Organização e Arquitetura de Computadores II

Introdução à Hierarquia de Memória

Capítulo 2, 6 do Jean-Loup Baer
Capítulo 4.2, 5.1, 5.2 do Monteiro
Capítulo 4 do Stallings
Capítulos 2.2 e 2.3 do Tanenbaum e Austin
Capítulo 5 do Hennessy e Patterson
Capítulo 7 do Patterson e Hennessy

Última alteração: 26/11/2017

Prof. Ney Laert Vilar Calazans

Baseado em notas de aulas originais do Prof. Dr. César Marcon

Índice

Introdução

- O desempenho de sistemas computacionais depende de três elementos fundamentais
  - **Computação**
  - **Comunicação**
  - **Armazenamento (Memórias)**
- Requisitos ideais de uma memória
  - **Tamanho ilimitado**
  - **Acesso instantâneo para escrita ou leitura de informações**
- Os requisitos de uma memória → Contraditórios
  - **Quanto maior a memória maior será o seu tempo de acesso**
- Solução
  - **Criar uma hierarquia de memória**
    - **"Ilusão" para o processador, de forma que a memória pareça grande e rápida o suficiente para não ser gargalo no sistema**
    - **Acesso transparente aos níveis de memória**

Exemplo de Estudante em uma Biblioteca

- **Na biblioteca, pesquisa tem o seguinte algoritmo**
  1. Ir até a estante de livros
  2. Procurar livro desejado
  3. Levar livro até a cadeira
  4. Consultar livro
  5. Se não concluiu pesquisa, ir para 1
- **Considerações**
  - Consulta com 10 livros
  - 1 minuto para ir e voltar da cadeira a estante
  - 1 minuto procurando o livro na estante
  - 30 segundos para consultar a informação desejada no livro
- **Tempo de consulta de cada livro**
  - 2 minutos e 30 segundos
- **Tempo total consumido**
  - 25 minutos

Exemplo de Estudante com Mesa Vazia

- **Novo algoritmo**
  1. Ir até a estante de livros
  2. Procurar livros desejados
  3. Levar livros até a mesa
  4. Consultar livros
  5. Se não terminou ir para 4
- **Consideração**
  - A mesa tem espaço para os 10 livros
- **Tempo total de pesquisa**
  - 16 minutos (1 para deslocamento, 10 para procura na estante e 5 para pesquisa de material nos livros)
- **Problemas possíveis**
  - Nem todos livros desejados cabem na mesa
  - Outro aluno ocupa parte da mesa, fazendo uma pesquisa diferente

Exemplo de Estudante com Folha de Rascunho

- **O acesso à folha é mais rápido que à mesa, mas na folha cabem menos informações**
  - Colocar na folha trechos dos livros que podem interessar
- **Este processo de seleção da informação pode continuar ...**

Princípios Fundamentais

- **Porque o tempo de acesso melhora em média?**
  - Princípio da Localidade ➔ Trabalho restrito a um grupo de livros
    - Localidade Espacial: A pesquisa de um determinado grupo de livros está localizada muito próxima
    - Localidade Temporal: De tempos em tempos o estudante volta a consultar um livro que já tinha consultado antes

Índice

Aspectos Importantes de Hierarquias de Memória

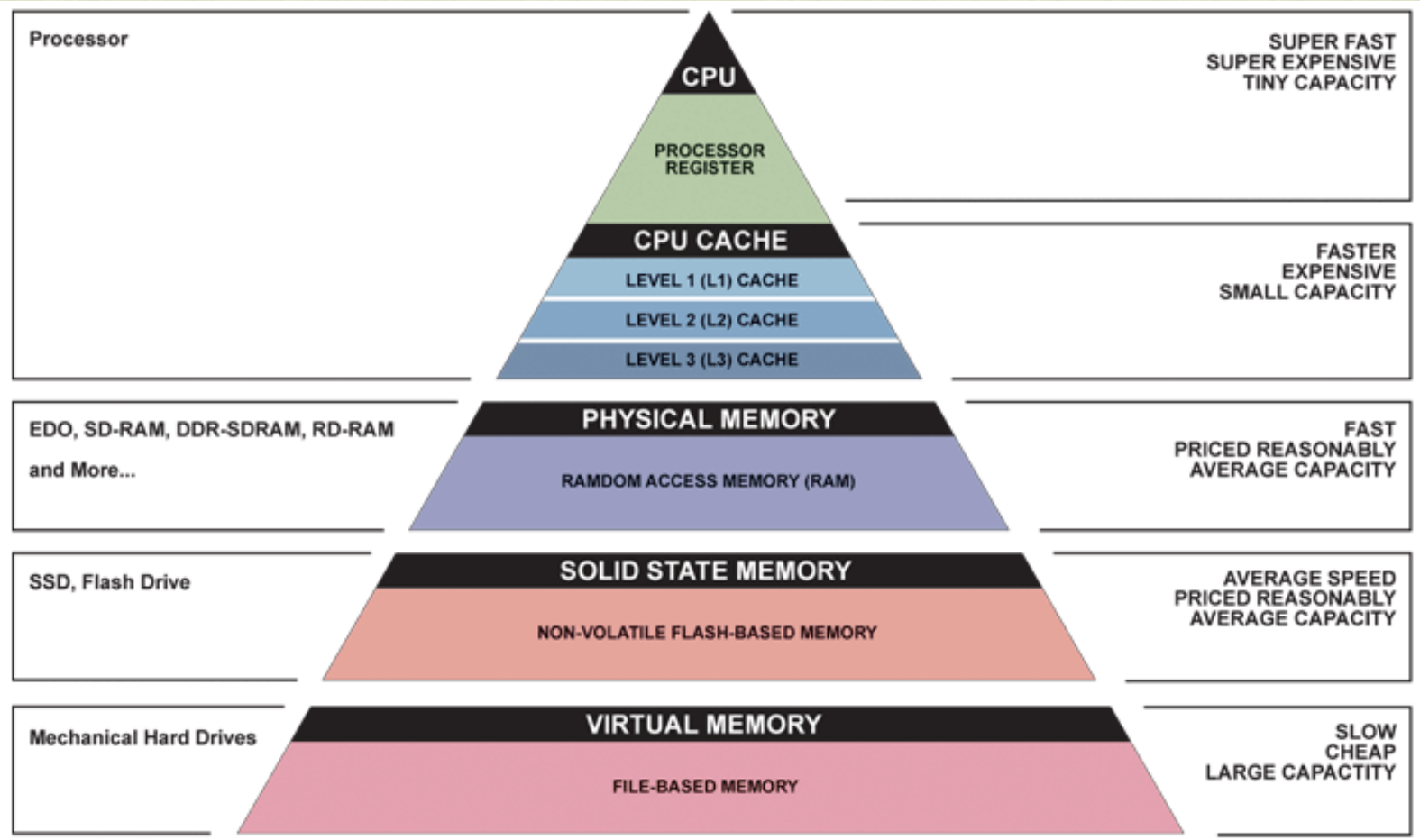
- **Ilusão de uma memória ilimitada e rápida**
  - Obtida devido aos níveis de acesso
- **Informações → transferidas para níveis mais altos**
- **Usa princípio de localidade espacial e temporal**
  - Se um endereço foi referenciado, existe grande probabilidade do endereço seguinte ser referenciado
    - Ex.: Execução Sequencial (Localidade Espacial)
  - Se um endereço foi referenciado, é provável que ele seja referenciado novamente em pouco tempo
    - Ex.: Loops (Localidade Temporal)
  - A forma de descrever um programa pode afetar o desempenho
- **Analise as seguintes situações, seja em termos de dados ou código**
  - Programa com apenas instruções seguidas de “go to” (vai para posição de memória)
  - Escolha da ordem de acesso a grandes vetores

Aspectos Importantes de Hierarquias de Memória

- **Transferência entre níveis por grupos de palavras (bloco, página)**
  - Diminuição de custo de transferência
  - Antecipar acessos (considerando o princípio da localidade espacial)
- **Movimentação de dados entre níveis necessita de *mecanismos***
- **Nas decisões estratégicas os mecanismos usam *políticas***
  - Ex.: Movimento de dados para um nível superior que já está cheio Quem retirar?
  - Decisão errada pode afetar desempenho do sistema como um todo
- **Tempo médio de acesso é reduzido quando os mecanismos conseguem manter as informações nos níveis mais altos**

Níveis de Hierarquias de Memória

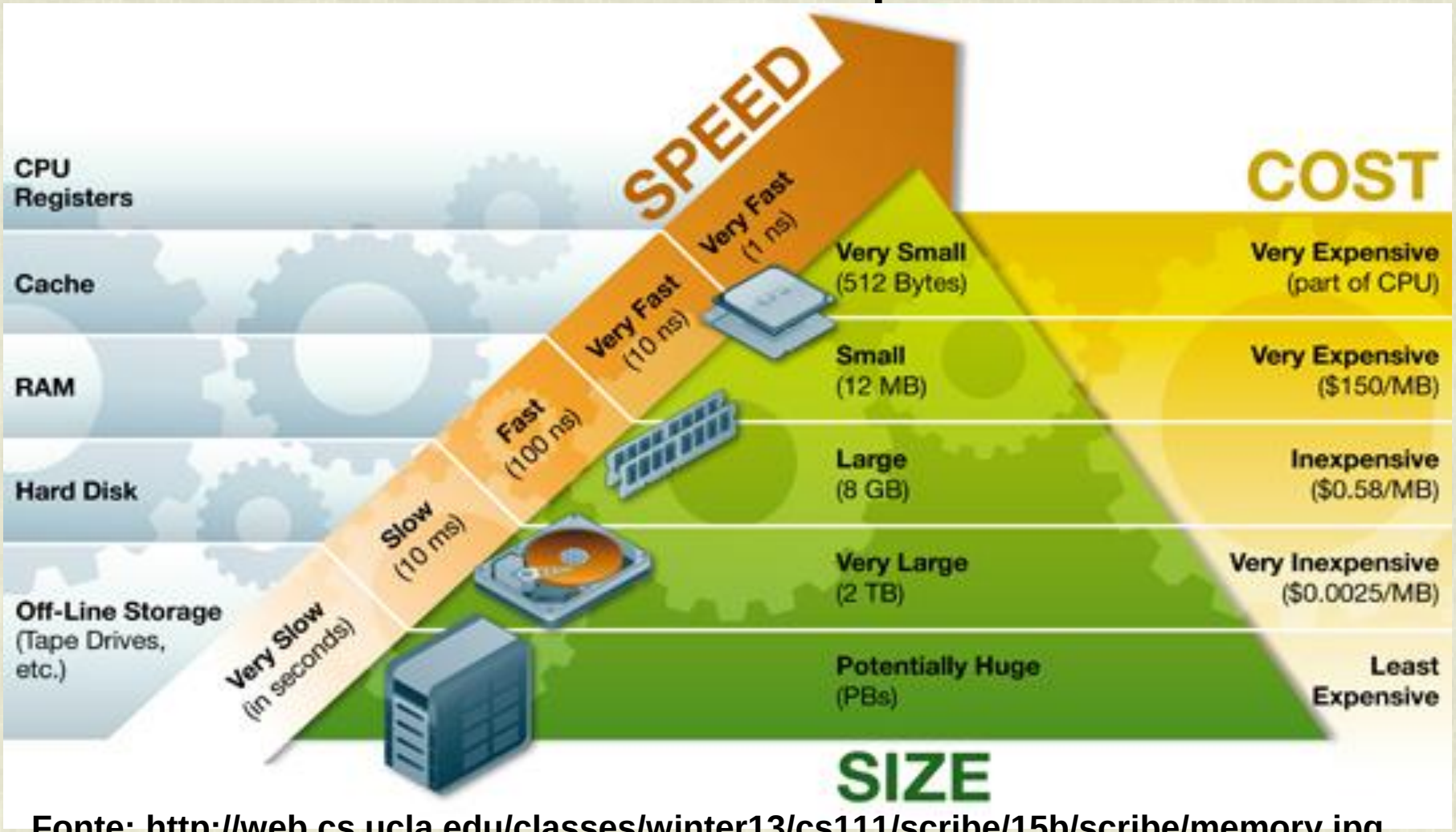
- **Níveis intermediários usados para amortizar a diferença de velocidade entre processador e memória**



▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

Comparação de Características da Hierarquia de Memória

- Comparação de **tempo de acesso**, **tamanho** e **custo**, entre diferentes níveis da hierarquia de memória



Fonte: http://web.cs.ucla.edu/classes/winter13/cs111/scribe/15b/scribe/memory.jpg

Definições Importantes

- **Hit – dado encontrado no nível procurado**
- **Miss – dado não encontrado no nível procurado**
- **Hit-rate – percentual de hits no nível, Ex.: 70%**
- **Miss-rate – percentual de *misses* no nível, Ex.: 30% (complementar ao Hit-rate)**
- **Hit-time – tempo de acesso ao nível incluindo tempo de ver se é *hit* ou *miss***
- **Miss-penalty – tempo médio gasto para que o dado não encontrado no nível seja transferido dos níveis mais baixos**
- **Average Memory Access Time (AMAT) – tempo médio efetivo para acessar um dado em certo nível de memória ➔ composição do Hit-time, Miss-rate e Miss-penalty do nível de memória inferior**
  - AMAT = L_1 Hit-time + L_1 Miss-rate * L_1 Miss-penalty
  - L_1 Miss-penalty = L_2 Hit-time + L_2 Miss-rate * L_2 Miss-penalty
  - L_2 Miss-penalty = L_3 Hit-time + L_3 Miss-rate * L_3 Miss-penalty
  - ...

Exercícios

1. Porque a implementação de uma memória quase-ideal, i.e. tamanho ilimitado e tempo de acesso desprezível, é tecnicamente contraditória?
2. Comente sobre o tempo de acesso, tamanho e custo ($/byte) dentro da hierarquia de memória
3. Porque não é necessário para o processador saber onde estão fisicamente os dados na hierarquia de memória?
4. Comente sobre os princípios de localidade que permitem a utilização eficiente da memória cache. Porque na ausência de cada um destes princípios a utilização de uma cache seria ineficiente?
5. Cite alguns problemas básicos do uso de memória cache e comente
6. O que você entende por cache hit e por cache miss?
7. Dadas as características da hierarquia de memória, o que provavelmente demorará mais tempo, o hit-time ou o miss-penalty? Justifique a resposta
8. Calcule o tempo médio efetivo de acesso (AMAT) a uma cache com Hit-ratio = 80%, Hit-time = 2 µs e Miss-penalty = 10 µs

Resposta de Exercícios

8. Calcule o tempo médio efetivo de acesso (AMAT) a uma cache com Hit-ratio = 80%, Hit-time = 2 µs e Miss-penalty = 10 µs

AMAT = Hit-time + (1- Hit-rate)* Miss-penalty

AMAT = 2 + (1 – 0.8) * 10

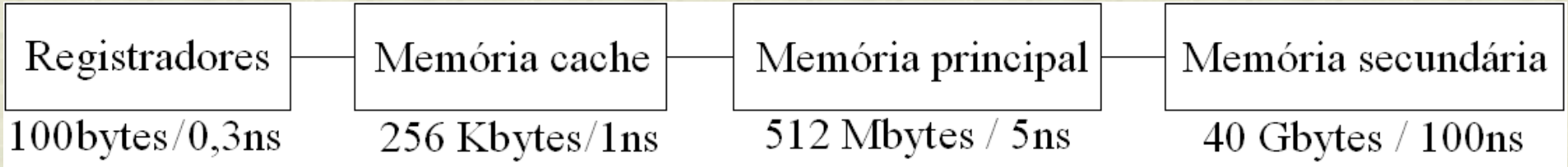
AMAT = 2 + 0.2 * 10

AMAT = 2 + 2

AMAT = 4 µs

Exercícios

9. **(ENADE 2005)** O grande desejo de todos os desenvolvedores de programas é utilizar quantidades ilimitadas de memória que, por sua vez, sejam extremamente rápidas. Infelizmente, isso não corresponde à realidade, como tenta representar a figura abaixo, que descreve uma hierarquia de memória: para cada elemento, estão indicados os tamanhos típicos disponíveis para armazenamento de informação e o tempo típico de acesso à informação armazenada. Como pode ser visto no diagrama abaixo, registradores do processador e memória cache operam com tempos distintos, o mesmo ocorrendo com a memória principal com relação à memória cache, e com a memória secundária com relação à memória principal



Considerando as informações acima apresentadas, responda às seguintes perguntas:

1. Que características um programa deve ter para que o uso de memória cache seja muito vantajoso?
2. Se registradores do processador e a memória cache operassem com os mesmos tempos de acesso, ainda haveria vantagem em se utilizar a memória cache? E se a memória cache e a memória principal operassem com os mesmos tempos de acesso, ainda haveria vantagem em se utilizar a memória cache? Justifique suas respostas

Resposta de Exercícios

9. (ENADE 2005) O grande ...

Considerando as informações acima apresentadas, responda às seguintes perguntas:

1. Que características um programa deve ter para que o uso de memória cache seja muito vantajoso?

   O programa deve ter trechos pequenos que sejam executados várias vezes, e os dados devem estar localizados próximos uns dos outros OU dados e instruções devem ter localidade espacial (próximos uns dos outros) e localidade temporal (serem usados várias vezes em um certo instante de tempo)

2. Se registradores do processador e a memória cache operassem com os mesmos tempos de acesso, ainda haveria vantagem em se utilizar a memória cache? E se a memória cache e a memória principal operassem com os mesmos tempos de acesso, ainda haveria vantagem em se utilizar a memória cache? Justifique suas respostas

   Se cache e processador operassem com os mesmos tempos, ainda assim seria vantajoso utilizar cache, porque o seu objetivo é justamente fornecer dados e instruções na velocidade do processador, simulando uma memória principal rápida. Se cache e memória operassem com os mesmos tempos, não haveria mais razão para se usar o cache, quer estivessem no cache ou na memória principal

Exercícios

10. (POSCOMP 2007) Um processador tem a seguinte hierarquia de memória: uma cache com latência de acesso de 1ns e uma memória principal com latência de acesso de 100ns. O acesso à memória principal somente é realizado após o valor não ser encontrado na cache. A MAIOR taxa de cache miss aceitável para que o tempo médio de acesso à memória seja menor ou igual à 2ns é

a) 10%
b) 5%
c) 50%
d) 1%
e) 2%

Resposta de Exercícios

10. (POSCOMP 2007) Um processador tem a seguinte hierarquia de memória: uma cache com latência de acesso de 1ns e uma memória principal com latência de acesso de 100ns. O acesso à memória principal somente é realizado após o valor não ser encontrado na cache. A MAIOR taxa de cache miss aceitável para que o tempo médio de acesso à memória seja menor ou igual à 2ns é

a) 10%
b) 5%
c) 50%
d) 1%
e) 2%

Exercícios

11. **(POSCOMP 2003 - 24)** A interposição de um circuito de memória cache entre o processador e a memória principal (RAM)

a) Aumenta o tráfego de instruções e/ou dados no barramento de memória

b) Aumenta o tráfego de instruções e/ou dados entre memória e disco

c) Diminui o tráfego de instruções e/ou dados no barramento de memória

d) Diminui o tráfego de instruções e/ou dados entre memória e disco

e) Permite acessos concorrentes à memória RAM

Resposta de Exercícios

11. **(POSCOMP 2003 - 24)** A interposição de um circuito de memória cache entre o processador e a memória principal (RAM)

a) Aumenta o tráfego de instruções e/ou dados no barramento de memória

b) Aumenta o tráfego de instruções e/ou dados entre memória e disco

c) Diminui o tráfego de instruções e/ou dados no barramento de memória

d) Diminui o tráfego de instruções e/ou dados entre memória e disco

e) Permite acessos concorrentes à memória RAM

Exercícios

12. **(POSCOMP 2011 - 42)** Ao medir o desempenho de um certo sistema, verificou-se que este passava muito tempo com a CPU ociosa e tinha um alto volume de acessos a disco. Assinale a alternativa que apresenta a solução traduzida na melhoria de desempenho desse sistema

a) Troca da CPU por uma mais rápida

b) Aumento na capacidade de memória do sistema

c) Aumento na capacidade de armazenamento do disco

d) Uso de memória cache

e) Troca do sistema operacional

Resposta de Exercícios

12. **(POSCOMP 2011 - 42)** Ao medir o desempenho de um certo sistema, verificou-se que este passava muito tempo com a CPU ociosa e tinha um alto volume de acessos a disco. Assinale a alternativa que apresenta a solução traduzida na melhoria de desempenho desse sistema

a) Troca da CPU por uma mais rápida

b) Aumento na capacidade de memória do sistema

c) Aumento na capacidade de armazenamento do disco

d) Uso de memória cache

e) Troca do sistema operacional

Exercícios

13. **(POSCOMP 2013, Questão 44)** A memória do computador é organizada em níveis. Assinale a alternativa que apresenta, corretamente, as estruturas encontradas no nível mais alto dessa hierarquia.

a) Cache L1

b) Cache L2

c) Disco rígido

d) Memória DRAM

e) Registradores do processador

Resposta de Exercícios

13. **(POSCOMP 2013, Questão 44)** A memória do computador é organizada em níveis. Assinale a alternativa que apresenta, corretamente, as estruturas encontradas no nível mais alto dessa hierarquia.

a) Cache L1

b) Cache L2

c) Disco rígido

d) Memória DRAM

e) Registradores do processador