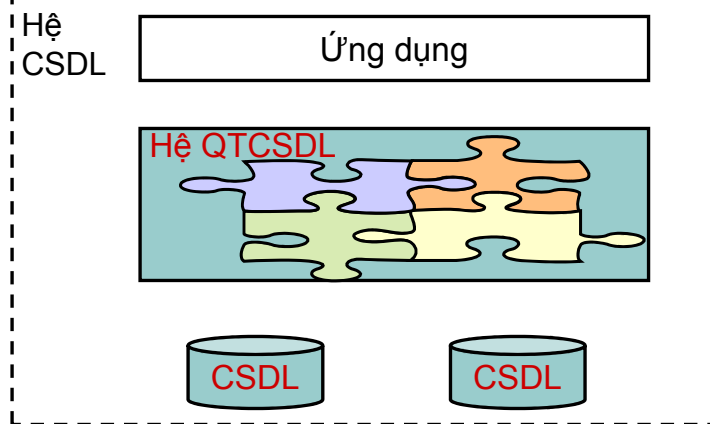


Tổ chức dữ liệu vật lý

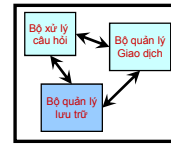
Vũ Tuyết Trinh

trinhvt@it-hut.edu.vn

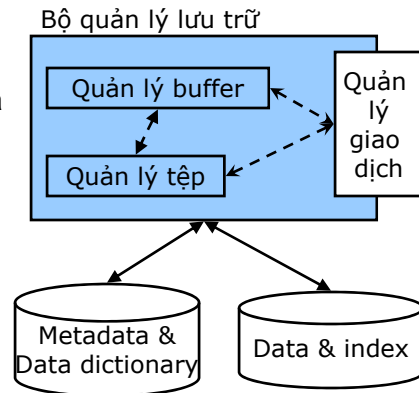
Bộ môn Các hệ thống thông tin, Khoa Công nghệ thông tin
Đại học Bách Khoa Hà Nội



Quản lý lưu trữ



- Tổ chức tệp: sắp xếp các bản ghi trên thiết bị nhớ ngoài
 - RID (*record id*): xác định địa chỉ vật lý của các bản ghi
 - chỉ số: cấu trúc dữ liệu xác định sự tương ứng giữa RID của bản ghi và giá trị của trường (khóa)
- Vùng nhớ đệm: trung gian giữa thiết bị nhớ ngoài và bộ nhớ trong (có thể sử dụng cho cả DL và chỉ số)



Tổ chức bộ nhớ ngoài

- Mục đích: giảm thiểu truy xuất đến dữ liệu không cần thiết trên thiết bị nhớ ngoài
- Các vấn đề cần quan tâm
 - Cấu trúc lưu trữ
 - Các phép toán (thêm, xoá, sửa, tìm kiếm)



Các thiết bị nhớ ngoài

- Đĩa từ, băng từ, trống từ, ...
- Đĩa từ: được tổ chức thành từng trang
 - Chi phí truy nhập đến các trang bất kỳ là tương đương
 - Chi phí đọc nhiều trang liên nhau < chi phí đọc các trang đó theo thứ tự bất kỳ
- Băng từ:
 - chỉ có thể đọc được các trang liên nhau
 - rẻ hơn đĩa từ nhưng chi phí truy nhập thường lớn hơn
- ...



Đĩa từ vs. bộ nhớ trong

- Tốc độ truy nhập bộ
ms vs. ns (~1000 lần)
- Kích thước
GB vs. 10x MB (~ 100 lần với cùng chi phí)
- Lưu trữ
ổn định (kể cả khi mất điện) vs. tạm thời
- Phân chia block
4KB vs. 1Byte

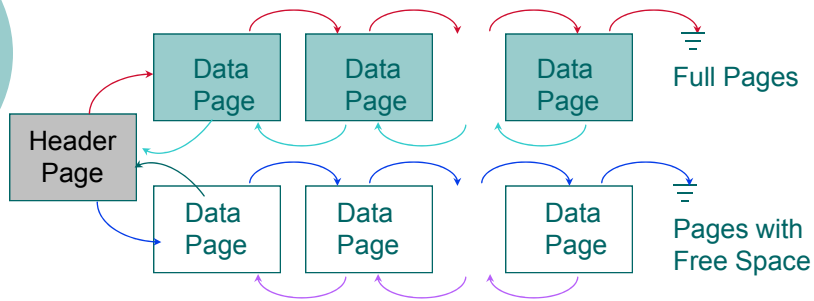
Nội dung

- ✓ Tổng quan về tổ chức bộ nhớ ngoài
- Tổ chức tệp đồng
- Tổ chức tệp băm
- Tổ chức tệp chỉ dẫn
- Cây cân bằng

Tổ chức tệp đồng (*Heap File*)

- Lưu trữ kế tiếp các bản ghi trong các trang không tuân theo một thứ tự đặc biệt nào
- Để thực hiện các phép toán, cần:
 - Ghi nhớ số trang trong 1 tệp
 - Ghi nhớ không gian trống trên các trang
 - Ghi nhớ các bản ghi trên các trang
- Có các con trỏ trỏ tới tất cả các trang của tệp và các con trỏ này được lưu trữ ở bộ nhớ trong.

Cài đặt tệp đồng bằng danh sách

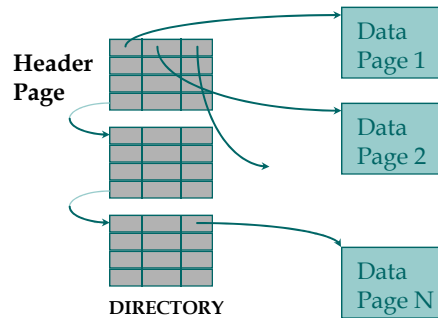


- Cần lưu trữ HeaderPage và tên của tệp
- Mỗi trang gồm dữ liệu và 2 con trỏ

Các phép toán

- Tìm kiếm 1 bản ghi
- Thêm 1 bản ghi
- Xoá 1 bản ghi
- Sửa đổi một bản ghi

Sử dụng trang danh bạ



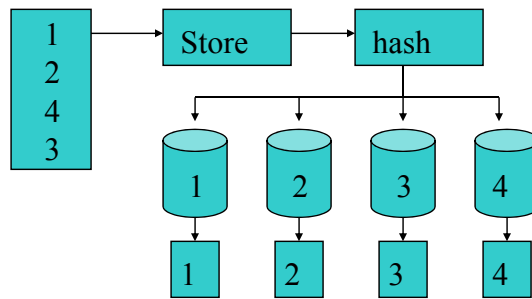
- Lưu thông tin về số byte còn trống trên trang đó
- Danh bạ là 1 tập các trang

Tổ chức tệp băm (*Hash File*)

- Mục đích
 - Sử dụng chỉ số để hạn chế số lượng phép truy xuất đĩa bằng các phân nhóm các bản ghi (giả thiết n nhóm)
 - *Mapping* giá trị khoá với vị trí của (nhóm) bản ghi tương ứng
- Dựa trên bảng băm (*hash table*)
 - Hàm băm (*hash function*)
 - Cụm (*bucket*)

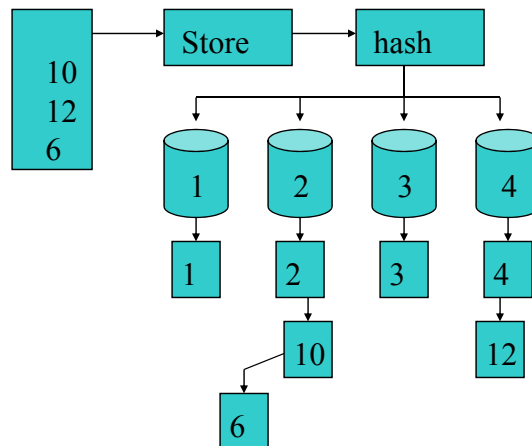
Ví dụ

$$h(x) = x \bmod 4$$



Ví dụ tiếp

$$h(x) = x \bmod 4$$





Các phép toán

- Tìm kiếm 1 bản ghi
- Thêm 1 bản ghi
- Xoá 1 bản ghi
- Sửa đổi một bản ghi

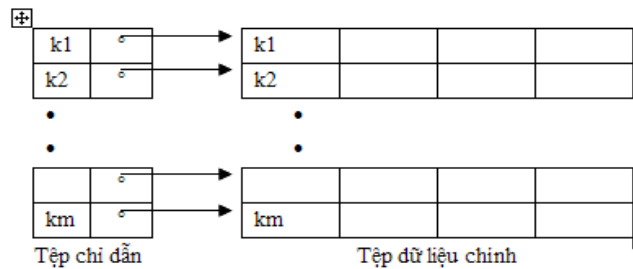


Tiêu chí chọn hàm băm

- Phân bố các bản ghi tương đối đồng đều (theo các cụm)
- Hạn chế việc sử dụng nhiều trang bộ nhớ cho 1 cụm

Tổ chức tệp chỉ dẫn (*Index File*)

- Tệp chỉ dẫn theo khoá được chọn trong bản ghi
- Tệp chỉ dẫn bao gồm các cặp (k,d), trong đó k là giá trị của khoá của bản ghi đầu tiên, d là địa chỉ của khối (hay con trỏ khối).
- Tệp chỉ dẫn được sắp xếp theo giá trị của khoá.



Các phép toán

- Tìm kiếm 1 bản ghi
- Thêm 1 bản ghi
- Xoá 1 bản ghi
- Sửa đổi một bản ghi

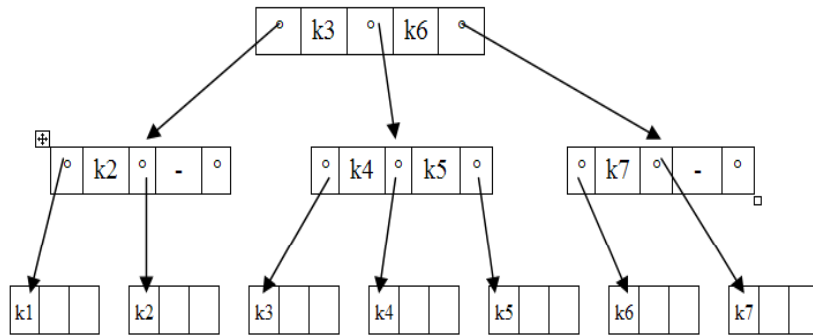
Tìm kiếm 1 bản ghi

- Tìm kiếm tuần tự
 - Duyệt tệp chỉ dẫn từ bản ghi đầu tiên đến khi tìm thấy bản ghi có khoá k cần tìm
 - Nhận xét
 - chậm đối với các tệp chỉ dẫn nói chung.
 - Thích hợp với các tệp chỉ dẫn nhỏ đủ để lưu ở bộ nhớ trong
- Tìm kiếm nhị phân
 - Chia đôi tệp chỉ dẫn đã sắp xếp để hạn chế số bản ghi cần duyệt
 - Tại mỗi lần chia hạn chế được $\frac{1}{2}$ số bản ghi cần xem xét

Cây cân bằng (*BalanceTree*)

- B-tree cân bằng được tổ chức theo cấp m, có các tính chất sau đây:
 - Gốc của cây hoặc là một nút lá hoặc ít nhất có hai con.
 - Mỗi nút (trừ nút gốc và nút lá) có từ $\lfloor m/2 \rfloor$ đến m con.
 - Mỗi đường đi từ nút gốc đến bất kỳ nút lá nào đều có độ dài như nhau.

Ví dụ



Nhận xét

- Cấu trúc của mỗi nút trong B-tree
($p_0, k_1, p_1, k_2, \dots, k_n, p_n$)
 - p_i ($i=1..n$) là con trỏ tới khối i của nút có k_i là khoá đầu tiên của khối đó.
 - Các khoá k trong một nút được sắp xếp theo thứ tự tăng dần.
- Mọi khoá trong cây con, trỏ bởi p_i đều nhỏ hơn k_{i+1}
- Mọi khoá trong cây con, trỏ bởi p_n đều lớn hơn k_n .



Các phép toán

- Tìm kiếm 1 bản ghi
- Thêm 1 bản ghi
- Xoá 1 bản ghi
- Sửa đổi một bản ghi



So sánh các cách tổ chức dữ liệu

- Tệp đồng
- Tệp băm
- Tệp chỉ dẫn
- Cây cân bằng

Kết luận

- Truy cập đến CSDL thường liên quan đến một phần nhỏ các bản ghi trong một tệp dữ liệu hay một vài trường (đặc biệt là các trường khoá) của các bản ghi dữ liệu.
 - Xác định các yêu cầu này cho phép thiết kế dữ liệu vật lý hiệu quả thông qua việc sử dụng các tổ chức lưu trữ đặc biệt
- Tệp chỉ dẫn được tạo lập trên khoá tìm kiếm để tăng hiệu quả của lưu trữ dữ liệu
 - Hiệu quả của các cấu trúc chỉ dẫn khác nhau phụ thuộc vào điều kiện áp dụng chúng

