

Prodotto tra una matrice sparsa ed un multivettore

Luca Capotombolo, Ludovico Zarrelli

1 Luglio 2023

Contents

1	Introduzione	2
1.1	Contesto	2
1.2	Matrici di test	2
1.3	Implementazione Seriale	2
2	Funzioni ausiliarie per il preprocessing dei dati	2
3	Discussione implementazione CSR	3
3.1	Funzioni ausiliarie per la memorizzazione nel formato CSR . . .	3
3.2	OpenMP	4
3.3	CUDA	4
4	Discussione implementazione ELLPACK	7
4.1	Funzioni ausiliarie per la memorizzazione nel formato ELLPACK	7
4.2	OpenMP	7
4.3	CUDA	7
5	Misurazione delle prestazioni	7
6	Suddivisione del lavoro	7
7	Istruzioni per la compilazione	7

1 Introduzione

1.1 Contesto

Il progetto prevede la realizzazione di un nucleo di calcolo per il prodotto tra una matrice sparsa ed un multivettore:

Nel progetto sono state utilizzate le matrici nel formato *MatrixMarket*. Per determinare la tipologia di matrice rappresentata nel file Matrix Market sono state utilizzate le funzioni di I/O disponibili all'indirizzo:

$$Y = AX$$

dove A è una matrice sparsa memorizzata nei formati *CSR* e *ELLPACK*. Il nucleo di calcolo è stato sviluppato in C ed è stato parallelizzato per sfruttare le risorse di calcolo disponibili con parallelizzazione *OpenMP* e *CUDA*.

1.2 Matrici di test

Per quanto riguarda le matrici di test, sono state scaricate nel formato *MatrixMarket* e sono state utilizzate le funzioni per l'I/O disponibili all'indirizzo:

<http://math.nist.gov/MatrixMarket/>

1.3 Implementazione Seriale

I risultati ottenuti dall'esecuzione parallela SpMM sono stati confrontati direttamente con i risultati ottenuti dall'esecuzione seriale. Per quanto riguarda le implementazioni seriali, sono state sviluppate tre different versioni:

- Prodotto seriale utilizzando il formato CSR
- Prodotto seriale utilizzando il formato ELLPACK con *zero padding*
- Prodotto seriale utilizzando il formato ELLPACK senza *zero padding*

Per avere una misura dell'errore, abbiamo deciso di calcolare due differenti errori che ci forniscono diverse informazioni:

- X : mi fornisce delle informazioni su
- Y : mi fornisce delle informazioni su

2 Funzioni ausiliarie per il preprocessing dei dati

La funzione *create_matrix_coo* ha il compito di leggere i dati da uno specifico file nel formato *MatrixMarket* per poi allocare e popolare la matrice nel formato *COO*. All'interno di questa funzione vengono gestite le possibili tipologie di matrici che sono supportate dal programma:

- *Simmetrica Pattern*
- *Simmetrica Reale*
- *Generale Pattern*
- *Generale Reale*

Le strutture dati che vengono utilizzate per la rappresentazione della matrice *COO* sono le seguenti:

- *I*: array contenente gli indici di riga degli elementi non zero
- *J*: array contenente gli indici di colonna degli elementi non zero
- *val*: array contenente i valori degli elementi non zero

In questo modo, l'elemento *i-esimo* di questi tre array rappresenta rispettivamente l'indice di riga, l'indice di colonna e il valore di uno specifico elemento non zero della matrice.

3 Discussione implementazione CSR

Passiamo alla discussione relativa al formato *CSR*. Vedremo gli aspetti riguardanti la conversione della matrice dal formato *COO* al formato *CSR* e gli aspetti che riguardano il prodotto SpMM sia con *OpenMP* che con *CUDA*.

3.1 Funzioni ausiliarie per la memorizzazione nel formato CSR

Le strutture dati che vengono utilizzate per il formato *CSR* sono le seguenti:

- *as*: è l'array dei coefficienti non zero
- *ja*: è l'array degli indici di colonna dei coefficienti non zero
- *irp*: è l'array dei puntatori all'inizio di ciascuna riga

Durante lo sviluppo del progetto, abbiamo avuto delle difficoltà nella conversione della matrice dal formato *COO* al formato *CSR*. Più precisamente, le nostre difficoltà erano dovute al grande tempo richiesto per eseguire la conversione per le matrici di grande dimensione.

La prima implementazione dell'algoritmo di conversione è rappresentata dalla funzione *coo_to_CSR_parallel* che si comporta bene per le matrici di piccole/medie dimensione ma ha un tempo di esecuzione elevato per le matrici di grande dimensione. Il primo step dell'algoritmo consiste nel computare il numero di non zeri per ogni riga popolando la struttura dati di appoggio *nz_per_row* che poi viene utilizzato per inizializzare la struttura dati *irp*. A questo punto, le tre strutture dati precedentemente elencate vengono popolate utilizzando un

doppio ciclo *for*. Per velocizzare l'esecuzione, questo blocco di codice è stato parallelizzato cercando di bilanciare il più possibile il carico di lavoro. Per liberare la memoria precedentemente occupata, al termine dell'inizializzazione delle tre strutture dati *as*, *ja* e *irp* vengono deallocate le strutture dati per la rappresentazione nel formato *COO*. (Da sistemare il codice con quel -1)

Il problema della prima implementazione è probabilmente rappresentato dalla presenza dei due cicli *for*. Per ridurre il tempo richiesto nella conversione dal formato *COO* al formato *CSR* in modo da trattare anche le matrici molto grandi, abbiamo deciso di implementare una seconda versione dell'algoritmo di conversione nella funzione *coo_to_CSR_parallel_optimization*. Il vantaggio di questa seconda versione consiste nell'utilizzo di un singolo ciclo che va ad iterare su tutti i non zeri della matrice. Per gestire la concorrenza, abbiamo introdotto una minima sezione critica che, comunque, ci consente di avere dei tempi di esecuzione molto inferiori rispetto alla versione precedente.

3.2 OpenMP

Il prodotto parallelo per il formato *CSR* con parallelizzazione *OpenMP* è implementato nella funzione *parallel_product_CSR*. La funzione *compute_chunk_size* ha il compito di calcolare la dimensione del chunk che deve essere assegnato a ciascun thread. Lo spazio delle iterazioni viene suddiviso a seconda del numero di processori che sono disponibili sul device. L'obiettivo è quello di assegnare un blocco di righe contiguo ai threads nello spazio di iterazione in modo da evita il più possibile il problema del *False Cache Sharing*. Ad ogni thread viene associato un certo numero di righe e ogni riga è associata ad un unico thread. In questo modo, posso utilizzare una variabile privata *partial_sum* per computare le somme parziali.

3.3 CUDA

Per quanto riguarda il prodotto parallelo per il formato *CSR* in *CUDA*, abbiamo deciso di scrivere differenti versioni del kernel. In realtà, le prime tre versioni del kernel potrebbero essere viste come la stessa versione con solamente delle modifiche minimali che le distinguono. Abbiamo deciso di riportare queste modifiche minimali differenziando così le tre versioni del kernel poiché hanno un grande impatto prestazionale. Poiché stiamo facendo calcolo ad alte prestazioni, è importante notare come queste differenze, che sembrano essere minimali, in realtà hanno un gran peso sulle prestazioni.

L'idea che c'è dietro alle prime tre versioni del kernel è quella di far sì che un singolo thread computi un singolo elemento della matrice finale *Y*. Inizialmente, viene allocata la memoria necessaria sul *CUDA Device* e vengono copiati i dati dall'host verso il device. Il numero di threads per blocco che viene utilizzato è pari a 1024. Nel kernel vengono calcolati gli indici di riga e di colonna che identificano l'elemento che deve essere calcolato dal thread. Come primo step viene calcolato l'identificativo globale del thread. Successivamente, vengono calcolati i due indici nel seguente modo:

$$row = tid / K$$

$$col = tid \% K$$

Una volta determinata la riga *row* della matrice sparsa *A* e la colonna *col* della matrice densa *X*, viene eseguito un controllo per verificare se il thread sta calcolando effettivamente un elemento della matrice. Questo controllo mi consente di gestire una griglia di blocchi che ha un numero totale di threads strettamente maggiore del numero di elementi della matrice che bisogna calcolare.

La prima versione del kernel non è ottimizzata poiché nel calcolo del singolo prodotto scalare si accede in memoria ogni volta che il thread modifica il risultato intermedio. Questi accessi continui portano a delle prestazioni molto basse. Un modo evidente per risolvere questo problema prestazionale consiste nello scrivere in memoria direttamente il risultato finale della computazione rappresentato dal valore della variabile *partial_sum*. Questa modifica permette di aumentare notevolmente le prestazioni del prodotto parallelo. A questo punto, è stata apportata un'ulteriore modifica minimale che consiste nel pre-computarsi le due variabili *start* e *end* e successivamente eseguire il ciclo. La terza versione riesce a guadagnare in termini prestazionali rispetto alla seconda versione.

Dopo aver implementato queste tre versioni del kernel abbiamo provato a distribuire la computazione richiesta per il calcolo di un singolo elemento della matrice. Questa quarta versione è stata ispirata al *CSR Vector* descritto nei vari paper. L'idea è quella di distribuire il calcolo di un elemento della matrice finale *Y* tra i threads di uno stesso warp. Più precisamente, ogni thread in un warp computa un risultato parziale per uno stesso elemento *y* della matrice *Y*. Viene eseguita una riduzione parallela che coinvolge i thread di uno stesso warp sfruttando la *shared memory*. Poiché tutti i threads di un warp contribuiscono al calcolo dello stesso elemento, è sufficiente che un solo thread tra essi scriva il risultato ottenuto dalla riduzione in memoria globale. Per fare ciò, abbiamo deciso che solamente il thread con identificativo 0 all'interno del warp ha il compito di scrivere il risultato in memoria globale. Tuttavia, questa nuova versione del kernel ci ha portato ad una notevole riduzione delle prestazioni. Il motivo per cui le prestazioni si sono ridotte così drasticamente è dovuto al numero molto piccolo di non zero per riga. Facendo un'analisi generale delle matrici di test, abbiamo potuto osservare come gran parte di esse abbiano un numero di non zero medio che è inferiore rispetto alla dimensione del warp. Questo implica il fatto che ci siano mediamente un numero non piccolo di thread all'interno di un warp che non danno un contributo effettivo al calcolo dell'elemento.

Nelle prime tre versioni, ogni elemento veniva computato da un singolo thread che eseguiva il prodotto scalare riga *row* per colonna *cols*. Nella versione successiva, invece di utilizzare un unico thread, abbiamo provato a distribuire la computazione tra i differenti thread di uno stesso warp. Tuttavia,

abbiamo visto che assegnando un singolo warp alla computazione di un singolo elemento y della matrice finale Y si ha una grande perdita nelle prestazioni. A questo punto, abbiamo pensato di seguire un approccio intermedio che consiste nel vedere i due casi implementati finora come dei *casi estremi*. Invece di assegnare un thread per elemento o un warp per elemento abbiamo provato a scegliere un numero intermedio di thread come compromesso tra i due differenti approcci.

Più precisamente, dato un *warp* di 32 thread, abbiamo creato 16 differenti *subwarps*, ognuno con il compito di computare un singolo elemento della matrice finale Y . Di conseguenza, ogni elemento y della matrice finale Y è computato da due differenti threads che costituiscono il subwarp. Nella versione *CSR_Vector_Sub_warp* si ragiona in termini di identificativo del subwarp. Come primo step si determina l'identificativo globale del thread per poi computare l'identificativo globale del *subwarp* a cui il thread appartiene. Poiché un *subwarp* ha una dimensione pari a 2, ogni thread che appartiene al *subwarp* ha un proprio identificativo locale al suo interno. Più precisamente, se la dimensione è pari a 2 allora i possibili indici dei threads all'interno di un subwarp sono 0 e 1. Dopo aver calcolato l'indice del subwarp è possibile determinare la riga e la colonna dell'elemento della matrice assegnato al subwarp. Poiché il subwarp si occupa di computare le somme parziali per un singolo elemento della matrice Y allora possiamo utilizzare direttamente una sola area di memoria associata ad ogni thread. Da notare che questo non sarebbe stato possibile se il subwarp avesse calcolato più elementi distinti della matrice finale Y . In questo caso, sarebbe stato necessario un meccanismo che permettesse di distinguere le somme parziali per i differenti elementi. In questo kernel viene utilizzata la *shared memory* per mantenere i risultati delle somme parziali. Più precisamente, ad ogni thread viene assegnata una specifica entry nell'array *vals* in cui mantenere i risultati delle somme parziali per il calcolo dell'elemento. Dopo aver calcolato i propri contributi, viene eseguita una riduzione parallela nella memoria condivisa. Infine, solamente il thread con indice 0 all'interno del subwarp scriverà il risultato in memoria globale.

- 4 **Discussione implementazione ELLPACK**
 - 4.1 Funzioni ausiliarie per la memorizzazione nel formato ELLPACK
 - 4.2 OpenMP
 - 4.3 CUDA
- 5 **Misurazione delle prestazioni**
- 6 **Suddivisione del lavoro**
- 7 **Istruzioni per la compilazione**