Module 4

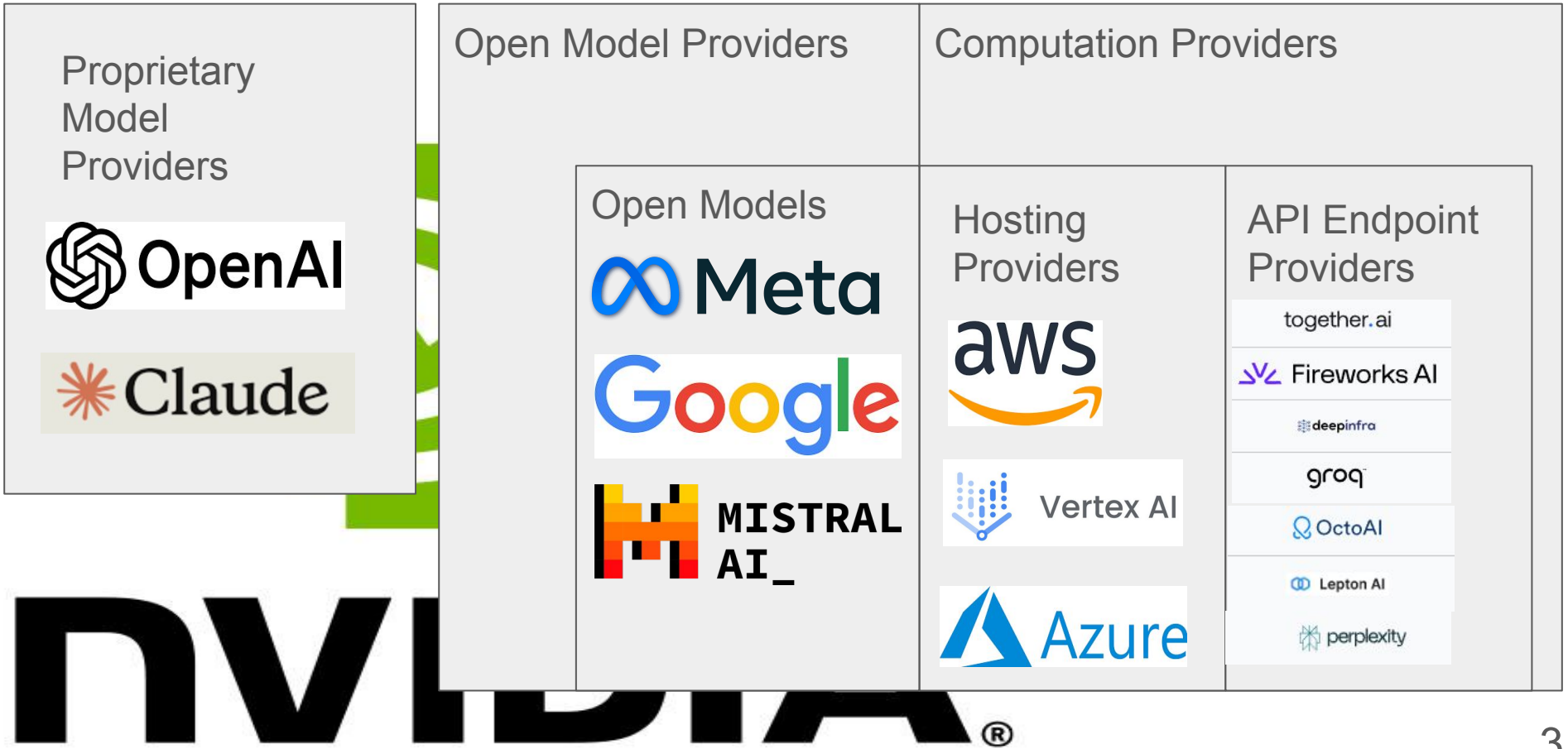
LLM Model Deployment

Wei Dong
wdong@aaalgo.com

Syllabus

- Overview
- Proprietary and Hosted Models
- Local Model Deployment
  - Floating Point Format and Quantization
  - Page Attention and vLLM
  - Flash Attention
- Nvidia and its GPUs

The LLM Deployment Landscape



Proprietary Model Pricing

	1 million token pricing		
OpenAI	gpt-4o-mini	gpt-4o	
	$0.15/0.60	$5/$15	
Google	Gemini 1.5 Flash	Gemini 1.5 Pro	
	$0.15/$0.60	$3.5-7.0/$10.50-$21	
Anthropic	Claude 3 Haiku	Claude 3 Sonnet	Claude 3 Opus*
	$0.25/$1.25	$3/$15	$15/$75
Mistral.ai	Mistral Nemo	Mistral Large	
	$0.3/$3	$3/$9	

Detour: The Claude Family of Models

Claude 3 released on Mar 4, 2024

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	**86.8%** 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	**50.4%** 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	**84.9%** 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	**83.1** 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	**86.8%** 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	**96.4%** 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	**95.4%** 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Detour: Claude Family of Models

Claude 3.5 released on

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning *GPQA, Diamond*	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge *MMLU*	88.7%** 5-shot 88.3% 0-shot CoT	86.8% 5-shot 85.7% 0-shot CoT	— 88.7% 0-shot CoT	85.9% 5-shot —	86.1% 5-shot —
Code *HumanEval*	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math *MGSM*	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	—
Reasoning over text *DROP, F1 score*	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations *BIG-Bench-Hard*	93.1% 3-shot CoT	86.8% 3-shot CoT	—	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving *MATH*	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
Grade school math *GSM8K*	96.4% 0-shot CoT	95.0% 0-shot CoT	—	90.8% 11-shot	94.1% 8-shot CoT

* Claude 3.5 Sonnet scores 67.2% on 5-shot CoT GPQA with maj@32
** Claude 3.5 Sonnet scores 90.4% on MMLU with 5-shot CoT prompting

Detoure: Llama 3.1 vs Claude 3.5

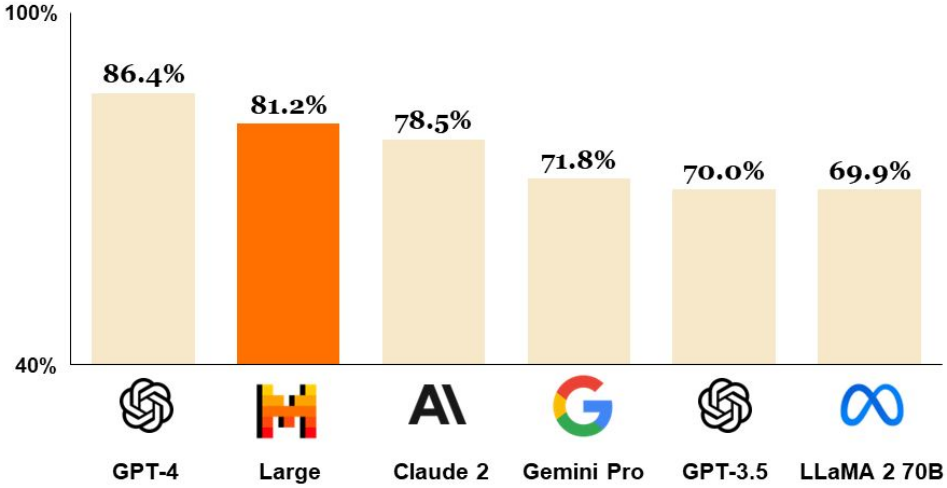
Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	**72.3**	61.1	**83.6**	76.9	70.7	87.3	82.6	85.1	89.1	**89.9**
	MMLU (0-shot, CoT)	**73.0**	72.3△	60.5	**86.0**	79.9	69.8	88.6	78.7◁	85.4	**88.7**	88.3
	MMLU-Pro (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2	73.3	62.7	64.8	74.0	**77.0**
	IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9	**88.6**	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0	89.0	73.2	86.6	90.2	**92.0**
	MBPP EvalPlus (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0	88.6	72.8	83.6	87.8	**90.5**
Math	GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6	**96.8**	92.3◇	94.2	96.1	96.4◇
	MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1	73.8	41.1	64.5	**76.6**	71.1
Reasoning	ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7	**96.9**	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	**46.7**	33.3	30.8	51.1	–	41.4	53.6	**59.4**
Tool use	BFCL	**76.1**	–	60.4	84.8	–	**85.9**	88.5	86.5	88.3	80.5	**90.2**
	Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2	**58.7**	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	**95.2**	–	**95.2**	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	**83.4**	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	**100.0**	**100.0**	90.8
Multilingual	MGSM (0-shot, CoT)	**68.9**	53.2	29.9	**86.9**	71.1	51.4	**91.6**	–	85.9	90.5	**91.6**

Detour: Anthropic

- US-based public-benefit company
- Founded in 2021 by former members of OpenAI
- $4 billion investment from Amazon and $2 billion from Google
- Timeline
  - 2019 – OpenAI transition from non-profit to for-profit
  - 2020 – OpenAI announced GPT-3
  - 2020-2021 – 11 employees left OpenAI to found Anthropic

Detour: Mistral AI

- French company
- Founded in April 2023 by former Meta and Google employees
- API only models
- Open-weights Models



Feb 26, 2024

LLM as Service [source]

PROVIDER	COST IN	COST OUT	RATE LIMIT	THROUGHPUT P50	THROUGHPUT P90	TTFT P50	TTFT P90	
DEEPINFRA mixtral-8x7b	$0.27/M #1	$0.27/M #1	Unlimited	58tks/s	76tks/s	0.42s	0.98s	
REPLICATE mixtral-8x7b	$0.3/M #2	$1/M	10 RPS	39tks/s	52tks/s	0.38s	0.48s	
FIREWORKS mixtral-8x7b	$0.5/M #5	$0.5/M #2	600 RPM	83tks/s #2	98tks/s #5	0.24s #2	0.33s #2	
LEPTON mixtral-8x7b	$0.5/M #3	$0.5/M #2	10 RPM	72tks/s #3	122tks/s #2	0.26s #3	0.57s	
PERPLEXITY mixtral-8x7b	$0.6/M	$0.6/M #3	24 RPM	129tks/s #1	135tks/s #1	0.13s #1	0.16s #1	
TOGETHER mixtral-8x7b	$0.6/M	$0.6/M #3	100 RPS	71tks/s	87tks/s	0.51s	0.69s	
REPLICATE llama2-70b-chat	$0.65/M	$2.75/M	10 RPS	39tks/s	50tks/s	0.4s	0.53s	
DEEPINFRA llama2-70b-chat	$0.7/M	$0.9/M	Unlimited	18tks/s	23tks/s	0.43s	1.12s	
ANTHROPIC claude-instant-1.2	$0.8/M	$2.4/M	Customized	57tks/s	63tks/s	0.31s	0.4s #3	
TOGETHER llama2-70b-chat	$0.9/M	$0.9/M	100 RPS	43tks/s	46tks/s	0.49s	0.94s	
OPENAI gpt-3.5-turbo	$1/M	$2/M	10K RPM	69tks/s	81tks/s	0.4s	0.54s	
ANTHROPIC claude-2.1	$8/M	$24/M	Customized	24tks/s	27tks/s	0.42s	0.51s	
OPENAI gpt-4-turbo	$10/M	$30/M	10K RPM	25tks/s	34tks/s	0.57s	0.85s	
OPENAI gpt-4	$30/M	$60/M	10K RPM	21tks/s	30tks/s	0.56s	0.78s	

How Much is 1 Million Tokens

API usually costs $0.1-10/million tokens

1 English Word ~ 1.5 tokens

Reference	Average Length	1M Token Equivalence
English Novels	70K - 100K words	6-8 full-length novels
Research Paper	5000-10000 words	~ 100 research papers
Wikipedia Pages	~600 words	~1000 articles
Chat Messages	20-30 words	2K - 3K messages

Open-Weights Model Deployment

Meet Llama 3.1

The open source AI model you can fine-tune, distill and deploy anywhere. Our latest instruction-tuned model is available in 8B, 70B and 405B versions.

Start building | Download models | Try 405B on Meta AI

Llama 3.1 models

Documentation Hub

405B

Flagship foundation model driving widest variety of use cases.

Download

70B

Highly performant, cost effective model that enables diverse use cases.

Download

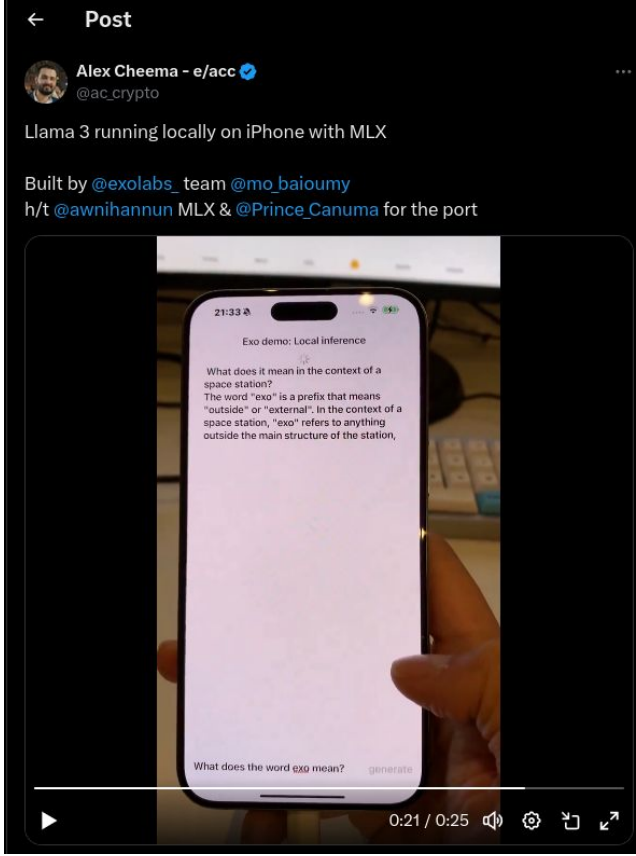
8B

Light-weight, ultra-fast model you can run anywhere.

Download

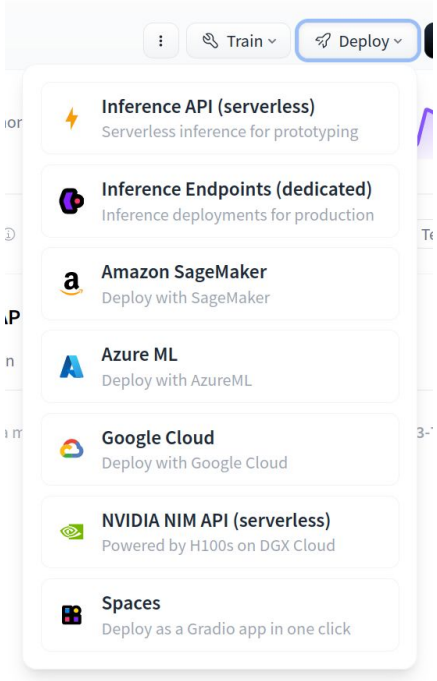
Target Inference Hardware Configuration

- 8B model – Consumer & Hobbyist
  - iPhone & PC
  - High-end gaming GPU (8-12G)
  - Low-end data center GPU (16-24G)
- 70B model – Professionals
  - 40GB-80GB GPUs (A100)
- 405B model – Big Teams
  - Workstation w/ 3 x A100(80GB)



Estimated Cost of Hosting Llama 70B

Provider	Instance Type	GPU Memory	Cost/Hour	Throughput	1M Token Cost
AWS	G5.12xlarge 4x A10g	96G	$5.672	~50/s	~$30
	G5.48xlarge 8x A10g	192G	$16.288	~100/s	~$45
AWS Bedrock API Service	??	??	$13-$21	??	??
Google Vertex AI	G2-standard-96		$4.6	~50/s	~$25



AWS instance rental cost is about buying the hardware every quarter.

Tech Details

Bits and Bytes

- 70B = 70 Billion Parameters

	1	8 Billion	70 Billion	405 Billion
32-bit floating point	4 bytes	32GB	280GB	1620GB
16-bit floating point (fp16/float16, bf16/bfloat16)	2 bytes	16GB	140GB	810GB
8-bit quantization	1 byte	8GB	70GB	405GB
4-bit quantization	½ byte	4GB	35GB	203GB

Model Deployment Optimizations

- Quantization
- Page-Attention

Floating Point Numbers

$$12.345 = \underbrace{12345}_{\text{significand}} \times \underbrace{10^{\overbrace{-3}^{\text{exponent}}}}_{\text{base}}$$

Scientific Notation

$$1.543125 \times 2^3 = 1.10001011\ldots_2 \times 2^{100_2}$$

No need to store leading 1

Conversion to Binary

	Sign	Exponent	Mantissa
Value:	+1	2^3	1 + 0.5431250333786011
Encoded as:	0	130	4556063
Binary:	☐	☑☐☐☐☐☐☑☐	☑☐☐☐☑☐☑☑☐☐☐☐☑☑☐☐☐☑☑☑

Decimal Representation	12.345
Value actually stored in float:	12.34500026702880859375
Error due to conversion:	0.00000026702880859375
Binary Representation	01000001010001011000010100011111
Hexadecimal Representation	4145851f

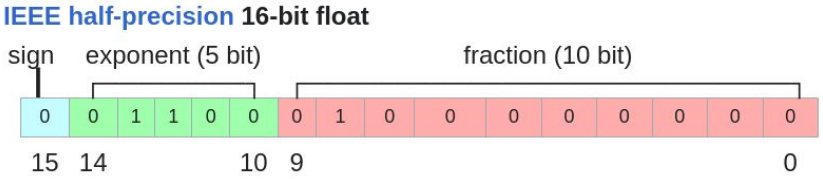
1

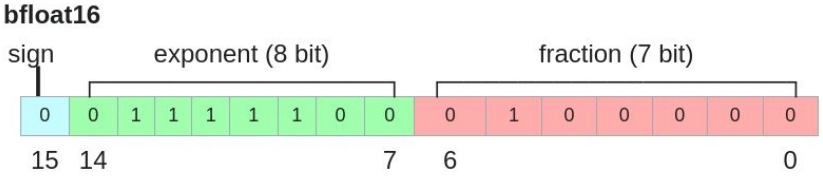
-1

https://www.h-schmidt.net/FloatConverter/IEEE754.html

IEEE 754 and bfloat16

Type	Bits				Exponent bias	Bits precision	Number of decimal digits
	Sign	Exponent	Significand	Total			
Half (IEEE 754-2008)	1	5	10	16	15	11	~3.3
Single	1	8	23	32	127	24	~7.2
Double	1	11	52	64	1023	53	~15.9
x86 extended precision	1	15	64	80	16383	64	~19.2
Quad	1	15	112	128	16383	113	~34.0



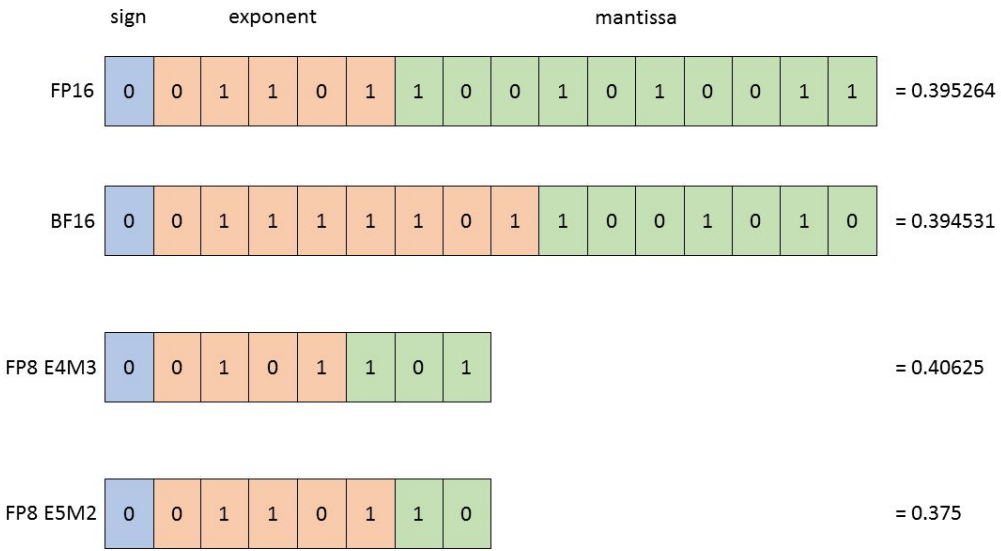


Take-away: bfloat16 has
- Same range as float32 (8-bit exponent)
- Less precision

Bfloat16 was introduced by Google brain.

FP8

- Jointly developed by Nvidia, Arm and Intel
- Less frequently used



Quantization

absmax quantization

$$\mathbf{X}_{\text{quant}} = \text{round}\left(\frac{127}{\max|\mathbf{X}|} \cdot \mathbf{X}\right)$$

$$\mathbf{X}_{\text{dequant}} = \frac{\max|\mathbf{X}|}{127} \cdot \mathbf{X}_{\text{quant}}$$

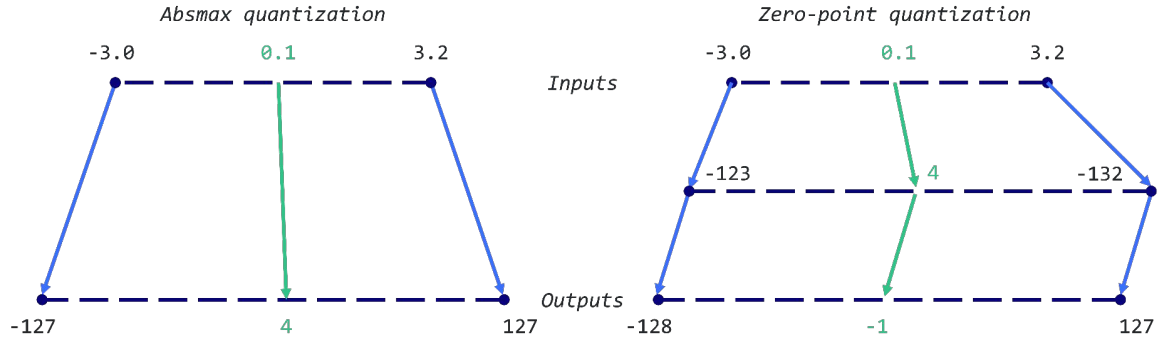
zero-point quantization

$$\text{scale} = \frac{255}{\max(\mathbf{X}) - \min(\mathbf{X})}$$

$$\text{zeropoint} = -\text{round}(\text{scale} \cdot \min(\mathbf{X})) - 128$$

$$\mathbf{X}_{\text{quant}} = \text{round}\left(\text{scale} \cdot \mathbf{X} + \text{zeropoint}\right)$$

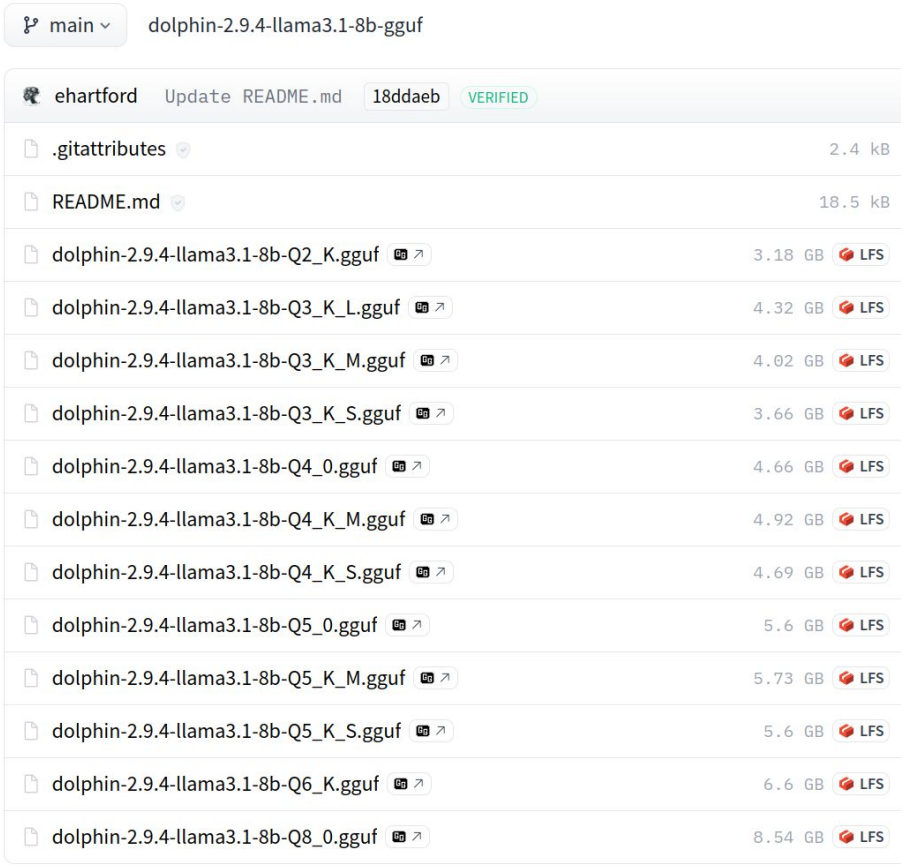
$$\mathbf{X}_{\text{dequant}} = \frac{\mathbf{X}_{\text{quant}} - \text{zeropoint}}{\text{scale}}$$



https://mlabonne.github.io/blog/posts/Introduction_to_Weight_Quantization.html

Less than 8-bit Quantization

- Popularized by llama.cpp
- Qn_0 n-bit legacy GPTQ
- Qn_K K-quantization
  - As fast or faster than legacy
- Qn_I a new SOTA method



GPTQ: Layerwise Quantization

- $\text{argmin}_{\widehat{\mathbf{W}}} ||\mathbf{WX} - \widehat{\mathbf{W}}\mathbf{X}||_2^2.$
- Optimize W layer-by-layer, w.r.t. a small set of samples (X)

Optimal Brain Damage

Yann Le Cun, John S. Denker and Sara A. Solla
AT&T Bell Laboratories, Holmdel, N. J. 07733

ABSTRACT

We have used information-theoretic ideas to derive a class of practical and nearly optimal schemes for adapting the size of a neural network. By removing unimportant weights from a network, several improvements can be expected: better generalization, fewer training examples required, and improved speed of learning and/or classification. The basic idea is to use second-derivative information to make a tradeoff between network complexity and training set error. Experiments confirm the usefulness of the methods on a real-world application.

$$w_q = \text{argmin}_{w_q} \frac{(\text{quant}(w_q) - w_q)^2}{[\mathbf{H}_F^{-1}]_{qq}},$$

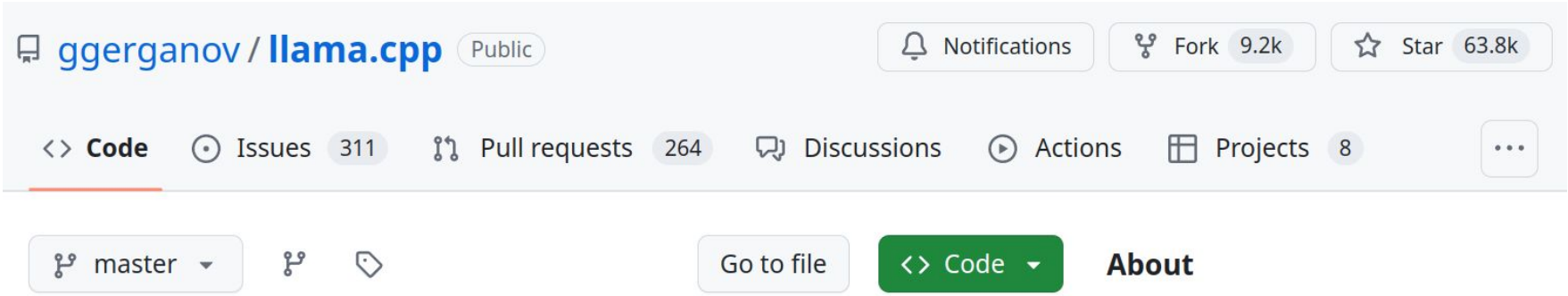
$$\boldsymbol{\delta}_F = -\frac{w_q - \text{quant}(w_q)}{[\mathbf{H}_F^{-1}]_{qq}} \cdot (\mathbf{H}_F^{-1})_{:,q}.$$

$$\mathbf{H}_F = 2\mathbf{X}_F\mathbf{X}_F^\top,$$

Frantar, et al. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers, 2022

Llama.CPP - Pure C++ Llama Inference Engine

- Started as one-person project
- Enabled LLM inference on CPU
- Aggressive quantization for memory efficiency
- Supports grammar-constrained generation



Nvidia and It’s GPUs

History of NVidia

- 1993: Founded by
  - Jensen Huang from LSI and AMD
  - Chris Malachowsky from Sun
  - Curtis Priem from IBM and Sun
- Founding vision: accelerated computing will be the future (of video gaming)
- 1995: release of 1st product – NV1
  - In late 1990s, 70 startup companies worked on graphics cards, two survived: Nvidia and ATI (merged into AMD)
- 2007: release of CUDA

Nvidia Microarchitectures

Consumer Product	Graphics Cards
Ada Lovelace (2022)	GeForce 40 series
Ampere (2020)	GeForce 30 series
Turing (2018)	GeForce 16 series GeForce 20 series
Pascal (2016)	GeForce 10 series Quadro P Tesla P4
Maxwell (2014)	GeForce 700 series GeForce 800M series GeForce 900 series
Kepler (2012)	GeForce 600 series GeForce 700 series GeForce 800M series
Fermi (2010)	GeForce 400 series GeForce 500 series
Tesla (2006)	GeForce 8 series GeForce 9 series GeForce 100 series GeForce 200 series GeForce 300 series
Curie (2004)	GeForce 6 series GeForce 7 series
Rankine (2003)	GeForce 5 series
Kelvin (2001)	GeForce 3 series GeForce 4 series
Celsius (1999)	GeForce 256 GeForce 2 series
Fahrenheit (1998)	STG-2000 RIVA 128 RIVA TNT RIVA TNT2

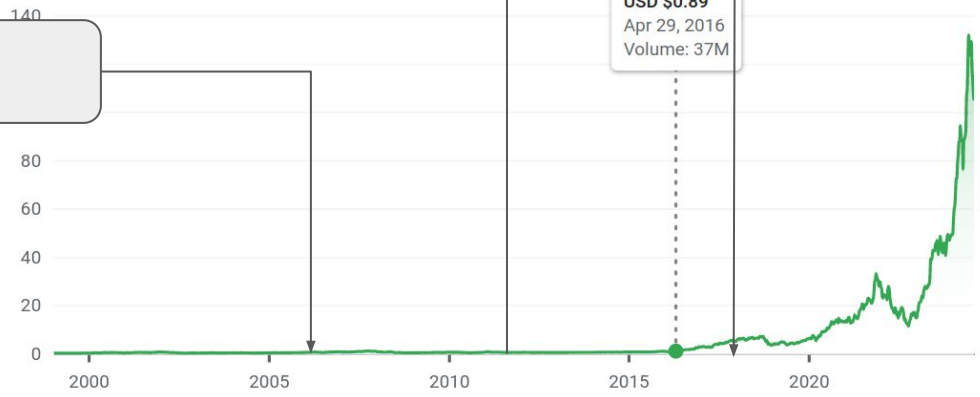
bfloat16

Professional Product	Graphics Cards
Blackwell (2024)	B100 B200
Hopper (2022)	H100 H200
Ampere (2020)	A100
Volta (2017) (Pred. - Pascal)	Tesla V Titan V Quadro GV100

Tensor Cores

CNN breakthrough

CUDA



NVDA Stock Price (normalized)

CUDA Compute Capabilities

- Features supported by GPU hardware

Feature Support	Compute Capability					
(Unlisted features are supported for all compute capabilities)	5.0, 5.2	5.3	6.x	7.x	8.x	9.0
Atomic functions operating on 32-bit integer values in global memory (Atomic Functions)	Yes					
Atomic functions operating on 32-bit integer values in shared memory (Atomic Functions)	Yes					
Atomic functions operating on 64-bit integer values in global memory (Atomic Functions)	Yes					
.....						
Unified Memory Programming (Unified Memory Programming)	Yes					
Dynamic Parallelism (CUDA Dynamic Parallelism)	Yes					
Half-precision floating-point operations: addition, subtraction, multiplication, comparison, warp shuffle functions, conversion	No	Yes				
Bfloat16-precision floating-point operations: addition, subtraction, multiplication, comparison, warp shuffle functions, conversion	No				Yes	
Tensor Cores	No			Yes		

Full Table

```
$ nvidia-smi --query-gpu=compute_cap --format=csv
compute_cap
9.0
```

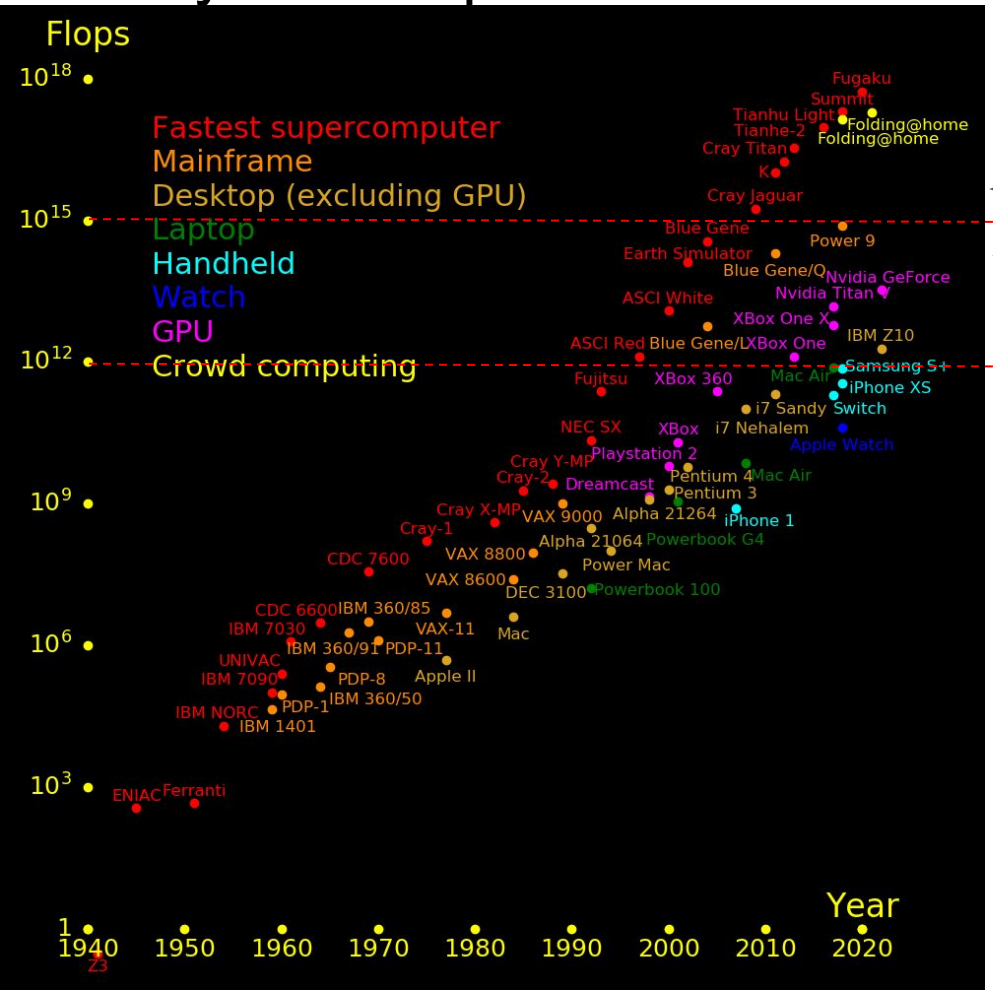
Significance of Bfloat16

	A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM
FP64	9.7 TFLOPS			
FP64 Tensor Core	19.5 TFLOPS			
FP32	19.5 TFLOPS			
Tensor Float 32 (TF32)	156 TFLOPS			
BFLOAT16 Tensor Core	312 TFLOPS			
FP16 Tensor Core	312 TFLOPS			
INT8 Tensor Core	624 TOPS			

[Source]

INT8 is not suitable for training.

History of Computation in FLOPS



H100, ~2000 TFLOPS, or 2 PFLOPS (FP16) @ $40,000

A100, ~300 TFLOPS (FP16) @ $10,000

RTX 4080, ~50 TFLOPS (FP16) @ $1500

1T FLOPS

1 **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE ~1000 PFLOPS

2 **Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel ~1000 PFLOPS

3 **Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure ~500 PFLOPS

4 **Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu

https://top500.org/

Shopping Guide of Nvidia GPUs

Consumer, RTX 4090, etc.

* Mobile/laptop version w/ ~70% TFLOPS

Ampere	RTX 30 Series (e.g. 3080)	6-24 GB	4-40 TFLOPS	$200 ~ $1500
Ada Lovelace	RTX 40 Series (e.g. 4080)	8-24 GB	10-80 TFLOPS	$200 ~ $1500

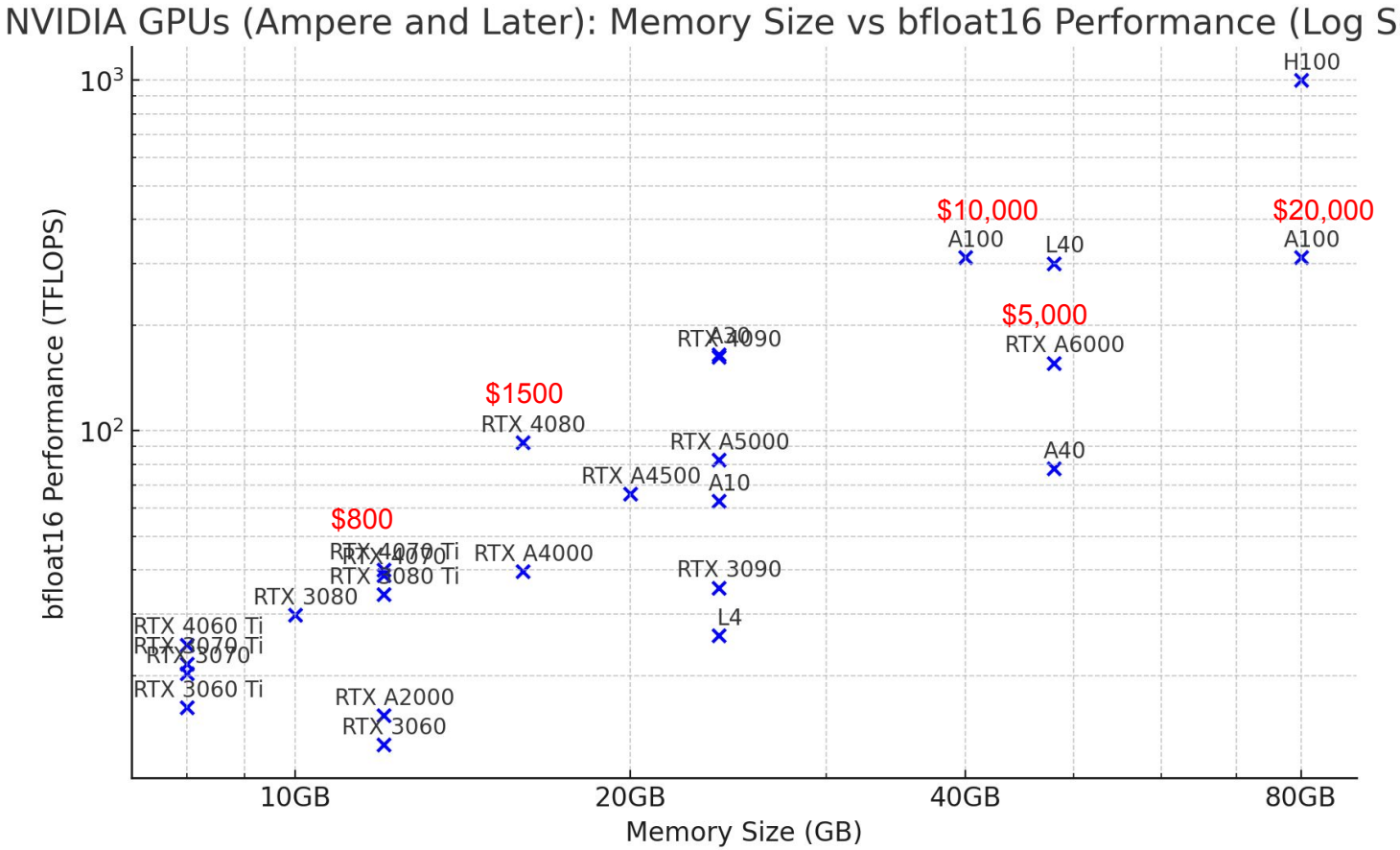
Workstations

Ampere	RTX Ax000 Series	6-48 GB	8-40 TFLOPS	$400 - $4000
Ada Lovelace	RTX x000 Ada Generation	16-48 GB	12-90 TFLOPS	$500 - $8000

Data Center

Ampere	A??	16-80 GB	20 - 300 TFLOPS	$10,000s
Hopper	H100	80 GB	700 - 1000 TFLOPS	$10,000s
Ada Lovelace	L4, L40	24 - 48 GB	100-350 TFLOPS	

Shopping Guide of Nvidia GPUs



Whole System Quotes – Feb, 2023

Qty	Description	Part Number	Unit Price	Ext. Price
1			$76,934.40	$76,934.40

Mercury GPU424 4U Server:
System: SMC 4125GS-TNRT1
CPU: (1) EPYC 9554 3.1GHz 64-Cores
Memory: 512GB DDR5-4800 (12x 48GB)
Storage: (8) 7.6TB NVMe SSD
OS Drives: (2) 480GB NVMe M.2 SSDs + Carrier
GPU: (8) L40S 48GB PCIe
*Includes NVidia EDU Discount $1400/L40S
Network: Dual Port 10Gb SFP+ + (2) SFP+ Optics
Power Supply: (2+2) Redundant
Power Cables: (4) C13/C14
BMC: Dedicated virtual KVM Port
Rail Kit
Warranty: Five-Year Bronze+ Advance Replacement Warranty
Toll-Free Phone Support Help Desk Avail. M-F 9am - 5pm PST
Free System Firmware Updates + Lifetime Technical Support.
Advance Replacement of All System Components with Prepaid Return Shipping.
Standard Return-to-Depot Warranty on Chassis.

AH-GPU424-SA02

Whole System Quotes – Feb, 2023

Qty	Description	Part Number	Unit Price	Ext. Price
1			$216,744.94	$216,744.94

Mercury GPU412 4U Server:
CPU: (2) Gold 6330 2.0GHz 28-Core
Memory: 1024GB DDR4-3200 (16x 64GB)
Storage: (8) 7.6TB NVMe SSD
GPU: (8) H100 80GB PCIe
*Price Include EDU Discount $5500/H100
Network: Two 10Gb/s LAN Ports
Power Supply: Three 80 Plus Platinum 2200W Redundant
Rail Kit
Warranty: Five-Year Bronze Advance Replacement Warranty
Advance Replacement of System and All Assoc. Components.
AHPC Covers RMA Shipping Costs To Customer,
Customer Covers RMA Shipping Costs back to AHPC
Toll-Free Phone Support Help Desk Avail. M-F 9am - 5pm PST
9 to 5 pm Local Time, Monday through Friday.
Free System Firmware Updates.

AH-GPU412-GX02

GPU Power Consumption

1 x A100	250W - 400W
1 x H100	350W - 700W – SPEC
1 DGX H100 system (8 x H100)	10 kW max
	Average kWH rate $0.1643 Monthly bill @ 50% load: ~$600
Average US household Usage	1.2 kW, or 10791 kWH/year
1 GPU is approaching 1 US household