Universidade Federal do Rio de Janeiro
Pós-Graduação em Informática IM-NCE/UFRJ

Microarquiteturas de Alto Desempenho

Pipeline

Gabriel P. Silva

Introdução

- **Pipeline é uma técnica de implementação de processadores que permite a sobreposição temporal das diversas fases de execução das instruções.**

- **Aumenta o número de instruções executadas simultaneamente e a taxa de instruções iniciadas e terminadas por unidade de tempo.**

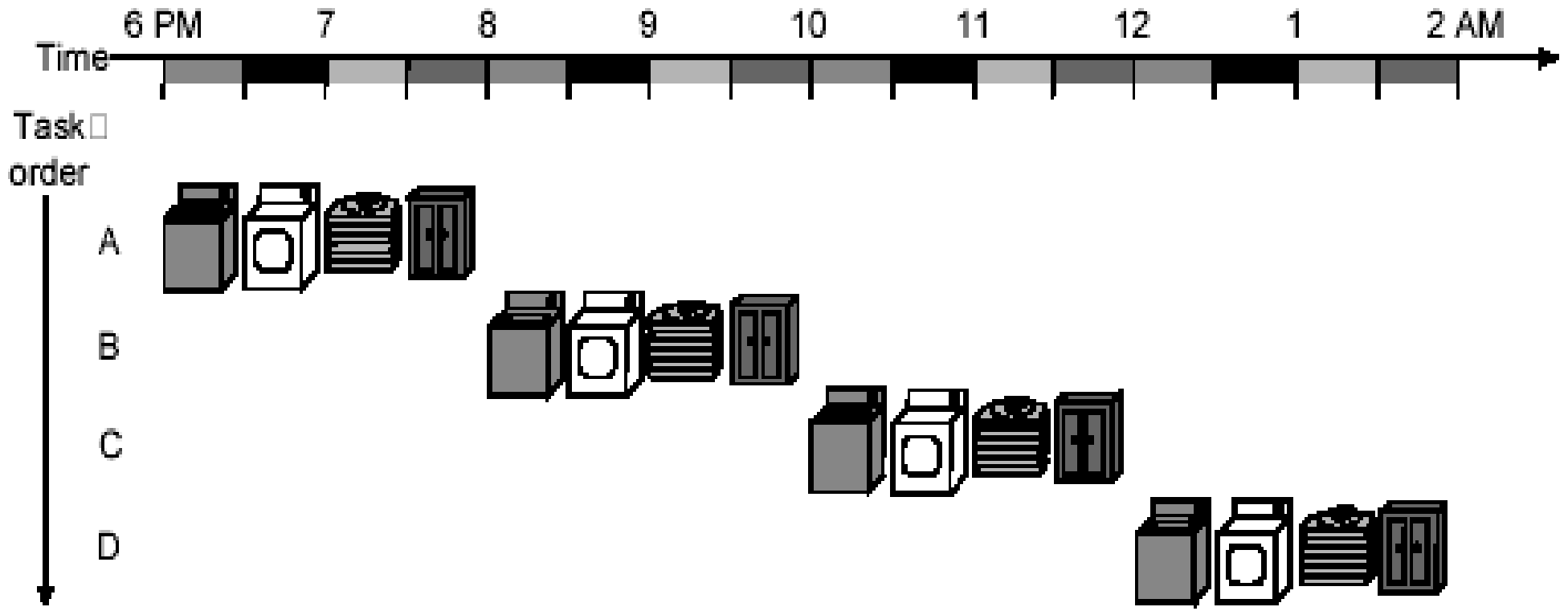
- **O pipeline não reduz o tempo gasto para completar cada instrução individualmente.**

Exemplo

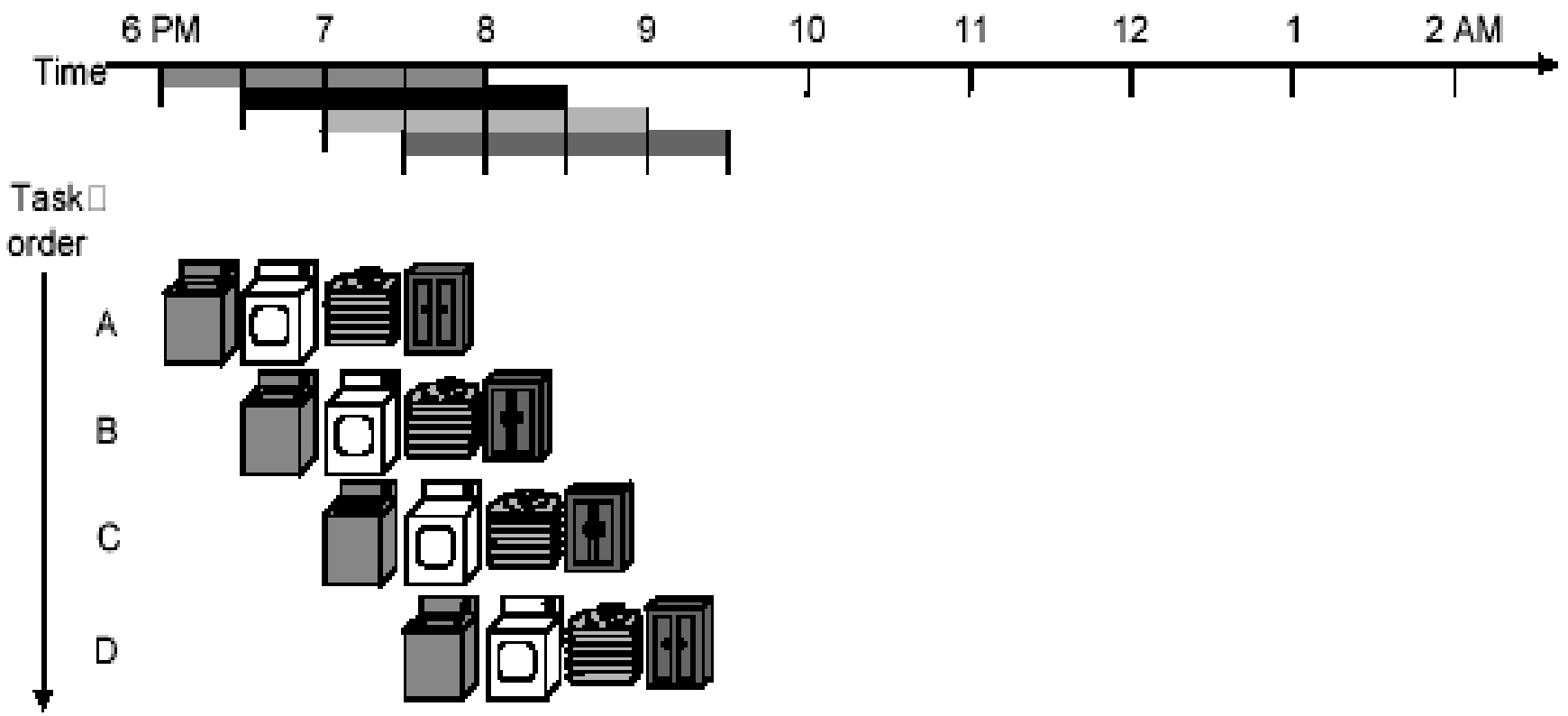
Vamos supor uma lavanderia, em que cada etapa possa ser realizada em 30 minutos:

2. **Colocar a roupa na máquina de lavar**
3. **Depois de lavada, colocá-la na máquina de secar roupa**
4. **Depois de seca, passar a ferro**
5. **Depois de passada, arrumá-la no armário**

Exemplo sem Pipeline



Exemplo com Pipeline



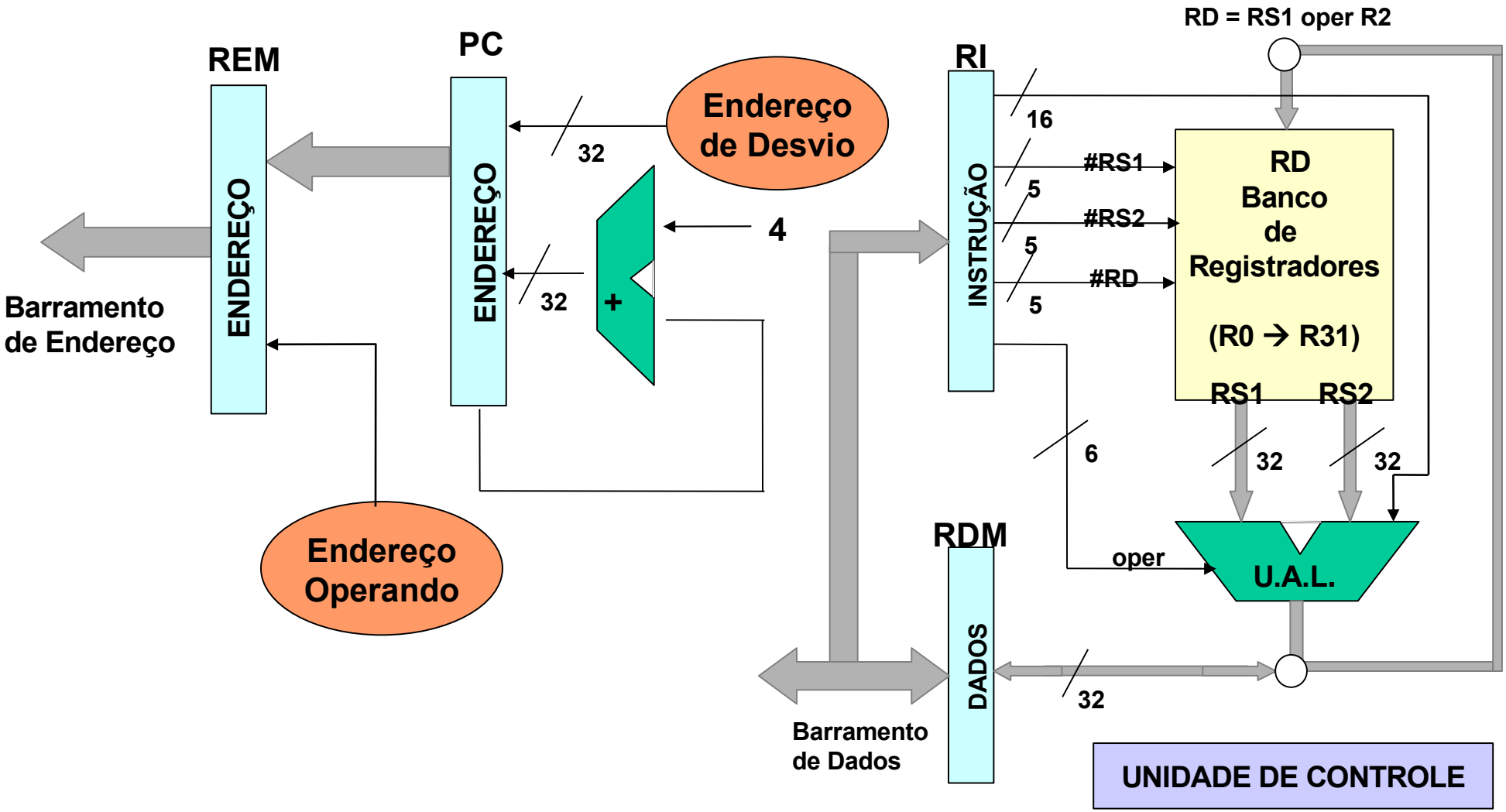
Exemplo

- **Supondo-se que cada uma destas etapas leve 30 minutos para ser realizada, a lavagem de um cesto de roupas continuará levando 2 horas para ser realizada.**
- **Entretanto, podemos iniciar a lavagem de um cesto de roupas a cada 30 minutos, até que tenhamos 4 cestos sendo lavados simultaneamente, um em cada etapa do "pipeline".**
- **Depois das primeiras 2 horas, teremos um cesto de roupa lavada a cada 30 minutos. Ao final do dia teremos lavado muito mais cestos de roupa do que sem o uso de pipeline**

Pipeline

- Não melhora a latência de cada tarefa individualmente
- Melhora o *throughput* de todo o trabalho
- Várias tarefas executam simultaneamente usando recursos diferentes
- *Speedup* potencial = número de estágios do pipeline

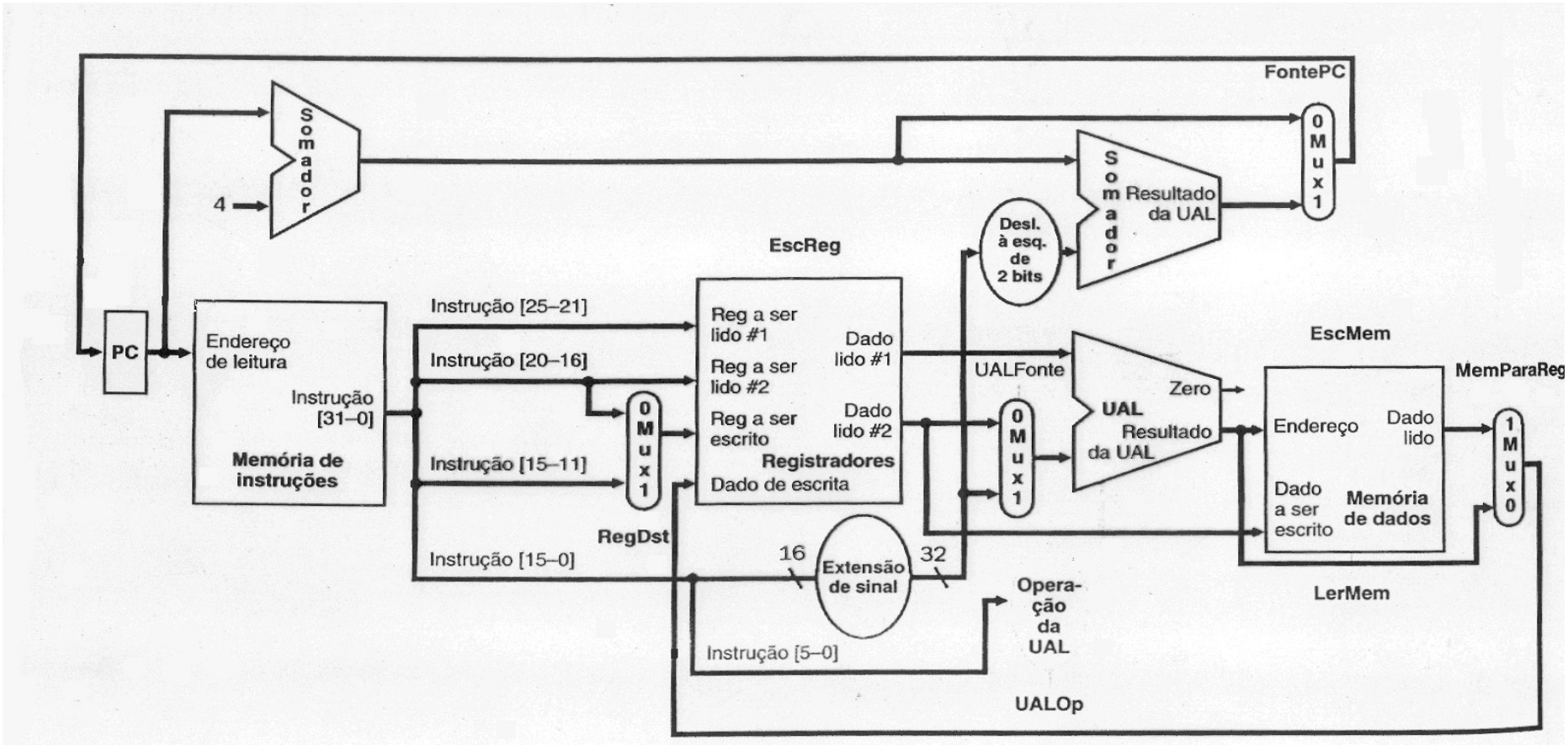
Arquitetura Básica



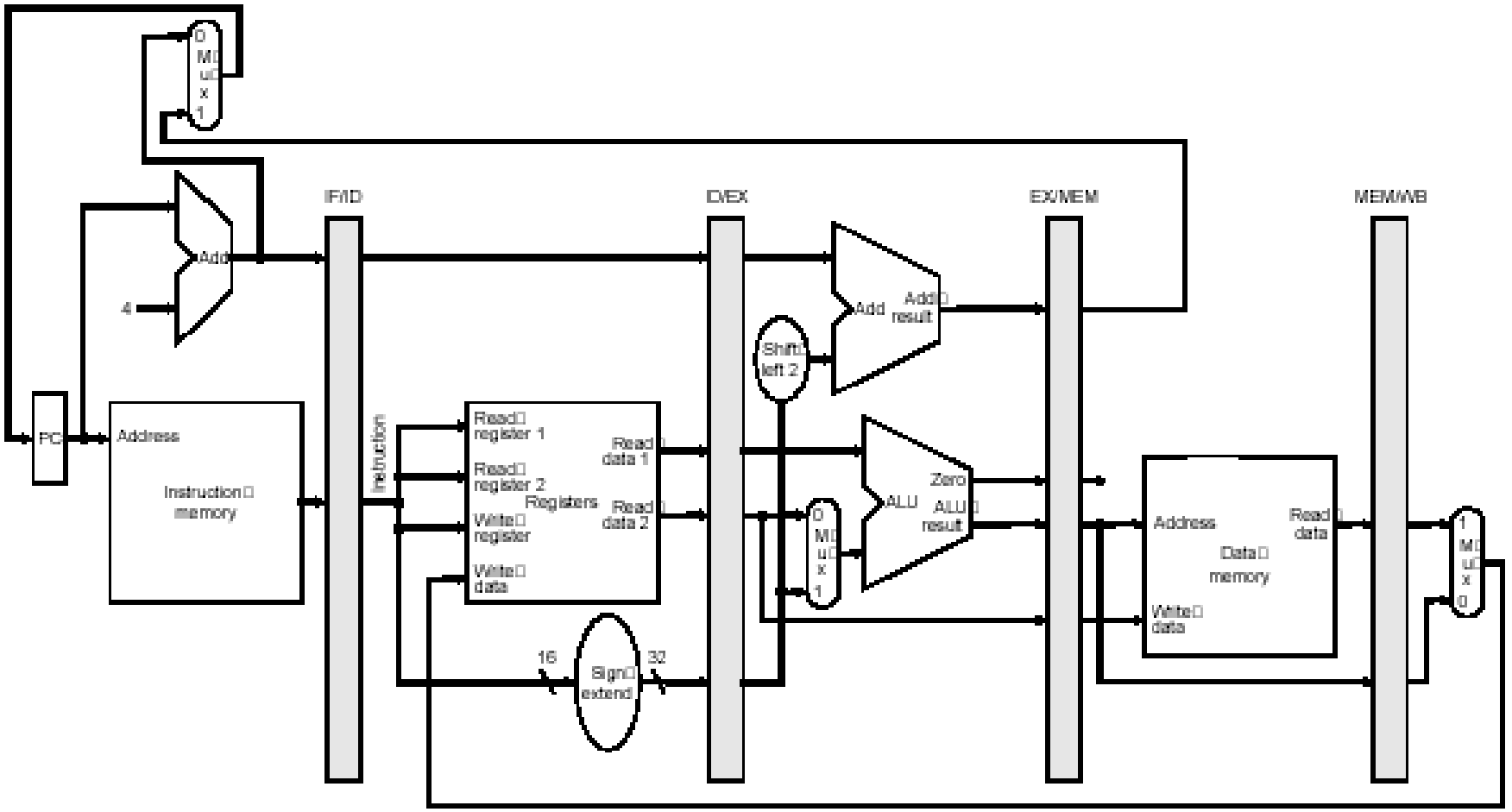
Exemplo de Pipeline de Instruções

- **Divisão da Execução da Instrução em 5 estágios:**
  - **Busca da Instrução na Memória (B)**
  - **Leitura dos Registradores e Decodificação da Instrução (D)**
  - **Execução da Instrução / Cálculo do Endereço (E)**
    Acesso a um Operando na Memória (M)
  - **Escrita de um Resultado em um Registrador (W)**

Arquitetura sem Pipeline



Arquitetura com Pipeline

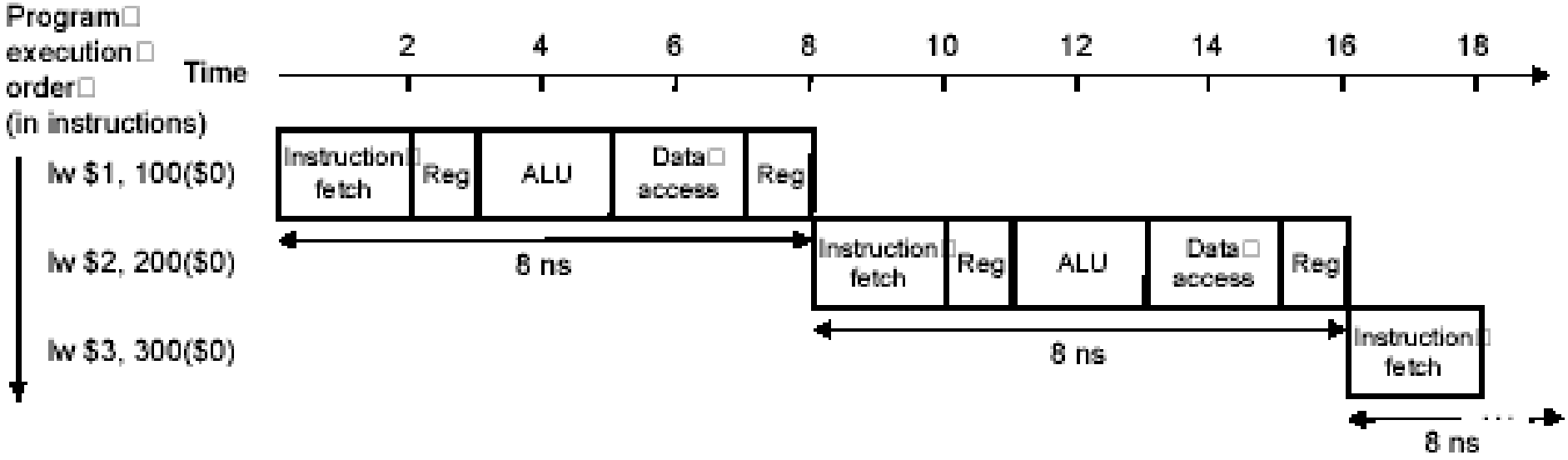


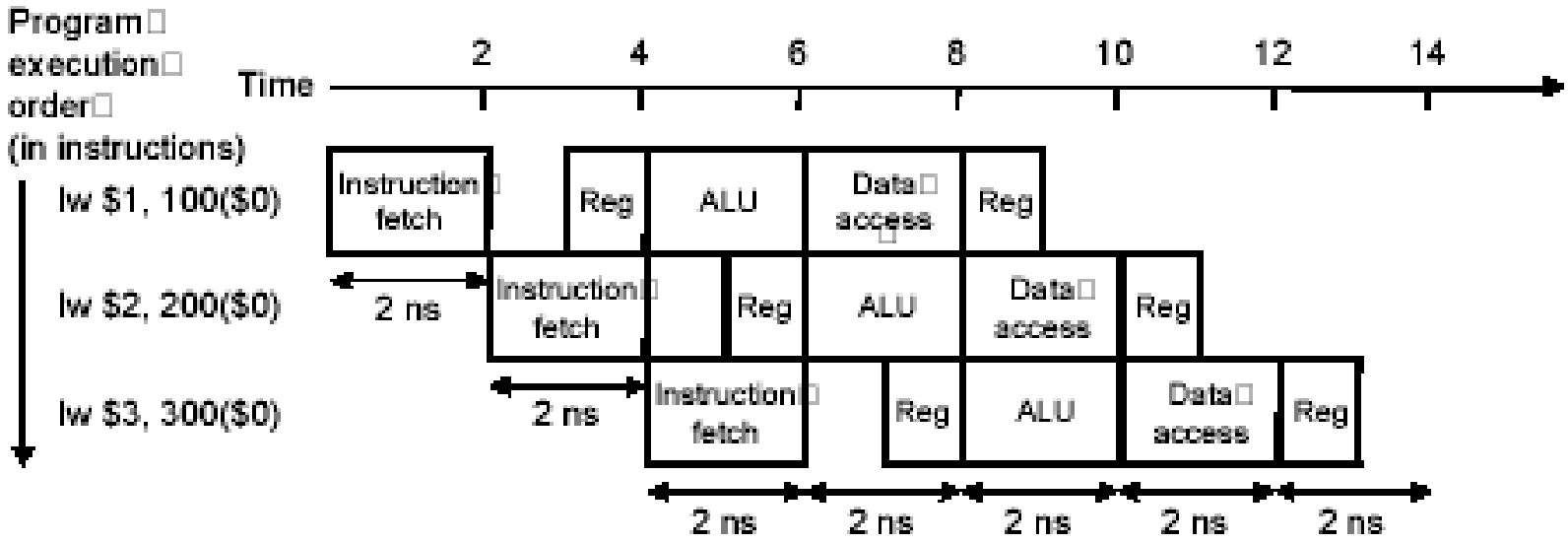
Exemplo de Pipeline de Instruções

Classe da Instrução	Busca da Instrução	Leitura Operando	Operação da ULA	Acesso à Memória	Escrita do Resultado	Total
Load Word (lw)	2 ns	1 ns	2 ns	2 ns	1 ns	8 ns
Store Word (sw)	2 ns	1 ns	2 ns	2 ns		7 ns
Aritméti-cas (add, sub, and)	2 ns	1 ns	2 ns		1ns	6 ns
Branch (beq)	2 ns	1 ns	2 ns			5 ns

Exemplo de Pipeline de Instruções

Classe da Instrução	Busca da Instrução	Leitura Operando	Operação da ULA	Acesso à Memória	Escrita do Resultado	Total
Load Word (lw)	2 ns	1 ns	2 ns	2 ns	1 ns	10 ns
Store Word (sw)	2 ns	1 ns	2 ns	2 ns		10 ns
Aritméti-cas (add, sub, and)	2 ns	1 ns	2 ns		1ns	10 ns
Branch (beq)	2 ns	1 ns	2 ns			10 ns


Program execution order (in instructions)
Time
2
4
6
8
10
12
14
16
18
lw $1, 100($0)
lw $2, 200($0)
lw $3, 300($0)
Instruction fetch
Reg
ALU
Data access
Reg
8 ns
8 ns
8 ns


Program execution order (in instructions)
Time
2
4
6
8
10
12
14
lw $1, 100($0)
lw $2, 200($0)
lw $3, 300($0)
Instruction fetch
Reg
ALU
Data access
Reg
2 ns
2 ns
2 ns
2 ns
2 ns
2 ns
2 ns

Características dos Pipelines de Instrução

- **O tempo do ciclo do relógio do processador deve ser igual ou maior que o tempo de execução do estágio mais lento do "pipeline".**
- **Deve-se procurar dividir a execução da instrução em estágios com o mesmo tempo.**
- **O pipeline deve ser mantido sempre "cheio" para que o desempenho máximo seja alcançado.**
- **De um modo geral, com o uso do pipeline, cada instrução ainda leva o mesmo tempo para ser executada.**
- **Algumas instruções contudo podem ter o seu tempo de execução aumentado, pois atravessam estágios em que não realizam nenhuma operação útil.**

Características dos Pipelines de Instrução

- **O tempo gasto no processamento de M instruções em um pipeline com K estágios e ciclo de máquina igual a t é dado por:**

$$T = [K + (M - 1)] * t$$

- **Se M >> K (caso comum), T é aproximadamente M * t**

Características dos Pipelines de Instrução

- **Um programa tem 1.000.000 de instruções. Em uma arquitetura sem pipeline, o tempo médio de execução de cada instrução é 6,5 ns. Qual o ganho na execução deste programa em um processador com pipeline de 5 estágios com ciclo de 2 ns?**

T_1 = 6,5 ns * 1.000.000 =~ 6,5 ms
(sem pipeline)

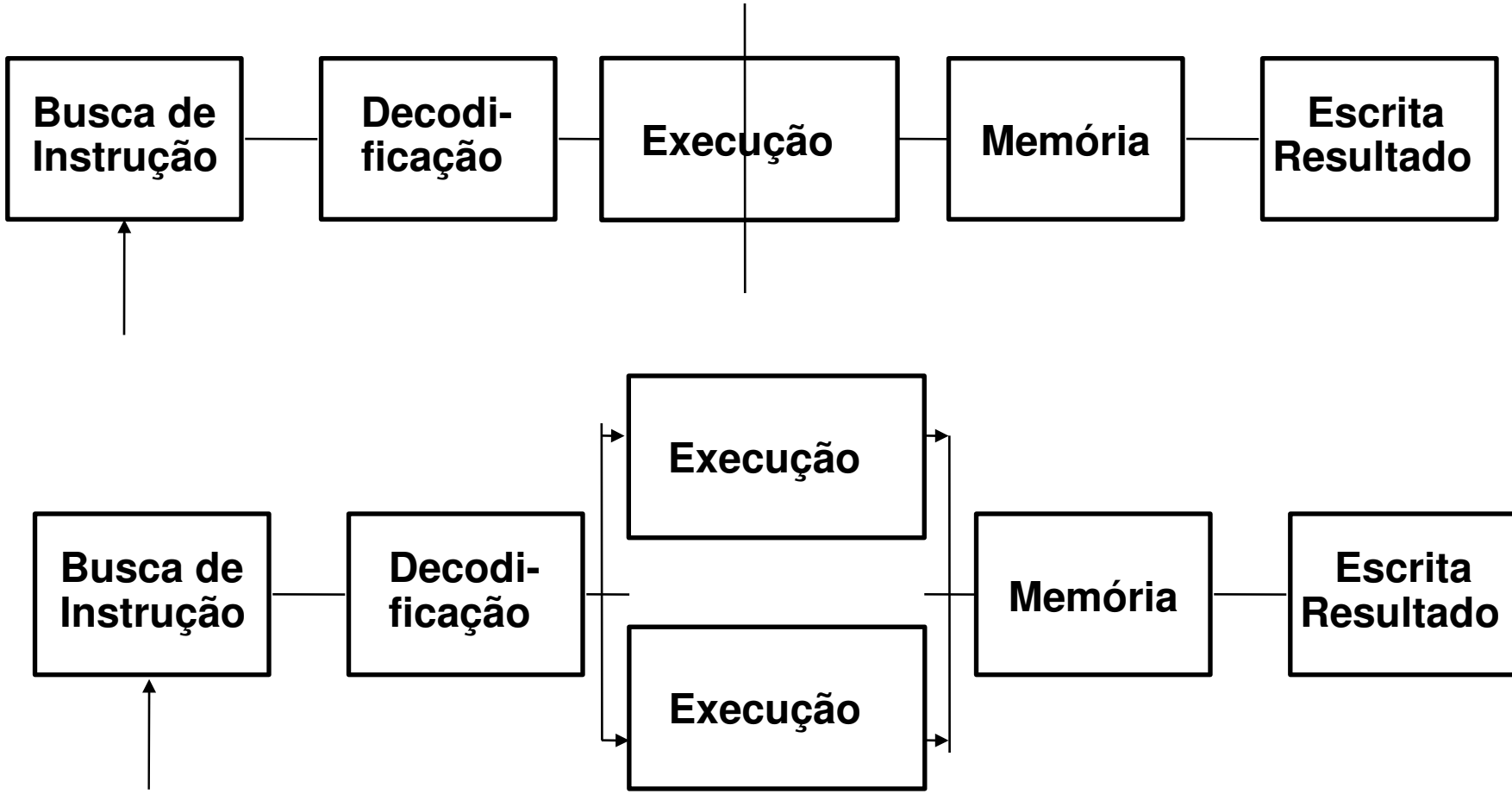
T_2= (5 + 999.999)* 2 ns =~ 2 ms
(com pipeline)

Ganho = 6,5 ms / 2 ms = 3,25

Problemas no Uso de Pipeline

- **Estágios podem ter tempos de execução diferentes:**
  - **Solução 1: Implementar esses estágios como um pipeline onde cada sub-estágio possua tempo de execução semelhante aos demais estágios do pipeline principal**
  - **Solução 2: Replicar esse estágio, colocando réplicas em paralelo no estágio principal. O número de réplicas é dado pela razão entre o tempo do estágio mais lento e os demais.**
- **O sistema de memória é incapaz de manter o fluxo de instruções no pipeline**
  - **O uso de memória cache com alta taxa de acerto e tempo de acesso compatível com o tempo de ciclo do pipeline**

Problema no Uso de Pipeline



Problemas no Uso de Pipelines

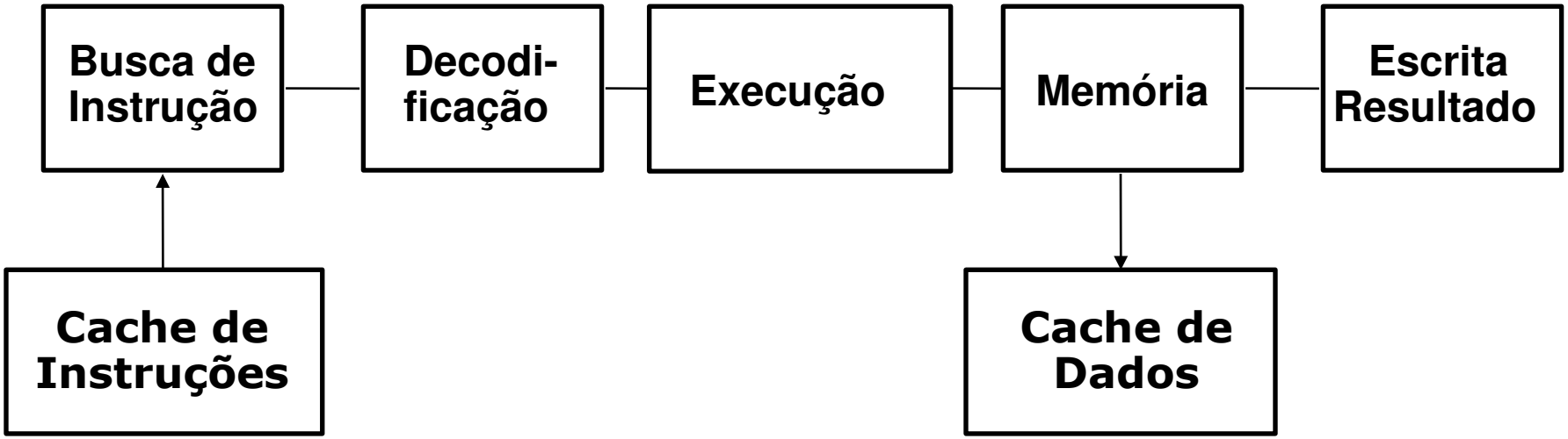
- **Dependências ou Conflitos ("Hazards")**
  - **Conflitos Estruturais**
    - **Pode haver acessos simultâneos ao mesmo recurso feitos por 2 ou mais estágios**
  - **Dependências de Dados**
    - **As instruções dependem de resultados de instruções anteriores, ainda não completadas**
  - **Dependências de Controle**
    - **A próxima instrução não está no endereço subseqüente ao da instrução anterior**
- **Tratamento de Exceções**

Conflitos Estruturais

- **Busca da instrução e leitura/escrita de dados simultâneamente à memória:**
  - **Uso de arquitetura "Harvard" com caches de dados e instrução separados**
- **Acesso simultâneo ao banco de registradores**
  - **Uso de banco de registradores com múltiplas portas**
- **Uso simultâneo de uma mesma unidade funcional**
  - **Replicação da unidade funcional ou implementação "pipelined" dessa unidade**

Conflitos Estruturais

- **Problema: Acessos simultâneos à memória por 2 ou mais estágios**



- **Soluções**
  - **Caches separadas de dados e instruções**

Conflitos por Dados

- **Problema: instruções consecutivas podem fazer acesso aos mesmos operandos:**
  - **execução da instrução seguinte depende de operando calculado pela instrução anterior:**

```
dadd R1, R2, R3
dsub R4, R1, R6
```

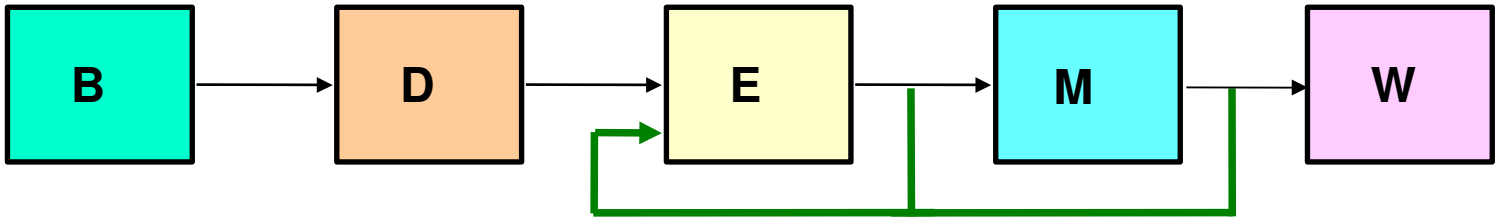
- **Tipos de dependências de dados**
  - **Dependência verdadeiras**
  - **Dependências falsas**
    - **antidependência**
    - **dependência de saída**

Dependências de Dados

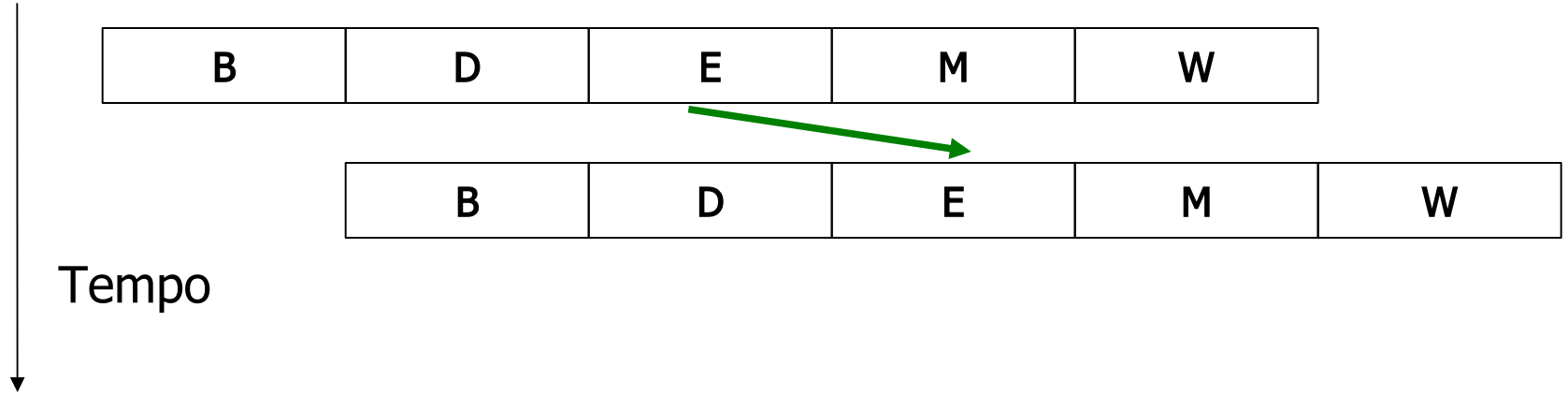
- **Dependências Verdadeiras (Direta):**
  - **O pipeline precisa ser parado durante certo número de ciclos ("interlock");**
  - **Colocação de instruções de "nop" ou escalonamento adequado das instruções pelo compilador;**
  - **Colocar caminhos de dados alternativos entre os estágios de pipeline: adiantamento dos dados. Resolve a maioria dos casos.**
- **Dependências Falsas:**
  - **Não é um problema em pipelines onde a ordem de execução das instruções é mantida;**
  - **Problema em processadores superescalares;**
  - **A renomeação dos registradores é uma solução usual para este problema.**

Adiantamento dos Dados

- **Caminho interno dentro do pipeline entre a saída de um estágio e a entrada da ALU**
  - **Evita a *parada* do pipeline**



Adiantamento do Resultado



Escalonamento das Instruções

- **Exemplo:**

  X = Y - W

  Z = K + L

- **Código Gerado:**

```
ld    r1, mem[Y]
ld    r2, mem[W]
dsub  r3, r1, r2      ➔ Situação de Interlock
sd    r3, mem[X]  ➔ Situação de Interlock
ld    r4, mem[K]
ld    r5, mem[L]
dadd  r6, r4, r5  ➔ Situação de Interlock
sd    r6, mem[Z]  ➔ Situação de Interlock
```

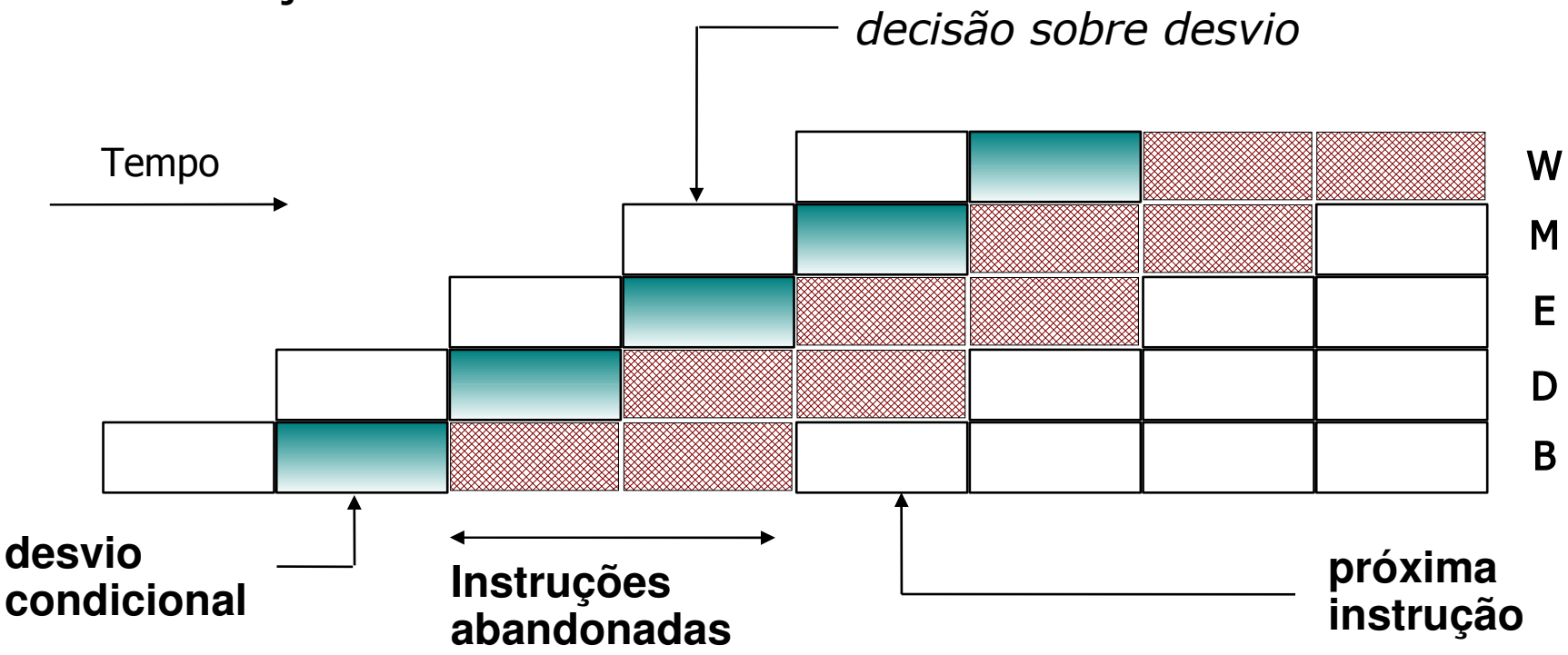
Escalonamento das Instruções

- **Código Otimizado:**

```
ld    r1, mem[Y]
ld    r2, mem[W]
ld    r4, mem[K]
dsub    r3, r1, r2
ld    r5, mem[L]
sd    r3, mem[X] ➔ Adiantamento
dadd    r6, r4, r5
sd    r6, mem[Z] ➔ Adiantamento
```

Conflitos de Controle

- **Efeito de desvios condicionais**
  - **Se o desvio ocorre, pipeline precisa ser esvaziado**
  - **Não se sabe se desvio ocorrerá ou não até o momento de sua execução**



Soluções para os Conflitos de Controle

- **Uso do Desvio Atrasado**
  - **A instrução após o desvio é sempre executada → preenchimento útil do "delay slot" nem sempre é possível**
- **Congelar o pipeline até que o resultado do desvio seja conhecido**
  - **Insere "bolhas" no pipeline → solução ruim quando o pipeline é muito longo**
- **Predição Estática de Desvios**
  - **O compilador faz uma predição se o desvio vai ser tomado ou não → geração de "bolhas" quando a predição é errada, baixa taxa de acertos**
- **Predição Dinâmica de Desvios**
  - **Existem mecanismos em "hardware" que fazem a predição baseada no comportamento daquele desvio no passado → idem, alta taxa de acertos**

Preenchimento do "delay slot"

- **Exemplo 1:**

dadd r1, r2, r3	beq r2, r0, label
beq r2, r0, label	dadd r1, r2, r3
delay slot	
3 ciclos	2 ciclos

- **Exemplo 2:**

dsub r4, r5, r6	dadd r1, r2, r3
dadd r1, r2, r3	beq r1, r0, label
beq r1, r0, label	dsub r4, r5, r6
delay slot	
4 ciclos	3 ciclos

Preenchimento do "delay slot"

- **Para facilitar o trabalho do compilador no preenchimento do "delay slot" muitas arquiteturas permitem o uso do "delay slot" com a opção de anulação automática dessa instrução se o desvio condicional não for tomado.**
- **Desse modo, uma instrução do endereço alvo pode ser movida para o "delay slot", o que é muito útil no caso de "loops". Nesse caso, está implícita uma previsão de desvio estática que diz que o desvio será sempre tomado.**

Predição Estática do Desvio

- **Três abordagens podem ser adotadas:**
  - **Assumir que todos os desvios são tomados ("branch taken");**
  - **Os desvios para trás são assumidos como tomados ("branch taken") e os desvios para frente são assumidos como não tomados ("branch not taken") ;**
  - **Fazer a predição com base em resultados coletados de experiências de "profile" realizadas anteriormente.**

Predição Dinâmica do Desvio

- **Pequena memória endereçada pela parte baixa do endereço das instruções de desvio;**
- **A memória contém 1 bit (bit de predição) que diz se o desvio foi tomado ou não da última vez;**
- **Se a predição for errada, o bit correspondente é invertido na memória**
- **Problemas:**
  - **Instruções de desvio diferentes podem mapear para uma mesma posição do buffer**
  - **O esquema pode falhar quando a decisão do desvio se alterna a cada execução**

Tratamento de Exceções

- **Exemplos de Exceções:**
  1. **Interrupção de dispositivos de E/S**
  2. **Chamadas ao Sistema Operacional**
  3. **Breakpoints**
  4. **Operações Aritméticas (Overflow e Underflow)**
  5. **Falha de página**
  6. **Erros de endereçamento de memória**
  7. **Violação de proteção de memória**
  8. **Instrução inválida**
  9. **Falha de alimentação**

Classificação das Exceções

Síncronas	2, 3, 4, 5, 6, 7, 8
Assíncronas	1, 9

Solicitadas pelo usuário	2, 3
Fora do controle do usuário	1, 4, 5, 6, 7, 8, 9

No meio da instrução	4, 5, 6, 7, 8, 9
Entre instruções	1, 2, 3

Encerram a execução do programa	6, 7, 8, 9
Permitem a continuação do programa	1, 2, 3, 4, 5

Modelo de Exceções Precisas

- **Definição:**
  - **Ao ser detectada uma exceção, todas as instruções anteriores à ocorrência da exceção podem ser completadas e as posteriores serão anuladas e reiniciadas após o tratamento da exceção**
- **Requisito Básico:**
  - **Instruções só mudam o estado da máquina quando há garantia de que elas concluirão sem gerar exceções.**
- **Conseqüências:**
  - **Difícil de implementar, sem perda de desempenho, em pipelines que implementam instruções complexas**
  - **Algumas arquiteturas possuem então dois modos de operação: com ou sem exceções precisas**