NTNU
Det skapende universitet

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i
TDT4225 Lagring og behandling av store datamengder
Kontinuasjonseksamen

Faglig kontakt under eksamen: Kjell Bratbergsengen

Tlf.: 906 17 185 / 7359 3439

Eksamensdato: Fredag 15. august 2014

Eksamenstid (fra-til): 0900 - 1300

Hjelpemiddelkode/Tillatte hjelpemidler: *D*
Ingen trykte eller håndskrevne hjelpemiddel er tillatt.
Bestemt, enkel kalkulator tillatt

Annen informasjon:

Sensur innen: mandag 1. september 2014.

Målform/språk: Bokmål

Antall sider: 3

Antall sider vedlegg: 0

Kontrollert av:

Sven Erik Bratsberg

_____

Dato Sign

TDT4225 Lagring og behandling av store datamengder
Fredag 15. august 2014, kl. 0900-1300

Oppgave 1, diskkontrollere, magnetisk disk og flashdisk (SSD) (10 %)

a) Beskriv oppgavene til en kontroller for magnetiske disker.

b) Beskriv oppgavene til en kontroller for flashminnebasert "disk".

Oppgave 2, RAID 1, speilte disker (20 %)

Blokk nr	Disk 0
0	
1	
2	
3	
4	
5	

Blokk nr	Disk 1
0	
1	
2	
3	
4	
5	

Figuren viser et RAID 1-system hvor de to diskene skal inneholde identiske data. Diskene er av samme type. Hver disk har en kapasitet på 2 TB, antall rotasjoner per sekund (ω) er 250 og hvert spor lagrer 1000 KB.

a) Hvor stor er hver disks maksimale lese- og skrivehastighet?

For spørsmålene b, c og d gjelder: Systemet er satt opp til å bruke en blokkstørrelse på 64 KB, og adressene er vilkårlige, dvs. "randomisert" lesing og skriving.

b) Hvor mange blokker kan systemet *lese* per sekund?

c) Hvor mange blokker kan systemet *skrive* per sekund?

d) Hvor mange blokker kan systemet *oppdatere* per sekund?

e) Hvis en av diskene må skiftes ut, hvor lang tid tar det å overføre dataene til den nye disken? Anta at hele disken må skrives og at annen trafikk er blokkert i perioden.

Oppgave 3, lagringsstrukturer (15 %)

a) Beskriv hvordan en R-trefil for to-dimensjonale objekter er organisert.

b) Beskriv i korte trekk hvordan en søker etter objekter som ligger helt eller delvis innenfor et gitt areal.

c) Et nytt objekt skal settes inn. Hvordan bestemmes hvilken datablokk som skal få objektet?

d) Hvis ingen relevant datablokk har plass for det nye objektet, hvordan skaffes ny plass?

Oppgave 4, sortering (15 %)

Du skal sortere en fil med 200 millioner poster, postlengde er100 byte, derav utgjør nøkkel 12 byte. Til rådighet har du 200 MB arbeidslager. CPU-hastigheten er slik at du kan regne med at gjennomsnittstiden for å sammenligne to poster er ett mikrosekund.

Alle filer (innfil, mellomlagerfiler og resultatfil) ligger på samme *flashdisk* (SSD) som har følgende parametre: Fast blokkstørrelse er 4 KB, lesing og skriving tar begge 0,1 ms per blokk.

a) Hvor lang tid trenger CPU for å gjøre initiell sortering?

b) Anta at IO og CPU arbeider parallelt.
Hvor lang tid tar hele sorteringen?

Oppgave 5, Bloomfilter (20 %, c-spørsmålet teller halvparten)

a) Forklar hvordan Bloomfilter virker.

b) Du skal lagre 10 millioner poster med unike nøkler. Hvor mye plass vil du da avsette til Bloomfilteret? Forklar hvordan du kommer fram til svaret.

c) Et kredittkortselskap har utstedt ca. 300 millioner kredittkort. Ca. 5 % av kortene er blokkert pga. mislighold, tyveri, og lignende. For hver transaksjon må en kontrollere om kortet er gyldig. Kredittkortets nummer er på 16 siffer. Skisser et system (datastrukturer) som effektivt kan brukes til å utføre gyldighetskontrollen. Beregn hvor stor plass (arbeidslager og sekundærlager) løsningen din krever og hvor mange kort den kan kontrollere per sekund.

Oppgave 6, Foreningsalgoritmer (20 %)

Tabelldata:	Resultat	Basetabeller	
	R	**A**	**B**
Nøkkellengde i byte	8	8	8
Postlengde i byte	200	600	300
Antall poster	250 000	300 000	1 200 000

Finn den beste algoritmen for å gjøre operasjonen **R=A* B** (likhetsforening av tabellene **A** og **B**). Resultatpostene i **R** henter 100 byte fra hver av operandene. Tilgjengelig arbeidslager WS er 10 MB. Beregn samlet transportvolum for de algoritmene du prøver.

Kjb/4.8.2014