Seminário – Nvidia

Universidade Federal de Ouro Preto (UFOP) - Departamento de Computação (DECOM) - BCC263 - Arquitetura de Computadores

Docente: Ricardo Augusto Rabelo Oliveira
Discentes: Felipe Braz, Lucas Chagas, Matheus Peixoto, Nicolas Mendes, Pedro Henrique Oliveira e Pedro Morais.

- Histórico e posicionamento no mercado
- Produtos (Mercado Industrial)
- Produtos (Mercado Aeroespacial)
- Produtos (Mercado Militar)
- Mercado
- Pergunta proposta

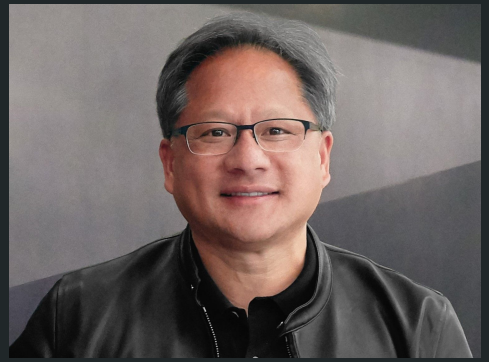
Histórico e posicionamento no mercado

História da empresa

Criada em 1993 por Jensen Huang, Chris Malachowsky e Curtis Priem

Foco em videogames e aceleração de processamento.

Primeiro projeto para a SGS-Thomson Microelectronics.



Jensen Huang

Curtis Priem

Chris Malachowsky

História da empresa

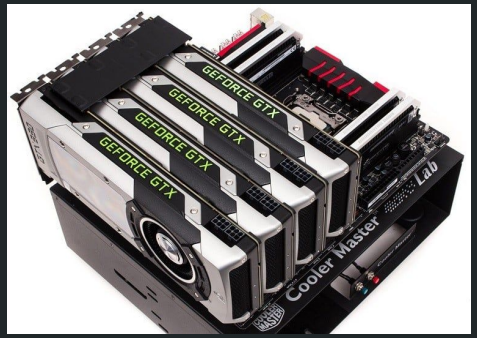
Lançam, em 1997, o NV1, que executava jogos de SEGA Saturno.

Lançamento da linha RIVA com compatibilidade com DirectX.

Em 1999, lançam a linha GeForce e QUADRA.

No início dos anos 2000 fecham uma parceria com a NASA.

Em 2004 lançam o suporte a SLI e o CUDA.



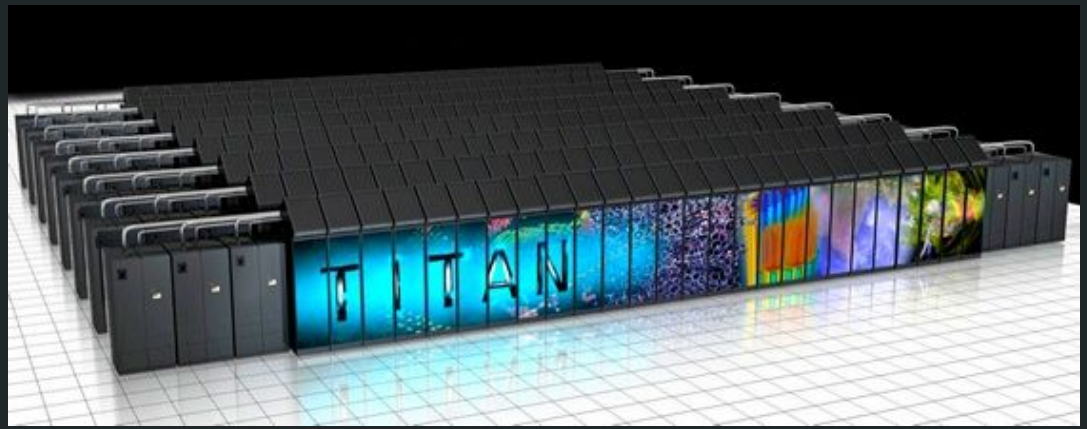
NVIDIA
CUDA

História da empresa

Lançamento da linha Tesla, usada no supercomputador Titan.

A partir de 2011, parceria com empresas automotivas para produzir os chips dos carros.

Em 2016, lançamento dos servidores e computadores DGX-1, possibilitando a criação do ChatGPT.



Supercomputador Titan

DGX-1

Posicionamento no mercado

Divisão em três grandes grupos:

- **Videogames**: Receita de 2,24 bilhões de dólares. Último lançamento sendo a linha RTX 4000
- **Data Center:** Receita de 4,2 bilhões de dólares. Parceria com o Google para usarem a NVIDIA L4 Tensor Core GPU no Google Cloud e outras parcerias para venda de serviços baseados na NVIDIA H100 Tensor Core GPU.
- **Setor automotivo:** Receita de 296 milhões de dólares. Parceria com a BYD para usarem a NVIDIA DRIVE Orin.

Obs: Dados retirados do segundo balanço fiscal da NVIDIA no segundo trimestre de 2023

Produtos - Mercado Industrial

Mercado Industrial

A NVIDIA tem desempenhado um papel extremamente importante no cenário industrial para transformá-lo, através da integração de inteligência artificial e aprendizado de máquina.Tal integração tem otimizado diversos processos e melhorado a segurança em ambientes industriais.

A causa da grande influência da NVIDIA deve-se a capacidade de processamento das GPUs, que permite a criação de sistemas de IA incrivelmente poderosos.

GPU NVIDIA L4 Tensor Core

- Possui suporte DirectX12 Ultimate.
- Possui 240 núcleos tensores para aprimorar a velocidade de aplicativos de aprendizado de máquina.
- 60 núcleos de aceleração ray tracing.
- Memória GDDR6 de 24 GB e é conectada através de uma interface de memória de 192 bits.
- Opera em uma frequência de 795 MHz mas pode ser aumentada para 2040 MHz
- Transistores de 5 nm
- A memória funciona a uma frequência de 1563 MHz.

GPU NVIDIA L4 Tensor Core

Esta GPU oferece alto desempenho em gráficos, vídeos e inteligência artificial, possuindo uma aceleração extremamente eficiente e uma economia de energia incrível para diversos usos, sendo uma placa bem compacta, que proporciona alta produtividade e economia.

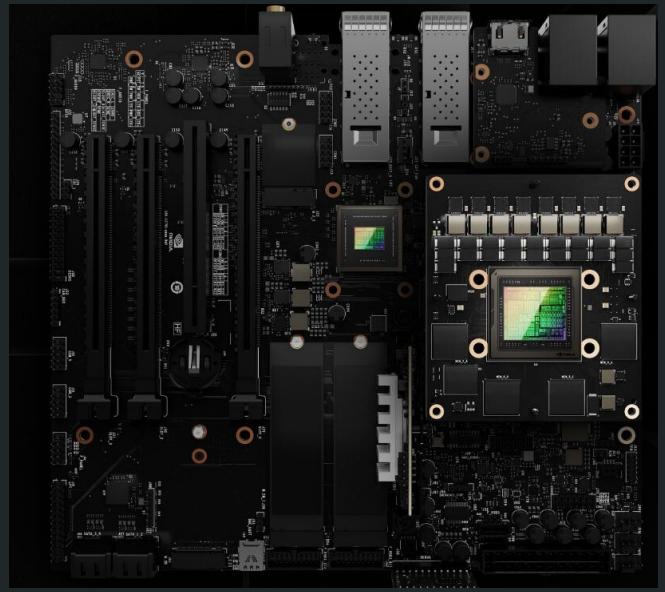
Em relação a utilização desta placa no mercado, diversas empresas grandes a utilizam, como descript, Twitter, SnapChat e o Google Cloud.

Data Sheet NVIDIA L4 Tensor Core

Specifications	
FP32	30.3 teraFLOPs
TF32 Tensor Core	120 teraFLOPS*
FP16 Tensor Core	242 teraFLOPS*
BFLOAT16 Tensor Core	242 teraFLOPS*
FP8 Tensor Core	485 teraFLOPs*
INT8 Tensor Core	485 TOPs*
GPU memory	24GB
GPU memory bandwidth	300 GB/s
NVENC | NVDEC | JPEG decoders	2 | 4 | 4
Max thermal design power (TDP)	72W
Form factor	1-slot low-profile, PCIe
Interconnect	PCIe Gen4 x16 64GB/s
Server options	Partner and NVIDIA-Certified Systems with 1–8 GPUs

NVIDIA IGX Orin

- Possui uma CPU 12-core ARM.
- A GPU possui uma frequência 1185 GHz, enquanto a CPU opera com uma frequência de 1996 GHz.
- Possui uma GPU Ampere 1024 CUDA, 64 Tensor.
- Oferece 275 tera operações por segundo de desempenho em IA.
- Possui uma unidade de microcontrolador de seguranca.



NVIDIA IGX Orin

Em relação ao seu uso, percebe-se uma grande utilização no meio industrial,como também no meio médico, permitindo uma relação entre máquinas e humanos de extrema segurança e eficiência, tornando o ambiente de trabalho mais dinâmico e inteligente.

Em relação a utilização desta plataforma por algumas empresas e centros médicos, há alguns parceiros da NVIDIA, como Siemens, Canonical e Activ Surgical.

Data Sheet NVIDIA IGX Orin

	IGX Orin Developer Kit
AI Performance	248 TOPS
SOM (System on Module)	GPU: 2,048-core NVIDIA Ampere architecture with 64 Tensor Cores CPU: 12-core Arm® Cortex®-A78AE v8.2
NVIDIA ConnectX-7	NVIDIA ConnectX-7 2x QSFP PCIe Gen 5.0 compatible, 32 lanes
Safety MCU (sMCU)	Infineon Aurix TC397
NVIDIA BMC (Baseboard Management Controller) Module	Aspeed AST2600 Microchip ERoT
GPU Max Frequency	1.185 GHz
CPU Max Frequency	1.996 GHz
DL Accelerator	2x NVDLA 2.0 Engines
DLA Max Frequency	1.4 GHz
Vision Accelerator	1x PVA v2
Memory	64GB 256-bit LPDDR5 204.8GB/s
Storage	64GB eMMC 5.1
Discrete GPU card	NVIDIA A6000 (optional)
Wireless	802.11 a/b/g/n/ac BT 5.0
Video Encode	2x 4K60 (H.265) 4x 4K30 (H.265) 8x 1,080p60 (H.265) 16x 1,080p30 (H.265)
Video Decode	1x 8K30 (H.265) 3x 4K60 (H.265) 7x 4K30 (H.265) 11x 1,080p60 (H.265) 22x 1,080p30 (H.265)
HDMI-IN	HDMI 2.0b input (up to 4Kp60)
PCIe	2x PCIe Gen5 (from ConnectX-7 PCIe switch) x8 lanes within 16x physical connector x16 lanes within x16 physical connector
USB	1x USB 3.2 Gen2 Type-C connectors 4x USB 3.2 Gen2 Type-A connectors

Data Sheet NVIDIA IGX Orin

Ethernet	2x RJ45 (up to 1 GbE) 2x QSFP ports (up to 100GB per port/up to 25 Gb/s per channel)
Display	One DisplayPort 1.4a output HDMI 2.0b input (up to 4kp60)
Audio (AU)	Three 3.5mm AU jack (MIC, line-in, speaker out)
Power	100 ~ 240 VAC 100 VAC: 8.5 A 240 VAC: 3.5 A
Mechanical	262.7 x 382 x 151.2 mm 6,700 grams

Produtos - Mercado Aeroespacial

Mercado Aeroespacial

A adoção de produtos da NVIDIA no mercado aeroespacial representa um marco significativo na evolução tecnológica dessa indústria.

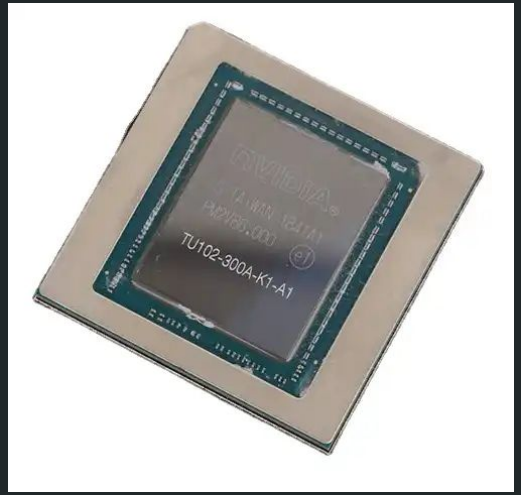
Sólidas parcerias com agências espaciais.

> Ex: NASA, aprimorou a qualidade das simulações de voo e facilitou tarefas, como navegação, monitoramento de missões e processamento de dados em tempo real.

Elevação do padrão da experiência espacial e destaque da importância da tecnologia de ponta na exploração e pesquisa fora da Terra.

GPU NVIDIA Quadro RTX 8000

- Utiliza processo de fabricação de 12nm e é baseada no processador gráfico TU102-875-A1.
- Área de matriz de 754 mm^2 e 18,6 bilhões de transistores.
- Possui 576 núcleos tensores que aceleram aplicativos de aprendizado de máquina.
- 72 núcleos de aceleração ray tracing.
- Memória GDDR6 de 48 GB funcionando a 1750 Mhz.
- Opera em uma frequência de 1395 MHz, mas pode ser aumentada para 1770 MHz.



GPU NVIDIA Quadro RTX 8000

Utilizações na área aeroespacial:

- Simulações.
- Treinamentos de piloto.
- Análise de dados de satélites.
- Missões espaciais.

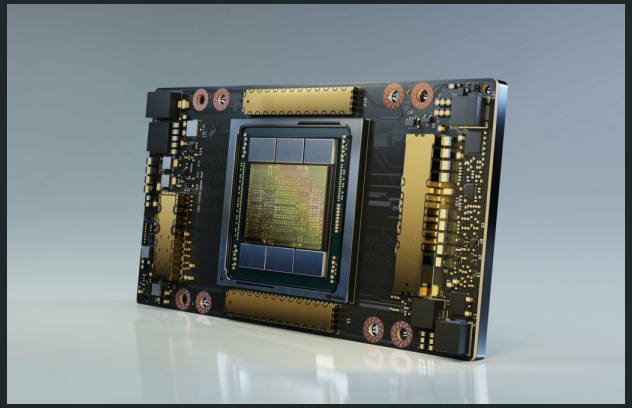
Utilizada pela NASA para acelerar a análise de imagens solares coletadas pelo Solar Dynamics Observatory.

Data Sheet GPU NVIDIA Quadro RTX 8000

SPECIFICATIONS	
GPU Memory	48 GB GDDR6
Memory Interface	384-bit
Memory Bandwidth	672 GB/s
ECC	Yes
NVIDIA CUDA Cores	4,608
NVIDIA Tensor Cores	576
NVIDIA RT Cores	72
Single-Precision Performance	16.3 TFLOPS
Tensor Performance	130.5 TFLOPS
NVIDIA NVLink	Connects 2 Quadro RTX 8000 GPUs[5]
NVIDIA NVLink bandwidth	100 GB/s (bidirectional)
System Interface	PCI Express 3.0 x 16
Power Consumption	Total board power: 295 W Total graphics power: 260 W
Thermal Solution	Active
Form Factor	4.4" H x 10.5" L, Dual Slot, Full Height
Display Connectors	4xDP 1.4, VirtualLink (1)
Max Simultaneous Displays	4x 3840 x 2160 @ 120 Hz, 4x 5120x2880 @ 60 Hz, 2x 7680x4320 @ 60 Hz
Encode / Decode Engines	1X Encode, 1X Decode
VR Ready	Yes
Graphics APIs	DirectX 12.0[7], Shader Model 5.1[7], OpenGL 4.6[8], Vulkan 1.1[8]
Compute APIs	CUDA, DirectCompute, OpenCL™

GPU NVIDIA A100 Tensor Core

- Utiliza processo de fabricação de 7nm e é baseada no processador gráfico GA100.
- Área de matriz de 826 mm^2 e 54,2 bilhões de transistores.
- Possui 432 núcleos tensores que aceleram aplicações de aprendizado de máquina.
- Memória HBM2e de 80 GB funcionando a 1593 Mhz.
- Opera em uma frequência de 1275 MHz, mas pode ser aumentada para 1410 MHz.
- Possuindo a maior largura de banda de memória do mundo, atinge 2 terabytes por segundo.



GPU NVIDIA A100 Tensor Core

Amplamente adotada em aplicações aeroespaciais, devido aos ganhos de desempenho em várias tarefas computacionais. Ela possibilita:

- Simulações mais ágeis e precisas.
- Processamento de imagem e sinal.
- Aprendizado de máquina.
- Computação de alto desempenho.

Utilizada pela Silicon Mechanics, a qual oferece sistemas de GPU personalizados, incluindo soluções voltadas para a indústria aeroespacial.

Data Sheet GPU NVIDIA A100 Tensor Core

	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS | 1248 TOPS*	
GPU Memory	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,935GB/s	2,039GB/s
Max Thermal Design Power (TDP)	300W	400W***
Multi-Instance GPU	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe dual-slot air cooled or single-slot liquid cooled	SXM
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 600GB/s ** PCIe Gen4: 64GB/s	NVLink: 600GB/s PCIe Gen4: 64GB/s
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs	NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4,8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs

Produtos - Mercado Militar

Mercado Militar

Existem poucos registros sobre produtos militares, pois são informações extremamente privilegiadas se forem parar em mãos erradas. Porém uma das parcerias da NVIDIA com a DARPA, deu origem a Virtual Eye. O notebook que executa esse programa contém duas placas NVIDIA K20.

GPU NVIDIA Tesla K20X

- A placa suporta DirectX 12.
- Possui uma memória GDDR5 de 6GB com uma interface de memória de 384 bits.
- A GPU opera a uma frequência de 732 MHz, enquanto a memória funciona a uma frequência de 1300 MHz
- Possui 7080 milhões de transistores.

Data Sheet GPU NVIDIA Tesla K20X

Specifications	Tesla K20X
Generic SKU reference	699-22081-0200-xxx
Chip	GK110
Package size GPU	45 mm × 45 mm 2397-pin S-FCBGA
Processor clock	732 MHz
Memory clock	2.6 GHz
Memory size	6 GB
Memory I/O	384-bit GDDR5
Memory configuration	24 pieces of 64M ×16 GDDR5 SDRAM
Display connectors	None
Power connectors	•8-pin PCI Express power connector •6-pin PCI Express power connector
Board power	235 W
Idle power	25 W
Thermal cooling solution	Passive heat sink
Mean time between failures (MTBF)	•Uncontrolled environment: 128440 hours at 35 °C •Controlled environment: 208861 hours at 35 °C

https://www.youtube.com/watch?v=uZGJ2P2BaIg&t=3s

Mercado

Norte adotado para análise de mercado

Relatório fiscal da Nvidia entregue à receita americana e publicado no site de relações com investidores referente ao **terceiro trimestre de 2021.**

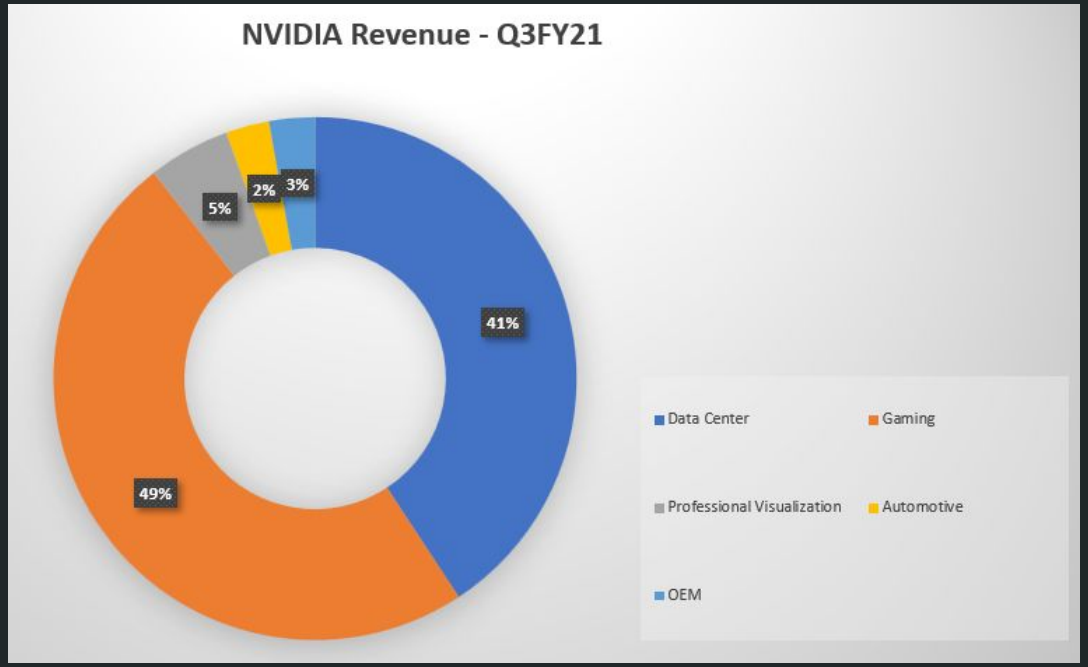
Divisão dos principais mercados atuantes (segundo o relatório)

- Indústria automotiva
- Jogos
- *Data center*
- Fabricação de equipamentos originais (da sigla OEM, em inglês)
- Visualização profissional

Distribuição de faturamento por mercado (Q3FY21)

- 45% da receita líquida no mercado de jogos
- 41% mercado de *data center*
- 8% visualização profissional
- 3% indústria automotiva
- 3% fabricação de equipamentos originais (OEM)

Distribuição de faturamento por mercado (Q3FY21) - Gráfico



Segmentos de Mercado tratados minuciosamente neste seminário

- Indústria automotiva
- Fabricação de produtos originais (OEM)
- Visualização professional
- *Data center*

Indústria automotiva

- Grandes players do segmento que possuem relações com a Nvidia:

Mercedes-Benz, Jaguar, Volvo, Land Rover e Hyundai Motor Group

Indústria automotiva (Campos de atuação)

- **Plataformas de Processamento Automotivo:** A Nvidia desenvolve plataformas de computação automotiva (NVIDIA DRIVE, por exemplo) que são utilizadas pelas montadoras como base para desenvolvimento de sistemas avançados de assistência ao motorista (ADAS).

Indústria automotiva (Campos de atuação)

- **Desenvolvimento de Processamento Automotivo:** A Nvidia disponibiliza variadas de desenvolvimento de software e simulação de prática de testes, utilizados no processo produtivo das montadoras.

Indústria automotiva (Campos de atuação)

- **Automação veicular:** Atualmente, a Nvidia colabora com as montadoras na construção de sistemas de direção autônoma, em que basicamente, por meio de modelos de deep learning, o mesmo se encarrega de tomar decisões de modo proativo, utilizando dados provenientes de múltiplos sensores no projeto do automóvel.

Fabricação de produtos originais (OEM)

- Grandes players do segmento que possuem relações com a Nvidia:

Lenovo, Acer, Asus, Supermicro, MSI, Adobe, Dell, Siemens, Boeing e Northrop Grumman.

Fabricação de produtos originais (OEM)[Campos de atuação]

- **Fornecimento de plataformas de desenvolvimento:** A Nvidia empreende múltiplas plataformas que servem, assim como o NVIDIA DRIVE, para desenvolvimento de sistemas mais complexos e voltados para as especificidades de cada produto final.

Fabricação de produtos originais (OEM)[Campos de atuação]

- **Produto de componentes-chaves:** A Nvidia provê múltiplos componentes essenciais que são integrados à versões finais de produtos. Em termos de hardware, essa relação se dá por meio de unidades de processamento de IA e GPUs majoritariamente. Esse produtos vão do mercado de consumidor final até contextos industriais, por exemplo.

Fabricação de produtos originais (OEM)[Campos de atuação]

- **Criação de ecossistemas:** A Nvidia, ao compor múltiplas cadeias produtivas de vários produtos, dispõe de um cenário ideal para a difusão de um ecossistema tecnológico abrangente, pois, ao possuir os componentes da marca, muitos produtos possuem um ponto comum que viabiliza essa integração

Data Center

- Arquitetura “ampere”
- Adaptação das empresas a inteligência artificial
- Aceleração de gpu
- Líder no mercado

Visualização Profissional

A mais de 20 anos a Nvidia possui ampla atuação na área de computação visual profissional, quando se diz a respeito de aplicativos profissionais relacionados com Manufatura, Imagens médicas e científicas e Energia até Mídia e entretenimento, através dos modelos de placa de video “Quadro”, que utiliza de tecnologias avançadas de visualização que visa proporcionar a melhor experiência visual e aumento considerável de produtividade em workflows.

Visualização Profissional

Empresas que a Nvidia possui parceria na área:

- **Lockheed Martin** (setor militar - projetos de simulação e visualização avançados para práticas militares)
- **General Dynamics** (setor militar - sistemas de análise de risco e monitoramento de campo)
- **Rockwell Automation** (setor industrial - sistema de análise e gestão de processos industriais)

Arquitetura Ampere

- Superior a seu antecessor (arquitetura turing)
- Tensor Cores de nova geração (Foco em I.A)
- Extremamente hábil a lidar com redes convolucionais
- Eficiência energética

Pergunta

Com base no que foi apresentado neste seminário e ao decorrer da disciplina, qual fato fez com que a NVIDIA conseguisse alcançar um grande sucesso no mercado de IA?

Resposta

A NVIDIA teve muito sucesso na área de IA devido ao desenvolvimento de chips gráficos especiais, chamados GPUs, que são muito bons em realizar os cálculos necessários para a inteligência artificial de forma rápida e eficiente. Isso tornou possível treinar e executar modelos de IA de maneira mais eficaz, impulsionando o progresso nessa área.

Referência das Imagens:

https://nvidianews.nvidia.com/bios/jensen-huang
https://ecse.rpi.edu/about/hall-of-fame/curtis-r-priem
https://nvidianews.nvidia.com/bios/chris-a-malachowsky
https://www.cgdirector.com/nvlink-vs-sli/
https://medium.com/geekculture/introduction-to-cuda-7bf6909ea57c
https://materiaincognita.com.br/titan-luta-para-provar-que-e-o-supercomputador-mais-poderoso-da-terra/
https://www.amax.com/products/nvidia-products/nvidia-dgx-1/
https://www.tecmundo.com.br/tecnologia-militar/106543-tecnologia-militar-permite-soldados-vejam-obstaculos.htm
https://www.techpowerup.com/gpu-specs/l4.c4091
https://www.nvidia.com/pt-br/edge-computing/products/igx/
https://cdn.wccftech.com/wp-content/uploads/2020/05/NVIDIA-GA100-GPU-Ampere-1030x667.jpg