Machine Learning Acceleration and Optimization: Use Cases

Francisco Almeida

falmeida@ull.edu.es

High Performace Computing Group

Universidad de La Laguna

Tenerife

Fondo Social Europeo Plus
Cofinanciado por la Unión Europea
Fondos Europeos
Gobierno de Canarias
Consejería de Economía, Conocimiento y Empleo
Agencia Canaria de Investigación Innovación y Sociedad de la Información
Gobierno de España
Ministerio de Ciencia, Innovación y Universidades

[TED2021-131019B-I00.](TED2021-131019B-I00.)

Outline

- Evolution of architectures the problem of Energy Consumption
- Evolution of AI Models the problem of Resource Consumption
- Quantization
- Pruning
- Conclusions

Preliminary Concepts

Universidad de La Laguna

Performance Metrics

- FLOP: Number of Floating Point Operations
- FLOPS: Floating Point Operations per Second
  - It not always represent well the capacity of a computer
  - Commonly accepted by the scientific community
- In the Supercomputing context the LINPACK test is used for calculating

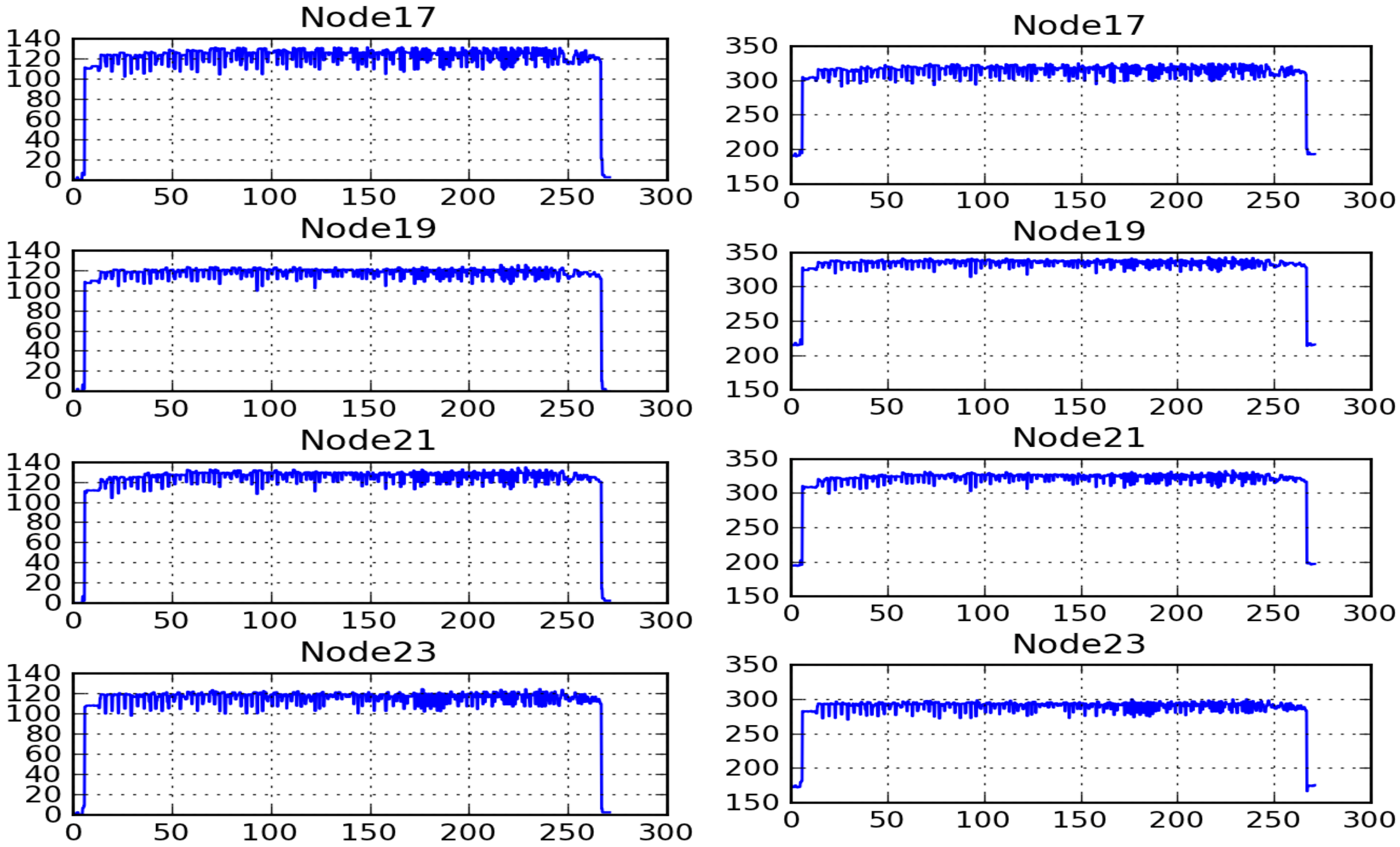
Nombre	Unidad	Flops
KiloFLOPS	Kflops	10^3
MegaFLOPS	Mflops	10^6
GigaFLOPS	Gflops	10^9
TeraFLOPS	Tflops	10^{12}
PetaFLOPS	Pflops	10^{15}
ExaFLOPS	Eflops	10^{18}
ZettaFLOPS	Zflops	10^{21}
YottaFLOPS	Yflops	10^{24}

Universidad de La Laguna

Performance Metrics

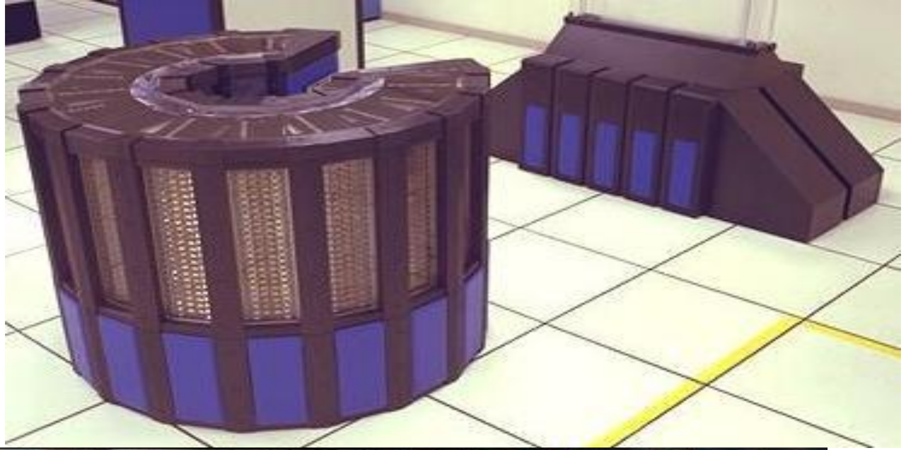
- Units for energy measurement
  - Watt (W) – Power (P)
- Energy (E)
  - E = P * T
  - Joule (J) – Watt second
  - Wh – Watt hour
  - kWh – KiloWatt hour
- Processor i9 → 1,3 Tflop/s (10^{12})

  → 125-250 W
- Self Estimation:
  - Sibiu city (68000 homes) would consume 250 GWh at the year?



Universidad de La Laguna

Supercomputer



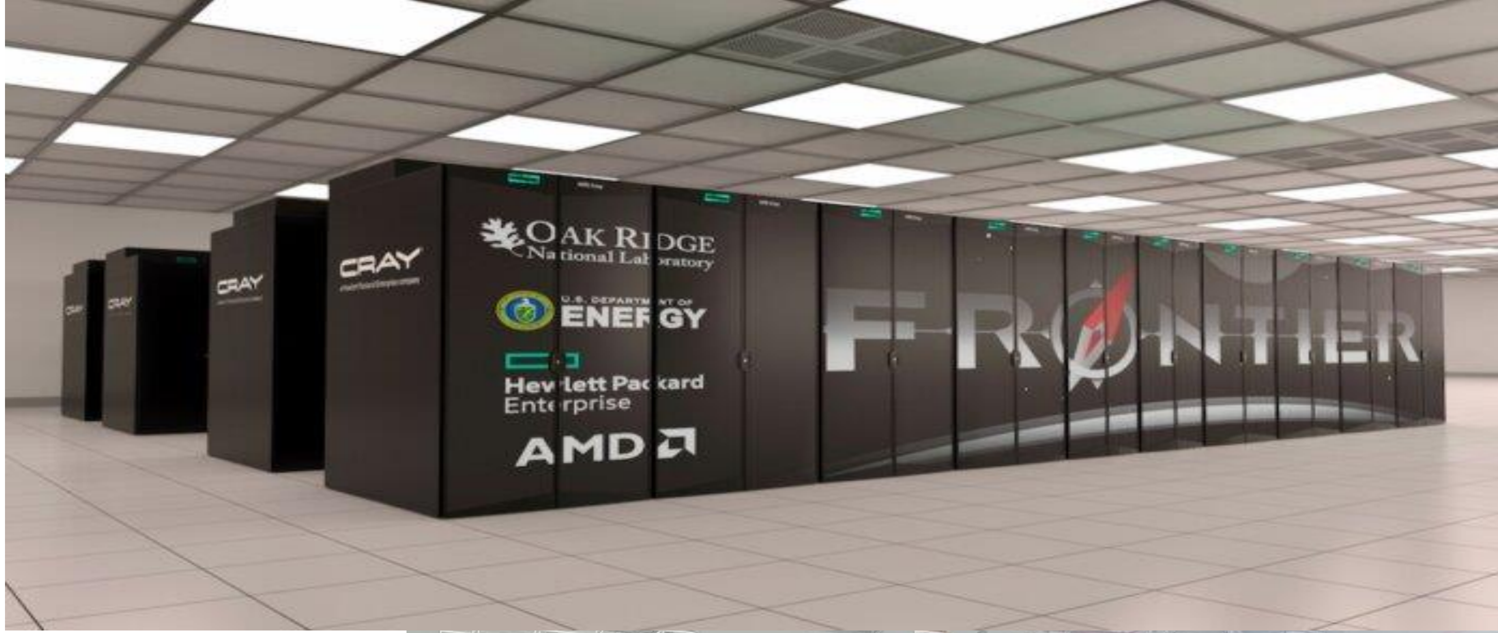
- Gigascale → 10^9 → 1985 → Cray 2
  - NASA
- Terascale → 10^{12} → 1997 →
  - Intel ASCI Red System
  - Sandia National Laboratory

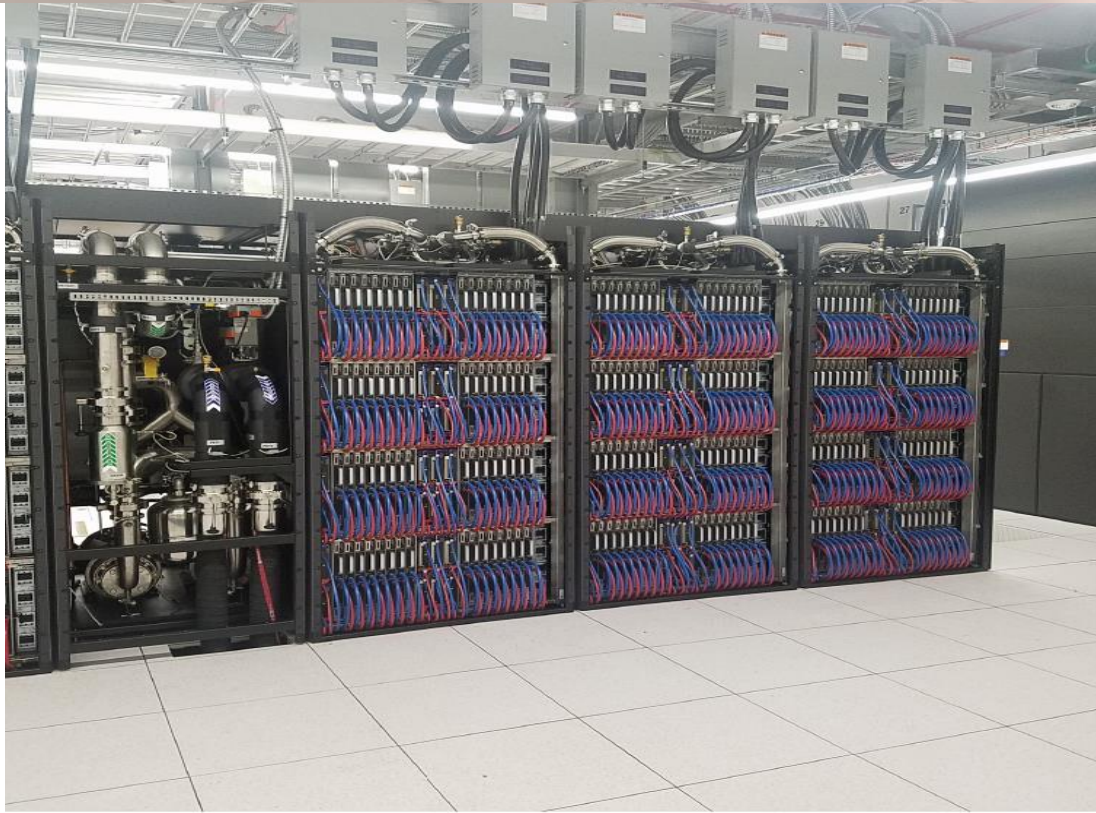
- Petascale → 10^{15} –> 2008
  - IBM RoadRunner
  - Los Alamos National Lab
- Exascale ?
  - First estimate 2015; 67MW-200MW
  - 2008 → Not before 2020; 20MW

Universidad de La Laguna

Frontier - 2022

- Oak Ridge National Laboratory – USA
- US$600 million
- Processor AMD EPYC “Trento” 64core integrated 4x MI250 “Instinct” GPUs
- 9,408 CPUs, 37,632 GPUs,
- 8.730.112 cores
- 1.12 Exaflops
- Peak performance 1.26 Exaflops
- Power: 21.100 kW
- 14 years later





Universidad
de La Laguna

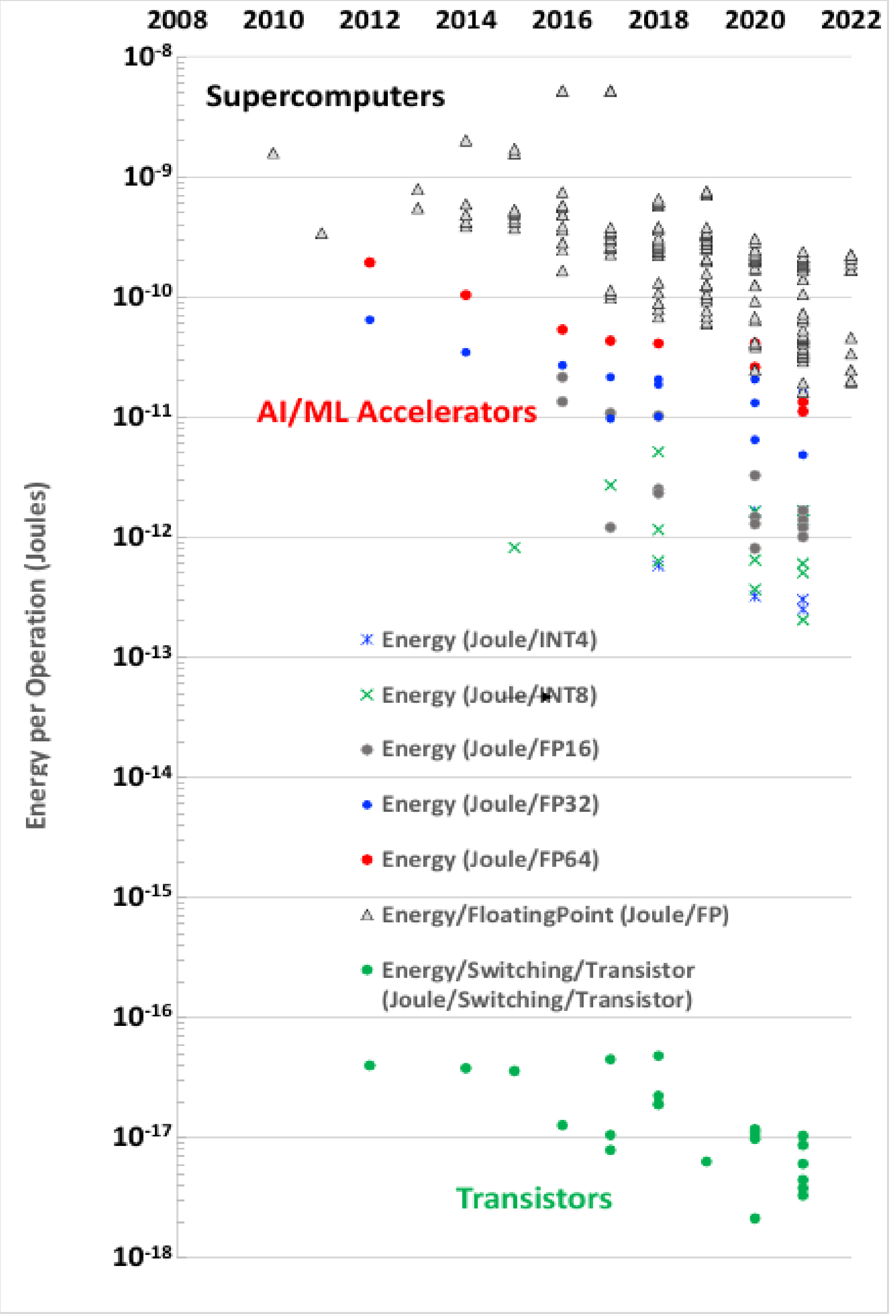
Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers and Compute-Intensive Applications

Shankar, Reuther
SLAC National Laboratory. Stanford University, CA, USA
MIT Lincoln Laboratory Supercomputing Center (LLSC), MA, USA

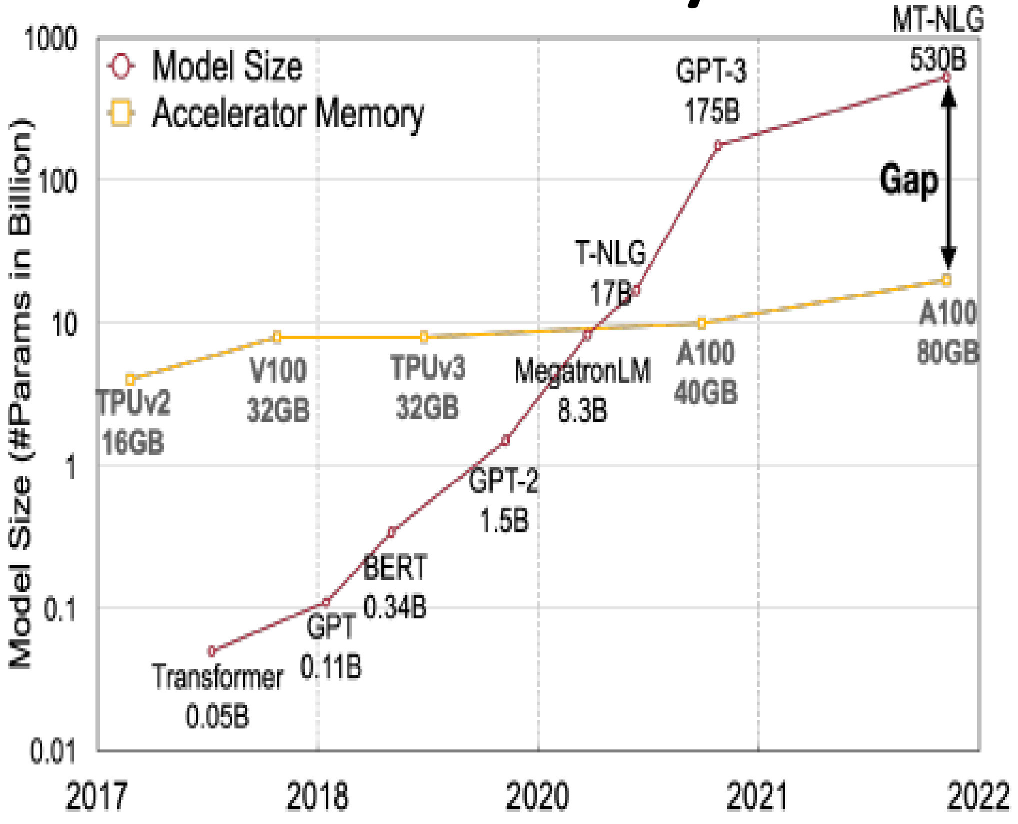
- Energy efficiency due to geometric scaling is slowing down
- Innovations in architectures can provide higher energy efficiency that that obtained by geometrical scaling
- Shift towards accelerating development of domain-specific specialized architectures

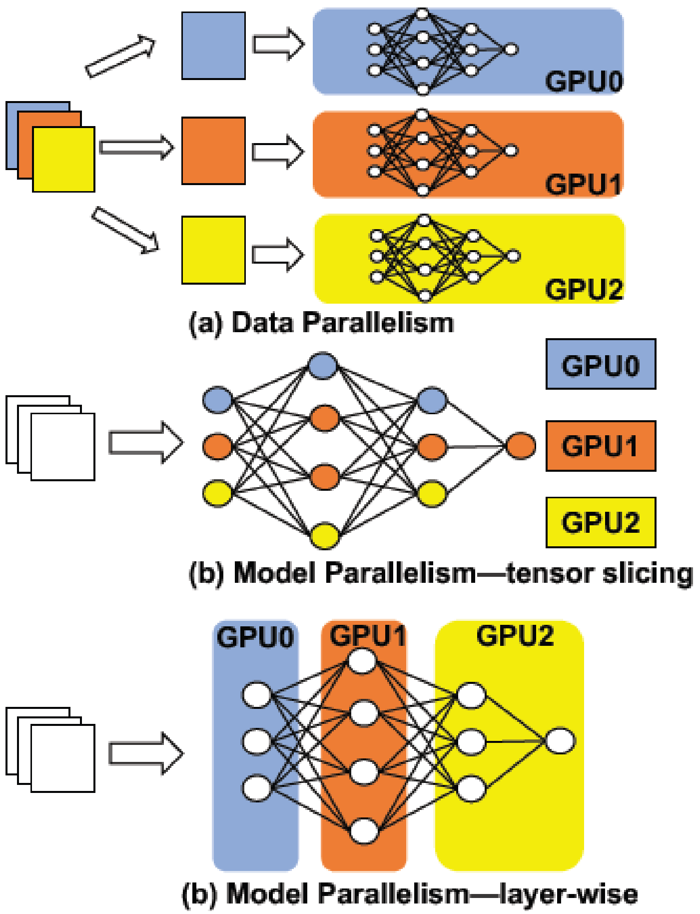
- Energy should be an additional design variable that bridges architecture and algorithms in addition to hardware and technology

Universidad de La Laguna



Model Size Increasing vs GPU Memory Increasing





Parallelizing DNN Training on GPUs: Challenges and Opportunities
Xu, Zhang, Tang
University of Pittsburgh. Pittsburgh, PA, USA

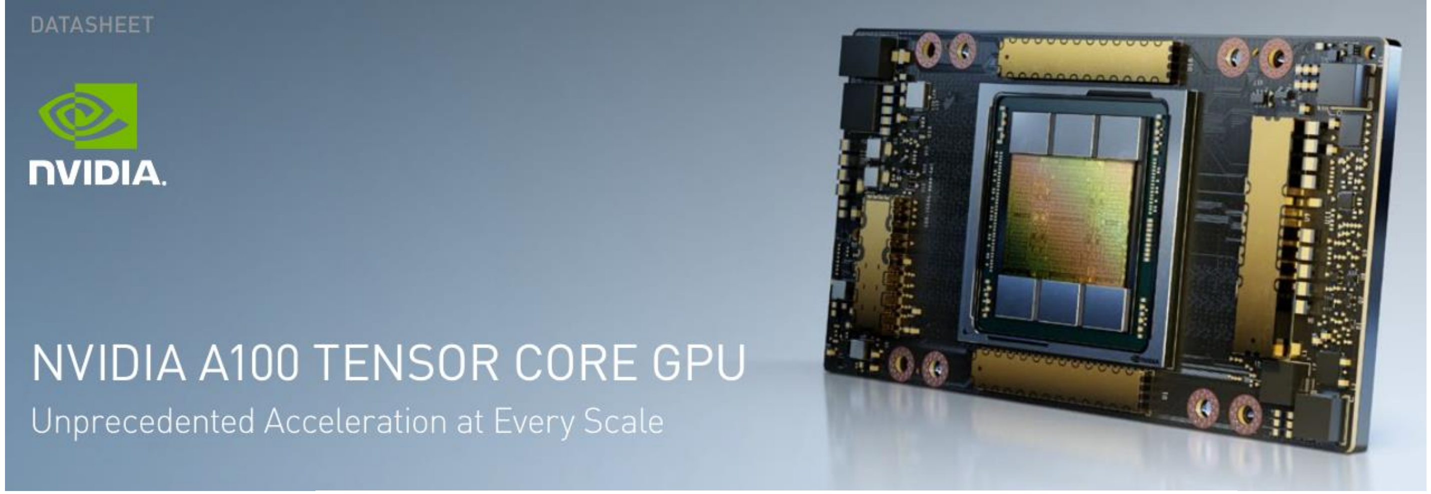
SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models
Xiao, Lin, Seznec, Wu, Demouth, Han
https://github.com/mit-han-lab/smoothquant

Universidad de La Laguna

GPT-4

- ~2e25 FLOP of training compute
  - 21.5 million Exaflop

- ~20,000 A100 for 90 to 100 days
  - 17 GWh-50 GWh



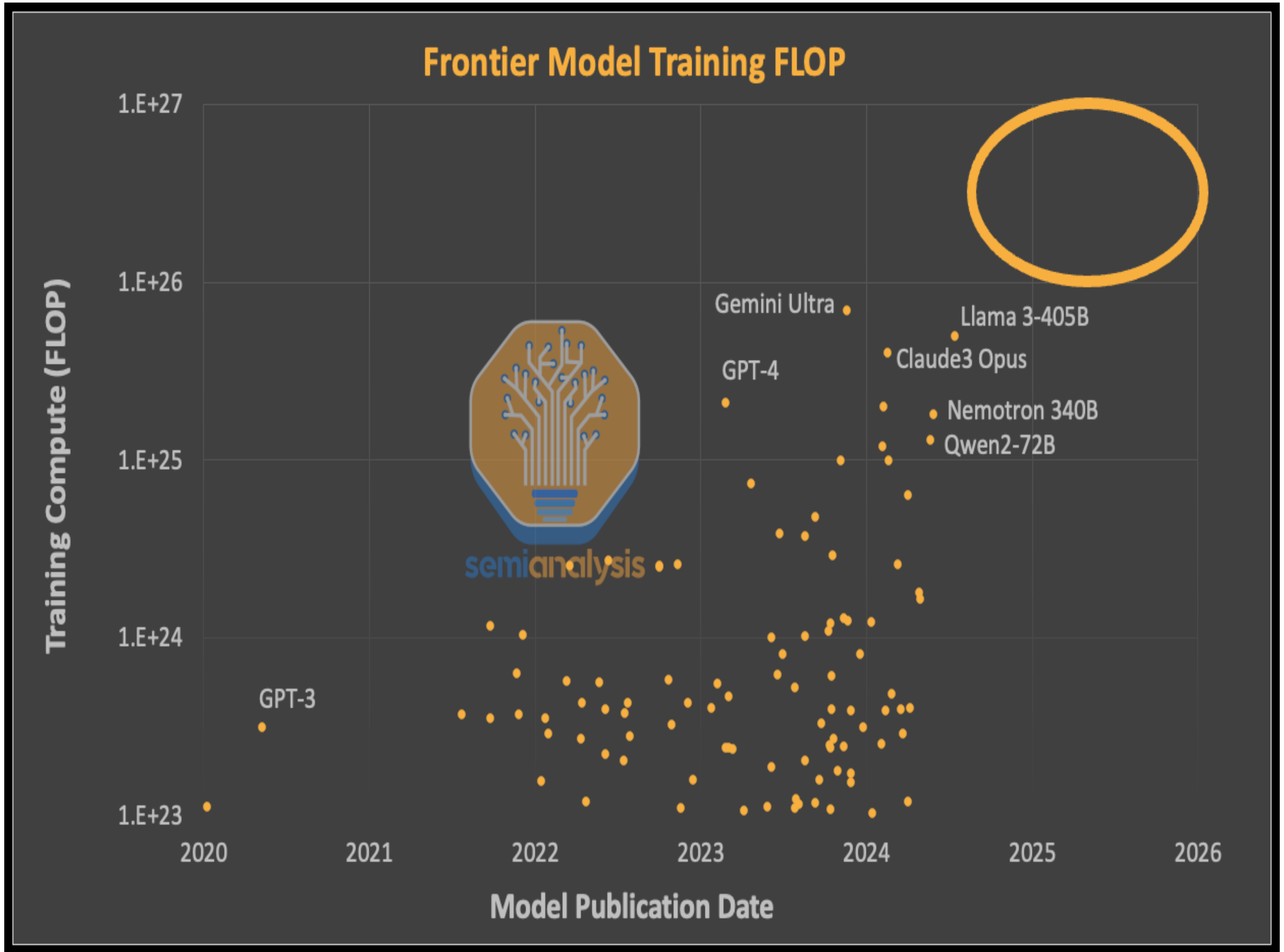
NVIDIA A100 TENSOR CORE GPU SPECIFICATIONS (SXM4 AND PCIE FORM FACTORS)

	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS | 1248 TOPS*	
GPU Memory	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,935GB/s	2,039GB/s
Max Thermal Design Power (TDP)	300W	400W***

Universidad de La Laguna

Billion Parameter Models

- Similar magnitudes of training computing for Gemini, Nemotron or Llama with less computational power



Universidad de La Laguna

On the race for Trillion Parameter Models: A 100K H100 Cluster

- A 100,000 H100 cluster would only take four days using FP8 to train GPT-4.
- On a 100k H100 cluster training run for 100 days, you can achieve an effective FP8 Model FLOP of ~6e26 (600 million ExaFLOP).
- Note that the poor reliability of hardware reduces MFU significantly.
- A 100,000 GPU cluster will require 150MW in datacenter capacity and guzzle down 1.59 Terawatt hours in a single year, costing ~200 million euros.

Semianalysis

Technical Specifications

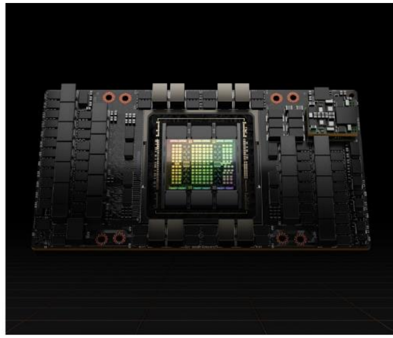
	H100 SXM	H100 NVL
FP64	34 teraFLOPS	30 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	60 teraFLOPS
FP32	67 teraFLOPS	60 teraFLOPS
TF32 Tensor Core*	989 teraFLOPS	835 teraFLOPS
BFLOAT16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP8 Tensor Core*	3,958 teraFLOPS	3,341 teraFLOPS
INT8 Tensor Core*	3,958 TOPS	3,341 TOPS
GPU Memory	80GB	94GB
GPU Memory Bandwidth	3.35TB/s	3.9TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Max Thermal Design Power (TDP)	Up to 700W (configurable)	350-400W (configurable)

Datasheet

NVIDIA

NVIDIA H100 Tensor Core GPU

Extraordinary performance, scalability, and security for every data center.



Universidad de La Laguna

Post

Elon Musk
@elonmusk

This weekend, the @xAI team brought our Colossus 100k H100 training cluster online. From start to finish, it was done in 122 days.

Colossus is the most powerful AI training system in the world. Moreover, it will double in size to 200k (50k H200s) in a few months.

Excellent work by the team, Nvidia and our many partners/suppliers.

Traducir post

5:53 p. m. · 2 sept. 2024 · **15,2 M** Reproducciones

7.794 Reposts **1.274** Citas **75,3 mil** Me gusta **3.851** Elementos guardados

3 mil

Elon Musk unveils Colossus: World's most advanced AI Supercomputer

Fri 06 Sep 2024 Science-Tech

The Brew News Team | < 1 min read

Universidad de La Laguna

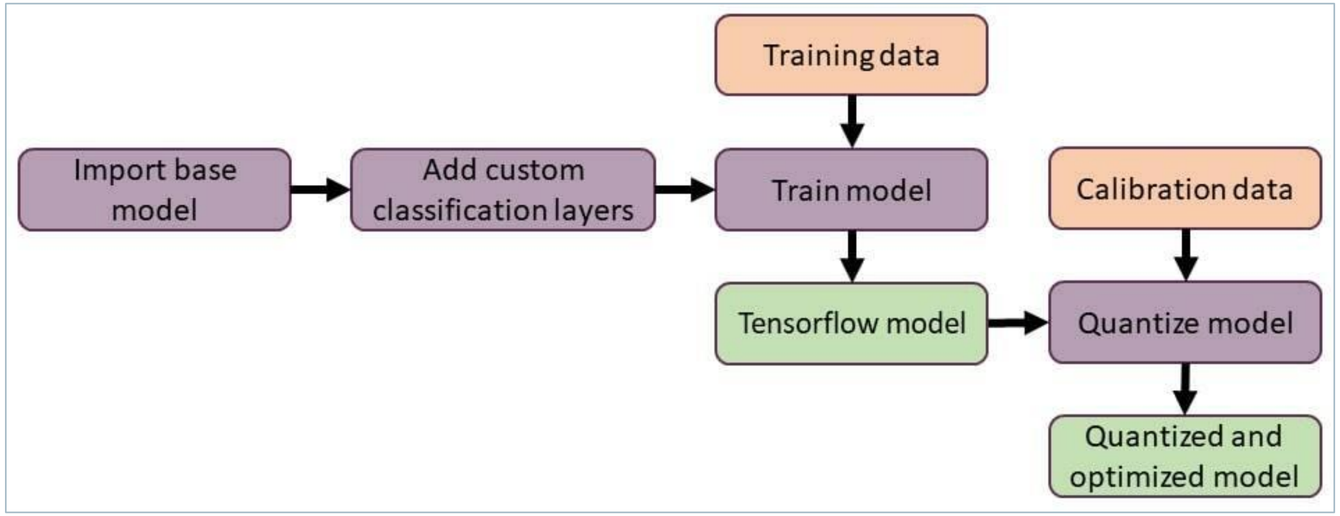
Is it sustainable?

Optimizations?

Universidad de La Laguna

Quantization

- The process of constraining an input from a continuous or otherwise large set of values (such as the real numbers) to a discrete set (such as the integers).
- A common way to achieve this is by rounding or truncating.
- Quantization can reduce memory and accelerate inference.
- Weights are easy to quantize while activations are not.

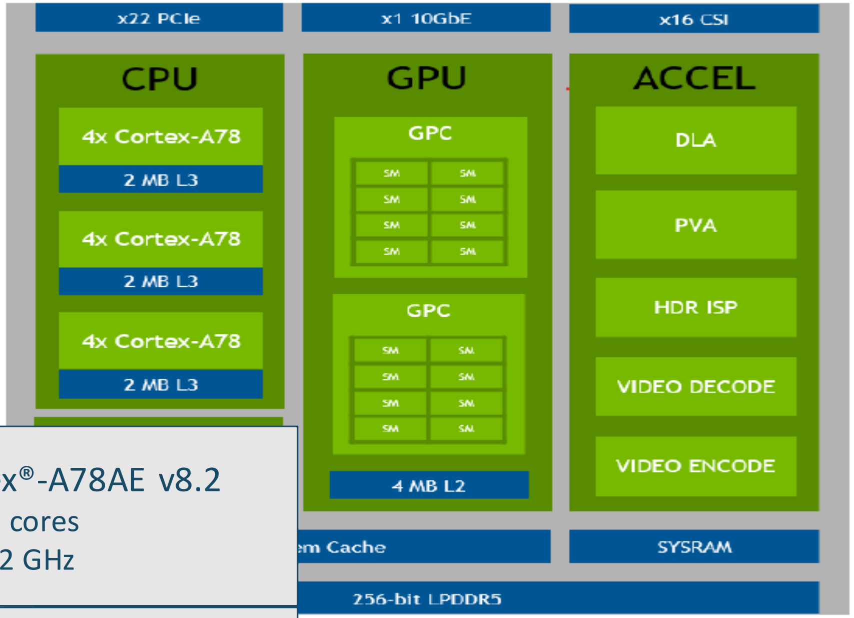


Post Training Quantization – PTQ
Aware Training Quantization - ATQ

FP32
FP16
INT8

Running in Resource Limited Architectures

Processor unit	DC-CPU	DC-GPU	Jetson-CPU	Jetson-GPU	Jetson-DLA
Name	i7-1260P	RTX3080	CortexA78AE	GA10B	NVDLA v2.0
Manufacturer	Intel	NVIDIA	ARM	NVIDIA	NVIDIA
Cores	12	8704 CUDA 272 Tensor	12	2048 CUDA 64 Tensor	-
Frequency	4,7 GHz	1,71 GHz	2,2 GHz	1,3 GHz	1,6 GHz
Memory/Cache	18 MB	10 GB	3 MB L2 6 MB L3	Integrated	-
Energy	20∼64 W	320 W	-	-	-



CPU	Arm® Cortex®-A78AE v8.2 12 cores 2,2 GHz
GPU	GA10B 2048 CUDA cores 64 Tensor cores 1,3 GHz
Power	15-50 W 50W: Maximum performance
DLA	Deep Learning Accelerator 2x NVDLA v2 1,6 GHz

- Desktop Computers
- Edge Nodes
- IoT Devices

Universidad de La Laguna

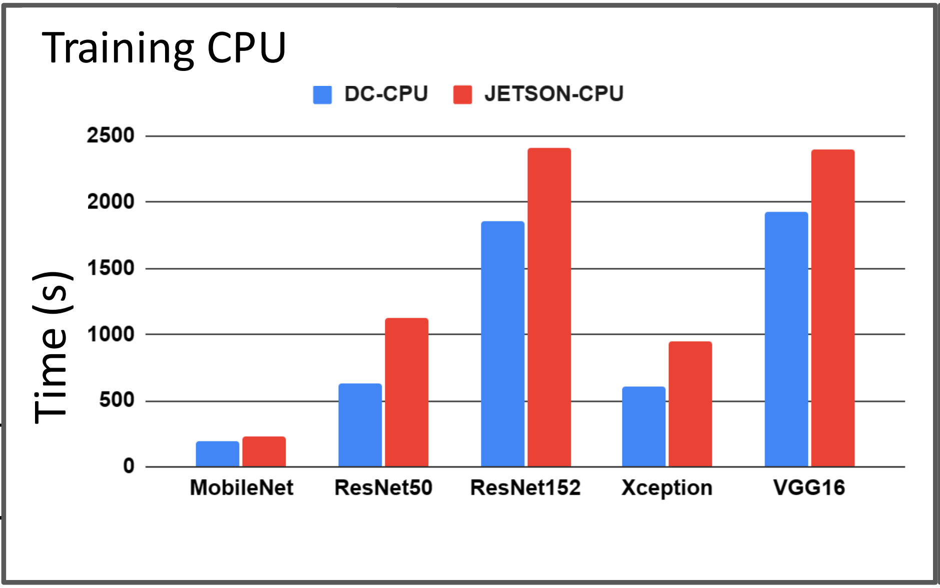
Notation

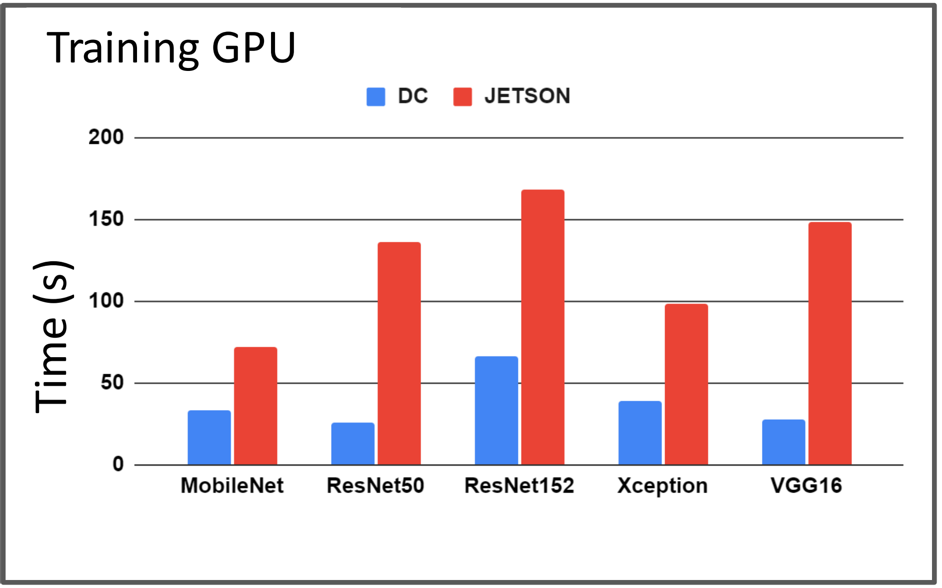
- Accuracy: How often a model predicts the outcome
- Precision: Numerical Precision
- Performance: Efficiency in terms of hardware/software
  - Processing Units
  - Running Time
  - Energy consumption

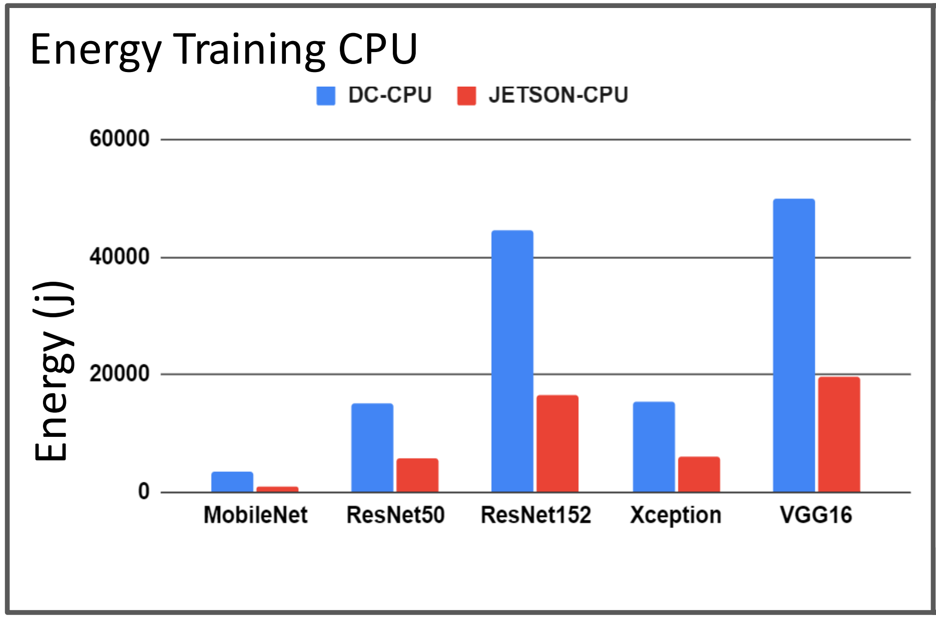
Universidad de La Laguna

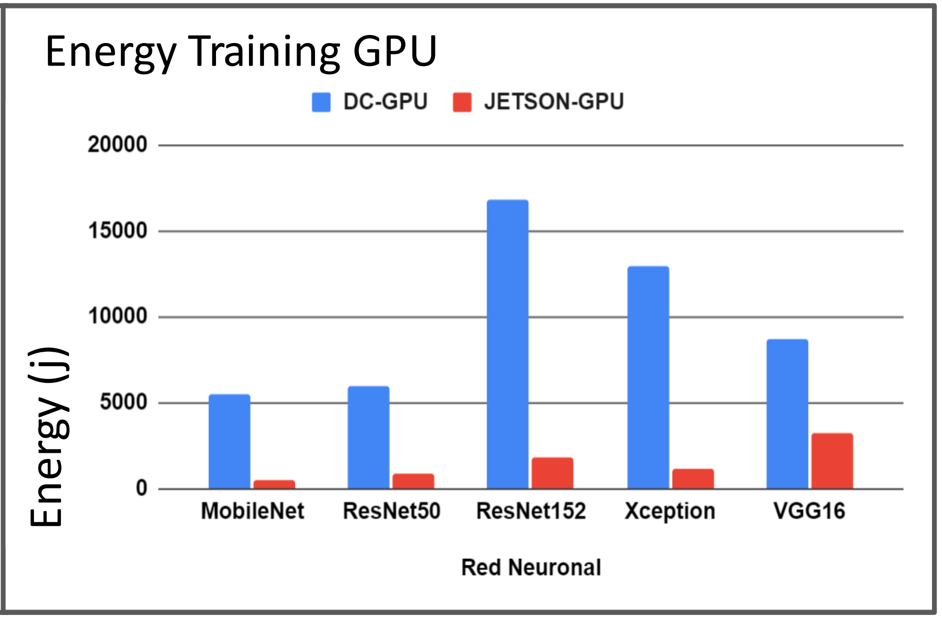
Quantization

Network	Depth	Parameters
MobileNet	105	3.538.984
ResNet50	107	25.636.712
ResNet152	311	60.419.944
Xception	81	22.910.480
Vgg16	16	138.357.544



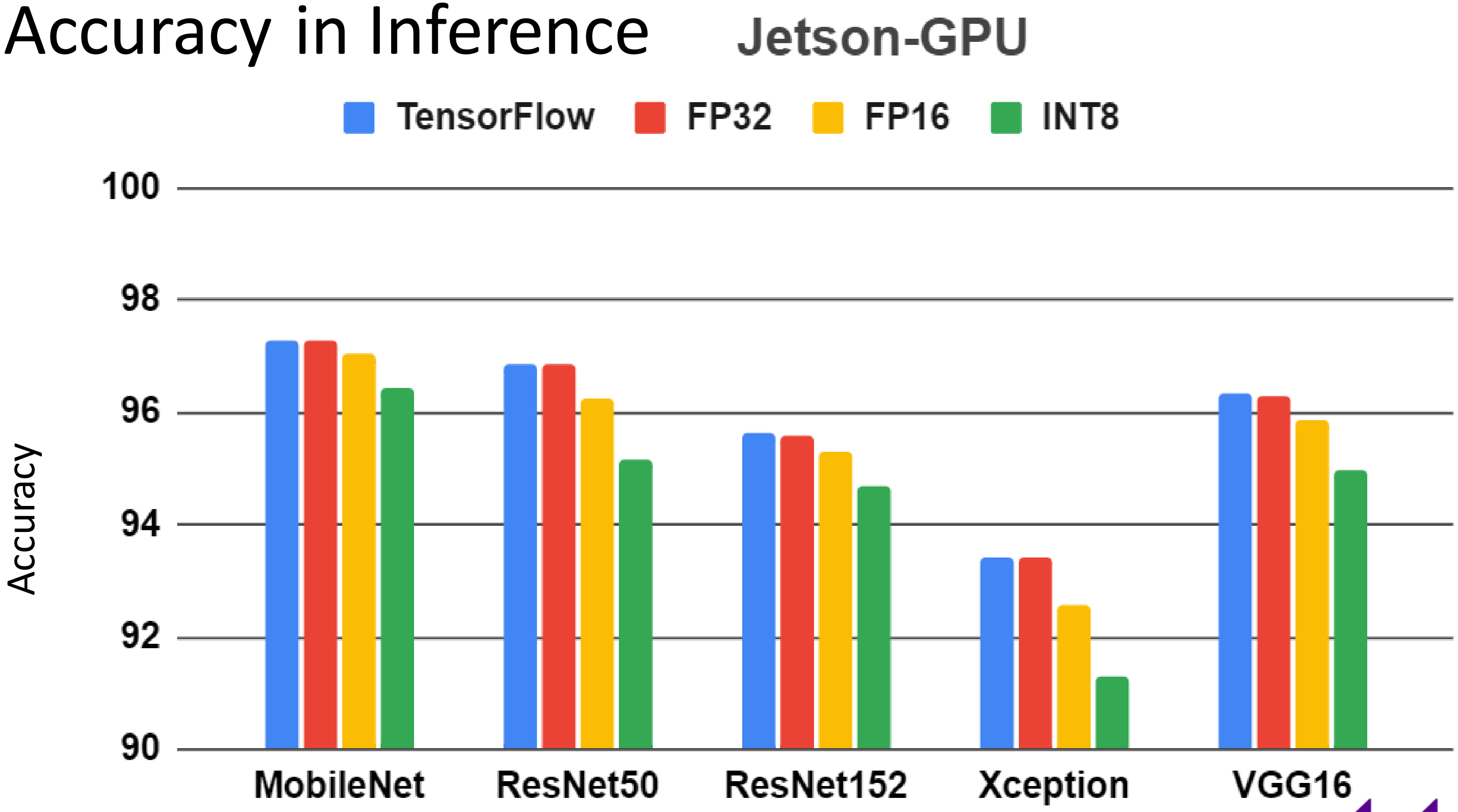






Universidad de La Laguna

Accuracy in Inference



Universidad de La Laguna

Quantization - “Extreme”

- *SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models*

Xiao, Lin; Massachusetts Institute of Technology

Seznec, Wu, Demouth, Han; NVIDIA

2024

- *BitNet: Scaling 1-bit Transformers for Large Language Models*

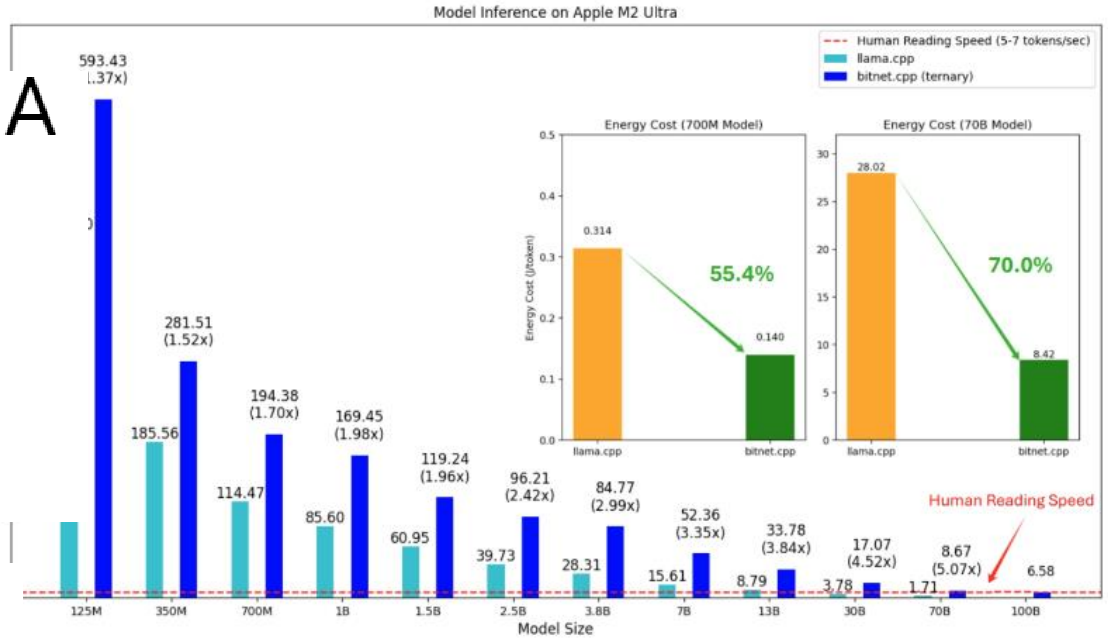
Wang, Ma, Dong, Huang, et all.

Microsoft Research, University of Chinese Academy of Sciences,Tsinghua University.
2023

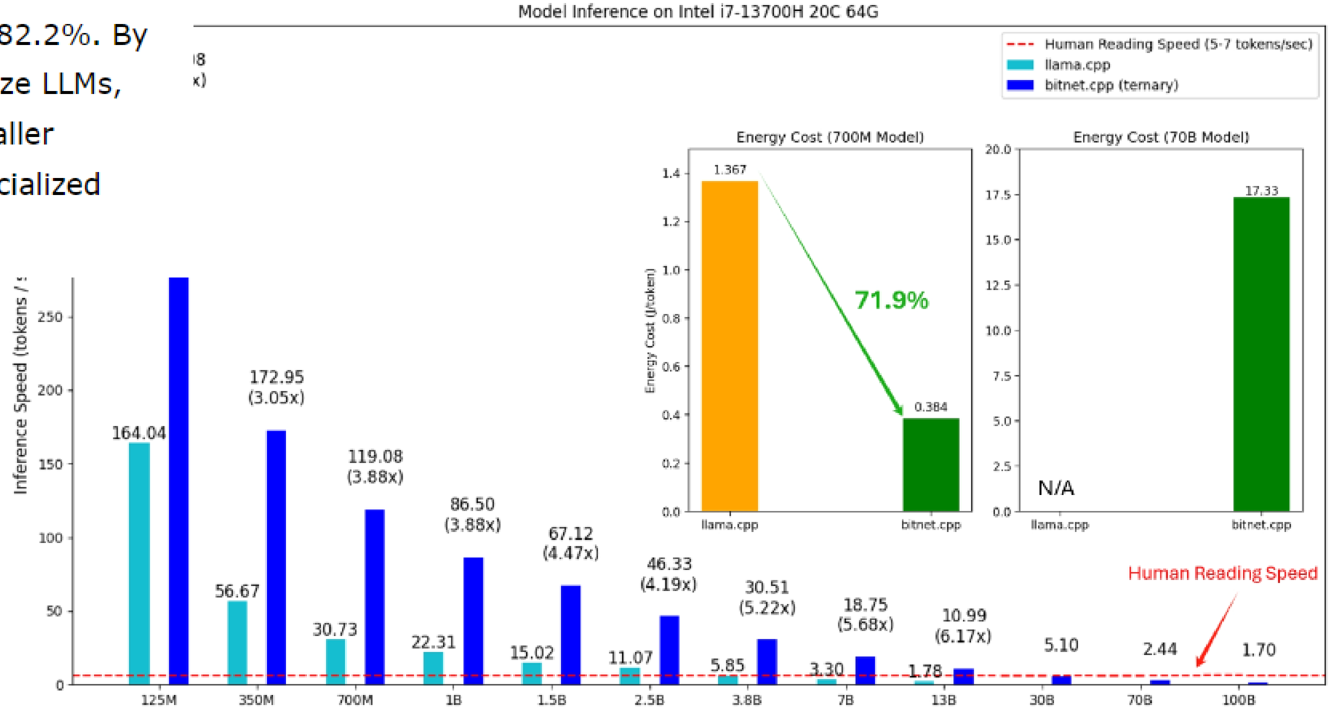
Universidad de La Laguna

Microsoft Open-Sources bitnet.cpp: A Super-Efficient 1-bit LLM Inference Framework that Runs Directly on CPUs

By **Asif Razzaq** - October 18, 2024



Microsoft recently open-sourced bitnet.cpp, a super-efficient 1-bit LLM inference framework that runs directly on CPUs, meaning that even large 100-billion parameter models can be executed on local devices without the need for a GPU. With bitnet.cpp, users can achieve impressive speedups of up to 6.17x while also reducing energy consumption by 82.2%. By lowering the hardware requirements, this framework could potentially democratize LLMs, making them more accessible for local use cases and enabling individuals or smaller businesses to harness AI technology without the hefty costs associated with specialized hardware.

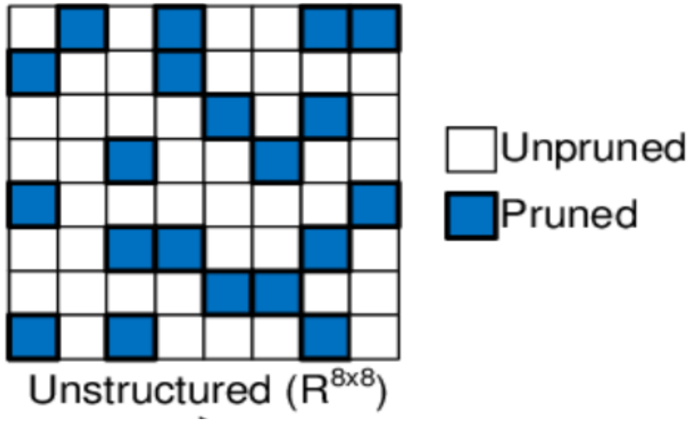


Pruning

- Removes individual connections (weights) of the network
- Technique used to reduce memory usage, which can also reduce the computational load when combined with compressed storage formats and efficient sparse kernels
- Many Criteria:

$$w_{i,j} = \begin{cases} 0 & \text{si } |w_{i,j}| < T \\ w_{i,j} & \text{si } |w_{i,j}| \geq T \end{cases}$$

- Nonstructured: Independent of their location
- Structured: Removes complete components (layers, heads)
- Semi-structured: Prune groups of weights



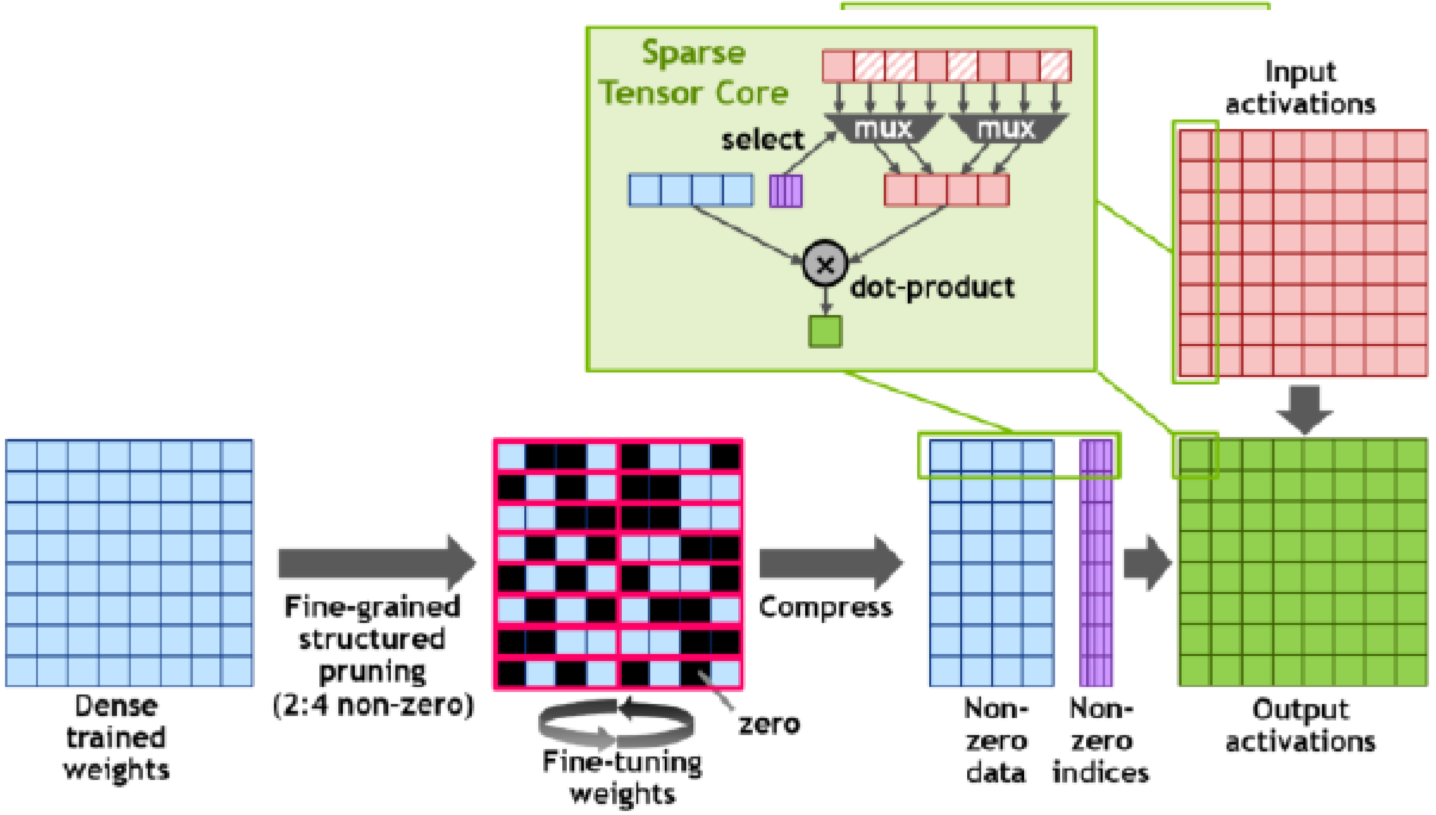
Universidad de La Laguna

Pruning

- Last generations of NVIDIA GPUs have extended their TCUs to also handle row-wise 2:4 sparsity. These updated TCUs include hardware support for sparse computation, and are referred to as Sparse Tensor Cores (SPTCs).

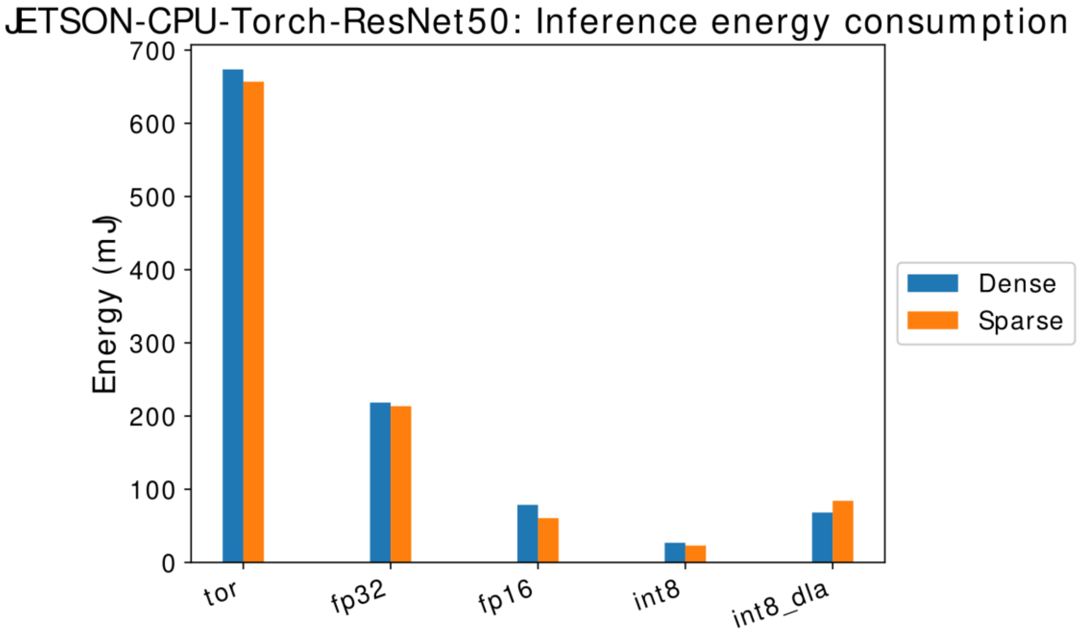
- To exploit SPTCs, the first argument in tensor operations must be stored in NVIDIA's N:M sparse format, N represents the maximum number of non-zero elements in a block of M values.

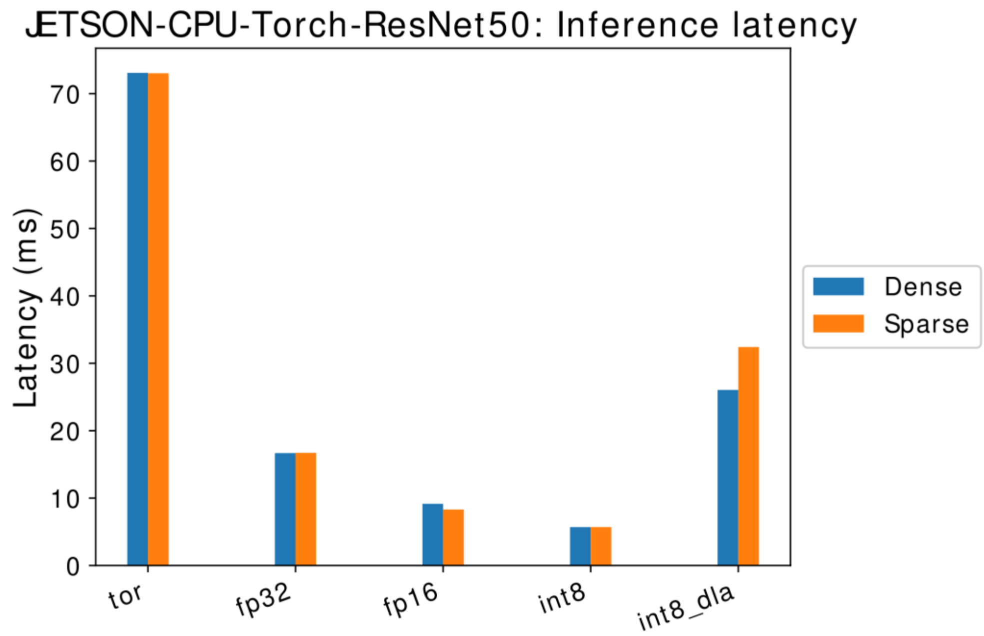
Ampere GPU 3rd Generation Tensor Core Sparsity



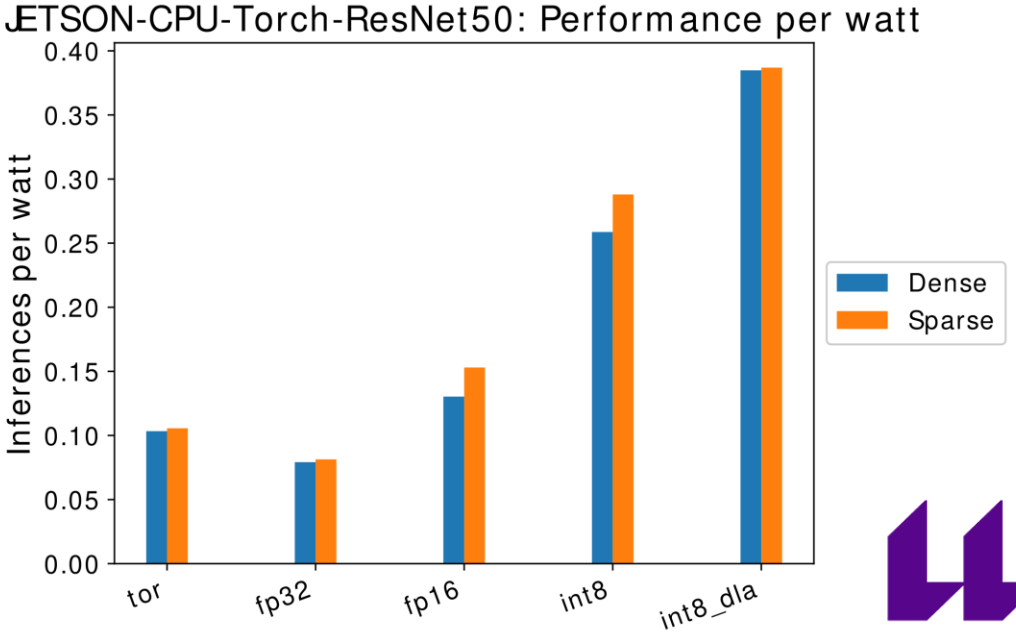
The 2:4 format and its mapping to SPTCs

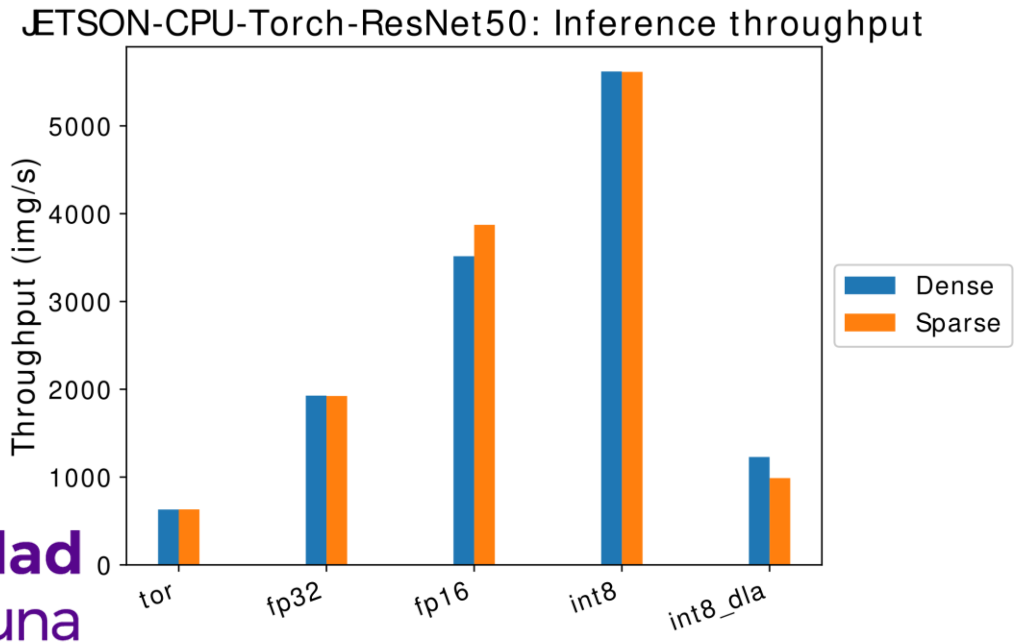
Universidad de La Laguna



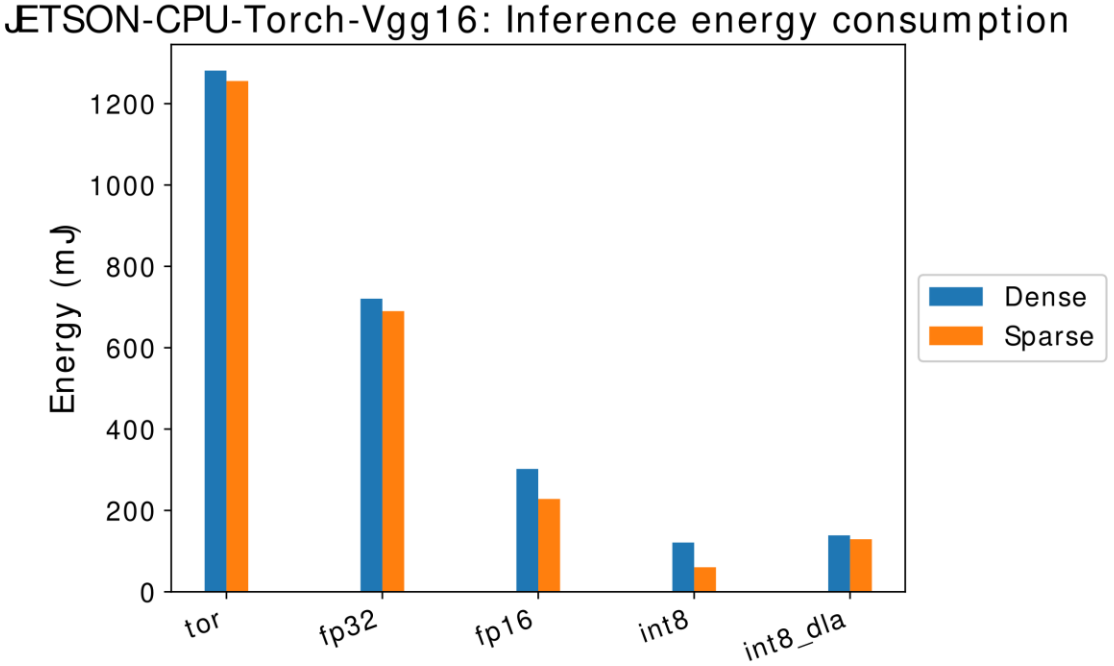


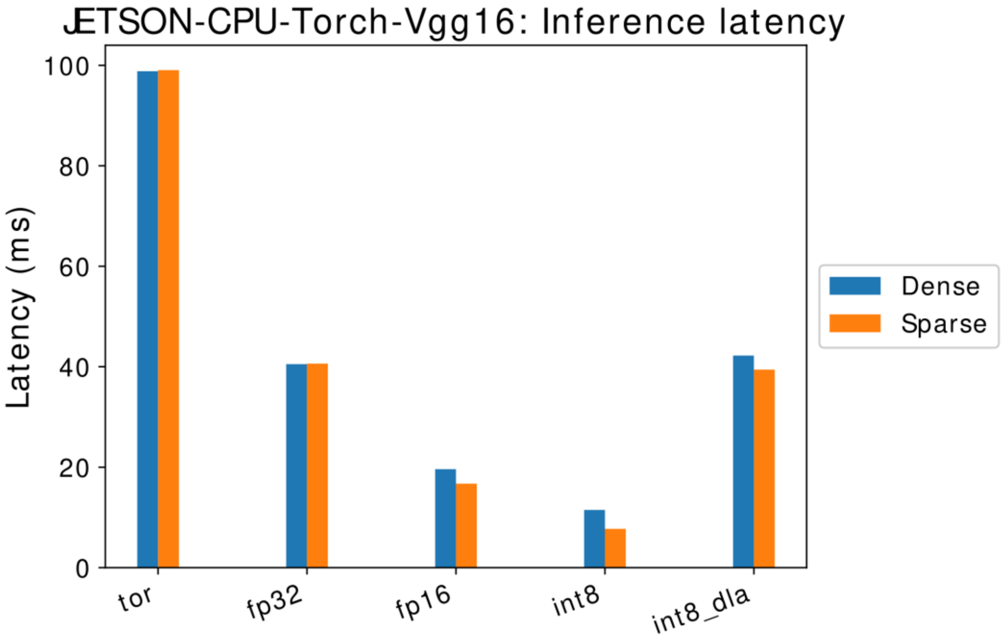
Prune



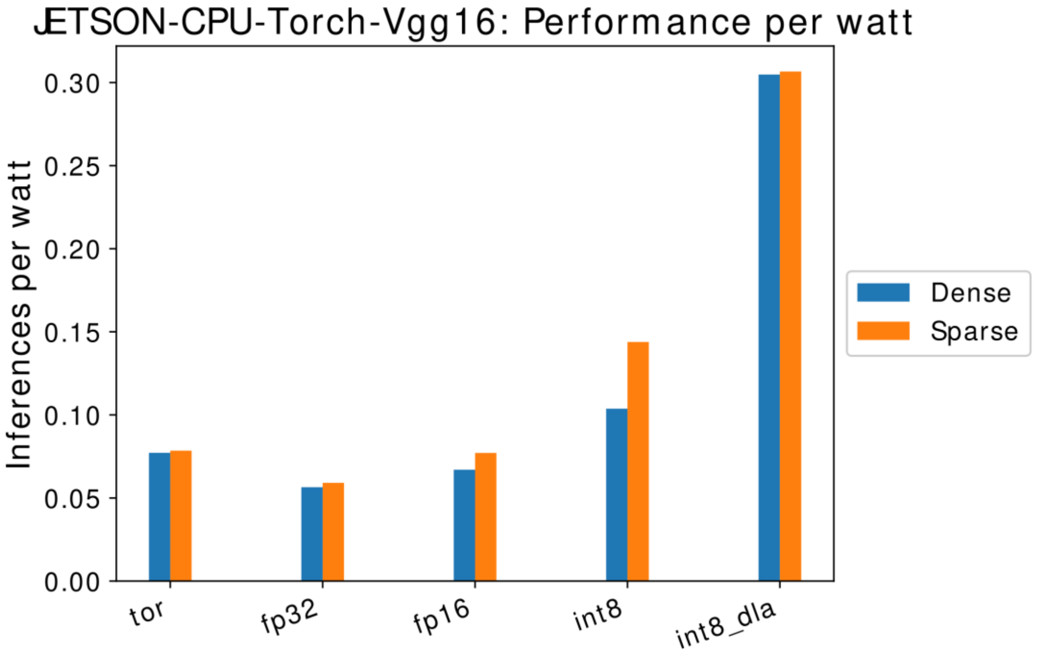


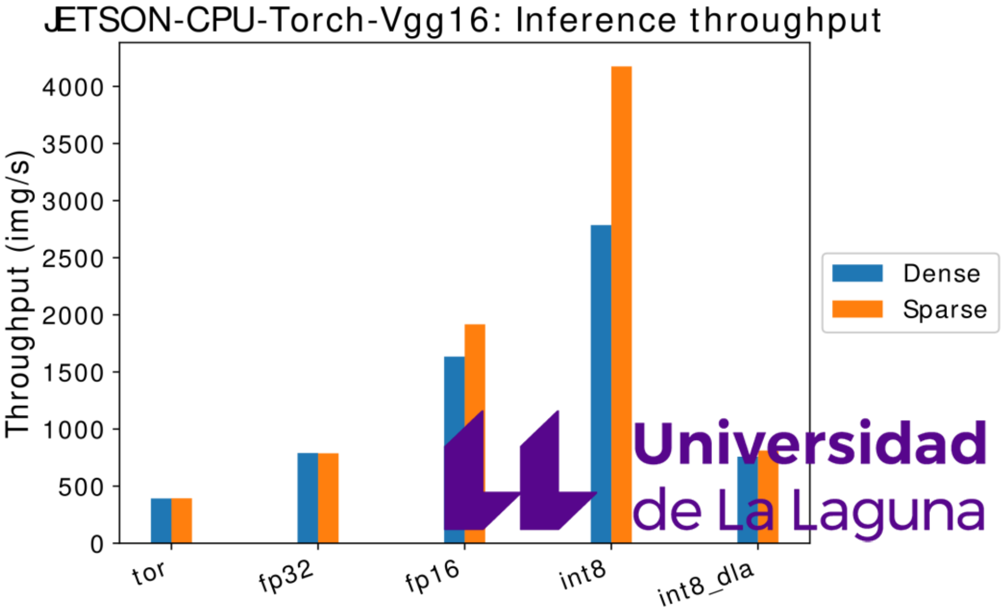
Universidad
de La Laguna





Prune





Universidad de La Laguna

Optimized Models

- GPT-4 Turbo, GPT-4o, GPT-4o mini
- Lightweight Llama 3.2 – Pruning and Distillation

Universidad de La Laguna

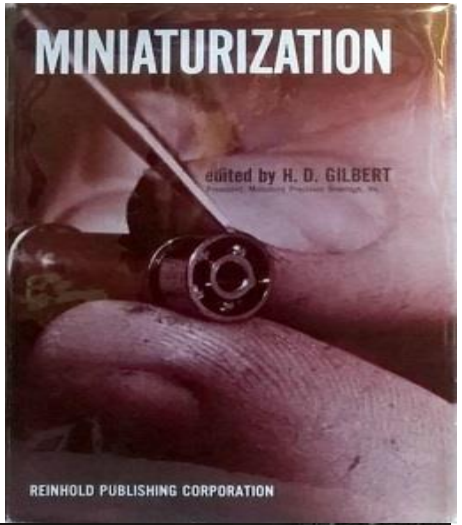
Conclusions

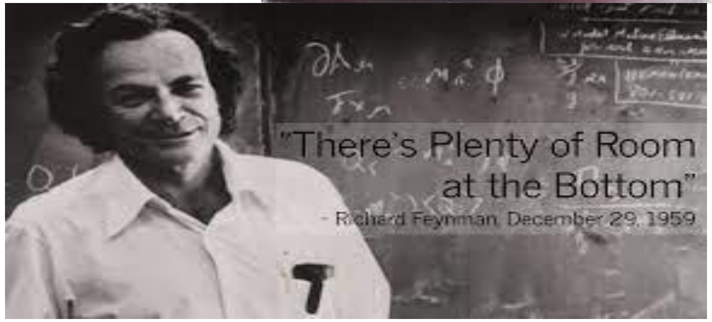
Universidad de La Laguna

The Energy is a big Issue

Universidad de La Laguna

- There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics.
  - Lecture at American Physical Society
  - 1959
  - Richard Feynman
  - Nobel Price Phisics 1965





- There's plenty of room at the Top: What will drive computer performance after Moore's law? - 2020
  - Leiserson et all.

Stay hungry stay foolish

We should never stop learning and should always try new things

Steve Jobs

Machine Learning Acceleration and Optimization: Use Cases

Francisco Almeida

falmeida@ull.edu.es

High Performace Computing Group

Universidad de La Laguna

Tenerife

Universidad de La Laguna

Fondo Social Europeo Plus
Cofinanciado por la Unión Europea
Fondos Europeos
Gobierno de Canarias
Consejería de Economía, Conocimiento y Empleo
Agencia Canaria de Investigación Innovación y Sociedad de la Información
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
TED2021-131019B-I00.