Green AI

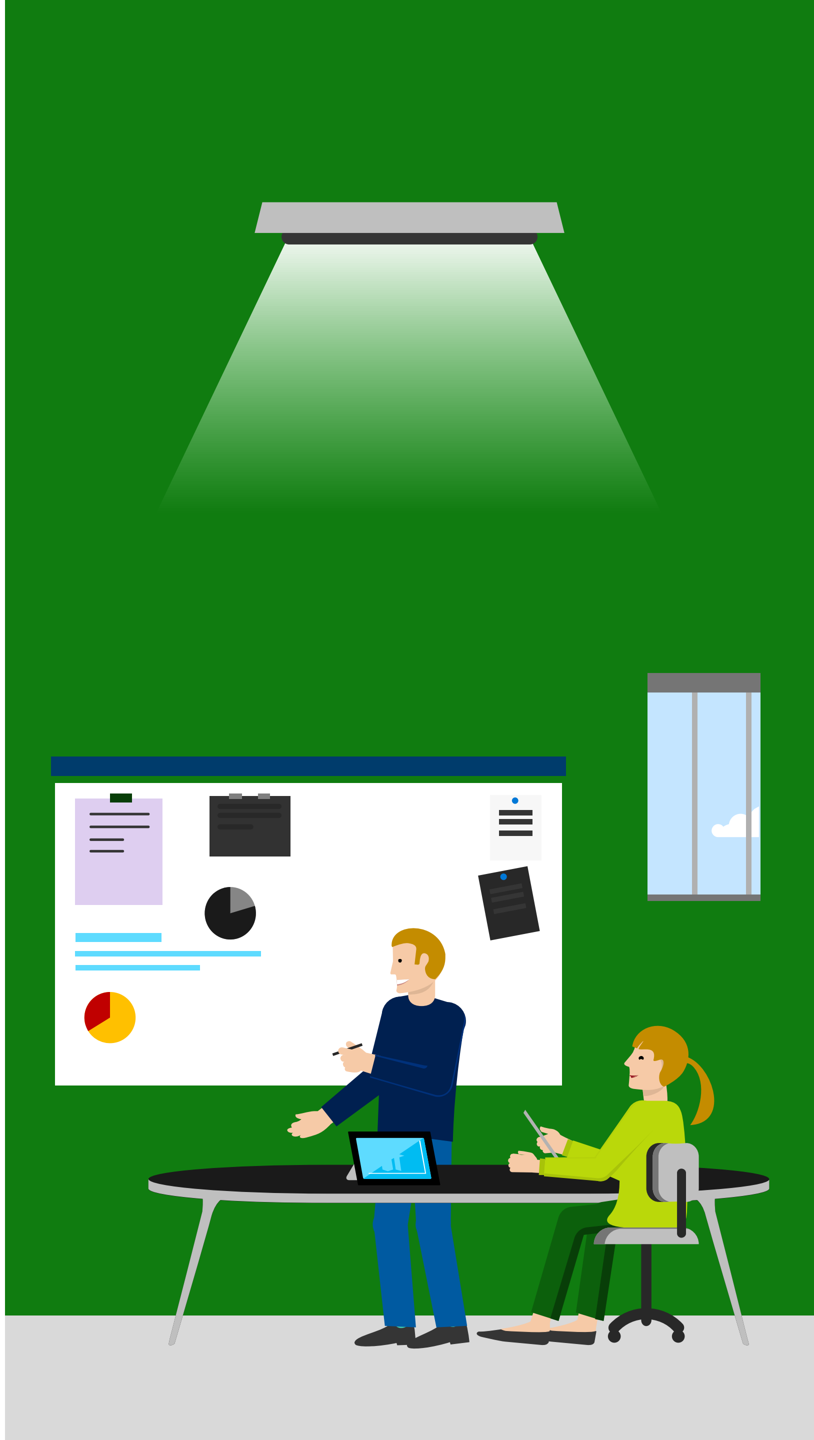
How to reduce the carbon footprint of ML workflows?

Agenda

1 Why Green AI?

2 What is Green AI?

IMPACT AI Understand – Measure - Reduce



Quick Overview

Artificial Intelligence

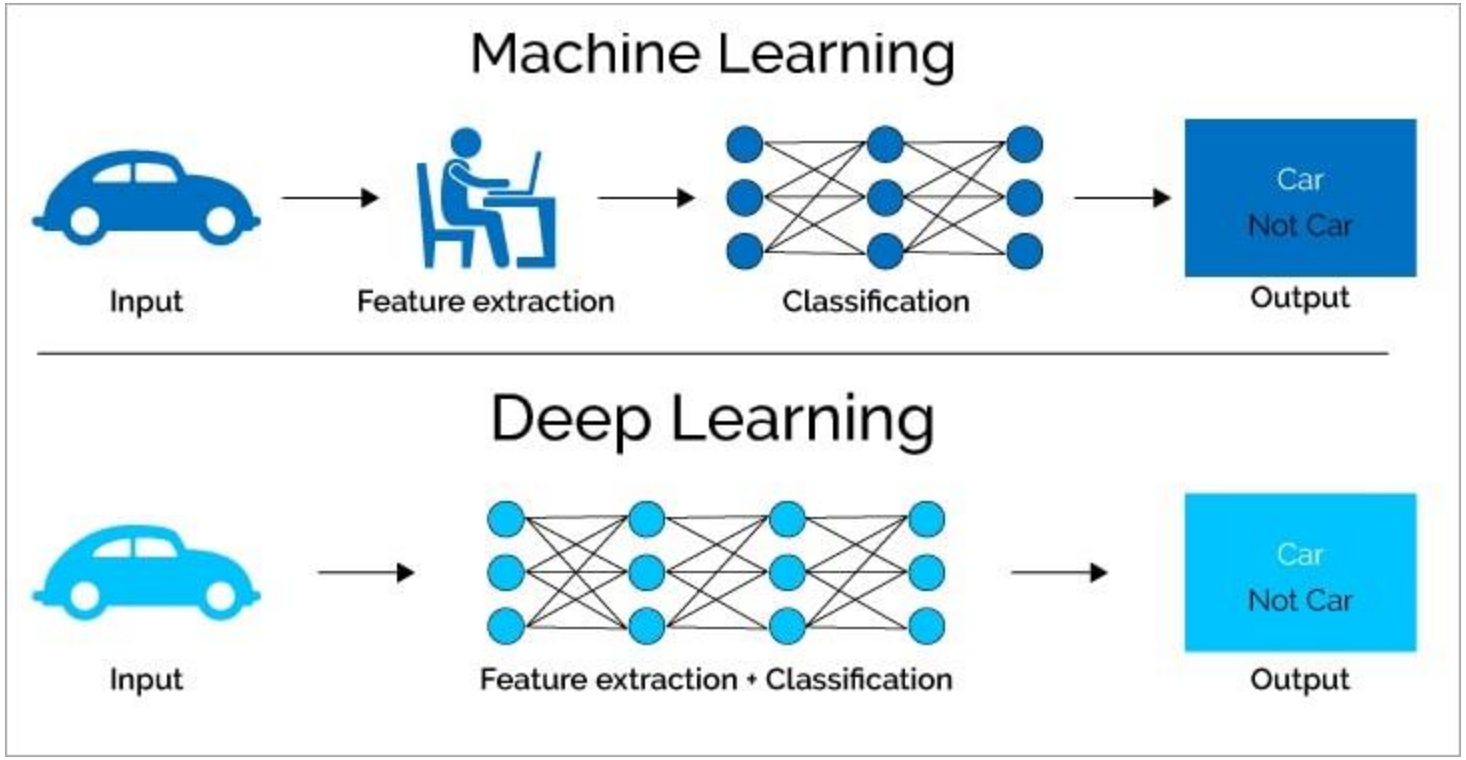
Any technique that enables computers to mimic human intelligence. It includes *machine learning*

Machine Learning

A subset of AI that includes techniques that enable machines to improve at tasks with experience. It includes *deep learning*

Deep Learning

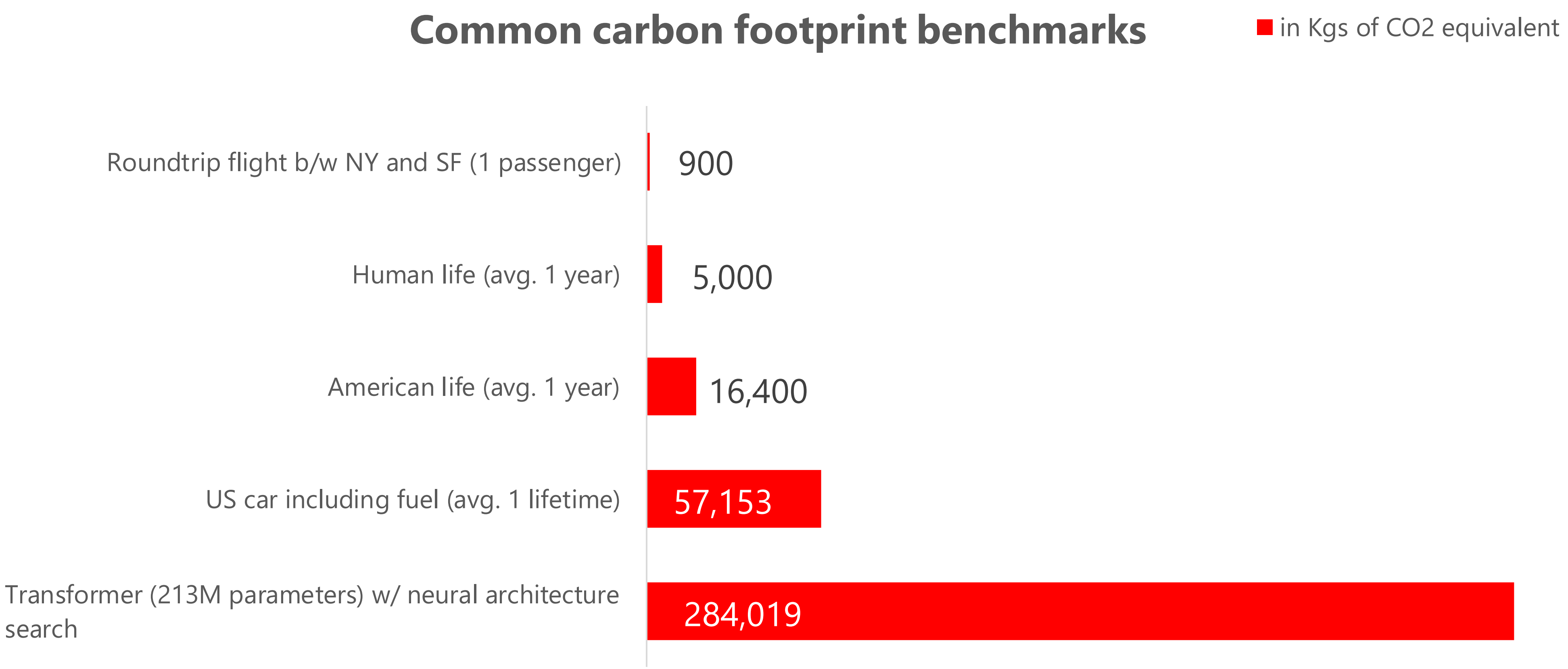
A subset of machine learning based on neural networks that permit a machine to train itself to perform a task.



Why Green AI?

The Staggering Cost of AI

- Computational costs have increased **300.000X** from 2012 to 2018
- Only **11%** of firms are seeing a 'significant' ROI on their AI workloads(wired)
- GPT3 training emits as much carbon as **3X** round-trip transcontinental flights (SF<>NYC)

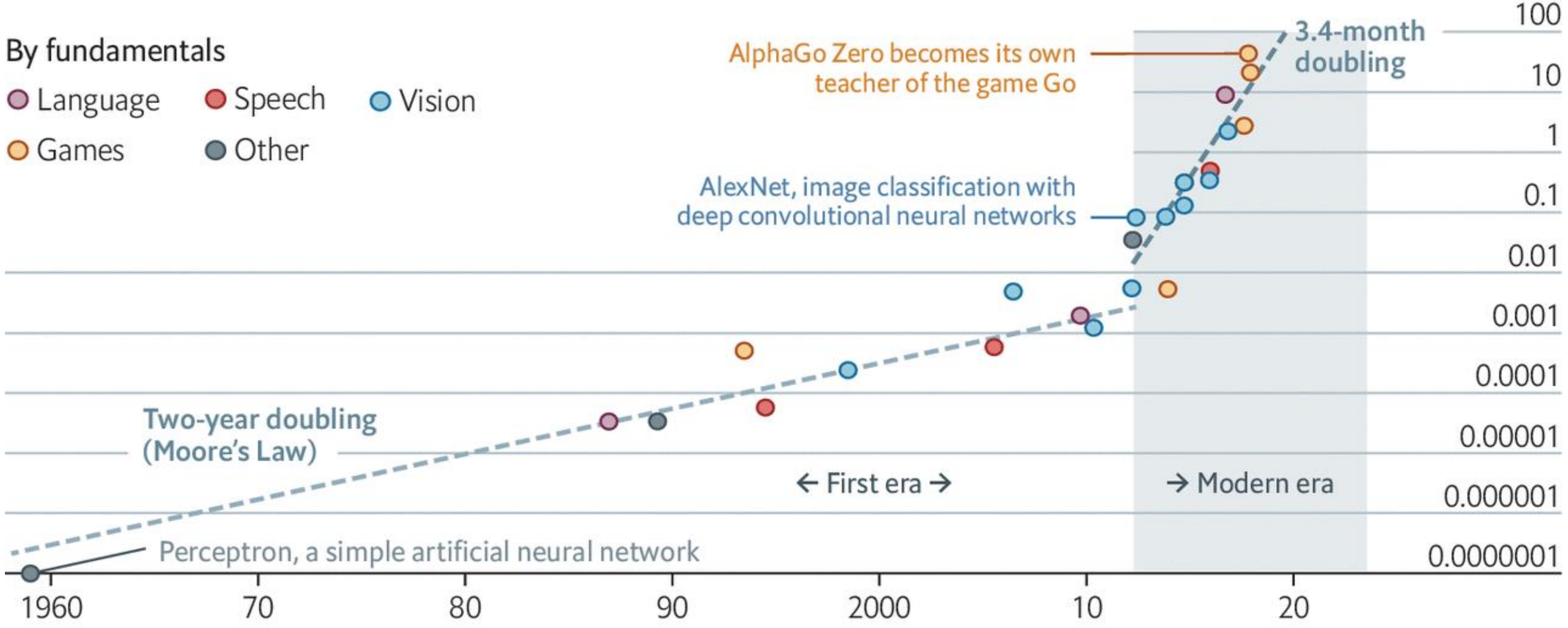


Source: Strubell

Deep and steep

Computing power used in training AI systems

Days spent calculating at one petaflop per second*, log scale

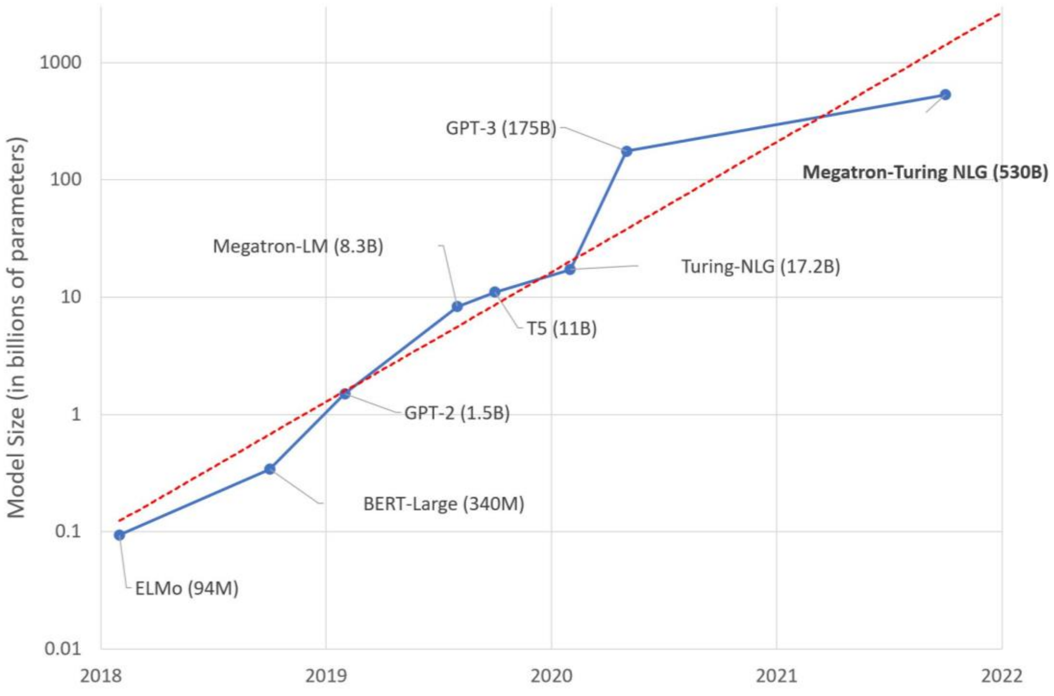


Source: OpenAI

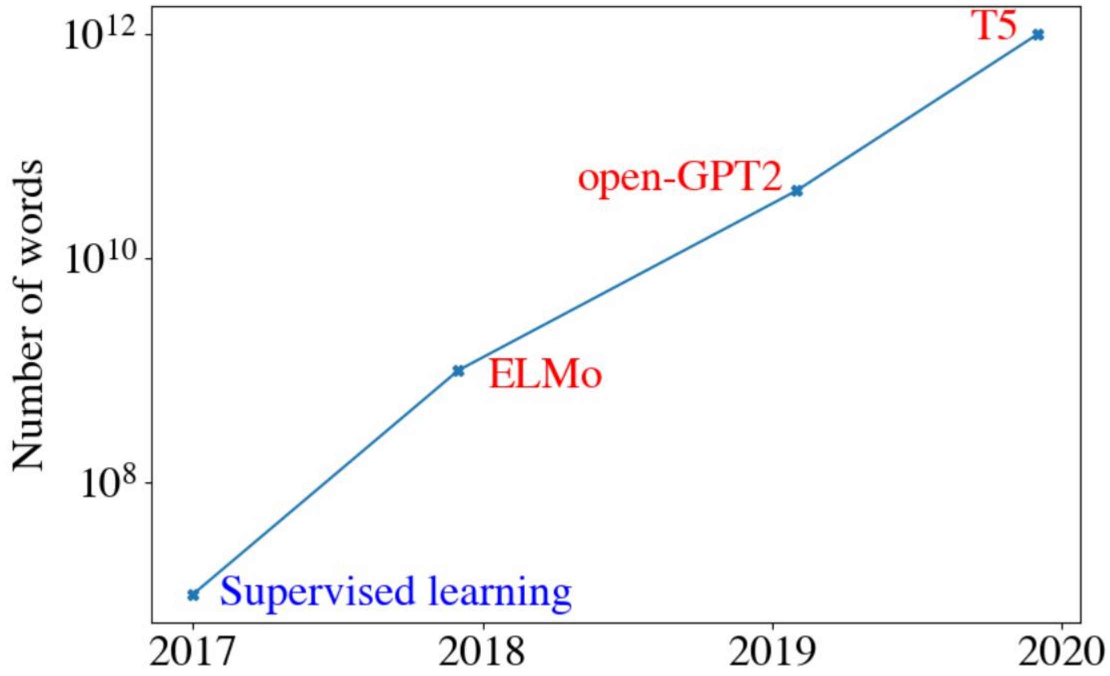
*1 petaflop=10^{15} calculations

The Economist

Bigger Models – Larger Datasets

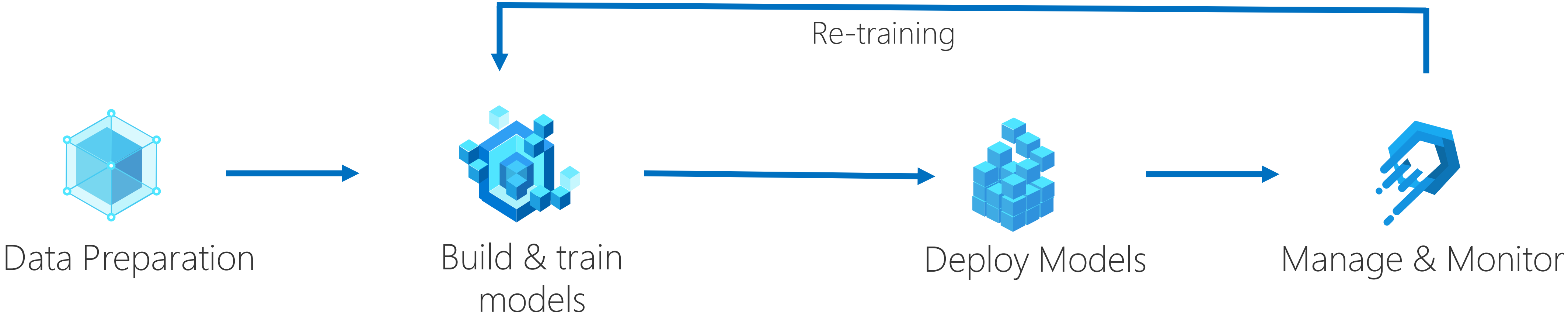


Parameter counts of several recently released pretrained language models



More Data 100.000x in 3 Years!

AI and Power



Hardware and power needed:

DL training requires specialized GPU hardware
GPUs are power-hungry (often 250-350W)
Inferences may use GPUs, FPGAs or CPUs (typical CPU ~135W, some up to 280W)

Problems with Big Models

NCE MACHINE LEARNING PROGRAMMING VISUALIZATION AI PICKS MC

Too big to deploy: How GPT-2 is breaking servers

A look at the bottleneck around deploying massive models to production

Caleb Kaiser Follow
Jan 31 · 7 min read

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

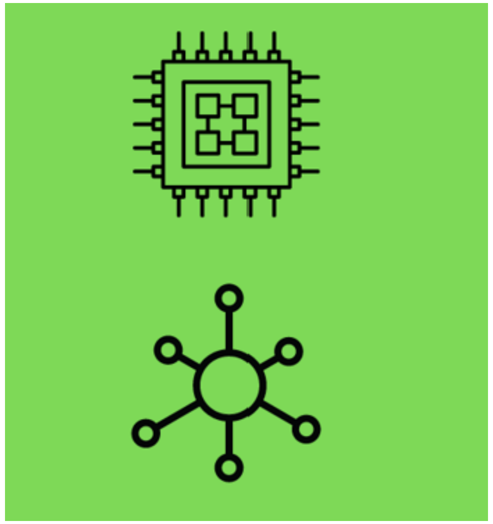
Inclusiveness

Adoption

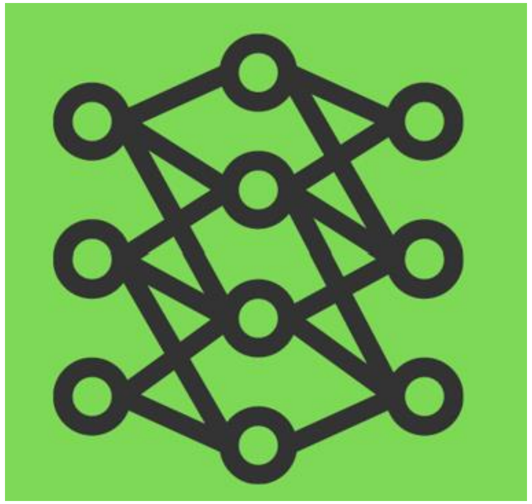
Environment

What is Green AI?

What is green AI?



Alternate deployment strategies



Elevate smaller models

Carbon-efficiency
and
Carbon awareness

The Imperative for Sustainable AI Systems (thegradient.pub)

Understand – Emission scopes

Greenhouse Gas Emissions (GHG) measurement

Scope
1

Direct emissions **created by your activities,** like consumption of gas, fuel oil or even leaks of refrigerants, present in the cooling and air conditioning circuits of data centers in particular

Scope
2

Indirect emissions from the production of electricity or heat you use to power buildings or processes

Scope
3

Indirect value chain emissions from all other activities in which you're engaged.
- Manufacturing, delivery and end of life of IT equipment related to the training and production of AI and edge equipment on which AI is deployed
- Purchases of technical and IT services and services dedicated to AI projects (software license, outsourcing, etc.)
- Use of the products / services targeted by the AI project

Understand - Considerations

All 3 scopes of GHG emissions

Entire life cycle of AI, from ideation and design to inference.

Impact of all the infrastructures and services associated with the AI project.

Include the Green AI approach within a more global Green IT approach.

Carbon is not only environmental impact of AI

Green AI - impact-ai

Peter Drucker "You can't manage what you can't measure"

Measure – ML lifecycle cost metrics framework

- Training: ~12% of models makes it to production
- Inference: 80-90% of carbon cost (NVIDIA)

Cost Metric	Training	Inference
Dollars	Jobs/pipelines	Operational Cost
Runtime	Core-seconds by SKU	Core-seconds by SKU
Energetic	GPU energy	GPU energy
Utilization	GPU Utilization (%) GPU Memory Utilization (%)	GPU Utilization (%) GPU Memory Utilization (%)

Operational Lifecycle Analysis Monitoring

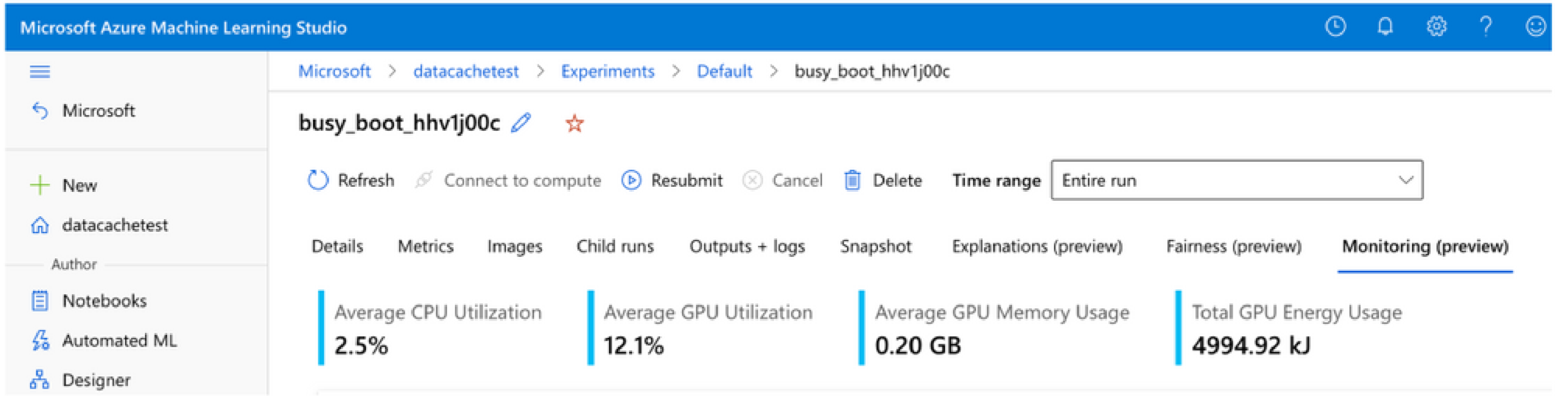
- Monitoring Capabilities: training/inference for cost ($, energy, carbon)
- Tools: Cost/benefit tradeoffs to optimize ROI

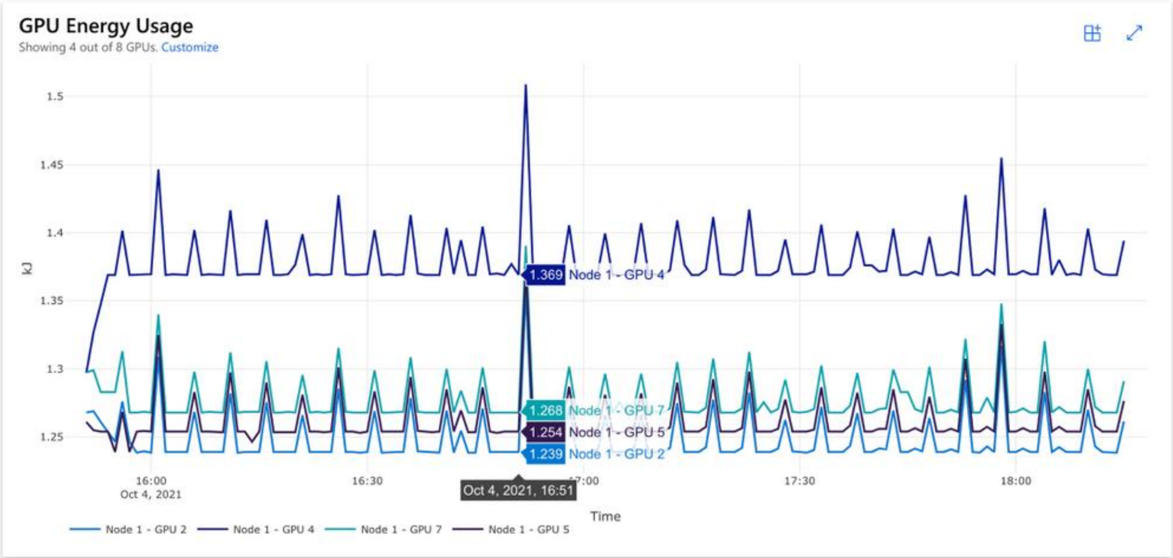
Measure – Training energy in Azure ML

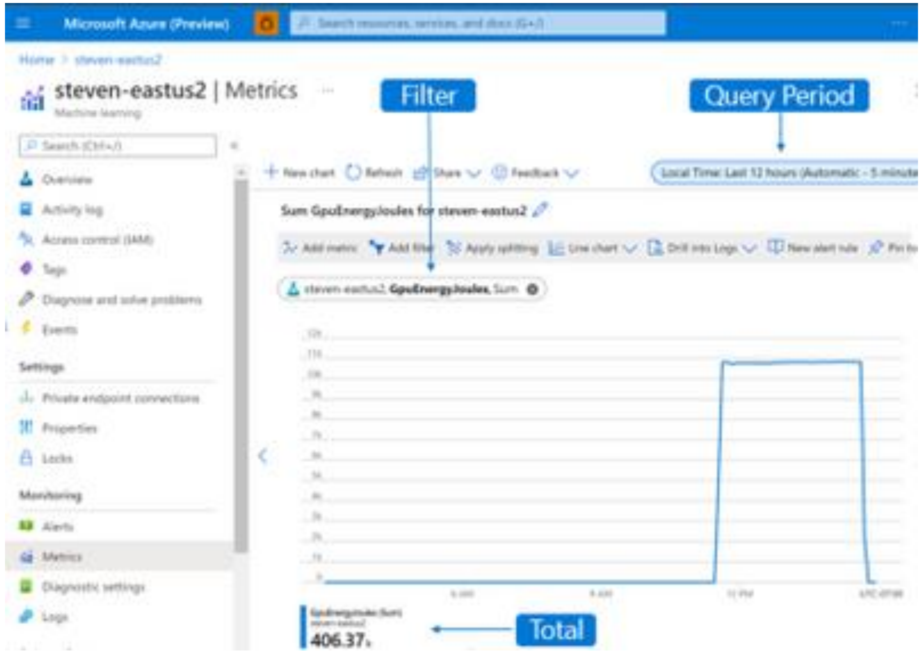
Opt-in metrics now let users sort to find the most expensive jobs & pipelines

- **Energy:** GPU energy consumed per job/pipeline (also avail in Azure Monitor)
- **Utilization**: GPU utilization, memory

Measure – Energy in Azure ML







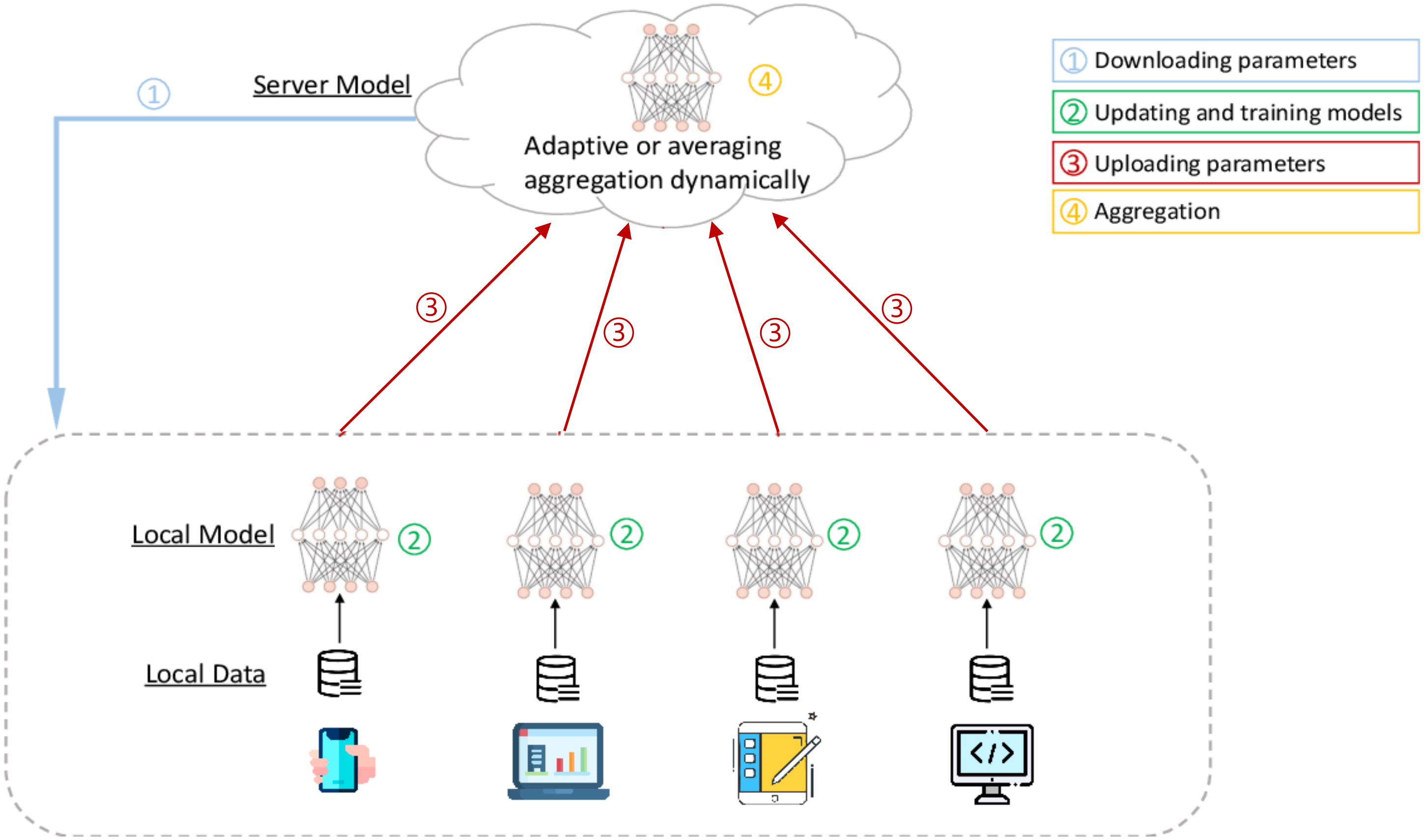
Reduce – Attenuate hardware impact

1. Use specialized hardware like ASICs/ FPGA to accelerate the run times of these jobs
2. Obtain higher utilization rates on existing hardware
3. Optimizing the use of existing hardware like general-purpose CPUs → reduce the demand for manufacturing new hardware

Scenarios & configurations on Azure	Supported DNN models	Regional support
+ Image classification and recognition scenarios + TensorFlow deployment (requires Tensorflow 1.x) + Intel FPGA hardware	- ResNet 50 - ResNet 152 - DenseNet-121 - VGG-16 - SSD-VGG	- East US - Southeast Asia - West Europe - West US 2

Reduce – Federated learning

Collaborative machine learning without centralized training data



https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Reduce – Federated learning

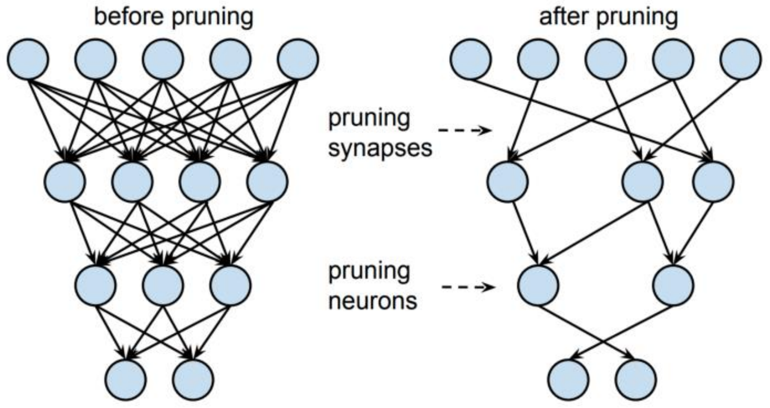
Country/CO2(g)	**V100** $PUE = 1.67$	**K80**	**V100** $PUE = 1.11$	**K80**	**FL** IID	**FL** non-IID
USA	1.6	5.2	1.1	3.5	0.5	1.0
China	2.9	9.2	1.9	6.2	0.9	1.7
France	0.2	0.8	0.2	0.5	0.1	0.1

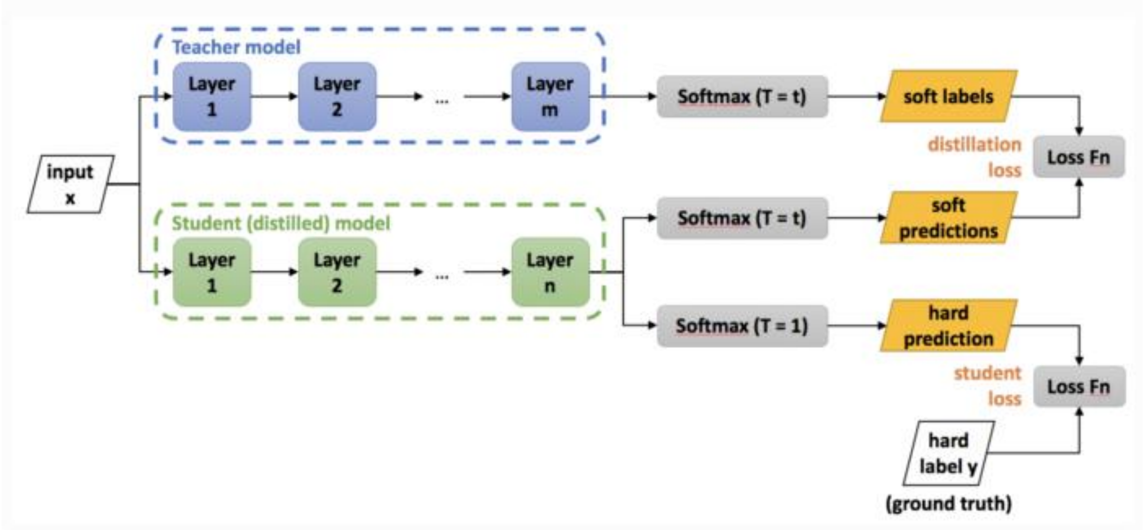
Table 4: CO_2 emissions (expressed in grams, *i.e.* **lower is better**) for centralized training and FL on Fashion-MNIST. Emissions are calculated once the top-1 accuracy on the test set reaches 90%. The number of epoch reported on the FL column relates to the number of local epoch done per client. “IID” and “non-IID” terms are employed to distinguish between clients that have an evenly distributed set of samples containing all the classes (IID) and clients that have more samples of certain classes (non-IID).

Can federated learning save the planet?

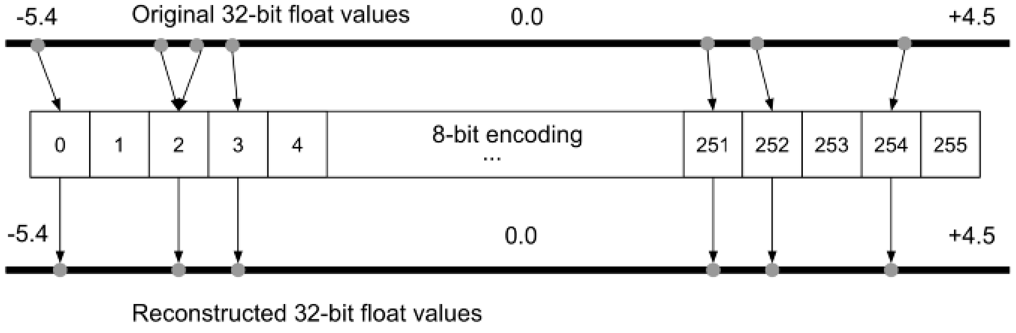
Reduce – Elevating smaller models

Pruning



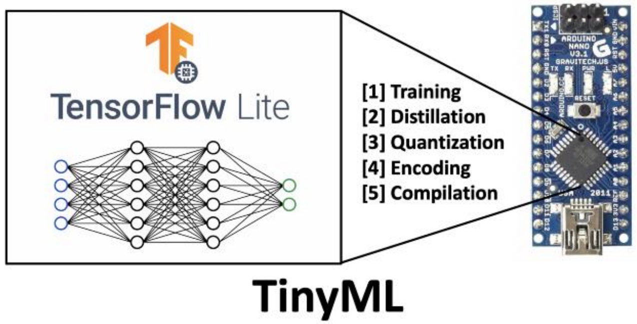


Knowledge distillation Source: ArXiv.



Quantization & Factorization

Reduce – Tiny ML



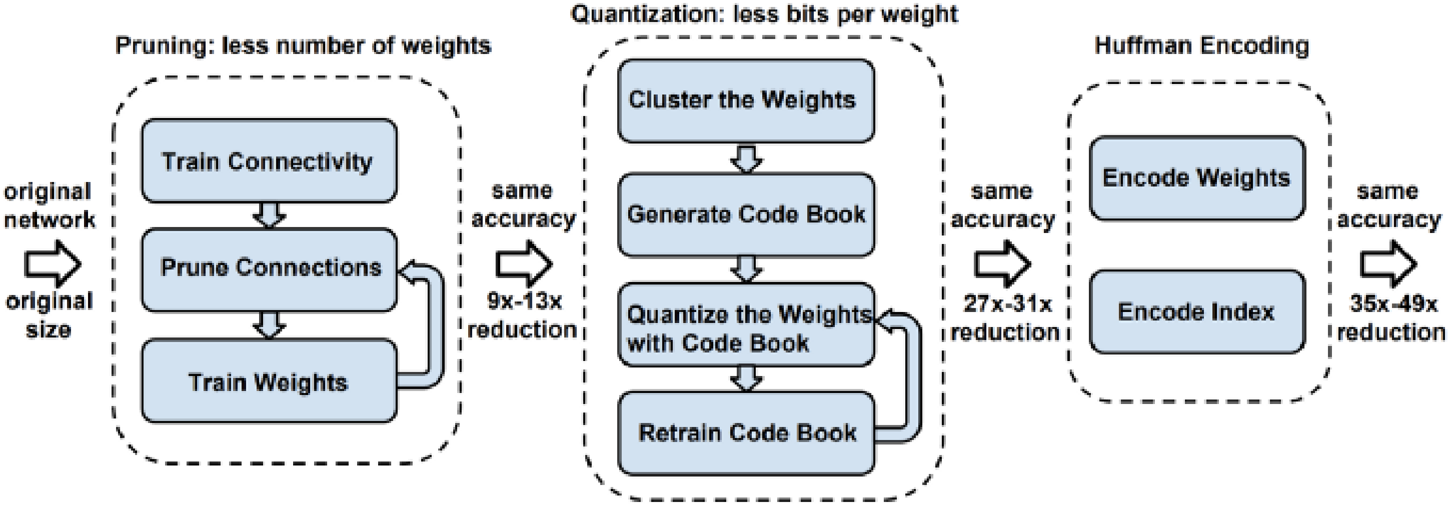


Diagram of the deep compression process. Source: ArXiv.

Reduce - Efficient model training

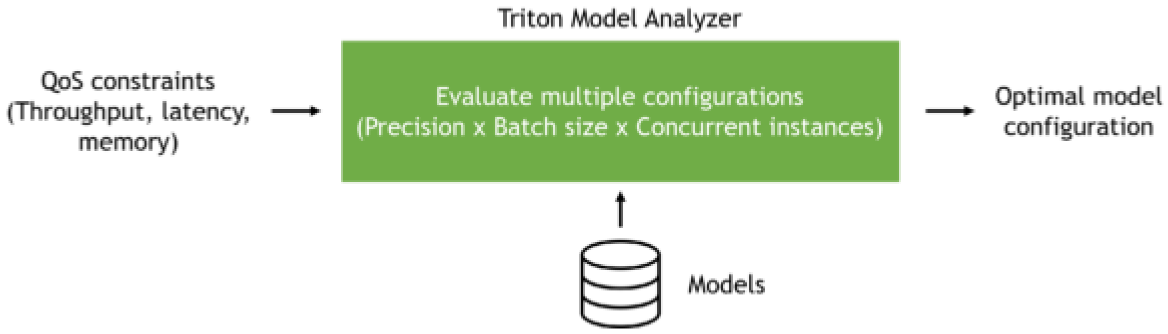
- Use feature stores

- Training refined into two stages for LLM:
  - Pre-training of a general model
  - Fine-tuning to produce accurate outcomes on a specific task

- Neural architecture search (NAS) and Hyperparameter Optimization (HPO) can also be used to satisfy different objective functions, such as computational efficiency or cost.

- muTransfer, that can transfer training hyperparameters across model sizes
  Enables equivalent accuracy levels while using at least an order of magnitude (~10x) less compute, with no limit to the efficiency gain as the target model size grows.

Reduce – Efficient inferencing



Overview of NVIDIA Triton Model Analyzer

Online Result Summary

Model: bert-large

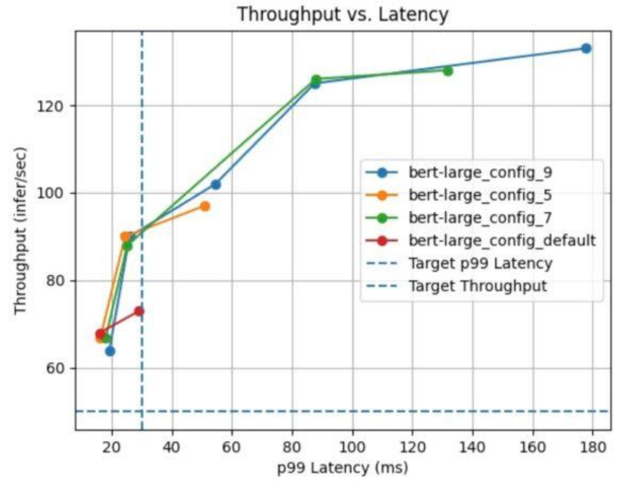
GPU(s): A100-SXM4-40GB

Total Available GPU Memory: 39.6 GB

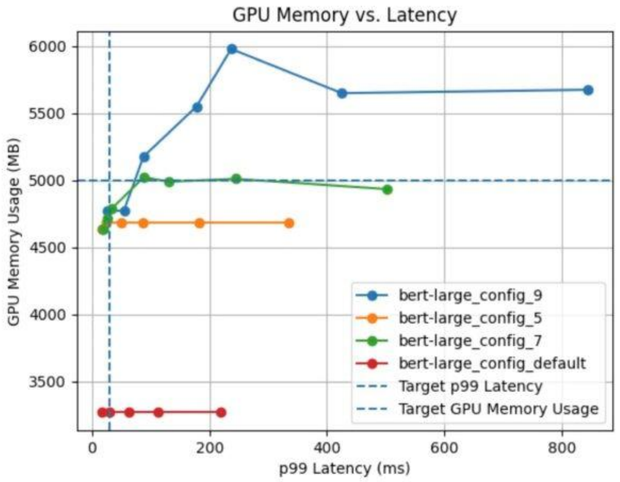
Constraint targets: Min Throughput : 50 infer/sec, Max p99 Latency : 30 ms, Max GPU Memory Usage : 5000 MB

In 161 measurement(s), config bert-large_config_9 (2/GPU model instance(s) with max batch size of 16 and dynamic batching enabled) on platform pytorch_libtorch delivers maximum throughput under the given constraints on GPU(s) A100-SXM4-40GB.

Curves corresponding to the 3 best model configuration(s) out of a total of 23 are shown in the plots.



Throughput vs. Latency curves for 3 best configurations.



GPU Memory vs. Latency curves for 3 best configurations.

Think Green

- Red AI
  - Big models, large datasets
  - Inclusiveness, adoption, environment

- Green AI
  - Enhance reporting of computational budget
  - Promote efficiency as a core evaluation for AI

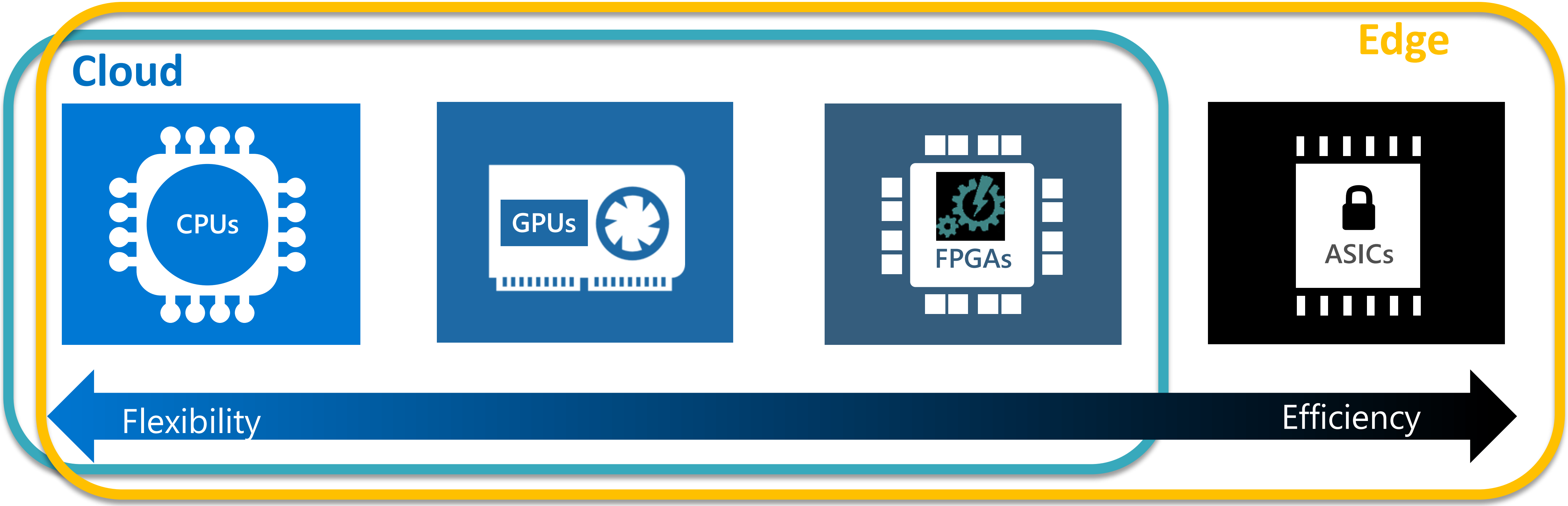

QUESTIONS
ANSWERS

Merci de votre participation !

Azure ML

Silicon Alternatives

FPGAs vs. CPU, GPU, and ASIC



TRAINING
CPUs and GPUs

INFERENCING
CPUs, GPUS, FPGAs

TRAINING

(HEAVY EDGE)
CPUs and GPUs

INFERENCING
CPUs, GPUS, FPGAs

Project Brainwave on Azure

Enables real-time AI calculations using FPGAs. Benefits include:

Performance

Excellent inference at low batch sizes

Ultra-low latency | 10x < CPU/GPU

Flexibility

Rapidly adapt to evolving ML

Inference-optimized numerical precision

Exploit sparsity, deep compression

Scale

World's largest cloud investment in FPGAs

Multiple Exa-Ops of aggregate AI capacity

Runs on Microsoft's scale infrastructure

Low cost

$0.21/million images on Azure FPGA

Project Brainwave on Azure

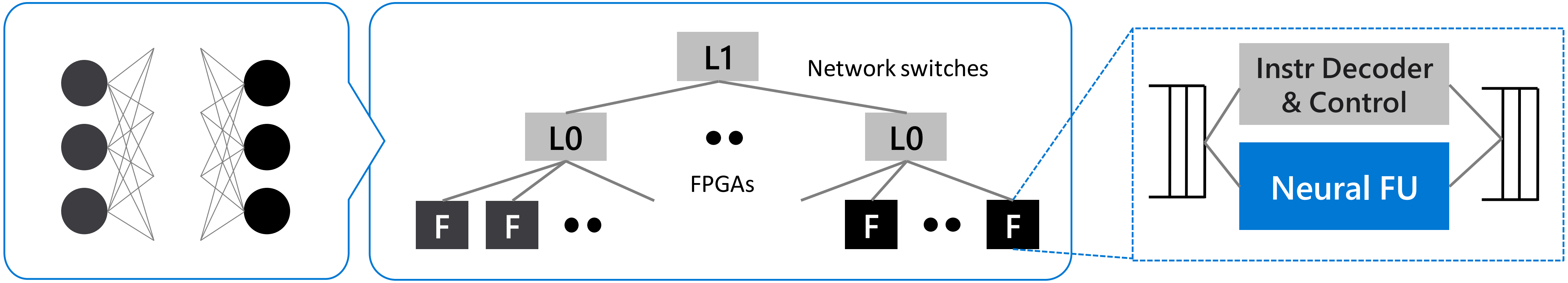
Capabilities Overview

A Scalable FPGA-Powered DNN Serving Platform

Fast: Ultra-low latency, high-throughput serving of DNN models at low batch sizes

Flexible: Future proof, adaptable to fast-moving AI space and evolving model types

Friendly: Turnkey deployment of TensorFlow/CNTK/Caffe/etc.



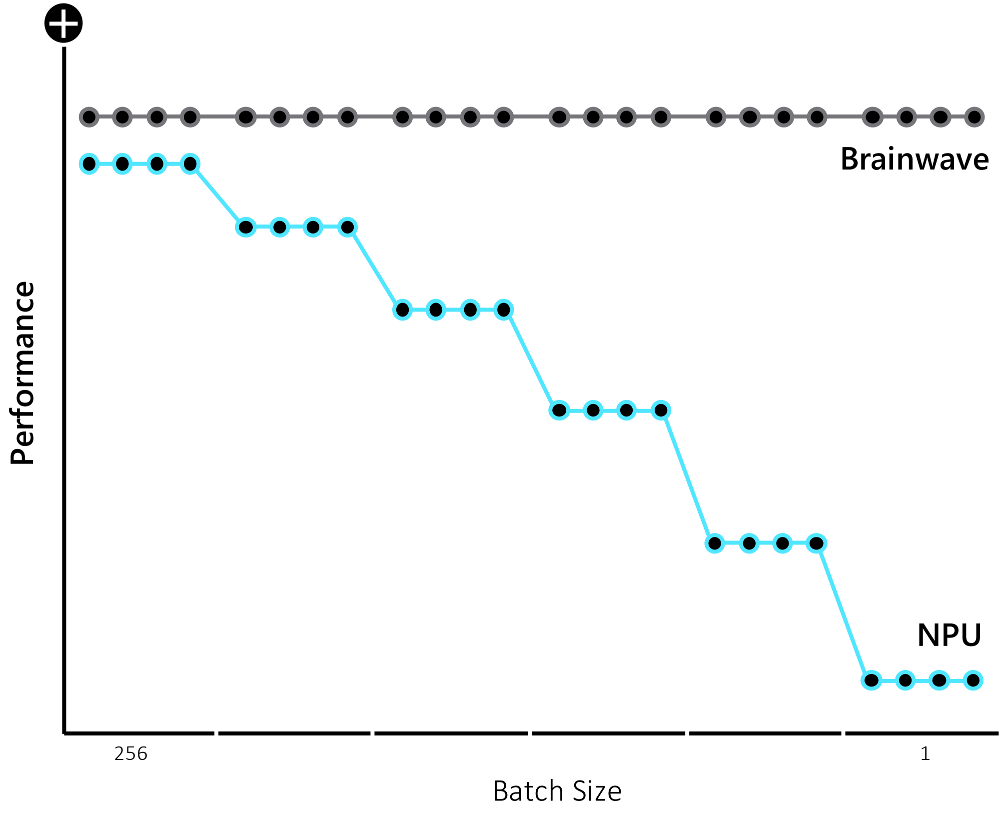
Pretrained DNN Model
in TensorFlow, CNTK, etc.

Scalable DNN Hardware
Microservice

BrainWave
Soft DPU

Project Brainwave

Advantages



Brainwave delivers the ideal combination:

High hardware utilization

Low latency

Low batch sizes

What is currently supported on Azure?

Today, Project Brainwave supports

Image classification and recognition scenarios

TensorFlow deployment

DNNs: ResNet 50, ResNet 152, VGG-16, SSD-VGG, and DenseNet-121

Intel FPGA hardware

Using this FPGA-enabled hardware architecture, trained neural networks run quickly and with lower latency.

Project Brainwave can parallelize pre-trained deep neural networks (DNN) across FPGAs to scale out your service.

The DNNs can be pre-trained, as a deep featurizer for transfer learning, or fine-tuned with updated weights

Here is the workflow for creating an image recognition service in Azure using supported DNNs as a featurizer for deployment on Azure FPGAs:

Use the Azure Machine Learning SDK for Python to create a service definition, which is a file describing a pipeline of graphs (input, featurizer, and classifier) based on TensorFlow. The deployment command will automatically compress the definition and graphs into a ZIP file and upload the ZIP to Azure Blob storage. The DNN is already deployed on Project Brainwave to run on the FPGA.

Register the model using the SDK with the ZIP file in Azure Blob storage.

Deploy the service with the registered model using SDK

How to deploy to FPGAs on Azure

Workflow for creating an image recognition service in Azure using supported DNNs as a featurizer for deployment on Azure FPGAs:

Use the Azure Machine Learning SDK for Python to create a service definition

- A file describing a pipeline of graphs (input, featurizer, and classifier) based on TensorFlow.
- Deployment command automatically compresses the definition and graphs into a ZIP file
- Deployment command automatically uploads the ZIP to Azure Blob storage.
- The DNN is already deployed on Project Brainwave to run on the FPGA.
- Model registration using the SDK with the ZIP file in Azure Blob storage.
- Deploy the service with the registered model using SDK.