Programação concorrente e Distribuída

Aula 2: Arquitetura de Máquinas Paralelas e Distribuídas

Prof: Álvaro L. Fazenda
(alvaro.fazenda@unifesp.br)

Classificação

- Importante para análise das possibilidades de arquiteturas e seus efeitos
- Muitas formas possíveis de classificação:
- Clássicas:
  - **Fluxo de instruções e fluxo de dados (Taxonomia de Flynn, 1972)**
  - Modelo de memória
  - Outros tipos:
    - Dispersão dos processadores
    - Estrutura de interconexão
    - Sincronismo
    - Modernas

Taxonomia de Flynn

- (“Some Computer organization and their effectiveness” - M. Flynn - IEEE Transactions on Computers, 1972)
- Mais aceita de forma universal
- O conceito central foi classificar arquiteturas pelo número de fluxos (stream) de dados e instruções
  - “Stream in this context simply means a sequence of items (instructions or data) as executed or operated on by a processor”

Taxonomia de Flynn (cont.)

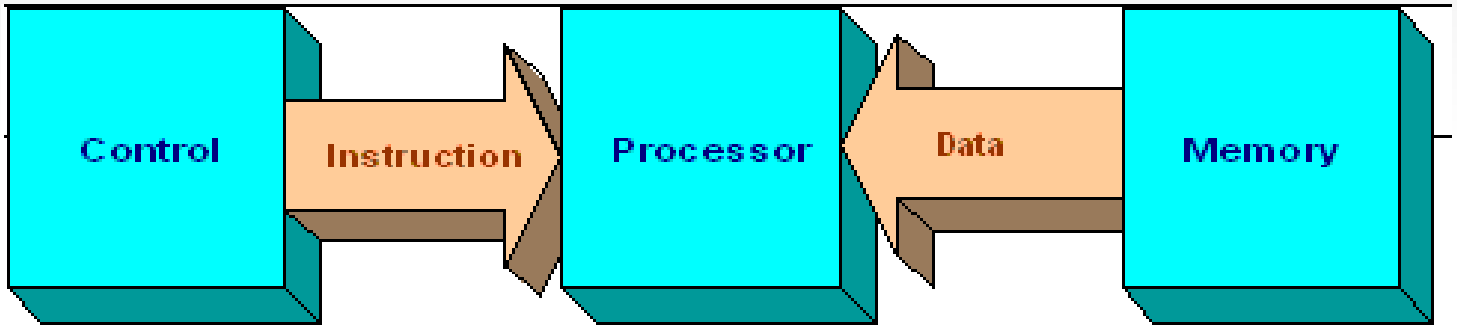
- (Fluxo de Instruções)x(Fluxo de Dados)

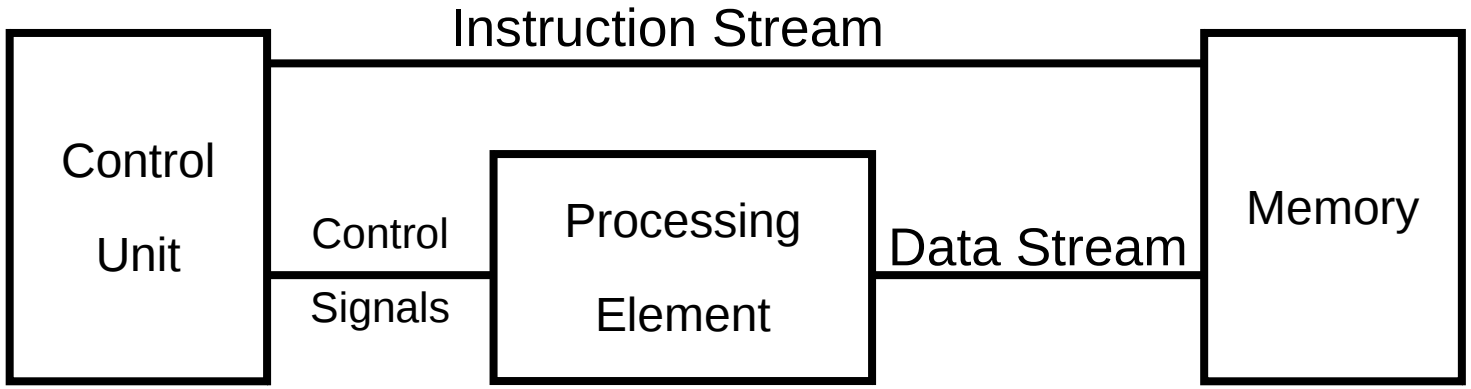
  - 4 casos possíveis:

    - *SISD (Single Instruction, Single Data)*
    - *SIMD (Single Instruction, Multiple Data)*
    - *MISD (Multiple Instructions, Single Data)*
    - *MIMD (Multiple Instructions, Multiple Data)*

SISD

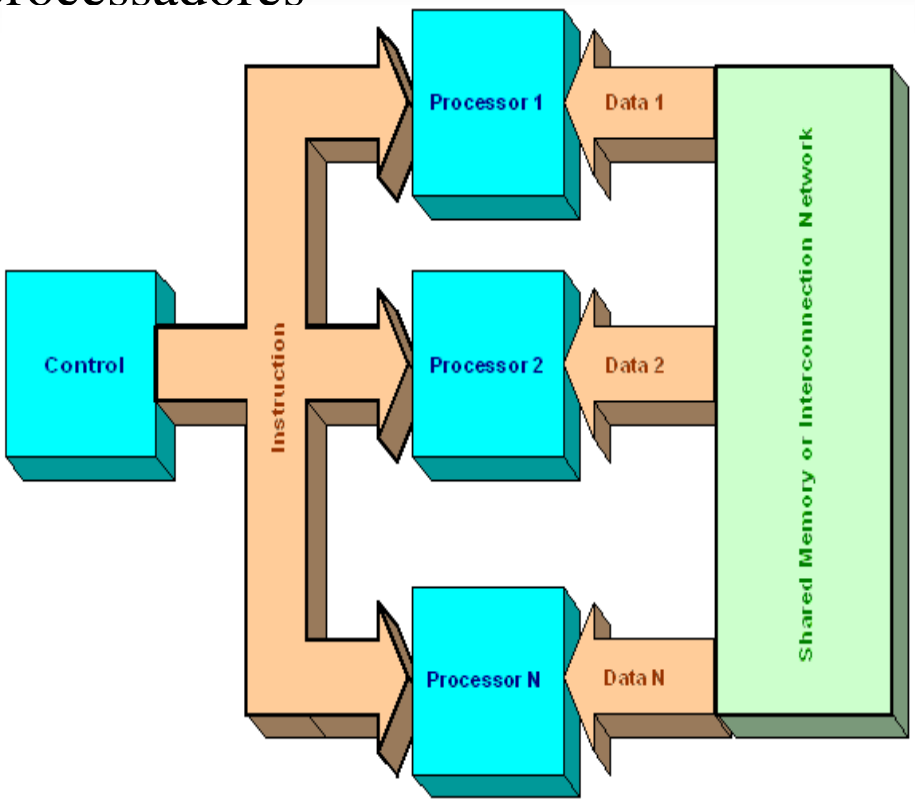
- Máquina convencional
  - Arquitetura clássica de Von Neumann





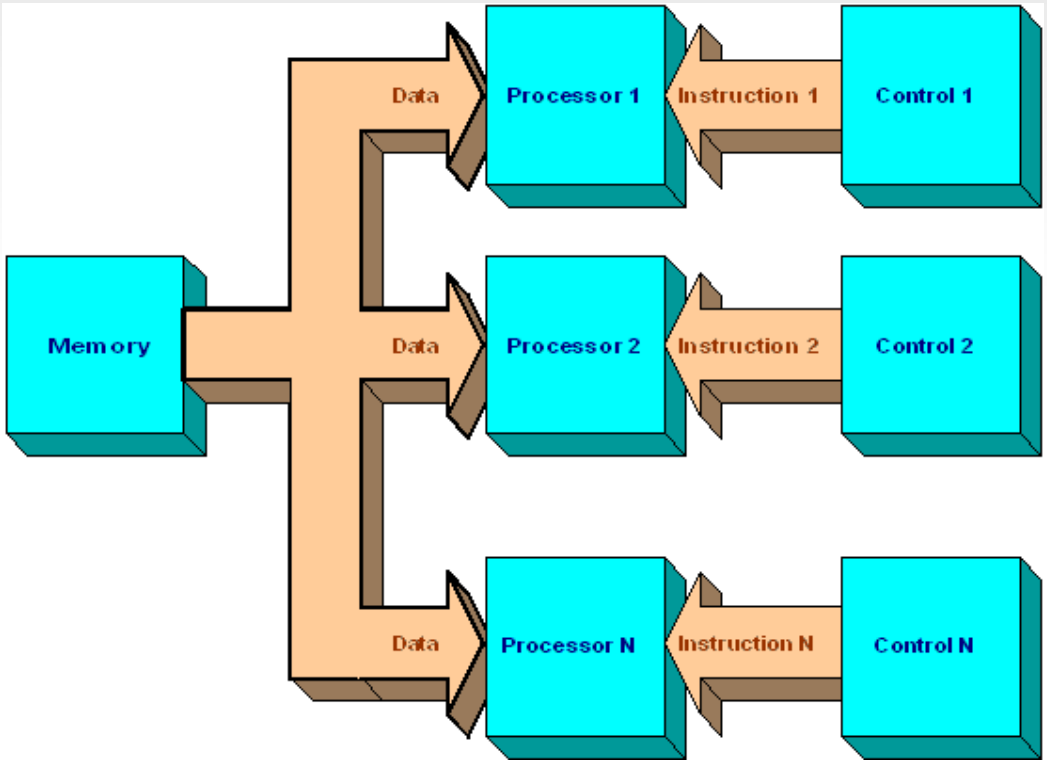
SIMD

- Mesma instrução executada simultaneamente sobre diversos conjuntos distintos de dados
  - Cada Processador atua exclusivamente sobre sua memória
  - Unidade de controle UC central
    - Envia a mesma instrução decodificada (sinais) para todos os processadores
    - Obtém instruções e dados da própria memória (exclusiva) e/ou das memórias dos diversos processadores
  - Exemplos:
    - Máquinas *array*
    - Máquinas vetoriais
    - Atualmente GPUs



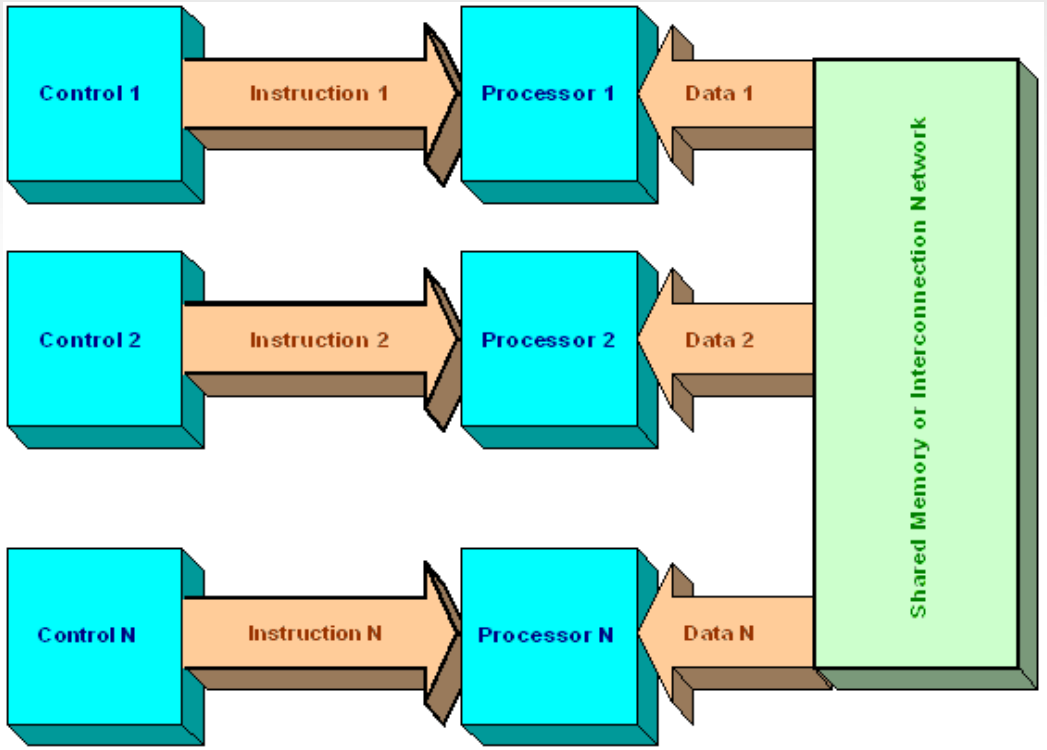
MISD

- Utilidade prática duvidosa
  - Máquinas sistólicas?



MIMD

- Cada processador pode executar seu próprio fluxo de instruções
- Troca de dados através de memória e/ou envio de mensagens através de rede
  - Clássicas máquinas paralelas
    - CM-2, MasPar, etc
  - Atuais Multicores baseados em x86



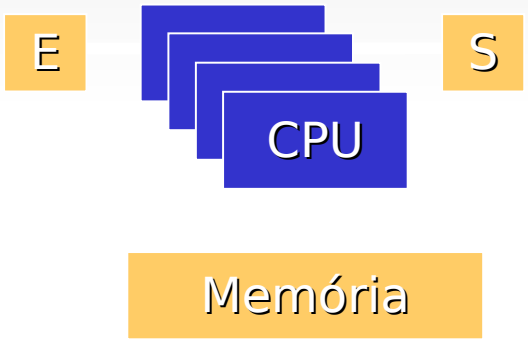
Classificação por modelo de memória

- Dois principais tipos:
  - Memória compartilhada (*shared/central memory*)
  - Memória Distribuída (*distributed memory*)

Arq. Memória Compartilhada

- Conhecido como: Multiprocessadores

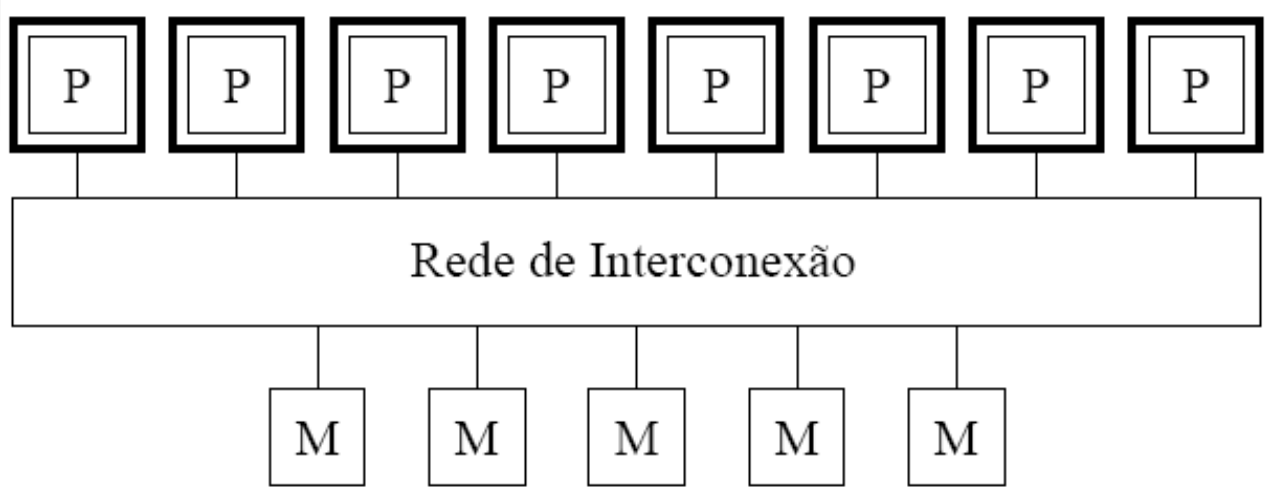
- Todos os processadores podem acessar uma área comum de memória
  - Espaço único de endereçamento
  - Operações de *LOAD/STORE*



- Vantagem:
  - Facilidade de programação e para troca de dados entre processadores (PThreads, OpenMP)
  - Comum, atualmente, em Processadores *Multicore*
- Desvantagem:
  - *Hardware* ainda caro para mais de 8 processadores
    - Tendência ao barateamento com o avanço do multicore

Multiprocessadores

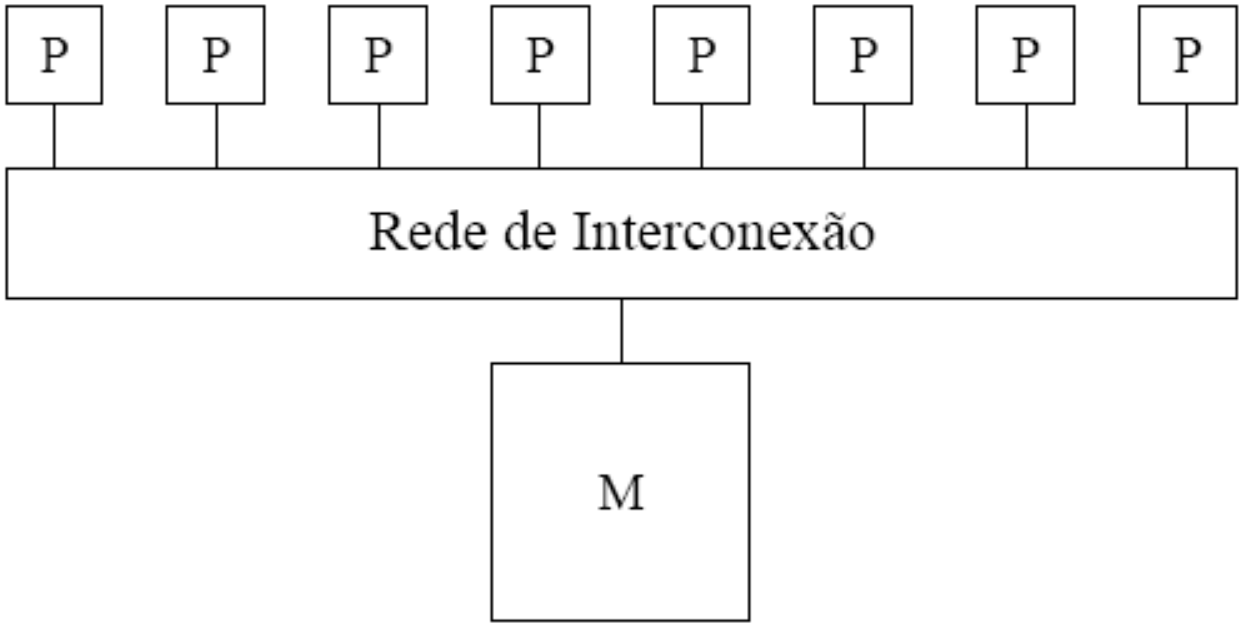
- Máquina paralela construída, geralmente, a partir da replicação de processadores de uma arquitetura convencional
- Todos processadores (*P*) acessam memórias compartilhadas (*M*) através de uma infra-estrutura de comunicação
  - Possui apenas um espaço de endereçamento
  - Comunicação entre processos de forma bastante eficiente (load e store)



- Em relação ao tipo de acesso às memórias do sistema, multiprocessadores podem ser sub-classificados como:
  - UMA
  - NUMA (e NCC-NUMA, CC-NUMA)
  - COMA

Acesso Uniforme à Memória
(*Uniform Memory Access* - UMA)

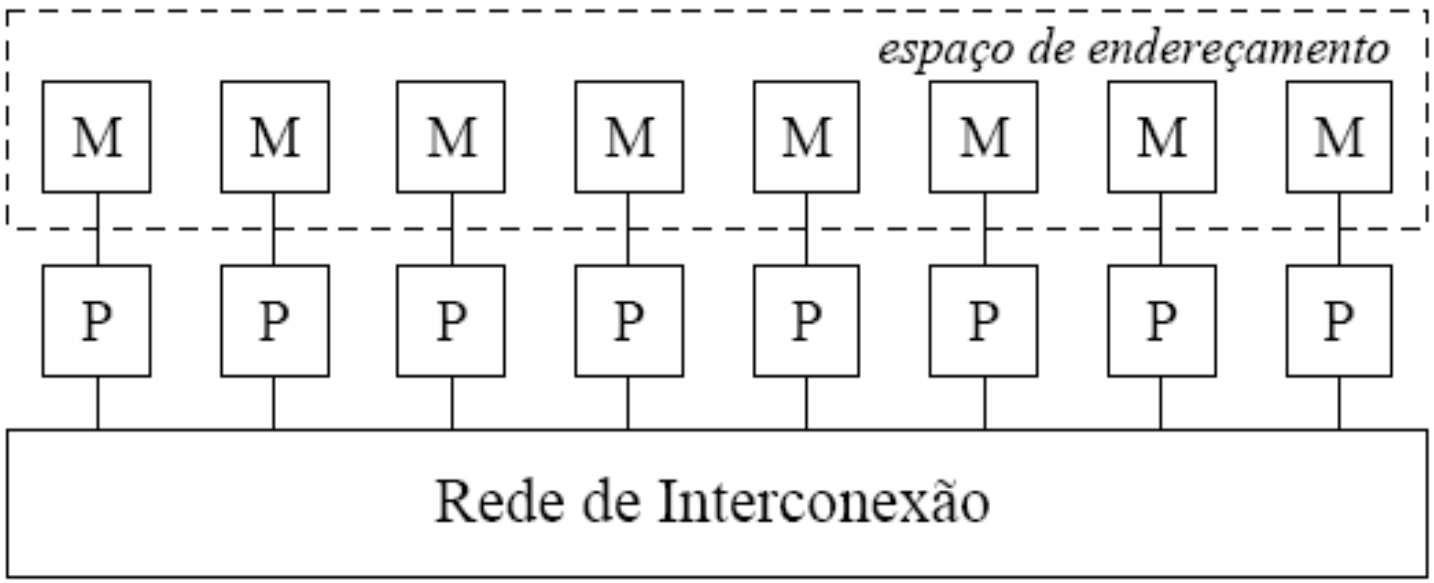
- Memória centralizada
  - Encontra-se à mesma distância de todos processadores
- Latência de acesso à memória
  - Igual para todos processadores
- Infra-estrutura de comunicação
  - Barramento é a mais usada → suporta apenas uma transação por vez
  - Outras infra-estruturas também se enquadram nesta categoria, se mantiverem uniforme o tempo de acesso à memória



Acesso Não Uniforme à Memória

(*Non-Uniform Memory Access* - NUMA)

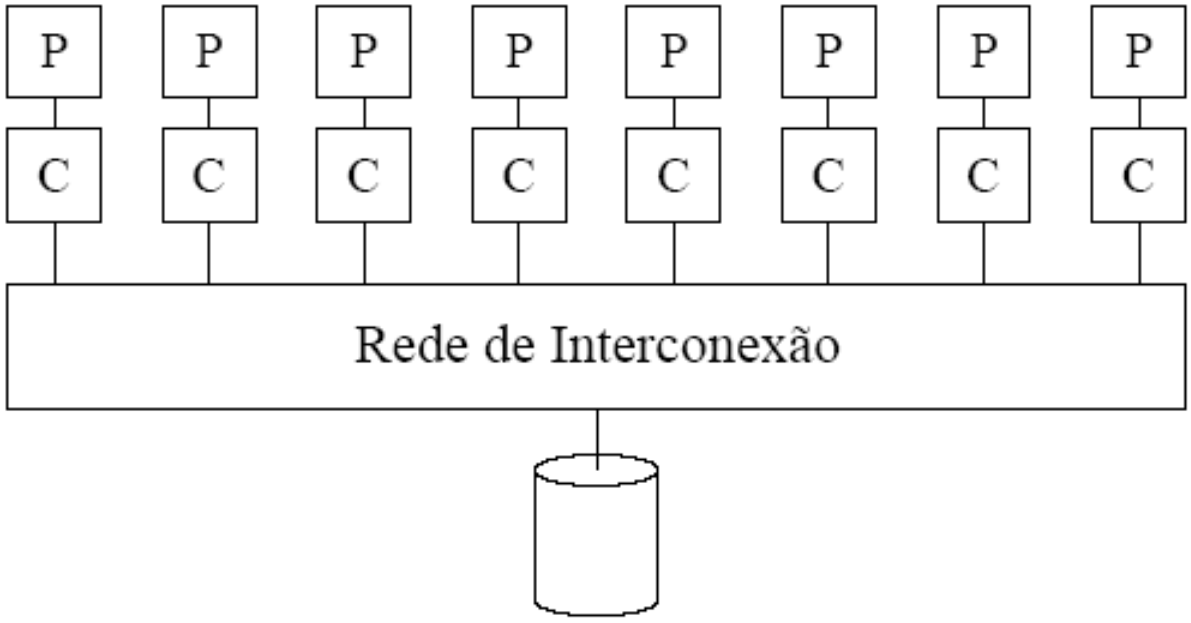
- Memória local distribuída
- Espaço de endereçamento único
  - Implementada com múltiplos módulos associados a cada processador
- Comunicação processador-memórias não locais através da infra-estrutura de comunicação
- Tempo de acesso à memória local < tempo de acesso às demais → Acesso não uniforme → Distância das memórias variável → depende do endereço



Arquiteturas de Memória Somente com *Cache*

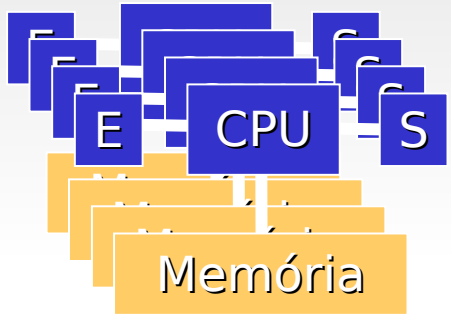
(*Cache-Only Memory Architecture* - COMA)

- Memórias locais estão estruturadas como memórias *cache*
  - São chamadas de COMA *caches*
    - Mais capacidade que uma *cache* tradicional
- Arquiteturas COMA têm suporte de hardware para manter coerência entre *cache* e memória principal através dos múltiplos nós
  - Geralmente reduz tempo global para pegar informações



Arq. Memória Distribuída

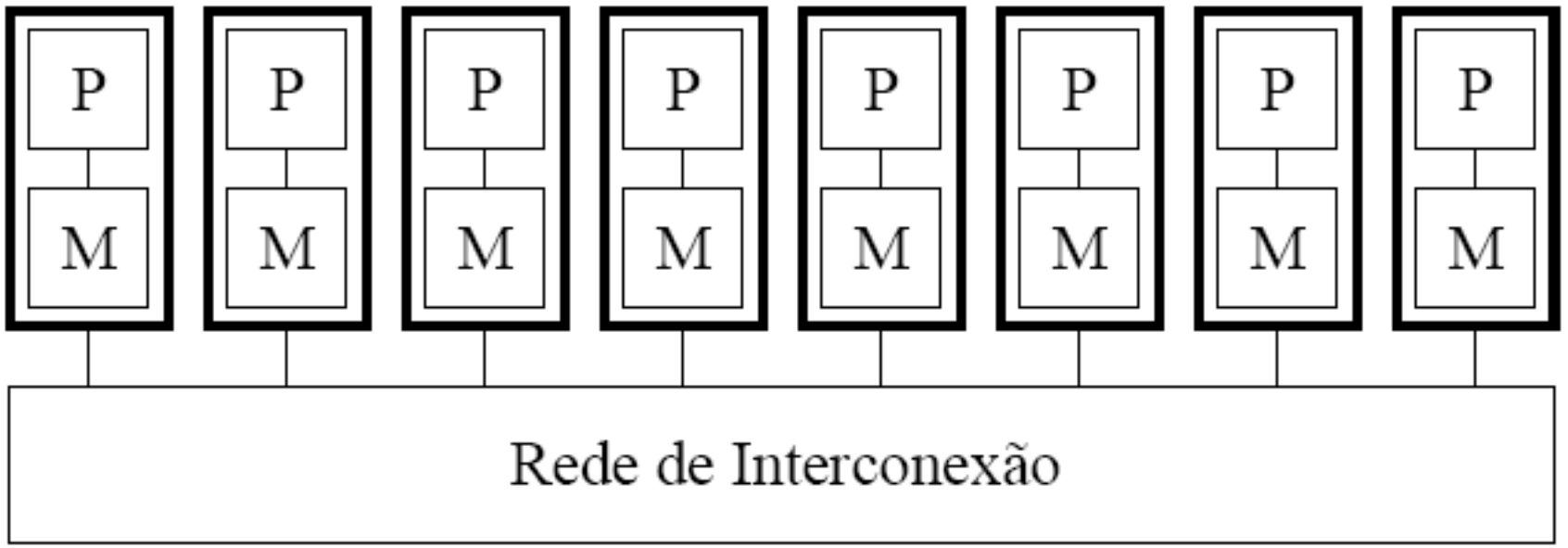
- Conhecidos por Multicomputadores
- Cada processador tem sua área de dados local
- Interconexão através de rede de dados
  - Troca de mensagens entre processos
    - Operações de *SEND/RECEIVE*
- Vantagens:
  - Fácil implementação de um Cluster
  - Fácil expansão (porém limitada pelo tipo de rede)
- Desvantagens:
  - Granularidade de Comunicação/Computação é crítica para desempenho
    - Rede pode limitar ganhos de desempenho
  - Programação, geralmente, mais complexa



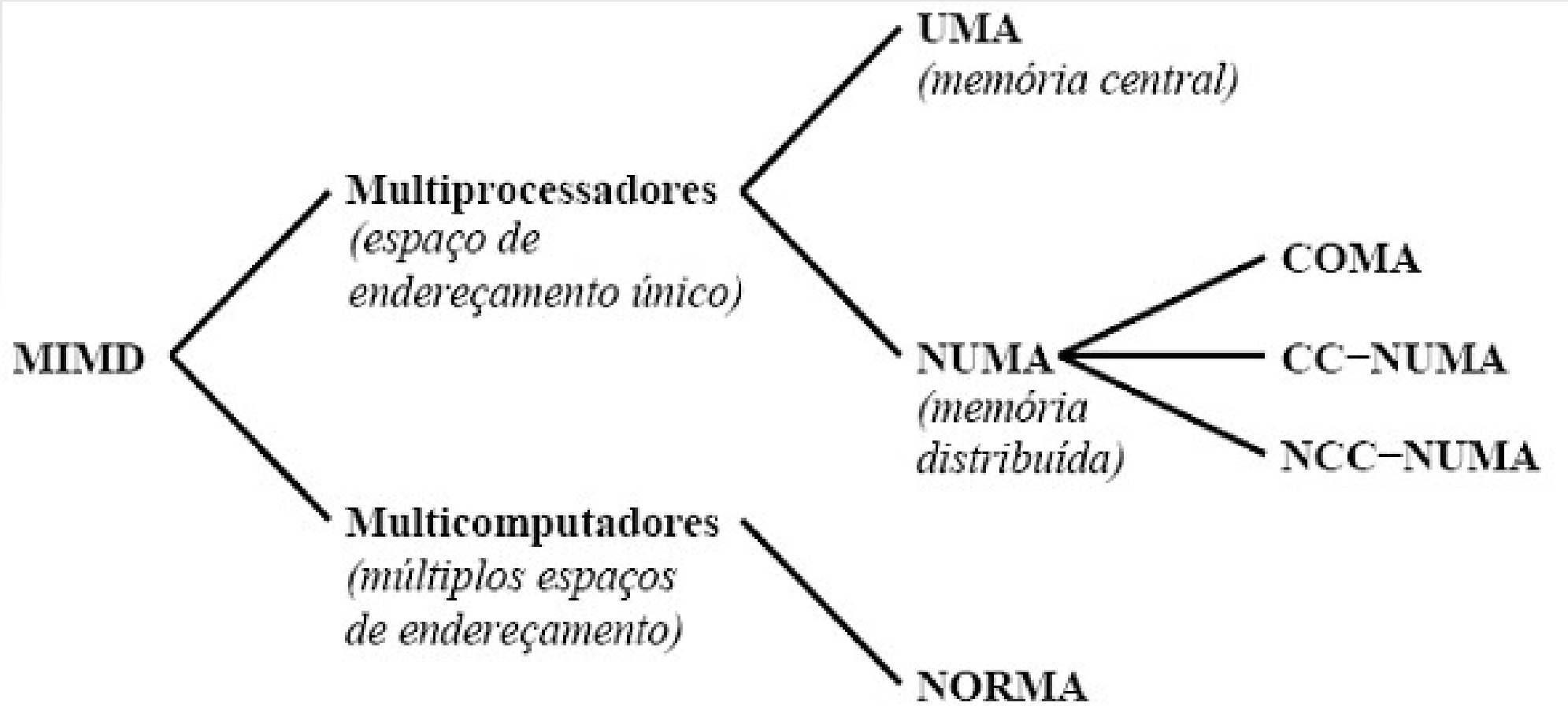
Sem Acesso à Memória Remota

(*Non-Remote Memory Access* - NORMA)

- Multicomputadores são classificados como NORMA
  - Cada nó só consegue endereçar sua memória local



Visão Geral da Classificação Segundo Modelo de Memória



Classificação por dispersão dos processadores

- Sistemas com dispersão geográfica
  - Sistemas distribuídos
    - *Cluster* e/ou rede de computadores
    - Multicomputadores
- Sistemas confinados
  - Máquinas paralelas
    - Multiprocessadores

Classificação por sincronismo

- Síncronos
  - Processadores operam sincronizadamente sob o controle de um único relógio global comum

- Assíncronos
  - Ausência completa de base de tempo comum a todos os processadores
    - Programador deve indicar explicitamente os pontos de sincronização

Classificações Modernas

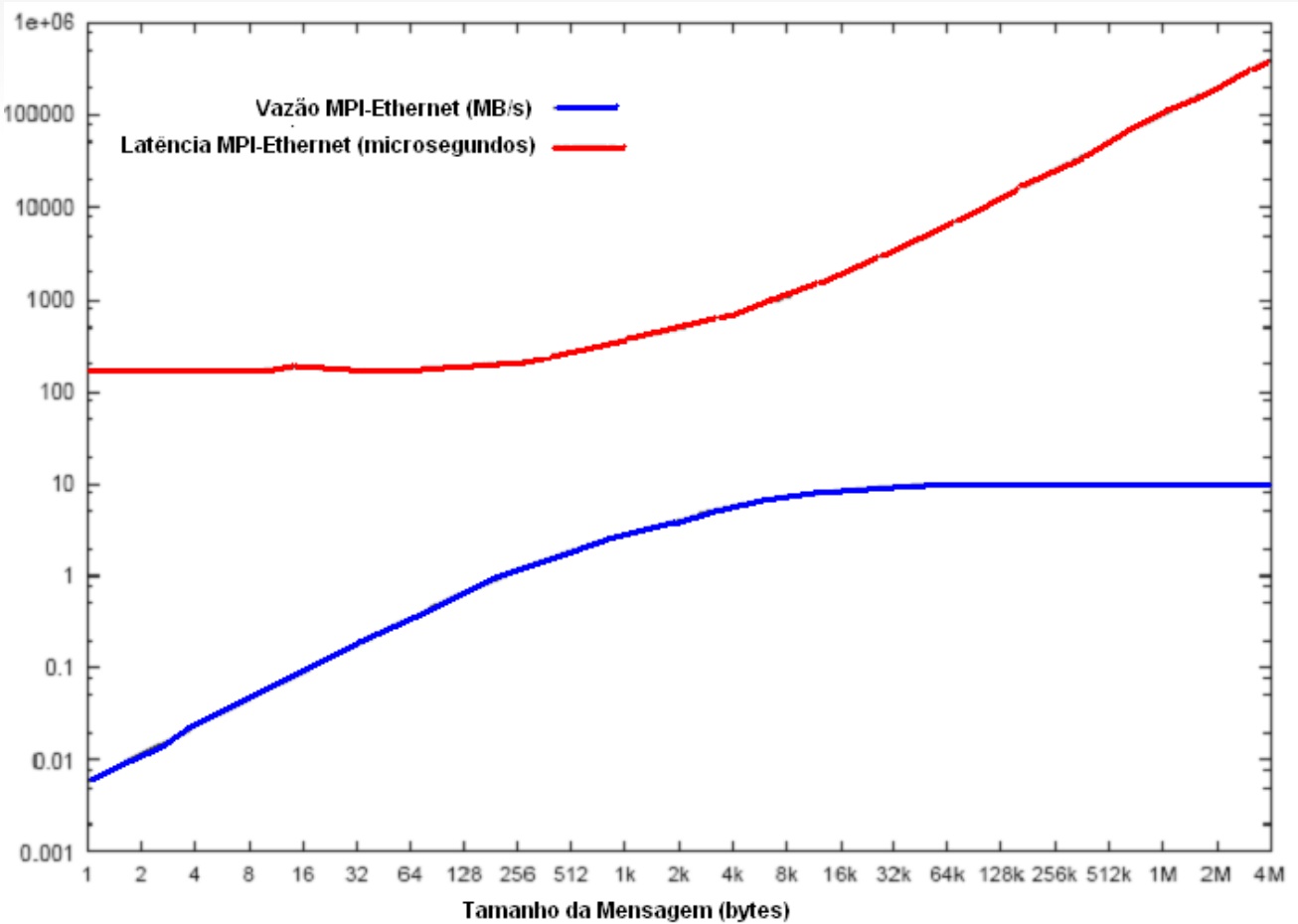
- Subdivisão dos diversos tipos de arquitetura em:
  - **PVP - Parallel Vector Processor**
  - **SMP - Symmetric Multiprocessor**
  - **MPP - Massively Parallel Processors**
  - **NOW - Network of Workstations**
    - **COW – Cluster of Workstations**

Redes para interconexão

- Todo computador paralelo necessita de uma rede de interconexão
  - Comunicação entre os seus diversos recursos de processamento, armazenamento e entrada/saída.
  - Aspectos que devem ser considerados:
    - latência (tempo de trânsito de uma mensagem pela rede de comunicação, inclui tempo de empacotar e desempacotar dados mais tempo de envio propriamente dito)
    - Vazão (Expressa a capacidade da rede de “bombear” dados entre dois pontos. Unidade: Quantidade de dados por unidade de tempo, exemplo: 10 MBytes/segundo (10MB/s))
    - conectividade (quantidade de vizinhos que cada processador possui)
    - confiabilidade (conseguida, por exemplo, através de caminhos redundantes)
    - escalabilidade: possibilidade de acréscimo de dispositivos sem a necessidade de alteração das características da rede

Desempenho da Rede de Interconexão - Exemplo

- Latência de 1 mensagem de 1 byte entre máquinas rodando GNU/Linux ligadas por Fast-Ethernet (100 Mbit/s) é de aproximadamente 150 µs
- A melhor vazão, obtida com uma mensagem de aproximadamente 64 KB, é em torno de 10 MB/s. Próximo do limite teórico (12,5 MB = 100 Mbits/s)

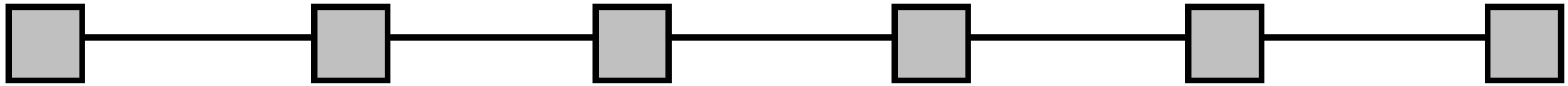


Tipos de redes

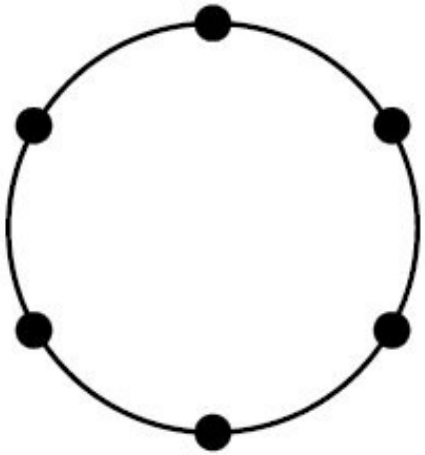
- Redes estáticas
  - ligações fixas entre os componentes
    - Diversas topologias possíveis
- Redes dinâmicas
  - conexões são feitas sob demanda
    - não existem ligações fixas entre os componentes
  - bloqueantes ou não bloqueantes
  - três tipos básicos:
    - Barramento, Matriz de chaveamento, Rede multinível

Redes estáticas

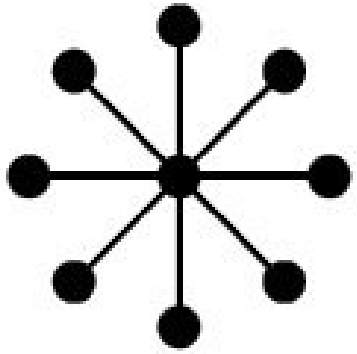
- Array linear
  - Sem caminhos alternativos



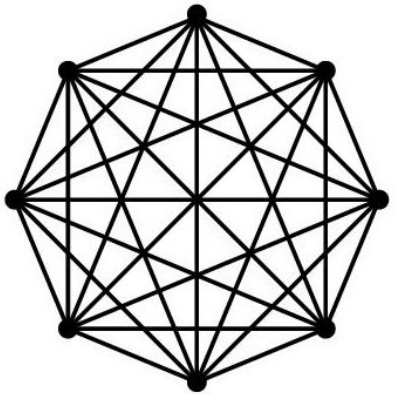
- Anel



- Estrela

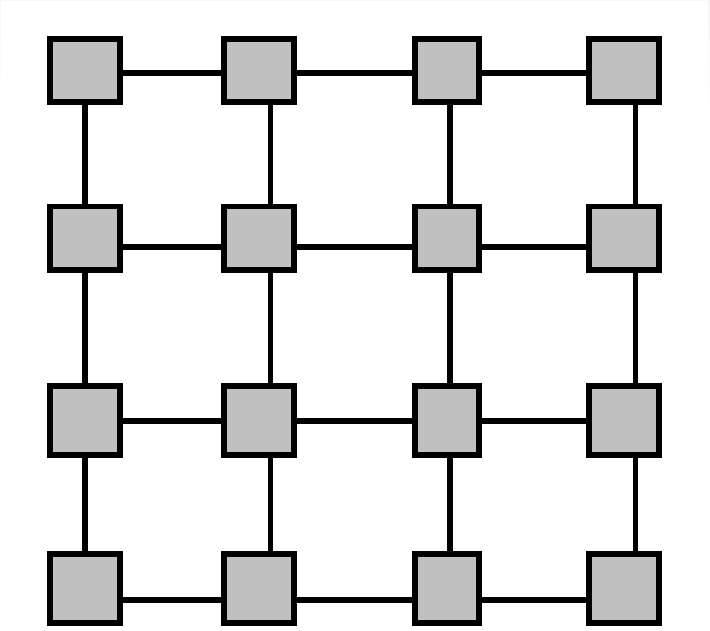


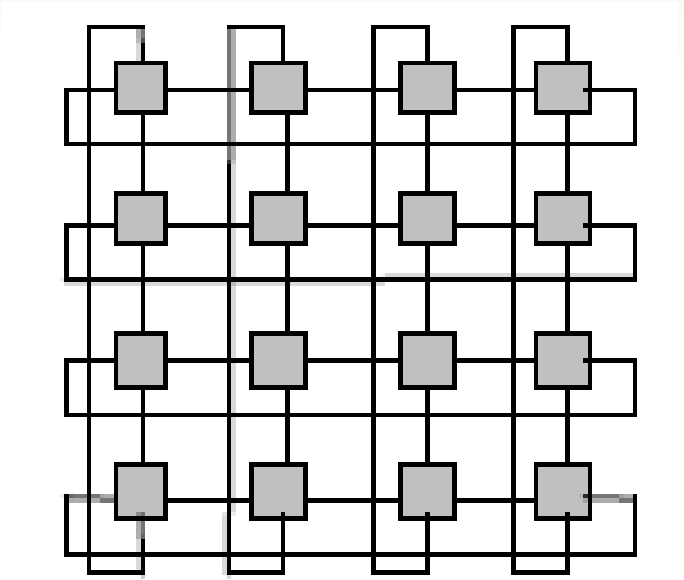
- Totalmente conectada



Redes estáticas
Malha

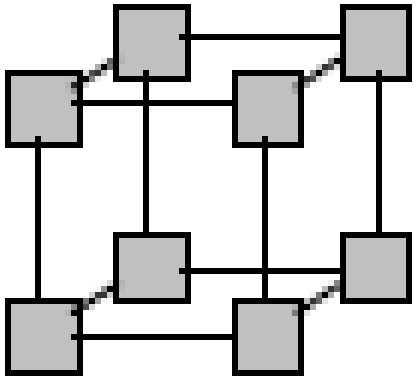
- canais de comunicação entre os processadores vizinhos





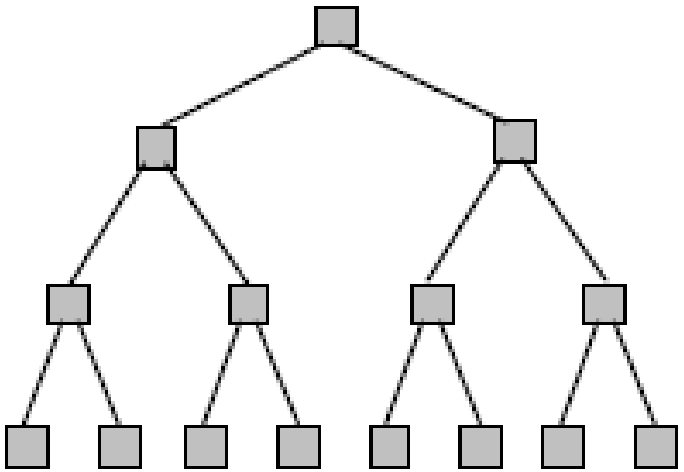
Redes estáticas
Hipercubo

- tamanhos do hipercubo são definidos por potências de 2
- escalabilidade restrita a potências de 2



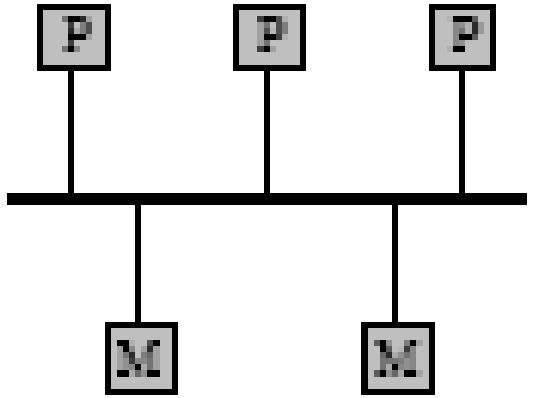
Redes estáticas
Arvore

- diâmetro cresce de forma linear com a altura h
- grau de nó máximo 3
- sem caminhos alternativos
- nó raiz é um gargalo



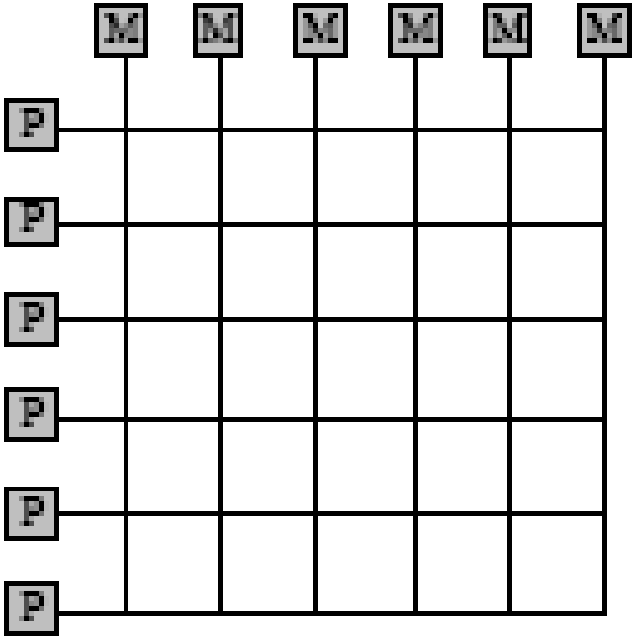
Redes Dinâmicas
Barramento

- todos os processadores estão conectados em um único barramento compartilhado
  - necessidade de aguardar que o barramento esteja livre
  - Existência de Colisões
  - viável para um pequeno número de processadores e/ou algoritmos com pouca comunicação



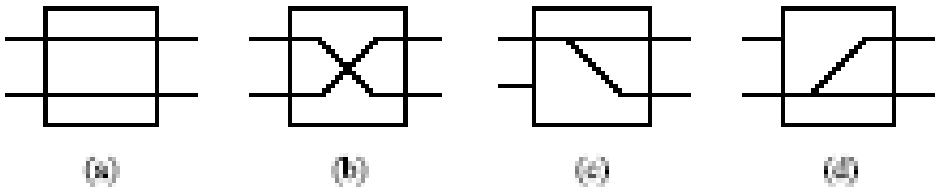
Redes dinâmicas - Matriz de Chaveamento (crossbar)

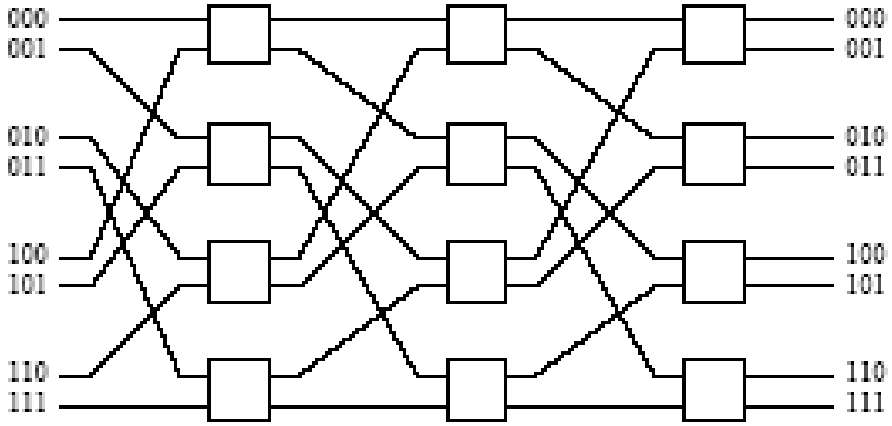
- alternativa não bloqueante de interconexão
  - Escalabilidade limitada por aspectos econômicos



Redes dinâmicas - Redes Multinível

- Conexões através da ligação de pequenas matrizes de chaveamento
  - tenta reduzir a probabilidade de conflitos entre conexões de diferentes pares
  - As matrizes chaveadoras presentes na maioria das redes multinível têm tamanho 2×2 e permitem no mínimo 2 e, na maioria das vezes, 4 posições de chaveamento





Redes dinâmicas – Roteamento de Mensagens

- Roteamento é o procedimento de condução de uma mensagem, através de nós intermediários, até seu destino
- não possuem ligações diretas entre todos os componentes de um sistema
  - Mensagem pode precisar trafegar por nós intermediários para chegar ao seu destino
- Baixo custo

Redes dinâmicas - Roteamento de Mensagens - Chaveamentos

- Chaveamento de circuito
  - Estabelece-se inicialmente um caminho fixo da origem ao destino, e só depois são enviadas todas as mensagens
    - usada por poucas máquinas paralelas, pois a comunicação entre dois nós tem pouca duração
- Chaveamento de pacotes
  - cada mensagem decide, a cada nó, qual a direção que irá seguir na rede
    - caminho dinâmico
  - elimina o custo inicial de estabelecimento de circuito, mas embute um custo adicional para o roteamento de cada mensagem em cada um dos nós visitados
  - reagem mais rapidamente a congestionamentos e falhas na rede

Tecnologia de rede (I)

- Gigabit Ethernet
  - extensão dos padrões 10/100 Mbps Ethernet
  - Atende a necessidade criada pelo aumento da vazão
  - baixo custo
  - Usada em aproximadamente 50% das máquinas do TOP500 até junho/2011
  - Usada em 7.6% das máquinas do TOP500 em junho/2016

Tecnologia de rede (II)

- Myrinet
  - Desenvolvida pela Myricom
  - portas e interfaces *full-duplex* alcançando 1.28 Gb/s para cada link;
  - controle de fluxo, de erro, e monitoramento contínuo dos links;
  - baixa latência, *switches crossbar* com monitoramento para aplicação de alta disponibilidade;
  - suporte a qualquer configuração de topologia;
  - latência: 13 a 21 us

Tecnologia de rede (III)

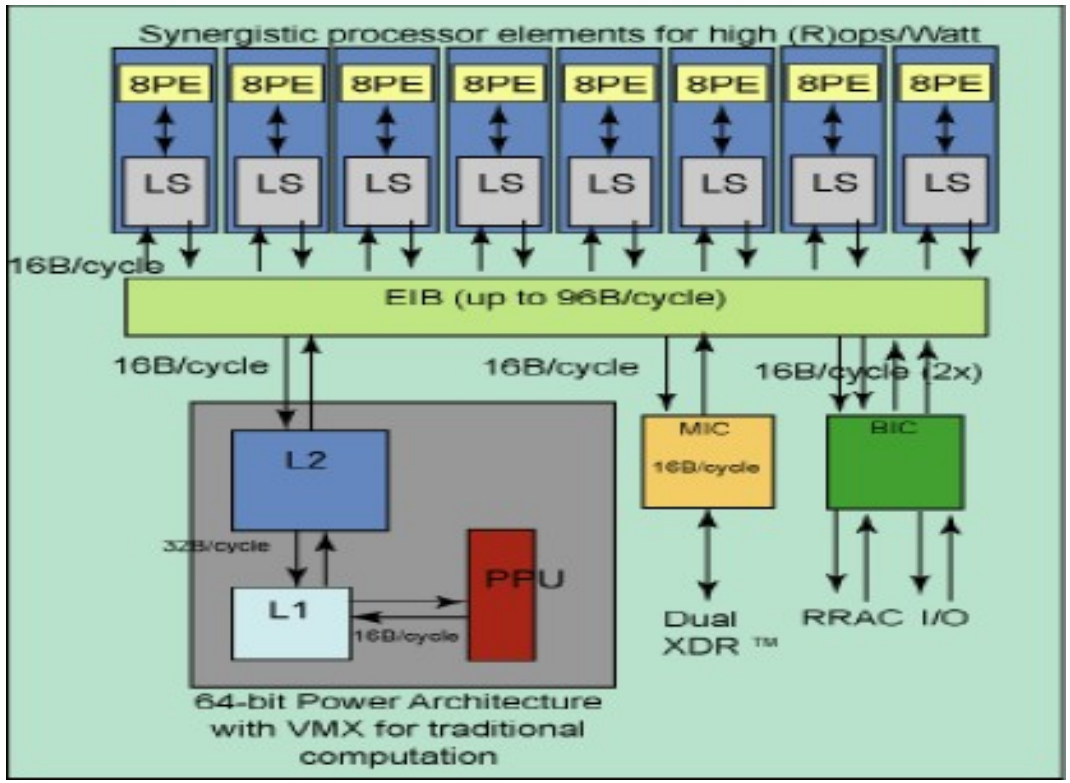
- Arquitetura Infiniband, ou IBA (*Infiniband Architecture)*
  - Padronização: Infiniband Trade Association, 2002
  - Usada em, aproximadamente, 40% dos supercomputadores do TOP500 até junho/2011
  - Usada em 38,8% das máquinas do TOP500 em junho/2016
  - Não necessita de recursos do processador
  - surgiu devido à necessidade de se melhorar o desempenho dos dispositivos de E/S e das comunicações, devido ao aumento da capacidade de processamento.
  - utiliza uma estrutura hierárquica, com comunicação do tipo ponto-a-ponto.
  - Vantagens: baixa latência e boa vazão (1,3 us e 2,5Gb/s até 10Gb/s)

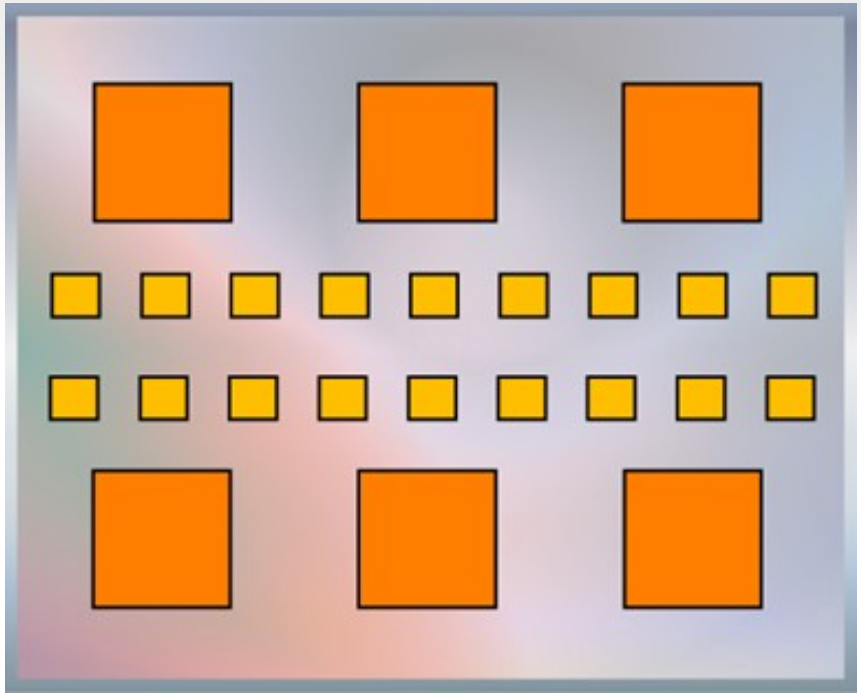
Máquinas híbridas implicam em paralelismo de múltiplos níveis

- Em todos os níveis, as arquiteturas provêm paralelismo importante.
- Linguagens de programação / algoritmos não são concebidos para aproveitar este paralelismo
  - Algoritmos: o modelo de referência é a máquina de Turing...
  - Modelos paralelos não capturam todos os parâmetros ou são impraticáveis
    - PRAM (memória compartilhada infinita)
    - BSP / CSM (fortemente síncrono)
    - LogP (poucos resultados)
- Só se sabe explorar paralelismo “trivial”
  - Mestre/escravo, Task Farm, ...

Nível 1 – Multicores ou manycores

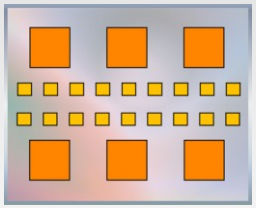
- **Composto de cores homogêneos ou heterogêneos**

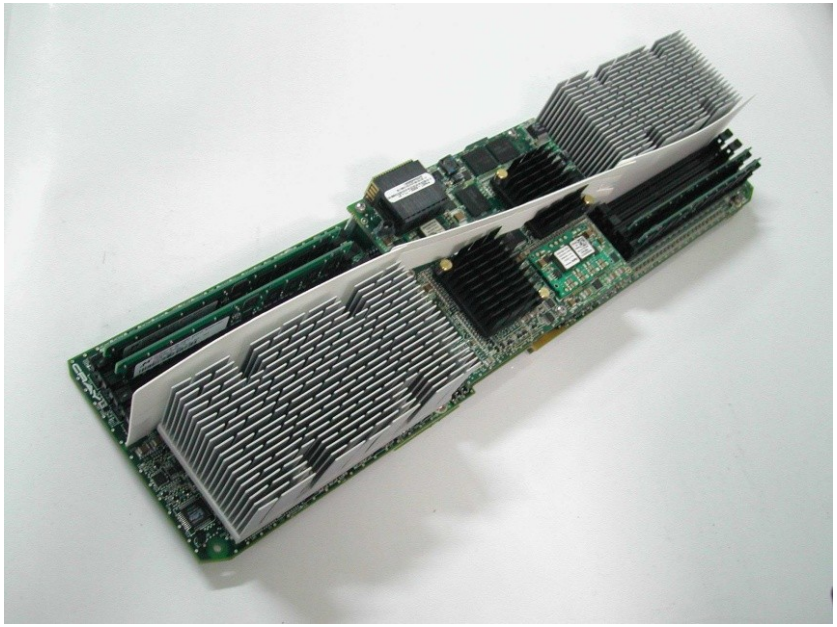


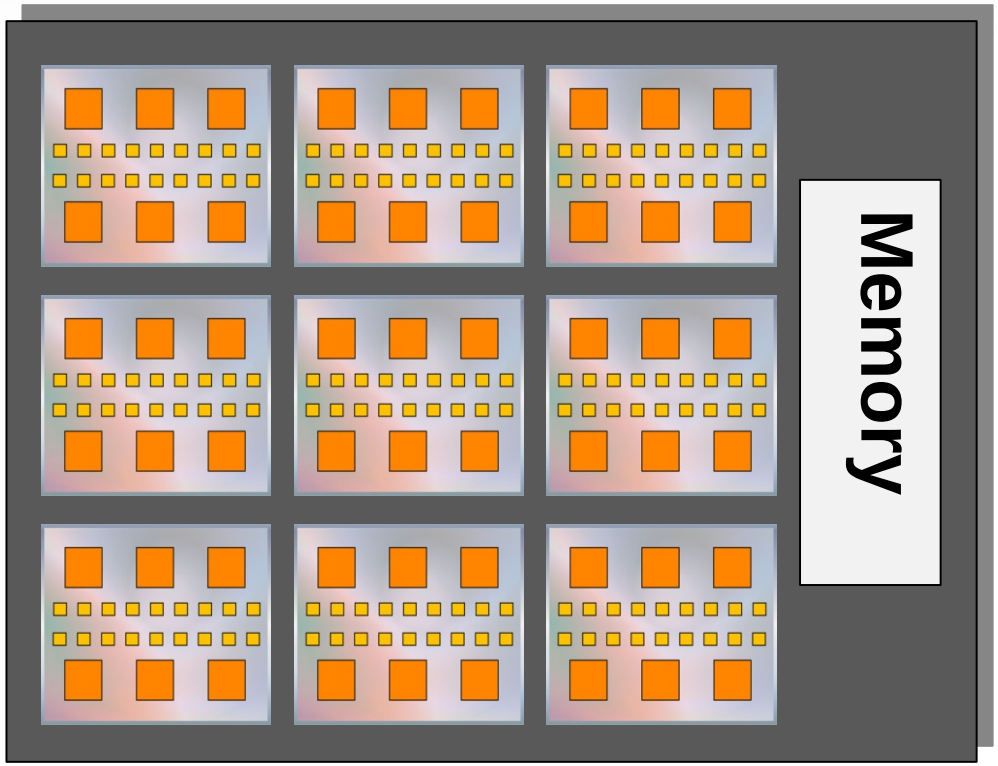


Nível 2 – *Board SMP*

- **Board composto de múltiplos chips que compartilham memória**
  - **SMP**

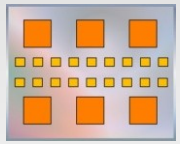


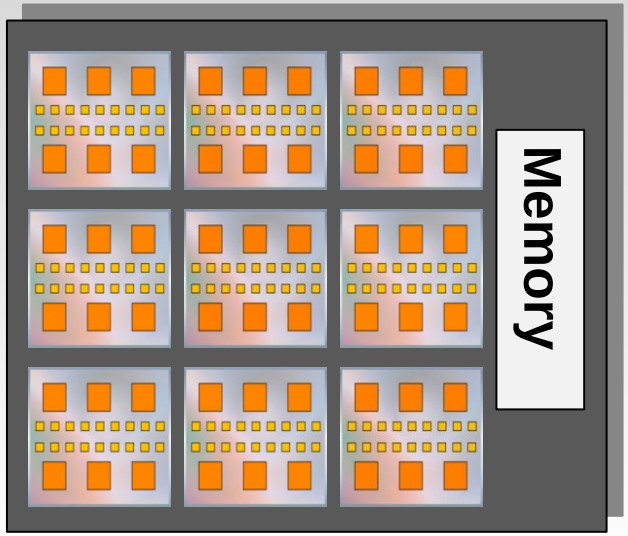


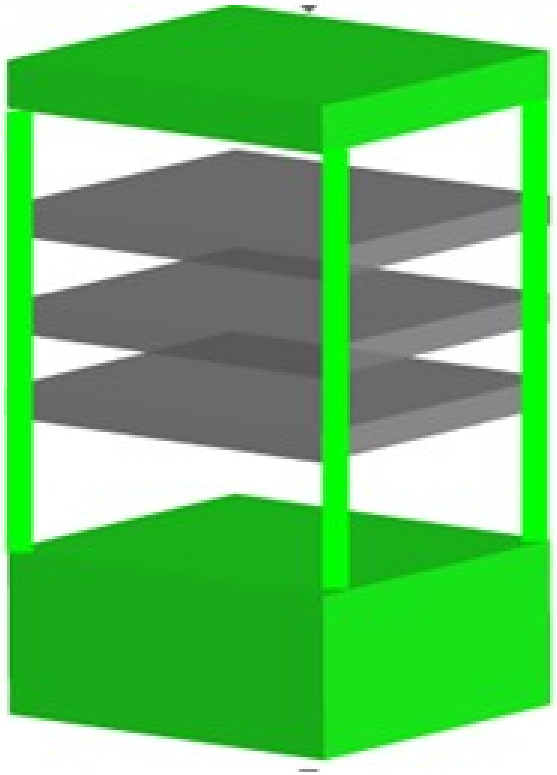


Nível 3 – Rack

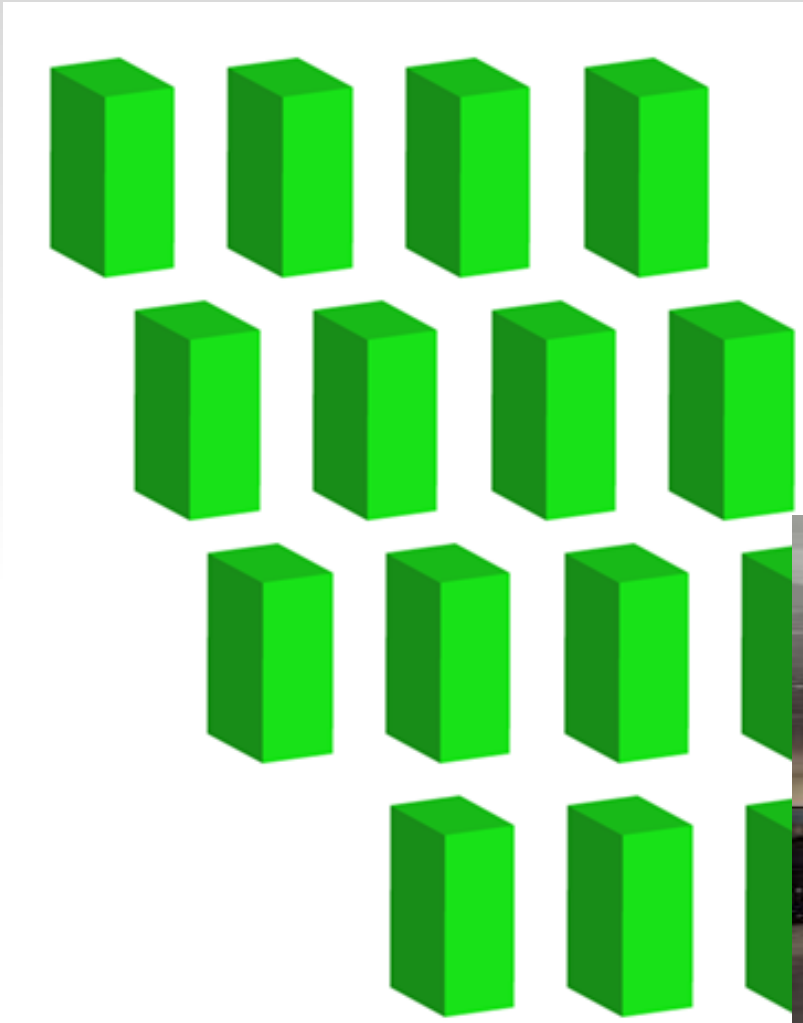
- ***Rack*** **composto de múltiplos** ***boards***
  - **Múltiplos** ***boards*** **tem memória distribuída**

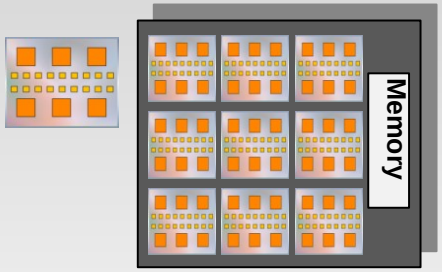


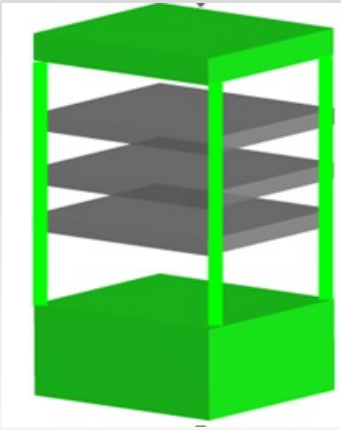




Nível 4 – Cluster de racks

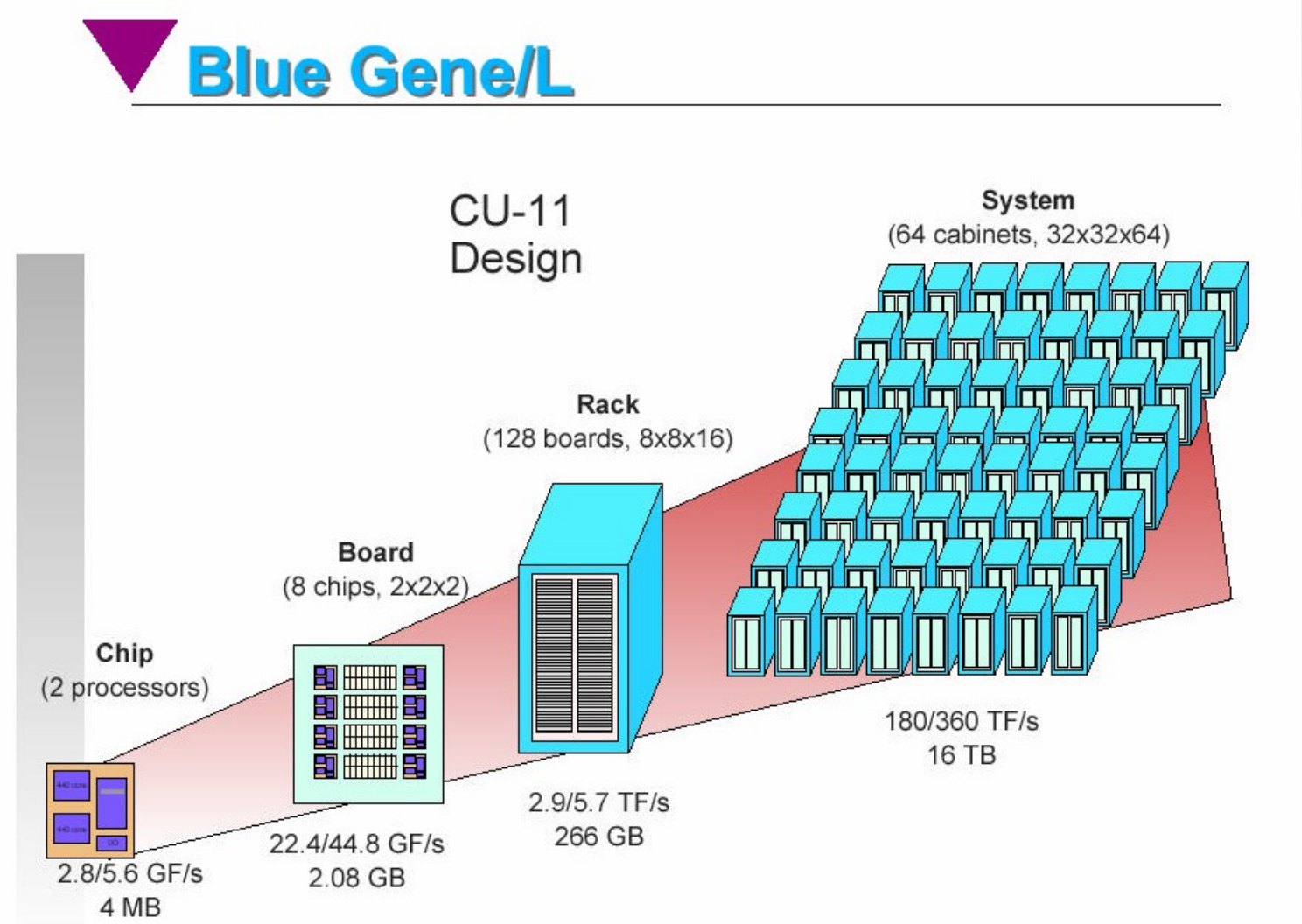






milhares de cores

Exemplo de arquitetura com paralelismo de múltiplos níveis



Nível 5 - Grid

- Vários clusters de racks trocando informações
  - Via internet
  - Fraco acoplamento das tarefas neste nível

Granularidade no paralelismo de múltiplos níveis

Granularidade ou tamanho do grau em paralelismo define a carga de trabalho (*workload*) executada entre nós paralelos em relação ao tempo/carga de comunicação entre eles