Índices: Hash

Daniel de Oliveira Capanema

Adaptado Prof. Kutova

Tabela de dispersão

- As tabelas de dispersão (*hash*) em disco também podem ser usadas como índices, ao invés das árvores.
- Nessas tabelas, o curso de acesso é O(1).
- A posição do registro é determinada por uma função de dispersão (ou função *hash*).

Função de dispersão

- h(chave) → endereço

$$h(3204) = 504$$

chave — endereço

- Depende do número de endereços e da natureza da chave.
- Registros do índice devem ser de tamanho fixo.
- Quantidade fixa de endereços
  (depende do tratamento de colisões)

Tabela de dispersão

h(1) = 2
h(2) = 8
h(3) = 6
h(4) = 5
h(5) = 0

	Código	Endereço
0	5	432
1		
2	1	0
3		
4		
5	4	324
6	3	216
7		
8	2	108
9		

	Código	Título	Autor	Preço
0	1	Java Web Services	Martin Kalin	34,87
108	2	Web Design	Maurício Samy	45,50
216	3	Web Services em PHP	Lorna Jane	33,90
324	4	Programação Java para	Décio	93,22
432	5	Desenvolvimento Web	Qian et al	118,90

Exemplos de função de dispersão

- Elevar a chave ao quadrado e pegar um grupo de dígitos do meio:

  $A = h(453) \rightarrow 453^2 = 205209 \rightarrow A = 52$

  (dois dígitos foram escolhidos pois o arquivo possui apenas 100 endereços)

Exemplos de função de dispersão

- Mudar a chave para outra base:

  $A = h(453) \rightarrow 453_{10} = 382_{11} \rightarrow$

  $382 \bmod 99 = 85 \rightarrow A = 85$

(99 é a quantidade de endereços no arquivo)

Exemplos de função de dispersão

- Multiplicar o valor ASCII das letras e usar o resto da divisão pelo número de endereços

Chave	Cálculo	Endereço
JOÃO	74 x 79 = 5846	846
CARLOS	67 x 65 = 4355	355
GILBERTO	71 x 73 = 5183	183

Colisões

DESEJÁVEL

PIOR CASO

ACEITÁVEL

Tratamento de colisões

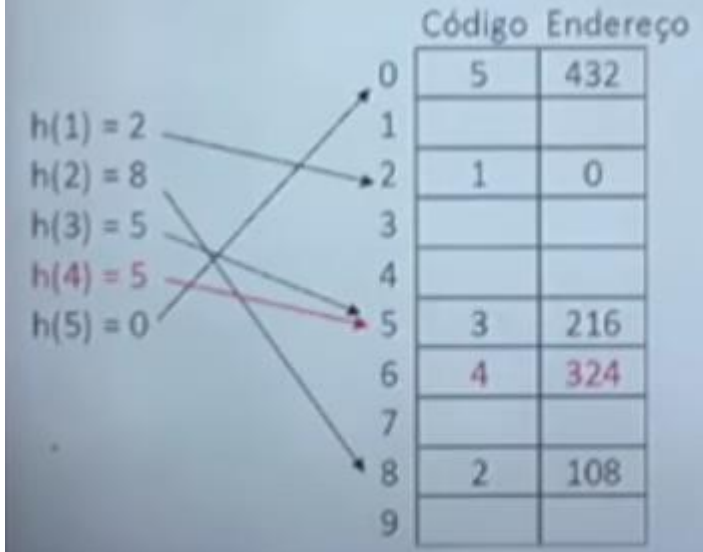
- Alternativas:
  - **Encadeamento interno** – usa outras posições vazias dentro da própria da tabela *hash*
  - **Encadeamento externo** – usa uma área extra, além da tabela *hash*, como uma área de extensão, ou um segundo arquivo.

Encadeamento interno

- Endereçamento aberto - uma nova posição **dentro da área da tabela** será procurada
  - Sondagem linear
  - Sondagem quadrática
  - Duplo *hash* (*double hashing*)

Encadeamento interno

- Sondagem linear – as próximas posições são sondadas (circularmente), até que uma posição livre seja encontrada.



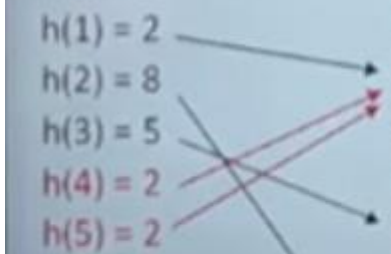
	Código	Endereço
0	5	432
1		
2	1	0
3		
4		
5	3	216
6	4	324
7		
8	2	108
9		

	Código	Título	Autor	Preço
0	1	Java Web Services	Martin Kalin	34,87
108	2	Web Design Responsivo	Maurício Samy Silva	45,50
216	3	Web Services em PHP	Lorna Jane Mitchell	33,90
324	4	Programação Java para a Web	Décio Heinzelmann	93,22
432	5	Desenvolvimento Web Java	Qian et al	118,90

Regra: $h(k,i) = [h(k) + i] mod\ n$

Encadeamento interno

- Sondagem quadrática – a distância até a próxima posição a ser sondada é determinada pelo quadrado da tentativa



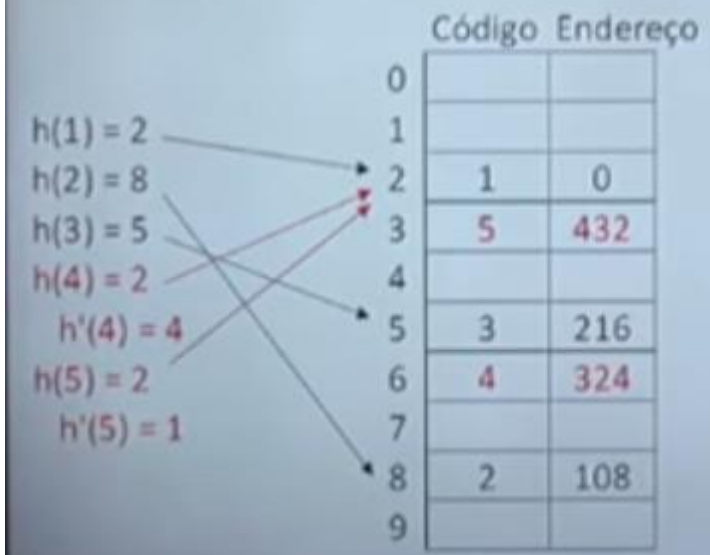
	Código	Endereço
0		
1		
2	1	0
3	4	324
4		
5	3	216
6	5	432
7		
8	2	108
9		

	Código	Título	Autor	Preço
0	1	Java Web Services	Martin Kalin	34,87
108	2	Web Design Responsivo	Maurício Samy Silva	45,50
216	3	Web Services em PHP	Lorna Jane Mitchell	33,90
324	4	Programação Java para a Web	Décio Heinzelmann	93,22
432	5	Desenvolvimento Web Java	Qian et al	118,90

Regra: $h(k,i) = [h(k) + i^2] \bmod n$

Encadeamento interno

- Duplo *hash* – a distância até a próxima posição a ser sondada é determinada por uma segunda função *hash*



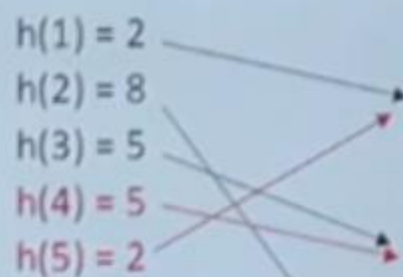
	Código	Endereço
0		
1		
2	1	0
3	5	432
4		
5	3	216
6	4	324
7		
8	2	108
9		

	Código	Título	Autor	Preço
0	1	Java Web Services	Martin Kalin	34,87
108	2	Web Design Responsivo	Maurício Samy Silva	45,50
216	3	Web Services em PHP	Lorna Jane Mitchell	33,90
324	4	Programação Java para a Web	Décio Heinzelmann	93,22
432	5	Desenvolvimento Web Java	Qian et al	118,90

Regra: $h(k,i) = [h(k) + i * h'(k)] \, mod \, n$

Encadeamento externo

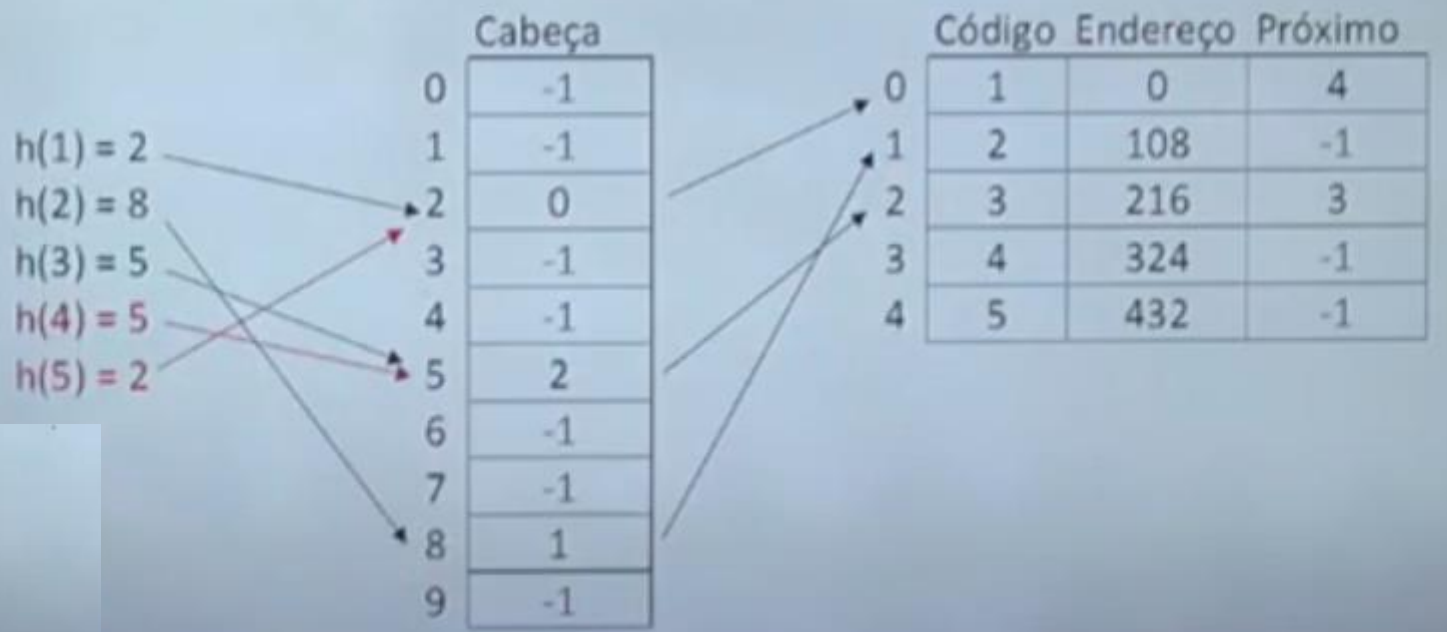
- Área de extensão – os registros colididos são armazenados em uma área de extensão



	Código	Endereço	Próximo
0			-1
1			-1
2	1	0	11
3			-1
4			-1
5	3	216	10
6			-1
7			-1
8	2	108	-1
9			-1
10	4	324	-1
11	5	432	-1

Encadeamento externo

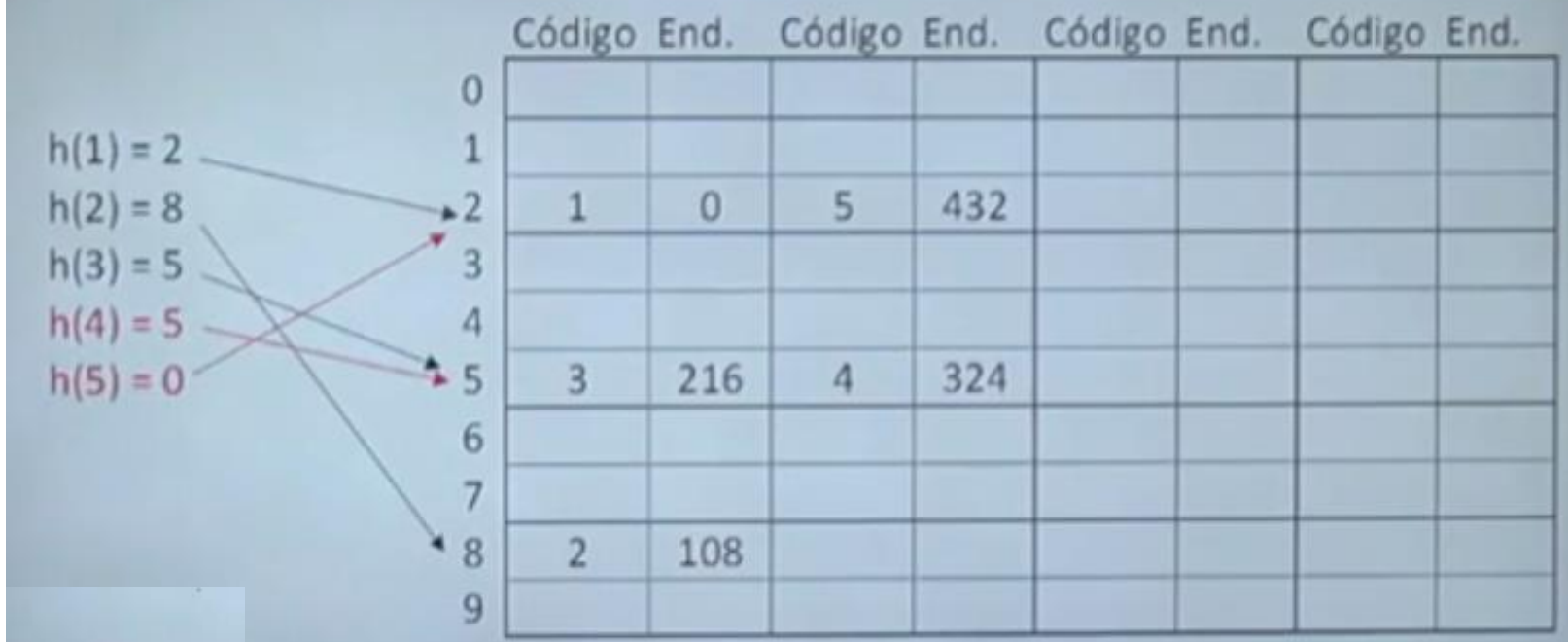
- Lista encadeada – todos os registros são armazenados em uma lista encadeada (outro arquivo)



Buckets (cesto)

- Da mesma forma que no caso da árvore B, é importante otimizar o acesso ao disco.
- Assim, cada *posição* no índice, pode conter mais de uma entrada (ou registro)
- Exemplo:
  - Registro no índice = 12 bytes
  - Setor do HD = 4096 bytes = 341,33 registros

Buckets (cesto)



	Código	End.	Código	End.	Código	End.	Código	End.
0								
1								
2	1	0	5	432				
3								
4								
5	3	216	4	324				
6								
7								
8	2	108						
9								

Buckets (cesto)

- Tratamento de colisões (quando o cesto está cheio)
  - Alocar o registro no próximo cesto em que houver espaço disponível (usando endereçamento aberto)
  - Usar uma das técnicas anteriores (considerando que cada posição equivale a um novo cesto)

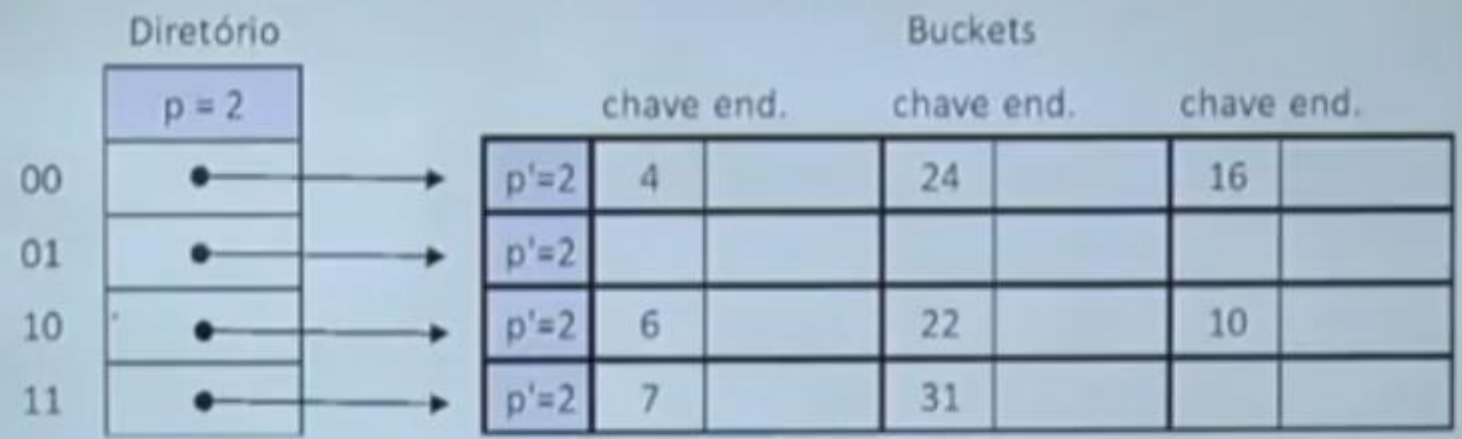
Tabela *hash* dinâmica

- Quando o arquivo de dados cresce ou diminui com frequência (muitas inclusões e exclusões), o índice também precisará ser ajustado.
- Uma tabela *hash* estática, para crescer, precisa reposicionar todos os registros.

Tabela *hash* dinâmica

- Uma tabela *hash* dinâmica é uma tabela *hash* em que apenas alguns registros afetados (aqueles do *bucket*) precisam ser reposicionados.

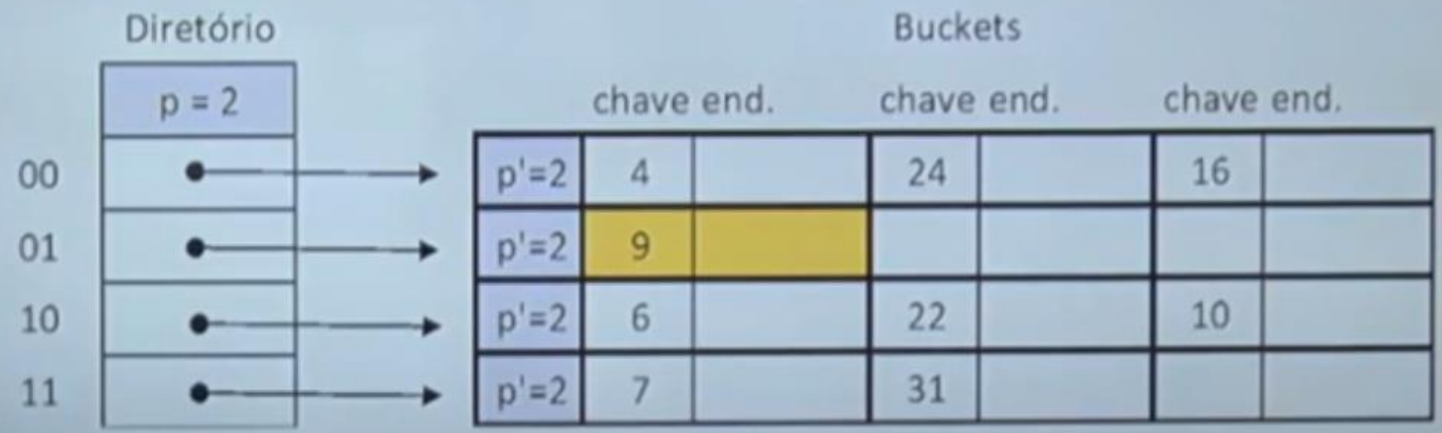
Hash extensível



$$h(k) = k \bmod 2^p$$

Hash extensível

Adicionar chave 9:



$h(k) = k \bmod 2^p$

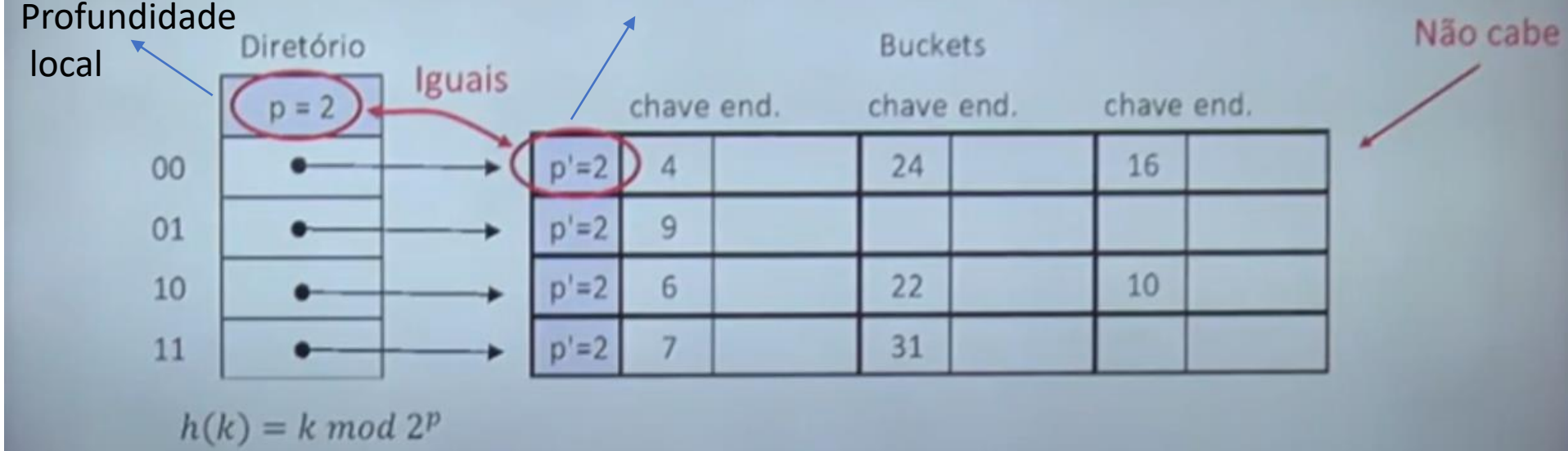
Hash extensível

Adicionar chave 20:

Profundidade local

Profundidade local

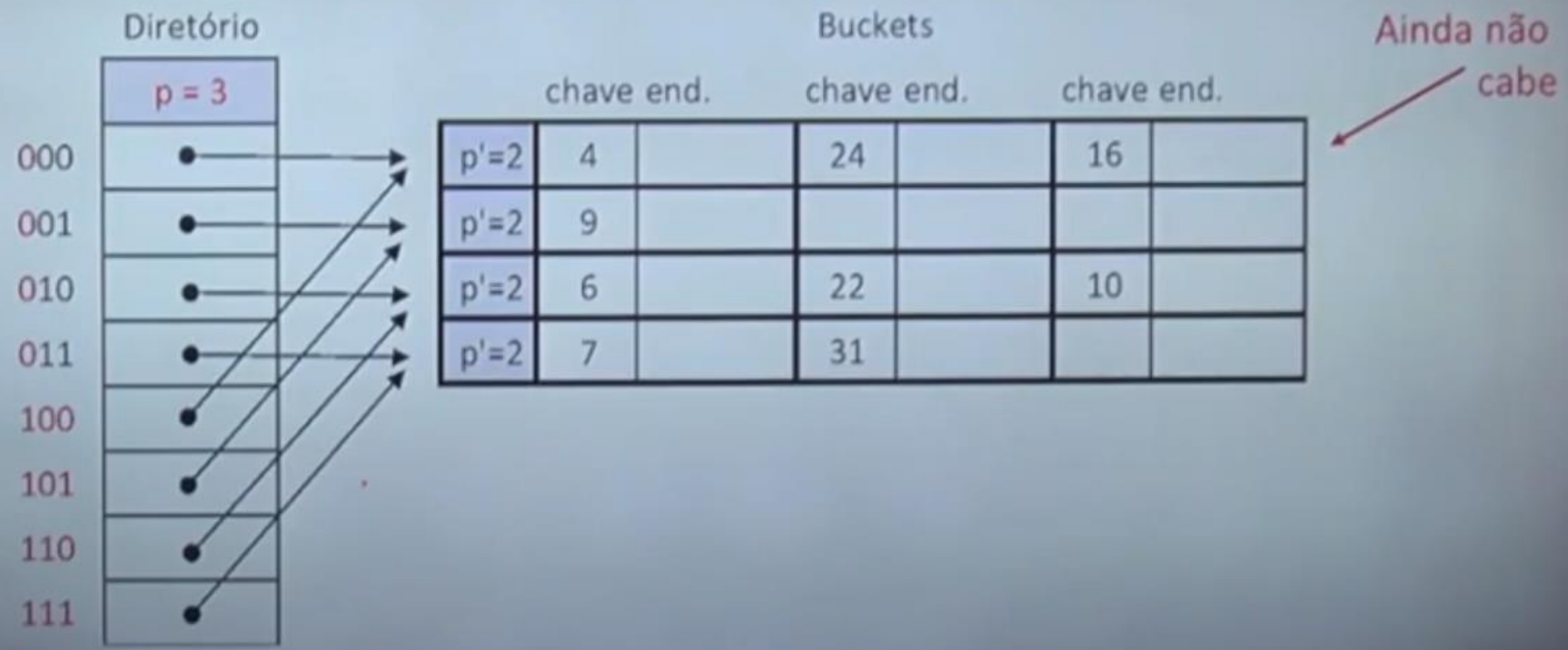
Não cabe



$h(k) = k \bmod 2^p$

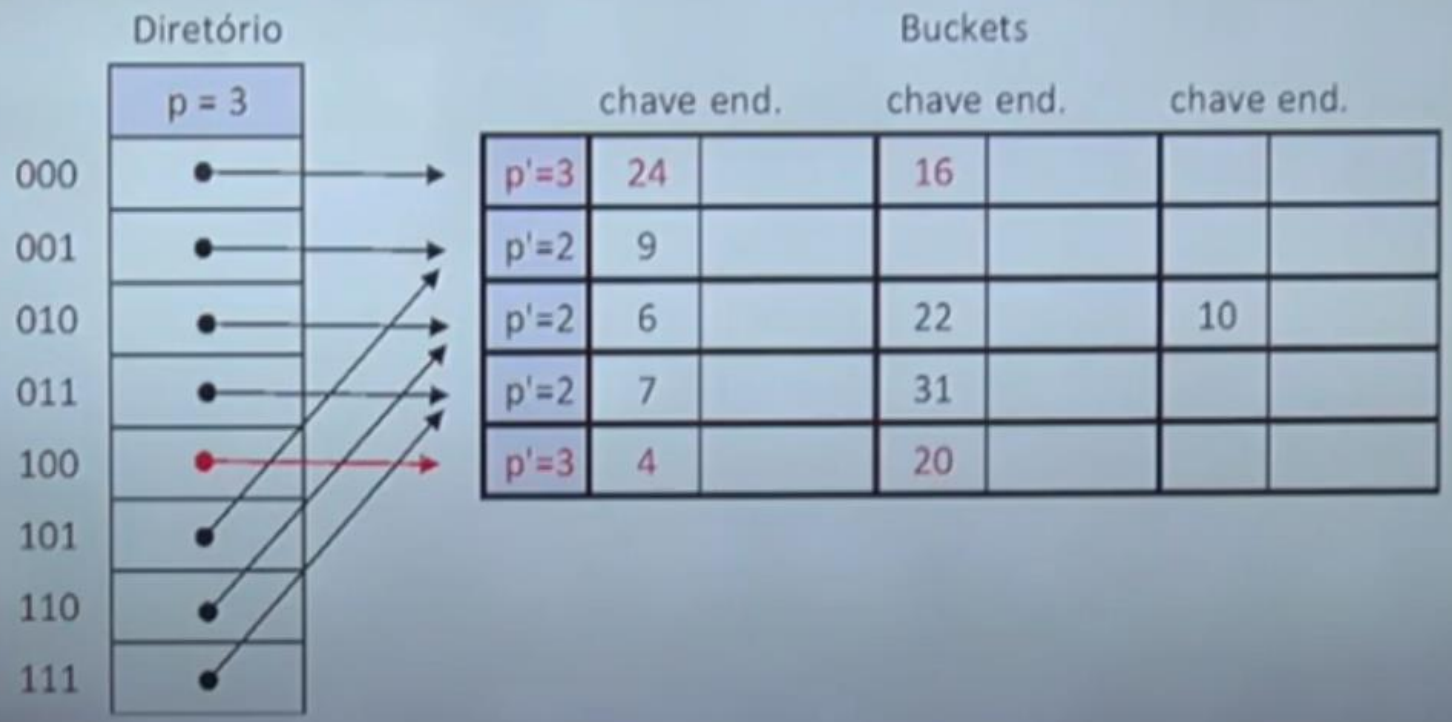
Hash extensível

Adicionar chave 20:



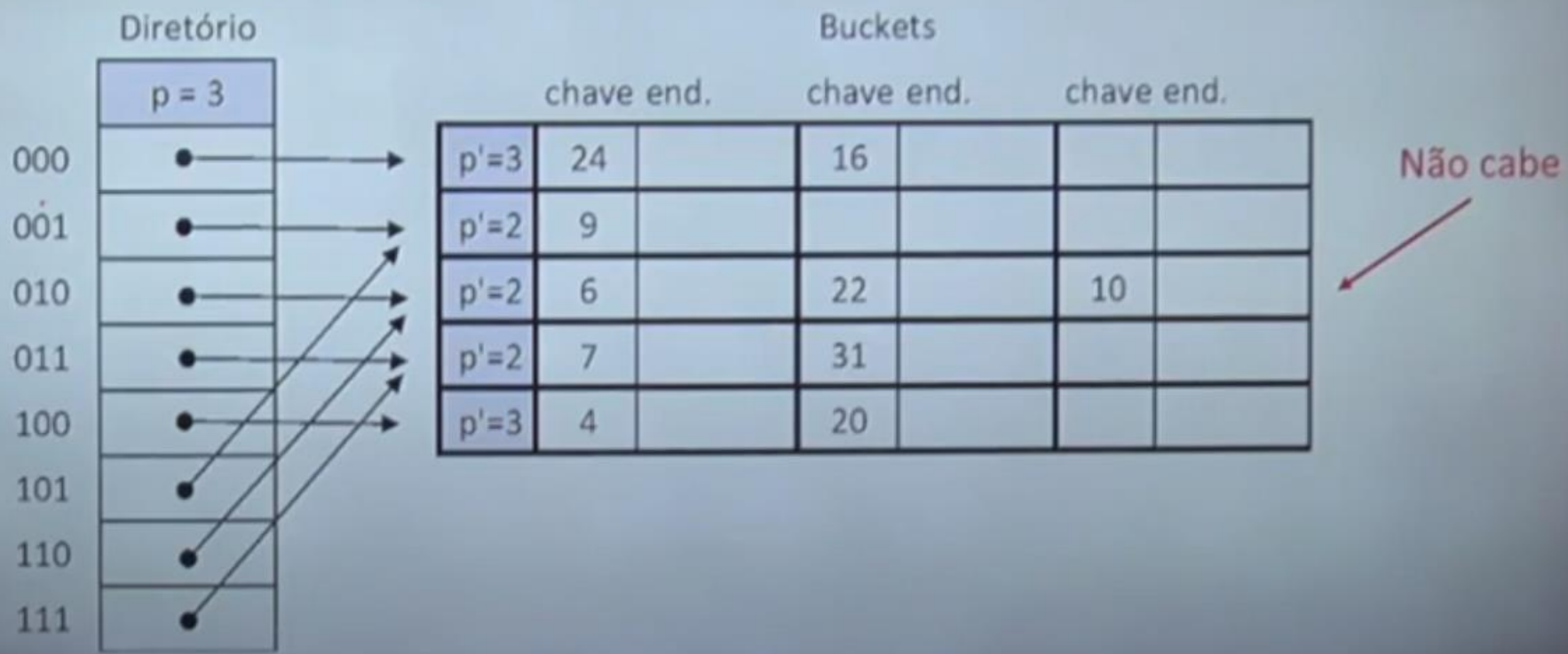
Hash extensível

Adicionar chave 20:



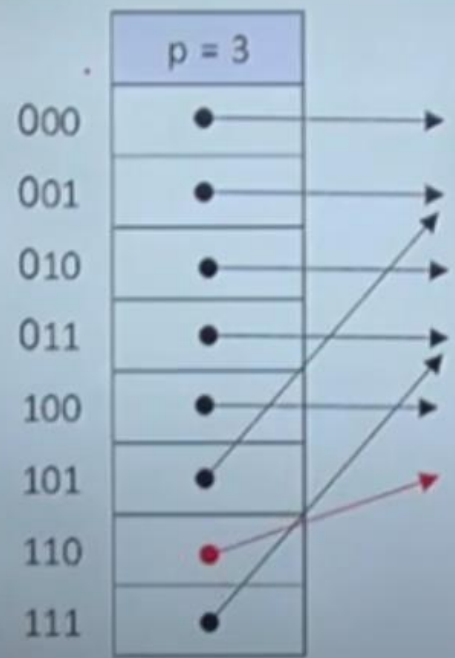
Hash extensível

Adicionar chave 26:



Hash extensível

Adicionar chave 26:



	chave end.		chave end.		chave end.	
p'=3	24		16			
p'=2	9					
p'=3	10		26			
p'=2	7		31			
p'=3	4		20			
p'=3	6		22			

Vantagens do *hash* extensível

- O diretório cresce, sem precisarmos reposicionar todos os registros (do índice)
- O índice (lista de *buckets*) cresce de acordo com a necessidade
- Como não há encadeamento dos *buckets*, não há perda de eficiência

Índices invertidos

- Um índice invertido é um índice em que uma parte do conteúdo de um registro (como uma palavra de um campo) é usada na localização do próprio registro.
- Essa é uma solução para permitir, entre outras, a busca de texto em arquivos, como fazem as máquinas de busca na Web.

Índices invertidos

Buscar Dados?

Cód.	Título	...
1	Implementação de sistemas de bancos de dados	...
2	Sistemas de bancos de dados	...
3	Estruturas de dados e seus algoritmos	...
4	Dominando algoritmos	...
5	Estruturas de dados em Java	...
6	Core Java	...
7	Biblioteca do programador Java	...

Busca sequencial ?

- A busca sequencial (testando cada registro) é lenta demais para um grande volume de dados
  - Google: pesquisa em bilhões de páginas em uma fração de segundo

Índice invertido

- Todos os "termos" são identificados e, para cada um deles, criamos uma lista dos registros em que aparecem

- Índice invertido = listas invertidas

Índice invertido

Cód.	Título	...
1	Implementação de sistemas de bancos de dados	...
2	Sistemas de bancos de dados	...
3	Estruturas de dados e seus algoritmos	...
4	Dominando algoritmos	...
5	Estruturas de dados em Java	...
6	Core Java	...
7	Biblioteca do programador Java	...

Termos
implementação
de
sistemas
bancos
dados
estruturas
e
seus
algoritmos
dominando
em
java
core
biblioteca
do
programador

Índice invertido

Cód.	Título	...
1	Implementação de sistemas de bancos de dados	...
2	Sistemas de bancos de dados	...
3	Estruturas de dados e seus algoritmos	...
4	Dominando algoritmos	...
5	Estruturas de dados em Java	...
6	Core Java	...
7	Biblioteca do programador Java	...

Termos	Registros
algoritmos	3 4
bancos	1 2
biblioteca	7
core	6
dados	1 2 3 5
dominando	4
estruturas	3 5
implementação	1
java	5 6 7
programador	7
seus	3
sistemas	1 2

Exemplo de consulta

Termos	Registros
algoritmos	3 4
bancos	1 2
biblioteca	7
core	6
dados	1 2 3 5
dominando	4
estruturas	3 5
implementação	1
java	5 6 7
programador	7
seus	3
sistemas	1 2

- Consulta: "estruturas de dados"

Faço buscas por termos;
Comparo em pares;
Encontro a resposta pela interseção

Vantagens

- Os índices invertidos (ou listas invertidas) podem ser construídos para qualquer conjunto de informações dos registros.
- Esses índices são estruturas adequadas para consultas combinadas (vários campos).