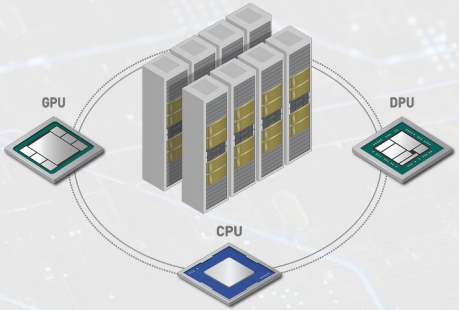

# رویکرد مناسب ایران در زیرساخت‌های پردازشی هوش مصنوعی

سید محمد محمدزاده ضیابری
رئیس کمیسیون هوش مصنوعی
سازمان نظام صنفی رایانه‌ای

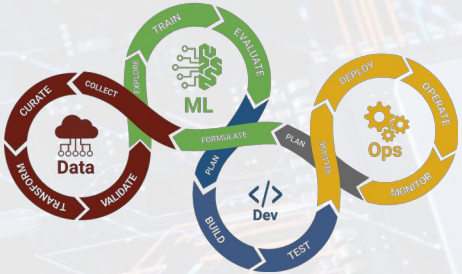
# ارکان زیرساختی هوش مصنوعی




زیرساخت پردازشی

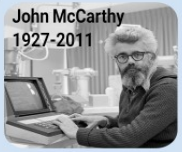
زیرساخت داده‌ای


زیرساخت یادگیری ماشین




زیرساخت مهندسی هوش

**۱۹۰۰**
اختراع لامپ خلاء
**۱۹۲۰**
اولین ربات جهان
**۱۹۳۸**
تولد منطق دیجیتال
**۱۹۴۸**
تولد علم سایبرنتیک




**۱۹۵۶**
**تولد کلمه هوش مصنوعی**

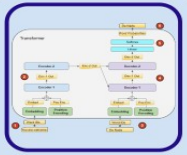
جان مک‌کارتی در مقاله‌ای با عنوان پروپوزال دارتموث برای اولین بار کلمه «هوش مصنوعی» را پیشنهاد کرد

**۱۹۸۰**
ربات نوازنده Wabot2
**۱۹۸۲**
شبکه‌های هاپفیلد
**۱۹۸۵**
خلق پردازش موازی
**۱۹۸۹**
شبکه‌های کانولوشن

**۱۹۹۷**
**شکست کاسپاروف**

ابررایانه دیپ‌بلو از شرکت IBM توانست ابرشطرنج‌باز جهان را در مسابقه شطرنج شکست دهد!

**۲۰۰۳-۲۰۰۵**
مسابقات دارپا
**۲۰۰۶**
شبکه‌های باور عمیق
**۲۰۰۸**
مغز گوگل
**۲۰۱۱**
معرفی سیری اپل



**۲۰۱۷**
شبکه‌های ترنسفرمر

"
Attention
Is
All
You
Need
"

**۲۰۱۷**
شبکه‌های wavenet
**۲۰۱۸**
GPT-1
**۲۰۱۹**
مدل زبانی BERT
**۲۰۲۰**
GPT-3

**۲۰۲۲**
ChatGPT

چت‌جی‌پی‌تی تنها در طول دو ماه توانست بیشترین میزان محبوبیت و نصب را در میان همه نرم‌افزارها کسب کند!


**۲۰۲۵**
deepseek

چت بات چینی دیپ‌سیک تنها در دو هفته توانست تعداد نصب چت‌جی‌پی‌تی را پشت سر گذاشته و به پر استفاده‌ترین نرم‌افزار هوش مصنوعی بدل شود!

خزان اول

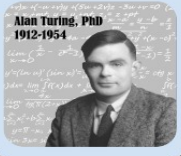
خزان دوم

| ۱۹۰۰ - ۱۹۵۰ | ۱۹۵۰ - ۱۹۷۳ | ۱۹۸۰ - ۱۹۸۷ | ۱۹۹۳ - ۲۰۰۰ | ۲۰۰۰ - ۲۰۱۲ | ۲۰۱۲ - ۲۰۱۷ | ۲۰۱۷ - ۲۰۲۱ | ۲۰۲۱ - ۲۰۲۵ | ۲۰۲۵ - ... |
|---|---|---|---|---|---|---|---|---|

۲۰۱۰

**۱۹۵۰**
**تست تورینگ**

ماشین زمانی شایسته هوشمندنامیده‌شدن است که بتواند انسان را به نحوی فریب دهد که انسان فرض شود




**۱۹۵۷**
اولین برنامه هوشمند
**۱۹۵۸**
شبکه پرسپترون
**۱۹۶۱**
اولین ربات صنعتی
**۱۹۶۶**
چت‌بات الیزا


**۱۹۸۶**
**پس‌انتشار خطا**

شبکه‌های چندلایه با قابلیت پس‌انتشار خطا انقلابی در قابلیت‌های شبکه‌های عصبی مصنوعی ایجاد کردند

**۱۹۹۶**
تولد ربوکاپ
**۱۹۹۷**
شبکه‌های LSTM
**۱۹۹۹**
تولد سگ‌های آیبو
**۲۰۰۰**
تولد ربات آسیمو

**۲۰۱۲**
AlexNet

الکس‌نت شبکه عصبی مصنوعی ۸ لایه با ۵ لایه کانولوشن توانست شبکه عصبی ImageNet را شکست دهد و از GPU برای آموزش شبکه استفاده شد.

**۲۰۱۳**
DeepMind
**۲۰۱۴**
تولد شبکه‌های مولد
**۲۰۱۵**
Tensorflow
**۲۰۱۶**
Alpha Go

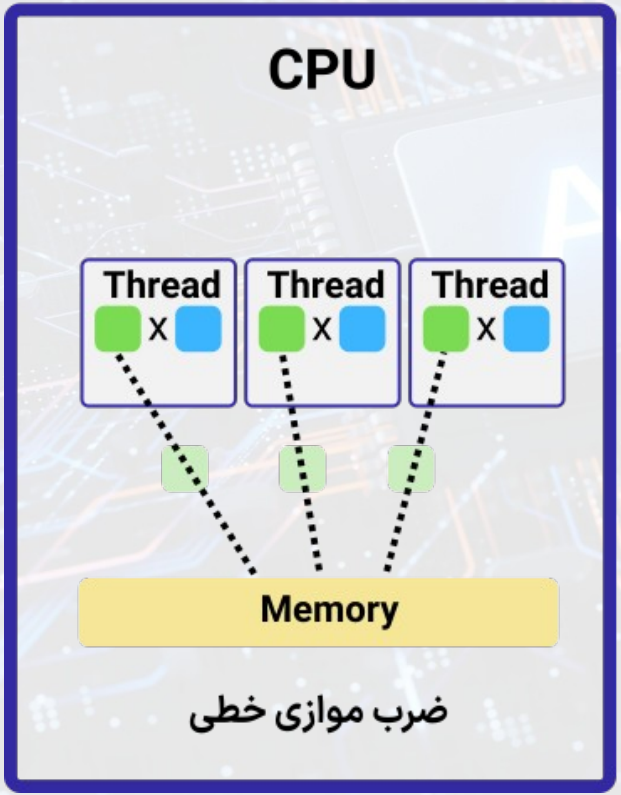

**۲۰۲۱**
DALL-E

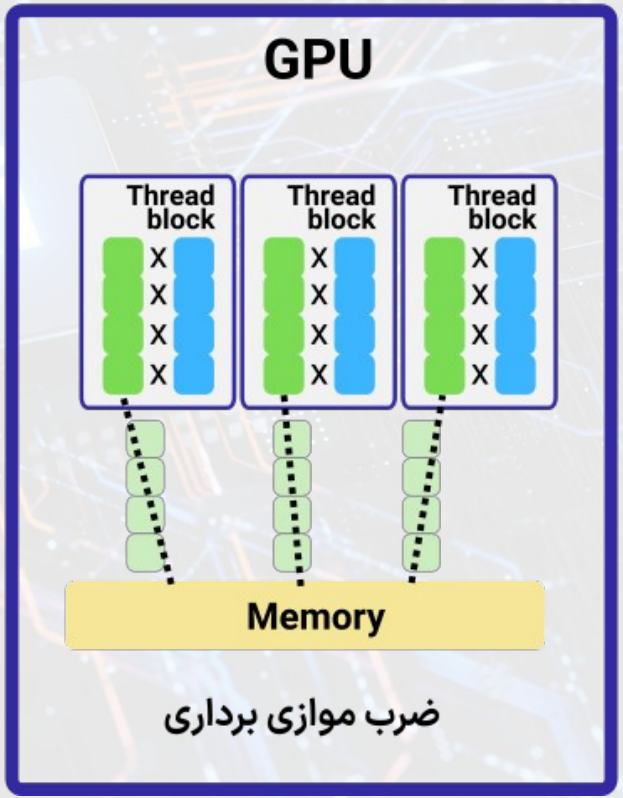
خلق تصویر از روی توضیحات،قدرت هوش مصنوعی را در انجام وظایف چند وجهی به نمایش می‌گذارد

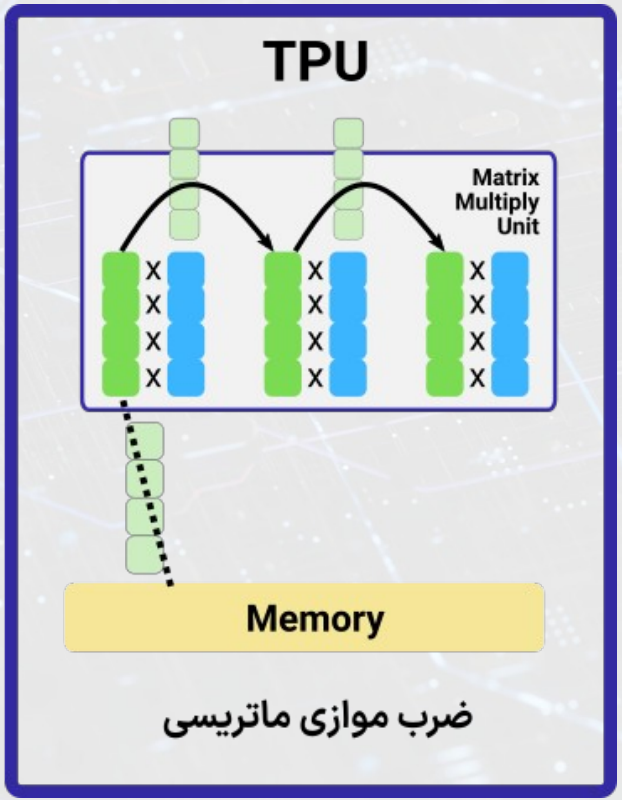

**۲۰۲۰**
خودرو خودران تسلا
**۲۰۲۱**
قاعده‌گذاری و اخلاق
**۲۰۲۴**
مدل تولید ویدیو sora
**۲۰۲۴**
EU AI Act

**+۲۰۲۵**
AGI
ASI
تکینگی
ابرانسان‌ها
و
فراتر از آن

# انواع پردازنده‌های هوش مصنوعی



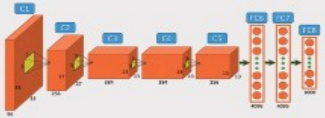



# تغییرات سهام Nvidia

Nvidia's RIVA 128 is released, its first commercially successful chip. 1997

Nvidia's first GPU, the GeForce 256, is released.

Nvidia releases Tesla, their first HPC-focused GPU architecture. 2007



AlexNext won the ImageNet Visual Recognition Challenge, using CUDA and Nvidia GPUs. 2012

Nvidia acquires Mellanox, the leader in InfiniBand networking. 2020

ChatGPT released. 2022

DGX Cloud released. 2023

Nvidia surpasses $35B in quarterly revenue, up 5x in just 6 quarters. 2024

Nvidia is Founded 1993

Nvidia's first chip, the NV1, is released. 1995

Nvidia IPOs at a $600M valuation. 1999

Programmable shaders are released with GeForce 3. 2001

CUDA is released, introducing GPU computing. 2006

The Kepler architecture is released. 2010

The Fermi architecture is released.

Maxwell architecture released. 2014

OpenAI founded. 2015

Pascal architecture released. 2016

Volta architecture released. 2017

Ampere architecture released. 2019

Turing architecture released. 2018

Hopper architecture released, Grace CPU chip announced, and DGX server released.

1993 1997 2001 2005 2009 2013 2017 2021 2024

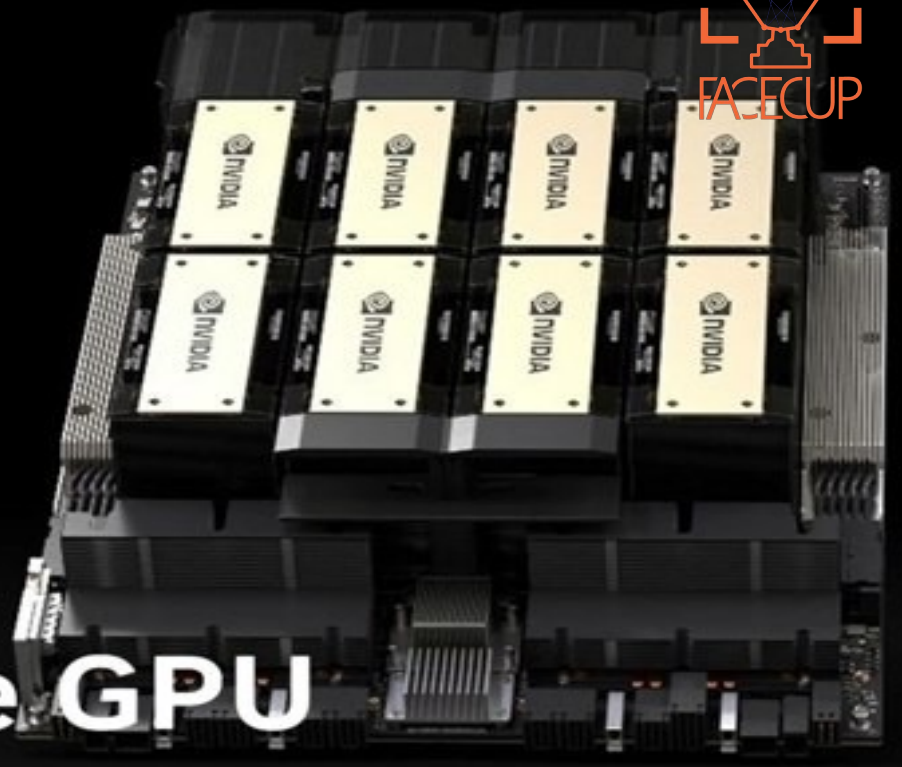
Passed $1B in annual revenue

Passed $10B in annual revenue

NVIDIA

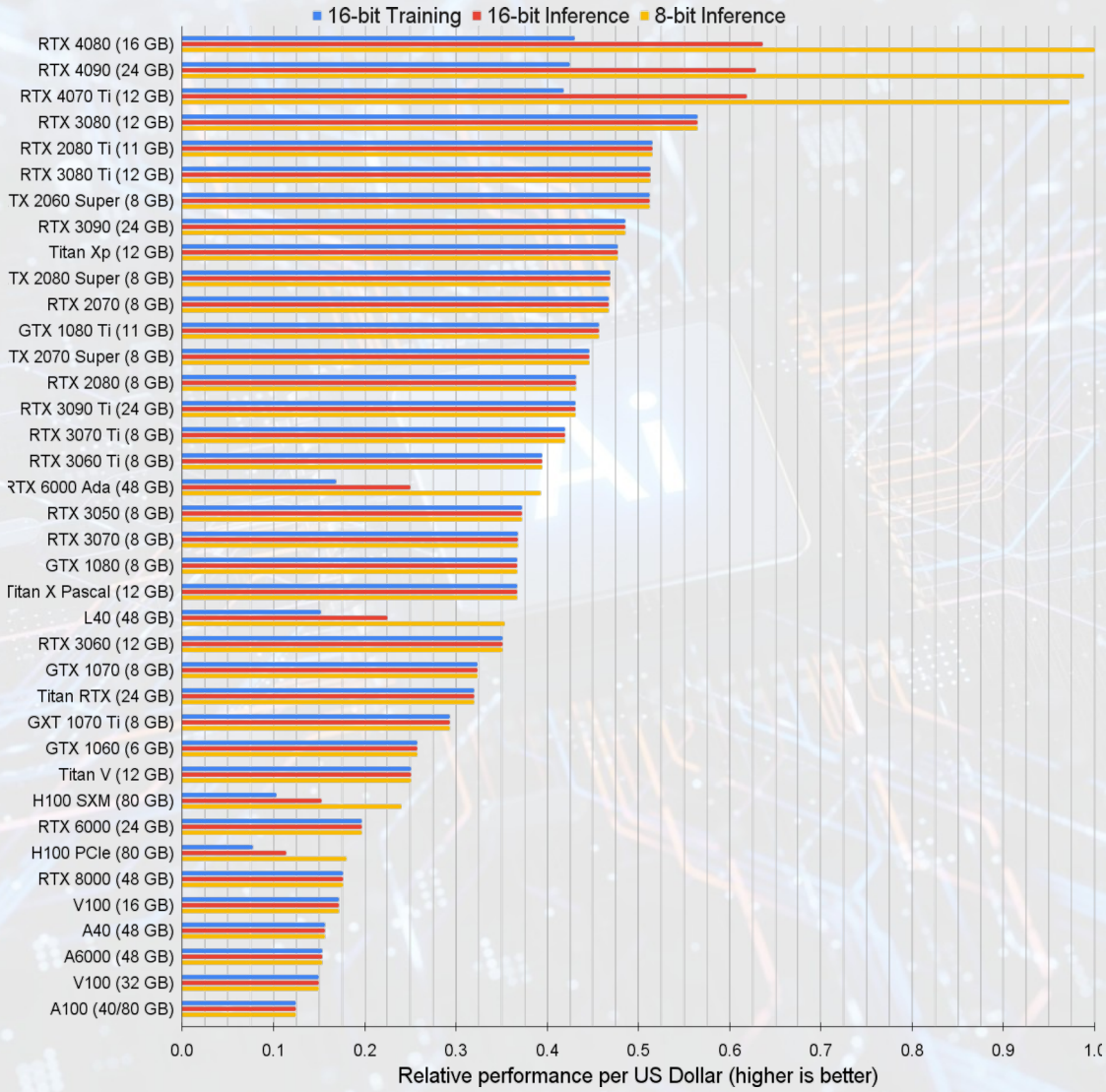
# H200

## Tensor Core GPU




FACECUP

کمیسیون هوش مصنوعی و علم داده
سازمان نظام صنفی رایانه‌ای استان تهران

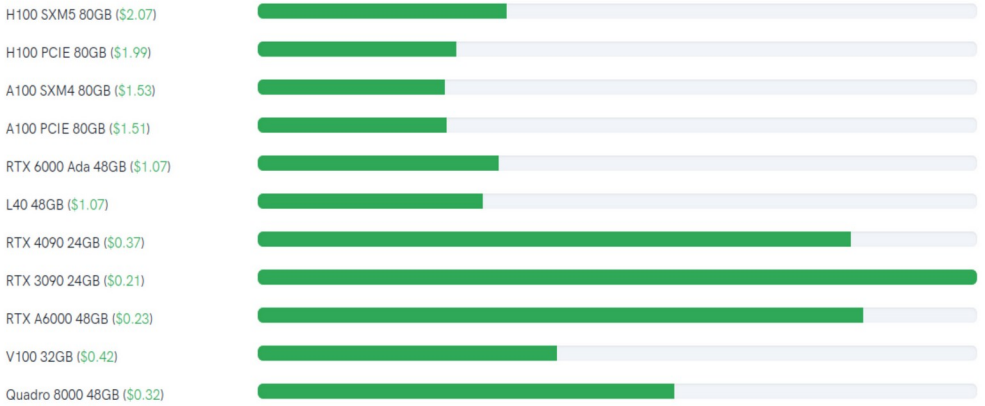



# کارایی به قیمت کارت‌های گرافیک در پردازش‌های ۱۶ بیت

## LLM Inference Throughput

Relative tokens per second on Mistral 7B. Half precision (FP16). The more, the better.

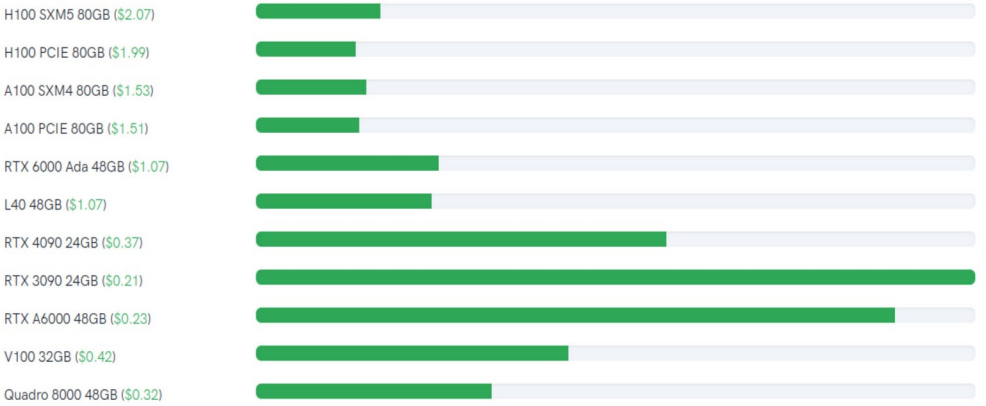
Mistral 7B | Throughput per dollar




## LLM Inference Throughput

Relative tokens per second on Mistral 7B. Half precision (FP16). The more, the better.

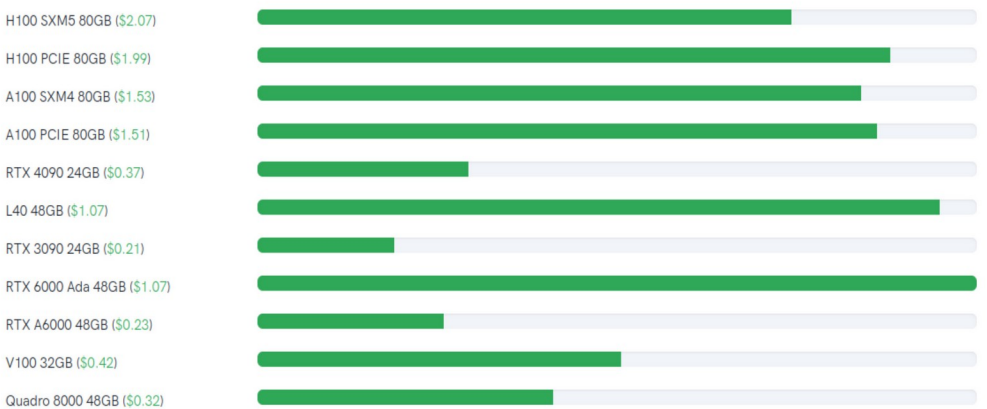
Facebook OPT-125M | Throughput per dollar




## LLM Batch Latency

Time taken to process one batch of tokens, p90, Mistral 7B. Half precision (FP16). The lower, the better.

Latency per dollar



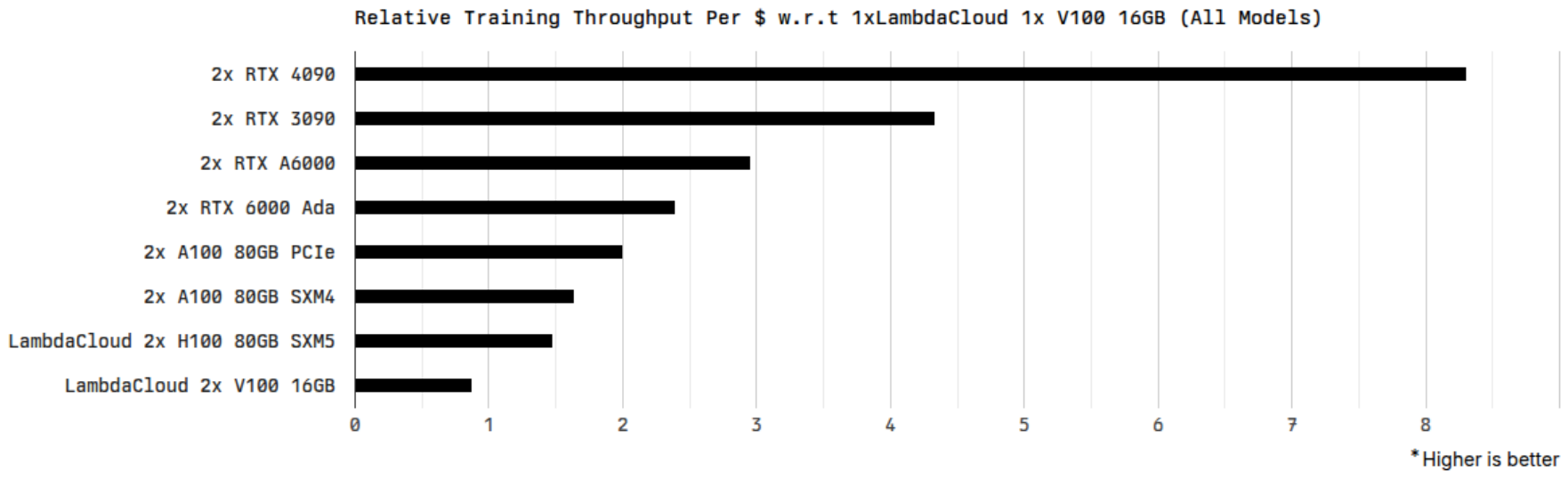
Lambda

AI Factories Enterprise Pricing Products Company



# PyTorch Training GPU Benchmarks

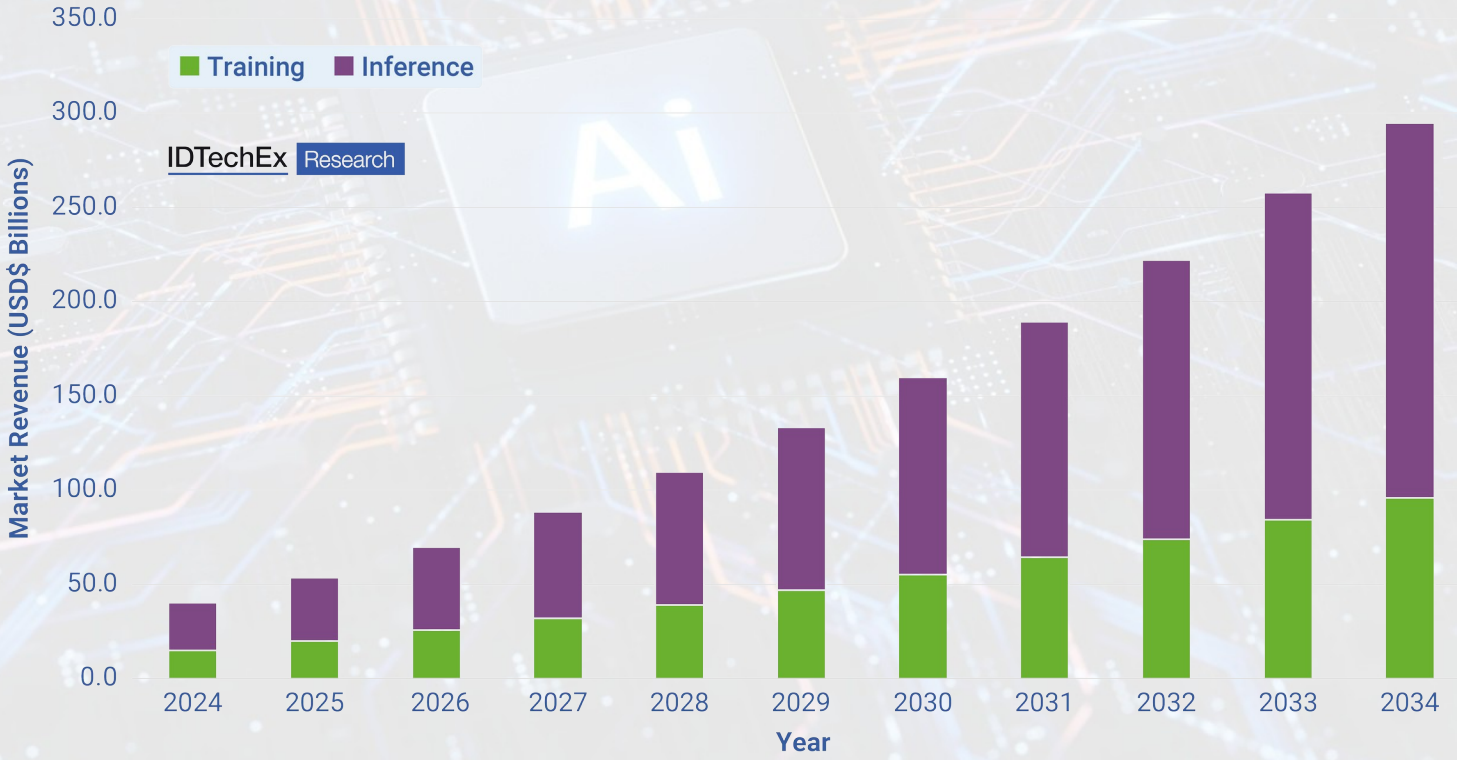
| Visualization | Metric | Precision | Number of GPUs | Model |
|---|---|---|---|---|
| chart | throughput/$ | fp16 | 2x | All Models |

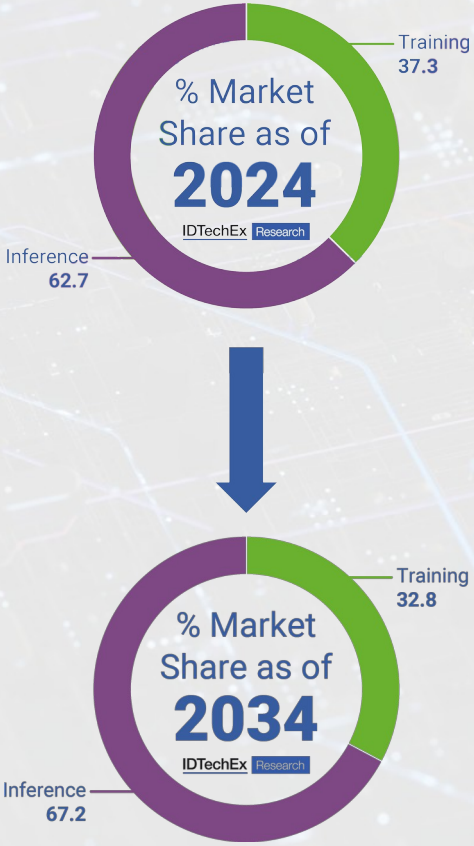



https://lambda.ai/gpu-benchmarks

# سرمایه‌گذاری‌ها از سال ۲۰۱۷ تاکنون

# پیش‌بینی تغییرات بازار در دو حوزه آموزش و استنتاج

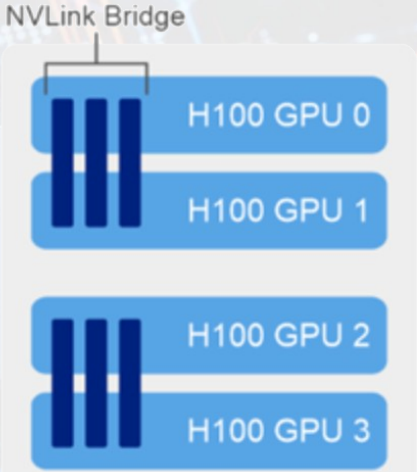


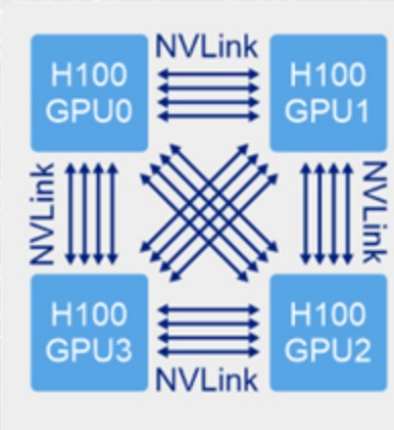
# کارت متناسب با کاربرد

# انواع اتصالات کارت‌ها

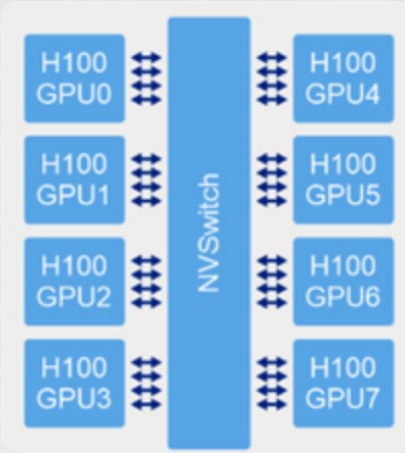

کارت‌های مستقل PCI




زوج کارت PCI متصل با NVLink



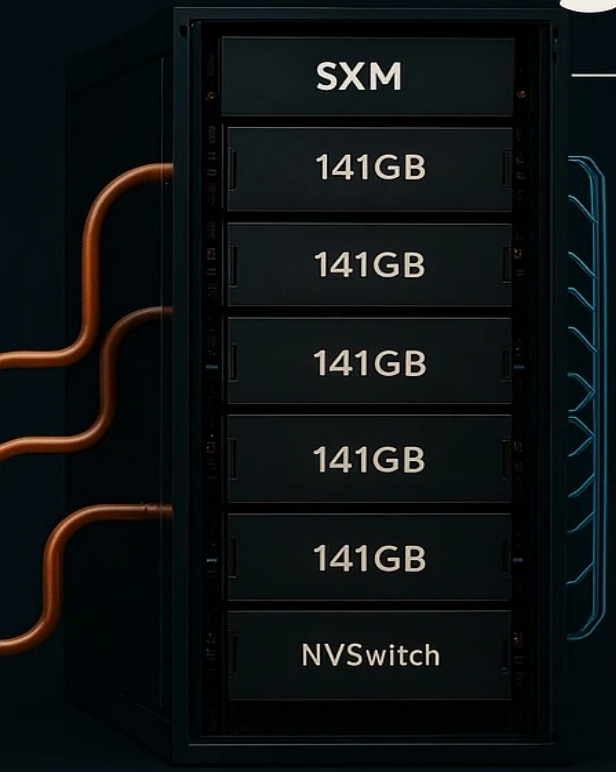

اتصال Mesh بین کارت‌های SXM

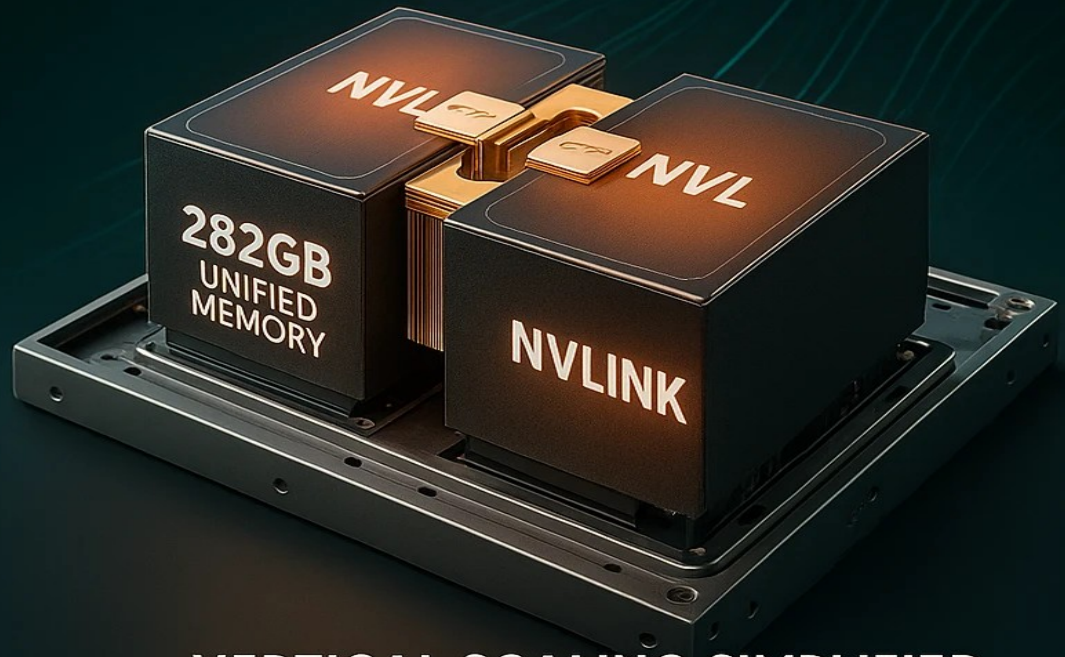



اتصال بین سرورها با NVSwitch

# SXM vs NVL

## DESIGN MATTERS




**HORIZONTAL GPU DENSITY**

**H200**

**VERTICAL SCALING SIMPLIFIED**

**H200 NVL**

# مقایسه SXM با NVL

| Feature | H200 SXM | H200 NVL |
|---|---|---|
| Total Memory | 141GB per GPU | 282GB unified pool |
| Memory Bandwidth | 4.8 TB/s per GPU | 9.6 TB/s total |
| NVLink Speed | 900 GB/s (between GPUs) | 1.8 TB/s (internal bridge) |
| Ideal For | Models fitting within 141GB per GPU | AI models >141GB |
| **Llama 70B, GPT-4 inference** | Requires sharding across GPUs<br>Communication delays between GPUs | 1.6–2× throughput<br>Entire model fits in 282GB unified memory |
| **Image upscaling, Financial modeling** | **Matches NVL per-GPU performance** | **No advantage**<br>(Same ~67 TFLOPS FP64/GPU) |

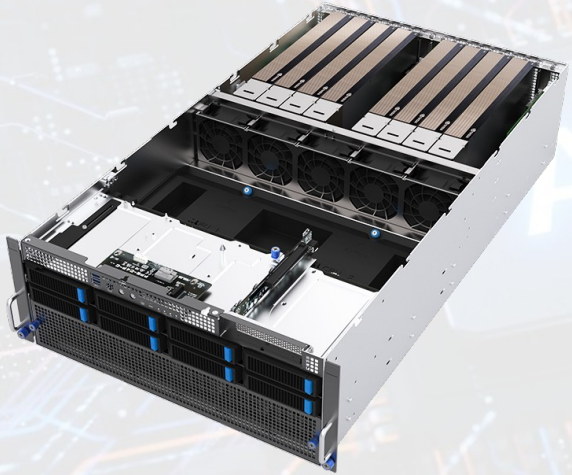


https://uvation.com/articles/nvidia-h200-sxm-vs-h200-nvl-choosing-the-right-ai-powerhouse

# انواع سرورهای Nvidia

| DGX ™ | HGX ™ | MGX ™ | EGX ™ |
| --- | --- | --- | --- |
|  |  |  |  |
| سرور آماده همراه نرم‌افزار هوش مصنوعی SXM | سرور منعطف برای کاربردهایی مختلف SXM | سیستم‌های مناسب برای پردازش لبه PCIe | سکوی مرجع ماژولار و انعطاف‌پذیر PCIe |

# انواع دیگر سرورها



ASUS

SUPERMICRO

GIGABYTE

# مقایسه سرورهای PCI با هم

| سازنده | بیشترین تعداد کارت | پشتیبانی از RTX4090 | مهمترین مزایا | مهمترین معایب |
|---|---|---|---|---|
| NVIDIA | ۸ | خیر | عملکرد بالا با NVLink، بهینه‌سازی کامل برای AI/HPC، مقیاس‌پذیری آسان در کلاسترها | گران‌قیمت (از ۲۰۰,۰۰۰ دلار)، انعطاف‌پذیری کم برای سفارشی‌سازی |
| HEWLETT PACKARD | ۸ | خیر | پشتیبانی enterprise قوی، امنیت بالا، ادغام با hybrid cloud | رشد کندتر در بازار AI، گران‌تر برای حجم کاری بالا، پشتیبانی محدود از کارت‌ها |
| SUPERMICRO | ۲۰ | بله | هزینه رقابتی، انعطاف‌پذیری در PCIe و پشتیبانی از GPUهای متنوع | مصرف برق بالا در کانفیگ‌های متراکم، پشتیبانی ضعیف‌تر نسبت به enterprise |
| GIGABYTE | ۱۶ | بله | ارتقاپذیری آسان، دارای گزینه‌های liquid cooling | بهینه‌سازی کمتر برای workloads سنگین AI نسبت به NVIDIA |
| ASUS | ۸ | بله | خنک‌سازی عالی، PCIe 5.0 برای آینده‌نگری، کارایی بالا در inference | کمتر شناخته‌شده |
| RIG | ۲۰ | بله | آزادی عمل بالا، هزینه بسیار پایین | فاقد استانداردهای مراکز داده |

# انواع خدمات پردازشی

- **Dedicated**: سرورهای اختصاصی به صورت تکی و خوشه‌های پردازشی
- **VM/Cloud**: ماشین‌های مجازی یا ابری با پردازنده گرافیکی اختصاصی به صورت تکی یا خوشه‌ای
- **VPDC**: مرکز داده اختصاصی به صورت مجازی شامل تمامی امکانات پردازشی و شبکه به صورت مدیریت‌شده


- **Sliced GPU**: پردازنده گرافیکی قطعه‌بندی شده و تخصیص داده‌شده به صورت اختصاصی (صرفا برای کارت‌های حرفه‌ای)
- **Shared GPU**: فناوری اشتراک‌گذاری واقعی پردازنده گرافیکی به صورت کاملا پویا قابل بهره‌برداری بر روی تمامی کارت‌ها
- **GPUDN**: شبکه پردازش هوش مصنوعی نزدیک به مصرف‌کننده مورد استفاده در مناطق محروم و با شبکه ضعیف

# جمع‌بندی

- مشکلات بومی راه‌کار بومی نیاز دارد
- هیچ گزینه حلال‌المسایلی وجود ندارد!
- ترندهای جهانی لزوماً برای ایران مناسب نیستند
- تفکیک دقیق و الزامی زیرساخت‌های استنتاج از آموزش
- با مهندسی هوش مصنوعی می‌توان نیاز به سخت‌افزار را بهینه کرد
- بهره‌وری باید اصل اول در انتخاب سخت‌افزار باشد نه ترند بودن یا مقالات