Esercitazione 3

Obiettivi

- pattern per la parallelizzazione di task
- utilizzare primitive per la gestione di processi (fork)
- comunicazione tra processi tramite pipe
- strategie di serializzazione dei dati
- evitare ritardi in lettura/scrittura mediante IO asincrono e select

Premessa

Questo laboratorio utilizza le primitive fork() e pipe() dei sistemi operativi Unix-like. Pertanto è necessario servirsi di un ambiente di esecuzione adeguato (una virtual machine con Linux, ad esempio). A meno di utilizzare system call particolari, il codice dovrebbe funzionare senza grossi problemi all'interno di Mac OSX.

Introduzione

MapReduce è un pattern utilizzato per processare grandi dataset quando il tipo di analisi da effettuare è parallelizzabile.
Il sistema è composto da tre tipi di attori:

1. un coordinatore
2. un pool di processi "mapper", che ricevono dal coordinatore task di lavoro in parallelo e restituiscono al coordinatore una lista di risultati, ognuno associato ad una chiave
   ```
   task -> [(key_1, result_1),..., (key_n,result_n)]
   ```
3. un pool di processi "reducer", che ricevono dal coordinatore i risultati dei mapper, assieme a un accumulatore che si ricorda "la riduzione" dei risultati precedenti associati alla stessa chiave, e restituiscono un accumulatore aggiornato con il valore ricevuto
   ```
   (key, result, acc)  -> (key, acc)
   ```

Questo pattern, ad esempio, può essere usato per contare la frequenza di simboli in un flusso di dati.
Prendiamo questa sequenza di test:

```
aaa bbb aaa ccc bbb aaa aaa
ccc bbb bbb aaa ccc bbb aaa
```

Il coordinatore leggerà una riga per volta e la passerà a un mapper, che restituirà una lista di risultati formati da coppie ordinate (simbolo, occorrenze). Per l'esempio precedente avremo:

```
aaa bbb aaa ccc bbb aaa aaa -> [(aaa, 4), (bbb, 2), (ccc, 1)]
ccc bbb bbb aaa ccc bbb ddd -> [(aaa, 1), (bbb, 3), (ccc, 2), (ddd, 1)]
```

Con due o più mapper l'elaborazione delle linee può essere fatta in parallelo distribuendo, a turno, ogni nuovo task ad un mapper differente.

Man mano che il coordinatore riceve i risultati, li passa ai reducer, assieme all'accumulatore corrispondente (che in questo caso esegue la somma, partendo dal valore iniziale zero). Continuando con i dati dell'esempio:

```
(aaa, 4, 0) -> (aaa, 4)
(bbb, 2, 0) -> (bbb, 2)
(ccc, 1, 0) -> (ccc, 1)
(aaa, 1, 4) -> (aaa, 5)
(bbb, 3, 2) -> (bbb, 5)
(ccc, 2, 1) -> (ccc, 3)
(ddd, 1, 0) -> (ddd, 1)
```

Leggendo il valore degli accumulatori abbiamo in tempo reale il risultato per ogni chiave.

Problema

Realizzare un sistema semplificato di MapReduce costituito da un coordinatore e due processi figli, un mapper e un reducer che comunicano con il coordinatore mediante pipe.

L'input è da processare è un file di log da leggere per righe (allegato a parte), mentre le funzioni di map e reduce devono poter essere facilmente specificabili con delle funzioni che possono essere passate come lambda

Formato del file di log

```
<indirizzo ip> - - [<data ora>] "<metodo> <url> <protocollo>" <codice risultato>
<dimensione risultato>

Esempio:
209.17.96.34 - - [01/Jan/2020:00:19:59 +0100] "GET / HTTP/1.1" 200 20744
194.55.187.131 - - [01/Jan/2020:00:44:56 +0100] "GET /manager HTTP/1.1" 404 1999
```

Le funzioni map e reduce sono definite nel seguente modo generico:

```
template<class MapperInputT, class ResultT>
std::vector<ResultT> map(const MapperInputT& input);

template<class ReducerInputT, class ResultT>
ResultT reduce(const ReducerInputT &input);
```

La classe MapperInputT wrappa l'input del file (di solito una stringa)
La classe ResultT contiene due attributi: una chiave e un valore (key, result)
La classe ReducerInputT contiene tre attributi: chiave, valore da ridurre, accumulatore corrente (key, value, acc)

1. Inizialmente implementare in modalità single process il processo di map reduce.
   Possibile pseudo codice:

```
function mapreduce(input, mapfun, reducefun){
      accs = {}
      while line = readline(input){
            results = mapfun(MapperInputT(line))
            for key, result in results {
                  new_key, new_acc = reducefun(
                        ReducerInputT(key, result, accs[key]));
                  accs[new_key] = new_acc
            }
      }
      return accs;
}
```

2. Realizzare le funzioni di map e reduce per:
   - contare quante richieste ci sono state per ogni indirizzo ip
   - contare quante richieste ci sono state per ogni ora del giorno (ignorando la data)
   - contare le url più visitate
   - per ogni codice di risultato (200, 201, …, 404, …, 500,...) raggruppare gli ip che hanno fatto richieste ottenendo quel codice e stampare quelli che hanno tentato attacchi richiedendo url con errori (codice 400, 404, 405). (suggerimento: la chiave può essere costruita ad hoc, tipo “codice-ip”)
3. testare con il file di input fornito, tenendo conto dei tempi di esecuzione
4. Il passo successivo consiste nel suddividere il lavoro su tre processi, un coordinare, un mapper e un reducer utilizzando la funzione fork() per creare i sotto-processi e collegandoli tra loro tramite delle pipe. Per fare ciò, prima di tutto occorre ragionare e ri-organizzare la struttura del codice.
   Su un piano concettuale, cosa è necessario cambiare per far comunicare i tre moduli?
   Posso invocare le funzioni direttamente? Perché?
   Attraverso una pipe è possibile passare direttamente gli oggetti MapperInputT, ReducerInputT o ResultT? Sarebbe possibile passa un oggetto string o un vector?
5. Per prima cosa, ci occupiamo di rendere possibile la comunicazione.
   Con le pipe abbiamo due file descriptor e non esiste un modo portabile per ottenere da questi degli input e degli output stream C++. Pertanto per comunicare occorre usare le le funzioni di basso livello write e read e trasformare gli oggetti in sequenze di byte facilmente analizzabili e che permettano di sapere esattamente quando inizia e finisce ogni oggetto e ogni suo attributo.
   Le possibilità sono molteplici:
   - Una codifica usando qualche formato standard, come ad esempio Protocol Buffer (https://developers.google.com/protocol-buffers) o JSON. Per JSON è possibile definire un oggetto corrispondente ad ogni parametro e sfruttare il fatto che si possono serializzare in una singola riga, quindi ogni oggetto scritto sulla pipe sarà separato da un newline (il JSON garantisce che dentro il messaggio non vi siano newline) Ad esempio per l’array di risultati:

`[ {"key": "val1", "result": {...}}, {"key": "val2", "result": {...}}, … ]`
Serializzare/ deserializzare gli oggetti in JSON usando libreria JSON di boost https://github.com/CPPAlliance/json

- Si possono serializzare gli oggetti e i loro contenuti tramite sequenze di byte. Per realizzare la serializzazione definiamo per MapperInputT, ReducerInputT o ResultT delle funzioni di serializzazione:

```
std::vector<char> serialize() const;
void deserialize(std::shared_ptr<char*>);
```

`serialize()` scriverà nel vettore in uscita il contenuto dell'oggetto corrente in questo modo (le lunghezze vanno scritte in binario!):
`<lunghezza totale><array attributi in ordine: <lunghezza attributo> <attributo>>`

Opzionale: provare ad implementare le funzioni serialize / deserialize come interfaccia usando il Curiously Recurring Template Pattern

6. Misurare i tempi delle serializzazioni e confrontarli. Attenzione: testare le funzioni di serializzazione a parte, non scrivendo ancora sulle pipe

7. Il passo successivo è scrivere e leggere gli oggetti sulle pipe. Da qui in avanti useremo la serializzazione binaria che abbiamo definito al passo precedente. Per comunicare possiamo usare solo i metodi di basso livello read e write, che normalmente sono bloccanti e quindi è facile fare una comunicazione sincrona con mapper e producer. Implementare in C++ il seguente pseudocodice:

```
function mapreduce(input_file, mapfun, reducefun){
      // create mapper_pipe and reducer_pipe
      // fork mapper and reducer
      accs = {}
      while line = readline(input_file){
            // MapperInputT serve per rendere serializzabile line,
            // che presumibilmente sarà una stringa
            write_to_pipe(mapper_pipe, MapperInputT(line))
            results = read_from_pipe(mapper_pipe)
            for key, result in results {
                  write_to_pipe(
                        reducer_pipe,
                        MapperInputT(key, result, accs[key]))
                  new_key new_acc = read_from_pipe(reducer_pipe)
                  accs[new_key] = new_acc
            }
      }
      return accs;
}
```

Esempio di definizione delle funzioni read e write sulle pipe:

```
template<class T>
void write_to_pipe(int fd, const T& obj);
template<class T>
T read_from_pipe(int fd);
```

Suggerimenti:

- nella la write usare reinterpret_cast per ottenere un char* da un vettore
- nella read leggete prima la lunghezza dell'oggetto e poi in una seconda read esattamente i byte indicati

Quale problema presenta questo approccio sincrono in termini di prestazioni?

8. Per sfruttare in modo opportuno il possibile parallelismo, il coordinatore dovrà occuparsi rifornire il più velocemente possibile di lavoro i due task subordinati, inviando al mapper i dati presenti nel file di ingresso e al reducer i dati via via prodotti dal mapper senza fermarsi ad aspettare i singoli risultati.
   Poiché i tempi di elaborazione dei singoli compiti non sono prevedibili e poiché le pipe hanno una capacità limitata di contenere dati "in transito", il coordinatore dovrà sforzarsi di leggere, il più frequentemente possibile, da ciascuno dei canali in ingresso dei dati e riversarli nel corrispondente canale di uscita, senza attendere se non vi sono dati o non c'è un messaggio completo.

   Occorre però evitare di rimanere bloccati quando il canale di uscita è già "saturo". Di conseguenza, tutte le operazioni di I/O, read e write, devono essere effettuate in modo non bloccante. Si legga attentamente il documento https://www.geeksforgeeks.org/non-blocking-io-with-pipes-in-c/ allo scopo di capire meglio la natura del problema.

   Inoltre vedere in dettaglio il comportamento di write e read non bloccanti sulle pipe:
   https://linux.die.net/man/3/write
   https://linux.die.net/man/3/read

   Come si comportanto read e write quando il buffer è vuoto o pieno? Che succede ad esempio se in una read nel buffer vi sono meno byte di quelli richiesti o se è vuoto? Quali modifiche al codice precedente sarebbero necessarie necessarie?
   (Non implementare codice a questo passo, limitarsi a descrivere i problemi da superare)

9. Per gestire la complessità dell'IO asincrono si può utilizzare la **select** (o sue implementazioni alternative come pselect, poll o epoll), che permette di testare prima su quali file descriptor si può leggere e scrivere senza rimanere bloccati e poi fare le operazioni di IO solo nelle pipe corrispondenti. In questo modo si possono preservare i benefici di read e write bloccanti, senza pagarne troppo il prezzo in termini di prestazioni.
   Per il suo utilizzo fare riferimento a questo link:

https://www.tutorialspoint.com/unix_system_calls/_newselect.htm

Il meccanismo base della select è che riceve in ingresso un set di file descriptor da osservare (fd_set) e un timeout. La funzione ritorna:

- -1 in caso di errore
- 0 se è scaduto il timeout senza che nessun file descriptor potesse essere usato in modo non bloccante
- un valore maggiore di zero altrimenti. In questo caso fd_set contiene la lista dei file descriptor la cui read e write non è bloccante, e si possono testare con la macro FD_ISSET

Occorre pertanto riorganizzare il codice del coordinatore per fare sì che lettura e scrittura ad ogni ciclo siano opzionali e fatte soltanto se in quel momento la pipe non è bloccante.
Trasformare il codice sincrono precedente scritto, seguendo questo pseudocodice

```
input_queue = [ ]
result_queue = [ ]
while (true) {
        fd_set = (input, mapper_pipe_in, mapper_pipe_out, reducer_pipe_in,
                reducer_pipe_out)
        ret = select(&fd_set, time_out)
        switch(ret){
                case -1:
                        // handle error
                        break;
                case 0:
                        // nothing to do
                        break
                default:
                        if(is_set(input, fd_set) && input_queue.lenhgth < MAX_LEN){
                                // read from input and queue in input_queue
                        }
                        // test mapper input and write input if any
                        // test mapper output, read and queue results
                        // test reducer input and write queue results if any
                        // test reducer output and read results
        }
```

10. Misurare il tempo totale di esecuzione e confrontarlo con quello single process. Discutere il risultato
11. Infine gestire la terminazione corretta del pattern MapReduce, in quanto il padre non ha nozione di quando i figli hanno terminato il loro carico di lavoro. Quali strategie adottare?
    Ricordare che, nell'ordine:

- i figli devono sapere che non c’è più input da processare
- il padre deve sapere quando i figli non hanno più task residui da inviare.