INSTITUTO DE INFORMÁTICA
Universidade Federal de Goiás

Arquitetura de Computadores

Paralelismo

INSTITUTO DE INFORMÁTICA
UFG

Bruna Michelly de O. Silva - brunamichelly@ufg.br

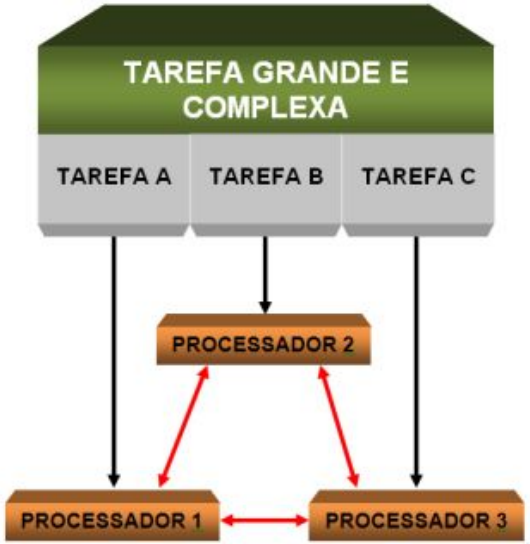
UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Conteúdo

- **Paralelismo**
- Métricas de Desempenho
- Classificação de Computadores

Paralelismo

Paralelismo é uma técnica de computação na qual **múltiplas tarefas são executadas simultaneamente**, aumentando a eficiência e o desempenho de sistemas de computação.



Motivação para Computação Paralela

1 **Volume de Dados**

O crescimento exponencial de dados exige processamento mais rápido e eficiente.

2 **Demanda por Velocidade**

Aplicações modernas requerem respostas em tempo real e análises complexas.

3 **Limites Físicos**

A computação paralela supera limitações de velocidade de processadores individuais.

4 **Problemas Complexos**

Simulações científicas e modelos complexos exigem grande poder computacional.

5 **Eficiência Energética**

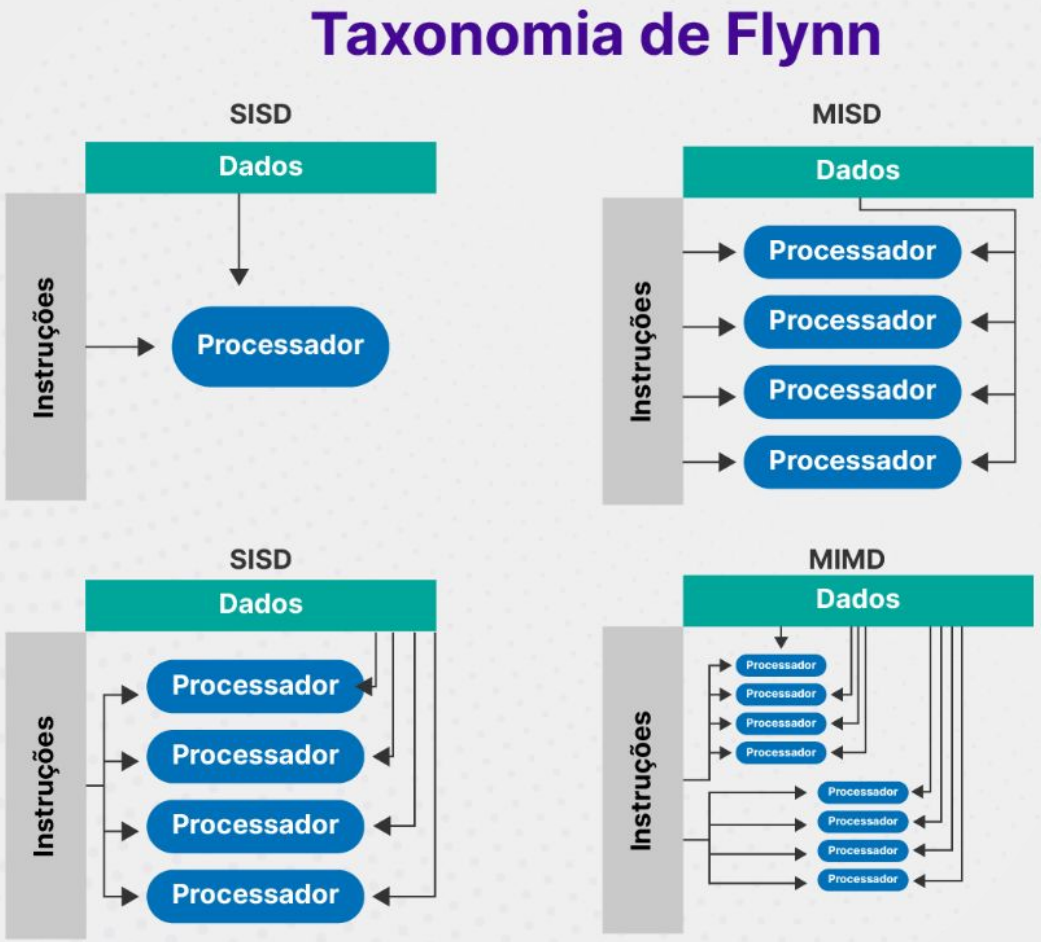
A computação paralela pode consumir menos energia do que sistemas sequenciais equivalentes.

6 **Aplicações Emergentes**

Novas áreas como inteligência artificial e aprendizado de máquina se beneficiam da computação paralela.

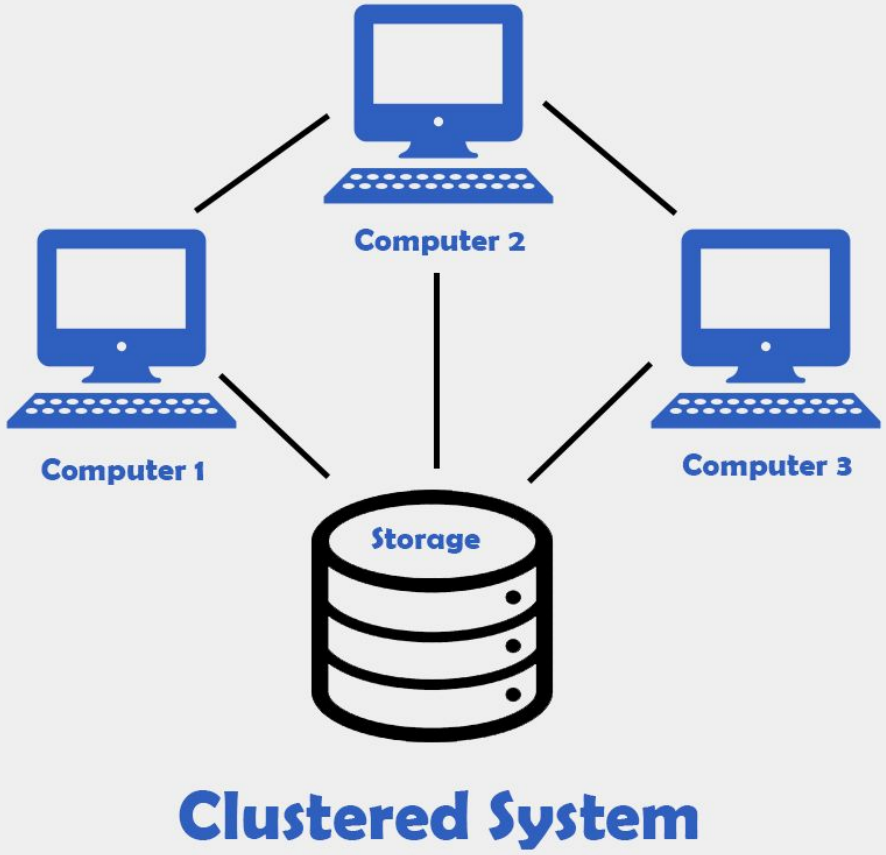
Taxonomia de Flynn

Taxonomia de Flynn classifica arquiteturas de computadores com base na natureza dos fluxos de instruções e dados.



Clusters

Conjunto de computadores **interligados** para funcionar como um **sistema único**



Grid Computing

Rede de computadores que compartilham recursos e trabalham em conjunto

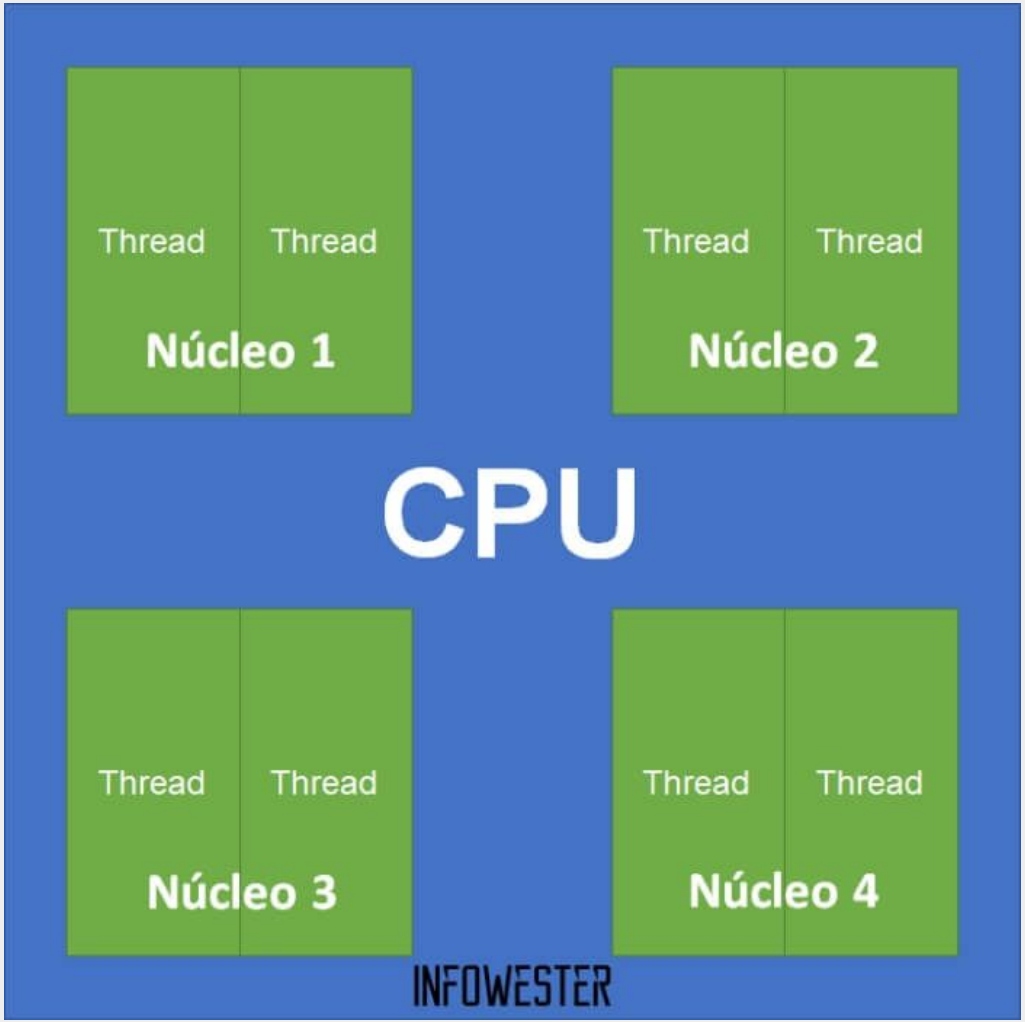
Clusters e Grid Computing

Em uma arquitetura de grid computing, vários clusters podem ser integrados, formando uma "super grade".

Cada cluster, com seus nós interconectados e geralmente localizado no mesmo espaço físico, **atua como uma unidade dentro do grid.**

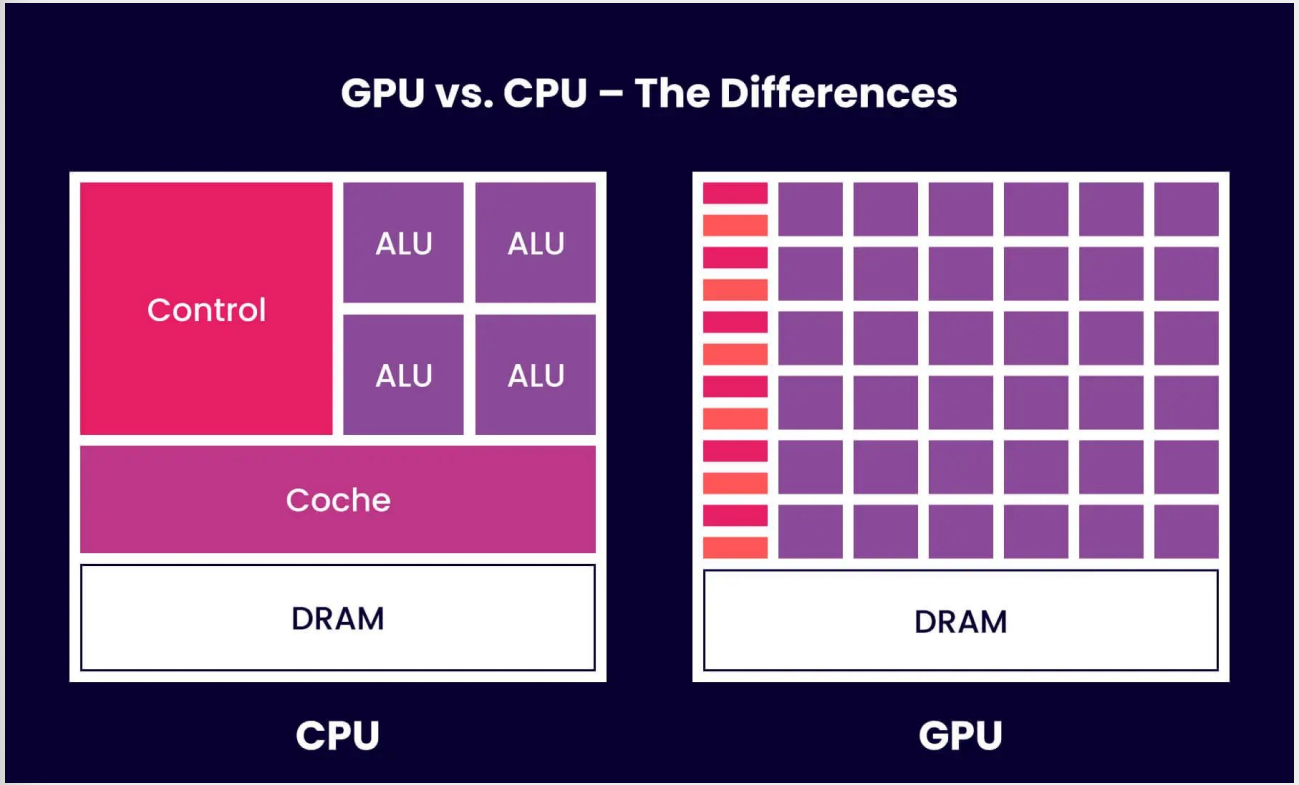
Arquiteturas Multicore

Múltiplos núcleos de processamento independentes são integrados em um único chip.



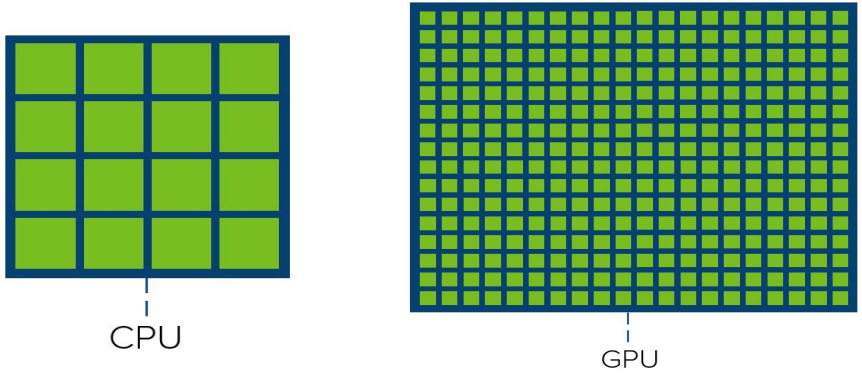
Arquitetura de GPU

Projetada especificamente para lidar com operações que envolvem processamento massivo de dados em paralelo.



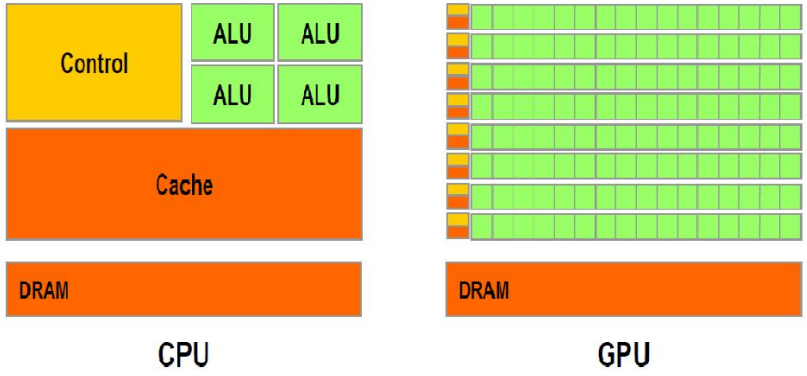
Estrutura e Componentes de uma GPU

- GPUs modernas contêm **milhares de núcleos de processamento pequenos e eficientes**.
- Cada núcleo executa operações aritméticas e lógicas simples, mas a enorme quantidade permite que grandes volumes de dados sejam processados ao mesmo tempo.



Estrutura e Componentes de uma GPU

Cada núcleo possui seu **próprio conjunto de registradores** para armazenar dados temporários rapidamente durante as operações, **reduzindo o acesso à memória principal** e aumentando a velocidade de execução

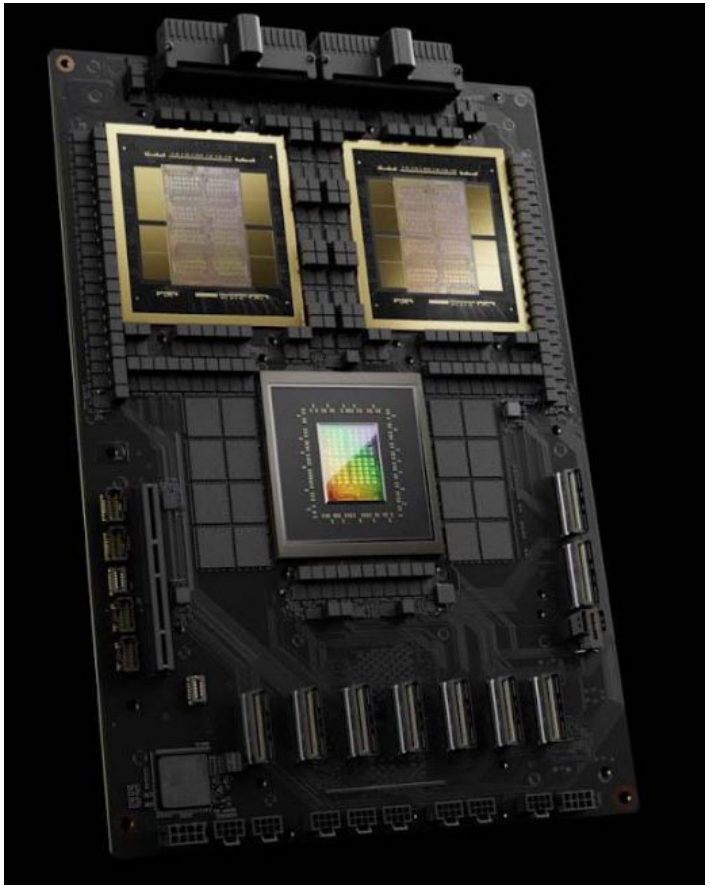


Arquitetura GPU

- As GPUs são projetadas para executar computação massivamente paralela.
- Elas dividem uma tarefa em milhares de partes e distribuem essas partes entre seus núcleos.
- Por exemplo, ao processar uma imagem, cada pixel pode ser manipulado simultaneamente em diferentes núcleos, acelerando o processamento.

Arquitetura GPU

GPUs são ideais para o treinamento de redes neurais profundas. Elas aceleram os cálculos necessários para o aprendizado de máquina, como multiplicações de matrizes e operações de ponto flutuante.



Computação Paralela e Distribuída

A união da computação paralela e distribuída abre caminho para soluções mais rápidas e poderosas, conectando e potencializando recursos em um mundo interligado.

- Paralelismo
- **Métricas de Desempenho**
- Classificação de Computadores

Métricas de Desempenho

- Métricas de desempenho são ferramentas essenciais para avaliar a eficiência e a eficácia de sistemas de computação.
- Elas permitem aos engenheiros e cientistas de computação medir, analisar e otimizar o desempenho de hardware, software e redes.

Principais Métricas de Desempenho

- As métricas de desempenho são vitais para entender e melhorar a eficiência de sistemas de computação.
- Elas fornecem informações detalhadas sobre como os recursos do sistema estão sendo utilizados e onde podem ser feitas otimizações.
- A seleção adequada das métricas de desempenho e o uso de ferramentas apropriadas são essenciais para garantir que os sistemas de computação atendam às necessidades de desempenho e eficiência.

Principais Métricas de Desempenho

- Throughput
- Latência
- Tempo de Execução
- Ciclos por Instrução (CPI)
- Utilização da CPU
- Memória (Uso e Latência)
- Velocidade de Clock
- Eficácia Energética

Throughput (Vazão)

- Refere-se à **quantidade de trabalho** que o sistema realiza em um determinado período. No contexto de um processador, throughput pode ser medido como o número de instruções processadas por segundo.
- Throughput é normalmente medido em operações por segundo. Por exemplo, se um processador executa 10 bilhões de instruções em 1 segundo, o throughput é de 10 Giga Instructions Per Second (GIPS).

Latência

- Latência é o tempo de espera necessário para que uma operação específica seja concluída. Em processamento, é o tempo entre o início e o fim de uma operação.
- Em programas/sistemas, o tempo total de execução de uma instrução pode ser medido com ferramentas de benchmarking.

Ciclos por Instrução (CPI)

- Ciclos por Instrução (CPI) é uma métrica que indica quantos ciclos de clock são necessários, em média, para executar uma instrução. Quanto menor o CPI, mais eficiente é o processador.

CPI = Total de Ciclos de Clock / Total de Instruções

Utilização da CPU

- Percentual de tempo em que a CPU está ocupada executando instruções, em vez de ficar ociosa. Uma alta utilização indica que a CPU está trabalhando constantemente.
- A utilização da CPU é dada como uma porcentagem, calculada por:

(tempo total de CPU em uso / tempo total de amostragem) * 100

Velocidade de Clock

- A velocidade de clock é a frequência em que o processador opera, geralmente medida em GHz. Velocidades de clock mais altas permitem que mais instruções sejam executadas por segundo.
- A velocidade de clock é dada pelo número de ciclos por segundo. Por exemplo, um processador com velocidade de 3 GHz realiza 3 bilhões de ciclos por segundo.

Atividade!

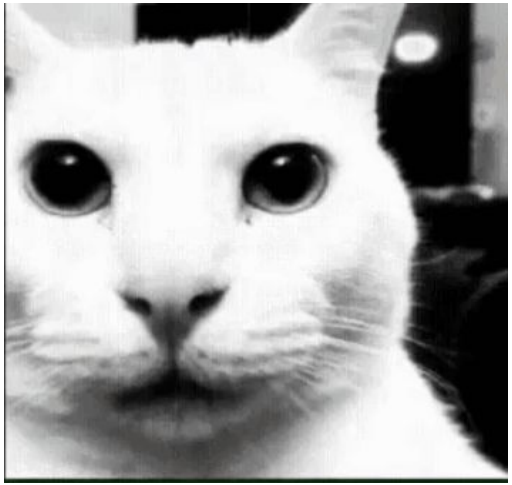
Explorando o Desempenho do Computador

- A velocidade de clock é a frequência em que o processador opera, geralmente medida em GHz. Velocidades de clock mais altas permitem que mais instruções sejam executadas por segundo.
- A velocidade de clock é dada pelo número de ciclos por segundo. Por exemplo, um processador com velocidade de 3 GHz realiza 3 bilhões de ciclos por segundo.

Referências

- Differences Between Grid, Cluster, Utility & Cloud Computing | Giva
- Deep Learning com GPU: Por que usar? Quando usar? Quais os primeiros passos? | by Denise Marti | Medium
- NVIDIA Blackwell Architecture Technical Overview
- Processador Intel® Core™ i9-13900K

Devolução das Provas!



Terceira Avaliação
03 de dezembro

Prof. Bruna Michelly - brunamichelly@ufg.br