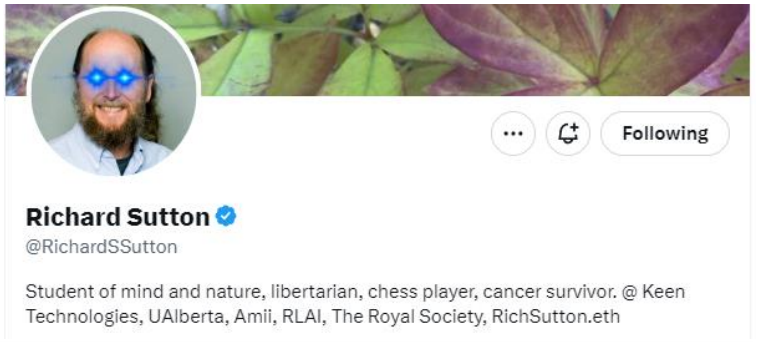
# NVIDIA GPU Architectures

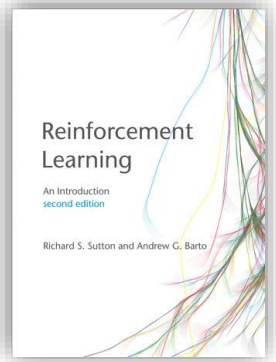
2023-11-03

손형욱

hyounguk.shon@kaist.ac.kr

# The Bitter Lesson




## The Bitter Lesson

### Rich Sutton

### March 13, 2019

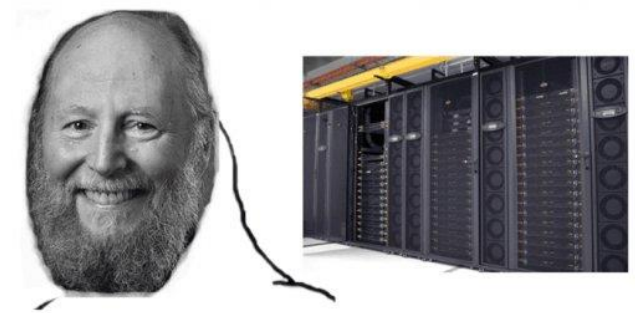
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

## Modeling vs. Computing power

Modeling = designing better architecture and learning algorithm

Sutton, Richard. "The bitter lesson." *Incomplete Ideas (blog)* 13.1 (2019). http://www.incompleteideas.net/IncIdeas/BitterLesson.html

# The Bitter Lesson

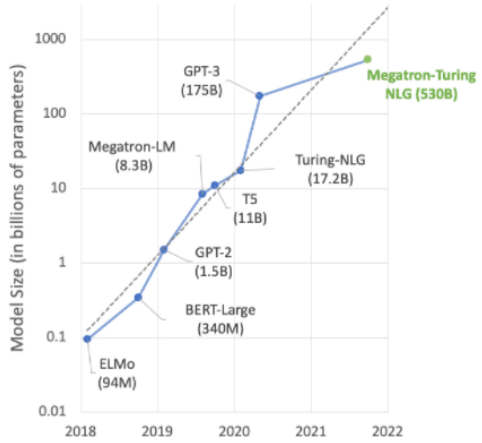

**70년간의 AI연구로부터 깨달을 수 있는 가장 큰 교훈은 바로 계산을 이용하는 일반적인 방법론이야말로 궁극적으로, 큰 격차를 두고도 가장 효율적이라는 것이다.** 이에 대한 궁극적인 이유는 바로 (무어의 법칙) 유닛 당 계산 비용이 꾸준히 감소하고 있기 때문이다. ...
첫 번째 교훈은 일반적인 방법론, 즉 사용 가능한 계산이 매우 커짐에 따라 계속해서 확장되는 방법의 위대한 힘이다. 이렇게 마구 확장하는 것처럼 보이는 두가지 방법은 바로 탐색과 학습이다. 두 번째 교훈은 정신의 실제 내용물은 엄청나게 복잡하다는 것이다. 우리는 정신의 내용에 대해서 단순한 방식으로 생각하는 방법을 찾는 것을 멈춰야 한다. 예를 들어 공간, 물체, 여러 행위자(agent), 혹은 대칭성에 대해서 생각하는 단순한 방법들 말이다.

# The Bitter Lesson

General and scalable algorithm beats other methods

CNN
RNN

Transformer




Google DeepMind 2023-10-26

## ConvNets Match Vision Transformers at Scale

Samuel L Smith[1], Andrew Brock[1], Leonard Berrada[1] and Soham De[1]
[1]Google DeepMind

**Many researchers believe that ConvNets perform well on small or moderately sized datasets, but are not competitive with Vision Transformers when given access to datasets on the web-scale. We challenge this belief by evaluating a performant ConvNet architecture pre-trained on JFT-4B, a large labelled dataset of images often used for training foundation models. We consider pre-training compute budgets between 0.4k and 110k TPU-v4 core compute hours, and train a series of networks of increasing depth and width from the NFNet model family. We observe a log-log scaling law between held out loss and compute budget. After fine-tuning on ImageNet, NFNets match the reported performance of Vision Transformers with comparable compute budgets. Our strongest fine-tuned model achieves a Top-1 accuracy of 90.4%.**

**Yann LeCun** ∞
@ylecun

Compute is all you need.
For a given amount of compute, ViT and ConvNets perform the same.

Quote from this DeepMind article: "Although the success of ViTs in computer vision is extremely impressive, in our view there is no strong evidence to suggest that pre-trained ViTs outperform pre-trained ConvNets when evaluated fairly."

arxiv.org/abs/2310.16764

7:56 AM · Oct 27, 2023 · **463.2K** Views

https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/

# The Bitter Lesson

## The “it” in AI models is the dataset.

Posted on June 10, 2023 by jbetker

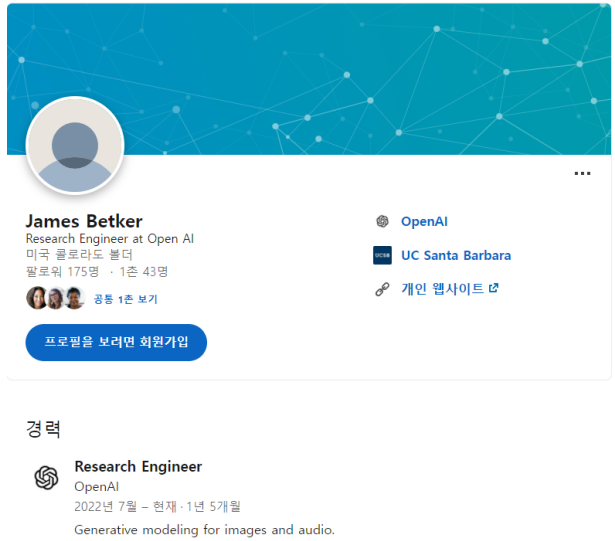
I’ve been at OpenAI for almost a year now. In that time, I’ve trained a **lot** of generative models. More than anyone really has any right to train. As I’ve spent these hours observing the effects of tweaking various model configurations and hyperparameters, one thing that has struck me is the similarities in between all the training runs.

It’s becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distributions that don’t matter, like what photos humans are likely to take or words humans commonly write down.

What this manifests as is – trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point. Sufficiently large diffusion conv-unets produce the same images as ViT generators. AR sampling produces the same images as diffusion.

This is a surprising observation! It implies that model behavior is not determined by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivery compute to approximating that dataset.

Then, when you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it’s not the model weights that you are referring to. It’s the dataset.

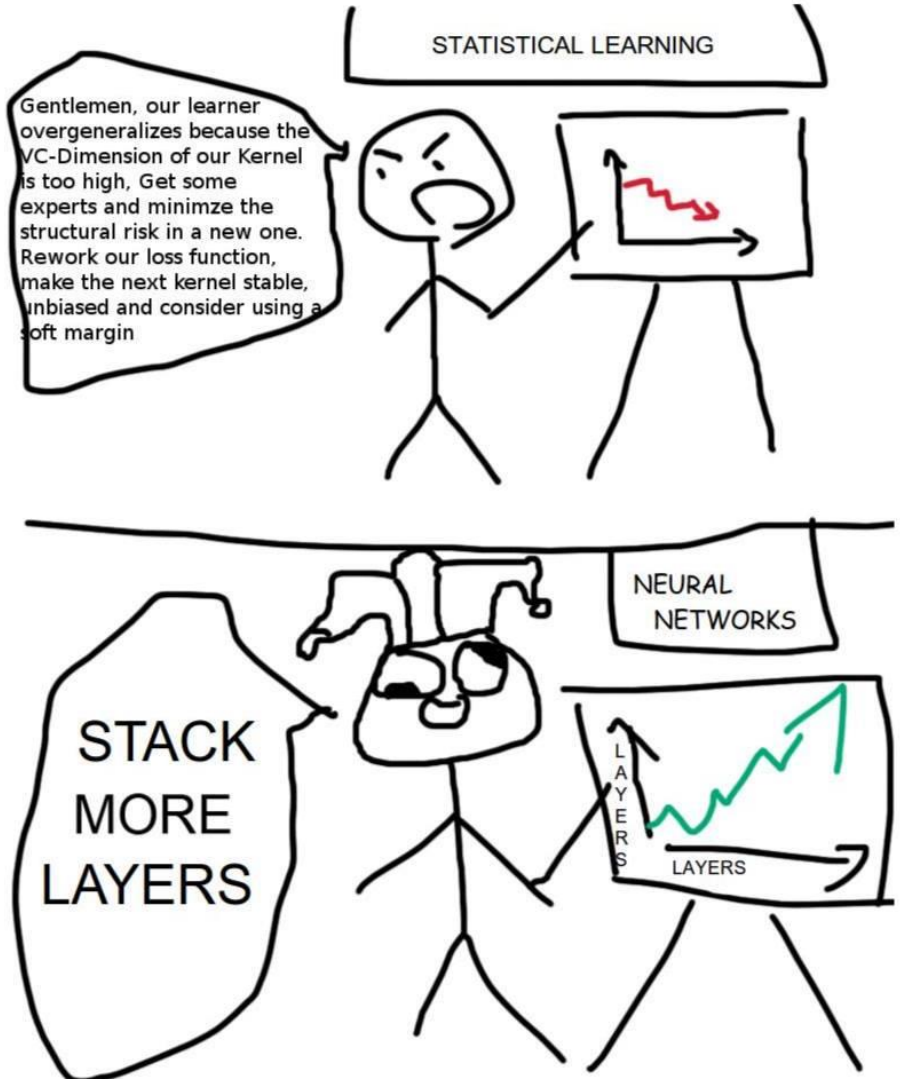



“충분한 모델 사이즈와 학습 시간이 있다면 어떤 아키텍쳐든 최종적으로 동일한 솔루션으로 수렴한다. 충분히 큰 diffusion conv-unet, ViT, AR 모델은 같은 이미지를 생성한다. 이는 아키텍쳐/하이퍼파라미터/옵티마이저와 무관하게, 오직 데이터셋에 의해서만 모델이 결정된다는 것을 말한다.”

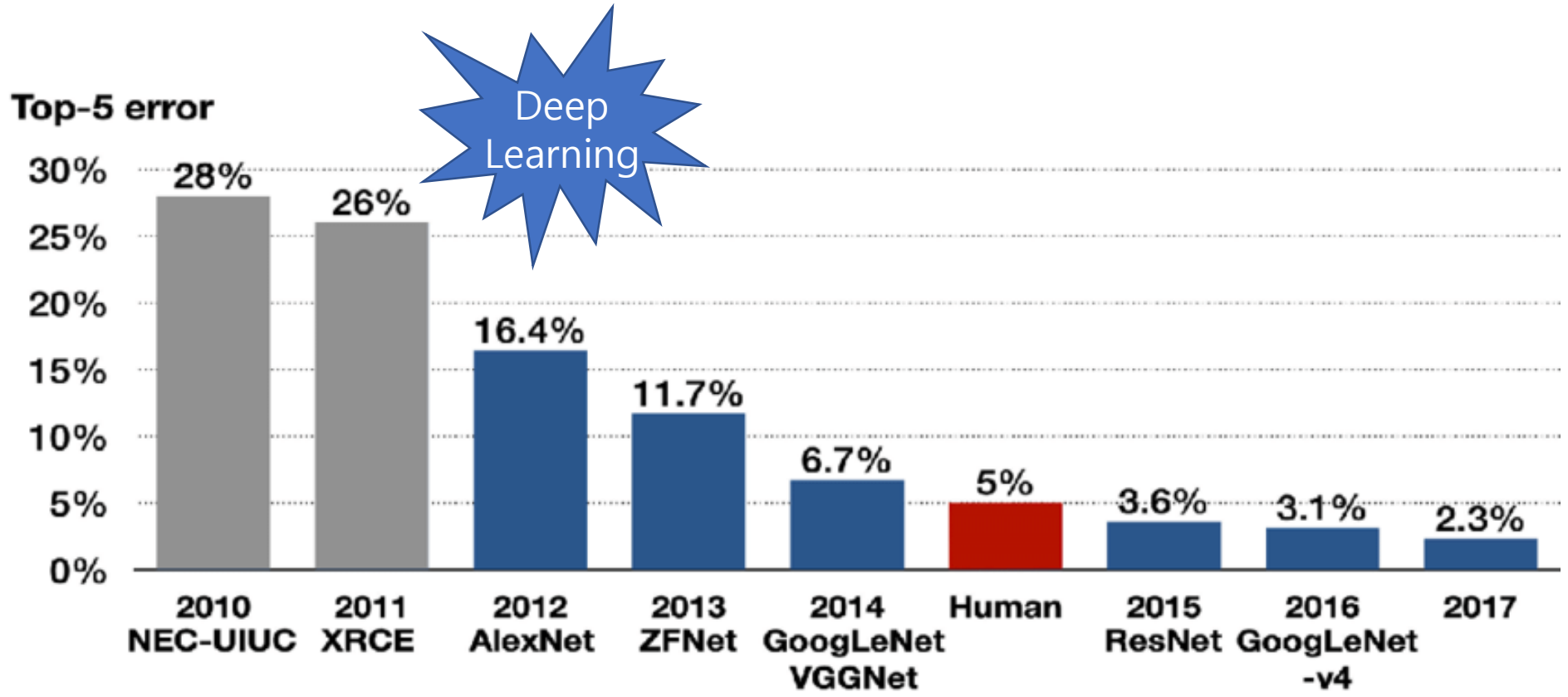
# The Bitter Lesson

Boundary of computing power
***is*** the state of machine learning.

# Deep Learning in summary



# History of ImageNet Challenge




General-purpose GPU (GPGPU)

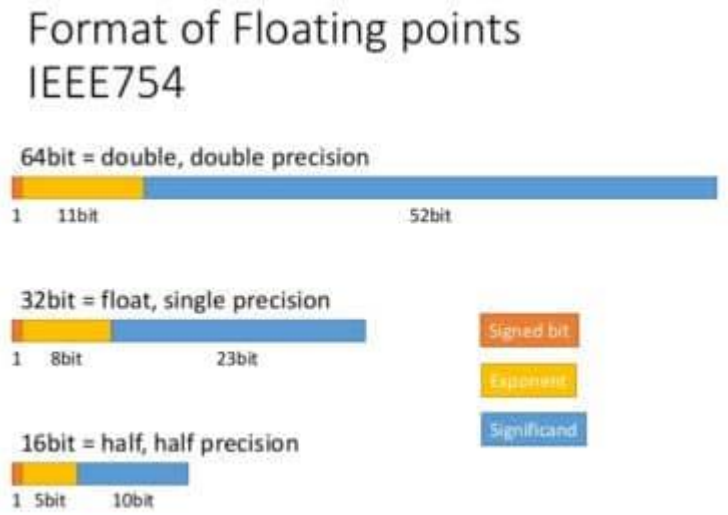
AI accelerator

# NVIDIA GPU Architectures

| | Maxwell (2014) | Pascal (2016) | Volta (2017) | Turing (2018) | Ampere (2020) | Hopper (2022) | Ada (2022) | Blackwell (2024?) |
|---|---|---|---|---|---|---|---|---|
| "Tesla" (datacenter) | M60 | P100 | V100 | T40 | A100 | H100 | L40 | |
| "Quadro" (professional) | M6000 | P6000 | | RTX 8000 | RTX A6000 | | RTX 6000 Ada | |
| "GeForce" (consumer) | Titan X<br>GTX 980Ti | Titan Xp<br>GTX 1080Ti | Titan V | Titan RTX<br>RTX 2080Ti | RTX 3090<br>RTX 3080 | | RTX 4090<br>RTX 4080 | |

# Arithmetic features for Deep Learning

| | Pascal (2016) | Volta (2017) | Ampere (2020) | Hopper (2022) |
|---|---|---|---|---|
| Training | FP16 | | TF32 | FP8 |
| Inference | FP16 | INT8 | | |



Arithmetics

Memory

Interconnect

# NVIDIA Pascal Architecture

NVIDIA

Whitepaper

# NVIDIA Tesla P100

***The Most Advanced Datacenter Accelerator Ever Built***

***Featuring Pascal GP100, the World's Fastest GPU***

Figure 1. NVIDIA Tesla P100 with Pascal GP100 GPU

NVIDIA GPU architecture with Deep Learning in mind


Figure 7. Pascal GP100 Full GPU with 60 SM Units


Figure 8. Pascal GP100 SM Unit

# FP16 arithmetic support

Reduce FP precision ->

- Less memory footprint,
- Faster arithmetic,
- Less data transfer,
- Sometimes better generalization

## Extreme Performance for High Performance Computing and Deep Learning

Tesla P100 was built to deliver exceptional performance for the most demanding compute applications, delivering:

- 5.3 TFLOPS of double precision floating point (FP64) performance
- 10.6 TFLOPS of single precision (FP32) performance
- 21.2 TFLOPS of half-precision (FP16) performance

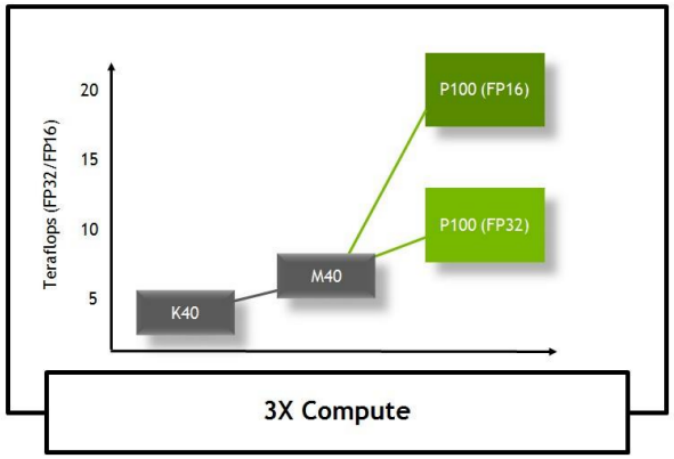



Figure 3. Tesla P100 Significantly Exceeds Compute Performance of Past GPU Generations

## Support for FP16 Arithmetic Speeds Up Deep Learning

Deep learning is one of the fastest growing fields of computing. It is a critical ingredient in many important applications, including real-time language translation, highly accurate image recognition, automatic image captioning, autonomous driving object recognition, optimal path calculations, collision avoidance, and others. Deep learning is a two-step process.

- First, a neural network must be trained.
- Second, the network is deployed in the field to run inference computations, where it uses the results of previous training to classify, recognize, and generally process unknown inputs.

Compared to CPUs, GPUs can provide tremendous performance speedups for Deep Learning training and inference.

Unlike other technical computing applications that require high-precision floating-point computation, deep neural network architectures have a natural resilience to errors due to the backpropagation algorithm used in their training. In fact, to avoid overfitting a network to a training dataset, approaches such as dropout aim at ensuring a trained network generalizes well and is not overly reliant on the accuracy of (or errors in) any given unit's computation.

Storing FP16 data compared to higher precision FP32 or FP64 reduces memory usage of the neural network and thus allows training and deploying of larger networks. Using FP16 computation improves performance up to 2x compared to FP32 arithmetic, and similarly FP16 data transfers take less time than FP32 or FP64 transfers.

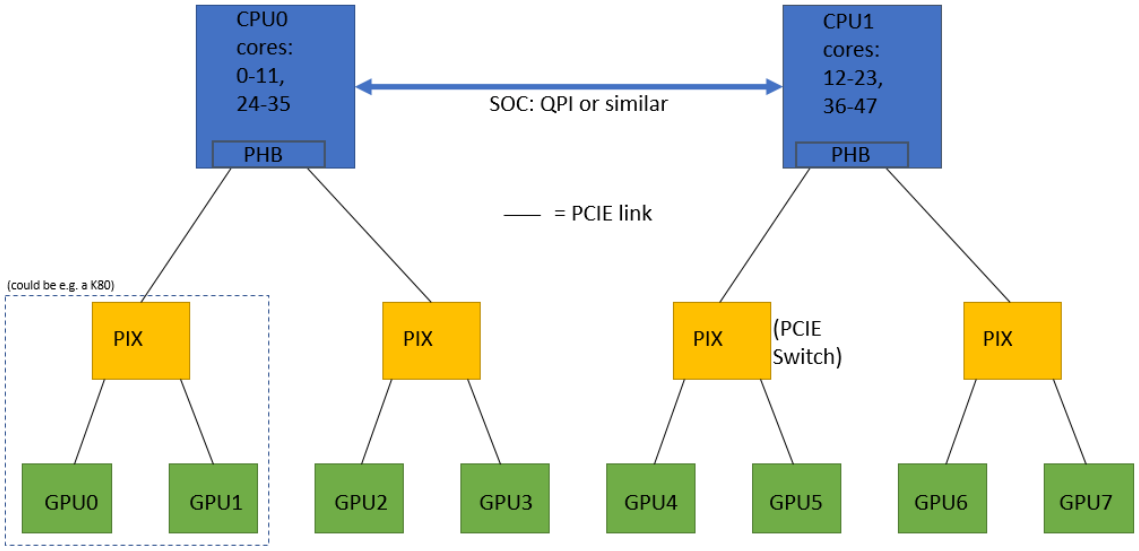
**Note:** In GP100, two FP16 operations can be performed using a single paired-operation instruction.

Architectural improvements in GP100, combined with support for FP16 datatypes allow significantly reduced Deep Learning processing times compared to what was achievable just last year.

# GPU-to-GPU communication

`$ nvidia-smi topo -m`

## Typical systems vs. NVLink



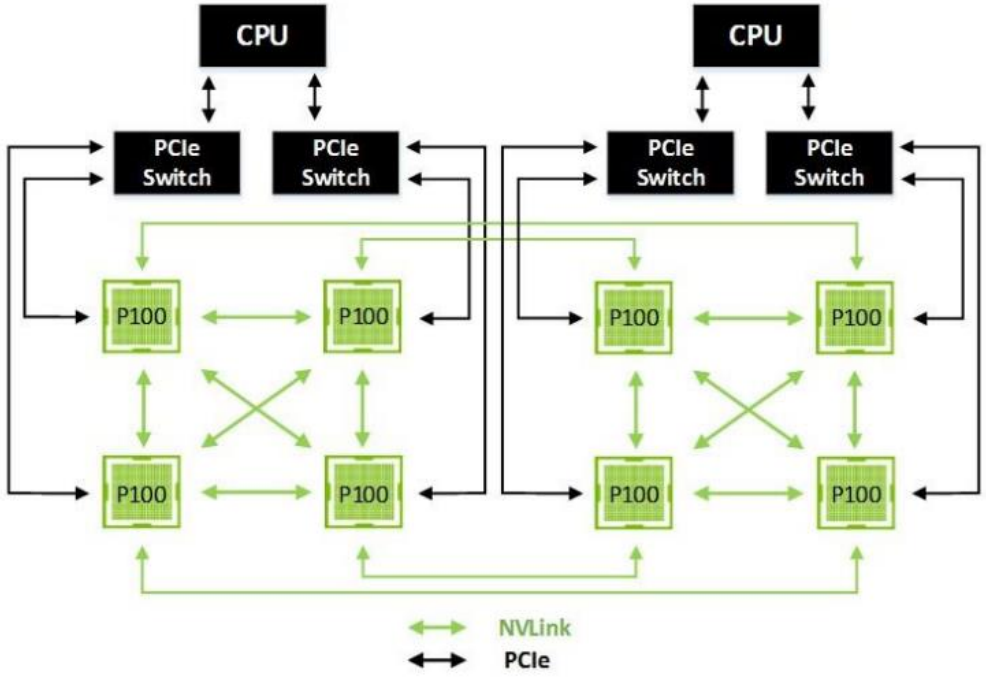




Figure 4. NVLink Connecting Eight Tesla P100 Accelerators in a Hybrid Cube Mesh Topology

# NVIDIA Volta Architecture

NVIDIA

# NVIDIA TESLA V100 GPU ARCHITECTURE

***THE WORLD'S MOST ADVANCED DATA CENTER GPU***


Figure 1. NVIDIA Tesla V100 SXM2 Module with Volta GV100 GPU

https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

# “Volta” V100 architecture

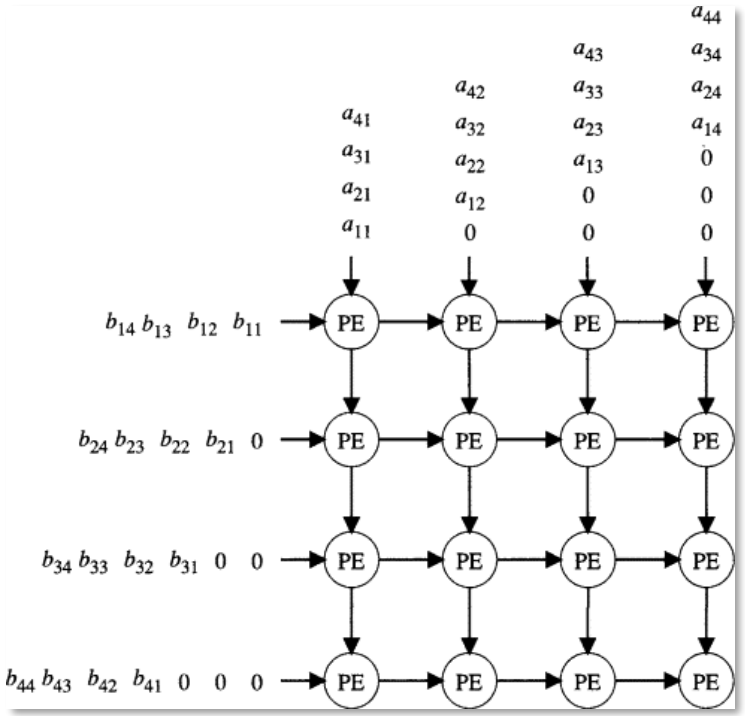

# TABLE OF CONTENTS



딥러닝 알고리즘이 본격적으로 반영된 아키텍쳐
- 딥러닝 학습 가속에 포커싱된 하드웨어 기능
- Tensor core
- SLI -> NVLink
- DGX-1 딥러닝용 슈퍼컴퓨터

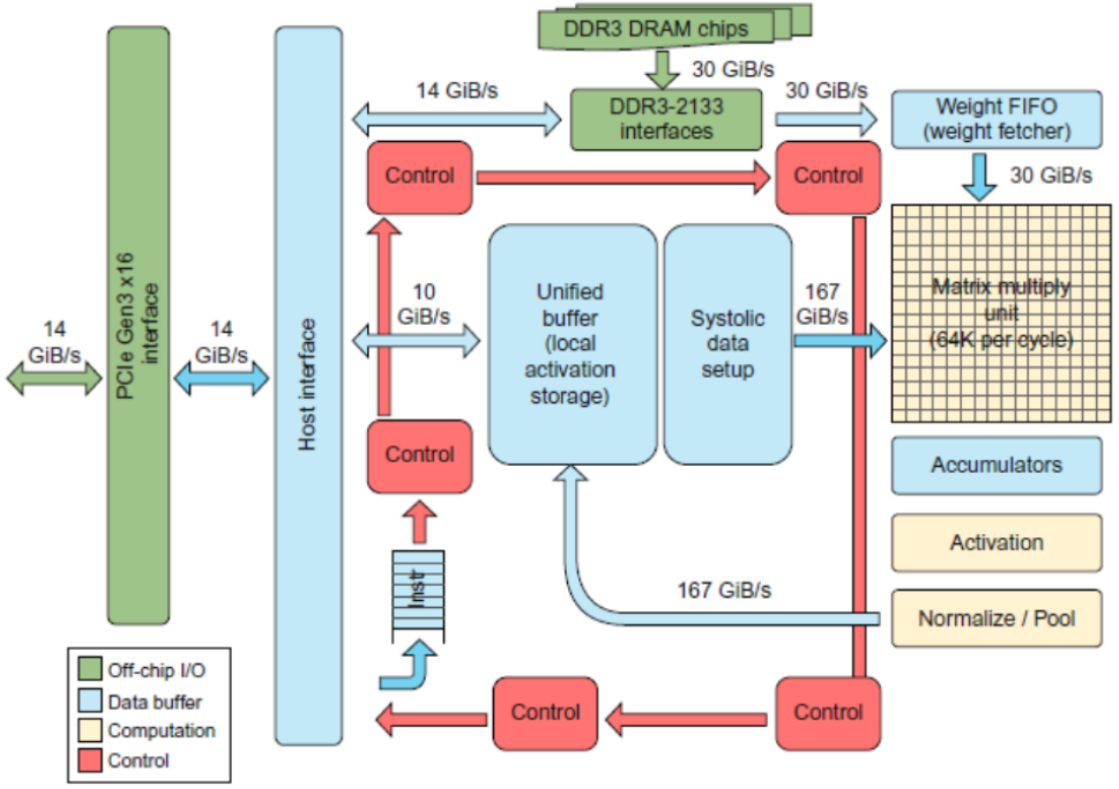
# 1st-gen Tensor Cores

Systolic arrays



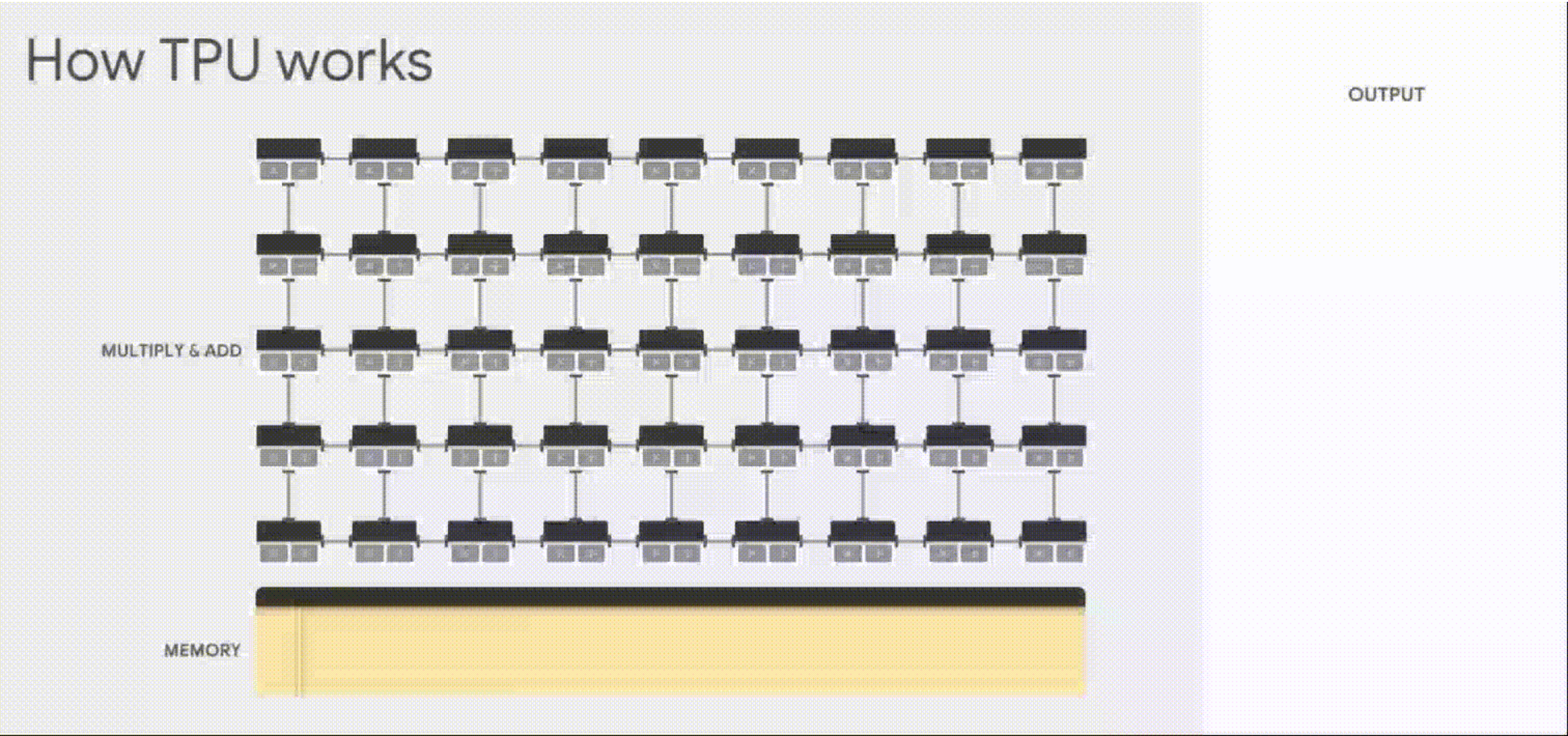

PE = multiply-and-accumulate unit

## Tensor Processing Unit (TPU)



https://tpudemo.com

9x5 FC layer
Batch size= 5



# Mixed Precision for AI training

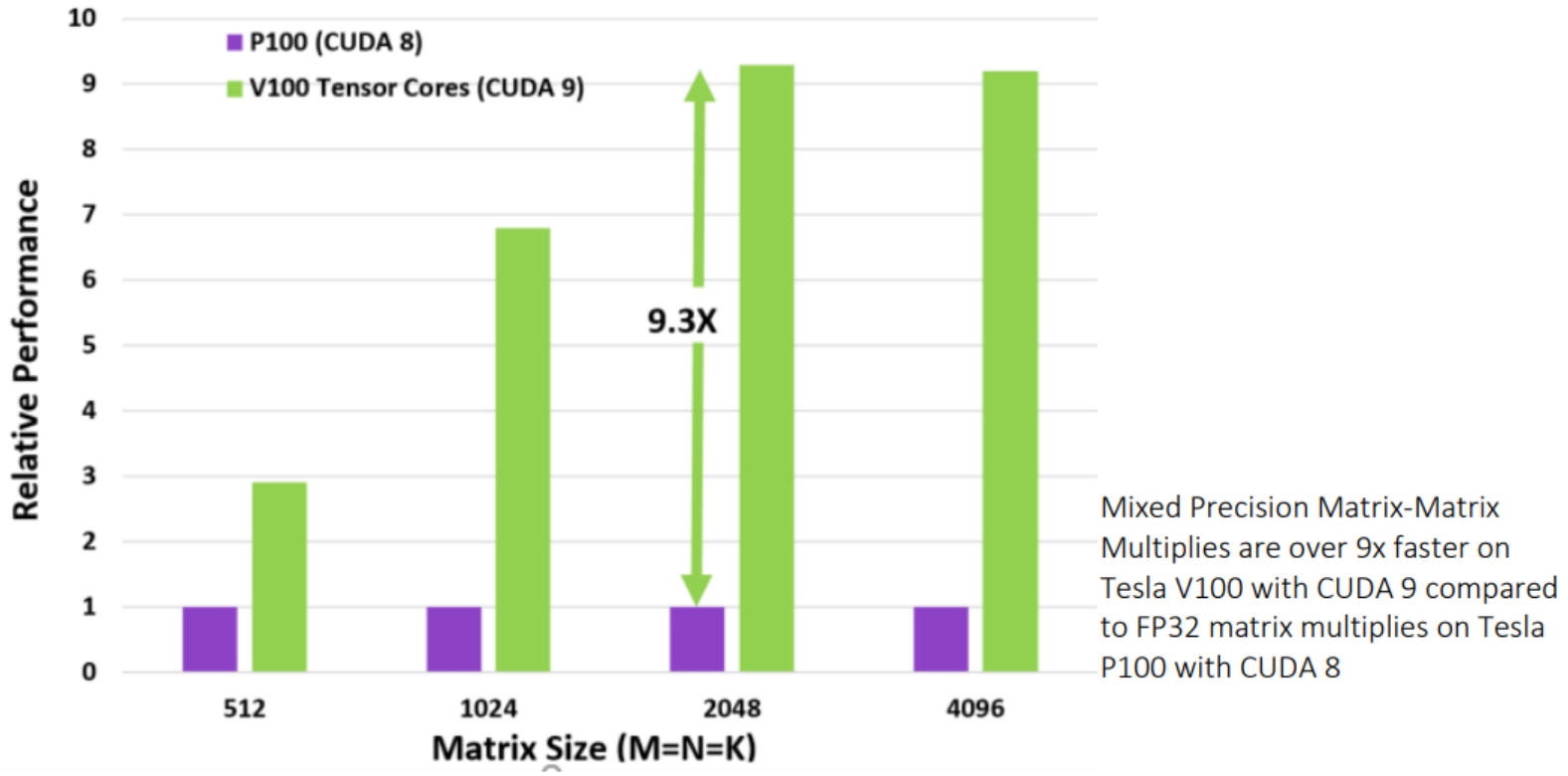



Figure 7. cuBLAS Mixed Precision (FP16 Input, FP32 Compute)

# Mixed Precision GEMM

Each Tensor Core operates on a 4x4 matrix and performs the following operation:

$D = A \times B + C$

where **A**, **B**, **C**, and **D** are 4x4 matrices (Figure 8). The matrix multiply inputs **A** and **B** are FP16 matrices, while the accumulation matrices **C** and **D** may be FP16 or FP32 matrices (see Figure 8).

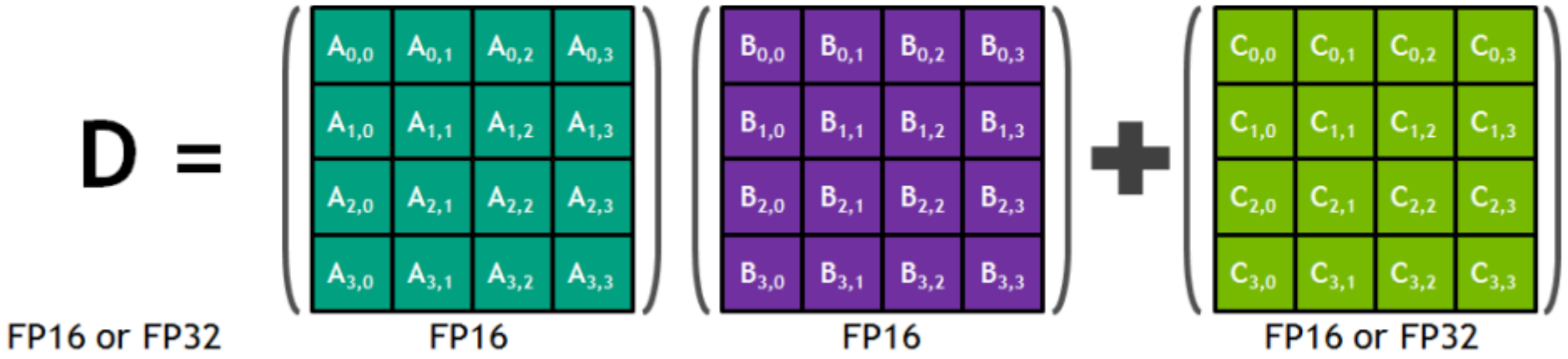
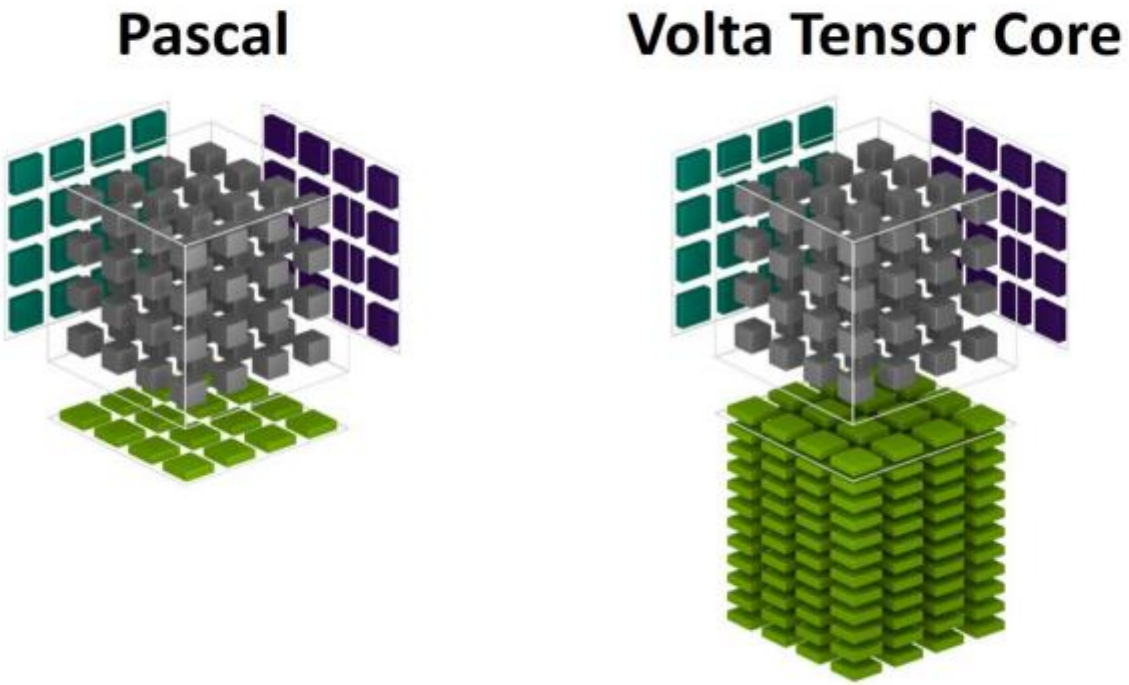



Figure 8. Tensor Core 4x4 Matrix Multiply and Accumulate

Figure 10 shows the 4x4 matrix multiplication (using the two source 4x4 matrices outside the cube) requiring 64 operations (represented by the cube) to generate a 4x4 output matrix (shown below the cube). The Volta-based V100 accelerator with Tensor Cores can perform such calculations at 12x faster rate than Pascal-based Tesla P100.



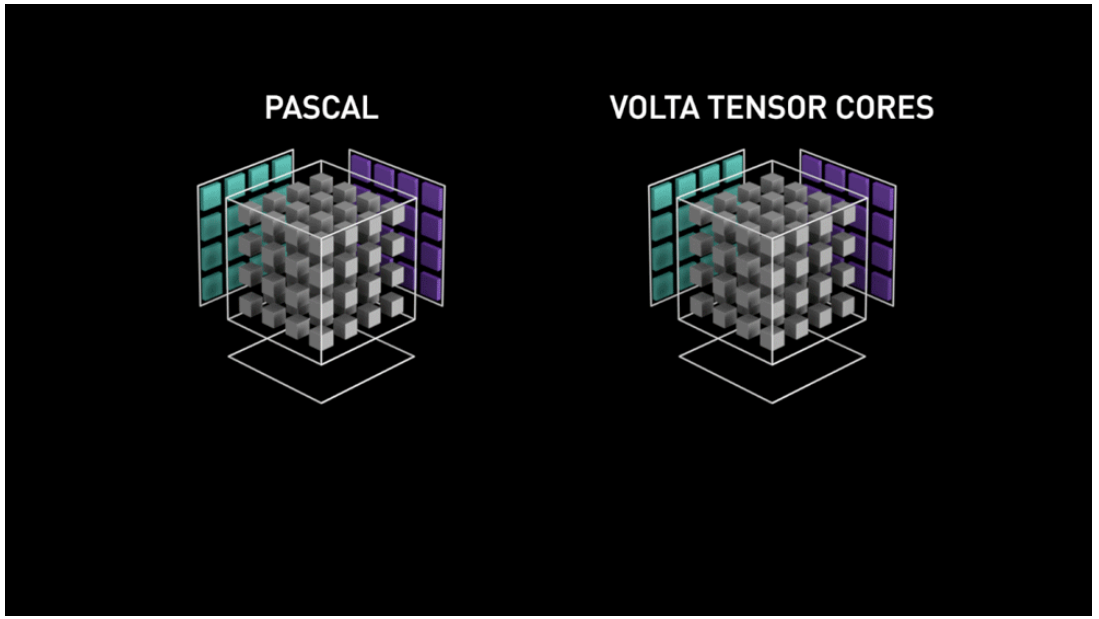




Figure 10. Pascal and Volta 4x4 Matrix Multiplication

# Fast computation in low-precision datatypes

| Standard format | Low-precision format |
|---|---|
| FP64, FP32 | FP16, BF16, TF32, … |

- Low-precision floating point format
  - Shrink from 32-bit format to 16-bit representation
  - Precision vs. Dynamic range
  - Saves memory cost and communication bandwidth
  - Reduced power cost = denser arithmetic units
- Low-precision Neural Nets
  - Inference is quite robust to weight truncation
  - Training gets instable due to limited dynamic range
- Mixed Precision Training (arXiv:1710.03740)
  - NVIDIA's recipe for FP16 training
  - Mixed use of FP16 and FP32 depending on layer type
  - ~2x training throughput

# Mixed Precision Training

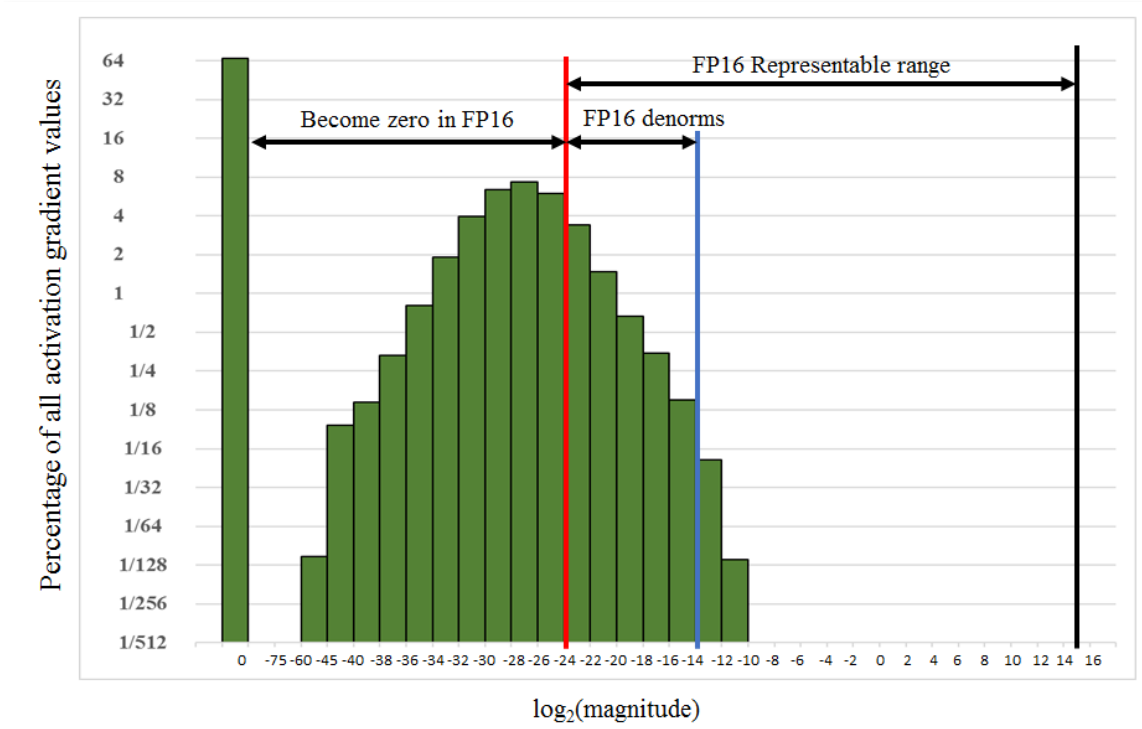
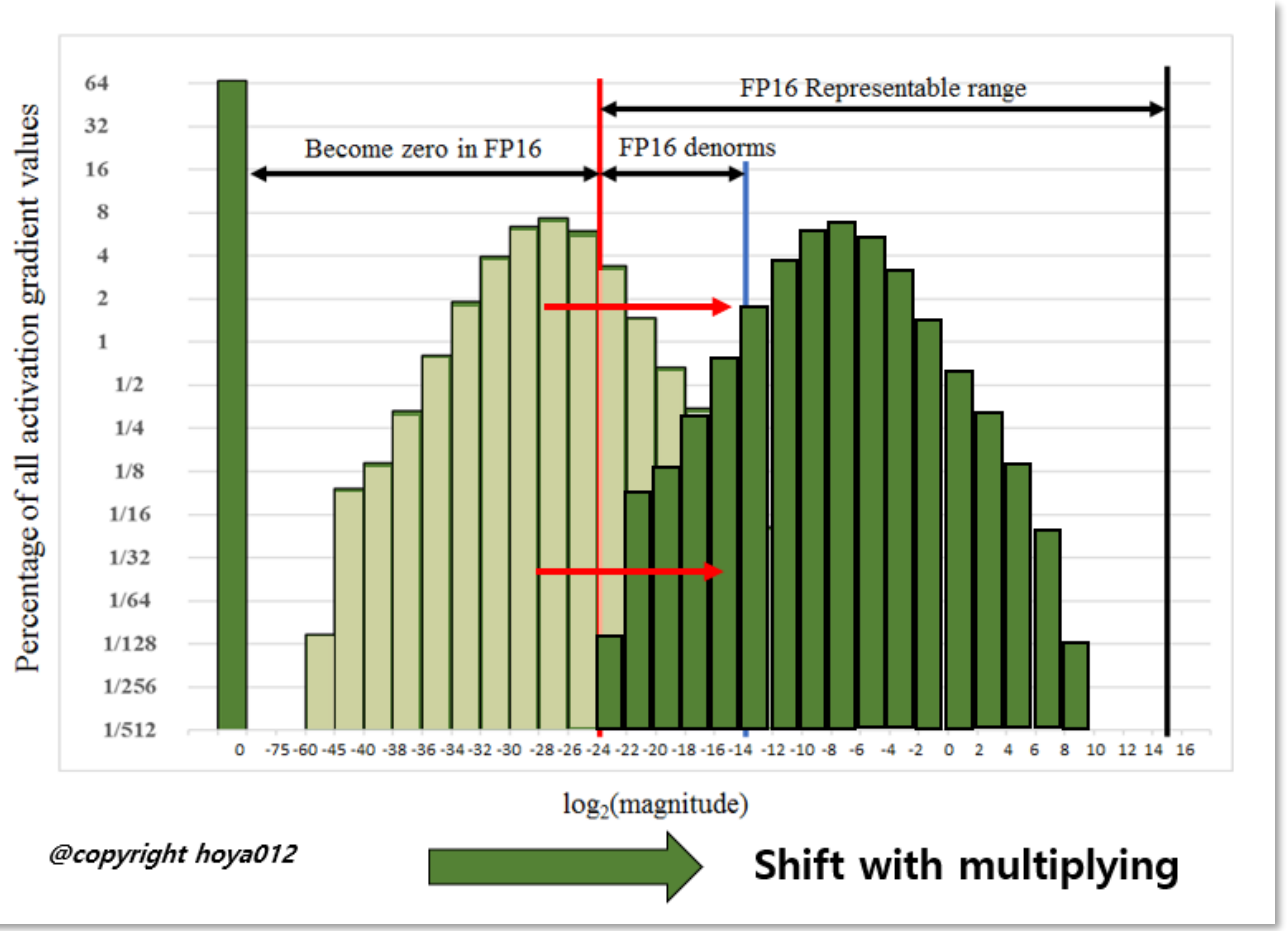



Figure 2. A histogram of activation gradients recorded when training the Multibox SSD detector network in single precision. The Y-axis is the percentage of all values on a log scale. The X-axis is the log scale of absolute values, as well as a special entry for zeros. For example, in this training session 66.8% of values were zero, whereas 4% of values were between $2^{-32}$ and $2^{-30}$.



# NVIDIA Ampere GPU Architecture

https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf


Figure 6. GA100 Full GPU with 128 SMs (A100 Tensor Core GPU has 108 SMs)


Figure 7. GA100 Streaming Multiprocessor (SM)

## A100 Tensor Cores Support All DL Data Types

In addition to FP16 precision introduced on the Volta Tensor Core, and the INT8, INT4 and binary 1-bit precisions added in the Turing Tensor Core, the A100 Tensor Core adds support for TF32, BF16 and FP64 formats. (FP64 double-precision MMA is discussed in the next section).

Volta GPU architecture introduced Tensor Cores that operate on IEEE FP16 data types, providing 8x more math throughput compared to V100 FP32. Results are accumulated into FP32 for mixed precision training or FP16 for inference. From a raw architectural performance perspective, if both A100 and V100 were operating at the same clock speed, a single A100 SM delivers 2x FP16 Tensor Core performance compared to the V100 SM, and 16x compared to standard V100 (and A100) FP32 FFMA operations.

Turing architecture extended Tensor Cores to handle more inference use cases by adding INT8, INT4, and Binary support. On Turing these provided 16x, 32x, and 128x more math throughput when compared to FP32. The A100 SM delivers 2x INT8, INT4, and Binary Tensor Core performance compared to the Turing SM, respectively, and 32x, 64x, and 256x compared to A100 FP32 FFMA.

NVIDIA Ampere architecture adds three additional formats to Tensor Cores – BF16, TF32 and FP64. BF16 is an alternative to IEEE FP16, and includes an 8-bit exponent, 7-bit mantissa, and 1 sign-bit. Both FP16 and BF16 have been shown to successfully train neural networks in mixed-precision mode, matching FP32 training results without hyper-parameter adjustment. Both FP16 and BF16 modes of Tensor Cores provide 16x more math throughput than FP32 in A100 GPUs.

Today, the default math for AI training is FP32, without tensor core acceleration. The NVIDIA Ampere architecture introduces new support for TF32, enabling AI training to use tensor cores by default with no effort on the user's part. Non-tensor operations continue to use the FP32 datapath, while TF32 tensor cores read FP32 data and use the same range as FP32 with reduced internal precision, before producing a standard IEEE FP32 output. TF32 includes an 8-bit exponent (same as FP32), 10-bit mantissa (same precision as FP16) and 1 sign-bit.

As with Volta, Automatic Mixed Precision (AMP) enables users to use mixed precision with FP16 for AI training with just a few lines of code changes. Using AMP, A100 delivers a further 2X faster Tensor Core performance over TF32.

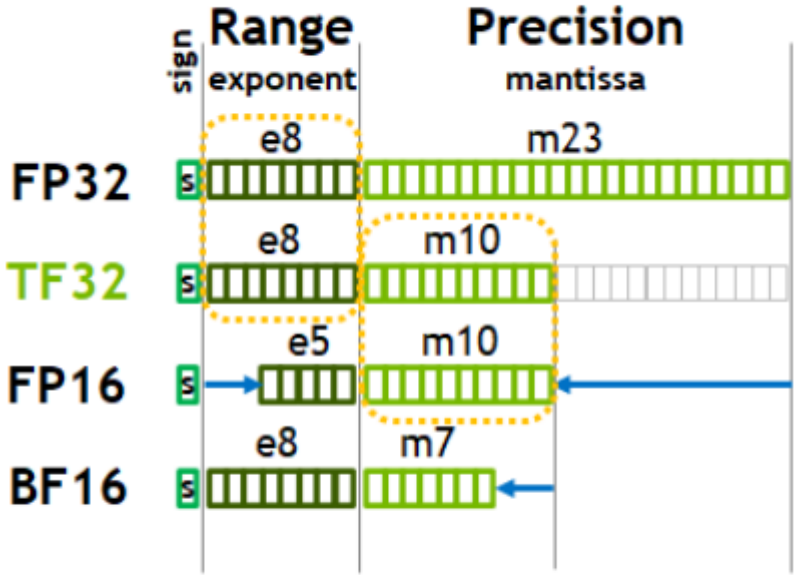
# 2nd-gen Tensor Cores

Table 3. A100 Tensor Core Input / Output Formats and Performance vs FP32 FFMA.

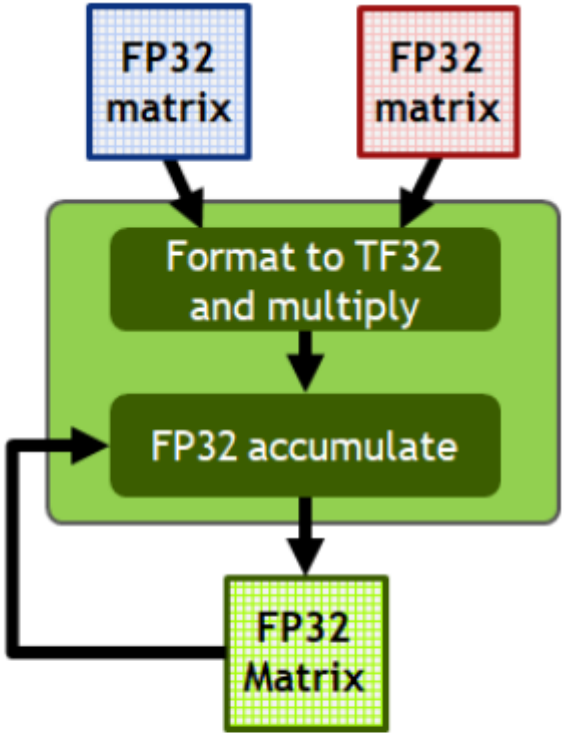
| | INPUT OPERANDS | ACCUMULATOR | TOPS | X-factor vs. FFMA | SPARSE TOPS | SPARSE X-factor vs. FFMA |
|---|---|---|---|---|---|---|
| V100 | FP32 | FP32 | 15.7 | 1x | - | - |
| | FP16 | FP32 | 125 | 8x | - | - |
| A100 | FP32 | FP32 | 19.5 | 1x | - | - |
| | TF32 | FP32 | 156 | 8x | 312 | 16x |
| | FP16 | FP32 | 312 | 16x | 624 | 32x |
| | BF16 | FP32 | 312 | 16x | 624 | 32x |
| | FP16 | FP16 | 312 | 16x | 624 | 32x |
| | INT8 | INT32 | 624 | 32x | 1248 | 64x |
| | INT4 | INT32 | 1248 | 64x | 2496 | 128x |
| | BINARY | INT32 | 4992 | 256x | - | - |
| | IEEE FP64 | | 19.5 | 1x | - | - |

**Note:** TOPS column indicates TFLOPS for floating-point ops and TOPS for integer ops. X-factors compare MMA ops with and without sparsity to standard FP32 FFMA ops. (Sparse TOPS represents effective TOPS / TFLOPS using the new Sparsity feature.)

# 2nd-gen Tensor Cores



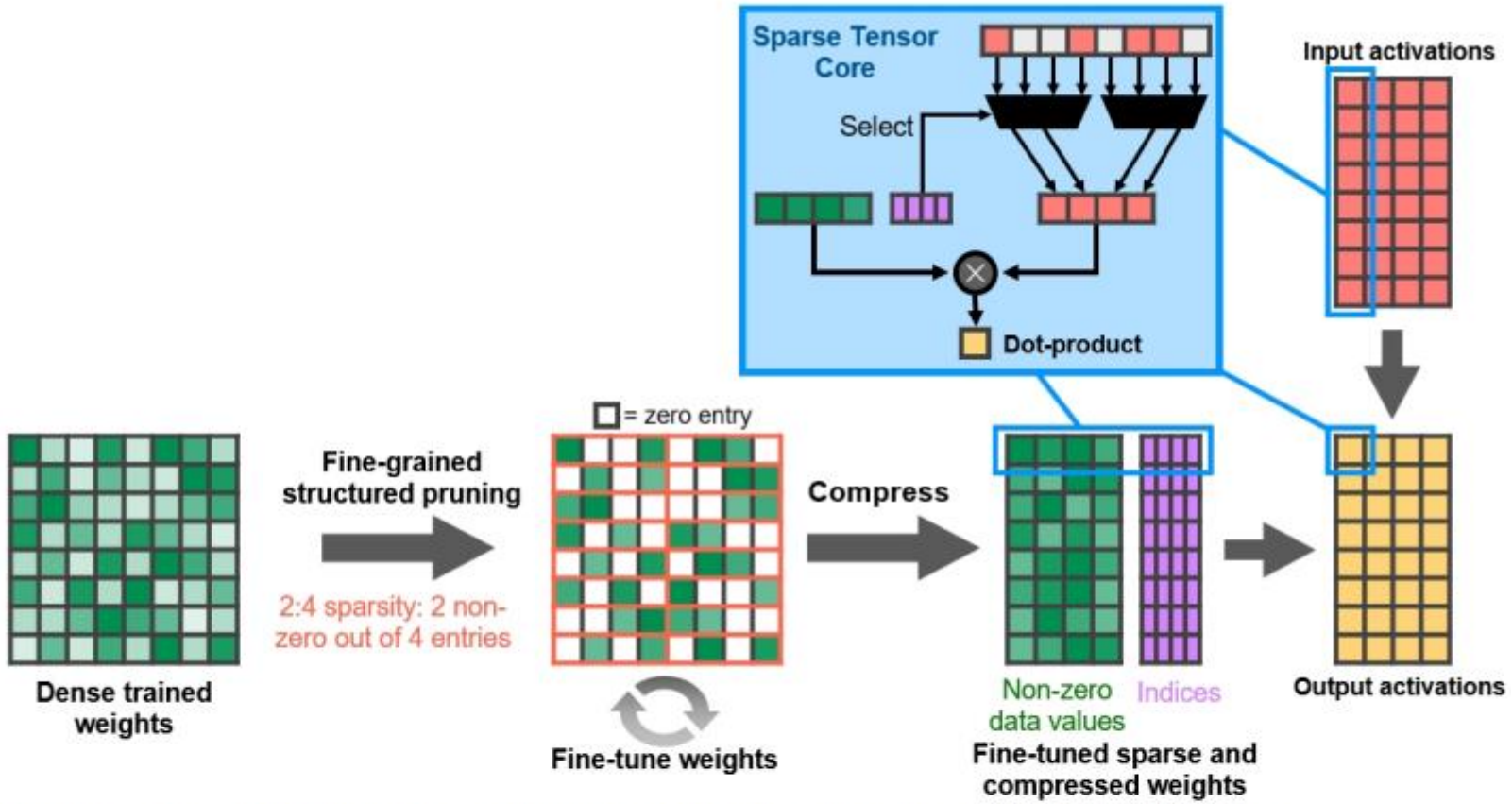

TF32 computation




TensorFloat-32 (TF32) provides the range of FP32 with the precision of FP16, 8x precision vs. BF16 (left). A100 accelerates tensor math with TF32 while supporting FP32 input and output data (right), enabling easy integration into DL and HPC programs and automatic acceleration of DL frameworks.

Figure 9. TensorFloat-32 (TF32)

# Fine-grained Structured Sparsity




A100 Fine-Grained Structured Sparsity prunes trained weights with a 2-out-of-4 non-zero pattern, followed by a simple and universal recipe for fine-tuning the non-zero weights. The weights are compressed for a 2x reduction in data footprint and bandwidth, and the A100 Sparse Tensor Core doubles math throughput by skipping the zeros.

Figure 12. A100 Fine-Grained Structured Sparsity

https://timdettmers.com/2022/08/17/llm-int8-and-emergent-features/

# INT8

We have demonstrated for the first time that multi-billion parameter transformers can be quantized to Int8 and used immediately for **inference** without performance degradation.

## LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale


Tim Dettmers[λ*] Mike Lewis[†] Younes Belkada[§∓] Luke Zettlemoyer[†λ]

University of Washington[λ]
Facebook AI Research[†]
Hugging Face[§]
ENS Paris-Saclay[∓]

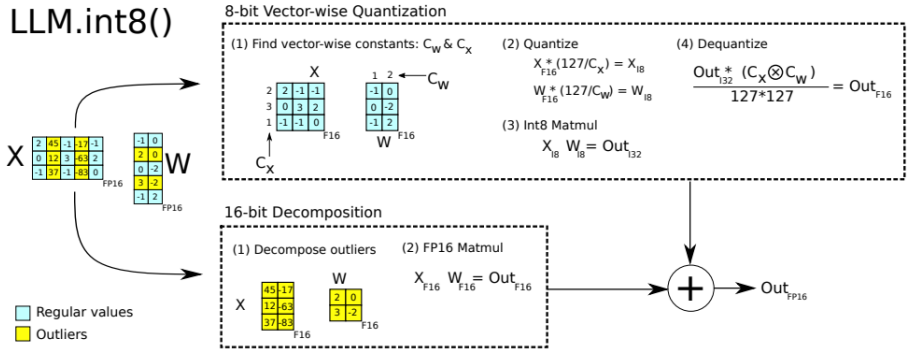




Figure 2: Schematic of LLM.int8(). Given 16-bit floating-point inputs $\mathbf{X}_{f16}$ and weights $\mathbf{W}_{f16}$, the features and weights are decomposed into sub-matrices of large magnitude features and other values. The outlier feature matrices are multiplied in 16-bit. All other values are multiplied in 8-bit. We perform 8-bit vector-wise multiplication by scaling by row and column-wise absolute maximum of $\mathbf{C}_x$ and $\mathbf{C}_w$ and then quantizing the outputs to Int8. The Int32 matrix multiplication outputs $\mathbf{Out}_{i32}$ are dequantization by the outer product of the normalization constants $\mathbf{C}_x \otimes \mathbf{C}_w$. Finally, both outlier and regular outputs are accumulated in 16-bit floating point outputs.

Table 2: Different hardware setups and which methods can be run in 16-bit vs. 8-bit precision. We can see that our 8-bit method makes many models accessible that were not accessible before, in particular, OPT-175B/BLOOM.

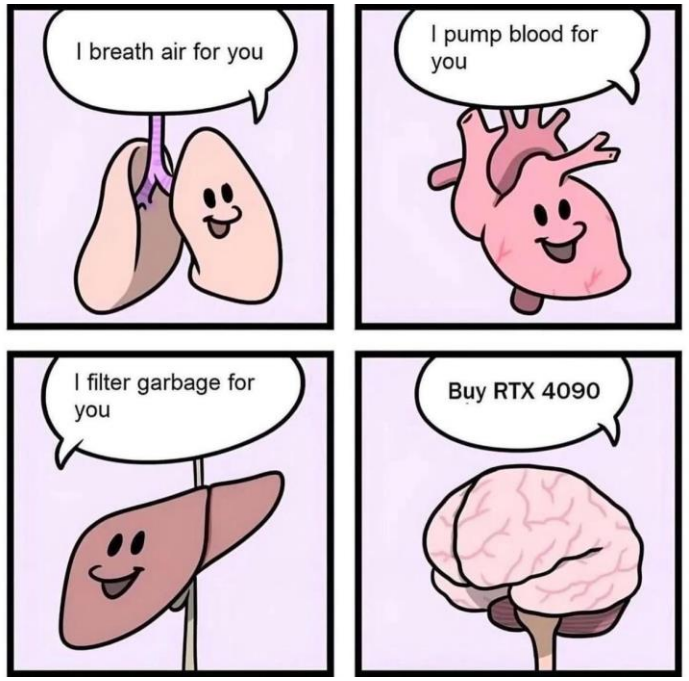
| | | | Largest Model that can be run | |
|---|---|---|---|---|
| Class | Hardware | GPU Memory | 8-bit | 16-bit |
| Enterprise | 8x A100 | 80 GB | **OPT-175B / BLOOM** | **OPT-175B / BLOOM** |
| Enterprise | 8x A100 | 40 GB | **OPT-175B / BLOOM** | OPT-66B |
| Academic server | 8x RTX 3090 | 24 GB | **OPT-175B / BLOOM** | OPT-66B |
| Academic desktop | 4x RTX 3090 | 24 GB | **OPT-66B** | OPT-30B |
| Paid Cloud | Colab Pro | 15 GB | **OPT-13B** | GPT-J-6B |
| Free Cloud | Colab | 12 GB | **T0/T5-11B** | GPT-2 1.3B |


Dettmers, Tim, et al. "Llm. int8 (): 8-bit matrix multiplication for transformers at scale." *arXiv preprint arXiv:2208.07339* (2022).

NVIDIA H100 Tensor Core GPU Architecture

EXCEPTIONAL PERFORMANCE, SCALABILITY, AND SECURITY FOR THE DATA CENTER

Includes final GPU / memory clocks and final TFLOPS performance specs. V1.04




# NVIDIA Hopper GPU Architecture

https://resources.nvidia.com/en-us-tensor-core


*Figure 3. GH100 Full GPU with 144 SMs*


*Figure 4. GH100 streaming multiprocessor*

# 3rd-gen Tensor Core

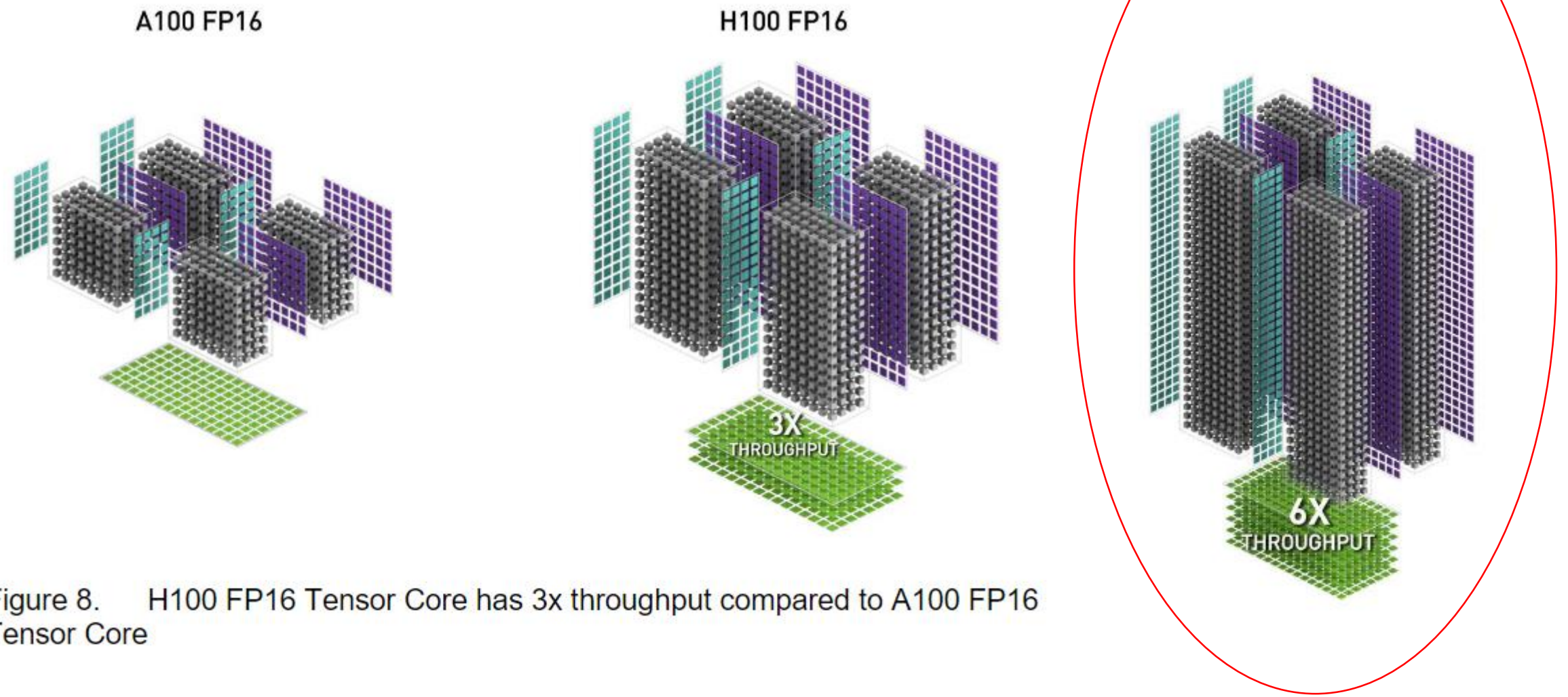
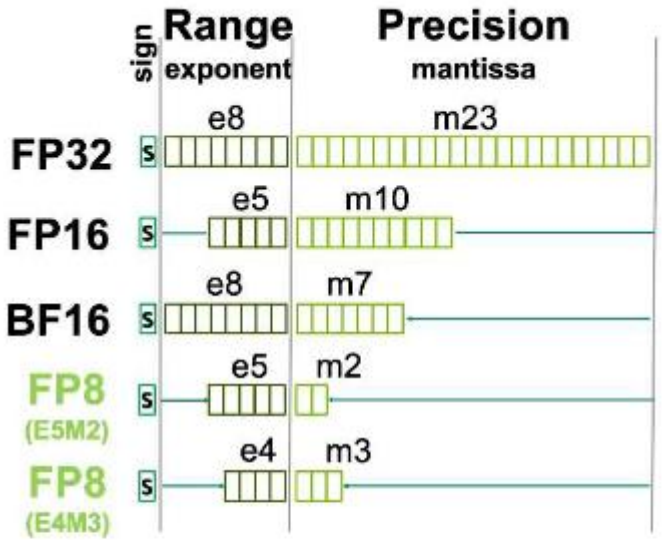
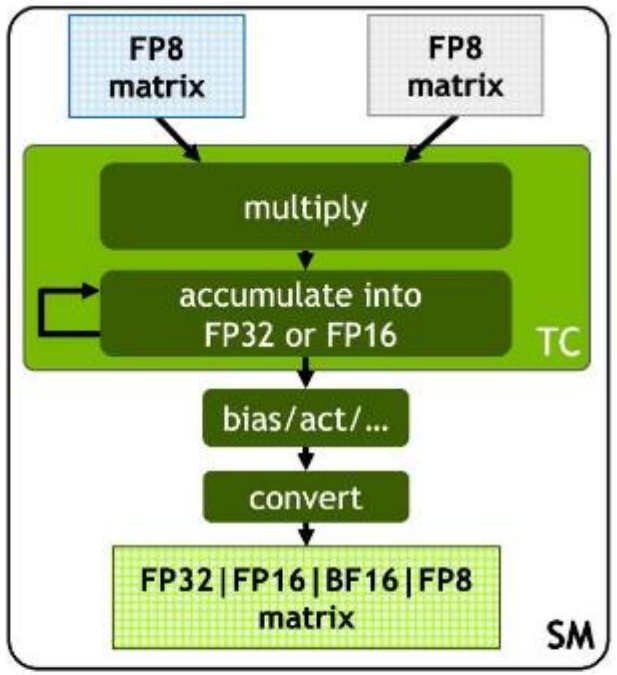



Figure 8. H100 FP16 Tensor Core has 3x throughput compared to A100 FP16 Tensor Core

# FP8 data format




Allocate 1 bit to either range or precision




Support for multiple accumulator and output types

Figure 9. New Hopper FP8 Precisions - 2x throughput and half the footprint of H100 FP16 / BF16

https://docs.nvidia.com/deeplearning/transformer-engine/user-guide/examples/fp8_primer.html

# FP8 training



## FP8 FORMATS FOR DEEP LEARNING

**Paulius Micikevicius, Dusan Stosic, Patrick Judd, John Kamalu, Stuart Oberman, Mohammad Shoeybi, Michael Siu, Hao Wu**
NVIDIA
{pauliusm, dstosic, pjudd, jkamalu, soberman, mshoeybi, msiu, skyw}@nvidia.com

**Neil Burgess, Sangwon Ha, Richard Grisenthwaite**
Arm
{neil.burgess, sangwon.ha, richard.grisenthwaite}@arm.com

**Naveen Mellempudi, Marius Cornea, Alexander Heinecke, Pradeep Dubey**
Intel
{naveen.k.mellempudi, marius.cornea, alexander.heinecke, pradeep.dubey}@intel.com

**ABSTRACT**

FP8 is a natural progression for accelerating deep learning training inference beyond the 16-bit formats common in modern processors. In this paper we propose an 8-bit floating point (FP8) binary interchange format consisting of two encodings - E4M3 (4-bit exponent and 3-bit mantissa) and E5M2 (5-bit exponent and 2-bit mantissa). While E5M2 follows IEEE 754 conventions for representatio of special values, E4M3's dynamic range is extended by not representing infinities and having only one mantissa bit-pattern for NaNs. We demonstrate the efficacy of the FP8 format on a variety of image and language tasks, effectively matching the result quality achieved by 16-bit training sessions. Our study covers the main modern neural network architectures - CNNs, RNNs, and Transformer-based models, leaving all the hyperparameters unchanged from the 16-bit baseline training sessions. Our training experiments include large, up to 175B parameter, language models. We also examine FP8 post-training-quantization of language models trained using 16-bit formats that resisted fixed point int8 quantization.

Table 2: Image Classification Models, ILSVRC12 Validation Top-1 Accuracy

| Model | Baseline | FP8 |
|---|---|---|
| VGG-16 | 71.27 | 71.11 |
| VGG-16 BN | 73.95 | 73.69 |
| Inception v3 | 77.23 | 77.06 |
| DenseNet 121 | 75.59 | 75.33 |
| DenseNet 169 | 76.97 | 76.83 |
| Resnet18 | 70.58 | 70.12 |
| Resnet34 | 73.84 | 73.72 |
| Resnet50 v1.5 | 76.71 | 76.76 |
| Resnet101 v1.5 | 77.51 | 77.48 |
| ResNeXt50 | 77.68 | 77.62 |
| Xception | 79.46 | 79.17 |
| MobileNet v2 | 71.65 | 71.04 |
| DeiT small | 80.08 | 80.02 |

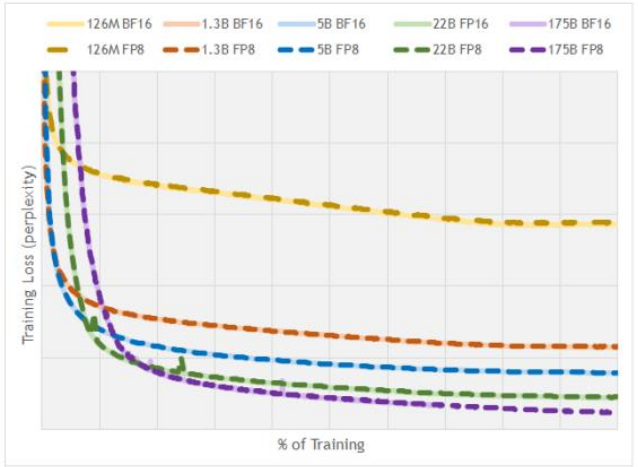



Figure 1: Training loss (perplexity) curves for various GPT-3 models. x-axis is normalized number of iterations.

Table 5: Post training quantization of models trained in 16-bit floating point. For F1 metrics higher is better, for perplexity lower is better. Best 8-bit result is bolded.

| Model | Dataset (metric) | 16-bit FP | int8 | E4M3 |
|---|---|---|---|---|
| BERT Base | SQuAD v1.1 (F1) | 88.19 | 76.89 | **88.09** |
| BERT Large | SQuAD v1.1 (F1) | 90.87 | 89.65 | **90.94** |
| GPT3 126M | wikitext103 (perplexity) | 19.01 | 28.37 | **19.43** |
| GPT3 1.3B | wikitext103 (perplexity) | 10.19 | 12.74 | **10.29** |
| GPT3 6.7B | wikitext103 (perplexity) | 8.51 | 10.29 | **8.41** |

Micikevicius, Paulius, et al. "FP8 formats for deep learning." *arXiv preprint arXiv:2209.05433* (2022).

# FP8 mixed precision training

PyTorch

```python
import torch
import transformer_engine.pytorch as te
from transformer_engine.common import recipe

# Set dimensions.
in_features = 768
out_features = 3072
hidden_size = 2048

# Initialize model and inputs.
model = te.Linear(in_features, out_features, bias=True)
inp = torch.randn(hidden_size, in_features, device="cuda")

# Create an FP8 recipe. Note: All input args are optional.
fp8_recipe = recipe.DelayedScaling(margin=0, interval=1, fp8_format=recipe.Format.E4M3)

# Enable autocasting for the forward pass
with te.fp8_autocast(enabled=True, fp8_recipe=fp8_recipe):
    out = model(inp)

loss = out.sum()
loss.backward()
```

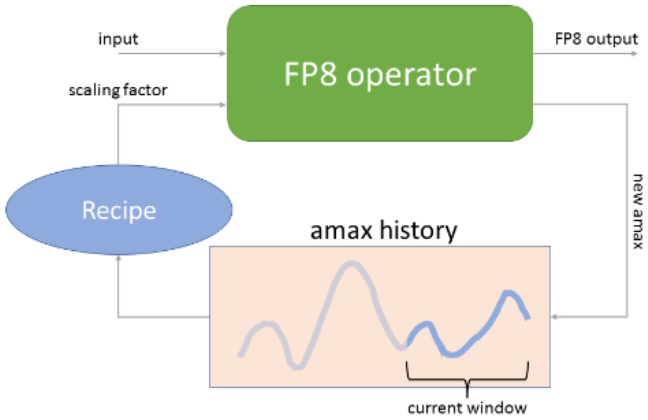
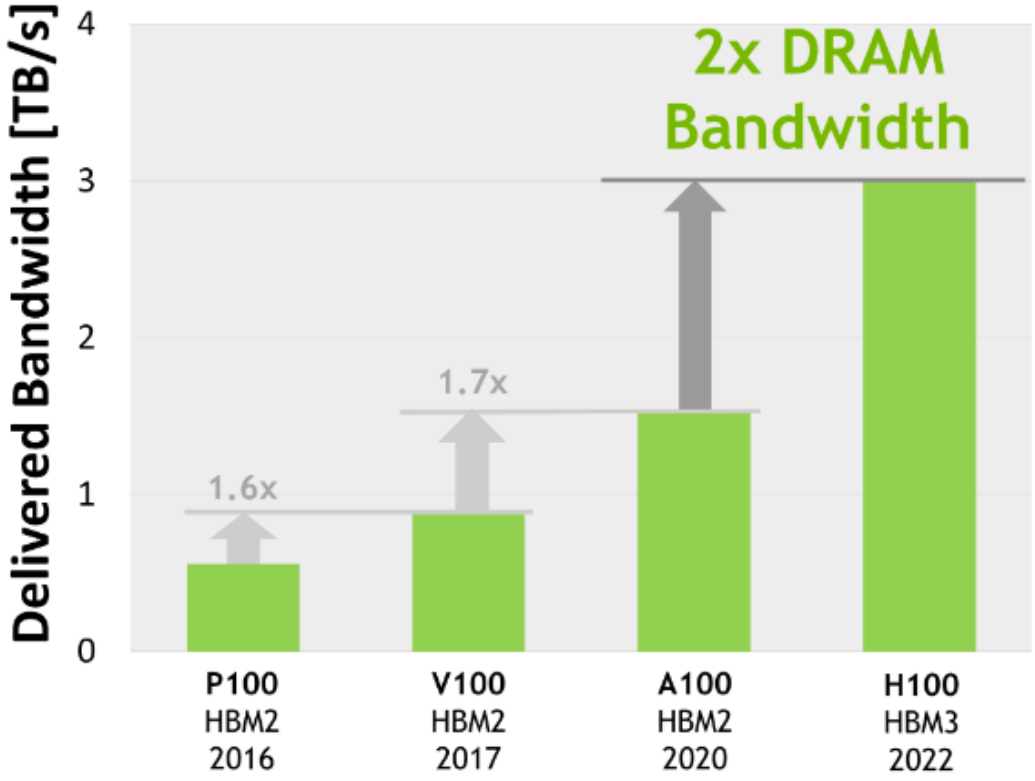



Figure 3: Delayed scaling strategy. The FP8 operator uses scaling factor obtained using the history of amaxes (maximums of absolute values) seen in some number of previous iterations and produces both the FP8 output and the current amax, which gets stored in the history.

https://docs.nvidia.com/deeplearning/transformer-engine/user-guide/examples/fp8_primer.html

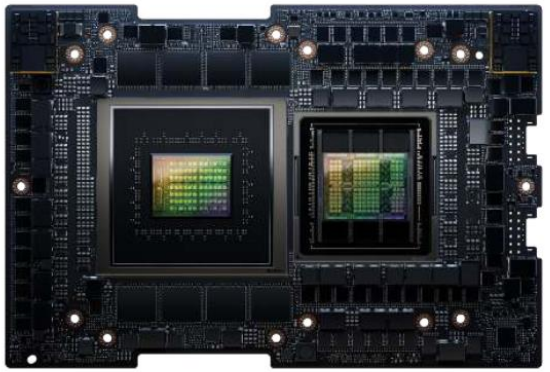
# HBM3 Memory




Memory data rates not finalized and subject to change in the final product.

Figure 21. World's First HBM3 GPU Memory Architecture, 2x Delivered Bandwidth

Inside NVIDIA's First GPU-CPU Superchip



The NVIDIA® GH200 Grace Hopper architecture brings together the groundbreaking performance of the NVIDIA Hopper GPU with the versatility of the NVIDIA Grace™ CPU, connected with a high bandwidth and memory coherent NVIDIA NVLink Chip-2-Chip (C2C)® interconnect in a single Superchip, and support for the new NVIDIA NVLink Switch System.

# NVIDIA Grace Hopper Superchip Architecture

https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

# Grace Hopper Superchip

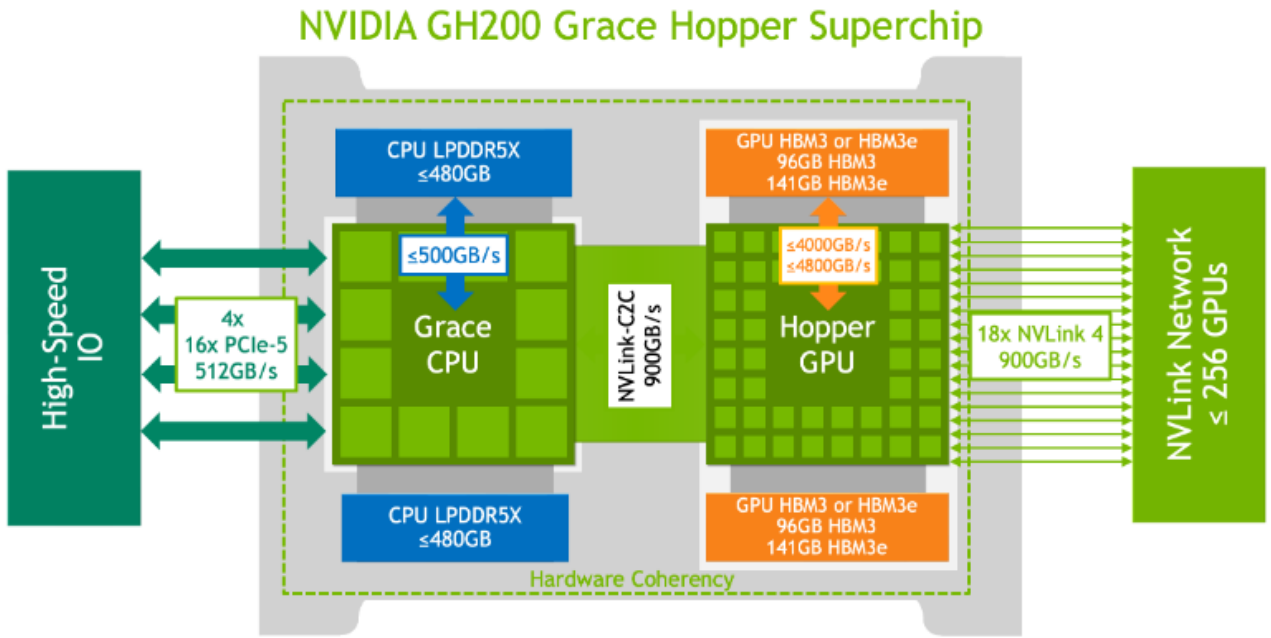



Figure 1. NVIDIA GH200 Grace Hopper Superchip Logical Overview

# NVIDIA DGX GH200 AI Supercomputer

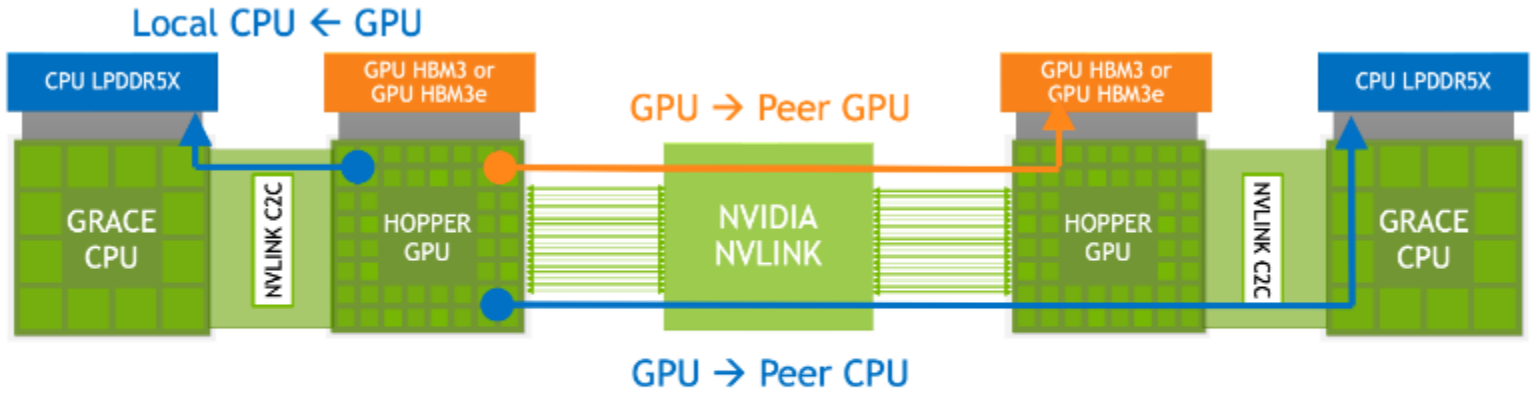



Figure 5. Memory Accesses across NVLink-connected Grace Hopper Superchips

# Hardware Accelerated Memory Coherency

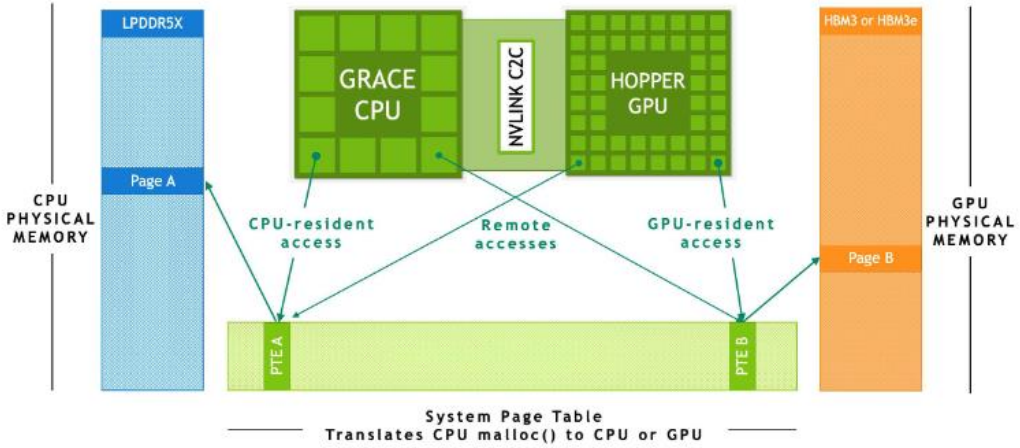



Figure 8. ATS in an NVIDIA Grace Hopper Superchip System

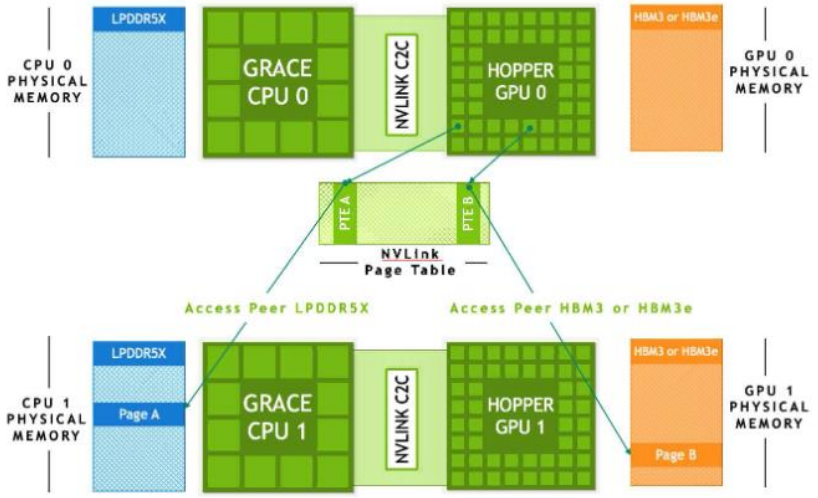



Figure 10. GPU threads address peer memory from other superchips in the NVLink Switch network

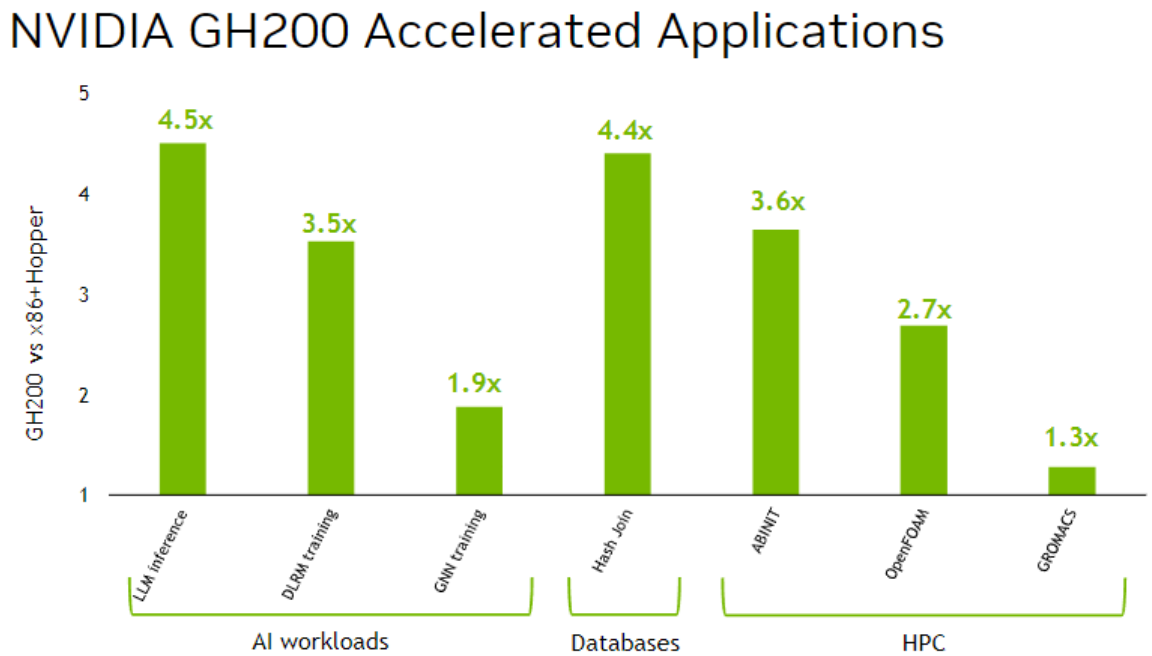



Figure 12. Performance Simulations for GH200 with HBM3 vs x86+Hopper for end-user applications. Datasets and details described in the individual application section below.

# LLM inference with offloading

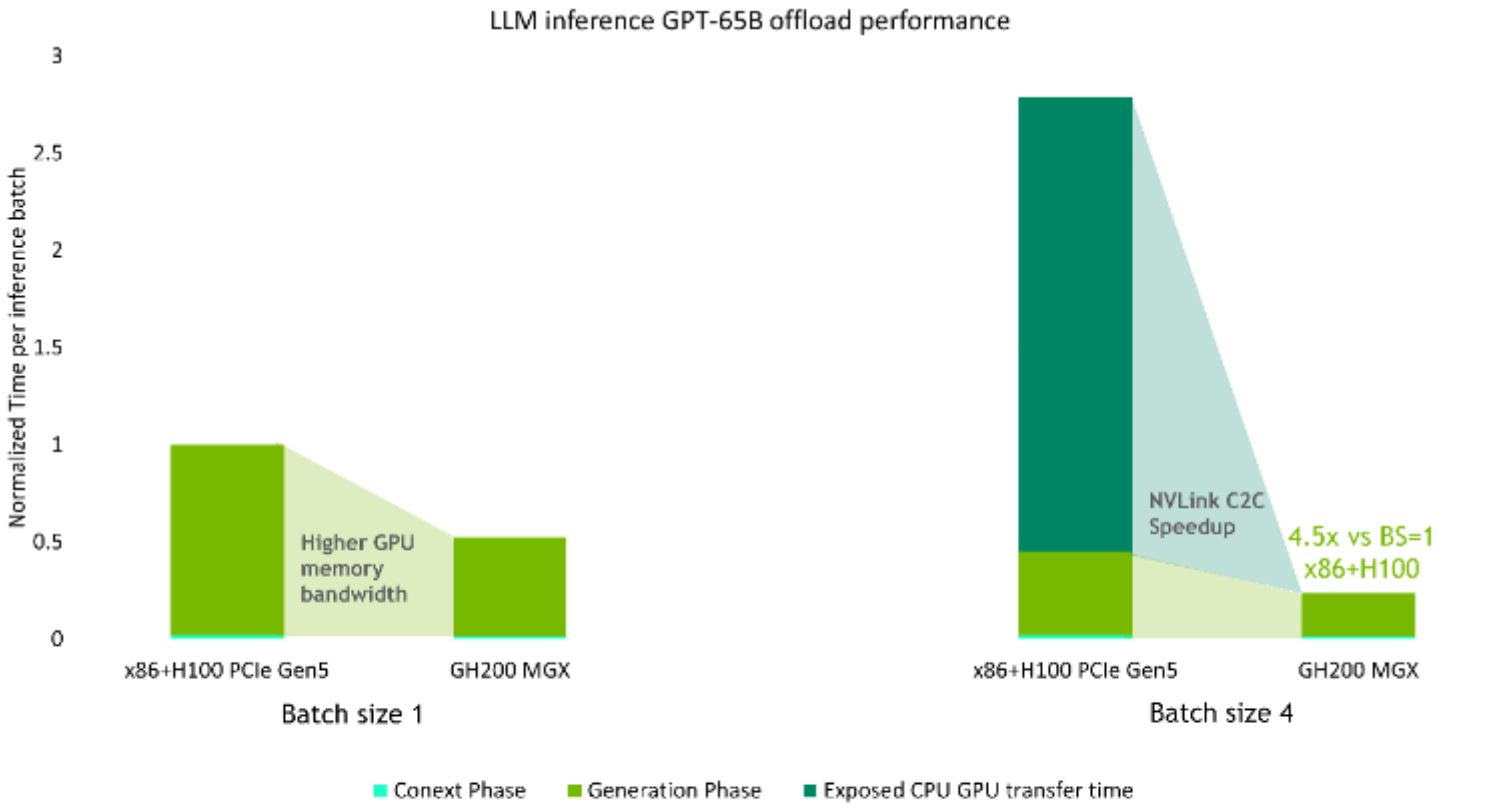
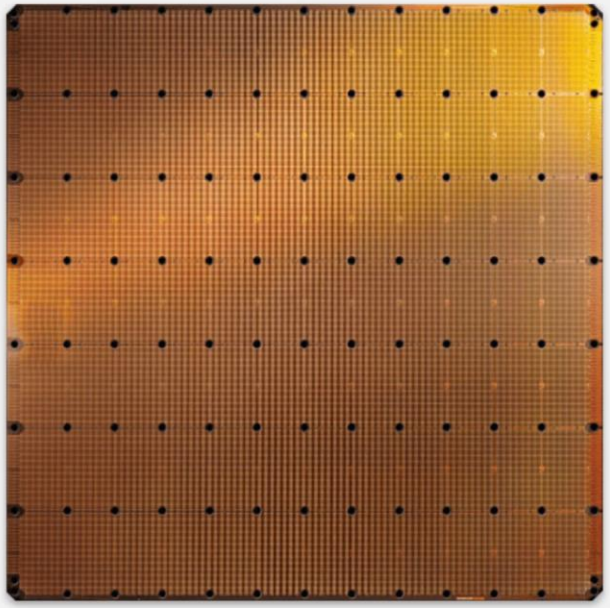



Figure 13. A performance simulation for LLM inference GPT3 65B LLM Model implemented with offloading.

# Future of AI accelerators

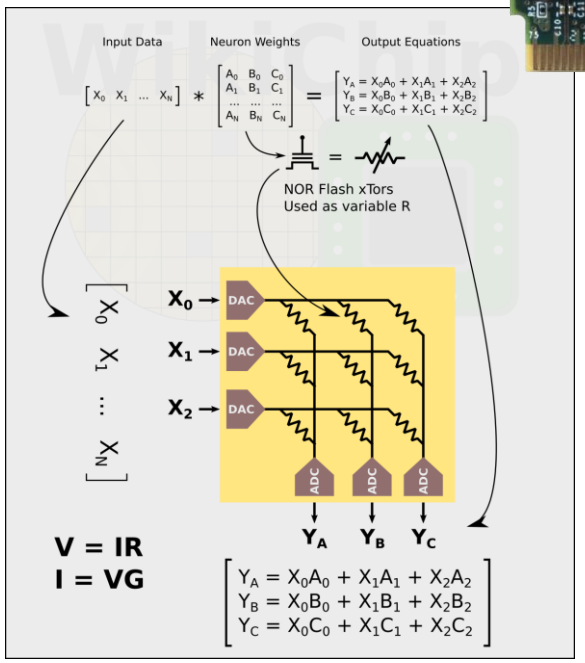
cerebras



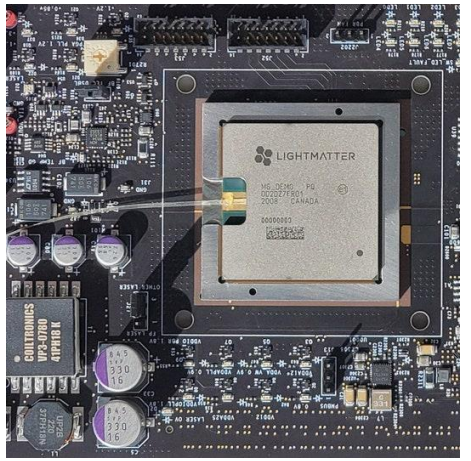
**Cerebras WSE-2**
46,225mm² Silicon
2.6 Trillion transistors

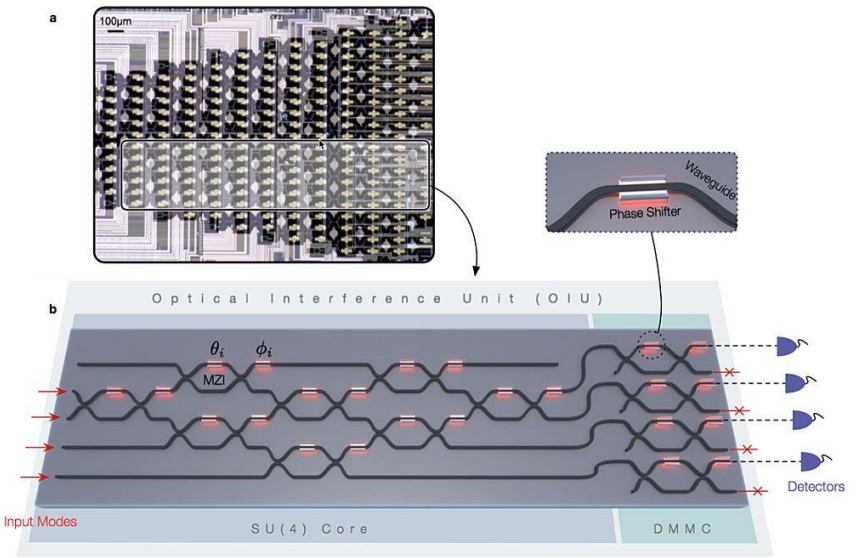
**Largest GPU**
826mm² Silicon
54.2 Billion transistors

Analog accelerators


Photonic accelerators



# ANN vs. Biological brains

## # of synaptic weights in a biological brain

On average, the human brain contains about 100 billion neurons and many more neuroglia which serve to support and protect the neurons. Each neuron may be connected to up to 10,000 other neurons, passing signals to each other via as many as 1,000 trillion synapses. 2019. 5. 30.

arxiv.org
https://arxiv.org › q-bio

Basic Neural Units of the Brain: Neurons, Synapses ... - arXiv

### Background [ edit ]

*Further information: GPT-3 § Background, and GPT-2 § Background*

OpenAI introduced the first GPT model (GPT-1) in 2018, publishing a paper called "Improving Language Understanding by Generative Pre-Training."[7] It was based on the transformer architecture and trained on a large corpus of books.[8] The next year, they introduced GPT-2, a larger model that could generate coherent text.[9] In 2020, they introduced GPT-3, a model with 100 times as many parameters as GPT-2, that could perform various tasks with few examples.[10] GPT-3 was further improved into GPT-3.5, which was used to create the chatbot product ChatGPT.

Rumors claim that GPT-4 has 1.76 trillion parameters, which was first estimated by the speed it was running and by George Hotz.[11]

https://en.wikipedia.org/wiki/GPT-4

## Estimated FLOPS of a biological brain = 1 petaFLOPS ~ 1 exaFLOPS

| | | | |
|---|---|---|---|
| (Neumann, 1958) | Storing all impulses over a lifetime. | | $10^{20}$ bits |
| (Wang, Liu et al., 2003) | Memories are stored as relations between neurons. | | $10^{8432}$ bits (See footnote [17]) |
| (Freitas Jr., 1996) | $10^{10}$ neurons, 1,000 synapses, firing 10 Hz [Level 4] | $10^{14}$ bits/second | |
| (Bostrom, 1998) | $10^{11}$ neurons, $5{\cdot}10^{3}$ synapses, 100 Hz, each signal worth 5 bits. [Level 5] | $10^{17}$ operations per second | |
| (Merkle, 1989a) | Energy constraints on Ranvier nodes. | $2{\cdot}10^{15}$ operations per second ($10^{13}$-$10^{16}$ ops/s) | |
| (Moravec, 1999; Morevec, 1988; Moravec, 1998) | Compares instructions needed for visual processing primitives with retina, scales up to brain and 10 times per second. Produces 1,000 MIPS neurons. [Level 3] | $10^{8}$ MIPS | $8{\cdot}10^{14}$ bits. |
| (Merkle, 1989a) | Retina scale-up. [Level 3] | $10^{12}$–$10^{14}$ operations per second. | |
| (Dix, 2005) | 10 billion neurons, 10,000 synaptic operations per cycle, 100 Hz cycle time. [Level 4] | $10^{16}$ synaptic ops/s | $4{\cdot}10^{15}$ bits (for structural information) |
| (Cherniak, 1990) | $10^{10}$ neurons, 1,000 synapses each. [Level 4] | | $10^{13}$ bits |
| (Fiala, 2007) | $10^{14}$ synapses, identity coded by 48 bits plus 2x36 bits for pre- and postsynaptic neuron id, 1 byte states. 10 ms update time. [Level 4] | 256,000 terabytes/s | $2{\cdot}10^{16}$ bits (for structural information) |
| (Seitz) | 50-200 billion neurons, 20,000 shared synapses per neuron with 256 distinguishable levels, 40 Hz firing. [Level 5] | $2{\cdot}10^{12}$ synaptic operations per secon | $4{\cdot}10^{15}$ - $8{\cdot}10^{15}$ bits |
| (Malickas, 1996) | $10^{11}$ neurons, $10^{2}$-$10^{4}$ synapses, 100-1,000 Hz activity. [level 4] | $10^{15}$-$10^{18}$ synaptic operations per secon | |
| | $1{\cdot}10^{11}$ neurons, each with $10^{4}$ compartments running the basic Hodgkin-Huxley equations with 1200 FLOPS each (based on (Izhikevich, 2004)). Each compartment would have 4 dynamical variables and 10 parameters described by one byte each. | $1.2{\cdot}10^{18}$ FLOPS | $1.12{\cdot}10^{28}$ bits |

| Technical Specifications | | | |
|---|---|---|---|
| | **H100 SXM** | **H100 PCIe** | **H100 NVL[1]** |
| **FP64** | 34 teraFLOPS | 26 teraFLOPS | 68 teraFLOPS |
| **FP64 Tensor Core** | 67 teraFLOPS | 51 teraFLOPS | 134 teraFLOPS |
| **FP32** | 67 teraFLOPS | 51 teraFLOPS | 134 teraFLOPS |
| **TF32 Tensor Core** | 989 teraFLOPS[2] | 756 teraFLOPS[2] | 1,979 teraFLOPS[2] |
| **BFLOAT16 Tensor Core** | 1,979 teraFLOPS[2] | 1,513 teraFLOPS[2] | 3,958 teraFLOPS[2] |
| **FP16 Tensor Core** | 1,979 teraFLOPS[2] | 1,513 teraFLOPS[2] | 3,958 teraFLOPS[2] |
| **FP8 Tensor Core** | 3,958 teraFLOPS[2] | 3,026 teraFLOPS[2] | 7,916 teraFLOPS[2] |
| **INT8 Tensor Core** | 3,958 TOPS[2] | 3,026 TOPS[2] | 7,916 TOPS[2] |
| **GPU memory** | 80GB | 80GB | 188GB |
| **GPU memory bandwidth** | 3.35TB/s | 2TB/s | 7.8TB/s[3] |
| **Decoders** | 7 NVDEC<br>7 JPEG | 7 NVDEC<br>7 JPEG | 14 NVDEC<br>14 JPEG |
| **Max thermal design power (TDP)** | Up to 700W (configurable) | 300-350W (configurable) | 2x 350-400W (configurable) |
| **Multi-instance GPUs** | Up to 7 MIGs @ 10GB each | Up to 7 MIGs @ 10GB each | Up to 14 MIGs @ 12GB each |
| **Form factor** | SXM | PCIe<br>> dual-slot<br>> air-cooled | 2x PCIe<br>> dual-slot<br>> air-cooled |
| **Interconnect** | NVLink:<br>> 900GB/s PCIe<br>> Gen5: 128GB/s | NVLink:<br>> 600GB/s PCIe<br>> Gen5: 128GB/s | NVLink:<br>> 600GB/s PCIe<br>> Gen5: 128GB/s |
| **Server options** | NVIDIA HGX™ H100 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs<br><br>NVIDIA DGX™ H100 with 8 GPUs | Partner and NVIDIA-Certified Systems with 1–8 GPUs | Partner and NVIDIA-Certified Systems with 2-4 pairs |
| **NVIDIA Enterprise** | Add-on | Included | Included |

[1] Preliminary specifications. May be subject to change. Specifications shown for 2x H100 NVL PCIe cards paired with NVLink Bridge.

[2] With sparsity.

# We're hitting the Memory Wall

FP32 6.691 TFLOPS (Titan X) -> FP16/32 165 TFLOPS (4090)

https://docs.google.com/spreadsheets/d/1Q-qqgOXtQqOxd3JjShiW_fNRpz6Kxd0YyC5__P5gAh4/edit#gid=0

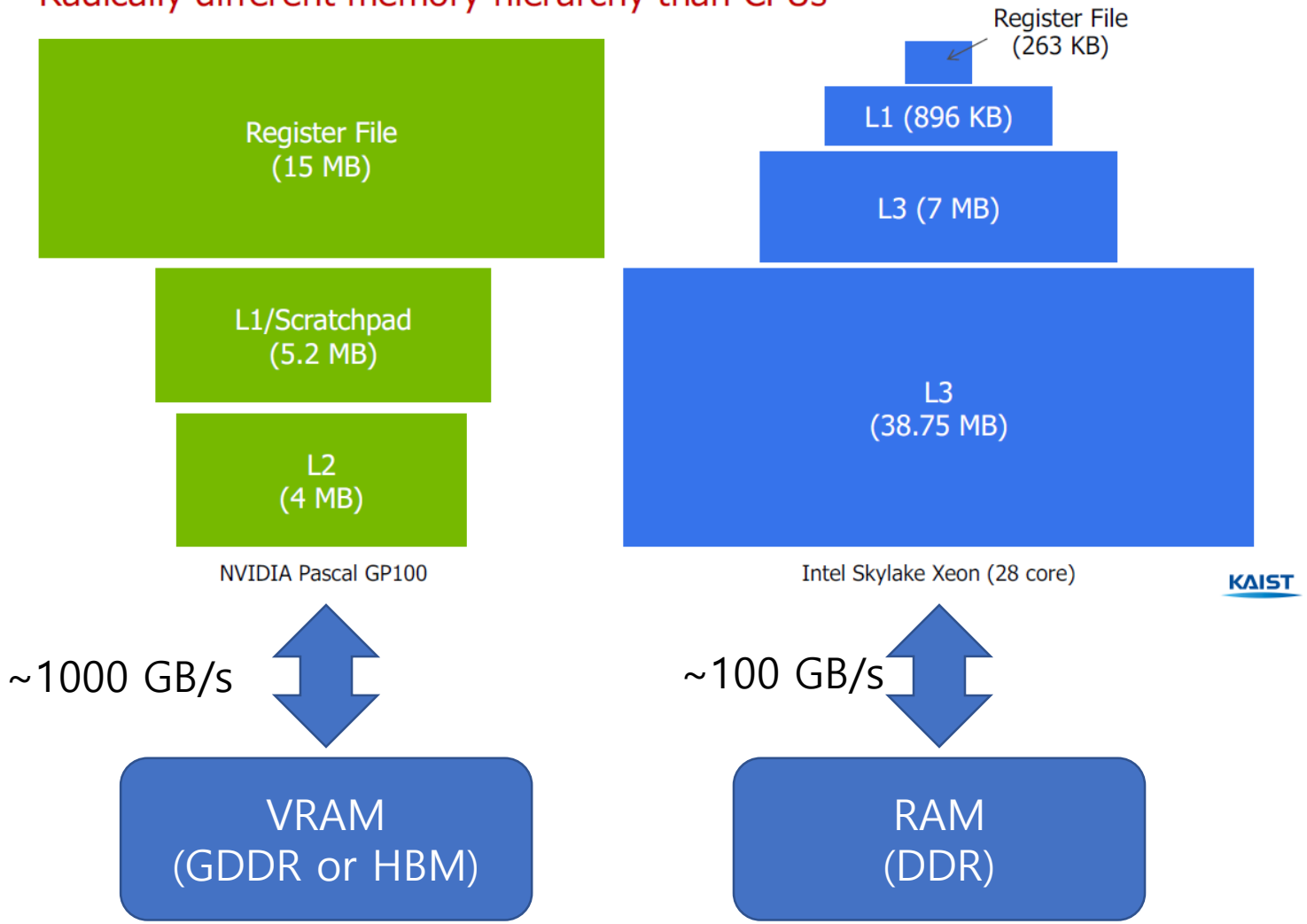
Arithmetic compute power: 25X
Memory capacity: 2X
Memory bandwidth: 3X

| | | | | | | Peak FP TFLOPS | | | Peak Tensor TFLOPS | | | | Processor Specs | | | Memory Specs | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Manufacturer | Architecture | Category | Product | Price | GPU | FP32 | FP16 | BF16 | FP16/FP32 | FP16/FP16 | BF16/FP32 | TF32 | SMs | CUDA Cores | Tensor Cores | Mem. Capacity | Mem. Type | Bandwidth | Max TDP |
| NVIDIA | Maxwell 2.0 | GeForce | TITAN X | $999 | GM200 | 6.69 | | | | | | | 24 | 3072 | | 12 GB | GDDR5 | 337 GBps | 250 W |
| | | | | | | Peak FP TFLOPS | | | Peak Tensor TFLOPS | | | | Processor Specs | | | Memory Specs | | | |
| Manufacturer | Architecture | Category | Product | Price | GPU | FP32 | FP16 | BF16 | FP16/FP32 | FP16/FP16 | BF16/FP32 | TF32 | SMs | CUDA Cores | Tensor Cores | Mem. Capacity | Mem. Type | Bandwidth | Max TDP |
| NVIDIA | Ada Lovelace | GeForce | RTX 4090 | $1,599 | AD102 | 82.58 | 82.58 | 82.58 | 165.20 | 330.30 | 165.20 | 82.60 | 128 | 16384 | 512 | 24 GB | GDDR6X | 1,018 GBps | 450 W |

Bottleneck is the time spent on moving data around.
(especially during inference)

# GPU memory hierarchy

Radically different memory hierarchy than CPUs




KAIST

# 이제, GPU 는 전략물자

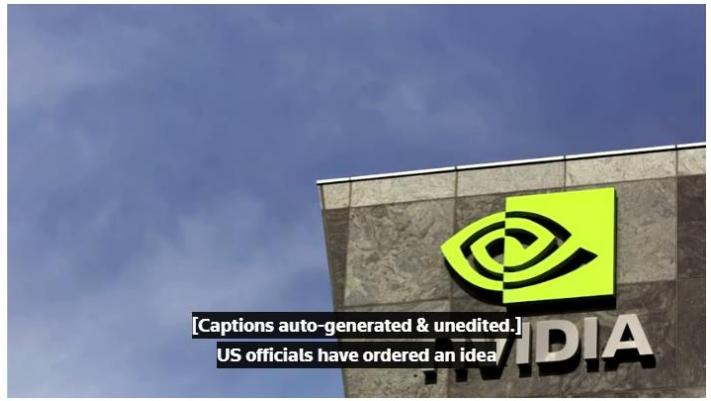
REUTERS® World Business Markets Sustainability Legal Breakingviews Technology In

Technology

## U.S. officials order Nvidia to halt sales of top AI chips to China

By **Stephen Nellis** and **Jane Lee**

September 1, 2022 6:13 PM GMT+9 · Updated a year ago




Sept 1 (Reuters) - Chip designer Nvidia Corp (NVDA.O) said on Wednesday that U.S. officials told it to stop exporting two top computing chips for artificial intelligence work to China, a move that could cripple Chinese firms' ability to carry out advanced work like image recognition and hamper Nvidia's business in the country.

The announcement signals a major escalation of the U.S. crackdown on China's technological capabilities as tensions bubble over the fate of Taiwan, where chips for Nvidia and almost every other major chip firm are manufactured.

REUTERS® World Business Markets Sustainability Legal Breakingviews Technology Inv

Disrupted

## Exclusive: Nvidia offers new advanced chip for China that meets U.S. export controls

By **Jane Lee**

November 8, 2022 10:32 AM GMT+9 · Updated a year ago


OAKLAND, Calif., Nov 7 (Reuters) - U.S. chip maker Nvidia Corp (NVDA.O) is offering a new advanced chip in China that meets recent export control rules aimed at keeping cutting-edge technology out of China's hands, the company confirmed on Monday.

Nvidia responded to Reuters' reporting that Chinese computer sellers are advertising products with the new chip.

The chip, called the A800, represents the first reported effort by a U.S. semiconductor company to create advanced processors for China that follow new U.S. trade rules. Nvidia has said the export limitations could cost it hundreds of millions of dollars in revenue.

**NVIDIA**

| Product Name | GPU Chip | Released | Bus | Memory | GPU clock | Memory clock | Shaders / TMUs / ROPs |
|---|---|---|---|---|---|---|---|
| Tesla Ampere (Axx) | | | | | | | |
| A30X | GA100 | Apr 12th, 2021 | PCIe 4.0 x8 | 24 GB, HBM2e, 3072 bit | 1035 MHz | 1593 MHz | 3584 / 224 / 96 |
| A30 PCIe | GA100 | Apr 12th, 2021 | PCIe 4.0 x16 | 24 GB, HBM2e, 3072 bit | 930 MHz | 1215 MHz | 3584 / 224 / 96 |
| A100 PCIe 40 GB | GA100 | Jun 22nd, 2020 | PCIe 4.0 x16 | 40 GB, HBM2e, 5120 bit | 765 MHz | 1215 MHz | 6912 / 432 / 160 |
| A100X | GA100 | Jun 28th, 2021 | PCIe 4.0 x8 | 80 GB, HBM2e, 5120 bit | 795 MHz | 1593 MHz | 6912 / 432 / 160 |
| A100 PCIe 80 GB | GA100 | Jun 28th, 2021 | PCIe 4.0 x16 | 80 GB, HBM2e, 5120 bit | 1065 MHz | 1512 MHz | 6912 / 432 / 160 |
| A100 SXM4 40 GB | GA100 | May 14th, 2020 | PCIe 4.0 x16 | 40 GB, HBM2e, 5120 bit | 1095 MHz | 1215 MHz | 6912 / 432 / 160 |
| A100 SXM4 80 GB | GA100 | Nov 16th, 2020 | PCIe 4.0 x16 | 80 GB, HBM2e, 5120 bit | 1275 MHz | 1593 MHz | 6912 / 432 / 160 |
| A800 PCIe 40 GB | GA100 | Nov 8th, 2022 | PCIe 4.0 x16 | 40 GB, HBM2e, 5120 bit | 765 MHz | 1215 MHz | 6912 / 432 / 160 |
| A800 PCIe 80 GB | GA100 | Nov 8th, 2022 | PCIe 4.0 x16 | 80 GB, HBM2e, 5120 bit | 1065 MHz | 1512 MHz | 6912 / 432 / 160 |
| A800 SXM4 80 GB | GA100 | Aug 11th, 2022 | PCIe 4.0 x16 | 80 GB, HBM2e, 5120 bit | 1155 MHz | 1593 MHz | 6912 / 432 / 160 |

**NVIDIA**

| Product Name | GPU Chip | Released | Bus | Memory | GPU clock | Memory clock | Shaders / TMUs / ROPs |
|---|---|---|---|---|---|---|---|
| Tesla Hopper (Hxx) | | | | | | | |
| H100 CNX | GH100 | Mar 22nd, 2022 | PCIe 5.0 x16 | 80 GB, HBM2e, 5120 bit | 690 MHz | 1593 MHz | 14592 / 456 / 24 |
| H100 PCIe 80 GB | GH100 | Mar 22nd, 2022 | PCIe 5.0 x16 | 80 GB, HBM2e, 5120 bit | 1095 MHz | 1593 MHz | 14592 / 456 / 24 |
| H100 PCIe 96 GB | GH100 | Mar 22nd, 2022 | PCIe 5.0 x16 | 96 GB, HBM3, 5120 bit | 1665 MHz | 1313 MHz | 16896 / 528 / 24 |
| H100 SXM5 64 GB | GH100 | Mar 22nd, 2022 | PCIe 5.0 x16 | 64 GB, HBM3, 3072 bit | 1665 MHz | 1313 MHz | 16896 / 528 / 24 |
| H100 SXM5 80 GB | GH100 | Mar 22nd, 2022 | PCIe 5.0 x16 | 80 GB, HBM3, 5120 bit | 1590 MHz | 1313 MHz | 16896 / 528 / 24 |
| H100 SXM5 96 GB | GH100 | Mar 22nd, 2022 | PCIe 5.0 x16 | 96 GB, HBM3, 5120 bit | 1665 MHz | 1313 MHz | 16896 / 528 / 24 |
| H800 SXM5 | GH100 | Mar 22nd, 2022 | PCIe 5.0 x16 | 80 GB, HBM3, 5120 bit | 1095 MHz | 1313 MHz | 16896 / 528 / 24 |

https://www.reuters.com/technology/nvidia-says-us-has-imposed-new-license-requirement-future-exports-china-2022-08-31/

https://www.reuters.com/technology/us-restricts-exports-some-nvidia-chips-middle-east-countries-filing-2023-08-30/

FEDERAL REGISTER
The Daily Journal of the United States Government

https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor

Rule

# Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification

A Rule by the Industry and Security Bureau on 10/13/2022

PUBLISHED DOCUMENT

Start Printed Page 62186

## AGENCY:

Bureau of Industry and Security, Department of Commerce.

## ACTION:

Interim final rule; request for comments.

## SUMMARY:

In this rule, the Bureau of Industry and Security (BIS) is amending the Export Administration Regulations (EAR) to implement necessary controls on advanced computing integrated circuits (ICs), computer commodities that contain such ICs, and certain semiconductor manufacturing items. In addition, BIS is expanding controls on transactions involving items for supercomputer and semiconductor manufacturing end uses, for example, this rule expands the scope of foreign-produced items subject to license requirements for twenty-eight existing entities on the Entity List that are located in China. BIS is also informing the public that specific activities of "U.S. persons" that 'support' the "development" or "production" of certain ICs in the PRC require a license. Lastly, to minimize short term impact on the semiconductor supply chain from this rule, BIS is establishing a Temporary General License to permit specific, limited manufacturing activities in China related to items destined for use outside China and is identifying a model certificate that may be used in compliance programs to assist, along with other measures, in conducting due diligence. `

DOCUMENT DETAILS

**Printed version:**
PDF

**Publication Date:**
10/13/2022

**Agencies:**
Bureau of Industry and Security

**Dates:**
a. Effective on October 7, 2022, are the following instructions: 7 (Sec. 740.2), 9 (Sec. 740.10), 11 (Sec. 742.6), 17 (Sec. 744.23), and 25 (supplement no. 1 to part 774).

**Effective Date:**
10/07/2022

**Document Type:**
Rule

**Document Citation:**
87 FR 62186

**Page:**
62186-62215 (30 pages)

**CFR:**
15 CFR 734
15 CFR 736
15 CFR 740
15 CFR 742
15 CFR 744
15 CFR 762

## A. 추가 수출 통제: 특정 고급 컴퓨팅 집적 회로(IC), 슈퍼컴퓨터 최종 용도; 엔터티 목록 수정

이 규칙에 따라 BIS는 특정 고급 컴퓨팅 반도체 칩과 해당 칩이 포함된 컴퓨터 상품에 대해 새로운 수출 통제를 부과합니다. 또한 이 규칙은 중국에 위치하거나 중국으로 향하는 "슈퍼컴퓨터"용 특정 품목에 대한 최종 사용 통제를 구현합니다.

고급 컴퓨팅 항목과 '슈퍼컴퓨터'를 사용하여 인공지능(AI) 애플리케이션을 비롯한 데이터 처리 및 분석 기능을 향상할 수 있습니다. 중국은 엑사급 슈퍼컴퓨팅 기능을 빠르게 개발하고 있으며 2030년까지 AI 분야의 세계적 리더가 되겠다는 의지를 발표했습니다. 이러한 고급 시스템은 다양한 용도로 사용되는 정교한 데이터 처리 및 분석이 가능하며 고급 IC를 통해 지원됩니다. 이들 시스템 중국은 군사 의사결정, 계획, 군수뿐 아니라 인지 전자전, 레이더, 신호 정보 등에 사용되는 자율 군사 시스템의 속도와 정확성을 향상시키기 위한 군사 현대화 노력을 위해 사용되고 있습니다. , 재밍. 또한 이러한 첨단 컴퓨팅 항목과 "슈퍼컴퓨터"는 중국에서 핵무기, 극초음속 및 기타 첨단 미사일 시스템과 같은 WMD를 포함한 무기 설계 및 테스트 계산을 개선하고 전장 효과를 분석하는 데 사용되고 있습니다. 또한, 중국은 막대한 양의 데이터를 효율적으로 처리하여 구현된 고급 AI 감시 도구를 기본 인권을 고려하지 않고 시민을 모니터링, 추적, 감시하는 등의 목적으로 사용하고 있습니다. 이 규칙을 통해 BIS는 핵무기 개발, 고급 정보 수집 및 분석 촉진, 감시 등 군사 현대화를 위한 중국의 고급 컴퓨팅에 대한 접근을 제한함으로써 미국 국가 안보와 외교 정책 이익을 보호하려고 합니다. BIS는 미국 국가 안보 및 외교 정책 이익에 반하는 용도로 고급 컴퓨팅 칩을 확보하거나 AI 및 "슈퍼컴퓨터" 기능을 추가로 개발하는 중국의 능력을 제한하기 위해 EAR 및 미국인 활동에 해당하는 항목에 대한 통제를 부과할 계획입니다.

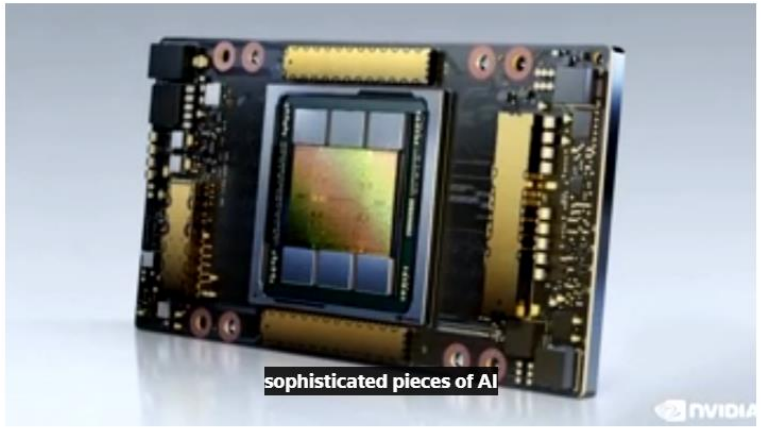
REUTERS® World Business Markets Sustainability Legal Breakingviews Technology Investi

Technology

# US curbs AI chip exports from Nvidia and AMD to some Middle East countries

By **Stephen Nellis** and **Max A. Cherney**

September 1, 2023 2:26 AM GMT+9 · Updated 2 months ago




Aug 30 (Reuters) - The U.S. expanded the restriction of exports of sophisticated Nvidia (NVDA.O) and Advanced Micro Devices (AMD.O) artificial-intelligence chips beyond China to other regions including some countries in the Middle East.

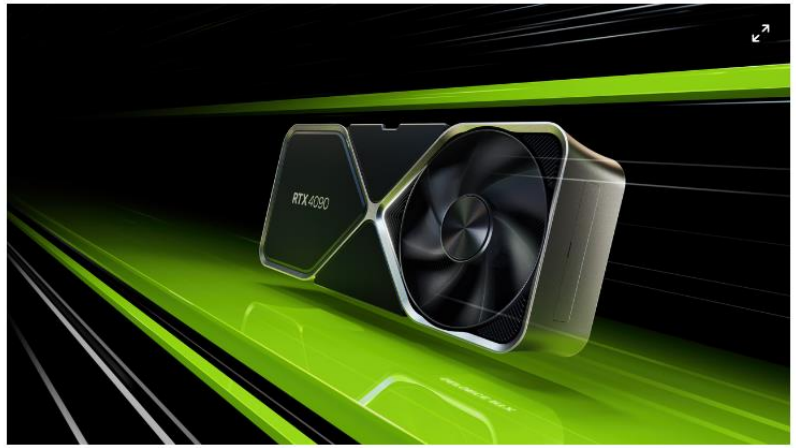
Nvidia said in a regulatory filing this week that the curbs, which affect its A100 and H100 chips designed to speed up machine-learning tasks, would not have an "immediate material impact" on its results.

Rival AMD also received an informed letter with similar restrictions, a person familiar with the matter told Reuters, adding that the move has no material impact on its revenue.

U.S. officials usually impose export controls for national security reasons. A similar move announced last year signaled an escalation of the U.S. crackdown on China's technological capabilities, but it was not immediately clear what risks were posed by exports to the Middle East.

# 게임용 GPU 'RTX', 수출금지 지정으로 중국내 가격 폭등

박찬 기자 입력 2023.10.23 18:09 댓글 0 좋아요 0




RTX 4090 GPU (사진=엔비디아)

미국의 AI 칩 수출 통제 강화 이후 중국에서 엔비디아의 소비자용 GPU 'RTX 4090'의 가격이 3배 가까이 치솟았다. 이전에 중국 시장에 풀렸던 재고를 대상으로 사재기가 벌어지고 있다.

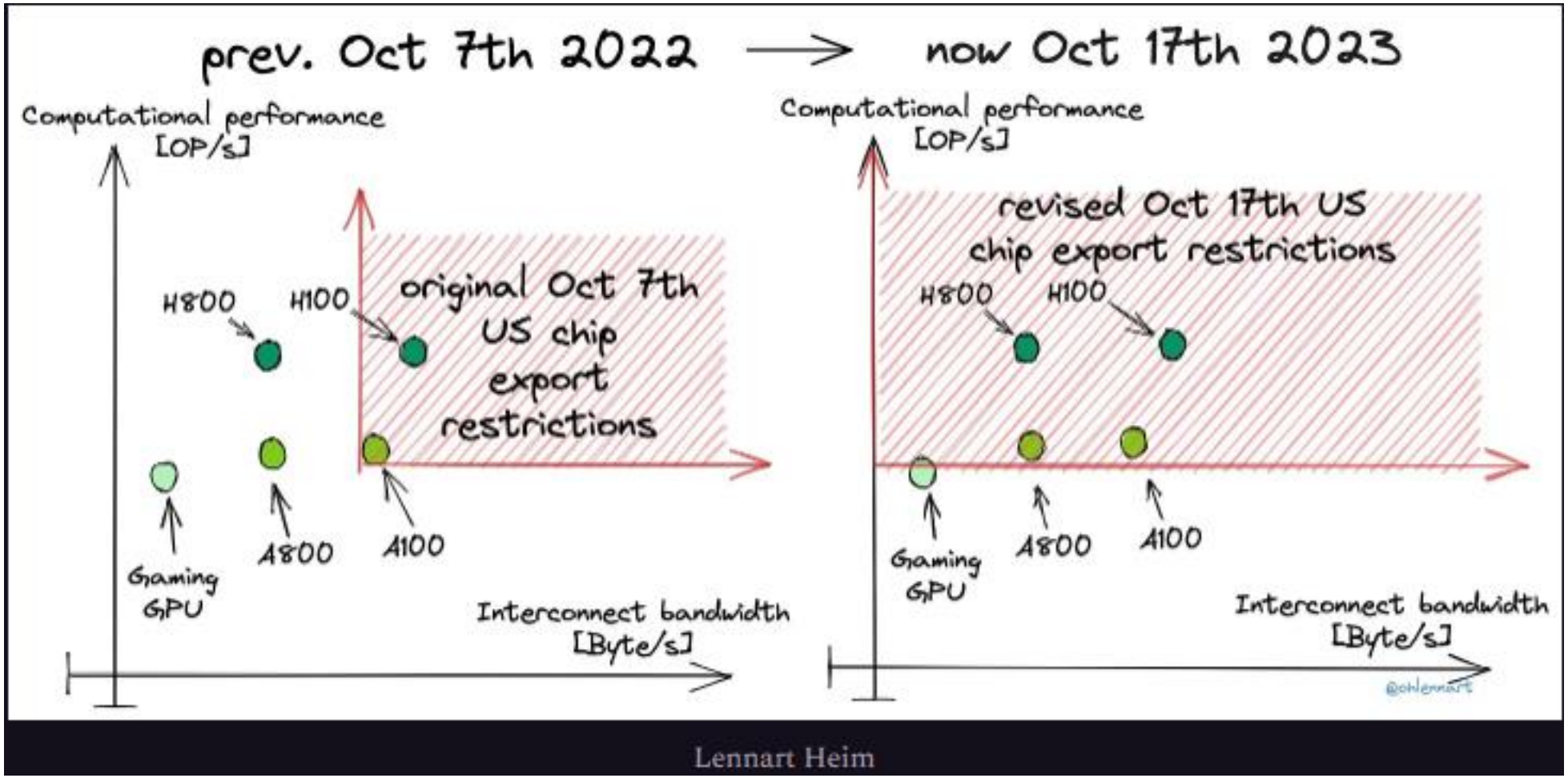
블룸버그는 22일(현지시간) 중국 알리바바가 운영하는 오픈마켓 타오바오에서 RTX 4090가 3970달러(약 537만원)에 거래되고 있다고 보도했다.

이는 엔비디아의 소비자 판매가 1599달러(약 216만원)의 2.5배 수준이다. RTX 4090은 현존하는 PC용 그래픽카드 중 최고 사양으로 평가받는다. 지난해 9월 출시된 후 중국의 게이머와 그래픽 디자이너 사이에서 베스트셀러 제품으로 팔리고 있었다.

그러나 지난 17일 미국 상무부가 저사양 AI 반도체까지 중국에 수출하지 못하도록 대중국 반도체 수출통제 조치를 강화하자 RTX 4090도 타격을 받았다.

엔비디아는 RTX 4090의 성능이 미국 당국의 기준치를 초과한 것으로 밝혀진 이후 미국 정부 허가 없이 중국으로 수출하는 것이 금지됐다고 밝혔다.

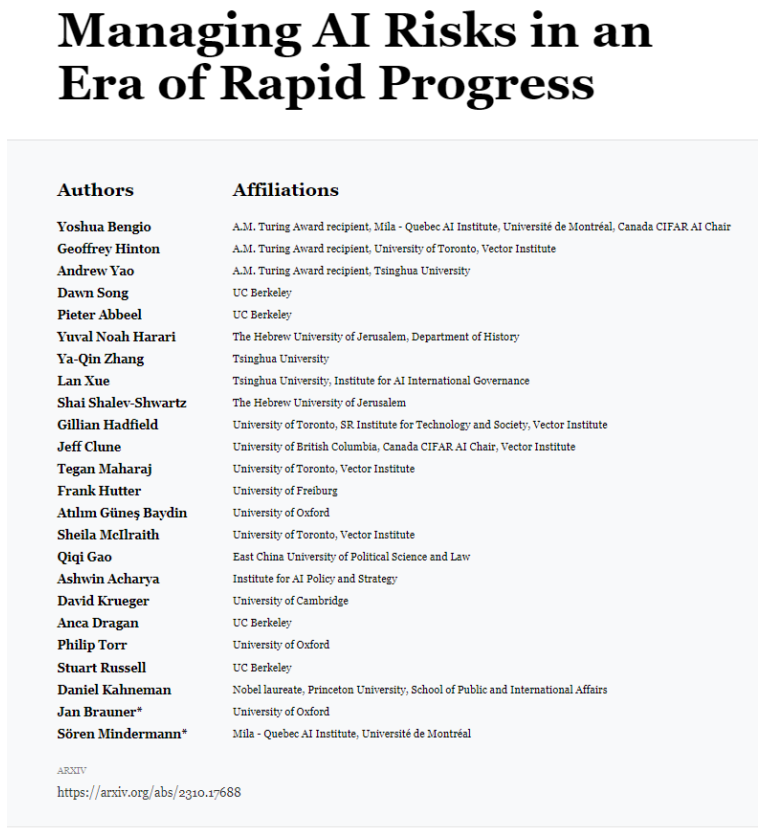
https://www.reuters.com/technology/us-restricts-exports-some-nvidia-chips-middle-east-countries-filing-2023-08-30/




https://www.semianalysis.com/p/wafer-wars-deciphering-latest-restrictions

# Beyond beating AI benchmarks


https://managing-ai-risks.com/



# Managing AI Risks in an Era of Rapid Progress

| Authors | Affiliations |
|---|---|
| Yoshua Bengio | A.M. Turing Award recipient, Mila - Quebec AI Institute, Université de Montréal, Canada CIFAR AI Chair |
| Geoffrey Hinton | A.M. Turing Award recipient, University of Toronto, Vector Institute |
| Andrew Yao | A.M. Turing Award recipient, Tsinghua University |
| Dawn Song | UC Berkeley |
| Pieter Abbeel | UC Berkeley |
| Yuval Noah Harari | The Hebrew University of Jerusalem, Department of History |
| Ya-Qin Zhang | Tsinghua University |
| Lan Xue | Tsinghua University, Institute for AI International Governance |
| Shai Shalev-Shwartz | The Hebrew University of Jerusalem |
| Gillian Hadfield | University of Toronto, SR Institute for Technology and Society, Vector Institute |
| Jeff Clune | University of British Columbia, Canada CIFAR AI Chair, Vector Institute |
| Tegan Maharaj | University of Toronto, Vector Institute |
| Frank Hutter | University of Freiburg |
| Atılım Güneş Baydin | University of Oxford |
| Sheila McIlraith | University of Toronto, Vector Institute |
| Qiqi Gao | East China University of Political Science and Law |
| Ashwin Acharya | Institute for AI Policy and Strategy |
| David Krueger | University of Cambridge |
| Anca Dragan | UC Berkeley |
| Philip Torr | University of Oxford |
| Stuart Russell | UC Berkeley |
| Daniel Kahneman | Nobel laureate, Princeton University, School of Public and International Affairs |
| Jan Brauner* | University of Oxford |
| Sören Mindermann* | Mila - Quebec AI Institute, Université de Montréal |

ARXIV
https://arxiv.org/abs/2310.17688

arXiv paper PDF | Policy supplement

**Abstract**

In this short consensus paper, we outline risks from upcoming, advanced AI systems. We examine large-scale social harms and malicious uses, as well as an irreversible loss of human control over autonomous AI systems. In light of rapid and continuing AI progress, we propose urgent priorities for AI R&D and governance.

### Reorienting technical R&D

We need research breakthroughs to solve some of today's technical challenges in creating AI with safe and ethical objectives. Some of these challenges are unlikely to be solved by simply making AI systems more capable [22, 31, 32, 33, 34, 35]. These include:

- Oversight and honesty: More capable AI systems are better able to exploit weaknesses in oversight and testing [32, 36, 37]—for example, by producing false but compelling output [35, 38].
- Robustness: AI systems behave unpredictably in new situations (under distribution shift or adversarial inputs) [39, 40, 34].
- Interpretability: AI decision-making is opaque. So far, we can only test large models via trial and error. We need to learn to understand their inner workings [41].
- Risk evaluations: Frontier AI systems develop unforeseen capabilities only discovered during training or even well after deployment [42]. Better evaluation is needed to detect hazardous capabilities earlier [43, 44].
- Addressing emerging challenges: More capable future AI systems may exhibit failure modes we have so far seen only in theoretical models. AI systems might, for example, learn to feign obedience or exploit weaknesses in our safety objectives and shutdown mechanisms to advance a particular goal [24, 45].

Given the stakes, we call on major tech companies and public funders to allocate at least one-third of their AI R&D budget to ensuring safety and ethical use, comparable to their funding for AI capabilities. Addressing these problems [34], with an eye toward powerful future systems, must become central to our field.

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

https://intel.ks.group/p/ai-executive-order-key-takeaways

https://twitter.com/iScienceLuvr/status/1718986424769552461

OCTOBER 30, 2023

# Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM ▸ PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

My Administration places the highest urgency on governing the development and use of AI safely and responsibly, and is therefore advancing a coordinated, Federal Government-wide approach to doing so. The rapid speed at which AI capabilities are advancing compels the United States to lead in this moment for the sake of our security, economy, and society.


In particular, the EO is focused on models that:

1. Substantially lower the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;
2. Enable powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber-attacks; or
3. Permit the evasion of human control or oversight through means of deception or obfuscation.

Reporting requirements apply to large computing clusters and models trained using a quantity of computing power just above the current state of the art and at the level of ~25-50K clusters of H100 GPUs. These parameters can change at the discretion of the Commerce Secretary, but the specified size and interconnection measures are intended to bring only the most advanced "frontier" models into the scope of future reporting and risk assessment.

Sec. 5. Promoting Innovation and Competition.

5.1. Attracting AI Talent to the United States. (a) Within 90 days of the date of this order, to attract and retain talent in AI and other critical and emerging technologies in the United States economy, the Secretary of State and the Secretary of Homeland Security shall take appropriate steps to:

(i) streamline processing times of visa petitions and applications, including by ensuring timely availability of visa appointments, for noncitizens who seek to travel to the United States to work on, study, or conduct research in AI or other critical and emerging technologies; and

(ii) facilitate continued availability of visa appointments in sufficient volume for applicants with expertise in AI or other critical and emerging technologies.