CompF04:
AI Hardware White Paper

Javier Duarte (UCSD)
Nhan Tran (Fermilab)

Today's agenda

This talk will present outline of current imagined AI Hardware thrusts

We have two brief talks on synergistic areas
- Allison Deiana McCarn (IF04)
- Dylan Rankin (CompF03)

Overlaps are good! Demonstrates importance across different subgroups and frontiers. Important for identifying broader themes

White paper scope

https://www.overleaf.com/7445234346cypyrnbjmkck

AI Hardware: The need for accelerating machine learning, and more generally artificial intelligence (AI), is leading to an explosion in different hardware architectures designed to optimize AI/ML. This area includes all the AI/ML-only hardware, i.e. architectures that are not general purpose.

Scope is limited to “offline” AI hardware - we will not cover real-time, online hardware though there is a large overlap. We will acknowledge that in our report but not focus on it

Outline

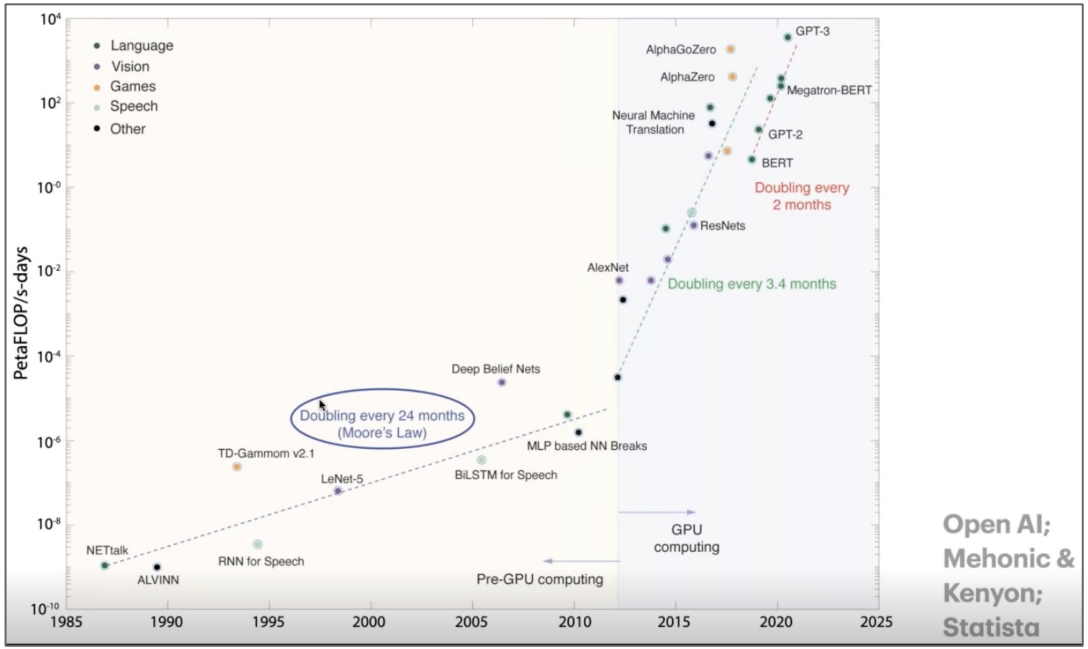
https://www.overleaf.com/7445234346cypyrnbjmkck

Executive summary

Science drivers

Hardware taxonomy

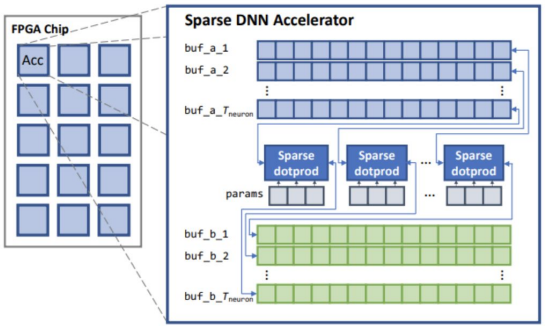
Software ecosystem and integration



Science drivers

Enumerate broad range of tasks

- Different AI tasks map more naturally on to different architectures
- **Image inputs, sparsity, point clouds, time-series data**
  - Maps on to different architectural challenges, e.g. sparsity:
    - MV vs. SpMV vs. SpMSpV
- Task scaling
  - How many parameters, bit ops, batch size?
- Different needs for training and inference
  - HEP known more for large inference workloads



FPGA based Sparse DNN Accelerator from HPEC 2019[1]

https://arxiv.org/pdf/2110.13041.pdf

Hardware taxonomy

Taxonomy of Compute Architectures for Deep Learning

A broad range of hardware architectures to deploy machine learning algorithms exists today. We can broadly classify them by the following criteria:

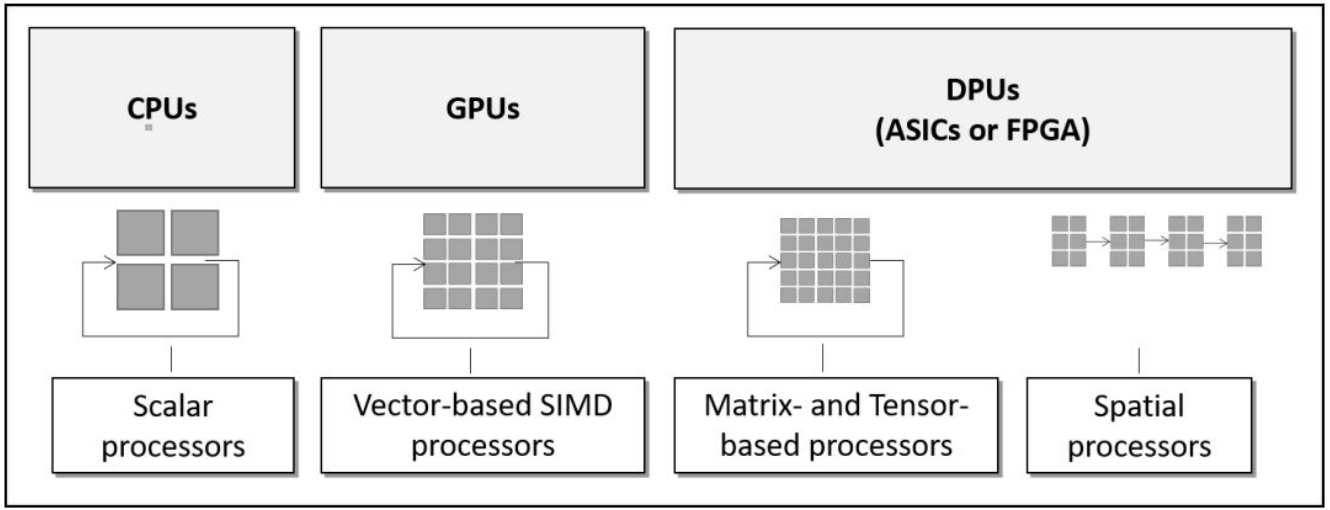
1. Basic type of compute operation
2. Inherent support for specific numerical representations
3. External memory capacity (which is mostly relevant for training workloads) [3]
4. External memory access bandwidth
5. Power consumption in the form of thermal design power (TDP)
6. Level of parallelism in the architecture and the degree of specialization

Server-class	Throughput	Latency	Power	Ext. Mem. Bandwidth	HW specialization	Ease of Use	Training/Inference
Conventional							
CPU	Medium	High	High	Medium	Low	High	Both
DPU-MPE	High	Medium-High	Medium	High	Medium	Low-Medium	Inference
DPU-Spatial	High	Low	Medium	High	High	Low	Inference
GPU (NVIDIA A100)	High	High	High	High	Medium	High	Both
Speculative							
Cerebras CS-1	Very High	Medium	High	Very High	Medium	Medium	Both

https://arxiv.org/pdf/2110.13041.pdf

Hardware taxonomy

Conventional CMOS



GPU:
Nvidia, AMD, Intel

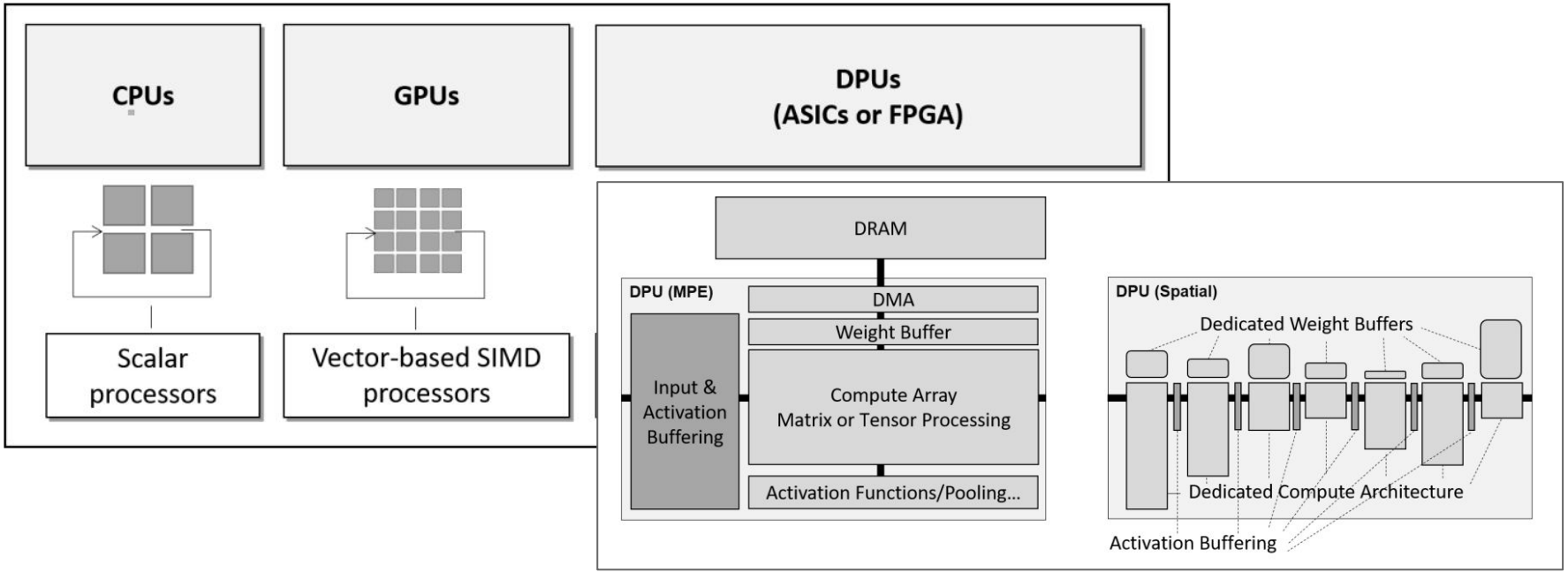
FPGA:
Intel, AMD/Xilinx

ASIC:
Google TPU, Intel Habana Gaudi/Goya

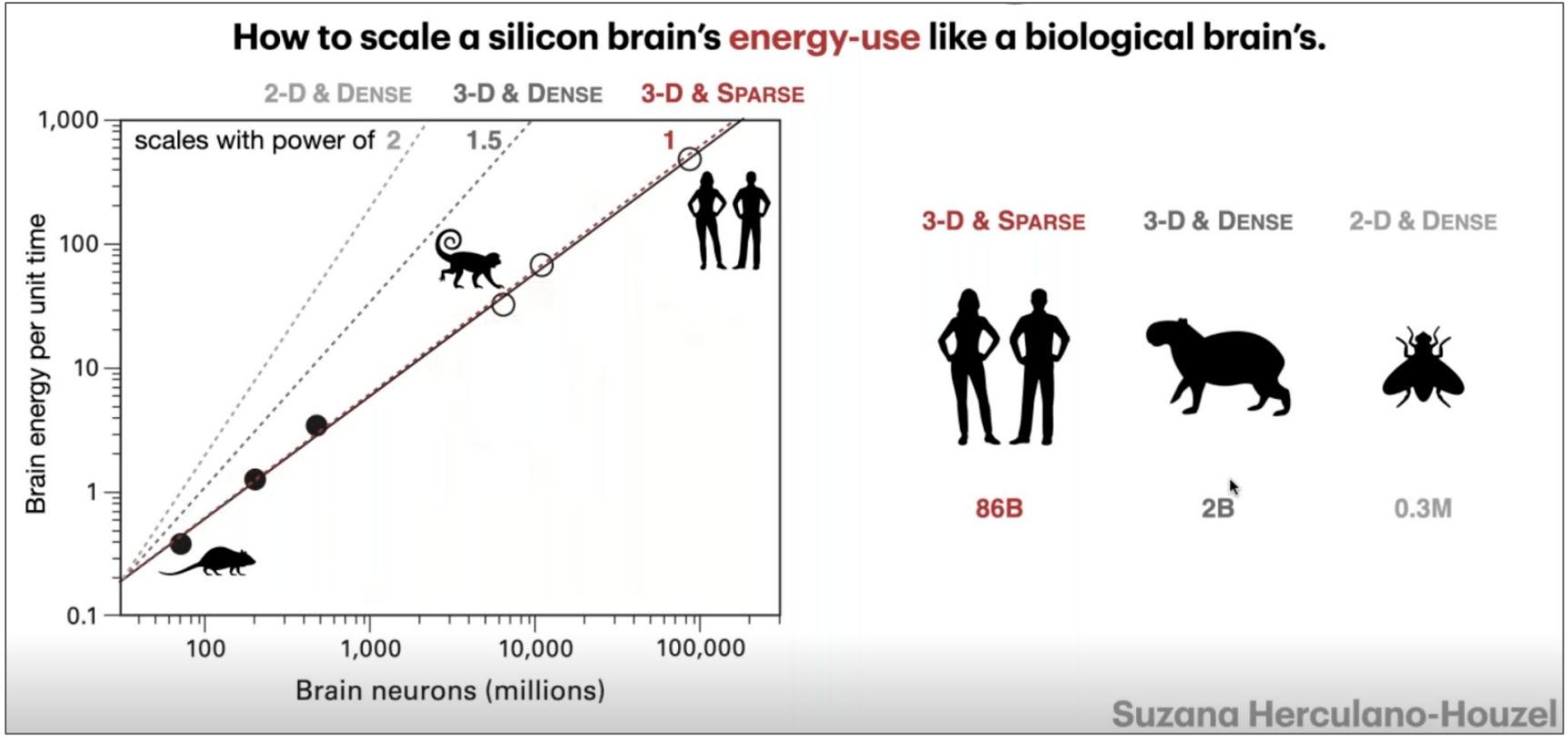
https://arxiv.org/pdf/2110.13041.pdf

Hardware taxonomy

Conventional CMOS



Beyond Conventional CMOS



Snapshot of novel archs

n.b. LIST INCOMPLETE:

Large-scale dataflow - e.g. IPU, SambaNova

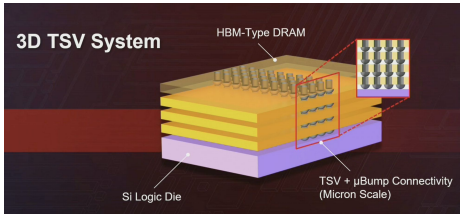
Wafer-scale integration, e.g. Cerebras

3D stacking, e.g. TSMC

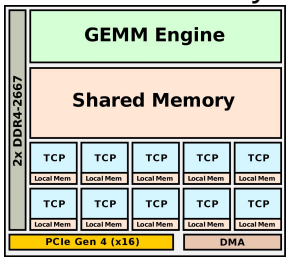
Analog, e.g. Mythic

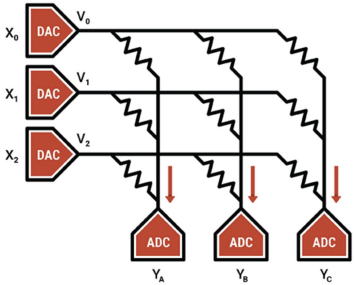
Spiking/Neuromorphic, e.g. IBM True North

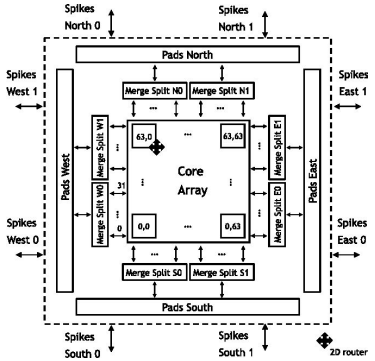
Photonic, e.g. lightelligence

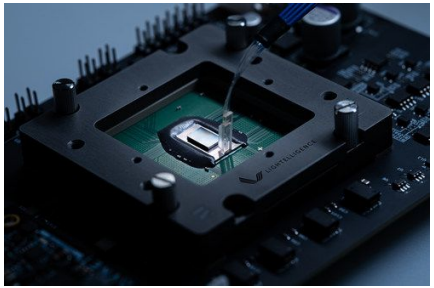


Goya









SW and integration

Rely heavily on industry-supported ecosystems for connecting AI software frameworks (TF, PyTorch, etc.) to novel hardware

- Can be a pain point, especially for custom operations and specialized packages, e.g. PyTorch Geometric

Abstraction of hardware explored via NVidia Triton and S.O.N.I.C

- Containerization tools
- Other tools (TFaaS?) or custom tools

Key R&D areas

We are largely at the whim of industry, but...

- We should focus on understanding the synergy between our *unique* AI tasks and available hardware platforms
  - e.g. FPGA better for small-batch, GPU better for large-batch? When are spatial data flow architectures faster than MPEs? Can CPUs do well on sparse architectures? When are spiking NN models good for physics applications?
- We should understand how best to integrate AI hardware into our SW frameworks and computing infrastructure
  - On-prem, HPC, cloud

Contributing

Let us know if you would like to get involved in contributing to the AI hardware white paper!