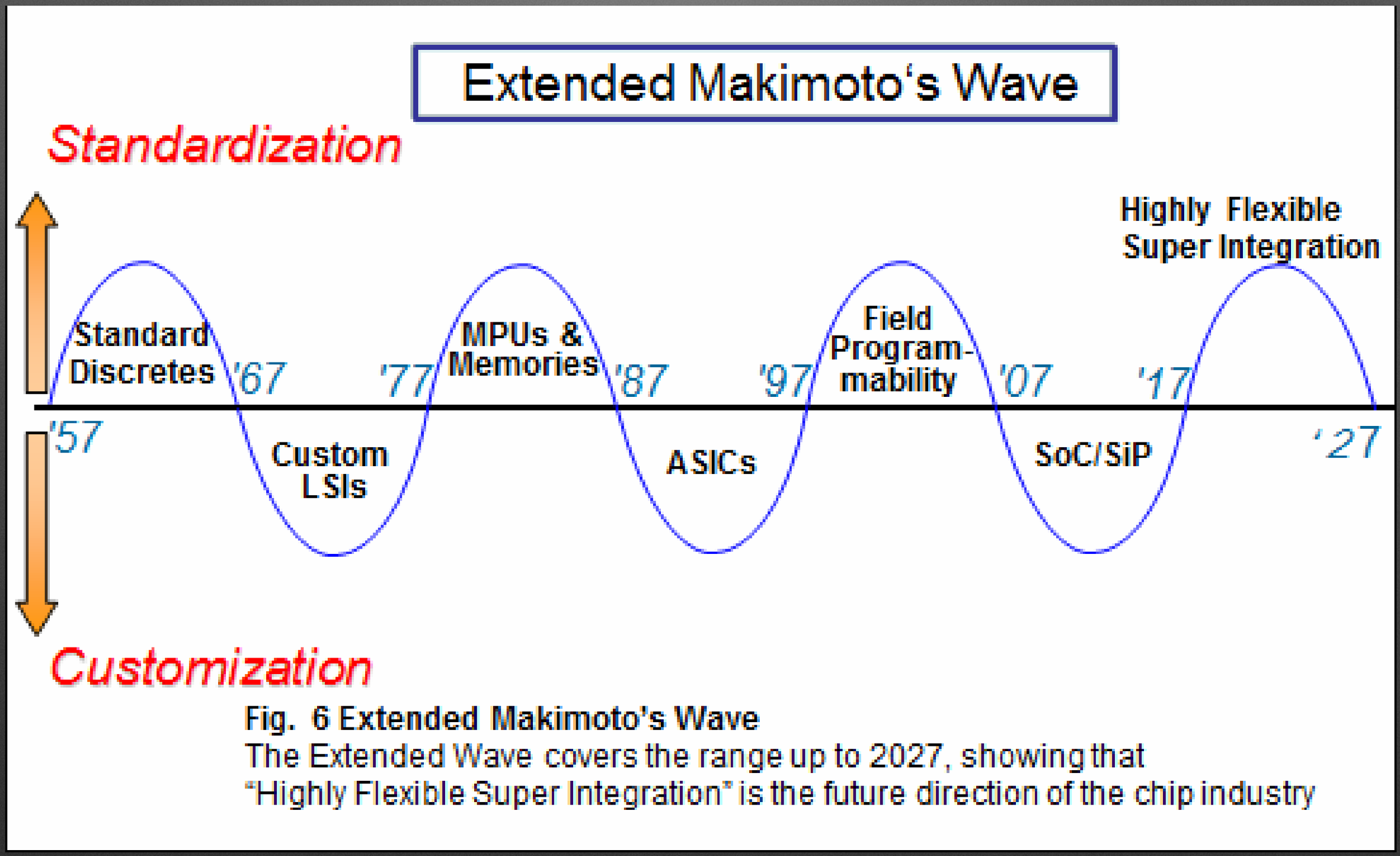
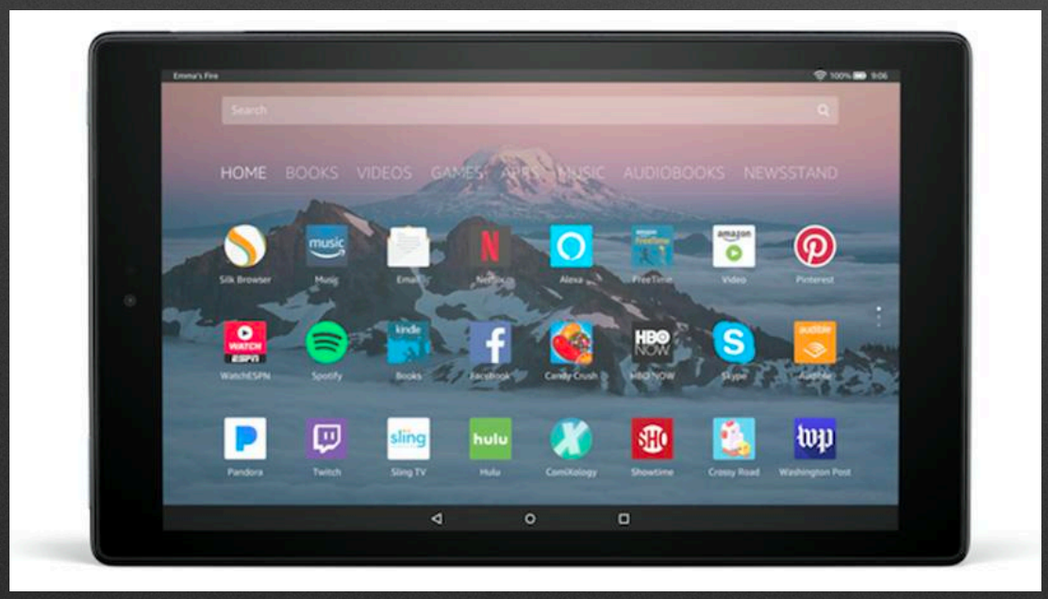


Fig. 6 Extended Makimoto's Wave
The Extended Wave covers the range up to 2027, showing that
“Highly Flexible Super Integration” is the future direction of the chip industry

http://www.shmj.or.jp/makimoto/pdf/makimoto_04_02.pdf

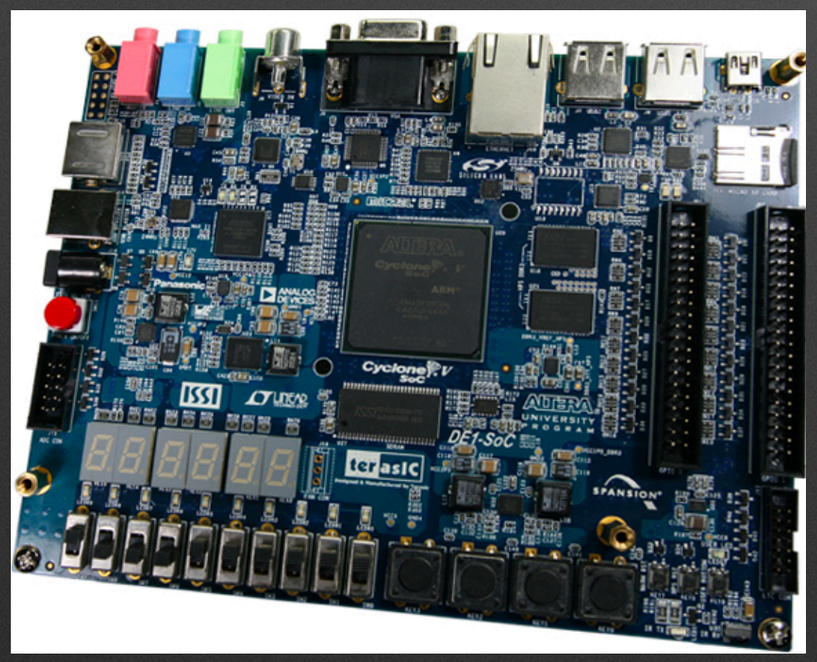
Integration of …

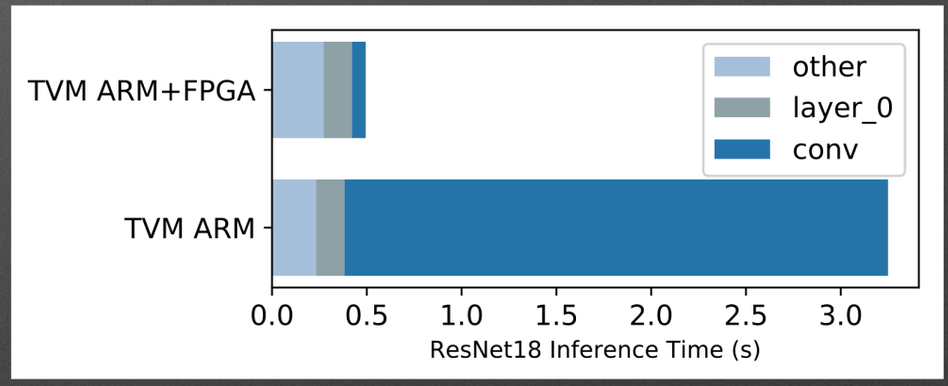
- ***Processor and GPU***
  - ***Game consoles : energy efficiency and performance***
  - ***Smartphones : UI and game performance***

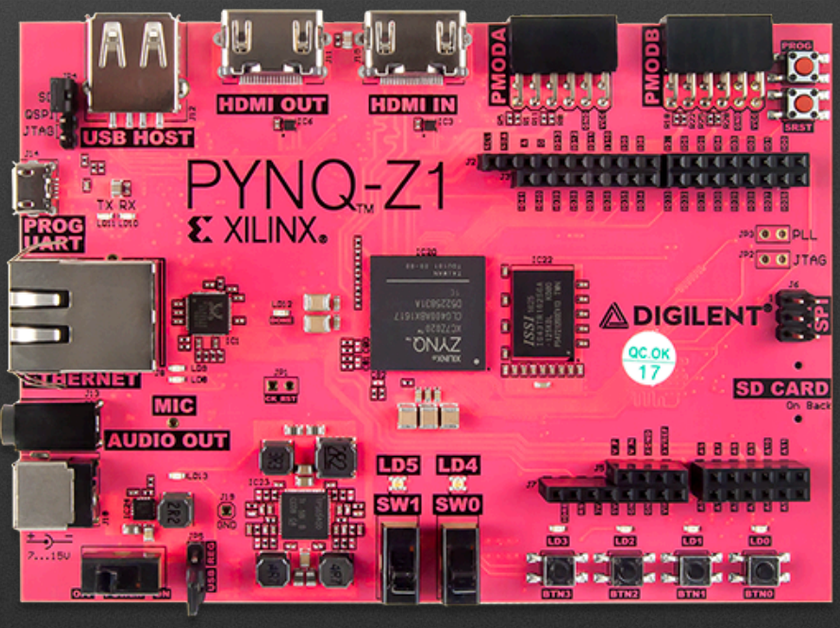


Integration of …

- ***Processor and FPGA***
  - *Very useful for education of logic and FPGA*
  - *Prototyping of processors*
  - *Replace ARM SoC with FPGA SoC for efficiency*
  - *Real custom computing with CPU*

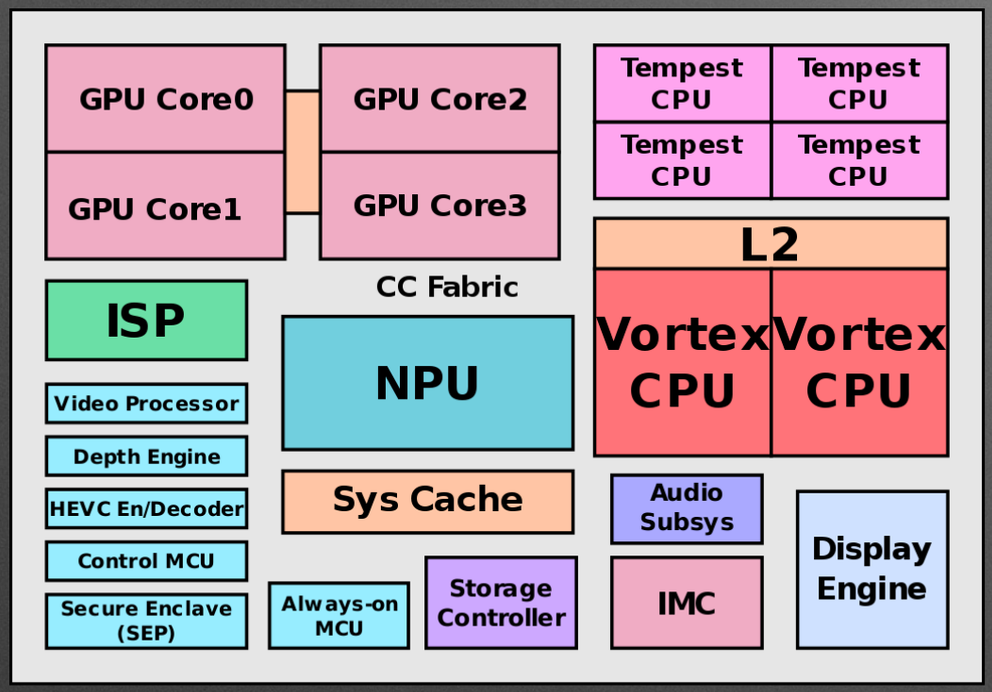


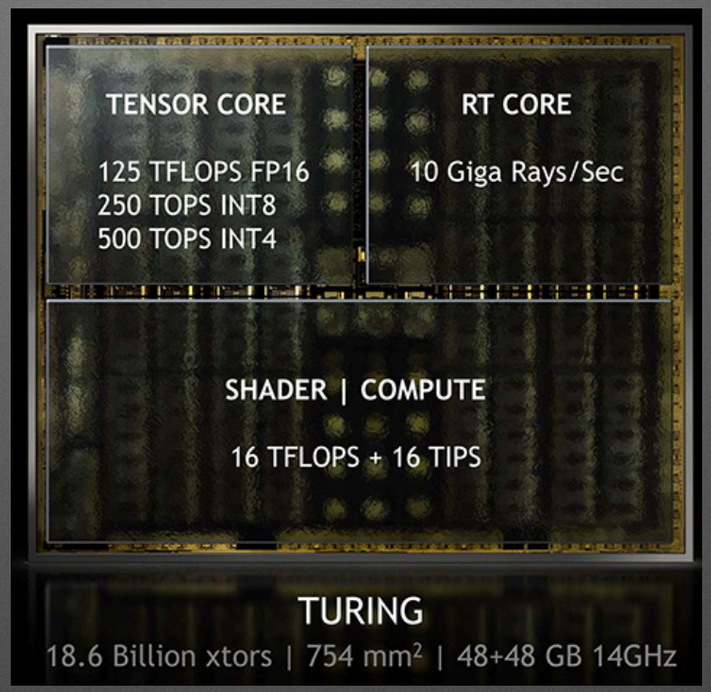


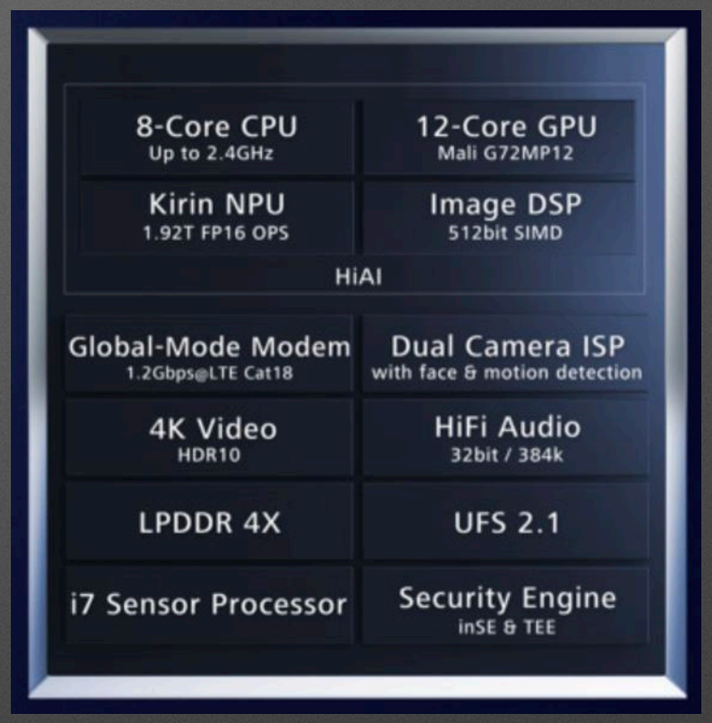


Integration of …

- ***Processor and "AI processor"***







Apple A12 ***NVIDIA TURING*** ***HiSilicon Kirin***

So many "Neural Processor" (NPU) Development

- ***Google TPU, TPUv2, TPUv3***
- ***Fujitsu Deep Learning Unit***
- ***DMP ZIA DV700***
- ***Microsoft Project Brainwave (FPGA)***
- ***Baidu XPU (FPGA)***

TOP500: System Configuration

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,414,592	148,600.0	200,794.9	10,096
2	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM / NVIDIA / Mellanox	1,572,480	94,640.0	125,712.0	7,438
3	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	Texas Advanced Computing Center/Univ. of Texas United States	**Frontera** - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR Dell EMC	448,448	23,516.4	38,745.9	
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray/HPE	387,872	21,230.0	27,154.3	2,384
7	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray/HPE	979,072	20,158.7	41,461.2	7,578
8	National Institute of Advanced Industrial Science and Technology (AIST) Japan	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR Fujitsu	391,680	19,880.0	32,576.6	1,649
9	Leibniz Rechenzentrum Germany	**SuperMUC-NG** - ThinkSystem SD650, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path Lenovo	305,856	19,476.6	26,873.9	
10	DOE/NNSA/LLNL United States	**Lassen** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Tesla V100 IBM / NVIDIA / Mellanox	288,288	18,200.0	23,047.2	

Power9 + V100

Power9 + V100

SW26010

Xeon + Matrix-2000

Xeon

Xeon + P100

Xeon Phi

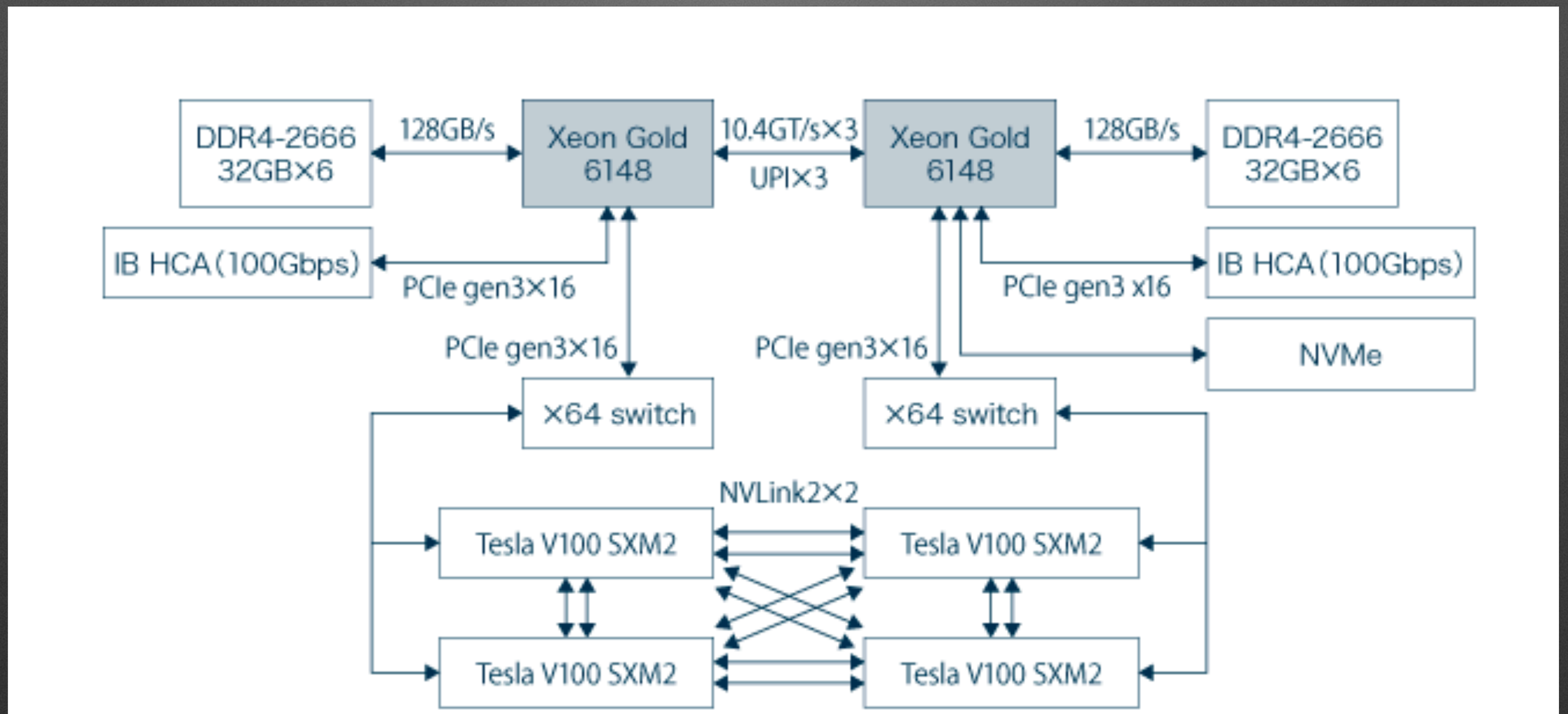
Xeon + V100

Xeon

Power9 + V100

Typical GPU based system

- ***Each node is Server CPU(Xeon) + GPU(Volta)***
- ***No memory sharing, long latency…***



Typical HPC systems are Not Integrated yet

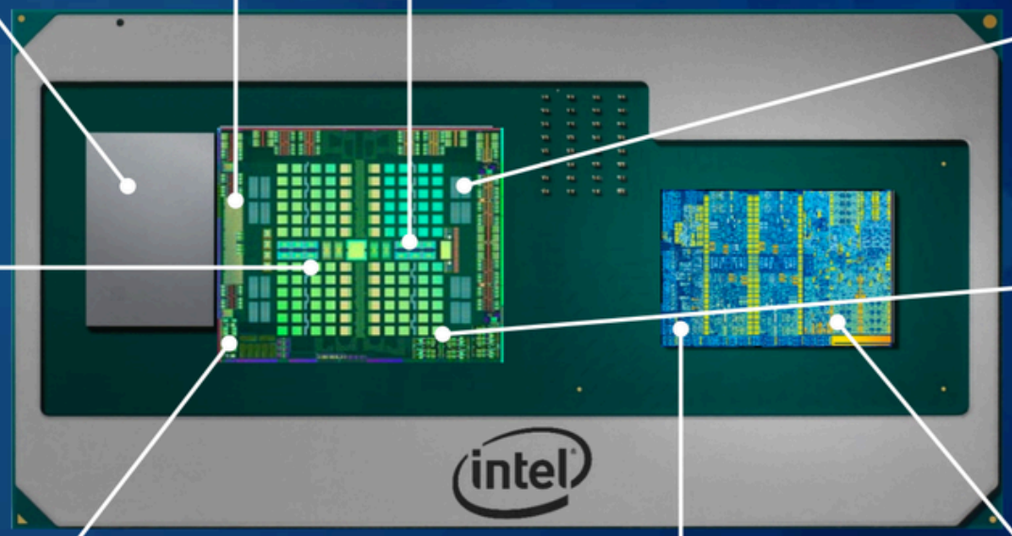
TWO AMAZING GRAPHICS SUBSYSTEMS ON ONE SMALL PACKAGE

High Bandwidth Cache Controller

High Bandwidth Cache
- 4GB Capacity
- 1024 bit bus width
- Low power

Quad Geometry Engines

Vega Pixel Engine
- Up to 16 Render Back Ends
- Up to 64 Pixels/Clock



Compute Units
- Up to 24 Compute Units
- Asynchronous Dispatch
- Per Compute Unit Power Gating
- Vulkan® & DirectX™ 12 Ready
- Supports Radeon Shader Intrinsics

Radeon™ Multimedia Engine
- 4K60 encode / decode with Radeon ReLive
- HEVC, H264 HDR enc/dec

Intel® Quick Sync Video
- VP9 & HEVC 10b HW enc/dec
- H264 HW enc/dec

MEDIA

DISPLAY

Radeon™ Display Engine
- 6 Displays
- Up to 4K resolution
- Display Port 1.4 w/ HDR
- HDMI 2.0b with HDR10 support

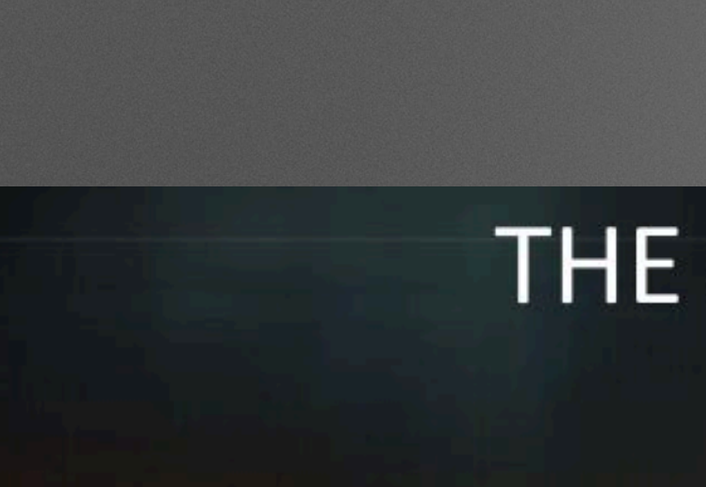
Intel® Gfx Display Engine
- 3 Displays
- 4K resolution
- eDP /PSR for long battery life

CPU/GPU integrated for consumer PCs

THE ARCHITECTURE ADVANTAGE

INTEGRATION: “ZEN” & “VEGA”

UNITED WITH INFINITY FABRIC



“ZEN” CORE COMPLEX

“VEGA” GRAPHICS

MULTIMEDIA ENGINES

INFINITY FABRIC

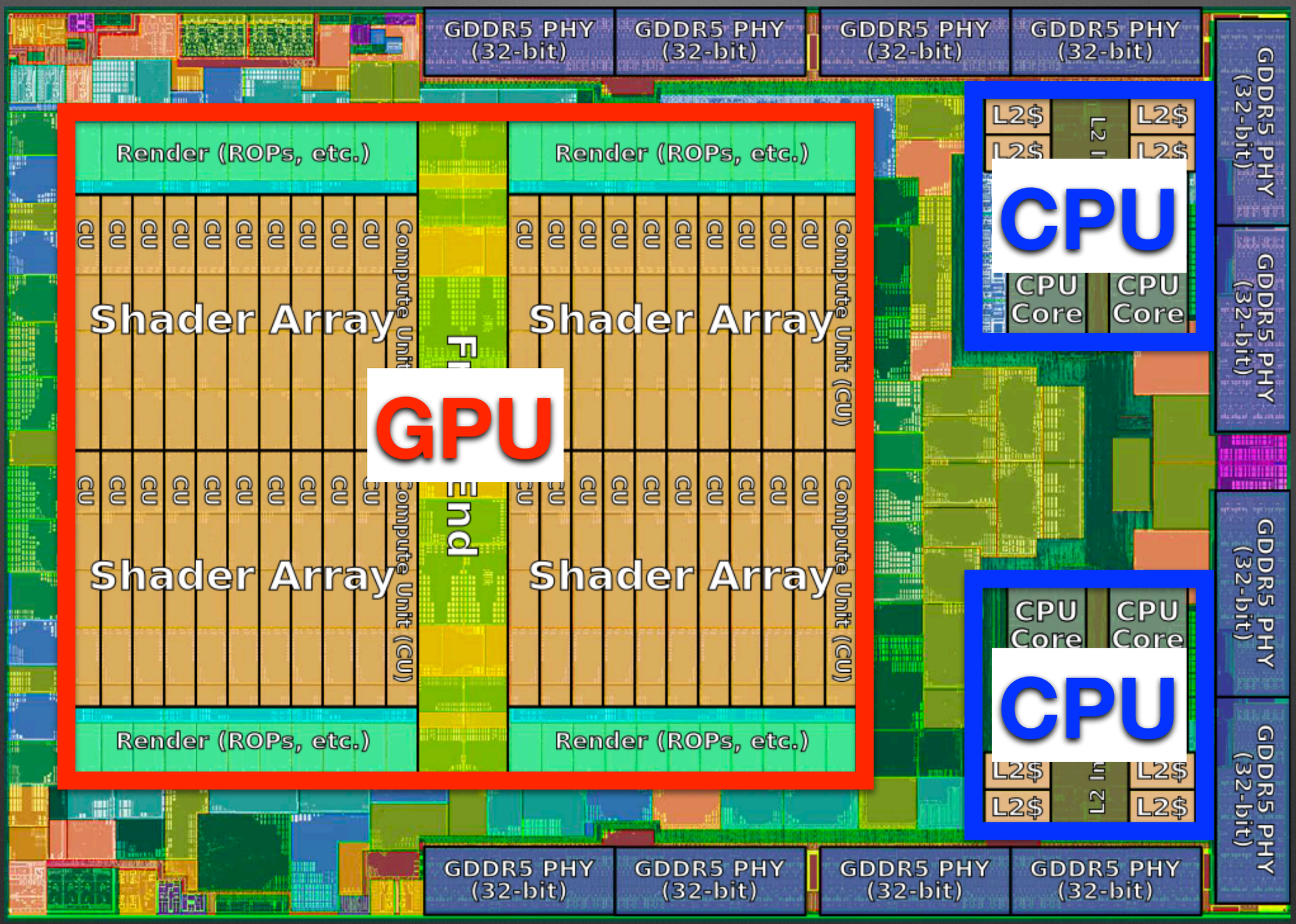
I/O AND SYSTEM HUB

DISPLAY ENGINE

DDR4 MEMORY CONTROLLERS

CPU/GPU integrated for consumer PCs

Scorpio Engine (Xbox One)



In HPC: No integrated processor

- Integration of CPU and accelerator is a natural consequence as cost effective and efficient architectures
  - *CPU dominated by Intel*
  - *GPU dominated by NVIDIA*

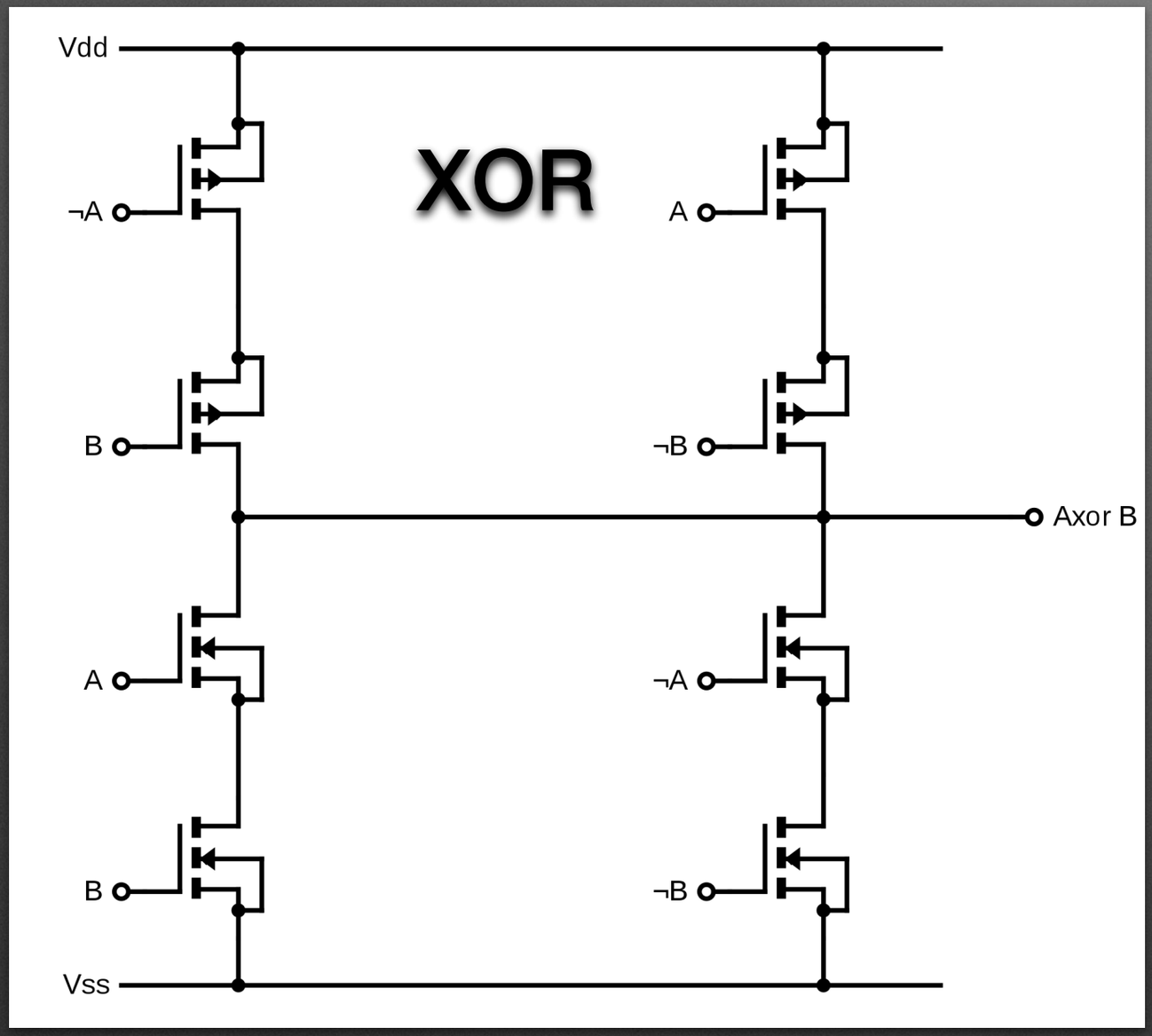
Intel has no GPU for HPC & NVIDIA has no CPU for HPC
Development of integrated processors will be necessary
Chance for AMD
Research using FPGA...

Rise of FPGA in HPC

- "Cloud" has a node with FPGA
  - Amazon AWS F1
  - Alibaba Cloud
  - Microsoft Project BrainWave (ML)

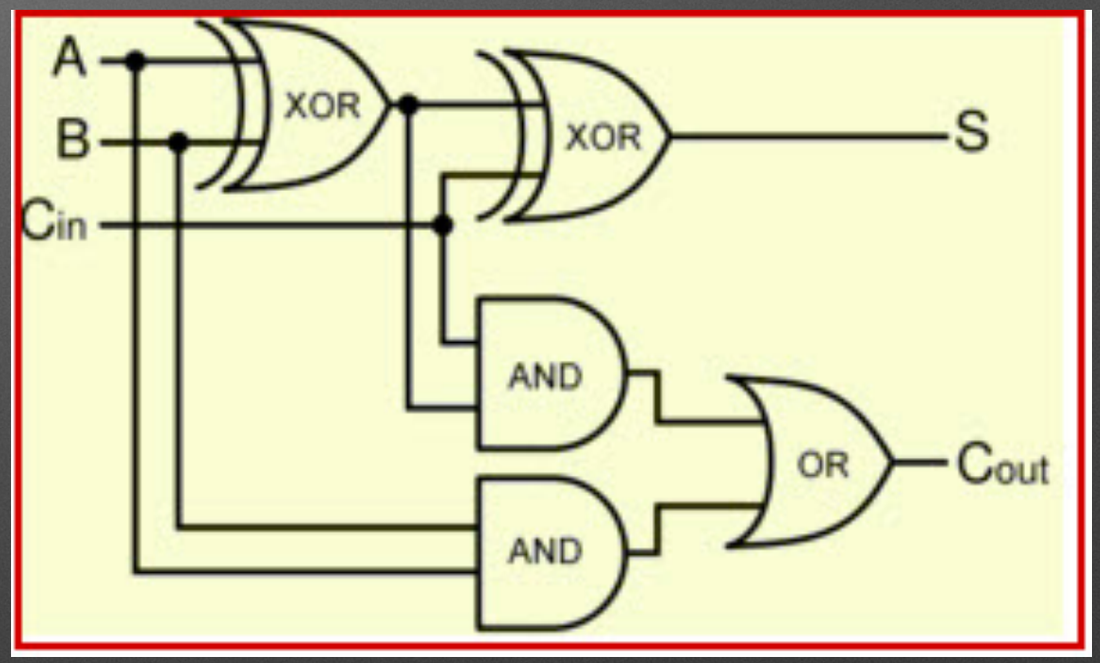
- FPGA based HPC deployment
  - Cray CS500 in Parderborn Univ. (Germany)
    - Out of 272 nodes, 32 nodes has Stratix10
  - Tsukuba U. will deploy "FPGA+GPU" system

“Computer” is based on logic



XOR ciruit using transisters

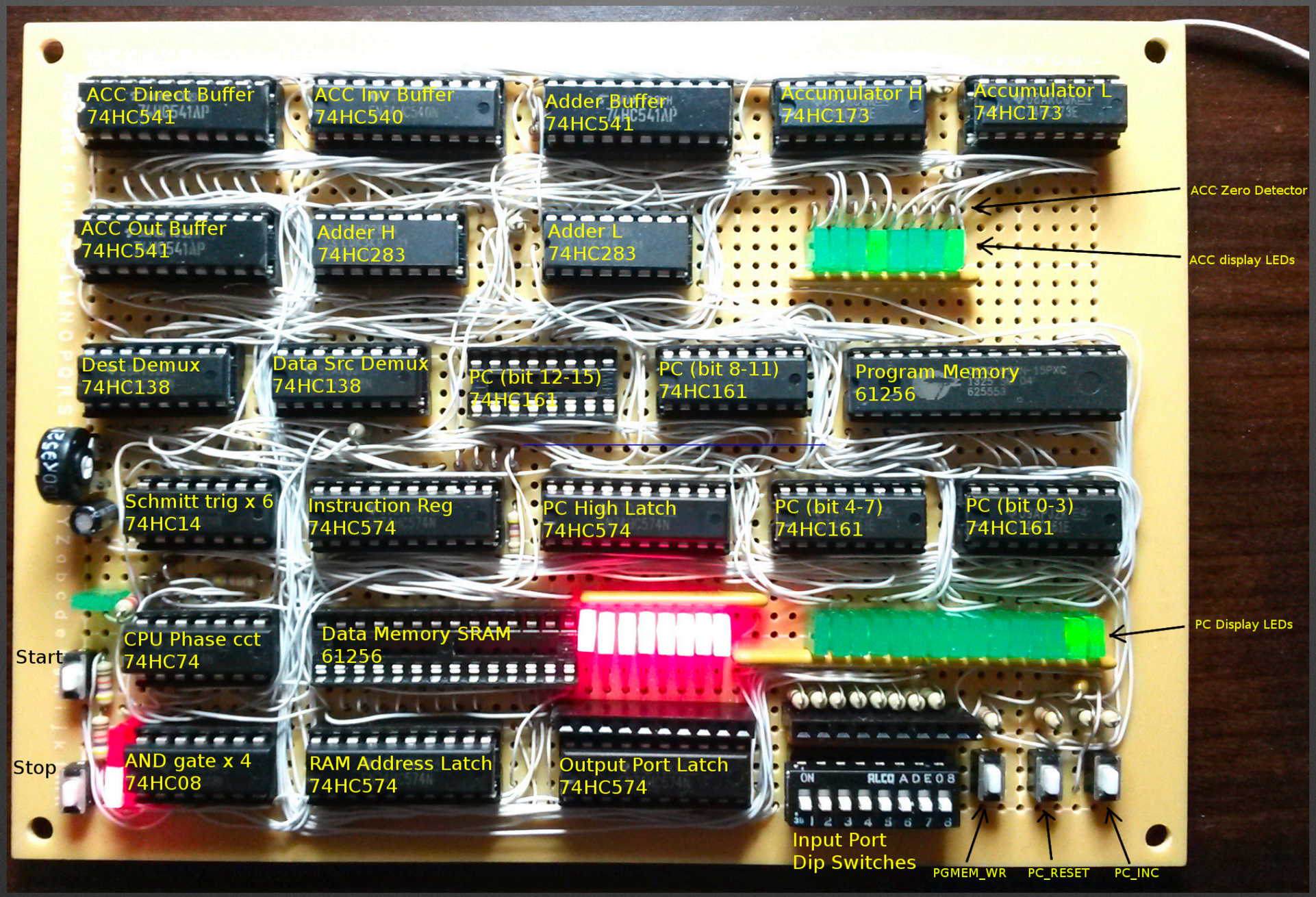
Full Adder



XOR, AND and OR

Handmade CPU

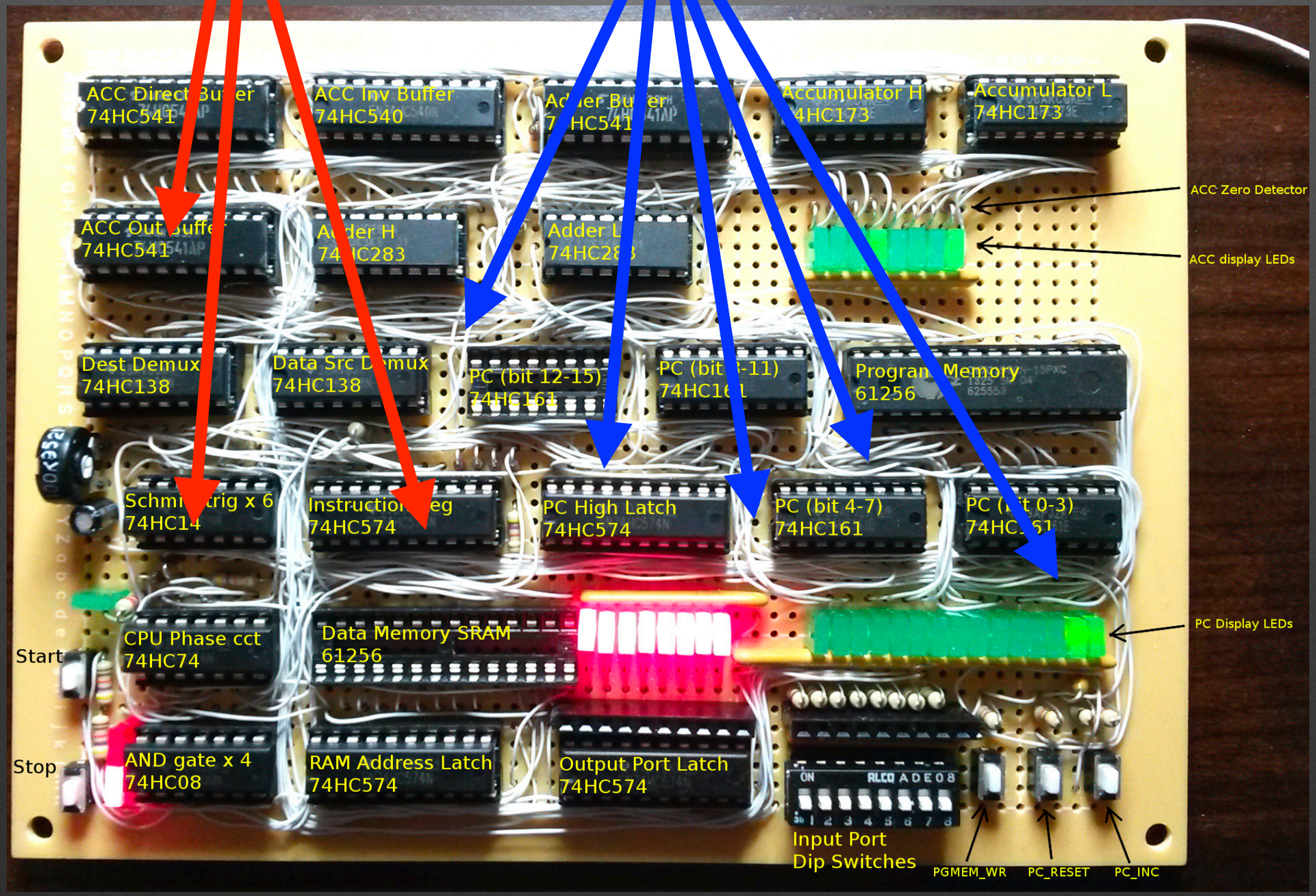
Using logic ICs



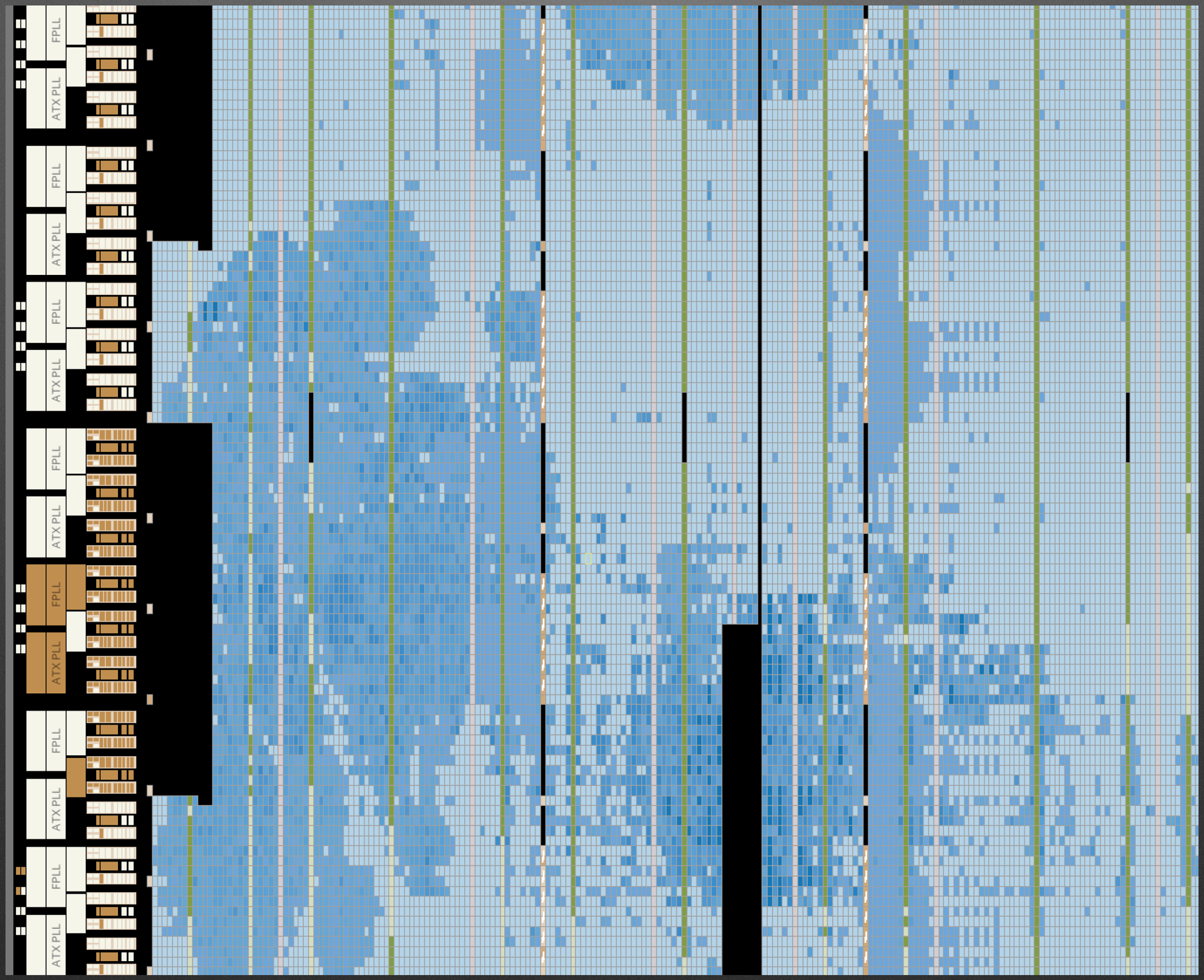
http://digitarworld.uw.hu/ttlcpu.html

FPGA: Reconfigurable

Inside FPGA, we can change the logic and connection network



FPGA internal structure



Recent FPGA has FP units

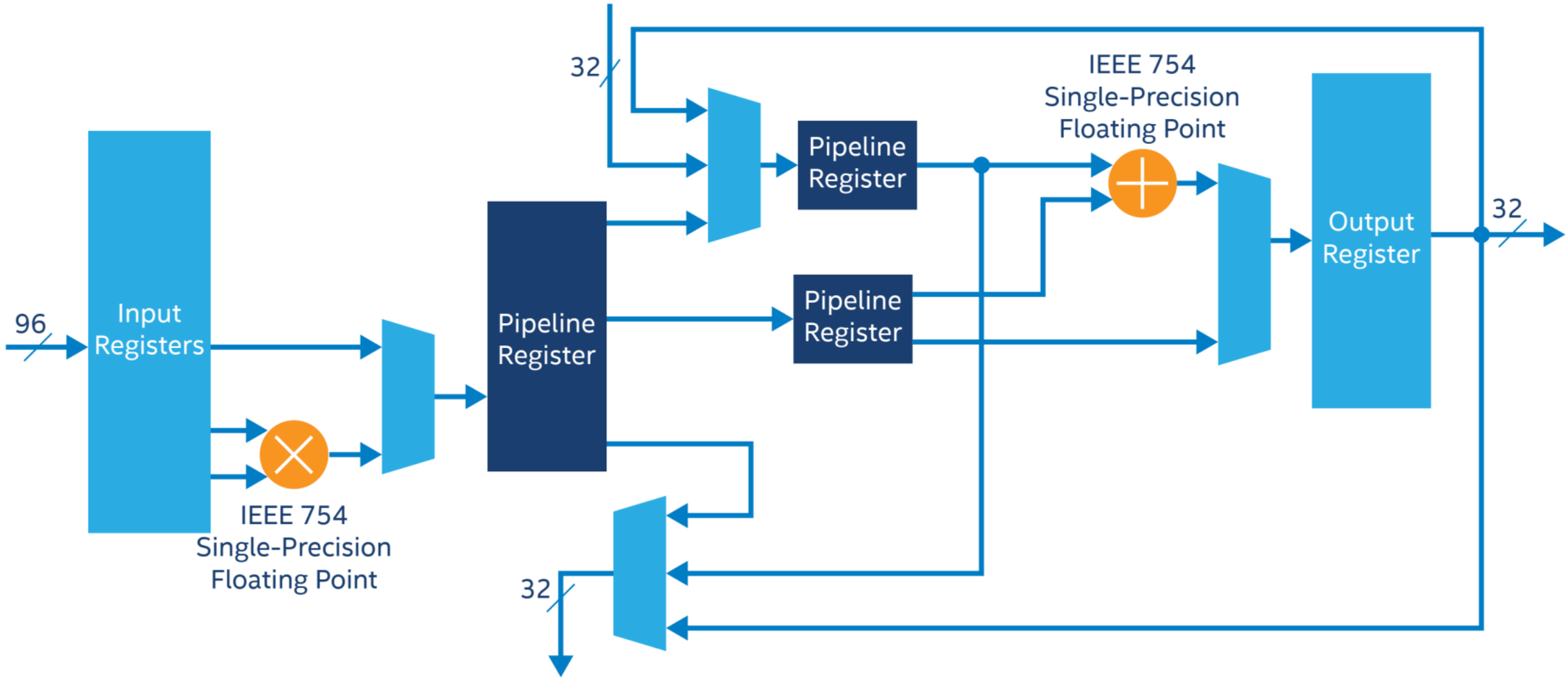
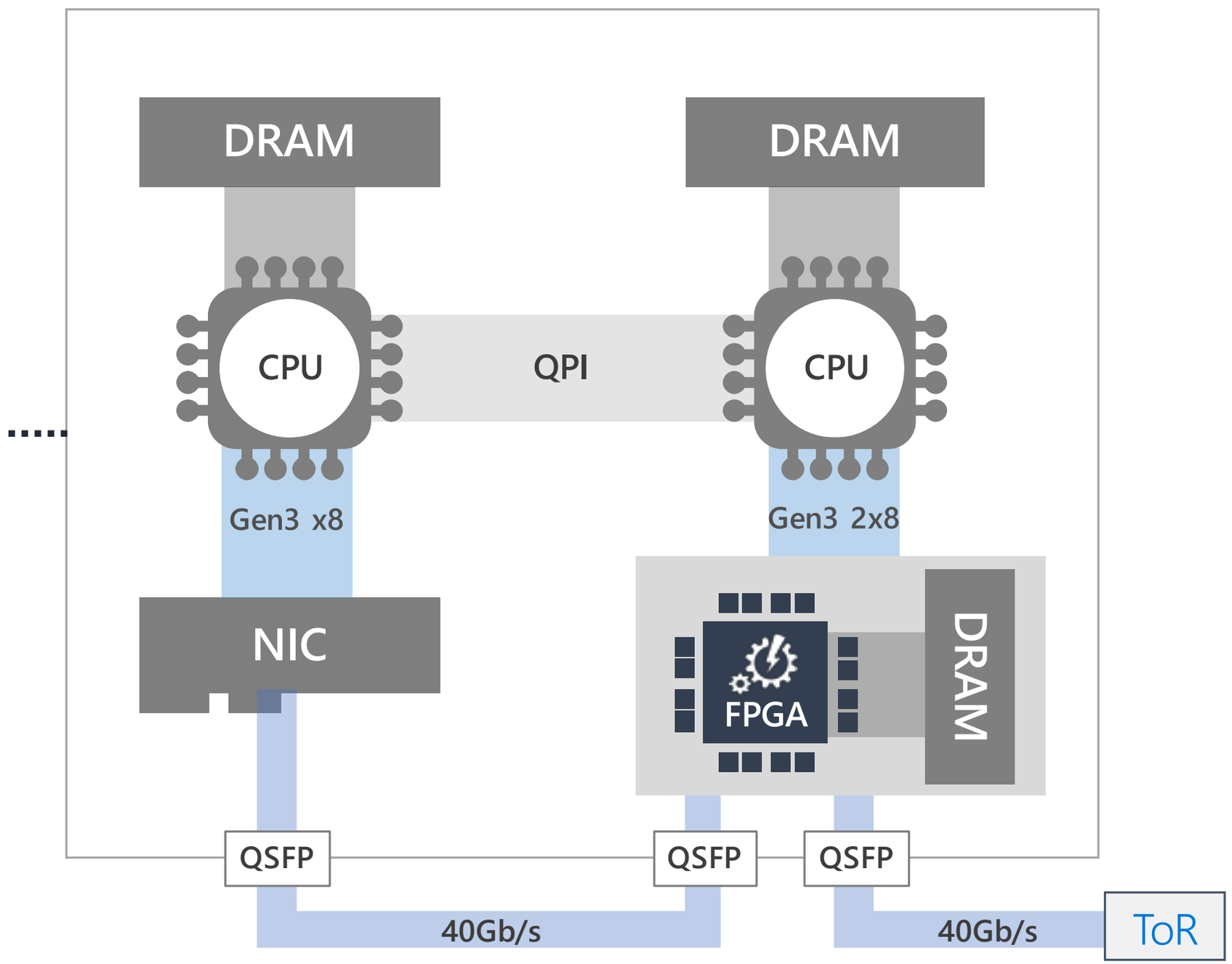


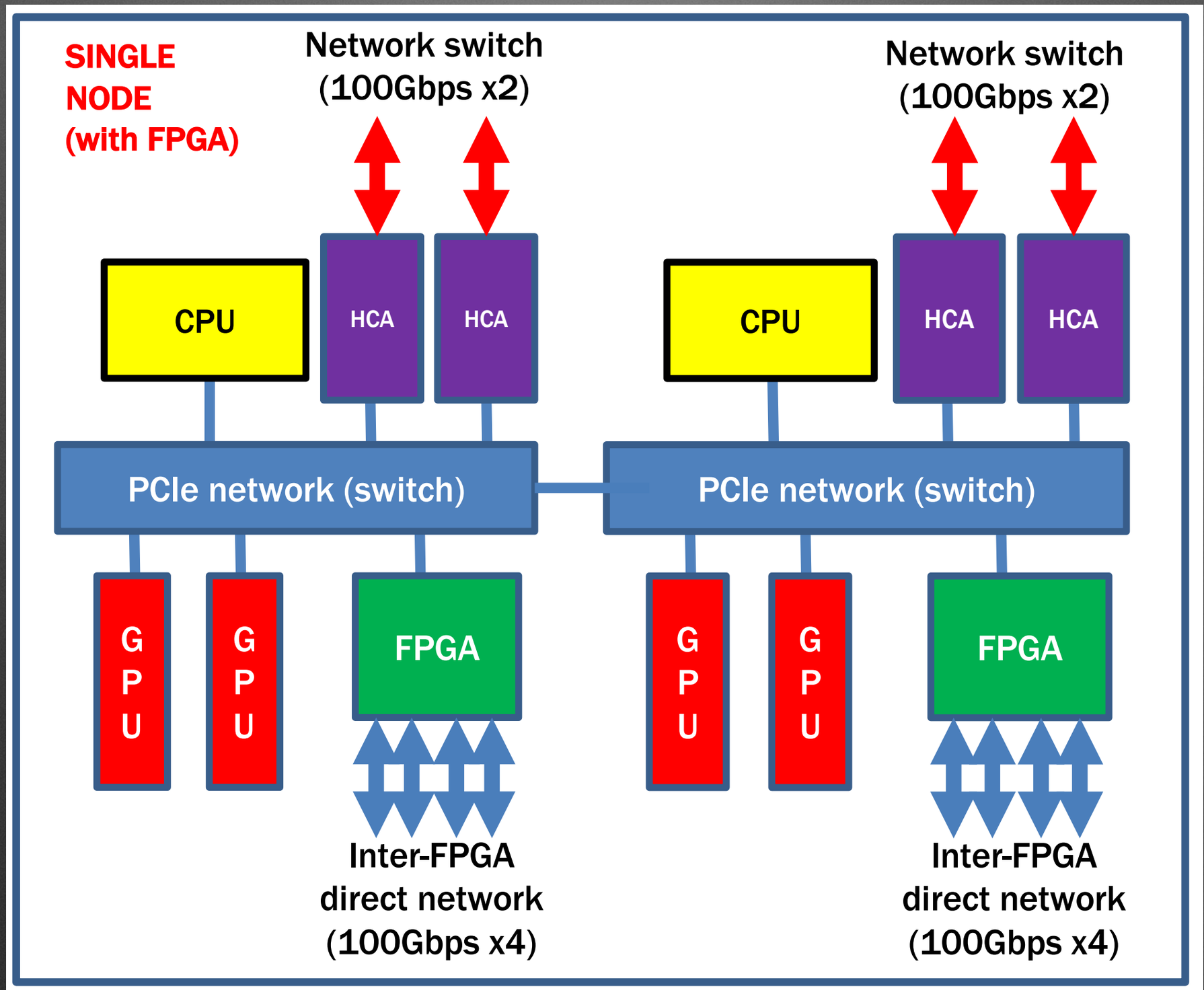
Figure 3. Floating-Point DSP Block Architecture in FPGAs

Arria10 ~ 1.5 TFLOPS in 32bit FP
Stratix10 ~ 10 TFLOPS in 32bit FP

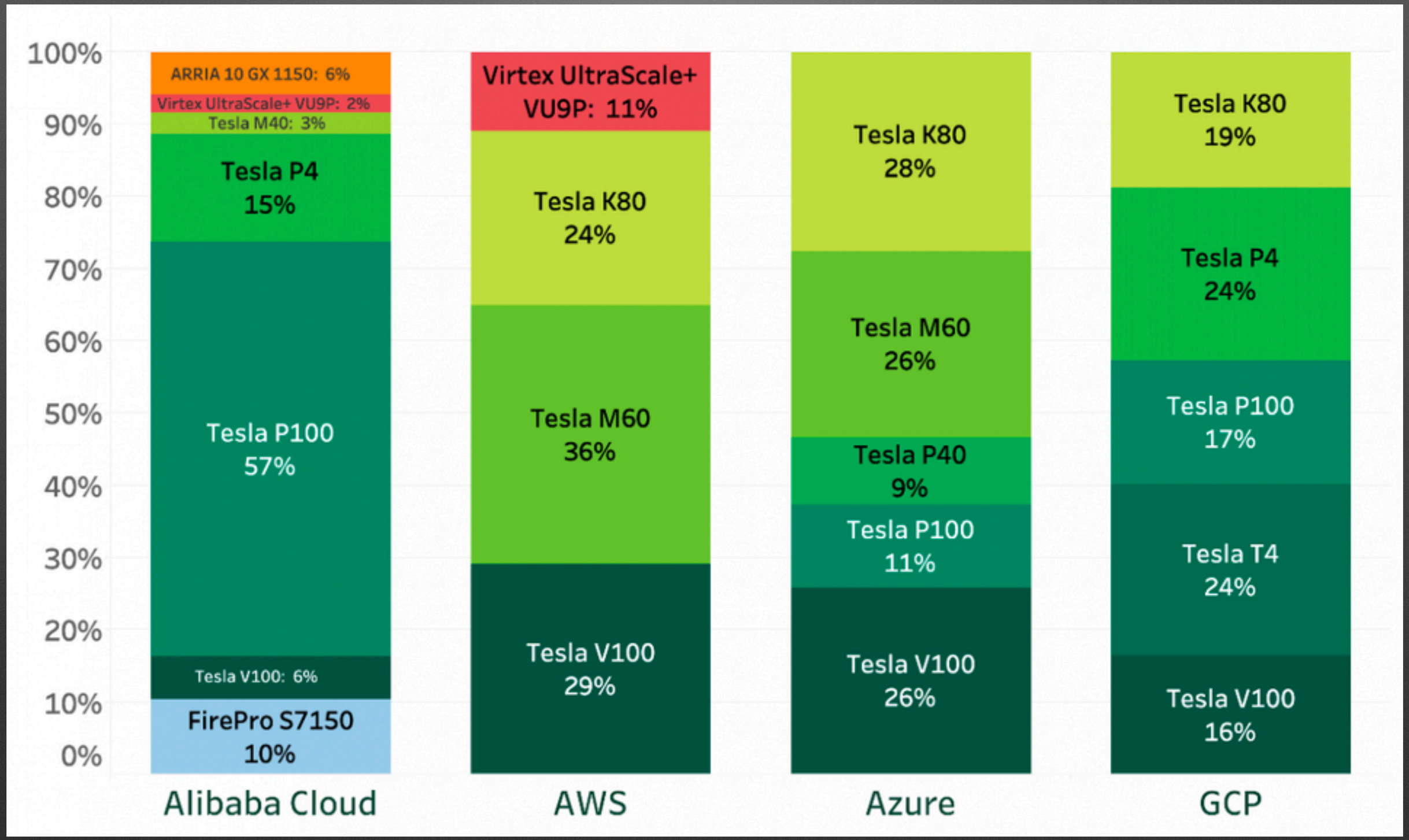
Catapult servers (MS)



Cygnus (U.Tsukuba)



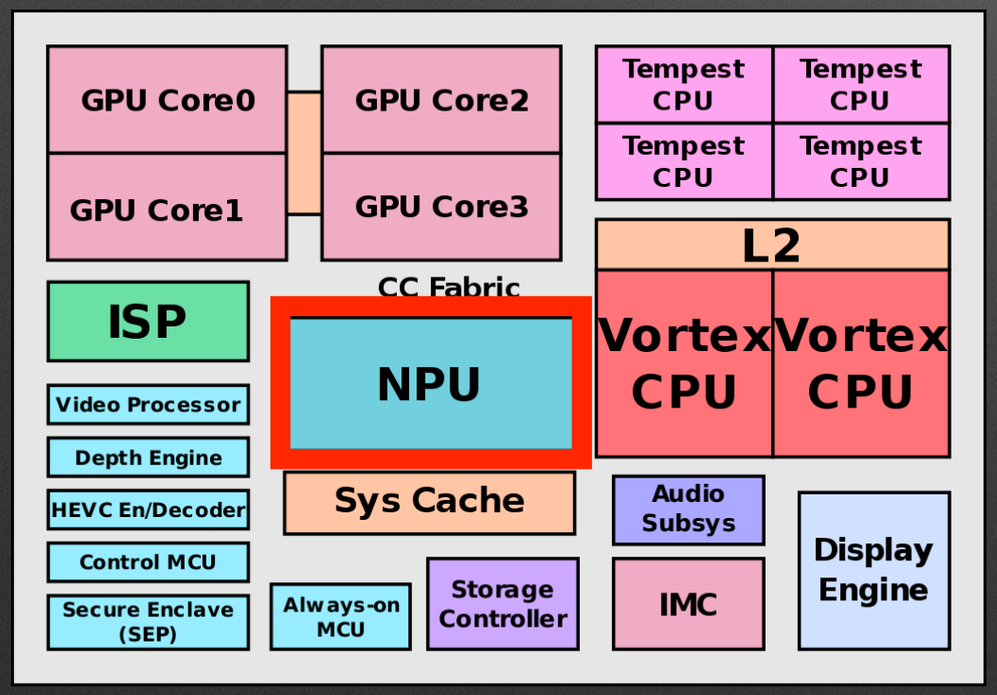
AI accelerators in Clouds

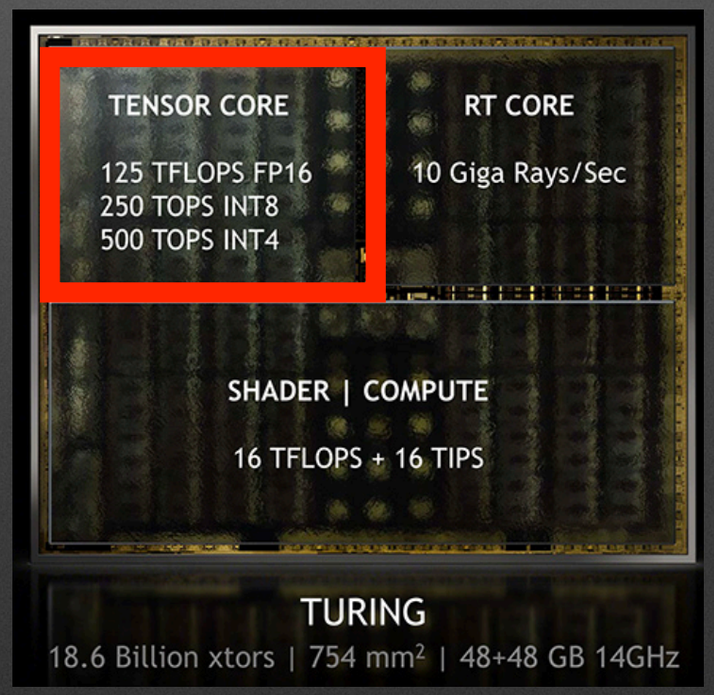


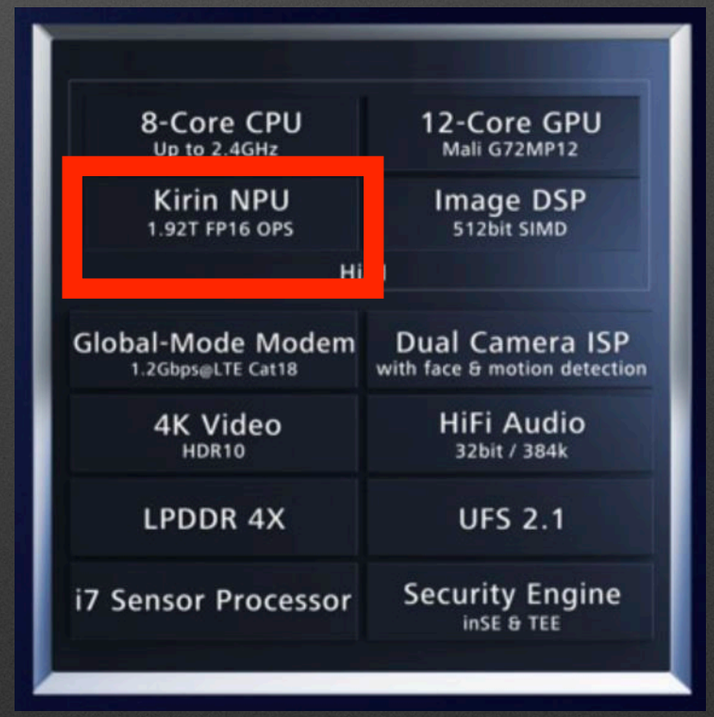
https://www.forbes.com/sites/paulteich/2019/06/17/nvidia-dominates-the-market-for-cloud-ai-accelerators-more-than-you-think/#138ac05a5edb

Deep Learning and CNN

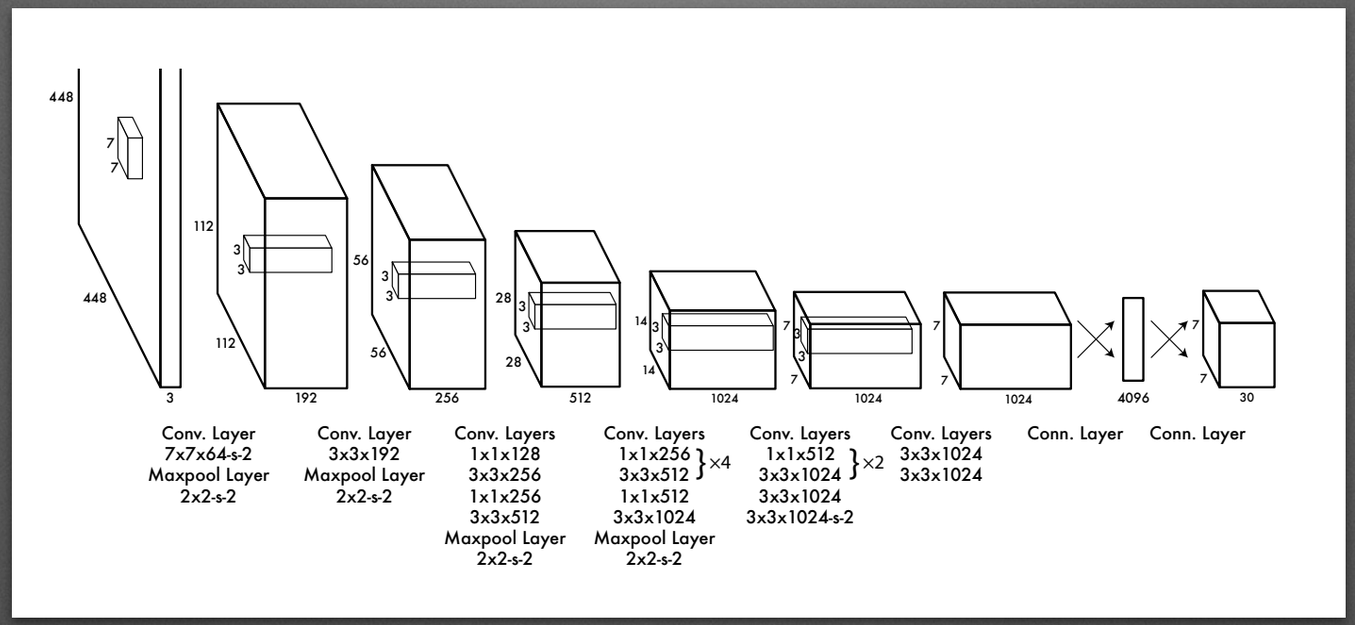
- Effective in Image classification and recognition, Image generation, Speech recognition and synthesis, Game of Go etc.
- Accelerated by GPU, FPGA and NPU

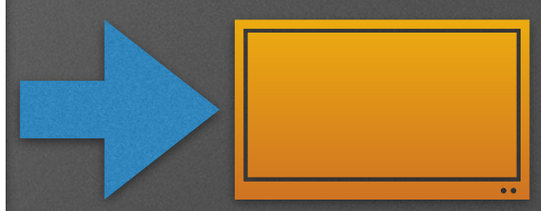






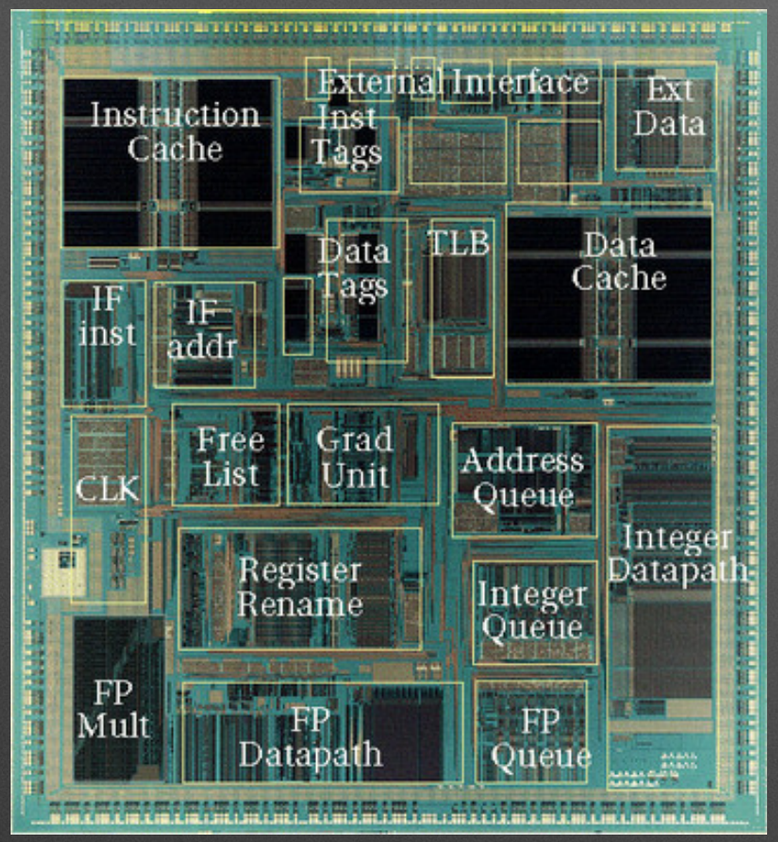
Object Detection Example





In a nutshell : arithmetic units

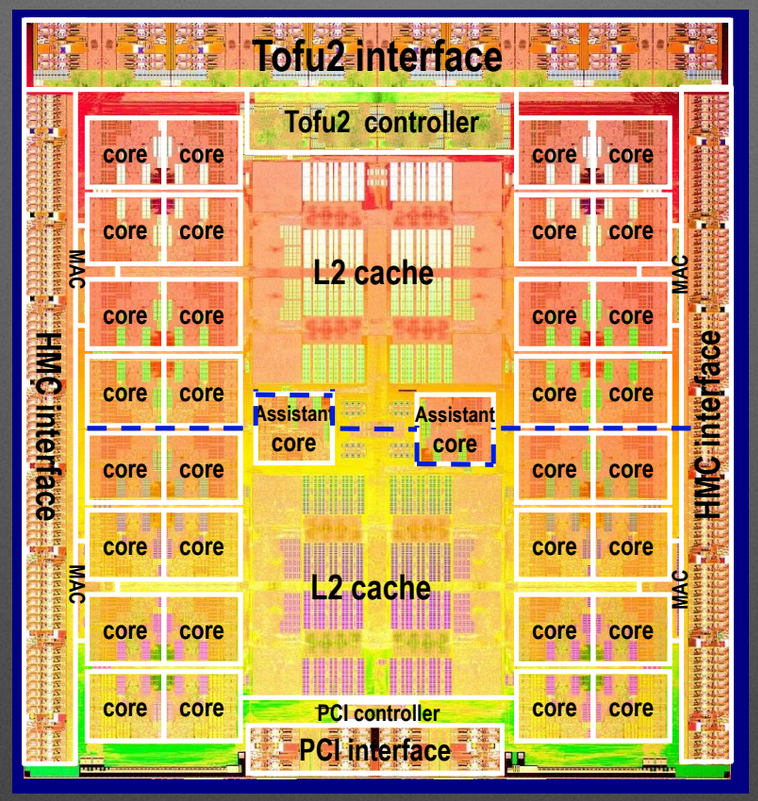
MIPS R10000 (1995)



http://cpudb.stanford.edu/processors/1400

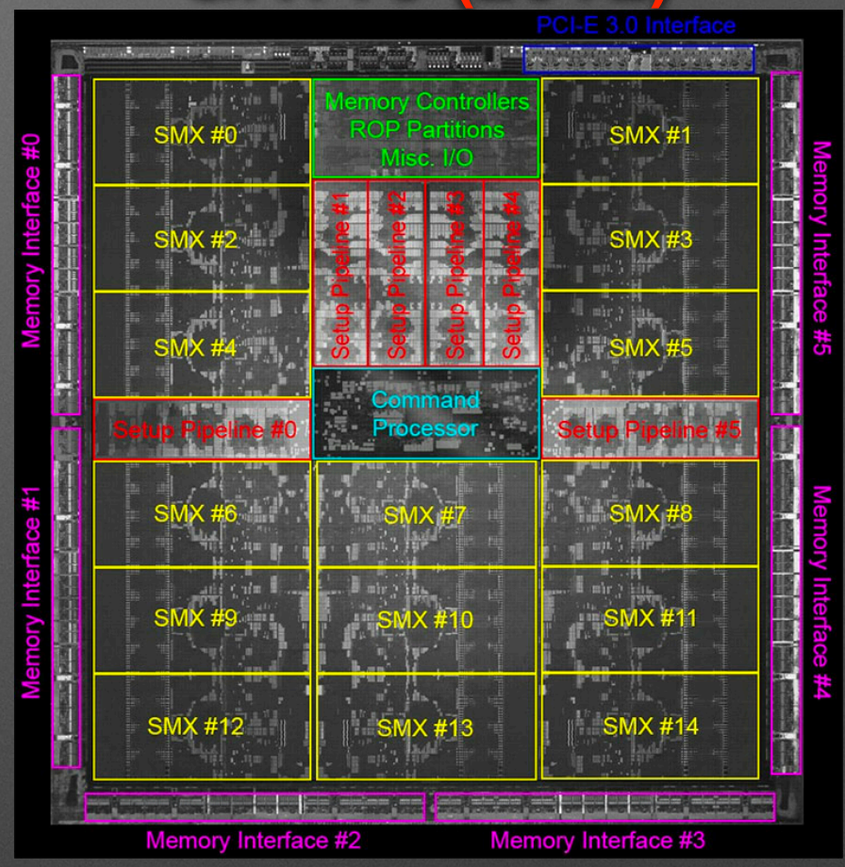
2 FP units

SPARC64 Xifx (2014)



544 FP units

GK110 (2012)

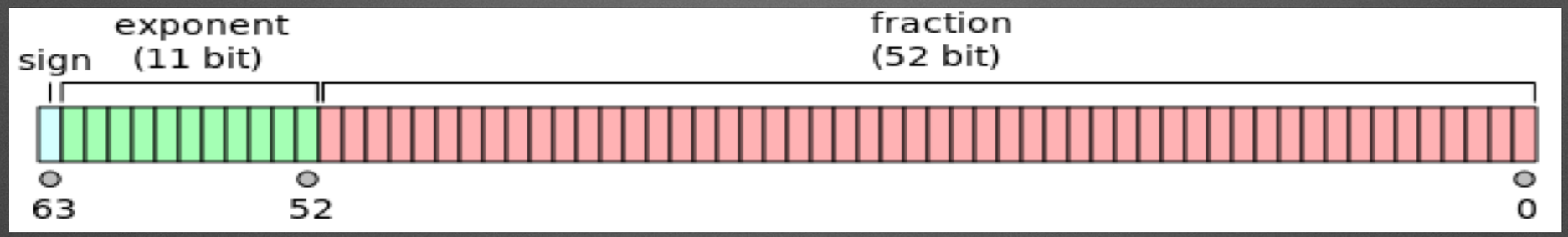


960 FP units

HPC processors and GPU/NPU have so many arithmetic units; relatively less memory

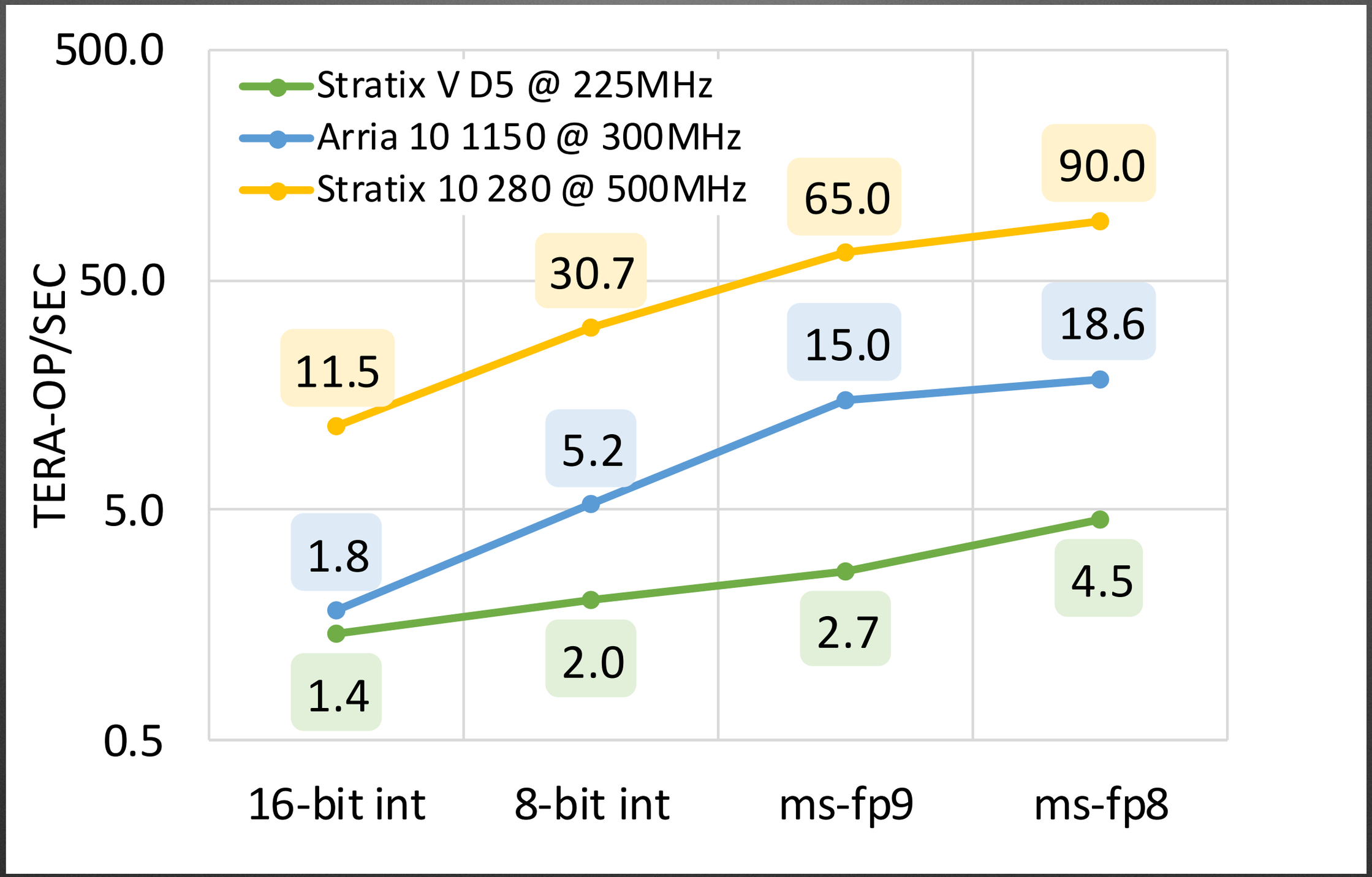
Encoding for arithmetic operations

- *A standard for FP is IEEE 754*
  - *16, 32, 64 and 128 bits*



- *For AI processors*
  - *Various exotic encoding is be proposed*
  - *half precision (16 bit)*
  - *fp8, fp9, fp10 etc… (MS, Google, Baidu)*
  - *Flexpoint (Intel)*
  - *Int8, Int9 etc. Even Int1 (binary) or Int2 (ternary)*

Comparison in various formats



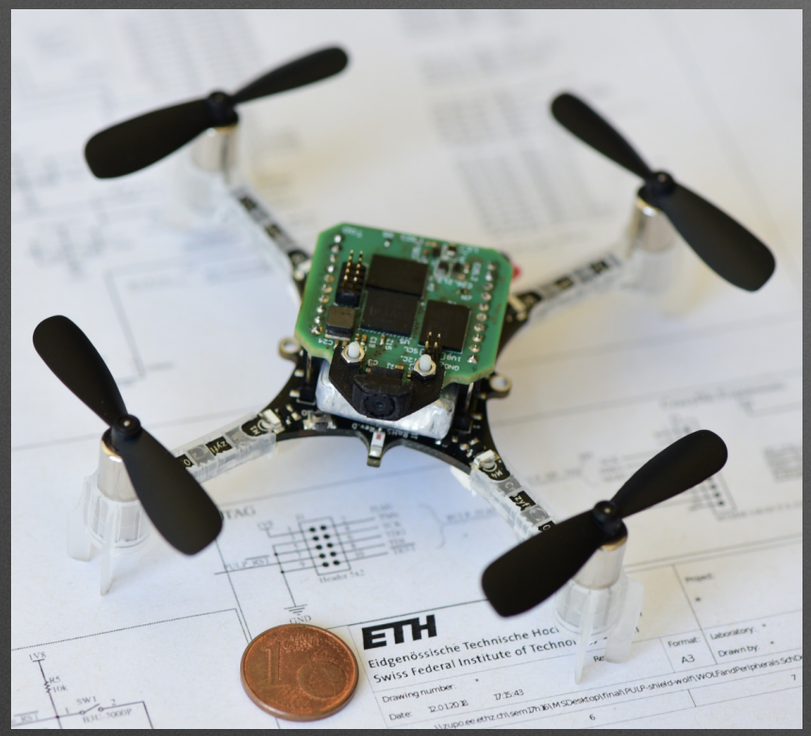
CPU vs. FPGA in ML

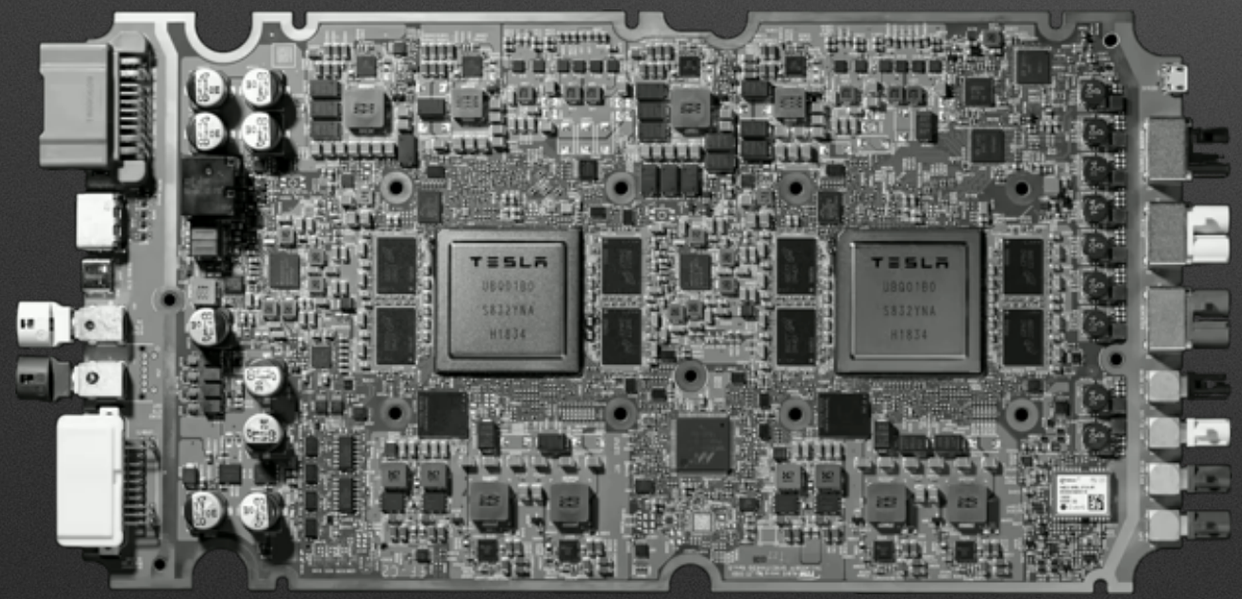
Table 1. Comparison of CPU-only vs. Brainwave-accelerated TP1 and DeepScan DNN models in Bing production.

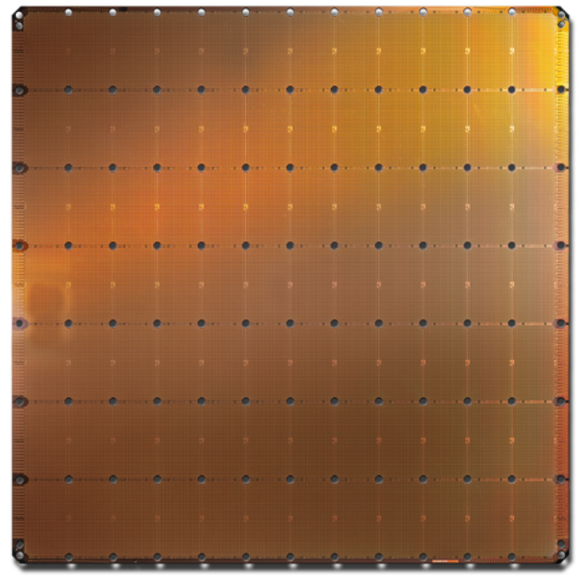
Bing TP1			
	CPU-only	**Brainwave-accelerated**	**Improvement**
Model details	GRU 128x200 (x2) + W2Vec	LSTM 500x200 (x8) + W2Vec	Brainwave-accelerated model is > 10X larger and > 10X lower latency
End-to-end latency per Batch 1 request at 95%	9 ms	0.850 ms	
Bing DeepScan			
	CPU-only	**Brainwave-accelerated**	**Improvement**
Model details	1D CNN + W2Vec *(RNNs removed)*	1D CNN + W2Vec + GRU 500x500 (x4)	Brainwave-accelerated model is > 10X larger and 3X lower latency
End-to-end latency per Batch 1 request at 95%	15 ms	5 ms	

https://www.microsoft.com/en-us/research/publication/serving-dnns-real-time-datacenter-scale-project-brainwave/

Domain Specific Architectures







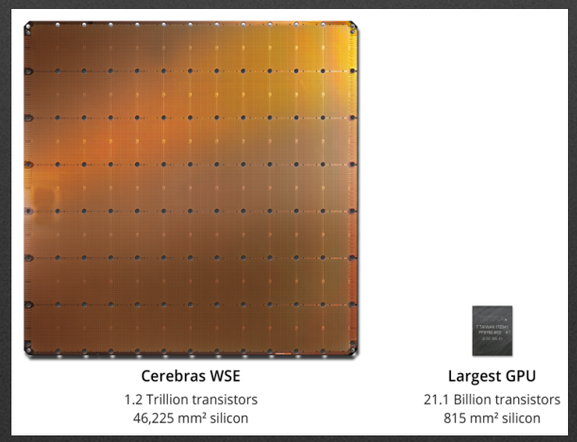
Cerebras WSE

1.2 Trillion transistors
46,225 mm² silicon

Super large AI processors

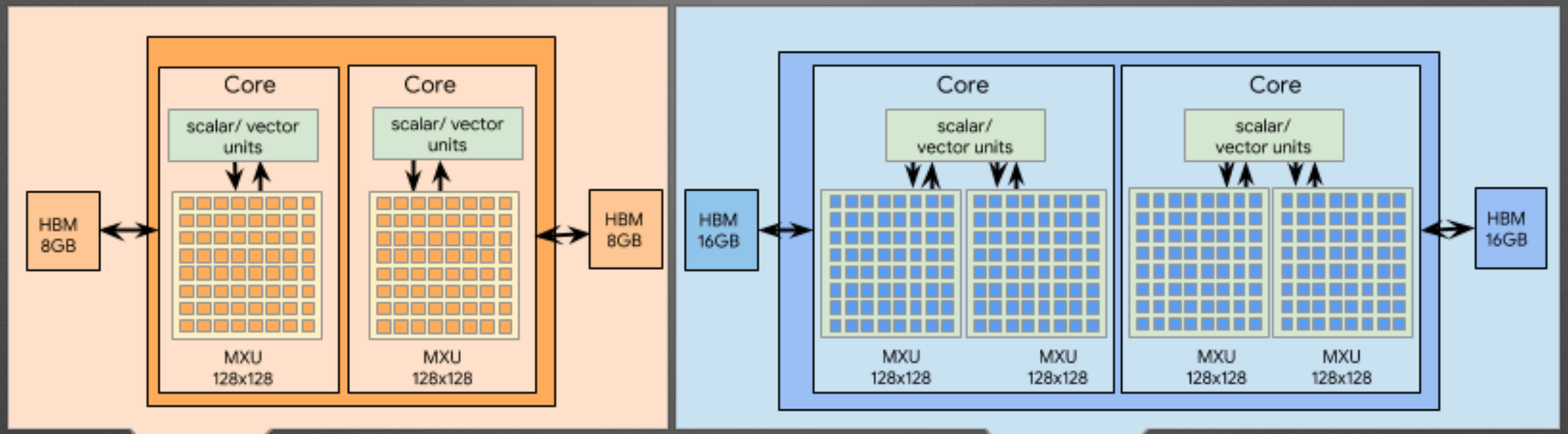
	Cerebras WSE	Largest GPU	Cerebras Advantage
Chip size	46,225 mm²	815 mm²	56.7 X
Cores	400,000	5,120	78 X
On chip memory	18 Gigabytes	6 Megabytes	3,000 X
Memory bandwidth	9 Petabytes/S	900 Gigabytes/S	10,000 X
Fabric bandwidth	100 Petabits/S	300 Gigabits/S	33,000 X

Figure 6: This summary table provides an overview of the magnitude of advancement made by the WSE

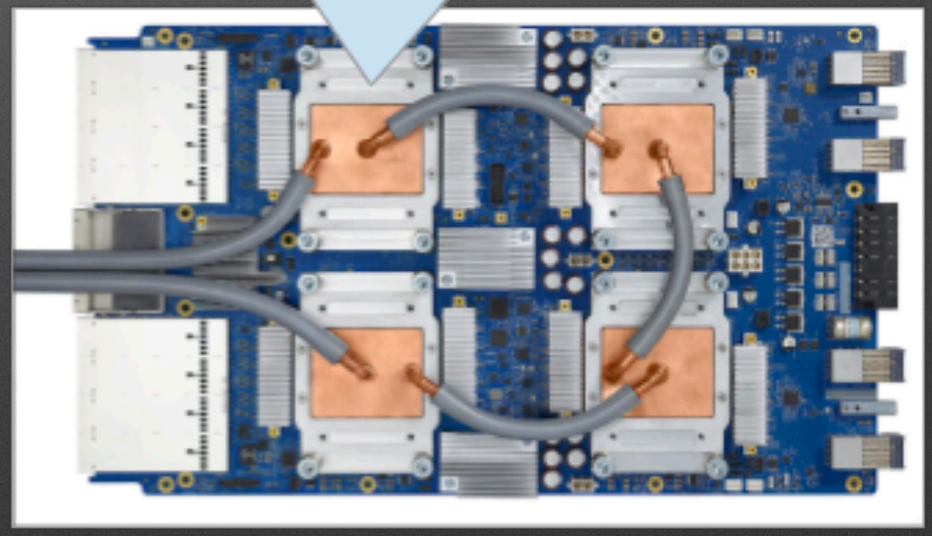


https://www.cerebras.net/introducing-the-cerebras-cs-1-the-industrys-fastest-artificial-intelligence-computer/

Google TPU

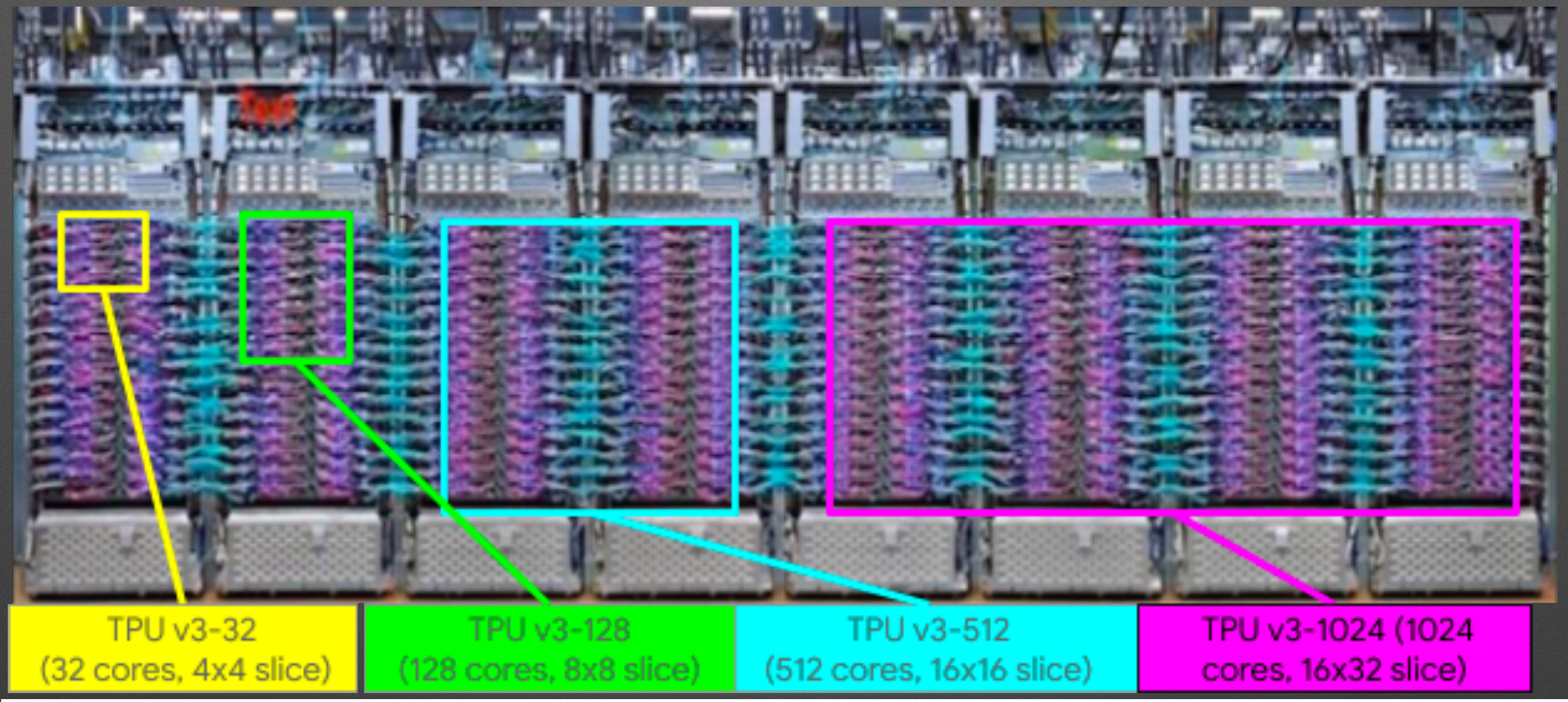


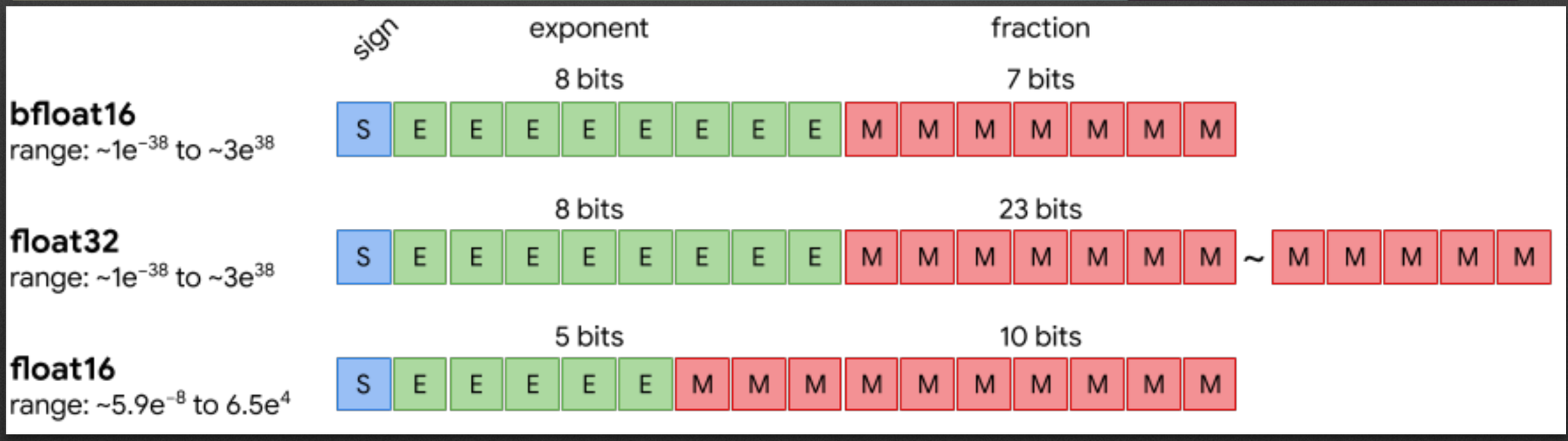
TPU v2 - 4 chips, 2 cores per chip



TPU v3 - 4 chips, 2 cores per chip

Google TPUv3 Pod





Graphcore IPU

IPU-Tiles™

1216 IPU-Tiles™ each with an independent IPU-Core™ and tightly coupled In-Processor-Memory™

IPU-Core™

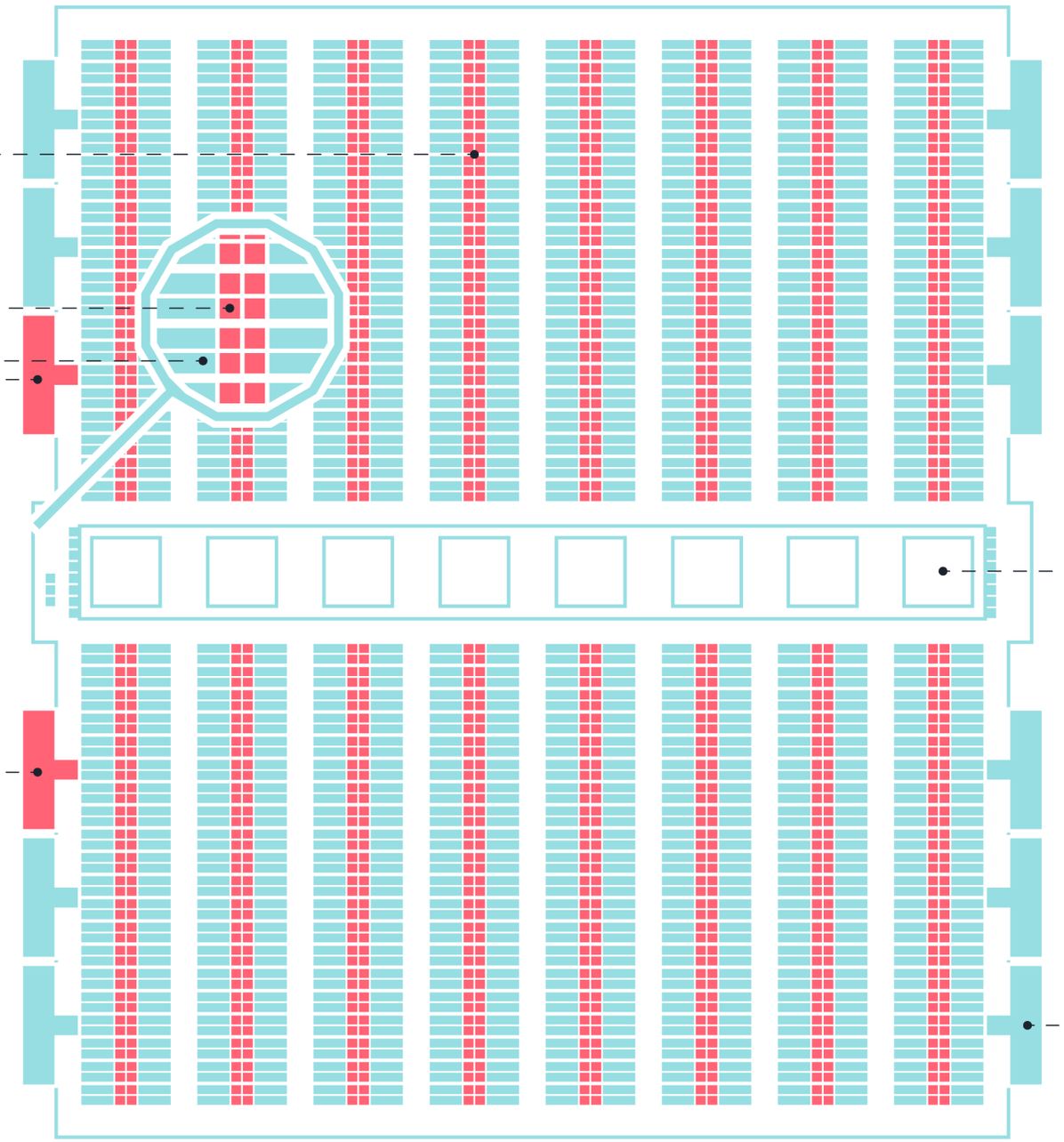
1216 IPU-Cores™ with 7296 programs executing in parallel

In-Processor-Memory™

300MB In-Processor-Memory™
45TB/s memory bandwidth
Whole model held on-chip

PCIe

PCIe Gen4 x16
64 GB/s bidirectional bandwidth to host



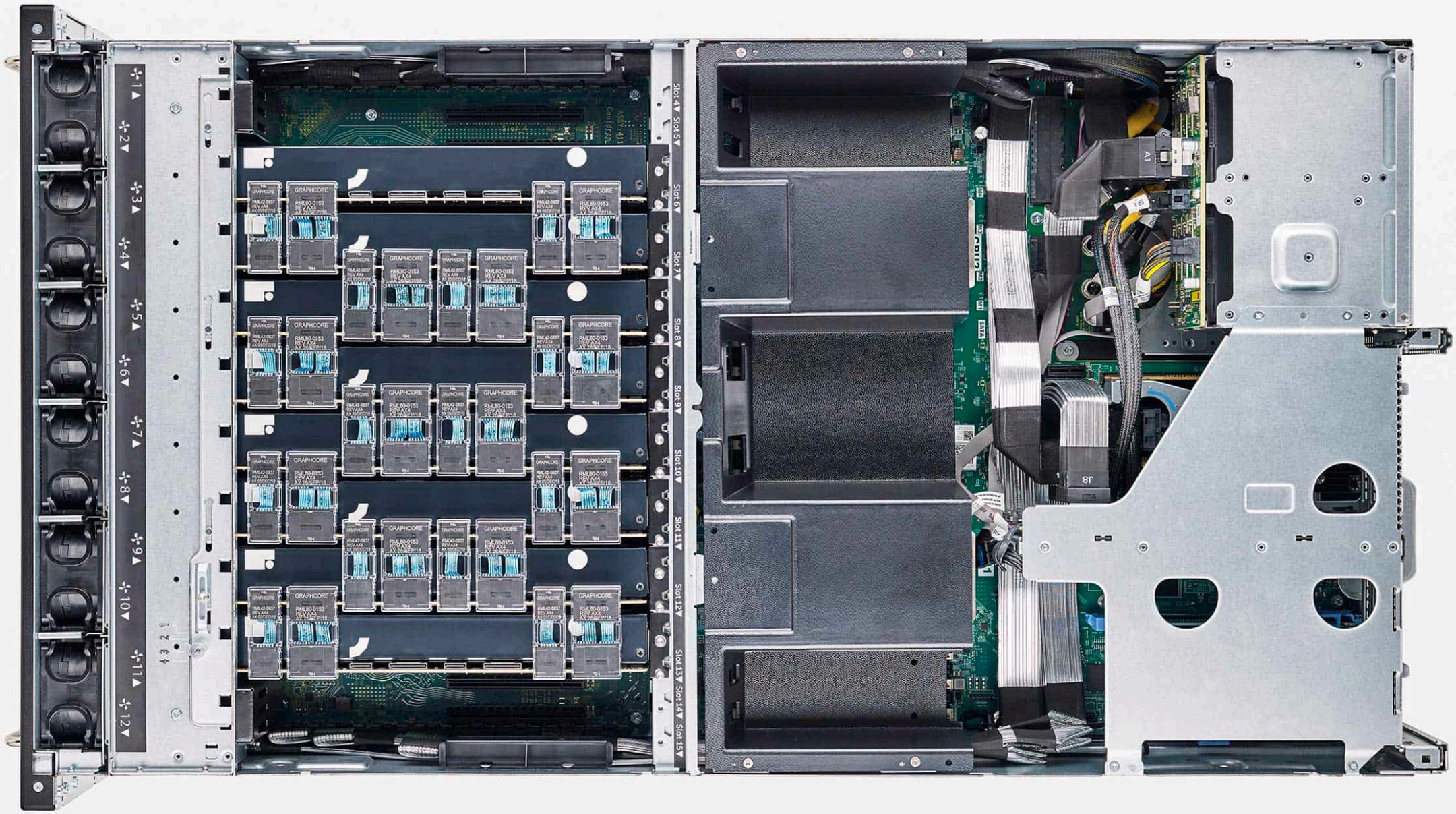
IPU-Exchange™

8 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern

IPU-Links™

80 IPU-Links, 320GB/s chip to chip bandwidth

IPU server


GRAPHCORE
Slot 4
Slot 5
Slot 6
Slot 7
Slot 8
Slot 9
Slot 10
Slot 11
Slot 12
Slot 13
Slot 14
Slot 15
J8