

Appunti del corso di Metodi numerici per l'informatica

A.A. 2021/2022

Indice

1	Introduzione	6
1.1	Tassonomia di Flynn	7
1.2	Principi del calcolo parallelo	7
1.2.1	Speed-up	7
1.2.2	Efficienza	8
1.2.3	Granularità	8
1.2.4	Accessi in memoria	9
1.2.5	Carico non bilanciato	9
1.2.6	Legge di Amdahl	9
1.3	Cenni finali	10
2	Somma di N numeri in parallelo	11
2.1	Somma prima strategia di parallelizzazione	11
2.1.1	Lavagna	11
2.2	Somma seconda strategia di parallelizzazione	12
2.2.1	Distribuzione dei dati	12
2.2.2	Lavagna	13
2.3	Somma terza strategia di parallelizzazione	13
2.3.1	Lavagna	14
2.4	Legge di amdhal dimostrazione	14
3	Prodotto matrice-vettore	15
3.1	Algoritmo seriale(sequenziale)	15
3.2	Prima strategia suddivisione in blocchi di righe	15
3.2.1	Funzionamento algoritmo	16
3.2.2	Valutazione dell'algoritmo	16
3.3	Seconda strategia distribuzione per blocchi di colonne	17
3.3.1	Funzionamento algoritmo	17
3.3.2	Lavagna	17
3.4	Topologie di processori	18
3.4.1	Topologie	18
3.4.2	Parallelizzazione del problema: distribuzione a griglia cartesiana	19
3.4.3	Sottogriglie	22
3.5	Terza strategia distribuzione per schema a blocchi	23
3.5.1	Algoritmo	23
3.5.2	Lavagna	23

4	GPGPU (General Purpose GPU)	25
4.1	GPU	25
4.1.1	Architettura CPU vs GPU	25
4.1.2	CUDA (Compute Unified Device Architecture)	25
4.1.3	Memorie di una GPU	27
4.1.4	Architetture della GPU	27
4.1.5	Modello di esecuzione di CUDA	27
4.1.6	Proprietà di CUDA	27
4.1.7	Latenza	28
4.1.8	Compute Capability	28
4.1.9	Come funziona uno streaming multi processor (SM)	30
4.2	Somma di due Vettori	32
4.2.1	Codice	32
4.2.2	Lavagna	34
4.3	Esempio di capability	35
4.3.1	Configurazione di un kernel	35
4.4	Somma di due matrici	36
4.4.1	Introduzione	36
4.4.2	Esempio con FERMI	36
4.4.3	Esempio con KEPLER	36
4.4.4	Lavagna	37
4.5	Memorie di una GPU	38
4.6	Schema delle memorie	38
4.6.1	Memoria globale	38
4.6.2	Memoria condivisa	38
4.6.3	Registri	38
4.6.4	Memoria locale	38
4.6.5	Memoria costante	38
4.6.6	Memoria Texture	39
4.7	Shared memory e prodotto scalare	39
4.7.1	Shared memory	39
4.7.2	Organizzazione della shared memory	40
4.7.3	Lavagna tipi di accessi al banco	41
4.8	Prodotto scalare	42
4.8.1	Kernel strategia 1 (banale)	42
4.8.2	Come calcolare C a partire da V	42
4.8.3	Riduzione	42
4.9	Allocazione statica	43
4.9.1	Esempio 1	43
4.9.2	Esempio 2	43
4.10	Allocazione dinamica	43
4.10.1	Esempio 3	43
4.11	Info su shared memory	44
4.12	Lavagne prodotto scalare	45
4.12.1	Prima strategia (banale)	45
4.12.2	Seconda strategia	45
4.12.3	Terza strategia	49

5	Sistemi di Raccomandazione	50
5.1	Big Data: cosa significa?	50
5.2	Sistemi di raccomandazione: conoscere per suggerire	50
5.2.1	Elementi di un sistema di raccomandazione	50
5.2.2	Tipi di sistemi di raccomandazione	51
5.2.3	Problemi dei sistemi di raccomandazione	51
5.2.4	Sistemi collaborative-filtering	51
5.2.5	Formalizziamo il problema	52
5.2.6	Similarità	52
5.2.7	SR Memory based di tipo K-NN	53
5.2.8	Sistemi di raccomandazione Model-Based	54
5.2.9	SR model- based	54
5.2.10	Overfitting dei dati	55
5.2.11	Metodi per trovare U e V	55
5.2.12	Metodi basati sulla SVD	57
5.2.13	SVD Troncata	57
5.2.14	SVD compressione immagini come scegliere K	59
6	MPI	60
6.1	Funzionalità della libreria MPI	60
6.1.1	Le funzioni MPI	60
6.1.2	Struttura di un programma MPI	61
6.2	Comunicazione uno a uno	61
6.2.1	Spedizione	61
6.2.2	Ricezione	62
6.3	Comunicazioni collettive	62
6.3.1	Broadcast	62
6.3.2	Scatter	62
6.3.3	Gather	63
6.3.4	AllGather	63
6.4	Operazioni collettive	63
6.4.1	Reduce	63
7	CUDA	64
7.0.1	Allocazione della memoria sulla GPU	64
7.0.2	Deallocazione della memoria sulla GPU	64
7.1	Scambio dei dati fra CPU e GPU	64
7.2	Somma di due vettori	65
7.3	Le funzioni CUDA	65
7.3.1	Configurazione del kernel CUDA	65
7.3.2	Impostazione della memoria della GPU ad un dato valore	65
7.3.3	Variabili architettura di memoria	66
7.4	Compilazione in cuda	66
7.4.1	Specifiche di compilazione	66
7.4.2	Architettura virtuale e reale	67
7.5	Misura dei tempi: gli eventi	67
7.5.1	Uso degli eventi	67
7.5.2	Ottimizzazione mediante profiling	67

8	Libreria CUBLAS	69
8.1	Libreria BLAS	69
8.1.1	3 livelli di Blas	69
8.2	Programmare con CUBLAS	70
8.2.1	Funzione cublasSetVector	70
8.2.2	Funzione cublasGetVector	70
8.2.3	Funzione cublasSdot	71
8.2.4	Avvio delle CUBLAS	71
8.2.5	Gestione degli errori in Cublas	71
8.2.6	Valore di ritorno dell CUBLAS	72
8.2.7	Compilare con CUBLAS	72
8.3	Cenni sulle matrici in CUBLAS	73
8.3.1	Memorizzazione delle matrici per Cublas	73
8.3.2	Funzione cublasSetMatrix	73
8.3.3	Funzione cublasGetMatrix	74
8.3.4	Funzione cublasSgemv	74

Capitolo 1

Introduzione

Il calcolo sequenziale risolve un problema tramite un algoritmo, le cui istruzioni sono eseguite in sequenza mentre con il calcolo parallelo il problema viene suddiviso in sequenze discrete di istruzioni che vengono eseguite una dopo l'altra.

Es:

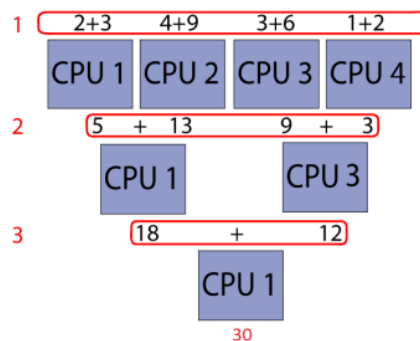


Figura 1.1: Esempio di comunicazione

- Parallelismo temporale (catena di montaggio, pipeline): la presenza della pipeline aumenta il numero di istruzioni contemporaneamente in esecuzione sfruttando anche i tempi morti di un processore per iniziare un nuovo ciclo computazionale. Quindi, introducendo il pipelining nel processore, aumenta il throughput (numero di istruzioni eseguite nell'unità di tempo) ma non si riduce la latenza della singola istruzione (tempo di esecuzione della singola istruzione, dal suo inizio fino al suo completamento), le pipeline devono operare in modo sincrono: questo comporta che lo stadio più lento determini la lunghezza di ogni fase della pipeline;
- Parallelismo spaziale al problema esposto: consiste nell'esecuzione contemporanea della stessa operazione su dati diversi, quindi, sono necessarie più unità aritmetico-logiche (ALU);
- Parallelismo asincrono.

1.1 Tassonomia di Flynn

La tassonomia di Flynn ci mostra quelle che sono le quattro architetture di elaboratori possibili:

- **SISD** (single instruction single data): un singolo set di dati una singola istruzione per volta esempio calcolatori sequenziali, nessun parallelismo possibile, le operazioni vengono eseguite sequenzialmente, su un solo dato alla volta.
- **SIMD** (single instruction multiple data): architetture composte da molte unità di elaborazione che eseguono contemporaneamente la stessa istruzione ma lavorano su insieme di dati diversi.
- **MISD (multiple instruction single data)**: più flussi di istruzioni lavorano contemporaneamente su un unico flusso di dati. Non viene usata perché è inutile.
- **MIMD** (multiple instruction multiple data): più flussi di istruzioni lavorano contemporaneamente su set di dati diversi (il vero parallelismo) dove ho bisogno sia di più unità di elaborazione che più unità di controllo. Abbiamo due tipi di memorie MIMD:
 - **Memoria distribuita**: ogni processore ha la propria memoria e può accedere ai dati di un altro processore con lo scambio di messaggi, un problema è la comunicazione tra processori.
 - **Memoria condivisa**: tutti i processori accedono alla stessa memoria ma ci sono problemi di sincronizzazione e di sicurezza.

1.2 Principi del calcolo parallelo

- Valutazione di un algoritmo parallelo (Speed-Up e Efficienza);
- Trovare e sfruttare la granularità;
- Preservare la località dei dati;
- Bilanciamento del carico computazionale;
- Coordinamento e sincronizzazione;

1.2.1 Speed-up

Lo speed-up misura la riduzione del tempo di esecuzione rispetto all'algoritmo su di 1 processore.

Dato un problema di dimensione fissata, siano T_s : tempo di esecuzione seriale (in sec) e T_p : tempo di esecuzione parallelo usando p processori (in sec):

$$S_p = \frac{T_s}{T_p} \quad (1.1)$$

dove T_s è il tempo algoritmo sequenziale e T_p è il tempo dell'algoritmo parallelo.

Si parla di *speed-up superlineare* quando $S_p > p$, può accadere nei seguenti casi:

1. Gestione ottimale della gerarchia di memoria (ad es. aumento della cache rispetto ad un calcolatore monoprocesso);
2. Differente ottimizzazione del codice;
3. Ordine di visita di grafi nei problemi di ricerca (suddividendo il grafo, uno dei processori potrebbe trovare prima l'oggetto della ricerca).

Speedup assoluto: Tempo di esecuzione sequenziale/tempo di esecuzione parallelo su p processori.

Speedup relativo: Tempo di esecuzione parallelo su 1 processore/tempo di esecuzione parallelo su p processori Restituiscono valori diversi perché un codice parallelo eseguito su un solo processore inserisce overhead superflui, e dunque può essere più lento del codice sequenziale.

Speedup scalato: Misura la scalabilità, ossia la capacità di un algoritmo parallelo di risolvere problemi «grandi», indichiamo con N la dimensione del problema, idealmente $SS_p = 1$:

$$SS_p = \frac{p * T_s(N)}{T_p(pN)} \quad (1.2)$$

1.2.2 Efficienza

L'efficienza misura quanto l'algoritmo sfrutta il parallelismo del calcolatore, ovvero le risorse che ha a disposizione.¹ L'efficienza è una misura relativa al numero di processori utilizzati:

$$E_p = \frac{S_p}{p} \quad o \quad E_p = \frac{T_s}{p * T_p}$$

Idealmente, $E_p=1$. L'efficienza ideale sarebbe: $E_p^{ideale} = \frac{S_p^{ideale}}{p} = 1$

Esempio speed-up e efficienza

Di seguito è riportata una tabella che mostra i diversi valori di speed-up e efficienza sul calcolo della somma in parallelo.

p	Sp	Ep
2	1,88	0,94
4	3,00	0,75
8	3,75	0,47

Tabella 1.1:

Se si rapporta lo speed-up al numero di processori, notiamo che il maggiore sfruttamento dei processori è per $p=2$ perché è “il più vicino” allo speed-up ideale.

Quindi da questo deduciamo che l'utilizzo di un maggior numero di processori NON è sempre una garanzia di sviluppo di algoritmi paralleli “efficaci”.

1.2.3 Granularità

Parallelismo a grana fine: il programma è suddiviso in un gran numero di piccoli compiti, assegnati ai vari processori

- il parallelismo a grana fine facilita il bilanciamento del carico;
- Il numero di processori richiesto per eseguire il programma è alto: questo aumenta l'overhead comunicazione e sincronizzazione.

Parallelismo a grana grossa: il programma è suddiviso in compiti di grandi dimensioni, assegnati ai vari processori.

- ciò potrebbe causare squilibri nella distribuzione del carico tra i processori, o addirittura far sì che una significativa parte del calcolo venga svolta in sequenziale;

¹Nell'algoritmo della somma ad ogni passo di computazione il numero totale di processori attivi si dimezza

- il vantaggio di questo tipo di parallelismo è basso numero di comunicazioni e fasi di sincronizzazione.

Overhead: se la quantità di computazione parallela è consistente, l'overhead è la barriera più importante che impedisce di ottenere valori ottimali di speedup.

Overhead di un algoritmo parallelo:

- costo dello starting di un processo o thread;
- costo della comunicazione di dati condivisi;
- costo della sincronizzazione;
- computazione extra (ridondante).

Ciascuna di queste può essere dell'ordine di millisecondi (= milioni di flops) su alcuni sistemi.

Quindi, in conclusione, un algoritmo deve avere unità di lavoro sufficientemente grandi per una veloce esecuzione parallela, cioè elevata granularità, ma non così grandi che non ci sia abbastanza lavoro da svolgere in parallelo.

1.2.4 Accessi in memoria

- Le memorie più grandi sono lente, quelle veloci sono piccole;
- Le gerarchie di memoria sono grandi e veloci mediamente;
- Processori paralleli, nel complesso, hanno memorie grandi e veloci, gli accessi lenti a dati remoti sono chiamate 'comunicazioni';
- Un algoritmo dovrebbe eseguire la maggior parte delle istruzioni su dati locali.

1.2.5 Carico non bilanciato

Un cattivo bilanciamento del carico si verifica quando ci sono alcuni processori che non eseguono operazioni a causa di insufficiente parallelismo (in quella fase) oppure compiti di dimensione diversa, come ad esempio per problemi che hanno una struttura non omogenea.

1.2.6 Legge di Amdahl

Il miglioramento nelle performances che può essere realizzato mediante il parallelismo è limitato dalla frazione di calcolo sequenziale inevitabilmente presente. Sia α = frazione seriale dell'algoritmo che non può essere eseguita in parallelo. Ad esempio: inizializzazione dei cicli, lettura/scrittura da unità di memoria, overhead della chiamata a funzioni. Il tempo di esecuzione parallelo è

$$T_p = \alpha T_s + (1 - \alpha) \frac{T_s}{p} \quad (1.3)$$

La legge di Amdahl dà una limitazione allo speedup in termini di α

$$\begin{aligned} T_p &= \alpha T_s + (1 - \alpha) \frac{T_s}{p} \text{ allora } S_p = \frac{T_s}{\alpha T_s + \frac{(1-\alpha)T_s}{p}} \\ &= \frac{1}{\alpha + \frac{(1-\alpha)}{p}} \leq \frac{1}{\alpha} \end{aligned}$$

Per esempio se $\alpha = 10\%$, allora il massimo speed-up è 10, anche se usiamo un numero enorme di processori.

Se la dimensione n del problema è fissata, al crescere del numero di processori p :

$$S_p = \frac{1}{\alpha + \frac{(1-\alpha)}{p}} \leq \frac{1}{\alpha} \quad \text{per } p : \rightarrow \infty$$

Quindi se la dimensione del problema resta invariata, al crescere del numero di processori lo speed-up resta illimitato.

Per migliorare la legge di Amdahl dobbiamo aggiungere l'overhead quindi diventerà:

$$S_p = \frac{T_s}{\alpha T_s + \frac{(1-\alpha)T_s}{p} + T_{overhead}} \rightarrow \frac{1}{\alpha + \frac{T_{overhead}}{T_s}} \quad \text{per } p : \rightarrow \infty$$

Tuttavia, se p è fissato, al crescere della dimensione n del problema, la parte sequenziale $\alpha \rightarrow 0$, da cui:

$$S_p = \frac{1}{\alpha + \frac{(1-\alpha)}{p}} \rightarrow p \quad \text{per } n : \rightarrow \infty$$

Aumentando la dimensione n del problema si possono ottenere speed up ottimali MA non è possibile aumentare in maniera indefinita n : le risorse (hardware) sono limitate!

Seconda legge di Amdahl Aumentando il numero p di processori e mantenendo fissata la dimensione n del problema si riesce ad utilizzare in maniera efficiente l'ambiente di calcolo parallelo, se $p \leq p_m$.

Aumentando la dimensione n del problema e mantenendo fisso il numero di p processori le prestazioni dell'algoritmo parallelo non degradano se $n \leq n_{max}$ superato questo valore potrei avere problemi con la memoria hardware!

1.3 Cenni finali

- **Complessità di Tempo $T(n)$** : numero di operazione eseguite dall'algoritmo. È importante notare come il numero complessivo di operazioni determina anche il numero dei passi temporali, ovvero del tempo di esecuzione. In un ambiente parallelo, il numero delle operazioni non è legato al numero di passi temporali (posso fare più operazioni in un solo passo temporale).
- Indichiamo con **$T(P)$** il numero di passi temporali su P processori. Se moltiplico $T(n)$ o $T(p)$ con il tempo per singola operazione T_{calc} ottengo il tempo di esecuzione.
- **Complessità di Spazio $S(n)$** : numero di variabili utilizzate dall'algoritmo

Capitolo 2

Somma di N numeri in parallelo

Problema: Vogliamo calcolare la somma di N numeri $a_0 + a_1 + \dots + a_{N-1}$. Sappiamo che su di un calcolatore mono-processore la somma è calcolata eseguendo le n-1 addizioni una per volta secondo un ordine prestabilito.

2.1 Somma prima strategia di parallelizzazione

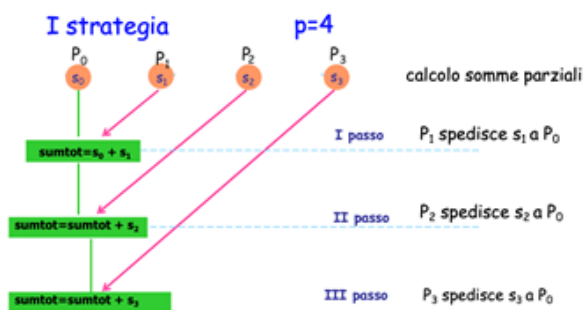
Abbiamo che ogni processore:

- Inizialmente calcola la propria somma parziale.

Ad ogni passo:

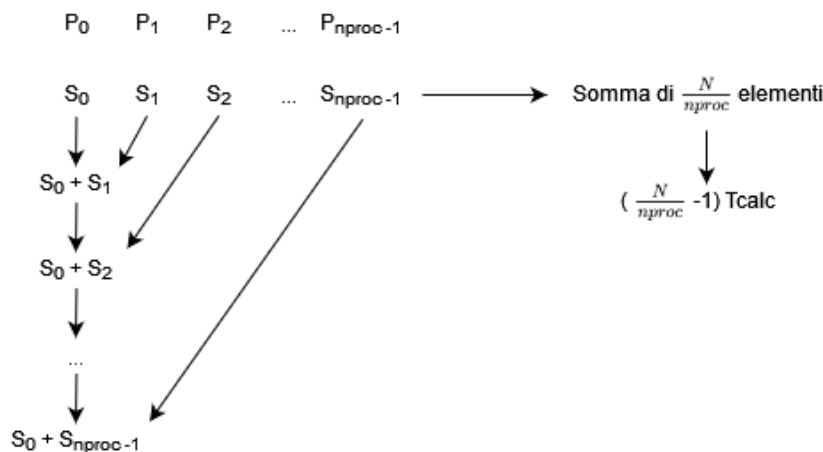
- Ciascun processore invia tale valore ad un unico processore prestabilito.

Tale processore contiene la somma totale.



2.1.1 Lavagna

Abbiamo N -> numeri e Nproc -> processori



In totale abbiamo $nproc - 1$ passi, questo a costo $(nproc - 1) (1 T_{calc} + 1 T_{com})$.
Quindi possiamo dire che:

$$T_P^I = \left(\frac{N}{N_{proc}} - 1\right) T_{calc} + (nproc - 1)(T_{calc} + T_{com})$$

$$T_S = N - 1 T_{calc}$$

$$S_P = \frac{T_S}{T_P} = \frac{N - 1 T_{calc}}{\left(\frac{N}{N_{proc}} - 1\right) T_{calc} + (nproc - 1)(T_{calc} + T_{com})}$$

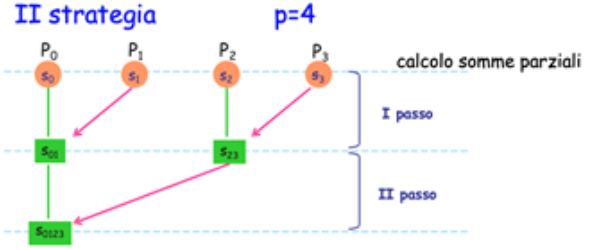
2.2 Somma seconda strategia di parallelizzazione

Abbiamo che ogni processore:

- Inizialmente calcola la propria somma parziale.

Ad ogni passo:

- Coppie distinte di processori comunicano contemporaneamente: in ogni coppia, un processore invia all'altro la propria somma parziale, che provvede all'aggiornamento della somma.



Il risultato è in un unico processore prestabilito

2.2.1 Distribuzione dei dati

Per la Distribuzione dei dati abbiamo 2 casi:

1. N multiplo di $nproc$, $nloc = \frac{N}{N_{proc}}$ con $nloc = nlocale$

$$P_0 : nloc = nloc \text{ dati } a_0, a_1, \dots, a_{nloc-1}$$

$$P_1 : nloc = nloc \text{ dati } a_0, a_1, \dots, a_{2nloc-1}$$

...

$$P_{nproc-1} : nloc = nloc \text{ dati } a_{N-nloc}, a_1, \dots, a_{N-1}$$

2. N non è multiplo di $nproc$, $nloc = \frac{N}{N_{proc}}$ con $nloc = nlocale$ e $r = \text{resto}(N, nproc)$

$$P_0 : nloc = nlocgen + 1 \text{ dati}$$

$$P_{1nloc} = nlocgen + 1 \text{ dati}$$

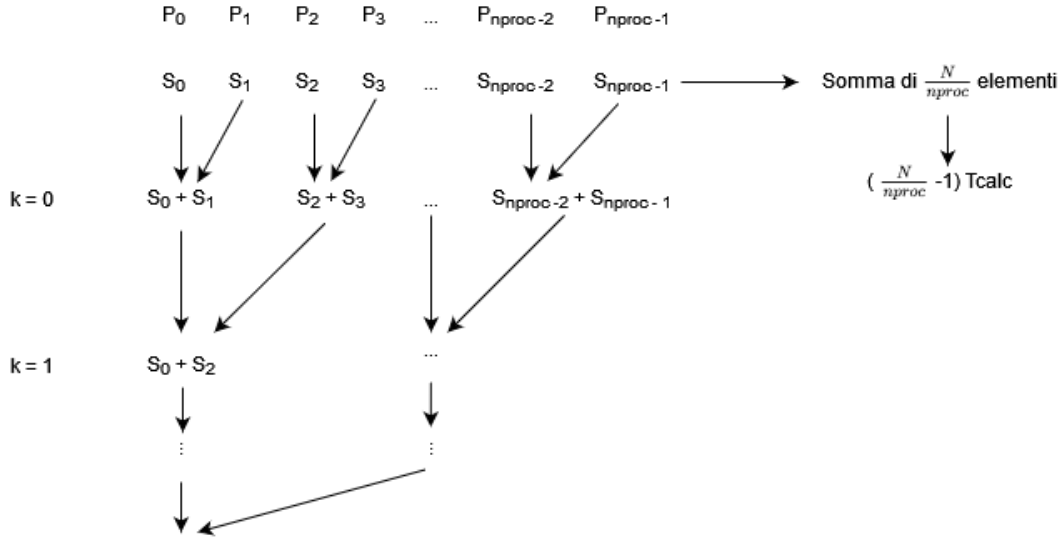
...

$$P_{r-1} : nloc = nlocgen + 1 \text{ dati}$$

$$P_r : nloc = nlocgen + 1 \text{ dati}$$

2.2.2 Lavagna

Abbiamo N e $N_{proc} = 2^P$ Ad ogni passo abbiamo che $DIST = 2^k$ e :



- Se $RESTO(menum, 2^{k+1}) = 0$ ricevo da $menum + DIST$ e $sommaloc = sommaloc + sommari-
cevuta$
- Altrimenti se $RESTO(menum, 2^{k+1}) = 2^k$ invia a $P_{menum} - DIST$

Solo P_0 ha il risultato finale.

Se vedo $nproc$ come 2^k e visto che ad ogni passo dimezzo il numero di processori usati ho $\log_2 nproc$ passi.

$$T_P^{II} = \left(\frac{N}{N_{proc}} - 1\right)T_{calc} + \log_2 nproc(T_{calc} + T_{com})$$

$$S_P = \frac{T_S}{T_P} = \frac{N-1T_{calc}}{\left(\frac{N}{N_{proc}} - 1\right)T_{calc} + \log_2 nproc(T_{calc} + T_{com})}$$

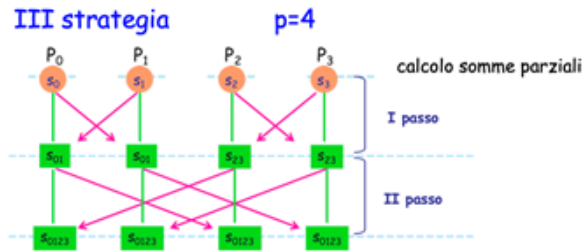
2.3 Somma terza strategia di parallelizzazione

Abbiamo che ogni processore:

- Inizialmente calcola la propria somma parziale.

Ad ogni passo:

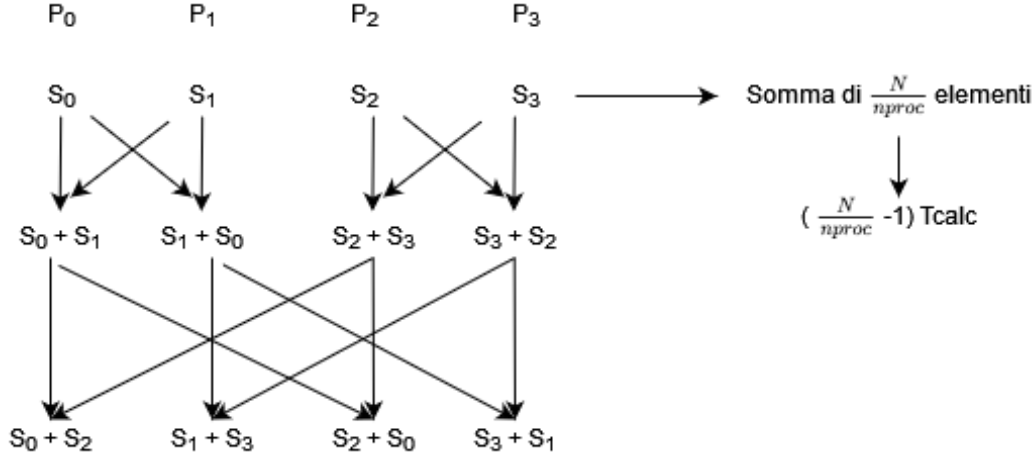
- coppie distinte di processori comunicano contemporaneamente, in ogni coppia i processori si scambiano le proprie somme parziali. Il risultato è in tutti i processori



Il risultato è in tutti i processori.

2.3.1 Lavagna

Esempio con $N = 16$ e $nproc = 4$



Tutti i processi hanno il risultato finale. Mi calcolo i k passi da fare con $\log_2 nproc$.
Ad ogni passo ho $\log_2 nproc$ processi che lavorano.

Se $RESTO(menum, 2^{k+1}) < 2^k$ invio e ricevo da $P_{menum} + DIST$

Se $RESTO(menum, 2^{k+1}) \geq 2^k$ invio e ricevo da $P_{menum} - DIST$

$T_P^{III} = (\frac{N}{N_{proc}} - 1)T_{calc} + \log_2 nproc(T_{calc} + T_{com})$ è uguale alla seconda strategia

2.4 Legge di amdhal dimostrazione

La legge di amdhal da una limitazione allo speed-up in termini di α , dove α è la parte non parallelizzabile.

$$T_P = \begin{cases} \alpha T_S - > \text{parte non parallelizzabile} \\ (1 - \alpha) T_S - > \text{parte parallelizzabile} \end{cases} \quad (2.1)$$

Quindi $T_P = \alpha T_S + \frac{(1-\alpha)T_S}{nproc}$

$$S_P = \frac{T_S}{T_P} = \frac{1}{\alpha T_S + \frac{(1-\alpha)T_S}{p}} = \frac{1}{\alpha + \frac{(1-\alpha)}{p}} \leq \frac{1}{\alpha}$$

$$\text{Notiamo che : } \frac{1}{\alpha T_S + \frac{(1-\alpha)T_S}{p}} = \begin{cases} p - > \infty \text{ allora } \frac{1}{\alpha + (\frac{1-\alpha}{p})=0} = \frac{1}{\alpha} \\ n - > \infty \text{ allora } \frac{1}{\alpha=0 + \frac{1-\alpha}{p}} = p \end{cases} \quad (2.2)$$

- $(1) =$ parte non parallelizzabile α
- $n - > \infty$ inciderà pochissimo
- $(1 - \alpha)$ è più grande

Legge di amdhal generalizzata: $T_S = \alpha_1 T_S + \alpha_2 T_S + \dots + \alpha_p T_S$
Ogni $\alpha_n T_S$ con $n = 1, 2, \dots, p$ rappresenta i processori utilizzati.

Capitolo 3

Prodotto matrice-vettore

Progettazione di un algoritmo parallelo per architettura MIMD a memoria distribuita per il calcolo del prodotto di una matrice (densa) A per un vettore x : $y = Ax$, $A \in R^{m \times n}$, $x \in R^n$, $y \in R^m$.

3.1 Algoritmo seriale(sequenziale)

ogni elemento di ogni riga si moltiplica per ogni elemento del vettore, sommandolo al successivo per riga.

```
for(int i = 0; i < n; i++){  
    for(int j = 0; i < m; j++){  
        y[i] = a[i][j] + y[i];  
    }  
}
```

Costo: 1 moltiplicazione + 1 addizione (= 2tcalc) per ogni $i=1, \dots, m$ e per ogni $j=1, \dots, n$. L'algoritmo sequenziale richiede $m \times n$ moltiplicazione e $m \times n$ addizioni $T_S = 2mntcalc$.

3.2 Prima strategia suddivisione in blocchi di righe

Suddividiamo la matrice A in blocchi di righe. ogni processo ottiene una o più righe della matrice e l'intero vettore, effettua la moltiplicazione e la somma di ogni elemento della riga per ogni elemento del vettore e restituisce una o più celle di quello che sarà il vettore finale.

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 \\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 \\ y_3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 \\ y_4 &= a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 \end{aligned}$$



La

matrice A viene distribuita in blocchi di righe fra i processori. Il vettore x è fornito interamente a tutti i processori. Se la decomposizione dei dati avviene mediante un partizionamento in blocchi di righe della matrice allora il vettore prodotto finale (il vettore y) viene calcolato in parallelo, in blocchi distribuiti fra i processori.

3.2.1 Funzionamento algoritmo

1. Inizializzazione/acquisizione dei dati:
 - Il processo root (quello con identificativo me=0) inizializza (oppure prende in input) la matrice A ed il vettore x;
 - Il processo root calcola la dimensione locale local_m=m/nproc.
2. Distribuzione dei dati, il processo root distribuisce a tutti i processi:
 - m, n e local_m broadcast;
 - la sottomatrice local_A della matrice A scatter per righe), se stesso incluso;
 - l'intero vettore x broadcast.
3. Calcolo dei prodotti parziali:
 - Ciascun processo calcola il prodotto parziale local_y = local_A x.
4. Combinazione dei risultati:
 - Il processo root raccoglie gli elementi local_y calcolati da ogni processo nel vettore risultato y gather.
5. Stampa del risultato finale: il solo processo root stampa il vettore y.

3.2.2 Valutazione dell'algoritmo

Sia p il numero di processori $p \leq m$, ogni processo esegue un prodotto mat-vet tra una matrice (m/p) e un vettore di lunghezza n:

$$T_P = \frac{2mn}{nproc} T_{calc}$$

$$S_P = \frac{T_S}{T_P} = \frac{2mn T_{calc}}{(\frac{2mn}{nproc}) T_{calc}} = p$$

$$E_P = \frac{S_P}{p} = 1$$

NOTA: le fasi di comunicazione riguardano unicamente la distribuzione iniziale dei dati e la raccolta finale dei risultati e quindi non sono state considerate nella valutazione dell'algoritmo.

3.3 Seconda strategia distribuzione per blocchi di colonne

Ogni processo ottiene una o più colonne della matrice e un o più celle del vettore, moltiplica ogni elemento della colonna per il rispettivo elemento del vettore, creando un vettore di moltiplicazioni/somme parziali lungo quanto la lunghezza della colonna, che poi verrà sommato agli altri vettori parziali per ottenere quello risultante finale.

$$\begin{aligned}
 y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 \\
 y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 \\
 y_3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 \\
 y_4 &= a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4
 \end{aligned}$$

$A_j = \text{colonna } j\text{-sima della matrice } A$

$$y = A_1x_1 + A_2x_2 + A_3x_3 + A_4x_4 = r_0 + r_1 + r_2 + r_3$$

3.3.1 Funzionamento algoritmo

1. Inizializzazione/acquisizione dei dati:
 - Il processo root (con identificativo 0) inizializza (o legge da file) la matrice A ed il vettore v;
 - Il processo root calcola la dimensione locale $\text{local_n} = n / \text{nproc}$.
2. Distribuzione dei dati il processo root distribuisce a tutti i processi:
 - m, n e local_n (broadcast);
 - la sottomatrice local_A della matrice A (scatter per colonne), se stesso incluso;
 - il sottovettore local_x del vettore x (scatter), se stesso incluso.
3. Calcolo dei prodotti parziali:
 - Ciascun processo calcola un vettore $\text{local_y} = \text{local_A} * \text{local_x}$.
4. Combinazione dei risultati:
 - Mediante un'operazione di riduzione*, vengono sommati in parallelo tutti i vettori local_y , ottenendo il vettore risultato y, e salvati dal processo root (o da tutti i processi).
5. Stampa del risultato finale: il solo processo root stampa il vettore y.

3.3.2 Lavagna

Supponiamo $\text{nproc} = 2$. $A(m \times n) = \begin{bmatrix} A_{00} & A_{01} & A_{02} & A_{03} \\ A_{10} & A_{11} & A_{12} & A_{13} \\ A_{20} & A_{21} & A_{22} & A_{23} \\ A_{30} & A_{31} & A_{32} & A_{33} \end{bmatrix}$ $x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $w = [w_0 \ w_1 \ w_2 \ w_3]$ la

dimensione di w = m.

P_0 avrà: $\begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \\ A_{20} & A_{21} \\ A_{30} & A_{31} \end{bmatrix}$ size: $m \times \frac{n}{\text{nproc}}$ e $\begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$ size: $\frac{n}{\text{nproc}}$

P_1 avrà: $\begin{bmatrix} A_{02} & A_{03} \\ A_{12} & A_{13} \\ A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix}$ $\begin{bmatrix} x_2 \\ x_3 \end{bmatrix}$

Ogni processo fa una moltiplicazione matrice-vettore tra una matrice $m * \frac{n}{\text{nproc}}$ e un vettore $\frac{n}{\text{nproc}}$ producendo un vettore w di dimensione m, questi vettori w saranno poi sommati con una strategia di somma o una reduce. Prima della somma abbiamo:

- $T_P = 2m \frac{n}{\text{nproc}} T_{calc}$

- Se uso la prima strategia per la somma:

$$T_P^I = 2m \frac{n}{nproc} T_{calc} + (nproc - 1)m(T_{calc} + T_{com}), \text{ m indica una somma per ogni riga}$$

- Se usiamo la seconda o terza strategia:

$$T_P^{II} = 2m \frac{n}{nproc} T_{calc} + (\log_2 nproc)m(T_{calc} + T_{com})$$

$$\begin{aligned} \bullet S_P &= \frac{T_S}{T_P} = \frac{2mnT_{calc}}{\frac{2mn}{nproc} + (\log_2 nproc)m(T_{calc} + T_{com})} = \\ &= \frac{1}{\frac{1}{nproc} + \frac{(\log_2 nproc)}{2n}} \quad \text{abbiamo due casi possibili:} \end{aligned}$$

$$\begin{cases} n \text{ fissato} \lim_{nproc \rightarrow \infty} \frac{1}{\frac{1}{nproc} + \frac{\infty}{2n} = \frac{1}{\infty}} = 0 \\ nproc \text{ fissato} \lim_{n \rightarrow \infty} \frac{1}{\frac{1}{nproc} + \frac{\log_2 nproc}{2m} = \frac{1}{\frac{1}{nproc}}} = nproc \end{cases} \quad (3.1)$$

$nproc$ = speed-up ideale

3.4 Topologie di processori

La disposizione più naturale per i processori è una griglia bidimensionale di $p \times q$ processori:

$$\begin{bmatrix} P_{00} & P_{01} & \dots & P_{0,q-1} \\ P_{10} & P_{11} & \dots & P_{1,q-1} \\ \dots & \dots & \dots & \dots \\ P_{p-1,0} & P_{p-1,1} & \dots & P_{p-1,q-1} \end{bmatrix}. \text{ Alcuni termini importanti:}$$

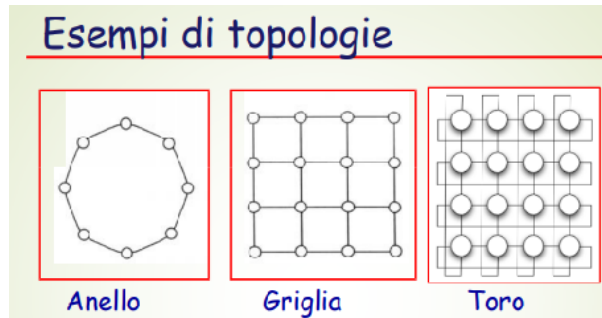
- Contesto (communicator): insieme di processi che possono comunicare reciprocamente attraverso lo scambio di messaggi;
- Una topologia è la geometria “virtuale” in cui si immaginano disposti i processori. La topologia “virtuale” in cui sono disposti i processori può non avere alcun nesso con la disposizione “reale” degli stessi! Infatti è struttura imposta sui processi che fanno parte di un contesto al fine di indirizzare specifici schemi di comunicazione.

3.4.1 Topologie

La topologia di comunicazione identifica gli schemi (principali) di comunicazione che avvengono in un codice parallelo. Le topologie virtuali sono uno strumento utile per: risparmiare tempo, evitare errori, strutturare il codice e sono utili quando gli schemi di comunicazione seguono una o più strutture precise.

- Topologia lineare: Di default, MPI assegna a ciascun processo in un contesto un identificativo da 0 a $n-1$: topologia lineare ($P_0 - > P_1 - > \dots - > P_{n-1}$). Con adeguate funzioni, MPI supporta anche topologie cartesiane e a grafo (ridefinizione del contesto).

- Topologia cartesiana: Ogni processo è identificato da coordinate cartesiane ed è connesso ai vicini da una griglia virtuale. Ai bordi ci può essere o no periodicità e i processi sono identificati dalle loro coordinate.
- altre topologie:



3.4.2 Parallelizzazione del problema: distribuzione a griglia cartesiana

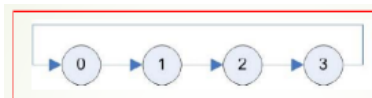
Si può dividere il dominio computazionale a “blocchi”. Questo tipo di suddivisione del dominio computazionale è preferibile perché:

- È compatibile con la geometria del problema;
- Permette una maggiore granularità;
- Riduce la comunicazione.

Però è più complessa da trattare (inizialmente).

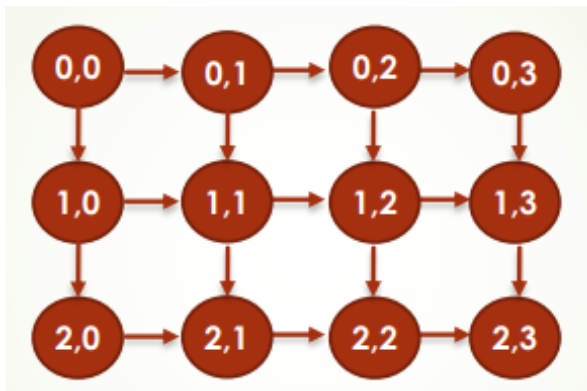
Circular shift

topologia cartesiana 1D (conviene quella ad anello).



Topologia cartesiana 2D

Ad ogni processo è associata una coppia di indici: coordinate nello spazio 2D. Funzione per la creazione di una griglia di processori (periodica e non): `MPI_Cart_create`. Definizione di tipo Communicator: `MPI_Comm comm_grid`.



```

    /* Scopo: definizione di una topologia a griglia bidimensionale nproc=row*col */
int main(int argc, char **argv) {
    int menum,nproc,menu_grid,row,col;
    int dim,*ndim,reorder,*period;
    int coordinate[2];

    MPI_Comm comm_grid; /* definizione di tipo communicator */
    MPI_Init(&argc,&argv);
    MPI_Comm_rank(MPI_COMM_WORLD,&menu);
    MPI_Comm_size(MPI_COMM_WORLD,&nproc); /* Numero di righe della griglia di processo */

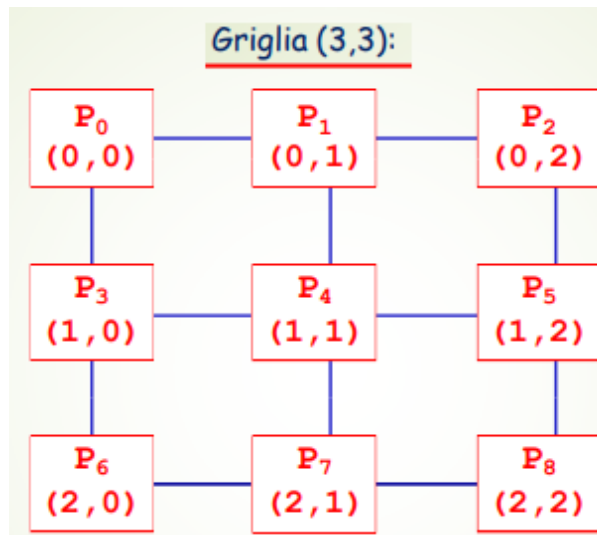
    if (menu == 0) {
    printf("Numero di righe della griglia ");
    scanf("%d",&row); }
    /* Spedizione di row da parte di 0 a tutti i processori */
    MPI_Bcast(&row,1,MPI_INT,0,MPI_COMM_WORLD);
    dim = 2; /* Numero di dimensioni della griglia */
    col = nproc/row; /* Numero di colonne della griglia */
    /* vettore contenente le lunghezze di ciascuna dimensione*/
    ndim = (int*)calloc(dim,sizeof(int));
    ndim[0] = row;
    ndim[1] = col;
    /* vettore contenente la periodicit  delle dimensioni */
    period = (int*)calloc(dim,sizeof(int));
    period[0] = period[1] = 0;
    reorder = 0;

    / * Definizione della griglia bidimensionale */
    MPI_Cart_create(MPI_COMM_WORLD, dim, ndim, period, reorder, &comm_grid);
    MPI_Comm_rank(comm_grid, &menu_grid); /* id nella griglia */
    /* Definizione delle coordinate di ciascun processo nella griglia bidimensionale */
    MPI_Cart_coords(comm_grid, menu_grid, dim, coordinate);
    /* Stampa delle coordinate */
    printf("Processore %d coordinate nella griglia (%d,%d) \n", menu,*coordinate,
        *(coordinate+1)); MPI_Finalize();
    return 0; }

```

Nel programma: `MPI_Cart_Create(MPI_COMM_WORLD,dim,ndim,period,reorder,&comm_grid);` Ogni processo dell'ambiente `MPI_COMM_WORLD` definisce la griglia denominata `comm_grid`, di dimensione 2 (`dim`) e non periodica lungo le due componenti (`period[i]=0`, $i=0,1$). Il numero di righe e di colonne della griglia sono memorizzati rispettivamente nella prima e nella seconda componente del vettore `ndim`. Gli identificativi dei processi non possono essere riordinati (`reorder=0`). Sintassi: `MPI_Cart_Create(MPI_COMM Comm_old, int dim, int *ndim, int *period, int reorder, MPI_COMM New Comm);`

- `comm_old`: contesto di input
- `dim`: numero di dimensioni della griglia
- `*ndim`: vettore di dimensione `dim`. La i -esima dimensione ha lunghezza `ndim[i]`.
- `*period`: vettore di dimensione `dim`. Se `period[i]=1`, la i -esima dimensione della griglia   periodica; non lo   se `period[i]=0`.
- `reorder`: permesso di riordinare gli identificativi (1=s ; 0=no)



- `*new_comm`: contesto di output associato alla griglia

Note su MPI_Cart: Se `reorder=1`, MPI potrebbe riassegnare gli id dei processi del nuovo contesto (per potenziale guadagno nelle prestazioni, grazie a vicinanza fisica dei processi). Se `reorder=0`, l'id del processo nel nuovo contesto è identico a quello che aveva nel vecchio contesto

. Se la dimensione complessiva della griglia: è più piccola dei processi disponibili, i processi rimasti fuori avranno in output `MPI_COMM_NULL`, se è più grande, la chiamata a `MPI_Cart` produrrà un errore.

Nel programma: `MPI_Cart_coords(comm_grid, menum, dim, coordinate);`

Ogni processo `menum` calcola le proprie `dim` (2) coordinate (`coordinate[i]`, $i=0,1$) nel contesto `comm_grid`.

Sintassi: `MPI_Cart_coords(MPI_Comm comm_grid, int menum_grid, int dim, int *coordinate);`

Operazione collettiva che restituisce a ciascun processo di `comm_grid` con identificativo `menum_grid`, le sue coordinate all'interno della griglia predefinita. `coordinate` è un vettore di dimensione `dim`, i cui elementi rappresentano le coordinate del processo all'interno della griglia, output.

Funzione che restituisce l'identificativo (`rank`) date le coordinate: `MPI_Cart_rank(MPI_Comm comm_grid, int icoords, int rank);` è un'operazione collettiva che restituisce a ciascun processo di `comm_grid`, date le coordinate `icoords`, l'identificativo `rank` associato all'interno della griglia predefinita.

Abbiamo due tipi di periodicità:

- `period[0] = 1` e `period[1] = 0`: periodicità sulle colonne;
- `period[0] = 0` e `period[1] = 1`: periodicità sulle righe;

Effetti della periodicità: Se imponiamo la periodicità, ogni riferimento prima della prima o dopo l'ultima componente di ogni riga/colonna si riavvolgerà ciclicamente. Se non c'è periodicità, ogni riferimento all'infuori del range definito, produrrà un identificativo negativo (`MPI_PROC_NULL`, che è -1).

3.4.3 Sottogriglie

Spesso si vuole operare su una parte di una topologia, si divide la topologia in sottotopologie che formano griglie di dimensioni minori e ogni sottogriglia genera un nuovo contesto in cui eseguire operazioni collettive.

Funzione per la creazione di una sottogriglia di processori: `MPI_Cart_sub(MPI_Comm comm_grid, int* rdims, MPI_Comm* new_griglia);`

- `comm_grid`: contesto della topologia;
- `int* rdims`: vettore di dimensione `dim` (dimensione della topologia):
 - se `rdims[i]=1`: la *i*-ma dimensione varia;
 - se `rdims[i]=0`, la *i*-ma dimensione è fissata
- `new_griglia`: nuovo contesto della sottogriglia

Il numero di sottotopologie è il prodotto del numero di processi relativi alle dimensioni eliminate. Esempio:

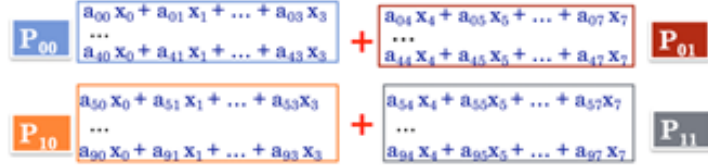
```
/* Crea una topologia 2D cartesiana */
MPI_Cart_create(MPI_COMM_WORLD, ndim, dims, period, reorder, &comm2D);
MPI_Comm_rank(comm2D, &id2D);
MPI_Cart_coords(comm2D, id2D, ndim, coords2D); /* crea le sottogriglie di righe*/
belongs[0] = 0;
belongs[1] = 1; /*dim. Variabile*/
MPI_Cart_sub(comm2D, belongs, &commrow); /* crea le sottogriglie di colonne*/
belongs[0] = 1; /* dim. Variabile*/
belongs[1] = 0;
MPI_Cart_sub(comm2D, belongs, &commcol);
```

Note su `MPI_Cart_sub`:

- Le sottogriglie hanno dimensione pari alla dimensione della griglia di origine, lungo la direzione considerata;
- Ogni sottogriglia è un contesto, che eredita le proprietà della griglia cartesiana, in particolare è ancora una griglia cartesiana (=> ogni processo ha sia id che coordinate);
- `MPI_Cart_sub` è una routine collettiva. È richiamata da tutti i processi del contesto di origine;
- Restituisce il contesto a cui il processo chiamante appartiene.

3.5 Terza strategia distribuzione per schema a blocchi

Ogni processo ottiene una sottomatrice e un sotto vettore, ne calcola il prodotto, che poi viene sommato o concatenato a quello degli altri.



3.5.1 Algoritmo

Le operazioni da effettuare sono le seguenti:

1. Inizializzazione/acquisizione dei dati:
 - Il processo root (con identificativo (0,0)) inizializza (o legge da file) la matrice A ed il vettore v;
2. P_{00} distribuisce il vettore x suddiviso a tutti i processi della sua riga;
3. I processi della 1° riga che hanno il pezzo di x che gli serve e lo mandano in broadcast alla propria colonna;
4. P_{00} distribuisce i pezzi di matrice grandi $n \times \frac{m}{p}$ ai processi della 1° colonna della griglia;
5. i processi della 1° colonna distribuiscono pezzi di matrice grandi $\frac{m}{p} \times \frac{n}{q}$ alla propria riga;
6. ogni processo fa un mat-vet di $\underbrace{\frac{m}{p} \times \frac{n}{q}}_A \times \underbrace{\frac{m}{p}}_x$;
7. ogni processo della 1° colonna somma i risultati parziali della propria riga;
8. Ogni processo della 1° colonna avrà un pezzo di w che verranno unita da P_{00} (gather).

3.5.2 Lavagna

$$A(m \times n) = \begin{bmatrix} a_{00} & \dots & a_{0,n-1} \\ \dots & \dots & \dots \\ a_{m-1,0} & \dots & a_{m-1,n-1} \end{bmatrix} x = \begin{bmatrix} x_0 \\ \dots \\ x_{n-1} \end{bmatrix} w = [w_0 \quad \dots \quad w_{m-1}]$$

A ha dimensione $p \times q$ processi, x ha dimensione n e w ha dimensione m.

Divido la matrice a blocchi di righe e colonne, ogni processo effettuerà una moltiplicazione matrice-vettore tra una matrice $\frac{m}{p} \times \frac{n}{q}$ e un vettore $\frac{n}{q}$ e successivamente ogni processo della prima colonna della griglia somma tutti i vettori parziali grandi $\frac{m}{p}$ con una delle strategie per la somma in parallelo.

- $T_P^I = 2 \frac{m}{p} \frac{n}{q} T_{calc} + (q-1) \frac{m}{p} (T_{cal} + T_{com})$
- $T_P^{II} = 2 \frac{m}{p} \frac{n}{q} T_{calc} + (\log_2 q) \frac{m}{p} (T_{cal} + T_{com})$

$\frac{m}{p}$ perchè ogni processore in contemporanea somma su $\frac{m}{p}$ righe.

Otteniamo così w_0 parziali che saranno uniti da P_{00} con una gather.

$$\bullet S_{pq}^I = \frac{T_S}{T_{pq}^I} = \frac{2mn}{\frac{2mn}{pq} + (q-1)\frac{m}{p}} = \frac{1}{\frac{1}{pq} + \frac{(q-1)}{2np}}$$

$$\bullet E_{pq}^I = \frac{S_{pq}^I}{pq} = \frac{1}{pq} * \frac{1}{\frac{1}{pq} + \frac{(q-1)}{2np}} = \frac{1}{1 + \frac{q(q-1)}{2n}}$$

$$\text{se } n- > \infty \text{ allora } \frac{1}{1+0}=1$$

$$\text{se } q- > \infty \text{ allora } \frac{1}{1+\infty}=0$$

$$\bullet S_{pq}^{II} = \frac{2mn}{\frac{2mn}{pq} + (\log_2 q)\frac{m}{p}} = \frac{1}{\frac{1}{pq} + \frac{(\log_2 q)}{2np}}$$

$$\bullet E_{pq}^{II} = \frac{1}{pq} * \frac{1}{\frac{1}{pq} + \frac{(\log_2 q)}{2np}} = \frac{1}{1 + \frac{q(\log_2 q)}{2n}}$$

$$\text{se } n- > \infty \text{ allora } \frac{1}{1+0}=1$$

$$\text{se } q- > \infty \text{ allora } \frac{1}{\infty}=0$$

Capitolo 4

GPGPU (General Purpose GPU)

4.1 GPU

La GPU (Graphics Processing Unit) è un microprocessore altamente parallelo (many-core) dotato di memoria privata ad alta banda ed è specializzata per le operazioni di rendering grafico 3D. La GPU è vista come un coprocessore matematico con una propria area di memoria (device memory).

4.1.1 Architettura CPU vs GPU

Le GPU sono processori specializzati per problemi che possono essere classificati come intense data-parallel computations:

- lo stesso algoritmo è eseguito su molti elementi differenti in parallelo;
- controllo di flusso molto semplice (control unit ridotta);
- limitata località spaziale (parallelismo a granularità fine) e alta intensità aritmetica che nasconde le latenze di load/store (cache ridotta);
- i thread GPU non hanno costo di attivazione/disattivazione quindi sono leggeri;
- la GPU richiede migliaia di threads per la piena efficienza.

Le applicazioni con un'elevata logica di controllo del processo di calcolo vengono eseguite efficientemente in modo sequenziale dalle tradizionali CPU, ma con pessime prestazioni dall'architettura parallela delle GPU (es. gestione dei database, compressione dei dati, algoritmi ricorsivi).

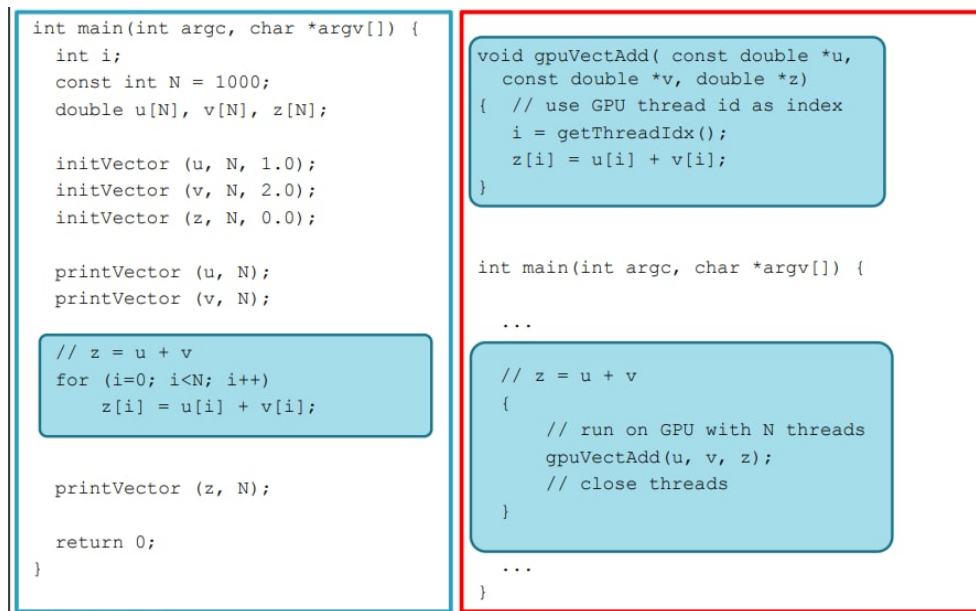
4.1.2 CUDA (Compute Unified Device Architecture)

Il cuore di cuda è basato su un set di istruzioni chiamato PTX (Parallel Thread Execution). La GPU è divisa in due parti:

- data-parallel: definisce una funzione kernel che viene eseguita identicamente da tutti i thread sul dispositivo.
- computational-intensive.

Un'applicazione CUDA combina parti sequenziali (a carico dell'host) e parti parallele, dette kernel (a carico del device).

Il parallelismo su cui si basa è SPMD e SIMT.



Quando si esegue un kernel CUDA sul device GPU bisogna specificare il numero di thread che devono essere lanciati, i thread sono raggruppati in blocchi che sono identificati da un set di coordinate cartesiane 1D, 2D, 3D.

I blocchi di thread costituiscono gli elementi di una griglia, ogni blocco all'interno della griglia può essere identificato mediante un set di coordinate cartesiane bidimensionale.

In ogni CUDA kernel è possibile utilizzare le seguenti variabili per identificare le coordinate del thread corrente e del blocco:

- **threadIdx**: coordinate del thread all'interno del blocco;
- **blockIdx**: coordinate del blocco all'interno della griglia;
- **blockDim**: dimensione del blocco all'interno del thread;
- **gridDim**: dimensione della griglia all'interno del blocco.

La scelta di come calcolare l'indice varia in base alla dimensione della griglia.

A una dimensione abbiamo diverse possibilità:

- $\text{index} = \text{threadIdx.x} + \text{blockIdx.x} * \text{blockDim.x}$ (utilizzato dalla prof) ;
- $\text{index} = \text{numBlocks} * (\text{N} + \text{numThreads} - 1) / \text{numThreads.x}$;
- $\text{index} = (\text{blockIdx} \% x - 1) * \text{blockDim} \% x + \text{threadIdx} \% x$.

A due dimensioni abbiamo:

```

i = blockIdx.x * blockDim.x + threadIdx.x;
j = blockIdx.y * blockDim.y + threadIdx.y;
index = j * gridDim.x * blockDim.x + i;

```

4.1.3 Memorie di una GPU

Per ottenere prestazioni ottimali è fondamentale fare un buon uso delle memorie:

- **Global memory:** memoria ad accesso lento, condivisa da tutti i blocchi thread;
- **Shared memory:** memoria ad accesso veloce, condivisa da tutti i thread di un blocco;
- **Registri:** i dati all'interno dei registri sono visibili solo al thread che li ha scritti e vengono rimossi alla fine del thread.
- **Local memory:** simile ad un registro ma è più lento, vengono memorizzate variabili di dimensione maggiore (array);
- **Constant memory:** è una memoria molto efficiente quando più thread devono accedere allo stesso valore contemporaneamente;
- **Texture memory:** Visibile da ogni thread, si sola lettura

4.1.4 Architetture della GPU

NVIDIA Fermi

È composta da 16 SM (Streaming processor) composto da: 32/48 CUDA core ognuno con una sua ALU sia per interi sia per floating point completamente pipelined, ha registri di 32 bit, 16 unità load/store. Ha anche 4-6 GB di memoria con ECC, ha due cache, L1 configurabile per SM e L2 comune a tutti gli SM. 2 controller indipendenti per il trasferimento dati da/verso host. Uno scheduler globale per distribuire i blocchi thread ad ogni scheduler degli SM.

NVIDIA Kepler

Composta da 192 core CUDA, 4 warp scheduler, 32 unità load/store, i registri sono a 32 bit.

4.1.5 Modello di esecuzione di CUDA

Al lancio di un kernel CUDA ogni blocco della griglia è assegnato ciclicamente a uno SM. Il numero massimo di blocchi in carico a ciascun SM dipende dalle caratteristiche hardware del multiprocessore (memoria shared e registri) e da quante risorse richiede il kernel da eseguire; eventuali blocchi non assegnati vengono allocati non appena un SM completa un blocco (non è possibile fare alcuna ipotesi sull'ordine di esecuzione dei blocchi! (nessuna sincronizzazione!))

I blocchi, una volta assegnati, non migrano, i thread in ciascun blocco sono raggruppati in gruppi di 32 thread di indice consecutivo (detti warp). Lo scheduler seleziona da uno dei blocchi a suo carico dei warp pronti per l'esecuzione e le istruzioni vengono dispacciate per warp, ogni CUDA core elabora uno dei 32 thread del warp.

4.1.6 Proprietà di CUDA

Scalabilità Trasparente

Un kernel CUDA scala su un qualsiasi numero di Streaming Multiprocessor e lo stesso codice può girare al meglio su architetture diverse (non essendo ottimizzato le prestazioni fanno cagare).

Gestione dei Registri

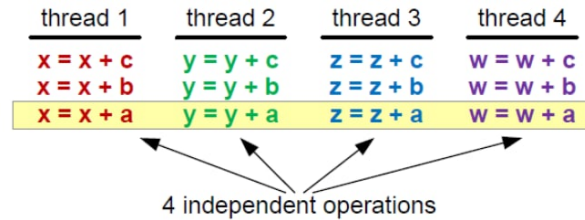
I registri sono partizionati dinamicamente tra tutti i blocchi assegnati ad un SM. I thread di un blocco accedono solo ai registri che gli sono stati assegnati. Lo zero-overload schedule è realizzato grazie al fatto che le informazioni di content switch dei warp rimangono inalterate nei registri del blocco.

4.1.7 Latenza

Una latenza è il numero necessario di cicli affinché un'istruzione venga completata. Possiamo limitarle saturando la pipeline di calcolo e la bandwidth. Questo si ottiene fornendo allo scheduler un numero sufficiente di operazioni da svolgere. Due possibili paradigmi:

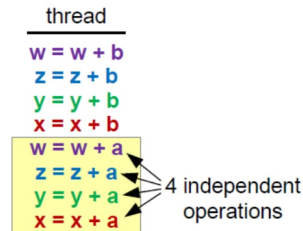
Thread-Level Parallelism (TLP)

Generalmente si consiglia di fornire molti thread per SM in modo da garantire il parallelismo minimo per coprire le latenze aritmetiche.



Instruction-Level Parallelism (ILP)

Abbiamo che ogni processore: Una seconda alternativa è sfruttare il parallelismo interno alle istruzioni dello stesso thread, fornendo allo scheduler più operazioni indipendenti da accodare nella pipeline. Lo scheduler non passa a un nuovo thread se ci sono altre istruzioni da poter istanziare.



4.1.8 Compute Capability

La compute capability specifica il set di istruzioni e features che si vuole supportare per la generazione del codice PTX infatti il codice PTX, pur descrivendo una macchina virtuale, viene esteso e arricchito nel tempo con nuove istruzioni per il nuovo hardware. È identificata da dal codice “compute_Xy”:

- (X): identifica l'architettura base del chip
- (y): identifica varianti con più o meno features

Infine specifica il set di istruzioni disponibili in compilazione del PTX code.

compute capability	feature support
compute_10	very basic features
compute_13	basic + double precision + atomics
compute_20	FERMI architecture (double precision)
compute_30	KEPLER K10 architecture (single precision)
compute_35	KEPLER K20, K20X, K40 architectures

Technical specifications	Compute capability (version)																	
	1.0	1.1	1.2	1.3	2.x	3.0	3.2	3.5	3.7	5.0	5.2	5.3	6.0	6.1	6.2	7.0 (7.2?)	7.5	
Maximum number of resident grids per device (concurrent kernel execution)	t.b.d.				16		4	32				16	128	32	16	128		
Maximum dimensionality of grid of thread blocks	2				3													
Maximum x-dimension of a grid of thread blocks	65535					$2^{31} - 1$												
Maximum y-, or z-dimension of a grid of thread blocks	65535																	
Maximum dimensionality of thread block	3																	
Maximum x- or y-dimension of a block	512				1024													
Maximum z-dimension of a block	64																	
Maximum number of threads per block	512				1024													
Warp size	32																	
Maximum number of resident blocks per multiprocessor	8					16				32							16	
Maximum number of resident warps per multiprocessor	24	32	48		16				64							32		
Maximum number of resident threads per multiprocessor	768	1024	1536		2048												1024	
Number of 32-bit registers per multiprocessor	8 K	16 K	32 K	64 K			128 K	64 K										
Maximum number of 32-bit registers per thread block	N/A			32 K	64 K	32 K	64 K				32 K	64 K		32 K	64 K			
Maximum number of 32-bit registers per thread	124			63		255												
Maximum amount of shared memory per multiprocessor	16 KB			48 KB				112 KB	64 KB	96 KB	64 KB		96 KB	64 KB	96 KB (of 128)	64 KB (of 96)		
Maximum amount of shared memory per thread block	48 KB														48/96 KB		64 KB	
Number of shared memory banks	16			32														
Amount of local memory per thread	16 KB			512 KB														
Constant memory size	64 KB																	
Cache working set per multiprocessor for constant memory	8 KB											4 KB		8 KB				
Cache working set per multiprocessor for texture memory	6 – 8 KB			12 KB		12 – 48 KB			24 KB	48 KB	N/A	24 KB	48 KB	24 KB	32 – 128 KB	32 – 64 KB		
Maximum width for 1D texture reference bound to a CUDA array	8192			65536														
Maximum width for 1D texture reference bound to linear memory	2^{27}																	
Maximum width and number of layers for a 1D layered texture reference	8192 × 512			16384 × 2048														
Maximum width and height for 2D texture reference bound to a CUDA array	65536 × 32768			65536 × 65535														
Maximum width and height for 2D texture reference bound to a linear memory	65000^2																	
Maximum width and height for 2D texture reference bound to a CUDA array supporting texture gather	N/A			16384^2														
Maximum width, height, and number of layers for a 2D layered texture reference	8192 × 8192 × 512			16384 × 16384 × 2048														
Maximum width, height and depth for a 3D texture reference bound to linear memory or a CUDA array	2048 ³					4096 ³												
Maximum width and number of layers for a cubemap layered texture reference	N/A			16384 × 2046														
Maximum number of textures that can be bound to a kernel	128				256													
Maximum width for a 1D surface reference bound to a CUDA array	Not supported				65536													
Maximum width and number of layers for a 1D layered surface reference					65536 × 2048													
Maximum width and height for a 2D surface reference bound to a CUDA array					65536 × 32768													
Maximum width, height, and number of layers for a 2D layered surface reference					65536 × 32768 × 2048													
Maximum width, height, and depth for a 3D surface reference bound to a CUDA array					65536 × 32768 × 2048													
Maximum width and number of layers for a cubemap layered surface reference					32768 × 2046													
Maximum number of surfaces that can be bound to a kernel					8	16												
Maximum number of instructions per kernel	2 million				512 million													

[35]

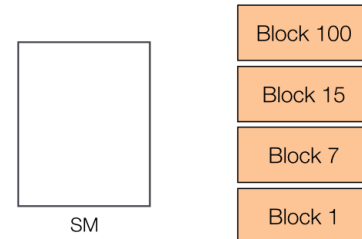
Architecture specifications	Compute capability (version)														
	1.0	1.1	1.2	1.3	2.0	2.1	3.0	3.5	3.7	5.0	5.2	6.0	6.1, 6.2	7.0, 7.2	7.5
Number of ALU lanes for integer and single-precision floating-point arithmetic operations	8 ^[36]				32	48	192			128		64	128	64	
Number of special function units for single-precision floating-point transcendental functions	2				4	8	32			16		32	16		
Number of texture filtering units for every texture address unit or <i>render output unit</i> (ROP)	2				4	8	16			8 ^[37]					
Number of warp schedulers	1				2		4					2	4		
Max number of instructions issued at once by a single scheduler	1					2 ^[38]					1				
Number of tensor cores	N/A													8 ^[37]	
Size in KB of unified memory for data cache and shared memory per multi processor	t.b.d.													128	96 ^[39]

4.1.9 Come funziona uno streaming multi processor (SM)

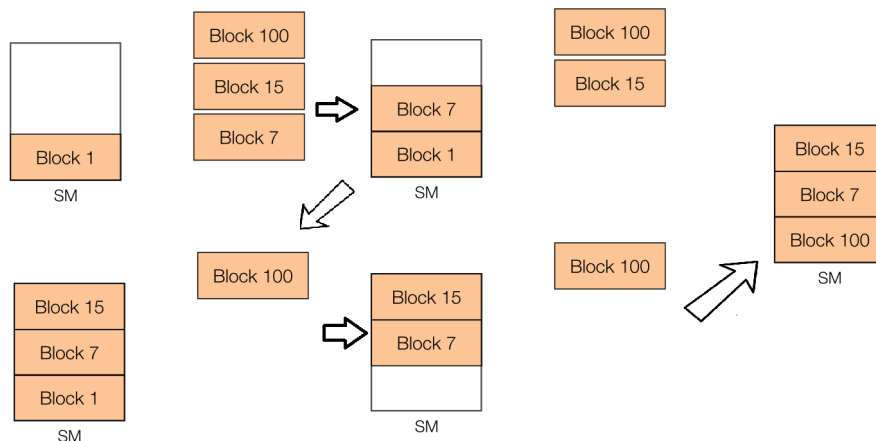
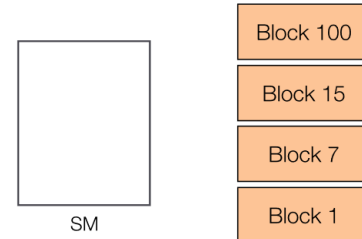
I thread CUDA sono raggruppati in blocchi di thread. I thread dello stesso blocco vengono eseguiti contemporaneamente nello stesso SM. Gli SM hanno memoria condivisa, quindi i thread all'interno di un blocco di thread possono comunicare.

La totalità dei thread di un blocco deve essere eseguita prima che ci sia spazio per schedulare un altro blocco di thread.

L'hardware pianifica i blocchi di thread sugli SM disponibili, non c'è nessuna garanzia sull'ordine di esecuzione e se un SM ha più risorse, l'hardware pianificherà più blocchi.



L'hardware pianifica i blocchi di thread sugli SM disponibili, non c'è nessuna garanzia sull'ordine di esecuzione e se un SM ha più risorse, l'hardware pianificherà più blocchi.



Warps

Il warp è l'unità fondamentale di esecuzione. Ogni warp può eseguire un'istruzione su 16 cuda core, 16 load/store units. Ci vogliono 2 cicli per completare un'istruzione o una load/store per tutto il warp. All'interno degli SM, i thread vengono lanciati in gruppi di 32 chiamati appunto warps. I warp condividono la parte di controllo (warp scheduler). In qualsiasi momento, viene eseguito solo un warp per SM. I thread in un warp eseguiranno la stessa istruzione. Sono usati gli Half warp per la computer capability 1.X.

Es: FERMI ARCHITECTURE: massimo numero di thread attivabili 1024 (thread massimi per blocco) $\times 8$ (blocchi massimi per SM) $\times 32$ (SM) $= 262144$ Lo SM implementa un warp scheduling con zero-overhead. L'architettura Fermi usa due warp scheduling.

4. GPGPU (General Purpose GPU)

Per sfruttare al massimo il parallelismo intrinseco del SM, i thread dello stesso warp devono eseguire la stessa istruzione: se questa condizione non si verifica si parla di divergenza dei threads. Se i thread dello stesso warp stanno eseguendo istruzioni diverse l'unità di controllo non può gestire tutto il warp: deve seguire le successioni di istruzioni per ogni singolo thread (o per sottoinsiemi omogenei di threads) in modo seriale (perdita del parallelismo).

Esempio di divergenza

```
int sum = 0;
if (threadIdx.x < 5 {
    sum = sum +10; *
} else {
    sum = sum + 3; #
}
```

In un warp avremo quindi che:

- Da 0 a 4 thread eseguiranno *;
- da 5 a 31 eseguiranno #.

Questo porta ad avere una divergenza. Come possiamo risolvere questa divergenza?

Risoluzione della divergenza

```
int WARP_SIZE = 32;
int sum = 0;
if (threadIdx.x/WARP_SIZE < 5 {
    sum = sum +10; *
} else {
    sum = sum + 3; #
}
```

Questo comporta che:

Warp	Thread	threadIdx.x/WARP_SIZE	Eseguono
1	0...31	0	*
2	32...63	1	*
...
z	256...282	8	#

Tabella 4.1:

4.2 Somma di due Vettori

Si vogliono sommare due vettori sfruttando come parallelismo la gpu.

4.2.1 Codice

Inanzitutto identifichiamo le parti data-parallel e i dati coinvolti da trasferire.

<pre> int main(int argc, char *argv[]) { int i; /*****/ const int N = 1000; double u[N], v[N], z[N]; /*****/ initVector (u, N, 1.0); initVector (v, N, 2.0); initVector (z, N, 0.0); printVector (u, N); printVector (v, N); /*****/ // z = u + v for (i=0; i<N; i++) z[i] = u[i] + v[i]; /*****/ printVector (z, N); return 0; } </pre>	<pre> program vectoradd integer :: i integer, parameter :: N=1000 real(kind(0.0d0)), dimension(N):: u, v, z call initVector (u, N, 1.0) call initVector (v, N, 2.0) call initVector (z, N, 0.0) call printVector (u, N) call printVector (v, N) ! z = u + v do i = 1,N z(i) = u(i) + v(i) end do call printVector (z, N) end program </pre>
---	---

Poi implementiamo il Kernel CUDA e scriviamo l'algoritmo per il singolo elemento.

<pre> const int N = 1000; double u[N], v[N], z[N]; // z = u + v for (i=0; i<N; i++) z[i] = u[i] + v[i]; </pre>	diventa:	<pre> void gpuVectAdd (const double *u, const double *v, double *z, int N) { // index is a unique identifier // of each GPU thread int index = ... ; if (index < N) z[index] = u[index] + v[index]; } </pre>	La
---	----------	---	----

scelta delle dimensioni della griglia è dettata dal problema e dal mapping che si utilizza nel kernel CUDA tra l'indice degli elementi da elaborare e gli identificativi dei thread CUDA.

Nel caso di un problema lineare come l'elaborazione degli elementi di un array, possiamo scegliere:

- blocchi di thread monodimensionali (1D)
- griglia monodimensionale di blocchi (1D)

Ogni thread della griglia elabora in parallelo un solo elemento del vettore per determinare la griglia in modo da generare un numero di thread sufficiente a ricoprire tutta la superficie da elaborare. Scelta la dimensione del blocco, la dimensione della griglia si può adattare al problema, alcuni thread potrebbero uscire dal dominio (nessun problema).

4. GPGPU (General Purpose GPU)

```
dim3 numThreads(32);
dim3 numBlocks( ( N + numThreads - 1 ) /
    numThreads.x );
gpuVectAdd<<numBlocks, numThreads>>>(
    u_dev, v_dev, z_dev, N );
```

```
__global__ void gpuVectAdd (const double *u,
    const double *v, double *z, int N) {
    // index is a unique identifier of each
    GPU thread
    int index = blockIdx.x * blockDim.x +
        threadIdx.x ;
    if (index < N)
        z[index] = u[index] + v[index];
}
```

Modificare il codice in modo da girare sulla GPU e allocare la memoria sul device.

```
double *u_dev, *v_dev, *z_dev;
cudaMalloc((void **)&u_dev, N * sizeof(double));
cudaMalloc((void **)&v_dev, N * sizeof(double));
cudaMalloc((void **)&z_dev, N * sizeof(double));
```

trasferire i dati necessari dall'host al device

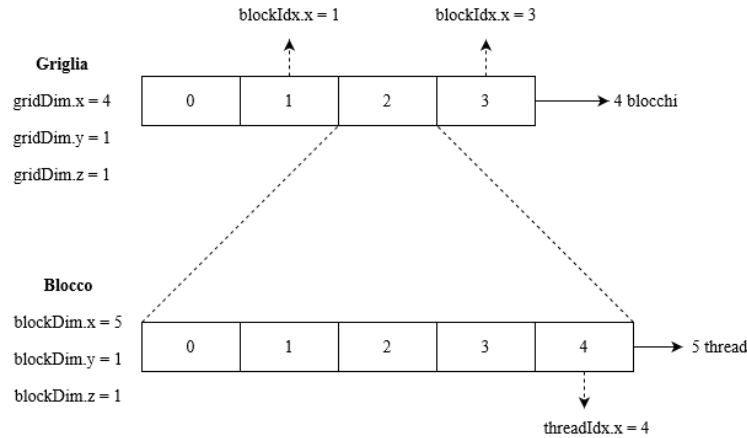
```
cudaMemcpy(u_dev, u, sizeof(u), cudaMemcpyHostToDevice);
cudaMemcpy(v_dev, v, sizeof(v), cudaMemcpyHostToDevice);
```

Porting finale del codice somma di due vettori in cuda c

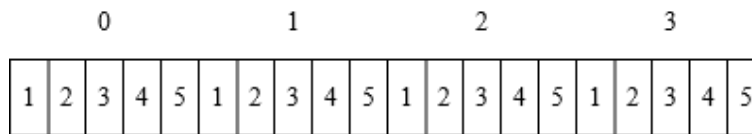
```
double *u_dev, *v_dev, *z_dev;
cudaMalloc((void **)&u_dev, N * sizeof(double));
cudaMalloc((void **)&v_dev, N * sizeof(double));
cudaMalloc((void **)&z_dev, N * sizeof(double));
cudaMemcpy(u_dev, u, sizeof(u), cudaMemcpyHostToDevice);
cudaMemcpy(v_dev, v, sizeof(v), cudaMemcpyHostToDevice);
dim3 numThreads( 256); // 128-512 are good choices
dim3 numBlocks( (N + numThreads.x - 1) / numThreads.x );
gpuVectAdd<<numBlocks, numThreads>>>( u_dev, v_dev, z_dev, N );
cudaMemcpy(z, z_dev, N * sizeof(double), cudaMemcpyDeviceToHost);
```

4.2.2 Lavagna

Abbiamo una griglia una griglia 1D e blocchi di 1D:



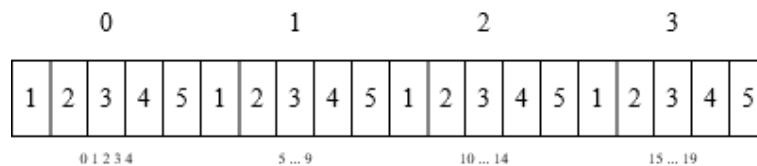
se rappresentiamo la griglia con tutti i thread ci troviamo davanti a:



Abbiamo quindi 20 thread attivabili, ora ritornando alla somma di due vettori, abbiamo che $z = u + v$, con u, v, z vettori di dimensione N .

Abbiamo 3 possibili casi di distribuzione dei dati:

1. Abbiamo $N = 20$



$$\text{index} = \text{threadIdx.x} + \text{blockIdx.x} * \text{blockDim.x}$$

2. $N > 20$: abbiamo delle componenti che rimangono fuori non è accettabile;
3. La situazione di quest'ultimo caso è quella che capita più spesso ed è quella in cui siamo noi a scegliere i numeri dei thread da utilizzare. Quindi se $N < 20$ e $N = 18$
 Associamo ai thread una coppia di componenti:

```
if(index < N)
    z(i) = u(i) + v(i)
```

Problema del terzo caso, con la formula da noi utilizzata potrebbero esserci dei casi in cui nessun blocco ha assegnato delle componenti e questo porta alla perdita del parallelismo.

Per risolvere questo problema scegliamo in modo arbitrario la dimensione.

1. Fissiamo blockDim.x ad un valore, in questo esempio mettiamolo a 32.
2. Scelgo gridDim.x in funzione di N e blockDim

Es:

- $N = 64 \rightarrow \text{gridDim.x} = \frac{N}{\text{blockDim.x}} = \frac{64}{32}$
- $N = 70 \rightarrow \text{gridDim.x} = \frac{70}{32} = 2$ elementi fuori, per non avere elementi fuori possiamo applicare quest'altra formula: $\text{gridDim.x} = \frac{N}{\text{blockDim}} + 1$

Quindi scegliamo in modo che almeno un thread nell'ultimo blocco lavori sempre.

4.3 Esempio di capability

4.3.1 Configurazione di un kernel

Vogliamo configurare un kernel con compute capability di tipo 2.0 (dati reperibili sulla tabella 4.1.8). Innanzitutto dobbiamo tenere conto di:

- Massimo numero di blocchi: 65535
- Massimo numero di thread: 1024
- Massimo numero blocchi in SM: 8
- Massimo numero di thread in SM : 1536
- Massimo numero di registri in SM: 32K

Abbiamo anche due limitazioni il numero di blocchi/thread da utilizzare e la capienza della memoria. Vogliamo ottenere una configurazione ottimale, ma non tutte le configurazioni sono ottimali per via delle limitazioni che possono verificarsi. Una configurazione si dice ottimale se vengono utilizzati il massimo numero di blocchi per ogni SM e se vengono utilizzati il massimo numero di thread per ogni SM.

Il nostro esempio si baserà su di una geometria 1D.

...
-----	-----	-----	-----

Per prima cosa vediamo per il nostro SM quanti thread può contenere all'interno di ogni blocco: $\frac{\#thread}{\#blocchi} = \frac{1536}{8} = 192$. Quindi blockDim = 192 è la scelta massima di # blocchi e #thread per ogni SM e quindi sarebbe la scelta ottimale; ma per vedere se è vero dobbiamo tener conto della memoria. Supponiamo \forall thread occorrono 30 registri, \forall SM abbiamo 30×1536 (8×192) registri = 46 080. Notiamo quindi che abbiamo superato il numero di registri disponibili che è 32K quindi non è una scelta ottima.

Ora supponiamo che si attivano solo 5 blocchi:

$5 \times 192 \times 30 = 28800$, in questo caso va bene perchè non superiamo 32k.

Quindi se scegliamo blockDim = 192 e si attivano 5 blocchi \forall SM abbiamo che alla fine saranno attivi 5×192 thread = 960 thread \forall SM < 1536.

Un'altro esempio: Proviamo un'altra scelta sempre con compute capability 2.0, ma stavolta con blockDim = 128. $8 \times 128 = 1024$ thread < 1536. È ottima? $1024 \times 30 = 30720 < 32K$ quindi è OK. Si attiveranno per ogni \forall 8 blocchi e 1024 thread, la soluzione quindi possiamo dire che è buona ma non è ottima.

4.4 Somma di due matrici

Si vuole effettuare la somma di due matrici con il calcolo parallelo su GPU.

4.4.1 Introduzione

Innanzitutto qual è la dimensione ottimale di un blocco di thread (BLOCK_DIM) su architetture Fermi e Kepler? Esistono infatti dei vincoli strutturali:

- Un blocco di CUDA thread può contenere al massimo 1024 thread;
- La griglia di blocchi non può contenere più di 65535 blocchi per Fermi o 231-1 blocchi per Kepler;
- BLOCK_DIM deve essere scelto in modo da saturare le risorse a disposizione dello Streaming Multiprocessor (SM) e da elevare il grado di parallelismo potenziale.

Infine si deve controllare che vengano rispettati i vincoli sull'utilizzo della memoria.

4.4.2 Esempio con FERMI

Nell'architettura FERMI ogni SM può gestire fino a 1536 thread.

- Numero massimo di 8 blocchi residenti per SM e blocco $8 \times 8 = 64$ thread : $1536/64 = 24$ blocchi per occupare tutto un SM. Con un massimo di 8 blocchi per SM : $64 \times 8 = 512$ thread per SM, su di un totale di 1536 disponibili;
- blocco $16 \times 16 = 256$ thread : $1536/256 = 6$ blocchi $256 \times 6 = 1536$ thread per SM (Piena occupazione dello SM!);
- blocco $32 \times 32 = 1024$ thread : $1536/1024=1.5 \Rightarrow 1$ blocco per SM $1024 \times 1 = 1024$ thread per SM.

Quindi possiamo dire che BLOCK_DIM = 16 è scelta ottimale.

4.4.3 Esempio con KEPLER

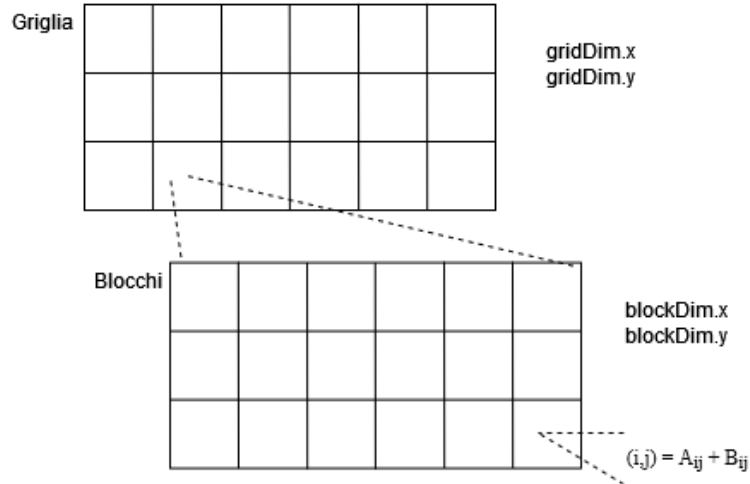
ogni SM può gestire max 2048 thread; max 16 blocchi per SM.

- $8 \times 8 = 64$ thread : $2048/64 = 32$ blocchi per occupare tutto un SM, con un massimo di 16 blocchi per SM : $64 \times 16 = 1024$ thread per SM su un totale di 2048 disponibili;
- $16 \times 16 = 256$ thread : $2048/256 = 8$ blocchi e $256 \times 8 = 2048$ thread per SM (Piena occupazione dello SM!);
- $32 \times 32 = 1024$ thread : $2048/1024 = 2$ blocchi e $1024 \times 2 = 2048$ thread per SM (Piena occupazione dello SM ma minore parallelismo potenziale).

Quindi possiamo dire che BLOCK_DIM = 16 è scelta ottimale.

4.4.4 Lavagna

Supponiamo di avere Due Mtrici A,B enteambe di dimensione $n \times n$ e di volerle sommare: $A + B = C$. Quindi in questo caso non utilizziamo Geometrie 1D ma 2D:



Ogni thread sommerà $C_{i,j} = A_{i,j} + B_{i,j}$.

Per comodità anche in questo esercizio usiamo array monodimensionali. La relazione tra l'indice dell'array monodimensionale dei dati del problema e le coordinte del thread è la seguente:

- $i = \text{blockIdx.x} * \text{blockDim.x} + \text{threadIdx.x}$;
- $j = \text{blockIdx.y} * \text{blockDim.y} + \text{threadIdx.y}$;
- $\text{index} = j * \text{gridDim.x} * \text{blockDim.x} + i$;

Altro esempio: Sia A matrice $M \times N$, supponiamo di voler assegnare un elemento di A a ciascun thread, la corrispondenza è: $A(\text{row}, \text{col})$ con:

- $\text{row} = \text{blockIdx.y} * \text{blockDim.y} + \text{threadIdx.y}$;
- $\text{col} = \text{blockIdx.x} * \text{blockDim.x} + \text{threadIdx.x}$.

4.5 Memorie di una GPU

Un buon uso delle memorie è fondamentale per ottenere buone prestazioni nel calcolo su GPU

4.6 Schema delle memorie

4.6.1 Memoria globale

La memoria globale è la memoria principale della GPU, è accessibile da tutti i thread sia in lettura che in scrittura. È il canale di comunicazione con l'host (cudaMemcpy opera qui). Questa tipologia di memoria è dell'ordine dei Gb ma è lenta. In questa memoria sono presenti tutti gli argomenti del kernel.

4.6.2 Memoria condivisa

E' la memoria che offre le migliori prestazioni, per cui è consigliabile massimizzarne l'uso ed è un canale di comunicazione tra tutti i thread di un blocco. Le variabili della memoria condivisa vivono solo nel kernel e si dichiarano nel seguente modo:

```
__shared__ float variable;
```

La grandezza è nell'ordine dei KB ma è veloce.

4.6.3 Registri

Nei registri vengono memorizzate tutte le variabili locali del kernel. E' la memoria con i migliori tempi di accesso.

```
float variable;
```

Il registro è un'area di memoria privata di ogni thread ed è veloce.

4.6.4 Memoria locale

Memoria usata nel kernel, per memorizzare: array o altre variabili, qualora si superi la memoria dei registri. E'una memoria lenta.

4.6.5 Memoria costante

Memoria da usare quando si deve accedere (in lettura) molte volte ad una stessa variabile, in quanto riduce i tempi di attesa

```
__constant__ float variable;
```

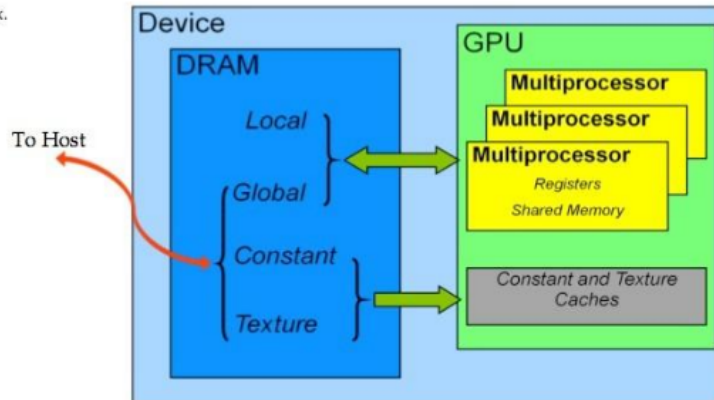
E' una memoria di sola lettura visibile da tutti i thread della griglia e dall'host, è molto veloce.

4.6.6 Memoria Texture

E' una memoria veloce di sola lettura visibile a tutti i thread ottimizzata per la località spaziale 2D.

Memory	Location on/off chip	Cached	Access	Scope	Lifetime
Register	On	n/a	R/W	1 thread	Thread
Local	Off	†	R/W	1 thread	Thread
Shared	On	n/a	R/W	All threads in block	Block
Global	Off	†	R/W	All threads + host	Host allocation
Constant	Off	Yes	R	All threads + host	Host allocation
Texture	Off	Yes	R	All threads + host	Host allocation

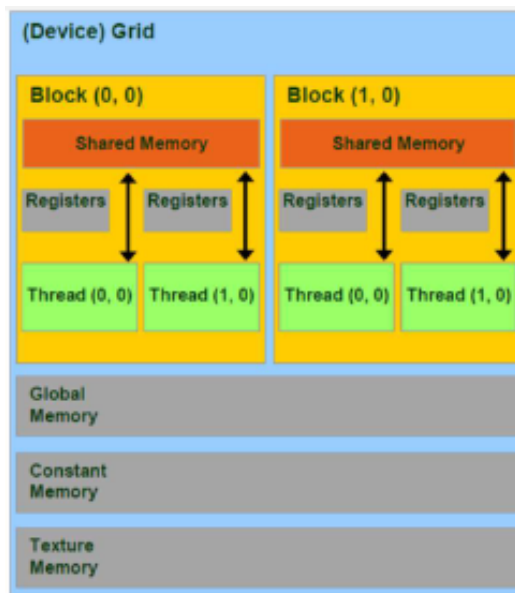
† Cached only on devices of compute capability 2.x.



4.7 Shared memory e prodotto scalare

4.7.1 Shared memory

La memoria condivisa (Shared Memory) è una memoria veloce residente su ciascuno Streaming Multiprocessor. È accessibile in lettura e scrittura dai soli thread del blocco: canale di comunicazione tra i thread di un blocco. La sua durata è limitata al kernel e non mantiene il suo stato tra il lancio di un kernel e l'altro. Bassa Latenza: 2 cicli di clock, throughput: 4 bytes per banco ogni 2 cicli infine di default ha 48 KB (Configurabile : 16/48 KB).



In esecuzione una risorsa critica che limita il parallelismo a livello device. Viene ripartita tra tutti i blocchi residenti in un SM, maggiore è la shared memory richiesta da un kernel, minore è il numero di blocchi attivi concorrenti. L'accesso è per warp e il caso migliore è una transazione x 32 thread, il caso peggiore 32 transazioni.

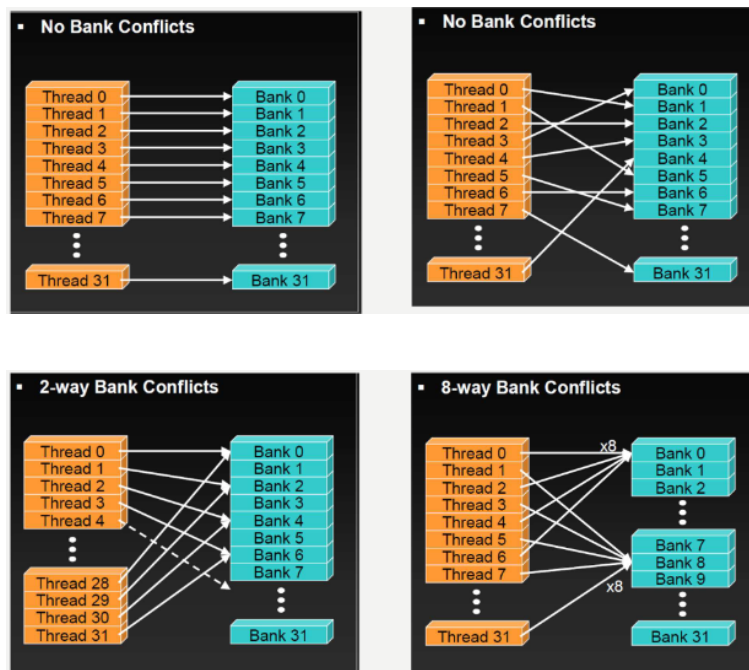
Viene utilizzata nelle seguenti modalità:

- Si caricano i dati nella memoria shared;
- Si sincronizza (se necessario);
- Si opera sui dati nella memoria shared;
- Si sincronizza (se necessario);
- Si scrivono i risultati nella memoria globale.

4.7.2 Organizzazione della shared memory

La shared memory è organizzata in 32 banchi da 4-byte di ampiezza ciascuno (dette word, ciascuna può contenere 1 int, 1 float, 2 short, 4 char, 1 half double, ...). I dati vengono distribuiti ciclicamente su banchi successivi ogni 4-byte, e gli accessi alla shared memory avvengono per warp. Abbiamo diversi tipi di accessi:

- Multicast : se n thread del warp accedono allo stesso elemento, l'accesso è eseguito in una singola transazione;
- Broadcast : se tutti i thread del warp accedono allo stesso elemento, l'accesso è eseguito in una singola transazione;
- Bank Conflict : se due o più thread differenti (dello stesso warp) tentano di accedere a dati differenti, residenti sullo stesso banco. Ogni conflitto viene servito e risolto serialmente.



Conflitto doppio (ottuplo): due (otto) thread differenti dello stesso warp tentano di accedere a dati differenti, residenti sullo stesso banco.

4.7.3 Lavagna tipi di accessi al banco

Memorizziamo in un array shared di 66 elementi: `__shared__ int a[66];`

Il banco sarà:

0	a(0)	a(32)	a(64)		
1	a(1)	a(33)	a(65)		
...			
...			
31	a(31)	a(63)			

Tabella 4.2:

Ricordando che siamo sempre all'interno del kernel abbiamo che:

- Broadcast: $x = a(2) + 5$; Tutti accedono alla stessa variabile: OK;
- Multicast:

```
if(threadIdx.x < 5)
    x = a(5) +1; // thread 0,1,2,3,4 accedono alla stessa variabile: OK
```

- Bank conflict: $x = a(2 * threadIdx.x) + 10$; Supponiamo che il banco seguente sia del warp 1:

Thread	Variabile	Banco
0	a(0)	0
1	a(2)	2
...
15	a(30)	30
16	a(32)	0
...
30	a(60)	28
31	a(62)	30

Tabella 4.3:

I conflitto li abbiamo sui thread 0 e 16 perchè accedono a elementi diversi ma presenti nello stesso banco. Questo succede anche ai thread 15 e 31.

Osservazioni

- Latency hiding: il ritardo che intercorre tra la richiesta avanzata dai thread alla Shared memory e l'ottenimento effettivo dei dati non è in generale un problema anche in caso di bank conflict: se vi sono molti thread in esecuzione, lo scheduler passa a un altro warp in attesa che quelli sospesi completino il trasferimento dei dati dalla shared memory. In questo modo il ritardo potrebbe essere irrilevante;
- Efficienza massima: Il modo più semplice per avere prestazioni elevate è quello di fare in modo che un warp acceda a word consecutive in memoria shared;
- Caching: con lo scheduling efficace, anche in presenza di conflitti, le prestazioni sono di gran lunga migliori rispetto a quelle in cui il cammino che i dati percorrono passa attraverso la cache L2 o peggio, la global memory.

4.8 Prodotto scalare

Implementare in cuda il prodotto scalare fra due array di float, $c = a \cdot b$, dove a e b sono array di uguali dimensioni e con N elementi (N costante definita opportunamente o letta dall'utente). Preso un vettore v come segue:

$$v = (a_1 b_1, a_2 b_2, \dots, a_n b_n)$$

abbiamo

$$c = \sum_{i=0}^n a_i$$

4.8.1 Kernel strategia 1 (banale)

```
__global__ void dotProdGPU(float *v, float *a, float *b, int N){
    //global index
    int idx=threadIdx.x+blockIdx.x*blockDim.x;
    if idx< N
        v[idx]=a[idx]*b[idx];
}
```

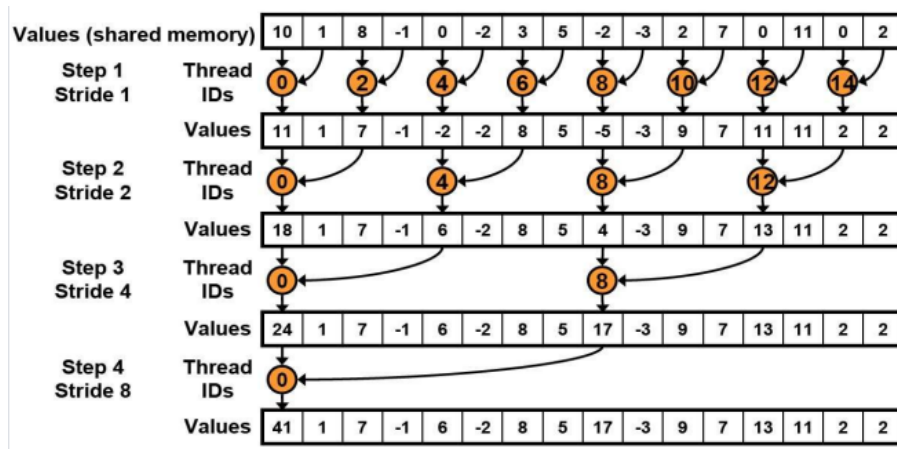
4.8.2 Come calcolare C a partire da V

Possiamo calcolare C banalmente in maniera seriale sull'host ma così facendo abbiamo un basso parallelismo, un uso solo della memoria globale (lenta) e inoltre dobbiamo trasferire V. Invece è possibile migliorare le prestazioni sfruttando il parallelismo sulla GPU, mediante la memoria condivisa.

4.8.3 Riduzione

Nel calcolo parallelo, capita frequentemente di ridurre più valori, calcolati in parallelo, in un singolo valore. Tra le possibili operazioni di riduzione sono: ADD, MUL, MIN, MAX.

Schema di riduzione in CUDA



Come possiamo vedere i thread attivi ad ogni passo dimezzano quindi questo codice è altamente divergente ottenendo prestazioni molto deludenti.

4.9 Allocazione statica

4.9.1 Esempio 1

```
__global__ void staticKernel(...){
    __shared__ int s[30];
    ...
}
```

4.9.2 Esempio 2

Nel main

```
const int n = 64;
```

Nel kernel

```
__global__ void staticKernel(...){
    __shared__ int s[n];
    ...
}
```

4.10 Allocazione dinamica

4.10.1 Esempio 3

```
__global__ void dynamicKernel(...){
    extern __shared__ int M[ ];
    ...
}
```

Chiamata nel main

```
sizeM=m*sizeof(int);
dynamicKernel<<<grid, block, sizeM>>>(...)
...
```

Nel caso di allocazione dinamica, va aggiunto un terzo parametro nella configurazione, dato dalla dimensione in byte dell'area di memoria shared (per ogni blocco di thread)

4.11 Info su shared memory

Compilando il programma precedente, otteniamo:

```
!nvcc -Xptxas -v shared.cu
ptxas info  : 0 bytes gmem
ptxas info  : Compiling entry function '_Z14dynamicReversePii' for 'sm_30'
ptxas info  : Function properties for _Z14dynamicReversePii
    0 bytes stack frame, 0 bytes spill stores, 0 bytes spill loads
ptxas info  : Used 7 registers, 332 bytes cmem[0]
ptxas info  : Compiling entry function '_Z13staticReversePii' for 'sm_30'
ptxas info  : Function properties for _Z13staticReversePii
    0 bytes stack frame, 0 bytes spill stores, 0 bytes spill loads
ptxas info  : Used 7 registers, 256 bytes smem, 332 bytes cmem[0]
```

Nessuna info a priori su memoria shared per l'allocazione dinamica

memoria **per thread** memoria shared **per blocco**

4.12 Lavagne prodotto scalare

4.12.1 Prima strategia (banale)

Abbiamo innanzi tutto due vettori:

- $a = (a_0, \dots, a_{n-1})$
- $b = (b_0, \dots, b_{n-1})$

il vettore risultante finale sarà: $c = a * b = \sum_{i=0}^{n-1} a_i b_i = \sum_{i=0}^{n-1} v_i$
con $v_0 = (a_0 b_0, \dots, a_{n-1} b_{n-1})$, infine avremo che:

1. Device: calcoliamo v

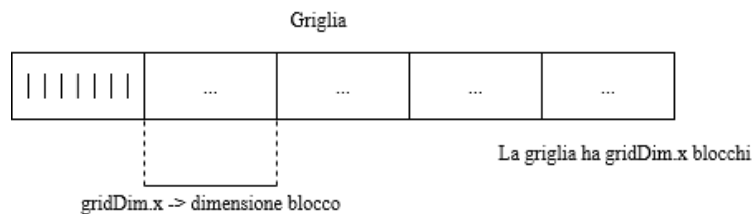
```
__global__ void dotProdGPU(float *v, float *a, float *b, int N){
    //global index
    int idx=threadIdx.x+blockIdx.x*blockDim.x;
    if idx< N
        v[idx]=a[idx]*b[idx];
}
```

2. Host: calcoliamo $\sum_{i=0}^{n-1} a_i b_i = \sum_{i=0}^{n-1} v_i \rightarrow c$

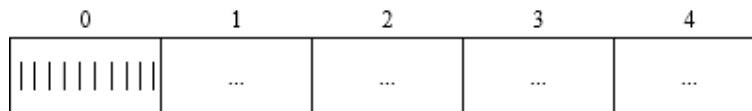
4.12.2 Seconda strategia

1. Device: calcoliamo v
2. Device: calcoliamo $\sum_{i=0}^{n-1} a_i b_i$

Immaginiamo di avere una griglia 1D:



Esempio: prendiamo come $n = 50$ e che la griglia abbia 5 blocchi con ciascuno 10 thread:



1. Ogni thread calcola una componente di v

$$2. c = \sum_{i=0}^{49} a_i b_i = \underbrace{\sum_{i=0}^9 a_i b_i}_{c_0} + \underbrace{\sum_{i=10}^{19} a_i b_i}_{c_1} + \underbrace{\sum_{i=20}^{29} a_i b_i}_{c_2} + \underbrace{\sum_{i=30}^{39} a_i b_i}_{c_3} + \underbrace{\sum_{i=40}^{49} a_i b_i}_{c_4}$$

Ogni sommatoria sarà assegnata quindi ad un blocco:

Blocchi	0	1	2	3	4
c	0	1	2	3	4

Tabella 4.4:

Per svolgere in parallelo questi calcoli abbiamo bisogno che i thread cooperino, ma coopereranno soltanto se condividono dei dati, ed è possibile solo all'interno del blocco.

Ogni blocco esegue la somma ad esso assegnato in parallelo usando la memoria condivisa.

In generale

$$c = \sum_{i=0}^{n-1} a_i b_i = \sum_{i=0}^{blockDim.x-1} a_i b_i + \sum_{i=blockDim.x}^{2blockDim.x-1} a_i b_i + \dots$$

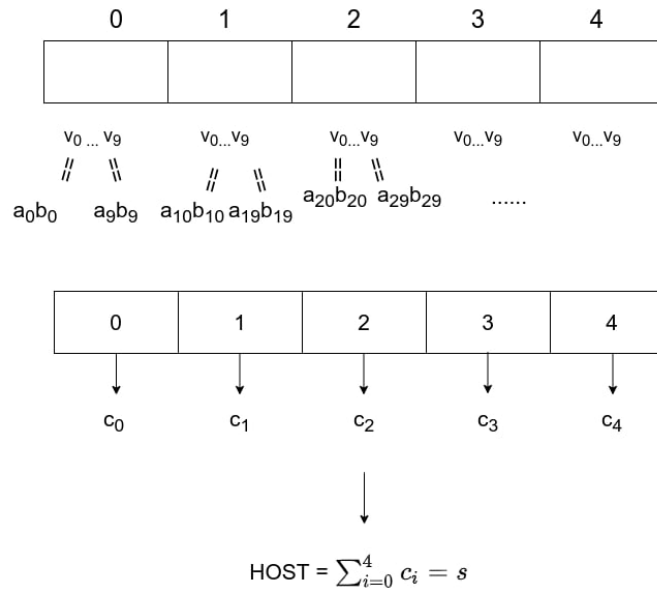
Kernel

```

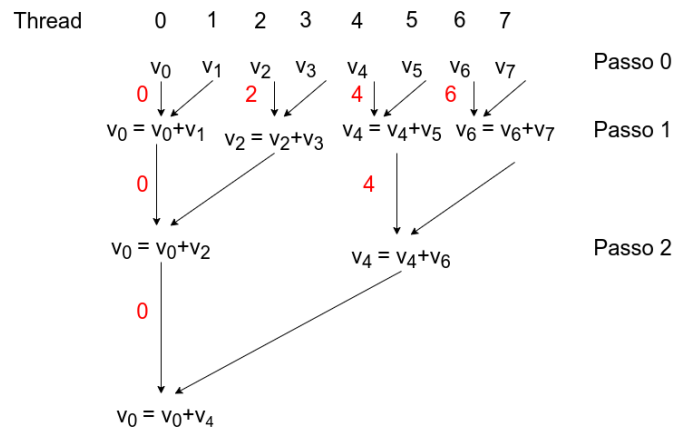
prodotto (a, b, c, n) // a e b input c e n saranno dati in output
__shared__ float v[blockDim.x];
index = threadIdx.x + blockIdx.x*blockDim.x;
id = threadIdx.x;
v(id) = a(index) * b(index) // a v e' associata una coppia di elementi di a e b
sincronizzazione
somma in parallelo  $\sum_{i=0}^{blockDim.x} v_i \rightarrow Sum$ 
sincronizzazione
c[blockIdx.x] = Sum
END

```

Dopo l'esecuzione del codice Kernel l'host effettua una memmcopy ed esegue $\sum_{i=0}^{gridDim.x-1} c_i$



Somma parallela all'interno del blocco



$p = \log_2 blockDim.x$ è il numero di passi.

Ad ogni passo K con $k = 0, 1, \dots, p-1$:

- $Dist = 2^k, id = threadIdx.x$
- Se $resto(id, 2^{k+1}) = 0$

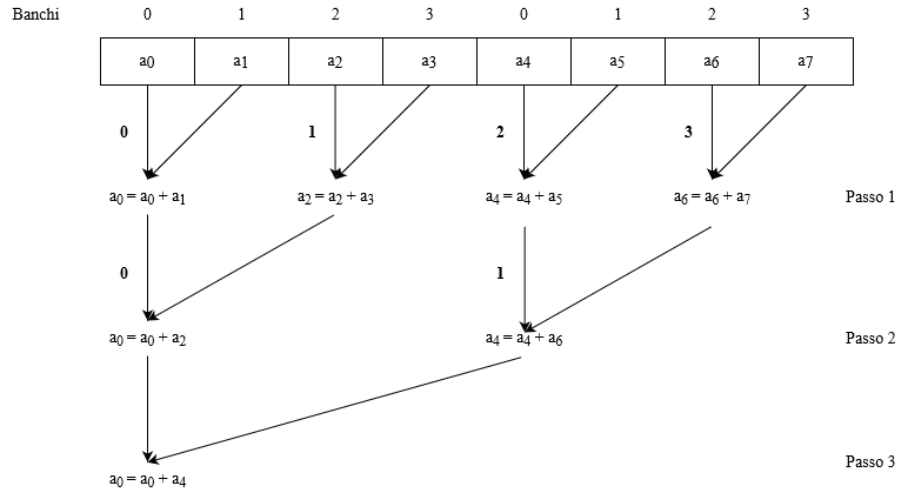
$$v(id) = v(id) + v(id + Dist)$$
- Sincronizzazione

```
if(id = 0) //per via della memoria condivisa conviene farlo fare da un solo thread
    c(blockIdx.x) = v(0)
```

Nella seconda strategia ad ogni passo si crea della divergenza, perchè abbiamo thread dello stesso warp che lavorano e altri che non eseguono nessun compito. Prestazioni non ottimali.

Seconda strategia bank conflict

Iporizziamo di avere un Warp di 4 thread e una memoria condivisa con 4 banchi, infine supponiamo di avere un blocco composto da 8 thread:



La situazione che abbiamo è la seguente:

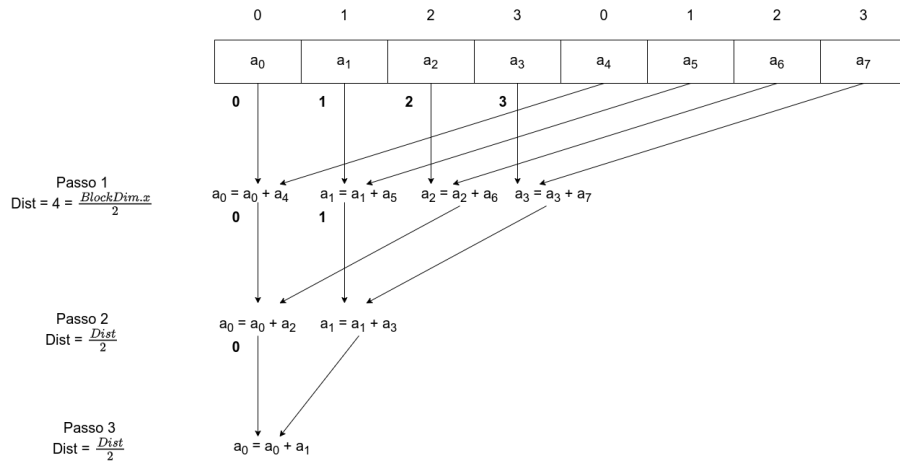
Passo	Thread	Variabile	Banco	
1	0	$a(0)$	0	Conflitto doppio (accedono a elementi diversi dello stesso banco)
		$a(1)$	1	
	2	$a(4)$	0	
		$a(5)$	1	
	1	$a(2)$	2	Conflitto doppio
		$a(3)$	3	
	3	$a(6)$	2	
		$a(7)$	3	
2	0	$a(0)$	0	Conflitto doppio
		$a(2)$	2	
	1	$a(4)$	0	
		$a(6)$	2	

Tabella 4.5:

Possiamo osservare quindi diversi conflitti di accesso ai banchi. Quindi presentiamo ora una nuova strategia senza divergenze (o almeno usciranno da un certo punto in poi) e senza conflitti di banco.

4.12.3 Terza strategia

Supponiamo sempre di avere $\text{blockDim.x} = 8$, 4 banchi e 4 thread.



Dallo schema notiamo che nel primo warp tutti i thread da 0 a 3 lavorano, mentre dal warp 2 in poi nessun thread lavora. Analizziamo poi l'eventuale presenza di conflitti:

Passo	Thread	Variabile	Banco	
1	0	$a(0)$	0	Nessun conflitto
		$a(4)$	0	
	1	$a(1)$	1	
		$a(5)$	1	
	2	$a(2)$	2	
		$a(6)$	2	
2	3	$a(3)$	3	Nessun conflitto
		$a(7)$	3	
	0	$a(0)$	0	
		$a(2)$	2	
	1	$a(1)$	1	Nessun conflitto
		$a(3)$	3	

Tabella 4.6:

Per la distanza poniamo inizialmente $\text{Dist} = \frac{\text{BlocchiDim.x}}{2}$ per poi nei passi successivi fare $\text{Dist} = \frac{\text{Dist}}{2}$. $p = \log_2 \text{blockDim.x}$ è il numero di passi e blockDim.x potenza di 2. Ad ogni passo K con $k = 0, 1, \dots, p-1$:

- $\text{Dist} = \frac{\text{Dist}}{2}, id = \text{threadIdx.x}$
- if $id < \text{Dist}$
 $v(id) = v(id) + v(id + \text{Dist})$
- Sincronizzazione

```
if(id = 0) //per via della memoria condivisa conviene farlo fare da un solo thread
    c(blockIdx.x) = v(0)
```

Capitolo 5

Sistemi di Raccomandazione

5.1 Big Data: cosa significa?

Insieme di dati la cui grandezza e complessità richiedono metodi specifici per l'immagazzinamento, l'analisi e l'estrazione dei valori.

I Big Data vengono descritti secondo il modello delle 4 V basato sulla tassonomia di Laney del 2001.

Il modello delle 4 V è dato da:

- **Volume:** ordine degli zettabyte;
- **Variety:** dati molto disomogenei e di varia natura;
- **Velocity:** i dati nascono e vengono acquisiti sempre più rapidamente;
- **Veracity:** In rete, non tutto è attendibile.

Al giorno d'oggi i Big Data provengono da varie fonti tra cui: Web e social, Macchine, Sensori, Transazioni, Internet of Things.

5.2 Sistemi di raccomandazione: conoscere per suggerire

Problema: compiere scelte all'interno di una grossa mole di dati, l'abbondanza di dati genera il "paradosso della scelta".

Un Sistema di raccomandazione orienta le nostre scelte nella grande mole di dati di cui dispone la rete, raccomanda oggetti di nostro gradimento (e-commerce), suggerisce utenti o pagine da seguire in un social network utilizzando una strategia win-win

5.2.1 Elementi di un sistema di raccomandazione

- **Item:** sono gli oggetti che ci vengono suggeriti dal sistema e vengono rappresentati con un set di attributi e proprietà;
- **User:** sono gli utenti di un sistema di raccomandazione;
- **Rating:** è la valutazione che l'utente fornisce del prodotto considerato. Può essere un numero o una valutazione binaria (sì/no, like/nessun like, pollice in su o in giù).

5.2.2 Tipi di sistemi di raccomandazione

Le principali classi di sistemi di raccomandazione sono

- **Sistemi Content-Based (CB):** utilizzano gli attributi dei prodotti che gli utenti hanno votato in passato per calcolare la similarità (similitudine tra classi del prodotto) con altri item. Ad es., se un utente ha valutato positivamente un film del genere commedia, gli verrà suggerito un altro film dello stesso genere;
- **Sistemi Collaborative-Filtering (CF):** raccomandano agli utenti prodotti apprezzati da altri utenti ritenuti simili. Analogamente si può tenere conto della similarità (similarità tra prodotti con le stesse valutazioni) tra prodotti. Questi sistemi si sono molto diffusi grazie al Netflix Prize.

Sistemi di raccomandazione meno diffusi o utilizzati in maniera ibrida insieme ai precedenti sono:

- **Sistemi demografici:** forniscono raccomandazioni in base al profilo demografico dell'utente;
- **Sistemi community-based:** forniscono raccomandazioni in base alle preferenze degli amici dell'utente. Difatti, le persone tendono ad apprezzare maggiormente le raccomandazioni di amici rispetto a quelle di utenti simili, ma sconosciuti. Inoltre tali sistemi si sono molto diffusi grazie al successo dei social-network.

5.2.3 Problemi dei sistemi di raccomandazione

- **Cold Start Problem:** causato dall'incapacità di gestire i nuovi prodotti o utenti che vengono inseriti nel sistema. Ad es. nei sistemi content-based occorre avere una conoscenza dei gusti dell'utente; nei sistemi collaborative-filtering non si può stabilire la similarità di nuovi utenti e/o prodotti;
- **Scalabilità:** spesso che è necessaria una grande potenza di calcolo per effettuare previsioni in tempo reale, a causa della grande mole di dati da analizzare.
- **Sparsità:** poiché i prodotti sono moltissimi, normalmente anche gli utenti più attivi avranno valutato solo una minuscola percentuale di essi.

5.2.4 Sistemi collaborative-filtering

In base all'algoritmo utilizzato, i sistemi CF si dividono a loro volta in:

- **Sistemi Model-based:** si basano su un modello (statistico) che permette al sistema di imparare a riconoscere schemi complessi su un set di dati di training;
- **Sistemi Memory-based:** si basano sul concetto di similarità tra utenti, cioè si assume che ogni utente faccia parte di un gruppo di persone con interessi simili. Analogamente, possono basarsi sul concetto di similarità tra prodotti.

5.2.5 Formalizziamo il problema

Formalizzazione matematica del problema della raccomandazione

- U insieme di utenti
- I insieme di item
- $r : U \times I \rightarrow R$ funzione di rating

$\forall u \in U$, determinare $i_i \in I$ tale che $i_u = \operatorname{argmax}_i (u, i), i \in I$

La matrice di Rating $R : r_{kl} = (u_k, i_l), u_k \in U, i_l \in I$

è la valutazione che l'utente u_k ha assegnato all'item i_l . E' una matrice molto sparsa, infatti la maggior parte dei rating non è nota e quindi vanno stimati. Quindi si cerca di mutuare le informazioni mancanti da quelli note per utenti dai gusti simili.

Principalmente per risolvere questo problema, si impiegano le Decomposizioni di matrici.

5.2.6 Similarità

Nei metodi model e memory-based si parla di similarità tra utenti o prodotti, con riferimento alle valutazioni. Ci sono varie misure di similarità, le più usate sono:

- Correlazione di Pearson

$$\operatorname{sim}(a, b) = \frac{\sum_{p \in I_{a,b}} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in I_{a,b}} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in I_{a,b}} (r_{b,p} - \bar{r}_b)^2}}$$

con $r_{a,p}$ = voto dell'utente a al prodotto p , \bar{r}_a = voto medio dell'utente a , $I_{a,b}$ = insieme dei prodotti votati sia da a che da b , $a, b \in U$.

- Coseno di similarità

$$\operatorname{sim}(a, b) = \frac{\vec{r}_a \cdot \vec{r}_b}{\|\vec{r}_a\|_2 \cdot \|\vec{r}_b\|_2} = \frac{\sum_{p \in I_{a,b}} r_{a,p} r_{b,p}}{\sqrt{\sum_{p \in I_{a,b}} r_{a,p}^2} \sqrt{\sum_{p \in I_{a,b}} r_{b,p}^2}}$$

\vec{r}_a ed \vec{r}_b hanno dimensione 2, il coseno di similarità corrisponde al coseno dell'angolo formato da essi.

- Distanza euclidea

$$\operatorname{sim}(a, b) = \sqrt{\sum_{p \in I_{a,b}} (r_{a,p} - r_{b,p})^2}$$

\vec{r}_a (\vec{r}_b) = vettore delle valutazioni dell'utente a (b) $\in I_{a,b}$

Corrisponde alla distanza euclidea tra i due vettori \vec{r}_a ed \vec{r}_b

5.2.7 SR Memory based di tipo K-NN

Essi forniscono un'approssimazione \hat{R} della matrice di rating R mediante la tecnica k-NN: k-Nearest Neighbours. Sono di due tipi:

- **User-based**

$$\hat{r}_{ki} = \mu_k + \frac{\sum_{v \in N_k^n(i)} \text{sim}(k,v)(r_{vi} - \mu_v)}{\sum_{v \in N_k^n(i)} \text{sim}(k,v)}$$

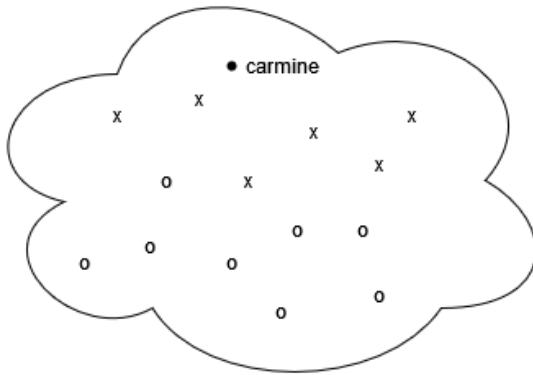
$N_k^n(i)$ = n utenti più simili a k che hanno espresso il voto per il prodotto i,
 μ_k = voto medio dell'utente k.
- **Item-based**

$$\hat{r}_{ui} = \mu_i + \frac{\sum_{j \in N_i^n(u)} \text{sim}(i,j)(r_{uj} - \mu_j)}{\sum_{j \in N_i^n(u)} \text{sim}(i,j)}$$

$N_i^n(u)$ = n prodotti più simili a i che sono stati votati dall'utente u
 μ_i = voto medio del prodotto i.

Lavagna sistema di raccomandazione K-NN

Supponiamo di avere degli utenti:



Quali sono gli utenti più simili?

- indichiamo con "x" i 6 utenti più simili;
- indichiamo con "o" i 7 utenti diversi

A priori sono già stati calcolati gli indici di similarità.

$$\hat{R}_{carmine,i} = \bar{R} + \frac{\sum_x R_{x,i} \cdot \text{sim}(x, \text{carmine})(r_{x,i} - \mu_x)}{\sum_x \text{sim}(x, \text{carmine})}$$

- $\hat{R}_{carmine,i}$: stima della valutazione di carmine per il prodotto i;
- \bar{R} : voto medio;
- $\text{sim}(x, \text{carmine})$: se x è più simile a carmine il suo voto avrà più peso, altrimenti avrà meno peso.

5.2.8 Sistemi di raccomandazione Model-Based

Sono basati sulla fattorizzazione di matrici. Infatti approssimano la matrice di rating R , di dimensione $n_u \times n_v$, con la matrice

$$\hat{R} = U \times V, \text{ ossia } \hat{r}_{ki} = u_k \times v_i \text{ (prodotto scalare)}$$

$$\begin{bmatrix} \dots \\ u_k \\ \dots \end{bmatrix}_{n_u \times d}, V = [\dots v_i \dots]_{d \times n_v}$$

- u_k vettore riga che rappresenta d fattori latenti* dell'utente k
- v_i vettore colonna che rappresenta d fattori latenti del prodotto i.

* n.b. sono caratteristiche virtuali, non corrispondenti a caratteristiche reali.

5.2.9 SR model- based

In genere, si desidera trovare U e V tale che sia minimo $\|\hat{R} - R\|^2$, cioè

$$\min_{U,V} \sum_{(k,i) \in K} (r_{ki} - \hat{r}_{ki})^2 = \sum_{(k,i) \in K} (r_{ki} - u_k \cdot v_i)^2 \quad [*]$$

con k = insieme per cui le valutazioni sono note. Quindi diciamo che per misurare la distanza tra la matrice \hat{R} e R usiamo la norma quadratica.

Lavagna SR Model-Based

La decomposizione della matrice R in $U * V$ è usata per ridurre il numero delle incognite e quindi ridurre la dimensione del problema.

$$R = \begin{bmatrix} \dots & \dots & \dots \\ \dots & r_{ki} & \dots \\ \dots & \dots & \dots \end{bmatrix}_{n_u * n_v}$$

- r_{ki} : valutazione dell'utente k per il prodotto i;
- con ; n_u : # utenti;
- n_v : # prodotti.

Il # incognite si calcola con $n_u * n_v$, se prendiamo per esempio $n_u = 100$ e $n_v = 10000$ otteniamo così 1 milione.

Decomponiamo ora R nel seguente modo: $\hat{R}_{n_u \times n_v} = U_{n_u \times d} * V_{d \times n_v}$, d è un valore arbitrario.

Otteniamo così che $r_{ki} = \sum_{l=1}^d u_{kl} * v_{li} = u_k + v_i$ Quindi abbiamo che U e v saranno composte nel seguente modo:

$$\bullet U = \begin{bmatrix} u_{1,1} & \dots & u_{1,d} \\ \dots & \dots & \dots \\ u_{n_u,1} & \dots & u_{n_u,d} \end{bmatrix} = \begin{bmatrix} u_1 \\ \dots \\ u_k \\ \dots \\ u_{n_u} \end{bmatrix} \rightarrow \text{vettore riga}$$

$u_k(u_{k,1}, \dots, u_{k,d})$ vettore dei fattori latenti dell'utente k;

$$\bullet V = \begin{bmatrix} v_{1,1} & \dots & v_{1,n_v} \\ \dots & \dots & \dots \\ v_{d,1} & \dots & v_{d,n_v} \end{bmatrix} = [v_1 \quad \dots \quad v_i \quad \dots \quad v_{n_v}] \rightarrow \text{vettore colonna}$$

$$v_i = \begin{bmatrix} v_{1,i} \\ \dots \\ v_{d,i} \end{bmatrix} \text{ vettore colonna dei } d \text{ fattori latenti associati al prodotto } i.$$

Verifichiamo ora se la dimensione del problema si sia ridotta, calcoliamo il numero di incognite come segue: $d(nU + nv) = 10(100 + 10000) = 10(10100) = 101000$. Abbiamo ridotto di molto la dimensione del problema è stata ridotta, ma vale solo d piccolo. QUindi possiamo dire che per d piccolo abbiamo che $d(nU + nv) \ll nU * nV$. Il nostro obiettivo è quello di trovare un \hat{R} molto vicina a R .

5.2.10 Overfitting dei dati

Il problema posto in questi termini potrebbe soffrire dell'overfitting dei dati, cioè avere più parametri liberi dei dati a disposizione, ossia delle condizioni da imporre. Per questo si ricorre ad una regolarizzazione del problema:

$$\min_{U,V} \sum_{(k,i) \in K} |r_{ki} - u_k \cdot v_i|^2 + \lambda (\|U\|_F + \|V\|_F) [**]$$

$$\text{con } \|U\|_F = \|U\|_2 = \text{norma di Frobenius} = \sqrt{\sum_{i,j} (u_{ij})^2} \text{ e}$$

λ parametro di controllo ($0 < \lambda < 1$, di solito $\lambda < 0.2$). In tal modo si preferiscono le soluzioni con norma piccola.

5.2.11 Metodi per trovare U e V

I metodi più diffusi, con molteplici varianti, sono:

- Discesa del gradiente stocastico;
- Metodo dei minimi quadrati alternati.

Discesa gradiente stocastico

E' un metodo iterativo. Ad ogni iterazione aggiorna u_k e v_i nel seguente modo:

$$\begin{cases} u_k = u_k + 2\eta e_{ki} x v_i \\ v_i = v_i + 2\eta e_{ki} x u_k \end{cases} \quad \text{con } (k,i) \in K$$

$$e_{ki} = r_{ki} - \hat{R}_{ki} - u_k x v_i$$

- $\nabla e_{ki}^2 = -(2e_{ki} x v_i, 2e_{ki} x u_k)$. Quindi u_k e v_i vanno nella direzione opposta al grandiente ∇e_{ki}^2 , ossia nel verso in cui e_{ki}^2 diminuisce.
- **Learning rate** $\nabla \in (0,002)$. Impedisce che la soluzione vari troppo da un'iterazione e l'altra, in modo da evitare di «perdere» il minimo.

Lavagna gradiente stocastico

Se abbiamo $f(x,y) = x^2 + 3y$ sappiamo che:

- $\frac{\delta f}{\delta x} =$ derivata di f rispetto a x considerando y come costante $= 2x + 0 = 2x$;
- $\frac{\delta f}{\delta y} =$ derivata di f rispetto a y considerando x come costante $= 0 + 3 = 3$;

Il gradiente di una funzione quindi sarà: $\nabla f =$ gradiente di $f = (\frac{\delta f}{\delta x}, \frac{\delta f}{\delta y}) = (2x, 3)$, questo ci dà la direzione in cui f cresce. **Nota: ∇ si legge nablà.**

Quindi ora possiamo dire che l'errore $= e_{ki} = r_{ki} - \hat{r}_{ki} = r_{ki} - u_n * v_i$.

∇e_{ki}^2 sarà:

- $\frac{\delta e_{ki}^2}{\delta u_n} = 2e_{ki} * \frac{\delta e_{ki}}{\delta u_n} = 2e_{ki} * (-v_i) = -2e_{ki} * v_i$;
- $\frac{\delta e_{ki}^2}{\delta v_i} = 2e_{ki} * \frac{\delta e_{ki}}{\delta v_i} = 2e_{ki} * (-u_n) = -2e_{ki} * u_n$;

dopo \rightarrow learning rate Dal grafico seguente possiamo notare che alla variazione dei passi cioè η cambia di quando ci avviciniamo al valore minimo che cerchiamo.



Minimi quadrati alternati

E' un metodo iterativo così schematizzabile:

1. Inizializzo $V = \begin{bmatrix} \text{valutazioni medie prodotti} \\ \dots \\ \text{numeri casuali piccoli} \\ \dots \end{bmatrix}$, $v_i(i) =$ valutazione media prodotto i
2. Fissato V , trovo U che minimizza $[*]$ o $[**]$
3. Fissato U , trovo V che minimizza $[*]$ o $[**]$
4. Ripeto 2 e 3 fino a convergenza.

5.2.12 Metodi basati sulla SVD

SVD si basa sulla decomposizione ai valori singolari. Le sue proprietà sono che i valori singolari sono unici, le matrici U e V che useremo non sono univocamente determinate, e questo comporta che la decomposizione a valori singolari non è unica.

lavagna decomposizione SVD

Data una matrice A reale di dimensione $m \times n$, esistono tre matrici U, Σ e V tali che: $A = U\Sigma V^T$

$$\underset{m \times n}{A} = \underset{m \times m}{U} \underset{m \times n}{\Sigma} \underset{n \times n}{V^T}$$

- U è una matrice ortogonale (cioè $UU^T = U^TU = I$) di dimensione $m \times m$;
- Σ è una matrice 'diagonale' di dimensioni $m \times n$. Gli elementi di Σ sono elementi diagonali, detti valori singolari, sono $\sigma_1 \geq \sigma_2 \geq \dots \sigma_p \geq 0$, $comp = \min(m, n)$;
- V^T è la trasposta di una matrice ortogonale V di dimensioni $n \times n$.

Consideriamo:

- ($m > n$) con $\Sigma = \begin{bmatrix} \sigma_1 & \dots & 0 \\ 0 & \ddots & 0 \\ \dots & \dots & \sigma_n \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 \end{bmatrix}$, dove $\begin{bmatrix} \sigma_1 & \dots & 0 \\ 0 & \ddots & 0 \end{bmatrix}$ ha n righe e $\begin{bmatrix} \dots & \dots & \sigma_n \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 \end{bmatrix}$ ha m - n righe
- ($m < n$) con $\Sigma = \begin{bmatrix} \sigma_1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n & 0 & \dots & 0 \end{bmatrix}$

Se approssimiamo in questo modo notiamo che A ha dimensione $m \times n$ elementi, mentre la nostra decomposizione $m^2 + \min(m, n) + n^2$ elementi, ma questo non porta a nessuna riduzione.

5.2.13 SVD Troncata

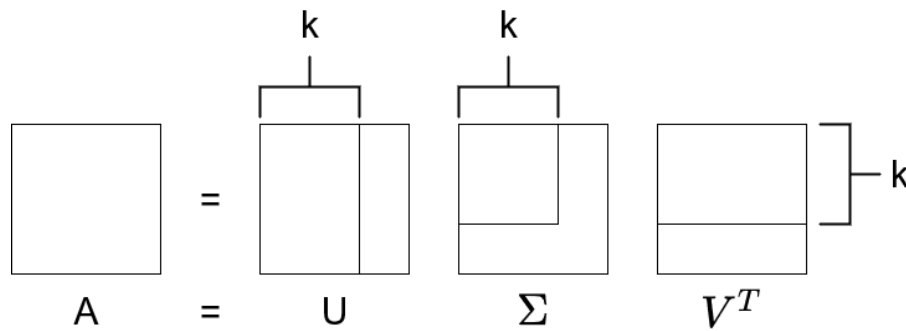
E' l'approssimazione della matrice di partenza con una di rango inferiore (SVD troncata), ottenuta utilizzando un numero minore di valori singolari.

Lavagna

Sia A una matrice $m \times n$ e sia $A = U\Sigma V^T$ la sua decomposizione ai valori singolari, con $\sigma_1 \geq \sigma_2 \geq \dots \sigma_r \geq \sigma_{r+1} = \dots = \sigma_p = 0$. sia $k \leq r$ e sia A_k la matrice $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$

$$\underset{m \times n}{A} = \underset{m \times m}{U} \underset{m \times n}{\Sigma} \underset{n \times n}{V^T} = \begin{bmatrix} \vdots & & & & \\ & \ddots & & & \\ & & \sigma_1 & \dots & \\ & & \vdots & \ddots & \\ & & & & \sigma_k & \dots \\ & & & & & \ddots \\ & & & & & & \sigma_n & \dots \\ & & & & & & & \ddots \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_k & & \\ & & & \ddots & \\ & & & & \sigma_n \end{bmatrix} \begin{bmatrix} \vdots & & & & \\ & \ddots & & & \\ & & \sigma_1 & \dots & \\ & & \vdots & \ddots & \\ & & & & \sigma_k & \dots \\ & & & & & \ddots \\ & & & & & & \sigma_n & \dots \\ & & & & & & & \ddots \end{bmatrix}$$

Per ridurre le dimensioni del problema e creare così un'approssimazione di A decidiamo di tagliare le matrici nel seguente modo:



Prendiamo quindi k elementi da ogni matrice: $A \approx A_k = U_k \Sigma_k V_k^T$ la dimensione del problema quindi ora è: $m \times n \gg m \times k + k + k \times n$ se $k \ll m, n$. Riduciamo di molto quindi le dimensioni del problema.

Ma A_k approssima bene A ? l'errore commesso si calcola facendo la norma 2:

$\|A - A_k\|_2 = \min_{B \text{ di rango } k} \|A - B\|_2 = \sigma_{k+1}$, sappiamo quindi misurare l'errore commesso e questo è il $k+1$ elemento non preso. Infine possiamo dire che A_k è la migliore approssimazione possibile tra tutte le matrici di rango k (che hanno k righe o colonne linearmente indipendenti).

La matrice ottenuta tramite l'applicazione della SVD troncata è molto meno complessa della matrice di rating di partenza, e ci permette di calcolare velocemente la vicinanza di un oggetto/servizio da raccomandare ad un utente.

Esempio SR basato su SVD troncata

Quando utilizziamo la SVD troncata, approssimiamo la matrice di rating osservando solo le caratteristiche più importanti, ossia quelle corrispondenti ai valori singolari più grandi. Nell'esempio seguente, calcoliamo U , V , and Σ (mediante alcune routine di algebra lineare, in MATLAB o Python ad esempio) ma conserviamo solo le due caratteristiche (fattori) principali, prendendo solo le prime due colonne di U e di V .

$$\text{SVD: } M_k = U_k \times \Sigma_k \times V_k^T$$

U_k	Dim1	Dim2
Alice	0.47	-0.30
Bob	-0.44	0.23
Mary	0.70	-0.06
Sue	0.31	0.93

V_k^T	Terminator	Die Hard	Twins	Eat Pray Love	Pretty Woman
Dim1	-0.44	-0.57	0.06	0.38	0.57
Dim2	0.58	-0.66	0.26	0.18	-0.36

Σ_k	Dim1	Dim2
Dim1	5.63	0
Dim2	0	3.23

Previsione: $\hat{r}_{ui} = \bar{r}_u + U_k(\text{Alice}) \times \Sigma_k \times V_k^T(\text{EPL})$
 $= 3 + (0.47 \times 5.63 \times 0.38 - 0.3 \times 3.23 \times 0.18)$
 $= 3 + 0.84 = 3.84$

Spiegazione

$$\hat{R} = \bar{R} + U_k \Sigma_k V_k^T$$

Dove \hat{R} è la matrice delle valutazioni medie, $\hat{R} = \hat{r} e_n^T$ con $\hat{r}_u = \text{val media dell'utente } u$, $(e_n)_i = 1$

- Ricordiamo $(U_k \Sigma_k V_k^T)_{ui} = \sum_{l=1}^k \sigma_l (u_l)_u (v_l^T)_i$
- Percio, se $u = \text{Alice}$ e $i = \text{Eat Pray Love (EPL)}$:
 $\hat{r}_{ui} = \bar{r}_{\text{Alice}} + \sigma_1 (u_1)_{\text{Alice}} (v_1^T)_{\text{EPL}} + \sigma_2 (u_2)_{\text{Alice}} (v_2^T)_{\text{EPL}}$

Osservazioni sulla riduzione della dimensionalità del problema

- Fattorizzazione delle matrici: Si genera una approssimazione di rango basso della matrice poi si determinano i fattori latenti e poi la proiezione di utenti e prodotti nello spazio k-dimensionale;
- La qualità della previsione può decrescere perchè le valutazioni originali non sono state considerate;
- Le previsioni potrebbero migliorare filtrando un certo rumore nei dati o trovando alcune correlazioni nei dati;
- Il parametro da scegliere per la giusta riduzione della matrice dipende dal numero dei valori singolari nell'approccio SVD. I parametri possono essere determinati e affinati sulla base di esperimenti in un certo dominio. Alcuni autori suggeriscono di considerare tra 20 e 100 fattori per ottenere buone stime

5.2.14 SVD compressione immagini come scegliere K

Nella compressione delle immagini con SVD per scegliere il parametro k si applica la regola del pollice.

Applicando la regola del pollice si conserva l'80-90% dell'«energia» totale: $\frac{\sum_{j=1}^k \sigma_j^2}{\sum_{j=1}^n \sigma_j^2}$

Capitolo 6

MPI

6.1 Funzionalità della libreria MPI

- Funzioni di management della comunicazione
 - Definizione/ identificazione di gruppi di processi(task) coinvolti nella comunicazione;
 - Definizione /gestione dell'identità del singolo processo all'interno del gruppo.
- Funzioni di scambio di messaggi
 - Inviare/ ricevere dati da un processo;
 - Inviare/ ricevere dati da un gruppo di processi.
- Definizione di costanti e valori di default;
- Unità fondamentale è il processo: il programma parallelo è costituito da una collezione di processi (task) autonomi:
 - ogni processo esegue un proprio algoritmo utilizzando i dati residenti nella propria memoria;
 - quando è necessario comunica con gli altri processi attraverso uno scambio di messaggi;

In ambiente MPI un programma è visto come un insieme di componenti (o processi) concorrenti. Il sistema lancia n processi e ogni processo esegue una copia di prog.exe e ha il suo spazio di memoria locale.

6.1.1 Le funzioni MPI

MPI è una libreria che comprende:

- Funzioni per definire l'ambiente;
- Funzioni per comunicazioni uno a uno;
- Funzioni per comunicazioni collettive
- Funzioni per operazioni collettive;
- Tutte le funzioni MPI iniziano con MPI_ seguito da una maiuscola e da minuscole;
- Header file per un programma C `# include<mpi.h>`

Tutte le routine MPI restituiscono un indicatore di errore: `error = MPI_Xxxx(parametri,...)`
Questo indicatore di errore può per esempio assumere uno di questi valori:

- `MPI_SUCCESS`, nessun errore;
- `MPI_ERR_ARG`, argomento errato;
- `MPI_ERR_RANK`, identificativo errato.

6.1.2 Struttura di un programma MPI

```
#include <stdio.h>
#include <mpi.h>
main(int argc, char *argv[]){
    int me, nproc;
    ::ulteriori dichiarazioni di variabili

    MPI_Init(&argc,&argv);
    MPI_Comm_size (MPI_COMM_WORLD,&nproc);
    MPI_Comm_rank (MPI_COMM_WORLD, &me);
    ::corpo del programma

    MPI_Finalize();
    return 0;
}
```

MPI_Init(&argc,&argv): `argc` e `argv` sono gli argomenti del `main`.

Questa routine deve essere chiamata prima di ogni altra routine MPI. Serve per inizializzare l'ambiente di esecuzione di MPI, definire l'insieme dei processi attivati (contesto) e inizializzare il contesto `MPI_COMM_WORLD`

MPI_Comm_rank(comm, &me): Questa routine assegna ad ogni processo del contesto un proprio numero identificativo.

- Input: `MPI_Comm comm`: contesto a cui appartiene il processo (`MPI_COMM_WORLD` è il nome del contesto globale)
- Output: `int me` : identificativo assegnato al processo

MPI_Comm_size(comm, &nproc): numero dei processi del communicator

- Input: `MPI_Comm comm`: contesto a cui appartiene il processo;
- Output: `int nproc`: numero di processi del contesto.

MPI_Finalize(): Termina un programma MPI.

6.2 Comunicazione uno a uno

6.2.1 Spedizione

La spedizione di un messaggio avviene con la seguente routine:

MPI_Send(msg, count, datatype, dest, tag, comm)

- void *msg indirizzo del primo elemento da spedire
- int count numero di elementi da spedire;
- MPI_Datatype datatype tipo degli elementi da spedire;
- int dest identificativo del destinatario del messaggio;
- int tag identificativo del messaggio;
- MPI_Comm comm contesto a cui appartengono i processi.

6.2.2 Ricezione

La ricezione di un messaggio avviene con la seguente routine:

MPI_Recv(msg, count, datatype, source, tag, comm, &status)

- void *msg : indirizzo del primo elemento da ricevere;
- int count: numero di elementi da ricevere (consecutivi);
- MPI_Datatype datatype: tipo degli elementi da ricevere;
- int source: identificativo del mittente del messaggio. Può assumere il valore MPI_ANY_SOURCE per ricevere da qualsiasi mittente;
- int tag: identificativo del messaggio. Può assumere il valore MPI_ANY_TAG per ricevere qualsiasi messaggio;
- MPI_Comm comm: contesto a cui appartengono i processi;
- MPI_Status status: valore che permette di conoscere il mittente (se nascosto in MPI_ANY_SOURCE), il TAG e la dimensione esatta di un messaggio

NOTA: si può effettuare anche una spedizione e una ricezione in un'unica chiamata

MPI_Sendrecv(sendmsg, sendcount, sendtype, dest, sendtag, recvmsg, recvcount, recvtype, source, recvtag, comm, &status);

6.3 Comunicazioni collettive

6.3.1 Broadcast

È possibile trasmettere un messaggio da un processo a tutti gli altri processi di un contesto:

MPI_Bcast(msg, count, datatype, root, comm);

int ROOT: identificativo del mittente

6.3.2 Scatter

È possibile trasmettere messaggi diversi da un processo a tutti i processi del contesto.

MPI_Scatter(sendmsg, sendcount, sendtype, recvmsg, recvcount, recvtype, root, comm);

- send_data : array di dati residente nel processo root di tipo send_data_datatype;
- send_count : numero elementi inviati ad ogni processo. Se send_count =2 , allora il processo 0 riceve i primi due elementi dell'array, il processo 1 riceve il terzo e il quarto elemento dell'array e così via. Di solito, send_count = (dimensione totale dell'array)/(numero dei processi);

- `recv_data` buffer di dati che può contenere `recv_count` elementi di tipo `recv_datatype`;
- `root` : processo che sta inviando;
- `communicator` : il contesto nel quale risiedono `root` e i processi che ricevono.

6.3.3 Gather

È possibile trasmettere un messaggio da ogni processo di un contesto ad un solo processo del contesto.

`MPI_Gather(sendmsg, sendcount, sendtype, recvmsg, recvcount, recvtype, root, comm);`

`int ROOT`: identificativo del destinatario

6.3.4 AllGather

Per trasmettere un messaggio da ogni processo di un contesto a tutti i processi del contesto.

`MPI_Allgather(sendmsg, sendcount, sendtype, recvmsg, recvcount, recvtype, comm);`

6.4 Operazioni collettive

6.4.1 Reduce

Operazioni utilizzate per ridurre un insieme di numeri con una funzione di somma in parallelo, non sappiamo quale strategia venga usata. **`MPI_Reduce(sendmsg, recvmsg, count, type, op, root, comm);`**

`MPI_Allreduce(sendmsg, recvmsg, count, type, op, comm);`

Capitolo 7

CUDA

Compilazione: \$ nvcc -o provaMemcpy provaMemcpy.cu

Esecuzione: \$./ provaMemcpy

7.0.1 Allocazione della memoria sulla GPU

- `cudaError_t cudaMalloc (void ** devPtr, size_t size);`
 - `devPtr` è un puntatore all'area di memoria da allocare sul device
 - `size` è la dimensione in bytes dell'area da allocare

7.0.2 Deallocazione della memoria sulla GPU

- `cudaError_t cudaFree (void * devPtr);`
 - `devPtr` è un puntatore all'area di memoria del device da deallocare

7.1 Scambio dei dati fra CPU e GPU

- `cudaError_t cudaMemcpy (void * dest, void * src, size_t nBytes, enum cudaMemcpyKind);`
 - `dest` è un puntatore all'area di memoria in cui effettuare la copia
 - `src` è un puntatore all'area di memoria da copiare
 - `nBytes` è il numero di byte da copiare
 - `kind` indica la direzione della copia; è una variabile enumerativa, che può assumere questi valori:
 - `cudaMemcpyHostToHost` : dall'host all'host
 - `cudaMemcpyHostToDevice` : dall'host al device
 - `cudaMemcpyDeviceToHost` : dal device all'host
 - `cudaMemcpyDeviceToDevice` : dal device al device

Questa funzione è bloccante: non inizia prima che siano completate tutte le CUDA calls precedenti e non termina se la copia non è completa.

7.2 Somma di due vettori

Per calcolare la somma di due vettori useremo un kernel e una funzione:

- kernel sommaGPU: calcola in parallelo la somma di due vettori;
- function sommaCPU : calcola in seriale la somma di due vettori.

I risultati ottenuti in sequenziale e in parallelo verranno poi confrontati per testare la validità del kernel.

7.3 Le funzioni CUDA

Le function CUDA hanno lo stesso prototipo delle usuali function C preceduto da uno di questi qualificatori:

- `__global__` : per le function richiamate dall'host ed eseguite sul device, ovvero i kernel;
- `__device__` : per le function richiamate dal device ed eseguite sul device;
- `__host__` : (opzionale) per le function canoniche eseguite sulla CPU.

Per dichiarare un kernel CUDA: `__global__ void nomeKernel (parametri);`

7.3.1 Configurazione del kernel CUDA

es. `sommaGPU<<gridDim, blockDim>>(a_d, b_d, c_d, N);`

- `gridDim` è il numero di blocchi di ogni dimensione della griglia (1D, 2D o 3D)
- `blockDim` è il numero di thread di ogni dimensione del singolo blocco (1D, 2D o 3D)
- `dim3` è un tipo predefinito di CUDA, vettore di tre interi; ogni componente è accessibile attraverso i campi `x`, `y`, `z`; i campi non assegnati sono posti a 1:

`dim3 threadIdx`: identificativo del thread all'interno del blocco ;

`dim3 blockIdx`: identificativo del blocco all'interno della griglia;

`dim3 blockDim`: numero di thread contenuti in un singolo blocco;

.

7.3.2 Impostazione della memoria della GPU ad un dato valore

- `cudaError_t cudaMemset(void* devPtr,int value,size_t count)`
 - `devPtr` è un puntatore all'area di memoria del device da impostare;
 - `value` è il valore da assegnare a quest'area di memoria;
 - `count` è il numero di byte di quest'area di memoria.

7.3.3 Variabili architettura di memoria

Di seguito saranno elencate i tipi da assegnare alle avariabile per usare una determinata memoria:

- Global memory: `__device__` float variable;
- Shared memory: `__shared__` float variable;
- Registri: float variable;
- Locale memory: float variable[10];
- Constant memory: `__constant__` float variable;
- Texture memory:

```
texture<type, dim> text_var; // inizializzazione
cudaChannelFormatDesc(); // opzioni
cudaBindTexture2D(...): // bind (legare)
text2D(tex_var, x_index,y_index); //fetch (recupero)
```

7.4 Compilazione in cuda

Ogni file sorgente contenente estensioni CUDA deve essere compilato con un compilatore CUDA compliant: `nvcc` per CUDA C (NVIDIA)

Il compilatore processa il sorgente separando codice device dall'host. il codice host viene rediretto a un compilatore standard di default (ad es. `gcc`) il codice device viene tradotto in PTX, a partire dal PTX prodotto è possibile:

- produrre codice oggetto binario (`cubin`) specializzato per una particolare architettura GPU
- produrre un eseguibile che include PTX e/o codice binario (`cubin`)

Quando si specifica un'architettura virtuale, `nvcc` posticipa la fase di assemblaggio del codice PTX all'esecuzione dell'applicazione, quando l'architettura reale è nota. Ad esempio, il comando seguente genera un codice binario che funziona perfettamente se lanciato su architetture con compute capability 5.0 o successive:

```
nvcc x.cu gpu architecture =compute_50 gpu code=compute_50
```

7.4.1 Specifiche di compilazione

Al compilatore va sempre specificata:

- l'architettura virtuale con cui generare il PTX code;
- l'architettura reale per creare il codice oggetto (`cubin`):
 » `nvcc arch =compute_30 code=sm_30,sm_31`
- `nvcc` ammette l'uso dell'abbreviazione `arch sm_XX` , ad es:
 » `nvcc arch =sm_30` ; equivalente a: `nvcc arch =compute_30 code=sm_30`

7.4.2 Architettura virtuale e reale

- **L'architettura virtuale** è un'indicazione della compute capability che deve avere l'architettura reale su cui andrà eseguito il codice. Richiedere un'architettura virtuale meno performante, lascia più scelta per l'architettura reale su cui potrà essere eseguito il codice.
- **L'architettura reale** dovrebbe essere scelta quale la migliore possibile. Ovviamente ciò è possibile solo se è nota l'architettura fisica su cui sarà eseguito il codice.

Nota: Per chi usa google colabatory usare il doppio trattino nelle istruzioni successive esempio: `nvcc -arch=sm_30` invece di `nvcc -arch=sm_30`. In compilazione si può usare solo un'architettura virtuale e una lista di architetture reali.

7.5 Misura dei tempi: gli eventi

Gli eventi sono una sorta di marcatori che possono essere utilizzati nel codice per:

- misurare il tempo trascorso elapsed) durante l'esecuzione di chiamate CUDA (precisione a livello di ciclo di clock);
- bloccare la CPU fino a quando le chiamate CUDA precedenti l'evento non siano state completate.

7.5.1 Uso degli eventi

```
cudaEvent_tstart, stop;
cudaEventCreate(&start);
cudaEventCreate(&stop);
...
cudaEventRecord(start);
kernel<<<grid , block>>>(...);
cudaEventRecord(stop);
cudaEventSynchronize(stop);
/* assicura che tutti i thread siano arrivati all'evento stop prima di registrare il
   tempo*/
float elapsed;
// tempo tra i due eventi in millisecondi
cudaEventElapsedTime(&elapsed , start, stop);
...
cudaEventDestroy(start);
cudaEventDestroy(stop);
```

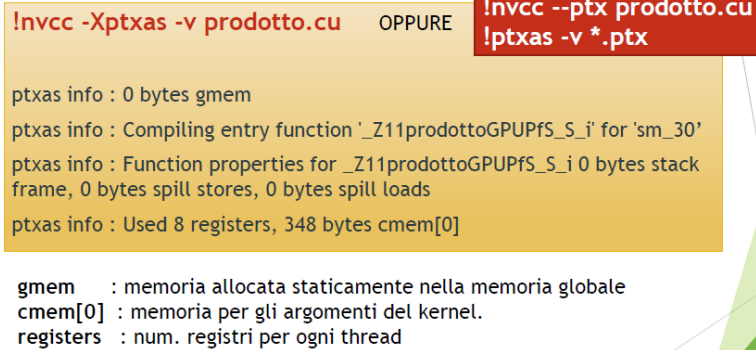
7.5.2 Ottimizzazione mediante profiling

Profiling significa analizzare le prestazioni del programma misurando:

- La complessità in spazio e(o in tempo dell'algoritmo;
- La frequenza e la durata delle chiamate a funzione.

In cuda in genere un'implementazione naïve non dà alte prestazioni, ma i tool di profiling aiutano a trovare colli di bottiglia da rimuovere.

Informazioni sulla memoria usata: `"nvcc -ptax-options=-v nomeFile.cu"` ci fornisce informazioni sulla memoria usata dal programma, come ad esempio i registri usati da ogni thread.



The screenshot shows a terminal window with a yellow background. At the top, two red boxes contain alternative compilation commands: `!nvcc -Xptxas -v prodotto.cu` and `!nvcc --ptx prodotto.cu !ptxas -v *.ptx`, separated by the word "OPPURE". Below these, the output of the compilation is shown in black text. At the bottom, a legend defines the terms used in the output: `gmem` for global memory, `cmem[0]` for kernel argument memory, and `registers` for the number of registers per thread.

```
!nvcc -Xptxas -v prodotto.cu    OPPURE    !nvcc --ptx prodotto.cu
!ptxas -v *.ptx

ptxas info : 0 bytes gmem
ptxas info : Compiling entry function '_Z11prodottoGPUPfS_S_i' for 'sm_30'
ptxas info : Function properties for _Z11prodottoGPUPfS_S_i 0 bytes stack
frame, 0 bytes spill stores, 0 bytes spill loads
ptxas info : Used 8 registers, 348 bytes cmem[0]

gmem      : memoria allocata staticamente nella memoria globale
cmem[0]   : memoria per gli argomenti del kernel.
registers : num. registri per ogni thread
```

Profiling in CUDA 5.0 e oltre

» `nvprof ./programma` È disponibile da CUDA 5 in poi. Attualmente sulla multiGPU è installata la versione di CUDA 4.0. Permette di valutare le singole chiamate a CUDA in termini di tempo.

Profiling versione precedenti CUDA 5.0

Per versioni di CUDA precedenti, impostare: » `%export COMPUTE_PROFILE=1` poi eseguire il programma normalmente, ad esempio: » `./tempi`. Così viene creato il file `cuda_profile_0.log`, contenente le informazioni sui tempi di trasferimento dati in microsecondi, del tipo riportato nella diapositiva successiva.

Nota: se si vuole usare `nvprof`, bisogna reimpostare `COMPUTE_PROFILE` a 0.

Capitolo 8

Libreria CUBLAS

Gran parte dei problemi del calcolo scientifico richiedono di risolvere uno o più problemi dell'algebra lineare numerica (ALN) :

- risoluzione di sistemi lineari;
- ricerca di autovalori e/o autovettori;
- calcolo della SVD (valori e vettori singolari).

La risoluzione di questi problemi ha generalmente un considerevole costo computazionale. Dunque, implementare in modo efficiente gli algoritmi per i problemi dell'algebra lineare è estremamente importante.

Gli algoritmi di algebra lineare numerica hanno in comune un insieme relativamente piccolo e stabile di operazioni di base e sono essenzialmente di tre tipi:

- $w = v1 \cdot v2$: prodotto scalare di due vettori (operazioni che lavorano su vettori e producono uno scalare);
- $w = A*v$: prodotto di una matrice per un vettore (operazioni tra vettori e matrici);
- $B = A1*A2$: prodotto di due matrici (operazioni tra matrici).

8.1 Libreria BLAS

La Basic Linear Algebra Subprogram è tra le prime librerie scritte (1979) per l'esecuzione di operazioni di base del calcolo matriciale che ottimizza gli accessi alla memoria.

8.1.1 3 livelli di Blas

Il fatto che le cpu e le memorie abbiano subito miglioramenti è rispecchiato anche nelle modifiche delle BLAS:

- 1979 BLAS1, concepite per eseguire solo operazioni scalari tra due vettori;
- 1988 BLAS2, dove si eseguono anche operazioni tra matrici e vettori;
- 1989 BLAS3, in cui l'operazione fondamentale è il prodotto tra due matrici.

8.2 Programmare con CUBLAS

La libreria blas è utilizzata da numerose librerie di livello più alto, perciò ne sono state prodotte diverse versioni, ottimizzate per varie piattaforme di calcolo. Noi vedremo nel dettaglio CUBLAS che è l'implementazione in CUDA C di BLAS.

Il modello di programmazione cublas si basa fondamentalmente su tre step principali:

- Creazione delle strutture dati in GPU (matrici o vettori). Per effettuare questa operazione è necessario prima di tutto allocare in GPU lo spazio necessario a contenere i dati, e poi sfruttare questo spazio per copiare i dati da CPU in una vettore o in una matrice attraverso le istruzioni `cublasSetVector()` e `cublasSetMatrix()`;
- Modifica dei dati caricati sul device attraverso le funzioni messe a disposizione dalle CUBLAS. Queste si dividono in BLAS di livello 1,2,3 come abbiamo visto nel dettaglio in precedenza e queste a loro volta sono suddivise in funzioni a singola precisione (operanti su float), doppia precisione (operanti su double) e funzioni su numeri complessi;
- Aggiornamento dei dati in CPU attraverso le primitive `cublasGetVector()` e `cublasGetMatrix()` e deallocazione dello spazio in GPU.

8.2.1 Funzione `cublasSetVector`

La funzione `cublasSetVector` copia `n` elementi da un vettore `x` nello spazio di memoria dell'host a un vettore `y` nello spazio di memoria della GPU.

```
cublasStatus_t cublasSetVector(int n, int elemSize, const void *x,  
                               int incx, void *y, int incy)
```

I parametri della funzione sono :

- `n` - il numero degli elementi da copiare dall'array `x`;
- `elemSize` - il numero di byte di ogni elemento dei vettori;
- `x` - puntatore all'array allocato sull'host;
- `incx` - la spaziatura di memorizzazione tra elementi consecutivi nell'array `x`;
- `y` - puntatore all'array allocato sulla GPU;
- `incy` - la spaziatura di memorizzazione tra elementi consecutivi nell'array `y`.

8.2.2 Funzione `cublasGetVector`

La funzione `cublasGetVector` copia `n` elementi da un vettore `x` nello spazio di memoria della GPU a un vettore `y` nello spazio di memoria dell'host.

```
cublasStatus_t cublasGetVector(int n, int elemSize, const void *x,  
                               void *y, int incy)  
                               int incx,
```

I parametri della funzione sono :

- `n` - il numero degli elementi da copiare dall'array `x`;
- `elemSize` - il numero di byte di ogni elemento dei vettori;
- `x` - puntatore all'array allocato sulla GPU;

- incx - la spaziatura di memorizzazione tra elementi consecutivi nell'array x;
- y - puntatore all'array allocato sull' host;
- incy - la spaziatura di memorizzazione tra elementi consecutivi nell'array y.

8.2.3 Funzione cublasSdot

La funzione cublasSdot calcola il prodotto scalare tra due vettori .

```
cublasStatus_t cublasSdot (cublasHandle_t handle, int n, const float *x,  
                           int incx, const float *y, int incy, float *result)
```

I parametri della funzione sono :

- handle - l'handle al contesto della libreria CUBLAS;
- n - numero degli elementi presenti nei vettori x e y;
- x - puntatore al vettore allocato sulla GPU;
- incx - la spaziatura di memorizzazione tra elementi consecutivi nell'array x;
- y - puntatore al vettore allocato sulla GPU;
- incy - la spaziatura di memorizzazione tra elementi consecutivi nell'array y;
- result - indirizzo della variabile nella quale sarà memorizzato il risultato.

8.2.4 Avvio delle CUBLAS

Per ogni programma è necessario avviare le cublas attraverso cublasCreate(&handle) prima di utilizzare qualsiasi operazione CUBLAS (in modo da creare un handle specifico della libreria per la gestione delle informazioni e relativo contesto in cui essa opera), il contesto così creato deve essere poi passato a tutte le successive chiamate di funzione di libreria e dovrebbe essere distrutto alla fine usando cublasDestroy(handle).

```
cublasHandle_t handle;  
cublasCreate(&handle);  
// il tuo codice  
cublasDestroy(handle);
```

8.2.5 Gestione degli errori in Cublas

CUBLAS inoltre è dotato di un sistema per recuperare e comprendere gli errori che avvengono in GPU durante l'esecuzione delle operazioni. Ogni funzione cuBLAS, infatti restituisce un oggetto status di tipo cublasStatus_t contenente le informazioni sui possibili errori. Di conseguenza è buona norma controllare lo stato dell'operazione prima di andare avanti:

```
if (cudaStat!=CUBLAS_STATUS_SUCCESS) {  
    printf ("CUBLAS error/n");  
    return EXIT_FAILURE;  
}
```

8.2.6 Valore di ritorno dell CUBLAS

tutte le funzioni della libreria cuBLAS restituiscono il loro stato, che può avere i seguenti valori:

Valore	Significato
CUBLAS_STATUS_SUCCESS	L'operazione è stata completata con successo.
CUBLAS_STATUS_NOT_INITIALIZED	La libreria cuBLAS non è stata inizializzata. Ciò è solitamente causato dalla mancanza di una precedente chiamata <code>cusblasCreate()</code> , da un errore nell'API Runtime CUDA richiamato dalla routine cuBLAS o da un errore nella configurazione dell'hardware.
CUBLAS_STATUS_ALLOC_FAILED	L'allocazione delle risorse non è riuscita nella libreria CUBLAS. Ciò è solitamente causato da un errore di <code>cudaMalloc()</code> .
CUBLAS_STATUS_INVALID_VALUE	Alla funzione è stato passato un valore o un parametro non supportato (una dimensione del vettore negativa, ad esempio).
CUBLAS_STATUS_ARCH_MISMATCH	La funzione richiede una caratteristica assente dall'architettura del dispositivo; solitamente causato da capacità di calcolo inferiori a 5.0.
CUBLAS_STATUS_MAPPING_ERROR	Un accesso allo spazio di memoria della GPU non è riuscito.
CUBLAS_STATUS_EXECUTION_FAILED	Il programma GPU non è stato eseguito. Ciò è spesso causato da un errore di avvio del kernel sulla GPU, che può essere causato da più motivi.
CUBLAS_STATUS_INTERNAL_ERROR	Un'operazione cuBLAS interna non è riuscita. Questo errore è solitamente causato da un errore di <code>cudaMemcpyAsync()</code> .
CUBLAS_STATUS_NOT_SUPPORTED	La funzione richiesta non è supportata.
CUBLAS_STATUS_LICENSE_ERROR	La funzionalità richiesta richiede una licenza ed è stato rilevato un errore durante il tentativo di controllare la licenza corrente

Tabella 8.1:

8.2.7 Compilare con CUBLAS

Per compilare un codice CUDA-C che usa la libreria CUBLAS, utilizzando il compilatore `nvcc` basta inserire un collegamento `-lcublas` per la libreria insieme alla compilazione del programma.

```
nvcc nomefile.cu -lcublas -o nomeprogramma
```

Poi per eseguirlo basta utilizzare: `./nomeprogramma`

8.3 Cenni sulle matrici in CUBLAS

8.3.1 Memorizzazione delle matrici per Cublas

Cublas per memorizzare le matrici adotta la stessa notazione adottata da Fortran e quindi memorizza gli elementi per colonne. Si supponga di avere la seguente matrice A:

$$\begin{bmatrix} 1 & 2 & 8 \\ 3 & 6 & 2 \\ 1 & 4 & 5 \\ 2 & 9 & 7 \end{bmatrix}$$

In C e C++ la memorizzazione avviene per righe, di conseguenza gli elementi di una stessa riga sono contigui in memoria.

1	2	8	3	6	1	2	4	5	2	9	7
A(0)	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)	A(11)

Per poter utilizzare le funzioni definite in cuBLAS, la stessa matrice deve essere invece memorizzata per colonne, quindi gli elementi presenti sulla stessa colonna saranno contigui:

1	3	1	2	2	6	4	9	8	2	5	7
A(0)	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	A(9)	A(10)	A(11)

Nel manuale viene fornita una macro per ovviare a questo problema nell'accesso all'elemento presente nella riga i e nella colonna j (per matrici già costruite, ovvero intervenendo a programma già avviato, non è possibile utilizzarla):

```
#define IDX2C(i,j,ld) (((j)*(ld))+(i)) // ld = numero di righe
```

In ogni caso se non si vuole rinunciare alla semantica di array multidimensionale si può considerare di chiamare le funzioni per le matrici trasposte.

8.3.2 Funzione cublasSetMatrix

La funzione cublasSetMatrix copia un riquadro di elementi da una matrice A nello spazio di memoria dell'host a una matrice B nello spazio di memoria della GPU.

```
cublasStatus_t cublasSetMatrix(int rows, int cols, int elemSize, const  
                               void *A, int lda, void *B, int ldb)
```

I parametri della funzione sono :

- rows - numero di righe da copiare dalla matrice A;
- cols - numero di colonne da copiare dalla matrice A;
- elemSize - il numero di byte di ogni elemento della matrice;
- A - puntatore alla matrice allocata sull'host;
- lda - numero di righe della matrice allocata per A anche se ne viene utilizzata solo una sottomatrice;
- B - puntatore alla matrice allocata sulla GPU;
- ldb - numero di righe della matrice allocata per B anche se ne viene utilizzata solo una sottomatrice.

8.3.3 Funzione cublasGetMatrix

La funzione cublasGetMatrix copia un riquadro di elementi da una matrice A nello spazio di memoria della GPU a una matrice B nello spazio di memoria dell'host.

```
cublasStatus_t cublasGetMatrix(int rows, int cols, int elemSize, const
                               void *A, int lda, void *B, int ldb)
```

I parametri della funzione sono :

- rows - numero di righe da copiare dalla matrice A;
- cols - numero di colonne da copiare dalla matrice A;
- elemSize - il numero di byte di ogni elemento della matrice;
- A - puntatore alla matrice allocata sulla GPU;
- lda - numero di righe della matrice allocata per A anche se ne viene utilizzata solo una sottomatrice;
- B - puntatore alla matrice allocata sull'host;
- ldb - numero di righe della matrice allocata per B anche se ne viene utilizzata solo una sottomatrice.

8.3.4 Funzione cublasSgemv

La funzione cublasSgemv calcola il prodotto matrice per vettore secondo la seguente formula:

$$y = \alpha * op(A) * x + \beta * y$$

```
cublasStatus_t cublasSgemv(cublasHandle_t handle, cublasOperation_t
                           trans, int m, int n, const float *alpha, const float *A,
                           int lda, const float *x, int incx, const float *beta,
                           float *y, int incy)
```

I parametri della funzione sono :

- handle - l'handle al contesto della libreria CUBLAS;
- trans - operazione sulla matrice A: con CUBLAS_OP_N allora $op(A) = A$ mentre con CUBLAS_OP_T allora $op(A) = A^T$;
- m: numero di righe della matrice e n: numero di colonne della matrice A;
- alpha - scalare utilizzato per la moltiplicazione;
- A - puntatore alla matrice allocata sulla GPU;
- lda - leading dimension della matrice A (numero di righe nel caso di memorizzazione per colonne);
- x - puntatore al vettore allocato sulla GPU;
- incx - lo spazio tra due elementi consecutivi nell'array x;
- beta - scalare utilizzato per la moltiplicazione;
- y - puntatore al vettore risultato allocato sulla GPU;
- incy - lo spazio tra due elementi consecutivi nell'array y.