Arquiteturas com Memória Distribuída

Arquitetura de Computadores

Emilio Francesquini
e.francesquini@ufabc.edu.br

2020.Q1

Centro de Matemática, Computação e Cognição
Universidade Federal do ABC

UFABC

Disclaimer

UFABC

- Estes slides foram preparados para o curso de **Arquitetura de Computadores** na UFABC.
- Este material pode ser usado livremente desde que sejam mantidos, além deste aviso, os créditos aos autores e instituições.
- O conteúdo destes slides foi baseado no conteúdo do livro *Computer Organization And Design: The Hardware/Software Interface*, 5th Edition.
- Contém algumas figuras por Peter Pacheco, *An Introduction to Parallel Programming* disponíveis em: `https://www.cs.usfca.edu/~peter/ipp/`

Arquiteturas com Memória Distribuída

Sistema com memória compartilhada

UFABC

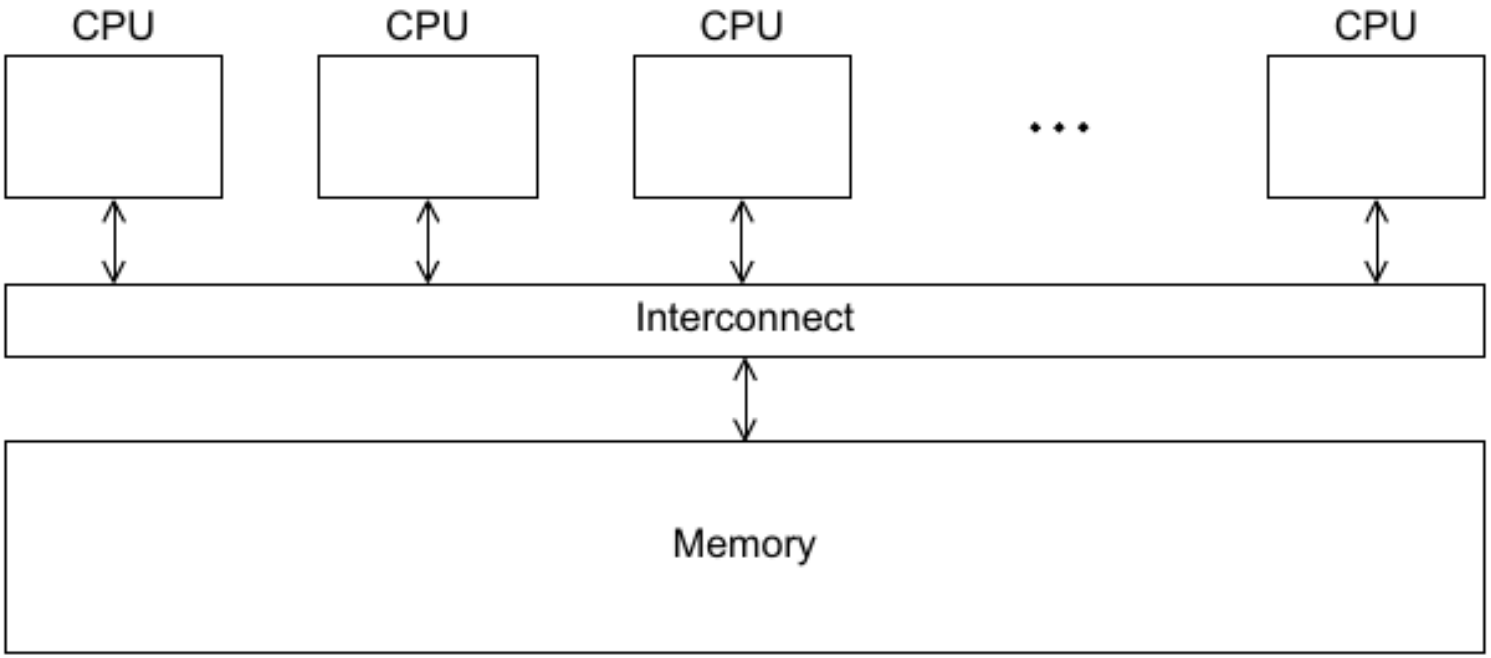
- Um conjunto de processadores autônomos conectados a um sistema de memória por uma rede de interconexão
- Cada processador pode acessar cada uma das posições de memória
- Os processadores se comunicam, tipicamente, implicitamente pelo acesso a estruturas de dados compartilhadas na memória

Sistema com memória compartilhada (2)

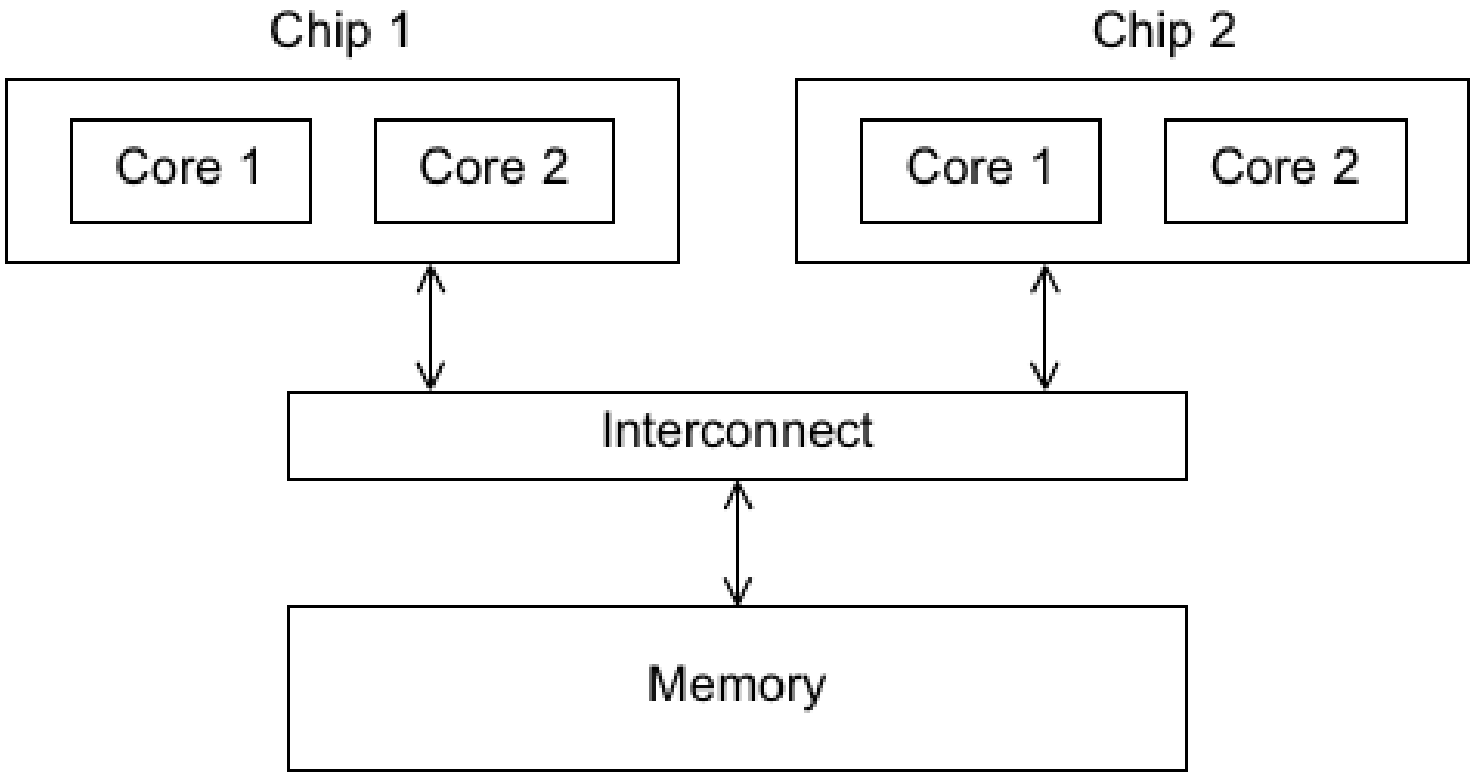
UFABC

- Os sistemas com memória compartilhada mais comuns oferecem um ou mais processadores multicore
  - Um ou mais cores ou processadores em um único chip

Sistema com memória compartilhada (2)

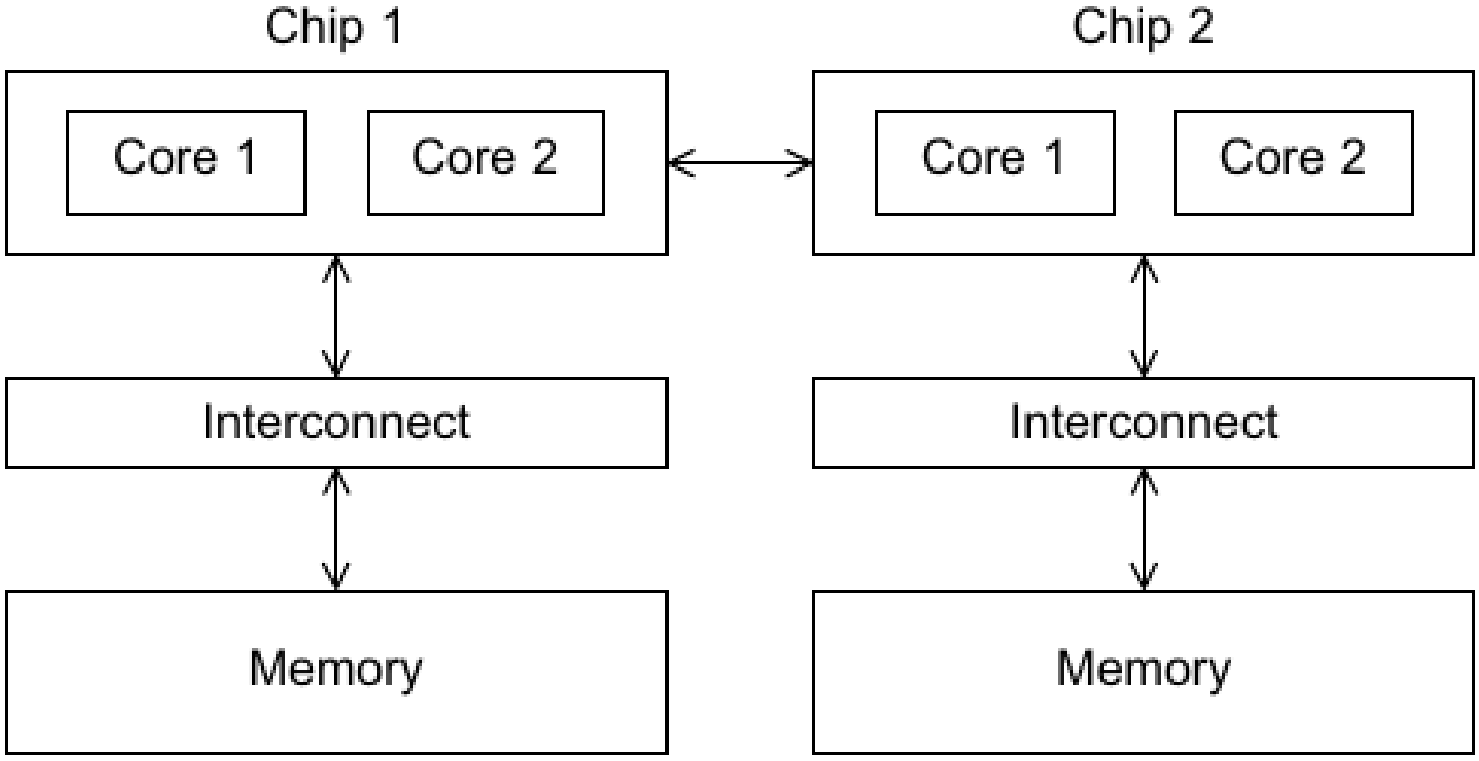


Uniform Memory Access - O tempo para acessar cada uma das posições de memória é sempre o mesmo, independentemente do core ou do endereço

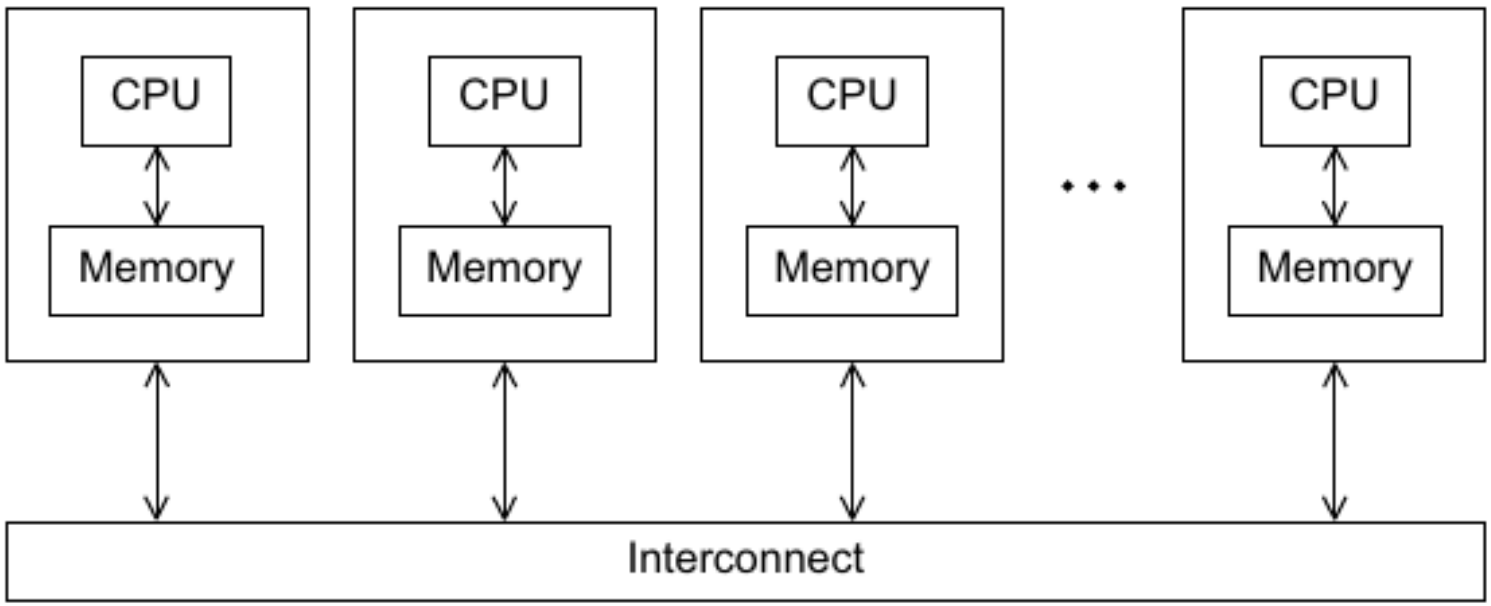


UFABC

Non-Uniform Memory Access - O tempo de acesso pode variar entre cada um dos cores ou dependendo do endereço



Sistemas com memória distribuída



Sistemas com memória distribuída

- Clusters - Modelo mais popular
  - Um conjunto de computadores comuns (com memória compartilhada)
  - Conexão usando redes
- Nós/Nodos/Nodes de um cluster são cada uma das unidades computacionais ligadas por uma rede de interconexão
- Também chamados de sistemas híbridos

Lembrando algumas definições

UFABC

- Para toda troca de dados é importante sabermos quanto tempo que os dados alcancem o seu destino
- Latência - O tempo que leva entre a origem **começar** a transmitir os dados e o destino **começar** a receber o primeiro byte
- Largura de banda - A taxa pela qual o destinatário recebe os dados após ter começado a receber o primeiro byte

Tempo de transmissão

Tempo = Latência (s) + Tamanho (bytes) / Banda (bytes/s)

Clusters

- Cada nó tem sua memória privada e uma cópia do SO
- Apropriados para aplicações com tarefas independentes
  - Servidores web, bancos de dados, simulações
  - Casamento perfeito para paralelismo no nível de requisições (*request level parallelism*)
- Tem como vantagens um baixo custo, alta disponibilidade, desempenho
  - Permitem vender o tempo livre e ainda ter lucro! IaaS, SaaS, ...
  - Baixa banda de comunicação
    - em comparação com aquela disponível em uma máquina UMA/SMP

Exemplo de aplicação: MapReduce - Word Count

UFABC

- Somar contar a ocorrência de palavras em um conjunto de documentos.
- Map - Primeiro distribui os documentos entre os processadores e faz uma separação inicial.

```
1 map(String key, String value):
2   //key: document name
3   //value: document contents
4   for each word w in value:
5     EmitIntermediate(w, “1”); //Produce list of all
      ↪  words
```

Exemplo de aplicação: MapReduce - Word Count

UFABC

- Reduce - Agrupa os resultados parciais.

```
1  reduce(String key, Iterator values):
2    //key: a word
3    //values: a list of counts
4    int result = 0;
5    for each v in values:
6      result += ParseInt(v); //get integer from key-value
       ↪  pair
```

UFABC

- Computadores espalhados por amplas regiões geográficas
  - Tipicamente conectados pela Internet.
  - Pacotes de trabalho são enviados por um coordenador e resultados são enviados de volta.
- Teve seu exemplo de sucesso no projeto SETI@Home
  - Hoje em dia tem muitos outros exemplos para procura de fármacos, números primos e simulações diversas.

Redes de interconexão

Redes de interconexão

UFABC

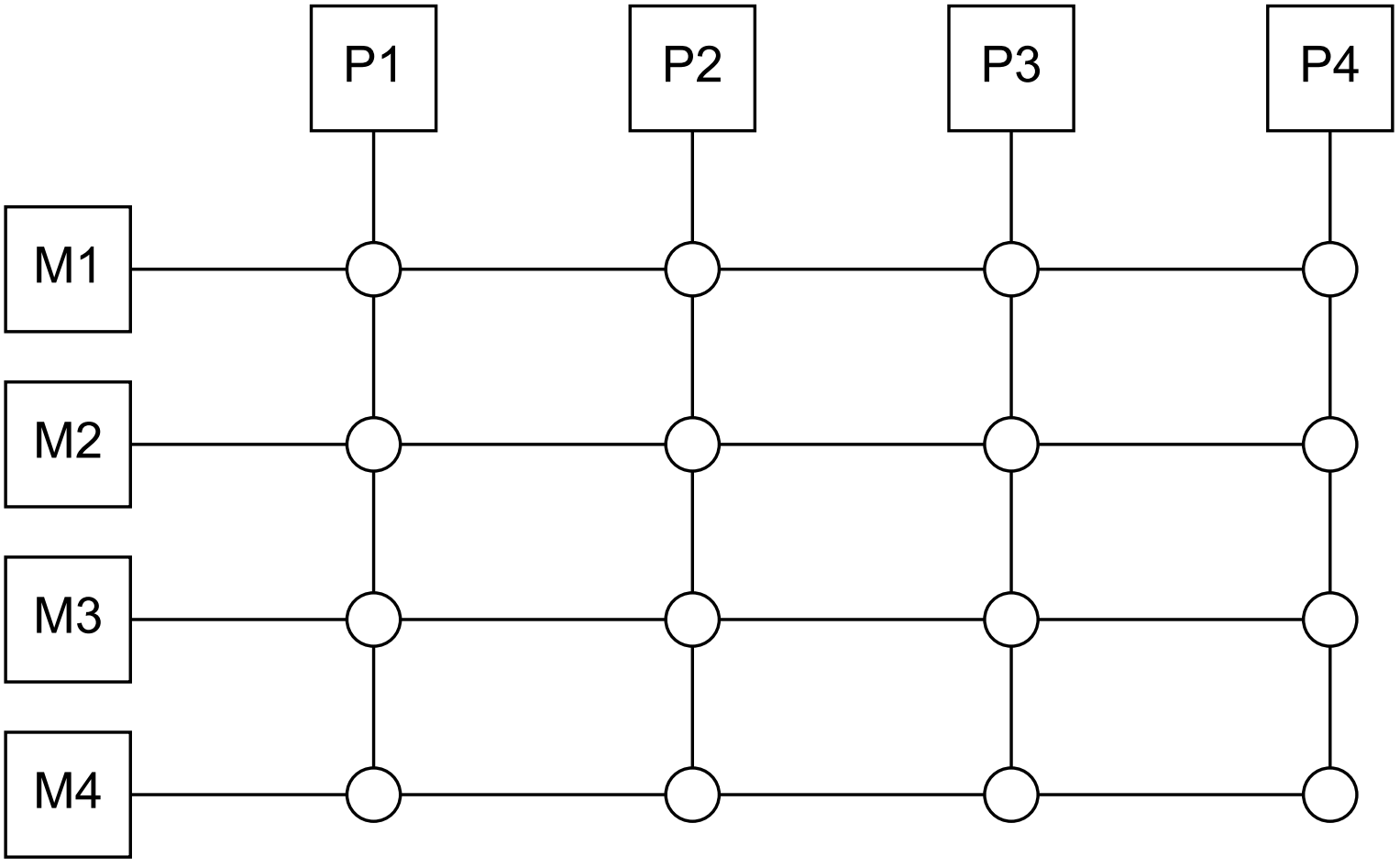
- Afetam o desempenho tanto de sistemas distribuídos quanto de memória compartilhada
- Duas categorias
  - Interconexões de memória compartilhada
  - Interconexões de memória distribuída

UFABC

- Barramento / *Bus*
  - Uma coleção de ligações (fios e cabos) paralelos em conjunto com um hardware que controla o acesso ao barramento
  - As vias de conexão são compartilhadas entre os dispositivos conectados
  - Conforme o número de dispositivos aumenta, também aumenta a contenção e, consequentemente, há uma diminuição no desempenho

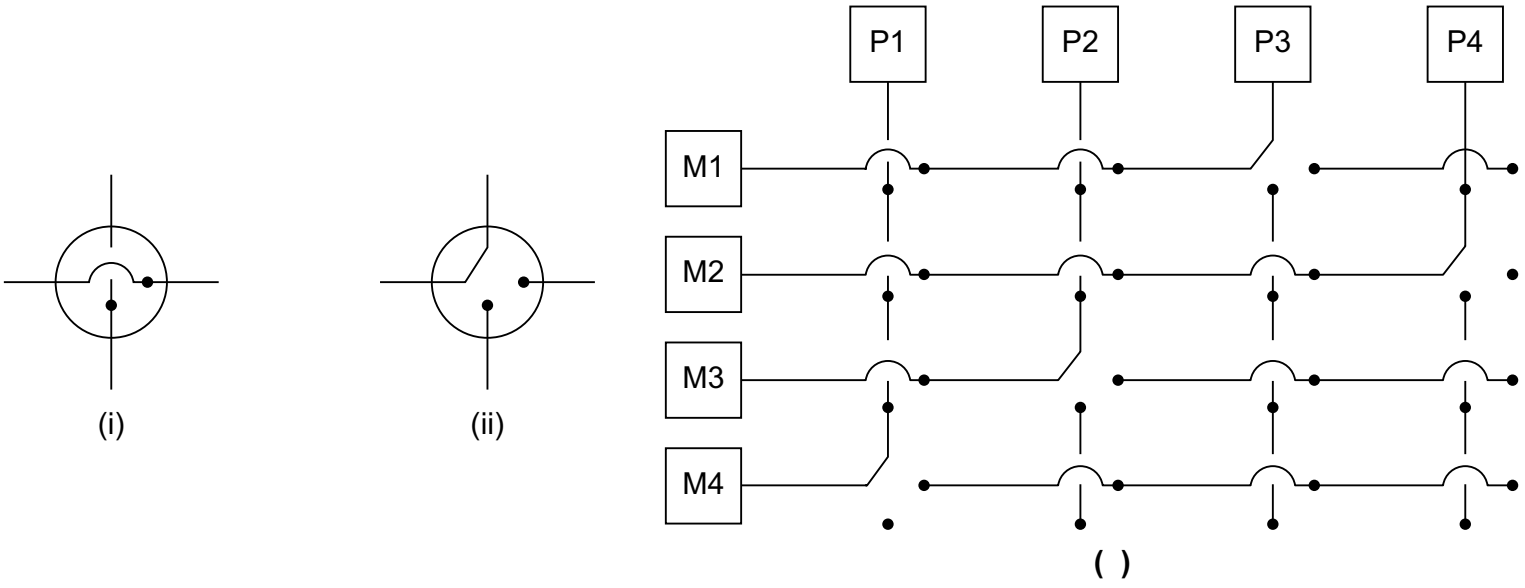
UFABC

- Interconexão chaveada / *Switched interconnect*
  - Utiliza chaveadores (*switches*) para controlar o fluxo e o roteamento dos dados entre os dispositivos
  - Crossbar
    - Permite comunicação simultânea entre vários dispositivos diferentes
    - Mais rápido que barramentos
    - Contudo, o custo é mais elevado do que barramentos



Crossbar conectando 4 processadores e 4 módulos de memória

Crossbar Switch



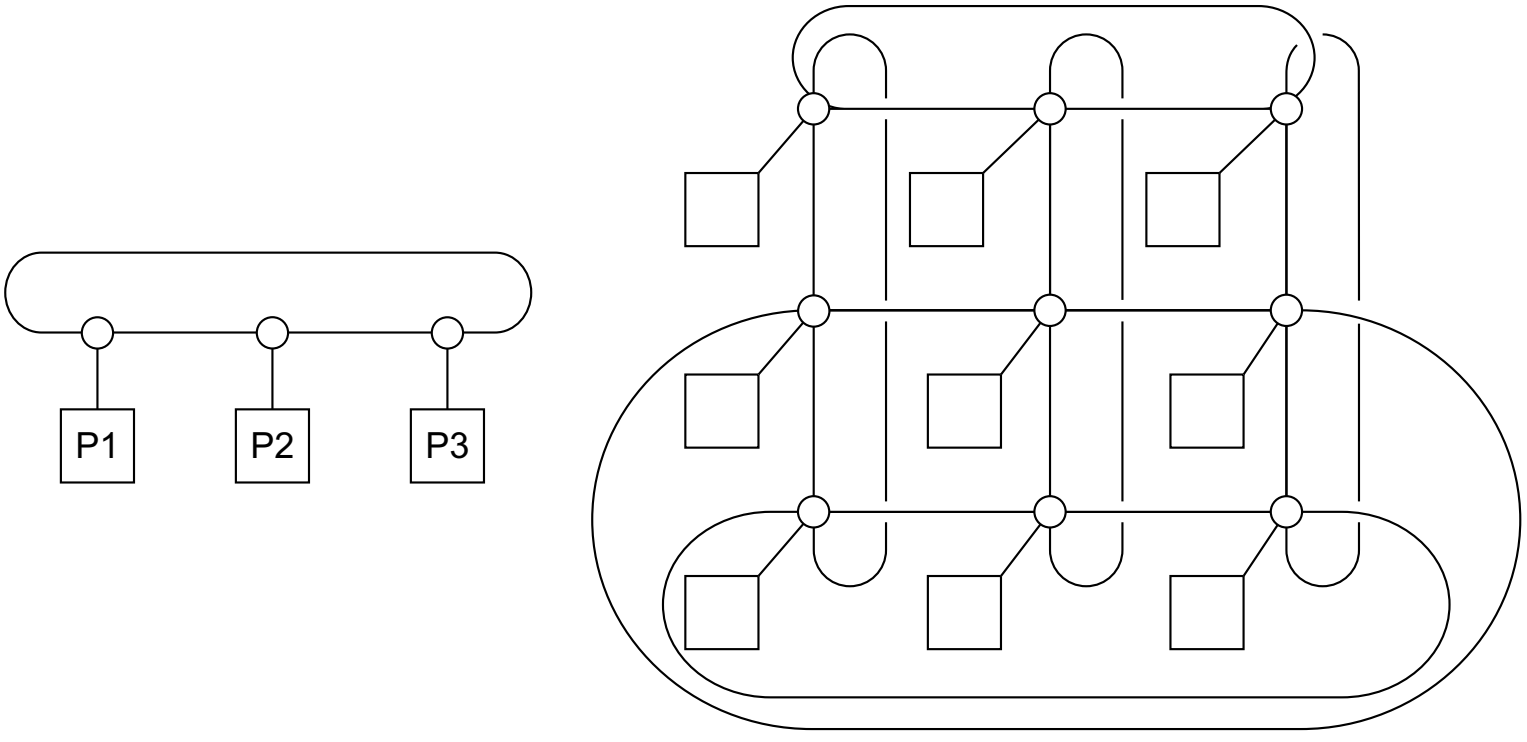
Acesso simultâneo à memória por vários processadores

Interconexões de memória distribuída

UFABC

- Dois grupos
  - Interconexão direta
    - Cada switch é diretamente conectado a um par processador/memória e os switches são conectados uns aos outros
  - Interconexão indireta
    - Os switches podem não estar conectados diretamente a um processador

Interconexão direta



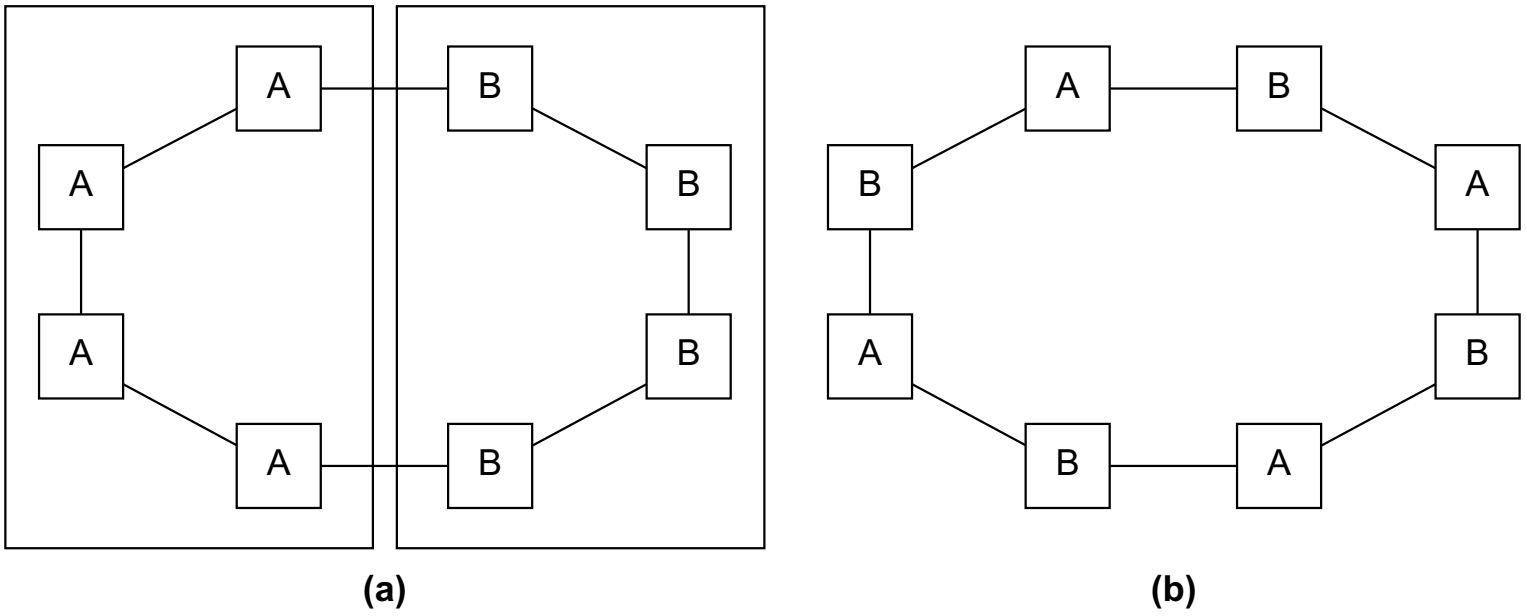
Conexão em (a) anel e (b) toroidal

Largura da bisseção

UFABC

- Medida do ”número de comunicações simultâneas”ou da conectividade
- Quantas comunicações podem ocorrer ao mesmo tempo através de cada parte?

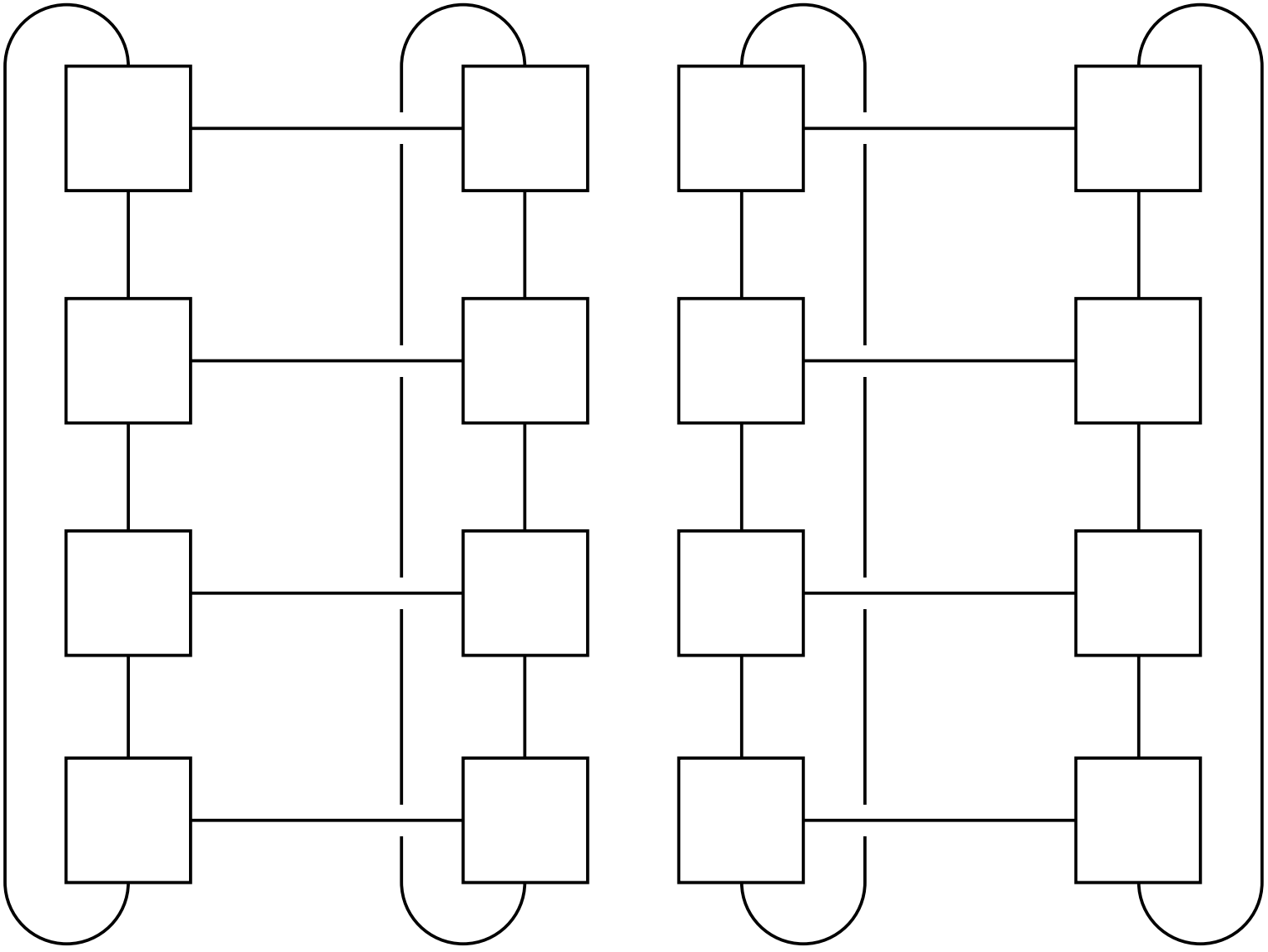
Exemplo: Anel



(a) 2 conexões (b) 4 conexões

Bisseção de uma interconexão toroidal

UFABC



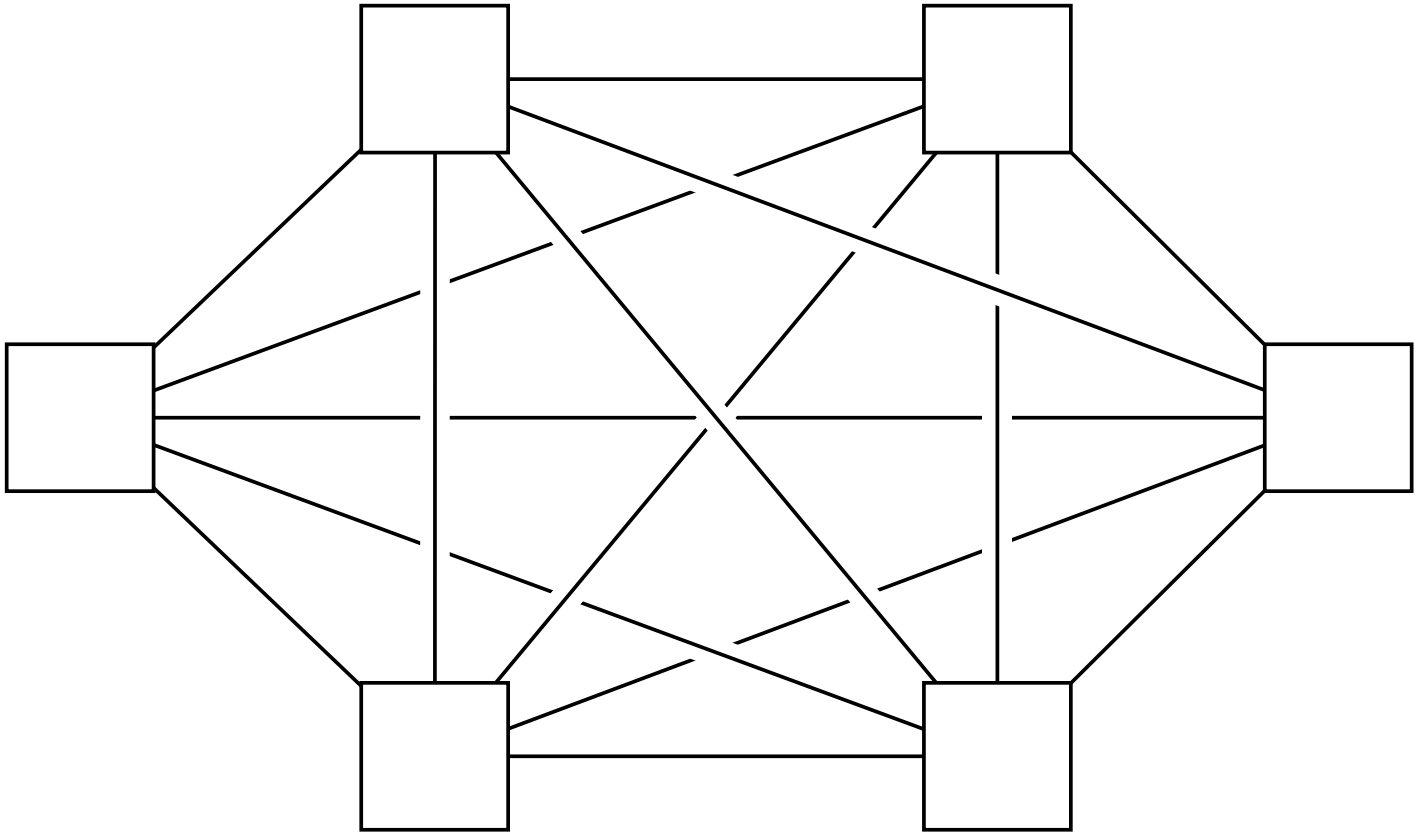
UFABC

- Largura de banda / *Bandwidth*
  - Taxa de transmissão de um link
  - Normalmente dada em megabits ou megabytes por segundo
- Largura de banda da bisseção
  - Medida da qualidade da rede
  - Em vez de contar o número de links que juntam as duas metades, soma a largura de banda dos links

Rede completamente conexa

UFABC

- Cada switch está conectado a todos os outros switches
- Impraticável para valores grandes de p

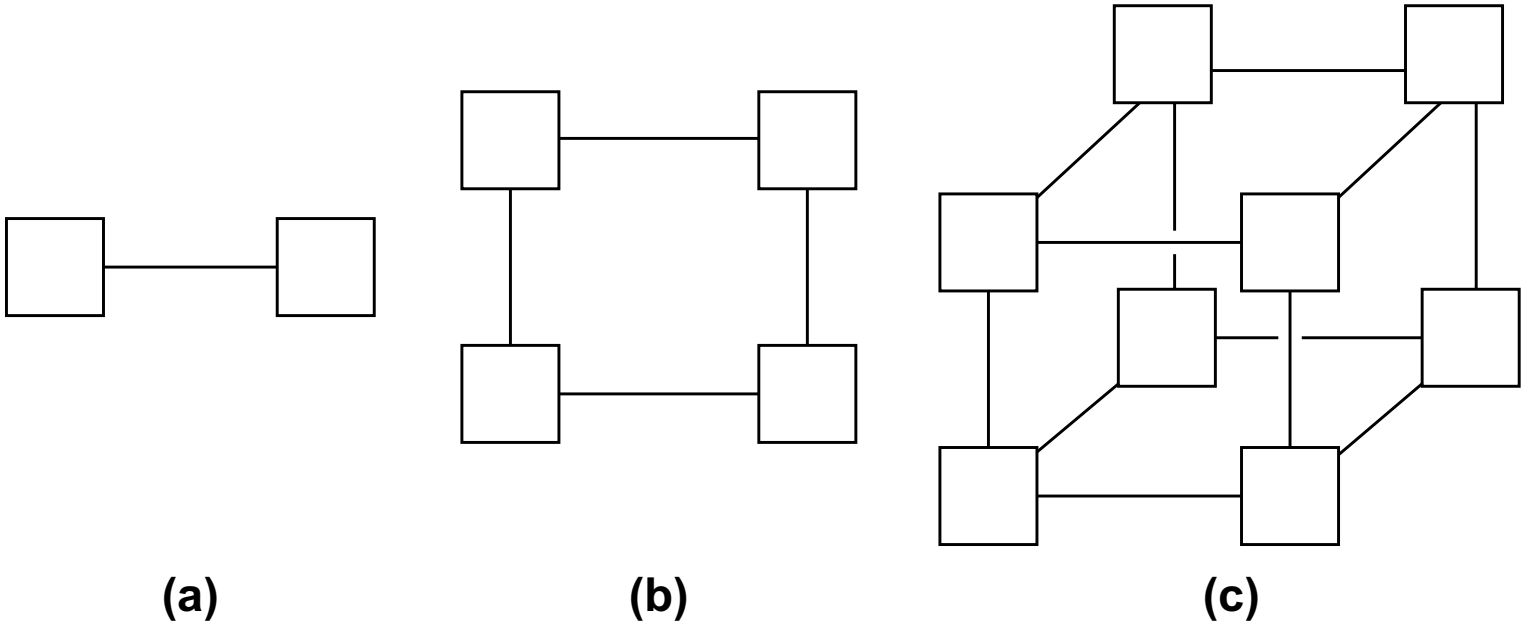


- Largura de bisseção = $\frac{p^2}{4}$

UFABC

- Interconexão direta e altamente conectada
- Construída por indução
  - Um hipercubo unidimensional é um sistema completamente conectado com dois processadores
  - Um hipercubo bidimensional é construído a partir de dois hipercubos unidimensionais pela junção dos switches ”correspondentes”
  - Da mesma maneira, um hipercubo tridimensional é um hipercubo construído a partir de dois hipercubos bidimensionais

Hipercubo



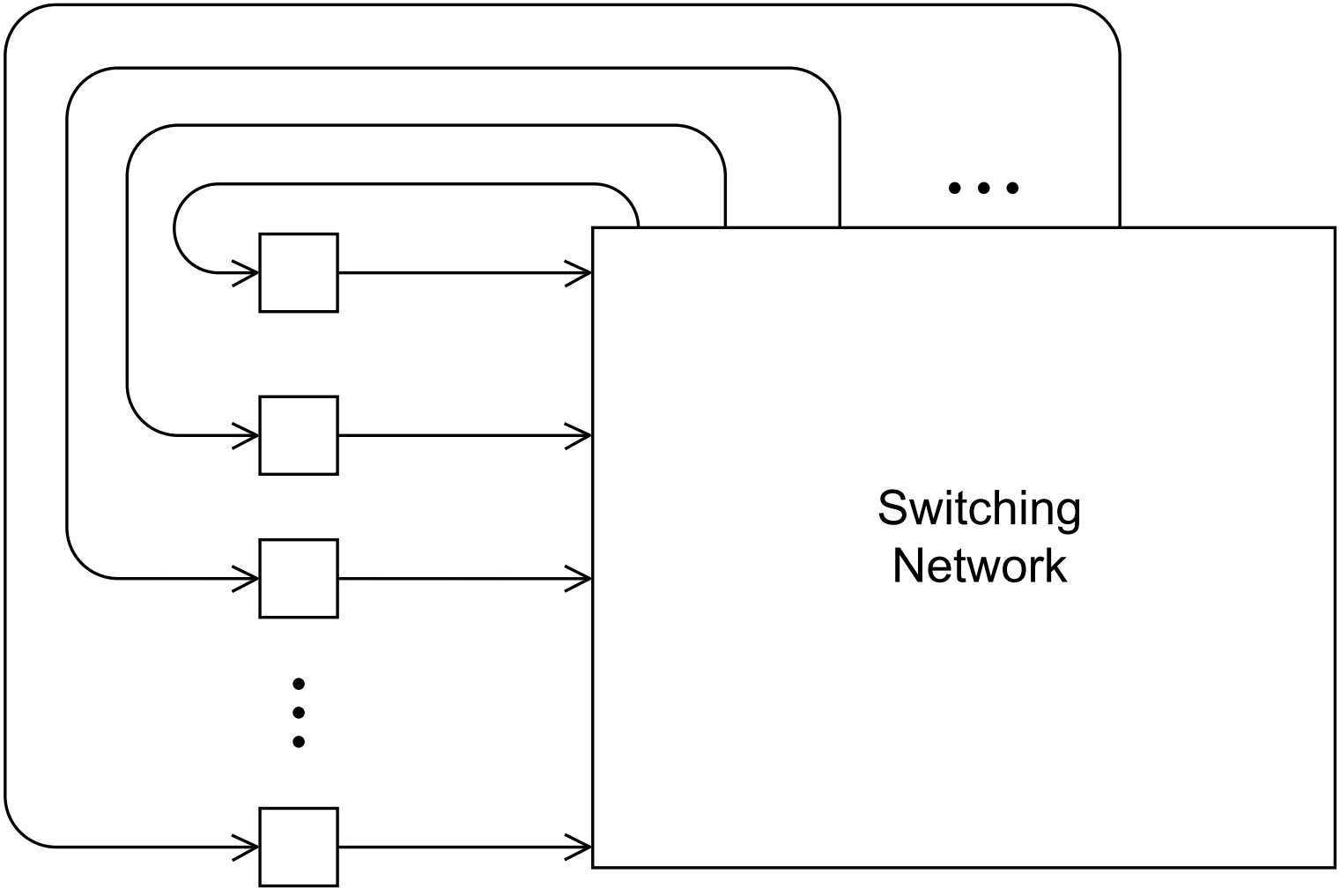
(a) 1D (b) 2D (c) 3D

Interconexões indiretas

UFABC

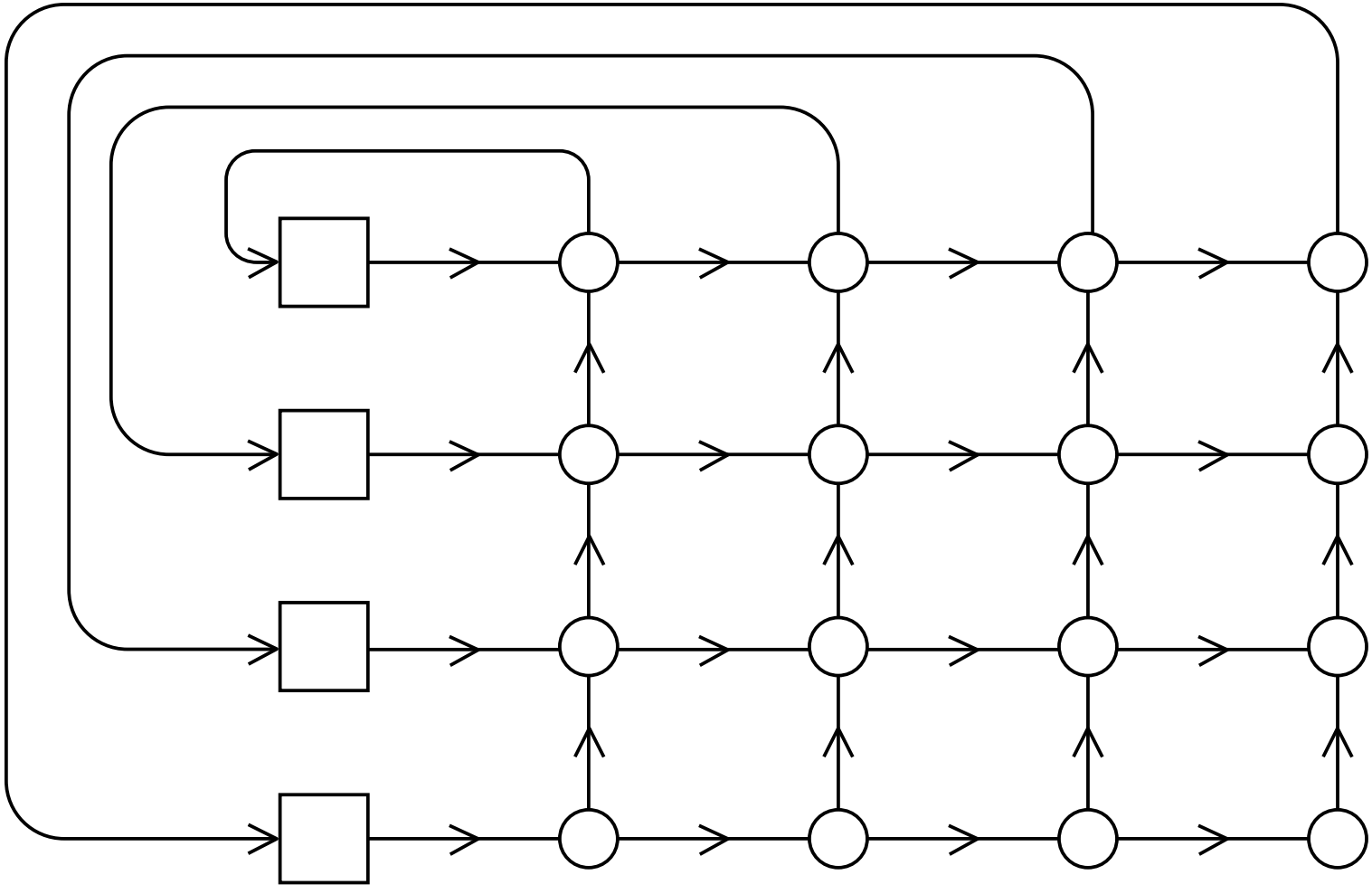
- Exemplos mais simples de redes de interconexão indiretas
  - Crossbar
  - Omega Network
- Geralmente são mostradas como redes com links unidirecionais e com uma coleção de processadores, cada um deles com um link de entrada e outro de saída, e uma rede chaveada

Uma rede indireta genérica



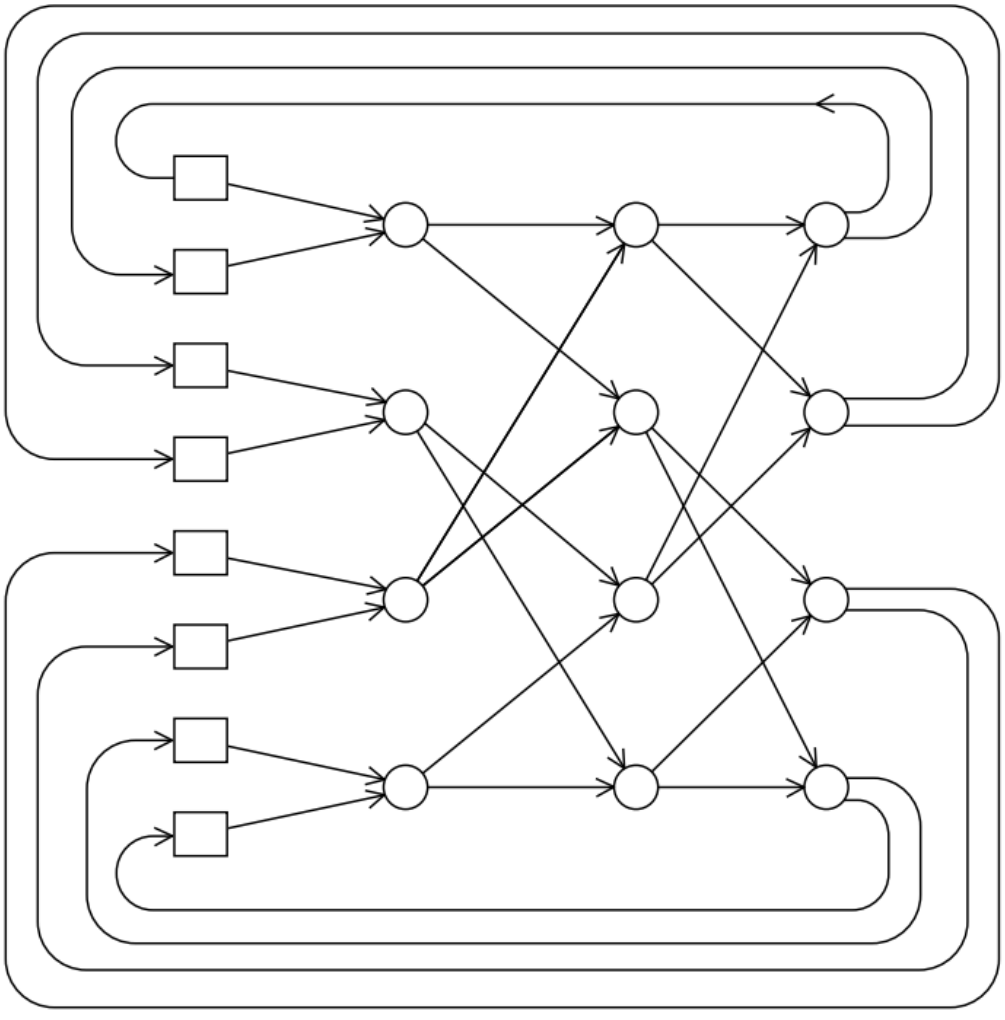
Interconexão Crossbar para memória distribuída

UFABC



Omega Network

UFABC



Um Switch em uma rede Omega

UFABC