

Università di Bologna, Laurea in Ingegneria e Scienze Informatiche
Corso di High Performance Computing

Progetto d'esame 2020/2021

prof. Moreno Marzolla

Versione 1.0 del 27/11/2020

Prima versione di questo documento

1. L'operatore SKYLINE

Consideriamo un insieme $P = \{P_0, P_2, \dots, P_{N-1}\}$ di N punti in uno spazio a D dimensioni; il punto P_i ha coordinate $(P_{i,0}, P_{i,1}, \dots, P_{i,D-1})$. Diciamo che P_i domina P_j se si verificano entrambe le condizioni seguenti:

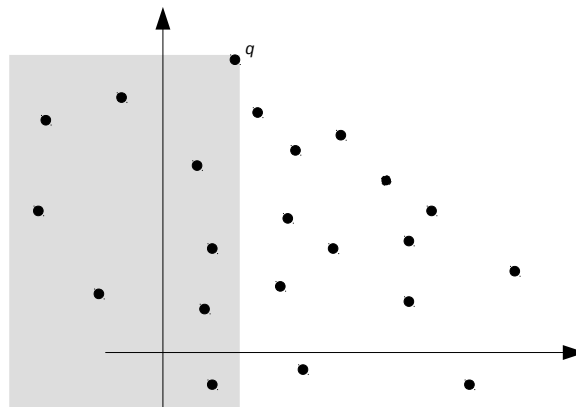
1. Per ogni dimensione $k = 0, \dots, D-1$, $P_{i,k} \geq P_{j,k}$
2. Esiste almeno una coordinata per cui la disuguaglianza di cui al punto precedente vale in senso stretto

(Si presti attenzione al fatto che in letteratura la relazione di dominanza viene talvolta definita con l'operatore \leq anziché \geq come facciamo noi). Detto in altri termini, P_i domina P_j se e solo se ogni coordinata di P_i ha valore maggiore o uguale alla corrispondente coordinata di P_j , ed esiste almeno una coordinata di P_i che abbia valore strettamente maggiore della corrispondente coordinata di P_j . Ad esempio, nel caso $D = 3$ si ha che $(3, 5, 4)$ domina $(3, 4, 4)$; $(3, 5, 4)$ domina $(2, -1, 3)$. Si noti che data una coppia di punti, non è detto che uno dei due domini l'altro. Ad esempio, se $s = (3, 5, 4)$ e $t = (3, 6, 3)$, si ha che s non domina t e t non domina s .

Lo skyline $Sk(P)$ dell'insieme P è composto da tutti i punti che non sono dominati da alcun altro punto (si noti che per definizione un punto non domina se stesso):

$$Sk(P) = \{s \in P : \text{non esiste alcun punto } t \text{ in } P \text{ tale che } t \text{ domini } s\}$$

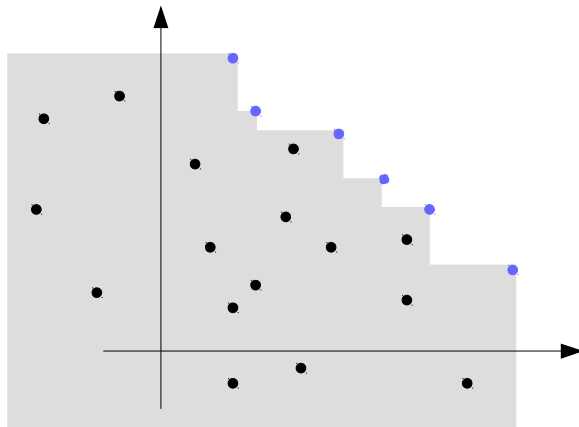
Per capire in cosa consiste l'operatore skyline, è utile considerare un semplice esempio in due dimensioni. La figura seguente rappresenta un insieme di punti sul piano:



I punti dominati da q sono quelli che si trovano all'interno dell'area grigia "proiettata" dal punto q verso il basso e verso sinistra; per definizione, nessuno dei punti nell'area grigia fa parte dello skyline, in quanto sono tutti dominati da un altro punto (q in questo caso).

L'operatore skyline viene utilizzato in diversi problemi di ottimizzazione ed è fornito come estensione del linguaggio SQL [1, 2]. Esistono diversi algoritmi per determinare lo skyline. In questo progetto consideriamo l'algoritmo banale che procede per eliminazione: si considerano uno

alla volta tutti i punti dell'insieme P , e per ogni punto q si eliminano da P tutti i punti da esso dominati. I punti che rimangono alla fine di questo processo di eliminazione sono tutti e soli quelli che fanno parte dello skyline. Nell'esempio in due dimensioni, lo skyline è composta dai punti in blu:



L'area grigia è composta dall'unione delle proiezioni di tutti i punti presenti; quindi, i punti che cadono all'interno dell'area grigia sono dominati da qualche altro punto (in particolare, da almeno uno dei punti blu), e non possono far parte dello skyline.

L'input del programma da sviluppare consiste in una matrice $P[0..N-1, 0..D-1]$, dove ogni riga della matrice rappresenta le D coordinate di un punto. Lo pseudocodice seguente determina e stampa lo skyline applicando il meccanismo appena descritto:

```

SKYLINE(  $P_0, P_1, \dots, P_{N-1}$  )
  bool S[0..N - 1];
  integer i, j;
  for i ← 0 to N - 1 do
    S[i] ← true;
  endfor
  for i ← 0 to N - 1 do
    if S[i] then
      for j ← 0 to N - 1 do
        if (S[j] and  $P_i$  domina  $P_j$ ) then
          S[j] ← false;
        endif
      endfor
    endif
  endfor
  for i ← 0 to N - 1 do
    if S[i] then
      print  $P_i$ 
    endif
  endfor

```

Al termine dell'esecuzione, $S[i]$ è *true* se e solo se il punto i -esimo fa parte dello skyline.

L'algoritmo proposto non è il più efficiente possibile per calcolare lo skyline, ma ha il vantaggio di essere comprensibile. Lo scopo di questo progetto non è di migliorare lo stato dell'arte, ma di dimostrare la capacità di parallelizzare in modo efficiente un algoritmo utile in pratica.

2. Specifica del progetto

Scopo del progetto è la realizzazione di due implementazioni parallele dell'algoritmo per il calcolo

dello skyline visto nella sezione precedente. La prima deve essere realizzata usando OpenMP; la seconda deve essere realizzata, a scelta, usando MPI oppure CUDA (uno dei due, non entrambi). I nomi degli eseguibili devono essere `omp-skyline` (per la versione OpenMP), `mpi-skyline` (per la versione MPI) e `cuda-skyline` (per la versione CUDA). Il programma viene lanciato dalla riga di comando nel modo seguente (si fa l'esempio della versione OpenMP; le altre versioni sono analoghe):

```
./omp-skyline < input_file > output_file
```

Il programma deve leggere l'input da standard input e scrivere il risultato su standard output. Nel caso di MPI, il programma verrà lanciato usando `mpirun`; il processo 0 (il master) dovrà leggere l'input e stampare l'output, in quanto è l'unico che ha accesso a `stdin/stdout`.

L'input è costituito da $N + 2$ righe di testo: la prima riga contiene il numero di dimensioni D , seguito da eventuali altri caratteri che devono essere ignorati fino alla fine della riga; la seconda contiene il numero di punti N ; le successive N righe contengono ciascuna le D coordinate reali di un punto, separate da spazi o tab.

D [eventuali altri caratteri da ignorare]

N

P_{00} P_{01} P_{02} ... $P_{0\ D-1}$

P_{10} P_{11} P_{12} ... $P_{1\ D-1}$

...

$P_{N-1\ 0}$ $P_{N-1\ 1}$ $P_{N-1\ 2}$... $P_{N-1\ D-1}$

Al termine dell'esecuzione, il programma deve stampare su standard output le coordinate dei punti dello skyline, nell'ordine in cui i punti compaiono nel file di input. L'output deve avere lo stesso formato dell'input: se lo skyline è composto da $K \leq N$ punti, l'output del programma conterrà $K + 2$ righe di testo; nella prima riga è presente il numero di dimensioni D ; sulla seconda comparirà il valore K ; nelle K righe successive sono elencate le coordinate dei punti che fanno parte dello skyline, separate da spazi. I punti dello skyline possono essere elencati in un ordine qualsiasi.

Si tenga inoltre presente quanto segue:

- Si assuma $D \geq 2$
- Si assuma $N \leq 1024 \times 1024$; il valore di N sarà comunque sempre scelto in modo tale da non esaurire la memoria a disposizione (sia nella versione per CPU che in quella per GPU).
- Verranno accettati programmi che impongono vincoli sulla dimensione del dominio (es., che funzionano correttamente solo se il numero di punti è multiplo di...); la relazione deve indicare chiaramente l'esistenza di tali vincoli. Ovviamente programmi soggetti a vincoli riceveranno una valutazione inferiore rispetto a quelli che funzionano correttamente nel caso generale.

3. Cosa consegnare

Viene richiesto di consegnare due versioni parallele del programma:

1. La prima, chiamata `omp-skyline.c`, deve utilizzare OpenMP;
2. La seconda versione deve utilizzare, a scelta, MPI oppure CUDA (uno dei due, non entrambi). Il file deve chiamarsi rispettivamente `mpi-skyline` oppure `cuda-skyline`

Oltre ai due programmi di cui sopra, è obbligatorio includere:

3. Una relazione in formato PDF della lunghezza massima di 6 facciate che descriva le strategie di parallelismo adottate e discuta la scalabilità ed efficienza dei programmi realizzati.

Ulteriori requisiti:

- I sorgenti devono essere adeguatamente commentati. All'inizio di ogni file deve essere indicato cognome, nome e numero di matricola dell'autore/autrice.
- Includere un file README contenente le istruzioni per la compilazione e l'esecuzione dei programmi consegnati; chi lo desidera può usare un Makefile per la compilazione.
- I programmi verranno compilati e testati sul server `isi-raptor03.csr.unibo.it`.
- Tutti i programmi devono compilare senza *warning*. Per le versioni OpenMP e MPI suggerisco di usare i flag `-std=c99 -Wall -Wpedantic` del compilatore, come per i programmi visti durante il corso.
- La relazione non deve superare la lunghezza di sei facciate in formato A4, contando tutte le pagine (inclusi eventuali frontespizi). Il formato della relazione è libero, ma nell'archivio contenente lo schema di progetto è presente uno schema in formato LibreOffice che a propria scelta può essere usato come base. La relazione non deve illustrare il codice riga per riga (per quello ci sono già i sorgenti), ma discutere le strategie di parallelizzazione adottate, e illustrare la scalabilità dei programmi mostrando i grafici di speedup ed efficienza. Nella relazione va indicato il proprio cognome, nome e numero di matricola.

4. Suggerimenti

Il tempo di esecuzione dell'algoritmo SKYLINE dipende sia dai dati di input che dal risultato (questo tipo di algoritmi si dicono *output-sensitive*). In particolare, dipende dal numero N di punti, dalla dimensione D dello spazio dei punti, e dal numero K di punti che fanno parte dello skyline (è parte del progetto determinare, se la si ritiene necessaria, l'espressione del costo asintotico dell'algoritmo).

Il programma seriale fornito va inteso come un possibile punto di partenza; è possibile modificarlo per adattarlo alle proprie esigenze, oppure partire da zero con una implementazione diversa. L'unica cosa che non può cambiare è il formato dell'input e dell'output.

Nella directory `datasets/` vengono forniti alcuni file di input che è possibile utilizzare per testare l'algoritmo. Le caratteristiche dei file di input sono riassunte nella seguente tabella:

File	Numero punti (N)	Dimensioni (D)	Punti skyline (K)
circle	1000	2	15
test1	100000	3	26
test2	100000	4	10352
test3	100020	10	24892
test4	100009	8	17458
test5	100000	20	96973
test6	100100	50	100050
test7	100400	200	100200
worst	100000	10	100000

Alcuni input potrebbero richiedere molto tempo per essere elaborati; non è obbligatorio usarli tutti, quindi è possibile limitarsi a quelli che possono essere gestiti in tempi ragionevoli sulla propria macchina. Nel caso in cui si vogliano generare input di dimensioni diverse, si faccia riferimento ai comandi presenti nel Makefile. La maggior parte dei file sono stati generati usando il comando

rbox che è parte del pacchetto `qhull-bin` (già installato sul server).

Invito a prestare la massima attenzione alla stesura della relazione. Una relazione scadente è uno dei motivi più frequenti di penalizzazione della valutazione dei progetti. Vanno assolutamente evitati errori grossolani come “~~ghost-shell~~” invece di “ghost cell”, “~~pull-di-thread~~” invece di “pool di thread”. Anche se la relazione non è una tesi di laurea, invito a prendere visione dei [consigli per la stesura della tesi](#) sulla mia pagina Web.

5. Modalità di svolgimento del progetto

- Il progetto deve essere svolto individualmente. Non è consentito condividere il codice o la relazione con altri studenti.
- È ammesso l'uso di porzioni di codice reperito in rete o tramite altre fonti purché (i) la provenienza di ogni frammento di codice scritto da terzi sia chiaramente indicata in un commento, e (ii) la licenza di tale codice ne consenta il riutilizzo. Il codice usato dal docente durante le lezioni o i laboratori può essere usato liberamente senza necessità di indicarne la fonte.
- Sono disponibile a fornire chiarimenti sulle specifiche del progetto (cioè sul presente testo), ma non guarderò il vostro codice per nessun motivo se non dopo la consegna; in particolare, non farò debug del vostro codice. Lo svolgimento di questo progetto è una parte importante dell'esame e va quindi affrontato con la massima serietà.

6. Consegna e validità del progetto

Il progetto può essere consegnato in qualsiasi momento, prima o dopo aver sostenuto la prova scritta. Le eventuali valutazioni positive (del progetto e/o della prova scritta) restano valide fino alla sessione d'esame di settembre 2021 inclusa; dopo tale data inizierà la nuova edizione del corso e tutti i voti in sospeso saranno persi.

Il progetto si consegna una volta sola. Non è possibile apportare modifiche o correzioni dopo la consegna: chi vuole migliorare la valutazione dovrà consegnare un nuovo progetto su nuove specifiche.

La consegna deve avvenire tramite la piattaforma <https://virtuale.unibo.it/>, mediante la quale è possibile caricare un unico archivio in formato `.zip` oppure `.tar.gz` contenente sia i sorgenti che la relazione. L'archivio dovrà essere denominato con il cognome e nome dell'autore (es., `MarzollaMoreno.zip` oppure `MarzollaMoreno.tar.gz`), e dovrà contenere una directory con lo stesso nome (es., `MarzollaMoreno/`) contenente a sua volta i file secondo il seguente layout:

`MarzollaMoreno/src/omp-skyline.c`

`MarzollaMoreno/src/mpi-skyline.c` (se si svolge la versione MPI)

`MarzollaMoreno/src/cuda-skyline.cu` (se si svolge la versione CUDA)

`MarzollaMoreno/src/Makefile` (facoltativo)

`MarzollaMoreno/README` (facoltativo)

`MarzollaMoreno/Relazione.pdf`

Includere ogni altro file necessario alla compilazione del codice, come ad esempio `hpc.h` o simili.

Prestare attenzione che la piattaforma Virtuale limita la dimensione dei file caricati a 20MB. **Non allegare i file di input forniti**: sono molto voluminosi e in ogni caso ne ho già una copia.

7. Valutazione dei progetti

Un progetto verrà considerato **sufficiente** se soddisfa almeno i seguenti requisiti minimi:

- I programmi compilano ed eseguono correttamente su istanze di input anche soggette a vincoli (es., dimensione del dominio multipla di...) sul server `isi-raptor03.csr.unibo.it`.
- La relazione dimostra un livello sufficiente di padronanza degli argomenti trattati.

Ulteriori aspetti che potranno comportare una valutazione superiore:

- Qualità del codice, in termini di correttezza, chiarezza, ed efficienza (es., uso appropriato delle primitive e/o dei pattern di programmazione concorrente adeguati).
- Generalità (es., il programma funziona con qualunque dimensione del dominio).
- Qualità della relazione, in termini di correttezza, chiarezza, e presentazione. Per ottenere una buona valutazione è richiesto che la relazione contenga lo studio di speedup ed efficienza dei programmi realizzati. Non è obbligatorio misurare i tempi di esecuzione sul server. Verrà valutata la capacità di interpretare i risultati ottenuti piuttosto che i risultati stessi.

Sebbene il progetto possa essere consegnato in qualsiasi momento, effettuerò tre sessioni di valutazione al termine delle sessioni d'esame di gennaio/febbraio 2021, giugno/luglio 2021 e settembre 2021. Questo significa che alla fine di febbraio 2021, luglio 2021 e settembre 2021 valuterò tutti i progetti ricevuti fino a quel momento. Chi avesse delle scadenze, ad esempio legate a richieste di borse di studio o per potersi laureare, è pregato di segnalarmelo all'atto della consegna. L'unico vincolo è che il progetto venga consegnato almeno **10 giorni lavorativi prima della data entro la quale si richiede la correzione** (sottolineo **lavorativi**) per consentirmi di far fronte anche ad eventuali altri miei impegni accademici.

8. Checklist per la consegna

Prima di consegnare il progetto, assicurarsi che i requisiti fondamentali elencati nella lista seguente siano soddisfatti:

<input type="checkbox"/>	Vengono consegnate due versioni del programma, una che usa OpenMP, e una che usa (a scelta) MPI oppure CUDA.
<input type="checkbox"/>	I sorgenti compilano ed eseguono correttamente sul server <code>isi-raptor03.csr.unibo.it</code> .
<input type="checkbox"/>	La relazione è in formato PDF e ha lunghezza minore o uguale a 6 facciate.
<input type="checkbox"/>	I sorgenti e la relazione indicano chiaramente cognome, nome e numero di matricola dell'autore/dell'autrice.
<input type="checkbox"/>	Il progetto viene consegnato in un unico archivio <code>.zip</code> oppure <code>.tar.gz</code> , nominato con nome e cognome dell'autore, che include i sorgenti e la relazione in formato PDF.

9. Riferimenti bibliografici

[1] S. Borzsony, D. Kossmann and K. Stocker, "The Skyline operator," *Proceedings 17th International Conference on Data Engineering*, Heidelberg, Germany, 2001, pp. 421-430, doi: 10.1109/ICDE.2001.914855.

[2] Wikipedia, *Skyline Operator* https://en.wikipedia.org/wiki/Skyline_operator