QXD-005 - Arquitetura de Computadores

Memória Interna Moderna

Prof. Pedro Botelho

Nas Aulas Passadas…

- Visão de Alto Nível do Computador
- Memória Cache
- Introdução a Memória Interna
- Questões:
  - Quais são os tipos de memórias usadas nos computadores **atualmente**?
  - Quais são as tecnologias **modernas** utilizadas nas memórias?

Nesta Aula…

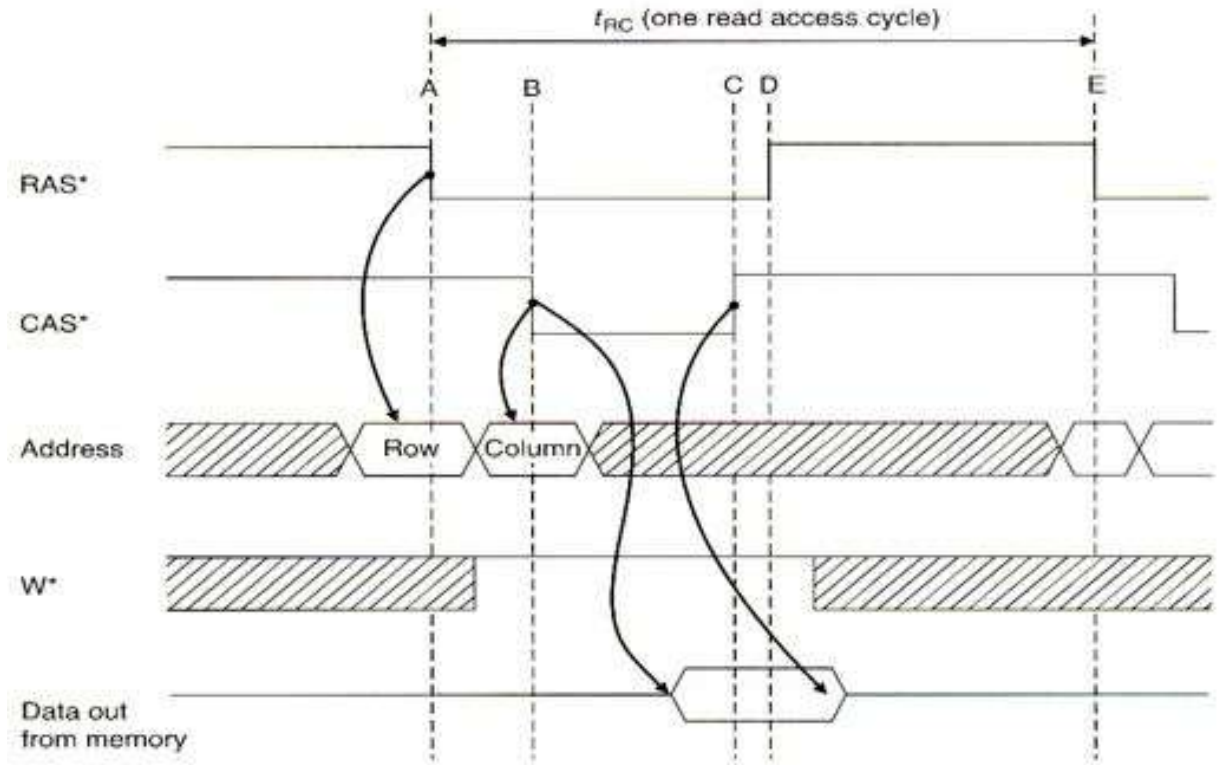
- Memórias DRAM Melhoradas
  - SDRAM
  - DDR
- Implementação da Memória Principal
- Memória Flash
- Aplicações e Estudo de Caso
  - Computadores Pessoais
  - Sistemas Embarcados

Memória Interna Moderna

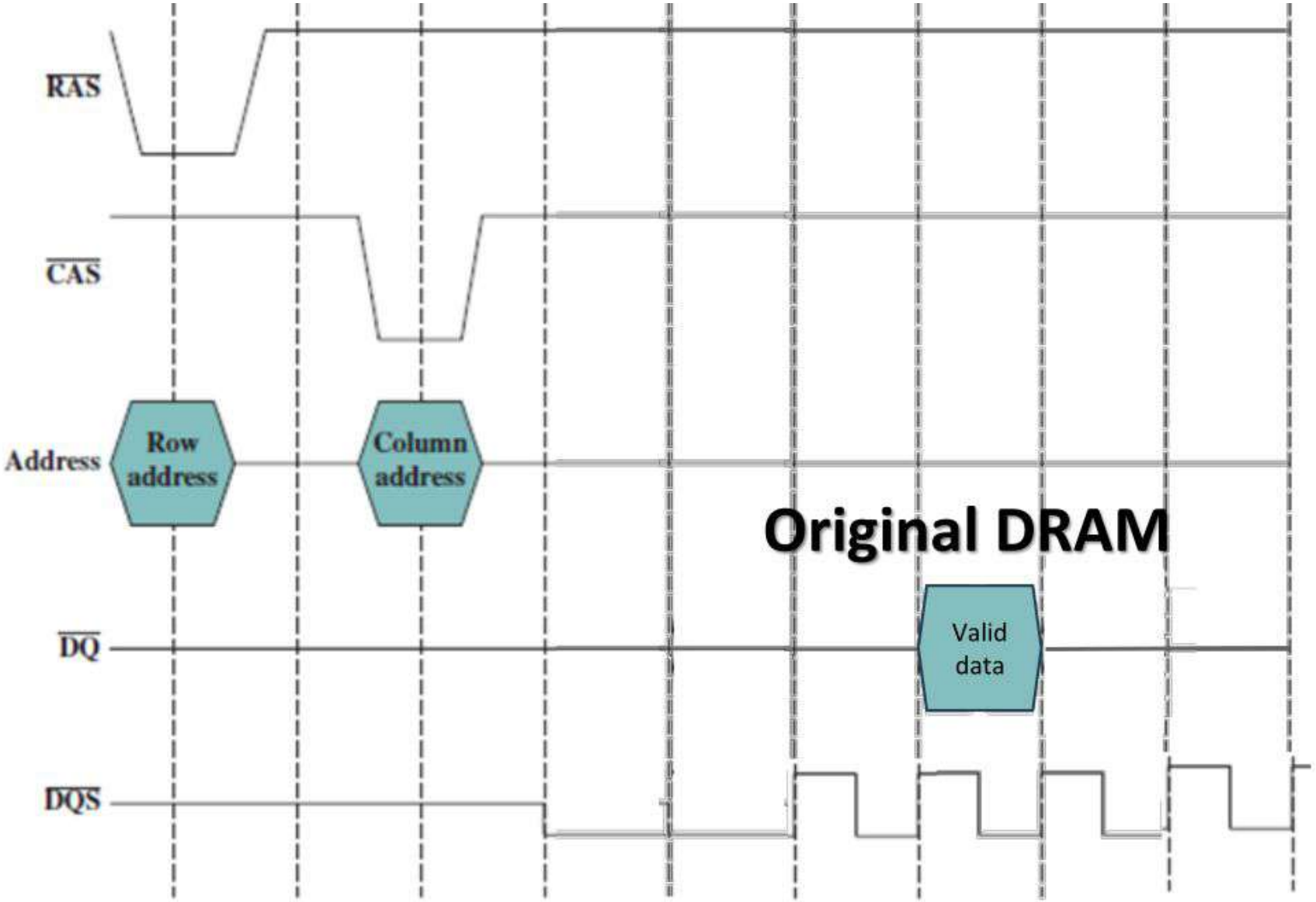
Memórias DRAM Melhoradas

Memória DRAM Assíncrona (Tradicional)

- Memória principal é um **gargalo crítico**: Mais lento que o processador
  - Caches (SRAM) internas à CPU não resolvem o problema: Cara e pequena
- Memória tradicional leva um tempo para realizar operação: Processo deve esperar
  - Têm-se um **tempo de acesso** para a DRAM realizar funções internas

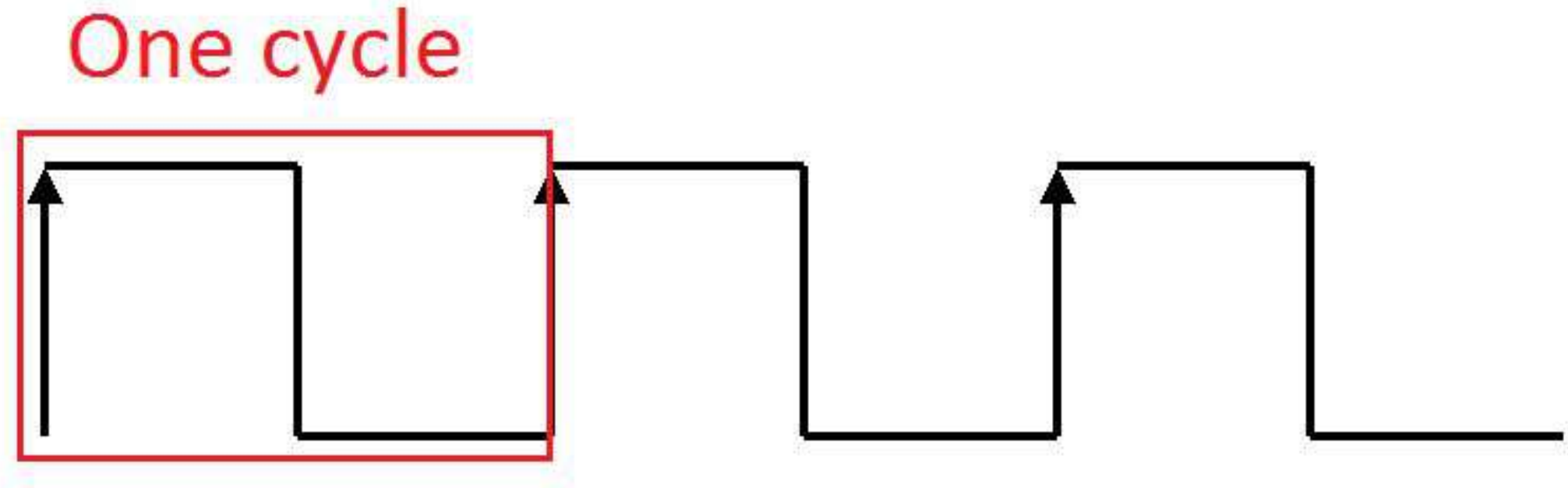


Temporização da Memória DRAM



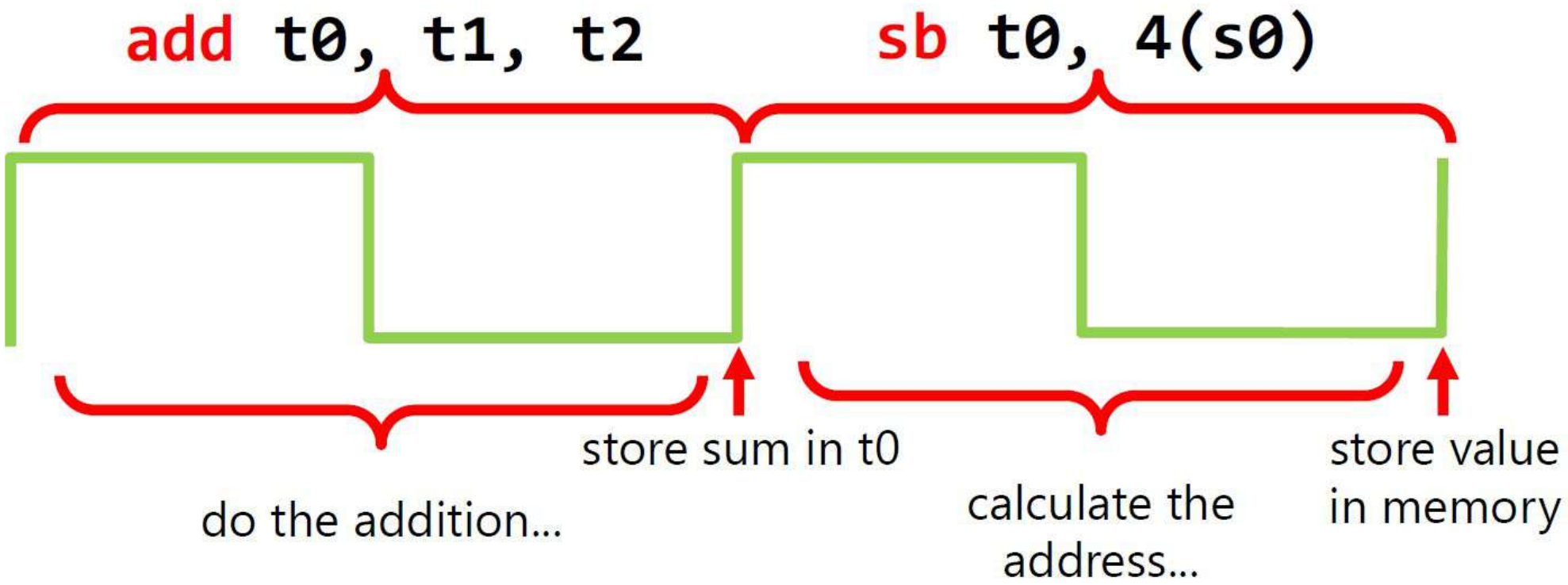
Recapitulando: O *Clock* do Processador

- Operações no processador são temporizadas conforme um sinal periódico: O ***clock***
  - Ex: Um computador que funciona a 10MHz realizaria 10 milhões de operações por segundo
  - Mas tem que esperar ocasionalmente pela memória principal DRAM!
- **Pergunta**: Como fazer o processador parar de esperar a memória?
- **Solução**: Sincronizar a memória com o *clock* do sistema!



Exemplo: Ciclo de Clock no RISC-V

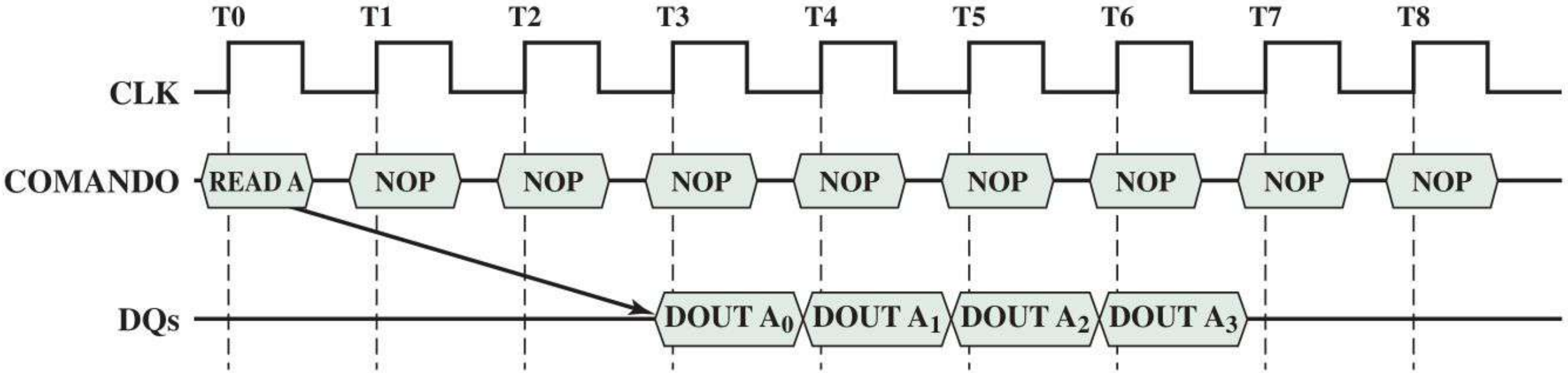
- Processador executa uma **instrução** a cada **ciclo** de *clock:*



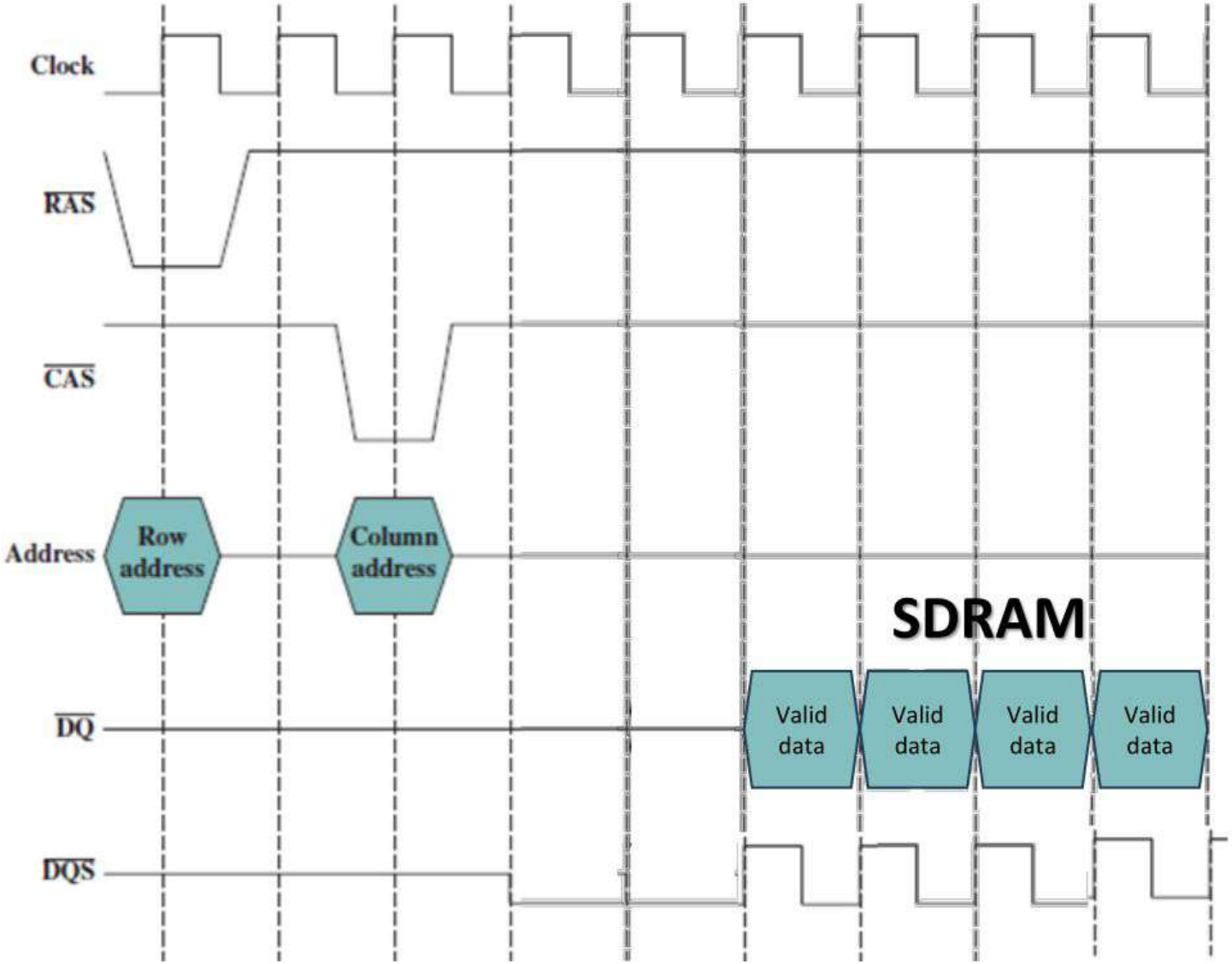
In this example each instruction **ends on the rising edge of the clock**
since that's when the registers store their values

Memória DRAM Síncrona (SDRAM)

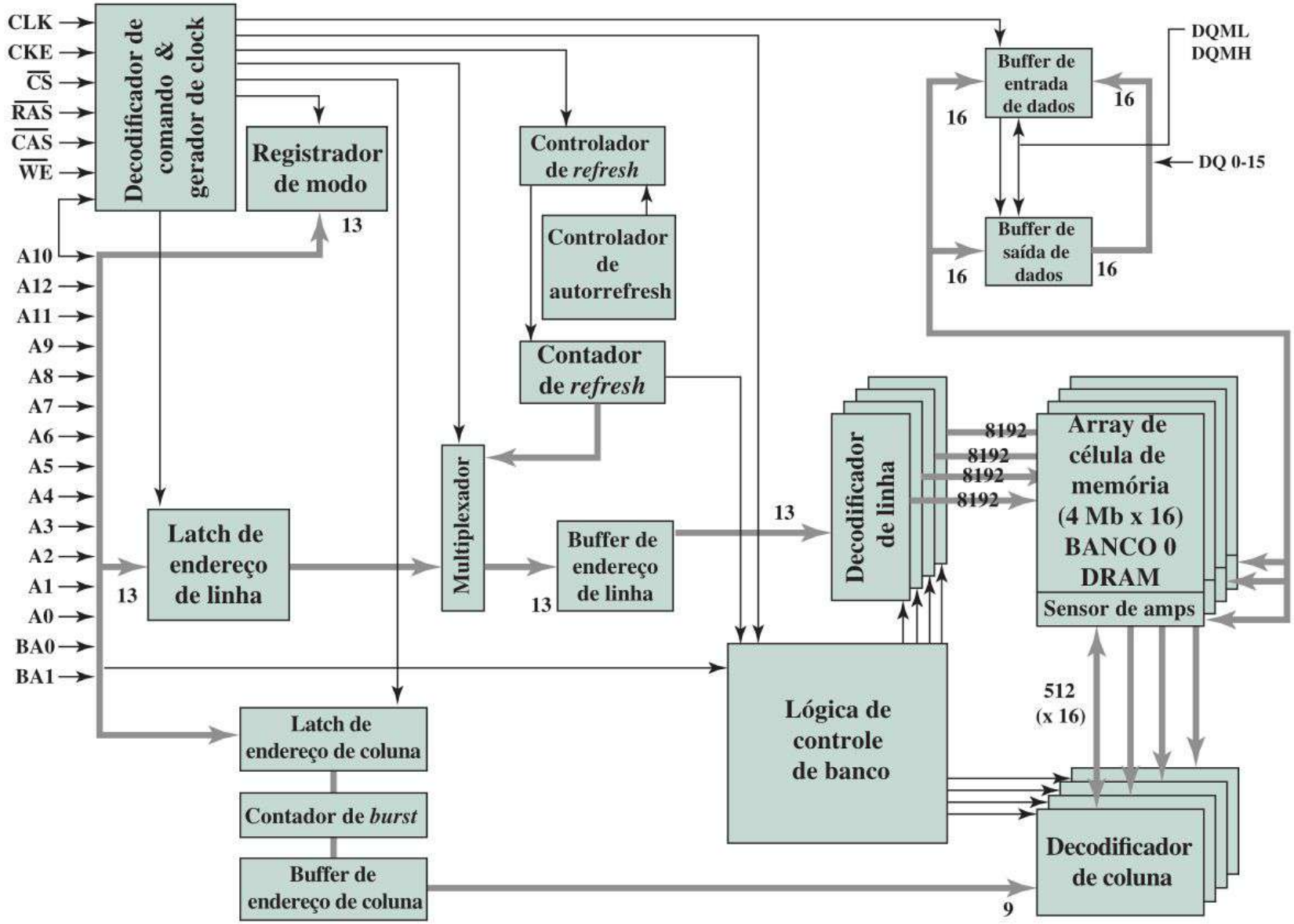
- Sincronizadas pelo *clock* da placa-mãe: Sem tempos de espera
  - Lê/escreve na borda de subida do *clock*
- **Latência de CAS (CL)**: Tempo entre o endereço de coluna e o dado ficar disponível
- **Modo Rajada (*Burst*)**: Dados acessados em série após *setup* do endereço
  - Acesso em Bloco: Permite configurar RAS/CAS apenas para o primeiro byte



Temporização da Memória SDRAM

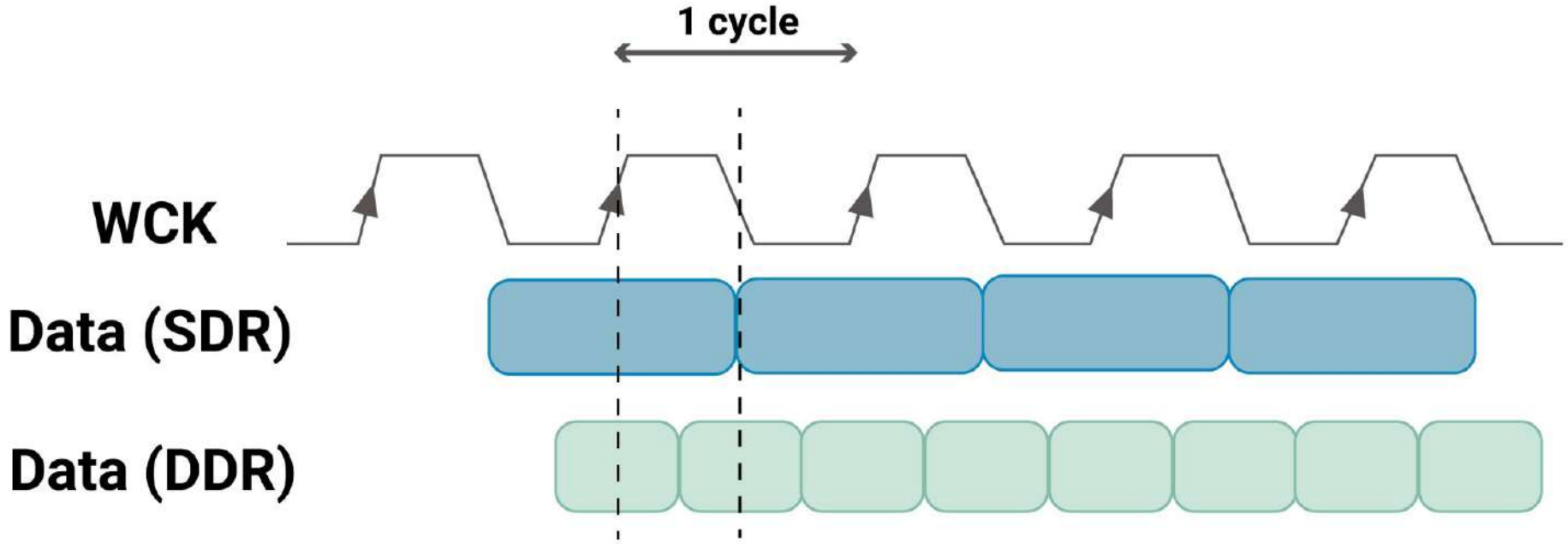


Organização Interna de uma SDRAM de 256Mbits

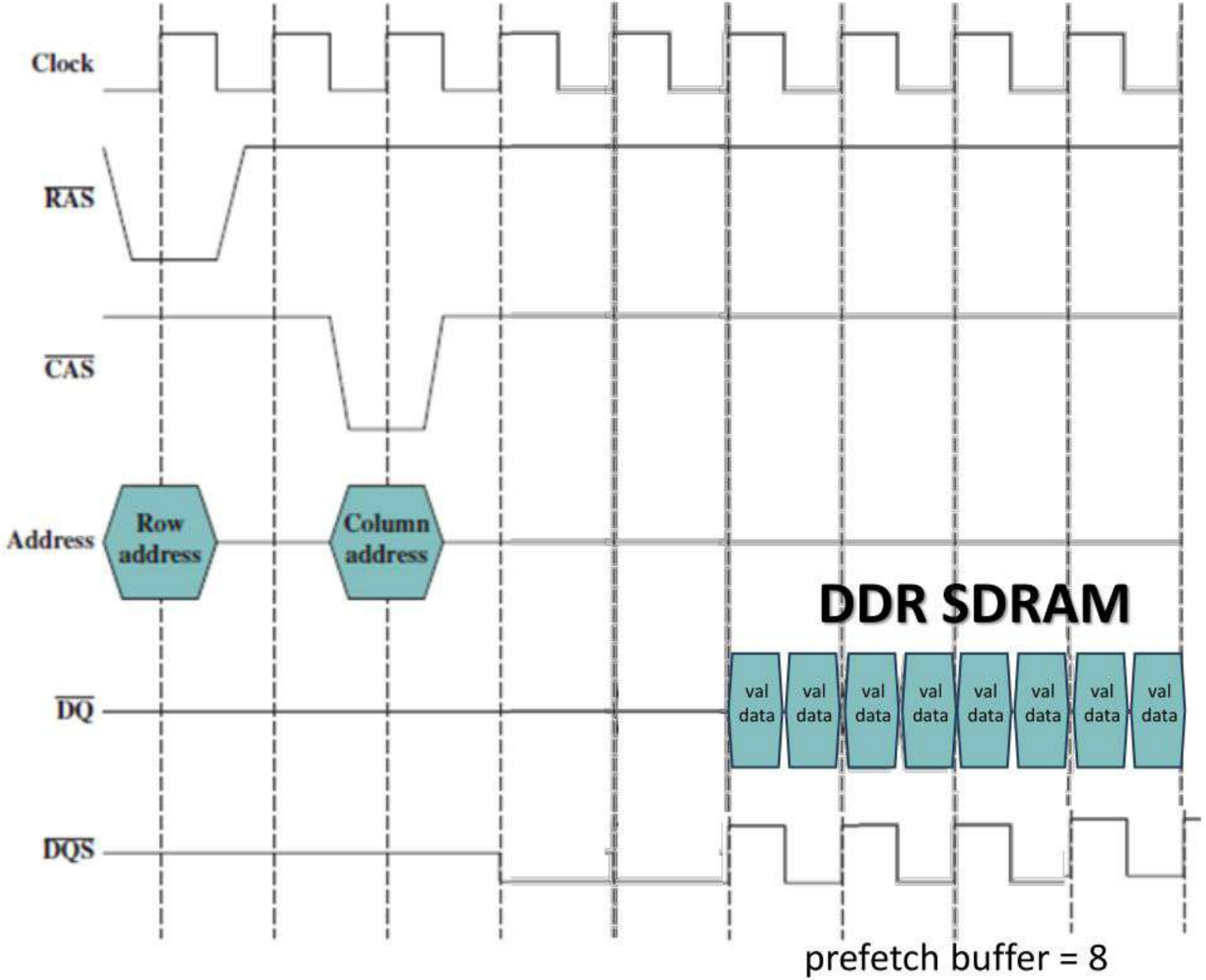


Memória SDRAM de Taxa Dupla de Dados (DDR SDRAM)

- Alcança taxas mais altas de dados que a SDRAM de três formas:
  - Acessa dados na borda de **subida** e de **descida** do *clock* (duas vezes por ciclo)
  - Frequência de *clock* do **barramento** mais alta: Taxa de transferência maior
  - ***Buffering***: Dados são pré-buscados (*prefetching*), ficando pronto para serem acessados
- *Low-Power* DDR (LPDDR): Projetada para dispositivos com restrições energéticas
  - Memórias LPDDR já vem soldadas nos dispositivos finais



Temporização da Memória DDR SDRAM

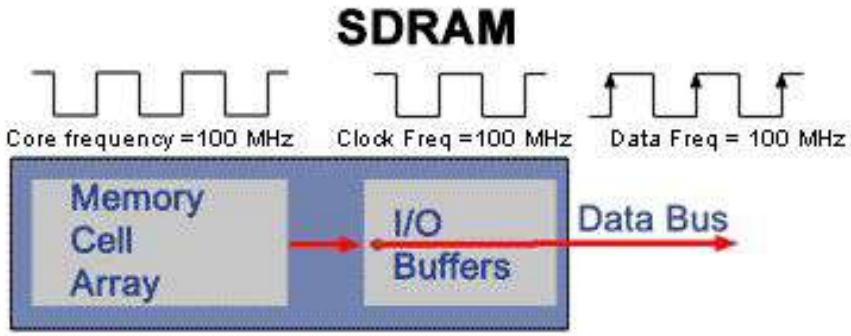


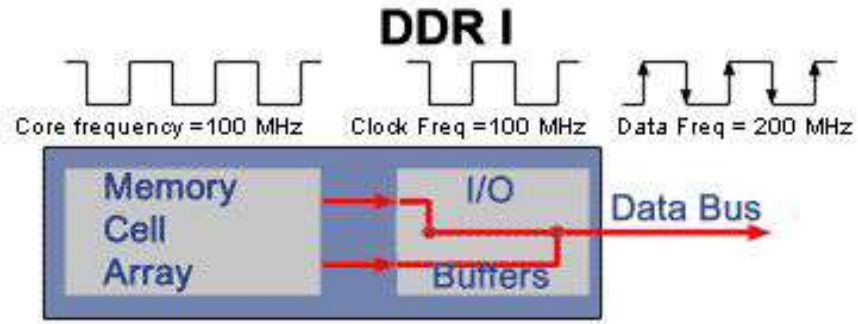
Memória Interna Moderna

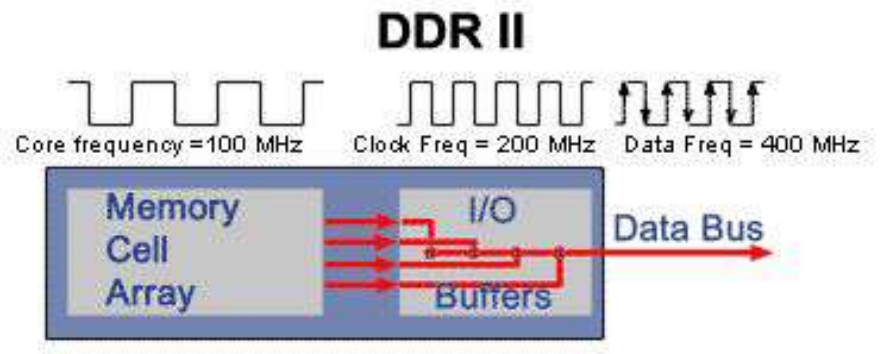
Implementação da Memória Principal

Implementação das Memórias DDR

- Ideia do DDR → Ler dados na borda de **subida** e **descida**
  - Chip SDRAM e barramento com mesma frequência (100MHz)
  - Mas **taxa de dados dupla** (200Mbps)
  - Chip DRAM então é mais lento que a interface de E/S
- Como ler uma memória com o **dobro de taxa**?
  - Ler dois bits de uma vez → Dados acessados em **paralelo**
  - Arranjo de células de memória fornecem *n* bits
  - Devem ser **serializados** pela interface DDR
- **Prefetch Buffer**: "Memória cache" no chip SDRAM
  - Multiplexa os bits pro barramento serial (rajada, ou ***burst***)
  - Permite altas taxas de transferência
- DDR2 permite leitura de 4 palavras de 64 bits (dobro)
  - Frequência do barramento é maior!







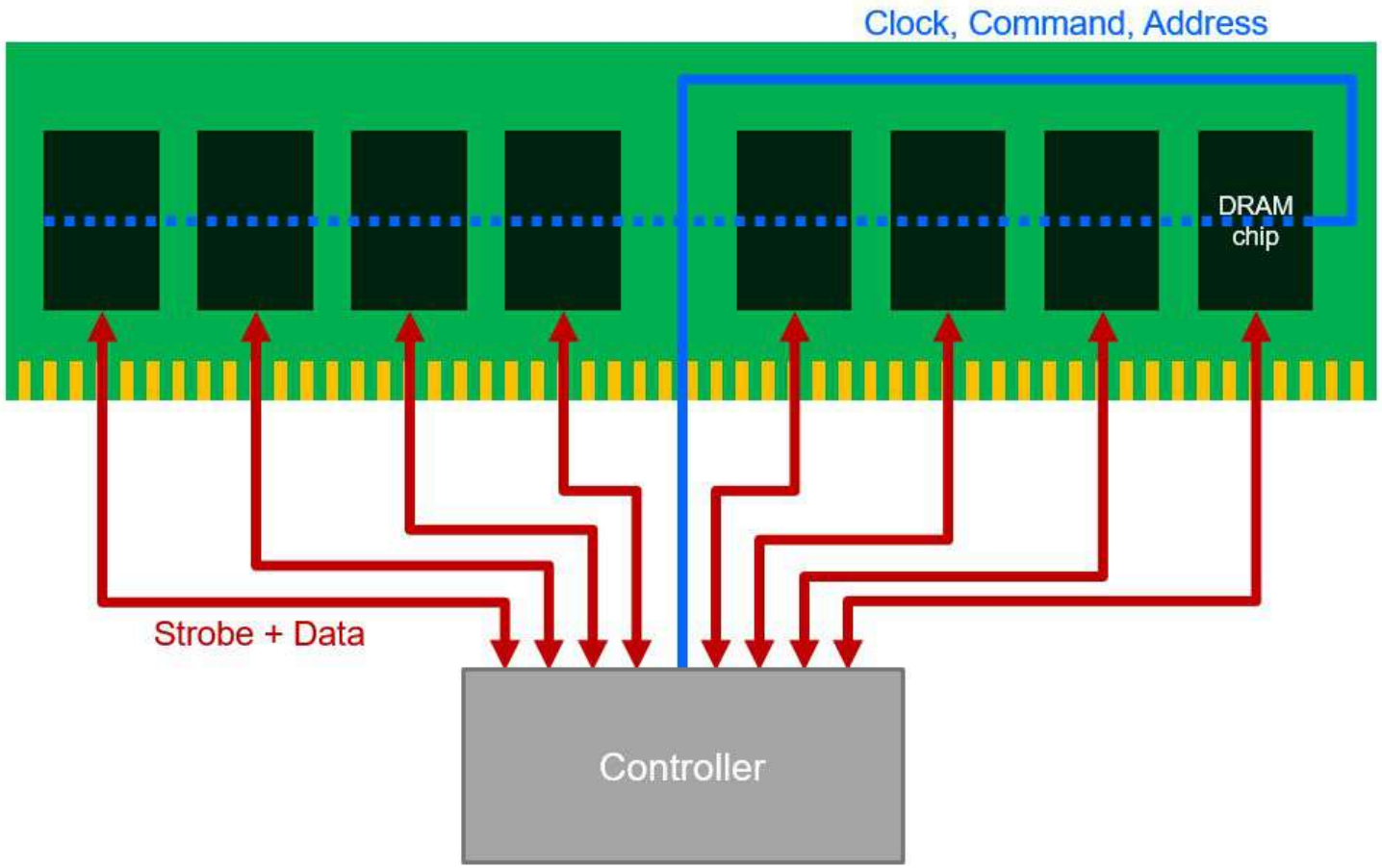
Velocidade das Memórias DDR

- **DDR2** permite leitura de 4 palavras de 64 bits (dobro)
- **DDR3** permite leitura de 8 palavras de 64 bits (2× DDR2)
- A cada geração a frequência do **barramento** da memória DDR fica 2× mais rápido!
  - Frequência interna muda pouquíssimo (ou nem aumenta)

DDR SDRAM Standard	Internal rate (MHz)	Bus clock (MHz)	Prefetch	Data rate (MT/s)	Transfer rate (GB/s)	Voltage (V)
SDRAM	100-166	100-166	1n	100-166	0.8-1.3	3.3
DDR	133-200	133-200	2n	266-400	2.1-3.2	2.5/2.6
DDR2	133-200	266-400	4n	533-800	4.2-6.4	1.8
DDR3	133-200	533-800	8n	1066-1600	8.5-14.9	1.35/1.5
DDR4	133-200	1066-1600	8n	2133-3200	17-21.3	1.2

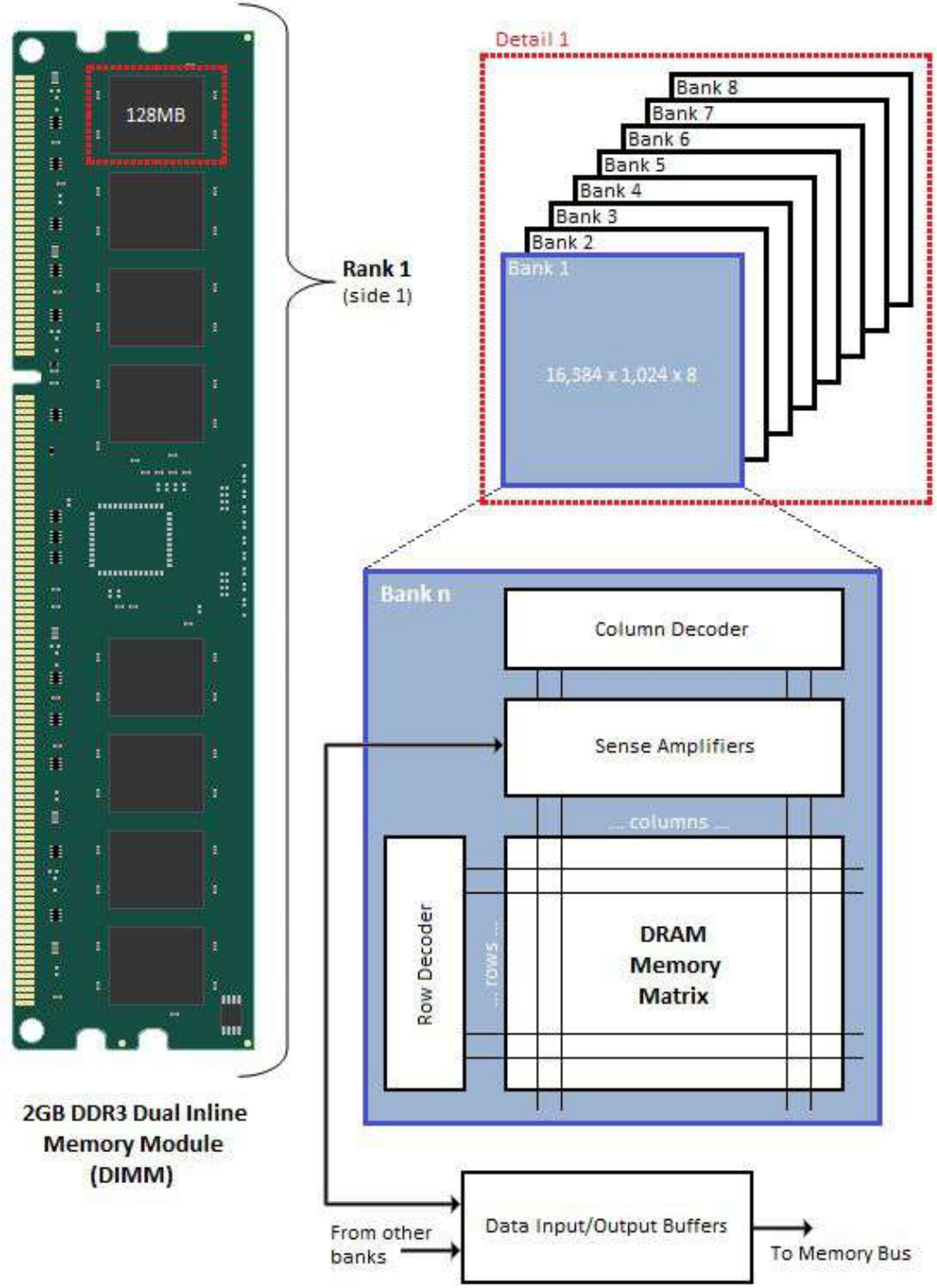
Chips de Memória DRAM em um Pente

- Pente de Memória: Coleção de chips DRAM
  - Cada chip fornece 8 bits de dados (8 chips × 8 bits = 64 bits)
- Dado de 64 bits “montado” pelo controlador de memória ⟶ Linha da cache!



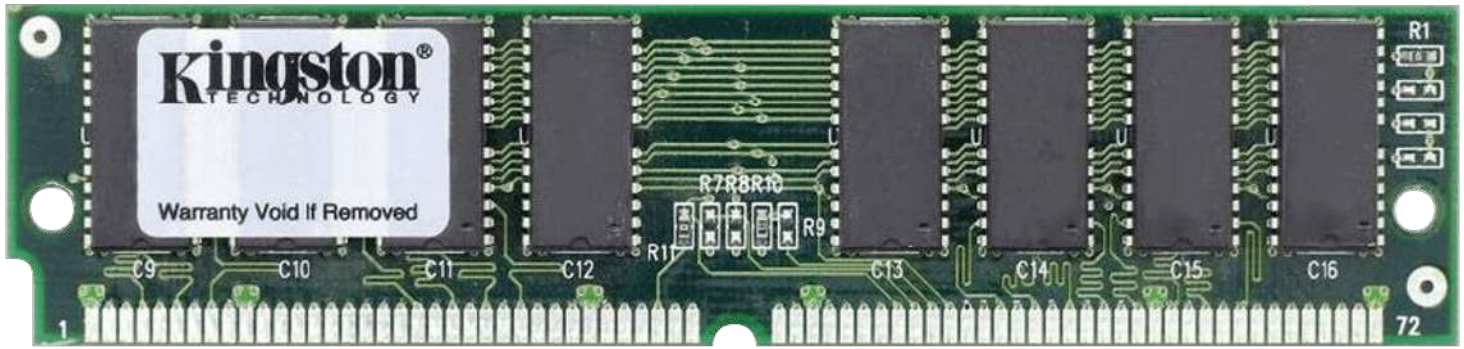
Memória Intercalada

- **Rank**: Conjunto de chips que participam da transferência (paralelismo)
  - Transferência de 64 bits envolvendo 8 chips de 8 bits
  - Pentes modernos possuem 2 ranks (DIMM)
  - Ex: Memória de 2GB dividida em 16 chips de 128MB
- **Bank**: Agrupamento de múltiplos chips em um só
  - Cada banco realiza operações de forma independente
  - Ex: Um chip de 128MB dividido em 8 bancos de 16MB
- **Requisição intercalada** de 8 palavras de 8 bancos
  - (no DDR3)
  - Alto Paralelismo: Múltiplas requisições simultâneas
  - 1ª palavra requisitada ao banco 1, 2ª ao banco 2, …
  - Não espera um terminar para requisitar o próximo



Módulos de Memória SIMM

- *Single In-line Memory Module*: Pinos apenas de um lado
  - Conjunto de chips DIP de memória montados em uma placa de circuito impresso (PCB)
  - Permite até 32 bits por acesso
- Variantes de 30 pinos ou 72 pinos
  - 72 pinos permite mais chips e maior largura de dados
  - “Notch” (corte) impede o chip novo de usar um *slot* antigo

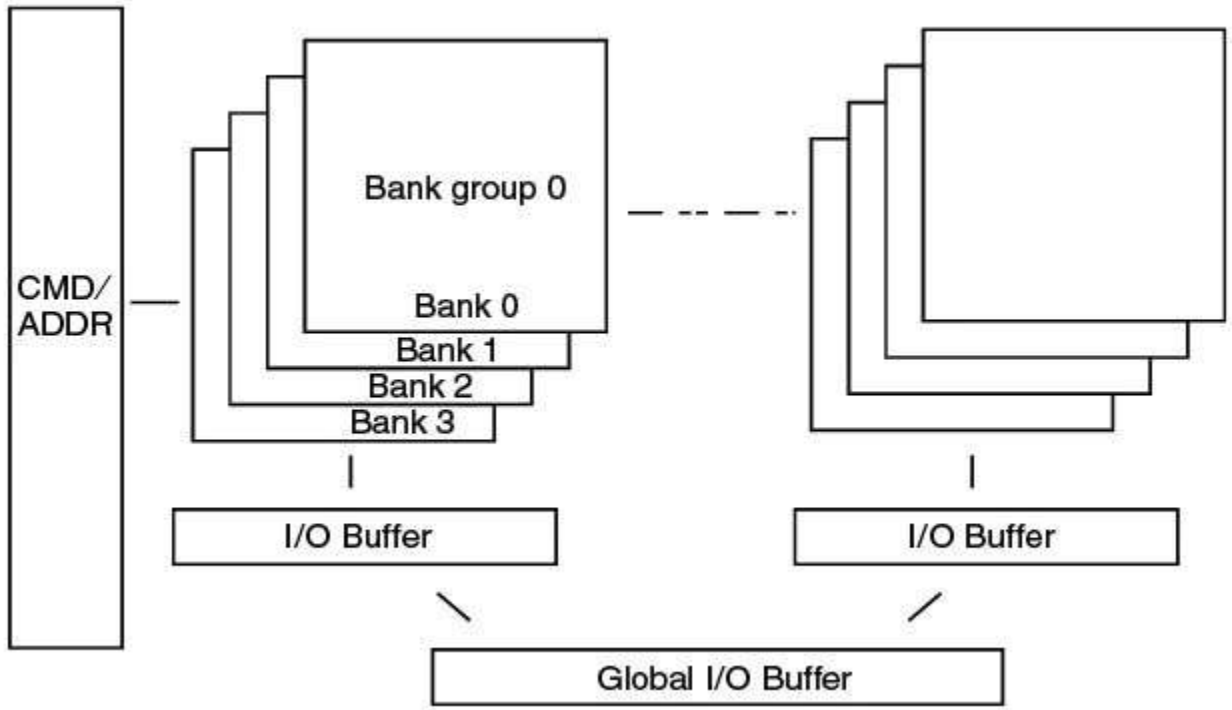


Módulos de Memória DIMM

- *Dual In-line Memory Module*: Pinos de ambos os lados (e.g. 288 pinos para DDR4)
  - Contém dois **ranks**: Requisição do **rank 1** enquanto transfere o **rank 0**
  - Usado em desktops e servidores: Maior performance, capacidade e suporte à ECC
  - Usado com memórias SDRAM e DDR: Permite 64 bits por ciclo
  - Permite rajada de até 8 transferências (64 bytes)
- *Small Outline* DIMM: Menos pinos (e.g. 260 pinos para DDR4)
  - Algumas desvantagens em relação ao DIMM: Menos recursos
  - Usado em notebooks: Tamanho reduzido e menor gasto energético

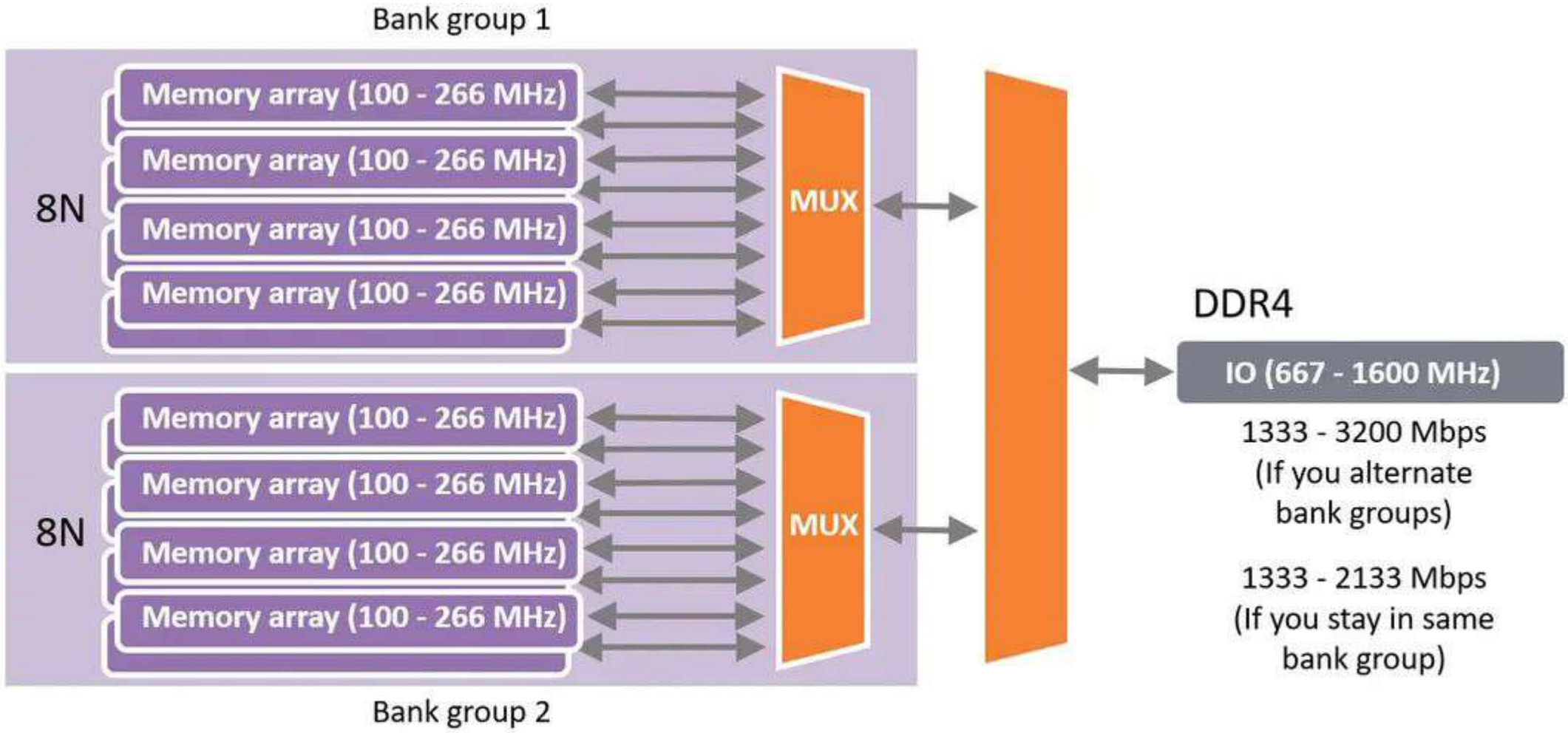
Evolução para o DDR4

- DDR3 permite leitura em **rajada** de 8 palavras de 64 bits (2× DDR2)
  - Até 8 bancos por chip (64 bytes)
- Não foi aumentado o tamanho do *buffer* para 16 no DDR4
  - Linha da cache possui apenas 64 bytes → Busca de uma segunda palavra seria útil?
  - Conceito de **grupos de bancos** → Permite requisições em **paralelo**



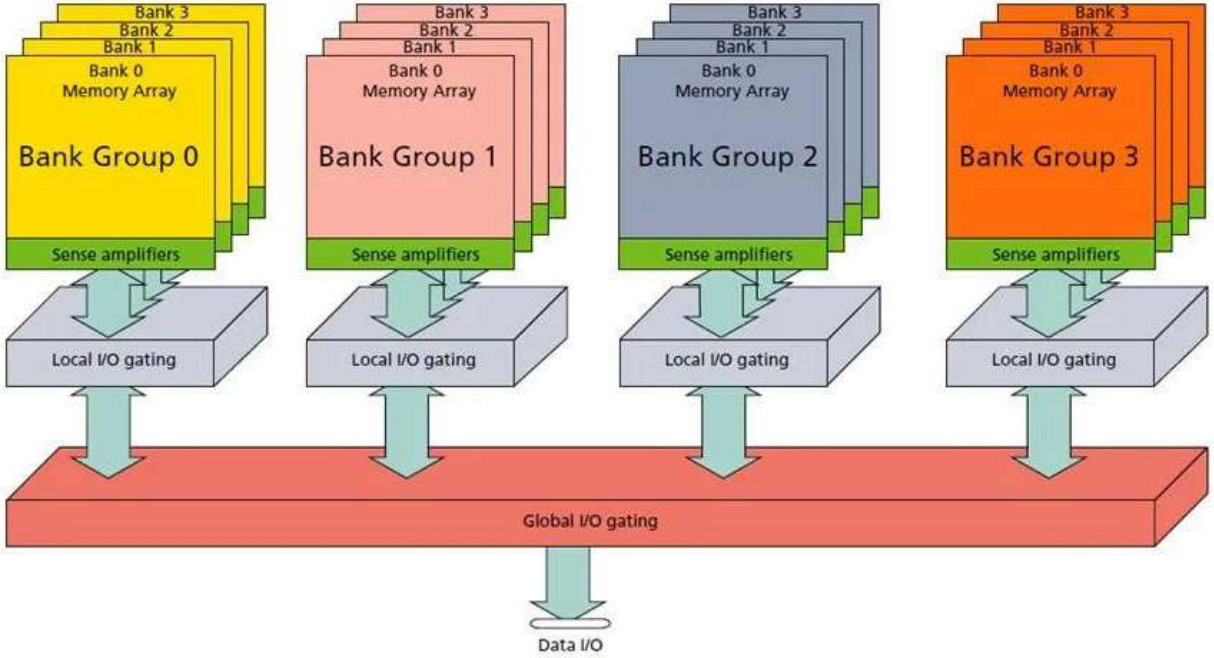
Exemplo: DDR4 com 2 grupos de bancos

- Cada grupo tem um *prefetch buffer* de 8n



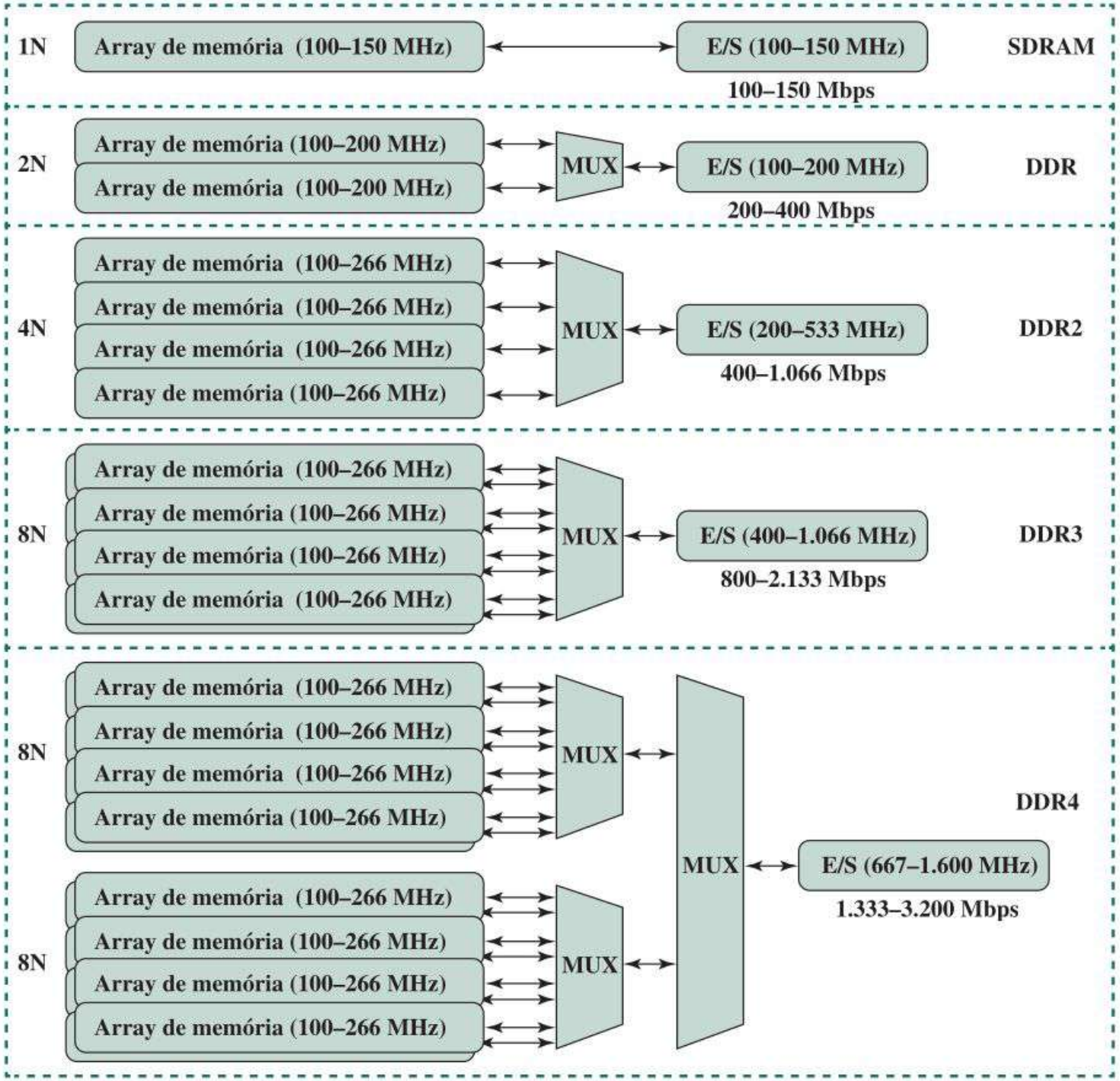
Grupo de Bancos no DDR4

- No DDR3: Requisita operação para primeiro banco, depois para o segundo, e por aí vai
  - **(intercalação lenta)**
- No DDR4: Requisições em paralelo para vários bancos (transferência continua banco-a-banco)
  - **(intercalação rápida)**
  - Até **4 grupos** de **4 bancos** podem ser usados em um *chip*: **16 bancos**



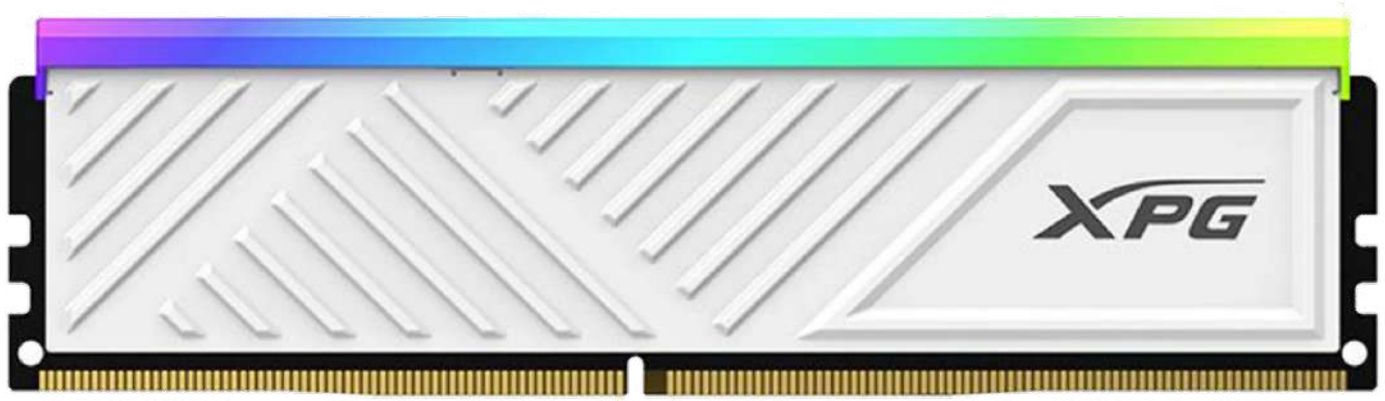
Resumo das Gerações

- **DDR** permite leitura de 2 palavras de 64 bits (**2× SDRAM**)
- **DDR2** permite leitura de 4 palavras de 64 bits (**2× DDR**)
- **DDR3** permite leitura de 8 palavras de 64 bits (**2× DDR2**)
- **DDR4** permite leitura de
  - …8 palavras de 64 bits (8×8B)
  - …4 palavras de 64 bits (4×8B)
  - Menor latência, logo **2× DDR3**
- E o **DDR5**? Calma…



Exemplo: Memória DDR4 SDRAM de 8GB

- **Pente** DDR4 de 8GB dividido em 2 **ranks** de 16 **chips** de 512MB cada
- Cada chip fornece **8 bits** → Possuem **16 bancos**, divididos em **4 grupos** de **4 bancos**
  - Burst de 8 palavras de 64 bits **cada** → 64 bytes
- Pode requisitar de uma vez…
  - Até **4 requisições** intercaladas em paralelo para cada **grupo de bancos** (intercalação rápida)
  - Então, **4 requisições** intercaladas sequenciais para cada **banco** do grupo (intercalação lenta)

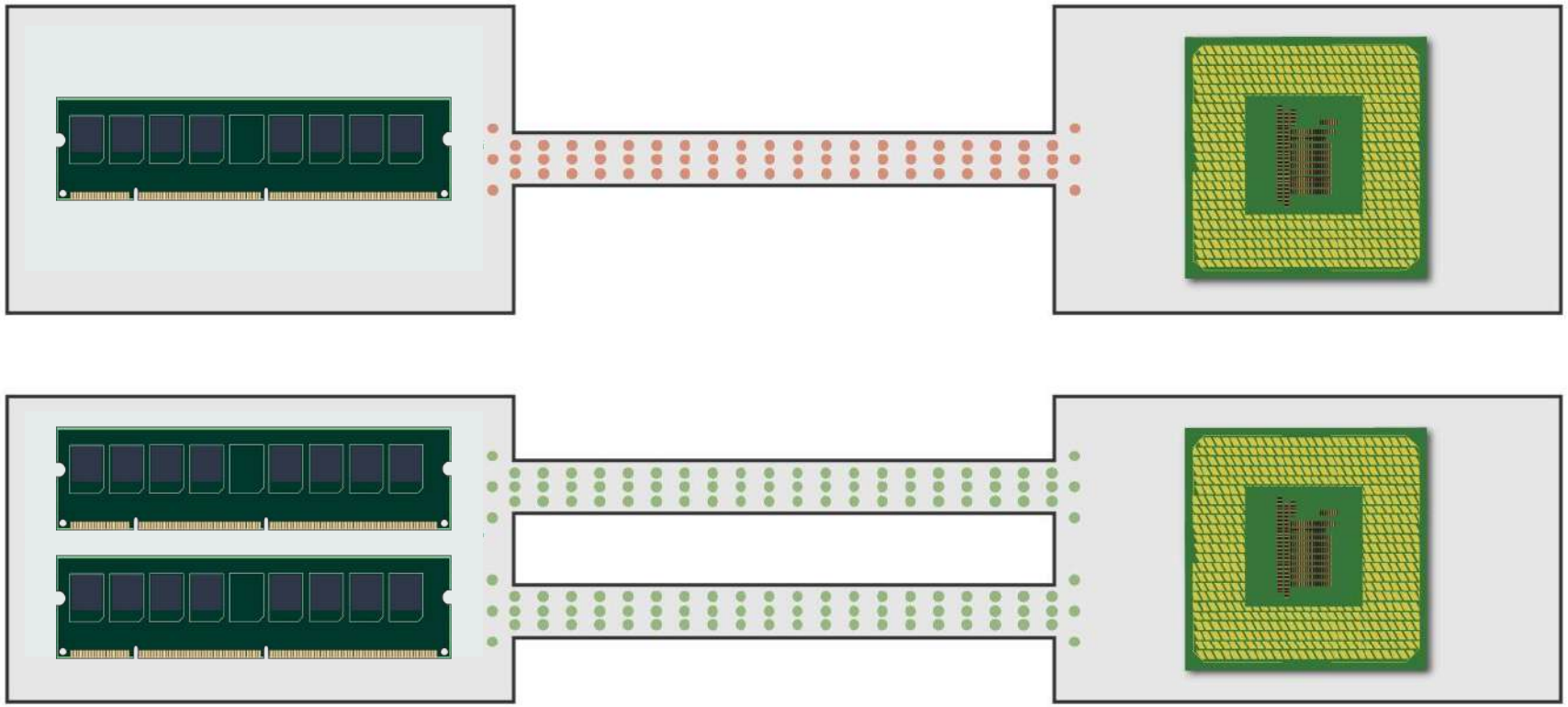


Gerações das Memórias DDR

DDR SDRAM Standard	Bus clock (MHz)	Internal rate (MHz)	Prefetch (# burst)	Max Transfer Rate (GB/s)	First release
 **DDR**	100–200	100–200	2n	3.2	2000
 **DDR2**	200–533	100–266	4n	8.5	2004
 **DDR3**	400–1066	200-533	8n	17.0	2007
 **DDR4**	800-1600	200-400	8n	25.6	2011
 **DDR5**	1600-3200	400-800	16n	51.2	2020

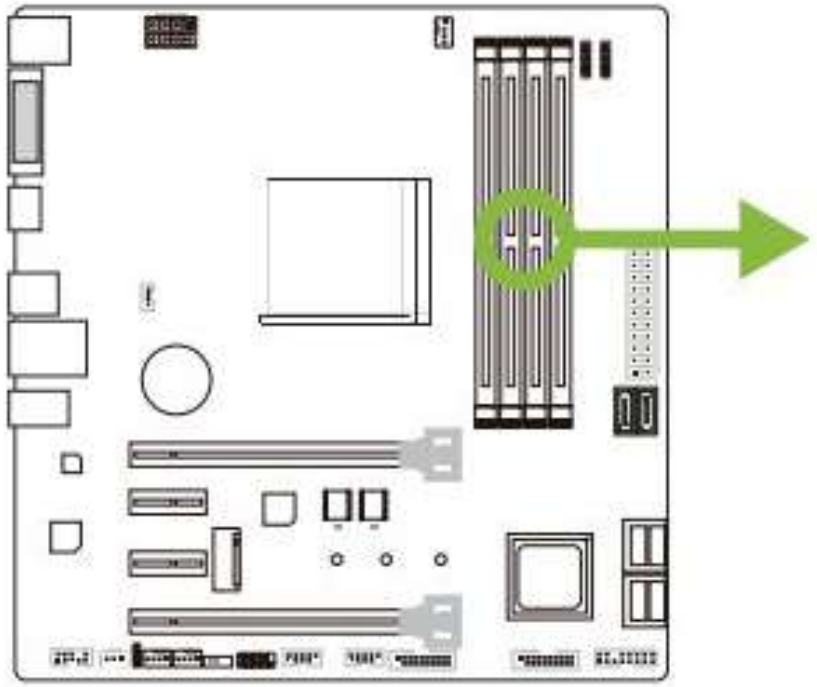
Dual Channel

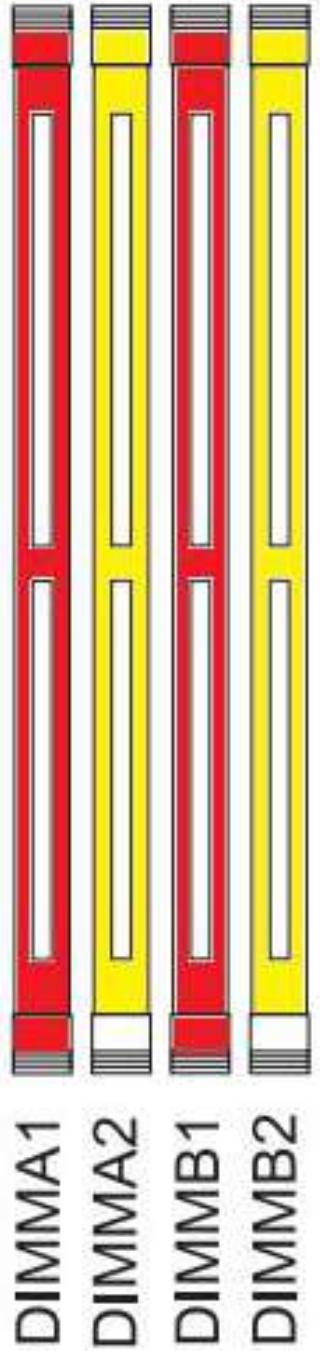
- Permite a CPU a acessar dados de **dois módulos de RAM** ao mesmo tempo
  - **Single Channel**: um pente⟶ 64 bits de barramento de dados
  - **Dual Channel**: dois pentes⟶ 128 bits de barramento de dados (duas portas **independentes**)
- Duplica a **largura de banda** ao invés da frequência (como no DDR)



Restrições do Dual Channel

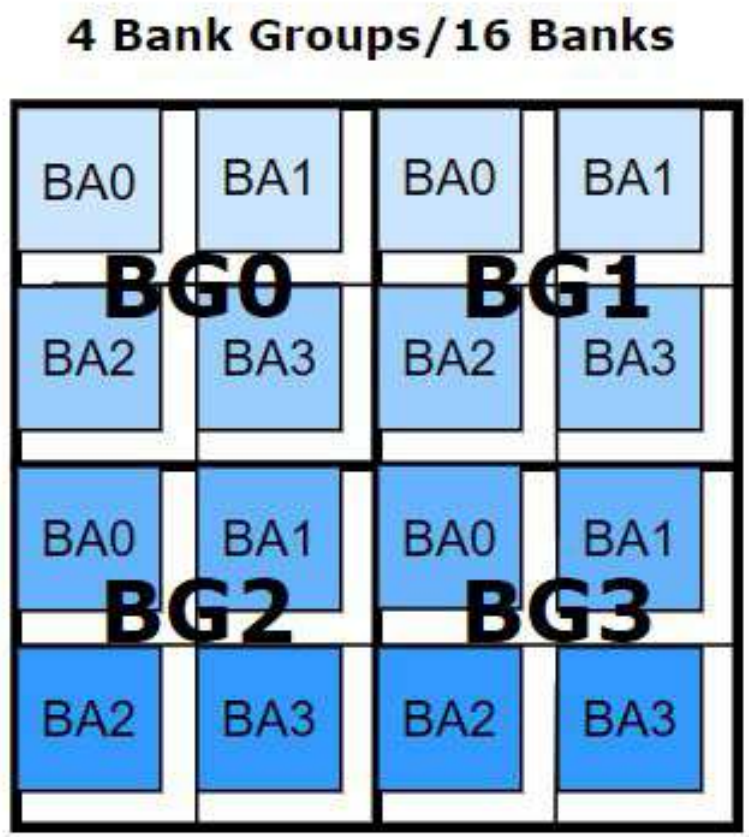
- Idealmente, os dois módulos devem ter a mesma **capacidade**
  - Alguns sistemas suportam modo híbrido (flex)
  - Se tamanhos forem diferentes, uma parte funciona em dual channel e o restante em single channel
- Ambos devem ter a mesma **frequência**
  - Se não, usa a **frequência menor** para ambos
- É recomendado que os dois módulos tenham a mesma **latência** (ex: CL16) para evitar instabilidades
- Os módulos devem ser da mesma geração

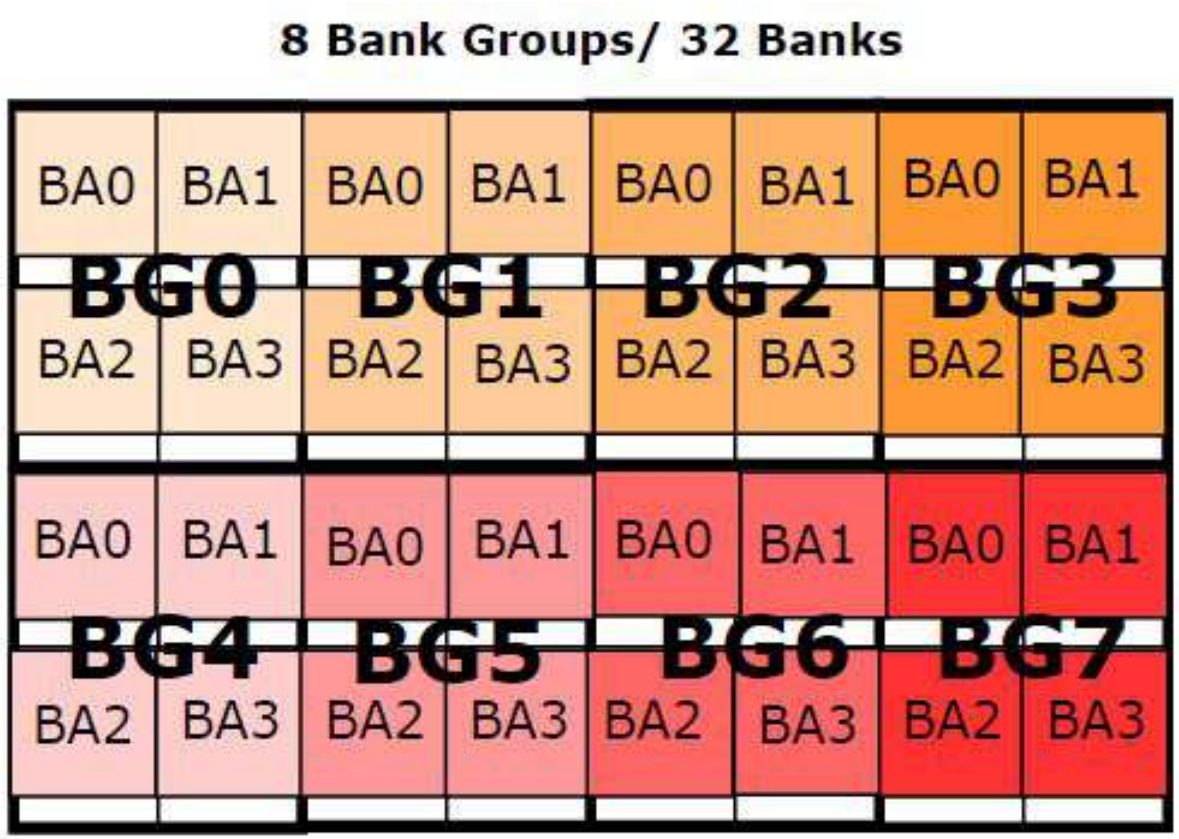




E o DDR5?

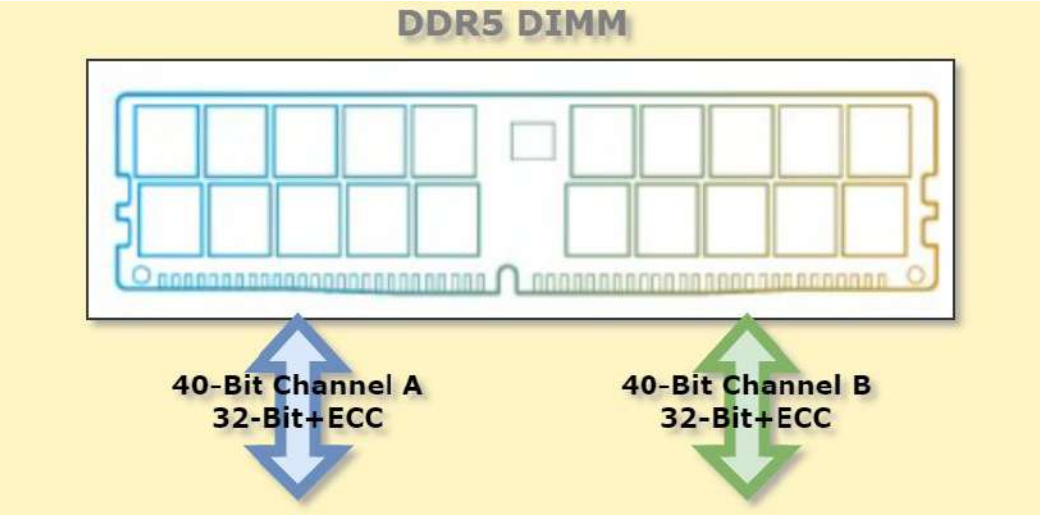
- DDR5 permite até 8 grupos de bancos, com 4 bancos cada
- Como obter *speedup* de 2x em relação ao DDR4?



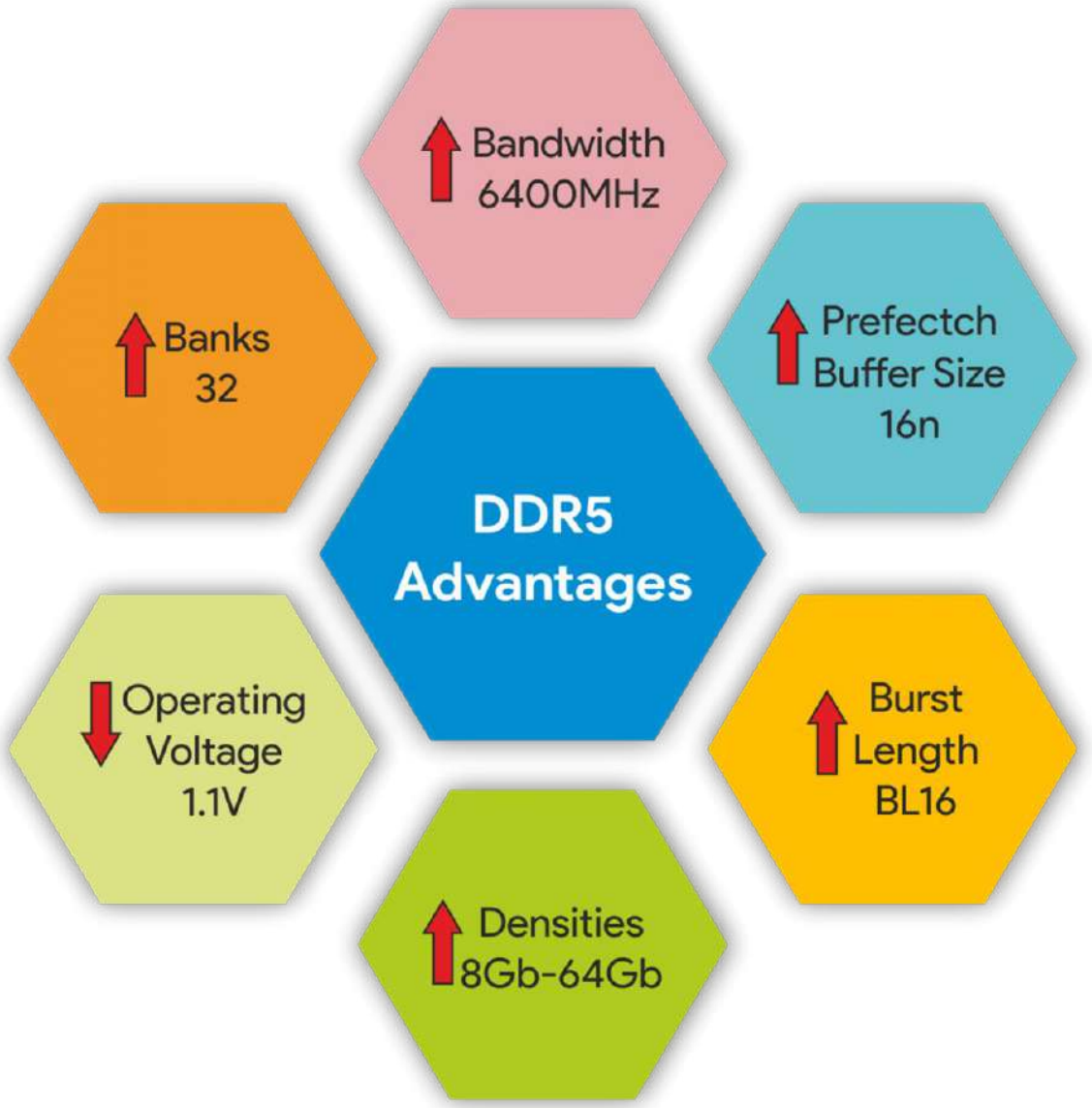


Channel Splitting no Memória DDR5

- Divide o barramento de 64 bits em **dois barramentos independentes de 32 bits**
  - Cada canal oferece 32 bits (64/2), permitindo um prefetch buffer de 16n (8n×2) e rajada de 64 bytes
  - **DDR4**: 64 bits (barramento) × 8 transferências (BL8) = 64 bytes (a 3200 MT/s)
  - **DDR5**: 32 bits (sub-canal) × 16 transferências (BL16) = 64 bytes (a 6400 MT/s)
- Suporta então **dois canais** por módulo (*dual channel per* DIMM):
  - **DDR4**: Canal singular de 72 bits (64 bits para dados e 8 bits para ECC)
  - **DDR5**: Canal duplo de 80 bits (32 bits para dados e 8 bits para ECC em ambos)



Vantagens da Memória DDR5

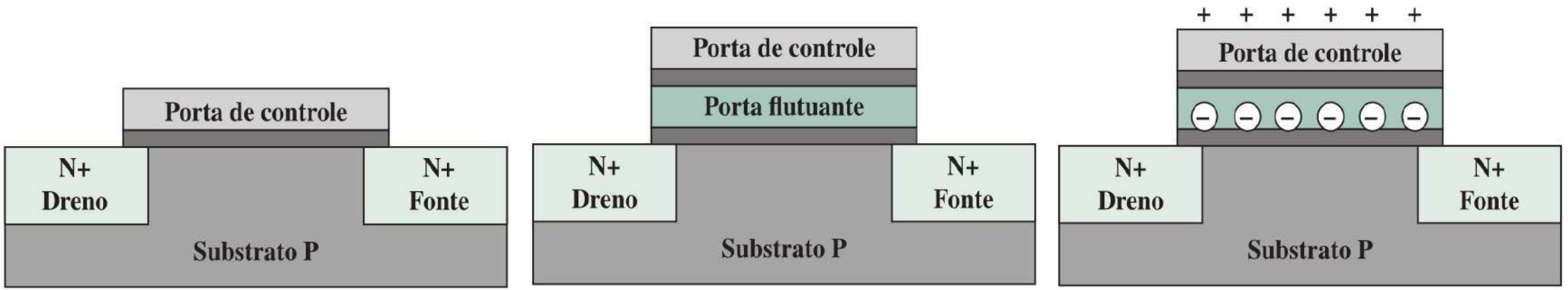


Memória Interna Moderna

Memória Flash

Memória Flash

- Criado pela Toshiba na decada de 1980 a partir de um **erro eletrônico** em transistores
  - Elétrons presos (portadora quente), ou não, mantém **dois estados**: Célula não-volátil
  - Impedem fluxo de elétrons (estado 0) ↔Operação normal (estado 1)
- Alta densidade: Um transistor flash por bit
- Dois tipos: Usada tanto como memória interna e externa
  - **NAND**: Acesso em **blocos** (páginas), alta velocidade de **escrita**, mais **barato** (memória externa)
  - **NOR**: Acesso em **byte**, alta velocidade de **leitura**, mais **caro** (memória interna)



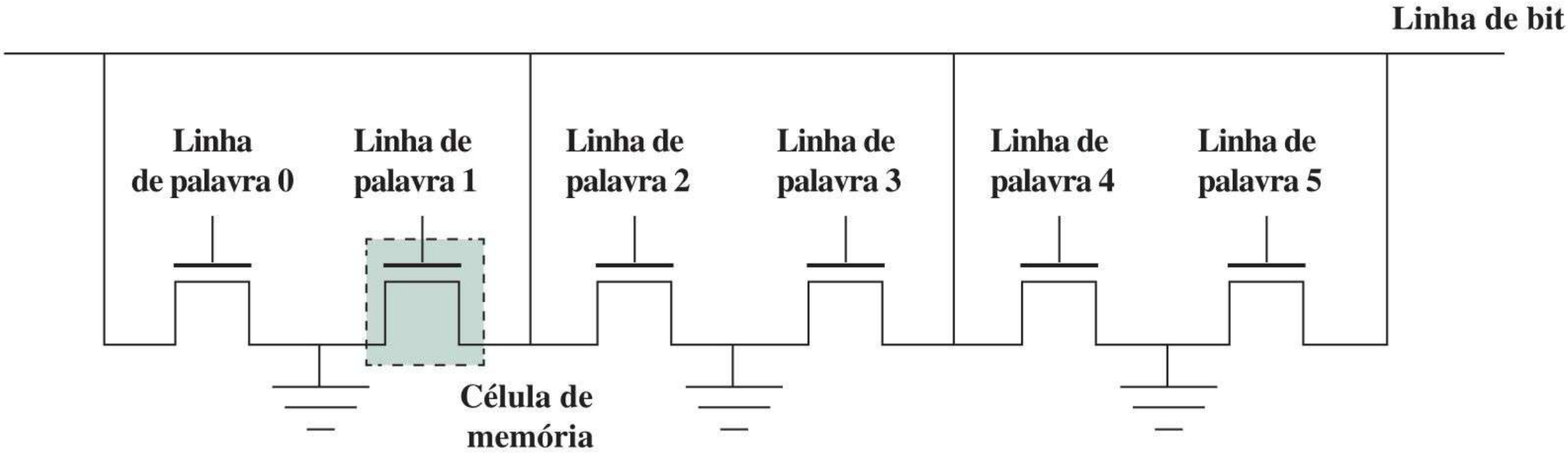
(a) Estrutura do transistor

(b) Célula de memória flash no estado 1

(c) Célula de memória flash no estado 0

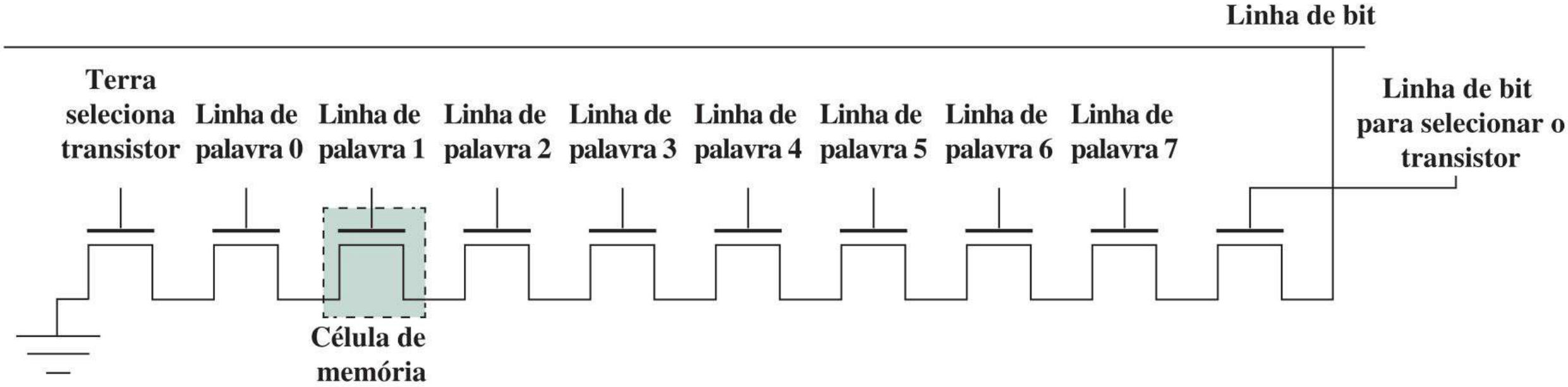
Memória Flash NOR

- Células conectadas em paralelo: Lidas, gravadas e apagadas individualmente
- Linha de Palavra:
  - Seleciona célula para leitura



Memória Flash NAND

- Células conectadas em série: 16 ou 32 transistores em série
  - Linha de bit só vai pra zero se todos forem ligados (comportamento da porta NAND)
- Leitura de uma célula: Todas as outras são selecionadas e a linha é verificada
  - Seleciona célula para leitura

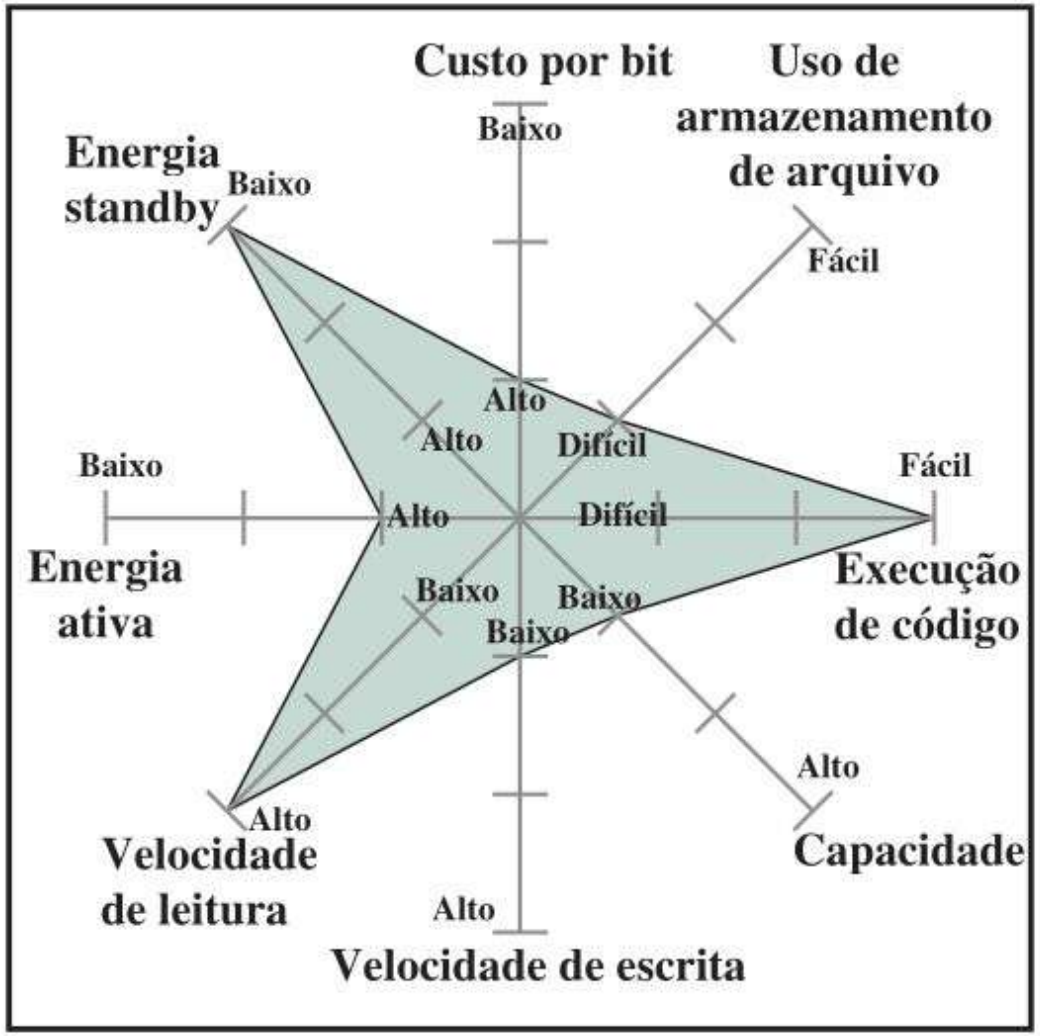


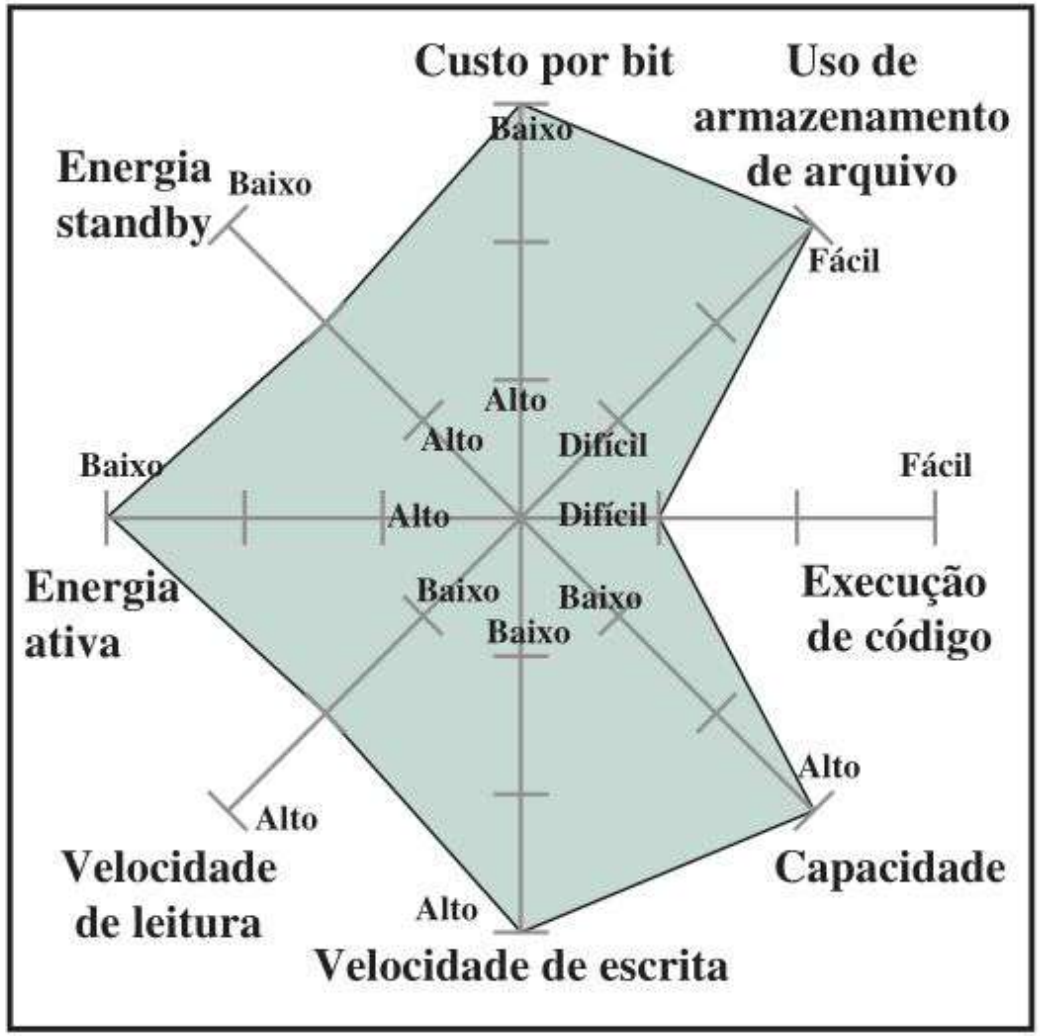
Comparação dos Tipos de Memória Flash

- **Flash NOR**: Acesso aleatório de alta velocidade
  - Permite leitura/escrita a nível de byte
  - Muito usada como memória interna para sistemas embarcados
- **Flash NAND**: Acesso baseado em **blocos** com vários bytes
  - Blocos divididos em **páginas**: Podem ser escritas e lidas (apagamento apenas a nível de bloco)
  - Velocidade mais alta de gravação e densidade maior
  - Mais adequada para memória externa e.g. pendrives, cartões de memória, SSDs.

Operation	Area
Read	Page
Program (Write)	Page
Erase	Block

Comparação dos Tipos de Memória Flash





Implementações de Memória Flash NAND

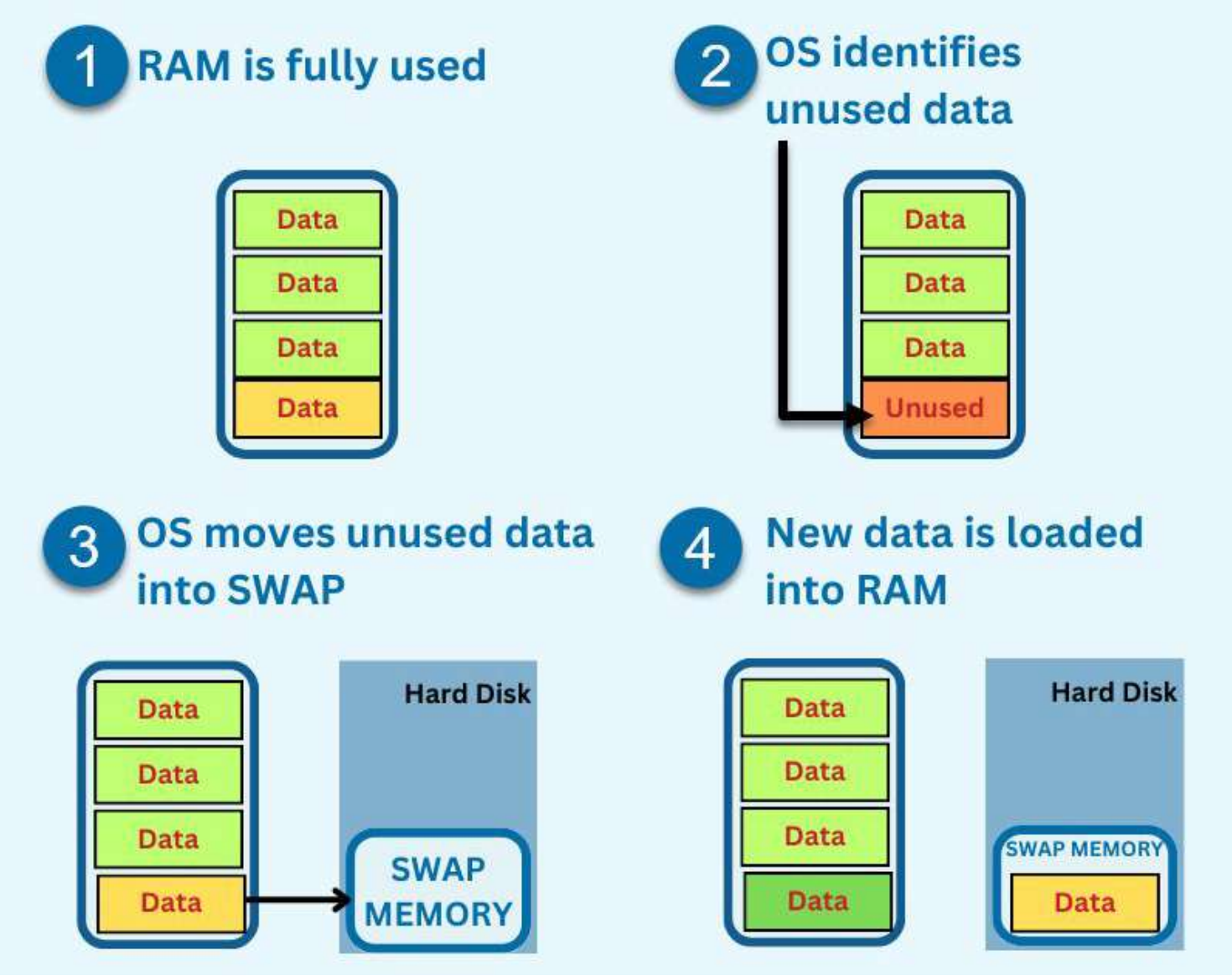
Memória Interna Moderna

Aplicações e Estudo de Caso

Caso Geral: Tamanho ou Velocidade?

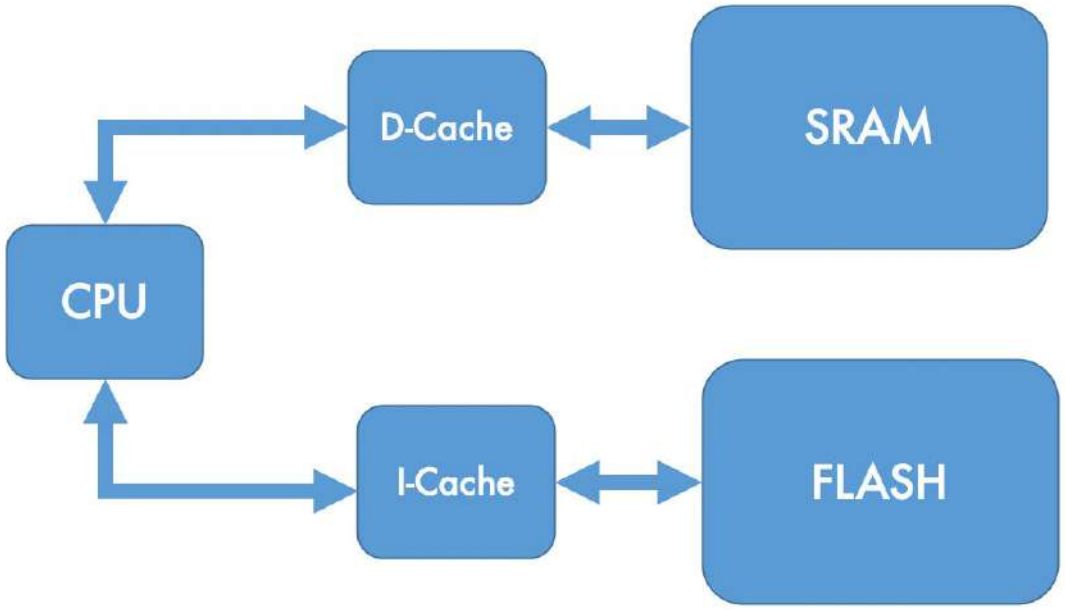
- É melhor ter uma memória **maior** ou uma memória **mais rápida**?
- Resposta: Depende
  - Primeiro, tenha RAM suficiente (maior capacidade)
  - Depois, preocupe-se em ter RAM rápida (maior velocidade)
- Se o sistema tiver pouca memória RAM → *Swap* será realizado
  - Dados escritos da RAM para o armazenamento em massa (HDD ou SSD)
  - Muito lento!
- Logo…
  - Tem pouca RAM (frequente uso de *swap*): Mais RAM é melhor
  - Tem RAM suficiente (sem swap): RAM mais rápida pode trazer ganhos

Swap entre RAM e HD



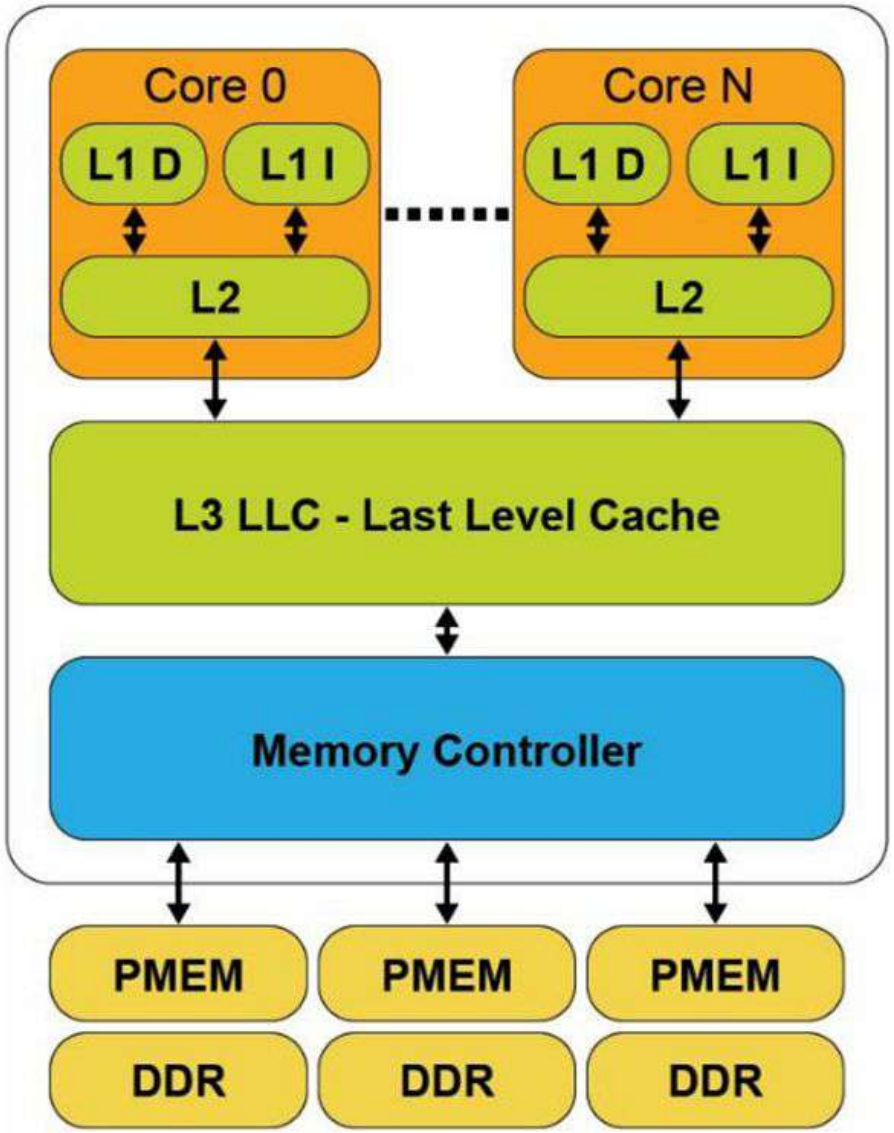
Arquitetura de Von Neumann *versus* Harvard

- Como é implementada a memória principal atualmente?
- **Arquitetura x86** → Primariamente **Von Neumann** (memória unificada)
  - Cache L1 separadas (arquitetura de Harvard?)
- **Arquiteturas Modernas** (ARM, RISC-V) → Primariamente **Harvard** (memória separada)
  - Arquitetura de Harvard Modificada → Endereços mapeados para diferentes memórias

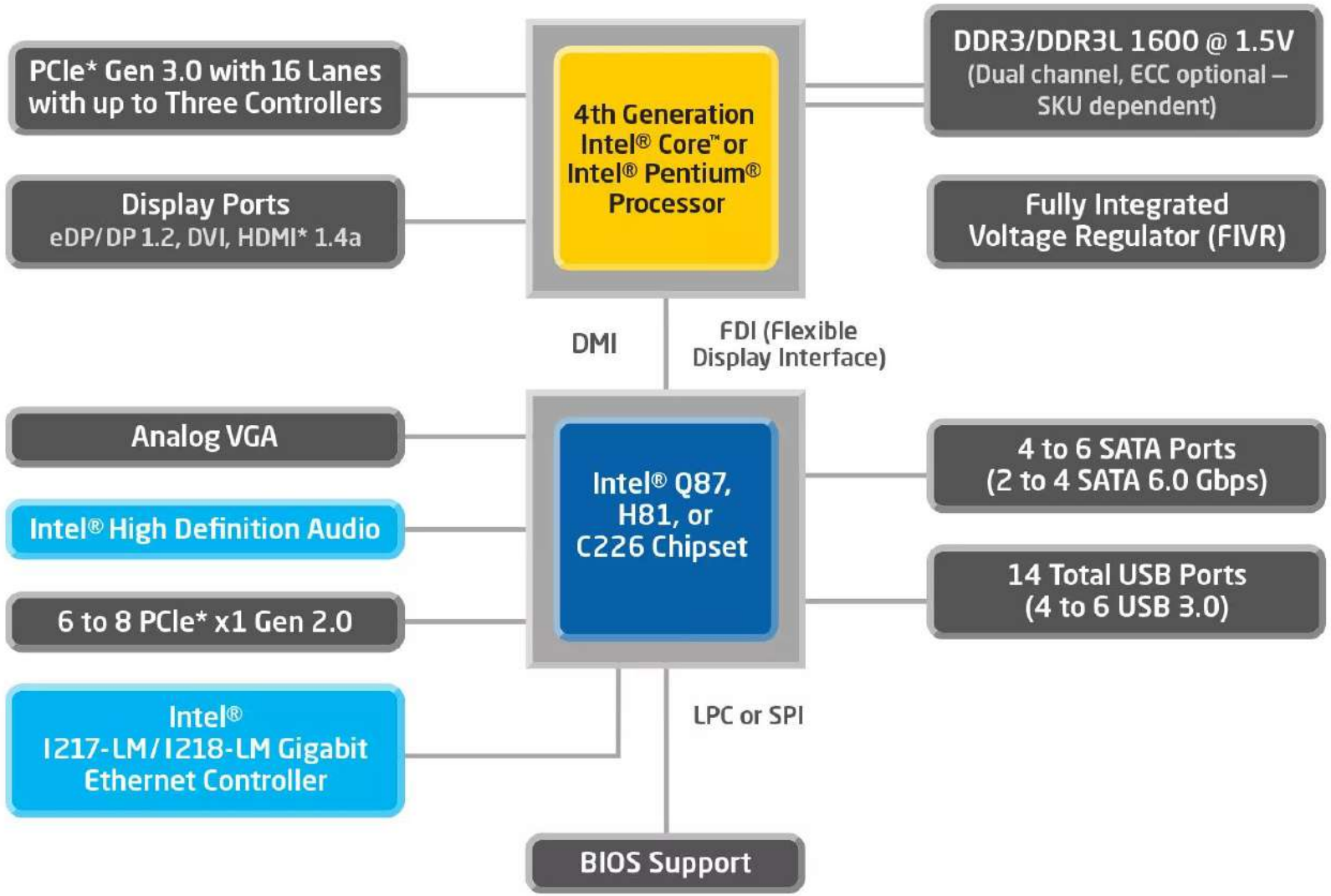


Caso 1: Computadores x86

- Arquitetura de processamento geral
- Acesso à memória baseado em **caches** (SRAM)
  - **L1** separado para dados e instruções
  - **L2** unificada dedicada para um núcleo
  - **L3** unificada compartilhada pelos núcleos
- Memória RAM DDR5 de alta performance
- *Firmware* armazenado em memória não volátil
  - BIOS em ROM (antigamente)
  - UEFI em FLASH (atualmente)



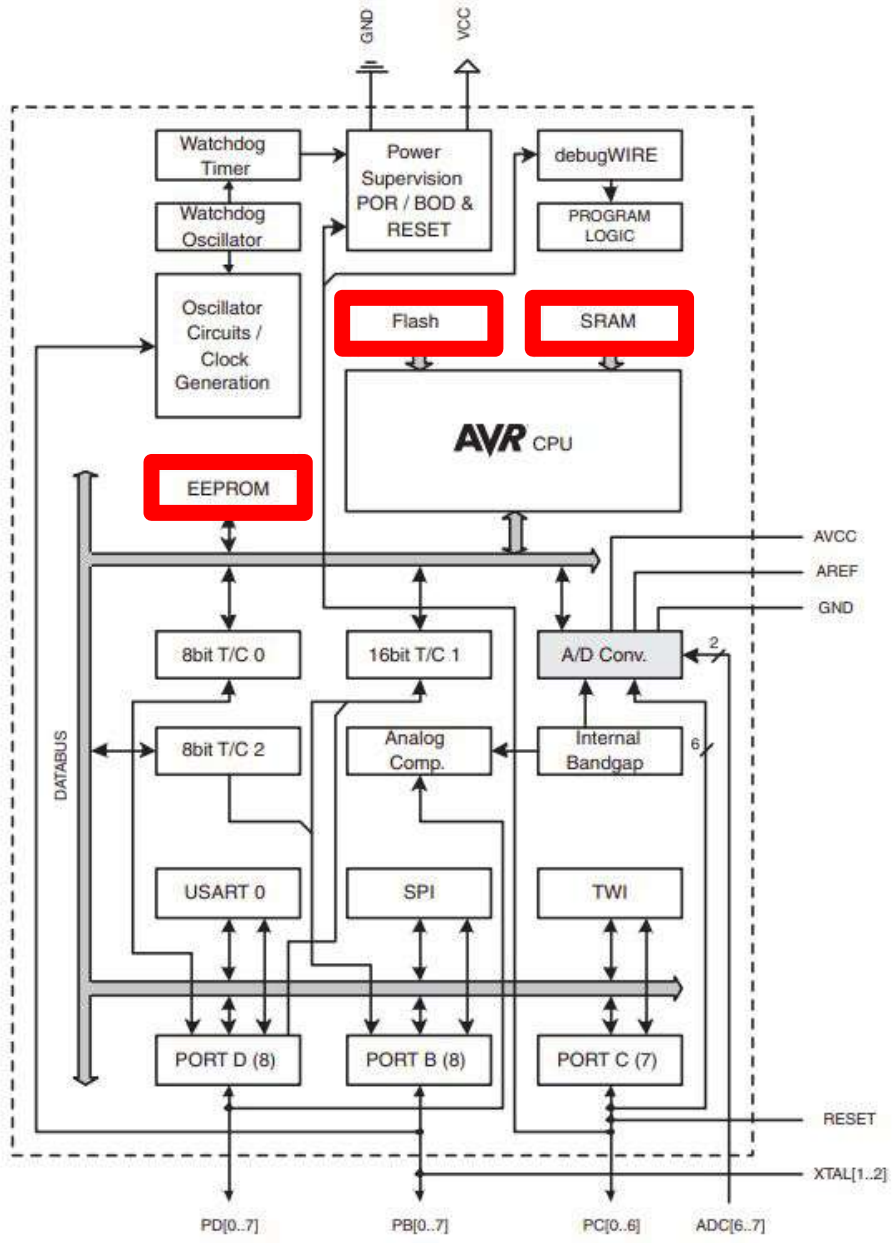
Exemplo: Processador Intel de 4ª Geração



Caso 2: Microcontroladores AVR

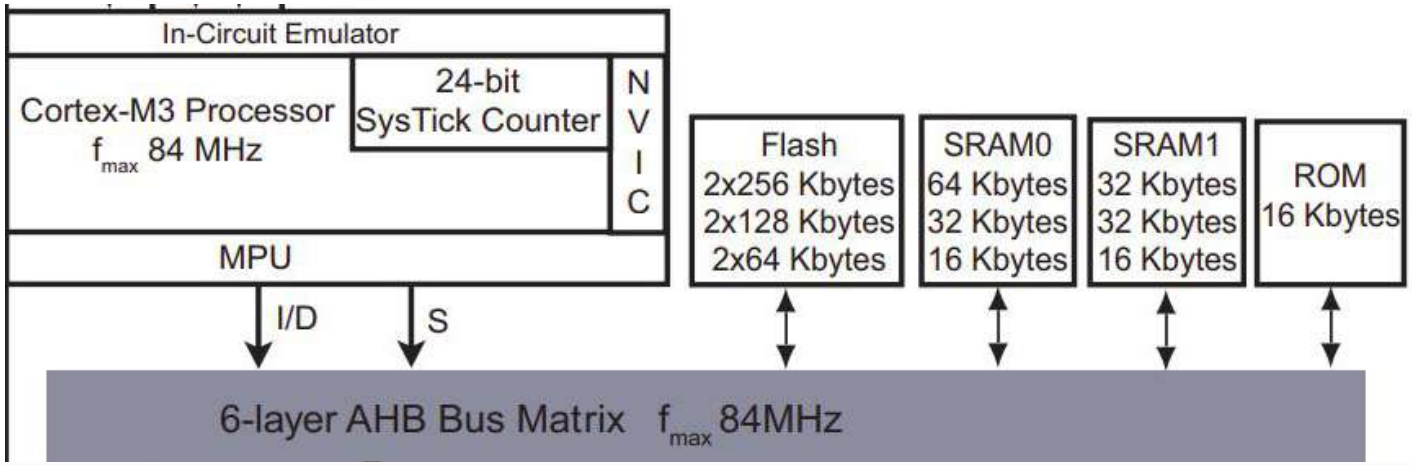
- Microcontroladores de baixa potência
  - Foco em sistemas embarcados simples
  - Sensível ao preço
- Tamanho X Preço (*trade-off*)
- Versões com diferentes configurações e preços:

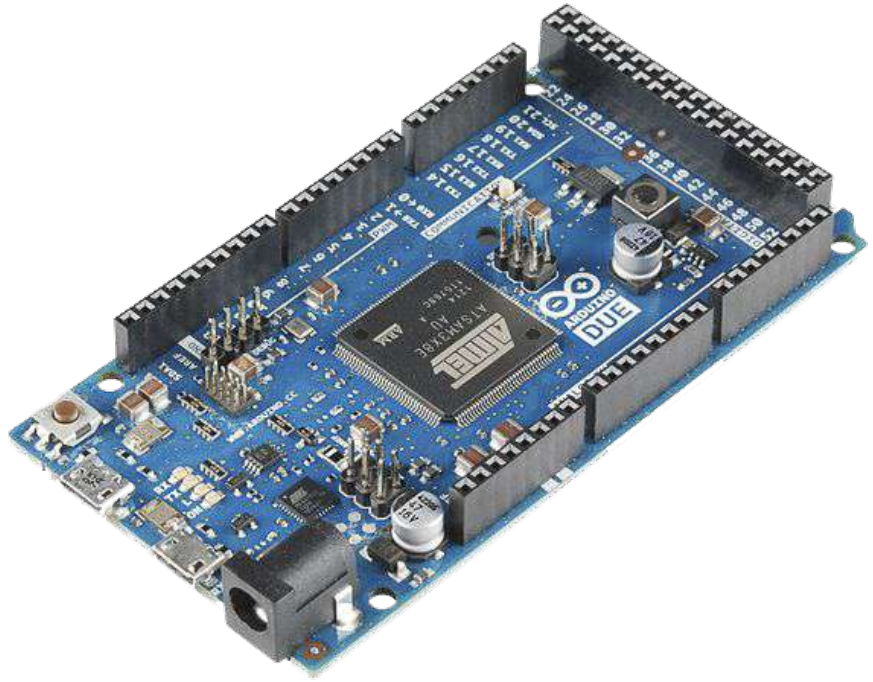
Device	Flash	EEPROM	RAM
ATmega48A	4KBytes	256Bytes	512Bytes
ATmega48PA	4KBytes	256Bytes	512Bytes
ATmega88A	8KBytes	512Bytes	1KBytes
ATmega88PA	8KBytes	512Bytes	1KBytes
ATmega168A	16KBytes	512Bytes	1KBytes
ATmega168PA	16KBytes	512Bytes	1KBytes
ATmega328	32KBytes	1KBytes	2KBytes
ATmega328P	32KBytes	1KBytes	2KBytes



Caso 3: Microcontroladores ARM

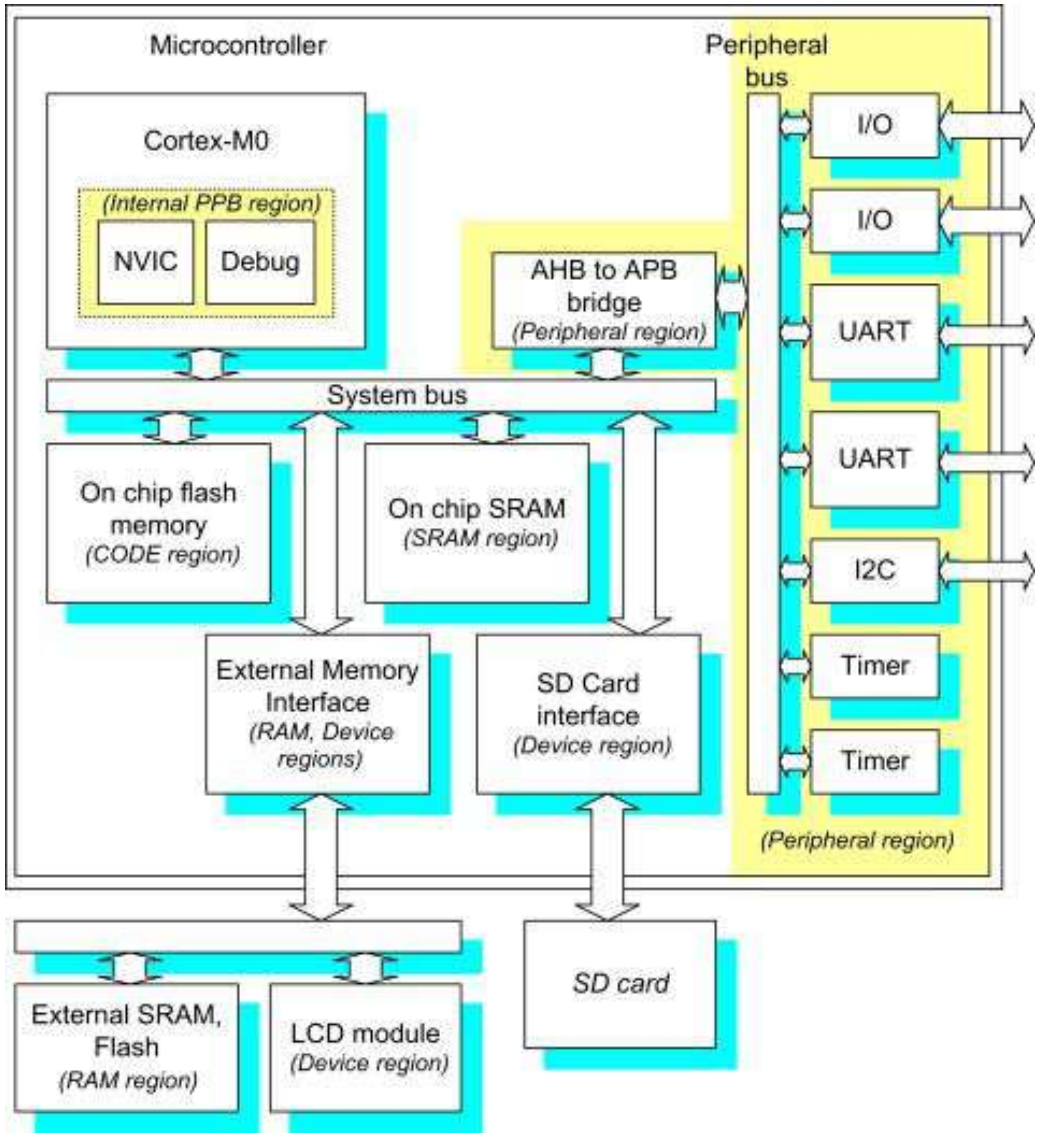
- Microcontroladores de alta performance
  - Foco em sistemas embarcados avançados
  - Processamento X Eficiência Energética
- Código maior →Memória maior (dimensionamento)





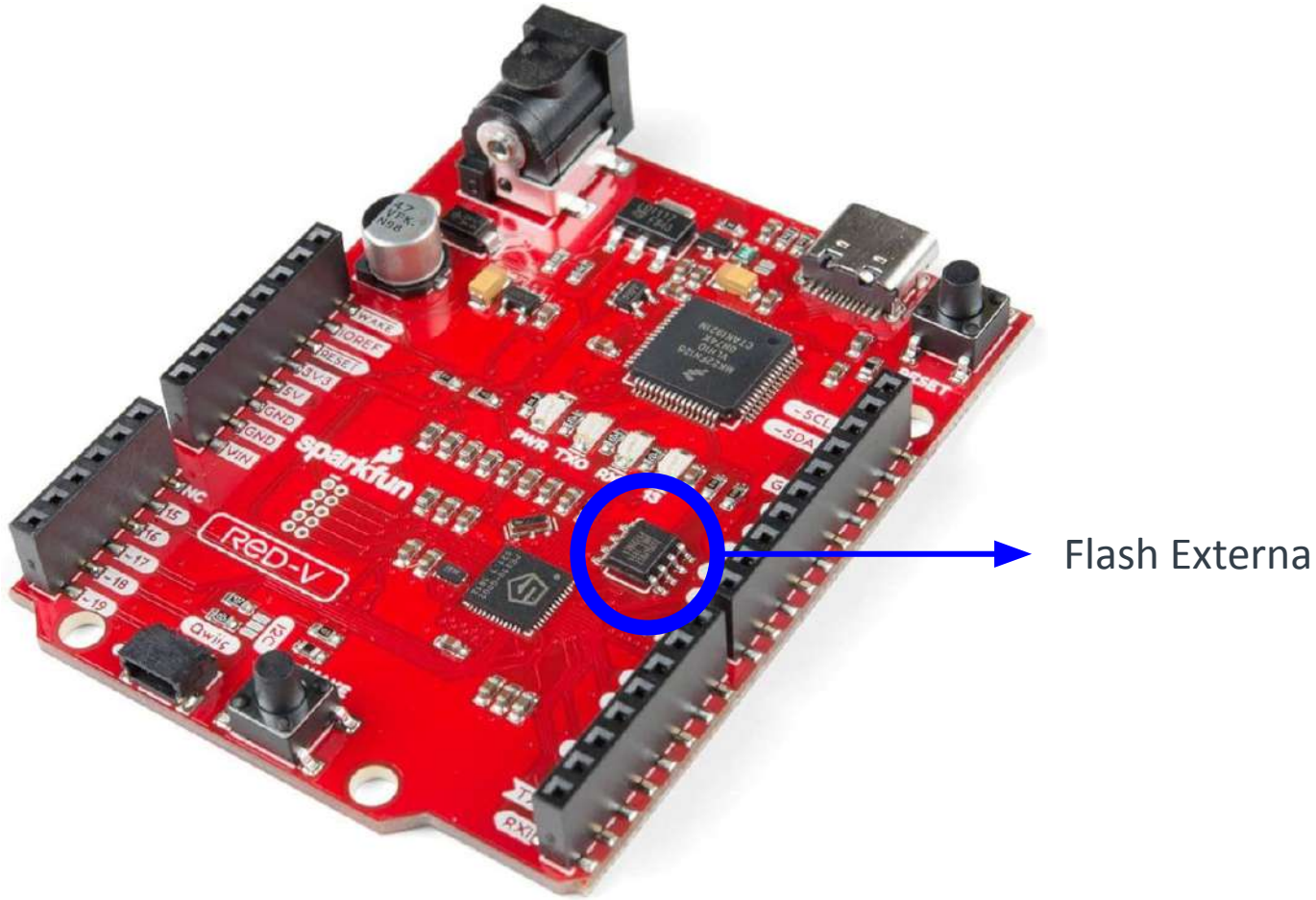
Feature	SAM3X8E	SAM3X8C	SAM3X4E	SAM3X4C	SAM3A8C	SAM3A4C
Flash	2 x 256 Kbytes	2 x 256 Kbytes	2 x 128 Kbytes	2 x 128 Kbytes	2 x 256 Kbytes	2 x 128 Kbytes
SRAM	64 + 32 Kbytes	64 + 32 Kbytes	32 + 32 Kbytes	32 + 32 Kbytes	64 + 32 Kbytes	32 + 32 Kbytes

Exemplo de Memórias de um Microcontrolador Simples

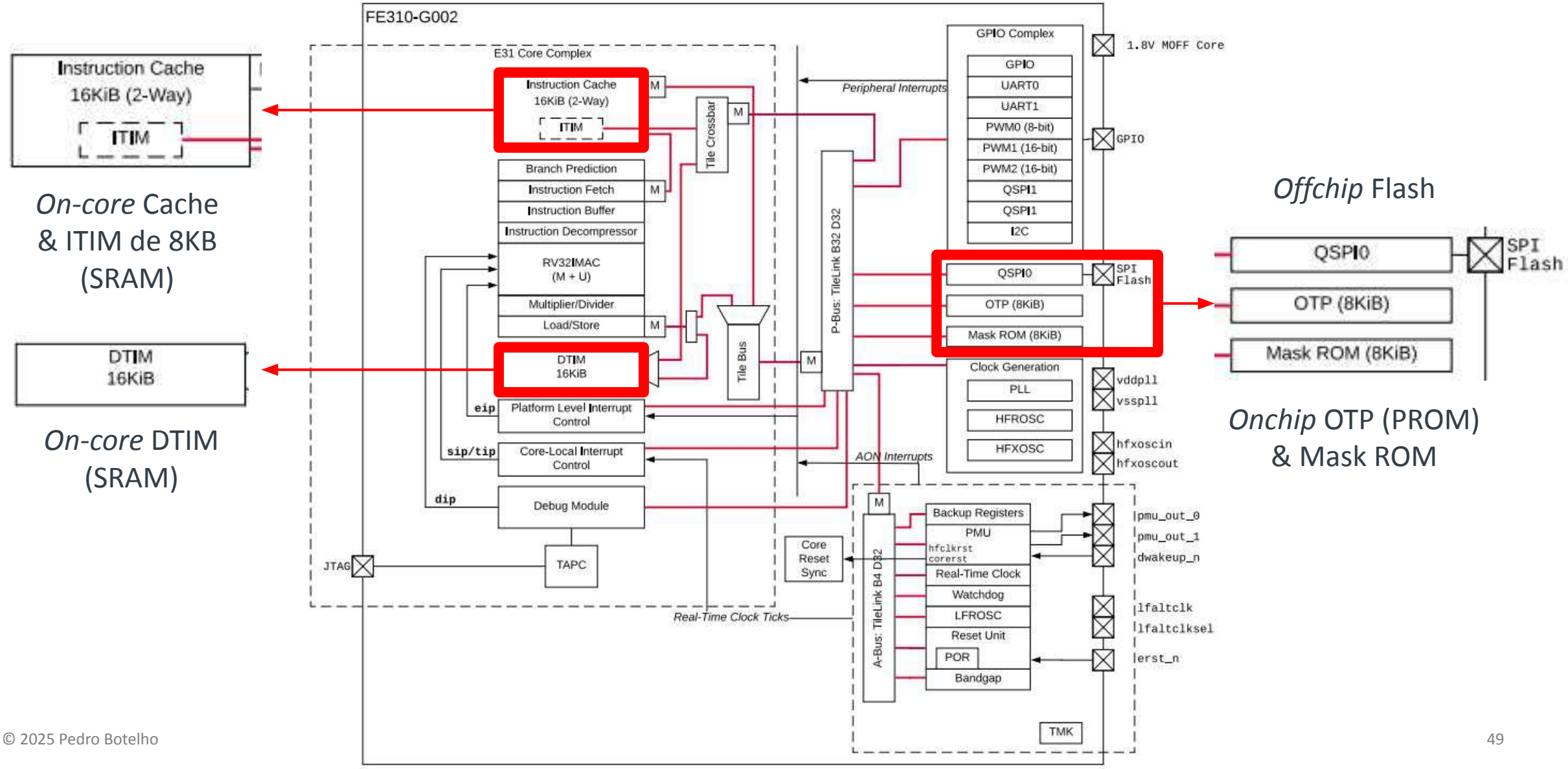


Caso 4: Microcontroladores RISC-V

- Microcontroladores leves, eficientes, personalizáveis e *open-source*
- Ex: Sparkfun RED-V com SiFive FE310 contém Memória Flash externa

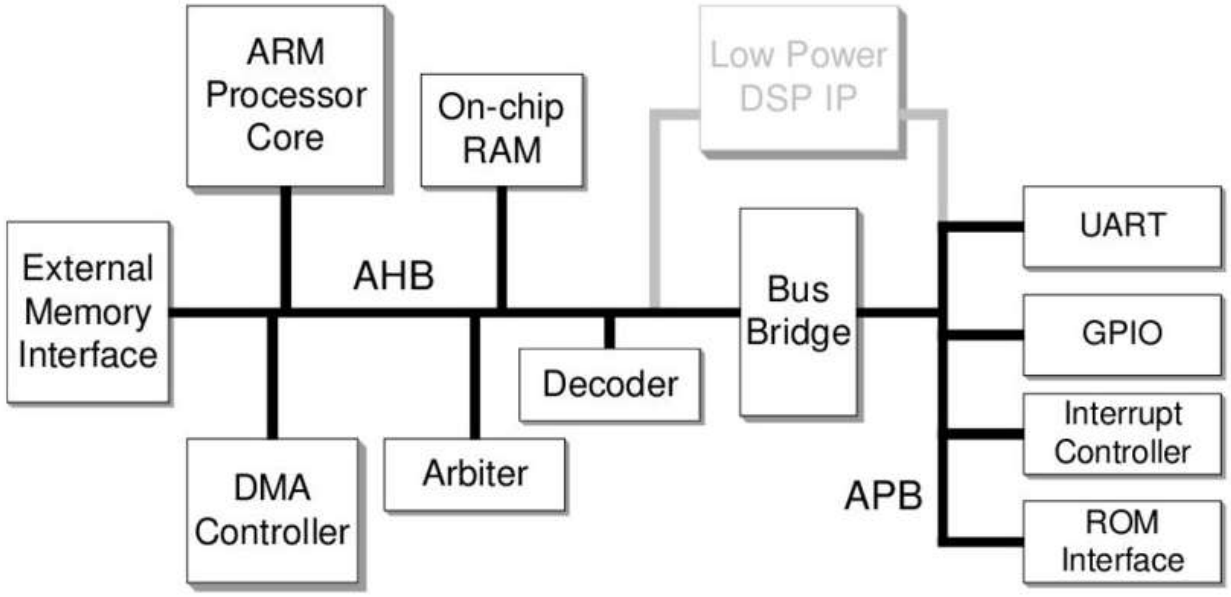


Exemplo: Memórias do Microcontrolador FE310

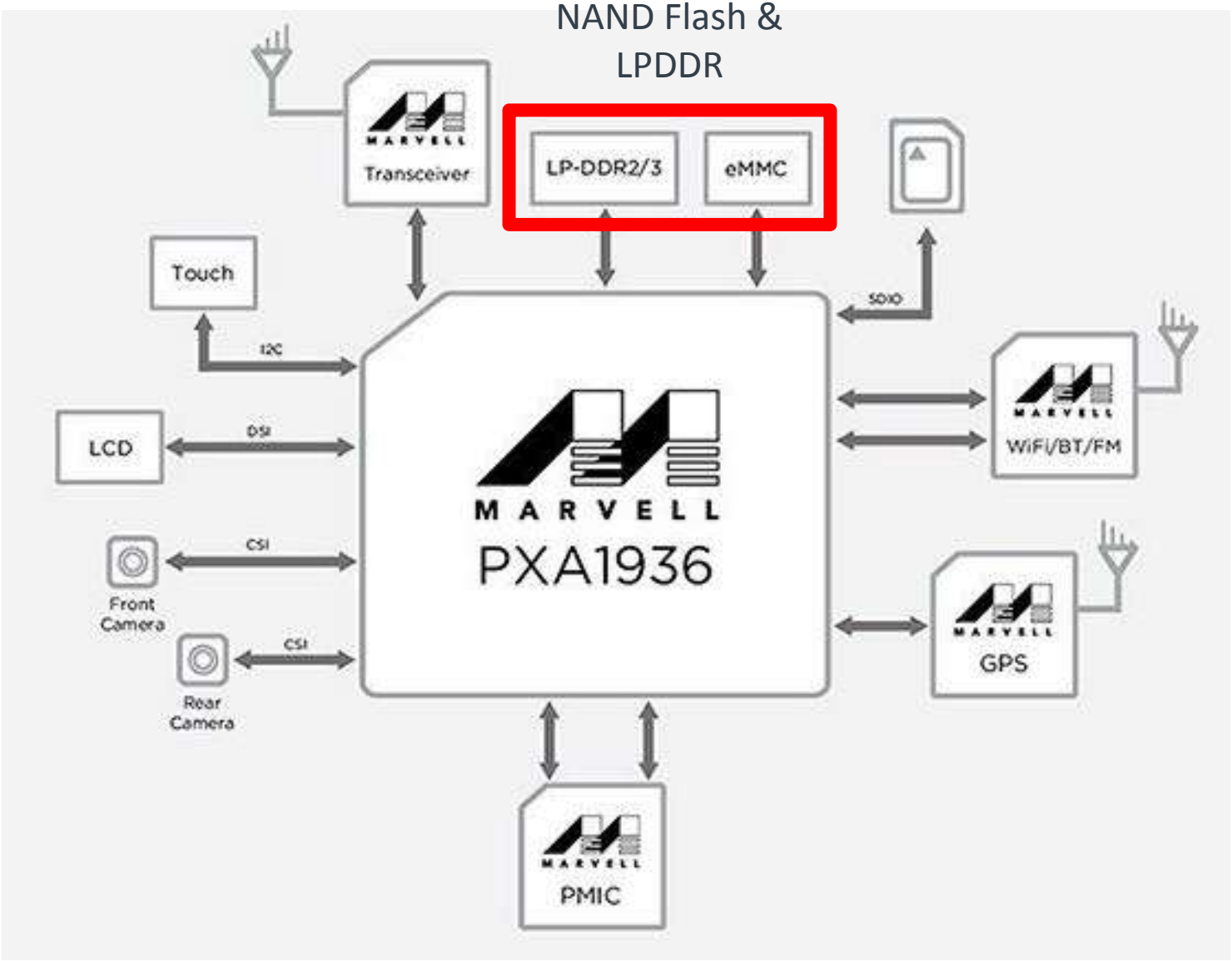


Caso 5: Smartphones ARM

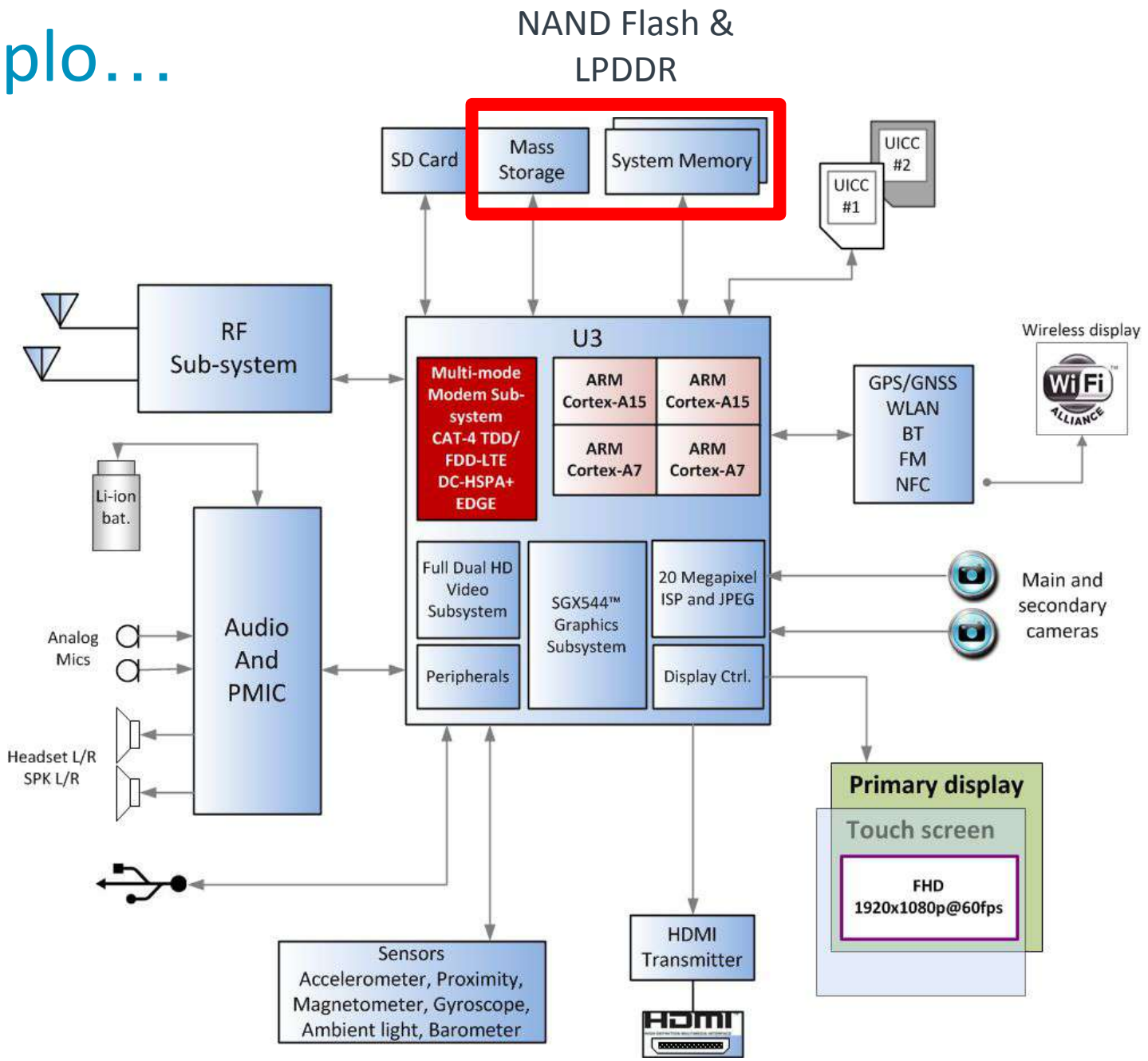
- Contém processadores de alta eficiência energética
  - Alto Processamento X Baixo Consumo
- Organização de Memória semelhante ao x86 (caches e memória LPDDR)
- Diferença: Todos os componentes encapsulados juntos (SoC) → **Limitações**
  - **Memória SRAM** pequena interna ao SoC ou uma **memória LPDDR** maior externa?



Exemplo: Memórias de um SoC de Smartphone



Outro Exemplo…



Memória Interna Moderna

Exercícios

Exercícios

1) Pesquise na literatura as principais diferenças entre EPROM, EEPROM e memória flash.

2) Qual é a diferença entre DRAM e SRAM em termos de velocidade, tamanho, custo e aplicação?

3) Considere uma RAM dinâmica convencional, uma SDRAM com buffer de pré-busca igual a 4 e uma DDR com buffer de pré-busca igual a 8, como mostrado nos slides anteriores, e que todos operam a 1 GHz e com 8 bits de dados cada. Calcule a taxa de dados teórica máxima que cada uma dessas variantes pode sustentar em bytes/segundo. Admita que um único relógio inativo precede uma nova configuração de endereço de linha.

Memória Interna Moderna

Conclusão

Resumo da Aula

- Tecnologias modernas trouxeram melhorias às memórias DRAM
  - Sincronização com o clock (SDRAM)
  - Taxa dupla de dados (DDR), bem como bufferização e maior frequência
- A memória Flash é não-volátil e muito rápida, porém mais cara
- Diferentes sistemas tem diferentes necessidades quando se fala de memória
  - Criticidade em relação ao custo da memória
  - **Microcontroladores**: Memória pequena e de baixíssima potência (criticidade altíssima)
  - **Smartphones**: Memória maior e de baixa potência (criticidade alta ou média)
  - **Computadores**: Memória grande e rápida (criticidade baixa em laptops e nula em PC)

Conclusão

- Nessa Aula:
  - Memória Interna Moderna
- Bibliografia Principal:
  - Arquitetura e Organização de Computadores; Stallings, W.; 10ª Edição (Capítulo 5)
- Próxima Aula:
  - Memória Externa