IIT Kharagpur | IIT Madras | IIT Goa | IIT Palakkad

Applied Accelerated AI

Introduction to AI Systems Hardware

Dr. Satyajit Das
Assistant Professor
Dept. of Data Science
Dept. of Computer Science and Engineering
IIT Palakkad

National Supercomputing Mission

Centre for Development of Advanced Computing

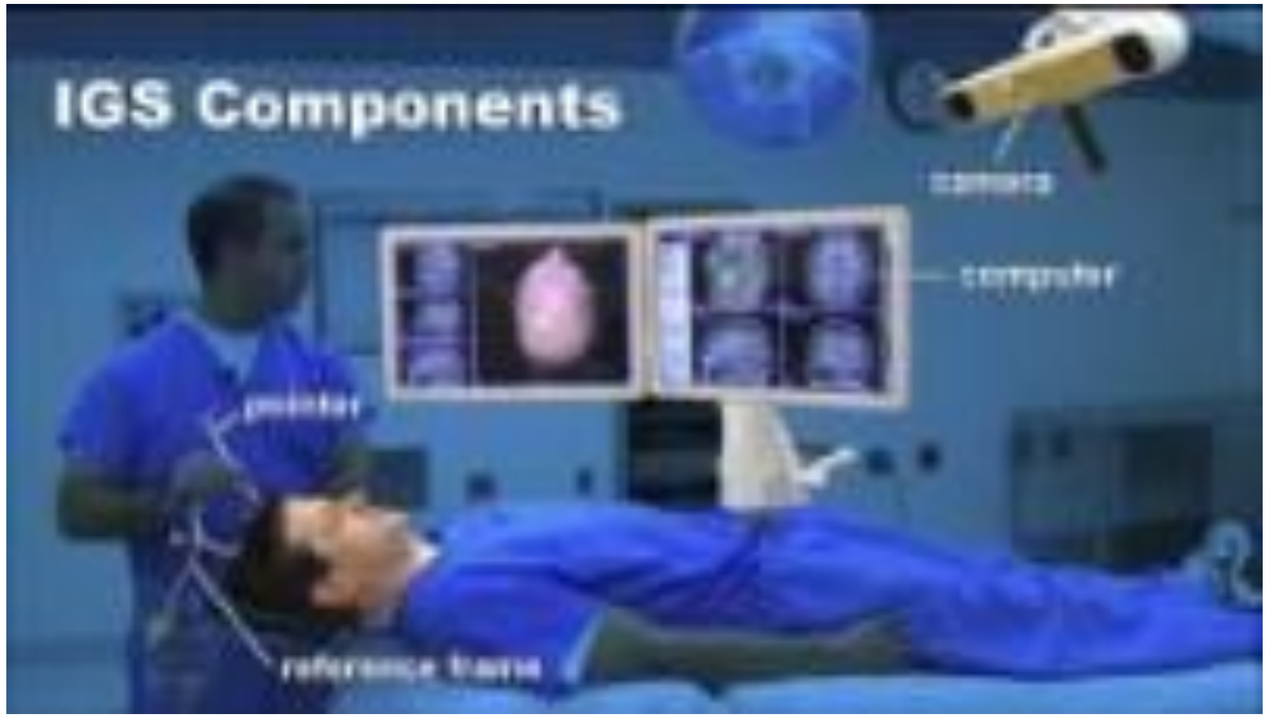
Contents

- Introduction to AI
- Introduction to Computing Systems
- Specialized Computation Engines

Contents

- Introduction to AI
- Introduction to Computing Systems
- Specialized Computation Engines


INTRODUCING
amazon go

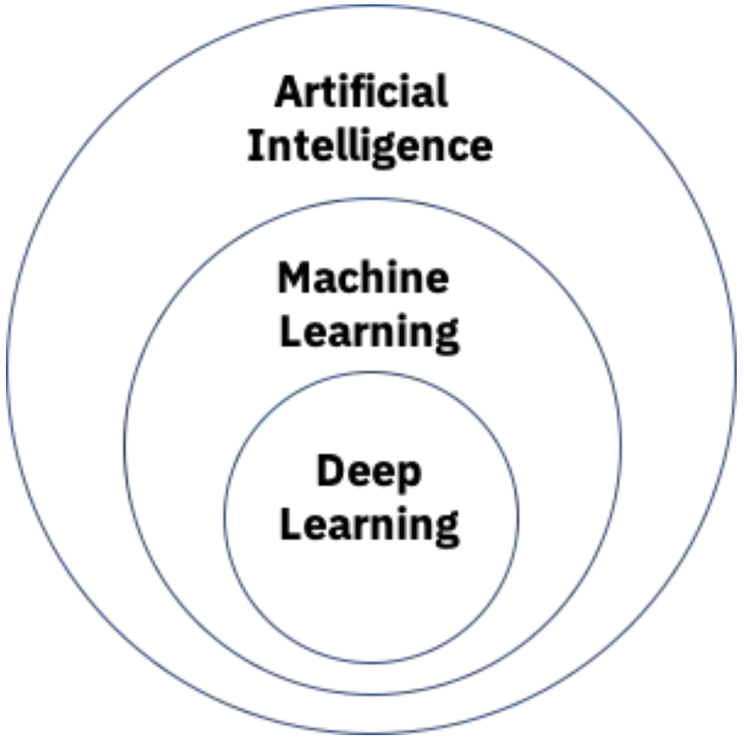

IGS Components
camera
computer
pointer
reference frame

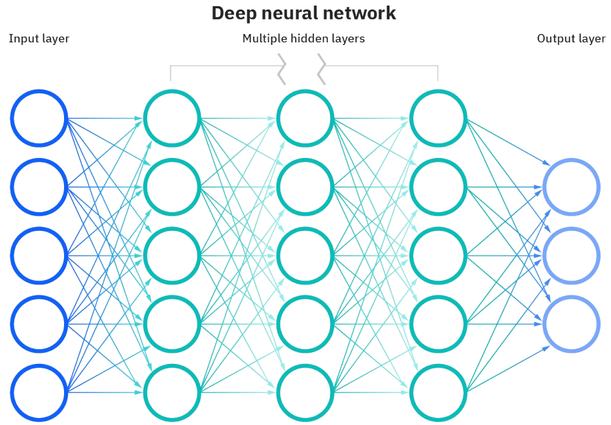
What is artificial intelligence?

- “…It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.” - John McCarthy

Deep learning vs. machine learning

- Deep learning automates much of the feature extraction piece of the process
- Classical machine learning more dependent on human intervention to learn



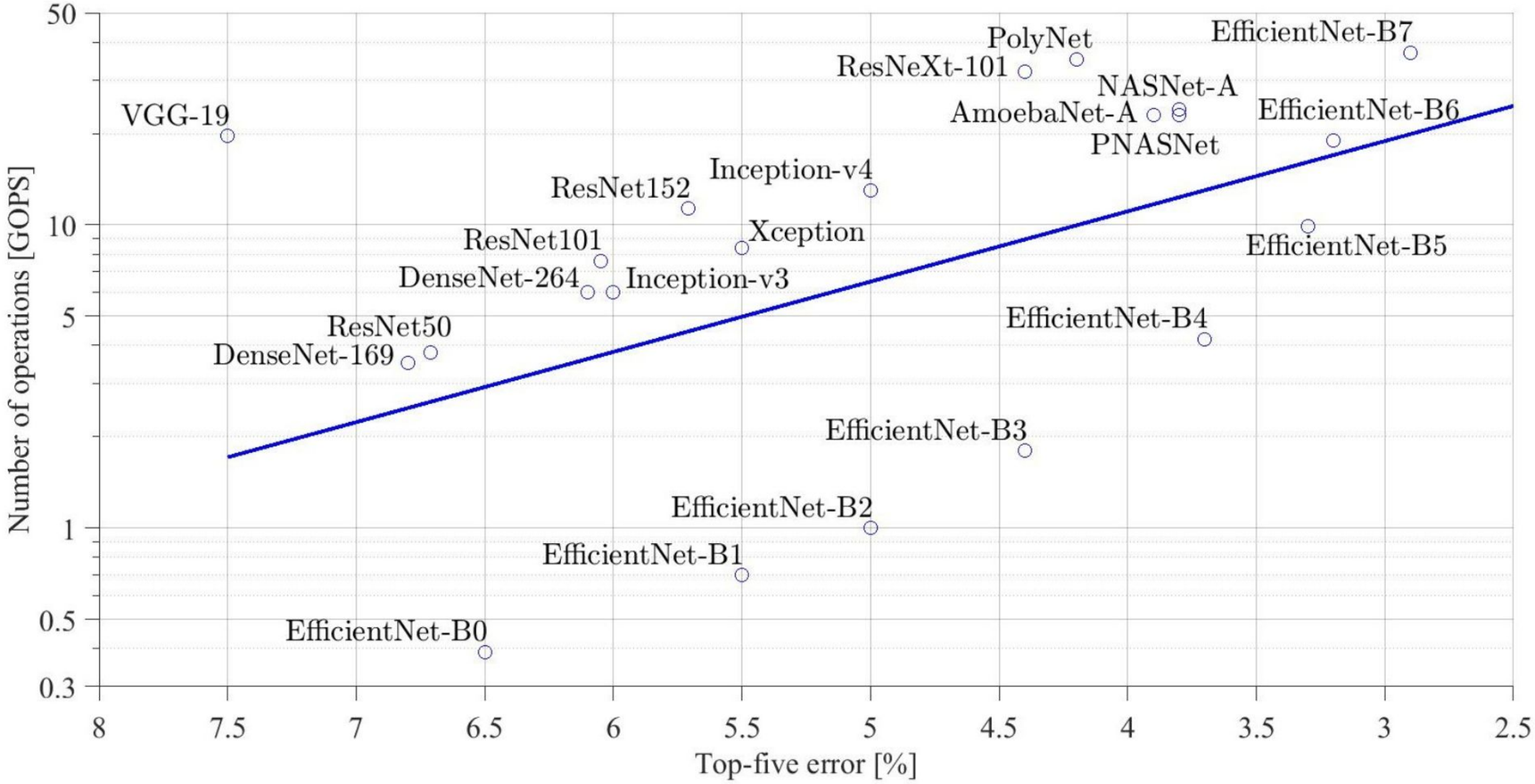


[Courtesy: IBM]

Factors for Advancement of AI

- Algorithmic innovation
- Data (which can be either supervised data or interactive environments)
- Amount of compute available for training

Computation Need for AI Algorithms

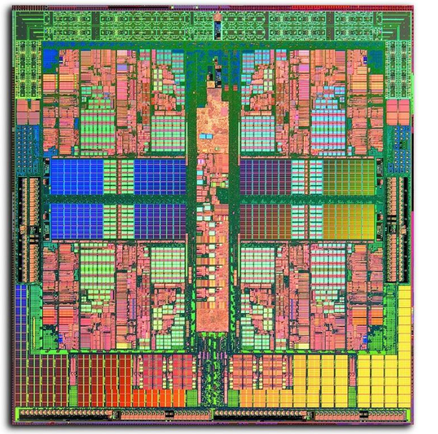


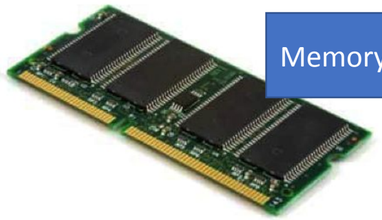
Baischer, L., Wess, M. and TaheriNejad, N., 2021. Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. *arXiv preprint arXiv:2104.09252*.

Contents

Traditional Computing Systems

- Computation
- Communication
- Storage/memory





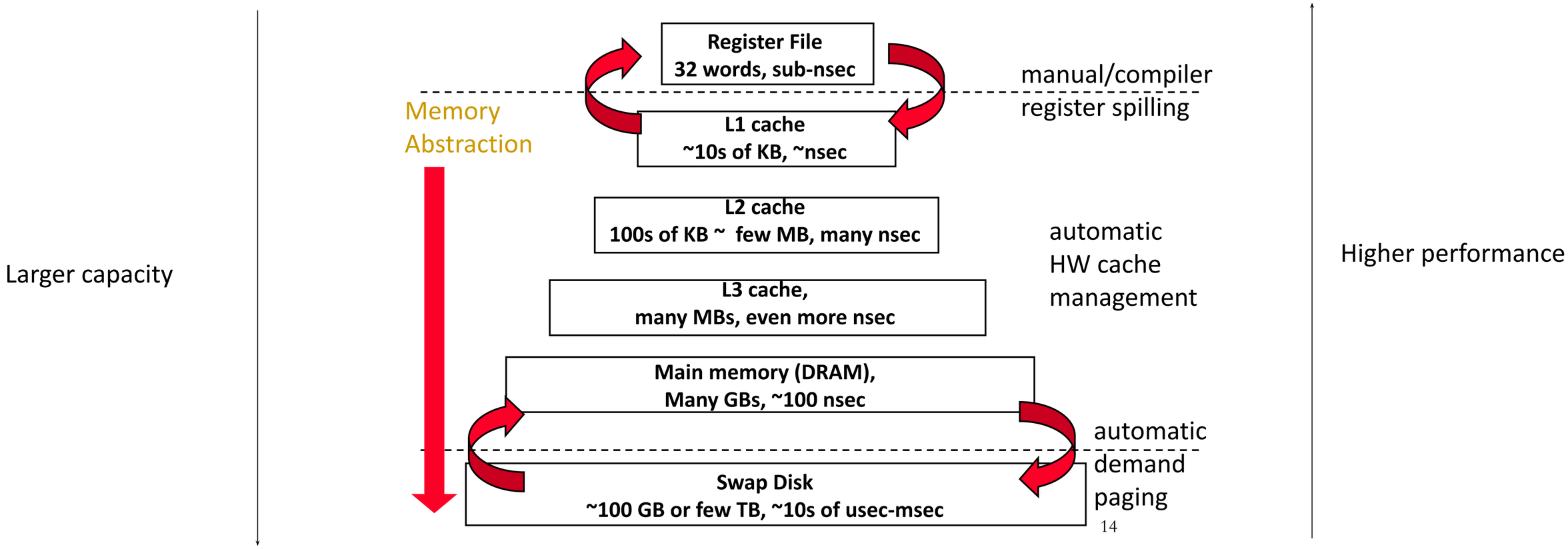
Memory system

Communication Unit

Memory

Storage system

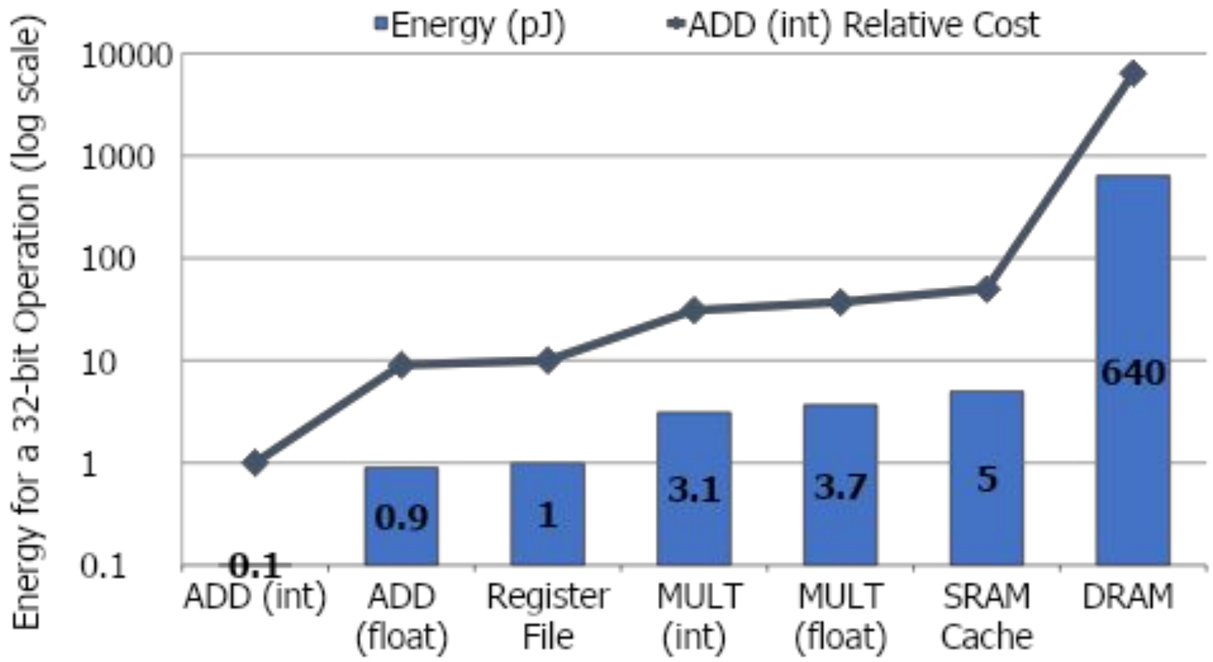
Modern Memory Hierarchy



Credit: Onur Mutlu

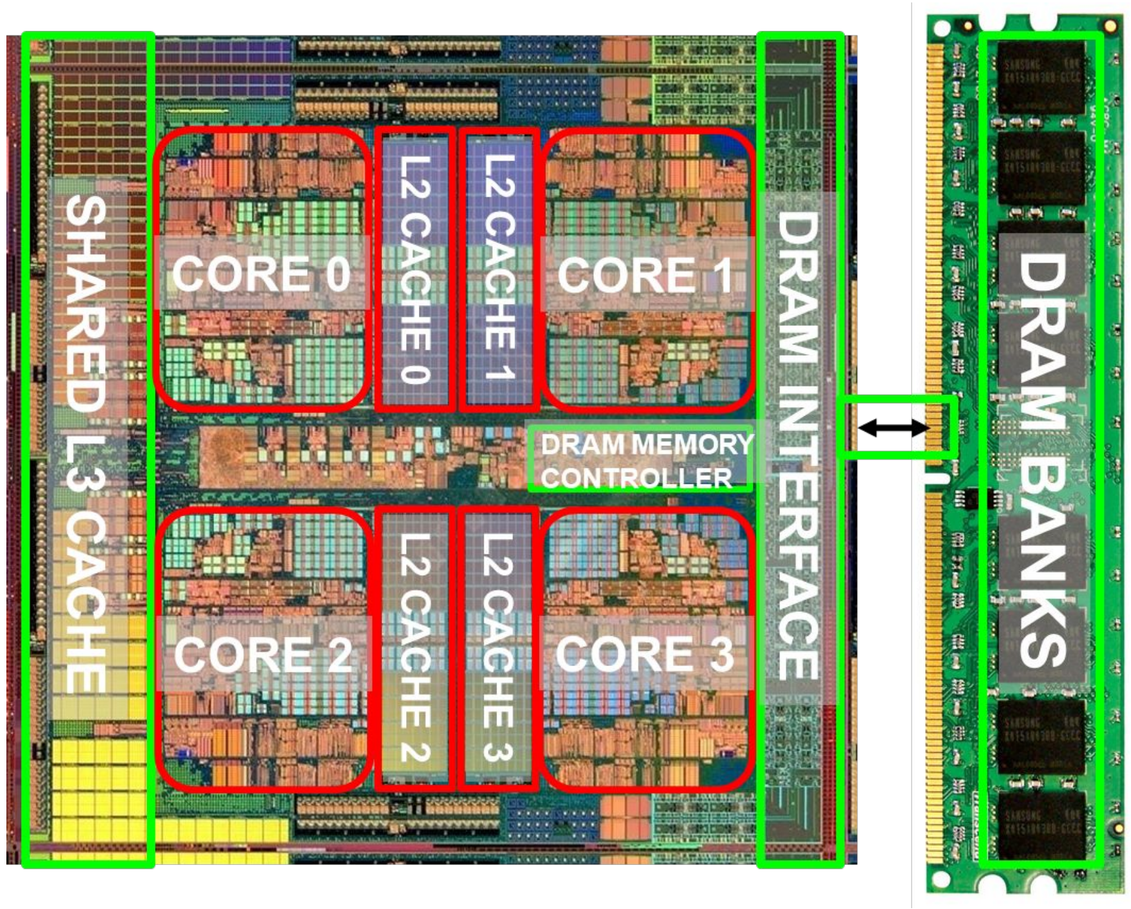
Key trends

- Data access is a major bottleneck
  - AI algorithms are extremely data hungry
- Energy consumption is a key limiter
  - Data movement energy dominates compute
  - Especially true for off-chip to on-chip movement



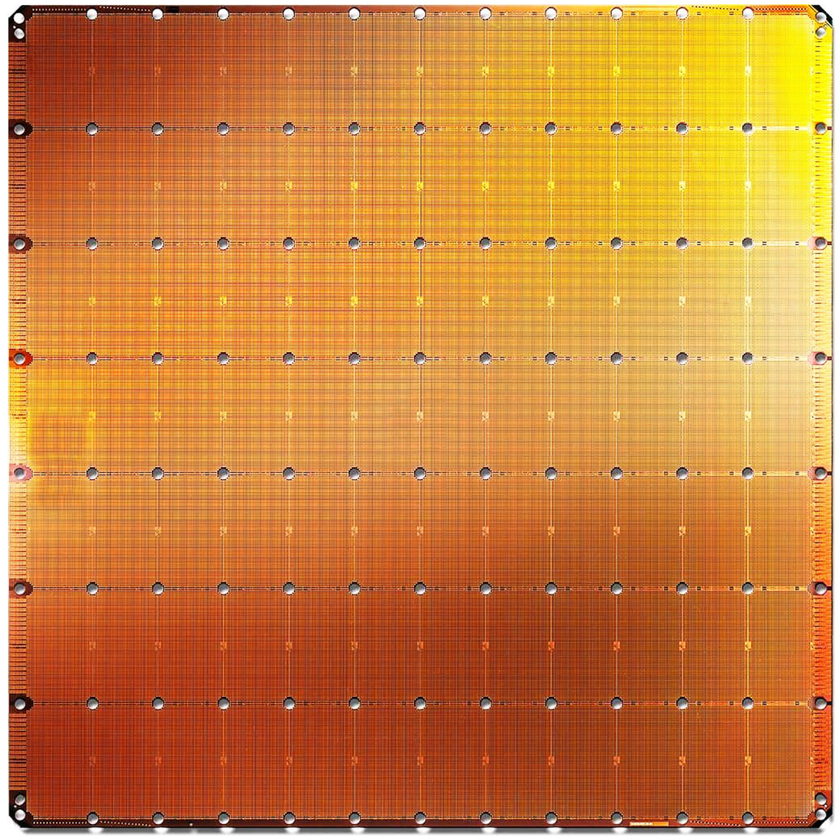
Han+, "EIE: Efficient Inference Engine on Compressed Deep Neural Network," ISCA 2016.

Modern Memory Systems



Credit: Onur Mutlu

Cerebras's Wafer Scale Engine (2019)



- The largest ML accelerator chip
- 400,000 cores
- **18 GB of on-chip memory**
- **9 PB/s memory bandwidth**

Cerebras WSE
1.2 Trillion transistors
46,225 mm^2

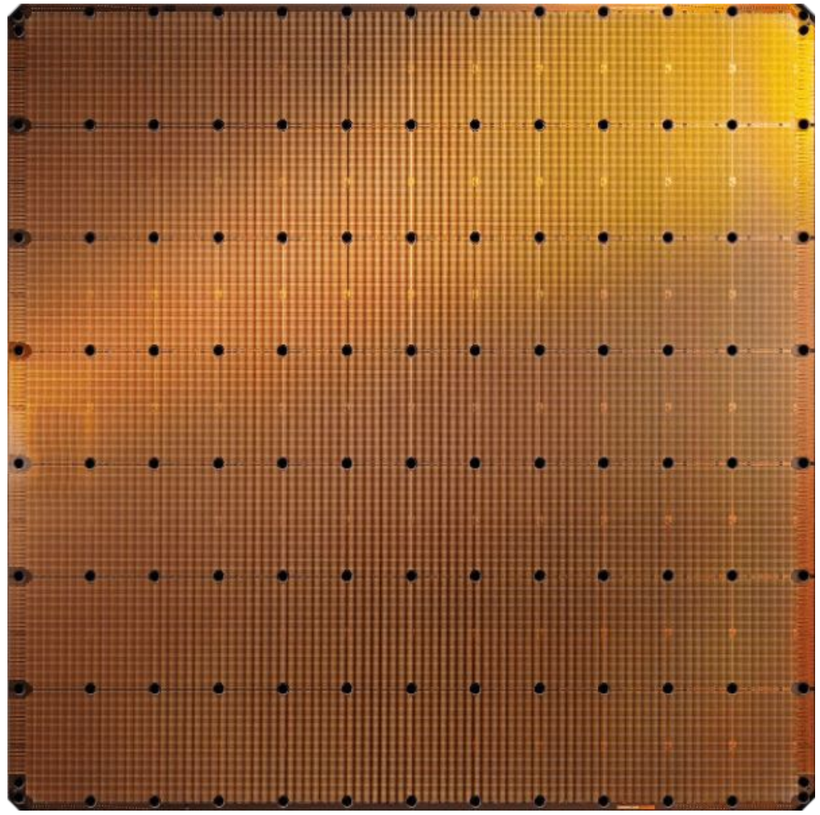
Largest GPU
54.2 Billion transistors
826 mm^2
NVIDIA Ampere GA100

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning

https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

Credit: Onur Mutlu

Cerebras's Wafer Scale Engine-2 (2021)



- The largest ML accelerator chip
- 850,000 cores
- **40 GB of on-chip memory**
- **20 PB/s memory bandwidth**

Cerebras WSE-2
2.6 Trillion transistors
46,225 mm^2

https://cerebras.net/product/#overview

Largest GPU
54.2 Billion transistors
826 mm^2
NVIDIA Ampere GA100

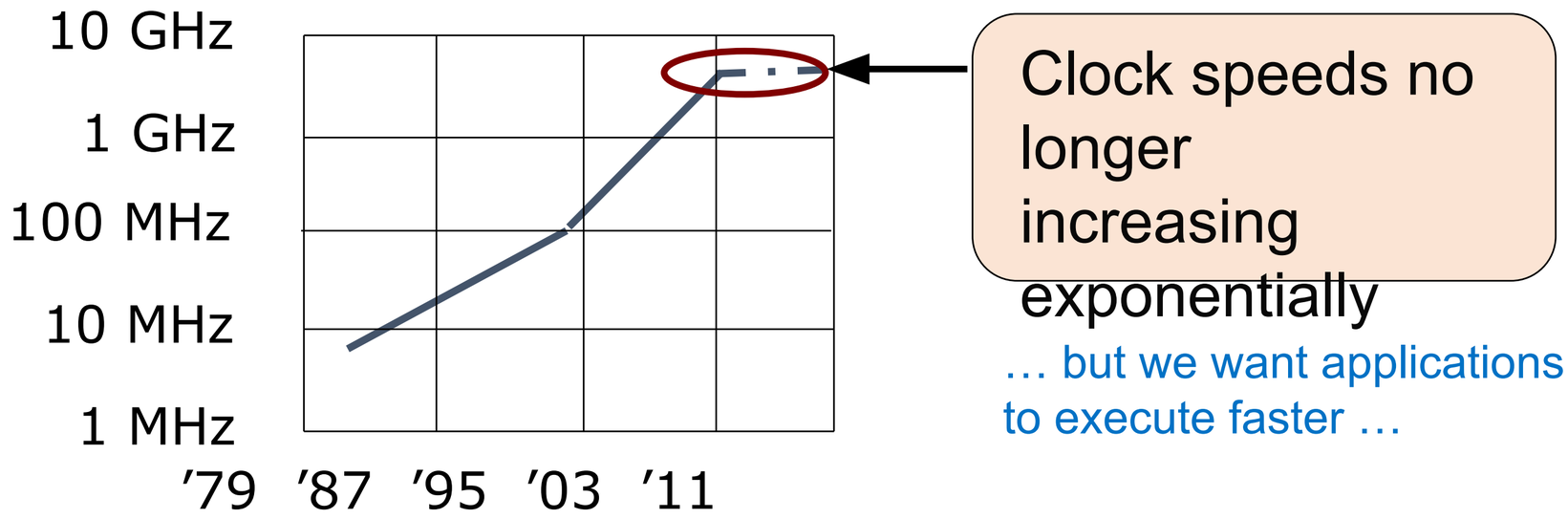
Credit: Onur Mutlu

Contents

- Introduction to AI
- Introduction to Computing Systems
- Specialized Computation Engines

Computation speed hitting wall

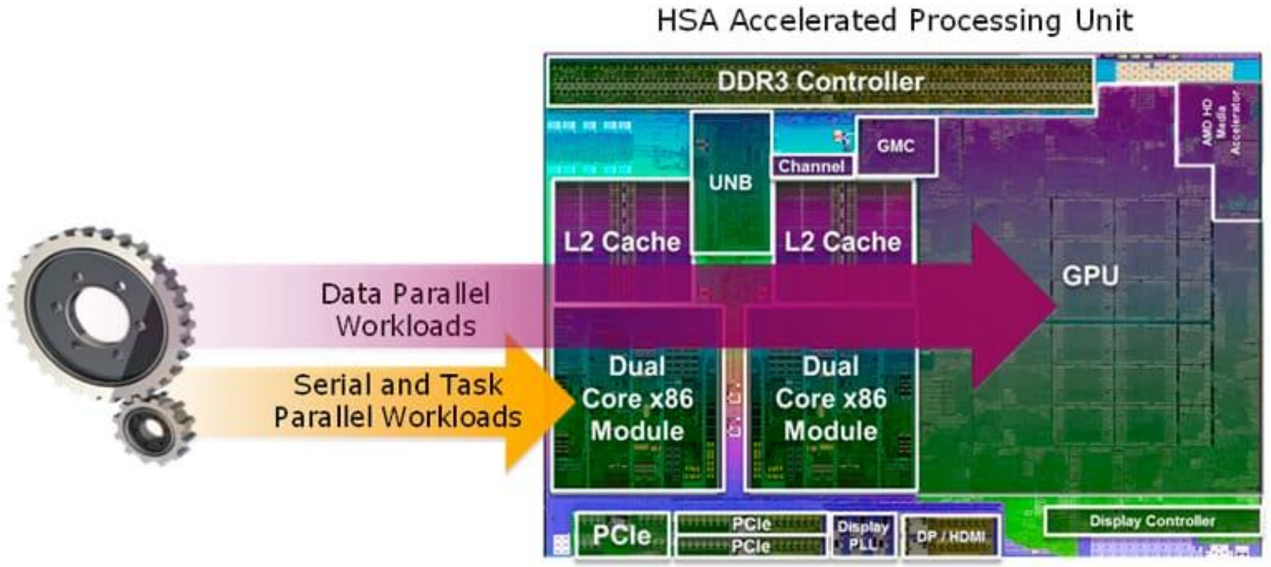
- “The free lunch is over: A Fundamental Turn Toward Concurrency” – Herb Sutter, Dr. Dobb’s Journal, March 2005



[Courtesy: Intel]

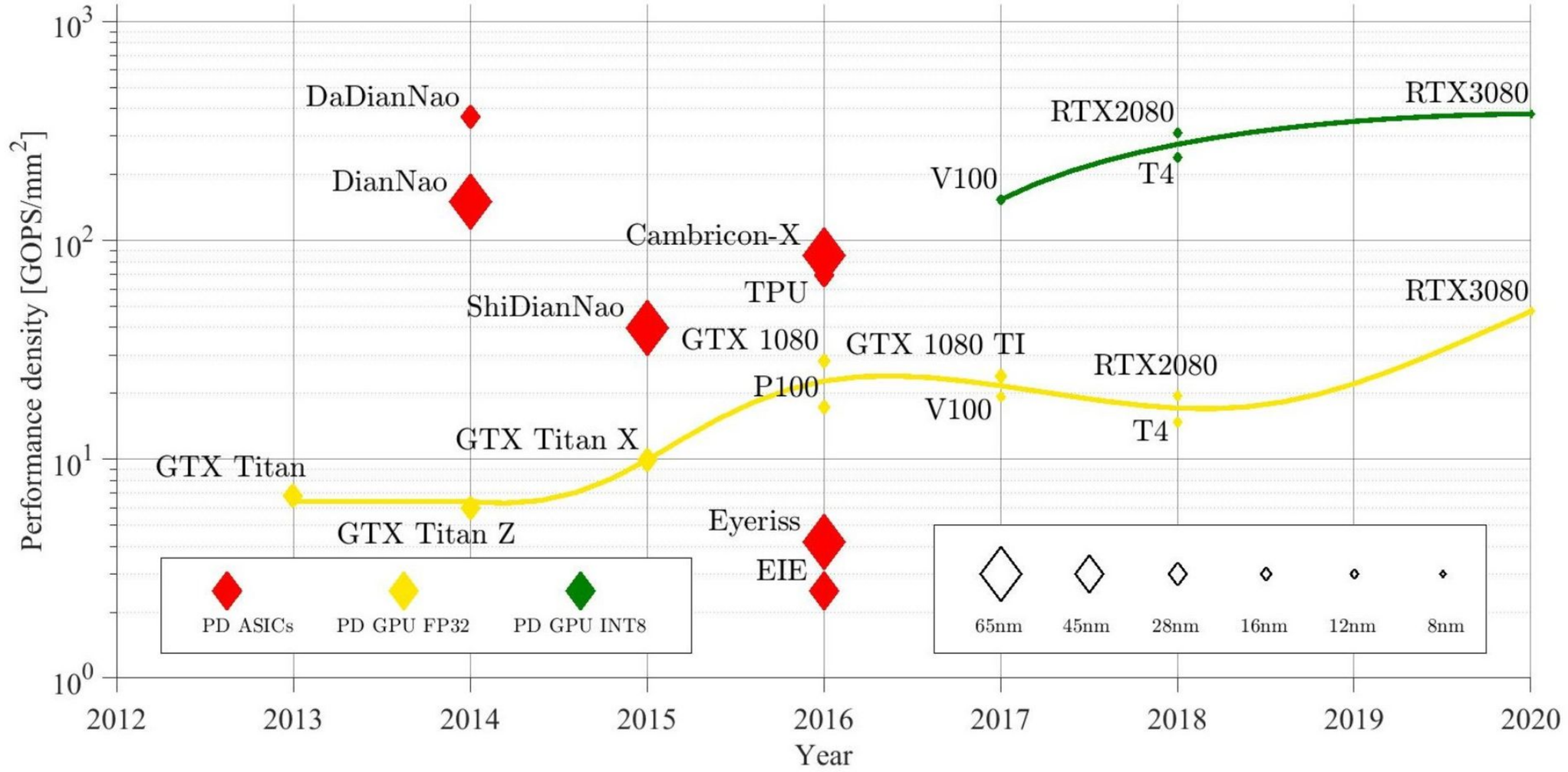
How can we accomplish high computational density

- Through parallel-computing…
- Attempt to speed solution of a particular task by
  - Dividing task into sub-tasks
  - Executing sub-tasks simultaneously on multiple processors and
  - Specialized tasks in **accelerators**



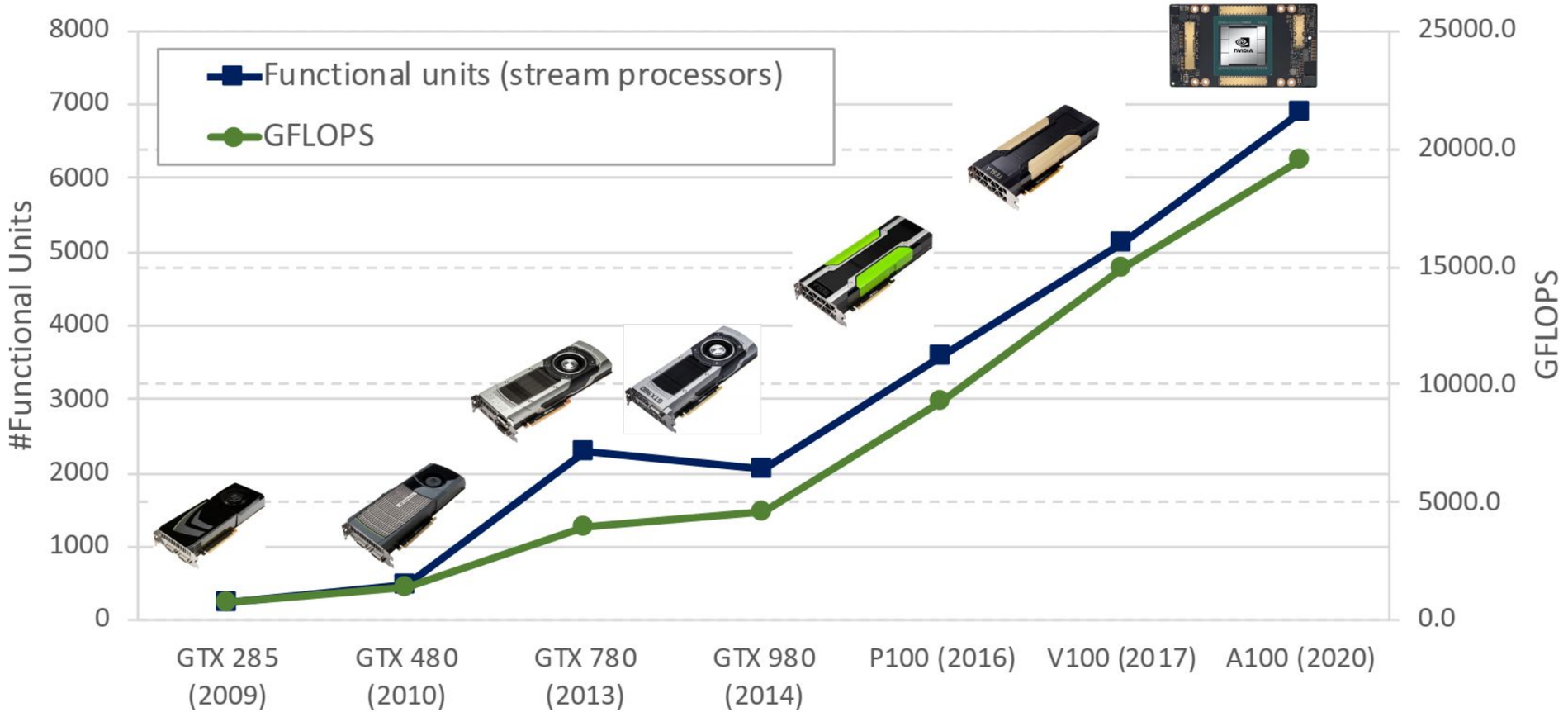
[Courtesy: AMD]

Specialized computation engines



Baischer, L., Wess, M. and TaheriNejad, N., 2021. Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. *arXiv preprint arXiv:2104.09252*.

Evolution of NVIDIA GPUs



NVIDIA V100 vs A100

- NVIDIA-terminology:
  - 5120 stream processors
  - “SIMT execution”
- Generic terminology:
  - 80 cores
  - 64 SIMD functional units per core
  - Tensor cores for Machine Learning

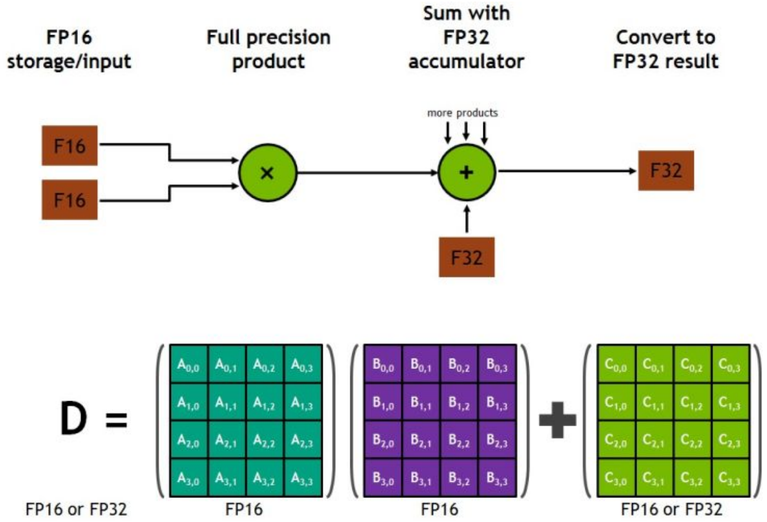
[NVIDIA, “NVIDIA Tesla V100 GPU Architecture. White Paper,” 2017]

https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

- NVIDIA-speak:
  - 6912 stream processors
  - “SIMT execution”
- Generic speak:
  - 108 cores
  - 64 SIMD functional units per core
  - Tensor cores for Machine Learning
    - Support for sparsity
    - New floating point data type (TF32)
- https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/

https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

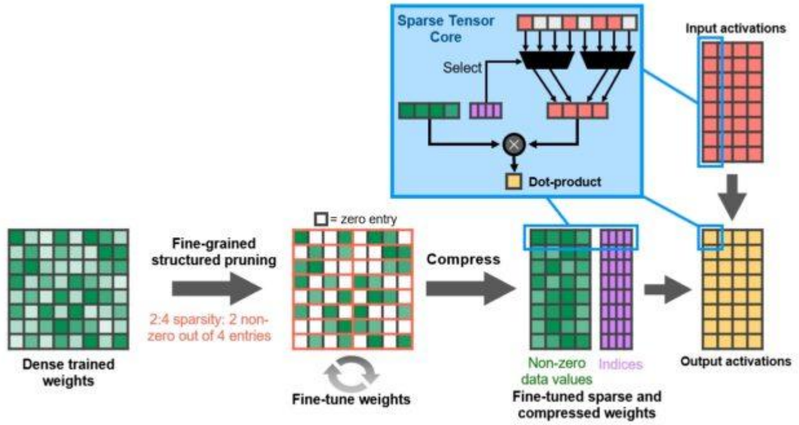
NVIDIA V100 core



- 15.7 TFLOPS Single Precision
- 7.8 TFLOPS Double Precision
- 125 TFLOPS for Deep Learning (Tensor cores)

Inside Volta: The World's Most Advanced Data Center GPU | NVIDIA Developer Blog

NVIDIA A100 Core



- 19.5 TFLOPS Single Precision
- 9.7 TFLOPS Double Precision
- 312 TFLOPS for Deep Learning (Tensor cores)

https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

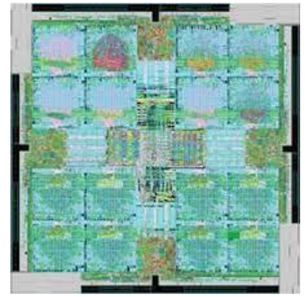
Overview of GPU based accelerators

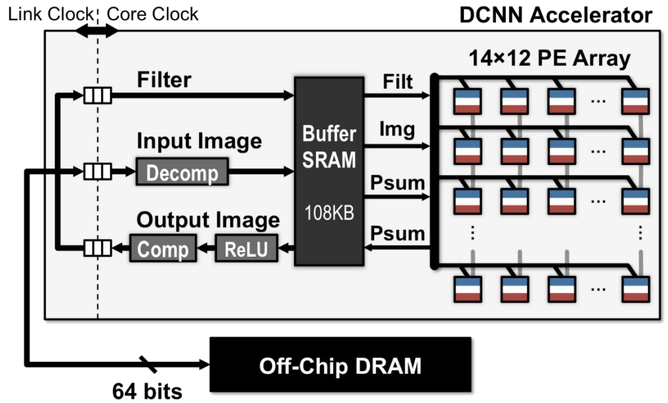
Name	Area [mm^2]	feature size [nm]	Quanti-zation	Bit width	Tensor unit	Throughput [TOPS][(a)]	Freq. [MHz]	Power [W]	[$\frac{GOPS}{mm^2}$][(b)]
V100[1] [37]	815	12	float	64		7.8	1530	300	9.57
			float	32		15.7	1530	300	19.26
			mixed	32-8	X	125	1530	300	153.37
T4[1] [38]	545	12	float	32		8.1	1590	70	14.81
			float	16	X	65	1590	70	119.26
			fixed	8	X	130	1590	70	238.53
			fixed	4	X	260	1590	70	477.06
RTX 2080[2] [38]	545	12	float	32		10.6	1710	225	19.45
			float	16	X	84.8	1710	225	155.6
			fixed	8	X	169.6	1710	225	311.19
			fixed	4	X	322.2	1710	225	591.2

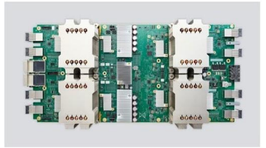
Baischer, L., Wess, M. and TaheriNejad, N., 2021. Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. *arXiv preprint arXiv:2104.09252*.

Overview of FPGA based accelerators

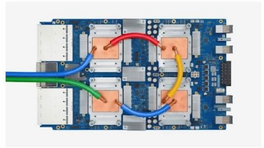
Name	Area [mm^2]	feature size [nm]	Quanti-zation	Bit width	Throughput [GOPS][(a)]	Frequency [MHz]	Power [mW]	[$\frac{GOPS}{mm^2}$][(b)]
DaDianNao [7]	4335*	28	fixed	16	1586288*	606	48380*	366*
EIE [18]	40.8	45	fixed	16	102	800	590	2.5
Cambricon-X [61]	6.38	65	fixed	16	544	1000	954	85.26
Eyeriss [8]	16	65	fixed	16	67.2**	200	278	4.2
TPU [25]	331***	28	fixed	8	92000	700	40000	69.48





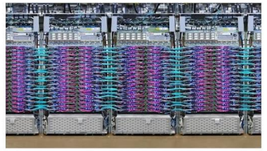


Cloud TPU v2
180 teraflops
64 GB High Bandwidth Memory (HBM)



Cloud TPU v3
420 teraflops
128 GB HBM

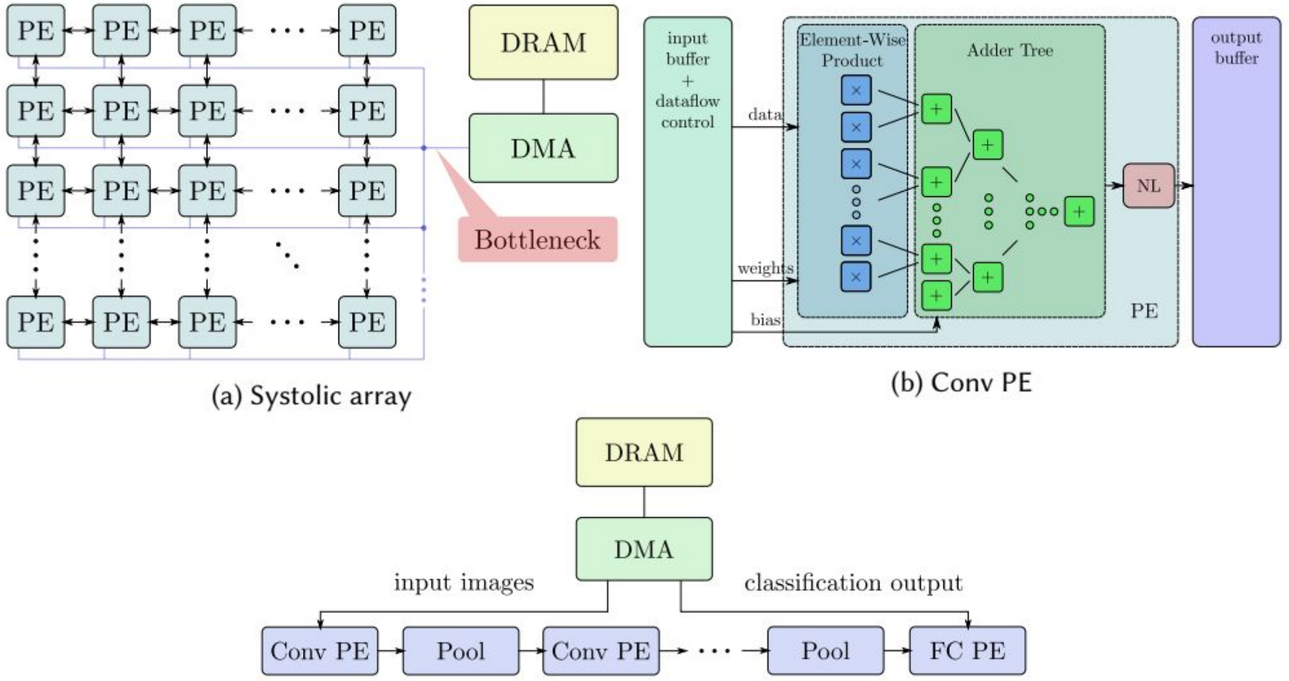
Cloud TPU v2 Pod
11.5 petaflops
4 TB HBM
2-D toroidal mesh network



Cloud TPU v3 Pod
100+ petaflops
32 TB HBM
2-D toroidal mesh network

Baischer, L., Wess, M. and TaheriNejad, N., 2021. Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. *arXiv preprint arXiv:2104.09252*.

Typical FPGA based accelerators



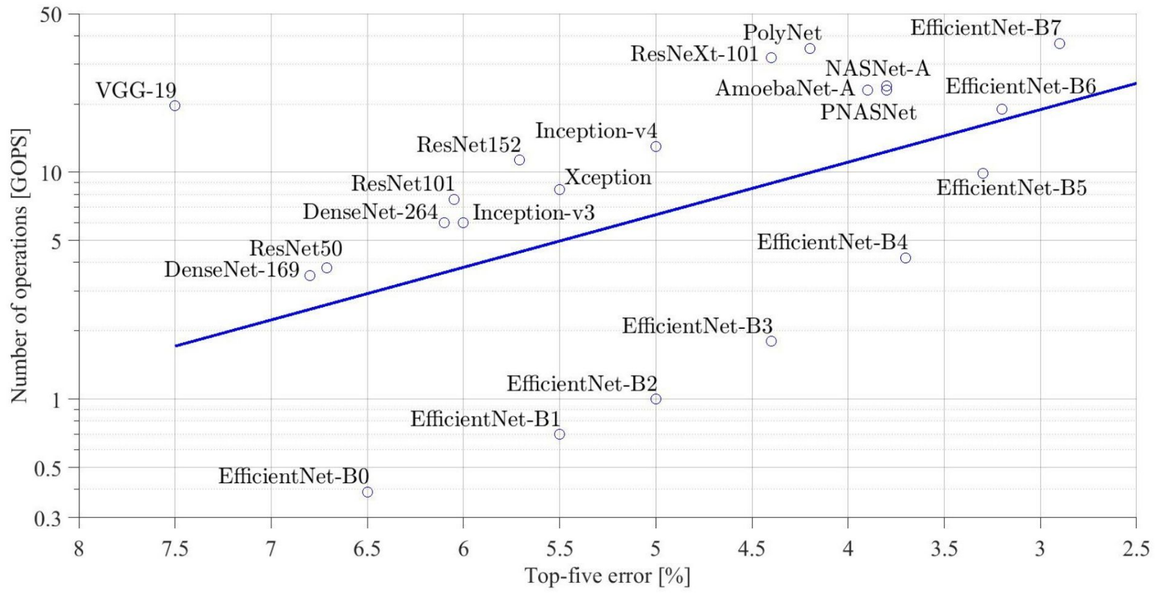
Baischer, L., Wess, M. and TaheriNejad, N., 2021. Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. *arXiv preprint arXiv:2104.09252*.

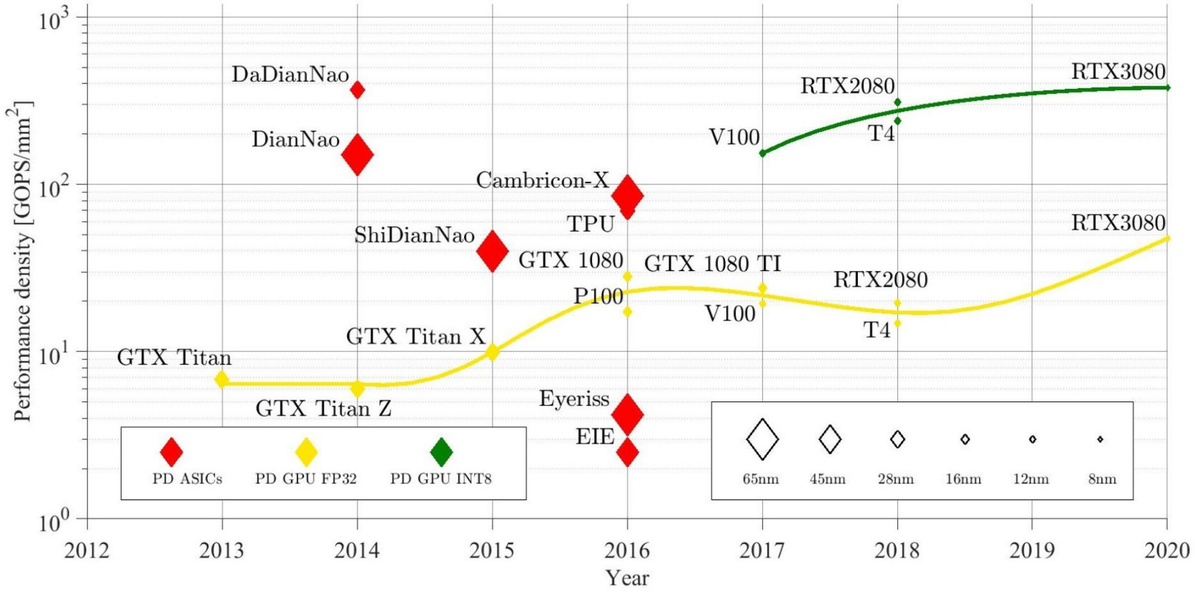
Market share of different technologies

- As of the third quarter of 2017
  - GPU
    - Nvidia represented 72.8%
    - with the rest by AMD
  - FPGA
    - Xilinx 53%
    - Altera 36%
    - Microsemi 7%
    - Lattice Semiconductor (3%)

Hwang, Tim. "Computational power and the social impact of artificial intelligence." *Available at SSRN 3147971* (2018).

The Gap…





Baischer, L., Wess, M. and TaheriNejad, N., 2021. Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. *arXiv preprint arXiv:2104.09252*.

References and further reading

- Kim & Mutlu, “Memory Systems,” Computing Handbook, 2014.
- Baischer, L., Wess, M. and TaheriNejad, N., 2021. Learning on Hardware: A Tutorial on Neural Network Accelerators and Co-Processors. arXiv preprint arXiv:2104.09252.

Thank You