Formação Inteligência Artificial

Programação Paralela em GPU

Hardware para Construção de Modelos em GPU

Data Science Academy

Data Science Academy

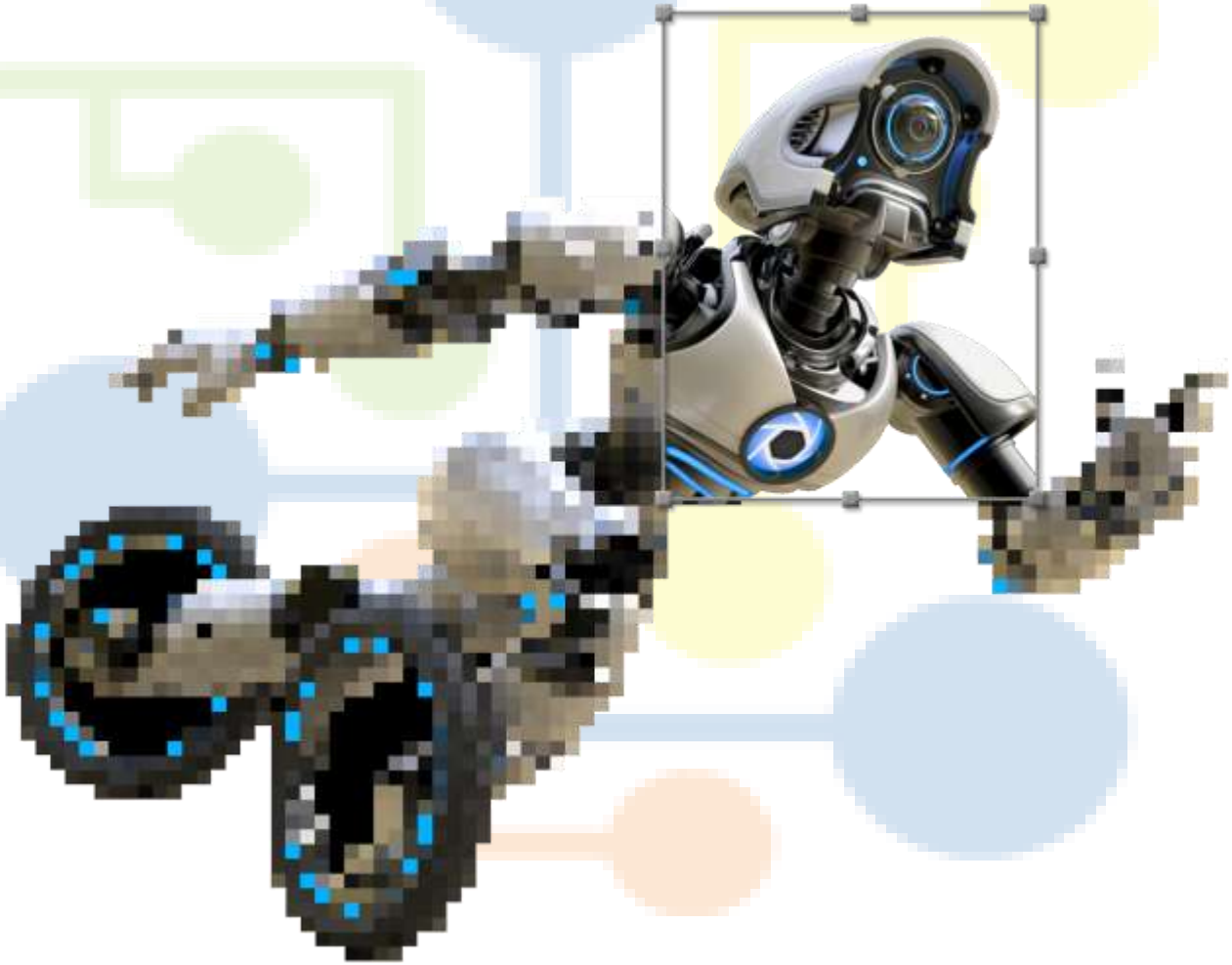
Uma Breve História das Placas de Vídeo

Data Science Academy

Uma Breve História das Placas de Vídeo

Data Science Academy

Uma Breve História das Placas de Vídeo



Data Science Academy

Uma Breve História das Placas de Vídeo

OpenGL®

Data Science Academy

CUDA e OpenCL

Data Science Academy

CUDA e OpenCL

(Compute Unified Device Architecture)

Data Science Academy

CUDA e OpenCL

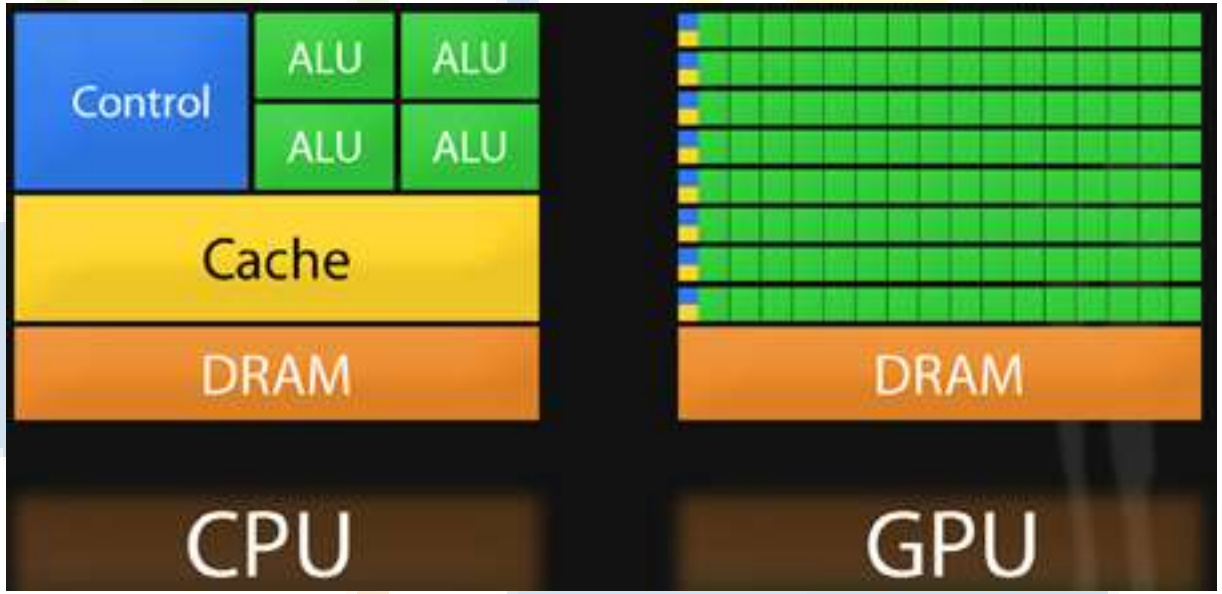
(Open Computing Language)

Data Science Academy

Arquitetura das GPUs Atuais

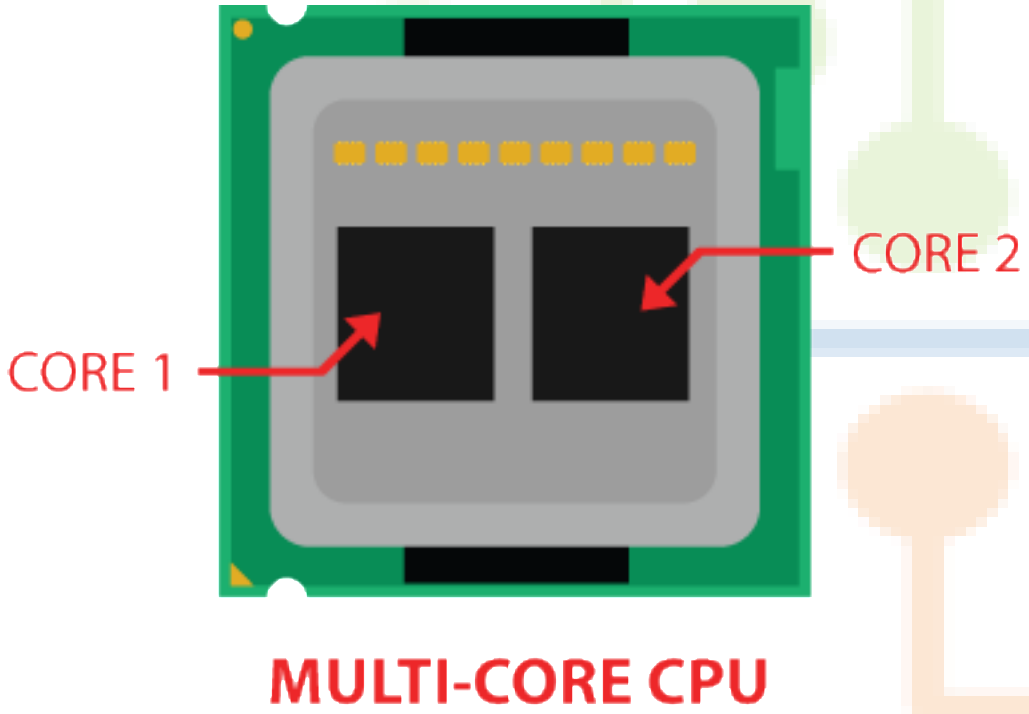
Data Science Academy

Arquitetura das GPUs Atuais



Data Science Academy

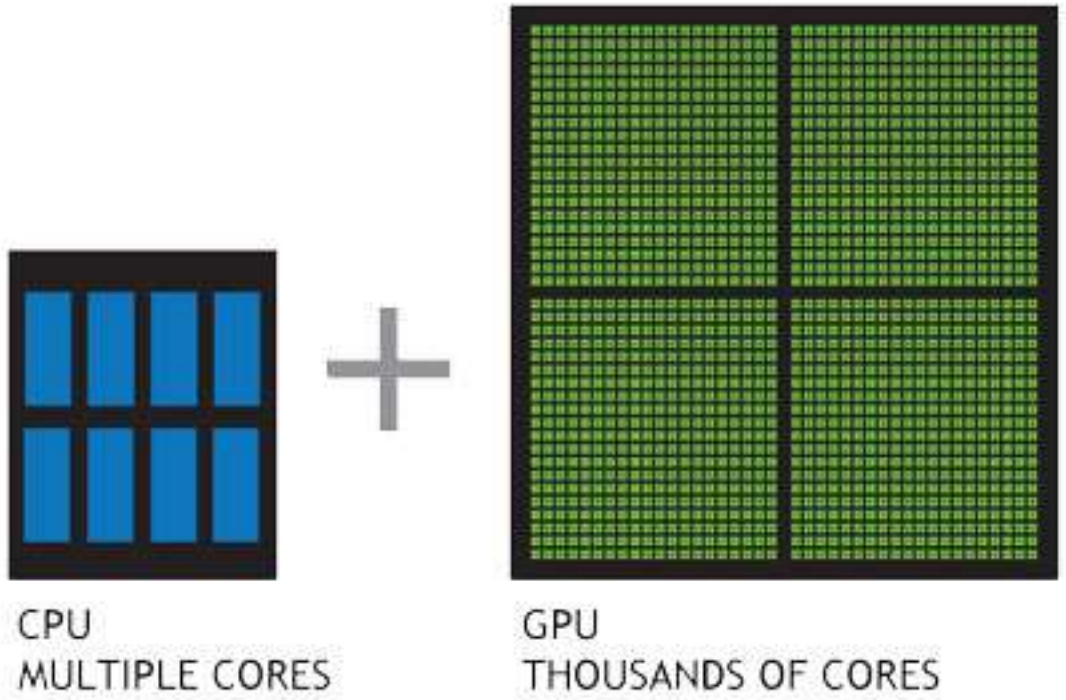
Arquitetura das GPUs Atuais



O clock rate normalmente se refere à frequência na qual um núcleo (core) de um processador multicore, está sendo executado e é usado como um indicador da velocidade do processador.

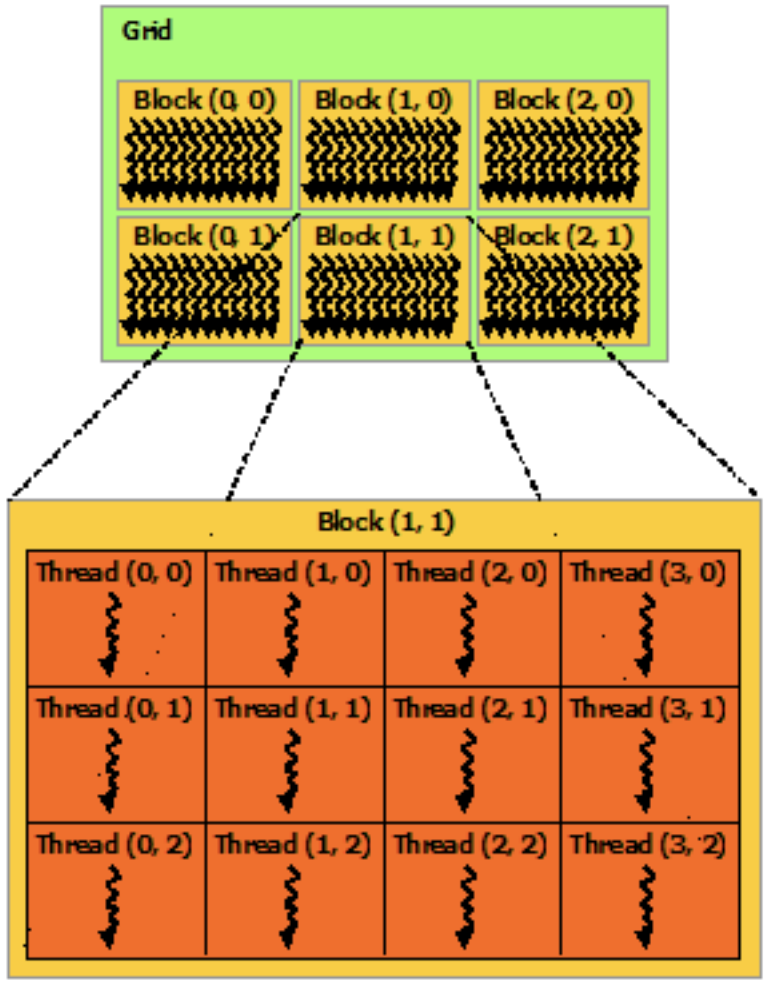
Data Science Academy

Arquitetura das GPUs Atuais



Data Science Academy

Arquitetura das GPUs Atuais



Como aumentar a capacitade computacional?

- Clocks mais rápidos
- Mais processamento por ciclo de clock
- Mais núcleos (cores)

Data Science Academy

Arquitetura das GPUs Atuais

CPU

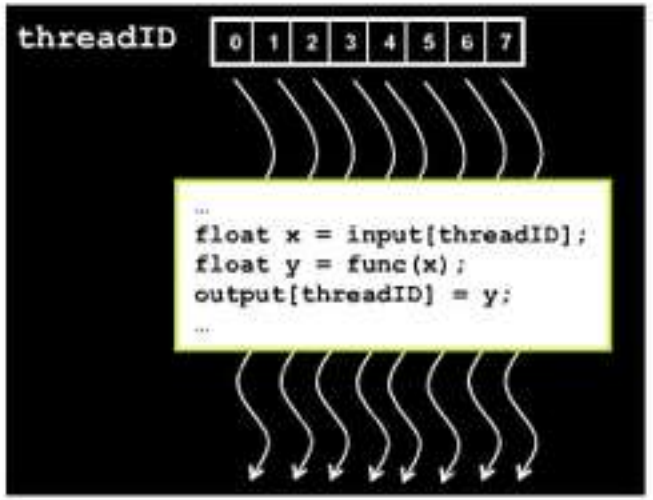
- Maior complexidade
- Maior flexibilidade
- Maior custo em termos de consumo de energia

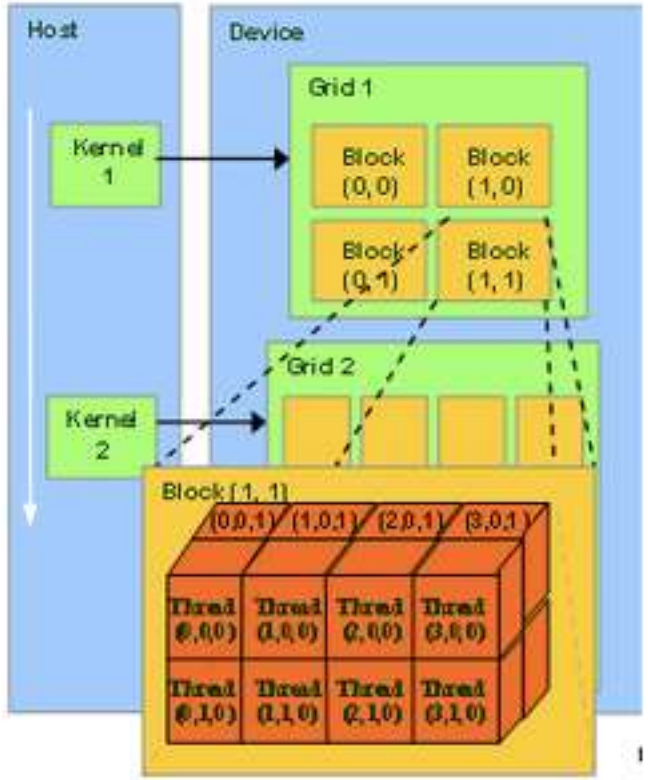
GPU

- Maior simplicidade
- Requer mais hardware para computação
- Menor custo em termos de consumo de energia (potencialmente)
- Modelo de Programação mais restritivo

Data Science Academy

Arquitetura das GPUs Atuais





Data Science Academy

Latência x Throughput

Data Science Academy

Latência x Throughput

Latência

Quantidade de tempo necessária para concluir uma tarefa

CPU

Throughput

Total de tarefas concluídas por unidade de tempo

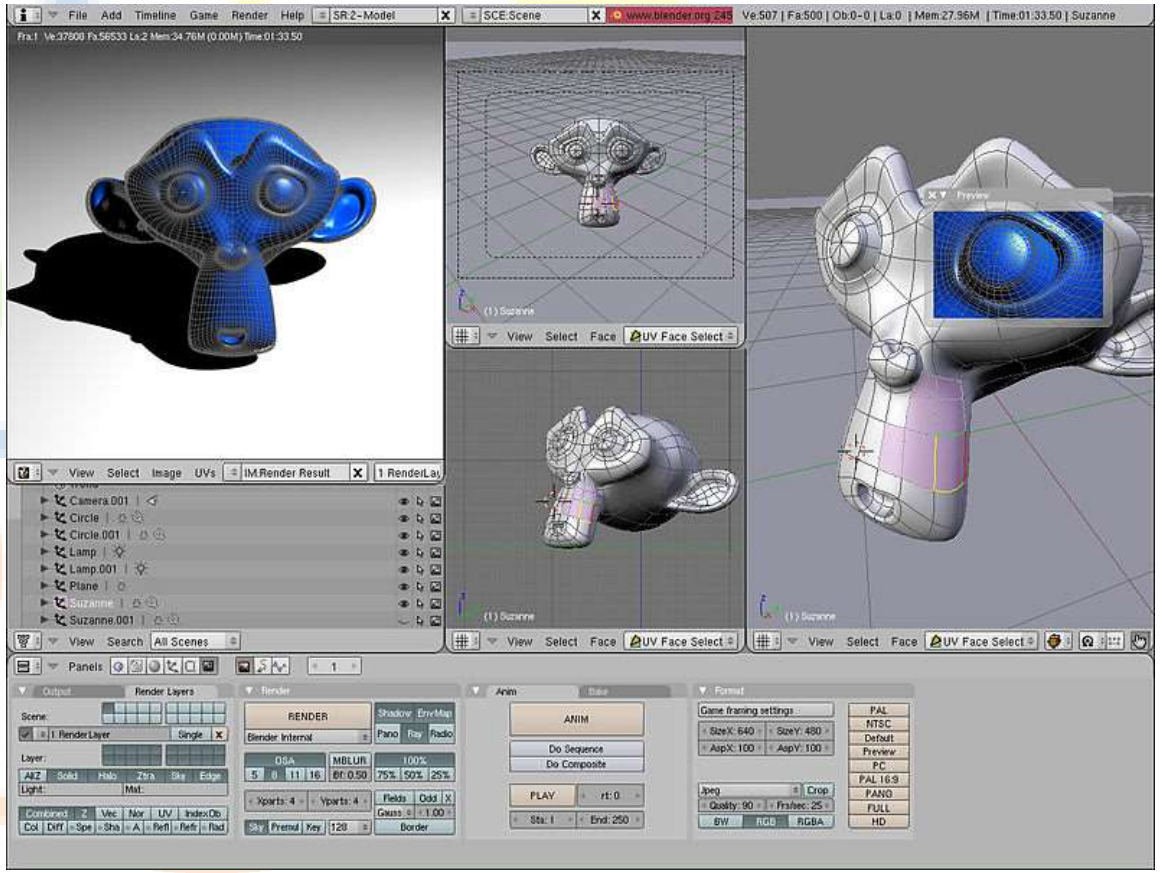
GPU

Data Science Academy

Latência x Throughput

Em computação gráfica, por exemplo, estamos mais preocupados com a quantidade de pixels por segundo (throughput) do que a latência de um pixel em particular.

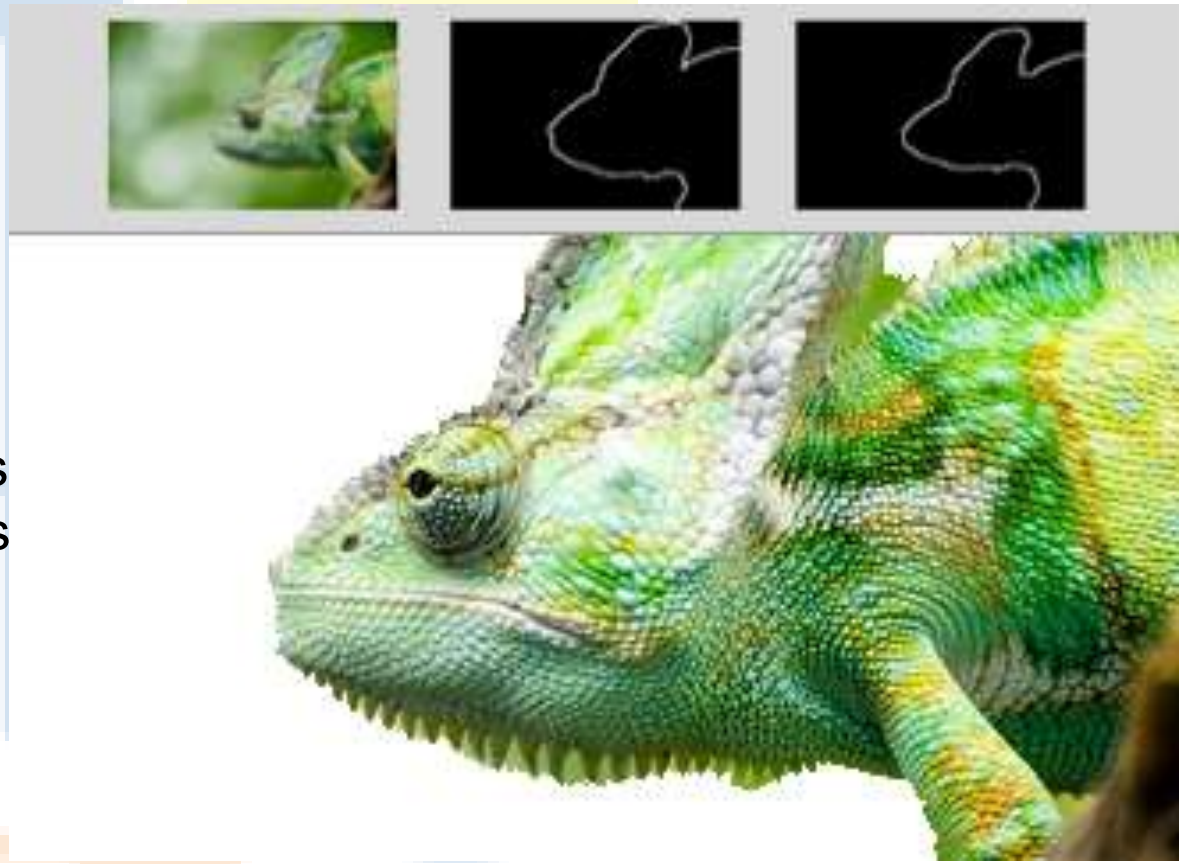
Não seria problema, se o processamento de um pixel em particular levar um pouco mais de tempo, se ao final o throughput de pixels for maior.



Data Science Academy

Latência x Throughput

Outro exemplo é o processamento de imagens, onde o throughput também é mais importante, pois estamos mais preocupados com pixels produzidos por segundo (throughput), do que o tempo de um pixel individual (latência).



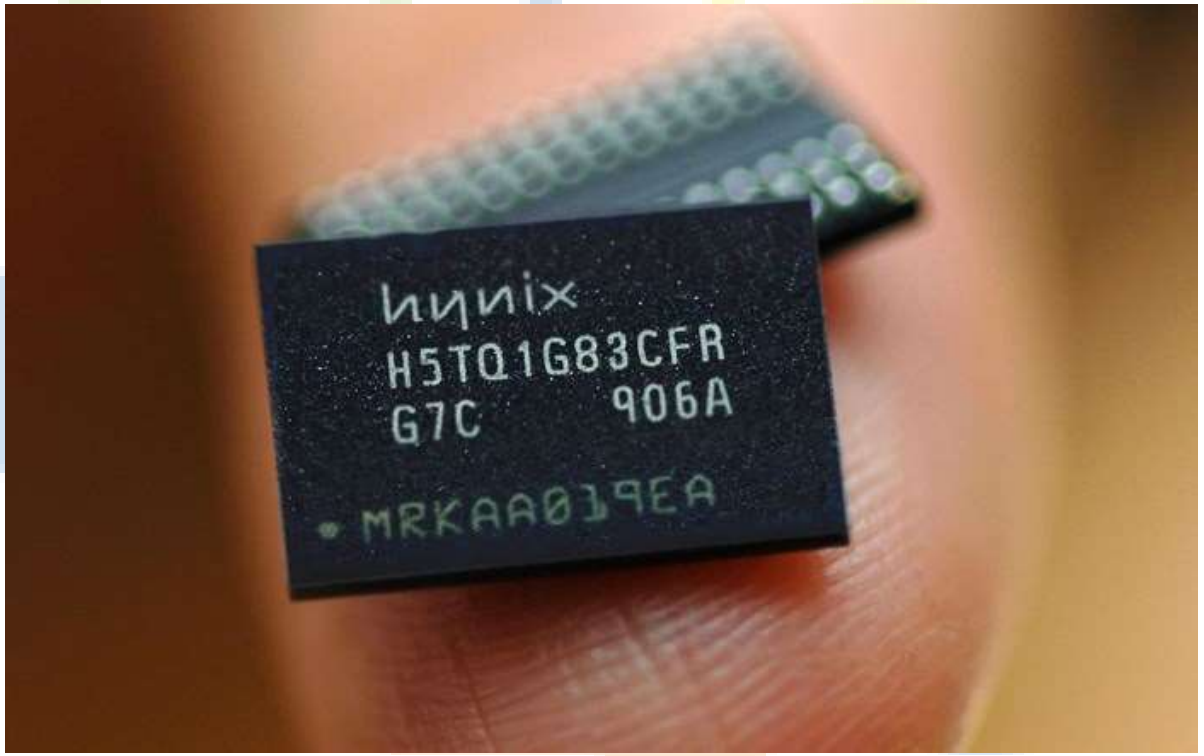
Data Science Academy

Largura de Banda da Memória na GPU

Data Science Academy

Largura de Banda da Memória na GPU

Largura de Banda da Memória na GPU



Data Science Academy

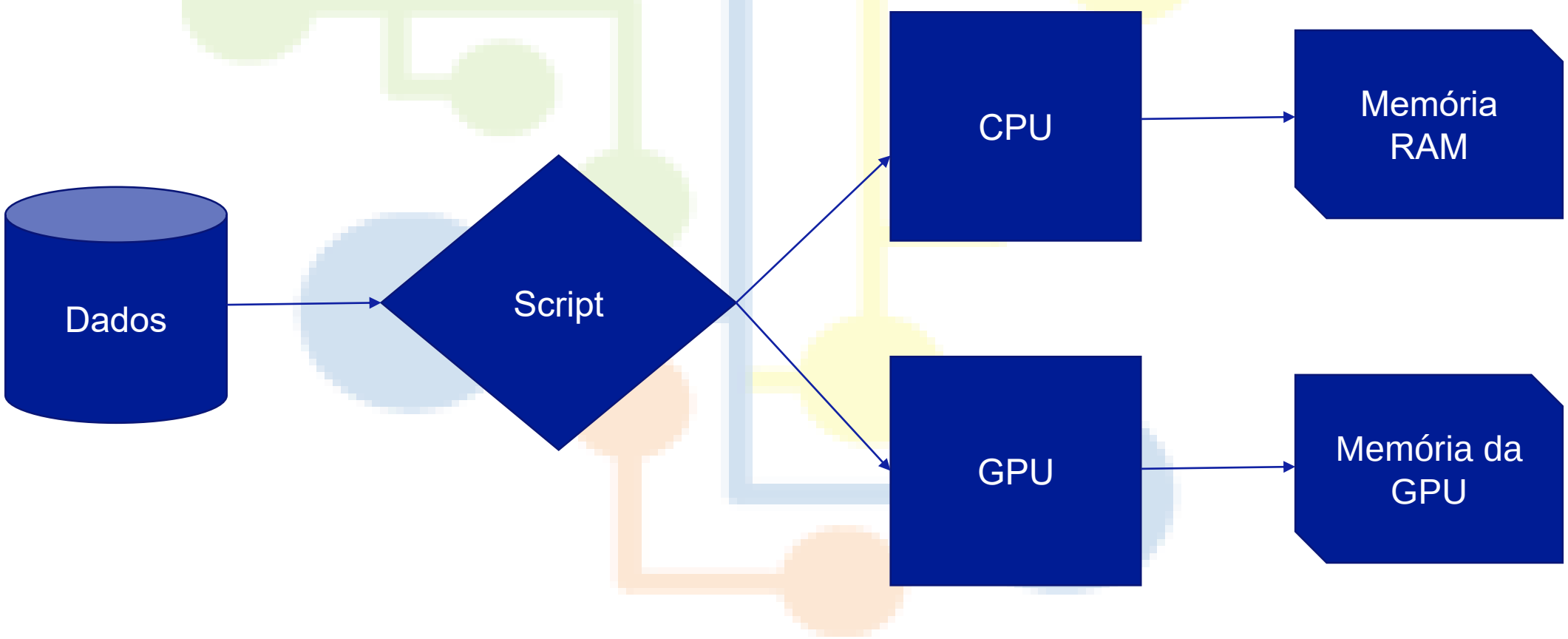
Largura de Banda da Memória na GPU

A quantidade de dados que a GPU pode transferir é justamente a chamada “Largura de banda” (que também é associada ao termo Taxa de transferência).

A largura de banda é a quantidade de dados que pode ser lida ou escrita na memória em um determinado intervalo de tempo.

Data Science Academy

Largura de Banda da Memória na GPU



Data Science Academy

Largura de Banda da Memória na GPU

DDR3

Memória RAM

GDDR5

Memória da GPU

Data Science Academy

Como Funciona uma GPU

Data Science Academy

Como Funciona uma GPU

CPU

Boa para obter pequenas quantidades de memória rapidamente
50 GB/s

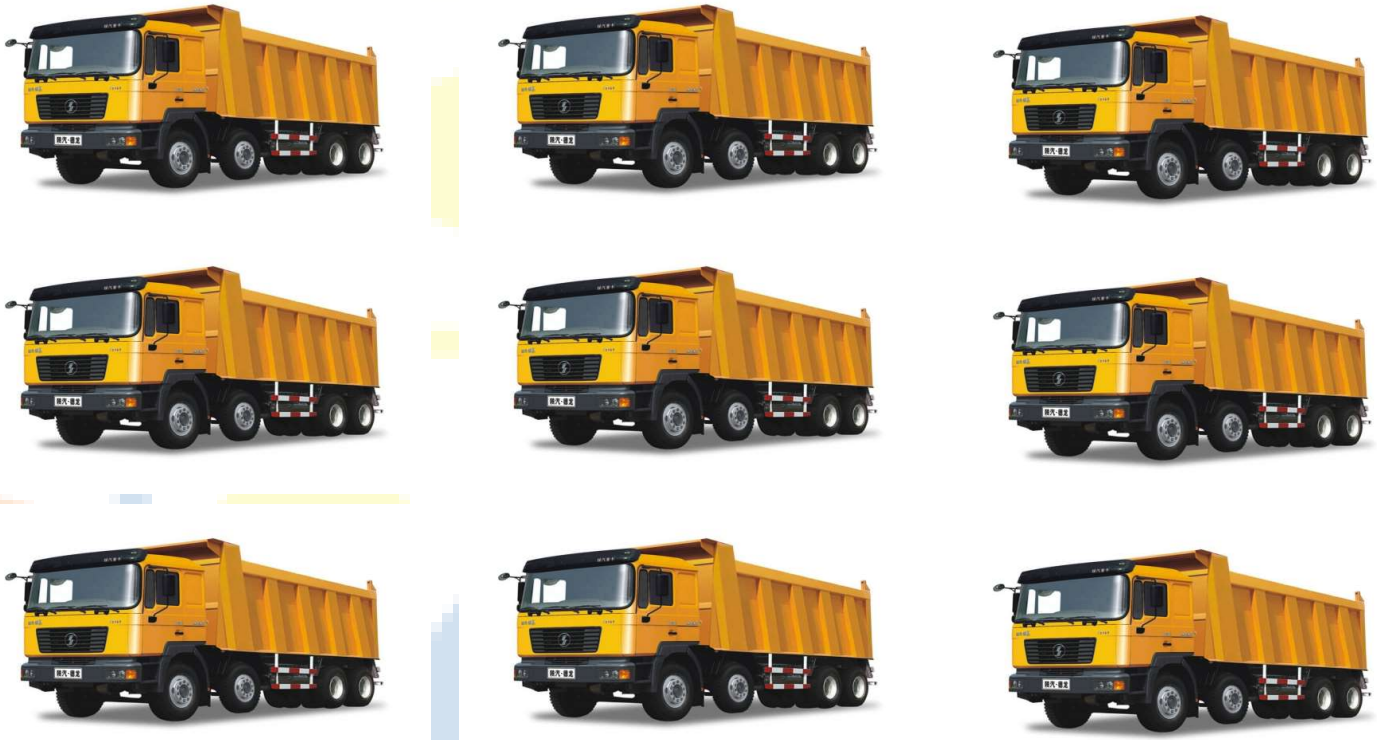
GPU

Boa para obter grandes quantidades de memória (como multiplicação de matrizes)
750 GB/s

Data Science Academy

Como Funciona uma GPU

CPU



GPU

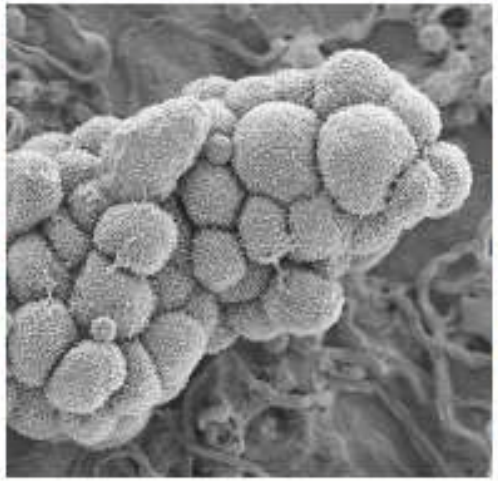
Data Science Academy

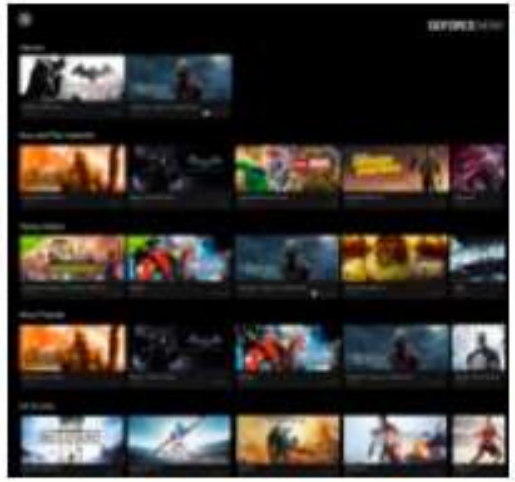
Por que as GPUs são Ideais para Deep Learning?

Data Science Academy

Por que as GPUs são Ideais para Deep Learning?

DEEP LEARNING EVERYWHERE





INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation

MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

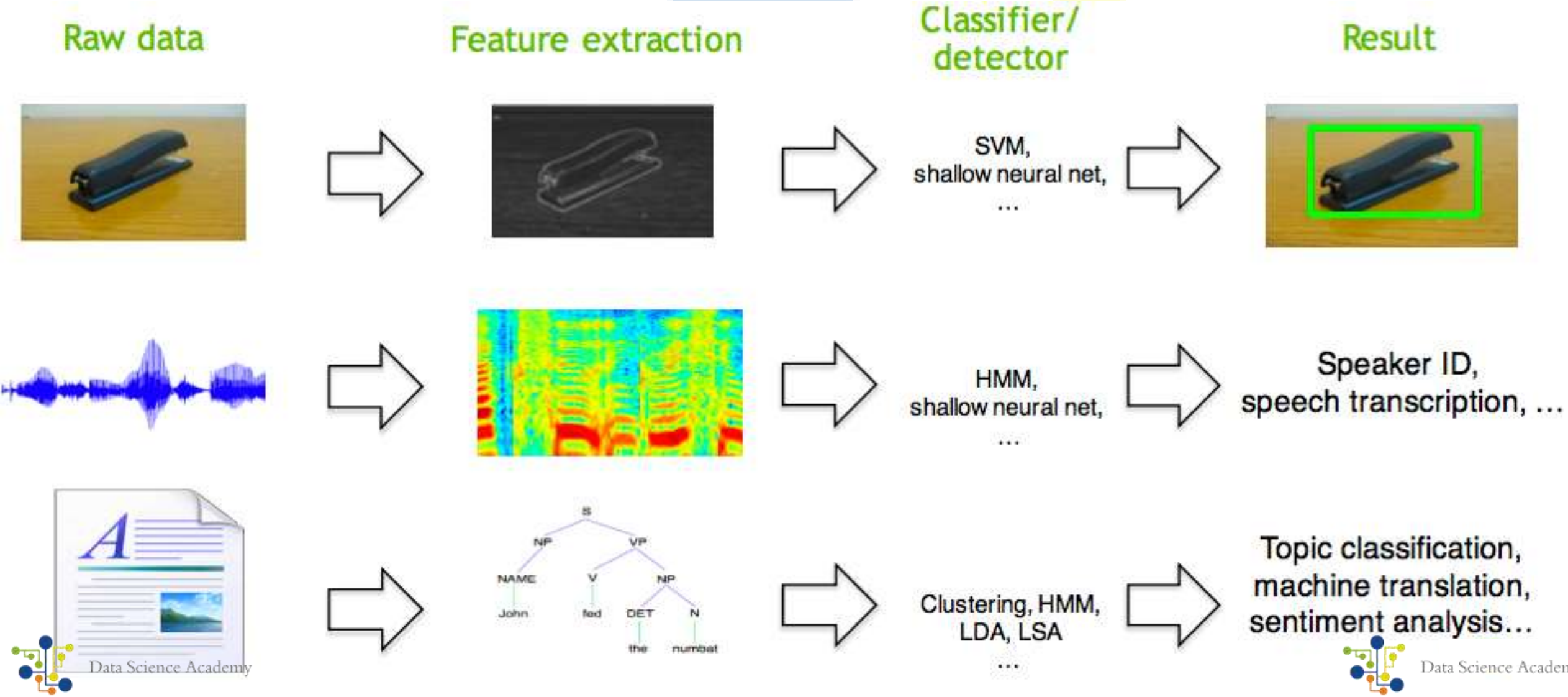
SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Por que as GPUs são Ideais para Deep Learning?



Por que as GPUs são Ideais para Deep Learning?

Por que as GPUs são Ideais para Deep Learning?

Em resumo, as GPUs funcionam bem com os cálculos Redes Neurais Profundas, porque:

1. GPUs têm muitos mais recursos e uma banda mais rápida para a memória
2. Os cálculos DNN se encaixam bem com a arquitetura GPU.

A velocidade computacional é extremamente importante porque o treinamento de Redes Neurais Profundas pode variar de dias a semanas, mas com o uso de GPUs, estamos reduzindo isso a horas.

Na verdade, muitos dos sucessos em Deep Learning poderiam não ter sido descobertos, se não fosse a disponibilidade de GPUs.

Data Science Academy

Tipos de Paralelismo

Data Science Academy

Tipos de Paralelismo

Instruction-Level Parallelism

Data Parallelism

Task Parallelism

Data Science Academy

Tipos de Paralelismo

Thread-level Parallelism

Data Science Academy

O Tipo de Paralelismo na GPU

Data Science Academy

O Tipo de Paralelismo na GPU

Em GPUs, o tipo de paralelismo oferecido é o Data Parallelism, mais precisamente baseado no modelo SIMT(Single Instruction Multiple Thread).

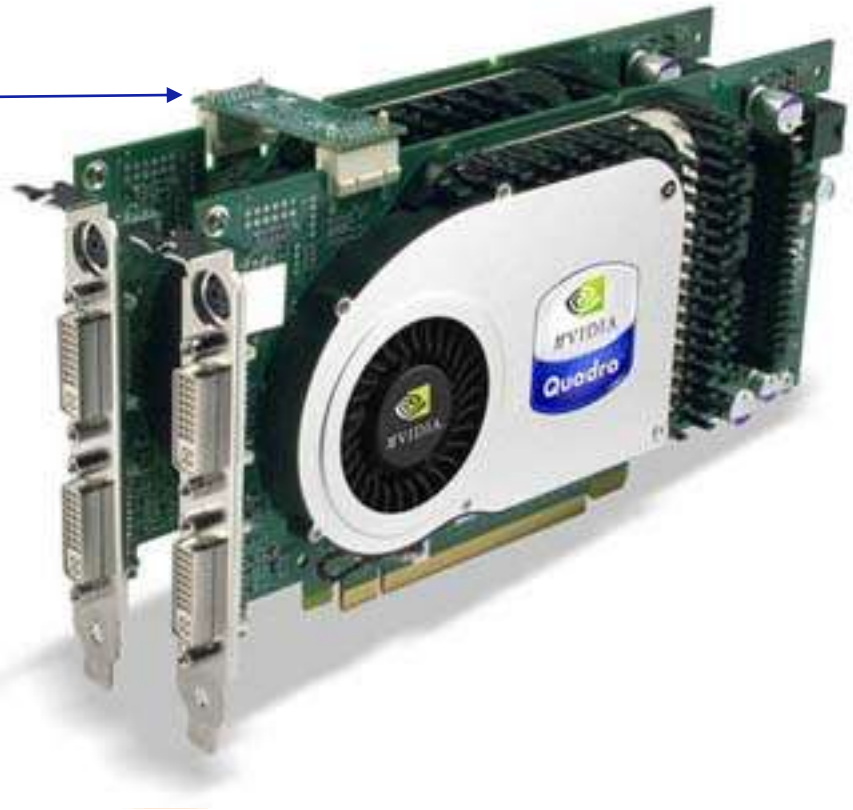
Data Science Academy

SLI – Scalable Link Interface

Data Science Academy

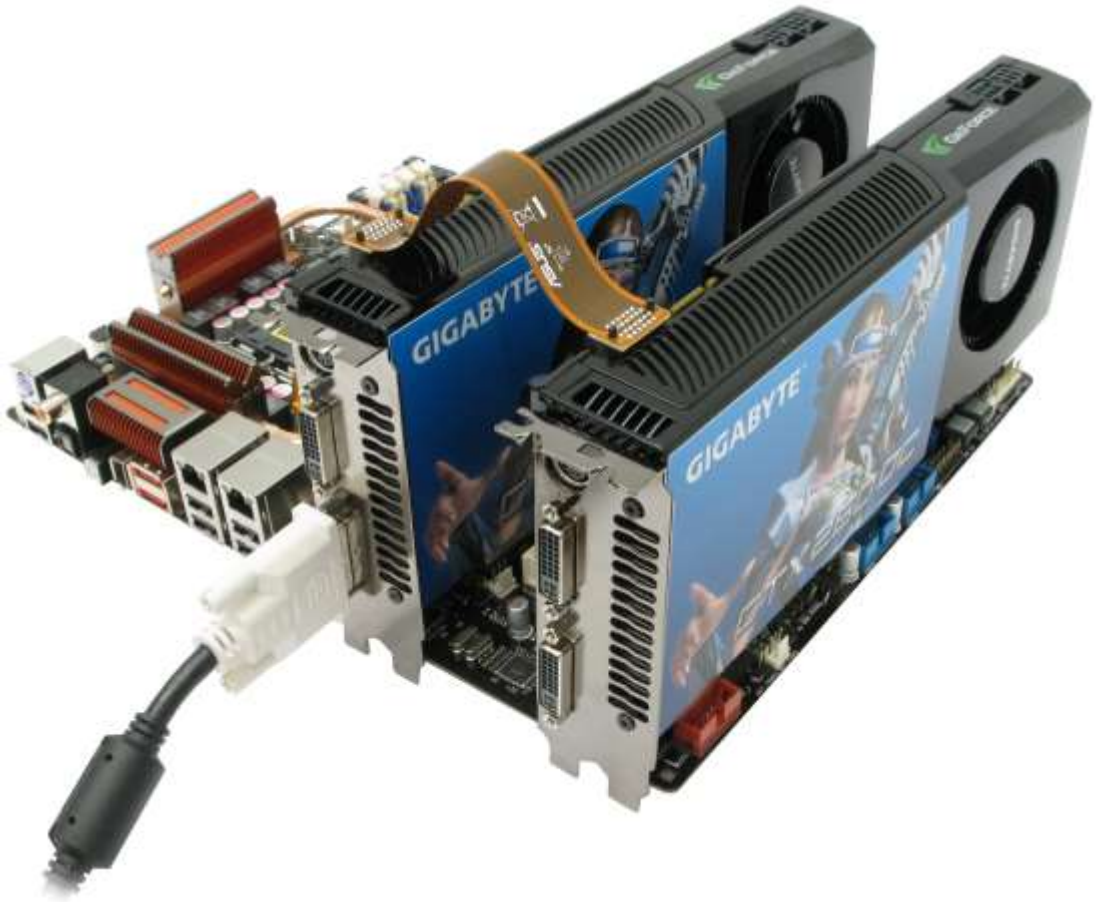
SLI – Scalable Link Interface

Conector SLI



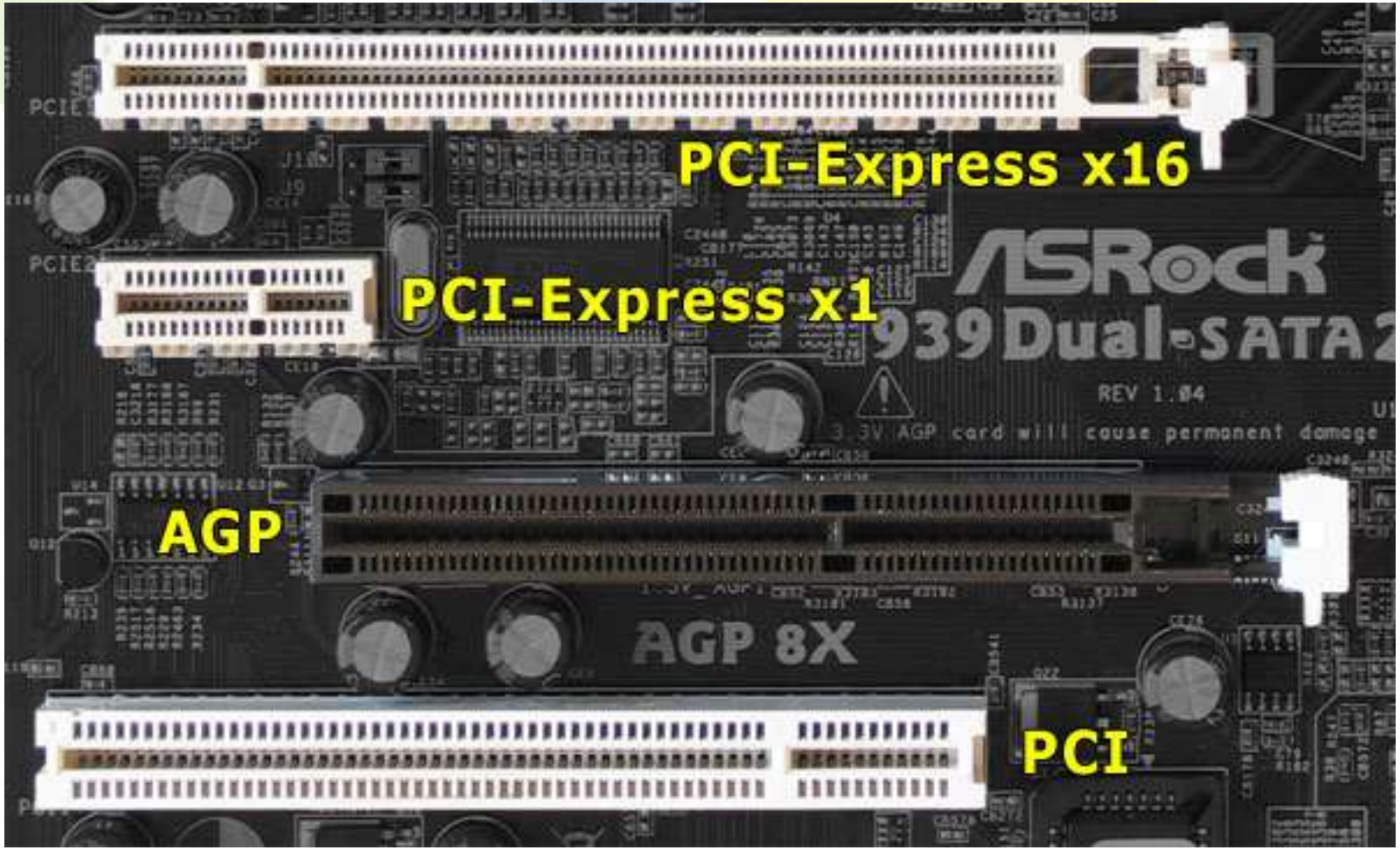
Data Science Academy

SLI – Scalable Link Interface

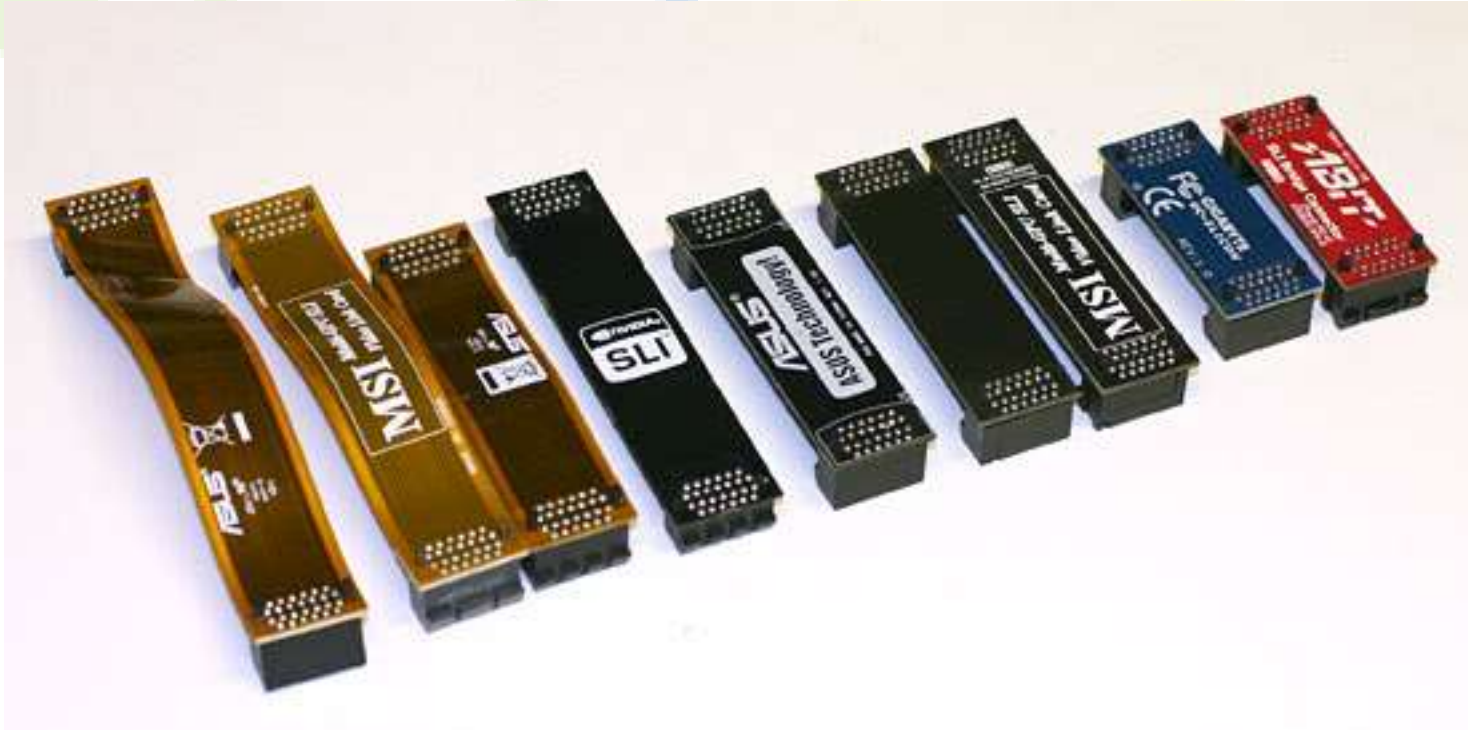


Data Science Academy

SLI – Scalable Link Interface

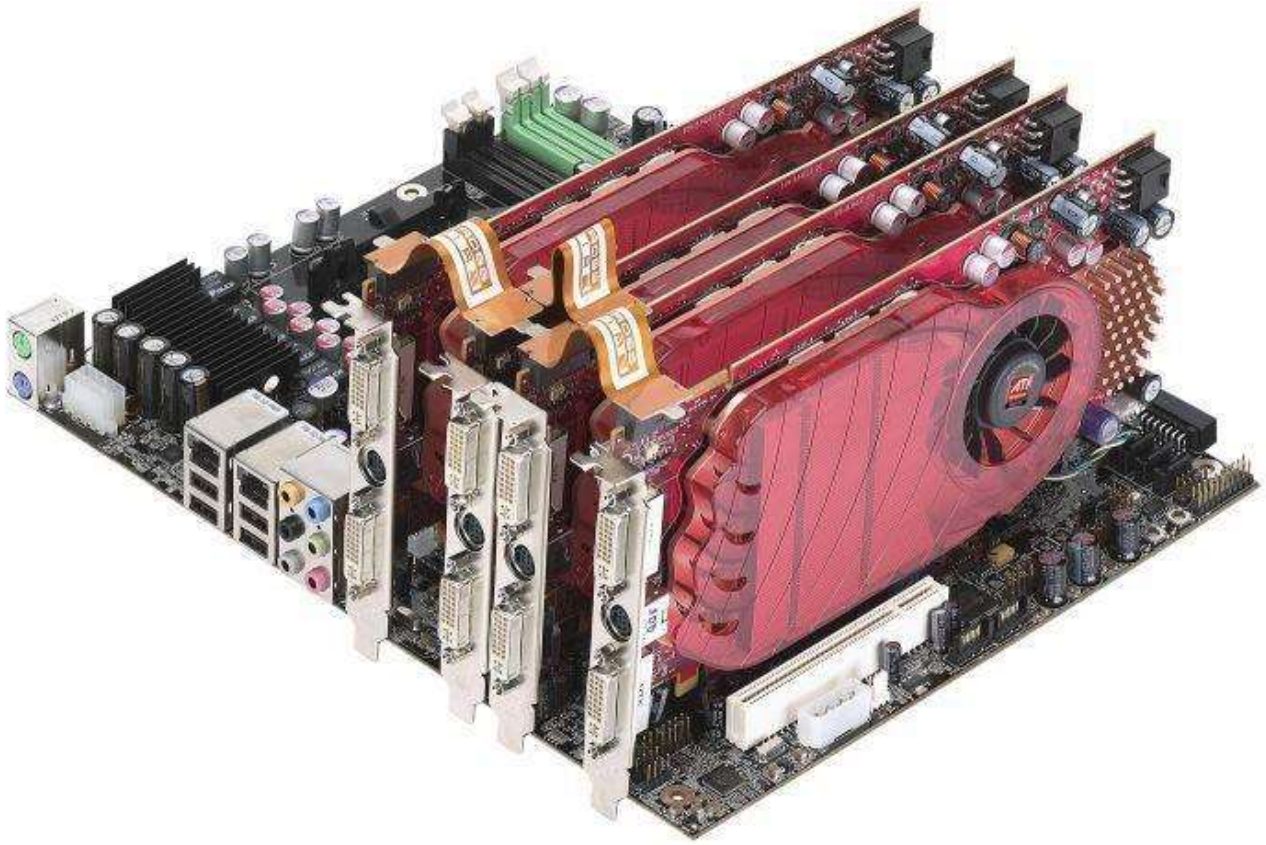


SLI – Scalable Link Interface



SLI x Crossfire

SLI x Crossfire

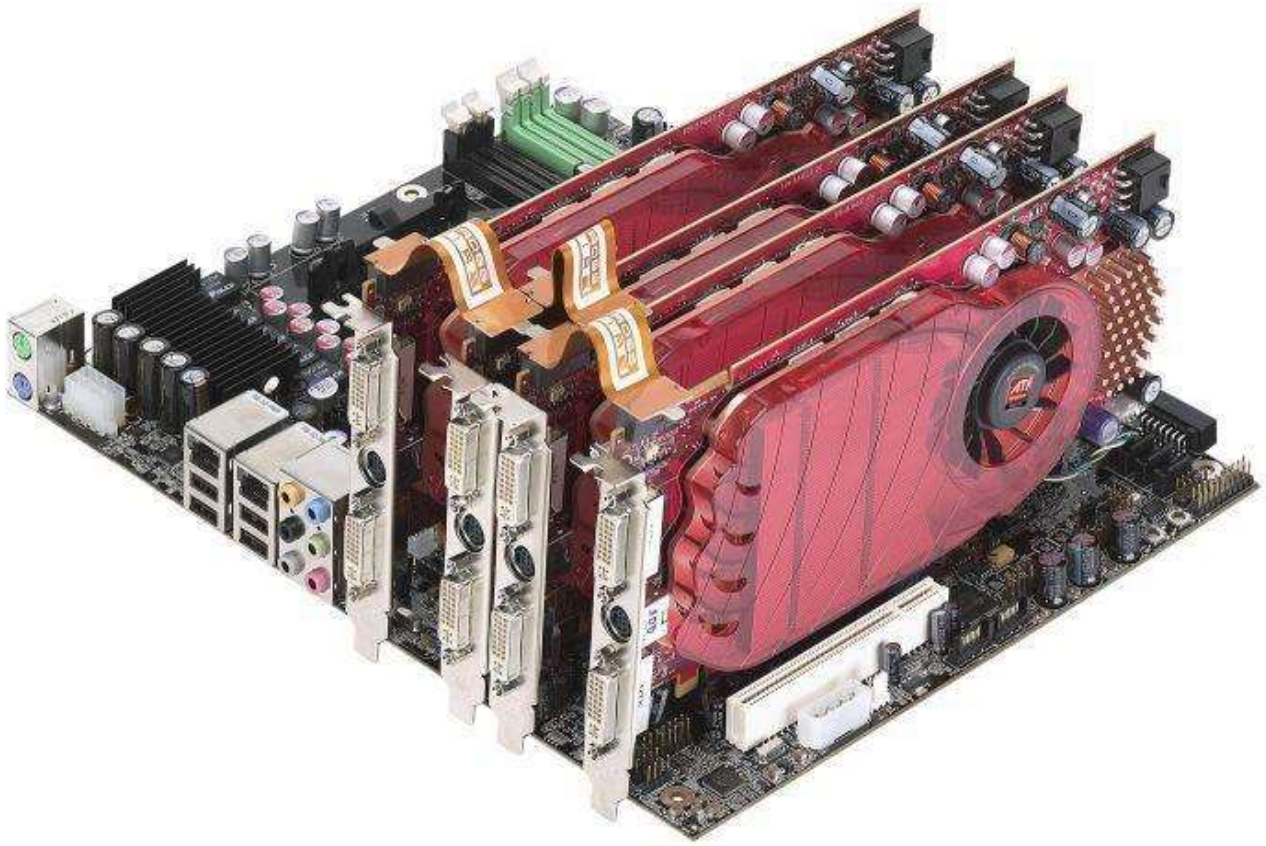


Data Science Academy

SLI x Crossfire

Data Science Academy

SLI x Crossfire



Data Science Academy

SLI x Crossfire

Data Science Academy


Obrigado