SSC0510
Arquitetura de Computadores

14ª Aula – Arquiteturas Paralelas

Arquitetura MIMD com Memória Distribuída

Profa. Sarita Mazzini Bruschi
sarita@icmc.usp.br

Comunicação em MIMD com Memória Distribuída

- A transmissão de mensagens é uma maneira de comunicação em máquinas MIMD com Memória Distribuída
- Uma mensagem pode ser definida como um bloco de informações relacionadas que trafegam entre os processadores
- Como nas máquinas MIMD com memória compartilhada, os programas são particionados em pequenas partes, cada uma podendo ser executada por um processador de maneira concorrente

Comunicação em MIMD com Memória Distribuída

- O tamanho do processo que está sendo executado é determinado pelo programador e pode ser descrito através de sua granulosidade:

$$Granulosidade = \frac{tempo\,de\,computação}{tempo\,de\,comunicação}$$

Comunicação em MIMD com Memória Distribuída

- Granulosidade grossa (*coarse granularity*)*:* cada processo gasta uma grande quantidade de tempo processando instruções sequenciais
- Granulosidade média (*medium granularity*): balanço entre computação e comunicação
- Granulosidade fina (*fine granularity*): cada processo executa poucas instruções sequenciais

Comunicação em MIMD com Memória Distribuída

- Fator importante: Rede de Interconexão
- Fatores de projeto das redes de interconexão:
  - *Bandwidth* do link de comunicação
    - Número de bits que podem ser transmitidos por unidade de tempo (bits/segundo)
    - Links podem ser unidirecionais ou bidirecionais e transmitir um ou vários bits de cada vez
  - Latência da rede
    - Tempo para completar a transferência de uma mensagem
    - Pode ser estimada a partir do tempo de determinação do caminho, que depende do número de nós, e do tempo de transmissão, que depende do tamanho da mensagem

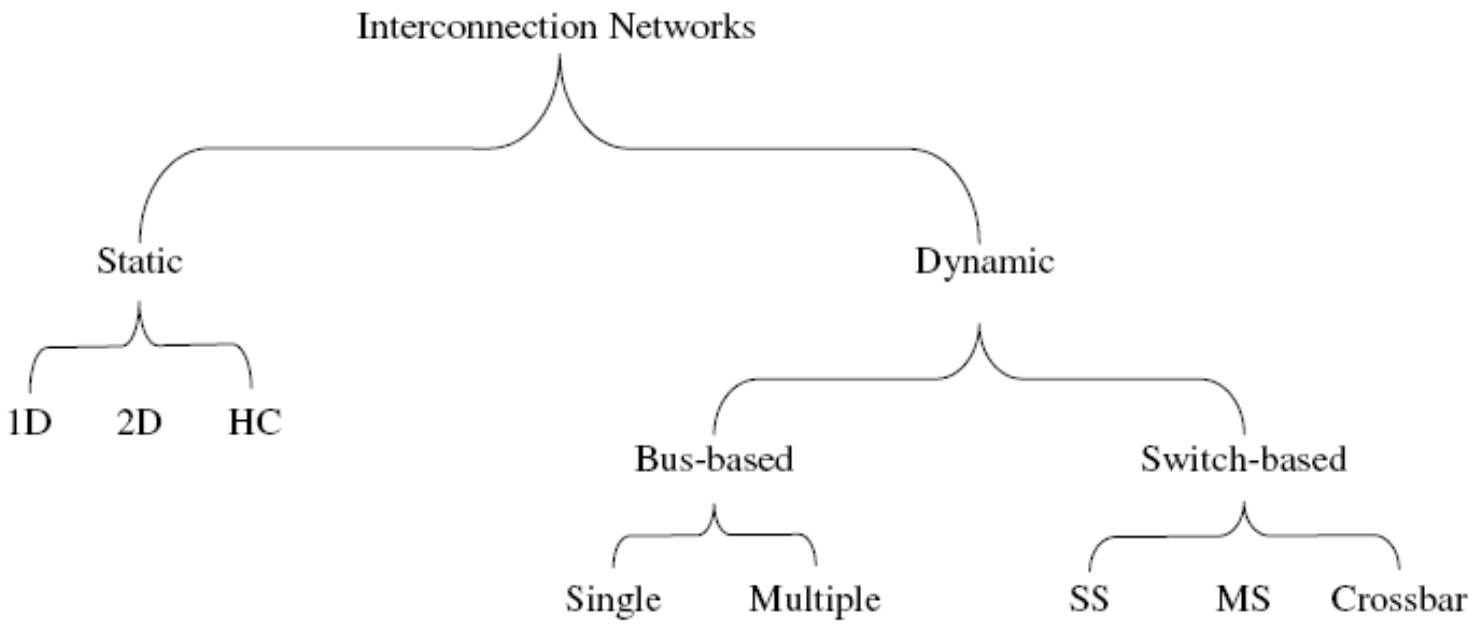
Redes de Interconexão

- Critérios utilizados para a classificação das redes:
  - Modo de operação
    - Síncrona: um único clock é usado por todos os componentes do sistema
    - Assíncrona: não precisa de um clock global (utiliza protocolos de *handshaking*)
  - Estratégia de controle
    - Centralizada: uma única unidade de controle central é utilizada para supervionar e controlar a operação de todos os componentes do sistema.
    - Descentralizada: a função de controle é distribuída entre os componentes do sistema.

Redes de Interconexão

- Critérios utilizados para a classificação das redes:
  - Técnicas de comutação:
    - Circuito: um caminho completo tem que ser estabelecido antes de se iniciar a transmissão, a qual se manterá fixa durante todo o período de comunicação
    - Pacote: comunicação feita através de mensagens, divididas em unidades menores denominadas pacotes
  - Topologia
    - Estática: existem caminhos fixos e diretos entre os nós
    - Dinâmica: conexões entre entradas e saídas são feitas utilizando elementos de conexão, denominados chaves. Dependendo da configuração desses elementos, diferentes interconexões podem ser estabelecidas

Redes de Interconexão
Topologias



1D – Uma dimensão
2D – Duas dimensões
HC – Hiper Cubo

SS – Single-stage
MS – Multi-stage

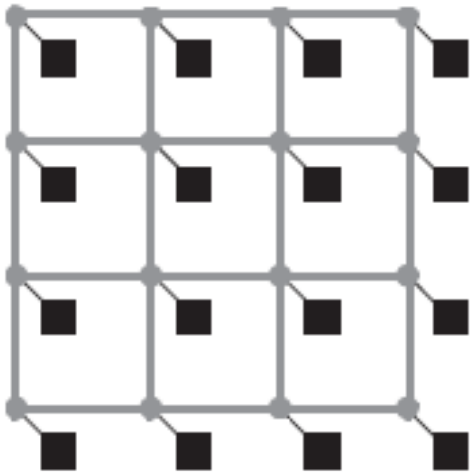
Redes de Interconexão
Topologias

- Critérios para avaliação de uma topologia:
  - **Número total de ligações entre os componentes**
  - **Grau do nó**: quantas ligações diretas cada componente possui
  - **Diâmetro**:maior distância entre dois componentes quaisquer da rede

Redes de Interconexão
Topologias - Exemplos

- Topologia estática
  - Anel

  - Malha bi-dimensional

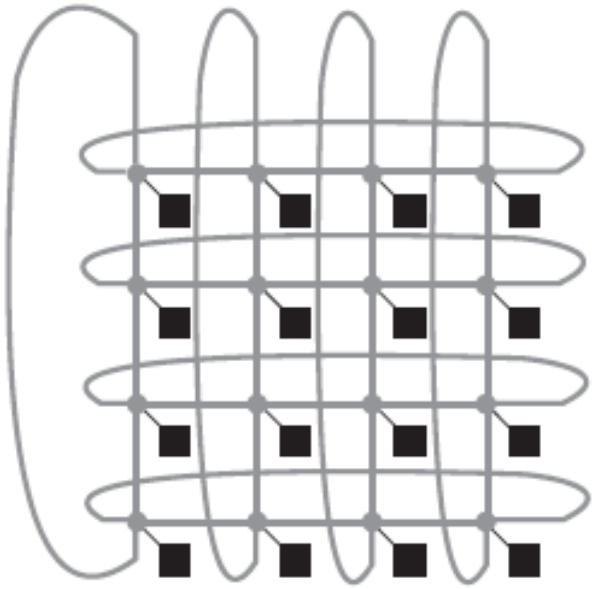


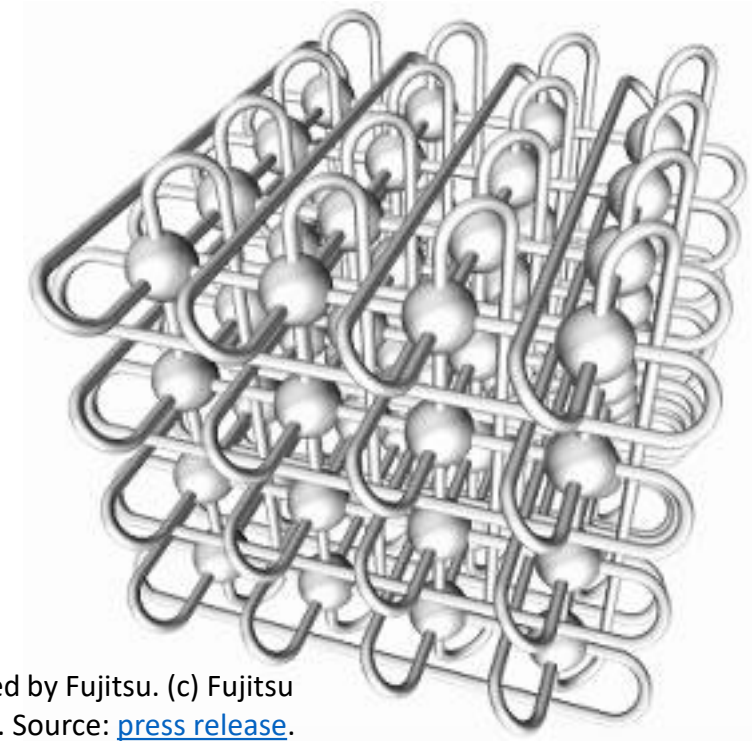
Processador/Memória

Chaves

Redes de Interconexão
Topologias - Exemplos

- Topologia estática
  - Torus (malha bi-dimensional)



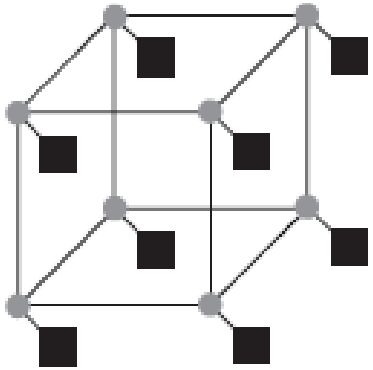


3D torus rendered by Fujitsu. (c) Fujitsu and RIKEN, 2009. Source: press release.

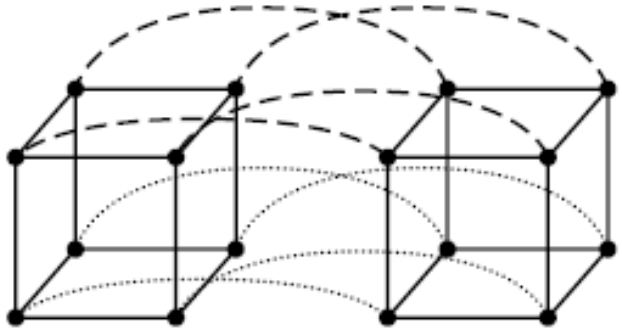
http://www.fujitsu.com/global/about/businesspolicy/tech/k/whatis/network/index.html

Redes de Interconexão
Topologias - Exemplos

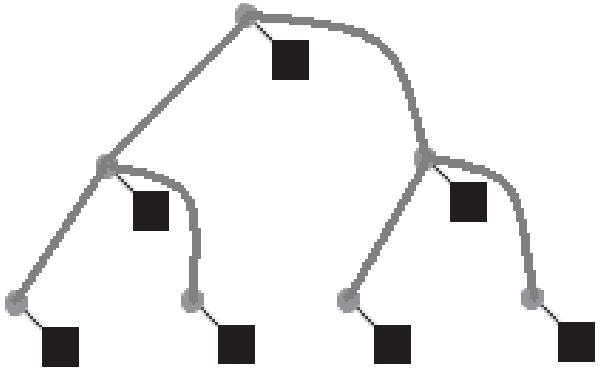
- Topologia estática
  - Cubo



Hipercubo 4D

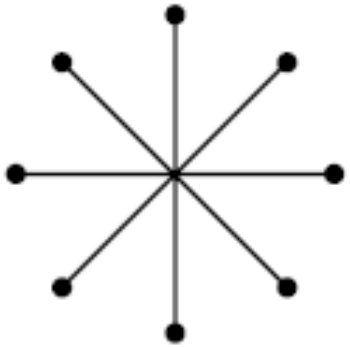


  - Árvore binária

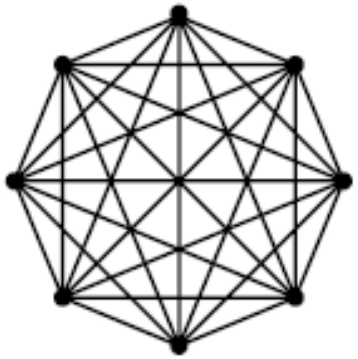


Redes de Interconexão
Topologias - Exemplos

- Topologia estática
  - Estrela

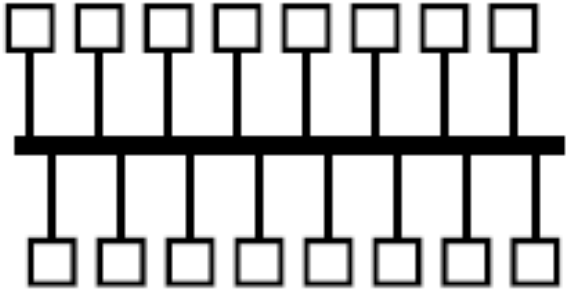


  - Totalmente conectada

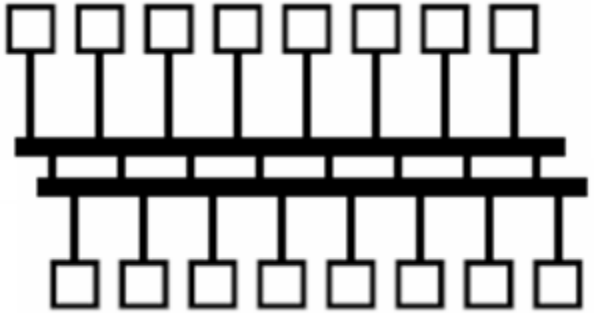


Redes de Interconexão
Topologias - Exemplos

- Topologia Dinâmica
  - Barramento
    - Single



    - Multiple

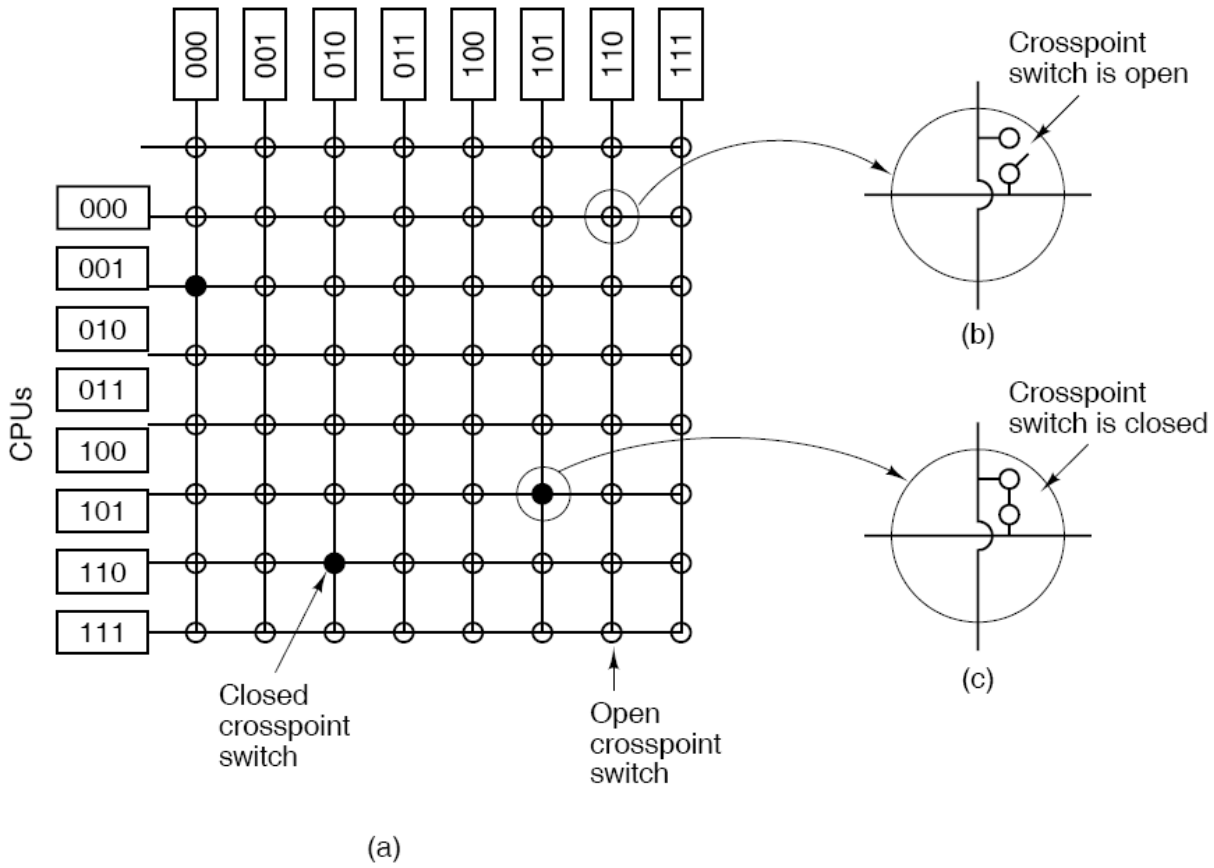


Redes de Interconexão
Topologias - Exemplos

- Topologia Dinâmica
  - Crossbar

  Proc/Mem

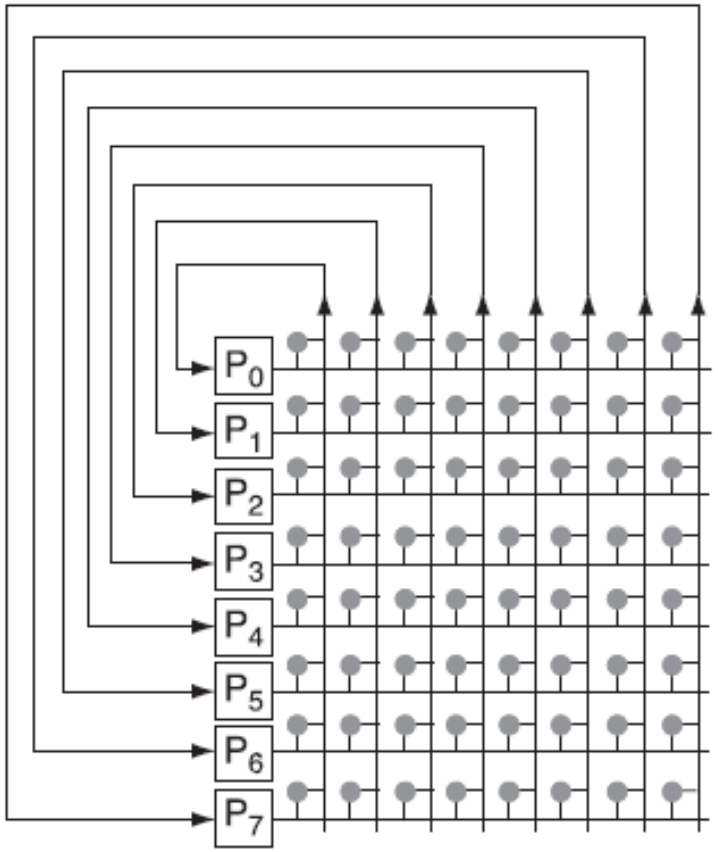
  (arquiteturas MIMD – Mem. Compartilhada)



Redes de Interconexão
Topologias - Exemplos

- Topologia Dinâmica
  - Crossbar

  Proc/Proc

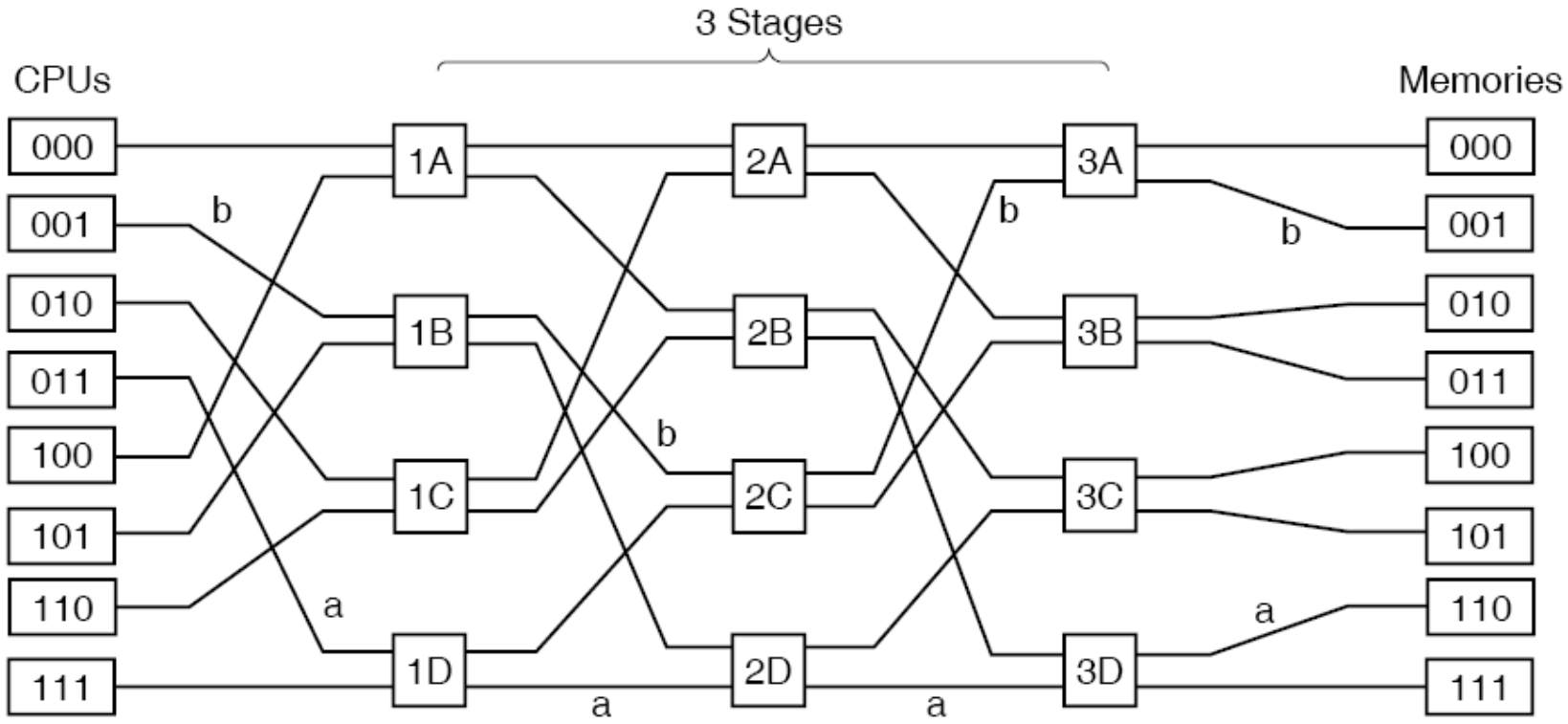


a. Crossbar

Redes de Interconexão
Topologias - Exemplos

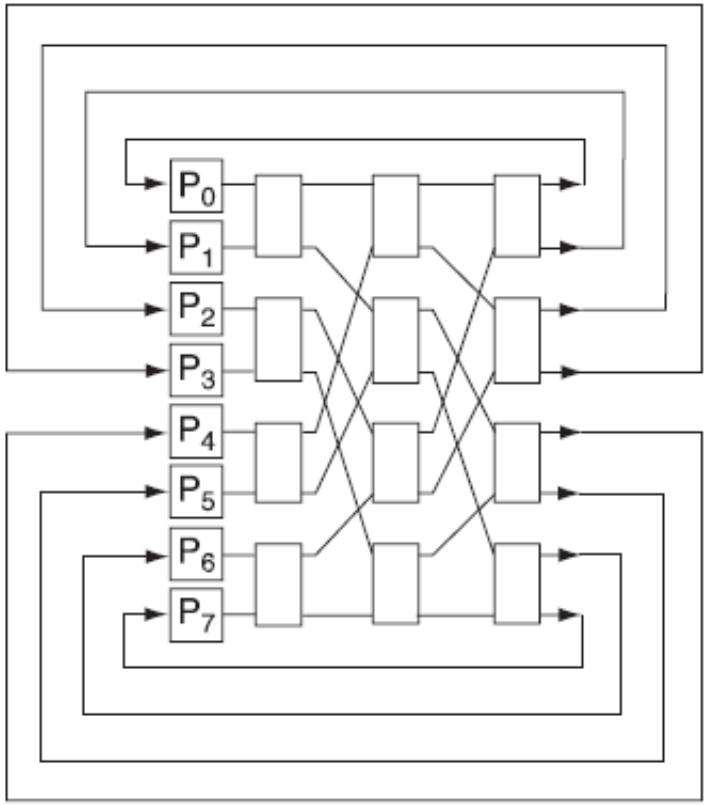
- Topologia Dinâmica

  Omega (*Multistage*) – Proc/Mem (arquiteturas MIMD – Mem. Compartilhada)

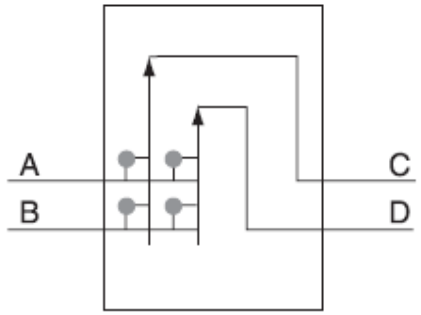


Redes de Interconexão
Topologias - Exemplos

- Topologia Dinâmica
  - Omega (*Multistage*) Proc/Proc



b. Rede Omega



Comunicação em MIMD com Memória Distribuída

- As mensagens são transmitidas através de primitivas *send* e *receive*
- Cada mensagem percorre o seguinte caminho:
  - nó de processamento origem
  - interface de rede do nó de origem
  - rede de interconexão
  - interface de rede do nó destino
  - nó de processamento destino
- A interação do nó de processamento com a sua interface de rede é realizada através de chamadas ao sistema operacional.
- As primitivas podem ser classificadas como:
  - Síncronas ou assíncronas
  - Bloqueantes ou não bloqueantes

Comunicação em MIMD com Memória Distribuída

- A latência total na comunicação por troca de mensagens ser definida pela soma das seguintes parcelas:
  - latência total = overhead de software + overhead da interface da rede + latência da rede de interconexão
- O *overhead* de software é devido a um conjunto de fatores:
  - complexidade do protocolo de comunicação,
  - processamento de interrupções
  - troca de contexto devido às chamadas ao sistema
  - cópias dos dados entre "buffers", etc

Arquitetura MIMD com Memória Distribuída

- As arquiteturas MIMD com memória distribuída, ou multicomputadores, podem variar na estrutura e no tamanho
- Dois estilos podem ser observados:
  - MPP – *Massively Parallel Processors*
  - Clusters

MPP – *Massively Parallel Processors*

- Supercomputadores muito caros
- Utilizado em computação científica, engenharia e indústria para cálculos de larga escala ou para uma grande quantidade de transações por segundo
- Utilizado também em *data warehousing* (manipulação de uma grande quantidade de dados) e *big data*

MPP – *Massively Parallel Processors*

- Composto por CPUs padrões
  - Intel Pentium
  - Sun UltraSPARC
  - IBM PowerPC
- O que faz a diferença nos MPPs é a rede de interconexão proprietária de alta performance projetada para mover mensagens com baixa latência a uma alta *bandwidth*
- Outra característica é a alta capacidade de E/S
- Hardware e software especiais para tratar tolerância a falhas (monitoramento do sistema, detecção e recuperação das falhas)
- Exemplos: BlueGene (IBM), K Computer (Fujitsu)

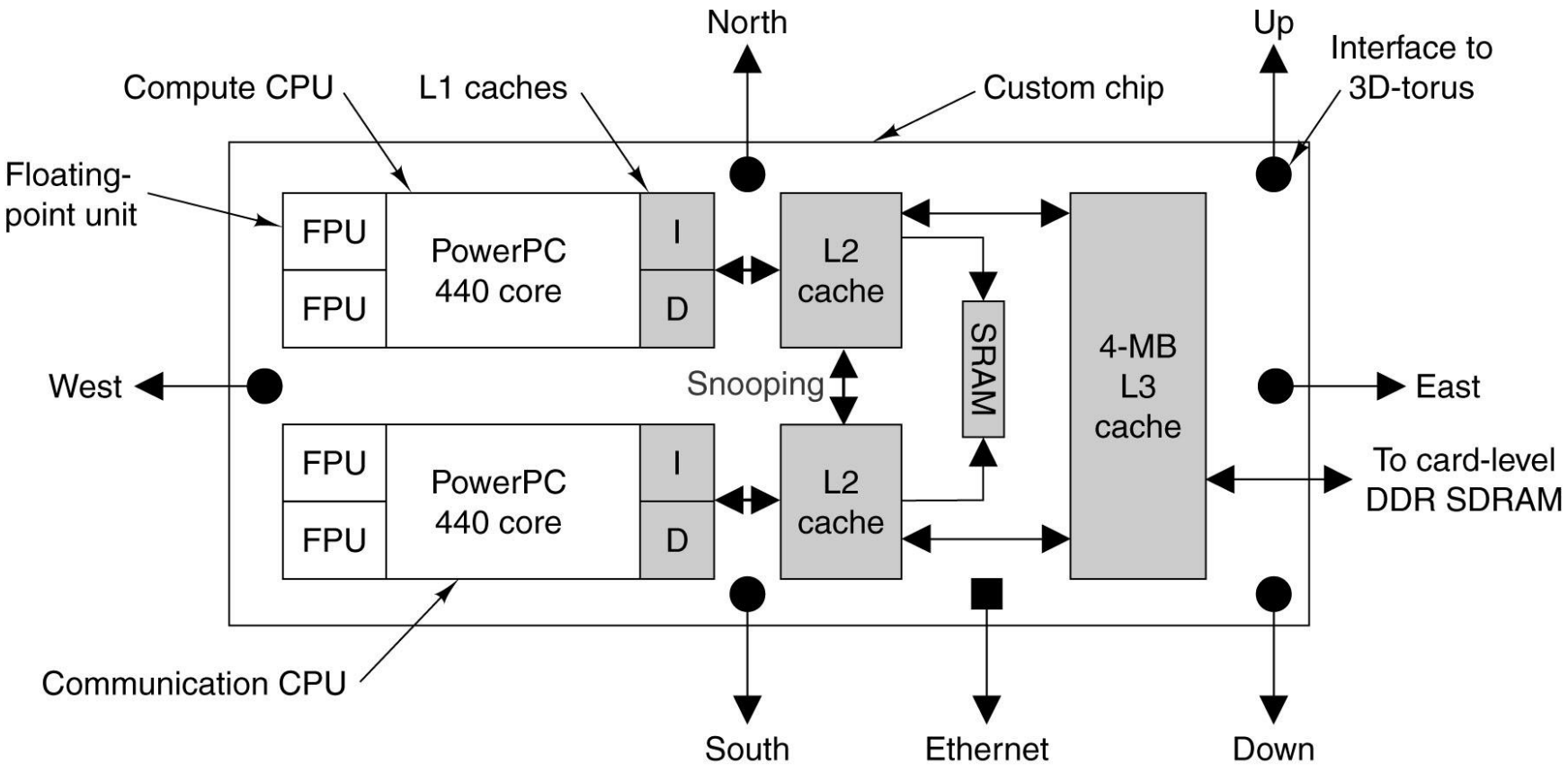
MPP – *Massively Parallel Processors* BlueGene

- Sistema projetado pela IBM, em 1999, para resolver problemas de computação intensiva relacionados à vida
- Primeiro: BlueGene/L
  - Início em 2001
  - Uma parceria da IBM com o Departamento de Energia dos EUA
  - Projetado não só ser o MPP mais rápido, mas também para ser o mais eficiente em termos de teraflops/dolar, teraflops/watt, teraflops/m^3
  - Um quarto (16384 processadores) disponível em nov/2004
  - http://domino.research.ibm.com/comm/research_projects.nsf/pages/bluegene.index.html

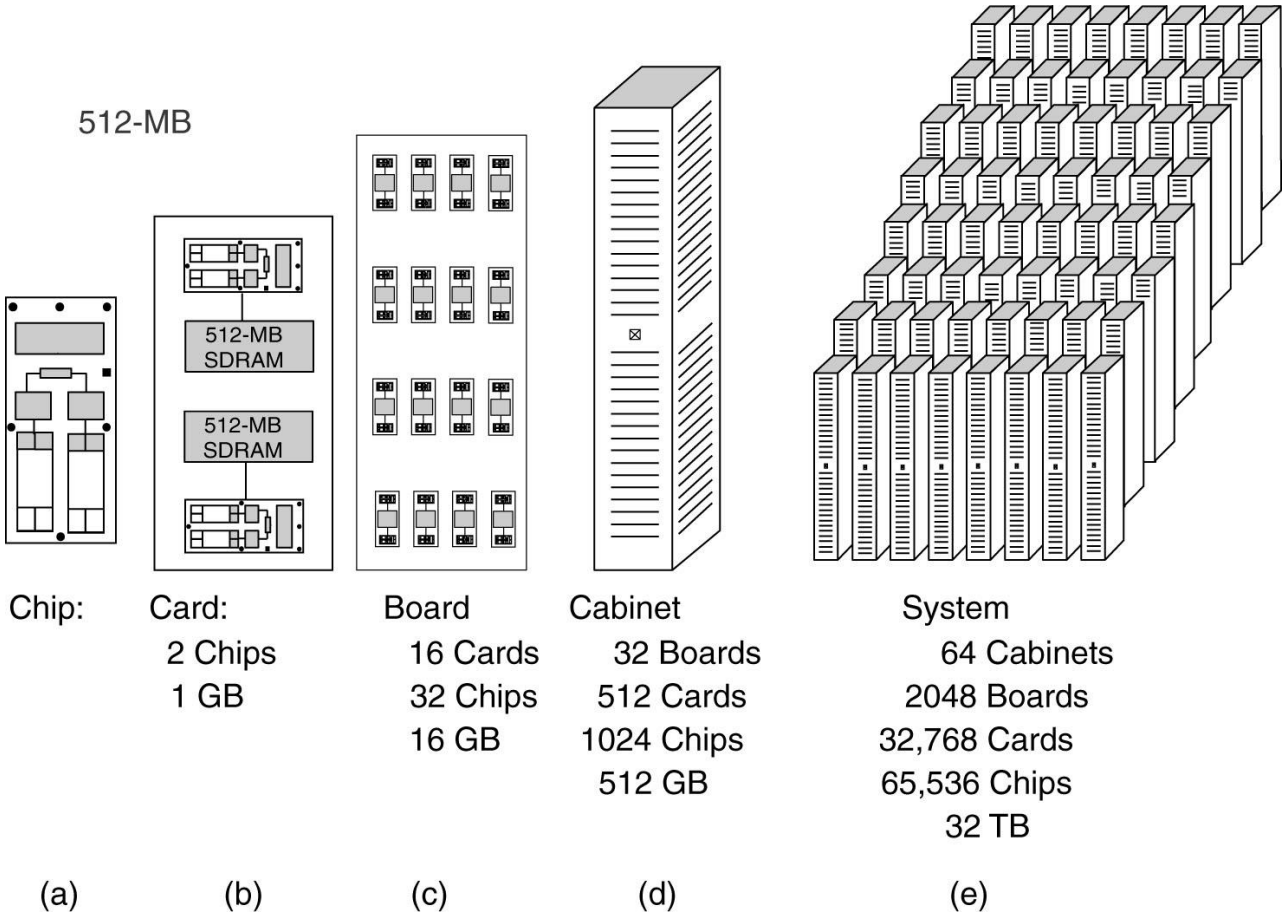
MPP – *Massively Parallel Processors* BlueGene

- O “coração” do BlueGene é um chip que contém dois cores do processador PowerPC 440, a 700 MHz
  - PowerPC 440 é um processador pipeline com 2 unidades superescalares

MPP – *Massively Parallel Processors* BlueGene



MPP – *Massively Parallel Processors* BlueGene



The BlueGene/L. (a) Chip. (b) Card. (c) Board. (d) Cabinet. (e) System.

MIMD - Clusters

- Arquitetura alternativa às SMPs
- Um grupo de computadores conectados
  - trabalham juntos como um recurso unificado
  - fornecem a ilusão de ser uma máquina paralela
  - cada computador é chamado de nó
- Benefícios:
  - Alto desempenho
  - Escalabilidade incremental
  - Alta disponibilidade
  - Boa relação custo/desempenho

MIMD - Clusters

- A unidade básica do cluster é um único computador, também chamado de nó.
- Os cluster podem aumentar de tamanho pela adição de outras máquinas.
- O cluster como um todo será mais poderoso quanto mais rápidos forem os seu computadores individualmente e quanto mais rápida for a rede de interconexão que os conecta.

MIMD - Clusters

- Um cluster típico possui:
  - Rede mais rápida e próxima do que uma rede local;
  - Protocolos de comunicação de baixa latência;
  - Conexão mais frouxa que um SMP

MIMD - Clusters

- Se você tiver dois ou mais computadores, existe uma grande chance de que, em um determinado instante, pelo menos um deles não esteja fazendo nada.
- E infelizmente, quando você realmente precisar de capacidade de processamento, toda aquele que estiver disponível provavelmente não será suficiente.
- A idéia por detrás do uso de clusters é espalhar as cargas entre todos os computadores disponíveis, usando ao máximo os recursos que estão livres nas outras máquinas.

MIMD - Clusters

- Além disso, o sistema operacional de um cluster deve fazer o melhor uso do hardware disponível em resposta às mudanças de condições da computação.
- Isto será um grande desafio se o cluster for composto de diferentes tipos de computador (um cluster “heterogêneo”) , se um grande número de máquinas deixar e entrar no cluster aleatoriamente e se as cargas não puderem ser previstas com antecipação.

MIMD - Clusters

- Dois tipos de clusters podem ser considerados:
  - Centralizados
    - Cluster de estações de trabalho ou PCs, instalados em um *rack* em uma sala
    - Normalmente são máquinas homogêneas e não possuem outros periféricos que não placas de rede e discos
    - Chamados de COW (*Cluster of Workstation*)

MIMD - Clusters

- Tipos de Clusters (cont.)
  - Descentralizados
    - Estações de trabalho ou PCs espalhados por prédios, campus, cidade, etc.
    - Muitos ficam ociosos por boa parte do dia
    - Normalmente são conectados por LAN e são heterogêneos
    - Também conhecido como *grid* (grades)

MIMD - Clusters

- Existem 3 tipos de clusters:
  - Tolerante à falhas
  - Balanceamento de Carga
  - Computação de Alto Desempenho
- Clusters Tolerantes à Falhas consistem de dois ou mais computadores conectados em rede com um software de monitoração (*heartbeat*) instalado entre os dois
  - Assim que uma máquina falhar, as outras não mais receberação as respostas das mensagens de *heartbeat* e tentam assumir o trabalho

MIMD - Clusters

- Cluster com Balanceamento de Carga utilizam o conceito de, por exemplo, quando um pedido chega para um servidor Web, o cluster verifica qual a máquina menos carregada e envia o pedido para esta máquina.
- Na realidade na maioria das vezes um cluster com balanceamento de carga é também um cluster tolerante à falha com a funcionalidade extra de balanceamento de carga e um número maior de nós.

MIMD - Clusters

- A última variação de cluster é o de alto desempenho: as máquinas são configuradas especialmente para oferecer o maior desempenho possível.
- Estes tipos de clusters também tem algumas funcionalidades para balanceamento de carga, já que eles tentam espalhar os processos por máquinas diferentes para obter maior desempenho.
- Mas o que ocorre normalmente é que um processo é paralelizado e que as rotinas (ou threads) é que podem executar em paralelo em máquinas diferentes.

MIMD - Clusters

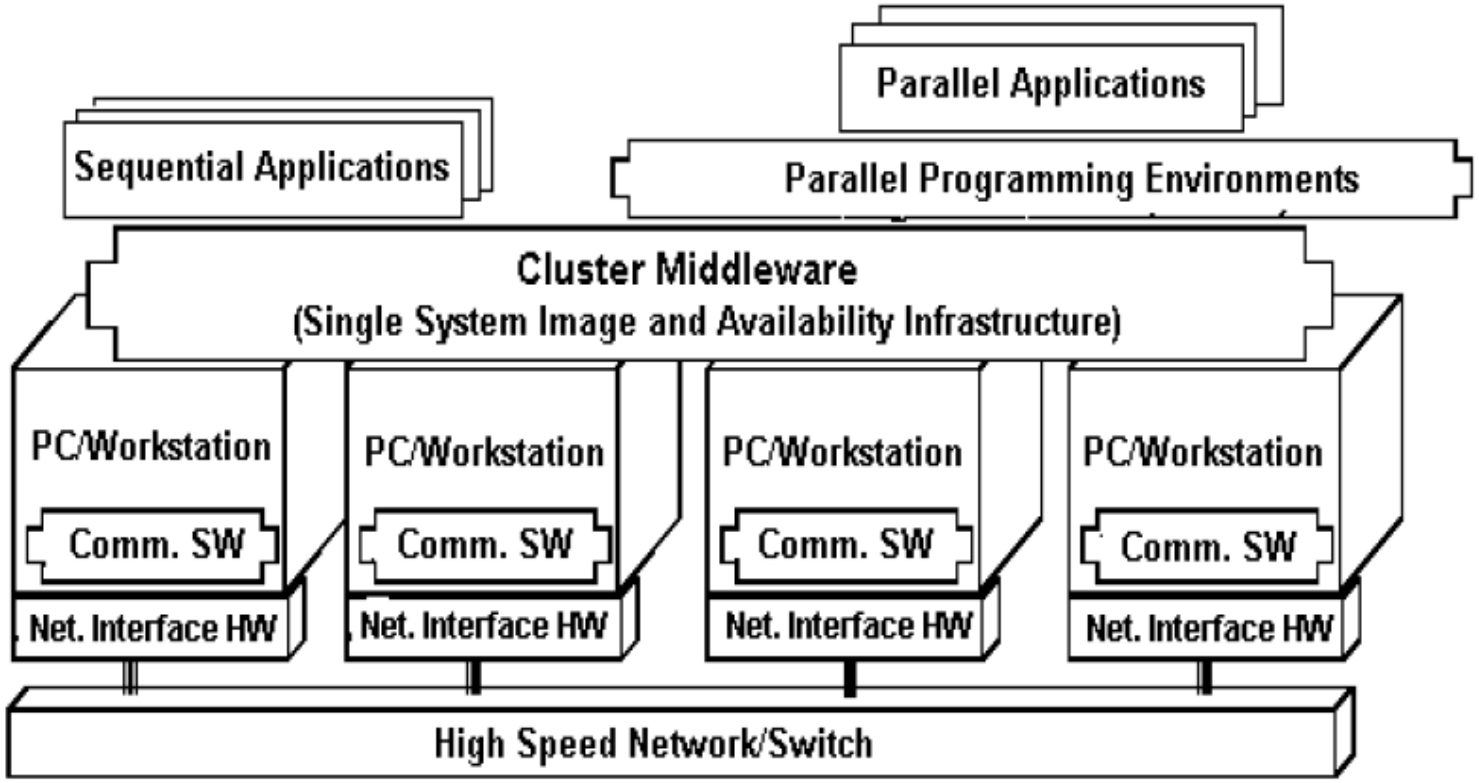
- Os supercomputadores tradicionais foram construídos por um pequeno número de fabricantes, com um alto orçamento destinado ao projeto.
- Muitas universidades não podem arcar com os custos de um supercomputador, então o uso de clusters se torna um alternativa interessante.
- Com o uso de hardware mais barato e disponível no mercado, sistemas com desempenho similar aos supercomputadores podem ser construídos.

MIMD - Clusters

- O desempenho dos componentes dos PCs e estações de trabalho é próximo do desempenho daqueles usados nos supercomputadores:
  - Microprocessadores
  - Redes de Interconexão
  - Sistemas Operacionais
  - Ambientes de Programação
  - Aplicações
- A taxa de melhoria de desempenho dos componentes ao longo do tempo é muito alta.

MIMD - Clusters

- Arquitetura de um Cluster



MIMD - Clusters

- Maiores desafios:
  - Escalabilidade (física e da aplicação)
  - Disponibilidade (gerenciamento de falhas)
  - Imagem única do sistema (faz o usuário perceber que é um único sistema)
  - Comunicação rápida (redes e protocolos de comunicação)
  - Balanceamento de carga (CPU, Rede, Memória, Discos)
  - Segurança e Encriptação (clusters de clusters)
  - Programabilidade (administração e controle)
  - Aplicabilidade (aplicações voltadas para o cluster)

MIMD - Clusters

- Ambientes de Programação
  - Threads (PCs, SMPs, NOW)
    - POSIX Threads
    - Java Threads
  - MPI
  - PVM
  - Software DSMs
    - SHMEM da Cray/SGI

MIMD - Clusters

- Aplicações
  - Podem ser sequenciais quando se beneficiam do balanceamento de carga
  - Podem ser paralelas/distribuídas quando se utilizam dos ambientes de programação existentes
  - Servidores Web
  - Data-Mining