CUDA程序设计

Lec 1: Course Overview

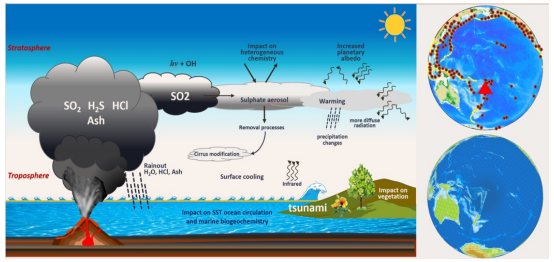
陈俊仕

2024 Fall

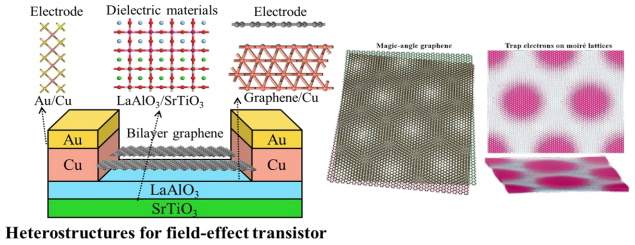
计算机科学与技术学院
School of Computer Science and Technology

自我介绍

- 科研方向
  - 高性能计算、并行编程模型和运行时系统
  - 国产超算系统上科学工程应用的大规模并行算法设计和优化
  - 涉及领域
    - 第一性原理材料模拟、计算生物学、流体力学、大气模式、等离子体模拟



海底火山喷发耦合模拟



第一性原理电子结构计算

EAST

CFETR

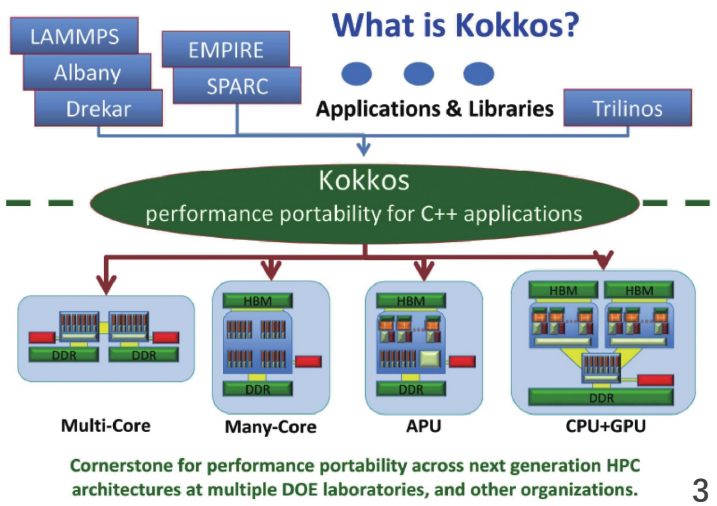
托克马克全装置等离子体模拟

课程目标

- 了解GPU的处理器架构
  - GPU与CPU架构的差异
- 掌握基于CUDA的并行编程模型
  - 其它编程模型（not present）
    - HIP、OpenCL、Vulkan、OpenACC、SYCL、Kokkos…
- 培养GPU并行计算系统的分析能力、编程能力
  - 性能热点分析、性能优化方法

NVIDIA CUDA

AMD ROCm



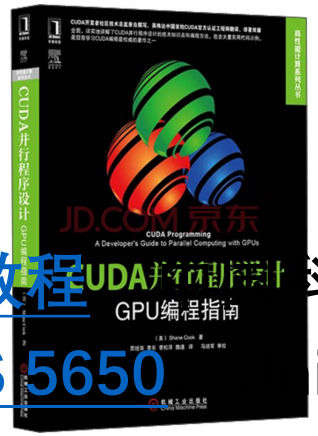
OpenCL™ Vulkan® OpenACC More Science, Less Programming SYCL™

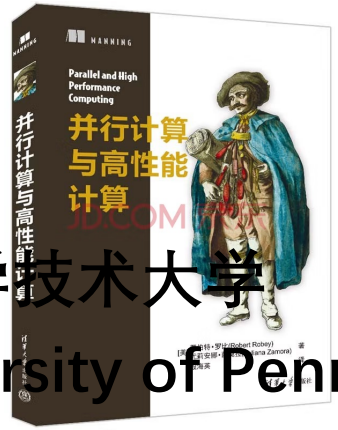
KHRONOS GROUP

VRM CONSORTIUM

参考教材

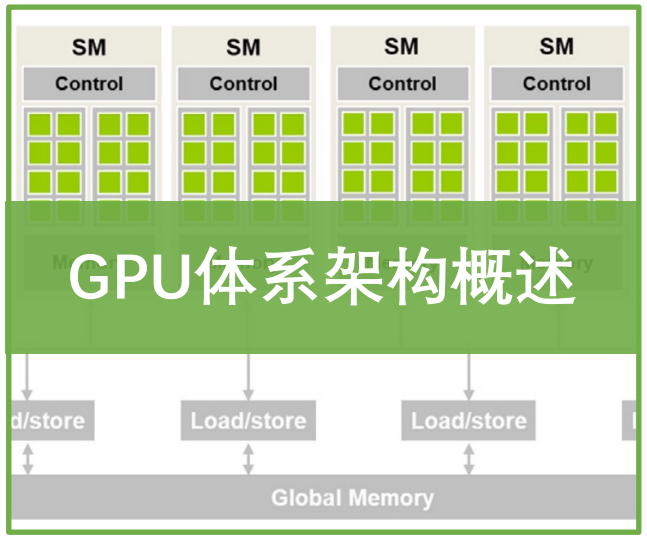
- 参考材料
  1. Programming Massively Parallel Processors: A Hands-on Approach, 4th Edition，Wenmei W. Hwu, etc.
  2. CUDA Programming Guide， NVIDIA Corp.
  3. CUDA Best Practice Guide, NVIDIA Corp.



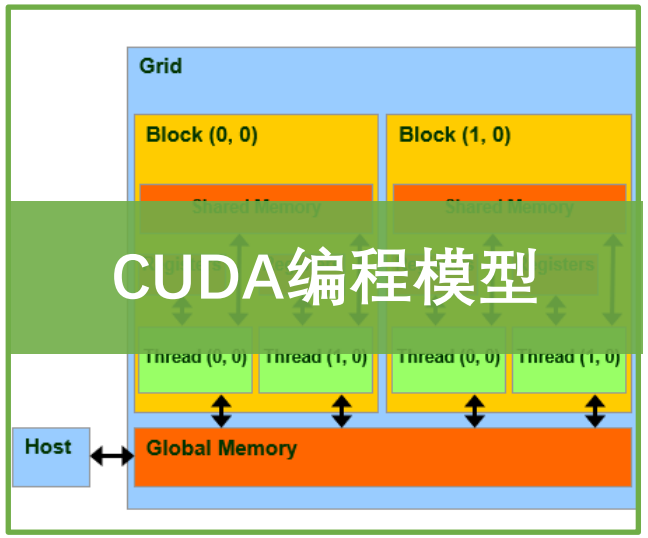


- 其TA课程
  - 周斌，NVIDIA CUDA初级教程，科学技术大学
  - Shehzan Mohammed，CIS 5650，University of Pennsylvania
  - CMPS 297S/396AA, American University of Beirut

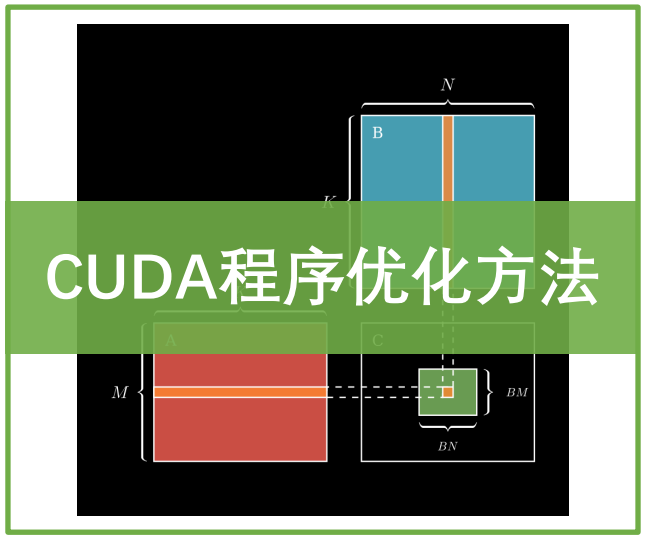
课程内容



GPU体系架构概述



CUDA编程模型



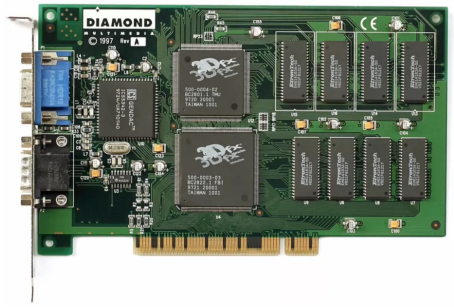
CUDA程序优化方法

GPU – Graphics Processing Unit

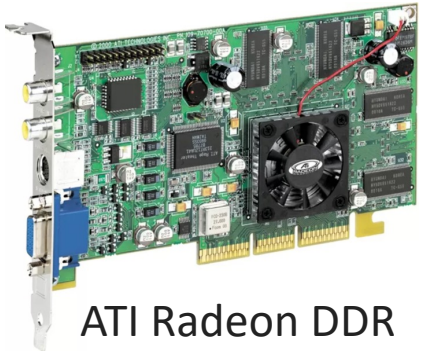
- The Early Days of 3D Consumer Graphics
  - 1976 - 1995
- 3Dfx Voodoo: The Game-changer
  - 1995 - 1999
- The Nvidia vs. ATI Era Begins
  - 2000 - 2006
- The Modern GPU: Stream Processing Units
  - 2006 - 2013
- Pushing GPU Technology Into New Territory
  - 2013-Now

https://www.techspot.com/article/650-history-of-the-gpu/

Atari 2600 (1977)

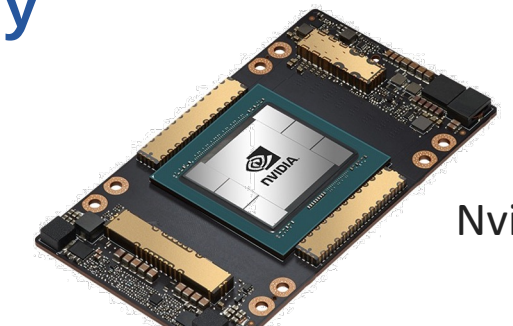


Diamond 3dfx Voodoo1



ATI Radeon DDR

Nvidia GeForce GTX Titan



Nvidia HGX A100

GPGPU - General Purpose GPU

- **2007年以前，GPU平台上的编程方法**
  - 图形APIs，如OpenGL、Direct3D等

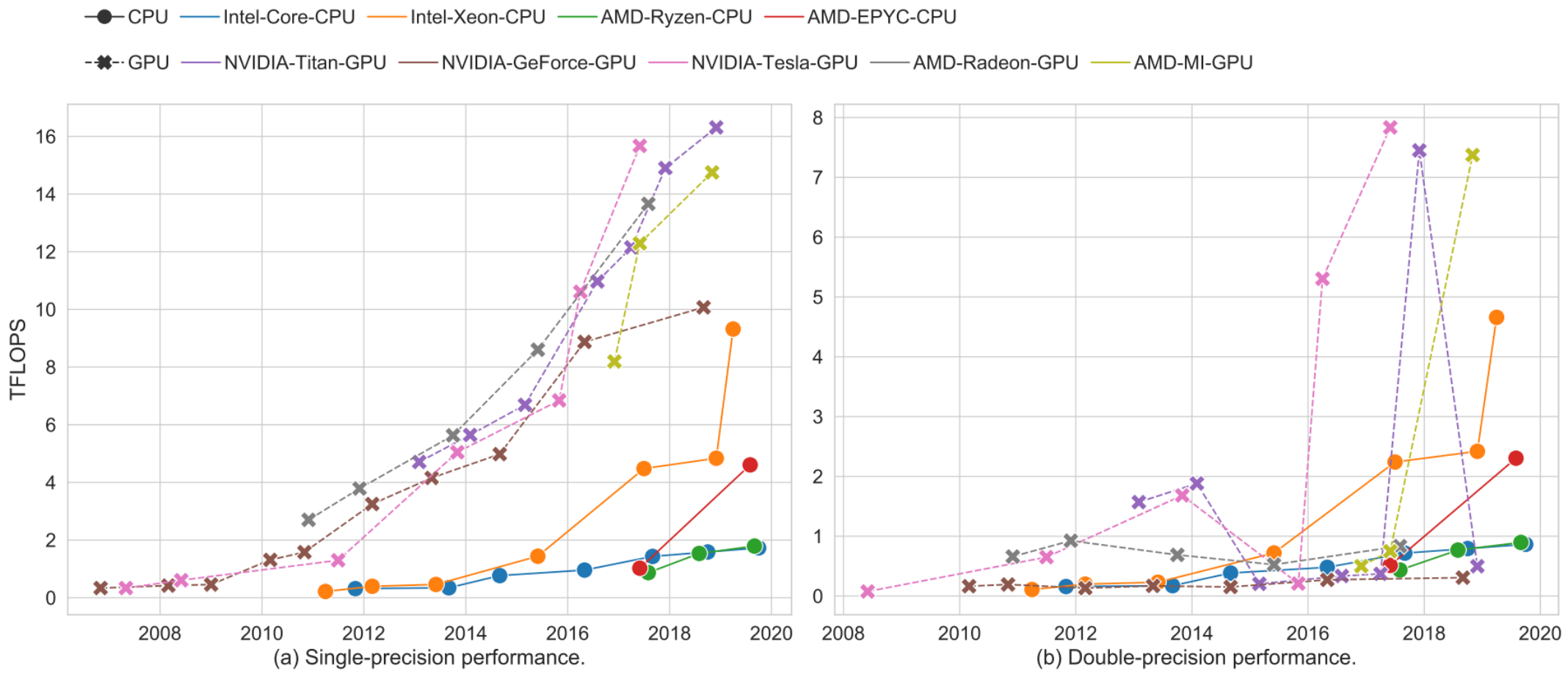
```
sampler mySamp;
Texture2D<float3> myTex;
float3 lightDir;

float4 diffuseShader(float3 norm,float2 uv) {
  float3 kd = myTex.Sample(mySamp, uv);
  kd *= clamp(dot(lightDir, norm), 0.0, 1.0);
  return float4(kd, 1.0);
}
```

- **2007年，NVIDIA推出CUDA**
  - **C**ompute **U**nified **D**evices **A**rchitectured，统一计算架构
  - 面向GPU架构的通用编程接口
  - 在C语言的基础上针对GPU架构的扩展

```
__global__ void vecAddKernel(float *A, float *B, float *C, int n) {
  int i = threadIdx.x + blockDim.x * blockIdx.x;
  if (i < n) {
    C[i] = A[i] + B[i];
  }
}
```

CPU vs. GPU

- single/double-precision performance of CPUs and GPUs



Summarizing CPU and GPU Design Trends with Product Data, arXiv:1911.11313

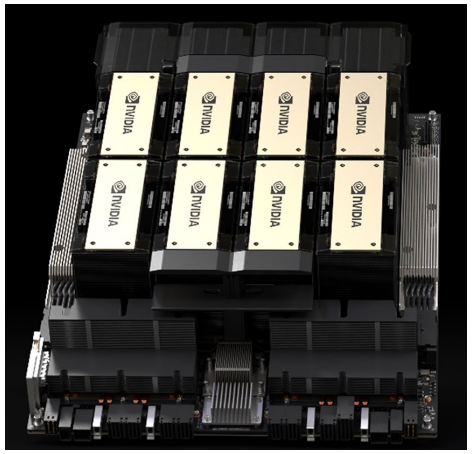
CPU vs. GPU

Intel i9 14900K

- Freq. 6GHz max
- 8+16=32 cores
- FP64: 1.23 TFLOPS
- Mem. BW: 89.6 GB/s

NVIDIA H200

- Freq. 1.83GHz
- 16896 CUDA cores
- FP64: 34 TFLOPS, 67 TFLOPS (TC)
- TF32: 989 TFLOPS (TC)
- Mem. BW: 4.8TB/s (HBM3e)

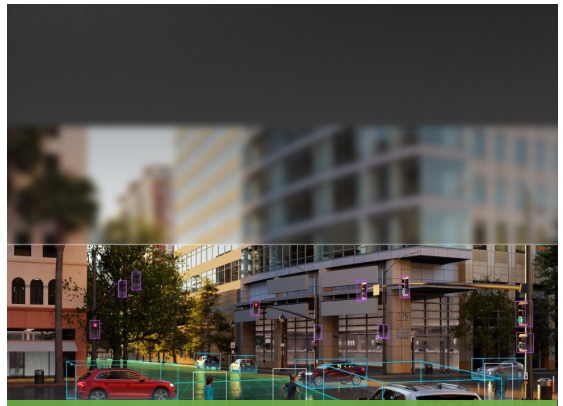


https://www.intel.com/content/dam/support/us/en/documents/processors/APP-for-Intel-Core-Processors.pdf

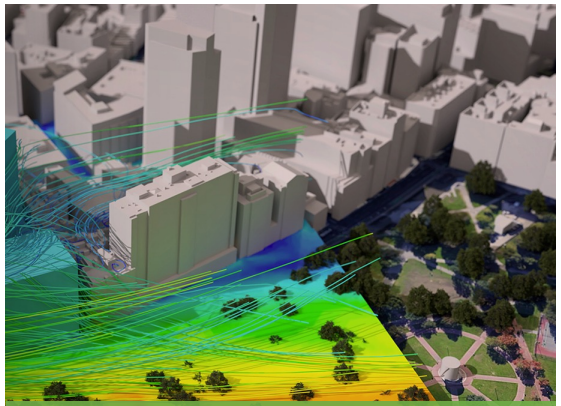
Common Uses of GPU Parallel Computing

Computer Graphics, Animation, and Games

High-performance computing

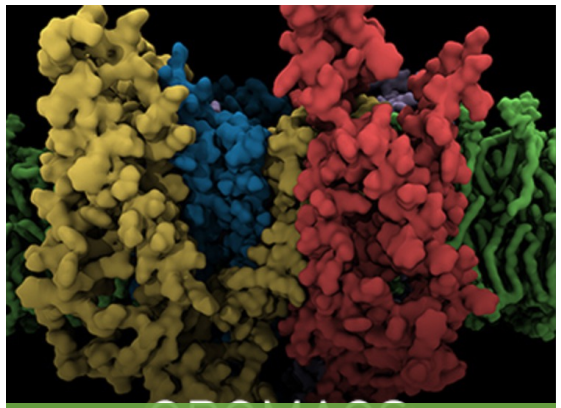


Machine Learning and AI

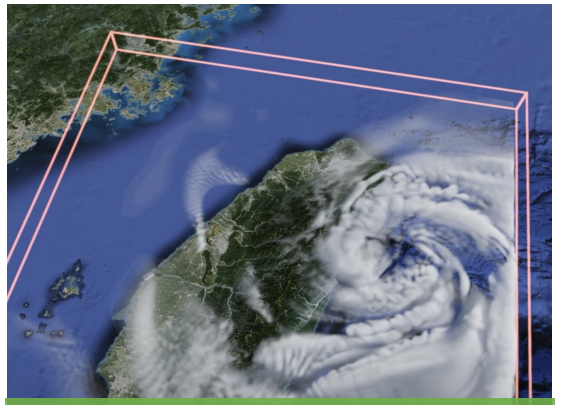


Computer Aided Modelling

Data Mining



Scientific Computing



Simulation

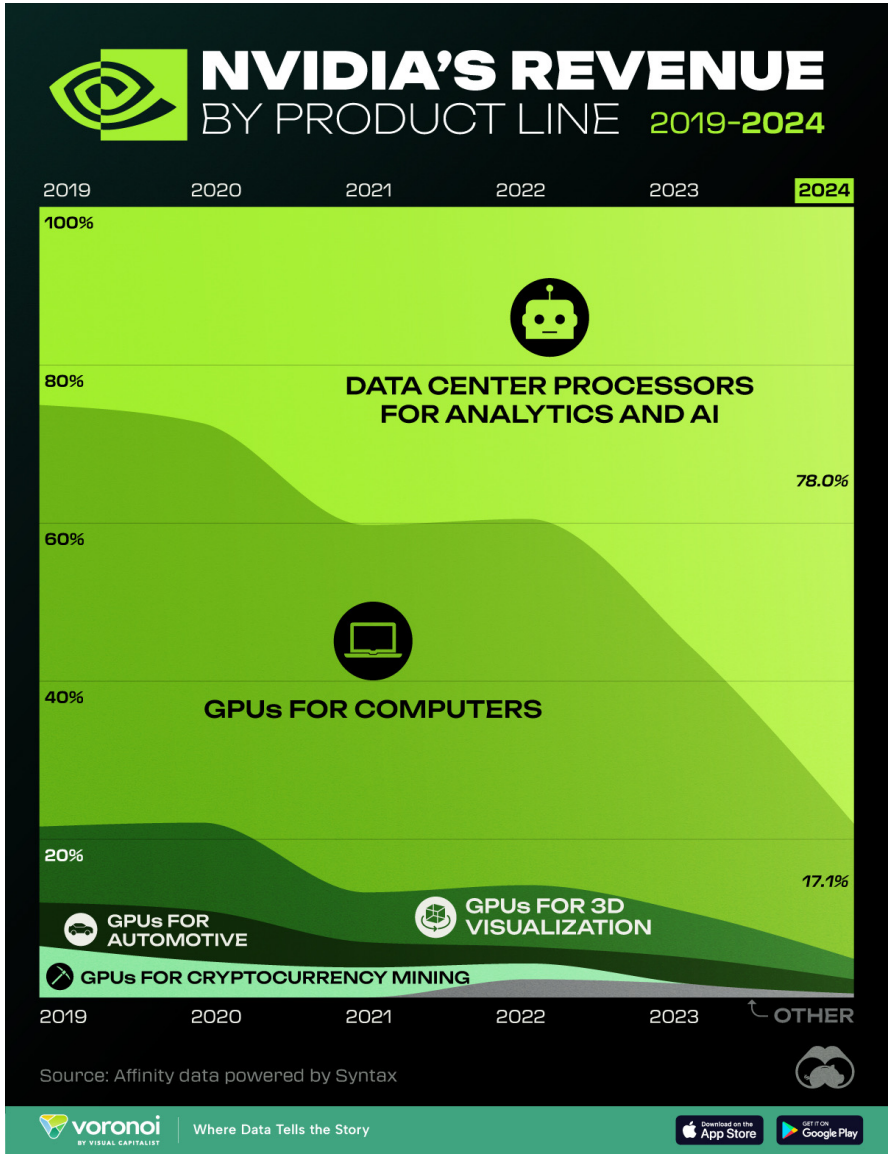
Top supercomputers in TOP500 List (2024)

Rank	System		Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	Frontier	US, AMD EPYC, AMD MI250X	8,699,904	1,206.00	1,714.81	22,786
2	Aurora	US, Intel Xeon, Intel GPU	9,264,128	1,012.00	1,980.01	38,698
3	Eagle	US, Intel Xeon, NV H100	2,073,600	561.20	846.84	
4	Fugaku	Japan, A64FX	7,630,848	442.01	537.21	29,899
5	LUMI	Finland, AMD EPYC, AMD MI250X	2,752,704	379.70	531.51	7,107
6	Alps	Switzerland, Cray, NV GH200	1,305,600	270.00	353.75	5,194
7	Leonardo	Italy, Intel Xeon, NV A100	1,824,768	241.20	306.31	7,494
8	MareNostrum 5 ACC	Spain, Intel Xeon, NV H100	663,040	175.30	249.44	4,159
9	Summit	US, IBM Power9, NV GV100	2,414,592	148.60	200.79	10,096
10	Eos NVIDIA DGX SuperPOD	US, Intel Xeon, NV H100	485,888	121.40	188.65	

Top supercomputers in Green500 List (2024)

Rank	TOP500 Rank	System		Cores	Rmax (PFlop/s)	Power (kW)	Energy Eff. (GFlops/watts)
1	189	JEDI	Germany，NV GH200 Superchip	19,584	4.50	67	72.733
2	128	Isambard-AI phase 1	UK, NV GH200 Superchip	34,272	7.42	117	68.835
3	55	Helios GPU	Poland, NV GH200 Superchip	89,760	19.14	317	66.948
4	328	Henri	US, NV H100 80GB PCIe	8,288	2.88	44	65.396
5	71	preAlps	Switzerland, NV GH200 Superchip	81,600	15.47	240	64.381
6	299	HoreKa-Teal	Germany, NV H100 94GB SXM5	13,616	3.12	50	62.964
7	54	Frontier TDS	US, AMD Instinct MI250X	120,832	19.20	309	62.684
8	11	Venado	US, NV GH200 Superchip	481,440	98.51	1,662	59.287
9	20	Adastra	France, AMD Instinct MI250X	319,072	46.10	921	58.021
10	28	Setonix – GPU	Australia, AMD Instinct MI250X	181,248	27.16	477	56.983

NVIDIA’s Revenue (2019-2024)



Product Line	2024	2023	2022	2021	2020	2019
Data Center Processors for Analytics and AI	78.0%	55.6%	39.4%	40.2%	27.3%	25.0%
GPUs for Computers	17.1%	33.6%	46.3%	46.5%	50.5%	53.3%
GPUs for 3D Visualization	2.6%	5.7%	7.8%	6.3%	11.1%	9.6%
GPUs for Automotive	1.8%	3.3%	2.1%	3.2%	6.4%	5.5%
GPUs for Cryptocurrency Mining	0.0%	0.0%	2.0%	3.8%	4.6%	6.5%
Other	0.5%	1.7%	2.3%	0.0%	0.0%	0.0%

https://www.visualcapitalist.com/nvidia-revenue-by-product-line/