UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MICROELETRÔNICA

Controle Adaptativo para Acesso à Memória Compartilhada em Sistemas em Chip

Apresentação de Defesa de Tese de Doutorado

Eng. MSc. Alexsandro Cristovão Bonatto
Orientador: Prof. Dr. Altamiro Susin

15 de Agosto de 2014

Sumário da Apresentação

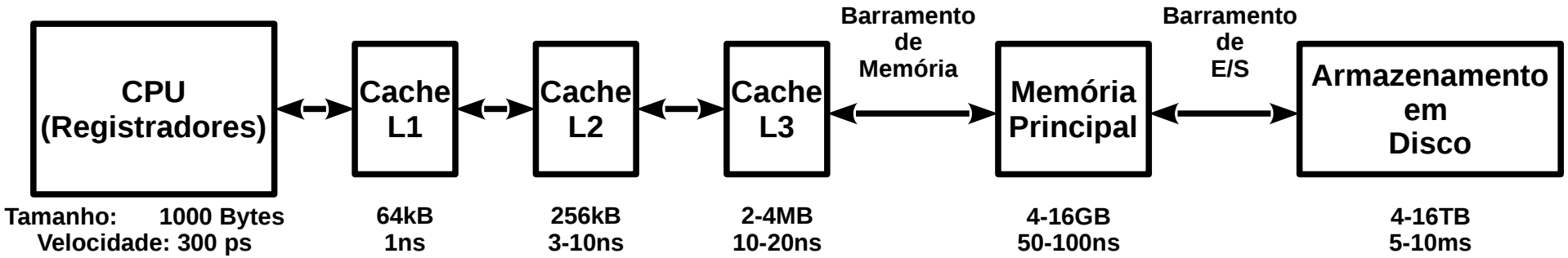
- Contextualização do Problema
- Funcionamento da DRAM
- Metodologia
- Resultados
- Comentários Finais

Sumário da Apresentação

- **Contextualização do Problema**
- Funcionamento da DRAM
- Metodologia
- Resultados
- Comentários Finais

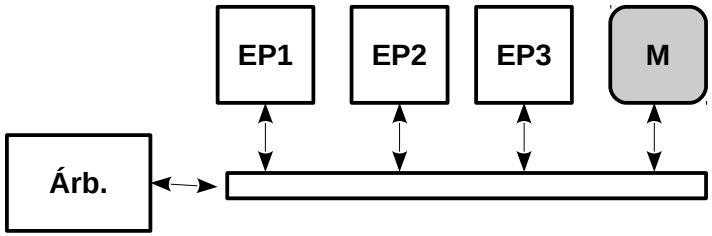
Contextualização do Problema

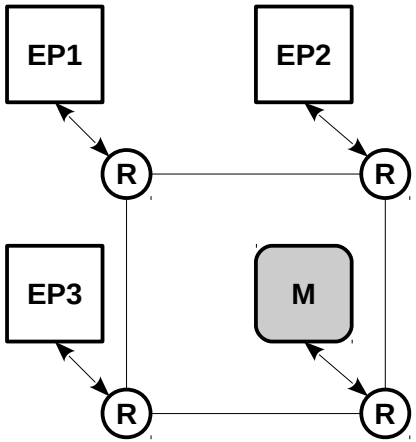
- Projeto de sistemas-em-Chip (SoCs), formados a partir de um número crescente de elementos de processamento (EPs);
- Integração de memórias de alta capacidade:
  - Memórias externas ao chip: tipicamente DRAM
  - Interfaces síncronas de comunicação: DDR (*Double Data Rate*)
- Limitações de velocidade de acesso aos dados: *Memory wall*
- Hierarquia de memórias: cache, DRAM e disco

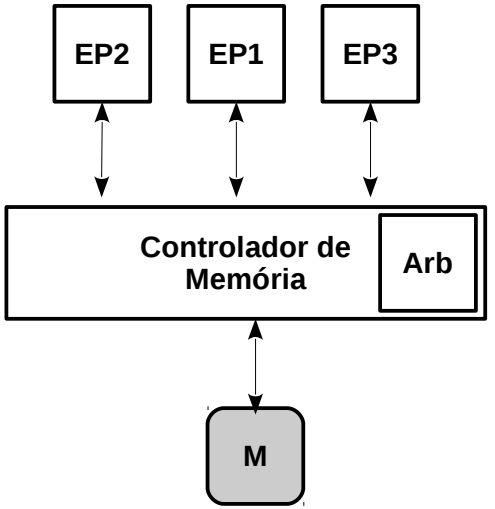


Estruturas de Comunicação

- Barramento, rede em chip, controlador multi-cliente

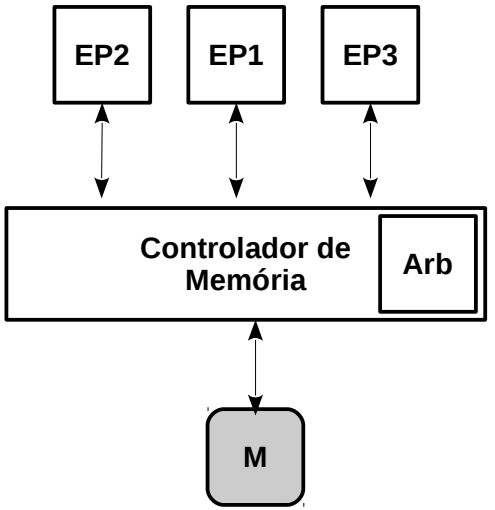






Controlador de Memória Multi-Cliente

- Melhorar a eficiência através da especialização das partes do circuito (ESMAEILZADEH et al., 2013).
- Considerar os aspectos específicos de acesso à memória (SEICULESCU et al., 2011);
- Explorar a organização de dados para aumentar a eficiência computacional (NEWMAN, 2014);
- Garantir os valores de largura de banda mínima e de latência máxima (AKESSON; GOOSSENS; RINGHOFER, 2007);
- Reproduzir um comportamento predizível no tempo (controlável) (MUTLU, 2013); (AKESSON; GOOSSENS, 2011).



Hipóteses

- Otimização do acesso ao canal de memória é atingida a partir do(a):
  - Controle de transações em sequências de rajadas: aumento da largura de banda e redução da potência dissipada;
  - Divisão do acesso ao canal de memória satisfazendo qualidade de serviço dos clientes;
  - Controle dos tempos de resposta: clientes com requisitos de tempo real;
- Controle adaptativo de prioridades;
- Diretrizes de projeto orientadas às características da memória:
  - Análise do comportamento da memória;
  - Implementação de regras de qualidade de serviço.

Contribuições da Tese

- Controle adaptativo para acesso à memória compartilhada:
  - Avaliação em tempo-real dos requisitos dos clientes;
  - Adaptação dinâmica de prioridades;
  - Estimação dos piores casos de acesso.

Contribuições da Tese

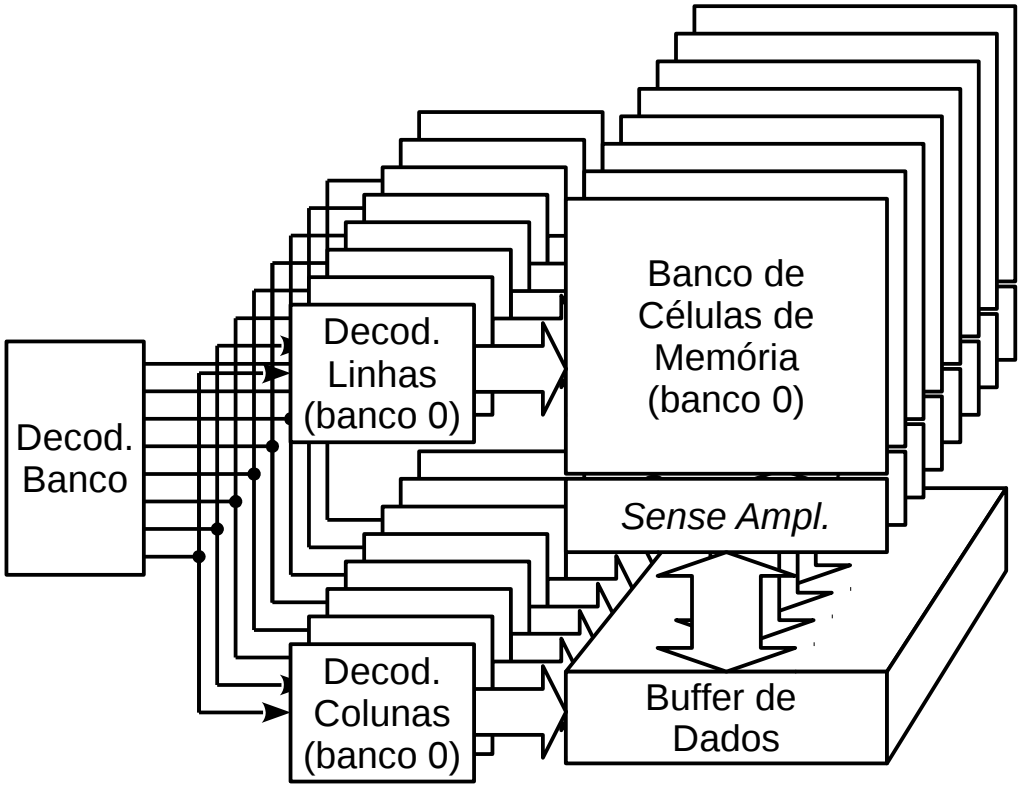
- Controle adaptativo para acesso à memória compartilhada:
  - Avaliação em tempo-real dos requisitos dos clientes;
  - Adaptação dinâmica de prioridades;
  - Estimação dos piores casos de acesso.
- Arquitetura de um subsistema de memória para SoCs com características de:
  - Adaptatividade: implementa um controle adaptativo;
  - Previsibilidade: apresenta um comportamento previsível;
  - Escalabilidade: implementa múltiplas interfaces;
  - Heterogeneidade: suporta interfaces com acessos diferenciados.

Sumário da Apresentação

- Contextualização do Problema
- **Funcionamento da DRAM**
- Metodologia
- Resultados
- Comentários Finais

DRAM
Arquitetura Interna

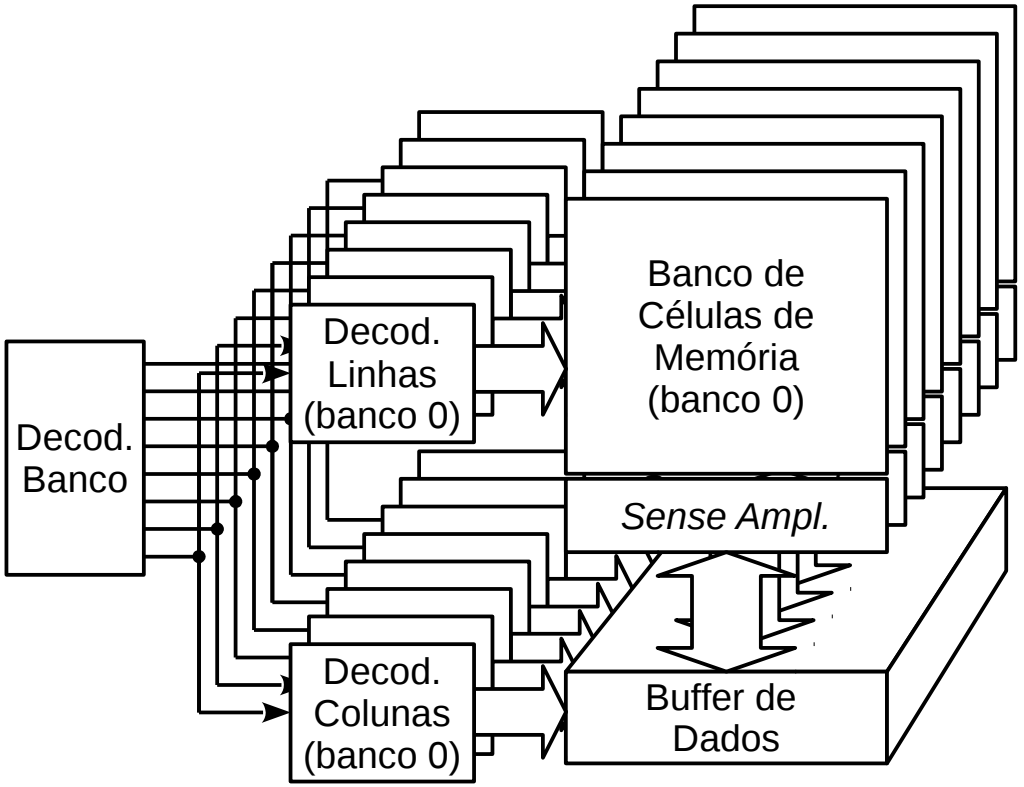
- Memória organizada em bancos, linhas e colunas:



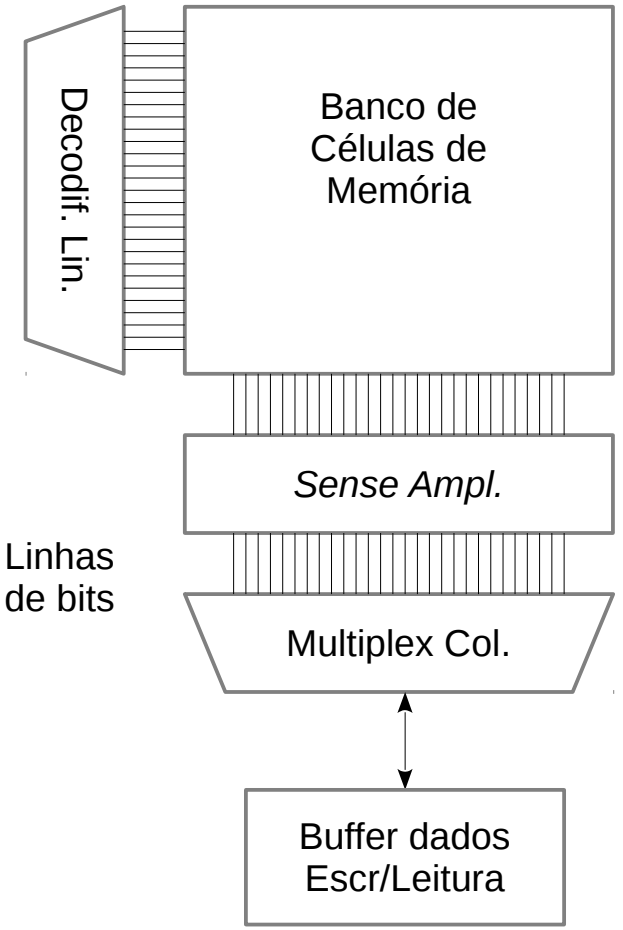
x8 bancos

DRAM
Arquitetura Interna

- Memória organizada em bancos, linhas e colunas:

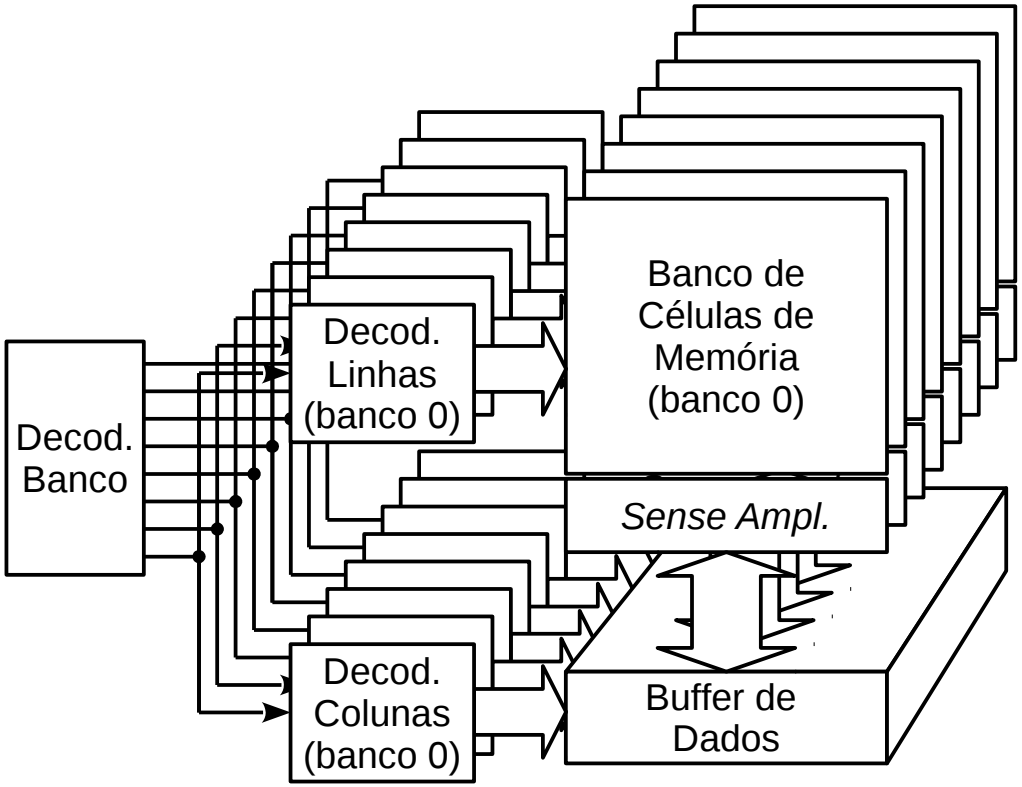


x8 bancos

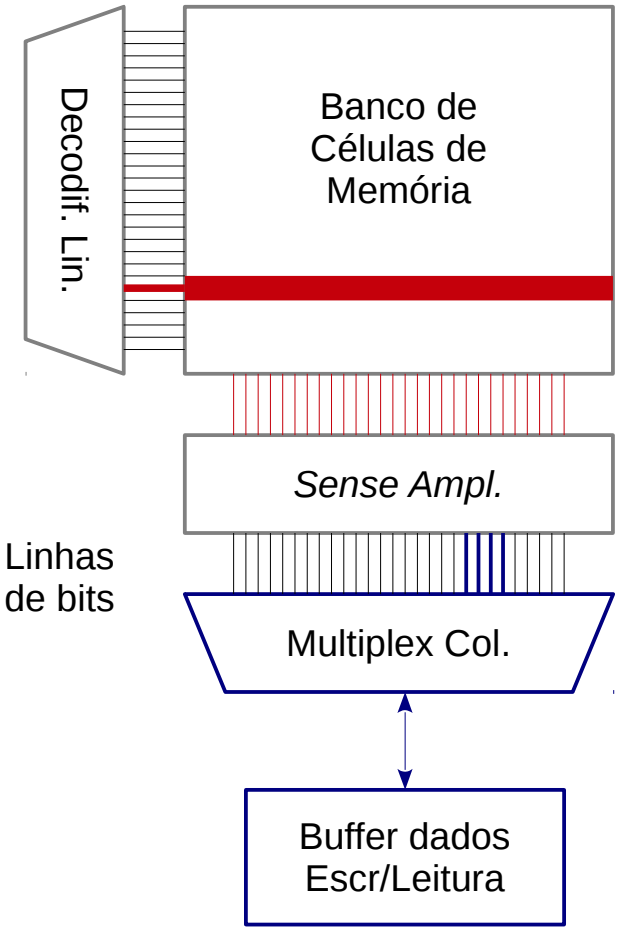


DRAM
Arquitetura Interna

- Memória organizada em bancos, linhas e colunas:



x8 bancos



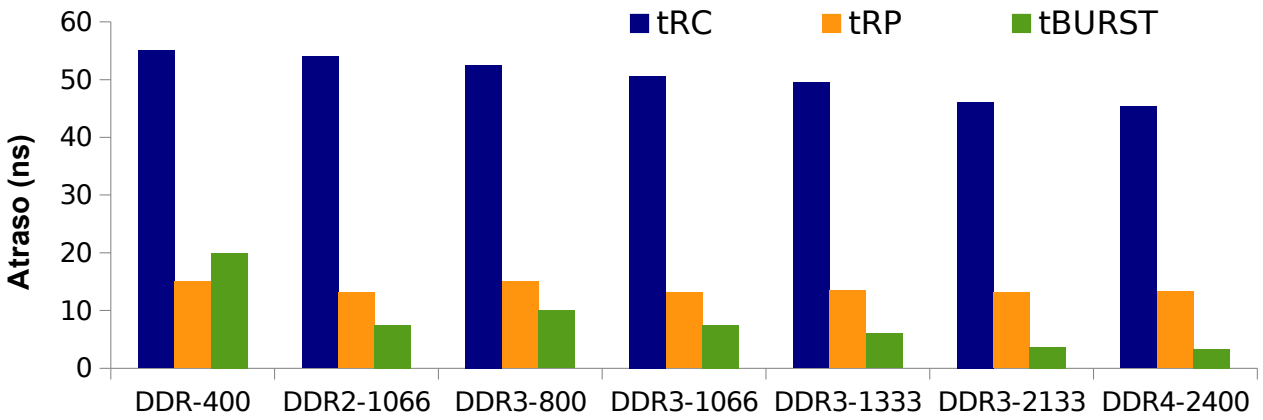
Dispositivos de Memória Dinâmica

- Gerações de DRAM:
  - Computação: SDRAM, DDR, DDR2, DDR3, DDR4.
  - Gráfica: GDDR, GDDR2, GDDR3, GDDR5.
  - Portáteis: LPDDR, LPDDR2
- Taxas de dados crescentes:
  - DDR: 400 Mtps*
  - DDR2: 1066 Mtps
  - DDR3: 2133 Mtps
  - DDR4: 4800 Mtps
- Pequena melhora na velocidade do núcleo

*Mega-transferências por segundo

Dispositivos de Memória Dinâmica

- Gerações de DRAM:
  - Computação: SDRAM, DDR, DDR2, DDR3, DDR4.
  - Gráfica: GDDR, GDDR2, GDDR3, GDDR5.
  - Portáteis: LPDDR, LPDDR2
- Taxas de dados crescentes:
  - DDR: 400 Mtps*
  - DDR2: 1066 Mtps
  - DDR3: 2133 Mtps
  - DDR4: 4800 Mtps
- Pequena melhora na velocidade do núcleo

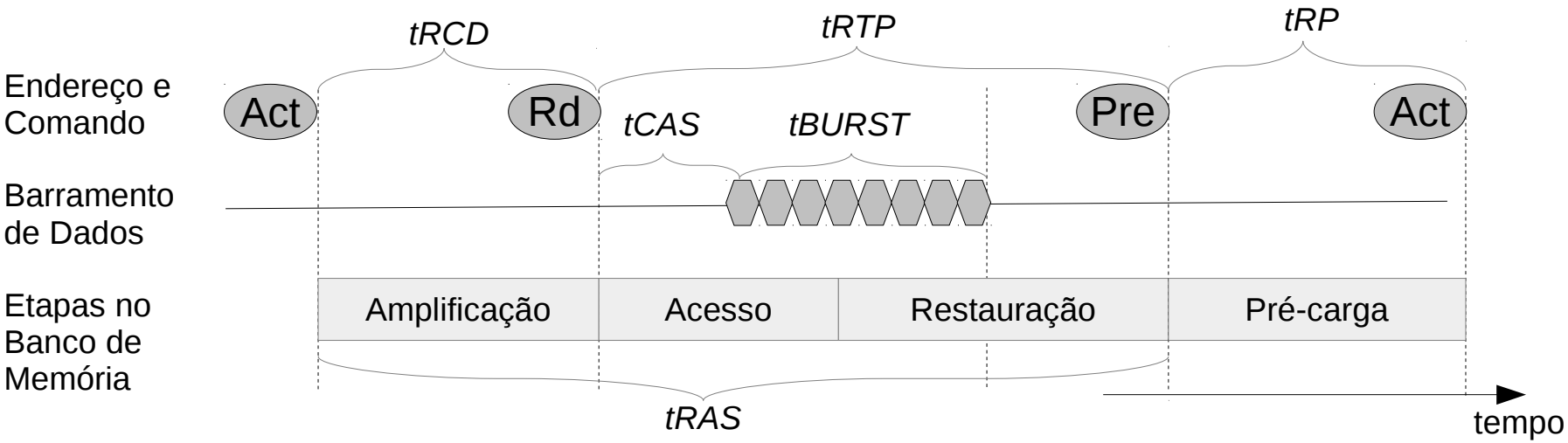


Redução de 17,6% para tRC, 11,2% para tRP e 12x para tBURST

*Mega-transferências por segundo

Operações para Acesso aos Dados
Etapas da Leitura de Dados

- Diagrama de tempo para leitura de uma rajada BL8 no modo de página fechada:



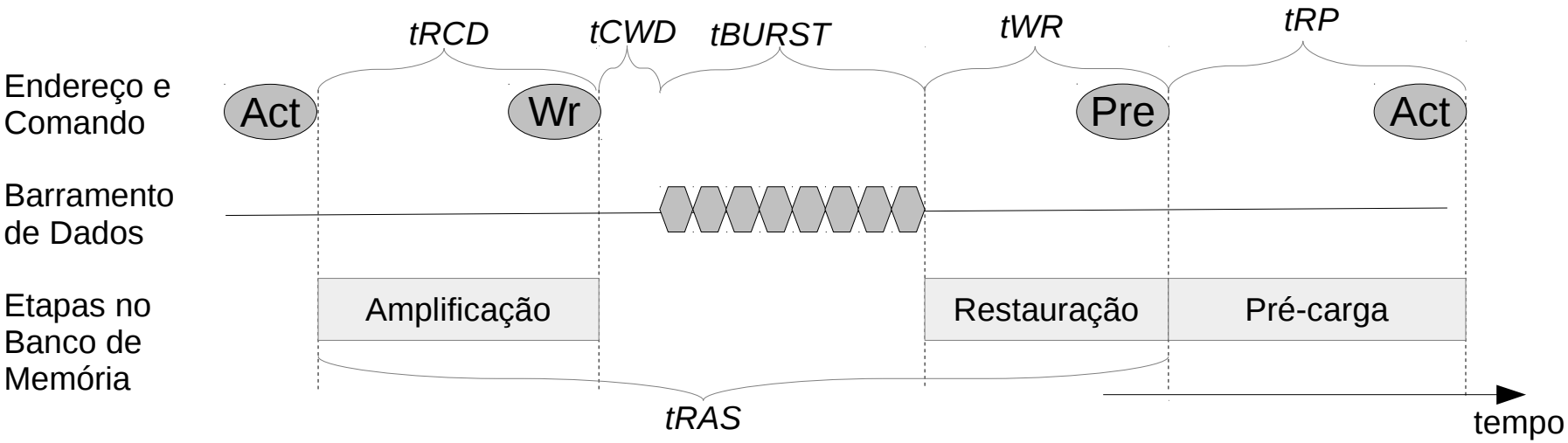
Ciclo de linha → tRC = tRAS + tRP

Tempo para leitura de um bloco de dados:

$t_r(n) = max(tRC; tRCD + (n-1) \cdot tCCD + tRTP + tRP)$

Operações para Acesso aos Dados
Etapas da Escrita de Dados

- Diagrama de tempo para escrita de uma rajada BL8 no modo de página fechada:



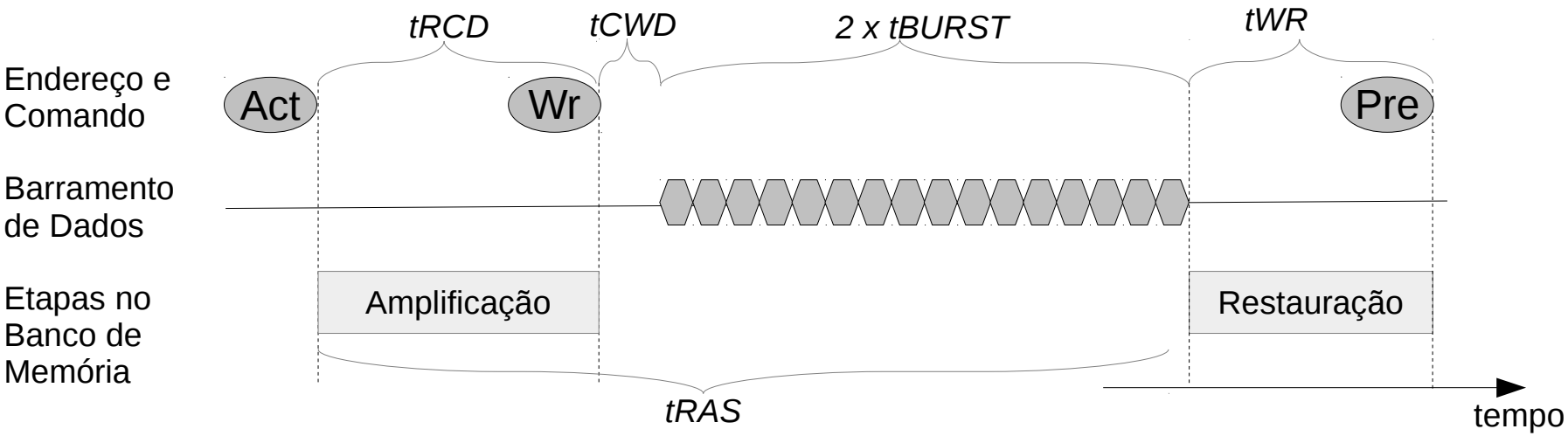
Ciclo de linha → tRC = tRAS + tRP

Tempo para escrita de um bloco de dados:

$t_w(n) = tRCD + tCWD + n \cdot tBURST + tWR + tRP$

Operações para Acesso aos Dados
Sequência de Rajadas de Dados

- Diagrama de tempo para escrita de uma rajada BL8 no modo de página fechada:



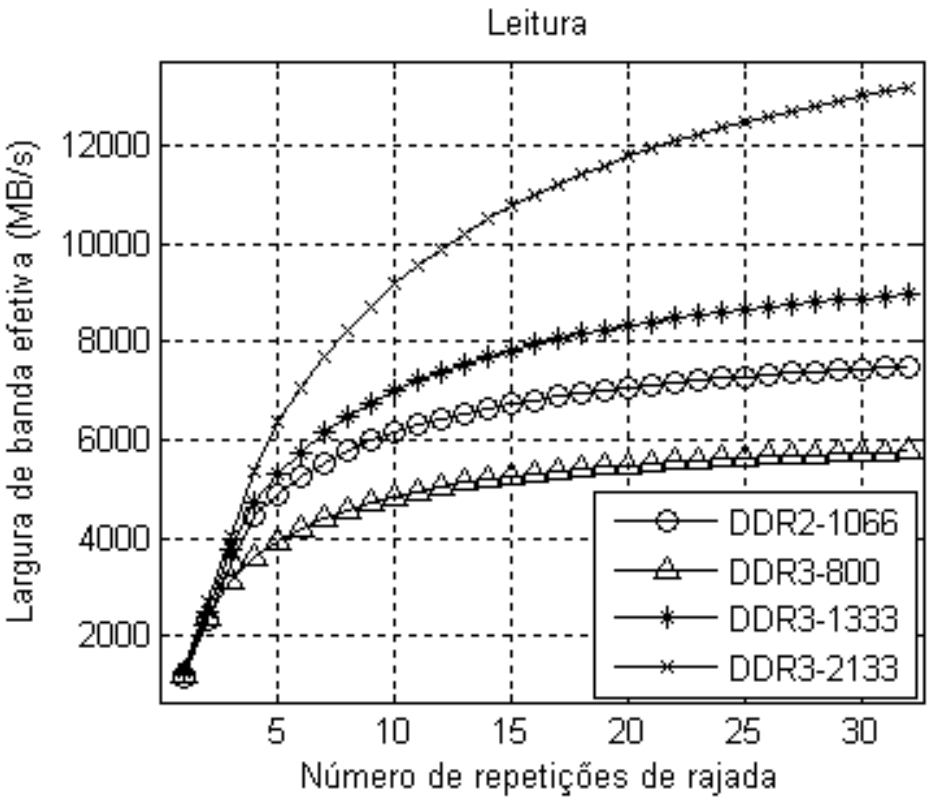
Ciclo de linha → tRC = tRAS + tRP

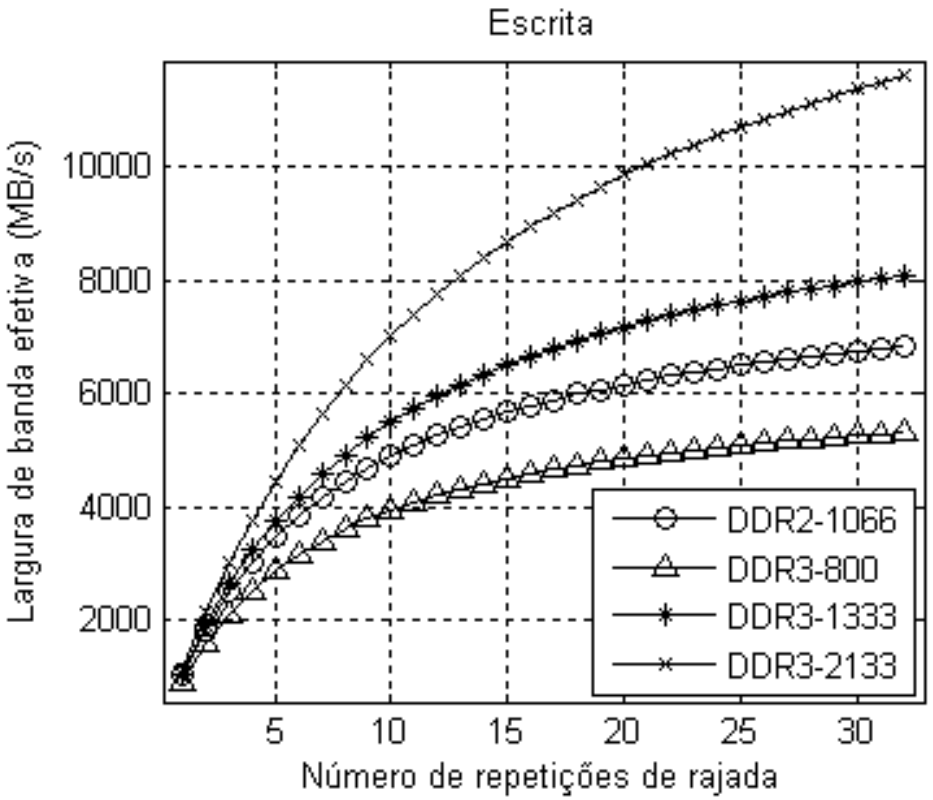
Tempo para escrita de um bloco de dados:

$t_w(n) = tRCD + tCWD + n \cdot tBURST + tWR + tRP$

Largura de Banda
Sequência de Rajadas de Dados

- Largura de banda sustentada aumenta com o aumento da repetição do número de rajadas





Sumário da Apresentação

- Contextualização do Problema
- Funcionamento da DRAM
- **Metodologia**
- Resultados
- Comentários Finais

Metodologia
Descrição

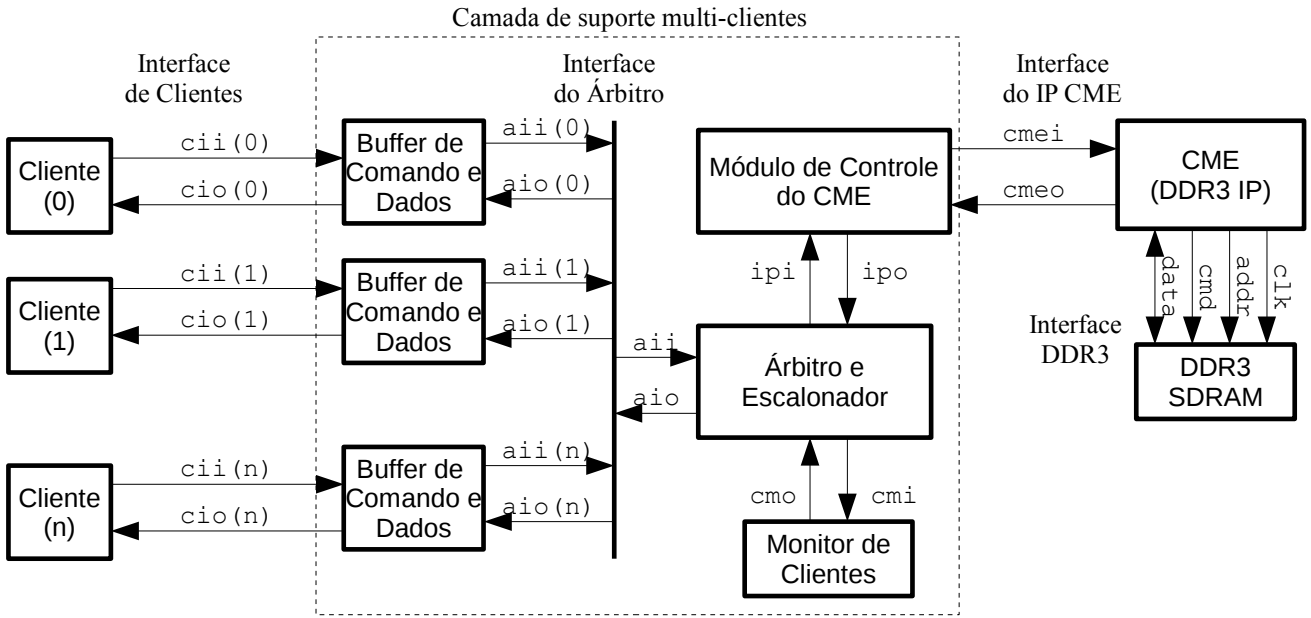
- *Memory-centric design*:
  - Avaliação das características da memória para traçar diretrizes ao projeto de partes do SoC, com objetivo de melhorar o desempenho no compartilhamento do canal de memória.

Metodologia
Descrição

- *Memory-centric design*:
  - Avaliação das características da memória para traçar diretrizes ao projeto de partes do SoC, com objetivo de melhorar o desempenho no compartilhamento do canal de memória.
- Arquitetura do controlador de memória:
  - Acessos em sequências de rajadas;
  - Adaptação dinâmica de prioridades;
  - Análise dos piores casos em tempo de execução.
- Tamanho da transação (em sequências de rajadas);
- Prazo de conclusão (*deadline*);
- Granularidade mínima: menor tamanho contíguo de acesso em uma transação.

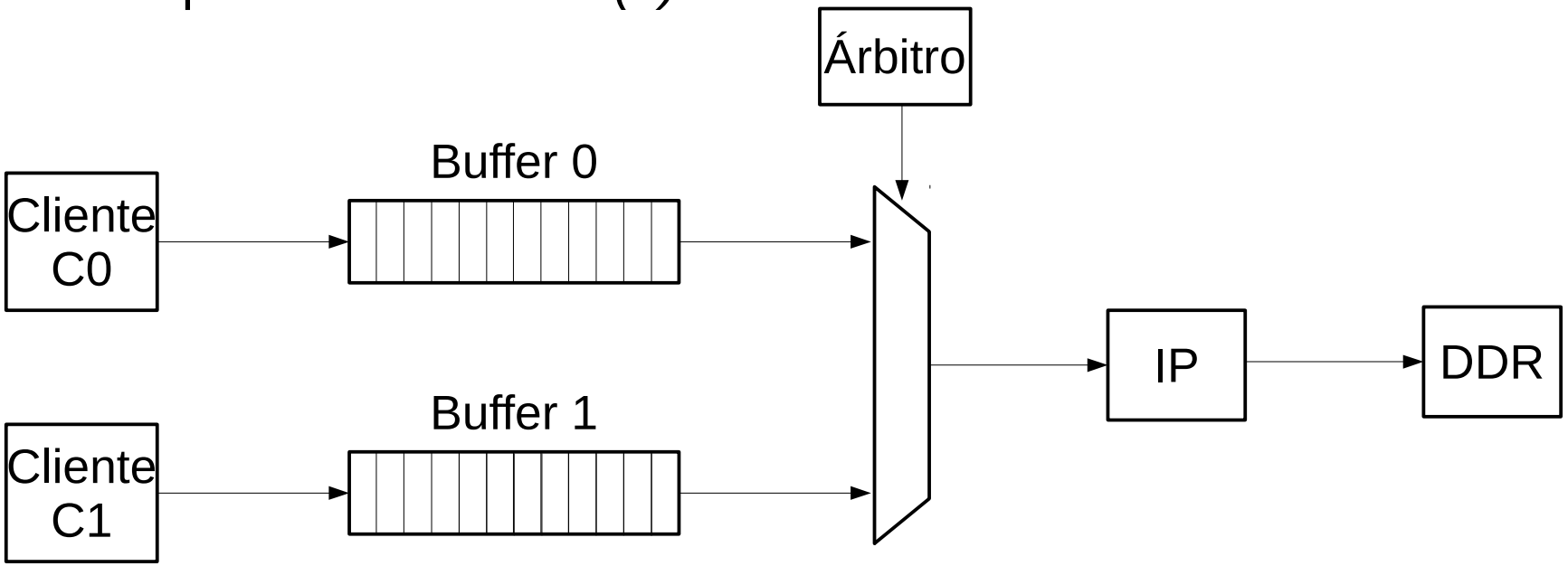
Metodologia
Arquitetura do Subsistema de Memória

- Elementos principais formadores do subsistema de memória:
  - Memórias temporárias (buffers ou filas)
  - Multiplexadores para dados, comandos e endereços
  - Árbitro para gerenciamento dos acessos
  - Monitor de clientes: gera estatísticas dos acessos
  - Interface física com a memória (PHY)



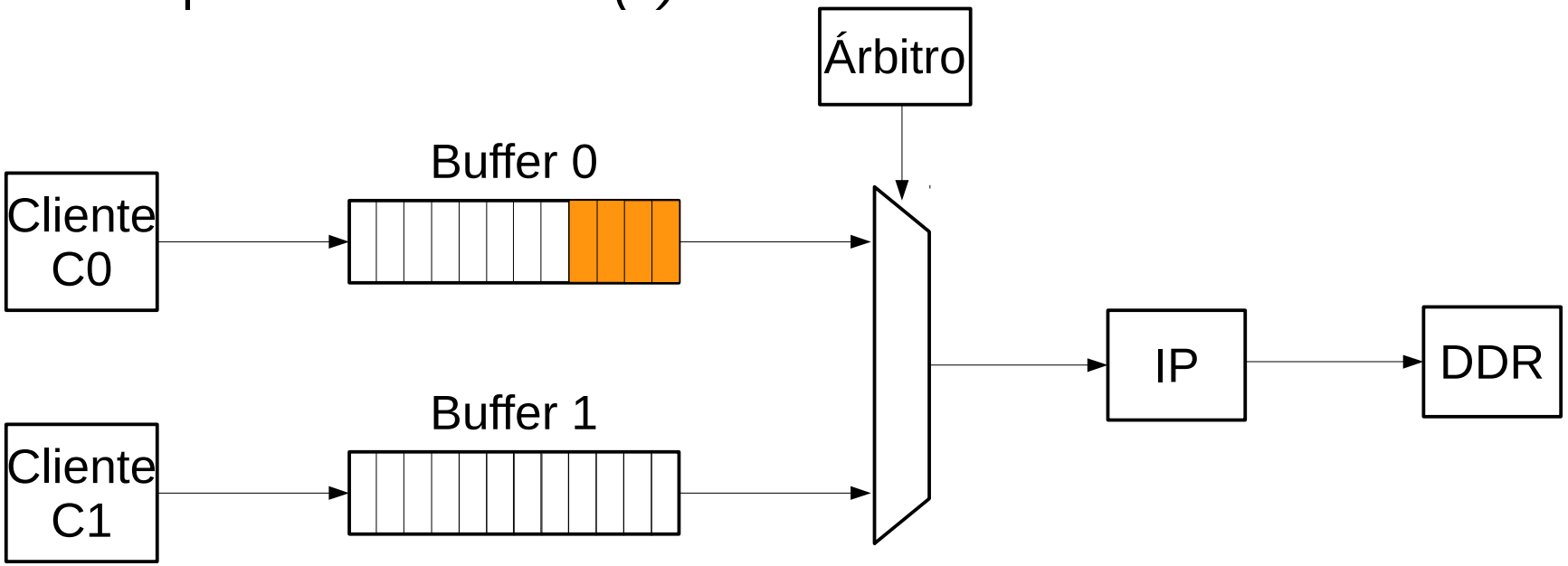
Funcionamento
Transações e Granularidade

- Adiciona informações das transações solicitadas pelos clientes:
  - Tamanho da transação: *l(n)*
  - Prazo para conclusão: *dl(n)*



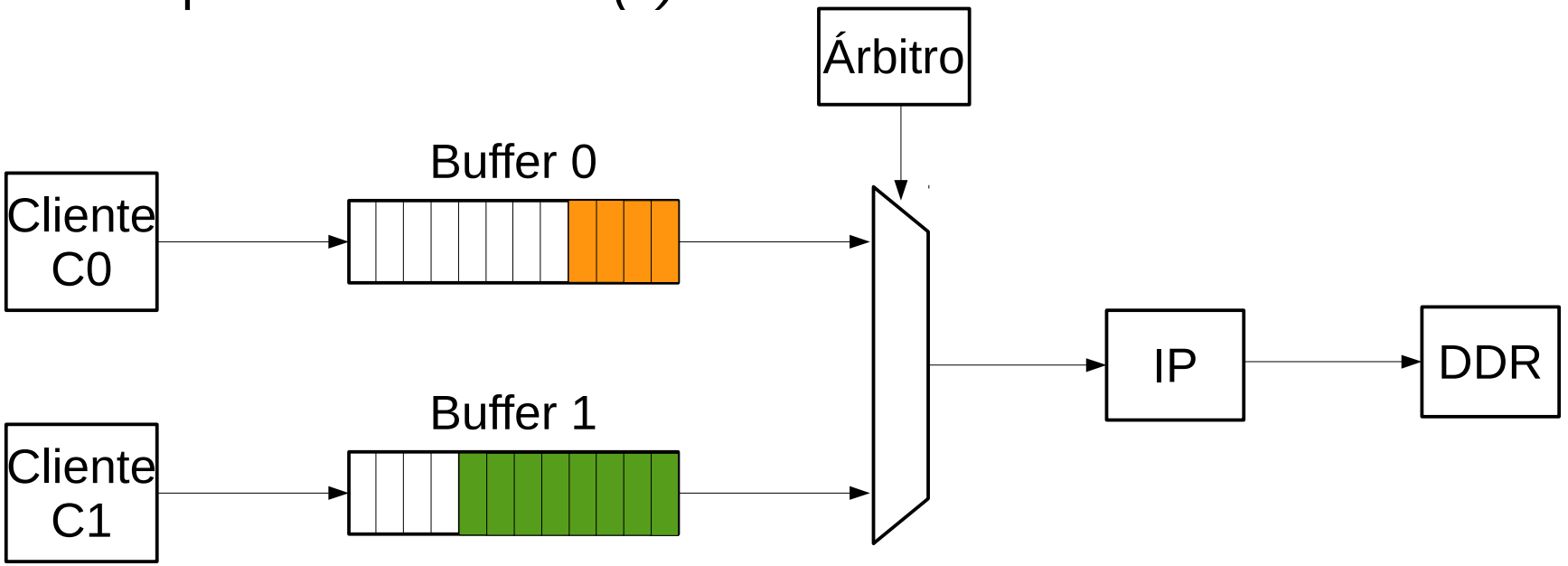
Funcionamento
Transações e Granularidade

- Adiciona informações das transações solicitadas pelos clientes:
  - Tamanho da transação: *l(n)*
  - Prazo para conclusão: *dl(n)*



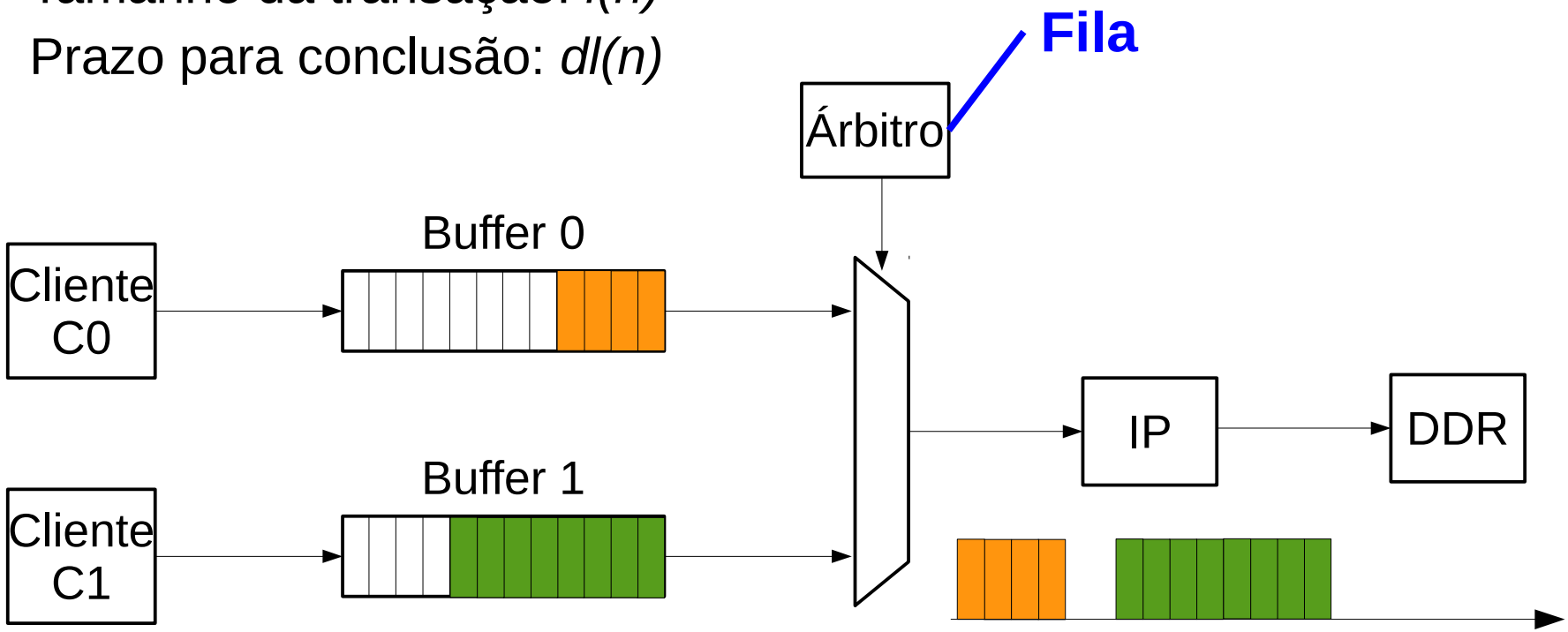
Funcionamento
Transações e Granularidade

- Adiciona informações das transações solicitadas pelos clientes:
  - Tamanho da transação: *l(n)*
  - Prazo para conclusão: *dl(n)*



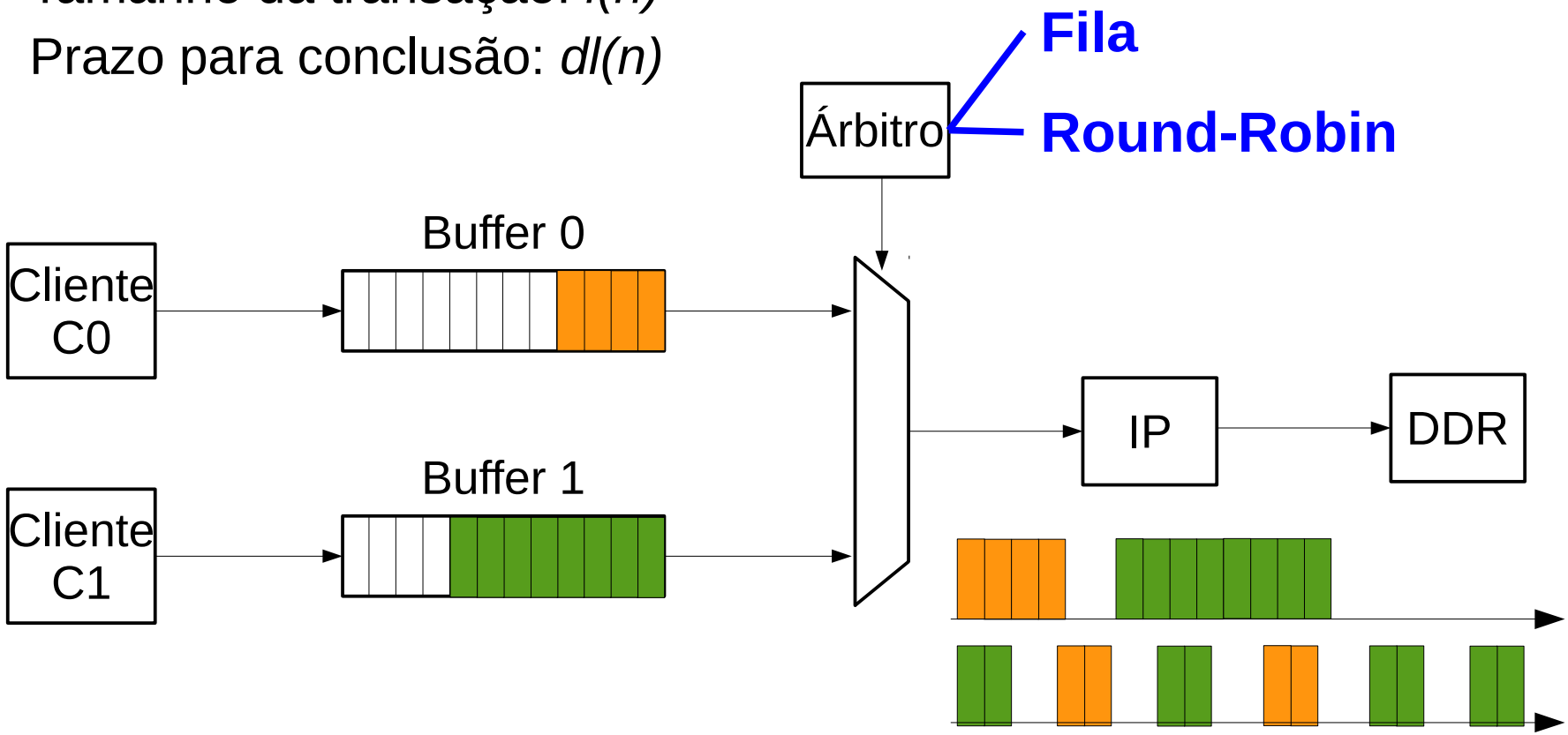
Funcionamento
Transações e Granularidade

- Adiciona informações das transações solicitadas pelos clientes:
  - Tamanho da transação: *l(n)*
  - Prazo para conclusão: *dl(n)*



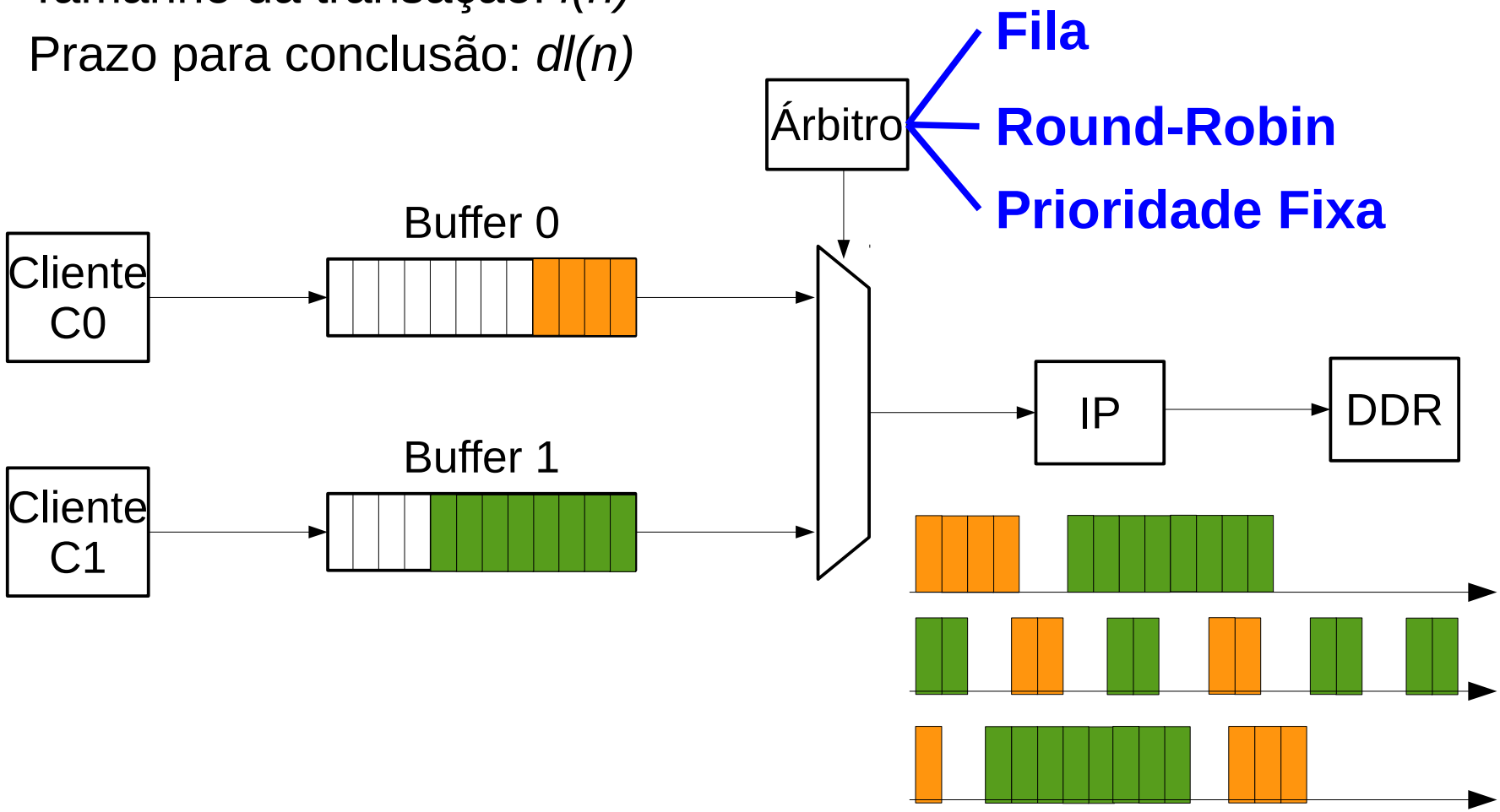
Funcionamento
Transações e Granularidade

- Adiciona informações das transações solicitadas pelos clientes:
  - Tamanho da transação: *l(n)*
  - Prazo para conclusão: *dl(n)*



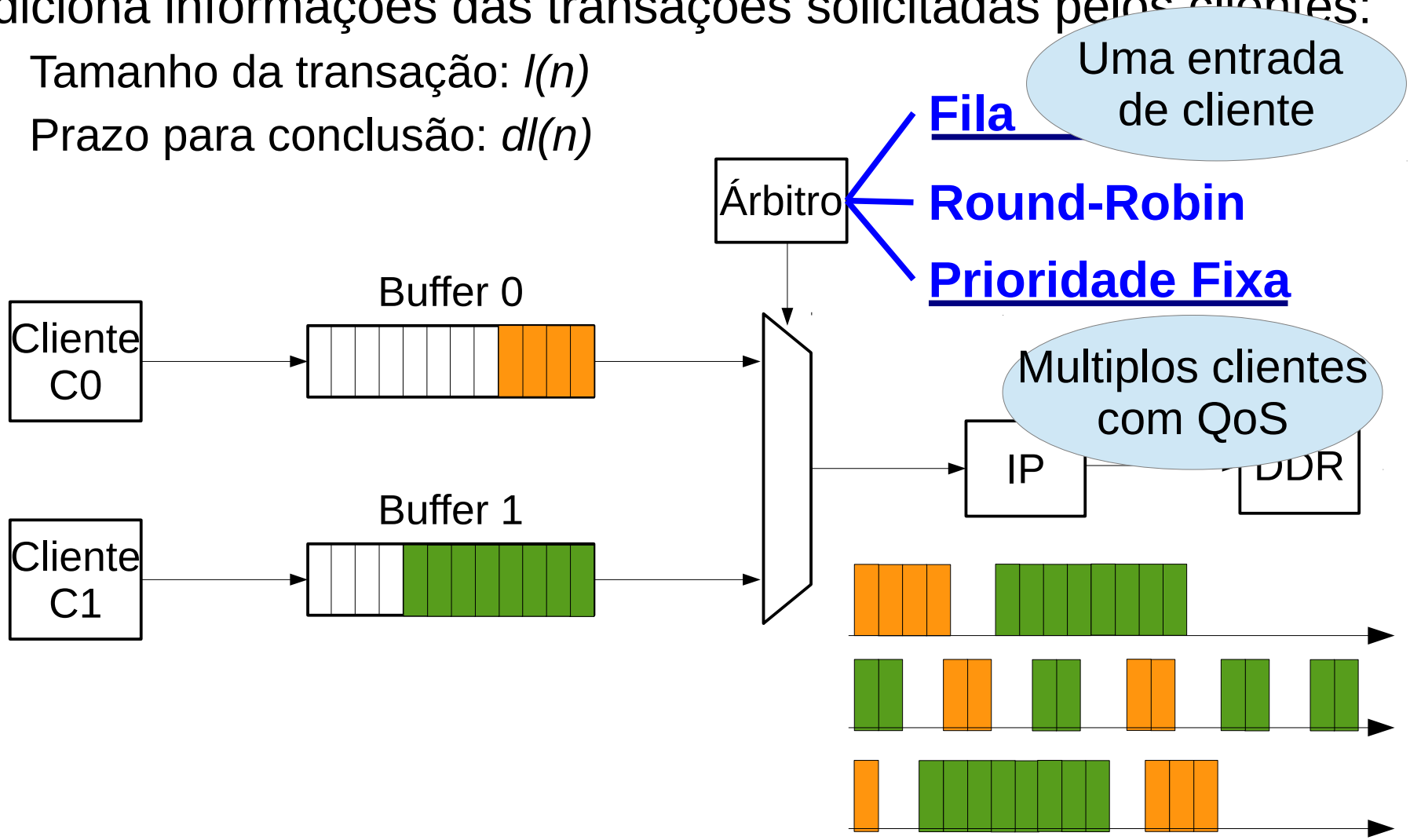
Funcionamento
Transações e Granularidade

- Adiciona informações das transações solicitadas pelos clientes:
  - Tamanho da transação: *l(n)*
  - Prazo para conclusão: *dl(n)*



Funcionamento
Transações e Granularidade

- Adiciona informações das transações solicitadas pelos clientes:
  - Tamanho da transação: *l(n)*
  - Prazo para conclusão: *dl(n)*

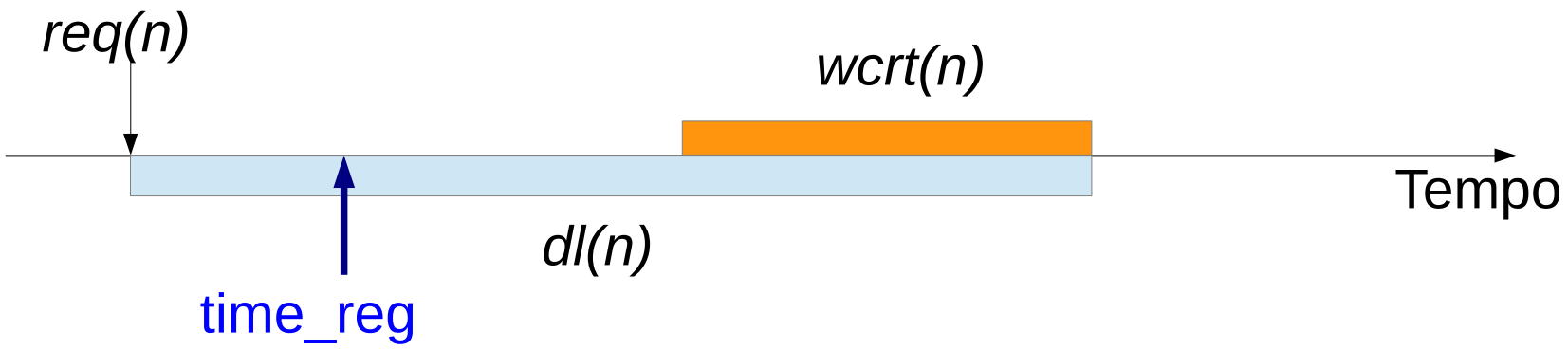


Prazo de Conclusão
Análise do pior caso

- Árbitro com classificação dinâmica de prioridades;
- Interrupção baseada nos prazos de conclusão;
- Cálculo do tempo de resposta no pior caso (WCRT).

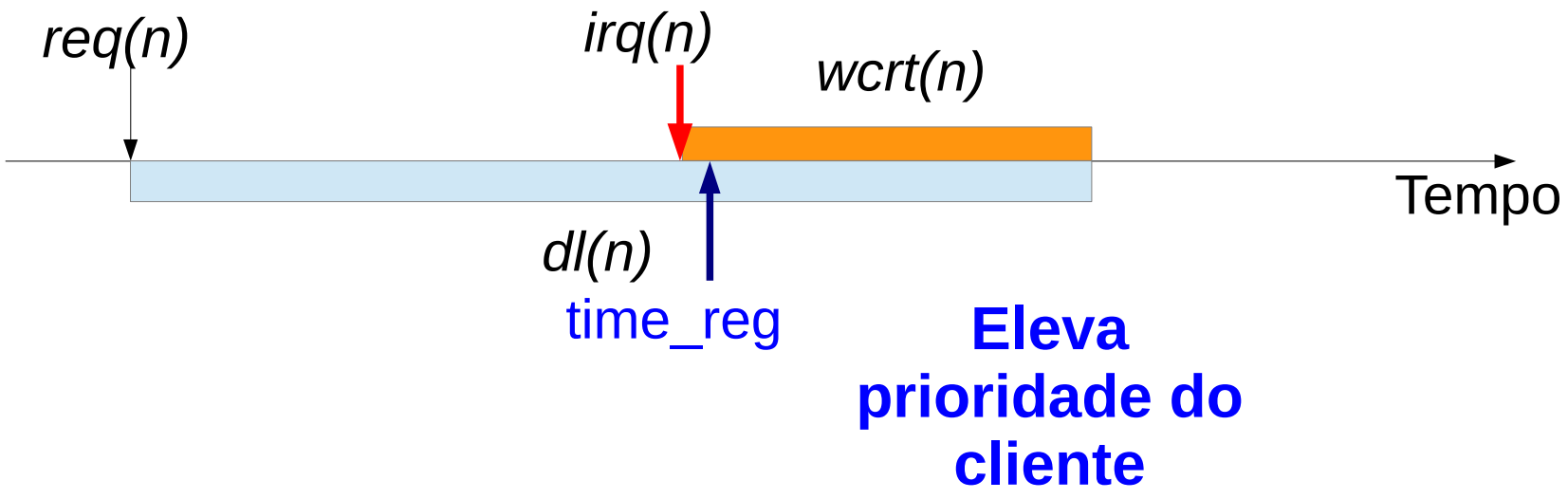
Prazo de Conclusão
Análise do pior caso

- Árbitro com classificação dinâmica de prioridades;
- Interrupção baseada nos prazos de conclusão;
- Cálculo do tempo de resposta no pior caso (WCRT).



Prazo de Conclusão
Análise do pior caso

- Árbitro com classificação dinâmica de prioridades;
- Interrupção baseada nos prazos de conclusão;
- Cálculo do tempo de resposta no pior caso (WCRT).



Análise do pior caso
Worst-Case Response Time

- Modelo analítico utilizado para determinar o pior caso:
  - Cálculo do tempo de resposta para o conjunto de clientes;
  - Gera uma estimativa confiável dos tempos de acesso (SHAH; KNOLL; AKESSON, 2013);
  - Baseado nos parâmetros da DRAM;
  - Depende do algoritmo de arbitragem implementado.
- Controlador deve ser implementado com comportamento previsível:
  - Política de página fechada;
  - Não implementa reordenamento de bancos ou comandos.

Pior Caso do Tempo de Resposta

- WCRT para escrita e leitura:

$$wcrt_E(l) = tRBC(l) + tED(l) + tAA$$

$$wcrt_L(l) = tRBC(l) + tLD(l) + tAA$$

$$tAA = tRFC + tWR + tRP + tRCD$$

- Para *n* clientes:

$$wcrt_E(l) = \sum tRBC(l) + tED(l) + tAA$$

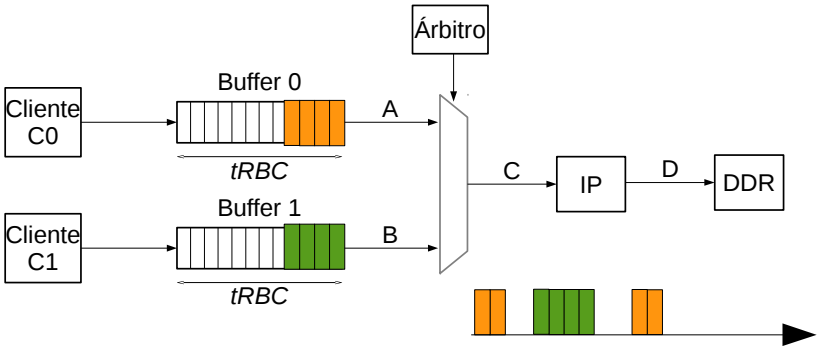
$$wcrt_L(l) = \sum tRBC(l) + tLD(l) + tAA$$

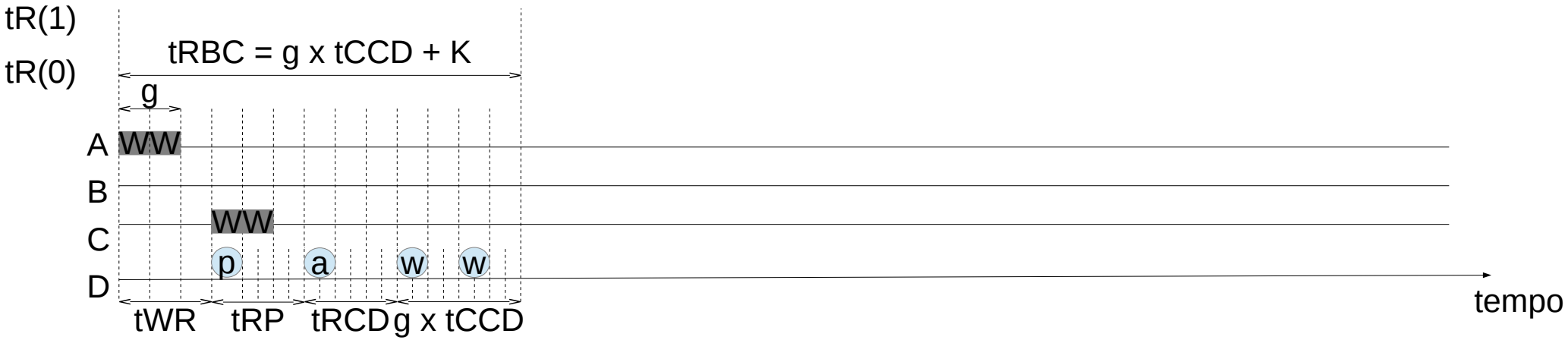
Modelo de Atrasos com Preempção

$$tRBC(l) = (\lceil l(0)/g \rceil - 1) \cdot (l(0) \cdot tCCD + K)$$

$$tED(l) = l(1) \cdot tCCD + K$$

$$K = tWR + tRP + tRCD$$



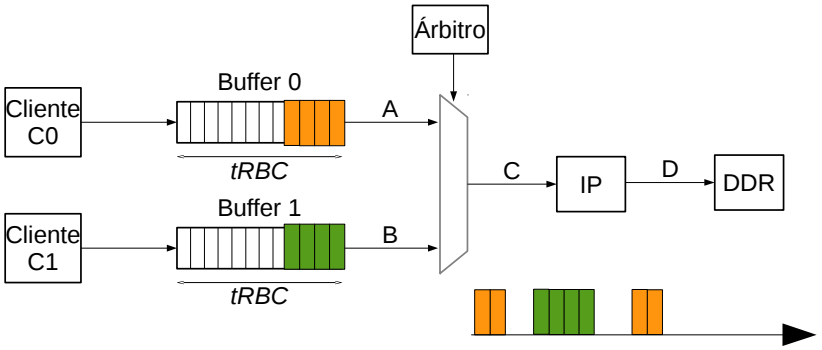


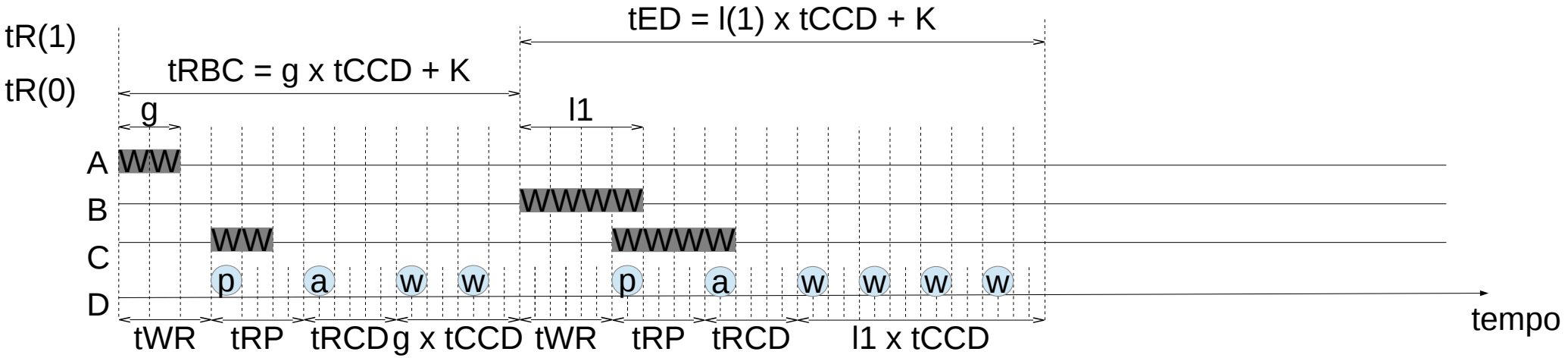
Modelo de Atrasos com Preempção

$$tRBC(l) = (\lceil l(0)/g \rceil - 1) \cdot (l(0) \cdot tCCD + K)$$

$$tED(l) = l(1) \cdot tCCD + K$$

$$K = tWR + tRP + tRCD$$



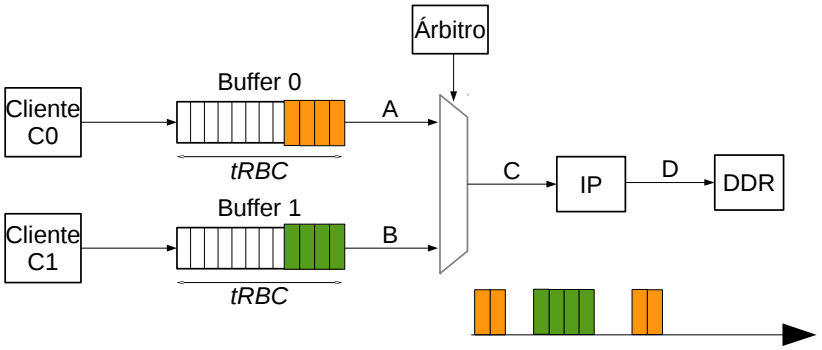


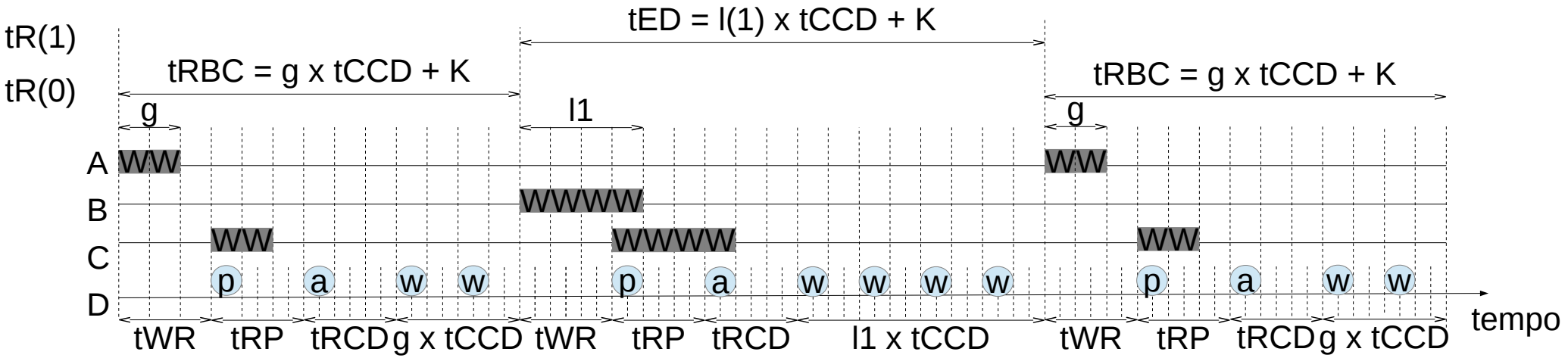
Modelo de Atrasos com Preempção

$$tRBC(l) = (\lceil l(0)/g \rceil - 1) \cdot (l(0) \cdot tCCD + K)$$

$$tED(l) = l(1) \cdot tCCD + K$$

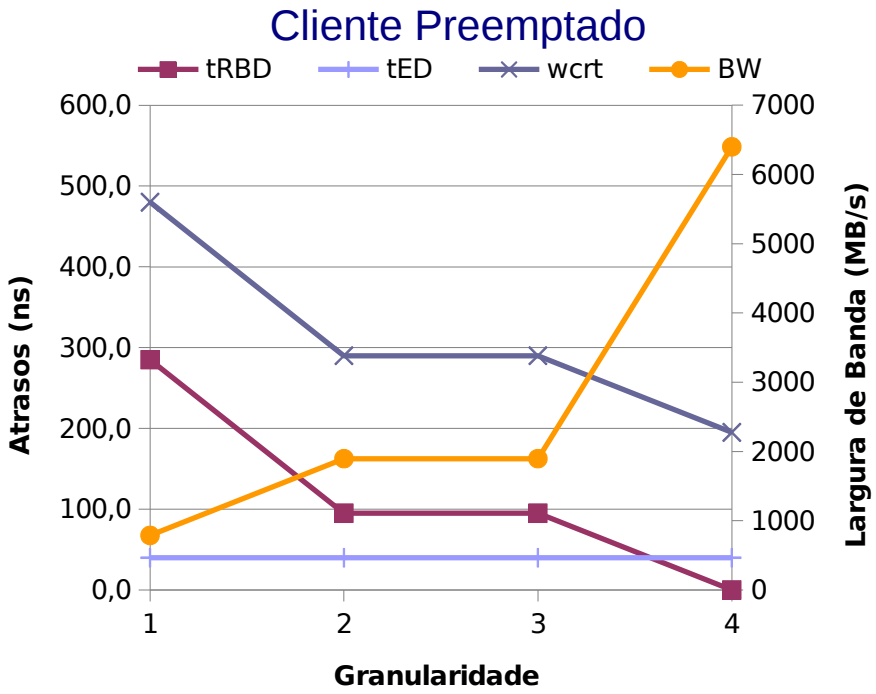
$$K = tWR + tRP + tRCD$$

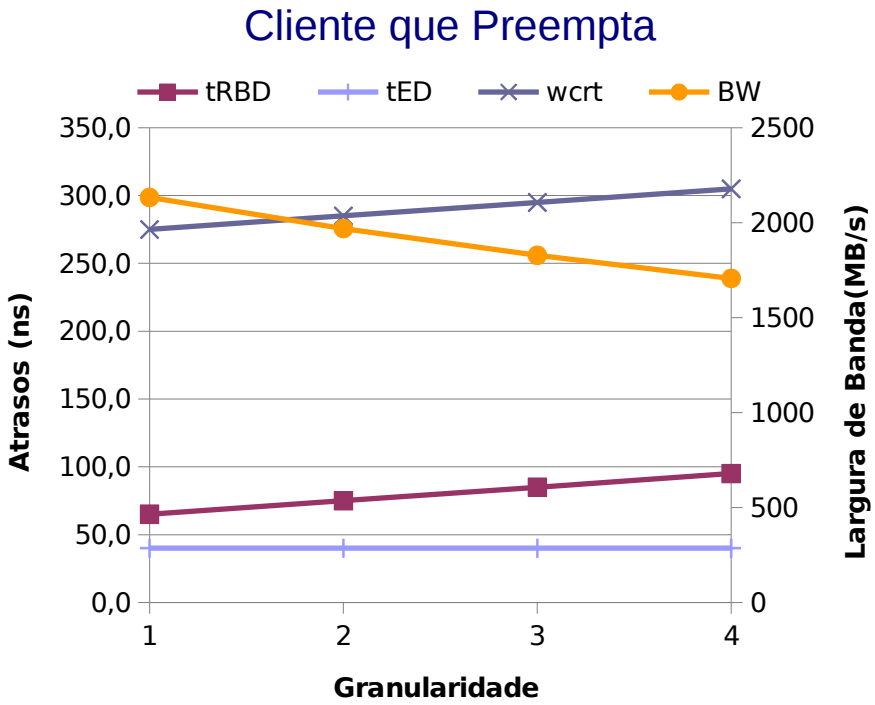




Análise da Granularidade
Tempo de Resposta e Largura de Banda

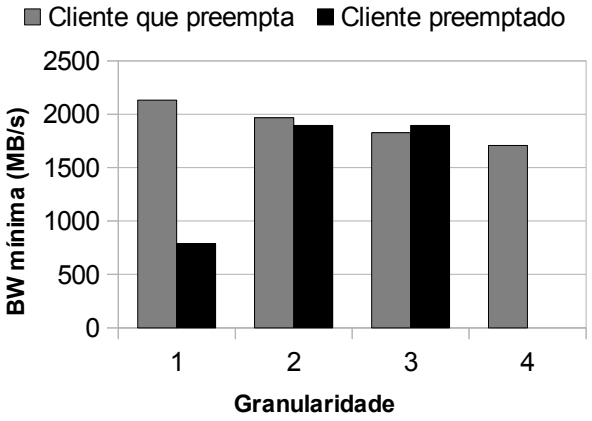
- Varia conforme a granularidade mínima de acessos:
  - Caso avaliado: *n=2, l(0)=l(1)=4*
  - *g=4* → preempção desabilitada
  - Largura de banda varia de 787 MB/s para 1896 MB/s

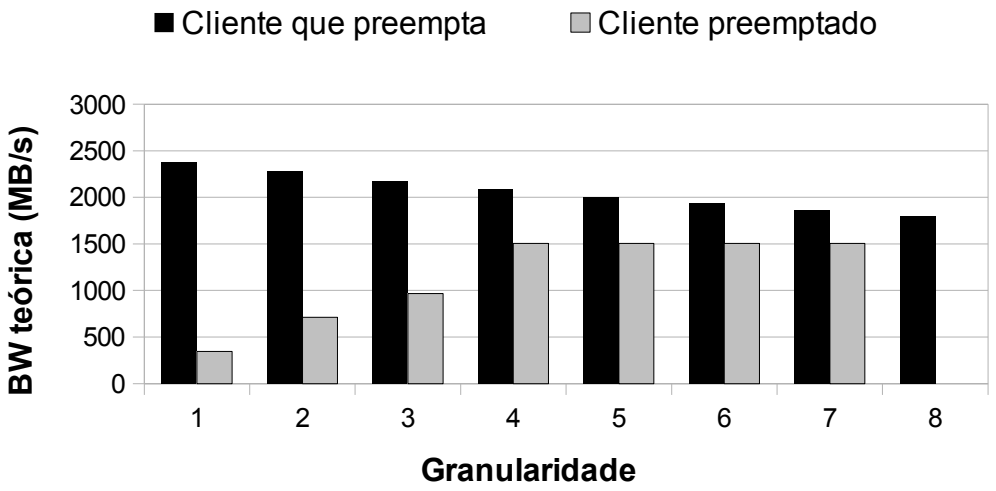


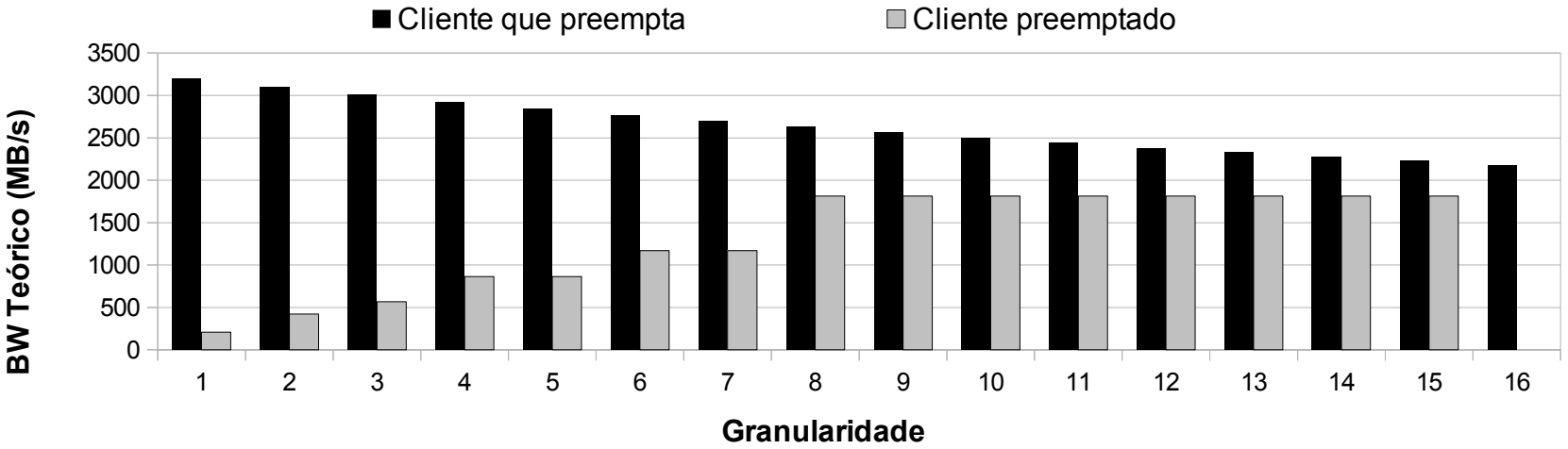


Análise da Granularidade
Diferentes Tamanhos de Transação

- Caso avaliado: 2 clientes

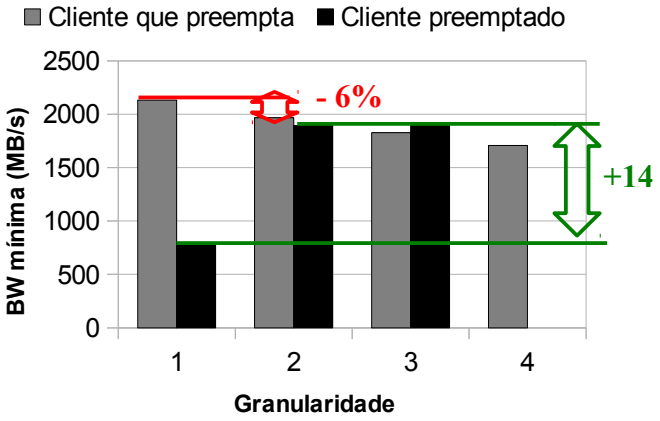


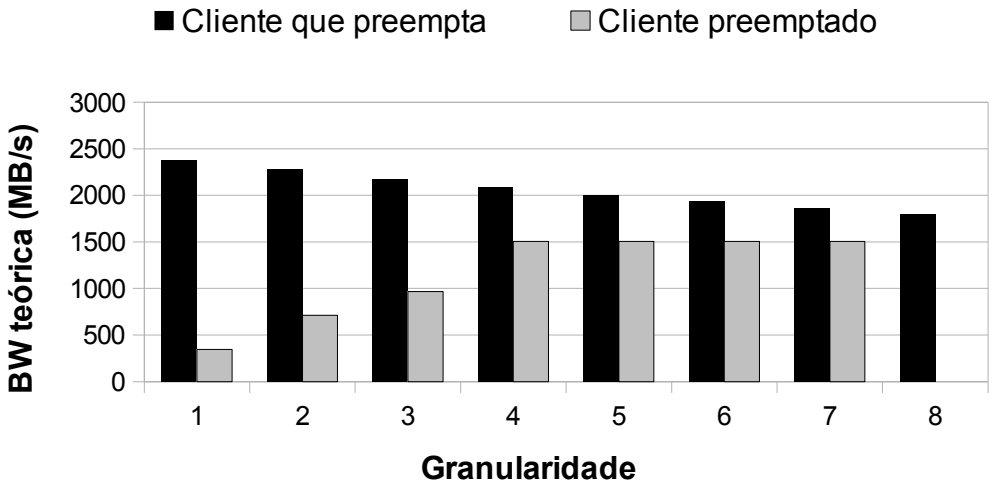


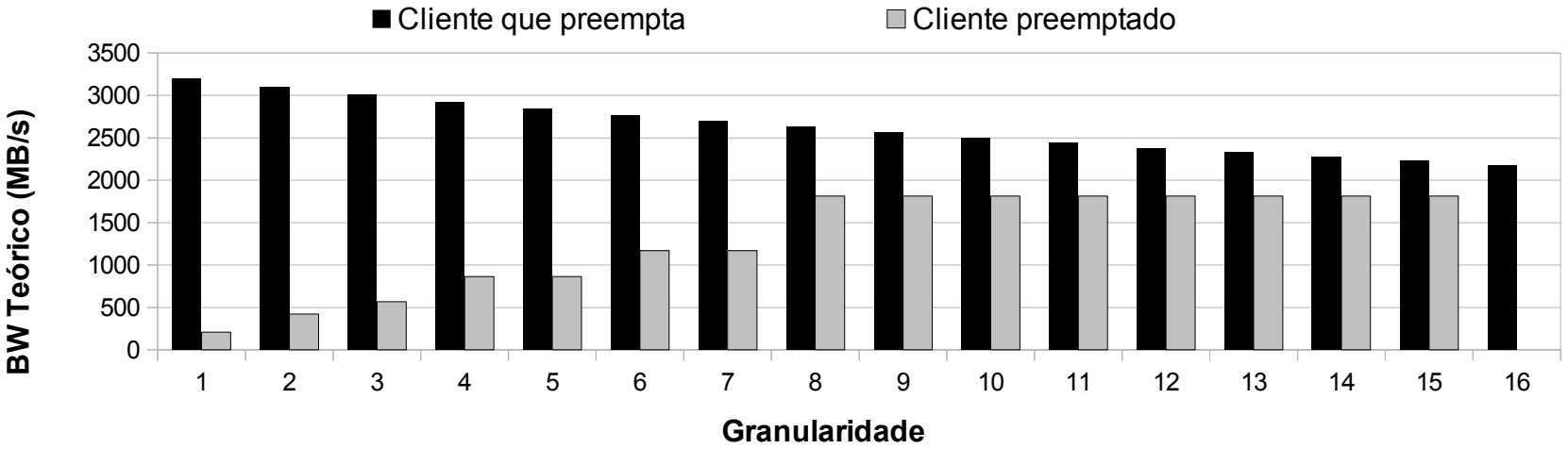


Análise da Granularidade
Diferentes Tamanhos de Transação

- Caso avaliado: 2 clientes

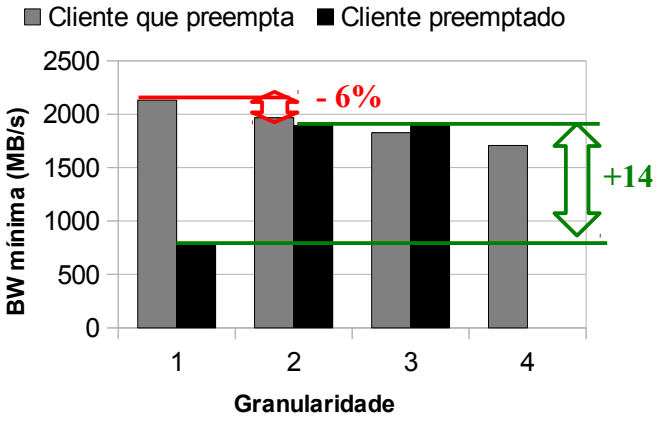


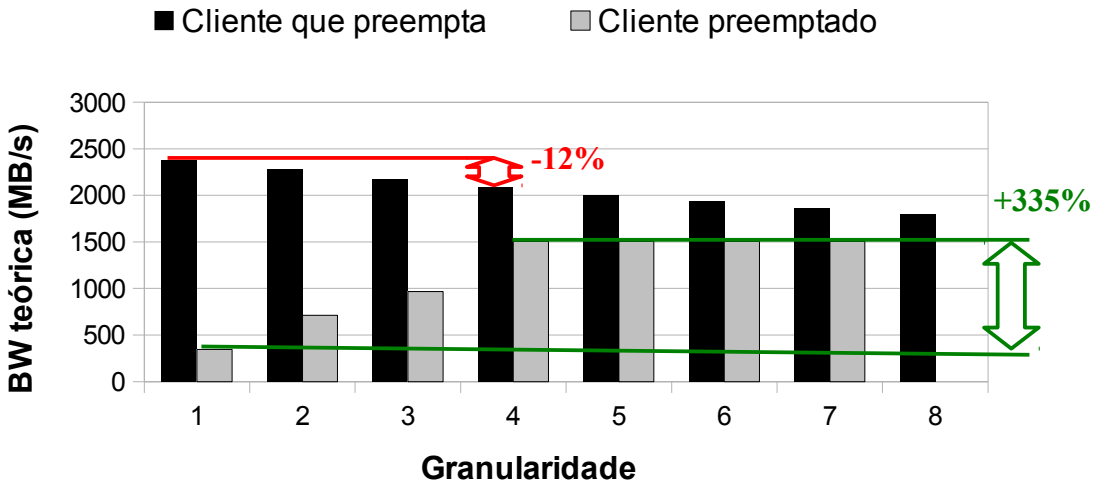


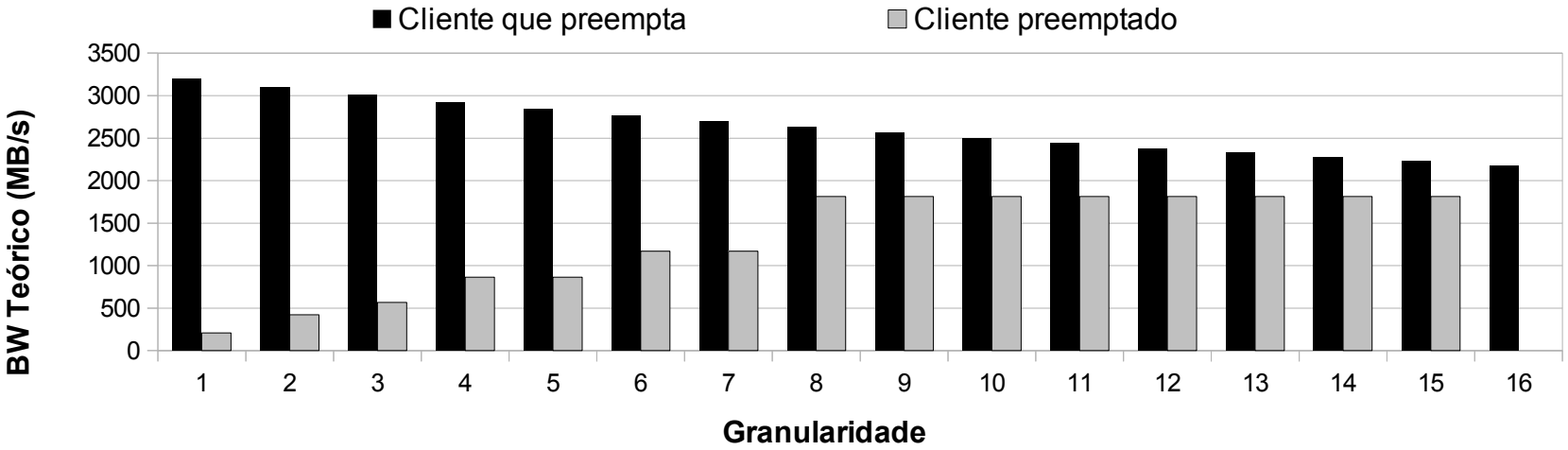


Análise da Granularidade
Diferentes Tamanhos de Transação

- Caso avaliado: 2 clientes

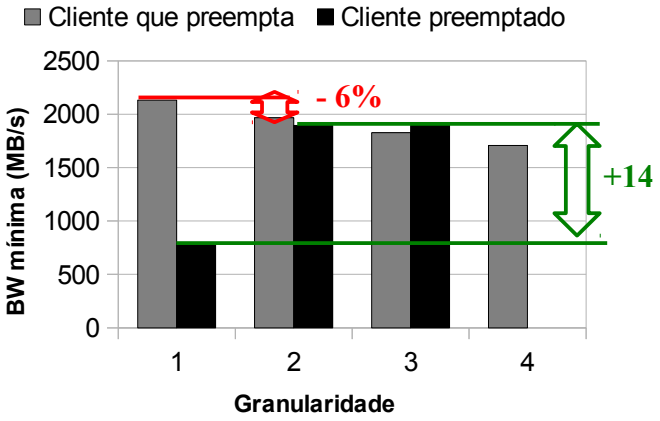


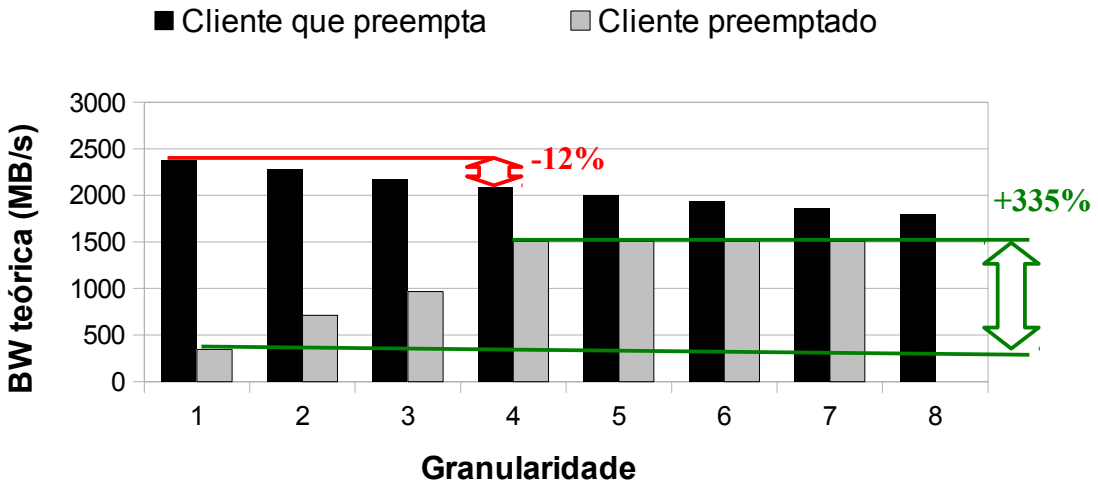


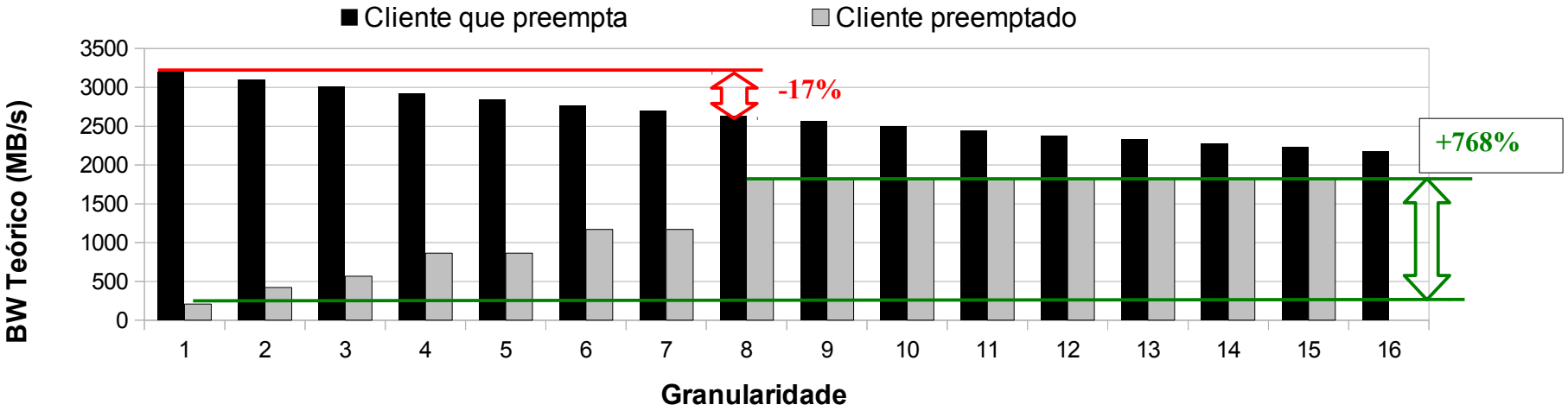


Análise da Granularidade
Diferentes Tamanhos de Transação

- Caso avaliado: 2 clientes



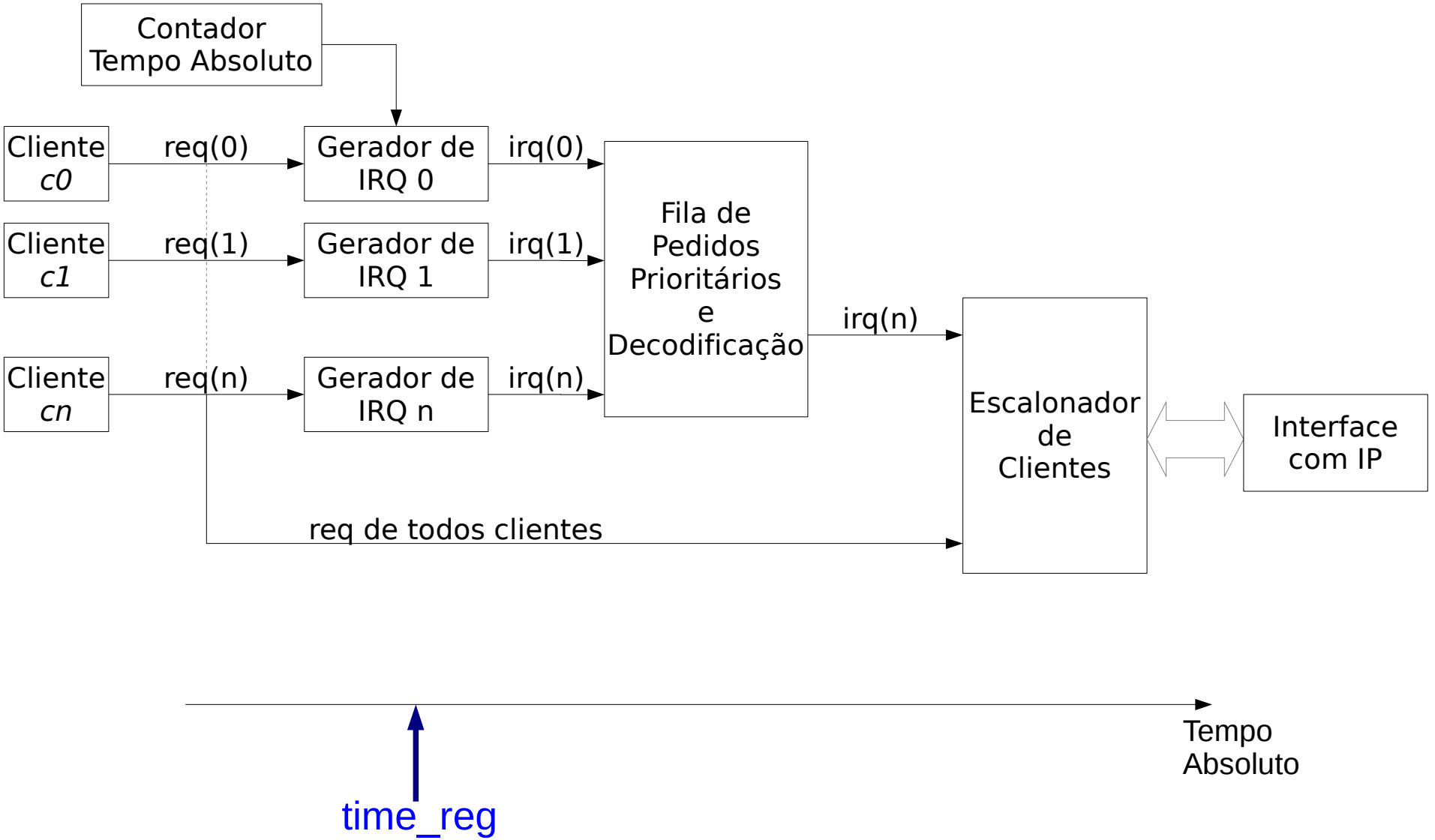




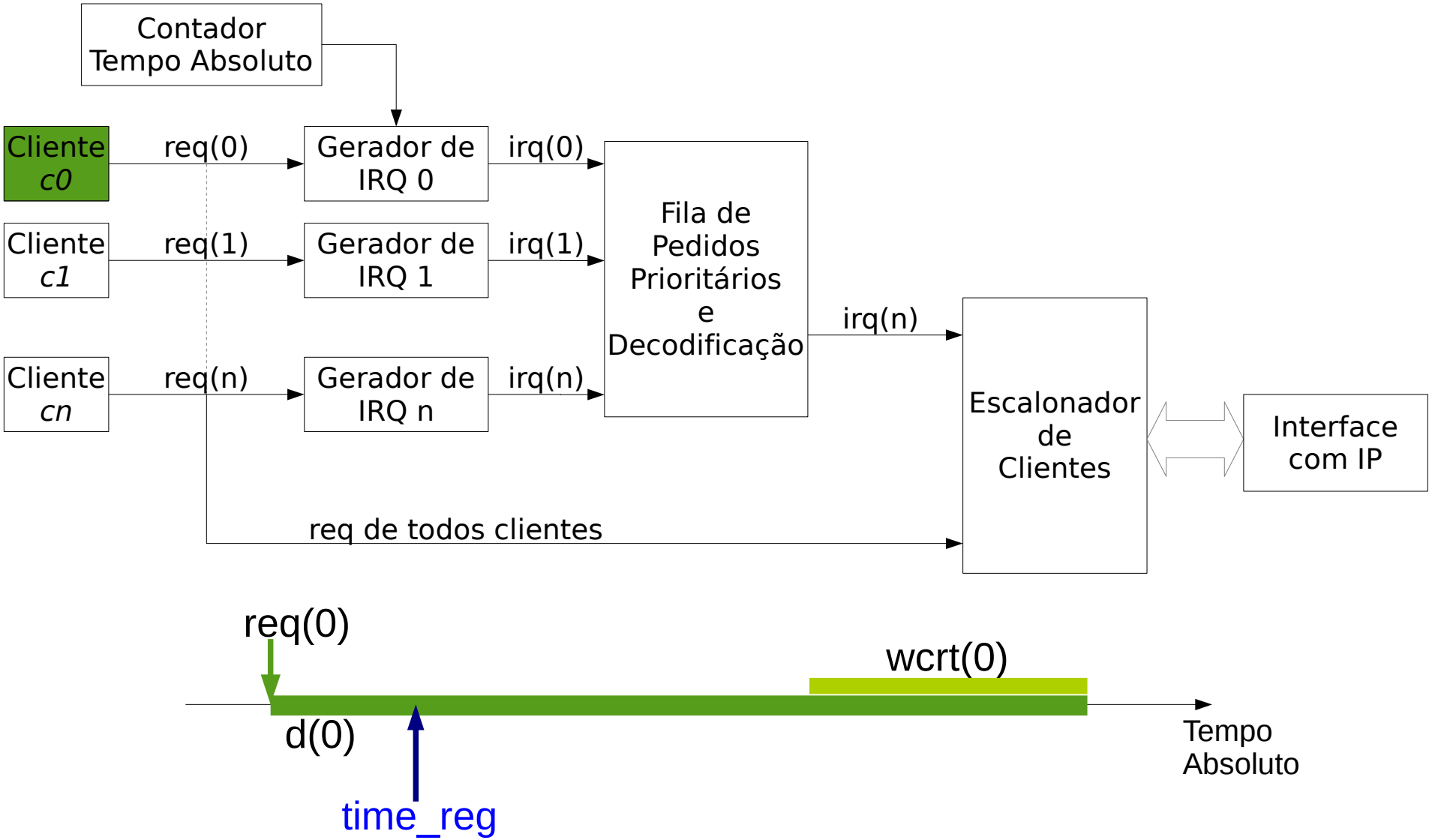
Árbitro Adaptativo
Características da implementação

- Controle de acessos dos clientes:
  - Preempção com granularidade mínima $g(n) = l(n)/2$
- Prioridades adaptativas:
  - Classificação em tempo de execução.
- Análise do WCRT:
  - Implementação com hardware dedicado;
  - $3n+1$ ciclos de relógio para um conjunto de n clientes;
  - Cálculo é realizado a cada mudança dos parâmetros dos clientes.

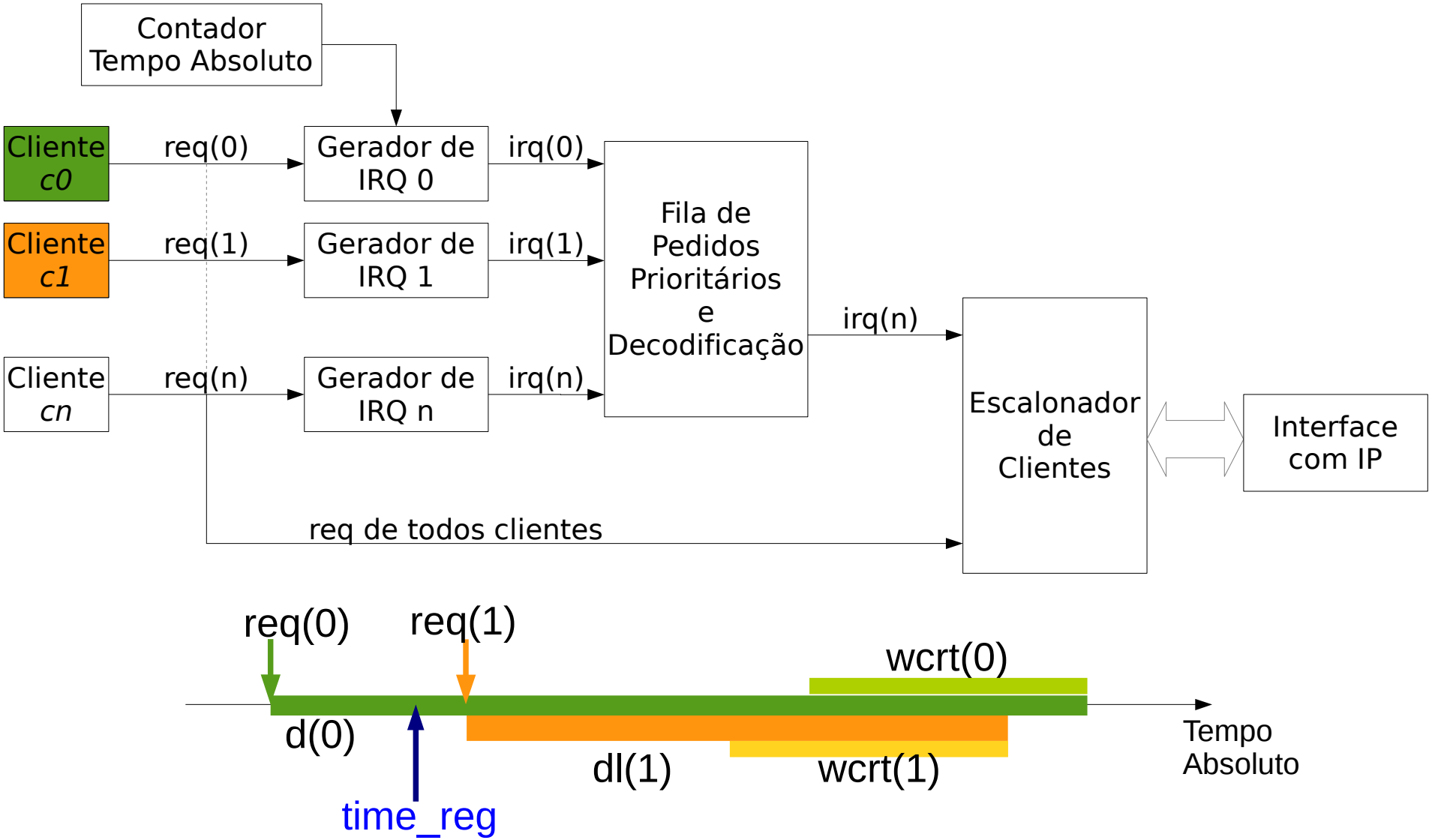
Fila de Pedidos de Interrupção
Implementação e funcionamento



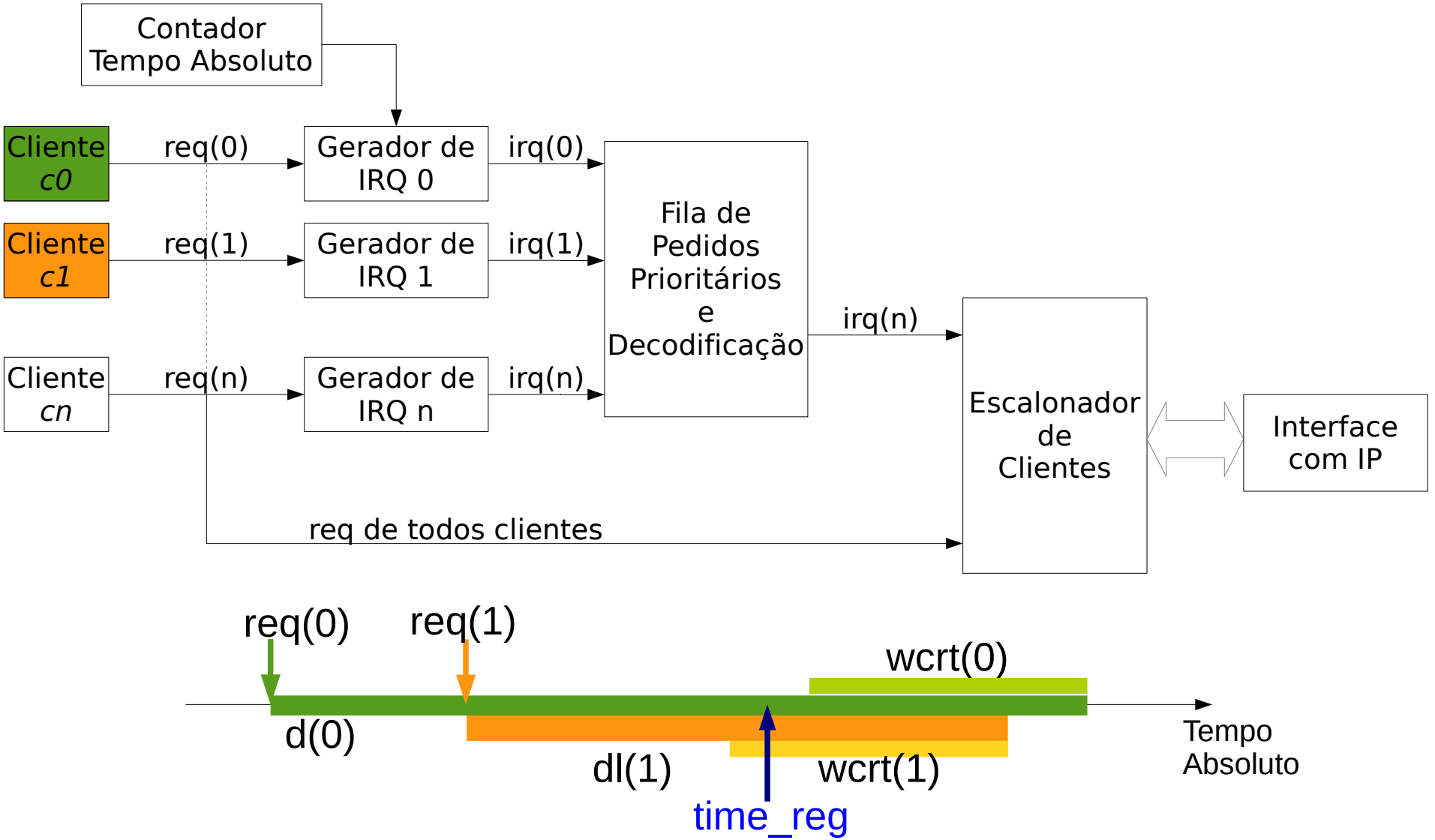
Fila de Pedidos de Interrupção
Implementação e funcionamento



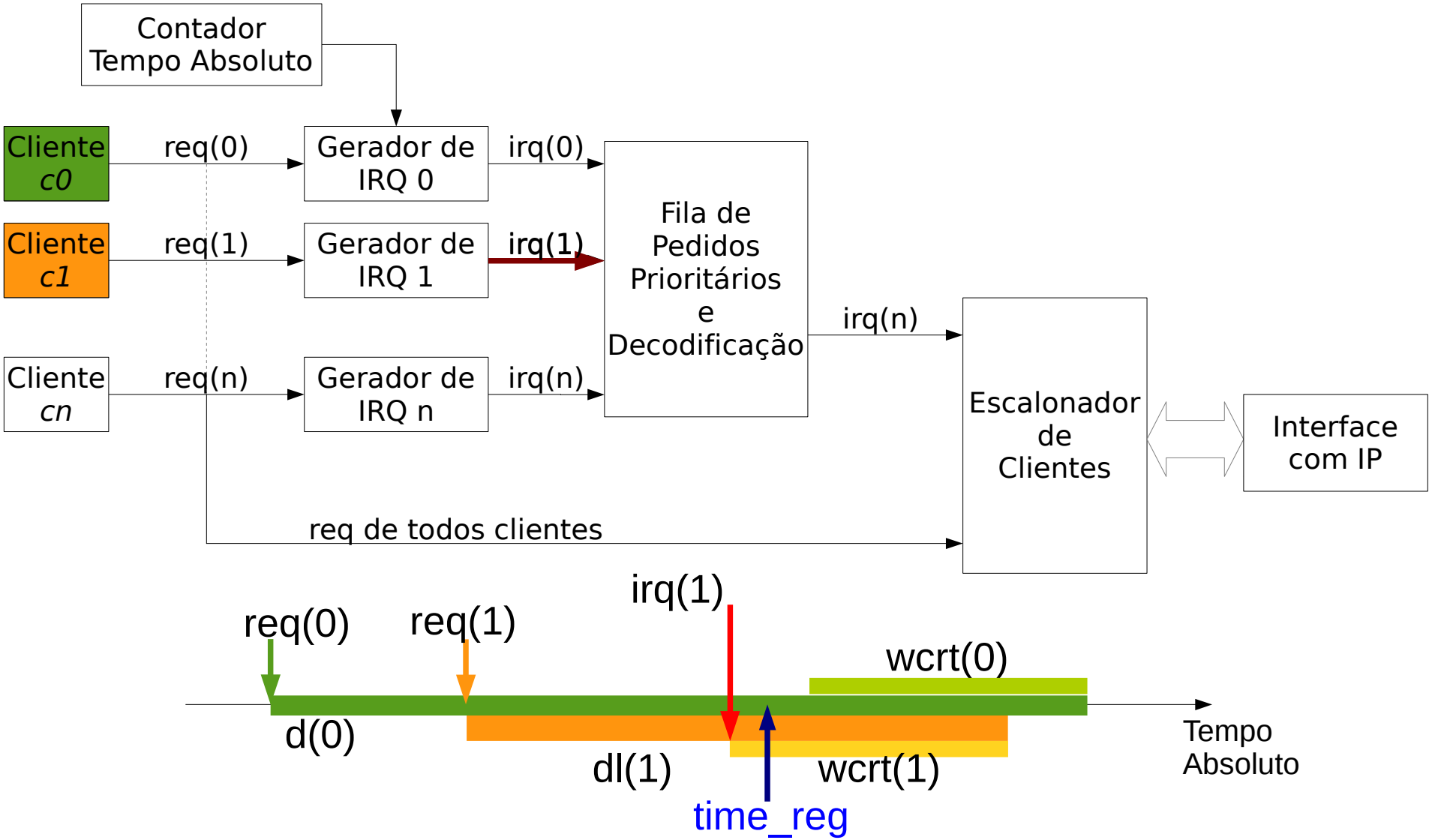
Fila de Pedidos de Interrupção
Implementação e funcionamento



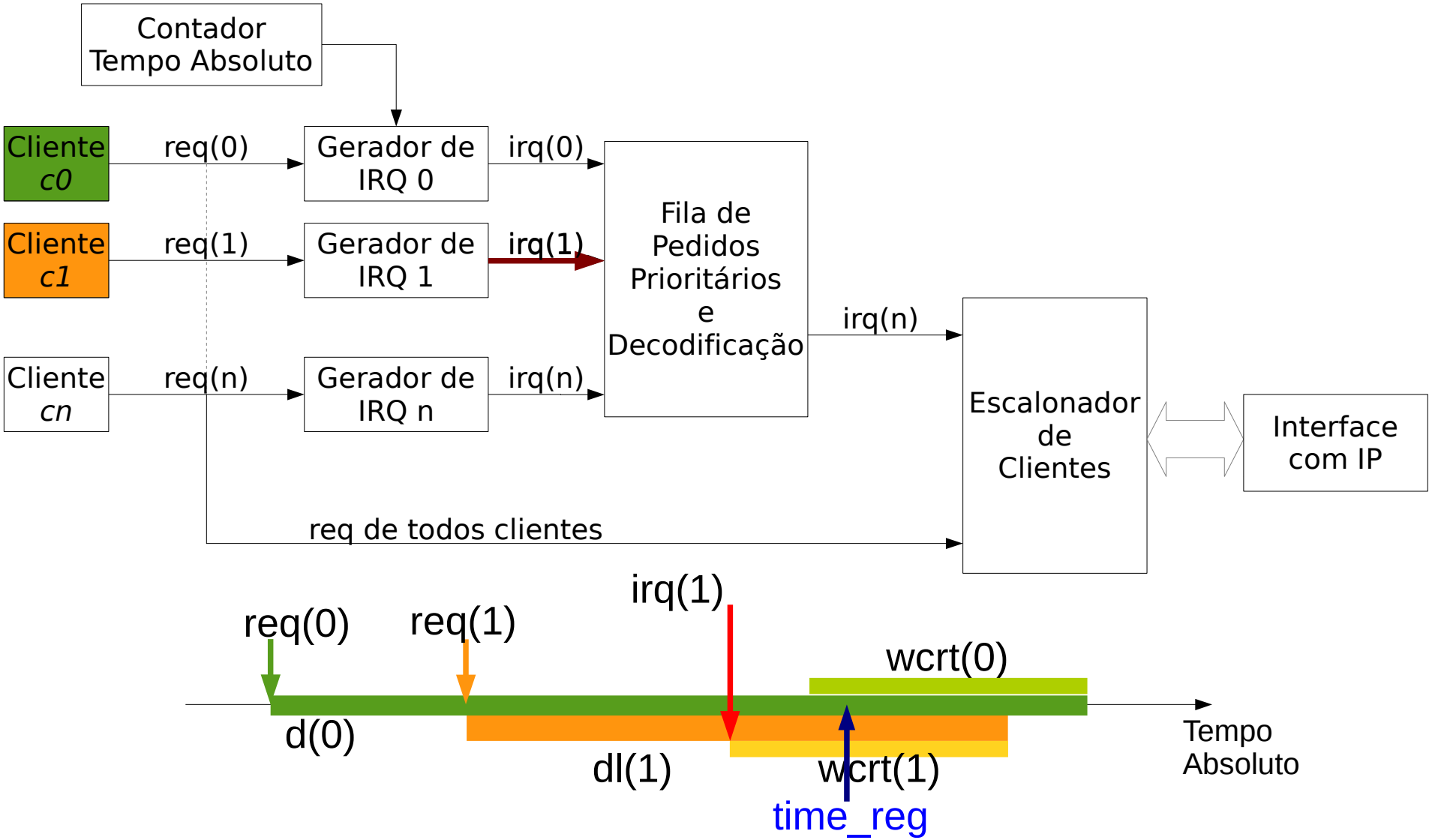
Fila de Pedidos de Interrupção
Implementação e funcionamento



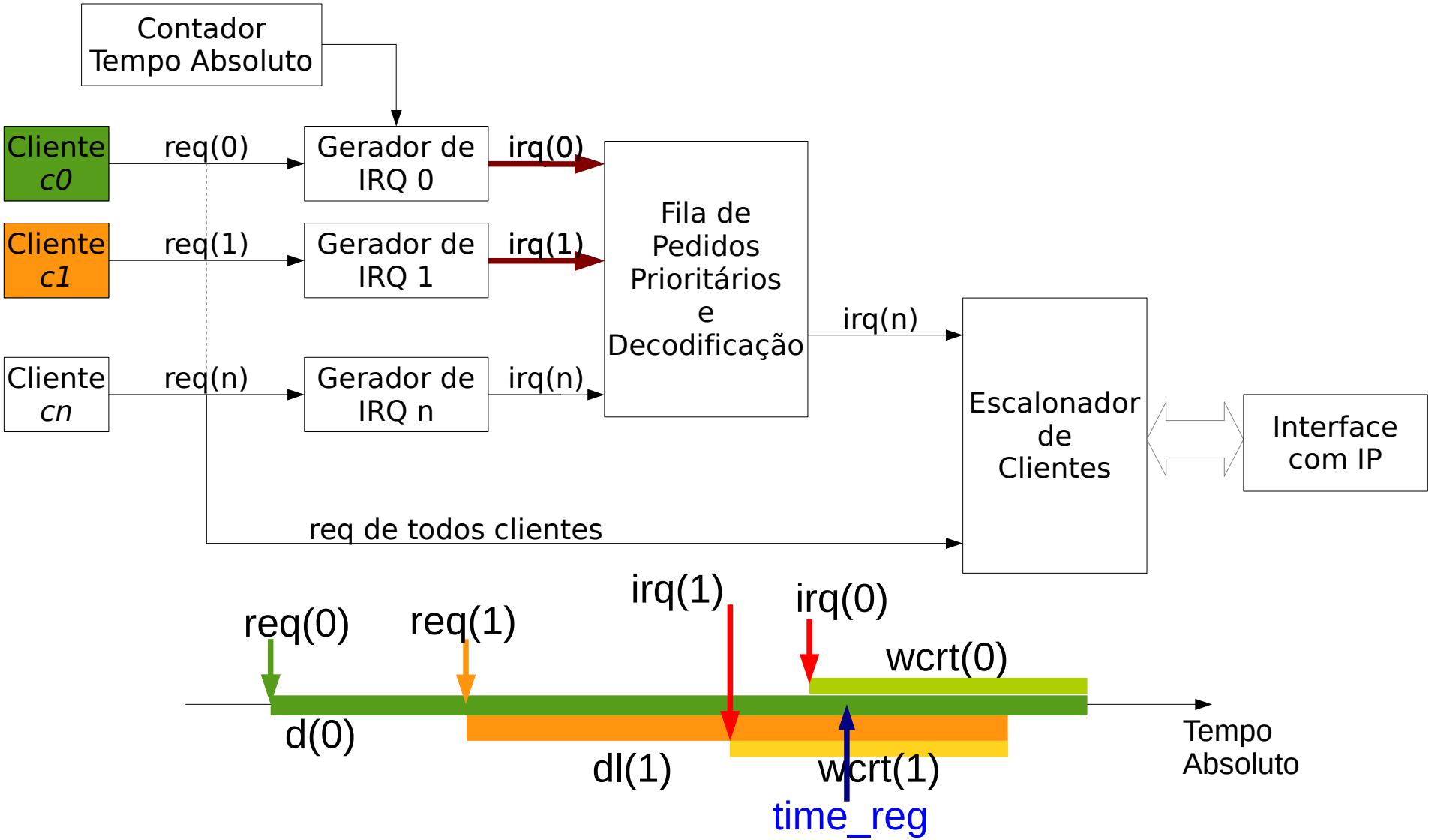
Fila de Pedidos de Interrupção
Implementação e funcionamento



Fila de Pedidos de Interrupção
Implementação e funcionamento



Fila de Pedidos de Interrupção
Implementação e funcionamento



Sumário da Apresentação

- Contextualização do Problema
- Funcionamento da DRAM
- Metodologia
- **Resultados**
- Comentários Finais

Análise do Pior Caso
Tempo de Resposta e Largura de Banda

- Simulação para 2, 4 e 8 clientes compartilhando igualmente o canal de memória:

- DDR3-800 64-bits:

100% BW peak: 6400 MB/s

100% BW sustentada:

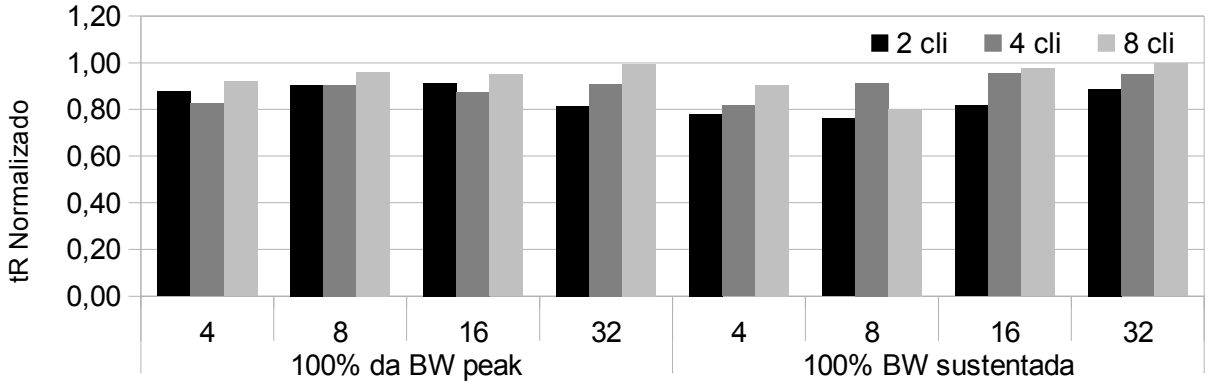
l(n)=4 3064 MB/s

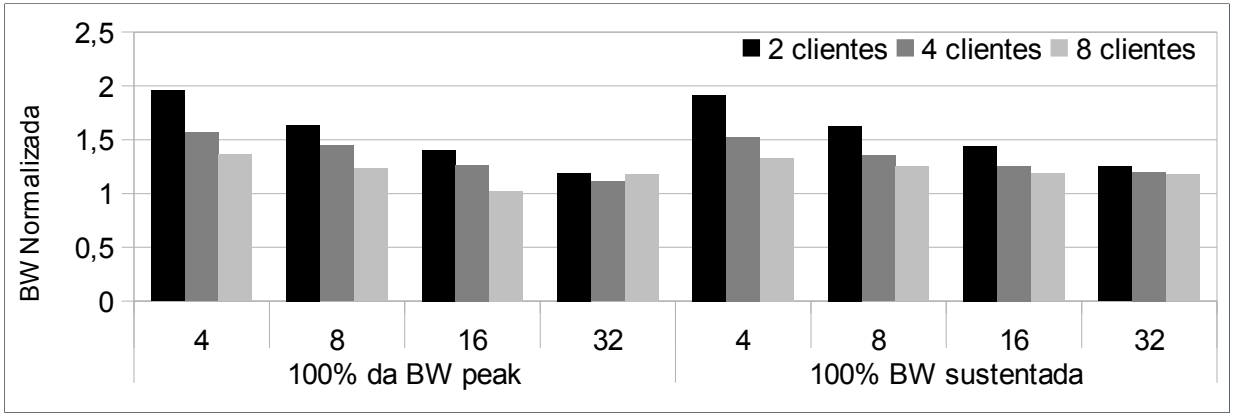
l(n)=8 4098 MB/s

l(n)=16 4952 MB/s

l(n)=32 5542 MB/s

- Modelo garante piores casos de tempo de resposta e largura de banda mínima prevista.

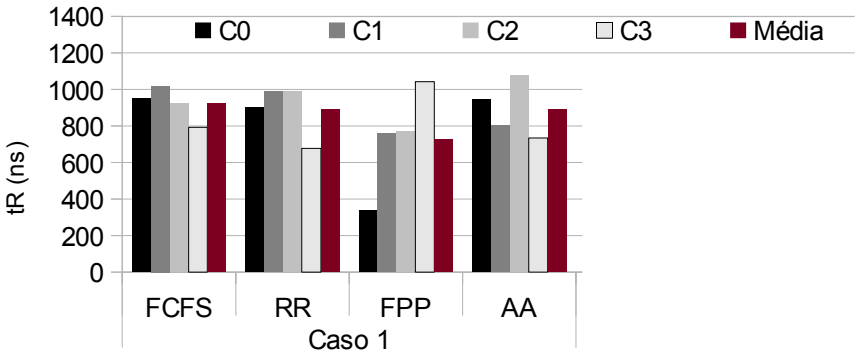


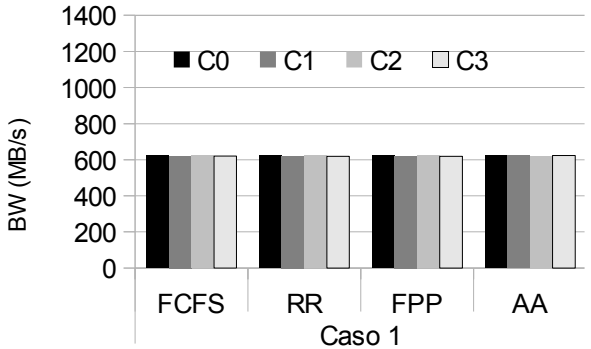


Controle do tempo de resposta
Análise dos piores casos

Caso 1		
	dl (ns)	*BW (MB/s)*
C0	5000	618
C1	5000	618
C2	5000	618
C3	5000	618

- Cada cliente ocupa 12,5% da lagura de banda sustentada para 16 rajadas (4952MB/s);
- Intervalos de acesso de 1655 ns;
- *WCRT = 1055 ns.*

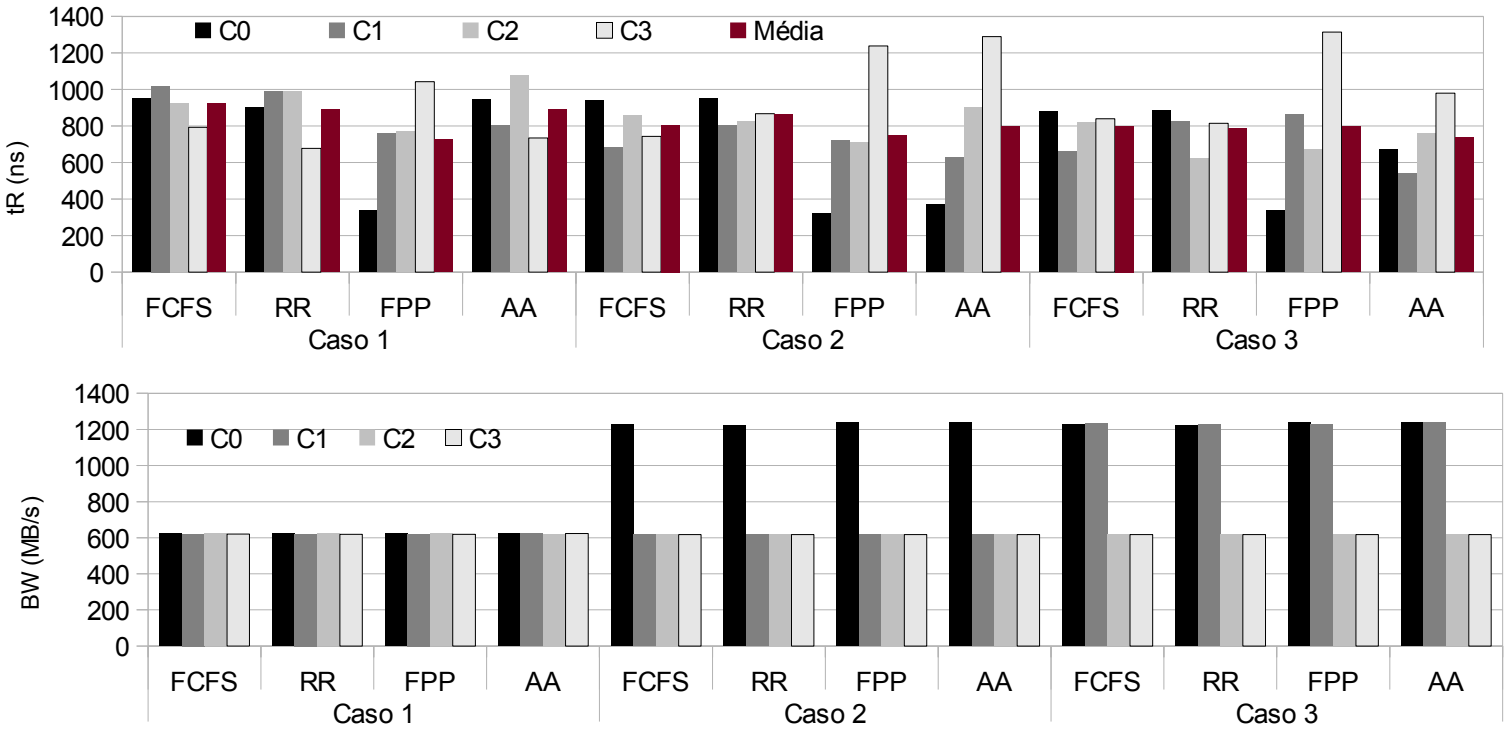




Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 1		
C0	5000	618
C1	5000	618
C2	5000	618
C3	5000	618
Caso 2		
C0	**700**	**1236**
C1	5000	618
C2	5000	618
C3	5000	618
Caso 3		
C0	**700**	**1236**
C1	**700**	**1236**
C2	5000	618
C3	5000	618

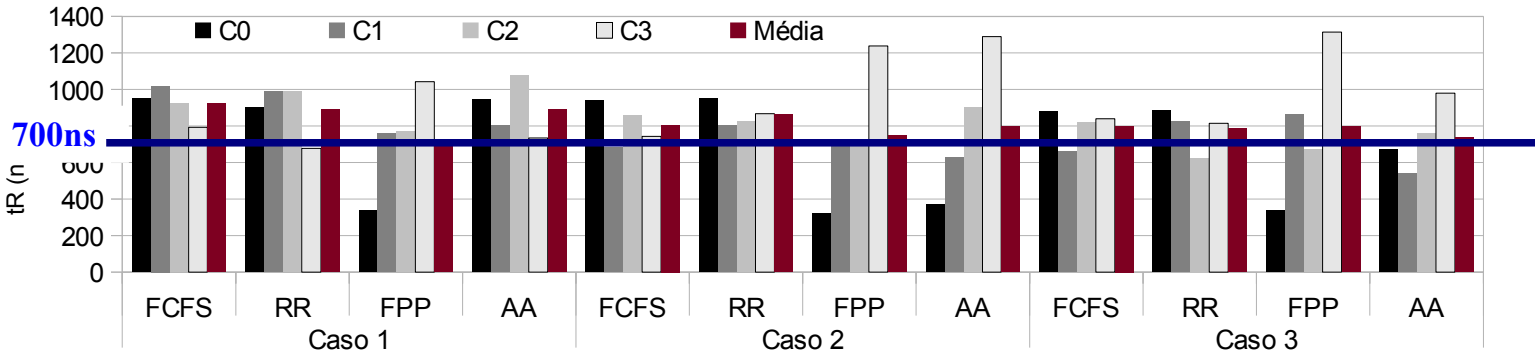
- Cada cliente ocupa 12,5% da lagura de banda sustentada para 16 rajadas (4952MB/s);
- Intervalos de acesso de 1655 ns;
- *WCRT = 1055 ns.*

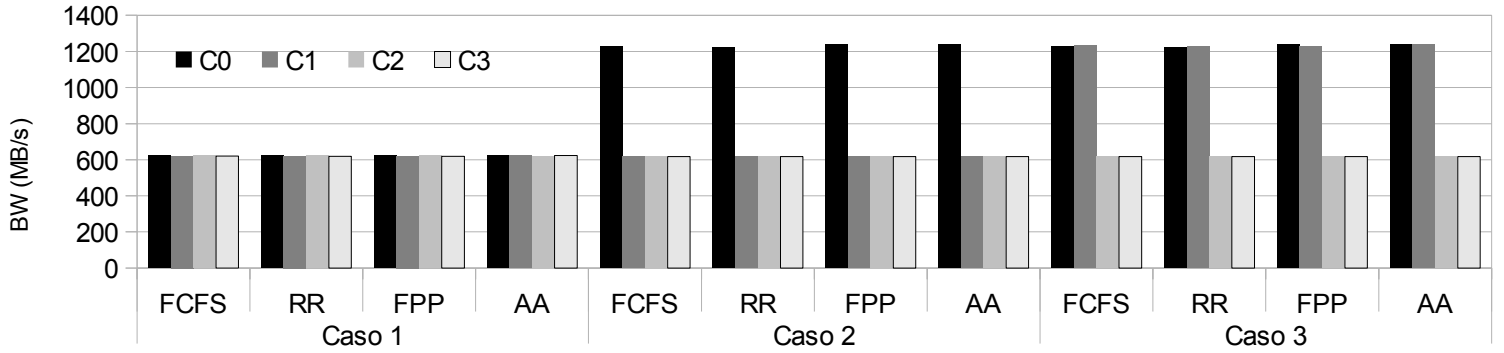


Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 1		
C0	5000	618
C1	5000	618
C2	5000	618
C3	5000	618
Caso 2		
C0	**700**	**1236**
C1	5000	618
C2	5000	618
C3	5000	618
Caso 3		
C0	**700**	**1236**
C1	**700**	**1236**
C2	5000	618
C3	5000	618

- Cada cliente ocupa 12,5% da lagura de banda sustentada para 16 rajadas (4952MB/s);
- Intervalos de acesso de 1655 ns;
- *WCRT = 1055 ns.*

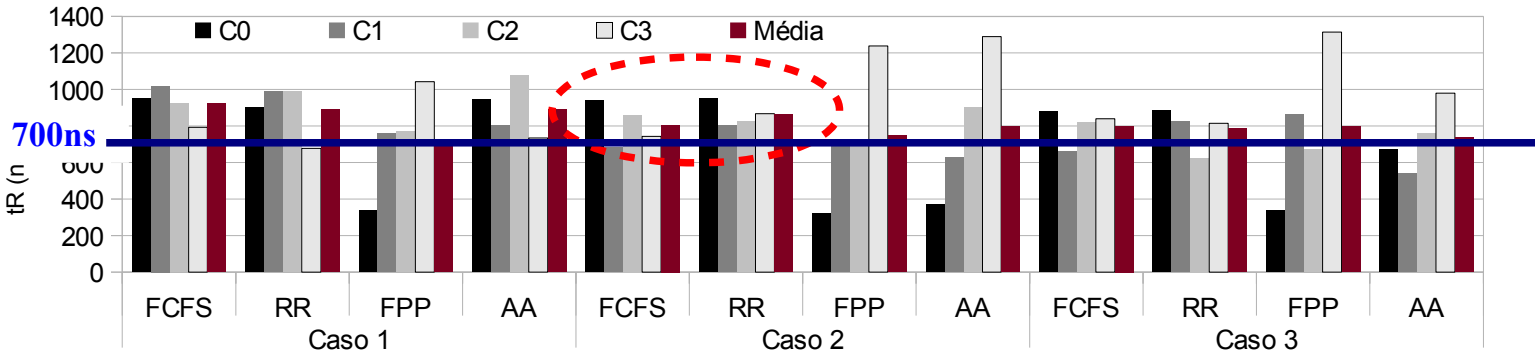


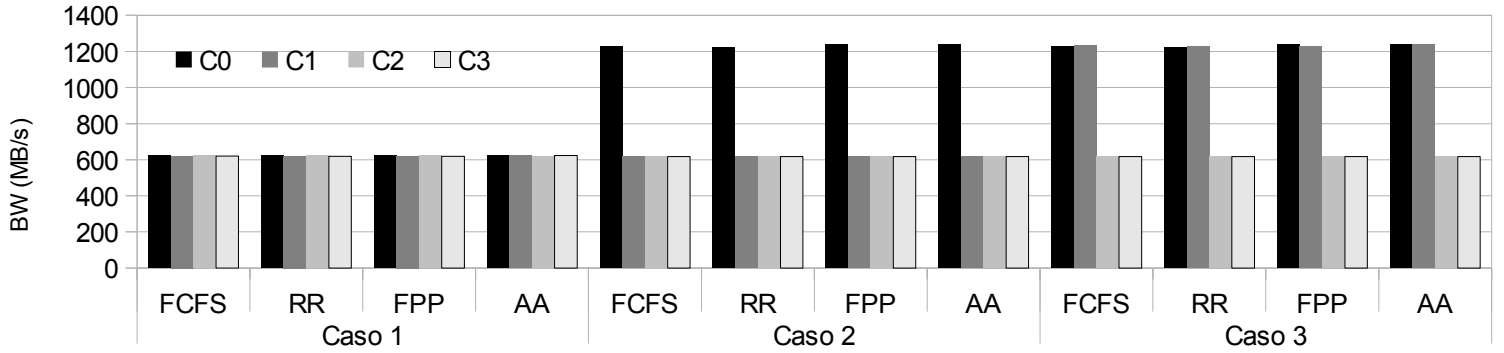


Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 1		
C0	5000	618
C1	5000	618
C2	5000	618
C3	5000	618
Caso 2		
C0	**700**	**1236**
C1	5000	618
C2	5000	618
C3	5000	618
Caso 3		
C0	**700**	**1236**
C1	**700**	**1236**
C2	5000	618
C3	5000	618

- Cada cliente ocupa 12,5% da lagura de banda sustentada para 16 rajadas (4952MB/s);
- Intervalos de acesso de 1655 ns;
- *WCRT = 1055 ns.*

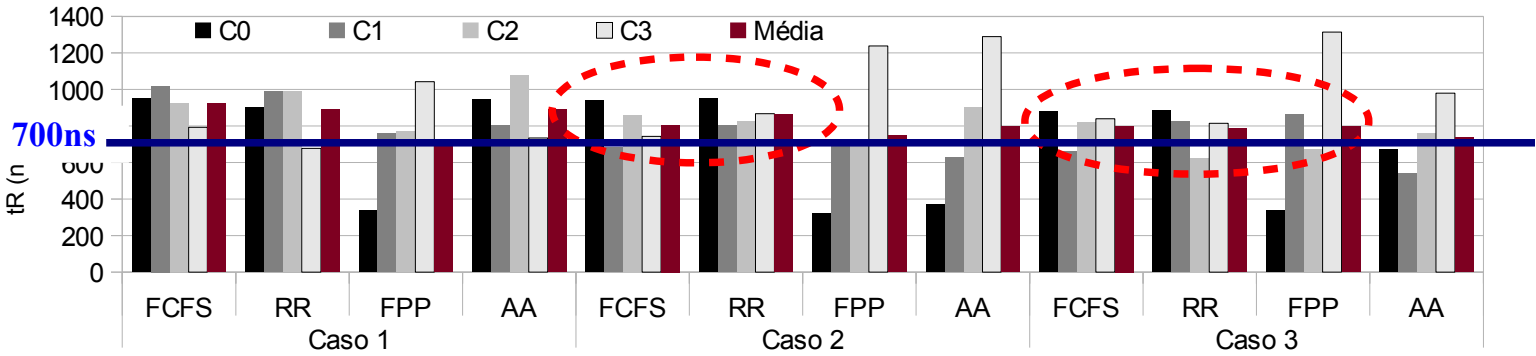


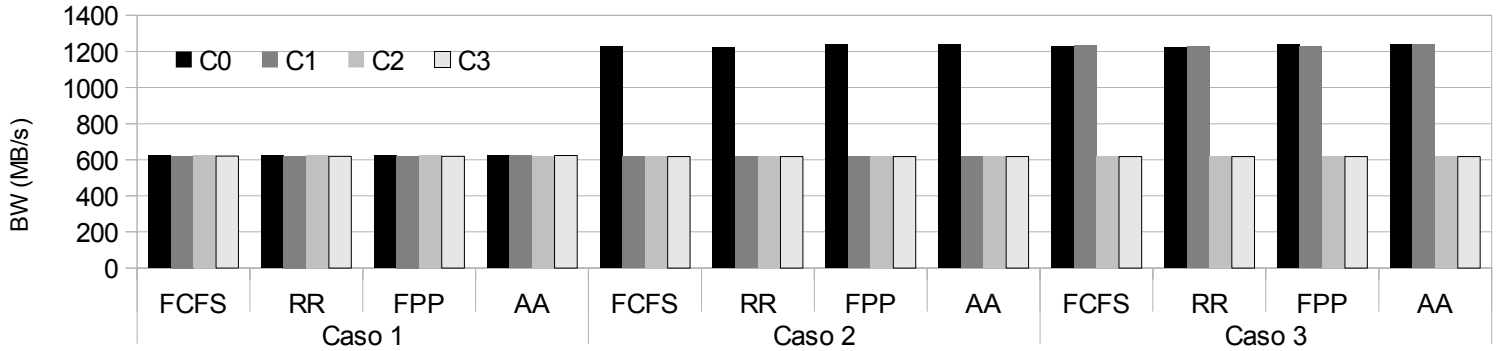


Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 1		
C0	5000	618
C1	5000	618
C2	5000	618
C3	5000	618
Caso 2		
C0	**700**	**1236**
C1	5000	618
C2	5000	618
C3	5000	618
Caso 3		
C0	**700**	**1236**
C1	**700**	**1236**
C2	5000	618
C3	5000	618

- Cada cliente ocupa 12,5% da lagura de banda sustentada para 16 rajadas (4952MB/s);
- Intervalos de acesso de 1655 ns;
- *WCRT = 1055 ns.*

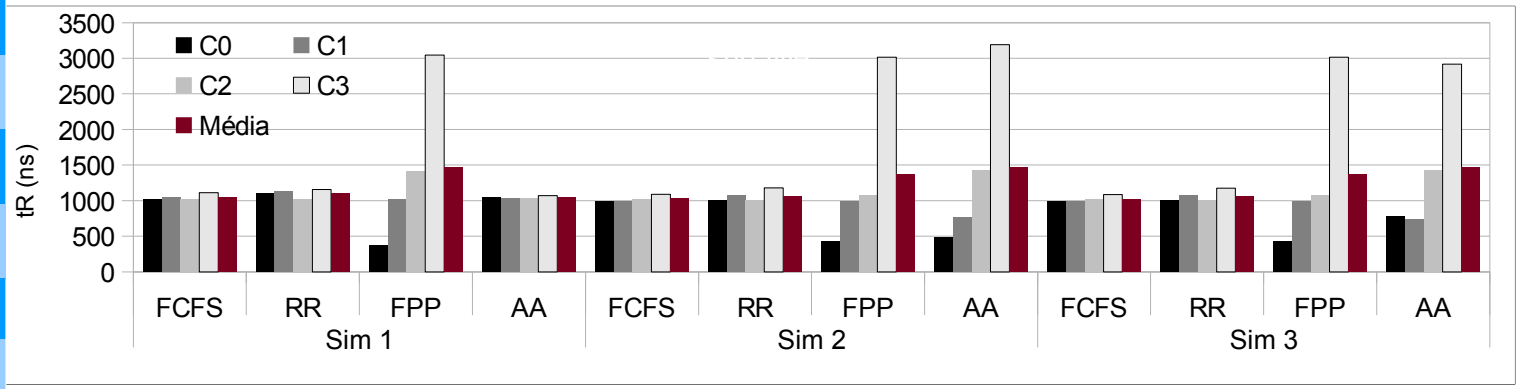


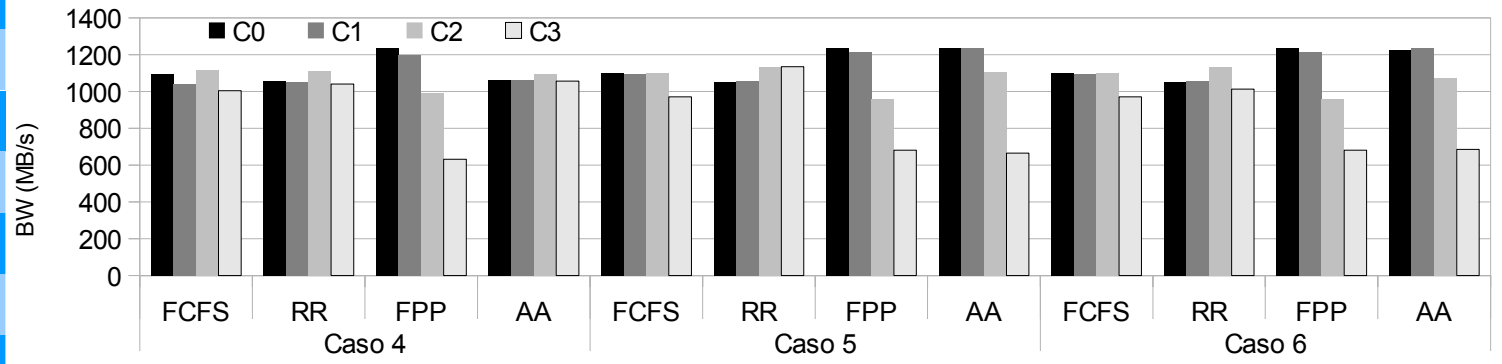


Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 4		
C0	5000	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 5		
C0	**800**	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 6		
C0	**800**	1236
C1	**800**	1236
C2	5000	1236
C3	5000	1236

- Cada cliente ocupa 25% da lagura de banda sustentada para 16 rajadas (4950MB/s)
- Intervalos de acesso de 825 ns
- *WCRT = 1055 ns*

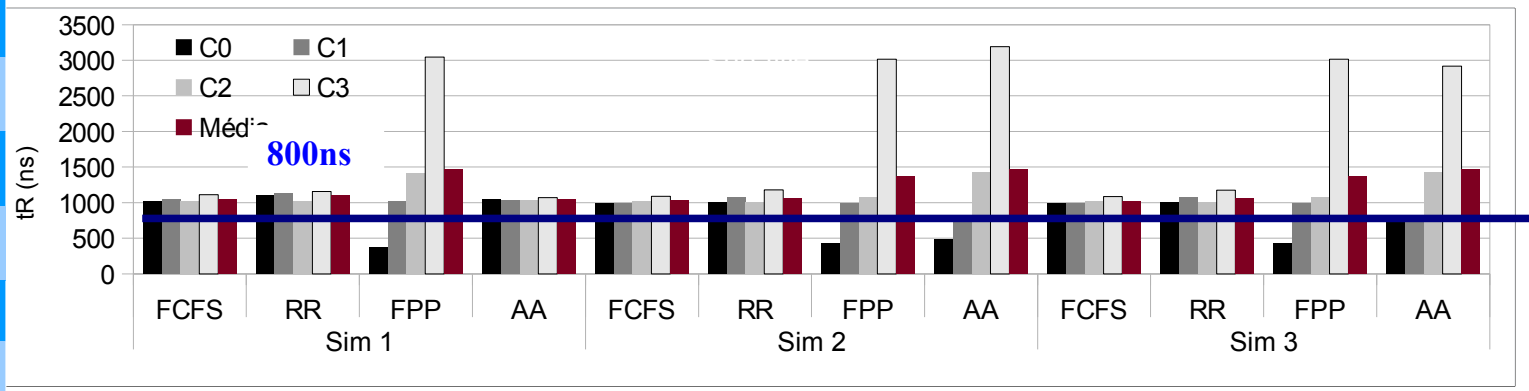


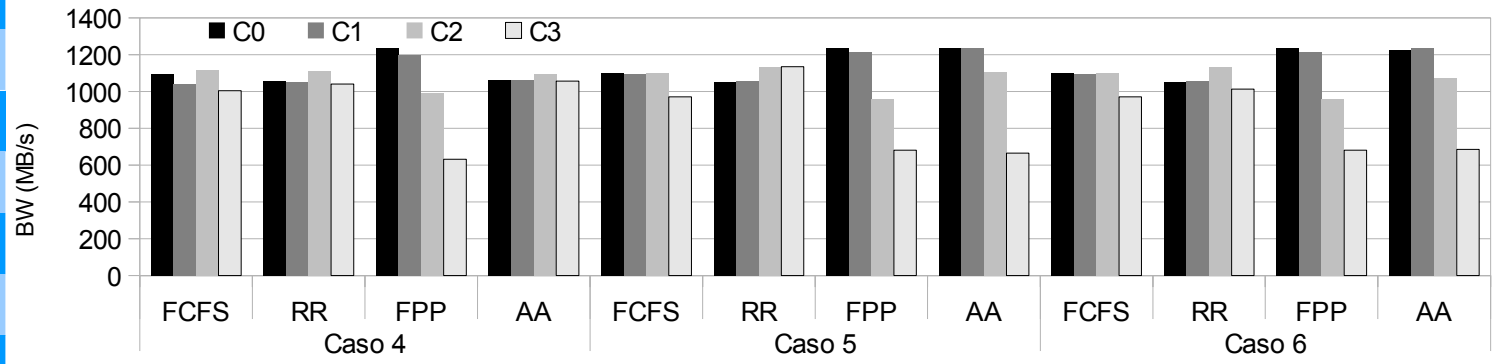


Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 4		
C0	5000	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 5		
C0	**800**	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 6		
C0	**800**	1236
C1	**800**	1236
C2	5000	1236
C3	5000	1236

- Cada cliente ocupa 25% da lagura de banda sustentada para 16 rajadas (4950MB/s)
- Intervalos de acesso de 825 ns
- *WCRT = 1055 ns*

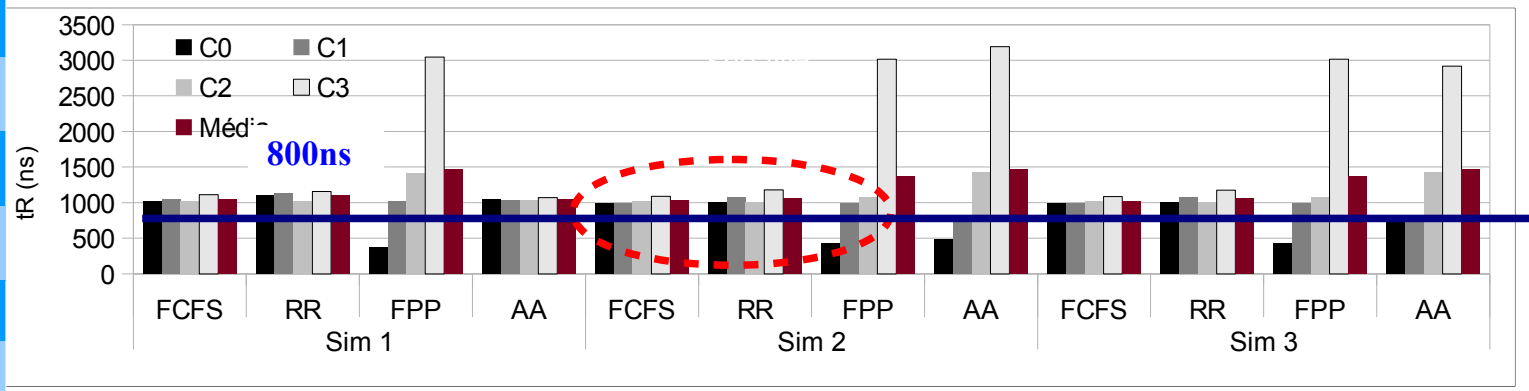


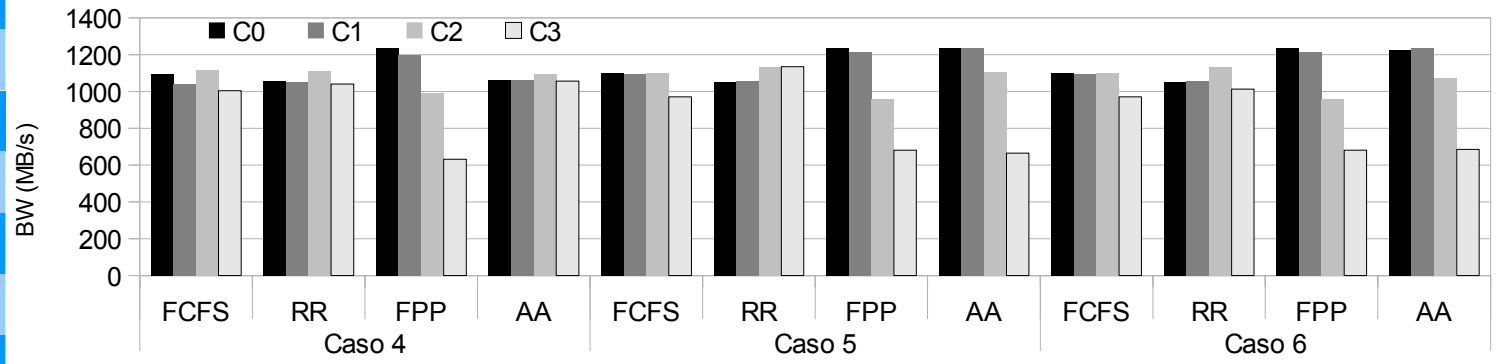


Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 4		
C0	5000	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 5		
C0	**800**	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 6		
C0	**800**	1236
C1	**800**	1236
C2	5000	1236
C3	5000	1236

- Cada cliente ocupa 25% da lagura de banda sustentada para 16 rajadas (4950MB/s)
- Intervalos de acesso de 825 ns
- *WCRT = 1055 ns*

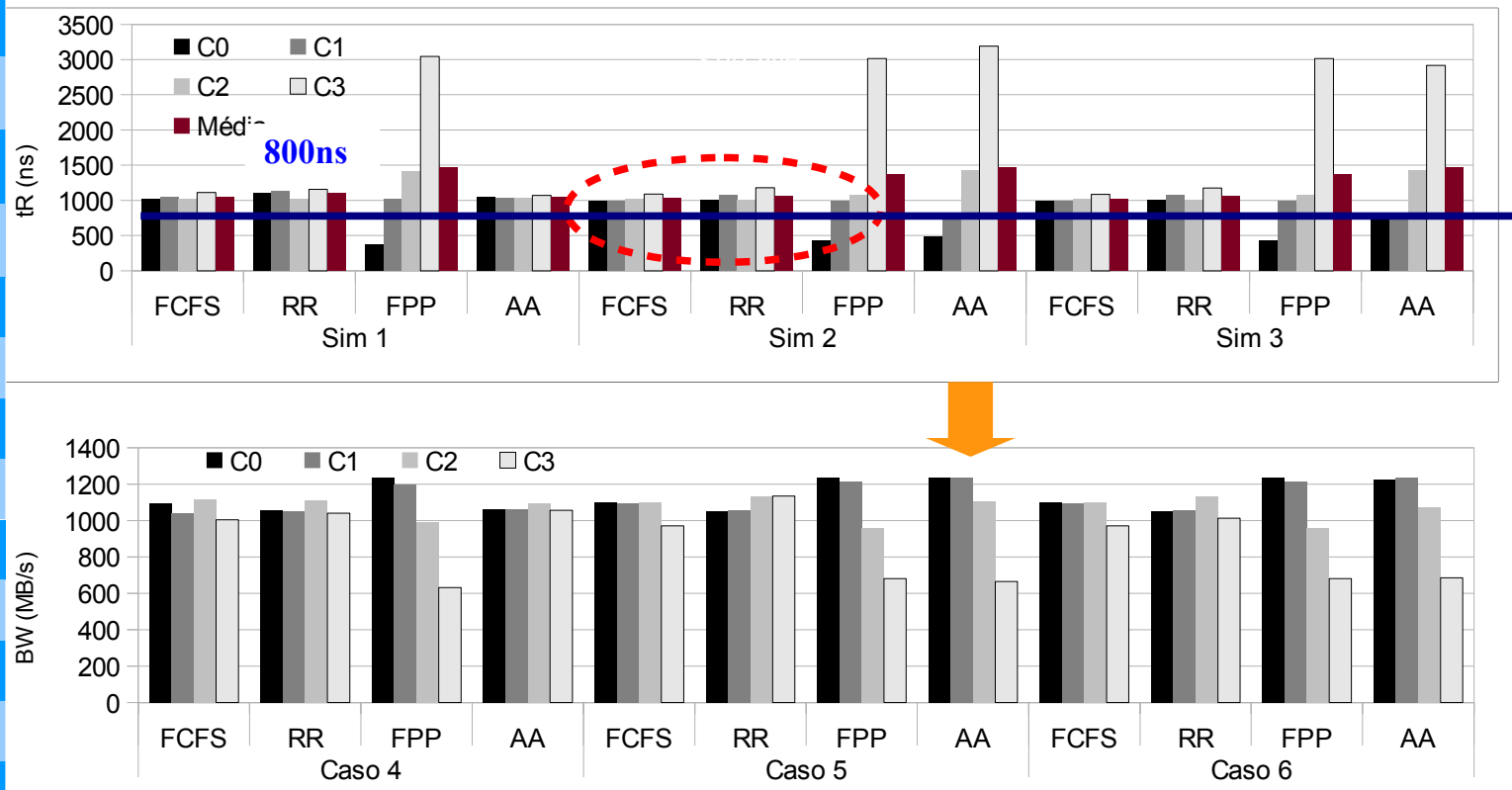




Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 4		
C0	5000	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 5		
C0	800	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 6		
C0	800	1236
C1	800	1236
C2	5000	1236
C3	5000	1236

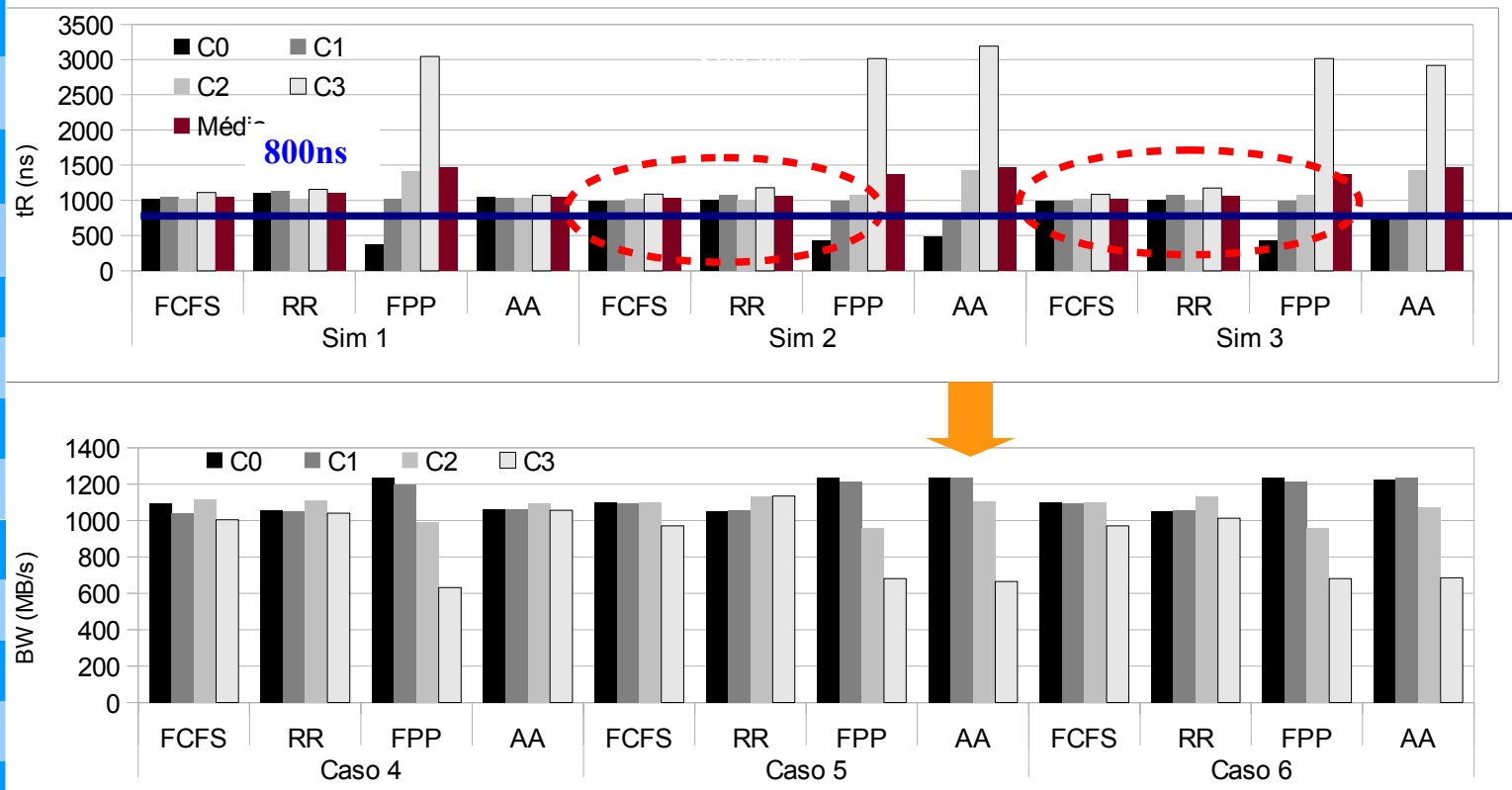
- Cada cliente ocupa 25% da lagura de banda sustentada para 16 rajadas (4950MB/s)
- Intervalos de acesso de 825 ns
- *WCRT = 1055 ns*



Controle do tempo de resposta
Análise dos piores casos

	dl (ns)	BW (MB/s)
Caso 4		
C0	5000	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 5		
C0	**800**	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 6		
C0	**800**	1236
C1	**800**	1236
C2	5000	1236
C3	5000	1236

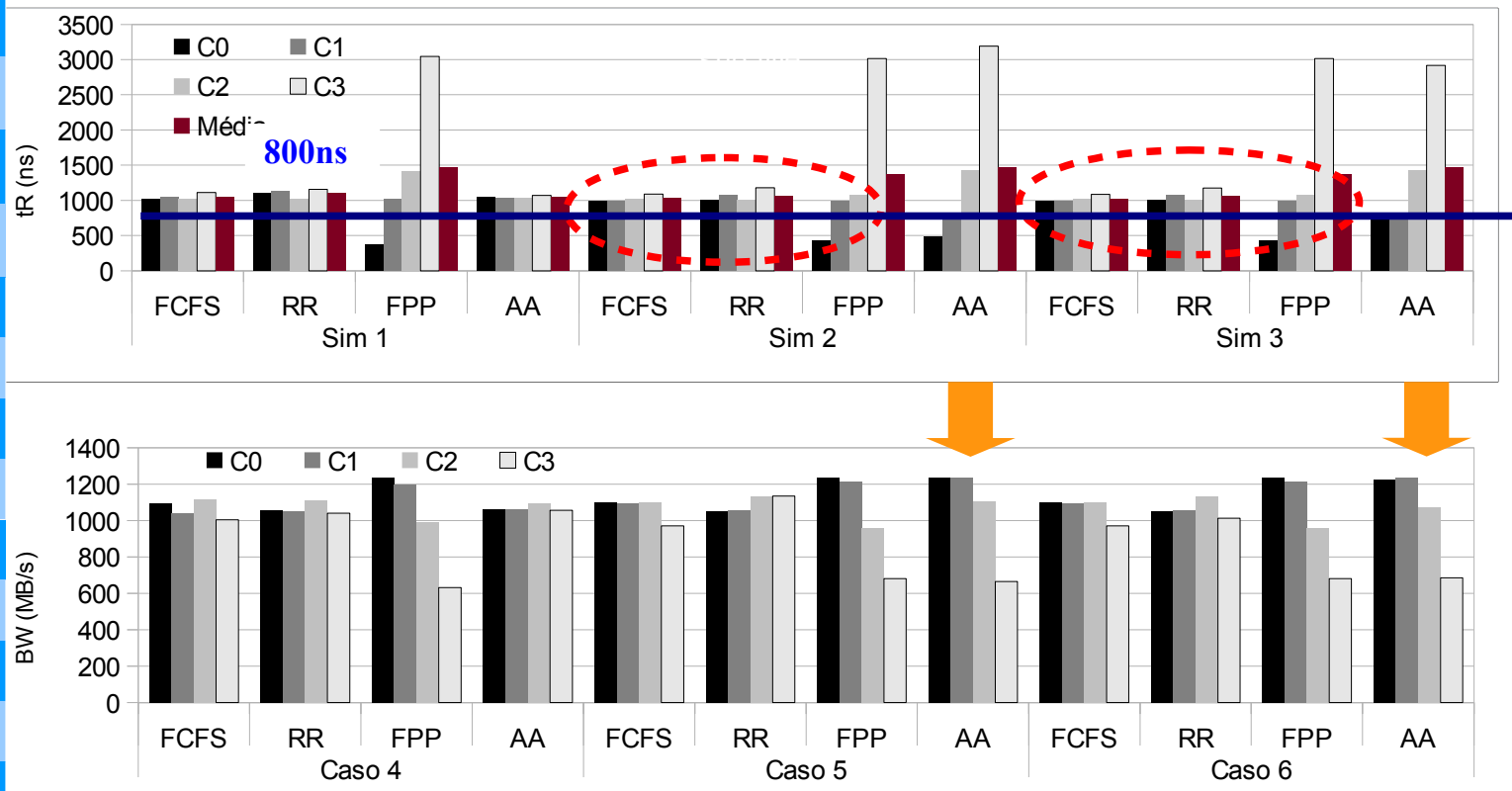
- Cada cliente ocupa 25% da lagura de banda sustentada para 16 rajadas (4950MB/s)
- Intervalos de acesso de 825 ns
- *WCRT = 1055 ns*



Controle do tempo de resposta
Análise dos piores casos

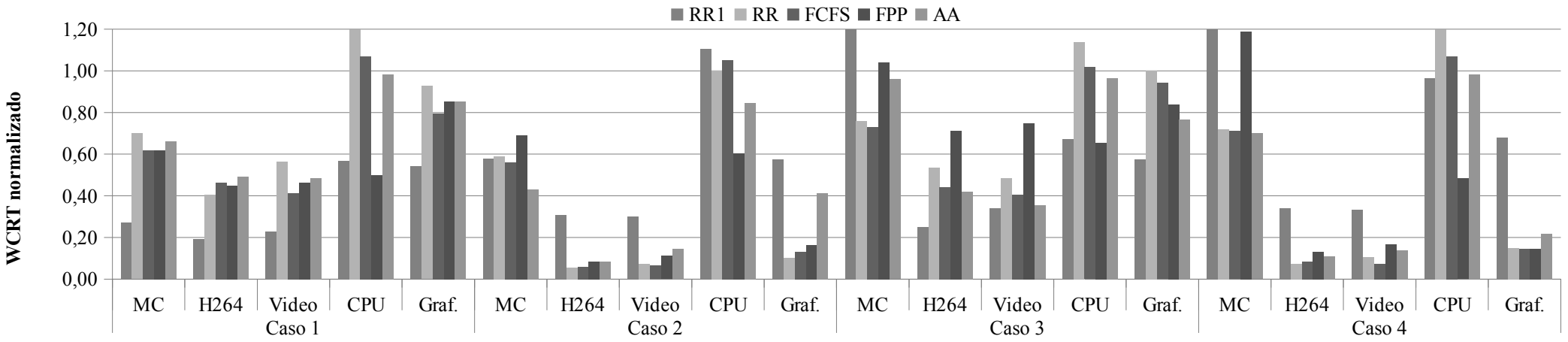
	dl (ns)	BW (MB/s)
Caso 4		
C0	5000	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 5		
C0	**800**	1236
C1	5000	1236
C2	5000	1236
C3	5000	1236
Caso 6		
C0	**800**	1236
C1	**800**	1236
C2	5000	1236
C3	5000	1236

- Cada cliente ocupa 25% da lagura de banda sustentada para 16 rajadas (4950MB/s)
- Intervalos de acesso de 825 ns
- *WCRT = 1055 ns*



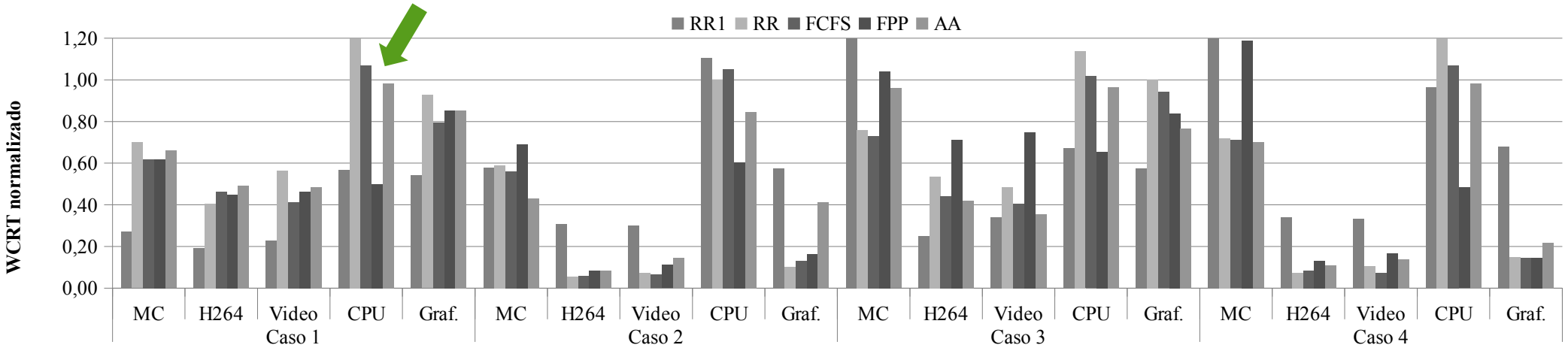
Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	I(n)	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



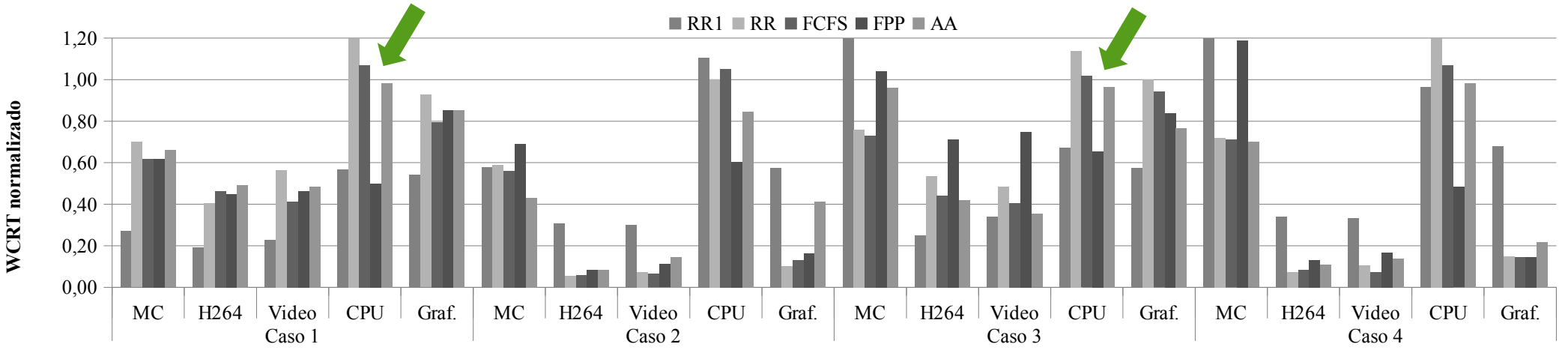
Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	I(n)	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



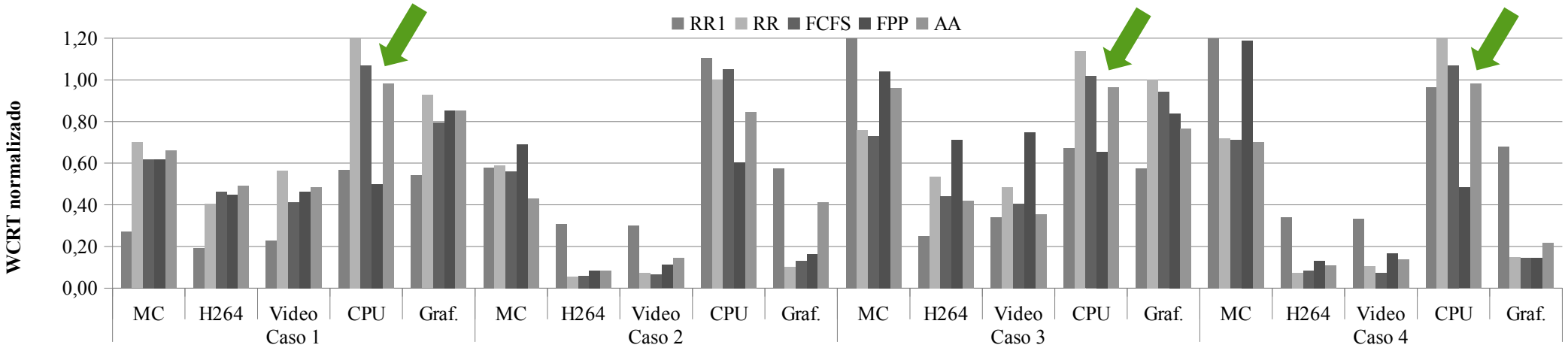
Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	I(n)	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



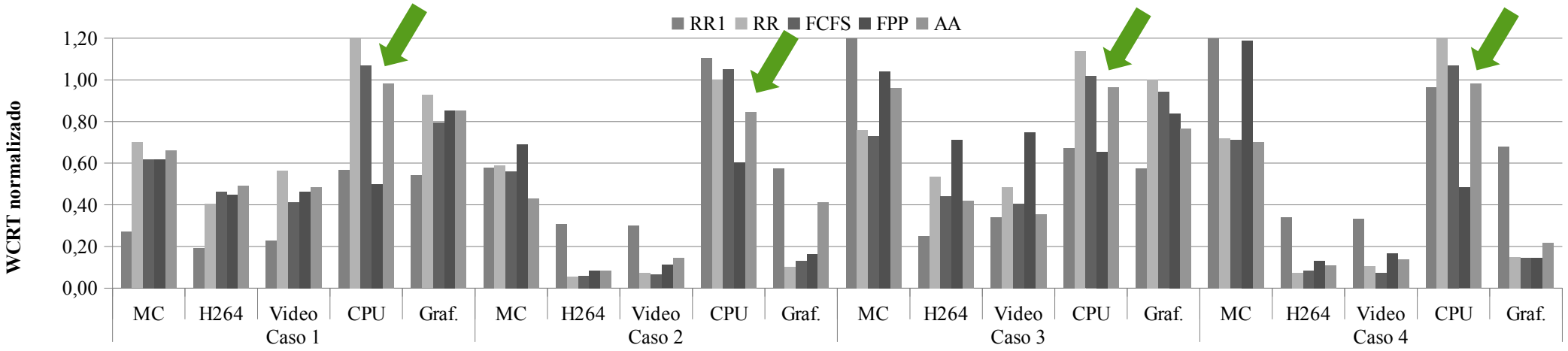
Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	l(n)	*dl(n)*	*l(n)*	*dl(n)*	*l(n)*	*dl(n)*	*l(n)*	*dl(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



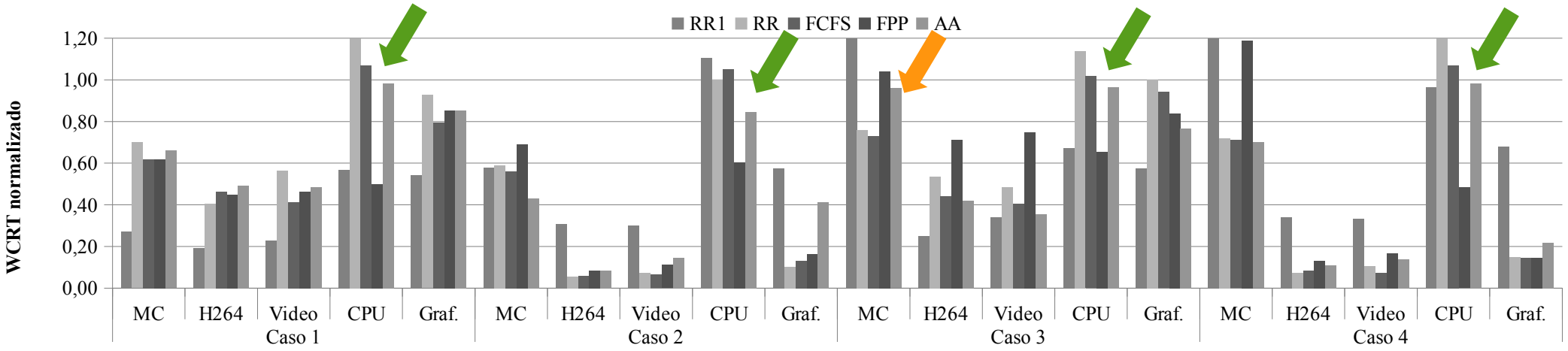
Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	I(n)	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



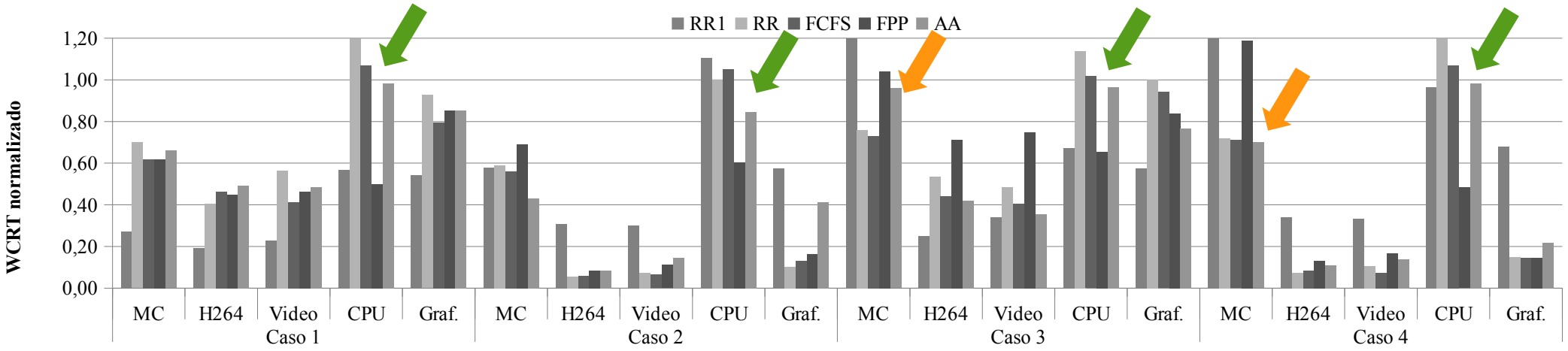
Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	I(n)	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



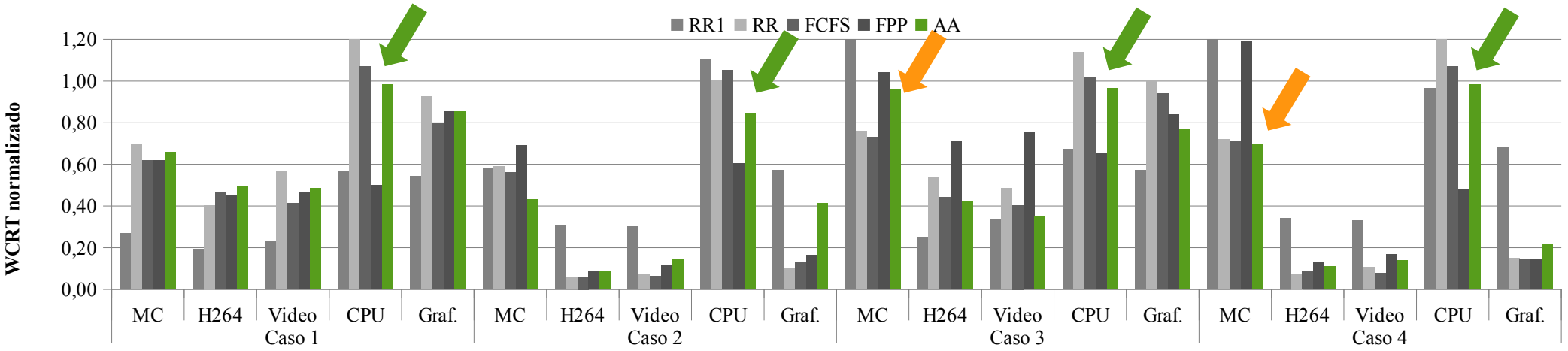
Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	I(n)	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*	*I(n)*	*dI(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



Set-top Box de Televisão Digital

	Caso 1		Caso 2		Caso 3		Caso 4	
	l(n)	*dl(n)*	*l(n)*	*dl(n)*	*l(n)*	*dl(n)*	*l(n)*	*dl(n)*
CPU	1	290 ns	1	290 ns	1	290 ns	1	290 ns
MC	1	500 ns	1	500 ns	18	500 ns	18	500 ns
H264	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Video	1	680 ns	6	4100 ns	1	680 ns	6	4100 ns
Gráfico	1	340 ns	6	2050 ns	1	340 ns	6	2050 ns
WCRT	530 ns		680 ns		700 ns		850 ns	



Sumário da Apresentação

- Contextualização do Problema
- Funcionamento da DRAM
- Metodologia
- Resultados
- **Comentários Finais**

Comentários Finais

- Implementação e validação de um controle adaptativo de acesso à memória compartilhada:
  - Controlador de memória multi-clientes;
  - Árbitro adaptativo com prioridades dinâmicas:
    - Adaptação é baseada nas características de acesso.
- Principais características:
  - Comportamento predizível no tempo:
    - Esquema de análise do pior caso, granularidade e preempção.
  - Avaliação do funcionamento em tempo de execução;
  - Escalabilidade das interfaces;
  - Suporte à clientes com acesso heterogêneos.

Comentários Finais

- Metodologia de projeto de SoCs que explora aspectos funcionais da memória:
  - Permite prever características de funcionamento do subsistema de memória:
    - Largura de banda;
    - Tempos de resposta;
    - Potência dissipada (não avaliada nesse trabalho).
- Modelo de comportamento do sistema:
  - Possibilita prever piores casos de execução:
    - Análise dos tempos de resposta.
  - Permite implementar garantias de prazo e largura de banda:
    - Desejável para clientes do tipo tempo-real.

Comentários Finais

- Limitações do sistema:
  - Implementação em hardware:
    - Frequência máxima de operação limitada pelo número de clientes;
    - Tempo de adaptação: Compromisso entre área e frequência.
  - Verificação de clientes “ativos”: melhorias podem ser inseridas para verificar clientes ativos
    - Reduz estimativa de pior caso do tempo de resposta.

Lista de Publicações

- BONATTO, A.C.; PEREIRA, F., BORIN, A., NEGREIROS, M., SUSIN, A.A. **Adaptive shared memory control for multimedia systems-on-chip.** SBCCI 2014
- BONATTO, A.C.; SUSIN, A.A. **Run-Time SoC Memory Subsystem Mapping of Heterogeneous Clients.** ISCAS 2014
- BONATTO, A.; SUSIN, A. **Memory Subsystem Architecture Design for Multimedia Applications**; ISVLSI 2013 (PhD Student Forum).
- BONATTO, A. et al. **Towards an Efficient Memory Architecture for Video Decoding Systems**. SBESC 2012
- BONATTO, A.; SOARES, A.; SUSIN, **A. Multichannel SDRAM controller design for H.264/AVC video decoder**. SPL 2011.
- BONATTO, A.; SOARES, A.; SUSIN, A. **High Efficiency Reference Frames Storage for H.264/AVC Decoder Hardware Implementation.** In: LASCAS 2010
- BONATTO, A. C. et al. **A 720p H.264/AVC decoder ASIC implementation for digital television set-top boxes**. SBCCI 2010.

Lista de Publicações

- SOARES, A. B.; BONATTO, A. C.; SUSIN, A. A. Development of a SoC for Digital Television Set-Top Box: architecture and system integration issues. **International Journal of Reconfigurable Computing,** 2013.
- NEGREIROS, M. et al. **Towards a video processing architecture for SBTVD.** SPL 2012.
- SOARES, A. B.; BONATTO, A. C. a.; SUSIN, A. A. **Integration issues on the development of an h.264/AVC video decoder SoC for SBTVD set top box**. SBCCI 2011.

Controle Adaptativo para Acesso à Memória Compartilhada em Sistemas em Chip

Referências

- AKESSON, B.; GOOSSENS, K. **Architectures and modeling of predictable memory controllers for improved system integration.** DATE 2011.
- AKESSON, B.; GOOSSENS, K.; RINGHOFER, M. **Predator: a predictable sdram memory controller**. In: HARDWARE/SOFTWARE CODESIGN AND SYSTEM SYNTHESIS (CODES+ISSS), 2007 5TH IEEE/ACM/IFIP INTERNATIONAL CONFERENCE ON. Anais. [S.l.: s.n.], 2007. p.251–256.
- ESMAEILZADEH, H. et al. **Power challenges may end the multicore era**. Commun. ACM, New York, NY, USA, v.56, n.2, p.93–102, Feb. 2013.
- HENRIKSSON, T et al.. **Network Calculus Applied to Verification of Memory Access Performance in SoCs**. ESTIMedia 2007.
- NEWMAN, L. H. Piz Daint Supercomputer Shows the Way Ahead on Efficiency. IEEE Spectrum, [S.l.], Jan. 2014.
- SEICULESCU, C. et al. **A DRAM Centric NoC Architecture and Topology Design Approach.** ISVLSI 2011.
- SHAH, H.; KNOLL, A.; AKESSON, B. **Bounding SDRAM interference: detailed analysis vs. latency-rate analysis.** DATE 2013.