



OLF 2023

Empire Nvidia and Its Challengers: The alternatives for ML today

Liang Yan

09/09/2023



Self-Introduction

Liang Yan

Sr. Software Engineer II, DigitalOcean, Louisville KY

- Over ten years industry experience on Virtualization, mostly on GPU, Network and Live Migration.
- Current focus on Distributed Machine Learning Infra for Large Model
- Opensource and Arm64 Board Enthusiast

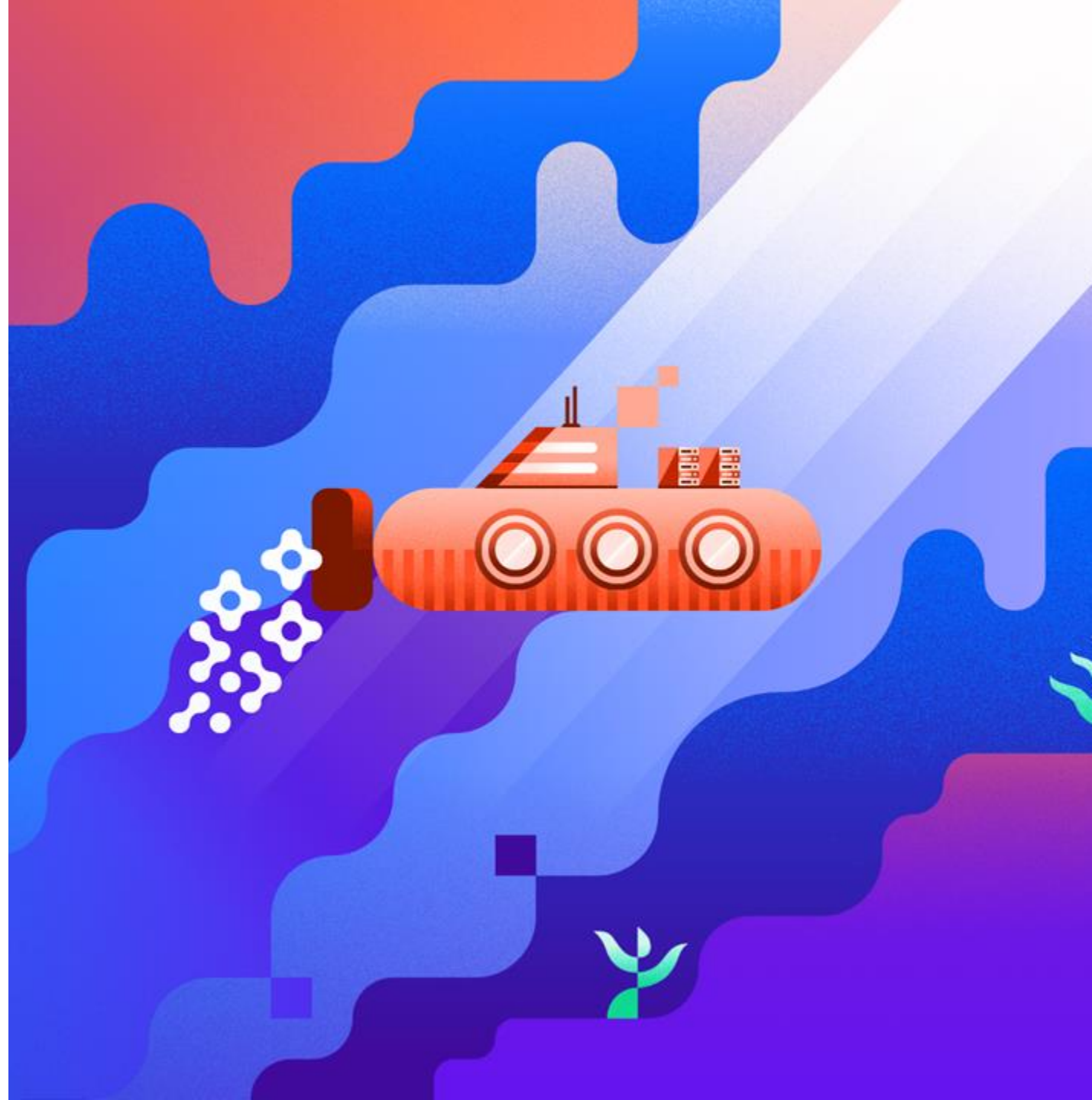


<https://www.linkedin.com/in/lyantech>



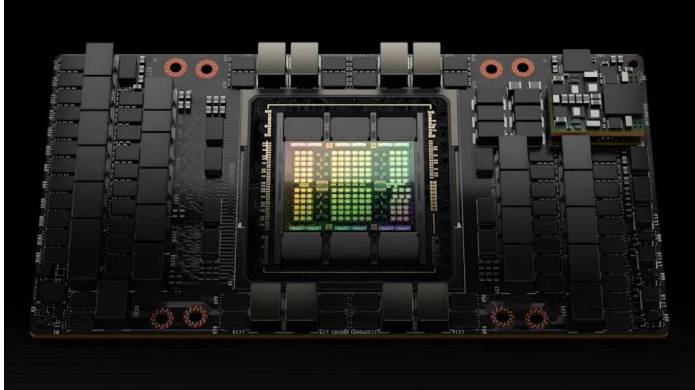
Agenda

- Where the story begins
- Nvidia GPGPU history
- The Challengers
- Software Ecosystem
- To continue





The story begins



- CoreWeave, [a cloud computing startup](#) backed by Nvidia and Magnetar Capital, has secured a \$2.3 billion debt facility using Nvidia's sought-after H100 GPUs as collateral.
- Nvidia is raking in nearly 1,000% (about 823%) in profit percentage for each H100 GPU accelerator it sells, according to estimates made in a [recent social media post](#) from Barron's senior writer [Tae Kim](#).



NVIDIA GPU

Currie (2004)	
Tesla (2006)	
Fermi(2010)	40nm 256 CUDA core
Kepler(2012)	28nm SM->SMX
Maxwell(2014)	SMX->SMM
Pascal(2016)	TSMC 16nm, NVLink 1.0, HBM 1.0
Volta (2017)	Tensor Core 1.0, NVLink 2.0, HBM 2.0
Turing (2018)	RT Core 1.0, Tensor Core 2.0, HBM2.0
Ampere (2020)	TSMC 7nm, Tensor 3.0, RT 2.0, NVLink 3.0
Hopper (2022)	TSMC 4nm, Tensor 4.0, NVLink 4.0 HBM 3.0, Transfer engine
Ada Lovelace (2022)	TSMC 4nm, RT core 3.0
Blackwell(2024 ?)	TSMC 3nm, GDDR7 PCIe 5



Other GPUs

AMD	Intel
Instinct MI 300X/A	Falcon Shores 2: 2026
MI 250X 250/210	Falcon Shores: 2025
MI100	ARC Xe HPG 2022
MI 60/50	SG1 Xe LP 2020
MI25	DG1 Xe MP
MI6	VCA2 EOL 2017



Hardware Accelerator Landscape

	GPU	FPGA	ASIC
Vendors	NVIDIA®, AMD®, INTEL®	Xilinx®, INTEL® (Altera)	Google TPU, AI Chips
Development Frameworks	OpenCL, CUDA	OpenCL	OpenCL, TensorFlow
Machine Learning Lifecycle	Training	Inference	Inference

FPGA: Field-Programmable Gate Array
ASIC: Application-Specific Integrated Circuit
TPU: Tensor Processing Unit
CUDA: Compute Unified Device Architecture

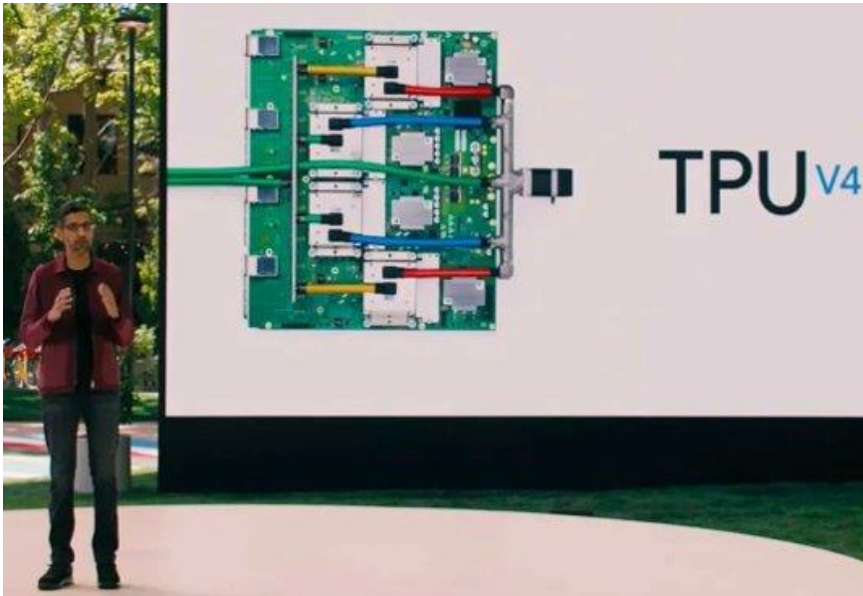


ASIC: Google TPU

Tensor processing unit : Google in 15 months

Nvidia

Volta: Tensorcore



TENSOR CORE
Mixed Precision Matrix Math
4x4 matrices


$$D = \begin{matrix} \text{FP16 or FP32} & \begin{matrix} \text{FP16} & \text{FP16} & \text{FP16} & \text{FP16} \end{matrix} & + & \begin{matrix} \text{FP16 or FP32} \end{matrix} \\ \begin{matrix} \begin{matrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{matrix} & \begin{matrix} \begin{matrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{matrix} & \begin{matrix} \begin{matrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{matrix} \end{matrix} \end{matrix}$$

$$D = AB + C$$





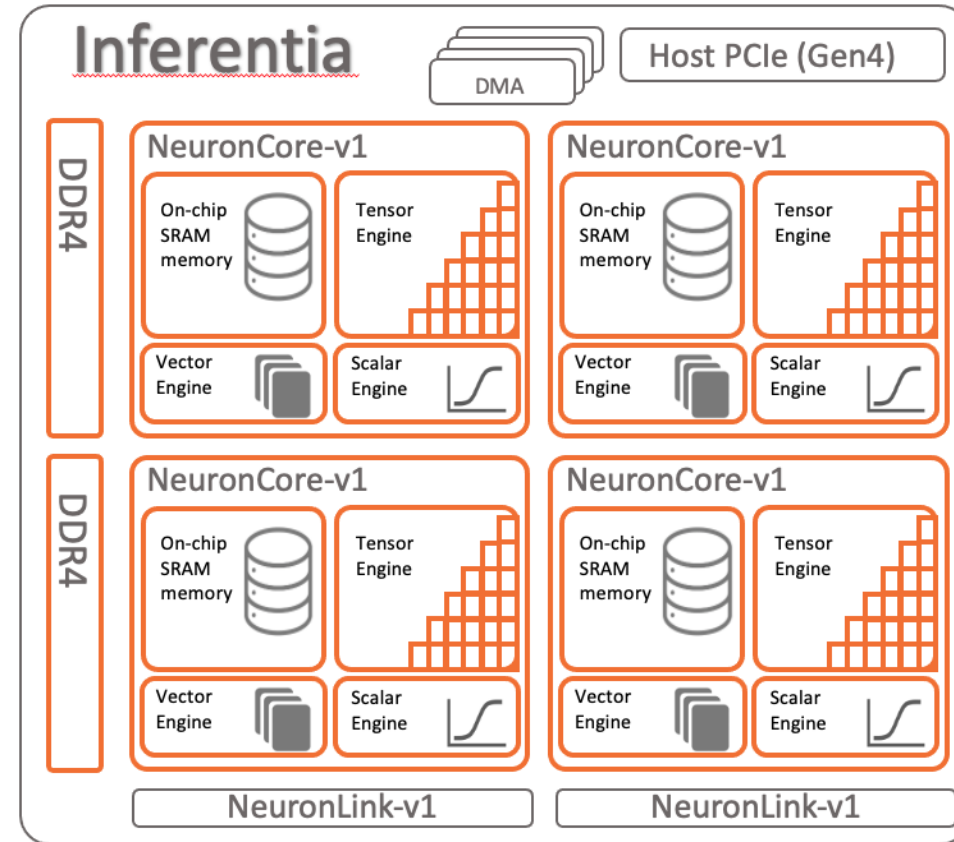
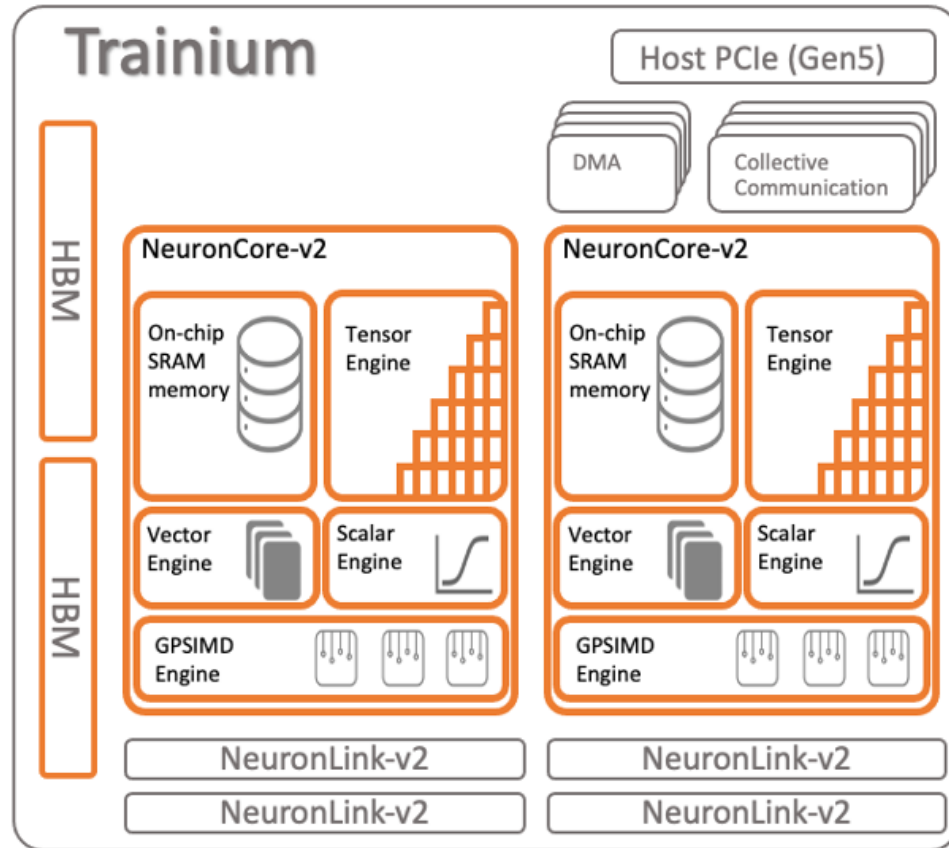
ASIC: Microsoft Athena

- Since 2019
- Release in 2024
- TSMC 5nm
- 500 Engineers





ASIC: AWS Trainium & Inferentia



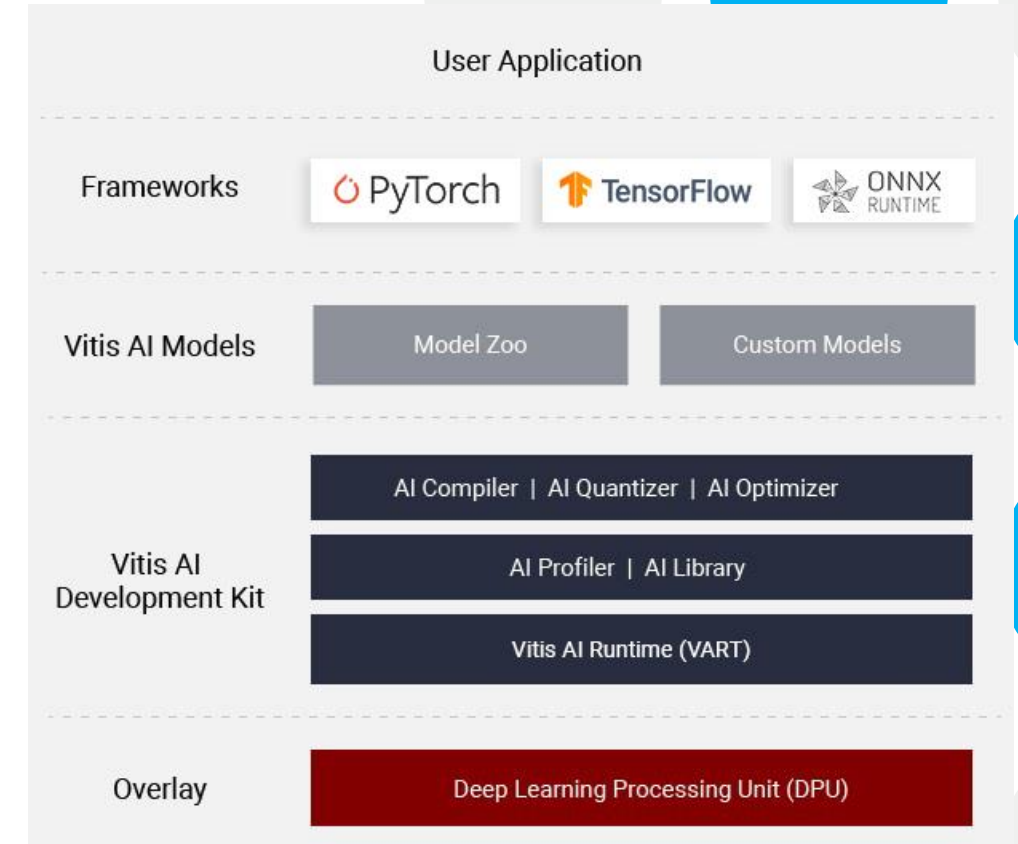
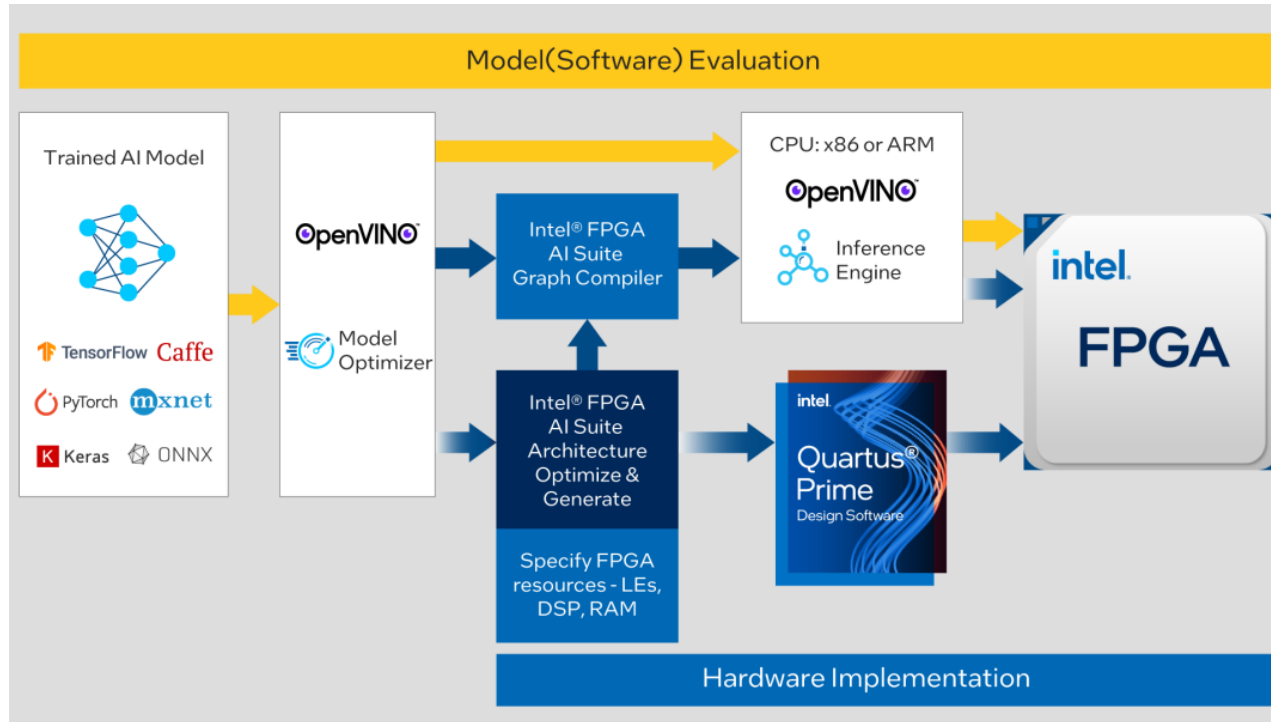


AI Chip Startup

Habana Labs	2016	Israel	Intel acquire 2019
Graphcore	2016	UK	IPU
Cerebras	2015	US	WSE-2
Mythic	2012	US	M1108 AMP
LightMatter	2017	US	Enviser
ReuReality	2018	Israel	NR1
Esperanto	2014	US	ET-SoC-1
Cambrian	2016	CN	MLU



Inte vs AMD FPGA





ML Stack: Compilation

	Frontend compiler		Backend compiler
JAX		CUDA	NVCC
MLIR		ROCm	Radeon Open Compute platform
Glow	Pytorch	OneAPI	CPU + GPU + FPGA + AI Accelerator
Inducer	Pytorch	OpenCL	Academic
Triton(*)		Apache TVM	



ML Stack: Framework

Pytorch2	ROCm, Graphecore
Tensorflow2	
DL4J	
MXNET	
Keras	
CNTK	
DeepSpeed	

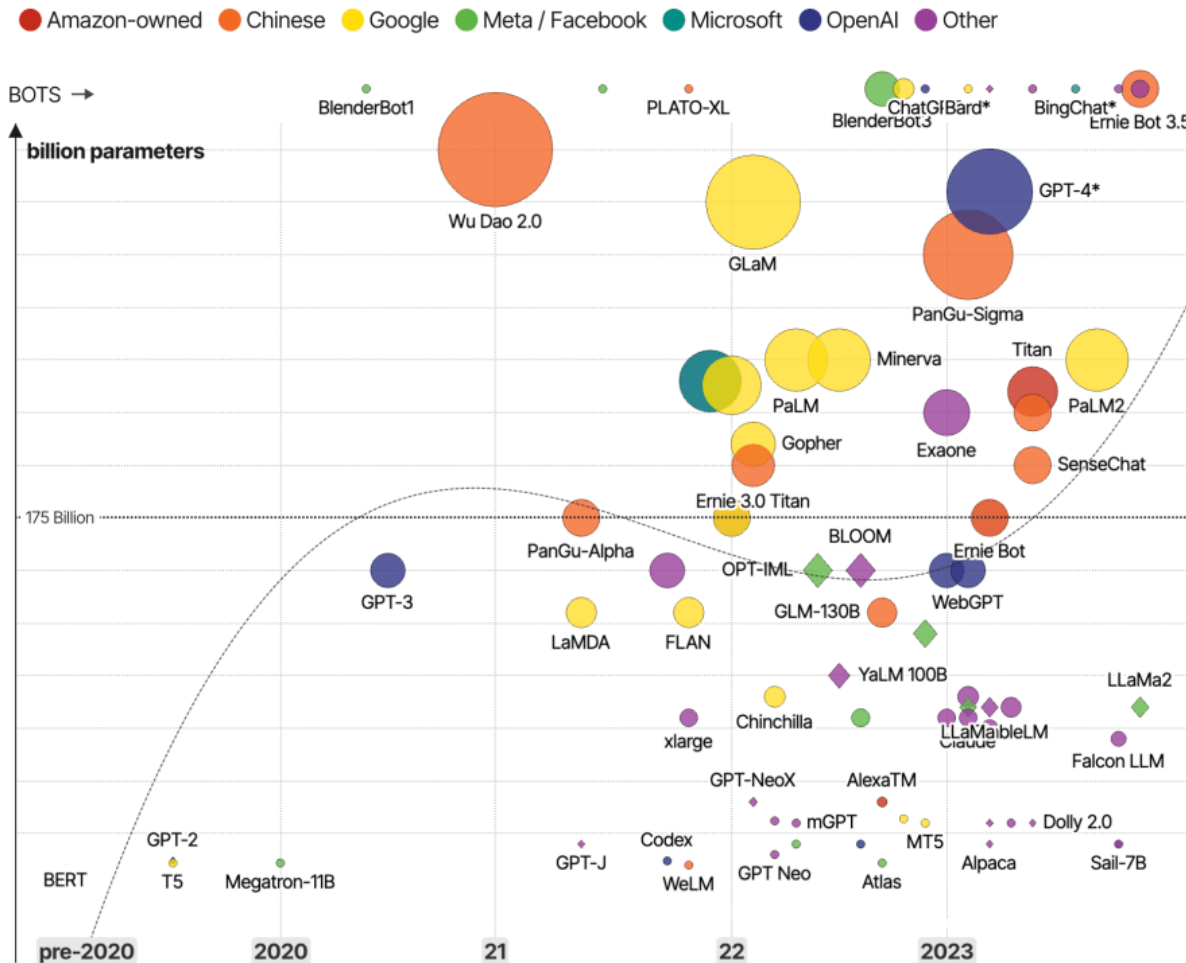


ML Stack: Large Model

Transformer	Google
Falcon	180B, TII
Gemini	Google
Claude 2	Anthropic
ChatGPT	OpenAI
LLaMA2	Meta
Bard	Google
PaLM 2	Google
Stable Diffusion	
Alexnet	Hinton, the beginning of the Era



ML Stack: Large Model



David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 27th Jul 23

source: news reports, [LifeArchitect.ai](https://lifeaiarchitect.ai)
* = parameters undisclosed // see [the data](#)



To continue

Ecosystem!

Ecosystem!

Ecosystem!





Reference

1. <https://www.theverge.com/2023/8/8/23824661/coreweave-nvidia-debt-gpu-ai-chips-collateral>
2. <https://www.tomshardware.com/news/nvidia-makes-1000-profit-on-h100-gpus-report>
3. <https://www.xilinx.com/products/design-tools/vitis/vitis-platform.html>
4. <https://www.intel.com/content/www/us/en/software/programmable/fpga-ai-suite/overview.html/>
5. <https://informationisbeautiful.net/>
6. <https://awsdocs-neuron.readthedocs-hosted.com/en/latest/index.html>
7. https://mp.weixin.qq.com/s/piPj1b2QdKUuf1UC_AiwVw
8. <https://metaverse-family.com/hangyezixun/5270.html>
9. <https://www.theinformation.com/articles/google-and-microsofts-other-ai-race-server-chips/>
10. <https://wallpaperaccess.com/to-be-continued/>

Q & A

Thanks!

