Formação Inteligência Artificial

Programação Paralela em GPU

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Computação em GPU

Data Science Academy

Data Science Academy

Computação em GPU

As GPUs são processadores especializados em processar imagem e são extremamente poderosos devido a sua arquitetura paralela e sua eficiência, tanto no acesso a memória como nas operações vetoriais e de interpolação

Computação em GPU

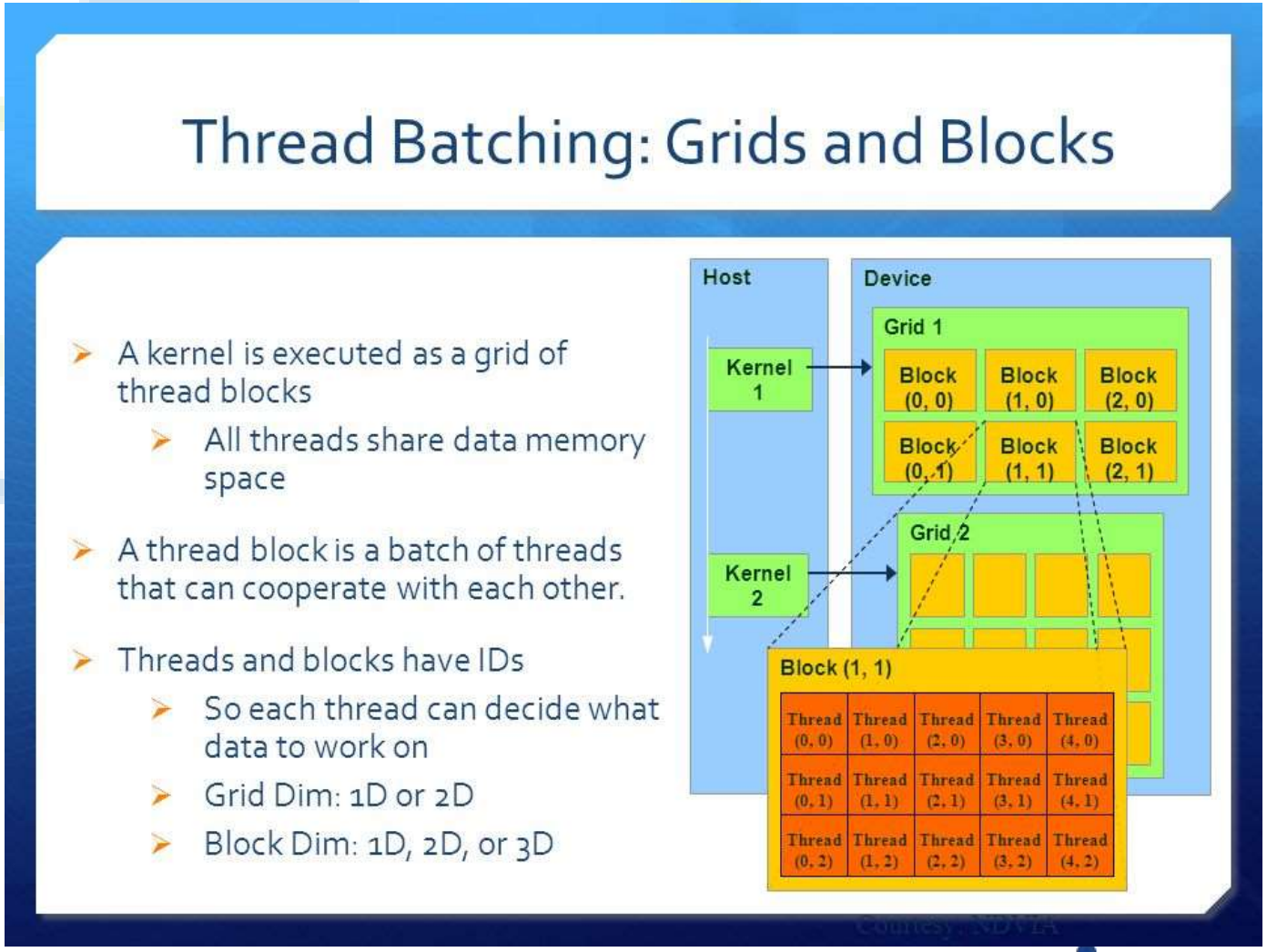
OpenCL™

PyCUDA

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

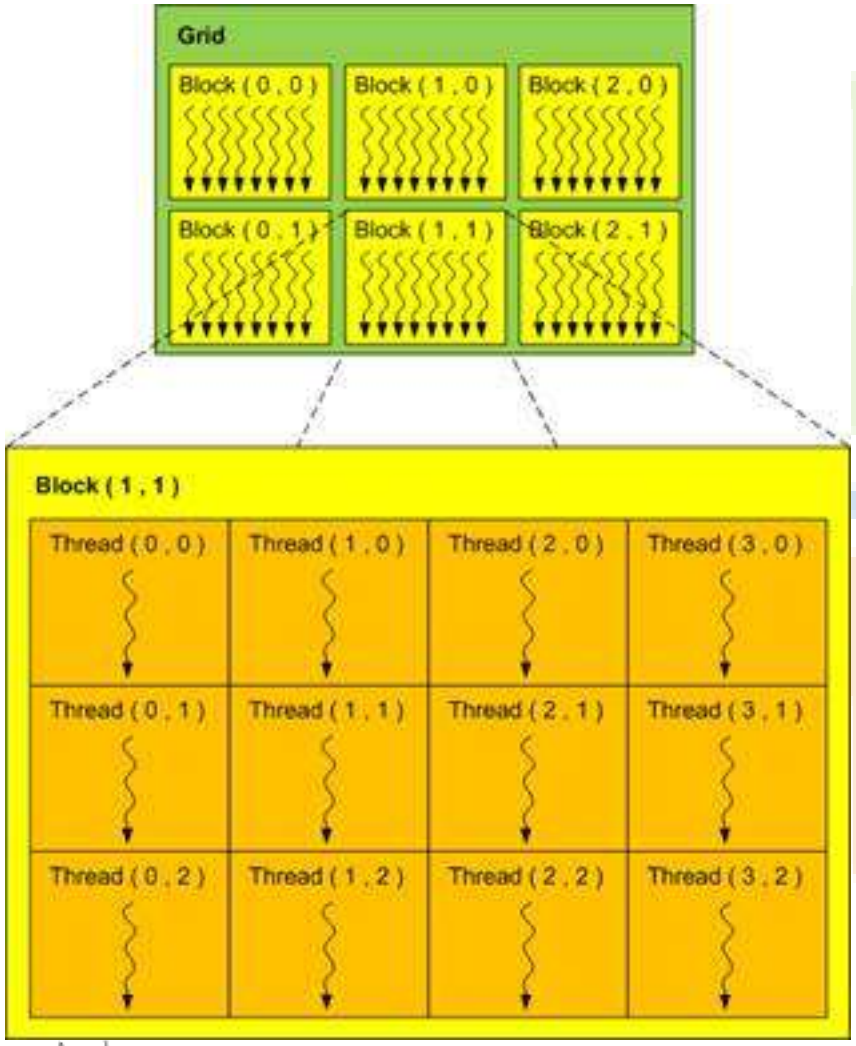
Computação em GPU

Uma função CUDA é chamada de kernel



Data Science Academy

Computação em GPU

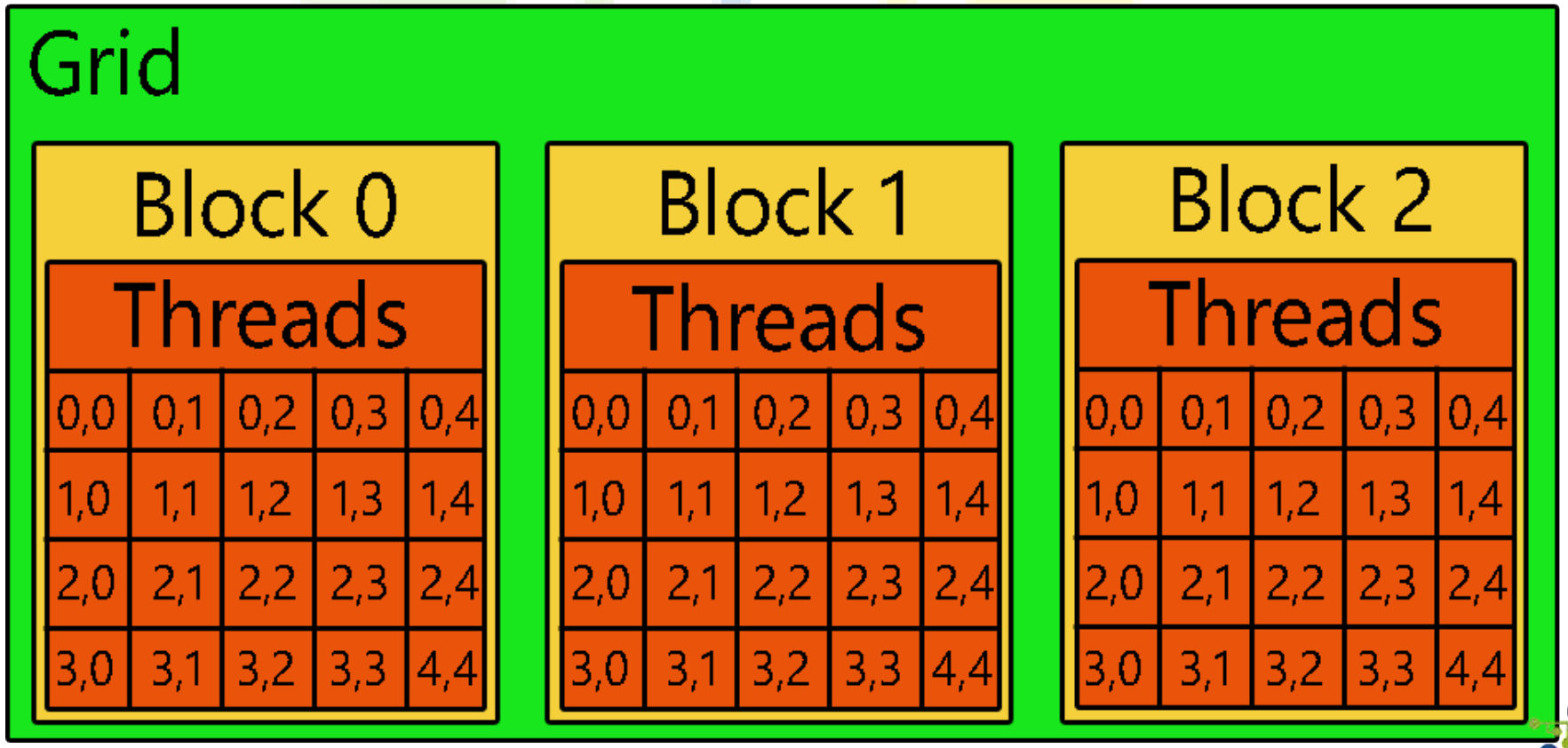


As **threads** são organizadas em grupos que são chamados de **blocks**, que por sua vez são organizados e chamados de **grids.**

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Computação em GPU

1D Grid of 2D Blocks



Data Science Academy

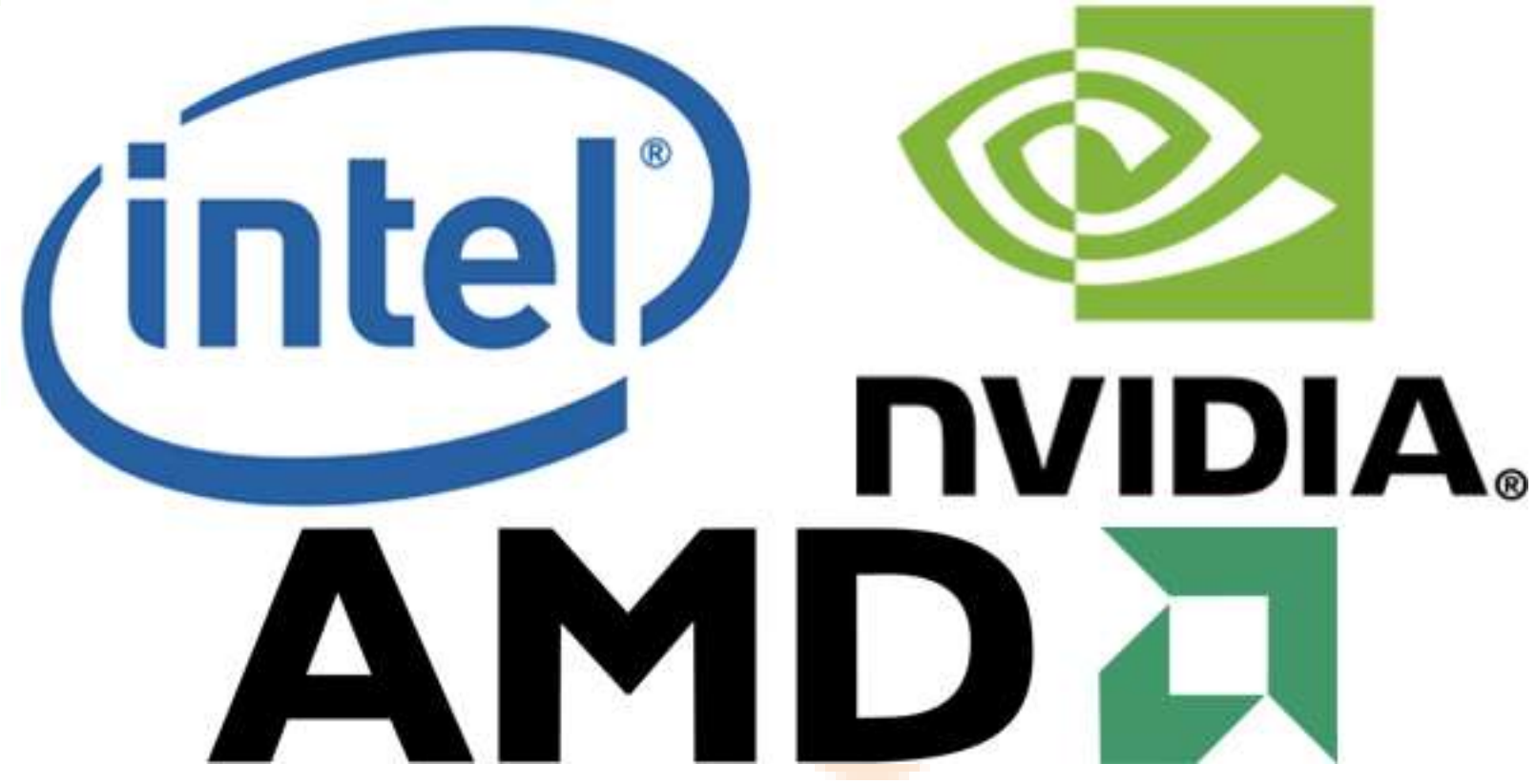
Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

História da Computação em GPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

História da Computação em GPU



Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

História da Computação em GPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

História da Computação em GPU

Sem GPU
O processamento gráfico era feito pela própria CPU

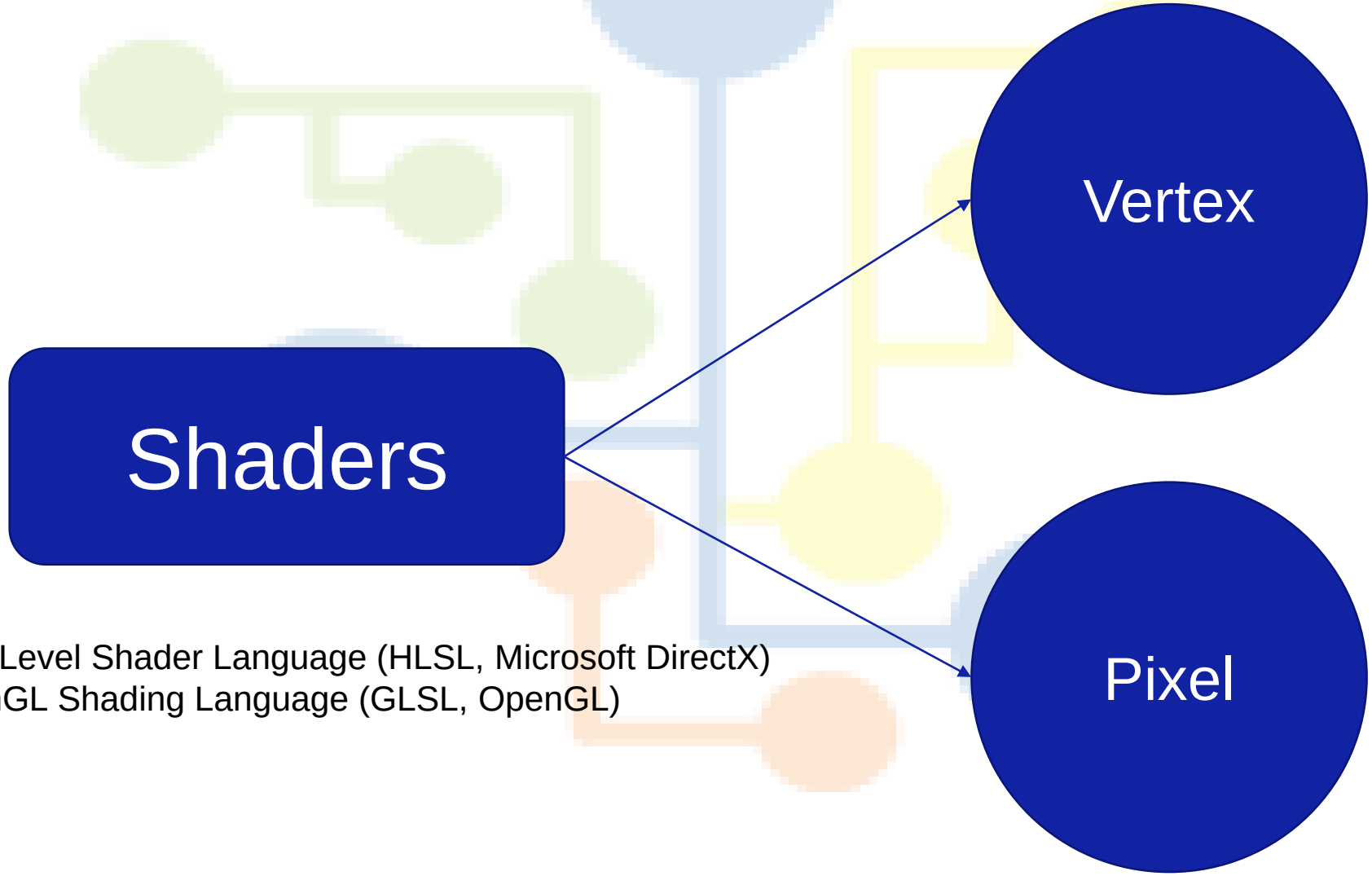
GPU Dedicada
Processamento gráfico separado (PCI, AGP)

GPU Programáveis
Shaders

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

História da Computação em GPU



High Level Shader Language (HLSL, Microsoft DirectX)
OpenGL Shading Language (GLSL, OpenGL)

Data Science Academy

Data Science Academy Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

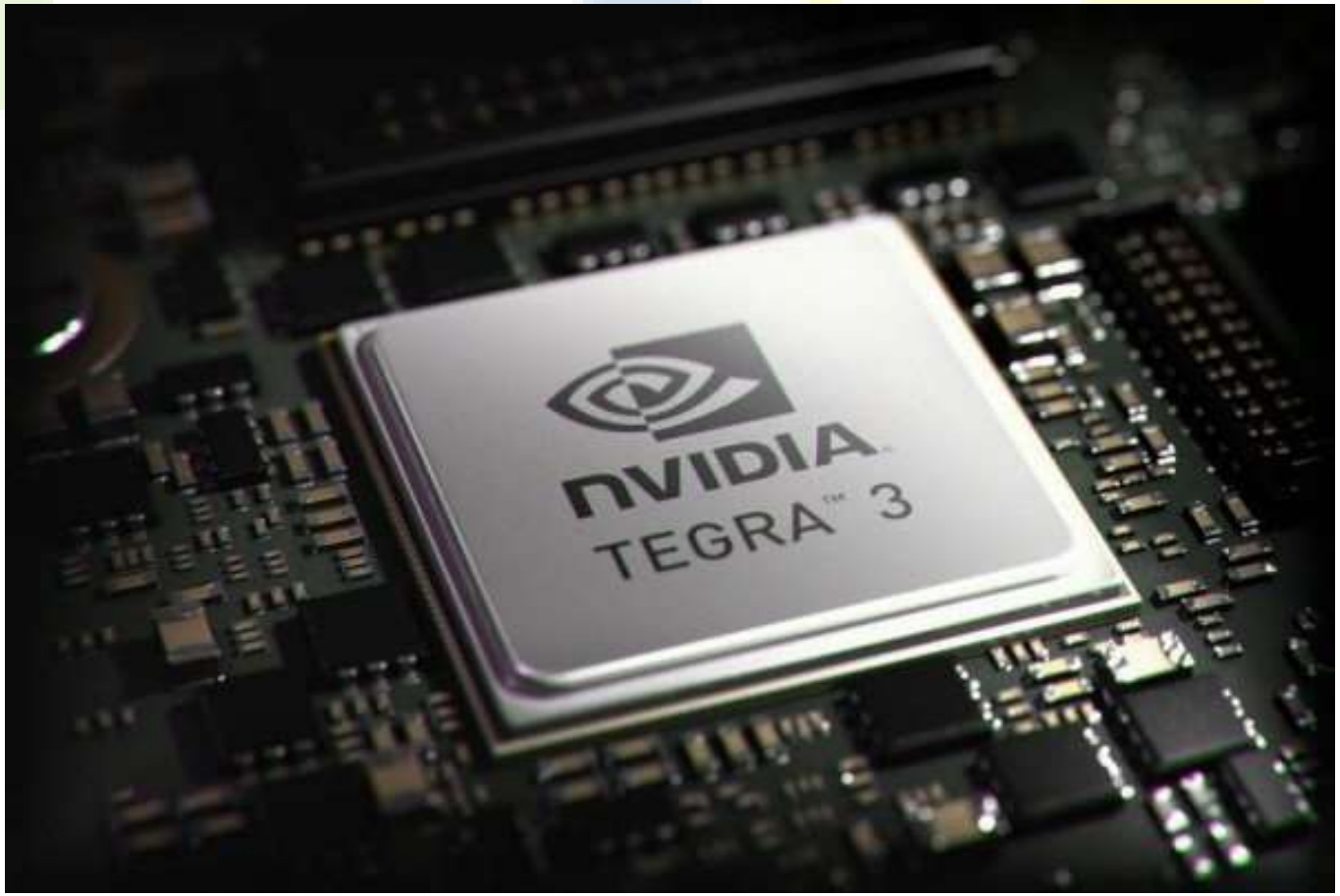
História da Computação em GPU

Na segunda metade da década de 2000, os shaders passaram a ser chamados de kernels, que suportam várias linhas de código, além da introdução ao suporte a ponto flutuante de precisão dupla.

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

História da Computação em GPU



Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

História da Computação em GPU

Tais processadores se chamam Streaming Processors, ou Processadores de Fluxo (numa tradução livre), e GPUs modernas contém até milhares deles.

A GPU Titan X por exemplo possui 3584 desses Streaming Processors, ou na nomenclatura da Nvidia, CUDA cores.

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU

General Purpose Graphical Processing Unit

Unidade de Processamento Gráfico de Propósito Geral

Data Science Academy

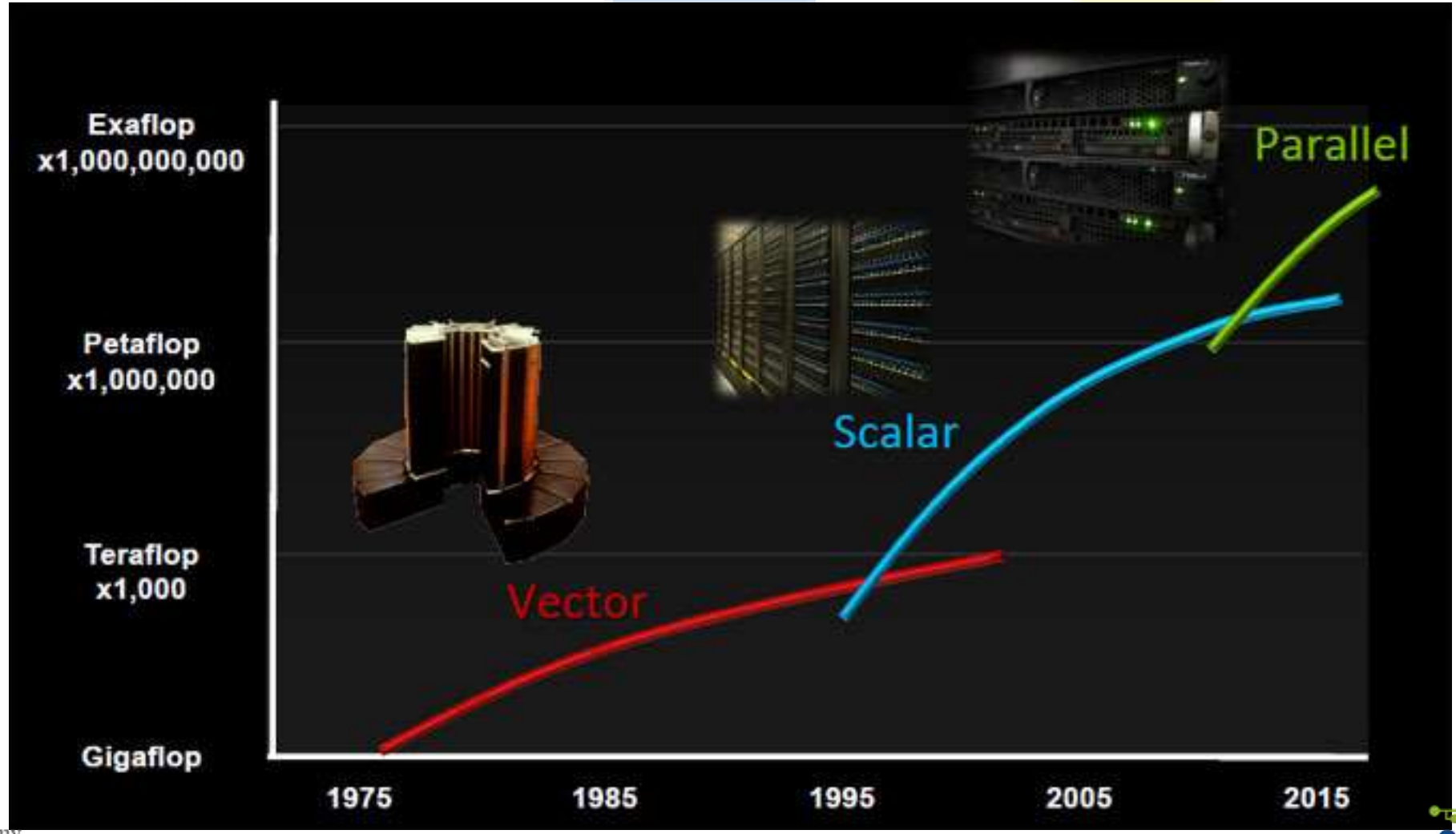
Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

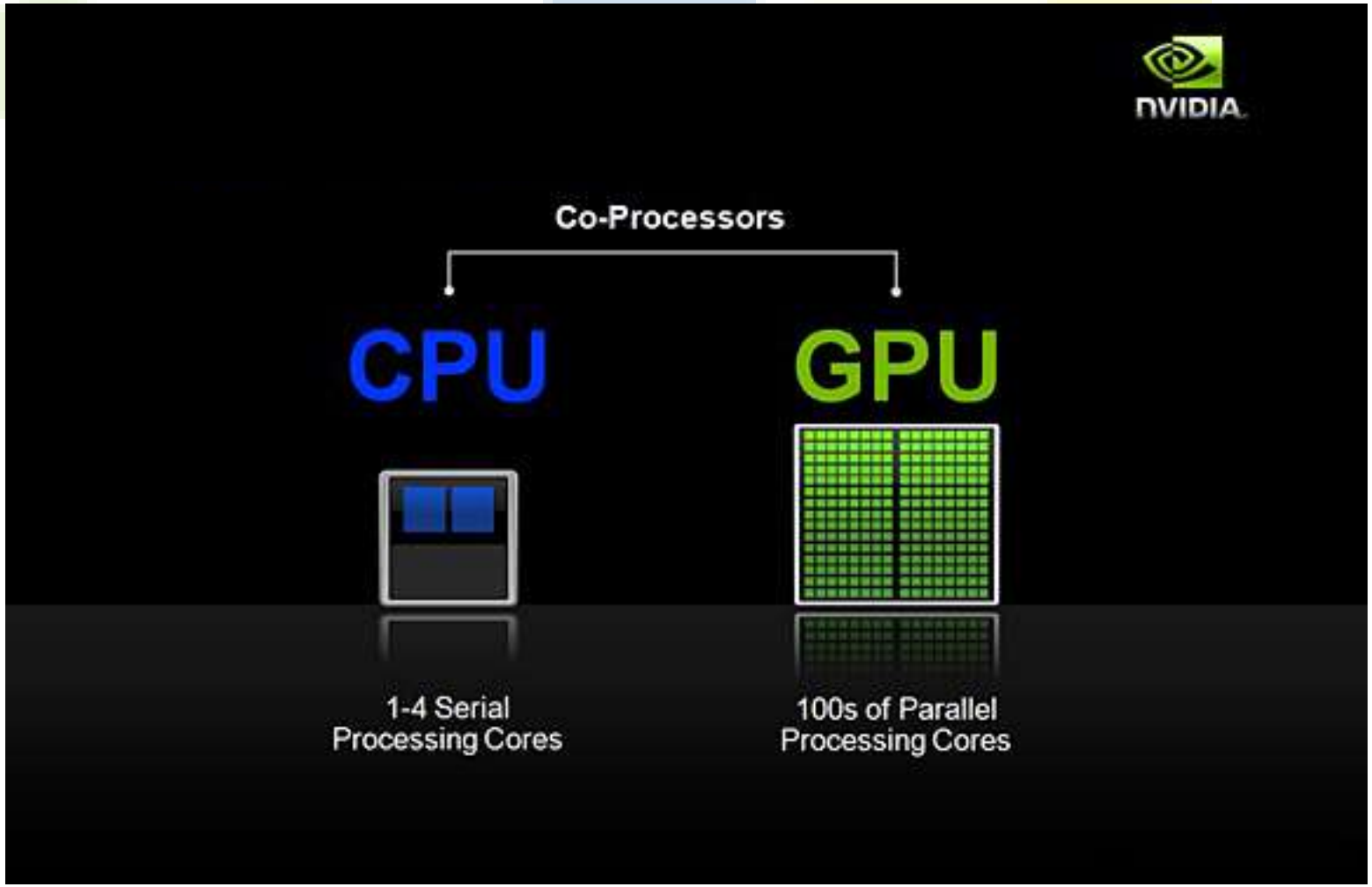
GPGPU



Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU



Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU

Arquitetura Altamente Paralelizada

- Muitas Threads concorrentes (SIMT)
- Throughput maior que as CPUs
- Mais FLOPS (floating-point operations per second)

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU

GPGPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU Frameworks

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU Frameworks

Compute Unified Driver Architecture (CUDA)

- Desenvolvido pela Nvidia
- Extensões para linguagens de programação C e C++
- Suporte para outras linguagens, através de módulos Fortran, PyCUDA, Matlab

Open Computing Language (OpenCL)

- Suporta diversas GPUs (incluindo as Nvidia)
- A forma de realizar computação de alto nível em GPUs ATI/AMD

C++ Accelerated Massive Programming (AMP)

- Superset C++
- Parte do MSVC++ da Microsoft
- Suporta GPUs Nvidia e ATI/AMD

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

GPGPU Frameworks

Alea, Aparapi, Brook

Data Science Academy

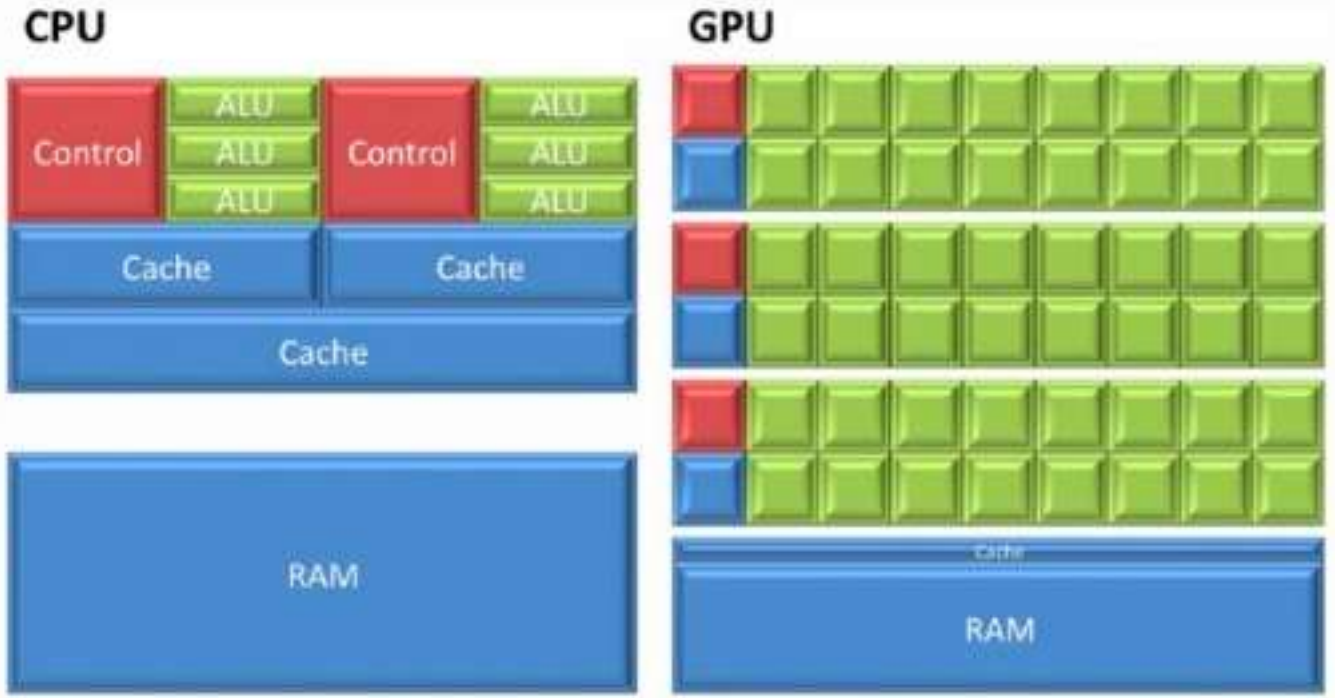
Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Arquitetura da GPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

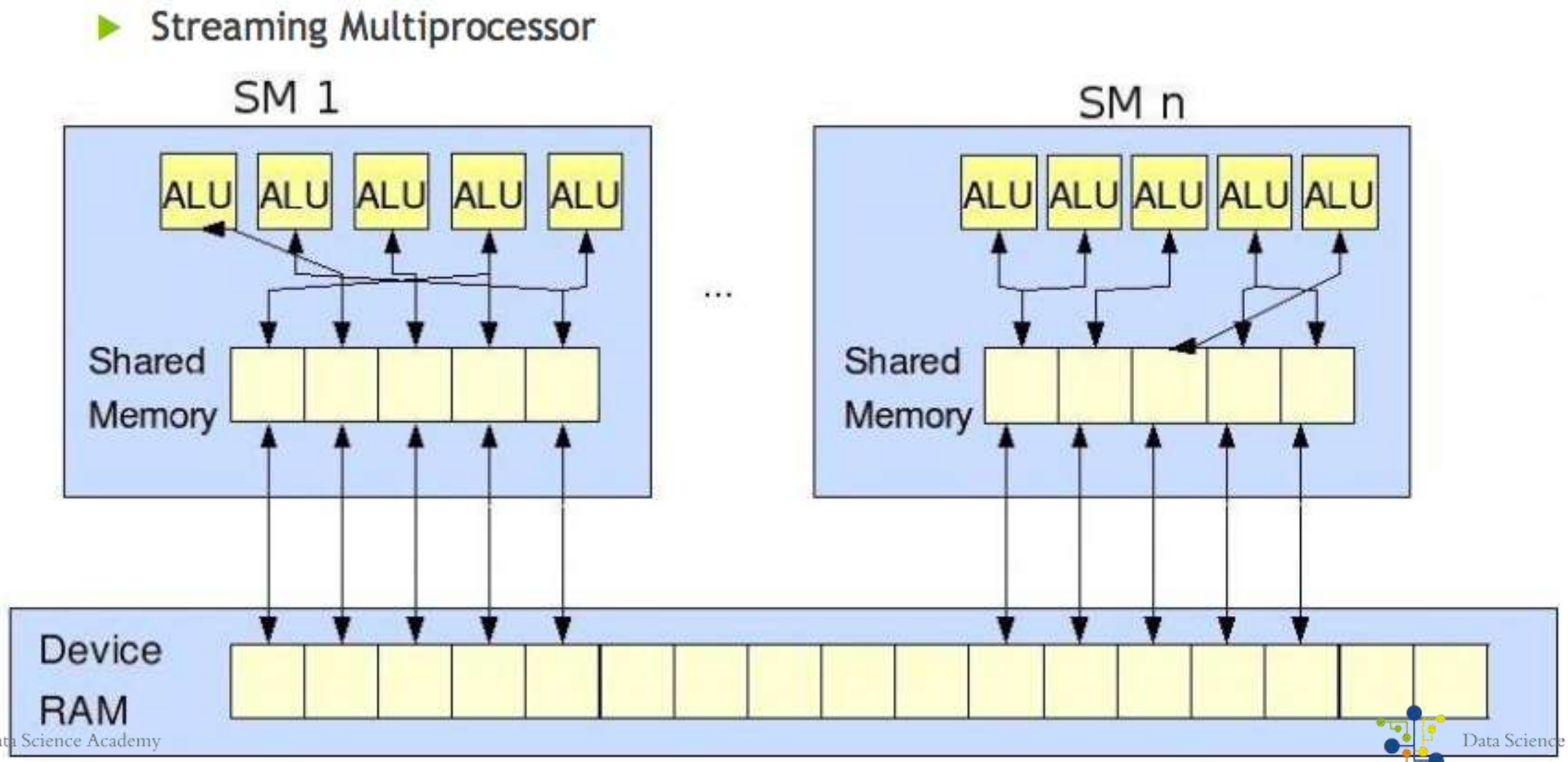
Arquitetura das GPUs



Data Science Academy

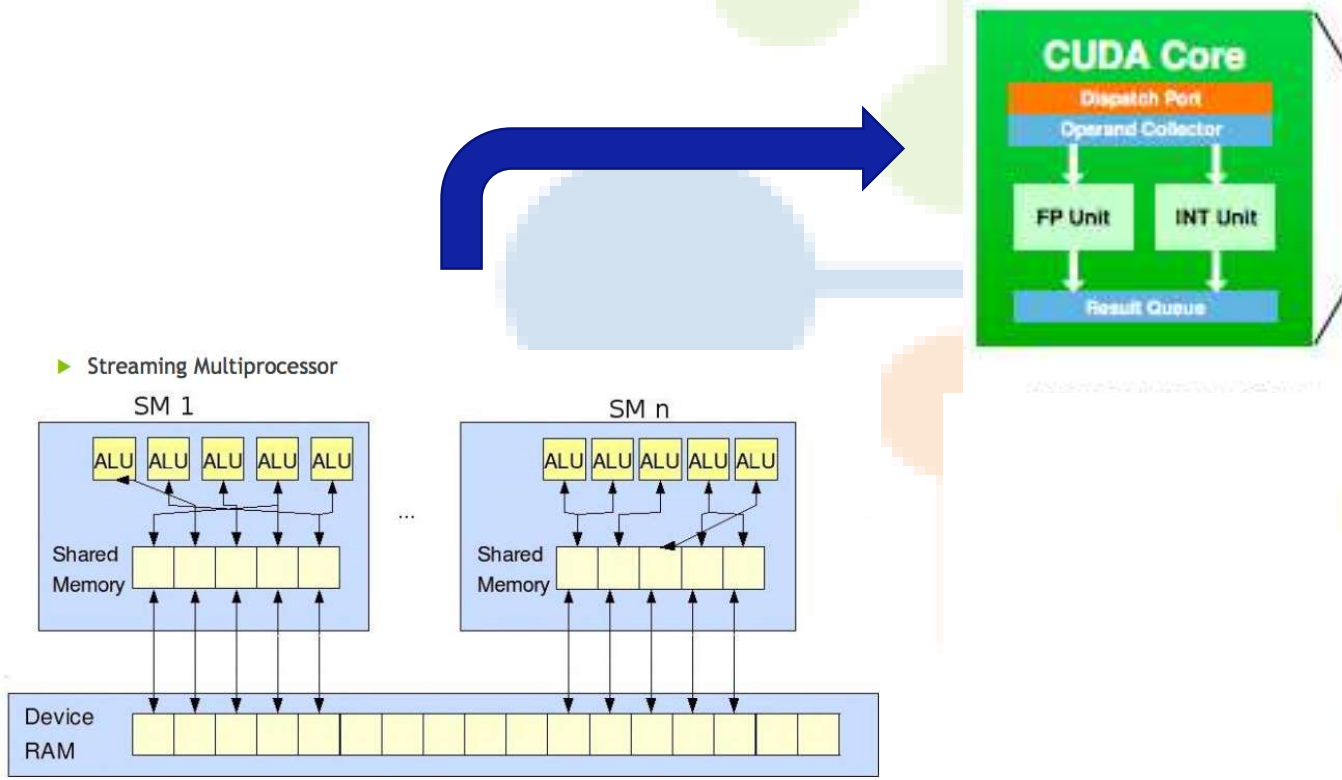
Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

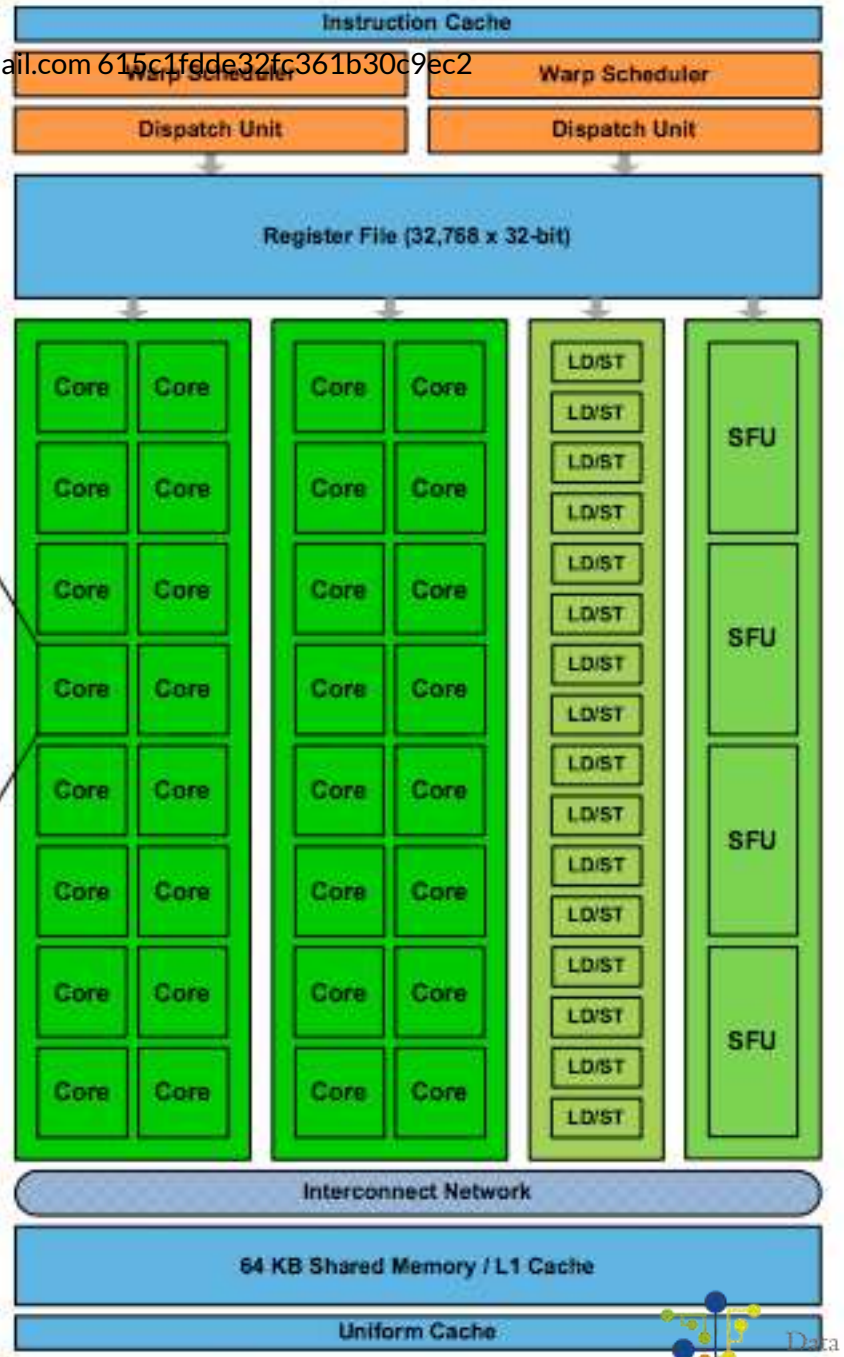
Arquitetura das GPUs



Arquitetura das GPUs

Cada SM possui vários núcleos, chamados de CUDA cores. Cada CUDA core possui pipelines completos de operações aritméticas (ALU - Arithmetic Logic Unit) e de pontos flutuantes (FPU - Floating Point Unit).

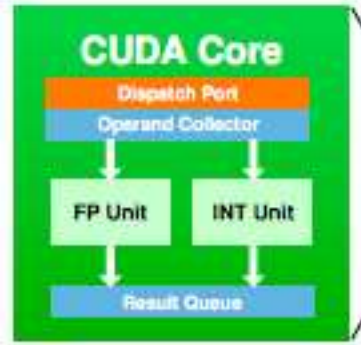


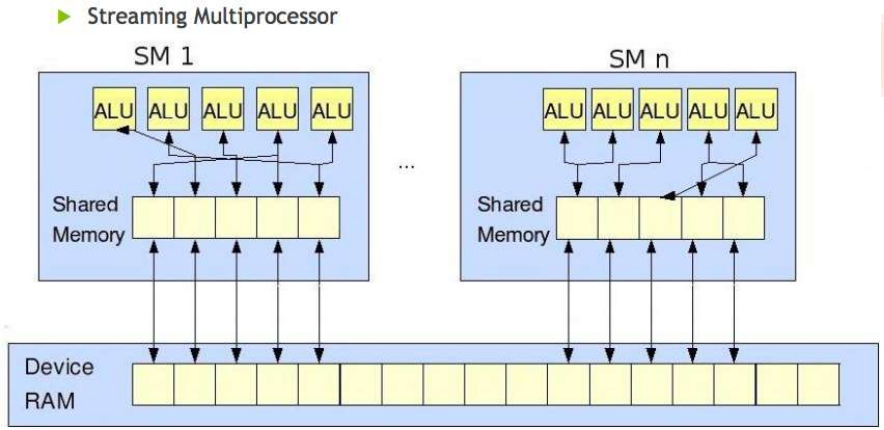


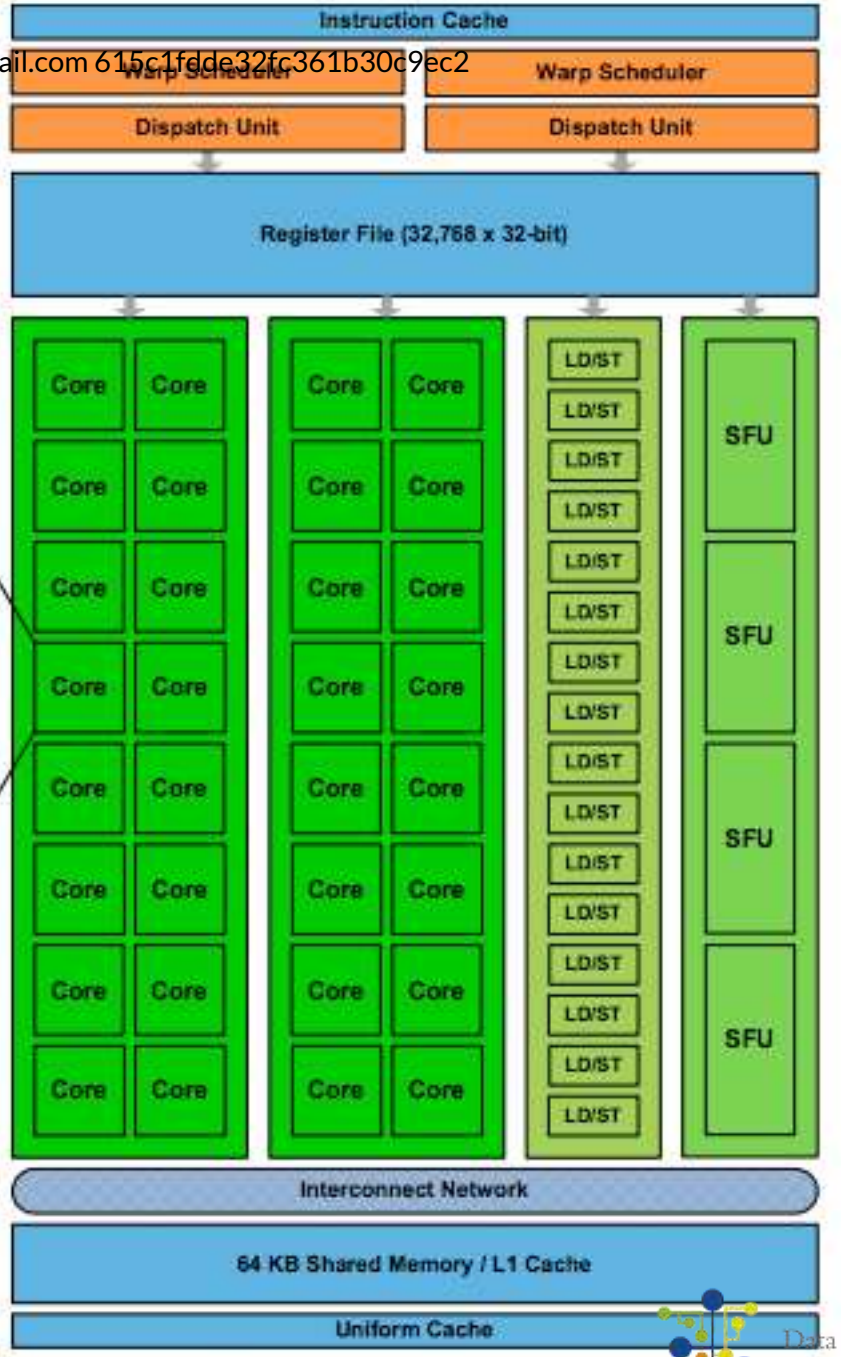
Fermi Streaming Multiprocessor (SM)

Arquitetura das GPUs

Streaming Processor ou Shader Unit







Fermi Streaming Multiprocessor (SM)

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Arquitetura das GPUs

Arquitetura da GPU Fermi

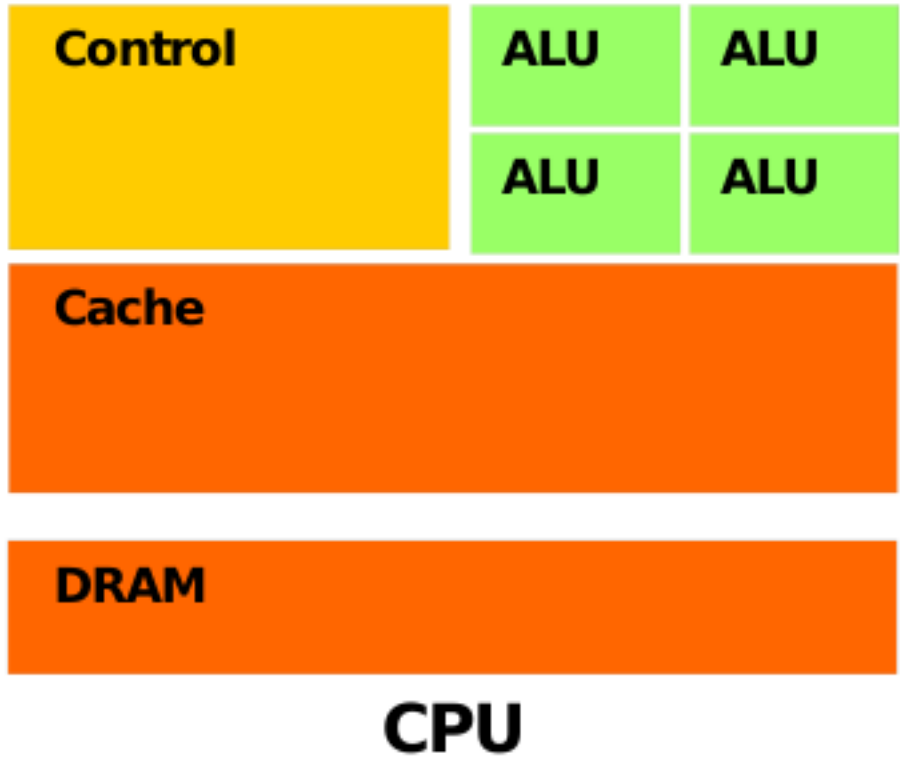
- 16 Multiprocessadores:
- 32 processadores
- 16 unidades Load/Store
- 4 unidades funções especiais
- 64 kB Memória compartilhada
- 32768 registradores
- Cache de constantes e textura

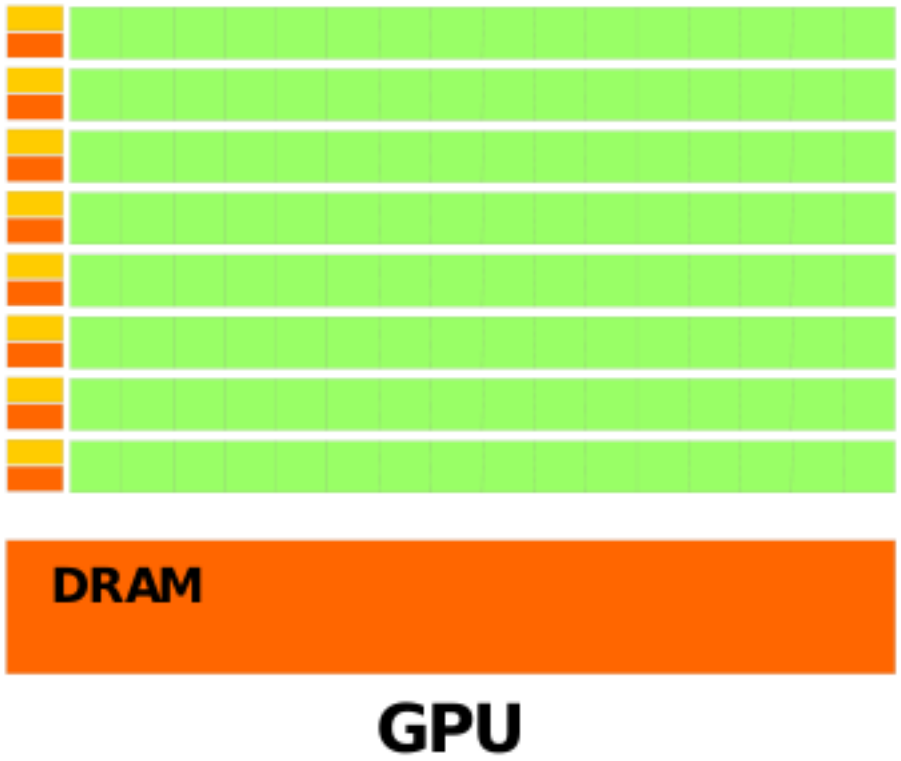
Data Science Academy


PCI Express 3.0 Host Interface
Data Science Academy
Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2
Memory Controller
Memory Controller
Memory Controller
Memory Controller
Memory Controller
Memory Controller
SMX
L2 Cache
Data Science Academy
Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Arquitetura das GPUs

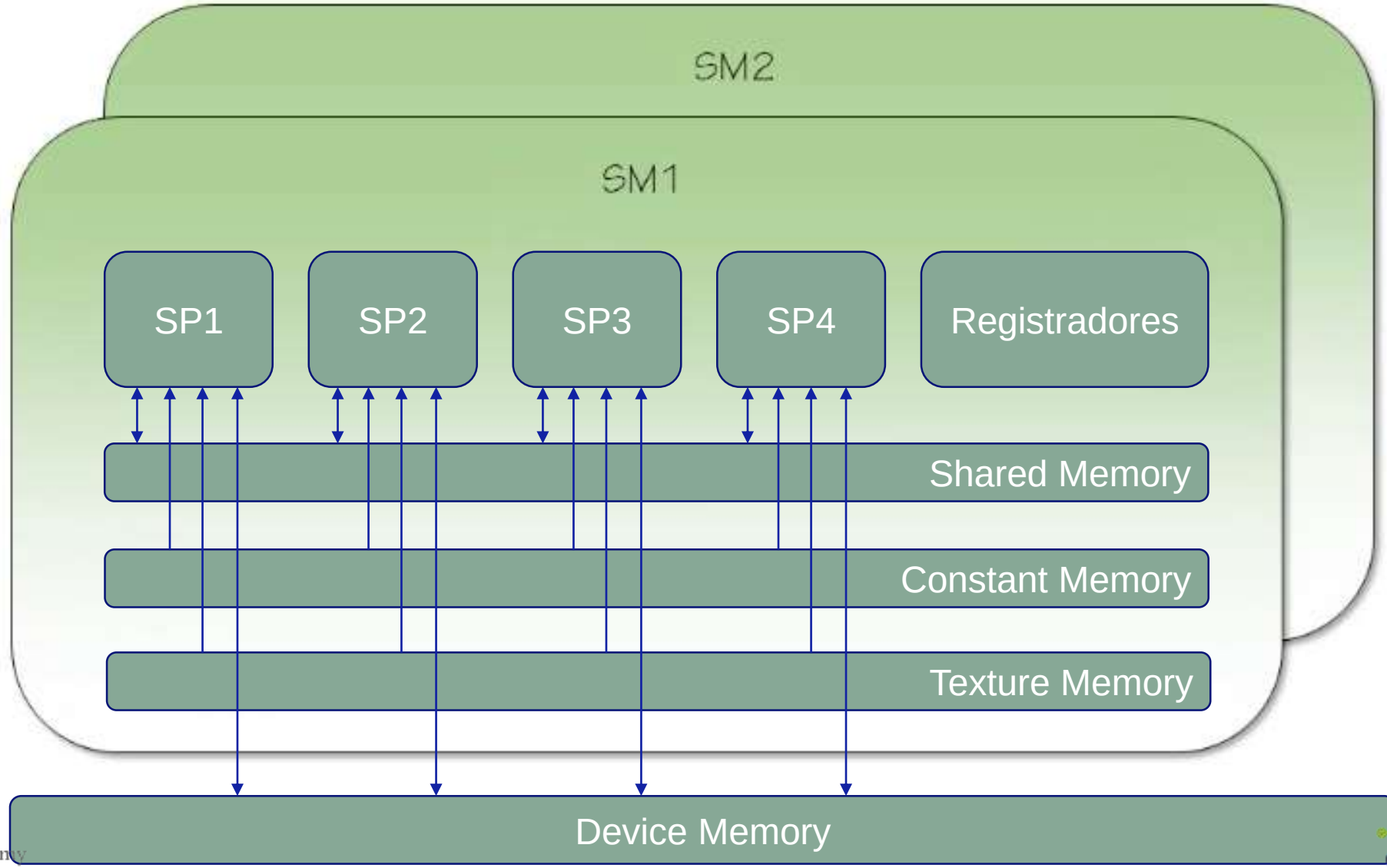




Data Science Academy

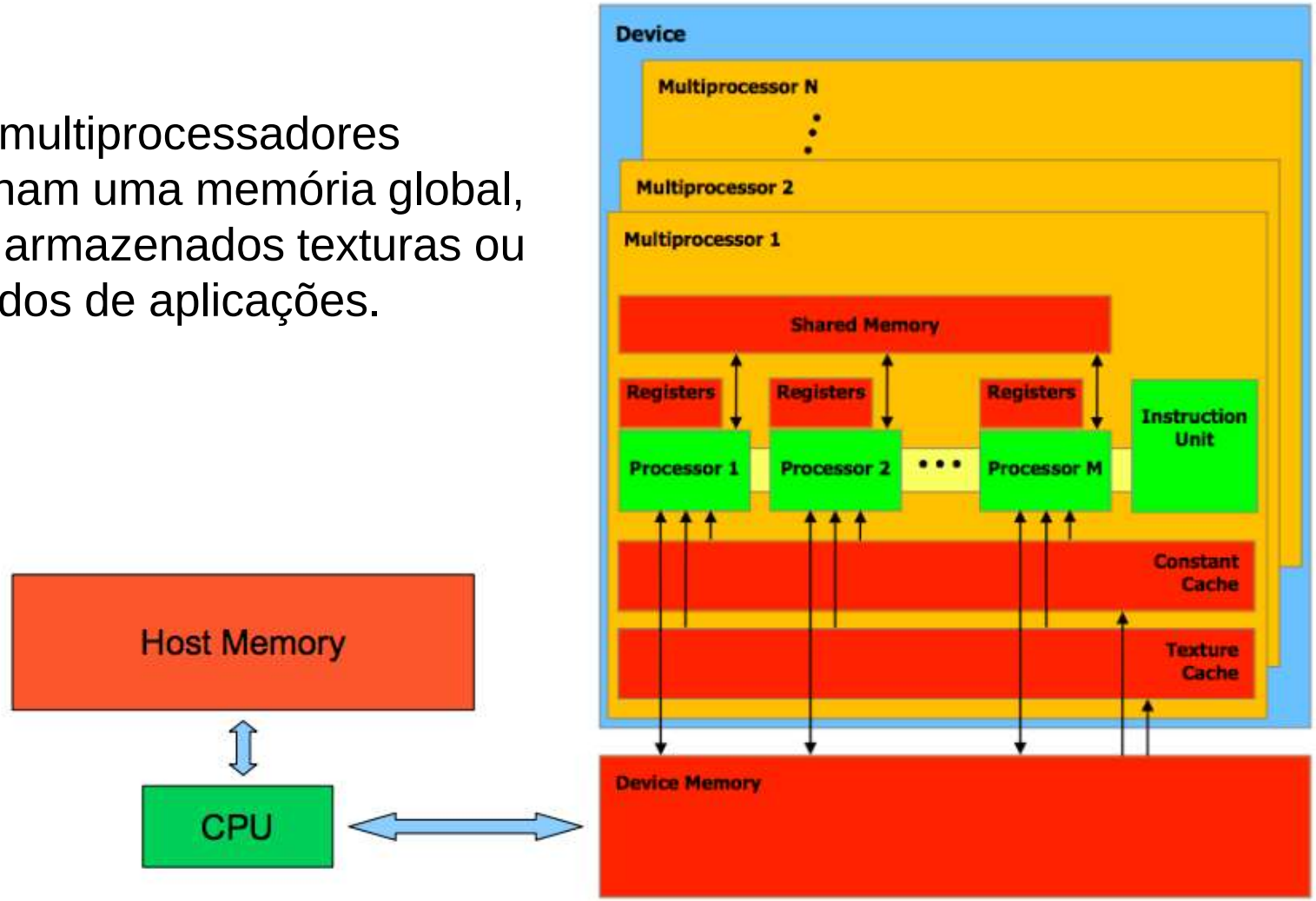
Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Arquitetura das GPUs



Arquitetura das GPUs

Os multiprocessadores compartilham uma memória global, onde são armazenados texturas ou dados de aplicações.



Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Compute Capability

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Compute Capability

CUDA-Enabled GeForce Products

GeForce Desktop Products

GPU	Compute Capability
NVIDIA TITAN Xp	6.1
NVIDIA TITAN X	6.1
GeForce GTX 1080 Ti	6.1
GeForce GTX 1080	6.1
GeForce GTX 1070	6.1
GeForce GTX 1060	6.1
GeForce GTX 1050	6.1
GeForce GTX TITAN X	5.2
GeForce GTX TITAN Z	3.5

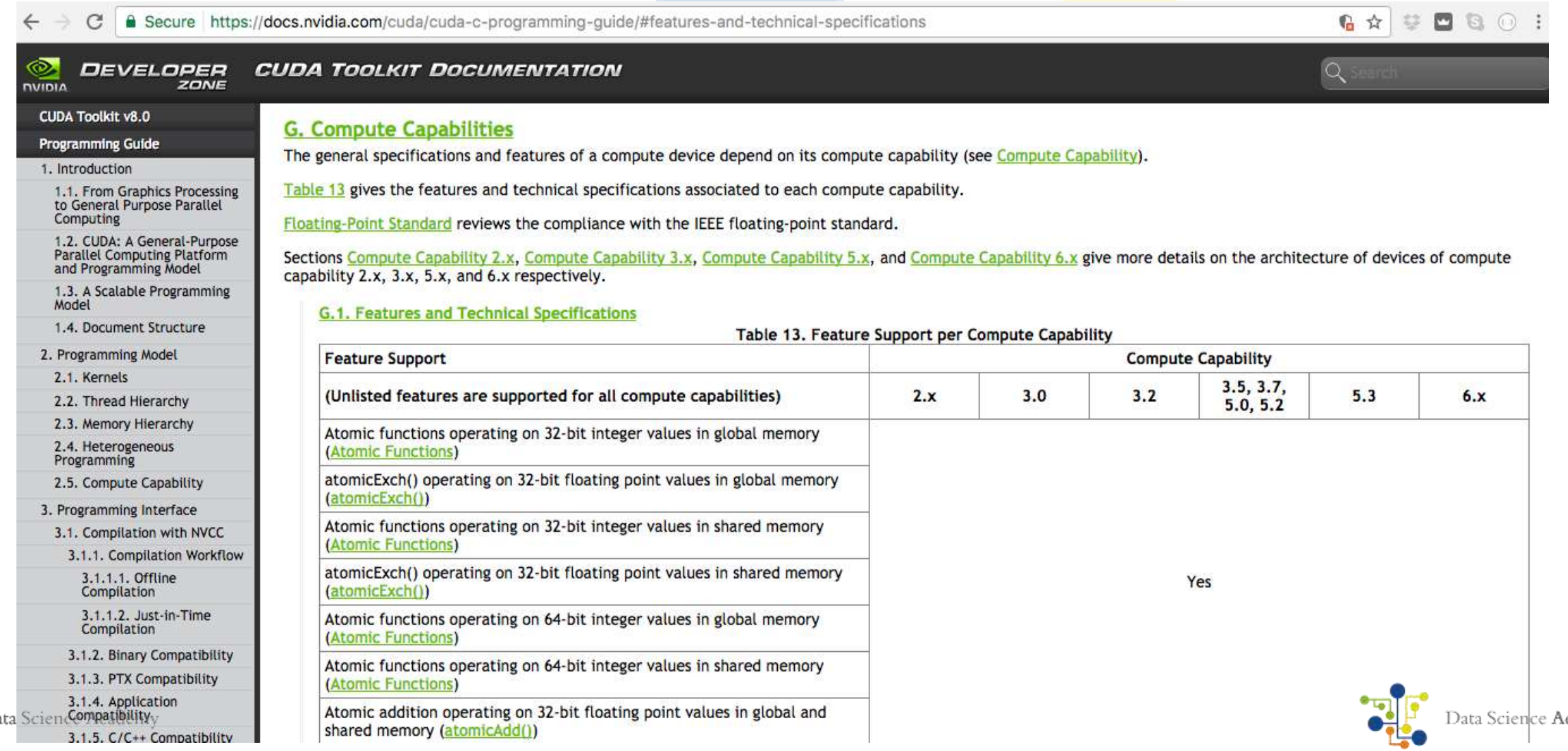
GeForce Notebook Products

GPU	Compute Capability
GeForce GTX 1080	6.1
GeForce GTX 1070	6.1
GeForce GTX 1060	6.1
GeForce GTX 980	5.2
GeForce GTX 980M	5.2
GeForce GTX 970M	5.2
GeForce GTX 965M	5.2
GeForce GTX 960M	5.0
GeForce GTX 950M	5.0

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Compute Capability



Secure https://docs.nvidia.com/cuda/cuda-c-programming-guide/#features-and-technical-specifications

NVIDIA DEVELOPER ZONE CUDA TOOLKIT DOCUMENTATION

G. Compute Capabilities

The general specifications and features of a compute device depend on its compute capability (see Compute Capability).

Table 13 gives the features and technical specifications associated to each compute capability.

Floating-Point Standard reviews the compliance with the IEEE floating-point standard.

Sections Compute Capability 2.x, Compute Capability 3.x, Compute Capability 5.x, and Compute Capability 6.x give more details on the architecture of devices of compute capability 2.x, 3.x, 5.x, and 6.x respectively.

G.1. Features and Technical Specifications

Table 13. Feature Support per Compute Capability

Feature Support	Compute Capability					
(Unlisted features are supported for all compute capabilities)	2.x	3.0	3.2	3.5, 3.7, 5.0, 5.2	5.3	6.x
Atomic functions operating on 32-bit integer values in global memory (Atomic Functions)	Yes					
atomicExch() operating on 32-bit floating point values in global memory (atomicExch())						
Atomic functions operating on 32-bit integer values in shared memory (Atomic Functions)						
atomicExch() operating on 32-bit floating point values in shared memory (atomicExch())						
Atomic functions operating on 64-bit integer values in global memory (Atomic Functions)						
Atomic functions operating on 64-bit integer values in shared memory (Atomic Functions)						
Atomic addition operating on 32-bit floating point values in global and shared memory (atomicAdd())						

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Compute Capability

Compute Capability é um número que indica o que a GPU pode fazer:

- Range atual: 1.0, 1.1, 1.2, 1.3, 2.0, 2.1, 3.0, 3.5, 3.7, 5.0, 5.3, 6.0, 6.1, 6.2, 7.0
- Titan X e GTX 1080 TI possuem Compute Capability igual a 6.1
- Tesla Volta (nova arquitetura anunciada em Maio/2017) possui Compute Capability igual a 7.0

Compute Capability afeta o suporte ao hardware a APIs:

- Número de CUDA cores por SM
- Máximo de registradores 32 bits por SM
- Máximo de instruções por kernel
- Suporte a operações de precisão dupla

Data Science Academy

Como Escolher uma GPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Como Escolher uma GPU

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Como Escolher uma GPU

- Especificações de Performance (FLOPS)
- Precisão Single x Double
- Total de Memória da GPU
- Compute Capability
- Arquitetura
- GPUs Simples / GPUs Workstation / GPUs Servidores (Tesla)
- Certifique-se que sua Motherboard/PSU suportam a GPU
- Mais de uma GPU?

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Como Escolher uma GPU

MacBook Pro (late 2016)
Radeon Pro 455

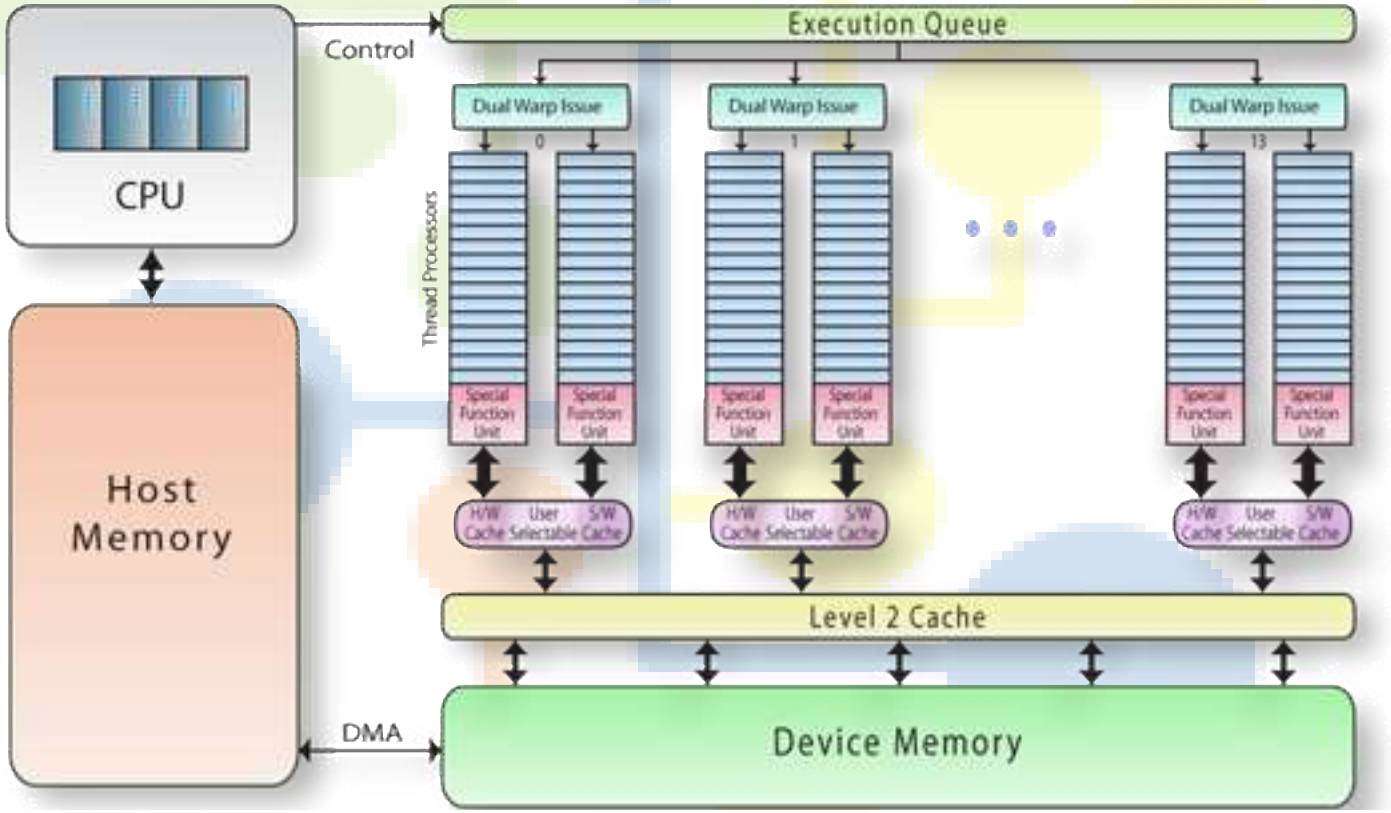
Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Modelo de Programação em GPU

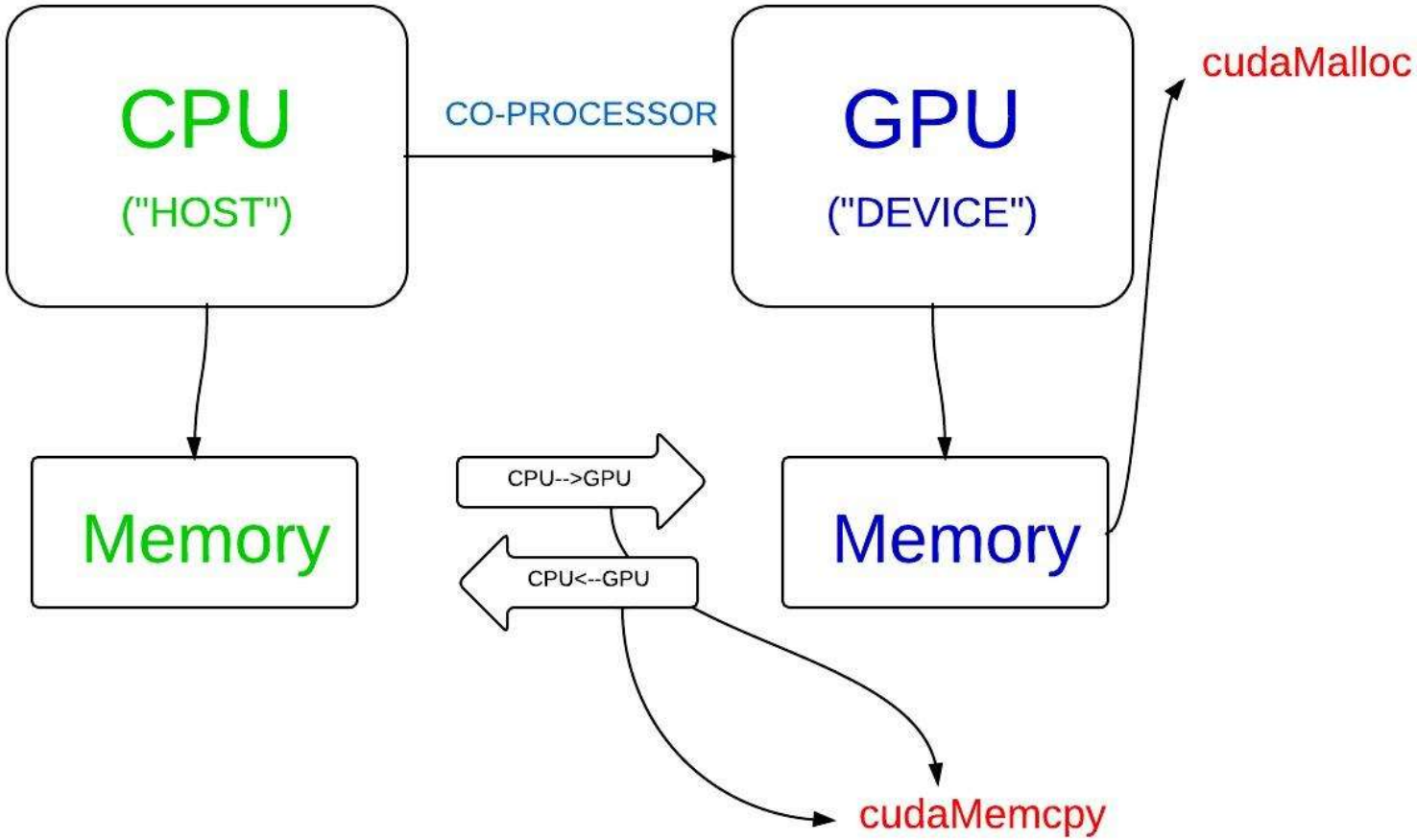
Data Science Academy

Modelo de Programação em GPU



Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

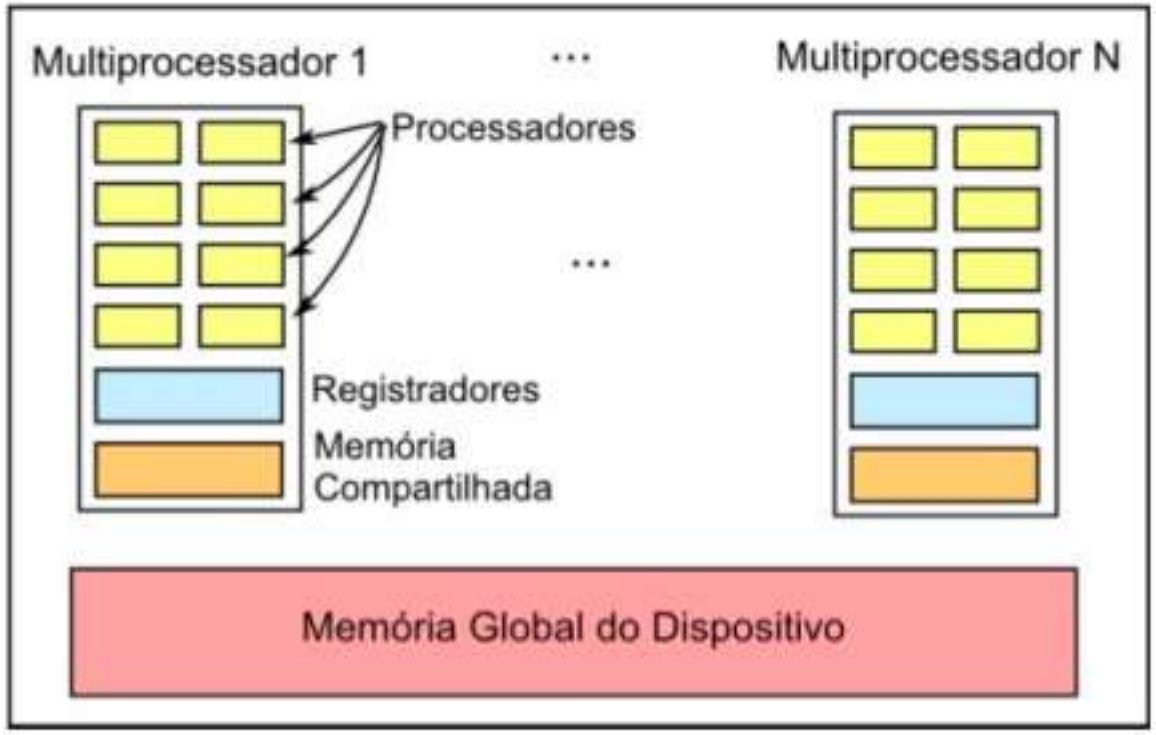
Modelo de Programação em GPU



Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Modelo de Programação em GPU



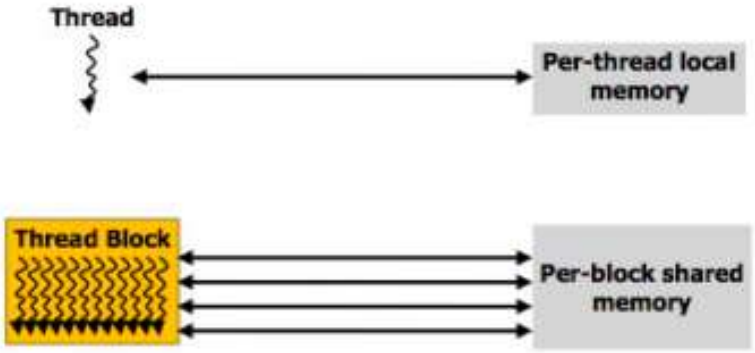
Em uma GPU, 7 operações de ponto flutuante podem ser executadas por um núcleo de processamento ao mesmo tempo que 1 byte demora para ser transferido da memoria externa para a GPU

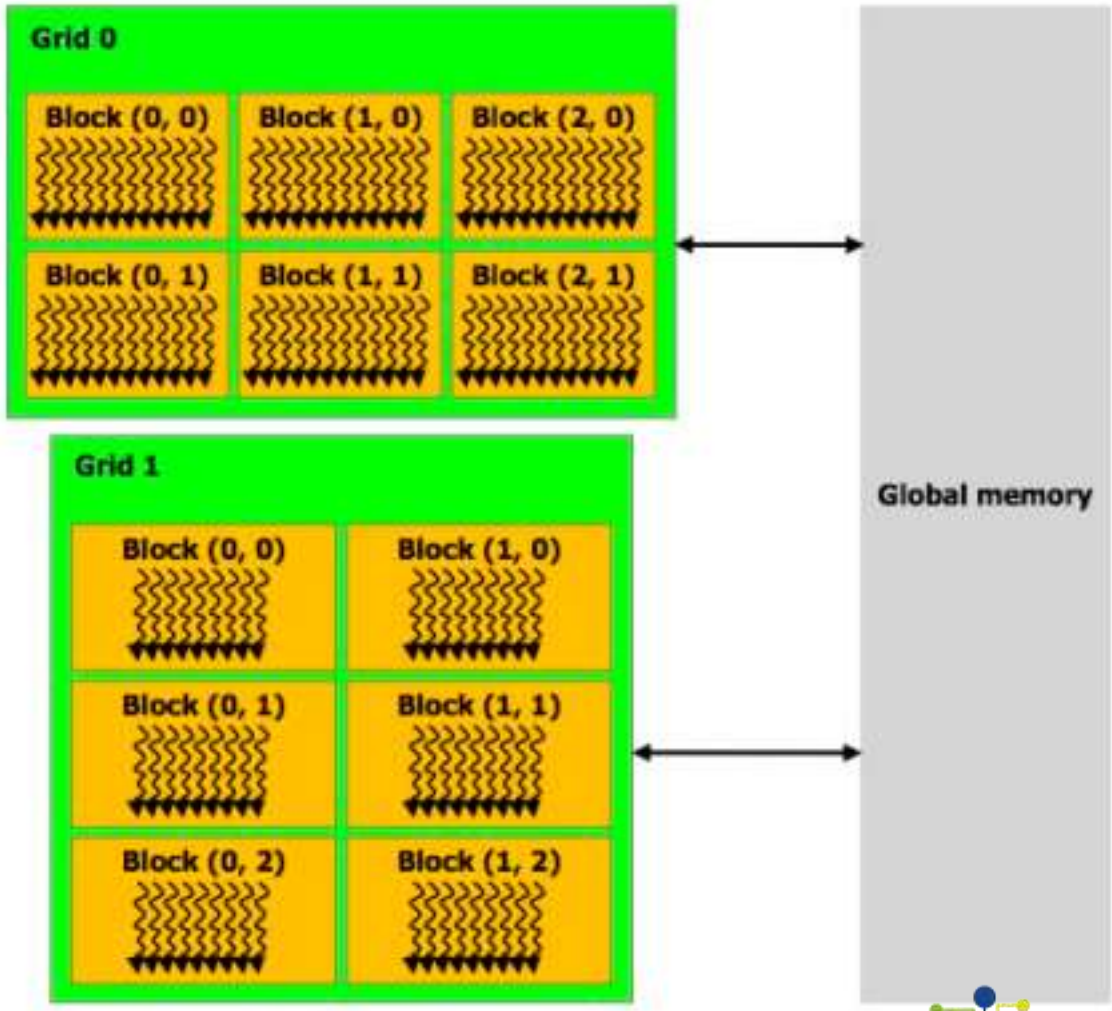
Data Science Academy

Modelo de Programação em GPU

Cada execução do kernel é composta por:

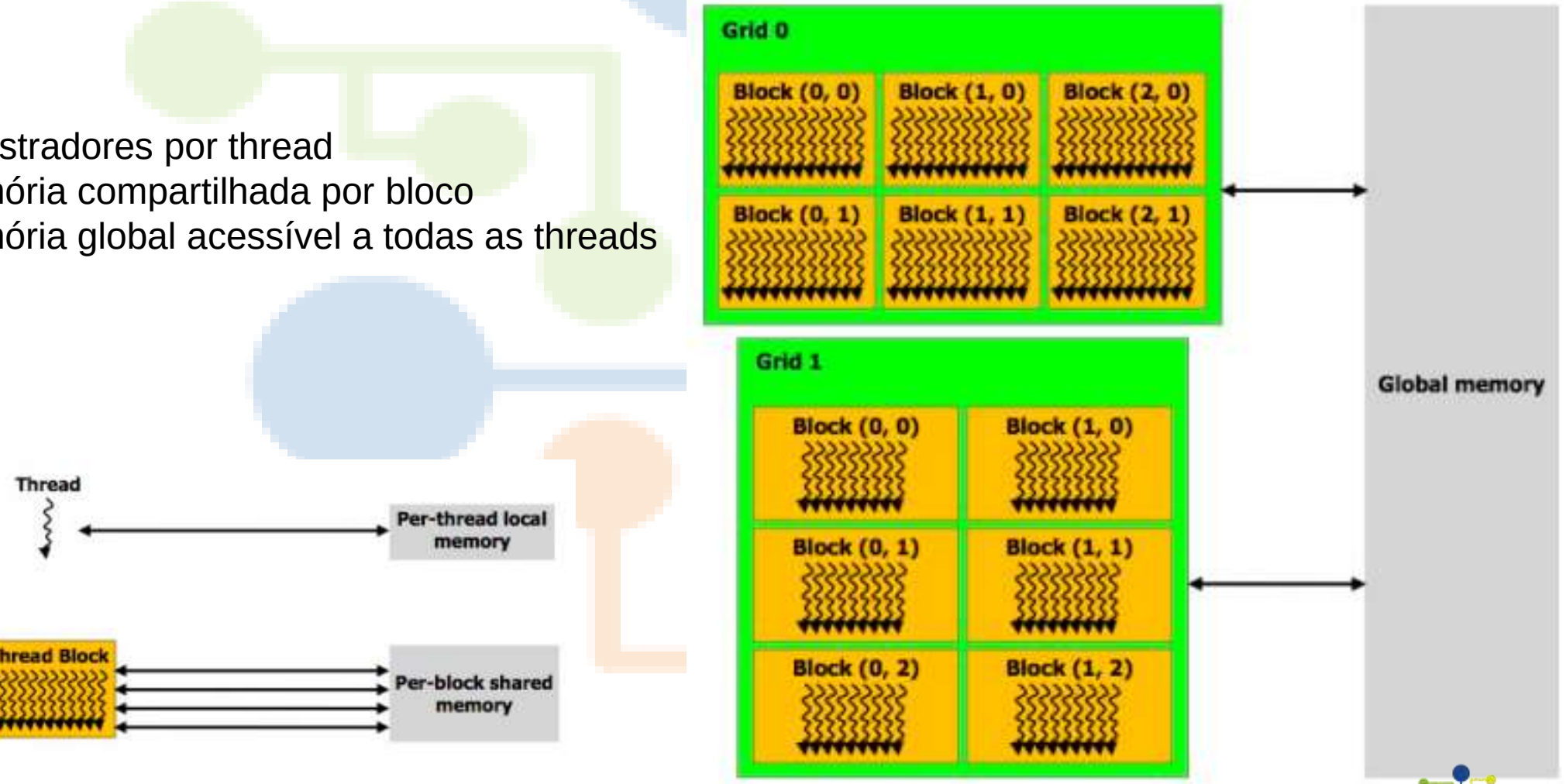
Grids → Blocks→ Threads





Modelo de Programação em GPU

- Registradores por thread
- Memória compartilhada por bloco
- Memória global acessível a todas as threads



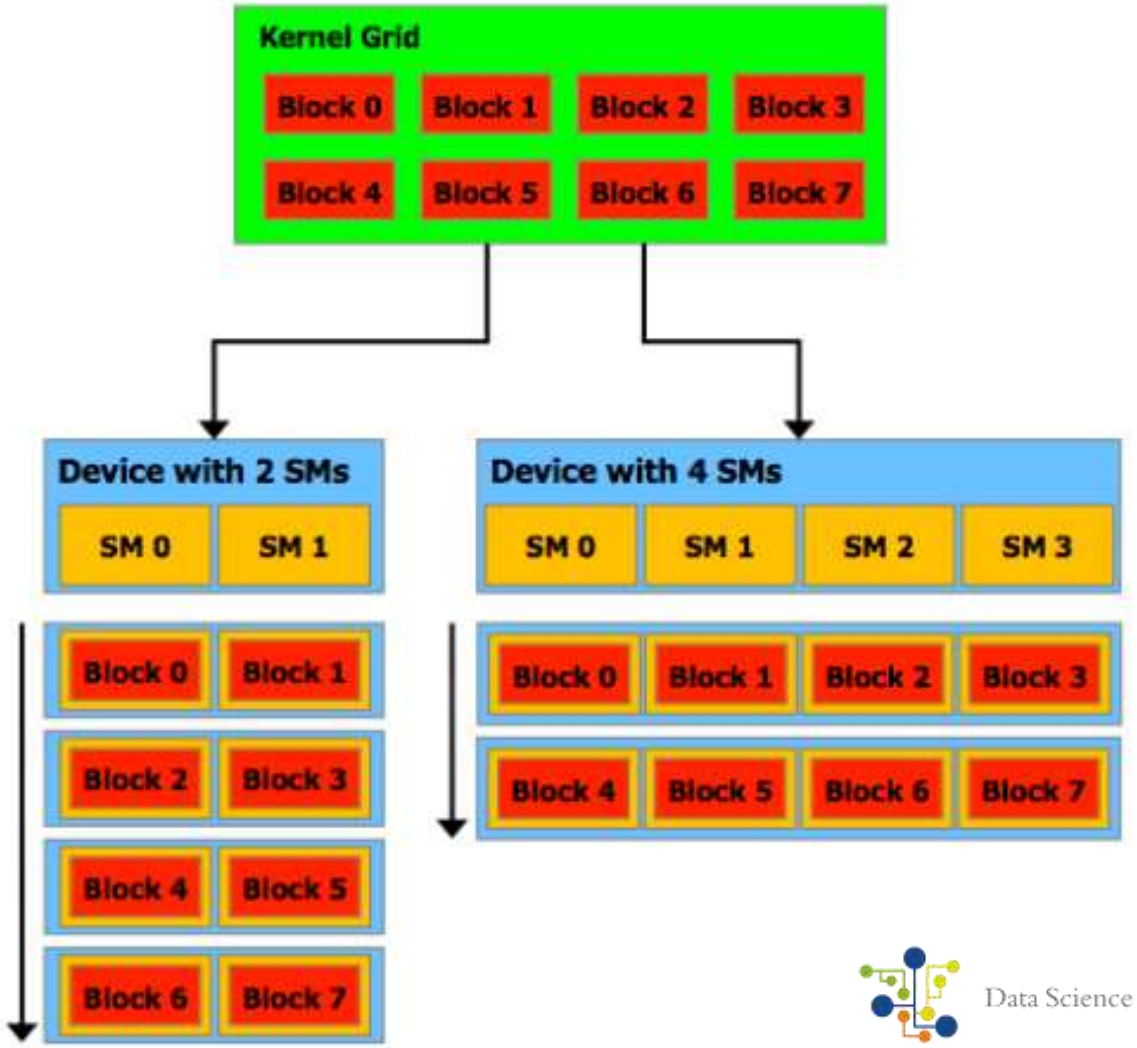
Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Modelo de Programação em GPU

E como ocorre a execução de aplicações?

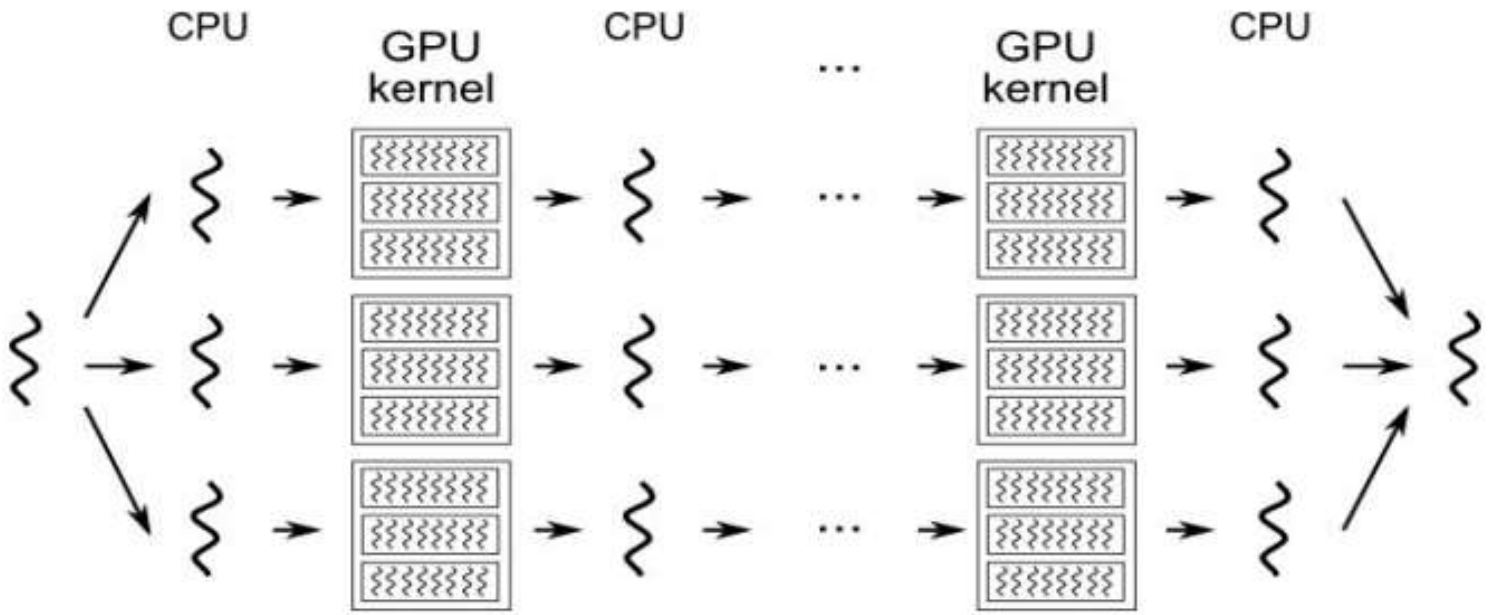
Cada bloco é alocado a um multiprocessador (Streaming Multiprocessor) da GPU, que pode executar 1 ou mais blocos simultaneamente.

Cada multiprocessador executa 16 threads no modelo SIMT: Single Intruction, Multiple Thread



Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Modelo de Programação em GPU



A execução do programa controlado pela CPU pode lançar kernels, que são trechos de código executados em paralelo por múltiplas threads na GPU.

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Problemas em Programação Paralela
Deadlock e Starvation

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Deadlock e Starvation

Deadlock

A espera B e B espera A e ambos esperam um ao outro

Starvation

A espera B mas nunca recebe uma resposta

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Deadlock e Starvation

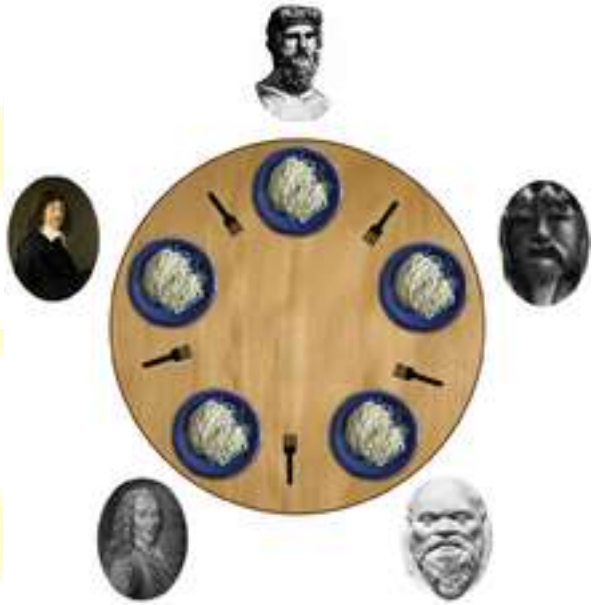
Deadlock causa Starvation, mas Starvation não causa Deadlock

Data Science Academy

Data Science Academy raphaelbsfontenelle@gmail.com 615c1fdde32fc361b30c9ec2

Deadlock e Starvation

Deadlock



Dinning Philosophers Problem
Starvation

Data Science Academy

Obrigado

Data Science Academy