AI Hardware Basics

Liang Cheng, Ph.D.

Department of Electrical Engineering and Computer Science

THE UNIVERSITY OF TOLEDO

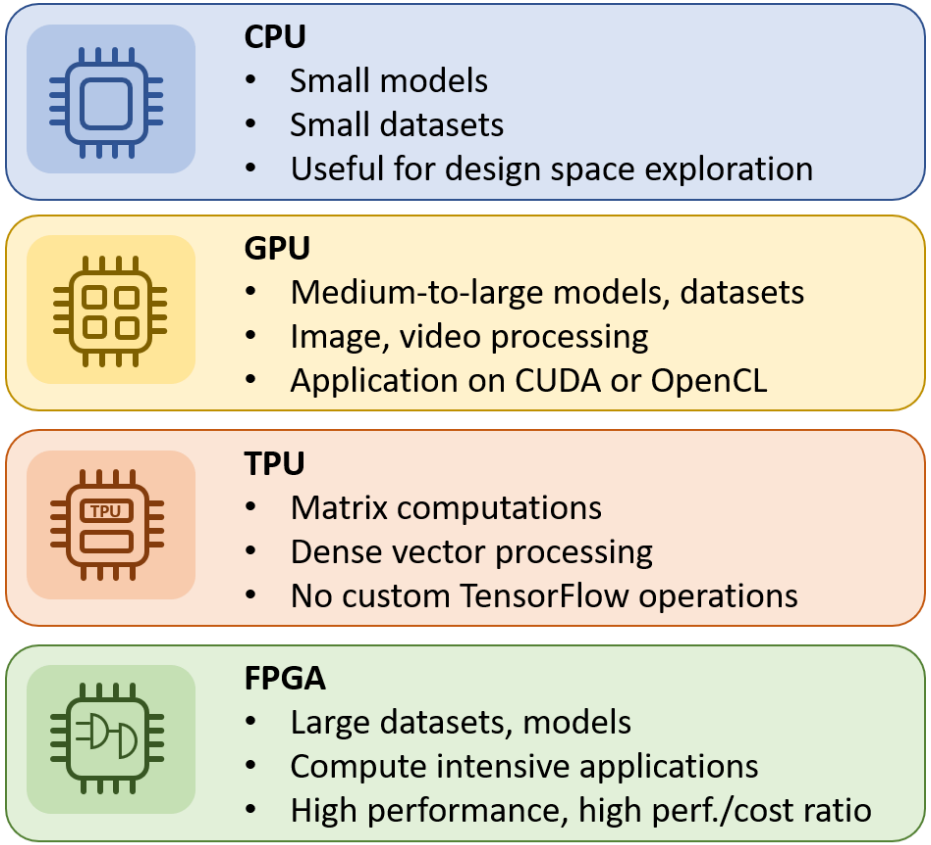
Key Components

- Neural Processing Unit (NPU)
  - Responsible for executing AI algorithms and performing computations for neural networks
- Memory
  - Storing and retrieving large amounts of data used by neural networks
- Interconnects
  - For data transfer between AI hardware components
- Power Management

THE UNIVERSITY OF TOLEDO

Design Considerations

- Performance
- Scalability
- Flexibility
- Energy Efficiency



CPU, GPU, FPGA or TPU: Which one to choose for my Machine learning training? | by InAccel | Medium

- CPU
  - Optimized for sequential logic
- GPU
  - Optimized for parallel computing
- TPU
  - Optimized for tensor operations

THE UNIVERSITY OF TOLEDO

Comparison

Observation	ParaDnn*	Proof	Insight/Explanation
1. TPU does not exploit the parallelism from the model depth (layer count).	✓	Fig 2	To design/upgrade new specialized systems, architects need to consider interactions between the operation mix from key workloads (arithmetic intensity) and system configurations (FLOPS, memory bandwidth/ capacity, and intra-chip and host-device interconnect). TPU serves as a great example.
2. Many FC and CNN operations are bottlenecked by TPU memory bandwidth.	✓	Fig 3	
3. TPU suffers large overheads due to inter-chip communication bottlenecks.	✓	Fig 4	
4. TPU performance can be improved by $\geq 34\%$ by improving data infeed.	-	Fig 5	
5. TPU v3 optimizes compute-bound MatMuls by 2.3×, memory-bound ones by 3×, and large embeddings by $> 3\times$, compared to v2.	✓	Fig 6	
6. The largest FC models prefer CPU due to memory constraints.	✓	Fig 7	Need for model parallelism on GPU and TPU.
7. Models with large batch size prefer TPU. Those with small batch size prefer GPU.	-	Fig 8 Fig 10	Large batches pack well on systolic arrays; warp scheduling is flexible for small batches.
8. Smaller FC models prefer TPU and larger FC models prefer GPU.	✓	Fig 8	FC needs more memory bandwidth per core (GPU).
9. TPU speedup over GPU increases with larger CNNs.	✓	Fig 10	TPU architecture is highly optimized for large CNNs.
10. TPU achieves 2× (CNN) and 3× (RNN) FLOPS utilization compared to GPU.	✓	Fig 11	TPU is optimized for both CNN and RNN models.
11. GPU performance scales better with RNN embedding size than TPU.	✓	Fig 10	GPU is more flexible to parallelize non-MatMuls.
12. Within seven months, the software stack specialized for TPU improved by up to 2.5× (CNN), 7× (FC), and 9.7× (RNN).	✓	Fig 12	It is easier to optimize for certain models than to benefit all models at once.
13. Quantization from 32 bits to 16 bits significantly improves TPU and GPU performance.	-	Fig 5 Fig 12	Smaller data types save memory traffic and enable larger batch sizes, resulting in super-linear speedups.
14. TensorFlow and CUDA teams provide substantial performance improvements in each update.	✓	Fig 12	There is huge potential to optimize compilers even after the hardware has shipped.

* Without ParaDnn the insights are not revealed, and/or lack deep explanations.

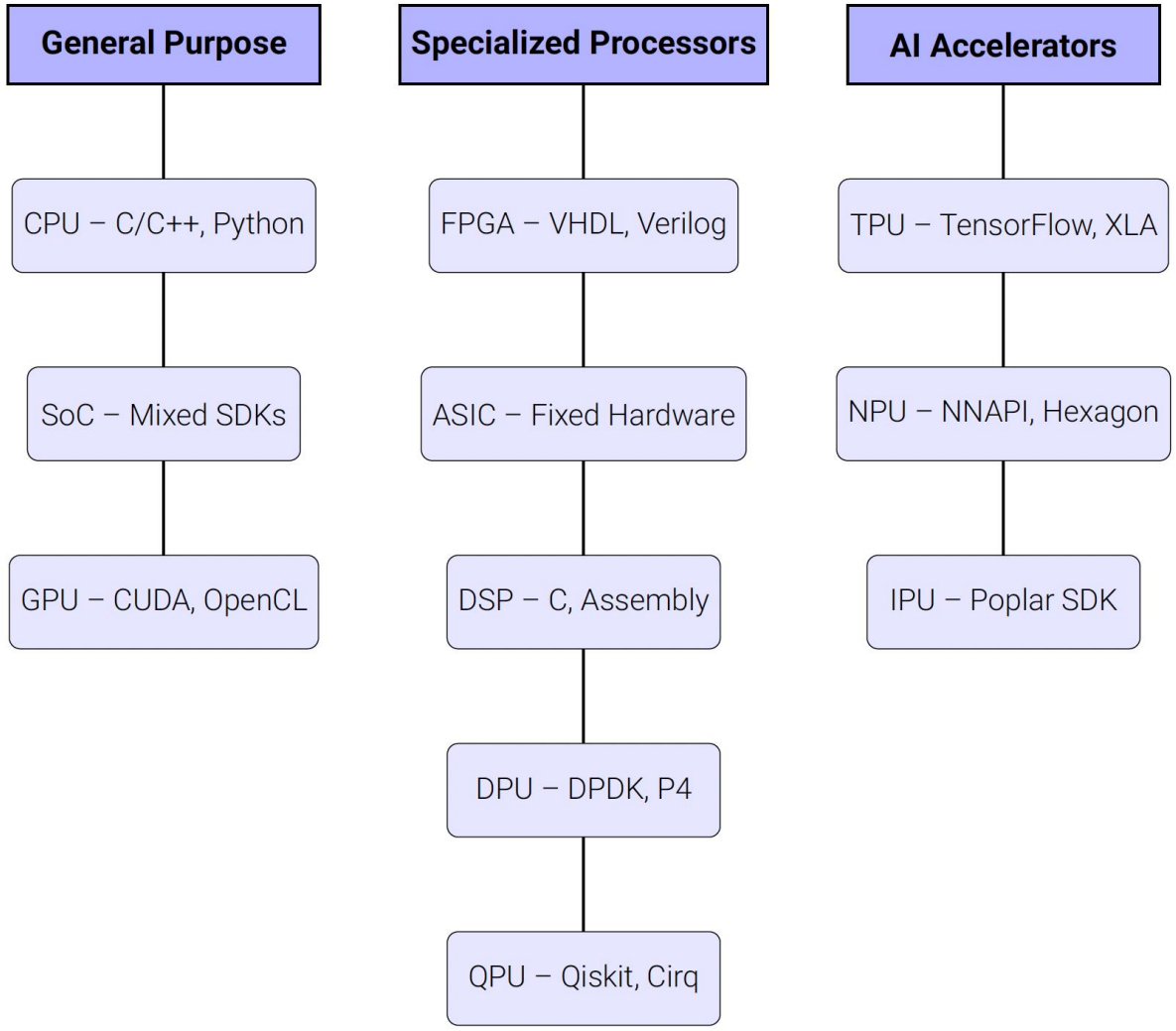
Wang, Y. E., Wei, G.-Y., and Brooks, D., "Benchmarking TPU, GPU, and CPU Platforms for Deep Learning", arXiv e-prints, Art. no. arXiv:1907.10701, 2019. doi:10.48550/arXiv.1907.10701.

A recent comparison study is reported by Alahmari and Nassef: Comprehensive Comparative Study of CPU, GPU, and TPU Performance for Neural Network Architectures Across Diverse Datasets, 2025.

THE UNIVERSITY OF TOLEDO

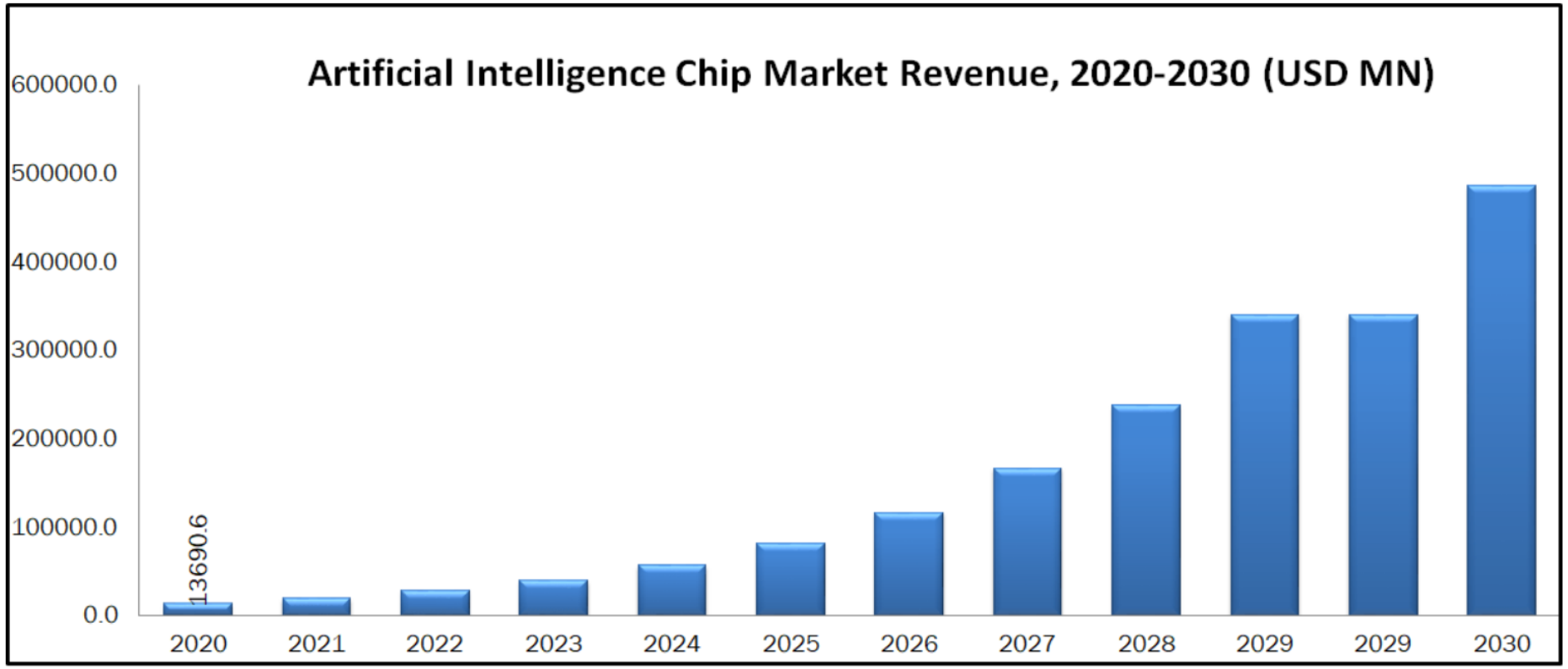
Related Discussions

- Brief Introduction to Modern Processing Devices
  - CPUs, SoCs, GPUs, DSPs, TPUs, NPUs, IPUs, DPUs, QPUs
    - Problems to solve
    - Architectural design
    - Programming tools
      - By Manuel José Fernández Iglesias, 2025



THE UNIVERSITY OF TOLEDO

Market Trend



Artificial Intelligence (Ai) Chip Market Size, Share, Growth, Forecast till 2030 (decisionforesight.com)

THE UNIVERSITY OF TOLEDO