Otimização de Memória em GPU

Stencil, Tiling e Shared Memory

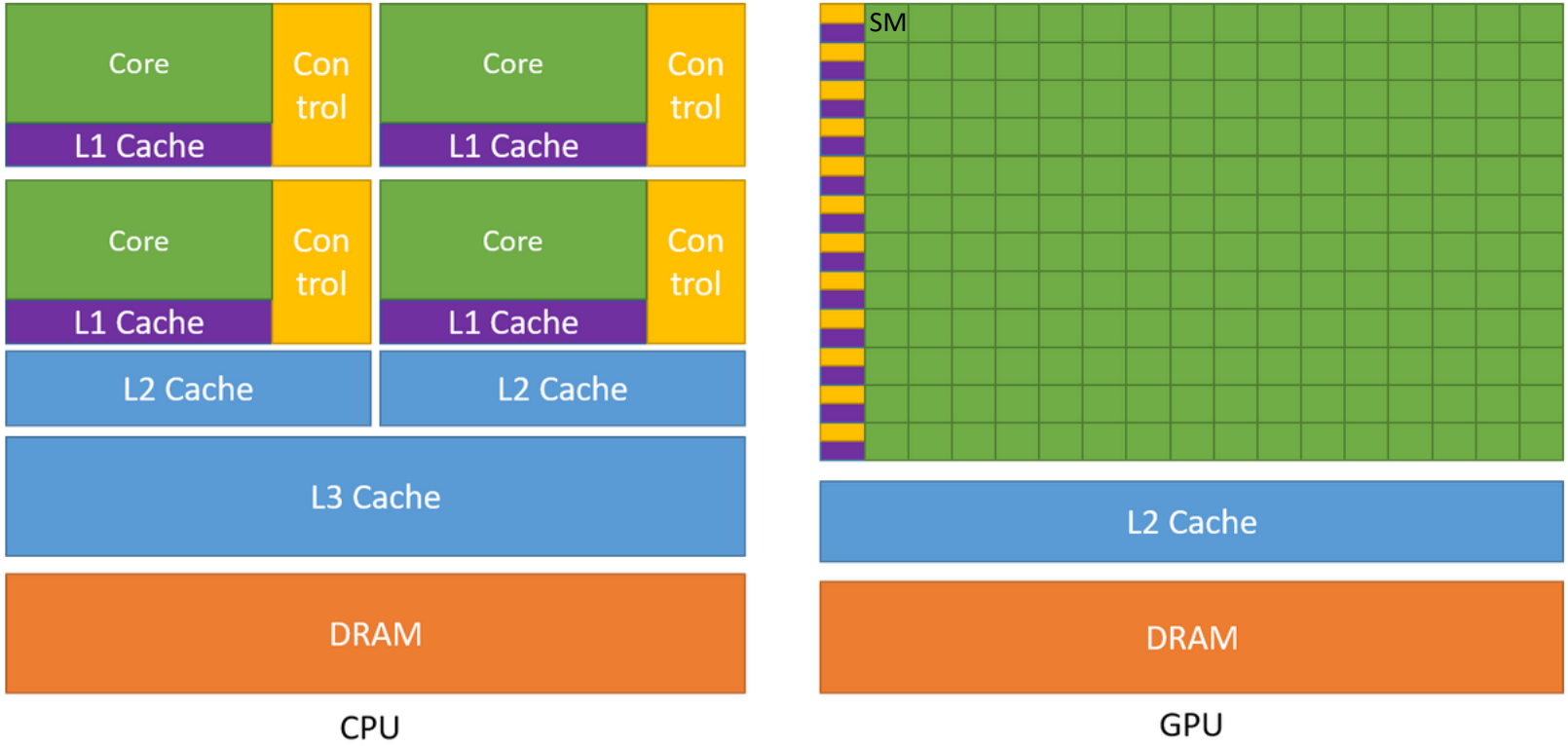
Engenharia de Computação
Lícia Sales Costa Lima
2026-1

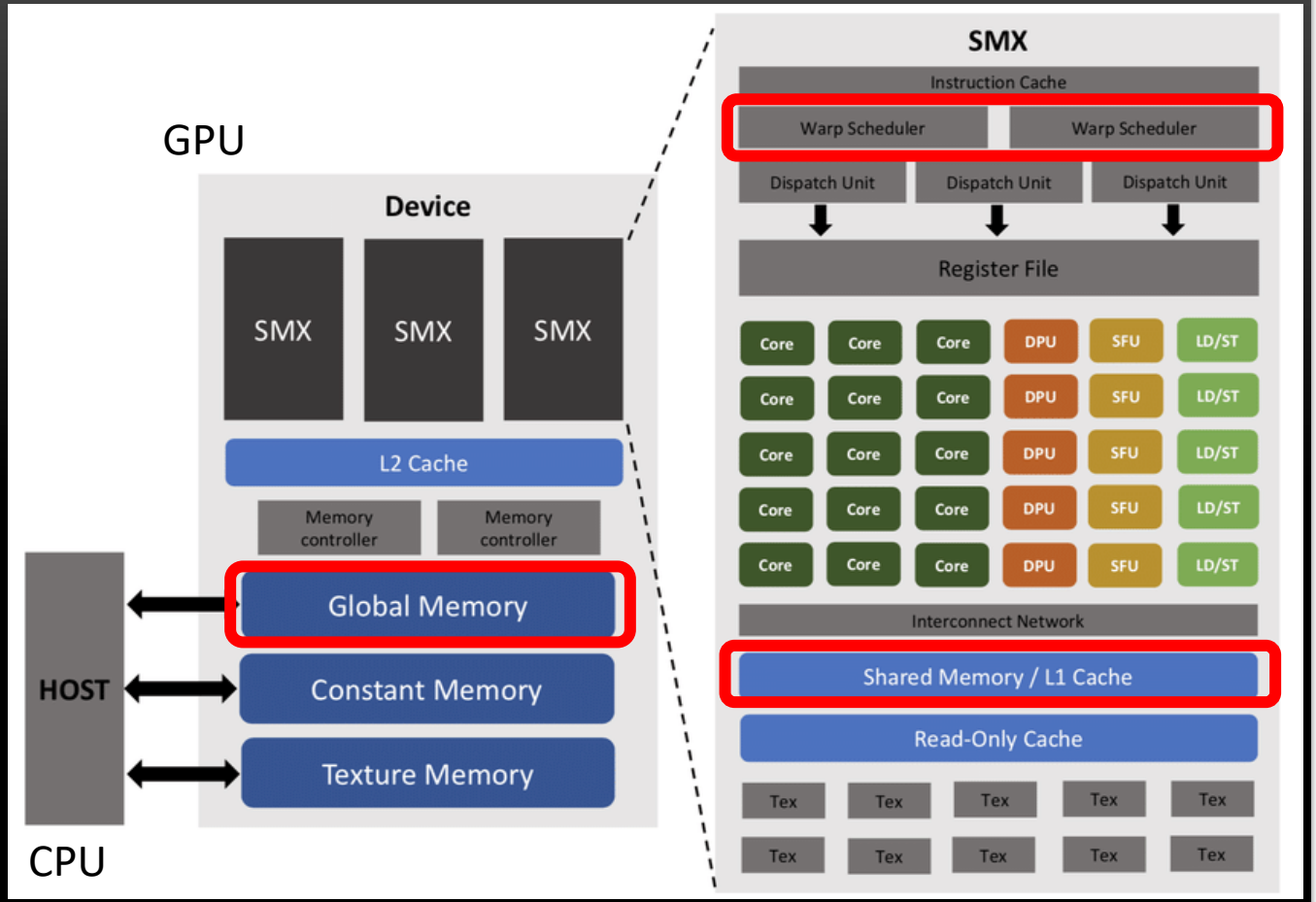
Objetivos da Aula

- Compreender a organização de hardware das GPUs NVIDIA (SMs, Cores, Warps).
- Entender a hierarquia de threads em CUDA (Grids, Blocos, Threads).
- Analisar o padrão de computação Stencil e suas aplicações.
- Implementar a técnica de Tiling com Memória Compartilhada para otimização.

Olhando para o Hardware

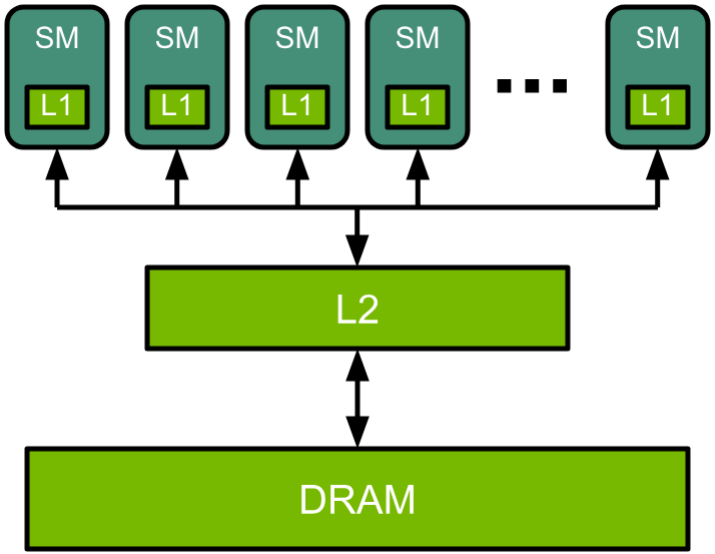
CPU x GPU


Core
Con trol
L1 Cache
Core
Con trol
L1 Cache
Core
Con trol
L1 Cache
Core
Con trol
L1 Cache
L2 Cache
L2 Cache
L3 Cache
DRAM
CPU
SM
L2 Cache
DRAM
GPU


GPU
Device
SMX
SMX
SMX
L2 Cache
Memory controller
Memory controller
Global Memory
Constant Memory
Texture Memory
HOST
CPU
SMX
Instruction Cache
Warp Scheduler
Warp Scheduler
Dispatch Unit
Dispatch Unit
Dispatch Unit
Register File
Core
DPU
SFU
LD/ST
Interconnect Network
Shared Memory / L1 Cache
Read-Only Cache
Tex

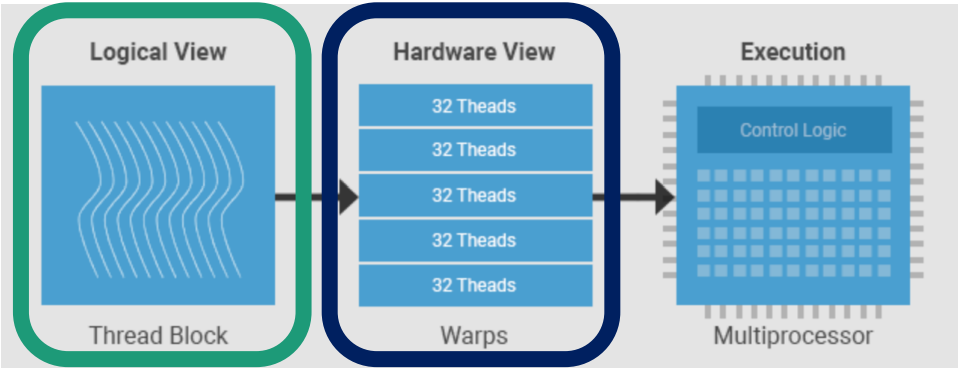
SM - Streaming Multiprocessors

- **Unidade Fundamental:** O SM é o "cérebro" da GPU, responsável por gerenciar e executar grupos de threads.
- **Recursos On-chip:** Cada SM possui seus próprios núcleos CUDA, registradores e memória compartilhada.
- **Escalonamento:** Alterna entre warps para minimizar os efeitos de latência nos acessos à memória global.
- **Persistência:** Um bloco de threads alocado a um SM permanece nele até o fim da execução.



WARPs

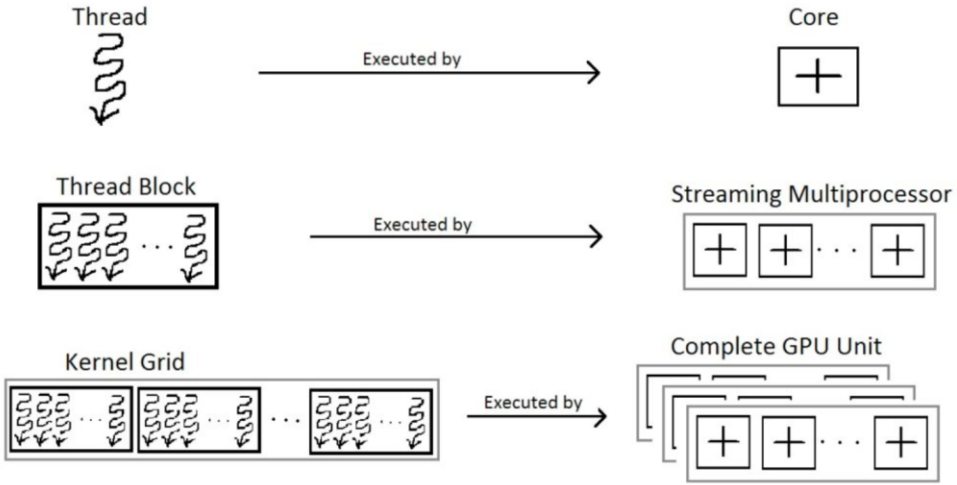
- **Warp:** Unidade primária de execução composta por 32 threads.
- **SIMT:** *Single Instruction, Multiple Thread* — todas as threads executam a mesma instrução simultaneamente.



Regra de Ouro

Para maximizar a ocupação e evitar desperdício de recursos, sempre utilize **blocos com múltiplos de 32 threads**.
Isso garante que nenhum warp seja lançado com threads inativas desnecessariamente.

Hierarquia de Execução em CUDA



Mapeamento de Hardware

Software (CUDA)	Hardware (GPU)
Thread	CUDA Core
Thread Block	Streaming Multiprocessor (SM)
Kernel Grid	GPU Inteira

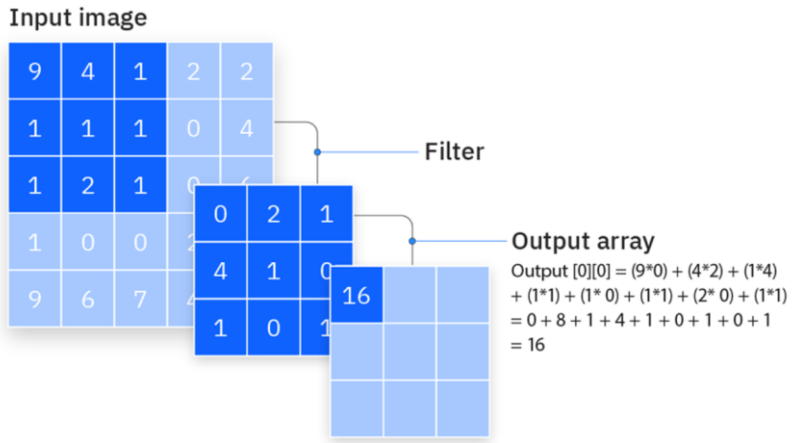
Padrão Stencil e Convolução

Insper

O Padrão de Computação Stencil

Definição

Um stencil é um padrão de computação geométrica onde o valor de um elemento é calculado com base em seus vizinhos.



Aplicações Comuns

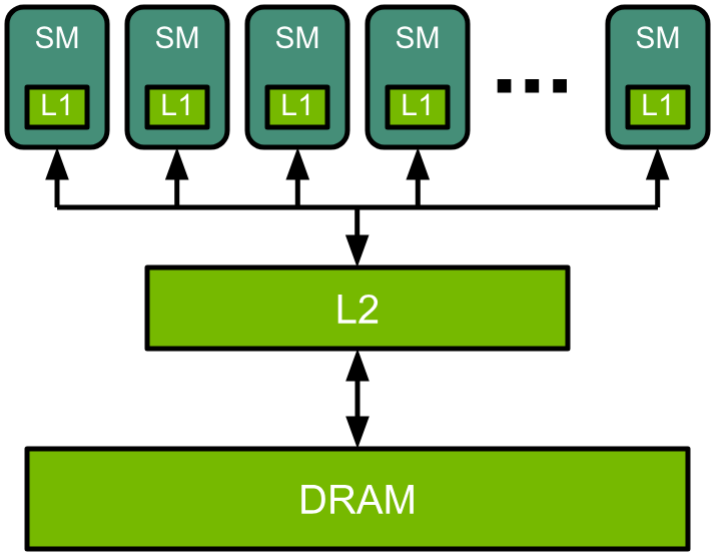
- **Processamento de Imagens:** Filtros de convolução (Blur, Sobel, Sharpen).
- **Simulações Físicas:** Resolução de Equações Diferenciais Parciais (Equação do Calor).
- **Dinâmica de Fluidos:** Métodos de Lattice Boltzmann.

Convolução 2D

O Problema da Memória Global

- **Redundância:** Cada pixel da imagem de entrada é lido múltiplas vezes por threads vizinhas.
- **Latência:** A memória global é lenta (centenas de ciclos de clock).
- **Gargalo:** A performance é limitada pela largura de banda da memória, não pelo processamento.

Cada thread busca seus vizinhos na memória global a cada iteração.

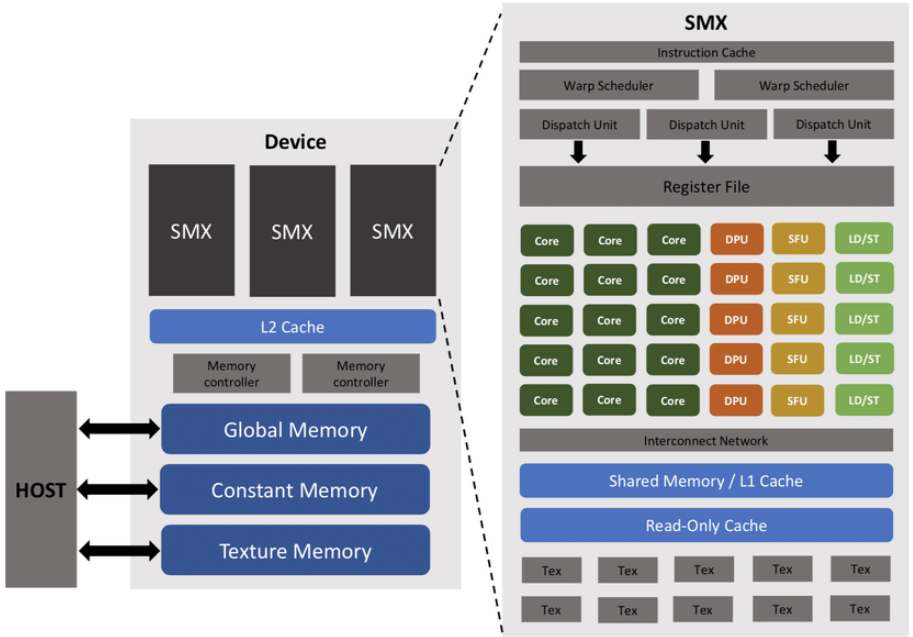


Técnica de Otimização: Tiling

O Conceito

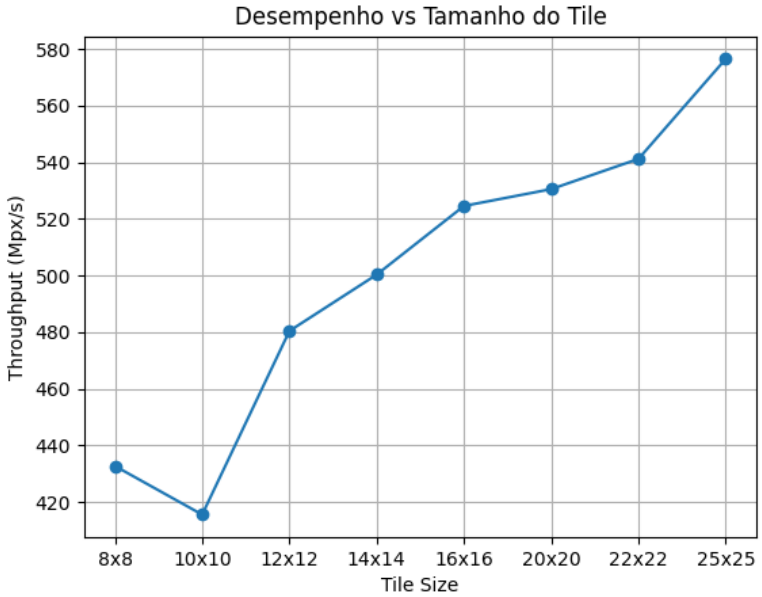
Dividir o problema global em blocos menores (tiles) que cabem em memórias rápidas.

> Memória Global: **LENTA**
>
> Shared Memory: **RÁPIDA**
>
> A Shared Memory é acessível por todas as threads de um mesmo bloco.



Insper

Resultados no Franky




```
Stencil 2D em GPU (3840x2160), 8.3 Mpx

============================================================
ANÁLISE TEÓRICA DE OCUPAÇÃO (CUDA)
------------------------------------------------------------
Threads ideais por bloco : 768
TILE_SIZE sugerido ≈ 25
Grade mínima recomendada : 36 blocos totais
============================================================

INICIANDO EXPERIMENTO DE DESEMPENHO GPU
   (média de 20 execuções por configuração)

Tile      Threads/Bloco   Warps   Tempo (ms)     Mpx/s
8x8       100             4       19.167         432.74
10x10     144             5       19.958         415.59
12x12     196             7       17.265         480.43
14x14     256             8       16.578         500.32
16x16     324             11      15.812         524.56
20x20     484             16      15.635         530.50
22x22     576             18      15.327         541.18
25x25     729             23      14.388         576.48
32x32     1156            37      0.000          161999998.11
```