Christof Teuscher

ECE 410/510: Hardware for AI and ML

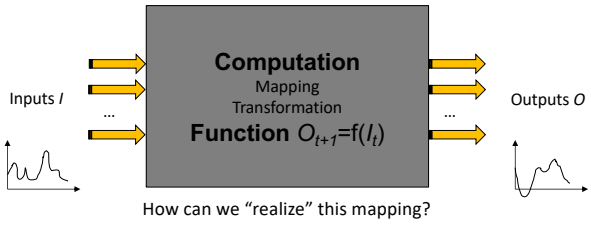
Hardware for AI and ML Overview + Co-design

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

What is a computer?

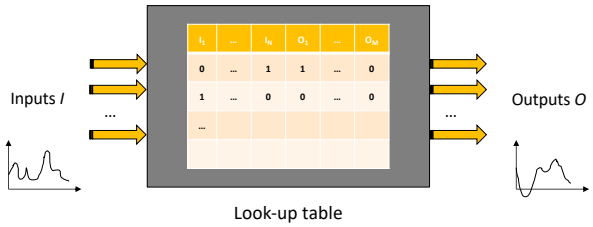
What's the job of a computer architect / chip designer?

What is computation? (1)



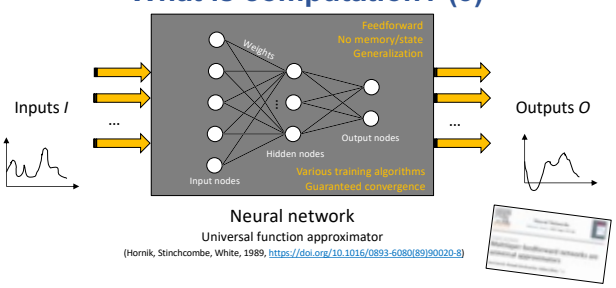
How can we "realize" this mapping?

What is computation? (2)



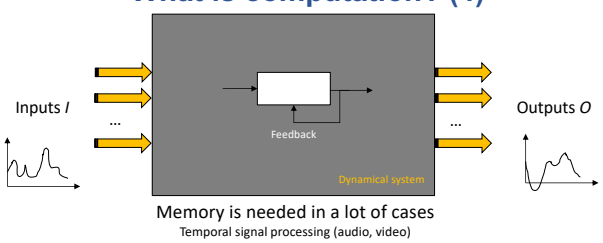
Look-up table
Simple, fast

What is computation? (3)



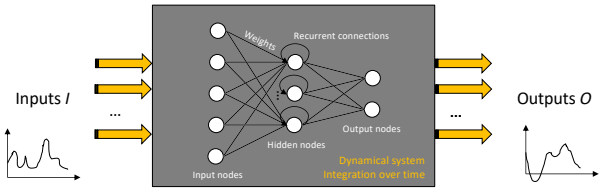
Neural network
Universal function approximator
(Hornik, Stinchcombe, White, 1989, https://doi.org/10.1016/0893-6080(89)90020-8)

What is computation? (4)



Memory is needed in a lot of cases
Temporal signal processing (audio, video)
Timeseries prediction
Control tasks

What is computation? (5)

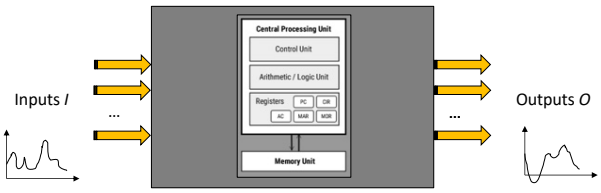


Recurrent neural networks (RNN)

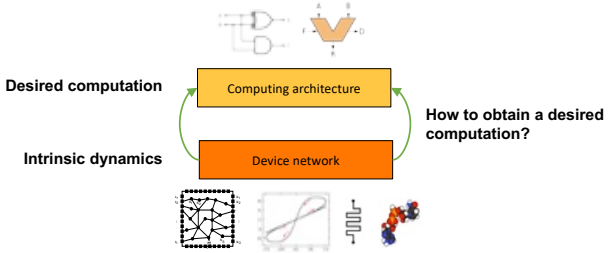
Training is challenging, various algorithms exist (backpropagation through time), convergence not guaranteed, many weights, slow learning. RNNs are also universal function approximators

(Schaefer & Zimmerman, 2007, https://doi.org/10.1142/S0129065707001111)

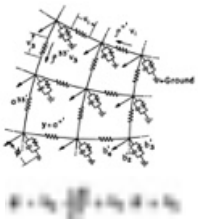
What is computation? (6)



Intrinsic vs designed computation (2)



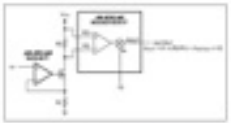
Intrinsic vs designed computation (3)



vs

Kron, Gabriel. "**Electric circuit models of partial differential equations.**" Electrical Engineering 67, no. 7 (1948): 672-684.

Intrinsic vs designed computation (4)



https://www.analog.com/en/resources/design-notes/using-a-generic-0-to-1v-analog-multiplier-to-ensure-accurate-power-management.html

vs

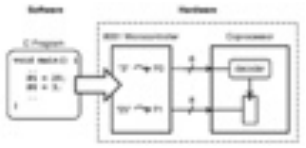
What are the design trade-offs?

- **Performance**
- **Flexibility** (general purpose vs application-specific)
- **Energy efficiency**
- **Power density** (avoid burning chip, cost of cooling technology)
- **Design complexity**
- **Design cost**
- **Scalability**
- **Shrinking design schedules** (time-to-market)

What is (HW/SW) co-design?

- The practice of taking the **best from software** design and the **best from hardware** design to solve design problems.
- HW/SW co-design deals with HW/SW **interfaces**.
- HW/SW co-design is the design of cooperating HW and SW components in a **single design effort**.

- Why design new HW if there is already one available to do the job? E.g., a RISC processor?
- From a flexibility and design effort perspective, as much as possible should be in software. [Easy, flexible, available libraries, etc.]



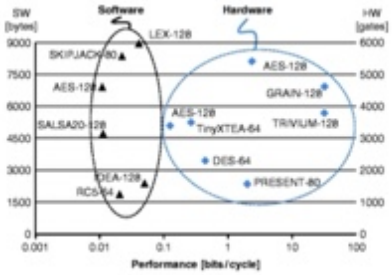
Schaumont, 2013

Why is HW/SW co-design important for AI/ML

- **Performance optimization**: Traditional hardware design assumes fixed software, while traditional software design assumes fixed hardware. Co-design breaks this barrier by simultaneously optimizing both, leading to significant performance improvements for AI workloads.
- **Specialized acceleration**: AI algorithms have unique computation patterns (like matrix multiplications and convolutions) that benefit from custom hardware accelerators tailored to these specific operations.
- **Memory bottlenecks**: AI models face severe memory bandwidth constraints. Co-designing hardware memory hierarchies with software that efficiently schedules operations can minimize data movement and maximize throughput.
- **Energy efficiency**: By tailoring hardware precisely to the needs of ML workloads and optimizing software to take advantage of hardware capabilities, co-design dramatically reduces power consumption, which is critical for both data centers and edge devices.

Cryptography on small embedded platforms

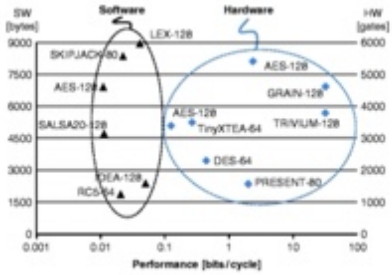
What do we learn here?



Schaumont, 2013

Cryptography on small embedded platforms

Hardware crypto-architectures have, on the average, a higher relative performance compared to embedded processors.

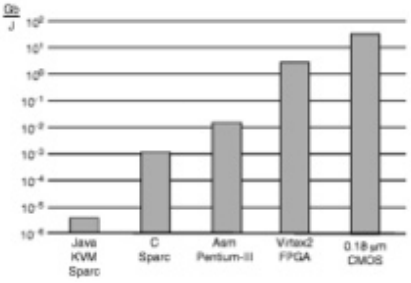


Schaumont, 2013

Cryptography on small embedded platforms

AES for different target platforms

What do we learn here?

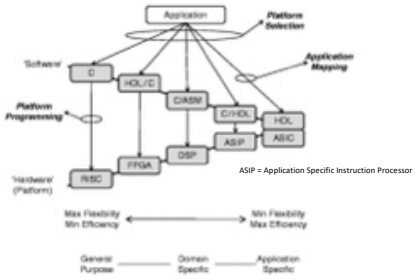


Schaumont, 2013

Driving factors in HW/SW co-design + dualism

Schaumont, 2013

The HW/SW co-design space

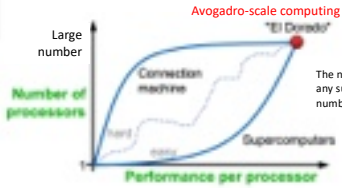


ASIP = Application Specific Instruction Processor

Schaumont, 2013

The goals and drivers have changed

CM 1 & 2: Up to 65,536 1-bit processors

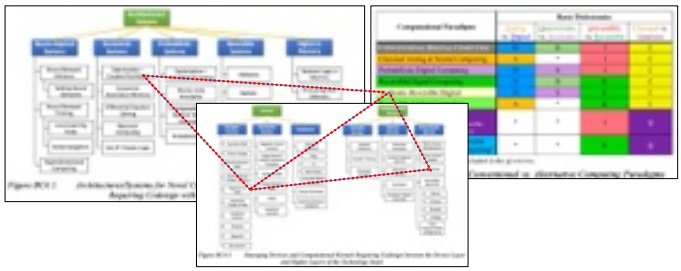


The number of units in one mole of any substance is called Avogadro's number or Avogadro's constant.
6.02×10^{23}

The goals and drivers have changed

https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0061

Deep co-design across the entire stack



What is Artificial Intelligence (AI)? What is Machine Learning (ML)?

AI vs ML

- **Machine Learning (ML)** is one technique used to achieve **Artificial Intelligence** (AI).
- All ML is AI, but not all AI is ML.
- AI is the destination (e.g., intelligent machines), while ML is one path to get there (learning from data).

- **Artificial Intelligence (AI):** Broader concept encompassing machines that can perform tasks requiring human intelligence. Ultimate goal: Artificial General Intelligence (AGI)
- **Machine Learning (ML):** Subset of AI focused specifically on algorithms that improve through experience. Example: neural networks

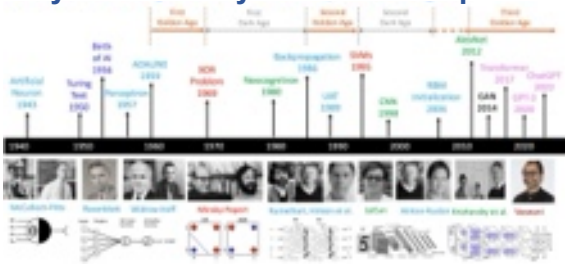
AI vs ML

Source: Tignis

A very brief history of AI

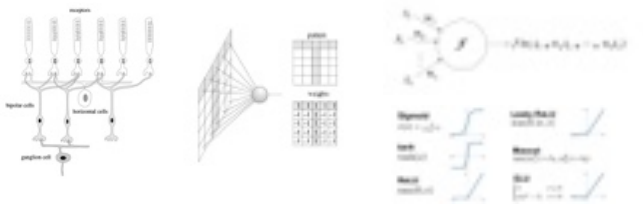
Strong AI = AGI = Artificial General Intelligence

A very brief history of AI with deep learning



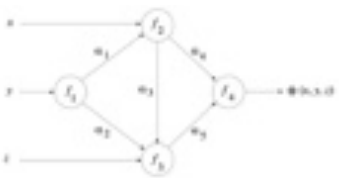
https://medium.com/@lmpo/a-brief-history-of-ai-with-deep-learning-26f7948bc87b

An abstract neuron



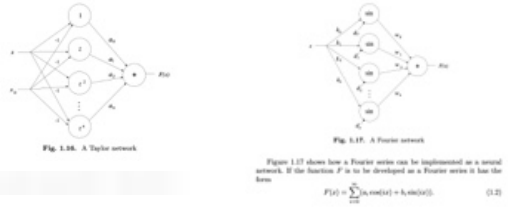
Rojas, 1986

A neural network



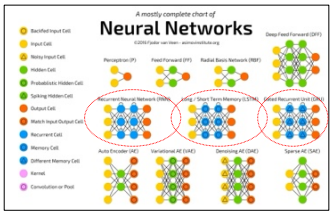
Rojas, 1986

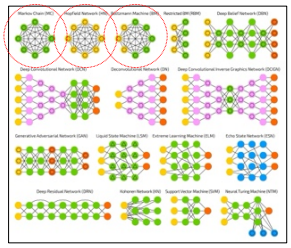
Special types of neural networks



Rojas, 1986

Types of neural networks

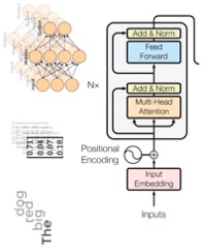




(Fjodor van Veen, 2016, asimovinstitute.org)

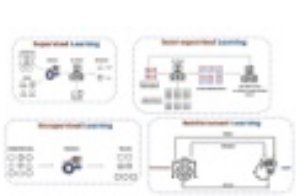
Deep neural networks

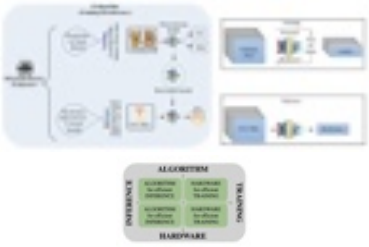
Transformers are deep neural networks



https://builtin.com/artificial-intelligence/transformer-neural-network

Learning





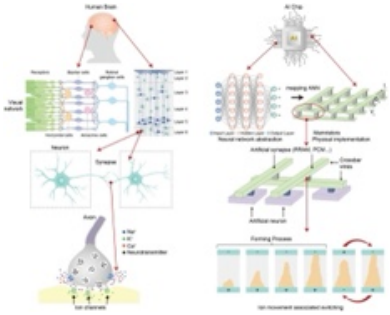
What is harder? Inference or training?

Mishra et al., 2023

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

There are two general approaches to developing artificial general intelligence (AGI):

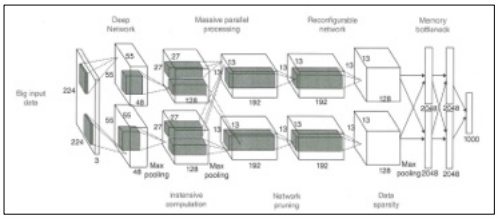
1. computer-science-oriented and
2. neuroscience-oriented



Source: "Bridging Biological and Artificial Neural Networks with Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges"

Neuromorphic Computing with Large Scale Spiking Neural Networks

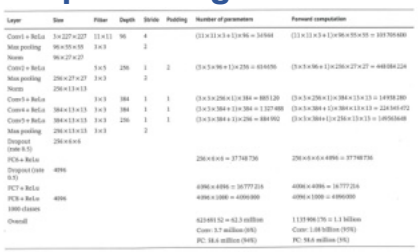
AlexNet: A classical deep convolutional neural network



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

Liu & Law, Artificial Intelligence Hardware Design, 2021

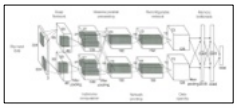
Deep learning HW issues



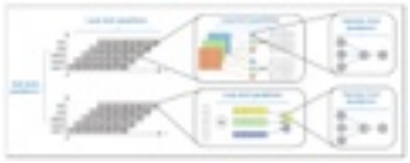
FC = fully connected

Liu & Law, Artificial Intelligence Hardware Design, 2021

Exploiting parallelism

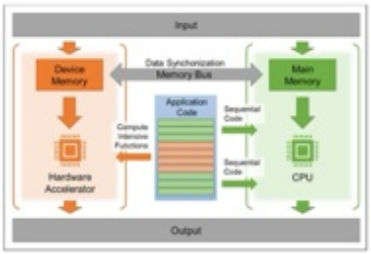


Does that come at a cost?



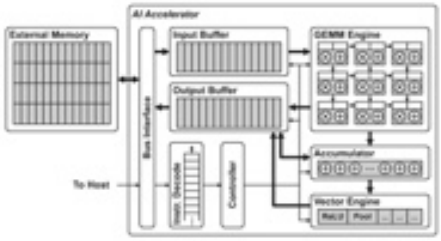
Hardware for AI/ML

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

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