

Programação Multicore

Paralelismo de Hardware e Software

Demetrios A. M. Coutinho - NADIC/IFRN

Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte

12-13 de maio de 2023



Agenda

Part 1 - **Contexto e Motivação**

Part 2 - **Paralelismo de Hardware e Software**

Part 3 - **Computação Paralela com OpenMP**

Part 4 - **Computação Paralela em CUDA**



Next

1 Arquiteturas Paralelas

2 Processadores Multicore

3 Desempenho de Algoritmos Paralelos

4 Resumo



Classificação

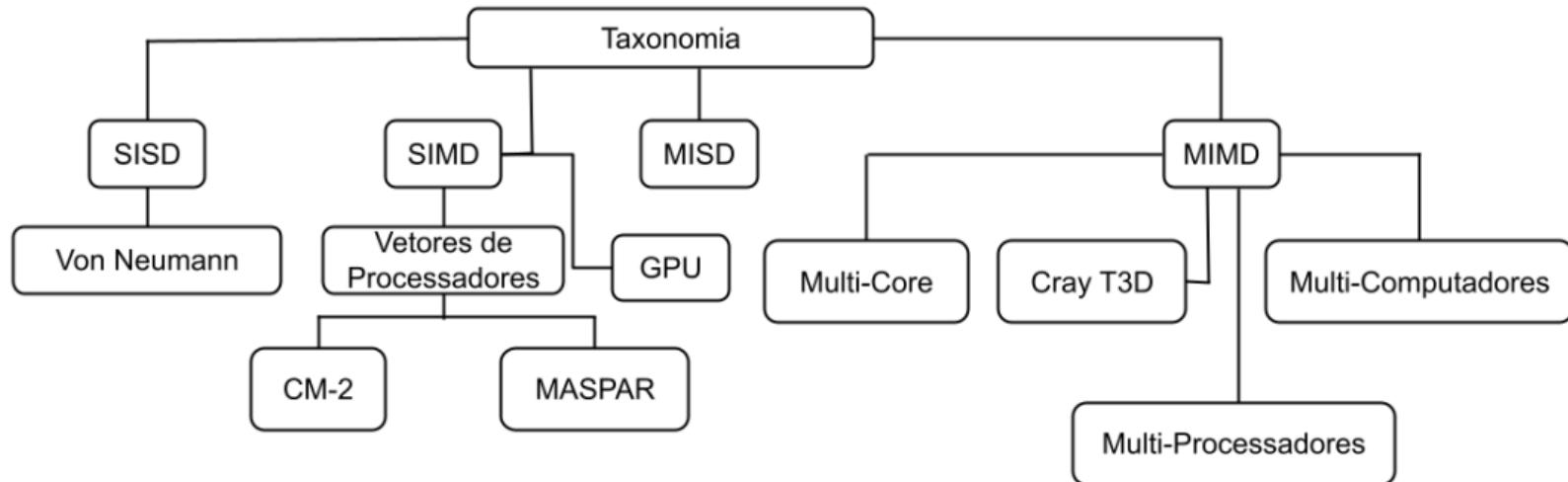
A taxonomia de Flynn (1972)

é uma maneira comum de se caracterizar arquiteturas de computadores paralelos.

	SD (Single Data)	MD (Multiple Data)
SI (Single Instruction)	SISD	SIMD
MI (Multiple Instruction)	MISD	MIMD



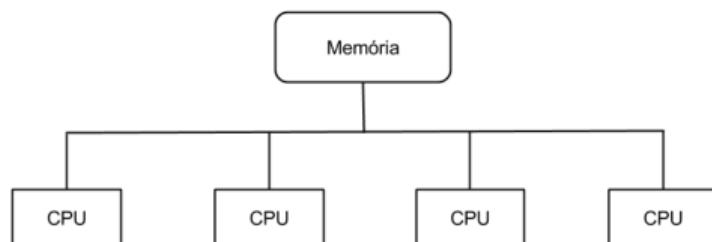
Ramificações



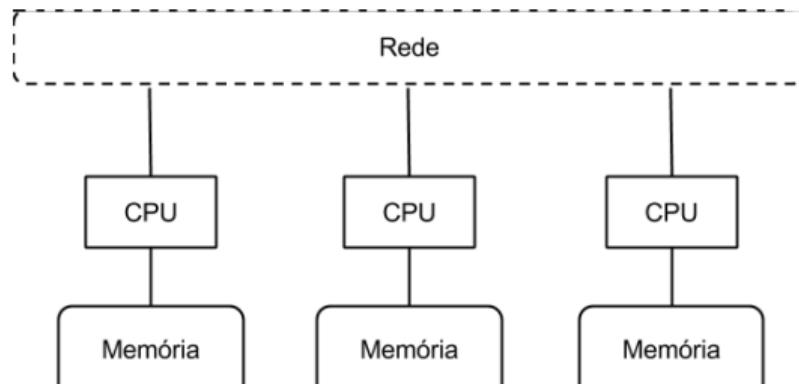
Arquitetura de Memória

- ▶ A arquitetura MIMD da classificação de Flynn é muito genérica para ser utilizada na prática.
- ▶ Então, ela geralmente é decomposta de acordo com a organização de memória.
- ▶ Existem dois principais tipos de sistemas paralelos, diferindo quanto a comunicação, são eles: **memória compartilhada** e **memória distribuída**.

Arquitetura de Memória



(a) Memória Compartilhada.



(b) Memória Distribuída.

Next

1 Arquiteturas Paralelas

2 Processadores Multicore

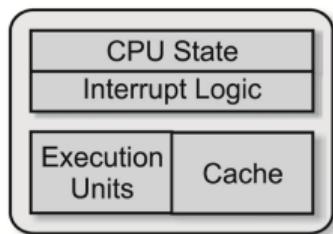
3 Desempenho de Algoritmos Paralelos

4 Resumo

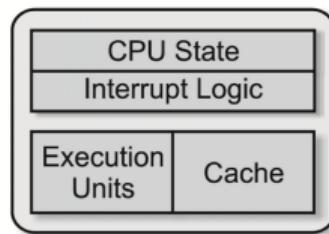


Arquiteturas Multicore

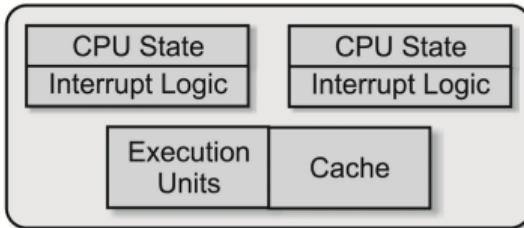
Figure: Comparação entre diferentes microarquiteturas de processadores.



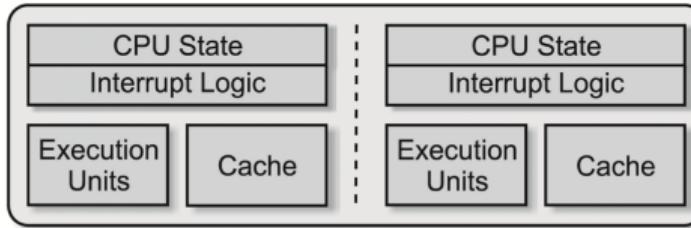
A) Single-Core



B) Multi-processador



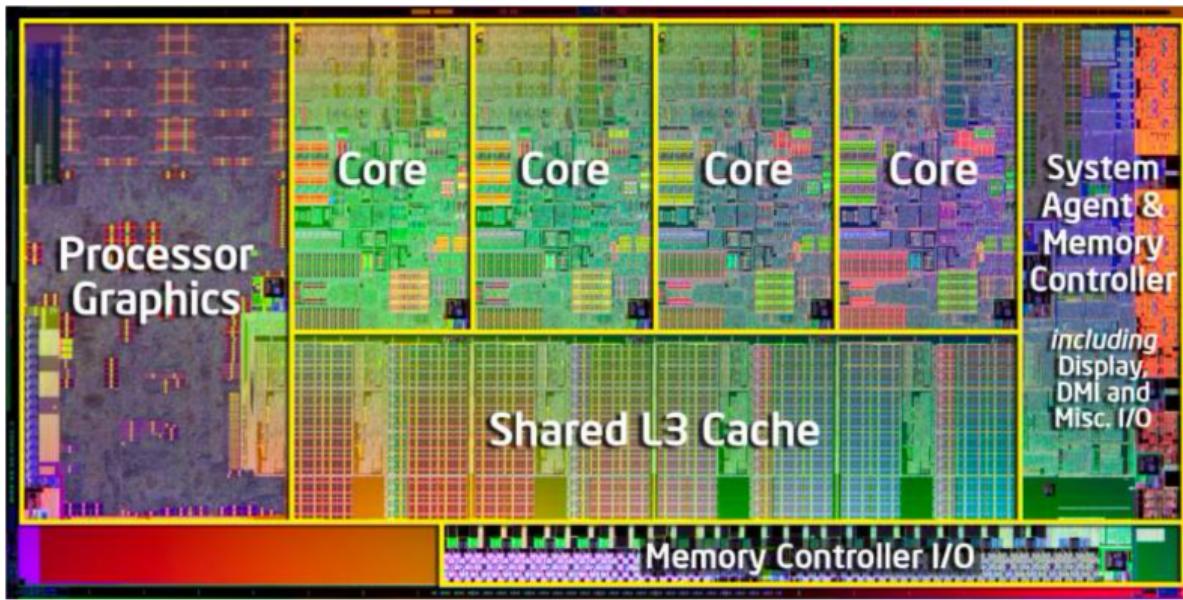
C) Hyper-Threading



D) Multi-core

Arquiteturas Multicore

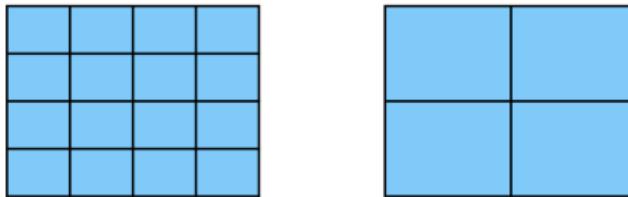
Figure: Estrutura interna (*die map*) dos processadores multicore i7 de 2 geração



Classificação de Processadores Multicore

- ▶ Com base nessa possibilidade de diferenciação na complexidade dos núcleos, pode ser realizada uma classificação (simples, mas importante) dos vários tipos de processadores multicore.

Multicore Homogêneo

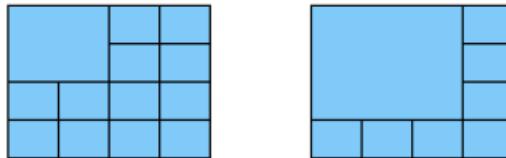


(a) Representação simplificada de processadores multicore homogêneos.



(b) Intel I9.

Multicore Heterogêneo



(a) Representação simplificada de processadores multicore heterogêneos.



(b) NVIDIA Jetson TX2 MSoC.

Multicore Dinâmico

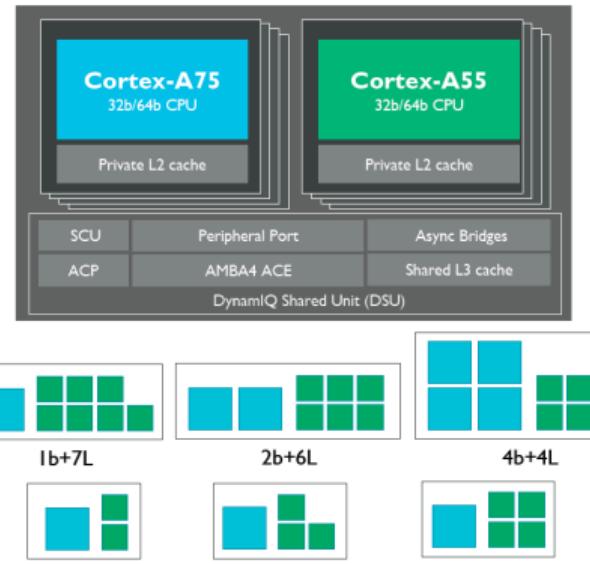
ON		off	off
off		off	off
off	off	off	off
off	off	off	off

Execução da porção serial

off		ON	ON
off		ON	ON
ON	ON	ON	ON
ON	ON	ON	ON
ON	ON	ON	ON

Execução da porção paralela

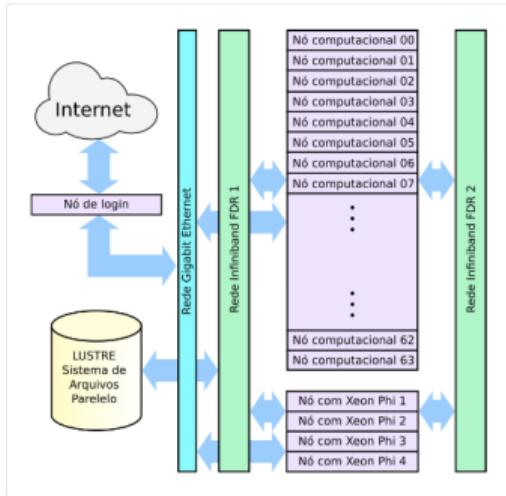
(a) Representação simplificada de um processador multicore dinâmico em diferentes momentos.



(b) ARM DynamIQ.

Supercomputador

Figure: Estrutura do supercomputador do NPAD - Núcleo de Processamento de Alto Desempenho.



64 nós computacionais em lâmina, cada um configurado com:

- 2 x CPU Intel Xeon Sixteen-Core E5-2698v3 de 2.3 GHz/40M cache/ 9.6 GT/s
- 128 GB RAM DDR4 2133 RDIMM (8 x 16GB)
- 1 x chip dual-port Infiniband FDR 4x on-board
- 1 x 120 GB SSD HD for local scratch
- Porta gigE dedicada para gerenciamento

4 x nós com Xeon Phi configurados com:

- 2 x CPU Intel Xeon Sixteen-Core E5-2698v3 de 2.3 GHz/40M cache/ 9.6 GT/s
- 128 GB RAM DDR4 2133 RDIMM (8 x 16GB)
- 1 x 120 GB SSD HD for local scratch
- 2 ports Infiniband FDR 4x
- Porta gigE dedicada para gerenciamento
- 1 x Intel Xeon Phi 5110P

2 nós com GPU, cada um configurado com:

- 2 x CPU Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.10GHz/40M cache/ 9.6 GT/s QPI
- 8 x Tesla V100-SXM2-16GB/ 5,210 GPU Cores/ 4 X DP 1.4
- 512 GB RAM DDR4 2400 DIMM (16 x 32GB)
- 1 x 1920 GB SSD HD for local scratch
- 16 x 2.5" hot-swap SAS/SATA drive bay

2 nós KNL, cada um configurado com:

- 1 x Intel(R) Xeon(R) CPU 7250F @ 1.40GHz/ 32M cache L2
- 6 x 16 GB RAM 1200 MHz DIMM
- 8 x 2 GB RAM 7200 MHz DIMM
- 1 x 800 GB SSD HD for local scratch/ 6Gb/s

1 x nó de login configurado com:

- 2 x CPUs Intel Xeon Six-Core E5-2620v3 de 2.4 GHz/15M cache/ 7.2 GT/s
- 128 GB RAM DDR4 2133 RDIMM (8 x 8GB)

Interconexão:

- 8 x Switch Blades 56 Gbps FDR 4x Standard
- 2 x Planos de Rede Infiniband FDR Quad-Linked Backplane
- Topologia Enhanced Hypercube

Armazenamento:

- Sistema de arquivos paralelo Lustre com 60 Terabytes

Servidor NADIC

Descrição

- ▶ AMD Ryzen 7 5800X.
 - ▶ 8 núcleos homogêneos com 16 threads.
 - ▶ 3.8GHz - 4.7GHz (Boost).
 - ▶ L2 Cache 4MB - L3 Cache 32MB.
- ▶ Memória RAM DD4 64GB.
- ▶ Placa de Vídeo RTX 3080
 - ▶ 1440MHz - 1710MHz (Boost).
 - ▶ 8704 Cuda cores
 - ▶ 10 GB GDDR6X.



Arquitetura Paralela

Oráculo

As tecnologias atuais e as futuras serão cada vez mais paralelas!



Next

1 Arquiteturas Paralelas

2 Processadores Multicore

3 Desempenho de Algoritmos Paralelos

4 Resumo



Desempenho de Algoritmos Paralelos

- ▶ É importante entender até que ponto é vantajoso utilizar programas paralelos.
- ▶ O objetivo é determinar os benefícios do paralelismo aplicados a um problema considerado.
- ▶ E também o quanto o algoritmo é capaz de continuar eficiente.



Speedup e Eficiência

$$S = \frac{T_S}{T_P}$$

- ▶ O **speedup** S diz quantas vezes o algoritmo paralelo é mais rápido que o algoritmo serial.

Speedup e Eficiência

$$E = \frac{S}{P} = \frac{\frac{T_S}{T_P}}{P} = \frac{T_S}{P T_P}$$

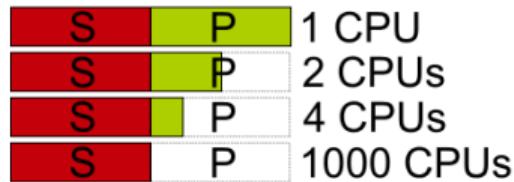
- ▶ A eficiência é uma medida normalizada de *speedup* que indica o quanto efetivamente cada processador é utilizado.
- ▶ P é número de *threads*.
- ▶ Um *speedup* com valor igual a P, tem uma eficiência igual 1, ou seja, todos processadores são utilizados e a eficiência é linear.

Lei de Amdahl

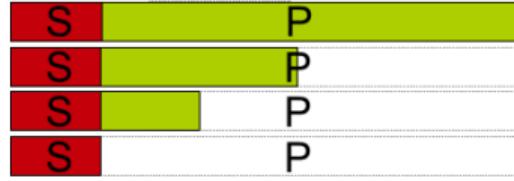
- ▶ Segundo a Lei de Amdahl (do inglês, *Amdahl's law*), a velocidade de processamento paralelo é limitada a porção sequencial do programa.
- ▶ Essa porção do programa que não pode ser paralelizada limitará o aumento de velocidade disponível com o paralelismo.

Lei de Amdahl

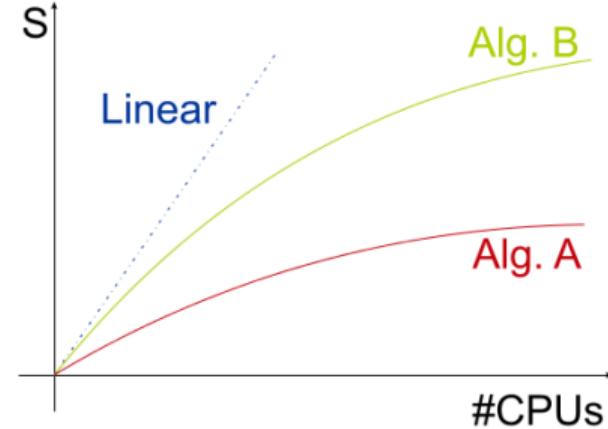
Alg. A



Alg. B



Tamanho fixo do problema

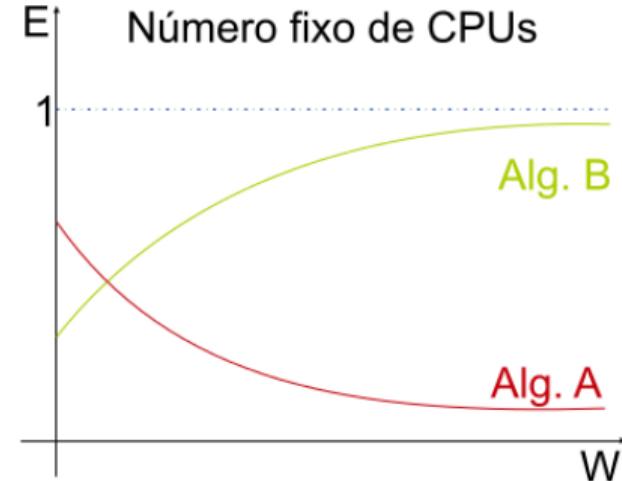
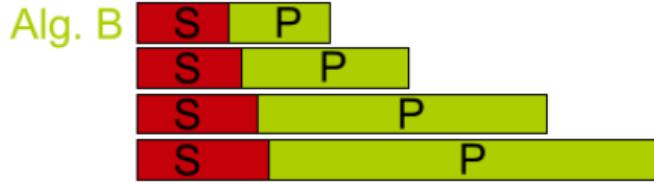
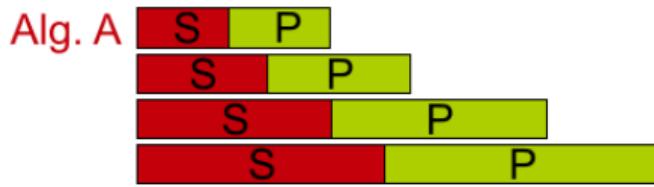


Escalabilidade de Gustafson

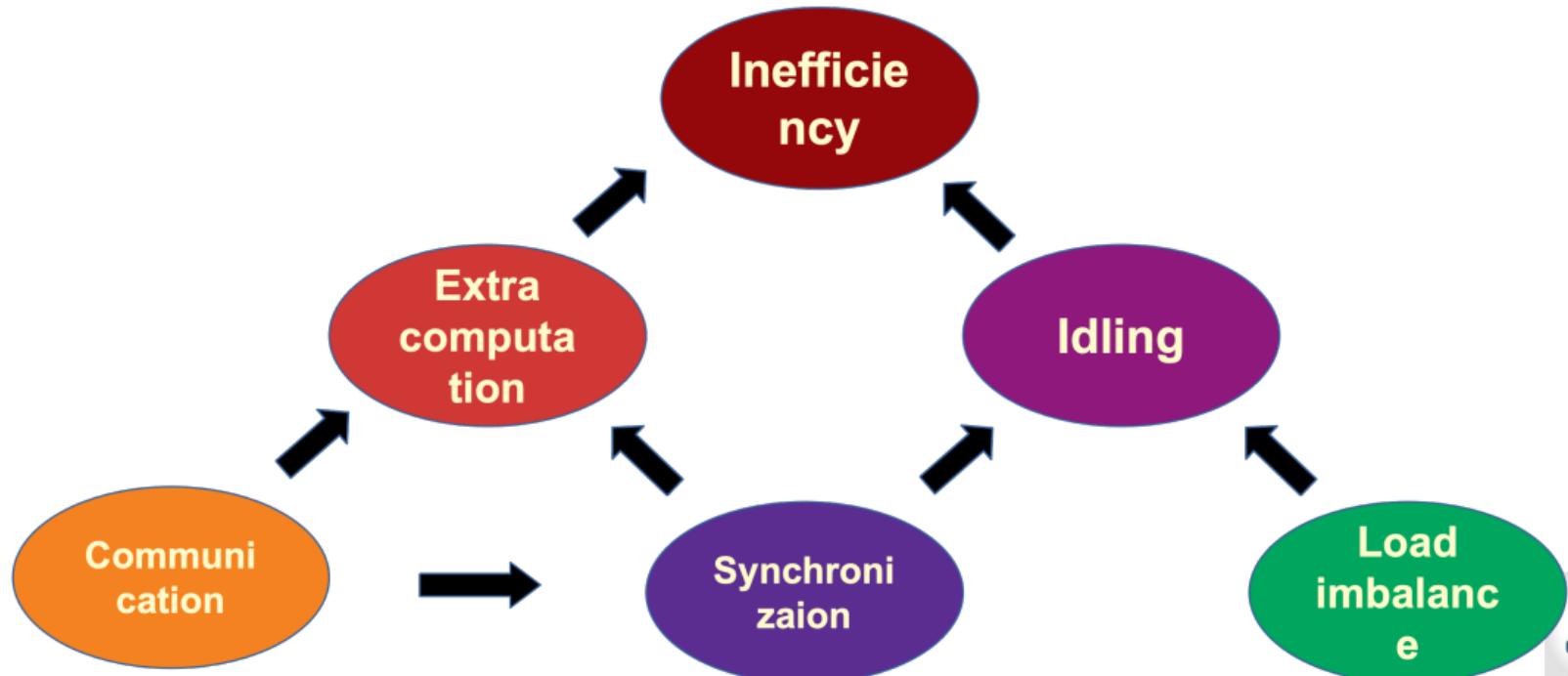
- ▶ A palavra escalável tem uma grande variedade de significados em diversas áreas.
- ▶ Informalmente, a tecnologia é escalável quando ela pode lidar com problemas cada vez maiores sem perdas de desempenho.



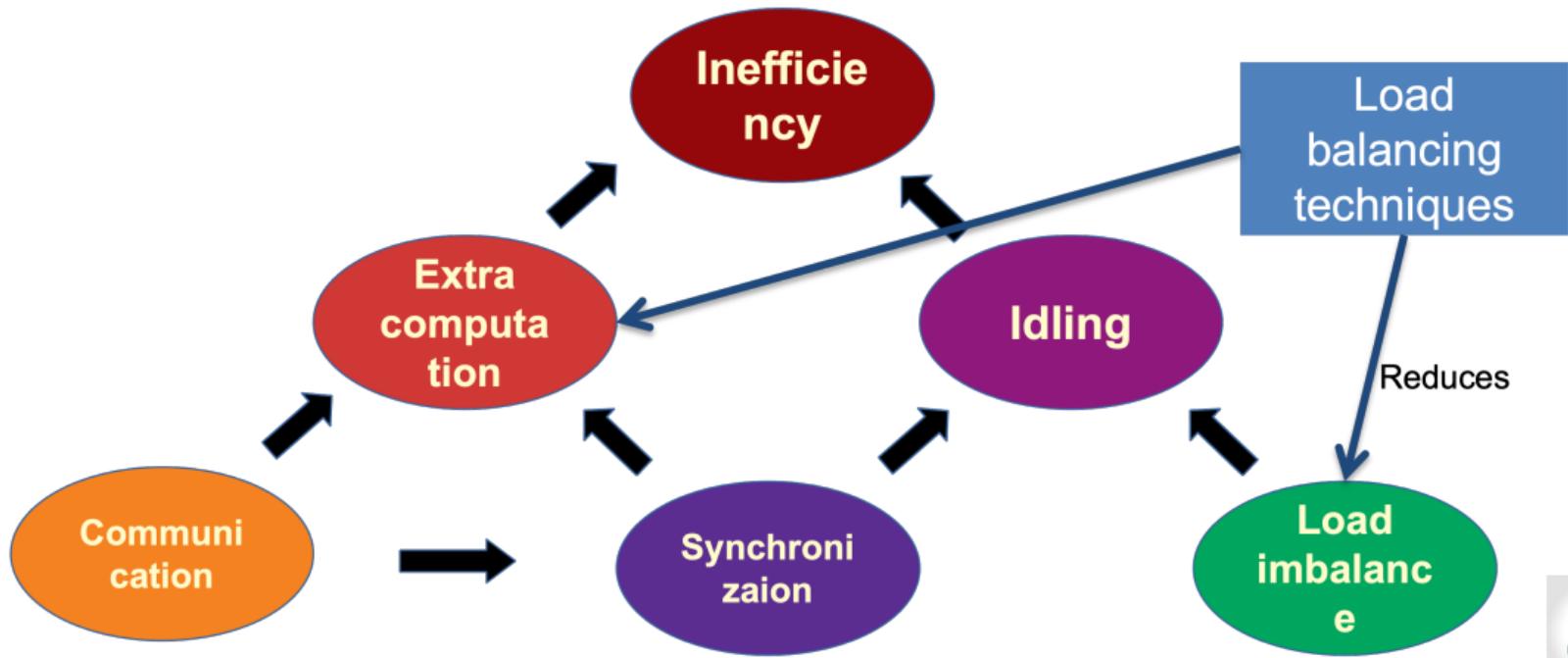
Escalabilidade de Gustafson



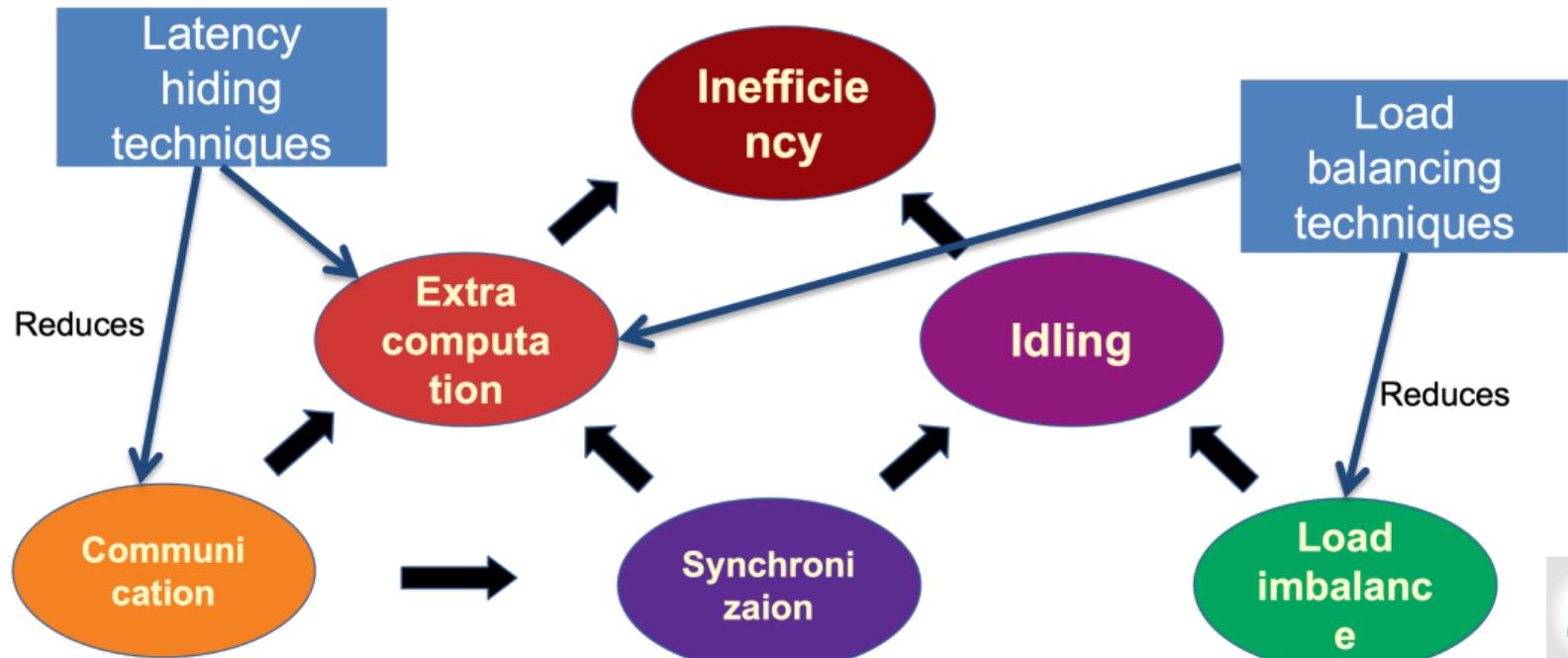
Eficiência no Desempenho



Eficiência no Desempenho



Eficiência no Desempenho



Next

1 Arquiteturas Paralelas

2 Processadores Multicore

3 Desempenho de Algoritmos Paralelos

4 Resumo



O que vimos

</> Multicore

Programação multicore maximiza recursos em sistemas modernos.

Arquiteturas Paralelas

Entender arquiteturas paralelas otimiza desempenho e eficiência.

Eficiência e Escalabilidade

Fatores cruciais ao desenvolver aplicações paralelas.

Agenda

Obrigado pela atenção 🙌

Próximos tópicos

Part 1 - **Contexto e Motivação**

Part 2 - **Parallelismo de Hardware e Software**

Part 3 - **Computação Paralela com OpenMP**

Part 4 - **Computação Paralela em CUDA**