شرکت‌های اصلی تولیدکننده تراشه‌های هوش مصنوعی و GPU

هایده بهگام

این مقاله به کمک هوش مصنوعی نگارش شده است.

۵ مرداد۱۴۰۴

هوش مصنوعی تحولی اساسی درزندگی انسانها ایجادکرده وتوجه بسیاری ازمردم جهان را به خودجلب است.این فن آوری که ازدوبخش سخت افزاروزنرم افزار تشکیل شده ، باعث تحولات اساسی دراقتصاد، سیاست ، آموزش ، امورنظامی و حتی ارزشهای انسانی شده ومی شود.باتوجه به سوالات بسیاری که درارتباط با آن مطرح است وترس ها وامیدهایی که ایجاد کرده، برآن شدم دراین مقاله به معرفی اجمالی شرکتهایی بپردازم که درایجادوتولید سخت افزارهای هوش مصنوعی پیشتازندوبنابراین سکان هدایت این فن آوری فعلادردست آنهاست.

شرکت‌های اصلی تولیدکننده تراشه‌های هوش مصنوعی و GPU شامل موارد زیر هستند:

۱. **NVIDIA**

- رهبر بازار در تولید GPUهای مخصوص هوش مصنوعی و یادگیری عمیق.
- محصولات کلیدی:
- H100, A100برای دیتاسنترها و مدل‌های بزرگ مانند (ChatGPT)

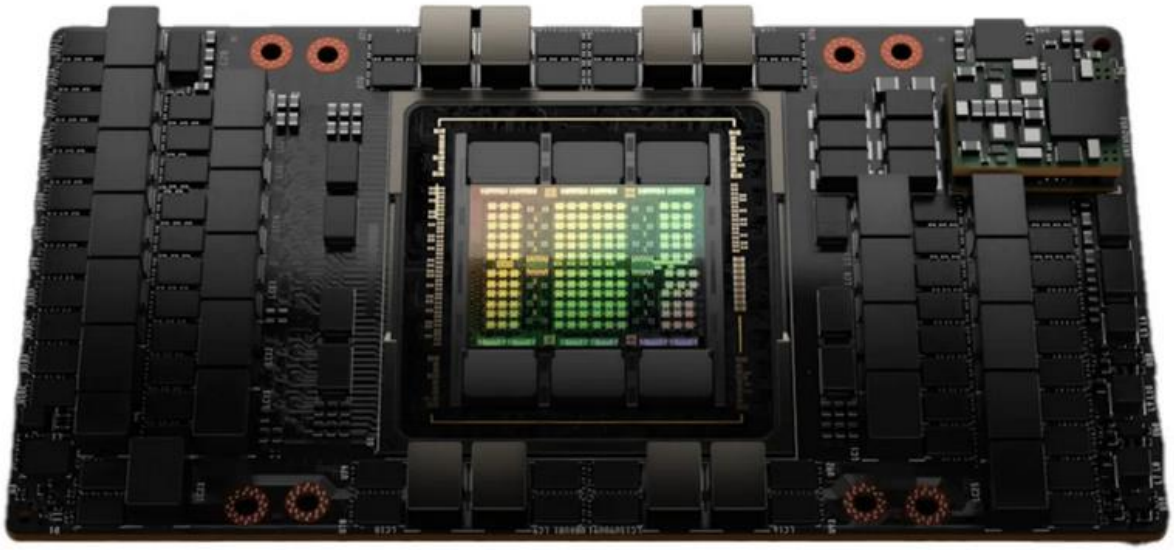
این تراشه‌های انویدیا به‌طور اختصاصی برای کاربردهای هوش مصنوعی، یادگیری عمیق (Deep Learning) و محاسبات در**مراکز داده** طراحی شده‌اند. آنها پایه‌گذار مدل‌های بزرگ زبانی (LLM)مانندChatGPT ، GeminiوLLaMA هستند.

درجدول زیرقابلیت های این تراشه ها باهم مقایسه شده است:

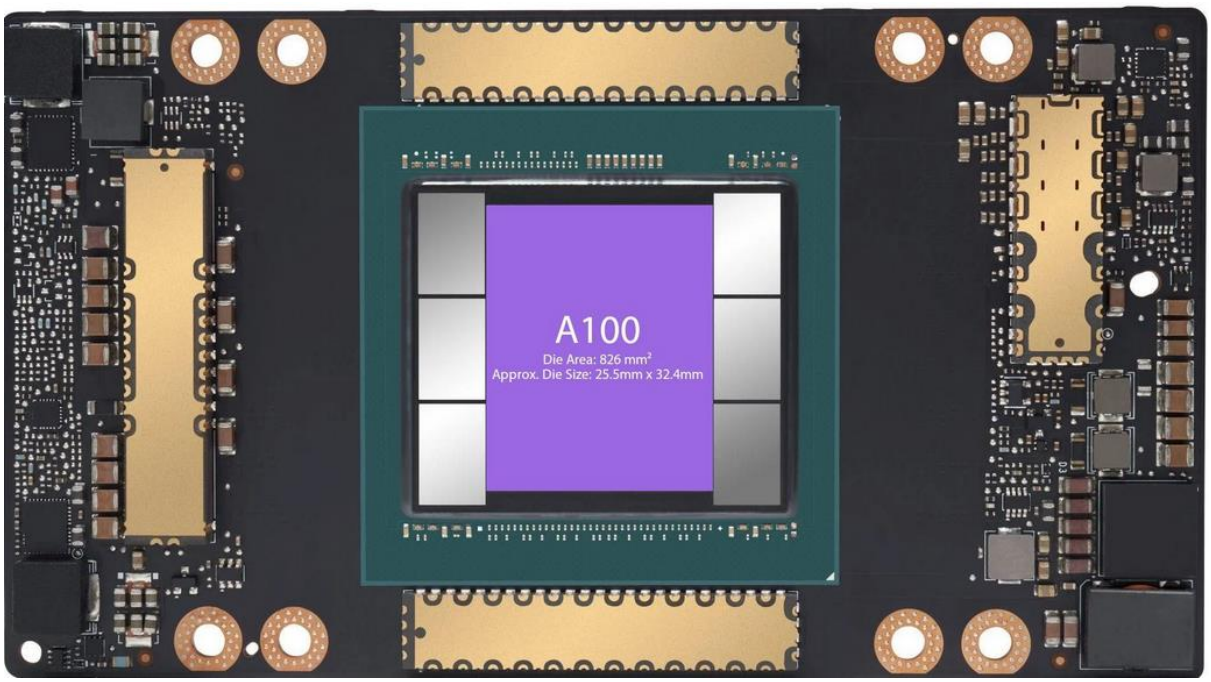
مقایسه کلیدی A100 و H100

ویژگی	NVIDIA H100 (Hopper)	NVIDIA A100 (Ampere)
معماری	Hopper (جدیدتر)	Ampere
ساخت تراشه	4nm (TSMC)	7nm (TSMC)
حافظه (VRAM)	80GB HBM3	40GB/80GB HBM2e

ویژگی	NVIDIA H100 (Hopper)	NVIDIA A100 (Ampere)
پهنای باند حافظه	3TB/s	2TB/s
محاسبات FP64 (HPC)	60 TFLOPS	19.5 TFLOPS
محاسبات FP16 (AI)	**2000 TFLOPS**	624 TFLOPS
تنسور کور (Tensor Core)	نسل ۴ (پشتیبانی از FP8)	نسل ۳ (FP16, TF32)
سرعت NVLink	900GB/s	600GB/s
نسخه PCIe	PCIe 5.0	PCIe 4.0
توان مصرفی (TDP)	700W	400W
کاربرد اصلی	مدل‌های فوق‌پیشرفته (GPT-4, Gemini)	هوش مصنوعی عمومی و HPC



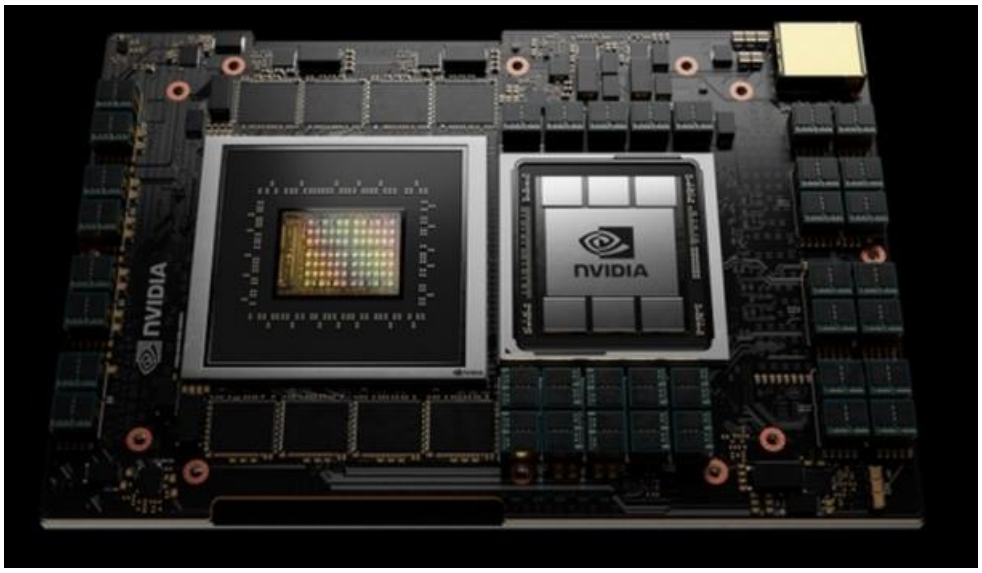
NVIDIA H100



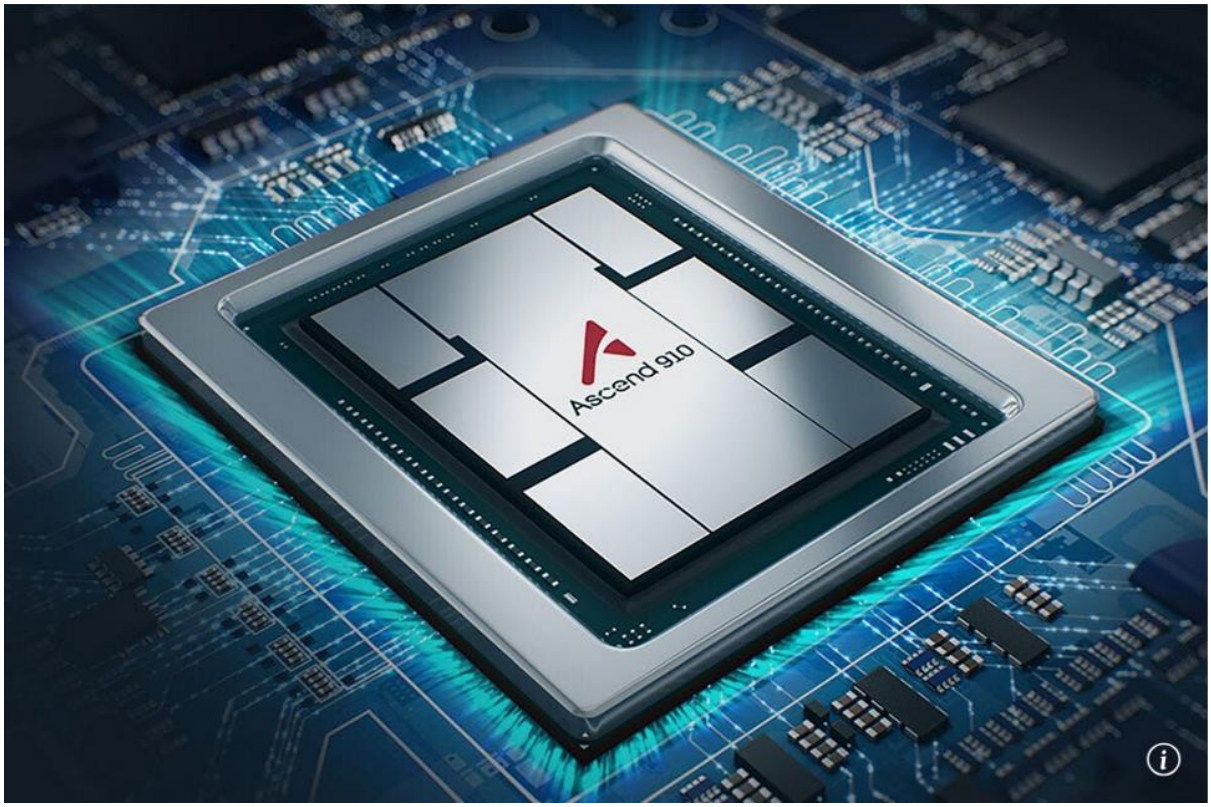
NVIDIA A100

H100از A100 قدرتمندتر است.زیرا معماری آن بهتراست.حافظه آن بیشتراست.مصرف انرژی آن کمتراست.سرعت NVLink آن بیشتراست(اتصال چندین GPU با سرعت بالاتر)

NVIDIA H20: تراشه جدید هوش مصنوعی برای بازار چین (با محدودیت‌های صادراتی آمریکا)است.



انویدیا ، H20را به‌عنوان یک تراشه کاهش‌یافته (Downgraded) برای رفع محدودیت‌های صادراتی آمریکا به چین طراحی کرده است. این کارت ، جایگزین H100و A800 در بازار چین شده و برای حفظ حضور انویدیا در این بازار بزرگ، با رقبای چینی مانند Ascend 910Bهواوی رقابت می‌کند.



H20 مشخصات فنی کلیدی

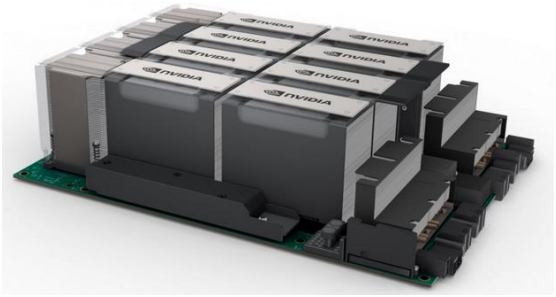
ویژگی	NVIDIA H20	NVIDIA H100 (مقایسه)
معماری	Hopper (کاهش‌یافته)	Hopper (کامل)
محاسبات FP16 (AI)	~148 TFLOPS	2000 TFLOPS
حافظه (VRAM)	96GB HBM3	80GB HBM3
پهنای باند حافظه	~4TB/s	3TB/s
سرعت NVLink	900GB/s	900GB/s
توان مصرفی (TDP)	400W	700W
قیمت (تقریبی)	~20,000 دلار	~30,000-40,000 دلار
محدودیت صادراتی	تایید شده برای چین	ممنوع برای چین

یکی ازعلت های ساخت H20 تحریم های دولت آمریکا برای فروش تراشه های **پیشرفته** مانند **H100** و **A100** به چین است (مقاله مرداد۱۴۰۴ نگارش شده است.) چین به دنبال جایگزینی مانند Ascend 910Bاست و انویدیا می‌خواهد سهم بازار خود را درچین حفظ کند.

H20 عملکردی پایین‌تر از H100 دارد تا از محدودیت‌های محاسباتی آمریکا عبور کند.

لازم به ذکر است که انویدیا ازنامگذاری H بعلاوه عدد برای تراشه های هوپر استفاده می کند. مانند H100 یا H20یا H200.

H200جدیدترین عضو خانواده معماری هوپر (Hopper) انویدیا است که به‌طور اختصاصی برای مدل‌های عظیم هوش مصنوعی (LLM) و ابرمحاسبات (HPC) طراحی شده است. این تراشه جانشین H100محسوب می‌شود و با حافظه‌ی HBM3e و ظرفیت بالاتر، تحولی در پردازش داده‌های حجیم ایجاد کرده است. فعلاً یک **تراشه انحصاری برای سازمان‌های بزرگ** است و قیمت آن برای اکثر شرکت‌ها بسیار بالا است . **چین، روسیه و ایران به‌طور رسمی نمی‌توانند H200 را خریداری کنند.هند**،نیاز به مجوزهای خاص برای خرید باحجم بالادارد.



H200

کشورهای اصلی تحویل گیرنده آمریکا،کانادا،آلمان، فرانسه، بریتانیا، سوئد،ژاپن، کره جنوبی، سنگاپور،استرالیا، نیوزیلندوبرخی کشورهای خلیج فارس مانند امارات وعربستان برای پروژه های خاص هستند.

مقایسه‌ی H100 با H200 🔻

ویژگی	H200	H100
معماری	Hopper (نسخه ارتقا یافته)	Hopper
حافظه (VRAM)	**141GB HBM3e**	80GB HBM3
پهنای باند حافظه	**4.8TB/s**	3TB/s
محاسبات FP16 (AI)	~2000 TFLOPS	~2000 TFLOPS
محاسبات FP64 (HPC)	~60 TFLOPS	~60 TFLOPS

ویژگی	H200	H100
توان مصرفی (TDP)	700W	700W
تنسور کور (Tensor Core)	نسل ۴ (پشتیبانی از FP8)	نسل ۴
کارایی در مدل‌های LLM	**تا ۲ برابر سریع‌تر**	پایه‌ی عملکرد

قیمت هرکارت آن بین۳۵۰۰۰تا۴۰۰۰۰دلاراست که درچین بامحدودیت صادرات ۵۰ درصد گرانتراست.(مقاله ، مرداد۱۴۰۴ نگارش شده است)

- RTX 4090, RTX 6000 Ada برای کاربردهای ایستگاه‌های کاری و هوش مصنوعی

مقایسه RTX 4090 و RTX 6000 Ada برای ایستگاه‌های کاری و هوش مصنوعی

معرفی کلی

ویژگی	RTX 4090 (مصرف‌کننده)	RTX 6000 Ada (حرفه‌ای)
معماری	Ada Lovelace	Ada Lovelace
حافظه (VRAM)	24GB GDDR6X	**48GB GDDR6 ECC**
پهنای باند حافظه	1 TB/s	960 GB/s
CUDA Cores	16,384	18,176
توان محاسباتی (FP32)	82.6 TFLOPS	91.1 TFLOPS

ویژگی	(مصرف‌کننده) RTX 4090	(حرفه‌ای) RTX 6000 Ada
توان مصرفی (TDP)	450W	300W
قیمت (تقریبی)	~1,600 دلار	~6,800 دلار
هدف بازار	گیمینگ/هوش مصنوعی سبک	ایستگاه‌های کاری حرفه‌ای

کاربردهای کلیدی:

RTX 4090 (هوش مصنوعی سبک‌تر و گیمینگ)

- مدل‌های زبانی کوچک (7B-13B پارامتر) : اجرای محلی LLaMA 2 ، Mistral .
- D رندرینگ و طراحی ۳: **Blender ، Maya، Unreal Engine**.
- یادگیری عمیق مبتدی **:آموزش مدل‌های کامپیوتر ویژن با PyTorch/TensorFlow**
- محدودیت **:**حجم کم حافظه (24GB) **برای مدل‌های بزرگ.**

RTX 6000 Ada(هوش مصنوعی حرفه‌ای و رندرینگ صنعتی)

- مدل‌های بزرگ (+70B پارامتر) : **اجرای GPT-3.5 ، Stable Diffusion XL**
- شبیه‌سازی‌های مهندسی**ANSYS :، .SOLIDWORKS**
- پردازش تصویر پزشکی **MRI/CT :با نرم‌افزارهای مانند Horos**
- مزیت ECC Memory **:برای محاسبات بدون خطا بعلاوه** NVLink **برای اتصال چند**GPU .

محدودیت‌های صادراتی کارت‌های گرافیک NVIDIA RTX 4090 و RTX 6000 Ada بر اساس تحریم‌های ایالات متحده علیه چین (شامل چین، هنگ کنگ و ماکائو) و روسیه اعمال شده‌اند. این محدودیت‌ها به دلیل نگرانی‌های امنیتی و جلوگیری از استفاده نظامی یا هوش مصنوعی پیشرفته توسط این کشورها وضع شده‌اند.

این کارت‌های گرافیک به دلیل قدرت پردازشی بالا در هوش مصنوعی و محاسبات پیشرفته ، مشمول قوانین کنترل صادرات BIS (اداره صنعت و امنیت ایالات متحده) شده‌اند. هدف اصلی این تحریم‌ها، جلوگیری از تقویت توان محاسباتی چین و روسیه در زمینه‌های نظامی و هوش مصنوعی است.

- فروش مستقیم این کارت‌ها به چین و روسیه ممنوع است.

- برخی شرکت‌ها نسخه‌های ضعیف‌تر (مانند RTX 4090Dبرای چین) را عرضه کرده‌اند که محدودیت‌های پردازشی دارند.

- قاچاق این کارت‌ها به کشورهای تحریم‌شده ممکن است منجر به جریمه برای شرکت‌های فروشنده شود.

- **Jetson برای هوش مصنوعی** در دستگاه‌های لبه . Edge AI یا **هوش مصنوعی لبه** به معنای اجرای مدل‌های هوش مصنوعی دردستگاههای محلی مانندموبایل است وبدون اتصال به اینترنت نیزکارمی کند.هوش مصنوعی لبه باسرعت بیشتری داده ها راپردازش می کندزیرانیازی به ارسال داده ها به فضای ابری نیست وبعلاوه به همین دلیل درمحیط های دورافتاده یا محیطهای امنیتی مفیدتراست هزینه پهنای باند ندارد.داده ها ازدستگاه خارج نمی شوند وحفظ حریم خصوصی داریم.

ازکاربردهای هوش مصنوعی لبه می توان به مواردزیراشاره کرد:

🛠 الف) رباتیک:

- کنترل ربات‌های صنعتی، پهپادها و سیستم‌های تحویل خودکار.
- پردازش تصویر برای **هدایت ربات‌ها در محیط‌های دینامیک**.

🚗 ب)خودروهای خودران:

- پردازش داده‌های سنسورهای **لیدار، رادار و دوربین** در خودروهای بدون راننده.
- اجرای مدل‌های **تشخیص اشیا، مسیریابی و جلوگیری از تصادف.**

🛡 پ)پزشکی و تشخیص تصویر:

- پردازش تصاویر **MRI و اشعه X** در دستگاه‌های پزشکی.

- **تشخیص بیماری‌ها** با کمک هوش مصنوعی روی دستگاه‌های قابل حمل.

ت) نظارت هوشمند:

- **تشخیص چهره، پلاک خودرو و رفتارهای مشکوک** در دوربین‌های امنیتی.
- پردازش ویدئو **بدون نیاز به ارسال داده به ابر.**

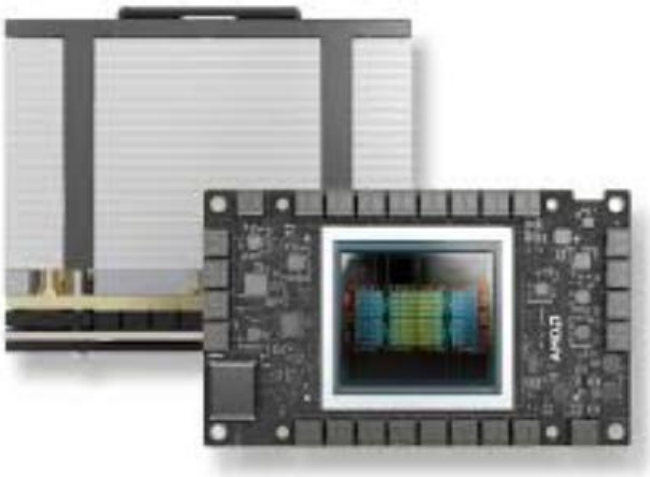
ث) اینترنت اشیا:(IoT)

- اجرای مدل‌های هوش مصنوعی روی **دستگاه‌های لبه (Edge Devices)** مانند سنسورهای هوشمند.

۲. AMD (Advanced Micro Devices)

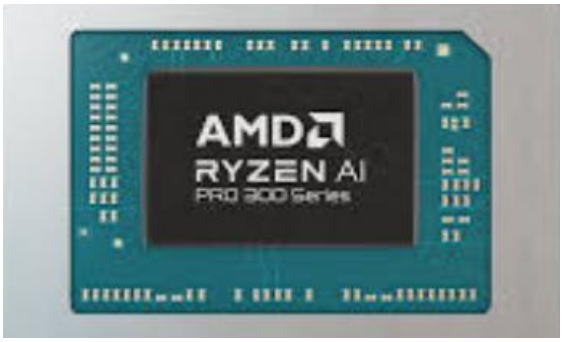
AMD رقیب اصلی NVIDIA در بازار GPU و پردازنده‌های هوش مصنوعی است.

- محصولات کلیدی:
- Instinct MI300X (برای هوش مصنوعی و محاسبات سطح بالا)
- Radeon RX 7900 XTX(برای گیمینگ و پردازش موازی)
- Ryzen AI (پردازنده‌های دارای واحد NPU برای هوش مصنوعی محلی)



Instinct MI300X

Radeon RX 7900 XTX



Ryzen AI

وضعیت محدودیت صادرات محصولاتAMD درسال ۲۰۲۴ بصورت زیراست:

برخی از **تراشه‌های پیشرفته AMD** مخصوصاً مدل‌های مخصوص **هوش مصنوعی و ابرمحاسبات** تحت **محدودیت‌های صادراتی آمریکا** قرار دارند، اما محدودیت‌ها نسبت به NVIDIA کمتر است.

تراشه‌های AMD با محدودیت صادراتی

تراشه	محدودیت صادرات به چین/روس/ایران	دلیل محدودیت
Instinct MI300X	✓بله (ممنوع بدون مجوز)	قدرت محاسباتی بالا192GB HBM3 ، ۵٫۳TB/s
Instinct MI250X	✓بله	رقیب مستقیمA100/H100

دلیل محدودیت	محدودیت صادرات به چین/روس/ایران	تراشه
استفاده در ابررایانه‌ها	✅بله (مدل‌های پیشرفته)	**EPYC 9xx4 (Zen 4)**
نگرانی امنیتی	✅بله	**Ryzen 7045HX لپتاپ‌های نظامی**

🔷مقایسه با محدودیت‌های انویدیا

نمونه تراشه‌های ممنوعه	محدودیت‌ها	شرکت
حتی نسخه‌های ضعیف‌شده (H20, A800) نیاز به مجوز دارند	بسیار سخت‌گیرانه (H100, A100, H200)	**NVIDIA**
MI300X, MI250X, EPYCپیشرفته	متوسط (تمرکز بر مدل‌های نظامی/ابررایانه‌ها)	**AMD**

AMDهم مانند انویدیا، تراشه‌های خاصی برای بازار چین عرضه کرده است، مانند:

Instinct MI309(نسخه ضعیف‌شده MI300X) .

"China Special Edition" (EPYC با کاهش فرکانس و حافظه)

تحریم‌ها باعث شده که چین به سمت هواوی و تراشه‌های داخلی **مانند**Ascend 910B **برود.**

AMDهمچنان می‌تواند تراشه‌های مصرفی (Ryzen, Radeon) را به چین بفروشد.

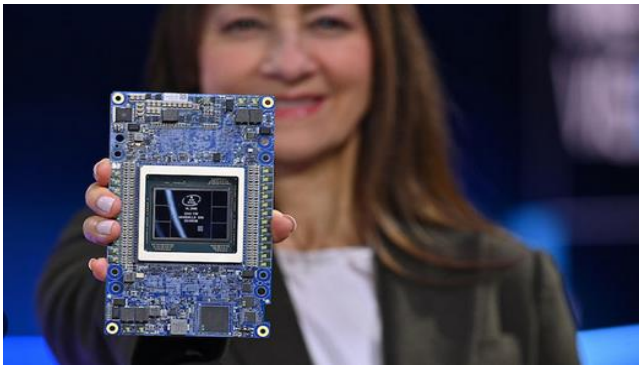
ایران و روسیه دسترسی بسیار محدودی به محصولات جدید AMD دارند.

۳. **Intel**

Intel در حال سرمایه‌گذاری گسترده در حوزه هوش مصنوعی و GPUهای مخصوص آن است.

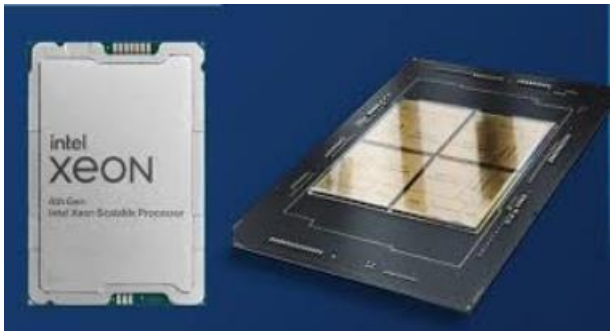
- محصولات کلیدی:

- Habana Gaudi 2 (برای آموزش مدل‌های هوش مصنوعی)



- Intel Arc GPU (برای پردازش گرافیکی و هوش مصنوعی)

- Xeon Sapphire Rapids (با قابلیت‌های شتاب دهنده هوش مصنوعی)



برخی از تراشه‌های هوش مصنوعی و پردازنده‌های پیشرفته اینتل **نیزتحت** محدودیت‌های صادراتی آمریکا **قرار دارند.**Habana Gaudi 2 برای صادرات به چین ،روسیه وایران نیازبه مجوزداردزیرا تقریبا رقیب **H100/A100 است.**Ponte Vecchio (Max Series GPU) برای چین محدودیت صادراتی داردزیرا درابررایانه ها وHPC مورداستفاده قرارمی گیرد. **مدل‌های خاص Xeon Phi** بخاطر کاربردهای آن درامورنظامی وتحقیقاتی برای هرسه کشور ایران روسیه چین ممنوعیت صادراتی دارد. Core Ultra (AI NPU) بدون محدودیت صادراتی است.

۴. **Googleبا تراشه‌های TPU**

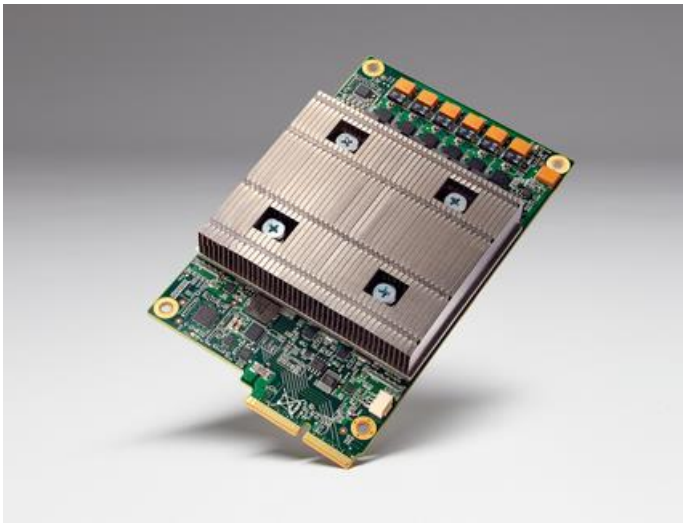
TPUیک **شتاب‌دهنده ویژه هوش مصنوعی** طراحی شده توسط گوگل است که به‌طور اختصاصی برای **اجرای مدل‌های مبتنی بر TensorFlow** بهینه شده است. این تراشه‌ها درمراکز داده ابری گوگل (Google Cloud)استفاده می‌شوند و برای **آموزش و استنتاج مدل‌های بزرگ** طراحی شده‌اند.

جدول نسل‌های مختلفTPU :

نسل	سال عرضه	معماری	کارایی کلیدی	کاربرد اصلی
TPU v1	2016	اختصاصی	92 TFLOPS (FP16)	AlphaGo
TPU v2	2017	اختصاصی	180 TFLOPS	مدل‌های ترنسفورمر اولیه
TPU v3	2018	اختصاصی	420 TFLOPS +خنک‌کننده مایع	BERT، GPT-2
TPU v4	2021	اختصاصی	275 TFLOPS (بهینه‌شده برای NLP)	GPT-3، مدل‌های چندوجهی
Edge TPU	2018	کم‌مصرف	4 TOPS (INT8)	دستگاه‌های لبه(IoT)

جدول مقایسه معماری TPU با GPU :

ویژگی	TPU	GPUمثلاً(NVIDIA A100)
هدف طراحی	بهینه برایTensorFlow	عمومی‌ترCUDA) ، (PyTorch
دقت محاسباتی	عمدتاًINT8/FP16	FP16/FP32/FP64
انعطاف‌پذیری	محدود (فقطTensorFlow)	بالا (چندفریمورک)
مصرف انرژی	بهینه‌تر برای بارهای خاص	مصرف بالاتر در مدل‌های مشابه
قیمت	ارزان‌تر در محاسبه/تسک	گران‌تر

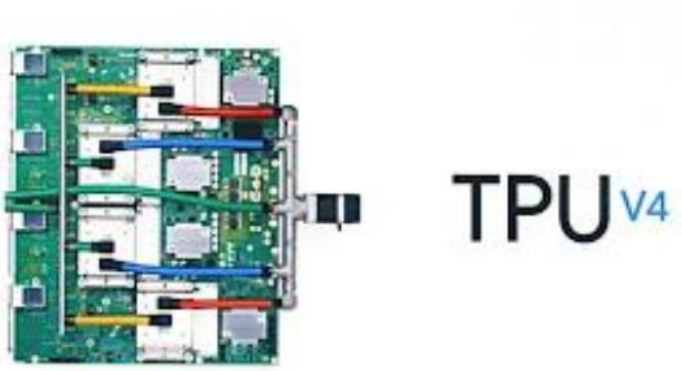


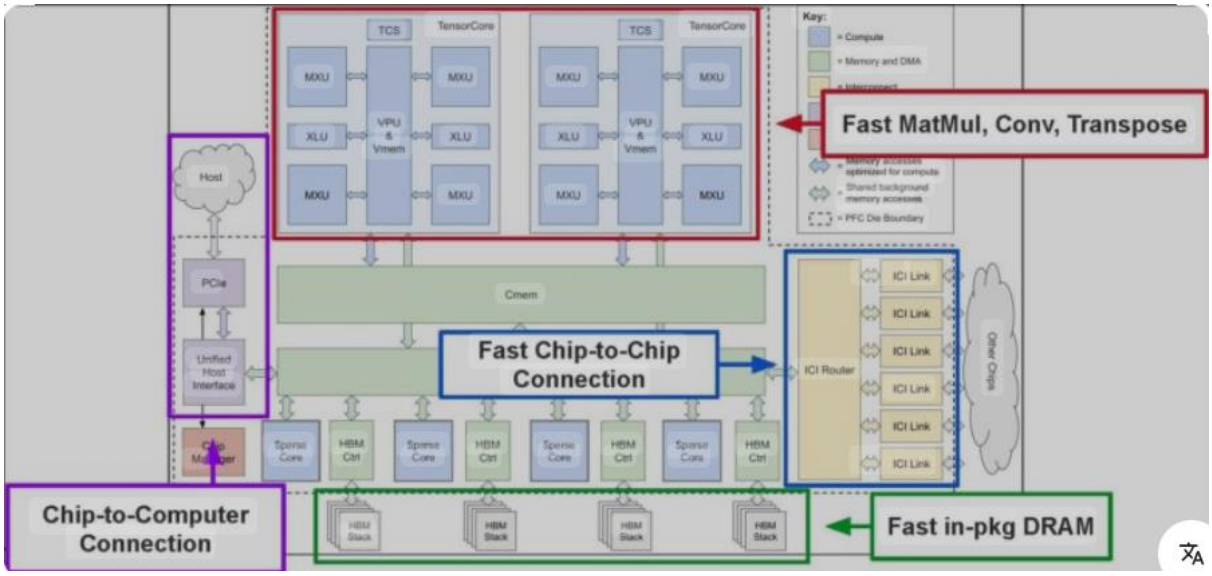
TPU

کاربردهای کلیدیTPU عبارتنداز:

۱. **آموزش مدل‌های بزرگ زبانی(LLM)**

گوگل از TPUv4 برای آموزش **Gemini** و **BERT** استفاده کرده است.





۲. **پردازش تصویر در مقیاس بزرگ**

مدل‌های بینایی ماشینGoogle Photos

۳. **یادگیری تقویتی(RL)**

پروژه‌هایی مانند AlphaGo و.AlphaFold

۴. **پردازش زبان طبیعی(NLP)**

ترجمه ماشینی.Google Translate

امکان خرید سخت‌افزار TPU وجود ندارد و فقط به صورت سرویس ابری ارائه می‌شود.

هزینه استفاده:

TPU v2: ~$4.50/hour

TPU v4: ~$6.50/hour

برخی نسخه‌های پیشرفته TPU در چین وروسیه در دسترس نیستند.

TPU مهم است زیرا برای TensorFlow بهینه سازی شده و مدل‌ها رابا **کارایی بالاتر و هزینه کمتر** نسبت بهGPU اجرا می کند.امکان استفاده از **هزاران هسته TPU** به صورت موازی وجوددارد(مثلاً در **Podهای TPUv4 ، ۴۰۹۶ هسته** وجوددارد)**بعلاوه** گوگل هر ۲-۳ سال ، نسل جدیدی از TPU را عرضه می‌کند .TPUگوگل یک **راه‌حل قدرتمند و مقرون‌به‌صرفه** برای اجرای مدل‌های هوش مصنوعی در مقیاس

بزرگ است، اما انعطاف‌پذیری کمتری نسبت به GPU دارد. اگر از **TensorFlow/JAX** استفاده می‌کنید، TPU می‌تواند انتخاب ایده‌آلی باشد!

آیا می‌خواهید بدانید چگونه از TPU در Google Cloud استفاده کنید؟

راهنمای گام‌به‌گام استفاده از TPU در Google Cloud

پیش‌نیازها

۱. حساب Google Cloud (GCP)

- اگر ندارید، از [صفحه ثبت‌نام GCP](https://cloud.google.com/) یک حساب ایجاد کنید (اعتبار رایگان ۳۰۰ دلاری برای کاربران جدید).

۲. پروژه GCP

- در کنسول GCP، یک پروژه جدید بسازید.

۳. فعال‌سازی سرویس‌های موردنیاز

- Compute Engine API و Cloud TPU API را از بخش API & Services فعال کنید.

روش‌های استفاده از TPU

۱. استفاده از سرویس‌های مدیریت‌شده (راحت‌تراست.)

- Google Colab رایگان برای TPU v2/v3

- به colab.research.google.com بروید.

- محیط جدید بسازید و از کد زیر استفاده کنید:

python Copy Download

```python
import tensorflow as tf
resolver = tf.distribute.cluster_resolver.TPUClusterResolver(tpu='')
tf.config.experimental_connect_to_cluster(resolver)
tf.tpu.experimental.initialize_tpu_system(resolver)
print("All devices: ", tf.config.list_logical_devices('TPU'))
```

توجه کنیدکه TPU فقط در مناطق خاصی مانند "us-central1" یا "europe-west4" در دسترس است.

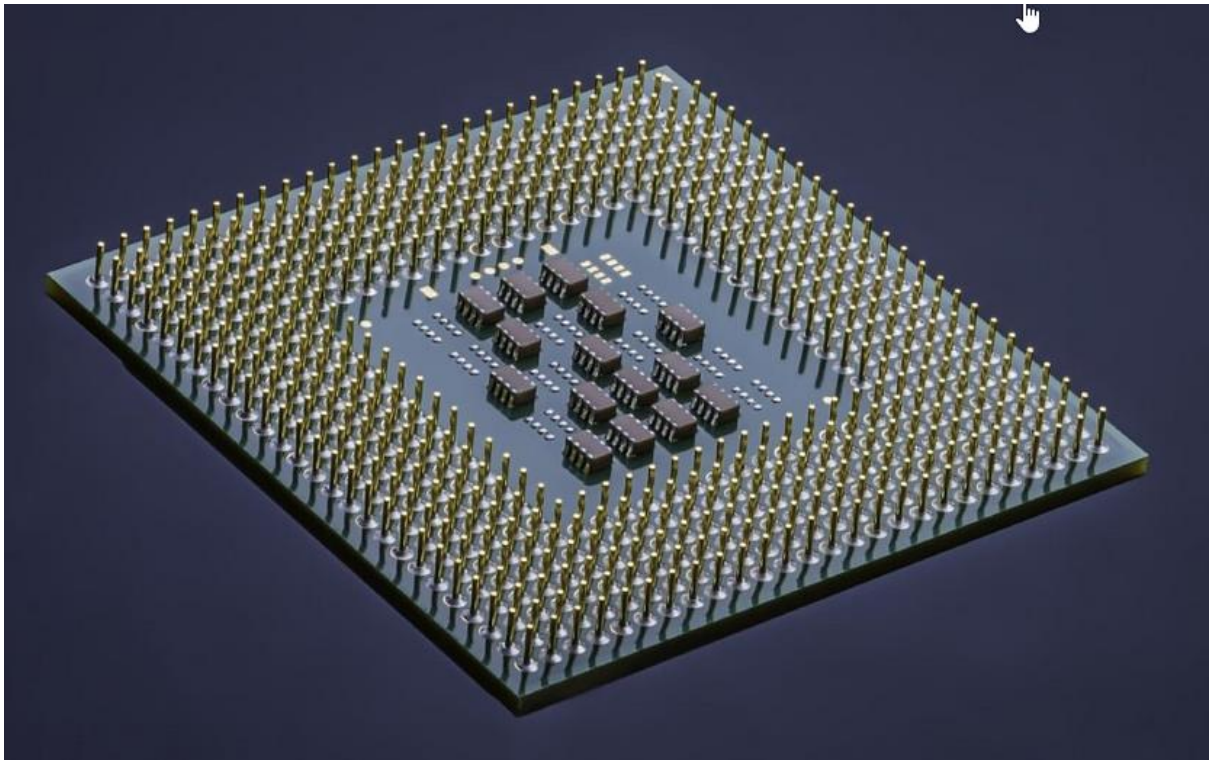
۵. Amazon با تراشه‌های AWS Trainium & Inferentia

شرکت آمازون (Amazon)نیز با عرضه تراشه‌های اختصاصی خود برای هوش مصنوعی و یادگیری ماشین، به رقابت با شرکت‌هایی مانند NVIDIA ، Google و Huawei پرداخته است. دو تراشه اصلی AWS در این زمینه عبارتند از:

۱. AWS Trainium

- هدف: آموزش (Trainingمدل‌های بزرگ هوش مصنوعی با کارایی بالا و هزینه کمتراست).
- ویژگی‌های کلیدی:
- بهینه‌شده برای مدل‌های عظیم (مانند مدل‌های زبانی مانند GPT-3 و مدل‌های توصیه‌گر).
- پشتیبانی از دقت ترکیبی (Mixed Precision – FP16, BF16, INT8) برای افزایش سرعت.
- یکپارچه با Amazon SageMakerو EC2(مثلاً نمونه‌های Trn1)
- مقیاس‌پذیری بالابا قابلیت اتصال چندین تراشه برای آموزش مدل‌های بسیار بزرگ.
- مقایسه با رقبا:
- رقیب مستقیم NVIDIA A100/H100 و Google TPU v4 در بخش آموزش مدل‌ها.
- ادعای ۴۰٪ صرفه‌جویی در هزینه نسبت به GPUهای مشابه.

۲. AWS Inferentia (و Inferentia2)



- هدف: استنتاج (Inference) سریع و مقرون‌به‌صرفه مدل‌های هوش مصنوعی.

- ویژگی‌های کلیدی:

- طراحی شده برای کاهش تأخیر (Latency)و افزایش توان عملیاتی (Throughput).

- پشتیبانی از مدل‌های پیچیده مانند تبدیل‌گرها (Transformers) و تشخیص تصویر.

- Inferentia2 (نسخه جدیدتر) بهبودهای چشمگیری در کارایی و مصرف انرژی دارد.

- یکپارچه با AWS SageMaker, EC2 Inf1/Inf2 instances و Lambda .

- مقایسه با رقبا:

- رقیب NVIDIA T4, A10G و Google TPU برای استنتاج .

- ادعای کاهش ۷۰٪ هزینه استنتاج در مقایسه با GPUها.

مزایای کلیدی AWS Trainium & Inferentia :

بهینه‌شده برای AWS: یکپارچه با خدمات ابری Amazon (مانند SageMaker, EC2)

مقرون‌به‌صرفه: کاهش هزینه‌های عملیاتی نسبت به GPUهای سنتی.

مقیاس‌پذیری: امکان آموزش و استنتاج مدل‌های بسیار بزرگ.

پشتیبانی از چارچوب‌های رایج: مانند TensorFlow, PyTorch, MXNet .

رقابت با Huawei Ascend 910B و دیگر تراشه‌ها

- در بخش آموزش:

Trainium با Huawei Ascend 910B و NVIDIA H100 رقابت می کند.تمرکز Trainium بر کاهش هزینه‌های AWS برای مشتریانش است.

- در بخش استنتاج:

Inferentia2 با Google TPU v4 و NVIDIA L4/T4 رقابت می کند.

Inferentia برای استنتاج مدل‌های بزرگ (مانند ChatGPT-like models) بسیار کارآمد است.

آمازون با Trainium و Inferentia به دنبال کاهش وابستگی به سخت‌افزارهای خارجی

(مثل NVIDIA) و ارائه راه‌حل‌های مقرون‌به‌صرفه برای هوش مصنوعی در اکوسیستم AWS است. این تراشه‌ها برای شرکت‌هایی که از ابر AWS استفاده می‌کنند، گزینه‌های جذابی هستند.

دسترسی به سرویس‌های هوش مصنوعی آمازون (AWS AI/ML) ممکن است محدودیت‌هایی داشته باشد که بستگی به عوامل مختلفی دارد، از جمله محدودیت‌های جغرافیایی و تحریم‌ها ،محدودیت‌های حساب کاربری AWS.بعلاوه برای دسترسی به برخی قابلیت‌ها (مثل تراشه‌های اختصاصی Trainium/Inferentia)، AWS ممکن است تأیید هویت سخت‌گیرانه‌تری اعمال کند.

اگر هزینه مصرفی از حد معینی عبور کند، AWS ممکن است دسترسی را موقتاً محدود کند.

استفاده ازهوش مصنوعی آمازون ممکن است بامحدودیت‌های فنی همراه باشدبعلاوه برخی ازمحدودیت های قانونی وامنیتی به دلیل مسائل حریم خصوصی نیز وجوددارد.

راه‌حل‌های احتمالی برای دور زدن محدودیت‌ها (با ریسک)

- استفاده از حساب AWS در مناطق غیرتحریمی (با پرداخت ارزی و تأیید هویت).

- استفاده از سرورهای واسط (Proxy/VPN اما AWS ممکن است حساب را مسدود کند).

- مهاجرت به سرویس‌های جایگزین مثل Google Cloud TPUیا Microsoft Azure AI (اگر در دسترس باشند).

دسترسی به هوش مصنوعی AWS برای همه کاربران یکسان نیست و بستگی به موقعیت جغرافیایی، نوع حساب AWS و قوانین محلی دارد. اگر در کشوری تحت تحریم هستید، احتمالاً دسترسی مستقیم به تراشه‌هایی مثل Trainium/Inferentiaیا سرویس‌هایی مثل Bedrock را نخواهید داشت.

۶. Microsoft همکاری با AMD و طراحی تراشه‌های اختصاصی

مایکروسافت نیز مانند سایر غول‌های فناوری، سرمایه‌گذاری قابل توجهی در توسعه تراشه‌های اختصاصی هوش مصنوعی انجام داده است. اگرچه مایکروسافت به اندازه شرکت‌هایی مانند NVIDIA، گوگل (TPU) یا آمازون (Trainium/Inferentia) در طراحی تراشه‌های سفارشی پیشرو نیست، اما همکاری‌های استراتژیک و پروژه‌های داخلی مهمی دارد. در اینجا مهم‌ترین اقدامات مایکروسافت در زمینه سخت‌افزار هوش مصنوعی را بررسی می‌کنیم:

۱. همکاری با NVIDIA و استفاده از تراشه‌های H100/A100

مایکروسافت عمدتاً برای زیرساخت ابر Azure خود از تراشه‌های NVIDIA مانند H100، A100 و L40S استفاده می‌کند. همچنین، در پروژه‌هایی مانند Microsoft Copilot و Azure OpenAI Serviceاین تراشه‌ها نقش کلیدی دارند.

۲. تراشه‌های اختصاصی مایکروسافت (Microsoft Maia & Cobalt)

در نوامبر ۲۰۲۳، مایکروسافت از دو تراشه جدید رونمایی کرد که نشان‌دهنده حرکت به سمت طراحی سخت‌افزارهای اختصاصی هوش مصنوعی است:

Microsoft Maia 100

- هدف: بهینه‌شده برای اجرای مدل‌های بزرگ هوش مصنوعی LLM (مانند ChatGPT و Microsoft Copilot) .

Microsoft Cobalt 100

- هدف: یک پردازنده سرور (CPU) بهینه‌شده برای بارهای کاری ابری و هوش مصنوعی.

۳. همکاری با AMD و Intel

- مایکروسافت با AMD روی تراشه‌های Instinct MI300X همکاری دارد که برای هوش مصنوعی طراحی شده‌اند.

- Intel از تراشه‌های Habana Gaudi (برای آموزش مدل‌ها) در Azure استفاده می‌کند.

۴. پروژه‌های تحقیقاتی (مثل Project Olympus)

مایکروسافت در حال تحقیق روی "سخت‌افزارهای ماژولار" برای دیتاسنترهاست تا انعطاف‌پذیری بیشتری در اجرای مدل‌های هوش مصنوعی داشته باشد.

۵. Azure Boost (بهینه‌سازی شبکه و ذخیره‌سازی برای هوش مصنوعی)

این فناوری مستقیماً یک تراشه نیست، اما با افزایش کارایی I/O در سرورهای Azure، پردازش مدل‌های هوش مصنوعی را سریع‌تر می‌کند.

اگر به دنبال استفاده از سخت‌افزارهای مایکروسافت برای هوش مصنوعی هستید، Azure بهترین گزینه است، چون به‌روزترین تراشه‌ها (حتی Maia و Cobalt) ابتدا در آن راه‌اندازی می‌شوند.

۷. **Qualcomm**

Snapdragon X Elite و Snapdragon 8 Gen 3تراشه‌های هوش مصنوعی کوالکام برای عصر جدید محاسبات AI

شرکت کوالکام (Qualcomm) با سری تراشه‌های Snapdragon خود، به‌ویژه در نسل‌های اخیر، تحولی در پردازش هوش مصنوعی در دستگاه‌های همراه، لپ‌تاپ‌ها ،اینترنت اشیا و سایر گجت‌های هوشمند ایجاد کرده است. در اینجا به بررسی مهم‌ترین تراشه‌های هوش مصنوعی کوالکام می‌پردازیم:

Snapdragon X Elite برای لپ‌تاپ‌های AI

- پردازنده ۱۲ هسته‌ای :معماری Oryon (بر پایه ARM)با فرآیند ۴ نانومتری
- موتورAI : Hexagon NPU با قدرت پردازش ۴۵ TOPS (تریلیون عملیات در ثانیه)
- کاربردها:
- اجرای مدل‌های LLM (مانند ChatGPT محلی)
- پردازش تصویر و ویدئوی هوشمند (مثل حذف نویز و بهبود کیفیت)
- بهینه‌سازی باتری با هوش مصنوعی

Snapdragon 8 Gen 3 (برای گوشی‌های هوشمند)

- موتورHexagon NPU : AI با ۶۰٪ بهبود عملکرد نسبت به نسل قبل

- قابلیت‌های کلیدی:

- اجرای مدل‌های تولید محتوا (مثل Stable Diffusion) مستقیماً روی گوشی

- پردازش تصویر چند فریمی (HDRبا AI)

- تشخیص صدا و زبان طبیعی پیشرفته (مثل دستیارهای صوتی هوشمندتر)

- دستگاه‌های شاخص: سامسونگ گلکسی S24 ، شیائومی ۱۴، OnePlus 12

سامسونگ گلکسی S24

OnePlus 12

Snapdragon 7 Gen 3,6 و Gen 1(تراشه‌های میان‌رده و اقتصادی)

- تمرکز: هوش مصنوعی در دستگاه‌های مقرون‌به‌صرفه

- ویژگی‌ها:

- پردازش تصویر مبتنی بر AI (مثل Portrait Mode)

- بهینه‌سازی گیمینگ با یادگیری ماشین

Snapdragon 8s Gen 3 (نسخه سبک‌تر پرچمدار)

- هدف: ارائه قابلیت‌های هوش مصنوعی پرچمدار با قیمت پایین‌تر

- تفاوت با ۸ : در Gen 3 قدرت NPU کمی کاهش یافته، اما همچنان از مدل‌های محلی تا ۱۰ میلیارد پارامترپشتیبانی می‌کند.

مقایسه با رقبا

تراشه	قدرت NPU (TOPS)	کاربرد اصلی	رقبای مستقیم
Snapdragon X Elite	45 TOPS	لپ‌تاپ‌های AI	Apple M4, Intel Core Ultra
Snapdragon 8 Gen 3	~50 TOPS	گوشی‌های هوشمند پرچمدار	Apple A17 Pro, Tensor G3
Snapdragon 7 Gen 3	~20 TOPS	گوشی‌های میان‌رده	Dimensity 7200, Exynos 1380

کاربردهای عملی هوش مصنوعی در Snapdragon

مدل‌های زبانی محلی (LLM): اجرای ChatGPT-like models بدون نیاز به اتصال به اینترنت.

عکاسی هوشمند(AI): بهبود عکس‌ها با حذف نویز، سوپررزولوشن.

رابط کاربری هوشمند: تشخیص صدا، ترجمه همزمان و پیش‌بینی متن.

گیمینگ هوشمند: افزایش نرخ فریم و کاهش تاخیر با یادگیری ماشین.

تراشه‌های Snapdragon، به‌ویژه X Elite8 و Gen 3 کوالکام را به یکی از پیشگامان هوش مصنوعی در لبه (Edge AI) تبدیل کرده‌اند. اگر به دنبال دستگاهی با قابلیت‌های هوش مصنوعی بدون وابستگی به ابرهستید، محصولات مجهز به این تراشه‌ها گزینه‌های عالی هستند.

۸. IBM (با پردازنده‌های Power10 و شتاب‌دهنده‌های AI

- برای کاربردهای هوش مصنوعی در سازمان‌های بزرگ و ابررایانه‌ها.

۹. **شرکت‌های نوظهور و تخصصی:**

- Cerebras Systems (با تراشه‌های عظیم Wafer-Scale برای هوش مصنوعی)
- Graphcore (تراشه‌های IPU برای هوش مصنوعی)
- SambaNova Systems (پلتفرم‌های هوش مصنوعی مبتنی بر تراشه‌های سفارشی)

خلاصه:

- NVIDIAو AMD دو غول اصلی در تولید GPUهای عمومی و هوش مصنوعی هستند.
- (TPU) Googleو Amazon (Trainium/Inferentia)تراشه‌های تخصصی برای ابرهای خود دارند.
- Intel و Qualcomm نیز در حال رقابت برای حضور پررنگ‌تر در بازار هستند.
- شرکت‌های نوپایی مانند Cerebras و Graphcore نیز فناوری‌های نوینی ارائه می‌دهند.