PA152

Implementace databázových systémů

Shrnutí

trvalé uložení –> disk

Moore

- rychlost procesoru >> rychlost disku

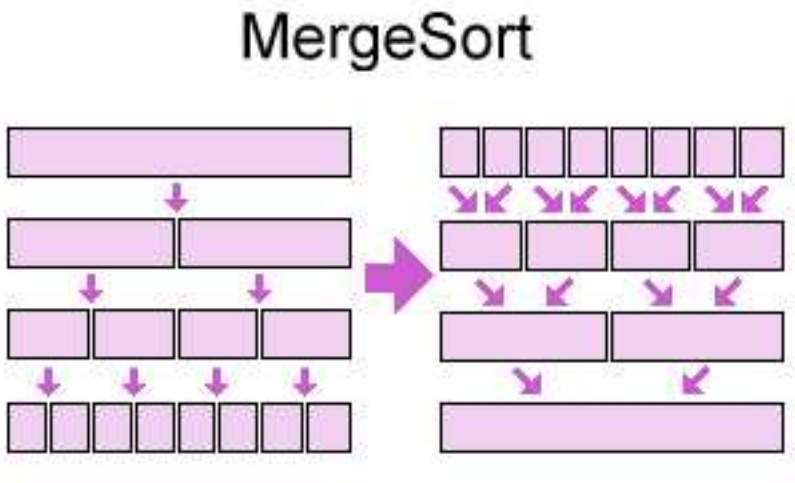
Optimalizace algoritmů

- minimalizace počtu náhodných
- přístupů k disku

Třídění

QuickSort (a jiné)
- předpokládají data v operační paměti

MergeSort
- třídění sléváním
  - rekurzivní dělení na menší posloupnosti
  - slévání setříděných postí



TPMMS

Two−Phase, Multiway Merge−Sort

Dvoufázové třídění vícecestným sléváním

- 1. fáze
  - třídění "bloků" v op. paměti
- 2. fáze
  - slévání všech proudů

Dvoufázové třídění – 1. fáze

- opakované
  - naplnění "celé" paměti daty
    - sekvenční čtení
  - třídění v paměti
  - zápis celého proudu na disk
    - sekvenční zápis
- 1 čtení + 1 zápis pro každý blok dat

Dvoufázové třídění – 2. fáze

- 1 blok paměti pro každý proud
- blok(y) paměti pro výstup
- načtení vstupních bufferů
- opakované
  - nalezení nejmenší hodnoty ze všech proudů
  - přesun do výstupního bloku
  - aktualizace čela použitého proudu
  - plný výst. buffer –> zápis
  - prázdný vstupní buffer –> čtení
- 1 čtení + 1 zápis pro každý blok dat

Dvoufázové třídění – analýza

- B = počet bloků s tříděnými daty
- 1. fáze: B * (1 čtení + 1 zápis)
- 2. fáze: B * (1 čtení + 1 zápis)
- celkem: max 4*B náhodných přístupů na disk

Složitost třídění

- O(n log n)
  - počet porovnání

Dvoufázové třídění–omezení

- parametry (v bajtech)
  - K = velikost bloku
  - M = velikost dostupné paměti
  - R = velikost záznamu
- max počet bufferů: M/K
- max počet proudů: M/K – 1
- počet záznamů v každém kroku 1. fáze: M/R
- max počet tříděných záznamů (M/R)(M/K – 1) ~ MM/RK
- max velikost tříděných dat: MM/K

Dvoufázové třídění–omezení

- max velikost tříděných dat: MM/K

- M (paměť) = 128MB = 2^27
- K (blok) = 128kB = 2^17
- MM/K = 2^(27*2 – 17) = 2^37 = 128GB

- pokud to nestačí
  - –> třífázové třídění

Třífázové třídění

- 1. fáze
  - stejně jako u dvoufázové třídění
- 2. fáze
  - třídíme pouze tolik proudů, kolik se vejde do paměti
  - každý běh vytvoří nový větší proud
- 3. fáze
  - třídíme všechny "velké" proudy z 2. fáze
- DÚ: Jaké jsou limity třífázového třídění?

Optim. přístupu na disk

Omezení náhodného přístupu

- umístění bloků čtených často po sobě na stejný cylindr
- zvětšení velikosti bloku
- prefetching, double buffering

Rozložení dat na více disků

- snížení průměrné doby vystavení hlaviček disku

Optim. přístupu na disk

Zrcadlení disků

- pro čtení je možné použít kterýkoliv disk

Plánování přístupu

- algoritmus výtahu
- asynchronní přístup

Výpadky disků

- Občasný výpadek
  - chyba při čtení/zápisu, opakování OK
- Vada média
  - trvalá hodnota nějakého bitu
- Chyba zápisu
  - chyba při zápisu i následném čtení
- Zničení disku
  - totální výpadek

Ošetření výpadků disků

- kontrolní součty
- samoopravné kódy

- stabilní uložení
  - zrcadlení nebo uložení na dvou místech stejného disku

Zotavení ze zničení

- neexistuje absolutně jistá strategie
- pouze minimalizace rizika

- MTTF – mean time to failure
  - průměrná doba funkčnosti
  - za jak dlouhou dobu u 50% disků daného typu dojde k totální mu výpadku
  - ~ 10 let
- použití redundance –> zvýšení MTTF celku

MTTF –– použití

- průměrná doba funkčnosti N let
- 50 % disků –– výpadek za N let
- 100 % disků –– výpadek za 2N let
- pravděpodobnost výpadku za rok 1/(2N)

- systém složený z více disků
  - MTTF závisí na
    - struktuře systému
    - MTTF jednotlivých částí
    - rychlosti výměny vadné části

RAID

Redundant Arrays of Idependent Disks

- diskové pole
  - RAID level 1
    - zrcadlení
  - RAID level 4
    - paritní disk
  - RAID level 5
    - paritní bloky
  - RAID level 6
    - opravné kódy

RAID level 1

- zrcadlení disku
  - poloviční kapacita
  - zrychlené čtení

- výpočet MTTF
  - 2 stejné disky, MTTF 3 roky
  - výměna vadného – 3,5 dne
  - výpadek – oba disky během 3,5 dne
    - p(výpadku disku za rok) = 1/6
    - p(výp. d. za 3,5 dne) = 1/6 * 3,5/365 = 1/600
    - p(výp. celku za rok) = 2 * 1/6 * 1/600 = 1/1800
    - MTTF = 900 let

RAID level 4

- datové disky + paritní disk
- kontrolní součet
  - součet modulo 2 na každém bitu
- čtení
  - datový disk
- zápis
  - čtení datový + paritní
  - zápis datový + paritní
  - velká zátěž paritního disku
- MMTF
    - p(výp. 2 disků) * (kombinace disků)
    - 3+1 disk –> MTTF = 300 let

RAID level 4 – obnova dat

- zvládne výpadek kteréhokoliv disku
- čtení během obnovy datového
  - čtení všech disků
  - výpočet hodnoty
- zápis
  - ohled na postup obnovy dat

RAID level 4 – příklad

- disk1: 11110000
- disk2: 10101010
- disk3: 00111000
- disk4: 01100010

RAID level 5

- paritní bloky
  - každý disk je paritní jen pro část bloků
- odstranění zátěže partiního disku

- čtení/zápis/obnova
  - stejné jako RAID level 4

- MMTF
  - stejně jako RAID level 4
  - v praxi delší

RAID level 6

- více redundantních disků
- samoopravné kódy
  - Hammingův kód
- zvládá výpadek více disků

- čtení
  - datový disk
- zápis
  - čtení datový + všechny redundantní
  - zápis datový + všechny redundantní

RAID level 6 – příklad

	datové				redundantní		
disk	1	2	3	4	5	6	7
	1	1	1	0	1	0	0
	1	1	0	1	0	1	0
	1	0	1	1	0	0	1

disk	obsah
1	11110000
2	10101010
3	00111000
4	01000001
5	01100010
6	00011011
7	10001001