Formação Inteligência Artificial

Data Science Academy

Programação Paralela em GPU

Programação Paralela em CUDA
Parte 2

Data Science Academy

Programação Paralela em CUDA - Parte 2

Data Science Academy

Qualificadores

Data Science Academy

Qualificadores

__global__	__device__	__host__
Define um Kernel Executa na **GPU**, sendo chamado pela **CPU**	Pode ser usado para variáveis Executa na **GPU**, sendo chamado pela **GPU**	Para tarefas na CPU Executa na **CPU**, sendo chamado pela **CPU**

Data Science Academy

Qualificadores

Os qualificadores também podem ser combinados em uma única instrução:

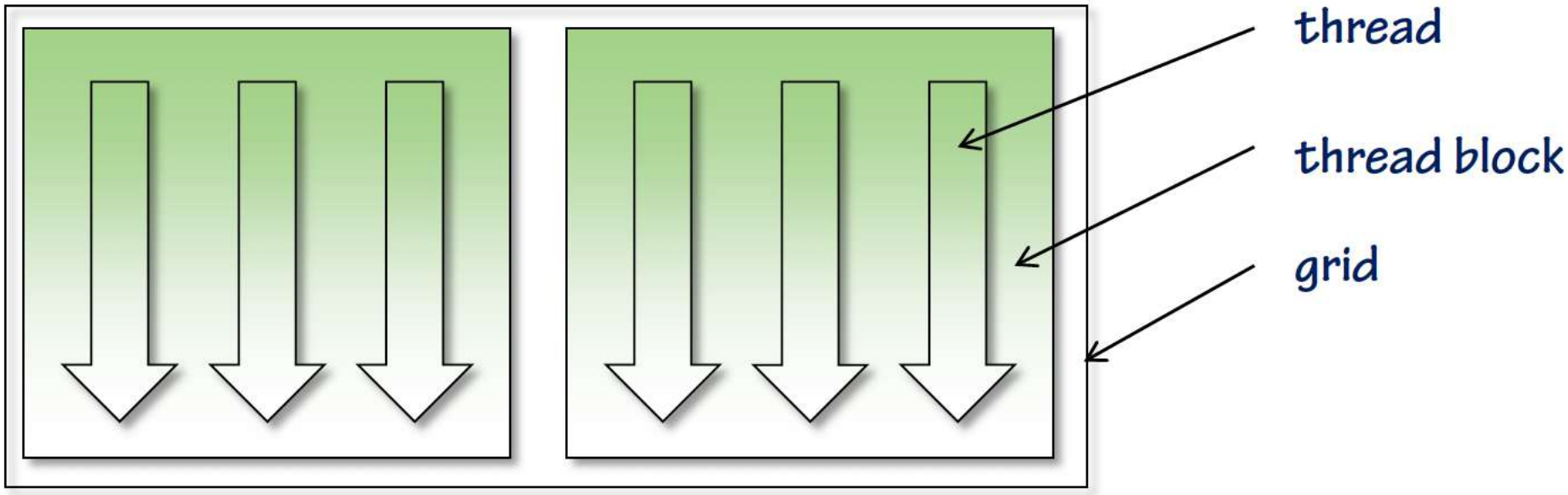
__host__ __device__ nome_func()

Data Science Academy

Modelo de Execução

Data Science Academy

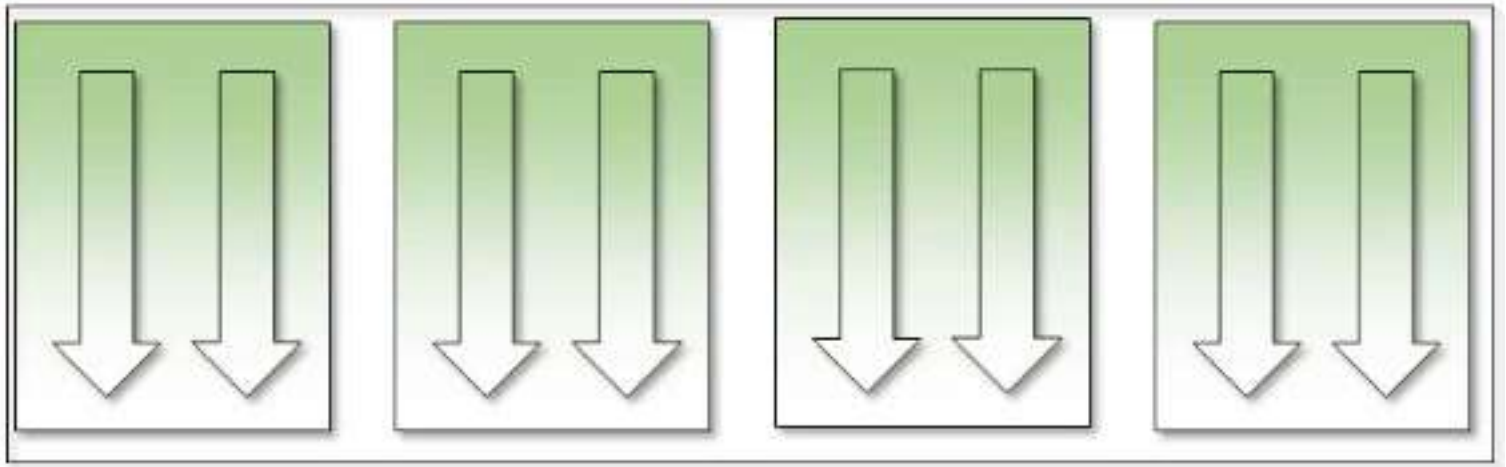
Modelo de Execução



addArraysGPU <<< 1, count >>> (da, db, dc)

Data Science Academy

Modelo de Execução



addArraysGPU <<< 4, 2 >>> (da, db, dc)

Data Science Academy

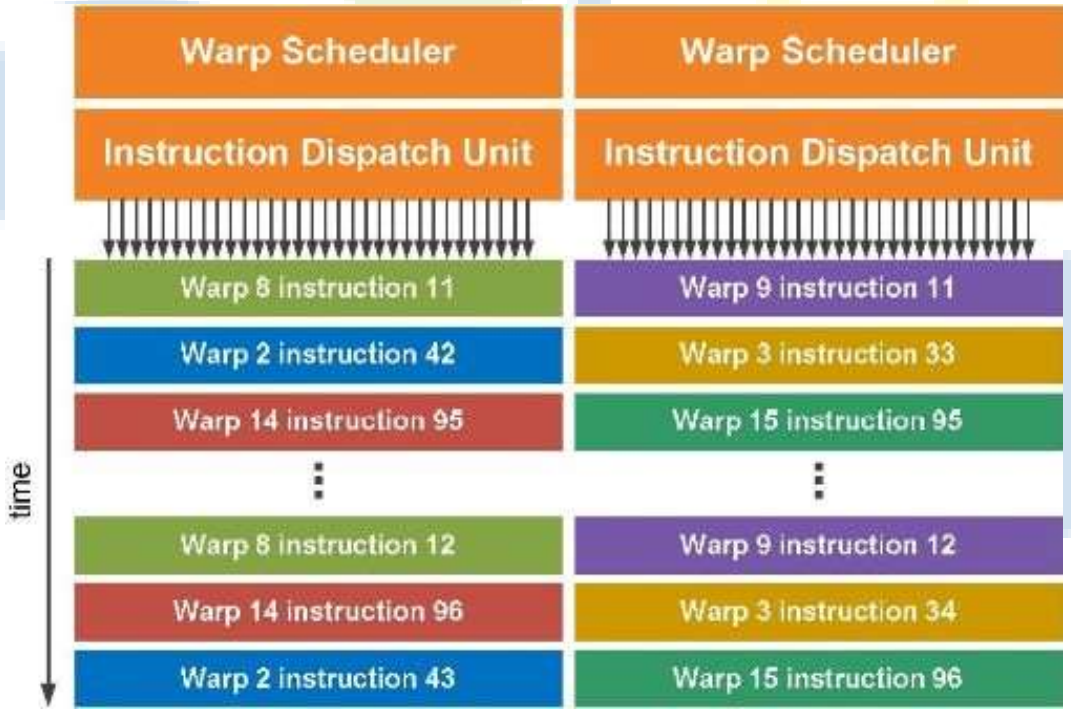
Modelo de Execução

- Os blocos de threads estão programados para serem executados nos SMs disponíveis.
- Cada SM executa um bloco por vez.

Modelo de Execução

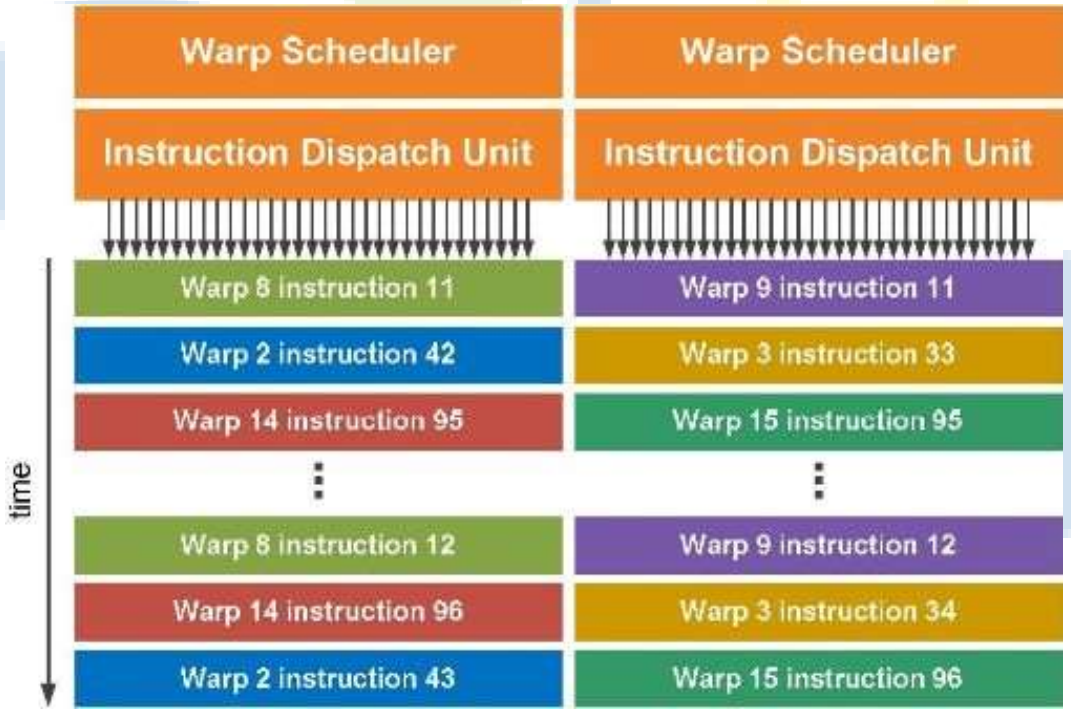
- Cada bloco de thread é dividido em warps (a variável warp size da GPU define quantos warps podem ser usados).

- Os warps são executados de forma paralela (o Warp Watch nos permite visualizar a execução dos warps).



Modelo de Execução

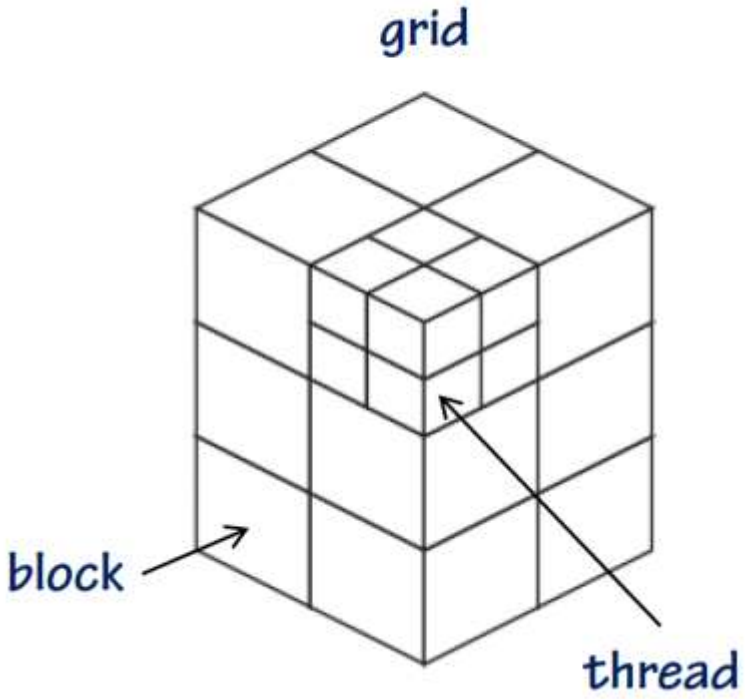
- Se um warp leva muito tempo para sua execução, o scheduler da GPU pode iniciar um novo warp.
- O número de threads por warps depende da capacidade de cálculo (Compute Capability).
- Todas as warps são tratadas em paralelo.



Dimensões de Grids e Blocos

Data Science Academy

Dimensões de Grids e Blocos



Definimos execução como:

<<< a, b >>>

Um grid de **a** blocos com **b** threads

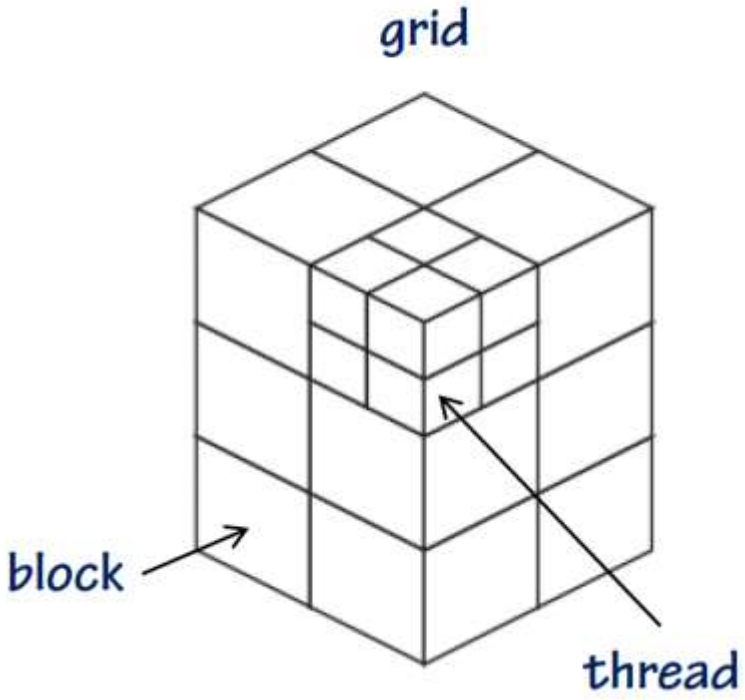
<<< a, b >>>

(a, 1, 1) blocos por (b, 1, 1) threads

Estrutura dim3

Data Science Academy

Dimensões de Grids e Blocos



Definimos execução como:

<<< a, b >>>

Um grid de **a** blocos com **b** threads

Um grid é uma estrutura 3D

Definimos como (A x B x C) blocos de (x x y x z) threads

Dimensões de Grids e Blocos

Parâmetros de execução e posição atual:

- BlockIdx
  - Onde estamos no grid
- GridDim
  - O tamanho do grid
- ThreadIdx
  - Posição da therad atual no bloco de thread
- BlockDim
  - Tamanho do bloco de thread
- Limitações
  - Tamanhos de grid e bloco
  - Número de threads

MAX_BLOCK_DIM_X	512
MAX_BLOCK_DIM_Y	512
MAX_BLOCK_DIM_Z	64
MAX_GRID_DIM_X	65535
MAX_GRID_DIM_Y	65535
MAX_GRID_DIM_Z	1

MAX_THREADS_PER_BLOCK	512
MAX_THREADS_PER_MULTIPROCESSOR	1024

Tratamento de Erros

Data Science Academy

Tratamento de Erros

- CUDA não gera mensagem de erro
  - Falha silenciosa

- As funções retornam cudaError_t
  - Verifique status cudaSuccess
  - Para o código de erro use cudaGetErrorString()

Obrigado

Data Science Academy