Artificial Intelligence and HPC

Giulio Cesare M. Santo
Nikolas Kronemberger
Welberth Nascimento

Summary

1. Context
2. High Performance Modeling
3. HPC Hardware for AI

Context

What is AI?

- Science and Engineering of designing intelligent machines
- Aims to utilize computers to simulate the human capacities of decision making and problem solving.

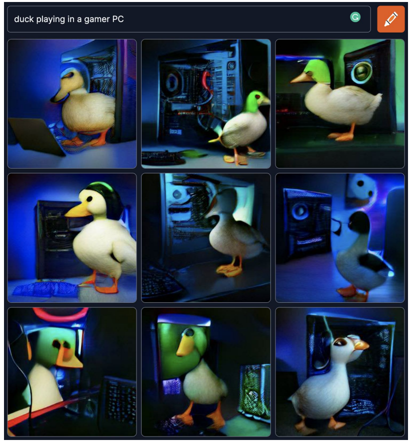
$$\int \frac{\sec^2 t}{1 + \sec^2 t - 3 \tan t} \, dt$$

$$\int \frac{x^4}{(1 - x^2)^{5/2}} \, dx$$

$$\int \frac{x\,dx}{\sqrt{x^2 + 2x + 5}}$$

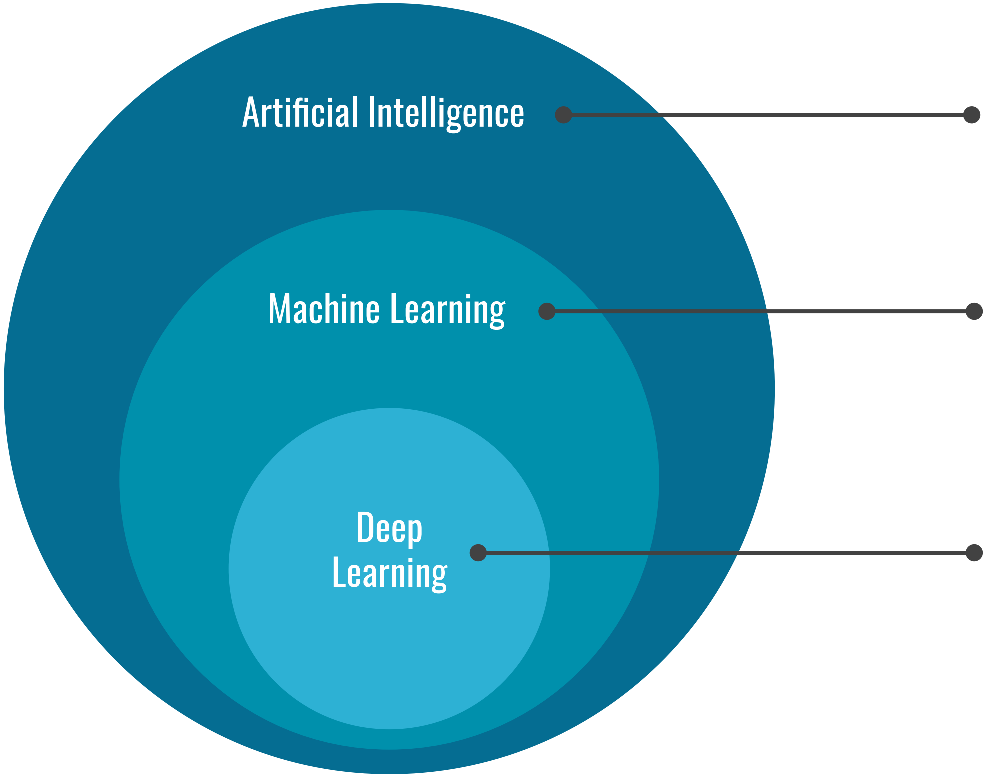
SAINT - 1961

Deep Blue - 1996



Dall-E - 2021

What is AI?



Models/programs with the following steps as objectives : **perception, reasoning and action.**

Models that learn by examples.

Subgroup of ML models that have a **multilayer neural network** architecture and need **many examples** to learn.

Challenges in AI - Model Complexity

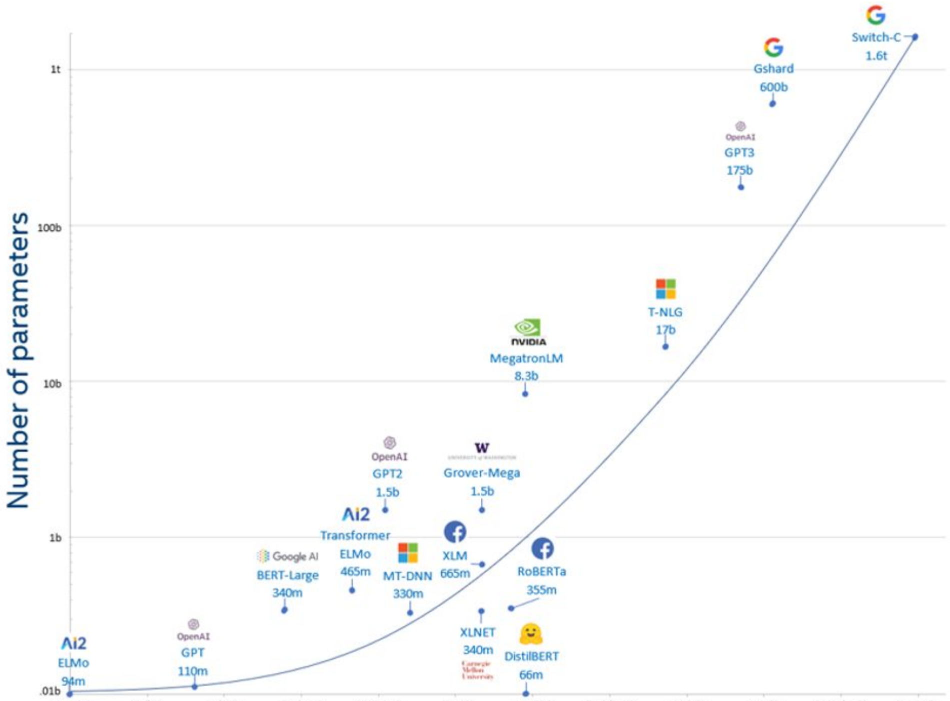
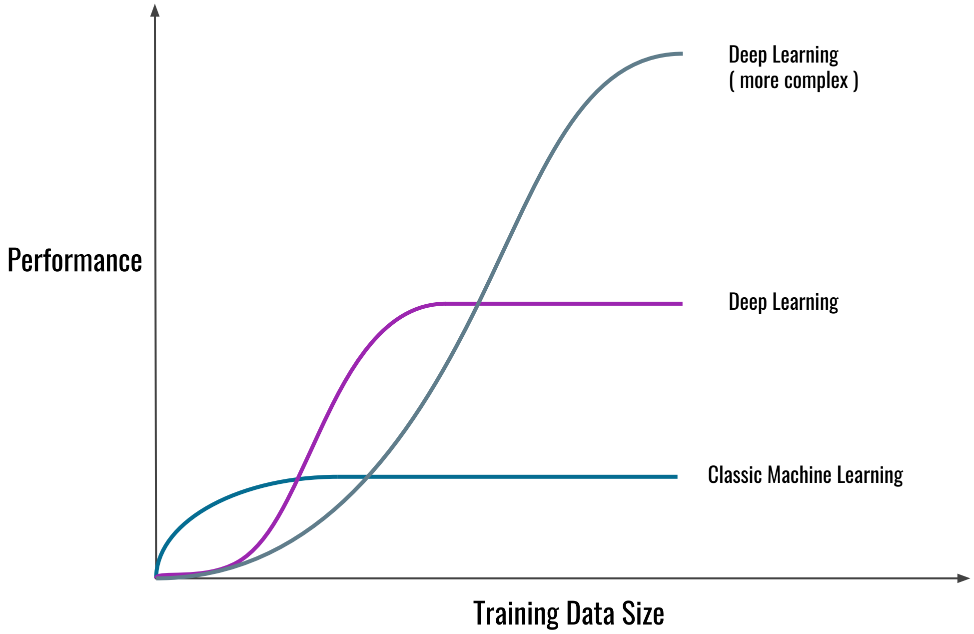


Figure 1: Exponential growth of number of parameters in DL models

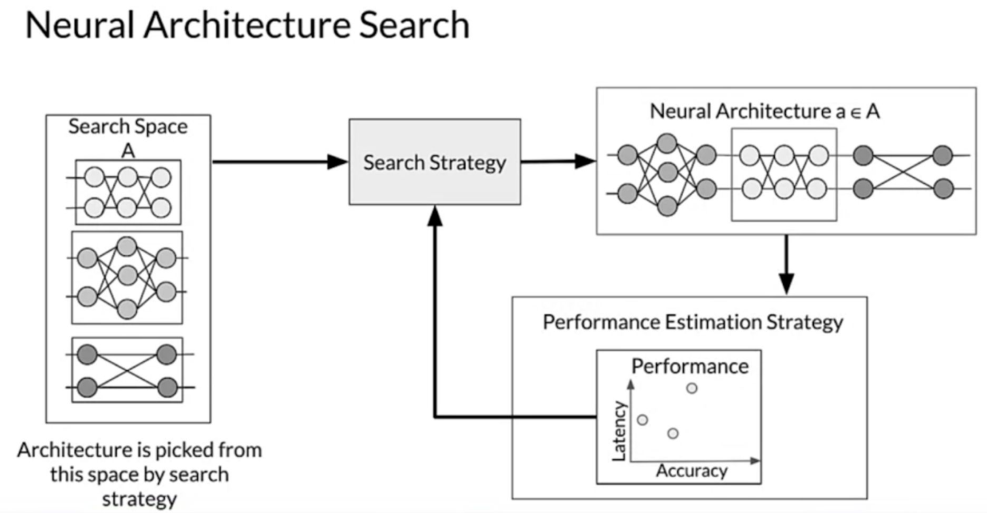
- Increasing number of model's neurons and layers have increased the number of training parameters exponentially.
- **Increasing operations** in forward and backpropagation steps

Challenges in AI - Data Quantity



- Complex models need **more training data** to achieve optimal performance
- Number of training images:
  - First Generation Generative models ~100 k
  - Dall-E ~ 400 M
- Need for **more epochs**
- More demanding ETL and Train stages

Challenges in AI - Model Optimization



- Search for a combination of hyperparameters that optimizes performance.
- Grid search, Random search, Bayesian search
- Increased model complexity leads to increased hyperparameters and architecture options. This substantially **increases the search space size**.

Source: Coursera's Machine Learning Modeling Pipelines in Production (Deeplearning.ai) course from the Machine Learning Engineering for Production (MLOps) Specialization

The need for HPC

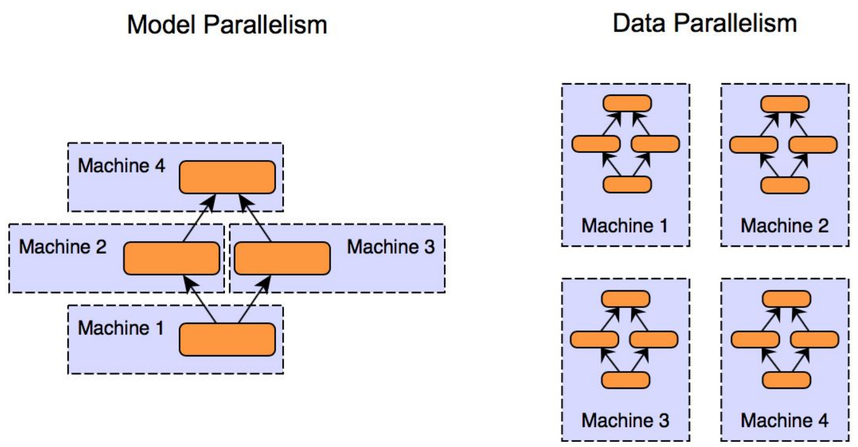
- With increased model complexity, training models become more time consuming:
  - More parameters to train
  - More data used in each epoch
  - Bigger search space for optimization
  - More epochs are needed
- State of the art DL models take several weeks to train
- Parallelism has been successfully used to:
  - Train models faster
  - Search for optimal hyperparameters combinations
  - Faster inference time

High Performance Modeling

Parallelism

Data Parallelism: models replicated into different accelerators (GPUs/TPUs) and dat split between them;

Model Parallelism: when models are too large to fit on a single device, they can be divided into partitions, each on a different accelerator



Reference: Intro Distributed Deep Learning

Source: Coursera's Machine Learning Modeling Pipelines in Production (Deeplearning.ai) course from the Machine Learning Engineering for Production (MLOps) Specialization

Parallelism

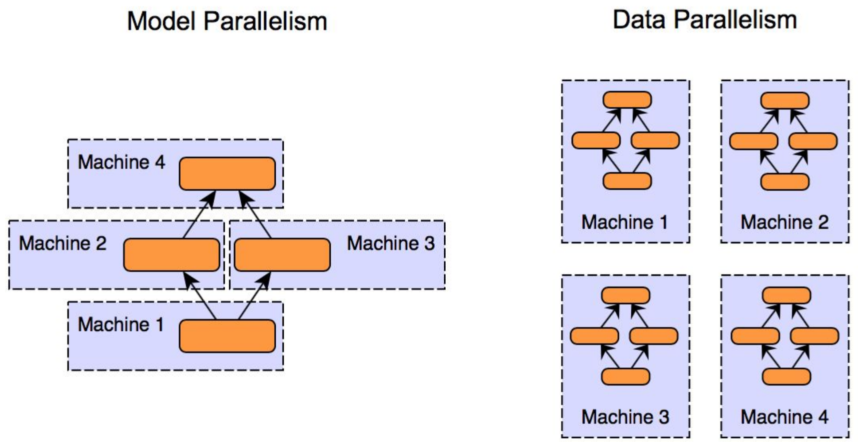
Distributed training using data parallelism

Synchronous training

- All workers train and complete updates in sync
- Supported via all-reduce architecture (cross-device communication)

Asynchronous Training

- Each worker trains and completes updates separately
- Supported via parameter server architecture
- More efficient, but can result in lower accuracy and slower convergence



Reference: Intro Distributed Deep Learning

Source: Coursera's Machine Learning Modeling Pipelines in Production (Deeplearning.ai) course from the Machine Learning Engineering for Production (MLOps) Specialization

Parallelism

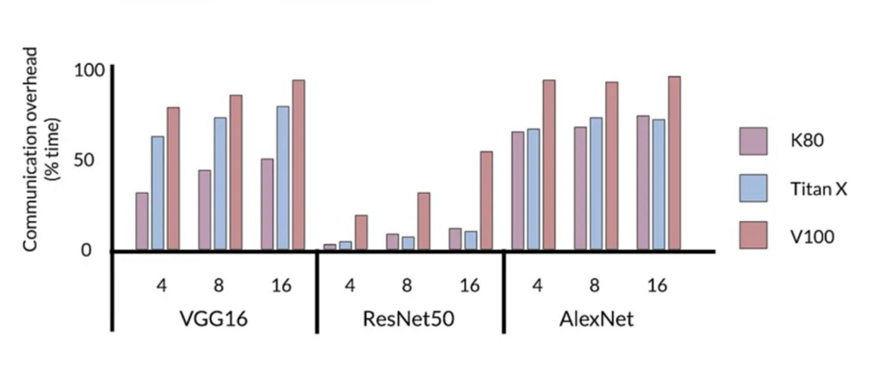
Challenges in Data Parallelism

Fault Tolerance

Failures in one worker would cause failure of distribution strategies;

Possible solution

Save training state and restore upon restart from job failure



(Lots of cross-communication/synchronization)

Source: Coursera's Machine Learning Modeling Pipelines in Production (Deeplearning.ai) course from the Machine Learning Engineering for Production (MLOps) Specialization

High Performance Ingestion

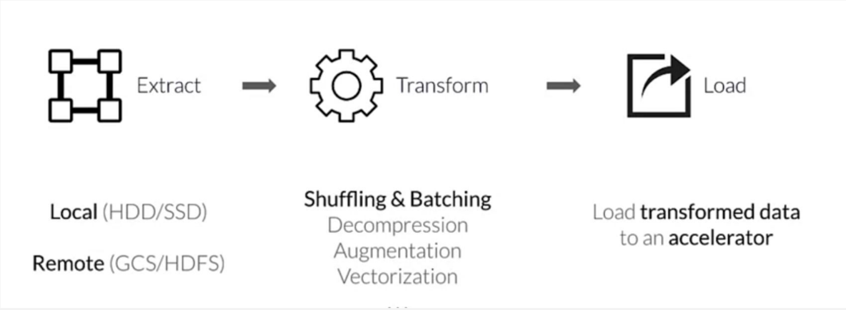
Input Pipelines

Accelerators are expensive. We need to keep then running (avoid under-utilization);

Avoid inefficiencies to make the most of the hardware available

Parallel processing of data, to maximize **compute, I/O and Network Resources**

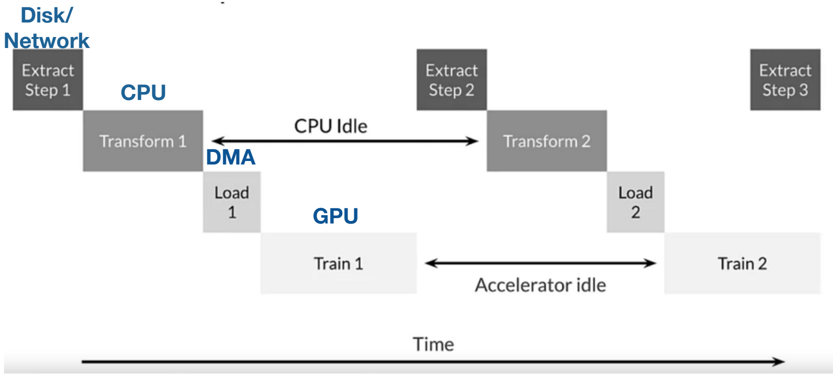
ETL (Extract, Transform, Load)



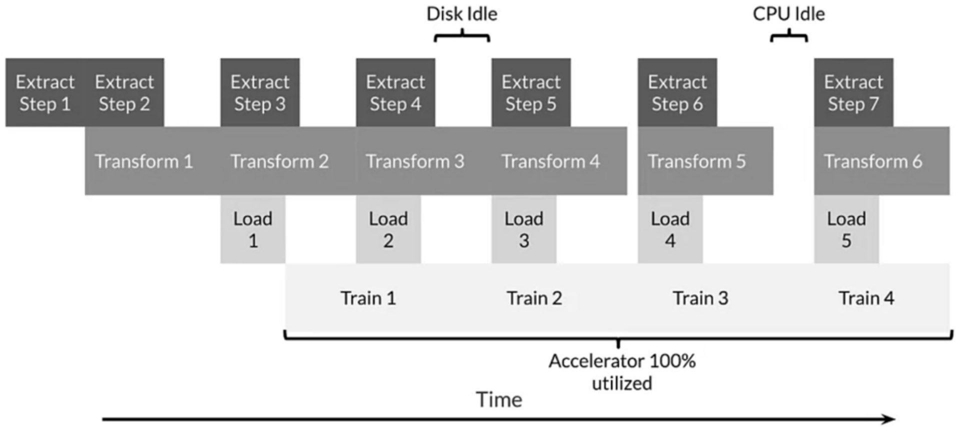
Source: Coursera's [Machine Learning Modeling Pipelines in Production (Deeplearning.ai)] course from the [Machine Learning Engineering for Production (MLOps) Specialization]

High Performance Ingestion

Inefficient ETL



Improved ETL



Source: Coursera's [Machine Learning Modeling Pipelines in Production (Deeplearning.ai)] course from the [Machine Learning Engineering for Production (MLOps) Specialization]

The Raise of Giant Models

Issues Training Huge Networks

- GPU Memory not Increasing fast as Networks;
- State of the art Image Models Reached Amount of Memory Available in Cloud TPUs;
- Need for large-scale Training of Giant Neural Networks.

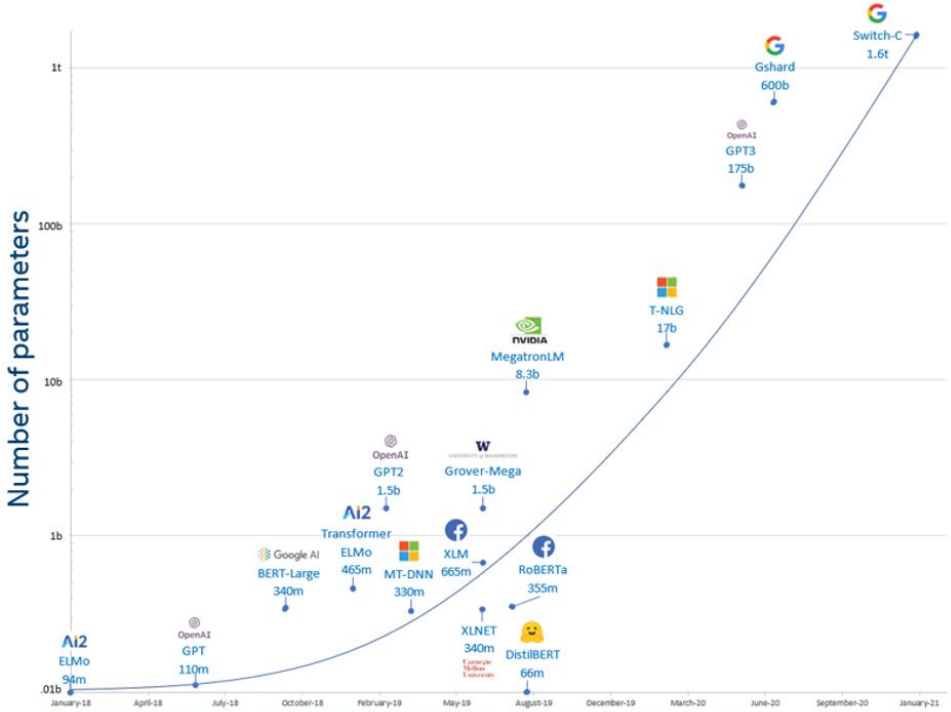


Figure 1: Exponential growth of number of parameters in DL models

Source: Coursera's [Machine Learning Modeling Pipelines in Production (Deeplearning.ai)] course from the [Machine Learning Engineering for Production (MLOps) Specialization]

The Raise of Giant Models

Overcoming Memory Constraints

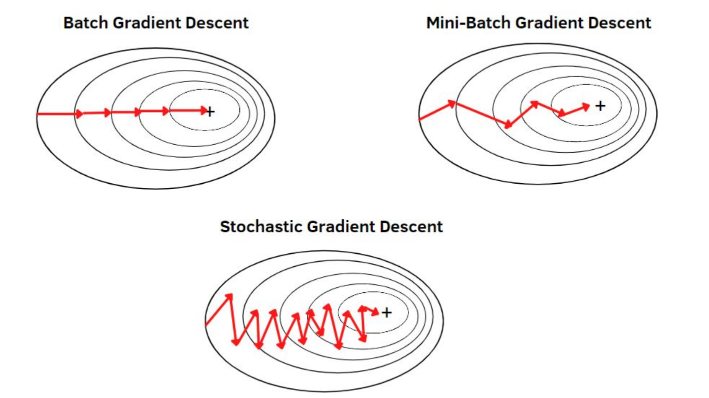
Gradient Accumulation (Mini-Batch)

Memory Swap
Copy activations to the CPU/Memory, than back to the Accelerator, and back and forth.

Data Parallelism

Model Parallelism

Pipeline Parallelism

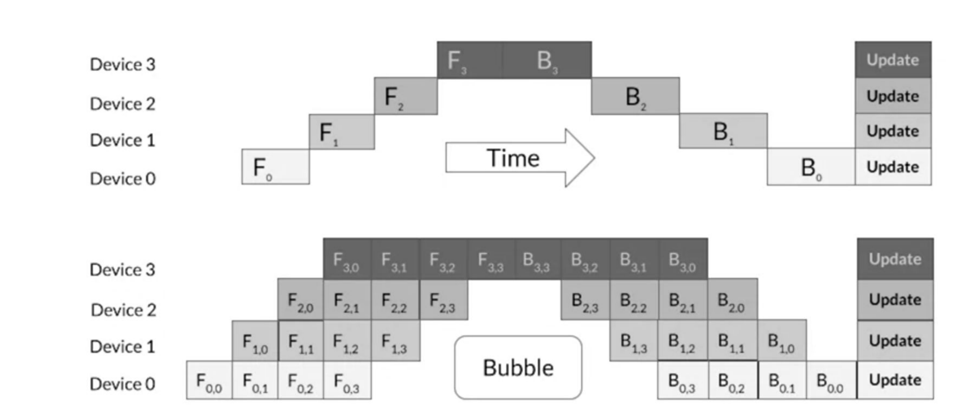


Source: Coursera's [Machine Learning Modeling Pipelines in Production (Deeplearning.ai)]() course from the [Machine Learning Engineering for Production (MLOps) Specialization]()

The Raise of Giant Models

Pipeline Parallelism

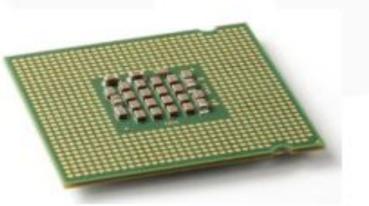
- Integrate both Model and Data Parallelism
- Divide Mini-batch into micro-batches
- Workers work on different micro-batches in parallel
- Allow models with large amount of parameters



Source: Coursera's [Machine Learning Modeling Pipelines in Production (Deeplearning.ai)] course from the [Machine Learning Engineering for Production (MLOps) Specialization]

HPC Hardware for AI

CPU and GPU for AI



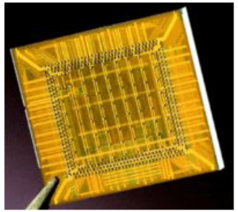
- The **C**entral **P**rocessing **U**nit (CPU) is a general-purpose processing unit with usually 4-16 cores.
- CPUs run complex tasks and facilitate system management.
- Versatility, ease of programming.
- Optimized for sequential processing with limited parallelism.
- Work well with mixed data inputs, such as systems that use both audio and text, and extract, transform, and load (ETL) processes.

- **G**raphics **P**rocessing **U**nits (GPUs) are highly parallel cores (100s or 1,000s) for high-speed graphics rendering.
- Originally designed for graphics; now used in a wide range of computationally intensive applications.
- They deliver high-performance processing, and typically have higher power consumption than CPUs.
- Facilitates both neural network training and AI inferencing.

Source: https://blog.adlinktech.com/2021/02/19/embedded-hardware-processing-ai-edge-gpu-vpu-fpga-asic/

FPGA and ASIC for AI

- **F**ield-**P**rogrammable **G**ate **A**rray (FPGA), which are configurable logic gates, consume less power than CPUs and GPUs.
- They can be the best choice when a high degree of flexibility is required.
- Decreased latency - larger memory bandwidths result in lower latency than GPUs.
- Relatively difficult to program.
- Poor performance for sequential operations; not good for floating-point operations



- **A**pplication-**S**pecific **I**ntegrated **C**ircuits (ASICs) are custom logic designed using a manufacturer's circuit libraries.
- Vision processing units (VPUs), image and vision processors, and co-processors.
- Tensor processing units (TPUs), such as the first TPU developed by Google for its machine learning framework, TensorFlow.
- Neural compute units (NCUs), including those from ARM.
- Longest development time; high cost; cannot be changed without redesigning the silicon.

Source: https://blog.adlinktech.com/2021/02/19/embedded-hardware-processing-ai-edge-gpu-vpu-fpga-asic/

Types of Computing Device for AI

Applications	CPU	FPGA	GPU	ASIC	Comments
Vision & image processing		✓	✓	✓	FPGA may give way to ASIC in high-volume applications
AI training			✓		GPU parallelism well-suited for processing terabyte data sets in reasonable time
AI inference	✓	✓	✓	✓	Everyone wants in! FPGAs perhaps leading; high-end CPUs (e.g., Intel's Xeon) and GPUs (e.g., Nvidia's T4) address this market
High-speed Search	✓	✓	✓	✓	Microsoft's Bing uses FPGAs; Google uses TPU ASIC; CPU needed for coordination & control
Industrial motor control	(✓)	✓		✓	Many motor-control MCUs and ASICs available; FPGAs offer a quick-turn ASIC alternative
Supercomputer HPC	✓		✓		Majority of TOP500 supercomputers uses some combination of CPUs and GPUs
General-purpose computing	✓		(✓)		CPU most versatile, flexible option; GPUs beginning to perform some tasks
Embedded control	✓	✓		✓	CPUs (→ MCU) dominant in low-cost, space-constrained, low-power, mobile applications
Prototyping, low-volume		✓			FPGAs best choice for low-volume, high-end applications; also pre-silicon validation, post-silicon validation and firmware development

Source: https://www.arrow.com/en/research-and-events/articles/fpga-vs-cpu-vs-gpu-vs-microcontroller

How AI Researchers are comparing Hardware?

MLPerf is a consortium of AI leaders from academia, research labs, and industry whose mission is to “build fair and useful benchmarks” that provide unbiased evaluations of training and inference performance for hardware, software, and services—all conducted under prescribed conditions.

Image Classification

Assigns a label from a fixed set of categories to an input image, i.e., applies to computer vision problems. details.

Object Detection (Lightweight)

Finds instances of real-world objects such as faces, bicycles, and buildings in images or videos and specifies a bounding box around each. details.

Object Detection (Heavyweight)

Detects distinct objects of interest appearing in an image and identifies a pixel mask for each. details.

Biomedical Image Segmentation

Performs volumetric segmentation of dense 3D images for medical use cases. details.

Automatic Speech Recognition (ASR)

Recognize and transcribe audio in real time. details.

Natural Language Processing (NLP)

Understands text by using the relationship between different words in a block of text. Allows for question answering, sentence paraphrasing, and many other language-related use cases. details.

Recommendation

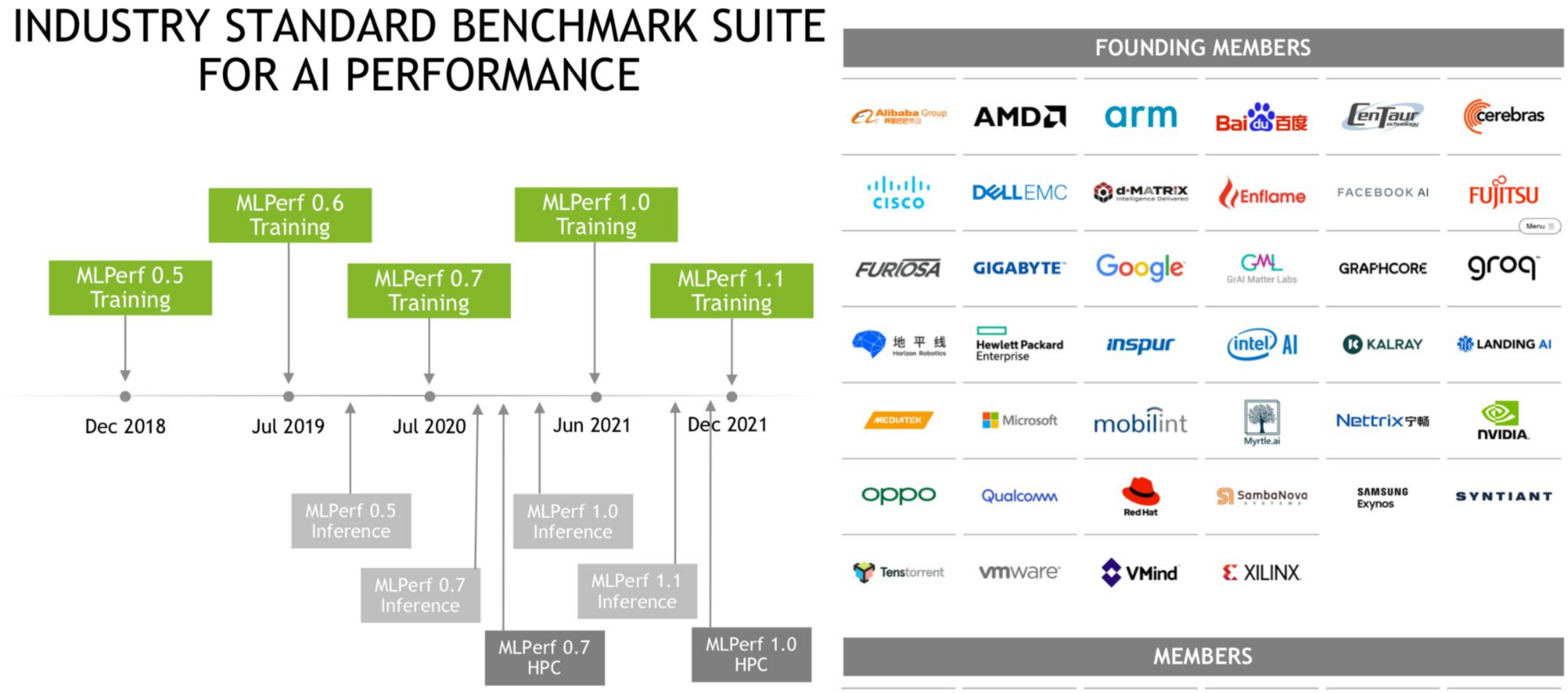
Delivers personalized results in user-facing services such as social media or e-commerce websites by understanding interactions between users and service items, like products or ads. details.

Reinforcement Learning

Evaluates different possible actions to maximize reward using the strategy game Go played on a 19x19 grid. details.

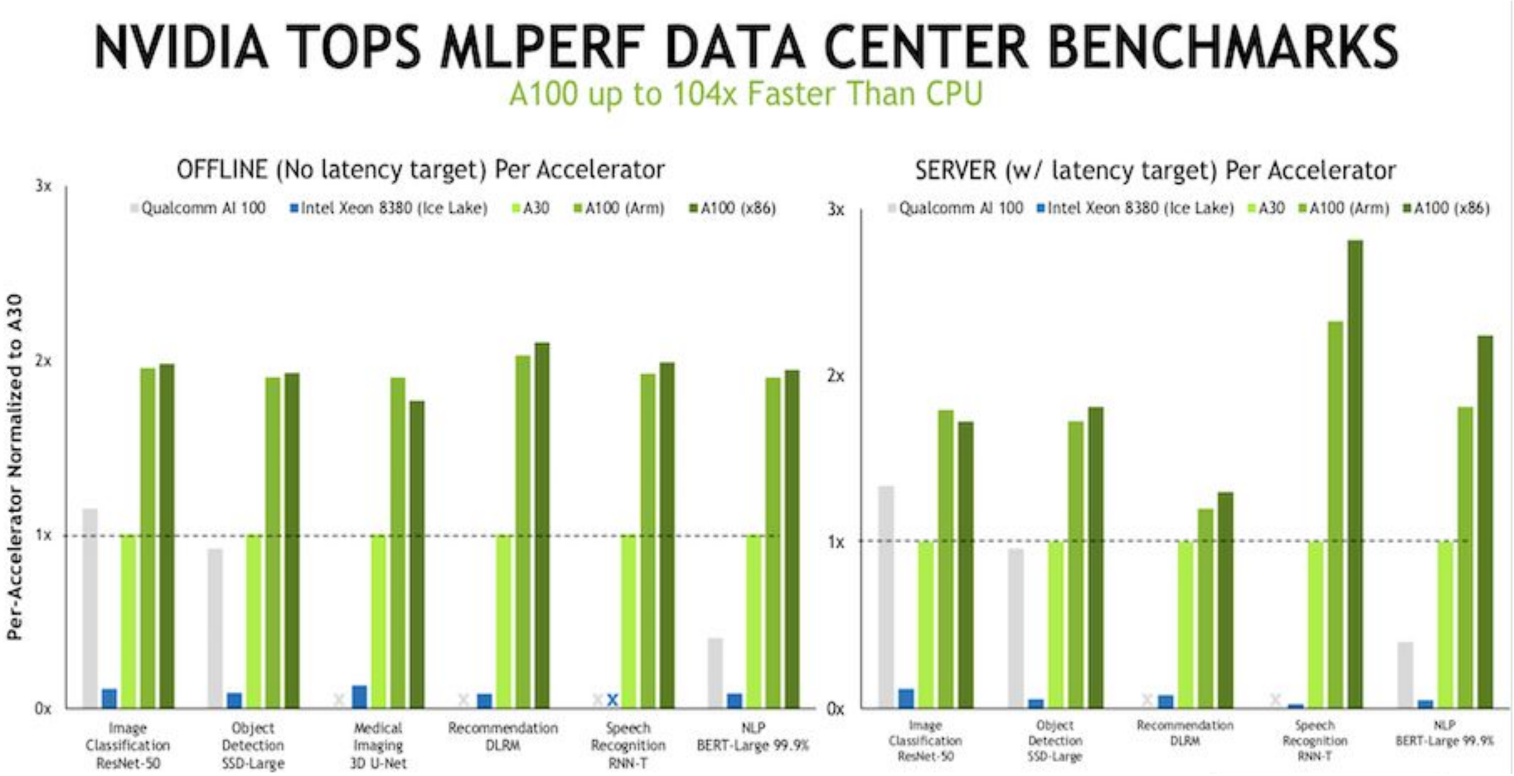
Source:
https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/#:~:text=MLPerf%20is%20a%20consortium%20of,all%20conducted%20under%20prescribed%20conditions.

How AI Researchers are comparing Hardware?



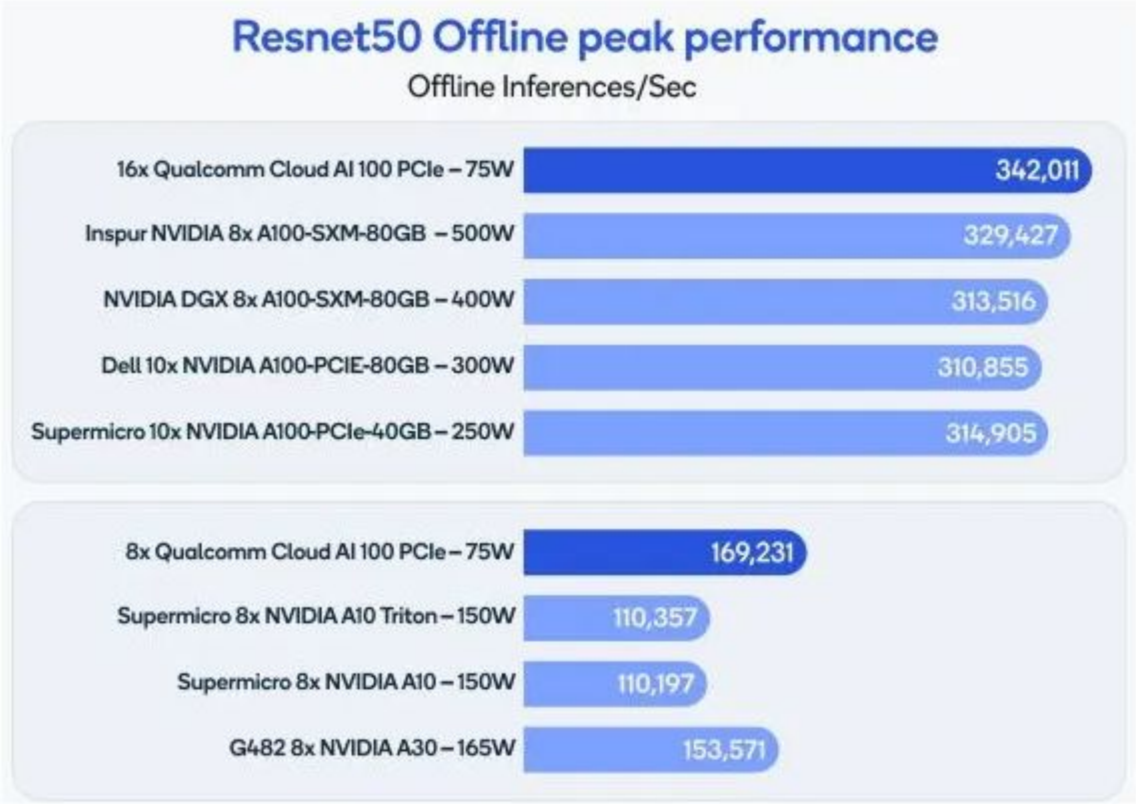
Source:
https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/#:~:text=MLPerf%20is%20a%20consortium%20of,all%20conducted%20under%20prescribed%20conditions.

How AI Researchers are comparing Hardware?



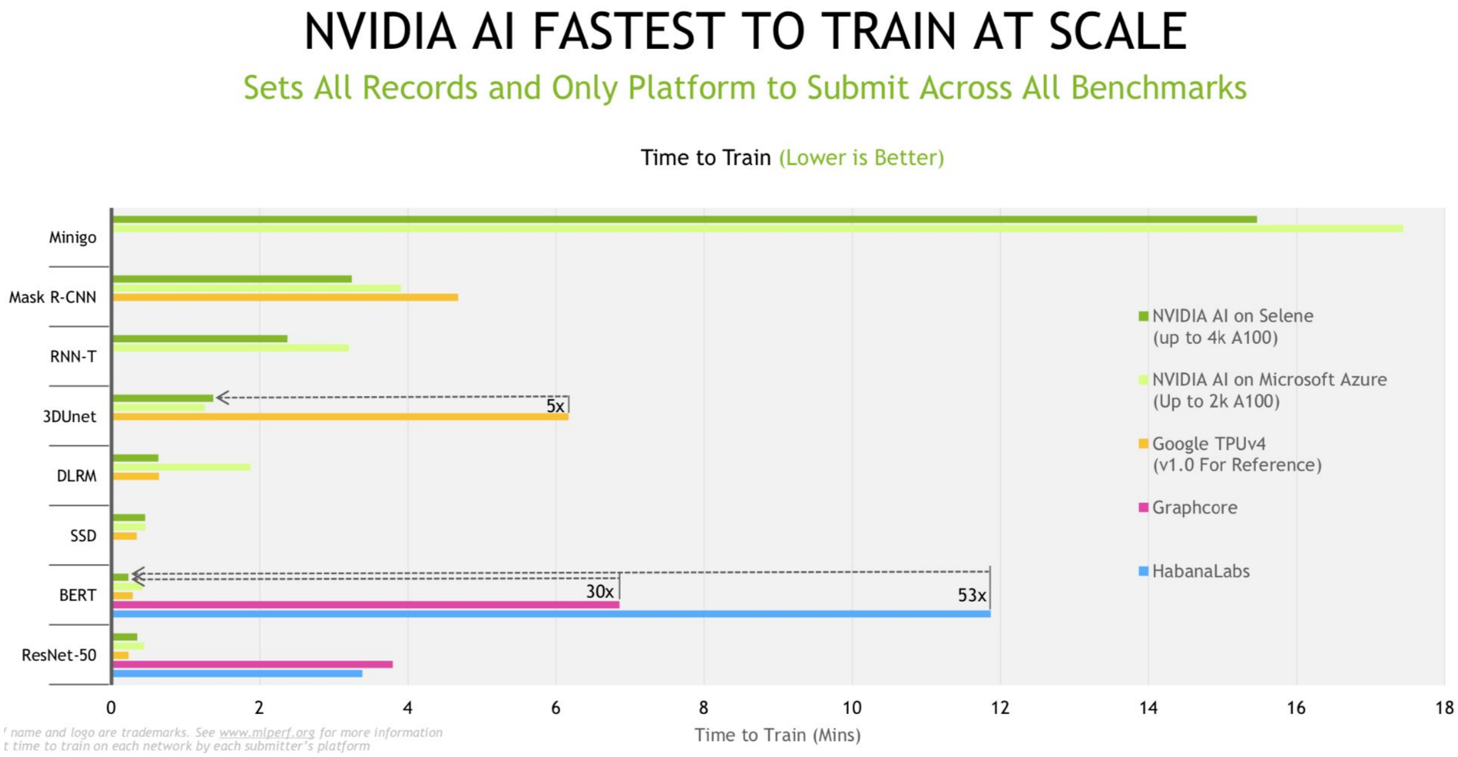
Source: https://www.hpcwire.com/2021/09/22/the-latest-mlperf-inference-results-nvidia-gpus-hold-sway-but-here-come-cpus-and-intel/#foobox-1/0/Nvidia_Mlperf_Datacenter.png

How AI Researchers are comparing Hardware?



Source: https://www.eetasia.com/qualcomm-takes-on-nvidia-for-mlperf-inference-title/

How AI Researchers are comparing Hardware?



Source:
https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/#:~:text=MLPerf%20is%20a%20consortium%20of,all%20conducted%20under%20prescribed%20conditions.

Thank you!