


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



# Arquiteturas de Sistemas de Processamento Paralelo



## Redes de Interconexão



Gabriel P. Silva



# **Redes de Interconexão Estáticas**

# Redes de Interconexão Estáticas

## ◆ Topologias

- Unidimensionais e Bidimensionais
- Tridimensionais e Hipercúbicas

## ◆ Técnicas de Chaveamento

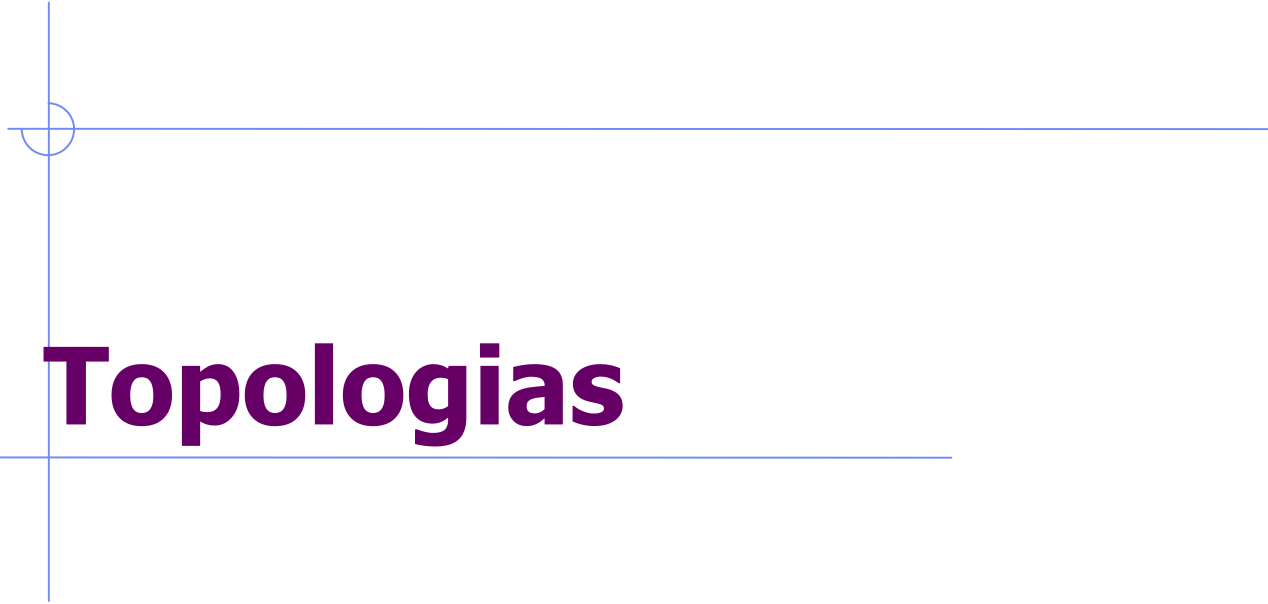
- Chaveamento por pacote (store-and-forward)
- Chavamento por circuito
- Virtual Cut-Through
- Wormhole

## ◆ Algoritmos de Roteamento

- Determinístico
- Adaptativo

# Redes de Interconexão Estáticas

- ◆ Normalmente utilizadas em arquiteturas paralelas por troca de mensagens (multicomputadores).
- ◆ Redes de interconexão estáticas são redes com topologia baseada em grafos, onde cada nó é um elemento processador e cada aresta do grafo representa um “link” entre dois elementos processadores.



# Topologias



# Topologias de Redes de Interconexão Estáticas

## ◆ Linha

- Cada processador está conectado aos seus vizinhos da esquerda e da direita.
- A mensagem é repetidamente passada para o próximo nó até chegar ao seu destino.

## ◆ Bi-dimensional

### ■ Anel

- ◆ Quando o primeiro e últimos nós da topologia em linha estão interconectados

### ■ Estrela

- ◆ Um nó atua como nó de controle ao qual todos demais nós estão conectados
- ◆ Por manipular toda a comunicação entre os nós, o nó central é o gargalo do sistema.

# Topologias de Redes de Interconexão Estáticas

## ◆ Bi-dimensional

### ■ Árvore

- ◆ Uma árvore binária de profundidade  $d$  tem  $2^d - 1$  nós.
- ◆ As redes em árvore sofrem de um gargalo de comunicação nos níveis mais altos da árvore binária.
- ◆ Este problema pode ser resolvido aumentando a capacidade de comunicação dos “links” que estão mais perto da raiz. Este rede é chamada de “Árvore Gorda”.

# Topologias de Redes de Interconexão Estáticas

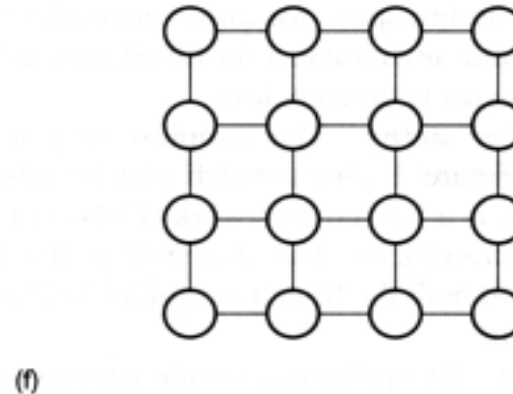
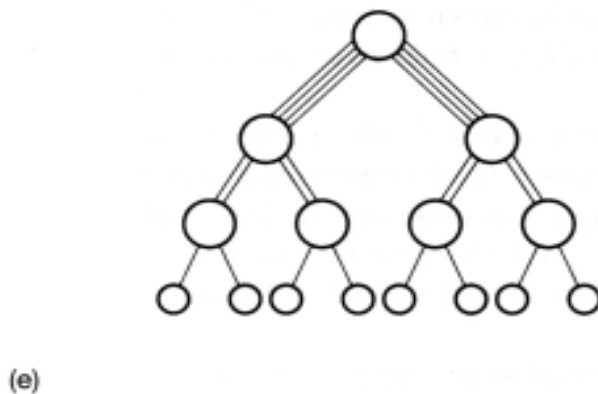
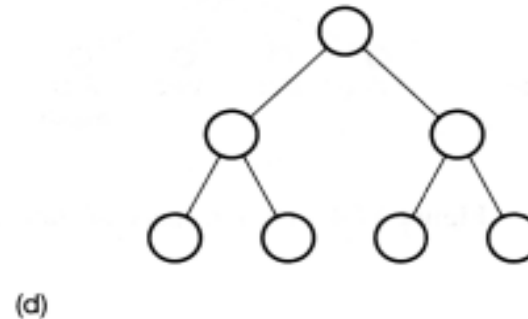
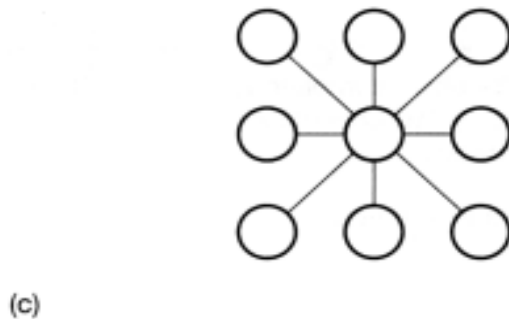
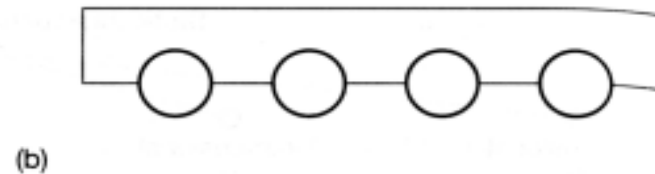
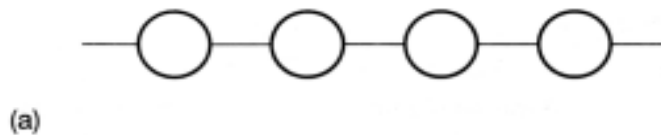
## ◆ Bi-dimensional

### ■ Malha Bi-dimensional

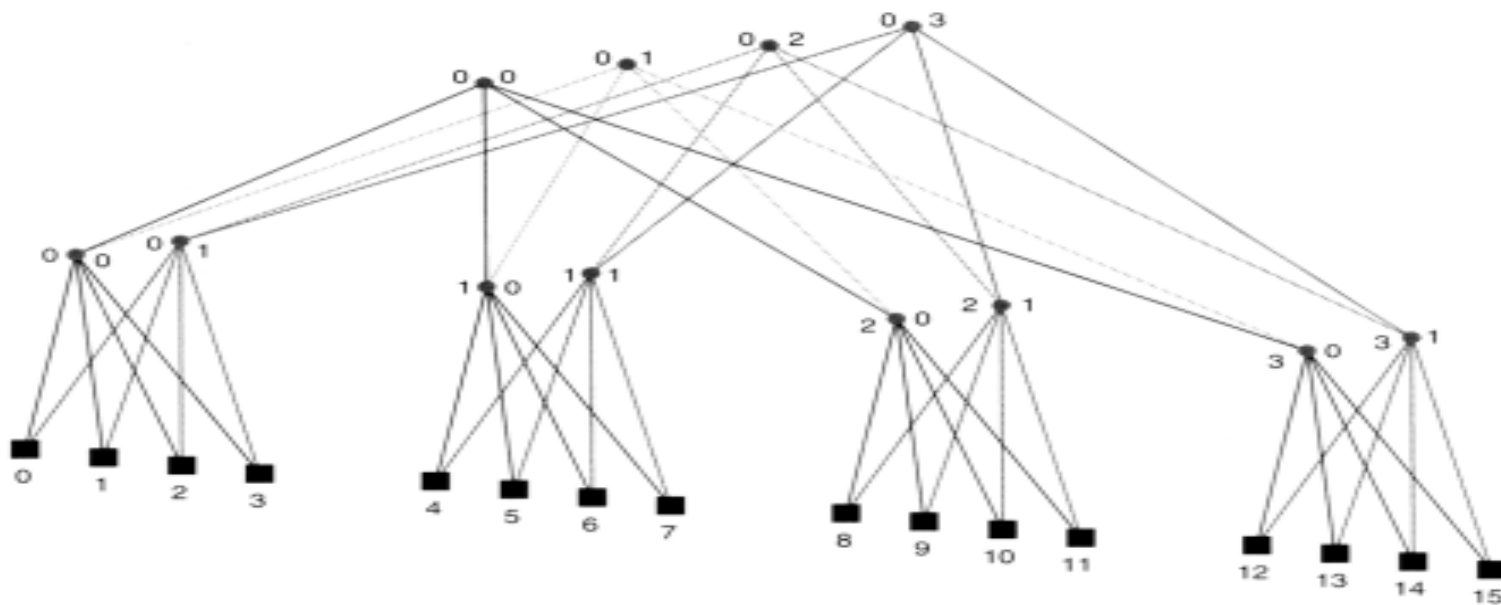
- ◆ Cada processador tem quatro vizinhos aos quais está conectado por um “link”.
- ◆ A malha bidimensional é uma extensão do vetor linear.
- ◆ Se as duas dimensões da malha não forem iguais, temos uma malha retangular.



# Topologias de Redes de Interconexão Estáticas



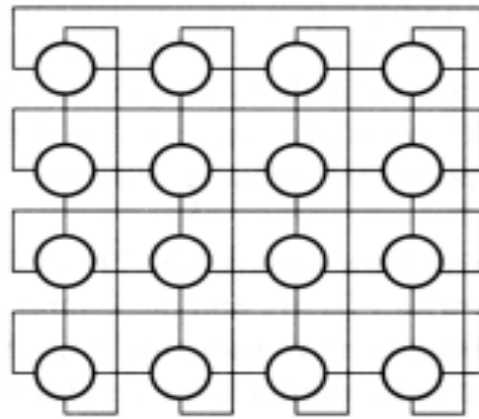
# Topologias de Redes de Interconexão Estáticas



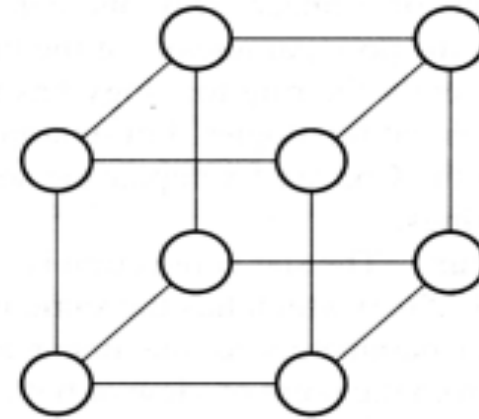
# Interconexão Estática Tri-dimensional

- ◆ **Toro (Malha Conectada nas 2 dimensões)**
- ◆ **Cubo 3-D**
- ◆ **3-cube-connected cycle**
- ◆ **Totalmente conectada**
  - Todos os nós estão conectados entre si por um “link” direto.
  - O número de arestas do grafo totalmente conectado é dado por:
$$d = n(n-1)/2$$
  - Este tipo de rede é muito pouco utilizado devido aos altos custos de comunicação.

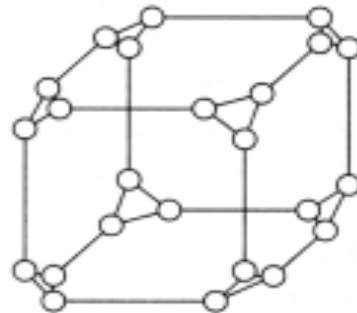
# Interconexão Estática Tri-dimensional



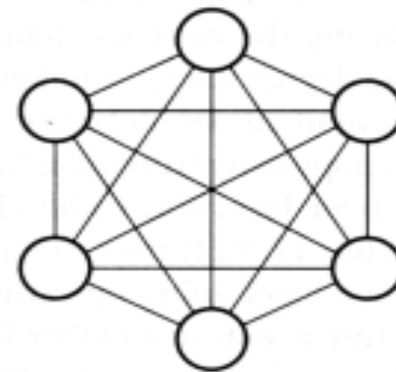
(g)



(h)



(i)



(j)

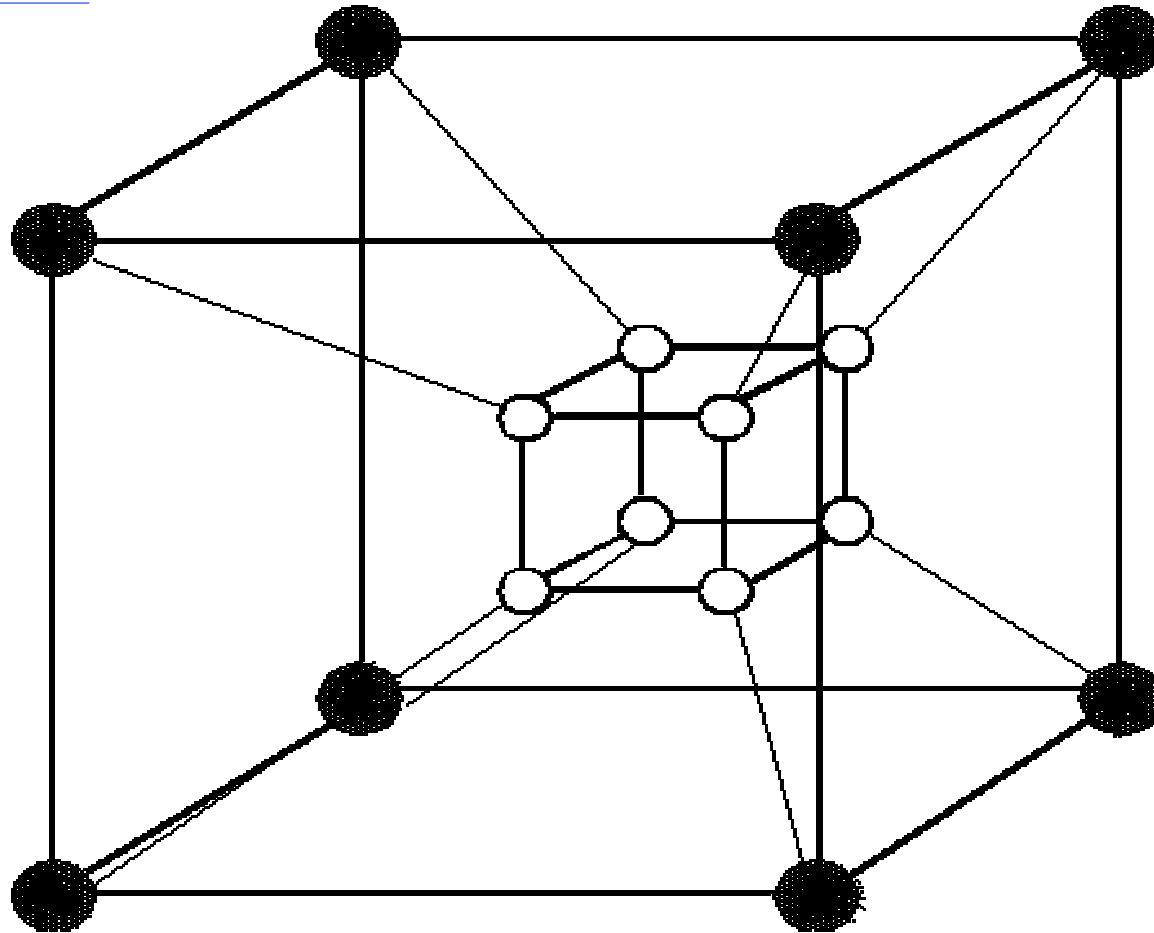
# Redes de Interconexão Hipercúbica

- ◆ Um hipercubo é uma malha multidimensional de nós processadores com exatamente dois nós em cada dimensão.
- ◆ Um hipercubo com dimensão  $d$  possui um total de  $n = 2^d$  processadores:
  - $d = 0$  é um hipercubo de dimensão zero, com apenas um nó;
  - $d = 1 \rightarrow n = 2$ , um hipercubo com 2 nós conectados por um “link”;
  - Um hipercubo de dimensão  $d+1$  consiste de dois hipercubos de dimensão  $d$ .

# Propriedades da Rede Hiperbólica

1. Dois nós estão conectados por um “link” se sua numeração binária difere apenas de **um** “bit”.
2. Cada nó está diretamente conectado a outros **d** processadores .
3. O número total de posições de bits diferentes entre dois nós é chamada de Distância de Hamming, **HD**, entre eles. Esta distância é caminho mais curto para uma mensagem trafegar entre esses dois nós. Por exemplo, a HD entre o nó 3(011) e o nó 5(101) é 2.

# Rede Hipercúbica



<http://www.esm-metz.fr/metz/personnel/vialle/noe/NOE-HyperCube-3/java/hypercube.html>

# Propriedades das Redes de Interconexão Estáticas

- 1) **Grau do nó:** Número de canais que incidem em um nó da rede.
- 2) **Diâmetro:** é a distância máxima entre quaisquer dois nós da rede.
- 3) **Conectividade:** é a medida da multiplicidade de caminhos entre dois nós quaisquer.
- 4) **Largura da Biseção:** é definida como o número mínimo de “links” de comunicação que necessitam ser removidos para particionar a rede em duas metades iguais.



# Propriedades das Redes de Interconexão Estáticas

- 5) **Largura de Banda da Biseção:** é definido como o volume mínimo de comunicação entre duas metades da rede com igual número de nós.
- 6) **Largura do Canal:** número de bits de cada “link” físico de comunicação.
- 7) **Taxa do Canal:** taxa de pico de transmissão dos bits através de cada “link” físico.
- 8) **Custo:** O custo de uma rede pode ser avaliado pela contagem do número de “links” requeridos por toda a rede de interconexão.

# Parâmetros de Desempenho

- **Banda Passante:**
  - Taxa máxima com que a rede é capaz de transmitir a informação.
- **Latência:**
  - Intervalo de tempo gasto por uma mensagem para atravessar a rede;
  - $\text{Latência} = \text{overhead} + (\text{tam. mensagem} / \text{banda passante})$ .
- **Escalabilidade**

# Propriedades das Redes de Interconexão Estáticas

Topologia	Grau Nó	Diâmetro	Largura Bisseção	Custo (Links)
Linha	1 ou 2	$N-1$	1	$N-1$
Anel	2	$N/2$	2	$N$
Estrela	1 ou $N-1$	2	1	$N-1$
Árvore Binária	1, 2 ou 3	$2 \log_2^* ((N+1)/2)$	1	$N-1$
Malha 2-D	2, 3, ou 4	$2(N^{1/2}-1)$	$N^{1/2}$	$2(N-N^{1/2})$

# Propriedades das Redes de Interconexão Estáticas

Topologia	Grau Nó	Diâmetro	Largura Bisseção	Custo (Links)
<b>Toro</b>	4	$(N^{1/2} - 1)$	$2 * (N^{1/2})$	$2 * N$
<b>Cubo 3-D</b>	3, 4, 5 ou 6	$3(N^{1/3} - 1)$	$N^{2/3}$	$2(N - N^{2/3})$
<b>Hipercubo</b>	$\log_2 N$	$\log_2 N$	$N/2$	$(N \log_2 N) / 2$
<b>Comp. Conectada</b>	$N - 1$	1	$(N^2) / 4$	$N(N-1) / 2$

# Exemplos de Arquiteturas com Redes Estáticas

- ◆ Cray T3-E: Torus 3-D, 600 Mbytes/ s por link.
- ◆ Cray T3-D: Torus 3-D, 300 Mbytes/ s por link.
- ◆ Intel iPSC-2: Hipercubo, 2.8 Mbytes/ s por link.
- ◆ Chaos Router: Torus 2-D, 360 Mbytes/ s por link.
- ◆ MIT M-Machine: Malha 3-D, 800 Mbytes/ s por link.

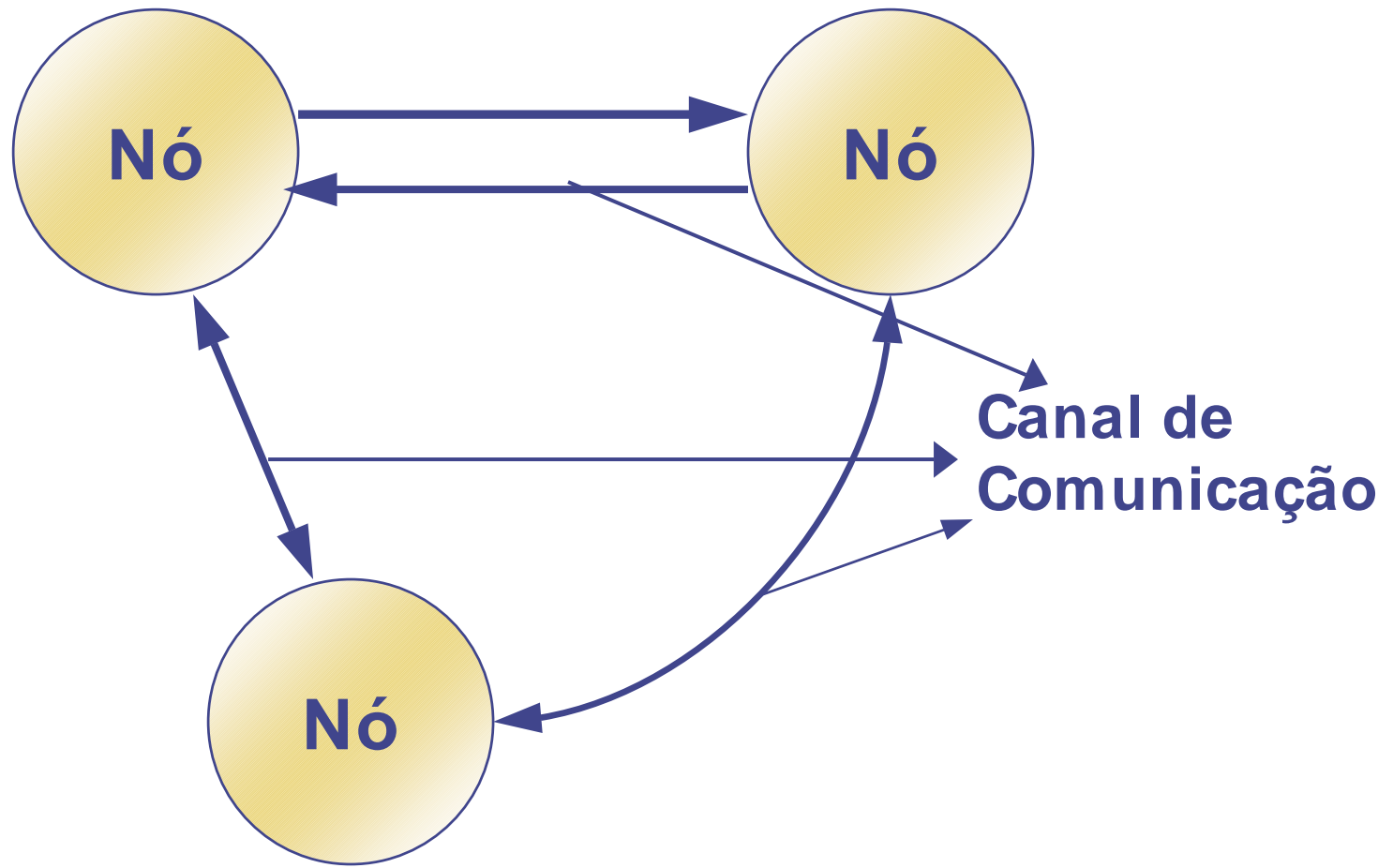


# **Técnicas de Roteamento**

# Elementos de uma Rede de Interconexão Estática

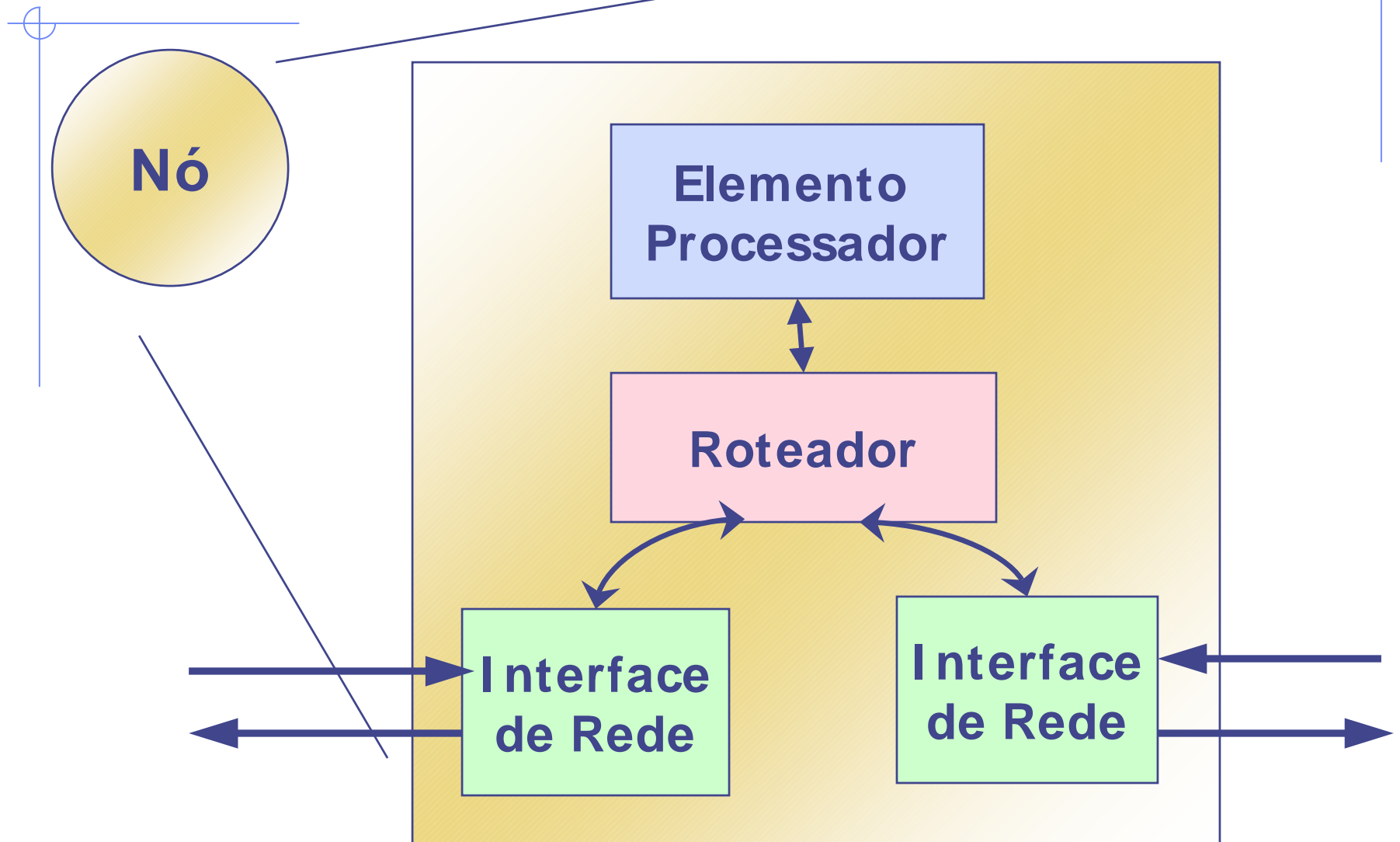
- ◆ **Nós:** são os elementos ativos da rede, que realizam a computação e o roteamento das mensagens. São compostos por:
  - Interface de Rede
  - Roteador
  - Elemento Processador
- ◆ **Canais:** conexões ponto-a-ponto por onde trafegam as mensagens.

# Elementos de uma Rede de Interconexão Estática





# Elementos de uma Rede de Interconexão Estática



# Transmissão da Mensagem

- ◆ Como as mensagens são transmitidas através da rede de interconexão?
- ◆ Existem dois métodos básicos:
  - Chaveamento de Pacotes
    - ◆ A mensagem é dividida em pacotes que são enviados individualmente pela rede.
  - Chaveamento por Circuito
    - ◆ Um circuito físico é estabelecido entre o destino e a origem para a transmissão da mensagem.

# Chaveamento de pacotes

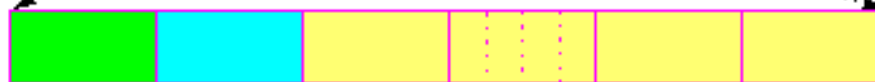
- ◆ Conhecido também como “store-and-forward”
- ◆ A mensagem é dividida em pacotes que são enviados independentemente através da rede de conexão
- ◆ O pacote será transmitido para o nó vizinho apenas se houver espaço disponível para o armazenamento
- ◆  $L = (P / B) * D$ 
  - $P \rightarrow$  comprimento do pacote
  - $B \rightarrow$  largura de banda do canal
  - $D \rightarrow$  distância entre os nós
- ◆ A latência é proporcional à distância entre os nós

# Formato da Mensagem

message



packet

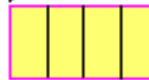


dest.  
adr.

seq.  
num.

data  
flit

flit



phit



# Formato da Mensagem

- ◆ **mensagem:** A unidade de comunicação do ponto de vista do programador. Seu tamanho é limitado apenas pelo espaço na memória de usuário.
- ◆ **pacote:** Menor unidade de comunicação de tamanho fixo contendo informação de roteamento (p. ex., endereço de destino) e de sequenciamento no seu cabeçalho. Seu tamanho é da ordem de centenas a dezenas de bytes ou palavras.

# Formato da Mensagem

- ◆ **flit:** Menor unidade de informação no nível do “link”, com o tamanho de uma ou várias palavras. Os Flits podem ser de diversos tipos e o protocolo para o envio de um flit consome diversos ciclos.
- ◆ **phit:** A menor unidade de informação no nível físico que é transferida através de um link físico em um ciclo.

# Chaveamento por circuito

- ◆ Todo um caminho é estabelecido pelo envio de uma pequena mensagem de “sonda” antes do envio da mensagem principal.
- ◆ Os canais que constituem o circuito são reservados exclusivamente.
- ◆  $L = (P / B) * D + M / B$ 
  - $P \rightarrow$  comprimento da mensagem de “sonda”
  - $M \rightarrow$  comprimento da mensagem
  - Se  $P \ll M$ , latência independe da distância

# Virtual cut-through

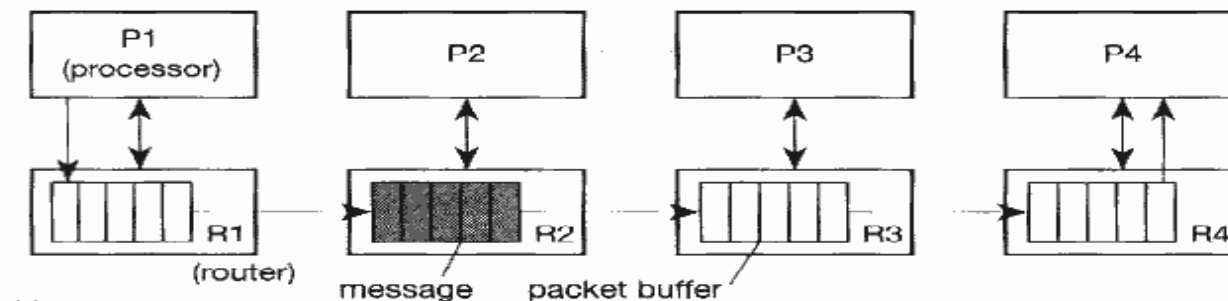
- ◆ Solução de compromisso.
- ◆ A mensagem é dividida em pequenas unidades chamadas “flow control digits” ou “flits”.
- ◆ Os flits são enviados, em modo pipeline, enquanto os canais estiverem disponíveis. Se algum canal requisitado estiver ocupado, os flits são armazenados nos nós intermediários .
- ◆  $L = (HF / B) D + M / B$ 
  - $HF \rightarrow$  comprimento do flit de cabeçalho
  - Se  $HF \ll M$ , então independe da distância



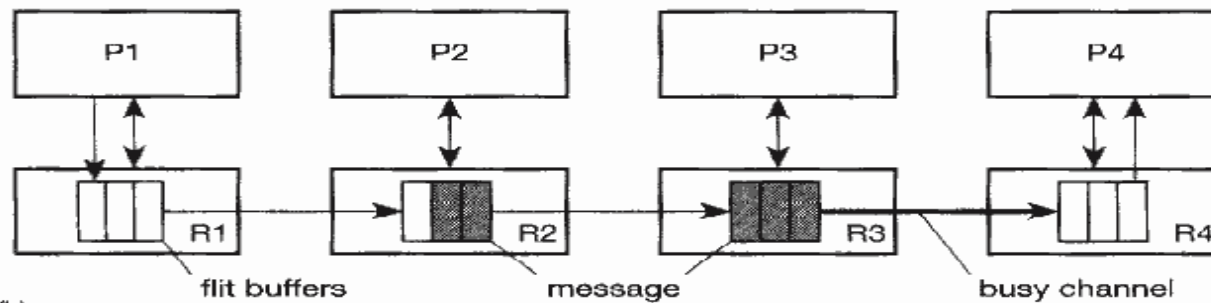
# Wormhole Routing

- ◆ É um caso especial do “virtual cut-through” onde a capacidade de armazenamento nos nós intermediários é igual a 1 flit.
- ◆ Pode realizar replicação de pacotes, enviando cópias de flits para diversos canais de saída para implementar “multicast” e “broadcast”.
- ◆ Com o uso de múltiplos “buffers” para cada canal, é possível implementar canais virtuais, para que diversas mensagens possam compartilhar o mesmo canal físico.
- ◆  $L = (HF / B) * D + M / B$ 
  - $HF \rightarrow$  comprimento do flit de cabeçalho
  - Se  $HF \ll M$ , então independe da distância

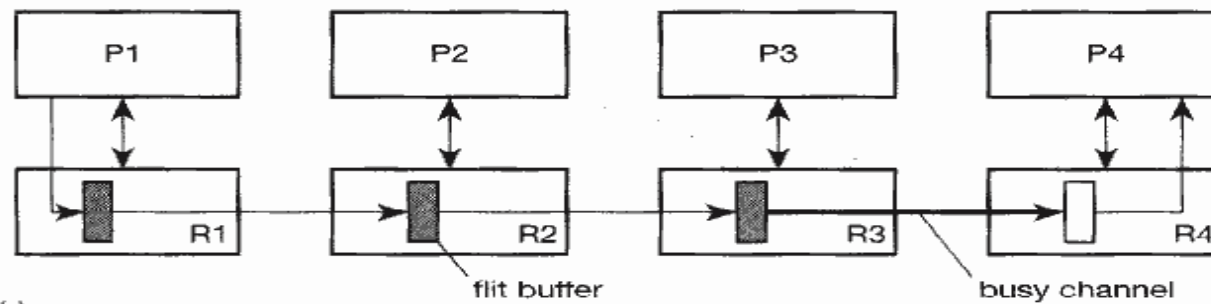
# Técnicas de Chaveamento



(a)

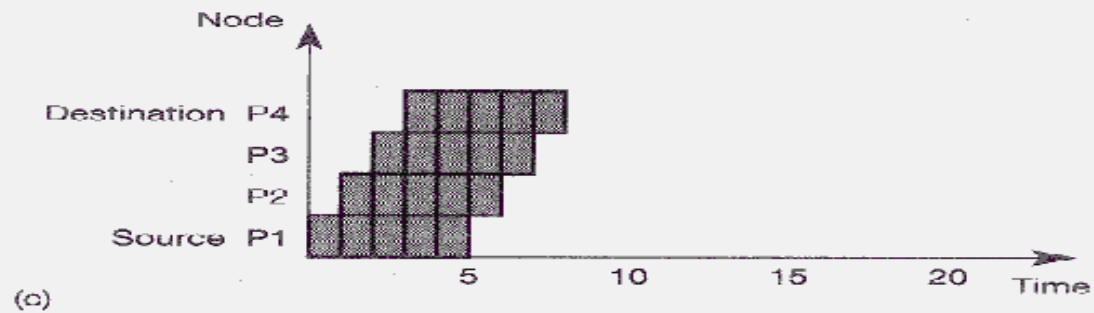
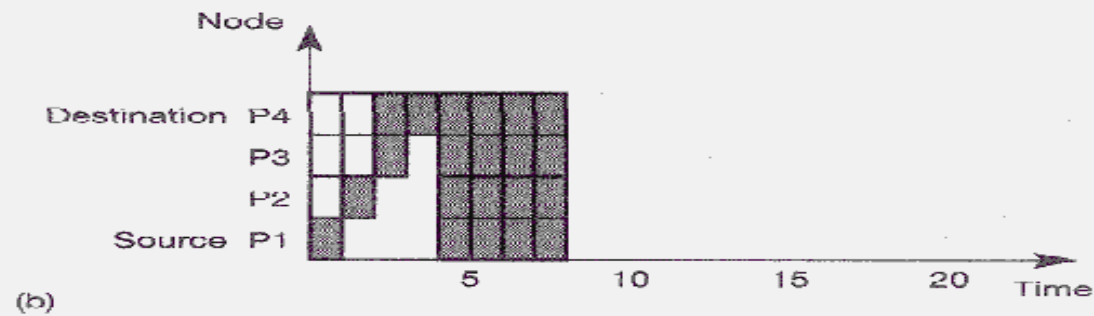
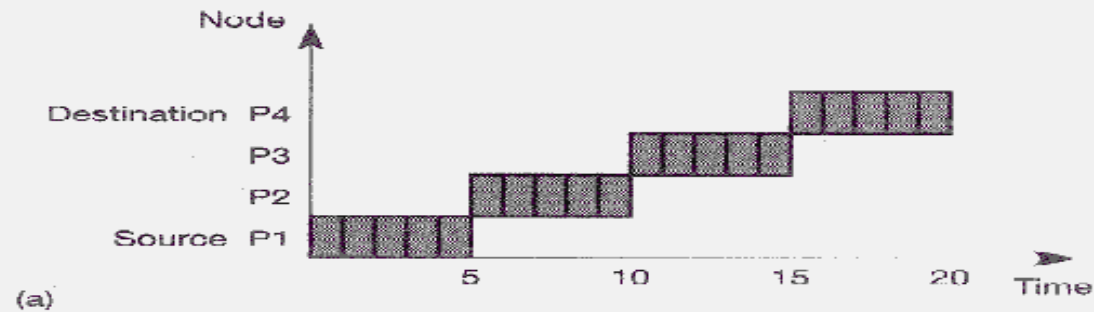


(b)

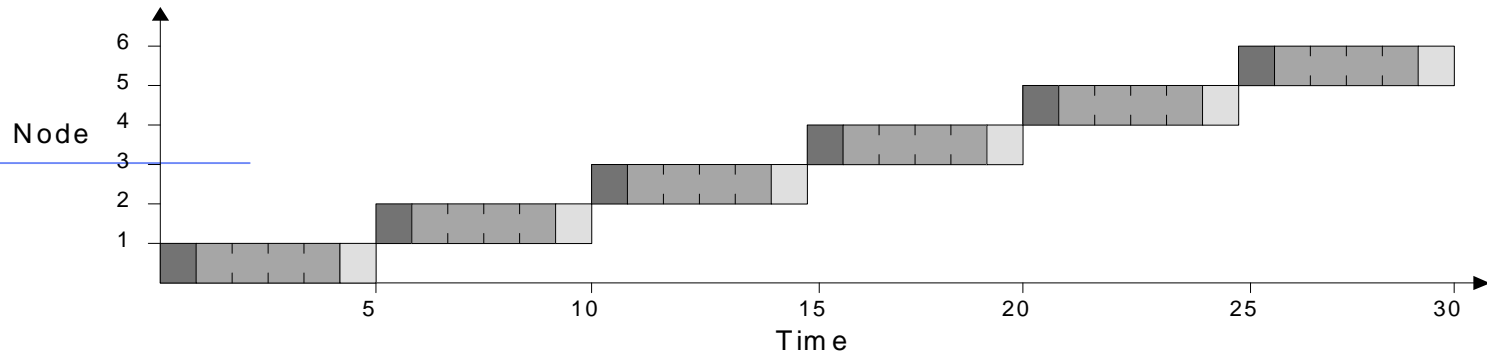


(c)

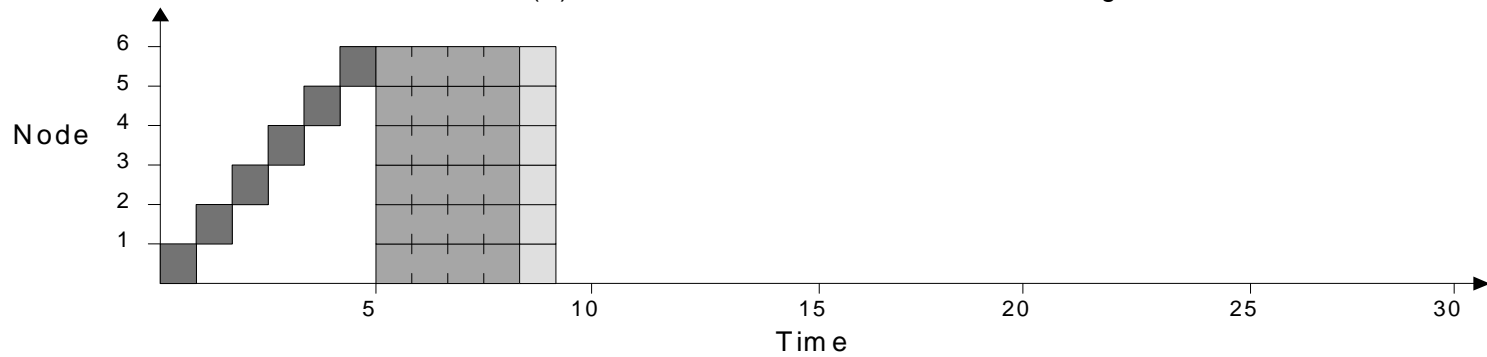
# Técnicas de Chaveamento



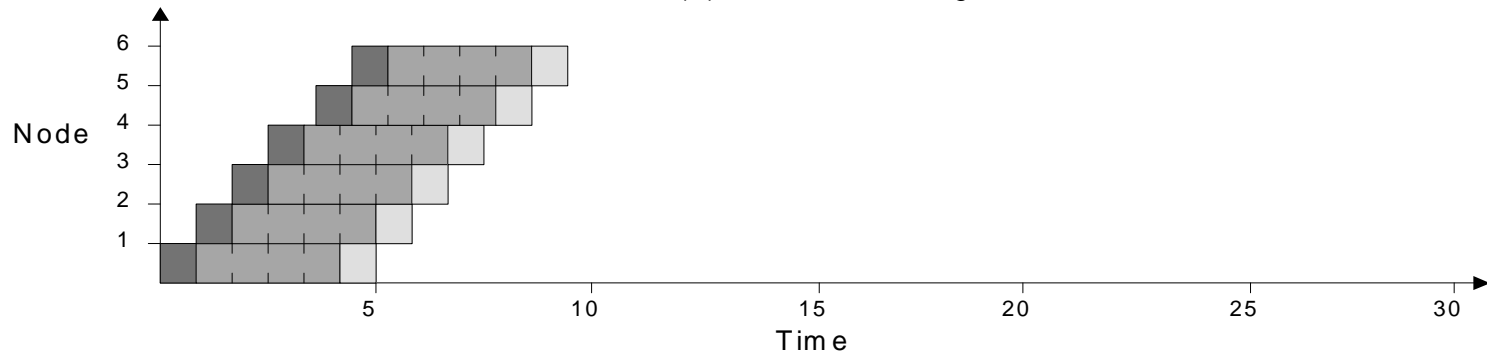
Header or Probe    Message Body    Tail



(a) Store-and-Forward Packet Switching

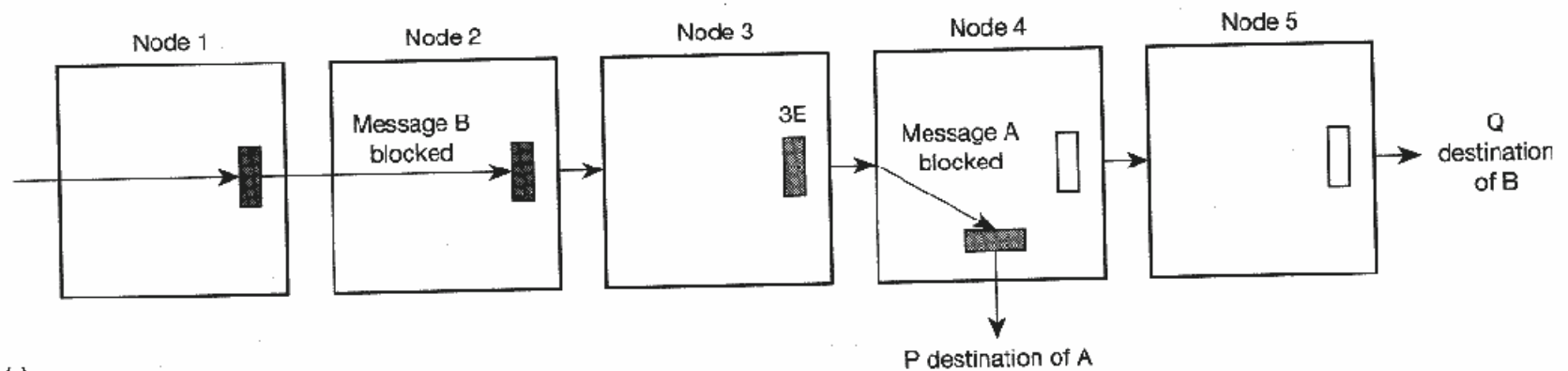


(b) Circuit Switching

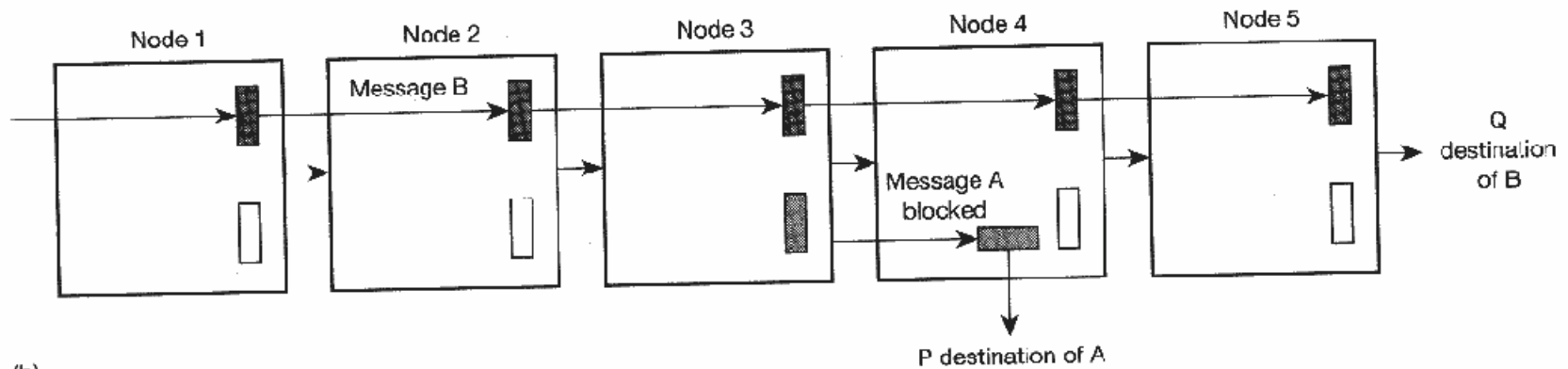


(c) Wormhole Routing or Virtual Cut-Through

# Roteamento – Canais Virtuais



(a)



(b)

# Roteamento – Canais Virtuais

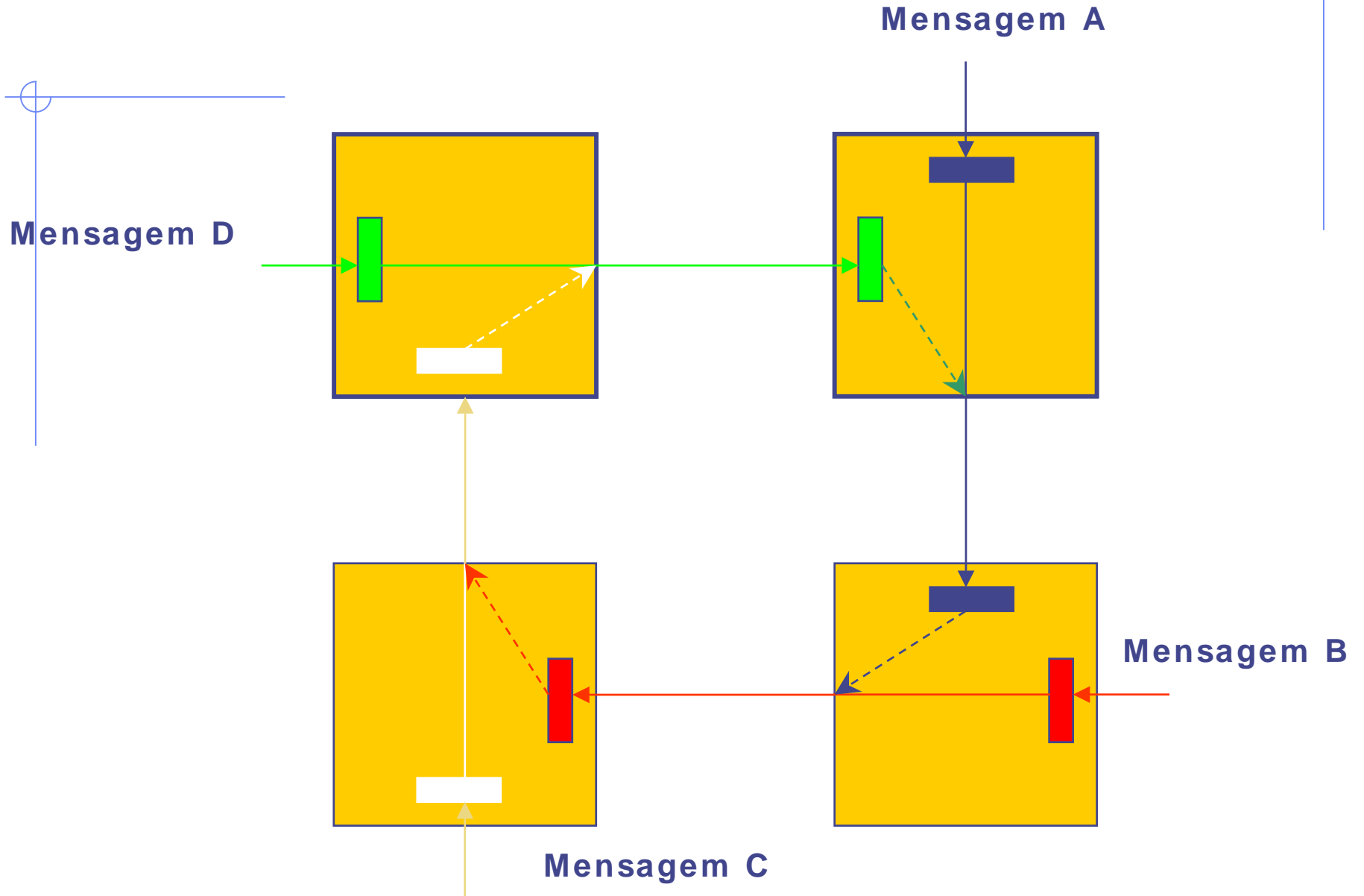
## ◆ Vantagens

- Aumentam o “throughput” da rede pela redução do tempo de ociosidade do canal físico
- Evitam a ocorrência de “deadlock”
- Facilitam o mapeamento da topologia de comunicação dos processos em uma topologia física específica
- Podem garantir a largura da banda para certas funções de sistemas, como monitoramento e depuração

# Roteamento – Deadlock

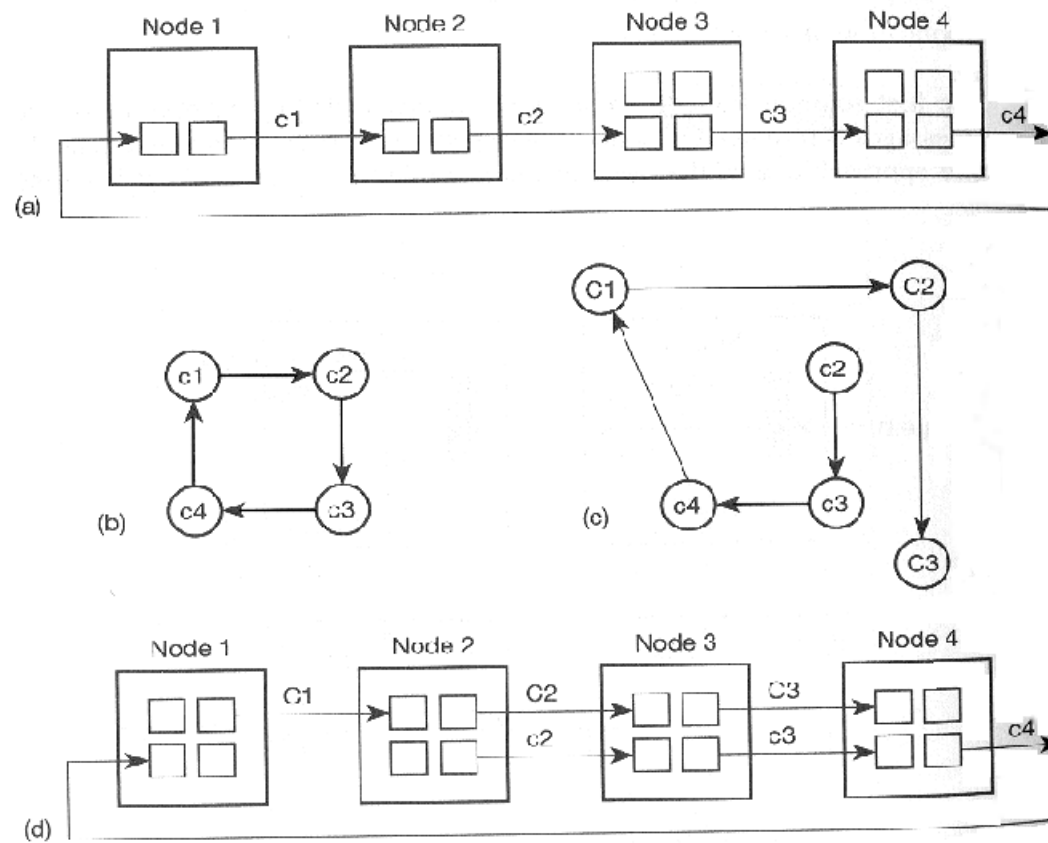
- ◆ “Deadlock” é uma situação onde um subconjunto de mensagens está mutuamente bloqueado, esperando por um “buffer” ser liberado por alguma das outras mensagens deste subconjunto
- ◆ Métodos de resolução de deadlock, que são causados pelo estabelecimento de ciclos fechados:
  - Preempção das mensagens por re-roteamento
  - Preempção das mensagens por descarte
  - Uso de canais virtuais.
- ◆ Desde que o número de canais virtuais seja suficiente, é sempre possível quebrar os ciclos fechados nos caminhos de transmissão de pacotes e, conseqüentemente, evitar a ocorrência de deadlocks.

# Roteamento – Deadlock





# Roteamento – Deadlock



**Figure 17.12** Deadlock avoidance by virtual channels. (a) Physical link interconnections; (b) dependency graph for interconnection (a); (c) dependency graph for interconnection (d); (d) virtual channel interconnections.

# Roteamento – Multicast e Broadcast

- ◆ Algumas redes de interconexão possuem recursos de hardware no roteamento de mensagens para suportar diferentes tipos de operações de comunicação.
- ◆ Todas as redes suportam a comunicação ponto-a-ponto ou “unicast”
- ◆ Operações coletivas
  - “broadcasting” (um nó origem envia uma mesma mensagem para todos os outros)
  - “multicasting” (um nó origem envia a mesma mensagem para um grupo especificado de nós destino).

# Broadcasting

- ◆ Algumas redes de interconexão possuem recursos de hardware no roteamento de mensagens para suportar diferentes tipos de operações de comunicação.
- ◆ Todas as redes suportam a comunicação ponto-a-ponto ou “*unicast*”.
- ◆ Operações coletivas:
  - “*broadcasting*” (um nó origem envia uma mesma mensagem para todos os outros)
  - “*multicasting*” (um nó origem envia a mesma mensagem para um grupo especificado de nós destino).

# Algoritmos

## ◆ Determinísticos

- O caminho é completamente determinado pelo endereço dos nós fonte e destino. Os nós intermediários, mesmo no caso de um congestionamento, não podem redirecionar as mensagens.

## ◆ Adaptativos

- No roteamento adaptativo os nós intermediários levam em conta o estado atual da rede para determinar a direção para qual a mensagem deve ser enviada.

# Algoritmos Determinísticos

## Roteamento “street-sign”

- Utilizado no roteamento do Iwarp
- Do tipo “Source routing”, ou seja, a informação de roteamento é montada no nó fonte.

## ◆ Roteamento ordenado por dimensão

- Utilizado na “J-Machine”
- Aplicado em Malhas N-dimensionais
- Do tipo “Roteamento Distribuído”

## ◆ Roteamento por tabela de busca

- Do tipo “Roteamento Distribuído”
- Em cada nó existe uma tabela indicando para qual vizinho a mensagem deve ser roteada, de acordo com o endereço destino
- O IMS T9000 utiliza uma variante deste algoritmo chamada de “interval labelling”.

# Algoritmos Adaptativos

## ◆ Profitable

- Seleciona apenas aqueles canais que garantidamente levam a mensagem mais próxima do seu destino
- Resultam em um caminho de menor comprimento
- Não sofrem de “livelock”
- São mais fáceis de demonstrar que são livres de “deadlock”

## ◆ Misrouting

- Seleciona indistintamente qualquer dos canais
- São vantajosos quando há canais defeituosos na rede

# Algoritmos Adaptativos

## ◆ Progressivo

- Mensagens não podem voltar no caminho que elas já percorreram

## ◆ Backtracking

- As mensagens podem voltar e explorar todas as opções entre os nós fonte e destino
- Os cabeçalhos devem conter informação para assegurar que não incorrerão em “livelock”
- São livres de “deadlock”
- Não pode ser usado com “wormhole”
- Implica em cabeçalhos muito longos e maior latência

# Algoritmos Adaptativos

- ◆ Os protocolos adaptativos podem ser ainda completamente ou parcialmente adaptativos
- ◆ Exemplos:
  - “turn model” → parcialmente adaptativo, progressivo, misrouting
  - “west-first” → parcialmente adaptativo, progressivo, misrouting



This document was created with Win2PDF available at <http://www.daneprairie.com>.  
The unregistered version of Win2PDF is for evaluation or non-commercial use only.