

# Seminário MAD

Performance Analysis of GPU Accelerators  
with Realizable Utilization of  
Computational Density

Laura Costa - Livia - Luca Gonzaga - Sophia Carrazza - Vitor Militão - Temístocles Schwartz

# **INFORMAÇÕES BÁSICAS:**

**Publicado em:** 2012 Symposium on Application Accelerators in High Performance Computing;

**Data da Conferência:** 10-11 Julho 2012;

**Publisher:** IEEE;

**Autores:** Justin W. Richardson, Alan D. George, Herman Lam.



## CONTEXTO

O estudo é direcionado à sistemas de alto desempenho, que utilizam ou podem utilizar GPUs para acelerar a computação.

## PROBLEMA

Mesmo com a expansão do poder computacional das GPUs, as aplicações não estão alcançando o desempenho teórico esperado.

# CONTEXTO

## Limitadores

- Complexidade dos códigos
- Ferramentas disponíveis
- Características da aplicação
- Gargalos da arquitetura

## Métricas

As métricas de desempenho dos dispositivos não levam em consideração os fatores específicos da aplicação.

# BACKGROUND

O artigo utiliza a métrica de Densidade Computacional (CD), que permite avaliar CPUs, GPUs e FPGAs. Esta métrica é uma **medida de capacidade computacional teórica de um dispositivo de realizar operações por unidade de tempo**, e pode ser utilizada em vários sistemas e arquiteturas.

(operações de ponto flutuante por segundo (FLOPs))



$$CD = f \times \sum_i \frac{N_i}{CPI_i}$$

unidades de execução disponíveis

ciclos de clock por instrução

# PROBLEMAS DA CD

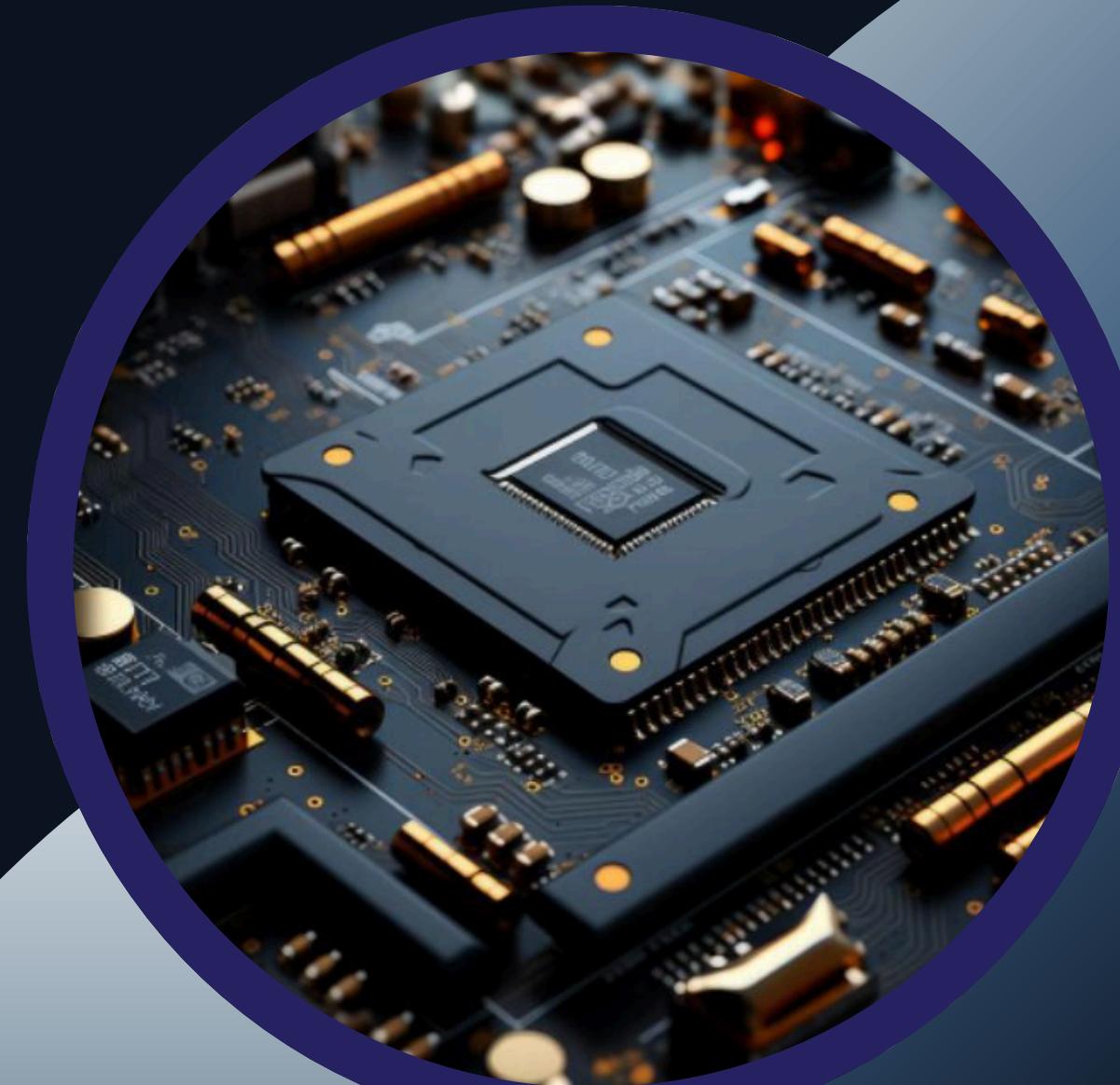
- Indica o potencial bruto de processamento do dispositivo, mas...
- Não leva em consideração limitadores práticos como eficiência do software e capacidade de paralelização da aplicação.



## UTILIZAÇÃO REALIZÁVEL

Diversos fatores podem reduzir o desempenho de um dispositivo, e esta métrica tem como objetivo **quantificar a diferença aproximada entre o desempenho teórico do desempenho real que o usuário pode esperar alcançar.**

A métrica permite que os desenvolvedores estimem o desempenho projetado de suas aplicações em um dispositivo específico, sem que seja necessário realizar os benchmarks.



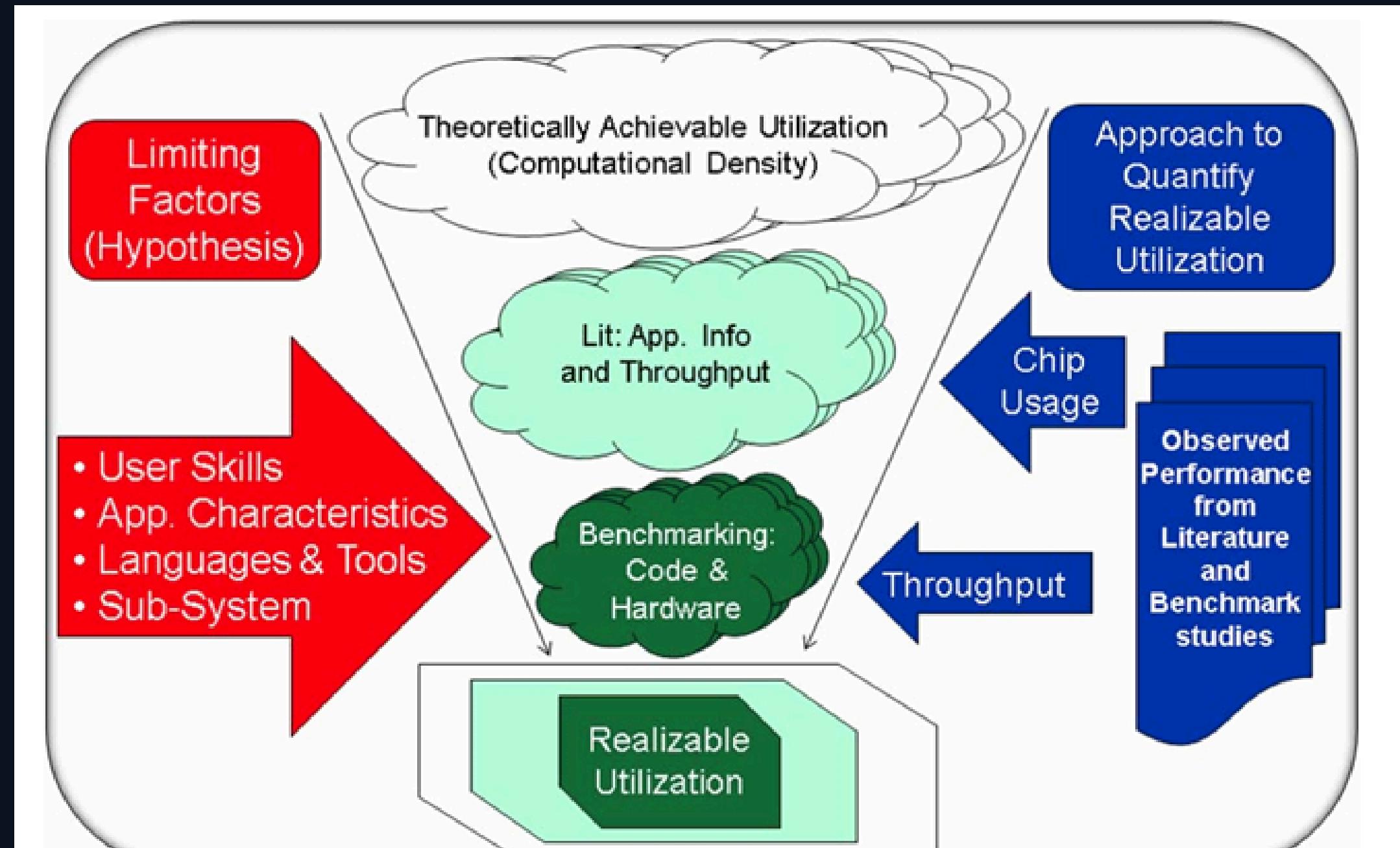


Fig. 1. Concept Diagram of Realizable Utilization

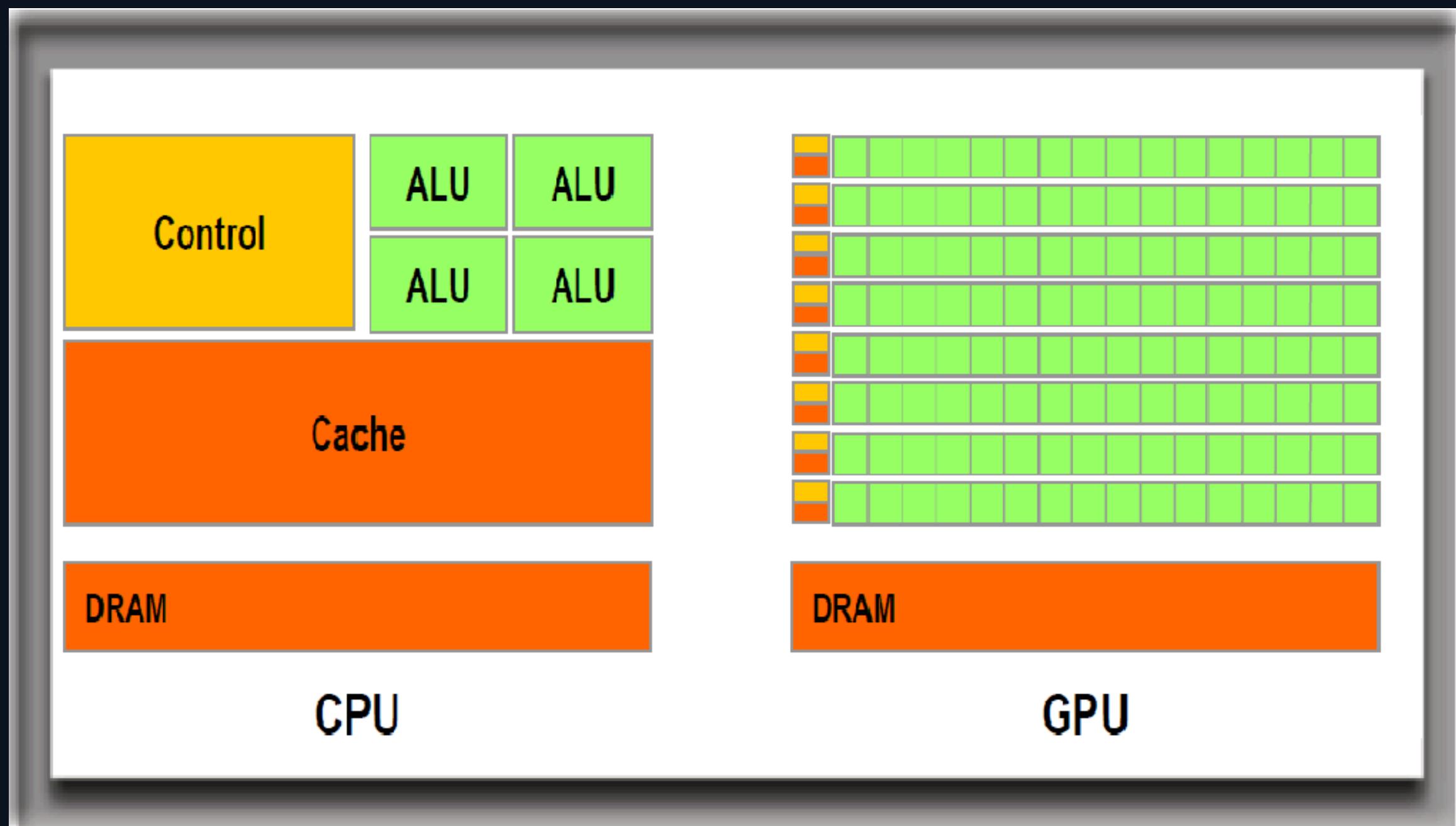
# UTILIZAÇÃO REALIZÁVEL

$$U_{realized} = \frac{R_{throughput}}{\alpha \cdot CD_{device}}$$

$$CD_{SPFP} = f \times \sum_i \frac{N_i}{CPI_i}$$

(SPFP: single-precision floating point)

# CPU VS GPU



# CPU'S VS GPU'S

## Comparação entre Tecnologias

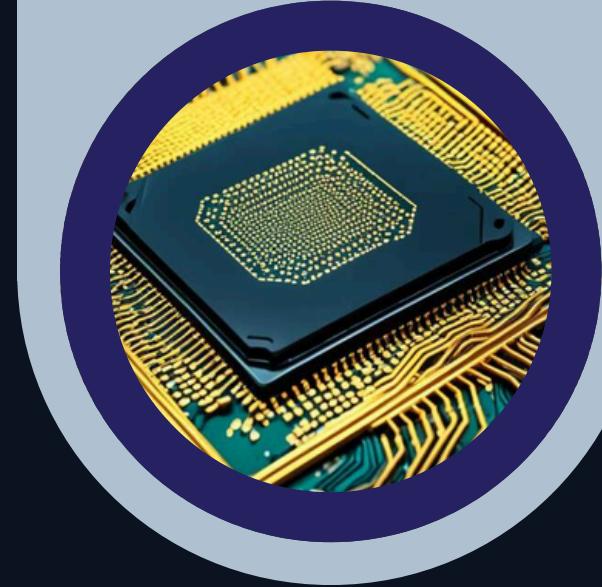
- **GPUs:** Altamente paralelizáveis, mas apresentam baixo aproveitamento da capacidade em muitas aplicações.
- **CPUs:** Mais flexíveis e eficientes em tarefas com menor paralelismo, alcançando melhor utilização realizável.

# USO DO RU PARA PROFISSIONAIS

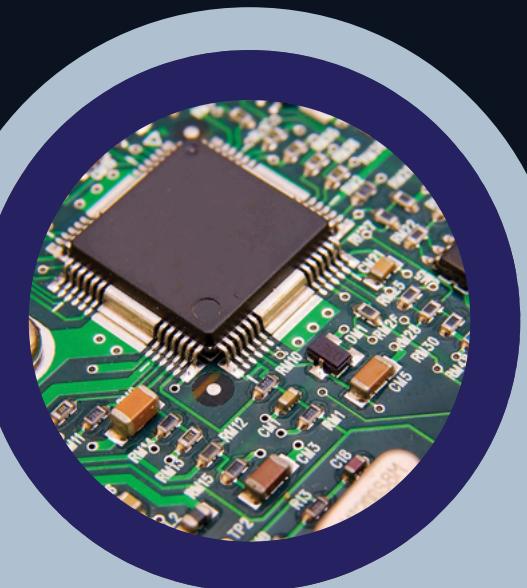


Permite que os desenvolvedores comparem dispositivos antes de escolherem uma plataforma;

Oferece feedback sobre o desenvolvimento e a otimização dos kernels;



Acontecem durante o processo de design do dispositivo, antes de ser fabricado;



Ajuda a selecionar plataformas de aceleração de aplicações antes de investir em hardware.



# RESULTADOS - 3 TIPOS DE KERNELS ARITMÉTICOS

- **Multiplicação de matrizes:** melhores pontuações obtidas nas GPUs da Série GeForce (mas as RUs diminuem significativamente à medida que o intervalo de CD aumenta)
- **Decomposição de matrizes:** A GPU com maior desempenho foi a GeForce 8800 GTX, com uma pontuação de 55,56%.
- **Simulações K-CORP:** mostraram os valores de RUs mais altos para GPUs e possuem a maior faixa de pontuações (1% a 99%)

# ANÁLISE GERAL E CONCLUSÕES

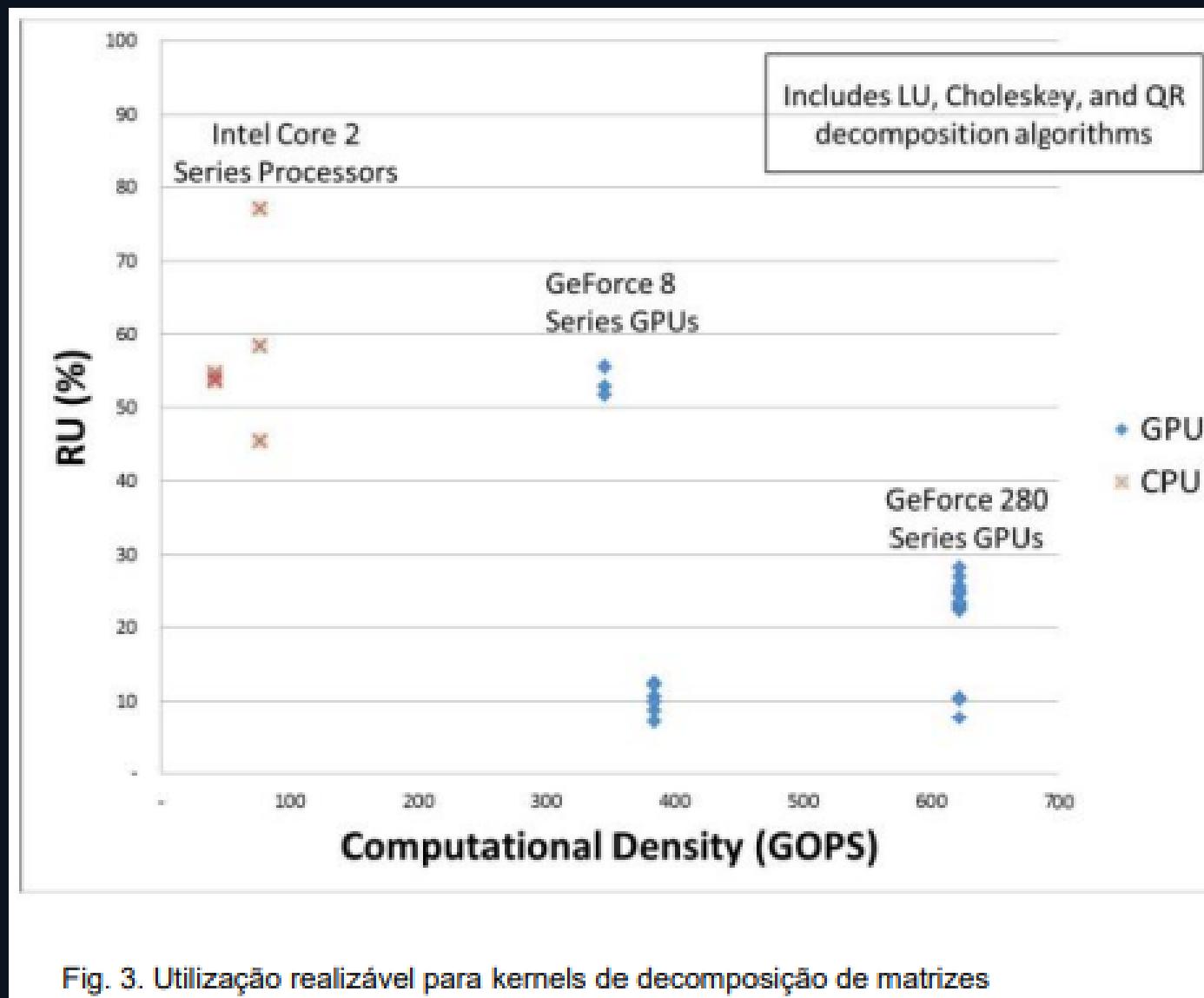


Fig. 3. Utilização realizável para kernels de decomposição de matrizes

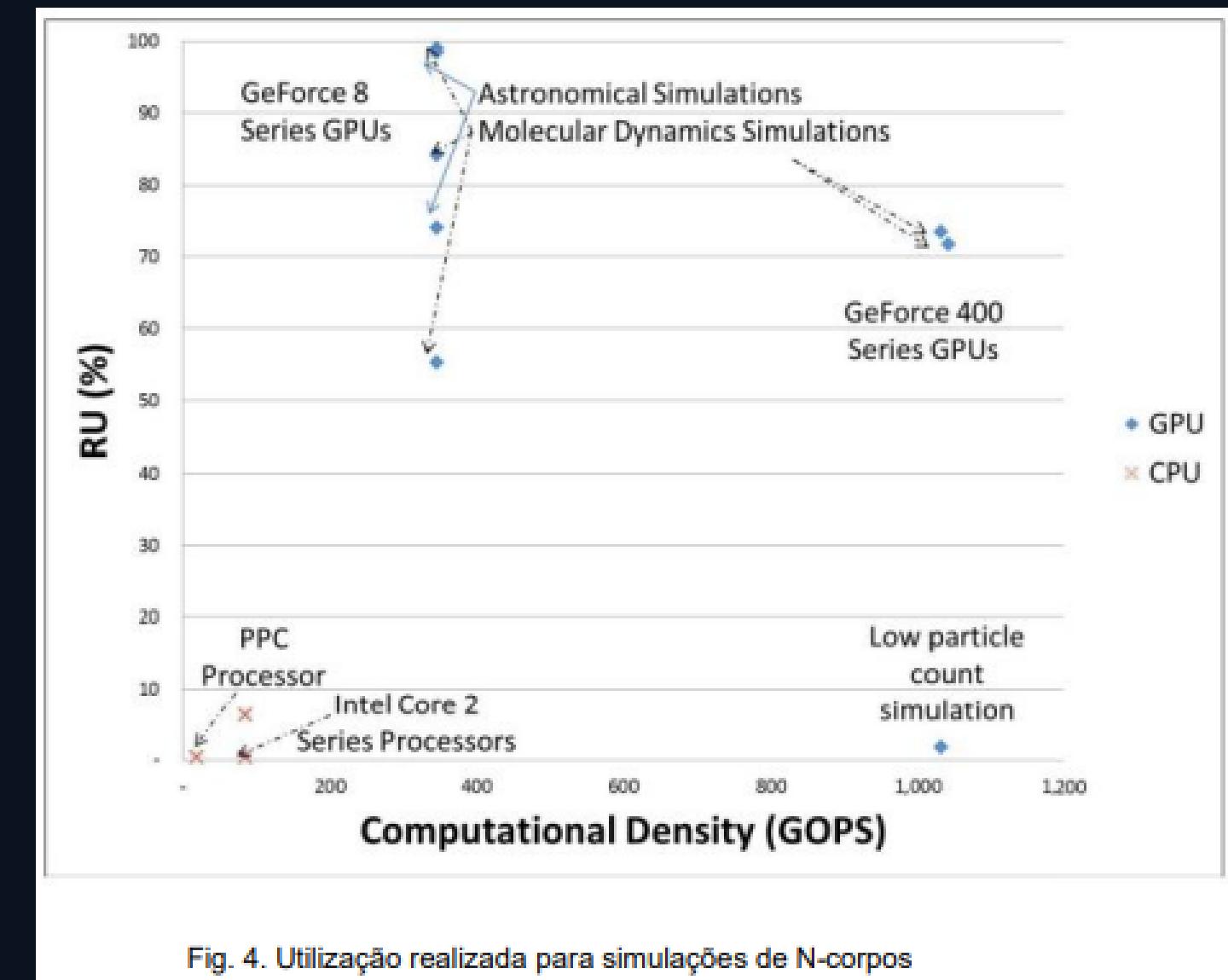


Fig. 4. Utilização realizada para simulações de N-corpos

## ANÁLISE GERAL E CONCLUSÕES

Para kernels de multiplicação de matriz, reforça a tendência de que mesmo que o desempenho bruto esteja aumentando nos hardwares, o desempenho não está acompanhando as capacidades teóricas. E isto pode ser observado por pontuações CD mais altas.

Demonstra que os aplicativos não são capazes de capitalizar totalmente os recursos computacionais adicionados.

# TRABALHOS FUTUROS

## Inclusão de Novos Dispositivos

Incorporar processadores ARM e FPGAs à análise, além de outros dispositivos emergentes, para expandir o escopo das métricas RU e CD.

## Análise de Consumo de Energia e Memória

Considerar características de consumo de energia e uso de memória como fatores complementares na avaliação de desempenho, permitindo uma análise mais holística da eficiência dos dispositivos.

## Estudos Avançados de Benchmarking

Realizar estudos mais detalhados para identificar os fatores específicos que levam às pontuações RU mais baixas, explorando como características do hardware, ferramentas ou aplicações impactam o desempenho realizável.

Muito  
Obrigado!