

Algoritmi paralleli e distribuiti

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
 Parte I — Algoritmi paralleli	6
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
 Parte II — Algoritmi paralleli a memoria distribuita	37
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52

Parte III — Algoritmi distribuiti	54
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62
3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

Introduzione

Algoritmi 101

Un **algoritmo** è una sequenza finita di istruzioni che non sono ambigue e che terminano, ovvero restituiscono un risultato. Gli **algoritmi sequenziali** avevano un solo esecutore, mentre gli algoritmi di questo corso utilizzano un **pool di esecutori**.

Le problematiche da risolvere negli algoritmi sequenziali si ripropongono anche qua, ovvero:

- **progettazione**: utilizzo di tecniche per la risoluzione, come *Divide et Impera*, *programmazione dinamica* o *greedy*;
- **valutazione delle prestazioni**: complessità spaziale e temporale;
- **codifica**: implementare con opportuni linguaggi di programmazione i vari algoritmi presentati.

Un **algoritmo parallelo** è un algoritmo **sincrono** che risponde al motto «*una squadra in cui batte un solo cuore*», ovvero si hanno più entità che obbediscono ad un clock centrale, che va a coordinare tutto il sistema.

Abbiamo la possibilità di condividere le risorse in due modi:

- memoria, formando le architetture
 - **a memoria condivisa**, ovvero celle di memoria fisicamente condivisa;
 - **a memoria distribuita**, ovvero ogni entità salva parte dei risultati parziali sul proprio nodo;
- uso di opportuni collegamenti.

Qualche esempio di architettura parallela:

- **supercomputer**: cluster di processori con altissime prestazioni;
- **GPU**: usate in ambienti grafici, molto utili anche in ambito vettoriale.

Un **algoritmo distribuito** è un algoritmo **asincrono** che risponde al motto «*ogni membro del pool è un mondo a parte*», ovvero si hanno più entità che obbediscono al proprio clock personale. Abbiamo anche in questo caso dei collegamenti ma non dobbiamo supportare una memoria condivisa o qualche tipo di sincronizzazione, quindi dobbiamo utilizzare lo **scambio di messaggi**.

Qualche esempio di architettura distribuita:

- **reti di calcolatori**: rete internet;
- **reti di sensori**: sistemi con limitate capacità computazionali che rispondono a messaggi *ack*, *recover*, *wake up*, eccetera.

La risorsa tempo

Il **tempo** è una variabile fondamentale nell'analisi degli algoritmi. Definiamo prima la funzione

$$T(x) = \# \text{operazioni elementari sull'istanza } x.$$

Questo valore dipende fortemente dall'istanza x . Per risolvere questo problema, visto che noi vogliamo ragionare nel caso **worst case**, ovvero il caso che considera la situazione peggiore possibile, così da avere dei bound ragionevoli, definiamo ora il tempo come la funzione

$$t(n) = \max(\{T(x) \mid x \in \Sigma^n\}),$$

dove n è la grandezza dell'input. Abbiamo quindi raggruppato le istanze di grandezza n e abbiamo preso tra queste il tempo massimo. Spesso saremo interessati al *tasso di crescita* di $t(n)$, definito tramite **funzioni asintotiche**, e non ad una sua valutazione precisa.

Date due funzioni $f, g : \mathbb{N} \rightarrow \mathbb{N}$, le principali funzioni asintotiche sono:

- **O grande**: $f(n) = O(g(n)) \iff f(n) \leq c \cdot g(n) \quad \forall n \geq n_0$;
- **omega grande**: $f(n) = \Omega(g(n)) \iff f(n) \geq c \cdot g(n) \quad \forall n \geq n_0$;

- **theta grande:** $f(n) = \Theta(g(n)) \iff c_1 \cdot g(n) \leq f(n) \leq c_2 \cdot g(n) \quad \forall n \geq n_0$.

Il tempo $t(n)$ dipende da due fattori molto importanti: il **modello di calcolo** e il **criterio di costo**.

Un **modello di calcolo** mette a disposizione le **operazioni elementari** che usiamo per formulare i nostri algoritmi. Modelli di calcolo sono le macchine RAM, le MdT, le macchine WHILE, eccetera.

Esempio 0.1 (*Funzione palindroma*): La funzione *palindroma*, che dice se una stringa x di lunghezza n data in input è palindroma, ha tempo:

- $O(n^2)$ in una MdT;
- $O(n)$ in una architettura con memoria ad accesso casuale.

Anche le dimensioni dei dati in gioco contano: il **criterio di costo uniforme** afferma che le operazioni elementari richiedono una unità di tempo, mentre il **criterio di costo logaritmico** afferma che le operazioni elementari richiedono un costo che dipende dal numero di bit degli operandi, ovvero dalla sua dimensione.

Con tutte queste nozioni possiamo creare una serie di **classi di complessità**. Un problema è **risolto efficientemente** in tempo se e solo se è risolto da una MdT in tempo polinomiale. Abbiamo tre principali classi di equivalenza per gli algoritmi sequenziali:

- P , classe dei problemi di decisione risolti in tempo polinomiale su una MdT;
- FP , classe delle funzioni risolte in tempo polinomiale su una MdT;
- NP , classe dei problemi di decisione risolti in tempo polinomiale su una MdTnd.

Per quanto riguarda gli algoritmi paralleli, definiamo una nuova classe di complessità, ovvero NC , classe delle funzioni che ammettono algoritmi paralleli efficienti. Un problema appartiene alla classe NC se viene risolto in tempo *polilogaritmico* e in spazio *polinomiale*.

Teorema 0.1: Vale

$$NC \subseteq FP.$$

Dimostrazione 0.1: Per ottenere un algoritmo sequenziale da uno parallelo, faccio eseguire in sequenza ad una sola identità il lavoro delle entità che prima lavoravano in parallelo.

Visto che lo spazio di un problema NC è polinomiale, ovvero abbiamo un numero polinomiale di processori, abbiamo un «programma unico» con una grandezza polinomiale.

Ogni parte di questo «super programma» ha tempo polinomiale, il tempo totale è un polinomio per un polilogaritmo, quindi il tempo è polinomiale. ■

Come per il famosissimo problema $P = NP$, qui il dilemma aperto è se vale $NC = FP$, ovvero se posso parallelizzare ogni algoritmo sequenziale efficiente. Per ora sappiamo che $NC \subseteq FP$, e che i problemi che appartengono a FP ma non a NC sono detti P -completi.

Parte I – Algoritmi paralleli

1. Architetture parallele

Prima di vedere l'architettura parallela che ci accompagnerà durante tutta questa parte del corso, diamo una brevissima introduzione delle due architetture principali che possiamo utilizzare quando ci troviamo in un ambiente di calcolo parallelo.

1.1. Memoria condivisa

L'architettura a **memoria condivisa** utilizza una memoria centrale che permette lo scambio di informazioni tra un numero n di processori, ognuno dei quali possiede anche una «*memoria personale*», formata da registri. Un **clock** centrale e comune coordina tutti i processori, che comunicano attraverso la memoria centrale in tempo costante $O(1)$, permettendo quindi una forte parallelizzazione.

1.2. Memoria distribuita

L'architettura a **memoria distribuita** utilizza una rete di interconnessione centrale che permette lo scambio di informazioni tra un numero n di processori, ognuno dei quali possiede anche una «*memoria personale*», formata da registri. Un **clock** centrale e comune coordina tutti i processori, che comunicano attraverso la rete di interconnessione in un tempo che dipende dalla distanza tra i processori.

1.3. Modello PRAM

Vediamo finalmente l'architettura che ci accompagnerà durante tutta questa parte del corso.

1.3.1. Struttura

Il **modello PRAM** (*Parallel RAM*) utilizza una memoria M formata da registri $M[i]$ e una serie di processori P_i che si interfacciano con essa. Ogni processore P_i è una **RAM sequenziale**, ovvero contiene una unità di calcolo e una serie di registri $R[i]$.

La comunicazione avviene con la memoria centrale M tramite due primitive che lavorano in tempo costante. Esse sono **LOAD** e **STORE**, usate rispettivamente per caricare un dato dalla memoria e salvare un dato nella memoria.

Le operazioni di ogni processore avvengono invece in locale, cioè con i dati della propria memoria privata. Il tempo di ogni processore P_i è scandito da un clock centrale, che fa eseguire ad ogni processore la «*stessa istruzione*» $istruzione_i$.

Infatti, andiamo a definire il **passo parallelo** nel seguente modo

Passo parallelo

- 1: for all $i \in \mathbb{I}$ par do
 - 2: L $istruzione_k$
-

Il passo parallelo fa eseguire a tutti i processori con indice in \mathbb{I} la k -esima istruzione, altrimenti fa eseguire una **nop**, l'operazione nulla.

L'istruzione eseguita dipende dal tipo di architettura:

- **SIMD** (*Single Instruction Multiple Data*) indica l'esecuzione della stessa istruzione ma su dati diversi;
- **MIMD** (*Multiple Instruction Multiple Data*) indica l'esecuzione di istruzioni diverse sempre su dati diversi.

Abbiamo diverse architetture anche per quanto riguarda l'accesso alla memoria:

- **EREW** (*Exclusive Read Exclusive Write*) indica una memoria con lettura e scrittura esclusive;

- **CREW** (*Concurrent Read Exclusive Write*) indica una memoria con lettura simultanea e scrittura esclusiva;
- **CRCW** (*Concurrent Read Concurrent Write*) indica una memoria con lettura e scrittura simultanee.

Per quanto riguarda la scrittura simultanea, abbiamo diverse modalità:

- **common**: i processori possono scrivere solo se scrivono lo stesso dato;
- **random**: si sceglie un processore a caso;
- **max/min**: si sceglie il processore con il dato massimo/minimo;
- **priority**: si sceglie il processore con priorità maggiore.

La politica EREW è la più semplice, ma si può dimostrare che

$$\text{Algo(EREW)} \iff \text{Algo(CREW)} \iff \text{Algo(CRCW)} .$$

1.3.2. Risorse di calcolo

Essendo i singoli processori delle RAM, abbiamo ancora le risorse di tempo $t(n)$ e spazio $s(n)$, ma dobbiamo aggiungere:

- $p(n)$ numero di processori richiesti su input di lunghezza n nel caso peggiore;
- $T(n, p(n))$ tempo richiesto su input di lunghezza n e $p(n)$ processori nel caso peggiore.

Notiamo come $T(n, 1)$ rappresenta il **tempo sequenziale** $t(n)$.

Ogni processore P_i esegue una serie di istruzioni nel passo parallelo, che possono essere più o meno in base al processore e al numero $p(n)$ di processori. Indichiamo con

$$t_i^{(j)}(n)$$

il tempo che impiega il processore j -esimo per eseguire l' i -esimo passo parallelo su un input di lunghezza n . Quello che vogliamo ricavare è il **tempo complessivo** del passo parallelo: visto che dobbiamo aspettare che ogni processore finisca il proprio passo, calcoliamo il tempo di esecuzione $t_i(n)$ dell' i -esimo passo parallelo come

$$t_i(n) = \max \left\{ t_i^{(j)}(n) \mid 1 \leq j \leq p(n) \right\}.$$

Banalmente, il tempo complessivo di esecuzione del programma è la somma di tutti i tempi dei passi paralleli, quindi

$$T(n, p(n)) = \sum_{i=1}^{k(n)} t_i(n).$$

Notiamo subito come:

- T dipende da $k(n)$, numero di passi paralleli;
- T dipende n , dimensione dell'input;
- T dipende da $p(n)$, numero di processori, perché diminuire/aumentare i processori causa un aumento/diminuzione dei tempi dei passi paralleli.

Confrontando $T(n, p(n))$ con $T(n, 1)$ abbiamo due casi:

- $T(n, p(n)) = \Theta(T(n, 1))$, caso che vogliamo evitare;
- $T(n, p(n)) = o(T(n, 1))$, caso che vogliamo trovare.

Introduciamo lo **speed-up**, il primo parametro utilizzato per l'analisi di un algoritmo parallelo. Esso viene definito come la quantità

$$S(n, p(n)) = \frac{T(n, 1)}{T(n, p(n))}.$$

Ad esempio, $S = 4$ indica che l'algoritmo parallelo è 4 volte più veloce dell'algoritmo sequenziale, ma questo ricade nel caso $T(n, p(n)) = \Theta(T(n, 1))$, poiché il fattore che definisce la complessità si semplifica. Vogliamo avere $S \rightarrow \infty$, poiché è la situazione di *o* piccolo che tanto desideriamo.

Questo primo parametro è ottimo ma non basta: stiamo considerando il numero di processori? **NO**, questo perché $p(n)$ non compare da nessuna parte, e quindi noi potremmo avere $S \rightarrow \infty$ perché stiamo utilizzando un numero spropositato di processori.

Esempio 1.3.2.1: Ad esempio, nel **problema di soddisfacibilità (SODD)** potremmo utilizzare 2^n processori, ognuno dei quali risolve un assegnamento. Con una serie di *OR* paralleli andiamo a vedere se siamo riusciti ad ottenere un assegnamento valido di variabili, tutto questo in tempo

$$T(n, p(n)) = \log_2(2^n) = n.$$

Introduciamo quindi la variabile di **efficienza**, definita come

$$E(n, p(n)) = \frac{S(n, p(n))}{p(n)} = \frac{T(n, 1)^*}{T(n, p(n)) \cdot p(n)},$$

dove $T(n, 1)^*$ indica il miglior tempo sequenziale ottenibile.

Per comodità, da qui in avanti chiameremo $E(n, p(n))$ semplicemente con E .

Teorema 1.3.2.1: Vale

$$0 \leq E \leq 1.$$

Dimostrazione 1.3.2.1: La dimostrazione di $E \geq 0$ risulta banale visto che si ottiene come rapporto di tutte quantità positive o nulle.

La dimostrazione di $E \leq 1$ richiede di sequenzializzare un algoritmo parallelo, ottenendo un tempo $\hat{T}(n, 1)$ che però «fa peggio» del miglior algoritmo sequenziale $T(n, 1)^*$, quindi

$$T(n, 1)^* \leq \hat{T}(n, 1) \leq p(n) \cdot t_1(n) + \dots + p(n) \cdot t_{k(n)}(n).$$

La somma di destra rappresenta la sequenzializzazione dell'algoritmo parallelo, che richiede quindi un tempo uguale a $p(n)$ volte il tempo che prima veniva eseguito al massimo in un passo parallelo. Risolvendo il membro di destra otteniamo

$$T(n, 1)^* \leq \sum_{i=1}^{k(n)} p(n) \cdot t_i(n) = p(n) \sum_{i=1}^{k(n)} t_i(n) = p(n) \cdot T(n, p(n)).$$

Se andiamo a dividere tutto per il membro di destra otteniamo

$$T(n, 1)^* \leq p(n) \cdot T(n, p(n)) \implies \frac{T(n, 1)^*}{p(n) \cdot T(n, p(n))} \leq 1 \implies E \leq 1. \quad \blacksquare$$

Se $E \rightarrow 0$ abbiamo dei problemi, perché nonostante un ottimo speed-up stiamo tendendo a 0, ovvero il numero di processori utilizzati è eccessivo. Devo quindi ridurre il numero di processori p senza degradare il tempo, passando da p a $\frac{p}{k}$.

L'algoritmo parallelo ora non ha più p processori, ma avendone di meno, per garantire l'esecuzione di tutte le istruzioni, vado a raggruppare in gruppi di k le istruzioni sulla stessa riga, così che ogni processore dei $\frac{p}{k}$ totali a disposizione esegua k istruzioni.

Il tempo per eseguire un blocco di k istruzioni ora diventa $k \cdot t_i(n)$ nel caso peggiore, mentre il tempo totale diventa

$$T\left(n, \frac{p}{k}\right) \leq \sum_{i=1}^{k(n)} k \cdot t_i(n) = k \sum_{i=1}^{k(n)} t_i(n) = k \cdot T(n, p(n)).$$

Calcoliamo l'efficienza con questo nuovo numero di processori, per vedere se è migliorata:

$$E\left(n, \frac{p}{k}\right) = \frac{T(n, 1)^*}{\frac{p}{k} \cdot T\left(n, \frac{p}{k}\right)} \geq \frac{T(n, 1)^*}{\frac{p}{k} \cdot k \cdot T(n, p(n))} = \frac{T(n, 1)^*}{p(n) \cdot T(n, p(n))} = E(n, p(n)).$$

Notiamo quindi che diminuendo il numero di processori l'**efficienza aumenta**.

Possiamo dimostrare infine che la nuova efficienza è comunque limitata superiormente da 1:

$$E(n, p(n)) \leq E\left(n, \frac{p}{k}\right) \leq E\left(n, \frac{p}{p}\right) = E(n, 1) = 1.$$

Questa idea di ridurre il numero di processori è quella che viene definita come **principio di Wyllie**: se $E \rightarrow 0$ quando $T(n, p(n)) = o(T(n, 1))$ allora è $p(n)$ che sta crescendo troppo, e dobbiamo quindi ridurre il numero di processori che utilizziamo.

2. Sommatoria

Cerchiamo un algoritmo parallelo per il calcolo di una **sommatoria**.

Il problema:

- prende in **input** una serie di numeri $M[1], \dots, M[n]$;
- restituisce in **output**, dentro il registro $M[n]$, la somma dei valori $M[1], \dots, M[n]$.

Un buon algoritmo sequenziale è quello che utilizza $M[n]$ come **accumulatore**, lavorando in tempo

$$T(n, 1) = n - 1$$

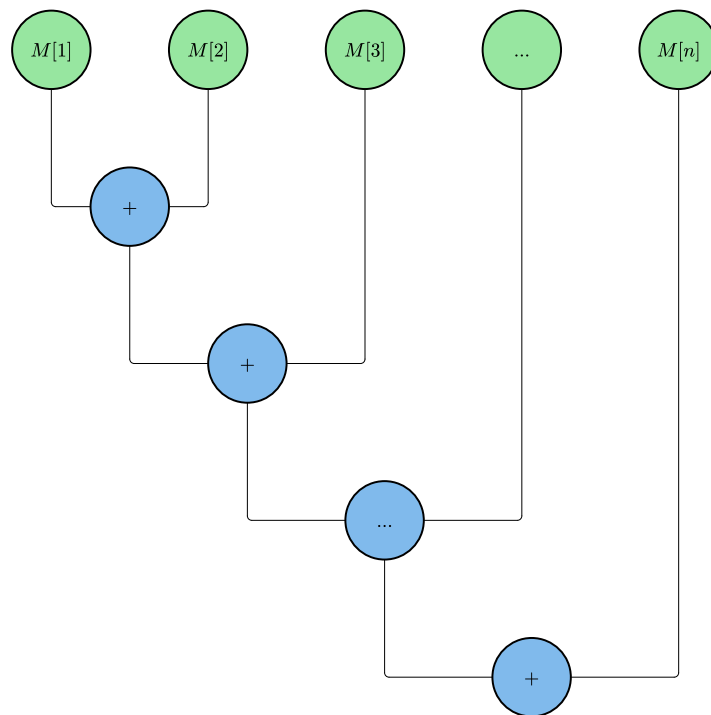
senza usare della memoria aggiuntiva.

Sommatoria sequenziale

```
1: for  $i = 1$  to  $n$  do  
2:    $M[n] = M[n] + M[i]$   
3: output  $M[n]$ 
```

2.1. Prima versione

Un primo approccio parallelo potrebbe essere quello di far eseguire ad ogni processore una somma.



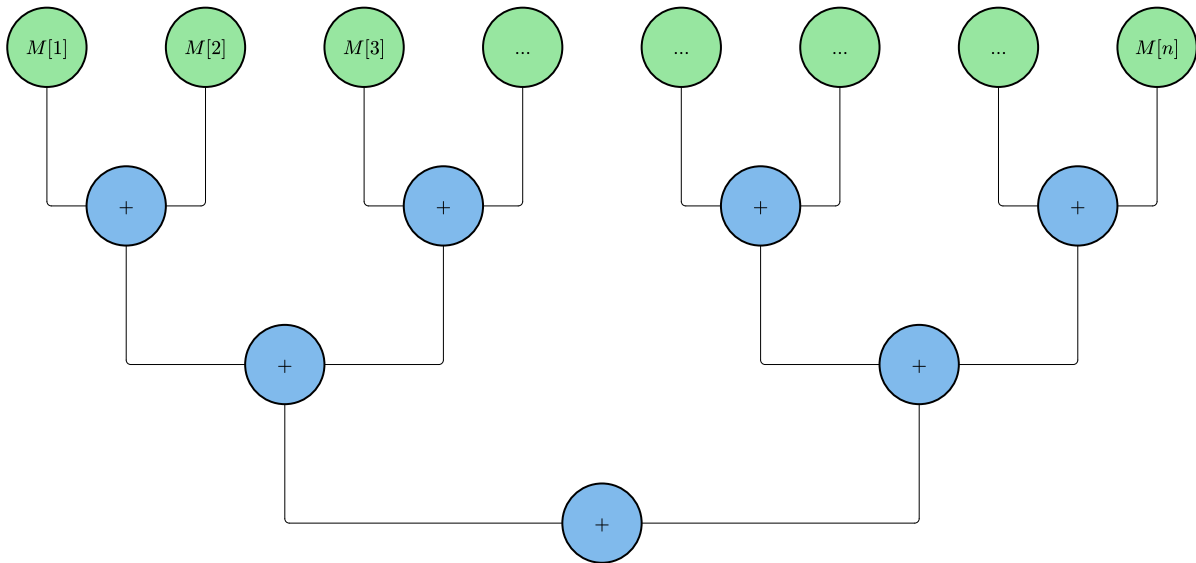
Usiamo $n - 1$ processori, ma abbiamo dei problemi:

- l'albero che otteniamo ha altezza $n - 1$;
- ogni processore deve aspettare la somma del processore precedente, quindi $T(n, n - 1) = n - 1$.

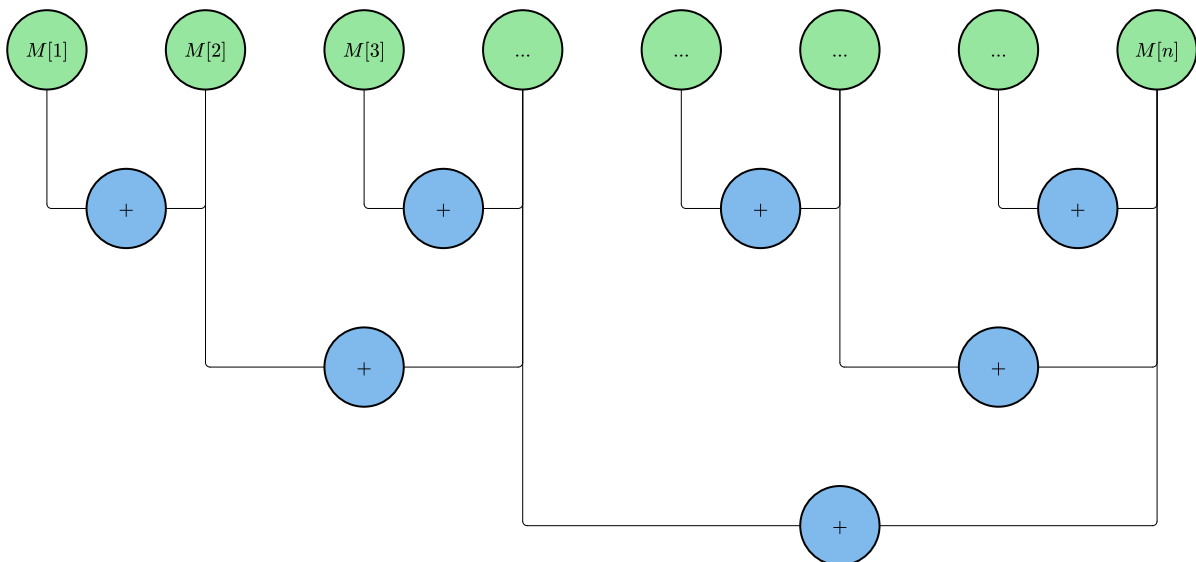
L'efficienza che otteniamo è

$$E(n, n - 1) = \frac{n - 1}{(n - 1)(n - 1)} \rightarrow 0.$$

Una soluzione migliore considera la *proprietà associativa* della somma per effettuare delle somme 2 a 2 e abbassare il tempo.



Quello che otteniamo è un albero binario, sempre con $n - 1$ processori ma altezza dell'albero logaritmica in n . Il risultato di ogni somma viene scritto nella cella di indice maggiore, quindi vediamo la rappresentazione corretta.



Quello che possiamo fare è sommare, ad ogni passo i , gli elementi che sono a distanza i : partiamo sommando elementi adiacenti a distanza 1, poi 2, fino a sommare al passo $\log(n)$ gli ultimi due elementi a distanza $\frac{n}{2}$.

Sommatoria parallela

- 1: for $i = 1$ to $\log(n)$ do
- 2: for $k = 1$ to $\frac{n}{2^i}$ par do

Sommatoria parallela

```
3:  L  L  M[2ik] = M[2ik] + M[2ik - 2i-1]  
4: return M[n]
```

Teorema 2.1.1: L'algoritmo di sommatoria parallela è EREW.

Dimostrazione 2.1.1: Dobbiamo mostrare che al passo parallelo i il processore a , che utilizza $2^i a$ e $2^i a - 2^{i-1}$, legge e scrive celle di memoria diverse rispetto a quelle usate dal processore b , che utilizza $2^i b$ e $2^i b - 2^{i-1}$.

Mostriamo prima che $2^i a \neq 2^i b$, ma questo è banale se $a \neq b$.

Mostriamo infine che $2^i a \neq 2^i b - 2^{i-1}$. Supponiamo per assurdo che siano uguali, allora

$$2 \cdot \frac{2^i a}{2^i} = 2 \cdot \frac{2^i b - 2^{i-1}}{2^i} \implies 2a = 2b - 1 \implies a = \frac{2b - 1}{2}$$

ma questo è assurdo perché $a \in \mathbb{N}$. ■

Teorema 2.1.2: L'algoritmo di sommatoria parallela è corretto.

Dimostrazione 2.1.2: Per dimostrare che l'algoritmo è corretto mostriamo che al passo parallelo i , nella cella $2^i k$, ho i $2^i - 1$ valori precedenti sommati a $M[2^i k]$, ovvero che

$$M[2^i k] = M[2^i k] + \dots + M[2^i(k-1) + 1].$$

Notiamo che se $i = \log_2(n)$ allora ho un solo processore attivo, quindi abbiamo $k = 1$ e otteniamo la definizione di sommatoria, ovvero

$$M[n] = M[n] + \dots + M[1].$$

Dimostriamo per induzione.

Passo base: se $i = 1$ allora $M[2k] = M[2k] + M[2k - 1]$.

Passo induttivo: supponiamo sia vero per $i - 1$, dimostriamo che vale per i .

Sappiamo che al generico passo i eseguiamo l'operazione $M[2^i k] = M[2^i k] + M[2^i k - 2^{i-1}]$.

Andiamo a riscrivere i due fattori della somma in un modo a noi più comodo.

Il primo fattore è

$$M[2^i k] = M[2^{i-1} \cdot (2k)] \stackrel{\text{HP-I}}{=} M[2^{i-1} \cdot (2k)] + \dots + M[2^{i-1} \cdot (2k - 1) + 1].$$

Il secondo fattore è

$$M[2^i k - 2^{i-1}] = M[2^{i-1}(2k - 1)] \stackrel{\text{HP-I}}{=} M[2^{i-1}(2k - 1)] + \dots + M[2^{i-1}(2k - 2) + 1].$$

Notiamo ora che il primo e il secondo fattore sono contigui: infatti, l'ultima cella del primo fattore è un indice superiore rispetto alla prima cella del secondo fattore. Inoltre, l'ultima cella del secondo fattore può essere riscritta come

$$M[2^i(k - 1) + 1],$$

quindi abbiamo ottenuto esattamente quello che volevamo dimostrare. ■

Se n è potenza di 2 usiamo $p(n) = \frac{n}{2}$ processori con un tempo $T(n, \frac{n}{2}) = 4 \log(n)$, dovuto alle microistruzioni che vengono fatte in ogni passo parallelo.

Se n non è potenza di 2 dobbiamo «allungare» il nostro input fino a raggiungere una dimensione uguale alla potenza di 2 più vicina, aggiungendo degli zeri in coda, ma questo non va ad intaccare le prestazioni perché la nuova dimensione è limitata da $2n$. Infatti, con lunghezza $2n$ usiamo $p(n) = \frac{2n}{2} = n$ processori e abbiamo tempo $T(n, n) = 4 \log(2n) \leq 5 \log(n)$.

In poche parole, in entrambi i casi abbiamo $p(n) = O(n)$ e $T(n, p(n)) = O(\log(n))$.

Se però calcoliamo l'efficienza otteniamo

$$E(n, n) = \frac{n - 1}{5n \log(n)} \rightarrow 0,$$

quindi dobbiamo trovare una soluzione migliore, anche se E tende a 0 lentamente.

2.2. Seconda versione [ottimizzata]

Usiamo il **principio di Wyllie**: vogliamo arrivare ad avere $E \rightarrow C \neq 0$, diminuendo il numero di processori utilizzati. Andiamo quindi ad utilizzare p processori raggruppando i numeri presenti in M in gruppi grandi $\Delta = \frac{n}{p}$, ognuno associato ad un processore.

Al primo passo dell'algoritmo ogni processore esegue la somma sequenziale dei Δ valori contenuti nel proprio gruppo, ovvero $M[k\Delta] = M[k\Delta] + \dots + M[(k - 1)\Delta + 1]$. Successivamente, si esegue l'algoritmo sommatoria proposto prima sulle celle di memoria $M[\Delta], M[2\Delta], \dots, M[p\Delta]$, e in quest'ultima viene inserito il risultato finale.

In questa versione ottimizzata usiamo $p(n) = p$ processori con un tempo $T(n, p)$ formato dal primo passo parallelo «di ottimizzazione» sommato al tempo di sommatoria, quindi

$$T(n, p) = \frac{n}{p} + 5 \log(p).$$

Andiamo a calcolare l'efficienza

$$E = \frac{n - 1}{p \cdot \left(\frac{n}{p} + 5 \log(p) \right)} = \frac{n - 1}{n + \underbrace{5p \log(p)}_n} \approx \frac{n}{2n} = \frac{1}{2} \neq 0.$$

Per arrivare ad avere questa efficienza che tanto volevamo, dobbiamo imporre $5p \log(p) = n$, quindi

$$p = \frac{n}{5 \log(n)}.$$

Con questa assunzione riusciamo ad ottenere un tempo

$$T(n, p(n)) \leq 10 \log(n).$$

Diamo un **lower bound**: sommatoria possiamo visualizzarlo usando un albero binario, dove le foglie sono i dati di input e i livelli sono i passi paralleli. Il livello con più nodi dà il numero di processori e l'altezza dell'albero dà il tempo dell'algoritmo. Se abbiamo altezza h , abbiamo massimo 2^h foglie, quindi

$$\text{foglie} = n \leq 2^h \implies h \geq \log(n),$$

quindi ho sempre tempo logaritmico.

2.3. Operazione iterata

La sommatoria può essere uno schema per altri problemi.

L'**operazione iterata** è una operazione associativa **OP** sulla quale definiamo un problema che:

- prende in **input** una serie di valori $M[1], \dots, M[n]$;
- restituisce in **output**, dentro il registro $M[n]$, l'operazione **OP** calcolata su tutti gli $M[i]$.

Abbiamo visto che sommatoria ammette soluzioni efficienti con $p(n) = O\left(\frac{n}{\log(n)}\right)$ processori utilizzati e tempo $T(n, p(n)) = O(\log(n))$.

Con modelli PRAM più potenti (*non EREW*) possiamo ottenere un tempo costante per AND e OR.

Supponiamo una CRCW-PRAM e vediamo il problema **AND-iterato**, ovvero

$$M[n] = \bigwedge_i M[i].$$

Come detto poco fa, il tempo è costante perché la PRAM è più potente.

\bigwedge -iterato

```

1: for  $1 \leq k \leq n$  par do
2:   if  $M[k] = 0$  then
3:      $M[n] = 0$ 
```

Serve CW con **politica common**, quindi scrivono i processori se e solo se il dato da scrivere è uguale per tutti. In realtà, anche le altre politiche (*random o priority*) vanno bene.

In questo specifico caso abbiamo $p(n) = n$ processori e tempo $T(n, n) = 3$. Calcoliamo l'efficienza di questo problema come

$$E = \frac{n-1}{3n} \rightarrow \frac{1}{3}.$$

Per il problema **OR-iterato** vale la stessa cosa, basta modificare leggermente il codice proposto.

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

3. Applicazioni di sommatoria

Il problema sommatoria lo possiamo utilizzare come modulo per risolvere dei problemi più complessi. In questo capitolo vediamo quattro problemi molto semplici.

3.1. Prodotto interno di vettori

Questo problema:

- prende in **input** due vettori $x, y \in \mathbb{N}^n$;
- restituisce in **output** il **prodotto scalare**.

Il prodotto scalare è un numero reale definito dalla formula

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i.$$

Il miglior tempo sequenziale è $T(n, 1) = 2n - 1$, formato da n prodotti e $n - 1$ somme finali.

Il modulo sommatoria viene usato per eseguire le somme finali in parallelo. L'algoritmo che vedremo si articola in due fasi:

- eseguo $\log(n)$ prodotti in sequenza delle componenti e la somma dei valori del blocco in sequenza;
- effettuo la somma di $p = \frac{n}{\log(n)}$ prodotti in parallelo.

Per sommatoria ci serviranno $p = c_1 \frac{n}{\log(n)}$ processori, con tempo $t = c_2 \log(n)$. Per la prima fase ci serviranno invece $p = \frac{n}{\log(n)}$ processori, con tempo $t = c_3 \log(n)$.

Ma allora utilizziamo in totale $p = \frac{n}{\log(n)}$ processori con tempo $t = \log(n)$. L'efficienza è

$$E = \frac{2n - 1}{\frac{n}{\log(n)} \cdot \log(n)} \rightarrow C \neq 0.$$

3.2. Prodotto matrice vettore

Questo problema:

- prende in **input** una matrice $A \in \mathbb{N}^{n \times n}$ e un vettore $x \in \mathbb{N}^n$;
- restituisce in **output** il prodotto Ax .

Per questo problema usiamo il prodotto interno di vettori come modulo.

Il migliore tempo sequenziale è $T(n, 1) = n(2n - 1) = 2n^2 - n$.

Per l'approccio parallelo, l'idea è usare il modulo del prodotto interno in parallelo per n volte. Il vettore se è acceduto simultaneamente dai moduli precedenti ci obbliga ad avere una politica CREW.

Questa idea utilizza $p(n) = \frac{n^2}{\log(n)}$ processori con tempo $T(n, p(n)) = \log(n)$. "efficienza vale

$$E = \frac{n^2}{\frac{n^2}{\log(n)} \log(n)} \rightarrow C \neq 0.$$

3.3. Prodotto matrice matrice

Questo problema:

- prende in **input** due matrici $A, B \in \mathbb{N}^{n \times n}$;
- restituisce in **output** il prodotto matriciale AB .

Usiamo ancora il prodotto interno come modulo per risolvere questo problema.

Il miglior tempo sequenziale è $T(n, 1) = n^{2.8}$, ottenuto con l'**algoritmo di Strassen**.

Come prima, l'idea è fare dei prodotti interni paralleli, solo che in questo caso sono n^2 . Ci servirà ancora la politica CREW, per via dell'accesso simultaneo alle righe di A e alle colonne di B .

Questo algoritmo usa $p(n) = \frac{n^3}{\log(n)}$ processori con tempo $T(n, p(n)) = \log(n)$. L'efficienza vale

$$E = \frac{n^{2.80}}{\frac{n^3}{\log(n)} \log(n)} \rightarrow 0.$$

L'efficienza tende a 0 ma lentamente, quindi lo accettiamo come risultato.

3.4. Potenza di matrice

Questo ultimo problema:

- prende in **input** una matrice $A \in \mathbb{N}^{n \times n}$;
- restituisce in **output** la potenza A^t , con $t = 2^k$.

Questo problema si risolve come prodotto iterato con tempo $T(n, 1) = n^{2.80} \log(n)$.

Potenza di matrice sequenziale

```
1: for  $i = 1$  to  $\log(n)$  do
2:    $A = A \cdot A$ 
```

L'approccio parallelo per $\log(n)$ volte esegue il prodotto $A \cdot A$, anche questo con politica CREW.

Per questo problema utilizziamo $p(n) = \frac{n^3}{\log(n)}$ processori con tempo $T(n, p(n)) = \log^2(n)$.

Purtroppo, l'efficienza che otteniamo è

$$E = \frac{n^{2.8} \log(n)}{\frac{n^3}{\log(n)} \cdot \log^2(n)} = \frac{n^{2.8}}{n^3} \rightarrow 0.$$

Come prima però, accettiamo l'efficienza visto che tende a 0 lentamente.

4. Somme prefisse

Anche il problema delle **somme prefisse** userà il modulo della sommatoria per essere risolto.

Questo problema

- prende in **input** una serie di numeri $M[1], \dots, M[n]$;
- restituisce in **output**, nella cella k -esima, la somma di tutti gli elementi precedenti più se stesso.

In poche parole, nella cella k -esima abbiamo la quantità

$$M[k] = \sum_{i=1}^k M[i].$$

Assumiamo, per semplicità, che n sia una potenza di 2.

Il migliore algoritmo sequenziale somma nella cella i quello che c'è nella cella $i - 1$.

Algoritmo sequenziale furbo

```
1: for  $k = 2$  to  $n$  do  
2:    $M[k] = M[k] + M[k - 1]$ 
```

Il tempo di questo algoritmo è $T(n, 1) = n - 1$.

4.1. Prima versione

Vediamo una prima proposta parallela. Al modulo sommatoria passo tutti i possibili prefissi: un modulo somma i primi due, un modulo i primi tre, eccetera.

Con questo approccio abbiamo un paio di problemi:

- l'algoritmo non è EREW, ma questo è facilmente risolvibile;
- l'algoritmo usa $p(n) = \frac{n^2}{\log(n)}$ processori con tempo $T(n, p(n)) = \log(n)$.

Con questo dispendio di tempo e spazio, l'efficienza vale

$$E = \frac{n - 1}{\frac{n^2}{\log(n)} \log(n)} \rightarrow 0.$$

4.2. Seconda versione [pointer doubling]

Usiamo il **pointer doubling**, un algoritmo ideato da **Kogge-Stone** nel 1973.

Dobbiamo stabilire dei legami tra i numeri, e ognuno di questi legami viene preso in carico da un processore, che ne fa la somma. Quest'ultima viene poi inserita nella cella di indice maggiore.

Alla prima iterazione ho dei link tra una cella e la successiva. Alla seconda iterazione ho dei link tra una cella e quella due posizioni dopo. Alla terza iterazione ho dei link tra una cella e quella quattro posizioni dopo. Alla quarta iterazione eccetera.

Notiamo due cose:

- la distanza dei link raddoppia ad ogni iterazione;
- alcuni processori non hanno dei successori, ovvero qualche processore non riesce a linkarsi con nessun altro elemento del vettore.

L'algoritmo termina quando non riesco più a mettere dei link tra celle di memoria.

Vediamo un po' di fatti che riguardano questo algoritmo:

- ogni volta raddoppiamo la distanza dei link, quindi l'algoritmo lavora in **tempo logaritmico**.

- prima di eseguire una somma e un aggiornamento di link al passo i , in quel momento erano attivi $n - 2^{i-1}$ processori.
- alla fine del passo i , il numero di processori senza successore è 2^i .

Ci serve sicuramente un vettore di successori, che usiamo per ricavare le celle da sommare tra loro.

Sia S tale vettore. Sia $S[k]$ il successore di $M[k]$. Come inizializzo S ?

Prima della prima iterazione assegno

- $S[k] = k + 1$;
- $S[n] = 0$.

Algoritmo di Kogge-Stone

```

1: for  $i = 1$  to  $\log(n)$  do
2:   for  $1 \leq k \leq n - 2^{i-1}$  par do
3:      $M[S[k]] = M[k] + M[S[k]]$ 
4:      $S[k] = S[S[k]]$ 

```

L'algoritmo incredibilmente è EREW perché accediamo sì alle stesse celle, ma in momenti diversi.

Teorema 4.2.1: L'algoritmo di Kogge-Stone è corretto.

Dimostrazione 4.2.1: Siamo in una EREW-PRAM, quindi il processore P_i lavora su $M[i]$ e $M[S[i]]$, e se considero $i \neq j$ allora $S[i] \neq S[j]$ e quindi abbiamo successori diversi (*accettiamo il caso in cui entrambi i successori sono nulli*).

Devo dimostrare che nella cella k -esima cella ho la somma degli elementi precedenti più l'elemento stesso. Vale la proprietà che dice che all' i esimo passo vale

$$M[t] = \begin{cases} M[t] + \dots + M[1] & \text{se } t \leq 2^i \\ M[t] + \dots + M[t - 2^i + 1] & \text{se } t > 2^i \end{cases}$$

Se $i = \log(n)$, ovvero sono all'ultima iterazione, siamo nel primo caso della funzione definita a tratti e quindi vale

$$M[t] = M[t] + \dots + M[1].$$

■

Mostriamo che vale la proprietà descritta nella dimostrazione precedente.

Dimostrazione 4.2.2: Dimostriamo questa proprietà per induzione su i

Caso base: se $i = 1$ allora:

- se $t \leq 2$ allora $M[1] = M[1]$ e $M[2] = M[2] + M[1]$;
- se $t > 2$ allora $M[t] = M[t] + M[t - 1]$.

Passo induttivo: assunto vero per $i - 1$ e dimostro vero per i .

Prima di iniziare il passo i -esimo il vettore S contiene

$$S[k] = \begin{cases} k + 2^{i-1} & \text{se } k \leq n - 2^{i-1} \\ 0 & \text{se } k > n - 2^{i-1} \end{cases}$$

Le celle con indice $\leq 2^{i-1}$ hanno la proprietà vera per ipotesi induttiva.

Concentriamoci sugli indici che avanzano da analizzare.

Se $2^{i-1} < t \leq 2^i$ allora possiamo scrivere t come $t = 2^{i-1} + a$ e quindi

$$\begin{aligned} M[a + 2^{i-1}] &= \underbrace{M[a]}_{a \leq 2^{i-1}} + \underbrace{M[a + 2^{i-1}]}_{a + 2^{i-1} > 2^{i-1}} = \\ &= M[1] + \dots + M[a] + | + M[a + 1] + \dots + M[a + 2^{i-1}]. \end{aligned}$$

Se invece $t > 2^i$ allora possiamo scrivere t come $t = a + 2^i$ e quindi

$$\begin{aligned} M[a + 2^i] &= M[(a + 2^{i-1}) + 2^{i-1}] = \underbrace{M[a + 2^{i-1}]}_{a + 2^{i-1} > 2^{i-1}} + \underbrace{M[a + 2^i]}_{a + 2^i > 2^{i-1}} = \\ &= M[a + 1] + \dots + M[a + 2^{i-1}] + | + M[a + 2^{i-1} + 1] + \dots + M[a + 2^i]. \end{aligned}$$

In entrambi gli indici mancanti la proprietà risulta vera. ■

Valutiamo infine questo algoritmo: esso utilizza $p(n) = n - 1$ processori con un utilizzo di tempo uguale a $T(n, p(n)) = 8 \log(n)$. Il fattore logaritmico viene dal passo parallelo. L'efficienza è quindi

$$E = \frac{n - 1}{(n - 1)8 \log(n)} = \frac{1}{8 \log(n)} \rightarrow 0.$$

Anche se l'efficienza tende a 0 lentamente, non siamo soddisfatti di questa soluzione.

Sfruttiamo il **principio di Wyllie** per far sparire il fattore $\log(n)$ dal denominatore. Usiamo quindi $p(n) = O\left(\frac{n}{\log(n)}\right)$ processori, ottenendo sempre un tempo logaritmico ma andremo ad avere

$$E = \frac{n - 1}{c_1 \frac{n}{\log(n)} c_2 \log(n)} = \frac{n - 1}{c_1 c_2 n} \rightarrow C \neq 0.$$

Questo problema può essere usato come modulo per risolvere il problema **OP-prefissa**, dove non devo più utilizzare la somma come operazione associativa ma devo usare una operazione **OP** generica.

5. Valutazione di polinomi

Questo problema

- prende in **input** un polinomio $p(x) = a_0 + a_1x + \dots + a_nx^n$ e un valore $\alpha \in \mathbb{R}$;
- restituisce in **output** la valutazione di p nel valore α , ovvero $p(\alpha)$.

In memoria ho il valore α e i coefficienti del polinomio $A[0], \dots, A[n]$.

L'algoritmo sequenziale esegue circa n^2 operazioni, tra somme e prodotti.

Con il metodo di **Ruffini-Horner** possiamo abbassare il numero di operazioni. L'idea che hanno avuto questi pazzi è quella di eseguire un raccoglimento di x iterativo, ottenendo

$$p(x) = a_0 + x(a_1 + \dots(a_{n-2} + x(a_{n-1} + a_nx))\dots).$$

Partiamo con $p = a^n$. Ad ogni passo dell'algoritmo calcoliamo somma+prodotto di una parentesi e assegniamo questo valore di nuovo a p . Vale quindi che

$$p = a_i + p\alpha.$$

Algoritmo di Ruffini-Horner

```
1:  $p = a_n$ 
2: for  $i = 1$  to  $n$  do
3:    $p = a_{n-i} + p\alpha$ 
4: output  $p$ 
```

Contando le 2 operazioni elementari, questo tempo sequenziale vale $T(n, 1) = 2n$.

Per l'algoritmo parallelo dobbiamo:

- costruire il vettore delle potenze di α , chiamato Q , e tale che $Q[k] = \alpha^k \mid 0 \leq k \leq n$;
- eseguire il prodotto interno tra A e Q .

Il vettore Q possiamo calcolarlo grazie al problema **replica**: dobbiamo mettere α in tutti gli elementi di Q da 1 a n , e poi applicare il prodotto prefisso.

Replica

```
1: for  $k = 1$  to  $n$  par do
2:    $Q[k] = \alpha$ 
```

L'algoritmo scritto è CREW, perché α è in memoria e quindi ho accesso simultaneo al dato, usa $p(n) = n$ processori con tempo $T(n, p(n)) = 2$. L'efficienza vale

$$E = \frac{n}{2n} = \frac{1}{2} \neq 0.$$

Possiamo modificare leggermente il modulo replica, perché se poi lo dobbiamo utilizzare prima del prodotto prefisso molti processori vengono inutilizzati. Abbassiamo quindi il numero di processori con il **principio di Wyllie**, raggruppando in gruppi di $\log(n)$ elementi l'input. Il k -esimo processore carica α nelle celle di indice

$$(k-1)\log(n) + 1 \quad | \quad \dots \quad | \quad k\log(n).$$

Replica migliorato

```
1: for  $k = 1$  to  $\frac{n}{\log(n)}$  par do
2:   for  $i = 1$  to  $\log(n)$  do
   L  $Q[(k-1)\log(n) + i] = \alpha$ 
```

Questa nuova versione usa $p(n) = \frac{n}{\log(n)}$ processori con tempo $t = c \log(n)$. L'efficienza vale

$$E = \frac{n}{c \frac{n}{\log(n)} \log(n)} = \frac{1}{c} \neq 0.$$

Anche questa versione però rimane CREW per l'accesso ad α . Noi però vorremmo un EREW, quindi:

- costruiamo il vettore $[\alpha, 0, \dots, 0]$;
- eseguiamo somme prefisse

Replica ancora migliore

```
1:  $Q[1] = \alpha$ 
2: for  $k = 2$  to  $n$  par do
3:   L  $Q[k] = 0$ 
4: SommePrefisse( $Q$ )
```

Posso anche ridurre i processori con il **principio di Wyllie**, usando

- $p(n) = \frac{n}{\log(n)}$ processori e tempo $T(n, p(n)) = \log(n)$ per costruire Q ;
- $p(n) = \frac{n}{\log(n)}$ processori e tempo $T(n, p(n)) = \log(n)$ per le somme prefisse.

Finalmente abbiamo un algoritmo EREW.

Valutazioni di polinomi parallela

```
1: Replica( $Q, \alpha$ )
2: ProdottoPrefisso( $Q$ )
3: ProdottoInterno( $A, Q$ )
```

Tutti i moduli utilizzati in questa ultima versione dell'algoritmo usano $p(n) = \frac{n}{\log(n)}$ processori con tempo $T(n, p(n)) = \log(n)$. L'efficienza, anche in questo caso, tende a $C \neq 0$.

6. Ricerca

Questo problema:

- prende in **input** una serie di valori $M[1], \dots, M[n]$ e un valore α ;
- restituisce in **output** 1 se $\alpha \in M$, 0 altrimenti, mettendo il risultato della valutazione in $M[n]$.

Il migliore algoritmo sequenziale ha tempo $T(n, 1) = n$ se consideriamo un array generico. Nel caso di array ordinato, tramite **ricerca dicotomica**, abbassiamo il tempo ad un fattore logaritmico. Considerando invece un algoritmo quantistico su un array non ordinato, il tempo vale $T(n, 1) = \sqrt{n}$, ma questo è dato dall'uso dell'**interferenza quantistica**.

Vediamo una prima versione CRCW parallela che utilizza una flag F .

CRCW con flag

```
1:  $F = 0$ 
2: for  $k = 1$  to  $n$  par do
3:   if  $M[k] == \alpha$ 
4:      $F = 1$ 
5:  $M[n] = F$ 
```

L'uso della flag è necessario: senza flag, se a fine algoritmo vale $M[n] == 1$ avrei due casi:

- ho avuto esito positivo della ricerca;
- il valore 1 era il valore originale che avevo in memoria.

Con questa versione ho quindi CR in α e CW in F . Vengono utilizzati $p(n) = n$ processori con tempo costante. L'efficienza vale

$$E = \frac{n}{nC} = \frac{1}{C} \neq 0.$$

Vediamo una seconda versione del problema, utilizzando un CREW (quindi senza l'uso di flag).

CREW con MAX-iterato

```
1: for  $k = 1$  to  $n$  par do
2:    $M[k] = (M[k] == \alpha ? 1 : 0)$ 
3: MAX-iterato( $M$ )
```

Trasformiamo M in un vettore booleano e poi vediamo il massimo valore presente in esso. Come prima, abbiamo la CR per l'accesso al valore α .

Stiamo utilizzando $p(n) = n$ processori, sempre con tempo costante. Con il **principio di Wyllie** andiamo a ridurre i processori a $p(n) = \frac{n}{\log(n)}$ e il tempo a $T(n, p(n)) = \log(n)$, che sono uguali a quelli del MAX-iterato. L'efficienza è quindi

$$E = \frac{n}{\frac{n}{\log(n)} \log(n)} = C \neq 0.$$

La terza versione del problema usa invece una politica EREW, che a noi piace molto.

EREW con replica e MAX-iterato

```
1: Replica( $A, \alpha$ )
2: for  $k = 1$  to  $n$  par do
3:    $\llcorner M[k] = (M[k] == A[k] ? 1 : 0)$ 
4: MAX-iterato( $M$ )
```

Usando sempre il **principio di Wyllie**, ogni modulo di questo algoritmo usa $p(n) = \frac{n}{\log(n)}$ processori con tempo $T(n, p(n)) = \log(n)$. L'efficienza vale quindi

$$E = \frac{n}{\frac{n}{\log(n)} \log(n)} = C \neq 0.$$

Il problema di ricerca ha alcune **varianti**:

- numero di volte in cui α compare dentro M , risolto usando il modulo sommatoria al posto del modulo MAX-iterato, così da contare effettivamente quante volte α è presente;
- indice massimo di α dentro M , risolto assegnando $M[k] = k$ quando cerchiamo α dentro M , mantenendo poi il MAX-iterato alla fine;
- posizione minima di α dentro M , risolto usando una **OP iterata** tale che

$$\text{OP}(x, y) = \begin{cases} \min(x, y) & \text{se } x \neq 0 \wedge y \neq 0 \\ \max(x, y) & \text{altrimenti} \end{cases}.$$

7. Ordinamento

Detto anche problema del **ranking**, questo problema:

- prende in **input** una serie di valori $M[1], \dots, M[n]$;
- restituisce in **output** una **permutazione** $p : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ tale che

$$M[p(1)] \leq \dots \leq M[p(n)],$$

con $p(i)$ indice dell'elemento del vettore M che va in posizione i .

Potremmo anche ordinare direttamente in M , ma noi decidiamo di usare gli indici.

In genere, gli algoritmi di ordinamento sono basati sui **confronti**, ovvero dei check nella forma

$$M[i] \leq M[j] ? \quad \text{SI} : \text{NO} .$$

Teorema 7.1: Gli algoritmi di ordinamento basati sui confronti hanno tempo

$$T(n) = \Theta(n \log(n)).$$

Dimostrazione 7.1: Dimostriamo i due bound di questi algoritmi:

- upper bound: esistono algoritmi, tipo il merge sort, che impiegano al massimo $n \log(n)$ passi per essere eseguiti;
- lower bound: costruiamo un albero binario di decisione, dove ogni nodo è un possibile confronto. Il numero di foglie che otteniamo è il numero di permutazioni dell'input, che sono $n!$. L'altezza dell'albero è il numero di confronti che devo fare, perché un cammino dalla radice ad una foglia percorre tutti i confronti che devono essere fatti per ordinare l'input in quel preciso modo. Sappiamo che il numero di foglie di un albero binario è al massimo 2^h , con h altezza dell'albero, quindi

$$2^h \geq \# \text{foglie} = n! \xrightarrow{\text{STIRLING}} h \geq \log_2 \left(\sqrt{2\pi n} \left(\frac{n}{e} \right)^n \right) \Rightarrow h \geq n \log(n).$$

Quindi vale $\Theta(n \log(n))$. ■

7.1. Counting sort

Vediamo un primo algoritmo di ordinamento, basato sul **conteggio**. Il **counting sort** conta i confronti. Sequenzialmente, il tempo è $T(n, 1) = \Theta(n^2)$ perché deve confrontare tutte le coppie. Assumiamo, per semplicità, che n sia potenza di 2 e che gli elementi siano tutti diversi tra loro.

Nel counting sort, $M[i]$ va in posizione k se e solo se ci sono k elementi $\leq M[i]$ in M .

Per effettuare questo conteggio, usiamo un vettore V che contiene i vari valori di k per ogni valore in M . Questo vettore è la **permutazione inversa** di p :

- prima, data una posizione, mi veniva detto che indice doveva finire in quel posto;
- ora, dato un indice, mi viene detto in che posizione devo finire.

Counting Sort sequenziale

1: for $i = 1$ to n

Counting Sort sequenziale

```
2:  L  $V[i] = 0$ 
3:  for  $i = 1$  to  $n$ 
4:    for  $j = 1$  to  $n$ 
5:      if  $M[j] \leq M[i]$  then
6:        L  $V[i] = V[i] + 1$ 
7:  for  $i = 1$  to  $n$ 
8:    L  $F[V[i]] = M[i]$ 
9:  for  $i = 1$  to  $n$ 
10: L  $M[i] = F[i]$ 
```

I primi due cicli for andrebbero già bene, ma abbiamo aggiunto gli ultimi due for per ordinare effettivamente il vettore. Il numero di confronti è $O(n^2)$, visto il doppio ciclo for, quindi anche il tempo $T(n, p(n))$ assume quel valore.

La versione parallela utilizza un processore per ogni coppia di indici (i, j) che calcola $M[j] \leq M[i]$ e aggiorna una matrice booleana $V[i, j]$ con il risultato del confronto. Notiamo come l' i -esima riga individua gli elementi di M che sono $\leq M[i]$ quando la cella contiene il valore 1.

Per ottenere il numero di elementi complessivo, che effettivamente ci interessa, eseguo la sommatoria parallela sulle righe, ottenendo un vettore colonna $V[i, n]$ che coincide con il vettore V della versione sequenziale precedente.

Counting Sort parallelo

```
1: for  $i \leq n \wedge j \leq n$  par do
2:  L  $V[i, j] = (M[j] \leq M[i] ? 1 : 0)$ 
3: for  $i = 1$  to  $n$  par do
4:  L Sommatoria( $V[i, 1], \dots, V[i, n]$ )
5: for  $i = 1$  to  $n$  par do
6:  L  $M[V[i]] = M[i]$ 
```

Questo algoritmo non assolutamente EREW, vista la lettura concorrente, ma la scrittura non lo è invece, visto che scriviamo in celle ogni volta diverse. L'algoritmo è quindi CREW.

Per il primo ciclo for usiamo $p(n) = n^2$ processori con tempo $T(n, p(n)) = 4$, che però con il **principio di Wyllie** trasformiamo in $p(n) = \frac{n^2}{\log(n)}$ processori con tempo $T(n, p(n)) = \log(n)$.

Per il secondo ciclo for usiamo $p(n) = \frac{n^2}{\log(n)}$ processori, ovvero n modulo sommatoria da $\frac{n}{\log(n)}$ processori ciascuno, con tempo $T(n, p(n)) = \log(n)$.

Per il terzo e ultimo ciclo for usiamo $p(n) = n$ processori con tempo $T(n, p(n)) = 3$.

I processori totali sono quindi $p(n) = \frac{n^2}{\log(n)}$ con tempo $T(n, p(n)) = \log(n)$. L'efficienza vale

$$E = \frac{n \log(n)}{\frac{n^2}{\log(n)} \log(n)} = \frac{\log(n)}{n} \rightarrow 0$$

e nemmeno lentamente, quindi dobbiamo trovare una soluzione alternativa.

7.2. Bitonic sort

Un algoritmo parallelo migliore è il **bitonic sort**, che avrà comunque efficienza $E \rightarrow 0$ ma molto lentamente. Un altro algoritmo parallelo ottimo è quello di **Cole**, del 1988, ma non lo vedremo, nonostante abbia $E \rightarrow C \neq 0$.

Per il bitonic sort prendiamo spunto dal merge sort, che usa la tecnica Divide et Impera.

Merge sort

```
1: if  $|A| > 1$ 
2:    $A_s = \text{MergeSort}(A[1], \dots, A[\frac{n}{2}])$ 
3:    $A_d = \text{MergeSort}(A[\frac{n}{2} + 1], \dots, A[n])$ 
4:    $A = \text{Merge}(A_s, A_d)$ 
5: return  $A$ 
```

La routine di merge effettua l'unione dei due array ordinati A_s e A_d scorrendoli in sequenza e mettendo i valori ordinati in A . Nel caso peggiore, questa routine impiega tempo $T_m(n) = n$. Il tempo di complessivo di merge sort è invece $T(n) = n \log(n)$.

Possiamo sfruttare l'idea che ci dà il merge sort? Dividiamo continuamente il vettore fino ad arrivare ad un elemento. Qua devo fare il merge, se usassi un approccio parallelo avrei $\log(n) - 1$ passi paralleli essendo un albero. Purtroppo, il merge non è parallelizzabile, e quindi avrei sempre e comunque un tempo $T(n, p(n)) = n \log(n)$.

Quando l'operazione di merge è facile? Supponiamo A_s e A_d ordinate con gli elementi di A_s tutti minori di A_d . La routine di merge è facilissima: basta concatenare le due sequenze.

Cercheremo di ottenere questa situazione usando delle sequenze di numeri particolari, che sono dette **sequenze bitoniche**.

Abbiamo due operazioni fondamentali:

- **reverse**, per fare il reverse di una sequenza;
- **minmax**, che costruisce i vettori A_{\min} e A_{\max} effettuando i seguenti passi:
 - divide a metà il vettore A nei vettori $A_{\min}(sx)$ e $A_{\max}(dx)$;
 - prese le posizioni a distanza $\frac{n}{2}$ in A , nel vettore A_{\min} viene messa la componente più piccola delle due e nel vettore A_{\max} viene messa la componente più grande delle due.

Reverse

```
1: for  $1 \leq k \leq \frac{n}{2}$  par do
2:    $\text{Swap}(A[k], A[n - k + 1])$ 
```

Vengono utilizzati $p(n) = \frac{n}{2}$ processori con tempo $T(n, p(n)) = 4$.

MinMax

```
1: for  $1 \leq k \leq \frac{n}{2}$  par do
2:   if  $A[k] > A[k + \frac{n}{2}]$ 
3:      $\text{Swap}(A[k], A[k + \frac{n}{2}])$ 
```

Vengono utilizzati $p(n) = \frac{n}{2}$ processori con tempo $T(n, p(n)) = 5$.

Vediamo ora le due sequenze numeriche che useremo per risolvere questo problema.

Definizione 7.2.1 (*Sequenza unimodale*): Una sequenza A è **unimodale** se e solo se

$$\exists k \in \{1, \dots, n\} \mid A[1] > A[2] > \dots > A[k] < A[k+1] < \dots < A[n]$$

oppure

$$\exists k \in \{1, \dots, n\} \mid A[1] < A[2] < \dots < A[k] > A[k+1] > \dots > A[n].$$

In poche parole, esiste un indice che mi individua un valore che mi fa da minimo/massimo e la sequenza è decrescente/crescente poi crescente/decescente. In soldoni, il vettore **non è perfettamente ordinato**.

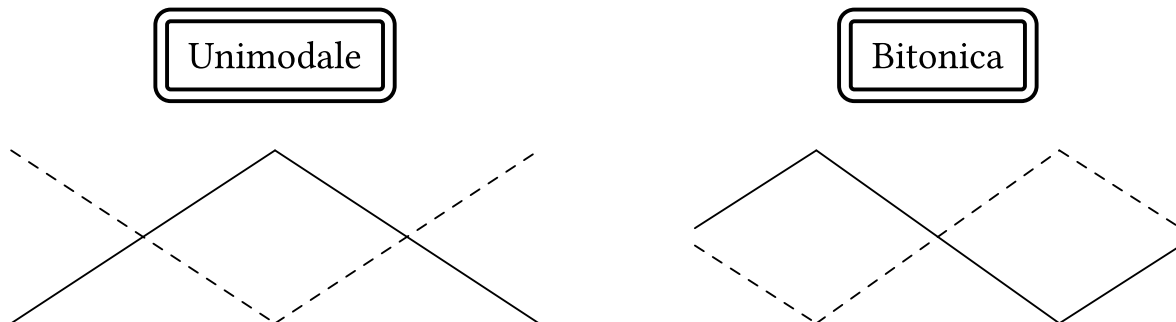
Definizione 7.2.2 (*Sequenza bitonica*): Una sequenza A è **bitonica** se e solo se esiste una permutazione ciclica di A che mi dà una sequenza unimodale, ovvero se

$$\exists k \in \{1, \dots, n\} \mid A[k], \dots, A[n], A[1], \dots, A[k-1]$$

è una sequenza unimodale.

Graficamente, una sequenza unimodale ha un picco *massimo/minimo*, mentre una sequenza bitonica ha due picchi:

- un *minimo+massimo* con i valori della coda iniziale più piccoli dei valori della coda finale;
- un *massimo+minimo* con i valori della coda iniziale più grandi dei valori della coda finale.



Vediamo finalmente l'algoritmo per ordinare sequenze bitoniche, ideato da **Batcher** nel 1968.

Osserviamo che:

- una sequenza unimodale è anche bitonica, grazie alla permutazione identità;
- siano A e B due sequenze ordinate, allora $A \cdot \text{Reverse}(B)$ è unimodale.

Vediamo delle proprietà ora.

Lemma 7.2.1: Sia A una sequenza bitonica. Se eseguo MinMax su A ottengo:

- A_{\min} e A_{\max} bitoniche;

- ogni elemento di A_{\min} minore di ogni elemento di A_{\max} .

Le osservazioni e le proprietà ci suggeriscono un approccio Divide et Impera:

- MinMax divide il problema di n elementi su istanze più piccole grazie alla prima parte del lemma;
- ordinando A_{\min} e A_{\max} la fusione di due sequenze ordinate avviene per concatenazione grazie alla seconda parte del lemma.

Bitonic merge sequenziale

```

1:  $A_{\min} \mid A_{\max} = \text{MinMax}(A)$ 
2: if  $|A| > 2$ 
3:   BitonicMerge( $A_{\min}$ )
4:   BitonicMerge( $A_{\max}$ )
5: return  $A$ 

```

Dobbiamo stare attenti a considerare **solo sequenze bitoniche**.

Teorema 7.2.1: L'algoritmo di bitonic merge è corretto.

Dimostrazione 7.2.1: Dimostriamolo per induzione su n .

Caso base: se $n = 2$ con MinMax scambio i due elementi se sono disordinati, poi ritorno il vettore A ordinato senza fare altro.

Passo induttivo: supponiamo vero per $n = 2^k$, mostriamo vero per $n = 2^{k+1}$

Calcolando MinMax su input di lunghezza 2^{k+1} otteniamo due sequenze di lunghezza 2^k , ma le sequenze di lunghezza 2^k bitonic merge le riesce a ordinare perfettamente per ipotesi induttiva. Grazie al lemma precedente, ogni elemento di A_{\min} è minore di ogni elemento di A_{\max} , quindi il vettore A è ordinato. ■

Vediamo l'implementazione parallela di bitonic merge:

- eseguiamo MinMax su input di lunghezza

$$\frac{n}{2^{i-1}},$$

partendo con $i = 1$, ottenendo di volta in volta due sequenze bitoniche A_{\min} e A_{\max} con i valori più piccoli nella prima e i valori più grandi nella seconda;

- al passo finale l'input è di lunghezza 2, quindi fermo le esecuzioni di MinMax;
- le sequenze di 2 elementi sono ordinate, e ora avviene una normalissima concatenazione di tutte le sequenze ottenute nei vari passi.

L'algoritmo è EREW perché lavora ogni volta su elementi diversi delle sequenze.

L'algoritmo termina i passi paralleli quando

$$\frac{n}{2^{i-1}} = 2 \Rightarrow i = \log(n).$$

Il MinMax ha costo $T(n, p(n)) = 5$ quindi $T(n, p(n)) = 5 \log(n)$. Il primo passo richiede $p(n) = \frac{n}{2}$ processori, il secondo $p(n) = \frac{n}{4} + \frac{n}{4} = \frac{n}{2}$ processori, eccetera. Il numero di processori che andiamo ad utilizzare è quindi $p(n) = \frac{n}{2}$.

Se vogliamo calcolare il tempo tramite l'equazione di ricorrenza, essa è

$$T(n) = \begin{cases} 5 & \text{se } n = 2 \\ T(\frac{n}{2}) + 5 & \text{altrimenti} \end{cases}$$

Non mettiamo costanti davanti $T(n)$ perché sto lavorando in parallelo. Il tempo che otteniamo, dopo aver risolto questa equazione di ricorrenza, è ancora $T(n, p(n)) = 5 \log(n)$. L'efficienza vale

$$E = \frac{n \log(n)}{\frac{n}{2} 5 \log(n)} \rightarrow C \neq 0.$$

7.3. Bitonic sort generico

Quello che abbiamo fatto per ora è ordinare sequenze bitoniche in modo efficiente. Vediamo, per completezza, il bitonic sort di Batcher applicato a qualunque sequenza.

Bitonic sort generico sequenziale

```

1:  $A_{\min} \mid A_{\max} = \text{MinMax}(A)$ 
2: if  $|A| > 2$ 
3:   BitSort( $A_{\min}$ )
4:   BitSort( $A_{\max}$ )
5:   BitMerge( $A_{\min} \cdot \text{Reverse}(A_{\max})$ )
6: return  $A$ 

```

Teorema 7.3.1: L'algoritmo di bitonic sort generico sequenziale è corretto.

Dimostrazione 7.3.1: Dimostriamolo per induzione su n .

Caso base: se $n = 2$ con MinMax scambio i due elementi se sono disordinati, poi ritorno il vettore A ordinato senza fare altro.

Passo induttivo: supponiamo vero per $n = 2^k$, mostriamo vero per $n = 2^{k+1}$

Calcolando MinMax su input di lunghezza 2^{k+1} otteniamo due sequenze di lunghezza 2^k , ma le sequenze di lunghezza 2^k bitonic sort le riesce a ordinare perfettamente per ipotesi induttiva. La chiamata finale a BitMerge avviene con una sequenza unimodale, che è al tempo stesso bitonica, ma sappiamo che BitMerge è corretto.

Quindi il vettore A è ordinato. ■

Vediamo l'implementazione parallela del bitonic sort:

- come prima, eseguiamo MinMax su input di lunghezza

$$\frac{n}{2^{i-1}},$$

partendo con $i = 1$, ottenendo di volta in volta due sequenze A_{\min} e A_{\max} ;

- al passo finale l'input è di lunghezza 2, quindi fermo le esecuzioni di MinMax, visto che le sequenze di 2 elementi sono ordinate;
- presa una coppia di sequenze A_{\min} e A_{\max} , applichiamo Reverse ad A_{\max} e passiamo le due sequenze a BitMerge.

L'algoritmo è EREW perché lavoriamo sempre su dati diversi, usando letture e scritture esclusive.

Il tempo per la prima fase, ovvero la fase di applicazione di MinMax, è come quello del bitonic merge precedente, quindi mi fermo al passo $i = \log(n)$. Nella seconda fase ho sempre $\log(n)$ passi da moltiplicare per il costo del bitonic merge, che è $\log(n)$, quindi il costo è $\log^2(n)$.

Il tempo totale è quindi $T(n, p(n)) = \log^2(n)$. Per quanto riguarda i processori, ne abbiamo sempre $p(n) = \frac{n}{2}$, utilizzati per intero ad ogni passo dell'algoritmo.

Per il tempo possiamo usare anche l'equazione di ricorrenza, che è

$$T(n) = \begin{cases} 5 & \text{se } n = 2 \\ T(\frac{n}{2}) + \underbrace{5}_{\text{MinMax}} + \underbrace{4}_{\text{Reverse}} + \underbrace{5 \log(n)}_{\text{BitMerge}} & \text{altrimenti} \end{cases}$$

ancora senza la costante sul $T(\frac{n}{2})$ perché sono in un ambiente parallelo.

L'efficienza per questo algoritmo vale

$$E = \frac{n \log(n)}{\frac{n}{2} 5 \log^2(n)} = \frac{C}{\log(n)} \rightarrow 0$$

molto lentamente. Per questa sua proprietà, si preferisce usarlo su istanze molto piccole.

7.4. Osservazioni

Un buon algoritmo sequenziale non implica un buon algoritmo parallelo: un esempio è il merge sort, come abbiamo visto durante il tentativo di parallelizzazione. Ma anche un buon algoritmo parallelo non implica un buon algoritmo sequenziale: un esempio è il bitonic sort. Vediamo perché.

Per la routine di bitonic merge, il tempo è definito dall'equazione di ricorrenza

$$T_m(n) = \begin{cases} O(1) & \text{se } n = 2 \\ 2T_m(\frac{n}{2}) + O(n) & \text{se } n > 2 \end{cases}$$

che, una volta risolta, ci dà tempo $T_b(n) = O(n \log(n))$.

Per la routine di bitonic sort, il tempo è definito dall'equazione di ricorrenza

$$T_s(n) = \begin{cases} O(1) & \text{se } n = 2 \\ 2T_s(\frac{n}{2}) + O(n \log(n)) & \text{se } n > 2 \end{cases}$$

che, una volta risolta, ci dà tempo $T_s(n) = O(n \log^2(n))$.

Vediamo come l'algoritmo parallelo sia molto buono perché ha tempo $T(n, p(n)) = O(\log^2(n))$ mentre l'algoritmo sequenziale ha tempo $T(n) = O(n \log^2(n))$, che è peggio del merge sort.

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

8. Navigazione

8.1. Cicli euleriani

Vediamo un po' di basi di teoria dei grafi.

Un grafo diretto D è una coppia V, E dove $E \subseteq V^2$. Indichiamo un arco con $(v, e) \in E$. Un cammino è una sequenza di archi e_1, \dots, e_k tale che per ogni coppia di lati consecutivi il nodo pozzo del primo coincide con il nodo sorgente del secondo. Un ciclo è un cammino tale che il nodo pozzo di e_k è il nodo sorgente di e_1 .

Un ciclo è euleriano quando ogni arco in E compare una e una sola volta. Un cammino euleriano è la stessa cosa. Un grafo è euleriano se contiene un ciclo euleriano.

Il problema base è, dato un grafo D , è euleriano?

Diamo notazioni:

- $\forall v \in V$ definiamo $\rho^-(v) = |\{(w, v) \in E\}|$ numero di archi entranti in v ed è detto grado di entrata di v ;
- $\forall v \in V$ definiamo $\rho^+(v) = |\{(v, w) \in E\}|$ numero di archi uscenti da v ed è detto grado di uscita di v

Teorema 8.1.1 (di Eulero (1736)): Un grafo D è euleriano se e solo se

$$\forall v \in V \quad \rho^-(v) = \rho^+(v).$$

Vediamo un problema simile: molto simile al ciclo hamiltoniano, ovvero un ciclo è hamiltoniano se e solo se è un ciclo dove ogni vertice in V compare una e una sola volta. D è hamiltoniano se e solo se contiene un ciclo hamiltoniano.

Per euleriano ho un algoritmo efficiente in $O(n^3)$ con $n = |V|$ mentre per hamiltoniano ho un problema NP completo.

Abbiamo una tecnica del ciclo euleriano: viene usata per costruire algoritmi paralleli efficienti che gestiscono strutture dinamiche come alberi binari.

Per trasformare un albero in una tabella con righe nodi e colonne figlio sx dx e il padre etichetta i nodi e poi popola.

Molti problemi ben noti usano alberi binari:

- ricerca;
- costruzione di dizionari;
- query.

Fondamentale in questi problemi è la navigazione dell'albero (ricerca, manutenzione, modifica, cancellazione, inserimento).

Possiamo operare su queste strutture in parallelo con algoritmi efficienti.

Idea: usiamo delle liste che contengono dei puntatori ai nodi dell'albero, e le possiamo usare bene in parallelo (ad esempio Kogge-Stone per le somme prefisse).

Usiamo un vettore S dei successori dell'albero, gli elementi sono i nodi dell'albero.

Associamo ad un albero binario un ciclo euleriano: sostituisco ogni ramo dell'albero con un doppio arco orientato. In questo modo navigo l'albero seguendo il ciclo euleriano.

Così abbiamo un ciclo, noi vogliamo un cammino. Ogni vertice v viene espanso in tre vertici $(v, s), (v, c), (v, d)$ sinistra centro destra. Con questi nuovi vertici creo un cammino: quando devo scendere di altezza collego al nodo s , se non posso scendere scorro tutti, se devo salire mi collego a c o a d (in ordine).

Terzo e ultimo passo è costruire una lista dal cammino euleriano. Quindi avrò $S((v, x))$ con $1 \leq v \leq n$ e $x \in \{s, c, d\}$. Costruiamo questa lista a partire dalla tabella con delle regole per nodi foglia o nodi interni.

Se sono in un nodo foglia v allora

$$\begin{aligned} S[(v, s)] &= (v, c) \\ S[(v, c)] &= (v, d) \\ S[(v, d)] &= \begin{cases} (\text{pad}(v), c) & \text{se } v = \text{sin}(\text{pad}(v)) \\ (\text{pad}(v), d) & \text{se } v = \text{des}(\text{pad}(v)) \end{cases} \end{aligned}$$

Se sono in un nodo interno v allora

$$\begin{aligned} S[(v, s)] &= (\text{sin}(v), s) \\ S[(v, c)] &= (\text{des}(v), s) \\ S[(v, d)] &= \begin{cases} (\text{pad}(v), c) & \text{se } v = \text{sin}(\text{pad}(v)) \\ (\text{pad}(v), d) & \text{se } v = \text{des}(\text{pad}(v)) \end{cases} \end{aligned}$$

Ultima regola è uguale per tutti.

Diamo un algoritmo parallelo per costruire S :

- un processore per ogni vertice, ovvero per ogni riga della tabella;
- il processore deve costruire $S[v, \dots]$
- le letture sono concorrenti, ho accesso a tutta la riga di v ma anche alle righe dei padri. Possiamo eliminare la concorrenza, ci fidiamo. Ad esempio, si fa con i nodi pari/dispari con piccoli accorgimenti per leggere solo dalla propria parte.

Algoritmo EREW con $p(n) = n$ e $T(n, p(n)) = O(1)$. Cambiamo con Wyllie e otteniamo $p(n) = \frac{n}{\log(n)}$ e $T = \log(n)$.

L'array S è utile per risolvere i problemi:

- attraversamento in pre-ordine;
- calcolare la profondità dei nodi.

Abbiamo bisogno di due definizioni:

- $\forall v \in V$ allora $N(v)$ indica l'ordine di attraversamento di v in pre-ordine;
- $\forall v \in V$ allora $P(v)$ indica la profondità di v nell'albero.

La radice ha $N(v) = 1$ mentre la foglia più a destra ha $N(v) = n$. La radice ha $P(v) = 1/0$, il figlio della radice uno in più.

8.2. Attraversamento in pre-ordine

Dai una definizione di pre-ordine (prima radice, poi sx, poi dx).

Definiamo un array A tale che

$$A[(v, x)] = \begin{cases} 1 & \text{se } x = s \\ 0 & \text{altrimenti} \end{cases} \quad \forall v \in V.$$

Ora, su (A, S) andiamo ad applicare somme prefisse. Dentro la cella $A[(v, s)]$ avremo $N(v)$ perché quando facciamo il cammino e visitiamo un nuovo nodo andiamo sempre nel suo nodo sinistro.

L'algoritmo calcola A e S , calcola somme prefisse su A e S . L'output è nel nodo $A[(v, s)]$.

L'algoritmo è EREW con $p(n) = \frac{n}{\log(n)}$ e $T(n, p(n)) = \log(n)$ per entrambi i passi, quindi ottengo

$$E = \frac{n}{\frac{n}{\log(n)} \log(n)} \rightarrow C \neq 0$$

ottimale si gode.

Per la profondità dei nodi ci serve un array tale che

$$A[(v, x)] = \begin{cases} 1 & x = s \\ 0 & x = c \\ -1 & x = d \end{cases}.$$

Anche su questo vettore applichiamo le somme prefisse. Troviamo $P(v)$ nella cella $A[(v, d)]$.

L'algoritmo parallelo per la profondità calcola A e S , calcolo le somme prefisse su (A, S) , l'output è in $A[(v, s)]$ se partiamo da 1, altrimenti in $A[(v, d)]$ se partiamo da 0.

Abbiamo efficienza

$$E = \frac{n}{\frac{n}{\log(n)} \log(n)} \rightarrow C \neq 0$$

quindi anche lui ottimale efficiente.

Parte II – Algoritmi paralleli a memoria distribuita

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

1. Introduzione

Osservazioni finali sulle PRAM:

- interesse teorico
 - processori sono uguali e alla pari
 - il tempo è strettamente legato alla computazione (comunicazione costante)
- interesse pratico
 - realizzazione fisica dei multicore

Multicore ha portato l'interesse del calcolo parallelo da ambiti scientifici ad un ambiente più ampio, tipo consumatore o informatico.

Prima del 2000 per aumentare le prestazioni si aumentava il clock con problemi:

- di assorbimento di energia (> 100W)
- di raffreddamento

Dopo il 2000 arrivano i multicore, si aumenta il grado di parallelismo con:

- clock di minor frequenza
- minor assorbimento di energia
- vantaggi sul raffreddamento

Questo porta allo sviluppo teorico in ambito di algoritmi paralleli (scrittura, riscrittura, manipolazione di software per i multicore).

1.1. Architetture parallele a memoria distribuita

Architetture parallele a memoria distribuita erano i paradigmi usati prima del multicore, usato dai supercomputer (anni 60 cray e intel paragon, mentre attuali cray, blue gene, red storm, earth simulator, tianhe-2)

Sono supercomputer a memoria distribuita, ovvero sono grafi con nodi processori e archi reti di connessioni. Alle PRAM manca la memoria condivisa.

I processori sono RAM sequenziali con:

- elementi di calcolo, hanno istruzioni per il calcolo e la loro memoria privata
- router, hanno istruzioni per la comunicazione di send e receive

La comunicazione avviene in parallelo, ma se p_1, \dots, p_k mandano contemporaneamente dati a p essi sono fatti in modo simultaneo, ma p lavora sequenzialmente quindi deve fare k receive, quindi servono $k + 1$ passi per la comunicazione (send parallela e k receive).

I collegamenti sono di tipo full-duplex, ovvero comunicazione diretta, archi non orientati. Se c'è collegamento diretto la comunicazione costa 2 passi (send e receive).

Abbiamo anche un clock centrale che scandisce il tempo per tutti i processori.

Il programma, come nelle PRAM, è un PAR DO, quindi

```

for k in I par do
    istruzione k

```

con anche send e receive (architettura SIMD single instruction multiple data)

Cambiano input e output: non abbiamo più la memoria condivisa come la PRAM, quindi l'input viene distribuito tra i processori, mentre l'output o viene messo in un processore dedicato o si legge in un certo ordine tra i vari processori.

Le risorse di calcolo sono:

- numero di processori: può essere la lunghezza dell'input ma ci sono tecniche per abbassare il numero
- tempo, dato da:
 - tempo di calcolo
 - tempo di comunicazione, può essere rilevante ed è legato alla rete di connessioni

Abbiamo i seguenti parametri di rete: data l'architettura $G = (V, E)$ definiamo:

- grado di G : per ogni vertice calcoliamo

$$\gamma = \max\{\rho(v) \mid v \in V\}$$

dove $\rho(v)$ è il numero di archi incidenti su v ; un valore alto permette buone comunicazioni ma rende più difficile la realizzazione fisica

- diametro di G : definiamo

$$\delta = \max\{d(v, w) \mid v, w \in V \wedge v \neq w\}$$

come il massimo tra tutte le distanze minime da v e w ; valori bassi di δ sono da preferire, ma aumentano il parametro γ

- ampiezza di bisezione di G : sia β il minimo numero di archi in G che tolti mi dividono i nodi in circa due metà; esso rappresenta la capacità di trasferire le informazioni in G , ancora una volta β alto si preferisce ma incrementa γ

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

2. Max e ordinamento

Problemi che vedremo sulle architetture parallela a memoria distribuita:

- max: comunicazione a coppie di processori δ basso comunicazione veloce
- ordinamento: spostamenti di parti dell'input β alto ordinamento efficiente

Valgono i seguenti limiti inferiori per i tempi di soluzione

Lemma 2.1: Il tempo richiesto per MAX in G è almeno δ

Dimostrazione 2.1: Ogni coppia di processori deve comunicazione, quindi servono almeno δ passo parallelo ■

Lemma 2.2: Il tempo richiesto per ORDINAMENTO in G è almeno

$$\frac{n}{2} \frac{1}{\beta}$$

Dimostrazione 2.2: Divido il grafo in due metà:

- in $\frac{n}{2}$ ho i numeri più alti
- in $\frac{n}{2}$ ho i numeri più bassi

Noi vogliamo crescente, nel caso peggiore sono tutto decrescente

Quanti trasferimenti devo fare? Posso trasferire da una nuvola all'altra in β , quindi facendo $\frac{n}{2}$ ci metto $\frac{n}{2} \frac{1}{\beta}$. Perché ho a disposizione β ponti. ■

Per analizzare questi problemi abbiamo bisogno dei confrontatori/comparatori e delle loro primitive.

Sono dei ponti che collegano due fili, sopra mettono il minimo e sotto il massimo dei due valori

if $A[i] > A[j]$ then SWAP($A[i]$, $A[j]$) con $i < j$

Ci sono alcuni che fanno il contrario, quindi max sopra min sotto e il confronto ha il minore. Se minimo sopra cerchio, se minimo sotto cerchio vuoto.

Possiamo creare reti di confrontatori, ovvero metto un input per ogni filo e metto una serie di confrontatori. Una sorting network è una rete di confrontatori che ordina

Metti esempio per 3 e 4 elementi

Una rete generale di ordinamento è data dal bubble sort, ovvero ad ogni fase l'elemento più pesante viene spinto verso il basso.

Idea parallela: i confrontatori che agiscono su fili diversi vengono messi in un passo parallelo. In quello da 4 elementi ho $T(n) = \#step$ e $p(n) = 4$.

Una rete di confrontatori la indichiamo con $R(x_1, \dots, x_n) = (y_1, \dots, y_n)$

Si dice che R è una sorting network se e solo se

$$\forall (x_1, \dots, x_n) \in \mathbb{N}^n \quad R(x_1, \dots, x_n) = (y_1, \dots, y_n)$$

con

$$y_1 < \dots < y_n$$

Sono anche dette reti di ordinamento test/swap oblivious (confronti non dipendono dall'input ma sono fissati a priori)

Ci chiediamo se R sia una sorting network. Per saperlo usiamo il principio 0 – 1 (zero uno, knuth nel 1972)

Formalmente,

$$\forall x \in \{0, 1\}^n \quad R(x) \text{ ordinato} \longrightarrow \forall y \in \mathbb{N}^n \quad R(y) \text{ ordinato}$$

Se vale un booleano vale anche per tutti gli altro

Vale anche esiste + non ordinato implica esiste + non ordinato

Introduciamo uno strumento: f-shift su R

Abbiamo una rete R e una funzione f . Prima applico f e poi R , oppure prima R e poi f , ovvero

$$R(f(x_1), \dots, f(x_n)) = f(R(x_1, \dots, x_n)) = (f(y_1), \dots, f(y_n))$$

Per essere vero f deve essere monotona crescente, perché i confrontatori trovano minimo e massimo e f non cambia niente

Teorema 2.1: Se R non corretta esiste $x \in \mathbb{N}^n$ tale che $R(x)$ non ordina, quindi esistono k, s tali che $y_k > y_s$ ma $k < s$

Dimostrazione 2.3: Definiamo $g = \mathbb{N} \longrightarrow \{0, 1\} = \begin{cases} 1 & \text{se } x \geq y_k \\ 0 & \text{altrimenti} \end{cases}$

Vediamo come sia monotona crescente

Ora applico f a R e ottengo

$$R(g(x_1), \dots, g(x_n))$$

ma per la regola dello shift questa cosa è uguale a

$$(g(y_1), \dots, g(y_n))$$

vettore binario che non è ordinato perché $g(y_k)$ vale 1 mentre in $g(y_s)$ vale 0 perché più piccolo.

Ma allora non ho ordinato



Per testare una R e capire se è sorting network mi basta valutare R solo su input binari, molto più facile da fare

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

3. Array lineari

Architettura parallela a memoria distribuita. Ci sono n processori p_1, \dots, p_n collegati su una riga. I parametri di questa rete sono:

- $\gamma = 2$ grado del grafo (ottimo per la realizzazione)
- $\delta = n - 1$ diametro (lower bound per max, non sono soddisfatto perché sequenziale ci mette n)
- $\beta = 1$ ampiezza di bisezione (lower bound per ordinamento)

Sulla PRAM abbiamo max che usa $p = \frac{n}{\log(n)}$ e $T = \log(n)$ mentre per ordinamento abbiamo $p = n$ e $T = \log(n)$ (per Cole) quindi dobbiamo abbassare i processori.

Problemi che affronteremo sono shuffle, max e ordinamento

Abbiamo una serie di processori P_n che tengono i dati $A[n]$. Se servono dati di altri li chiedono.

Vediamo la primitiva per shuffle.

Per risolvere shuffle facciamo swap contiguo, ovvero $\text{swap}(k, k+1)$. Ovvero devo avere P_k con $A[k+1]$ e P_{k+1} con $A[k]$

Avrò bisogno di 3 passi paralleli, ovvero doppia send per mandare, doppia receive per ricevere e poi assegnare $A[k]=A[k+1]$ e $A[k+1]=A[k]$

Vediamo shuffle

In input ho $A[1], \dots, A[s], A[s+1], \dots, A[2s]$ di lunghezza pari, e in output voglio $A[1], A[s+1], A[2], A[s+2], \dots, A[s], A[2s]$ ovvero voglio primo elemento prima metà + primo seconda metà, poi secondo elemento, poi terzo, eccetera.

Idea: scambio a metà, poi quelli vicini, poi vicini ancora, eccetera. Creo un albero di swap contigui

Il numero di processori è $2(s-1)$ e il tempo parallelo è $3(s-1)$. Un algoritmo sequenziale per questo è $\Theta(s^2)$ (senza memoria, sennò n) quindi

$$E = \frac{s^2}{ss} = C \neq 0$$

quindi mega ottimale cazzo

Vediamo massimo MAX

Dobbiamo mandare dati ai vari processori. Se devo mandare da i a j devo fare $\text{SEND}(i, j)$, lo faccio in un numero di passi $d(i, j)$, entra nel costo del problema. Quindi faccio send, poi receive-send, poi solo receive in j .

In totale sono $2d(i, j) = 2 \text{abs}(i-j)$

Vediamo come la trasmissione non è più costante come nelle PRAM

In input ho $A[1], \dots, A[n]$ e vogliamo in P_n il valore

$$\max\{A[i] \mid 1 \leq i \leq n\}$$

Il tempo per MAX su array lineari è limitato inferiormente da n , mentre sulle PRAM era $\log(n)$. Il sequenziale invece n .

Idea:

- si considera l'algoritmo per sommatoria delle PRAM
- riduciamo i processori per abbassare $\Theta(n)$ su array a n processori

Al j -esimo passo confrontiamo i numeri a distanza 2^{j-1} , selezioniamo il massimo e lo memorizziamo nel processore $2^j t$

Il numero di passi è $\log(n)$. Confronto ogni volta $2_j t - 2^{j-1}$

Vediamo il codice

```

for  $j = 1$  to  $\log(n)$ 
  for  $k \in \{2^j t - 2^{j-1} \mid 1 \leq t \leq \frac{n}{2^j}\}$  par do
    SEND( $k, k + 2^{j-1}$ )
  for  $k \in \{2^j t \mid 1 \leq t \leq \frac{n}{2^j}\}$  par do
    if ( $A[k] < A[k - 2^{j-1}]$ ) then
       $A[k] = A[k - 2^{j-1}]$ 

```

Fase di send e fase di compare

Vediamo il tempo:

- send è due volte la distanza tra i processori, quindi ho $2 \cdot 2^{j-1} = 2^j$ per $j = 1, \dots, \log(n)$
- compare vale 2, solo confronto e assegnamento, ma sempre per $j = 1, \dots, \log(n)$

Quindi il totale è

$$\sum_{j=1}^{\log(n)} 2^j + 2 = \sum_{i=1}^{\log(n)} 2^i + 2 \log(n) = 2^{\log(n)+1} - 1 - \underbrace{1}_{\text{parto da 0}} + 2 \log(n) = 2n - 2 + 2 \log(n) = O(n).$$

Ho n processori, quindi $E \rightarrow 0$ non va bene

Riduciamo i processori da n a p in questo modo operiamo sul parametro δ distanza massima tra i processori. Ogni processore prende $\frac{n}{p}$ elementi e non più uno. Ora ho array di p elementi e quindi cambia anche δ che si riduce.

Nuovo algoritmo:

- un processore seleziona il max sequenziale tra i suoi $\frac{n}{p}$ numeri
- si esegue MAX su p processori

Prestazioni:

- processori p
- tempo $O\left(\frac{n}{p}\right) + O(p)$

Efficienza ho numeratore n e denominatore

$$p \left(O\left(\frac{n}{p}\right) + O(p) \right) = O(n) + O(p^2) = O(n)$$

Vorremmo $p^2 = n$ per avere $O(n)$ e semplificare, quindi

$$p = \sqrt{n}$$

per avere efficienza non nulla

Quindi MAX lo risolviamo bene su array lineari con $p = \sqrt{n}$ e tempo $T = \left(O\left(\frac{n}{\sqrt{n}}\right)\right) + O(\sqrt{n}) = O(\sqrt{n})$

Vediamo invece ordinamento

Abbiamo bisogno di swap tra due processori generici, non più contigui. Scambiamo P_i e P_j , abbiamo diverse soluzioni

USO LA SEND, quindi devo fare una send i,j POI una send j,i , poi alla fine di tutto assegnamento

Il tempo è $2d(i,j) + 2d(j,i) + 1 = 4d(i,j) + 1$ ma tempo brutto

USO LA SEND SIMULTANEA abbiamo due casi:

- distanza tra i processori dispari $2k+1$ ovvero i processori sono pari $2k+2$, divisi a metà $k+1$ processori. Quelli in mezzo sono $k+1$ e $k+1+1$. I dati dai bordi arrivano simultanei con $\text{send}(i, k+1)$ e $\text{send}(j, k+1+1)$, poi faccio send simultanea (ho un full duplex) e poi le send opposte
- distanza tra i processori pari: come esercizio

Il tempo del primo caso è $2k$ (distanza bordi) + $2 + 2k$ (send di nuovo) + 1 (assegnamento) = $4k+3 = 2(2k+1) + 1 = 2d(i,j) + 1$

Il tempo nel secondo caso è $2d(i,j) + 3$

Altra primitiva per l'ordinamento che ci serve è minmax, ovvero processore di indice minimo metto valore minimo, processore di indice massimo metto valore massimo

In poche parole implementa il confrontatore, ovvero $P_k = \min\{A[k], A[k+1]\}$ e altro bla bla bla

La minmax ha 4 come costo del passo parallelo.

Ordinamento ha in input $A[1], \dots, A[n]$ assegnati ai processori P_1, \dots, P_n e in output voglio $\text{cont}(P_1) < \dots < \text{cont}(P_n)$

Diamo un algoritmo test/swap oblivious descritto da una SN, si chiama ODD/EVEN sorting network.

Abbiamo una colonna di confrontatori dispari (sopra è su un dispari) seguita una colonna di confrontatori pari. Una colonna si chiama round perché è un passo parallelo. Alterno per n volte questi confrontatori.

Per la correttezza di ODD/EVEN usiamo il principio 0/1, ovvero

$$\{0, 1\}^n \rightsquigarrow \text{ODD/EVEN} \rightsquigarrow 0^j 1^e \quad | \quad j + e = n$$

a patto di fare esattamente n round.

Ogni 1 deve scendere di $n - e = j$ posizioni. Contiamo gli uni a partire da 1 a partire dal basso. Vediamo quanto ci mettono a scendere di j posizioni. Notiamo che i ha i ritardi e poi scende di j posizioni. Quindi il primo uno che abbiamo ci mette $j + i$, ma i è esattamente il numero di 1 dell'input, è il ritardo.

Regola: l' i -esimo uno dal basso impiega $n - e + i$ passi per posizionarsi correttamente. In generale è un al più.

Dato che $i \leq e$ si ottiene

$$n - e + \underbrace{e}_{i \text{ di prima}} = n.$$

Quindi n passi sono necessari (visto dall'input) e sufficienti.

Implementazione sequenziale è $n \frac{n}{2} \approx n^2$

In parallelo usiamo Haberman del 1972

Abbiamo n passi paralleli o round di confrontatori paralleli con $\text{minmax}(k, k+1)$

```
1: for  $i = 1$  to  $n$ 
2:   for  $k \in \{2t - (i \% 2) \mid 1 \leq t \leq \frac{n}{2}\}$  par do
3:     minmax( $k, k+1$ )
```

Il tempo è $4n = O(n)$ mentre efficienza è

$$\frac{n \log(n)}{nn} \rightarrow 0$$

non va bene riduco i processori

Osservazione: per n processori in un array lineare l'ordinamento vuole tempo $\Omega\left(\frac{n}{2^\beta}\right)$ e qui $\beta = 1$ quindi ci mettevo n , non va bene

Riduco i processori a p quindi ognuno prende $\frac{n}{p}$ dati e li ordina in sequenza in un tempo $O\left(\frac{n}{p} \log\left(\frac{n}{p}\right)\right)$

L'algoritmo è una versione che non usa minmax ma merge-split: avviene tra due processori contigui. Essa fa:

- processore di sinistra spedisce $\frac{n}{p}$ dati ordinati al processore di destra in un tempo $O\left(\frac{n}{p}\right)$
- il processore di destra riceve e fonde i nuovi $\frac{n}{p}$ dati con i suoi $\frac{n}{p}$ ordinati (operazione di merge) in tempo $O\left(\frac{n}{p}\right)$
- il processore di destra invia i dati più piccoli $\frac{n}{p}$ al processore di sinistra in un tempo $O\left(\frac{n}{p}\right)$ (operazione di split)

```
1: for  $i = 1$  to  $p$ 
2:   for  $k \in \{2t - (i \% 2) \mid 1 \leq t \leq \frac{p}{2}\}$  par do
3:     merge-split( $k, k+1$ )
```

Ogni primitiva di merge-split viene ripetuta p volte

Il tempo parallelo è $\frac{n}{p} \log\left(\frac{n}{p}\right) + p \frac{n}{p} = O(n)$. Il denominatore è $p\left(\frac{n}{p} \log\left(\frac{n}{p}\right) + n\right) = n \log\left(\frac{n}{p}\right) + np$ e noi vogliamo che sia $n \log(n)$ per avere una frazione uguale. Quindi prendiamo $p = \log(n)$. Efficienza quindi $E \rightarrow C \neq 0$

Osserviamo come il tempo sia rimasto $O(n)$ ma abbiamo abbassato il tempo parallelo. Ciò si spiega in quanto la riduzione dei processori agisce sul diametro e non sull'ampiezza di bisezione, che rimane sempre $\beta = 1$.

Possiamo migliorare questi costi

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

4. Architettura MESH

Usate dai supercomputer, array bidimensionale ovvero griglia di processori

Avendo n processori abbiamo un quadrato $m \times m$ con $m = \sqrt{n}$ disposti come una matrice. Se n non è un quadrato perfetto prendiamo

$$(\lfloor \sqrt{n} \rfloor + 1)^2 \leq 2n$$

per $n \geq 6$ quindi va bene lo stesso

Parametri di rete:

- $\gamma = 4$ (secondo lei è ρ)
- $\delta = 2\sqrt{n}$ perché devo fare la scala
- $\beta \sim \sqrt{n}$ (due diversi tagli)

I nostri lower bound diventano $\Omega(\sqrt{n})$ per max e $\Omega(\sqrt{n})$ per l'ordinamento, quindi tanta roba

Vediamo il massimo: idea immediata se usiamo la mesh come array lineare di n processori, facendo tipo serpentello, ma si ottiene $\Omega(n)$ per il tempo contro $\Omega(\sqrt{n})$ che abbiamo adesso in una mesh

Pensiamo ad un algoritmo righe-colonna, ovvero non penso al serpentello ma anche ad alte connessioni. Ogni riga è un array lineare di \sqrt{n} processori, in più anche l'ultima colonna è un array lineare di \sqrt{n} processori. Sposto i massimi di ogni riga in fondo e poi max dei massimi.

```

1:  $m = \sqrt{n}$ 
2: for  $i = 1$  to  $m$  par do
3:   MAX(Pi1, Pi2, dots, Pim)
   in Pim ho il massimo ora
4: MAX(P1m, P2m, dots, Pmm)
```

Tempo per la parte 1 è \sqrt{n} , idem per la seconda parte, quindi eseguo in tempo $O(\sqrt{n})$ quindi siamo soddisfatti. Efficienza purtroppo è

$$\frac{n}{n\sqrt{n}} \rightarrow 0$$

quindi riduciamo i processori

Passiamo da n a p , ma andiamo avanti prossima volta

Riprendiamo ancora la mesh

Riduciamo i processori da n a p , ogni processore calcola il max tra $\frac{n}{p}$ dati in tempo $O(\frac{n}{p})$. Poi, si attiva l'algoritmo di prima sulla griglia $\sqrt{p} \times \sqrt{p}$, che viene eseguito però in $O(\sqrt{p})$

Il tempo totale è $T = \frac{n}{p} + \sqrt{p}$

Il denominatore invece è

$$p \left(\frac{n}{p} + \sqrt{p} \right) = n + p^{\frac{3}{2}}$$

Vogliamo un denominatore n per avere una costante per l'efficienza, quindi scelgo $p^{\frac{3}{2}} = n$ ovvero

$$p = n^{\frac{2}{3}}$$

Il tempo totale diventa quindi

$$\frac{n}{p} + \sqrt{p} = n^{1-\frac{2}{3}} + \sqrt{n^{\frac{2}{3}}} = n^{\frac{1}{3}} + n^{\frac{1}{2} \cdot \frac{2}{3}} = \sqrt[3]{x} + \sqrt[3]{x} = O(\sqrt[3]{x})$$

Otteniamo efficienza ok

Limite teorico dobbiamo rifare i conti: con \sqrt{p} il limite è $\Omega(\sqrt{p})$ quindi $\Omega(\sqrt[3]{n})$ ma noi siamo esattamente qua quindi tanta tanta roba

4.1. Ordinamento LS3

Ricercatori anni 80

Ricorsivo con divide et impera. Non ordiniamo più a serpente.

Sia M il quadrato dei processori:

- dividi: divido in 4 quadrati M_1, M_2, M_3, M_4 sx alto dx alto sx basso dx basso (italiano) di dimensione $\frac{m}{2}$ con $m = \sqrt{n}$
- ordina: faccio l'ordinamento a serpente in parallelo
- fondi: mi arrivano 4 matrici, le fondo dando la matrice totalmente ordinata

Vediamo LS3 parallelo

LS3sort

```

1: if  $|M| == 1$ 
2:   return  $M$ 
3: LS3sort( $M_1$ )
4: LS3sort( $M_2$ )
5: LS3sort( $M_3$ )
6: LS3sort( $M_4$ )
7: LS3merge(...)
```

Ci serve vedere anche la merge

Ci servono shuffle e odd/even. Data una riga i della mesh, essa è un array lineare di processori. Su questa riga facciamo lo shuffle. Lo shuffle viene fatto su M , formato dalle matrici M_k ordinate. Quindi lo shuffle viene fatto sulla riga di M .

Lo shuffle intramezzava prima metà con la seconda metà, cioè metteva vicini gli elementi con lo stesso indice delle due righe della matrice.

Dopo lo shuffle devo eseguire ODD-EVEN tra due colonne adiacenti i e $i+1$. Vogliamo vederlo come array lineare: lo costruiamo a serpente. Su questo facciamo ODD-EVEN quindi ordiniamo, sono $2\sqrt{n}$ e quindi avremo quel numero di round

LS3merge

```
1: for  $i = 1$  to  $\sqrt{n}$  par do
2:   L SHUFFLE(i)
3:   for  $i = 1$  to  $\frac{\sqrt{n}}{2}$  par do
4:     L ODD-EVEN(2i-1, 2i)
5:   esegui i primi  $2\sqrt{n}$  passi di ODD-EVEN sull'intera mesh a serpente
```

Il tempo per sta roba è $O(\sqrt{n})$ per lo shuffle, $O(\sqrt{n})$ per ODD-EVEN, e ancora $O(\sqrt{n})$ per l'ultima esecuzione, quindi il tempo per la merge è $T_m(n) = h\sqrt{n}$

Risolviamo l'equazione di ricorrenza per sto schifo:

$$T(n) = \begin{cases} 1 & \text{se } n = 1 \\ T(\frac{n}{4}) + h\sqrt{n} & \text{altrimenti} \end{cases}$$

Ma allora

$$\sqrt{\frac{n}{4^2}} + h\sqrt{\frac{n}{4}} + h\sqrt{n} = \dots = \sum_{i=0}^{\log_4(n)-1} h\sqrt{\frac{n}{4^i}} + 1 = h\sqrt{n} \sum_{i=0}^{\log_4(n)-1} \sqrt{\frac{1}{4^i}} + 1 = h\sqrt{n} \sum_{i=0}^{\log_2 \frac{n}{\log_4(n)}-1} \frac{1}{2^i} + 1 = h\sqrt{n} \sum_{i=0}^{\frac{\log(n)}{2}-1} \left(\frac{1}{2}\right)^i + 1 = h$$

Se aggiungo un 4 ci esce $n\sqrt{n}$ per il sequenziale (peggio del merge sort)

Processori sono $p(n) = n$ e il tempo è $T(n) = O(\sqrt{n})$ quindi

$$E = \frac{n \log(n)}{n\sqrt{n}} \rightarrow 0$$

che non ci piace

Possiamo migliorare riducendo i processori, ma non lo vedremo

Con una versione del bitonic sort su mesh usa processori $O(\log^2(n))$ e tempo $T(n) = O\left(\frac{n}{\log(n)}\right)$ che è efficiente, ma come tempo è peggiore di LS3.

Parte III – Algoritmi distribuiti

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

1. Introduzione

Abbiamo un grafo orientato, dove i nodi sono entità e le frecce sono link/conessioni (non per forza full-duplex). Non abbiamo un clock globale. Ogni entità possiede:

- memoria locale
- capacità locale
- capacità di comunicazione
- clock locale proprio

Entità sono processori, processi, sensori, switch, eccetera

Nella memoria locale:

- registro di input, il registro è $\text{valore}(x) = \text{input dell'entità } x$
- registro di stato, il registro è $\text{stato}(x) = \text{stato dell'entità } x$, ovvero è il valore attuale dell'entità ed è cambiata localmente dalla stessa x

Per il clock locale, è possibile settare o resettare una sveglia

Proprietà delle entità:

- sono reattive: all'accadere di un evento compiono una azione
 - Eventi:
 - interni al sistema: ricezione di messaggi, sveglia
 - esterni al sistema: impulso spontaneo (START)
 - AZIONI:
 - sequenza finita di operazioni indivisibili (inizio l'azione e la porto a termine, non si blocca, un esempio è nil)
- seguono delle regole: una regola è un oggetto della forma $\text{stato } x \text{ evento} \rightarrow \text{azione}$. Sia x una entità, definiamo $B(x)$ l'insieme delle regole a cui è soggetta x . Questo insieme deve essere completo e non ambiguo, ovvero è praticamente il codice di x

Se E insieme delle entità che cooperano tra loro, allora

$$B(E) = \bigcup_{x \in E} B(x)$$

che è il comportamento del sistema, ed è importante che sia omogeneo, ovvero

$$\forall x, y \in E \quad B(x) = B(y)$$

ovvero il codice deve essere uguale per tutte le entità

Il codice è detto algoritmo distribuito o protocollo per E , ma solo se $B(E)$ omogeneo

Lemma 1.1: È sempre possibile ottenere $B(E)$ omogeneo

Dimostrazione 1.1: Idea è utilizzare un registro locale aggiuntivo che differenzia quelle entità che alla stessa coppia stati x evento hanno azioni diverse. Questo è il registro ruolo(x) = ruolo di x . La regola viene modificata in

$$\text{stato} \times \text{evento} \longrightarrow \text{if ruolo}(x) = a \text{ then } A_a \text{ else } A_b$$



Proprietà della rete:

- la comunicazione avviene usando una etichettatura sui link. Per l'entità x , l'etichettatura è denotata con λx (nome di ogni link). x si trova in G , indichiamo con:
 - $N_{in}(x)$ vicini di ingresso ad x
 - $N_{out}(x)$ vicini di uscita di x
- assiomi della rete:
 - ritardo finito di comunicazione, in assenza di errori un messaggio è spedito prima o poi arriverà
 - orientamento locale, ogni entità riesce a distinguere tra i suoi vicini N_{in} e N_{out} grazie alla conoscenza di λx

Parametri della rete:

- numero di entità: n
- numero di link: m
- diametro della rete: d (delta delle parallele)

Assioma abbiamo anche:

- restrizioni: dichiarate al momento della scrittura del codice, sono proprietà positive della rete su cui facciamo affidamento

Restrizioni sulla comunicazione:

- link bidirezionali: connessioni full-duplex, ovvero

$$\forall x \quad N_{in}(x) = N_{out}(x) \wedge \lambda_x(x, y) = \lambda_x(y, x) \implies N(x)$$

- ordinamento dei messaggi: i messaggi sullo stesso link vengono prelevati con la politica FIFO

Restrizioni sull'affidabilità:

- rilevazione di errori, a livello di entità e di link, quindi l'abbiamo avuta
- affidabilità parziale (non ci saranno errori in futuro)
- affidabilità totale: non ci sono stati errori e non ce ne saranno in futuro

Restrizioni sulla topologia di rete:

- connettività del grafo, abbiamo fortemente connesso (per ogni coppia di entità abbiamo cammino bidirezionale che le collega) oppure connesso (grafo non diretto)

Restrizioni sul tempo:

- tempi di comunicazione unitari
- clock sincronizzati

Tali restrizioni a volte vengono considerate per il calcolo della prestazioni ideali del codice distribuito

Misure di complessità:

- tempo: intervallo tra la prima entità che si attiva e l'ultima che termina

- quantità di comunicazione: numero di messaggi spediti (se sono omogenei, ovvero stessa grandezza) e numero di bit spediti

Esecuzioni diverse dello stesso codice distribuito può portare a tempi diversi, quindi il tempo non va bene

Risolviamo quindi con il tempo ideale: tempo misurato considerando comunicazioni unitarie e clock sincroni

Il tempo causale (caso peggiore) è il tempo misurato considerando la catena più lunga di comunicazione richiesta dal codice. Questo difficile da calcolare.

Definizione di un problema: esso è una tripla P_{init} , P_{final} e R , dove i primi due sono due predicati che descrivono le configurazioni del sistema all'inizio e alla fine, mentre R sono le restrizioni del sistema.

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

2. Broadcasting

Pinit una entità tiene I:

$$\exists x \in E \mid \text{valore}(x) \wedge \forall y \neq x \quad \text{valore}(y) = \emptyset$$

Pfinal: tutte le entità ce l'hanno, ovvero

$$\forall x \in E \quad \text{valore}(x) = I$$

Le restrizioni sono:

- link bidirezionali BL
- affidabilità totale TR (total reliability)
- connettività CN
- queste tre sono R
- unico iniziatore UI (parte solo il primo che contiene x, detta I sta cosa)

Tutto detto RI

Vediamo un algoritmo distribuito / protocollo

Diamo l'insieme di regole stato times evento \rightarrow azione:

- stato_t(x) stato di x al tempo t
- evento è impulso spontaneo, sveglia o messaggio
- azione è mini programma indivisibile

L'esecuzione di un protocollo genera una sequenza di configurazioni successive del sistema

Sia $\Sigma(t)$ il contenuto dei registri delle entità al tempo t , sia futuro(t) eventi già generati al tempo t ma che non sono ancora processati

Indichiamo con $C(t)$ la configurazione del sistema del tempo t , definita dalla coppia $\Sigma(t), \text{futuro}(t)$

Definiamo $C(0) = (\Sigma(0), \text{Futuro}(0))$ registri inizializzati + impulso spontaneo

L'esecuzione del protocollo è una sequenza di configurazioni successive tale che $C(0)$ a $C(f)$ con il protocollo

Quando una C soddisfa P lo scriviamo con $C \in P$

Dobbiamo definire come un protocollo risolve un problema (definita dalla tripla)

Usiamo due stati $S = \{\text{iniziatore}, \text{inattivo}\}$. Abbiamo Sinit = stati delle entità in $C(0)$ e Sterm = stati delle entità in $C(f)$

2.1. Prima versione

$S = \{\text{iniziatore}, \text{inattivo}\}$, Sinit = {iniziatore, inattivo} e Sfinal = {inattivo}

Iniziatore, se ricevo impulso spontaneo

- send(M) to N(x)
- become inattivo

Inattivo, se ricevo M

- processa M (preleva informazione e mettila in valore)
- send(M) to N(x)

Il messaggio è $M = (t, o, d, I)$ con t tipologia del mex, o e d sono origine e destinatario, I informazione

Abbiamo un problema

Problema dell'altra volta: protocollo corretto ma non termina, perché dopo lo stato iniziale tutto diventa inattivo e gli inattivi ricevono, processano e mandano ancora in giro. Non funziona anche se modifichiamo imponendo di non mandare a chi me l'ha mandato.

Infatti Futuro(t) non è mai il vuoto, la computazione non termina

Modifichiamo gli stati: Sstart subset Sinit stati che fanno iniziare il protocollo, mentre Sfinal subset Sterm stati per cui la sola azione possibile è quella nulla

Abbiamo ora Sinit = {iniziatore, inattivo}, Sstart = {iniziatore}, Sterm e Sfinal = {inattivo}

Grazie a questo stato il protocollo termina

La soluzione per P è tale che

$$\forall C(0) \in \text{Pinit} \quad \exists t' \mid \forall t > t' \quad C(t) \in \text{Pfinal} \text{ (correttezza)}$$

e anche

$$\forall x \in E \quad \text{stato}_t(x) \in \text{Sfinal} \text{ (terminazione)}$$

2.2. Seconda versione [flooding]

Abbiamo $S = \{\text{iniziatore, inattivo, finito}\}$, Sstart = {iniziatore}, Sfinal = {finito}

Abbiamo iniziatore con impulso spontaneo

- send(M) to N(x)
- become finito

Abbiamo inattivo con ricezione di M

- processa M
- send M to N(x)-sender
- become finito

Le coppie stato x evento non indicano fanno nil

Vediamo la complessità:

- $M[F] = \sum_{x \in E} (N(x) - 1) + \underbrace{1}_{\text{iniziatore}} = 2m - n + 1$
- $T[F] \leq d$ diametro della rete

Abbiamo anche lower bound:

- $T[\text{broadcast} / \text{RI}] \geq d$, tempo causale (caso peggiore)
- $M[\text{broadcast} / \text{RI}] \geq m$ per un teorema

Il protocollo è ottimale

Teorema 2.2.1: $M[\text{broadcast} / \text{RI}] \geq m$

Dimostrazione 2.2.1: Per assurdo, risolvo il problema con meno di m messaggi. Sia A il protocollo che non manda messaggi su (x, y) . A è corretto e deve lavorare bene su ogni G , quindi anche G' ottenuto da G mettendo un nuovo nodo z non iniziatore, tolgo arco xy e aggiungo xz e zy con

$$\lambda_x(x, z) = \lambda_x(x, y)$$

e

$$\lambda_y(y, z) = \lambda_y(y, x)$$

Se eseguo A su G' allora z non riceve mai il messaggio I quindi A non è corretto. ■

2.3. Problema wake-up

Broadcast parto da una e mando a tutti, in wake-up è generale, ho tot entità attive che devono mandare. Rilassiamo il vincolo unico iniziatore quindi.

Protocollo wFlood, ho $S = \{\text{dormiente}, \text{attivo}\}$, $S_{\text{init}} = S_{\text{start}} = \{\text{dormiente}\}$, $S_{\text{term}} = S_{\text{final}} = \{\text{attivo}\}$

Se dormiente e impulso spontaneo faccio

- send(W) to N(x)
- become attivo

Se dormiente e ricevo W

- send(W) to N(X) - sender
- become attivo

Il resto è nil

Costi di sta roba:

- $T[\text{wFlood}] \text{ lt.eq } d$
- $2m - n + 1 \text{ lt.eq } M[\text{wFlood}] \text{ lt.eq } 2m$ (1 entità e tutte le entità)

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

3. Traversal

Ogni entità della rete deve essere visitata MA sequenzialmente, cioè una dopo l'altra. Applicazioni è la gestione delle risorse condivise. Versione ristretta del wake up.

Si parte da una visitata, le altre sono dormienti/unvisited, finale ho tutte visited ma una alla volta, in sequenza, ad ogni unità di tempo se ne aggiunge una nuova alla volta

Protocollo depth-first traversal, ovvero una visita in profondità, «si scende sempre verso il vicino non ancora visitato»

Usiamo un messaggio particolare, un token T . In ogni istante di tempo deve viaggiare al più un token. Quando un nodo lo riceve diventa visitato.

Passi:

- un nodo che riceve T per la prima volta:
 - ricorda il sender
 - fa una lista dei vicini non visitati
 - invia T ad uno di essi
 - aspetta un messaggio da quest'ultima di return/back-edge (svegliata da T / già ricevuto il token)
- il vicino che riceve T :
 - se è il primo T ripete il punto 1
 - altrimenti (già visitato) spedisce back-edge
- solo dopo aver finito la lista dei vicini non visitati, un nodo deve inviare la return al sender

Abbiamo tre tipi di messaggi:

- T token (ordine)
- B back-edge (già visitato)
- R return (finito con i vicini)

Vediamo DF-traversal:

- $S = \{\text{initiator, idle, visited, done}\}$
- $S_{\text{init}} = \{\text{initiator, idle}\}$
- $S_{\text{term}} = \{\text{done}\}$

Restrizioni sono RI

Initiator

- se impulso spontaneo
 - initiator = true
 - unvisited = $N(x)$
 - visit

Idle

- se receiving(T)
 - entry = sender
 - unvisited = $N(x) - \text{sender}$
 - initiator = false

- visit

Procedura VISIT

- if unvisited non vuoto then
 - next = unvisited
 - send(T) to next
 - become visited
- else
 - if (initiator == false) then
 - send(return) to entry
 - become done

Visited

- se receiving(return)
 - visit
- se receiving(back-edge)
 - visit
- se receiving(T)
 - unvisited = unvisited - sender
 - send(back-edge) to sender

Per tutto il resto c'è mastercard

Complessità: osserviamo che se x, y sono identità, sul loro canale passa il token + return o back-edge

Traversal è sequenziale, perché le attivo una alla volta, ma allora $T[\text{DF-traversal}]$ e $M[\text{DF-traversal}]$ sono $2m$ (M perché mando due messaggi per m archi), mentre il tempo idem perché sono sequenziale

Vediamo i lower bound di traversal:

- $M[\text{traversal}] \geq m$ (vale per broadcast)
- $T[\text{traversal}] \geq n-1$ (ogni nodo viene visitato in sequenza)

Nel caso di un grafo connesso si passa a

$$n - 1 \leq m \leq \frac{n(n-1)}{2} = O(n^2)$$

quindi DF-traversal ottimo per messaggi ma nel caso peggiore non ottimo per il tempo, passiamo da n teorico a n^2 peggiore

Osservazione: il problema per il costo del tempo è che ad ogni istante viaggia un solo messaggio → mettiamo concorrenza, aggiungendo una quantità opportuna di messaggi $O(m)$ che possano rendere più veloce il protocollo

Possiamo evitare di inviare T su un link back-edge?

Idea: un nodo non visitato che riceve T comunica l'evento ai vicini mandando un messaggio visited, così i vicini tolgono quel nodo dagli unvisited

Abbiamo evitato i link back-edge? NO, ma migliorato lo stesso

La nuova complessità è:

- messaggi
 - $2n-2$ per T + return
 - $2m - (n-1)$ per visited
 - $2(m - (n-1))$ per errori sull'invio dei T su back-edge

In totale è $O(m)$

Tempo ideale non ho ritardi e gli errori non capitano. Inoltre, visited viaggia assieme a T, quindi il tempo diventa $2(n-1)$ (TR) = $O(n)$ che è finalmente il lower bound

DF* è protocollo con ste modifiche, ed è ottimo per messaggi e tempo

$S = \{\text{initiator, idle, available, visited, done}\}$, $S_{\text{init}} = \{\text{initiator, idle}\}$, $S_{\text{term}} = \{\text{done}\}$

Initiator

- spontaneo
 - initiator = true
 - unvisited = $N(x)$
 - next = unvisited
 - send T to next
 - send visited to $N(x)$ - next
 - become visited

Idle

- riceve T
 - unvisited = $N(x)$
 - first-visit
- riceve visited
 - unvisited = $N(x)$ - sender
 - become available

Available

- ricevet T
 - first-visit
- riceve visited
 - unvisited = unvisited - sender

Visited

- riceve visited
 - unvisited = unvisited - sender
 - if (next = sender) then visit
- riceve T
 - unvisited = unvisited - sender
 - if (next = sender) then visit
- riceve return
 - visit

Procedura first-visit (prima volta che si riceve il token per mandare tutto assieme)

- initiator = false
- entry = sender
- unvisited = unvisited - sender
- if non vuoto then
 - next = unvisited
 - send(T) to next
 - send(visited) to $N(x)$ - {entry,next}
 - become visited
- else
 - send(return) to entry

- send(visited) to $N(x)$ - entry
- become done

Procedura visit (già mandato il visited)

- if unvisited non vuoto then
 - next = unvisited
 - send(T) to next
- else
 - if not initiator then
 - send(return) to entry
 - become done

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

4. Spanning tree

Osservazione: broadcast, wp e tr sono $\Theta(m)$ con m = numero di link e n entità

Ma noi sappiamo che

$$n - 1 \leq m \leq \dots = O(n^2)$$

ma $n-1$ è un albero e l'altro è un grafo completo. Noi non scegliamo la rete, possiamo costruire una sotto-rete

Perché importante? Strategia:

- al posto di usare tutta G usiamo una sottorete per minimizzare la complessità di comunicazione, quale sottorete? ALBERO

Attenzione ai costi:

- costruzione dell'albero
- costo originale sull'albero

Ad esempio, BD sull'albero è esattamente $n-1$ su alberi perché hai

$$2m - n + 1 = 2(n - 1) - n + 1 = 2n - 2 - n + 1 = n - 1$$

Vogliamo costruire sottorete tale che

- coinvolge tutte le entità
- le entità sono connesse
- è priva di cicli

La soluzione distribuita richiede la conoscenza dell'albero all'interno della rete, ovvero ogni entità vedrà una piccola parte dell'albero (noi siamo migliori)

Definiamo $\forall x \in E$ la roba $\text{Tree-N}(x) \subset N(x)$, sottoinsieme di vicini che partecipano all'albero e che sono collegati direttamente a x

Diciamo che un arco (x,y) appartiene agli archi $\text{link}(\text{Tree-N}(x))$ se e solo se y sta in $\text{Tree-N}(x)$. Link è un insieme di archi.

Infine, Tree è union di x in E di $\text{link}(\text{Tree-N}(x))$

Dobbiamo anche dire chi è la radice

Noi usiamo restrizioni RI con il protocollo Shout:

- ogni entità vede solo i suoi $\text{Tree-N}(x)$ e tiene traccia del padre

La radice è l'entità che inizia il protocollo

Strategia Shout: CHIEDI!

Vediamo:

- la $s(\text{root})$ spedisce Q ai suoi vicini e attende le risposte
- ogni entità x diversa da s che riceve Q per:
 - la prima volta risponde YES e invia Q ai suoi vicini e si mette in attesa
 - una volta successiva alla prima risponde NO

- serve memorizzare l'entità padre e le entità che mi rispondono yes
- entità termina quando riceve tutte le risposte

In pratica è flooding con reply. Dobbiamo

- mandare messaggi Q yes no
- aggiornare variabili root, parent, tree-N(x), counter
- aggiornare lo stato per raggiungere la terminazione

Abbiamo quindi

- stati iniziatore, inattivo, attivo, finito
- Sinit = {iniziatore, inattivo}
- Sterm = {finito}

Dobbiamo definire le azioni per iniz, inat e attivo

Iniziatore:

- se impulso spontaneo
 - root = true
 - counter = 0
 - tree-N(x) = vuoto
 - send(Q) to N(x)
 - become attivo

Inattivo

- se riceve Q
 - root = false
 - parent = sender
 - counter = 1
 - tree-N(x) = sender
 - send yes to sender
 - if counter = |N(x)| then
 - become finito
 - else
 - send(Q) to N(x) - sender
 - become attivo

Attivo

- se ricevo Q
 - send no to sender
- se riceve yes
 - Tree-N(x) U= {sender}
 - counter += 1
 - if counter == |N(x)| then
 - become finito
- se riceve no
 - counter += 1
 - if counter == |N(x)| then
 - become finito

Correttezza di Shout

- terminazione: in assenza di errori ricevuto un numero di risposte pari ai Q inviati, diventando finito
- tutte le entità sono presenti, grazie al flooding di Q

- le entità sono connesse: grazie al fatto che al primo Q rispondo con yes
- è priva di cicli: ogni entità risponde yes una e una sola volta, tranne la radice che risponde sempre no

Vediamo i costi:

- $M[\text{Shout}] = 2 M[\text{flooding}] (Q + \text{risposta}) = 2[2m - (n-1)]$ circa $4m$
- $T[\text{Shout}] = T[\text{flooding}] + 1 \text{ lt.eq } d + 1$ (+1 è risposta ultimo Q)

I lower bound sono:

- $M[\text{SPT} / \text{RI}] \text{ gt.eq } m$
- $T[\text{SPT} / \text{RI}] \text{ gt.eq } d$

Vediamo Shout++: posso eliminare qualche messaggio? Per il tempo ci siamo, i messaggi non tanto

Yes si tengono, sono necessari, quelli no li cancelliamo. Questo perché se prendo no vuol dire che il bro ha già ricevuto un Q, quindi mi basta vedere un Q e interpretarlo come no

Se la risposta è no ho già ricevuto un Q a cui ha detto sì, inviano altri Q in giro, quindi se ricevuto un Q lo interpreto come no e basta

Nuovo costo è $2m$, un q in una direzione e nell'altra yes oppure q, quindi $M[\text{Shout}++] = 2m$ (q-q o q-yes)

Altra soluzione usa il protocollo traversal, che però costruisce l'albero in sequenza, e a noi piace in parallelo. Tree sono i link su cui viaggiano i return (solo un padre)

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

5. Election

Rompere la simmetria

Individuare una entità specifica tra tante autonome e omogenee. Tale è leader e le altre sono follower. Applicazioni: per certi lavori serve una unità centrale che diventi coordinatrice per le altre entità.

Risultato di impossibilità

Lemma 5.1: Impossibile deterministicamente individuare un leader sotto le restrizioni R

Dimostrazione 5.1: Idea della prova: siano $x, t \in E$ omogenee. Esse sono nello stato e iniziate nello stesso modo. Eseguono stesso algoritmo e sono ancora in simmetria. Ma allora non ho trovato un leader. ■

Risultato di possibilità

Lemma 5.2: Sotto RI la starting entità diventa subito leader, il problema però è risolto dall'esterno e non dal sistema

Nuova restrizione: initial distinct values (ID), con IR notazione $R \cup \{ID\}$, ovvero ho $id(x)$ = nome di x o valore di x (si confonde)

Strategia di soluzione:

- elect minimum:
 - trova $id(x)$ minimo e fai x leader
 - $\forall y \neq x \in E$ y diventa follower
- elect minimum initiator:
 - trova $id(x)$ minimo tra le sole entità initiator ed eleggi x leader
 - same secondo punto

Primo risolviamo in una topologia ring, ad anello

Topologia ring: le entità sono disposte ad anello, ovvero ho $A = (x_0, \dots, x_{n-1})$ e questo ha $m = n$

Aggiungiamo una restrizione, ovvero ogni entità x sa di essere in un ring

Chiameremo per ora $N(x)$ -sender come OTHER, perché ho solo un altro vicino se tolgo il sender

Protocollo All the Way

I messaggi viaggiano intorno all'anello, inoltrati dalle entità nella stessa direzione. Messaggi sono («elect», $id(x)$, ...)

Quando x riceve E da y:

- inoltra E
- inoltra E' con $id(x)$ al posto di $id(y)$

Questi verso OTHER

Ogni entità x vede $id(y)$ for all y eq. not x in E e può calcolare il minimo

Quando facciamo terminare le entità?

Risposta parziale: una volta che x riceve un msg E con il proprio $id(x)$ sa che E ha fatto il giro e quindi non lo inoltra più

Può terminare?

- sì: se supponiamo message ordering (prelevo sui link secondo FIFO, ma noi non ce l'abbiamo)
- solo se ne ha visti n diversi: se si suppone che le entità siano a conoscenza della dimensione (ma noi non ce l'abbiamo)
- no: giusto, dobbiamo riempire in maniera opportuna i msg E per far terminare correttamente le altre entità (un contatore)

Come usare il counter su $E = (elect, id(x), counter)$

- inizio ho counter = 1 per x
- ogni altra entità y diversa da x che inoltra E somma 1 a counter
- quando E ritorna a X, il counter sarà uguale a $n = abs(A)$
- se x ha ricevuto n diversi id può terminare
- altrimenti aspetta, riceve altri messaggi e li inoltra, controlla per verificare se è arriva a n id diversi

Allora

- stati {asleep, awake, leader, follower}
- Sinit = {asleep}
- Sterm = {leader, follower}

Asleep

- spontaneo
 - initialize
 - become awake
- ricevono (elect, value, counter)
 - initialize
 - send (elect, value, counter + 1) to other
 - $min = \min\{min, value\}$
 - count = count + 1
 - become awake

Procedura initialize

- count = 0
- size = 1
- know = false
- send(elect, $id(x)$, size) to right
- min = $id(x)$

Awake

- ricevo (elect, value, counter)
 - if value diverso $id(x)$ then
 - send (elect, value, counter + 1) to other

- min = Min{min, value}
- count = count + 1
- if know = true then check
- ▶ else
 - size = counter
 - know = true
 - check

Procedura CHECK

- if count == size then
 - ▶ if min = id(x) then
 - become leader
 - ▶ else
 - become follower

Complessità

- $M[\text{All the way} / \text{IR union Ring}] = n^2$

Troppo costoso, vediamo versione due

Solo gli initiator generano E, mentre le altre inoltrano e basta

Problema di terminazione: da parte delle entità non initiator. Quando gli initiator hanno finito il calcolo del leader mandano messaggio di fine altri $\approx n$ messaggi di fine

Complessità:

- $M[\text{Min}] = nk + n$ dove k sono gli initiator
- $T[\text{Min}] \leq 3n - 1$ perché vanno in parallelo, per il tempo consideriamo caso peggiore, ovvero solo 2 initiator si attivano (n per il ciclo del primo, n per il secondo, n per il check). Lo raggiungiamo con 2 bro che si svegliano in momenti diversi

Indice

Introduzione	4
Algoritmi 101	4
La risorsa tempo	4
1. Architetture parallele	7
1.1. Memoria condivisa	7
1.2. Memoria distribuita	7
1.3. Modello PRAM	7
1.3.1. Struttura	7
1.3.2. Risorse di calcolo	8
2. Sommatoria	11
2.1. Prima versione	11
2.2. Seconda versione [ottimizzata]	14
2.3. Operazione iterata	15
3. Applicazioni di sommatoria	17
3.1. Prodotto interno di vettori	17
3.2. Prodotto matrice vettore	17
3.3. Prodotto matrice matrice	17
3.4. Potenza di matrice	18
4. Somme prefisse	19
4.1. Prima versione	19
4.2. Seconda versione [pointer doubling]	19
5. Valutazione di polinomi	22
6. Ricerca	24
7. Ordinamento	26
7.1. Counting sort	26
7.2. Bitonic sort	28
7.3. Bitonic sort generico	31
7.4. Osservazioni	32
8. Navigazione	34
8.1. Cicli euleriani	34
8.2. Attraversamento in pre-ordine	36
1. Introduzione	39
1.1. Architetture parallele a memoria distribuita	39
2. Max e ordinamento	42
3. Array lineari	46
4. Architettura MESH	51
4.1. Ordinamento LS3	52
1. Introduzione	56
2. Broadcasting	60
2.1. Prima versione	60
2.2. Seconda versione [flooding]	61
2.3. Problema wake-up	62

3. Traversal	64
4. Spanning tree	69
5. Election	73
6. Routing	77

6. Routing

Scopo: x vuole spedire messaggio a y, vuole cammino in G da x a y

Problema shortest path: scopo determinare cammino migliore (di costo minimo) tra x e y in G

Applicazioni: comunicazione, cruciale in una computazione di un sistema distribuito

Per risolvere il primo basta fare broadcast da x del messaggio a tutte le y in E, ma totalmente inefficiente

Scegliamo quindi un cammino tra i tanti possibili tra x e y. Ovviamente se è il migliore possibile lo preferiamo

Secondo problema: richiede l'uso della memoria per registrare informazioni sui costi di G per ogni entità al fine di calcolare i cammini minimi verso ogni altra entità

Shortest path

Mettiamo restrizioni IR. Le strategie differiscono dal tipo di info che le entità tengono in memoria.

Full routing table: ogni strategia alla fine ha bisogno di questa tabella per risolvere il problema

La tabella ha come righe le destination, mentre sulle colonne il path minimo e il costo (momento reti)

Protocollo gossiping (tanta memoria)

Idea: ogni entità costruisce una mappa del sistema G, una matrice che è la matrice di adiacenza di G e all'occorrenza di calcola le righe della full routing table

Questa la chiamo MAP(G). Nella cella i,j ho peso arco i,j

Per costruire la MAP(G) in un distribuito ogni entità x diffonde le proprie informazioni sui vicini ad ogni altra y in E

Map-Gossip

- costruzione di un albero T per G
- ogni entità acquisisce dai vicini id e i costi del link
- ogni entità diffonde le sue informazioni a tutte le altre usando i link di T

Complessità:

- comunicazione di n^2 messaggi
 - spanning tree T è $O(m + n \log(n))$ il migliore (non l'abbiamo visto)
 - prendo info dai vicini $2m$
 - broadcast delle info su T $2m(n - 1)$
- quindi $M[\text{Map-Gossip}]$ circa $2mn$ ovvero $O(n^2)$ se sparso, tempo difficile da calcolare

Richiede tanta memoria

Protocollo Iterated-Construction

Strategia: ogni x costruisce la FRT a più riprese senza usare MAP

Inizialmente la FRT contiene info solo sui vicini, mentre i non vicini hanno infinito

Notiamo che la FRT non deve contenere per forza l'intero shortest path per arrivare a z, basta sapere il vicino coinvolto nello shortest per z

Ovvero

$$\forall z \in E \mid z \neq x \quad \begin{cases} \text{costo dello SP per } z \\ \text{primo link dello SP che si traduce in un nodo} \end{cases}$$

Definiamo il distance vector, ovvero la FRT ristretta alle colonne, con solo destination e cost. Indichiamo sta roba con V

Con $V[z]$ indichiamo il cammino minimo da x a z

Iterazione:

- ogni entità diffonde la propria V ai suoi vicini
- sulla base delle info che gli arrivano dai vicini stabilisce se sono stati trovati cammini minimi migliori di quelli della propria FRT e in tal caso la aggiorna

Il numero di iterazioni è $n - 1$ m si dimostra per induzione

Come fa x ad individuare ad ogni iterazione il cammino minimo per arrivare a z?

Sia $V_y^{i[z]}$ il costo del cammino da y a z alla i-esima iterazione. Alla i+1-esima questo costo arriva ai vicini di y.

Sia x uno dei vicini, x si calcola, alla i-esima,

$$w[z] = \min_{y \in N(x)} \{ \vartheta(x, y) + V_y^{i[z]} \}$$

dove theta è il costo del link x,y

Se $w[z] < V_x^{i[z]}$ allora x sceglie $w[z]$ come costo per lo sp per z, aggiorna la FRT e memorizza anche il vicino che ha dato questo costo minimo

Vantaggi: la memoria, meno spazio della map, i DV sono lineari, le MAP erano quadratiche

Complessità:

- $M[IR] = 2m \cdot n$ (n-1) ovvero n-1 iterazioni, mando n volte V e 2m sono i link del grafo
- $T[IR] = (n-1) \cdot \tau(n)$ tempo ideale per trasmettere V, che è $O(1)$ quando G consente messaggi lunghi, altrimenti $O(n)$

Se $\tau(n) = O(1)$ diventa tempo lineare in n e $M = O(mn)$

Nell'altro caso abbiamo $O(mn^2)$ per i messaggi e tempo quadratico

