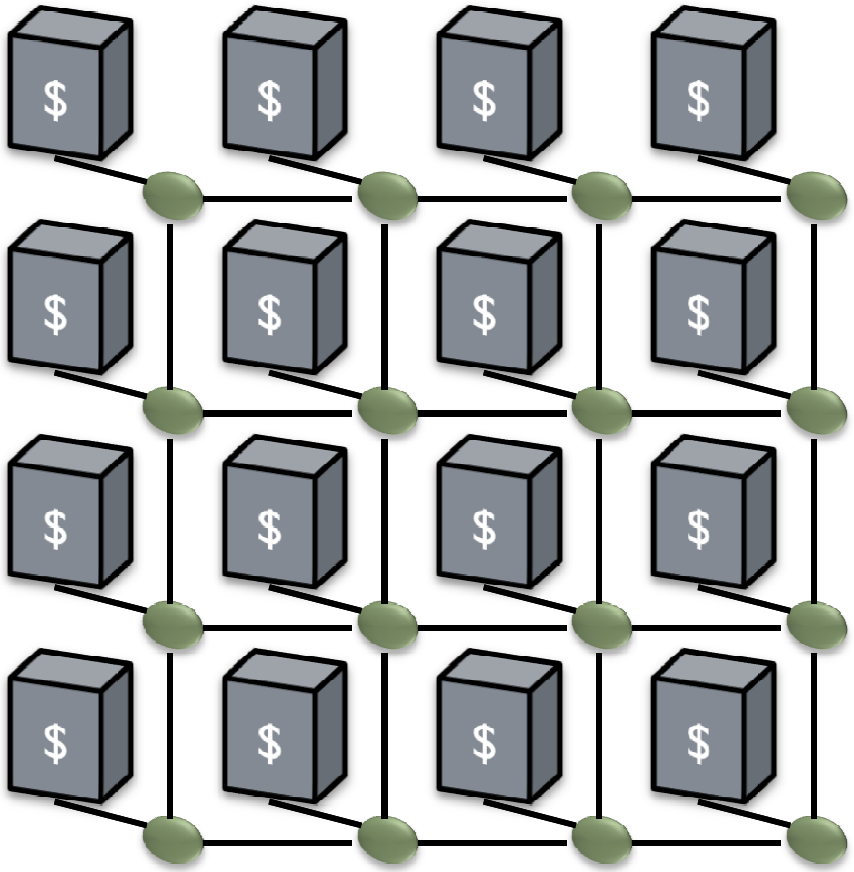
Survey on NUCA: Non-Uniform Cache Architecture

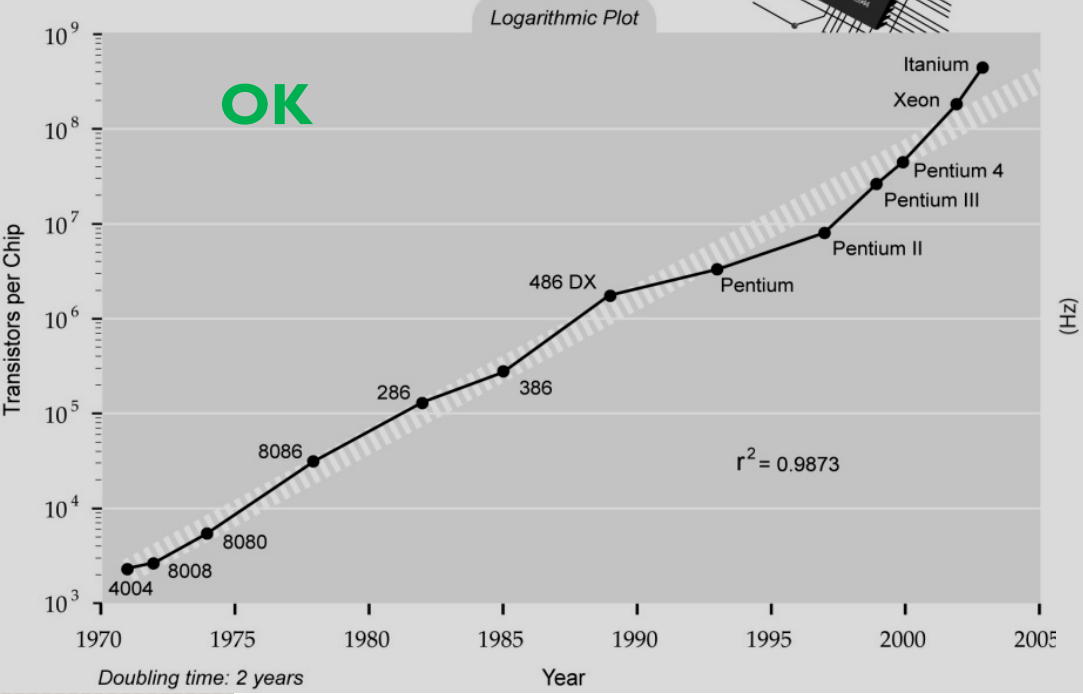
Marco A. Z. Alves
Philippe O. A. Navaux



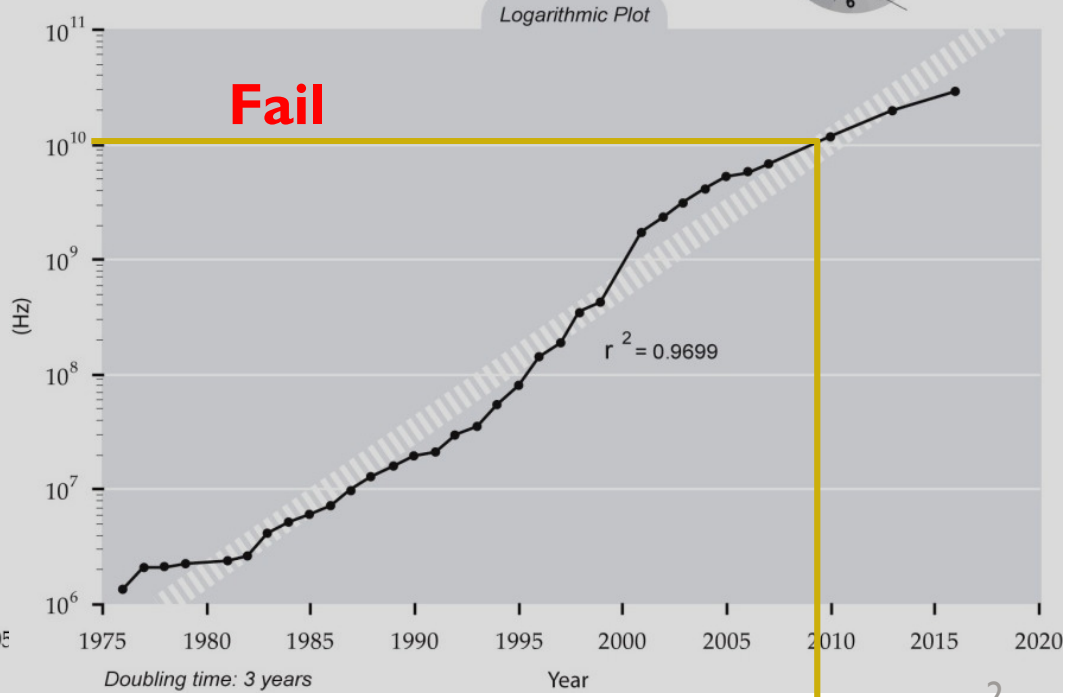
Introdução

- Aumento no Nível de Integração: Chips 65, 45, 32nm
- ☺ Crescimento no poder de criação (super lei de Moore)
- ☹ Wire-Delay, Consumo de Potência (limites físicos)

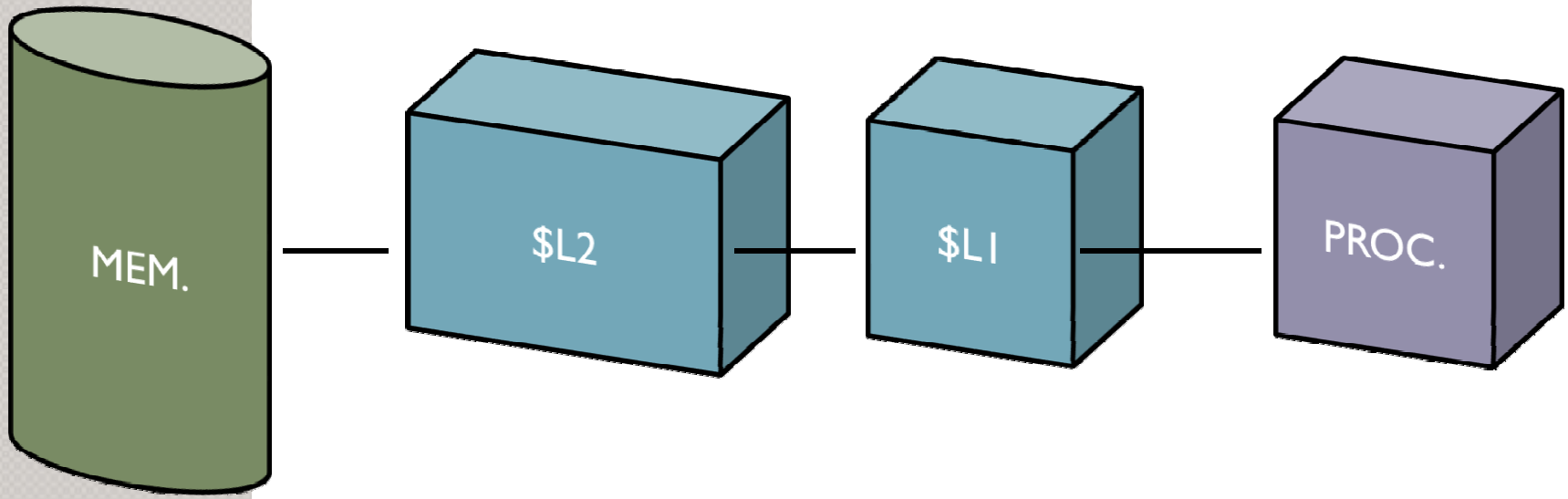
Transistores por Microprocessador



Freqüência do Microprocessador

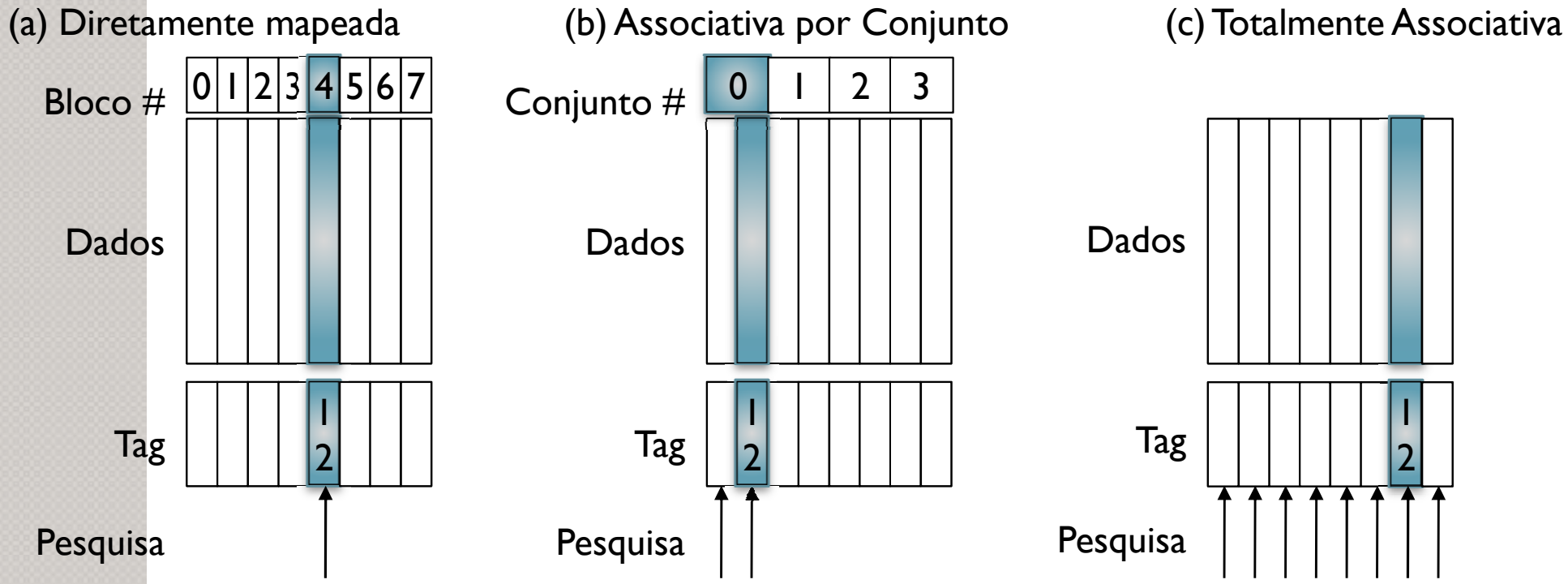


MEMÓRIAS CACHE UNIFORMES



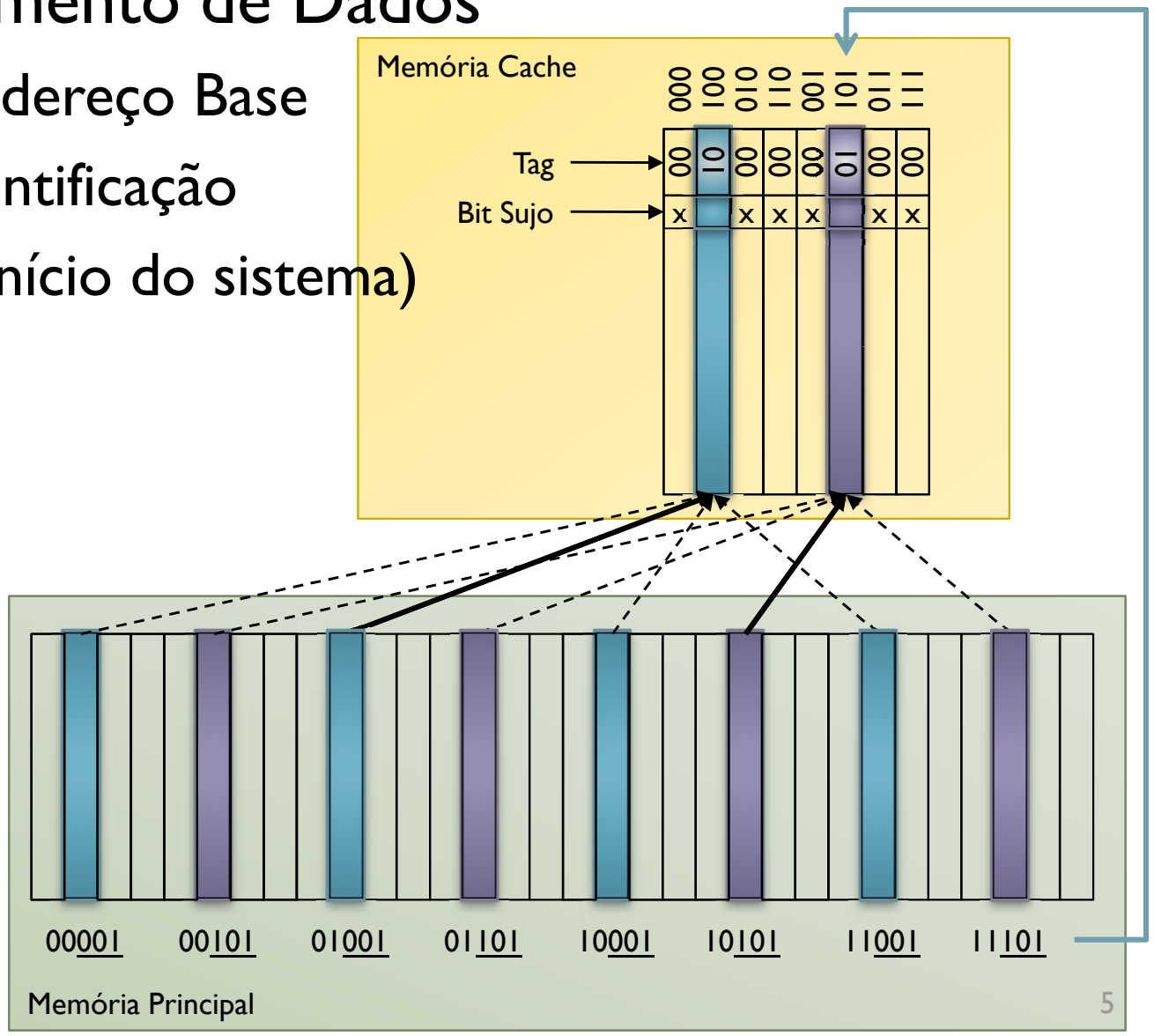
Memórias Cache Uniformes

- Estratégias de Mapeamento de Dados
  - Mapeamento direto (barato)
  - Associativo por Conjunto (médio)
  - Totalmente Associativo (caro)



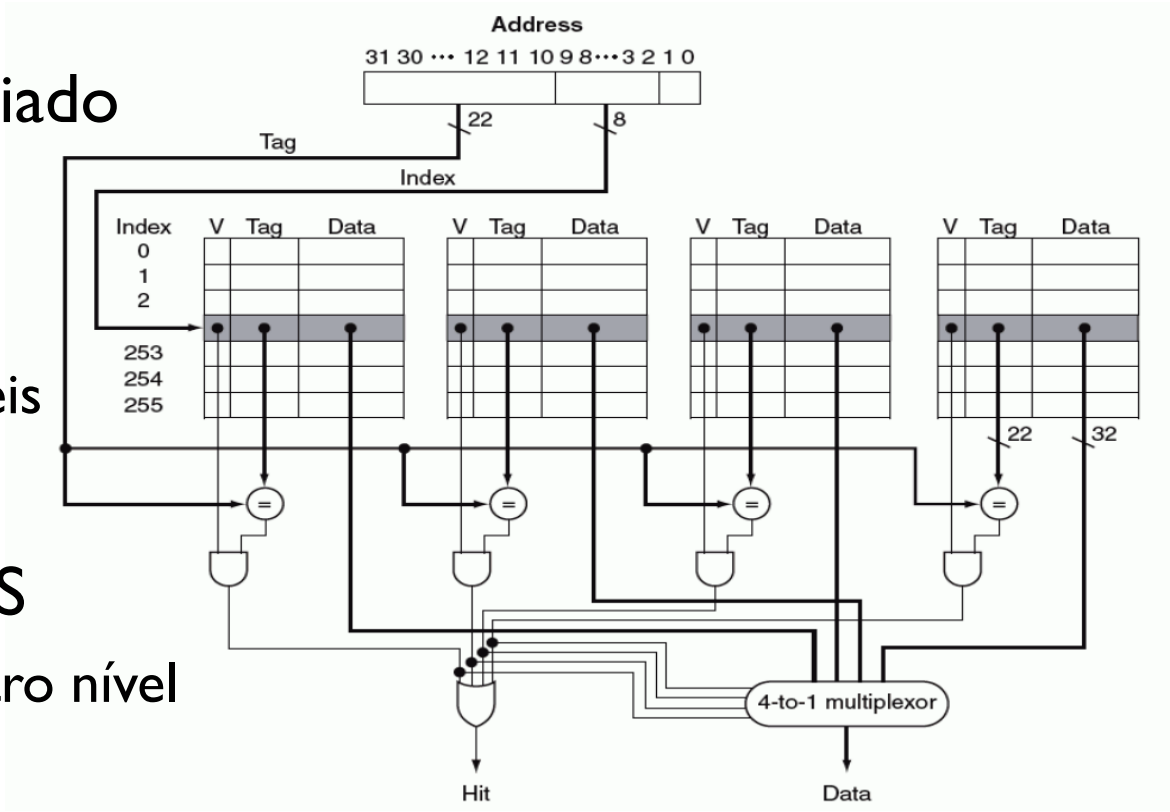
Memórias Cache Uniformes

- Armazenamento de Dados
  - Muitos Endereço Base
  - Tag de identificação
  - Bit Sujo (início do sistema)



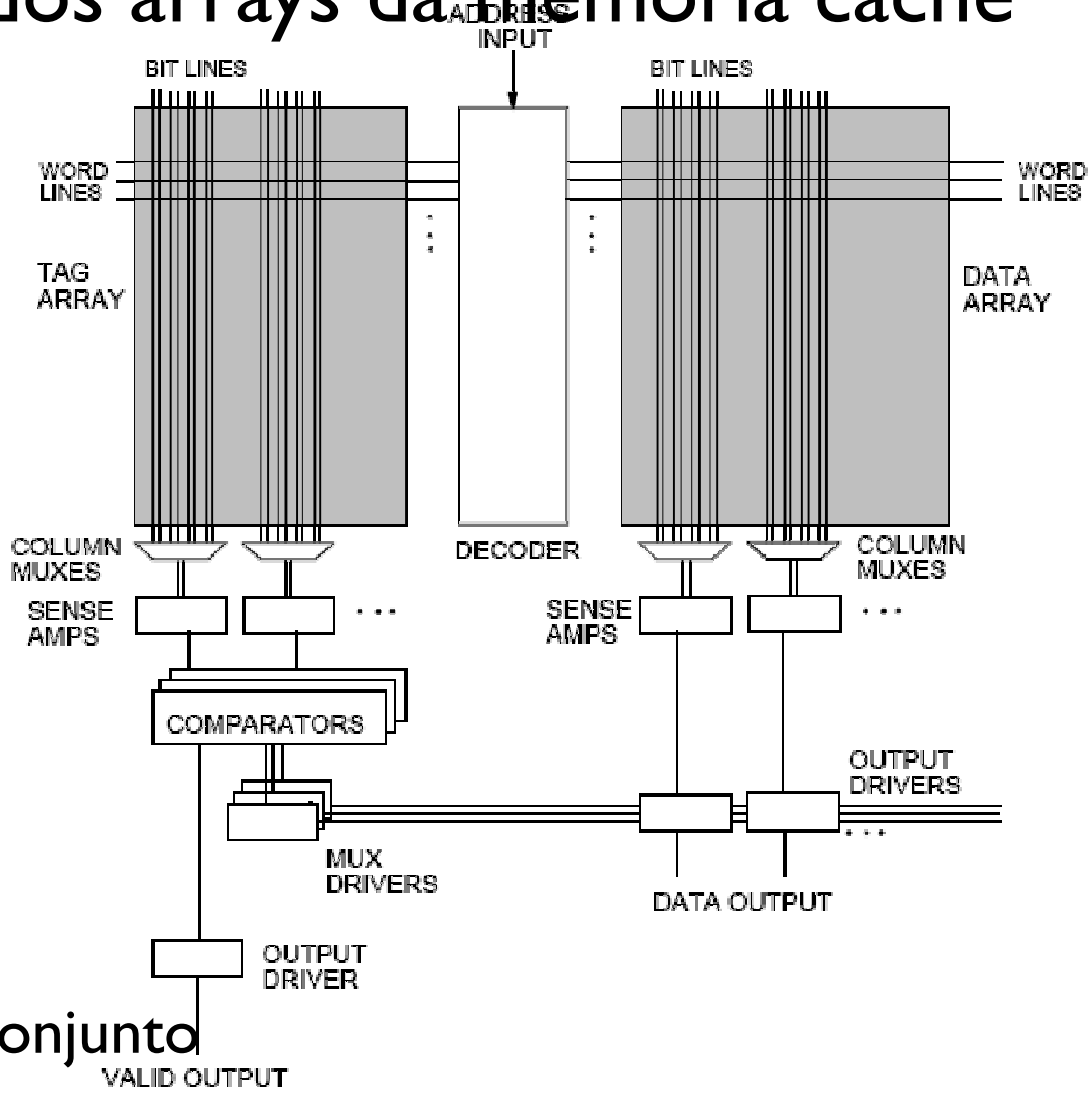
Memórias Cache Uniformes

- Localizando um Bloco na Memória Cache
  - Endereço é enviado
  - Se sinalizar HIT
    - Dados disponíveis
  - Se sinalizar MISS
    - Pesquisa em outro nível



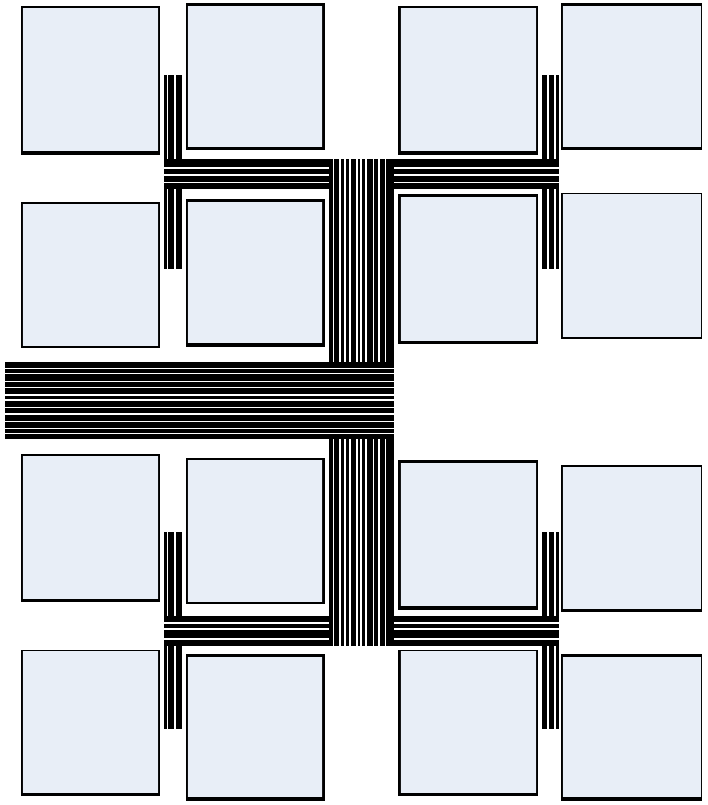
Memórias Cache Uniformes

- Estrutura interna dos arrays da memória cache
  - Array de Tags
  - Array de Dados
  - Decoder
    - Recebe endereço
    - Ativa os vetores
  - Sense Amps
    - Detecta os valores
  - Comparators
    - Mux ativa saída do conjunto

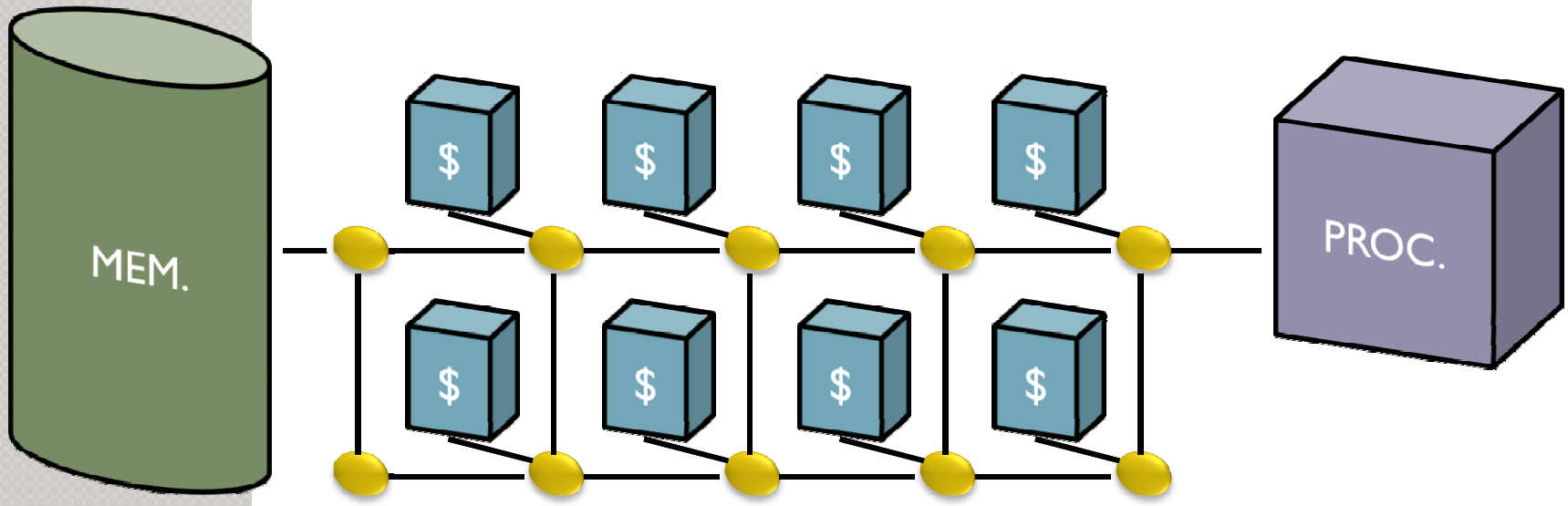


Memórias Cache Uniformes

- Organização dos sub-blocos
  - Layout H-Tree
  - Área total deve considerar blocos e interconexão
  - Tempo de acesso único
  - Problemas:
    - Latência
    - Largura de Banda
    - Área
    - Interconexão
    - Etc.



MEMÓRIAS CACHE NÃO-UNIFORMES

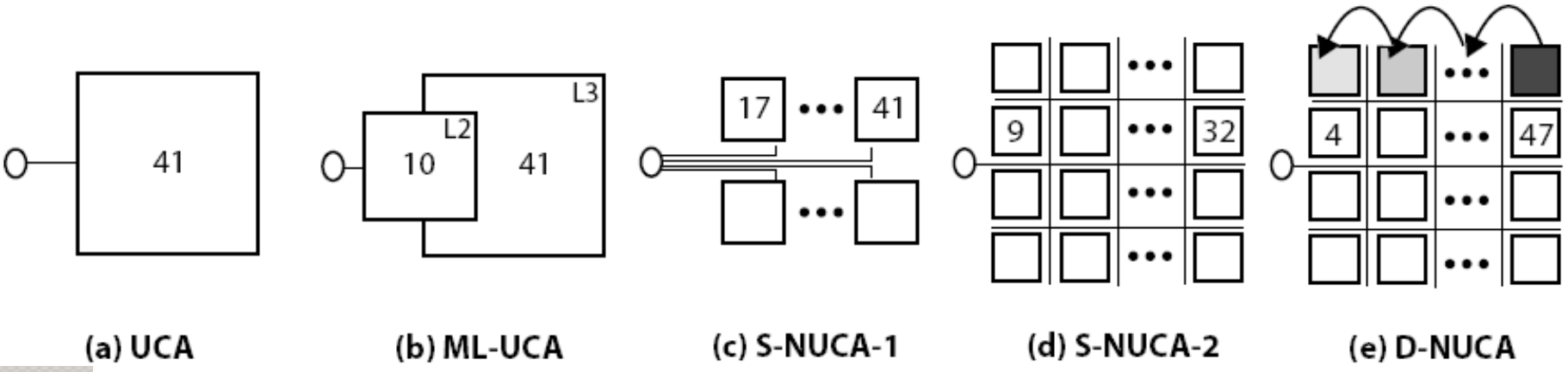


Memórias Cache Não-Uniformes

- Origem das NUCA
  - Tempo de acesso a grandes memórias cache não apropriadas
  - Problemas de atraso do fio entre outros problemas físicos
  - Isolamento dos blocos NUCA podem ser convenientes para futuras tecnologias de integração
- Funcionamento Básico
  - Blocos próximos ao processador vão ser acessados mais rapidamente

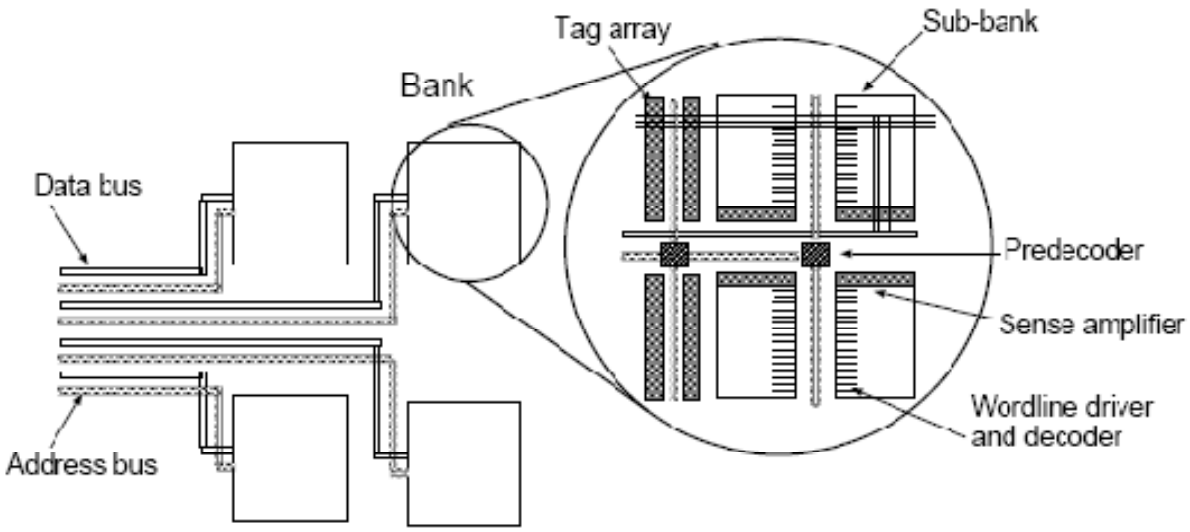
Memórias Cache Não-Uniformes

- Tipos básicos de NUCA com típicos tempos de acesso
  - A) Memória UCA
  - B) Memória UCA com níveis 2 e 3
  - C) NUCA estática com barramento
  - D) NUCA estática com NoC
  - E) NUCA dinâmica com NoC



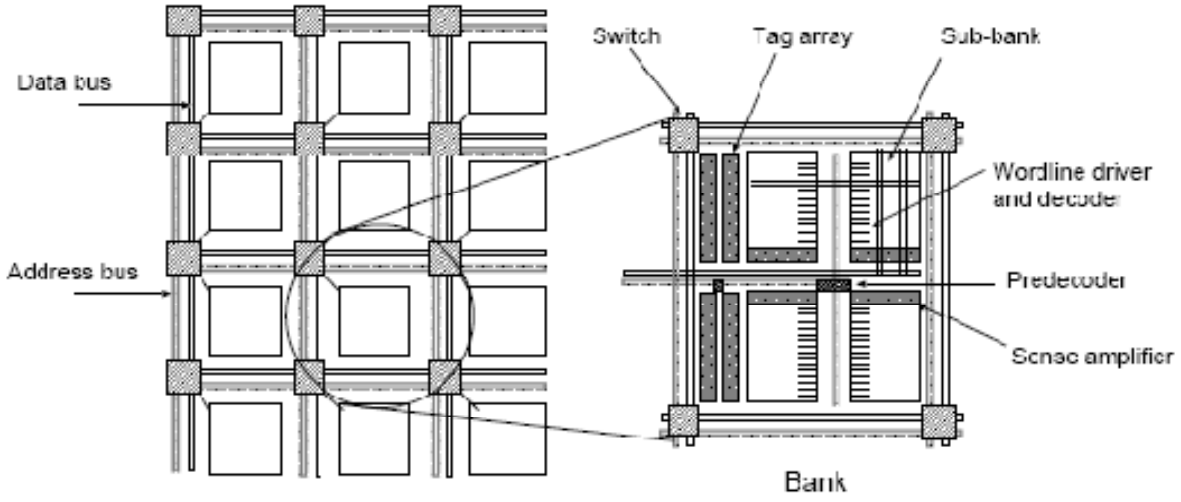
Memórias Cache Não-Uniformes

- Interconexão dos Blocos NUCA
  - Estrutura da S-NUCA 1
    - Barramentos conectando os bancos



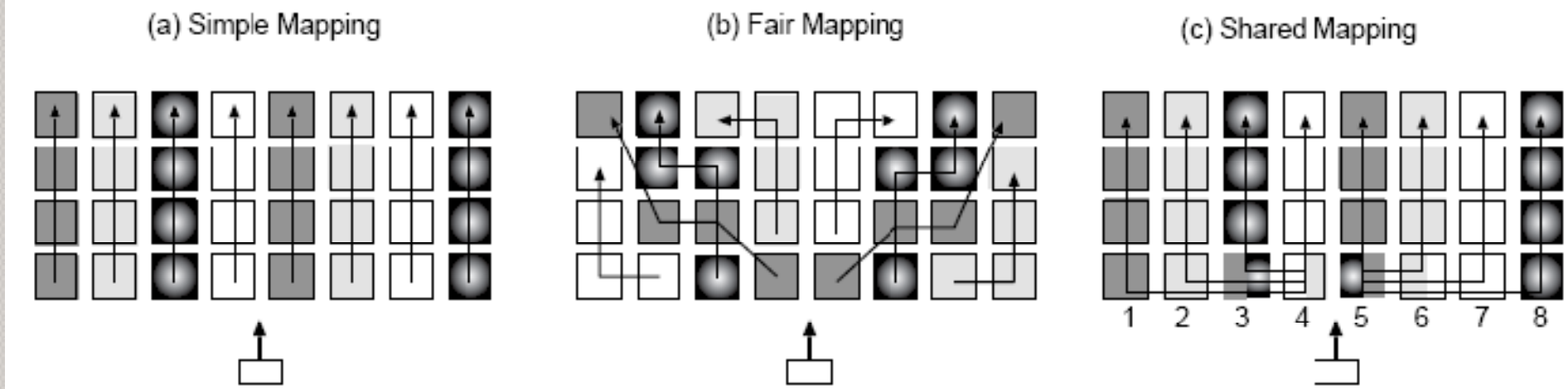
Memórias Cache Não-Uniformes

- Interconexão dos Blocos NUCA
  - Estrutura S-NUCA 2 e D-NUCA
    - Utiliza redes de interconexão intra-chip



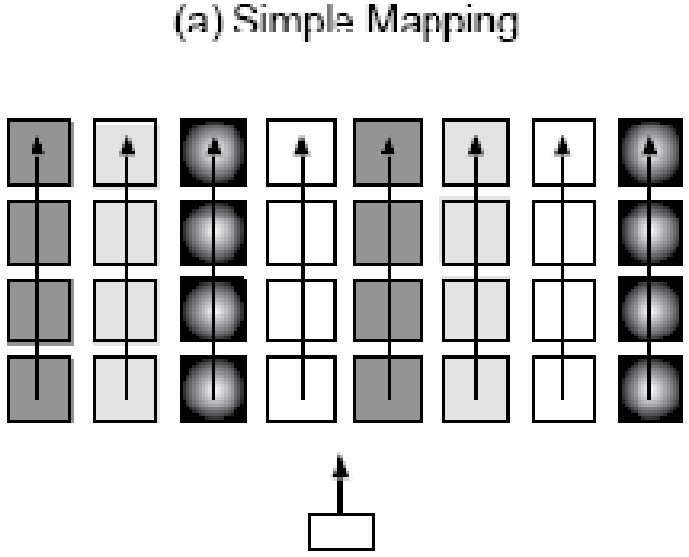
Memórias Cache Não-Uniformes

- Políticas de Mapeamento de Blocos
  - A) Mapeamento Simples
  - B) Mapeamento balanceado
  - C) Mapeamento Compartilhado



Memórias Cache Não-Uniformes

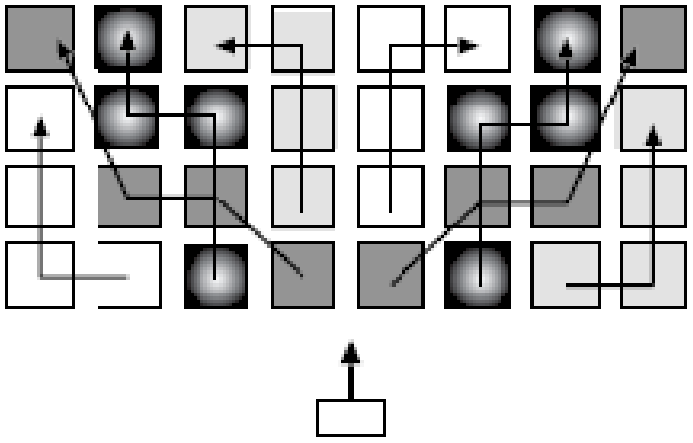
- Mapeamento Simples
  - Colunas são conjuntos associativos
  - Cada Linha é uma via do conjunto associativo
  - ☹ Linhas podem não corresponder ao número de vias associativas
  - ☹ Tempo de acesso as vias diferente dentro de um mesmo coniunto associativo



Memórias Cache Não-Uniformes

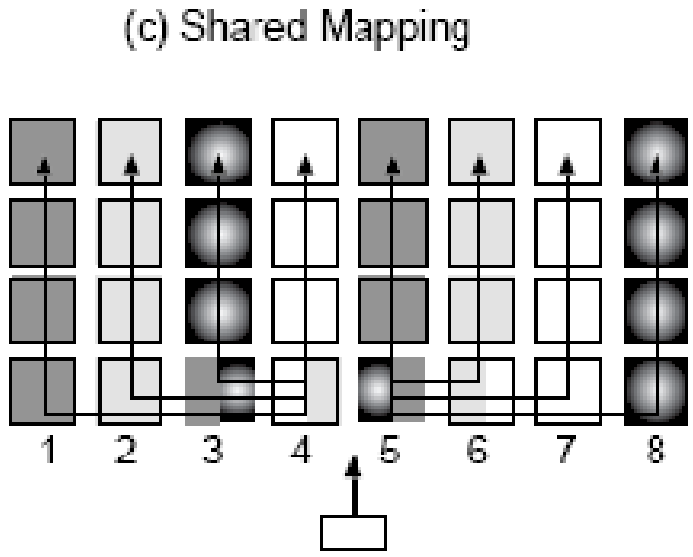
- Mapeamento Balanceado
  - Tenta equalizar o tempo de acesso aos conjuntos associativos
  - ☺ Tempo de acesso equilibrado
  - ☹ Sobrecusto de complexidade adicional
  - ☹ Sobrecusto de roteamento

(b) Fair Mapping



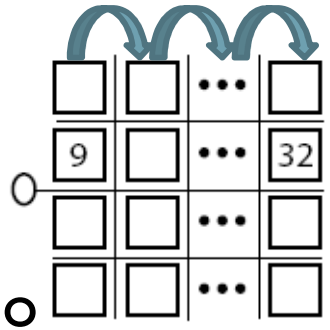
Memórias Cache Não-Uniformes

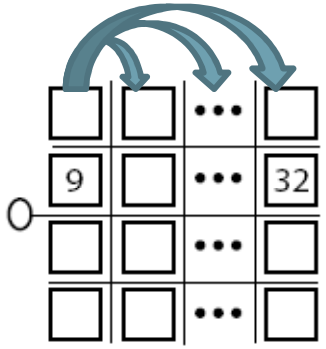
- Mapeamento Compartilhado
  - Compartilha os bancos próximos ao processador
  - ☺ Provê rápido tempo de acesso a todos conjuntos
  - ☹ Complexidade do mapeamento



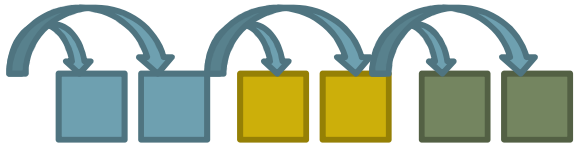
Memórias Cache Não-Uniformes

- Busca de Dados em NUCA
  - Incremental Search
    - ☺ Baixo consumo ☹ Baixo desempenho
  - Multicast Search
    - ☺ Alto Desempenho ☹ Alto Consumo

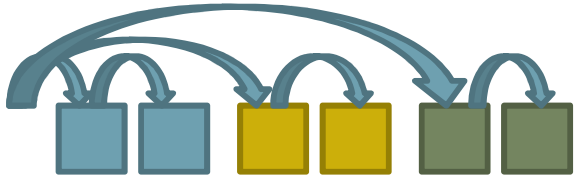




  - Limited Multicast
    - Dividido em sub-grupos de procura paralela

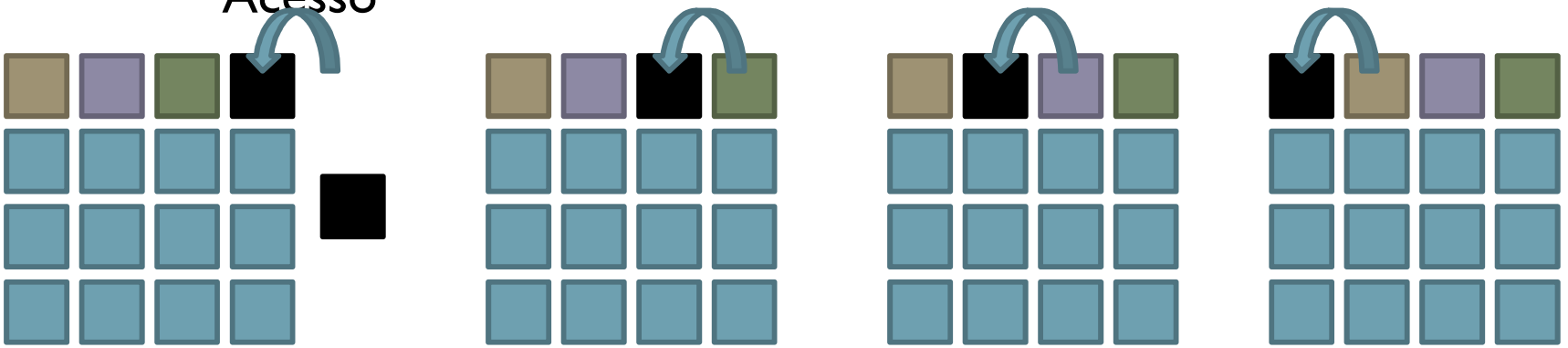


  - Partitioned Multicast
    - Dividido em sub-grupos paralelos



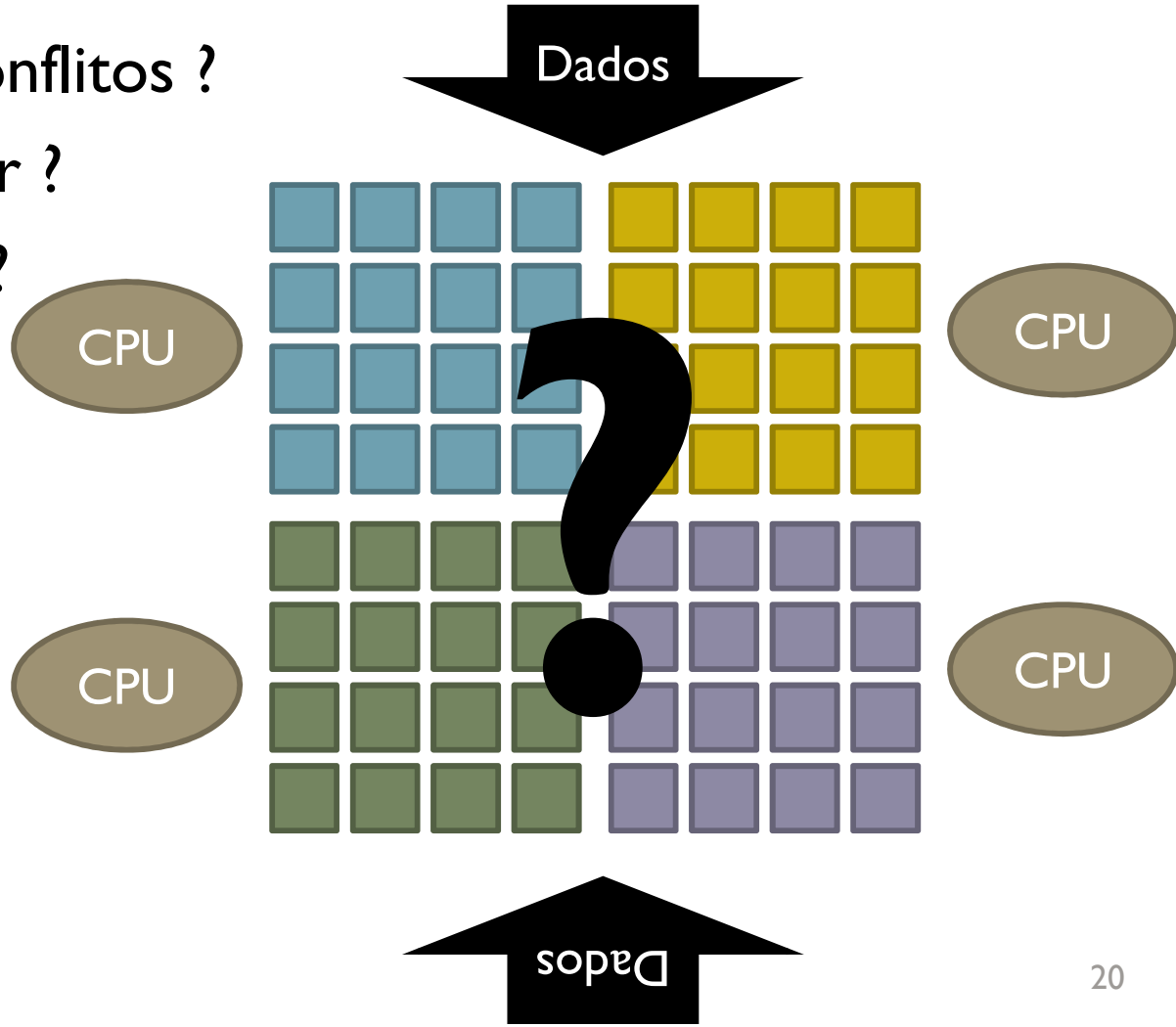
Memórias Cache Não-Uniformes

- Migração de Dados em Memórias NUCA dinâmicas
  - Tenta maximizar os acessos a bancos próximos, de menor latência
  - LRU – Bancos próximos com MRU (Dados Usados mais recentemente)
    - ☹ Grandes movimentações de dados
  - Generational Promotion
    - Remove a Vítima ou Troca de Lugar com a Vítima
  - 1° Acesso 2° Acesso 3° Acesso 4° Acesso



Memórias Cache Não-Uniformes

- Memórias NUCA em processadores Multi-Core
  - Como manter coerente ?
  - Como evitar conflitos ?
  - Como organizar ?
  - Como mapear ?



Considerações Finais

- Problemas físicos estão cada vez mais claros
- Utilização de many-core parece ser uma solução
- Dúvidas:
  - Como trocar informações ?
  - Como prover dados ?
  - Como programar ?
  - Como reduzir o consumo ?
  - Como manter crescimento do desempenho ?
- Soluções de suporte a todos os núcleos são necessárias
  - Redes de Interconexão Intra-Chip - NoC
  - Memórias Cache Não-Uniforme - NUCA

Memórias NUCA
Non-Uniform Cache Architecture

Marco Antonio Zanata Alves

marco.zanata@inf.ufrgs.br

Physical Area Analysis

Exp.	Organization	Total Logical Size	Slice Logical Size	Associativity	Line Size	Normalized Physical Area	Latency	Penalty	Normalized Cache Misses
1	1Core/L2	32MB	1MB	8 Ways S.A.	64 Bytes	100%	1.6 ns	4 Cycles	100%
1	2Cores/L2	16MB	1MB	8 Ways S.A.	64 Bytes	50%	1.6 ns	4 Cycles	112%
2	2Cores/L2	32MB	2MB	8 Ways S.A.	64 Bytes	78%	2.1 ns	5 Cycles	89%
3	2Cores/L2	16MB	1MB	16 Ways S.A.	64 Bytes	50%	2.6 ns	6 Cycles	108%
4	2Cores/L2	16MB	1MB	8 Ways S.A.	128 Bytes	73%	2.6 ns	6 Cycles	68%

- ☹ Increase on physical area size for same cache size.
- ☹ Increase on data access latency
- ☺ Balance between data access latency and reduction on cache misses (Efficient Cache)

Global Conclusions

- The traditional techniques as increase on cache size, associativity, and line size did not work:
  - Cache size cannot increase the performance.
  - High associativity cannot increase the performance.
- Any change in cache parameters increase the latency.
- Need a good trade-off between cache latency and cache misses.
- Future works:
  - Developing non-uniform cache architecture (NUCA) considering our results, in order to hide high latency problems.