

25

Sortering III.

Basale forudsætninger.

Sortering ved fletning
med tre bånd, i to faser.

Sortering ved fletning,
med fire bånd, i én fase (balanceret fletning).

Polyfase fletning med tre bånd.

Generaliseret polyfase fletning.

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

I de to forrige forelæsninger om sortering har vi set på *intern sortering*, altså sortering som udelukkende foregår ved databehandling i det primære lager af maskinen.

Når store datamængder på eksterne lagre, typisk magnetbånd, skal sorteres, må vi ty til andre sorteringsteknikker. Vi taler i sådanne tilfælde om *ekstern sortering*.

Med dagens teknologi, først og fremmest arbejdslagre på mange megabytes (og virtuelle arbejdslagre på endnu flere megabytes), er ekstern sortering ikke helt så vigtig, som det før har været tilfældet. På trods af dette vil vi i denne forelæsning give et overblik over eksterne sorteringsteknikker. Dette hører med til en sømmelig datalogisk opdragelse, og der er mange spændende teknikker involveret, som er et nærmere bekendtskab værdig.

Basale forudsætninger.

Data læses og skrives strengt sekventielt.

Operationer i en abstrakt datatype SEKVENS:

- Read_next_object: Object
- Write_next_object (obj: Object)
- Rewind

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

I de to tidligere forelæsninger om sortering har vi antaget, at sorteringen forløb på instanser af en abstrakt datatype LIST. Vi har også antaget, at der er et forholdsvis righoldigt repertoire af effektive operationer i LIST, f.eks. aflæsning og skrivning af i-te element i listen. Som omtalt er LIST typisk baseret på arrays, når vi taler intern sortering.

I denne forelæsning vil vi på en meget radikal måde indsnævre operationerne på lister. I stedet for lister vil vi tale om SEKVENS. Som nævnt ovenfor vil vi kun forudsætte tre forskellige operationer i denne datatype: Læs objekt, skriv næste objekt, og positioner sekvensen i starttilstand. Vi har tidligere i forelæsningen om "filer og indekser" stiftet bekendtskab en sådan datatype.

Som en yderligere begrænsning vil vi antage, at skrivning umuliggør efterfølgende læsning, medmindre sekvensen først er repositioneret til starttilstand.

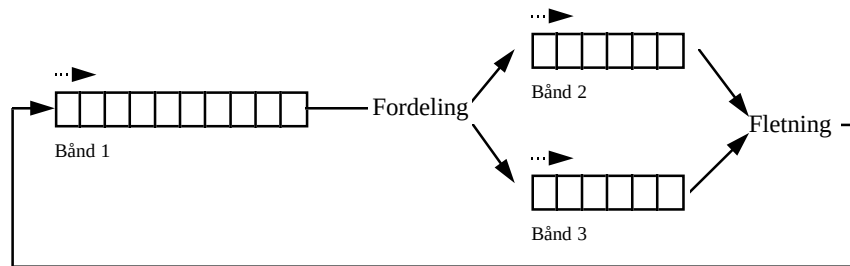
Vi har tidligere i noterne vedtaget at tale om poster på en fil, i stedet for objekter eller records. I dette kapitel vil vi dog tillade os tale om objekter repræsenteret på en fil.

Sortering ved fletning med tre bånd i to faser (1).

Givet: $S = (O_1 O_2 \dots O_n)$ på bånd. Endvidere givet to andre bånd.

Metode:

1. Fordel objekterne ligeligt på bånd 2 og bånd 3.
2. **Flet** objekterne fra bånd 2 og 3 til sorterede delsekvenser af dobbelt længde på bånd 1.
3. Gentag ovenstående to trin indtil bånd 1 er sorteret.



PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

Vi vil her og i det følgende tale om **bånd**, hvorpå vi antager at sekvenser af data er repræsenterede. Man kan naturligvis tænke på bånd som **magnetbånd**. For hvert bånd der opereres på er vi selvfølgelig nødt til at have en **magnetbåndstation**. I praksis er antallet af stationer den begrænsede ressource.

Den her skitserede sorteringsteknik bør sammenlignes med sortering ved fletning, som vi studerede sidste gang (se forelæsningen "sortering 2"). Merge-sort er en top-down realisering af ideen om sortering ved fletning, som kan beskrives elegant ved brug af rekursion. Merge-sort er dog ikke, som formuleret i forrige forelæsning, rettet mod en praktisk situation, hvor vi har et fast antal magnetbåndstationer til rådighed.

Formuleringen af sortering ved fletning, som skitseret på denne slide, er en iterativ, praktisk anlagt (rettet mod sortering på ydre, sekventielle medier), bottom-up formulering af sorteringsidéen.

Sortering ved fletning med tre bånd i to faser (2).

Sortering af n objekter.

Antal fordelingsfaser og fletningsfaser: $\log_2 n$

Antal læse+skriveoperationer: $2n \log_2 n$

Svagheder:

Små sorterede sekvenser i begyndelsen af processen.

For mange unyttige kopieringsoperationer.

Eksempel:

Sortering af $n = 2^{17} = 131.072$ objekter

$\log_2 n = 17$ faser.

$2n \log_2 n = 4.456.448$ læse+skriveoperationer.

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

I regnestykket på denne side betegner en læse+skriveoperation én læseoperation efterfulgt af én skrive operation.

Den næste eksterne sorteringsmetode, vi vil studere, har løsninger på de to svagheder, vi har identificeret på denne slide. Bemærk dog, at vi går fra 3 til minimum 4 bånd(stationer).

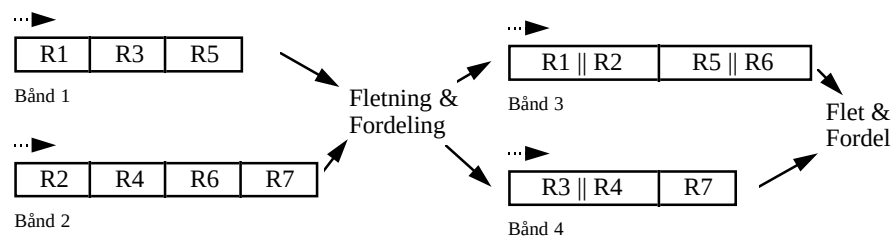
Det er imidlertid under visse omstændigheder muligt at optimisere sorteringen, som vist på forrige slide, uden at kræve et ekstra bånd. Dette vender vi tilbage til senere i denne forelæsning under den såkaldte polyfasesortering.

Sortering ved fletning med fire bånd i én fase (1).

Givet: $S = (O_1 \ O_2 \ \dots \ O_n)$ på et bånd, samt tre andre bånd.

Metode:

1. Etabler sorterede "runs" (ved intern sortering), og fordel disse runs ligeligt på bånd 1 og 2.
2. **Flet** to runs fra bånd 1 og 2, og fordel skiftevis disse på bånd 3 og 4.
3. Gentag trin 2, med alternerende rollefordeling (kilde/mål) af båndene.



PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

Sorteringsteknikken på denne slide er kendt som "balanceret tovejs fletning".

Ved at benytte intern sortering af delsekvenser fra bånd opnår vi, at den eksterne sorteringsproces starter med sorterede "runs" af en betydelig længde. Dette er klart en stor forbedring i forhold til den første metode, vi studerede, idet vi antager, at intern sortering er meget mere effektiv end ekstern sortering.

Vi antager at alle runs er ca. lige lange i begyndelsen. Dette er ikke en nødvendig antagelse. Vi antager også at ca. halvdelen af runs initialt placeres på bånd 1, og den anden halvdel på bånd 2.

Som vi vil se under polyfase sortering, er den ligelige fordeling af runs ikke altid den bedste antagelse.

Ved at benytte fire bånd kan vi slå flette og fordelingsprocessen sammen. Når vi fletter to runs skriver vi skiftevis det resulterende run ud på bånd 3 og 4. Efter hver gennemført flet-og-fordel proces har vi runs af den dobbelte længde i forhold til den forrige situation (modulo de runs, som ikke kan flettes, fordi de ikke har en "partner").

I stedet for at tale om konkrete bånd, kunne vi tale om et par af kilde-bånd og et par af mål-bånd. Efter en endt flet-og-fordel proces skifter kilde- og mål-bånd roller.

Vi har ikke på denne slide gjort rede for, hvornår bånd skal "rewindes". Det er dog helt oplagt, hvornår dette skal ske.

Den eksterne sorteringsteknik kan generaliseres til anvendelse af flere end 4 bånd (og båndstationer). Antag vi har $X + Y$ bånd og -stationer ($X \geq 2, Y \geq 2$), hvor vi i begyndelsen har sekvenser af sorterede runs på X kilde-bånd. Disse flettes og fordeles nu til de Y målbånd, og processen fortsætter. Hvis vi antager at $X = Y$ skal vi gennemløbe $\log X(R)$ faser, hvor R er antallet af initiale runs, og hvor $\log X$ er X -tals logaritmen.

Sortering ved fletning med fire bånd i én fase (2).

Sortering af n objekter, med r initiale runs.

Antal splitfaser og fletfaser: $\log_2 r$

Antal læse/skrive-operationer: $n \log_2 r$

Eksempel:

Sortering af $n = 2^{17} = 131.072$ objekter

Intern sortering, som resulterer i 512 initiale runs á 256 objekter.

$\log_2 512 = 9$ faser

$9n = 1.179.648$ læse+skrive operationer.

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

I regnestykket på denne slide har vi *ikke* medtaget bidraget fra intern sortering, samt læse/skrive operationerne omkring denne.

Polyfasesortering med tre bånd (1).

Givet: $S = (O_1 \ O_2 \ \dots \ O_n)$ på bånd 1 samt to andre bånd.

Metode:

1. Etabler r (*et Fibonacci tal*) sorterede "runs" (ved intern sortering), og fordel disse runs i Fibonacci forhold på to kilde-bånd.
2. **Flet** par af runs fra de to kilde-bånd, og udskriv dem på mål-båndet.
Forsæt med fletningen indtil ét af kilde-båndene ikke indeholder flere runs.
3. Mål-båndet *spoles tilbage*, og bliver kildebånd i næste fase.
Det kilde-bånd, som blev tømt, *spoles tilbage* og bliver mål-bånd i næste fase.
Det kilde-bånd, som ikke blev tømt, forbliver kildebånd i næste fase.
4. Iterér punkterne 2 og 3 indtil mål-båndet indeholder ét run med alle objekter, som nu er sorteret.

De to kildebånd bliver aldrig tømt samtidigt under polyfasesortering.

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

Når r er Fibonaccitallet F_i , og $i \geq 2$, da kan vi opskrive r som $F_{i-1} + F_{i-2}$. $F_1 = 1$, og $F_0 = 0$. Det omtalte Fibonacci forhold i trin 1 af algoritmen er forholdet F_{i-1}/F_{i-2} .

I trin 2 er vi interesserede i, at kun ét bånd løber tør for runs i hver fase. Dette sikres netop af, at vi arbejder med fordelinger efter Fibonacci tal.

Bemærk, at vi ved polyfase sortering bliver nødt til at være omhyggelig med at fortælle, hvornår bånd skal spoles tilbage.

Ligesom i den forrige beskrevne sorteringsteknik benytter polyfase fletningen på denne slide initialt sorterede runs. I litteraturen bliver der brugt en del kræfter på at "vride" eksisterende, interne sorteringsalgoritmer, således at de rent faktisk kan anvendes til at sortere flere data, end der kan rummes i arbejdslageret. Dette er beskrevet i afsnit 7.11.6 i Weiss, og vi er motiverede for at se lidt på dette, idet det giver os endnu en anvendelse af minimum-hobe.

Næste slide viser skematisk, hvordan polyfase sortering forløber i termer af antal runs, og længden af runs.

Polyfasesortering med tre bånd (2).

i, i, i

3 runs,
af relativ længde i.

Fase	Bånd 1	Bånd 2	Bånd 3
1.	1, 1, 1, 1, 1, 1, 1, 1	1, 1, 1, 1, 1	
2.	1, 1, 1		2, 2, 2, 2, 2
3.		3, 3, 3	2, 2
4.	5, 5	3	
5.	5		8
6.		13	

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

Denne slide illustrerer, hvordan polyfase sortering forløber i tilfældet, hvor de to kilde-bånd hhv. indeholder 5 og 8 runs.

Der er ingen begrænsning på længden af de initiale runs. Disse kan indeholde f.eks. 1 objekt, 1000 objekter, eller 7346 objekter hver. Vi kan endog benytte runs med 0 objekter. Det kan være nyttigt at inddrage et antal runs af længde 0, hvis ikke det initiale antal runs er et Fibonacci tal. Hvis vi således starter med R "lange" runs (som stammer fra intern sortering) vil vi inddrage S tomme runs ($S \geq 0$) således at $S + R$ er et Fibonacci tal.

Det er vigtigt at observere, at ikke alle objekter bliver behandlet i hver af faserne. Pointen i teknikken er jo netop, at i hver fase er der en rest af runs, som kommer til at indgå i næste fase.

Polyfasesortering med tre bånd (3).

Sortering af n objekter, med r initiale runs.

Antal splitfaser og fletfaser: $2.078 \ln r + 0.67$ (asymptotisk værdi).

Antal læse/skrive-operationer: $n * (1.504 \ln r + 0.992)$ (asymptotisk værdi).

Eksempel:

Sortering af $n = 131.072$ objekter.

Intern sortering, som resulterer i $r = 512$ initiale runs á 256 objekter.

$2.078 \ln r + 0.672 = 14$ faser

$n * (1.504 \ln r + 0.992) = 1.359.798$ læse+skrive operationer.

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

Formlerne på denne slide stammer fra Knuth. Vi udleder ikke disse formler på dette kursus.

Formlen for antallet af faser giver et korrekt result, anvendt på $r = 512$, idet det 13. Fibonacci tal er 377, og det 14. er 610.

Hvis vi sammenligner det totale antal læse+skriveoperationer med de tidligere resultater kan vi se, at det står sig pænt i sammenligning med den allerførste, naive teknik, som vi studerede.

I forhold til sortering med 4 bånd, kan vi næsten gøre det lige så godt med 3-bånds polyfasesortering.

Generaliseret polyfasesortering.

Polyfase sortering med 3 bånd kan umiddelbart generaliseres til $T > 3$ bånd:

- I hver fase har vi $T-1$ kildebånd og ét målbånd.
- Den nødvendige initiale fordeling kan beskrives af *generaliserede Fibonacci tal*.

Eksempler på initiale fordelinger:

$T = 7$, 6 bånd flettes	Sum	$T = 6$, 5 bånd flettes.	Sum
1 0 0 0 0 0	1	1 0 0 0 0	1
1 1 1 1 1 1	6	1 1 1 1 1	5
2 2 2 2 2 1	11	2 2 2 2 1	9
4 4 4 4 3 2	21	4 4 4 3 2	17
8 8 8 7 6 4	41	8 8 7 6 4	33
16 16 15 14 12 8	81	16 15 14 12 8	65
32 31 30 28 24 16	161	31 30 28 24 16	129
63 62 60 56 48 32	321		

PS1 -- Sortering III

© Kurt Nørmark, Aalborg Universitet, 1994.

Noter

I eksemplerne forneden angiver hver række af tal fordelinger af runs.

Fordelingerne er frembragt med det formål for øje, at kun netop ét bånd må blive tomt i hver fase.

Man kan relativt let indse, at én række fremkommer fra den forrige ved at flytte første tal hen som det sidste tal i næste række, samt ved at addere første tal til hvert af de efterfølgende, og forskyde resultatet én plads til venstre i næste række (prøv efter!)

De generaliserede Fibonacci tal af orden hhv. 5 og 4 forekommer i 1. søjle af tabellerne.

Fibonacci tal af orden n fremkommer ved at addere de $n+1$ tidligere tal for at få det næste. Følgen starter med ét 1-tal, og resten 0'er.

De kursiverede summer af rækker angiver totale antal af runs, som kan fordeles således at optimale polyfase fletteprocesser kan finde sted (med hhv. 7 bånd og 6 bånd).