

Desempenho do Computador

Arquitetura e Organização de Computadores

Prof. Lucas de Oliveira Teixeira

UEM

Introdução

Introdução

- Cada vez os computadores são usados em aplicações altamente intensivas, como processamento de imagem, reconhecimento de voz, simulações, redes neurais, entre outros.
- Com isso, os fabricantes sempre buscam aumentar o desempenho dos seus processadores.
- Assim, o desempenho dos processadores normalmente quadruplica o seu desempenho a cada três anos (isso diminuiu um pouco recentemente).

Desempenho

- O aumento do desempenho nos processadores ocorre em duas frentes: melhorias na tecnologia e na arquitetura e organização dos computadores.

Tecnologia:

- A melhoria na tecnologia permite uma redução no tamanho do transistor.
- Isso aumenta a quantidade de transistores em um chip e a distância entre os transistores.
- Com isso, é possível aumentar a velocidade (clock) de operação dos processadores.

Arquitetura e organização:

- A melhoria na organização do processador também possui um impacto enorme no desempenho.
- As técnicas mais utilizadas: empacotamento de instruções, aumento da largura de banda, pipeline, memória cache, previsão de desvio, reorganização de instruções, execução especulativa e múltiplos núcleos.

Empacotamento de instruções:

- Criar instruções mais complexas (que fazem mais coisa).
- Assim, os programas ficam menores e o processador precisa buscar menos instruções.

Aumento da largura de banda:

- Aumentar a largura de banda do barramento.
- Com isso, mais informações podem ser transmitidas por vez.

Pipeline:

- A execução de uma instrução envolve diferentes etapas: busca da instrução, decodificação, busca de operandos, execução e escrita do resultado.
- Todas as instruções precisam passar por essas etapas.
- O pipeline organiza a execução de cada etapa de forma independente, assim ele consegue executar várias instruções ao mesmo tempo.

Memória cache:

- O processador precisa de um fluxo constante de instruções e dados para executar, porém a evolução da memória principal não acompanhou a velocidade do processador, ou seja, a memória é mais lenta que o processador.
- A memória cache é localizada dentro do processador e contém trechos da memória que foram recentemente utilizados.
- A ideia é simples: a probabilidade do processador precisar de instruções ou dados próximos aquele em que ele está executando é alta.

Previsão de desvio:

- Com o uso do pipeline, o processador não sabe se um desvio condicional será tomado ou não até a etapa de execução.
- Com isso, novas instruções entram no pipeline e podem ser desnecessárias se o desvio for tomado.
- Diante disso, o processador realiza a previsão de desvios para tentar de antemão decidir se o desvio será tomado ou não.
- A taxa de acerta é alta, uma vez que a maioria dos desvios representam laços de repetição.

Reorganização de instruções:

- Com o uso do pipeline, o processador não sabe o resultado de uma operação aritmética ou lógica até a etapa de execução.
- Assim, se instruções imediatamente posteriores precisarem do resultado da operação, elas terão que ficar esperando.
- O processador tenta reorganizar as instruções para evitar o máximo possível essa dependência entre os operandos.

Execução especulativa:

- Uma melhoria sobre a previsão de desvio é a execução especulativa.
- A ideia é executar instruções antes que suas dependências sejam resolvidas.
- Para a dependência de dados, é possível fazer a predição de valores e endereços.
- Para a dependência de controle de fluxo, é possível utilizar a previsão de desvio ou simplesmente executar ambos os caminhos do desvio.

Múltiplos núcleos:

- Um chip comum possui vários núcleos de processamento independentes.
- Essa técnica foi adotada por causa da dificuldade em aumentar a frequência de funcionamento do processador (difícilmente um processador passa de 4 GHz).
- Cada um é um processador completo e normalmente todos possuem as caches L1 e L2 separadas, e a cache L3 unificada.
- Isso permite que múltiplas threads ou programas sejam executados de forma paralela.

Avaliação de Desempenho

- Avaliar o desempenho de dois processadores diferentes é uma tarefa complexa.
- O desempenho não depende só da velocidade do processador, depende também do conjunto de instruções, compilador, linguagem de alto nível, habilidade do programador, entre outros.

Taxa de clock:

- Um pulso de clock é um sinal de sincronização entre as unidades do processador.
- Relembre, um computador é uma máquina discreta então ele sempre funciona em passos discretos.
- Quanto mais rápido o ciclo de clock, mais rápido o processador executa os programas.
- Essa velocidade não é arbitrária, mas sim depende do aspecto físico do processador.

Métricas:

- CPI: Ciclos por instrução.
- MIPS: Milhões de instruções por segundo.

Avaliação de Desempenho

Por exemplo, determine o CPI, a taxa MIPS e o tempo de execução do programa abaixo considerando um processador de 40 MHz:

Tipo de instrução	Quantidade de instruções	CPI
Aritmética de inteiros	45000	1
Transferência de dados	32000	2
Ponto flutuante	15000	2
Transferência de controle	8000	2

- Para determinar o CPI, precisamos calcular o total de instruções e o total de ciclos necessários para a execução do programa.
- Total de instruções:
 $45000 + 32000 + 15000 + 8000 = 100000$ instruções.
- Total de ciclos:
 $(45000 \times 1) + (32000 \times 2) + (15000 \times 2) + (8000 \times 2) =$
 $45000 + 64000 + 30000 + 16000 = 155000$ ciclos.
- CPI: $155000/100000 = 1,55$

- Para determinar a taxa MIPS, precisamos saber o CPI e a frequência do processador.
- $\text{MIPS} = 40\text{MHz} / 1,55 = 40 \times 10^6 / 1,55 = 25,8$ milhões de instruções por segundo.

- Para determinar o tempo de execução, precisamos saber o total de ciclos gasto pelo programa e a frequência do processador.
- Total de ciclos: 155000 ciclos.
- Tempo de execução:
 $155000 / 40\text{MHz} = 155000 / 40000000 = 0,0039$ segundos.

Avaliação de Desempenho

Por exemplo, considere um programa com 2 milhões de instruções executando em um processador de 500 MHz, determine o CPI, a taxa MIPS e o tempo de execução do programa considerando a distribuição abaixo:

Tipo de instrução	Quantidade de instruções	CPI
Aritmética e lógica	60%	1
Load/store com acerto de cache	18%	2
Desvio	12%	4
Load/store com falha de cache	10%	8

Avaliação de Desempenho

Tipo de instrução	Quantidade de instruções	CPI
Aritmética e lógica	1 200 000	1
Load/store com acerto de cache	360 000	2
Desvio	240 000	4
Load/store com falha de cache	200 000	8

- Para determinar o CPI, precisamos calcular o total de instruções e o total de ciclos necessários para a execução do programa.
- Total de instruções: 2000000 instruções.
- Total de ciclos:
 $(1200000 \times 1) + (360000 \times 2) + (240000 \times 4) + (200000 \times 8) =$
 $1200000 + 720000 + 960000 + 1600000 = 4480000$ ciclos.
- CPI: $4480000 / 2000000 = 2,24$

- Para determinar a taxa MIPS, precisamos saber o CPI e a frequência do processador.
- $\text{MIPS} = 500\text{MHz} / 2,24 = 500 \times 10^6 / 2,24 = 223,2$ milhões de instruções por segundo.

- Para determinar o tempo de execução, precisamos saber o total de ciclos gasto pelo programa e a frequência do processador.
- Total de ciclos: 4480000 ciclos.
- Tempo de execução:
$$4480000 / 500\text{MHz} = 4480000 / 5000000000 = 0,00896$$
 segundos.

Avaliação de Desempenho

Por exemplo, considere duas máquinas diferentes ambas com uma taxa de clock de 200MHz e a seguinte distribuição de instruções:

Máquina	Tipo de instrução	Quantidade de instruções	CPI
A	Aritmética	8000000	1
	Load e Store	4000000	3
	Desvio	2000000	4
	Outros	4000000	3
B	Aritmética	10000000	1
	Load e Store	8000000	2
	Desvio	2000000	4
	Outros	4000000	3

Qual máquina é melhor? Quanto?

- Quantidade de instruções de A =
 $8000000 + 4000000 + 2000000 + 4000000 = 18000000$
instruções.
- Quantidade de ciclos de A = $(8000000 \times 1) + (4000000 \times 3) + (2000000 \times 4) + (4000000 \times 3) = 40000000$ ciclos.
- CPI A = $40000000 / 18000000 = 2,22$.

- Quantidade de instruções de B =
 $10000000 + 8000000 + 2000000 + 4000000 = 24000000$
instruções.
- Quantidade de ciclos de B = $(10000000 \times 1) + (8000000 \times 2) + (2000000 \times 4) + (4000000 \times 3) = 46000000$ ciclos.
- CPI B = $46000000 / 24000000 = 1,91$.

- A máquina B possui um CPI menor, logo é capaz de executar mais instrução por ciclo de clock.
- $\text{CPI A} / \text{CPI B} = 2,22 / 1,91 = 1,16$.
- Assim, a máquina B possui um speedup de 1,16 em relação à máquina A (nesse caso não vamos trabalhar com porcentagem de aumento).