Arquiteturas de alto desempenho

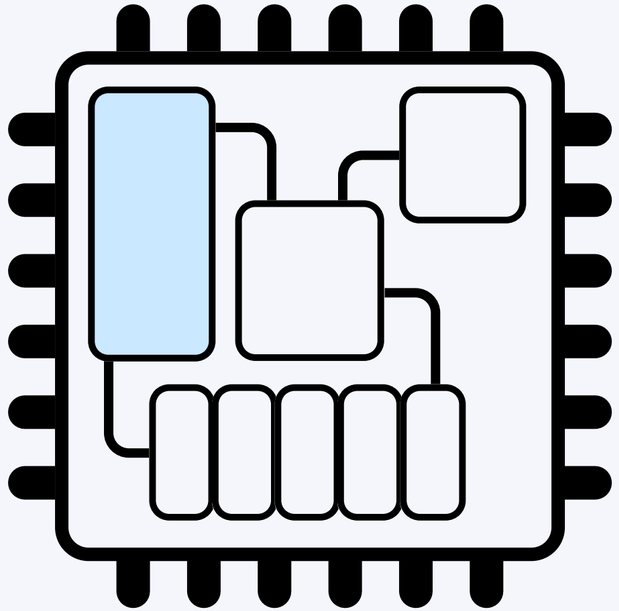
Guilherme Silva Castro

GPUs

Sumario

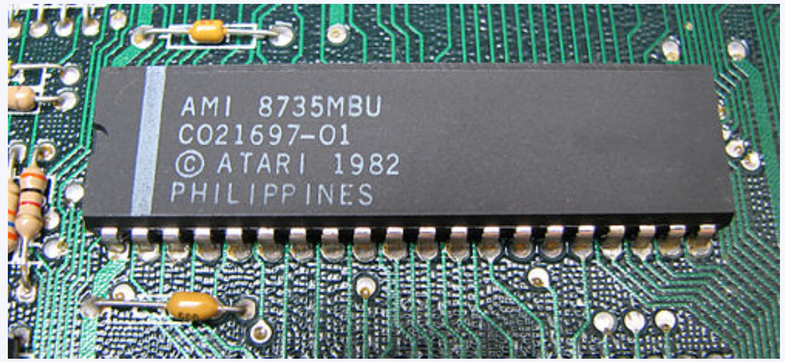
O que são GPUs

Graphics Processing Units



História das GPUs

- Os primeiros chips surgiram no final da década de 1970 e início da década de 1980
- Comecaram em jogos de arcade

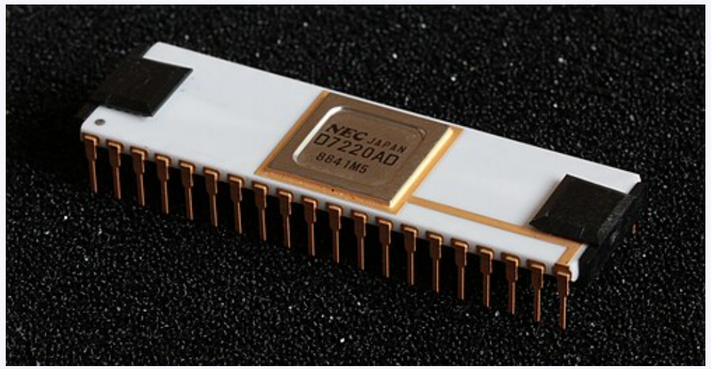


Atari 130XE

História das GPUs

Década de 1980:

- NEC µPD7220
- Hitachi lançou o ARTC HD63484
- Padrão VGA estabelecido



NEC µPD7220

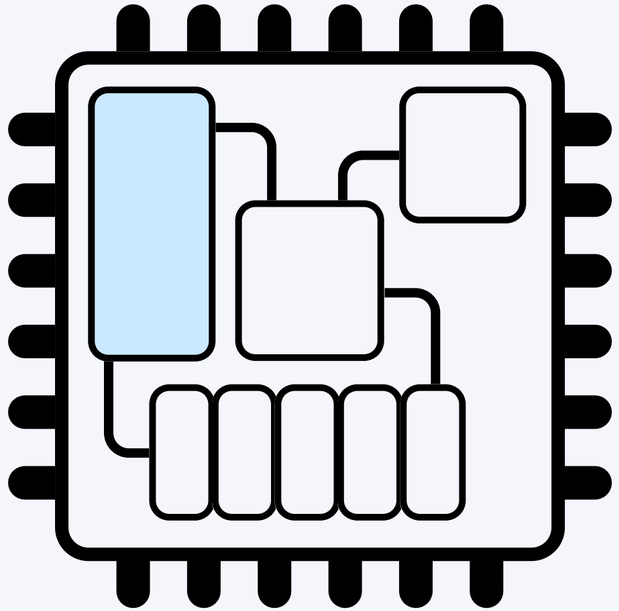
História das GPUs

GPUs comerciais foram popularizadas por conta do mercado de jogos

rtx4060Ti

GPUs de Estudo/Pesquisa (Data Center/Profissional)

são otimizadas para tarefas de computação de alto desempenho (HPC) e inteligência artificial

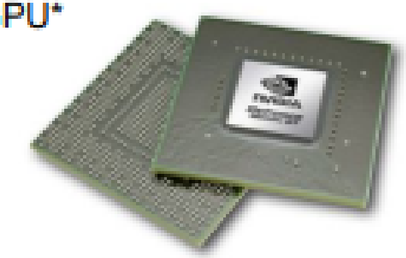


GPU vs. CPU

GPUs - Tarefas sem dependência de dados.

CPUs - Tarefas com dependência de dados.

GPU*



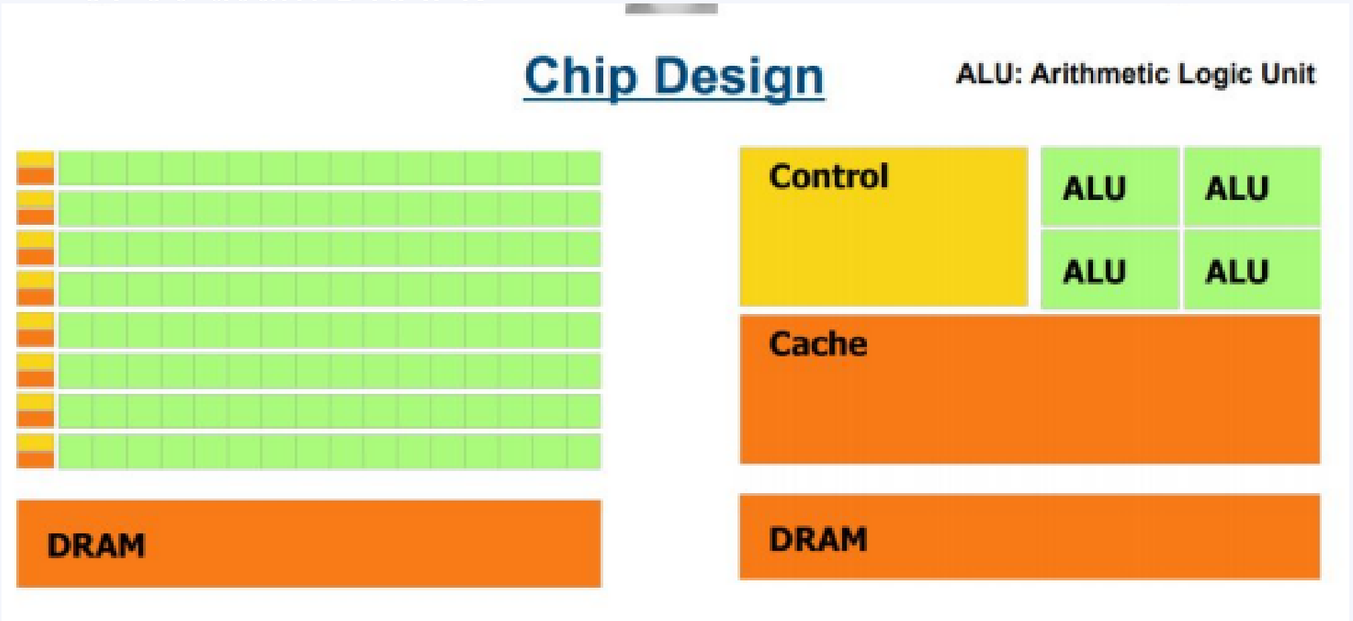
High throughput
and
reasonable latency

Quad_Core*

Low latency
and
reasonable throughput

GPU vs. CPU

GPUs - Memórias com alta largura de banda;
GPUs usam GDDR (Graphics DRAM); (Maior largura de banda que CPUs);
CPUs usam DRAM.



Arquiteturas de GPUs e Paralelismo

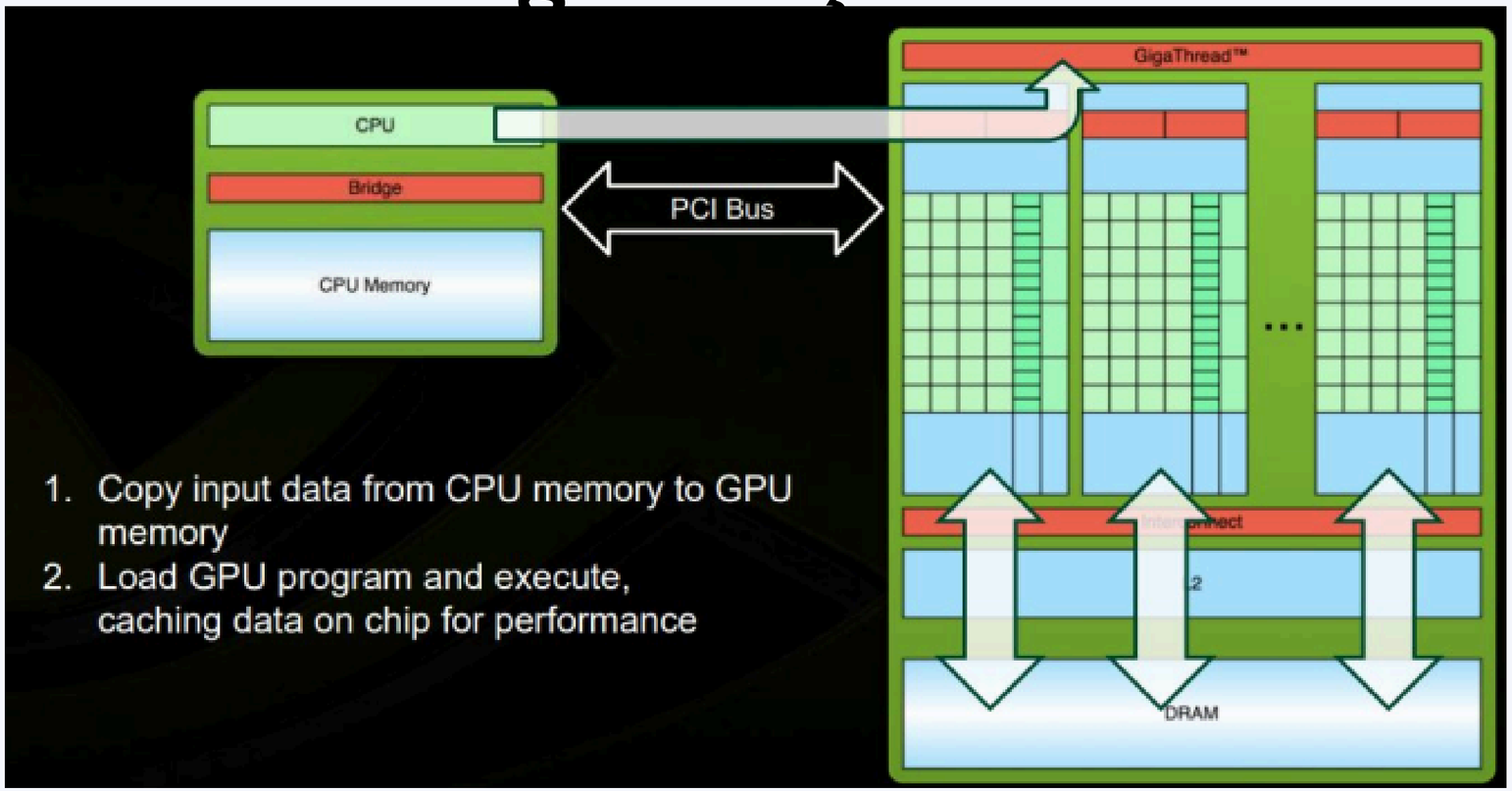
Núcleos e Streaming Multiprocessors (SMs)

Ao contrário das CPUs que possuem poucos núcleos de processamento potentes, as GPUs contam com milhares de núcleos menores e mais simples. Esses núcleos são agrupados em unidades maiores, conhecidas como Streaming Multiprocessors (SMs) na arquitetura NVIDIA CUDA, ou Compute Units (CUs) na arquitetura AMD GCN/RDNA. Cada SM/CU é capaz de executar centenas, ou até milhares, de threads (pequenas unidades de execução) simultaneamente.

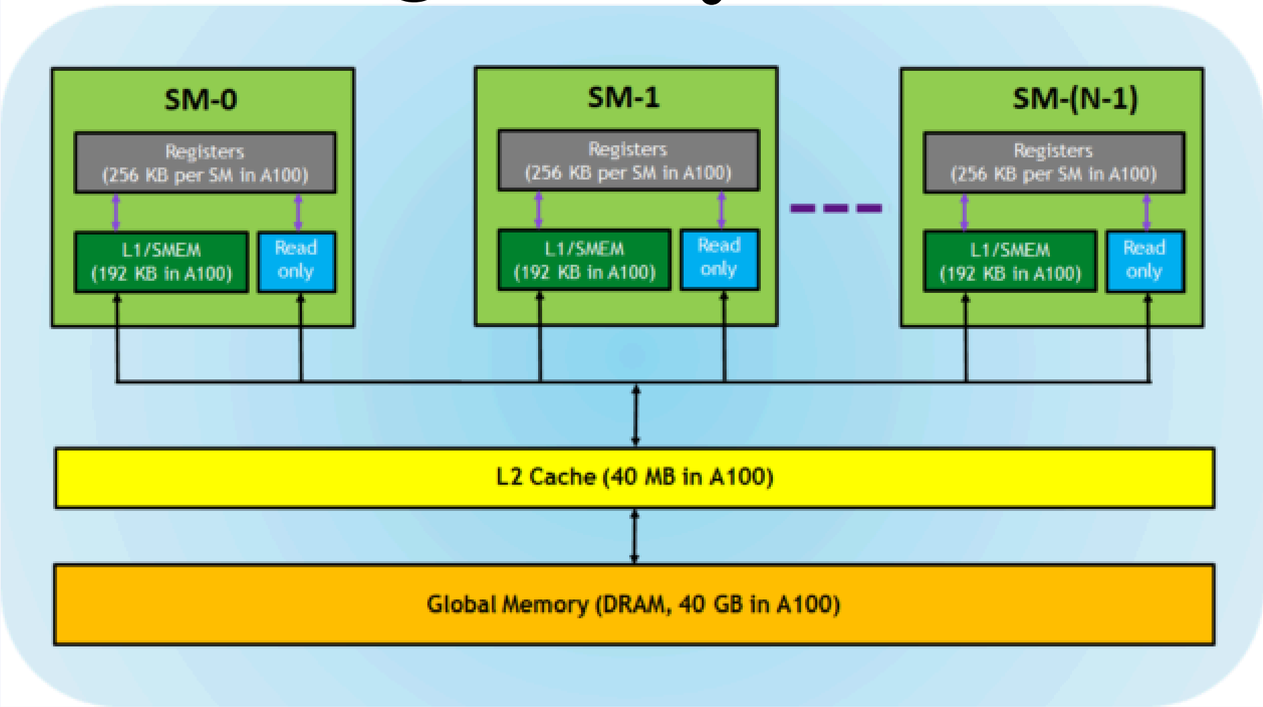
Modelos de Programação Paralela

CUDA é uma plataforma de computação paralela e um modelo de programação que permite aos desenvolvedores usar uma linguagem de programação (como C++) para escrever programas que podem ser executados diretamente nos núcleos da GPU.

Modelos de Programação Paralela



Modelos de Programação Paralela



GPUs e IAs

Treinamento de Redes Neurais Profundas
Inferência em Tempo Real