
Aula 02: Arquitetura de Computadores – Questões de desempenho

Prof. Hugo Puertas de Araújo
hugo.puertas@ufabc.edu.br
Sala: 509-2 (5º andar / Torre 2)



Arquitetura de Computadores

■ Objetivos de aprendizagem

- Compreender as principais questões de desempenho que se relacionam com o projeto do computador.
- Explicar as razões de mudar para a organização multicore e entender a relação entre recursos de cache e de processador em um chip único.
- Fazer distinção entre organizações multicore, MIC e GPGPU.
- Resumir algumas das questões relacionadas à avaliação do desempenho do computador.
- Discutir os benchmarks do SPEC.
- Explicar as diferenças entre as médias aritmética, harmônica e geométrica.

Elaboração do projeto visando o desempenho

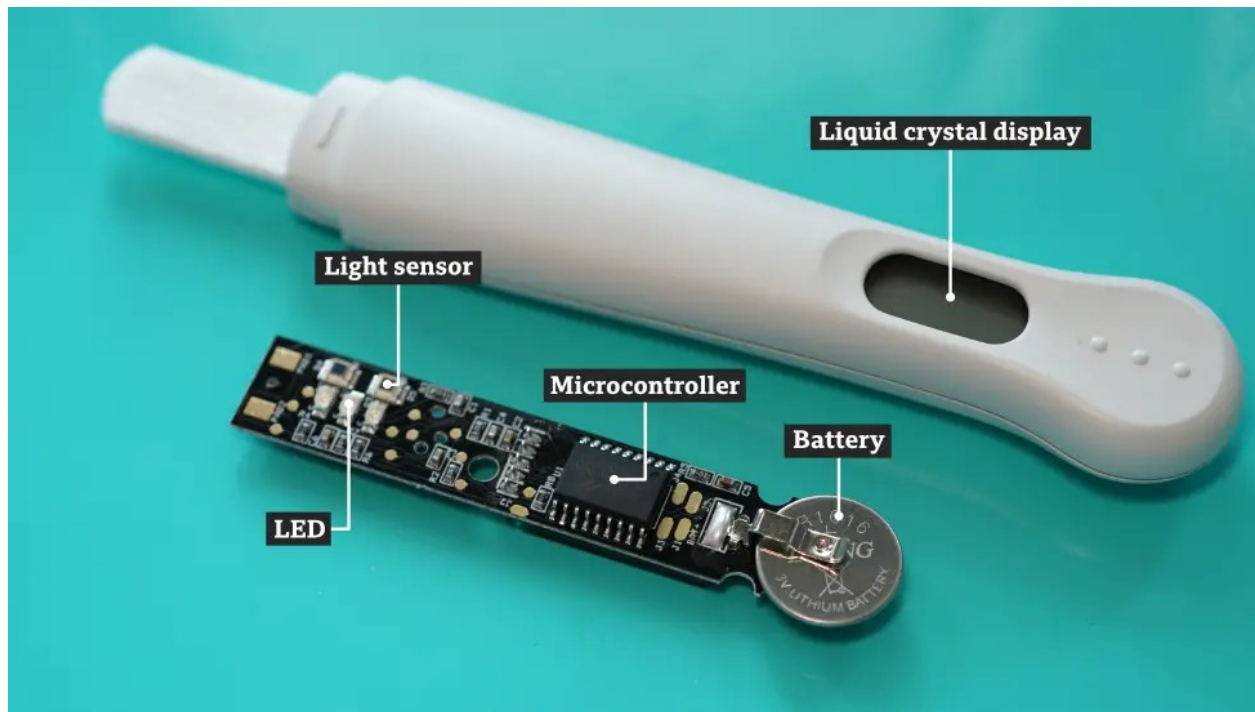
15 anos atrás



Mesmo desempenho



Elaboração do projeto visando o desempenho



■ Elaboração do projeto visando o desempenho

- O que dá aos processadores Intel x86 ou computadores mainframe da IBM uma potência incrível é a busca implacável de velocidade pelos fabricantes de chip de processador.
- Consequentemente, os projetistas de processadores deverão aparecer com técnicas ainda mais elaboradas para alimentar o monstro. Por exemplo:
 - ❖ Realização de pipeline
 - ❖ Predição de desvio
 - ❖ Execução superescalar
 - ❖ Análise de fluxo de dados
 - ❖ Execução especulativa

Balanço do desempenho

- Outra área de foco de projeto é o tratamento dos dispositivos de E/S.
- À medida que os computadores se tornam mais rápidos e mais capazes, aplicações sofisticadas são desenvolvidas para dar suporte ao uso de periféricos com demandas intensas de E/S.
- A chave em tudo isso é o equilíbrio.
- Os projetistas constantemente lutam para equilibrar as demandas de fluxo e processamento dos componentes do processador, memória principal, dispositivos de E/S e estruturas de interconexão.

■ Melhorias na organização e na arquitetura do chip

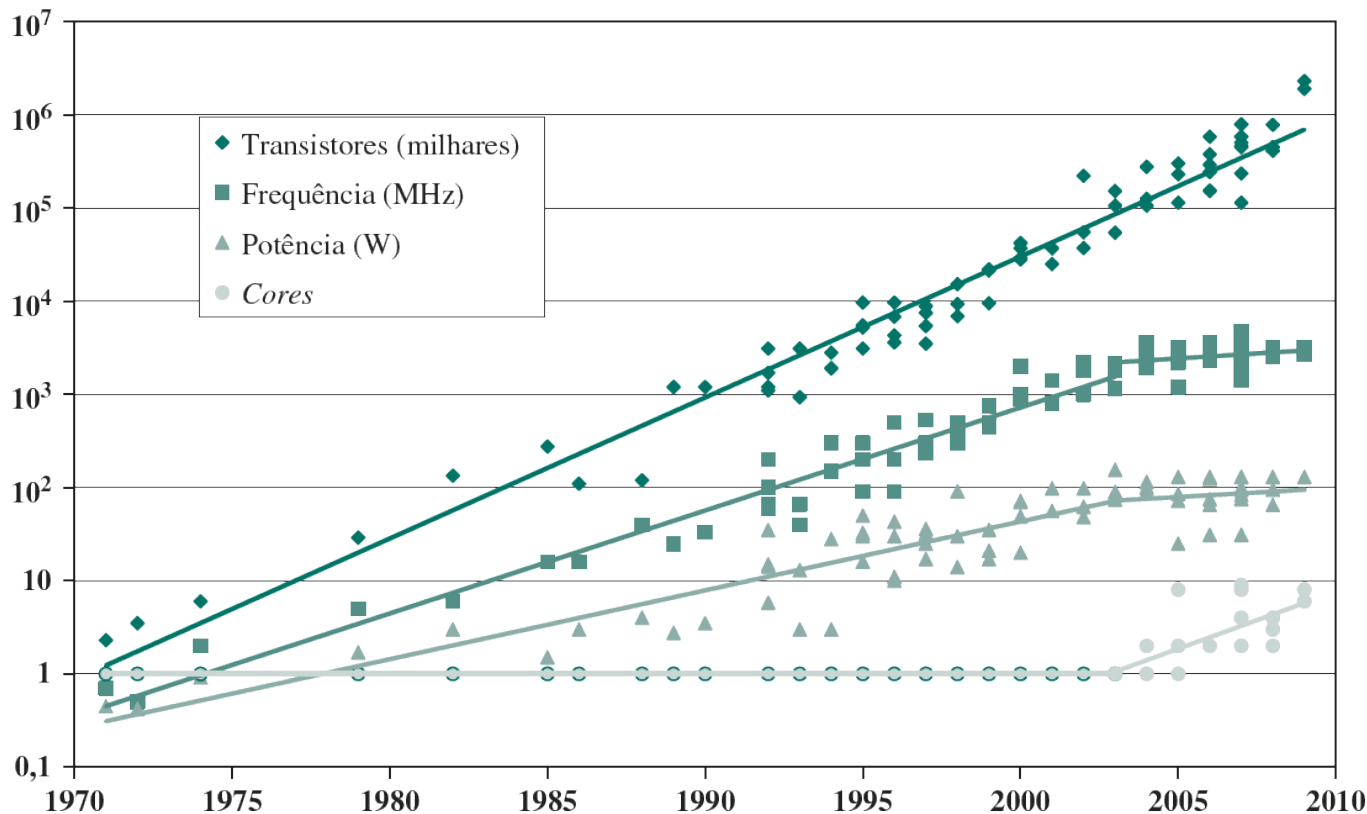
- Permanece a necessidade de aumentar a velocidade do processador e para isso, existem três técnicas:
 - i. Aumentar a velocidade de hardware do processador.
 - ii. Aumentar o tamanho e a velocidade das caches interpostas entre o processador e a memória principal.
 - iii. Fazer mudanças na organização e na arquitetura do processador, que aumentam a velocidade efetiva da execução da instrução.

Melhorias na organização e na arquitetura do chip

- À medida que a velocidade do clock e a densidade lógica aumentam, diversos obstáculos tornam-se mais significativos:
 - i. **Potência**: à medida que a densidade da lógica e a velocidade do clock em um chip aumentam, também aumenta a densidade de potência (Watts/cm²).
 - ii. **Atraso de RC**: o atraso aumenta à medida que o produto RC aumenta.
 - iii. **Latência e taxa de transferência da memória**: a velocidade de acesso à memória (latência) e a taxa de transferência limitam as velocidades do processador.

Melhorias na organização e na arquitetura do chip

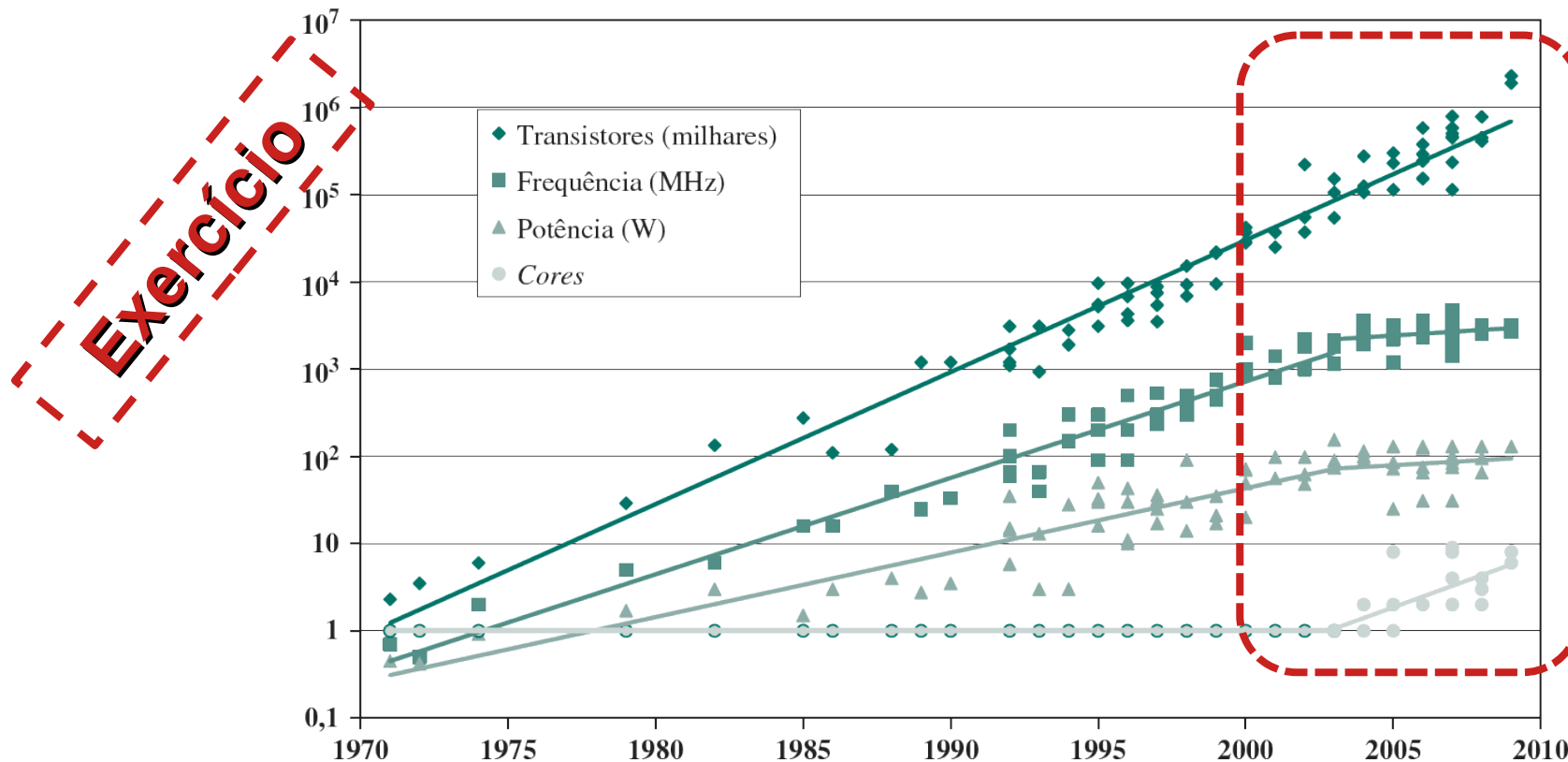
Tendências dos processadores:



Melhorias na organização e na arquitetura do chip

Tendências dos processadores:

O que aconteceu aqui?



■ Multicore, MICs e GPGPUs

- O uso de diversos processadores em um único chip, também chamado de múltiplos cores ou multicore, proporciona o potencial para aumentar o desempenho sem aumentar a frequência do clock.
- A estratégia é usar dois processadores simplificados no chip em vez de um processador mais complexo.
- Os chips de dois cores foram rapidamente seguidos por chips de quatro cores, oito, dezesseis e assim por diante.
- Os fabricantes de chip estão agora no processo de dar um enorme salto, com mais de 50 cores por cada um.

■ Multicore, MICs e GPGPUs

- O salto no desempenho têm levado ao surgimento de um novo termo: muitos cores integrados (MIC — do inglês, Many Integrated Cores).
- A estratégia do multicore e do MIC envolve uma coleção homogênea de processadores de uso geral em um único chip.
- Ao mesmo tempo, os fabricantes de chip estão buscando outra opção de desenvolvimento:
 - ❖ Um chip com múltiplos processadores de uso geral mais unidades de processamento gráfico (GPUs — do inglês, Graphics Processing Units).

■ Multicore, MICs e GPGPUs

- Em termos gerais, uma GPU é um core desenvolvido para desempenhar operações paralelas em dados gráficos.
- Tradicionalmente encontrada em uma placa plug-in (adaptador de vídeo), ela é usada para codificar e renderizar gráficos 2D e 3D, bem como para processar vídeos.
- Quando uma grande gama de aplicações é suportada por um processador, o termo GPUs de computação de uso geral (GPGPUs — em inglês, General-Purpose computing on GPUs) é usado.

Lei de Amdahl

- A lei de Amdahl foi proposta primeiro por Gene Amdahl em 1967 (AMDAHL, 1967, 2013) e lida com o potencial *speedup* de um programa usando múltiplos processadores em comparação com um único processador.
- Considere um programa sendo executado em um único processador de modo que uma fração $(1 - f)$ do tempo de execução envolve o código, que é inerentemente sequencial, e uma fração f que envolve o código que é infinitamente paralelizável sem sobrecarga no escalonamento.
- Considere que T é o tempo de execução total do programa que usa um único processador.

Lei de Amdahl

- Então, o speedup (aumento de velocidade) mediante o uso de um processador paralelo com N processadores que exploram completamente a parte paralela do programa se dá da seguinte forma:

$$\begin{aligned} \text{Speedup} &= \frac{\text{Tempo para executar o programa em um único processador}}{\text{Tempo para executar o programa em } N \text{ processadores paralelos}} \\ &= \frac{T(1-f) + Tf}{T(1-f) + \frac{Tf}{N}} = \frac{1}{(1-f) + \frac{f}{N}} \end{aligned}$$

Lei de Amdahl – Exercício

- Calcule o speedup alcançado se uma tarefa tiver 35% do seu tempo de execução passível de paralelização, sendo então executada em 2 processadores ao mesmo tempo.
- Qual seria o speedup da tarefa acima caso fosse possível utilizar 5 processadores em paralelo?

Lei de Amdahl – Exercício

- Calcule o speedup alcançado se uma tarefa tiver 35% do seu tempo de execução passível de paralelização, sendo então executada em 2 processadores ao mesmo tempo.

$$S = \frac{1}{1 - 0,35 + \frac{0,35}{2}} = 1,21$$

- Qual seria o speedup da tarefa acima caso fosse possível utilizar 5 processadores em paralelo?

$$S = \frac{1}{1 - 0,35 + \frac{0,35}{5}} = 1,39$$

Lei de Amdahl – Exercício

- Qual é o speedup máximo que se obtém para uma tarefa que é 90% paralelizável?

Lei de Amdahl – Exercício

- Qual é o speedup máximo que se obtém para uma tarefa que é 90% paralelizável?

$$S = \frac{1}{1 - 0,90 + \frac{0,90}{N}} = \frac{1}{0,10} = 10$$

$$P / N \rightarrow \infty$$

Lei de Amdahl

- A lei de Amdahl pode ser generalizada para avaliar um desenvolvimento ou uma melhoria de técnica em um sistema computacional.
- Considere qualquer aumento em uma característica de um sistema que resulta em um speedup.
- O speedup pode ser expresso como:

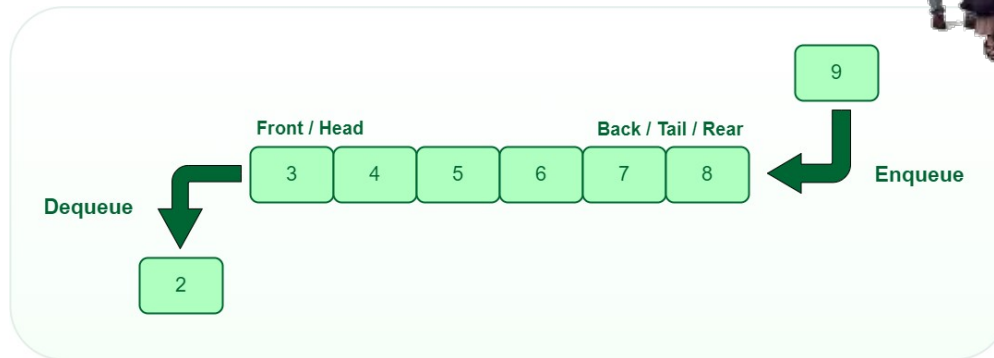
$$Speedup = \frac{\text{Desempenho depois do aumento}}{\text{Desempenho antes do aumento}} = \frac{\text{Tempo de execução antes do aumento}}{\text{Tempo de execução depois do aumento}}$$

Lei de Little

- A lei de Little relaciona esses três fatores (L: número de itens na fila; I: taxa de chegada de novos itens; W: tempo médio de atendimento de um item) como: $L = I \cdot W$.
- Para entender a fórmula de Little, considere o seguinte argumento, que foca na experiência de um único item:
 - ❖ Quando o item chega, ele encontra uma média de L itens à frente dele, um sendo servido e o restante em uma fila.
 - ❖ Quando o item deixar o sistema antes de ser servido, ele deixará para trás em média o mesmo número de itens no sistema, nomeadamente L, porque L é definido como número médio de itens em espera.

Lei de Little

- Para entender a fórmula de Little, considere o seguinte argumento, que foca na experiência de um único item:
 - ❖ Quando o item chega, ele encontra uma média de L itens à frente dele, um sendo servido e o restante em uma fila.
 - ❖ Quando o item deixar o sistema antes de ser servido, ele deixará para trás em média o mesmo número de itens no sistema, nomeadamente L , porque L é definido como número médio de itens em espera.



Lei de Little

- Além disso, o tempo médio que o item esteve no sistema foi W .
- Desde que os itens chegaram a uma taxa de I , podemos dizer que no tempo W um total de $I \cdot W$ itens deve ter chegado.
- Assim, $L = I \cdot W$.
- Para resumir, sob condições estáveis, o número médio de itens em um sistema de enfileiramento é igual à taxa média em que os itens chegam, multiplicados pelo tempo médio que um item gasta no sistema.

■ Medidas básicas de desempenho do computador

■ Velocidade de clock

- ❖ A velocidade de um processador é ditada pela frequência de pulso produzida pelo clock, medida em ciclos por segundo, ou Hertz (Hz).
- ❖ A taxa de pulsos é conhecida como frequência do clock ou velocidade de clock.
- ❖ Um incremento (ou pulso) do clock é conhecido como um ciclo de clock ou um período do clock.
- ❖ O tempo entre os pulsos é o tempo de ciclo.

■ Medidas básicas de desempenho do computador

■ Taxa de execução de instrução

- ❖ Um parâmetro importante é a média de ciclos por instrução (CPI — do inglês, Cycles Per Instruction) para um programa.
- ❖ Podemos calcular um CPI geral como a seguir:

$$CPI = \frac{\sum_{i=1}^n (CPI_i \times I_i)}{I_c}$$

← Total de instruções

- ❖ O tempo T do processador necessário para executar determinado programa pode ser expresso como:

$$T = I_c \times CPI \times \tau$$

← Período do clock

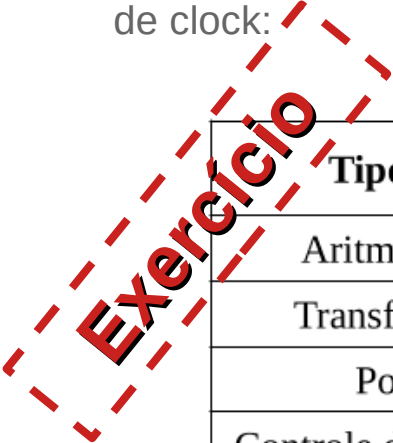
Medidas básicas de desempenho do computador

$$CPI = \frac{\sum_{i=1}^n (CPI_i \times I_i)}{I_c}$$

$$T = I_c \times CPI \times \tau$$

$$MIPS = \frac{I_c}{T \cdot 10^6}$$

- Um programa de benchmark é executado em um processador a 40 MHz. O programa executado consiste em 100.000 execuções de instrução, com os seguintes tipos de instruções e número de ciclos de clock:



Tipo de instrução	Número de instruções	Ciclos por instrução
Aritmética de inteiros	45.000	1
Transferência de dados	32.000	2
Ponto flutuante	15.000	2
Controle de fluxo de execução	8.000	2

- Determine o CPI efetivo, a taxa de MIPS e o tempo de execução para esse programa.

■ Medidas básicas de desempenho do computador

$$CPI = \frac{\sum_{i=1}^n (CPI_i \times I_i)}{I_c} = 1,55 \quad T = I_c \times CPI \times \tau = 3,87 \text{ ms} \quad MIPS = \frac{I_c}{T \cdot 10^6} = 25,8$$

- Um programa de benchmark é executado em um processador a 40 MHz. O programa executado consiste em 100.000 execuções de instrução, com os seguintes tipos de instruções e número de ciclos de clock:

Tipo de instrução	Número de instruções	Ciclos por instrução
Aritmética de inteiros	45.000	1
Transferência de dados	32.000	2
Ponto flutuante	15.000	2
Controle de fluxo de execução	8.000	2

- Determine o CPI efetivo, o tempo de execução e a taxa de MIPS para esse programa.

■ Cálculo da média

- Dado um conjunto de números reais n (x_1, x_2, \dots, x_n), as três médias são definidas da seguinte maneira:

Média aritmética

$$MA = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Média geométrica

$$MG = \sqrt[n]{x_1 \times \dots \times x_n} = \left(\prod_{i=1}^n x_i \right)^{1/n} = e^{\left(\frac{1}{n} \sum_{i=1}^n \ln(x_i) \right)}$$

Média harmônica

$$MH = \frac{n}{\left(\frac{1}{x_1} \right) + \dots + \left(\frac{1}{x_n} \right)} = \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i} \right)} \quad x_i > 0$$

Média aritmética

- Uma MA é uma medida apropriada se a soma de todas as medidas for um valor significativo e interessante.
- A MA de todas as execuções é uma boa medida dos desempenhos de sistemas em simulações e um bom número para uso em comparação entre sistemas.
- A MA usada como uma variável baseada em tempo (por exemplo, segundos), como tempo de execução de programa, tem uma propriedade importante que é diretamente proporcional ao tempo total.
- Se o tempo total dobra, o valor médio segue o mesmo caminho.

Média harmônica

- Para algumas situações, a taxa de execução de um sistema pode ser vista como uma medida útil do valor do sistema.
- Pode ser a taxa de execução de instruções, medida em MIPS ou MFLOPS, ou uma taxa de execução de programa, que mede a taxa na qual um dado tipo de programa pode ser executado.
- Considere como gostaríamos que a média calculada se comportasse.
- Não faz sentido dizer que gostaríamos que a taxa média fosse proporcional à taxa total, onde a taxa total é definida como a soma das taxas individuais.

Média harmônica

- Vemos que a taxa de execução de MA é proporcional à soma dos inversos dos tempos de execução, que não é a mesma conforme for inversamente proporcional à soma de tempos de execução.
- Assim, a MA não tem a propriedade desejada.
- A MH produz o seguinte resultado:

$$MH = \frac{n}{\sum_{i=1}^n \left(\frac{1}{R_i} \right)} = \frac{n}{\sum_{i=1}^n \left(\frac{1}{Z/t_i} \right)} = \frac{nZ}{\sum_{i=1}^n t_i}$$

Média geométrica

- Com relação às mudanças nos valores, a MG confere peso igual para todos os valores no conjunto de dados.
- Para a MG de uma razão, a MG das razões iguala a razão das MGs:

$$MG = \left(\prod_{i=1}^n \frac{Z_i}{t_i} \right)^{1/n} = \frac{\left(\prod_{i=1}^n Z_i \right)^{1/n}}{\left(\prod_{i=1}^n t_i \right)^{1/n}}$$

- Pode haver casos em que a MA de um conjunto de dados é maior que aquela do outro conjunto, mas a MG é menor.

■ Média geométrica

- Uma propriedade da MG que tem tido apelo na análise de benchmark é que ela proporciona resultados consistentes quando mede o desempenho relativo das máquinas.
- De fato isso é para o que os benchmarks são usados em primeiro lugar.
- Os resultados, como temos visto, são expressos em termos de valores que são normalizados para a máquina de referência.
- É seguro dizer que nenhum número único pode proporcionar todas as informações necessárias para comparar resultados entre sistemas.

Benchmarks e SPEC

- Os benchmarks proporcionam orientações para os clientes que tentam decidir qual sistema comprar.
- Pode ser útil para vendedores e desenvolvedores na determinação de como desenvolver sistemas para atingir as metas de benchmark.
- Weicker (1990) lista as características desejadas de um programa de benchmark:
 - i. É escrito em uma linguagem de alto nível, tornando-o portátil entre diferentes máquinas.
 - ii. Representa um tipo particular de estilo de programação, como programação de sistemas, programação numérica ou programação comercial.
 - iii. Pode ser medido com facilidade.
 - iv. Tem ampla distribuição.

Benchmarks e SPEC

- Para entender melhor os resultados publicados de um sistema usando CPU2006, definimos os seguintes termos usados na documentação da SPEC:
 - ❖ Benchmark: um programa escrito em uma linguagem de alto nível que pode ser compilado e executado em qualquer computador que implemente o compilador.
 - ❖ Sistema em teste: é o sistema a ser avaliado.
 - ❖ Máquina de referência: cada benchmark é executado e medido em sua máquina para estabelecer o tempo de referência para tal benchmark.

Benchmarks e SPEC

- ❖ Métrica de base: o compilador padrão com mais ou menos configurações padrão deve ser usado em cada sistema em teste para atingir resultados comparativos.
- ❖ Métrica de pico: possibilita aos usuários tentar otimizar o desempenho do sistema ao otimizar a saída do compilador.
- ❖ Métrica de velocidade: é simplesmente uma medida do tempo que leva para a execução de um benchmark compilado.
- ❖ Métrica de taxa: é uma medida de quantas tarefas um computador pode cumprir em certa quantidade de tempo.