Helsingin yliopisto, Tietojenkäsittelytieteen laitos
Tietokannan hallinta, kurssikoe 16.5.2003

Kirjoita jokaiseen erilliseen vastauspaperiin kurssin nimi, tentin päiväys, oma nimesi (selvästi), syntymäaikasi ja nimikirjoituksesi

Anna kunkin tehtävän (1, 2, 3) vastaus erillisellä paperilla.

1. Tarkastellaan levyasemaa, jossa on 20 levypintaa, Levy jakautuu 1000 sylinteriin. Uran keskimääräinen kapasiteetti on 50 lohkoa a' 4KB. Hakuvarren keskimääräinen kohdistusaika on 4 ms. Levyn pyörimisnopeus on 12000 kierrosta minuutissa, eli yhteen kierrokseen menee aikaa 5 ms. Olkoon taulussa 200000 riviä. Rivin keskipituus on 400 tavua. Taulu toteutetaan avaimeen perustuvana hajautusrakenteena ilman soluhakemistoa. Perussolun koko on yksi lohko. Alkuperäinen hajautusalue on kooltaan 40 000 lohkoa. Ylivuotosivuille on varattu tilaa 10000 lohkoa, Hajautusfunktio jakaa tietueet melko tasaisesti ja ylivuotosivuja on käytössä vain 200.
   a) Mistä tekijöistä muodostuu satunnaisen tietueen hakuaika avaimen perusteella haettaessa ja mikä tämä aika on tässä tapauksessa?

Hajautuksessa lasketaan hakuavaimen perusteella soluosoite ja haetaan tietue solusta. Solusta haettaessa on luettava ainakin kotisolu. Lisäksi voidaan joutua lukemaan ylivuotosivuja.
Sivun satunnaissaantiaika = kohdistusaika+pyörähdysviive+siirtoaika.
Kohdistuaaika= 4ms.
Pyörähdysviive= 0.5*5ms= 2.5 ms.
Siirtoaika= 5ms/50 =0.1 ms
Sivun satunnaissaantiaika = 4+2.5+0.1=6.6 ms.
Keskimääräinen haettaessa olemassa olevaa riviä= 6.6 ms + 200/40200*6.6 ms
= 6.6 + 0.03ms eli noin 6.6 (tietenkin ylivuotoketju jonkin sivun tapauksessa voi olla pidempi kuin 1 , mutta todennäköisyys osua ylivuotosivulle on pieni 0.03 ja hajautuksen on kerrottu olevan tasaista)

   b) Miten käsittely kannattaa tehdä ja mikä on käsittelyaika, jos tiedostosta haetaan tietuejoukko jonkin muun hakutiedon kuin avaimen perusteella.

Luetaan peräkkäin kaikki hajautusalueen 40000 lohkoa ja ylivuotoalueen 200 lohkoa. Voidaan siis hyödyntää peräkkäisyyttä. Oletetaan että sijoittelu on optimaalista.

Kokonaishakuajan alaraja: Hajautusalue: kohdistusaika + pyörähdysviive +siirtoaika +
Siirtymiset sylinteriltä seuraavaan (lasketaan tässä keskimääräisinä) +
Ylivuotoalue : . kohdistusaika + pyörähdysviive +siirtoaika =
4 ms + 2.5 ms + 40000*0.1 ms + (40000/ 1000)*(4+2.5) +
4ms+2.5ms + 200*0.1 ms =
6.5+4000+260 + 6.5+20 = 4293 eli noin 4.3 s

   (6p = 3+3)

2. Tarkastellaan tauluja R(A, B,…) ja S(X, …, Z ->R). (Z on viiteavain tauluun R). Yhteen R-riviin kytkeytyy keskimäärin 20 S-riviä. Taulun R koko on noin 500 riviä (tiedostossa 200 datasivua a´4KB). Taulussa S on 10000 riviä (tiedostossa 5000 datasivua). Sarakkeen A

arvojen keskipituus on 30 tavua. Sivukoko on 4KB ja osoite esitetään 10 tavulla. Taulut on toteutettu kasarakenteina. Puskuritilaa on käytettävissä 150 sivua. Tarkastellaan kyselyä

```
select * from R,S where R.A=S.Z and S.Z=x;
```

a) Mikä on tuloksen oletettu koko riveinä?

Keskimäärin 20 riviä. Tämä voidaan päätellä joko suoraan tehtävänannosta ("Yhteen R-riviin kytkeytyy keskimäärin 20 S-riviä" ja kyselyn valintaehto rajaa käytännössä R-taulusta valittavaksi vain yhden rivin) tai laatimalla kyselypuun ja optimoimalla sitä.

b) Miten toteuttaisit kyselyn, jos käytössä ei ole indeksejä (kuvaa periaate)? Paljonko levyhakuja tällöin tarvitaan?

1. Luetaan S-sivut muistiin (5000 hakua) ja poimitaan samalla eroon n. 20 ehdon S.Z = x täyttävää monikkoa.

2. Jaksottaista sisäkkäisten silmukoiden menetelmää käyttäen pienempi taulu 20 riviä on jo muistissa. Luetaan R-sivuja muistiin ja verrataan niiden sisältämiä monikoita
muistissa oleviin kahteenkymmeneen S-monikkoon. Ainoa vastaava R-monikko tulee vastaan keskimäärin puolessavälissä, sadan levyhaun jälkeen (max. 200:n). (järjestelmä tietää että A on R:n avain)

3. Muodostetaan lopputulos. Tämä vaati keskimäärin 5100 levyhakua, max 5200.

c) Miten taulujen avaimiin ja viiteavaimiin perustuvat indeksit vaikuttavat kyselyn toteutukseen? Mikä on tällöin tehokkain toteutussuunnitelma ja mitä luokkaa olisi levyhakujen määrä?

Jos molemmille tauluille on olemassa sopivat ISAM-hakemistot (B+-puilla lopputulos osapuilleen sama):

Osoitetietue vie tilaa 10t+30t=40t.
Yhdelle sivulle menee 4000t/40t=100 osoitetietuetta.

S-taulun ISAM:
- Pohjataso 10000 sivua / 100 = 100 sivua
- Juuritaso 100 sivua / 100 = 1 sivu

R-taulun ISAM:

- Pohjataso 500 sivua / 100 = 5 sivua
- Juuritaso 5 sivua / 100 < 1 sivu

Haku:

1. Haetaan S:n juuri puskuriin (1), haetaan sen viittaamat pohjatason sivut (1-2) ja niiden osoittamat datatason sivut (max. 20), joilta löytyvät kaivatut S:n monikot.

2. Haetaan R:n juuri puskuriin (1), haetaan sen viittaama pohjatason sivu (1) ja sen osoittama datasivu (1), jolta löytyy kaivattu R:n monikko.

3. Muodostetaan lopputulos. Tämä vaati maksimissaan 26 hakua.

(10p jaettu 2+4+4)

3. a) Mikä on tarkistuspiste (checkpoint)?
   b) Mitä tietokantaloki sisältää?
   c) Mitä tietokantalokille tapahtuu sitoutumisen yhteydessä? (9p)

---

4. **Kotikoetehtävä.** ***Tähän tehtävään vastaavat vain ne, jotka haluavat korvata laskuharjoituspisteensä.*** Vastaus on lähetettävä sähköpostitse osoitteeseen *Harri.Laine@cs.helsinki.fi* viimeistään ***perjantaina 23.5. klo 18.00***. Vastaus voi olla upotettuna sähköpostiviestiin tai postin liitetiedosto. Se voi olla myös linkki tiedostoon. Tiedostomuotoina käyvät ascii teksti, MSWord:n doc muoto, Postscript, PDF tai HTML.

   Selvitä miten tietokannanhallintajärjestelmissä on toteutettu suurten sarakearvojen (clob, blob) ja vaihtuvapituisten (varchar) sarakearvojen talletus. Käsittele vähintään kolmea yleisesti tunnettua järjestelmää. Selvityksen laajuus noin 3 sivua. (5p)