Christof Teuscher

ECE 410/510: Hardware for AI and ML

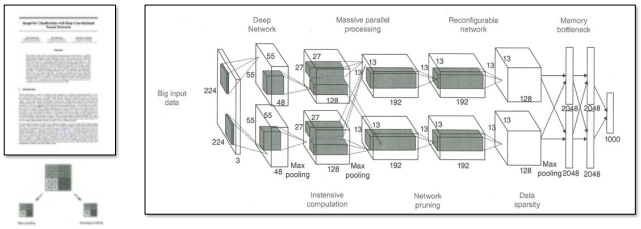
GPUs + CUDA

Portland State University
Department of Electrical and Computer Engineering (ECE)
www.teuscher-lab.com
teuscher@pdx.edu

teuscher Lab teuscher-lab.com

Portland State UNIVERSITY

AlexNet: A classical deep convolutional neural network



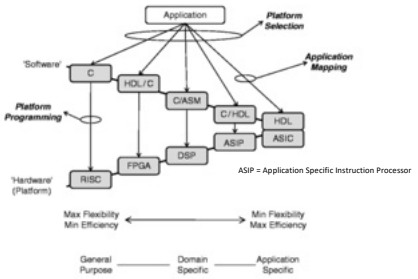
AlexNet paper: https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Liu & Law, Artificial Intelligence Hardware Design, 2021

Flexibility vs perfomance

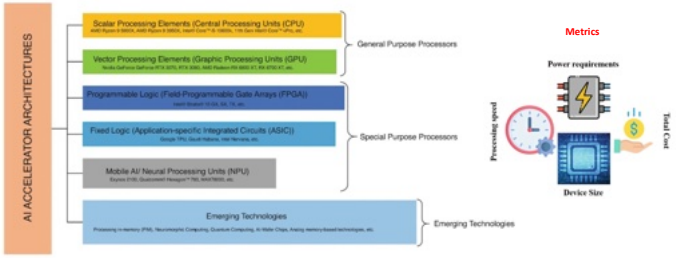
1. Why is running a neural network simulation on a general purpose processor inefficient?
2. How can we improve performance?
3. What is a better processor architecture?

The HW/SW co-design space



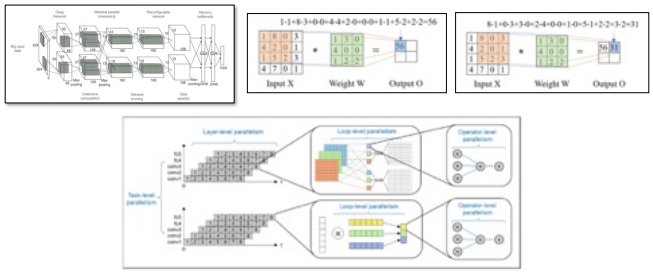
Schaumont, 2013

Hardware for AI/ML



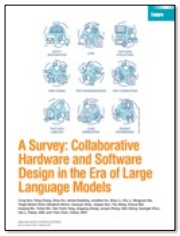
Mishra et al., 2023

Exploiting parallelism



Paper discussion

- What did you learn?



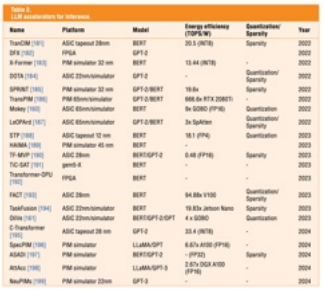
Paper discussion (1)

- The difference between these papers shows how the field of co-design has evolved **from general principles to specialized approaches** for complex AI systems over nearly three decades.
- LLMs face unique challenges compared to traditional deep learning systems, requiring extensive computational resources, high energy consumption, and complex software optimizations.
- LLMs differ from CNNs in their memory requirements and processing patterns, requiring unique co-design methodologies.

Paper discussion (2)

- Co-design is about a systems perspective: meeting system-level objectives.
- Co-design is the key to designing efficient systems.
- Designing HW is increasingly like designing SW.
- Flexibility vs performance trade-off.
  - E.g., reconfigurability increases usability, but does not increase performance.
- Co-design tools: modeling, validation, implementation.
- Identify critical segments of SW and run them on special HW.
- LLMs require unique optimization strategies.
  - GPT-3: 175 billion parameters, 350GB of GPU memory just for parameters
- GPUs are the cornerstone of modern deep learning infrastructure.

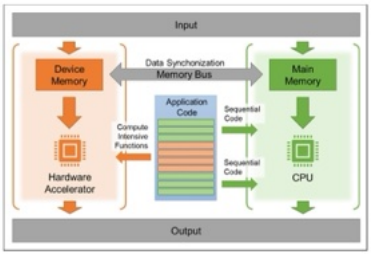
Paper discussion (3)



Guo et al., 2025

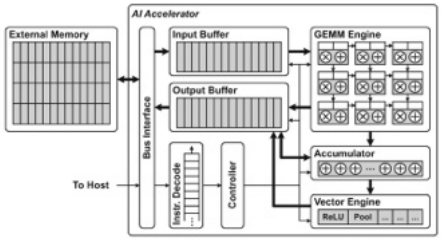
Hardware for AI/ML

General co-processing mechanism



Mishra, Ashutosh; Cha, Jaekwang; Park, Hyunbin; Kim, Shiho. Artificial Intelligence and Hardware Accelerators, 2025

General AI accelerator architecture



GEMM = General Matrix Multiply

Tensor cores!

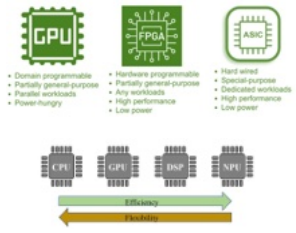
A vector is a 1st-order tensor
A matrix is a 2nd-order tensor.

Mishra, Ashutosh; Cha, Jaekwang; Park, Hyunbin; Kim, Shiho. Artificial Intelligence and Hardware Accelerators, 2025

Hardware for AI/ML

Processor	Power Consumption	Strengths	Limitations
CPU	High	• Flexible • General-purpose processing • Complex instructions and tasks • System management	• Possible memory access bottlenecks • Few cores (4–16)
GPU	High	• Parallel cores (~1000 s of cores) • High-performance AI processing	• Power consumption • Large footprint
FPGA	Medium	• Configurable logic gates • Flexible • In-field re-programma-bility	• Programming complexity
ASIC	Low	• Custom logic designed with libraries • Faster processing • Small footprint	• Fixed function • Expensive custom design
Vision Processing Unit (VPU)	Ultra-low	• Dedicated image and vision co-processor • Small footprint	• Limited dataset and batch size • Limited network support
Tensor Processing Unit (TPU)	Low to medium	• Specialized tool support • Optimized for Tensor-Flow	• Proprietary design • Limited framework support

Mishra et al., 2023



Hardware for AI/ML

Processor	Power Consumption	Strengths	Limitations
CPU	High	• Flexible • General-purpose processing • Complex instructions and tasks • System management	• Possible memory access bottlenecks • Few cores (4–16)
GPU	High	• Parallel cores (~1000 s of cores) • High-performance AI processing	• Power consumption • Large footprint
FPGA	Medium	• Configurable logic gates • Flexible • In-field re-programma-bility	• Programming complexity
ASIC	Low	• Custom logic designed with libraries • Faster processing • Small footprint	• Fixed function • Expensive custom design
Vision Processing Unit (VPU)	Ultra-low	• Dedicated image and vision co-processor • Small footprint	• Limited dataset and batch size • Limited network support
Tensor Processing Unit (TPU)	Low to medium	• Specialized tool support • Optimized for Tensor-Flow	• Proprietary design • Limited framework support

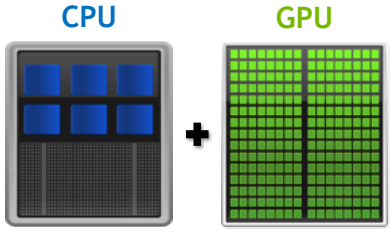
Mishra et al., 2023

Graphics Processing Units (GPUs)

A bit of history

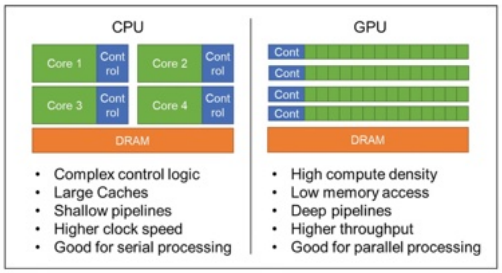
- Over several decades, GPUs evolved from a single core, fixed function
- hardware used for graphics solely, to a set of programmable parallel cores.
- The history of modern GPUs starts in 1995 with the introduction of the first 3D add-in cards.
- 1996: **3DFx's Voodoo** graphics card took over about 85% of the market. Cards that could only render 2D became obsolete very fast.
- 1999: **NVIDIA's GeForce 256**, the "world's first GPU."

Multi-core vs GPU



NVIDIA

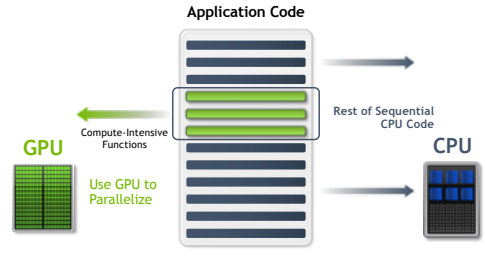
Multi-core vs GPU



SIMT (Single Instruction, Multiple Threads) is a parallel processing model used in GPUs where multiple threads execute the same instruction simultaneously, but on different data.

Mishra, Ashutosh; Cha, Jaekwang; Park, Hyunbin; Kim, Shiho. Artificial Intelligence and Hardware Accelerators, 2025

Small Changes, Big Speed-up

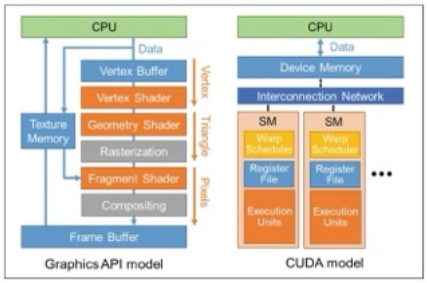


NVIDIA

CUDA model

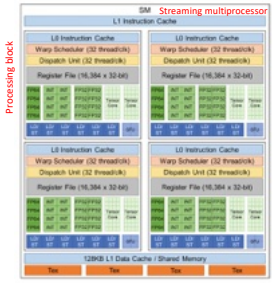
General-purpose computing on GPU (**GPGPU**)

Early GPGPU was implemented by formulating data and operations in graphics primitives and translated to pass through a fixed graphics pipeline.

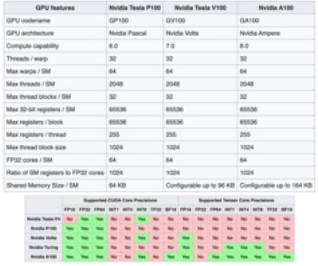


Mishra, Ashutosh; Cha, Jaekwang; Park, Hyunbin; Kim, Shiho. Artificial Intelligence and Hardware Accelerators, 2025

NVIDIA Volta Microarchitecture (2013-2017)



Tex = Texture units
SFU = Special Function Unit



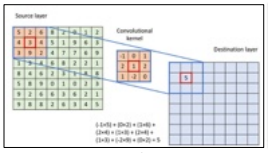
- First chip to feature tensor cores
- A GV100 SM is partitioned into four processing blocks

Warp

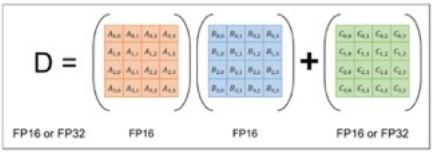
- A "warp" is a group of threads that execute instructions simultaneously on a GPU (Graphics Processing Unit).
- The warp scheduler is a component of the GPU architecture that manages the execution of these thread groups.
- A warp typically consists of 32 threads in NVIDIA GPUs (AMD calls similar groupings "wavefronts"). These threads execute the same instruction at the same time in a **SIMT** (Single Instruction, Multiple Thread) architecture.
- The warp scheduler is responsible for:
  - Deciding which warps are ready to execute
  - Allocating compute resources to warps
  - Managing warp context switching when threads need to wait for memory operations.
  - Hiding memory latency by switching between warps

Tensor cores

- Each tensor core executes MMA (matrix multiplication and accumulation, D = A × B + C) on a 4 × 4 floating-point matrix.
- They are built to enhance deep learning by boosting matrix arithmetic.
- Low-precision operations often supported.



Convolution



MMA

Mishra et al., 2023

Streaming vs tensor cores

SM Cores (Streaming Multiprocessors)

- These are the traditional GPU computing units
- Handle general-purpose parallel computing tasks
- Process standard graphics rendering workloads
- Execute regular FP32 (single precision) and INT32 operations
- Responsible for most traditional GPU workloads

Tensor Cores

- Specialized hardware accelerators designed specifically for matrix operations
- Optimized for deep learning and AI workloads
- Dramatically accelerate matrix multiply-accumulate operations (critical for neural networks)
- Can perform mixed-precision calculations (like FP16 and FP32)
- Provide order-of-magnitude speedups for deep learning training and inference
- Introduced with NVIDIA's Volta architecture and improved in subsequent generations

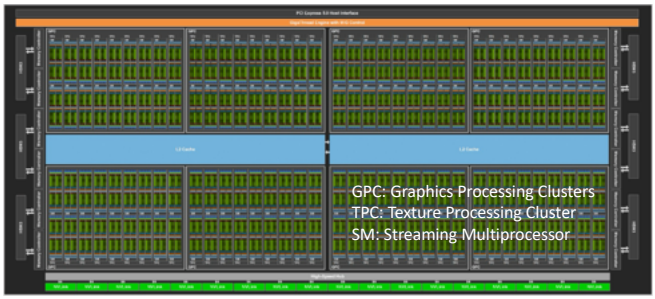
Graphics Card	GeForce RTX 3090	GeForce RTX 4090	GeForce RTX 5090
GPU Codename	GA102	AD102	GB202
GPU Architecture	NVIDIA Ampere	NVIDIA Ada Lovelace	NVIDIA Blackwell
GPCs	7	11	11
TPCs	41	64	85
SMs	82	128	170
CUDA Cores / SM	128	128	128
CUDA Cores / GPU	10496	16384	21760
Tensor Cores / SM	4 (3rd Gen)	4 (4th Gen)	4 (5th Gen)
Tensor Cores / GPU	328 (3rd Gen)	512 (4th Gen)	680 (5th Gen)
GPU Boost Clock (MHz)	1695	2520	2407
RT Cores	82 (2nd Gen)	128 (3rd Gen)	170 (4th Gen)
RT TFLOPS	69.5	191	317.5
Frame Buffer Memory Size and Type	24 GB GDDR6X	24 GB GDDR6X	32 GB GDDR7
Memory Interface	384-bit	384-bit	512-bit
Memory Clock (Data Rate)	19.5 Gbps	21 Gbps	28 Gbps
Memory Bandwidth	936 GB/sec	1008 GB/sec	1792 GB/sec
ROPs	112	176	176

Consumer-grade GPUs

Pixel Fill-rate (Gigapixels/sec)	189.8	443.5	423.6
Texture Units	328	512	680
Texel Fill-rate (Gigatexels/sec)	555.96	1290.2	1636.76
L1 Data Cache/Shared Memory	10496 KB	16384 KB	21760 KB
L2 Cache Size	6144 KB	73728 KB	98304 KB
TGP (Total Graphics Power)	350 W	450 W	575 W
Manufacturing Process	Samsung 8 nm 8N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process	TSMC 4nm 4N NVIDIA Custom Process
PCI Express Interface	Gen 4	Gen 4	Gen 5

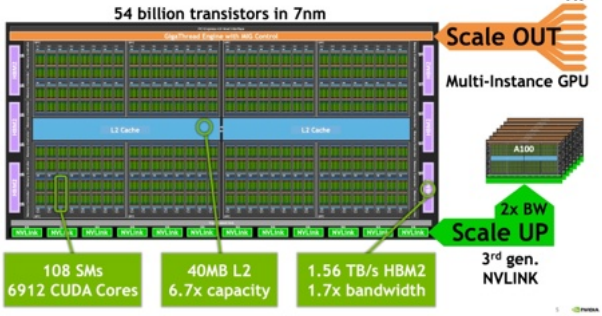
GPC: Graphics Processing Clusters
TPC: Texture Processing Cluster
SM: Streaming Multiprocessor
RT: Ray Tracing
ROP: Render Output Unit

RTX 5090



A100 TENSOR-CORE GPU

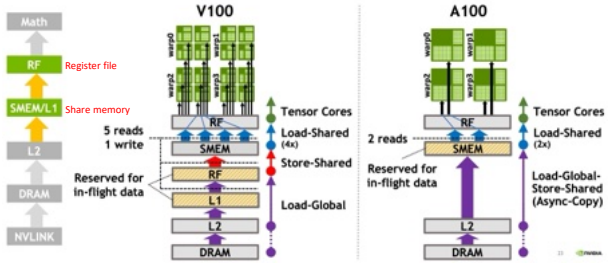
54 billion transistors in 7nm



V100 A100

A100 data movement efficiency

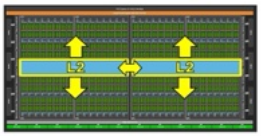
3x SMEM/L1 bandwidth, 2x in-flight capacity



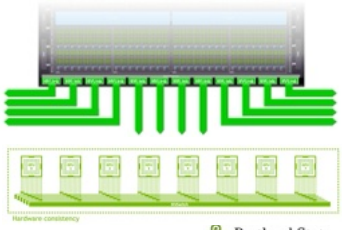
in-flight data = data that is actively being processed or transferred within the system but hasn't yet reached its final destination

A100 data movement efficiency

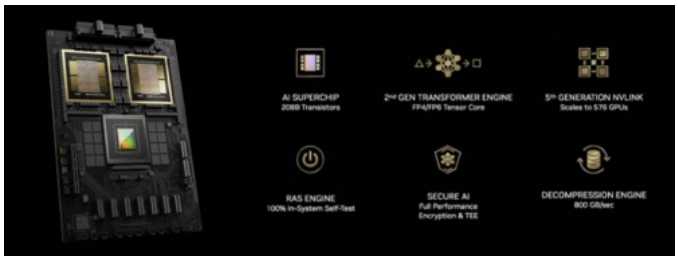
Split L2 with hierarchical crossbar - 2.3x increase in bandwidth over V100, lower latency



Third Generation NVLink
50 Gbit/sec per signal pair
12 links, 25 GB/s in/out, 600 GB/s total
2x vs. V100



GB100 $40k, GB200: $70k



RAS (Reliability, Availability, and Serviceability)

Datacenter

Die	GB100	
Variant(s)		
Release date	Dec 2024	
Cores	CUDA Cores	
	TMUs	
	ROPs	
	RT Cores	
	Tensor Cores	
Streaming Multiprocessors		
Cache	L1	
	L2	
Memory interface	8192-bit	
Die size		
Transistor count	104 bn.	
Transistor density		
Package socket	SXM6	
Products	B100	
	B200	

Consumer

Die						
Variant(s)						
Release date						TBA
Cores	CUDA Cores					
	TMUs					
	ROPs					
	RT Cores					
	Tensor Cores					
SMs						
GPCs						
Cache	L1					
	L2					
Memory interface						
Die size						
Transistor count						
Transistor density						
Products						
Consumer	Desktop					
	Mobile					
Workstation	Desktop					
	Mobile					
Server						